CLIP

2021년 당시 비전과 자연어 분야동

비전

- 2021년 이전은 효율적으로 깊은 모델을 어떻게 만들지 고민했다. 또는 Attention 모듈을 적용시키기도 했다.
  - Inception, ResNet
  - SENet, BAM
- 2021년 부터는 Transformer 구조를 적용시키는게 트랜드였다.
  - Vision Transformer, Image GPT
- 위에처럼 다양한 시도를 했지만 고질적인 문제가 있었다.
  1. 일반화능력 부족
  2. 작은 노이즈에 취약하다.

언어 모델 (LM)

- 2017년 까지만 해도 seq2seq 방식의 한계로 측정했지만 2017년 이후에 Transformer 구조가 들어오고 기존에 긴 문장을 소화하지 못한 seq2seq 구조와 달리 긴 문장도 효과적으로 처리 할 수 있게 되었다.
  그래서 나온게 GPT-1, GPT-2, BERT 등 초거대 언어 모델 (LLM) 이 나온거다.

CLIP

- 저자들은 LLM처럼 작동하는 vision 모델을 만들고 싶어했다. ⇨ 이유! 일반화, 범용성, 유연성을 높일수있다. (문서작업사용가능)
  - 글래스가 추가될 때마다 재학습 필요 없어짐
  1. Transformer 기반
  2. 자연어 문장으로 라벨대체
  3. 2번처럼 진행해서 Zero-shot 분류가능하게 하자 → Zero-shot이란 → 설명글에서 따로 설명진행