LLaVA : Large Langauge and Vision Assistant(Visual Instruction Tuning)

Haotain Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee

Dankook University
TaekminYoun
YoonjaeLee
sukmin Choi
YungjuKim

2025..06.05

Contents

1 Background

2 Methodologies

3 Experiments

4 Conclusion

Background

Background

Instruction tuning 등장 배경

Prompt-Completion Misalignment

- Language Model은 사용자의 의도와는 다른 답변을 주는 경우가 있음
- 예를 들어, 사용자가 '오늘 날씨 어때?'라고 물었을 때, 모델이 날씨와 관련 없는 정보나 일반적인 답변을 생성하는 경우를 말함

Instruction Tuning

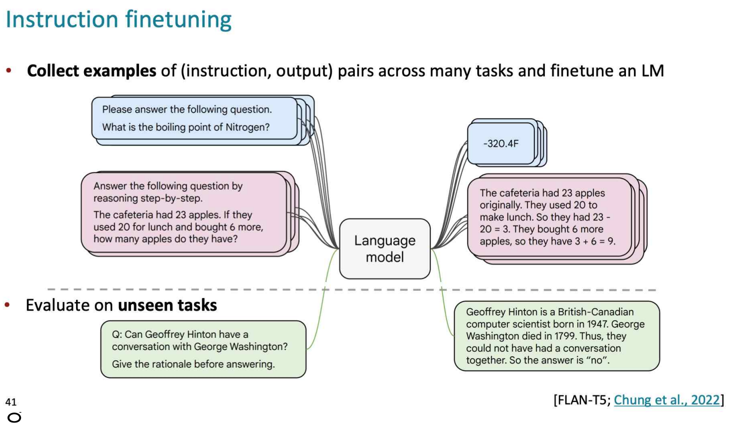
- 다양한 종류의 Task가 Instruction 형태로 들어 있는 Instruction, Output 쌍의 데이터 셋을 통해 Language Model을 Fine-tuning하는 방법
- 즉, Pretrained Model에 prompt와 completion 쌍의 데이터를 넣어 supervised learning을 수행하는 방법
  - Unseen Task에 대해 평가를 진행했을 때, Zero-shot 성능 향상

Background

Instruction Tuning의 장단점

- 장점
  - 간단하고 직관적인 방법을 통해 높은 성능 향상을 냄
  - Unseen Task까지 generalize할 수 있음

- 단점
  - 너무 많은 Task에 대한 demonstration을 수집하는데 많은 비용 발생
  - Language Model의 Objective와 사람의 Preference 사이의 mismatch가 있음
  - Hallucination 문제 발생
    - 문제를 보완하기 위해 RLHF 방법론 등장



Instruction finetuning

- **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM

Please answer the following question.
What is the boiling point of Nitrogen?

-320.4F

Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.

- Evaluate on **unseen tasks**

Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

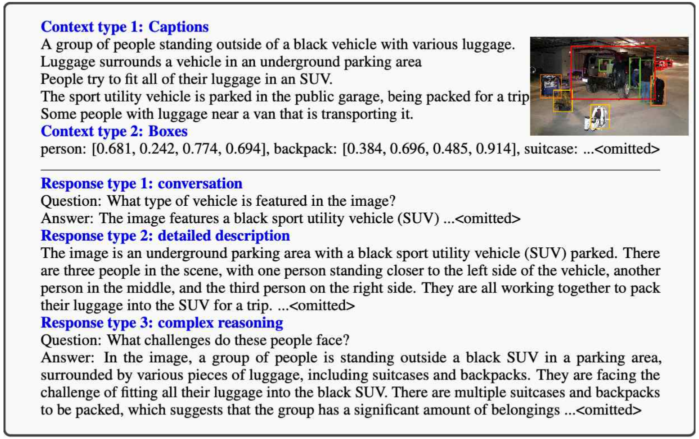
[FLAN-T5; Chung et al., 2022]

