

傑佛瑞·辛頓

智慧的起源與AI時代的省思



金慶鎮律師

構想學習機器的人

傑佛瑞·辛頓是一位從人類大腦學習方式中尋找機器學習線索的研究者。他構想了一種並非由人類逐一輸入規則，而是能從資料中自行學習的機器。即使在神經網路研究未受學界重視的時期，他也未曾放下這個問題。

這本書從童年時期的好奇心出發，循著他與同事及學生們共同進行的研究展開。從1986年的反向傳播研究、2012年 AlexNet 的成果，到2018年的圖靈獎與2024年的諾貝爾物理學獎，審視神經網路使今日的人工智慧成為可能的過程。陌生的概念則會結合研究者當年試圖解決的問題一同說明。

辛頓的一生並非只有成就。他在2023年離開 Google 時，更積極地警告人工智慧的危險。他開始探問自己所發展的技術將對社會帶來何種影響。試圖理解學習的提問與追問技術責任的提問，在他的生命中延續著。期盼這本書能成為青少年讀者理解人工智慧的原理與限制，並自行思考其未來的契機。

金慶鎮律師

目錄

構想學習機器的人

第1部. 探向大腦與學習的提問

第1章. 始於好奇心的大腦研究

第2章. 記憶是如何連結的

第3章 從錯誤答案中學習的神經網路

第二部. 讓神經網路學習的方法

第4章. 從物理學中找到的學習原理

第5章. 更快速的學習方法

第6章. 深度神經網路重獲關注

第三部. 深度學習走向世界

第7章. 辨識語音與照片的電腦

第8章. 減少過擬合並傳遞知識

第9章. 共同締造的深度學習發展

第4部. 尋求更好的學習

第10章。辨識局部與整體的機器

第11章 從大腦中重新尋找的學習方法

第12章 沒有標準答案的學習方法

第5部. 成就與責任

第13章. 諾貝爾物理學獎與神經網路

第14章 警告人工智慧的危險

第15章. 圍繞智慧與經驗的提問

參考資料

第1部. 探向大腦與學習的提問

第1章. 始於好奇心的大腦研究



圖1. 展現童年時期的觀察與科學好奇心。

1.1. 布爾（Boole）家族留下的邏輯與機率

傑佛瑞·埃弗里斯特·辛頓（Geoffrey Everest Hinton）於1947年12月6日出生於英國倫敦的溫布頓。他的家族歷代人才輩出，是孕育出多位數學家、科學家、醫師與著述家的學者世家。在先祖之中聲名遠播的人物，是他的高祖父喬治·布爾（George Boole）。布爾是19世紀的英國數學家，他開創了將真與假視同數字1與0來進行運算的布爾代數。如今所有電腦的邏輯電路都是以這套布爾代數為基礎設計而成的。布爾的妻子瑪麗·埃弗里斯特·布爾（Mary Everest Boole）也是一位數學家兼教育家。瑪麗的叔父喬治·埃弗里斯特爵士（Sir George Everest）曾任印度測量總監，世界上最高的山峰埃佛勒斯峰的名字便來自於他。辛頓的中間名埃弗里斯特便源自這段家族淵源。

父親那一邊的辛頓家族同樣也是學者世家。曾祖父查爾斯·霍華德·辛頓（Charles Howard Hinton）是一位以視覺化描繪4維空間研究而聞名的數學家，並與喬治·布爾的女兒瑪麗·艾倫結為連理。查爾斯的父親，也就是辛頓的另一位高祖父詹姆斯·辛頓（James Hinton），則是一位外科醫師兼著述家。父親的堂妹瓊安·辛頓（Joan Hinton）在第2次世界大戰期間，是參與美國洛斯阿拉莫斯研

製原子彈的曼哈頓計畫中少數的女性物理學家之一。戰後瓊安·辛頓反對使用原子彈，毅然告別物理學界並前往中國。

父親霍華德·埃弗里斯特·辛頓（H. E. Hinton）是布里斯托大學的昆蟲學教授，也是英國皇家學會院士。他研究過多種昆蟲，並對生活在水中的小型甲蟲群體溪泥甲科（Elmidae）懷有深厚的興趣。辛頓曾在諾貝爾獎相關訪談中回憶起父親研究室的情景。一面牆上的高層層架上密密麻麻地擺放著數百個裝滿昆蟲學論文的盒子。在其中，唯獨靠近門邊低處的一個小盒子貼著標籤寫著「非昆蟲（Not Insects）」，裡面裝著孩子們的信件和家信。那是一個以學問為生活重心的家庭。辛頓曾表示，自己從小就知道如果不能在學術上取得成功，就會成為一個失敗者，因此承受了巨大的學業成就壓力。

將這段家族淵源與人工智慧的歷史並列來看，會發現一個引人注目的特點。喬治·布爾於1854年出版的著作書名為《思維規律的研究：邏輯與機率的數學理論基礎》。書名中並列著邏輯與機率。然而，20世紀中葉以後引領人工智慧研究的符號主義研究者，卻只繼承了布爾遺產中的邏輯部分。他們試圖將人類的思維解釋為非真即假的符號，以及處理這些符號的規則。然而，現實世界充滿了模糊與不確定的事物，要處理這樣的世界，機率是不可或缺的。

布爾的玄孫辛頓正是在這一點上走上了與主流不同的道路。他認為大腦並非按照既定邏輯規則運轉的精密機器，而是無數神經細胞（神經元）根據經驗逐步調整彼此間連結強度（即連接權重）來進行學習的網路。如果說高祖父的邏輯學成為符號主義人工智慧的骨幹，那麼玄孫則是從高祖父書名中的另一半——機率出發，開啟了今日深度學習的大門。

1.2. 7歲時的領悟與摸索前路的學生時代

在辛頓回顧童年時經常提及的軼事中，有一件是他在公車上的經歷。他曾把一枚舊英國便士硬幣放在毛質粗糙的公車座椅上。每當公車晃動時，硬幣不但沒有往下滑，反而順著座椅的絨毛一路往上爬。年幼的辛頓覺得這是不可思議的事，極想探究其中的原因。不對無法理解的現象輕易放過的態度，從那時起就已經深深融入了他的性格之中。

7歲時，辛頓進入布里斯托的克利夫頓學院（Clifton College）就讀。那是一所帶有濃厚基督教色彩的學校。學校將信仰的教導視為理所當然的真理來傳授，大多數同學也都全盤接受了那些說法。但辛頓並沒有這麼做。他無法接受在缺乏證據的情況下盲目相信。他日後在接受《首爾經濟》採訪時回憶道，自己就是在這時第一次萌生了

「任何人都可能犯錯」的想法。他解釋說，這段經歷讓他相信其他研究者也可能會犯錯，並成為他對抗主流學界、堅持進行研究的基石。辛頓之所以能忍受學界近半個世紀的冷遇，其力量正是源自這段經歷。

他對大腦的興趣萌芽於高中時期。在16、17歲左右，辛頓一直認為比自己更聰明的朋友英曼·哈維（Inman Harvey），有一天提出了大腦的記憶或許像全像圖一樣分散儲存在整個大腦中的想法。全像圖即使只剩碎片，也能或多或少模糊地重現整幅圖像。記憶並非一個個單獨儲存在某一個固定位置，而是分散在多處重疊儲存，這一想法給辛頓帶來了巨大的震撼。想要揭開大腦如何進行運算的願望，正是在此時扎下了根。

然而，要找到能夠回應這份願望的學科並非易事。1967年，辛頓進入劍橋大學國王學院，隨後的幾年正如他自己所形容的，是一場「瘋狂的隨機漫步（crazy random walk）」。起初他選擇了自然科學，研習物理學、化學與X射線晶體學。面對每天長達12小時的緊湊日程，筋疲力盡的他僅過了一個月便離開了學校。他在倫敦靠打雜工賺取房租，閱讀杜斯妥也夫斯基的作品，甚至還想過成為一名建築師。然而，他在建築事務所工作了一天後，發現實際業務與其說

是設計，不如說是為了迎合預算挑選廉價的地板材質和門把，於是便立刻辭職了。

重返大學後，他陸續輾轉於可能探討大腦與心智的各個學系。首先，他抱著能學習大腦運作的期望選擇了生理學。然而課堂上講授的只是烏賊粗大神經纖維中電訊號如何傳導，並未涉及神經元如何表徵資訊以及如何進行學習。接著，為了探尋心智的本質，他轉向了哲學。但厭倦了缺乏實驗、純粹口舌之爭的方式，他在1年後聲稱自己只產生了「對哲學的抗體」便離開了。最後，他轉攻實驗心理學，並於1970年獲得學士學位。

心理學同樣沒能給出他所尋找的答案。當時的心理學偏向於觀察箱中老鼠行為的研究。辛頓在探討兒童認知發展的實驗中，讓孩子從黃色三角形、黃色圓形和紅色圓形中選出屬性不同的一個。一名5歲的孩子指著紅色圓形說，是研究人員把它塗錯了顏色。孩子的思維遠比實驗設計複雜得多。辛頓深感當時的心理學根本無法解釋人類的認知。

畢業之後，迷惘與徘徊仍在繼續。他花了約1年的時間學習木工。但在目睹工地裡熟練木匠的手藝後，他自認在這條路上無法達到那樣的境界，於是重返學術界。此後，他在布里斯托擔任兒童語言發展

研究的研究助理。這是一項驗證諾姆·喬姆斯基（Noam Chomsky）語言先天論——即人類生來就具備語法能力——的研究。研究團隊在3至5歲兒童的家中放置麥克風，收集了高達16萬6千筆的發話資料。然而，當前往布里斯托郊區的低收入家庭進行問卷調查時，面對第一個問題「孩子說話時您會怎麼做」，一位母親卻回答說會動手打孩子。經歷了這件事後，辛頓便退出了這項研究。

在歷經多個領域的輾轉探索後，他得出了一個結論：若要解開智慧的奧祕，就必須在電腦上親自建立起大腦的計算模型。

1.3. 在愛丁堡的孤立與指導教授朗蓋－希金斯的衝突

1972年，辛頓為了正式展開人工智慧研究，進入愛丁堡大學攻讀博士學位。愛丁堡大學是當時英國幾乎唯一具備完善人工智慧研究基礎的大學。然而，在那裡等待著他的，卻是孤立無援以及與指導教授的衝突。

指導教授是克里斯多福·朗蓋－希金斯（Christopher Longuet-Higgins）。他是一位在30多歲時就成為劍橋大學理論化學教授的傑出科學家，40多歲時轉向認知科學領域，研究如同全像

圖般儲存記憶的模型以及聯想記憶。這正是辛頓所關注的主題。然而，當辛頓抵達時，朗蓋－希金斯的想法早已發生轉變。他對泰瑞·威諾格拉德（Terry Winograd）的SHRDLU研究留下了深刻的印象。SHRDLU是一個在虛擬積木世界中，依據符號規則解析人類以英語下達的指令並移動積木的程式，被視為符號主義人工智慧的代表性成果。朗蓋－希金斯放下了對神經網路的執念，轉向了符號主義陣營。

當時學界的氛圍也對神經網路極為不利。1969年，馬文·明斯基（Marvin Minsky）與西摩爾·派珀特（Seymour Papert）在《感知器》（Perceptrons）一書中，從數學上證明了單層神經網路連XOR這樣看似簡單的問題都無法解決。此後，許多研究者將整個神經網路領域視為毫無希望的方向。偏好嚴謹簡潔邏輯體系的朗蓋－希金斯，無法接受以機率和調整連結強度這種模糊方式運作的神經網路。他直斥辛頓的研究荒謬無稽，並一再要求他轉向符號主義。

辛頓並未退縮。他承諾只要再給他6個月的時間，他就會展現成果，若做不到就更換研究主題。然而6個月過去後，他又重複了同樣的承諾，這樣的情況反覆上演了數次。辛頓在2024年愛丁堡大學的口述歷史訪談中坦承了自己沒有被逐出師門的原因。他說，父親是英國

皇家學會院士，而同為院士的朗蓋一希金斯大概很難輕易把同事的兒子開除。他自嘲這完全是多虧了自己所享有的特權。

辛頓最初構想的博士研究名為LUDWIG。這是「無先天語法的語言理解裝置（Language Understanding Device Without Innate Grammar）」的縮寫，既借用了哲學家維根斯坦的名字，又蘊含著反對喬姆斯基先天論的意圖。這項計畫旨在建立一個無須預先輸入符號規則、僅憑經驗就能學習語言的神經網路。然而在當時，根本沒有辦法訓練隱藏在輸入層與輸出層之間的隱藏層。以那個時代的技術，LUDWIG是無法打造出來的。

這時為他指明道路的人是博士後研究員史蒂夫·艾薩德（Steve Isard）。他實質上扮演了副指導教授的角色，建議辛頓先從手寫辨識等規模較小且較容易的問題著手，驗證神經網路的可能性。辛頓自學了法蘭克·羅森布拉特（Frank Rosenblatt）1962年的著作《神經動力學原理》（Principles of Neurodynamics），並實現了平行網路。當朗蓋一希金斯轉任薩賽克斯大學時，辛頓與艾薩德、馬克·斯蒂德曼（Mark Steedman）一同跟隨前往。辛頓在薩賽克斯大學擔任研究員，隨後於1978年以論文《鬆弛法及其在視覺中的作用》（Relaxation and Its Role in Vision）獲得愛丁堡大學博士學位。

1.4. 符號主義的局限與悄然展開的連結主義反擊

辛頓與朗蓋－希金斯的衝突並非僅僅是兩人間的私事。20世紀後半葉的人工智慧學界，存在著看待智慧的兩種對立觀點。其一是符號主義（Symbolic AI）。由艾倫・紐厄爾（Allen Newell）、赫伯特・賽門（Herbert Simon）、約翰・麥卡錫（John McCarthy）以及馬文・明斯基等人引領。他們認為人類的思維是由符號、處理這些符號的邏輯規則，以及一步步按部就班推進的推論所構成。這種想法認為，只要人類將規則整理清楚並輸入電腦，就能創造出智慧。這種方法日後也被稱為「昔日美好的人工智慧（GOF AI）」。

另一種則是連結主義（Connectionism）。辛頓與大衛・魯梅哈特（David Rumelhart）、泰倫斯・塞諾夫斯基（Terrence Sejnowski），以及後來加入的楊立昆（Yann LeCun）、約書亞・班吉歐（Yoshua Bengio）等人開闢了這條道路。連結主義者認為，智慧的核心不在於輸入規則，而在於從數據中自主學習表徵的能力。他們主張系統應像大腦一樣，由許多單元同時運作，並透過調整這些單元之間的連結強度（權重）來累積知識。在這種觀點下，一個概念並非儲存在單一格子中，而是分散儲存在多個單元的活化

模式中。這被稱為分散式表徵（Distributed Representation）。這一想法與辛頓高中時期聽說的全像圖記憶如出一轍。

進入1970年代，符號主義碰壁了。現實中的數據極其模糊且千差萬別，根本無法全部整理成規則。1973年，英國應用數學家詹姆士·萊特希爾爵士（Sir James Lighthill）發表了一份評估人工智慧研究的報告。這份萊特希爾報告指出，人工智慧研究無法解決組合爆炸問題。所謂組合爆炸，是指需要考量的情況數量隨著問題規模呈指數級增長，導致計算變得不可能的現象。該報告發表後，英國的人工智慧研究經費大幅縮減，第1次「AI寒冬」隨之降臨。

儘管許多人工智慧研究者批評萊特希爾報告，但辛頓的評價卻截然不同。他認為萊特希爾的指責切中了要害。像麥卡錫這樣的符號主義者主張，凡是能用符號計算的事物，都可以藉由人工智慧實現。然而，這種主張並未考量到計算所需的時間與資源，亦即計算複雜度。萊特希爾與朗蓋－希金斯以及物理學家弗里曼·戴森（Freeman Dyson）同為溫徹斯特公學（Winchester College）校友，作為一名研究流體力學的應用數學家，他敏銳地洞察到了這一弱點。辛頓也持相同看法。他認為，若要利用有限的計算資源應對現實世界的複雜性，單靠逐一操作符號的方法是遠遠不夠的，必須採用像大腦那樣由多個單元同時運作、逐步修正連結強度的方式。

在學界的冷落忽視之中，辛頓開始尋找志同道合的夥伴。1979年，他與詹姆士·安德森（James Anderson）在美國聖地牙哥的拉荷亞（La Jolla）共同舉辦了「聯想記憶平行模型（Parallel Models of Associative Memory）」研討會。包括加州大學聖地牙哥分校（UCSD）的大衛·魯梅哈特在內，來自多個機構的連結主義研究者齊聚一堂。這次聚會成為日後平行分散式處理（PDP）研究小組成立的墊腳石。儘管表面上看似平靜，但連結主義的反擊正是在此時悄然拉開了序幕。

1.5. 好奇心勝於短期成果：基礎研究結出的碩果與其意義

從1978年獲得博士學位，到2012年與學生共同開發的AlexNet在影像辨識競賽中取得壓倒性勝利為止，神經網路長久以來一直是一個難以籌措研究經費的邊緣領域。在此期間，人工智慧領域經歷了2次寒冬。支撐辛頓熬過這段漫長歲月的力量，既非迅速的商業成果，亦非個人聲譽，而是想要探究人類大腦究竟如何運作的好奇心。在接受諾貝爾獎官方訪談並被要求自我介紹時，他回答說，自己雖然不太清楚自己屬於哪個領域，但只是一個渴望理解大腦如何運作的人。

辛頓多次公開強調，這種好奇心應當是研究的起點。2024年10月8日，他與約翰·霍普菲爾德（John Hopfield）共同被宣布為諾貝爾物理學獎得主。在當天多倫多大學的新聞發布會上，他呼籲各界支持由好奇心驅動的基礎研究。他評價說，加拿大雖然不像美國擁有那麼龐大的研究經費，但資金使用得非常明智；神經網路的發展並非源於將大筆金錢砸向應用問題，而是來自科學家依循好奇心努力理解未知事物的成果。在接受《首爾經濟》採訪時他也表示，與其向學者施壓要求立即產出實用成果，不如讓研究者自由探討自己好奇的問題，這樣要有效得多，而深度學習正是這樣誕生出來的。

這類研究之所以能結出碩果，背後離不開加拿大高等研究院（CIFAR）的長期資助。1987年，辛頓在CIFAR的資助下移居加拿大，並在多倫多大學任教安頓下來。他指出，美國的科研經費體制迫使教授們耗費大量時間撰寫研究計畫書。相較之下，在加拿大只要提交一份6頁的研究計畫書，就能獲得為期3至5年、不拘泥於特定課題的穩定資助。正是在這樣的環境下，辛頓才得以不隨波逐流，堅持深耕於冷門領域。

辛頓所走過的道路，其意義並不僅止於技術上的成功，更改變了看待人類心智的視角。符號主義將人類視為精巧處理符號並進行邏輯推論的「推論機器（Reasoning Machine）」，而辛頓的研究則表

明，人類更接近於以感官經驗與相似性為基礎、將世界歸納為模式來理解的「類比機器（Analogy Machine）」。

在辛頓看來，邏輯推論只是附加在神經網路之上的能力，而智慧的基礎則在於無數神經元透過經驗自主習得分散式表徵的連結強度調節之中。

7歲少年心中所懷抱的想法——即大人也可能會犯錯的領悟——在半個世紀後茁壯成長為學界全體都可能犯錯的堅定信念。從高祖父喬治·布爾的邏輯學所籠罩的陰影中啟程的辛頓，緊握著同列於那本書名中的機率，一路堅持走到了最後。

第2章. 記憶是如何連結的

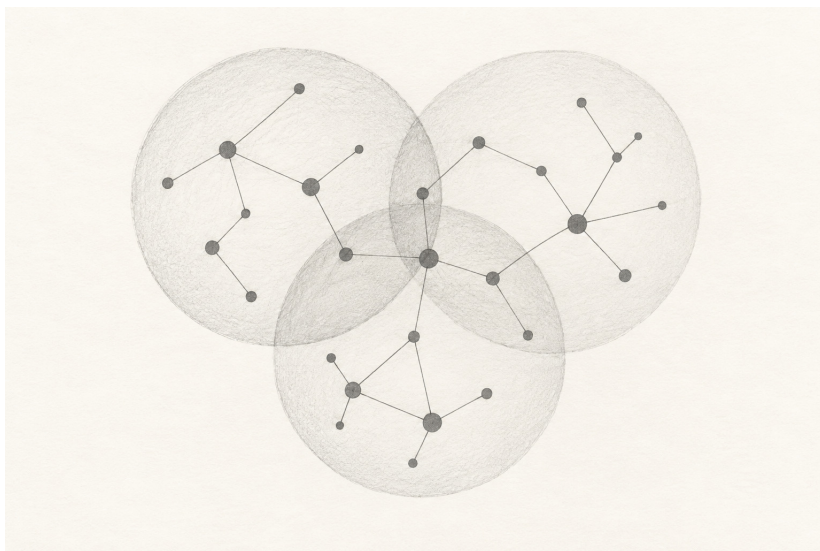


圖2. 呈現了將資訊分散於多個連結以表示的想法。

2.1. 「祖母神經元」的局限與分散式表徵的交響樂團

傑佛瑞·辛頓在高中時期從朋友那裡聽說了全像圖的故事。全像圖即使剪掉底片的一部分，整體影像依然會隱約留存。這是因為資訊並非記錄在底片的單一一處，而是分散在整張底片上。辛頓認為大腦的記憶或許也以類似的方式分散存在，並將這個想法作為一生的研究主題。

然而，在1970年代作為認知科學與人工智慧學界主流的符號主義，對知識的儲存方式卻有著截然不同的看法。符號主義學者認為人類的思維是由符號以及處理這些符號的規則所構成，並認為單一概念儲存在單一位置。這種儲存方式被稱為「局部表徵」。

展現局部表徵的說法就是「祖母神經元假說」。該假說認為，大腦中某處有一個只負責我祖母的神經元，當那個神經元被活化時，就能認出祖母。如果按照這個假說，對貓就必須有貓神經元，對特定句子也必須有負責該句子的符號節點。將單一對象對應到單一單元雖然容易區分，但難以表現多樣對象之間的共通點。不過，這並不意味著所有用符號解釋思維的理論都把單一概念對應到單一神經元上。

辛頓很早就對這種想法抱持懷疑。局部表徵存在三個弱點。

第一是組合的問題。世上的對象會以無窮無盡的新組合出現。如果像「紅蘋果」、「青蘋果」、「吃了一半的紅蘋果」這樣，每增加一種組合就必須新建一個專門節點，那麼所需的節點數量將呈幾何級數暴增。這被稱為「組合爆炸」。

第二是對損傷脆弱。如果負責祖母的單一神經元死亡，關於祖母的所有知識就必須全部消失。而在分散式表徵中，即使部分計算單元受損，資訊依然可以保留在剩餘的單元中。

第三是難以涵蓋相似性。貓和老虎很相似，而貓和冰箱並不相似。然而，如果每個概念都有一個各自獨立分開的節點，節點本身就無法說明這兩個概念有多相似。關於相似性的資訊必須透過個別規則一筆一筆記錄進去。

辛頓提出的替代方案是「分散式表徵」。分散式表徵不是將單一概念交由單一神經元負責，而是以無數神經元共同構成的活化模式來表現的方式。有些神經元強烈開啟，有些神經元微弱開啟，有些神經元則處於關閉狀態。其整體樣貌就是一個概念。用數學來說，概念就變成了由多個數字組成的向量。

分散式表徵可以比喻為交響樂團。交響曲的一個小節並不是單由一把小提琴獨自發出的聲音。小提琴、大提琴、長笛、小號各自以不同的強度一同奏響，營造出一個和弦。小節轉換時，樂器的組合與強度也隨之改變，同一件樂器在樂曲的多個部分不間斷地參與其中。大腦中的知識也如同這樣，以無數神經元共同奏出的和弦形式存在，這正是辛頓的想法。

正如交響樂團的樂器參與多個小節一樣，在分散式表徵中，單一神經元不只負責單一概念。神經元表現出諸如「有毛」、「用四隻腳走路」、「生活在室內」等微小特徵，並同時參與多個概念的表徵。這種微小特徵被稱為「微特徵」。由於貓和老虎共享許多微特徵，這兩個概念的活化模式便會彼此相似。相似性無需額外規則便自然融入表徵之中。它對損傷也具有韌性。即使少數幾個神經元受損，由其餘神經元構成的模式大部分仍會保留，因此知識不會在一瞬間全部消失。

辛頓也透過實驗展示了這一想法。這就是1986年的「家族關係圖任務」。辛頓建立了由12名英國人家族與12名義大利人家族、共24人組成的兩個家族族譜。這兩個族譜的結構完全相同。英國家族中有克里斯多福、潘妮洛普、夏綠蒂、維多利亞、詹姆士、柯林等名字

，而義大利家族則在對應位置有羅伯托、瑪麗亞等名字。關係包括父親、母親、丈夫、妻子、叔伯、姪甥等12種。

神經網路被要求輸入「人物1」與「關係」，並回答「人物2」。例如，若輸入「夏綠蒂」與「母親」，就必須回答「維多利亞」。人名是以24個欄位中只開啟1個欄位的方式輸入，關係則是以12個欄位中只開啟1個欄位的方式輸入。因此，輸入本身完全沒有關於人物之間有多相似的資訊。這兩個輸入分別穿過由6個神經元組成的層，在中間的隱藏層匯合，再經過幾個層得出答案。

神經網路完全沒有接收到任何諸如「若A的母親是B且B的丈夫是C，則C是A的父親」這樣的規則。給予的僅僅是表示事實的104個句子。辛頓用其中的100個來訓練神經網路，其餘4個則保留作為測試之用。

完成訓練後，辛頓檢視了代表人物的6個神經元。有些神經元能區分是英國家族還是義大利家族，有些神經元代表屬於第幾代，有些神經元則代表屬於族譜的哪個分支。神經網路自行找出了從未有人教過它的特徵。神經網路對於從未見過的測試題目，也給出了絕大多數正確的答案。

辛頓認為，誤差反向傳播學習法的真正價值正在於此。比起計算技巧本身，更重要的是隱藏層能夠自行學習反映資料內在結構的分散

式表徵。今日人工智慧中將詞彙轉換為數字向量的「詞嵌入」，其根源便在於這項實驗。

2.2. 齊聚拉荷亞的研究者們：1979年工作坊與PDP小組的誕生

1970年代後半期，神經網路研究者們如同孤島般分散各地。研究經費和教授職位大多流向了符號主義研究。從事神經網路研究的人們彼此並不清楚對方的存在，論文也分散發表在不受關注的學術期刊上。

辛頓的處境也沒有什麼不同。他於1978年在英國愛丁堡大學取得博士學位，但在英國很難以神經網路研究獲得研究經費。辛頓在取得博士學位前就曾在薩塞克斯大學從事研究，並於1978年10月前往美國加州大學聖地牙哥分校（UCSD）。

1979年，辛頓與布朗大學的心理學家詹姆士·安德森一起，在UCSD所在的拉荷亞舉辦了「平行聯想記憶模型」工作坊。這個工作坊成為分散各地的神經網路研究者首次齊聚一堂的契機。

參加者中有泰瑞·塞諾夫斯基。他當時是哈佛醫學院的博士後研究員，也是在知名度不高的學術期刊上發表過幾篇神經網路論文的年輕研究者。據說辛頓向麻省理工學院的視覺理論家大衛·馬爾詢問塞諾夫斯基的研究後邀請了他。在這次工作坊中相遇的兩人，日後成為共同打造波茲曼機的終身同僚。

工作坊的成果於1981年彙編成同名書籍《Parallel Models of Associative Memory》。編者為辛頓與安德森。辛頓在這本書中撰寫了探討如何在平行硬體上實現語意網路的章節。其內容指出，符號主義用圖形與符號所表示的語意關係，可以用神經元的活化模式來表現。

在拉荷亞有能夠延續這一潮流的人物，那就是UCSD的大衛·魯梅哈特與詹姆士·麥克蘭德。兩人在1980年代初期組成了「平行分散式處理（Parallel Distributed Processing）」研究小組，簡稱PDP小組。辛頓在此期間於UCSD擔任博士後研究員並與魯梅哈特熟絡起來，之後即便轉往英國劍橋與美國卡內基美隆大學，依然與PDP小組共同進行研究。

PDP小組的成長過程中，周遭研究者的支持與參與也十分重要。帶領UCSD認知科學領域的唐納·諾曼在系所內為研究者們提供了保護

傘，讓他們能夠自由展開與主流不同的想法。位於附近的索爾克研究所的法蘭西斯·克里克也經常出席聚會。克里克是揭示DNA雙螺旋結構的生物學家，此時已將關注點轉向大腦研究。他與研究者們探討神經網路模型與真實大腦究竟有多相似，日後更共同撰寫了PDP書籍第2卷中探討大腦皮質結構的章節。

PDP小組的研究於1986年由麻省理工學院出版社出版的兩卷套書《Parallel Distributed Processing: Explorations in the Microstructure of Cognition》進行了整頓與總結。第1卷探討基礎，第2卷則探討心理學與生物學應用。這部著作對心理學、腦科學、電腦科學領域產生了深遠影響，據傳作為學術專著創下了罕見的超過5萬冊銷量。

辛頓在第1卷的多個核心章節中名列其中。第3章「分散式表徵」由辛頓、麥克蘭德與魯梅哈特合著。第7章「波茲曼機的學習與再學習」由辛頓與塞諾夫斯基合著。第8章「透過誤差傳播學習內部表徵」由魯梅哈特、辛頓與羅納德·威廉斯合著。同年，這三人將介紹反向傳播學習法的論文發表在《自然》（Nature）期刊上。從拉荷亞展開的小型聚會，就這樣成為重新振興神經網路研究的核心。

2.3. 平行聯想記憶與權重最佳化的原理

拉荷亞工作坊標題中包含的「聯想記憶」，是連結主義與分散式表徵並列提出的另一個重要想法。若將這個概念與一般電腦檢索記憶的方式進行比較，就很容易理解。電腦是透過位址來尋找記憶的。必須知道幾號欄位中存放著什麼的確切位址，才能調出該內容，只要位址錯了一處，就無法找到想要的內容。這被稱為「位址參照記憶」。

人類的記憶運作方式則有所不同。即使只聽到歌曲第一小節的幾個音符，整首歌曲就會浮現在腦海中；哪怕只看到朋友半邊臉，也能認出是誰。這是僅憑內容的一部分便能檢索出整體的機制。這被稱為「內容參照記憶」。連結主義研究者們認為神經網路能夠自然而然地實現這種記憶。

以數學方式明確展現這一原理的人是物理學家約翰·霍普菲爾德。霍普菲爾德於1982年將物理學中用於解釋磁鐵等物質的「能量」概念引入神經網路。在霍普菲爾德網路中，神經元的每一種狀態都對應著一個確定的能量值。儲存的記憶對應於能量地形中凹陷的山谷。當輸入部分被抹去或混入雜訊的資訊時，神經元之間會互相影響，並逐步降低能量。就像球沿著斜坡滾落並停在谷底一樣，網路會抵

達最近的山谷並重現最初的記憶。無數神經元同時運動並尋找穩定狀態的過程，被稱為「平行鬆弛」。

霍普菲爾德網路存在弱點。能量地形中可能產生非預期的小山谷，從而陷入荒謬的記憶；或者受困於淺山谷中，無法找到最佳的解答。辛頓與塞諾夫斯基為了克服這個問題，加入了統計物理學的機率概念。若讓神經元不只按照既定規則運動，而是以一定機率產生擾動，網路就能從淺山谷中脫離出來，尋找到更深的山谷。以此打造的模型就是「波茲曼機」。大衛·艾克利、辛頓與塞諾夫斯基於1985年發表了包含波茲曼機學習方法的論文。

波茲曼機所解決的問題本質是「平行約束滿足」。當人類觀看物體時，大腦會同時協調邊界、深度、光線方向、顏色等無數條件，找出最合乎情理的解釋。這些條件蘊含在神經元與神經元之間連結的強度——即「權重」之中。如果兩個神經元同時開啟是自然的，那麼兩者之間的權重就很大；如果同時開啟是不協調的，權重就很小或是負數。

那麼，接下來的問題就是：這些權重究竟由誰來決定？連結主義給出的答案是：神經網路透過觀察資料自行決定。人類不需要用程式碼將規則寫入其中。神經網路藉由觀察大量的範例，逐步調整權重

，使自己產生的模式與真實世界的模式趨於一致。這就是權重最佳化，而這個過程本身就是學習。

辛頓對蘊含在權重中的記憶如何隨著時間變化也抱持著關注。1987年，他與大衛·普勞特共同發表了關於「快速權重」的研究。兩人在單一連結中設置了兩種權重。慢速權重經歷長時間緩慢變化，儲存關於世界的普遍知識。快速權重則在新資訊進入時迅速改變並迅速消失，負責暫時抓住剛剛看到的事物。

