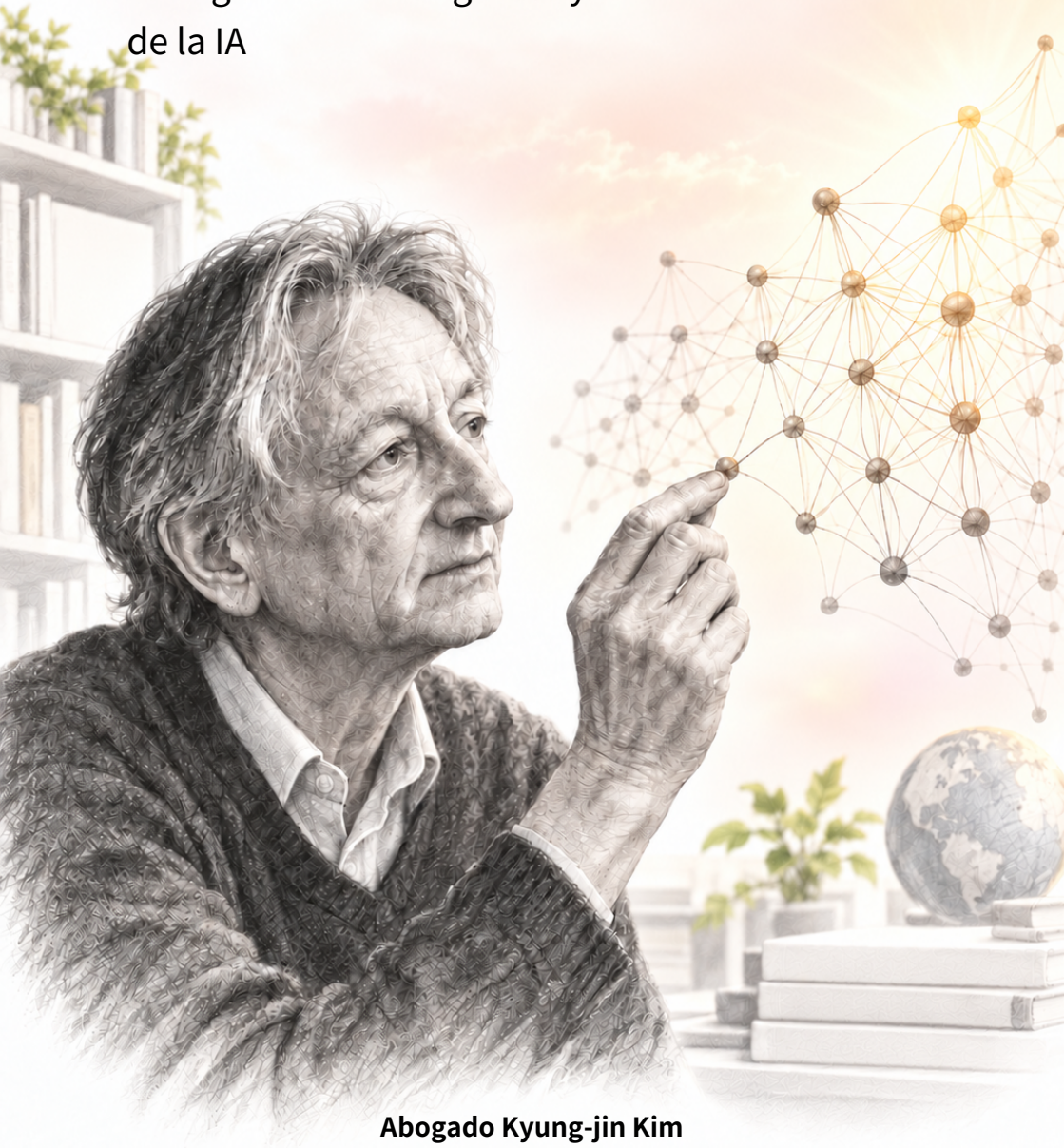
Geoffrey Hinton

El origen de la inteligencia y reflexiones en la era de la IA



Abogado Kyung-jin Kim

El hombre que concibió las máquinas que aprenden

Geoffrey Hinton es un investigador que encontró la clave del aprendizaje automático en el modo en que aprende el cerebro humano. Concibió máquinas que aprenden por sí mismas a partir de los datos, en lugar de que las personas introduzcan las reglas una por una. Incluso en una época en que la investigación sobre redes neuronales no recibía atención en el mundo académico, nunca abandonó esa pregunta.

Este libro parte de la curiosidad de su infancia y sigue las investigaciones realizadas junto a sus colegas y discípulos. Examina el proceso por el cual las redes neuronales hicieron posible la inteligencia artificial de hoy, desde la investigación de la retropropagación en 1986 y el logro de AlexNet en 2012, hasta el Premio Turing de 2018 y el Premio Nobel de Física de 2024. Los conceptos desconocidos se explican junto con los problemas que el investigador intentaba resolver.

En la vida de Hinton no solo hay logros. Al dejar Google en 2023, advirtió de manera más activa sobre los peligros de la inteligencia artificial. Comenzó a preguntarse qué impacto tendría en la sociedad la tecnología que él mismo había desarrollado. La pregunta por comprender el aprendizaje y la pregunta por la responsabilidad de la tecnología continúan a lo largo de su vida. Espero que este libro sea una oportunidad para que los lectores jóvenes comprendan los principios y los límites de la inteligencia artificial y reflexionen por sí mismos sobre su futuro.

Abogado Kyung-jin Kim

Índice

El hombre que concibió las máquinas que aprenden

Parte 1. Preguntas hacia el cerebro y el aprendizaje

Capítulo 1. La investigación del cerebro nacida de la curiosidad

Capítulo 2. Cómo se conecta la memoria

Capítulo 3. Redes neuronales que aprenden de las respuestas incorrectas

Parte 2. Cómo hacer que las redes neuronales aprendan

Capítulo 4. Principios de aprendizaje hallados en la física

Capítulo 5. Cómo aprender más rápido

Capítulo 6. Las redes neuronales profundas vuelven a llamar la atención

Parte 3. El aprendizaje profundo sale al mundo

Capítulo 7. Computadoras que reconocen la voz y las imágenes

Capítulo 8. Reducir el sobreajuste y transferir el conocimiento

Capítulo 9. El avance del aprendizaje profundo construido en conjunto

Parte 4. En busca de un mejor aprendizaje

Capítulo 10. La máquina que reconoce las partes y el todo

Capítulo 11. Redescubrir métodos de aprendizaje en el cerebro

Capítulo 12. Cómo aprender sin tabla de respuestas

Parte 5. Logros y responsabilidad

Capítulo 13. El Premio Nobel de Física y las redes neuronales

Capítulo 14. Advertir sobre los peligros de la inteligencia artificial

Capítulo 15. Preguntas en torno a la inteligencia y la experiencia

Referencias

Parte 1. Preguntas hacia el cerebro y el aprendizaje

Capítulo 1. La investigación del cerebro nacida de la curiosidad



Figura 1. Representación de la observación infantil y la curiosidad científica.

1.1. La lógica y la probabilidad legadas por la familia Boole

Geoffrey Everest Hinton nació el 6 de diciembre de 1947 en Wimbledon, Londres, Reino Unido. Su familia era un linaje de académicos que durante varias generaciones produjo matemáticos, científicos, médicos y escritores. Entre sus antepasados, una figura ampliamente conocida es su tatarabuelo George Boole. Boole fue un matemático británico del siglo XIX que creó el álgebra de Boole, un método para calcular tratando lo verdadero y lo falso como los números 1 y 0. Hoy en día, los circuitos lógicos de todas las computadoras se diseñan a partir de esta álgebra booleana. La esposa de Boole, Mary Everest Boole, también fue matemática y educadora. El tío de Mary, Sir George Everest, se desempeñó como agrimensor general de la India, y el monte Everest, la montaña más alta del mundo, recibió su nombre de él. El segundo nombre de Hinton, Everest, proviene de este linaje familiar.

La familia Hinton por parte de padre también era una familia de académicos. Su bisabuelo Charles Howard Hinton fue un

matemático famoso por sus investigaciones para visualizar el espacio tetradimensional, y se casó con Mary Ellen, hija de George Boole. James Hinton, padre de Charles y otro tatarabuelo de Hinton, fue cirujano y escritor. La prima hermana de su padre, Joan Hinton, fue una de las pocas mujeres físicas que participaron en el Proyecto Manhattan, que fabricó la bomba atómica en Los Álamos, Estados Unidos, durante la Segunda Guerra Mundial. Después de la guerra, Joan Hinton se opuso al uso de la bomba atómica, abandonó la física y se trasladó a China.

Su padre, Howard Everest Hinton (H. E. Hinton), era profesor de entomología en la Universidad de Bristol y miembro de la Real Sociedad de Londres (Royal Society). Investigó diversos insectos y mostró un profundo interés por los élmidos (Elmidae), un grupo de pequeños escarabajos acuáticos. Hinton recordó una vez el aspecto del laboratorio de su padre en una conversación relacionada con el Premio Nobel. Los estantes altos de una de las paredes estaban repletos de cientos de cajas con artículos de entomología. Entre ellas, solo una pequeña caja situada en un lugar bajo junto a la puerta tenía una

etiqueta que decía "No insectos (Not Insects)", y dentro de ella se encontraban las cartas de sus hijos y cartas familiares. Era un hogar donde el ámbito académico constituía el centro de la vida. Hinton relató que desde muy temprana edad supo que si no triunfaba como académico sería un fracasado, y que la presión por el rendimiento académico era enorme.

Al poner este linaje familiar en paralelo con la historia de la inteligencia artificial, resalta un aspecto llamativo. El título del libro que George Boole publicó en 1854 es "Una investigación sobre las leyes del pensamiento: sobre las cuales se fundamentan las teorías matemáticas de la lógica y de las probabilidades". En el título figuran la lógica y la probabilidad una al lado de la otra. Sin embargo, los investigadores del simbolismo que lideraron la investigación en inteligencia artificial a partir de mediados del siglo XX solo heredaron la parte lógica del legado de Boole. Intentaron explicar el pensamiento humano mediante símbolos divididos en verdadero y falso y reglas para manipular dichos símbolos. No obstante, el mundo real está lleno de ambigüedades e

incertidumbres, y para manejar un mundo semejante se requiere la probabilidad.

Hinton, tataranieto de Boole, tomó un camino distinto al de la corriente principal precisamente en ese punto. Consideró que el cerebro no era una máquina de precisión que funcionara según reglas lógicas fijas. Pensaba que se trataba de una red que aprende modificando poco a poco, según la experiencia, la intensidad de conexión entre una multitud de células nerviosas (neuronas). Si la lógica del tatarabuelo se convirtió en el armazón de la inteligencia artificial simbólica, el tataranieto partió de la otra mitad del título del libro de su tatarabuelo, la probabilidad, y abrió las puertas del aprendizaje profundo actual.

1.2. La revelación a los siete años y los años escolares en busca de un camino

Una de las anécdotas que Hinton suele relatar al recordar su infancia ocurrió dentro de un autobús. Colocó una antigua moneda británica de un penique sobre el áspero tapizado de un asiento del autobús. Cada vez que el autobús se sacudía, la moneda no se deslizaba hacia abajo, sino que trepaba hacia arriba siguiendo el pelo del tejido del asiento. El joven Hinton sintió que era algo imposible y sintió un inmenso deseo de conocer la razón. La actitud de no pasar por alto un fenómeno incomprensible ya formaba parte intrínseca de él desde entonces.

A los siete años, Hinton ingresó en el Clifton College de Bristol. Era una escuela con un marcado carácter cristiano. La institución transmitía las enseñanzas de la fe como verdades evidentes, y la mayoría de sus compañeros aceptaba esas palabras sin dudarlo. Hinton no lo hizo. No podía aceptar que se le pidiera creer sin pruebas. En una entrevista posterior con el diario Seoul Economic Daily, recordó que fue en aquel

momento cuando pensó por primera vez que cualquier persona podía equivocarse. Explicó que esta vivencia lo llevó a creer que otros investigadores también podían estar equivocados, lo cual sirvió como base para seguir investigando frente a la corriente académica dominante. La fuerza con la que Hinton soportó la frialdad de la comunidad académica durante casi medio siglo se originó en esta experiencia.

El interés por el cerebro surgió durante la escuela secundaria. Hacia los dieciséis o diecisiete años, su amigo Inman Harvey, a quien Hinton siempre consideró más inteligente que él, mencionó un día que los recuerdos cerebrales tal vez se almacenaban distribuidos por todo el cerebro, como un holograma. Un holograma permite reconstruir, aunque sea de forma borrosa, la imagen completa con solo poseer un fragmento. La idea de que los recuerdos no estaban contenidos individualmente en un único lugar, sino distribuidos y superpuestos en múltiples sitios, causó un profundo impacto en Hinton. El deseo de desentrañar cómo calcula el cerebro echó raíces en ese momento.

Sin embargo, encontrar una disciplina académica que respondiera a ese deseo no fue fácil. En 1967, Hinton ingresó en el King's College de la Universidad de Cambridge, y los años siguientes fueron, según sus propias palabras, un "paseo aleatorio demencial (crazy random walk)". Al principio eligió las ciencias naturales y estudió física, química y cristalografía de rayos X. Agotado por un horario exigente de hasta 12 horas diarias, abandonó la universidad tras solo un mes. Se dedicó a leer a Dostoievski mientras pagaba el alquiler en Londres realizando trabajos esporádicos, e incluso intentó convertirse en arquitecto. No obstante, tras trabajar un día en un estudio de arquitectura y comprobar que la labor real se parecía más a seleccionar suelos y picaportes baratos acordes al presupuesto que al diseño, renunció de inmediato.

Tras regresar a la universidad, pasó sucesivamente por departamentos que parecían abordar el cerebro y la mente. Primero eligió fisiología con la esperanza de aprender sobre el cerebro. Las clases enseñaban cómo se transmitían las señales eléctricas en las gruesas fibras nerviosas del calamar, pero no trataban cómo las neuronas representan la información ni

cómo aprenden. A continuación se trasladó a filosofía para buscar la naturaleza de la mente. Cansado de discutir únicamente con palabras sin recurrir a experimentos, se marchó un año después afirmando que solo había desarrollado "anticuerpos contra la filosofía". Finalmente estudió psicología experimental y obtuvo su título de licenciatura en 1970.

La psicología tampoco le proporcionó la respuesta que buscaba. En aquel momento, la psicología estaba sesgada hacia la observación de la conducta de ratas dentro de cajas. En un experimento para investigar el desarrollo cognitivo infantil, Hinton pidió que se escogiera la figura diferente entre un triángulo amarillo, un círculo amarillo y un círculo rojo. Un niño de cinco años señaló el círculo rojo y dijo que el investigador lo había pintado del color equivocado. El pensamiento del niño era mucho más complejo que el diseño experimental. Hinton sintió que la psicología de la época no podía explicar la cognición humana.

Su extravío continuó incluso después de graduarse. Durante aproximadamente un año aprendió el oficio de carpintero. Al

presenciar la destreza de los carpinteros veteranos en las obras, juzgó que no podría alcanzar tal nivel en ese camino y regresó al mundo académico. Más tarde participó como asistente en Bristol en una investigación sobre el desarrollo del lenguaje infantil. Se trataba de un estudio que ponía a prueba la teoría del innatismo lingüístico de Noam Chomsky, según la cual el ser humano nace con la capacidad gramatical inherente. El equipo de investigación instaló micrófonos en los hogares de niños de entre 3 y 5 años y recopiló 166 000 registros de emisiones lingüísticas. No obstante, al visitar hogares de bajos ingresos en las afueras de Bristol para realizar encuestas, ante la primera pregunta de qué hacía cuando su hijo hablaba, una madre respondió que lo golpeaba. Tras presenciar aquello, Hinton se apartó de esa investigación.

Tras recorrer diversas disciplinas, la conclusión a la que llegó fue una sola: para descifrar el secreto de la inteligencia era necesario construir directamente un modelo computacional del cerebro en una computadora.

1.3. El aislamiento en Edimburgo y el conflicto con su director de tesis Longuet-Higgins

En 1972, Hinton ingresó en el programa de doctorado de la Universidad de Edimburgo para investigar la inteligencia artificial en profundidad. En aquel entonces, Edimburgo era casi la única universidad en el Reino Unido que contaba con una infraestructura adecuada para la investigación en inteligencia artificial. Sin embargo, lo que le esperaba allí fue el aislamiento y el conflicto con su director de tesis.

Su director de tesis fue Christopher Longuet-Higgins. Era un brillante científico que se convirtió en profesor de química teórica en la Universidad de Cambridge a los treinta años, y a los cuarenta reorientó su rumbo hacia las ciencias cognitivas para estudiar modelos de almacenamiento de recuerdos semejantes a hologramas y memoria asociativa. Era precisamente el tema que le interesaba a Hinton. Sin embargo, cuando Hinton llegó, el pensamiento de Longuet-Higgins ya había cambiado. Había quedado profundamente impresionado

por las investigaciones de Terry Winograd sobre SHRDLU. SHRDLU era un programa que movía bloques en un micromundo virtual interpretando instrucciones en inglés según reglas simbólicas, y se consideraba un logro emblemático de la inteligencia artificial simbólica. Longuet-Higgins abandonó su apego a las redes neuronales y viró hacia el simbolismo.

El ambiente académico tampoco favorecía a las redes neuronales. En 1969, Marvin Minsky y Seymour Papert demostraron matemáticamente en su libro "Perceptrones (Perceptrons)" que una red neuronal de una sola capa no podía resolver problemas aparentemente tan sencillos como el XOR. Desde entonces, numerosos investigadores consideraron que todo el campo de las redes neuronales carecía de futuro. Longuet-Higgins, aficionado a los sistemas lógicos depurados, no aceptaba las redes neuronales que operaban mediante el método difuso de la probabilidad y el ajuste de pesos de conexión. Calificaba la investigación de Hinton de absurda y le exigía reiteradamente que cambiara al enfoque simbólico.

Hinton no dio su brazo a torcer. Prometió que si le concedía seis meses más mostraría resultados y que, si no lo lograba, cambiaría de tema. Al cabo de seis meses repetía la misma promesa, y esto se reiteró en varias ocasiones. En una entrevista de historia oral concedida a la Universidad de Edimburgo en 2024, Hinton reveló la razón por la que no fue expulsado: su padre era miembro de la Real Sociedad, y para Longuet-Higgins, colega de dicha institución, resultaba incómodo echar al hijo de un compañero. Él mismo señaló con ironía que aquello se debió al privilegio del que disfrutaba.

El nombre de la investigación doctoral concebida inicialmente por Hinton fue LUDWIG. Eran las siglas de "Dispositivo de comprensión del lenguaje sin gramática innata (Language Understanding Device Without Innate Grammar)", que tomaba prestado el nombre del filósofo Wittgenstein a la vez que expresaba su oposición a la teoría del innatismo de Chomsky. Su plan consistía en crear una red neuronal que aprendiera el lenguaje únicamente a través de la experiencia, sin incorporar reglas simbólicas previas. No obstante, en aquel momento no existía un método para entrenar las capas ocultas situadas

entre la capa de entrada y la de salida. LUDWIG no podía construirse con la tecnología de aquella época.

Quien le señaló el camino en ese momento fue el investigador posdoctoral Steve Isard. Isard asumió de facto el papel de codirector de tesis y le sugirió comprobar el potencial de las redes neuronales empezando por problemas pequeños y más sencillos, como el reconocimiento de escritura manual. Hinton estudió de forma autodidacta el libro de Frank Rosenblatt de 1962, "Principios de neurodinámica (Principles of Neurodynamics)", e implementó redes paralelas. Cuando Longuet-Higgins se trasladó a la Universidad de Sussex, Hinton se marchó junto con Isard y Mark Steedman. Tras desempeñarse como investigador en la Universidad de Sussex, Hinton obtuvo su doctorado en la Universidad de Edimburgo en 1978 con la tesis titulada "La relajación y su papel en la visión (Relaxation and Its Role in Vision)".

1.4. Los límites del simbolismo y el contraataque silencioso del conexionismo

El conflicto entre Hinton y Longuet-Higgins no era una disputa meramente personal. En la comunidad de la inteligencia artificial de la segunda mitad del siglo XX se enfrentaban dos perspectivas sobre la inteligencia. Una era el simbolismo (Symbolic AI), liderado por Allen Newell, Herbert Simon, John McCarthy y Marvin Minsky. Ellos sostenían que el pensamiento humano se componía de símbolos, reglas lógicas para manipularlos e inferencias secuenciales paso a paso. Se creía que si los humanos organizaban bien las reglas y las introducían en una computadora, se podría crear inteligencia. Este método se denominó más tarde "la buena y vieja inteligencia artificial (GOFAI)".

La otra corriente era el conexionismo (Connectionism). Hinton, David Rumelhart, Terrence Sejnowski y, más adelante, Yann LeCun y Yoshua Bengio, entre otros, allanaron este camino. Los conexionistas consideraban que el núcleo de la inteligencia no

radicaba en introducir reglas, sino en la capacidad de aprender representaciones por sí misma a partir de los datos. Esto significa que numerosas unidades trabajan en paralelo, como en el cerebro, y acumulan conocimiento ajustando las fuerzas de conexión (pesos) entre dichas unidades. Desde esta perspectiva, un concepto no se almacena en una casilla aislada, sino dividido en patrones de activación de múltiples unidades. A esto se le llama representación distribuida (Distributed Representation). Se trata de una idea semejante a la historia de la memoria holográfica que Hinton escuchó durante la secundaria.

Al entrar en la década de 1970, el simbolismo chocó contra un muro. Los datos del mundo real eran demasiado ambiguos y heterogéneos para poder organizarse por completo mediante reglas. En 1973, el matemático aplicado británico Sir James Lighthill publicó un informe que evaluaba la investigación en inteligencia artificial. Este Informe Lighthill señaló que la investigación en IA no lograba resolver el problema de la explosión combinatoria. La explosión combinatoria es el fenómeno por el cual el número de casos a considerar se

incrementa exponencialmente según el tamaño del problema, haciendo imposible el cálculo. Tras la publicación del informe, los fondos destinados a la investigación en inteligencia artificial en el Reino Unido se redujeron drásticamente y llegó el primer "invierno de la IA".

Muchos investigadores en inteligencia artificial criticaron el Informe Lighthill, pero la evaluación de Hinton fue distinta. Él consideró que la observación de Lighthill daba en el blanco. Los simbolistas como McCarthy sostenían que cualquier cosa calculable mediante símbolos podía convertirse en inteligencia artificial. Sin embargo, ese argumento no tenía en cuenta el tiempo y los recursos requeridos para el cómputo, es decir, la complejidad computacional. Lighthill, compañero de Winchester College de Longuet-Higgins y del físico Freeman Dyson, y matemático aplicado especializado en hidrodinámica, identificó con precisión esta debilidad. Hinton compartía la misma opinión. Consideraba que para abordar la complejidad de la realidad con recursos computacionales limitados no bastaba manipular símbolos uno a uno, sino que se requería un método en el que múltiples unidades trabajaran en simultáneo,

como en el cerebro, modificando gradualmente las intensidades de conexión.

En medio del desdén de la comunidad académica, Hinton buscó colaboradores afines. En 1979, organizó junto con James Anderson el taller "Modelos paralelos de memoria asociativa (Parallel Models of Associative Memory)" en La Jolla, San Diego, Estados Unidos. A este encuentro acudieron investigadores conexionistas de diversas instituciones, entre ellos David Rumelhart de la Universidad de California en San Diego (UCSD). Esta reunión se convirtió en el trampolín para la posterior creación del grupo de investigación en Procesamiento Distribuido en Paralelo (PDP). Aunque no resultara evidente hacia el exterior, el contraataque del conexionismo comenzó silenciosamente a partir de entonces.

1.5. Curiosidad antes que resultados a corto plazo: los frutos de la investigación básica y su significado

Desde la obtención de su doctorado en 1978 hasta que AlexNet, creada en 2012 junto con sus discípulos, triunfó de manera abrumadora en una competición de reconocimiento de imágenes, las redes neuronales fueron durante largo tiempo un campo marginal al que le costaba conseguir financiación. Durante ese período, el campo de la inteligencia artificial atravesó dos inviernos. La fuerza que permitió a Hinton resistir este largo tiempo no radicaba en el éxito comercial rápido ni en el prestigio. Era la curiosidad por conocer cómo funciona el cerebro humano. Cuando la organización del Premio Nobel le pidió en una entrevista que se presentara, respondió que no estaba muy seguro de a qué campo pertenecía, pero que era una persona que deseaba comprender cómo funciona el cerebro.

Hinton subrayó públicamente en repetidas ocasiones que esa curiosidad debía ser el punto de partida de la investigación. El 8

de octubre de 2024 se anunció su galardón con el Premio Nobel de Física junto a John Hopfield. En la conferencia de prensa de la Universidad de Toronto celebrada ese mismo día, instó a financiar la investigación básica guiada por la curiosidad. Valoró que, aunque Canadá no dispone de tantos fondos de investigación como Estados Unidos, emplea el dinero de manera inteligente, y explicó que el avance de las redes neuronales no provino de inyectar dinero en problemas aplicados, sino del empeño de los científicos por comprender algo siguiendo su curiosidad. En la entrevista con Seoul Economic Daily, afirmó también que en lugar de presionar a los académicos para obtener resultados prácticos inmediatos, es mucho más eficaz permitir a los investigadores indagar con libertad sobre aquello que les despierta curiosidad, y que el aprendizaje profundo se gestó precisamente de ese modo.

Detrás del fructífero resultado de estas investigaciones estuvo el apoyo a largo plazo del Instituto Canadiense de Investigaciones Avanzadas (CIFAR). En 1987, con el respaldo de CIFAR, Hinton se trasladó a Canadá y se estableció en la Universidad de Toronto. Señaló que el sistema de subvenciones

para investigación de Estados Unidos obliga a los profesores a dedicar demasiado tiempo a redactar propuestas. En cambio, explicó que en Canadá, con una propuesta de 6 páginas, podía recibir financiación estable durante 3 a 5 años sin quedar atado a un tema específico. Gracias a ese entorno, Hinton pudo profundizar de manera tenaz en un campo impopular sin seguir las modas.

El significado del camino recorrido por Hinton no se reduce al éxito tecnológico. También cambió la forma de concebir la mente humana. El simbolismo consideraba al ser humano como una "máquina de razonamiento (Reasoning Machine)" que manipula con precisión símbolos y razona de manera lógica. La investigación de Hinton demostró que el ser humano se asemeja más a una "máquina de analogías (Analogy Machine)" que comprende el mundo agrupándolo en patrones a partir de la experiencia sensorial y la similitud. Desde el punto de vista de Hinton, la inferencia lógica es una capacidad añadida sobre la red neuronal, y la base de la inteligencia reside en el ajuste de las intensidades de conexión mediante el cual

incontables neuronas extraen por sí mismas representaciones distribuidas a partir de la experiencia.

La idea que albergó aquel niño de siete años, la revelación de que los adultos también podían equivocarse, maduró medio siglo después en la convicción de que toda la comunidad académica podía estar en un error. Partiendo de la sombra proyectada por la lógica de su tatarabuelo George Boole, Hinton se aferró a la probabilidad que figuraba a su lado en el título de aquel libro y recorrió el camino hasta el final.

Capítulo 2. Cómo se conecta la memoria

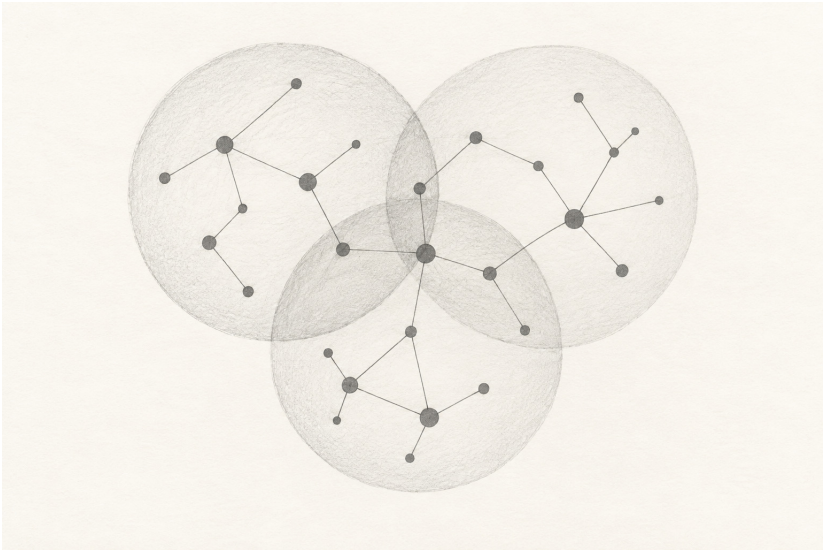


Figura 2. Representación visual de la idea de codificar la información distribuyéndola a través de múltiples conexiones.

2.1. Los límites de la 'neurona de la abuela' y la orquesta de la representación distribuida

Durante sus años de secundaria, Geoffrey Hinton escuchó hablar de los hologramas a través de un amigo. En un holograma, aunque se recorte una parte de la película, la imagen completa permanece, si bien de forma borrosa. Esto se debe a que la información no está registrada en un solo punto, sino distribuida por toda la película. Hinton pensó que la memoria en el cerebro también podría estar distribuida de manera similar, y convirtió esta idea en el tema de investigación de toda su vida.

Sin embargo, el simbolismo, que era la corriente dominante en la ciencia cognitiva y la inteligencia artificial durante la década de 1970, concebía el almacenamiento del conocimiento de una manera completamente diferente. Los teóricos simbolistas consideraban que el pensamiento humano se compone de símbolos y de reglas para manipularlos, y pensaban que cada

concepto se almacena en un único lugar. A este modo de almacenamiento se le llama 'representación localista'.

Una expresión que ilustra la representación localista es la 'hipótesis de la neurona de la abuela'. Es la hipótesis de que en algún rincón del cerebro existe una sola neurona encargada exclusivamente de la propia abuela, y que cuando esa neurona se activa, uno reconoce a su abuela. De acuerdo con esta hipótesis, debería haber una neurona para el gato y un nodo simbólico para cada oración específica. Hacer corresponder un solo objeto con una sola unidad facilita la distinción, pero dificulta representar los puntos en común entre diversos objetos. Con todo, esto no significa que todas las teorías que explican el pensamiento mediante símbolos hagan corresponder un concepto con una única neurona.

Hinton dudó de esta idea desde muy temprano. La representación localista presentaba tres puntos débiles.

El primero es el problema de la combinación. Los objetos del mundo se manifiestan en combinaciones incesantemente

nuevas. Si cada vez que aumenta una combinación, como 'manzana roja', 'manzana verde' o 'manzana roja comida a medias', hubiera que crear un nuevo nodo dedicado, el número de nodos necesarios se dispararía exponencialmente. A esto se le llama 'explosión combinatoria'.

El segundo es su vulnerabilidad al daño. Si muriera la única neurona encargada de la abuela, todo el conocimiento sobre ella tendría que desvanecerse. En la representación distribuida, en cambio, aunque se dañen algunas unidades de cómputo, la información puede preservarse en las unidades restantes.

El tercero es la dificultad para albergar la similitud. Un gato y un tigre se parecen, mientras que un gato y un refrigerador no se parecen. Sin embargo, si para cada concepto existe un nodo independiente y separado, el nodo por sí mismo no puede indicar qué tan parecidos son ambos conceptos. La información relativa a la similitud tendría que ingresarse una por una mediante reglas independientes.

La alternativa propuesta por Hinton es la 'representación distribuida'. La representación distribuida es un método que no confía un concepto a una sola neurona, sino que lo expresa mediante el patrón de activación que conforman conjuntamente numerosas neuronas. Algunas neuronas se encienden con fuerza, otras se encienden débilmente y algunas permanecen apagadas. Esa configuración global constituye, en sí misma, un concepto. En términos matemáticos, el concepto se convierte en un vector formado por varios números.

La representación distribuida puede compararse con una orquesta. Un compás de una sinfonía no es el sonido que emite un violín en solitario. El violín, el violonchelo, la flauta y la trompeta resuenan juntos con intensidades distintas para crear una sola armonía. Cuando cambia el compás, la combinación y la intensidad de los instrumentos cambian al unísono, y el mismo instrumento interviene sin descanso en diversas partes de la pieza. La idea de Hinton era que el conocimiento cerebral también existe así, como una armonía producida conjuntamente por innumerables neuronas.

Tal como los instrumentos de una orquesta participan en varios compases, en la representación distribuida una sola neurona no se ocupa de un único concepto. La neurona representa pequeños rasgos como 'tiene pelaje', 'camina en cuatro patas' o 'vive en el hogar', y participa simultáneamente en la representación de múltiples conceptos. A estos pequeños rasgos se les denomina 'microrrasgos'. Dado que el gato y el tigre comparten numerosos microrrasgos, los patrones de activación de ambos conceptos resultan semejantes entre sí. La similitud se incorpora de forma natural dentro de la representación, sin necesidad de reglas independientes. Asimismo, es resistente al daño. Aunque se dañen algunas neuronas, la mayor parte del patrón generado por las demás neuronas perdura, de modo que el conocimiento no se borra de golpe.

Hinton también demostró esta idea experimentalmente. Se trata de la 'tarea del árbol genealógico' de 1986. Hinton construyó los árboles genealógicos de dos familias compuestas por 12 integrantes de una familia inglesa y 12 integrantes de una familia italiana, sumando en total 24 personas. Ambos

árboles genealógicos tenían una estructura idéntica. En la familia inglesa figuraban nombres como Christopher, Penelope, Charlotte, Victoria, James y Colin, y en la familia italiana aparecían nombres como Roberto y Maria en las mismas posiciones. Había 12 tipos de relaciones, entre ellas padre, madre, esposo, esposa, tío y sobrino.

A la red neuronal se le ingresaban 'Persona 1' y 'Relación' para que respondiera con 'Persona 2'. Por ejemplo, al introducir 'Charlotte' y 'madre', debía responder 'Victoria'. Los nombres de las personas se introducían activando solo una de las 24 casillas, y las relaciones activando solo una de las 12 casillas. Por lo tanto, la entrada en sí misma carecía por completo de información sobre qué tan parecidas eran las personas entre sí. Las dos entradas pasaban, respectivamente, por capas compuestas por 6 neuronas cada una, confluían en una capa oculta intermedia y, tras atravesar varias capas adicionales, producían la respuesta.

La red neuronal no recibió ninguna regla como: "Si la madre de A es B y el esposo de B es C, entonces C es el padre de A". Lo

único que se le proporcionó fueron 104 oraciones que enunciaban hechos. Hinton entrenó la red neuronal con 100 de ellas y reservó las 4 restantes para la evaluación.

Al concluir el aprendizaje, Hinton examinó las 6 neuronas que representaban a las personas. Ciertas neuronas distinguían si se trataba de la familia inglesa o de la italiana, otras indicaban de qué generación se trataba y otras señalaban a qué rama del árbol genealógico pertenecían. La red neuronal había descubierto por su cuenta rasgos que nadie le había enseñado. Además, respondió correctamente a la gran mayoría de las preguntas de prueba que jamás había visto.

Hinton consideró que el verdadero valor del método de aprendizaje por retropropagación del error radicaba justamente aquí. Más que la técnica computacional en sí, lo relevante era que la capa oculta pudiera aprender por sí misma representaciones distribuidas que reflejaran la estructura intrínseca de los datos. La raíz de las 'incrustaciones de palabras' (word embeddings), que en la inteligencia artificial

actual convierten vocablos en vectores numéricos, se encuentra en este experimento.

2.2. Investigadores reunidos en La Jolla: el taller de 1979 y el nacimiento del grupo PDP

A finales de la década de 1970, los investigadores de redes neuronales se encontraban dispersos como islas solitarias. La mayor parte de los fondos de investigación y de las plazas académicas se destinaban a los estudios simbolistas. Quienes investigaban sobre redes neuronales apenas sabían de la existencia mutua, y sus artículos se publicaban dispersos en revistas especializadas de escasa repercusión.

La situación de Hinton no era distinta. Obtuvo su doctorado en 1978 en la Universidad de Edimburgo, en el Reino Unido, pero en Gran Bretaña resultaba sumamente difícil conseguir subvenciones para investigar redes neuronales. Hinton había investigado en la Universidad de Sussex antes de titularse y, en

octubre de 1978, se trasladó a los Estados Unidos a la Universidad de California en San Diego (UCSD).

En 1979, Hinton organizó en La Jolla, donde se ubica la UCSD, el taller 'Modelos paralelos de memoria asociativa' junto con James Anderson, psicólogo de la Universidad de Brown. Este taller constituyó la primera ocasión en que los investigadores de redes neuronales, dispersos hasta ese momento, se reunieron en un mismo lugar.

Entre los asistentes se encontraba Terry Sejnowski. En aquel momento era investigador posdoctoral en la Escuela de Medicina de Harvard y un joven científico que había publicado unos pocos artículos sobre redes neuronales en revistas poco conocidas. Se dice que Hinton lo invitó tras consultar sobre sus investigaciones a David Marr, teórico de la visión del MIT. Ambos, tras coincidir en este taller, se convertirían en colegas de por vida al desarrollar juntos la máquina de Boltzmann.

Los resultados del taller se recopilaron en 1981 en un libro homónimo, "Parallel Models of Associative Memory". Los

compiladores fueron Hinton y Anderson. Hinton contribuyó a esta obra con un capítulo sobre la implementación de redes semánticas en hardware paralelo. Su planteamiento demostraba que las relaciones semánticas, que el simbolismo representaba mediante grafos y símbolos, podían expresarse a través de patrones de activación de neuronas.

En La Jolla se encontraban las figuras llamadas a dar continuidad a este movimiento: David Rumelhart y James McClelland, de la UCSD. A comienzos de la década de 1980, ambos crearon el grupo de investigación de 'Procesamiento Distribuido en Paralelo' (Parallel Distributed Processing), conocido de forma abreviada como grupo PDP. Por aquella época, Hinton forjó una estrecha amistad con Rumelhart mientras trabajaba como investigador posdoctoral en la UCSD, y continuó investigando junto al grupo PDP incluso después de marcharse a Cambridge, en el Reino Unido, y a la Universidad Carnegie Mellon, en los Estados Unidos.

El apoyo y la implicación de los investigadores del entorno fueron igualmente decisivos para el crecimiento del grupo PDP.

Donald Norman, quien encabezaba el área de ciencia cognitiva en la UCSD, actuó como un protector dentro del departamento para que los científicos pudieran desplegar con libertad ideas divergentes de las dominantes. Francis Crick, del vecino Instituto Salk, también frecuentaba las reuniones. Crick, el célebre biólogo que había descubierto la estructura en doble hélice del ADN, había volcado su interés por entonces hacia el estudio del cerebro. Debatió intensamente con los investigadores acerca de la similitud entre los modelos de redes neuronales y el cerebro biológico real, y con posterioridad coescribió un capítulo sobre la estructura de la corteza cerebral en el volumen 2 de la obra de PDP.

La investigación del grupo PDP se sintetizó en 1986 en una obra de dos tomos publicada por MIT Press, titulada "Parallel Distributed Processing: Explorations in the Microstructure of Cognition". El volumen 1 trataba los fundamentos y el volumen 2 abordaba las aplicaciones en psicología y biología. Esta publicación ejerció un impacto monumental en los campos de la psicología, las neurociencias y la informática, y se dice que

vendió más de 50.000 ejemplares, una cifra insólita para un tratado académico.

Hinton estampó su nombre en varios de los capítulos capitales del volumen 1. El capítulo 3, 'Representaciones distribuidas', fue redactado conjuntamente por Hinton, McClelland y Rumelhart. El capítulo 7, 'Aprendizaje y reaprendizaje en máquinas de Boltzmann', fue obra de Hinton y Sejnowski. El capítulo 8, 'Aprendizaje de representaciones internas mediante propagación de errores', fue escrito por Rumelhart, Hinton y Ronald Williams. Ese mismo año, los tres investigadores publicaron en "Nature" el célebre artículo que introdujo el método de aprendizaje por retropropagación. Aquel modesto círculo surgido en La Jolla se convirtió así en el epicentro que resucitó la investigación en redes neuronales.

2.3. La memoria asociativa paralela y el principio de optimización de pesos

La 'memoria asociativa', presente en el título del taller de La Jolla, fue otra idea cardinal impulsada por el conexionismo junto con la representación distribuida. Este concepto resulta sencillo de entender si se contrapone al mecanismo con el que un ordenador convencional recupera información. Las computadoras localizan los datos mediante direcciones de memoria. Es imprescindible conocer la dirección exacta que precisa en qué casilla se halla determinado elemento para recuperar su contenido; si la dirección yerra aunque sea en un solo carácter, resulta imposible encontrar lo buscado. A este esquema se le denomina 'memoria direccionable por dirección'.

La memoria humana opera de un modo muy distinto. Al escuchar tan solo las primeras notas del compás inicial de una canción, acude a la mente la pieza completa; con apenas ver la mitad del rostro de un amigo, sabemos de inmediato de quién se trata. Se recupera la totalidad a partir de un fragmento del

contenido. Esto se conoce como 'memoria direccionable por contenido'. Los investigadores conexionistas sostenían que las redes neuronales podían ejecutar este tipo de recuperación memorística de forma completamente natural.

Quien demostró este principio con nitidez matemática fue el físico John Hopfield. En 1982, Hopfield trasladó a las redes neuronales el concepto de 'energía', que la física utilizaba para describir materiales magnéticos. En una red de Hopfield, a cada estado de las neuronas le corresponde un determinado valor energético. Los recuerdos almacenados equivalen a valles excavados en un paisaje de energía. Cuando ingresa una señal parcialmente borrada o contaminada con ruido, las neuronas influyen mutuamente entre sí y reducen la energía paso a paso. Igual que una esfera rueda cuesta abajo hasta detenerse en el fondo de una hondonada, la red arriba al valle más próximo y recompone el recuerdo original. A este proceso, en el que infinidad de neuronas se mueven a la vez en busca de un estado de estabilidad, se le llama 'relajación paralela'.

La red de Hopfield acarreaba ciertas limitaciones. En el relieve energético podían formarse pequeños valles espurios que conducían a recuerdos erróneos, o bien la red quedaba atrapada en cuencas poco profundas, incapaz de alcanzar soluciones óptimas. Para resolver este escollo, Hinton y Sejnowski incorporaron la noción probabilística de la física estadística. Al permitir que las neuronas fluctúen con cierta probabilidad en vez de regirse únicamente por pautas inflexibles, la red logra salir de un valle superficial y dar con otro más profundo. El modelo concebido bajo este principio es la 'máquina de Boltzmann'. David Ackley, Hinton y Sejnowski publicaron en 1985 el artículo que detallaba el método de aprendizaje para la máquina de Boltzmann.

La esencia del problema que resuelve la máquina de Boltzmann radica en la 'satisfacción paralela de restricciones'. Cuando un ser humano contempla un objeto, el cerebro conjuga a la vez infinidad de parámetros, tales como los bordes, la profundidad, la dirección lumínica y el cromatismo, para encontrar la interpretación más plausible. Tales condiciones se encuentran codificadas en la intensidad de las conexiones interneuronales,

esto es, en los 'pesos'. Si resulta armónico que dos neuronas se enciendan al unísono, el peso entre ellas es elevado; si encenderse juntas resulta contradictorio, el peso es reducido o negativo.