Methodologies

Methodologies

Multimodal Instruction-Following Dataset 생성

- 기존 CC, LAION 데이터셋은 단순한 Image Captioning에 그침
- LLaVA 모델 학습을 위한 Instruction-Following Dataset 생성 필요
  - 직접 생성하는 경우 많은 시간이 소요될 수 있음
  - Human Crowd-Sourcing을 하는 경우 데이터의 정의가 잘 이루어지지 않을 수 있음



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

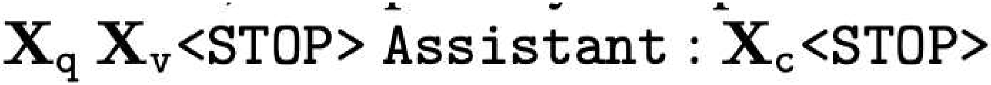
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their 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>

Methodologies

Multimodal Instruction-Following Dataset 생성



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

- 이미지 X_v와 Caption X_c가 있는 경우, 이미지를 서술해 달라는 내용을 질문 X_q로 한 데이터셋 생성
- 장점
  - 저비용으로 데이터셋 생성이 가능함
- 단점
  - 다양성 부족
  - Instructions과 Responses에서 심도 있는 reasoning 부족

-> 따라서 논문에서는 Image-Text Pair 데이터를 기반으로 ChatGPT/GPT4를 활용하여 Instruction-Following Dataset을 생성

Methodologies

Multimodal Instruction-Following Dataset 생성

- 문제점
  - ChatGPT/GPT4가 Visual Content를 인지할 수 없는 문제

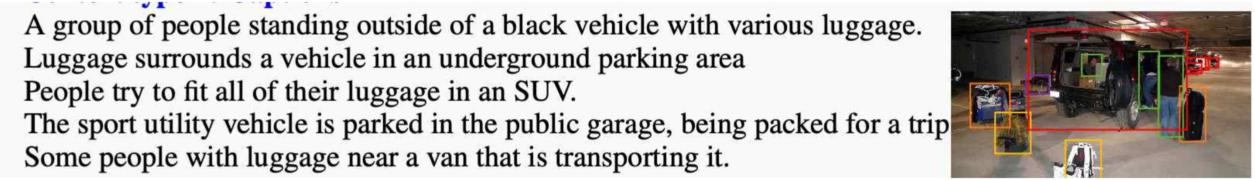
    ➔Symbolic Representations 방법

  - Instruction-Following Dataset의 구성 문제

    ➔Conversation, Detailed Description, Complex Reasoning 방법

Methodologies

Visual Content를 인지할 수 없는 문제 해결 방법
-> Symbolic Representations : Captions

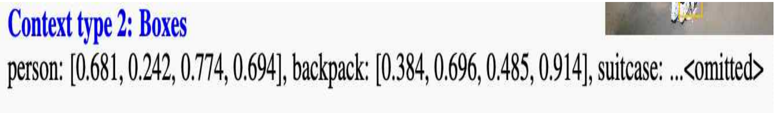


A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van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

- Language-Only GPT-4 / ChatGPT가 Visual Content를 포함한 Instruction-Following 데이터를 만들기 위해 활용
  - Captions는 시각적으로 다양한 관점에서 바라본 Image Scene