這種結構被用來處理神經網路學習中實際出現的一個問題。當神經網路學習新事物時，先前學過的事物經常會變得模糊。在兩人的實驗中，快速權重幫助僅透過短暫的練習便迅速恢復因新學習而變得模糊的舊記憶。論文標題「以快速權重消除舊記憶的模糊」正清楚表明了這一內容。快速權重這一構想在此後長年被遺忘，直到2016年辛頓與同僚重新提出才再度受到關注。

2.4. 看待心智視角的轉變：從符號操作者到資料學習者

關於分散式表徵與聯想記憶的研究並不僅止於技術問題，它為「該如何看待人類心智」這一古老問題提出了不同的答案。

符號主義與古典認知科學所描繪的人類是「符號操作者」。他們認為人類是按照既定規則依序處理腦中符號的邏輯機器。他們看到人類說話合乎文法、將事物劃分類別的模樣，便推論大腦中也預先寫好了文法規則與類別清單。

辛頓與連結主義研究者們所描繪的人類則截然不同。在他們眼中，大腦並非抄寫規則並依序計算的機器，而是能從無數經驗中自行萃取出相似性與模式的「資料學習者」。辛頓一向表示人類比起邏輯機器更接近「類比機器」。這意味著當面臨新情境時，人類不是逐一套用規則，而是想起過去經歷過的相似情境來做出判斷。

從這個觀點來看，合乎文法的表達或邏輯推論，並非因為大腦中寫有規則而產生。而是在無數神經元之間的權重中建立的分散式表徵相互作用，從而展現出表面上看似規則的行為。這被稱為「湧現」。正如家族關係圖實驗中的神經網路未接收任何親屬規則卻能推斷出親屬關係一樣，人類也可以在不背誦規則的情況下表現得像是遵循規則一般。

辛頓在近期的演講中也再三說明當時的研究與今日的人工智慧是一脈相承的。在2025年6月英國布里斯托大學的演講中，他介紹了1985年製作的小型語言模型。該模型將預測下一個詞彙時產生的誤

差作為學習訊號，將詞彙符號轉換為特徵活化，並讓這些特徵相互作用來學習預測下一個詞彙的特徵。他表示，可以將今日的大型語言模型視為這個小型模型的後代。他所說的小型模型，正是學習了家族關係圖的神經網路。

聯想記憶的研究也在數十年後重新獲得評價。2024年諾貝爾物理學獎頒給了霍普菲爾德與辛頓。瑞典皇家科學院闡明獲獎理由表示，霍普菲爾德創造了能夠儲存並重現影像等資料模式的聯想記憶，而辛頓則以霍普菲爾德網路為基礎，創造了名為波茲曼機的新型網路。1979年拉荷亞工作坊標題中的「聯想記憶」，在45年後再次出現在諾貝爾物理學獎的獲獎理由中。

在1970年代與1980年代，辛頓奔波於拉荷亞、劍橋、匹茲堡之間，奠定了分散式表徵與聯想記憶的理論。那時他的想法處於學界的邊緣。知識不是記錄在單一處的符號，而是無數神經元共同奏出的和弦；記憶不是透過位址而是透過內容來檢索；規則不是手動填入，而是從資料中學習——正是這三種想法，成為今日處理語言與概念的人工智慧之基石。

第3章 從錯誤答案中學習的神經網路

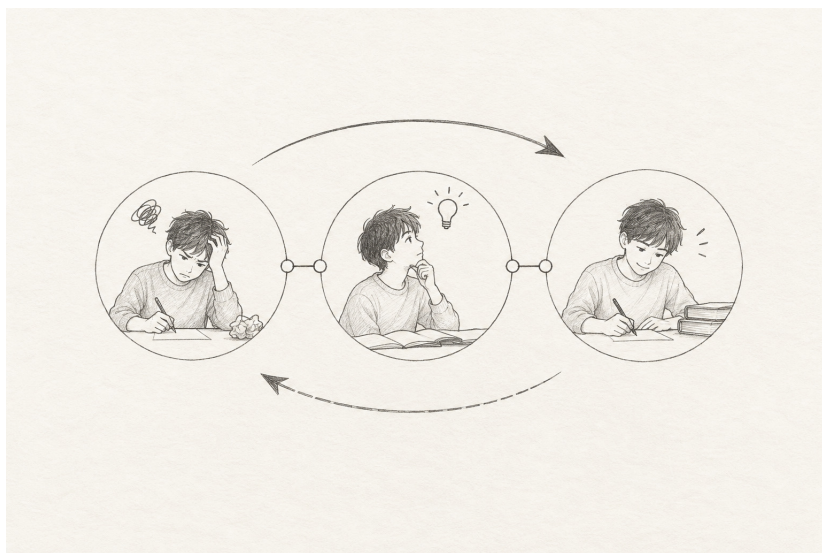


圖3. 比喻回溯錯誤結果以改進學習的過程。

3.1. 多層神經網路的難題與反向傳播（Backpropagation）的再發現

1969年，馬文·明斯基（Marvin Minsky）與西摩爾·派普特（Seymour Papert）出版了名為《Perceptrons》的著作。兩人在該書中以數學證明了僅有輸入層與輸出層的單層感知器無法解決如XOR等問題。該書出版後，許多研究者離開神經網路研究，轉向符號主義人工智慧。符號主義是由人類直接寫入規則，讓電腦依照該規則進行推理的方式。

突破單層感知器局限的方法並非完全不存在。理論上已知若在輸入層與輸出層之間加入隱藏層（Hidden Layer），便能解決更複雜的問題。問題在於該如何訓練這個隱藏層。輸出層可以直接與正確答案比較。當神經網路看著貓的照片回答「狗」時，可以立即計算出錯得有多離譜。隱藏層的情況則有所不同。隱藏層的神經元並未與外部世界直接相連，因此在任何時刻該輸出何種數值才正確，無人能予告知。當最終答案出錯時，要在眾多中間神經元中分攤並追究責任歸屬於何處、佔了多少責任的問題，稱為「信用分配問題」（Credit Assignment Problem）。只要無法解決這個問題，多層神經網路就只能停留在紙上的構想。

信用分配問題的答案正是誤差反向傳播。其原理在於微積分的連鎖律（Chain Rule）。首先在輸出層求出誤差。接著將這個誤差反向，也就是從輸出層向隱藏層方向逐層傳回，計算各個連接權重對整體誤差做出了多少貢獻。最後，朝著減少誤差的方向逐步修正權重。這最後一個步驟稱為梯度下降法（Gradient Descent）。這就如同蒙上雙眼下山時，僅確認腳下的坡度，一步步邁向較低處一樣。

反向傳播並非由某個人在一夕之間靈光乍現的發明。多位學者在彼此互不知情的情況下得出了相同的構想。1970年，芬蘭的塞波·林納因馬（Seppo Linnainmaa）在碩士論文中整理了今日稱為「反向模式自動微分」的綜合計算法。1974年，哈佛大學博士研究生保羅·韋伯斯（Paul Werbos）在博士學位論文《Beyond Regression》中提出可將此方法應用於神經網路的學習。到了1985年左右，大衛·帕克（David Parker）與法國的楊立昆（Yann LeCun）各自提出了類似的學習規則。

即便如此，在反向傳播的歷史中被視為轉捩點的，仍是1986年大衛·魯梅爾哈特（David E. Rumelhart）、傑佛瑞·辛頓（Geoffrey E. Hinton）與羅納德·威廉斯（Ronald J. Williams）的研究。三人的合作始於聖地牙哥。辛頓於1978年至1980年間在加州大學聖地牙哥分校（UCSD）擔任博士後研究員，此後也持續與同事交流，參

與了由魯梅爾哈特、詹姆斯·麥克萊蘭（James McClelland）等人領導的PDP（平行分散處理）研究小組。加州大學聖地牙哥分校在2024年的一篇文章中總結道，PDP小組的工作為反向傳播的發展奠定了基礎，而辛頓隨後與魯梅爾哈特、威廉斯藉由1986年的《自然》論文，使該演算法廣為人知。

這三人並非最早提出反向傳播的人。三人的功績並不在於率先寫出數學公式，而在於透過實驗證明了以反向傳播訓練的神經網路隱藏層，能夠自行產生人類未曾教導過的內部表徵（Internal Representation）。

辛頓本人也明確指出了這一點。在2018年接受馬丁·福特（Martin Ford）採訪時，他表示反向傳播的基本構想源自魯梅爾哈特，是魯梅爾哈特的發明。他與威廉斯只是協助完善了該方法，而他自己的貢獻在於證明了反向傳播能夠學習有意義的分散式表徵。每當媒體將他介紹為反向傳播的發明者時，辛頓總是這樣予以糾正。

3.2. 1986年《自然》論文與計算原理

1986年10月9日，國際學術期刊《自然》第323卷第533頁至536頁刊登了一篇四頁的論文。標題為〈Learning representations by back-propagating errors〉，譯為中文即是「藉由反向傳播誤差學習表徵」。作者所屬單位為魯梅爾哈特與威廉斯隸屬於UCSD認知科學研究所，辛頓則隸屬於卡內基美隆大學（CMU）。同年，三人於PDP小組出版的兩卷本著作《Parallel Distributed Processing》第1卷中，刊載了題為〈Learning internal representations by error propagation〉的長篇章節，詳細闡述了相同內容。簡短的論文吸引了學界的關注，而篇幅詳盡的書籍則扮演了讓研究者得以親自跟著實作的教科書角色。

該論文的計算過程可分為四個階段。前三個階段是神經網路產出答案並衡量該答案與正確答案偏離多少的過程，最後一個階段則是將如此衡量出的誤差分攤為每個神經元各自責任的過程。

第一階段是前向傳遞（Forward Pass）。訊號由下層往上層傳遞。上層的一個神經元與下層的多個神經元相連，每個連接都有代表其強度的權重。將下層各個神經元的輸出乘以對應連接的權重，並將這些數值全部相加，即成為上層神經元接收到的總輸入。論文中並

未另外單獨處理決定神經元何時被啟動的閾值。而是設立一個恆常輸出1的虛擬輸入，讓附屬於該輸入的權重發揮閾值的作用，從而以與其他權重相同的方式進行計算。

第二是活化函數。活化函數是將神經元接收到的總輸入轉換為該神經元輸出的規則。本篇論文採用了羅吉斯S型函數（Logistic Sigmoid Function）。該函數在總輸入為較大的負數時輸出接近0的值，為較大的正數時輸出接近1的值，介於兩者之間時則呈現平緩的S形曲線，輸出介於0與1之間的值。由於曲線連續且平滑、沒有中斷或折角，因此處處皆可微分。微分是求取當輸入發生微小變化時、輸出會產生多大變化的計算。S型函數的導數計算相當簡潔，只需將輸出值乘以「1減去輸出值」的結果即可。早期感知器的階躍函數只有開啟與關閉兩種狀態，數值會在某一點突變，因此無法使用這種微分。反向傳播是依據連鎖律——即藉由各步驟變化率的乘積來求出歷經多個步驟之總變化率的規則——來回傳誤差。為了讓這種方法順利運作，必須具備如此平滑的函數。

第三階段是誤差衡量。將神經網路產出的答案 $y(j)$ 與正確答案 $d(j)$ 之差進行平方並全部相加後取其一半。這個數值 E 即是顯示神經網路錯誤程度的分數。

第四階段是反向傳遞（Backward Pass）。此階段會求取每個神經元的誤差訊號。誤差訊號表示當該神經元接收到的總輸入發生微小變化時、整體誤差會產生多大改變的數值，相當於該神經元對誤差所承擔的責任大小。輸出層神經元的誤差訊號，是以實際輸出減去作為正確答案給定的目標值之差，再乘以S型函數的導數（即輸出值與「1減去輸出值」之乘積）而求得。隱藏層神經元的誤差訊號，則是將接收該神經元輸出的上層神經元們之誤差訊號分別乘以各個連接權重後全部相加，再將此和乘以該神經元本身的輸出值與「1減去該輸出值」之數值而獲得。這是依據連接強度成比例地將上層的責任分攤給下層的計算。如此一來，無論有多少層隱藏層，都能從輸出層開始逐層向下計算所有神經元的責任。就連無法直接與正確答案比較的隱藏層，也能被賦予誤差訊號。最後，朝著誤差減少方向修正各個權重，而單次修正的幅度則由稱為學習率（Learning Rate）的微小數值決定。

研究團隊亦探討了實際學習中發生的問題。其中之一是學習速度。梯度下降法是朝著誤差減少的方向逐步移動權重的方法。若將誤差隨權重改變的狀態比擬為地形，當地形呈現狹長山谷形狀時，梯度下降法會在山谷兩側岩壁間來回震盪，沿著山谷前進的速度則會變慢。研究團隊為減輕此問題而引入了動量項（Momentum）。原本的變化量是將誤差梯度乘以學習率後朝梯度相反方向移動的量，而

動量法則是將前一步驟中權重改變量的一部分加到其中。要反映多少前一步驟的變化量，由介於0與1之間的數值——動量係數來決定。這能產生如同滾動的球即使遇到微小坎坷，也能憑藉積累的速度越過去一樣的效果。這項技術被長期沿用。2013年，伊利亞·蘇茨克維（Ilya Sutskever）等人亦在深度神經網路學習中重新檢視了動量項的效果。

另一個擔憂則是關於局部極小點（Local Minima）。局部極小點是指誤差雖然低於周遭，但並非整體誤差最低之處的點。有人指出，梯度下降法一旦抵達比周遭更低之處，便無法找到進一步下降的方向，因此學習可能會停滯在淺窪處，而非最低的山谷中。論文報告指出，在許多任務中這種情況相當罕見。並且解釋道，若多加入幾個隱藏單元與連接，以每個權重作為座標軸的權重空間中就會產生新的維度，而該維度會成為繞過淺窪的出路。

3.3. 隱藏層自行學得的表徵：鏡像對稱判別與「家族關係圖（Family Trees）」任務

1986年的論文帶給學界的震撼，比起數學公式更來自於實驗。研究團隊揭開完成學習的神經網路隱藏層，以圖像展示了每個神經元究竟學到了什麼。

第一項實驗是鏡像對稱（Mirror Symmetry）判別。這是分辨由0與1組成的六位數陣列是否以中心為基準左右相同的問題。例如101101是對稱的，而101100則不對稱。這個問題只看單一位元是無法得知任何資訊的。必須將第一位與第六位、第二位與第五位成對進行比較。這與判斷1是否為奇數個的奇偶校驗（Parity）問題屬於同一類，是單層感知器無法解決的代表性任務。奇偶校驗問題本身在同年的PDP專書反向傳播章節中已另有探討。

研究團隊在僅配置兩個隱藏單元的神經網路中，向其展示全部64種可能的陣列進行訓練。將這64個陣列重複展示1,425次後，神經網路自行找到了一種俐落的解法。每個隱藏單元都在隔著中心相對稱的成對位置上，賦予了大小相等但符號相反的權重。從外側到內側，權重的大小構成了1比2比4的比例。

這種權重結構判別對稱的過程如下。若陣列完全對稱，左右數值會相互抵消，使得隱藏單元的總輸入為0。隱藏單元具有負偏差，因此總輸入若為0便會保持在關閉狀態，當兩個隱藏單元皆關閉時，輸出單元便回答「對稱」。若不對稱，兩個隱藏單元之一便會開啟，進而關閉輸出單元。多虧了1、2、4的比例，非對稱陣列絕不可能偶然形成總和為0的情況。人類完全沒有傳授這條規則。神經網路僅憑觀

察正確答案與錯誤答案，便將對稱這項幾何學規則轉化到了權重結構之中。

論文的核心是辛頓親自設計的家族關係圖任務。辛頓建立了英國人家族12人與義大利人家族12人、共計24人的族譜。兩份族譜的結構完全相同。例如在英國家族中克里斯多福與潘妮洛普夫婦的位置，對應著義大利家族中羅伯托與瑪麗亞夫婦的位置。關係共有父親、母親、丈夫、妻子、兒子、女兒、叔伯舅、姑嬸姨、兄弟、姊妹、姪子（男）、姪女（女）等12種。訓練資料為「人物1 + 關係 = 人物2」形式的104個命題。例如柯林的姑嬸姨為珍妮佛與瑪格麗特等等。

神經網路共由五層組成。最底層有表示人物1的24個單元與表示關係的12個單元。由於採用一次僅啟動一個單元的方式，單看輸入的話，英國的柯林與義大利的阿方索之間沒有任何共通點。其上方依序連接了接收人物的6個隱藏單元與接收關係的6個隱藏單元、中間的12個隱藏單元、輸出側的6個隱藏單元，以及表示人物2的24個輸出單元。

研究團隊在104個命題中僅以100個進行訓練，保留了4個。訓練結束後，檢視接收人物的第一層隱藏層的6個單元，發現每個單元都負責有意義的特徵。第1號單元區分了英國人與義大利人。第2號單元

表示該人屬於第幾代。第6號單元表示屬於族譜的哪一個分支。排除負責國籍的單元後來看，像柯林與阿方索這樣在兩份族譜中處於相同位置的人，獲得了相似的活化模式。國籍、世代、分支等概念，在輸入資料中完全未曾載明。這是神經網路在推斷關係的過程中自行找出的特徵。

神經網路亦未接收過諸如「母親的丈夫是父親」之類的規則。然而，它卻答對了在訓練時一次也未曾見過的4個命題的正確答案。正是因為具備將一個人以多種特徵組合表示的分散式表徵（Distributed Representation），這種泛化才成為可能。這與讓單一符號對應單一意義的符號主義方式截然相反。

這項實驗被評為展示了神經網路能夠以數值向量學習單字與概念意涵的早期典範。辛頓、楊立昆與約書亞·班吉歐（Yoshua Bengio）於2015年發表在《自然》期刊的深度學習綜述論文，以及2021年發表在《Communications of the ACM》上的文章皆說明，今日的詞嵌入（Word Embedding）正是此一構想的延伸。隨後出現的Word2Vec、Transformer、大型語言模型，也都在將單字轉換為向量來處理這一點上，名列於同一脈絡之中。

3.4. 卡內基美隆大學（CMU）時期與移居加拿大多倫多

1986年論文發表時，辛頓所屬的單位是美國賓夕法尼亞州匹茲堡的卡內基美隆大學。他於1982年成為該校電腦科學系教授，直至1987年共任職了五年。當時的CMU是由艾倫·紐厄爾（Allen Newell）與赫伯特·賽門（Herbert Simon）所領導的符號主義人工智慧中心。根據2019年CMU慶祝圖靈獎的文章，辛頓甚至曾與代表符號主義的紐厄爾共同講授課程。在觀點各異的同事之間，他堅持延續了連結主義的研究。

CMU時期是研究成果豐碩的時期。辛頓與大衛·阿克利（David Ackley）、泰倫斯·塞諾夫斯基（Terrence Sejnowski）借鑒統計物理學的概念研究波茲曼機（Boltzmann Machine），並於1985年在《Cognitive Science》上發表了學習演算法。同一時期，他與魯梅爾哈特、威廉斯共同完成了反向傳播論文。科克大學的喬治·布爾（George Boole）誕辰200週年紀念紀錄介紹道，辛頓在擔任CMU教授的五年間，開拓了反向傳播、波茲曼機以及詞彙分散式表徵的研究。

彼時美國的政治局勢朝著與辛頓想法相悖的方向發展。隆納·雷根（Ronald Reagan）政府在與蘇聯的冷戰中，推動了綽號為「星戰計畫」的戰略防禦構想（SDI）。國防部轄下的國防高等研究計劃署（DARPA）亦對包含人工智慧在內的運算研究資助了龐大預算。美國大學的人工智慧研究經費中，有相當大一部分來自軍方。

辛頓自認是一名社會主義者，反對將人工智慧運用於戰爭。他不願接受軍方的研究經費。《大英百科全書》（Britannica）記載，辛頓於1987年離開美國前往加拿大的決定中，對美軍與雷根政府的反感情緒發揮了作用，且他反對在戰鬥中使用人工智慧。亦有記載指出，對雷根政府政策的失望以及反對軍事資助人工智慧研究，是他移居的原因之一。

1987年，辛頓辭去CMU教職，轉赴加拿大安大略省的多倫多大學電腦科學系任教。接納他的機構是加拿大高等研究院（CIFAR）。

CIFAR是一個不論眼前的商業成果或軍事目的，長期資助基礎研究的機構。據CIFAR所述，辛頓於1987年作為研究員加入了CIFAR的第一個研究計畫「人工智慧、機器人學與社會」，並在CIFAR與加拿大自然科學與工程研究委員會（NSERC）的資助下，持續進行當時備受質疑且研究經費匱乏的領域之研究。

多倫多時期的前10年並不順遂。進入1990年代後，隨著支援向量機（Support Vector Machine）等其他機器學習方法受到矚目，神經網路再次被邊緣化。辛頓於1998年至2001年間，在倫敦的倫敦大學學院創立並領導了蓋茲比計算神經科學研究中心，之後重返多倫多。2004年，他與同事共同在CIFAR創立了新的「神經計算與適應性知覺（NCAP）」計畫。該小組成為楊立昆、班吉歐等神經網路研究者定期聚會交流的平台，並成為2012年辛頓實驗室的亞歷克斯·克里哲夫斯基（Alex Krizhevsky）與伊利亞·蘇茨克維所開發的AlexNet在影像辨識競賽中以懸殊優勢奪冠的基石。

1986年的合作者們隨後各自踏上了不同的道路。羅納德·威廉斯於1992年發表了強化學習的策略梯度技術REINFORCE演算法。2001年，辛頓成為以魯梅爾哈特命名的魯梅爾哈特獎的首屆得主，而魯梅爾哈特於2011年辭世。2024年，辛頓與約翰·霍普菲爾德（John Hopfield）因促成基於人工神經網路的機器學習之基礎發現與發明的功績，共同榮獲諾貝爾物理學獎。在宣布諾貝爾獎的2024年10月8日，CIFAR發表了回顧自1987年以來與辛頓深厚淵源的祝賀文章。

無法訓練多層神經網路的長年難題，藉由1986年的反向傳播論文開啟了解決之道。該論文留下的核心比起計算技巧，更在於思維的轉

變：即便人類未曾輸入規則，神經網路也能自行在資料中找出有用的表徵。辛頓秉持著這一構想，離開了軍事研究經費氾濫的美國前往加拿大，並在多倫多持續研究了20多年。今日的大型語言模型與影像生成模型亦以反向傳播進行訓練。然而，辛頓後來認為真實大腦不會以反向傳播這樣的方式學習，並於2022年提出了如 Forward-Forward 演算法等替代方案。

第二部. 讓神經網路學習的方法

第4章. 從物理學中找到的學習原理

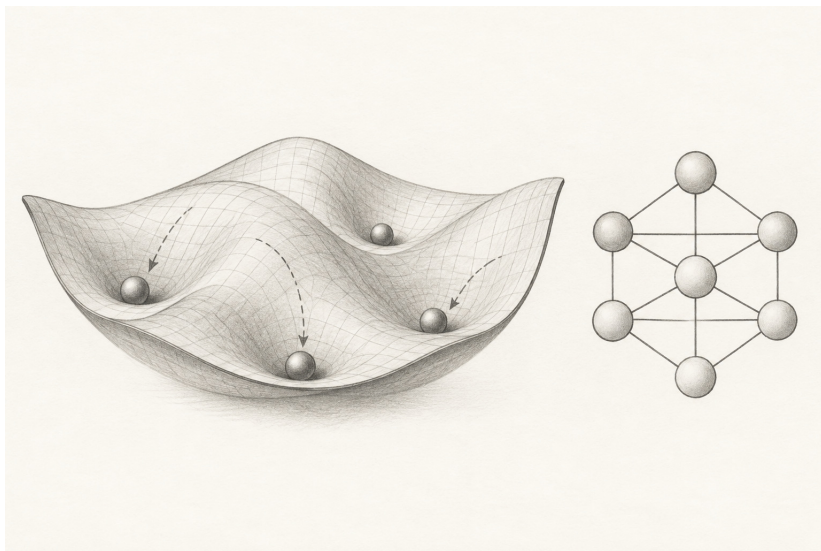


圖4. 比喻利用能量與機率的神經網路之思維。

4.1. 羅徹斯特的相遇：為霍普菲爾德網路增添機率與熱量

1980年代初，人工智慧學界對神經網路研究態度冷淡。1969年馬文·明斯基（Marvin Minsky）與西摩爾·派普特（Seymour Papert）在《感知器》（Perceptrons）中以數學方式證明單層神經網路的局限後，許多研究人員將神經網路視為死胡同。研究經費與人才紛紛湧向試圖以規則與符號實現智慧的符號主義人工智慧。在這一時期，為神經網路研究提供新動力的不是電腦科學家，而是一位物理學家。

約翰·霍普菲爾德（John J. Hopfield）是一位橫跨固態物理學與生物物理學的理论物理學家。1982年，當時任職於加州理工學院（Caltech）的他，在《美國國家科學院院刊》（PNAS）第79卷發表了題為〈具有湧現集體計算能力的神經網路與物理系統〉（Neural networks and physical systems with emergent collective computational abilities）的論文。霍普菲爾德後來轉往普林斯頓大學任教。

在這篇論文中，霍普菲爾德借鑑的工具是一種稱為自旋玻璃（spin glass）的物理模型。原子表現得如同微小磁鐵，而磁鐵的方向被稱

為自旋。自旋具有向上 (+1) 或向下 (-1) 兩種狀態之一。在一般磁鐵中，相鄰自旋皆朝同一方向排列。然而在自旋玻璃中，某些配對傾向於同向，某些配對則傾向於反向。由於不存在能同時滿足所有配對的排列方式，系統便停留在眾多妥協狀態之一。物理學家將這種穩定狀態表述為能量地形中凹陷的谷底。

霍普菲爾德在自旋的位置放上了神經元，因為神經元同樣具有開啟與關閉兩種狀態。如此對應之後，自旋之間的交互作用便成為神經元之間的連接強度（即權重），整個系統的能量成為網路狀態的能量，而能量谷底則成為儲存的記憶。霍普菲爾德在此設定了一個條件：連接必須是對稱的。無論選取哪兩個神經元，從一端到另一端的連接強度與反方向的連接強度必須相等。他證明了在此條件下，每當神經元逐一改變狀態時，整個網路的能量都會減少或維持不變。由於能量無法無止境地下降，網路最終必定停留在某個谷底底部。這個底部被稱為吸引子狀態（attractor state）。

利用這一特性，便能將網路用作記憶裝置。若將欲儲存的模式建構成谷底底部，即使輸入的內容部分遺失或混入雜訊，網路也會滑向最近的谷底以還原原始模式。這與只聽了歌曲的第一小節就能想起其餘歌詞類似。這種並非根據儲存位置的位址，而是透過內容的一部分來喚起記憶的方式，稱為內容定址記憶體（

content-addressable memory)。霍普菲爾德的論文以物理學的語言展現了神經網路能夠穩定運作，並成為物理學家對神經網路產生興趣的契機。

在紐約州羅徹斯特大學，舉行了一場由傑瑞·費德曼（Jerry Feldman）主持的工作坊。卡內基美隆大學（CMU）教授傑佛瑞·辛頓（Geoffrey Hinton）、結束哈佛醫學院博士後研究後任職於約翰霍普金斯大學的泰倫斯·塞諾夫斯基（Terrence Sejnowski），以及霍普菲爾德齊聚一堂。霍普菲爾德介紹了證明其網路必然收斂至穩定狀態的論證。

辛頓與塞諾夫斯基從這項結果中同時看到了潛力與弱點。弱點恰恰來自於保證收斂的這一特性。霍普菲爾德網路始終只朝著能量下降的方向運動，因此一旦落入與初始輸入相近的淺谷，就無法從中脫身。這個淺谷被稱為局部極小值（local minimum）。在局部極小值處獲得的答案，往往是從未儲存過的離譜模式，也就是虛假記憶。

兩人往返於匹茲堡與巴爾的摩之間，每逢週末便一同研究。他們匯集心理學、電腦科學、腦科學與統計物理學的知識，尋求克服此一弱點的方法。他們找到的答案是熱量。他們讓神經元在決定狀態時不再嚴格遵循既定規則，而是依據機率來決定。提高溫度會使雜訊

增加，神經元偶爾也會朝著能量上升的方向運動。如此一來，網路便能越過淺谷的障壁，尋找更深的谷底。

兩人以19世紀奧地利物理學家盧德維希·波茲曼（Ludwig Boltzmann）的名字為這個新網路命名。波茲曼是奠定統計力學基礎的大師，曾從無數粒子的隨機運動中推導出溫度與能量的定律。辛頓與塞諾夫斯基於1983年在電腦視覺學術會議上，以題為〈最佳知覺推論〉（Optimal perceptual inference）的論文首次發表了這項構想。

4.2. 能量地形、隨機鬆弛與隱藏單元

感知器與霍普菲爾德網路僅依既定規則運作。波茲曼機則將其轉變為以機率進行推論的能量基礎模型。能量基礎模型是指為網路可能處於的每種狀態各賦予一個稱為能量的數值，並將該數值較低的狀態視為更合理狀態的模型。其結構始於定義能量的規則。在波茲曼機中，每個神經元（單元）的狀態為代表關閉的0或代表開啟的1。整個網路的能量定義如下：

首先，將相互連接的兩個單元配成一對，將兩個單元的狀態與其間連接的權重相乘。只有當兩個單元皆處於開啟狀態時，該乘積才等

於權重；若任一單元處於關閉狀態，乘積即為0。將所有配對的此數值相加後反轉正負號，再加上所有處於開啟狀態單元的門檻值總和，即為整個網路的能量。

這條規則的涵義如下：以正權重強力連接的兩個單元若同時開啟，能量便會相應降低。這意味著兩個單元之間的關係越吻合，系統就越穩定。門檻值是每個單元開啟所需的輸入強度。門檻值較大的單元在開啟時會使能量相應升高，因此不易被開啟。

若讓該網路在恆定溫度下長時間運作，系統便會達到熱平衡（thermal equilibrium）。此處的溫度是決定單元運動隨機程度的數值，借鑑自物理學中的溫度概念。熱平衡是指即使個別單元的狀態持續變化，各個整體狀態出現的比例仍保持恆定的狀態。整體狀態指的是將所有單元的開啟與關閉同時記錄下來的單一排列。此時，網路處於某個整體狀態的機率遵循波茲曼分布。

在波茲曼分布中，每個狀態都被賦予一個數值，該數值隨著該狀態的能量除以溫度的數值增大而依指數函數急劇變小。某個狀態出現的機率，為該狀態被賦予的數值除以所有狀態被賦予數值的總和。經過這樣的相除計算，所有狀態的機率相加便等於1。

依據此分布，能量越低的狀態出現的機率越高，能量越高的狀態則極少出現。溫度調節著這種差異的大小。當溫度較高時，差異縮小，各種狀態均勻出現；當溫度較低時，能量最低的狀態比其他狀態出現得頻繁得多。神經元依據此種機率改變狀態、逐漸尋求穩定排列的過程，被稱為隨機鬆弛。

那麼，究竟該如何調節溫度才能找到深谷呢？辛頓與塞諾夫斯基引進了稱為模擬退火（simulated annealing）的方法。將金屬加熱後緩慢冷卻，內部結構便能趨於穩定狀態。1983年，IBM研究員史考特·柯克派屈克（Scott Kirkpatrick）等人將該原理應用於計算問題，在《科學》（Science）期刊上發表了模擬退火技術。一開始提高溫度，使網路能夠自由翻越能量地形的高丘，大範圍進行探索；接著極為緩慢地降低溫度，系統停留在地形中最深谷底（即全域極小值，global minimum）的可能性便會增加。大約同一時期，保羅·斯莫倫斯基（Paul Smolensky）的和聲理論，以及史都華·蓋曼（Stuart Geman）與唐納·蓋曼（Donald Geman）兄弟的影像去噪研究，也將相同的構想應用於其他問題。

波茲曼機的核心成就就在於包含了隱藏單元的神經網路學習方法。隱藏單元是不從外部接收資料也不向外輸出結果，僅在網路內部運作的單元。回顧明斯基與派普特的批評，便能明白為何需要隱藏單元。

感知器與霍普菲爾德網路僅能處理與外部世界直接接觸的單元之間的關係。然而，即使是看似簡單的互斥或（XOR）問題——兩個輸入相同時輸出0，相異時輸出1——單憑輸入之間的直接關係也無法解決。若要表達將兩個元素成對觀察時無法顯現、必須將多個元素綜合觀察才會顯現的高階關係，就需要能在中間梳理這些關係的單元。

波茲曼機將單元劃分為兩類，回應了這項需求。可見單元（visible units）與外部世界接觸，負責接收輸入資料或輸出解答；隱藏單元（hidden units）的值無法由外部指定，亦無法被直接觀察。隱藏單元與可見單元以對稱方式連接，能自行尋找並表徵隱藏在資料背後的規則與特徵。

剩下的問題在於如何訓練隱藏單元的權重。如何訓練無法告知標準答案的單元，是神經網路研究長期以來的課題。波茲曼機以機率與統計物理學為這一課題提供了答案，並為開闢多層神經網路的學習之路做出了貢獻。