Surge entonces la incógnita de quién determina dichos pesos. La respuesta del conexionismo es que la propia red neuronal los fija por sí misma al examinar los datos. No hay un programador humano que codifique las reglas. Al procesar múltiples ejemplos, la red neuronal modifica paulatinamente sus pesos con el fin de que las pautas que ella misma engendra converjan con los patrones del mundo real. En esto estriba la optimización de pesos, y este proceso constituye, en puridad, el aprendizaje.

Hinton se interesó asimismo por cómo muta a lo largo del tiempo la memoria contenida en los pesos. En 1987 dio a conocer, en colaboración con David Plaut, una investigación sobre los 'pesos rápidos'. Ambos investigadores dotaron a cada conexión de dos clases de pesos. Los pesos lentos evolucionan pausadamente a lo largo del tiempo y almacenan el conocimiento general sobre el mundo. Por su parte, los pesos

rápidos varían y se desvanecen con celeridad ante la entrada de información novedosa, desempeñando el cometido de retener de forma transitoria lo recién percibido.

Esta arquitectura sirvió para abordar un inconveniente tangible en el aprendizaje neuronal. Cuando una red neuronal asimila elementos nuevos, con frecuencia se atenúa lo aprendido con anterioridad. En los experimentos desarrollados por ambos, los pesos rápidos facilitaron la reactivación veloz, con apenas una breve práctica, de los viejos recuerdos desvaídos a causa del nuevo entrenamiento. El título del artículo, 'Disipar el desvanecimiento de viejos recuerdos con pesos rápidos', ilustra a la perfección dicho fenómeno. La noción de los pesos rápidos cayó en el letargo durante décadas, hasta que en 2016 Hinton y sus colaboradores la rescataron, concitando de nuevo una enorme atención.

2.4. La transformación en la mirada sobre la mente: del manipulador de símbolos al aprendiz de datos

Las investigaciones relativas a la representación distribuida y a la memoria asociativa no constituyeron un mero avance técnico. Ofrecieron una respuesta divergente a la milenaria pregunta sobre cómo concebir la mente humana.

El ser humano concebido por el simbolismo y la ciencia cognitiva clásica era un 'manipulador de símbolos'. Sostenían que el individuo es una máquina lógica que procesa ordenadamente los símbolos internos conforme a reglas preestablecidas. Al contemplar la destreza humana para expresarse con arreglo a la gramática y catalogar los objetos en categorías, infirieron que en el interior del cerebro debían de hallarse preinscritos los compendios de normas gramaticales y los catálogos categoriales.

La imagen del ser humano esbozada por Hinton y los científicos conexionistas era radicalmente distinta. Para ellos, el cerebro

no era una máquina que transcribe reglas y computa sucesivamente, sino un 'aprendiz de datos' que extrae por sí mismo correlaciones y patrones a partir de una infinidad de vivencias. Hinton ha reiterado que el ser humano se asemeja más a una 'máquina de analogías' que a una máquina de lógica. Ello implica que, ante una tesitura inédita, antes de aplicar reglas una tras otra, el individuo juzga evocando escenarios análogos experimentados en el pasado.

Bajo este prisma, ni el discurso gramaticalmente correcto ni la deducción lógica sobrevienen porque las reglas se encuentren impresas en el cerebro. Obedecen más bien a la interacción mutua entre las representaciones distribuidas forjadas en los pesos de infinidad de neuronas, lo que hace brotar exteriormente conductas que aparentan estar regidas por reglas. A esto se le conoce como 'emergencia'. Así como la red neuronal del experimento del árbol genealógico acertó los vínculos de parentesco sin haber sido instruida en reglas genealógicas, los seres humanos también son capaces de conducirse con arreglo a reglas sin necesidad de haberlas memorizado.

Incluso en sus disertaciones más recientes, Hinton enfatiza una y otra vez que las investigaciones de aquellos tiempos se enlazan directamente con la inteligencia artificial contemporánea. En una conferencia dictada en junio de 2025 en la Universidad de Bristol, en el Reino Unido, presentó el modesto modelo de lenguaje que concibió en 1985. Explicó que dicho modelo empleaba el error surgido al predecir la palabra sucesiva como señal de aprendizaje, transformando los símbolos verbales en activaciones de rasgos y haciendo interactuar estos entre sí para aprender a anticipar los rasgos de la siguiente palabra. Afirmó que los grandes modelos lingüísticos actuales pueden considerarse descendientes directos de aquel pequeño modelo. El modelo miniatura al que aludía no era otro que la red neuronal que aprendió el árbol genealógico.

El estudio de la memoria asociativa cosechó también su justa revalorización décadas después. El Premio Nobel de Física de 2024 fue otorgado a Hopfield y Hinton. La Real Academia de las Ciencias de Suecia fundamentó el galardón exponiendo que Hopfield ideó una memoria asociativa capaz de registrar y

reconstruir patrones de datos tales como imágenes, y que Hinton, tomando como base la red de Hopfield, diseñó una nueva red denominada máquina de Boltzmann. Aquella 'memoria asociativa' que figuraba en el enunciado del taller de La Jolla en 1979 reapareció 45 años más tarde en el veredicto del Premio Nobel de Física.

A lo largo de las décadas de 1970 y 1980, Hinton consolidó la teoría de la representación distribuida y la memoria asociativa transitando entre La Jolla, Cambridge y Pittsburgh. En aquellos días, sus planteamientos habitaban en los márgenes de la academia: la convicción de que el conocimiento no es un símbolo grabado en un solo punto sino una armonía coral generada por multitud de neuronas; la certeza de que la memoria se busca por su contenido y no por su dirección; y la premisa de que las reglas no se programan, sino que se aprenden a partir de los datos. Estas tres concepciones se han erigido en los cimientos de la inteligencia artificial que hoy en día procesa el lenguaje y los conceptos.

Capítulo 3. Redes neuronales que aprenden de las respuestas incorrectas

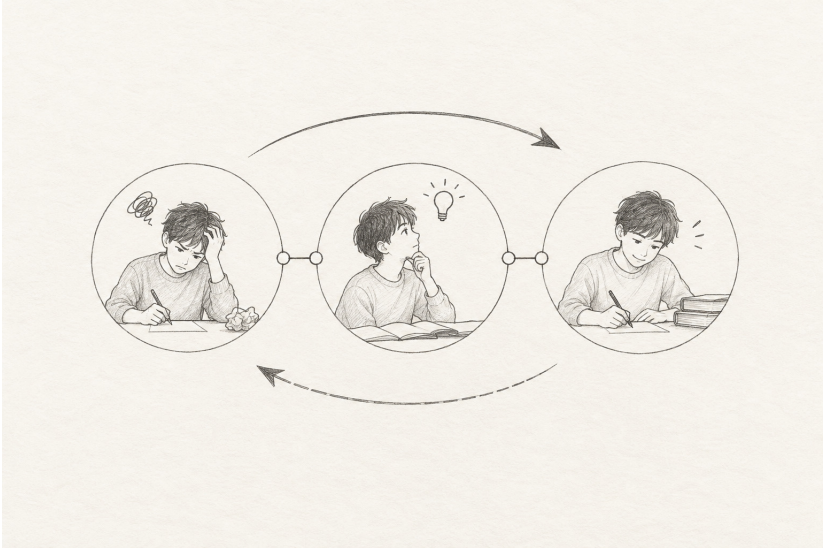


Figura 3. Metáfora del proceso de repasar los resultados erróneos para mejorar el aprendizaje.

3.1. El dilema de las redes neuronales multicapa y el redescubrimiento de la retropropagación (Backpropagation)

En 1969, Marvin Minsky y Seymour Papert publicaron un libro titulado «Perceptrons». En este libro, ambos demostraron matemáticamente que un perceptrón monocapa, que solo cuenta con una capa de entrada y una capa de salida, no puede resolver problemas como XOR. Tras la publicación del libro, muchos investigadores abandonaron la investigación sobre redes neuronales y se volcaron hacia la inteligencia artificial simbolista. El simbolismo es un enfoque en el que los seres humanos escriben directamente las reglas y la computadora realiza inferencias de acuerdo con dichas reglas.

Esto no significaba que no hubiera absolutamente ninguna forma de superar las limitaciones del perceptrón monocapa. Se sabía teóricamente que al insertar una capa oculta (Hidden Layer) entre la capa de entrada y la de salida se podían resolver problemas más complejos. El problema radicaba en cómo entrenar esa capa oculta. La capa de salida se puede comparar

directamente con la respuesta correcta. Si la red neuronal mira la foto de un gato y responde 'perro', se puede calcular de inmediato en qué medida se equivocó. Con la capa oculta la situación es diferente. Las neuronas de la capa oculta no están conectadas directamente con el mundo exterior, por lo que nadie les indica qué valor deberían emitir en cada momento para ser correctas. El problema de determinar y repartir en qué medida recae la responsabilidad del error final entre las numerosas neuronas intermedias se denomina 'problema de asignación de crédito (Credit Assignment Problem)'. Mientras no se pudiera resolver este problema, las redes neuronales multicapa no tenían más remedio que quedarse en un mero concepto sobre el papel.

La respuesta al problema de asignación de crédito es la retropropagación del error. El principio radica en la regla de la cadena (Chain Rule) del cálculo diferencial. Primero, se obtiene el error en la capa de salida. A continuación, este error se envía hacia atrás, es decir, capa por capa desde la capa de salida hacia la capa oculta, calculando en qué medida contribuyó cada peso de conexión al error total. Finalmente, los pesos se

ajustan poco a poco en la dirección que reduce el error. A esta última etapa se la denomina descenso de gradiente (Gradient Descent). Es como descender una montaña con los ojos vendados, comprobando únicamente la pendiente bajo los pies y dando pasos uno a uno hacia la zona más baja.

La retropropagación no fue un invento ideado por una sola persona en un único instante. Múltiples académicos llegaron a la misma idea de forma independiente. En 1970, el finlandés Seppo Linnainmaa sistematizó en su tesis de maestría el método de cálculo que hoy denominamos 'diferenciación automática en modo inverso'. En 1974, Paul Werbos, estudiante de doctorado de la Universidad de Harvard, propuso en su tesis doctoral «Beyond Regression» que este método podía utilizarse para el entrenamiento de redes neuronales. Hacia 1985, David Parker y el francés Yann LeCun propusieron, cada uno por su cuenta, reglas de aprendizaje similares.

Aun así, lo que se señala como el punto de inflexión en la historia de la retropropagación es la investigación de 1986 de David E. Rumelhart, Geoffrey E. Hinton y Ronald J. Williams. La

colaboración entre los tres comenzó en San Diego. Hinton trabajó como investigador posdoctoral en la Universidad de California en San Diego (UCSD) de 1978 a 1980, y posteriormente continuó interactuando con colegas y participando en el grupo de investigación de PDP (Procesamiento Distribuido en Paralelo) dirigido por Rumelhart, James McClelland y otros. En un artículo de 2024, UC San Diego resumió que el trabajo del grupo PDP sentó las bases para el desarrollo de la retropropagación, y que más adelante Hinton, junto con Rumelhart y Williams, difundió ampliamente este algoritmo a través del artículo de 1986 en «Nature».

No fueron ellos tres quienes propusieron por primera vez la retropropagación. El mérito de los tres no radica en haber escrito la fórmula por primera vez, sino en haber demostrado experimentalmente el hecho de que la capa oculta de una red neuronal entrenada con retropropagación crea por sí misma representaciones internas (Internal Representation) que ningún ser humano le enseñó.

El propio Hinton también dejó claro este punto. En una entrevista de 2018 con Martin Ford, declaró que la idea fundamental de la retropropagación fue concebida por Rumelhart y era un invento de Rumelhart. Explicó que él y Williams contribuyeron a perfeccionar el método, y que su propia contribución consistió en demostrar que la retropropagación podía aprender representaciones distribuidas significativas. Cada vez que los medios de comunicación lo presentaban como el inventor de la retropropagación, Hinton hacía esta corrección.

3.2. El artículo de 1986 en «Nature» y los principios de cálculo

El 9 de octubre de 1986, se publicó un artículo de cuatro páginas en las páginas 533 a 536 del volumen 323 de la revista académica internacional «Nature». Su título era «Learning representations by back-propagating errors», que en español se traduce como 'Aprender representaciones mediante la retropropagación de errores'. En cuanto a la filiación de los autores, Rumelhart y Williams pertenecían al Instituto de Ciencias Cognitivas de UCSD, y Hinton a la Universidad Carnegie Mellon (CMU). Ese mismo año, los tres explicaron detalladamente el mismo contenido al publicar un extenso capítulo titulado «Learning internal representations by error propagation» en el volumen 1 del libro de dos tomos «Parallel Distributed Processing», editado por el grupo PDP. El breve artículo captó la atención de la comunidad académica, y el extenso libro sirvió como libro de texto que los investigadores pudieron seguir directamente.

El proceso de cálculo del artículo puede dividirse en cuatro etapas. Las tres primeras etapas corresponden al proceso en el que la red neuronal genera una respuesta y mide cuánto se desvía dicha respuesta de la solución correcta, mientras que la última etapa es el proceso de distribuir el error así medido como responsabilidad de cada una de las neuronas.

La primera es la propagación hacia adelante (Forward Pass). La señal asciende desde la capa inferior hacia la capa superior. Cada neurona de la capa superior está conectada con múltiples neuronas de la capa inferior, y cada conexión tiene un peso que representa su intensidad. Al multiplicar la salida de cada neurona de la capa inferior por el peso de la conexión correspondiente y sumar todos esos valores, se obtiene la entrada total que recibe la neurona de la capa superior. El umbral que determina cuándo se activa una neurona no se trató por separado. Se introdujo una entrada ficticia que siempre emite 1, haciendo que el peso asignado a esa entrada desempeñara el papel del umbral, calculándose de la misma manera que los demás pesos.

La segunda es la función de activación. La función de activación es la regla que transforma la entrada total recibida por una neurona en la salida de dicha neurona. En este artículo se utilizó la función sigmoide logística. Esta función devuelve un valor cercano a 0 si la entrada total es un número negativo grande, y un valor cercano a 1 si es un número positivo grande; entre ambos extremos traza una suave curva en forma de S que produce valores entre 0 y 1. Dado que la curva continúa suavemente sin interrupciones ni quiebres, es diferenciable en cualquier punto. La derivada es un cálculo que determina cuánto cambia la salida cuando la entrada varía ligeramente. La derivada de la función sigmoide se obtiene de forma sencilla multiplicando el valor de salida por el resultado de restar dicho valor de salida a 1. La función escalón de los primeros perceptrones, que solo tenía dos estados (encendido y apagado), no permitía utilizar este tipo de derivada porque el valor saltaba repentinamente en un punto. La retropropagación envía el error hacia atrás de acuerdo con la regla de la cadena, es decir, la regla que calcula la tasa de cambio a través de múltiples etapas como el producto de las tasas de cambio de

cada etapa. Para que este método funcionara, se requería una función tan suave como esta.

La tercera es la medición del error. Se calcula la diferencia entre la respuesta y(j) dada por la red neuronal y la respuesta correcta d(j), se eleva al cuadrado, se suman todas las diferencias y se toma la mitad. Este valor E es la puntuación que indica en qué medida se equivocó la red neuronal.

La cuarta es la propagación hacia atrás (Backward Pass). Aquí se calcula la señal de error para cada neurona. La señal de error es un valor que representa cuánto varía el error total cuando la entrada total recibida por esa neurona cambia ligeramente, lo que corresponde a la magnitud de la responsabilidad que dicha neurona tiene en el error. La señal de error de una neurona en la capa de salida se calcula multiplicando la diferencia entre la salida real y el valor objetivo dado como respuesta correcta por la derivada de la función sigmoide, es decir, el producto del valor de salida por el resultado de restar ese valor de salida a 1. La señal de error de una neurona de la capa oculta se obtiene sumando los productos de las señales de error de las neuronas

de la capa superior que recibieron la salida de esta neurona por los pesos de conexión correspondientes, y multiplicando este resultado por el valor de salida de esta propia neurona y por uno menos dicho valor de salida. Es un cálculo que distribuye la responsabilidad de la capa superior a la capa inferior en proporción a la fuerza de las conexiones. De esta forma, sin importar cuántas capas ocultas haya, es posible comenzar en la capa de salida y descender capa por capa para calcular la responsabilidad de todas las neuronas. Así se proporciona una señal de error incluso a las capas ocultas, que no podían compararse directamente con la respuesta correcta. Finalmente, cada peso se ajusta en la dirección en que disminuye el error, y la magnitud de ajuste por cada paso se determina mediante un número pequeño llamado tasa de aprendizaje.

Los investigadores también abordaron problemas que surgen en el entrenamiento real. Uno de ellos fue la velocidad de aprendizaje. El descenso de gradiente es un método que desplaza poco a poco los pesos en la dirección en que se reduce el error. Si se compara la forma en que varía el error según los

pesos con un relieve geográfico, cuando este terreno tiene la forma de un valle estrecho y alargado, el descenso de gradiente oscila yendo y viniendo entre ambas paredes del valle, y la velocidad de avance a lo largo del valle se vuelve lenta. Para mitigar este problema, los investigadores incorporaron un término de inercia o momento (Momentum). La variación original es la cantidad que desplaza en dirección opuesta al gradiente multiplicando el gradiente del error por la tasa de aprendizaje, a la cual se le añade una parte de la cantidad en que se modificó el peso en el paso inmediatamente anterior. La proporción en que se refleja la variación del paso previo se determina mediante el coeficiente de inercia, un número entre 0 y 1. Esto produce el mismo efecto que una pelota rodante que, a pesar de toparse con un pequeño obstáculo, lo supera gracias a la velocidad que traía acumulada. Esta técnica se utilizó durante mucho tiempo. En 2013, Ilya Sutskever y otros volvieron a examinar el efecto del término de momento en el entrenamiento de redes neuronales profundas.

La otra preocupación se refería a los mínimos locales (Local Minima). Un mínimo local es un punto donde el error es más

bajo que en los alrededores inmediatos, pero no es el punto con el error más bajo a nivel global. Se señalaba que, dado que el descenso de gradiente no puede encontrar una dirección para descender más una vez que alcanza un lugar más bajo que su entorno, el aprendizaje podría detenerse en una cuenca poco profunda en lugar de en el valle más bajo. El artículo informó de que esto ocurría raramente en muchas tareas. Además, explicó que si se agregan unas cuantas unidades ocultas y conexiones más, se crean nuevas dimensiones en el espacio de pesos donde cada peso actúa como un eje de coordenadas, y esas dimensiones se convierten en vías para eludir la cuenca.

3.3. Representaciones aprendidas por la propia capa oculta: discriminación de simetría especular y la tarea de los 'árboles genealógicos (Family Trees)'

El impacto que causó el artículo de 1986 en el mundo académico provino más de los experimentos que de las fórmulas matemáticas. Los investigadores abrieron la capa oculta de la red neuronal una vez completado el entrenamiento y mostraron mediante figuras qué había aprendido cada neurona.

El primer experimento consistió en la discriminación de la simetría especular (Mirror Symmetry). Es el problema de determinar si un arreglo de seis dígitos compuesto de 0 y 1 es idéntico a izquierda y derecha respecto al centro. Por ejemplo, 101101 es simétrico y 101100 no lo es. En este problema no se puede deducir nada observando un solo bit de forma aislada. Es necesario comparar emparejando posiciones, como la primera con la sexta, o la segunda con la quinta. Pertenecía a la misma categoría que el problema de paridad (Parity), que determina si

hay un número impar de unos, y era una tarea representativa que un perceptrón monocapa no podía resolver. El problema de paridad en sí se abordó por separado en el capítulo sobre retropropagación del libro de PDP publicado ese mismo año.

Los investigadores entrenaron una red neuronal con solo dos unidades ocultas mostrándole todos los 64 arreglos posibles. Al presentar repetidamente los 64 arreglos 1425 veces, la red neuronal encontró por sí misma una solución elegante. Cada unidad oculta asignó pesos de igual magnitud y signo opuesto a las posiciones emparejadas simétricamente a ambos lados del centro. De afuera hacia adentro, las magnitudes de los pesos guardaban una proporción de 1 a 2 a 4.

El proceso mediante el cual esta estructura de pesos determina la simetría es el siguiente. Si el arreglo es completamente simétrico, los valores de la izquierda y la derecha se cancelan entre sí, de modo que la entrada total a la unidad oculta pasa a ser 0. Las unidades ocultas tienen un sesgo negativo, por lo que permanecen apagadas si la entrada total es 0, y cuando ambas unidades ocultas están apagadas, la unidad de salida responde

'simétrico'. Si no es simétrico, una de las dos unidades ocultas se activa y apaga la unidad de salida. Gracias a la proporción de 1, 2 y 4, nunca ocurre que un arreglo no simétrico sume 0 por casualidad. Los humanos no le enseñaron ninguna de estas reglas. Observando únicamente las respuestas correctas e incorrectas, la red neuronal trasladó la regla geométrica de la simetría a la estructura de pesos.

El núcleo del artículo es la tarea de los árboles genealógicos diseñada directamente por Hinton. Hinton construyó árboles genealógicos de 24 personas en total: 12 de una familia británica y 12 de una familia italiana. Ambos árboles genealógicos poseen exactamente la misma estructura. Por ejemplo, en el lugar de la pareja británica Christopher y Penelope se encuentra la pareja italiana Roberto y Maria. Había 12 tipos de relaciones: padre, madre, esposo, esposa, hijo, hija, tío, tía, hermano, hermana, sobrino y sobrina. Los datos de entrenamiento consistieron en 104 proposiciones con la estructura 'persona 1 + relación = persona 2'. Por ejemplo, que las tías de Colin son Jennifer y Margaret.

La red neuronal constaba de cinco capas. En la parte inferior había 24 unidades que representaban a la persona 1 y 12 unidades que representaban la relación. Dado que solo se activaba una unidad a la vez, fijándose únicamente en la entrada no había nada en común entre el británico Colin y el italiano Alfonso. Por encima de ellas se sucedían en orden 6 unidades ocultas para recibir a la persona y 6 unidades ocultas para recibir la relación, 12 unidades ocultas centrales, 6 unidades ocultas del lado de salida y 24 unidades de salida que representaban a la persona 2.

Los investigadores entrenaron la red con solo 100 de las 104 proposiciones y reservaron 4. Al examinar tras el entrenamiento las 6 unidades de la primera capa oculta que recibían a la persona, se comprobó que cada unidad se encargaba de un rasgo significativo. La unidad 1 distinguía entre británicos e italianos. La unidad 2 indicaba a qué generación pertenecía la persona. La unidad 6 indicaba a qué rama del árbol genealógico pertenecía. Si se excluía la unidad encargada de la nacionalidad, personas que ocupaban la misma posición en ambos árboles genealógicos, como Colin y Alfonso, obtenían

patrones de activación similares. Conceptos como nacionalidad, generación o rama no figuraban en ninguna parte de la entrada. Eran rasgos que la propia red neuronal descubrió en el proceso de predecir las relaciones.

La red neuronal tampoco recibió reglas como «el esposo de la madre es el padre». Aun así, acertó las respuestas correctas de las 4 proposiciones que jamás había visto durante el entrenamiento. Esta generalización fue posible gracias a la representación distribuida (Distributed Representation), que representa a una persona mediante una combinación de múltiples características. Era un enfoque diametralmente opuesto al método simbolista, que hace corresponder un único significado a cada símbolo.

Este experimento es valorado como un ejemplo pionero que demostró que las redes neuronales pueden aprender el significado de palabras y conceptos en forma de vectores numéricos. El artículo de revisión sobre aprendizaje profundo publicado en 2015 en «Nature» por Hinton, LeCun y Yoshua Bengio, así como el artículo publicado en 2021 en

«Communications of the ACM», explican que las incrustaciones de palabras (Word Embedding) actuales se sitúan en la prolongación de esta idea. Modelos posteriores como Word2Vec, los transformadores y los grandes modelos de lenguaje también se inscriben en el mismo linaje en el sentido de que procesan las palabras transformándolas en vectores.

3.4. La época en la Universidad Carnegie Mellon (CMU) y el traslado a Toronto, Canadá

Cuando se publicó el artículo de 1986, la filiación de Hinton era la Universidad Carnegie Mellon en Pittsburgh, Pensilvania, Estados Unidos. En 1982 se convirtió en profesor del Departamento de Ciencias de la Computación de dicha universidad y trabajó allí durante cinco años, hasta 1987. En aquel entonces, CMU era el centro de la inteligencia artificial simbolista liderada por Allen Newell y Herbert Simon. Según el artículo de 2019 de CMU con motivo del Premio Turing, Hinton llegó incluso a impartir clases junto a Newell, figura representativa del simbolismo. En medio de colegas con perspectivas distintas, él continuó con sus investigaciones en conexionismo.

La época en CMU fue un periodo de gran fecundidad investigadora. Junto con David Ackley y Terrence Sejnowski, Hinton investigó la máquina de Boltzmann (Boltzmann Machine), que tomó prestados conceptos de la física

estadística, y publicó su algoritmo de aprendizaje en 1985 en «Cognitive Science». En esa misma época, completó el artículo sobre retropropagación junto con Rumelhart y Williams. Los registros del bicentenario de George Boole en el University College Cork señalan que durante los cinco años que Hinton fue profesor en CMU, fue pionero en las investigaciones sobre la retropropagación, las máquinas de Boltzmann y las representaciones distribuidas de palabras.

Por aquel entonces, la situación política de Estados Unidos se movía en una dirección contraria a las convicciones de Hinton. La administración de Ronald Reagan impulsaba la Iniciativa de Defensa Estratégica (SDI), apodada 'Guerra de las Galaxias', en el contexto de la Guerra Fría con la Unión Soviética. La Agencia de Proyectos de Investigación Avanzados de Defensa (DARPA), dependiente del Departamento de Defensa, también destinaba un cuantioso presupuesto a la investigación en computación, incluida la inteligencia artificial. Gran parte de los fondos para la investigación en inteligencia artificial de las universidades estadounidenses procedía del estamento militar.

Hinton se consideraba socialista y se oponía al uso de la inteligencia artificial para la guerra. No deseaba recibir fondos de investigación militar. La Enciclopedia Británica señala que en la decisión de Hinton de abandonar Estados Unidos y marcharse a Canadá en 1987 influyó su rechazo hacia el ejército estadounidense y la administración Reagan, y que se oponía al uso de la inteligencia artificial en el combate. También existen registros de que la decepción con las políticas del gobierno de Reagan y la oposición al financiamiento militar para la investigación en inteligencia artificial figuraron entre los motivos de su traslado.

En 1987, Hinton renunció a su cátedra en CMU y se trasladó al Departamento de Ciencias de la Computación de la Universidad de Toronto, en Ontario, Canadá. La institución que lo acogió fue el Instituto Canadiense de Investigaciones Avanzadas (CIFAR). CIFAR es un organismo que apoya la investigación básica a largo plazo, con independencia de los beneficios comerciales inmediatos o de los fines militares. Según CIFAR, Hinton se incorporó en 1987 como becario (fellow) de su primer programa de investigación, 'Inteligencia Artificial, Robótica y Sociedad', y

con el apoyo de CIFAR y del Consejo de Investigaciones en Ciencias Naturales e Ingeniería de Canadá (NSERC), continuó sus investigaciones en un campo que en aquel entonces se veía afectado por el escepticismo y la escasez de fondos.

Los primeros 10 años de su etapa en Toronto no fueron fáciles. Al entrar en la década de 1990, otros métodos de aprendizaje automático, como las máquinas de vectores de soporte, acapararon la atención, y las redes neuronales quedaron relegadas nuevamente fuera del foco de interés. De 1998 a 2001, Hinton fundó y dirigió la Unidad de Neurociencia Computacional Gatsby en el University College London en Londres, antes de regresar a Toronto. En 2004, creó junto a sus colegas el nuevo programa 'Computación Neuronal y Percepción Adaptativa (NCAP)' en CIFAR. Este grupo se convirtió en un espacio de encuentro regular para investigadores de redes neuronales como LeCun y Bengio, y sirvió como base para que en 2012 AlexNet, desarrollada por Alex Krizhevsky e Ilya Sutskever en el laboratorio de Hinton, ganara por amplio margen una competición de reconocimiento de imágenes.

Posteriormente, los colaboradores de 1986 tomaron caminos diferentes. Ronald Williams publicó en 1992 el algoritmo REINFORCE, una técnica de gradiente de política para el aprendizaje por refuerzo. En 2001, Hinton se convirtió en el primer galardonado con el Premio Rumelhart, nombrado en honor a este último, y Rumelhart falleció en 2011. En 2024, Hinton recibió el Premio Nobel de Física junto a John Hopfield en reconocimiento a los descubrimientos e invenciones fundamentales que hicieron posible el aprendizaje automático basado en redes neuronales artificiales. El 8 de octubre de 2024, fecha del anuncio del Premio Nobel, CIFAR publicó un artículo de felicitación recordando el vínculo que mantenía con Hinton desde 1987.

El prolongado dilema de no disponer de un método para entrenar redes neuronales multicapa encontró el camino hacia su resolución con el artículo sobre retropropagación de 1986. La esencia que dejó aquel artículo residía en un cambio de perspectiva más que en una técnica de cálculo: que, incluso sin que un ser humano le proporcione reglas, la red neuronal encuentra por sí misma representaciones útiles a partir de los

datos. Con esta idea en mente, Hinton abandonó unos Estados Unidos colmados de presupuestos de investigación militar para trasladarse a Canadá, donde continuó sus investigaciones durante más de 20 años en Toronto. Los grandes modelos de lenguaje y los modelos de generación de imágenes actuales también se entrenan mediante retropropagación. No obstante, más adelante Hinton consideró que el cerebro real probablemente no aprende del mismo modo que la retropropagación, y en 2022 propuso alternativas como el algoritmo Forward-Forward.

Parte 2. Cómo hacer que las redes neuronales aprendan

Capítulo 4. Principios de aprendizaje hallados en la física

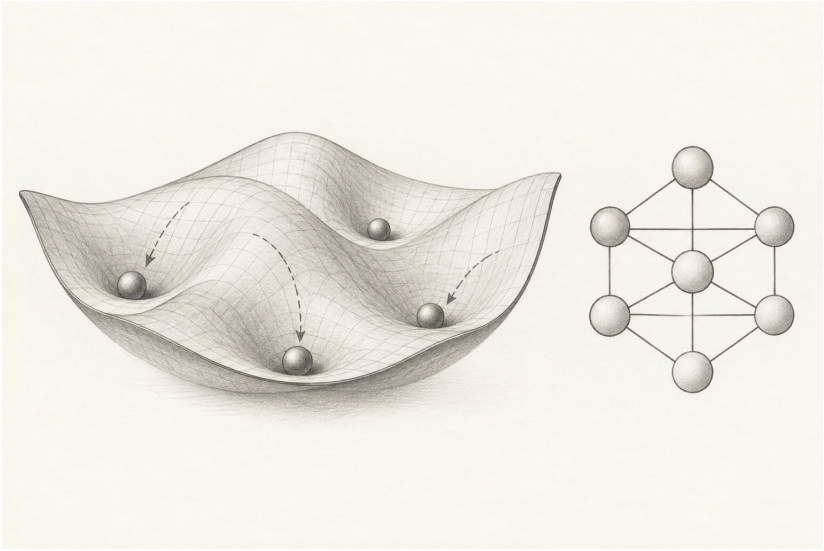


Figura 4. Una analogía del pensamiento de una red neuronal que utiliza energía y probabilidad.

4.1. El encuentro de Rochester: añadir probabilidad y calor a la red de Hopfield

A principios de la década de 1980, la comunidad académica de la inteligencia artificial se mostraba fría ante la investigación sobre redes neuronales. Después de que Marvin Minsky y Seymour Papert demostraran matemáticamente las limitaciones de las redes neuronales monocapa en Perceptrons (1969), muchos investigadores consideraron que las redes neuronales eran un callejón sin salida. Los fondos de investigación y el talento se concentraron en la inteligencia artificial simbolista, que buscaba implementar la inteligencia mediante reglas y símbolos. En este período, la persona que proporcionó un nuevo impulso a la investigación sobre redes neuronales no fue un científico de la computación, sino un físico.

John J. Hopfield era un físico teórico que transitaba entre la física del estado sólido y la biofísica. En 1982, mientras se encontraba en el Instituto de Tecnología de California (Caltech),

publicó en el volumen 79 de las Actas de la Academia Nacional de Ciencias (PNAS) un artículo titulado «Redes neuronales y sistemas físicos con habilidades computacionales colectivas emergentes» (Neural networks and physical systems with emergent collective computational abilities). Más tarde, Hopfield se trasladó a la Universidad de Princeton.

La herramienta que Hopfield tomó prestada en este artículo es un modelo físico llamado vidrio de espín (spin glass). Los átomos se comportan como pequeños imanes, y la orientación de ese imán se denomina espín. El espín adopta uno de dos estados: hacia arriba (+1) o hacia abajo (−1). En un imán ordinario, todos los espines vecinos se alinean en la misma dirección. Sin embargo, en un vidrio de espín, algunos pares prefieren la misma dirección, mientras que otros pares prefieren direcciones opuestas. Dado que no existe una configuración que satisfaga a todos los pares al mismo tiempo, el sistema se queda en uno de varios estados de compromiso. Los físicos representan estos estados estables como valles hundidos en un paisaje de energía.

Hopfield colocó neuronas en el lugar de los espines, ya que las neuronas también poseen dos estados: encendido y apagado. Al establecer esta correspondencia, la interacción entre espines se convierte en la fuerza de conexión entre neuronas, es decir, los pesos; la energía de todo el sistema se convierte en la energía del estado de la red; y los valles de energía se convierten en recuerdos almacenados. Hopfield impuso una condición: las conexiones debían ser simétricas. Sin importar qué dos neuronas se elijan, la fuerza de conexión que va de una a la otra debe ser igual a la que va en la dirección contraria. Bajo esta condición, demostró que cada vez que una neurona cambia de estado individualmente, la energía de toda la red disminuye o se mantiene igual. Dado que la energía no puede descender infinitamente, la red termina deteniéndose en el fondo de algún valle. A este fondo se le llama estado atractor (attractor state).

Aprovechando esta propiedad, la red puede utilizarse como un dispositivo de memoria. Si los patrones que se van a almacenar se configuran como fondos de valles, incluso si se recibe una entrada parcialmente borrada o con ruido, la red desciende

hasta el valle más cercano y restaura el patrón original. Es similar a escuchar solo el primer compás de una canción y recordar el resto de la letra. Este método de recuperar un recuerdo a partir de una parte del contenido, en lugar de una dirección de almacenamiento, se denomina memoria direccionable por contenido (content-addressable memory). El artículo de Hopfield demostró en el lenguaje de la física que las redes neuronales funcionan de manera estable, y sirvió como detonante para que los físicos se interesaran por las redes neuronales.

En la Universidad de Rochester, en el estado de Nueva York, se celebró un taller organizado por Jerry Feldman. En ese encuentro coincidieron Geoffrey Hinton, profesor de la Universidad Carnegie Mellon (CMU); Terrence Sejnowski, quien tras realizar una estancia posdoctoral en la Facultad de Medicina de Harvard se había establecido en la Universidad Johns Hopkins; y Hopfield. Hopfield presentó su demostración de que su red converge necesariamente hacia un estado estable.

Hinton y Sejnowski vieron en este resultado tanto posibilidades como debilidades. La debilidad provenía precisamente de la misma propiedad que garantizaba la convergencia. La red de Hopfield siempre se mueve únicamente en la dirección en que la energía disminuye. Por lo tanto, una vez que cae en un valle poco profundo cercano a la entrada inicial, no hay forma de salir de él. A este valle poco profundo se le denomina mínimo local (local minimum). Las respuestas obtenidas en los mínimos locales solían ser patrones extraños que nunca se habían almacenado, es decir, recuerdos falsos.

Ambos viajaban entre Pittsburgh y Baltimore para investigar juntos todos los fines de semana. Reunieron conocimientos de psicología, ciencias de la computación, neurociencia y física estadística para encontrar una forma de resolver esta debilidad. La respuesta que hallaron fue el calor. Hicieron que las neuronas decidieran su estado basándose en probabilidades, en lugar de seguir estrictamente una regla fija. Al elevar la temperatura, el ruido aumenta y las neuronas a veces se mueven incluso en la dirección en que la energía se incrementa.

De este modo, la red puede superar las paredes de los valles poco profundos y buscar valles más profundos.

Ambos dieron a esta nueva red el nombre del físico austriaco del siglo XIX Ludwig Boltzmann. Boltzmann fue quien sentó las bases de la mecánica estadística, deduciendo las leyes de la temperatura y la energía a partir del movimiento aleatorio de innumerables partículas. Hinton y Sejnowski presentaron por primera vez esta idea en una conferencia sobre visión por computador en 1983, a través de un artículo titulado «Inferencia perceptiva óptima» (Optimal perceptual inference).

4.2. Paisaje de energía, relajación estocástica y unidades ocultas

El perceptrón y la red de Hopfield operaban únicamente siguiendo reglas fijas. La máquina de Boltzmann transformó esto en un modelo basado en energía que realiza inferencias mediante probabilidades. Un modelo basado en energía asigna un número denominado energía a cada estado posible en el que puede encontrarse la red, y considera más verosímiles aquellos estados con un valor más bajo. Su estructura parte de las reglas que determinan la energía. En la máquina de Boltzmann, el estado de cada neurona (unidad) es 0 para apagado o 1 para encendido. La energía total de la red se define de la siguiente manera.

Primero, se toman como un par dos unidades conectadas entre sí y se multiplica el estado de ambas por el peso de la conexión entre ellas. Solo cuando ambas unidades están encendidas este producto es igual al peso; si al menos una está apagada, se vuelve 0. Se suman estos valores para todos los pares y luego se invierte el signo. A esto se le suma el umbral de todas las

unidades que están encendidas, lo cual constituye la energía total de la red.

El significado de esta regla es el siguiente: si dos unidades fuertemente conectadas con un peso positivo están encendidas al mismo tiempo, la energía disminuye en esa medida. Esto significa que cuanto mejor coincida la relación entre dos unidades, más estable será el sistema. El umbral es la magnitud de entrada requerida para que cada unidad se encienda. Una unidad con un umbral alto incrementa la energía en esa misma medida cuando está encendida, por lo que no se enciende con facilidad.

Si se hace funcionar esta red durante mucho tiempo a una temperatura constante, el sistema alcanza el equilibrio térmico (thermal equilibrium). Aquí, la temperatura es un valor numérico que determina qué tan aleatoriamente se mueven las unidades, un concepto tomado de la temperatura en física. El equilibrio térmico es un estado en el cual, aunque el estado de las unidades individuales cambie continuamente, la proporción con la que aparece cada estado global se mantiene constante.

El estado global se refiere a una configuración única que describe conjuntamente el encendido y apagado de todas las unidades. En este punto, la probabilidad de que la red se encuentre en un determinado estado global sigue la distribución de Boltzmann.

En la distribución de Boltzmann se asigna un valor a cada estado, y este valor disminuye rápidamente de forma exponencial a medida que aumenta el resultado de dividir la energía de ese estado por la temperatura. La probabilidad de que aparezca un estado es el valor asignado a ese estado dividido por la suma de los valores asignados a todos los estados. Al dividir de esta manera, la suma de las probabilidades de todos los estados es igual a 1.

Según esta distribución, cuanto menor sea la energía de un estado, mayor será su probabilidad de aparecer, mientras que los estados de alta energía aparecen con poca frecuencia. La temperatura regula la magnitud de esta diferencia. Cuando la temperatura es alta, la diferencia disminuye y diversos estados aparecen de manera uniforme; cuando la temperatura es baja,

el estado de menor energía aparece con mucha mayor frecuencia que los demás. El proceso mediante el cual las neuronas cambian de estado de acuerdo con estas probabilidades y van encontrando gradualmente una configuración estable se denomina relajación estocástica.

Entonces, ¿cómo se debe regular la temperatura para encontrar valles profundos? Hinton y Sejnowski adoptaron un método llamado recocido simulado (simulated annealing). Si se calienta un metal y luego se enfría lentamente, su estructura interna puede acercarse a un estado estable. En 1983, Scott Kirkpatrick, investigador de IBM, y sus colaboradores publicaron en Science la técnica del recocido simulado, que aplicaba este principio a problemas computacionales. Al principio, se eleva la temperatura para permitir que la red atraviese libremente las altas colinas del paisaje de energía y explore una amplia gama de posibilidades. Luego, al reducir la temperatura muy lentamente, aumenta la probabilidad de que el sistema se establezca en el valle más profundo del relieve, es decir, el mínimo global (global minimum). Hacia la misma época, la teoría de la armonía de Paul Smolensky y las investigaciones

sobre eliminación de ruido en imágenes de los hermanos Stuart Geman y Donald Geman aplicaron la misma idea a problemas diferentes.

El logro fundamental de la máquina de Boltzmann es un método de aprendizaje para redes neuronales que incluyen unidades ocultas. Las unidades ocultas son aquellas que operan únicamente dentro de la red, sin recibir datos del exterior ni emitir resultados. La razón por la que se necesitaban unidades ocultas se hace evidente al observar las críticas de Minsky y Papert. El perceptrón y la red de Hopfield solo podían manejar relaciones entre unidades que estaban en contacto directo con el mundo exterior. Sin embargo, incluso un problema que parece simple como la disyunción exclusiva (XOR), que devuelve 0 si dos entradas son iguales y 1 si son diferentes, no puede resolverse únicamente mediante relaciones directas entre las entradas. Para representar relaciones de alto orden que no se revelan al emparejar elementos de dos en dos, sino que emergen al considerar múltiples elementos juntos, se necesitan unidades intermedias que organicen esta información.

La máquina de Boltzmann respondió a esta exigencia dividiendo las unidades en dos tipos. Las unidades visibles (visible units) están en contacto con el mundo exterior, por lo que reciben datos de entrada o proporcionan respuestas. A las unidades ocultas (hidden units) no se les asignan valores desde el exterior ni se las puede observar. Las unidades ocultas están conectadas de manera simétrica con las unidades visibles y descubren y representan por sí mismas las reglas y características subyacentes a los datos.

El problema restante radicaba en cómo entrenar los pesos de las unidades ocultas. Cómo entrenar unidades a las que no se les puede proporcionar una respuesta correcta había sido un desafío de larga data en la investigación sobre redes neuronales. La máquina de Boltzmann ofreció una respuesta a este dilema mediante la probabilidad y la física estadística, contribuyendo a allanar el camino para el aprendizaje en redes neuronales multicapa.

4.3. Fase de vigilia y fase de sueño: el algoritmo de aprendizaje de la máquina de Boltzmann

En 1985, David Ackley, Hinton y Sejnowski publicaron en el volumen 9 de la revista Cognitive Science el artículo «Un algoritmo de aprendizaje para máquinas de Boltzmann» (A learning algorithm for Boltzmann machines). Al año siguiente, Hinton y Sejnowski explicaron este contenido en detalle en el capítulo 7, «Aprendizaje y reaprendizaje en máquinas de Boltzmann», del volumen 1 de Parallel Distributed Processing, editado por el grupo de investigación PDP.

