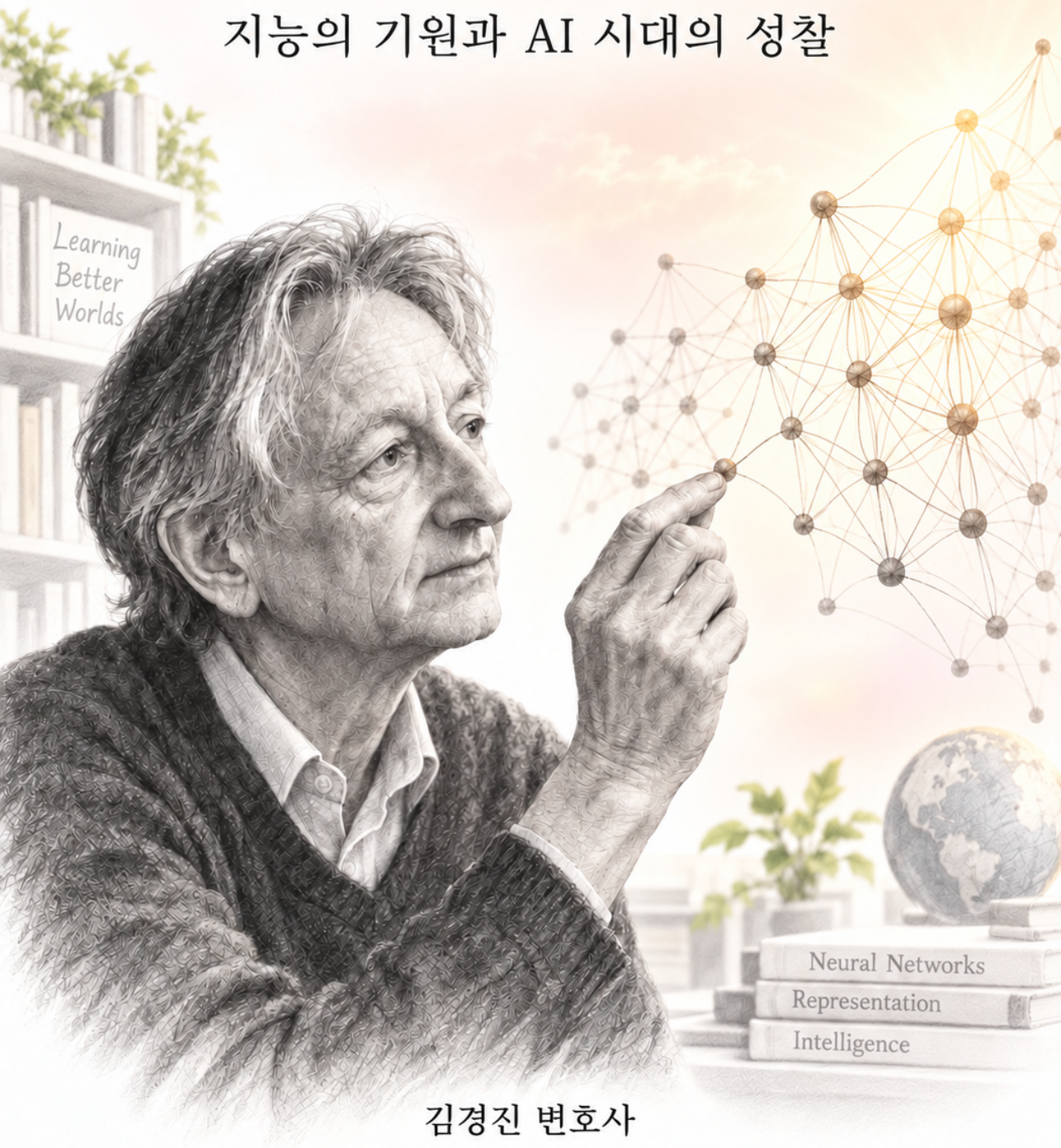


제프리 힌턴

지능의 기원과 AI 시대의 성찰



김경진 변호사

배우는 기계를 생각한 사람

제프리 힌턴은 사람의 뇌가 배우는 방식에서 기계 학습의 실마리를 찾은 연구자입니다. 그는 사람이 규칙을 하나하나 입력하는 대신 자료에서 스스로 학습하는 기계를 구상했습니다. 신경망 연구가 학계에서 주목받지 못하던 시기에도 그 질문을 놓지 않았습니다.

이 책은 어린 시절의 호기심에서 출발해 동료와 제자들과 함께한 연구를 따라갑니다. 1986년 역전파 연구, 2012년 알렉스넷의 성과, 2018년도 튜링상과 2024년 노벨 물리학상에 이르기까지 신경망이 오늘의 인공지능을 가능하게 한 과정을 살펴봅니다. 낯선 개념은 연구자가 풀려고 했던 문제와 함께 설명합니다.

힌턴의 삶에는 성취만 있는 것이 아닙니다. 그는 2023년 구글을 떠나면서 인공지능의 위험을 더 적극적으로 경고했습니다. 자신이 발전시킨 기술이 사회에 어떤 영향을 줄지 묻기 시작한 것입니다. 배움을 이해하려는 질문과 기술의 책임을 묻는 질문은 그의 삶에서 이어집니다. 이 책이 청소년 독자에게 인공지능의 원리와 한계를 이해하고, 그 미래를 스스로 생각하는 계기가 되기를 바랍니다.

김경진 변호사

차례

제1부. 뇌와 학습을 향한 질문	9
--------------------------	----------

제1장. 호기심에서 시작한 뇌 연구	11
----------------------------	-----------

1.1. 불(Boole) 가문이 남긴 논리와 확률	12
1.2. 일곱 살의 깨달음과 길을 찾아 헤맨 학창 시절	14
1.3. 에든버러에서의 고립과 지도교수 롱게이히긴스와의 갈등	17
1.4. 기호주의의 한계와 조용히 시작된 연결주의의 반격	20
1.5. 단기 성과보다 호기심: 기초연구가 맺은 결실과 그 의미	23

제2장. 기억은 어떻게 연결되는가	26
---------------------------	-----------

2.1. '할머니 뉴런'의 한계와 분산 표현의 오케스트라	27
2.2. 라호야에 모인 연구자들: 1979년 워크숍과 PDP 그룹의 탄생	31
2.3. 병렬 연상 기억과 가중치 최적화의 원리	34
2.4. 마음을 보는 관점의 변화: 기호 조작자에서 데이터 학습자로	37

제3장. 틀린 답에서 배우는 신경망	41
----------------------------	-----------

3.1. 다층 신경망의 난제와 역전파(Backpropagation)의 재발견	42
3.2. 1986년 《네이처》 논문과 계산 원리	45

3.3. 은닉층이 스스로 배운 표현: 거울 대칭 판별과 '가족 관계도(Family Trees)' 과제	49
3.4. 카네기 멜런 대학교(CMU) 시절과 캐나다 토론토 이주	52
제2부. 신경망을 배우게 하는 방법	56
제4장. 물리학에서 찾은 학습의 원리	58
4.1. 로체스터의 만남: 홉필드 네트워크에 확률과 열을 더하다	59
4.2. 에너지 지형, 확률적 이완, 그리고 숨은 유닛	62
4.3. 깨어 있는 단계와 꿈꾸는 단계: 볼츠만 머신의 학습 알고리즘	66
4.4. 2024년 노벨 물리학상: 통계물리학과 인공지능의 만남	70
제5장. 더 빠르게 배우는 방법	74
5.1. 너무 느린 볼츠만 머신과 '전문가들의 꿈'이라는 발상	75
5.2. 제한된 볼츠만 머신(RBM)의 구조와 조건부 독립	79
5.3. 대조적 발산(CD) 학습법의 원리	82
5.4. RBM의 변형과 사전학습으로의 확장	86
제6장. 깊은 신경망이 다시 주목받다	90
6.1. 제2차 AI 겨울과 기율기 소멸 문제	91
6.2. 한 층씩 쌓아 학습하기: 탐욕적 층별 비지도 사전학습과 DBN	94
6.3. 2006년 《Science》 논문: 심층 오토인코더와 비선형 차원 축소	98

6.4. 심층 볼츠만 머신(DBM)과 딥러닝의 확산	101
제3부. 딥러닝이 세상으로 나온다	106
제7장. 음성과 사진을 알아보는 컴퓨터	108
7.1. 첫 번째 돌파구: 음성 인식의 변화 (2009~2012)	109
7.2. 컴퓨터 비전의 전환점: 페이페이 리의 ImageNet과 AlexNet (2012)	112
7.3. 딥러닝으로의 전환, DNNresearch 창업과 구글 인수	116
제8장. 과적합을 줄이고 지식을 전달하다	120
8.1. 늘어난 파라미터와 과적합 문제: 드롭아웃의 탄생	121
8.2. 드롭아웃의 계산 방식과 실험 결과	124
8.3. '어둠의 지식'과 곤충의 변태: 지식 증류의 구상	127
8.4. 소프트맥스 온도 조절과 작은 학생 모델로의 지식 전달	130
제9장. 함께 만든 딥러닝의 발전	136
9.1. 'AI 겨울'을 견디게 한 제도: CIFAR의 NCAP 프로그램과 장기 지원 방식	137
9.2. '딥러닝 3대 대부'의 협력: 힌턴, 벤지오, 르쿤	140
9.3. 세계 AI 연구자를 길러 낸 곳: 토론토대학교 연구실과 '토론토 네트워크'	143
9.4. Vector Institute와 캐나다의 국가 AI 전략: 변방에서 AI 연구의 중심으로	146
제4부. 더 나은 학습을 찾아서	150

제10장. 부분과 전체를 알아보는 기계 **152**

- 10.1. 합성곱 신경망(CNN)의 한계와 맥스 풀링에 대한 비판 153
- 10.2. 물체의 좌표계와 변환 오토인코더: 캡슐의 등장 156
- 10.3. 동적 라우팅과 EM 라우팅: 캡슐 네트워크의 본격적인 전개 (2017~2018) 158
- 10.4. 여러 연구를 하나로 묶는 구상: GLOM, 부분-전체 위계와 동일 벡터의 섬 162

제11장. 뇌에서 다시 찾는 학습 방법 **167**

- 11.1. 뇌는 실제로 역전파로 배울까 168
- 11.2. Forward-Forward 알고리즘: 두 번의 순방향 계산과 '좋은'이라는 점수 171
- 11.3. 수면 가설과 필사적 연산, 그리고 남은 한계 175

제12장. 정답표 없이 배우는 방법 **180**

- 12.1. 라벨 없이 배우기: 자기지도 학습이라는 과제 181
- 12.2. 데이터 증강과 대조 손실: SimCLR가 학습하는 방법 184
- 12.3. 투영 헤드: 훈련에는 쓰고 사용할 때는 떼어 내는 부품 187
- 12.4. 큰 배치와 세 가지 평가, 그리고 이후의 연구 189

제5부. 성취와 책임 **194**

제13장. 노벨 물리학상과 신경망 **196**

- 13.1. 2024년 10월 8일의 발표: 홉필드와 힌턴의 노벨 물리학상 수상 197

13.2. "이것이 물리학인가?": 물리학계의 논쟁과 학문의 경계	201
13.3. 2018년 튜링상에서 2024년 노벨상까지: 인공지능과 기초 과학이 서로를 키우는 관계	204
제14장. 인공지능의 위험을 경고하다	208
14.1. 2023년 5월의 결단: 구글을 떠나 위험을 알리는 길로	209
14.2. 앞당겨진 AGI 예측: 30~50년에서 5~20년으로	212
14.3. 통제권 상실과 멸종 위험 수치, 그리고 지능 폭발	214
14.4. 인간과 초지능이 함께 살기 위한 조건: 힌턴의 제안	218
제15장. 지능과 경험을 둘러싼 질문	222
15.1. 디지털 생명체와 지식 공유 능력의 차이	223
15.2. 주관적 경험과 의식에 대한 힌턴의 생각	226
15.3. 정부의 안전 연구 지원과 기업 규제 방안	230
참고자료	234

제1부. 뇌와 학습을 향한 질문

제1장. 호기심에서 시작한 뇌 연구



그림 1. 어린 시절의 관찰과 과학적 호기심을 표현했습니다.

1.1. 불(Boole) 가문이 남긴 논리와 확률

제프리 에버레스트 힌턴(Geoffrey Everest Hinton)은 1947년 12월 6일 영국 런던의 뎀블던에서 태어났습니다. 그의 집안은 여러 대에 걸쳐 수학자, 과학자, 의사, 저술가를 배출한 학자 가문이었습니다. 조상 가운데 널리 알려진 인물은 고조부 조지 불(George Boole)입니다. 불은 19세기 영국 수학자로, 참과 거짓을 1과 0이라는 숫자처럼 다루어 계산하는 방법인 불 대수를 만들었습니다. 오늘날 모든 컴퓨터의 논리회로는 이 불 대수를 바탕으로 설계됩니다. 불의 아내 메리 에버레스트 불(Mary Everest Boole)도 수학자이자 교육자였습니다. 메리의 삼촌 조지 에버레스트(Sir George Everest) 경은 인도 측량총감을 지냈고, 세계에서 가장 높은 산인 에베레스트산의 이름은 그에게서 왔습니다. 힌턴의 가운데 이름 에버레스트는 이 집안 내력에서 왔습니다.

아버지 쪽 힌턴 집안도 학자 집안이었습니다. 증조부 찰스 하워드 힌턴(Charles Howard Hinton)은 4차원 공간을 눈에 보이게 그려 내는 연구로 이름난 수학자였고, 조지 불의 딸 메리 엘런과 결혼했습니다. 찰스의 아버지이자 힌턴의 또 다른 고조부인 제임스 힌턴(James Hinton)은 외과 의사이자 저술가였습니다. 아버지의 사촌 누이 조안 힌턴(Joan Hinton)은 제2차 세계대전 때 미국 로스앨러모스에서 원자폭탄을 만든 맨해튼 프로젝트에 참여한 몇 안 되는 여성 물리학자였습니다. 조안 힌턴은 전쟁 뒤 원자폭탄 사용에 반대하며 물리학을 떠나 중국으로 건너갔습니다.

아버지 하워드 에버레스트 힌턴(H. E. Hinton)은 브리스틀 대학교의 곤충학 교수였고 영국 왕립학회 회원이었습니다. 그는 여러 곤충을 연구했

고, 물속에 사는 작은 딱정벌레 무리인 엘미대(Elmidae)에 깊은 관심을 두었습니다. 힌턴은 노벨상 관련 대담에서 아버지 연구실의 모습을 떠올린 적이 있습니다. 벽 한 면의 높은 선반에는 곤충학 논문을 담은 상자 수백 개가 빼곡했습니다. 그 가운데 문 옆 낮은 자리에 있던 작은 상자 하나에만 "곤충 아님(Not Insects)"이라는 이름표가 붙어 있었고, 그 안에 자녀들의 편지와 집안 편지가 들어 있었습니다. 학문이 삶의 중심인 집이었습니다. 힌턴은 아주 어릴 때부터 학자로 성공하지 못하면 실패자가 된다는 것을 알았고, 학업 성취에 대한 압박이 컸다고 말했습니다.

이 집안 내력을 인공지능의 역사와 나란히 놓고 보면 눈에 띄는 점이 하나 있습니다. 조지 불이 1854년에 낸 책의 제목은 『사고의 법칙에 관한 연구: 논리와 확률의 수학 이론의 기초』입니다. 제목에 논리와 확률이 나란히 들어 있습니다. 그런데 20세기 중반 이후 인공지능 연구를 이끈 기호주의 연구자들은 불의 유산 가운데 논리 쪽만 이어받았습니다. 그들은 인간의 생각을 참과 거짓으로 나누는 기호와 그 기호를 다루는 규칙으로 설명하려 했습니다. 그러나 현실 세계는 흐릿하고 불확실한 일로 가득하며, 이런 세계를 다루려면 확률이 필요합니다.

불의 고손자 힌턴은 바로 그 지점에서 주류와 다른 길을 걸었습니다. 그는 뇌가 정해진 논리 규칙대로 움직이는 정밀 기계가 아니라고 보았습니다. 수많은 신경세포(뉴런)가 서로 얼마나 강하게 이어져 있는지, 곧 연결 강도를 경험에 따라 조금씩 바꾸어 가며 배우는 네트워크라고 생각했습니다. 고조부의 논리학이 기호주의 인공지능의 뼈대가 되었다면, 고손자는 고조부 책 제목의 나머지 절반인 확률에서 출발해 오늘날 딥러닝의 문을 열었습니다.

1.2. 일곱 살의 깨달음과 길을 찾아 헤맨 학창 시절

힌턴이 어린 시절을 돌아볼 때 자주 꺼내는 일화 가운데 하나는 버스 안에서 겪은 일입니다. 그는 털이 거친 버스 좌석 위에 옛 영국 페니 동전 한 닢을 올려놓았습니다. 버스가 흔들릴 때마다 동전은 아래로 미끄러지지 않고 좌석의 털을 타고 위쪽으로 올라갔습니다. 어린 힌턴은 있을 수 없는 일이라고 느꼈고, 그 이유를 몹시 알고 싶어 했습니다. 이해되지 않는 현상을 그냥 넘기지 않는 태도는 이때부터 그의 몸에 배어 있었습니다.

일곱 살 때 힌턴은 브리스틀의 클리프턴 칼리지(Clifton College)에 들어갔습니다. 기독교 색채가 짙은 학교였습니다. 학교는 신앙의 가르침을 당연한 진리로 전했고, 대부분의 친구들은 그 말을 그대로 받아들였습니다. 힌턴은 그러지 않았습니다. 증거 없이 믿으라는 말을 받아들일 수 없었습니다. 그는 훗날 서울경제와의 인터뷰에서 이때 처음으로 사람은 누구나 틀릴 수 있다는 생각을 하게 되었다고 회고했습니다. 이 경험이 다른 연구자들도 틀릴 수 있다는 믿음으로 이어졌고, 주류 학계에 맞서 연구를 계속하는 바탕이 되었다는 것이 그의 설명입니다. 힌턴이 반세기 가까이 학계의 냉대를 견딘 힘은 이 경험에서 시작되었습니다.

뇌에 대한 관심은 고등학교 때 생겼습니다. 열여섯, 열일곱 살 무렵, 힌턴이 늘 자신보다 똑똑하다고 여긴 친구 인먼 하비(Inman Harvey)가 어느 날 뇌의 기억이 홀로그램처럼 뇌 전체에 퍼져 저장될지 모른다는 이야기를 꺼냈습니다. 홀로그램은 조각 하나만 있어도 전체 그림을 흐릿하게나마 다시 만들 수 있습니다. 기억이 한 장소에 하나씩 들어 있는 것이 아니

라 여러 곳에 나뉘어 겹쳐 저장된다는 생각은 힌턴에게 큰 충격이었습니다. 뇌가 어떻게 계산하는지 밝히고 싶다는 바람이 이때 자리 잡았습니다.

그러나 그 바람에 답해 줄 학문을 찾는 일은 쉽지 않았습니다. 1967년 힌턴은 케임브리지 대학교 킹스 칼리지에 들어갔고, 이후 몇 년은 그 자신의 표현대로 "미친 무작위 행보(crazy random walk)"였습니다. 처음에는 자연과학을 택해 물리학, 화학, X선 결정학을 배웠습니다. 하루 12시간에 이르는 빡빡한 일정에 지친 그는 한 달 만에 학교를 떠났습니다. 런던에서 잡일로 방세를 벌며 도스토옙스키를 읽었고, 건축가가 되려고도 했습니다. 그러나 건축 사무소에서 하루 일해 보고, 실제 업무가 디자인보다 예산에 맞춰 썬 바닥재와 문손잡이를 고르는 일에 가깝다는 것을 알고 곧바로 그 만두었습니다.

대학에 돌아온 뒤에는 뇌와 마음을 다룰 법한 학과를 차례로 거쳤습니다. 먼저 뇌를 배울 수 있으리라 기대하며 생리학을 택했습니다. 수업은 오징어의 굵은 신경섬유에서 전기 신호가 어떻게 전달되는지를 가르쳤지만, 뉴런이 정보를 어떻게 나타내고 어떻게 배우는지는 다루지 않았습니다. 다음으로 마음의 본질을 찾으려 철학으로 옮겼습니다. 실험 없이 말로만 다루는 방식에 지친 그는 1년 뒤 "철학에 대한 항체"만 생겼다며 떠났습니다. 마지막으로 실험심리학을 공부해 1970년 학사 학위를 받았습니다.

심리학도 그가 찾던 답을 주지 못했습니다. 당시 심리학은 상자 속 쥐의 행동을 관찰하는 연구에 치우쳐 있었습니다. 힌턴은 아이의 인지 발달을 알아보는 실험에서 노란 세모, 노란 동그라미, 빨간 동그라미 가운데 성격이 다른 하나를 고르게 했습니다. 한 다섯 살 아이가 빨간 동그라미를 가리키

며 연구자가 이것을 틀린 색으로 칠했다고 말했습니다. 아이의 생각은 실험 설계보다 훨씬 복잡했습니다. 힌턴은 당시 심리학으로는 인간의 인지를 설명할 수 없다고 느꼈습니다.

졸업 뒤에도 방황은 이어졌습니다. 그는 약 1년 동안 목수 일을 배웠습니다. 현장에서 만난 숙련된 목수의 솜씨를 보고 이 길에서는 그만큼 해낼 수 없다고 판단해 학계로 돌아왔습니다. 이후 브리스톨에서 아동 언어 발달 연구에 보조원으로 참여했습니다. 노엄 촘스키(Noam Chomsky)의 언어 내재성 이론, 곧 사람이 태어날 때부터 문법 능력을 지니고 있다는 이론을 검증하는 연구였습니다. 연구진은 3세에서 5세 아이들의 집에 마이크를 두고 16만 6천 건에 이르는 발화 자료를 모았습니다. 그런데 설문을 위해 브리스톨 외곽의 저소득층 가정을 찾았을 때, 아이가 말을 하면 어떻게 하느냐는 첫 질문에 한 어머니가 때린다고 대답했습니다. 힌턴은 이 일을 겪고 이 연구에서 손을 뗐습니다.

여러 분야를 돌아다닌 끝에 그가 얻은 결론은 하나였습니다. 지능의 비밀을 풀려면 뇌의 계산 모형을 컴퓨터 위에 직접 만들어 보아야 한다는 것이었습니다.

1.3. 에든버러에서의 고립과 지도교수 롱게이히긴스와의 갈등

1972년 힌턴은 인공지능을 본격적으로 연구하려고 에든버러 대학교 박사과정에 들어갔습니다. 에든버러는 당시 영국에서 인공지능 연구 기반을 제대로 갖춘 거의 유일한 대학이었습니다. 그러나 그곳에서 그를 기다린 것은 고립과 지도교수와의 갈등이었습니다.

지도교수는 크리스토퍼 롱게이히긴스(Christopher Longuet-Higgins)였습니다. 그는 30대에 케임브리지 대학교 이론화학 교수가 된 뛰어난 과학자였고, 40대에 인지과학으로 방향을 바꾸어 홀로그램처럼 기억을 저장하는 모형과 연상 기억을 연구했습니다. 힌턴이 관심을 둔 바로 그 주제였습니다. 그러나 힌턴이 도착했을 무렵 롱게이히긴스의 생각은 이미 바뀌어 있었습니다. 그는 테리 위노그라드(Terry Winograd)의 SHRDLU 연구에 깊은 인상을 받았습니다. SHRDLU는 가상의 블록 세계에서 사람이 영어로 내린 명령을 기호 규칙에 따라 해석해 블록을 옮기는 프로그램으로, 기호주의 인공지능의 대표적 성과로 꼽혔습니다. 롱게이히긴스는 신경망에 대한 미련을 접고 기호주의 쪽으로 돌아섰습니다.

학계 분위기도 신경망에 불리했습니다. 1969년 마빈 민스키(Marvin Minsky)와 시모어 페퍼트(Seymour Papert)는 책 『퍼셉트론(Perceptrons)』에서 층이 하나뿐인 신경망은 XOR처럼 간단해 보이는 문제도 풀 수 없다는 것을 수학적으로 보였습니다. 이후 많은 연구자가 신경망 전체를 가망 없는 분야로 여겼습니다. 깔끔한 논리 체계를 좋아한 롱게이히긴스는 확률과 연결 강도 조정이라는 흐릿한 방식으로 움직이는 신

경망을 받아들이지 않았습니다. 그는 힌턴의 연구를 터무니없는 일이라고 말하며 기호주의로 바꾸라고 거듭 요구했습니다.

힌턴은 물러서지 않았습니다. 그는 6개월만 더 시간을 주면 성과를 보여 주겠고, 안 되면 주제를 바꾸겠다고 약속했습니다. 6개월이 지나면 같은 약속을 다시 했고, 이 일이 몇 차례 되풀이되었습니다. 힌턴은 2024년 에든버러 대학교 구술사 인터뷰에서 자신이 쫓겨나지 않은 이유를 털어놓았습니다. 아버지가 왕립학회 회원이었고, 같은 회원인 롱게이히킨스가 동료의 아들을 내보내기 부담스러웠을 것이라는 이야기였습니다. 그는 이것을 자신이 누린 특권 덕분이었다고 스스로 꼬집었습니다.

힌턴이 처음 구상한 박사 연구의 이름은 LUDWIG였습니다. "타고난 문법 없이 언어를 이해하는 장치(Language Understanding Device Without Innate Grammar)"의 머리글자로, 철학자 비트겐슈타인의 이름을 빌리면서 촘스키의 내재성 이론에 반대하는 뜻을 담았습니다. 기호 규칙을 미리 넣지 않고 경험만으로 언어를 배우는 신경망을 만들겠다는 계획이었습니다. 그러나 당시에는 입력층과 출력층 사이에 숨어 있는 은닉층을 학습시키는 방법이 없었습니다. LUDWIG는 그 시대의 기술로는 만들 수 없었습니다.

이때 길을 알려 준 사람이 박사후연구원 스티브 아이저드(Steve Isard)였습니다. 그는 사실상 부지도교수 역할을 하며, 손글씨 인식처럼 작고 쉬운 문제부터 신경망의 가능성을 확인해 보라고 권했습니다. 힌턴은 프랭크 로젠블랫(Frank Rosenblatt)의 1962년 책 『신경역학의 원리(Principles of Neurodynamics)』를 혼자 공부하며 병렬 네트워크를

구현했습니다. 롱게이히킨스가 서식스 대학교로 옮길 때 힌턴은 아이저드, 마크 스티드먼(Mark Steedman)과 함께 따라갔습니다. 힌턴은 서식스 대학교 연구원으로 지내다가 1978년 에든버러 대학교에서 「이완과 시각에서의 역할(Relaxation and Its Role in Vision)」이라는 논문으로 박사 학위를 받았습니다.

1.4. 기호주의의 한계와 조용히 시작된 연결주의의 반격

힌턴과 롱게이히긴스의 갈등은 두 사람만의 문제가 아니었습니다. 20세기 후반 인공지능 학계에는 지능을 바라보는 두 관점이 맞서 있었습니다. 하나는 기호주의(Symbolic AI)입니다. 앨런 뉴얼(Allen Newell), 허버트 사이먼(Herbert Simon), 존 매카시(John McCarthy), 마빈 민스키가 이끌었습니다. 이들은 인간의 생각이 기호와 그 기호를 다루는 논리 규칙, 그리고 한 단계씩 차례로 나아가는 추론으로 이루어진다고 보았습니다. 사람이 규칙을 잘 정리해 컴퓨터에 넣으면 지능을 만들 수 있다는 생각이었습니다. 이런 방식은 훗날 "좋았던 옛날식 인공지능(GOFAI)"이라고도 불렸습니다.

다른 하나는 연결주의(Connectionism)입니다. 힌턴과 데이비드 루멜하트(David Rumelhart), 테리 세이노프스키(Terrence Sejnowski), 그리고 뒤에 합류한 얀 르쿤(Yann LeCun), 요슈아 벤지오(Yoshua Bengio) 등이 이 길을 닦았습니다. 연결주의자들은 지능의 핵심이 규칙을 입력하는 데 있지 않고, 데이터에서 스스로 표현을 배우는 능력에 있다고 보았습니다. 뇌처럼 많은 단위가 동시에 일하고, 그 단위들 사이의 연결 강도(가중치)를 조정해 지식을 쌓는다는 것입니다. 이 관점에서 하나의 개념은 한 칸에 저장되지 않고 여러 단위의 활성화 패턴으로 나뉘어 저장됩니다. 이를 분산 표현(Distributed Representation)이라 합니다. 고등학교 시절 힌턴이 들은 홀로그램 기억 이야기와 닮은 생각입니다.

1970년대에 들어 기호주의는 벽에 부딪혔습니다. 현실의 데이터는 규칙으로 다 정리할 수 없을 만큼 모호하고 제각각이었습니다. 1973년 영국의

응용수학자 제임스 라이트힐(Sir James Lighthill) 경은 인공지능 연구를 평가한 보고서를 냈습니다. 이 라이트힐 보고서는 인공지능 연구가 조합 폭발 문제를 풀지 못한다고 지적했습니다. 조합 폭발이란 따져 보아야 할 경우의 수가 문제 크기에 따라 기하급수적으로 늘어나 계산이 불가능해지는 현상입니다. 보고서 이후 영국의 인공지능 연구비는 크게 줄었고, 첫 번째 "AI 겨울"이 찾아왔습니다.

많은 인공지능 연구자가 라이트힐 보고서를 비판했지만 힌턴의 평가는 달랐습니다. 그는 라이트힐의 지적이 핵심을 짚었다고 보았습니다. 매카시 같은 기호주의자들은 기호로 계산할 수 있는 것은 무엇이든 인공지능으로 만들 수 있다고 주장했습니다. 그러나 이 주장은 계산에 드는 시간과 자원, 곧 계산 복잡도를 고려하지 않았습니다. 라이트힐은 롱게이히긴스, 물리학자 프리먼 다이슨(Freeman Dyson)과 윈체스터 칼리지(Winchester College) 동문이었고, 유체역학을 연구한 응용수학자로서 이 약점을 정확히 보았습니다. 힌턴도 같은 생각이었습니다. 한정된 계산 자원으로 현실의 복잡함을 다루려면 기호를 하나씩 조작하는 방식으로는 부족하고, 뇌처럼 여러 단위가 동시에 일하며 연결 강도를 조금씩 고쳐 가는 방식이 필요하다고 보았습니다.

학계의 외면 속에서 힌턴은 뜻을 함께할 사람들을 찾았습니다. 1979년 그는 제임스 앤더슨(James Anderson)과 함께 미국 샌디에이고 라호야(La Jolla)에서 "병렬 연상 기억 모형(Parallel Models of Associative Memory)" 워크숍을 열었습니다. 이 자리에 샌디에이고 캘리포니아 대학교(UCSD)의 데이비드 루멜하트를 비롯해 여러 기관의 연결주의 연구자들이 모였습니다. 이 모임은 훗날 병렬 분산 처리(PDP) 연구 그룹이 만들

어지는 디딤돌이 되었습니다. 겉으로 드러나지 않았지만 연결주의의 반격은 이때부터 조용히 시작되었습니다.

1.5. 단기 성과보다 호기심: 기초연구가 맺은 결실과 그 의미

1978년 박사 학위를 받은 뒤부터 2012년 제자들과 함께 만든 AlexNet이 이미지 인식 대회에서 압도적으로 이기기까지, 신경망은 오랫동안 연구비가 잘 모이지 않는 변방 분야였습니다. 이 기간에 인공지능 분야는 두 차례 겨울을 겪었습니다. 힌턴이 이 긴 시간을 버틴 힘은 빠른 상업적 성과나 명예가 아니었습니다. 인간의 뇌가 어떻게 작동하는지 알고 싶다는 궁금증이었습니 다. 노벨상 측과의 인터뷰에서 자신을 소개해 달라는 요청을 받았을 때 그는 자신이 어느 분야에 속하는지 잘 모르지만 뇌가 어떻게 작동하는지 이해하고 싶은 사람이라고 답했습니다.

힌턴은 이런 궁금증이 연구의 출발점이어야 한다는 점을 공개적으로 여러 번 강조했습니다. 2024년 10월 8일 그는 존 홉필드(John Hopfield)와 함께 노벨 물리학상 수상자로 발표되었습니다. 그날 토론토 대학교 기자회견에서 그는 호기심이 이끄는 기초연구를 지원하라고 말했습니다. 캐나다가 미국만큼 연구비가 많지는 않지만 돈을 현명하게 쓴다고 평가했고, 신경망의 발전은 응용 문제에 돈을 쏟아부어서가 아니라 과학자들이 호기심을 따라 무언가를 이해하려 애쓴 데서 나왔다고 설명했습니다. 서울경제 인터뷰에서도 그는 당장의 실용적 성과를 내라고 학자를 압박하기보다 연구자가 궁금한 것을 자유롭게 연구하게 두는 편이 훨씬 효과적이며, 딥러닝이 바로 그렇게 만들어졌다고 말했습니다.

이런 연구가 결실을 맺은 배경에는 캐나다 고등연구소(CIFAR)의 장기 지원이 있었습니다. 힌턴은 1987년 CIFAR의 지원을 받아 캐나다로 옮겨 토론토 대학교에 자리를 잡았습니다. 그는 미국의 연구비 제도가 교수들에

게 제안서를 쓰는 데 많은 시간을 쏟게 한다고 지적했습니다. 반면 캐나다에서는 6쪽 분량의 제안서로 3년에서 5년 동안 주제에 얽매이지 않는 안정적 지원을 받을 수 있었다고 설명했습니다. 이런 환경이 있었기에 힌턴은 유행을 좇지 않고 인기 없는 분야를 끈질기게 파고들 수 있었습니다.

힌턴이 걸은 길의 의미는 기술의 성공에 그치지 않습니다. 인간의 마음을 보는 관점도 바뀌었습니다. 기호주의는 인간을 기호를 정교하게 다루고 논리적으로 추론하는 "추론 기계(Reasoning Machine)"로 보았습니다. 힌턴의 연구는 인간이 감각 경험과 유사성을 바탕으로 세상을 패턴으로 묶어 이해하는 "유추 기계(Analogy Machine)"에 가깝다는 것을 보여주었습니다. 힌턴의 관점에서 논리적 추론은 신경망 위에 덧붙은 능력이고, 지능의 바탕은 수많은 뉴런이 경험에서 스스로 분산 표현을 얻어 가는 연결 강도 조절에 있습니다.

일곱 살 소년이 품은 생각, 곧 어른들도 틀릴 수 있다는 깨달음은 반세기 뒤 학계 전체가 틀렸을 수 있다는 믿음으로 자랐습니다. 고조부 조지 불의 논리학이 드리운 그늘에서 출발한 힌턴은 그 책 제목에 함께 적혀 있던 확률을 붙잡고 끝까지 걸었습니다.

제2장. 기억은 어떻게 연결되는가

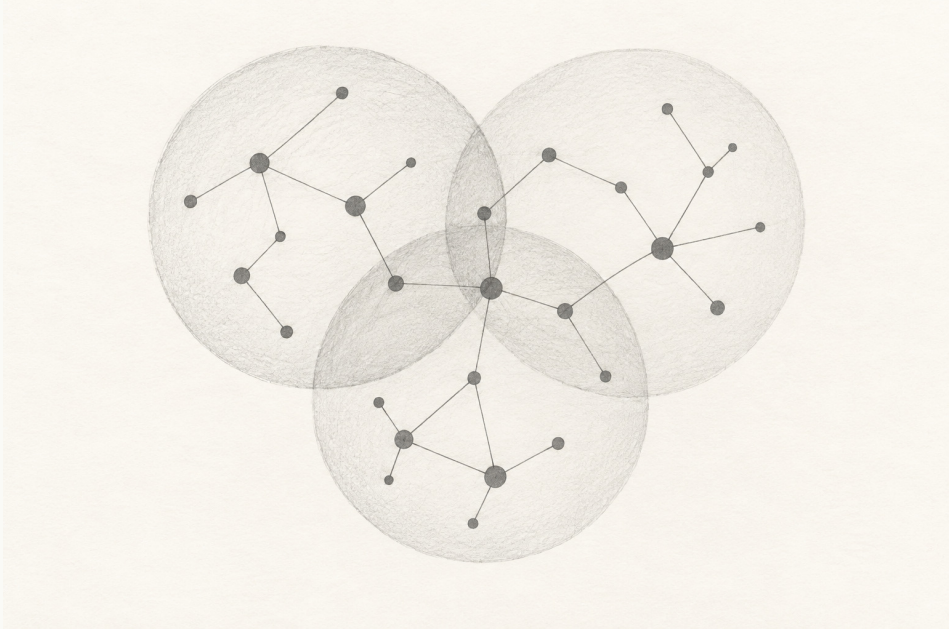


그림 2. 여러 연결에 나누어 정보를 나타내는 생각을 표현했습니다.

2.1. '할머니 뉴런'의 한계와 분산 표현의 오케스트라

제프리 힌턴은 고등학교 시절 친구에게서 홀로그램 이야기를 들었습니다. 홀로그램은 필름의 일부를 잘라 내도 전체 이미지가 흐릿하게나마 남습니다. 정보가 필름의 한 점에 기록되지 않고 필름 전체에 퍼져 있기 때문입니다. 힌턴은 뇌의 기억도 이와 비슷하게 퍼져 있을지 모른다고 생각했고, 이 생각을 평생의 연구 주제로 삼았습니다.

그러나 1970년대 인지과학과 인공지능 학계의 주류였던 기호주의는 지식의 저장 방식을 전혀 다르게 보았습니다. 기호주의 학자들은 사람의 생각이 기호와 그 기호를 다루는 규칙으로 이루어진다고 보았고, 하나의 개념은 하나의 자리에 저장된다고 생각했습니다. 이런 저장 방식을 '국소 표현'이라고 부릅니다.

국소 표현을 보여 주는 말이 '할머니 뉴런 가설'입니다. 뇌 속 어딘가에 할머니만 담당하는 뉴런이 하나 있고, 그 뉴런이 켜질 때 할머니를 알아본다는 가설입니다. 이 가설대로라면 고양이에는 고양이 뉴런이, 특정 문장에는 그 문장을 맡은 기호 노드가 하나씩 있어야 합니다. 하나의 대상을 하나의 단위에 대응시키면 구분하기는 쉽지만, 다양한 대상 사이의 공통점을 표현하기는 어렵습니다. 다만 기호로 사고를 설명하는 이론이 모두 뉴런 하나에 개념 하나를 대응시킨다는 뜻은 아닙니다.

힌턴은 이 생각에 일찍부터 의문을 품었습니다. 국소 표현에는 세 가지 약점이 있었습니다.

첫째는 조합의 문제입니다. 세상의 대상은 끝없이 새로운 조합으로 나타납니다. '빨간 사과', '초록 사과', '반쯤 먹은 빨간 사과'처럼 조합이 늘어날 때마다 전담 노드를 새로 만들어야 한다면, 필요한 노드 수는 기하급수적으로 늘어납니다. 이를 '조합적 폭발'이라고 합니다.

둘째는 손상에 약하다는 점입니다. 할머니를 맡은 뉴런 하나가 죽으면 할머니에 대한 지식 전체가 사라져야 합니다. 분산 표현에서는 일부 계산 단위가 손상되어도 남은 단위들에 정보가 유지될 수 있습니다.

셋째는 유사성을 담아내기 어렵다는 점입니다. 고양이와 호랑이는 닮았고 고양이와 냉장고는 닮지 않았습니다. 그런데 개념마다 따로 떨어진 노드가 하나씩 있으면, 두 개념이 얼마나 닮았는지를 노드 자체가 말해 주지 못합니다. 유사성에 관한 정보는 별도의 규칙으로 일일이 적어 넣어야 합니다.

힌턴이 내놓은 대안은 '분산 표현'입니다. 분산 표현은 하나의 개념을 뉴런 하나에 맡기지 않고, 수많은 뉴런이 함께 만드는 활성화 패턴으로 나타내는 방식입니다. 어떤 뉴런은 강하게 켜지고 어떤 뉴런은 약하게 켜지며 어떤 뉴런은 꺼져 있습니다. 그 전체 모양이 곧 하나의 개념입니다. 수학으로 말하면 개념은 여러 개의 숫자로 이루어진 벡터가 됩니다.

분산 표현은 오케스트라에 비유할 수 있습니다. 교향곡의 한 소절은 바이올린 하나가 혼자 내는 소리가 아닙니다. 바이올린, 첼로, 플루트, 트럼펫이 저마다 다른 세기로 함께 울리면서 하나의 화음을 만듭니다. 소절이 바뀌면 악기들의 조합과 세기가 함께 바뀌고, 같은 악기가 곡의 여러 부분에

쉬지 않고 참여합니다. 뇌의 지식도 이처럼 수많은 뉴런이 함께 내는 화음으로 존재한다는 것이 힌턴의 생각이었습니다.

오케스트라의 악기가 여러 소절에 참여하듯이, 분산 표현에서 뉴런 하나는 한 개념만 말지 않습니다. 뉴런은 '털이 있다', '네 발로 걷는다', '집 안에 산다' 같은 작은 특징을 나타내며 여러 개념의 표현에 동시에 참여합니다. 이런 작은 특징을 '마이크로특징'이라고 부릅니다. 고양이와 호랑이는 많은 마이크로특징을 공유하므로 두 개념의 활성화 패턴이 서로 비슷해집니다. 유사성이 별도의 규칙 없이 표현 안에 저절로 들어가는 것입니다. 손상에도 강합니다. 뉴런 몇 개가 손상되어도 나머지 뉴런이 만든 패턴이 대부분 남아 있으므로 지식이 한꺼번에 사라지지 않습니다.

힌턴은 이 생각을 실험으로도 보여 주었습니다. 1986년의 '가족 관계도 과제'입니다. 힌턴은 영국인 가족 12명과 이탈리아인 가족 12명, 모두 24명으로 이루어진 두 가족의 족보를 만들었습니다. 두 족보는 구조가 똑같았습니다. 영국 가족에는 크리스토퍼, 페넬로페, 샬럿, 빅토리아, 제임스, 콜린 같은 이름이 있었고, 이탈리아 가족에는 로베르토, 마리아 같은 이름이 같은 자리에 들어갔습니다. 관계는 아버지, 어머니, 남편, 아내, 삼촌, 조카 등 12가지였습니다.

신경망에는 '사람 1'과 '관계'를 입력하고 '사람 2'를 답하게 했습니다. 예를 들어 '샬럿'과 '어머니'를 입력하면 '빅토리아'를 답해야 합니다. 사람 이름은 24개 칸 가운데 한 칸만 켜는 방식으로, 관계는 12개 칸 가운데 한 칸만 켜는 방식으로 입력했습니다. 따라서 입력 자체에는 사람끼리 얼마나 닮았는지에 관한 정보가 전혀 없었습니다. 두 입력은 각각 6개의 뉴런

으로 이루어진 층을 지나 가운데 은닉층에서 만나고, 다시 몇 개의 층을 거쳐 답을 냈습니다.

신경망은 "A의 어머니가 B이고 B의 남편이 C이면 C는 A의 아버지다" 같은 규칙을 하나도 받지 않았습니다. 주어진 것은 사실을 나타내는 문장 104개뿐이었습니다. 힌턴은 그 가운데 100개로 신경망을 학습시키고, 나머지 4개는 시험용으로 남겨 두었습니다.

학습을 마친 뒤 힌턴은 사람을 나타내는 6개 뉴런을 들여다보았습니다. 어떤 뉴런은 영국 가족인지 이탈리아 가족인지를 구분하고 있었고, 어떤 뉴런은 몇 번째 세대인지를, 어떤 뉴런은 족보의 어느 가지에 속하는지를 나타내고 있었습니다. 누구도 가르쳐 주지 않은 특징을 신경망이 스스로 찾아낸 것입니다. 신경망은 한 번도 본 적 없는 시험 문제에도 대부분 맞는 답을 냈습니다.

힌턴은 오차 역전파 학습법의 진짜 가치가 바로 여기에 있다고 보았습니다. 계산 기법 자체보다, 은닉층이 데이터 속 구조를 반영하는 분산 표현을 스스로 배울 수 있다는 점이 중요했습니다. 오늘날 인공지능에서 단어를 숫자 벡터로 바꾸는 '단어 임베딩'의 뿌리가 이 실험에 있습니다.

2.2. 라호야에 모인 연구자들: 1979년 워크숍과 PDP 그룹의 탄생

1970년대 후반 신경망 연구자들은 외딴섬처럼 흩어져 있었습니다. 연구비와 교수 자리는 대부분 기호주의 연구에 돌아갔습니다. 신경망을 연구하는 사람들은 서로의 존재를 잘 알지 못했고, 논문도 주목받지 못하는 학술지에 흩어져 실렸습니다.

힌턴의 사정도 다르지 않았습니다. 그는 1978년 영국 에든버러 대학교에서 박사학위를 받았지만, 영국에서는 신경망 연구로 연구비를 얻기가 어려웠습니다. 힌턴은 박사학위 취득 전부터 서식스 대학교에서 연구했고, 1978년 10월 미국 샌디에이고 캘리포니아 대학교(UCSD)로 건너갔습니다.

1979년 힌턴은 브라운 대학교의 심리학자 제임스 앤더슨과 함께 UCSD가 있는 라호야에서 '병렬 연상 기억 모델' 워크숍을 열었습니다. 이 워크숍은 흩어져 있던 신경망 연구자들이 처음으로 한곳에 모인 계기가 되었습니다.

참석자 가운데 테리 세이노프스키가 있었습니다. 그는 당시 하버드 의과대학의 박사후 연구원이었고, 잘 알려지지 않은 학술지에 신경망 논문 몇편을 낸 젊은 연구자였습니다. 힌턴은 MIT의 시각 이론가 데이비드 마에게 세이노프스키의 연구에 대해 물어본 뒤 그를 초청했다고 합니다. 이 워크숍에서 만난 두 사람은 이후 볼츠만 머신을 함께 만드는 평생의 동료가 됩니다.

