Visual Instruction Tuning

Haotian Liu[1*], Chunyuan Li[2*], Qingyang Wu[3], Yong Jae Lee[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

- Problem / objective
  - Propose Vision Language Model, called **LLaVA**

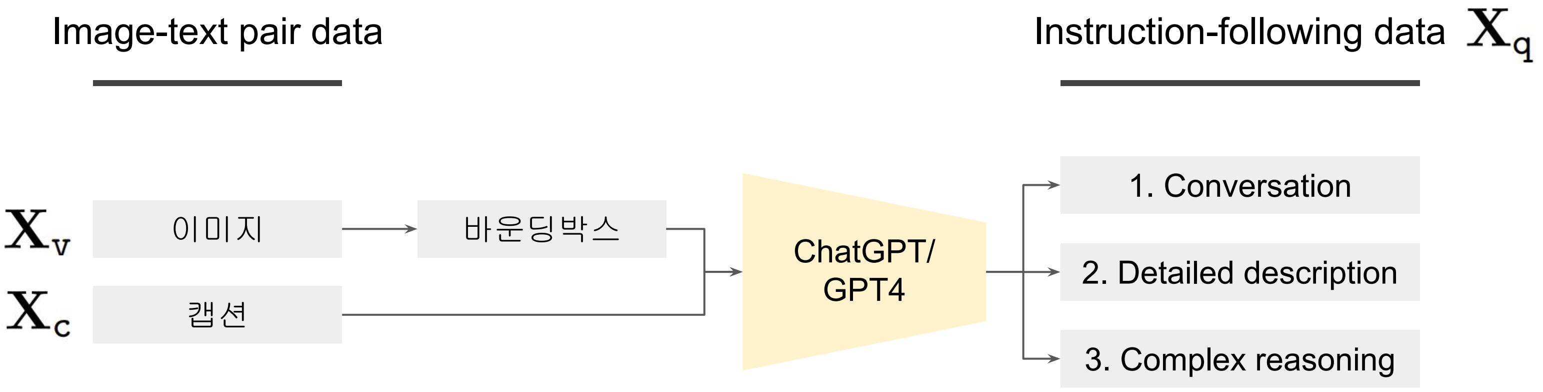
- Contribution / Key idea
  - Multimodal instruction-following data
  - Large multimodal models
  - Multimodal instruction-following benchmark
  - Open-source

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

NeurIPS 2023 Oral

GPT-assisted Visual Instruction Data Generation

- Instruction following 데이터
  - 형태 : $\mathbf{X}_{\mathrm{q}}\ \mathbf{X}_{\mathrm{v}}\texttt{<STOP>}\ \texttt{Assistant : }\mathbf{X}_{\mathrm{c}}\texttt{<STOP>}$
  - 3 종류 : 1. Conversation / 2. Detailed description / 3. Complex reasoning
  - 각각 58K, 23K, 77K 개씩 총 158K 개의 Instruction-following 데이터 생성.



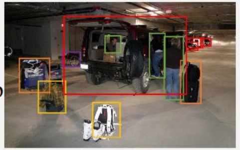
전유진

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

NeurIPS 2023 Oral

GPT-assisted Visual Instruction Data Generation

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Table 1: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.

