DAY 46–55 · 5부 안을 들여다보기

5부 총정리 — 안을 들여다보기

점 하나에서 내 폰까지 · 용어까지 · DAY 46-55

시험 없음 — 읽고 넘어가는 복습 자료

점 하나는 곱하고 더하는 것밖에 못 해

1. 신경망은 숫자를 주고받는 점이 층으로 쌓인 구조야. 점을 노드, 한 겹을 층이라고 불러.

2. 점 하나는 들어온 숫자마다 **가중치**를 곱해 전부 더해. 이 값을 가중합이라고 해.

3. 거기에 문턱 역할을 하는 함수를 한 번 통과시켜 다음 층으로 넘겨.

문턱이 하는 일 — 활성함수

1. 문턱 함수를 활성함수라고 해. 신호를 그대로 흘릴지, 꺾을지, 끊을지를 정하는 자리야.

2. 활성함수가 없으면 층을 몇 개 쌓아도 결국 **한 번 곱해 더한 것**과 같아져.

3. 곧게 비례하는 계산만 겹치면 아무리 겹쳐도 곧은 계산 하나로 줄어들기 때문이야.

가중치는 어떻게 정해지나 — 역전파

1. 처음 가중치는 아무 값이야. 답을 내 보고 틀린 만큼 거꾸로 되짚어 고쳐.

2. 틀린 양을 뒤에서 앞으로 되돌려 가중치를 조금씩 고치는 방법을 **역전파**라고 해. 1986년에 널리 알려졌어.

3. 이걸 수없이 되풀이하면, 남은 숫자들이 곧 배운 내용이야.

시선은 나눠 갖는 거야

1. 어텐션은 단어마다 질의·키·값 세 벌을 만들어 두고, 질의와 키를 맞춰 점수를 내는 계산이야.

2. 그 점수를 **합이 1인 몫**으로 바꾸는 단계를 소프트맥스라고 해.

3. 몫이라 총량이 정해져 있어 — 한쪽을 세게 보면 나머지는 반드시 약해져.

왜 이게 뜻을 정하나

1. '그것'의 질의가 앞에 나온 어떤 명사의 키와 잘 맞으면 그 자리에 시선이 몰려.

2. 값이 그 몫만큼 섞여 들어와서 **문맥이 섞인 새 숫자 묶음**이 만들어져.

3. 같은 단어라도 문장이 다르면 다른 값이 나오는 이유가 이거야.

한 줄 읽기에서 한꺼번에 보기로

1. 2014년 번역 연구가 처음 썼어. 긴 문장에서 멀리 있는 단어를 놓치는 문제를 풀려는 장치였지.

2. 2017년 논문은 순서대로 읽는 부분을 아예 빼고 **어텐션만 남겼어** — 트랜스포머야.

3. 한꺼번에 보니 계산을 나눠 돌리기 쉬워졌고, 그래서 모델을 크게 키울 수 있었어.

뜻을 자리로 바꿔 놓는다

1. 단어를 숫자 묶음으로 바꾼 값을 임베딩이라고 해. 묶음 하나에 수백에서 수천 개 숫자가 들어가.

2. 학습 규칙은 간단해 — **주변에 같이 나오는 말이 비슷하면 숫자도 비슷하게** 맞춘다는 거야.

3. 사람이 뜻을 적어 주지 않아. 엄청난 양의 글을 읽으며 스스로 맞춰 가.

자리가 되면 계산이 된다

1. 뜻이 자리가 되면 두 단어의 거리를 계산으로 잴 수 있어.

2. 2013년 연구는 자리 사이의 **방향까지 뜻을 담는다**는 걸 보였어 — 왕에서 남자를 빼고 여자를 더하면 여왕 근처에 닿는다는 식으로.

3. 방향이 관계를 담는다는 게 이 방식이 오래 쓰이는 이유야.

통째 단어에서 조각으로

1. 지금 모델은 단어를 통째로 쓰지 않고 토큰이라는 조각으로 잘라 써.

2. 처음 보는 말도 조각으로 쪼개면 다룰 수 있어서 **모르는 단어 문제**가 많이 줄었어.

3. 그래서 글자 수가 아니라 토큰 수로 길이를 세는 거야.

글자가 아니라 뜻을 잰다

1. 문장도 임베딩으로 바꿀 수 있어. 두 묶음 사이 각도를 재면 뜻이 얼마나 겹치는지 나와.