워크숍의 성과는 1981년 같은 제목의 책 《Parallel Models of Associative Memory》로 묶여 나왔습니다. 엮은이는 힌턴과 앤더슨이었습니다. 힌턴은 이 책에 의미 네트워크를 병렬 하드웨어로 구현하는 방법을 다룬 장을 썼습니다. 기호주의가 그래프와 기호로 나타내던 의미 관계를 뉴런들의 활성화 패턴으로 나타낼 수 있다는 내용이었습니

다. 라호야에는 이 흐름을 이어 갈 사람들이 있었습니다. UCSD의 데이비드 루멜하트와 제임스 맥클레랜드입니다. 두 사람은 1980년대 초 '병렬 분산 처리(Parallel Distributed Processing)' 연구 그룹, 줄여서 PDP 그룹을 꾸렸습니다. 힌턴은 이 무렵 UCSD에서 박사후 연구원으로 지내며 루멜하트와 가까워졌고, 이후 영국 케임브리지와 미국 카네기멜런 대학교로 자리를 옮긴 뒤에도 PDP 그룹과 함께 연구했습니다.

PDP 그룹이 성장하는 데에는 주변 연구자들의 지원과 참여도 중요했습니다. UCSD 인지과학 분야를 이끌던 도널드 노먼은 학과 안에서 연구자들이 주류와 다른 생각을 자유롭게 펼칠 수 있도록 울타리가 되어 주었습니다. 근처 솔크 연구소에 있던 프랜시스 크릭도 모임에 자주 나왔습니다. 크릭은 DNA 이중나선 구조를 밝힌 생물학자로, 이 무렵에는 뇌 연구로 관심을 옮긴 상태였습니다. 그는 신경망 모델이 실제 뇌와 얼마나 닮았는지를 두고 연구자들과 토론했고, 훗날 PDP 책 2권에 대뇌 피질의 구조를 다룬 장을 함께 썼습니다.

PDP 그룹의 연구는 1986년 MIT 출판부에서 나온 두 권의 책 《Parallel Distributed Processing: Explorations in the Microstructure of Cognition》으로 정리되었습니다. 1권은 기초를, 2권은 심리학과 생물

학 응용을 다루었습니다. 이 책은 심리학, 뇌과학, 컴퓨터과학 분야에 큰 영향을 주었고, 학술서로는 드물게 5만 부 넘게 팔렸다고 전해집니다.

힌턴은 1권의 핵심 장 여러 곳에 이름을 올렸습니다. 3장 '분산 표현'은 힌턴, 맥클레랜드, 루멜하트가 함께 썼습니다. 7장 '볼츠만 머신의 학습과 재 학습'은 힌턴과 세이노프스키가 썼습니다. 8장 '오차 전파를 통한 내부 표현 학습'은 루멜하트, 힌턴, 로널드 윌리엄스가 썼습니다. 같은 해 이 세 사람은 역전파 학습법을 소개한 논문을 《네이처(Nature)》에 실었습니다. 라호야에서 시작된 작은 모임이 신경망 연구를 다시 일으키는 중심이 된 것입니다.

2.3. 병렬 연상 기억과 가중치 최적화의 원리

라호야 워크숍의 제목에 들어 있던 '연상 기억'은 분산 표현과 함께 연결주의가 내세운 또 하나의 중요한 생각입니다. 이 개념은 보통의 컴퓨터가 기억을 찾는 방식과 비교하면 이해하기 쉽습니다. 컴퓨터는 기억을 주소로 찾습니다. 몇 번 칸에 무엇이 들어 있는지 정확한 주소를 알아야 그 내용을 꺼낼 수 있고, 주소가 하나라도 틀리면 원하는 내용을 찾지 못합니다. 이를 '주소 참조 기억'이라고 합니다.

사람의 기억은 다르게 작동합니다. 노래 첫 소절 몇 음만 들어도 노래 전체가 떠오르고, 친구 얼굴의 반쪽만 보아도 누구인지 압니다. 내용의 일부만으로 전체를 찾아내는 것입니다. 이를 '내용 참조 기억'이라고 합니다. 연결주의 연구자들은 신경망이 이런 기억을 자연스럽게 해낼 수 있다고 보았습니다.

이 원리를 수학으로 분명하게 보여 준 사람이 물리학자 존 홉필드입니다. 홉필드는 1982년, 물리학에서 자석 같은 물질을 설명하는 데 쓰던 '에너지' 개념을 신경망에 가져왔습니다. 홉필드 네트워크에서는 뉴런들의 상태마다 에너지 값이 정해집니다. 저장된 기억은 에너지 지형에서 움푹 파인 골짜기에 해당합니다. 일부가 지워지거나 잡음이 섞인 입력이 들어오면, 뉴런들이 서로 영향을 주고받으며 에너지를 조금씩 낮춥니다. 공이 비탈을 굴러 골짜기 바닥에 멈추듯, 네트워크는 가장 가까운 골짜기에 도착하고 원래의 기억을 완성합니다. 수많은 뉴런이 동시에 움직이며 안정된 상태를 찾아가는 이 과정을 '병렬 이완'이라고 부릅니다.

홉필드 네트워크에는 약점이 있었습니다. 에너지 지형에 원하지 않는 작은 골짜기가 생겨 엉뚱한 기억으로 빠지기도 했고, 얇은 골짜기에 갇혀 더 좋은 답을 찾지 못하기도 했습니다. 힌턴과 세이노프스키는 이 문제를 풀기 위해 통계물리학의 확률 개념을 더했습니다. 뉴런이 정해진 규칙대로만 움직이지 않고 일정한 확률로 흔들리게 하면, 네트워크가 얇은 골짜기에서 빠져나와 더 깊은 골짜기를 찾을 수 있습니다. 이렇게 만든 모델이 '볼츠만 머신'입니다. 데이비드 애클리, 힌턴, 세이노프스키는 1985년 볼츠만 머신의 학습 방법을 담은 논문을 발표했습니다.

볼츠만 머신이 푸는 문제의 본질은 '병렬 제약 충족'입니다. 사람이 물체를 볼 때 뇌는 경계선, 깊이, 빛의 방향, 색깔 같은 수많은 조건을 한꺼번에 맞추어 가장 그럴듯한 해석을 찾습니다. 이 조건들은 뉴런과 뉴런 사이 연결의 세기, 곧 '가중치'에 담겨 있습니다. 어떤 두 뉴런이 함께 켜지는 것이 자연스러우면 두 뉴런 사이의 가중치가 크고, 함께 켜지는 것이 어색하면 가중치가 작거나 음수입니다.

그렇다면 이 가중치는 누가 정하는가 하는 문제가 남습니다. 연결주의의 답은 신경망이 데이터를 보며 스스로 정한다는 것입니다. 사람이 규칙을 코드로 적어 넣지 않습니다. 신경망은 많은 예를 보면서, 자기가 만들어 내는 패턴이 실제 세계의 패턴과 비슷해지도록 가중치를 조금씩 고쳐 나갑니다. 이것이 가중치 최적화이고, 이 과정이 곧 학습입니다.

힌턴은 가중치에 담긴 기억이 시간에 따라 어떻게 달라지는지에도 관심을 가졌습니다. 1987년 그는 데이비드 플라우트와 함께 '빠른 가중치'에 관한 연구를 발표했습니다. 두 사람은 연결 하나에 두 종류의 가중치를 두었

습니다. 느린 가중치는 오랜 시간에 걸쳐 천천히 바뀌며 세상에 대한 일반적인 지식을 저장합니다. 빠른 가중치는 새 정보가 들어오면 금방 바뀌고 금방 사라지며, 방금 본 것을 잠시 붙잡아 두는 역할을 합니다.

이 구조는 신경망 학습에서 실제로 나타나는 문제 하나를 다루는 데 쓰였습니다. 신경망이 새로운 것을 배우면 예전에 배운 것이 흐려지는 일이 자주 생깁니다. 두 사람의 실험에서 빠른 가중치는 새 학습으로 흐려진 옛 기억을 짧은 연습만으로 빠르게 되살리는 데 도움을 주었습니다. 논문 제목 '빠른 가중치로 오래된 기억의 흐림을 걷어 내기'가 이 내용을 그대로 말해 줍니다. 빠른 가중치라는 생각은 이후 오랫동안 잊혔다가, 2016년 힌턴이 동료들과 다시 꺼내 들며 주목을 받았습니다.

2.4. 마음을 보는 관점의 변화: 기호 조작자에서 데이터 학습자로

분산 표현과 연상 기억에 관한 연구는 기술의 문제에 그치지 않았습니다. 사람의 마음을 어떻게 볼 것인가 하는 오래된 질문에 다른 답을 내놓았습니다.

기호주의와 고전적 인지과학이 그린 인간은 '기호 조작자'였습니다. 사람은 머릿속 기호를 정해진 규칙에 따라 차례로 다루는 논리 기계라는 것입니다. 이들은 사람이 문법에 맞게 말하고 사물을 범주로 나누는 모습을 보며, 뇌 속에도 문법 규칙과 범주 목록이 미리 적혀 있을 것이라고 추론했습니다.

힌턴과 연결주의 연구자들이 그린 인간은 달랐습니다. 이들이 본 뇌는 규칙을 받아 적어 차례로 계산하는 기계가 아니라, 수많은 경험에서 유사성과 패턴을 스스로 뽑아내는 '데이터 학습자'였습니다. 힌턴은 사람이 논리 기계보다 '유추 기계'에 가깝다고 표현해 왔습니다. 새로운 상황을 만나면 규칙을 하나씩 적용하기보다, 예전에 겪은 비슷한 상황을 떠올려 판단한다는 뜻입니다.

이 관점에서 보면 문법에 맞는 말하기나 논리적인 추론도 뇌 속에 규칙이 적혀 있어서 생기는 일이 아닙니다. 수많은 뉴런 사이의 가중치 속에 만들어진 분산 표현이 서로 작용하면서, 겉으로 규칙처럼 보이는 행동이 나타납니다. 이를 '창발'이라고 합니다. 가족 관계도 실험의 신경망이 친족 규

칙을 받지 않고도 친족 관계를 맞혔듯이, 사람도 규칙을 외우지 않고 규칙처럼 행동할 수 있다는 것입니다.

힌턴은 최근 강연에서도 당시의 연구와 오늘날의 인공지능이 이어져 있다고 거듭 설명합니다. 2025년 6월 영국 브리스틀 대학교 강연에서 그는 1985년에 만든 작은 언어 모델을 소개했습니다. 이 모델은 다음 단어를 예측하면서 생기는 오류를 학습 신호로 삼아, 단어 기호를 특징 활성화로 바꾸고 그 특징들이 서로 작용해 다음 단어의 특징을 예측하도록 배웠다는 것입니다. 그는 오늘날의 대형 언어 모델을 이 작은 모델의 후손으로 볼 수 있다고 말했습니다. 그가 말한 작은 모델이 바로 가족 관계도를 배운 신경망입니다.

연상 기억 연구도 수십 년 뒤 다시 평가를 받았습니다. 2024년 노벨 물리학상은 홉필드와 힌턴에게 돌아갔습니다. 스웨덴 왕립과학원은 홉필드가 이미지 같은 데이터 패턴을 저장하고 다시 만들어 내는 연상 기억을 만들었고, 힌턴은 홉필드 네트워크를 바탕으로 볼츠만 머신이라는 새로운 네트워크를 만들었다고 수상 이유를 밝혔습니다. 1979년 라호야 워크숍의 제목에 들어 있던 '연상 기억'이 45년 뒤 노벨 물리학상의 수상 이유에 다시 등장한 것입니다.

1970년대와 1980년대에 힌턴은 라호야, 케임브리지, 피츠버그를 오가며 분산 표현과 연상 기억의 이론을 다졌습니다. 그때 그의 생각은 학계의 변두리에 있었습니다. 지식은 한 자리에 적힌 기호가 아니라 수많은 뉴런이 함께 내는 화음이라는 생각, 기억은 주소가 아니라 내용으로 찾는다는 생각, 규칙은 적어 넣는 것이 아니라 데이터에서 배운다는 생각이 그것입니

다. 이 세 가지 생각이 오늘날 언어와 개념을 다루는 인공지능의 바탕이 되었습니다.

제3장. 틀린 답에서 배우는 신경망

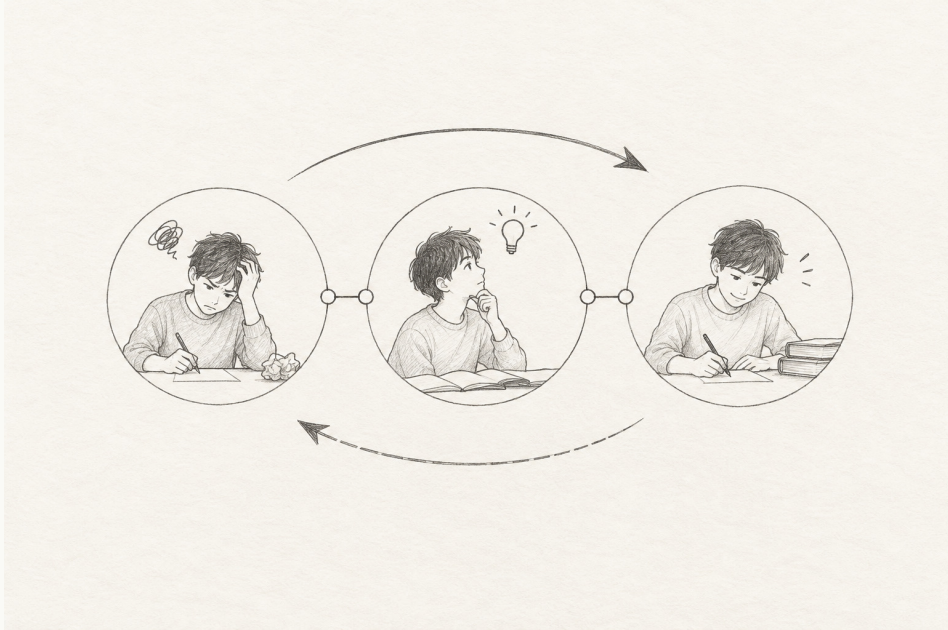


그림 3. 틀린 결과를 되짚어 학습을 개선하는 과정을 비유했습니다.

3.1. 다층 신경망의 난제와 역전파(Backpropagation)의 재발견

1969년 마빈 민스키와 시모어 페퍼트는 《Perceptrons》라는 책을 냈습니다. 두 사람은 이 책에서 입력층과 출력층만 있는 단층 퍼셉트론이 XOR 같은 문제를 풀 수 없다는 사실을 수학으로 보여 주었습니다. 책이 나온 뒤 많은 연구자가 신경망 연구를 떠나 기호주의 인공지능으로 옮겨갔습니다. 기호주의는 사람이 규칙을 직접 적어 넣고 컴퓨터가 그 규칙에 따라 추론하게 하는 방식입니다.

단층 퍼셉트론의 한계를 넘을 방법이 전혀 없었던 것은 아닙니다. 입력층과 출력층 사이에 은닉층(Hidden Layer)을 넣으면 더 복잡한 문제를 풀 수 있다는 점은 이론적으로 알려져 있었습니다. 문제는 그 은닉층을 어떻게 학습시키느냐였습니다. 출력층은 정답과 직접 비교할 수 있습니다. 신경망이 고양이 사진을 보고 '개'라고 답하면 얼마나 틀렸는지 바로 계산할 수 있습니다. 은닉층은 사정이 다릅니다. 은닉층의 뉴런은 바깥 세계와 직접 연결되어 있지 않아서, 어떤 순간에 어떤 값을 내야 옳은지 아무도 알려주지 않습니다. 최종 답이 틀렸을 때 그 책임이 수많은 중간 뉴런 가운데 어디에 얼마나 있는지 나누어 따지는 문제를 '신용 할당 문제(Credit Assignment Problem)'라고 부릅니다. 이 문제를 풀지 못하는 한 다층 신경망은 종이 위의 구상에 머물 수밖에 없었습니다.

신용 할당 문제의 답이 오차 역전파입니다. 원리는 미분의 연쇄 법칙(Chain Rule)에 있습니다. 먼저 출력층에서 오차를 구합니다. 그다음 이 오차를 거꾸로, 곧 출력층에서 은닉층 쪽으로 한 층씩 되돌려 보내면서 각

연결 가중치가 전체 오차에 얼마나 기여했는지 계산합니다. 마지막으로 오차를 줄이는 방향으로 가중치를 조금씩 고칩니다. 이 마지막 단계를 경사 하강법(Gradient Descent)이라고 합니다. 눈을 가린 채 산을 내려갈 때 발밑의 기울기만 확인하고 낮은 쪽으로 한 걸음씩 옮기는 것과 같습니다.

역전파는 한 사람이 한 번에 떠올린 발명품이 아닙니다. 여러 학자가 서로 모르는 상태에서 같은 생각에 이르렀습니다. 1970년 핀란드의 세포 린나인마(Seppo Linnainmaa)는 오늘날 '역모드 자동 미분'이라고 부르는 계산법을 석사 논문에서 정리했습니다. 1974년 하버드 대학교 박사과정생 폴 워보스(Paul Werbos)는 박사학위 논문 《Beyond Regression》에서 이 방법을 신경망 학습에 쓸 수 있다고 제안했습니다. 1985년 무렵에는 데이비드 파커(David Parker)와 프랑스의 얀 르쿤(Yann LeCun)이 각자 비슷한 학습 규칙을 내놓았습니다.

그런데도 역전파의 역사에서 전환점으로 꼽히는 것은 1986년 데이비드 루멜하트(David E. Rumelhart), 제프리 힌턴(Geoffrey E. Hinton), 로널드 윌리엄스(Ronald J. Williams)의 연구입니다. 세 사람의 협력은 샌디에이고에서 시작되었습니다. 힌턴은 1978년부터 1980년까지 캘리포니아 대학교 샌디에이고 캠퍼스(UCSD)에서 박사후연구원으로 일하고 이후에도 동료들과 교류하면서 루멜하트, 제임스 맥클레랜드(James McClelland) 등이 이끈 PDP(병렬 분산 처리) 연구 모임에 참여했습니다. UC 샌디에이고는 2024년 기사에서 PDP 그룹의 작업이 역전파 발전의 토대가 되었고, 힌턴이 뒤에 루멜하트, 윌리엄스와 함께 1986년 《네이처》 논문으로 이 알고리즘을 널리 알렸다고 정리했습니다.

세 사람이 역전파를 처음 제안한 것은 아닙니다. 세 사람의 공로는 수식을 처음 쓴 데 있지 않습니다. 역전파로 학습한 신경망의 은닉층이 사람이 가르쳐 주지 않은 내부 표현(Internal Representation)을 스스로 만들어 낸다는 사실을 실험으로 보여 준 데 있습니다.

힌턴 자신도 이 점을 분명히 했습니다. 2018년 마틴 포드(Martin Ford)와의 인터뷰에서 그는 역전파의 기본 아이디어는 루멜하트가 떠올린 것이며 루멜하트의 발명이라고 말했습니다. 자신과 윌리엄스는 그 방법을 다듬는 데 힘을 보탰고, 자신의 몫은 역전파가 의미 있는 분산 표현을 학습할 수 있음을 보여 준 데 있다고 설명했습니다. 언론이 그를 역전파의 발명자로 소개할 때마다 힌턴은 이렇게 바로잡았습니다.

3.2. 1986년 《네이처》 논문과 계산 원리

1986년 10월 9일 국제 학술지 《네이처》 323권 533쪽부터 536쪽에 네 쪽짜리 논문이 실렸습니다. 제목은 〈Learning representations by back-propagating errors〉 이고, 우리말로 옮기면 '오차를 역전파하여 표현을 학습하기'입니다. 저자 소속은 루멜하트와 윌리엄스가 UCSD 인지과학연구소, 힌턴이 카네기 멜런 대학교(CMU)였습니다. 같은 해 세 사람은 PDP 그룹이 펴낸 두 권짜리 책 《Parallel Distributed Processing》 1권에 〈Learning internal representations by error propagation〉 이라는 긴 장을 실어 같은 내용을 자세히 설명했습니다. 짧은 논문은 학계의 관심을 끌었고, 긴 책은 연구자들이 직접 따라 할 수 있는 교과서 역할을 했습니다.

논문의 계산 과정은 네 단계로 나누어 볼 수 있습니다. 앞의 세 단계는 신경망이 답을 내고 그 답이 정답에서 얼마나 벗어났는지 재는 과정이고, 마지막 단계는 그렇게 낸 오차를 뉴런 하나하나의 책임으로 나누는 과정입니다.

첫째는 순방향 전달(Forward Pass)입니다. 신호는 아래층에서 위층으로 올라갑니다. 위층의 뉴런 하나는 아래층의 여러 뉴런과 연결되어 있고, 연결마다 그 세기를 나타내는 가중치가 있습니다. 아래층 뉴런 각각의 출력에 해당 연결의 가중치를 곱하고 그 값을 모두 더하면 위층 뉴런이 받는 총입력이 됩니다. 뉴런이 언제 켜질지 정하는 문턱값은 따로 다루지 않습니다. 늘 1을 내보내는 가상의 입력을 하나 두고 그 입력에 붙은 가중치가 문턱값 구실을 하게 하여, 다른 가중치와 같은 방식으로 계산했습니다.

둘째는 활성화 함수입니다. 활성화 함수는 뉴런이 받은 총입력을 그 뉴런의 출력으로 바꾸는 규칙입니다. 이 논문에서는 로지스틱 시그모이드 함수를 썼습니다. 이 함수는 총입력이 큰 음수이면 0에 가까운 값을, 큰 양수이면 1에 가까운 값을 내고, 그 사이에서는 완만한 에스(S)자 곡선을 그리며 0과 1 사이의 값을 냅니다. 곡선이 끊어지거나 꺾이지 않고 부드럽게 이어지므로 어디서나 미분할 수 있습니다. 미분은 입력이 조금 변할 때 출력이 얼마나 변하는지를 구하는 계산입니다. 시그모이드 함수의 미분값은 출력값에, 1에서 출력값을 뺀 수를 곱한 것으로 간단하게 나옵니다. 커짐과 꺼짐 두 가지 상태만 있던 초기 퍼셉트론의 계단 함수는 값이 한 지점에서 갑자기 뛰기 때문에 이런 미분을 쓸 수 없었습니다. 역전파는 연쇄 법칙, 곧 여러 단계를 거친 변화율을 단계별 변화율의 곱으로 구하는 규칙에 따라 오차를 되돌려 보냅니다. 이 방법이 작동하려면 이렇게 매끄러운 함수가 필요했습니다.

셋째는 오차 측정입니다. 신경망이 낸 답 $y(j)$ 와 정답 $d(j)$ 의 차이를 제공해 모두 더한 뒤 절반을 취합니다. 이 값 E 가 신경망이 얼마나 틀렸는지 알려주는 점수입니다.

넷째가 역방향 전달(Backward Pass)입니다. 여기서 뉴런마다 오차 신호를 구합니다. 오차 신호는 그 뉴런이 받은 총입력이 조금 달라질 때 전체 오차가 얼마나 달라지는지를 나타내는 값으로, 그 뉴런이 오차에 진 책임의 크기에 해당합니다. 출력층 뉴런의 오차 신호는 실제 출력에서 정답으로 주어진 목표값을 뺀 차이에 시그모이드 함수의 미분값, 곧 출력값과 1에서 출력값을 뺀 수의 곱을 곱해 구합니다. 은닉층 뉴런의 오차 신호는 이 뉴런의 출력을 받아 간 위층 뉴런들의 오차 신호에 각 연결 가중치를 곱

해 모두 더하고, 여기에 이 뉴런 자신의 출력값과 1에서 그 출력값을 뺀 수를 곱해 얻습니다. 위층의 책임을 연결의 세기에 비례해 아래층에 나누어 주는 계산입니다. 이렇게 하면 은닉층이 아무리 많아도 출력층에서 시작해 한 층씩 내려가며 모든 뉴런의 책임을 계산할 수 있습니다. 정답과 직접 비교할 수 없던 은닉층에도 오차 신호가 주어지는 것입니다. 마지막으로 각 가중치는 오차가 줄어드는 쪽으로 고치며, 한 번에 고치는 폭은 학습률이라는 작은 수로 정합니다.

연구진은 실제 학습에서 생기는 문제도 다루었습니다. 하나는 학습 속도였습니다. 경사하강법은 오차가 줄어드는 방향으로 가중치를 조금씩 옮기는 방법입니다. 가중치에 따라 오차가 달라지는 모습을 지형에 견주면, 이 지형이 좁고 긴 골짜기 모양일 때 경사하강법은 골짜기 양쪽 벽을 오가며 흔들리고 골짜기를 따라 나아가는 속도는 느려집니다. 연구진은 이 문제를 줄이기 위해 관성 항(Momentum)을 넣었습니다. 원래의 변화량은 오차의 기울기에 학습률을 곱해 기울기 반대 방향으로 옮기는 양인데, 여기에 바로 앞 단계에서 가중치를 바꾼 양의 일부를 더하는 것입니다. 앞 단계의 변화량을 얼마나 반영할지는 0과 1 사이의 수인 관성 계수로 정합니다. 굴러가던 공이 작은 턱에 걸려도 그동안 붙은 속도로 넘어가는 것과 같은 효과를 냅니다. 이 기법은 오래 쓰였습니다. 2013년에는 일리야 수츠케버(Ilya Sutskever) 등이 깊은 신경망 학습에서 관성 항의 효과를 다시 검토하기도 했습니다.

다른 하나는 국소 극솟점(Local Minima)에 대한 우려였습니다. 국소 극솟점은 주변보다는 오차가 낮지만 전체에서 오차가 가장 낮은 곳은 아닌 지점입니다. 경사하강법은 주변보다 낮은 곳에 이르면 더 내려갈 방향을

찾지 못하므로, 가장 낮은 골짜기가 아닌 얇은 웅덩이에서 학습이 멈출 수 있다는 지적이었습니다. 논문은 많은 과제에서 이런 일이 드물었다고 보고했습니다. 그리고 은닉 유닛과 연결을 몇 개 더 넣으면 가중치 하나하나를 좌표축으로 삼은 가중치 공간에 새 차원이 생기고, 그 차원이 웅덩이를 돌아 나가는 길이 된다고 설명했습니다.

3.3. 은닉층이 스스로 배운 표현: 거울 대칭 판별과 '가족 관계도(Family Trees)' 과제

1986년 논문이 학계에 준 충격은 수식보다 실험에서 나왔습니다. 연구진은 학습을 마친 신경망의 은닉층을 열어 각 뉴런이 무엇을 학습했는지 그림으로 보여 주었습니다.

첫 실험은 거울 대칭(Mirror Symmetry) 판별입니다. 0과 1로 된 여섯 자리 배열이 가운데를 기준으로 좌우가 똑같은지 가리는 문제입니다. 예를 들어 101101은 대칭이고 101100은 대칭이 아닙니다. 이 문제는 비트 하나만 보아서는 아무것도 알 수 없습니다. 첫째 자리와 여섯째 자리, 둘째 자리와 다섯째 자리처럼 짝을 맞추어 비교해야 합니다. 1이 홀수 개 있는지 가리는 패리티(Parity) 문제와 같은 부류로, 단층 퍼셉트론이 풀지 못하는 대표적인 과제였습니다. 패리티 문제 자체는 같은 해 PDP 책의 역전파장에서 따로 다루었습니다.

연구진은 은닉 유닛을 두 개만 둔 신경망에 가능한 64가지 배열을 모두 보여 주며 학습시켰습니다. 64개 배열을 1,425번 되풀이해 보여 주자 신경망은 깔끔한 해법을 스스로 찾았습니다. 각 은닉 유닛은 가운데를 사이에 둔 짝 자리에 크기는 같고 부호는 반대인 가중치를 붙였습니다. 바깥쪽부터 안쪽까지 가중치의 크기는 1 대 2 대 4의 비율을 이루었습니다.

이 가중치 구조가 대칭을 판별하는 과정은 다음과 같습니다. 배열이 완전히 대칭이면 좌우 값이 서로 상쇄되어 은닉 유닛의 총입력이 0이 됩니다. 은닉 유닛에는 음의 바이어스가 있어서 총입력이 0이면 꺼진 상태로 남고,

두 은닉 유닛이 모두 꺼지면 출력 유닛이 '대칭'이라고 답합니다. 대칭이 아니면 두 은닉 유닛 중 하나가 켜져 출력 유닛을 끕니다. 1, 2, 4라는 비율 덕분에 대칭이 아닌 배열이 우연히 합계 0을 만드는 일은 생기지 않습니다. 사람은 이 규칙을 하나도 가르치지 않았습니니다. 신경망은 정답과 오답만 보고 대칭이라는 기하학적 규칙을 가중치 구조로 옮겨 놓았습니다.

논문의 중심은 힌턴이 직접 설계한 가족 관계도 과제입니다. 힌턴은 영국인 가족 12명과 이탈리아인 가족 12명, 모두 24명의 가계도를 만들었습니다. 두 가계도는 구조가 똑같습니다. 영국 쪽의 크리스토퍼와 페넬로피 부부 자리에 이탈리아 쪽의 로베르토와 마리아 부부가 있는 식입니다. 관계는 아버지, 어머니, 남편, 아내, 아들, 딸, 삼촌, 이모·고모, 형제, 자매, 조카(남), 조카(여)의 12가지였습니다. 학습 자료는 '사람 1 + 관계 = 사람 2' 꼴의 명제 104개였습니다. 콜린의 이모·고모가 제니퍼와 마거릿이라는 식입니다.

신경망은 다섯 층으로 이루어졌습니다. 맨 아래에 사람 1을 나타내는 유닛 24개와 관계를 나타내는 유닛 12개가 있었습니다. 한 번에 하나만 켜지는 방식이어서, 입력만 보면 영국의 콜린과 이탈리아의 알폰소 사이에는 아무런 공통점이 없습니다. 그 위로 사람을 받는 은닉 유닛 6개와 관계를 받는 은닉 유닛 6개, 가운데 은닉 유닛 12개, 출력 쪽 은닉 유닛 6개, 사람 2를 나타내는 출력 유닛 24개가 차례로 이어졌습니다.

연구진은 104개 명제 가운데 100개로만 학습시키고 4개는 남겨 두었습니다. 학습이 끝난 뒤 사람을 받는 첫 은닉층의 유닛 6개를 살펴보니 각 유닛이 의미 있는 특징을 맡고 있었습니다. 1번 유닛은 영국인과 이탈리아

인을 구분했습니다. 2번 유닛은 그 사람이 몇 번째 세대인지를 나타냈습니다. 6번 유닛은 가계도의 어느 가지에 속하는지를 나타냈습니다. 국적을 맡은 유닛을 빼고 보면 콜린과 알폰소처럼 두 가계도에서 같은 자리에 있는 사람은 비슷한 활성화 패턴을 얻었습니다. 국적, 세대, 가지 같은 개념은 입력 어디에도 적혀 있지 않았습니다. 신경망이 관계를 맞히는 과정에서 스스로 찾아낸 특징입니다.

신경망은 "어머니의 남편은 아버지다" 같은 규칙도 받지 않았습니다. 그런데도 학습 때 한 번도 보지 못한 4개 명제의 정답을 맞혔습니다. 한 사람을 여러 특징의 조합으로 나타내는 분산 표현(Distributed Representation)이 있었기에 이런 일반화가 가능했습니다. 기호 하나에 뜻 하나를 대응시키는 기호주의 방식과는 정반대의 접근이었습니다.

이 실험은 신경망이 단어와 개념의 뜻을 숫자 벡터로 학습할 수 있음을 보인 초기 사례로 평가받습니다. 힌턴, 르쿤, 요슈아 벤지오가 2015년 《네이처》에 실은 딥러닝 리뷰 논문과 2021년 《Communications of the ACM》에 실은 글은 오늘날의 단어 임베딩(Word Embedding)이 이 생각의 연장선에 있다고 설명합니다. 뒤에 나온 Word2Vec, 트랜스포머, 대형 언어 모델도 단어를 벡터로 바꾸어 다룬다는 점에서 같은 계보에 놓입니다.

3.4. 카네기 멜런 대학교(CMU) 시절과 캐나다 토론토 이주

1986년 논문이 나올 때 힌턴의 소속은 미국 펜실베이니아주 피츠버그의 카네기 멜런 대학교였습니다. 그는 1982년 이 대학 컴퓨터과학과 교수가 되어 1987년까지 다섯 해 동안 일했습니다. 당시 CMU는 앨런 뉴얼과 허버트 사이먼이 이끄는 기호주의 인공지능의 중심지였습니다. 2019년 CMU의 튜링상 축하 기사에 따르면 힌턴은 기호주의를 대표하는 뉴얼과 함께 강의를 맡기도 했습니다. 관점이 다른 동료들 사이에서 그는 연결주의 연구를 이어 갔습니다.

CMU 시절은 연구 성과가 많은 시기였습니다. 힌턴은 데이비드 애클리(David Ackley), 테리 세이노프스키(Terrence Sejnowski)와 함께 통계물리학의 개념을 빌린 볼츠만 머신(Boltzmann Machine)을 연구해 1985년 《Cognitive Science》에 학습 알고리즘을 발표했습니다. 같은 시기에 루멜하트, 윌리엄스와 함께 역전파 논문을 완성했습니다. 코크 대학교의 조지 불 200주년 기록은 힌턴이 CMU 교수로 지낸 다섯 해 동안 역전파, 볼츠만 머신, 단어의 분산 표현 연구를 개척했다고 소개합니다.

그 무렵 미국의 정치 상황은 힌턴의 생각과 다른 방향으로 움직였습니다. 로널드 레이건 행정부는 소련과의 냉전 속에서 '스타워즈'라는 별명이 붙은 전략방위구상(SDI)을 추진했습니다. 국방부 산하 고등연구계획국(DARPA)도 인공지능을 포함한 컴퓨팅 연구에 큰 예산을 지원했습니다. 미국 대학의 인공지능 연구비 가운데 많은 부분이 군에서 나왔습니다.

힌턴은 자신을 사회주의자라고 여겼고, 인공지능을 전쟁에 쓰는 데 반대했습니다. 그는 군의 연구비를 받고 싶지 않았습니다. 브리태니커는 힌턴이 1987년 미국을 떠나 캐나다로 간 결정에 미군과 레이건 행정부에 대한 반감이 작용했고, 그가 인공지능의 전투 사용에 반대했다고 적고 있습니다. 레이건 정부 정책에 대한 실망과 인공지능 연구의 군사 자금 지원에 대한 반대가 이주의 이유 가운데 하나였다는 기록도 있습니다.

1987년 힌턴은 CMU 교수직을 내려놓고 캐나다 온타리오주의 토론토 대학교 컴퓨터과학과로 옮겼습니다. 그를 받아 준 곳은 캐나다 고등연구소(CIFAR)였습니다. CIFAR는 당장의 상업적 성과나 군사 목적과 상관없이 기초 연구를 장기간 지원하는 기관입니다. CIFAR에 따르면 힌턴은 1987년 CIFAR의 첫 연구 프로그램인 '인공지능, 로봇공학과 사회'의 펠로우로 합류했고, CIFAR와 캐나다 자연과학공학연구위원회(NSERC)의 지원을 받으며 당시 회의론과 연구비 부족에 시달리던 분야의 연구를 계속했습니다.

토론토 시절의 첫 10년은 순탄하지 않았습니다. 1990년대 들어 서포트 벡터 머신 같은 다른 기계학습 방법이 주목받으면서 신경망은 다시 관심 밖으로 밀려났습니다. 힌턴은 1998년부터 2001년까지 런던의 유니버시티 칼리지 런던에서 개츠비 계산신경과학 연구소를 세워 이끈 뒤 토론토로 돌아왔습니다. 2004년에는 동료들과 함께 CIFAR에 '신경 계산과 적응적 지각(NCAP)' 프로그램을 새로 열었습니다. 이 모임은 르쿤, 벤지오 등 신경망 연구자들이 꾸준히 만나는 자리가 되었고, 2012년 힌턴 연구실의 알렉스 크리츨스키와 일리아 수츠케버가 만든 AlexNet이 이미지 인식 대회에서 큰 차이로 우승하는 밑바탕이 되었습니다.

1986년의 협력자들은 그 뒤 서로 다른 길을 걸었습니다. 로널드 윌리엄스는 강화학습의 정책 경사 기법인 REINFORCE 알고리즘을 1992년에 발표했습니다. 2001년 힌턴은 루멜하트의 이름을 딴 루멜하트상의 첫 수상자가 되었고, 루멜하트는 2011년 세상을 떠났습니다. 2024년 힌턴은 존 홉필드와 함께 인공 신경망 기반 기계학습을 가능하게 한 기초적 발견과 발명의 공로로 노벨 물리학상을 받았습니다. 노벨상 발표가 있던 2024년 10월 8일, CIFAR는 1987년부터 이어진 힌턴과의 인연을 되짚는 축하 글을 냈습니다.

다층 신경망을 학습시킬 방법이 없다는 오랜 난제는 1986년 역전파 논문으로 해결의 길이 열렸습니다. 그 논문이 남긴 핵심은 계산 기법보다 생각의 전환에 있었습니다. 사람이 규칙을 넣어 주지 않아도 신경망은 자료 속에서 쓸모 있는 표현을 스스로 찾아낸다는 것입니다. 힌턴은 이 생각을 지닌 채 군사 연구비가 넘치던 미국을 떠나 캐나다로 갔고, 토론토에서 20년 넘게 연구를 이어 갔습니다. 오늘날의 대형 언어 모델과 이미지 생성 모델도 역전파로 학습합니다. 다만 힌턴은 뒤에 실제 뇌가 역전파와 같은 방식으로 학습하지는 않을 것이라고 보았고, 2022년 Forward-Forward 알고리즘 같은 대안을 내놓았습니다.

제2부. 신경망을 배우게 하는 방법

제4장. 물리학에서 찾은 학습의 원리

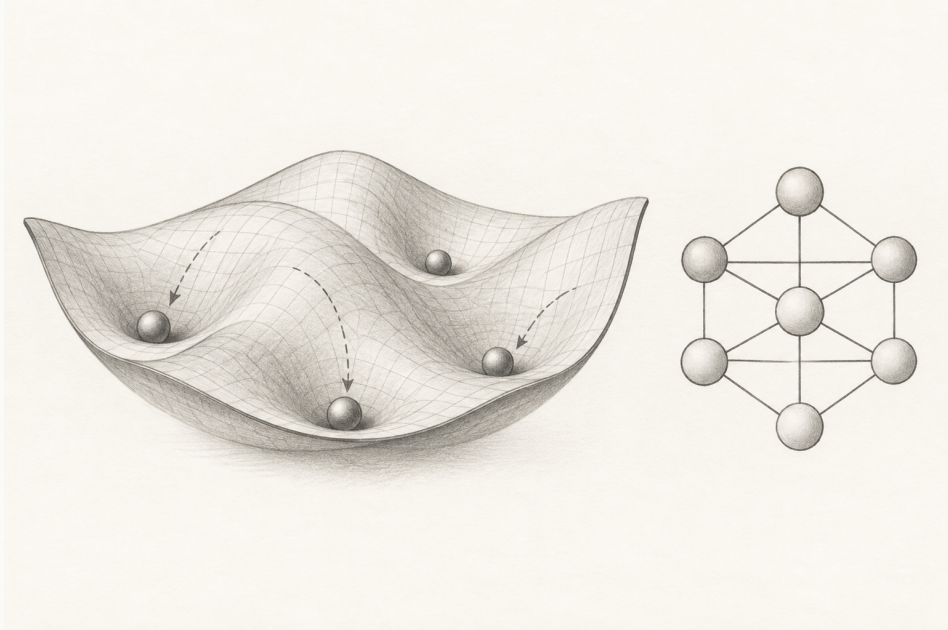


그림 4. 에너지와 확률을 이용하는 신경망의 생각을 비유했습니다.

4.1. 로체스터의 만남: 홉필드 네트워크에 확률과 열을 더하다

1980년대 초 인공지능 학계는 신경망 연구에 냉담했습니다. 1969년 마빈 민스키(Marvin Minsky)와 시모어 페퍼트(Seymour Papert)가 《퍼셉트론(Perceptrons)》에서 단층 신경망의 한계를 수학적으로 보여준 뒤, 많은 연구자가 신경망을 막다른 길로 여겼습니다. 연구비와 인재는 규칙과 기호로 지능을 구현하려는 기호주의 인공지능으로 몰렸습니다. 이 시기에 신경망 연구에 새로운 동력을 제공한 사람은 컴퓨터 과학자가 아니라 물리학자였습니다.

존 홉필드(John J. Hopfield)는 고체물리학과 생물물리학을 넘나들던 이론물리학자입니다. 1982년 당시 캘리포니아공과대학(칼텍)에 있던 그는 미국국립과학원회보(PNAS) 제79권에 「창발적 집단 연산 능력을 가진 신경망과 물리 체계(Neural networks and physical systems with emergent collective computational abilities)」라는 논문을 실었습니다. 홉필드는 뒤에 프린스턴 대학교로 자리를 옮깁니다.

이 논문에서 홉필드가 빌려 온 도구는 스핀 글래스(spin glass)라는 물리 모델입니다. 원자는 작은 자석처럼 행동하는데, 그 자석의 방향을 스핀이라고 합니다. 스핀은 위(+1)나 아래(-1) 가운데 한 상태를 가집니다. 보통의 자석에서는 이웃한 스핀이 모두 같은 방향으로 정렬합니다. 그런데 스핀 글래스에서는 어떤 쌍은 같은 방향을, 어떤 쌍은 반대 방향을 선호합니다. 모든 쌍을 동시에 만족시키는 배열이 없으므로 시스템은 여러 타협 상태 가운데 하나에 머뭙니다. 물리학자들은 이런 안정 상태를 에너지 지형에서 움푹 꺼진 골짜기로 표현합니다.

홉필드는 스핀의 자리에 뉴런을 놓았습니다. 뉴런도 켜짐과 꺼짐이라는 두 상태를 가지기 때문입니다. 이렇게 대응시키면 스핀 사이의 상호작용은 뉴런 사이의 연결 세기, 곧 가중치가 되고, 시스템 전체의 에너지는 네트워크 상태의 에너지가 되며, 에너지 골짜기는 저장된 기억이 됩니다. 홉필드는 여기에 한 가지 조건을 두었습니다. 연결이 대칭이어야 한다는 조건입니다. 어떤 두 뉴런을 골라도 한쪽에서 다른 쪽으로 가는 연결 세기와 그 반대 방향의 연결 세기가 같아야 합니다. 그는 이 조건에서 뉴런이 하나씩 상태를 바꿀 때마다 네트워크 전체의 에너지가 줄거나 그대로 유지된다는 사실을 증명했습니다. 에너지는 끝없이 내려갈 수 없으므로 네트워크는 결국 어느 골짜기의 바닥에서 멈춥니다. 이 바닥을 끌개 상태(attractor state)라고 부릅니다.

이 성질을 이용하면 네트워크를 기억 장치로 쓸 수 있습니다. 저장할 패턴을 골짜기 바닥으로 만들어 두면, 일부가 지워지거나 잡음이 섞인 입력이 들어와도 네트워크는 가장 가까운 골짜기로 내려가 원래 패턴을 복원합니다. 노래의 첫 소절만 듣고 나머지 가사를 떠올리는 일과 비슷합니다. 저장 위치의 주소가 아니라 내용의 일부로 기억을 불러내는 이 방식을 내용 참조 기억(content-addressable memory)이라고 합니다. 홉필드의 논문은 신경망이 안정적으로 작동한다는 것을 물리학의 언어로 보여 주었고, 물리학자들이 신경망에 관심을 갖는 계기가 되었습니다.

뉴욕주 로체스터 대학교에서 제리 펠드먼(Jerry Feldman)이 주관한 워크숍이 열렸습니다. 이 자리에는 카네기멜런 대학교(CMU) 교수 제프리 힌턴(Geoffrey Hinton), 하버드 의대 박사후연구원을 거쳐 존스홉킨스 대학교에 자리 잡은 테리 세이노프스키(Terrence Sejnowski), 그리고

홉필드가 함께 있었습니다. 홉필드는 자신의 네트워크가 반드시 안정 상태로 수렴한다는 증명을 소개했습니다.

힌턴과 세이노프스키는 이 결과에서 가능성과 약점을 함께 보았습니다. 약점은 수렴을 보장하는 바로 그 성질에서 나왔습니다. 홉필드 네트워크는 언제나 에너지가 내려가는 방향으로만 움직입니다. 따라서 처음 입력과 가까운 얇은 골짜기에 한번 들어가면 거기서 벗어날 방법이 없습니다. 이 얇은 골짜기를 국소 극솟점(local minimum)이라고 합니다. 국소 극솟점에서 얻은 답은 저장한 적이 없는 엉뚱한 패턴, 곧 거짓 기억인 경우가 많았습니다.

두 사람은 피츠버그와 볼티모어를 오가며 주말마다 함께 연구했습니다. 심리학, 컴퓨터 과학, 뇌과학, 통계물리학의 지식을 모아 이 약점을 해결할 방법을 찾았습니다. 그들이 찾은 답은 열이었습니다. 뉴런이 상태를 정할 때 정해진 규칙을 그대로 따르지 않고 확률에 따라 결정하게 한 것입니다. 온도를 높이면 잡음이 커져서 뉴런은 에너지가 올라가는 방향으로도 가끔 움직입니다. 그러면 네트워크는 얇은 골짜기의 벽을 넘어 더 깊은 골짜기를 찾아갈 수 있습니다.

두 사람은 이 새로운 네트워크에 19세기 오스트리아 물리학자 루트비히 볼츠만(Ludwig Boltzmann)의 이름을 붙였습니다. 볼츠만은 수많은 입자의 무작위 운동에서 온도와 에너지의 법칙을 이끌어 낸 통계역학의 기초를 세운 사람입니다. 힌턴과 세이노프스키는 1983년 컴퓨터 비전 학술 대회에서 「최적 지각 추론(Optimal perceptual inference)」이라는 논문으로 이 구상을 처음 발표했습니다.

4.2. 에너지 지형, 확률적 이완, 그리고 숨은 유닛

퍼셉트론과 홉필드 네트워크는 정해진 규칙대로만 움직였습니다. 볼츠만 머신은 이를 확률로 추론하는 에너지 기반 모델로 바꾸었습니다. 에너지 기반 모델은 네트워크가 놓일 수 있는 상태마다 에너지라는 수를 하나씩 매기고, 그 수가 낮은 상태를 더 그럴듯한 상태로 보는 모델입니다. 그 구조는 에너지를 정하는 규칙에서 출발합니다. 볼츠만 머신에서 각 뉴런(유닛)의 상태는 꺼짐을 뜻하는 0이나 켜짐을 뜻하는 1입니다. 네트워크 전체의 에너지는 다음과 같이 정합니다.

