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

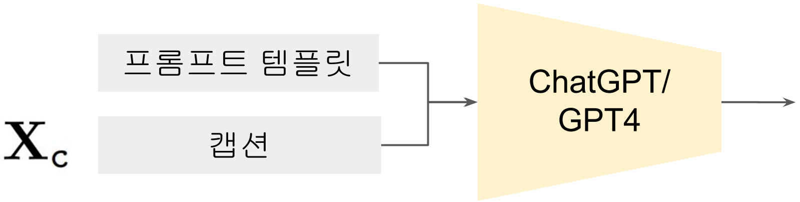
- Problem / objective
  - Propose Vision Language Model, called **LLaVA**

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

GPT-assisted Visual Instruction Data Generation

- **Instruction-following 데이터 생성 방법 1**
  - 형태 : $\mathbf{X}_{\mathrm{q}}\ \mathbf{X}_{\mathrm{v}}$<STOP> Assistant : $\mathbf{X}_{\mathrm{c}}$<STOP>
  - 간단하지만 다양성 및 깊이있는 추론 부족

Image-text pair data $\mathbf{X}_{\mathrm{v}}\ \mathbf{X}_{\mathrm{c}}$



Instructions $\mathbf{X}_{\mathrm{q}}$

- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo’s key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

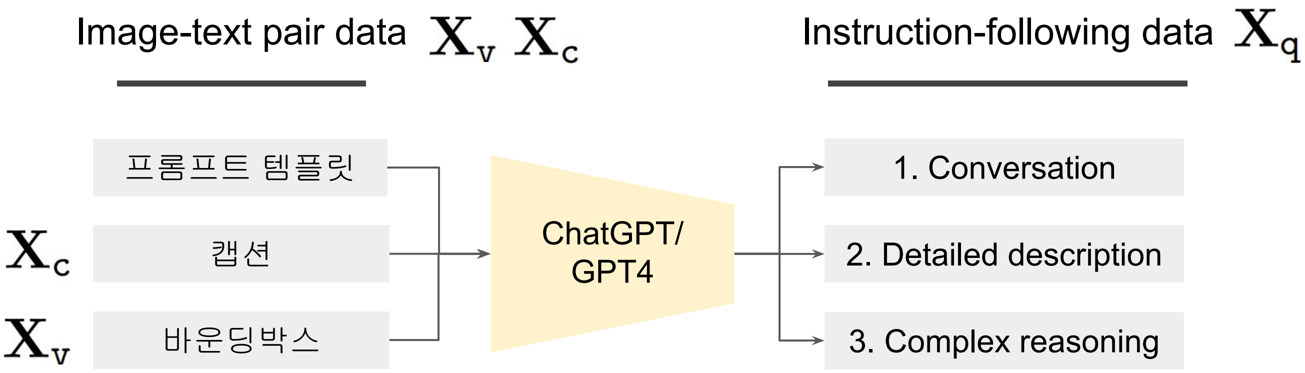
Table 11: The list of instructions for brief image description.

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 생성 방법 2**
  - 형태 : $\mathbf{X}_q\ \mathbf{X}_v$<STOP> Assistant : $\mathbf{X}_c$<STOP>
  - 종류 : 1. Conversation / 2. Detailed description / 3. Complex reasoning
  - 각각 58K, 23K, 77K 개씩 총 158K 개의 Instruction-following 데이터 생성.



[Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing sy*]

Instruction-following data - 1) Conversation

시스템 메시지

You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question.
Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image.
Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details.
Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

There is a movie theater that displays the show times above the doors.
A red fire hydrant is deep in the snow.
The fire hydrant is in the snow near a recently plowed sidewalk.
This city has had a very hard winter with snow.
A hotel for dogs in the snow in winter.

캡션

ChatGPT/
GPT4

Conversation 데이터

Question:
What color is the fire hydrant in the image?
===
Answer:
The fire hydrant in the image is red.
===
Question:
Is the fire hydrant fully covered by snow or partially visible?
===
Answer:
The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.
===
Question:
What is the condition of the sidewalk near the fire hydrant?
===
Answer:
The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.
===
Question:
What does the image suggest about the winter season in the city?
===
Answer:
The image suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the image indicates that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents.

