Seeing Understanding

CNN(합성곱 신경망, ResNet, VGG), Vision Tranformers(VIT)를 통해 **시각정 특징 추출**
YOLO, Faster R-CNN을 통해 **객체의 위치**를 네모난 상자로 표시, 객체 윤곽을 Mask R-CNN을 통해 픽셀 단위 **분할**
추출한 시각적 특징으로 사진 **인코더**, RNN이나 Transformer로 **디코더(문장 추출), 어텐션 메커니즘** 활용
멀티모달 학습으로 [이미지-텍스트] 쌍으로 학습

Vision Language(비전 언어, VL)

CNN이나 ViT를 통해 이미지를 읽고 특징벡터 추출
RNN이나 Tranformer를 통해 텍스트 분석, 임베딩 벡터 추출
크로스-어텐션 메커니즘을 통해 텍스트 인코딩에서 주목한 단어와 시각 인코딩에서 추출한 영역(픽셀)을 매칭
CLIP학습을 통해 텍스트와 이미지의 거리를 매핑
오픈소스 : All-Seeing
이미지와 텍스트간 쌍으로 연결하여 이미지 파악 후 강력한 쌍을 찾아 이미지->텍스트 파악
오픈소스2 : sglang
LLM(텍스트 데이터 학습 + 학습한 데이터 기반 질문 이해 및 답)과 VLM을 빠르고 효율적으로 사용하는 기술

SSD

이미지 속 객체의 위치와 종류를 파악(객체 검출 알고리즘)
한번의 단계만으로 여러 객체들의 위치와 클래스를 동시에 파악
홈 cctv, 자율주행, OCR 등에 자주 사용
오픈소스 : ssd-tansorflow