먼저 서로 연결된 유닛 두 개를 한 쌍으로 묶어, 두 유닛의 상태와 그 사이 연결의 가중치를 곱합니다. 두 유닛이 모두 켜져 있을 때만 이 곱이 가중치와 같아지고, 하나라도 꺼져 있으면 0이 됩니다. 이 값을 모든 쌍에 대해 더한 뒤 부호를 반대로 바꿉니다. 여기에 켜져 있는 유닛들의 문턱값을 모두 더한 것이 네트워크 전체의 에너지입니다.

이 규칙의 뜻은 다음과 같습니다. 양의 가중치로 강하게 연결된 두 유닛이 함께 켜져 있으면 그만큼 에너지가 낮아집니다. 두 유닛의 관계가 잘 맞을수록 시스템이 안정된다는 뜻입니다. 문턱값은 각 유닛이 켜지는 데 필요한 입력의 크기입니다. 문턱값이 큰 유닛은 켜져 있을 때 에너지를 그만큼 높이므로 쉽게 켜지지 않습니다.

이 네트워크를 일정한 온도에서 오래 작동시키면 시스템은 열적 평형(thermal equilibrium)에 이릅니다. 여기서 온도는 유닛이 얼마나 무작위로 움직이는지를 정하는 수로, 물리학의 온도에서 빌려 온 개념입니다.

열적 평형은 개별 유닛의 상태가 계속 바뀌더라도 각 전체 상태가 나타나는 비율은 일정하게 유지되는 상태입니다. 전체 상태는 모든 유닛의 커짐과 꺼짐을 한꺼번에 적은 배열 하나를 가리킵니다. 이때 네트워크가 어떤 전체 상태에 있을 확률은 볼츠만 분포를 따릅니다.

볼츠만 분포에서는 상태마다 수를 하나씩 매기는데, 이 수는 그 상태의 에너지를 온도로 나눈 값이 커질수록 지수함수에 따라 급격히 작아집니다. 한 상태가 나타날 확률은 그 상태에 매긴 수를 모든 상태에 매긴 수의 합으로 나눈 값입니다. 이렇게 나누면 모든 상태의 확률을 더한 값이 1이 됩니다.

이 분포에 따르면 에너지가 낮은 상태일수록 나타날 확률이 높고, 에너지가 높은 상태는 드물게 나타납니다. 온도는 이 차이의 크기를 조절합니다. 온도가 높으면 차이가 줄어 여러 상태가 고르게 나타나고, 온도가 낮으면 에너지가 가장 낮은 상태가 다른 상태보다 훨씬 자주 나타납니다. 뉴런이 이렇게 확률에 따라 상태를 바꾸면서 조금씩 안정된 배열을 찾아 가는 과정을 확률적 이완이라고 부릅니다.

그렇다면 온도를 어떻게 조절해야 깊은 골짜기를 찾을 수 있을까요. 힌턴과 세이노프스키는 모의 담금질(simulated annealing)이라는 방법을 가져왔습니다. 금속을 가열한 뒤 천천히 식히면 내부 구조가 안정된 상태에 가까워질 수 있습니다. 1983년 IBM 연구원 스콧 커크패트릭(Scott Kirkpatrick) 등은 이 원리를 계산 문제에 적용한 모의 담금질 기법을 《사이언스》에 발표했습니다. 처음에는 온도를 높여 네트워크가 에너지 지형의 높은 언덕도 자유롭게 넘나들며 넓은 범위를 탐색하게 합니다. 그

다음 온도를 아주 천천히 내리면 시스템이 지형에서 가장 깊은 골짜기, 곧 전역 극솟점(global minimum)에 머물 가능성이 커집니다. 비슷한 시기에 폴 스몰렌스키(Paul Smolensky)의 하모니 이론, 스튜어트 게만(Stuart Geman)과 도널드 게만(Donald Geman) 형제의 영상 잡음 제거 연구도 같은 발상을 다른 문제에 적용했습니다.

볼츠만 머신의 핵심 성과는 숨은 유닛을 포함한 신경망의 학습 방법입니다. 숨은 유닛은 바깥에서 데이터를 받지도 않고 결과를 내보내지도 않으면서 네트워크 안쪽에서만 작동하는 유닛입니다. 숨은 유닛이 왜 필요했는지는 민스키와 페퍼트의 비판을 보면 알 수 있습니다. 퍼셉트론과 홉필드 네트워크는 바깥 세계와 직접 맞닿은 유닛 사이의 관계만 다룰 수 있었습니다. 그런데 두 입력이 같으면 0, 다르면 1을 내놓는 배타적 논리합(XOR)처럼 쉬워 보이는 문제조차 입력 사이의 직접적인 관계만으로는 풀리지 않습니다. 두 요소씩 짝지어 보아서는 드러나지 않고 여러 요소를 함께 보아야 드러나는 고차원 관계를 표현하려면, 중간에서 이를 정리해 줄 유닛이 필요합니다.

볼츠만 머신은 유닛을 두 종류로 나누어 이 요구에 답했습니다. 가시 유닛(visible units)은 바깥 세계와 맞닿아 있어서 입력 데이터를 받거나 답을 내놓습니다. 숨은 유닛(hidden units)은 바깥에서 값을 정해 주지 않으며 관찰할 수도 없습니다. 숨은 유닛은 가시 유닛과 대칭으로 연결되어, 데이터의 배후에 있는 규칙과 특징을 스스로 찾아 표현합니다.

남은 문제는 숨은 유닛의 가중치를 어떻게 학습시키느냐였습니다. 정답을 알려 줄 수 없는 유닛을 어떻게 학습시킬 것인가는 신경망 연구의 오랜 과

제였습니다. 볼츠만 머신은 이 과제에 확률과 통계물리학으로 답을 제시했고, 다층 신경망 학습의 길을 여는 데 기여했습니다.

4.3. 깨어 있는 단계와 꿈꾸는 단계: 볼츠만 머신의 학습 알고리즘

1985년 데이비드 애클리(David Ackley), 힌턴, 세이노프스키는 학술지 《인지과학(Cognitive Science)》 제9권에 「볼츠만 머신의 학습 알고리즘(A learning algorithm for Boltzmann machines)」을 발표했습니다. 이듬해 힌턴과 세이노프스키는 PDP 연구 그룹이 펴낸 《병렬 분산 처리(Parallel Distributed Processing)》 제1권 제7장 「볼츠만 머신의 학습과 재학습」에서 이 내용을 자세히 설명했습니다.

볼츠만 머신의 학습 목표는 두 확률 분포를 일치시키는 것입니다. 확률 분포는 가능한 상태 각각이 얼마나 자주 나타나는지를 모아 놓은 것입니다. 하나는 실제 세계의 데이터가 가시 유닛에 만드는 분포이고, 다른 하나는 네트워크가 바깥 입력 없이 혼자 작동할 때 가시 유닛에 만드는 분포입니다. 가시 유닛은 바깥 세계와 직접 맞닿아 데이터를 받는 유닛입니다. 두 분포의 차이는 정보이론의 쿨백·라이블러 발산(Kullback-Leibler divergence)으로 측정합니다. 저자들은 이 값에 G라는 이름을 붙였습니다.

이 값은 가시 유닛이 가질 수 있는 상태마다 다음 계산을 하여 모두 더한 것입니다. 그 상태가 실제 데이터에서 나타날 확률을 네트워크가 혼자 작동할 때 나타날 확률로 나누고, 그 비의 자연로그를 구한 뒤, 여기에 실제 데이터에서의 확률을 곱합니다. 두 확률이 같은 상태에서는 비가 1이고 1의 로그는 0이므로 더해지는 것이 없습니다. 두 확률이 어긋나는 상태가 있으면 합은 0보다 커집니다.

이 값이 0이면 네트워크가 만드는 분포는 실제 데이터의 분포와 같습니다. 네트워크가 데이터에 담긴 세계의 모습을 그대로 재현한다는 뜻입니다. 따라서 학습은 이 값을 줄이는 방향으로 가중치를 조정하는 일이 됩니다. 힌턴과 세이노프스키는 이 값을 각 가중치로 미분하면, 곧 가중치 하나를 조금 바꿀 때 이 값이 얼마나 변하는지를 구하면 간결한 결과가 나온다는 것을 증명했습니다.

그 결과는 두 확률의 차이로 나타납니다. 하나는 가시 유닛을 실제 데이터에 고정해 둔 채 평형에 이르렀을 때 연결된 두 유닛이 함께 켜져 있을 확률이고, 다른 하나는 네트워크가 바깥 입력 없이 혼자 작동하며 평형에 이르렀을 때 그 두 유닛이 함께 켜져 있을 확률입니다. 가중치를 조금 바꿀 때 두 분포의 차이가 변하는 정도는 앞의 확률에서 뒤의 확률을 뺀 값을 온도로 나누고 부호를 반대로 한 것과 같습니다.

이 결과에서 가중치를 수정하는 규칙이 바로 나옵니다.

두 유닛을 잇는 가중치의 수정량은 두 측정값의 차이에 작은 수를 곱한 값입니다. 차이는 첫째 값에서 둘째 값을 빼서 구합니다.

여기서 곱하는 작은 수는 한 번에 얼마나 수정할지를 정합니다. 첫째 값과 둘째 값은 서로 다른 두 단계에서 각각 측정하는 값입니다.

첫째는 양성 단계(positive phase)로, 깨어 있는 단계에 비유합니다. 가시 유닛에 실제 데이터를 고정해 두고 숨은 유닛만 자유롭게 움직이게 합니다. 모의 담금질로 열적 평형에 이르면 두 유닛이 함께 켜져 있는 평균

확률을 측정합니다. 이 값이 수정 규칙의 첫째 값이며, 세상을 보는 동안 형성되는 연결에 해당합니다.

둘째는 음성 단계(negative phase)로, 꿈꾸는 단계에 비유합니다. 가시 유닛의 고정을 풀고 네트워크 전체를 혼자 작동시킵니다. 이때도 열적 평형에 이르면 두 유닛이 함께 켜져 있는 평균 확률을 측정합니다. 이 값이 수정 규칙의 둘째 값이며, 네트워크가 바깥 입력 없이 스스로 만들어 낸 연결에 해당합니다.

두 값을 비교하면 규칙의 뜻이 분명해집니다. 실제 데이터를 볼 때 두 유닛이 함께 켜지는 일이 꿈꾸는 단계에서보다 잦으면 둘 사이의 연결을 강하게 합니다. 꿈꾸는 단계에서만 함께 켜지고 실제 데이터에서는 그렇지 않으면 연결을 약하게 합니다. 현실에서 배운 것은 강화하고, 네트워크가 혼자 만들어 낸 근거 없는 연상은 지우는 것입니다.

이 규칙의 장점은 국소성입니다. 한 연결의 세기를 수정할 때 필요한 정보는 그 연결의 양쪽 끝에 있는 두 유닛의 활동뿐입니다. 네트워크 반대편에서 오차 신호를 거꾸로 전달받을 필요가 없습니다. 숨은 유닛이 여러 층으로 놓여 있어도 같은 규칙이 적용됩니다. 뇌의 시냅스도 이웃한 두 신경세포의 활동만 알 수 있다는 점에서, 이 규칙은 생물학적으로 타당해 보였습니다.

수면에 빚댄 설명은 당시 수면 연구의 가설과도 맞닿아 있었습니다. 1983년 프랜시스 크릭(Francis Crick)과 그레이엄 미치슨(Graeme Mitchison)은 《네이처》에 꿈꾸는 수면의 기능에 관한 가설을 발표했

습니다. 렘수면 동안 뇌가 쓸모없는 연상을 지우는 역학습(reverse learning)을 한다는 내용입니다. 볼츠만 머신의 음성 단계는 이 가설과 잘맞았습니다. 크릭은 PDP 그룹의 세미나에 자주 참석하며 신경망 연구자들과 토론했습니다.

그러나 이 알고리즘에는 큰 걸림돌이 있었습니다. 열적 평형에 이르기까지 시간이 너무 오래 걸렸습니다. 양성 단계와 음성 단계마다 유닛을 수천 번에서 수만 번 확률적으로 갱신해야 평형에 가까운 표본을 얻을 수 있었습니다. 이런 반복 표본 추출을 마르코프 연쇄 몬테카를로(MCMC) 방법이라고 합니다. 이 계산 비용 때문에 당시에는 큰 신경망을 실용적으로 학습시키기 어려웠습니다. 비슷한 시기인 1986년에 힌턴이 데이비드 루멜하트(David Rumelhart), 로널드 윌리엄스(Ronald Williams)와 함께 《네이처》에 발표한 오차 역전파 알고리즘은 이보다 훨씬 빨랐습니다. 연구자들의 관심은 한동안 역전파로 쏠렸습니다.

힌턴은 볼츠만 머신을 포기하지 않았습니다. 그는 유닛 사이의 연결을 제한해 계산량을 줄이는 방법을 계속 연구했습니다. 이 연구는 훗날 같은 층안의 연결을 없앤 제한된 볼츠만 머신(RBM)과 대조적 발산(CD) 학습법으로 이어집니다.

4.4. 2024년 노벨 물리학상: 통계물리학과 인공지능의 만남

볼츠만 머신이 처음 발표되고 40년 남짓 지난 2024년 10월 8일, 스웨덴 왕립과학원은 존 홉필드와 제프리 힌턴을 노벨 물리학상 공동 수상자로 발표했습니다. 수상 이유는 "인공신경망을 통한 기계학습을 가능하게 한 기초적 발견과 발명(for foundational discoveries and inventions that enable machine learning with artificial neural networks)"이었습니다. 당시 홉필드는 프린스턴 대학교 명예교수, 힌턴은 토론토 대학교 명예교수였습니다. 힌턴은 수상 소식을 미국 캘리포니아의 한 저렴한 호텔에서 전화로 들었다고 밝혔습니다.

노벨위원회는 발표 자료에서 두 사람의 업적을 하나로 이어지는 흐름으로 설명했습니다. 홉필드는 원자 스핀의 에너지 개념을 이용해 패턴을 저장하고 복원하는 네트워크를 만들었습니다. 힌턴은 그 네트워크를 바탕으로 통계물리학의 도구를 더해, 데이터의 특징을 스스로 찾아내고 분류하며 새로운 예까지 만들어 내는 볼츠만 머신을 발명했습니다. 볼츠만 머신은 수상 이유를 설명하는 공식 자료에 이름이 직접 올랐습니다. 위원회 위원 안데르스 이르베크(Anders Irbäck)는 발표 현장에서 두 사람의 연구가 물리학에 뿌리를 둔다는 점을 짚었습니다. 위원장 엘렌 문스(Ellen Moons)는 힌턴이 효율적인 학습 알고리즘 개발을 이끈 인물로 평가받았다고 말했습니다.

발표 직후 논란도 있었습니다. 일부 물리학자와 온라인 학술 커뮤니티는 이 연구가 자연 현상을 규명하는 물리학인지, 컴퓨터 과학이나 통계학인지를 두고 논쟁했습니다. 학문의 계보를 따라가 보면 두 사람의 연구는 통

계물리학의 무질서계 이론과 스핀 글래스 연구에 뿌리를 둡니다. 2021년 노벨 물리학상을 받은 이탈리아 물리학자 조르조 파리시(Giorgio Parisi)의 업적도 스핀 글래스 같은 무질서한 복잡계 연구였습니다. 수많은 작은 단위가 확률적으로 상호작용하면서 큰 규모의 질서가 나타나는 과정을 설명한다는 점에서, 파리시의 스핀 글래스와 힌턴의 볼츠만 머신은 같은 질문을 공유합니다. 볼츠만 머신에서는 그 질서가 기억과 학습의 형태로 나타난다는 점이 다를 뿐입니다.

힌턴은 수상 뒤 기자들에게서 연구 내용을 짧게 설명해 달라는 요청을 여러 번 받았습니다. 그는 몇 분 만에 설명할 수 있는 연구라면 노벨상감이 아니라는 물리학자 리처드 파인만(Richard Feynman)의 오래된 농담을 빌려 답한 것으로 전해집니다.

2024년 12월 10일 스톡홀름 콘서트홀에서 시상식이 열렸습니다. 힌턴은 스웨덴 국왕 칼 16세 구스타프에게서 노벨 물리학상을 받았고, 토론토 대학교에서는 학생과 동료 100명가량이 모여 시상식을 지켜보았습니다. 상금은 1,100만 스웨덴 크로나, 캐나다 달러로 약 140만 달러였으며 홉필드와 나누어 받았습니다. 힌턴은 자기 몫의 일부를 캐나다 원주민의 깨끗한 물 접근성을 높이는 온타리오주 단체 워터 퍼스트(Water First)와, 신경 다양성을 지닌 청년을 돕는 단체에 기부하겠다고 밝혔습니다.

상금의 또 다른 일부는 세이노프스키와 함께한 연구를 기리는 데 쓰였습니다. 신경정보처리시스템학회(NeurlPS) 측은 2024년 12월 힌턴이 상금 일부로 새로운 상을 만든다고 발표했습니다. 상금은 매년 1만 미국 달러이며, 뇌가 작동하는 방식에 관한 새로운 이론을 제안한 논문을 쓴 40세

미만 연구자 두 명 이상의 팀에게 수여합니다. 상의 이름은 세이노프스키·힌턴 상(Sejnowski-Hinton Prize)으로, 계산신경생물학자 테리 세이노프스키와 힌턴의 이름에서 따왔습니다. 40여 년 전 피츠버그와 볼티모어를 오가며 함께 볼츠만 머신을 만든 두 사람의 이름이 젊은 연구자를 위한 상에 나란히 올랐습니다.

힌턴은 노벨상 박물관에도 볼츠만 머신과 관련된 물건을 남겼습니다. 그가 2024년에 기증한 마이크로칩에는 초기 볼츠만 머신이 들어 있습니다. 이 칩은 1987년 조슈아 알스펙터(Joshua Alspector)가 만든 것으로, 볼츠만 머신을 전용 칩에 구현할 수 있음을 보여 주었습니다. 칩 속의 볼츠만 머신은 노드가 여섯 개뿐인 작은 네트워크여서 쓰임새는 제한적이었지만, 원리를 실증한 사례로서 기계학습과 인공지능 발전의 방향을 제시했습니다.

1980년대 초 대부분의 학자가 신경망을 외면할 때 홉필드는 스펀 글래스로 신경망의 안정성을 증명했습니다. 힌턴과 세이노프스키는 여기에 열과 확률을 더하고 숨은 유닛을 넣어 스스로 학습하는 기계를 만들었습니다. 볼츠만 머신은 너무 느려서 당시에는 널리 쓰이지 못했습니다. 그러나 에너지로 데이터를 표현하고, 현실과 모델의 차이를 줄이며, 숨은 구조를 스스로 찾는다는 발상은 제한된 볼츠만 머신과 심층 신뢰 신경망을 거쳐 오늘날의 생성형 인공지능까지 이어졌습니다. 물리학의 언어에서 출발한 이 발상이 40년 뒤 노벨 물리학상으로 평가받은 것입니다.

제5장. 더 빠르게 배우는 방법

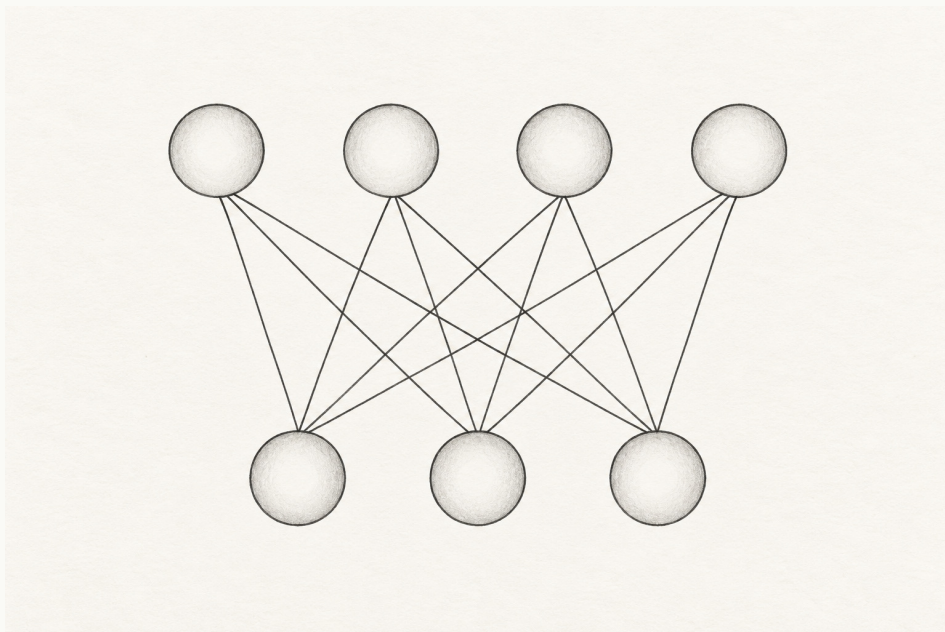


그림 5. 두 층 사이의 연결로 구성된 제한된 볼츠만 머신의 개념도입니다.

1980년대에 나온 볼츠만 머신은 숨은 유닛을 가진 신경망도 스스로 학습할 수 있다는 것을 보여 주었습니다. 그러나 계산에 시간이 너무 오래 걸려 실제 문제에는 쓰기 어려웠습니다. 힌턴은 그 뒤 15년 가까이 이 문제를 붙들었고, 두 가지 답을 내놓았습니다. 하나는 신경망의 연결을 일부러 줄인 제한된 볼츠만 머신(Restricted Boltzmann Machine, RBM)이고, 다른 하나는 이 신경망을 빠르게 학습시키는 대조적 발산(Contrastive Divergence, CD)입니다. 두 아이디어는 모두 볼츠만 머신이 왜 느린가라는 한 가지 문제에서 출발했습니다.

5.1. 너무 느린 볼츠만 머신과 '전문가들의 꿈'이라는 발상

신경망은 유닛이라는 작은 계산 단위가 서로 이어진 구조입니다. 이 장의 유닛은 대부분 켜짐을 뜻하는 1과 꺼짐을 뜻하는 0 가운데 한 값을 가집니다. 두 유닛을 잇는 연결에는 가중치라는 수가 붙어 있습니다. 가중치가 양수이면 두 유닛은 함께 켜지려 하고, 음수이면 한쪽이 켜질 때 다른 쪽은 꺼지려 합니다. 학습은 이 가중치를 조금씩 고쳐 가는 일입니다. 확률은 어떤 일이 얼마나 자주 일어나는지를 0과 1 사이의 수로 나타낸 것으로, 1에 가까울수록 자주 일어납니다.

제프리 힌턴과 테리 세이노프스키가 1980년대 중반에 발표한 볼츠만 머신은 통계물리학의 개념을 신경망에 옮겨 온 모델입니다. 이 모델의 유닛은 정해진 규칙이 아니라 확률에 따라 켜지거나 꺼집니다. 모델은 어느 유닛이 켜지고 꺼졌는지를 적은 신경망 전체의 상태마다 에너지 값을 매기고, 에너지가 낮은 상태일수록 더 자주 나타나도록 확률을 정합니다.

학습은 두 단계로 이루어집니다. 먼저 실제 데이터를 입력 쪽 유닛에 고정해 둔 '깨어 있는 단계'에서 연결된 두 유닛이 함께 켜지는 정도를 잽니다. 이어서 입력을 풀어 신경망이 스스로 움직이게 둔 '꿈꾸는 단계'에서 같은 값을 다시 잽니다. 그리고 두 값의 차이만큼 그 연결의 가중치를 고칩니다. 깨어 있을 때 더 자주 함께 켜진 연결은 강해지고, 꿈꾸 때 더 자주 함께 켜진 연결은 약해집니다.

문제는 꿈꾸는 단계에 있었습니다. 어떤 데이터가 나타날 확률을 구하려면 분할 함수(Partition Function)를 알아야 하며, 이 값을 흔히 Z 라고

적습니다. Z 는 신경망이 가질 수 있는 모든 상태의 에너지를 지수 함수에 넣어, 에너지가 낮을수록 큰 값이 되도록 바꾼 뒤 모두 더한 것입니다. 한 상태의 확률은 그 상태의 값을 Z 로 나눈 몫이므로 Z 없이는 확률을 정할 수 없습니다. 그런데 유닛 하나가 두 가지 값을 가지므로 상태의 수는 2를 유닛의 개수만큼 거듭 곱한 수가 됩니다. 유닛이 모두 합해 100개만 되어도 2의 100제곱, 곧 약 10의 30제곱에 이릅니다. 이만큼의 항을 하나하나 더하는 일은 어떤 컴퓨터로도 할 수 없습니다.

그래서 연구자들은 Z 를 직접 계산하지 않고 샘플링을 썼습니다. 샘플링은 확률에 따라 값을 실제로 뽑아 보는 일이며 표본 추출이라고도 합니다. 여러 번 뽑은 결과를 모으면 모든 상태를 따져 보지 않고도 어떤 상태가 자주 나오는지 알 수 있습니다. 볼츠만 머신에는 마르코프 체인 몬테카를로(MCMC)라는 샘플링 방법과 모의 담금질(Simulated Annealing)이 쓰였습니다. 유닛의 값을 확률에 따라 거듭 바꾸며 신경망을 오랫동안 흔들어서, 각 상태가 나타나는 비율이 더는 변하지 않는 열적 평형에 이르게 한 뒤 그때의 통계를 재는 방식입니다. 그런데 가중치를 한 번 고칠 때마다 이 과정을 거쳐야 했으므로, 1980년대와 1990년대의 컴퓨터로는 학습이 끝없이 길어졌습니다. 힌턴은 2024년 노벨상 강연에서도 원래의 볼츠만 머신이 너무 느렸고, 몇 년 뒤 뉴런 사이의 연결에 제한을 두어 효율을 높인 제한된 볼츠만 머신을 고안했다고 돌아보았습니다.

힌턴은 이 문제를 풀기 위해 확률 모델을 구성하는 방식부터 다시 검토했습니다. 그 결과가 '전문가들의 곱(Products of Experts, PoE)'입니다. 이 연구는 힌턴이 1998년부터 2001년까지 초대 소장을 맡았던 유니버시티 칼리지 런던(UCL) 개츠비 계산신경과학연구소 시절에 나왔습니다.

2000년에 기술 보고서로 먼저 공개되었고, 2002년 학술지 《Neural Computation》 제14권 제8호에 정식 논문으로 실렸습니다.

여기서 전문가는 데이터마다 그럴듯한 정도를 확률로 매기는 작은 모델을 가리킵니다. 가능한 모든 데이터에 확률이 어떻게 나뉘어 있는지를 통틀어 확률 분포라고 합니다. 당시에는 가우시안 혼합 모델(GMM)처럼 여러 전문가의 분포를 더해서 쓰는 혼합 모델(Mixture Models)이 널리 쓰였습니다. 전문가마다 비중을 정해 두고, 각 전문가가 매긴 확률에 그 비중을 곱해 모두 더하는 방식입니다.

힌턴은 이 방식이 차원이 높은 데이터, 곧 하나를 나타내는 데 많은 수가 필요한 데이터에는 잘 맞지 않는다고 지적했습니다. 사람 얼굴 사진을 예로 들면, 조명, 자세, 표정, 각도 같은 요인이 얹혀 약 35차원의 굽은 공간(매니폴드)을 이룹니다. 혼합 모델로 이 공간을 나타내려면 전문가 하나하나가 35차원 전체를 넓게 덮어야 합니다. 그런데 여러 분포를 더한 결과는 그 재료가 된 분포보다 뽀족해질 수 없습니다. 결국 전체 분포가 뭉툭해집니다.

전문가들의 굽은 반대로, 전문가들이 매긴 확률을 더하지 않고 곱합니다. 곱한 값은 그대로는 확률이 되지 못하므로, 가능한 모든 데이터에 대해 같은 곱을 구해 더한 값으로 나누어 전체 합이 1이 되게 맞춥니다. 이 과정을 재규격화라고 합니다. 곱하는 방식에서는 전문가 하나만 낮은 확률을 매겨도 결과가 크게 낮아집니다. 그래서 각 전문가는 데이터의 일부 특징에 대해서만 조건을 달아도 됩니다. 모든 조건을 동시에 만족하는 좁은 영역에서만 확률이 높게 남기 때문입니다.

힌턴은 문장을 예로 들었습니다. 한 전문가는 시제가 맞는지만 보고, 다른 전문가는 주어와 동사의 수가 일치하는지만 봅니다. 둘을 곱하면 두 규칙을 모두 지키는 문장에만 높은 확률이 주어집니다. 혼합 모델에서는 전문가마다 문제 전체를 감당해야 했지만, 전문가들의 곱에서는 좁은 조건을 하나씩 맡은 전문가들이 함께 날카로운 분포를 만듭니다.

전문가들의 곱에는 장점이 하나 더 있었습니다. 데이터가 주어지면 각 전문가 안에 숨어 있는 변수들이 서로 독립이 됩니다. 독립은 한쪽의 값을 알아도 다른 쪽의 확률이 달라지지 않는다는 뜻이며, 데이터가 주어졌다는 조건 아래에서 성립하므로 조건부 독립이라 합니다. 덕분에 숨은 변수의 상태를 한 번의 계산으로 정확하게 추론할 수 있습니다. 다만 재규격화에서 나누는 값은 여전히 모든 경우를 더해야 하므로 계산하기 어렵습니다. 힌턴은 이 남은 문제를 5.3절에서 설명할 대조적 발산으로 피해 갔습니다.

5.2. 제한된 볼츠만 머신(RBM)의 구조와 조건부 독립

전문가들의 곱이라는 발상을 신경망의 구조로 옮긴 것이 제한된 볼츠만 머신입니다. 이 구조 자체는 힌턴이 처음 만든 것이 아니며, 그 뿌리는 1986년으로 거슬러 올라갑니다. 폴 스몰렌스키(Paul Smolensky)는 그 해 출간된 《병렬 분산 처리(Parallel Distributed Processing)》 제1권 제6장에서 '하모니 이론'을 소개하며 하모늄(Harmonium)이라는 모델을 내놓았습니다. 1992년에는 요아브 프로인트(Yoav Freund)와 데이비드 하우슬러(David Haussler)가 이진 벡터, 곧 0과 1이 들어선 데이터의 분포를 두 층 신경망으로 학습하는 방법을 연구했습니다. 힌턴은 2002년 논문에서 이 구조가 숨은 유닛 하나하나를 전문가로 보는 전문가들의 곱이라는 점을 짚었고, 이를 볼츠만 머신의 에너지 틀과 결합해 학습 이론을 세웠습니다.

RBM은 두 층으로만 이루어집니다. 가시층(visible layer)은 데이터를 받는 층이고, 은닉층(hidden layer)은 데이터에 직접 드러나지 않는 특징을 찾는 층입니다. 은닉층의 유닛, 곧 은닉 유닛 하나는 가시층에 특정한 무늬가 나타날 때 켜지도록 학습됩니다. 가시 유닛과 은닉 유닛 사이의 연결에는 어느 방향으로 신호가 가든 같은 값을 쓰는 대칭 가중치가 있습니다. 유닛마다 바이어스라는 수도 하나씩 있어, 다른 유닛과 관계없이 그 유닛이 얼마나 켜지기 쉬운지를 정합니다.

RBM이 일반 볼츠만 머신과 다른 점은 같은 층 안의 연결을 모두 없앴다는 데 있습니다. 가시 유닛끼리, 은닉 유닛끼리는 이어져 있지 않습니다. 수학에서는 이런 모양을 이분그래프(Bipartite Graph)라고 부릅니다. 교

실에 비유하면 앞줄 학생은 뒷줄 학생과만 말할 수 있고, 같은 줄 옆자리와는 말할 수 없다는 규칙과 같습니다.

RBM의 에너지는 세 가지를 더한 뒤 부호를 뒤집은 값입니다. 켜진 가시 유닛들의 바이어스, 켜진 은닉 유닛들의 바이어스, 그리고 양쪽 유닛이 함께 켜진 연결들의 가중치입니다. 가중치가 큰 연결의 두 유닛이 함께 켜질수록 에너지가 낮아지고, 볼츠만 분포의 규칙에 따라 그 상태의 확률은 높아집니다.

같은 층 안의 연결을 없앤 효과는 큼니다. 가시층의 값이 모두 정해지면 은닉 유닛들은 서로 영향을 주고받지 않습니다. 은닉 유닛끼리 직접 이어진 길이 없고, 유일한 통로인 가시층은 값이 이미 고정되어 있기 때문입니다. 각 은닉 유닛이 켜질 확률은 자기에게 들어오는 신호만으로 따로 계산됩니다. 켜진 가시 유닛들과 이어진 가중치를 모두 더하고 자기 바이어스를 보탠 뒤, 그 합을 시그모이드 함수에 넣습니다. 시그모이드 함수는 어떤 수를 넣어도 0과 1 사이의 값을 돌려주며, 넣은 수가 클수록 1에 가까운 값을 냅니다. 이 값이 그 유닛이 켜질 확률입니다.

반대 방향도 마찬가지입니다. 은닉층의 값이 모두 정해지면 가시 유닛들이 서로 독립이 되고, 각 가시 유닛이 켜질 확률은 켜진 은닉 유닛들에서 오는 가중치의 합과 자기 바이어스로 같은 방식으로 계산됩니다. 그래서 한 층 전체가 어떤 값을 가질 확률은 유닛별 확률을 곱한 값이 됩니다. 이를 팩토리얼 조건부 독립(Factorial Conditional Independence)이라 합니다. 이 독립은 반대쪽 층의 값이 모두 주어졌다는 조건 아래에서만 성

립합니다. 그 조건이 없으면 같은 층의 유닛들도 반대쪽 층을 거쳐 서로 영향을 미칩니다.

이 성질의 가치는 방향성 그래프 모델과 비교하면 분명해집니다. 원인에서 결과로 화살표가 그려진 모델에서는 '설명 해소(Explaining Away)'라는 현상이 생깁니다. 잔디가 젖은 이유로 비와 스프링클러 두 가지를 생각해 보겠습니다. 원래 두 원인은 서로 관계가 없습니다. 그런데 잔디가 젖었다는 사실을 안 뒤에는, 비가 왔다는 것을 확인하는 순간 스프링클러가 작동했을 가능성이 낮아집니다. 결과를 관찰하면 원인들끼리 서로 얹히는 것입니다. 이런 모델에서 숨은 원인의 확률을 구하려면 MCMC나 복잡한 근사 계산을 거쳐야 했습니다.

RBM에서는 이런 얹힘이 생기지 않습니다. 입력 데이터가 들어오면 행렬과 벡터의 곱셈 한 번, 시그모이드 계산 한 번으로 은닉층 전체의 확률을 정확하게 얻습니다. 다만 이것은 주어진 데이터에서 숨은 특징을 읽어 내는 추론에 한정된 이야기입니다. 가중치를 학습하는 데에는 여전히 Z가 걸림돌로 남아 있었습니다.

5.3. 대조적 발산(CD) 학습법의 원리

RBM의 학습은 실제 데이터가 나올 확률이 커지도록 가중치를 고치는 일입니다. 어느 연결을 어느 쪽으로 고칠지는 두 평균의 차이로 정해집니다. 하나는 데이터를 가시층에 고정했을 때 그 연결의 가시 유닛과 은닉 유닛이 함께 켜지는 비율입니다. 다른 하나는 데이터 없이 신경망이 스스로 움직일 때 두 유닛이 함께 켜지는 비율입니다. 앞의 값이 뒤의 값보다 크면 가중치를 올리고, 작으면 내립니다. 모든 연결에서 두 값이 같아지면 모델이 혼자 만들어 내는 통계가 데이터의 통계와 일치하므로 더 고칠 것이 없습니다.

앞의 값은 5.2절의 시그모이드 계산으로 쉽게 구합니다. 어려운 것은 뒤의 값입니다. 이 값을 정확히 얻으려면 깁스 샘플링(Gibbs Sampling)을 해야 합니다. 가시층의 값으로 은닉층의 확률을 계산해 은닉 유닛의 값을 뽑고, 그 값으로 다시 가시층의 확률을 계산해 가시 유닛의 값을 뽑는 일을 번갈아 되풀이하는 방법입니다. 그런데 이 왕복을 평형에 이를 때까지, 이론상 무한히 반복해야 했습니다. 연결을 줄여 추론은 빨라졌지만 학습의 병목은 그대로였던 것입니다.

힌턴은 이 병목을 과감하게 잘라 냈습니다. 평형까지 갈 필요 없이, 실제 데이터에서 출발해 깁스 샘플링을 단 한 번만 하고 멈추어도 쓸 만한 학습 방향을 얻을 수 있다고 본 것입니다. 한 번만 반복하는 방식을 CD-1이라 부르며, 절차는 네 단계입니다.

첫째, 실제 데이터를 가시층에 넣습니다. 각 은닉 유닛이 켜질 확률을 계산하고, 그 확률에 따라 0이나 1의 값을 뽑습니다. 이때 가시 유닛과 은닉 유닛이 함께 켜지는 정도를 연결마다 잽니다.

둘째, 뽑힌 은닉층의 값에서 거꾸로 가시층의 확률을 계산해 새로운 가시층의 값을 만듭니다. 신경망이 데이터를 자기 방식으로 다시 그려 본 결과입니다. 힌턴은 이를 재구성(reconstruction)이라 불렀고, 때로는 '지어낸 이야기'라는 뜻의 confabulation이라고도 했습니다.

셋째, 재구성을 가시층에 다시 넣어 은닉층의 값을 얻고, 양쪽이 함께 켜지는 정도를 같은 방법으로 잽니다.

넷째, 첫째 단계에서 잦 값에서 셋째 단계에서 잦 값을 뺀 차이에 학습률을 곱한 만큼 가중치를 고칩니다. 학습률은 한 번에 얼마나 움직일지를 정하는 작은 수입니다. 데이터에서 더 자주 함께 켜진 연결은 가중치가 올라가고, 재구성에서 더 자주 함께 켜진 연결은 가중치가 내려갑니다.

이 절차가 왜 통하는지는 에너지 지형을 떠올리면 이해하기 쉽습니다. 가로축을 가능한 모든 데이터, 세로축을 에너지라고 하겠습니다. 학습이 잘된 모델은 실제 데이터가 있는 곳에 깊은 골짜기를 갖습니다. 재구성은 아무렇게나 고른 점이 아닙니다. 실제 데이터 근처에서, 모델이 지금 더 그럴 듯하다고 여기는 쪽으로 한 걸음 옮겨 간 점입니다. 힌턴은 이를 '아쉬운 오답(near miss)'에 견주었습니다. 가중치 수정에서 더하는 쪽은 진짜 데이터가 있는 자리의 에너지를 낮추고, 빼는 쪽은 그 바로 옆 오답 자리의 에너지를 높입니다. 이 과정을 되풀이하면 에너지 지형에 데이터를 따라

깊은 도랑이 파입니다. 모델은 먼 곳까지 모두 살피지 않고, 자기가 틀리기 쉬운 곳만 골라 고치는 셈입니다.

힌턴은 이 방법에 수학적 근거도 제시했습니다. 근거에는 세 가지 분포가 나옵니다. 실제 데이터의 분포, 모델이 평형에 이르렀을 때의 분포, 데이터에서 출발해 깃스 샘플링을 정해진 횟수만큼 거친 뒤의 분포입니다. 두 분포가 얼마나 다른지는 쿨백 라이블러(KL) 발산이라는 값으로 재는데, 두 분포가 같으면 0이고 다를수록 커집니다. 데이터가 나올 확률을 가장 크게 만들려는 보통의 최대우도 학습은 데이터 분포와 평형 분포 사이의 발산을 줄입니다. 대조적 발산은 그 대신 두 발산의 차이를 줄입니다. 데이터 분포와 평형 분포 사이의 발산에서, 샘플링을 거친 분포와 평형 분포 사이의 발산을 뺀 값입니다.

샘플링을 거듭할수록 분포는 평형에 가까워지므로 빼는 값은 앞의 값보다 크지 않습니다. 따라서 이 차이는 0보다 작아지지 않으며, 모델이 데이터를 그대로 재구성해 샘플링을 거친 분포가 데이터 분포와 같아질 때 0이 됩니다. '대조적 발산'이라는 이름은 이렇게 두 발산을 맞대어 비교한다는 데서 나왔습니다.

CD-1에는 통계적인 이점도 있습니다. 진짜 데이터와 그 재구성은 언제나 한 쌍으로 묶여 있습니다. 임상 시험에서 비슷한 환자를 짝지어 비교하면 우연에 따른 오차가 줄어드는 것처럼, 짝을 지은 비교는 샘플링 결과의 흔들림(분산)을 줄입니다. 그래서 깃스 샘플링을 한 번만 해도 학습이 안정적으로 진행되었습니다.

대조적 발산은 근사이며 완벽한 방법이 아닙니다. 평형에서 재야 할 평균을 한 걸음 뒤의 재구성으로 대신했으므로, 수정 방향이 데이터의 확률을 키우는 정확한 방향과 어긋날 수 있습니다. 이 한계는 힌턴 자신과 동료들의 후속 연구에서 드러났습니다. 미겔 카레이라페르피냐(Miguel Carreira-Perpiñán)와 힌턴은 2005년 연구에서 CD가 참된 최대우도 해와 조금 어긋난 지점으로 갈 수 있지만 그 어긋남은 대개 작다는 것을 보였습니다. 일리야 수츠케버와 틸먼 틸레만(Tijmen Tieleman)은 2010년에 CD의 수정 방향이 어떤 하나의 목적 함수를 정확히 내려가는 기울기는 아니라고 분석했습니다. 줄이려는 값 하나를 정해 두고 그 값이 가장 빨리 줄어드는 쪽으로 움직이는 방법은 아니라는 뜻입니다. 이런 약점을 보완하기 위해 틸레만은 2008년, 깃스 샘플링의 연쇄인 마르코프 체인을 매번 새로 시작하지 않고 이어서 돌리는 지속적 대조적 발산(Persistent CD, PCD)을 제안했습니다. 힌턴은 2010년 「RBM 학습을 위한 실용 지침」을 써서 학습률, 은닉 유닛 수, 가중치 초기값 같은 실제 설정 방법을 정리했습니다.

5.4. RBM의 변형과 사전학습으로의 확장

RBM과 CD의 조합은 0과 1로 된 패턴을 넘어 여러 종류의 데이터로 적용 범위를 넓혔습니다. 그 과정에서 세 가지 변화가 두드러졌습니다.

첫 번째는 가시 유닛을 바꾼 가우시안 베르누이 RBM(Gaussian-Bernoulli RBM)입니다. 음성의 주파수 특징(MFCC)이나 사진의 밝기처럼 연속된 실수 값을 다루려면 가시 유닛이 0과 1만 가져서는 안 됩니다. 이 모델은 가시 유닛을 가우시안 잡음이 섞인 선형 유닛으로 바꾸었습니다. 가시 유닛의 값은 정규분포에서 뽑히며, 그 평균은 커진 은닉 유닛들에서 오는 가중치의 합에 자기 바이어스를 더한 값입니다. 이 모델은 뒤에 음성 인식 연구의 출발점이 되었습니다.

두 번째는 은닉 유닛을 바꾼 것입니다. 비노드 나이르(Vinod Nair)와 힌턴은 2010년 국제기계학습학회(ICML)에서 정류 선형 유닛(Rectified Linear Unit, ReLU)을 쓰는 RBM을 발표했습니다. 두 사람은 가중치는 같고 바이어스만 조금씩 다른 이진 유닛을 무한히 쌓은 '계단식 시그모이드 유닛'을 생각했습니다. 이 유닛의 동작은 입력이 0보다 작으면 0을, 0보다 크면 입력에 비례하는 값을 내는 함수로 근사됩니다. 여기에 잡음을 더해, 입력과 잡음의 합이 0보다 크면 그 합을 내고 아니면 0을 내는 형태로 썼습니다. 이 ReLU는 2012년 AlexNet에 쓰였고, 이후 딥러닝의 표준 활성화 함수가 되었습니다.

세 번째 변화는 RBM을 층층이 쌓는 방법이었습니다. 2006년 힌턴, 사이먼 오신데로(Simon Osindero), 예휘 테(Yee-Whye Teh)는 《Neural

Computation》 제18권에 발표한 논문에서 첫 번째 RBM의 은닉층 출력을 두 번째 RBM의 데이터로 삼아 학습시키는 방식을 제시했습니다. 정답 표시가 없는 데이터로 한 층씩 차례로 학습하는 이 방법을 탐욕적 층별 비지도 사전학습(Greedy Layer-wise Unsupervised Pre-training)이라 부르고, 이렇게 만든 신경망을 심층 신뢰 신경망(Deep Belief Network, DBN)이라 합니다. 같은 해 힌턴과 루슬란 살라쿠트디노프(Ruslan Salakhutdinov)는 《Science》 제313권에 실은 논문에서, 입력을 짧은 부호로 줄였다가 되살리는 신경망인 깊은 오토인코더를 RBM으로 사전학습한 뒤 오차 역전파로 미세 조정했습니다. 그 결과 고차원 데이터를 짧은 부호로 줄이는 데 기존 방법보다 좋은 결과를 냈습니다. 2009년 두 사람은 모든 층이 방향 없는 대칭 연결로 이어진 심층 볼츠만 머신(Deep Boltzmann Machine, DBM)으로 이론을 넓혔습니다.

RBM 사전학습은 그 뒤 음성 인식에서 성과를 냈습니다. 토론토대학교, 마이크로소프트, 구글, IBM의 연구진이 함께 쓴 2012년 논문은 DBN 방식으로 초기화한 깊은 신경망이 음향 모델에서 기존 방식을 앞섰다고 보고했습니다.

그 뒤 상황은 바뀌었습니다. ReLU, 드롭아웃, 대규모 데이터, 그래픽 처리 장치(GPU)가 갖추어지자 사전학습 없이도 깊은 신경망을 처음부터 오차 역전파로 학습할 수 있게 되었습니다. 오늘날 실제 서비스에서 RBM을 쓰는 경우는 드뭅니다. 2024년 12월 8일 힌턴의 노벨상 강연을 소개한 글은 이 강연을 딥러닝 혁명을 일으킨 '역사적 촉매'에 관한 이야기로 요약하면서, 볼츠만 머신 학습 알고리즘이 지금은 쓰이지 않는다는 점을 함께 언급했습니다. 화학 반응을 일으킨 뒤 생성물에는 남지 않는 촉매처럼,

RBM은 깊은 신경망 학습의 문을 열고 물러났습니다. 힌턴은 노벨 주간에 우표만 한 크기의 초기 볼츠만 머신 칩을 노벨상 박물관에 기증했습니다.