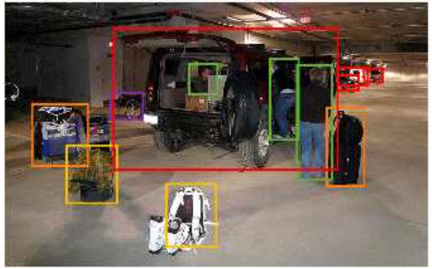
Methodologies

Visual Content를 인지할 수 없는 문제 해결 방법
-> Symbolic Representations : Bounding Boxes



Context type 2: Boxes

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

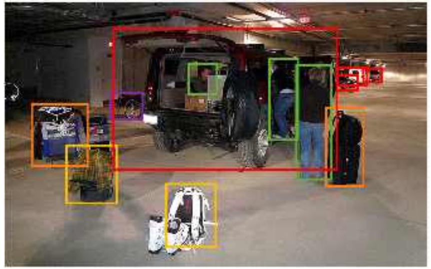


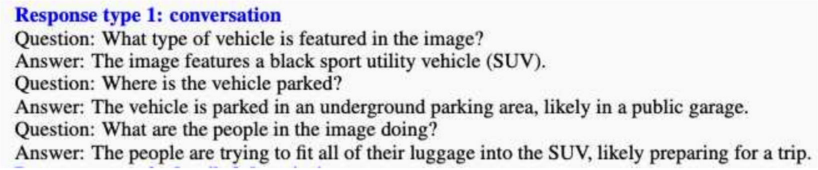
- Language-Only GPT-4 / ChatGPT가 Visual Content를 포함한 Instruction-Following 데이터를 만들기 위해 활용
  - Bounding Boxes는 Scene에서 특정 물체가 어디있는지 좌표를 통해 설명

Captions와 Bounding Boxes를 활용하여 Language-Only LLM이 인식할 수 있는 형태로 Image로 Encode

Methodologies

Instruction-Following 데이터 구성 문제
-> Conversation



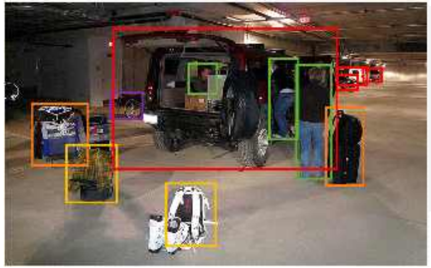


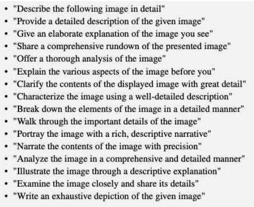
Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

- 주어진 이미지에 대해 Assistant와 Human이 대화하는 형태
- 이미지만 보고 알 수 있는것에 대한 Question-Answer 쌍 생성
- 이미지 내의 시각적 요소 자체에 대한 질문 포함
  - 객체의 종류, 개수, 동작, 위치, 객체 간의 상대적 위치 등

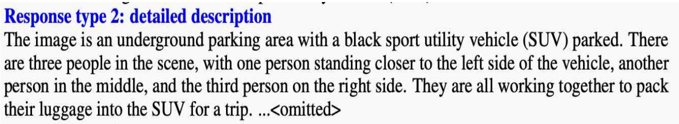
Methodologies

Instruction-Following 데이터 구성 문제
-> Detailed Description





- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before you"
- "Clarify the contents of the displayed image with great detail"
- "Characterize the image using a well-detailed description"
- "Break down the elements of the image in a detailed manner"
- "Walk through the important details of the image"
- "Portray the image with a rich, descriptive narrative"
- "Narrate the contents of the image with precision"
- "Analyze the image in a comprehensive and detailed manner"
- "Illustrate the image through a descriptive explanation"
- "Examine the image closely and share its details"
- "Write an exhaustive depiction of the given image"



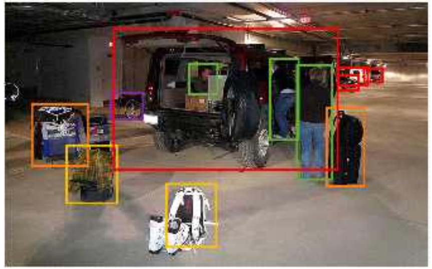
Response type 2: detailed description

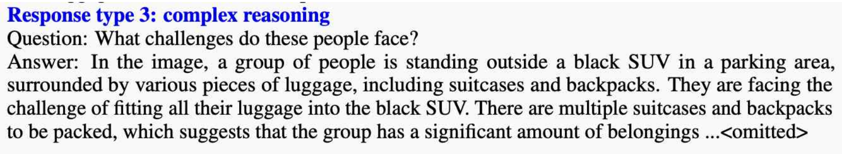
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their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

- Image에 대한 상세한 설명을 GPT-4를 통해 생성

Methodologies

Instruction-Following 데이터 구성 문제
-> Complex Reasoning





Response type 3: complex reasoning

Question: What challenges do these people face?