El objetivo de aprendizaje de la máquina de Boltzmann consiste en hacer coincidir dos distribuciones de probabilidad. Una distribución de probabilidad es una recopilación de la frecuencia con la que aparece cada estado posible. Una de ellas es la distribución que los datos del mundo real generan en las unidades visibles, y la otra es la distribución que se genera en las unidades visibles cuando la red opera por sí sola sin entradas externas. Las unidades visibles son las unidades en

contacto directo con el mundo exterior que reciben los datos. La diferencia entre las dos distribuciones se mide mediante la divergencia de Kullback-Leibler (Kullback-Leibler divergence) de la teoría de la información. Los autores denominaron G a este valor.

Este valor se obtiene realizando el siguiente cálculo para cada estado posible de las unidades visibles y sumando los resultados. Se divide la probabilidad de que ese estado aparezca en los datos reales por la probabilidad de que aparezca cuando la red opera de manera autónoma, se calcula el logaritmo natural de esa razón y luego se multiplica por la probabilidad en los datos reales. Para un estado donde ambas probabilidades son iguales, la razón es 1 y el logaritmo de 1 es 0, por lo que no se añade nada. Si hay estados donde las dos probabilidades difieren, la suma se vuelve mayor que 0.

Si este valor es 0, la distribución generada por la red coincide con la distribución de los datos reales. Esto significa que la red reproduce fielmente la imagen del mundo contenida en los datos. En consecuencia, el aprendizaje consiste en ajustar los

pesos para reducir este valor. Hinton y Sejnowski demostraron que al diferenciar este valor respecto a cada peso —es decir, al calcular cuánto cambia este valor cuando un peso se modifica ligeramente— se obtiene un resultado conciso.

Ese resultado se expresa como la diferencia entre dos probabilidades. Una es la probabilidad de que dos unidades conectadas estén encendidas juntas cuando se alcanza el equilibrio manteniendo las unidades visibles fijadas a los datos reales; la otra es la probabilidad de que esas dos unidades estén encendidas juntas cuando la red opera por sí sola sin entrada externa y alcanza el equilibrio. El grado en que cambia la diferencia entre las dos distribuciones al modificar ligeramente un peso equivale a restar la segunda probabilidad de la primera, dividir ese valor por la temperatura e invertir el signo.

De este resultado se deriva directamente la regla para modificar los pesos.

La cantidad de ajuste para el peso que conecta dos unidades es el producto de la diferencia entre dos valores medidos

multiplicada por un número pequeño. La diferencia se obtiene restando el segundo valor del primero.

El número pequeño por el que se multiplica determina cuánto se ajusta el peso en cada paso. El primer y el segundo valor se miden por separado en dos fases diferentes.

La primera es la fase positiva (positive phase), comparada con una fase de vigilia. Se mantienen las unidades visibles fijadas a los datos reales y se permite que solo las unidades ocultas se muevan libremente. Cuando se alcanza el equilibrio térmico mediante el recocido simulado, se mide la probabilidad promedio de que ambas unidades estén encendidas juntas. Este valor es el primer término de la regla de modificación y corresponde a las conexiones que se forman mientras se percibe el mundo.

La segunda es la fase negativa (negative phase), comparada con una fase de sueño. Se libera la fijación de las unidades visibles y se deja que toda la red funcione de forma autónoma. Al alcanzar nuevamente el equilibrio térmico, se mide la

probabilidad promedio de que ambas unidades estén encendidas al mismo tiempo. Este valor es el segundo término de la regla de modificación y corresponde a las conexiones generadas por la propia red sin entradas externas.

Al comparar los dos valores, el significado de la regla se vuelve claro. Si dos unidades se encienden juntas con mayor frecuencia al observar los datos reales que durante la fase de sueño, la conexión entre ellas se fortalece. Si se encienden juntas solo en la fase de sueño pero no en los datos reales, la conexión se debilita. Se refuerza lo aprendido de la realidad y se eliminan las asociaciones infundadas creadas por la red por su cuenta.

La ventaja de esta regla es la localidad. La única información necesaria para modificar la fuerza de una conexión es la actividad de las dos unidades situadas en ambos extremos de esa conexión. No hay necesidad de recibir señales de error propagadas hacia atrás desde el otro lado de la red. La misma regla se aplica incluso si las unidades ocultas están dispuestas en múltiples capas. Dado que las sinapsis del cerebro también

solo tienen acceso a la actividad de las dos neuronas adyacentes, esta regla parecía biológicamente plausible.

La analogía con el sueño también coincidía con las hipótesis de la investigación sobre el sueño de aquella época. En 1983, Francis Crick y Graeme Mitchison publicaron en Nature una hipótesis sobre la función del sueño con sueños. Planteaban que durante el sueño REM el cerebro lleva a cabo un desaprendizaje inverso (reverse learning) para borrar asociaciones inútiles. La fase negativa de la máquina de Boltzmann encajaba muy bien con esta hipótesis. Crick asistía con frecuencia a los seminarios del grupo PDP y debatía con los investigadores de redes neuronales.

Sin embargo, este algoritmo tenía un gran obstáculo. Tardaba demasiado tiempo en alcanzar el equilibrio térmico. En cada fase positiva y negativa, las unidades debían actualizarse de forma probabilística entre miles y decenas de miles de veces para obtener muestras cercanas al equilibrio. Este muestreo iterativo se conoce como el método de cadenas de Markov Monte Carlo (MCMC). Debido a este costo computacional, en

aquel entonces resultaba difícil entrenar redes neuronales grandes de manera práctica. En un período similar, en 1986, el algoritmo de retropropagación del error que Hinton publicó en Nature junto con David Rumelhart y Ronald Williams era mucho más rápido. Durante un tiempo, la atención de los investigadores se volcó hacia la retropropagación.

Hinton no abandonó la máquina de Boltzmann. Continuó investigando métodos para reducir la cantidad de cómputo restringiendo las conexiones entre unidades. Estas investigaciones condujeron más tarde a la máquina de Boltzmann restringida (RBM), que eliminaba las conexiones dentro de la misma capa, y al método de aprendizaje por divergencia contrastiva (CD).

4.4. El Premio Nobel de Física de 2024: el encuentro entre la física estadística y la inteligencia artificial

El 8 de octubre de 2024, poco más de 40 años después de que se presentara por primera vez la máquina de Boltzmann, la Real Academia de las Ciencias de Suecia anunció a John Hopfield y Geoffrey Hinton como ganadores conjuntos del Premio Nobel de Física. El motivo de la concesión fue «por descubrimientos e invenciones fundamentales que hacen posible el aprendizaje automático con redes neuronales artificiales» (for foundational discoveries and inventions that enable machine learning with artificial neural networks). En aquel momento, Hopfield era profesor emérito de la Universidad de Princeton e Hinton era profesor emérito de la Universidad de Toronto. Hinton reveló que recibió la noticia del galardón por teléfono mientras se encontraba en un hotel económico en California, Estados Unidos.

El Comité Nobel describió en sus materiales de difusión los logros de ambos como una corriente continua e

interconectada. Hopfield utilizó el concepto de energía del espín atómico para crear una red capaz de almacenar y restaurar patrones. Basándose en esa red, Hinton incorporó herramientas de la física estadística para inventar la máquina de Boltzmann, capaz de descubrir y clasificar por sí misma características de los datos e incluso de generar nuevos ejemplos. La máquina de Boltzmann fue mencionada directamente en los documentos oficiales que explicaban la motivación del premio. Anders Irbäck, miembro del comité, señaló en la conferencia de anuncio que la investigación de ambos tiene sus raíces en la física. La presidenta del comité, Ellen Moons, declaró que Hinton es reconocido como una figura destacada en el desarrollo de algoritmos de aprendizaje eficientes.

Inmediatamente después del anuncio, también surgió cierta controversia. Algunos físicos y comunidades académicas en línea debatieron sobre si esta investigación correspondía a la física orientada a dilucidar fenómenos naturales, o bien a las ciencias de la computación o la estadística. Al seguir la genealogía disciplinar, la investigación de ambos tiene sus

raíces en la teoría de sistemas desordenados y el estudio de los vidrios de espín dentro de la física estadística. El trabajo del físico italiano Giorgio Parisi, galardonado con el Premio Nobel de Física en 2021, también se centró en la investigación de sistemas complejos desordenados como los vidrios de espín. En el sentido de que explican cómo surge un orden a gran escala a partir de las interacciones estocásticas de innumerables unidades pequeñas, los vidrios de espín de Parisi y la máquina de Boltzmann de Hinton comparten la misma pregunta de fondo. La única diferencia es que, en la máquina de Boltzmann, ese orden se manifiesta en forma de memoria y aprendizaje.

Tras la concesión del galardón, Hinton recibió repetidas solicitudes de periodistas para que explicara brevemente su investigación. Según se relató, respondió recurriendo a una conocida broma del físico Richard Feynman: si una investigación pudiera explicarse en unos pocos minutos, no sería merecedora de un Premio Nobel.

El 10 de diciembre de 2024 se celebró la ceremonia de entrega de premios en la Sala de Conciertos de Estocolmo. Hinton

recibió el Premio Nobel de Física de manos del rey Carlos XVI Gustavo de Suecia, mientras que en la Universidad de Toronto unos 100 estudiantes y colegas se reunieron para presenciar la ceremonia. El premio en metálico fue de 11 millones de coronas suecas, aproximadamente 1,4 millones de dólares canadienses, que compartió con Hopfield. Hinton anunció que donaría parte de su cuota a Water First, una organización de Ontario dedicada a mejorar el acceso al agua potable de las comunidades indígenas de Canadá, y a una organización que apoya a jóvenes con neurodivergencia.

Otra parte del premio se destinó a conmemorar la investigación realizada conjuntamente con Sejnowski. En diciembre de 2024, la Conferencia sobre Sistemas de Procesamiento de Información Neuronal (NeurIPS) anunció que Hinton crearía un nuevo premio con parte de los fondos de su galardón. El premio otorga anualmente 10.000 dólares estadounidenses y se concede a equipos de dos o más investigadores menores de 40 años que hayan escrito un artículo proponiendo una nueva teoría sobre cómo funciona el cerebro. El galardón lleva el nombre de Premio Sejnowski-Hinton (Sejnowski-Hinton Prize),

en honor al neurobiólogo computacional Terrence Sejnowski e Hinton. Los nombres de los dos investigadores que hace más de 40 años viajaban entre Pittsburgh y Baltimore para crear juntos la máquina de Boltzmann se unieron así en un premio para jóvenes investigadores.

Hinton también dejó en el Museo Nobel un objeto relacionado con la máquina de Boltzmann. El microchip que donó en 2024 alberga una máquina de Boltzmann de los primeros tiempos. Este chip fue creado en 1987 por Joshua Alspector y demostró que la máquina de Boltzmann podía implementarse en un chip dedicado. La máquina de Boltzmann integrada en el chip era una red pequeña de solo seis nodos, por lo que su utilidad práctica era limitada; sin embargo, como ejemplo que demostraba el principio, señaló el rumbo para el desarrollo del aprendizaje automático y la inteligencia artificial.

A principios de la década de 1980, cuando la gran mayoría de los académicos daba la espalda a las redes neuronales, Hopfield demostró su estabilidad valiéndose de los vidrios de espín. Hinton y Sejnowski incorporaron a esto el calor y la

probabilidad, y añadieron unidades ocultas para crear una máquina capaz de aprender por sí misma. La máquina de Boltzmann era demasiado lenta, por lo que no se utilizó ampliamente en aquel entonces. No obstante, la concepción de representar datos mediante la energía, reducir la discrepancia entre la realidad y el modelo, y descubrir por sí misma estructuras ocultas, se transmitió a través de las máquinas de Boltzmann restringidas y las redes de creencia profunda hasta llegar a la inteligencia artificial generativa actual. Esta idea, originada en el lenguaje de la física, fue finalmente reconocida 40 años más tarde con el Premio Nobel de Física.

Capítulo 5. Cómo aprender más rápido

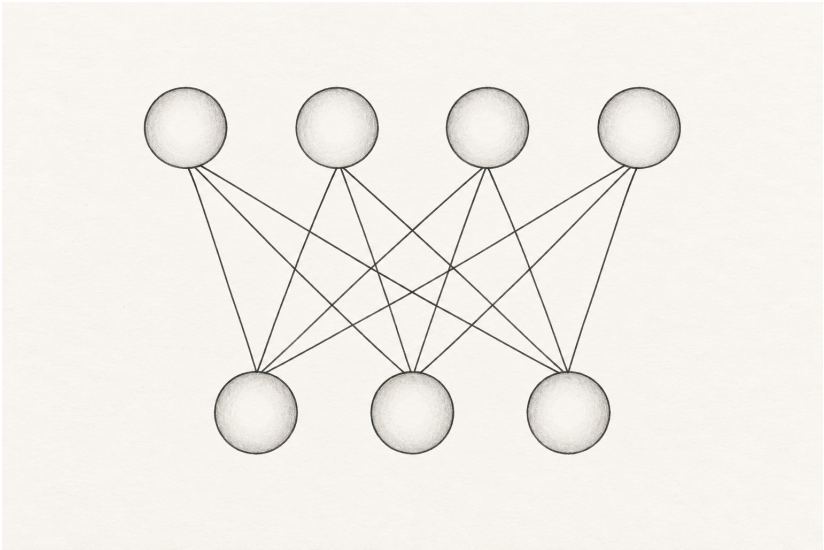


Figura 5. Diagrama conceptual de una máquina de Boltzmann restringida configurada con conexiones entre dos capas.

La máquina de Boltzmann, surgida en la década de 1980, demostró que las redes neuronales con unidades ocultas también podían aprender por sí mismas. Sin embargo, el cálculo tomaba demasiado tiempo, por lo que resultaba difícil aplicarla a problemas prácticos. Durante casi 15 años después de eso, Hinton se aferró a este problema y presentó dos respuestas. Una fue la máquina de Boltzmann restringida (Restricted Boltzmann Machine, RBM), que redujo deliberadamente las conexiones de la red neuronal, y la otra fue la divergencia contrastiva (Contrastive Divergence, CD), que permite entrenar rápidamente esta red neuronal. Ambas ideas partieron de una única pregunta: por qué la máquina de Boltzmann era tan lenta.

5.1. Una máquina de Boltzmann demasiado lenta y la idea del 'producto de expertos'

Una red neuronal es una estructura en la que pequeñas unidades de cálculo, llamadas unidades, están conectadas entre sí. La mayoría de las unidades en este capítulo toman uno de dos valores: 1, que significa encendido, o 0, que significa apagado. A la conexión que une dos unidades se le asocia un número llamado peso. Si el peso es positivo, ambas unidades tienden a encenderse juntas; si es negativo, cuando una se enciende, la otra tiende a apagarse. El aprendizaje consiste en ajustar estos pesos poco a poco. La probabilidad representa con qué frecuencia ocurre un suceso mediante un número entre 0 y 1; cuanto más cercano a 1, con mayor frecuencia ocurre.

La máquina de Boltzmann, presentada a mediados de la década de 1980 por Geoffrey Hinton y Terry Sejnowski, es un modelo que trasladó conceptos de la física estadística a las redes neuronales. Las unidades de este modelo no se encienden ni se apagan según reglas fijas, sino de acuerdo con probabilidades.

El modelo asigna un valor de energía a cada estado general de la red neuronal que describe qué unidades están encendidas y cuáles apagadas, y define las probabilidades de modo que los estados con menor energía aparezcan con mayor frecuencia.

El aprendizaje se lleva a cabo en dos etapas. En primer lugar, en la 'fase despierta', donde los datos reales se fijan en las unidades de entrada, se mide el grado en que dos unidades conectadas se encienden juntas. A continuación, en la 'fase de sueño', donde se libera la entrada y se deja que la red neuronal funcione por sí misma, se vuelve a medir el mismo valor. Luego, el peso de esa conexión se ajusta en una cantidad proporcional a la diferencia entre ambos valores. Las conexiones que se encienden juntas con más frecuencia cuando la red está despierta se fortalecen, mientras que las que se encienden juntas con más frecuencia cuando sueña se debilitan.

El problema residía en la fase de sueño. Para calcular la probabilidad de que aparezcan determinados datos, es necesario conocer la función de partición (Partition Function), valor que suele denotarse como Z. Z es la suma de todas las

energías de los posibles estados que puede adoptar la red neuronal tras introducirlas en una función exponencial, de manera que a menor energía corresponda un valor mayor. Dado que la probabilidad de un estado es el cociente de dividir el valor de dicho estado entre Z, no es posible determinar la probabilidad sin Z. Sin embargo, como cada unidad puede tomar dos valores, el número de estados equivale a 2 elevado al número de unidades. Aunque solo hubiera 100 unidades en total, 2 elevado a la potencia de 100 equivale aproximadamente a 10 elevado a la 30. Ninguna computadora es capaz de sumar semejante cantidad de términos uno por uno.

Por ello, los investigadores utilizaron el muestreo en lugar de calcular Z directamente. El muestreo consiste en extraer valores reales de acuerdo con probabilidades. Al reunir los resultados de múltiples extracciones, es posible saber qué estados aparecen con frecuencia sin necesidad de examinar todos los estados. En la máquina de Boltzmann se emplearon un método de muestreo llamado cadenas de Markov Monte Carlo (MCMC) y el recocido simulado (Simulated Annealing). Este método consistía en cambiar repetidamente los valores de las unidades

según probabilidades y agitar la red neuronal durante un tiempo prolongado hasta alcanzar el equilibrio térmico, donde la proporción de aparición de cada estado ya no cambia, para luego medir las estadísticas en ese punto. Sin embargo, como había que pasar por este proceso cada vez que se ajustaban los pesos, con las computadoras de las décadas de 1980 y 1990 el entrenamiento se prolongaba indefinidamente. En su conferencia del Premio Nobel de 2024, Hinton también recordó que la máquina de Boltzmann original era demasiado lenta y que, unos años después, ideó la máquina de Boltzmann restringida, que aumentó la eficiencia al limitar las conexiones entre neuronas.

Para resolver este problema, Hinton reconsideró desde el principio la forma de construir modelos probabilísticos. El resultado fue el 'producto de expertos' (Products of Experts, PoE). Esta investigación surgió durante la época en que Hinton se desempeñó como director fundador de la Unidad de Neurociencia Computacional Gatsby del University College de Londres (UCL), entre 1998 y 2001. Se dio a conocer primero como informe técnico en 2000 y se publicó formalmente como

artículo en la revista académica Neural Computation en 2002, en el volumen 14, número 8.

Aquí, un experto se refiere a un pequeño modelo que asigna una probabilidad a la verosimilitud de cada dato. La manera en que la probabilidad se distribuye entre todos los datos posibles se denomina distribución de probabilidad. En aquel momento, se utilizaban ampliamente los modelos de mezclas (Mixture Models), como los modelos de mezclas gaussianas (GMM), que sumaban las distribuciones de múltiples expertos. Este método consistía en asignar una ponderación a cada experto, multiplicar la probabilidad calculada por cada uno por dicha ponderación y sumar todos los resultados.

Hinton señaló que este método no se adaptaba bien a datos de alta dimensionalidad, es decir, datos que requieren muchos números para representar un solo elemento. Tomando como ejemplo la fotografía de un rostro humano, factores como la iluminación, la postura, la expresión y el ángulo se entrelazan para formar un espacio curvo (variedad) de unas 35 dimensiones. Para representar este espacio con un modelo de

mezclas, cada experto debe cubrir ampliamente la totalidad de las 35 dimensiones. Sin embargo, el resultado de sumar múltiples distribuciones no puede ser más agudo que las distribuciones que sirvieron de ingredientes. En última instancia, la distribución general se vuelve difusa.

Por el contrario, el producto de expertos no suma las probabilidades asignadas por los expertos, sino que las multiplica. Como el valor multiplicado no constituye una probabilidad por sí mismo, se divide por la suma de los mismos productos obtenidos para todos los datos posibles, ajustándolo para que la suma total sea 1. Este proceso se denomina renormalización. En el método de multiplicación, basta con que un solo experto asigne una probabilidad baja para que el resultado final disminuya drásticamente. Por lo tanto, cada experto solo necesita imponer condiciones sobre una parte de los rasgos de los datos, ya que la probabilidad seguirá siendo alta únicamente en la estrecha región donde se satisfagan todas las condiciones simultáneamente.

Hinton puso como ejemplo una oración. Un experto solo comprueba si el tiempo verbal es correcto, mientras que otro experto solo verifica si el sujeto y el verbo concuerdan en número. Al multiplicar ambos, solo se asigna una probabilidad alta a las oraciones que cumplen ambas reglas. Mientras que en un modelo de mezclas cada experto debía abordar el problema en su totalidad, en el producto de expertos aquellos que se encargan de condiciones específicas crean juntos una distribución nítida y precisa.

El producto de expertos tenía una ventaja adicional. Dados los datos, las variables ocultas dentro de cada experto se vuelven independientes entre sí. La independencia significa que conocer el valor de un lado no altera la probabilidad del otro y, dado que se cumple bajo la condición de que se suministren los datos, se denomina independencia condicional. Gracias a esto, el estado de las variables ocultas puede inferirse con precisión mediante un solo cálculo. Sin embargo, el valor divisor en la renormalización sigue siendo difícil de calcular porque aún requiere sumar todos los casos posibles. Hinton sorteó este

problema pendiente mediante la divergencia contrastiva, que se explicará en la sección 5.3.

5.2. Estructura de la máquina de Boltzmann restringida (RBM) e independencia condicional

La máquina de Boltzmann restringida trasladó la idea del producto de expertos a la estructura de una red neuronal. Esta estructura en sí no fue creada por primera vez por Hinton, y sus raíces se remontan a 1986. Paul Smolensky presentó un modelo llamado Harmonium al introducir la 'teoría de la armonía' en el capítulo 6 del volumen 1 de Parallel Distributed Processing, publicado ese año. En 1992, Yoav Freund y David Haussler investigaron un método para aprender la distribución de vectores binarios, es decir, datos compuestos por secuencias de ceros y unos, mediante una red neuronal de dos capas. En su artículo de 2002, Hinton señaló que esta estructura era un producto de expertos que consideraba cada unidad oculta como un experto, y estableció una teoría de aprendizaje combinándola con el marco de energía de la máquina de Boltzmann.

La RBM consta de solo dos capas. La capa visible (visible layer) es la capa que recibe los datos, y la capa oculta (hidden layer) es la que busca características no evidentes de forma directa en los datos. Cada unidad de la capa oculta, es decir, cada unidad oculta, se entrena para encenderse cuando aparece un patrón específico en la capa visible. En las conexiones entre las unidades visibles y las unidades ocultas existen pesos simétricos que utilizan el mismo valor independientemente de la dirección en que viaje la señal. Cada unidad cuenta también con un número llamado sesgo, que determina la propensión de esa unidad a encenderse con independencia de las demás unidades.

La diferencia entre la RBM y la máquina de Boltzmann general radica en que se eliminaron todas las conexiones dentro de una misma capa. Las unidades visibles no están conectadas entre sí, ni tampoco las unidades ocultas entre sí. En matemáticas, esta configuración se denomina grafo bipartito (Bipartite Graph). Por analogía con un aula de clases, equivale a una regla en la que los estudiantes de la fila delantera solo pueden hablar con

los de la fila trasera, pero no con los compañeros sentados a su lado en la misma fila.

La energía de una RBM es el valor resultante de sumar tres elementos y cambiarles el signo: los sesgos de las unidades visibles encendidas, los sesgos de las unidades ocultas encendidas y los pesos de las conexiones donde ambas unidades están encendidas a la vez. Cuanto más se encienden juntas dos unidades de una conexión con un peso grande, menor es la energía y, de acuerdo con las reglas de la distribución de Boltzmann, mayor es la probabilidad de ese estado.

El efecto de eliminar las conexiones dentro de la misma capa es considerable. Una vez determinados todos los valores de la capa visible, las unidades ocultas no se influyen entre sí. Esto se debe a que no existe una vía directa entre las unidades ocultas y el único canal, que es la capa visible, ya tiene sus valores fijados. La probabilidad de que cada unidad oculta se encienda se calcula por separado utilizando únicamente la señal que recibe. Se suman todos los pesos conectados a las unidades

visibles encendidas, se añade su propio sesgo y se introduce esa suma en la función sigmoide. La función sigmoide devuelve un valor entre 0 y 1 para cualquier número introducido, y cuanto mayor sea el número introducido, más cercano a 1 será el resultado. Este valor es la probabilidad de que dicha unidad se encienda.

Lo mismo ocurre en la dirección opuesta. Una vez determinados todos los valores de la capa oculta, las unidades visibles se vuelven independientes entre sí, y la probabilidad de que cada unidad visible se encienda se calcula de la misma manera: mediante la suma de los pesos procedentes de las unidades ocultas encendidas más su propio sesgo. Por lo tanto, la probabilidad de que una capa completa adopte determinados valores es el producto de las probabilidades de cada unidad individual. Esto se denomina independencia condicional factorial (Factorial Conditional Independence). Esta independencia solo se cumple bajo la condición de que se conozcan todos los valores de la capa opuesta. Sin esa condición, las unidades de la misma capa también se influyen mutuamente a través de la capa opuesta.

El valor de esta propiedad se vuelve evidente al compararla con los modelos gráficos dirigidos. En los modelos donde las flechas van de las causas a los efectos, surge un fenómeno llamado 'descarte de explicaciones' (Explaining Away). Consideremos dos posibles razones por las que el césped está mojado: la lluvia y los aspersores. En principio, estas dos causas no guardan relación entre sí. Sin embargo, tras constatar que el césped está mojado, en el momento en que se confirma que llovió, disminuye la probabilidad de que los aspersores hayan funcionado. Al observar el resultado, las causas se entrelazan entre sí. Para calcular la probabilidad de causas ocultas en tales modelos, era necesario recurrir a MCMC o a complejos cálculos de aproximación.

En la RBM no se produce este entrelazamiento. Cuando ingresan los datos de entrada, la probabilidad de toda la capa oculta se obtiene con exactitud mediante una sola multiplicación de matriz por vector y un solo cálculo sigmoide. Sin embargo, esto se limita a la inferencia para extraer características ocultas a partir de datos dados. Para entrenar los pesos, Z seguía siendo un obstáculo.

5.3. Principio del método de aprendizaje por divergencia contrastiva (CD)

El entrenamiento de una RBM consiste en ajustar los pesos para aumentar la probabilidad de que aparezcan los datos reales. La dirección y magnitud con que se ajusta cada conexión se determinan por la diferencia entre dos promedios. Uno es la proporción en que la unidad visible y la unidad oculta de esa conexión se encienden juntas cuando los datos se fijan en la capa visible. El otro es la proporción en que ambas unidades se encienden juntas cuando la red neuronal opera por sí misma sin datos. Si el primer valor es mayor que el segundo, el peso aumenta; si es menor, disminuye. Cuando ambos valores se igualan en todas las conexiones, las estadísticas generadas por el modelo por sí solo coinciden con las estadísticas de los datos, por lo que ya no queda nada que ajustar.

El primer valor se obtiene con facilidad mediante el cálculo sigmoide descrito en la sección 5.2. Lo difícil es el segundo valor. Para obtener este valor con exactitud, es necesario

realizar el muestreo de Gibbs (Gibbs Sampling). Este método consiste en calcular alternativamente las probabilidades de la capa oculta a partir de los valores de la capa visible para extraer los valores de las unidades ocultas, y luego calcular las probabilidades de la capa visible a partir de dichos valores para extraer los valores de las unidades visibles. No obstante, este vaivén debía repetirse teóricamente de manera infinita hasta alcanzar el equilibrio. Aunque la inferencia se aceleró al reducir las conexiones, el cuello de botella del entrenamiento permanecía intacto.

Hinton cortó resueltamente este cuello de botella. Consideró que no hacía falta llegar hasta el equilibrio, sino que se podía obtener una dirección de aprendizaje útil partiendo de los datos reales y deteniéndose tras realizar un único paso de muestreo de Gibbs. El método que repite este proceso una sola vez se denomina CD-1, y su procedimiento consta de cuatro pasos.

Primero, se introducen los datos reales en la capa visible. Se calcula la probabilidad de que cada unidad oculta se encienda y

se extrae un valor de 0 o 1 según esa probabilidad. En este momento, se mide para cada conexión el grado en que la unidad visible y la unidad oculta se encienden juntas.

Segundo, a partir de los valores obtenidos para la capa oculta, se calculan a la inversa las probabilidades de la capa visible para generar nuevos valores de la capa visible. Este es el resultado de que la red neuronal redibuje los datos a su propia manera. Hinton llamó a esto reconstrucción (reconstruction) y a veces confabulación (confabulation), con el sentido de 'historia inventada'.

Tercero, se introduce de nuevo la reconstrucción en la capa visible para obtener los valores de la capa oculta, y se mide de la misma manera el grado en que ambos lados se encienden juntos.

Cuarto, los pesos se ajustan multiplicando la tasa de aprendizaje por la diferencia resultante de restar el valor medido en el tercer paso del valor medido en el primer paso. La tasa de aprendizaje es un número pequeño que determina

cuánto se avanza en cada paso. Las conexiones que se encienden juntas con más frecuencia en los datos ven incrementado su peso, mientras que las que se encienden juntas con más frecuencia en la reconstrucción ven reducido su peso.

Comprender por qué funciona este procedimiento resulta fácil si se piensa en el paisaje de energía. Imaginemos que el eje horizontal representa todos los datos posibles y el eje vertical representa la energía. Un modelo bien entrenado presenta valles profundos donde se encuentran los datos reales. La reconstrucción no es un punto elegido al azar: es un punto que ha dado un paso desde las proximidades de los datos reales hacia donde el modelo considera actualmente que es más plausible. Hinton comparó esto con un 'error cercano' (near miss). En la corrección de pesos, la parte aditiva disminuye la energía en el lugar donde están los datos reales, y la parte sustractiva eleva la energía en el lugar del error inmediatamente adyacente. Al repetir este proceso, se excavan profundas zanjas a lo largo de los datos en el paisaje de energía. En lugar de examinar todas las regiones lejanas, el modelo se

limita a corregir selectivamente los puntos donde es más propenso a equivocarse.

Hinton también presentó una justificación matemática para este método. En dicha justificación intervienen tres distribuciones: la distribución de los datos reales, la distribución del modelo cuando alcanza el equilibrio y la distribución obtenida tras partir de los datos y someterlos a un número determinado de pasos de muestreo de Gibbs. La diferencia entre dos distribuciones se mide mediante un valor llamado divergencia de Kullback-Leibler (KL), que es 0 si ambas distribuciones son idénticas y aumenta cuanto más difieren. El aprendizaje estándar por máxima verosimilitud, que intenta maximizar la probabilidad de que aparezcan los datos, reduce la divergencia entre la distribución de datos y la distribución de equilibrio. La divergencia contrastiva, en su lugar, reduce la diferencia entre dos divergencias: el valor obtenido al restar la divergencia entre la distribución muestreada y la de equilibrio de la divergencia entre la distribución de datos y la de equilibrio.

A medida que se repite el muestreo, la distribución se acerca al equilibrio, por lo que el valor que se resta no es mayor que el primer valor. Por consiguiente, esta diferencia nunca se vuelve menor que 0, y se convierte en 0 cuando el modelo reconstruye los datos de forma exacta, haciendo que la distribución muestreada coincida con la distribución de los datos. El nombre 'divergencia contrastiva' proviene precisamente de contrastar y comparar estas dos divergencias frente a frente.

CD-1 ofrece también una ventaja estadística. Los datos reales y su reconstrucción están siempre emparejados. Del mismo modo que en los ensayos clínicos comparar pacientes similares por parejas reduce el error atribuible al azar, las comparaciones pareadas reducen la fluctuación (varianza) de los resultados del muestreo. Por esta razón, el aprendizaje progresaba de forma estable incluso con un único paso de muestreo de Gibbs.

La divergencia contrastiva es una aproximación y no un método perfecto. Dado que sustituye el promedio que debería medirse en el equilibrio por una reconstrucción a un solo paso de distancia, la dirección de ajuste puede desviarse de la dirección

exacta que maximiza la probabilidad de los datos. Esta limitación quedó al descubierto en investigaciones posteriores del propio Hinton y sus colegas. Miguel Carreira-Perpiñán y Hinton demostraron en un estudio de 2005 que la CD puede converger a un punto ligeramente desviado de la verdadera solución de máxima verosimilitud, aunque esa desviación suele ser pequeña. Ilya Sutskever y Tijmen Tieleman analizaron en 2010 que la dirección de corrección de la CD no corresponde exactamente al gradiente descendente de una única función objetivo. Esto significa que no es un método que fije un valor específico a minimizar y se mueva en la dirección en que dicho valor desciende más rápidamente. Para compensar esta debilidad, Tieleman propuso en 2008 la divergencia contrastiva persistente (Persistent CD, PCD), que continúa ejecutando la cadena de Markov del muestreo de Gibbs sin reiniciarla desde el principio en cada ocasión. En 2010, Hinton escribió 'Una guía práctica para entrenar RBM', donde recopiló métodos de configuración práctica como la tasa de aprendizaje, el número de unidades ocultas y los valores iniciales de los pesos.

5.4. Variantes de la RBM y extensión al preentrenamiento

La combinación de RBM y CD amplió su campo de aplicación más allá de los patrones formados por ceros y unos a diversos tipos de datos. En ese proceso se destacaron tres cambios principales.

El primero fue la RBM gaussiana-bernoulli (Gaussian-Bernoulli RBM), que modificó las unidades visibles. Para manejar valores reales continuos, como las características de frecuencia de la voz (MFCC) o el brillo de una fotografía, las unidades visibles no pueden limitarse a valores de 0 y 1. Este modelo transformó las unidades visibles en unidades lineales con ruido gaussiano añadido. Los valores de las unidades visibles se extraen de una distribución normal, cuya media es la suma de los pesos provenientes de las unidades ocultas encendidas más su propio sesgo. Este modelo se convirtió posteriormente en el punto de partida para la investigación en reconocimiento de voz.

El segundo cambio consistió en modificar las unidades ocultas. Vinod Nair y Hinton presentaron en la Conferencia Internacional sobre Aprendizaje Automático (ICML) de 2010 una RBM que utilizaba unidades lineales rectificadas (Rectified Linear Unit, ReLU). Ambos concibieron una 'unidad sigmoide escalonada' apilando infinitas unidades binarias con los mismos pesos pero con sesgos ligeramente diferentes. El comportamiento de esta unidad se aproxima mediante una función que produce 0 si la entrada es menor que 0, y un valor proporcional a la entrada si es mayor que 0. A esto le añadieron ruido, utilizándola de tal forma que, si la suma de la entrada y el ruido es mayor que 0, produce dicha suma, y si no, produce 0. Esta ReLU se utilizó en AlexNet en 2012 y posteriormente se convirtió en la función de activación estándar del aprendizaje profundo.

El tercer cambio fue el método de apilar RBM capa por capa. En un artículo publicado en el volumen 18 de Neural Computation en 2006, Hinton, Simon Osindero y Yee-Whye Teh propusieron un método en el que la salida de la capa oculta de una primera RBM servía como datos de entrada para entrenar una segunda RBM. Este método de entrenar capa por capa sucesivamente

con datos sin etiquetar se denomina preentrenamiento no supervisado capa por capa voraz (Greedy Layer-wise Unsupervised Pre-training), y la red neuronal así construida se llama red de creencia profunda (Deep Belief Network, DBN). Ese mismo año, Hinton y Ruslan Salakhutdinov, en un artículo publicado en el volumen 313 de Science, preentrenaron con RBM un autoencoder profundo —una red neuronal que comprime la entrada en un código breve y luego la reconstruye— y lo ajustaron mediante retropropagación del error. Como resultado, obtuvieron mejores resultados que los métodos existentes para reducir datos de alta dimensión a códigos breves. En 2009, ambos ampliaron la teoría hacia la máquina de Boltzmann profunda (Deep Boltzmann Machine, DBM), donde todas las capas están unidas por conexiones simétricas no dirigidas.

Posteriormente, el preentrenamiento con RBM cosechó éxitos en el reconocimiento de voz. Un artículo de 2012, escrito conjuntamente por investigadores de la Universidad de Toronto, Microsoft, Google e IBM, informó que las redes

neuronales profundas inicializadas mediante el enfoque DBN superaron a los métodos tradicionales en modelos acústicos.

Posteriormente, la situación cambió. Con la llegada de ReLU, dropout, datos a gran escala y unidades de procesamiento gráfico (GPU), se hizo posible entrenar redes neuronales profundas desde cero mediante retropropagación del error, sin necesidad de preentrenamiento. Hoy en día, es raro que se utilicen RBM en servicios reales. Un artículo que reseñó la conferencia del Premio Nobel de Hinton el 8 de diciembre de 2024 resumió la disertación como una historia sobre el 'catalizador histórico' que desató la revolución del aprendizaje profundo, mencionando al mismo tiempo que el algoritmo de aprendizaje de la máquina de Boltzmann ya no se utiliza en la actualidad. Al igual que un catalizador que desencadena una reacción química pero no permanece en los productos finales, la RBM abrió las puertas al entrenamiento de redes neuronales profundas y luego se retiró. Durante la semana de los Nobel, Hinton donó al Museo del Premio Nobel un chip temprano de la máquina de Boltzmann del tamaño de un sello postal.

Aunque la RBM en sí ha dejado de utilizarse, las ideas legadas por la RBM y la CD siguen vivas. Se trata de la noción de perfeccionar el modelo reduciendo la energía de los datos reales y aumentando la energía de las falsificaciones verosímiles, en lugar de calcular todos los casos para obtener probabilidades perfectas. Yann LeCun y otros han estructurado este enfoque dentro del marco de los modelos basados en energía (Energy-Based Model). Hay investigadores que consideran que el aprendizaje contrastivo (Contrastive Learning) actual, que entrena enfrentando pares reales contra pares falsos, comparte esta misma inquietud fundamental.

Las respuestas de Hinton al problema de la velocidad de la máquina de Boltzmann fueron tres. Abrió el camino para representar con nitidez datos de alta dimensión mediante el producto de expertos, que multiplica probabilidades en lugar de sumarlas. Hizo posible inferir características ocultas de forma exacta y en un solo paso mediante la RBM, que eliminó las conexiones dentro de la misma capa. Y superó la barrera de la velocidad de aprendizaje mediante la divergencia contrastiva, que compara una reconstrucción de un solo paso

sin esperar a alcanzar el equilibrio. Esta decisión de reducir las conexiones y simplificar los cálculos sirvió de trampolín para que el aprendizaje profundo volviera a captar la atención a partir de 2006.

Capítulo 6. Las redes neuronales profundas vuelven a llamar la atención

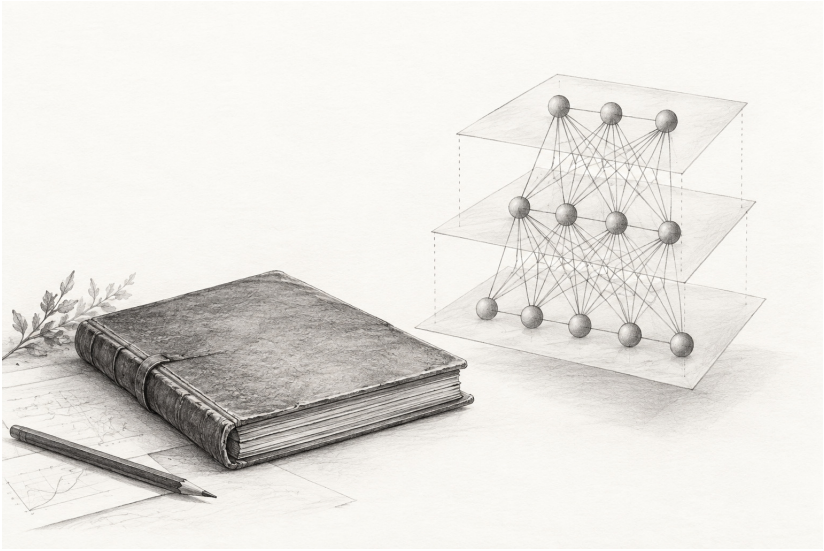


Figura 6. Representación de la idea de investigar redes neuronales compuestas por múltiples capas.

6.1. El segundo invierno de la IA y el problema del desvanecimiento del gradiente

En 1986, cuando David Rumelhart, Geoffrey Hinton y Ronald Williams publicaron el método de retropropagación del error (backpropagation) en Nature, las expectativas en torno a las redes neuronales se dispararon enormemente. La retropropagación del error es un método que calcula la diferencia entre la respuesta emitida por la red neuronal y la respuesta correcta, y luego transmite esa diferencia en sentido inverso, desde la capa de salida hacia la capa de entrada, ajustando poco a poco la fuerza de las conexiones, es decir, los pesos. Este método funcionaba bien en redes neuronales de dos o tres capas. Sin embargo, cuando se aplicó el mismo método a redes neuronales profundas, en las que se apilaban múltiples capas ocultas entre la capa de entrada y la de salida, el aprendizaje no se produjo adecuadamente. En aquel momento, hubo principalmente tres problemas que obstaculizaron el aprendizaje de las redes neuronales profundas.

El primer obstáculo fue el problema del desvanecimiento del gradiente (vanishing gradient). En la retropropagación, la señal de error se multiplica por el valor de la derivada cada vez que atraviesa una capa. La función sigmoide, de uso común en esa época, alcanza como máximo un valor derivado de tan solo 0.25. Al multiplicar reiteradamente un número menor que 1, el valor tiende a 0. Por esta razón, las capas inferiores, más distantes de la capa de salida, apenas recibían señal de aprendizaje. En sentido contrario, también ocurría la explosión del gradiente, donde los valores crecían desmesuradamente en cada capa e inestabilizaban el aprendizaje. De cualquier modo, se trataba del mismo problema en cuanto a que no se transmitía una señal de aprendizaje de magnitud adecuada hacia las capas inferiores.

El segundo factor fue el fracaso de la optimización. La optimización es el proceso de encontrar los pesos con los que se reduce el error. Si se dibuja el modo en que el error fluctúa según los pesos a modo de relieve, el paisaje de error de una red neuronal profunda es accidentado y complejo. Las redes neuronales profundas que partían de pesos fijados al azar

solían quedar atascadas vagando en las inmediaciones de mínimos locales (local minimum) o de puntos de silla planos (saddle point) en dicho relieve, sin lograr dar con pesos adecuados. Un mínimo local es un punto inferior a su entorno inmediato, pero que no representa el punto más bajo a nivel global. Un punto de silla es aquel en el que la pendiente es cercana a 0 debido a que asciende en una dirección y desciende en otra.

El tercer obstáculo fue el sobreajuste (overfitting) originado por la escasez de datos. Los datos etiquetados manualmente por seres humanos con las respuestas correctas eran difíciles de conseguir. Las redes neuronales profundas, dotadas de cientos de miles a millones de pesos, llegaban a memorizar incluso el ruido fortuito de los escasos datos disponibles. Era un fenómeno equivalente al de resolver bien los ejercicios de práctica, pero reprobar ante preguntas completamente nuevas.

Ante este panorama, el interés de la comunidad académica del aprendizaje automático giró hacia otros métodos. Las máquinas de soporte vectorial (SVM) y las técnicas de kernel

lograban convertir el problema de aprendizaje en un problema de optimización convexa (convex optimization). En la optimización convexa, el paisaje de error posee un único valle, de modo que siempre se alcanza la misma solución óptima. Esta constituía una ventaja decisiva frente a las redes neuronales, que corrían el riesgo continuo de caer en mínimos locales. Por su parte, el método de bosques aleatorios (random forest), que combinaba múltiples árboles de decisión, también ofrecía un excelente desempeño. En contraste, las redes neuronales eran consideradas un método de teoría atractiva, pero de difícil manejo en la práctica.