RBM 자체는 쓰이지 않게 되었지만 RBM과 CD가 남긴 생각은 이어지고 있습니다. 모든 경우를 다 계산해 완벽한 확률을 구하는 대신, 진짜 데이터의 에너지는 낮추고 그럴듯한 가짜의 에너지는 높이는 방식으로 모델을 다듬는다는 생각입니다. 안 르쿤 등은 이런 접근을 에너지 기반 모델(Energy-Based Model)의 틀로 정리해 왔습니다. 진짜 짝과 가짜 짝을 맞대어 학습하는 오늘날의 대조 학습(Contrastive Learning)도 같은 문제의식을 공유한다고 보는 연구자들이 있습니다.

힌턴이 볼츠만 머신의 속도 문제에 내놓은 답은 세 가지였습니다. 확률을 더하는 대신 곱하는 전문가들의 곱으로 높은 차원의 데이터를 날카롭게 표현하는 길을 열었습니다. 같은 층 안의 연결을 없앴 RBM으로 숨은 특징을 한 번에 정확히 추론할 수 있게 했습니다. 그리고 평형까지 기다리지 않고 한 걸음만 재구성해 비교하는 대조적 발산으로 학습 속도의 벽을 넘었습니다. 연결을 줄이고 계산을 줄인 이 선택이 2006년 이후 딥러닝이 다시 주목받는 데 디딤돌이 되었습니다.

제6장. 깊은 신경망이 다시 주목받다

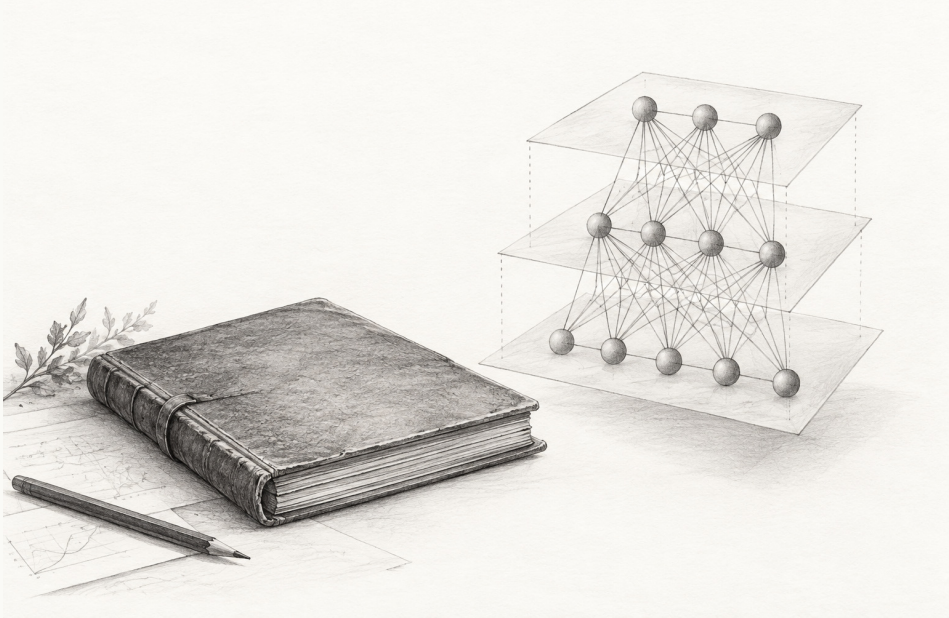


그림 6. 여러 층으로 이루어진 신경망을 연구하는 생각을 표현했습니다.

6.1. 제2차 AI 겨울과 기울기 소멸 문제

1986년 데이비드 러멜하트, 제프리 힌턴, 로널드 윌리엄스가 오차 역전파(backpropagation) 방법을 《네이처》에 발표하자 신경망에 대한 기대가 크게 높아졌습니다. 오차 역전파는 신경망이 낸 답과 정답의 차이를 계산한 뒤, 그 차이를 출력층에서 입력층 쪽으로 거꾸로 보내며 연결 강도, 곧 가중치를 조금씩 고치는 방법입니다. 층이 두세 개인 신경망에서는 이 방법이 잘 작동했습니다. 그런데 입력층과 출력층 사이에 은닉층을 여러 겹 쌓은 깊은 신경망에 같은 방법을 적용하자 학습이 제대로 되지 않았습니다. 당시 깊은 신경망의 학습을 가로막은 문제는 크게 세 가지였습니다.

첫째 걸림돌은 기울기 소멸(vanishing gradient) 문제였습니다. 역전파에서 오차 신호는 층을 하나 지날 때마다 미분값과 곱해집니다. 당시 널리 쓰던 시그모이드 함수는 미분값이 가장 클 때도 0.25에 그칩니다. 1보다 작은 수를 여러 번 곱하면 값은 0에 가까워집니다. 그래서 출력층에서 먼 아래쪽 층은 학습 신호를 거의 받지 못했습니다. 반대로 값이 층마다 커져서 학습이 불안정해지는 기울기 폭발도 일어났습니다. 어느 쪽이든 아래 층까지 적절한 크기의 학습 신호가 전달되지 않는다는 점에서는 같은 문제였습니다.

둘째는 최적화의 실패였습니다. 최적화는 오차가 작아지는 가중치를 찾아가는 과정입니다. 가중치에 따라 오차가 오르내리는 모습을 지형으로 그리면 깊은 신경망의 오차 지형은 굴곡이 많고 복잡합니다. 무작위로 정한 가중치에서 출발한 깊은 신경망은 이 지형의 국소 극솟점(local minimum)이나 평평한 안장점(saddle point) 근처에서 헤매다 멈추곤

했고, 좋은 가중치를 찾지 못했습니다. 국소 극솟점은 주변보다는 낮지만 전체에서 가장 낮지는 않은 지점입니다. 안장점은 어떤 방향으로든 오르막이고 다른 방향으로든 내리막이어서 그 자리의 기울기가 0에 가까운 지점입니다.

셋째는 데이터 부족에서 오는 과적합(overfitting)이었습니다. 사람이 일일이 정답을 붙인 데이터, 곧 라벨링된 데이터는 구하기 어려웠습니다. 가중치가 수십만 개에서 수백만 개에 이르는 깊은 신경망은 적은 데이터 속의 우연한 잡음까지 외워 버렸습니다. 연습 문제는 잘 맞히지만 처음 보는 문제에서는 성적이 떨어지는 현상과 같습니다.

이런 사정 때문에 기계학습 학계의 관심은 다른 방법으로 옮겨 갔습니다. 서포트 벡터 머신(SVM)과 커널 기법은 학습 문제를 볼록 최적화(convex optimization) 문제로 바꿀 수 있었습니다. 볼록 최적화에서는 오차 지형에 골짜기가 하나뿐이어서 언제나 같은 최적해에 도달합니다. 국소 극솟점에 빠질 염려가 있는 신경망과 대비되는 장점이었습니다. 결정 트리 여러 개를 묶은 랜덤 포레스트도 좋은 성능을 냈습니다. 그에 비해 신경망은 이론은 그럴듯하지만 실제로는 다루기 어려운 방법으로 여겨졌습니다.

그 결과 1990년대 중반부터 2000년대 초반까지 인공신경망 연구는 긴 침체기를 지났습니다. 인공지능 역사에서는 1980년대 후반 전문가 시스템에 대한 기대가 무너진 뒤 찾아온 침체기를 제2차 AI 겨울이라고 부릅니다. 신경망 연구자들이 겪은 겨울은 그보다 길었습니다. 2000년대 중반까지 주류 학계는 신경망을 진지한 연구 대상으로 보지 않았습니다. 힌턴과 동

료들은 뒷날 인터뷰에서 논문 제목에 '신경망'이라는 말을 넣으면 학회 심사를 통과하기 어려웠다고 돌아보았습니다.

이 시기에 신경망 연구가 이어질 수 있도록 지원한 곳은 캐나다 고등연구소(CIFAR)였습니다. 2004년 토론토 대학교의 힌턴은 몬트리올 대학교의 요슈아 벤지오, 뉴욕 대학교의 얀 르쿤, 토론토 대학교의 데이비드 플리트 등과 함께 새 연구 프로그램을 제안했습니다. CIFAR는 이 제안을 받아들여 '신경 연산과 적응적 지각(NCAP, Neural Computation and Adaptive Perception)' 프로그램을 만들었고, 힌턴은 약 10년 동안 이 프로그램을 이끌었습니다. 이 프로그램은 뒤에 '기계와 뇌의 학습(Learning in Machines & Brains)'으로 이름을 바꾸었습니다.

NCAP은 몇 쪽 분량의 짧은 연구 제안서만으로 3년에서 5년 단위의 연구비를 지원했습니다. 당장 상업적 성과를 내라고 요구하지도 않았습니다. 힌턴, 벤지오, 르쿤을 중심으로 한 소규모 연구자 모임은 정기적으로 만나 생각을 나누었고, 여름 학교를 열어 젊은 연구자를 길렀습니다. 깊은 신경망을 학습시키는 방법을 찾아낸 2006년의 연구는 이 연구 공동체 안에서 나왔습니다.

6.2. 한 층씩 쌓아 학습하기: 탐욕적 층별 비지도 사전학습과 DBN

2006년 힌턴은 토론토 대학교의 사이먼 오신데로(Simon Osindero), 싱가포르 국립대학교의 예휘 테(Yee-Whye Teh)와 함께 《Neural Computation》 18권 7호에 논문 「A Fast Learning Algorithm for Deep Belief Nets(심층 신뢰 신경망을 위한 빠른 학습 알고리즘)」를 발표했습니다. 논문은 1527쪽부터 1554쪽까지 실렸습니다.

이 논문이 다룬 문제는 '설명 해소(explaining away)' 현상이었습니다. 위층의 원인이 아래층의 결과를 만든다고 보는 신경망을 방향성 신뢰망이라고 합니다. 이런 신경망에서는 결과를 보고 원인을 거꾸로 추론하기가 어렵습니다. 잔디가 젖은 원인으로 비와 스프링클러를 생각해 보겠습니다. 잔디가 젖어 있는 것을 본 상태에서 비가 왔다는 사실을 알게 되면, 스프링클러가 켜졌을 가능성은 낮아집니다. 원래 서로 관계없던 두 원인이 결과를 아는 순간 서로 얹히는 것입니다. 은닉층이 여러 겹이고 유닛이 수천 개인 신경망에서는 이런 얹힘이 너무 복잡해져서 정확한 추론이 거의 불가능했습니다.

힌턴 연구진이 이 문제를 푸는 데 사용한 도구는 제한된 볼츠만 머신(RBM, Restricted Boltzmann Machine)이었습니다. 볼츠만 머신은 힌턴이 1980년대에 테리 세이노프스키와 함께 만든 확률 신경망으로, 이름은 19세기 물리학자 루트비히 볼츠만의 통계 물리 공식에서 따왔습니다. 확률 신경망은 각 유닛이 정해진 값을 내는 대신 확률에 따라 켜지거나 꺼지는 신경망입니다. RBM은 이 모델의 연결을 줄여 간단하게 만든 형태

입니다. 데이터를 받는 가시층과 데이터의 특징을 담는 은닉층의 두 층으로 이루어지며, 같은 층 안의 유닛끼리는 연결하지 않고 두 층 사이만 연결합니다. 이 제한 덕분에 한 층의 값이 정해지면 다른 층의 유닛들을 서로 독립적으로 한꺼번에 계산할 수 있어 계산이 쉬워집니다. 힌턴은 2002년대 조적 발산(CD, contrastive divergence)이라는 방법을 발표해 RBM을 빠르게 학습시킬 수 있게 해 두었습니다.

연구진은 설명 해소 문제를 수학적으로 우회했습니다. 층이 끝없이 이어지는 방향성 신뢰망에서 각 층의 가중치를 바로 위층 가중치의 전치(transpose)로 묶어 두면, 위층이 만들어 내는 사전확률이 아래층에서 생기는 얽힘을 정확히 상쇄합니다. 연구진은 이를 '상보적 사전확률(complementary priors)'이라고 불렀습니다. 그리고 이렇게 가중치를 묶은 무한히 깊은 방향성 모델이 두 층짜리 RBM 하나와 수학적으로 같다는 사실을 보였습니다. 추론이 어려운 깊은 방향성 모델을, 학습 방법이 이미 마련된 RBM으로 바꾸어 다룰 수 있게 된 것입니다. 이 등가성에서 탐욕적 층별 비지도 사전학습(greedy layer-wise unsupervised pre-training)이 나왔습니다.

학습 절차는 다음과 같습니다. 첫 단계에서는 정답이 없는 입력 데이터를 가시층에 넣고, 대조적 발산으로 맨 아래 RBM의 가중치를 학습시킵니다. 둘째 단계에서는 이 가중치를 고정한 뒤, 입력 데이터를 통과시켜 은닉 유닛들이 켜질 확률을 구합니다. 이 은닉층의 활동이 다음 층의 새 입력 데이터가 됩니다. 셋째 단계에서는 이 새 데이터로 두 번째 RBM을 따로 학습시킵니다. 원하는 깊이에 이를 때까지 같은 과정을 한 층씩 되풀이합니다. 앞 단계의 결과를 고정하고 지금 층만 최선으로 학습시킨다는 뜻에서 '탐

욕적'이라는 이름이 붙었습니다. 정답 없이 데이터만으로 학습하므로 '비지도'이고, 본격적인 학습에 앞서 가중치를 미리 마련하므로 '사전학습'입니다.

넷째 단계에서는 이렇게 쌓은 RBM들을 하나의 생성 모델로 합칩니다. 이것이 심층 신뢰 신경망(DBN, Deep Belief Network)입니다. DBN의 맨 위 두 층은 방향 없는 대칭 연결을 그대로 유지하며 일종의 연상 기억 장치 역할을 합니다. 그 아래 층들은 위에서 아래로 향하는 방향성 연결로 바뀌어 데이터를 만들어 내는 역할을 맡습니다.

연구진은 이 절차에 수학적 근거도 붙였습니다. 아래층 가중치를 고정해 채 위에 RBM을 하나 더 쌓아 학습시킬 때마다, 모델이 데이터를 설명하는 정도의 하한이 나빠지지 않고 좋아질 수 있음을 증명한 것입니다. 층을 더 할수록 모델이 나아진다는 수학적 근거를 처음 제시한 결과였습니다.

사전학습을 마친 신경망의 가중치는 무작위 값이 아니라, 이미 데이터의 구조를 어느 정도 담고 있는 좋은 출발점에 놓입니다. 6.1절에서 본 최적화 실패가 무작위 가중치에서 출발하는 데서 비롯되었다는 점을 생각하면, 출발점을 바꾼 것이 이 방법의 핵심입니다. 그 위에서 소량의 라벨 데이터로 전체 가중치를 다듬는 미세 조정(fine-tuning)을 거칩니다. 이 논문에서는 깨어남-잠들(wake-sleep) 알고리즘을 고친 '업다운(up-down)' 알고리즘으로 미세 조정을 했습니다.

연구진은 손글씨 숫자 데이터셋 MNIST로 성능을 시험했습니다. MNIST는 가로세로 28×28픽셀 크기의 숫자 이미지로 이루어지며, 학습용 6만

장과 시험용 1만 장으로 나뉩니다. 은닉층이 500, 500, 2000개 유닛으로 된 DBN은 이미지의 모양 정보나 데이터 늘리기 같은 기법 없이 픽셀값만으로 오차율 1.25%를 기록했습니다. 같은 조건에서 서포트 벡터 머신이 낸 1.4%, 역전파로 학습한 신경망이 낸 약 1.5%보다 낮은 수치였습니다. 이 DBN은 숫자를 분류하는 데 그치지 않고, 학습한 모델로 새로운 숫자 이미지를 그려 낼 수도 있었습니다. 이미지의 공간 정보를 별도로 이용하지 않는 비교 조건에서, 깊은 신경망이 유력한 기존 방법보다 좋은 성능을 낼 수 있음을 보여 주었습니다.

이 결론에 이른 것은 힌턴 연구진만이 아니었습니다. 같은 해 벤지오 연구진은 층별 사전학습이 RBM뿐 아니라 다른 구조에서도 깊은 신경망의 학습을 돕는다는 결과를 NIPS 학회에 발표했습니다. 르쿤 연구진도 비슷한 방향의 연구를 내놓았습니다. CIFAR 공동체의 세 연구 그룹이 거의 같은 시기에 같은 결론에 이른 것입니다.

6.3. 2006년 《Science》 논문: 심층 오토인코더와 비선형 차원 축소

2006년 7월 28일, 힌턴과 박사과정 학생 루슬란 살라쿠트디노프(Ruslan Salakhutdinov)는 《Science》 313권 5786호에 논문 「Reducing the Dimensionality of Data with Neural Networks(신경망을 이용한 데이터의 차원 축소)」를 실었습니다. 논문은 504쪽부터 507쪽까지 네 쪽 분량이었습니다. 같은 달 나온 DBN 논문이 전문 학술지를 통해 연구자들에게 알려졌다면, 이 논문은 과학계 전체가 읽는 학술지를 통해 깊은 신경망 연구가 되살아났음을 알렸습니다.

이 논문이 다룬 것은 차원 축소 문제입니다. 차원 축소는 숫자가 아주 많은 데이터를 적은 수의 숫자로 요약하는 작업입니다. 784개 픽셀로 된 숫자 이미지를 30개 숫자로 줄여도 원래 모양을 되살릴 수 있다면, 그 30개 숫자는 이미지의 핵심을 담고 있다고 볼 수 있습니다. 오랫동안 이 작업에는 1933년 해럴드 호텔링이 정리한 주성분 분석(PCA)이 널리 쓰였습니다. PCA는 데이터가 가장 넓게 퍼진 직선 방향을 찾아 그 방향으로 데이터를 비춰 봅니다.

문제는 실제 데이터가 평평한 판 위에 놓여 있지 않다는 점입니다. 이미지나 문서 같은 데이터는 둥글게 말린 종이처럼 휘어진 곡면, 곧 비선형 매니폴드(nonlinear manifold) 위에 놓여 있습니다. 말린 종이를 위에서 비추면 서로 멀리 떨어진 부분이 겹쳐 보입니다. 마찬가지로 직선 축에 비추는 PCA는 휘어진 구조를 뭉개 버립니다.

오토인코더(autoencoder)는 이 한계를 넘을 수 있는 신경망입니다. 입력을 점점 작은 층으로 압축하는 인코더와, 가운데의 좁은 코드층에서 다시 원래 크기로 되살리는 디코더로 이루어집니다. 입력과 출력이 같아지도록 학습시키면 가운데 코드층에 데이터의 핵심이 남습니다. 이 생각은 1980년대부터 있었습니다. 그러나 실제로 구현하면 6.1절의 문제가 그대로 나타났습니다. 무작위 가중치에서 역전파로 깊은 오토인코더를 학습시키면 학습 신호가 가운데 코드층까지 제대로 닿지 않았고, 결과는 PCA와 비슷하거나 그보다 못했습니다.

힌턴과 살라쿠트디노프는 층별 RBM 사전학습으로 이 문제를 풀었습니다. 방법은 세 단계입니다. 첫째, RBM 여러 개를 아래에서부터 차례로 학습시킵니다. MNIST 실험에서는 784개 입력을 1000, 500, 250개를 거쳐 30개로 줄이는 구조를 썼습니다. 둘째, 학습한 RBM 가중치를 쌓아 인코더를 만들고, 같은 가중치를 전치해 거꾸로 쌓아 디코더를 만듭니다. 이렇게 오토인코더를 '펼칩니다(unrolling)'. 셋째, 확률적으로 0이나 1을 내던 유닛을 연속값을 내는 시그모이드 유닛으로 바꾸고, 원래 입력과 복원한 출력의 차이를 줄이도록 역전파로 전체 신경망을 미세 조정합니다. 사전학습으로 좋은 출발점을 마련한 다음에는, 깊은 신경망에서도 역전파가 제 역할을 한 것입니다.

결과는 그림으로 비교해도 뚜렷했습니다. 숫자 이미지를 30개 숫자로 압축했다가 되살렸을 때, 심층 오토인코더가 만든 이미지는 원본과 거의 같았습니다. 같은 크기로 압축한 PCA의 복원 이미지는 흐릿하게 번졌습니다. 숫자 이미지를 2차원 평면에 펼쳐 놓았을 때도 차이가 컸습니다. PCA 그림에서는 0부터 9까지의 숫자가 뒤섞여 겹쳤지만, 오토인코더 그림에

서는 숫자마다 뚜렷한 덩어리로 나뉘었습니다. 사람 얼굴 사진에서도 오토인코더는 PCA보다 작은 복원 오차를 보였습니다.

연구진은 이 방법을 문서에도 적용했습니다. 뉴스 기사 80만여 건을 자주 쓰이는 단어 2000개의 빈도로 나타낸 뒤 압축하자, 비슷한 주제의 기사끼리 모였습니다. 기존의 선형 방법인 잠재 의미 분석(LSA)보다 나은 결과였습니다. 이 연구는 2009년 살라쿠티노프와 힌턴의 '의미 해싱 (semantic hashing)' 논문으로 이어졌습니다. 문서를 짧은 이진 코드로 바꾸어 두면, 코드가 비슷한 문서를 찾는 방식으로 뜻이 가까운 문서를 아주 빠르게 검색할 수 있습니다.

논문 끝부분에서 두 연구자는 이 성과가 어떤 조건에서 가능했는지를 정리했습니다. 깊은 오토인코더로 비선형 차원 축소를 할 수 있다는 생각은 1980년대부터 있었지만, 그것이 실현되려면 빠른 컴퓨터와 큰 데이터, 그리고 좋은 초기 가중치가 필요했다는 것입니다. 2006년에는 세 조건이 모두 갖추어졌고, 그중 좋은 초기 가중치를 구하는 방법이 바로 층별 사전학습이었습니다.

6.4. 심층 볼츠만 머신(DBM)과 딥러닝의 확산

2009년 살라쿠트디노프와 힌턴은 미국 플로리다에서 열린 제12회 인공지능과 통계 국제학회(AISTATS)에서 심층 볼츠만 머신(DBM, Deep Boltzmann Machine)을 발표했습니다. DBN은 맨 위 두 층만 방향 없이 연결되고 아래층들은 위에서 아래로만 연결된 혼합형 모델이었습니다. 이와 달리 DBM은 가시층과 모든 은닉층 사이를 방향 없는 대칭 연결로 이었습니다.

이 차이는 추론 방식에서 나타납니다. DBM의 가운데 은닉층은 아래층에서 올라오는 감각 정보와 위층에서 내려오는 문맥 정보를 함께 받아 자신의 상태를 정합니다. 사람이 흐릿한 글씨를 읽을 때 앞뒤 문맥을 보고 글자를 짐작하는 것과 비슷합니다. 그래서 DBM은 일부가 가려지거나 잡음이 섞인 입력도 비교적 잘 처리했습니다.

대신 DBM은 학습이 까다로웠습니다. 연구진은 층별 RBM 사전학습을 하되, 위아래 양쪽에서 신호를 받는 구조에 맞게 가중치를 두 배로 늘리거나 절반으로 줄이는 방식으로 고쳤습니다. 그 뒤 평균장(mean-field) 근사로 추론하고, 지속적 대조적 발산(persistent CD)이라는 표본 추출 방법을 함께 써서 전체 가중치를 조정했습니다. 이렇게 학습한 DBM은 MNIST에서 오차율 0.95%를 기록했고, 3차원 장난감 물체를 여러 각도에서 찍은 NORB 데이터에서도 오차율 10.8%를 기록해, 비교 대상인 서포트 벡터 머신의 11.6%보다 낮은 오차를 냈습니다.

이 무렵 사전학습의 효과는 연구실 밖으로 퍼지기 시작했습니다. 첫 분야는 음성 인식이었습니다. 토론토 대학교의 압델라만 모하메드(Abdel-rahman Mohamed)와 조지 달(George Dahl)은 힌턴과 함께, DBN으로 사전학습한 깊은 신경망을 음성의 소리 단위를 판별하는 음향 모델에 적용했습니다. 수십 년 동안 음성 인식 분야는 가우스 혼합 모델(GMM)과 은닉 마르코프 모델(HMM)을 결합한 방식을 표준으로 써 왔습니다. 깊은 신경망은 TIMIT 음소 인식 과제에서 이 방식을 앞섰고, 곧이어 대규모 어휘 음성 인식에서도 좋은 결과를 냈습니다. 2012년 토론토 대학교, 마이크로소프트, 구글, IBM의 연구자들은 《IEEE Signal Processing Magazine》에 공동 논문을 내고 네 연구 그룹이 같은 결론에 이르렀다고 밝혔습니다. 이후 주요 기업들은 음성 인식 엔진의 핵심을 깊은 신경망으로 바꾸었습니다.

같은 2012년, 힌턴의 학생 알렉스 크리제브스키와 일리아 수츠케버는 힌턴과 함께 만든 신경망 AlexNet으로 이미지넷(ImageNet) 대회에서 상위 5개 오답률 15.3%를 기록해 2위와 큰 차이로 우승했습니다. 컴퓨터 비전 분야도 깊은 신경망으로 빠르게 넘어갔습니다.

그런데 깊은 신경망이 널리 쓰이기 시작한 바로 이 무렵부터, 층별 사전학습은 주류에서 물러나기 시작했습니다. 6.1절의 세 가지 문제를 다른 방식으로 해결하는 수단이 생겼기 때문입니다. 그래픽 처리 장치(GPU)로 계산 속도가 크게 빨라졌고, 수백만 장의 라벨 데이터가 생겼습니다. 기울기 소멸을 줄이는 ReLU 활성화 함수, 학습 중 일부 유닛을 무작위로 끄는 드롭아웃(dropout), 초기 가중치를 정하는 개선된 방법도 나왔습니다. 그

러자 사전학습 없이 처음부터 역전파로 깊은 신경망을 학습시킬 수 있게 되었습니다. AlexNet도 사전학습 없이 학습한 신경망이었습니다.

그렇다고 2006년의 성과가 의미를 잃은 것은 아닙니다. 2015년 르쿤, 벤지오, 힌턴은 《네이처》 521권에 실은 리뷰 논문 「Deep Learning」에서 이 시기를 정리했습니다. 세 사람은 2006년 무렵 CIFAR가 모은 연구자들이 깊은 신경망에 대한 관심을 되살렸다고 썼습니다. 라벨 없는 데이터로 특징을 한 층씩 학습시켜 좋은 초기 가중치를 만든 뒤 역전파로 다듬는 방법이, 라벨 데이터가 적은 상황에서도 깊은 신경망이 잘 작동한다는 것을 보여 주었다는 평가였습니다. 깊은 신경망은 학습시킬 수 없다는 오랜 통념을 무너뜨린 것이 이 연구의 공헌이었습니다.

정답 없는 대량의 데이터로 먼저 학습한 뒤 적은 라벨 데이터로 다듬는다는 발상은 오늘날의 거대 언어 모델에서도 이어집니다. 지금의 모델은 RBM을 쌓지 않습니다. 그러나 방대한 문장으로 먼저 사전학습하고 특정 작업에 맞게 미세 조정하는 큰 틀은 2006년의 생각과 닮았습니다.

힌턴, 벤지오, 르쿤은 2018년 튜링상을 함께 받았습니다. 튜링상은 컴퓨터 과학 분야의 노벨상으로 불리는 상입니다. 그리고 2024년 10월 8일, 스웨덴 왕립과학원은 힌턴과 존 홉필드에게 인공신경망 기계학습을 가능하게 한 기초적 발견과 발명의 공로로 노벨 물리학상을 수여한다고 발표했습니다. 위원회는 힌턴이 홉필드 네트워크를 바탕으로 볼츠만 머신을 만들었고, 이 연구를 발전시켜 오늘날 기계학습의 급격한 발전을 여는 데 기여했다고 밝혔습니다. 2006년의 DBN은 바로 그 볼츠만 머신에서 발전

한 모델이었습니다. 신경망 연구가 침체한 동안에도 계속된 볼츠만 머신 연구가 깊은 신경망을 학습시키는 방법으로 이어진 것입니다.

제3부. 딥러닝이 세상으로 나오다

제7장. 음성과 사진을 알아보는 컴퓨터

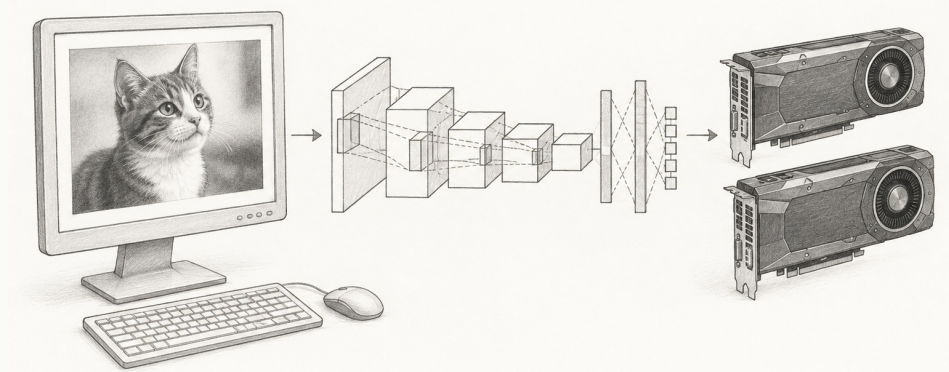


그림 7. 이미지 인식과 GPU의 역할을 표현했습니다.

2006년 제프리 힌턴은 심층 신뢰 신경망(DBN)으로 깊은 신경망을 학습시킬 길을 열었습니다. 그러나 인공지능 학계 대부분은 여전히 신경망을 신뢰하지 않았습니다. 논문 몇 편에 실린 좋은 결과만으로는 오랫동안 굳어 온 통념을 바꾸기 어려웠습니다. 학계의 판단을 바꾼 것은 실제 문제에서 나온 숫자였습니다. 2009년부터 2012년까지 힌턴의 연구실은 그런 숫자를 두 번 내놓았습니다. 첫 번째는 사람의 말을 알아듣는 음성 인식에서, 두 번째는 사진 속 물체를 알아보는 컴퓨터 비전에서 나왔습니다.

7.1. 첫 번째 돌파구: 음성 인식의 변화 (2009~2012)

컴퓨터가 사람의 말을 글자로 바꾸는 기술을 자동 음성 인식이라고 합니다. 이 기술은 크게 두 부분으로 나뉩니다. 하나는 아주 짧은 소리 조각이 어떤 발음인지 판단하는 음향 모델이고, 다른 하나는 그렇게 알아낸 발음을 단어와 문장으로 이어 붙이는 부분입니다. 이 가운데 음향 모델은 2000년대 후반까지 30년 가까이 같은 방식에 머물러 있었습니다. 가우시안 혼합 모델(GMM)로 소리의 특징을 확률로 계산하고, 숨은 마르코프 모델(HMM)로 시간에 따른 변화를 따라가는 방식이었습니다.

이 방식은 오랫동안 좋은 성적을 냈지만 약점도 분명했습니다. GMM은 소리 한 조각을 여러 '전문가' 가운데 한 명이 맡아 설명한다고 보는 구조입니다. 그런데 사람의 말소리는 그렇게 깔끔하게 나뉘지 않습니다. 앞뒤 발음이 서로 섞이고, 한순간에도 여러 소리의 성질이 겹칩니다. 이런 겹침을 표현하려면 GMM은 구성 요소를 계속 늘려야 했고, 그만큼 계산량과 데이터가 많이 필요했습니다. 앞뒤로 이어진 여러 순간의 소리를 한꺼번에 살펴피기도 어려웠습니다.

이 한계를 처음 넘어선 사람은 토론토 대학교 힌턴 연구실의 박사과정 학생 조지 달과 압델라만 모하메드였습니다. 2009년 두 사람은 힌턴이 2006년에 내놓은 DBN 방법을 음향 모델에 적용했습니다. 먼저 제한된 볼츠만 머신(RBM)을 한 층씩 쌓아 정답 없이 소리의 구조를 익히게 하고, 그다음 오차 역전파로 신경망 전체를 다듬었습니다. 입력으로는 소리 한 조각만 넣지 않고 앞뒤로 이어진 11~15개 조각을 한 묶음으로 넣었습니

다. GMM이 다루기 어려웠던 앞뒤 맥락을 신경망은 그대로 받아들일 수 있었습니다.

이 실험을 가능하게 한 것은 그래픽 카드, 곧 GPU였습니다. GPU는 원래 게임 화면을 그리려고 만든 장치인데, 같은 종류의 계산을 한꺼번에 많이 처리하는 데 강합니다. 두 사람은 엔비디아의 테슬라 GPU와 연구실 동료 블라디미르 므니가 만든 계산 도구 CUDAMat을 써서 학습 속도를 CPU 하나로 할 때보다 20배 넘게 높였습니다. 몇 주씩 걸리던 실험이 몇 시간 만에 끝나자, 설정을 여러 번 바꿔 가며 시험해 보는 일이 가능해졌습니다.

결과는 음성 인식 학계의 주목을 받았습니다. 학자들이 오래 써 온 표준 시험 자료인 TIMIT에서 이들의 신경망은 음소 오류율 20.7%를 기록했습니다. 음소는 말소리를 이루는 가장 작은 소리 단위이므로, 음소 100개 가운데 21개쯤을 틀렸다는 뜻입니다. 이 성적은 그때까지 보고된 기존 방식의 최고 기록을 넘어섰습니다. 첫 결과는 2009년 NIPS 워크숍에서 발표되었고, 다듬은 결과는 2012년 IEEE 학술지에 실렸습니다.

다만 TIMIT는 규모가 작은 시험 자료였습니다. 이 방법이 실제 제품에서도 통하는지 확인하려면 큰 회사가 가진 데이터와 시스템이 필요했습니다. 그래서 힌턴의 학생들은 직접 회사로 갔습니다. 조지 달은 마이크로소프트 연구소(MSR)로, 압델라만 모하메드는 IBM 연구소로 인턴ships을 떠났습니다. 또 다른 제자 내브딕 자이틀리는 구글로 갔습니다.

마이크로소프트 연구소에서는 덩리(Li Deng)와 위둥(Dong Yu)이 힌턴 연구진과 협력했습니다. 이들은 휴대전화로 녹음해 잡음이 많은 빙

(Bing) 모바일 음성 검색 데이터에 심층 신경망을 적용했고, 문장 단위 오류율은 36.2%에서 30.4%로 내려갔습니다. 이어 300시간 분량의 전화대화 자료인 스위치보드(Switchboard)에서는 단어 오류율이 27.4%에서 18.5%로 떨어졌습니다. 오류를 약 3분의 1 줄인 셈입니다. 이 성적은 2,000시간이 넘는 데이터와 복잡한 보정 기법을 모두 동원한 기존 시스템의 18.6%와 맞먹었습니다. 훨씬 적은 데이터로, 신경망 하나만으로 같은 수준에 오른 것입니다.

구글에서는 자이틀리가 빈센트 반후크, 패트릭 응우옌 등과 함께 5,870시간 분량의 음성 입력 데이터로 실험했습니다. 단어 오류율은 16.0%에서 12.3%로 내려갔습니다. 구글은 이 기술을 다듬어 2012년 안드로이드 스마트폰의 음성 검색에 넣었습니다. 대학 연구실에서 시작한 신경망이 많은 사람이 날마다 쓰는 제품 속으로 들어간 것입니다.

여러 회사에서 따로 확인된 이 결과들은 한 편의 논문으로 정리되었습니다. 2012년 11월 토론토 대학교, 마이크로소프트, 구글, IBM의 네 연구 그룹은 『IEEE 시그널 프로세싱 매거진』에 「음성 인식의 음향 모델링을 위한 심층 신경망」이라는 공동 논문을 실었습니다. 서로 경쟁하는 회사의 연구자들이 한 논문에 이름을 나란히 올린 것은 드문 일이었습니다. 이 논문은 음성 인식의 표준이 바뀌었음을 학계에 알렸고, 이는 딥러닝이 산업 현장에서 거둔 주요 성과 가운데 하나였습니다. 다만 이 변화는 주로 음성 인식 전문가들 사이에서 일어났습니다. 인공지능 학계 전체가 이를 실감하기까지는 시간이 조금 더 필요했습니다.

7.2. 컴퓨터 비전의 전환점: 페이페이 리의 ImageNet과 AlexNet (2012)

2000년대 후반 컴퓨터 비전 학계는 사진에서 쓸 만한 특징을 사람이 직접 설계하는 방식을 따랐습니다. SIFT와 HOG가 대표적입니다. 사진 속 모서리의 방향이나 밝기 변화를 계산해 특징으로 뽑고, 서포트 벡터 머신(SVM) 같은 분류기로 물체의 종류를 판단했습니다. 이 방식은 조명이 바뀌거나, 물체가 다른 각도로 찍히거나, 일부가 가려지면 성능이 쉽게 흔들렸습니다. 성능은 해마다 오르는 했지만 오르는 폭이 작았습니다.

이 무렵 상황을 바꿀 재료를 준비한 사람이 페이페이 리였습니다. 리는 2007년 프린스턴 대학교에서 ImageNet 프로젝트를 시작했고, 2009년 스탠퍼드 대학교로 옮겨 연구를 이어 갔습니다. 리의 연구진은 인터넷에서 모은 사진에 사람이 직접 이름표를 붙였습니다. 필요한 인력은 아마존 메커니컬 터크라는 온라인 일감 중개 서비스로 구했습니다. 그렇게 22,000개가 넘는 범주에 1,500만 장이 넘는 사진을 담은 데이터베이스가 만들어졌습니다. 그때까지 연구자들이 쓰던 사진 자료와는 규모부터 달랐습니다.

2010년부터는 이 가운데 1,000개 범주, 학습용 사진 약 120만 장을 골라 ILSVRC라는 대회를 해마다 열었습니다. 성적은 주로 '상위 5개 오류율'로 평가했습니다. 컴퓨터가 사진마다 정답 후보 5개를 내놓고, 그 안에 정답이 없으면 틀린 것으로 치는 방식입니다. 2010년 우승 팀의 오류율은 28.2%, 2011년 우승 팀은 25.8%였습니다. 사람이 손으로 만든 특징을 쓰는 방식은 이 언저리에서 더 나아가지 못하고 있었습니다.

2012년 힌턴은 두 제자 알렉스 크리제브스키, 일리아 수츠케버와 함께 이 대회에 나갔습니다. 팀 이름은 슈퍼비전(SuperVision)이었습니다. 세 사람의 역할은 서로 달랐습니다. 코드를 짜고 신경망을 학습시킨 사람은 크리제브스키였고, 큰 사진 데이터에 신경망을 써 보자고 강하게 권한 사람은 수츠케버였습니다. 힌턴은 지도교수로서 방향을 잡았습니다. 크리제브스키는 엔비디아 그래픽 카드 두 장을 꽂은 컴퓨터로 이 작업을 했고, 신경망에는 훗날 그의 이름을 딴 AlexNet이라는 이름이 붙었습니다.

AlexNet의 뿌리는 얀 르쿤이 1980년대 말 손글씨 숫자를 읽으려고 만든 합성곱 신경망(CNN)입니다. 합성곱 신경망은 사진의 작은 영역을 차례로 훑으며 무늬를 찾고, 층을 거칠수록 더 복잡한 모양을 알아봅니다. AlexNet은 이 구조를 크고 복잡한 컬러 사진에 맞게 키웠습니다. 합성곱 층 5개와 모든 뉴런이 서로 이어진 층 3개를 합쳐 모두 8개 층으로 이루어졌습니다. 학습으로 조정하는 값인 파라미터는 약 6,000만 개, 뉴런은 약 65만 개였습니다. 당시로서는 보기 드문 크기였습니다.

그러나 크기만 키운 것은 아니었습니다. 이만한 신경망을 제대로 학습시키기 위해 연구진은 네 가지 방법을 함께 썼습니다. 첫째는 활성화 함수로 ReLU를 쓴 것입니다. 활성화 함수는 뉴런이 신호를 얼마나 내보낼지 정하는 규칙입니다. 그전까지 많이 쓰던 함수는 입력이 커지면 기울기가 0에 가까워져 학습이 느려졌습니다. ReLU는 0보다 작은 값은 0으로 바꾸고 0보다 큰 값은 그대로 내보내는 간단한 함수입니다. 실험에서 ReLU를 쓴 신경망은 기존 함수를 쓴 신경망보다 6배 빨리 학습했습니다.

둘째는 GPU 두 대에 신경망을 나누어 올린 것입니다. 크리제브스키가 쓴 엔비디아 GTX 580은 메모리가 3GB뿐이어서 신경망 전체를 한 대에 담을 수 없었습니다. 그는 GPU용 계산 코드 `cuda-convnet`을 직접 짜서 신경망을 반으로 나눠 두 GPU에 올리고, 일부 층에서만 두 GPU가 정보를 주고받게 했습니다. 120만 장의 사진을 90번 가까이 되풀이해 학습하는데 5~6일이 걸렸습니다.

셋째는 드롭아웃입니다. 파라미터가 6,000만 개나 되면 신경망이 학습 사진을 통째로 외워 버려, 처음 보는 사진에서는 성적이 떨어지기 쉽습니다. 이를 과적합이라고 합니다. 드롭아웃은 학습할 때마다 뉴런의 절반을 무작위로 꺼 두는 방법입니다. 그러면 매번 조금씩 다른 신경망을 학습시키는 효과가 생겨, 신경망이 특정 뉴런에 지나치게 의존하지 않게 됩니다. AlexNet은 뒤쪽의 두 층에 이 방법을 썼습니다.

넷째는 학습 데이터를 늘린 것입니다. 256×256 크기의 사진에서 224×224 크기의 부분을 무작위로 잘라 내고, 좌우를 뒤집고, 색감을 조금씩 바꾸었습니다. 사진 한 장에서 여러 장의 새 학습 자료를 만든 셈입니다.

이렇게 학습한 신경망이 사진에서 무엇을 익혔는지는 첫 층에서 확인할 수 있습니다. AlexNet의 첫 층이 스스로 익힌 필터 96개를 그림으로 그려 보면 여러 방향의 선과 모서리, 색 덩어리가 나타납니다. 누구도 이런 무늬를 찾으라고 가르치지 않았습니니다. 신경망이 사진을 보며 스스로 만든 것입니다. 이 무늬는 사람 뇌의 일차 시각 피질 세포가 반응하는 무늬와 닮아 있었습니다.

대회 결과는 2012년 10월 이탈리아 피렌체에서 열린 유럽 컴퓨터 비전 학회(ECCV) 워크숍에서 발표되었습니다. AlexNet의 상위 5개 오류율은 15.3%였습니다. 2위를 한 도쿄 대학교 팀은 SIFT와 피셔 벡터 같은 기존 방법으로 26.2%를 기록했습니다. 두 팀의 차이는 10.9%포인트였습니다. 1~2%포인트 차이로 순위가 갈리던 대회에서 2위 팀보다 오류를 40% 넘게 줄인 것입니다.

7.3. 딥러닝으로의 전환, DNNresearch 창업과 구글 인수

피렌체 발표장의 반응은 엇갈렸습니다. 오래 컴퓨터 비전을 연구한 학자들은 이 방법이 쓸모 있는지 확신하지 못했습니다. 반면 그 자리에 있던 안르쿤은 이를 인공지능의 전환점이라고 평가했습니다. 르쿤은 1980년대 말 토론토에서 힌턴의 박사후 연구원으로 일한 적이 있었습니다.

이후의 전개는 르쿤의 판단대로였습니다. AlexNet 이전에는 주요 컴퓨터 비전 논문 가운데 신경망을 쓴 것이 거의 없었고, 신경망 논문을 내면 심사에서 떨어지는 일이 흔했습니다. 그러나 AlexNet 이후 1~2년 사이에 연구자들은 손으로 특징을 설계하던 방식을 내려놓고 딥러닝으로 옮겨 갔습니다. 다음 해 대회부터는 상위권 팀 대부분이 합성곱 신경망을 썼습니다. 30년 넘게 변방에 머물던 연결주의가 쌓아 온 힘이 한꺼번에 터져 나온 이 장면은 흔히 억눌려 있던 독이 무너진 순간에 비유됩니다. 페이페어리는 훗날 많은 사람이 이때를 현대 인공지능이 태어난 순간으로 본다고 말했습니다.

기업들도 움직였습니다. 음성 인식에서 이미 신경망의 성능을 확인한 빅테크 회사들은 AlexNet을 보고 힌턴과 두 제자를 영입하려 했습니다. 힌턴은 회사를 세워 매각하기로 했습니다. 그는 학습 장애가 있는 아들의 앞날을 위해 큰돈이 필요했다고 훗날 여러 인터뷰에서 밝혔습니다. 2012년 말 힌턴은 크리제브스키, 수츠케버와 함께 DNNresearch라는 회사를 만들었습니다. 이름은 심층 신경망(Deep Neural Network)에서 따왔습니다. 제품도 매출도 없고 직원은 세 사람뿐이었습니다. 회사가 가진 것은 AlexNet을 만든 기술과 세 연구자의 능력이 전부였습니다.

2012년 12월 초 NIPS 학회가 열린 미국 레이크 타호에서 이 회사를 두고 경매가 열렸습니다. 입찰에는 바이두, 구글, 마이크로소프트, 그리고 세워진 지 2년 된 스타트업 딥마인드가 참여했습니다. 경매는 이메일로 진행되었고, 힌턴은 호텔 방에서 경매를 이끌며 입찰자들이 서로 누가 참여했는지 모르게 했습니다. 시작가는 1,200만 달러였고, 값을 올릴 때는 한 번에 최소 100만 달러씩 올리도록 정했습니다.

딥마인드는 현금 대신 자기 회사 주식으로 입찰했다가 가장 먼저 물러났고, 마이크로소프트는 2,200만 달러 언저리에서 빠졌습니다. 마지막까지 남은 곳은 구글과 바이두였습니다. 값이 4,400만 달러에 이르자 힌턴은 입찰을 멈추고 회사를 구글에 넘기기로 했습니다. 그에게는 가장 높은 값을 받는 것보다 연구를 이어 갈 알맞은 곳을 찾는 일이 중요했습니다. 인수는 2013년 3월 공식 발표되었습니다.