전유진

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

Visual Instruction Tuning - Architecture

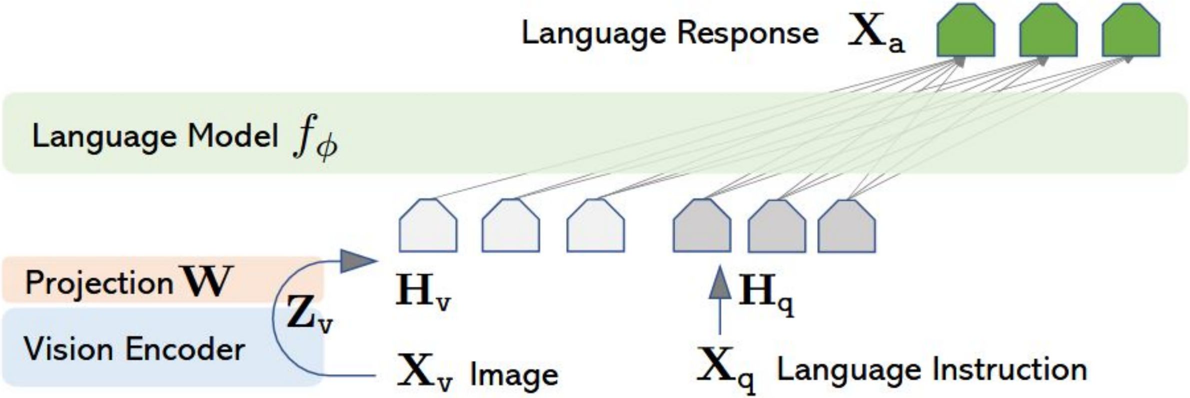


Figure 1: LLaVA network architecture.

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (

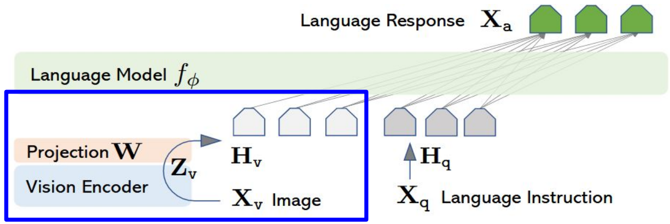
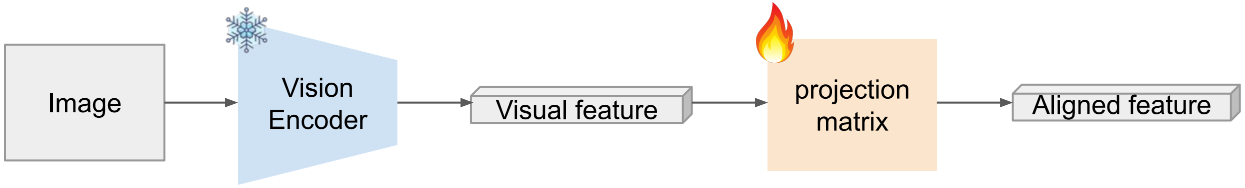


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Architecture

- Architecture



$\mathbf{X}_v$ $g(\cdot)$ $\mathbf{Z}_v = g(\mathbf{X}_v)$ $\mathbf{W}$ $\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v$

- How to connect 'image features' into 'the word embedding space'?
  - 선행 연구
    - Flamingo : gated cross-attention 제안.
    - BLIP-2 : Q-former 제안.
  - LLaVA
    - A single linear layer 통해 projection 하는 방법 제안.

[1] Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." *Advances in neural information processing systems* 35 (2022): 23716-23736.
[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2

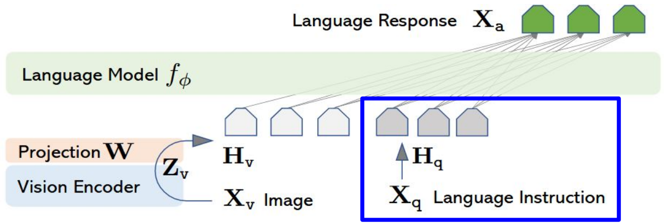


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- LLM 의 인풋 시퀀스

모델에게 역할을 알려주는 시작 프롬프트

각 질문/응답 경계를 구분하기 위한 토큰

Image

$\mathbf{X}_v$

Multi-turn conversation data

$(\boxed{\mathbf{X}_q^1}, \mathbf{X}_a^1, \cdots, \boxed{\mathbf{X}_q^T}, \mathbf{X}_a^T)$

$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}_a^1$ <STOP>
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}_a^2$ <STOP> · · ·

Table 2: The input sequence used to train the model. Only two conversation turns are illustrated here; in practice, the number of turns varies based on the instruction-following data. In our current implementation, we follow Vicuna-v0 [9] to set the system message $\mathbf{X}_{\texttt{system-message}}$ and we set <STOP> = ###. The model is trained to predict the assistant answers and where to stop, and thus only green sequence/tokens are used to compute the loss in the auto-regressive model.