2. 많이 쓰는 잣대가 **코사인 유사도**야 — 묶음의 크기는 빼고 방향만 비교해.

3. 방향이 같으면 1에 가깝고, 상관없으면 0 근처로 떨어져.

글자 세기와 뜻 재기는 다르다

1. '재밌어'와 '재미없어'는 글자가 거의 같지만 방향은 멀어.

2. 거꾸로 겹치는 글자가 하나도 없어도 **뜻이 같으면 가까워**.

3. 글자 겹침으로만 세는 옛 방식이 자주 틀리던 자리가 바로 이 둘이야.

문장 묶음을 잘 만드는 법

1. 2019년 연구는 두 문장이 같은 잣대에 놓이도록 따로 학습시키는 방법을 내놨어.

2. 덕분에 글 조각 수백만 개를 미리 숫자로 바꿔 두고, 질문이 오면 **가까운 것만 꺼내** 쓸 수 있게 됐어.

3. 이 꺼내 오기가 다음 날 배우는 책 펴놓고 답하기의 앞 절반이야.

외우기 대신 찾아서 펴 놓기

1. 찾아온 글을 옆에 펴 두고 답하게 하는 방식을 검색 보강 생성이라고 해. 줄여서 RAG라고 부르지.

2. 질문을 임베딩으로 바꿔 **가까운 글 조각**을 꺼내 오고, 그 조각을 질문과 함께 모델에 넣어.

3. 모델은 외운 것 대신 눈앞의 글을 읽고 답해.

왜 지어내기가 줄어드나

1. 모르는 걸 외운 기억으로 메우려 할 때 지어내기가 생겨.

2. 근거 글이 눈앞에 있으면 메울 자리가 줄고, **어디서 나온 말인지** 같이 보여 줄 수 있어.

3. 출처가 붙으니 읽는 사람이 직접 확인할 수도 있지.

약한 자리는 찾기 쪽이다

1. 모델을 다시 가르치지 않아도 되는 게 큰 장점이야 — **펴 놓는 글만 갈아 끼우면** 답이 따라서 새로워져.

2. 대신 찾아온 조각이 엉뚱하거나 틀리면 답도 같이 틀려.

3. 그래서 실제로 공을 들이는 곳은 답하기보다 찾기 쪽이야.

다 큰 모델을 조금 다듬는다

1. 크게 한 번 배운 모델을 기본 모델이라고 하고, 그 위에 필요한 일만 더 학습시키는 걸 파인튜닝이라고 해.

2. 안에 든 숫자를 전부 고치는 게 아니라 **아주 일부만** 고치는 방식이 널리 쓰여.

3. 학습을 처음부터 다시 하는 것과는 비용이 아예 다른 일이야.

LoRA — 고치는 숫자를 줄이는 방법

1. 2021년에 나온 LoRA는 원래 숫자를 얼려 두고, 옆에 작은 판 두 장을 덧붙여 그것만 학습해.

2. 그 결과 고치는 숫자를 **만분의 일**까지 줄였어.

3. 덧붙인 판만 갈아 끼우면 같은 기본 모델로 여러 일을 돌릴 수 있고.

어느 쪽을 쓰나 — 다듬기와 펴 놓기

1. 파인튜닝이 센 자리는 **말투·형식·일하는 방식**이야.

2. 새로 생긴 사실을 넣는 일은 앞에서 본 책 펴놓고 답하기가 맞아.

3. 둘은 경쟁이 아니라 역할이 달라서, 실제로는 같이 쓰는 경우가 많아.

정답 대신 어느 쪽이 더 나은지

1. 같은 질문에 답을 여러 개 내게 하고 사람이 더 나은 쪽을 고르게 해. 이 기록을 선호 데이터라고 불러.

2. 고른 기록으로 먼저 **보상 모델**을 만들어 — 답을 보면 사람이 몇 점 줄지 맞혀 보는 작은 모델이야.

3. 그 점수를 올리는 쪽으로 본 모델을 다듬는 단계가 강화학습이야.

왜 정답표를 안 쓰나

1. 좋은 답은 하나로 적을 수 없어. '친절하게'나 '짧게'는 채점표로 만들기 어렵지.

2. 그래서 점수를 적는 대신 **둘 중 어느 쪽이 나은지**만 물어. 사람이 훨씬 쉽게 답할 수 있는 질문이야.