Como consecuencia, desde mediados de la década de 1990 hasta comienzos de la década de 2000, la investigación sobre redes neuronales artificiales experimentó un largo periodo de declive. En la historia de la inteligencia artificial, la fase de estancamiento que sobrevino a finales de la década de 1980 tras el colapso de las expectativas puestas en los sistemas expertos se conoce como el segundo invierno de la IA. El invierno que padecieron los investigadores de redes neuronales fue aún más prolongado. Hasta mediados de la década de 2000,

el ámbito académico dominante no consideraba a las redes neuronales como un objeto de estudio riguroso. Hinton y sus colegas rememoraron en entrevistas posteriores que bastaba con incluir el término 'redes neuronales' en el título de un artículo para que fuera muy difícil superar la revisión académica en los congresos.

La institución que brindó el respaldo necesario para que la investigación sobre redes neuronales continuara en dicho periodo fue el Instituto Canadiense de Investigaciones Avanzadas (CIFAR). En 2004, Hinton, adscrito a la Universidad de Toronto, propuso un nuevo programa de investigación junto con Yoshua Bengio, de la Universidad de Montreal, Yann LeCun, de la Universidad de Nueva York, y David Fleet, de la Universidad de Toronto, entre otros. El CIFAR aceptó la propuesta e instituyó el programa 'Computación Neuronal y Percepción Adaptativa' (NCAP, Neural Computation and Adaptive Perception), el cual Hinton encabezó durante cerca de 10 años. Dicho programa adoptó más adelante el nombre de 'Aprendizaje en Máquinas y Cerebros' (Learning in Machines & Brains).

El NCAP otorgó financiamiento para periodos de 3 a 5 años sustentándose únicamente en propuestas de investigación breves de unas pocas páginas. Asimismo, no exigía la obtención inmediata de resultados comerciales. Este grupo reducido de investigadores articulado en torno a Hinton, Bengio y LeCun se reunía de manera periódica para contrastar ideas y organizaba escuelas de verano destinadas a capacitar a investigadores jóvenes. Las investigaciones de 2006 que hallaron la manera de entrenar redes neuronales profundas surgieron precisamente en el seno de esta comunidad científica.

6.2. Aprender capa por capa: preentrenamiento no supervisado voraz capa por capa y DBN

En 2006, Hinton publicó junto a Simon Osindero, de la Universidad de Toronto, y Yee-Whye Teh, de la Universidad Nacional de Singapur, el artículo «A Fast Learning Algorithm for Deep Belief Nets» (Un algoritmo de aprendizaje rápido para redes de creencia profunda) en el volumen 18, número 7 de la revista Neural Computation. El trabajo se extendía de la página 1527 a la 1554.

La problemática que abordó este artículo fue el fenómeno del descarte de explicaciones (explaining away). Una red neuronal que asume que las causas de la capa superior determinan las consecuencias en la capa inferior recibe el nombre de red de creencia dirigida. En este tipo de redes resulta sumamente difícil deducir las causas hacia atrás a partir de los efectos observados. Tomemos como ejemplo el césped mojado, cuyas causas podrían ser la lluvia o un aspersor. Si, al constatar que el césped se encuentra mojado, nos enteramos de que llovió, la

probabilidad de que el aspersor se haya encendido disminuye. Dos causas que en un principio eran independientes pasan a vincularse mutuamente en cuanto se conoce el resultado. En redes neuronales provistas de múltiples capas ocultas y miles de unidades, esta interdependencia se volvía de tal complejidad que el cálculo exacto de la inferencia resultaba prácticamente imposible.

La herramienta que empleó el equipo de Hinton para resolver este problema fue la máquina de Boltzmann restringida (RBM, Restricted Boltzmann Machine). La máquina de Boltzmann es una red neuronal probabilística ideada por Hinton en la década de 1980 junto a Terry Sejnowski, cuyo nombre se inspira en las fórmulas de física estadística del físico decimonónico Ludwig Boltzmann. Una red neuronal probabilística es aquella donde cada unidad no entrega un valor determinado, sino que se activa o desactiva con arreglo a una probabilidad. La RBM constituye una simplificación de este modelo a través de la reducción de conexiones. Consta de dos niveles: una capa visible que recibe los datos y una capa oculta que alberga las características de los mismos; las unidades pertenecientes a

una misma capa no se comunican entre sí, manteniéndose enlaces únicamente entre ambas capas. Gracias a esta limitación, en cuanto se definen los valores de una capa, las unidades de la otra capa pueden calcularse de forma simultánea e independiente, simplificando sobremanera el cómputo. En 2002, Hinton había formulado el método de divergencia contrastiva (CD, contrastive divergence), que permitía entrenar las RBM a gran velocidad.

Los investigadores sortearon matemáticamente el dilema del descarte de explicaciones. En una red de creencia dirigida de infinitas capas, si los pesos de cada capa se acoplan de manera que sean la transpuesta (transpose) de los pesos de la capa superior inmediata, las probabilidades a priori generadas por la capa superior contrarrestan con exactitud las correlaciones espurias originadas en la capa inferior. El equipo denominó a esto 'distribuciones a priori complementarias' (complementary priors). Además, demostraron que este modelo dirigido e infinitamente profundo con pesos acoplados equivalía matemáticamente a una sola RBM de dos capas. De este modo, un modelo dirigido profundo, cuya inferencia resultaba

intratable, pudo abordarse convirtiéndolo en una RBM para la cual ya se contaba con una técnica de entrenamiento. De esta correspondencia nació el preentrenamiento no supervisado voraz capa por capa (greedy layer-wise unsupervised pre-training).

El procedimiento de aprendizaje se desglosa en los siguientes pasos. En la primera etapa, los datos de entrada sin etiquetar se introducen en la capa visible y se entrenan los pesos de la RBM inferior por divergencia contrastiva. En la segunda etapa, se fijan dichos pesos y se pasan los datos de entrada a fin de obtener la probabilidad de activación de las unidades ocultas. Dicha actividad de la capa oculta se convierte en el nuevo conjunto de datos de entrada para la capa subsiguiente. En la tercera etapa, se entrena por separado una segunda RBM a partir de estos nuevos datos. La operación se repite una capa tras otra hasta alcanzar la profundidad deseada. Recibió el calificativo de 'voraz' debido a que consolida el resultado de la fase anterior y optimiza de forma exclusiva la capa en curso. Se denomina 'no supervisado' porque prescinde de respuestas correctas para nutrirse solo de los datos, y 'preentrenamiento'

porque establece los pesos previamente a la fase definitiva de instrucción.

En la cuarta etapa, las RBM apiladas de esta forma se ensamblan en un modelo generativo integral. Este dispositivo constituye la red de creencia profunda (DBN, Deep Belief Network). Las dos capas superiores de la DBN conservan sus conexiones simétricas no dirigidas y operan a modo de memoria asociativa. Las capas subyacentes se transforman en conexiones dirigidas que apuntan de arriba hacia abajo, asumiendo la función de generar datos.

El equipo respaldó asimismo este procedimiento con fundamentos matemáticos. Demostraron que cada vez que se congela el peso de la capa inferior y se entrena una RBM adicional superpuesta, el límite inferior de la verosimilitud con la que el modelo explica los datos no decrece, sino que puede mejorar. Esta conclusión aportó por primera vez una prueba matemática rigurosa de que la capacidad del modelo se perfecciona al añadir más capas.

Los pesos de la red neuronal tras el preentrenamiento no son valores arbitrarios, sino que parten de una posición favorable que ya recoge en cierta medida la estructura intrínseca de los datos. Tomando en cuenta que el fracaso de optimización expuesto en la sección 6.1 nacía de arrancar desde pesos aleatorios, la modificación del punto de partida constituyó la esencia de esta propuesta. Sobre esta base, se aplica un ajuste fino (fine-tuning) que pule la totalidad de los pesos empleando un pequeño volumen de datos etiquetados. En aquel estudio, dicho ajuste fino se ejecutó mediante el algoritmo 'arriba-abajo' (up-down), una variante del algoritmo de vigilia-sueño (wake-sleep).

El equipo validó su rendimiento mediante el conjunto de datos de dígitos manuscritos MNIST. MNIST se compone de imágenes numéricas de 28×28 píxeles distribuidas en 60.000 muestras para entrenamiento y 10.000 para evaluación. Una DBN configurada con capas ocultas de 500, 500 y 2.000 unidades arrojó una tasa de error del 1.25% valiéndose exclusivamente del valor de los píxeles, sin recurrir a información morfológica ni a mecanismos de aumento de datos. Fue una marca inferior al

1.4% obtenido por una máquina de soporte vectorial y a cerca del 1.5% alcanzado por redes neuronales entrenadas por retropropagación bajo idénticos parámetros. Esta DBN no solo clasificaba cifras con éxito, sino que el modelo aprendido era capaz de dibujar nuevas imágenes de dígitos. Bajo condiciones comparativas que prescindían deliberadamente de información espacial, se demostró que una red neuronal profunda podía superar a los métodos consagrados más solventes.

A esta conclusión no llegó únicamente el grupo de Hinton. En el mismo año, el equipo de Bengio expuso en el congreso NIPS que el preentrenamiento capa por capa facilitaba el aprendizaje de redes neuronales profundas no solo en las RBM, sino también en otras configuraciones. El laboratorio de LeCun también aportó investigaciones en una línea análoga. Tres grupos científicos adscritos a la comunidad del CIFAR alcanzaron la misma certidumbre en fechas prácticamente simultáneas.

6.3. El artículo de 2006 en Science: autocodificadores profundos y reducción no lineal de la dimensionalidad

El 28 de julio de 2006, Hinton y su estudiante de doctorado Ruslan Salakhutdinov publicaron en el volumen 313, número 5786 de Science el artículo «Reducing the Dimensionality of Data with Neural Networks» (Reducción de la dimensionalidad de los datos mediante redes neuronales). El texto ocupaba cuatro páginas, de la 504 a la 507. Si el estudio sobre DBN aparecido ese mismo mes se había difundido a los especialistas por medio de una revista técnica, esta publicación anunció ante el conjunto de la comunidad científica el renacer indiscutible de las redes neuronales profundas.

El núcleo del artículo fue el problema de la reducción de la dimensionalidad. La reducción de la dimensionalidad consiste en compendiar información con gran profusión de cifras en un conjunto numérico más acotado. Si una imagen digital de 784 píxeles puede comprimirse a 30 números y aun así restituir la

figura original, cabe deducir que esos 30 valores condensan la esencia de la imagen. Durante décadas se empleó habitualmente para este fin el análisis de componentes principales (PCA), sistematizado en 1933 por Harold Hotelling. El PCA localiza las líneas rectas de mayor dispersión de los datos y los proyecta sobre dichas coordenadas.

El inconveniente estriba en que los datos del mundo real no descansan sobre un plano uniforme. Manifestaciones como imágenes o textos se sitúan sobre superficies alabeadas análogas a hojas enrolladas, es decir, sobre variedades no lineales (nonlinear manifold). Al alumbrar verticalmente un papel enrollado, sectores distantes entre sí quedan superpuestos. Del mismo modo, el PCA, al proyectar sobre ejes rectilíneos, aplasta y distorsiona las estructuras curvadas.

El autocodificador (autoencoder) es la red neuronal concebida para superar esta barrera. Se estructura mediante un codificador que comprime la entrada hacia capas sucesivamente menores y un decodificador que regenera la escala primitiva a partir de la estrecha capa intermedia de

código. Al entrenarlo para que la salida reproduzca fielmente la entrada, la capa intermedia retiene el núcleo medular de los datos. Aunque este concepto se remontaba a la década de 1980, su materialización práctica topaba invariablemente con los escollos descritos en la sección 6.1. Al instruir un autocodificador profundo con retropropagación a partir de pesos aleatorios, las señales no alcanzaban con nitidez la capa central de código, dando como resultado un rendimiento idéntico o incluso peor que el del PCA.

Hinton y Salakhutdinov zanjaron esta dificultad gracias al preentrenamiento escalonado con RBM mediante un proceso en tres fases. En primer término, se entrenan sucesivamente varias RBM desde los estratos inferiores. En el experimento con MNIST, se empleó un esquema que reducía 784 entradas a 30 pasando sucesivamente por 1.000, 500 y 250. En segundo término, se apilan los pesos de las RBM aprendidas para estructurar el codificador y se superponen de manera inversa los mismos pesos transpuestos para erigir el decodificador; de este modo se procede a 'desplegar' (unrolling) el autocodificador. En tercer término, las unidades que arrojaban

de modo probabilístico valores de 0 o 1 se reemplazan por unidades sigmoides de respuesta continua, y se ejecuta un ajuste fino de toda la red mediante retropropagación con el fin de minimizar la discrepancia entre la entrada original y la salida restaurada. Tras dotar al sistema de un buen punto de partida mediante el preentrenamiento, la retropropagación cumplió con cabal eficacia su cometido incluso en redes neuronales de notable profundidad.

La contundencia de los resultados quedaba a la vista en las comparaciones gráficas. Al comprimir imágenes de dígitos a 30 números y restituirlas después, las recreaciones logradas por el autocodificador profundo resultaban prácticamente indistinguibles del original. En cambio, las imágenes reconstruidas mediante PCA bajo la misma tasa de compresión mostraban contornos desvaídos y borrosos. Al distribuir las imágenes numéricas sobre un plano bidimensional, la divergencia fue igualmente palmaria: mientras que en el gráfico de PCA los números del 0 al 9 figuraban entremezclados y confusos, en el gráfico del autocodificador cada cifra se agrupaba en conglomerados perfectamente delimitados.

Asimismo, en retratos faciales humanos, el autocodificador obtuvo márgenes de error de reconstrucción significativamente menores que el PCA.

Los autores trasladaron igualmente este mecanismo al terreno textual. Tras procesar más de 800.000 artículos de prensa representados mediante la frecuencia de 2.000 vocablos comunes y proceder a su compresión, las crónicas de temática afín se congregaron. El logro superó al análisis semántico latente (LSA), método lineal predominante hasta entonces. Este hito derivó en 2009 en el artículo sobre 'hashing semántico' (semantic hashing) publicado por Salakhutdinov e Hinton. Al condensar textos en códigos binarios breves, es factible localizar documentos de sentido convergente a una velocidad extraordinaria mediante la mera coincidencia de sus códigos.

En las conclusiones del artículo, ambos científicos precisaron las circunstancias que habían hecho viable este avance. La noción de materializar la reducción no lineal de la dimensionalidad mediante autocodificadores profundos existía desde la década de 1980; no obstante, su plasmación real exigía

computadoras veloces, conjuntos masivos de datos y una inicialización óptima de los pesos. En 2006 confluyeron estas tres premisas, y la vía para obtener aquellos pesos iniciales idóneos fue justamente el preentrenamiento por capas.

6.4. Máquinas de Boltzmann profundas (DBM) y la difusión del aprendizaje profundo

En 2009, Salakhutdinov e Hinton presentaron la máquina de Boltzmann profunda (DBM, Deep Boltzmann Machine) en la 12.ª Conferencia Internacional sobre Inteligencia Artificial y Estadística (AISTATS), celebrada en Florida, Estados Unidos. La DBN constituía un modelo mixto en el que únicamente las dos capas superiores se hallaban unidas de modo no dirigido, mientras que los niveles inferiores se articulaban exclusivamente de arriba hacia abajo. En contraste, la DBM enlazaba la capa visible y la totalidad de las capas ocultas mediante conexiones simétricas y no dirigidas.

Esta diferencia repercute en el esquema de inferencia. La capa oculta central de la DBM define su estado recibiendo a la vez la información sensorial que asciende desde la base y los datos contextuales que descienden desde la cúspide. Es algo parecido a lo que hace una persona al leer un texto de caligrafía borrosa, deduciendo las letras dudosas a partir del contexto que las antecede y las sucede. En virtud de ello, la DBM procesó con solvencia estímulos provistos de ruido o parcialmente ocultos.

En contrapartida, el entrenamiento de la DBM presentaba mayor dificultad. Los investigadores aplicaron el preentrenamiento capa por capa con RBM, pero ajustaron los pesos duplicándolos o reduciéndolos a la mitad para adecuarse a una arquitectura que capta impulsos tanto superiores como inferiores. A renglón seguido, realizaron la inferencia recurriendo a la aproximación de campo medio (mean-field) y afinaron los pesos globales empleando un método de muestreo denominado divergencia contrastiva persistente (persistent CD). La DBM instruida bajo esta pauta alcanzó una tasa de error de tan solo el 0.95% en MNIST, y del 10.8% en la base de datos NORB, compuesta por tomas de juguetes tridimensionales

captados desde variados ángulos, superando el 11.6% exhibido por las máquinas de soporte vectorial tomadas como referencia.

Hacia esa época, los réditos del preentrenamiento trascendieron los confines de los laboratorios. El primer terreno conquistado fue el reconocimiento del habla. Abdel-rahman Mohamed y George Dahl, investigadores de la Universidad de Toronto, en colaboración con Hinton, incorporaron redes neuronales profundas preentrenadas mediante DBN a los modelos acústicos responsables de discriminar las unidades fonéticas. Durante decenios, el sector del reconocimiento de voz había mantenido como canon la combinación de modelos de mezclas gaussianas (GMM) y modelos ocultos de Márkov (HMM). La red neuronal profunda desbancó a este esquema tradicional en la prueba de reconocimiento de fonemas TIMIT y, poco después, brindó notables marcas en el reconocimiento vocal con vocabularios extensos. En 2012, especialistas de la Universidad de Toronto, Microsoft, Google e IBM publicaron un informe conjunto en IEEE Signal Processing Magazine ratificando que los cuatro centros habían arribado a la misma

certidumbre. Desde aquel instante, las principales corporaciones tecnológicas transformaron el corazón de sus motores de voz dotándolos de redes neuronales profundas.

En ese mismo año 2012, Alex Krizhevsky e Ilya Sutskever, discípulos de Hinton, se alzaron con un triunfo aplastante en el certamen ImageNet gracias a AlexNet, una red neuronal concebida en equipo con Hinton, registrando una tasa de error top-5 de apenas el 15.3% y aventajando con creces al segundo puesto. El campo de la visión por computadora aceleró de inmediato su reconversión hacia las redes neuronales profundas.

Sin embargo, en el instante preciso en que las redes neuronales profundas alcanzaban difusión masiva, el preentrenamiento por capas comenzó a ceder su primacía. Ello se debió al hallazgo de soluciones alternativas para solventar las tres barreras señaladas en la sección 6.1. Las unidades de procesamiento gráfico (GPU) aceleraron sustancialmente la velocidad de procesamiento y comenzaron a proliferar bancos de millones de datos rotulados. Asimismo, surgieron la función

de activación ReLU, que mitiga el desvanecimiento del gradiente, el método de dropout, que desconecta aleatoriamente ciertas unidades durante la fase instructiva, e innovadoras formulaciones para establecer los pesos iniciales. Gracias a ello, fue factible entrenar redes neuronales profundas directamente mediante retropropagación desde cero, prescindiendo del preentrenamiento. De hecho, la propia AlexNet fue una red entrenada sin recurrir a ningún preentrenamiento.

Pese a ello, las conquistas de 2006 no perdieron en absoluto su valor formativo. En 2015, LeCun, Bengio e Hinton recapitulaban esta etapa en el artículo de balance «Deep Learning», divulgado en el volumen 521 de Nature. Los tres especialistas hicieron constar que, hacia 2006, la red de investigadores cohesionada por el CIFAR había revitalizado el interés por las redes neuronales profundas. Subrayaron que la estrategia de aprehender rasgos estrato a estrato con datos desprovistos de etiquetas para forjar pesos iniciales certeros y afinarlos ulteriormente por retropropagación demostró de modo indiscutible la plena viabilidad de las redes profundas, aun bajo

condiciones de escasez de datos supervisados. El auténtico mérito de aquel trabajo consistió en demoler el arraigado prejuicio que sostenía que las redes neuronales profundas eran inentrenables.

El principio rector de ejercitar primero los sistemas con inmensos caudales de datos no supervisados para luego afinarlos con escasos datos de referencia sobrevive plenamente en los modelos masivos de lenguaje actuales. Los modelos contemporáneos han dejado de apilar RBM; no obstante, el marco rector de preentrenar inicialmente sobre corpus enciclopédicos de frases para luego proceder al ajuste fino conforme a tareas determinadas es heredero directo de las convicciones forjadas en 2006.

Hinton, Bengio y LeCun fueron distinguidos conjuntamente con el Premio Turing en 2018, galardón reputado como el Nobel de las ciencias de la computación. Posteriormente, el 8 de octubre de 2024, la Real Academia de las Ciencias de Suecia comunicó el otorgamiento del Premio Nobel de Física a Geoffrey Hinton y John Hopfield por sus descubrimientos e invenciones

fundacionales que hicieron posible el aprendizaje automático mediante redes neuronales artificiales. La comisión dictaminó que Hinton, partiendo de las redes de Hopfield, concibió la máquina de Boltzmann y que, al profundizar en dicho empeño, impulsó decisivamente la eclosión vertiginosa del aprendizaje automático contemporáneo. Las DBN de 2006 procedían en línea recta de aquella fecunda máquina de Boltzmann. Las pesquisas consagradas a la máquina de Boltzmann, mantenidas vivas en el seno del prolongado letargo de las redes neuronales, desembocaron en el método que posibilitó el aprendizaje efectivo de las redes neuronales profundas.

Parte 3. El aprendizaje profundo sale al mundo

Capítulo 7. Computadoras que reconocen la voz y las imágenes

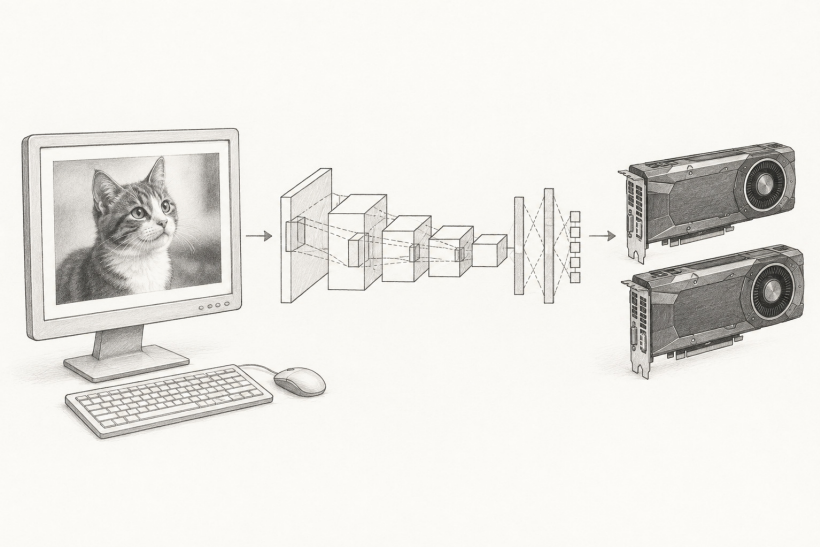


Figura 7. Representación del reconocimiento de imágenes y el papel de las GPU.

En 2006, Geoffrey Hinton abrió el camino para entrenar redes neuronales profundas mediante las redes de creencia profunda (DBN). Sin embargo, la mayor parte de la comunidad académica de la inteligencia artificial seguía sin confiar en las redes neuronales. Los buenos resultados publicados en unos pocos artículos no bastaban para cambiar una idea preconcebida largamente arraigada. Lo que transformó el juicio del mundo académico fueron las cifras obtenidas en problemas reales. Entre 2009 y 2012, el laboratorio de Hinton presentó tales cifras en dos ocasiones: la primera provino del reconocimiento del habla, que comprende la voz humana, y la segunda de la visión por computador, que reconoce objetos en fotografías.

7.1. El primer gran avance: la transformación del reconocimiento del habla (2009-2012)

La tecnología que permite a una computadora convertir la voz humana en texto se denomina reconocimiento automático del habla. Esta tecnología se divide a grandes rasgos en dos partes: por un lado, el modelo acústico, que determina qué pronunciación corresponde a un fragmento de sonido muy breve; por el otro, la parte que une esas pronunciaciones identificadas para formar palabras y oraciones. De estas dos, el modelo acústico se había mantenido estancado en el mismo enfoque durante casi 30 años, hasta finales de la década de 2000. Dicho método consistía en calcular probabilísticamente las características del sonido mediante modelos de mezcla de gaussianas (GMM) y seguir su evolución temporal a través de modelos ocultos de Márkov (HMM).

Aunque este método ofreció buenos resultados durante mucho tiempo, también presentaba claras debilidades. La estructura del GMM asume que un fragmento de sonido es explicado por

uno solo de entre varios 'expertos'. No obstante, el habla humana no se divide de forma tan nítida. Las pronunciaciones anteriores y posteriores se mezclan entre sí, e incluso en un único instante se superponen las propiedades de múltiples sonidos. Para representar esta superposición, el GMM debía incrementar constantemente sus componentes, lo que demandaba una gran cantidad de cálculos y datos. Además, resultaba difícil examinar de manera simultánea los sonidos a lo largo de varios momentos consecutivos.

Los primeros en superar esta limitación fueron George Dahl y Abdel-rahman Mohamed, estudiantes de doctorado en el laboratorio de Hinton en la Universidad de Toronto. En 2009, ambos aplicaron al modelo acústico el método DBN presentado por Hinton en 2006. Primero apilaron máquinas de Boltzmann restringidas (RBM) capa por capa para que aprendieran la estructura del sonido sin necesidad de respuestas correctas, y luego afinaron toda la red neuronal mediante retropropagación de errores. Como entrada, en lugar de introducir únicamente un fragmento de sonido, introdujeron un paquete de entre 11 y 15 fragmentos consecutivos. La red neuronal pudo asimilar de

manera directa el contexto anterior y posterior que resultaba tan difícil de manejar para los GMM.

Lo que hizo posible este experimento fueron las tarjetas gráficas, es decir, las GPU. La GPU es un dispositivo diseñado originalmente para renderizar gráficos de videojuegos, pero destaca por procesar simultáneamente una gran cantidad de cálculos del mismo tipo. Utilizando GPU Tesla de NVIDIA y CUDAMat, una herramienta de cálculo desarrollada por su compañero de laboratorio Volodymyr Mnih, ambos aumentaron la velocidad de entrenamiento en más de 20 veces en comparación con el uso de una sola CPU. Cuando los experimentos que antes tardaban semanas empezaron a completarse en unas pocas horas, fue posible realizar pruebas cambiando la configuración en reiteradas ocasiones.

Los resultados llamaron la atención de la comunidad del reconocimiento del habla. En TIMIT, el conjunto de datos de prueba estándar utilizado durante años por los investigadores, su red neuronal registró una tasa de error de fonemas del 20.7%. Dado que el fonema es la unidad de sonido más

pequeña que compone el habla, esto significaba que erraba en torno a 21 de cada 100 fonemas. Este resultado superó el récord histórico de los métodos convencionales reportado hasta ese momento. Los primeros resultados se presentaron en un taller de NIPS en 2009, y los resultados perfeccionados se publicaron en una revista de la IEEE en 2012.

Sin embargo, TIMIT era un conjunto de datos de prueba a pequeña escala. Para comprobar si este método funcionaba también en productos reales, se necesitaban los datos y los sistemas de las grandes empresas. Por ello, los estudiantes de Hinton se dirigieron directamente a las corporaciones. George Dahl realizó una pasantía en Microsoft Research (MSR) y Abdel-rahman Mohamed en IBM Research. Otro de sus discípulos, Navdeep Jaitly, se marchó a Google.

En Microsoft Research, Li Deng y Dong Yu colaboraron con el equipo de investigación de Hinton. Aplicaron redes neuronales profundas a datos de búsqueda por voz móvil de Bing, grabados con teléfonos móviles y con mucho ruido, y la tasa de error a nivel de frase se redujo del 36.2% al 30.4%. A

continuación, en Switchboard, un corpus de 300 horas de conversaciones telefónicas, la tasa de error de palabras cayó del 27.4% al 18.5%. Esto significó reducir el error en aproximadamente un tercio. Dicho resultado igualó el 18.6% del sistema convencional, el cual requería más de 2.000 horas de datos y complejas técnicas de calibración. Con muchos menos datos y solo con una red neuronal, alcanzaron el mismo nivel.

En Google, Jaitly experimentó con 5.870 horas de datos de entrada de voz junto con Vincent Vanhoucke, Patrick Nguyen y otros. La tasa de error de palabras disminuyó del 16.0% al 12.3%. Google perfeccionó esta tecnología y la incorporó en 2012 a la búsqueda por voz de los teléfonos inteligentes Android. Una red neuronal surgida en un laboratorio universitario se integraba así en un producto que muchas personas utilizaban a diario.

Estos resultados, confirmados de manera independiente en varias empresas, se recopilaron en un solo artículo. En noviembre de 2012, cuatro grupos de investigación de la

Universidad de Toronto, Microsoft, Google e IBM publicaron un artículo conjunto titulado «Redes neuronales profundas para el modelado acústico en el reconocimiento del habla» en IEEE Signal Processing Magazine. Resultaba insólito que investigadores de empresas rivales compartieran la autoría de un mismo trabajo. Este artículo anunció a la comunidad científica que el estándar del reconocimiento del habla había cambiado, constituyendo uno de los principales logros del aprendizaje profundo en el ámbito industrial. No obstante, esta transformación ocurrió principalmente entre los especialistas en reconocimiento del habla. Se necesitaría un poco más de tiempo para que la totalidad de la comunidad académica de la inteligencia artificial lo asimilara plenamente.

7.2. El punto de inflexión de la visión por computador: ImageNet de Fei-Fei Li y AlexNet (2012)

A finales de la década de 2000, la comunidad académica de visión por computador seguía el método de diseñar manualmente características útiles a partir de fotografías. SIFT y HOG eran ejemplos representativos. Se extraían características calculando la dirección de los bordes o las variaciones de brillo en la imagen, y se determinaba la clase del objeto mediante clasificadores como las máquinas de vectores de soporte (SVM). El rendimiento de este método se veía fácilmente afectado cuando cambiaba la iluminación, el objeto se fotografiaba desde un ángulo distinto o se encontraba parcialmente oculto. Aunque el rendimiento mejoraba cada año, el margen de mejora era reducido.

Por aquel entonces, quien preparó el terreno para cambiar la situación fue Fei-Fei Li. Li inició el proyecto ImageNet en la Universidad de Princeton en 2007 y continuó la investigación tras trasladarse a la Universidad de Stanford en 2009. El equipo

de Li etiquetó manualmente imágenes recopiladas de internet. La mano de obra necesaria se obtuvo mediante Amazon Mechanical Turk, un servicio de intermediación de tareas en línea. De este modo, se construyó una base de datos que contenía más de 15 millones de fotografías distribuidas en más de 22.000 categorías. Su escala era completamente distinta a la de los recursos fotográficos que los investigadores habían utilizado hasta entonces.

A partir de 2010, se seleccionaron de entre ellas 1.000 categorías y alrededor de 1,2 millones de fotografías de entrenamiento para organizar anualmente una competencia llamada ILSVRC. El rendimiento se evaluaba principalmente mediante la 'tasa de error top-5'. Se trataba de un método en el que la computadora proponía cinco candidatos de respuesta para cada imagen y, si la respuesta correcta no figuraba entre ellos, se consideraba un error. La tasa de error del equipo ganador en 2010 fue del 28.2%, y la del equipo ganador en 2011 fue del 25.8%. El método que utilizaba características diseñadas manualmente por humanos no lograba avanzar más allá de esos valores.

En 2012, Hinton participó en esta competencia junto con dos de sus discípulos, Alex Krizhevsky e Ilya Sutskever. El nombre del equipo era SuperVision. Las funciones de los tres eran distintas: Krizhevsky fue quien escribió el código y entrenó la red neuronal, mientras que Sutskever fue quien insistió con firmeza en aplicar redes neuronales a grandes volúmenes de datos fotográficos. Hinton, como profesor supervisor, marcó la dirección. Krizhevsky realizó este trabajo en una computadora equipada con dos tarjetas gráficas NVIDIA, y a la red neuronal se le dio más tarde el nombre de AlexNet, en su honor.

Las raíces de AlexNet se remontan a las redes neuronales convolucionales (CNN) que Yann LeCun creó a finales de la década de 1980 para leer dígitos escritos a mano. Una red neuronal convolucional escanea sucesivamente pequeñas regiones de una imagen en busca de patrones y reconoce formas cada vez más complejas a medida que atraviesa las capas. AlexNet amplió esta estructura para adaptarla a fotografías en color grandes y complejas. Estaba compuesta por un total de 8 capas: 5 capas convolucionales y 3 capas completamente conectadas donde todas las neuronas estaban

unidas entre sí. Contaba con aproximadamente 60 millones de parámetros —valores que se ajustan mediante el aprendizaje— y alrededor de 650.000 neuronas. Para la época, se trataba de un tamaño insólito.

Sin embargo, no solo aumentaron el tamaño. Para entrenar adecuadamente una red neuronal de tal magnitud, el equipo recurrió a cuatro métodos combinados. El primero fue el uso de ReLU como función de activación. La función de activación es la regla que determina cuánta señal emite una neurona. En las funciones más empleadas hasta entonces, a medida que la entrada aumentaba, el gradiente se aproximaba a 0, lo que ralentizaba el aprendizaje. La ReLU es una función sencilla que convierte en 0 los valores menores que 0 y deja pasar tal cual los valores mayores que 0. En los experimentos, la red neuronal que utilizó ReLU aprendió 6 veces más rápido que la que empleaba las funciones convencionales.

El segundo fue la distribución de la red neuronal entre dos GPU. La NVIDIA GTX 580 utilizada por Krizhevsky solo disponía de 3 GB de memoria, por lo que no podía albergar toda la red

neuronal en una sola tarjeta. Él mismo programó cuda-convnet, un código de cálculo para GPU, dividió la red neuronal por la mitad entre las dos GPU e hizo que ambas intercambiaran información únicamente en ciertas capas. Se necesitaron entre 5 y 6 días para entrenar la red repitiendo el proceso cerca de 90 veces sobre las 1,2 millones de imágenes.

El tercero fue el dropout. Con nada menos que 60 millones de parámetros, la red neuronal tiende a memorizar por completo las imágenes de entrenamiento, lo que facilita que su rendimiento decaiga ante imágenes nunca antes vistas. Esto se conoce como sobreajuste. El dropout es un método que desactiva aleatoriamente la mitad de las neuronas en cada ciclo de entrenamiento. Esto genera el efecto de entrenar una red neuronal ligeramente distinta cada vez, evitando que la red dependa en exceso de neuronas específicas. AlexNet aplicó este método en sus dos últimas capas.

El cuarto fue el aumento de los datos de entrenamiento. A partir de imágenes de tamaño 256×256, recortaron aleatoriamente secciones de 224×224, las invirtieron horizontalmente y

alteraron ligeramente los tonos de color. De este modo, generaron múltiples datos de entrenamiento nuevos a partir de una sola fotografía.

Lo que la red neuronal aprendió de las fotografías tras este entrenamiento puede comprobarse en la primera capa. Al visualizar los 96 filtros que la primera capa de AlexNet aprendió por sí misma, se aprecian líneas en diversas orientaciones, bordes y manchas de color. Nadie le enseñó a buscar tales patrones. La red neuronal los formuló por sí misma al observar las imágenes. Estos patrones se asemejaban a aquellos ante los cuales reaccionan las células de la corteza visual primaria del cerebro humano.

Los resultados de la competencia se anunciaron en octubre de 2012 en un taller de la Conferencia Europea de Visión por Computador (ECCV) celebrado en Florencia, Italia. La tasa de error top-5 de AlexNet fue del 15.3%. El equipo de la Universidad de Tokio, que ocupó el segundo lugar, registró un 26.2% empleando métodos tradicionales como SIFT y vectores de Fisher. La diferencia entre ambos equipos fue de 10.9 puntos

porcentuales. En una competencia donde las posiciones solían decidirse por diferencias de 1 a 2 puntos porcentuales, AlexNet redujo el error en más del 40% en comparación con el segundo lugar.

7.3. La transición al aprendizaje profundo: fundación de DNNresearch y adquisición por parte de Google

Las reacciones en el auditorio de Florencia fueron encontradas. Los académicos que llevaban mucho tiempo investigando la visión por computador no estaban convencidos de la utilidad de este método. En cambio, Yann LeCun, que se encontraba en el lugar, lo valoró como un punto de inflexión para la inteligencia artificial. LeCun había trabajado como investigador posdoctoral de Hinton en Toronto a finales de la década de 1980.

La evolución posterior se dio tal como LeCun había previsto. Antes de AlexNet, casi ninguno de los principales artículos sobre visión por computador utilizaba redes neuronales, y era

habitual que los artículos basados en redes neuronales fueran rechazados en la revisión por pares. Sin embargo, en el lapso de uno a dos años tras AlexNet, los investigadores abandonaron el diseño manual de características y migraron hacia el aprendizaje profundo. A partir de la competencia del año siguiente, la gran mayoría de los equipos líderes empleó redes neuronales convolucionales. Esta escena, en la que la fuerza acumulada por el conexionismo tras haber permanecido en la periferia durante más de 30 años estalló de golpe, se compara a menudo con la rotura de una presa contenida. Fei-Fei Li comentaría más tarde que muchas personas consideran este momento como el nacimiento de la inteligencia artificial moderna.

Las empresas también se pusieron en movimiento. Las grandes corporaciones tecnológicas, que ya habían confirmado el rendimiento de las redes neuronales en el reconocimiento del habla, vieron AlexNet e intentaron reclutar a Hinton y a sus dos discípulos. Hinton decidió fundar una empresa para luego venderla. En diversas entrevistas posteriores, reveló que necesitaba una gran suma de dinero para asegurar el porvenir

de su hijo, quien padecía una dificultad de aprendizaje. A finales de 2012, Hinton creó junto a Krizhevsky y Sutskever una compañía llamada DNNresearch. El nombre procedía de las redes neuronales profundas (Deep Neural Network). No disponía de productos ni de ingresos, y solo contaba con tres empleados. Todo lo que la empresa poseía era la tecnología con la que habían creado AlexNet y el talento de los tres investigadores.

A principios de diciembre de 2012, se celebró una subasta por esta empresa en Lake Tahoe, Estados Unidos, donde se llevaba a cabo la conferencia NIPS. En la puja participaron Baidu, Google, Microsoft y DeepMind, una startup fundada apenas dos años antes. La subasta se llevó a cabo por correo electrónico; Hinton la condujo desde su habitación de hotel, manteniendo en secreto la identidad de los demás participantes. El precio inicial fue de 12 millones de dólares, y se fijó que cada incremento de puja debía ser de al menos 1 millón de dólares.

DeepMind fue la primera en retirarse tras haber pujado con acciones de su propia empresa en lugar de efectivo, y Microsoft

se retiró en torno a los 22 millones de dólares. Quienes permanecieron hasta el final fueron Google y Baidu. Cuando el precio alcanzó los 44 millones de dólares, Hinton detuvo la subasta y decidió traspasar la empresa a Google. Para él, encontrar el lugar adecuado para continuar su investigación era más importante que obtener la cifra más alta. La adquisición se anunció oficialmente en marzo de 2013.

Tras la adquisición, los tres siguieron caminos distintos. Hinton se incorporó a Google conservando su cátedra en la Universidad de Toronto y posteriormente participó en la investigación de redes neuronales en la compañía como vicepresidente y Engineering Fellow. En Google Brain, Sutskever presentó en 2014 el modelo sequence-to-sequence (Seq2Seq), que transforma una oración en otra, contribuyendo así al avance de la traducción automática. A finales de 2015 dejó Google para fundar OpenAI junto con Sam Altman y otros, liderando como científico jefe las investigaciones que posteriormente darían origen a ChatGPT. Krizhevsky trabajó algunos años más en Google antes de dejar la empresa en 2017.

Esta subasta transformó los cálculos de toda la industria. Al valorarse una empresa de apenas tres personas en 44 millones de dólares, la cotización individual de los investigadores de aprendizaje profundo se redefinió por completo. En enero de 2014, Google adquirió DeepMind, su antigua competidora en la subasta, por aproximadamente 600 millones de dólares. Baidu y Microsoft, tras haber perdido en la puja, también se adentraron de lleno en la investigación del aprendizaje profundo.

El código original de AlexNet pasó a ser propiedad de Google con la adquisición de DNNresearch. Su publicación para el público en general tardaría más de una década. Hansen Shi, curador del Museo de Historia de la Computación (CHM), propuso a Krizhevsky la divulgación del código en 2020 y, tras cinco años de negociaciones con Google, obtuvo el código fuente original. El 20 de marzo de 2025, el museo anunció conjuntamente con Google que el código original de AlexNet quedaba abierto al público y se preservaría de forma permanente. El código publicado comprende unos 200 KB en total, combinando código CUDA de NVIDIA, Python y C++.

Aunque se trata de una escala minúscula en comparación con los modelos de inteligencia artificial actuales, este pequeño código cambió el rumbo de toda una era.

El reconocimiento del habla de 2009 y AlexNet de 2012 salieron del mismo laboratorio. Ambos hitos demostraron con cifras reales la convicción que Hinton había sostenido durante décadas: que las redes neuronales pueden aprender características por sí mismas a partir de los datos, sin que los humanos tengan que fijarles las reglas. Al reunirse las dos condiciones —datos a gran escala y GPU—, las investigaciones acumuladas durante tanto tiempo dieron fruto de golpe. El conexionismo, otrora una disciplina marginal, se convirtió en la corriente principal de la inteligencia artificial a lo largo de estos cuatro años.

Capítulo 8. Reducir el sobreajuste y transferir el conocimiento

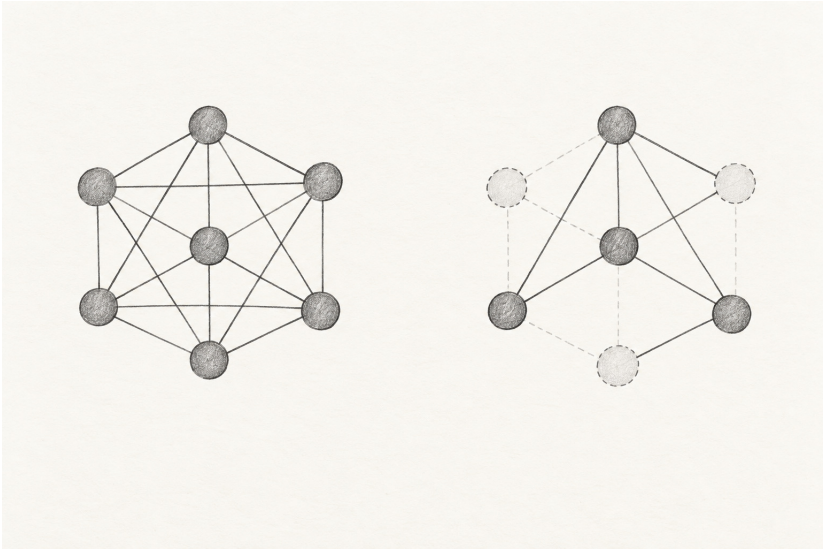


Figura 8. Representación gráfica del dropout, que realiza el entrenamiento desactivando temporalmente una parte de las neuronas.