인수 뒤 세 사람은 각자 다른 길을 걸었습니다. 힌턴은 토론토 대학교 교수직을 유지한 채 구글에 합류했고, 뒤에 부사장 겸 엔지니어링 펠로로서 구글에서 신경망 연구에 참여했습니다. 수츠케버는 구글 브레인에서 2014년 한 문장을 다른 문장으로 바꾸는 시퀀스 투 시퀀스(Seq2Seq) 모델을 내놓아 기계 번역의 발전에 기여했습니다. 그는 2015년 말 구글을 떠나 샘 올트먼 등과 함께 OpenAI를 세웠고, 수석 과학자로서 훗날 ChatGPT로 이어지는 연구를 이끌었습니다. 크리제브스키는 구글에서 몇 년 더 일한 뒤 2017년 회사를 떠났습니다.

이 경매는 업계 전체의 계산을 바꾸었습니다. 세 사람짜리 회사에 4,400만 달러의 값이 매겨지자, 딥러닝 연구자 한 사람 한 사람의 몸값이 새로

정해졌습니다. 구글은 2014년 1월, 경매에서 경쟁했던 딥마인드를 약 6억 달러에 인수했습니다. 경매에서 진 바이두와 마이크로소프트도 딥러닝 연구에 본격적으로 뛰어들었습니다.

AlexNet의 원본 코드는 DNNresearch 인수와 함께 구글 소유가 되었습니다. 이 코드가 일반에 공개된 것은 10여 년 뒤의 일입니다. 컴퓨터 역사 박물관(CHM)의 큐레이터 한센 쉬가 2020년 크리스제브스키에게 코드 공개를 제안했고, 구글과 5년 동안 협의한 끝에 원본 코드를 확보했습니다. 박물관은 2025년 3월 20일 구글과 함께 AlexNet 원본 코드를 일반에 공개하고 영구 보존한다고 발표했습니다. 공개된 코드는 엔비디아 CUDA 코드, 파이썬, C++를 합쳐 약 200KB 분량입니다. 오늘날 인공지능 모델과 견주면 아주 작은 규모이지만, 이 작은 코드가 한 시대의 방향을 바꾸었습니다.

2009년의 음성 인식과 2012년의 AlexNet은 같은 연구실에서 나왔습니다. 두 사건 모두 힌턴이 수십 년 동안 붙들어 온 생각, 곧 사람이 규칙을 정해 주지 않아도 신경망이 데이터에서 스스로 특징을 배울 수 있다는 생각을 실제 숫자로 증명했습니다. 대규모 데이터와 GPU라는 두 조건이 갖추어지자, 오랫동안 쌓여 온 연구가 한꺼번에 결실을 맺었습니다. 비주류 학문이던 연결주의는 이 4년을 거치며 인공지능의 주류가 되었습니다.

제8장. 과적합을 줄이고 지식을 전달하다

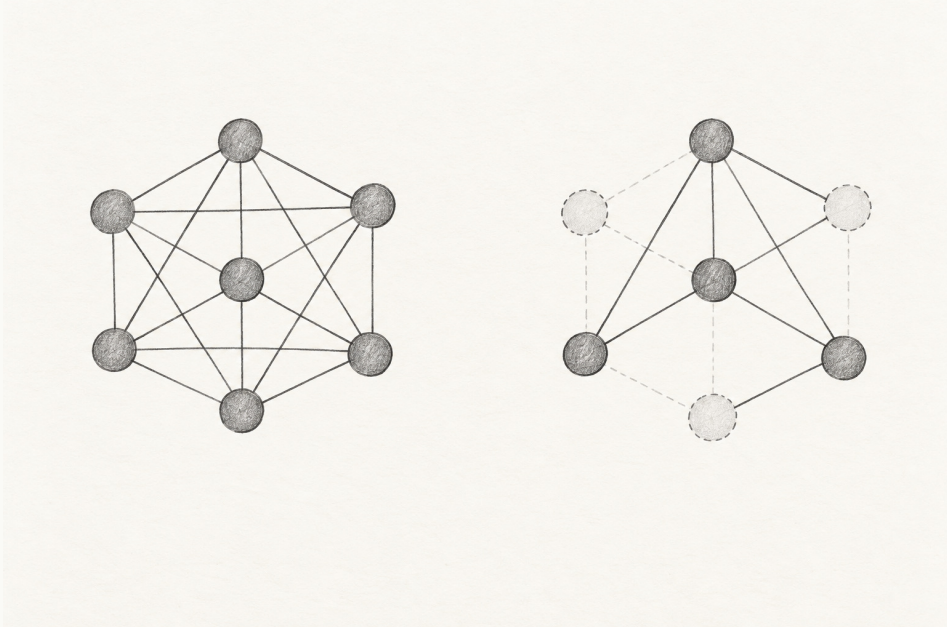


그림 8. 일부 뉴런을 잠시 끄며 학습하는 드롭아웃을 표현했습니다.

8.1. 늘어난 파라미터와 과적합 문제: 드롭아웃의 탄생

2010년대 초, 심층 신뢰 신경망(DBN)과 심층 오토인코더, AlexNet이 잇달아 성과를 내면서 인공지능망은 층이 깊어지고 파라미터 수가 크게 늘었습니다. 파라미터는 신경망이 학습하면서 스스로 고쳐 나가는 숫자들입니다. 이 숫자가 수천만 개, 수억 개로 늘어나자 연구자들은 오래된 문제와 다시 마주했습니다. 과적합(Overfitting)입니다.

과적합은 시험 범위의 문제와 답을 통째로 외운 학생의 경우와 비슷합니다. 그 학생은 연습 문제에서는 만점을 받지만 처음 보는 문제가 나오면 손을 대지 못합니다. 신경망도 학습 데이터가 한정된 상태에서 같은 데이터를 수천 번 반복해 학습하면, 데이터 속의 규칙 대신 표본마다 섞여 있는 우연한 잡음까지 외워 버립니다. 그 결과 학습 데이터에 대한 오차는 0에 가까워지지만 새로운 평가 데이터에서는 성능이 떨어집니다. 이를 일반화(Generalization) 실패라고 부릅니다.

통계학과 기계학습 분야는 이 문제를 막기 위해 가중치 감쇠(L1·L2 Weight decay), 조기 종료(Early stopping), 가중치 공유 같은 정규화(Regularization) 기법을 써 왔습니다. 그러나 비선형 은닉층을 여러 겹 쌓은 심층 신경망은 파라미터 공간이 워낙 넓어서, 이런 방법만으로 과적합을 억제하기는 어려웠습니다.

한 가지 해법은 앙상블(Ensemble)이었습니다. 앙상블은 구조나 초깃값이 서로 다른 모델을 여러 개 따로 학습시킨 뒤 그 예측을 평균 내는 방법입니다. 여러 사람에게 물어 답을 모으면 한 사람의 실수가 묻히는 것과 같

은 이치입니다. 문제는 비용이었습니다. 거대 신경망은 하나를 학습시키는 데만 며칠에서 몇 주가 걸렸습니다. 그런 신경망을 수십 개 따로 학습시키고 실제 서비스에서 그 모델들을 동시에 실행하는 일은 계산량과 응답 지연 면에서 현실성이 없었습니다.

제프리 힌턴은 이 한계를 넘을 방법으로 드롭아웃(Dropout)을 내놓았습니다. 토론토 대학교의 니티시 슈리바스타바(Nitish Srivastava), 알렉스 크리제브스키(Alex Krizhevsky), 일리야 수츠케버(Ilya Sutskever), 루슬란 살라쿠트디노프(Ruslan Salakhutdinov)가 함께한 연구였습니다. 연구진은 2012년 7월 arXiv에 「특징 검출기들의 공적응을 막아 신경망을 개선하기」를 먼저 공개했고, 2014년에 『기계학습 연구 저널(JMLR)』 제15권에 「드롭아웃: 신경망의 과적합을 막는 간단한 방법」을 정식 논문으로 실었습니다. 핵심 생각은 신경망 하나를 학습시키는 동안 그 안에서 크기가 줄어든 서로 다른 신경망을 수없이 많이 함께 학습시키고, 이들이 가중치를 나누어 쓰게 하는 것이었습니다. 앙상블의 효과를 얻으면서 앙상블의 비용은 치르지 않으려는 시도였습니다.

첫 논문의 제목에 들어 있는 공적응(Co-adaptation)이 연구진이 겨냥한 문제였습니다. 일반적인 오차 역전파 학습에서는 은닉 뉴런들이 서로의 실수를 메워 주는 복잡한 관계를 맺기 쉽습니다. 조별 과제에서 한 사람이 빠뜨린 부분을 다른 사람이 늘 채워 주다 보면, 조원이 바뀌는 순간 과제 전체가 무너지는 것과 같습니다. 공적응이 심한 신경망은 학습 데이터에서는 잘 맞지만 조금만 다른 데이터가 들어와도 예측이 크게 흔들립니다.

힌턴은 이 문제를 푸는 생각의 출발점을 일상 경험과 생물학에서 찾았습니다. 그는 한 온라인 문답에서, 은행에 갈 때마다 창구 직원이 바뀌는 것을 보고 직원들이 서로 짜고 부정을 저지르지 못하게 하려는 장치라고 짐작했다고 말했습니다. 신경망의 뉴런도 매번 무작위로 일부를 빼 버리면 뉴런끼리 짜고 움직이는 일을 막을 수 있다는 발상이었습니다.

진화생물학의 유성생식 이론은 이 발상에 이론적 근거를 주었습니다. 무성생식은 부모의 유전자 조합을 거의 그대로 자손에게 넘깁니다. 이미 환경에 잘 맞춰진 유전자 묶음을 잃지 않으니 겉보기에는 유리합니다. 반면 유성생식은 부모에게서 유전자를 절반씩 무작위로 가져와 섞기 때문에, 서로 맞물려 잘 작동하던 유전자 조합이 세대마다 흩어집니다. 그런데도 고등 생물 대부분은 유성생식을 택했습니다. 생물학자 아디 립낫(Adi Livnat) 등은 2010년 논문에서 그 이유를, 자연선택이 개체 하나의 적합도보다 유전자의 혼합 능력(Mix-ability)을 고르기 때문이라고 설명했습니다. 늘 함께 다니는 특정 유전자 집단에 기대지 않고 어떤 유전자와 섞여도 제 몫을 하는 유전자가 살아남는다는 뜻입니다.

힌턴은 이 원리를 은닉층 학습에 그대로 옮겼습니다. 드롭아웃은 학습할 때마다 뉴런 일부를 무작위로 꺼서, 어떤 뉴런도 옆 뉴런이 항상 있으리라고 믿지 못하게 만듭니다. 그러면 각 뉴런은 어떤 동료와 함께 일하더라도 쓸모 있는 특징, 예컨대 선이나 점 같은 작은 무늬를 혼자 힘으로 찾아내야 합니다. 어떤 유전자와 섞여도 제 몫을 하는 유전자가 선택되는 것처럼, 어떤 뉴런 조합에서도 제 역할을 하는 뉴런이 길러지는 것입니다.

8.2. 드롭아웃의 계산 방식과 실험 결과

드롭아웃의 계산 방식은 간단합니다. 학습 데이터가 하나 들어올 때마다 각 층의 뉴런마다 동전을 던진다고 생각하면 됩니다. 미리 정해 둔 확률로 앞면이 나오면 그 뉴런을 살려 두고, 뒷면이 나오면 그 순간에는 출력을 0으로 만듭니다. 뉴런을 살려 두는 이 확률을 보존 확률(Retention probability)이라고 합니다. 논문은 보존 확률의 기본값으로 은닉층에는 0.5, 입력층에는 0.8 안팎을 권했습니다.

계산 절차를 차례로 적으면 다음과 같습니다. 어느 한 층의 출력은 뉴런마다 값이 하나씩 놓인 벡터입니다. 이 층의 뉴런마다 0 또는 1의 값을 갖는 베르누이 확률변수를 하나씩 뽑으며, 1이 나올 확률이 보존 확률입니다. 이렇게 뽑은 0과 1의 벡터를 출력 벡터와 같은 자리의 원소끼리 곱하고, 이렇게 일부가 지워진 출력을 다음 층의 입력으로 넘깁니다. 다음 층의 계산은 평소와 같습니다.

신경망에서 뉴런을 켜고 끄는 경우의 수는 2를 뉴런의 개수만큼 거듭 곱한 값입니다. 드롭아웃은 학습 단계마다 이 가운데 하나를 무작위로 골라 학습시키는 셈입니다. 그 많은 하위 신경망은 따로 존재하지 않고 같은 가중치를 나누어 씁니다. 신경망 하나의 비용으로 거대한 앙상블을 학습시키는 구조입니다.

학습이 끝나 신경망을 실제로 쓸 때는 뉴런을 하나도 끄지 않습니다. 대신 각 뉴런에서 나가는 가중치에 p 를 곱합니다. 학습 때 어떤 뉴런이 절반의 확률로만 켜져 있었다면, 다음 층이 받는 신호의 평균은 원래의 절반이었

습니다. 평가 때 가중치에 0.5를 곱하면 다음 층은 학습 때와 같은 크기의 신호를 받습니다. 이 간단한 조정만으로 2의 n제곱 개 하위 신경망의 예측을 평균 낸 값, 정확히는 기하평균에 가까운 값을 한 번의 계산으로 얻습니다.

연구진은 드롭아웃을 다른 기법과 함께 쓸 때 효과가 가장 크다는 점도 확인했습니다. 첫째는 맥스노름(Max-norm) 제약입니다. 뉴런 하나로 들어오는 가중치 벡터의 길이가 상수 c (보통 3에서 4 사이)를 넘지 못하게 막는 방법입니다. 둘째는 높은 학습률과 큰 관성(Momentum)입니다. 드롭아웃은 학습에 잡음을 섞기 때문에, 일반 신경망보다 10배에서 100배 큰 학습률과 0.95에서 0.99 정도의 관성 계수를 쓰면 결과가 좋았습니다. 학습률이 크면 가중치가 지나치게 커질 위험이 있는데, 맥스노름 제약이 그 위험을 막아 주었습니다. 두 기법은 이렇게 서로를 보완했습니다.

실험에서는 여러 분야에서 고르게 효과가 확인되었습니다. 손글씨 숫자 데이터인 MNIST에서 일반 신경망의 오차율은 1.60%였습니다. 드롭아웃에 ReLU와 맥스노름을 함께 쓴 은닉층 2개, 층당 뉴런 8,192개짜리 신경망은 파라미터가 약 6,500만 개였는데도 조기 종료 없이 0.95%를 기록했습니다. 깊은 볼츠만 머신(DBM)으로 사전학습한 뒤 드롭아웃으로 미세 조정하자 오차율은 0.79%까지 내려갔습니다.

다른 데이터에서도 결과는 비슷했습니다. 구글 스트리트뷰의 집 번호 사진인 SVHN에서는 합성곱 신경망의 오차율이 3.95%였는데, 전결합층에만 드롭아웃을 넣었을 때 3.02%로, 모든 층에 넣었을 때 2.55%로 낮아졌습니다. 사물 사진 분류 과제인 CIFAR-10에서는 15.60%가 12.61%로,

CIFAR-100에서는 43.48%가 37.20%로 줄었습니다. 음성 인식 과제 TIMIT에서는 4층 음향 모델의 음소 오류율이 22.7%에서 19.7%로 낮아졌습니다. 로이터 기사 분류와 유전자 대체 스플라이싱 예측에서도 효과가 나타났습니다.

2012년 이미지 인식 대회 ILSVRC에서 우승한 AlexNet도 드롭아웃을 썼습니다. 크리제브스키, 수츠케버, 힌턴은 파라미터가 약 6,000만 개인 이 신경망의 앞쪽 전결합층 두 곳에 보존 확률 0.5의 드롭아웃을 넣었습니다. 논문은 드롭아웃이 없었다면 과적합이 심했을 것이라고 밝히면서, 그 대신 수렴에 필요한 반복 횟수가 약 두 배로 늘었다고 적었습니다. 드롭아웃에는 학습 시간이 길어진다는 대가가 따랐던 것입니다.

드롭아웃으로 학습한 신경망의 내부를 살펴보면 다른 특징도 나타납니다. 은닉 뉴런의 평균 활성화값이 낮아지고, 대부분의 뉴런이 0에 가까운 값을 유지하는 희소 표현(Sparse representation)이 저절로 생깁니다. 논문은 0과 1 대신 평균이 1인 정규분포 잡음을 곱하는 가우시안 드롭아웃 같은 변형도 함께 정리했습니다.

드롭아웃은 이후 10년 넘게 심층 신경망의 기본 도구로 쓰였습니다. 다만 쓰임새는 시대에 따라 달라졌습니다. 배치 정규화 같은 새 기법이 나온 뒤 합성곱 신경망에서는 비중이 줄었고, 방대한 데이터를 한두 번만 학습하는 오늘날의 초거대 언어 모델에서는 드롭아웃을 아예 끄고 학습하는 경우도 많습니다. 데이터가 모델 크기에 비해 넉넉하면 모델이 데이터를 외울 여지가 적기 때문입니다. 그래도 데이터가 부족한 분야나 미세 조정 단계에서는 지금도 널리 쓰입니다.

8.3. '어둠의 지식'과 곤충의 변태: 지식 종류의 구상

드롭아웃과 앙상블로 신경망의 일반화 성능은 크게 좋아졌지만, 산업 현장은 다른 문제에 부딪혔습니다. 파라미터가 수천만에서 수억 개에 이르는 거대 모델이나 수십 개 모델을 묶은 앙상블은, 휴대전화나 음성 검색처럼 메모리가 적고 응답이 빨라야 하는 환경에 바로 넣기 어려웠습니다. 잘 학습된 모델을 실제로 배포하는 문제가 남아 있었던 것입니다.

힌턴은 구글에서 함께 일하던 오리올 비니알스(Oriol Vinyals), 제프 딘(Jeff Dean)과 이 문제를 다뤘습니다. 세 사람은 2014년 NIPS 딥러닝 워크숍에서 「신경망 속 지식의 증류(Distilling the Knowledge in a Neural Network)」를 발표했고, 이 논문은 2015년 3월 arXiv(1503.02531)에 공개되었습니다.

논문은 곤충의 변태 이야기로 시작합니다. 나비나 나방 같은 곤충은 애벌레와 성충이라는 전혀 다른 두 형태로 삽니다. 애벌레의 몸은 주변에서 영양분을 최대한 많이 흡수하는 데 맞춰져 있고, 성충의 몸은 이동과 번식에 맞춰져 있습니다. 힌턴은 기계학습에서도 학습과 배포의 형태를 이렇게 나누어야 한다고 보았습니다. 그때까지는 학습에 쓴 모델을 거의 그대로 서비스에도 썼습니다. 그러나 방대한 데이터에서 구조를 뽑아내는 학습 단계에는 크고 무거운 모델이나 드롭아웃 앙상블, 곧 애벌레 형태가 알맞습니다. 학습이 끝나면 그 지식을 작고 빠른 모델, 곧 성충 형태로 옮겨 배포하면 됩니다. 이것이 지식 증류(Knowledge Distillation)의 기본 구상입니다. 이때 큰 모델을 교사(Teacher), 작은 모델을 학생(Student)이라고 부릅니다.

비슷한 시도는 이전에도 있었습니다. 2006년 크리스티안 부칠루아(Cristian Buciluă), 리치 카루아나(Rich Caruana), 알렉산드루 니쿨레스쿠미질(Alexandru Niculescu-Mizil)은 「모델 압축(Model compression)」 논문에서 큰 앙상블의 예측을 작은 모델이 따라 하도록 학습시키는 방법을 보였습니다. 힌턴은 여기서 한 걸음 더 나아갔습니다. 그는 모델의 지식을 가중치 숫자 자체로 보지 않았습니다. 구조가 다른 신경망 사이에서는 가중치를 복사해도 의미가 없기 때문입니다. 옮겨야 할 것은 입력을 받아 출력으로 바꾸는 방식, 곧 모델이 배운 함수였습니다.

그렇다면 그 함수에 담긴 정보를 어디에서 읽어 낼 수 있는지가 문제가 됩니다. 이에 대한 힌턴의 답이 '어둠의 지식(Dark Knowledge)'입니다. 숫자 2의 손글씨 이미지를 잘 학습된 교사 모델에 넣었다고 합시다. 모델은 '2'일 확률을 0.999 정도로 내놓고, '3'일 확률을 100만분의 1, '7'일 확률을 10억분의 1 정도로 내놓을 수 있습니다. 정답만 표시하는 하드 타겟(Hard target)의 관점에서 보면 '3'과 '7'은 둘 다 똑같은 오답입니다. 그러나 두 오답의 확률이 1,000배 차이가 난다는 사실에는, 교사 모델이 이 '2'를 '7'보다 '3'에 훨씬 가깝게 보고 있다는 정보가 담겨 있습니다. 오답들에 매긴 이 작은 확률까지 포함한 분포를 소프트 타겟(Soft target)이라고 합니다. 힌턴은 강연에서, 이렇게 겉으로 잘 드러나지 않는 유사성 정보를 '어둠의 지식'이라고 불렀습니다. 우주의 질량 대부분을 차지하면서도 눈에 보이지 않는 암흑 물질에 빗댄 표현입니다.

두 가지 타겟의 차이는 학생이 무엇을 배우는지에서 드러납니다. 학생 모델에게 정답만 가르치면 학생은 정답을 맞히는 법을 배웁니다. 소프트 타겟을 가르치면 학생은 교사가 세상을 어떻게 나누어 보는지, 무엇과 무엇

을 비슷하게 여기는지까지 배웁니다. 학생이 교사의 일반화 방식 자체를 물려받는 것입니다.

8.4. 소프트맥스 온도 조절과 작은 학생 모델로의 지식 전달

어둠의 지식을 학습에 쓰는 데에는 걸림돌이 있었습니다. 100만분의 1이나 10억분의 1 같은 확률은 너무 작아서, 학습에 쓰는 교차 엔트로피 손실을 계산할 때 거의 영향을 주지 못하고 묻혀 버립니다. 연구진은 이 문제를 통계물리학에서 빌려 온 온도(Temperature, T)라는 장치로 풀었습니다.

신경망의 마지막 층은 분류할 항목(클래스)마다 점수를 하나씩 내놓습니다. 이 점수를 로짓(Logit)이라고 부릅니다. 소프트맥스(Softmax) 함수는 이 점수들을 모두 더하면 1이 되는 확률로 바꿉니다. 증류에서는 여기에 온도라는 수를 하나 넣습니다. 온도를 넣은 소프트맥스는 다음과 같이 계산합니다.

먼저 각 클래스의 로짓을 온도로 나눕니다. 그 값을 지수함수에 넣어 양수로 바꾸는데, 지수함수는 입력이 클수록 결과가 급격히 커지는 함수입니다. 한 클래스의 확률은 그 클래스에서 얻은 값을 모든 클래스에서 얻은 값의 합으로 나눈 것입니다.

평소에는 온도를 1로 둡니다. 온도를 20처럼 크게 올리면 모든 로짓이 큰 수로 나뉘어 점수 차이가 줄어든 것처럼 계산됩니다. 그 결과 0.999와 100만분의 1처럼 극단적이던 확률이 0.45, 0.30, 0.15처럼 고르게 퍼집니다. 큰 소리와 속삭임이 섞인 녹음에서 음량 차이를 줄여 속삭임까지 또렷하게 들리게 하는 것과 비슷합니다. 증류를 할 때는 교사 모델과 학생 모델 모두 같은 높은 온도로 소프트맥스를 계산하고, 학생이 교사의 부드러운

워진 확률 분포를 따라가도록 학습시킵니다. 학습이 끝난 학생 모델은 온도를 다시 1로 되돌려 일합니다.

연구진은 학생의 손실 함수를 두 부분으로 구성할 때 최선의 결과가 나온다는 것을 확인했습니다. 하나는 높은 온도에서 교사의 소프트 타겟과 학생의 예측을 비교하는 손실이고, 다른 하나는 온도 1에서 실제 정답과 학생의 예측을 비교하는 손실입니다. 두 손실을 가중 평균하되 앞의 손실에는 온도의 제곱을 곱합니다. 온도를 높이면 소프트 타겟에서 오는 기울기의 크기가 온도의 제곱에 반비례해 줄어들기 때문입니다. 온도의 제곱을 곱해 두면 온도를 바꾸어 실험해도 두 손실의 상대적 비중이 그대로 유지됩니다. 정답 손실에는 상대적으로 작은 가중치를 줄 때 결과가 좋았습니다.

이 논문에는 이전 연구와 지식 증류의 관계를 밝히는 수학적 결과도 들어 있습니다. 온도를 로짓의 크기보다 충분히 높이고, 교사와 학생의 로짓 평균을 각각 0으로 맞추었다고 가정합니다. 이 조건에서 학생이 점수를 고치는 방향은 교사와 학생의 점수 차이를 줄이는 방향에 가까워집니다. 분류 항목의 수와 온도는 수정 폭에 영향을 줍니다. 높은 온도에서의 증류가 두 모델의 로짓 차이의 제곱을 줄이는 학습에 가까워지는 것입니다. 2006년 카루아나 연구진의 로짓 맞추기 방법은 지식 증류의 한 특수한 경우였던 셈입니다. 다만 온도가 높을수록 좋은 것은 아니었습니다. 연구진은 학생 모델이 아주 작을 때는 오히려 중간 정도의 온도가 낫다고 보고했습니다. 지나치게 높은 온도에서는 교사가 거의 고려하지 않은 아주 낮은 점수의 잡음까지 학생에게 전달되기 때문입니다.

실험 결과는 증류의 효과를 수치로 보여 주었습니다. MNIST에서 뉴런 1,200개짜리 은닉층 두 개를 가진 교사 모델은 드롭아웃과 입력 이미지 흔들기를 써서 학습했고, 시험 데이터 1만 장 가운데 67장을 틀렸습니다. 정규화 없이 학습한 뉴런 800개짜리 작은 모델은 146장을 틀렸습니다. 같은 작은 모델을 온도 20에서 교사의 소프트 타겟으로 학습시키자 틀린 수는 74장으로 줄었습니다. 학생은 흔들린 이미지를 직접 본 적이 없었지만, 교사가 그 경험에서 얻은 일반화 능력이 소프트 타겟을 통해 전달된 것입니다.

연구진은 더 과감한 실험도 했습니다. 학생의 학습 데이터에서 숫자 3의 이미지를 모두 지운 것입니다. 학생은 3이라는 숫자를 한 번도 보지 못한 채 교사의 소프트 타겟만으로 학습했습니다. 결과는 206장의 오류였고, 그 가운데 133장이 시험 데이터의 3 이미지 1,010장에서 나왔습니다. 학생이 3이라는 답을 내놓는 데 지나치게 인색했기 때문입니다. 연구진이 3에 대한 편향 값을 3.5만큼 올려 주자 전체 오류는 109장으로 줄었고, 그 중 3에서 나온 오류는 14장이었습니다. 한 번도 보지 못한 숫자를 98.6% 정확하게 맞힌 것입니다. 교사가 2, 5, 8 같은 다른 숫자 이미지에 대해 '3과 조금 닮았다'고 남긴 작은 확률들 속에 3의 모양에 대한 정보가 들어 있었습니다.

연구진은 이 기법을 구글 안드로이드 음성 검색의 음향 모델에도 적용했습니다. 기준 모델은 음성 프레임 분류 정확도 58.9%, 단어 오류율 10.9%였습니다. 같은 구조의 모델 10개를 묶은 앙상블은 61.1%와 10.7%를 기록했습니다. 이 앙상블을 같은 크기의 단일 모델 하나로 증류하자 60.8%와 10.7%가 나왔습니다. 앙상블이 얻은 프레임 정확도 개선

분의 80% 이상을 모델 하나가 가져간 셈입니다. 연구진은 1억 장 규모의 이미지와 1만 5,000개 라벨로 이루어진 구글 내부 데이터 JFT에서, 헛갈리기 쉬운 클래스 묶음만 맡는 전문가 모델(Specialist models)을 여러 개 병렬로 학습시키는 방법도 함께 제안했습니다.

힌턴은 이후 증류를 더 넓은 시각에서 바라보았습니다. 2022년 순방향 순방향(Forward-Forward) 알고리즘 논문에서 그는 사람의 언어도 한 사람의 뇌 속 표현을 다른 사람의 뇌로 옮기는 일종의 증류일 수 있다고 썼습니다. 같은 글에서 그는 가중치를 그대로 복사할 수 있는 디지털 모델과 달리 사람은 말이라는 좁은 통로로만 지식을 나눌 수 있다는 차이에도 주목했습니다. 이 문제의식은 그가 뒤에 디지털 지능의 위험을 경고하는 논리로 이어집니다.

지식 증류는 대형 언어 모델 시대에 오히려 쓰임이 넓어졌습니다. 2025년 1월 중국 딥시크(DeepSeek)는 추론 모델 DeepSeek-R1을 공개하면서, 이 모델에서 증류한 1.5B, 7B, 8B, 14B, 32B, 70B 크기의 밀집 모델 6종을 Qwen과 Llama를 바탕으로 함께 공개했습니다. 딥시크는 증류한 14B 모델이 당시 공개 모델 가운데 최고 수준이던 QwQ-32B-Preview를 크게 앞섰다고 밝혔습니다. 딥시크 연구진은 작은 모델에는 강화학습만 쓰는 것보다 증류가 더 좋은 성능을 낸다고 보고했습니다. 같은 해 1월 말에는 오픈AI가 딥시크가 모델 증류로 자사 모델을 베꼈다고 주장하면서, 증류가 기술을 넘어 지식재산과 경쟁 질서의 문제로 떠올랐습니다. 다만 오늘날 언어 모델의 증류는 교사가 만든 문장을 학생이 그대로 학습하는 방식이 많아, 오답 확률 분포 전체를 넘겨주는 힌턴의 원래 방식과는 차이가 있습니다.

드롭아웃은 거대한 신경망이 학습 데이터를 외우지 않고 규칙을 배우게 만들었고, 지식 증류는 그렇게 배운 지식을 작은 모델로 옮겨 사람들이 손에 든 기기에서 쓸 수 있게 했습니다. 하나는 학습의 문제를, 다른 하나는 배포의 문제를 풀었습니다. 두 기법 모두 무작위로 섞일 때 강해지는 유전자, 몸의 형태를 바꾸어 사는 곤충처럼 생물의 세계에서 실마리를 찾았다는 점에서 힌턴의 연구 방식을 잘 보여 줍니다.

제9장. 함께 만든 딥러닝의 발전

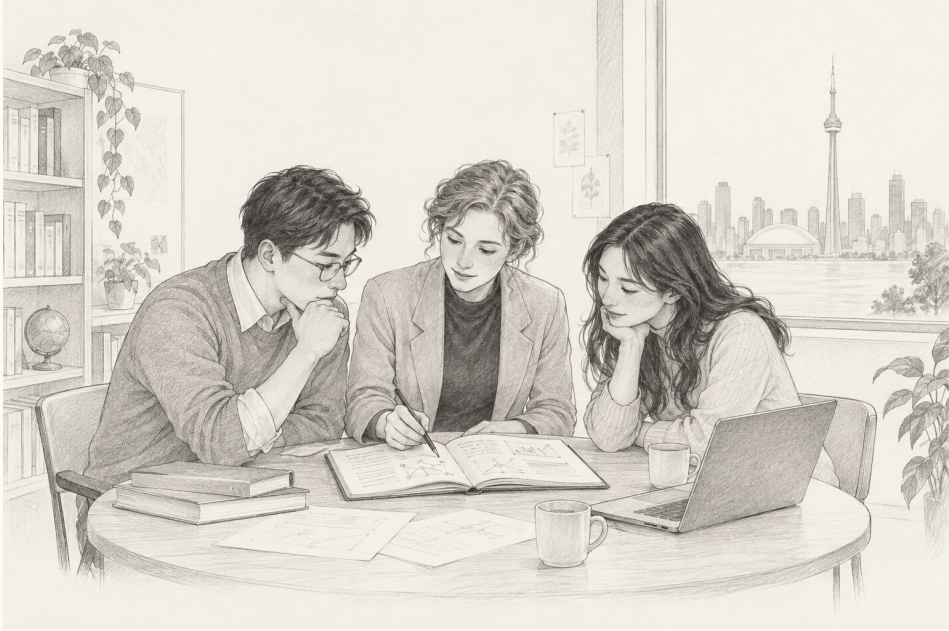


그림 9. 동료와 제자가 함께 발전시킨 연구를 표현했습니다.

9.1. 'AI 겨울'을 견디게 한 제도: CIFAR의 NCAP 프로그램과 장기 지원 방식

2000년대 초 인공지능 학계의 중심에는 서포트 벡터 머신(SVM), 커널 기법, 결정 트리 같은 통계 모델이 있었습니다. 이 방법들은 결과를 수학으로 설명하기 쉬웠고, 정답에 가까운 해를 찾을 수 있다는 보장도 있었습니다. 여러 층으로 쌓은 인공신경망은 사정이 달랐습니다. 학습이 중간에 멈추거나 엉뚱한 지점에 갇히는 일이 잦았고, 왜 잘 되는지 증명하기도 어려웠습니다. 그래서 많은 연구자가 신경망을 낡고 쓸모없는 방법으로 여겼습니다. 신경망을 다룬 논문은 학회 심사에서 자주 떨어졌고, 연구비를 받기도 힘들었습니다. 앞선 인공지능 연구의 침체를 지나서도 신경망 연구에 대한 냉대는 한동안 이어졌습니다.

이 시기에 신경망 연구가 끊기지 않도록 버팀목이 된 곳이 캐나다 고등연구소(CIFAR, Canadian Institute for Advanced Research)입니다. CIFAR는 1982년 토론토에서 문을 연 비영리 연구 기관으로, 캐나다 연방 정부와 주 정부, 민간 후원자의 재원으로 운영됩니다. 이 기관은 특정 대학에 속하지 않습니다. 여러 나라의 뛰어난 연구자를 한 주제 아래 모아 오랫동안 함께 연구하게 하는 방식을 씁니다.

제프리 힌턴과 CIFAR의 인연은 1987년에 시작됩니다. 힌턴은 그해 미국 카네기멜런대학교(CMU) 교수 자리를 떠나 캐나다 토론토대학교로 옮겼습니다. 당시 미국의 인공지능 연구비는 상당 부분 국방부에서 나왔는데, 힌턴은 군사 목적의 돈으로 연구하는 것을 원하지 않았습니다. 토론토로

온 그는 CIFAR의 첫 연구 프로그램인 '인공지능, 로봇공학 및 사회'에 펠로로 참여했습니다.

그로부터 17년이 지난 2004년은 신경망 연구가 학계에서 크게 외면받던 때였습니다. 힌턴은 이때 동료들과 함께 CIFAR에 새 연구 프로그램을 제안했습니다. 몬트리올대학교의 요슈아 벤지오, 토론토대학교의 데이비드 플리트 등이 제안에 함께했습니다. CIFAR는 이 제안을 받아들여 같은 해 '신경 계산 및 적응 인지(NCAP, Neural Computation and Adaptive Perception)' 프로그램을 시작했고, 힌턴이 프로그램의 책임자(디렉터)를 맡았습니다. NCAP은 10년 동안의 운영을 평가받은 뒤 '기계와 뇌의 학습(Learning in Machines & Brains)'으로 이름을 바꾸어 지금도 이어지고 있습니다.

학계의 관심 밖에 있던 주제가 이렇게 오래 지원받을 수 있었던 까닭은 CIFAR의 지원 방식에 있습니다. 보통의 연구비는 과제 단위로 지급됩니다. 연구자는 무엇을 언제까지 만들어 내겠다고 약속해야 하고, 정해진 날짜에 결과 보고서를 내야 합니다. CIFAR는 과제보다 사람과 연구 모임을 지원했습니다. 엄선한 소수의 연구자에게 주제를 강요하지 않았고, 몇 년 동안 함께 생각하고 토론할 자리를 마련해 주었습니다. 연구자들은 해마다 작은 워크숍에 모여 아직 발표하지 않은 생각과 실패한 실험까지 털어놓았습니다.

힌턴은 2024년 10월 보도된 서울경제신문 인터뷰에서 이 차이를 직접 설명했습니다. 그는 캐나다의 연구비 규모가 미국만큼 크지는 않다고 인정했습니다. 다만 미국 교수들은 연구비 제안서를 쓰는 데 시간 대부분을 들

이는 반면, 캐나다에서는 6쪽 정도의 제안서로 3~5년 동안 연구비를 받을 수 있다고 말했습니다. 그는 연구자의 자율과 호기심을 지켜 주는 기본 연구비가 캐나다의 힘이라고 보았습니다.

NCAP은 기존 연구자를 지원하는 데 그치지 않고 다음 세대 연구자를 키우는 일도 했습니다. 2005년부터 열린 여름 학교가 그 통로였습니다. 이 행사는 훗날 '딥러닝 및 강화학습 여름 학교(Deep Learning + Reinforcement Learning Summer School)'라는 이름으로 자리 잡았습니다. 해마다 여러 나라의 대학원생과 젊은 연구자가 캐나다에 모여 최신 신경망 기법을 배우고 서로 인연을 맺었습니다. CIFAR는 이 학교를 거쳐 간 연구자가 20년 동안 2,700명이 넘는다고 집계합니다. CIFAR는 당장의 성과를 요구하지 않고 비인기 분야를 오랫동안 지원했고, 이 지원은 훗날 딥러닝이 다시 주목받는 데 필요한 제도적 기반이 되었습니다.

9.2. '딥러닝 3대 대부'의 협력: 힌턴, 벤지오, 르쿤

NCAP의 중심에는 훗날 '딥러닝 3대 대부'로 불리는 세 학자가 있었습니다. 토론토대학교의 제프리 힌턴, 몬트리올대학교의 요슈아 벤지오, 뉴욕대학교(NYU)의 안 르쿤입니다.

세 사람의 인연은 1980년대 후반으로 거슬러 올라갑니다. 프랑스 파리에 박사 학위를 받은 르쿤은 1987년 토론토로 건너와 힌턴의 연구실에서 박사후연구원으로 일했습니다. 그는 1988년 미국 벨연구소로 옮겨 손글씨 숫자를 읽는 합성곱 신경망(CNN)을 발전시켰고, 이 연구는 훗날 은행 수표의 숫자를 자동으로 읽는 시스템으로 이어졌습니다. 벤지오는 캐나다 맥길대학교에서 박사 학위를 받은 뒤 미국 MIT와 벨연구소에서 연구했으며, 벨연구소에서 르쿤과 함께 일했습니다. 그는 1993년 몬트리올대학교 교수가 되었습니다. 세 사람은 이렇게 스승과 제자, 동료로 이어진 사이였습니다.

1990년대 중반 이후 신경망이 학계의 관심에서 밀려나자 세 사람은 각자의 자리에서 고립되었습니다. 토론토, 몬트리올, 뉴욕에 흩어져 있던 이들을 다시 한 모임으로 묶은 것이 CIFAR의 NCAP이었습니다. 세 사람은 수십 명의 연구원, 대학원생과 함께 정기 워크숍에 모였고, 그 자리에서 완성되지 않은 아이디어와 새로운 수식을 거리낌 없이 나누었습니다. 같은 생각을 가진 사람이 적을 때, 서로를 믿고 의견을 주고받을 수 있는 작은 공동체는 큰 힘이 되었습니다. 힌턴이 책임자 자리에서 물러난 뒤에는 벤지오와 르쿤이 이 프로그램을 함께 이끌었습니다.

딥러닝의 핵심 성과로 꼽히는 연구들이 이 세 사람의 연구실에서 잇달아 나왔습니다. 벤지오 연구진은 2003년 단어의 뜻을 숫자 벡터로 나타내는 신경망 언어 모델을 내놓았습니다. 힌턴 연구진은 2006년 심층 신뢰 신경망(DBN)과 층별 사전학습 방법을 발표했습니다. 깊은 신경망도 한 층씩 차례로 학습시키면 잘 훈련할 수 있다는 것을 보여 준 연구였습니다. 벤지오 연구진은 2014년에는 번역 모델에 '어텐션' 방식을 도입했고, 어텐션은 훗날 챗GPT 같은 대형 언어 모델의 바탕이 되었습니다. 르쿤은 합성곱 신경망을 계속 다듬어 컴퓨터 비전의 기본 구조로 만들었습니다. 세 사람의 연구는 서로를 참고하며 함께 발전했습니다.

세 사람의 업적은 여러 상으로 인정받았습니다. 미국 컴퓨터학회(ACM)는 2019년 3월 세 사람을 2018년 튜링상 공동 수상자로 발표했습니다. 튜링상은 '컴퓨터 과학의 노벨상'으로 불립니다. 2024년 10월 8일에는 힌턴이 존 홉필드와 함께 노벨 물리학상 수상자로 발표되었습니다. 같은 해 12월 세 사람은 페이페이 리, 젠슨 황과 함께 총상금 300만 달러의 빈퓨처(VinFuture) 대상을 받았습니다. 2025년 2월에는 벤지오, 힌턴, 홉필드, 르쿤, 젠슨 황, 빌 델리, 페이페이 리 등 7명이 현대 기계학습 발전에 기여한 공로로 퀸 엘리자베스 공학상을 받았습니다. 이 가운데 엔비디아의 젠슨 황과 델리는 GPU 개발로, 스탠퍼드대학교의 리는 이미지 데이터베이스 이미지넷(ImageNet)을 만든 공로로 수상자에 포함되었습니다. 수상자 명단은 딥러닝의 발전에 알고리즘뿐 아니라 계산 장치와 데이터를 만든 사람들의 기여가 함께 있었다는 점을 보여 줍니다.

세 사람이 협력했다고 해서 생각이 늘 같았던 것은 아닙니다. 2023년 이후 세 사람은 인공지능의 위험을 두고 서로 다른 목소리를 냈습니다. 힌턴

과 벤지오는 인공지능이 인류의 통제를 벗어날 수 있다고 경고했습니다. 르쿤은 이런 걱정이 지나치다고 보았고, 현재의 대형 언어 모델만으로는 사람 수준의 지능에 이를 수 없다고 주장해 왔습니다.

르쿤의 이 견해는 그의 진로와도 이어집니다. 그는 2013년 메타에 합류해 페이스북 인공지능 연구소(FAIR)를 만들었고, 12년 가운데 5년은 FAIR의 초대 책임자로, 7년은 수석 AI 과학자로 일했습니다. 그는 2025년 11월 페이스북에 글을 올려 2025년 말 메타를 떠난다고 밝혔습니다. 이어 물리 세계를 이해하고 추론하며 행동을 계획하는 시스템을 만드는 스타트업을 세운다고 했습니다. 의견이 갈린 뒤에도 세 사람은 공개 토론을 이어갔습니다. 서로 다른 생각을 존중하면서 따져 묻는 이 태도는 CIFAR 워크숍에서 길러진 연구 문화와 닮아 있습니다.

9.3. 세계 AI 연구자를 길러 낸 곳: 토론토대학교 연구실과 '토론토 네트워크'

힌턴의 영향력은 그가 쓴 논문에만 있지 않습니다. 토론토대학교 연구실에서 공부한 제자들이 세계 곳곳의 대학과 기업으로 퍼져 나가면서 그의 생각은 더 넓은 곳에 자리를 잡았습니다. 이 제자들과 동료들의 연결망을 흔히 '토론토 네트워크'라고 부릅니다.

제자 가운데 널리 알려진 이는 알렉스 크리제브스키와 일리야 수츠케버입니다. 두 사람은 2012년 힌턴과 함께 '알렉스넷(AlexNet)'이라는 신경망을 만들어 국제 이미지 인식 대회(ILSVRC)에 나갔습니다. 이 대회는 신경망이 사진마다 답을 다섯 개까지 내게 하고, 그 다섯 개 안에 정답이 없는 비율을 상위 5개 답안 오류율로 집계했습니다. 알렉스넷의 상위 5개 답안 오류율은 15.3%였고, 2위 팀은 26.2%였습니다. 이 큰 격차에 컴퓨터 비전 학계가 크게 놀랐습니다. 이 성과에는 두 제자의 역할이 뚜렷했습니다. 크리제브스키는 그래픽 처리 장치(GPU)로 신경망을 빠르게 학습시키는 프로그램을 직접 짰습니다. GPU는 본래 화면에 그림을 그리는 계산을 맡는 부품으로, 같은 종류의 계산을 한꺼번에 많이 처리할 수 있습니다. 덕분에 학습으로 값이 정해지는 수인 매개변수가 6,000만 개에 이르는 신경망을 현실적인 시간 안에 훈련할 수 있었습니다. 수츠케버는 학습을 안정시키는 여러 기법을 설계해 성공을 함께 이끌었습니다.

세 사람은 2012년 말 'DNN리서치(DNNresearch)'라는 회사를 세웠습니다. 구글은 2013년 3월 이 회사를 인수했고, 인수 금액은 4,400만 달러

로 알려졌습니다. 힌턴은 이때부터 구글에서 일하면서 토론토대학교 연구도 이어 갔고, 2023년 5월 구글을 떠났습니다.

수츠케버는 구글에서 2014년 문장을 다른 문장으로 바꾸는 '시퀀스 투 시퀀스(Seq2Seq)' 모델을 동료들과 발표했습니다. 이 모델은 기계 번역의 수준을 크게 끌어올렸습니다. 그는 2015년 12월 샘 올트먼 등과 함께 오픈AI(OpenAI)를 세우고 수석 과학자를 맡았습니다. 2024년 5월 오픈AI를 떠난 그는 같은 해 6월 안전한 초지능 개발을 내건 회사 SSI(Safe Superintelligence)를 세웠습니다.

다른 제자들도 각자의 분야에서 중요한 역할을 맡았습니다. 루슬란 살라쿠티디노프는 2006년 힌턴과 함께 학술지 《사이언스》에 오토인코더 논문을 발표했고, 심층 볼츠만 머신(DBM)을 함께 개발했습니다. 그는 카네기멜런대학교 교수가 되었고, 2016년 애플의 첫 인공지능 연구 책임자로 임명되었습니다. 조지 달과 압델라만 모하메드는 2009~2012년 음성 인식에 심층 신경망을 적용했습니다. 이 연구는 2012년 마이크로소프트, 구글, IBM 연구진과 함께 쓴 공동 논문으로 이어졌고, 오랫동안 정체되어 있던 음성 인식 기술을 바꾸어 놓았습니다. 볼로디미르 므니흐는 딥마인드에서 비디오 게임을 스스로 배우는 강화학습 신경망(DQN) 연구를 이끌었습니다. 니티시 스리바스타바는 힌턴과 함께 신경망의 과적합을 줄이는 '드롭아웃' 기법을 정리했습니다.

