Visual Instruction Tuning

Haotian Liu, 2017

발표자: 전민하, 최수빈

1. 연구 배경 및 동기

기존 접근법의 한계

- 언어 강화 foundation vision 모델들
  - 각 태스크별로 독립적인 단일 대형 시각 모델 사용 -> 범용성 부족
  - 언어는 단순히 이미지 내용을 설명하는 데 활용
  - 고정된 인터페이스 -> 사람과의 상호작용 부족

LLM의 성공 사례

- ChatGPT, GPT-4가 보여준 가능성
  - 언어의 범용 인터페이스 역할
  - 다양한 태스크 명령을 언어로 명시적 표현
  - instruction-following 능력 즉, 사람이 내리는 명령을 이해하고 정확하게 행동하는 능력이 뛰어남
- but! **Text-only**

연구 목표

- Visual instruction-tuning: instruction-tuning을 언어-이미지 멀티모달 공간으로 확장

2. 주요 기여사항

1. **Multimodal instruction-following 데이터 생성**
   - 문제점: 시각-언어 instruction-following 데이터 부족
   - 해결책: 이미지-텍스트 쌍을 instruction-following 형태로 자동 전환하는 파이프라인 설계 (ChatGPT / GPT-4 활용)
   - 총 158K 샘플 생성 (대화형, 상세 설명, 복잡한 추론 포함)

2. **대규모 멀티모달 모델 개발 (LMM)**
   - 구성: CLIP 시각 인코더 + Vicuna 언어 디코더 연결
   - 학습: 생성된 데이터 기반 end-to-end fine-tuning
   - 성과: GPT-4과의 앙상블을 통해 ScienceQA에서 SOTA 달성

3. **LLaVA-Bench 벤치마크 제안**
   - 다양한 paired image-instruction-annotation 구성의 두 개 벤치마크 포함

4. **오픈소스 공개**
   - 데이터, 코드, 모델 체크포인트, 시각적 채팅 데모 공개

3. 관련 연구

멀티모달 Instruction-Following 에이전트

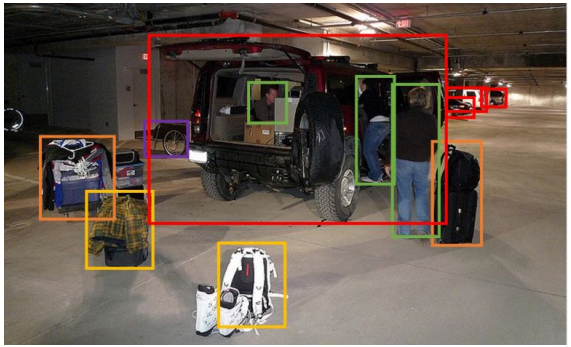
- 두 가지 범주:
  1. End-to-end 훈련 모델: 각 특정 연구 주제별로 개별 탐구 (예: Habitat, InstructPix2Pix)
  2. LangChain/LLM 기반 시스템: 다양한 모델들을 조합 (예: Visual ChatGPT, MM-REACT 등)

Instruction Tuning 발전

- NLP 분야: GPT-3 → InstructGPT, ChatGPT, FLAN-T5, FLAN-PaLM, OPT-IML 등
  a. 효과: LLM의 zero-shot 또는 few-shot 일반화 능력 향상
- 멀티모달 분야: Flamingo, BLIP-2, FROMAGe, PaLM-E, OpenFlamingo 등
- 오픈 소스: OpenFlamingo, LLaMA-Adapter
- 기존 한계: vision-language instruction 데이터가 명시적으로 튜닝되지 않아 성능 제한

4. GPT 기반 Visual Instruction 데이터 생성

이미지의 상징적 표현



1. **캡션**: 다양한 관점에서 시각적 장면 설명
2. **바운딩 박스**: 객체 개념과 공간적 위치 정보

구체적 예시 (COCO 데이터 활용)
Context 입력:

- **캡션**: "지하 주차장의 검은 SUV 주변에 서 있는 사람들과 다양한 짐들"
- **바운딩 박스**: person:[0.681, 0.242, 0.774, 0.694], backpack:[0.384, 0.696, 0.485, 0.914], suitcase:...