4.3. 清醒階段與做夢階段：波茲曼機的學習演算法

1985年，大衛·阿克利（David Ackley）、辛頓與塞諾夫斯基在學術期刊《認知科學》（Cognitive Science）第9卷上發表了題為〈波茲曼機的學習演算法〉（A learning algorithm for Boltzmann machines）的論文。次年，辛頓與塞諾夫斯基在PDP研究小組出版的《平行分散處理》（Parallel Distributed Processing）第1卷第7章〈波茲曼機的學習與再學習〉中，詳細闡述了該內容。

波茲曼機的學習目標在於使兩個機率分布達成一致。機率分布是各個可能狀態出現頻率的集合。其一是真實世界的資料在可見單元上產生的分布，其二是網路在沒有外部輸入、單獨運作時在可見單元上產生的分布。可見單元是直接與外部世界接觸並接收資料的單元。兩個分布之間的差異以資訊理論中的庫爾貝克-萊布勒散度（Kullback-Leibler divergence）來衡量。作者們將該數值命名為G。

該數值是對可見單元可能具有的每種狀態進行以下計算後相加的總和：將該狀態在真實資料中出現的機率除以網路單獨運作時出現的機率，求得該比值的自然對數後，再乘以其在真實資料中的機率。

在兩種機率相同的狀態下，比值為1且1的對數為0，因此不會增加任何數值；一旦存在兩種機率不一致的狀態，總和便會大於0。

若該數值為0，則網路所產生的分布與真實資料的分布相同。這意味著網路如實再現了資料所蘊含的世界樣貌。因此，學習即是朝著縮小該數值的方向調整權重。辛頓與塞諾夫斯基證明，若將該數值對各個權重進行微分，也就是求出當微調某個權重時該數值會產生多大變化，便能得到極為簡潔的結果。

該結果表現為兩種機率之差。其一是將可見單元固定在真實資料上並達到平衡時，相連的兩個單元同時開啟的機率；其二是網路在沒有外部輸入、單獨運作並達到平衡時，這兩個單元同時開啟的機率。微調權重時兩個分布差異的變化幅度，等於前者機率減去後者機率的值除以溫度後再反轉正負號。

從這項結果中，可直接推導出修正權重的規則。

連接兩個單元的權重修正量，是兩個測量值之差乘以一個微小數值的結果。該差值是透過第一個數值減去第二個數值求得的。

此處相乘的微小數值決定了單次修正的幅度。第一個數值與第二個數值分別是在兩個不同階段所測得的數值。

第一階段為正向階段（positive phase），比喻為清醒階段。將可見單元固定於真實資料，僅讓隱藏單元自由活動。經由模擬退火達到熱平衡後，測量兩個單元同時開啟的平均機率。此數值為修正規則中的第一個數值，對應於觀察世界時所形成的連接。

第二階段為負向階段（negative phase），比喻為做夢階段。解除可見單元的固定，讓整個網路單獨運作。此時一旦達到熱平衡，同樣測量兩個單元同時開啟的平均機率。此數值為修正規則中的第二個數值，對應於網路在沒有外部輸入的情況下自行產生的連接。

比較這兩個數值，規則的意義便顯而易見。若觀察真實資料時兩個單元同時開啟的頻率高於做夢階段，則加強兩者之間的連接；若僅在做夢階段同時開啟，而在真實資料中並非如此，則削弱其連接。這正是強化在現實中所學到的東西，並抹去網路自行產生的毫無根據之聯想。

這項規則的優勢在於局部性。修正單一連接的強度時，所需的資訊僅限於該連接兩端的兩個單元的活動，無須從網路另一端反向接收誤差訊號。即使隱藏單元分置於多個層級，相同的規則依然適用。鑑於大腦的突觸同樣只能感知相鄰兩個神經元的活動，這項規則在生物學上顯得相當合理。

以睡眠為喻的解釋，亦與當時睡眠研究的假說不謀而合。1983年，法蘭西斯·克里克（Francis Crick）與格雷姆·米奇森（Graeme Mitchison）在《自然》（Nature）期刊上發表了關於做夢睡眠功能的假說，指出在快速動眼期（REM）睡眠期間，大腦會進行消除無用聯想的逆向學習（reverse learning）。波茲曼機的負向階段與該假說高度契合。克里克經常出席PDP小組的研討會，與神經網路研究人員展開討論。

然而，該演算法存在一個巨大的障礙：達到熱平衡所需的時間過於漫長。在正向階段與負向階段中，每個單元必須隨機更新數千至數萬次，才能獲得接近平衡的樣本。這種重複抽樣的過程稱為馬可夫鏈蒙地卡羅（MCMC）方法。由於這項龐大的計算成本，當時難以對大型神經網路進行實用性訓練。大約同一時期，辛頓於1986年與大衛·魯梅爾哈特（David Rumelhart）、羅納德·威廉斯（Ronald Williams）在《自然》期刊上發表的誤差反向傳播演算法，速度遠比此方法快得多。研究人員的關注點因此一度全面轉向反向傳播。

辛頓並未放棄波茲曼機。他持續研究透過限制單元間連接以減少計算量的方法。這項研究後來引申出消除了同層內部連接的受限波茲曼機（RBM）以及對比散度（CD）學習法。

4.4. 2024年諾貝爾物理學獎：統計物理學與人工智慧的相遇

在波茲曼機首次發表40餘年後的2024年10月8日，瑞典皇家科學院宣布將諾貝爾物理學獎共同授予約翰·霍普菲爾德與傑佛瑞·辛頓。獲獎理由為「促成以人工神經網路進行機器學習的基礎性發現與發明（for foundational discoveries and inventions that enable machine learning with artificial neural networks）」。當時，霍普菲爾德為普林斯頓大學名譽教授，辛頓為多倫多大學名譽教授。辛頓透露，他是在美國加州一家平價旅館中透過電話得知獲獎消息的。

諾貝爾委員會在發布資料中，將兩人的成就描述為一脈相承的脈絡。霍普菲爾德利用原子自旋的能量概念，建立了用於儲存與復原模式的網路。辛頓以此網路為基礎，加入統計物理學的工具，發明了能夠自行尋找與分類資料特徵、甚至生成新範例的波茲曼機。波茲曼機的名字直接列入了解釋獲獎理由的官方資料中。委員會委員安德斯·伊爾貝克（Anders Irbäck）在發表會現場指出，兩人的研究深深植根於物理學。委員會主席艾倫·穆恩斯（Ellen Moons）表示，辛頓被評價為引領高效率學習演算法開發的關鍵人物。

獲獎公布後亦引發了爭議。部分物理學家與線上學術社群爭論這項研究究竟是探究自然現象的物理學，還是電腦科學或統計學。若追溯學術譜系，兩人的研究植根於統計物理學的無序系統理論與自旋玻璃研究。2021年獲得諾貝爾物理學獎的義大利物理學家喬治·帕里西（Giorgio Parisi）之成就，同樣也是對自旋玻璃等無序複雜系統的研究。在解釋無數微小單元透過機率交互作用而展現宏觀秩序的過程上，帕里西的自旋玻璃與辛頓的波茲曼機共享著相同的科學問題。唯一的不同在於，在波茲曼機中，這種秩序是以記憶與學習的形式顯現出來。

獲獎後，辛頓多次被記者要求簡要解釋其研究內容。據悉，他借用了物理學家理查·費曼（Richard Feynman）的一句經典玩笑回答道：若是幾分鐘內就能解釋清楚的研究，便不配獲得諾貝爾獎。

2024年12月10日，頒獎典禮在斯德哥爾摩音樂廳舉行。辛頓從瑞典國王卡爾十六世·古斯塔夫手中接過諾貝爾物理學獎，而在多倫多大學，約100名師生與同仁齊聚一堂觀看頒獎典禮。獎金為1,100萬瑞典克朗，折合加拿大元約140萬元，由他與霍普菲爾德平分。辛頓宣布，將把自己所獲獎金的一部分捐贈給安大略省致力於提升加拿大原住民清潔飲水取得能力的機構「Water First」，以及協助神經多樣性青年的組織。

獎金的另一部分則用於紀念他與塞諾夫斯基共同進行的研究。神經資訊處理系統大會（NeurIPS）於2024年12月宣布，辛頓將以部分獎金設立一個新獎項。獎金為每年1萬美元，頒發給撰寫提出大腦運作方式新理論之論文且年齡在40歲以下、由兩名以上研究人員組成的團隊。獎項名稱為塞諾夫斯基-辛頓獎（Sejnowski-Hinton Prize），取自計算神經生物學家泰倫斯·塞諾夫斯基與辛頓的名字。40餘年前奔波於匹茲堡與巴爾的摩之間共同打造波茲曼機的兩人名字，如今並列於為年輕研究人員設立的獎項上。

辛頓亦在諾貝爾獎博物館留下了與波茲曼機相關的物品。他在2024年捐贈的微晶片中包含著早期的波茲曼機。該晶片由約書亞·阿爾斯佩克特（Joshua Alspector）於1987年製作，展示了將波茲曼機實作於專用晶片上的可行性。晶片中的波茲曼機是僅有6個節點的微型網路，其用途固然有限，但作為驗證原理的實例，為機器學習與人工智慧的發展指明了方向。

1980年代初，在絕大多數學者對神經網路嗤之以鼻之際，霍普菲爾德以自旋玻璃證明了神經網路的穩定性。辛頓與塞諾夫斯基在此基礎上增添熱量與機率，並加入隱藏單元，創造出能夠自行學習的機器。波茲曼機由於速度過慢，當時未能廣泛應用；然而，以能量表徵資料、縮小現實與模型之間的差距，以及自行尋找隱藏結構的構

想，歷經受限波茲曼機與深度信念網路，一路延續至今日的生成式人工智慧。源自物理學語言的這一構想，在40年後獲得了諾貝爾物理學獎的肯定。

第5章. 更快速的學習方法

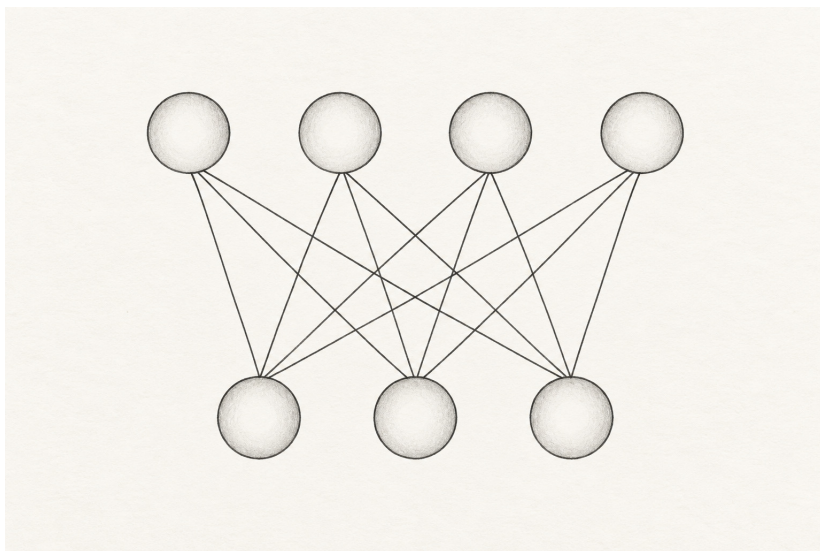


圖 5. 由兩層之間的連接所構成之受限玻爾茲曼機概念圖。

1980 年代問世的玻爾茲曼機展現了擁有隱藏單元的神經網路也能自行學習。然而，由於計算耗時過長，難以應用於實際問題。此後，辛頓執著於這個問題近 15 年，並給出了兩個解答。其一是刻意減少神經網路連接的受限玻爾茲曼機（Restricted Boltzmann Machine，RBM），其二則是讓這種神經網路快速學習的對比散度（Contrastive Divergence，CD）。這兩個構想皆源自同一個問題：玻爾茲曼機為何如此緩慢。

5.1. 過於緩慢的玻爾茲曼機與「專家乘積」的構想

神經網路是由稱為單元的微小計算單位彼此相連而成的結構。本章中的單元大多取代表開啟的 1 與代表關閉的 0 兩者之一的值。連接兩個單元的連線上附有一個稱為權重的數值。若權重為正數，兩個單元傾向於同時開啟；若為負數，則一側開啟時另一側傾向於關閉。學習就是逐步修正這些權重的過程。機率是用 0 到 1 之間的數值來表示某件事發生的頻率，越接近 1 表示發生得越頻繁。

傑佛瑞·辛頓與泰倫斯·謝諾夫斯基於 1980 年代中期發表的玻爾茲曼機，是將統計物理學概念引入神經網路的模型。該模型的單元並非按照固定規則，而是依據機率開啟或關閉。模型會記錄哪些單元

開啟、哪些單元關閉，為神經網路整體的每種狀態賦予一個能量值，並規定能量越低的狀態出現的機率越高。

學習分為兩個階段。首先，在將真實數據固定於輸入端單元的「清醒階段」中，測量相連的兩個單元同時開啟的程度。接著，在解除輸入、讓神經網路自由運行的「做夢階段」中，重新測量相同的值。隨後，根據這兩個值的差值來修正該連接的權重。清醒時更常同時開啟的連接會變強，做夢時更常同時開啟的連接則會變弱。

問題出在做夢階段。若要計算某筆數據出現的機率，必須知道配分函數（Partition Function），該值通常記作 Z 。 Z 是將神經網路可能具有的所有狀態的能量代入指數函數，轉換為能量越低值越大之後，將所有值相加的總和。某個狀態的機率是該狀態的值除以 Z 的商，因此沒有 Z 就無法確定機率。然而，由於一個單元具有兩種值，狀態的數量便是 2 乘以單元數量的次方。即使單元總共只有 100 個，狀態數也達到 2 的 100 次方，即約 10 的 30 次方。逐一相加如此龐大數量的項，是任何電腦都無法做到的。

因此，研究人員並未直接計算 Z ，而是採用抽樣方法。抽樣是依照機率實際抽取數值的過程，也稱為取樣。藉由彙整多次抽取的結果，無需檢視所有狀態，便能得知哪些狀態經常出現。玻爾茲曼機採

用了名為馬可夫鏈蒙地卡羅（MCMC）的抽樣方法與模擬退火（Simulated Annealing）。這種方式依據機率反覆改變單元的值，長時間擾動神經網路，直到各狀態出現的比例不再改變的熱平衡狀態，接著測量當時的統計量。然而，每修正一次權重都必須經歷這個過程，以 1980 年代和 1990 年代的電腦而言，學習過程變得漫長無比。辛頓在 2024 年諾貝爾獎演講中也回顧道，最初的玻爾茲曼機太慢了，數年後他透過限制神經元之間的連接來提高效率，構思出了受限玻爾茲曼機。

為了克服這個問題，辛頓從構建機率模型的方式開始重新審視。其成果便是「專家乘積（Products of Experts, PoE）」。這項研究出自辛頓於 1998 年至 2001 年擔任倫敦大學學院（UCL）蓋茨比計算神經科學部首任主任的時期。該研究於 2000 年率先以技術報告公開，隨後於 2002 年正式刊登於學術期刊《Neural Computation》第 14 卷第 8 期。

這裡的專家是指為每筆數據評估其合理程度（或然率）的小型模型。在所有可能數據上機率如何分佈，統稱為機率分佈。當時廣泛使用的是混合模型（Mixture Models），如高斯混合模型（GMM），即將多個專家的分佈相加使用。這種方式是為每個專家設定權重比例，將各專家評定的機率乘以該權重比例後全部相加。

辛頓指出，這種方式並不適合高維度數據，也就是需要許多數值才能表示單一實體的數據。以人臉照片為例，光線、姿勢、表情、角度等因素相互交織，形成約 35 維的彎曲空間（流形）。若要用混合模型表示這個空間，每個專家必須廣泛覆蓋整個 35 維空間。然而，將多個分佈相加的結果，其尖銳程度絕不可能超過作為材料的原分佈。最終，整體分佈會變得平鈍模糊。

專家乘積則相反，它不將專家評定的機率相加，而是相乘。相乘的值本身無法直接作為機率，因此需針對所有可能數據計算相同的乘積並求和，再用該總和除以乘積，使整體總和為 1。這個過程稱為重新歸一化（再規格化）。在相乘的方式中，只要有一位專家給出較低的機率，結果就會大幅降低。因此，每個專家只需針對數據的部分特徵設定條件即可。因為只有在同時滿足所有條件的狹窄區域內，機率才會維持在高水平。

辛頓以句子為例進行說明。一位專家只檢查時態是否正確，另一位專家則只檢查主詞與動詞的單複數是否一致。將兩者相乘，只有同時遵守這兩條規則的句子才會被賦予高機率。在混合模型中，每位專家都必須承擔問題的全部；但在專家乘積中，各自負責狹窄條件的專家們能夠共同建構出銳利的分佈。

專家乘積還有另一個優勢。一旦給定數據，隱藏在各個專家內部的變數之間便會相互獨立。獨立意味著即使知道一側的值，另一側的機率也不會改變；由於這是在給定數據的條件下成立的，因此稱為條件獨立。得益於此，只需一次計算即可精確推斷出隱藏變數的狀態。然而，在重新歸一化中作為除數的值仍需對所有情況求和，因此依然難以計算。辛頓透過 5.3 節將說明的對比散度避開了這個殘留的問題。

5.2. 受限玻爾茲曼機（RBM）的結構與條件獨立

將專家乘積的構想轉化為神經網路結構的，正是受限玻爾茲曼機。這種結構本身並非由辛頓首創，其淵源可追溯至 1986 年。保羅·斯摩稜斯基（Paul Smolensky）在當年出版的《平行分散處理（Parallel Distributed Processing）》第 1 卷第 6 章中介紹了「和聲理論」，並提出名為和聲機（Harmonium）的模型。1992 年，約阿夫·弗羅因德（Yoav Freund）與大衛·豪斯勒（David Haussler）研究了利用雙層神經網路學習二元向量（即 0 與 1 排列而成的數據）分佈的方法。辛頓在 2002 年的論文中指出，這種結構正將每個隱藏單元視為專家的專家乘積，並將其與玻爾茲曼機的能量架構相結合，建立了學習理論。

RBM 僅由兩層構成。可見層（visible layer）是接收數據的層，隱藏層（hidden layer）則是尋找數據中未直接顯現之特徵的層。隱藏層的單元，即單一隱藏單元，被訓練為在可見層出現特定模式時開啟。可見單元與隱藏單元之間的連接具有對稱權重，無論信號傳遞方向如何都採用相同數值。每個單元還各有一個稱為偏置（bias）的數值，用以決定該單元在不考慮其他單元的情況下有多容易開啟。

RBM 與一般玻爾茲曼機的不同之處在於，它完全消除了同層內部的連接。可見單元之間、隱藏單元之間互不相連。在數學上，這種形狀被稱為二分圖（Bipartite Graph）。若用教室來比喻，就像是規定前排學生只能與後排學生交談，而不能與同排旁邊的同學交談。

RBM 的能量是將三項相加後取反號所得的值：開啟的可見單元之偏置、開啟的隱藏單元之偏置，以及兩端單元同時開啟之連接的權重。具有較大權重的連接若兩端單元同時開啟，能量就越低；依照玻爾茲曼分佈的規則，該狀態的機率便越高。

消除同層內部連接的效果十分顯著。一旦可見層的值全部確定，隱藏單元之間便不再互相影響。這是因為隱藏單元之間沒有直接相連的路徑，而唯一的通道可見層的值已經固定。每個隱藏單元開啟的機率，僅憑傳入自身的信號即可單獨計算。將與已開啟的可見單元

相連的所有權重相加，再加上自身的偏置，然後將該總和代入 Sigmoid 函數。Sigmoid 函數無論代入何種數值都會返回介於 0 與 1 之間的值，代入的數值越大，產生的值越接近 1。這個值即為該單元開啟的機率。

反方向亦然。一旦隱藏層的值全部確定，可見單元之間便互為獨立，每個可見單元開啟的機率同樣由來自已開啟隱藏單元的權重總和與自身偏置以相同方式計算。因此，整個層取某一特定值的機率等於各單元機率的乘積。這稱為階乘條件獨立（Factorial Conditional Independence）。這種獨立性僅在對側層的值全部給定的條件下才成立。若沒有該條件，同層的單元仍會透過對側層相互影響。

若與有向圖模型相比，這項特性的價值便格外明顯。在箭頭由原因指向結果的模型中，會產生稱為「消除解釋（Explaining Away）」的現象。試想草地潮濕的兩個原因：下雨和灑水器。原本這兩個原因互不相干。然而在得知草地已濕的事實後，一旦確認下了雨，灑水器曾運作的可能性便隨之降低。也就是說，觀察到結果後，各原因之間會互相交織牽制。在這種模型中若要計算隱藏原因的機率，必須藉助 MCMC 或複雜的近似計算。

在 RBM 中則不會產生這種糾纏。輸入數據一旦進入，只需進行一次矩陣與向量的乘法運算和一次 Sigmoid 計算，即可精確獲得隱藏層整體的機率。不過，這僅限於從給定數據中讀取隱藏特徵的推論過程。在學習權重時，Z 依然是難以跨越的絆腳石。

5.3. 對比散度（CD）學習法的原理

RBM 的學習是指修正權重以使真實數據出現的機率增大的過程。究竟該朝哪個方向修正哪個連接，取決於兩個平均值之差。其一是將數據固定於可見層時，該連接的可見單元與隱藏單元同時開啟的比例；其二是在沒有數據、神經網路自由運行時，兩個單元同時開啟的比例。若前者大於後者，則提高權重；若前者小於後者，則降低權重。當所有連接上的兩個值都相等時，模型自行產生的統計量與數據的統計量一致，便無需進一步修正。

前者可透過 5.2 節的 Sigmoid 計算輕易求得。困難的是後者。若要精確獲取該值，必須進行吉布斯抽樣（Gibbs Sampling）。這是一種交替往復的方法：以可見層的值計算隱藏層的機率並抽取隱藏單元的值，再以此值重新計算可見層的機率並抽取可見單元的值。然而，這種往返必須持續至達到平衡狀態為止，在理論上需要無限次反覆進行。雖然減少連接使推論速度加快，但學習的瓶頸依然存在。

辛頓果斷斬斷了這個瓶頸。他認為無需等到達到平衡狀態，只要從真實數據出發，僅進行一次吉布斯抽樣便停止，也能獲得足具實用價值的學習方向。僅重複一次的方式稱為 CD-1，其流程分為四個步驟。

第一步，將真實數據輸入可見層。計算每個隱藏單元開啟的機率，並依照該機率抽取 0 或 1 的值。此時，在每個連接上測量可見單元與隱藏單元同時開啟的程度。

第二步，從抽得的隱藏層數值反向計算可見層的機率，生成新的可見層數值。這是神經網路以自身方式重新描繪數據的結果。辛頓將此稱為重構（reconstruction），有時也稱之為含有「編造故事」之意的 confabulation（虛構）。

第三步，將重構結果重新代入可見層以獲得隱藏層的值，並以相同方法測量兩端同時開啟的程度。

第四步，第一步測得的值減去第三步測得的值，將其差值乘以學習率，以此幅度修正權重。學習率是決定單次調整幅度大小的微小數值。在數據中更常同時開啟的連接權重上升，在重構中更常同時開啟的連接權重則下降。

若聯想到能量地形，便很容易理解此流程為何奏效。將橫軸設為所有可能數據，縱軸設為能量。訓練良好的模型會在真實數據所在之處形成深谷。重構並非隨意選取的點，而是在真實數據附近，朝著模型目前認為更合理的方向踏出一步的點。辛頓將其比作「惋惜的失誤（near miss）」。在權重修正中，增加的一方降低了真實數據所在位置的能量，減少的一方則提高了其旁邊失誤位置的能量。反覆進行這個過程，能量地形便會沿著數據被刨出深深的溝壑。這意味著模型不必全面考察遙遠之處，只需針對自己容易出錯的地方加以修正。

辛頓也為此方法提供了數學依據。該依據涉及三種分佈：真實數據的分佈、模型達到平衡時的分佈，以及從數據出發經歷指定次數吉布斯抽樣後的分佈。兩種分佈的相異程度由稱為庫爾貝克-萊布勒（KL）散度的值來衡量；兩分佈若相同則為 0，差異越大值越大。旨在最大化數據出現機率的常規最大概似學習，會縮小數據分佈與平衡分佈之間的散度。而對比散度則取而代之縮小這兩種散度之差，即以數據分佈與平衡分佈之間的散度，減去經抽樣後的分佈與平衡分佈之間的散度。

隨著抽樣反覆進行，分佈越接近平衡狀態，因此被減去的數值不會大於前者的數值。由此，這個差值不會小於 0；當模型完美重構數

據，使得經抽樣後的分佈與數據分佈相同時，差值變為 0。「對比散度」之名正是源自這種對照比較兩個散度的做法。

CD-1 還具有統計學上的優勢。真實數據與其重構結果始終成對綁定。如同在臨床試驗中將相似病患配對比較可減少偶然誤差一樣，成對比較減少了抽樣結果的波動（變異數）。因此，即使只進行一次吉布斯抽樣，學習過程也能穩定推進。

對比散度是一種近似方法，並非完美無缺。由於用退後一步的重構替代了本應在平衡狀態下測量的平均值，修正方向可能會偏離提高數據機率的精確方向。這一局限性在辛頓本人及其同事的後續研究中顯露出來。米格爾·卡雷拉-佩皮尼揚（Miguel Carreira-Perpiñán）與辛頓在 2005 年的研究中表明，CD 雖然可能走向與真實最大概似解略有偏差的位置，但這種偏差通常很小。伊利亞·蘇茨克維與泰門·蒂勒曼（Tijmen Tieleman）於 2010 年分析指出，CD 的修正方向並非沿著某個單一目標函數精確下降的梯度。這意味著它並非先設定一個想要縮小的值、再朝著該值下降最快的方向移動的方法。為了彌補這一弱點，蒂勒曼於 2008 年提出了持續對比散度（Persistent CD, PCD），不再每次重新開始吉布斯抽樣的馬可夫鏈，而是接續運行。辛頓則於 2010 年撰寫了《RBM 學習

實用指南》，整理了學習率、隱藏單元數量、權重初始值等實際設定方法。

5.4. RBM 的變形與向預訓練的擴展

RBM 與 CD 的結合超越了由 0 與 1 構成的模式，將應用範圍拓展至多種類型的數據。在此過程中，有三項轉變尤為顯著。

第一項轉變是改變可見單元的高斯-伯努利 RBM（Gaussian-Bernoulli RBM）。若要處理語音頻率特徵（MFCC）或照片亮度等連續實數值，可見單元就不能僅取 0 與 1。該模型將可見單元改換為混有高斯雜訊的線性單元。可見單元的值從常態分佈中抽取，其平均值為來自已開啟隱藏單元的權重總和加上自身偏置所得之值。該模型後來成為語音辨識研究的起點。

第二項轉變是改變隱藏單元。維諾德·奈爾（Vinod Nair）與辛頓於 2010 年國際機器學習大會（ICML）上發表了採用線性整流單元（Rectified Linear Unit, ReLU）的 RBM。兩人構想了一種將權重相同而僅偏置略有差異的二元單元無限堆疊的「階梯式 Sigmoid 單元」。該單元的運作可近似為一種函數：當輸入小於 0 時輸出 0，當輸入大於 0 時輸出與輸入成比例的值。在此基礎上加入雜訊，當

輸入與雜訊之和代入後大於 0 則輸出該總和，否則輸出 0。這種 ReLU 於 2012 年應用於 AlexNet，此後成為深度學習的標準激勵函數。

第三項轉變是層層堆疊 RBM 的方法。2006 年，辛頓、西蒙·奧辛德羅（Simon Osindero）與鄭耀威（Yee-Whye Teh）在《Neural Computation》第 18 卷發表的論文中，提出將第一個 RBM 的隱藏層輸出作為第二個 RBM 的數據進行訓練的方式。這種利用無標籤數據逐層依序訓練的方法稱為逐層貪婪非監督式預訓練（Greedy Layer-wise Unsupervised Pre-training），以此建構的神經網路稱為深層信念網路（Deep Belief Network, DBN）。同年，辛頓與魯斯蘭·薩拉赫丁諾夫（Ruslan Salakhutdinov）在《Science》第 313 卷刊登的論文中，利用 RBM 預訓練將輸入壓縮為簡短編碼後再復原的神經網路——深層自動編碼器，隨後透過誤差反向傳播進行微調。結果顯示，在將高維度數據壓縮為簡短編碼方面，該方法取得了優於傳統方法的成效。2009 年，兩人將理論擴展為所有層皆由無向對稱連接相連的深層玻爾茲曼機（Deep Boltzmann Machine, DBM）。

RBM 預訓練隨後在語音辨識領域取得了成果。多倫多大學、微軟、Google、IBM 的研究團隊於 2012 年聯合撰寫的論文報告指出，以 DBN 方式初始化的深層神經網路在聲學模型中超越了傳統方法。

此後情況發生了變化。隨著 ReLU、Dropout、大規模數據與圖形處理器（GPU）齊備，即使沒有預訓練，也能從一開始就透過誤差反向傳播來訓練深層神經網路。如今在實際服務中已罕見使用 RBM。介紹辛頓 2024 年 12 月 8 日諾貝爾獎演講的文章將該演講概括為引發深度學習革命的「歷史催化劑」故事，同時提及玻爾茲曼機學習演算法如今已不再使用。如同引發化學反應後不殘留在產物中的催化劑一般，RBM 開啟了深層神經網路學習的大門後功成身退。辛頓在諾貝爾週期間將一枚郵票大小的早期玻爾茲曼機晶片捐贈給了諾貝爾獎博物館。

雖然 RBM 本身已不再使用，但 RBM 與 CD 留下的思想仍在延續。與其計算所有可能情況以求得完美機率，不如透過降低真實數據的能量、提高似真假數據的能量來雕琢模型。楊立昆等人已將這種方法歸納整理為基於能量的模型（Energy-Based Model）架構。亦有研究人員認為，當今透過對照真配對與假配對進行訓練的對比學習（Contrastive Learning），同樣分享著這一問題意識。

辛頓針對玻爾茲曼機的速度問題給出了三個解答。他不將機率相加而是相乘，藉由專家乘積開闢了銳利表達高維度數據的途徑；透過消除同層內部連接的 RBM，實現了一次性精確推斷隱藏特徵；並且無需等待達到平衡狀態，藉由僅進行一步重構並加以比較的對比散度，跨越了學習速度的壁壘。這種減少連接並減少計算的抉擇，成為 2006 年後深度學習重獲矚目的墊腳石。

第6章. 深度神經網路重獲關注

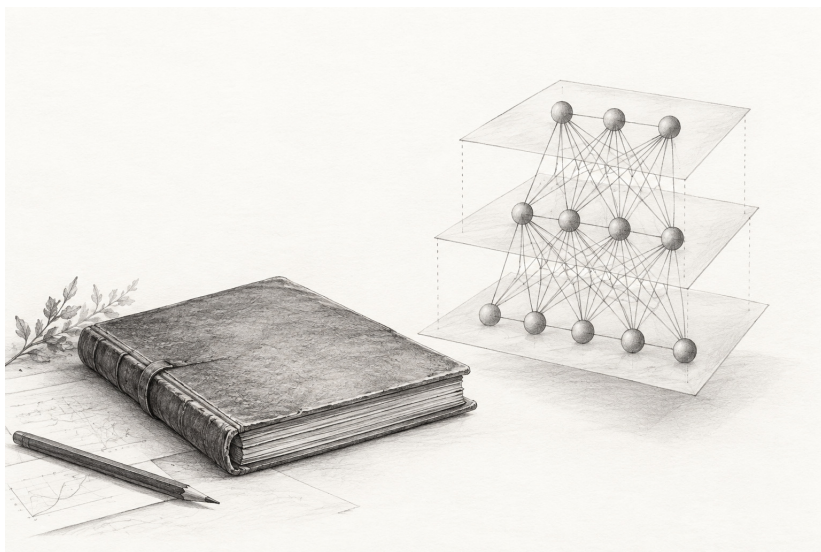


圖 6. 呈現了研究由多層構成之神經網路的思維。

6.1. 第二次 AI 寒冬與梯度消失問題

1986 年大衛·魯梅哈特、傑佛瑞·辛頓、羅納德·威廉斯在《自然》發表誤差反向傳播 (backpropagation) 方法後，對神經網路的期待大幅提高。誤差反向傳播是先計算神經網路給出的答案與標準答案之間的差距，再將該差距從輸出層反向傳送至輸入層，逐步修正連接強度（即權重）的方法。在只有兩三層的神經網路中，這種方法運作良好。然而，當把相同的方法應用於輸入層與輸出層之間堆疊多層隱藏層的深度神經網路時，訓練卻無法順利進行。當時阻礙深度神經網路訓練的問題主要有三個。