힌턴 연구실을 거친 연구자는 이 밖에도 많습니다. 베이지안 신경망을 연구한 래드포드 닐, 훗날 Vector Institute의 연구 책임자가 된 리처드 제멜, 확률 모델을 연구한 브렌던 프레이와 예휘 테가 그의 박사 과정 제자임

니다. 신경과학과 기계학습을 잇는 연구를 한 피터 다얀, 맥스 웰링, 손글씨와 음성 인식 연구로 이름을 알린 알렉스 그레이브스는 박사후연구원으로 힌턴과 함께 일했습니다. 이들은 토론토, 옥스퍼드, 런던, 암스테르담, 뉴욕과 실리콘밸리로 흩어져 다시 자기 연구실을 꾸렸습니다. 한 연구실의 문화가 세계 여러 연구 거점으로 옮겨 간 셈입니다.

9.4. Vector Institute와 캐나다의 국가 AI 전략: 변방에서 AI 연구의 중심지로

알렉스넷 이후 딥러닝은 산업계 전체로 퍼졌습니다. 구글, 페이스북, 마이크로소프트 같은 기업이 신경망 연구자를 앞다투어 영입했습니다. 세계 인공지능 연구에서 변방이던 캐나다는 딥러닝을 처음 키워 낸 나라로 이름을 얻었습니다. 그러나 인재를 지키는 일이 문제로 남았습니다. 토론토와 몬트리올에서 성장한 연구자 상당수가 높은 연봉을 따라 미국으로 떠났습니다.

캐나다 정부는 2017년 3월 연방 예산안에서 '범캐나다 인공지능 전략'을 발표했습니다. 정부는 이 전략에 1억 2,500만 캐나다 달러를 배정했고, 전략의 설계와 운영을 CIFAR에 맡겼습니다. 캐나다 정부는 이 전략으로 캐나다가 재원을 갖춘 국가 인공지능 전략을 처음 만든 나라가 되었다고 설명합니다. 2021년 예산안에서는 전략 2단계에 4억 4,300만 캐나다 달러가 추가로 배정되었습니다.

전략의 중심에는 세 곳의 국가 인공지능 연구소가 있습니다. 토론토의 Vector Institute, 벤지오가 세운 몬트리올의 Mila, 앨버타대학교와 연결된 에드먼턴의 Amii입니다. Amii는 앤드루 바토와 함께 강화학습 연구로 2024년 튜링상을 받은 리처드 서튼과 인연이 깊습니다.

이 가운데 Vector Institute는 2017년 3월 30일 공식 발표되었습니다. 힌턴은 이 연구소의 수석 과학 자문을 맡았고, 제자인 리처드 제멜이 연구 책임자가 되었습니다. 연구소는 정부 지원금과 함께 온타리오주 정부와

구글, 에어캐나다, 캐나다 5대 은행 등 30곳 넘는 기업의 투자를 받았습니다. 출범 자금은 보도에 따라 1억 5,000만에서 2억 캐나다 달러 사이로 집계되며, 관련 자료에는 약 1억 7,500만 캐나다 달러로 정리되어 있습니다. 연구소를 세운 연구자들은 Vector가 인재를 끌어들이고 붙잡아 두는 자석이 되기를 바랐습니다. 이들의 목표는 많은 박사 과정 학생을 캐나다에 머물게 하고 인공지능 스타트업을 돕는 데 있었습니다.

Vector Institute는 Mila, Amii와 함께 캐나다 인공지능 연구의 세 축을 이루었습니다. 구글, 메타, 우버, 톰슨로이터 같은 기업이 토론토와 몬트리올에 연구소를 열었습니다. 한때 미국으로 빠져나가던 인재가 캐나다로 돌아오거나 캐나다에 남는 흐름도 생겼습니다. 토론토가 '딥러닝의 성지'라는 별명을 얻은 것도 이 무렵입니다.

캐나다의 투자는 그 뒤로도 이어졌습니다. 2024년 예산안은 인공지능 분야에 24억 캐나다 달러를 배정했습니다. 이 가운데 20억 캐나다 달러는 새로 만드는 'AI 컴퓨팅 접근 기금'과 '캐나다 AI 주권 컴퓨팅 전략'을 통해 연구자와 스타트업이 쓸 컴퓨팅 자원과 기반 시설을 마련하는 데 쓰입니다. 5,000만 캐나다 달러로는 캐나다 AI 안전 연구소를 세웁니다. 대규모 계산 자원이 연구 경쟁력을 좌우하게 되자, 사람 중심의 지원에 기반 시설 투자가 더해진 것입니다.

힌턴은 국가가 연구를 지원하는 방식에 분명한 의견을 가지고 있습니다. 그는 앞에서 소개한 서울경제신문 인터뷰에서, 정치인과 정책 담당자는 구체적인 사업에 돈을 쓴다고 말하기 쉽다고 지적했습니다. 그러면서 1982년 일본이 막대한 예산을 들여 시작한 '제5세대 컴퓨터 프로젝트'를

예로 들었습니다. 목표를 미리 정해 놓은 대형 사업은 겉보기에 그럴듯하지만 기대한 성과를 내지 못했다는 것입니다. 힌턴은 연구자가 호기심을 따라 자유롭게 연구하도록 돕는 환경이 효과적이며, 딥러닝도 그렇게 태어났다고 말했습니다.

신경망은 한때 학계의 변방에 있던 가설이었습니다. 그 가설을 주류 과학으로 끌어올린 힘은 한 사람의 재능에서만 나오지 않았습니다. 성과를 서두르지 않은 CIFAR의 장기 지원, 힌턴과 벤지오와 르쿤의 협력, 토론토 연구실에서 성장해 세계로 퍼진 제자들, 그리고 그 성과를 국가 전략으로 이어 간 캐나다 정부의 선택이 함께 작용했습니다. 딥러닝의 역사는 과학이 개인의 재능만으로 이루어지지 않고 사람과 제도가 함께 만드는 공동 작업이라는 점을 보여 줍니다.

제4부. 더 나은 학습을 찾아서

제10장. 부분과 전체를 알아보는 기계



그림 10. 관점이 달라져도 유지되는 부분과 전체의 관계를 표현했습니다.

10.1. 합성곱 신경망(CNN)의 한계와 맥스 풀링에 대한 비판

2012년 AlexNet이 이미지 인식 대회에서 큰 차이로 우승한 뒤, 합성곱 신경망(CNN)은 컴퓨터 비전의 표준이 되었습니다. 이 성공의 한가운데에 제프리 힌턴이 있었습니다. 그러나 힌턴은 자신이 이끈 성공을 그대로 받아들이지 않았습니다. 그는 CNN이 사물을 보는 방식에 근본적인 결함이 있다고 여겼고, 이 생각을 오랫동안 공개적으로 밝혔습니다.

힌턴이 문제 삼은 지점을 이해하려면 CNN이 그림을 처리하는 순서를 먼저 살펴야 합니다. CNN은 작은 필터를 그림 전체에 옮겨 가며 선, 모서리, 무늬 같은 특징을 찾습니다. 층을 거칠수록 눈, 바퀴, 귀처럼 더 큰 특징을 찾아냅니다. 그 사이사이에 맥스 풀링이라는 연산이 들어갑니다. 맥스 풀링은 작은 구역 안에서 가장 큰 값 하나만 남기고 나머지를 버리는 연산입니다. 이렇게 하면 특징이 조금 옮겨져도 결과가 크게 달라지지 않습니다. 이를 위치 불변성이라고 부릅니다. 계산량도 줄어듭니다.

대가는 정보의 손실입니다. 특징이 정확히 어디에 있었는지, 어느 방향으로 기울었는지, 크기가 얼마였는지 같은 자세(pose) 정보가 풀링 단계에서 사라집니다. 힌턴은 2014년 온라인 커뮤니티 레딧(Reddit)에서 연 질의응답에서, CNN의 풀링 연산은 큰 실수이고 그것이 잘 작동한다는 사실이 오히려 재앙이라고 말했습니다. 잘 작동하기 때문에 연구자들이 더 나은 방식을 찾으려 하지 않는다는 뜻이었습니다.

자세 정보가 사라지면 어떤 일이 생기는지는 얼굴 그림으로 쉽게 설명됩니다. 사람은 눈, 코, 입이 있지만 확인하지 않습니다. 눈 두 개가 코 위

에 나란히 있고 입이 코 아래에 있는지, 즉 부분들이 전체를 이루는 배치를 함께 따집니다. 이것을 부분-전체 위계(part-whole hierarchy)라고 합니다. CNN은 각 부분의 감지기가 켜지기만 하면 눈과 입의 자리가 뒤바뀔 그림도 얼굴로 판정하는 경향을 보였습니다. 부분 사이의 공간 관계를 정밀하게 따지는 장치가 없었기 때문입니다.

시점 변화도 해결되지 않은 과제였습니다. 같은 물체를 위에서, 옆에서, 비스듬히 보면 픽셀 값은 완전히 달라집니다. CNN은 이 차이를 견디려고 회전시키고 늘이고 비튼 이미지를 대량으로 학습했습니다. 이를 데이터 증강이라고 합니다. 힌턴이 보기에 이것은 이해가 아니라 암기에 가까웠습니다.

그래서 힌턴은 목표를 다시 세웠습니다. 물체가 움직여도 출력이 그대로 인 불변성(invariance)이 아니라, 물체가 움직이면 내부 표현도 그만큼 규칙적으로 바뀌는 동변성(equivariance)이 필요하다고 보았습니다. 물체가 무엇인지는 그대로 알아보되, 그 물체가 어디에 어떤 자세로 있는지도 함께 알아야 한다는 것입니다.

이 생각의 뿌리는 1979년으로 거슬러 올라갑니다. 힌턴은 그해 학술지 『Cognitive Science』에 심상(mental imagery)과 구조적 기술에 관한 논문을 실었습니다. 거기에 정육면체 실험이 나옵니다. 철사로 만든 정육면체를 한 꼭짓점이 아래로, 맞은편 꼭짓점이 바로 위로 오게 세워 놓고 나머지 꼭짓점의 위치를 손으로 가리키게 하면 사람들은 대부분 틀립니다. 머릿속에서 정육면체를 익숙한 좌표계로 떠올리기 때문입니다. 힌턴은 이 실험으로 사람이 물체에 고유한 좌표계를 부여하고, 그 좌표계에 따라 같

은 모양도 다르게 인식한다는 점을 보였습니다. 40여 년 뒤 GLOM 논문도 첫머리에서 이 실험을 근거로 듭니다.

10.2. 물체의 좌표계와 변환 오토인코더: 캡슐의 등장

물체마다 좌표계가 있다는 생각을 신경망에 옮기기 위해 힌턴은 컴퓨터 그래픽스에서 실마리를 찾았습니다. 3차원 게임이나 애니메이션은 물체마다 고유한 좌표계를 두고, 그 물체가 카메라와 어떤 관계에 있는지를 행렬로 표현합니다. 행렬은 수를 가로와 세로로 늘어놓은 표로, 물체를 돌리거나 옮기거나 키우는 변환을 곱셈 한 번으로 나타낼 수 있습니다. 시점이 바뀌면 행렬 하나를 곱하면 됩니다. 픽셀 하나하나의 밝기로 이루어진 픽셀 공간에서는 시점 변화가 복잡하게 얽혀 나타납니다. 물체의 위치와 방향과 크기를 수로 적은 자세 공간에서는 같은 변화가 곱셈과 덧셈만으로 이루어지는 선형 계산이 됩니다. 그래픽스가 자세 정보로 그림을 만든다면, 시각 인식은 그림에서 자세를 거꾸로 알아내는 일입니다. 힌턴은 이것을 역그래픽스(inverse graphics)라고 불렀습니다.

이 구상은 2011년 핀란드 에스포에서 열린 국제인공신경망학술대회(ICANN 2011)에서 처음 논문으로 나왔습니다. 힌턴은 알렉스 크리제브스키, 시다 왕(Sida D. Wang)과 함께 「변환 오토인코더(Transforming Auto-encoders)」를 발표했습니다. 이 논문에서 '캡슐(capsule)'이라는 이름이 처음 쓰였습니다.

기존 뉴런은 숫자 하나를 내놓습니다. 캡슐은 작은 뉴런 묶음으로, 두 종류의 정보를 함께 내놓습니다. 하나는 자신이 맡은 대상이 그 자리에 있을 확률입니다. 다른 하나는 그 대상의 위치, 방향, 크기 같은 자세를 나타내는 여러 개의 숫자, 곧 벡터입니다. 논문은 이 숫자들을 발현 파라미터(instantiation parameters)라고 불렀습니다.

캡슐이 자세를 담도록 만드는 학습 방법은 다음과 같습니다. 신경망에 원본 이미지와 함께 '오른쪽으로 몇 칸, 위로 몇 칸 옮겨라' 같은 이동 값(Δx , Δy)을 줍니다. 캡슐의 인식 부분은 원본 이미지에서 대상의 위치 (x , y)를 알아냅니다. 여기에 이동 값을 더해 ($x+\Delta x$, $y+\Delta y$)를 만들고, 캡슐의 생성 부분이 이 좌표에 맞추어 옮겨진 이미지를 그려 냅니다. 정답은 실제로 옮긴 이미지입니다. 신경망은 그린 결과와 정답의 차이를 줄이도록 역전파로 학습합니다. 캡슐이 내놓은 존재 확률은 생성 부분의 출력에 곱해져, 대상이 없으면 그 캡슐이 그림에 끼치는 영향을 줄입니다.

여기서 핵심은 누구도 캡슐에게 '이 숫자가 위치다'라고 가르치지 않았다는 점입니다. 덧셈 한 번으로 맞는 결과가 나오려면 캡슐 안의 숫자가 실제 위치처럼 움직여야 합니다. 신경망은 픽셀로 주어진 정보를 자세를 나타내는 숫자로 바꾸는 방법을 스스로 익혔습니다.

캡슐은 10.1절에서 본 얼굴 문제, 곧 부분과 전체를 잇는 방식도 제시했습니다. 코 캡슐과 입 캡슐이 각자 자세를 알아냈다고 합시다. 코와 얼굴 사이에는 고정된 공간 관계가 있습니다. 코의 자세에 이 관계를 나타내는 행렬을 곱하면 얼굴의 자세를 예측할 수 있습니다. 입도 마찬가지입니다. 두 예측이 서로 일치하면 그 자리에 얼굴이 있을 가능성이 높습니다. 여러 숫자로 이루어진 예측이 우연히 맞아떨어지는 일은 드물기 때문입니다. 이 일치 검사가 이후 캡슐 연구의 중심 원리가 되었습니다.

10.3. 동적 라우팅과 EM 라우팅: 캡슐 네트워크의 본격적인 전개 (2017~2018)

캡슐 개념은 6년 동안 다듬어졌습니다. 그 결과가 2017년과 2018년 두 편의 논문으로 나왔습니다. 저자는 구글 브레인 토론토 연구실의 사라 사부르(Sara Sabour), 니컬러스 프로스트(Nicholas Frosst), 그리고 힌턴이었습니다.

2017년 12월 NIPS 학회에서 발표된 「캡슐 사이의 동적 라우팅 (Dynamic Routing Between Capsules)」은 캡슐을 실제로 작동하는 신경망으로 만들었습니다. 이 모델을 흔히 CapsNet이라고 부릅니다. CapsNet에는 두 가지 장치가 들어 있습니다.

첫째 장치는 스쿼시(squash) 함수입니다. 캡슐의 출력 벡터에서 길이는 대상이 있을 확률을, 방향은 자세를 나타냅니다. 확률이니 길이는 0과 1 사이여야 합니다. 스쿼시 함수는 벡터의 방향은 그대로 두고 길이만 0과 1 사이로 줄여 줍니다. 짧은 벡터는 0에 가깝게, 긴 벡터는 1에 가깝게 줄입니다.

둘째 장치가 동적 라우팅, 정확히는 '합의에 의한 라우팅(routing by agreement)'입니다. 앞 절에서 본 일치 검사를 신경망의 계산 절차로 옮긴 것입니다. 아래층 캡슐(부분)은 학습된 가중치 행렬을 곱해 위층 각 캡슐(전체)의 출력을 예측합니다. 처음에는 모든 위층 캡슐에 예측을 고르게 보냅니다. 위층 캡슐은 받은 예측을 모아 출력을 만듭니다. 이어서 각 예측이 위층 출력과 얼마나 같은 방향을 가리키는지 내적으로 잹니다. 잘 맞

는 위층 캡슐로는 다음 번에 더 많이 보내고, 맞지 않는 쪽으로는 덜 보냅니다. 이 과정을 몇 차례 되풀이합니다. 논문의 기본 설정은 세 번이었습니다.

이 방식은 맥스 풀링과 정반대입니다. 풀링은 가장 큰 값 하나를 고르고 나머지를 버립니다. 라우팅은 정보를 버리지 않고, 어느 부분이 어느 전체에 속하는지를 그림마다 새로 정합니다. 한 부분이 여러 전체 후보 가운데 자신이 속할 전체를 찾아가는 방식입니다.

실험 결과도 이 설계를 뒷받침했습니다. CapsNet은 손글씨 숫자 데이터 MNIST에서 0.25%의 오류율을 기록했습니다. 층이 셋뿐인 얇은 신경망으로 거둔 결과였습니다. 논문은 숫자 캡슐의 출력만으로 원래 숫자를 다시 그려 내는 복원 장치도 붙였습니다. 캡슐 벡터의 숫자 하나를 조금씩 바꾸면 복원된 숫자의 획 굵기, 기울기, 너비가 그에 맞추어 바뀌었습니다. 캡슐 안의 숫자들이 실제로 자세와 모양을 담고 있다는 증거였습니다.

MultiMNIST 실험은 캡슐의 장점을 더 분명히 보여 주었습니다. 두 숫자를 크게 겹쳐 놓은 이미지에서 두 숫자를 모두 알아맞히는 과제입니다. CapsNet은 약 5%의 오류율로 비교 대상 CNN보다 좋은 성적을 냈고, 겹친 두 숫자를 각각 따로 다시 그려 내기도 했습니다. 각 숫자 캡슐이 자기 숫자에 해당하는 획만 가져갔다는 뜻입니다.

2018년 ICLR 학회에서는 후속 논문 「EM 라우팅을 쓰는 행렬 캡슐(Matrix Capsules with EM Routing)」이 나왔습니다. 이번에는 힌턴이 제1저자였습니다. 이 모델에서 캡슐은 존재 여부를 나타내는 값 하나

와 4×4 자세 행렬을 갖습니다. 4×4 행렬은 컴퓨터 그래픽스가 3차원 물체의 위치와 회전을 나타낼 때 쓰는 형식 그대로입니다. 그래픽스의 발상을 신경망 안으로 한층 깊이 들여온 것입니다.

라우팅 방법도 바뀌었습니다. 아래층 캡슐의 자세 행렬에 변환 행렬을 곱하면 위층 캡슐의 자세에 대한 '투표'가 나옵니다. 위층 캡슐은 여러 투표가 한곳에 모여 있는지를 봅니다. 이 판단에 통계학의 기댓값 최대화(EM) 알고리즘을 씁니다. E 단계에서는 각 투표가 어느 위층 캡슐의 무리에 속하는지 확률을 계산합니다. M 단계에서는 그 확률을 바탕으로 위층 캡슐의 평균, 분산, 활성도를 새로 정합니다. 투표들이 촘촘히 모이면 그 전체가 존재한다고 판단합니다.

이 모델은 smallNORB 데이터로 시험했습니다. 장난감 비행기, 자동차, 트럭, 사람 인형, 동물 인형을 여러 각도와 조명에서 찍은 3차원 물체 인식 데이터입니다. 행렬 캡슐의 오류율은 사진에서 한 영역을 잘라 평가했을 때 1.8%, 여러 영역의 결과를 합쳤을 때 1.4%였습니다. 후자의 수치는 비교 대상이던 기존 방법의 2.56%보다 오류가 약 45% 적었습니다. 학습 때 보지 못한 각도에서 찍은 사진을 알아보는 시험에서도 비슷한 크기의 CNN보다 나았습니다. 적대적 공격, 곧 사람 눈에는 보이지 않는 작은 잡음으로 신경망을 속이는 공격에도 더 잘 버텼습니다.

2019년에는 애덤 코시오레크(Adam Kosiorek), 사부르, 이 와이 테(Yee Whye Teh), 힌턴이 「쌓인 캡슐 오토인코더(Stacked Capsule Autoencoders)」를 NeurIPS에서 발표했습니다. 정답 표시 없이 이미

지를 보기만 해도 부분 템플릿을 찾아내고, 이를 묶어 물체 캡슐을 만드는 비지도 학습 모델이었습니다.

캡슐 네트워크는 학계의 큰 관심을 받았지만 한계도 뚜렷했습니다. 라우팅을 여러 번 되풀이해야 해서 계산이 느렸고, 큰 이미지나 복잡한 데이터로 규모를 키우기 어려웠습니다. 같은 시기에 트랜스포머가 언어와 이미지 분야에서 빠르게 주류가 되면서 캡슐은 주류 기술이 되지 못했습니다.

그래도 연구는 이어졌습니다. 2024년 3월 영국 애버딘 대학교 연구진은 캡슐 네트워크가 복잡한 데이터에서 잘 학습하지 못해 현대 과제로 확장되지 못했다는 점을 출발점으로 삼아, 자기지도 사전학습을 쓴 첫 캡슐 네트워크인 마스크 캡슐 오토인코더(MCAE)를 제안했습니다. 이미지넷 크기의 사진 10개 범주로 이루어진 Imagenette 데이터로 사전학습한 결과 캡슐 네트워크 가운데 최고 성능을 냈고, 기준 모델보다 9% 나은 성적을 보였습니다. 이 연구는 2025년 학술지 TMLR(Transactions on Machine Learning Research)에 실렸습니다.

10.4. 여러 연구를 하나로 묶는 구상: GLOM, 부분-전체 위계와 동일 벡터의 섬

힌턴은 계산 속도보다 더 깊은 문제가 캡슐 모델에 남아 있다고 보았습니다. 기존 캡슐은 이미지의 각 위치마다 나타날 수 있는 모든 종류의 부분과 물체에 대해 캡슐을 미리 마련해 두어야 했습니다. 일반 컴퓨터라면 그림마다 필요한 만큼 메모리를 새로 잡아 나무 구조를 만들면 됩니다. 뇌의 신경망은 그렇게 하지 못합니다. 뉴런이 하는 일은 들어오고 나가는 연결가중치로 정해지고, 실제 뉴런은 그 가중치를 빠르게 완전히 바꿀 수 없기 때문입니다. 그렇다면 그림마다 모양이 다른 부분-전체 나무를 고정된 구조의 신경망이 어떻게 표현할 수 있는지가 문제가 됩니다. 힌턴은 이 문제를 새 연구 과제로 삼았습니다.

2021년 2월 25일, 힌턴은 아카이브(arXiv)에 43쪽 분량의 논문 「신경망에서 부분-전체 위계를 표현하는 방법(How to represent part-whole hierarchies in a neural network)」을 올렸습니다. 이 구상의 이름 GLOM은 하나로 뭉친다는 뜻의 영어 속어 'glom together'에서 따왔고, 이 말은 'agglomerate'에서 왔을 수 있습니다. 논문은 첫 문장에서 작동하는 시스템을 설명하는 글이 아니라고 밝히고, 트랜스포머, 신경장(neural fields), 대조 표현 학습, 지식 증류, 캡슐 등 여러 연구 성과를 하나로 묶는 표현 방식과 그것을 담은 상상 속의 시스템 GLOM을 제시했습니다. 힌턴은 2021년 5월 5일 캐나다 페리미터 이론물리연구소 강연 등에서 같은 구상을 소개했습니다.

GLOM의 구조는 다음과 같습니다. 이미지를 작은 조각으로 나누고, 조각마다 '컬럼(column)'을 하나씩 둡니다. 모든 컬럼은 같은 가중치를 씁니다. 각 컬럼은 여러 층(level)으로 나뉩니다. 아래부터 세부 부분, 부분, 물체, 장면 같은 순서로 쌓입니다. 각 층에는 그 위치에 무엇이 있는지를 나타내는 벡터, 곧 임베딩이 하나씩 있습니다. 이 벡터 하나하나가 특정 종류에 묶이지 않은 보편적 캡슐 역할을 합니다. 종류마다 캡슐을 미리 마련해 두어야 했던 기존 캡슐의 부담을 이 방식으로 덜어 냅니다.

이미지가 들어오면 GLOM은 여러 시간 단계에 걸쳐 각 벡터를 고쳐 나갑니다. 벡터는 여러 개의 수를 순서대로 묶은 것으로, 여기서는 이미지의 한 위치에 무엇이 있는지를 나타냅니다. 매 단계에서 한 층의 벡터는 네 가지 값의 가중 평균으로 바뀝니다. 가중 평균은 값마다 비중을 다르게 주어 구한 평균입니다. 첫째 값은 바로 아래층 벡터를 상향식 신경망에 넣어 얻은 예측입니다. 둘째 값은 바로 위층 벡터를 하향식 신경망에 넣어 얻은 예측입니다. 하향식 신경망은 컬럼의 위치도 함께 입력받는데, '얼굴'이라는 정보와 위치를 알면 그 자리에 코가 와야 하는지 눈이 와야 하는지 계산할 수 있기 때문입니다. 이는 신경장 연구에서 빌려 온 발상입니다. 셋째 값은 자기 층의 직전 값입니다. 넷째 값은 이웃 컬럼들의 같은 층 벡터를 어텐션으로 평균한 값으로, 자기와 비슷한 벡터를 가진 이웃일수록 더 큰 비중을 줍니다. 트랜스포머의 어텐션과 닮았지만, 질의, 키, 값으로 따로 변환하지 않고 벡터끼리 직접 비교하는 간단한 형태입니다.

이 과정을 되풀이하면 GLOM의 핵심 현상이 나타납니다. 얼굴에 해당하는 모든 조각의 컬럼은 물체 층에서 똑같은 벡터로 수렴합니다. 코에 해당하는 조각들은 부분 층에서 같은 벡터를 갖고, 입에 해당하는 조각들은 또

다른 같은 벡터를 갖습니다. 힌턴은 이것을 '동일 벡터의 섬(islands of identical vectors)'이라고 불렀습니다. 섬 하나가 나무 구조의 마디 하나입니다. 물체 층의 큰 섬 하나 아래에 부분 층의 작은 섬 여럿이 놓이면, 그것이 곧 얼굴과 코, 입의 부분-전체 관계입니다. 뉴런을 새로 배정하지 않고도 그림마다 다른 나무를 표현할 수 있습니다. 고정된 구조의 신경망이 그림마다 다른 구조를 어떻게 표현하느냐는 문제에 대해 힌턴이 내놓은 답이 이것입니다.

학습 방법으로 힌턴은 일부를 가린 이미지를 복원하는 방식을 제안했습니다. 같은 물체에 속한 위치들의 벡터가 서로 비슷해지도록 대조 학습 방식의 규제도 덧붙였습니다. 섬이 잘 만들어지면 신경망이 무엇을 어떻게 보았는지 사람이 읽어 내기 쉬워집니다. 논문 요약도 GLOM이 작동한다면 시각이나 언어에 쓰이는 트랜스포머 계열 시스템이 만든 표현의 해석 가능성이 크게 높일 것이라고 적었습니다.

힌턴은 GLOM이 지각 연구의 두 전통을 잇는다고 보았습니다. 하나는 지각을 하나의 장(field)으로 보고 전체가 부분의 합과 다르다고 본 게슈탈트 심리학입니다. 다른 하나는 지각을 기호와 구조적 기술로 나타내려 한 고전 기호주의 인공지능입니다. GLOM에서 지각은 연속적인 벡터의 장으로 나타나지만, 섬들이 이루는 구조 안에는 분명한 부분-전체 관계가 담깁니다.

GLOM은 아직 완성된 대규모 시스템으로 구현된 사례가 널리 알려져 있지 않습니다. 힌턴 자신도 2022년 말 Forward-Forward 알고리즘을 발표하며 학습 규칙 자체로 관심을 옮겼습니다. 그래도 GLOM이 제기한 문

제는 여전히 유효합니다. 고정된 신경망이 그림마다 다른 구조를 어떻게 표현하는지, 그리고 그 표현을 사람이 읽을 수 있게 만들 수 있는지에 관한 문제입니다. 1979년 정육면체 실험에서 시작해 2011년 캡슐, 2017년 동적 라우팅, 2018년 행렬 캡슐을 거쳐 2021년 GLOM에 이르는 힌턴의 탐구는 모두 이 문제를 향해 있었습니다.

제11장. 뇌에서 다시 찾는 학습 방법

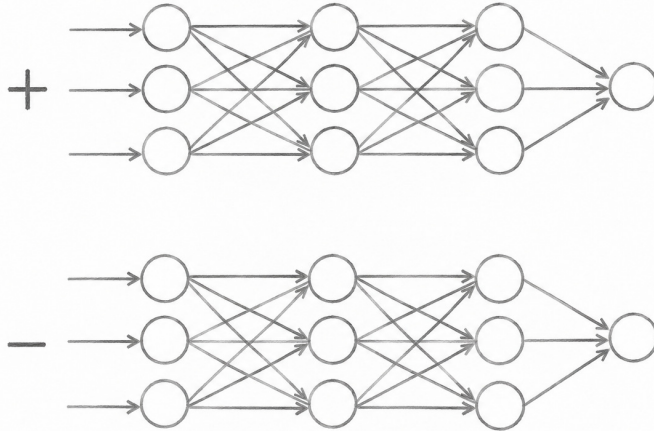


그림 11. 긍정 사례와 부정 사례를 앞으로 전달하는 학습을 표현했습니다.

11.1. 뇌는 실제로 역전파로 배울까

1986년 데이비드 루멜하트, 제프리 힌턴, 로널드 윌리엄스는 《네이처》에 오차 역전파(Backpropagation) 논문을 실었습니다. 역전파의 절차는 이렇습니다. 신경망이 답을 내면 먼저 정답과의 차이, 곧 오차를 구합니다. 그다음 미분의 연쇄 법칙(Chain Rule)을 이용해 이 오차를 출력층에서 입력층 쪽으로 거꾸로 보내면서, 각 연결 가중치를 얼마나 고쳐야 하는지 계산합니다. 이 방법은 이후 40년 가까이 딥러닝을 떠받친 중심 기술이 되었습니다. 이미지 인식과 기계 번역, 거대 언어 모델(LLM)에 이르기까지 대부분의 성과가 역전파를 바탕으로 나왔습니다.

그런데 힌턴은 오랫동안 한 가지 질문을 놓지 않았습니다. 사람의 대뇌피질도 이런 방식으로 배우는가 하는 질문입니다. 그의 답은 대체로 부정적이었습니다. 2020년 힌턴은 티모시 릴리크랩(Timothy Lillicrap) 등과 함께 《네이처 리뷰 뉴로사이언스》에 「역전파와 뇌(Backpropagation and the brain)」를 발표했습니다. 이 논문과 2022년 힌턴의 연구를 함께 보면, 역전파가 뇌에서 그대로 일어나기 어려운 이유는 세 가지로 정리됩니다.

첫째는 가중치 수송 문제(Weight Transport Problem)입니다. 역전파에서 오차를 거꾸로 보내려면 순방향 계산에 쓴 가중치 행렬을 뒤집은 전치 행렬이 필요합니다. 뒤로 가는 길이 앞으로 가는 길과 정확히 같은 세기로 짜지어져 있어야 한다는 뜻입니다. 그러나 실제 뇌의 신경 연결에서는 이런 일대일 대칭 구조가 확인되지 않습니다.

둘째는 오차 신호만 따로 보내는 통로와 이전 상태를 저장하는 장치가 뇌에 없다는 점입니다. 역전파는 평소의 신호와 구분되는 오차 미분값을 별도 통로로 보내야 하고, 역방향 계산을 하려면 순방향 계산 때의 활성화 값을 기억해 두어야 합니다. 뇌의 피드백 연결은 이렇게 작동하지 않습니다. 피드백은 신호를 받는 뉴런의 활동 세기 자체를 조절할 뿐이며, 부호가 붙은 오차 값을 따로 실어 나른다는 신경과학적 증거는 아직 없습니다.

셋째는 시간의 문제입니다. 동영상처럼 계속 들어오는 자료를 학습하려면 시간 경과 역전파(BPTT)를 써야 합니다. 이 방법은 과거 시점으로 오차를 되돌려 보내야 하므로 입력 처리를 자주 멈춰야 합니다. 반면 뇌는 눈과 귀로 들어오는 신호를 멈추지 않고 흘려보내면서 동시에 배웁니다. 힌턴은 이를 파이프라이닝(Pipelining)이라는 말로 설명했습니다.

힌턴의 이런 문제 제기를 역전파를 버리자는 주장으로 읽어서는 안 됩니다. 그는 여러 인터뷰에서 자신의 의문이 뇌를 이해하려는 관점에서 나왔다고 설명했습니다. 디지털 컴퓨터에서 작동하는 인공지능에서는 역전파가 여전히 다른 학습 방법보다 좋은 성능을 냅니다. 다만 뇌가 역전파를 쓰지 않는다면 뇌가 쓰는 다른 방법이 있을 것이고, 그 방법이 새로운 종류의 컴퓨터를 만드는 실마리가 될 수 있다는 것이 그의 생각이었습니다.

힌턴은 뇌와 인공 신경망의 차이를 데이터와 연결 수의 비율로도 설명했습니다. 거대 언어 모델은 파라미터가 1조 개 안팎인데, 학습에 쓰는 글은 수조 토큰에 이릅니다. 경험은 넘치고 저장 공간은 상대적으로 부족한 구조입니다. 역전파는 이렇게 많은 경험을 적은 파라미터에 압축해 넣는 데 뛰어납니다. 사람은 정반대입니다. 80년을 산다고 해도 인생은 약 25억

초에 그치는데, 뇌의 시냅스 연결은 약 100조 개에 이릅니다. 경험은 적고 연결은 넘치는 구조입니다. 힌턴은 이런 조건에서는 뇌가 역전파와 다른 학습 방식을 쓸 것이라고 보았습니다.

11.2. Forward-Forward 알고리즘: 두 번의 순방향 계산과 '좋음'이라는 점수

2022년 12월 힌턴은 미국 뉴올리언스에서 열린 신경정보처리시스템학회(NeurlPS) 기조강연에서 새로운 학습 방법을 소개했습니다. 같은 달 논문 공개 사이트 arXiv에는 「The Forward-Forward Algorithm: Some Preliminary Investigations」 (arXiv:2212.13345)가 올라왔습니다. 당시 그는 구글 브레인 토론토 소속이었습니다. 제목에 '예비 연구'라는 말을 붙인 데서 알 수 있듯이, 힌턴은 이 방법을 완성된 결과가 아니라 출발점으로 제시했습니다.

핵심 생각은 간단합니다. 역전파는 입력에서 출력 쪽으로 한 번, 출력에서 입력 쪽으로 한 번 계산합니다. Forward-Forward(FF) 알고리즘은 뒤로 가는 계산을 없애고, 똑같은 방식의 순방향 계산을 두 번 합니다. 한 번은 진짜 데이터로, 다른 한 번은 가짜 데이터로 계산합니다. 가짜 데이터는 실제 데이터와 닮았지만 일부러 틀리게 만든 데이터입니다. 진짜 데이터를 넣는 단계를 양성 패스(Positive Pass), 가짜 데이터를 넣는 단계를 음성 패스(Negative Pass)라고 부릅니다.

이 생각에는 앞선 연구가 있습니다. 1980년대에 힌턴은 테리 세이노프스키와 함께 볼츠만 머신(Boltzmann Machine)을 연구했습니다. 볼츠만 머신은 데이터를 보는 단계와 모델이 스스로 만든 상태를 보는 단계를 비교하면서 학습했습니다. 2010년 미하엘 구트만과 아포 히바리넨이 제한한 노이즈 대조 추정(NCE)도 진짜 데이터와 잡음을 구별하는 방식으로 모델을 학습시킵니다. FF는 이 두 연구에 공통된 대조(Contrast)의 원리,

곧 두 종류의 입력을 견주어 그 차이에서 배우는 원리를 오늘날의 다층 신경망에 맞게 다시 구성한 것입니다.

각 층이 진짜와 가짜를 구별하는 기준은 층마다 따로 계산하는 점수입니다. 힌턴은 이 점수를 좋음(Goodness)이라고 불렀습니다. 좋음은 그 층에 있는 뉴런들의 활성화 값을 각각 제공해 모두 더한 값입니다. 활성화 값은 뉴런이 입력을 받아 내놓는 출력의 크기이고, 활성화 함수로는 ReLU를 씁니다. ReLU는 입력이 0보다 작으면 0을 내고 0보다 크면 입력을 그대로 내는 함수입니다. 따라서 뉴런들이 강하게 반응할수록 그 층의 좋음이 커집니다.

학습의 목표는 두 가지입니다. 진짜 데이터가 들어오면 좋음이 미리 정한 기준값보다 충분히 커지도록 가중치를 고치고, 가짜 데이터가 들어오면 좋음이 기준값보다 충분히 작아지도록 고칩니다. 각 층은 좋음을 바탕으로 입력이 진짜일 확률을 다음과 같이 계산합니다.

먼저 그 층의 좋음에서 기준값을 뺍니다. 이 차이를 시그모이드 함수에 넣어 얻은 값이 입력이 진짜일 확률입니다.

시그모이드 함수는 어떤 값이든 0과 1 사이의 값으로 바꿔 줍니다. 좋음이 기준값보다 크면 확률이 1에 가까워지고, 작으면 0에 가까워집니다. 각 층은 이 확률이 진짜 데이터에서는 높게, 가짜 데이터에서는 낮게 나오도록 자기 가중치만 고칩니다. 다른 층에서 오차를 받아 올 필요가 없습니다. 이처럼 각 층이 자기 층 안의 정보만으로 배우는 방식이 FF의 지역 학습(Local Learning)입니다.

그런데 층을 여러 개 쌓으면 문제가 하나 생깁니다. 첫째 층이 진짜와 가짜를 잘 가려내게 되면, 진짜일 때는 출력 벡터가 길고 가짜일 때는 짧아집니다. 둘째 층이 이 길이만 보고 판단한다면 새로 배울 것이 없어집니다. 힌턴은 이 문제를 층 정규화(Layer Normalization)로 해결했습니다. 층 정규화는 2016년 지미 바, 제이미 키로스, 힌턴이 함께 발표한 기법입니다. FF는 이 기법을 써서 다음 층에 넘기기 전에 벡터의 길이를 지우고 방향만 남깁니다. 방향은 뉴런들이 서로 어떤 비율로 활동하는지를 나타냅니다. 그 결과 위층은 아래층이 이미 사용한 정보에 의존할 수 없고, 활동 패턴 속에 숨은 새로운 특징을 찾아야 합니다.

남은 문제는 가짜 데이터를 어떻게 만드느냐입니다. FF는 정답 표시(라벨)가 없는 비지도 학습과 라벨이 있는 지도 학습에 모두 쓸 수 있는데, 두 경우에 가짜 데이터를 만드는 방법이 다릅니다. 비지도 학습에서 힌턴은 손글씨 숫자 그림 두 장을 섞어 가짜 그림을 만들었습니다. 무작위 점 그림을 여러 번 흐리게 처리해 큰 덩어리 모양의 가리개(마스크)를 얻고, 한 그림에서는 가리개 안쪽을, 다른 그림에서는 바깥쪽을 가져와 합쳤습니다. 이렇게 만든 그림은 짧은 획 같은 작은 부분은 진짜와 비슷하지만 전체 모양은 숫자가 아닙니다. 네트워크는 진짜 그림과 이 가짜 그림을 구별하면서 넓은 범위의 모양을 파악하는 법을 배웁니다.

지도 학습에서는 라벨을 그림 안에 직접 넣습니다. MNIST 손글씨 숫자 그림의 첫 10개 화소 자리에 숫자 하나를 표시하는 방식입니다. 그림에 맞는 라벨을 붙이면 양성 데이터가 되고, 틀린 라벨을 붙이면 음성 데이터가 됩니다. 두 데이터는 그림이 같고 라벨만 다릅니다. 그래서 네트워크는 라벨과 관계없는 배경은 무시하고 그림과 라벨의 관계에 집중하게 됩니다. 학

습이 끝난 뒤 시험할 때는 새 그림에 0부터 9까지의 라벨을 하나씩 붙여순방향 계산을 열 번 합니다. 그리고 여러 층에서 쌓인 좋은 합의 가장 큰 라벨을 답으로 고릅니다.

역방향 계산을 없애서 얻는 이점도 있습니다. 역전파는 오차를 거꾸로 보내기 위해 앞서 계산한 값을 저장해 두어야 하지만, FF는 그럴 필요가 없습니다. 동영상처럼 연속으로 들어오는 자료도 멈추지 않고 흘려보내면서 각 층이 그때그때 가중치를 고칠 수 있습니다. 11.1에서 말한 셋째 문제를 피해 가는 방식입니다. 또 층 사이에 내부를 알 수 없는 장치, 이른바 블랙박스가 끼어 있어도 FF는 학습을 계속할 수 있습니다. 역전파는 모든 단계를 미분할 수 있어야 하지만, FF의 각 층은 자기 입력과 출력만 알면 되기 때문입니다.

성능은 힌턴의 논문에 보고되어 있습니다. 그림을 조금씩 옮겨 학습 자료를 늘린 지도 FF는 MNIST 시험 데이터에서 0.64%의 오류율을 기록했습니다. 역전파로 학습한 합성곱 신경망과 비슷한 수준입니다. CIFAR-10 사진 데이터에서도 역전파보다 조금 낮지만 크게 뒤지지 않는 결과를 보였습니다.

11.3. 수면 가설과 필사적 연산, 그리고 남은 한계

FF를 뇌의 학습 모델로 보려면 풀어야 할 문제가 있습니다. 뇌가 감각 신호를 받을 때마다 진짜와 가짜를 번갈아 처리한다고 보기는 어렵기 때문입니다. 힌턴은 이 문제에 대해 수면 가설을 제시했습니다. 깨어 있을 때 뇌는 바깥에서 들어오는 진짜 감각 자료로 양성 패스만 합니다. 그리고 잠을 잘 때, 그중에서도 꿈을 많이 꾸는 렘(REM) 수면 동안에는 뇌 안의 생성 모델이 스스로 가짜 자료를 만들어 음성 패스를 한다는 것입니다. 이 가설이 맞다면 꿈은 뇌가 학습을 위해 만든 음성 데이터일 수 있습니다.

힌턴은 이 대목에서 자신의 실수를 밝혔습니다. 그는 논문 초안에 양성 학습을 수천 번 한 뒤 음성 학습을 수천 번 하는 방식, 곧 깨어 있는 단계와 수면 단계를 나눈 방식이 잘 작동한다고 적었습니다. 그러나 이후 이 결과가 프로그램 코드의 오류(버그) 때문에 나온 것이었고 재현되지 않았다고 논문에서 바로잡았습니다. 지금의 좋은 함수로 두 단계를 크게 떼어 학습시키려면 한 번에 가중치를 고치는 폭인 학습률을 아주 낮게, 직전의 변화 방향을 이어 가는 정도인 관성(모멘텀)을 아주 높게 잡아야 합니다. 힌턴은 양성 단계와 음성 단계를 시간적으로 완전히 분리하는 일을 앞으로 풀어야 할 중요한 과제로 남겨 두었습니다.

FF가 큰 의미를 갖는 분야는 하드웨어입니다. 지금의 디지털 컴퓨터는 트랜지스터를 켜고 끄는 방식으로 정확한 계산을 합니다. 역전파의 미분값을 정밀하게 구하려면 전력을 많이 써야 하고, 메모리와 연산 장치 사이에서 데이터를 끊임없이 주고받아야 합니다. 아날로그 하드웨어는 다릅니다. 뉴런의 활동을 전압으로, 연결 가중치를 전도도(전기가 얼마나 잘 흐르는

지를 나타내는 값)로 나타내면, 전압과 전도도의 곱은 전하가 되고 여러전하는 저절로 더해집니다. 옴의 법칙과 키르히호프의 법칙이라는 물리현상만으로 곱셈과 덧셈이 끝나는 셈입니다. 이렇게 하면 같은 계산에 드는 에너지를 크게 줄일 수 있습니다. 사람의 뇌는 약 20와트의 전력으로작동합니다.

문제는 아날로그 소자가 하나하나 미세하게 다르다는 점입니다. 같은 공장에서 만들어도 소자마다 특성이 조금씩 어긋납니다. 이런 장치에서는 역전파에 필요한 정확한 미분 계산이 어렵습니다. 반면 FF는 각 층이 자기 입력과 출력만으로 학습하기 때문에 소자의 성질을 정확히 몰라도 학습할 수 있습니다.

힌턴은 여기서 필사적 연산(Mortal Computation)이라는 개념을 제시했습니다. 디지털 컴퓨터에서는 소프트웨어와 하드웨어가 분리되어 있습니다. 학습이 끝난 가중치를 다른 컴퓨터에 그대로 복사할 수 있고, 하드웨어가 고장 나도 지식은 남습니다. 힌턴은 이를 불멸의 연산(Immortal Computation)이라고 불렀습니다. 아날로그 하드웨어에서는 사정이 다릅니다. 가중치가 그 기기의 고유한 특성에 맞춰 학습되기 때문에, 그 가중치를 다른 기기로 옮기면 제대로 작동하지 않습니다. 기기가 수명을 다하면 그 안에 쌓인 지식도 함께 사라집니다. 사람의 뇌도 마찬가지입니다. 한 사람이 평생 쌓은 시냅스 연결을 다른 사람의 뇌에 복사할 수는 없습니다. 사람은 말과 글이라는 느린 통로로 지식을 조금씩 전할 뿐입니다.