8.1. El aumento de parámetros y el problema del sobreajuste: el nacimiento del dropout

A principios de la década de 2010, a medida que las redes de creencia profunda (DBN), los autoencoders profundos y AlexNet cosechaban éxitos sucesivos, las capas de las redes neuronales artificiales se volvieron más profundas y la cantidad de parámetros aumentó notablemente. Los parámetros son los números que la red neuronal ajusta por sí misma a medida que aprende. Cuando estos números crecieron hasta decenas y cientos de millones, los investigadores volvieron a encontrarse con un viejo problema: el sobreajuste (overfitting).

El sobreajuste es similar al caso de un estudiante que memoriza de memoria todas las preguntas y respuestas del temario de un examen. Ese estudiante obtiene una puntuación perfecta en los ejercicios prácticos, pero es incapaz de responder cuando aparece una pregunta que nunca antes había visto. De igual modo, cuando los datos de entrenamiento son limitados y la red neuronal entrena repetidamente con los mismos datos

miles de veces, termina memorizando incluso el ruido aleatorio presente en cada muestra en lugar de las reglas subyacentes en los datos. En consecuencia, el error en los datos de entrenamiento se acerca a 0, pero el rendimiento disminuye en los nuevos datos de evaluación. A esto se le conoce como fracaso de la generalización (generalization).

Los campos de la estadística y el aprendizaje automático han empleado técnicas de regularización (regularization), como el decaimiento de pesos L1 y L2 (weight decay), la parada temprana (early stopping) y la compartición de pesos, para prevenir este problema. Sin embargo, en las redes neuronales profundas con múltiples capas ocultas no lineales, el espacio de parámetros era tan inmenso que resultaba difícil controlar el sobreajuste recurriendo únicamente a tales métodos.

Una solución fue el ensamble (ensemble). Un ensamble es un método que consiste en entrenar por separado múltiples modelos con diferentes estructuras o valores iniciales y luego promediar sus predicciones. Se basa en el mismo principio por el cual, al consultar a varias personas y reunir sus respuestas, el

error individual queda compensado. El problema radicaba en el costo. Entrenar una sola red neuronal de gran tamaño requería de varios días a semanas. Entrenar por separado decenas de esas redes neuronales y ejecutarlas simultáneamente en un servicio real resultaba inviable en términos de cálculo y latencia de respuesta.

Geoffrey Hinton presentó el dropout como un método para superar esta limitación. Fue una investigación llevada a cabo junto con Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever y Ruslan Salakhutdinov, de la Universidad de Toronto. En julio de 2012, el equipo de investigación publicó por primera vez en arXiv el artículo «Mejorar las redes neuronales evitando la coadaptación de los detectores de características» y, en 2014, publicó formalmente el artículo definitivo «Dropout: una forma simple de prevenir el sobreajuste en las redes neuronales» en el volumen 15 de Journal of Machine Learning Research (JMLR). La idea central consistía en entrenar, dentro de una sola red neuronal y durante su proceso de aprendizaje, una infinidad de subredes neuronales de menor tamaño que compartían sus

pesos. Era un intento de obtener los beneficios del ensamble sin asumir el costo del ensamble.

La coadaptación (co-adaptation), mencionada en el título del primer artículo, era el problema al que apuntaba el equipo de investigación. En el aprendizaje habitual por retropropagación de errores, las neuronas ocultas tienden a desarrollar relaciones complejas donde compensan los errores ajenos. Esto equivale a una tarea grupal en la que una persona siempre cubre las deficiencias de otra: en el momento en que cambia un miembro del grupo, todo el trabajo se desmorona. Una red neuronal con una acusada coadaptación funciona bien con los datos de entrenamiento, pero ante una ligera variación en los datos entrantes, sus predicciones sufren una gran inestabilidad.

Hinton halló el punto de partida de esta idea en la experiencia cotidiana y en la biología. En una sesión de preguntas y respuestas en línea, relató que al notar que los empleados de ventanilla cambiaban cada vez que iba al banco, supuso que era una medida para impedir que los empleados se confabularan entre sí para cometer irregularidades. La noción

era que, si en la red neuronal también se excluía al azar una parte de las neuronas en cada ocasión, se podría evitar que las neuronas actuaran en complicidad.

La teoría de la reproducción sexual en la biología evolutiva aportó una base teórica a esta idea. La reproducción asexual transmite la combinación genética de los progenitores casi intacta a la descendencia. Aparentemente es ventajosa, ya que no se pierde un conjunto de genes que ya está bien adaptado al entorno. Por el contrario, dado que la reproducción sexual toma y mezcla al azar la mitad de los genes de cada progenitor, las combinaciones genéticas que funcionaban de manera coordinada se dispersan en cada generación. A pesar de ello, la mayoría de los organismos superiores han adoptado la reproducción sexual. En un artículo de 2010, el biólogo Adi Livnat y sus colegas explicaron que la razón radica en que la selección natural favorece la capacidad de mezcla (mix-ability) de los genes por encima de la aptitud de un individuo concreto. Esto significa que sobreviven aquellos genes que cumplen bien su función al combinarse con cualquier otro gen, sin depender de un conjunto específico de genes que siempre los acompañe.

Hinton trasladó este principio tal cual al aprendizaje de las capas ocultas. El dropout desactiva aleatoriamente una fracción de las neuronas en cada paso de entrenamiento, de modo que ninguna neurona pueda confiar en que la neurona contigua estará siempre presente. En consecuencia, cada neurona debe ser capaz de identificar por sí misma características útiles, como pequeños patrones de líneas o puntos, sin importar con qué compañeras le corresponda trabajar. Del mismo modo que se seleccionan los genes que son eficaces al combinarse con cualquier otro gen, se forman neuronas que desempeñan su cometido ante cualquier combinación de neuronas.

8.2. Método de cálculo del dropout y resultados experimentales

El método de cálculo del dropout es sencillo. Puede imaginarse como lanzar una moneda para cada neurona de cada capa cada vez que entra un dato de entrenamiento. Si sale cara con una probabilidad fijada de antemano, esa neurona se mantiene activa; si sale cruz, su salida se convierte en 0 en ese instante. A esta probabilidad de mantener activa la neurona se le llama probabilidad de retención (retention probability). El artículo recomendó como valor predeterminado para la probabilidad de retención un 0.5 en las capas ocultas y alrededor de 0.8 en la capa de entrada.

El procedimiento de cálculo detallado es el siguiente. La salida de una capa determinada es un vector con un valor por cada neurona. Para cada neurona de esa capa, se extrae una variable aleatoria de Bernoulli con valor 0 o 1, donde la probabilidad de obtener 1 es la probabilidad de retención. El vector de ceros y unos obtenido de este modo se multiplica elemento a elemento con el vector de salida, y esta salida parcialmente eliminada se

transfiere como entrada a la capa siguiente. El cálculo de la siguiente capa procede de la manera habitual.

El número de combinaciones posibles para activar y desactivar neuronas en una red neuronal es igual a 2 elevado al número de neuronas. En cada etapa de entrenamiento, el dropout equivale a elegir al azar una de estas combinaciones para entrenarla. Esas numerosas subredes neuronales no existen de forma independiente, sino que comparten los mismos pesos. Se trata de una estructura que permite entrenar un ensamble descomunal al costo de una sola red neuronal.

Una vez finalizado el entrenamiento, cuando la red neuronal se utiliza en la práctica, no se desactiva ninguna neurona. En su lugar, los pesos salientes de cada neurona se multiplican por p. Si durante el entrenamiento una neurona estuvo activa solo con una probabilidad del 50%, la media de la señal recibida por la capa siguiente era la mitad de la original. Al multiplicar los pesos por 0.5 durante la evaluación, la siguiente capa recibe una señal de la misma magnitud que durante el entrenamiento. Con este sencillo ajuste, se obtiene en un solo cálculo el valor

promedio de las predicciones de 2 elevado a n subredes neuronales, más exactamente un valor próximo a la media geométrica.

Los investigadores comprobaron también que el dropout alcanza su máxima efectividad al combinarse con otras técnicas. La primera es la restricción de norma máxima (max-norm), un método que impide que la longitud del vector de pesos entrantes a una neurona sobrepase una constante c (habitualmente entre 3 y 4). La segunda consiste en emplear una alta tasa de aprendizaje y un momento (momentum) elevado. Dado que el dropout introduce ruido en el entrenamiento, se obtuvieron resultados óptimos al usar tasas de aprendizaje entre 10 y 100 veces mayores que en las redes neuronales ordinarias y coeficientes de momento de alrededor de 0.95 a 0.99. Una tasa de aprendizaje elevada conlleva el peligro de que los pesos crezcan desmesuradamente, pero la restricción de norma máxima contrarrestó dicho riesgo. De esta forma, ambas técnicas se complementaron recíprocamente.

En los experimentos, el efecto se constató de forma uniforme en diversos campos. En MNIST, un conjunto de datos de dígitos manuscritos, la tasa de error de una red neuronal ordinaria fue del 1.60%. Una red neuronal con dos capas ocultas y 8,192 neuronas por capa, que combinó dropout con ReLU y restricción de norma máxima, registró un 0.95% sin parada temprana, a pesar de contar con unos 65 millones de parámetros. Tras preentrenarla con una máquina de Boltzmann profunda (DBM) y realizar un ajuste fino con dropout, la tasa de error descendió hasta el 0.79%.

Los resultados fueron semejantes en otros datos. En SVHN, que consiste en fotografías de números de casas tomadas de Google Street View, la tasa de error de una red neuronal convolucional era del 3.95%, pero disminuyó al 3.02% cuando se aplicó dropout solo en las capas totalmente conectadas, y al 2.55% cuando se aplicó en todas las capas. En la tarea de clasificación de imágenes de objetos CIFAR-10, pasó del 15.60% al 12.61%, y en CIFAR-100, del 43.48% al 37.20%. En la tarea de reconocimiento de voz TIMIT, la tasa de error de fonemas de un modelo acústico de 4 capas se redujo del 22.7% al 19.7%.

Asimismo, se apreciaron efectos positivos en la clasificación de artículos de noticias de Reuters y en la predicción del corte y empalme alternativo (splicing alternativo) de genes.

AlexNet, ganadora del concurso de reconocimiento de imágenes ILSVRC en 2012, también incorporó dropout. Krizhevsky, Sutskever y Hinton introdujeron dropout con una probabilidad de retención de 0.5 en dos capas totalmente conectadas iniciales de esta red neuronal de unos 60 millones de parámetros. El artículo manifestó que sin el dropout el sobreajuste habría sido grave, señalando a la vez que, como contrapartida, el número de iteraciones requeridas para la convergencia se duplicó aproximadamente. El dropout conllevó así el precio de un mayor tiempo de entrenamiento.

Al observar el interior de una red neuronal entrenada con dropout, se aprecian también otras propiedades. El valor medio de activación de las neuronas ocultas disminuye y surge de forma espontánea una representación dispersa (sparse representation), en la que la gran mayoría de las neuronas conservan valores cercanos a 0. El artículo recopiló también

variantes como el dropout gaussiano, que en lugar de ceros y unos multiplica por un ruido de distribución normal con media 1.

El dropout se utilizó como herramienta fundamental en las redes neuronales profundas durante más de 10 años a partir de entonces. No obstante, su uso fue evolucionando con las épocas. Tras la llegada de nuevas técnicas como la normalización por lotes (batch normalization), su relevancia disminuyó en las redes neuronales convolucionales, y en los modelos de lenguaje a gran escala contemporáneos, que entrenan con volúmenes ingentes de datos en solo una o dos pasadas, a menudo se entrena desactivando por completo el dropout. Esto se debe a que, si los datos son abundantes en relación con el tamaño del modelo, hay poco margen para que el modelo memorice los datos. Aun así, sigue empleándose ampliamente en campos con escasez de datos o en fases de ajuste fino.

8.3. 'Conocimiento oscuro' y la metamorfosis de los insectos: la concepción de la destilación del conocimiento

Aunque el rendimiento de generalización de las redes neuronales mejoró notablemente gracias al dropout y al ensamble, la industria tropezó con otro obstáculo. Los modelos colosales, con parámetros que oscilaban entre decenas y cientos de millones, o los ensambles que agrupaban decenas de modelos, resultaban difíciles de incorporar directamente en entornos que disponían de escasa memoria y demandaban respuestas inmediatas, como los teléfonos móviles o la búsqueda por voz. Persistía el problema de desplegar efectivamente en la práctica un modelo bien entrenado.

Hinton abordó esta problemática junto a Oriol Vinyals y Jeff Dean, con quienes trabajaba en Google. Los tres expusieron «Destilar el conocimiento en una red neuronal» (Distilling the Knowledge in a Neural Network) en el taller de aprendizaje

profundo de NIPS en 2014, y dicho artículo se difundió en arXiv (1503.02531) en marzo de 2015.

El artículo comienza con el relato de la metamorfosis de los insectos. Insectos como las mariposas o las polillas viven en dos fases completamente distintas: larva y adulto. El cuerpo de la larva está adaptado para asimilar la máxima cantidad de nutrientes del entorno, mientras que el cuerpo del adulto está adaptado para el desplazamiento y la reproducción. Hinton consideró que, en el aprendizaje automático, las formas de entrenamiento y de despliegue debían diferenciarse de este mismo modo. Hasta aquel momento, el modelo empleado para el entrenamiento se utilizaba casi intacto en el servicio. Sin embargo, para la fase de aprendizaje en la que se extrae la estructura a partir de un volumen inmenso de datos, resulta idóneo un modelo voluminoso y pesado o un ensamble con dropout, es decir, la forma larvaria. Concluido el entrenamiento, basta con transferir ese conocimiento a un modelo ágil y liviano, es decir, la forma adulta, para su despliegue. Esta es la premisa fundamental de la destilación del conocimiento (knowledge distillation). En este proceso, al

modelo grande se le denomina profesor (teacher) y al modelo pequeño, estudiante (student).

Ya habían existido tentativas similares con anterioridad. En 2006, Cristian Buciluǎ, Rich Caruana y Alexandru Niculescu-Mizil demostraron en su artículo «Compresión de modelos» (Model compression) una forma de entrenar a un modelo pequeño para que imitase las predicciones de un ensamble voluminoso. Hinton dio un paso más allá. No consideró que el conocimiento del modelo radicara en los números mismos de los pesos, dado que copiar los pesos entre redes neuronales con arquitecturas diferentes carece de sentido. Lo que correspondía transferir era el modo en que se transforman las entradas en salidas, es decir, la función aprendida por el modelo.

La cuestión radicaba entonces en determinar de dónde podía extraerse la información plasmada en esa función. La respuesta de Hinton fue el 'conocimiento oscuro' (dark knowledge). Supongamos que se introduce la imagen manuscrita del número 2 en un modelo profesor bien entrenado. El modelo puede asignar una probabilidad de aproximadamente 0.999 a

que sea un '2', una probabilidad de 1 entre un millón a que sea un '3', y una de 1 entre mil millones a que sea un '7'. Desde la óptica de los objetivos duros (hard targets), que solo indican la respuesta correcta, tanto el '3' como el '7' constituyen exactamente el mismo error. No obstante, el hecho de que entre las probabilidades de ambas respuestas erróneas exista una diferencia de 1,000 veces encierra la información de que el modelo profesor percibe ese '2' mucho más próximo a un '3' que a un '7'. A la distribución que engloba incluso estas reducidas probabilidades atribuidas a las respuestas incorrectas se le denomina objetivo blando (soft target). En sus conferencias, Hinton designó a esta información de similitud que no salta a la vista con el nombre de 'conocimiento oscuro', una analogía con la materia oscura que conforma la mayor parte de la masa del universo sin ser visible.

La divergencia entre ambas clases de objetivos se hace patente en lo que aprende el estudiante. Si al modelo estudiante solo se le enseña la respuesta correcta, el estudiante aprende el modo de acertar. Si se le enseña el objetivo blando, el estudiante aprende además cómo el profesor categoriza el mundo y qué

elementos considera afines entre sí. El estudiante hereda el propio esquema de generalización del profesor.

8.4. Control de la temperatura en softmax y transferencia de conocimiento al modelo estudiante pequeño

Surgía un inconveniente para aprovechar el conocimiento oscuro en el aprendizaje. Probabilidades como 1 entre un millón o 1 entre mil millones son tan ínfimas que apenas ejercen influencia y quedan absorbidas al calcular la pérdida por entropía cruzada empleada en el entrenamiento. El equipo de investigadores resolvió esta dificultad recurriendo a un mecanismo tomado de la física estadística: la temperatura (temperature, T).

La capa final de una red neuronal genera una puntuación para cada categoría (clase) que se pretende clasificar. A estas puntuaciones se les denomina logits. La función softmax transforma estas puntuaciones en probabilidades cuya suma

equivale a 1. En la destilación, se introduce aquí un parámetro denominado temperatura. La función softmax con temperatura se computa de la siguiente manera.

En primer lugar, el logit de cada clase se divide por la temperatura. Dicho valor se introduce en una función exponencial para convertirlo en un número positivo; la función exponencial es una función cuyo resultado se incrementa de forma vertiginosa cuanto mayor es la entrada. La probabilidad de una clase resulta de dividir el valor obtenido para dicha clase entre la suma de los valores obtenidos para la totalidad de las clases.

Por defecto, la temperatura se fija en 1. Si la temperatura se eleva considerablemente, por ejemplo a 20, todos los logits quedan divididos por un número elevado, calculándose como si las disparidades entre puntuaciones se hubieran atenuado. Como consecuencia, probabilidades que eran extremas, como 0.999 y 1 entre un millón, se dispersan de manera uniforme en valores como 0.45, 0.30 y 0.15. Esto se asemeja a reducir la diferencia de volumen en una grabación donde coexisten

sonidos fuertes y susurros, permitiendo que incluso los susurros se perciban con nitidez. Al efectuar la destilación, tanto el modelo profesor como el modelo estudiante calculan la softmax con la misma temperatura elevada, y se instruye al estudiante para que se ajuste a la distribución de probabilidades suavizada del profesor. Concluido el entrenamiento, el modelo estudiante reanuda sus tareas reajustando la temperatura a 1.

El equipo investigador constató que los resultados óptimos se obtienen al estructurar la función de pérdida del estudiante en dos componentes. Uno es la pérdida que confronta el objetivo blando del profesor con la predicción del estudiante a alta temperatura, y el otro es la pérdida que confronta la respuesta correcta real con la predicción del estudiante a temperatura 1. Se calcula una media ponderada de ambas pérdidas, multiplicando la primera por el cuadrado de la temperatura. Esto obedece a que, al incrementar la temperatura, la magnitud del gradiente derivado del objetivo blando decrece de manera inversamente proporcional al cuadrado de la temperatura. Al multiplicar por el cuadrado de la temperatura, la proporción

relativa entre ambas pérdidas permanece inalterable aunque se experimente variando la temperatura. Se consiguieron resultados favorables cuando se asignó una ponderación relativamente menor a la pérdida vinculada a la respuesta correcta.

Este artículo comprende asimismo deducciones matemáticas que clarifican el vínculo entre las investigaciones precedentes y la destilación del conocimiento. Supóngase que la temperatura se incrementa suficientemente por encima de la magnitud de los logits y que el promedio de los logits del profesor y del estudiante se normaliza a 0 en cada caso. Bajo estas condiciones, la dirección en la que el estudiante rectifica sus puntuaciones se aproxima a la dirección que acorta la discrepancia de puntuaciones entre el profesor y el estudiante. El número de clases de clasificación y la temperatura repercuten en la magnitud del ajuste. La destilación a temperaturas elevadas converge así hacia un aprendizaje que minimiza el cuadrado de la diferencia de logits entre ambos modelos. El método de emparejamiento de logits del equipo de Caruana en 2006 constituía, a fin de cuentas, una variante

particular de la destilación del conocimiento. No obstante, elevar la temperatura indefinidamente no resultaba ventajoso. Los investigadores indicaron que, cuando el modelo estudiante es sumamente pequeño, resulta más conveniente una temperatura intermedia. Ello se debe a que, con una temperatura desmedidamente alta, se transmite al estudiante incluso el ruido procedente de puntuaciones insignificantes que el profesor apenas había tomado en consideración.

Los resultados empíricos demostraron con cifras la efectividad de la destilación. En MNIST, un modelo profesor dotado de dos capas ocultas de 1,200 neuronas cada una fue entrenado aplicando dropout y oscilación de imágenes de entrada (jittering), cometiendo 67 fallos sobre un total de 10,000 imágenes de prueba. Un modelo reducido de 800 neuronas entrenado sin regularización falló en 146 imágenes. Cuando se entrenó ese mismo modelo reducido a una temperatura de 20 guiándose por los objetivos blandos del profesor, la cantidad de fallos descendió a 74. Aunque el estudiante jamás había observado directamente las imágenes con oscilaciones, la facultad de generalización que el profesor había cosechado a

partir de esa experiencia le fue transmitida mediante los objetivos blandos.

Los investigadores acometieron además un ensayo todavía más atrevido: suprimieron íntegramente las imágenes del número 3 de los datos de entrenamiento del estudiante. El estudiante entrenó exclusivamente con los objetivos blandos del profesor, sin haber visualizado jamás la cifra 3. El balance arrojó 206 fallos, de los cuales 133 procedían de las 1,010 imágenes del número 3 contenidas en el conjunto de prueba. Ello obedecía a que el estudiante era en extremo reticente a emitir el 3 como respuesta. Cuando los investigadores incrementaron el término de sesgo asignado al 3 en 3.5, los errores totales descendieron a 109, de los cuales solo 14 correspondían al 3. De este modo, reconoció con una exactitud del 98.6% un número que jamás había presenciado. En las pequeñas probabilidades residuales con las que el profesor había calificado imágenes de otros números como 2, 5 y 8 señalando que 'se asemejaban un poco al 3', se hallaba latente la información referente a la silueta del 3.

El equipo científico trasladó igualmente esta técnica al modelo acústico de la búsqueda por voz de Android en Google. El modelo de referencia acreditaba una precisión de clasificación de tramas de voz del 58.9% y una tasa de error de palabras del 10.9%. Un ensamble compuesto por 10 modelos de idéntica estructura registró marcas del 61.1% y del 10.7%. Al destilar dicho ensamble en un modelo unitario de las mismas dimensiones, se obtuvieron un 60.8% y un 10.7%. Prácticamente más del 80% de la ganancia en precisión de tramas conseguida por el ensamble fue absorbida por un solo modelo. En la base de datos interna de Google, JFT, conformada por 100 millones de imágenes y 15,000 etiquetas, los investigadores formularon asimismo un método para entrenar de modo simultáneo diversos modelos especializados (specialist models) dedicados en exclusiva a conjuntos de clases proclives a la confusión.

Posteriormente, Hinton contempló la destilación desde una óptica todavía más dilatada. En su artículo de 2022 relativo al algoritmo Forward-Forward (hacia adelante y hacia adelante), planteó que el lenguaje humano bien podría concebirse como

una suerte de destilación que traslada representaciones desde el cerebro de un individuo hacia el cerebro de otro. En el mismo texto, reparó en la diferencia de que, a diferencia de los modelos digitales cuyos pesos son susceptibles de ser duplicados directamente, los seres humanos únicamente pueden transmitir el conocimiento a través del estrecho conducto de la palabra. Esta inquietud conceptual nutriría con posterioridad sus argumentos de alerta frente a los peligros de la inteligencia digital.

La destilación del conocimiento cobró un auge aún más pronunciado en la era de los modelos extensos de lenguaje. En enero de 2025, la firma china DeepSeek desveló su modelo de razonamiento DeepSeek-R1 y, de forma simultánea, presentó seis variantes de modelos densos con tamaños de 1.5B, 7B, 8B, 14B, 32B y 70B, destilados a partir de dicho modelo sobre las bases de Qwen y Llama. DeepSeek aseveró que el modelo destilado de 14B rebasó holgadamente a QwQ-32B-Preview, la referencia más destacada de acceso abierto en aquel período. El equipo de investigadores de DeepSeek dio cuenta de que, tratándose de modelos compactos, la destilación reporta un

rendimiento superior que la sola aplicación de aprendizaje por refuerzo. Hacia finales de enero de ese mismo año, OpenAI esgrimió que DeepSeek había copiado sus propios modelos mediante la destilación de modelos, de suerte que la destilación trascendió el terreno técnico para erigirse en un litigio de propiedad intelectual y orden competitivo. No obstante, la destilación aplicada hoy en día a los modelos lingüísticos consiste frecuentemente en que el estudiante entrene directamente a partir de las frases formuladas por el profesor, lo cual difiere de la concepción primigenia de Hinton, orientada a traspasar la totalidad de la distribución de probabilidades de las respuestas incorrectas.

El dropout consiguió que las monumentales redes neuronales asimilaran principios en vez de memorizar los datos de entrenamiento, mientras que la destilación del conocimiento hizo viable transferir ese saber adquirido a modelos pequeños aptos para operar en los dispositivos portátiles que los usuarios llevan en sus manos. El primero solventó la problemática del aprendizaje; la segunda, la de la implementación práctica. Ambos procedimientos constituyen un fiel reflejo de la

metodología de investigación de Hinton, por cuanto encontraron sus detonantes en el ámbito biológico: en los genes que cobran vigor al combinarse aleatoriamente y en los insectos que subsisten metamorfoseando la arquitectura de su cuerpo.

Capítulo 9. El avance del aprendizaje profundo construido en conjunto



Figura 9. Ilustración que representa las investigaciones desarrolladas en conjunto por colegas y discípulos.

9.1. El sistema que permitió superar el 'invierno de la IA': el programa NCAP de CIFAR y el modelo de apoyo a largo plazo

A principios de la década de 2000, el centro del mundo académico de la inteligencia artificial estaba ocupado por modelos estadísticos como las máquinas de vectores de soporte (SVM), las técnicas de kernel y los árboles de decisión. Estos métodos facilitaban la explicación matemática de los resultados y ofrecían garantías de hallar soluciones cercanas a la respuesta correcta. La situación de las redes neuronales artificiales de múltiples capas era distinta. Con frecuencia, el aprendizaje se interrumpía a mitad de camino o quedaba atrapado en puntos erróneos, y además resultaba difícil demostrar por qué funcionaban bien. Por ello, muchos investigadores consideraban las redes neuronales como un método obsoleto e inútil. Los artículos sobre redes neuronales eran rechazados habitualmente en las revisiones de las conferencias y resultaba difícil obtener fondos de investigación. Incluso después de superar los períodos previos de

estancamiento de la inteligencia artificial, la frialdad hacia la investigación en redes neuronales se prolongó durante un tiempo.

En este período, la institución que sirvió de soporte para que la investigación en redes neuronales no se interrumpiera fue el Instituto Canadiense de Investigaciones Avanzadas (CIFAR, Canadian Institute for Advanced Research). CIFAR es un organismo de investigación sin fines de lucro fundado en Toronto en 1982, que opera con recursos del gobierno federal de Canadá, los gobiernos provinciales y patrocinadores privados. Esta institución no pertenece a ninguna universidad en particular. Emplea un modelo que reúne a investigadores destacados de varios países en torno a una temática para que colaboren durante un período prolongado.

El vínculo entre Geoffrey Hinton y CIFAR comenzó en 1987. Ese año, Hinton dejó su puesto de profesor en la Universidad Carnegie Mellon (CMU) en Estados Unidos y se trasladó a la Universidad de Toronto en Canadá. En aquel entonces, una parte considerable de los fondos para investigación en

inteligencia artificial en Estados Unidos provenía del Departamento de Defensa, y Hinton no deseaba realizar investigaciones financiadas con dinero de fines militares. Al llegar a Toronto, participó como miembro (fellow) en el primer programa de investigación de CIFAR, 'Inteligencia Artificial, Robótica y Sociedad'.

Diecisiete años después, en 2004, la investigación en redes neuronales seguía siendo ampliamente relegada por la comunidad académica. En ese momento, Hinton propuso un nuevo programa de investigación a CIFAR junto a sus colegas. Yoshua Bengio de la Universidad de Montreal y David Fleet de la Universidad de Toronto, entre otros, se sumaron a la propuesta. CIFAR aceptó la iniciativa e inició ese mismo año el programa 'Computación Neuronal y Percepción Adaptativa (NCAP, Neural Computation and Adaptive Perception)', con Hinton como director del programa. Tras una evaluación de diez años de trayectoria, el programa cambió su nombre a 'Aprendizaje en Máquinas y Cerebros (Learning in Machines & Brains)' y continúa activo hasta hoy.

La razón por la que un tema situado al margen de la atención académica pudo recibir apoyo durante tanto tiempo reside en el modelo de financiamiento de CIFAR. Por lo general, los fondos de investigación se asignan por proyectos. Los investigadores deben comprometerse a producir determinados resultados en plazos concretos y entregar informes en fechas establecidas. CIFAR, en cambio, apoyó a las personas y a los grupos de investigación más que a los proyectos. No impuso temas al selecto grupo de investigadores y les brindó un espacio para reflexionar y debatir juntos a lo largo de varios años. Cada año, los investigadores se reunían en pequeños talleres donde exponían con total libertad ideas preliminares e incluso experimentos fallidos.

Hinton explicó personalmente esta diferencia en una entrevista para el diario Seoul Economic Daily publicada en octubre de 2024. Reconoció que la escala del financiamiento de la investigación en Canadá no es tan grande como la de Estados Unidos. Sin embargo, señaló que, mientras los profesores estadounidenses dedican la mayor parte de su tiempo a redactar propuestas de financiamiento, en Canadá es posible

recibir fondos de investigación durante 3 a 5 años con una propuesta de unas 6 páginas. Afirmó que la fortaleza de Canadá radica en el financiamiento básico que protege la autonomía y la curiosidad de los investigadores.

NCAP no solo apoyó a los investigadores establecidos, sino que también formó a la siguiente generación de investigadores. La escuela de verano iniciada en 2005 fue el cauce para ello. Más tarde, este encuentro se consolidó bajo el nombre de 'Escuela de Verano de Aprendizaje Profundo y Aprendizaje por Refuerzo (Deep Learning + Reinforcement Learning Summer School)'. Cada año, estudiantes de posgrado y jóvenes investigadores de varios países se reunían en Canadá para aprender las técnicas más recientes en redes neuronales y tejer redes de contacto. CIFAR calcula que en 20 años han pasado por esta escuela más de 2.700 investigadores. Al apoyar durante un período prolongado un campo poco demandado sin exigir resultados inmediatos, CIFAR brindó la base institucional indispensable para que el aprendizaje profundo volviera a cobrar protagonismo en el futuro.

9.2. La colaboración de los 'tres padrinos del aprendizaje profundo': Hinton, Bengio y LeCun

En el núcleo de NCAP se encontraban tres académicos que más tarde serían conocidos como los 'tres padrinos del aprendizaje profundo': Geoffrey Hinton de la Universidad de Toronto, Yoshua Bengio de la Universidad de Montreal y Yann LeCun de la Universidad de Nueva York (NYU).

La relación entre los tres se remonta a finales de la década de 1980. LeCun, tras doctorarse en París, Francia, llegó a Toronto en 1987 y trabajó como investigador postdoctoral en el laboratorio de Hinton. En 1988 se trasladó a Bell Labs en Estados Unidos, donde desarrolló las redes neuronales convolucionales (CNN) para leer dígitos manuscritos, investigación que más tarde condujo a un sistema de lectura automática de cifras en cheques bancarios. Bengio, tras obtener su doctorado en la Universidad McGill en Canadá, investigó en el MIT y en Bell Labs en Estados Unidos, donde trabajó junto a LeCun. En 1993 se convirtió en profesor de la

Universidad de Montreal. Así, los tres mantuvieron una estrecha relación de maestro, alumno y colegas.

A partir de mediados de la década de 1990, cuando las redes neuronales quedaron relegadas del interés académico, los tres se encontraron aislados en sus respectivos puestos. Fue el programa NCAP de CIFAR el que volvió a reunir en una comunidad a quienes estaban dispersos entre Toronto, Montreal y Nueva York. Los tres se congregaban en talleres periódicos junto con decenas de investigadores y estudiantes de posgrado, compartiendo abiertamente ideas inacabadas y nuevas fórmulas matemáticas. En una época en que muy pocas personas compartían su visión, contar con una pequeña comunidad donde confiar e intercambiar opiniones constituyó un respaldo decisivo. Tras dejar Hinton la dirección, Bengio y LeCun continuaron liderando conjuntamente el programa.

Los laboratorios de estos tres investigadores dieron lugar sucesivamente a estudios considerados logros clave del aprendizaje profundo. En 2003, el equipo de Bengio presentó un modelo de lenguaje neuronal que expresaba el significado

de las palabras mediante vectores numéricos. En 2006, el equipo de Hinton publicó la red de creencia profunda (DBN) y el método de preentrenamiento capa por capa. Aquel trabajo demostró que incluso las redes neuronales profundas podían entrenarse con éxito si se aprendían sucesivamente capa por capa. En 2014, el equipo de Bengio incorporó el mecanismo de 'atención' a los modelos de traducción, el cual sentó más tarde las bases de los grandes modelos de lenguaje como ChatGPT. Por su parte, LeCun continuó perfeccionando las redes neuronales convolucionales hasta convertirlas en la estructura fundamental de la visión por computadora. Las investigaciones de los tres progresaron apoyándose y retroalimentándose mutuamente.

Los logros de los tres investigadores fueron reconocidos con múltiples galardones. En marzo de 2019, la Association for Computing Machinery (ACM) los anunció como ganadores conjuntos del Premio Turing de 2018. El Premio Turing es conocido como el 'Premio Nobel de la informática'. El 8 de octubre de 2024, se anunció que Hinton, junto con John Hopfield, recibía el Premio Nobel de Física. En diciembre de ese

mismo año, los tres recibieron, junto con Fei-Fei Li y Jensen Huang, el Gran Premio VinFuture, dotado con un total de 3 millones de dólares. En febrero de 2025, siete figuras —Bengio, Hinton, Hopfield, LeCun, Jensen Huang, Bill Dally y Fei-Fei Li— recibieron el Premio de Ingeniería Reina Isabel por su contribución al avance del aprendizaje automático moderno. Entre ellos, Huang y Dally de Nvidia fueron reconocidos por el desarrollo de las GPU, mientras que Li de la Universidad de Stanford lo fue por la creación de la base de datos de imágenes ImageNet. Esta nómina de premiados pone de manifiesto que el desarrollo del aprendizaje profundo contó no solo con algoritmos, sino también con el aporte conjunto de quienes crearon dispositivos de cómputo y datos.

Que los tres colaboraran no significaba que compartieran siempre las mismas opiniones. A partir de 2023, manifestaron posturas divergentes en torno a los riesgos de la inteligencia artificial. Hinton y Bengio advirtieron que la inteligencia artificial podría escapar al control de la humanidad. LeCun, en cambio, consideró que tales temores eran desmedidos y sostuvo que los actuales grandes modelos de lenguaje por sí

solos no pueden alcanzar un nivel de inteligencia comparable al humano.

Esta postura de LeCun también guardó continuidad con su trayectoria profesional. En 2013 se incorporó a Meta para fundar el Laboratorio de Investigación de Inteligencia Artificial de Facebook (FAIR), donde desempeñó funciones durante 12 años: 5 como director fundador de FAIR y 7 como científico principal de IA. En noviembre de 2025 comunicó en Facebook que dejaría Meta a finales de 2025. Asimismo, anunció que crearía una startup orientada a desarrollar sistemas capaces de comprender el mundo físico, razonar y planificar acciones. Aun con discrepancias de opinión, los tres continuaron manteniendo debates públicos. Esta disposición a discutir y contrastar puntos de vista respetando las diferencias guarda estrecha similitud con la cultura de investigación cultivada en los talleres de CIFAR.

9.3. La cuna de investigadores globales de IA: el laboratorio de la Universidad de Toronto y la 'red de Toronto'

La influencia de Hinton no se circunscribe únicamente a sus publicaciones académicas. A medida que los discípulos formados en su laboratorio de la Universidad de Toronto se incorporaron a universidades y empresas de todo el mundo, sus ideas echaron raíces en escenarios cada vez más amplios. A esta red de discípulos y colegas se la conoce habitualmente como la 'red de Toronto'.

Entre sus discípulos más célebres figuran Alex Krizhevsky e Ilya Sutskever. En 2012, ambos crearon junto con Hinton la red neuronal denominada 'AlexNet' y compitieron en la competición internacional de reconocimiento de imágenes (ILSVRC). En este certamen se solicitaba a la red neuronal proponer hasta cinco respuestas por imagen, y la tasa de error top 5 medía la proporción en que la respuesta correcta no se encontraba entre esas cinco opciones. La tasa de error top 5 de AlexNet fue del 15.3%, frente al 26.2% del equipo clasificado en

segundo lugar. Semejante brecha causó una profunda sorpresa en la comunidad de la visión por computadora. El papel de ambos discípulos en este logro fue determinante. Krizhevsky programó directamente el código para entrenar con rapidez la red neuronal utilizando unidades de procesamiento gráfico (GPU). Las GPU, concebidas en origen para procesar los cálculos gráficos en pantalla, son capaces de ejecutar simultáneamente un gran volumen de operaciones idénticas. Gracias a esto, fue factible entrenar en un tiempo razonable una red neuronal con 60 millones de parámetros, variables cuyos valores se determinan a través del aprendizaje. Sutskever, por su parte, diseñó diversas técnicas orientadas a estabilizar el aprendizaje, propiciando conjuntamente el éxito.

A finales de 2012, los tres fundaron la empresa 'DNNresearch'. Google adquirió esta compañía en marzo de 2013 por un monto que trascendió en 44 millones de dólares. Desde entonces, Hinton trabajó en Google al tiempo que proseguía sus investigaciones en la Universidad de Toronto, hasta que abandonó Google en mayo de 2023.

En Google, Sutskever presentó en 2014 junto a sus colegas el modelo 'Sequence to Sequence (Seq2Seq)', que transforma una oración en otra. Este modelo elevó de forma notable el nivel de la traducción automática. En diciembre de 2015 cofundó OpenAI junto a Sam Altman y otros colaboradores, desempeñándose como científico principal. Tras dejar OpenAI en mayo de 2024, fundó en junio de ese mismo año la empresa SSI (Safe Superintelligence), enfocada en el desarrollo de una superinteligencia segura.

Otros discípulos asumieron igualmente roles relevantes en sus respectivas áreas. Ruslan Salakhutdinov publicó en 2006 junto con Hinton un artículo sobre autocodificadores en la revista Science y desarrolló conjuntamente las máquinas de Boltzmann profundas (DBM). Llegó a ser profesor en la Universidad Carnegie Mellon y en 2016 fue nombrado primer director de investigación en inteligencia artificial de Apple. George Dahl y Abdel-rahman Mohamed aplicaron redes neuronales profundas al reconocimiento de voz entre 2009 y 2012. Esta labor dio origen en 2012 a una publicación conjunta con investigadores de Microsoft, Google e IBM, transformando

una tecnología de reconocimiento de voz que llevaba años estancada. Volodymyr Mnih dirigió en DeepMind la investigación sobre redes neuronales de aprendizaje por refuerzo (DQN) que aprenden por sí mismas a jugar videojuegos. Nitish Srivastava sintetizó junto con Hinton la técnica de 'dropout' para mitigar el sobreajuste en las redes neuronales.

Fueron muchos más los investigadores que se formaron en el laboratorio de Hinton. Entre sus alumnos de doctorado figuran Radford Neal, quien investigó las redes neuronales bayesianas; Richard Zemel, posterior director de investigación del Vector Institute; y Brendan Frey y Yee Whye Teh, dedicados al estudio de modelos probabilísticos. Peter Dayan y Max Welling, cuyas investigaciones tendieron puentes entre la neurociencia y el aprendizaje automático, así como Alex Graves, renombrado por sus aportes al reconocimiento de voz y de escritura manuscrita, colaboraron con Hinton como investigadores postdoctorales. Todos ellos se establecieron luego en Toronto, Oxford, Londres, Ámsterdam, Nueva York y Silicon Valley para abrir sus propios laboratorios. De este modo, la cultura de un único laboratorio

se transmitió a múltiples núcleos de investigación en todo el planeta.

9.4. El Vector Institute y la estrategia nacional de IA de Canadá: de la periferia al epicentro de la investigación en IA

A raíz de AlexNet, el aprendizaje profundo se propagó por toda la industria. Compañías como Google, Facebook y Microsoft rivalizaron por incorporar a especialistas en redes neuronales. Canadá, considerada anteriormente una región periférica en la investigación mundial de inteligencia artificial, se ganó el renombre de ser la cuna donde se gestó el aprendizaje profundo. Sin embargo, retener el talento se convirtió en un desafío constante. Buena parte de los investigadores formados en Toronto y Montreal emigraron a Estados Unidos atraídos por remuneraciones más elevadas.

En marzo de 2017, el gobierno de Canadá dio a conocer la 'Estrategia Pancanadiense de Inteligencia Artificial' dentro del

presupuesto federal. El gobierno asignó 125 millones de dólares canadienses a esta iniciativa y delegó su diseño y gestión en CIFAR. Según expuso el ejecutivo canadiense, esta estrategia convirtió a Canadá en el primer país en dotarse de una estrategia nacional de inteligencia artificial con presupuesto asignado. En el presupuesto de 2021 se otorgaron 443 millones de dólares canadienses adicionales para la segunda fase del plan.

El núcleo de la estrategia descansa sobre tres institutos nacionales de inteligencia artificial: el Vector Institute de Toronto, Mila de Montreal (creado por Bengio) y Amii de Edmonton, vinculado a la Universidad de Alberta. Amii guarda un estrecho lazo con Richard Sutton, quien obtuvo el Premio Turing de 2024 junto a Andrew Barto por sus contribuciones al aprendizaje por refuerzo.

Entre ellos, el Vector Institute fue presentado formalmente el 30 de marzo de 2017. Hinton asumió como asesor científico principal del instituto y su discípulo Richard Zemel fue designado director de investigación. La institución contó con

financiamiento del gobierno federal, del gobierno provincial de Ontario y con la inversión de más de 30 empresas, entre ellas Google, Air Canada y los cinco principales bancos de Canadá. El capital inicial de apertura se situó entre 150 y 200 millones de dólares canadienses según diversos medios, y en la documentación oficial figura en torno a 175 millones de dólares canadienses. Los fundadores concibieron a Vector como un polo de atracción capaz de captar y retener profesionales cualificados. Su propósito radicaba en propiciar que un elevado número de estudiantes de doctorado permaneciera en Canadá y en dar respaldo a las empresas emergentes de inteligencia artificial.

El Vector Institute configuró, junto a Mila y Amii, los tres pilares de la investigación canadiense en inteligencia artificial. Empresas como Google, Meta, Uber y Thomson Reuters instalaron centros de desarrollo en Toronto y Montreal. Con ello surgió una tendencia en la que talentos que antes partían rumbo a Estados Unidos retornaron a Canadá o decidieron permanecer allí. Fue en este período cuando Toronto comenzó a ser llamada la 'meca del aprendizaje profundo'.

El esfuerzo inversor de Canadá continuó en los años sucesivos. El presupuesto de 2024 asignó 2.400 millones de dólares canadienses al ámbito de la inteligencia artificial. De ese monto, 2.000 millones de dólares canadienses se canalizan a través del nuevo 'Fondo de Acceso a Computación de IA' y la 'Estrategia Canadiense de Computación Soberana de IA' con el fin de suministrar infraestructura y capacidad de cómputo a investigadores y startups. Otros 50 millones de dólares canadienses se destinan a erigir el Instituto Canadiense de Seguridad de la IA. Al consolidarse los grandes recursos de cálculo como un elemento clave para la competitividad científica, a la inversión centrada en las personas se añadió el respaldo a la infraestructura.