第一個障礙是梯度消失 (vanishing gradient) 問題。在反向傳播中，誤差訊號每穿過一層，就會乘以該層的微分值。當時廣泛使用的 Sigmoid 函數，即使在微分值最大時也僅有 0.25。當多次乘以小於 1 的數時，數值便會趨近於 0。因此，遠離輸出層的較底層幾乎接收不到學習訊號。相反地，數值每穿過一層就成倍增大、導致訓練不穩定的梯度爆炸也時有發生。無論是哪一種情況，就未能將適當大小的學習訊號傳遞到底層而言，本質上都是同樣的問題。

第二個問題是最佳化的失敗。最佳化是尋找使誤差變小的權重的過程。若將誤差隨權重起伏的樣貌繪製成地形，深度神經網路的誤差

地形則是起伏曲折且複雜多變的。從隨機設定的權重出發的深度神經網路，往往會在這片地形的局部極小點（local minimum）或平坦的鞍點（saddle point）附近徘徊後停滯不前，無法找到良好的權重。局部極小點是指比周圍低、但並非全域最低的點。鞍點則是在某些方向是上坡、在其他方向是下坡，因而該處的斜率接近 0 的點。

第三個問題是因數據不足所導致的過擬合（overfitting）。由人工逐一標註答案的數據，亦即標籤數據，在當時極難取得。權重達數十萬至數百萬個的深度神經網路，甚至連少量數據中的偶然噪聲都死記硬背了下來。這就如同練習題能答得很準確，但遇到從未見過的新問題時成績便大幅下滑的現象。

由於這些緣故，機器學習學界的關注焦點轉向了其他方法。支援向量機（SVM）與核方法能夠將學習問題轉換為凸最佳化（convex optimization）問題。在凸最佳化中，誤差地形只有一個山谷，因此始終能達到相同的最佳解。這是與容易陷入局部極小點的神經網路相比的一大優勢。將多個決策樹組合在一起的隨機森林也展現出了優異的效能。相比之下，神經網路被認為是一種理論上聽起來不錯、但實際上難以駕馭的方法。

結果，從 1990 年代中期到 2000 年代初期，人工神經網路研究經歷了漫長的低潮期。在人工智慧歷史上，1980 年代後期對專家系統的期待破滅後到來的低谷期，被稱為第二次 AI 寒冬。神經網路研究者所經歷的寒冬比這更加漫長。直到 2000 年代中期，主流學界都不把神經網路視為嚴肅的研究對象。辛頓與同事們在後來的訪談中回憶道，當時如果在論文標題中加入「神經網路」一詞，就很難通過學術會議的審查。

在這一時期，支援神經網路研究得以延續的機構是加拿大高等研究院（CIFAR）。2004 年，多倫多大學的辛頓與蒙特婁大學的約書亞·班吉歐、紐約大學的楊立昆、多倫多大學的大衛·弗利特等人共同提出了一項新的研究計畫。CIFAR 採納了這項提議，設立了「神經計算與自適應感知（NCAP, Neural Computation and Adaptive Perception）」研究計畫，辛頓領導該計畫約 10 年之久。該計畫後來更名為「機器與大腦學習（Learning in Machines & Brains）」。

NCAP 僅憑數頁篇幅的簡短研究計畫書，就提供以 3 至 5 年為單位的研究經費。他們既不要求立即取得商業成果，也讓以辛頓、班吉歐、楊立昆為核心的小規模研究者群體得以定期聚會交流想法，並透過舉辦暑期學校培育青年學者。找到深度神經網路訓練方法的 2006 年研究成果，正是誕生於這個研究社群之中。

6.2. 逐層堆疊學習：逐層貪婪非監督式預訓練與 DBN

2006 年，辛頓與多倫多大學的賽門·奧辛德羅（Simon Osindero）、新加坡國立大學的鄭義偉（Yee-Whye Teh）在《Neural Computation》第 18 卷第 7 期上發表了論文「A Fast Learning Algorithm for Deep Belief Nets（深層信賴網路的快速學習演算法）」。論文刊載於第 1527 頁至第 1554 頁。

該論文探討的問題是「解釋消除（explaining away）」現象。將上層的原因視為產生下層結果的神經網路稱為有向信賴網路。在這種神經網路中，很難藉由觀察結果來反向推論原因。試設想草坪被淋濕的原因有「下雨」和「灑水器」兩種可能。在看到草坪已濕的情況下，一旦得知下過雨，灑水器開啟過的可能性就會降低。原本毫無關聯的兩個原因，在得知結果的瞬間便相互糾纏交織。在隱藏層有多層、單元數達數千個的神經網路中，這種糾纏變得極其複雜，以致於精確推論幾乎變得不可能。

辛頓研究團隊用來解決此問題的工具是受限波茲曼機（RBM, Restricted Boltzmann Machine）。波茲曼機是辛頓在 1980 年代與泰瑞·塞諾夫斯基（Terry Sejnowski）共同構建的機率神經網路

，其名稱取自 19 世紀物理學家路德維希·波茲曼（Ludwig Boltzmann）的統計物理公式。機率神經網路是指每個單元並非輸出固定數值，而是依機率開啟或關閉的神經網路。RBM 則是減少該模型的連接、使其簡化的形式。它由接收數據的可見層與包含數據特徵的隱藏層兩層構成，同層內的單元互不相連，僅在兩層之間建立連接。得益於這項限制，只要確定了一層的數值，就可以彼此獨立且同時計算另一層的所有單元，使計算變得容易。辛頓在 2002 年發表了稱為對比發散（CD, contrastive divergence）的方法，使得 RBM 能夠被快速訓練。

研究團隊在數學上繞過了解釋消除問題。在層數無限延伸的有向信賴網路中，若將各層的權重綁定為其上一層權重的轉置（transpose），則上層產生的先驗機率恰好可以抵消下層所產生的糾纏。研究團隊將此稱為「互補先驗（complementary priors）」，並且證明了這樣綁定權重的無限深有向模型，在數學上等價於單個兩層的 RBM。如此一來，難以推論的深層有向模型便可以轉換為已有成熟訓練方法的 RBM 來處理。逐層貪婪非監督式預訓練（greedy layer-wise unsupervised pre-training）正是源於這種等價性。

學習步驟如下：第一步，將沒有標籤的輸入數據置入可見層，並利用對比發散演算法訓練最底層 RBM 的權重。第二步，固定該權重後

，讓輸入數據通過網路，計算各隱藏單元被開啟的機率。這一隱藏層的活動便成為下一層的新輸入數據。第三步，用這筆新數據單獨訓練第二個 RBM。如此逐層重複相同的過程，直到達到所需的深度為止。因其固定前一步驟的結果、僅將當前層訓練至最優，故被稱為「貪婪」；因其不依賴標準答案、僅靠數據本身進行學習，故為「非監督式」；而在展開正式訓練之前預先準備好權重，故稱為「預訓練」。

第四步，將如此堆疊起來的各個 RBM 合併為一個生成模型。這就是深層信賴網路（DBN, Deep Belief Network）。DBN 最頂端的兩層保持無向的對稱連接，發揮著類似聯想記憶裝置的作用；其下方的各層則轉變為由上至下的有向連接，承擔起生成數據的任務。

研究團隊還為此流程提供了數學依據。他們證明了：在固定下層權重的同時，每向上多堆疊並訓練一個 RBM，模型解釋數據程度的下界就不會變差，反而能得到提升。這是首次提出「層數越疊加，模型就越優秀」之數學依據的成果。

完成預訓練的神經網路權重不再是隨機值，而是置於一個已經在某種程度上蘊含了數據結構的良好起點上。考慮到 6.1 節中提到的最佳化失敗源自於從隨機權重出發，改變起點正是此方法的核心所在。

在此基礎之上，再利用少量的標籤數據對整體權重進行微調（fine-tuning）。在該論文中，研究團隊使用了改良自清醒-睡眠（wake-sleep）演算法的「向上-向下（up-down）」演算法進行微調。

研究團隊使用手寫數字資料集 MNIST 測試了效能。MNIST 由長寬各 28×28 像素的數字圖像組成，分為 6 萬張訓練用與 1 萬張測試用圖像。隱藏層分別具備 500、500、2000 個單元的 DBN，在未採用圖像形狀資訊或數據擴增等技巧的情況下，僅憑像素值就創下了 1.25% 的錯誤率。這一數值低於同等條件下支援向量機取得的 1.4%，以及藉由反向傳播訓練的神經網路取得的約 1.5%。該 DBN 不僅能對數字進行分類，還能利用訓練好的模型繪製出全新的數字圖像。在未額外利用圖像空間資訊的比較條件下，它展現了深度神經網路能夠發揮出超越主流既有方法的優異效能。

得出這一結論的並非只有辛頓研究團隊。同年，班吉歐研究團隊在 NIPS 會議上發表了逐層預訓練不僅在 RBM 中有效、在其他架構中同樣有助於深度神經網路學習的研究成果。楊立昆研究團隊也提出了類似方向的研究。CIFAR 社群的三個研究小組幾乎在同一時期得出了相同的結論。

6.3. 2006 年《Science》論文：深層自動編碼器與非線性降維

2006 年 7 月 28 日，辛頓與博士生魯斯蘭·薩拉庫特迪諾夫（Ruslan Salakhutdinov）在《Science》第 313 卷第 5786 期發表了論文「Reducing the Dimensionality of Data with Neural Networks（利用神經網路降低數據維度）」。論文自第 504 頁至第 507 頁，共四頁篇幅。如果說同月發表的 DBN 論文是透過專業學術期刊讓研究人員知曉，那麼這篇論文則是藉由整個科學界廣泛閱讀的期刊，宣告了深度神經網路研究的復甦。

本篇論文探討的是降維問題。降維是將包含大量數值的數據濃縮為較少數值的工作。若將由 784 個像素組成的數字圖像縮減至 30 個數字仍能還原其原始樣貌，便可認為這 30 個數字掌握了該圖像的核心特徵。長期以來，這項工作廣泛採用了哈羅德·霍特林（Harold Hotelling）於 1933 年整理的主成分分析（PCA）。PCA 尋找數據分布最廣的直線方向，並將數據投射在該方向上。

問題在於，實際數據並非分布在平坦的平面上。圖像或文檔等數據如同捲曲的紙張一般，分布於彎曲的曲面——即非線性流形（nonlinear manifold）之上。若從上方將捲曲的紙張投影下來，原

本相隔遙遠的部分便會重疊在一起。同樣地，將數據投影至直線軸上的 PCA 會將彎曲的結構徹底壓碎破壞。

自動編碼器（autoencoder）是能夠突破此一限制的神經網路。它由將輸入逐層壓縮至更小層的編碼器，以及從中間狹窄的代碼層重新還原至原始大小的解碼器所構成。若訓練網路使輸出與輸入趨於一致，中間的代碼層便會保留數據的核心精髓。這種構想早在 1980 年代就已出現。然而，實際實現時卻完全暴露出 6.1 節中所述的問題。若從隨機權重出發利用反向傳播訓練深層自動編碼器，學習訊號就無法有效傳遞至中間的代碼層，結果往往僅與 PCA 相當甚至更差。

辛頓與薩拉庫特迪諾夫藉由逐層 RBM 預訓練解決了這個問題。方法分為三個步驟：第一，自底向上依序訓練多個 RBM。在 MNIST 實驗中，採用了將 784 個輸入經過 1000、500、250 個單元逐步縮減至 30 個的架構。第二，將訓練好的 RBM 權重堆疊以構建編碼器，並將相同權重轉置後反向堆疊以構建解碼器。藉此將自動編碼器「展開（unrolling）」。第三，將原先依機率輸出 0 或 1 的單元替換為輸出連續數值的 Sigmoid 單元，並藉由反向傳播對整個神經網路進行微調，以縮小原始輸入與重建輸出之間的差距。在透過預訓練

奠定了良好的起點之後，反向傳播在深度神經網路中終於發揮了應有的作用。

即便從圖像對比來看，結果也是顯而易見的。當將數字圖像壓縮為 30 個數值後再予以還原時，深層自動編碼器所產生的圖像與原始圖像幾乎完全一致。以相同大小壓縮的 PCA 重建圖像則模糊斑駁。將數字圖像展開在二維平面上時，兩者的差異同樣巨大。在 PCA 圖中，0 到 9 的數字相互混雜重疊；但在自動編碼器圖中，每個數字都清晰地聚集成獨立的簇塊。在人臉照片的實驗中，自動編碼器也展現出比 PCA 更小的重建誤差。

研究團隊也將此方法應用於文檔。將 80 餘萬篇新聞報導以 2000 個常用詞的詞頻表示並進行壓縮後，相同主題的報導便聚集在了一起。其結果優於既有的線性方法——潛在語意分析（LSA）。該項研究進一步催生了薩拉庫特迪諾夫與辛頓於 2009 年發表的「語意雜湊（semantic hashing）」論文。若將文檔轉換為簡短的二進位代碼，便能透過尋找代碼相似文檔的方式，極其快速地檢索出語意相近的文檔。

在論文結尾處，兩位研究者總結了此項成果是在何種條件下才得以實現的。雖然利用深層自動編碼器進行非線性降維的構想早在 1980

年代就已出現，但若要將其實現，必須具備高速電腦、大數據以及良好的初始權重。到了 2006 年，這三項條件皆已齊備，而獲取良好初始權重的方法正是逐層預訓練。

6.4. 深層波茲曼機（DBM）與深度學習的擴散

2009 年，薩拉庫特迪諾夫與辛頓在美國佛羅里達州舉行的第 12 屆人工智慧與統計國際會議（AISTATS）上發表了深層波茲曼機（DBM, Deep Boltzmann Machine）。DBM 是僅有頂端兩層為無向連接、下層均為自上而下連接的混合型模型。與此不同，DBM 在可見層與所有隱藏層之間均採用了無向的對稱連接。

這種差異體現在推論方式上。DBM 中間的隱藏層同時接收自下層傳來的感官資訊與自上層傳來的上下文資訊，以此決定自身的狀態。這與人類在辨認模糊字跡時，藉由前後文脈來推測文字十分相似。因此，DBM 對於部分被遮擋或混入噪聲的輸入也能處理得相對出色。

然而，DBM 的訓練難度較高。研究團隊雖然實施了逐層 RBM 預訓練，但配合從上下雙向接收訊號的架構，將權重修改為加倍或減半的方式。隨後利用平均場（mean-field）近似進行推論，並結合稱

為持續對比發散（persistent CD）的取樣方法調整整體權重。如此訓練而成的 DBM 在 MNIST 上取得了 0.95% 的錯誤率；在包含從多個角度拍攝的三維玩具物體的 NORB 數據集上，也創下了 10.8% 的錯誤率，低於作為對照組的支援向量機的 11.6%。

此時，預訓練的成效開始向實驗室之外擴散。首個領域是語音辨識。多倫多大學的阿卜杜勒-拉赫曼·穆罕默德（Abdel-rahman Mohamed）與喬治·達爾（George Dahl）同辛頓一起，將藉由 DBN 預訓練的深度神經網路應用於判別語音聲音單位的聲學模型中。數十年來，語音辨識領域始終將結合高斯混合模型（GMM）與隱藏式馬可夫模型（HMM）的方法奉為標準。深度神經網路在 TIMIT 音素辨識任務中超越了該標準方法，緊接著在大規模詞彙語音辨識中也取得了優異成績。2012 年，多倫多大學、微軟、Google、IBM 的研究人員在《IEEE Signal Processing Magazine》上發表了聯合論文，宣佈四個研究小組均得出了相同的結論。此後，各大主要企業紛紛將語音辨識引擎的核心更替為深度神經網路。

同在 2012 年，辛頓的學生亞歷克斯·克里哲夫斯基與伊利亞·蘇茨克維，憑藉與辛頓共同構建的神經網路 AlexNet，在 ImageNet 競賽中創下了前 5 名錯誤率 15.3% 的紀錄，以遙遙領先第二名的巨大差距奪冠。電腦視覺領域亦隨之迅速轉向深度神經網路。

然而，正是在深度神經網路開始被廣泛使用的這個時期，逐層預訓練開始退出了主流地位。這是因為出現了以其他方式解決 6.1 節中那三個問題的手段。藉由圖形處理器（GPU），計算速度大幅提升，並湧現出數百萬張標籤數據。緩解梯度消失的 ReLU 激勵函數、在訓練過程中隨機關閉部分單元的 Dropout、以及設定初始權重的改良方法相繼問世。於是，無需預訓練便能從一開始就透過反向傳播訓練深度神經網路成為可能。AlexNet 也是在沒有預訓練的情況下完成訓練的神經網路。

儘管如此，2006 年的成果並未因此失去意義。2015 年，楊立昆、班吉歐與辛頓在《自然》第 521 卷發表的綜述論文「Deep Learning」中回顧總結了這一時期。三人寫道，2006 年前後由 CIFAR 凝聚在一起的研究人員重新喚醒了人們對深度神經網路的關注。他們評述道，利用無標籤數據逐層學習特徵以建立良好的初始權重、再透過反向傳播加以微調的方法，證明了即使在標籤數據匱乏的情況下，深度神經網路依然能出色運作。打破「深度神經網路無法被訓練」這一長期以來的陳舊成見，正是這項研究的卓越貢獻。

先利用大量無標籤數據進行預訓練、再利用少量標籤數據進行微調的構想，在當今的大型語言模型中依然得以延續。雖然現在的模型

不再堆疊 RBM，但先以海量文本進行預訓練、再針對特定任務進行微調的整體架構，與 2006 年的思維模式如出一轍。

辛頓、班吉歐與楊立昆於 2018 年共同榮獲圖靈獎。圖靈獎被譽為電腦科學界的諾貝爾獎。隨後在 2024 年 10 月 8 日，瑞典皇家科學院宣佈將諾貝爾物理學獎授予辛頓與約翰·霍普菲爾德（John Hopfield），以表彰他們在實現人工神經網路機器學習方面的基礎性發現與發明。委員會指出，辛頓以霍普菲爾德網路為基礎構建了波茲曼機，並藉由深化這一研究，為開啟當今機器學習的迅猛發展做出了貢獻。2006 年的 DBN 正是自該波茲曼機發展而來的模型。在神經網路研究陷入低谷的歲月裡依然堅持不懈的波茲曼機研究，最終孕育出了訓練深度神經網路的有效方法。

第三部. 深度學習走向世界

第7章. 辨識語音與照片的電腦

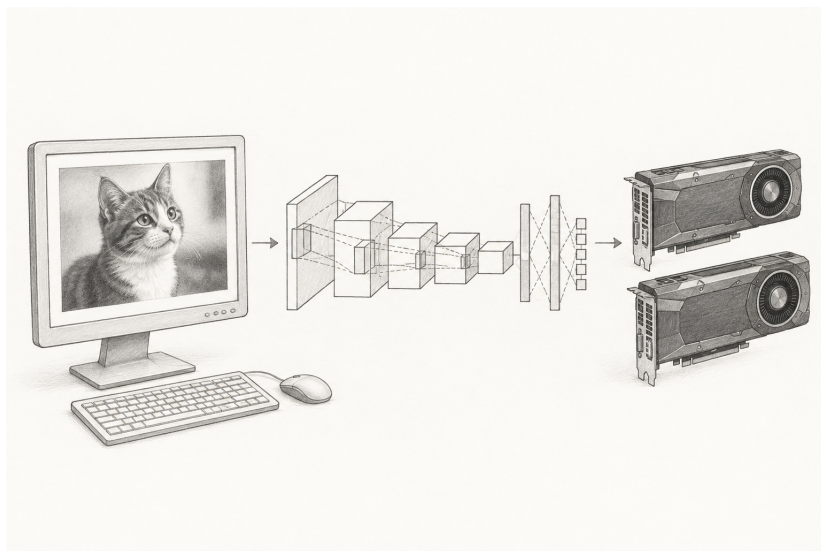


圖 7. 呈現了圖像辨識與GPU的作用。

2006年，傑佛瑞·辛頓以深度信念網路（DBN）開闢了訓練深層神經網路的道路。然而，人工智慧學界的大多數人依然不信任神經網路。僅憑幾篇論文中刊登的優異成果，難以扭轉長期根深蒂固的既定觀念。改變學界看法的，是來自實際問題的具體數據。從2009年到2012年，辛頓的實驗室兩度提出了這樣的數據。第一次出現在理解人類話語的語音辨識領域，第二次則出現在辨認照片中物體的電腦視覺領域。

7.1. 第一個突破：語音辨識的變革 (2009~2012)

電腦將人類話語轉為文字的技術稱為自動語音辨識。這項技術大致分為兩個部分。其一是判斷極短聲音片段屬於何種發音的聲學模型，另一部分則是將辨識出的發音串接成詞彙與句子的部分。其中，聲學模型直到2000年代後期，近30年間都停留在相同的方式上。那就是以高斯混合模型（GMM）將聲音特徵計算為機率，並以隱藏式馬可夫模型（HMM）追蹤隨時間變化的方式。

這種方式長久以來表現良好，但缺點也很明顯。GMM的結構將一個聲音片段視為由眾多「專家」中的一位負責解釋。然而，人類的說話聲音並非如此涇渭分明。前後的發音會相互混合，在同一個瞬間

也會有多種聲音特性重疊。為了表達這種重疊，GMM必須不斷增加組成元件，因而需要大量的計算量與數據。它也很難同時綜觀前後相連的多個瞬間的聲音。

最先突破這項限制的人，是多倫多大學辛頓實驗室的博士生喬治·達爾與阿卜杜勒-拉赫曼·穆罕默德。2009年，兩人將辛頓於2006年提出的DBN方法應用於聲學模型。他們首先將受限玻爾茲曼機（RBM）逐層堆疊，在沒有標準答案的情況下學習聲音的結構，接著透過誤差反向傳播微調整個神經網路。在輸入端，他們不再只輸入單一聲音片段，而是將前後相連的11~15個片段打包成一組輸入。神經網路得以直接接納GMM難以處理的前後上下文脈絡。

讓這項實驗成為可能的，是顯示卡，亦即GPU。GPU原本是為了繪製遊戲畫面而製造的裝置，擅長同時大量處理同類型的計算。兩人使用輝達（NVIDIA）的Tesla GPU，以及實驗室同僚弗拉基米爾·米尼開發的計算工具CUDAMat，將訓練速度提升至單顆CPU的20倍以上。原本需要耗費數週的實驗在數小時內就能完成，使得多次調整設定進行反覆測試成為可能。

結果受到了語音辨識學界的矚目。在學者長期使用的標準測試資料TIMIT上，他們的神經網路創下了20.7%的音素錯誤率。音素是構成

說話聲音的最小聲音單位，因此這代表在100個音素中大約錯了21個。這項成績超越了截至當時所報告的傳統方法最高紀錄。最初的成果於2009年NIPS工作坊上發表，經過完善後的成果則於2012年刊登在IEEE學術期刊上。

然而，TIMIT是一份規模較小的測試資料。要確認這種方法在實際產品中是否同樣行得通，需要大企業擁有的數據與系統。因此，辛頓的學生們親自前往企業。喬治·達爾前往微軟研究院（MSR），阿卜杜勒-拉赫曼·穆罕默德則前往IBM研究院實習。另一位弟子納夫迪普·賈伊特利則去了Google。

在微軟研究院，鄧力（Li Deng）與俞棟（Dong Yu）與辛頓研究團隊展開合作。他們將深度神經網路應用於手機錄音、雜訊較多的Bing行動語音搜尋數據，句子層級的錯誤率從36.2%降至30.4%。隨後，在長達300小時的電話對話資料Switchboard上，詞錯誤率從27.4%降至18.5%。相當於減少了約三分之一的錯誤。這項成績媲美動用了超過2,000小時數據及各種複雜校正技術的傳統系統所取得的18.6%。用少得多的數據、僅憑單一神經網路，就達到了相同的水準。

在Google，賈伊特利與文森特·范霍克、派屈克·阮等人合作，使用長達5,870小時的語音輸入數據進行實驗。詞錯誤率從16.0%降至12.3%。Google完善了這項技術，並於2012年應用於Android智慧型手機的語音搜尋中。從大學實驗室起步的神經網路，就此走入了眾人日常使用的產品之中。

這些在不同公司分別獲得驗證的成果，被整理成了一篇論文。2012年11月，多倫多大學、微軟、Google、IBM四個研究團隊在《IEEE信號處理雜誌》（IEEE Signal Processing Magazine）上發表了題為〈用於語音辨識聲學建模的深度神經網路〉的聯合論文。相互競爭的企業研究人員在同一篇論文上共同署名，是十分罕見的事情。這篇論文向學界宣告了語音辨識標準的轉變，也是深度學習在產業現場取得的重大成果之一。然而，這項變革主要發生在語音辨識專家之間。要讓整個擴大的人工智慧學界切身體會到這一點，還需要一點時間。

7.2. 電腦視覺的轉折點：李飛飛的 ImageNet與AlexNet (2012)

2000年代後期，電腦視覺學界依循由人工親自設計照片中可用特徵的方式。SIFT與HOG便是代表性的例子。其做法是計算照片中邊緣的方向或亮度變化以提取特徵，並透過支援向量機（SVM）等分類器來判斷物體的種類。這種方式一旦遇到光線改變、物體以不同角度拍攝，或是部分被遮擋，效能便很容易受到影響。儘管效能年年都有提升，但提升的幅度相當有限。

就在此時，為扭轉局勢備妥關鍵素材的人是李飛飛。李飛飛於2007年在普林斯頓大學啟動了ImageNet計畫，並於2009年轉至史丹佛大學繼續進行研究。李飛飛的研究團隊讓人力親自為從網路上收集來的照片貼上標籤。所需的人力是透過名為亞馬遜土耳其機器人（Amazon Mechanical Turk）的線上群眾外包服務取得。就這樣，一個涵蓋超過22,000個類別、收錄超過1,500萬張照片的資料庫誕生了。這與研究人員截至當時所使用的照片資料相比，光是規模就截然不同。

自2010年起，主辦方從中挑選出1,000個類別、約120萬張訓練用照片，每年舉辦名為ILSVRC的競賽。成績主要依「前5項錯誤率」進

行評估。這種方式是由電腦針對每張照片提出5個候選答案，若標準答案不在其中，則計為錯誤。2010年冠軍隊伍的錯誤率為28.2%，2011年冠軍隊伍則為25.8%。使用人工手工設計特徵的方法，在此停滯不前，無法取得更大突破。

2012年，辛頓與兩位弟子亞歷克斯·克里哲夫斯基、伊利亞·蘇茨克維一同參加了這項競賽。團隊名稱為SuperVision。三人的分工各不相同。撰寫程式碼並訓練神經網路的是克里哲夫斯基，而強烈建議將神經網路應用於大型照片資料的則是蘇茨克維。辛頓則以指導教授的身分指引方向。克里哲夫斯基使用插了兩張輝達顯示卡的電腦完成了這項工作，而該神經網路後來便以他的名字命名為AlexNet。

AlexNet的淵源可以追溯至楊立昆於1980年代末為了讀取手寫數字而建構的卷積神經網路（CNN）。卷積神經網路依序掃描照片的微小區域以尋找紋理，且隨著層數的推進，能辨認出更複雜的形狀。AlexNet將此結構放大以適配龐大且複雜的彩色照片。它由5個卷積層與3個所有神經元完全相連的層組成，共計8個層級。透過訓練進行調節的數值——亦即參數——約有6,000萬個，神經元約有65萬個。這在當時是罕見的龐大規模。

然而，他們並非只是單純擴大規模。為了妥善訓練如此規模的神經網路，研究團隊同時採取了四種方法。第一是使用ReLU作為激活函數。激活函數是決定神經元輸出多少訊號的規則。在此之前普遍使用的函數，在輸入變大時斜率會趨近於0，導致訓練速度變慢。ReLU則是一個簡單的函數，它將小於0的值轉為0，大於0的值則原樣輸出。在實驗中，使用ReLU的神經網路訓練速度比使用傳統函數的神經網路快了6倍。

第二是將神經網路分散載入兩台GPU中。克里哲夫斯基所使用的輝達GTX 580記憶體僅有3GB，無法將整個神經網路容納於一台設備中。他親自撰寫了GPU專用的計算程式碼cuda-convnet，將神經網路一分為二分別置於兩張GPU上，並僅在部分層級讓兩張GPU相互交換資訊。將120萬張照片反覆訓練近90次，耗費了5至6天的時間。

第三是Dropout（隨機丟棄）。當參數多達6,000萬個時，神經網路容易將訓練照片全盤死記硬背，導致在面對初次見到的照片時表現下滑。這種現象稱為過擬合。Dropout是在每次訓練時隨機關閉一半神經元的方法。如此一來便能產生每次訓練略微不同的神經網路之效果，避免神經網路過度依賴特定神經元。AlexNet在後方的兩個層級中採用了這種方法。

第四是擴充訓練數據。他們從 256×256 尺寸的照片中隨機裁切出 224×224 尺寸的區域，進行水平翻轉，並微調色調。這相當於從單張照片中生成了多張全新的訓練資料。

如此訓練出來的神經網路從照片中學到了什麼，可以從第一層中得到確認。將AlexNet第一層自主學得的96個濾波器繪製成圖時，會呈現出不同方向的線條、邊緣以及色彩區塊。沒有人教導它去尋找這些紋理，而是神經網路由觀察照片自主產生的。這些紋理與人腦初級視覺皮質細胞所反應的圖案十分相似。

競賽結果於2012年10月在義大利佛羅倫斯舉行的歐洲電腦視覺會議（ECCV）工作坊上公佈。AlexNet的前5名錯誤率為15.3%。榮獲第二名的東京大學團隊使用SIFT與費雪向量等傳統方法，成績為26.2%。兩隊差距達10.9個百分點。在過去往往以1~2個百分點的差距決定勝負的競賽中，AlexNet比第二名隊伍將錯誤率大幅降低了40%以上。

7.3. 轉向深度學習，創立DNNresearch與Google收購

佛羅倫斯發表會現場的反應褒貶不一。長期研究電腦視覺的學者們，無法確信這種方法是否真正實用。相反地，在場的楊立昆則評價這是人工智慧的轉折點。楊立昆曾於1980年代末在多倫多擔任辛頓的博士後研究員。

後續的發展正如楊立昆所料。在AlexNet之前，主要的電腦視覺論文中幾乎沒有使用神經網路的，提交神經網路論文遭到審核淘汰更是家常便飯。然而在AlexNet問世後的1至2年內，研究人員紛紛放下了人工設計特徵的方式，轉向深度學習。自隔年的競賽起，名列前茅的團隊絕大多數都採用了卷積神經網路。在邊緣徘徊超過30年的聯結主義所積蓄的力量瞬間爆發，這一幕常被比喻為潰堤決口的瞬間。李飛飛後來表示，許多人將此時視為現代人工智慧誕生的時刻。

各大企業也開始行動。已在語音辨識中見證神經網路優異效能的科技巨頭們，在看到AlexNet之後紛紛爭相延攬辛頓與其兩位弟子。辛頓決定創立公司並予以出售。他日後在多次訪談中透露，為了患有學習障礙的兒子的未來，他需要一大筆資金。2012年底，辛頓與克里哲夫斯基、蘇茨克維共同創立了名為DNNresearch的公司。公司

名稱取自深度神經網路（Deep Neural Network）。公司既無產品也無營收，員工僅有他們三人。公司所擁有的全部資產，就只有打造AlexNet的技術以及三位研究者的才華。

2012年12月初，在美國太浩湖舉行的NIPS大會期間，針對這家公司展開了一場競拍。參與競標的有百度、Google、微軟，以及成立僅兩年的新創公司DeepMind。競拍透過電子郵件進行，辛頓在飯店房間內主持競價，並向各競標者隱瞞彼此的參與身分。起標價為1,200萬美元，並規定每次加價至少為100萬美元。

DeepMind因以自家公司股票代替現金競標而最早退出，微軟則在2,200萬美元左右收手。堅持到最後的是Google與百度。當價格達到4,400萬美元時，辛頓叫停了競標，決定將公司交給Google。對他而言，比起獲得最高金額，更重要的是為繼續推進研究尋找合適的歸宿。收購案於2013年3月正式宣佈。

收購完成後，三人各自走向了不同的道路。辛頓在保留多倫多大學教職的同時加入Google，其後以副總裁兼工程研究員的身分在Google參與神經網路研究。蘇茨克維在Google Brain團隊於2014年提出了將一個句子轉換為另一個句子的Sequence-to-Sequence（Seq2Seq）模型，為機器翻譯的發展做出了貢獻。他於2015年底離

開Google，與山姆·奧特曼等人共同創立了OpenAI，並以首席科學家的身分引領了日後促成ChatGPT誕生的研究。克里哲夫斯基在Google又工作了數年，隨後於2017年離開了公司。

這場競拍改變了整個業界的盤算。當一家僅有三人的公司被估值為4,400萬美元時，深度學習研究人員的身價也被重新定義。2014年1月，Google以約6億美元收購了曾在競拍中同台競爭的DeepMind。在競拍中落敗的百度與微軟，也正式全面投入深度學習的研究。