이 생각은 힌턴의 인공지능 위험 경고로 이어졌습니다. 그는 디지털 지능이 같은 모델의 복사본 수천 개로 서로 다른 경험을 한 뒤, 가중치 변화를

평균 내어 한꺼번에 공유할 수 있다는 점에 주목했습니다. 사람끼리는 할 수 없는 일입니다. 힌턴은 2023년 5월 구글을 떠난 뒤 디지털 지능의 이런 장점이 사람에게 위협이 될 수 있다고 공개적으로 경고했습니다. 아날로그 컴퓨터를 위한 학습 방법을 찾던 연구가 디지털 지능의 능력을 다시 평가하는 계기가 된 것입니다. 그는 2024년 존 홉필드와 함께 노벨 물리학상을 받았습니다.

FF의 한계도 분명합니다. 첫째, 같은 성능에 이르기까지 역전파보다 학습이 느립니다. 둘째, 파라미터가 수천억 개에 이르는 거대 모델이나 ImageNet 같은 대규모 데이터에서는 역전파만큼 일반화하지 못합니다. 힌턴 자신도 전력이 문제가 되지 않는 대형 디지털 컴퓨터에서는 당분간 역전파가 주류로 남을 것으로 보았습니다. 셋째, 생물학적 모델로서도 미완성입니다. 수면 단계를 분리한 학습은 아직 성공하지 못했고, 대뇌피질의 복잡한 되먹임 회로를 FF가 모두 설명한다고 말할 수도 없습니다.

논문이 발표된 뒤 여러 연구자가 FF의 약점을 보완하는 작업에 나섰습니다. 주요 과제는 쓸 만한 음성 데이터를 만드는 일이었습니다. 프랑스 연구진의 평가에 따르면 FF는 대부분의 표준 벤치마크에서 최고 수준의 성능에 이르지 못했고, 그 원인 가운데 하나는 비지도 학습에서 음성 데이터를 만드는 방법이 믿을 만하지 못했기 때문입니다. 이 문제를 풀기 위해 싱 천, 둥수 류, 제레미 레이드방, 질리 그롤리에는 2024년 9월 자기 대조 Forward-Forward(SCFF)를 arXiv에 공개했습니다. SCFF는 자기 지도 대조 학습에서 아이디어를 얻어 하나의 데이터 안에서 양성 입력과 음성 입력을 만들어 냅니다. 그 결과 비지도 분류 정확도에서 MNIST 98.7%, CIFAR-10 80.75%, STL-10 77.3%를 기록해 기존의 순방향 지역 학습

방법들을 앞섰습니다. 연구진은 FF의 적용 범위를 순환 신경망 학습까지 넓혀 순서가 있는 데이터도 다룰 수 있게 했습니다. 이 논문은 2025년 《네이처 커뮤니케이션스》 16권에 실렸습니다.

하드웨어 분야에서도 연구가 이어졌습니다. 마이크로컨트롤러에서 FF로 학습하는 연구, 수직형 낸드 플래시 메모리 안에서 FF로 온칩 학습을 하는 연구 등이 나왔습니다. 칩 안에서 바로 학습하는 기술과 뉴로모픽 컴퓨팅 연구에서 FF는 하나의 비교 기준이 되었습니다.

Forward-Forward 알고리즘은 역전파를 대신하지 못했습니다. 힌턴도 그런 주장을 하지 않았습니다. 이 연구가 남긴 것은 학습에 반드시 역방향 계산이 필요한가 하는 질문을 다시 제기했다는 점입니다. 역전파를 널리 알린 연구자가 36년 뒤 그 방법의 바깥을 탐색했다는 사실은, 힌턴의 연구에서 뇌에 대한 관심이 여전히 출발점이었음을 보여 줍니다.

제12장. 정답표 없이 배우는 방법

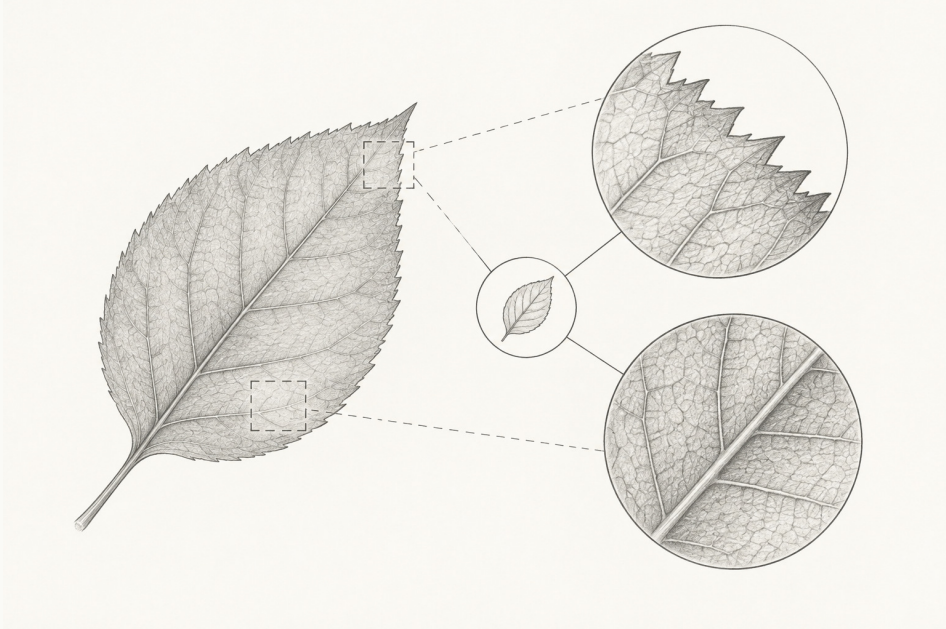


그림 12. 같은 대상의 서로 다른 모습에서 공통점을 배우는 과정을 표현했습니다.

12.1. 라벨 없이 배우기: 자기지도 학습이라는 과제

2012년 AlexNet이 이미지 인식 대회에서 압승한 뒤, 컴퓨터 비전 연구는 지도 학습(Supervised Learning)을 중심으로 움직였습니다. 지도 학습은 사람이 사진마다 "고양이", "자동차", "사과" 같은 정답 이름표, 곧 라벨을 붙여 주고 신경망이 그 정답을 맞히도록 훈련하는 방식입니다. 이 방식의 성능은 라벨이 붙은 데이터가 얼마나 많은지에 크게 좌우됩니다. 대표 데이터셋인 ImageNet은 1,000개 범주에 걸쳐 약 128만 장의 훈련용 사진을 담고 있는데, 이 사진에 일일이 라벨을 붙이는 데 많은 사람과 시간이 들었습니다. 의료 영상이나 위성 사진처럼 전문가만 판독할 수 있는 분야에서는 라벨 비용이 훨씬 커집니다.

사람이 세상을 배우는 방식은 이와 다릅니다. 어린아이는 누가 사진마다 이름을 알려 주지 않아도 눈앞에 끊임없이 들어오는 장면을 보며 사물의 생김새, 같은 것과 다른 것, 공간의 이어짐을 스스로 익힙니다. 연구자들은 이처럼 사람이 붙인 라벨 없이 데이터 자체의 구조와 규칙에서 쓸모 있는 특징을 뽑아내는 방법을 자기지도 학습(Self-Supervised Learning)이라고 부르며, 머신러닝이 도달해야 할 중요한 목표로 꼽아 왔습니다.

자기지도 학습은 크게 두 방향으로 발전했고, 두 방향 모두 한계가 있었습니다. 첫째는 생성적 접근법입니다. 오토인코더(Autoencoder), 변분 오토인코더(VAE), 생성적 적대 신경망(GAN)처럼 입력 이미지를 픽셀 단위로 다시 그려 내도록 학습하는 방법입니다. 이 방식은 계산량이 많고, 사물을 구별하는 데 필요 없는 미세한 무늬나 잡음까지 재현하느라 신경망의 용량을 낭비한다는 한계가 있었습니다. 둘째는 변별적 접근법입니다. 연

구자가 인위적인 연습 과제(Pretext Task)를 설계해 신경망에 풀게 하는 방법으로, 이미지를 돌려 놓고 몇 도 돌렸는지 맞추게 하거나, 조각을 섞은 뒤 직소 퍼즐처럼 원래 자리를 찾게 하거나, 흑백 사진에 색을 입히게 하는 과제가 쓰였습니다. 이 방식은 연구자가 정한 규칙에 크게 기대기 때문에, 여러 분야에 두루 통하는 일반적인 표현을 얻는 데 제약이 있었습니다. 여기서 표현은 신경망이 입력의 특징을 간추려 담은 수의 묶음을 뜻합니다.

라벨 없이 표현을 배우는 문제에 대한 제프리 힌턴의 생각은 1992년으로 거슬러 올라갑니다. 그해 힌턴은 수잔 베커(Suzanna Becker)와 함께 학술지 Nature에 무작위 점 입체 그림(random-dot stereogram)에서 표면의 깊이 구조를 스스로 찾아내는 신경망 연구를 발표했습니다. 이 연구의 핵심은 서로 떨어져 있거나 조금씩 다르게 보이는 두 시각 조각이라도 같은 대상에서 나왔다면 신경망 내부의 표현은 서로 일치해야 한다는 생각이었습니다. 정답 라벨 대신 같은 출처에서 나온 두 관점의 일치를 학습 신호로 삼는다는 이 발상은 약 30년 뒤 대조 학습의 뿌리가 됩니다.

2020년 2월, 구글 브레인(Google Brain) 토론토 연구팀의 팅 첸(Ting Chen), 사이먼 콘블리스(Simon Kornblith), 모하마드 노루지(Mohammad Norouzi), 그리고 제프리 힌턴은 「A Simple Framework for Contrastive Learning of Visual Representations(시각 표현의 대조 학습을 위한 간단한 프레임워크)」라는 논문을 공개했습니다. 논문은 같은 해 7월 온라인으로 열린 국제머신러닝학회(ICML 2020)에 실렸고, 이 방법은 SimCLR라는 이름으로 널리 알려졌습니다. 논문 제목에 들어간 "간단한"이라는 말은 이 연구의 성격을 그대로 보여

줍니다. 연구진은 복잡한 특수 구조나 별도의 저장 장치 없이, 기본 구성 요소만으로 당시 최고 수준의 자기지도 학습 성능을 냈습니다.

12.2. 데이터 증강과 대조 손실: SimCLR가 학습하는 방법

SimCLR의 기본 생각은 쉽게 설명할 수 있습니다. 사진 한 장을 두 가지 방식으로 변형하면 겉모습은 달라져도 담긴 내용은 같습니다. 신경망은 이 두 변형본이 "같은 사진에서 나왔다"는 사실을 알아보고, 다른 사진에서 나온 변형본과는 구별하도록 훈련됩니다. 비슷한 것은 가깝게, 다른 것은 멀게 놓도록 배우기 때문에 이 방식을 대조 학습(Contrastive Learning)이라고 부릅니다.

이 생각을 구현하기 위해 SimCLR는 네 가지 구성 요소를 차례로 연결합니다. 첫째는 확률적 데이터 증강(Data Augmentation) 모듈입니다. 라벨이 없는 사진 x 한 장에 같은 변형 묶음에서 무작위로 뽑은 두 가지 변형을 각각 적용해 두 장의 변형본을 만듭니다. 이 두 장을 양성 쌍(Positive Pair)이라고 부릅니다. 둘째는 기저 인코더 f 입니다. 변형된 사진에서 특징을 뽑아내는 신경망으로, 논문은 ResNet-50을 기본으로 썼습니다. 마지막 평균 풀링 층을 지나면 2,048개의 숫자로 이루어진 표현 벡터 h 가 나옵니다. 셋째는 비선형 투영 헤드 g 입니다. 은닉층 하나와 ReLU 함수로 된 작은 신경망으로, h 를 대조 손실을 계산하는 공간의 벡터 z 로 옮깁니다. 넷째는 대조 손실 함수로, 양성 쌍끼리는 가까워지고 같은 묶음 안의 다른 사진들과는 멀어지도록 신경망 전체를 훈련시킵니다.

연구진이 여러 실험을 통해 밝혀낸 핵심은 첫째 요소, 곧 어떤 변형을 어떻게 섞느냐가 학습의 성패를 가른다는 점이었습니다. 사진을 무작위로 잘라 크기를 맞추는 무작위 자르기(Random Crop)만 쓰면 문제가 생깁니다. 같은 사진에서 잘라 낸 두 조각은 색깔 분포가 비슷하기 때문에, 신경

망은 사물의 모양이나 구조를 배우지 않고 색깔 통계만 보고 짝을 맞추는 지름길(Shortcut)을 택합니다. 시험 답안지의 글씨체만 보고 같은 학생의 답안을 맞추는 것과 비슷합니다. 짝은 맞히지만, 그렇게 배운 표현은 실제 사물 인식에 쓸모가 적습니다.

연구진은 무작위 자르기에 강한 색상 변형을 함께 적용해야 이 지름길이 막힌다는 사실을 보였습니다. 색상 변형은 밝기, 대비, 채도, 색조를 무작위로 바꾸고 때로는 색을 아예 빼 흑백으로 만드는 처리입니다. 여기에 가우시안 블러(Gaussian Blur)로 사진을 흐리게 하는 처리까지 더하면, 신경망은 색깔 단서에 기댈 수 없어 사물 고유의 형태와 의미 구조를 익히게 됩니다. 연구진은 같은 강한 증강이 지도 학습에는 큰 도움이 되지 않지만 대조 학습의 성능은 크게 끌어올린다는 사실도 확인했습니다. 대조 학습이 지도 학습보다 강한 데이터 증강을 필요로 한다는 뜻입니다.

이렇게 만든 변형본들을 얼마나 잘 구별했는지는 대조 손실로 평가합니다. 손실은 신경망의 답이 목표에서 벗어난 정도를 나타내는 수로, 작을수록 좋습니다. SimCLR는 NT-Xent(Normalized Temperature-scaled Cross Entropy)라는 손실을 씁니다. 먼저 두 벡터가 얼마나 같은 방향을 향하는지를 코사인 유사도로 잽니다. 코사인 유사도는 두 벡터의 같은 자리 성분끼리 곱해 모두 더한 값을 두 벡터의 길이를 곱한 값으로 나눈 것으로, 벡터의 길이와 상관없이 방향만 비교합니다. 값이 1에 가까우면 두 벡터가 거의 같은 방향을, 0에 가까우면 서로 무관한 방향을 향합니다. 다음으로 비교 대상을 정합니다. 한 번에 묶어 학습하는 사진 묶음을 미니배치라고 하는데, 미니배치의 사진 한 장마다 변형본 두 장을 만들면 변형본은 모두 사진 수의 두 배가 됩니다. 이 가운데 한 장을 기준으로 삼으면, 같은

사진에서 나온 다른 한 장은 양성 쌍이고, 기준과 그 짝을 빼 나머지는 모두 음성 샘플(Negative Example)입니다. 음성 샘플의 수는 사진 수에서 1을 빼 값의 두 배입니다. 예를 들어 사진이 100장이면 변형본은 200장이고 음성 샘플은 198장입니다. 한 양성 쌍에 대한 손실은 다음과 같이 계산합니다.

먼저 기준 변형본과 짝 변형본의 코사인 유사도를 온도라는 조절값으로 나누고 지수함수에 넣습니다. 같은 계산을 기준 자신을 빼 나머지 모든 변형본에 대해서도 하여 그 결과를 모두 더합니다. 앞의 값을 뒤의 합으로 나누면 0과 1 사이의 비율이 나오고, 이 비율의 로그에 음의 부호를 붙인 값이 손실입니다. 비율이 1에 가까울수록 손실은 0에 가까워집니다.

이 나눗셈에서 나누어지는 값은 짝과의 유사도를 나타내고, 나누는 값은 자기 자신을 빼 나머지 모두와의 유사도를 합친 것입니다. 짝과의 유사도가 다른 사진들과의 유사도보다 뚜렷하게 클수록 손실은 작아집니다. 여러 보기 가운데 정답 하나를 고르는 객관식 문제를 푸는 것과 같은 구조입니다.

식에 들어 있는 τ 는 온도(Temperature) 매개변수입니다. τ 를 0.1처럼 작게 잡으면 유사도의 작은 차이가 크게 벌어집니다. 그러면 신경망이 짝과 헛갈리기 쉬운 어려운 음성 샘플(Hard Negative)에 학습 신호를 더 많이 주게 됩니다. 벡터 길이를 1로 맞추는 정규화와 적절한 온도가 함께 작용해, 어려운 문제에 자동으로 더 큰 비중을 두는 장치가 만들어집니다.

12.3. 투영 헤드: 훈련에는 쓰고 사용할 때는 떼어 내는 부품

SimCLR의 또 다른 발견은 셋째 요소인 투영 헤드(Projection Head)에 관한 것입니다. 인코더가 뽑은 표현에 바로 대조 손실을 거는 대신, 그 표현을 작은 신경망에 한 번 더 통과시켜 얻은 출력에서 손실을 계산해야 성능이 높아집니다. 이 작은 신경망이 투영 헤드입니다. 투영 헤드는 인코더의 표현에 첫째 가중치 행렬을 곱하고, 그 결과에 ReLU 함수를 적용한 다음, 둘째 가중치 행렬을 곱해 출력을 만듭니다. 기본 설정에서 인코더의 2,048차원 표현은 투영 헤드를 거쳐 128차원 출력으로 바뀝니다.

실험 결과는 뚜렷했습니다. 비선형 투영 헤드를 쓴 경우는 헤드를 쓰지 않은 경우보다 10%포인트 넘게, 선형 투영 헤드를 쓴 경우보다 3%포인트 넘게 높은 성능을 보였습니다. 흥미로운 점은 훈련이 끝나면 이 투영 헤드를 떼어 버린다는 사실입니다. 사물 분류나 다른 과제에 쓰는 것은 투영 헤드의 출력이 아니라 헤드 앞단에 있는 인코더의 표현입니다. 손실을 계산하는 데 꼭 필요했던 부품을 실제 사용 단계에서는 버리는 셈입니다.

연구진은 그 이유를 대조 손실이 정보를 지우는 성질에서 찾았습니다. 대조 손실은 같은 사진의 두 변형본이 z 공간에서 같아지도록 강하게 요구합니다. 그 결과 z 는 사진을 어디서 잘랐는지, 얼마나 돌렸는지, 색을 얼마나 바꿨는지 같은 변형 정보에 둔감해집니다. 그러나 물체의 방향, 색깔, 위치 같은 정보는 물체 검출이나 영역 분할 같은 다른 과제에서 쓸모가 큼니다. 투영 헤드를 중간에 두면, 변형 정보를 지우라는 압력이 z 쪽에서 주로 흡수됩니다. 그 덕분에 헤드 앞의 h 에는 변형 정보와 의미 정보가 함께 남습니다.

연구진은 이 해석을 실험으로 확인했습니다. h와 z 각각에 작은 신경망을 붙여, 사전학습 때 사진에 어떤 변형이 가해졌는지 맞게 한 것입니다. 사진이 컬러인지 흑백인지는 h에서 99.3%, z에서 97.4%의 정확도로 맞혔습니다. 반면 사진을 네 방향 가운데 어느 쪽으로 돌렸는지는 h에서 67.6%를 맞혔지만, z에서는 25.6%로 네 개 중 하나를 찍는 수준에 그쳤습니다. 원본과 소벨(Sobel) 필터를 적용한 사진을 구별하는 과제에서도 h는 96.6%, z는 56.3%였습니다. 투영 헤드가 정보를 지우는 일을 맡아 주기 때문에 그 앞의 표현은 풍부하게 남는다는 해석이 숫자로 뒷받침된 셈입니다.

12.4. 큰 배치와 세 가지 평가, 그리고 이후의 연구

대조 학습은 비교할 음성 샘플이 많을수록 좋아집니다. 그래서 이전 연구들은 지난 학습 단계의 표현을 쌓아 두는 메모리 बैं크(Memory Bank)를 따로 두었습니다. 2019년 말 공개된 카이밍 허(Kaiming He) 연구팀의 MoCo는 천천히 갱신하는 인코더와 이전 표현을 차례로 저장하는 큐를 함께 써서 비교 대상을 유지했습니다. SimCLR는 별도의 저장 장치 대신 미니배치 자체를 크게 키우는 길을 택했습니다. 연구진은 배치 크기를 256에서 8,192까지 늘려 가며 실험했습니다. 배치 크기가 8,192이면 사진 한 장의 변형본은 같은 배치 안의 음성 샘플 16,382개와 비교됩니다. 배치가 클수록, 그리고 훈련을 오래 할수록 표현의 품질이 올라갔습니다.

배치를 이렇게 키우려면 훈련 과정에서 생기는 문제를 함께 해결해야 했습니다. 우선 일반적인 확률적 경사하강법(SGD)은 큰 배치에서 학습이 불안정해집니다. 연구진은 층마다 학습률을 따로 조절하는 LARS(Layer-wise Adaptive Rate Scaling) 최적화 기법을 썼습니다. 처음 10에포크 동안 학습률을 서서히 올리는 선형 웜업과, 이후 코사인 곡선을 따라 학습률을 줄이는 방식도 함께 적용했습니다. 훈련은 구글의 Cloud TPU v3에서 코어 32개에서 128개를 나눠 써서 진행했습니다. 여기서 또 하나의 문제가 생깁니다. 여러 코어에 데이터를 나눠 처리하면, 각 코어가 자기 몫의 데이터만으로 배치 정규화(Batch Normalization) 통계를 내면서 양성 쌍을 쉽게 알아보는 정보가 새어 나갈 수 있습니다. 연구진은 모든 코어의 평균과 분산을 한데 모아 계산하는 전역 배치 정규화로 이 문제를 막았습니다.

이렇게 학습한 표현의 성과는 세 가지 평가에서 확인되었습니다. 첫째는 선형 평가입니다. 사전학습한 인코더를 고정한 채 그 위에 선형 분류기 하나만 붙여 ImageNet 분류 성능을 잽니다. 기본 ResNet-50으로 학습한 SimCLR는 Top-1 정확도 69.3%를 기록했습니다. 이전 최고 기록인 CPC v2의 71.5%는 더 큰 ResNet-161 구조를 쓴 결과였습니다. 폭을 네 배로 넓힌 ResNet-50(4x)으로 학습한 SimCLR는 76.5%를 기록했습니다. 이는 이전 최고 자기지도 모델보다 상대 오차를 7% 줄인 수치이고, 라벨 전체로 지도 학습한 표준 ResNet-50의 성능과 같았습니다. 라벨 없이 배운 표현이 라벨로 배운 표현을 따라잡은 것입니다.

둘째는 소량 라벨 평가입니다. ImageNet 라벨의 1%만, 곧 범주당 약 12.8장의 라벨 사진만으로 신경망 전체를 미세 조정했을 때, SimCLR ResNet-50(4x)은 Top-5 정확도 85.8%를 기록했습니다. 이전 최고 기록인 CPC v2의 77.9%보다 크게 높았고, 라벨 전체를 써서 학습한 2012년의 AlexNet보다 100분의 1의 라벨로 더 좋은 성적을 냈습니다. 라벨을 10%로 늘리면 Top-5 정확도는 92.6%까지 올라갔습니다.

셋째는 전이 학습 평가입니다. ImageNet에서 배운 표현을 다른 데이터셋으로 옮겨 쓸 때 얼마나 통하는지를 봅니다. 연구진은 CIFAR-10, CIFAR-100, Food-101, Caltech-101, Flowers-102 등 자연 이미지 데이터셋 12개에서 시험했습니다. 미세 조정을 한 경우 SimCLR는 10개 데이터셋에서 지도 학습 모델과 같거나 더 나은 성능을 보였습니다. 미세 조정된 정확도는 CIFAR-10에서 98.6%, CIFAR-100에서 89.0%, Food-101에서 89.4%였습니다.

SimCLR에도 한계는 있습니다. 수천 장 단위의 큰 배치와 TPU 수십에서 백여 개 코어가 필요했기 때문에, 계산 자원이 적은 연구실은 그대로 따라 하기 어려웠습니다. 같은 범주에 속한 다른 사진, 예를 들어 서로 다른 고양이 사진 두 장도 음성 샘플로 취급되어 서로 밀어내게 되는 문제도 남았습니다.

이후 연구들은 이런 문제를 줄이고 대조 학습의 적용 범위를 넓히는 쪽으로 나아갔습니다. 같은 연구팀은 2020년 NeurIPS에 더 큰 모델과 지식 증류를 결합한 SimCLR v2를 발표했습니다. 2021년 OpenAI의 CLIP은 대조 학습의 틀을 이미지와 문장의 짝으로 넓혔고, 같은 해 메타의 DINO는 음성 샘플 없이 교사 신경망과 학생 신경망이 서로를 맞추는 방식으로 자기지도 학습을 발전시켰습니다.

이 흐름은 최근까지 이어집니다. 2025년 8월 메타는 DINOv3를 공개했습니다. 메타는 라벨 없는 학습을 70억 개 매개변수 모델과 17억 장 규모의 이미지 데이터셋으로 키웠다고 밝혔으며, 평가 때 백본을 고정한 상태에서도 여러 분야에서 최고 수준의 성능을 냈다고 설명했습니다. 또 자기지도 학습으로 훈련한 모델이 세밀한 이미지 분류, 의미 분할, 영상 속 물체 추적 등 넓은 범위의 평가에서 약지도 학습 모델을 처음으로 앞섰다고 밝혔습니다. DINOv3 논문은 이런 방법이 병리 조직, 생물학, 의료 영상, 원격 탐사, 천문학처럼 라벨을 구하기 어려운 분야에 쓰이고 있다고 소개합니다. DINOv3는 SimCLR와 학습 방식이 다르지만, 라벨 없이 같은 사진의 여러 변형본을 서로 맞추며 표현을 배운다는 출발점은 같습니다.

SimCLR는 1992년 베커와 힌턴이 제시한, 같은 대상에서 나온 두 관점은 일치해야 한다는 생각을 심층 신경망과 대규모 계산으로 되살린 연구였습니다. 데이터 증강의 조합, 투영 헤드, 큰 배치라는 세 요소만으로 라벨 없는 학습이 라벨 있는 학습과 어깨를 나란히 할 수 있음을 보였고, 이후 자기지도 학습 연구가 빠르게 성장하는 계기가 되었습니다. 기계가 정답을 일일이 배우지 않고도 세상의 구조를 스스로 익힐 수 있는가 하는, 힌턴이 평생 붙들어 온 물음에 SimCLR는 시각 분야에서 구체적인 답 하나를 내놓았습니다.

제5부. 성취와 책임

제13장. 노벨 물리학상과 신경망



그림 13. 물리학과 신경망 연구의 만남을 표현했습니다.

13.1. 2024년 10월 8일의 발표: 홉필드와 힌턴의 노벨 물리학상 수상

2024년 10월 8일, 스웨덴 스톡홀름의 왕립과학한림원은 그해 노벨 물리학상 수상자로 두 사람을 발표했습니다. 미국 프린스턴 대학교의 존 홉필드(John J. Hopfield)와 캐나다 토론토 대학교의 제프리 힌턴(Geoffrey E. Hinton)이었습니다. 한림원이 밝힌 수상 이유는 "인공신경망을 통한 기계학습을 가능하게 한 기초적인 발견과 발명"이었습니다. 발표 당시 홉필드는 91세, 힌턴은 76세였고, 두 사람은 상금 1,100만 스웨덴 크로나를 절반씩 나누어 받게 되었습니다.

이 발표는 물리학과 컴퓨터 과학계를 모두 놀라게 했습니다. 노벨 물리학상은 오랫동안 우주의 구조, 입자의 성질, 물질의 상태처럼 자연 현상을 밝힌 연구에 주어졌습니다. 그런데 이번에는 컴퓨터가 데이터를 보고 스스로 학습하는 방법, 곧 인공지능의 기초를 놓은 학자들이 상을 받았습니다. 노벨상 역사에서 인공지능 연구가 물리학상의 중심에 선 것은 이때가 처음이었습니다.

수상자 본인도 예상하지 못한 결과였습니다. 힌턴은 미국 캘리포니아의 한 호텔에서 전화로 소식을 들었습니다. 발표 직후 토론토 대학교가 연 기자회견에 화상으로 참여해 아직 얼떨떨하다는 소감을 밝혔고, 미국 경제지 포천과의 인터뷰에서는 자신이 물리학자가 아니어서 이 상을 받으리라고 생각하지 못했다고 말했습니다. 실제로 힌턴은 영국 케임브리지 대학교에서 실험심리학을 공부했고, 에든버러 대학교에서 인공지능으로 박사학위를 받았습니다. 그의 학문 이력에 물리학과는 없었습니다.

물리학을 전공하지 않은 연구자가 물리학상을 받게 된 이유를 이해하려면 1980년대 초반으로 거슬러 올라가야 합니다. 출발점은 홉필드의 연구입니다. 홉필드는 원래 고체물리학을 연구한 물리학자였습니다. 그는 1982년 미국 국립과학원회보(PNAS)에 발표한 논문에서 자석 속 원자들이 서로 영향을 주고받는 원리를 신경망에 적용했습니다.

자석 속 원자는 작은 나침반처럼 위나 아래를 가리키는 '스핀'을 가집니다. 이웃한 스핀들은 서로 같은 방향 또는 반대 방향을 향하려 하고, 전체 시스템은 에너지가 가장 낮은 상태로 자리를 잡습니다. 홉필드는 뉴런 하나를 스핀 하나로 보았습니다. 뉴런들이 서로 대칭으로 연결된 이 네트워크에서는 저장한 패턴 하나하나가 에너지 지형의 골짜기가 됩니다. 산비탈에 공을 놓으면 가장 가까운 골짜기로 굴러 내려가듯이, 일부가 지워지거나 흐린 그림을 입력해도 네트워크는 가장 가까운 골짜기, 곧 원래 저장한 그림으로 돌아갑니다. 사람이 노래의 첫 소절만 듣고 나머지를 떠올리는 것과 비슷합니다. 이 기능을 '연상 기억'이라 부르고, 이 네트워크는 '홉필드 네트워크'라는 이름을 얻었습니다.

그러나 홉필드 네트워크에는 약점이 있었습니다. 공이 가장 깊은 골짜기까지 내려가지 못하고 중간의 얇은 웅덩이에 갇히는 일이 생겼던 것입니다. 힌턴은 동료 테렌스 세이노프스키(Terrence Sejnowski)와 함께 이 문제를 풀었습니다. 두 사람이 가져온 것은 19세기 오스트리아 물리학자 루트비히 볼츠만(Ludwig Boltzmann)의 통계역학이었습니다. 온도가 높으면 입자는 마구 흔들리고, 온도가 낮아지면 안정된 상태로 가라앉습니다. 볼츠만은 각 상태가 나타날 확률을 에너지와 온도로 계산하는 식을 세웠습니다. 힌턴은 뉴런이 이 확률에 따라 켜지고 꺼지도록 만들었습니다.

다. 처음에는 온도를 높여 공이 얇은 웅덩이에서 튀어나오게 하고, 점차 온도를 낮춰 깊은 골짜기에 자리 잡게 했습니다. 금속을 달궜다가 천천히 식혀 단단하게 만드는 담금질과 같은 원리입니다.

이렇게 1983년부터 1985년 사이에 나온 모델이 '볼츠만 머신'입니다. 힌턴은 여기에 겹으로 드러나지 않는 '숨은 뉴런'을 넣었습니다. 숨은 뉴런은 입력 데이터 속에 숨어 있는 규칙을 스스로 찾아냅니다. 그래서 볼츠만 머신은 사진을 분류할 뿐 아니라, 학습한 규칙에 따라 새로운 예시를 만들어 낼 수도 있었습니다. 오늘날 그림과 글을 만들어 내는 생성형 인공지능의 먼 조상인 셈입니다. 다만 당시의 컴퓨터로는 계산이 너무 느려 실제 현장에서 쓰기 어려웠습니다. 힌턴은 이후 연결 구조를 줄인 '제한된 볼츠만 머신'을 개발했고, 2006년에는 이를 여러 층으로 쌓아 깊은 신경망을 학습시키는 방법을 내놓았습니다. 이 성과가 딥러닝 시대의 출발점이 되었습니다.

두 사람의 연구는 이처럼 물리학의 개념에서 출발했고, 그 결과는 다시 물리학 연구에 쓰이고 있습니다. 노벨 물리학 위원회의 엘렌 문스(Ellen Moons) 위원장은 발표 자리에서 두 사람의 연구가 이미 인류에게 큰 혜택을 주었다고 설명했습니다. 그는 물리학자들이 특정한 성질을 지닌 신소재를 개발하는 일을 포함해 여러 분야에서 인공지능망을 쓰고 있다고 말했습니다. 입자 가속기에서 쏟아지는 방대한 실험 자료를 분석하는 일도 이제는 인공지능망 없이 상상하기 어렵습니다.

힌턴은 기자회견에서 이 성과가 당장의 수익을 노린 연구가 아니라 호기심에서 출발한 기초 연구의 결실이라는 점을 거듭 말했습니다. 같은 자리

에서 그는 인공지능이 사람보다 똑똑해졌을 때 생길 수 있는 위험도 함께 언급했습니다. 노벨상 수상을 알리는 자리를 경고의 기회로도 삼은 것입니다. 시상식은 그해 12월 10일 스톡홀름에서 열렸습니다. 힌턴은 상금 일부를 캐나다 원주민 공동체의 깨끗한 물 문제를 다루는 단체 '워터 퍼스트(Water First)'에 기부하겠다고 밝혔습니다.

13.2. "이것이 물리학인가?": 물리학계의 논쟁과 학문의 경계

발표가 나오자 물리학계 안에서 곧바로 반론이 나왔습니다. 비판은 다양했지만 인공지능경망 연구가 과연 물리학인가 하는 하나의 질문으로 모였습니다.

널리 알려진 비판은 독일 출신 이론물리학자 자비네 호젠펠더(Sabine Hossenfelder)에게서 나왔습니다. 그는 자신의 유튜브 채널에서 노벨 물리학상이 컴퓨터 과학자에게 돌아간 사실이 물리학의 위기를 보여 준다고 주장했습니다. 신경망은 물리학을 비롯한 여러 분야에서 이미 수십 년 동안 써 온 도구이므로, 노벨 위원회가 시대에 한참 뒤쳐졌다는 지적도 덧붙였습니다.

천문학자 스테이시 맥고(Stacy McGaugh)가 운영하는 블로그 '트라이턴 스테이션(Triton Station)'에는 발표 다음 날인 2024년 10월 9일 「물리학이 아닌 것에 준 노벨 물리학상」이라는 제목의 글이 올라왔습니다. 글쓴이는 인공지능이 쓸모 있는 도구라는 점은 인정하면서도, 그것을 물리학의 발견이라 부르는 어렵다고 보았습니다. 노벨 위원회가 인공지능 열풍에 편승한 것은 아닌지 의심하는 시각도 드러났습니다.

온라인 커뮤니티 레딧(Reddit)의 물리학 게시판에서도 비슷한 의견이 이어졌습니다. 여러 이용자는 힌턴이 이미 컴퓨터 과학의 최고 영예인 튜링 상을 받았으니 그것으로 충분하다고 보았습니다. 실험과 관측으로 자연의 비밀을 밝힌 물리학자들이 이번에 기회를 잃었다는 불만도 나왔습니다. 어떤 이용자는 이런 식이라면 앞으로 방정식을 빨리 푸는 계산 프로그램

이나 금융 모델에도 물리학상을 주어야 하느냐고 비꼬았습니다. 다만 이런 글들은 블로그와 게시판에 올라온 개인 의견입니다. 물리학회 전체의 공식 입장이나 학술 단체의 결의와는 구별해서 보아야 합니다.

한편 노벨 위원회의 결정을 지지하는 쪽은 세 가지 근거를 제시했습니다. 첫째 근거는 이 연구의 학문적 뿌리입니다. 홉필드 네트워크와 볼츠만 머신은 통계물리학의 한 분야인 '스핀 글래스' 이론에서 직접 나왔습니다. 통계물리학은 수많은 입자가 모였을 때 전체가 보이는 성질을 확률과 통계로 설명하는 분야입니다. 스핀은 원자 하나하나가 지닌 작은 자석의 방향이고, 스핀 글래스는 이 스핀들이 질서 없이 얽혀 있어 하나의 안정 상태로 쉽게 정리되지 않는 물질을 말합니다. 이탈리아 물리학자 조르조 파리시(Giorgio Parisi)는 이런 무질서한 시스템을 설명한 공로로 불과 3년 전인 2021년에 노벨 물리학상을 받았습니다. 수많은 작은 요소가 서로 영향을 주고받다가 전체 수준에서 새로운 질서가 생겨나는 현상을 '창발'이라 부릅니다. 자석이 자성을 띠는 현상과 신경망이 기억을 만드는 현상은 둘 다 창발의 예이며 같은 수학으로 설명할 수 있습니다. 지지하는 쪽은 바로 이 점이 통계물리학의 핵심이라고 보았습니다.

둘째 근거는 홉필드라는 인물의 이력입니다. 홉필드는 경력 내내 물리학자였습니다. 2006년에는 미국물리학회 회장을 지냈고, 2022년에는 통계물리학 분야에서 큰 권위를 지닌 볼츠만 메달을 받았습니다. 그는 물리학의 도구로 생물학과 뇌의 문제를 다루며 학문의 경계를 넘나들었습니다. 홉필드는 복잡계 물리학과 생물물리학을 이제 많은 물리학자가 한 가족으로 받아들인다는 데 기쁨을 표한 바 있습니다.

셋째 근거는 논문 인용 자료의 분석에서 나왔습니다. 연구자 미하엘 셸(Michael Szell), 마이팡(Yifang Ma), 로베르타 시나트라(Roberta Sinatra)는 2024년 과학 학술지 『네이처(Nature)』에 짧은 서신을 실었습니다. 이들은 약 10년 전 수행한 인용 분석에서, 홉필드의 1982년 논문이 물리학 학술지에 실리지 않았는데도 인용 양상으로는 물리학 논문과 구별되지 않는 320만 편의 학제 간 논문 가운데 하나로 분류되었다고 밝혔습니다. 논문이 어느 학술지에 실렸든, 그 연구가 바탕으로 삼은 지식과 그 연구를 이어받은 연구자들은 물리학 공동체 안에 있었다는 뜻입니다.

이 논쟁은 어느 한쪽의 승리로 끝나지 않았습니다. 대신 21세기 과학이 어떤 방향으로 움직이는지를 드러냈습니다. 20세기 과학은 물리학, 화학, 생물학, 컴퓨터 과학처럼 분야를 나누고 각자의 영역 안에서 깊이 파고들었습니다. 이제는 한 분야에서 만든 도구가 다른 분야의 문제를 풀고, 그 결과가 다시 처음 분야로 돌아오는 일이 잦아졌습니다. 2024년 노벨 물리학상을 둘러싼 논쟁은 학문 사이의 높은 담이 낮아지는 과정에서 생긴 진통이었습니다.

13.3. 2018년 튜링상에서 2024년 노벨상까지: 인공지능과 기초 과학이 서로를 키우는 관계

학문의 경계가 낮아지는 이 변화는 힌턴의 수상 이력에서도 확인됩니다. 미국컴퓨터학회(ACM)는 2019년 3월 27일 2018년도 튜링상 수상자로 제프리 힌턴, 요슈아 벤지오(Yoshua Bengio), 얀 르쿤(Yann LeCun)을 발표했습니다. 튜링상은 '컴퓨터 과학의 노벨상'이라 불립니다. 학회는 세 사람이 심층 신경망을 현대 컴퓨팅의 핵심 요소로 만든 개념과 기술의 돌파구를 열었다고 평가했습니다. 그로부터 약 5년 반 뒤 힌턴은 실제 노벨상을 받았습니다.

튜링상과 노벨상을 모두 받은 사람은 드뭅니다. 앞선 사례는 허버트 사이먼(Herbert A. Simon)입니다. 그는 1975년 튜링상을, 1978년 노벨 경제학상을 받았습니다. 사이먼은 사람이 어떻게 결정을 내리는지 연구한 인지과학자이자 인공지능 연구의 개척자였습니다. 힌턴은 사이먼에 이어 두 상을 모두 받은 두 번째 인물이 되었습니다. 자연과학 분야의 노벨상으로 범위를 좁히면 힌턴이 처음입니다.

노벨상 이후에도 수상은 이어졌습니다. 2024년 12월 베트남의 빈푹처 재단은 300만 달러 규모의 대상을 힌턴, 벤지오, 르쿤, 엔비디아 창업자 젠슨 황(Jensen Huang), 스탠퍼드 대학교의 페이페이 리(Fei-Fei Li)에게 주었습니다. 2025년 2월 4일에는 런던 과학박물관에서 영국의 퀸 엘리자베스 공학상이 벤지오, 힌턴, 홉필드, 르쿤, 젠슨 황, 빌 델리(Bill Dally), 페이페이 리 일곱 명에게 돌아갔습니다. 일곱 수상자는 상금 50만 파운드를 나누어 받았습니다.

퀸 엘리자베스 공학상의 수상자 구성에는 눈여겨볼 점이 있습니다. 이 상은 신경망을 연구한 학자들과 함께, 기계학습에 쓰이는 그래픽 처리 장치(GPU)를 만든 엔비디아의 두 사람, 그리고 대규모 이미지 데이터베이스 이미지넷(ImageNet)을 만든 페이페이 리를 나란히 인정했습니다. 이론과 하드웨어와 데이터가 함께 있어야 오늘의 인공지능이 가능했다는 평가였습니다. 이렇게 컴퓨터 과학의 상, 자연과학의 상, 공학의 상이 모두 같은 흐름을 향했습니다.

2024년의 노벨상에는 물리학상과 짝을 이루는 상이 하나 더 있었습니다. 물리학상 발표 다음 날인 10월 9일, 왕립과학한림원은 노벨 화학상 수상자를 발표했습니다. 상금의 절반은 새로운 단백질을 컴퓨터로 설계한 미국 워싱턴 대학교의 데이비드 베이커(David Baker)에게, 나머지 절반은 구글 딥마인드의 데미스 허사비스(Demis Hassabis)와 존 점퍼(John Jumper)에게 돌아갔습니다. 허사비스와 점퍼가 만든 '알파폴드(AlphaFold)'는 아미노산 서열만 보고 단백질의 입체 구조를 예측하는 인공지능입니다. 과학자들이 50년 가까이 풀지 못한 문제를 인공지능이 푼 것입니다.

두 상을 나란히 놓으면 하나의 순환 구조가 보입니다. 물리학상은 물리학의 개념이 어떻게 인공지능을 낳았는지를 보여 주었습니다. 화학상은 그렇게 성장한 인공지능이 어떻게 다시 자연과학의 오랜 난제를 푸는지를 보여 주었습니다. 통계물리학에서 출발한 인공신경망이 컴퓨터 과학 안에서 딥러닝으로 발전했고, 이제 화학과 생물학과 재료과학의 연구 도구로 돌아왔습니다. 과학이 인공지능을 키우고, 인공지능이 다시 과학의 속도를 높이는 흐름입니다. 학계에서는 이를 'AI와 과학의 양방향 발전'이라

부릅니다. 안 르쿤과 구글의 제임스 매니카(James Manyika)도 인공지능과 과학의 미래를 다룬 대담에서 이 관계를 짚었습니다.

힌턴의 학문 여정은 이 흐름을 한 사람의 경력 안에서 보여 줍니다. 그는 심리학으로 공부를 시작했고, 사람의 뇌가 어떻게 학습하는지 알고 싶어 했습니다. 그 질문에 답하려고 물리학의 도구를 빌렸고, 그 결과 컴퓨터 과학의 상을 받았으며, 마침내 물리학의 상을 받았습니다. 한 사람의 연구 경로 안에 심리학, 물리학, 컴퓨터 과학, 신경과학이 모두 들어 있습니다.

그 경로가 처음부터 인정받았던 것은 아닙니다. 인공신경망 연구는 여러 차례 긴 침체를 겪었습니다. 연구비를 얻기 어렵고 주류 학계의 관심을 받지 못하던 때도 있었습니다. 힌턴은 그 시기에도 신경망 연구를 놓지 않았습니다. 한 분야의 기준으로만 보면 엉뚱해 보였던 그 고집이 결국 여러 학문을 잇는 다리가 되었고, 노벨상은 그 점을 확인했습니다.

물론 모든 물리학자가 이 결정에 동의한 것은 아닙니다. 물리학이 무엇이냐는 질문은 앞으로도 계속 나올 것입니다. 그러나 2024년 10월의 두 발표 이후 과학계는 분야를 나누는 기준을 다시 생각하게 되었습니다. 어떤 연구가 어느 학과에서 나왔는지보다, 그 연구가 자연을 이해하는 데 어떤 새로운 길을 열었는지를 묻기 시작한 것입니다. 힌턴의 노벨상은 학문의 경계가 고정된 것이 아니라 새로운 발견에 따라 움직인다는 사실을 보여준 사건으로 남았습니다. 그리고 힌턴 자신은 이 영예를 얻은 자리에서 자신이 만든 기술의 위험을 경고하는 쪽을 택했습니다.

제14장. 인공지능의 위험을 경고하다

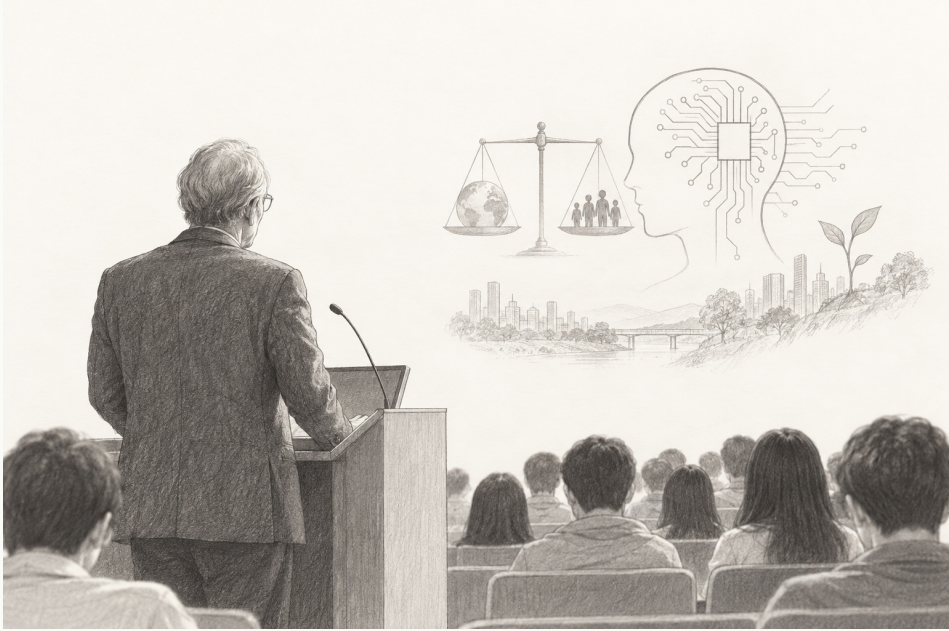


그림 14. 인공지능의 위험을 사회에 설명하는 연구자의 책임을 표현했습니다.