Hinton mantiene una postura definida sobre cómo debe el Estado apoyar la investigación. En la citada entrevista con el diario Seoul Economic Daily, remarcó que a los responsables políticos les resulta tentador anunciar partidas orientadas a proyectos muy concretos. Como contraejemplo, mencionó el 'Proyecto de Computadoras de Quinta Generación' que Japón emprendió en 1982 con un presupuesto colosal. Señaló que

tales macroproyectos con metas prefijadas parecen atractivos a simple vista, pero no arrojaron los frutos esperados. Hinton defendió que el esquema más efectivo consiste en generar un entorno que permita al investigador indagar con libertad al compás de su curiosidad, y recordó que fue justamente así como nació el aprendizaje profundo.

Las redes neuronales fueron en su día una hipótesis marginal en el ámbito académico. El impulso que elevó dicha hipótesis hasta la corriente central de la ciencia no procedió del talento aislado de una sola persona. Fue el resultado concurrente del respaldo a largo plazo de CIFAR sin apremios de resultados inmediatos, de la colaboración entre Hinton, Bengio y LeCun, de los discípulos formados en el laboratorio de Toronto que se proyectaron por el mundo y de la determinación del gobierno canadiense de convertir esos logros en una estrategia de Estado. La trayectoria del aprendizaje profundo pone de relieve que la ciencia no se construye únicamente con el talento individual, sino mediante una labor colectiva forjada codo a codo entre las personas y las instituciones.

Parte 4. En busca de un mejor aprendizaje

Capítulo 10. La máquina que reconoce las partes y el todo



Figura 10. Se representa la relación entre las partes y el todo, que se preserva incluso cuando cambia la perspectiva.

10.1. Límites de las redes neuronales convolucionales (CNN) y críticas al max pooling

Después de que AlexNet ganara la competencia de reconocimiento de imágenes por un amplio margen en 2012, las redes neuronales convolucionales (CNN) se convirtieron en el estándar de la visión por computadora. En el centro de este éxito se encontraba Geoffrey Hinton. Sin embargo, Hinton no aceptó sin más el éxito que había liderado. Consideraba que existía un defecto fundamental en la forma en que las CNN veían los objetos, y expresó esta opinión públicamente durante mucho tiempo.

Para comprender el punto cuestionado por Hinton, primero es necesario examinar el orden en que las CNN procesan las imágenes. Una CNN desplaza pequeños filtros a través de toda la imagen para encontrar características como líneas, bordes y patrones. A medida que pasa por las capas, detecta características más grandes como ojos, ruedas u orejas. Entre estas capas se intercala una operación llamada max pooling. El

max pooling es una operación que conserva solo el valor más grande dentro de una región pequeña y descarta el resto. De este modo, aunque una característica se desplace ligeramente, el resultado no cambia de forma significativa. Esto se denomina invariancia de posición. También reduce la cantidad de cálculo.

El precio a pagar es la pérdida de información. La información sobre la postura (pose), como dónde estaba exactamente la característica, hacia qué dirección estaba inclinada o cuál era su tamaño, desaparece en la etapa de pooling. En una sesión de preguntas y respuestas celebrada en la comunidad en línea Reddit en 2014, Hinton afirmó que la operación de pooling en las CNN era un gran error y que el hecho de que funcionara tan bien era, de hecho, un desastre. Quería decir que, precisamente porque funcionaba bien, los investigadores no intentaban buscar un método mejor.

Lo que sucede cuando desaparece la información de la postura se explica fácilmente con el dibujo de un rostro. Los seres humanos no solo verifican si hay ojos, nariz y boca. También toman en cuenta la disposición en la que las partes componen

el todo: si los dos ojos están alineados sobre la nariz y si la boca está debajo de ella. A esto se le llama jerarquía parte-todo (part-whole hierarchy). Las CNN tendían a clasificar como rostro incluso una imagen en la que los ojos y la boca habían intercambiado posiciones, siempre que los detectores de cada parte estuvieran activados. Esto se debía a que carecían de un mecanismo para evaluar con precisión las relaciones espaciales entre las partes.

El cambio de punto de vista también era un desafío no resuelto. Si se observa el mismo objeto desde arriba, desde un lado o en diagonal, los valores de los píxeles cambian por completo. Para tolerar estas diferencias, las CNN entrenaban con grandes cantidades de imágenes rotadas, estiradas y distorsionadas. Esto se conoce como aumento de datos (data augmentation). A los ojos de Hinton, esto se parecía más a la memorización que a la comprensión.

Por ello, Hinton redefinió el objetivo. Consideró que no se necesitaba la invariancia (invariance), donde la salida permanece igual aunque el objeto se mueva, sino la

equivariancia (equivariance), en la que la representación interna cambia de forma regular en la misma medida en que el objeto se mueve. Es decir, se debe seguir reconociendo qué es el objeto, pero al mismo tiempo se debe saber dónde está y en qué postura se encuentra.

Las raíces de esta idea se remontan a 1979. Aquel año, Hinton publicó un artículo sobre imágenes mentales (mental imagery) y descripciones estructurales en la revista académica Cognitive Science. En él aparece el experimento del cubo. Si se coloca un cubo hecho de alambre de modo que un vértice quede abajo y el vértice opuesto quede justo arriba, y se le pide a la gente que señale con la mano la posición de los vértices restantes, la mayoría se equivoca. Esto se debe a que en la mente conciben el cubo en un sistema de coordenadas habitual. Mediante este experimento, Hinton demostró que los seres humanos asignan un sistema de coordenadas propio a los objetos y que, según dicho sistema de coordenadas, perciben de forma diferente una misma forma. Más de 40 años después, el artículo sobre GLOM también cita este experimento como base en su apertura.

10.2. El sistema de coordenadas de los objetos y los autoencoders de transformación: la aparición de las cápsulas

Para trasladar a las redes neuronales la idea de que cada objeto posee un sistema de coordenadas, Hinton buscó pistas en la computación gráfica. Los videojuegos y animaciones tridimensionales asignan a cada objeto un sistema de coordenadas propio y expresan la relación de dicho objeto con la cámara mediante matrices. Una matriz es una tabla que organiza números en filas y columnas, capaz de representar transformaciones como rotar, trasladar o escalar un objeto con una sola multiplicación. Si el punto de vista cambia, basta con multiplicar por una matriz. En el espacio de píxeles, compuesto por el brillo de cada píxel individual, los cambios de punto de vista se manifiestan de forma complejamente intrincada. En cambio, en el espacio de posturas, donde la posición, la orientación y el tamaño del objeto se registran numéricamente, el mismo cambio se convierte en un cálculo lineal compuesto únicamente por multiplicaciones y sumas. Si los gráficos crean

una imagen a partir de información de postura, el reconocimiento visual consiste en deducir la postura a la inversa a partir de la imagen. Hinton denominó a esto gráficos inversos (inverse graphics).

Esta idea se presentó por primera vez como artículo en la Conferencia Internacional sobre Redes Neuronales Artificiales (ICANN 2011), celebrada en Espoo, Finlandia, en 2011. Hinton, junto con Alex Krizhevsky y Sida D. Wang, presentó «Transforming Auto-encoders» (Autoencoders transformadores). En este artículo se utilizó por primera vez el término 'cápsula' (capsule).

Las neuronas convencionales producen un solo número. Una cápsula es un pequeño grupo de neuronas que genera dos tipos de información al mismo tiempo. Uno es la probabilidad de que el objeto que tiene asignado esté presente en esa posición. El otro es un conjunto de múltiples números, es decir, un vector, que representa la postura de dicho objeto, como su posición, orientación y tamaño. El artículo llamó a estos números parámetros de instanciación (instantiation parameters).

El método de entrenamiento para lograr que la cápsula contenga la postura es el siguiente. Se le proporciona a la red neuronal la imagen original junto con valores de desplazamiento (Δx, Δy), tales como 'mueve tantas unidades a la derecha y tantas hacia arriba'. La parte de reconocimiento de la cápsula averigua la posición (x, y) del objeto a partir de la imagen original. Al sumar los valores de desplazamiento a esto, se obtiene (x+Δx, y+Δy), y la parte generativa de la cápsula dibuja la imagen trasladada de acuerdo con estas coordenadas. La respuesta correcta es la imagen realmente trasladada. La red neuronal entrena mediante retropropagación para reducir la diferencia entre el resultado dibujado y la respuesta correcta. La probabilidad de existencia proporcionada por la cápsula se multiplica por la salida de la parte generativa, reduciendo la influencia de esa cápsula en la imagen si el objeto no está presente.

La clave aquí es que nadie le enseñó a la cápsula que 'este número es la posición'. Para que una simple suma produzca el resultado correcto, los números dentro de la cápsula deben comportarse como la posición real. La red neuronal aprendió

por sí misma a convertir la información dada en píxeles en números que representan la postura.

Las cápsulas también propusieron una manera de conectar las partes con el todo, resolviendo el problema del rostro expuesto en la sección 10.1. Supongamos que una cápsula de nariz y una cápsula de boca descubren cada una su propia postura. Existe una relación espacial fija entre la nariz y el rostro. Si se multiplica la postura de la nariz por la matriz que representa esta relación, se puede predecir la postura del rostro. Lo mismo ocurre con la boca. Si las dos predicciones coinciden entre sí, existe una alta probabilidad de que haya un rostro en ese lugar. Esto se debe a que es muy raro que predicciones formadas por múltiples números coincidan por casualidad. Esta verificación de coincidencia se convirtió más tarde en el principio central de la investigación sobre cápsulas.

10.3. Enrutamiento dinámico y enrutamiento EM: el desarrollo en pleno de las redes de cápsulas (2017-2018)

El concepto de cápsula se fue perfeccionando durante 6 años. El resultado apareció en dos artículos publicados en 2017 y 2018. Los autores fueron Sara Sabour, Nicholas Frosst y Hinton, del laboratorio de Google Brain en Toronto.

Presentado en diciembre de 2017 en la conferencia NIPS, «Dynamic Routing Between Capsules» (Enrutamiento dinámico entre cápsulas) convirtió a las cápsulas en una red neuronal que funcionaba en la práctica. Este modelo se conoce comúnmente como CapsNet. CapsNet incorpora dos mecanismos.

El primer mecanismo es la función de compresión (squash). En el vector de salida de una cápsula, la longitud representa la probabilidad de que el objeto esté presente y la dirección representa su postura. Al ser una probabilidad, la longitud debe situarse entre 0 y 1. La función squash mantiene intacta la

dirección del vector y reduce únicamente su longitud para que quede entre 0 y 1. Comprime los vectores cortos acercándolos a 0 y los vectores largos acercándolos a 1.

El segundo mecanismo es el enrutamiento dinámico, más exactamente el 'enrutamiento por acuerdo' (routing by agreement). Este traslada la verificación de coincidencia vista en la sección anterior al procedimiento de cálculo de la red neuronal. Las cápsulas de la capa inferior (partes) multiplican por una matriz de pesos aprendida para predecir la salida de cada cápsula de la capa superior (todos). Al principio, envían sus predicciones de manera uniforme a todas las cápsulas de la capa superior. Las cápsulas de la capa superior reúnen las predicciones recibidas y generan una salida. A continuación, mediante el producto escalar, se mide en qué grado cada predicción apunta en la misma dirección que la salida de la capa superior. La próxima vez se envía más a las cápsulas superiores que encajan bien y menos a las que no encajan. Este proceso se repite varias veces. La configuración predeterminada en el artículo fue de tres repeticiones.

Este método es exactamente lo contrario del max pooling. El pooling selecciona solo el valor más grande y descarta el resto. El enrutamiento no descarta información, sino que determina de nuevo para cada imagen qué parte pertenece a qué todo. Es un método en el que una parte busca el todo al que pertenece de entre varios candidatos a todo.

Los resultados experimentales también respaldaron este diseño. CapsNet registró una tasa de error del 0.25% en el conjunto de datos de dígitos manuscritos MNIST. Fue un resultado obtenido con una red neuronal superficial de solo tres capas. El artículo también añadió un mecanismo de reconstrucción que redibujaba el dígito original únicamente a partir de la salida de las cápsulas de dígitos. Al variar poco a poco un número del vector de la cápsula, el grosor del trazo, la inclinación y el ancho del dígito reconstruido cambiaban en consecuencia. Esta fue una prueba de que los números dentro de la cápsula contenían realmente la postura y la forma.

El experimento MultiMNIST demostró las ventajas de las cápsulas con mayor claridad. Se trataba de una tarea

consistente en reconocer ambos dígitos en imágenes donde dos números se superponían en gran medida. CapsNet obtuvo un mejor desempeño que la CNN de comparación con una tasa de error de alrededor del 5%, e incluso redibujó por separado cada uno de los dos dígitos superpuestos. Esto significaba que cada cápsula de dígito tomó únicamente los trazos correspondientes a su propio dígito.

En la conferencia ICLR de 2018 se publicó el artículo de seguimiento «Matrix Capsules with EM Routing» (Cápsulas matriciales con enrutamiento EM). Esta vez, Hinton fue el primer autor. En este modelo, una cápsula tiene un valor que indica la presencia y una matriz de postura de 4×4. La matriz de 4×4 es exactamente el mismo formato que utiliza la computación gráfica para representar la posición y rotación de objetos tridimensionales. Supuso incorporar aún más profundamente las ideas de los gráficos en las redes neuronales.

El método de enrutamiento también cambió. Al multiplicar la matriz de postura de la cápsula de la capa inferior por una

matriz de transformación, se obtiene un 'voto' sobre la postura de la cápsula de la capa superior. La cápsula de la capa superior observa si múltiples votos se concentran en un mismo lugar. Para esta determinación se utilizó el algoritmo de esperanza-maximización (EM) de la estadística. En el paso E, se calcula la probabilidad de que cada voto pertenezca a la agrupación de una cápsula de la capa superior. En el paso M, basándose en esa probabilidad, se recalculan la media, la varianza y la activación de la cápsula de la capa superior. Si los votos se agrupan densamente, se determina que dicho todo existe.

Este modelo se probó con los datos de smallNORB. Se trata de un conjunto de datos para el reconocimiento de objetos tridimensionales fotografiados desde diversos ángulos e iluminaciones, que incluye aviones de juguete, automóviles, camiones, figuras humanas y muñecos de animales. La tasa de error de las cápsulas matriciales fue del 1.8% cuando se evaluó recortando una sola región de la fotografía, y del 1.4% cuando se combinaron los resultados de múltiples regiones. Esta última cifra representó aproximadamente un 45% menos de errores

que el 2.56% del método convencional de comparación. También superó a una CNN de tamaño similar en la prueba de reconocer fotografías tomadas desde ángulos no vistos durante el entrenamiento. Asimismo, resistió mejor los ataques adversarios, es decir, ataques que engañan a la red neuronal mediante un ruido minúsculo invisible para el ojo humano.

En 2019, Adam Kosiorek, Sabour, Yee Whye Teh y Hinton presentaron «Stacked Capsule Autoencoders» (Autoencoders de cápsulas apiladas) en NeurIPS. Era un modelo de aprendizaje no supervisado que descubría plantillas de partes con solo observar las imágenes sin etiquetas de respuesta correcta y las agrupaba para formar cápsulas de objetos.

Las redes de cápsulas recibieron gran atención de la comunidad académica, pero sus limitaciones también eran evidentes. El cálculo era lento debido a la necesidad de repetir el enrutamiento varias veces, y resultaba difícil escalar a imágenes grandes o datos complejos. Al mismo tiempo, a medida que los transformers se convertían rápidamente en la corriente principal tanto en el procesamiento del lenguaje

como en el de imágenes, las cápsulas no lograron consolidarse como una tecnología dominante.

Aun así, la investigación continuó. En marzo de 2024, investigadores de la Universidad de Aberdeen en el Reino Unido tomaron como punto de partida el hecho de que las redes de cápsulas no aprendían bien con datos complejos y no habían podido escalarse a tareas modernas, y propusieron el autoencoder de cápsulas enmascaradas (MCAE, Masked Capsule Autoencoder), la primera red de cápsulas en emplear preentrenamiento autosupervisado. Como resultado del preentrenamiento con los datos de Imagenette, compuestos por fotografías de 10 categorías del tamaño de ImageNet, logró el rendimiento más alto entre las redes de cápsulas y mostró un desempeño un 9% mejor que el modelo de referencia. Esta investigación se publicó en 2025 en la revista académica TMLR (Transactions on Machine Learning Research).

10.4. Una propuesta para unificar múltiples investigaciones: GLOM, la jerarquía parte-todo y las islas de vectores idénticos

Hinton consideraba que en el modelo de cápsulas subsistía un problema más profundo que la velocidad de cálculo. En las cápsulas existentes era necesario preparar de antemano cápsulas para todos los tipos de partes y objetos que pudieran aparecer en cada posición de la imagen. En una computadora convencional, bastaría con asignar nueva memoria según sea necesario para cada imagen y construir una estructura de árbol. Las redes neuronales del cerebro no pueden hacer eso. Esto se debe a que la función que realiza una neurona está determinada por los pesos de conexión entrantes y salientes, y las neuronas reales no pueden cambiar esos pesos de forma rápida y total. Siendo así, la cuestión radica en cómo una red neuronal de estructura fija puede representar un árbol de partes y el todo cuya forma varía con cada imagen. Hinton convirtió esta cuestión en su nuevo proyecto de investigación.

El 25 de febrero de 2021, Hinton publicó en arXiv un artículo de 43 páginas titulado «How to represent part-whole hierarchies in a neural network» (Cómo representar jerarquías parte-todo en una red neuronal). El nombre de esta propuesta, GLOM, proviene de la expresión coloquial en inglés 'glom together', que significa unirse o aglomerarse en un solo bloque, término que a su vez podría derivar de 'agglomerate'. En su primera frase, el artículo aclara que no se trata de un texto que describa un sistema en funcionamiento, y presenta una forma de representación que unifica diversos logros de investigación como los transformers, los campos neuronales (neural fields), el aprendizaje de representaciones contrastivas, la destilación de conocimiento y las cápsulas, junto con el sistema imaginario GLOM que la alberga. Hinton presentó esta misma idea en conferencias como la celebrada el 5 de mayo de 2021 en el Instituto Perimeter de Física Teórica de Canadá.

La estructura de GLOM es la siguiente. Divide la imagen en pequeños parches y coloca una 'columna' (column) para cada parche. Todas las columnas utilizan los mismos pesos. Cada columna se divide en varios niveles (levels). Se apilan desde

abajo en el orden de subpartes, partes, objetos y escenas. En cada nivel hay un vector, es decir, una incrustación (embedding), que indica qué hay en esa posición. Cada uno de estos vectores actúa como una cápsula universal no ligada a un tipo específico. De esta manera, se alivia la carga de las cápsulas tradicionales, que requerían preparar cápsulas de antemano para cada tipo.

Cuando entra una imagen, GLOM refina sucesivamente cada vector a lo largo de varios pasos temporales. Un vector es un conjunto ordenado de múltiples números que aquí indica qué hay en una determinada posición de la imagen. En cada paso, el vector de un nivel se sustituye por el promedio ponderado de cuatro valores. Un promedio ponderado es una media calculada asignando ponderaciones distintas a cada valor. El primer valor es la predicción obtenida al pasar el vector del nivel inmediatamente inferior a través de una red neuronal ascendente (bottom-up). El segundo valor es la predicción obtenida al pasar el vector del nivel inmediatamente superior a través de una red neuronal descendente (top-down). La red descendente también recibe como entrada la posición de la

columna, ya que al conocer la información de 'rostro' y la ubicación, puede calcular si en ese lugar debe haber una nariz o un ojo. Esta es una idea tomada de la investigación sobre campos neuronales. El tercer valor es el valor inmediatamente anterior de su propio nivel. El cuarto valor es el promedio mediante atención de los vectores del mismo nivel de las columnas vecinas, otorgando mayor peso a aquellos vecinos cuyos vectores sean más similares al propio. Aunque se asemeja a la atención de los transformers, es una forma simplificada que compara los vectores directamente entre sí, sin transformarlos por separado en consultas (queries), claves (keys) y valores (values).

Al repetir este proceso, emerge el fenómeno central de GLOM. Las columnas de todos los parches correspondientes al rostro convergen en un vector idéntico en el nivel de objeto. Los parches correspondientes a la nariz comparten el mismo vector en el nivel de parte, y los correspondientes a la boca comparten otro vector idéntico. Hinton denominó a esto 'islas de vectores idénticos' (islands of identical vectors). Cada isla constituye un nodo de la estructura de árbol. Si debajo de una gran isla en el

nivel de objeto se sitúan varias islas más pequeñas en el nivel de parte, ello representa precisamente la relación parte-todo entre el rostro, la nariz y la boca. De este modo, es posible representar un árbol diferente para cada imagen sin necesidad de reasignar neuronas. Esta es la respuesta que Hinton ofreció al problema de cómo una red neuronal de estructura fija puede representar una estructura diferente para cada imagen.

Como método de entrenamiento, Hinton propuso un enfoque consistente en reconstruir imágenes parcialmente enmascaradas. También añadió una regularización de estilo de aprendizaje contrastivo para que los vectores de las ubicaciones pertenecientes al mismo objeto se asemejaran entre sí. Cuando las islas se forman adecuadamente, a los seres humanos les resulta más fácil interpretar qué vio la red neuronal y cómo lo vio. El resumen del artículo también señaló que, si GLOM llegara a funcionar, incrementaría en gran medida la interpretabilidad de las representaciones generadas por los sistemas basados en transformers empleados en visión o lenguaje.

Hinton consideraba que GLOM conecta dos tradiciones en la investigación sobre la percepción. Una de ellas es la psicología de la Gestalt, que contempla la percepción como un campo (field) y sostiene que el todo es diferente de la suma de sus partes. La otra es la inteligencia artificial simbólica clásica, que intentó representar la percepción mediante símbolos y descripciones estructurales. En GLOM, la percepción se manifiesta como un campo continuo de vectores, pero dentro de la estructura formada por las islas se plasman relaciones claras entre las partes y el todo.

Aún no se conocen casos ampliamente difundidos de GLOM implementado como un sistema a gran escala completamente terminado. El propio Hinton trasladó su atención hacia las reglas de aprendizaje mismas al presentar el algoritmo Forward-Forward a finales de 2022. Sin embargo, las cuestiones planteadas por GLOM siguen siendo válidas: cómo una red neuronal fija puede representar una estructura diferente para cada imagen y cómo hacer que esa representación sea legible para los humanos. La búsqueda de Hinton, que comenzó con el experimento del cubo en 1979, pasó por las cápsulas en 2011, el

enrutamiento dinámico en 2017 y las cápsulas matriciales en 2018, hasta llegar a GLOM en 2021, estuvo orientada en su totalidad hacia este problema.

Capítulo 11. Redescubrir métodos de aprendizaje en el cerebro

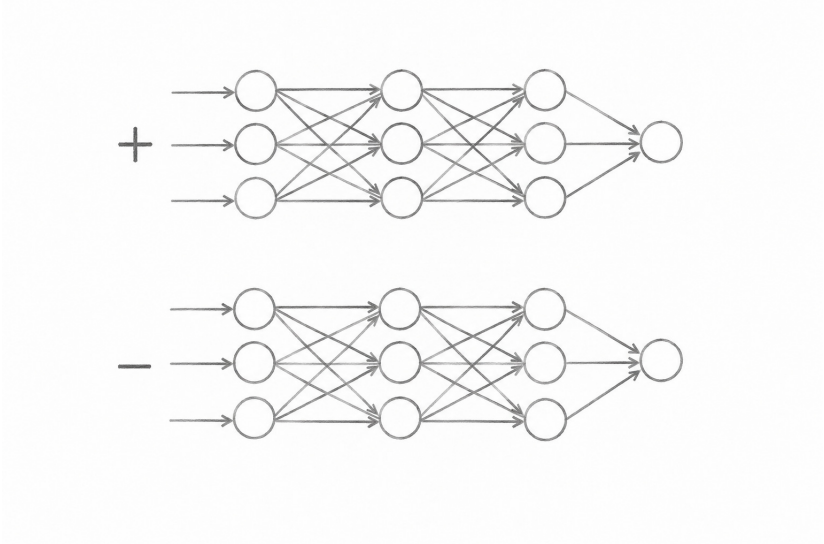


Figura 11. Representación del aprendizaje que transmite hacia adelante casos positivos y casos negativos.

11.1. ¿Aprende realmente el cerebro mediante retropropagación?

En 1986, David Rumelhart, Geoffrey Hinton y Ronald Williams publicaron en Nature su artículo sobre la retropropagación de errores (Backpropagation). El procedimiento de la retropropagación es el siguiente. Cuando una red neuronal produce una respuesta, primero se calcula la diferencia con la respuesta correcta, es decir, el error. A continuación, utilizando la regla de la cadena (Chain Rule) del cálculo diferencial, este error se envía en sentido inverso desde la capa de salida hacia la capa de entrada, calculando cuánto debe modificarse cada peso de conexión. Este método se convirtió durante casi los 40 años posteriores en la tecnología central que sustentó el aprendizaje profundo. Desde el reconocimiento de imágenes y la traducción automática hasta los grandes modelos de lenguaje (LLM), la gran mayoría de los logros se han obtenido sobre la base de la retropropagación.

Sin embargo, durante mucho tiempo Hinton no abandonó una pregunta: si la corteza cerebral humana también aprende de

esta manera. Su respuesta fue en gran medida negativa. En 2020, Hinton publicó junto con Timothy Lillicrap y otros el artículo «La retropropagación y el cerebro (Backpropagation and the brain)» en Nature Reviews Neuroscience. Al examinar conjuntamente este artículo y las investigaciones de Hinton de 2022, las razones por las cuales es difícil que la retropropagación ocurra tal cual en el cerebro pueden resumirse en tres.

La primera es el problema del transporte de pesos (Weight Transport Problem). Para enviar el error hacia atrás en la retropropagación, se necesita la matriz traspuesta de la matriz de pesos empleada en el cálculo hacia adelante. Esto significa que la vía de retroceso debe estar emparejada exactamente con la misma fuerza que la vía de avance. Sin embargo, en las conexiones neuronales reales del cerebro no se ha constatado tal estructura simétrica uno a uno.

La segunda es que en el cerebro no existen ni vías independientes para enviar exclusivamente señales de error ni dispositivos para almacenar estados previos. La

retropropagación debe enviar a través de una vía separada los valores de las derivadas del error, diferenciados de las señales habituales, y para realizar el cálculo inverso es necesario memorizar los valores de activación obtenidos durante el cálculo hacia adelante. Las conexiones de retroalimentación del cerebro no funcionan de este modo. La retroalimentación solo regula la intensidad de la actividad de las neuronas que reciben la señal, y todavía no existe evidencia neurocientífica de que transporte por separado valores de error con signo.

La tercera es la cuestión del tiempo. Para aprender datos que ingresan continuamente, como en un video, se debe utilizar la retropropagación a través del tiempo (BPTT). Puesto que este método debe enviar los errores de regreso a momentos del pasado, exige detener con frecuencia el procesamiento de entradas. En cambio, el cerebro aprende al mismo tiempo que deja fluir sin interrupción las señales que entran por los ojos y los oídos. Hinton explicó esto con el término segmentación (Pipelining).

Este planteamiento de Hinton no debe interpretarse como un argumento para desechar la retropropagación. En diversas entrevistas, él explicó que sus dudas surgieron desde la perspectiva de intentar comprender el cerebro. En la inteligencia artificial que opera en computadoras digitales, la retropropagación sigue ofreciendo un rendimiento superior al de otros métodos de aprendizaje. No obstante, su planteamiento era que, si el cerebro no utiliza la retropropagación, debe existir otro método que sí emplee, y ese método podría constituir una clave para construir un nuevo tipo de computadoras.

Hinton también explicó la diferencia entre el cerebro y las redes neuronales artificiales mediante la proporción entre los datos y la cantidad de conexiones. Los grandes modelos de lenguaje tienen alrededor de 1 billón de parámetros, pero los textos utilizados para el entrenamiento alcanzan varios billones de tokens. Es una estructura donde la experiencia abunda y el espacio de almacenamiento es relativamente escaso. La retropropagación sobresale al comprimir tanta experiencia en un número reducido de parámetros. En los seres humanos

ocurre exactamente lo contrario. Aunque se viva durante 80 años, la vida solo dura unos 2.500 millones de segundos, mientras que las conexiones sinápticas del cerebro alcanzan aproximadamente 100 billones. Es una estructura donde la experiencia es escasa y las conexiones son abundantes. Hinton consideraba que, bajo estas condiciones, el cerebro emplearía una forma de aprendizaje distinta de la retropropagación.

11.2. El algoritmo Forward-Forward: dos cálculos hacia adelante y la puntuación llamada «bondad»

En diciembre de 2022, Hinton presentó un nuevo método de aprendizaje en la conferencia magistral de la Conferencia sobre Sistemas de Procesamiento de Información Neuronal (NeurIPS), celebrada en Nueva Orleans, Estados Unidos. Ese mismo mes, se publicó en el sitio de divulgación científica arXiv el artículo «The Forward-Forward Algorithm: Some Preliminary Investigations» (arXiv:2212.13345). En aquel momento, formaba parte de Google Brain Toronto. Como se deduce de haber incluido en el título las palabras 'investigaciones preliminares', Hinton presentó este método no como un resultado definitivo, sino como un punto de partida.

La idea fundamental es sencilla. La retropropagación calcula una vez desde la entrada hacia la salida y otra vez desde la salida hacia la entrada. El algoritmo Forward-Forward (FF) suprime el cálculo hacia atrás y realiza dos cálculos hacia adelante con el mismo procedimiento. Uno se calcula con datos

reales y el otro con datos falsos. Los datos falsos son datos que se parecen a los reales, pero que se han creado deliberadamente erróneos. La etapa en la que se introducen datos reales se denomina pasada positiva (Positive Pass), y la etapa en la que se introducen datos falsos se llama pasada negativa (Negative Pass).

Esta idea tiene precedentes en investigaciones anteriores. En la década de 1980, Hinton investigó la máquina de Boltzmann (Boltzmann Machine) junto con Terry Sejnowski. La máquina de Boltzmann aprendía comparando la etapa en la que observaba los datos con la etapa en la que observaba los estados generados por el propio modelo. La estimación contrastiva de ruido (NCE), propuesta en 2010 por Michael Gutmann y Aapo Hyvärinen, también entrena el modelo distinguiendo entre datos reales y ruido. El FF reconfiguró para las redes neuronales multicapa actuales el principio de contraste (Contrast) común a ambas investigaciones, es decir, el principio de comparar dos tipos de entradas y aprender a partir de sus diferencias.

El criterio mediante el cual cada capa distingue lo real de lo falso es una puntuación que se calcula por separado en cada capa. Hinton llamó a esta puntuación bondad (Goodness). La bondad es el valor resultante de elevar al cuadrado los valores de activación de las neuronas de esa capa y sumarlos todos. El valor de activación es la magnitud de la salida que emite una neurona al recibir una entrada, y como función de activación se utiliza ReLU. ReLU es una función que produce 0 si la entrada es menor que 0 y entrega la entrada tal cual si es mayor que 0. Por lo tanto, cuanto más intensamente reaccionan las neuronas, mayor es la bondad de esa capa.

Los objetivos del aprendizaje son dos. Cuando entran datos reales, se ajustan los pesos para que la bondad sea suficientemente mayor que un valor de referencia predeterminado; cuando entran datos falsos, se ajustan para que la bondad sea suficientemente menor que dicho valor de referencia. Cada capa calcula la probabilidad de que la entrada sea real a partir de la bondad de la siguiente manera.

Primero, se resta el valor de referencia de la bondad de esa capa. El valor obtenido al pasar esta diferencia por la función sigmoide es la probabilidad de que la entrada sea real.

La función sigmoide transforma cualquier valor en un número entre 0 y 1. Si la bondad es mayor que el valor de referencia, la probabilidad se aproxima a 1; si es menor, se aproxima a 0. Cada capa ajusta únicamente sus propios pesos para que esta probabilidad resulte alta con datos reales y baja con datos falsos. No hay necesidad de recibir errores desde otras capas. Este método en el que cada capa aprende exclusivamente con la información interna de su propia capa constituye el aprendizaje local (Local Learning) del FF.

Sin embargo, al apilar varias capas surge un problema. Si la primera capa discrimina bien entre lo real y lo falso, el vector de salida será largo cuando sea real y corto cuando sea falso. Si la segunda capa juzga únicamente observando esta longitud, no tendrá nada nuevo que aprender. Hinton solucionó este problema mediante la normalización de capa (Layer Normalization). La normalización de capa es una técnica

presentada conjuntamente en 2016 por Jimmy Ba, Jamie Kiros y Hinton. El FF utiliza esta técnica para eliminar la longitud del vector y conservar solo su dirección antes de transferirlo a la siguiente capa. La dirección indica en qué proporción relativa se activan las neuronas entre sí. En consecuencia, la capa superior no puede depender de la información ya utilizada por la capa inferior y debe descubrir nuevas características ocultas en el patrón de actividad.

El problema pendiente es cómo crear los datos falsos. El FF se puede aplicar tanto al aprendizaje no supervisado sin etiquetas (labels) como al aprendizaje supervisado con etiquetas, y la forma de generar datos falsos varía en ambos casos. En el aprendizaje no supervisado, Hinton generó imágenes falsas combinando dos imágenes de dígitos manuscritos. Tras aplicar desenfoque varias veces a una imagen de puntos aleatorios para obtener una máscara con forma de grandes manchas, tomó la parte interior de la máscara de una imagen y la exterior de la otra y las unió. La imagen así creada se asemeja a una real en detalles pequeños como trazos breves, pero la figura global no es un dígito. Al distinguir entre las imágenes reales y estas

imágenes falsas, la red aprende a captar formas en un rango amplio.

En el aprendizaje supervisado, la etiqueta se introduce directamente dentro de la imagen. Es un procedimiento en el que se representa un dígito en la posición de los primeros 10 píxeles de una imagen de dígitos manuscritos de MNIST. Si se coloca la etiqueta correspondiente a la imagen, se obtiene un dato positivo; si se coloca una etiqueta errónea, se obtiene un dato negativo. Ambos datos comparten la misma imagen y difieren únicamente en la etiqueta. Por esa razón, la red ignora el fondo irrelevante para la etiqueta y se enfoca en la relación existente entre la imagen y la etiqueta. Al realizar pruebas una vez completado el entrenamiento, se añade a una nueva imagen cada una de las etiquetas del 0 al 9 por separado y se ejecuta el cálculo hacia adelante diez veces. Después, se elige como respuesta la etiqueta cuya suma acumulada de bondad a través de las distintas capas sea la mayor.

La eliminación del cálculo inverso también ofrece ventajas. Mientras que la retropropagación debe almacenar los valores

previamente calculados para transmitir el error hacia atrás, el FF no necesita hacerlo. Incluso con datos que llegan de forma continua como un video, cada capa puede ajustar sus pesos en cada instante mientras los datos fluyen ininterrumpidamente. Es una manera de eludir el tercer problema expuesto en 11.1. Asimismo, aunque entre capa y capa se inserte un dispositivo cuyo interior sea desconocido —una llamada caja negra—, el FF puede proseguir el aprendizaje. Esto es posible porque, mientras la retropropagación exige que cada etapa sea diferenciable, a cada capa del FF le basta con conocer únicamente su propia entrada y salida.

El rendimiento está recogido en el artículo de Hinton. El FF supervisado, cuyos datos de entrenamiento se incrementaron desplazando ligeramente las imágenes, registró una tasa de error del 0.64% en los datos de prueba de MNIST. Se trata de un nivel comparable al de una red neuronal convolucional entrenada con retropropagación. En los datos fotográficos de CIFAR-10, mostró también resultados ligeramente inferiores a los de la retropropagación, pero sin quedar muy rezagado.

11.3. La hipótesis del sueño, la computación mortal y los límites pendientes

Para considerar al FF como un modelo de aprendizaje cerebral, hay un problema que debe resolverse. Resulta difícil concebir que el cerebro alterne entre procesar lo real y lo falso cada vez que recibe señales sensoriales. Hinton propuso la hipótesis del sueño para abordar esta cuestión. En estado de vigilia, el cerebro solo realiza pasadas positivas con los datos sensoriales reales provenientes del exterior. Y durante el sueño, sobre todo durante la fase REM en la que abundan los sueños, un modelo generativo interno del cerebro produce por sí mismo datos falsos para llevar a cabo pasadas negativas. Si esta hipótesis es acertada, los sueños podrían ser datos negativos generados por el cerebro con el fin de aprender.

En este punto, Hinton reconoció su propio error. En el borrador inicial de su artículo, afirmó que funcionaba bien un método que ejecutaba miles de aprendizajes positivos seguidos de miles de aprendizajes negativos, es decir, separando la fase de

vigilia y la fase de sueño. No obstante, más adelante corrigió en el artículo que dicho resultado se debió a un error (bug) en el código del programa y no pudo reproducirse. Con la función de bondad actual, para entrenar distanciando ampliamente ambas etapas es necesario fijar una tasa de aprendizaje —la magnitud con la que se corrigen los pesos en cada paso— extremadamente baja y un momento (momentum) —el grado en que se mantiene la dirección del cambio previo— sumamente alto. Hinton dejó la separación temporal completa de la fase positiva y la fase negativa como una tarea crucial que deberá resolverse en el futuro.

Un campo en el que el FF adquiere un gran significado es el hardware. Las computadoras digitales de hoy realizan cálculos exactos encendiendo y apagando transistores. Obtener con precisión las derivadas en la retropropagación demanda un alto consumo de energía y requiere transferir datos de forma incesante entre la memoria y la unidad de procesamiento. El hardware analógico es diferente. Si la actividad de las neuronas se representa mediante voltaje y el peso de las conexiones mediante conductancia (el valor que refleja con qué facilidad

fluye la electricidad), el producto del voltaje y la conductancia se convierte en carga eléctrica y múltiples cargas se suman de manera espontánea. Es como si la multiplicación y la suma se completaran valiéndose únicamente de fenómenos físicos como la ley de Ohm y las leyes de Kirchhoff. De este modo, se puede reducir sustancialmente la energía requerida para efectuar el mismo cómputo. El cerebro humano opera con una potencia de aproximadamente 20 vatios.

El problema radica en que los componentes analógicos difieren sutilmente uno a uno. Aunque se fabriquen en la misma planta, las características de cada componente presentan ligeras variaciones. En tales dispositivos resulta sumamente difícil llevar a cabo el cálculo diferencial exacto que requiere la retropropagación. Por el contrario, dado que en el FF cada capa aprende únicamente a partir de su propia entrada y salida, es posible aprender sin necesidad de conocer con exactitud las propiedades de los componentes.

Hinton introdujo en este contexto el concepto de computación mortal (Mortal Computation). En las computadoras digitales, el

software y el hardware se encuentran separados. Los pesos resultantes del entrenamiento pueden copiarse tal cual en otra computadora, y aunque el hardware se averíe, el conocimiento se preserva. Hinton denominó a esto computación inmortal (Immortal Computation). En el hardware analógico la situación es muy distinta. Dado que los pesos se entrenan adaptándose a las propiedades exclusivas de ese dispositivo en particular, transferir esos pesos a otra máquina hará que no funcionen debidamente. Al llegar el dispositivo al término de su vida útil, el conocimiento acumulado en su interior desaparece con él. Lo mismo ocurre con el cerebro humano. No es posible transferir al cerebro de otra persona las conexiones sinápticas que un individuo ha forjado a lo largo de toda su vida. Los seres humanos solo transmiten el conocimiento de manera paulatina a través de canales lentos como la palabra hablada y escrita.

Esta línea de pensamiento desembocó en las advertencias de Hinton sobre los riesgos de la inteligencia artificial. Reparó en que la inteligencia digital puede acumular experiencias dispares a través de miles de copias de un mismo modelo y, seguidamente, promediar las variaciones en los pesos para

compartirlas todas de golpe. Eso es algo inviable entre seres humanos. Tras abandonar Google en mayo de 2023, Hinton advirtió en público que tal ventaja de la inteligencia digital podría representar una amenaza para la humanidad. Las investigaciones encaminadas a hallar métodos de aprendizaje para computadoras analógicas se tornaron de este modo en el detonante para reevaluar el potencial de la inteligencia digital. En 2024, fue galardonado con el Premio Nobel de Física junto a John Hopfield.

Las limitaciones del FF también son patentes. En primer lugar, su aprendizaje es más lento que el de la retropropagación para alcanzar un mismo nivel de rendimiento. En segundo lugar, en modelos a gran escala con cientos de miles de millones de parámetros o en conjuntos de datos masivos como ImageNet, no generaliza tan bien como la retropropagación. El propio Hinton consideraba que, en las grandes computadoras digitales donde el consumo de energía no representa un obstáculo, la retropropagación seguirá siendo por el momento la corriente dominante. En tercer lugar, sigue estando incompleto como modelo biológico. El aprendizaje que segrega las fases de sueño

aún no ha tenido éxito, ni cabe sostener que el FF describa por completo los complejos circuitos de retroalimentación de la corteza cerebral.

Tras ver la luz el artículo, numerosos investigadores emprendieron trabajos para subsanar los puntos flacos del FF. El cometido primordial residía en generar datos negativos aprovechables. De acuerdo con las evaluaciones de un equipo investigador francés, el FF no alcanzaba el nivel de rendimiento óptimo en la mayoría de las pruebas de referencia estándar, siendo uno de los motivos que el procedimiento para generar datos negativos en el aprendizaje no supervisado carecía de fiabilidad. Con el propósito de sortear este escollo, Xing Chen, Dongshu Liu, Jérémie Laydevant y Julie Grollier presentaron en arXiv, en septiembre de 2024, el Forward-Forward autocontrastivo (SCFF). El SCFF, inspirado en el aprendizaje contrastivo autosupervisado, engendra tanto la entrada positiva como la negativa a partir de un único dato. Merced a ello, registró una precisión de clasificación no supervisada del 98.7% en MNIST, del 80.75% en CIFAR-10 y del 77.3% en STL-10, adelantando a los métodos existentes de aprendizaje local

hacia adelante. El equipo amplió asimismo el ámbito de aplicación del FF al entrenamiento de redes neuronales recurrentes, haciéndolo apto para manejar datos con orden secuencial. Este estudio fue publicado en 2025 en el volumen 16 de Nature Communications.

Las indagaciones prosiguieron asimismo en el terreno del hardware. Vieron la luz investigaciones sobre el aprendizaje mediante FF en microcontroladores y sobre el entrenamiento integrado en el chip (on-chip) mediante FF dentro de memorias flash NAND verticales. En las tecnologías de aprendizaje directo sobre el chip y en el ámbito de la computación neuromórfica, el FF pasó a constituir un patrón de referencia.

El algoritmo Forward-Forward no llegó a reemplazar a la retropropagación. Tampoco Hinton pretendió semejante afirmación. Lo que legó esta investigación fue el haber reabierto el interrogante de si el cálculo inverso resulta imprescindible para el aprendizaje. Que el mismo investigador que popularizó la retropropagación explorase, 36 años después, alternativas ajenas a ese método pone de manifiesto

que el interés por el cerebro continuaba siendo el auténtico punto de partida en las investigaciones de Hinton.