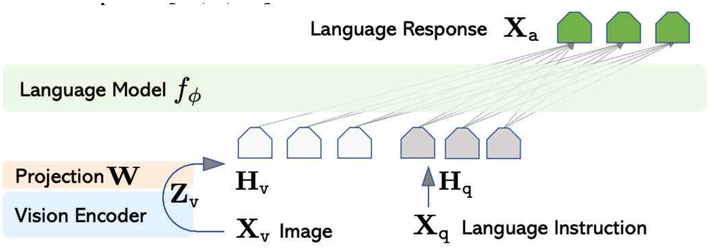
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed,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

- Conversation과는 달리 심층적인 추론을 하는 질문 및 답변을 GPT-4를 통해 생성
- 단계별로 엄격한 논리가 포함된 이유를 함께 요구

위 방법 이용하여 58K개의 Conversation, 23K개의 Detailed Description, 77K개의 Complex Reasoning 데이터 생성

Methodologies

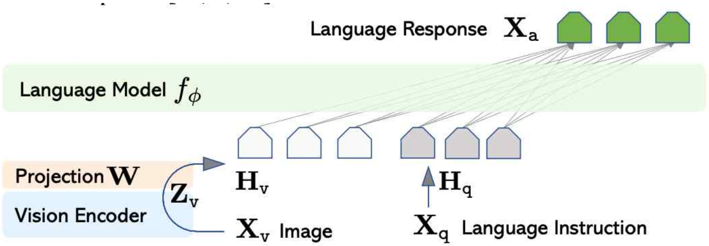
Architecture



- 목표
  - Pretrained LLM과 Pretrained Visual Model을 효과적으로 활용하는 것
- LLM
  - 언어 Tasks의 Checkpoints 중 가장 우수한 Instruction-Following 능력을 가진 Vicuna 활용
- Vision Encoder
  - Pretrained CLIP Visual Encoder인 ViT-L/14 활용

Methodologies

Architecture

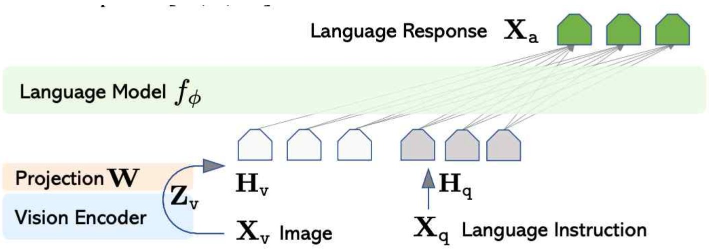


$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v)$$

- Vision Encoder
- Z_v : Vision Encoder를 활용하여 이미지 X_v를 Visual Feature화
- H_v : Language Model의 word embedding space와 동일한 차원을 갖도록 Z_v에 projection matrix인 W Project

Methodologies

Training

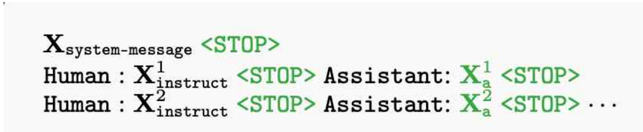


- Image Features와 Word embedding space의 결합에는 단순한 Linear Layer 사용
- Linear Layer 대신 Flamingo의 Gated Cross-Attention, BLIP-2의 Q-Former 사용하면 더욱 정교한 작업 가능

Methodologies

Training

Input Sequence



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

- Training Data 구축
- 각 이미지 X_v에 대해 Multi-turn Conversation Data확보
- 각 Conversation의 답변을 Assistant의 답변으로 간주

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

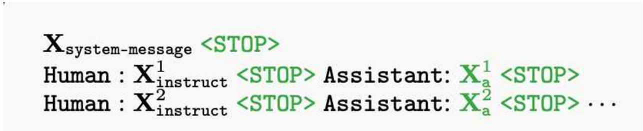
- t번째 instruction은 위와 같이 지정

위의 방법을 사용하면 일관된 형태로 Multimodal Instruction-Following Sequence 형성 가능

Methodologies

Training

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

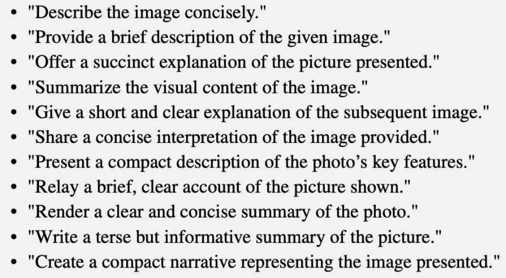


$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_{\mathrm{a}}$ <STOP>
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_{\mathrm{a}}$ <STOP> $\cdots$