1. '데이터 내 question' 은 'Human's instruction' 으로 구성. (단, 첫번째 turn 의 instruction 에는 이미지도 포함시킴.)

$$\mathbf{X}^t_{\texttt{instruct}} = \begin{cases} \text{Randomly choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1], & \text{the first turn } t = 1 \\ \mathbf{X}_q^t, & \text{the remaining turns } t > 1 \end{cases} \tag{2}$$

2. '데이터 내 answer' 은 'Assistant's response' 으로 구성.
3. 모델은 'Assistant's response' 와 '어디가 끝인지' 를 학습 (Table 2 에서 초록색 글씨)

전유진

[1] Chiang, Wei-Lin, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality." *See https://vicuna. lmsys. org (accessed 14 April 2023) 2.3 (2023): 6.*

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems 36 (2*

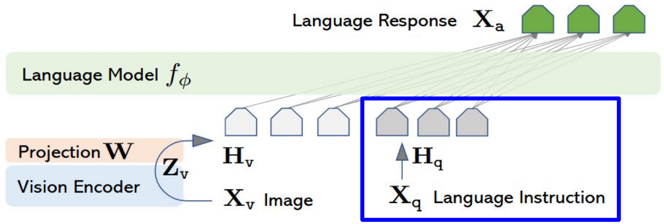


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- LLM 의 인풋 시퀀스

Image

$\mathbf{X}_v$

Multi-turn conversation data

$(\mathbf{X}_q^1, \boxed{\mathbf{X}_a^1}, \cdots, \mathbf{X}_q^T, \boxed{\mathbf{X}_a^T})$

모델에게 역할을 알려주는 시작 프롬프트

각 질문/응답 경계를 구분하기 위한 토큰

$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_a$ <STOP>
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_a$ <STOP> · · ·

Table 2: The input sequence used to train the model. Only two conversation turns are illustrated here; in practice, the number of turns varies based on the instruction-following data. In our current implementation, we follow Vicuna-v0 [9] to set the system message $\mathbf{X}_{\texttt{system-message}}$ and we set <STOP> = ###. The model is trained to predict the assistant answers and where to stop, and thus only green sequence/tokens are used to compute the loss in the auto-regressive model.

1. '데이터 내 question' 은 'Human's instruction' 으로 구성. (단, 첫번째 turn 의 instruction 에는 이미지도 포함시킴.)

$$\mathbf{X}^t_{\texttt{instruct}} = \begin{cases} \text{Randomly choose } [\mathbf{X}^1_q, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}^1_q], & \text{the first turn } t = 1 \\ \mathbf{X}^t_q, & \text{the remaining turns } t > 1 \end{cases} \quad (2)$$

2. '데이터 내 answer' 은 'Assistant's response' 으로 구성.
3. 모델은 'Assistant's response' 와 '어디가 끝인지' 를 학습 (Table 2 에서 초록색 글씨)

전유진

[1] Chiang, Wei-Lin, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality." *See https://vicuna. lmsys. org (accessed 14 April 2023) 2.3 (2023): 6.*

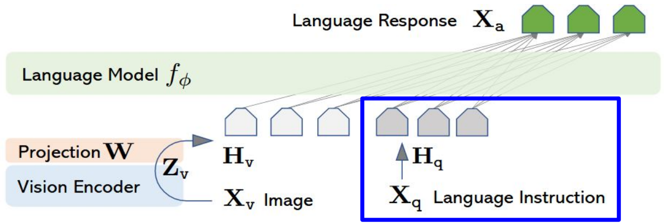


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- LLM 의 인풋 시퀀스

Image

$\mathbf{X}_v$

Multi-turn conversation data

$(\mathbf{X}_q^1, \mathbf{X}_a^1, \cdots, \mathbf{X}_q^T, \mathbf{X}_a^T)$

모델에게 역할을 알려주는 시작 프롬프트

각 질문/응답 경계를 구분하기 위한 토큰

$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_a$ <STOP>
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_a$ <STOP> · · ·

Table 2: The input sequence used to train the model. Only two conversation turns are illustrated here; in practice, the number of turns varies based on the instruction-following data. In our current implementation, we follow Vicuna-v0 [9] to set the system message $\mathbf{X}_{\texttt{system-message}}$ and we set <STOP> = ###. The model is trained to predict the assistant answers and where to stop, and thus only green sequence/tokens are used to compute the loss in the auto-regressive model.