3. 모은 건 맞고 틀림이 아니라 사람의 취향이라는 말이 이 뜻이야.

효과와 남은 문제

1. 2022년 연구에서, 이렇게 다듬은 작은 모델의 답을 사람들은 **130배 넘게 큰 모델**의 답보다 더 좋아했어.

2. 크기를 키우는 것보다 다듬는 게 더 큰 차이를 낸 드문 사례야.

3. 대신 누구의 취향을 모았느냐가 그대로 남아. 그래서 규칙을 글로 적어 두고 AI끼리 채점하게 하는 방법도 2022년에 나왔어.

계산의 모양이 원래 같았다

1. 그래픽카드(GPU)는 쉬운 계산을 동시에 아주 많이 하도록 만든 부품이야. 이걸 병렬 계산이라고 해.

2. 화면의 점 수백만 개를 동시에 칠하려면 **똑똑한 몇 명보다 단순한 수천 명**이 필요했어.

3. AI 학습의 속은 행렬 곱셈 — 쉬운 곱셈과 덧셈 수조 개를 동시에 하는 일이야.

숫자로 보는 카드 한 장

1. 학습에 많이 쓰는 카드 한 장 안에는 트랜지스터라는 작은 스위치가 **약 800억 개** 들어 있어.

2. 그 한 장이 1초에 약 1,000조 번 계산해.

3. 요즘 카드는 숫자를 거친 눈금으로 다뤄 더 많이 계산하는 길도 함께 써.

한 장으로는 안 돼서

1. 라마 3는 카드 **1만 6천 장**을 한 덩어리로 묶어 54일을 돌려서 만들었어.

2. 이렇게 묶은 덩어리를 클러스터라고 하고, 카드 사이를 잇는 선의 속도가 성능을 가르기도 해.

3. 2024년에는 10만 장 규모를 묶었다고 밝힌 곳도 나왔어.

한 번은 작고, 다 더하면 커

1. 2025년에 공개된 측정에서, 글로 한 번 물었을 때 전기 사용량 중간값은 **0.24와트시**였어.

2. TV를 9초 켜 두는 정도라, 한 번만 보면 작은 양이야.

3. 다만 이건 한 회사가 자사 기준으로 밝힌 값이고, 어디까지 포함했는지에 따라 값이 달라져.

건물 전체로 보면

1. 전 세계 데이터센터가 2024년에 쓴 전기는 **415테라와트시** — 전 세계 전기의 약 1.5%야.

2. 국제에너지기구는 2030년에 945테라와트시로 두 배가 넘을 거라고 봤어.

3. 작아 보이는 1.5%지만, 6년 만에 두 배가 되는 항목은 흔하지 않아.

그래서 어디를 봐야 하나

1. 한 번 값은 **빠르게 작아지고** 있어. 같은 답을 더 적은 계산으로 내는 쪽으로 계속 바뀌니까.

2. 그런데 질문 수가 더 빠르게 늘면 합계는 계속 커져.

3. 숫자를 볼 때는 한 번과 합계를 섞지 않는 것부터 챙겨야 해.

조각으로 나눠 쓴다

1. 요즘 폰 안에는 숫자 **30억 개쯤 되는 작은 모델**이 들어 있어. 폰 안에서 도는 걸 온디바이스라고 해.

2. 인터넷이 끊겨도 요약하고 다듬고 알림을 정리할 수 있어.

3. 서버에 있는 모델과는 크기가 수십에서 수백 배 차이 나.

어떻게 폰에 들어가나

1. 큰 모델의 숫자를 더 거친 눈금으로 줄여 담는 **양자화**를 써. 자리를 줄이면 속도도 올라가.

2. 메모리가 모자랄 땐 필요한 부분만 그때그때 불러 쓰는 방법도 같이 쓰여.

3. 그래도 작아진 만큼 못 하는 일이 생기니까, 전부를 폰에 넣지는 않아.

조각으로 나누는 기준

1. 폰이 먼저 해 보고 버거운 조각만 골라 밖으로 내보내는 방식이 기본이야.

2. 안에서 끝내면 **빠르고, 전기도 덜 들고, 내 글이 밖으로 안 나가**.

3. 앞에서 본 그 건물이 필요한 일은 여전히 남아. 둘을 섞어 쓰는 게 지금의 답이야.

안쪽은 마법이 아니라 계산이야.
겨누는 법과 돌리는 법이 전부지.



6부에서는 앞으로를 본다.