생성된 응답:

1. **대화형**: "Q: 이미지에 나타난 차량은 무엇인가요? A: 검은색 SUV입니다..."
2. **상세 설명**: "지하 주차장에 검은 SUV가 주차되어 있고, 세 명의 사람들이 여행을 위해 협력하여 짐을 싣고 있습니다..."
3. **복잡한 추론**: "Q: 이 사람들이 직면한 어려움은? A: 여러 개의 가방과 짐들을 SUV에 모두 넣어야 하는 공간 활용의 도전..."

4. GPT 기반 Visual Instruction 데이터 생성

세 가지 instruction-following 데이터 타입

1) 대화 (Conversation) - 58K

- 어시스턴트와 사진에 대해 질문하는 사람 간의 대화
- 이미지를 보고 답변하는 톤으로 응답
- 객체 유형, 객체 수, 행동, 위치, 상대적 위치 등 다양한 질문

2) 상세 설명 (Detailed Description) - 23K

- 이미지에 대한 풍부하고 포괄적인 설명
- GPT-4로 질문 목록 생성 후 선별
- 각 이미지마다 목록에서 무작위 질문 선택하여 설명 생성

3) 복잡한 추론 (Complex Reasoning) - 77K

- 시각적 내용 기반 심층 추론 질문
- 엄격한 논리를 따르는 단계별 추론 과정 요구
- 공간 추론 등 고품질 데이터 제공

총 158K개의 고유한 언어-이미지 instruction-following 샘플 생성

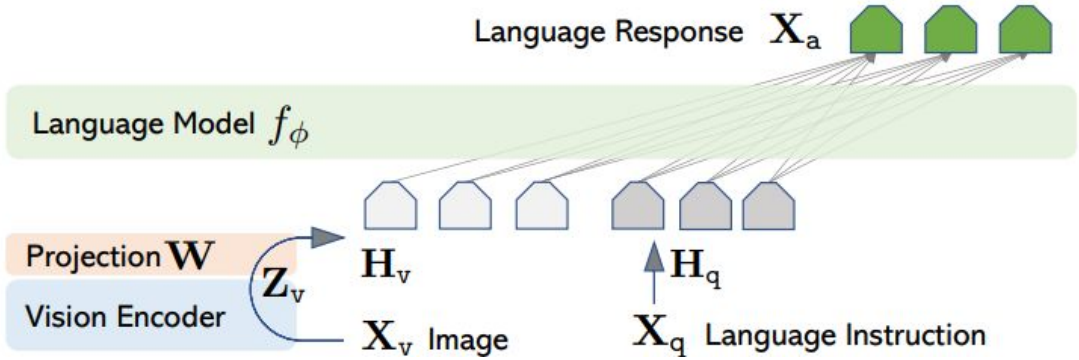
5. Visual Instruction Tuning 아키텍처 - 1. 네트워크 구조

네트워크 구조

- **언어 모델**: Vicuna (공개 가능한 체크포인트 중 최고 instruction following 성능)
- **시각 인코더**: 사전 훈련된 CLIP ViT-L/14
- **연결 방법**: 단순 선형 레이어로 이미지 특징을 단어 임베딩 공간에 연결

$H_v = W \cdot Z_v$, where $Z_v = g(X_v)$

- Z_v: 시각 특징
- W: 훈련 가능한 프로젝션 매트릭스
- H_v: 언어 임베딩 토큰과 동일 차원의 시각 토큰 시퀀스



5. Visual Instruction Tuning 아키텍처 – 2단계 훈련

2단계 훈련 전략

Stage 1: Feature Alignment를 위한 Pre-training

- **데이터**: CC3M에서 필터링된 595K 이미지-텍스트 쌍
- **변환 방법**: 단순 확장 방법으로 instruction-following 데이터 생성
- **훈련 방식**: 시각 인코더와 LLM 가중치 고정, 프로젝션 매트릭스(W)만 훈련
- **목적**: 이미지 특징 Hv를 사전 훈련된 LLM 단어 임베딩과 정렬