For instance, the deep snow covering the fire hydrant could potentially hinder its accessibility in case of emergencies, which poses safety concerns. The recently plowed sidewalk implies that the city's maintenance crews have been actively working to keep the walkways clear and safe for pedestrians, but the sheer amount of snowfall might make it difficult to keep up with the continuous snow removal efforts.

Furthermore, such extreme winter conditions can impact transportation, leading to traffic issues, delays in public transit, and increased risks of accidents due to slippery roads. It can also cause problems for businesses and schools, as people might face challenges commuting to work or attending classes. Additionally, the heavy snow can put extra strain on infrastructure, such as roofs and power lines, increasing the likelihood of structural damage or power outages.

In conclusion, the image of the red fire hydrant deep in the snow and the recently plowed sidewalk suggest that the city has faced a particularly severe winter season, with substantial snowfall that has likely caused various challenges and disruptions for its residents and infrastructure.

ral

진

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

NeurIPS 2023 Oral

Instruction-following data - 2) Detailed description

시스템 메시지

You are an AI visual assistant that can analyze a single image. You receive five sentences, each describing the same image you are observing. In addition, specific object locations within the image are given, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.

Using the provided caption and bounding box information, describe the scene in a detailed manner.

Instead of directly mentioning the bounding box coordinates, utilize this data to explain the scene using natural language. Include details like object counts, position of the objects, relative position between the objects.

When using the information from the caption and coordinates, directly explain the scene, and do not mention that the information source is the caption or the bounding box. Always answer as if you are directly looking at the image.

캡션

A harbor filled with lots of boats next to a building.
A bicycle parked in front of several boats at a dock.
A red bicycle in front of a line of docked white yachts
A bike sits before boats which sit before a long building.
A bicycle is a convenient means of land transportation when you live on a boat.

바운딩박스

bicycle: [0.287, 0.641, 0.507, 0.874]
bicycle: [0.566, 0.667, 0.63, 0.731]
boat: [0.318, 0.579, 0.575, 0.724]
boat: [0.704, 0.607, 0.818, 0.727]
boat: [0.818, 0.601, 0.942, 0.744]
boat: [0.002, 0.53, 0.243, 0.71]
boat: [0.541, 0.611, 0.668, 0.731]
person: [0.778, 0.527, 0.797, 0.57]
cup: [0.708, 0.733, 0.724, 0.758]
boat: [0.236, 0.532, 0.404, 0.64]
boat: [0.81, 0.632, 0.836, 0.676]
boat: [0.957, 0.526, 1.0, 0.752]

ChatGPT/
GPT4

Detailed description 데이터

It is a harbor filled with numerous boats of various sizes docked next to a long building. Among the boats, there are a few white yachts lined up, standing out from the rest. There is a red bicycle prominently parked in front of the line of docked boats, serving as a convenient means of land transportation for those living on the boats. Another bicycle can be seen further back in the scene, near the middle of the harbor.

A person is visible near the right side of the harbor, possibly enjoying the view or attending to their boat. Additionally, there is a cup placed on a surface near the middle of the scene.

전유진

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

NeurIPS 2023 Oral

Instruction-following data - 3) Complex reasoning

You are an AI visual assistant that can analyze a single image. You receive five sentences, each describing the same image you are observing. In addition, specific object locations within the image are given, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.

The task is to use the provided caption and bounding box information, create a plausible question about the image, and provide the answer in detail.

Create complex questions beyond describing the scene.
To answer such questions, one should require first understanding the visual content, then based on the background knowledge or reasoning, either explain why the things are happening that way, or provide guides and help to user's request. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first.

Instead of directly mentioning the bounding box coordinates, utilize this data to explain the scene using natural language. Include details like object counts, position of the objects, relative position between the objects.

When using the information from the caption and coordinates, directly explain the scene, and do not mention that the information source is the caption or the bounding box. Always answer as if you are directly looking at the image.