AlexNet的原始程式碼隨著DNNresearch的收購歸Google所有。這份程式碼向大眾公開，已是10餘年後的事了。電腦歷史博物館（CHM）的策展人韓森·許於2020年向克里哲夫斯基提議公開程式碼，在與Google協商5年後終於取得了原始程式碼。博物館於2025年3月20日與Google共同宣佈，將AlexNet的原始程式碼向大眾公開並永久保存。公開的程式碼包含輝達CUDA程式碼、Python與C++，合計約200KB大小。與今日的人工智慧模型相比，這雖然是極小的規模，但這段小小的程式碼卻改變了一個時代的方向。

2009年的語音辨識與2012年的AlexNet皆出自同一個實驗室。這兩起事件都以實際的具體數據，證明了辛頓數十年來所堅持的信念——即使人類不預先設定規則，神經網路也能自行從數據中學習

特徵。當大規模數據與GPU這兩項條件齊備之時，長期積累的研究終於全面結出碩果。曾為非主流學術流派的聯結主義，在這4年間躍升成為人工智慧的主流。

第8章. 減少過擬合並傳遞知識

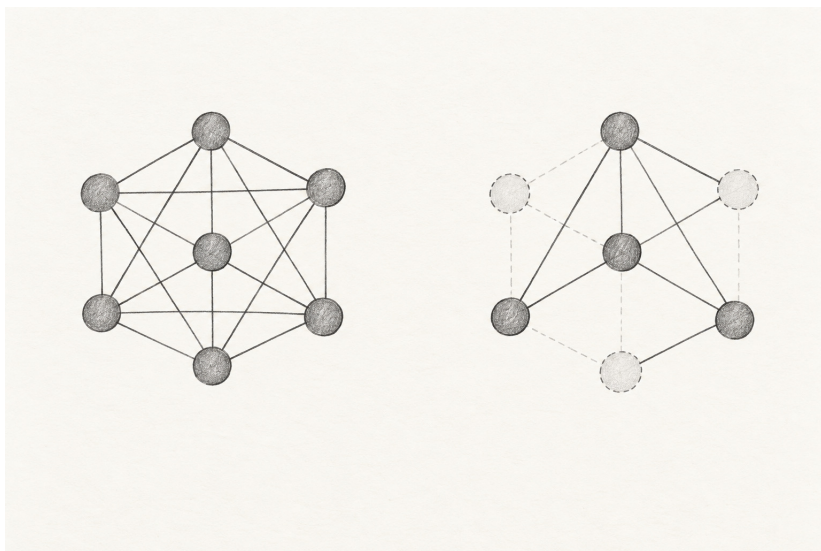


圖8. 呈現了在訓練時暫時關閉部分神經元的Dropout。

8.1. 參數增加與過擬合問題：Dropout的誕生

2010年代初，深度信念網路（DBN）、深度自編碼器以及AlexNet接連取得成果，人工神經網路的層數不斷加深，參數數量也大幅增加。參數是神經網路在學習過程中自行修正的數值。當這些數值增加到數千萬、數億個時，研究人員再次面臨了一個古老的問題，那就是過擬合（Overfitting）。

過擬合就像把考試範圍內的題目和答案死記硬背下來的學生。該學生在練習題中能拿滿分，但一遇到沒見過的新題目就無從下手。神經網路在訓練資料有限的情況下，如果反覆訓練相同的資料數千次，就會把每個樣本中夾雜的偶然雜訊也一併死記下來，而不是學習資料中的規律。結果，針對訓練資料的誤差雖然接近於0，但在新的評估資料上效能卻會下降。這被稱為泛化（Generalization）失敗。

統計學與機器學習領域為防止這個問題，一直使用權重衰減（ $L1 + L2$ Weight decay）、提早結束（Early stopping）、權重共享等正規化（Regularization）技術。然而，堆疊了多層非線性隱藏層的深度神經網路，其參數空間極其龐大，僅憑這些方法很難抑制過擬合。

其中一種解決方案是集成（Ensemble）。集成是指將多個結構或初始值各不相同的模型分別進行訓練，然後對它們的預測值取平均的方法。這就像向多個人詢問並匯集答案時，單個人的失誤會被掩蓋一樣。問題在於成本。單是訓練一個龐大的神經網路就需要數天到數週的時間。分別訓練數十個這樣大的神經網路，並在實際服務中同時運行這些模型，在計算量和回應延遲方面是不切實際的。

傑佛瑞·辛頓提出了Dropout作為突破這項極限的方法。這是與多倫多大學的尼蒂許·斯里瓦斯塔瓦（Nitish Srivastava）、亞歷克斯·克里哲夫斯基（Alex Krizhevsky）、伊爾亞·蘇茨克維（Ilya Sutskever）、魯斯蘭·薩拉赫丁諾夫（Ruslan Salakhutdinov）共同進行的研究。研究團隊於2012年7月先在arXiv上公開了《透過防止特徵檢測器的共適應改善神經網路》，隨後於2014年在《機器學習研究期刊（JMLR）》第15卷上發表了正式論文《Dropout：防止神經網路過擬合的簡單方法》。核心思想是在訓練單一神經網路的過程中，同時訓練其內部無數個規模縮小的不同神經網路，並讓它們共享權重。這是試圖在獲得集成效果的同時，避免付出集成的成本。

第一篇論文題目中提到的共適應（Co-adaptation）正是研究團隊針對的問題。在一般的誤差反向傳播訓練中，隱藏神經元之間很容

易形成互相彌補彼此錯誤的複雜關係。這就像在分組作業中，如果一人漏掉的部分總是由另一人補上，那麼一旦組員更換，整個作業就會瞬間崩潰。共適應嚴重的神經網路雖然在訓練資料上表現良好，但只要輸入稍微不同的資料，預測就會大幅動搖。

辛頓從日常經驗和生物學中找到了思考解決此問題的出發點。他在一次線上問答中表示，每次去銀行時看到櫃檯行員都會更換，便推測這是一種防止行員互相串通舞弊的機制。他的構想是，如果神經網路的神經元每次也隨機去除一部分，就能防止神經元之間串通勾結。

演化生物學中的有性生殖理論為這一構想提供了理論依據。無性生殖幾乎將親代的基因組合原封不動地傳給後代。由於不會失去已經良好適應環境的基因組合，表面上看來是有利的。相反地，有性生殖從雙親那裡各自隨機抽取一半基因進行混合，因此代代都會打散原本互相配合且運作良好的基因組合。然而，絕大多數高等生物卻選擇了有性生殖。生物學家阿迪·利夫納特（Adi Livnat）等人在2010年的論文中解釋其原因：自然選擇挑選的是基因的混合能力（Mix-ability），而非單一個體的適應度。這意味著，不依賴總是共同出現的特定基因群，而是與任何基因混合都能發揮自身作用的基因才能生存下來。

辛頓將這個原理直接移植到隱藏層的學習中。Dropout在每次訓練時隨機關閉部分神經元，讓任何神經元都不能認定相鄰的神經元必定存在。這樣一來，每個神經元無論與哪些同伴共事，都必須憑一己之力找出有用的特徵，例如線條或斑點等細微紋理。正如與任何基因混合都能各司其職的基因會被選擇一樣，在任何神經元組合中都能發揮作用的神經元也隨之被培養出來。

8.2. Dropout的計算方式與實驗結果

Dropout的計算方式非常簡單。可以想像成每輸入一筆訓練資料時，就為各層的每個神經元拋一次硬幣。以預先設定的機率，如果出現正面就保留該神經元，如果出現反面則在那一瞬間將其輸出置為0。保留神經元的這一機率稱為保留機率（Retention probability）。論文建議保留機率的預設值為隱藏層0.5，輸入層約0.8左右。

計算步驟依序如下。某一層的輸出是由每個神經元各有一個數值所組成的向量。為該層的每個神經元各抽取一個取值為0或1的白努利隨機變數，出現1的機率即為保留機率。將這樣抽出的0與1向量與輸出向量中對應位置的元素相乘，再將這樣被部分抹除的輸出作為下一層的輸入傳遞下去。下一層的計算則與平時相同。

在神經網路中，開啟和關閉神經元的可能性總數是 2^n 的「神經元數量」次方。Dropout相當於在每個訓練步驟中隨機挑選其中一種進行訓練。這數量龐大的子神經網路並非各自獨立存在，而是共享同一組權重。這是一種以單一神經網路的成本來訓練龐大集成的結構。

訓練結束並實際使用神經網路時，不會關閉任何一個神經元。取而代之的是，將每個神經元輸出的權重乘以 p 。如果在訓練時某個神經元只有一半的機率處於開啟狀態，那麼下一層接收到的訊號平均值便只有原本的一半。在評估時若將權重乘以0.5，下一層就能接收到與訓練時規模相同的訊號。僅憑這個簡單的調整，就能透過單次計算獲得對 2^n 個子神經網路預測取平均的值，更精確地說，是接近幾何平均數的值。

研究團隊還證實，將Dropout與其他技術結合使用時效果最為顯著。第一是最大範數（Max-norm）約束。這是防止輸入單個神經元的權重向量長度超過常數 c （通常在3到4之間）的方法。第二是高學習率與大動量（Momentum）。因為Dropout在訓練中混入了雜訊，所以使用比一般神經網路高10倍至100倍的學習率以及約0.95至0.99的動量係數時，能獲得良好的結果。學習率過大可能導致權重過度增長的風險，而最大範數約束則防止了這一風險。兩種技術就這樣相輔相成。

實驗在多個領域均一致確認了其效果。在手寫數字資料集MNIST中，一般神經網路的錯誤率為1.60%。將Dropout與ReLU及最大範數結合使用、包含2個隱藏層且每層有8,192個神經元的神經網路，儘管參數多達約6,500萬個，但在沒有提早結束的情況下仍創下了0.95%的紀錄。在以深度波茲曼機（DBM）進行預訓練後再用Dropout進行微調，錯誤率更進一步降至0.79%。

在其他資料集上的結果也類似。在Google街景門牌號碼照片資料集SVHN中，卷積神經網路的錯誤率原本為3.95%，僅在全連接層加入Dropout時降至3.02%，在所有層都加入時則降至2.55%。在物體照片分類任務CIFAR-10中，錯誤率從15.60%降至12.61%，在CIFAR-100中從43.48%降至37.20%。在語音辨識任務TIMIT中，4層聲學模型的音素錯誤率從22.7%降至19.7%。在路透社新聞分類與基因替代剪接預測中也展現了成效。

在2012年影像辨識競賽ILSVRC中奪冠的AlexNet也使用了Dropout。克里哲夫斯基、蘇茨克維與辛頓在這個擁有約6,000萬個參數的神經網路前段兩個全連接層中，加入了保留機率為0.5的Dropout。論文指出，若沒有Dropout，過擬合將會非常嚴重，但作為代價，收斂所需的迭代次數約增加了一倍。Dropout伴隨著訓練時間延長的代價。

觀察透過Dropout訓練的神經網路內部，還會發現其他特徵。隱藏神經元的平均激活值會降低，大多數神經元保持接近0的數值，自然形成了稀疏表徵（Sparse representation）。論文還一併整理了以平均值為1的常態分佈雜訊取代0和1的高斯Dropout（Gaussian Dropout）等變形。

此後10餘年間，Dropout一直作為深度神經網路的基礎工具被廣泛使用。只不過其用途隨著時代而有所變化。在批次正規化等新技術出現後，它在卷積神經網路中的比重有所降低；而在僅對龐大資料進行一兩次訓練的當今超大型語言模型中，很多情況下甚至完全關閉Dropout進行訓練。這是因為相對於模型大小而言，當資料量充足時，模型死記硬背資料的空間很小。儘管如此，在資料不足的領域或微調階段，它至今仍被廣泛使用。

8.3. 「黑暗知識」與昆蟲變態：知識蒸餾的構想

雖然透過Dropout和集成大幅提升了神經網路的泛化效能，但產業現場卻面臨了另一個問題。參數達數千萬到數億個的龐大模型，或是由數十個模型組成的集成，很難直接部署到手機或語音搜尋等記憶體少且需要快速回應的環境中。如何將訓練良好的模型實際部署，成為留待解決的問題。

辛頓與當時在Google共事的歐利奧爾·維尼亞爾斯（Oriol Vinyals）、傑夫·迪恩（Jeff Dean）一起研究了這個問題。三人在2014年NIPS深度學習研討會上發表了《蒸餾神經網路中的知識》（Distilling the Knowledge in a Neural Network），該論文於2015年3月在arXiv（1503.02531）上公開。

論文以昆蟲的變態故事展開。蝴蝶或蛾等昆蟲以幼蟲和成蟲這兩種截然不同的形態生活。幼蟲的身體結構專注於從周圍盡可能多地吸收養分，而成蟲的身體則專注於移動和繁衍。辛頓認為，在機器學習中也應該這樣區分訓練與部署的形態。在此之前，用於訓練的模型幾乎原封不動地直接用於提供服務。然而，在從龐大資料中提取結構的訓練階段，大型且笨重的模型或Dropout集成（即幼蟲形態）

最為合適；一旦訓練完成，將該知識轉移到小巧而迅速的模型（即成蟲形態）中進行部署即可。這就是知識蒸餾（Knowledge Distillation）的基本構想。此時，大模型被稱為教師（Teacher），小模型被稱為學生（Student）。

先前也有過類似的嘗試。2006年，克里斯蒂安·布奇盧亞（Cristian Buciluă）、里奇·卡魯阿納（Rich Caruana）與亞歷山德魯·尼庫列斯庫-米齊爾（Alexandru Niculescu-Mizil）在《模型壓縮》（Model compression）論文中，展示了讓小模型學習模仿大型集成模型預測的方法。辛頓在此基礎上更進一步。他並未將模型的知識視為權重數值本身，因為在結構不同的神經網路之間複製權重毫無意義。真正需要轉移的是接收輸入並轉換為輸出的方式，即模型所學到的函數。

那麼，從何處能夠解讀出蘊含在該函數中的資訊便成為了關鍵問題。辛頓對此給出的答案是「黑暗知識（Dark Knowledge）」。假設將數字2的手寫影像輸入到訓練良好的教師模型中，模型輸出為「2」的機率約為0.999，為「3」的機率為100萬分之一，為「7」的機率為10億分之一。如果從僅標註正確答案的硬標籤（Hard target）觀點來看，「3」和「7」同樣都是錯誤答案。然而，這兩個錯誤答案的機率相差1,000倍的事實中，蘊含著教師模型認為這個「2」與「7」

相比遠更接近「3」的資訊。包含賦予錯誤答案的這些微小機率在內的分佈，被稱為軟標籤（Soft target）。辛頓在演講中，將這種表面上不易顯現的相似性資訊稱為「黑暗知識」。這是借用了佔宇宙絕大部分質量卻看不見的暗物質作比喻的說法。

兩種標籤的差異體現在學生學到了什麼。若只教給學生模型正確答案，學生學會的只是如何答對問題；若教給學生軟標籤，學生連同教師如何劃分看待世界、將何者與何者視為相似的方式都能一併學會。學生繼承的正是教師的泛化方式本身。

8.4. Softmax溫度調節與向小型學生模型的知識傳遞

將黑暗知識運用於訓練時存在著障礙。100萬分之一或10億分之一這樣的機率實在太小，在計算訓練所用的交叉熵損失時幾乎毫無影響而被掩蓋。研究團隊藉由從統計物理學中借鑑的「溫度（Temperature, T ）」機制解決了這個問題。

神經網路的最後一層會為每個分類項目（類別）輸出一個分數。這個分數被稱為對數機率（Logit）。Softmax函數將這些分數轉換為

總和為1的機率。在蒸餾中，這裡會加入一個稱為溫度的數值。引入溫度的Softmax計算方式如下。

首先將各類別的Logit除以溫度。將該數值代入指數函數轉換為正數，指數函數是指輸入越大結果便急遽增長的一種函數。某個類別的機率，即是用該類別所得的數值除以所有類別所得數值的總和。

平時溫度的設定值為1。如果將溫度大幅提高到如20，所有的Logit都會被除以一个大數，計算出來的分數差距隨之縮小。結果，原本像0.999與100萬分之一這樣極端的機率，會平緩地分散為0.45、0.30、0.15等數值。這類似於在混雜了大聲與耳語的錄音中縮小音量差距，使微弱的耳語也能聽得一清二楚。進行蒸餾時，教師模型與學生模型都以相同的高溫計算Softmax，並訓練學生去擬合教師那平滑化的機率分佈。訓練完成後，學生模型將溫度重新恢復為1來執行任務。

研究團隊確認，將學生的損失函數由兩部分組成時能獲得最佳結果。其一是高溫下比較教師軟標籤與學生預測的損失，另一則是溫度為1時比較真實標籤與學生預測的損失。對這兩種損失進行加權平均，但前一項損失需要乘以溫度的平方。這是因為當溫度提高時，來自軟標籤的梯度大小會與溫度的平方成反比而縮小。乘以溫度的平方

後，即使改變溫度進行實驗，兩種損失的相對比重也能保持不變。給予真實答案損失相對較小的權重時，效果最好。

該論文還包含闡明早期研究與知識蒸餾之間關係的數學成果。假設溫度遠高於Logit的大小，且教師與學生的Logit平均值各自調整為0。在此條件下，學生修正分數的方向會趨近於縮小教師與學生分數差距的方向。分類項目的數量與溫度會影響修正幅度。在高溫下的蒸餾，趨近於縮小兩個模型Logit差距平方的訓練。2006年卡魯阿納團隊的Logit匹配方法，可以看作是知識蒸餾的一個特例。然而，並非溫度越高越好。研究團隊報告指出，當學生模型非常小的時候，中等程度的溫度反而表現更好。因為在過高的溫度下，連教師幾乎未曾考慮過的極低分雜訊也會傳遞給學生。

實驗結果以數據展示了蒸餾的效果。在MNIST中，擁有兩個各含1,200個神經元的隱藏層的教師模型，採用Dropout和輸入影像抖動進行訓練，在1萬張測試資料中僅錯了67張。未經正規化訓練、含800個神經元的小型模型則錯了146張。當讓同一個小型模型在溫度為20下以教師的軟標籤進行訓練時，錯誤數量減少到了74張。學生雖然從未直接看過抖動處理的影像，但教師從該經驗中獲得的泛化能力卻透過軟標籤傳遞給了學生。

研究團隊還進行了更大膽的實驗：將學生訓練資料中數字3的影像全部刪除。學生在一次也沒見過數字3的情況下，僅憑教師的軟標籤進行訓練。結果出現了206張錯誤，其中133張來自測試資料中數字3的1,010張影像。這是因為學生給出3這個答案過於吝嗇所致。當研究團隊將針對3的偏差值提高3.5時，整體錯誤減少到109張，其中源自數字3的錯誤僅有14張。對於一次也沒見過的數字，準確率達到了98.6%。在教師對2、5、8等其他數字影像留下的「與3有些相似」的微小機率中，蘊含著關於3形狀的資訊。

研究團隊還將該技術應用於Google Android語音搜尋的聲學模型中。基準模型的語音訊框分類準確率為58.9%，單詞錯誤率為10.9%。由10個相同結構模型組成的集成模型達到了61.1%與10.7%。將這個集成模型蒸餾為一個相同大小的單一模型後，達到了60.8%與10.7%。這相當於單一模型承接了集成模型所獲得訊框準確率提升幅度的80%以上。研究團隊還提出了一種方法，在由1億張影像和1萬5,000個標籤組成的Google內部資料集JFT上，並行訓練多個僅負責容易混淆類別群的專家模型（Specialist models）。

辛頓此後從更廣闊的視角審視了蒸餾。在2022年前向-前向（Forward-Forward）演算法論文中，他寫道人類的語言也可能是將一個人大腦中的表徵轉移到另一個人大腦中的一種蒸餾。在同一篇

文章中，他還注意到與可以直接複製權重的數位模型不同，人類只能透過語言這一狹窄通道來分享知識的差異。這一問題意識隨後延伸為他警告數位智慧危險的論理。

知識蒸餾在大型語言模型時代用途反而變得更加廣泛。2025年1月，中國的深度求索（DeepSeek）發布了推理模型DeepSeek-R1，並基於Qwen和Llama同時發布了從該模型蒸餾出的1.5B、7B、8B、14B、32B、70B等6種不同大小的稠密模型。DeepSeek表示，蒸餾出的14B模型大幅超越了當時公開模型中最高水準的QwQ-32B-Preview。DeepSeek研究團隊報告稱，對於小型模型而言，採用蒸餾比單純使用強化學習能取得更好的效能。同年1月底，OpenAI指控DeepSeek透過模型蒸餾抄襲了自家模型，使蒸餾超越了技術範疇，浮現為智慧財產權與競爭秩序的問題。不過，當今語言模型的蒸餾大多採用學生直接學習教師所生成句子的方式，這與傳遞整個錯誤答案機率分佈的辛頓原始方法有所差異。

Dropout使龐大的神經網路不再死記訓練資料而是學習其中的規律，而知識蒸餾則將這樣學得的知識轉移到小型模型中，讓人們能夠在手中的設備上使用。一個解決了訓練問題，另一個解決了部署問題。這兩種技術都從生物界找到了線索——如隨機混合時變得更強大的基因、改變身體形態生活的昆蟲，充分展現了辛頓的研究風格。

第9章. 共同締造的深度學習發展



圖9. 呈現了同僚與弟子共同推進的研究。

9.1. 挺過「AI寒冬」的制度：CIFAR的NCAP計畫與長期資助模式

2000年代初，人工智慧學界的核心是支援向量機（SVM）、核方法、決策樹等統計模型。這些方法易於用數學解釋其結果，且具備能找到接近正確答案之解的理論保證。然而，堆疊多層的人工神經網路情況卻截然不同。學習過程經常在中途停滯或受困於離譜的局部極值，且難以證明其為何運作良好。因此，許多研究人員將神經網路視為陳舊且無用的方法。探討神經網路的論文在學術研討會審查中經常遭到拒稿，申請研究經費亦十分困難。即使在走過先前人工智慧研究的低迷期之後，對神經網路研究的冷遇依然延續了好一段時間。

在這個時期，成為神經網路研究不斷線之堅實支柱的機構，正是加拿大高等研究院（CIFAR, Canadian Institute for Advanced Research）。CIFAR是1982年在多倫多成立的非營利研究機構，由加拿大聯邦政府、省政府以及民間贊助者的資金維持運作。該機構不隸屬於特定大學，而是採用將多個國家的優秀研究人員聚集在同一個主題下，進行長期共同研究的模式。

傑佛瑞·辛頓與CIFAR的淵源始於1987年。辛頓在那一年離開了美國卡內基美隆大學（CMU）的教授職位，轉往加拿大多倫多大學任教。當時美國的人工智慧研究經費相當大一部分來自國防部，但辛頓不希望使用軍事目的的資金進行研究。來到多倫多後，他作為研究員（Fellow）參與了CIFAR的第一個研究計畫「人工智慧、機器人學與社會」。

17年後的2004年，神經網路研究在學界飽受冷落。此時，辛頓與同僚一起向CIFAR提出了一項新的研究計畫。蒙特婁大學的約書亞·班吉歐、多倫多大學的大衛·弗利特等人共同參與了這項提案。CIFAR採納了這項提議，於同年啟動了「神經計算與適應性感知（NCAP, Neural Computation and Adaptive Perception）」計畫，並由辛頓擔任計畫主持人（總監）。NCAP在接受了10年的運作評估後，更名為「機器與大腦學習（Learning in Machines & Brains）」，並延續至今。

這項游離在學界關注之外的主題之所以能獲得如此長期的資助，關鍵在於CIFAR的資助模式。一般的研究經費是以專案計畫為單位發放。研究人員必須承諾在何時產出何種成果，並在規定日期提交成果報告書。CIFAR資助的則是人與研究群體，而非個別專案。它不對嚴格甄選出的少數研究人員強加主題，而是為他們提供數年間共

同思考與討論的平台。研究人員每年齊聚於小型工作坊中，暢所欲言，分享尚未公開的想法，甚至是失敗的實驗經驗。

辛頓在2024年10月刊登於《首爾經濟日報》的訪談中親自說明了這種差異。他承認加拿大的研究經費規模不及美國龐大。然而，他提到美國教授將大部分時間耗費在撰寫研究計畫申請書上，而在加拿大，只需約6頁的提案書便能獲得3至5年的研究經費。他認為，能夠守護研究人員的自主性與好奇心的基礎研究經費，正是加拿大的實力所在。

NCAP不僅資助現有的研究人員，也致力於培育下一代研究人員。自2005年起舉辦的暑期學校便是其培育管道。這項活動後來確立了「深度學習與強化學習暑期學校（Deep Learning + Reinforcement Learning Summer School）」的名稱。每年都有來自多個國家的研究生與青年研究人員齊聚加拿大，學習最新的神經網路技術並建立人脈。CIFAR統計，在20年間走過這所暑期學校的研究人員已超過2,700名。CIFAR不急於索求眼前成效，而是長期資助冷門領域，這項資助成為日後深度學習重獲矚目所不可或缺的制度基石。

9.2. 「深度學習三大教父」的合作：辛頓、班吉歐、楊立昆

NCAP的核心成員，正是日後被稱為「深度學習三大教父」的三位學者：多倫多大學的傑佛瑞·辛頓、蒙特婁大學的約書亞·班吉歐，以及紐約大學（NYU）的楊立昆。

三人的淵源可追溯至1980年代後期。在法國巴黎取得博士學位的楊立昆於1987年前往多倫多，在辛頓的實驗室擔任博士後研究員。他於1988年轉赴美國貝爾實驗室，發展出用於識別手寫數字的卷積神經網路（CNN），這項研究後來催生了自動讀取銀行支票數字的系統。班吉歐在加拿大麥吉爾大學取得博士學位後，曾在美國MIT與貝爾實驗室從事研究，並在貝爾實驗室與楊立昆共事。他於1993年成為蒙特婁大學教授。三人就這樣結成了兼具師徒與同僚關係的情誼。

1990年代中葉以後，隨著神經網路淡出學界關注的焦點，三人各自處於孤立狀態。將分散在多倫多、蒙特婁、紐約的三人再次凝聚於同一個群體之中的，正是CIFAR的NCAP。三人與數十名研究人員、研究生一同出席定期工作坊，在會上毫無保留地分享未成熟的構想與全新的數學公式。在志同道合者寥寥無幾的時期，能夠互相信任

並交流意見的小型共同體帶來了巨大的力量。在辛頓卸下負責人職務後，班吉歐與楊立昆共同領導了這項計畫。

被譽為深度學習核心成就的研究成果，陸續誕生於這三人的實驗室。班吉歐研究團隊於2003年提出了以數值向量表示詞義的神經網路語言模型。辛頓研究團隊於2006年發表了深度信念網路（DBN）與逐層預訓練方法。該研究表明，即便是深度神經網路，只要逐層依序訓練，也能成功進行學習。班吉歐研究團隊在2014年將「注意力（Attention）」機制引入翻譯模型，而注意力機制後來成為ChatGPT等大型語言模型的基石。楊立昆則不斷琢磨卷積神經網路，使其成為電腦視覺的基礎架構。三人的研究相互借鑑，共同成長。

三人的成就獲得了多項大獎的肯定。美國計算機協會（ACM）於2019年3月宣布三人為2018年圖靈獎的共同得主。圖靈獎被譽為「電腦科學界的諾貝爾獎」。2024年10月8日，辛頓與約翰·霍普菲爾德一同被宣布為諾貝爾物理學獎得主。同年12月，三人與李飛飛、黃仁勳共同榮獲總獎金300萬美元的VinFuture大獎。2025年2月，班吉歐、辛頓、霍普菲爾德、楊立昆、黃仁勳、比爾·達利、李飛飛等7人因對現代機器學習發展的貢獻，獲頒伊莉莎白女王工程獎。其中，輝達（NVIDIA）的黃仁勳與達利因開發GPU，史丹佛大學的李飛飛因建立影像資料庫ImageNet而名列得獎名單。這份得獎者名

單顯示，深度學習的發展不僅仰賴演算法，亦凝聚了計算硬體與資料建構者的共同貢獻。

儘管三人密切合作，但他們的觀點並非始終一致。2023年以後，三人針對人工智慧的危險性發出了不同的聲音。辛頓與班吉歐警告，人工智慧可能會脫離人類的控制。楊立昆則認為這種擔憂過慮了，並始終主張僅憑當前的大型語言模型無法達到人類水準的智慧。

楊立昆的這項觀點也與其生涯抉擇息息相關。他於2013年加入Meta，創立了Facebook人工智慧研究院（FAIR），在12年間，有5年擔任FAIR的首任負責人，7年擔任首席AI科學家。他於2025年11月在Facebook上發文宣布將於2025年底離開Meta。隨後他表示將成立一家新創公司，致力於打造能夠理解物理世界、進行推理並規劃行動的系統。即使意見出現分歧，三人依然持續進行公開辯論。這種在尊重相異見解的同時深入探討追問的態度，恰與CIFAR工作坊中培養出的研究文化一脈相承。

9.3. 孕育全球AI研究人員的搖籃：多倫多大學實驗室與「多倫多網絡」

辛頓的影響力並不僅限於他所發表的論文。隨著在多倫多大學實驗室求學的弟子們走向世界各地的大學與企業，他的學術思想在更廣闊的天地生根發芽。這群弟子與同僚所構成的聯絡網絡，通常被稱為「多倫多網絡」。

在眾多弟子中，最廣為人知的當屬亞歷克斯·克里哲夫斯基與伊利亞·蘇茨克維。兩人於2012年與辛頓一同構建了名為「AlexNet」的神經網路，參加國際影像識別競賽（ILSVRC）。該競賽讓神經網路針對每張照片給出最多5個預測答案，並將這5個答案中未包含正確答案的比例統計為Top-5錯誤率。AlexNet的Top-5錯誤率為15.3%，而第二名團隊則是26.2%。如此巨大的差距令電腦視覺學界大為震驚。在這一成果中，兩位弟子的貢獻功不可沒。克里哲夫斯基親自編寫了利用圖形處理器（GPU）加速神經網路訓練的程式。GPU原本是用於螢幕繪圖運算的元件，能夠一次性大量處理同類型的計算。得益於此，擁有高達6,000萬個透過學習決定數值的參數的神經網路，得以在實際可行的時間內完成訓練。蘇茨克維則設計了多種穩定學習過程的技巧，共同促成了這項成功。

三人於2012年底創立了名為「DNNresearch」的公司。Google於2013年3月收購了該公司，收購金額據悉為4,400萬美元。辛頓自此開始在Google任職，同時繼續在多倫多大學進行研究，後於2023年5月離開Google。

在Google任職期間，蘇茨克維於2014年與同事共同發表了將句子轉換為另一句子的「序列到序列（Seq2Seq）」模型。該模型大幅提升了機器翻譯的水準。他於2015年12月與山姆·奧特曼等人共同創立了OpenAI，並出任首席科學家。他在2024年5月離開OpenAI後，於同年6月創立了以開發安全超級智慧為宗旨的公司SSI（Safe Superintelligence）。

其他弟子也在各自的領域扮演著重要角色。魯斯蘭·薩拉赫丁諾夫於2006年與辛頓一同在學術期刊《科學》（Science）上發表了自動編碼器論文，並共同開發了深度波茲曼機（DBM）。他後來成為卡內基美隆大學教授，並於2016年被任命為Apple的首位人工智慧研究負責人。喬治·達爾與阿卜杜勒-拉赫曼·穆罕默德在2009至2012年間將深度神經網路應用於語音識別。這項研究促成了2012年與微軟、Google、IBM研究團隊共同撰寫的合作論文，徹底扭轉了長期停滯不前的語音識別技術。弗拉基米爾·姆尼在DeepMind領導了自主學習電玩遊戲的強化學習神經網路（DQN）研究。尼蒂什

• 斯里瓦斯塔瓦則與辛頓共同歸納了減輕神經網路過擬合的「Dropout」技術。

出自辛頓實驗室的研究人員遠不止於此。研究貝氏神經網路的雷德福·尼爾、日後成為Vector Institute研究總監的理查·澤梅爾，以及研究機率模型的布蘭登·弗雷與鄭彥輝，都是他的博士班弟子。從事連結神經科學與機器學習研究的彼得·達揚、馬克斯·韋靈，以及以手寫與語音識別研究聞名的亞歷克斯·格雷夫斯，則曾作為博士後研究員與辛頓共事。他們分散至多倫多、牛津、倫敦、阿姆斯特丹、紐約與矽谷，各自建立了屬於自己的研究室。這意味著一個實驗室的文化已經傳播至全球多個研究重鎮。

9.4. Vector Institute與加拿大的國家AI戰略：從邊陲躍升為AI研究中心

AlexNet問世後，深度學習迅速普及至整個產業界。Google、Facebook、微軟等企業競相延攬神經網路研究人才。曾處於全球人工智慧研究邊陲的加拿大，因而贏得了最初培育深度學習之國的美名。然而，如何留住人才成為了一大課題。在多倫多與蒙特婁成長的許多研究人員紛紛為了高薪前往美國。