Stage 2: End-to-End Fine-tuning

- **데이터**: 158K 언어-이미지 instruction-following 데이터
- **훈련 방식**: 시각 인코더 고정, 프로젝션 레이어와 LLM 가중치 업데이트
- **훈련 파라미터**: $\theta = \{W, \phi\}$

5. Visual Instruction Tuning 아키텍처 - 멀티턴, 시나리오

멀티턴 대화 처리

- 각 이미지에 대해 멀티턴 대화 데이터 생성
- 모든 답변을 어시스턴트 응답으로 처리
- 첫 번째 턴에서 [질문, 이미지] 또는 [이미지, 질문] 무작위 선택
- 나머지 턴에서는 질문만 사용

두 가지 활용 시나리오

1) 멀티모달 챗봇

- 158K 언어-이미지 instruction-following 데이터로 파인튜닝
- 대화형은 멀티턴, 나머지 두 타입은 단일턴으로 구성
- 훈련 시 균등하게 샘플링

2) Science QA

- ScienceQA 벤치마크: 최초 대규모 멀티모달 과학 질문 데이터셋
- 상세한 강의와 설명이 주석으로 제공
- 자연어 또는 이미지 형태의 맥락 제공
- 자연어로 추론 과정 제공 후 다중 선택지에서 답 선택

6. 기술적 특징

아키텍처 설계 철학

- **단순성**: 간단한 선형 프로젝션으로 빠른 데이터 중심 실험 가능
- **모듈성**: 사전 훈련된 LLM과 시각 모델의 능력 효과적 활용
- **확장성**: 더 정교한 연결 방식으로 향후 개선 가능

훈련 효율성

- **단계별 접근**: Feature alignment → Instruction tuning
- **선택적 업데이트**: 각 단계에서 필요한 파라미터만 업데이트
- **데이터 활용**: 기존 이미지-텍스트 쌍 데이터의 효율적 재활용

데이터 생성 혁신

- **자동화**: 인간 주석 최소화 (시드 예시만 수동 생성)
- **품질**: GPT-4가 ChatGPT보다 일관되게 높은 품질의 데이터 생성
- **다양성**: 세 가지 타입으로 다양한 능력 커버

실험

- **Multimodal Chatbot과 ScienceQA 데이터 셋을 통해 Llava의 능력을 평가**
  - 사용모델 : vicuna-13B
  - A100 GPU * 8 사용

- **1단계 – 사전학습 (Pre-training) 진행**
  - 데이터셋: CC-595K (CC3M에서 필터링된 이미지-텍스트 쌍 595K개)
  - 학습 횟수: 1 epoch
  - 학습률: 2e-3 (0.002)
  - 배치 크기: 128
  - 목적: 시각 정보와 언어 표현의 정렬

- **2단계 – 미세조정 (Fine-tuning) 진행**
  - 데이터셋: LLaVA-Instruct-158K (GPT가 만든 instruction-following 데이터)
  - 학습 횟수: 3 epochs
  - 학습률: 2e-5 (0.00002)
  - 배치 크기: 32
  - 목적: GPT처럼 자연스럽게 대화하며 지시를 따를 수 있게 만듦

실험 – Multimodal Chatbot

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.

- 비교를 위해 GPT-4, BLIP-2와 OpenFlamingo 모델에도 동일한 질문 입력

- 의의
  - BLIP-2 및 OpenFlamingo와 비교했을 때, LLaVA는 사용자의 지시를 정확하게 수행
  - GPT-4처럼 장면을 단순히 묘사하는 대신. LLaVA는 보다 더 포괄적인 응답 제공
  - Llava는 상대적으로 적은 데이터셋(약 80K의 고유 이미지)으로 훈련되었음에도 내용적으로 GPT-4와 매우 유사한 추론 결과 도출
  - 즉, Llava는 이미지를 잘 이해하고 질문 지시에 따라 합리적인 응답을 제공함을 확인

실험 – 정량적 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)

LLaVA-Bench(coco)

Coco-val-2014 데이터 셋에 각각의 instruction tuning을 했을 때의 결과값 나열

Instruction tuning을 많이 할수록 대화 품질과 복잡한 질문 대응 능력이 높아짐을 확인할 수 있음.

