11주차

Instruction

인공지능 연구의 핵심 목표 중 하나는 multi-modal vision-and-language instructions 를 처리할 수 있는 general-purpose assistant 를 만드는 것.

- 하지만 대부분의 기존 연구가 하나의 단일 대형 vision 모델에 의해 이루어졌고, 언어는 주로 이미지 내용을 묘사하는데만 사용

반면 LLM 연구에서는 언어가 훨신 폭넓은 역할을 할 수 있다는 것을 보여주었음.

→ 본 논문에서는 instruduction tunign 개념을 언어-이미지 멀티모달 공간으로 확장하려는 첫 시도록, visual instruction tuning 을 제안

1. Multimodal instruction-following data
    - vision-language data 의 부족을 해결하기 위해 data reformation 관점과 이를 위한 파이프라인 제공
    - 이미지-텍스트 쌍을 instruction-following format 으로 변환하기 위해 ChatGPT 와 GPT-4 사용
2. Large multimodal models
    - CLIP 의 시각 인코더와 Vicuna 의 언어 인코더를 연결하여 새로운 대규모 멀티모달 모델 (LLM) 구축
    - 생성된 명령 데이터로 end to end 학습 됨
3. Multimodal instruction-following benchmark
4. Open-source release

Relative Work

Multimodal Instruction-following Agents

1. end to end 로 학습된 모델들: 이러한 모델은 특정 연구 주제별로 개별적으로 학습
2. 모델 간 조정 시스템: 다양한 모델을 연결하여 인간의 지시를 따르도록 하는 시스템

→ 모두 명령 수행형 비전-언어 에이전트를 구축하는 목표는 같지만 본 논문에서는 **여러 작업을 동시에 수행할 수 있는 end-to-end 학습된 언어-비전 멀티모달 모델**의 개발에 초점

Instruction Tuning

- 기존 연구에서 LLM 이 인간의 지시를 이해하고 실제 과제를 수행할 수 있도록 하기 위해 GPT-3, T5, PaLM, OPT와 같은 모델들을 대상으로 다양한 명령 튜닝 방법이 연구되어 옴.

→ 이 접근을 비전 분야로 확장

- 최근 공개된 LLaMA, OpenFlamingo, LLaMA-Adapter 등: 이미지 입력을 활용할 수 있는 오픈소스 LLM 을 구축하기 위한 노력

→ 하지만 명령 수행 데이터를 직접 활용하지 않기 때문에, 명령 기반 데이터로 학습된 모델에 비해 성능이 부족한 경향.

⇒ vision-language 명령 수행 데이터를 기반으로 한 visual instruction tuning 제안

GPT-assisted Visual Instruction Data Generation

최근 멀티모달 데이터 (이미지-텍스트 쌍 등) 이 폭팔적으로 증가했지만 멀티모달 instruction-following data 로 한정하면 그 양은 여전히 부족

→ 본 연구에서는 **ChatGPT/GPT-4를 활용한 명령 수행 데이터 자동 생성 방법**을 제안

데이터 변환 방식

- 하나의 이미지와 그에 대응하는 텍스트가 주어졌을 때 모델에게 이미지를설명하게 하기 위해 여러 개의 질문을 구성

  Human: X_1, X_2, X_3 <STOP>
  Assistant: X_4 <STOP>

  하지만 단순히 텍스트로만 이루어진 지시문은 다양성과 deep reasoning 이 부족함

→ 언어 모델과 시각 입력을 결합하여 시각적 정보를 반영한 명령 수행 데이터를 자동 생성하도록 설계됨.

→ 이미지를 텍스트로 텍스트로 인코딩하기 위해 두 가지 주요 symbolic representation 사용

1. Captions: 이미지를 여러 관점에서 설명하는 문장이다.
2. Bounding boxes: 이미지 내 객체의 위치를 명시

데이터 유형

1. Conversation:

   모델이 마치 인간처럼 이미지를 '보고 있는 것처럼' 질문하고 답하는 형식.

2. Detailed Description:

   이미지에 대한 풍부하고 상세한 설명을 요구. 하나의 이미지에 대해 여러 개의 질문 목록을 만들고, 그중 하나를 무작위로 선택하여 GPT-4에게 묘사를 생성하도록 요청

3. Complex Reasoning: 미지 속 상황을 기반으로 **심층적 추론(in-depth reasoning)**을 수행하도록 설계

Visual Instruction Tuning

Architecture

- 목표: 사전 학습된 LLM과 비전 모델의 능력을 모두 효과적으로 결합하기

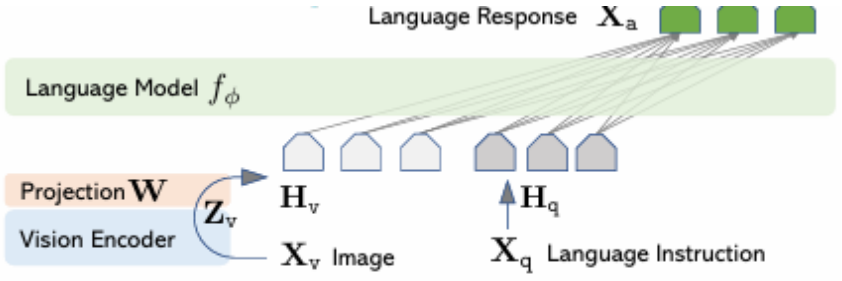

Figure 1: LLaVA network architecture.