1. '데이터 내 question' 은 'Human's instruction' 으로 구성. (단, 첫번째 turn 의 instruction 에는 이미지도 포함시킴.)

$$\mathbf{X}^t_{\texttt{instruct}} = \begin{cases} \text{Randomly choose } [\mathbf{X}^1_q, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}^1_q], & \text{the first turn } t = 1 \\ \mathbf{X}^t_q, & \text{the remaining turns } t > 1 \end{cases} \quad (2)$$

2. '데이터 내 answer' 은 'Assistant's response' 으로 구성.
3. 모델은 'Assistant's response' 와 '어디가 끝인지' 를 학습 (Table 2 에서 초록색 글씨)

전유진

[1] Chiang, Wei-Lin, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality." *See https://vicuna. lmsys. org (accessed 14 April 2023) 2.3 (2023): 6.*

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2

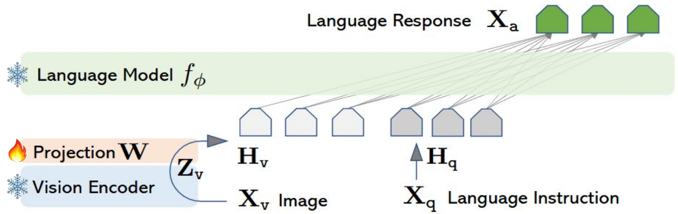


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- Stage 1: Pre-training for Feature Alignment.

1. CC3M 데이터셋의 (약 20%) 595K 개의 image-text pair 데이터 로부터 instruction-following 데이터 생성.

$$\mathbf{X}_q\ \mathbf{X}_v\texttt{<STOP>}\ \texttt{Assistant}: \mathbf{X}_c\texttt{<STOP>}$$

2. Visual Encoder 와 LLM 는 frozen 시키고, projection layer 만 학습.
   - Maximize following likelihood :

$$p(\mathbf{X}_a|\mathbf{X}_v, \mathbf{X}_{\texttt{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\mathbf{X}_v, \mathbf{X}_{\texttt{instruct},<i}, \mathbf{X}_{a,<i}), \quad (3)$$

   - 학습 파라미터 : $\boldsymbol{\theta} = \mathbf{W}$ (the projection matrix)
   - original caption 을 gt prediction $\mathbf{X}_a$ 으로 사용

전유진

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2

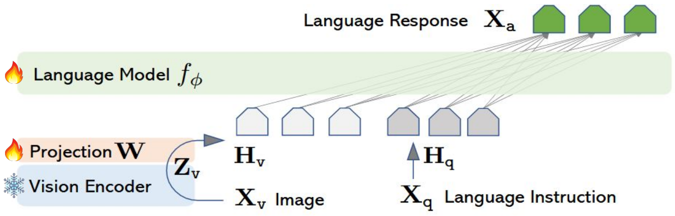


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- Stage 2: Fine-tuning End-to-End.

1. Visual Encoder 만 frozen 시키고, projection layer 와 LLM 학습.
   - Maximize following likelihood :

$$p(\mathbf{X}_a | \mathbf{X}_v, \mathbf{X}_{\mathtt{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i | \mathbf{X}_v, \mathbf{X}_{\mathtt{instruct},<i}, \mathbf{X}_{a,<i}), \quad (3)$$

   - 학습 파라미터 : $\boldsymbol{\theta} = \{\mathbf{W}, \phi\}$

2. 2가지 학습방법.
   - Multimodal Chatbot.
     - 158K 개의 instruction-following 데이터를 사용하여 파인튜닝.
     - 3가지 reponse 타입 (multi-turn 1개, single-turn 2개) 균등 샘플링 하여 학습.
   - Science QA.
     - single-turn 데이터로 학습. question & context as $\mathbf{X}_{\mathtt{instruct}}$, and reasoning & answer as $\mathbf{X}_a$

전유진

[1] Lu, Pan, et al. "Learn to explain: Multimodal reasoning via thought chains for science question answering." *Advances in Neural Information Processing Systems* 35 (2022): 2507-2521.

Experiments

- A100 지피유 8개로 학습.
- Vicuna 모델 하이퍼파라미터 그대로 사용.
- 학습 디테일
    - Stage 1: Pre-training for Feature Alignment.
        - 데이터셋 : filtered CC-595K dataset, 에포크 : 1, 학습률 : 2e-03, 배치사이즈 : 128.
    - Stage 2: Fine-tuning End-to-End.
        - 데이터셋 : LLaVA-Instruct-158K dataset, 에포크 : 3, 학습률 : 2e-05, 배치사이즈 : 32.
- 벤치마킹 2종류
    - Multimodal chatbot : Instruction-following 성능 평가.
    - ScienceQA : Visual reasoning 성능 평가.