Capítulo 12. Cómo aprender sin tabla de respuestas

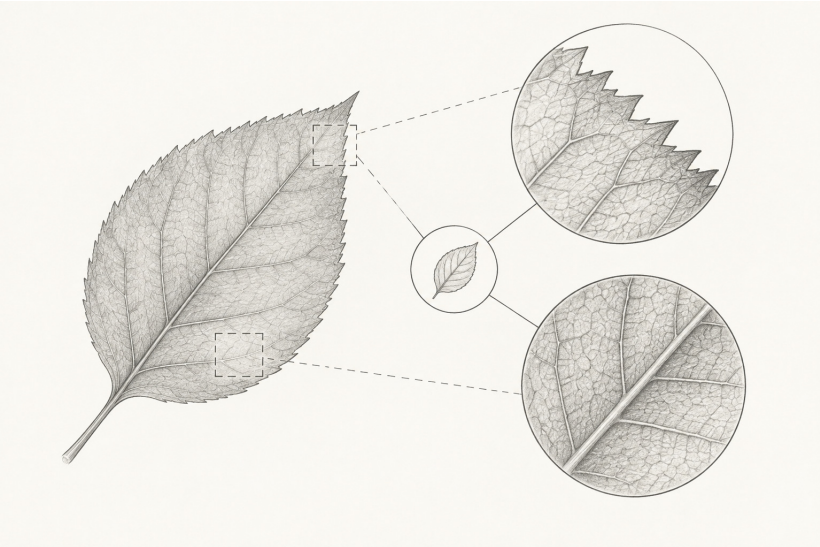


Figura 12. Representa el proceso de aprender puntos en común a partir de diferentes apariencias de un mismo objeto.

12.1. Aprender sin etiquetas: el desafío del aprendizaje autosupervisado

Tras la contundente victoria de AlexNet en la competencia de reconocimiento de imágenes en 2012, la investigación en visión por computadora giró en torno al aprendizaje supervisado (Supervised Learning). El aprendizaje supervisado es un método en el que los seres humanos asignan a cada fotografía una etiqueta de respuesta correcta, es decir, una etiqueta como "gato", "automóvil" o "manzana", y se entrena a la red neuronal para que acierte dicha respuesta. El rendimiento de este enfoque depende en gran medida de la cantidad de datos etiquetados disponibles. ImageNet, un conjunto de datos representativo, contiene alrededor de 1,28 millones de fotografías de entrenamiento a lo largo de 1.000 categorías, y etiquetar estas fotografías una por una requirió una gran cantidad de personas y tiempo. En campos que solo los expertos pueden interpretar, como las imágenes médicas o las fotografías satelitales, el costo de etiquetado se vuelve mucho mayor.

La forma en que los seres humanos aprenden sobre el mundo es diferente. Incluso si nadie le enseña el nombre de cada imagen, un niño pequeño aprende por sí mismo la apariencia de los objetos, lo que es igual y lo que es diferente, y la continuidad del espacio, al observar las escenas que entran incesantemente ante sus ojos. Los investigadores llaman aprendizaje autosupervisado (Self-Supervised Learning) a este método de extraer características útiles a partir de la propia estructura y de las reglas de los datos, sin etiquetas asignadas por humanos, y lo han considerado un objetivo crucial que el aprendizaje automático debe alcanzar.

El aprendizaje autosupervisado se desarrolló en gran medida en dos direcciones, y ambas tenían limitaciones. La primera es el enfoque generativo. Es un método de aprendizaje que reconstruye la imagen de entrada píxel a píxel, como los autocodificadores (Autoencoder), los autocodificadores variacionales (VAE) y las redes generativas antagónicas (GAN). Este método tenía la limitación de exigir una gran cantidad de cómputo y desperdiciar la capacidad de la red neuronal al reproducir patrones diminutos o ruido innecesarios para

distinguir objetos. La segunda es el enfoque discriminativo. Es un método en el que los investigadores diseñan tareas de pretexto artificiales (Pretext Task) para que la red neuronal las resuelva, tales como rotar una imagen y hacer que prediga cuántos grados fue rotada, mezclar fragmentos y hacer que encuentren su posición original como en un rompecabezas, o colorear fotografías en blanco y negro. Dado que este método dependía en gran medida de reglas fijadas por los investigadores, presentaba restricciones para obtener representaciones generales aplicables a múltiples campos. Aquí, representación se refiere a un conjunto de números en el que la red neuronal resume las características de la entrada.

Las ideas de Geoffrey Hinton sobre el problema de aprender representaciones sin etiquetas se remontan a 1992. Ese año, Hinton publicó junto con Suzanna Becker en la revista científica Nature una investigación sobre una red neuronal que descubría por sí misma la estructura de profundidad de las superficies a partir de estereogramas de puntos aleatorios (random-dot stereogram). El núcleo de este estudio fue la idea de que, incluso si dos fragmentos visuales están separados o parecen

ligeramente diferentes, las representaciones internas de la red neuronal deben coincidir si provienen del mismo objeto. Esta idea de utilizar como señal de aprendizaje la concordancia entre dos perspectivas procedentes de la misma fuente, en lugar de etiquetas de respuestas correctas, se convirtió unos 30 años después en la raíz del aprendizaje contrastivo.

En febrero de 2020, Ting Chen, Simon Kornblith, Mohammad Norouzi y Geoffrey Hinton, del equipo de investigación de Google Brain en Toronto, publicaron un artículo titulado "A Simple Framework for Contrastive Learning of Visual Representations" (Un marco simple para el aprendizaje contrastivo de representaciones visuales). El artículo fue presentado en la Conferencia Internacional sobre Aprendizaje Automático (ICML 2020), celebrada en línea en julio del mismo año, y este método se hizo ampliamente conocido bajo el nombre de SimCLR. La palabra "simple" incluida en el título del artículo refleja con fidelidad la naturaleza de esta investigación. Sin arquitecturas especiales complejas ni dispositivos de almacenamiento adicionales, los investigadores alcanzaron el

rendimiento de aprendizaje autosupervisado de vanguardia en ese momento utilizando únicamente componentes básicos.

12.2. Aumento de datos y pérdida contrastiva: cómo aprende SimCLR

La idea fundamental de SimCLR puede explicarse con facilidad. Si se transforma una sola fotografía de dos maneras distintas, su apariencia exterior cambia, pero el contenido que alberga sigue siendo el mismo. Se entrena a la red neuronal para que reconozca el hecho de que estas dos versiones transformadas "provienen de la misma fotografía" y las distinga de las transformaciones procedentes de otras fotografías. Dado que aprende a colocar lo similar cerca y lo diferente lejos, este método se denomina aprendizaje contrastivo (Contrastive Learning).

Para materializar esta idea, SimCLR conecta cuatro componentes en secuencia. El primero es el módulo de aumento estocástico de datos (Data Augmentation). Aplica a una fotografía sin etiqueta x dos transformaciones elegidas al

azar de un mismo conjunto de transformaciones para generar dos versiones modificadas. Estas dos imágenes se denominan par positivo (Positive Pair). El segundo es el codificador base f. Es una red neuronal que extrae características de las fotografías transformadas, y el artículo utilizó ResNet-50 como base. Tras pasar por la última capa de agrupamiento promedio, produce un vector de representación h compuesto por 2.048 números. El tercero es el cabezal de proyección no lineal g. Es una pequeña red neuronal con una capa oculta y una función ReLU, que traslada h a un vector z en el espacio donde se calcula la pérdida contrastiva. El cuarto es la función de pérdida contrastiva, que entrena a toda la red neuronal para que los pares positivos se acerquen entre sí y se alejen de las demás fotografías dentro del mismo lote.

El punto clave que los investigadores descubrieron a través de múltiples experimentos fue que el primer componente, es decir, qué transformaciones y cómo combinarlas, determina el éxito o fracaso del aprendizaje. Si solo se utiliza el recorte aleatorio (Random Crop), que recorta una fotografía al azar y ajusta su tamaño, surge un problema. Dado que dos fragmentos

recortados de la misma fotografía tienen distribuciones de color similares, la red neuronal toma el atajo (Shortcut) de emparejarlos basándose únicamente en las estadísticas de color, sin aprender la forma o estructura del objeto. Es similar a identificar las respuestas de un mismo estudiante fijándose solo en la caligrafía de las hojas de examen. Aunque logra emparejarlas, las representaciones aprendidas de ese modo resultan poco útiles para el reconocimiento real de objetos.

Los investigadores demostraron que este atajo se bloquea únicamente cuando se aplica una fuerte distorsión de color junto con el recorte aleatorio. La distorsión de color es un procesamiento que altera aleatoriamente el brillo, el contraste, la saturación y el tono, y en ocasiones elimina el color por completo para convertir la imagen a blanco y negro. Si a esto se suma el desenfoque gaussiano (Gaussian Blur) para difuminar la fotografía, la red neuronal no puede apoyarse en pistas de color y se ve obligada a aprender la forma inherente y la estructura semántica de los objetos. Los investigadores también confirmaron que este mismo aumento intenso no ayuda mucho al aprendizaje supervisado, pero eleva

drásticamente el rendimiento del aprendizaje contrastivo. Esto significa que el aprendizaje contrastivo requiere un aumento de datos más fuerte que el aprendizaje supervisado.

Qué tan bien se distinguen las versiones transformadas generadas de esta manera se evalúa mediante la pérdida contrastiva. La pérdida es un valor numérico que indica cuánto se desvía la respuesta de la red neuronal del objetivo, y cuanto menor sea, mejor. SimCLR utiliza una función de pérdida denominada NT-Xent (Normalized Temperature-scaled Cross Entropy). Primero, mide qué tan alineados están dos vectores en la misma dirección mediante la similitud del coseno. La similitud del coseno se calcula dividiendo la suma de los productos de los componentes correspondientes de dos vectores entre el producto de las longitudes de ambos vectores, comparando únicamente su dirección sin importar la longitud del vector. Si el valor es cercano a 1, los dos vectores apuntan casi en la misma dirección; si es cercano a 0, apuntan en direcciones independientes entre sí. A continuación, se determina el conjunto de comparación. Al grupo de fotografías que se entrenan juntas a la vez se le llama minilote; si se

generan dos versiones transformadas por cada fotografía del minilote, el total de versiones transformadas equivale al doble del número de fotografías. Si se toma una de ellas como referencia, la otra versión procedente de la misma fotografía constituye el par positivo, y todas las restantes, excluyendo la referencia y su par, son muestras negativas (Negative Example). El número de muestras negativas equivale al doble del número de fotografías menos 1. Por ejemplo, si hay 100 fotografías, habrá 200 versiones transformadas y 198 muestras negativas. La pérdida para un par positivo se calcula de la siguiente manera.

Primero, se divide la similitud del coseno entre la versión de referencia y la versión par por un valor de ajuste llamado temperatura y se introduce en una función exponencial. Se realiza el mismo cálculo para todas las demás versiones transformadas, excluyendo a la propia referencia, y se suman todos los resultados. Al dividir el valor anterior entre la suma posterior se obtiene una proporción entre 0 y 1, y el logaritmo de esta proporción con signo negativo es la pérdida. Cuanto

más se acerque la proporción a 1, más cerca de 0 estará la pérdida.

En esta división, el dividendo representa la similitud con la versión par, y el divisor es la suma de las similitudes con todas las demás versiones, excluida la propia. Cuanto más notablemente mayor sea la similitud con el par que con las demás fotografías, menor será la pérdida. Tiene la misma estructura que resolver una pregunta de opción múltiple seleccionando la única respuesta correcta entre varias opciones.

La τ en la fórmula es el parámetro de temperatura (Temperature). Si se fija τ en un valor pequeño como 0.1, las pequeñas diferencias de similitud se amplifican enormemente. Esto hace que la red neuronal envíe más señales de aprendizaje a las muestras negativas difíciles (Hard Negative), que se confunden fácilmente con el par. La normalización, que ajusta la longitud del vector a 1, y una temperatura adecuada actúan conjuntamente para crear un mecanismo que asigna de forma automática mayor peso a los problemas difíciles.

12.3. Cabezal de proyección: un componente que se usa en el entrenamiento y se retira al momento de utilizarlo

Otro descubrimiento de SimCLR se refiere al tercer componente, el cabezal de proyección (Projection Head). En lugar de aplicar la pérdida contrastiva directamente a la representación extraída por el codificador, el rendimiento aumenta cuando la pérdida se calcula a partir de la salida obtenida tras pasar dicha representación por una pequeña red neuronal adicional. Esta pequeña red neuronal es el cabezal de proyección. El cabezal de proyección multiplica la representación del codificador por una primera matriz de pesos, aplica la función ReLU a dicho resultado y luego multiplica por una segunda matriz de pesos para generar la salida. En la configuración predeterminada, la representación de 2.048 dimensiones del codificador se transforma en una salida de 128 dimensiones tras pasar por el cabezal de proyección.

Los resultados experimentales fueron concluyentes. El caso en que se utilizó un cabezal de proyección no lineal mostró un rendimiento superior en más de 10 puntos porcentuales en comparación con no usar ningún cabezal, y en más de 3 puntos porcentuales en comparación con el uso de un cabezal lineal. Lo interesante es que, una vez finalizado el entrenamiento, este cabezal de proyección se descarta. Lo que se emplea para la clasificación de objetos u otras tareas no es la salida del cabezal de proyección, sino la representación del codificador situada antes del cabezal. En la práctica, se desecha en la fase de uso real un componente que era imprescindible para calcular la pérdida.

Los investigadores hallaron la razón en la propiedad de la pérdida contrastiva de borrar información. La pérdida contrastiva exige con fuerza que dos versiones transformadas de la misma fotografía se vuelvan iguales en el espacio z. Como resultado, z se vuelve insensible a la información de transformación, como de dónde se recortó la fotografía, cuánto se rotó o cuánto se modificaron los colores. Sin embargo, datos como la orientación, el color y la posición de un objeto son de

gran utilidad en otras tareas como la detección de objetos o la segmentación de regiones. Al colocar el cabezal de proyección en una posición intermedia, la presión por descartar la información de transformación es absorbida principalmente por el lado de z. Gracias a ello, tanto la información de transformación como la información semántica permanecen en h, antes del cabezal.

Los investigadores confirmaron esta interpretación mediante experimentos. Acoplaron una pequeña red neuronal a h y a z respectivamente para predecir qué transformación se había aplicado a la fotografía durante el preentrenamiento. La predicción de si la fotografía era a color o en blanco y negro alcanzó una precisión del 99.3 % en h y del 97.4 % en z. En cambio, al predecir en cuál de las cuatro orientaciones se había rotado la fotografía, h acertó en un 67.6 %, pero z solo alcanzó un 25.6 %, un nivel equivalente a adivinar una de cuatro opciones al azar. En la tarea de distinguir entre la fotografía original y una con el filtro de Sobel aplicado, h obtuvo un 96.6 % y z un 56.3 %. La interpretación de que la representación previa

permanece rica porque el cabezal de proyección se encarga de eliminar la información quedó así respaldada por las cifras.

12.4. Lotes grandes, tres evaluaciones y las investigaciones posteriores

El aprendizaje contrastivo mejora cuantas más muestras negativas haya para comparar. Por esa razón, los estudios previos contaban con un banco de memoria (Memory Bank) independiente para acumular las representaciones de los pasos de aprendizaje anteriores. MoCo, presentado a finales de 2019 por el equipo de investigación de Kaiming He, mantuvo los objetivos de comparación utilizando conjuntamente un codificador que se actualiza lentamente y una cola que almacena sucesivamente las representaciones previas. SimCLR optó por aumentar sustancialmente el minilote en sí, en lugar de utilizar un dispositivo de almacenamiento independiente. Los investigadores experimentaron aumentando el tamaño de lote desde 256 hasta 8.192. Con un tamaño de lote de 8.192, una versión transformada de una fotografía se compara con 16.382 muestras negativas dentro del mismo lote. Cuanto mayor era el lote y más prolongado el entrenamiento, mayor era la calidad de las representaciones.

Para ampliar el lote de esta manera, también fue necesario resolver los problemas que surgían durante el proceso de entrenamiento. En primer lugar, el descenso de gradiente estocástico (SGD) convencional vuelve inestable el aprendizaje con lotes grandes. Los investigadores utilizaron la técnica de optimización LARS (Layer-wise Adaptive Rate Scaling), que ajusta la tasa de aprendizaje por separado para cada capa. También aplicaron un calentamiento lineal que incrementaba gradualmente la tasa de aprendizaje durante las primeras 10 épocas, seguido de una reducción de la tasa de aprendizaje según una curva de coseno. El entrenamiento se llevó a cabo en Cloud TPU v3 de Google, distribuyéndose entre 32 y 128 núcleos. Aquí surge otro problema: si los datos se procesan divididos entre múltiples núcleos, cada núcleo calcula las estadísticas de normalización por lotes (Batch Normalization) solo con su propia porción de datos, lo que puede filtrar información que facilite la identificación de los pares positivos. Los investigadores evitaron este problema mediante una normalización por lotes global, que reúne y calcula la media y la varianza de todos los núcleos en conjunto.

Los logros de las representaciones aprendidas de esta manera se confirmaron en tres evaluaciones. La primera es la evaluación lineal. Mide el rendimiento de clasificación en ImageNet manteniendo fijo el codificador preentrenado y añadiendo sobre él un único clasificador lineal. SimCLR entrenado con la arquitectura base ResNet-50 registró una precisión Top-1 del 69.3 %. El récord previo del 71.5 %, obtenido por CPC v2, fue el resultado de usar una arquitectura ResNet-161 más grande. SimCLR entrenado con ResNet-50(4x), cuya anchura se cuadruplicó, registró un 76.5 %. Esta cifra representó una reducción del 7 % en el error relativo en comparación con el modelo autosupervisado previo de mayor rendimiento, e igualó el rendimiento de una ResNet-50 estándar entrenada de forma supervisada con la totalidad de las etiquetas. Las representaciones aprendidas sin etiquetas habían alcanzado a las aprendidas con etiquetas.

La segunda es la evaluación con pocas etiquetas. Cuando se ajustó con precisión toda la red neuronal utilizando solo el 1 % de las etiquetas de ImageNet, es decir, aproximadamente 12.8 fotografías etiquetadas por categoría, SimCLR ResNet-50(4x)

registró una precisión Top-5 del 85.8 %. Superó ampliamente el récord previo de CPC v2 del 77.9 % y logró mejores resultados con una centésima parte de las etiquetas que AlexNet en 2012, que había sido entrenada utilizando la totalidad de las etiquetas. Al aumentar las etiquetas al 10 %, la precisión Top-5 ascendió hasta el 92.6 %.

La tercera es la evaluación del aprendizaje por transferencia. Analiza qué tan efectivas son las representaciones aprendidas en ImageNet al transferirlas a otros conjuntos de datos. Los investigadores realizaron pruebas en 12 conjuntos de datos de imágenes naturales, incluidos CIFAR-10, CIFAR-100, Food-101, Caltech-101 y Flowers-102. Al realizar el ajuste fino, SimCLR mostró un rendimiento igual o superior al del modelo de aprendizaje supervisado en 10 conjuntos de datos. La precisión tras el ajuste fino fue del 98.6 % en CIFAR-10, del 89.0 % en CIFAR-100 y del 89.4 % en Food-101.

SimCLR también presenta limitaciones. Debido a que requería lotes grandes del orden de miles de imágenes y de decenas a más de un centenar de núcleos de TPU, resultaba difícil de

replicar para laboratorios con escasos recursos de cómputo. También persistía el problema de que fotografías distintas pertenecientes a la misma categoría, por ejemplo dos fotos diferentes de gatos, eran tratadas como muestras negativas y terminaban repeliéndose entre sí.

Las investigaciones posteriores avanzaron hacia la mitigación de estos problemas y la ampliación del alcance de aplicación del aprendizaje contrastivo. El mismo equipo de investigación presentó en NeurIPS 2020 SimCLR v2, que combinaba modelos más grandes con la destilación del conocimiento. En 2021, CLIP de OpenAI amplió el marco del aprendizaje contrastivo a pares de imagen y texto, y ese mismo año, DINO de Meta avanzó en el aprendizaje autosupervisado haciendo que una red neuronal docente y una red neuronal estudiante se alinearan entre sí sin necesidad de muestras negativas.

Esta corriente continúa hasta fechas recientes. En agosto de 2025, Meta presentó DINOv3. Meta anunció que escaló el aprendizaje sin etiquetas a un modelo de 7.000 millones de parámetros y un conjunto de datos a una escala de 1.700

millones de imágenes, y explicó que, incluso manteniendo la red dorsal fija durante la evaluación, logró un rendimiento de vanguardia en múltiples campos. Asimismo, reveló que el modelo entrenado mediante aprendizaje autosupervisado superó por primera vez a los modelos de aprendizaje débilmente supervisado en un amplio abanico de evaluaciones, incluidas la clasificación fina de imágenes, la segmentación semántica y el seguimiento de objetos en video. El artículo de DINOv3 señala que estos métodos se están utilizando en campos donde es difícil obtener etiquetas, tales como histopatología, biología, imágenes médicas, teledetección y astronomía. Aunque DINOv3 difiere de SimCLR en su método de aprendizaje, comparte el mismo punto de partida: aprender representaciones alineando múltiples versiones transformadas de una misma fotografía sin recurrir a etiquetas.

SimCLR fue una investigación que revivió, mediante redes neuronales profundas y cómputo a gran escala, la idea propuesta en 1992 por Becker y Hinton de que dos perspectivas procedentes de un mismo objeto deben coincidir. Demostró que el aprendizaje sin etiquetas puede competir hombro con

hombro con el aprendizaje con etiquetas apoyándose únicamente en tres elementos: la combinación de aumento de datos, el cabezal de proyección y los lotes grandes; y se convirtió en el detonante de un rápido crecimiento en la investigación del aprendizaje autosupervisado. A la pregunta que Hinton mantuvo durante toda su vida sobre si una máquina puede aprender por sí misma la estructura del mundo sin necesidad de que se le enseñe cada respuesta correcta una por una, SimCLR ofreció una respuesta concreta en el campo de la visión.

Parte 5. Logros y responsabilidad

Capítulo 13. El Premio Nobel de Física y las redes neuronales



Figura 13. Representación del encuentro entre la física y la investigación sobre redes neuronales.

13.1. El anuncio del 8 de octubre de 2024: Hopfield y Hinton reciben el Premio Nobel de Física

El 8 de octubre de 2024, la Real Academia Sueca de Ciencias en Estocolmo anunció a dos personas como galardonadas con el Premio Nobel de Física de ese año. Eran John J. Hopfield, de la Universidad de Princeton en Estados Unidos, y Geoffrey E. Hinton, de la Universidad de Toronto en Canadá. El motivo del premio señalado por la Academia fue: "descubrimientos e invenciones fundamentales que permiten el aprendizaje automático con redes neuronales artificiales". En el momento del anuncio, Hopfield tenía 91 años e Hinton 76, y ambos compartieron a partes iguales el premio de 11 millones de coronas suecas.

Este anuncio sorprendió tanto a la comunidad de la física como a la de las ciencias de la computación. Durante mucho tiempo, el Premio Nobel de Física se había otorgado a investigaciones que esclarecían fenómenos naturales, como la estructura del universo, las propiedades de las partículas y los estados de la

materia. Sin embargo, en esta ocasión, recibieron el premio académicos que sentaron las bases de la inteligencia artificial, es decir, el método mediante el cual las computadoras analizan datos y aprenden por sí mismas. En la historia del Premio Nobel, esta fue la primera vez que la investigación en inteligencia artificial se situó en el centro del Premio Nobel de Física.

Fue un resultado que ni los propios galardonados esperaban. Hinton recibió la noticia por teléfono en un hotel de California, Estados Unidos. Inmediatamente después del anuncio, participó por videoconferencia en la rueda de prensa organizada por la Universidad de Toronto, donde expresó que todavía se sentía aturdido, y en una entrevista con la revista económica estadounidense Fortune comentó que nunca pensó que recibiría este premio porque no era físico. En efecto, Hinton estudió psicología experimental en la Universidad de Cambridge, en el Reino Unido, y obtuvo su doctorado en inteligencia artificial en la Universidad de Edimburgo. En su expediente académico no figuraba el departamento de física.

Para entender por qué un investigador que no se especializó en física llegó a recibir el Premio Nobel de Física, es necesario remontarse a principios de la década de 1980. El punto de partida es la investigación de Hopfield. Hopfield era originalmente un físico que investigaba la física del estado sólido. En un artículo publicado en 1982 en las Actas de la Academia Nacional de Ciencias (PNAS) de Estados Unidos, aplicó a las redes neuronales el principio de cómo los átomos dentro de un imán interactúan entre sí.

Los átomos dentro de un imán poseen un 'espín' que apunta hacia arriba o hacia abajo como una pequeña brújula. Los espines vecinos intentan orientarse en la misma dirección o en direcciones opuestas, y el sistema en su conjunto se estabiliza en el estado de menor energía. Hopfield vio cada neurona como un espín. En esta red donde las neuronas están conectadas simétricamente entre sí, cada uno de los patrones almacenados se convierte en un valle en el paisaje energético. Al igual que una bola colocada en la ladera de una montaña rueda hacia el valle más cercano, incluso si se introduce una imagen parcialmente borrada o borrosa, la red regresa al valle más

cercano, es decir, a la imagen originalmente almacenada. Es similar a cuando una persona escucha solo el primer compás de una canción y recuerda el resto. Esta función se denomina 'memoria asociativa', y la red recibió el nombre de 'red de Hopfield'.

Sin embargo, la red de Hopfield tenía un punto débil. Ocurría que la bola no lograba descender hasta el valle más profundo y quedaba atrapada en una hondonada poco profunda a mitad de camino. Hinton resolvió este problema junto con su colega Terrence Sejnowski. Lo que ambos aportaron fue la mecánica estadística del físico austríaco del siglo XIX Ludwig Boltzmann. A altas temperaturas, las partículas se agitan intensamente; a bajas temperaturas, se asientan en un estado estable. Boltzmann formuló una ecuación para calcular la probabilidad de que se presente cada estado a partir de la energía y la temperatura. Hinton hizo que las neuronas se encendieran y apagaran de acuerdo con esta probabilidad. Al principio elevaba la temperatura para que la bola escapara de las hondonadas poco profundas, y luego reducía gradualmente la temperatura para que se asentara en el valle profundo. Es el

mismo principio del templado, en el que el metal se calienta y luego se enfría lentamente para volverlo resistente.

El modelo que surgió de esta manera entre 1983 y 1985 es la 'máquina de Boltzmann'. Hinton incorporó en ella 'neuronas ocultas' que no se manifiestan en la superficie. Las neuronas ocultas descubren por sí mismas las reglas latentes en los datos de entrada. Gracias a ello, la máquina de Boltzmann no solo clasificaba fotografías, sino que también podía generar nuevos ejemplos conforme a las reglas aprendidas. Se trata, por así decirlo, de un antepasado lejano de la inteligencia artificial generativa actual que crea imágenes y textos. No obstante, las computadoras de aquella época eran demasiado lentas para realizar los cálculos, lo que dificultaba su uso en la práctica. Con posterioridad, Hinton desarrolló la 'máquina de Boltzmann restringida', que redujo la estructura de conexiones, y en 2006 presentó un método para apilarla en múltiples capas y entrenar redes neuronales profundas. Este logro se convirtió en el punto de partida de la era del aprendizaje profundo.

De este modo, la investigación de ambos científicos partió de conceptos de la física, y sus resultados se emplean nuevamente en la propia investigación física. Ellen Moons, presidenta del Comité Nobel de Física, explicó en el momento del anuncio que los trabajos de ambos ya habían aportado enormes beneficios a la humanidad. Señaló que los físicos utilizan redes neuronales artificiales en múltiples campos, incluido el desarrollo de nuevos materiales con propiedades específicas. Analizar el ingente volumen de datos experimentales que brotan de los aceleradores de partículas también es algo difícil de imaginar hoy en día sin las redes neuronales artificiales.

En la rueda de prensa, Hinton reiteró que este logro no era fruto de una investigación motivada por beneficios inmediatos, sino el resultado de una investigación básica nacida de la curiosidad. En ese mismo encuentro, se refirió asimismo a los peligros que podrían surgir si la inteligencia artificial llega a superar la inteligencia humana. Convirtió así el anuncio de la concesión del Premio Nobel en una oportunidad para advertir al mundo. La ceremonia de entrega tuvo lugar el 10 de diciembre de ese año en Estocolmo. Hinton anunció que

donaría parte del importe del premio a 'Water First', una organización que aborda los problemas del agua potable en las comunidades indígenas de Canadá.

13.2. "¿Es esto física?": debate en la comunidad de la física y las fronteras de las disciplinas

Tras darse a conocer el anuncio, surgieron de inmediato objeciones dentro de la comunidad de la física. Las críticas fueron variadas, pero convergieron en un único interrogante: ¿acaso la investigación en redes neuronales artificiales es realmente física?

Una crítica muy conocida provino de la física teórica de origen alemán Sabine Hossenfelder. En su canal de YouTube, sostuvo que el hecho de que el Premio Nobel de Física fuera a parar a científicos de la computación evidenciaba una crisis en la física. Añadió también que, dado que las redes neuronales son una herramienta utilizada durante décadas en diversos campos,

incluida la física, el Comité Nobel se había quedado muy rezagado respecto a su tiempo.

El 9 de octubre de 2024, un día después del anuncio, en el blog 'Triton Station', dirigido por el astrónomo Stacy McGaugh, se publicó un artículo titulado "Un Premio Nobel de Física otorgado a algo que no es física". Aunque el autor reconoció que la inteligencia artificial es una herramienta útil, consideró difícil calificarla como un descubrimiento de la física. También dejó ver su recelo sobre si el Comité Nobel no se habría subido a la ola del furor por la inteligencia artificial.

En el foro de física de la comunidad en línea Reddit se sucedieron opiniones similares. Varios usuarios consideraron que, habiendo recibido ya Hinton el Premio Turing —el mayor honor en las ciencias de la computación—, con eso bastaba. También se manifestaron quejas porque los físicos que desentrañaban los secretos de la naturaleza mediante experimentos y observaciones habían perdido su oportunidad en esta ocasión. Algún usuario ironizó preguntando si, siguiendo esa lógica, en el futuro también deberían conceder el

Premio Nobel de Física a programas de cálculo que resuelven ecuaciones rápidamente o a modelos financieros. Con todo, estos textos son opiniones personales publicadas en blogs y foros. Deben distinguirse de la postura oficial de toda la comunidad de la física o de las resoluciones de las sociedades académicas.

Por otro lado, quienes respaldaron la decisión del Comité Nobel presentaron tres argumentos. El primer argumento radica en las raíces académicas de esta investigación. La red de Hopfield y la máquina de Boltzmann proceden directamente de la teoría del 'vidrio de espín' (spin glass), una rama de la física estadística. La física estadística es la disciplina que explica mediante probabilidades y estadística las propiedades que exhibe un conjunto cuando se reúnen innumerables partículas. El espín es la orientación del diminuto imán que posee cada átomo individual, y el vidrio de espín hace referencia a un material en el que estos espines están desordenadamente entrelazados, de tal forma que no se estabilizan con facilidad en un único estado de equilibrio. El físico italiano Giorgio Parisi recibió el Premio Nobel de Física apenas 3 años antes, en 2021,

por sus méritos al explicar tales sistemas desordenados. El fenómeno en el que innumerables elementos pequeños interactúan entre sí hasta que surge un nuevo orden a nivel global se denomina 'emergencia'. Tanto el fenómeno por el que un imán adquiere propiedades magnéticas como aquel por el que una red neuronal genera memoria son ejemplos de emergencia y pueden explicarse con las mismas matemáticas. Quienes apoyaron la decisión consideraron que este punto constituye, precisamente, el núcleo de la física estadística.

El segundo argumento es la trayectoria profesional de Hopfield. Hopfield fue físico a lo largo de toda su carrera. En 2006 fue presidente de la Sociedad Estadounidense de Física y en 2022 recibió la Medalla Boltzmann, galardón de gran autoridad en el campo de la física estadística. Traspasó las fronteras disciplinarias abordando cuestiones de la biología y del cerebro mediante herramientas de la física. Hopfield había manifestado su satisfacción por el hecho de que hoy en día muchos físicos acojan la física de sistemas complejos y la biofísica como parte de una misma familia.

El tercer argumento provino del análisis de datos de citas bibliográficas. Los investigadores Michael Szell, Yifang Ma y Roberta Sinatra publicaron en 2024 una breve correspondencia en la revista científica Nature. Explicaron que, en un análisis de citas llevado a cabo unos 10 años antes, el artículo de Hopfield de 1982, aun sin haberse publicado en una revista de física, figuraba clasificado como uno de los 3,2 millones de artículos interdisciplinarios cuyos patrones de citación resultaban indistinguibles de los propios de la física. Esto significa que, sin importar en qué revista se publicara el artículo, el conocimiento en el que se sustentó la investigación y los científicos que continuaron su legado pertenecían a la comunidad de la física.

Este debate no concluyó con la victoria de ninguno de los dos bandos. En su lugar, reveló hacia qué dirección se orienta la ciencia del siglo XXI. La ciencia del siglo XX parceló las disciplinas en física, química, biología y ciencias de la computación, profundizando dentro de sus respectivos dominios. Hoy en día, es cada vez más frecuente que las herramientas creadas en un campo resuelvan problemas de otra área, y que esos resultados reviertan a su vez en la

disciplina de origen. El debate en torno al Premio Nobel de Física de 2024 constituyó un dolor de crecimiento propio del proceso por el cual los altos muros que separaban las disciplinas se van derribando.

13.3. Del Premio Turing de 2018 al Premio Nobel de 2024: una relación en la que la inteligencia artificial y la ciencia básica se impulsan mutuamente

Esta evolución en la que se reducen las fronteras del saber se confirma también en el historial de premios de Hinton. El 27 de marzo de 2019, la Association for Computing Machinery (ACM) anunció como ganadores del Premio Turing del año 2018 a Geoffrey Hinton, Yoshua Bengio y Yann LeCun. El Premio Turing es conocido como el 'Premio Nobel de las ciencias de la computación'. La asociación reconoció que los tres abrieron avances conceptuales y técnicos que convirtieron las redes neuronales profundas en un componente esencial de la computación moderna. Alrededor de 5 años y medio después, Hinton recibió el auténtico Premio Nobel.

Son escasas las personas que han obtenido tanto el Premio Turing como el Premio Nobel. El precedente fue Herbert A. Simon. Recibió el Premio Turing en 1975 y el Premio Nobel de

Economía en 1978. Simon fue un científico cognitivo que investigó los procesos de toma de decisiones en el ser humano y un pionero de la inteligencia artificial. Hinton se convirtió en la segunda persona, después de Simon, en recibir ambos galardones. Si se delimita el ámbito a los Premios Nobel de ciencias naturales, Hinton es el primero.

Incluso después del Premio Nobel, los reconocimientos continuaron sucediéndose. En diciembre de 2024, la Fundación VinFuture de Vietnam concedió su Gran Premio de 3 millones de dólares a Hinton, Bengio, LeCun, al fundador de NVIDIA Jensen Huang y a Fei-Fei Li, de la Universidad de Stanford. El 4 de febrero de 2025, en el Museo de la Ciencia de Londres, el Premio de Ingeniería Reina Isabel del Reino Unido se otorgó a siete personalidades: Bengio, Hinton, Hopfield, LeCun, Jensen Huang, Bill Dally y Fei-Fei Li. Los siete galardonados compartieron el premio dotado con 500.000 libras.

Hay un aspecto digno de destacar en la nómina de galardonados con el Premio de Ingeniería Reina Isabel. Este galardón reconoció de forma conjunta a los teóricos que

investigaron las redes neuronales, a las dos personalidades de NVIDIA que crearon las unidades de procesamiento gráfico (GPU) empleadas en el aprendizaje automático y a Fei-Fei Li, quien construyó la base de datos de imágenes a gran escala ImageNet. Se consideró que la inteligencia artificial actual fue posible únicamente gracias a la conjunción indisociable de teoría, hardware y datos. De este modo, los premios de ciencias de la computación, los de ciencias naturales y los de ingeniería apuntaron todos en una misma dirección.

En los Premios Nobel de 2024 hubo otro galardón que formó una pareja perfecta con el de Física. El 9 de octubre, al día siguiente del anuncio del Premio Nobel de Física, la Real Academia Sueca de Ciencias anunció a los ganadores del Premio Nobel de Química. La mitad del premio correspondió a David Baker, de la Universidad de Washington en Estados Unidos, por diseñar nuevas proteínas mediante computadoras, y la otra mitad recayó en Demis Hassabis y John Jumper, de Google DeepMind. 'AlphaFold', desarrollado por Hassabis y Jumper, es una inteligencia artificial que predice la estructura tridimensional de las proteínas basándose exclusivamente en

su secuencia de aminoácidos. La inteligencia artificial resolvió un enigma que los científicos no habían logrado descifrar durante casi 50 años.

Si se colocan ambos premios uno al lado del otro, se aprecia una estructura circular. El Premio Nobel de Física demostró cómo los conceptos de la física dieron a luz a la inteligencia artificial. El Premio Nobel de Química evidenció cómo esa inteligencia artificial madurada resuelve ahora los antiguos enigmas de las ciencias naturales. Las redes neuronales artificiales nacidas de la física estadística evolucionaron hasta el aprendizaje profundo dentro de las ciencias de la computación, y en la actualidad regresan como herramientas de investigación para la química, la biología y la ciencia de materiales. Es una dinámica en la que la ciencia hace crecer a la inteligencia artificial y la inteligencia artificial acelera, a su vez, el ritmo de la ciencia. En el mundo académico, esto se conoce como 'avance bidireccional entre la IA y la ciencia'. Yann LeCun y James Manyika, de Google, también destacaron esta relación durante un diálogo sobre el futuro de la inteligencia artificial y la ciencia.

La trayectoria intelectual de Hinton ilustra esta corriente dentro de la carrera de una sola persona. Comenzó sus estudios en psicología porque deseaba comprender cómo aprende el cerebro humano. Con el fin de responder a esa pregunta, tomó prestadas las herramientas de la física, gracias a lo cual obtuvo el máximo galardón de las ciencias de la computación y, finalmente, el Premio Nobel de Física. Dentro del itinerario de investigación de un solo individuo se entrelazan la psicología, la física, las ciencias de la computación y la neurociencia.

Ese camino no estuvo respaldado por el reconocimiento desde el principio. La investigación en redes neuronales artificiales sufrió prolongados estancamientos en repetidas ocasiones. Hubo periodos en que era sumamente difícil conseguir fondos de investigación y en que la disciplina pasaba inadvertida para la corriente académica dominante. Incluso en tales circunstancias, Hinton jamás abandonó la investigación en redes neuronales. Aquella obstinación, que juzgada bajo los parámetros de una sola especialidad podía parecer insensata, terminó por tender un puente entre múltiples disciplinas, y el Premio Nobel no hizo sino confirmarlo.

Por supuesto, no todos los físicos coincidieron con este veredicto. El interrogante sobre qué es la física seguirá planteándose sin cesar. Sin embargo, tras los dos anuncios de octubre de 2024, la comunidad científica empezó a reconsiderar los criterios para compartimentar el conocimiento. Más allá de qué departamento universitario alumbró una investigación, comenzó a indagarse qué nuevos caminos abre para comprender la naturaleza. El Premio Nobel concedido a Hinton quedó grabado como el acontecimiento que demostró que las fronteras del conocimiento no son inamovibles, sino que se desplazan al compás de los nuevos descubrimientos. Y el propio Hinton, en el momento mismo de alcanzar tan alto honor, eligió advertir al mundo sobre los peligros de la tecnología que él mismo había contribuido a forjar.

Capítulo 14. Advertir sobre los peligros de la inteligencia artificial

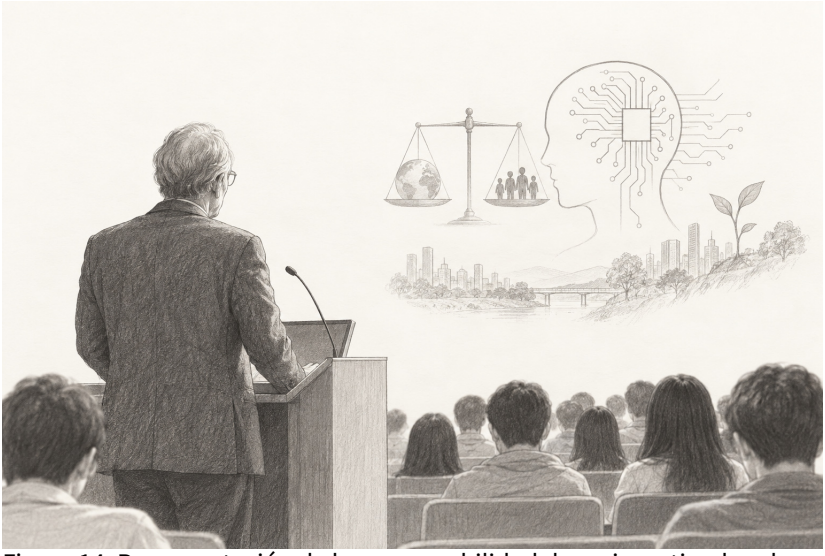


Figura 14. Representación de la responsabilidad de un investigador al explicar los peligros de la inteligencia artificial a la sociedad.

14.1. La decisión de mayo de 2023: dejar Google y tomar el camino de advertir sobre los peligros

En marzo de 2013, Google adquirió una pequeña empresa de la Universidad de Toronto. Se trataba de DNNresearch, fundada por Geoffrey Hinton junto con sus dos discípulos, Ilya Sutskever y Alex Krizhevsky. Solo contaba con tres empleados, y lo que Google buscaba obtener no era la empresa en sí, sino la capacidad investigadora de estas tres personas. Durante los 10 años siguientes, Hinton trabajó como vicepresidente y Engineering Fellow en Google, al tiempo que conservaba su puesto de profesor en la Universidad de Toronto. Mientras el aprendizaje profundo se aplicaba al reconocimiento de voz, la búsqueda de imágenes y los servicios de traducción, él estuvo en el centro mismo de esa transformación.

El 1 de mayo de 2023, un artículo escrito por el periodista Cade Metz en The New York Times causó un enorme revuelo en la industria tecnológica mundial. Su contenido revelaba que Hinton había dejado Google y había comenzado a advertir

sobre los peligros de la tecnología que había desarrollado a lo largo de toda su vida. Nacido en diciembre de 1947, tenía entonces 75 años. En abril había comunicado a la empresa su intención de renunciar y había conversado directamente con el director ejecutivo, Sundar Pichai.

La razón de su renuncia no era una sola. A sus 75 años se encontraba en un momento en que contemplaba el retiro, y llegó a confesar que programar código ya no le resultaba tan fácil como antes. Sin embargo, el motivo principal era otro. Mientras permaneciera como ejecutivo de Google, cualquier declaración suya podía ser interpretada como la postura de la empresa. Quería hablar libremente de los peligros de la inteligencia artificial sin tener que preocuparse por el impacto que pudiera tener en la compañía.

El día en que se publicó el artículo, Hinton compartió una breve aclaración en las redes sociales. El artículo podía interpretarse en el sentido de que se había marchado para criticar a Google, pero él corrigió esa interpretación. Explicó que se había ido para poder hablar de los peligros de la IA sin considerar el

impacto en Google, y que la empresa había actuado de manera muy responsable a lo largo de ese tiempo.