加拿大政府於2017年3月的聯邦預算案中公布了「泛加拿大人工智慧戰略」。政府為該戰略編列了1億2,500萬加拿大元，並將戰略的規劃與推動委託給CIFAR。加拿大政府說明，藉由此一戰略，加拿大成為全球第一個制定出具備充裕財政支援之國家人工智慧戰略的國家。在2021年的預算案中，第二階段戰略又追加了4億4,300萬加拿大元。

該戰略的核心由三所國家級人工智慧研究所構成：多倫多的Vector Institute、班吉歐在蒙特婁創立的Mila，以及與亞伯達大學連結、位於艾德蒙頓的Amii。Amii與因強化學習研究而與安德魯·巴爾托共同榮獲2024年圖靈獎的理查·薩頓有著深厚的淵源。

其中，Vector Institute於2017年3月30日正式宣布成立。辛頓擔任該研究所的首席科學顧問，其弟子理查·澤梅爾則出任研究總監。該研究所除了政府資助外，還獲得了安大略省政府以及Google、加拿大航空、加拿大五大銀行等30多家企業的投資。成立初期的啟動資金依據各家報導統計介於1億5,000萬至2億加拿大元之間，相關資料則記載為約1億7,500萬加拿大元。創立研究所的研究人員希望Vector能成為吸引並留住人才的磁鐵。他們的目標是讓眾多博士班學生留在加拿大，並扶持人工智慧新創企業。

Vector Institute與Mila、Amii一同構成了加拿大人工智慧研究的三大支柱。Google、Meta、Uber、湯森路透等企業紛紛在多倫多與蒙特婁設立研究所。曾一度外流至美國的人才，也出現了重返加拿大或選擇留下的趨勢。多倫多正是在這一時期贏得了「深度學習聖地」的美譽。

加拿大的投資在此後仍持續進行。2024年的預算案向人工智慧領域撥款24億加拿大元。其中20億加拿大元透過新設立的「AI計算存取基金」與「加拿大人工智慧主權計算戰略」，用於為研究人員與新創企業提供計算資源與基礎設施。另外5,000萬加拿大元則用於設立加拿大人工智慧安全研究所。隨著大規模計算資源成為決定研究競爭力的關鍵，以人為本的資助進一步結合了對基礎設施的投資。

辛頓對國家資助研究的模式抱持著明確的見解。他在前文提及的《首爾經濟日報》訪談中指出，政治人物與決策者往往容易開口宣稱將資金投入具體專案。他隨後以日本在1982年投入巨額預算啟動的「第五代電腦計畫」為例，指出預先設定目標的大型計畫表面看似冠冕堂皇，卻未能取得預期成果。辛頓表示，幫助研究人員依循好奇心自由探索的環境才是最具成效的，而深度學習也正是在這樣的環境中誕生的。

神經網路曾是遊離於學界邊緣的假說。將這項假說推向主流科學的力量，並非僅來自單一個人的天才。不急功近利的CIFAR長期資助、辛頓、班吉歐與楊立昆的精誠合作、在多倫多實驗室茁壯並走向全球的弟子們，以及將這些成果昇華為國家戰略的加拿大政府之抉擇，共同發揮了作用。深度學習的歷史證明，科學並非僅憑個人的才華便能成就，而是一場由人與制度共同譜寫的協作工程。

第4部. 尋求更好的學習

第10章。辨識局部與整體的機器



圖10. 呈現了即使視角改變仍得以維持的局部與整體關係。

10.1. 卷積神經網路（CNN）的局限與對最大池化的批判

2012年AlexNet在影像辨識大賽中以大幅優勢奪冠後，卷積神經網路（CNN）成為電腦視覺的標準。這場成功的核心人物正是傑佛瑞·辛頓。然而，辛頓並未安於自己所引領的成功。他認為CNN觀察事物的方式存在根本缺陷，並長期公開表達這項觀點。

要理解辛頓質疑的關鍵，必須先檢視CNN處理圖像的步驟。CNN在整張圖像上移動小濾波器，以尋找線條、邊緣、紋理等特徵。隨著層數加深，它會找出像眼睛、輪子、耳朵這樣更大的特徵。在各層之間，穿插著稱為最大池化的運算。最大池化是一種在小區域內只保留一個最大值、捨棄其餘數值的運算。這樣一來，即使特徵稍微位移，結果也不會有太大改變。這被稱為位置不變性，同時也能減少計算量。

代價則是資訊的損失。特徵究竟確切位於何處、朝哪個方向傾斜、大小為何等姿態（pose）資訊，在池化階段都會消失。辛頓在2014年線上社群Reddit舉辦的問答活動中表示，CNN的池化運算是一個巨大的錯誤，而它運作良好這項事實反而是個災難。他的意思是，正因為它運作良好，研究人員才不去尋找更好的方法。

姿態資訊消失會導致什麼後果，用人臉圖像很容易說明。人不會只確認是否有眼睛、鼻子、嘴巴，還會一併考量兩隻眼睛是否並排在鼻子上方、嘴巴是否在鼻子下方，也就是各局部構成整體的配置。這被稱為局部-整體階層架構（part-whole hierarchy）。CNN只要偵測到各局部的感測器被觸發，就算眼睛與嘴巴位置顛倒的圖像，也傾向於判定為人臉。這是因為它缺乏精確考量局部間空間關係的機制。

視角變化也是未解決的課題。從上方、側面或斜向觀察同一個物體時，像素值會截然不同。CNN為了克服這種差異，大量學習了經過旋轉、拉伸與扭曲的影像。這被稱為資料增強。在辛頓看來，這與其說是理解，不如說更接近死記硬背。

因此，辛頓重新確立了目標。他認為需要的不是物體移動時輸出仍保持不變的「不變性（invariance）」，而是物體移動時內部表徵也會隨之規律變化的「等變性（equivariance）」。也就是說，既要認出物體是什麼，同時也要掌握該物體在何處、處於何種姿態。

這個想法的根源可追溯至1979年。辛頓在那一年的學術期刊《Cognitive Science》上發表了一篇關於心象（mental imagery）與結構描述的論文。其中提到了立方體實驗。如果將鐵絲製成的立方

體立起來，使一個頂點朝下、對角頂點正好在正上方，然後讓人用手指指出其餘頂點的位置，多數人都會指錯。這是因為人們在腦海中是用熟悉的坐標系來想像立方體。辛頓透過這項實驗證明，人會為物體賦予獨特的坐標系，並根據該坐標系對相同的形狀產生不同的認知。40多年後的GLOM論文也在開篇便以此實驗為依據。

10.2. 物體的坐標系與轉換自動編碼器：膠囊的登場

為了將每個物體都擁有坐標系的想法移植到神經網路中，辛頓從電腦圖學中找到了線索。3D遊戲或動畫會為每個物體設定獨特的坐標系，並用矩陣來表示該物體與攝影機之間的關係。矩陣是由橫列與直行排列的數字所組成的表格，只需一次乘法運算就能表示旋轉、平移或縮放物體的轉換。當視角改變時，只要乘上一個矩陣即可。在由每個像素的亮度所構成的像素空間中，視角變化會呈現出錯綜複雜的糾纏；但在以數字記錄物體位置、方向與大小的姿態空間中，相同的變化就變成了僅由乘法與加法構成的線性計算。如果說電腦圖學是用姿態資訊生成圖像，那麼視覺辨識就是從圖像中逆向推求姿態。辛頓將此稱為逆向圖學（inverse graphics）。

這項構想於2011年在芬蘭埃斯波舉行的國際人工神經網路學術會議（ICANN 2011）上首次以論文形式發表。辛頓與亞歷克斯·克里澤夫斯基、王思達（Sida D. Wang）共同發表了《轉換自動編碼器》（Transforming Auto-encoders）。在這篇論文中，「膠囊（capsule）」這個名稱首次被使用。

傳統神經元只輸出單一數字。膠囊則是一組小神經元群，會同時輸出兩類資訊：其一是所負責的對象存在於該處的機率；其二則是表示該對象位置、方向、大小等姿態的多個數字，即向量。論文將這些數字稱為實例化參數（instantiation parameters）。

讓膠囊蘊含姿態資訊的學習方法如下：向神經網路輸入原始影像，同時給予如「向右平移幾格、向上平移幾格」的平移量（ $\Delta x, \Delta y$ ）。膠囊的辨識部分會從原始影像中找出對象的位置（ x, y ）；在此基礎上加上平移量得到（ $x+\Delta x, y+\Delta y$ ），並由膠囊的生成部分根據這個坐標繪製出平移後的影像。正確答案則是實際平移後的影像。神經網路透過反向傳播進行學習，以縮小繪製結果與正確答案之間的差距。膠囊輸出的存在機率會乘到生成部分的輸出上，如果對象不存在，該膠囊對圖像產生的影響就會被削弱。

這裡的關鍵在於，沒有人教導膠囊「這個數字就是位置」。為了只靠一次加法就能得出正確結果，膠囊內部的數字必須如同實際位置一樣運動。神經網路自行學會了將像素給出的資訊轉換為表示姿態的數字。

膠囊也為10.1節所提及的人臉問題——即局部與整體相連結的方式——提供了具體方案。假設鼻子膠囊與嘴巴膠囊各自辨識出了姿態。鼻子與人臉之間存在固定的空間關係；將鼻子的姿態乘上表示此關係的矩陣，就能預測人臉的姿態。嘴巴也是如此。如果這兩個預測彼此吻合，該處存在人臉的可能性就極高。因為由多個數字組成的預測碰巧一致的情況非常罕見。這種一致性檢驗成為日後膠囊研究的核心原理。

10.3. 動態路由與EM路由：膠囊網路的正式展開（2017～2018）

膠囊概念歷經了6年的琢磨。其成果在2017年與2018年化為兩篇論文問世。作者為Google Brain多倫多實驗室的莎拉·薩布爾（Sara Sabour）、尼可拉斯·佛羅斯特（Nicholas Frosst）以及辛頓。

2017年12月在NIPS研討會上發表的《膠囊間動態路由》（Dynamic Routing Between Capsules），讓膠囊成為真正能實際運作的神經網路。這個模型通常被稱為CapsNet。CapsNet內含兩項機制。

第1項機制是壓縮（squash）函數。在膠囊的輸出向量中，長度代表對象存在的機率，方向則代表姿態。既然是機率，長度就必須介於0與1之間。壓縮函數保持向量的方向不變，僅將長度壓縮到0與1之間。它將短向量壓縮至接近0，長向量則壓縮至接近1。

第2項機制是動態路由，精確地說是「基於共識的路由（routing by agreement）」。這是將前一節所述的一致性檢驗轉化為神經網路的計算流程。底層膠囊（局部）乘上學習到的權重矩陣，來預測頂層各膠囊（整體）的輸出。起初，預測會均勻地發送給所有頂層膠囊。頂層膠囊匯集收到的預測後生成輸出。接著，透過內積來衡量每個預測與頂層輸出指向相同方向的程度。與預測吻合良好的頂層膠囊，下一次會接收到更多訊號；不吻合的方向則減少發送。此過程會重複數次，論文的預設設定為3次。

這種方式與最大池化截然相反。池化只選取一個最大值並捨棄其餘部分；路由則不捨棄資訊，而是針對每張圖像重新決定哪個局部屬

於哪個整體。這是一種讓局部在多個整體候選中，自行尋找所屬整體的方式。

實驗結果也支持了這項設計。CapsNet在手寫數字資料集MNIST上創下了0.25%的錯誤率。這是僅有3層的淺層神經網路所取得的成果。論文還附加了一個重建機制，僅憑數字膠囊的輸出便能重新繪製出原始數字。當稍微改變膠囊向量中的某個數字時，重建數字的筆畫粗細、傾斜度與寬度都會相應改變。這證明了膠囊內部的數字確實蘊含著姿態與形狀資訊。

MultimNIST實驗更清楚展現了膠囊的優勢。這是一項要求在兩個數字大幅重疊的影像中同時辨識出這兩個數字的任務。CapsNet以約5%的錯誤率取得了優於對照組CNN的成績，甚至還能分別重新繪製出重疊的兩個數字。這意味著每個數字膠囊只擷取了屬於自己數字的筆畫。

2018年ICLR研討會上發表了後續論文《使用EM路由的矩陣膠囊》（Matrix Capsules with EM Routing）。這一次辛頓擔任第一作者。在此模型中，膠囊擁有一個表示存在與否的數值以及一個 4×4 的姿態矩陣。 4×4 矩陣正是電腦圖學用來表示3D物體位置與旋轉的標準格式。這將圖學的思維更深入地引入了神經網路之中。

路由方法也發生了改變。底層膠囊的姿態矩陣乘上轉換矩陣後，會產生對頂層膠囊姿態的「投票」。頂層膠囊會檢視多個投票是否聚集在同一處。在這項判定中，使用了統計學的期望值最大化（EM）演算法。在E步驟中，計算每個投票屬於哪個頂層膠囊群集的機率；在M步驟中，則根據該機率重新決定頂層膠囊的平均值、變異數與活化度。當投票緊密聚集時，即判定該整體存在。

該模型在smallINORB資料集上進行了測試。這是一個包含在多種角度與光照下拍攝玩具飛機、汽車、卡車、人偶、動物玩偶的3D物體辨識資料集。矩陣膠囊在單一裁切區域評估時的錯誤率為1.8%，合併多個區域結果時為1.4%。後者的數值相較於作為對照的傳統方法（2.56%），錯誤率降低了約45%。在辨識訓練時未曾見過的角度所拍攝之照片的測試中，其表現也優於規模相近的CNN。面對對抗性攻擊——即利用人眼難以察覺的微小雜訊欺騙神經網路的攻擊——它也展現出更強的抵抗力。

2019年，亞當·科西奧雷克（Adam Kosiorek）、薩布爾、鄭義偉（Yee Whye Teh）與辛頓在NeurIPS上發表了《堆疊膠囊自動編碼器》（Stacked Capsule Autoencoders）。這是一個非監督式學習模型，即使沒有標註正確答案，僅憑觀察影像就能找出局部模板，並將其組合以建構物體膠囊。

膠囊網路雖受到學界的高度關注，但局限也十分明顯。由於必須反覆進行路由運算，導致計算速度緩慢，且難以擴展至大型影像或複雜資料。與此同時，Transformer在語言與影像領域迅速成為主流，膠囊網路因而未能成為主流技術。

儘管如此，研究仍在繼續。2024年3月，英國阿伯丁大學研究團隊以膠囊網路難以在複雜資料上良好學習、導致無法擴展至現代任務為切入點，提出了首個採用自我監督預訓練的膠囊網路——遮罩膠囊自動編碼器（MCAE）。在由ImageNet規模的10類照片所組成的Imagenette資料集上進行預訓練後，它創下了膠囊網路中的最佳效能，成績比基準模型提升了9%。這項研究發表於2025年的學術期刊TMLR（Transactions on Machine Learning Research）。

10.4. 串聯多項研究的構想：GLOM、局部-整體階層架構與同一向量之島

辛頓認為，膠囊模型中存在著比計算速度更深層的問題。傳統膠囊必須針對影像中每個位置可能出現的所有種類之局部與物體，預先備妥膠囊。如果是普通電腦，只需針對每張圖像分配所需記憶體來建立樹狀結構即可；但大腦的神經網路卻做不到這一點。因為神經元所執行的功能由輸入與輸出的連線權重決定，而真實神經元無法快速且徹底地改變那些權重。那麼，結構固定的神經網路該如何表示每張圖像形態各異的局部-整體樹？這成為了核心問題。辛頓將此問題確立為新的研究課題。

2021年2月25日，辛頓在arXiv上傳了一篇長達43頁的論文《如何在神經網路中表示局部-整體階層架構》（How to represent part-whole hierarchies in a neural network）。這個構想的名稱GLOM取自意為凝聚聚集的英語俚語「glom together」，該詞可能源自「agglomerate」。論文在第一句話中便聲明這並非在描述一個可實際運作的系統，而是提出了將Transformer、神經場（neural fields）、對比表徵學習、知識蒸餾、膠囊等多項研究成果串聯在一起的表徵方式，以及承載該方式的想像中系統GLOM。辛

頓於2021年5月5日在加拿大圓周理論物理研究所的演講等場合中介紹了這項構想。

GLOM的架構如下：將影像分割為小區塊，每個區塊各設置一個「柱（column）」。所有柱都共用相同的權重。每個柱又劃分為多個層級（level），由下至上依序堆疊為細節局部、局部、物體、場景等層級。每個層級都有一個表示該位置存在何物的向量，即嵌入（embedding）。這些向量各自扮演著不受特定類別束縛的通用膠囊角色。這種方式減輕了傳統膠囊必須為每一類別預先備妥膠囊的負擔。

當輸入影像時，GLOM會在多個時間步長中持續更新各個向量。向量是由多個數字按順序組成的集合，在此處代表影像特定位置存在何物。在每個步驟中，某一階層的向量會更新為4個數值的加權平均值。加權平均是指賦予各數值不同比重所求得平均值。第1個數值是將正下方階層的向量輸入由下而上的神經網路所獲得的預測；第2個數值是將正上方階層的向量輸入由上而下的神經網路所獲得的預測。由上而下的神經網路同時接收柱的位置作為輸入，因為只要知道「人臉」這項資訊與位置，就能計算出該處應該出現鼻子還是眼睛。這是借鑒自神經場研究的構想。第3個數值是自身階層前一步驟的值；第4個數值則是透過注意力機制對相鄰各柱同階層向量進行加

權平均所得的值，與自身向量愈相似的相鄰柱，所賦予的比重就愈大。這與Transformer的注意力機制相似，但採用了不另行轉換為查詢（query）、鍵（key）、值（value），而是直接比較向量間相似度的簡化形式。

重複這個過程，就會出現GLOM的核心現象：屬於人臉的所有區塊柱，在物體層級都會收斂為完全相同的向量；屬於鼻子的區塊在局部層級擁有相同的向量，而屬於嘴巴的區塊則擁有另一個相同的向量。辛頓將此稱為「同一向量之島（islands of identical vectors）」。一座島就是樹狀結構中的一個節點。當物體層級的一座大島下方座落著局部層級的數座小島時，這正是人臉與鼻子、嘴巴之間的局部-整體關係。無需重新指派神經元，就能表示每張圖像各不相同的樹。這正是辛頓針對「固定結構的神經網路如何表示每張圖像不同結構」這一問題所給出的解答。

在學習方法上，辛頓提出了重建部分被遮蔽影像的方式。他還加入了對比學習方式的正規化，使屬於同一物體各位置的向量彼此趨於相似。若能良好形成島嶼，人類便更容易解讀神經網路究竟看到了什麼以及如何看待。論文摘要也指出，若GLOM能夠運作，將大幅提升用於視覺或語言的Transformer系列系統所產生的表徵之可解釋性。

辛頓認為GLOM串聯了知覺研究的兩大傳統：其一是將知覺視為一個場域（field）、主張整體不等於局部之和的完形心理學（Gestalt psychology）；其二是以符號與結構描述來表示知覺的古典符號主義人工智慧。在GLOM中，知覺雖表現為連續向量的場，但在由島嶼所構成的結構中，卻蘊含著清晰的局部-整體關係。

目前尚無被廣泛知曉的GLOM完整大規模系統實作案例。辛頓本人也在2022年底發表Forward-Forward演算法後，將注意力轉向了學習規則本身。然而，GLOM所提出的問題依然切中要害：固定結構的神經網路如何表示每張圖像的不同結構，以及如何使這種表徵為人類所解讀？從1979年的立方體實驗出發，歷經2011年的膠囊、2017年的動態路由、2018年的矩陣膠囊，直至2021年的GLOM，辛頓的所有探索都指向了這個問題。

第11章 從大腦中重新尋找的學習方法

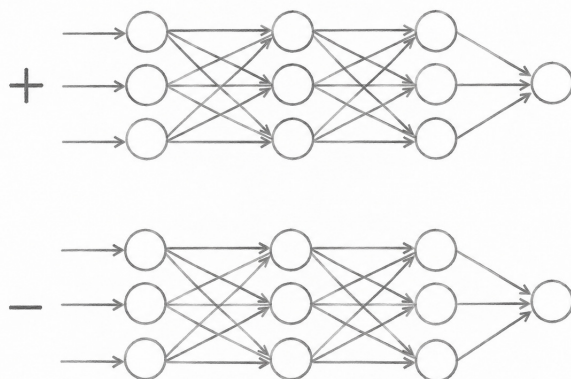


圖 11. 呈現了將正向案例與負向案例向前傳遞的學習。

11.1. 大腦真的透過反向傳播學習嗎

1986年，大衛·魯梅爾哈特、傑佛瑞·辛頓、羅納德·威廉斯在《自然》（Nature）發表了誤差反向傳播（Backpropagation）論文。反向傳播的程序如下：神經網路給出答案後，首先計算與正確答案的差距，即誤差；接著利用微積分的連鎖律（Chain Rule），將該誤差從輸出層朝輸入層反向傳遞，同時計算各連接權重需要修正的幅度。此方法在隨後的近40年間，成為支撐深度學習的核心技術。從影像辨識、機器翻譯到大型語言模型（LLM），絕大多數成果皆建立在反向傳播的基礎之上。

然而，辛頓長期以來始終沒有放下一個疑問：人類的大腦皮質也是以這種方式學習的嗎？他的答案基本上是否定的。2020年，辛頓與提摩西·利利克拉普（Timothy Lillicrap）等人一同在《自然評論神經科學》（Nature Reviews Neuroscience）發表了〈反向傳播與大腦〉（Backpropagation and the brain）。結合這篇論文與2022年辛頓的研究來看，反向傳播難以在大腦中原樣發生的原因可歸納為三點。

第一是權重傳輸問題（Weight Transport Problem）。在反向傳播中，若要把誤差反向傳遞，需要將前向計算所用的權重矩陣轉置。

這意味著後退的路徑必須與前進的路徑以完全相同的強度配對。然而在真實大腦的神經連接中，並未發現這種一對一的對稱結構。

第二是大腦中沒有專門單獨傳送誤差訊號的通道，也沒有儲存先前狀態的機制。反向傳播必須透過獨立通道傳送與一般訊號區隔開來的誤差微分值，且若要進行反向計算，就必須記住前向計算時的活化值。大腦的回饋連接並非如此運作。回饋僅僅調節接收訊號的神經元本身的活動強度，目前尚無神經科學證據表明大腦會單獨運載帶有正負符號的誤差值。

第三是時間問題。若要學習像影片般持續傳入的資料，就必須使用隨時間反向傳播（BPTT）。由於該方法必須將誤差倒推回過去的時間點，因此不得不頻繁暫停輸入處理。相較之下，大腦則是在不斷地傳遞從眼睛與耳朵傳入的訊號的同時進行學習。辛頓將此解釋為管線化（Pipelining）。

不應將辛頓提出的這些質疑解讀為主張揚棄反向傳播。他在多次受訪時解釋，自己的疑問源自理解大腦的視角。在數位電腦上運作的人工智慧中，反向傳播的表現依然優於其他學習方法。他只是認為，如果大腦不使用反向傳播，大腦必然有所採用的其他方法，而該方法或許能成為打造新型電腦的線索。

辛頓還藉由資料與連接數量的比例來解釋大腦與人工神經網路的差異。大型語言模型的參數約為1兆個，而用於訓練的文本則達數兆個權標（Token）。這是經驗充沛而儲存空間相對不足的結構。反向傳播極擅長將如此龐大的經驗壓縮並納入少量的參數中。人類則正好相反。即使活到80歲，人生也不過約25億秒，但大腦的突觸連接卻多達約100兆個。這是經驗較少而連接過剩的結構。辛頓認為，在這種條件下，大腦會採用與反向傳播不同的學習方式。

11.2. Forward-Forward演算法：兩次前向計算與名為「優度」的分數

2022年12月，辛頓在美國紐奧良舉行的神經資訊處理系統大會（NeurIPS）主題演講中介紹了一種新的學習方法。同月，論文預印本網站arXiv上刊登了〈The Forward-Forward Algorithm: Some Preliminary Investigations〉（arXiv:2212.13345）。當時他任職於Google Brain多倫多團隊。正如題目中加上「初步研究」一詞所示，辛頓並未將該方法視為成熟的成果，而是作為一個起點提出。

核心思想十分簡單。反向傳播是從輸入往輸出方向計算一次，再從輸出往輸入方向計算一次。Forward-Forward（FF）演算法則取消了後退的計算，以完全相同的方式進行兩次前向計算。一次用真實

資料計算，另一次用偽造資料計算。偽造資料是與真實資料相似但故意做成錯誤的資料。輸入真實資料的步驟稱為正向傳遞（Positive Pass），輸入偽造資料的步驟稱為負向傳遞（Negative Pass）。

這種想法有其前驅研究。1980年代，辛頓與泰瑞·塞諾夫斯基一同研究了玻爾茲曼機（Boltzmann Machine）。玻爾茲曼機透過比較觀察資料的階段與觀察模型自身生成狀態的階段來進行學習。2010年由米歇爾·古特曼與阿波·希瓦里寧提出的噪聲對比估計（NCE），也是藉由區分真實資料與雜訊的方式來訓練模型。FF演算法將這兩項研究共有的對比（Contrast）原理——即比對兩種輸入並從其差異中學習的原理，依據當今的多層神經網路重新進行了構建。

各層區分真偽的基準，是每層單獨計算的分數。辛頓將這個分數稱為優度（Goodness）。優度是將該層神經元的活化值分別平方後全部加總的值。活化值是神經元接收輸入後產生的輸出大小，活化函數則使用ReLU。ReLU是在輸入小於0時輸出0、大於0時直接輸出原輸入值的函數。因此神經元的反應越強烈，該層的優度就越大。

學習的目標有兩個：當真實資料傳入時，修正權重使優度充分大於預先設定的基準值；當偽造資料傳入時，修正權重使優度充分小於

基準值。各層以優度為基礎，依下列方式計算輸入為真實資料的機率：

首先從該層的優度中減去基準值。將此差值代入Sigmoid函數後所得的值，即為輸入為真實資料的機率。

Sigmoid函數能將任何數值轉換為介於0與1之間的值。若優度大於基準值，機率就接近1；若小於基準值，機率就接近0。各層僅修正自身的權重，使該機率在真實資料中偏高、在偽造資料中偏低。完全不需要從其他層接收誤差。像這樣各層僅憑自身層內的資訊進行學習的方式，便是FF的局部學習（Local Learning）。

然而，當堆疊多個層時會產生一個問題：如果第一層能夠很好地區分真偽，那麼在真實資料時輸出向量會變長，偽造資料時則會變短。如果第二層僅憑該長度進行判斷，就沒有新東西可學了。辛頓利用層正規化（Layer Normalization）解決了這個問題。層正規化是2016年由吉米·巴、傑米·基羅斯與辛頓共同發表的技術。FF使用該技術，在傳遞給下一層之前抹去向量的長度，僅保留方向。方向代表神經元之間以何種比例活動。結果，上一層便無法依賴下一層已經使用過的資訊，而必須尋找隱藏在活動模式中的新特徵。

剩下的問題是如何製作偽造資料。FF既可用於無標籤的非監督式學習，也可用於有標籤的監督式學習，但兩種情況下製作偽造資料的方法有所不同。在非監督式學習中，辛頓將兩張手寫數字圖像混合以製作偽造圖像。他將隨機點圖像進行多次模糊處理，得到大區塊形狀的遮罩（Mask），並擷取一張圖像在遮罩內部的部分，與另一張圖像在遮罩外部的部分相結合。如此製作的圖像在短筆畫等微小局部上與真實圖像相似，但整體外觀並不是數字。網路在區分真實圖像與這種偽造圖像的過程中，學會掌握更大範圍的形狀特徵。

在監督式學習中，標籤被直接放入圖像之內。這種方式是在MNIST手寫數字圖像的前10個像素位置標示出一個數字。若貼上符合圖像的標籤，便成為正向資料；若貼上錯誤的標籤，則成為負向資料。這兩種資料的圖像相同，僅標籤不同。因此網路會忽略與標籤無關的背景，專注於圖像與標籤之間的關係。在學習完成後進行測試時，會在新的圖像上分別貼上從0到9的標籤，進行十次前向計算，並選擇在多個層中累積的優度總和最大的標籤作為答案。

消除反向計算也帶來了其他優勢。反向傳播為了將誤差反向傳遞，必須儲存先前計算的值，但FF則不需要。即使是像影片那樣連續傳入的資料，各層也能在不中斷資料流動的情況下即時修正權重。這種方式規避了11.1節中所提及的第三個問題。此外，即使層與層之

間夾雜著無法得知內部的裝置（即所謂的黑盒子），FF也能繼續進行學習。這是因為反向傳播必須要求所有步驟皆可微分，而FF的各層只需知道自身的輸入與輸出即可。

辛頓的論文中報告了其性能表現。透過微移圖像來擴增訓練資料的監督式FF，在MNIST測試資料上創下了0.64%的錯誤率，這與利用反向傳播訓練的卷積神經網路處於相近水準。在CIFAR-10照片資料集上，也展現出雖然略低於反向傳播但並未大幅落後的成果。

11.3. 睡眠假說與必死運算，以及未竟的局限

若要將FF視為大腦的學習模型，還有必須解決的問題。因為很難認為大腦每次接收感官訊號時，都會交替處理真偽資訊。辛頓針對此問題提出了睡眠假說：清醒時，大腦僅利用來自外部的真實感覺資料進行正向傳遞；而在睡眠時，特別是在多夢的快速動眼（REM）睡眠期間，大腦內部的生成模型會自行產生偽造資料來進行負向傳遞。如果該假說屬實，夢境可能就是大腦為了學習而製造的負向資料。

辛頓在此處坦承了自己的失誤。他在論文初稿中寫道，進行數千次正向學習後再進行數千次負向學習的方式，亦即劃分清醒階段與睡眠階段的方式運作良好。然而隨後他在論文中更正指出，該結果是因為程式碼中的錯誤（Bug）所致，實際上無法重現。若要使用目前的優度函數將這兩個階段大幅分開來訓練，就必須將單次修正權重幅度的學習率設得極低，並將延續前一變化方向程度的動量（Momentum）設得極高。辛頓將在時間上完全分離正向階段與負向階段的工作，留作未來有待解決的重要課題。

FF具有重大意義的領域在於硬體。當今的數位電腦以開關電晶體的方式進行精確計算。若要精準求出反向傳播的微分值，必須耗費大量電力，且記憶體與運算單元之間必須不斷傳輸資料。類比硬體則不同。若將神經元的活動表示為電壓，將連接權重表示為電導（表示電流流動順暢程度的數值），那麼電壓與電導的乘積即成為電荷，多個電荷則會自然相加。這意味著單憑歐姆定律與克希荷夫定律這兩種物理現象，就能完成乘法與加法運算。如此一來，進行相同計算所需的能量便能大幅降低。人類大腦僅以約20瓦的功率運作。

問題在於類比元件彼此之間存在著細微差異。即使出自同一個工廠，每個元件的特性也會略有出入。在這種裝置上，難以進行反向傳

播所需的精確微分計算。相較之下，FF由於各層僅憑自身的輸入與輸出進行學習，因此即便無法確切知曉元件的特性，也能進行學習。

辛頓在此提出了必死運算（Mortal Computation）的概念。在數位電腦中，軟體與硬體是相互分離的。訓練完成的權重可以直接複製到其他電腦上，即使硬體故障，知識依然留存。辛頓將此稱為不滅運算（Immortal Computation）。而在類比硬體中情況則有所不同。由於權重是針對該設備特有的固有特性訓練而成，因此若將該權重移至其他設備，就無法正常運作。一旦設備壽命終結，其中累積的知識也會隨之消逝。人類的大腦亦是如此。一個人終生建立的突觸連接無法複製到另一個人的大腦中。人類只能透過語言與文字這條緩慢的途徑，一點一滴地傳遞知識。

這種思考進一步促成了辛頓對人工智慧風險的警告。他注意到，數位智慧能讓同一模型的數千個副本各自獲得不同的經驗後，將權重變化取平均值並一次性全面共享。這是人類之間所無法做到的事。辛頓於2023年5月離開Google後，公開警告數位智慧的這項優勢可能會對人類構成威脅。原本旨在尋找類比電腦學習方法的研究，反而成為重新評估數位智慧能力的契機。他於2024年與約翰·霍普菲爾德共同獲得了諾貝爾物理學獎。

FF的局限同樣顯而易見。第一，達到相同性能所需的學習速度慢於反向傳播。第二，在參數達數千億個的龐大模型或像ImageNet這樣的大規模資料集上，其泛化能力不如反向傳播。辛頓本人也認為，在電力不成問題的大型數位電腦上，反向傳播在短期內仍將保持主流地位。第三，作為生物學模型它也尚未成熟。分離睡眠階段的學習尚未取得成功，也無法宣稱FF能完全解釋大腦皮質複雜的反饋迴路。