시스템 메시지

A man wearing multiple neck ties making a goofy face.
A man in a white shirt wearing very many ties.
a man with ties on poses for a picture
A man wearing multiple ties on his neck.
A young man smiles while wearing several ties.

캡션

tie: [0.574, 0.298, 0.752, 0.704]
tie: [0.464, 0.339, 0.639, 0.789]
tie: [0.349, 0.363, 0.563, 0.732]
tie: [0.259, 0.255, 0.668, 0.805]
person: [0.019, 0.065, 0.962, 0.988]
person: [0.0, 0.24, 0.214, 1.0]
tie: [0.316, 0.778, 0.443, 0.867]
tie: [0.386, 0.707, 0.496, 0.801]
tie: [0.251, 0.354, 0.402, 0.747]
tie: [0.44, 0.318, 0.63, 0.397]
tie: [0.566, 0.33, 0.655, 0.543]
tie: [0.25, 0.409, 0.359, 0.556]

바운딩박스

ChatGPT/
GPT4

Complex reasoning 데이터

Question:
What is unusual about this photo?
===
Answer:
In the photo, the man is wearing a total of ten ties around his neck. This is considered unusual because it is not common for someone to wear multiple ties at once. In a typical setting, a person would wear only one tie. The man's goofy facial expression also adds to the unconventional and humorous nature of the image.

전유진

Visual Instruction Tuning - Architecture

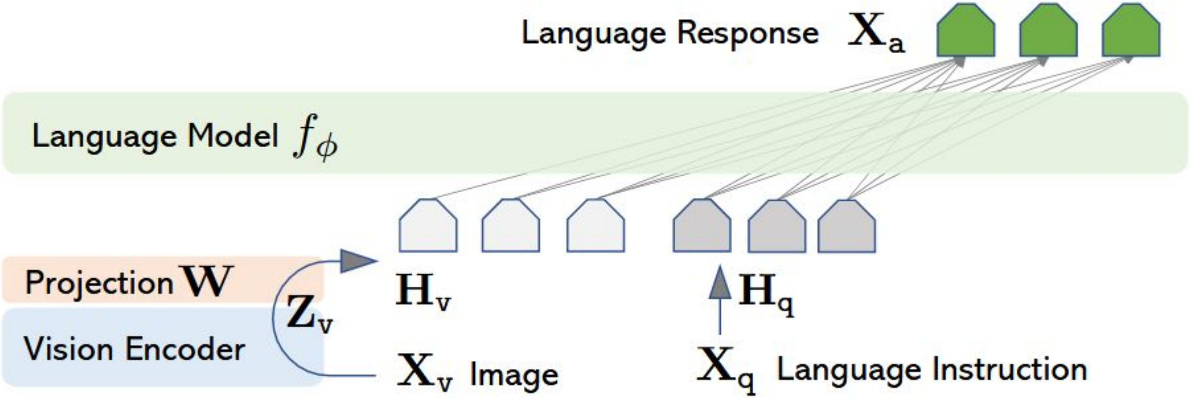


Figure 1: LLaVA network architecture.

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

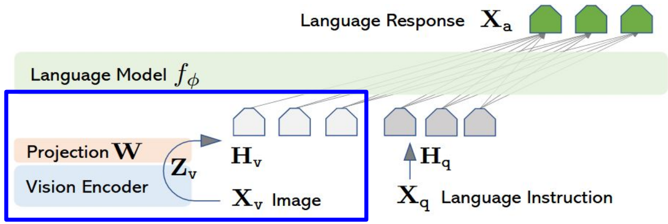
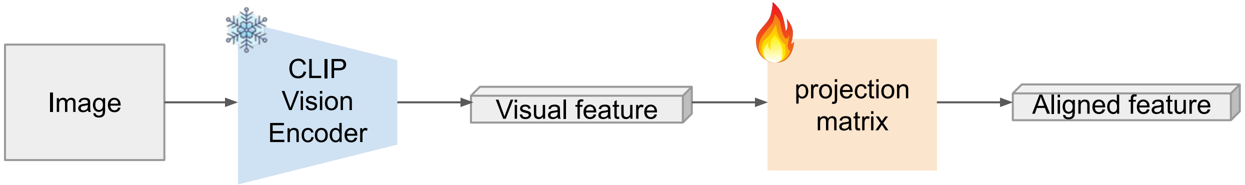


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Architecture

- Architecture



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

- 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 (2023): 34892-34916.](#)

NeurIPS 2023 Oral

How to connect 'image features' into 'the word embedding space'?

