10주차 요약

LLaVA

배경

Multimodal Model: 여러가지 다른 모달리티의 데이터를 동시에 이해하고 처리할 수 있는 인공지능

- In-context Learning (ICL): LLM 이 모델의 가중치 없데이트 없이 입력 프롬포트 내에서 제공되는 몇 가지 예시로부터 특정 태스크를 학습하는 능력
- Instruction Tuning: 사용자가 제공하는 자연어 insturduction 을 이해하고 그에 따라 원하는 작업을 수행하도록 하는 **방법론**

Motivation

여러 테스크를 범용적으로 수행할 수 있으며 멀티모달리티 데이터를 사용 가능하고 사람의 의도대로 지시를 따르는 모델을 만들자!

Method

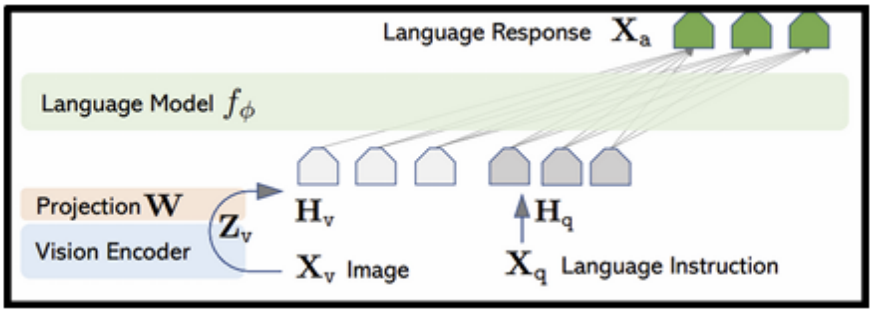
1. GPT-assisted Visual Insturduction Data Generation

   기존 image -text 데이터를 GPT 를 이용해서 instruction-following 형식의 데이터로 바꿈

   - 프롬포트의 context 에 캡션과 바운딩 박스 값들을 넣어 준다

   → 대화형식 데이터, 자세한 설명, 복잡한 추론 데이터 3가지 데이터 생성

2. Visual Instruduction Tuning



$$\mathbf{X}_{\texttt{system-message}}\ \texttt{<STOP>}\ \backslash\texttt{n}$$
$$\texttt{Human}: \mathbf{X}^1_{\texttt{instruct}}\ \texttt{<STOP>}\ \backslash\texttt{n}\ \texttt{Assistant:}\ \mathbf{X}^1_{\texttt{a}}\ \texttt{<STOP>}\ \backslash\texttt{n}$$
$$\texttt{Human}: \mathbf{X}^2_{\texttt{instruct}}\ \texttt{<STOP>}\ \backslash\texttt{n}\ \texttt{Assistant:}\ \mathbf{X}^2_{\texttt{a}}\ \texttt{<STOP>}\ \backslash\texttt{n}\ \cdots$$

- 초록 부분 토큰을 예측하도록 학습

1. Pre-training for Feature Alignment

   single turn 대화 데이터를 사용, 이미지에 대해 간략하게 설명하도록 요청하는 instruduction 를 주고 기존의 실제 캡션을 대답하도록 함.

   - 비주얼 인코더와 LLM weight 는 frozen
   - projection mtx 만 학습

2. Fine-Tuning End to End

   (1) 멀티모달 챗봇 데이터 또는 (2) Science QA 데이터 사용

   - 비주얼 인코더 frozen
   - LLM 과 projection layer 학습

Experiment

1. Multimodal Chatbot / LLaVA-Bench (COCO)

	Conversation	Detail description	Complex reasoning	All
Full data	83.1	75.3	96.5	85.1
Detail + Complex	81.5 (-1.6)	73.3 (-2.0)	90.8 (-5.7)	81.9 (-3.2)
Conv + 5% Detail + 10% Complex	81.0 (-2.1)	68.4 (-7.1)	91.5 (-5.0)	80.5 (-4.4)
Conversation	76.5 (-6.6)	59.8 (-16.2)	84.9 (-12.4)	73.8 (-11.3)
No Instruction Tuning	22.0 (-61.1)	24.0 (-51.3)	18.5 (-78.0)	21.5 (-63.6)

모든 유형의 데이터를 조합하여 학습했을 때 가장 높은 종합 성능. 특히 복합 추론 질문에서는 GPT-4 에 필적하는 점수

	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [28]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA†	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

다른 Vision-Language 모델들을 능가하는 Instrudction-Following 능력 갖춤

2. Science QA

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Representative & SoTA methods with numbers reported in the literature									
Human [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [34]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [59]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [61]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [61]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Results with our own experiment runs									
GPT-4†	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4† (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4† (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	**92.53**

→ LLaVA가 ScienceQA 벤치마크에서 단독으로도 매우높은성능(90.92%)을 달성하여 이전 SoTA에 근접

→ 나아가 LLaVA는 텍스트 전용 GPT-4를 "judge" 으로 활용하는 새로운앙상블 방식을 통해 92.53%의 정확도를 기록하며 이 데이터셋의 새로운SoTA가 됨

3. Abaltions

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)	-

ScienceQA 성능에 있어 시각 특징의 선택,추론과정의 순서,그리고 두단계의 사전학습, 모델의 크기가 모두 중요한 요소임