  - **X_v** : 원본 이미지
  - **g(X_v)** : CLIP이 이미지를 벡터로 바꾼 결과 → **Z_v**
  - W(projection matrix): CLIP이 만든 벡터를 LLM이 이해할 수 있는 언어공간으로 바꿔주는 변환기
  - **H_v = W · Z_v** : 변환된 벡터 → LLM의 입력으로 들어가는 visual tokens

$$\mathbf{H}_\mathrm{v} = \mathbf{W} \cdot \mathbf{Z}_\mathrm{v}, \text{ with } \mathbf{Z}_\mathrm{v} = g(\mathbf{X}_\mathrm{v})$$

Training

각 이미지 X_v에 대해, 다중 턴 대화 형식의 학습 데이터를 생성. 즉, (X_v, X_a¹, ... , X_q^{T}, X_a^{T})의 형태로, 총 T개의 턴으로 구성.

- 각 턴에서 명령 입력 X_q$^{t}_{instruct}$:

$$\mathbf{X}_{\mathrm{instruct}}^{t} = \begin{cases} \text{Randomly choose } [\mathbf{X}_\mathrm{q}^1, \mathbf{X}_\mathrm{v}] \text{ or } [\mathbf{X}_\mathrm{v}, \mathbf{X}_\mathrm{q}^1], & \text{the first turn } t = 1 \\ \mathbf{X}_\mathrm{q}^t, & \text{the remaining turns } t > 1 \end{cases} \tag{2}$$

→ 생성된 데이터는 자동회귀(auto-regressive) 방식으로 LLM을 학습시키는 통합된 형식으로 변환

- L개의 토큰 길이를 가지는 시퀀스에 대해, 목표 응답 X_a의 확률:

$$p(\mathbf{X}_\mathrm{a}|\mathbf{X}_\mathrm{v}, \mathbf{X}_{\mathrm{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(x_i|\mathbf{X}_\mathrm{v}, \mathbf{X}_{\mathrm{instruct},<i}, \mathbf{X}_{\mathrm{a},<i}),$$

- 예시

  System-message <STOP>
  Human: X_1_instruct <STOP> Assistant: X_1_a <STOP>
  Human: X_2_instruct <STOP> Assistant: X_2_a <STOP> ...

  - 모델은 어시스턴트의 답변과 종료 시점을 예측하도록 훈련
  - 손실 계산 시에는 초록색으로 표시된 sequence/token들만 사용

두 단계 학습 절차

1. Pre-training for Feature Alignment

   데이터 커버리지와 효율성의 균형을 맞추기 위해 CC3M 데이터셋을 59만 5천 개의 이미지-텍스트 쌍으로 필터링.

   - 각 샘플은 단일 턴 대화로 구성
     - 입력 X_instruct: “이 이미지를 간단히 설명하라”는 명령어
     - 정답 X_a: 원본 캡션

이 단계에서 시각 인코더와 LLM의 가중치는 고정(frozen)된 상태로 유지되며, 투사 행렬 W만 학습된다.

⇒ 시각적 특징 H_v가 LLM의 단어 임베딩 공간과 정렬되도록 학습. LLM에 호환되는 비전 토크나이저를 학습하는 과정이라고 할 수 있음.

2. Fine-tuning End-to-End: 이 단계에서는 시각 인코더를 고정한 채,투사 행렬 W와 LLM 파라미터 φ를 함께 업데이트
    - **Multimodal Chatbot**: 15만 8천 개의 언어-이미지 명령 데이터를 사용하여 학습, 대화형, 세부 묘사, 복합 추론의 세 유형의 데이터를 포함
    - **Science QA**:시각 기반의 과학 질의응답 벤치마크 데이터셋을 추가로 사용, 모델이 언어적 지식뿐 아니라 시각적 근거를 통해 답을 도출하도록 설계

Experiment

실험 설정

- 8개의 A100 GPU
- Vicuna의 하이퍼파라미터 설정
- 전처리된 CC-595K 하위 집합을 이용해 1 epoch 동안 학습(학습률 2e-3, 배치 크기 128)
- LLaVA-Instruct-158K 데이터셋으로 3 epoch 동안 파인튜닝(학습률 2e-5, 배치 크기 32).