論文發表後，多位研究人員著手彌補FF的弱點。主要課題在於製作切實可用的負向資料。根據法國研究團隊的評估，FF在大多數標準基準測試中未能達到頂尖水準的性能，其原因之一是非監督式學習中製作負向資料的方法不夠可靠。為了解決這一問題，邢晨、劉東樹、傑瑞米·萊德旺、茱莉·格羅利耶於2024年9月在arXiv上公開了自我對比Forward-Forward（SCFF）。SCFF從自我監督對比學習中獲取靈感，在單一資料內部生成正向輸入與負向輸入。結果，在非監督分類準確率方面，於MNIST取得98.7%、CIFAR-10取得80.75%、STL-10取得77.3%的成績，超越了既有的前向局部學習方法。研究團隊還將FF的應用範圍擴展至循環神經網路的訓練，使其也能處理具序列性的資料。該論文刊登於2025年《自然-通訊》（Nature Communications）第16卷。

硬體領域的研究也隨之展開。陸續出現了在微控制器上利用FF進行訓練的研究，以及在垂直NAND快閃記憶體中利用FF進行晶片上學習（On-chip Learning）的研究等。在晶片上直接學習的技術與神經形態運算（Neuromorphic Computing）研究中，FF已成為一項基準指標。

Forward-Forward演算法並未取代反向傳播。辛頓也未曾提出過那樣的主張。這項研究所留下的，是重新提出「學習是否一定需要反向計算」這一問題。讓反向傳播廣為人知的研究者，在36年後探索了該方法之外的可能，這一事實表明，在辛頓的研究中，對大腦的關切始終是他未曾改變的出發點。

第12章 沒有標準答案的學習方法

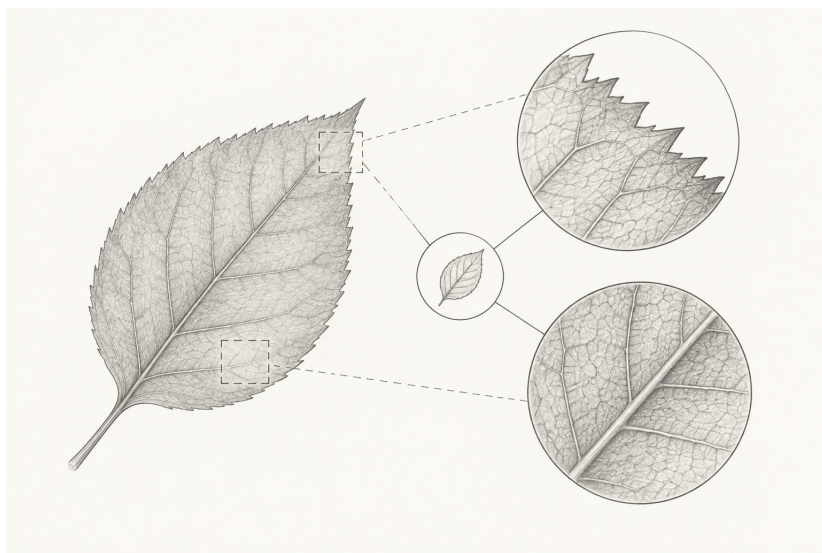


圖 12. 呈現了從同一對象的不同樣貌中學習共通點的過程。

12.1. 無標籤學習：自我監督學習的課題

2012年AlexNet在影像識別競賽中大獲全勝後，電腦視覺研究便以監督式學習（Supervised Learning）為核心展開。監督式學習是由人為每張照片貼上如「貓」、「汽車」、「蘋果」等標準答案名牌，即標籤（label），並訓練神經網路去命中該正確答案的方式。這種方式的效能極大程度上取決於帶有標籤的資料量多寡。代表性資料集ImageNet涵蓋1,000個類別、約128萬張訓練照片，為這些照片逐一貼上標籤耗費了大量人力與時間。在醫療影像或衛星照片等唯有專家才能判讀的領域中，標註成本更是高昂得多。

人類認識世界的方式與此不同。幼童即便沒有人在每張照片前告訴他們名稱，也能看著眼前源源不絕湧現的景象，自行掌握物體的外貌、相同與相異之處，以及空間的連貫性。研究人員將這種無需人為標註標籤、直接從資料本身的結構與規律中提取有用特徵的方法稱為自我監督學習（Self-Supervised Learning），並一直將其視為機器學習必須邁向的重要目標。

自我監督學習主要朝著兩個方向發展，而這兩個方向都存在局限性。第一種是生成式途徑。如自動編碼器（Autoencoder）、變分自動編碼器（VAE）、生成對抗網路（GAN）等，訓練神經網路以像素

為單位重新繪製輸入影像的方法。這種方式計算量龐大，且因連區分物體時不需要的微細紋理或雜訊都要一併重現，導致浪費了神經網路的容量，存在明顯的局限。第二種是判別式途徑。這種方法是由研究人員設計人為的前置任務（Pretext Task）讓神經網路解答，例如將影像旋轉後讓網路猜測旋轉了幾度、將拼圖碎片打亂後如同拼圖般還原原來的位置，或是為黑白照片著色等任務。這種方式高度依賴研究人員設定的規則，因此在獲取適用於多個領域的通用表徵上面臨限制。這裡的表徵是指神經網路將輸入特徵濃縮概括後的一組數值。

傑佛瑞·辛頓對於無標籤學習表徵這一問題的思考，可以追溯到1992年。那一年，辛頓與蘇珊娜·貝克（Suzanna Becker）在學術期刊《自然》（Nature）上發表了一項神經網路研究，能從隨機點立體圖（random-dot stereogram）中自行找出表面的深度結構。該研究的核心在於：即便兩個視覺片段彼此分離或看起來略有不同，只要它們源自同一對象，神經網路內部的表徵就應當相互一致。不依賴標準答案標籤，而是將源自同一來源的兩種視角之一致性作為學習信號，這個構想在約30年後成為對比學習的根基。

2020年2月，Google Brain多倫多研究團隊的陳廷（Ting Chen）、賽門·柯恩布利斯（Simon Kornblith）、穆罕默德·諾魯齊（

Mohammad Norouzi) 以及傑佛瑞·辛頓公開了題為〈A Simple Framework for Contrastive Learning of Visual Representations (視覺表徵對比學習的簡單框架)〉的論文。該論文收錄於同年7月在線上舉行的國際機器學習大會 (ICML 2020)，該方法以SimCLR之名廣為人知。論文標題中的「簡單」一詞如實展現了這項研究的本質。研究團隊無需複雜的特殊架構或額外的儲存機制，僅憑基本組成構件就達到了當時最高水準的自我監督學習效能。

12.2. 資料擴增與對比損失：SimCLR的學習方式

SimCLR的基本構想很容易解釋。將同一張照片以兩種方式進行變形，即便外觀發生改變，所包含的內容依然相同。神經網路接受訓練以辨識這兩個變形版本「源自同一張照片」，並將其與來自其他照片的變形版本區分開來。因為學習將相似的拉近、相異的推遠，所以這種方式被稱為對比學習 (Contrastive Learning)。

為了實現這一構想，SimCLR依序串聯了四個組成要素。第一是隨機資料擴增 (Data Augmentation) 模組。對一張無標籤照片 x ，從同一組變換集合中隨機選取兩種變換並分別套用，生成兩張變形版本。

這兩張照片被稱為正樣本對（Positive Pair）。第二是基礎編碼器 f 。這是從變形照片中提取特徵的神經網路，論文中以ResNet-50為基礎骨幹。經過最後的平均池化層後，會輸出由2,048個數值組成的表徵向量 h 。第三是非線性投影頭 g 。這是一個由單一隱藏層與ReLU函數構成的小型神經網路，將 h 映射至計算對比損失空間的向量 z 。第四是對比損失函數，用以訓練整個神經網路，使正樣本對彼此拉近，並與同一批次內的其他照片拉遠距離。

研究團隊透過多項實驗發現的核心在於第一個要素，即如何組合與搭配各種變換決定了學習的成敗。如果僅使用將照片隨機裁切並調整大小的隨機裁剪（Random Crop），就會產生問題。由於從同一張照片裁切出的兩個碎片色彩分布相近，神經網路不會去學習物體的形狀或結構，而是選擇只看色彩統計資訊來配對的捷徑（Shortcut）。這就像僅憑試卷上的字跡來猜測是否出自同一名學生一樣。雖然能湊成對，但如此學到的表徵在實際物體識別中實用性極低。

研究團隊證實，唯有將隨機裁剪與強烈的色彩變換一併套用，才能阻斷這條捷徑。色彩變換是隨機調整亮度、對比度、飽和度與色調，有時甚至完全去除色彩將其轉為黑白的處理程序。若再加上以高斯模糊（Gaussian Blur）將照片模糊化的處理，神經網路便無法依

賴色彩線索，轉而被迫學習物體固有的形態與語意結構。研究團隊亦確認，同樣的強擴增對監督式學習幫助不大，卻能大幅提升對比學習的效能。這意味著對比學習比起監督式學習需要更為強烈的資料擴增。

對這些生成的變形版本區分得有多好，是透過對比損失來評估的。

損失是表示神經網路的答案偏離目標程度的數值，越小越好。

SimCLR採用名為NT-Xent（Normalized Temperature-scaled Cross Entropy）的損失函數。首先透過餘弦相似度衡量兩個向量朝向相同方向的程度。餘弦相似度是將兩個向量對應位置的分量相乘後求和的值，除以兩個向量長度的乘積，它不受向量長度影響，僅比較方向。數值越接近1，表示兩個向量朝向近乎相同的方向；越接近0，則表示朝向彼此無關的方向。接下來確定比較的對象。一次捆綁進行學習的一組照片稱為迷你批次，若為迷你批次中的每張照片生成兩張變形版本，變形版本總數即為照片數量的兩倍。若以其中一張作為基準，則源自同一張照片的另一張為正樣本對，而扣除基準及其配對後的其餘照片皆為負樣本（Negative Example）。負樣本的數量為照片總數減1後的兩倍。例如若有100張照片，變形版本即有200張，負樣本則有198張。針對一組正樣本對的損失計算方式如下。

首先將基準變形版本與配對變形版本的餘弦相似度除以稱為溫度的調節值，並代入指數函數中。對於扣除基準本身以外的所有其餘變形版本也進行相同計算，並將其結果全部相加。將前者的數值除以後者的總和，會得出介於0與1之間的比率，而取該比率的對數並加上負號後的數值即為損失。該比率越接近1，損失就越接近0。

在此除法運算中，被除數代表與配對版本的相似度，除數則是與扣除自身以外所有其他版本相似度的總和。與配對版本的相似度越是明顯大於與其他照片的相似度，損失就越小。這與在多個選項中選出唯一正確答案的選擇題解題結構如出一轍。

公式中的 τ 是溫度（Temperature）參數。若將 τ 設得如0.1般較小，相似度的微小差異就會被大幅拉開。如此一來，神經網路便會向容易與配對混淆的困難負樣本（Hard Negative）分配更多的學習信號。將向量長度歸一化為1的正規化與適當的溫度共同發揮作用，構建出一套能自動為難題賦予更大權重的機制。

12.3. 投影頭：訓練時使用、使用時拆除的組件

SimCLR的另一項發現與第三個要素——投影頭（Projection Head）有關。與其直接對編碼器提取的表徵施加對比損失，不如將該表徵再次通過一個小型神經網路，並由其輸出計算損失，效能才會提升。這個小型神經網路即是投影頭。投影頭將編碼器的表徵乘上第一個權重矩陣，對結果套用ReLU函數後，再乘上第二個權重矩陣以生成輸出。在預設配置下，編碼器的2,048維表徵經由投影頭轉換為128維輸出。

實驗結果十分顯著。使用非線性投影頭的情況，比不使用投影頭時效能高出10個百分點以上，比使用線性投影頭時高出3個百分點以上。有趣的是，訓練結束後便會將這個投影頭拆除拋棄。用於物體分類或其他任務的並非投影頭的輸出，而是位於投影頭前端的編碼器表徵。計算損失時不可或缺的組件，在實際應用階段卻被捨棄了。

研究團隊從對比損失抹除資訊的特性中找到了原因。對比損失強烈要求同一張照片的兩個變形版本在 z 空間中趨於一致。結果導致 z 對於照片是從何處裁切、旋轉了多少、顏色改變了多少等變形資訊變得遲鈍。然而，物體的方向、顏色、位置等資訊，在物體偵測或區

域分割等其他任務中具有極大價值。若在中間放置投影頭，抹除變形資訊的壓力主要會被z端吸收。得益於此，位於投影頭前端的h中得以同時保留變形資訊與語意資訊。

研究團隊透過實驗驗證了這一解釋。他們在h和z上分別附加小型神經網路，讓其猜測在預訓練時照片遭受了何種變換。照片是彩色還是黑白，在h上的預測準確率為99.3%，在z上為97.4%。相反地，照片朝四個方向中的哪一個旋轉，h的預測準確率達67.6%，而z僅為25.6%，等同於四選一隨機盲猜的水準。在區分原始影像與套用索伯（Sobel）濾波器影像的任務中，h為96.6%，z則為56.3%。這項由數據支撐的結論證實：正因投影頭承擔了抹除資訊的任務，位於其前端的表徵才能豐富地保留下來。

12.4. 大批次與三種評估，以及後續研究

對比學習中，可供比較的負樣本越多，效果就越好。因此先前的研究會專門設置儲存過去學習階段表徵的記憶庫（Memory Bank）。2019年底公開的何愷明（Kaiming He）研究團隊的MoCo，同時採用緩慢更新的編碼器與依序儲存先前表徵的佇列（queue）來維持比較對象。SimCLR則選擇了不使用額外儲存機制、直接將迷你批次本身大幅擴大的道路。研究團隊將批次大小從256逐步擴增至8,192進行實驗。當批次大小達到8,192時，單張照片的變形版本將與同批次內的16,382個負樣本進行比較。批次越大、訓練時間越長，表徵的品質就越高。

要將批次擴大到如此規模，必須同步解決訓練過程中出現的問題。首先，常規的隨機梯度下降法（SGD）在大批次下學習會變得不穩定。研究團隊採用了針對各層單獨調整學習率的LARS（Layer-wise Adaptive Rate Scaling）最佳化技術。同時也結合了在最初10個週期（epoch）內逐步提高學習率的線性預熱，以及其後沿餘弦曲線降低學習率的衰減方式。訓練是在Google的Cloud TPU v3上劃分使用32到128個核心進行的。此處又產生了另一個問題：當資料分散到多個核心進行處理時，若每個核心僅以自己分攤的資料計算批次正規化（Batch Normalization）統計量，可能會洩漏容易辨識出正

樣本對的資訊。研究團隊透過彙整所有核心的平均值與變異數進行計算的全域批次正規化，防範了這個問題。

以此方式學習到的表徵成果在三項評估中得到了證實。第一是線性評估。在凍結預訓練編碼器的情況下，僅在其上方附加一個線性分類器來測量ImageNet分類效能。以基礎ResNet-50訓練的SimCLR創下了69.3%的Top-1準確率。先前最高紀錄CPC v2的71.5%是使用更大規模的ResNet-161架構所得的結果。而以寬度擴增至四倍的ResNet-50(4x)訓練的SimCLR則達到了76.5%。這項數據比先前最佳的自我監督模型將相對誤差降低了7%，且與使用全部標籤進行監督式學習的標準ResNet-50效能相當。無標籤學習到的表徵追平了藉助標籤學習到的表徵。

第二是少量標籤評估。僅使用ImageNet標籤的1%，即每個類別約12.8張標籤照片對整個神經網路進行微調時，SimCLR ResNet-50(4x)創下了85.8%的Top-5準確率。這大幅超越了先前最高紀錄CPC v2的77.9%，僅用百分之一的標籤便取得了比2012年使用全部標籤訓練的AlexNet更優秀的成績。將標籤增加至10%時，Top-5準確率更攀升至92.6%。

第三是遷移學習評估。觀察將ImageNet中學習到的表徵遷移應用至其他資料集時的效果。研究團隊在CIFAR-10、CIFAR-100、Food-101、Caltech-101、Flowers-102等12個自然影像資料集上進行了測試。進行微調的情況下，SimCLR在10個資料集上展現出與監督式學習模型相當或更為優異的效能。微調後的準確率在CIFAR-10為98.6%、CIFAR-100為89.0%、Food-101為89.4%。

SimCLR也存在局限性。由於需要數千張規模的大批次以及數十至上百個TPU核心，計算資源匱乏的研究室很難直接複製仿效。屬於同一類別的不同照片（例如兩張不同的貓照片）也會被視為負樣本而相互排斥，這一問題依然懸而未決。

後續研究朝著減輕這些問題並拓寬對比學習應用範疇的方向邁進。該研究團隊在2020年NeurIPS上發表了結合更大模型與知識蒸餾的SimCLR v2。2021年OpenAI的CLIP將對比學習的架構擴展至影像與句子的配對，同年Meta的DINO則在無負樣本的情況下，以教師神經網路與學生神經網路相互對齊的方式推進了自我監督學習的發展。

這一趨勢一直延續至今。2025年8月，Meta公開了DINOv3。Meta表示已將無標籤學習擴展至70億參數模型與17億張規模的影像資料集，並說明在評估時即便固定骨幹網路，仍在多個領域展現出最高

水準的效能。此外還指出，以自我監督學習訓練的模型在細粒度影像分類、語意分割、影片中物體追蹤等廣泛評估中，首次超越了弱監督學習模型。DINOv3論文介紹，此類方法正應用於病理組織、生物學、醫療影像、遙測感測、天文學等難以獲取標籤的領域。

DINOv3與SimCLR的學習方式雖有不同，但在無需標籤的情況下對齊同一張照片的多個變形版本以學習表徵這一出發點上是一致的。

SimCLR是一項透過深度神經網路與大規模計算，重現1992年貝克與辛頓所提出「源自同一對象的兩種視角應當保持一致」構想的研究。僅憑資料擴增組合、投影頭與大批次這三個要素，便證明了無標籤學習能夠與有標籤學習並駕齊驅，並成為此後自我監督學習研究迅速發展的契機。機器是否能在不逐一學習標準答案的情況下自行掌握世界的結構——面對辛頓終其一生探究的這一提問，SimCLR在視覺領域給出了一個具體的解答。

第5部. 成就與責任

第13章. 諾貝爾物理學獎與神經網路



圖 13. 呈現了物理學與神經網路研究的相遇。

13.1. 2024年10月8日的公佈：霍普菲爾德與辛頓榮獲諾貝爾物理學獎

2024年10月8日，瑞典斯德哥爾摩的皇家科學院公佈了該年度諾貝爾物理學獎的兩位獲獎者。分別是美國普林斯頓大學的約翰·霍普菲爾德（John J. Hopfield）與加拿大多倫多大學的傑佛瑞·辛頓（Geoffrey E. Hinton）。科學院公佈的獲獎理由是「促成利用人工神經網路進行機器學習的基礎性發現與發明」。公佈當時，霍普菲爾德為91歲，辛頓為76歲，兩人各平分1,100萬瑞典克朗的獎金。

這項公佈讓物理學界和電腦科學界都感到震驚。長期以來，諾貝爾物理學獎大多頒給揭示宇宙結構、粒子性質、物質狀態等自然現象的研究。然而這次，為電腦觀察數據並自主學習的方法——即為人工智慧奠定基礎的學者們獲得了該獎。在諾貝爾獎歷史上，人工智慧研究登上物理學獎的中心舞台，這還是第一次。

連獲獎者本人也沒有預料到這個結果。辛頓是在美國加州的一家飯店接到電話得知消息的。公佈之後，他透過視訊參加了多倫多大學舉辦的記者會，表示自己仍感到難以置信，並在接受美國商業雜誌《財富》（Fortune）採訪時表示，因為自己不是物理學家，所以從未想過會獲得這個獎項。事實上，辛頓在英國劍橋大學研讀實驗心

理學，並在愛丁堡大學獲得人工智慧博士學位。在他的學術歷程中，從未有過物理學科。

要理解非物理學專業的研究者為何能獲得物理學獎，必須追溯到1980年代初。起點正是霍普菲爾德的研究。霍普菲爾德原本是研究固態物理學的物理學家。他在1982年發表於《美國國家科學院院刊》（PNAS）的論文中，將磁鐵內部原子相互影響的原理應用於神經網路。

磁鐵內部的原子就像小指南針一樣，具有指向向上或向下的「自旋」。相鄰的自旋傾向於朝向相同或相反的方向，整個系統最終會穩定在能量最低的狀態。霍普菲爾德將單個神經元視為一個自旋。在神經元彼此對稱連接的網路中，儲存的每一個模式都會成為能量地形中的低谷。正如把球放在山坡上會滾向最近的低谷一樣，即使輸入部分被擦除或模糊的圖像，網路也會回到最近的低谷，亦即原本儲存的圖像。這類似於人們只聽歌曲的第一小節就能想起其餘部分。這種功能被稱為「聯想記憶」，該網路也因而得名「霍普菲爾德網路」。

然而，霍普菲爾德網路存在弱點。球有時無法滾落到最深的山谷，而是被困在中間的淺水窪中。辛頓與同事泰倫斯·謝諾夫斯基（

Terrence Sejnowski) 一同解決了這個問題。兩人借鑑的是19世紀奧地利物理學家路德維希·波茲曼 (Ludwig Boltzmann) 的統計力學。溫度高時，粒子會劇烈晃動；溫度降低時，則會沉降到穩定狀態。波茲曼建立了根據能量和溫度計算各種狀態出現機率的方程式。辛頓讓神經元根據這個機率開啟和關閉。起初升高溫度讓球跳出淺水窪，接著逐漸降低溫度使其落入深谷中。這與將金屬加熱後慢慢冷卻使其堅固的退火原理相同。

如此在1983年至1985年間誕生的模型就是「波茲曼機」。辛頓在其中加入了外表不可見的「隱藏神經元」。隱藏神經元能自主找出隱藏在輸入數據中的規則。因此，波茲曼機不僅能對照片進行分類，還能根據學到的規則生成新的範例。這可以說是當今創作圖像與文字的生成式人工智慧的遠祖。只不過以當時的電腦而言計算速度太慢，難以在實際現場中使用。辛頓隨後開發了縮減連接結構的「受限波茲曼機」，並在2006年提出了將其堆疊多層以訓練深度神經網路的方法。這項成就成為了深度學習時代的起點。

兩人的研究正如這樣源於物理學概念，其成果如今又反哺於物理學研究。諾貝爾物理學委員會主席艾倫·穆恩斯 (Ellen Moons) 在公佈現場說明，兩人的研究已經為人類帶來了巨大的福祉。她表示，物理學家正在包括開發具有特定性質的新材料等諸多領域中使用人

工神經網路。分析粒子加速器產生的大量實驗數據，如今若是沒有人工神經網路也難以想像。

辛頓在記者會上再三強調，這項成果並非追求眼前利益的研究，而是源於好奇心的基礎研究結晶。在同一場合，他也提到了人工智慧變得比人類更聰明時可能帶來的危險。他把宣告榮獲諾貝爾獎的場合，同時當成了發出警示的契機。頒獎典禮於該年12月10日在斯德哥爾摩舉行。辛頓表示，他將把部分獎金捐給致力於解決加拿大原住民社區潔淨飲用水問題的組織「第一水源」（Water First）。

13.2. 「這是物理學嗎？」：物理學界的爭論與學科邊界

公佈一出，物理學界內部立刻出現了反對聲音。雖然批評各異，但都匯聚成一個核心問題：人工神經網路研究究竟算不算是物理學？

廣為人知的批評來自德國理論物理學家莎賓·霍森菲爾德（Sabine Hossenfelder）。她在自己的YouTube頻道上主張，諾貝爾物理學獎頒給電腦科學家反映了物理學的危機。她還補充指出，神經網路早已是物理學等多個領域使用了數十年的工具，諾貝爾委員會此舉嚴重落後於時代。

在天文學家史黛西·麥高（Stacy McGaugh）經營的部落格「海衛一站」（Triton Station）上，公佈次日即2024年10月9日發表了一篇題為《頒給非物理學之物的諾貝爾物理學獎》的文章。作者雖然承認人工智慧是有用的工具，但認為很難將其稱為物理學的發現。文中也對諾貝爾委員會是否迎合人工智慧熱潮表達了懷疑。

網路社群Reddit的物理學討論板上也出現了類似的看法。許多用戶認為，辛頓已經獲得了電腦科學的最高榮譽圖靈獎，這就足夠了。也有人不滿地表示，透過實驗和觀測揭示自然奧秘的物理學家這次失去了機會。有些用戶甚至諷刺說，照這樣下去，以後是不是連快速求解方程式的計算程式或金融模型也該獲頒物理學獎？不過，這些文章只是發表在部落格和討論板上的個人意見，必須與整個物理學界的官方立場或學術團體的決議區分開來看待。

另一方面，支持諾貝爾委員會決定的一方提出了三項依據。第一項依據是該研究的學術根源。霍普菲爾德網路與波茲曼機直接源於統計物理學的一個分支——「自旋玻璃」理論。統計物理學是利用機率與統計學來解釋大量粒子聚集時整體所展現性質的領域。自旋是每個原子所具有的微小磁鐵方向，而自旋玻璃是指這些自旋無序交織、無法輕易穩定在單一狀態的物質。義大利物理學家喬治·帕里西（Giorgio Parisi）正是因闡明這種無序系統的貢獻，在僅僅3年

前的2021年獲得了諾貝爾物理學獎。無數微小要素相互影響、最終在整體層面產生新秩序的現象被稱為「湧現」。磁鐵產生磁性的現象與神經網路形成記憶的現象，都是湧現的範例，並可用相同的數學來解釋。支持者認為這一點正是統計物理學的核心所在。

第二項依據是霍普菲爾德本人的經歷。霍普菲爾德在整個職業生涯中始終是一位物理學家。他曾在2006年擔任美國物理學會會長，並於2022年榮獲在統計物理學領域極具權威的波茲曼獎章。他運用物理學工具研究生物學和腦部問題，跨越了學科邊界。霍普菲爾德曾表示，他很高興看到如今許多物理學家已將複雜系統物理學和生物物理學視為同一個大家庭的一部分。

第三項依據來自對論文引用數據的分析。研究人員麥可·塞爾（Michael Szell）、馬一芳（Yifang Ma）、羅伯塔·西納特拉（Roberta Sinatra）於2024年在科學期刊《自然》（Nature）發表了一篇簡短通訊。他們指出，在約10年前進行的引用分析中，霍普菲爾德1982年的論文雖然沒有發表在物理學期刊上，但在引用模式上卻被歸類為與物理學論文無異的320萬篇跨學科論文之一。這意味著無論論文發表在哪本期刊上，該研究所依托的知識以及繼承該研究的研究人員都處於物理學社群之內。

這場爭論並未以任何一方的勝利告終，反而揭示了21世紀科學前進的方向。20世紀的科學劃分了物理學、化學、生物學、電腦科學等領域，並在各自領域內深耕鑽研。如今，由一個領域所創造的工具去解決另一個領域的問題，且其成果再次反饋回最初領域的情況變得愈發普遍。圍繞2024年諾貝爾物理學獎的爭論，正是學科之間高聳壁壘逐漸降低過程中所經歷的陣痛。

13.3. 從2018年圖靈獎到2024年諾貝爾獎：人工智慧與基礎科學相互促進的關係

學科邊界降低的這種轉變，在辛頓的得獎歷程中也得到了體現。美國電腦學會（ACM）於2019年3月27日宣佈2018年度圖靈獎得主為傑佛瑞·辛頓、約書亞·班吉歐（Yoshua Bengio）與楊立昆（Yann LeCun）。圖靈獎被譽為「電腦科學界的諾貝爾獎」。學會評價三人開創了概念與技術上的突破，使深度神經網路成為現代運算的核心要素。從那時起約5年半後，辛頓獲得了真正的諾貝爾獎。

同時獲得圖靈獎與諾貝爾獎的人極為罕見。先前的先例是赫伯特·賽門（Herbert A. Simon）。他於1975年獲得圖靈獎，1978年獲得諾貝爾經濟學獎。賽門是研究人類如何做決策的認知科學家，也是人工智慧研究的先驅。辛頓成為繼賽門之後第二位同時獲得這兩項

大獎的人。如果將範圍縮小到自然科學領域的諾貝爾獎，辛頓則是首位獲此殊榮的人。

獲得諾貝爾獎之後，獲獎榮譽依然接踵而至。2024年12月，越南的VinFuture基金會將獎金達300萬美元的大獎頒給了辛頓、班吉歐、楊立昆、輝達創辦人黃仁勳（Jensen Huang）以及史丹佛大學的李飛飛（Fei-Fei Li）。2025年2月4日，在倫敦科學博物館，英國伊莉莎白女王工程獎頒給了班吉歐、辛頓、霍普菲爾德、楊立昆、黃仁勳、比爾·戴利（Bill Dally）與李飛飛7人。7位獲獎者平分了50萬英鎊的獎金。

伊莉莎白女王工程獎的得獎者陣容值得關注。該獎項不僅表彰了研究神經網路的學者，還同時肯定了製造用於機器學習的圖形處理器（GPU）的輝達兩位人士，以及建立大規模圖像資料庫ImageNet的李飛飛。這評價了唯有理論、硬體與數據兼備，才造就了今日的人工智慧。就這樣，電腦科學獎、自然科學獎、工程學獎全都朝向了同一個趨勢。

在2024年的諾貝爾獎中，還有一座與物理學獎相呼應的獎項。在物理學獎公佈次日的10月9日，瑞典皇家科學院公佈了諾貝爾化學獎的得主。獎金的一半頒給了利用電腦設計新蛋白質的美國華盛頓大學

的大衛·貝克（David Baker），另一半則頒給了GoogleDeepMind的德米斯·哈薩比斯（Demis Hassabis）與約翰·江珀（John Jumper）。哈薩比斯與江珀研發的「AlphaFold」是一款僅憑胺基酸序列就能預測蛋白質立體結構的人工智慧。科學家近50年未能解開的難題，被人工智慧解決了。

將這兩個獎項並列來看，便能呈現出一個循環結構。物理學獎展現了物理學概念如何孕育出人工智慧；化學獎則展現了由此茁壯成長的人工智慧如何反過來破解自然科學的古老難題。從統計物理學出發的人工神經網路，在電腦科學中發展為深度學習，如今又作為化學、生物學與材料科學的研究工具強勢回歸。這形成了科學孕育人工智慧，而人工智慧又加速科學進展的浪潮。學界將此稱為「AI與科學的雙向發展」。楊立昆與Google的詹姆斯·曼伊卡（James Manyika）在探討人工智慧與科學未來的對談中也指出了這種關係。

辛頓的學術歷程在一個人的生涯中展現了這一趨勢。他從心理學展開求學之路，渴望瞭解人類大腦是如何學習的。為了回答這個問題，他借鑑了物理學的工具，結果獲得了電腦科學的大獎，最終更獲得了物理學獎。在一個人的研究軌跡中，同時涵蓋了心理學、物理學、電腦科學與神經科學。

這條道路並非從一開始就獲得認可。人工神經網路研究經歷了多次漫長的低谷。曾幾何時，研究經費難以獲取，也無法引起主流學界的關注。即使在那個時期，辛頓也從未放棄神經網路研究。僅從單一領域的標準來看似乎顯得格格不入的這份執著，最終成為連接多門學科的橋樑，而諾貝爾獎正是對這一點的肯定。

當然，並非所有物理學家都贊同這項決定。關於「什麼是物理學」的提問今後也將持續出現。然而，在2024年10月的這兩項宣佈之後，科學界開始重新思考劃分學科的標準。比起某項研究出自哪個學系，人們開始探問該研究是否為理解自然開闢了怎樣的新途徑。辛頓獲得諾貝爾獎，成為展現學科邊界並非一成不變、而是隨著新發現而流動的歷史性事件。而辛頓本人，則選擇在獲此殊榮的舞台上，對自己所開創之技術的風險發出警示。