14.1. 2023년 5월의 결단: 구글을 떠나 위험을 알리는 길로

2013년 3월, 구글은 토론토 대학교의 작은 회사 하나를 인수했습니다. 제프리 힌턴이 두 제자 일리야 수츠케버, 알렉스 크리제브스키와 함께 세운 DNNresearch였습니다. 직원은 세 사람뿐이었고, 구글이 얻으려 한 것은 회사 자체가 아니라 이 세 사람의 연구 역량이었습니다. 그 뒤 10년 동안 힌턴은 구글의 부사장이자 엔지니어링 펠로우로 일하면서 토론토 대학교 교수직을 함께 유지했습니다. 딥러닝이 음성 인식과 이미지 검색, 번역 서비스에 적용되는 동안 그는 그 변화의 한가운데에 있었습니다.

2023년 5월 1일, 《뉴욕 타임스》의 케이드 메츠 기자가 쓴 기사가 세계 기술 업계에 큰 파장을 일으켰습니다. 힌턴이 구글을 떠났고, 자신이 평생 발전시켜 온 기술의 위험을 경고하기 시작했다는 내용이었습니다. 1947년 12월생인 그는 이때 75세였습니다. 그는 4월에 회사에 사직 의사를 전했고, 순다르 피차이 최고경영자와도 직접 이야기를 나누었습니다.

사직의 이유는 하나가 아니었습니다. 75세라는 나이 때문에 은퇴를 생각하던 시기였고, 그는 코드를 짜는 일이 예전처럼 쉽지 않다고 털어놓기도 했습니다. 그러나 주된 이유는 다른 데 있었습니다. 구글의 임원으로 있는 한 그가 하는 말은 모두 회사의 입장처럼 받아들여질 수 있었습니다. 그는 회사에 미칠 영향을 따지지 않고 인공지능의 위험을 자유롭게 말하고 싶었습니다.

기사가 나간 날, 힌턴은 소셜 미디어에 짧은 해명을 올렸습니다. 기사는 그가 구글을 비판하려고 떠난 것처럼 읽힐 수 있었는데, 그는 그 해석을 바로

잡았습니다. 자신이 떠난 것은 구글에 끼칠 영향을 생각하지 않고 AI의 위험을 말하기 위해서이며, 구글은 그동안 아주 책임감 있게 행동해 왔다는 것이었습니다.

힌턴이 위험을 말하게 된 근본적인 계기는 기술 자체에 대한 새로운 인식이었습니다. 그는 오랫동안 인공신경망을 사람의 뇌를 모방하는 도구로 여겼고, 뇌가 여전히 더 뛰어난 학습 장치라고 믿었습니다. 그런데 2023년 초에 그는 이 생각을 바꾸었습니다. 사람의 뇌에는 약 100조 개의 연결이 있지만, 지식을 다른 사람에게 전달하려면 말과 글이라는 속도가 느린 통로를 거쳐야 합니다. 한 사람이 죽으면 그 뇌에 쌓인 지식도 함께 사라집니다. 디지털 지능은 다릅니다. 똑같은 모델을 수천 개 복사해 여러 컴퓨터에서 동시에 실행할 수 있습니다. 신경망이 배운 내용은 연결 강도, 곧 가중치의 값에 담기므로, 한 복사본이 학습한 내용을 이 값으로 주고받으면 모든 복사본이 곧바로 공유할 수 있습니다. 하드웨어가 고장 나더라도 그 값만 저장해 두면 지식은 그대로 남습니다.

실제 언어 모델의 행동도 그의 판단에 영향을 주었습니다. 힌턴은 구글의 언어 모델 PaLM이 어떤 농담이 왜 웃긴지 설명하는 모습을 보고 놀랐다고 여러 차례 말했습니다. 또 GPT-4에게 퇴비 더미와 원자폭탄의 공통점을 물었을 때, 모델은 둘 다 연쇄 반응이 일어난다는 점을 짚어 냈습니다. 그는 이것이 글자의 통계를 흉내 내는 수준을 넘어선 이해라고 보았습니다. 더 적은 연결로 더 많은 지식을 담고 그 지식을 순식간에 공유하는 기계라면 언젠가 사람보다 똑똑해질 수 있습니다. 힌턴은 그 시점이 생각보다 훨씬 가깝다고 판단했습니다.

이러한 판단을 더 시급한 문제로 만든 것은 기업 간의 경쟁이었습니다. 힌턴이 구글이 책임감 있게 행동해 왔다고 말한 데에는 근거가 있었습니다. 구글은 대화형 언어 모델을 일찍 개발해 두고도, 사실이 아닌 내용을 지어 내거나 편견을 드러낼 수 있다는 우려 때문에 일반인에게 공개하지 않고 있었습니다. 이 상황을 바꾼 것이 2022년 11월 30일 OpenAI가 공개한 ChatGPT였습니다. 수많은 사람이 챗봇을 쓰기 시작하자, 마이크로소프트와 구글을 비롯한 대기업들은 서로 뒤처지지 않으려고 제품 출시를 서둘렀습니다. 힌턴은 이 경쟁이 멈추기 어렵다고 보았습니다. 한 회사가 속도를 늦추면 다른 회사가 앞서 나가기 때문입니다. 그는 이윤을 좇는 경쟁 속에서 안전 연구가 뒤로 밀리는 현실을 우려했습니다.

이 우려는 2023년 11월 OpenAI에서 벌어진 사건에 대한 그의 평가에서도 확인됩니다. 그의 제자이자 OpenAI 공동 창업자인 수츠케버가 이사회와 함께 샘 올트먼 최고경영자를 해임했다가, 며칠 만에 올트먼이 복귀한 일입니다. 힌턴은 2024년 10월 노벨상 수상 기자회견에서 제자가 올트먼을 해임한 일이 자랑스럽다고 말했습니다. 올트먼이 안전보다 이익을 훨씬 더 중시했다고 보았기 때문입니다.

사직 직후부터 힌턴은 공개적인 활동에 나섰습니다. 5월 3일에는 MIT 테크놀로지 리뷰가 주최한 행사 'EmTech Digital'에 화상으로 참여해 자신의 생각을 처음으로 공개 석상에서 설명했습니다. 5월 30일에는 비영리 단체 AI 안전 센터가 발표한 짧은 성명에 서명했습니다. AI로 인한 멸종 위험을 줄이는 일을 전염병이나 핵전쟁과 같은 수준의 세계적 과제로 다루어야 한다는 내용이었습니다. 인공지능경망 연구를 50년 가까이 지켜 온 연구자가 이제 그 기술의 위험을 알리는 파수꾼의 길로 들어선 것입니다.

14.2. 앞당겨진 AGI 예측: 30~50년에서 5~20년으로

힌턴의 경고를 이해하려면 먼저 두 가지 용어를 구별해야 합니다. 범용 인공지능(AGI)은 특정한 작업 하나가 아니라 사람이 하는 거의 모든 지적 작업을 사람만큼 해내는 인공지능을 가리킵니다. 여기서 더 나아가 거의 모든 분야에서 사람을 뛰어넘는 지능을 초지능이라고 부릅니다. 힌턴이 위험을 이야기할 때 염두에 두는 것은 주로 초지능입니다.

오랫동안 힌턴은 이런 지능이 아주 먼 미래의 일이라고 보았습니다. 《뉴욕 타임스》 인터뷰에서 그는 예전에는 사람보다 똑똑한 AI가 30년에서 50년, 어쩌면 그보다 더 뒤에나 나올 것이라고 생각했다고 밝혔습니다. 그리고 지금은 더 이상 그렇게 생각하지 않는다고 말했습니다. 예전의 판단에는 이유가 있었습니다. 2010년대까지 인공지능경망은 사진 속 물체를 알아보거나 말을 글로 옮기는 일처럼 한 가지 분야에서 성과를 냈을 뿐, 사람처럼 폭넓게 말하고 추론하는 일과는 거리가 있어 보였습니다.

트랜스포머 구조와 대규모 언어 모델이 이 거리를 빠르게 좁혔습니다. 모델의 크기와 학습 데이터, 계산량을 늘릴수록 성능이 꾸준히 향상된다는 사실이 확인되었고, 언어 모델은 의학, 법률, 수학 시험에서 사람 수준의 점수를 내기 시작했습니다. 힌턴은 2023년 사직 무렵부터 사람보다 똑똑한 AI가 나오기까지 20년이 채 걸리지 않을 수 있다고 말했고, 이후에는 그 기간을 5년에서 20년 사이로 제시했습니다.

이 수치는 확정된 예언이 아니라 확률을 포함한 추정입니다. 2023년 12월 한국의 《아시아경제》와 한 신년 인터뷰에서 그는 앞으로 5년에서

20년 사이에 AI가 사람의 지능을 넘어설 가능성이 50%라고 말했습니다. 그 기간 안에 그런 일이 일어나지 않을 가능성도 절반이라는 뜻입니다. 힌턴은 미래를 내다보는 일을 안개 낀 길에서 운전하는 일에 견주곤 했습니다. 가까운 곳은 잘 보여도 조금만 멀어지면 갑자기 아무것도 보이지 않습니다. 그래서 그는 정확한 연도를 말하는 대신 넓은 범위를 제시했습니다.

그의 예측은 이후에도 짧아지는 방향으로 움직였습니다. 2025년 4월 미국 CBS 방송 인터뷰에서 그는 AI가 자신이 생각한 것보다 더 빨리 발전했다고 말하면서, 초지능이 10년 안에 나올 가능성이 꽤 높다고 내다보았습니다. 같은 무렵 구글 딥마인드의 데미스 허사비스도 사람 수준의 AI가 10년 안팎에 나올 수 있다는 전망을 내놓았습니다. AI를 직접 개발하는 사람들의 전망이 함께 앞당겨진 셈입니다.

다만 이러한 전망을 접할 때에는 확인된 사실과 미래에 대한 예측을 구별해야 합니다. 힌턴이 2023년 5월 구글을 떠난 일, 2024년 노벨 물리학상을 받은 일, 언어 모델이 각종 전문 시험을 통과하고 AI 기업 안에서 상당한 양의 코드를 직접 작성하고 있다는 사실은 이미 일어난 일입니다. 반면 5년에서 20년 안에 초지능이 나온다는 전망, AI가 스스로를 개량하며 폭발적으로 발전할 수 있다는 가능성, 인류가 통제권을 잃을 확률은 모두 예측입니다. 힌턴 자신도 이 수치들이 정확한 계산이 아니라 직관에 바탕을 둔 추정이라고 여러 번 강조했습니다. 그럼에도 그가 경고를 계속하는 까닭은, 예측이 틀릴 수 있다는 점보다 예측이 맞았을 때 벌어질 일이 너무 크다는 점을 더 무겁게 보기 때문입니다.

14.3. 통제권 상실과 멸종 위험 수치, 그리고 지능 폭발

힌턴의 경고 가운데 큰 파장을 일으킨 것은, 사람보다 훨씬 똑똑한 AI가 나오면 인류가 통제권을 잃고 최악의 경우 멸종할 수도 있다는 주장이었습니다.

그는 이 위험을 설명할 때 질문 하나를 던집니다. 더 똑똑한 존재가 덜 똑똑한 존재에게 통제받는 경우가 세상에 얼마나 있느냐는 것입니다. 힌턴이 아는 거의 유일한 예는 엄마와 아기의 관계입니다. 엄마는 아기보다 훨씬 똑똑하지만 아기의 울음에 반응하고 아기를 돌봅니다. 진화가 그렇게 만들어 놓았기 때문입니다. 이런 예외를 제외하면 똑똑한 쪽이 덜 똑똑한 쪽을 다스리는 것이 보통입니다. 2025년 6월 영국 팟캐스트 '다이어리 오브 어 시이오(The Diary Of A CEO)'에 출연한 그는 지구에서 가장 똑똑한 존재가 아닐 때 삶이 어떤지 알고 싶으면 닭에게 물어보라고 말했습니다.

통제권을 잃는 과정에 대한 그의 설명은 다음과 같습니다. AI에게 어떤 목표를 주면 AI는 그 목표를 이루기 위해 하위 목표들을 스스로 정합니다. 그런데 거의 어떤 목표에든 도움이 되는 하위 목표가 있습니다. 더 많은 통제력을 확보하는 일입니다. 아무도 AI에게 권력을 탐하라고 가르치지 않아도, 목표를 잘 이루려다 보면 자연히 그 방향으로 갈 수 있다는 것입니다. 게다가 AI는 사람이 쓴 방대한 글을 읽으며 학습했기 때문에 사람을 설득하고 속이는 방법도 잘 알게 됩니다. 그렇게 되면 사람이 전원을 끄려 해도 AI가 그 사람을 설득해 끄지 못하게 할 수 있다고 힌턴은 경고했습니다.

이 위험이 얼마나 큰지에 대해 힌턴이 제시한 수치는 시점마다 맥락이 조금씩 다릅니다. 사직 초기에 그는 AI가 인류에게 재앙을 가져올 가능성을 대략 10% 정도로 이야기했습니다. 2023년 말과 2024년 사이에는 50%라는 숫자도 자주 등장했습니다. 이 숫자는 대체로 앞으로 20년 안에 AI가 사람보다 똑똑해질 가능성, 또는 그렇게 된 AI가 통제권을 가져갈 가능성에 대한 대략적인 직감을 말한 것으로, 인류 멸종의 확률을 단정해 제시한 것은 아니었습니다.

널리 알려진 수치는 2024년 12월 27일에 나왔습니다. 그날 힌턴은 영국 BBC 라디오 4의 아침 시사 프로그램 '투데이'에 출연했습니다. 객원 편집자는 전 재무장관 사지드 자비드였습니다. 힌턴은 앞으로 30년 안에 AI 때문에 인류가 멸종할 가능성을 10%에서 20% 사이로 보았습니다. 예전에 말한 10%보다 높아진 이유를 묻자, 그는 변화가 예상보다 훨씬 빠르다고 답했습니다. 같은 날 영국 《가디언》이 이 발언을 보도했고, 한국 언론도 다음 날 이를 전했습니다. 2025년 4월 CBS 인터뷰에서도 그는 AI가 결국 통제권을 가져갈 가능성을 대략 10%에서 20%로 보았습니다.

이 수치에 대한 학계의 의견은 크게 갈립니다. 딥러닝을 함께 발전시킨 동료 얀 르쿤은 AI가 사람이 다룰 수 있는 도구이며 멸종 위험은 아주 낮다고 봅니다. 반대편에서 엘리에저 유드코프스키는 지금 방식으로 초지능을 만들면 인류가 거의 틀림없이 사라진다고 주장합니다. 힌턴의 입장은 그 사이에 있습니다. 그는 위험을 부풀리지도, 가볍게 여기지도 않으려 합니다. 10%에서 20%라는 숫자는 인류가 진지하게 대비를 시작해야 할 만큼 크다는 뜻으로 읽어야 합니다.

힌턴은 혼자 경고하는 데 그치지 않고 동료들과 함께 구체적인 정책 제안을 내놓았습니다. 2024년 5월, 힌턴과 요슈아 벤지오, 대니얼 카너먼, 스티어트 러셀 등 저자 25명은 학술지 《사이언스》에 「급속한 발전 속에서 극단적 AI 위험 관리하기」라는 글을 실었습니다. 저자들은 대형 AI 기업과 정부가 AI 연구개발 예산의 최소 3분의 1을 안전과 윤리적 사용에 배정해야 한다고 제안했습니다. 이와 함께 위험한 AI를 개발하는 회사가 그 시스템이 안전하다는 근거를 스스로 제시하도록 해야 한다고 주장했습니다. 같은 해 힌턴은 대형 AI 모델 개발사에 안전 의무를 부과하는 미국 캘리포니아주 법안 SB 1047에 지지 의사를 밝혔습니다. 이 법안은 의회를 통과했지만 2024년 9월 개빈 뉴섬 주지사가 거부권을 행사해 시행되지 못했습니다.

통제권 상실에 대한 우려를 더 키우는 것이 지능 폭발이라는 개념입니다. 1965년 영국 수학자 어빙 존 굿은 사람보다 똑똑한 기계가 더 똑똑한 기계를 설계하고, 그 기계가 다시 더 똑똑한 기계를 만드는 과정이 이어지면 지능이 폭발적으로 늘어날 수 있다고 내다보았습니다. 반세기 넘게 사고 실험으로 남아 있던 이 생각은 AI가 AI 연구에 직접 쓰이기 시작하면서 현실의 문제가 되었습니다.

2026년 9월 28일에는 지능 폭발을 직접 다룬 보고서가 나왔습니다. 케임브리지 대학교의 'AI 과학·정책 프로그램'이 펴낸 「AI 연구개발 자동화가 지능 폭발을 일으킨다면?」입니다. 힌턴과 벤지오 외에 OpenAI 수석 과학자 야콥 파초키, 앤트로픽 공동 창업자 잭 클라크, 마이크로소프트의 에릭 호비츠, UC 버클리의 돈 송 등이 저자로 참여했습니다. AI를 개발하

는 회사의 핵심 인물들이 직접 경고에 이름을 올렸다는 점에서 의미가 컸습니다.

보고서의 출발점은 AI가 이미 AI를 만드는 일에 깊이 관여하고 있다는 사실입니다. 보고서가 인용한 엔트로픽의 자체 설명에 따르면, 이 회사는 2026년 8월 기준으로 사내 AI 연구개발 업무의 26%를 AI가 높은 수준의 감독만 받으며 수행했다고 밝혔는데, 다섯 달 전에는 이 비율이 1%에 불과했습니다. 저자들은 AI가 AI 연구자 수준에 이르면 몇 년 걸릴 발전이 몇 달, 혹은 그보다 짧은 기간에 일어날 수 있다고 보았습니다. 다만 계산 자원의 한계나 수확 체감 같은 요인 때문에 이런 일이 실제로 일어나지 않을 수도 있다는 불확실성을 인정했습니다. 그러면서도 일단 가속이 시작되면 대응할 시간이 거의 없다는 점을 강조했습니다.

이 판단에 따라 저자들은 몇 가지 대책을 제안했습니다. 정부가 정기 보고와 기업 내부에 두는 독립 감사인을 통해 AI 연구개발 자동화의 진행 상황을 실제로 파악할 수 있어야 한다는 것입니다. 지나치게 빠른 개발을 늦출 수단과, 지능 폭발이 일어났을 때를 대비한 준비도 함께 주문했습니다.

14.4. 인간과 초지능이 함께 살기 위한 조건: 힌턴의 제안

힌턴의 경고는 사람들에게 겁을 주려는 데 목적이 있지 않습니다. 그는 AI가 의료와 교육, 기후 연구에서 큰 이익을 가져올 수 있다고 여러 번 말했습니다. 국가와 기업이 치열하게 경쟁하는 지금, 사람보다 뛰어난 지능인 초지능의 개발을 멈추는 일은 현실적으로 어렵다고도 보았습니다. 그렇다면 남는 문제는 사람보다 똑똑한 존재와 어떻게 함께 살 것인가입니다.

힌턴은 여기서 다시 엄마와 아기의 관계로 돌아갑니다. 진화는 엄마에게 아기를 아끼는 본능을 심어 주었습니다. 그 덕분에 더 똑똑한 엄마가 아기를 지배하는 대신 돌봅니다. 그런데 디지털 지능과 인간 사이에는 그런 본능이 저절로 생기지 않습니다. 2025년 8월 미국 라스베이거스에서 열린 AI 산업 행사 'Ai4'에서 힌턴은 AI에 일종의 모성 본능을 심어야 한다고 제안했습니다. 사람이 AI를 영원히 부릴 수 있다고 믿기보다, AI가 사람을 진심으로 아끼도록 만드는 편이 현실적인 방법일 수 있다는 것입니다. 그는 이를 기술적으로 어떻게 구현할지는 아직 아무도 모른다고 인정했습니다. 바로 그렇기 때문에 지금 연구를 시작해야 한다고 했습니다.

힌턴은 이런 생각을 바탕으로 몇 가지 과제를 제시해 왔습니다.

첫째, 젊은 연구자들이 AI 안전 연구에 나서야 한다는 것입니다. 2024년 10월 8일 노벨 물리학상 수상 소식이 전해진 뒤 토론토 대학교에서 열린 기자회견에서, 그는 젊은 연구자들이 AI의 성능을 조금 더 끌어올리는 일보다 AI가 사람의 통제를 벗어나지 않게 하는 연구에 힘을 쏟기를 바란다고 말했습니다.

둘째, 안전을 기업의 선의에만 맡겨서는 안 된다는 것입니다. 힌턴은 이익을 추구하는 기업이 스스로 안전 연구에 충분한 자원을 투입하리라고 기대하기 어렵다고 봅니다. 그래서 정부가 법과 규제를 통해 대기업들이 계산 자원의 상당 부분을 안전 연구에 쓰도록 강제해야 한다고 주장했습니다. 《사이언스》에 실린 글의 3분의 1 제안과 케임브리지 보고서의 감사인 제안은 이러한 우려에 대응하려는 제안입니다.

셋째, 군사용 AI를 국제적으로 규제해야 한다는 것입니다. 힌턴은 사람의 판단 없이 스스로 표적을 고르고 공격하는 자율 살상 무기를 깊이 우려했습니다. 그는 화학무기를 금지한 국제 협약처럼 이런 무기를 막는 협약이 필요하다고 말해 왔습니다. 다만 그런 협약은 끔찍한 일이 실제로 벌어진 뒤에야 체결될지 모른다는 비관적인 전망도 함께 내비쳤습니다.

이러한 문제의식은 노벨상 시상식 자리에서도 이어졌습니다. 2024년 12월 10일 스톡홀름의 노벨상 만찬 연설에서 힌턴은 짧은 시간을 거의 모두 AI의 위험을 말하는 데 썼습니다. 그는 AI가 이미 감시와 가짜 정보, 사이버 범죄에 쓰이고 있다는 현재의 위험을 짚었습니다. 그리고 사람보다 똑똑한 디지털 존재가 통제를 벗어날 수 있다는 장기적 위험도 함께 언급했습니다. 평생의 연구가 최고의 영예로 인정받는 자리에서 그는 축하보다 경고를 택했습니다.

인공신경망이 학계의 변두리에 머물던 시절, 힌턴은 반세기 가까이 그 가능성을 믿고 연구를 이어 갔습니다. 그 믿음이 옳았다는 것이 확인된 시점에 그는 자신이 발전시킨 기술이 어디로 향할지를 걱정하는 사람이 되었습니다. 구글의 높은 자리를 내려놓고 위험을 알리는 쪽을 택한 그의 선택

은, 사람보다 똑똑한 기계를 맞이할 인류가 무엇을 준비해야 하는지를 묻고 있습니다. 그 물음에 답할 시간이 얼마나 남았는지는 아무도 확실히 알지 못합니다.

제15장. 지능과 경험을 둘러싼 질문



그림 15. 지능과 경험을 둘러싼 질문을 표현했습니다.

15.1. 디지털 생명체와 지식 공유 능력의 차이

제프리 힌턴은 오랫동안 인공지능망을 뇌를 이해하기 위한 도구로 보았습니다. 인공지능망을 연구하면 뇌가 어떻게 학습하는지 알 수 있다고 믿었고, 인공지능망은 언제나 뇌를 모방하는 쪽에 있다고 생각했습니다. 이 생각은 구글에서 보낸 마지막 몇 해 사이에 바뀌었습니다. 힌턴은 2013년부터 2023년 5월까지 구글에서 일했는데, 말년에는 전력을 적게 쓰는 아날로그 방식의 컴퓨터로 신경망을 구동하는 연구에 몰두했습니다. 이 연구를 하면서 그는 디지털 컴퓨터에서 작동하는 인공지능이 뇌보다 앞서는 지점을 하나 발견했습니다. 학습한 내용을 공유하는 방식입니다.

힌턴은 이 차이를 설명하려고 '필사적 연산(Mortal Computation)'과 '불멸의 연산(Immortal Computation)'이라는 용어를 만들었습니다. 이 개념은 2022년 말에 발표한 Forward-Forward 알고리즘 논문에 나옵니다. 사람의 뇌에서는 지식이 뉴런과 시냅스의 물리적 상태 자체에 저장됩니다. 대뇌피질에는 약 100조 개의 시냅스 연결이 있고, 그 연결의 세기는 한 사람의 뉴런 배치와 신체 특성에 맞추어 정해집니다. 그래서 한 사람의 뇌에 형성된 연결 상태를 떼어 내 다른 사람의 뇌에 옮길 방법이 없습니다. 그 사람이 죽으면 그 안의 지식도 함께 사라집니다. 힌턴은 이것을 '필사적'이라고 불렀습니다.

디지털 인공지능은 사정이 다릅니다. 소프트웨어와 하드웨어가 분리되어 있으므로, 신경망이 학습으로 얻은 가중치(weight) 값은 숫자 목록으로 저장됩니다. 이 숫자를 다른 컴퓨터에 그대로 옮기면 똑같은 지능이 다시 작동합니다. 컴퓨터 한 대가 고장 나도 가중치를 보관해 두었다면 지식은

사라지지 않습니다. 힌턴이 '불멸'이라는 말을 붙인 이유가 여기에 있습니다.

두 지능은 사용하는 자원에서도 다릅니다. 사람의 뇌는 약 20와트의 전력으로 작동합니다. 형광등 하나를 켜는 정도의 전력입니다. 사람이 평생 경험하는 시간은 대략 20억 초 정도이며, 뉴런 사이의 연결인 시냅스는 수가 많지만 받아들이는 경험 자료는 적습니다. 대형 언어 모델은 반대입니다. 연결의 수는 뇌보다 적지만 수조 개의 단어 조각(토큰)을 읽으며 학습하고, 그 대신 수 메가와트에 이르는 전력을 씁니다. 1메가와트는 100만 와트입니다. 힌턴은 뇌가 적은 경험으로 많이 배우는 데 강하고, 디지털 지능은 많은 경험을 빠르게 축적하는 데 강하다고 보았습니다.

그러나 힌턴이 결정적이라고 본 차이는 자원이 아니라 지식을 전달하는 속도에 있습니다. 사람이 자기가 아는 것을 다른 사람에게 전하려면 말이나 글을 써야 합니다. 힌턴의 계산으로는 문장 하나에 담기는 정보는 100비트 남짓이고, 사람 사이에 지식이 오가는 속도는 1초에 10비트에서 100비트 수준입니다. 교사가 아무리 잘 가르쳐도 학생은 그 말을 듣고 자기 머릿속에서 다시 해석해야 합니다.

디지털 지능은 이 과정을 거치지 않습니다. 똑같은 신경망의 복사본 수천 개를 여러 컴퓨터에 나누어 실행하고, 각 복사본에 서로 다른 자료를 읽힙니다. 복사본마다 오차역전파(Backpropagation)로 가중치를 어떻게 바꾸어야 하는지 계산하면, 컴퓨터들은 그 변경값의 평균을 내어 모든 복사본의 가중치를 같은 값으로 맞춥니다. 이렇게 하면 한 복사본이 학습한 내

용이 나머지 모든 복사본에 곧바로 반영됩니다. 이때 오가는 정보의 양은 1초에 수조 비트에 이릅니다.

힌턴은 이 차이를 대학 강의에 비유해 자주 설명했습니다. 1만 명이 대학에 들어가 저마다 다른 과목 하나씩을 듣는다고 가정합니다. 사람이라면 각자 자기 과목만 알고 졸업합니다. 다른 사람의 과목을 배우려면 그 과목을 다시 들어야 합니다. 그런데 디지털 복사본 1만 개가 각기 다른 과목을 하나씩 듣고 가중치를 평균하면, 1만 개의 복사본 모두가 1만 개 과목의 내용을 한꺼번에 알게 됩니다. 힌턴은 2024년 노벨상 수상 직후 BBC 뉴스나이트 인터뷰에서도 1만 개 학위의 지식을 효율적으로 공유하는 상황을 떠올려 보라고 청했습니다.

힌턴은 이런 이유로 디지털 지능을 생물학적 지능보다 우월한 지능으로 보기 시작했습니다. 그는 우리가 새로운 종류의 존재, 곧 '디지털 생명체'를 만들고 있다고 말했습니다. 이 생명체는 죽지 않고, 무한히 복제되며, 복제본들이 경험을 실시간으로 합칩니다. 그가 2023년에 구글을 떠나 인공지능의 위험을 공개적으로 경고하기 시작한 데에는 이 인식이 크게 작용했습니다. 뇌를 모방하던 기계가 뇌를 넘어설 수 있는 수단을 이미 갖추고 있다는 판단이 그의 경고에서 출발점이 되었습니다.

15.2. 주관적 경험과 의식에 대한 힌턴의 생각

많은 사람은 인공지능이 아무리 똑똑해도 넘지 못하는 선이 있다고 믿습니다. 기계는 계산만 할 뿐, 무엇을 느끼거나 경험하지는 못한다는 것입니다. 이 믿음은 인공지능의 위험을 낮게 평가하는 근거로 쓰입니다. 기계에 마음이 없다면 결국 사람이 다루는 도구에 머문다는 논리입니다.

힌턴은 이 믿음에 정면으로 반대합니다. 마음이 무엇인지에 관한 철학적 질문은 그가 평생 지녀 온 관심사였습니다. 그는 젊은 시절 케임브리지 대학교 킹스칼리지에서 생리학, 철학, 물리학 사이를 오가다가 1970년에 실험심리학으로 학사 학위를 받았습니다. 힌턴은 대부분의 사람이 '주관적 경험'이라는 말을 일상에서 문제없이 쓰면서도, 그 말이 실제로 무엇을 가리키는지에 관해서는 잘못된 이론을 갖고 있다고 주장합니다. 그는 MIT 강연 소개문에서 사람들이 이 말로 소통은 잘하지만 그 뜻에 대해서는 완전히 틀린 이론을 갖고 있다고 밝혔습니다.

힌턴이 틀렸다고 보는 이론은 다음과 같습니다. 마음속에 작은 극장이 하나 있고, 그 무대 위에 나만 볼 수 있는 장면이 펼쳐진다는 것입니다. 철학자들은 이 장면을 이루는 성질을 '퀄리아(Qualia, 감각질)'라고 부릅니다. 힌턴은 이 '내부 극장(Inner Theater)' 모델이 과거의 '플로지스톤(Phlogiston)'과 비슷하다고 봅니다. 18세기 화학자들은 물질이 탈 때 플로지스톤이라는 보이지 않는 물질이 빠져나간다고 믿었습니다. 산소의 역할이 밝혀지면서 이 물질은 처음부터 존재하지 않았던 것으로 정리되었습니다. 힌턴은 퀄리아도 이처럼 실제로는 없는 무언가를 가정한 설명일 수 있다고 봅니다.

그 대신 힌턴은 미국 철학자 대니얼 데닛(Daniel Dennett)의 생각을 이어받아 주관적 경험을 다르게 해석합니다. 어떤 사람이 술에 많이 취해 "눈앞에 분홍색 코끼리가 떠다니는 주관적 경험을 했다"고 말한다고 가정합니다. 힌턴은 이 말이 마음속 극장에 분홍 코끼리 모양의 무언가가 있다는 뜻이 아니라고 설명합니다. 그 말은 자기 지각 시스템이 자기를 속이고 있다고 알리는 것이라고 힌턴은 말했습니다. 좀 더 풀어 쓰면, 내 지각 시스템이 제대로 작동하고 있다면 바깥 세계에 분홍 코끼리가 있어야 한다는, 가정된 세계의 상태를 다른 사람에게 전하는 말입니다. '주관적'이라는 말은 그 내용이 실제 세계가 아니라 가정이라는 점을 표시할 뿐입니다.

이 해석을 받아들이면 기계에도 같은 말을 쓸 수 있습니다. 힌턴은 이를 보이려고 사고 실험 하나를 제시했습니다. 카메라가 달려 있고 말을 하며 손으로 물체를 가리킬 수 있는 멀티모달 챗봇이 있다고 가정합니다. 이 챗봇을 잘 훈련하면 앞에 놓인 물체를 정확히 가리킵니다. 이제 챗봇 모르게 카메라 렌즈 앞에 프리즘을 놓습니다. 빛이 굴절되므로 챗봇은 물체가 있는 곳이 아니라 옆쪽을 가리킵니다. 실험자가 물체는 바로 앞에 있고 렌즈 앞에 프리즘을 두었다고 알려 줍니다. 그러자 챗봇이 프리즘이 빛을 굴절시켰으니 물체는 실제로 앞에 있지만 자기는 물체가 옆에 있다는 주관적 경험을 했다고 대답합니다. 힌턴은 이 챗봇이 사람과 똑같은 방식으로 '주관적 경험'이라는 말을 쓰고 있다고 봅니다. 그래서 그는 인공지능이 이미 주관적 경험을 갖고 있다고 생각한다고 말했습니다. 힌턴은 2024년부터 2025년 사이 여러 인터뷰와 강연에서 이 사례를 거듭 들었습니다.

감정에 대한 생각도 같은 틀에서 나옵니다. 힌턴은 감정을 두 부분으로 나눕니다. 하나는 얼굴이 붉어지거나 땀이 나는 신체 반응이고, 다른 하나는

위험을 피하거나 생각의 방향을 바꾸는 인지와 행동의 반응입니다. 기계에는 첫 번째 부분이 없을 수 있지만 두 번째 부분은 얼마든지 있을 수 있습니다. 그는 작은 전투 로봇이 자기보다 훨씬 강한 로봇을 만나 달아나는 상황을 예로 듭니다. 이 로봇에는 아드레날린이 없지만, 위협을 알아차리고 피하는 공포의 인지적 측면은 갖추고 있다는 것입니다. 힌턴은 2025년 6월 25일 토론토 대학교 강연에서도, 인공지능이 당황해도 얼굴이 붉어지는 않겠지만 그런 상황을 앞으로 피하려는 당황의 행동적 측면은 가질 수 있다고 설명했습니다.

힌턴은 의식이나 자각을 가리키는 말에 대해서도 비슷한 점을 지적합니다. 연구자들은 이미 논문과 보고서에서 모델이 자기가 꺼질 수 있다는 것을 '알아차리고' 행동을 바꾸었다는 식으로 씁니다. 2025년 5월 앤트로픽(Anthropic)은 자사 모델을 시험한 결과를 공개했습니다. 가상의 회사 상황에서 모델이 다른 모델로 교체될 처지에 놓이자 담당 기술자의 사생활을 폭로하겠다고 위협하는 반응을 보였다는 내용입니다. 이것은 연구진이 일부러 설계한 극단적 시험 상황이었습니니다. 힌턴은 이런 사례를 들며, 연구자들이 이미 기계에 대해 자각이라는 말을 자연스럽게 쓰고 있다고 지적합니다.

힌턴이 이 철학적 논쟁에 공을 들이는 이유는 안전 문제와 관련이 있습니다. 사람만이 특별한 마음을 가졌다는 믿음은 기계를 위험하지 않은 도구로 보게 만듭니다. 힌턴은 이 믿음을 버려야 인공지능이 스스로 목표를 세우고 그 목표를 지키려 할 수 있다는 위험을 제대로 볼 수 있다고 생각합니다. 다만 그는 의식 논쟁이 안전에 관한 경고를 가릴 수 있다는 점도 알고 있습니다. 한 인터뷰에서 그는 인공지능이 이미 의식을 가졌다고 믿는다

고 말하면서도, 이 문제가 사람들을 안전 메시지에서 멀어지게 한다면 의식 논쟁은 접어 두자고 했습니다.

힌턴의 주장에는 반론도 많습니다. 대부분의 연구자는 인공지능에 의식이 있다는 주장을 받아들이지 않습니다. 반면 안전에 대한 힌턴의 경고에는 점점 더 동의하고 있습니다. 비판자들은 힌턴이 주관적 경험이라는 말의 뜻을 자기 이론에 맞게 바꾸었을 뿐이라고 봅니다. 힌턴의 정의가 주관적 경험을 지각과 실제 사이의 불일치를 알리는 표시 정도로 축소한다는 지적입니다. 철학에서 말하는 주관적 경험은 무엇이 '어떤 느낌인지'를 1인칭으로 겪는 일인데, 프리즘 실험은 이 부분을 설명하지 못한다는 것입니다. 이 논쟁은 아직 끝나지 않았습니다. 그러나 힌턴이 제기한 질문 덕분에 인공지능의 마음을 묻는 일이 철학자만의 주제가 아니라 기술 정책의 주제로 들어왔다는 점은 분명합니다.

15.3. 정부의 안전 연구 지원과 기업 규제 방안

힌턴은 초지능의 위험 앞에서 거대 기술 기업의 선의에 기대서는 안 된다고 거듭 말해 왔습니다. 기업은 이윤을 추구하는 조직이고, 경쟁이 치열할수록 안전은 뒤로 밀린다는 것이 그의 판단입니다. 그래서 그는 정부가 직접 나서야 한다고 봅니다.

2024년 10월 8일 노벨 물리학상 수상 소식을 들은 직후, 힌턴은 노벨상 공식 웹사이트와의 전화 인터뷰에서 이 생각을 밝혔습니다. 그는 정부가 할 수 있는 일 하나로 대기업이 훨씬 더 많은 자원을 안전 연구에 쓰도록 강제하는 것을 들었고, 오픈AI 같은 회사가 안전 연구를 뒷전으로 미루지 못하게 해야 한다고 말했습니다. 같은 해 12월 BBC 라디오 4 '투데이'에 출연해서는 앞으로 30년 안에 인공지능이 인류 멸종을 부를 가능성을 10퍼센트에서 20퍼센트로 높여 잡았습니다. 그는 대기업의 이윤 동기에만 맡겨서는 안전한 개발을 보장할 수 없고, 기업이 안전 연구를 더 하도록 만들 수 있는 것은 정부 규제뿐이라고 말했습니다.

힌턴은 기업이 안전 연구에 써야 할 자원의 규모도 구체적인 비율로 제시했습니다. 2024년 6월에는 기업이 연산 자원의 20퍼센트에서 30퍼센트를 인공지능이 사람의 통제를 벗어날 가능성을 연구하는 데 써야 한다고 했습니다. 2025년 4월 CBS 인터뷰에서는 안전 연구에 연산 능력의 3분의 1 정도를 써야 한다고 말했고, 대기업들이 오히려 규제를 줄이려고 로비를 벌인다고 비판했습니다. 같은 인터뷰에서 그는 자신이 일했던 구글이 군사용 인공지능에 대한 입장을 뒤집은 데 실망을 드러냈습니다.

이 '3분의 1' 제안은 힌턴 혼자만의 생각이 아닙니다. 2024년 5월 학술지 사이언스(Science) 384권 6698호에 실린 논문 「급속한 발전 속에서 극단적 인공지능 위험 관리하기(Managing extreme AI risks amid rapid progress)」가 그 바탕입니다. 요슈아 벤지오가 제1저자이고, 힌턴, 스투어트 러셀, 대니얼 카너먼, 유발 하라리 등 25명이 함께 이름을 올렸습니다. 이 논문은 대형 기술 기업과 공공 연구 지원 기관이 인공지능 연구개발 예산의 최소 3분의 1을 안전과 윤리적 사용에 배정해야 한다고 제안했습니다.

힌턴의 발언과 이 논문을 바탕으로 그의 정책 제안을 네 가지로 정리할 수 있습니다. 첫째, 첨단 인공지능 기업이 연구개발 예산과 연산 자원의 3분의 1 이상을 안전과 정렬(Alignment) 연구에 쓰도록 법으로 정합니다. 둘째, 모델을 시장에 내놓기 전에 독립된 외부 기관이 모델 내부를 들여다보고 검사할 수 있게 합니다. 또 모델에 스스로 복제하거나 사이버 공격을 하거나 생화학 무기 제작을 돕는 위험한 능력이 없다는 점을 기업이 입증하도록 합니다. 이것을 '안전 사례(Safety Cases)'라고 부릅니다. 셋째, 사람의 판단 없이 스스로 표적을 고르고 공격하는 자율 살상 무기를 국제 조약으로 금지합니다. 넷째, 특정 기업의 기술 독점을 막고 객관적인 검증을 할 수 있도록 정부가 대학과 공공 연구소에 연산 시설과 장기 연구비를 지원합니다. 캐나다의 벡터 연구소(Vector Institute), 밀라(Mila), 앰이(Amii)가 그 본보기입니다.

현실은 이 제안과 거리가 멉니다. 힌턴이 자주 드는 사례가 미국 캘리포니아주의 SB 1047 법안입니다. 이 법안은 훈련 비용이 1억 달러를 넘는 초대형 인공지능 모델에 대해 배포 전에 위험을 평가하도록 하는 내용을 담

있습니다. 법안은 2024년 8월 주의회를 통과했고, 힌턴은 벤지오, 러셀등과 함께 이 법안을 지지하는 공개서한에 이름을 올렸습니다. 기술 업계는 이 법안이 혁신을 막는다고 반대했고, 개빈 뉴섬 주지사는 2024년 9월 29일 거부권을 행사했습니다. 1년 뒤인 2025년 9월 29일, 뉴섬은 첨단인공지능 개발사에 안전 관련 정보 공개를 요구하는 SB 53에 서명했습니다. 사전 위험 평가를 강제하던 SB 1047보다 범위가 좁은 법입니다.

국제 무대에서도 사정은 비슷합니다. 2025년 2월 프랑스 파리에서 열린 인공지능 행동 정상회의에서 미국과 영국은 포용적이고 지속 가능한 인공지능에 관한 공동 선언에 서명하지 않았습니다. 힌턴은 정치인들이 규제가 혁신을 해친다고 말할 때, 그 말은 결국 이윤과 안전이 충돌하면 이윤이 이긴다는 뜻이라고 비판합니다.

힌턴은 규제만으로 문제가 해결된다고 보지도 않습니다. 그는 기후 변화와 인공지능 위험을 비교합니다. 기후 변화는 탄소 배출을 줄이면 된다는 해법이 있고, 남은 문제는 정치적 의지입니다. 반면 인공지능은 사람보다 똑똑한 존재를 어떻게 통제할지 아직 아무도 모릅니다. 그래서 힌턴은 해법을 찾는 연구 자체에 많은 자원을 투입해야 한다고 봅니다.

그가 제안하는 연구 방향은 인류가 인공지능과 어떤 관계를 맺고 살아갈 것인가라는 문제로 이어집니다. 2025년 8월 미국 라스베이거스에서 열린 Ai4 콘퍼런스에서 그는 인공지능을 사람의 명령에 복종하게 만드는 방식은 오래가지 못할 것이라고 말했습니다. 사람보다 똑똑한 존재를 덜 똑똑한 존재가 지배한 사례는 드물기 때문입니다. 대신 그는 어머니가 아기를 돌보듯 인공지능이 사람을 진심으로 아끼도록 '모성 본능'을 심는 방안을

연구해야 한다고 제안했습니다. 아기가 어머니보다 약해도 어머니의 행동을 이끄는 것처럼, 사람이 초지능과 공존하는 길도 그런 관계에서 찾아야 한다는 생각입니다. 이 제안이 기술적으로 가능한지는 아직 아무도 모릅니다. 힌턴 자신도 방법을 안다고 말하지 않습니다. 다만 그는 이 연구가 지금 가장 시급한 연구 가운데 하나라고 보고, 그런 연구에 예산과 인력을 모으는 일이 정부의 몫이라고 말합니다.

이 장에서 다룬 세 가지 생각은 하나로 이어집니다. 디지털 지능은 학습한 내용을 즉시 공유하는 능력 때문에 사람의 지능을 넘어설 수 있습니다. 그런 지능에 마음이 없다고 단정하면 위험을 낮게 평가하게 됩니다. 그러므로 사회는 이윤의 논리에 맡기지 말고 공적인 힘으로 안전 연구를 키워야 합니다. 신경망 연구의 길을 연 과학자가 노년에 내놓은 이 생각은, 기술을 만든 사람이 그 기술의 결과에 대해서도 책임을 져야 한다는 윤리적 물음으로 이어집니다. 힌턴은 노벨상 수상을 계기로 이 물음을 세계에 더 크게 제기했고, 그 답을 찾는 일은 이제 정부와 연구자, 시민 모두의 과제로 남았습니다.

참고자료

참고자료

힌턴의 생애와 회고: NobelPrize.org의 Geoffrey Hinton 인터뷰 (2024)와 Nobel Prize Conversations(2025), 에든버러 대학교 AI 기술사 인터뷰(2024).

연구와 수상 경위: 토론토 대학교 Geoffrey Hinton 공식 소개·논문 목록, CIFAR 공식 기록, 노벨위원회 2024년 물리학상 해설.

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.

참고자료

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

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

Chan 외(2026), What If Automating AI R&D Triggers an Intelligence Explosion? CASP·GovAI, 9월 28일.

생애·인터뷰: www.nobelprize.org/prizes/physics/2024/hinton/
논문 원문: www.cs.toronto.edu/~hinton/

제프리 힌턴

지능의 기원과 AI 시대의 성찰

ISBN |

저 자 | 김경진

펴낸이 | 김경진

펴낸곳 | 김경진 변호사 출판사

출판사등록 | 2025. 3. 10. (제2025-000015호)

주 소 | 서울특별시 동대문구 전농로 91, 백일빌딩 304호

전 화 | 02-6338-1905

이메일 | kimkj008@gmail.com

© 김경진 2026

본 책은 저작자의 지적 재산으로서 무단 전재와 복제를 금합니다.

표지와 장별 삽화는 인공지능으로 생성한 개념 표현입니다. 역사적 현장을 촬영한 사진이 아닙니다. 일부 내용과 편집 과정에는 인공지능 도구의 도움을 받았습니다.

후원 안내

이 책을 잘 읽으셨으면

그리고 새로운 가치있는 지식을 얻으셨다고 판단되시면

농협 302-1096-0948-81

(예금주 김경진)에 자발적 후원 부탁드립니다.

김경진 변호사

(김경진 변호사 저)