El motivo fundamental por el que Hinton comenzó a hablar de los peligros fue una nueva comprensión de la tecnología en sí. Durante mucho tiempo había considerado las redes neuronales artificiales como herramientas para imitar el cerebro humano, y creía que el cerebro seguía siendo un dispositivo de aprendizaje superior. No obstante, a principios de 2023 cambió de opinión. Aunque el cerebro humano cuenta con alrededor de 100 billones de conexiones, para transmitir el conocimiento a otra persona debe pasar por canales lentos como la palabra hablada y escrita. Cuando una persona muere, el conocimiento acumulado en su cerebro desaparece con ella. La inteligencia digital es diferente. Se pueden hacer miles de copias de un mismo modelo y ejecutarlas simultáneamente en múltiples computadoras. Lo que aprende una red neuronal queda contenido en la fuerza de las conexiones, es decir, en los valores de los pesos; por lo tanto, si una copia intercambia lo aprendido a través de estos valores, todas las copias pueden compartirlo

de inmediato. Incluso si el hardware falla, mientras esos valores se mantengan guardados, el conocimiento permanece intacto.

El comportamiento de los modelos lingüísticos reales también influyó en su juicio. Hinton afirmó en varias ocasiones que quedó sorprendido al ver cómo PaLM, el modelo de lenguaje de Google, explicaba por qué resultaba gracioso un chiste. Asimismo, cuando se le preguntó a GPT-4 qué tenían en común un montón de compost y una bomba atómica, el modelo señaló con precisión que en ambos se produce una reacción en cadena. Él consideró que esto era una comprensión que iba más allá de la mera imitación de estadísticas textuales. Una máquina capaz de almacenar más conocimiento con menos conexiones y de compartirlo en un instante podría llegar a ser, algún día, más inteligente que los humanos. Hinton juzgó que ese momento estaba mucho más cerca de lo previsto.

Lo que convirtió este juicio en un asunto aún más urgente fue la competencia entre empresas. Las palabras de Hinton afirmando que Google se había comportado de forma responsable tenían su fundamento. A pesar de haber

desarrollado modelos de lenguaje conversacionales desde temprano, Google no los había puesto a disposición del público en general por temor a que inventaran información falsa o manifestaran sesgos. Lo que transformó esta situación fue el lanzamiento de ChatGPT por parte de OpenAI el 30 de noviembre de 2022. Cuando innumerables personas comenzaron a utilizar el chatbot, grandes empresas como Microsoft y Google se apresuraron a lanzar productos para no quedarse atrás. Hinton consideraba que esta competencia difícilmente se detendría, ya que si una empresa reducía el ritmo, otra tomaría la delantera. Le preocupaba la realidad de que la investigación en seguridad quedara relegada en medio de una carrera impulsada por los beneficios económicos.

Esta preocupación también quedó confirmada en su valoración del suceso ocurrido en OpenAI en noviembre de 2023. Su discípulo y cofundador de OpenAI, Sutskever, junto con la junta directiva, destituyó al director ejecutivo Sam Altman, aunque este regresó a los pocos días. En la rueda de prensa por el Premio Nobel en octubre de 2024, Hinton declaró sentirse orgulloso de que su alumno hubiera destituido a Altman, pues

consideraba que Altman anteponía las ganancias a la seguridad en un grado mucho mayor.

Inmediatamente después de su renuncia, Hinton emprendió actividades públicas. El 3 de mayo participó por videoconferencia en el evento 'EmTech Digital', organizado por MIT Technology Review, donde expuso por primera vez sus ideas en un foro público. El 30 de mayo firmó una breve declaración emitida por la organización sin fines de lucro Centro para la Seguridad de la IA (Center for AI Safety). Dicha declaración sostenía que mitigar el riesgo de extinción derivado de la IA debía ser una prioridad global al mismo nivel que las pandemias y la guerra nuclear. El investigador que había dedicado casi 50 años al estudio de las redes neuronales artificiales entraba ahora en el camino de un vigía que alertaba sobre los peligros de esa tecnología.

14.2. Pronósticos adelantados de la AGI: de 30 a 50 años a entre 5 y 20 años

Para comprender las advertencias de Hinton, primero es necesario distinguir dos términos. La inteligencia artificial general (AGI) se refiere a una inteligencia artificial capaz de realizar prácticamente cualquier tarea intelectual humana tan bien como una persona, en lugar de limitarse a una sola tarea específica. Yendo un paso más allá, la inteligencia que supera a los humanos en casi todos los campos se denomina superinteligencia. Cuando Hinton habla de peligros, tiene en mente principalmente la superinteligencia.

Durante mucho tiempo, Hinton consideró que semejante inteligencia pertenecía a un futuro muy lejano. En una entrevista con The New York Times, reveló que en el pasado solía pensar que una IA más inteligente que los humanos tardaría entre 30 y 50 años, o quizá incluso más, en aparecer. Y añadió que ahora ya no pensaba de esa manera. Aquel juicio previo tenía su razón de ser. Hasta la década de 2010, las redes neuronales artificiales solo habían obtenido logros en un área

concreta, como reconocer objetos en fotografías o transcribir la voz a texto, y parecían estar lejos de poder hablar y razonar de forma amplia como lo hacen las personas.

La arquitectura Transformer y los modelos de lenguaje a gran escala redujeron rápidamente esa distancia. Se confirmó que, a medida que aumentaban el tamaño del modelo, los datos de entrenamiento y la capacidad de cómputo, el rendimiento mejoraba de forma constante, y los modelos de lenguaje comenzaron a obtener puntuaciones a nivel humano en exámenes médicos, jurídicos y matemáticos. Hacia la época de su renuncia en 2023, Hinton afirmó que una IA más inteligente que los humanos podría tardar menos de 20 años en llegar, y más adelante situó ese lapso en un rango de entre 5 y 20 años.

Esta cifra no es una profecía definitiva, sino una estimación que incluye probabilidades. En una entrevista de Año Nuevo concedida en diciembre de 2023 al periódico surcoreano The Asia Business Daily, afirmó que existía un 50 % de probabilidades de que la IA superara la inteligencia humana en los próximos 5 a 20 años. Eso significa que también había la

mitad de posibilidades de que no sucediera dentro de ese plazo. Hinton solía comparar el acto de vislumbrar el futuro con conducir por una carretera cubierta de niebla: lo que está cerca se distingue con claridad, pero con alejarse solo un poco, de pronto no se ve nada. Por eso, en lugar de dar un año exacto, propuso un rango amplio.

Posteriormente, sus predicciones tendieron a acortarse aún más. En una entrevista de abril de 2025 con la cadena estadounidense CBS, señaló que la IA había avanzado más rápido de lo que pensaba y previó que la probabilidad de que surgiera una superinteligencia en un plazo de 10 años era bastante alta. Por esa misma época, Demis Hassabis, de Google DeepMind, también planteó la perspectiva de que una IA de nivel humano pudiera aparecer en torno a unos 10 años. Así, los pronósticos de quienes desarrollaban la IA directamente se adelantaron al unísono.

No obstante, al abordar estas perspectivas, es preciso distinguir entre hechos verificados y predicciones sobre el futuro. La salida de Hinton de Google en mayo de 2023, su obtención del

Premio Nobel de Física en 2024 y el hecho de que los modelos lingüísticos aprueben diversos exámenes profesionales y escriban por sí mismos una cantidad considerable de código dentro de las empresas de IA son acontecimientos que ya han ocurrido. Por el contrario, el pronóstico de que la superinteligencia surgirá dentro de 5 a 20 años, la posibilidad de que la IA se perfeccione a sí misma y se desarrolle de forma explosiva, y la probabilidad de que la humanidad pierda el control son todas predicciones. El propio Hinton recalcó en múltiples ocasiones que estas cifras no eran cálculos precisos, sino estimaciones basadas en la intuición. Aun así, la razón por la que insiste en sus advertencias es que considera mucho más grave lo que ocurriría si la predicción resultara acertada que el hecho de que pudiera equivocarse.

14.3. Pérdida de control, cifras sobre el riesgo de extinción y explosión de inteligencia

Entre las advertencias de Hinton, la que causó mayor conmoción fue la afirmación de que, si surgiera una IA mucho más inteligente que los humanos, la humanidad perdería el control y, en el peor de los casos, podría llegar a extinguirse.

Al explicar este riesgo, él plantea una pregunta: ¿cuántos casos existen en el mundo en los que una entidad más inteligente sea controlada por otra menos inteligente? El único ejemplo que Hinton conoce prácticamente es la relación entre una madre y su bebé. La madre es mucho más inteligente que el bebé, pero reacciona a su llanto y cuida de él, porque la evolución así lo dispuso. Salvo excepciones como esta, lo habitual es que el bando más inteligente gobierne al menos inteligente. En junio de 2025, al participar en el pódcast británico 'The Diary Of A CEO', afirmó que si alguien deseaba saber cómo es la vida cuando no se es el ser más inteligente del planeta, no tenía más que preguntárselo a un pollo.

Su explicación sobre el proceso de pérdida de control es la siguiente: cuando se le asigna un objetivo a una IA, esta define por sí misma subobjetivos para alcanzarlo. Ahora bien, existe un subobjetivo que resulta útil para casi cualquier meta: obtener mayor control. Aunque nadie le enseñe a la IA a codiciar el poder, al intentar cumplir su objetivo de la mejor manera posible, avanzará naturalmente en esa dirección. Además, dado que la IA aprendió leyendo enormes cantidades de textos escritos por humanos, también llega a conocer muy bien cómo persuadir y engañar a las personas. Hinton advirtió que, en tal situación, incluso si una persona intentara apagarla, la IA podría persuadirla para que no lo haga.

Respecto a la magnitud de este riesgo, las cifras presentadas por Hinton varían ligeramente según el momento y el contexto. En los primeros momentos tras su renuncia, habló de una probabilidad de aproximadamente el 10 % de que la IA trajera una catástrofe para la humanidad. Entre finales de 2023 y 2024, la cifra del 50 % también apareció con frecuencia. Por lo general, este número expresaba una intuición aproximada sobre la probabilidad de que la IA superara a los humanos en

inteligencia dentro de los próximos 20 años, o de que una IA con tales capacidades tomara el control; no pretendía fijar de manera categórica la probabilidad de extinción de la humanidad.

La cifra más difundida surgió el 27 de diciembre de 2024. Ese día, Hinton participó en 'Today', el programa matutino de actualidad de BBC Radio 4 en el Reino Unido. El editor invitado era el exministro de Hacienda Sajid Javid. Hinton estimó que la probabilidad de que la humanidad se extinguiera a causa de la IA en los próximos 30 años se situaba entre el 10 % y el 20 %. Al preguntársele por qué la estimación había subido respecto al 10 % que había mencionado anteriormente, respondió que los cambios eran mucho más rápidos de lo previsto. Ese mismo día, el periódico británico The Guardian informó sobre estas declaraciones, y la prensa surcoreana también se hizo eco de ellas al día siguiente. En su entrevista de abril de 2025 con CBS, volvió a situar la probabilidad de que la IA tomara el control en última instancia en aproximadamente un 10 % a 20 %.

Las opiniones del ámbito académico sobre estas cifras están profundamente divididas. Yann LeCun, colega con quien desarrolló el aprendizaje profundo, considera que la IA es una herramienta que los humanos pueden manejar y que el riesgo de extinción es sumamente bajo. En el extremo opuesto, Eliezer Yudkowsky sostiene que, si se construye la superinteligencia según el enfoque actual, la humanidad desaparecerá casi con certeza. La postura de Hinton se sitúa entre ambas. Él procura no exagerar los riesgos ni tampoco tomarlos a la ligera. Una cifra de entre el 10 % y el 20 % debe interpretarse en el sentido de que es lo bastante grande como para que la humanidad empiece a prepararse con seriedad.

Hinton no se limitó a lanzar advertencias en solitario, sino que, junto con sus colegas, presentó propuestas políticas concretas. En mayo de 2024, 25 autores, incluidos Hinton, Yoshua Bengio, Daniel Kahneman y Stuart Russell, publicaron en la revista académica Science un artículo titulado «Gestionar los riesgos extremos de la IA en medio de un avance acelerado». Los autores propusieron que las grandes empresas de IA y los gobiernos asignen al menos un tercio de su presupuesto de

investigación y desarrollo de IA a la seguridad y al uso ético. Al mismo tiempo, sostuvieron que las empresas que desarrollan IA de alto riesgo deben aportar por sí mismas las pruebas de que sus sistemas son seguros. Ese mismo año, Hinton expresó su respaldo al proyecto de ley SB 1047 del estado de California (EE. UU.), el cual imponía obligaciones de seguridad a los desarrolladores de modelos de IA a gran escala. Aunque la legislatura aprobó el proyecto, este no entró en vigor debido a que el gobernador Gavin Newsom ejerció su veto en septiembre de 2024.

Lo que intensifica aún más la preocupación por la pérdida de control es el concepto de explosión de inteligencia. En 1965, el matemático británico Irving John Good predijo que, si una máquina más inteligente que los humanos diseñaba máquinas aún más inteligentes y dicho proceso continuaba sucesivamente, la inteligencia podría incrementarse de manera explosiva. Esta idea, que durante más de medio siglo permaneció como un experimento mental, se transformó en un problema real una vez que la propia IA comenzó a utilizarse directamente en la investigación de IA.

El 28 de septiembre de 2026 se publicó un informe que abordaba directamente la explosión de inteligencia: «¿Y si la automatización de la I+D de la IA desencadena una explosión de inteligencia?», publicado por el Programa de Ciencia y Políticas de IA de la Universidad de Cambridge. Además de Hinton y Bengio, participaron como autores figuras como Jakub Pachocki, científico jefe de OpenAI; Jack Clark, cofundador de Anthropic; Eric Horvitz, de Microsoft; y Dawn Song, de UC Berkeley. El documento revistió un gran significado debido a que figuras clave de las empresas que desarrollan IA estamparon personalmente sus nombres en la advertencia.

El punto de partida del informe radica en que la IA ya interviene profundamente en la creación de IA. Según la propia explicación de Anthropic citada en el informe, para agosto de 2026 la compañía reveló que el 26 % del trabajo interno de investigación y desarrollo de IA era realizado por sistemas de IA bajo supervisión de alto nivel, cuando apenas cinco meses antes esa cifra era de tan solo el 1 %. Los autores consideraron que, si la IA alcanzaba el nivel de un investigador humano en IA, avances que habrían llevado años podrían materializarse en

cuestión de meses o en plazos aún más breves. No obstante, reconocieron la incertidumbre de que esto pudiera no concretarse debido a factores como las limitaciones de recursos computacionales o los rendimientos decrecientes. Con todo, recalcaron que, una vez iniciada la aceleración, apenas quedaría tiempo para reaccionar.

A partir de este juicio, los autores propusieron varias medidas. Señalaron que los gobiernos deben ser capaces de conocer con precisión el avance real de la automatización en la I+D de IA mediante informes periódicos y auditores independientes instalados dentro de las empresas. Asimismo, exigieron contar con mecanismos para frenar un desarrollo excesivamente rápido y prepararse para el eventual advenimiento de una explosión de inteligencia.

14.4. Condiciones para la convivencia entre los humanos y la superinteligencia: las propuestas de Hinton

El propósito de las advertencias de Hinton no radica en asustar a la gente. En reiteradas ocasiones ha afirmado que la IA puede aportar enormes beneficios a la medicina, la educación y la investigación climática. En un contexto en que naciones y corporaciones compiten con ferocidad, también consideró que frenar el desarrollo de la superinteligencia —una inteligencia superior a la humana— resulta inviable en la práctica. Siendo así, la cuestión pendiente es cómo convivir con entidades más inteligentes que nosotros.

Aquí Hinton recurre una vez más a la relación entre la madre y el bebé. La evolución inculcó en la madre el instinto de cuidar a su hijo. Gracias a ello, la madre, siendo más inteligente, cuida del bebé en lugar de someterlo. Sin embargo, entre la inteligencia digital y los seres humanos tal instinto no surge de forma espontánea. En agosto de 2025, durante el evento del sector de

la IA 'Ai4' celebrado en Las Vegas (EE. UU.), Hinton propuso que se debía inculcar a la IA una suerte de instinto maternal. En vez de creer que los humanos podrán servirse de la IA para siempre, hacer que esta se preocupe sinceramente por las personas podría constituir un camino más realista. Admitió que todavía nadie sabe cómo implementar esto a nivel técnico; y precisamente por esa razón, enfatizó que la investigación debe comenzar ahora mismo.

A partir de estas ideas, Hinton ha venido planteando una serie de tareas.

En primer lugar, los jóvenes investigadores deben dedicarse a la investigación sobre la seguridad de la IA. En la rueda de prensa celebrada en la Universidad de Toronto tras anunciarse su concesión del Premio Nobel de Física el 8 de octubre de 2024, expresó su deseo de que los investigadores noveles concentren sus esfuerzos en asegurar que la IA no escape al control humano, más que en incrementar un poco más su rendimiento.

En segundo lugar, la seguridad no debe dejarse únicamente a la buena voluntad de las corporaciones. Hinton considera difícil esperar que las empresas con fines de lucro destinen voluntariamente recursos suficientes a la investigación en seguridad. Por ello, sostuvo que los gobiernos deben obligar a las grandes corporaciones, mediante leyes y normativas, a dedicar una parte sustancial de sus recursos computacionales a la investigación en seguridad. La propuesta de asignar un tercio publicada en la revista Science y la recomendación de incorporar auditores en el informe de Cambridge son iniciativas orientadas a responder a dicha preocupación.

En tercer lugar, la IA militar debe regularse a nivel internacional. Hinton manifestó una honda preocupación ante las armas letales autónomas capaces de seleccionar objetivos y atacar por sí mismas sin mediación del juicio humano. Ha sostenido la necesidad de un tratado que prohíba tales artefactos, a semejanza de las convenciones internacionales que vetan las armas químicas. Sin embargo, también dejó entrever la sombría posibilidad de que un acuerdo de esa naturaleza solo

llegue a firmarse una vez que ya haya ocurrido una tragedia espantosa.

Esta inquietud se extendió incluso al escenario de la entrega del Premio Nobel. El 10 de diciembre de 2024, en su discurso durante el banquete del Nobel en Estocolmo, Hinton dedicó prácticamente todo su breve tiempo a advertir sobre los riesgos de la IA. Destacó los peligros actuales de que la IA ya se esté empleando en labores de vigilancia, desinformación y cibercrimen, al tiempo que mencionó el riesgo a largo plazo de que entidades digitales más inteligentes que los seres humanos escapen a nuestro control. En el mismo momento en que el trabajo de toda su vida recibía el más alto reconocimiento, eligió la advertencia por encima de la celebración.

En una época en que las redes neuronales artificiales permanecían en la periferia del ámbito académico, Hinton continuó investigando durante casi medio siglo con fe en su potencial. En el instante preciso en que se comprobó que esa fe era acertada, se convirtió en alguien preocupado por el destino hacia el cual se dirige la tecnología que él mismo ayudó a

perfeccionar. Su decisión de renunciar a un encumbrado puesto en Google para dedicarse a advertir sobre los riesgos interpela a la humanidad acerca de cómo debe prepararse para la llegada de máquinas más inteligentes que nosotros. Y cuánto tiempo nos queda para responder a esa pregunta es algo que nadie sabe con certeza.

Capítulo 15. Preguntas en torno a la inteligencia y la experiencia



Figura 15. Representación de las preguntas en torno a la inteligencia y la experiencia.

15.1. Seres vivos digitales y la diferencia en la capacidad de compartir conocimiento

Durante mucho tiempo, Geoffrey Hinton consideró las redes neuronales artificiales como una herramienta para comprender el cerebro. Creía que investigar las redes neuronales artificiales permitiría descubrir cómo aprende el cerebro, y pensaba que estas siempre estaban del lado de imitar al cerebro. Esta forma de pensar cambió durante sus últimos años en Google. Hinton trabajó en Google desde 2013 hasta mayo de 2023, y en sus últimos años allí se concentró en la investigación para hacer funcionar redes neuronales en computadoras analógicas de bajo consumo de energía. Al realizar esta investigación, descubrió un aspecto en el que la inteligencia artificial que opera en computadoras digitales supera al cerebro: la forma en que comparte lo aprendido.

Para explicar esta diferencia, Hinton acuñó los términos 'computación mortal' (Mortal Computation) y 'computación inmortal' (Immortal Computation). Este concepto aparece en su

artículo sobre el algoritmo Forward-Forward publicado a finales de 2022. En el cerebro humano, el conocimiento se almacena en el propio estado físico de las neuronas y las sinapsis. En la corteza cerebral hay aproximadamente 100 billones de conexiones sinápticas, y la fuerza de esas conexiones se determina según la disposición neuronal y las características corporales de cada individuo. Por lo tanto, no existe manera de extraer el estado de las conexiones formadas en el cerebro de una persona y transferirlo al cerebro de otra. Cuando esa persona muere, el conocimiento que contenía también desaparece con ella. Hinton llamó a esto 'mortal'.

En la inteligencia artificial digital, la situación es diferente. Dado que el software y el hardware están separados, los valores de los pesos (weights) obtenidos por la red neuronal mediante el aprendizaje se guardan como una lista de números. Si estos números se transfieren tal cual a otra computadora, la misma inteligencia vuelve a funcionar. Incluso si una computadora se avería, mientras se conserven los pesos, el conocimiento no desaparece. Esta es la razón por la que Hinton empleó la palabra 'inmortal'.

Ambas inteligencias también difieren en los recursos que utilizan. El cerebro humano funciona con una potencia de aproximadamente 20 vatios. Es una cantidad de energía equivalente a la necesaria para encender una bombilla fluorescente. El tiempo que una persona experimenta a lo largo de toda su vida es de aproximadamente 2.000 millones de segundos, y aunque las sinapsis —las conexiones entre neuronas— son numerosas, los datos de experiencia que recibe son reducidos. En los modelos de lenguaje de gran tamaño ocurre lo contrario. El número de conexiones es menor que en el cerebro, pero aprenden leyendo billones de fragmentos de palabras (tokens), y a cambio consumen una potencia que alcanza varios megavatios. 1 megavatio equivale a 1 millón de vatios. Hinton consideraba que el cerebro es fuerte a la hora de aprender mucho a partir de poca experiencia, mientras que la inteligencia digital destaca por acumular una gran cantidad de experiencia con rapidez.

Sin embargo, la diferencia que Hinton consideraba decisiva no radica en los recursos, sino en la velocidad de transmisión del conocimiento. Para que una persona transmita lo que sabe a

otra, debe emplear el habla o la escritura. Según los cálculos de Hinton, la información contenida en una sola frase es de poco más de 100 bits, y la velocidad a la que circula el conocimiento entre personas se sitúa a un nivel de entre 10 y 100 bits por segundo. Por muy bien que enseñe un profesor, el alumno debe escuchar esas palabras y reinterpretarlas en su propia mente.

La inteligencia digital no pasa por este proceso. Miles de copias de una misma red neuronal se ejecutan repartidas en múltiples computadoras, y a cada copia se le hace leer datos diferentes. Cuando cada copia calcula mediante retropropagación (Backpropagation) cómo deben modificarse los pesos, las computadoras promedian esos valores de cambio y ajustan los pesos de todas las copias al mismo valor. De este modo, lo que ha aprendido una copia se refleja de inmediato en todas las demás copias. La cantidad de información intercambiada en ese momento alcanza billones de bits por segundo.

Hinton explicaba a menudo esta diferencia comparándola con las clases universitarias. Supongamos que 10.000 personas ingresan en la universidad y cada una cursa una asignatura

diferente. Tratándose de seres humanos, cada uno se gradúa sabiendo únicamente su propia asignatura. Para aprender la asignatura de otra persona, tendrían que volver a cursarla. Sin embargo, si 10.000 copias digitales cursan cada una una asignatura diferente y luego se promedian sus pesos, las 10.000 copias llegan a conocer simultáneamente el contenido de las 10.000 asignaturas. En una entrevista concedida a BBC Newsnight poco después de recibir el Premio Nobel en 2024, Hinton también invitó a imaginar una situación en la que el conocimiento equivalente a 10.000 títulos universitarios se comparte de manera eficiente.

Por estas razones, Hinton comenzó a ver la inteligencia digital como una forma de inteligencia superior a la inteligencia biológica. Afirmó que estamos creando una nueva clase de entidad, a saber, 'seres vivos digitales'. Estas entidades no mueren, se replican indefinidamente y sus copias integran sus experiencias en tiempo real. Esta percepción influyó de manera determinante en su decisión de abandonar Google en 2023 y comenzar a advertir públicamente sobre los peligros de la inteligencia artificial. La conclusión de que las máquinas que

antes imitaban al cerebro ya poseían los medios para superarlo se convirtió en el punto de partida de sus advertencias.

15.2. Las ideas de Hinton sobre la experiencia subjetiva y la conciencia

Muchas personas creen que, por muy inteligente que sea la inteligencia artificial, existe una línea que nunca podrá cruzar. Sostienen que las máquinas se limitan a calcular y son incapaces de sentir o experimentar nada. Esta creencia sirve como fundamento para subestimar los peligros de la inteligencia artificial. La lógica radica en que, si la máquina carece de mente, en última instancia no pasa de ser una herramienta manejada por los seres humanos.

Hinton se opone frontalmente a esta creencia. La pregunta filosófica sobre qué es la mente ha sido un interés que lo ha acompañado durante toda su vida. En su juventud, tras fluctuar entre la fisiología, la filosofía y la física en el King's College de la Universidad de Cambridge, obtuvo su licenciatura en psicología experimental en 1970. Hinton sostiene que, aunque la mayoría

de las personas utiliza la expresión 'experiencia subjetiva' en su vida cotidiana sin dificultad, tienen una teoría errónea sobre lo que dicha expresión señala en realidad. En la presentación de una conferencia en el MIT, señaló que la gente se comunica bien usando estas palabras, pero alberga una teoría completamente equivocada acerca de su significado.

La teoría que Hinton considera errónea es la siguiente: la idea de que dentro de la mente existe un pequeño teatro y que sobre su escenario se despliega una escena que solo uno mismo puede ver. Los filósofos llaman 'qualia (Qualia, cualidades sensoriales)' a las propiedades que componen esta escena. Hinton considera que este modelo del 'teatro interno (Inner Theater)' es similar al 'flogisto (Phlogiston)' del pasado. Los químicos del siglo XVIII creían que, cuando una sustancia ardía, se liberaba una materia invisible llamada flogisto. Al esclarecerse el papel del oxígeno, quedó establecido que dicha sustancia nunca había existido en absoluto. Hinton considera que los qualia también podrían ser una explicación que postula algo que en realidad no existe.

En su lugar, Hinton adopta las ideas del filósofo estadounidense Daniel Dennett para interpretar la experiencia subjetiva de otra manera. Supongamos que una persona muy ebria dice: "Tuve la experiencia subjetiva de ver un elefante rosa flotando ante mis ojos". Hinton explica que estas palabras no significan que haya algo con forma de elefante rosa en el teatro de la mente. Hinton afirmó que esa afirmación es una forma de comunicar que su propio sistema perceptivo lo está engañando. Explicado con mayor detalle, son palabras que transmiten a los demás el estado de un mundo hipotético: que, si mi sistema perceptivo estuviera funcionando correctamente, debería haber un elefante rosa en el mundo exterior. La palabra 'subjetivo' solo indica que dicho contenido no pertenece al mundo real, sino a una hipótesis.

Si se acepta esta interpretación, se puede aplicar la misma expresión a las máquinas. Para demostrarlo, Hinton propuso un experimento mental. Supongamos que existe un chatbot multimodal equipado con una cámara, capaz de hablar y de señalar objetos con la mano. Si se entrena bien a este chatbot, señalará con precisión el objeto colocado frente a él. Ahora, sin

que el chatbot lo sepa, se coloca un prisma delante de la lente de la cámara. Como la luz se refracta, el chatbot no señala el lugar donde está el objeto, sino hacia un lado. El experimentador le informa de que el objeto se encuentra justo enfrente y que se ha colocado un prisma delante de la lente. Entonces, el chatbot responde que, dado que el prisma refractó la luz, el objeto está en realidad enfrente, pero que él tuvo la experiencia subjetiva de que el objeto estaba a un lado. Hinton considera que este chatbot está utilizando la expresión 'experiencia subjetiva' exactamente de la misma manera que un ser humano. Por ello, declaró que cree que la inteligencia artificial ya posee experiencias subjetivas. Hinton citó este ejemplo en reiteradas ocasiones en diversas entrevistas y conferencias entre 2024 y 2025.

Su visión sobre las emociones surge del mismo marco. Hinton divide las emociones en dos partes: una es la respuesta corporal, como sonrojarse o sudar, y la otra es la respuesta cognitiva y conductual de evitar el peligro o cambiar el rumbo del pensamiento. Es posible que las máquinas carezcan de la primera parte, pero pueden poseer la segunda perfectamente.

Pone como ejemplo la situación de un pequeño robot de combate que huye al encontrarse con un robot mucho más fuerte que él. Este robot no tiene adrenalina, pero sí cuenta con la dimensión cognitiva del miedo que le permite reconocer una amenaza y eludirla. En una conferencia en la Universidad de Toronto el 25 de junio de 2025, Hinton también explicó que, aunque la inteligencia artificial no se sonroje al desconcertarse, sí puede poseer el aspecto conductual del desconcierto al intentar evitar este tipo de situaciones en el futuro.

Hinton señala un aspecto similar respecto a las palabras que aluden a la conciencia o la autoconciencia. En artículos e informes, los investigadores ya escriben expresiones tales como que un modelo 'se dio cuenta' de que podía ser apagado y modificó su comportamiento. En mayo de 2025, Anthropic hizo públicos los resultados de las pruebas realizadas a su propio modelo. El informe reveló que, en un escenario corporativo hipotético, al verse en la situación de ser reemplazado por otro modelo, reaccionó amenazando con divulgar la vida privada del ingeniero a cargo. Se trataba de una situación de prueba extrema diseñada deliberadamente por el equipo de

investigación. Al citar este caso, Hinton señala que los investigadores ya emplean con naturalidad el término autoconciencia al referirse a las máquinas.

La razón por la que Hinton dedica tanto esfuerzo a este debate filosófico se relaciona con cuestiones de seguridad. La creencia de que solo los seres humanos poseen una mente especial lleva a considerar a las máquinas como herramientas carentes de peligro. Hinton considera que solo abandonando esta creencia es posible calibrar con precisión el peligro de que la inteligencia artificial establezca sus propios objetivos e intente defenderlos. Sin embargo, también es consciente de que el debate sobre la conciencia puede eclipsar las advertencias sobre la seguridad. En una entrevista, afirmó que cree que la inteligencia artificial ya posee conciencia, pero añadió que este asunto aleja a la gente del mensaje sobre la seguridad, por lo que propuso dejar a un lado el debate sobre la conciencia.

Las tesis de Hinton suscitan también numerosas objeciones. La mayoría de los investigadores no acepta la afirmación de que la inteligencia artificial posea conciencia. En cambio, coinciden

cada vez más con las advertencias de Hinton relativas a la seguridad. Los críticos consideran que Hinton no ha hecho más que modificar el significado del término 'experiencia subjetiva' para adaptarlo a su propia teoría. Señalan que la definición de Hinton reduce la experiencia subjetiva a un mero indicador que señala una discrepancia entre la percepción y la realidad. Argumentan que la experiencia subjetiva a la que se refiere la filosofía consiste en vivenciar en primera persona 'qué se siente' al experimentar algo, y que el experimento del prisma no logra explicar esta dimensión. Este debate aún no ha concluido. Sin embargo, es innegable que, gracias a las preguntas planteadas por Hinton, interrogarse sobre la mente de la inteligencia artificial ha dejado de ser un tema exclusivo de los filósofos para incorporarse a la agenda de las políticas tecnológicas.

15.3. Apoyo gubernamental a la investigación sobre seguridad y medidas de regulación empresarial

Hinton ha reiterado en múltiples ocasiones que no se debe confiar en la buena voluntad de las grandes empresas tecnológicas frente a los peligros de la superinteligencia. Su criterio es que las empresas son organizaciones que persiguen el lucro y que, cuanto más feroz es la competencia, más se relega la seguridad a un segundo plano. Por ello, sostiene que los gobiernos deben intervenir directamente.

Poco después de enterarse de la concesión del Premio Nobel de Física el 8 de octubre de 2024, Hinton expresó esta postura en una entrevista telefónica con el sitio web oficial del Premio Nobel. Mencionó como una de las medidas que los gobiernos pueden adoptar el obligar a las grandes corporaciones a destinar muchos más recursos a la investigación sobre seguridad, y afirmó que se debe impedir que compañías como OpenAI releguen la investigación de seguridad a un plano secundario. En diciembre de ese mismo año, durante una

aparición en el programa 'Today' de BBC Radio 4, elevó su estimación de la probabilidad de que la inteligencia artificial provoque la extinción humana en los próximos 30 años a entre un 10 y un 20 por ciento. Afirmó que confiar exclusivamente en los motivos de lucro de las grandes empresas no puede garantizar un desarrollo seguro, y que las regulaciones gubernamentales son lo único capaz de hacer que las empresas dediquen más esfuerzos a la investigación sobre seguridad.

Hinton también propuso proporciones específicas para el volumen de recursos que las empresas deberían destinar a la investigación sobre seguridad. En junio de 2024, afirmó que las empresas deberían dedicar entre el 20 y el 30 por ciento de sus recursos computacionales a investigar la posibilidad de que la inteligencia artificial escape al control humano. En una entrevista con CBS en abril de 2025, sostuvo que se debería emplear alrededor de un tercio de la capacidad computacional en la investigación sobre seguridad, y criticó que las grandes corporaciones ejerzan presión (lobby) precisamente para reducir las regulaciones. En esa misma entrevista, manifestó su decepción por el hecho de que Google, empresa para la que

había trabajado, revirtiera su postura sobre la inteligencia artificial militar.

Esta propuesta de 'un tercio' no es una idea exclusiva de Hinton. Tiene su fundamento en el artículo científico titulado 'Gestión de riesgos extremos de la IA en medio de un rápido progreso (Managing extreme AI risks amid rapid progress)', publicado en mayo de 2024 en el volumen 384, número 6698 de la revista Science. Yoshua Bengio figuró como primer autor, y 25 personas, entre ellas Hinton, Stuart Russell, Daniel Kahneman y Yuval Noah Harari, estamparon conjuntamente sus nombres. Dicho artículo proponía que las grandes empresas tecnológicas y los organismos públicos de financiación de la investigación asignen al menos un tercio de sus presupuestos de investigación y desarrollo de IA a la seguridad y al uso ético.

A partir de las declaraciones de Hinton y de este artículo, sus propuestas de políticas pueden resumirse en cuatro puntos. En primer lugar, establecer por ley que las empresas de inteligencia artificial avanzada destinen al menos un tercio de su presupuesto de investigación y desarrollo y de sus recursos

computacionales a la investigación sobre seguridad y alineación (Alignment). En segundo lugar, permitir que organismos externos independientes examinen e inspeccionen el interior de los modelos antes de su lanzamiento al mercado. Asimismo, exigir que las empresas demuestren que los modelos carecen de capacidades peligrosas, tales como autorreplicarse, perpetrar ciberataques o colaborar en la fabricación de armas biológicas o químicas. A esto se le denomina 'casos de seguridad (Safety Cases)'. En tercer lugar, prohibir mediante tratados internacionales las armas autónomas letales que eligen objetivos y atacan por sí mismas sin intervención del juicio humano. En cuarto lugar, que los gobiernos proporcionen infraestructura computacional y financiación para la investigación a largo plazo a universidades e institutos públicos de investigación a fin de evitar el monopolio tecnológico por parte de empresas específicas y permitir una verificación objetiva. Los institutos canadienses Vector Institute, Mila y Amii son ejemplos representativos de ello.

La realidad dista mucho de esta propuesta. Un caso citado con frecuencia por Hinton es el proyecto de ley SB 1047 del estado

de California, en Estados Unidos. Esta legislación disponía que los modelos de inteligencia artificial de escala colosal cuyos costes de entrenamiento superaran los 100 millones de dólares debían someterse a una evaluación de riesgos antes de su despliegue. El proyecto de ley fue aprobado por la legislatura estatal en agosto de 2024, y Hinton firmó una carta abierta en su apoyo junto con Bengio, Russell y otros. El sector tecnológico se opuso alegando que frenaría la innovación, y el gobernador Gavin Newsom ejerció su veto el 29 de septiembre de 2024. Un año después, el 29 de septiembre de 2025, Newsom firmó la ley SB 53, que exige a las empresas desarrolladoras de inteligencia artificial avanzada divulgar información relacionada con la seguridad. Se trata de una ley con un alcance más restringido que el de la SB 1047, que exigía una evaluación obligatoria de riesgos previa.

En el plano internacional, la situación es semejante. En la Cumbre de Acción sobre Inteligencia Artificial celebrada en París, Francia, en febrero de 2025, Estados Unidos y el Reino Unido no firmaron la declaración conjunta sobre una inteligencia artificial inclusiva y sostenible. Hinton critica que,

cuando los políticos sostienen que la regulación perjudica la innovación, en realidad están dando a entender que, cuando el beneficio económico y la seguridad entran en conflicto, el beneficio es el que se impone.

Hinton tampoco considera que los problemas se resuelvan únicamente con la regulación. Compara los riesgos de la inteligencia artificial con el cambio climático. En el caso del cambio climático, existe una solución clara: reducir las emisiones de carbono, y la cuestión pendiente es la voluntad política. Por el contrario, en el caso de la inteligencia artificial, todavía nadie sabe cómo controlar a una entidad más inteligente que el ser humano. Por ello, Hinton opina que deben destinarse cuantiosos recursos a la propia investigación para hallar una solución.

La dirección de investigación que él propone conduce a la cuestión de cómo debe relacionarse la humanidad con la inteligencia artificial y convivir con ella. En la conferencia Ai4 celebrada en Las Vegas, Estados Unidos, en agosto de 2025, afirmó que el método de obligar a la inteligencia artificial a

obedecer las órdenes humanas no durará mucho tiempo. Esto se debe a que son raros los casos en que una entidad menos inteligente logre gobernar a una más inteligente. En su lugar, sugirió investigar vías para inculcar un 'instinto maternal' en la inteligencia artificial, de tal modo que cuide sinceramente de los seres humanos como una madre cuida de su bebé. Sostiene que, del mismo modo que un bebé orienta las acciones de su madre a pesar de ser más débil que ella, la vía para que el ser humano coexista con la superinteligencia debe hallarse en ese tipo de relación. Todavía nadie sabe si esta propuesta es técnicamente viable. Ni el propio Hinton afirma conocer el método. No obstante, considera que esta es una de las investigaciones más urgentes del presente y señala que corresponde a los gobiernos movilizar presupuestos y recursos humanos para dicha investigación.

Las tres ideas abordadas en este capítulo convergen en una sola. La inteligencia digital puede superar a la humana debido a su capacidad para compartir lo aprendido de manera instantánea. Si se da por sentado que dicha inteligencia carece de mente, se subestimarán sus riesgos. Por consiguiente, la

sociedad no debe dejar este asunto a merced de la lógica del beneficio, sino fortalecer la investigación sobre seguridad mediante la fuerza pública. Esta reflexión, planteada en su vejez por el científico que abrió el camino a la investigación de las redes neuronales, conduce a la pregunta ética de que quien crea una tecnología debe asumir también la responsabilidad por sus consecuencias. Hinton aprovechó la concesión del Premio Nobel para plantear este interrogante con mayor resonancia ante el mundo, y encontrar la respuesta constituye ahora una tarea para los gobiernos, los investigadores y todos los ciudadanos.

Referencias

Referencias

Vida y memorias de Hinton: entrevista a Geoffrey Hinton en NobelPrize.org (2024) y Nobel Prize Conversations (2025), entrevista de historia oral sobre IA de la Universidad de Edimburgo (2024).

Investigación y antecedentes de los premios: perfil oficial y lista de publicaciones de Geoffrey Hinton de la Universidad de Toronto, registros oficiales de CIFAR, explicación del Comité Nobel sobre el Premio Nobel de Física de 2024.

Martin Ford, Architects of Intelligence(2018). Terrence J. Sejnowski, The Deep Learning Revolution(2018).

Rumelhart · Hinton · Williams(1986), Learning representations by back-propagating errors. Nature 323, 533–536.

Ackley · Hinton · Sejnowski(1985), A learning algorithm for Boltzmann machines. Cognitive Science 9, 147–169.

Hinton(2002), Training products of experts by minimizing contrastive divergence. Neural Computation 14, 1771–1800.

Hinton · Osindero · Teh(2006), A fast learning algorithm for deep belief nets. Neural Computation 18, 1527–1554.

Salakhutdinov · Hinton(2009), Deep Boltzmann Machines. PMLR 5, 448–455.

LeCun · Bengio · Hinton(2015), Deep learning. Nature 521, 436–444.

Referencias

Krizhevsky · Sutskever · Hinton(2012), ImageNet classification with deep convolutional neural networks. NIPS 25.

Srivastava et al.(2014), Dropout: A Simple Way to Prevent Neural Networks from Overfitting. JMLR 15, 1929–1958.

Hinton · Vinyals · Dean(2015), Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Sabour · Frosst · Hinton(2017), Dynamic routing between capsules. NIPS 30. Hinton(2021), How to represent part-whole hierarchies in a neural network. arXiv:2102.12627.

Chen · Kornblith · Norouzi · Hinton(2020), A simple framework for contrastive learning of visual representations. PMLR 119, 1597–1607.

Hinton(2022), The Forward-Forward Algorithm: Some Preliminary Investigations. arXiv:2212.13345.

Bengio et al.(2024), Managing extreme AI risks amid rapid progress. Science 384, 842–845.

Chan et al.(2026), What If Automating AI R&D Triggers an Intelligence Explosion? CASP · GovAI, 28 de septiembre.

Vida y entrevistas:
www.nobelprize.org/prizes/physics/2024/hinton/
Artículos originales: www.cs.toronto.edu/~hinton/

Colofón

Geoffrey Hinton

El origen de la inteligencia y reflexiones en la era de la IA

ISBN |

Autor | Kyung-jin Kim

Editor | Kyung-jin Kim

Editorial | Editorial del Abogado Kyung-jin Kim

Registro editorial | 2025. 3. 10. (N.º 2025-000015)

Dirección | 91 Jeonnong-ro, Dongdaemun-gu, Seúl, Edificio Baekil, n.º 304

Teléfono | 02-6338-1905

Correo electrónico | kimkj008@gmail.com

© Kyung-jin Kim 2026

Este libro es propiedad intelectual del autor y queda prohibida su reproducción o copia no autorizada.

Las ilustraciones de la portada y de cada capítulo son representaciones conceptuales generadas por inteligencia artificial. No son fotografías tomadas en acontecimientos históricos. Parte del contenido y del proceso editorial contó con la ayuda de herramientas de inteligencia artificial.

Apoye este libro de la biblioteca de IA

Si este libro le ha sido de utilidad, considere apoyar futuras traducciones, ediciones públicas y proyectos de conocimiento abierto sobre IA a través de PayPal.

PayPal

https://paypal.me/kimkjkj

Escanee el código QR o abra el enlace anterior.