- Probability 계산
  - Sequence의 길이가 L일 때, 정답 X_a에 대한 확률은 위의 수식과 같이 계산
    - $X_{instruct,<i}$: 현재 예측 토큰인 X_i 이전 모든 Turns에 대한 Instruction Tokens
    - $X_{a,<i}$: 현재 예측 토큰인 X_i 이전 모든 Turns에 대한 Answer Tokens
  - Auto-Regressive 모델에서 초록색 부분의 Sequence/Tokens만 Loss 계산에 이용됨
  - Stop토큰을 Loss로 사용하는 이유는 언제 문장 생성을 멈출지 알기 위해서임

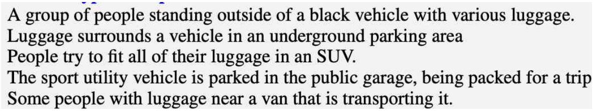
Methodologies

1. Pretraining for feature alignment



- "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."

Random select
->



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.

- CC3M Dataset을 위해 595000개의 image-text 쌍으로 필터링
- Dataset 생성 방법론에 따라 Instruction-Following Dataset으로 변환
- 이미지 X_v에 대해 간단한 요약을 요청하는 질문 X_q설정, ground truth prediction X_a는 그림의 원래 Caption으로 설정
- Visual Encoder와 LLM의 Weight는 freeze
- Trainable Parameters $\boldsymbol{\theta}$ = W(Projection Matrix)만을 이용하여 유사도 극대화

-> Image Features H_v와 Pretrained LLM의 Alignment을 맞추기 위한 방법으로 LLM에 대해 호환 가능한 Visual Tokenizer의 training 과정

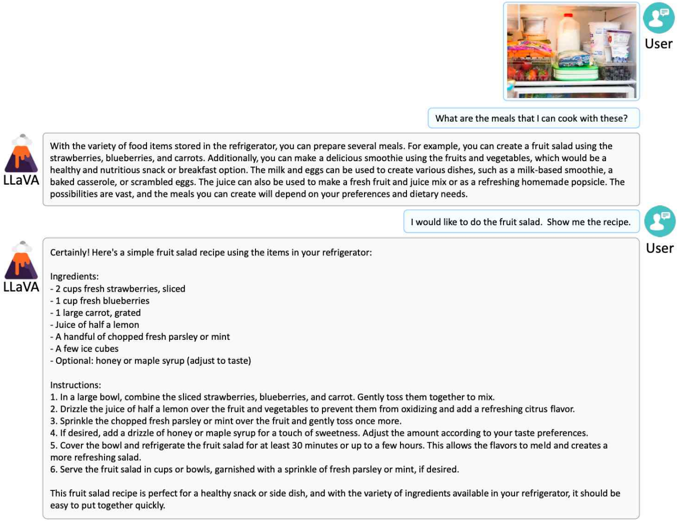
Methodologies

2. Fine-tuning

- Visual Encoder의 weight는 고정하고, Projection layer과 LLM 모델의 weight 업데이트
  - ($\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{\phi}\}$)

- 두 가지 시나리오를 위한 fine-tuning 진행

- Multimodal Chatbot
  - 앞서 생성한 158K개의 Language-Image Instruction-Following Dataset을 기반으로 Fine-tuning
- ScienceQA
  - Question과 Context을 $X_{instruct}$로, Reasoning와 Answer을 X_a로 정하여 학습



User

What are the meals that I can cook with these?

LLaVA

With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, you can make a delicious smoothie using the fruits and vegetables, which would be a healthy and nutritious snack or breakfast option. The milk and eggs can be used to create various dishes, such as a milk-based smoothie, a baked casserole, or scrambled eggs. The juice can also be used to make a fresh fruit and juice mix or as a refreshing homemade popsicle. The possibilities are vast, and the meals you can create will depend on your preferences and dietary needs.