[1] Chiang, Wei-Lin, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality." *See https://vicuna. lmsys. org (accessed 14 April 2023)* 2.3 (2023): 6.

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

Experiments - Multimodal Chatbot

- GPT-4 논문에서 사용한 예시 그대로 돌려봄. 비교를 위해 BLIP-2 와 OpenFlamingo 에도 해봄.

GPT-4 visual input example, Extreme Ironing:

User	What is unusual about this image?
	 Source: `https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg`
GPT-4	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Table 16. Example prompt demonstrating GPT-4's visual input capability. The prompt requires image understanding.

전유진

[1] Achiam, Josh, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).

Liu, H

systems 36 (2023): 34892-34916.

NeurIPS 2023 Oral

Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[Start a new conversation, and clear the history]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Table 3: Example prompt from GPT-4 paper [36] to compare visual reasoning and chat capabilities. Compared to BLIP-2 [28] and OpenFlamingo [5], LLaVA accurately follows the user's instructions, instead of simply describing the scene. LLaVA offers a more comprehensive response than GPT-4. Even when merely asked to describe the image, LLaVA identifies atypical aspects of the image.

- **LLaVA**

1. 80K 개 정도의 이미지 만으로 학습했는데, GPT-4 와 유사하게 잘하네.
2. LLaVA 의 out-of-domain 이미지인데도 잘하네.

- **BLIP-2, OpenFlamingo**

1. 이미지 묘사 정도만 잘하고, 사용자의 질문에 제대로 답하지를 못하네.

전유진

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

NeurIPS 2023 Oral

Experiments - Multimodal Chatbot

- LLaVA-Bench (COCO)
  - LLaVA 에서 새롭게 제안한 벤치마킹 방법1.
  - 모델의 alignment 능력 및 Instruction-following 성능 평가.
  - COCO-Val-2014 데이터셋에서 이미지 30개 랜덤 샘플링해서, 지피티 사용하여 생성한 90개 질문 (3개 카테고리) 으로 평가.

	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)

Table 4: Ablation on LLaVA-Bench (COCO) with different training data. We report relative scores *w.r.t.* a text-only GPT-4 model that uses ground truth image captions and bounding boxes as visual input. We prompt GPT-4 with the answers from our model outputs and the answers by GPT-4 (text-only), and let it compare between both responses and give a rating with an explanation.

: 역시 다 사용해서 학습해야 제일 좋네.

: Reasoning 능력 향상이 Conversation 능력에도 긍정적 영향을 주네. (상호보완)

: Stage 2 학습 해야겠다.

Experiments - Multimodal Chatbot

- LLaVA-Bench (In-the-Wild)
  - LLaVA 에서 새롭게 제안한 벤치마킹 방법2.
  - 어려운 task 및 새로운 도메인에 대한 generalizability 평가.
  - 다양한 도메인의 24개 이미지, 사람이 직접 만든 총 60개 질문으로 평가.

	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

Table 5: Instruction-following capability comparison using relative scores on LLaVA-Bench (In-the-Wild). The results are reported in the format of *mean* $\pm$ *std*. For the first three rows, we report three inference runs. LLaVA performs significantly better than others. † For a given set of LLaVA decoding sequences, we evaluate by querying GPT-4 three times; GPT-4 gives a consistent evaluation.

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

NeurIPS 2023 Oral

Experiments - ScienceQA

- 총 21K 개 이상의 멀티모달 객관식 과학 질문을 포함하는 벤치마크
- 다양한 도메인 : 3개 과목, 26개 주제, 127개 카테고리, 379개 능력
- train / val / test = 12,726 / 4,241 / 4,241

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**

Table 7: Accuracy (%) on Science QA dataset. Question categories: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. †Text-only GPT-4, our eval. Our novel model ensembling with the text-only GPT-4 consistently improves the model's performance under all categories, setting the new SoTA performance.

: SOTA 와 비슷.
: GPT-4가 실패한 경우에만 LLaVA의 출력을 사용
: GPT-4와 LLaVA가 다를 때, GPT-4에게 다시 물어 최종 정답 요청

전유진