Multimodel Chatbot

이미지 이해와 대화 능력을 평가하기 위해 LLaVA 기반 챗봇 데모를 개발, 이 모델이 시각적 입력을 얼마나 잘 해석하고 명령 수행 능력을 보이는지 확인

- GPT-4 논문에 사용된 이미지 이해 예시를 그대로 가져와 실험을 수행

Extreme Ironing 예시

→ LLaVA는 이 이미지를 묘사할 뿐 아니라 그 비정상적 상황을 정확히 포착하며 “불안정한 환경에서 다림질하는 행위가 비정상적이다”라고 응답

→ 이는 단순한 이미지 설명에 머무른 BLIP-2나 OpenFlamingo보다 추론 능력이 좋다.

정량 평가 (Quantitative Evaluation)

GPT-4를 이용하여 생성된 응답의 품질을 측정, 각 샘플은 (이미지, 정답 캡션, 질문)으로 이루어진 삼중쌍(triplet)으로 구성.

- GPT-4는 인간 평가자 역할을 하여 모델의 응답과 실제 정답을 비교하고 “도움이 되는가(helpfulness)”, “정확한가(accuracy)”, “설명 수준은 어떤가(level of detail)” 로 **1~10점 척도**로 평가
- **LLaVA-Bench (COCO)**

  COCO-Val-2014에서 무작위로 30개의 이미지를 선택하고, 각 이미지에 대해 세 가지 유형의 질문(대화형, 세부 묘사, 복합 추론)을 생성하여 총 90개의 질문 세트를 구축
  - 서로 다른 유형의 명령 수행 데이터가 모델 성능에 어떤 영향을 미치는지 분석

    → 명령 튜닝이 없는 경우보다 50포인트 이상 성능 향상

    → 세부 묘사 및 복합 추론 데이터를 추가하면 추가로 약 7포인트 향상

    → **모든 유형의 데이터를 함께 사용할 때 최고 성능(85.1%) 달성**
- **LLaVA-Bench (In-the-Wild)**

  보다 도전적인 상황에서의 일반화 성능을 측정하기 위해 총 24개의 실제 장면(실내, 예술, 만화, 스케치 등)을 수집하고 각각에 대해 세밀한 주석과 질문을 작성
  - BLIP, OpenFlamingo와 비교했을 때 LLaVA는 각각 +29%, +48%의 상대적 향상
  - GPT-4(text-only)가 접근 가능한 정답 라벨(ground-truth)을 사용할 때 81.7%였던 반면, LLaVA는 67.3%

- **모델 한계 (Limitations):** 다중 모달 정보 통합(multi-modal retrieval) 이 충분하지 않고, 이미지를 단일 맥락에서 해석할 수 있지만 GPT-4 수준 폭넓은 지식 연동은 어려움

ScienceQA dataset

Model Ensembling

GPT-4와 LLaVA의 출력을 결합하는 두 가지 방식의 앙상블

- **GPT-4 complement:** GPT-4가 답을 생성하지 못할 때 LLaVA의 출력을 사용
- **GPT-4 judge:** 두 모델의 답을 비교해 GPT-4가 더 나은 답을 선택

GPT-4 judge 방식이 가장 높은 성능(92.53%)을 기록하며 새로운 SOTA를 달성.

Ablation Study

Visual features	Before	Last
Best variant	90.92	89.96 (-0.96)
Predict answer first	-	89.77 (-1.15)
Training from scratch	85.81 (-5.11)	-
7B model size	89.84 (-1.08)	-

Table 8: Design choice ablations (%). The difference with the best variant is reported in red text.

- 모델 설계 요소별 성능 차이
  1. Visual features - CLIP의 마지막 레이어 대신 그 이전 레이어를 사용할 경우 0.96% 향상
  2. Predict answer first - 12 epoch에서 89.77% 정확도
  3. Training from scratch - 사전 학습을 건너뛰고 직접 ScienceQA에서 학습하면 성능이 5.11% 하락
  4. Model Size - 13B 대신 7B 모델을 사용하면 1.08% 감소 (89.84%).

Conclusion

visual instruction tuning 의 효과를 입증

→ 언어-이미지 명령 수행 데이터를 자동으로 생성하는 파이프라인을 제시하였으며, 이를 기반으로 LLaVA라는 멀티모달 모델을 학습

- LLaVA는 ScienceQA에서 파인튜닝 되었을 때 새로운 SOTA 수준의 정확도를 달성
- 멀티모달 챗 데이터로 학습되었을 때 뛰어난 시각적 대화 능력
- multimodal instruction-following capability 을 연구하기 위한 최초의 벤치마크를 제시