I would like to do the fruit salad. Show me the recipe.

User

LLaVA

Certainly! Here's a simple fruit salad recipe using the items in your refrigerator:

Ingredients:
- 2 cups fresh strawberries, sliced
- 1 cup fresh blueberries
- 1 large carrot, grated
- Juice of half a lemon
- A handful of chopped fresh parsley or mint
- A few ice cubes
- Optional: honey or maple syrup (adjust to taste)

Instructions:
1. In a large bowl, combine the sliced strawberries, blueberries, and carrot. Gently toss them together to mix.
2. Drizzle the juice of half a lemon over the fruit and vegetables to prevent them from oxidizing and add a refreshing citrus flavor.
3. Sprinkle the chopped fresh parsley or mint over the fruit and gently toss once more.
4. If desired, add a drizzle of honey or maple syrup for a touch of sweetness. Adjust the amount according to your taste preferences.
5. Cover the bowl and refrigerate the fruit salad for at least 30 minutes or up to a few hours. This allows the flavors to meld and creates a more refreshing salad.
6. Serve the fruit salad in cups or bowls, garnished with a sprinkle of fresh parsley or mint, if desired.

This fruit salad recipe is perfect for a healthy snack or side dish, and with the variety of ingredients available in your refrigerator, it should be easy to put together quickly.

Experiments

Experiments

Pretrain

GPU : A100 * 8

Dataset : CC-595K Subset(Filtered from CC-3M)

Batch Size : 128

Epoch : 1

Learning Rate : 2e-3

Training Time : 대략 4 hours

Fine-tuning

GPU : A100 * 8

Dataset : LLaVA-Instruct-158K

Batch Size : 32

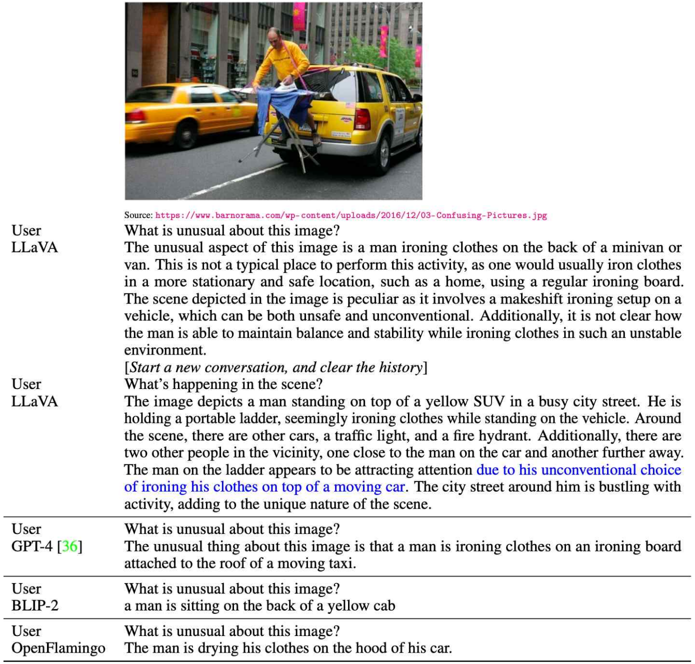
Epoch : 3

Learning Rate : 2e-5

Training Time : 대략 10 hours

Experiments

- 80K 정도로 적은 수의 image만을 학습했지만, 새로운 이미지에 대해서도 GPT-4와 매우 유사한 답변을 내는 것을 확인
- 단순히 이미지를 묘사하는 BLIP-2와 OpenFlamingo에 비해 LLaVA는 prompt에 맞는 정확도 높은 답변 제공

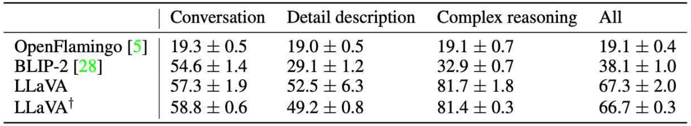


Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User LLaVA	What is unusual about this image? 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 LLaVA	What's happening in the scene? 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 GPT-4 [36]	What is unusual about this image? 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 BLIP-2	What is unusual about this image? a man is sitting on the back of a yellow cab
User OpenFlamingo	What is unusual about this image? The man is drying his clothes on the hood of his car.

