

● DAY 19 · 2 부 여기까지 온 길

사진과 소리를 어떻게 글자 옆에 세울까 — 보고 듣는 AI

하루 5 분 , AI 이야기 · 2026 년 9 월 25 일



사진도 소리도 조각으로 쪼개 줄을 세우면 , 글자와 똑같은 자리에 설 수 있어 .

한 줄로 말하면

- 1 AI 가 사진을 보고 소리를 듣게 된 비결은 눈과 귀를 단 게 아니야 . 사진도 소리도 **작은 조각으로 쪼개서** 글자 조각과 같은 줄에 세운 거야 .
- 2 글은 낱말 조각으로 , 사진은 격자로 자른 네모 조각으로 , 소리는 **아주 짧은 시간 토막**으로 잘라 . 잘라 놓으면 셋 다 그냥 ' 줄 세운 조각 ' 이 돼 .
- 3 그러면 **뼈대를 새로 만들 필요가 없어** . 조각들이 서로 쳐다보게 하는 트랜스포머에 그대로 넣으면 되거든 . 여러 감각을 한꺼번에 다룬다고 해서 멀티모달이라고 불러 .

무엇을 넣든 같은 길로 들어간다



처음부터 셋을 같이 배운 모델은 사진을 글로 옮겨 적는 단계를 건너뛰어서 더 빠르고 정확해 .

핵심 : 눈과 귀를 단 게 아니라 , 모든 것을 조각으로 바꿔 같은 줄에 세운 것이다

글 · 사진 · 소리를 조각으로 쪼개 보기

글을 넣으면

낱말 조각으로 잘린다

오늘 급식 은 떡볶이 야

글은 낱말이나 그 일부로 잘려 . ' 떡볶이 ' 가 한 덩이가 아니라 두세 조각으로 잘리기도 해 — 자주 안 쓰는 말일수록 잘게 쪼...

사진을 넣으면

격자 네모로 잘린다

하늘 하늘 구름 지붕 창문 담장
길 사람 길

사진은 바둑판처럼 격자로 잘라 . 조각마다 몇 번째 칸인지 번호를 같이 붙여 주면 , 글자 조각과 똑같이 한 줄로 세울 수 있어 .

소리를 넣으면

짧은 시간 토막으로 잘린다

0.0 초 0.1 초 0.2 초 0.3 초
0.4 초 0.5 초

소리는 시간을 따라 아주 짧게 토막 내 . 여기까지 오면 셋이 똑같아져 — 글 · 사진 · 소리가 모두 줄 세운 조각일 뿐이라 같은...

버튼을 눌러 재료를 바꿔 봐 . 무엇을 넣든 결국 조각이 줄지어 선다는 게 보이면 성공이야 . 주황색은 지금 눈여겨볼 조각이야 .

학습 .html 에서 직접 만져 볼 수 있어

세 문제만 풀어 보자

Q1

사진을 AI 에 넣을 때 실제로 하는 일은 ?

- ① 사진을 먼저 글로 번역한다
- ② 흑백으로 바꾼다
- ③ 격자로 잘라 조각을 만들어 줄 세운다
- ④ 크기를 두 배로 키운다

Q2

글 · 사진 · 소리를 한 뼈대로 다룰 수 있는 이유는 ?

- ① 모델에 눈과 귀를 달아서
- ② 사진이 원래 글자니까
- ③ 소리에는 뜻이 없으니까
- ④ 셋 다 조각으로 바꿔 같은 줄에 세우니까

Q3

소리는 어떤 식으로 쪼갤까 ?

- ① 짧은 시간 토막
- ② 글자 한 개씩
- ③ 색깔별로
- ④ 길이순 정렬

왜 그런지까지 알고 가자

Q1

③

바둑판처럼 잘라서 조각마다 몇 번째 칸인지 번호를 붙여 . 그러면 낱말 조각과 똑같은 모양이 되지 .

Q2

④

잘라 놓으면 셋 다 ' 줄 세운 조각 ' 이야 . 그래서 조각끼리 쳐다보게 하는 트랜스포머를 그대로 쓸 수 있어 .

Q3

①

소리는 시간을 따라 흐르니까 아주 짧은 시간 단위로 토막 내 . 그 토막이 곧 조각 하나야 .

**사진도 소리도 조각으로 쪼개 줄을 세우면 ,
글자와 똑같은 자리에 설 수 있어 .**



내일은 모두의 모델 경쟁 — 만드는 곳이 몇 군데에서 수십 곳으로 늘어난 이야기 .