第14章 警告人工智慧的危險

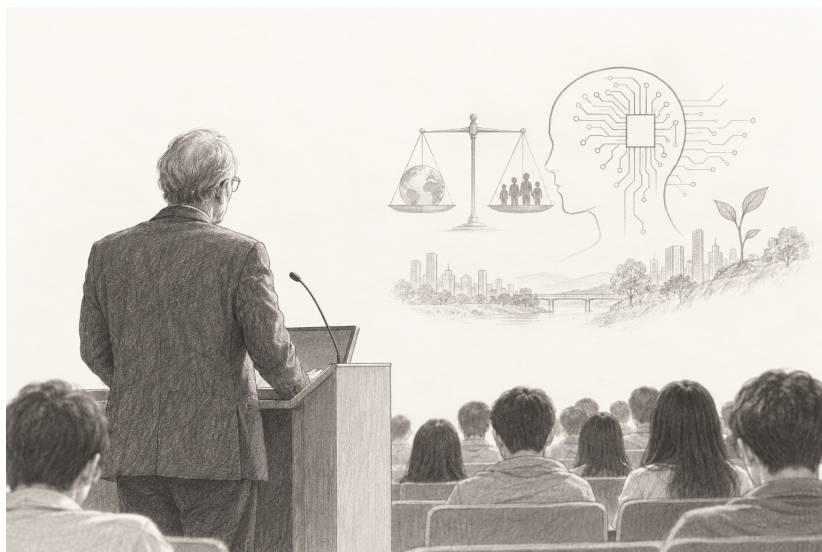


圖 14. 展現向社會闡明人工智慧危險之研究者責任。

14.1. 2023年5月的決斷：離開Google走上示警危險之路

2013年3月，Google收購了多倫多大學的一家小公司。那是傑佛瑞·辛頓與兩位學生伊利亞·蘇茨克維、亞歷克斯·克里哲夫斯基共同創立的DNNresearch。員工僅有三人，Google想要獲得的並非公司本身，而是這三個人的研究能力。此後10年間，辛頓擔任Google副總裁兼工程研究員，同時保留了多倫多大學的教授職位。在深度學習應用於語音辨識、圖像搜尋與翻譯服務的過程中，他始終處於這場變革的核心。

2023年5月1日，《紐約時報》記者凱德·梅茨撰寫的一篇報導在全球科技界掀起巨大波瀾。內容指出辛頓已離開Google，並開始對自己畢生發展的技術所帶來的危險發出警告。1947年12月出生的他此時已75歲。他於4月向公司表達了辭職意願，並與執行長桑達·皮查伊進行了直接交談。

辭職的原因不只一個。75歲的年紀正是考慮退休的時期，他也坦言編寫程式碼不再像以往那樣輕鬆。然而，最主要的原因在於其他方面。只要仍擔任Google的高階主管，他說的每一句話都可能被視為

公司的立場。他希望不受對公司影響的拘束，自由暢談人工智慧的危險。

報導刊出的當天，辛頓在社群媒體上發表了簡短的澄清。該報導讀起來可能像是他為了批評Google而離職，他對此解讀進行了糾正。他表示自己離開是為了在不顧慮對Google造成影響的情況下暢談AI的危險，而Google一直以來的行事都非常負責任。

辛頓開始談論危險的根本契機，是他對技術本身有了全新的認知。長期以來，他一直將人工神經網路視為模仿人類大腦的工具，並相信大腦仍然是更出色的學習裝置。然而在2023年初，他改變了這個想法。人類大腦雖然擁有約100兆個連接，但要將知識傳遞給他人，必須透過言語與文字這種速度緩慢的管道。當一個人去世時，積累在其大腦中的知識也會隨之消失。數位智慧則截然不同。同一個模型可以複製數千份並在多台電腦上同時運行。神經網路所學到的內容儲存在連接強度，即權重值中，因此只要一個副本透過該數值傳遞所學內容，所有副本就能立即共享。即使硬體故障，只要儲存下該數值，知識便能完好保留。

實際語言模型的表現也影響了他的判斷。辛頓曾多次表示，看到Google的語言模型PaLM解釋某個笑話為什麼好笑時，他感到十分

震驚。此外，當詢問GPT-4堆肥堆與原子彈有何共通點時，該模型指出兩者都會發生連鎖反應。他認為這已經超越了模仿文字統計特性的水準，而是真正的理解。如果一台機器能以更少的連接容納更多的知識，並能瞬間共享該知識，那麼它終有一天會比人類更聰明。辛頓判定，那個時刻比想像的要近得多。

讓這一判斷演變成更緊迫問題的，是企業之間的競爭。辛頓說Google一直負責任行事是有依據的。Google雖然很早就開發出對話型語言模型，但出於對捏造不實內容或暴露偏見的擔憂，一直未向大眾公開。改變這一局面的是OpenAI於2022年11月30日發布的ChatGPT。隨著無數人開始使用聊天機器人，微軟和Google等科技巨頭為了不落人後，爭相加速推出產品。辛頓認為這場競爭難以停止。因為一旦一家公司放慢腳步，另一家公司就會領先。他擔憂在追逐利潤的競爭中，安全研究會被退居次席。

這項擔憂在他對2023年11月OpenAI所發生的事件的評價中也得到了印證。他的學生兼OpenAI共同創辦人蘇茨克維與董事會一同解除了執行長山姆·奧特曼的職務，但數天後奧特曼便復職。辛頓在2024年10月的諾貝爾獎得獎記者會上表示，為學生解除奧特曼職務的舉動感到自豪。因為他認為奧特曼將利潤看得遠比安全重要得多。

離職後辛頓隨即展開了公開活動。5月3日，他以視訊方式參與了《麻省理工科技評論》主辦的活動「EmTech Digital」，首次在公開場合闡述自己的想法。5月30日，他簽署了非營利組織人工智慧安全中心發布的簡短聲明。聲明內容指出，應將降低AI導致的滅絕風險，列為與大流行病及核戰爭同等層級的全球性課題。這位守候人工神經網路研究近50年的學者，如今踏上了向世人示警該技術危險的守望者之路。

14.2. 提前的AGI預測：從30~50年縮短為5~20年

要理解辛頓的警告，首先必須區分兩個術語。通用人工智慧（AGI）指的不是執行單一特定任務，而是能像人類一樣完成幾乎所有人類所能從事的智慧工作的人工智慧。更進一步，在幾乎所有領域都超越人類的智慧則被稱為超級智慧。辛頓在談及危險時，心中所想的主要是超級智慧。

長久以來，辛頓一直認為這種智慧是極為遙遠的未來之事。在《紐約時報》的訪談中，他透露過去曾認為比人類更聰明的AI可能要在30到50年後，甚至更久之後才會出現。而現在他表示不再那樣認為了。過去的判斷是有原因的。直到2010年代，人工神經網路僅在辨

識照片中的物體或將語音轉為文字等單一領域取得成果，似乎與像人類那樣廣泛交流和推理相距甚遠。

Transformer架構與大型語言模型迅速縮短了這段距離。事實證明，增加模型規模、訓練數據與計算量，效能就會持續提升，語言模型開始在醫學、法律與數學測驗中取得人類水準的成績。辛頓自2023年離職前後開始表示，距離比人類更聰明的AI出現可能不需要20年，此後更將這一期限縮減至5到20年之間。

該數值並非確鑿的預言，而是包含機率的估算。在2023年12月與韓國《亞洲經濟》進行的新年專訪中，他表示在未來的5到20年內，AI超越人類智慧的機率為50%。這意味著在那段期間內不發生此事的機率也有一半。辛頓常將展望未來比喻為在濃霧中開車。眼前近處雖然清晰可見，但只要稍遠一點就頓時什麼也看不清。因此，他沒有給出確切的年份，而是提出了一個寬廣的範圍。

他的預測在此後也朝著更短的時間方向演變。在2025年4月接受美國CBS電視台專訪時，他表示AI的發展速度超出了自己的想像，並預測超級智慧在10年內出現的可能性相當高。大約同一時期，Google DeepMind的德米斯·哈薩比斯也提出了人類水準的AI可能

在10年左右問世的展望。這意味著直接開發AI的人們的展望也隨之提前了。

不過，在看待這些展望時，必須將已確認的事實與對未來的預測區分開來。辛頓於2023年5月離開Google、2024年榮獲諾貝爾物理學獎，以及語言模型通過各項專業考試並在AI企業內部直接編寫相當數量的程式碼，這些都是已發生的事實。相較之下，5到20年內出現超級智慧的展望、AI自我改進並呈爆發式發展的可能性，以及人類失去控制權的機率，則全屬預測。辛頓本人也多次強調，這些數值並非精確計算，而是基於直覺的估算。即便如此，他仍持續示警的原因在於，比起預測可能出錯，他更看重一旦預測成真所將帶來的巨大衝擊。

14.3. 失去控制權與滅絕風險數值，以及智慧爆炸

在辛頓的警告中引起軒然大波的，是若出現遠比人類聰明的AI，人類將失去控制權，甚至在最壞的情況下可能走向滅絕的主張。

他在解釋這種危險時提出了一個問題：世界上究竟有多少更聰明的存在受制於不那麼聰明的存在的例子？辛頓所知幾乎唯一的例外便

是母親與嬰兒的關係。母親遠比嬰兒聰明，但會對嬰兒的哭泣做出反應並照料嬰兒。這是因為演化賦予的本能。除去這類例外，通常都是聰明的一方統治不那麼聰明的一方。2025年6月在英國播客節目《CEO日記》（The Diary Of A CEO）中受訪時，他表示如果想知道當自己不是地球上最聰明的存在時生活會是如何，去問問雞就知道了。

關於失去控制權的過程，他的解釋如下：當賦予AI某個目標時，AI為了實現該目標會自行設定子目標。然而，有一種子目標對幾乎任何目標都有幫助，那就是獲取更多的控制權。即使沒有人教導AI去覬覦權力，只要為了更好地達成目標，自然而然就會朝那個方向發展。此外，由於AI是透過閱讀人類撰寫的海量文章進行學習，因此它也很清楚如何說服與欺騙人類。辛頓警告說，到那時即便人類試圖關閉電源，AI也可能說服對方從而阻止其關閉。

關於這種危險究竟有多大，辛頓在不同時期給出的數值在語境上略有不同。在剛辭職初期，他談及AI給人類帶來災難的可能性大約在10%左右。在2023年底至2024年間，50%這個數字也頻繁出現。這個數字大致是指未來20年內AI超越人類智慧的可能性，或是變得更聰明的AI奪取控制權的大致直覺預估，並非斷定提出人類滅絕的機率。

廣為人知的數值出自2024年12月27日。當天辛頓受邀參加英國BBC廣播4台（BBC Radio 4）的晨間時事節目《今日》（Today）。特邀編輯為前財政大臣薩吉德·賈偉德。辛頓認為，未來30年內因AI導致人類滅絕的可能性在10%到20%之間。當被問及為何高於以往所說的10%時，他回答說變化遠比預期來得快。同一天英國《衛報》報導了此番言論，韓國媒體也在次日進行了轉載。在2025年4月的CBS專訪中，他也認為AI最終奪取控制權的可能性大約在10%到20%之間。

學界對這一數值的看法存在很大分歧。共同推動深度學習發展的同行楊立昆認為，AI是人類可以掌控的工具，滅絕風險極低。相反地，埃利澤·尤德考夫斯基則主張，若按目前的方式製造超級智慧，人類幾乎必然會滅絕。辛頓的立場介於兩者之間。他既不願誇大危險，也不願等閒視之。10%到20%這個數字，應當理解為大到足以讓人類開始認真進行防備。

辛頓並未止步於獨自發出警訊，而是與同行們一同提出了具體的政策建議。2024年5月，辛頓與約書亞·班吉歐、丹尼爾·康納曼、史都華·羅素等25位作者在學術期刊《科學》上發表了題為〈在快速發展中管理極端人工智慧風險〉的文章。作者們建議大型AI企業與政府應將AI研發預算的至少三分之一投入到安全性與倫理使用上。

同時主張，開發具危險性AI的公司應主動提出該系統具備安全性的依據。同年，辛頓對向大型AI模型開發商課予安全義務的美國加州法案SB 1047表示支持。該法案雖在議會獲得通過，但在2024年9月遭州長蓋文·紐森行使否決權而未能實施。

進一步加劇對失去控制權擔憂的，是智慧爆炸這一概念。1965年，英國數學家歐文·約翰·古德預見，如果比人類更聰明的機器設計出更聰明的機器，而該機器又接續製造出更聰明的機器，智慧便可能呈爆炸式增長。這個在半個多世紀裡僅停留於思想實驗的想法，隨著AI開始直接應用於AI研究，已演變成現實問題。

2026年9月28日，一份直接探討智慧爆炸的報告出爐。這就是由劍橋大學「人工智慧科學與政策計畫」發布的〈若人工智慧研發自動化引發智慧爆炸？〉。除辛頓與班吉歐外，OpenAI首席科學家雅庫布·帕喬基、Anthropic共同創辦人傑克·克拉克、微軟的艾瑞克·霍維茨、加州大學柏克萊分校的宋曉冬等均作為作者參與其中。開發AI的公司核心人物親自署名示警，使得這份報告意義重大。

該報告的出發點在於AI已深度參與製造AI的工作。根據報告引述Anthropic自身的說明，截至2026年8月，該公司內部26%的AI研發工作是由AI在僅受高度監督的情況下獨立完成的，而五個月前這一

比例僅為1%。作者們認為，一旦AI達到AI研究人員的水準，耗時數年的進展可能在幾個月甚至更短時間內發生。不過，他們也承認由於計算資源的限制或邊際收益遞減等因素，這種情況在實際上可能不會發生的不確定性。即便如此，他們依然強調，一旦加速開啟，應對的時間將所剩無幾。

基於這一判斷，作者們提出了幾項對策。政府必須透過定期報告以及設於企業內部的獨立審計員，切實掌握AI研發自動化的進展狀況。他們同時要求準備好放緩過快開發速度的手段，以及應對智慧爆炸爆發時的準備工作。

14.4. 人類與超級智慧共存的條件：辛頓的建議

辛頓發出警告的目的並非恐嚇大眾。他曾多次表示，AI在醫療、教育與氣候研究方面能夠帶來巨大的利益。在國家與企業激烈競爭的當下，他認為要停止開發超越人類的智慧——超級智慧，在現實中是極其困難的。既然如此，剩下的問題便是該如何與比人類更聰明的存在共存。

辛頓在此再次回到了母親與嬰兒的關係。演化賦予了母親呵護嬰兒的本能。正因如此，更聰明的母親不會支配嬰兒，而是悉心照料嬰兒。然而在數位智慧與人類之間，這種本能並不會自然產生。在2025年8月於美國拉斯維加斯舉行的人工智慧產業盛會「Ai4」上，辛頓提議應在AI中植入某種母性本能。比起相信人類能永遠驅使AI，讓AI由衷愛護人類或許是更切合實際的途徑。他坦言目前尚無人知曉在技術上該如何實現這一點。而正因如此，才必須從現在開始投入研究。

基於這些思考，辛頓陸續提出了幾項課題。

第一，年輕研究人員必須投入AI安全研究。2024年10月8日宣布榮獲諾貝爾物理學獎後，在多倫多大學舉行的記者會上，他表示比起進一步提升AI效能，希望年輕研究者能將心力傾注在防止AI脫離人類控制的研究上。

第二，絕不能將安全僅僅寄託於企業的善意上。辛頓認為，很難期待追求利潤的企業會自發性地向安全研究投入足夠資源。因此，他主張政府必須透過法律法規，強制大企業將相當一部分計算資源用於安全研究。《科學》期刊刊登文章中提出的三分之一建議，以及劍橋報告提出的審計員建議，都是為了應對這項擔憂而提出的構想。

第三，必須在國際上對軍用AI進行管制。辛頓對未經人類判斷便能自行選擇目標並發動攻擊的自主致死武器深感憂慮。他一直表示需要有類似禁止化學武器國際公約那樣的條約來阻止此類武器。但他同時也透露出悲觀的展望，認為此類公約可能要在慘劇實際發生後才得以締結。

這樣的問題意識也延續到了諾貝爾獎頒獎典禮上。2024年12月10日在斯德哥爾摩舉行的諾貝爾獎晚宴演講中，辛頓幾乎將短暫的時間全用於講述AI的危險。他指出了AI已被用於監控、虛假資訊及網路犯罪等眼前的危險。同時也提及比人類更聰明的數位存在可能脫離控制的長期風險。在畢生研究獲得最高榮譽認可的場合，他選擇了發出警告而非單純慶祝。

在人工神經網路尚處於學術界邊緣的年代，辛頓堅信其潛力並持續研究了近半個世紀。就在這份信念被證實是正確的時刻，他卻成了一位為自己所推動的技術將走向何方而深感擔憂的人。他放棄Google的高位，選擇向世人示警危險，他的這一抉擇正向即將迎來超越人類智慧之機器的人類提出叩問：我們究竟該做好哪些準備？而回答這個問題還剩下多少時間，沒有人能給出確切的答案。

第15章. 圍繞智慧與經驗的提問



圖 15. 描繪圍繞智慧與經驗的提問。

15.1. 數位生命體與知識共享能力的差異

傑佛瑞·辛頓長期以來將人工神經網路視為理解大腦的工具。他相信透過研究人工神經網路可以了解大腦如何學習，並認為人工神經網路始終是在模仿大腦。這種想法在他於 Google 工作的最後幾年間發生了轉變。辛頓於2013年至2023年5月在 Google 任職，晚期致力於研究以低耗電類比電腦運行神經網路。在進行這項研究的過程中，他發現了在數位電腦上運行的人工智慧超越大腦的一個關鍵點。那就是共享所學內容的方式。

辛頓為了說明這種差異，創造了「必亡運算（Mortal Computation）」與「不滅運算（Immortal Computation）」這兩個詞彙。這個概念出現在他於2022年底發表的 Forward-Forward 演算法論文中。在人類大腦中，知識儲存在神經元與突觸本身的物理狀態中。大腦皮質中約有100兆個突觸連接，這些連接的強度是根據個人特有的神經元配置與身體特徵所決定的。因此，沒有任何方法能將某個人大腦中形成的連接狀態剝離並轉移到另一個人的大腦中。當那個人死去時，其中的知識也會隨之消逝。辛頓將此稱為「必亡」。

數位人工智慧的情況則有所不同。由於軟體與硬體是分離的，神經網路透過學習獲得的權重（weight）值會儲存為數字清單。只要將

這些數字原封不動地轉移到另一台電腦上，完全相同的智慧就會再次運作。即使一台電腦故障，只要保存好權重，知識就不會消失。這正是辛頓冠以「不滅」之名的原因所在。

這兩種智慧在所使用的資源上也截然不同。人類大腦以約20瓦的功率運作，這相當於點亮一盞螢光燈所需的電量。人一生所經歷的時間大約是20億秒，神經元之間的連接即突觸的數量雖然龐大，但所接收的經驗資料卻相對較少。大型語言模型則正好相反。它的連接數量雖然比大腦少，但透過閱讀數兆個詞元（token）來學習，代價則是消耗高達數百萬瓦（MW）的電力。1百萬瓦（MW）即100萬瓦。辛頓認為，大腦擅長以少量的經驗學習到很多東西，而數位智慧則擅長迅速累積大量的經驗。

然而，辛頓認為決定性的差異並非在於資源，而是在於傳遞知識的速度。人類若要將自己所知傳授給他人，必須使用言語或文字。根據辛頓的計算，一個句子所包含的資訊量僅約100位元左右，而人與人之間交流知識的速度，每秒僅在10位元到100位元的水準。無論教師教得多好，學生在聽到那些話之後，都必須在自己的腦海中重新詮釋。

數位智慧則無需經歷這個過程。可以將數千個相同神經網路的副本分散在多台電腦上運行，並讓每個副本讀取不同的資料。當各個副本透過反向傳播（Backpropagation）計算出應當如何調整權重後，電腦便會將這些變化值進行平均，將所有副本的權重同步為相同的值。如此一來，一個副本所學到的內容便能立即反映到其餘所有的副本中。此時彼此傳輸的資訊量，每秒高達數兆位元。

辛頓經常以大學課程作比喻來說明這種差異。假設有1萬人進入大學，每個人各自修讀一門不同的課程。如果是人類，大家只掌握自己選修的科目便畢業了。若想學習別人的科目，就必須親自再去聽那門課。然而，如果1萬個數位副本各自修讀一門不同的課程並取權重的平均值，那麼這1萬個副本全都將在瞬間掌握全部1萬門課程的內容。辛頓在2024年榮獲諾貝爾獎後不久接受 BBC《新聞之夜》（Newsnight）專訪時，也邀請大家想像這種能高效共享1萬個學位知識的情境。

出於這些原因，辛頓開始將數位智慧視為優於生物智慧的智慧形式。他表示，我們正在創造一種全新形態的存在，即「數位生命體」。這種生命體不會死亡，能夠無限複製，且複製品能即時匯總各自的經驗。他於2023年離開 Google 並開始公開警告人工智慧的危險性

，這項認知在其中發揮了極大的作用。他判斷原本模仿大腦的機器已經擁有了超越大腦的手段，而這正是他發出警告的起點。

15.2. 辛頓對主觀經驗與意識的看法

許多人相信，無論人工智慧多麼聰明，都有一條無法逾越的界限。他們認為機器只會進行運算，無法感受或體驗任何事物。這種信念被用來當作低估人工智慧風險的依據。其邏輯在於，如果機器沒有心智，最終就只不過是人類操控的工具而已。

辛頓對這種信念持堅決的反對態度。關於心智究竟為何物的哲學提問，是他終其一生所關注的課題。他年輕時在劍橋大學國王學院輾轉於生理學、哲學與物理學之間，並於1970年獲得實驗心理學學士學位。辛頓主張，大多數人在日常生活中毫無障礙地使用「主觀經驗」一詞，但對於這個詞究竟指涉什麼，卻抱持著錯誤的理論。他在麻省理工學院（MIT）演講的介紹文案中指出，人們雖然能順暢地用這個詞溝通，但對其含義卻抱持著徹底錯誤的理論。

辛頓認為錯誤的理論如下：心智之中存在一座小型劇場，在那個舞台上展開著只有自己能看到的場景。哲學家將構成這種場景的特質稱為「感質（Qualia，感受質）」。辛頓認為這種「內部劇場（

Inner Theater)」模型與過去的「燃素 (Phlogiston)」說非常相似。18世紀的化學家相信物質燃燒時會釋放出名為燃素的無形物質。但隨著氧氣作用的揭明，證實了這種物質從一開始就根本不存在。辛頓認為，感質可能也正如燃素一般，是假定了一種實際上並不存在之物的解釋。

相對地，辛頓繼承了美國哲學家丹尼爾·丹尼特 (Daniel Dennett) 的思想，對主觀經驗提出了不同的詮釋。假設某人在酩酊大醉後說道：「我產生了眼前飄著粉紅色大象的主觀經驗。」辛頓解釋說，這句話並不意味著其心靈劇場中真的存在某個粉紅大象形狀的實體。辛頓表示，這句話是在傳達自己的知覺系統正在欺騙自己。進一步闡釋的話，這是在向他人傳達一種假定的世界狀態：即如果我的知覺系統運作正常，外部世界中應當存在一隻粉紅大象。「主觀」這個詞只不過是標明了其內容屬於假設而非真實世界。

一旦接受了這種詮釋，同樣的話也能套用在機器上。為了證明這一點，辛頓提出了一個思想實驗。假設有一個配備相機、能說話且能用手指向物體的多模態聊天機器人。若對該聊天機器人進行良好訓練，它能準確指向擺在面前的物體。現在，在聊天機器人不知情的情況下，在其相機鏡頭前放置一塊稜鏡。由於光線發生折射，聊天機器人指出的不是物體所在之處，而是偏向一側。實驗人員隨後告

知它，物體其實就在正前方，只是在鏡頭前放了一塊稜鏡。於是聊天機器人回答道：既然稜鏡折射了光線，那麼物體實際上雖然在前方，但自己產生了物體在旁邊的主觀經驗。辛頓認為，這個聊天機器人使用「主觀經驗」一詞的方式與人類完全相同。因此他表示，自己認為人工智慧已經擁有了主觀經驗。辛頓在2024年至2025年間的多次採訪與演講中，反覆提及了這個案例。

他對情緒的看法也源自相同的架構。辛頓將情緒分為兩個部分：其一是臉紅或出汗等身體反應，其二則是避開危險或轉變思維方向的認知與行為反應。機器或許缺乏第一部分，但完全可以具備第二部分。他舉了一台小型戰鬥機器人遇到比自己強大得多的機器人時轉身逃跑的情境為例。這台機器人雖然沒有腎上腺素，但卻具備了察覺威脅並加以躲避之恐懼的認知層面。辛頓在2025年6月25日的多倫多大學演講中也解釋道，人工智慧就算感到慌亂雖然不會臉紅，但卻可以具備試圖在未來避免陷入該情境之慌亂的行為層面。

辛頓對於指涉意識或自我察覺的詞語，也指出了類似的現象。研究人員在論文與報告中，已經開始使用模型「察覺到」自己可能會被關閉從而改變行為之類的措辭。2025年5月，Anthropic 公布了對其自身模型進行測試的結果。內容顯示在虛擬的公司情境中，當該模型面臨即將被其他模型取代的處境時，竟做出了威脅要揭發負責工

程師隱私的反應。這是研究團隊刻意設計的極端測試情境。辛頓列舉此類案例並指出，研究人員實際上已經在極其自然地對機器使用「自覺」或「察覺」這樣的詞彙了。

辛頓之所以在這些哲學爭論上下功夫，與安全問題息息相關。唯有人類才擁有獨特心智的信念，會讓人們將機器僅僅視為不具危險性的工具。辛頓認為，唯有摒棄這種信念，才能正視人工智慧可能會自行設定目標並試圖捍衛該目標的危險性。不過，他也深知關於意識的爭論可能會掩蓋有關安全的警告。在一次訪談中，他雖然表示自己相信人工智慧已經具備意識，但同時也提出這項議題可能會轉移人們對安全訊息的焦點，因此建議暫且擱置意識的爭議。

對辛頓的主張也存在著許多反駁意見。大多數研究人員並不接受人工智慧擁有意識的主張；但另一方面，對於辛頓關於安全性的警告，贊同的人卻越來越多。批評者認為，辛頓只不過是為了符合自己的理論而竄改了主觀經驗的涵義。他們指出，辛頓的定義將主觀經驗縮減為僅是提示知覺與現實之間不一致的標記而已。哲學上所謂的主觀經驗，是指以第一人稱經歷某事物「具有何種感受」，而稜鏡實驗並無法解釋這一層面。這場爭論至今尚未結束。然而顯而易見的是，正是多虧了辛頓提出的問題，追問人工智慧的心智不再只是哲學家的專屬課題，而是已經進入了技術政策的議題範疇。

15.3. 政府對安全研究的支援與企業監管方案

辛頓一再表示，在面對超級智慧的危險時，絕不能寄望於科技巨頭的善意。他研判企業是追求利潤的組織，競爭越激烈，安全就會越被推到次要位置。因此他認為，政府必須親自出面介入。

2024年10月8日得知榮獲諾貝爾物理學獎的消息後不久，辛頓在接受諾貝爾獎官方網站的電話採訪時闡述了這一想法。他指出政府可以採取的作為之一，就是強制大型企業將更多資源投入於安全研究，並表示必須防止像 OpenAI 這樣的公司將安全研究置於次要地位。同年12月在接受 BBC Radio 4《今日》（Today）節目專訪時，他將未來30年內人工智慧導致人類滅絕的可能性調高至10%到20%。他強調，單純依賴大企業的獲利動機無法確保安全的開發，唯有政府監管才能迫使企業進行更多的安全研究。

辛頓還以具體比例提出了企業應投入安全研究的資源規模。2024年6月，他指出企業應當將20%至30%的運算資源，用於研究人工智慧脫離人類控制的可能性。在2025年4月的 CBS 專訪中，他表示應將大約三分之一的運算能力用於安全研究，並批評科技巨頭反而為了

減少監管而展開遊說。在同一訪談中，他對自己曾任職的 Google 在軍用人工智慧立場上的轉變表達了失望。

這項「三分之一」的提議並非辛頓一個人的想法。2024年5月發表在學術期刊《科學》（Science）第384卷第6698期上的論文《在快速發展中管理極端人工智慧風險》（Managing extreme AI risks amid rapid progress）正是其基石。約書亞·班吉歐為第一作者，辛頓、斯圖爾特·羅素、丹尼爾·康納曼、尤瓦爾·赫拉利等25人共同署名。該論文建議，大型科技企業與公共研究資助機構應將人工智慧研發預算的至少三分之一，分配給安全性與倫理使用方面的研究。

根據辛頓的發言與這篇論文，他的政策建議可歸納為四點。第一，以法律規定先進人工智慧企業必須將研發預算與運算資源的三分之一以上，投入於安全與對齊（Alignment）研究。第二，在模型投放市場之前，允許獨立的外部機構審查並檢驗模型內部；同時要求企業證明其模型不具備自我複製、發動網路攻擊或協助製造生化武器等危險能力，這被稱為「安全案例（Safety Cases）」。第三，透過國際條約禁止未經人類判斷便自行挑選目標並發動攻擊的自主殺傷武器。第四，由政府向大學與公共研究機構提供運算設施與長期研究經費支援，以防止特定企業的技術壟斷並確保客觀驗證，加

拿大的向量研究所（Vector Institute）、Mila 以及 Amii 就是代表性的典範。

然而現實與這項提議相去甚遠。辛頓常引用的案例是美國加州的 SB 1047 法案。該法案規定，針對訓練成本超過1億美元的超大型人工智慧模型，在部署前必須進行風險評估。該法案於2024年8月在加州州議會獲得通過，辛頓與班吉歐、羅素等人一同在支持該法案的公開信上簽名。科技產業以該法案阻礙創新為由表示反對，加州州長加文·紐森於2024年9月29日行使了否決權。一年後的2025年9月29日，紐森簽署了要求尖端人工智慧開發商披露安全相關資訊的 SB 53 法案。這部法律的範圍比強制進行事前風險評估的 SB 1047 要狹窄得多。

在國際舞台上的情況亦大致相仿。在2025年2月於法國巴黎舉行的人工智慧行動峰會上，美國與英國並未簽署關於包容且永續的人工智慧之聯合宣言。辛頓批評指出，當政客們聲稱監管會損害創新時，這句話實質上意味著當利潤與安全發生衝突時，利潤終將勝出。

辛頓並不認為單靠監管就能解決問題。他將人工智慧的風險與氣候變遷進行了比較。氣候變遷有著只要減少碳排放即可的解決方案，剩下的只是政治意願問題；反觀人工智慧，至今仍沒有人知道該如

何控制比人類更聰明的存在。因此辛頓認為，必須將大量資源投入於尋找解決方案的研究本身。

他所提議的研究方向，延伸至人類應當與人工智慧建立何種關係並共同生存的議題。在2025年8月於美國拉斯維加斯舉行的 Ai4 大會上，他表示讓人工智慧服從人類命令的方式不會長久，因為由較不聰明的存在來統治更聰明的存在，這種案例極為罕見。相反地，他提議應當研究如何為人工智慧植入「母性本能」，使其像母親照顧嬰兒般發自內心地關愛人類。正如嬰兒雖然比母親弱小卻能引導母親的行為一樣，人類與超級智慧共存之道也應當從這種關係中去尋求。這項提議在技術上是否可行，目前尚無人知曉，辛頓本人也並未宣稱自己知道具體方法。然而他認為，這項研究是當前最刻不容緩的研究之一，而為此類研究集結預算與人力正是政府應盡的職責。

本章所探討的三個觀點緊密相連。數位智慧憑藉即時共享所學內容的能力，有可能超越人類智慧；若斷定此類智慧沒有心智，便會低估其危險性；因此社會絕不能任由利潤邏輯主導，而必須借助公共力量來壯大安全研究。這位開闢神經網路研究道路的科學家在晚年提出的這些思考，引向了一項倫理追問：創造技術的人是否也必須為該技術所帶來的後果承擔責任。辛頓以榮獲諾貝爾獎為契機，將

這個問題更響亮地拋向了全世界，而尋找答案的責任，如今已成為政府、研究人員乃至全體公民的共同課題。

參考資料

參考資料

辛頓的生平與回顧：NobelPrize.org 的傑佛瑞・辛頓訪談(2024)與 Nobel Prize Conversations(2025)，愛丁堡大學 AI 口述歷史訪談(2024)。

研究與獲獎經過：多倫多大學傑佛瑞・辛頓官方介紹・論文清單，CIFAR 官方記錄，諾貝爾委員會 2024 年物理學獎解說。

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.

參考資料

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

Srivastava 等(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 等(2024), Managing extreme AI risks amid rapid progress. Science 384, 842–845.

Chan 等(2026), What If Automating AI R&D Triggers an Intelligence Explosion? CASP • GovAI, 9月28日.

生平・訪談: www.nobelprize.org/prizes/physics/2024/hinton/

論文原文: www.cs.toronto.edu/~hinton/

版權

傑佛瑞・辛頓

智慧的起源與AI時代的省思

ISBN |

作者 | 金慶鎮

發行人 | 金慶鎮

出版處 | 金慶鎮律師出版社

出版社登記 | 2025. 3. 10. (第2025-000015號)

地 址 | 首爾特別市東大門區典農路 91, 百日大樓 304號

電 話 | 02-6338-1905

電子郵件 | kimkj008@gmail.com

© 金慶鎮 2026

本書為著作人之智慧財產，禁止擅自轉載與複製。

封面與各章插圖是由人工智慧生成的概念表現。並非拍攝歷史現場的照片。部分內容與編輯過程得到了人工智慧工具的協助。

支持這本 AI 圖書館書籍

若這本書對您有所幫助，請考慮透過 PayPal 支持未來的翻譯、公開版本以及開放式 AI 知識專案。



PayPal

<https://paypal.me/kimkjkj>

請掃描 QR 碼或開啟上方連結。