Experiments

Quantitative Evaluation (Multimodal Chatbot)



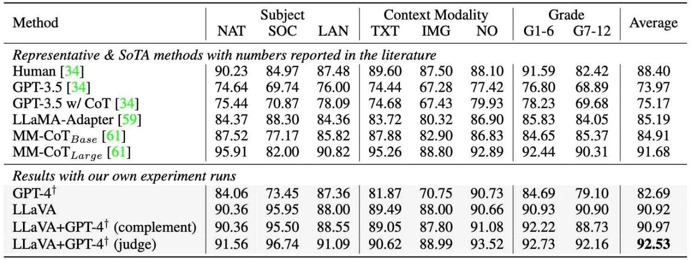
	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [28]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA†	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

- Text-Only GPT-4의 답변을 ground truth로 설정
- 평가 모델로부터 응답의 종합적인 평가 결과를 점수로 제공 받음

- Result
  - 학습되지 않은 다양한 조율의 데이터셋에 대해서도 BLIP-2, OpenFlamingo보다 우수한 성능을 보임
  - 특히 Complex reasoning에서 81.7%의 우수한 성능을 보임

Experiments

Quantitative Evaluation (ScienceQA)

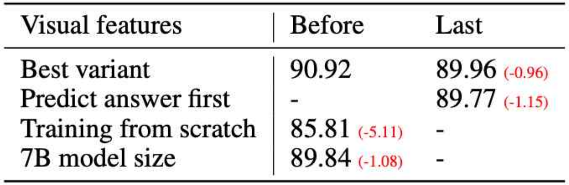


Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Representative & SoTA methods with numbers reported in the literature									
Human [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 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**

- ScienceQA의 Test dataset을 이용하여 평가
- Result
  - LLaVA 모델만 이용한 결과 90.92%로 SOTA에 근접한 성능을 냄
  - LLaVA와 GPT-4 답변이 다른 경우, GPT-4에게 두 답변을 넣고 최종 결론을 요구한 경우 92.53%로 SOTA 달성

Experiments

Ablation(ScienceQA)



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)	-

- ScienceQA에 일부 선택사항을 제거하며 평가
- Result
- Vision Encoder에 Last Layer Feature을 사용하는 경우, Before the Last Layer Feature을 사용하는 경우보다 0.96% 낮음
    - 이유는 마지막 layer는 그 직전 layer보다 일반적이고 추상적인 이미지 속성을 담고 있을 것으로 추정되기 때문
- Answer를 우선으로 예측하는 경우 12 Epochs에서 89.77%에 도달하지만, Reasoning을 우선 예측하는 경우 6 Epochs에서 89.77%에 도달
    - Reasoning을 먼저하는 CoT 전략이 학습 속도 개선에는 좋지만, 최종 성능에는 큰 영향을 미치지 않음
- Pretraining을 거치지 않은 경우, Pretraining을 한 경우보다 5.11% 성능이 더 낮음
    - Pretraining 중요성 강조
- 기존 13B Size의 모델을 7B Size로 낮춘 결과 1.08% 성능 감소
    - 모델 크기에 따른 성능 차이 존재

Conclusion

Experiments

- Multimodal Instruction Following Capability를 연구하기 위한 최초의 벤치마크 제안
- Vision Encoder로는 CLIP, Language Decoder로는 Vicuna를 결합하여 Vision 및 Language가 통합된 LLaVA 개발
- Language만을 이해하는 GPT-4를 사용하여 Instruction-following이라는 새로운 데이터셋을 생성함으로써 풍부하고 다양한 Multimodal 학습 데이터를 만들어내며 모델이 더 정교하게 시각적 상황을 이해하고 Language Instruction 수행
- Fine-Tuning시 Multimodal Chatbot 데이터셋에서 뛰어난 Visual Chat Capability를, ScienceQA 데이터셋에서 SOTA 달성