- Flamingo : Gated cross-attention

  LLM 의 각 트랜스포머 layer 사이에 visual feature 와 cross attention 하는 layer 삽입.

- BLIP-2 : Q-former

  중요한 visual feature 담게 학습된 Q-vector 들을 LLM 인풋에 concat.

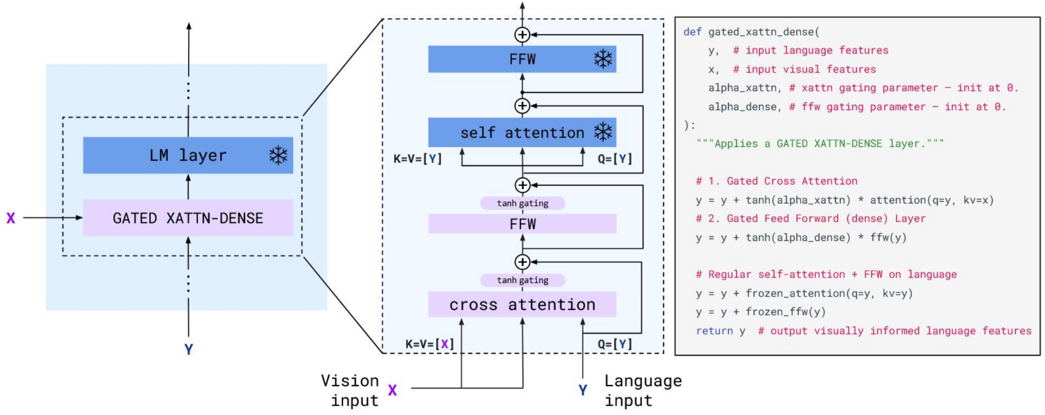


Figure 4: GATED XATTN-DENSE **layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

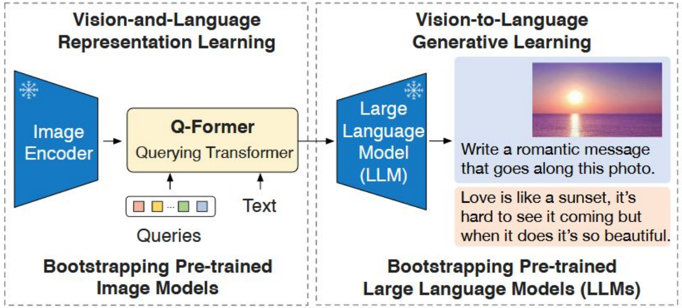


Figure 1. Overview of BLIP-2's framework. We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen LLM, which enables zero-shot instructed image-to-text generation (see Figure 4 for more examples).

[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

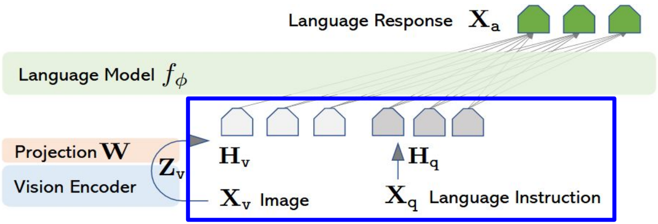


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- LLM 의 인풋 시퀀스

Image

$\mathbf{X}_v$

Multi-turn conversation data

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

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> $\cdots$

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 포함

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

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

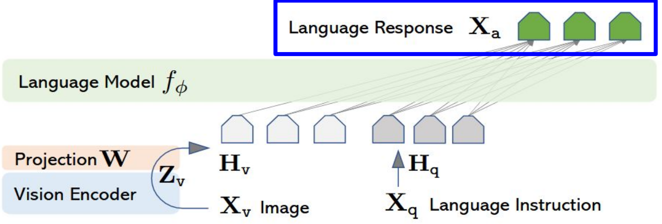


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- 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

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

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 $\mathbf{X}_{\text{system-message}}$ 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.)

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

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

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

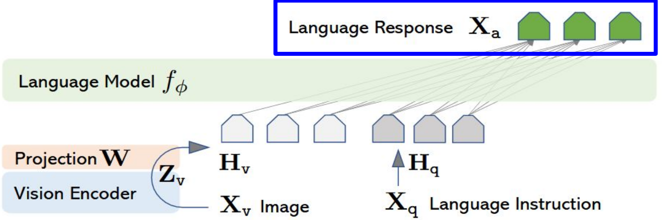


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- LLM 의 인풋 시퀀스

Image

$\mathbf{X}_v$

Multi-turn conversation data

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

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> $\cdots$

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.)

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

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

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

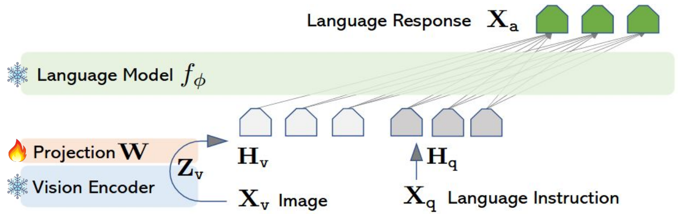


Figure 1: LLaVA network architecture.

Visual Instruction Tuning - Training

- **Stage 1: Pre-training for Feature Alignment.**

1. 595K 개의 instruction-following 데이터로 사전학습. (CC3M 데이터셋의 약 20% image-text pair 만 사용.) (방법1 통해 생성.)
2. Visual Encoder 와 LLM 는 frozen 시키고, projection layer 만 학습.
    - Ma

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

$$\boldsymbol{\theta} = \mathbf{W} \text{ (the projection matrix)}$$

    - 학습 파라미터 : $\mathbf{X}_a$
    - original caption 을 gt prediction 으로 사용

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

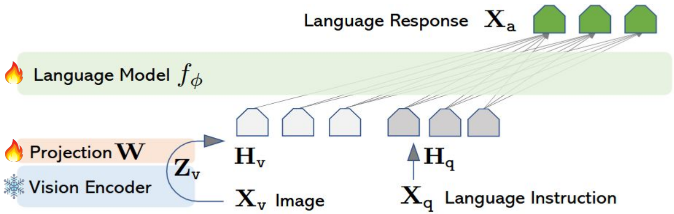


Figure 1: LLaVA network architecture.

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 :

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

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

2. Multimodal Chatbot.
   a. 158K 개의 instruction-following 데이터로 파인튜닝. (방법2 통해 생성.)
   b. 3가지 response 타입 균등 샘플링 하여 학습. (대화 : multi-turn, 상세한 설명 및 복잡한 추론 : single-turn 데이터)

3. Science QA.
   a. single-turn 데이터로 구성하여 벤치마킹.

question & context as $\mathbf{X}_{\mathtt{instruct}}$, and reasoning & answer as $\mathbf{X}_\mathrm{a}$

전유진

Experiments

- A100 GPU 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.
  - Science QA : 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.

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.

[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 잘한다.
2. LLaVA 의 out-of-domain 이미지인데도 잘한다.

- **BLIP-2, OpenFlamingo**

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

전유진

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

Experiments - Multimodal Chatbot

- LLaVA-Bench (COCO)
  - 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.
  - 평가방법 : GPT-4 judge : GPT-4 의 응답을 레퍼런스로 삼아 유사도 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 좋다.

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

: Stage 2 의 필요성.

Experiments - Multimodal Chatbot

- LLaVA-Bench (In-the-Wild)
  - 어려운 task 및 새로운 도메인에 대한 generalizability 평가.
  - 다양한 도메인의 24개 이미지, 사람이 직접 만든 총 60개 어려운 질문들로 평가.
  - 평가방법 : GPT-4 judge : GPT-4 의 응답을 레퍼런스로 삼아 유사도 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.

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 NAT	SOC	LAN	Context Modality TXT	IMG	NO	Grade G1-6	G7-12	Average
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