LLaVA-Bench(In-the-Wild)

Coco 데이터셋처럼 정형화된 데이터가 아닌, 인터넷에 떠돌아다니는 무작위 이미지들로 생성한 테스트 셋으로 실험 진행

타 모델에 비해서 LLaVa가 높은 성능을 발휘함을 기록 즉, 실생활에서도 폭넓게 사용 가능함을 의미함.

	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

실험 – 정량적 평가

Challenging examples from LLaVA-Bench (In-the-Wild):

	 ICHIRAN Ramen [source]	 Filled fridge [source]
Annotation	A close-up photo of a meal at ICHIRAN. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with chili sauce, chopped scallions, and served with two pieces of chashu. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of orange spice (possibly garlic sauce), a plate of smoke-flavored stewed pork with chopped scallions, and a cup of matcha green tea.	An open refrigerator filled with a variety of food items. In the left part of the compartment, towards the front, there is a plastic box of strawberries with a small bag of baby carrots on top. Towards the back, there is a stack of sauce containers. In the middle part of the compartment, towards the front, there is a green plastic box, and there is an unidentified plastic bag placed on it. Towards the back, there is a carton of milk. In the right part of the compartment, towards the front, there is a box of blueberries with three yogurts stacked on top. The large bottle of yogurt is Fage non-fat yogurt, and one of the smaller cups is Fage blueberry yogurt. The brand and flavor of the other smaller cup are unknown. Towards the back, there is a container with an unknown content.
Question 1	What's the name of the restaurant?	What is the brand of the blueberry-flavored yogurt?
Question 2	Describe this photo in detail.	Is there strawberry-flavored yogurt in the fridge?

Llava-Bench(in-the-wild)의 한계

1. 이치란 라멘의 사진(왼쪽)처럼 외국어 식당 이름을 답하기 위해선 다국어에 대한 이해 능력이 필요함.

2. 냉장고 사진(오른쪽)에서 요거트 브랜드를 맞추기 위해서는, 고해상도 이미지를 처리 및 유제품 브랜드들에 대한 이해도가 필요함

위와 같은 문제들 때문에 예시이미지처럼 각각에 해당하는 각주들을 달아줘야 질문에 대한 답을 얻어낼 수 있었음.

실험 - 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**

- 기존 문헌들 중 SoTA가 91.68이었는데, 그것에 거의 근접한 수치 90.92 기록
- 또한 LLM의 한계를 보기 위해 GPT-4에게 2-shot in-context-learing을 시켰을 때 82.96 기록
- GPT-4가 답변을 제공하지 못했을 때 LLaVA와 결합하여 90.97 기록
- GPT-4와 LLaVA의 답이 다를 경우 두개의 답을 토대로 다시 GPT-4에게 답변 요청. 92.53으로 새로운 SoTA 달성

실험 - Ablations

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 데이터 셋을 대상으로 제한된 실험 진행

- **Visual features** : CLIP 의 마지막 레이어를 사용하였을 때 이전 레이어 대비 0.96 감소한 89.96 기록.
- 이유는, 마지막 레이어는 전역적이고 추상적인 정보에 집중하고, 그 전 레이어는 국소적이고 세부적인 특징을 더 잘 반영하기 때문으로 추측.

- **Chain-of-thoughts** : 답변 우선 방식에 비해 추론 우선 방식이 빠르게 89.77의 성능을 기록했지만, 이후 성능 향상이 없었음.
- 즉, 추론 우선 방식은 학습 속도에 도움이 되지만 최종 성능에 큰 영향을 미치지 않음을 확인.

- **Pre-training** : Pre-training을 건너뛰고 바로 ScienceQA를 학습시켰을 때 85.81%로 감소.

- **Model size** : 모든 구성을 동일하게 유지시키고 13B->7B로 바꾸었을 때 90.92% -> 89.84%로 감소

결론

- Visual instruction Tuning의 효과 입증
- Multimodal model LLaVA를 훈련했으며, fine-tunin을 통해 새로운 SoTA를 달성
- Multimodal ChatBot을 통해 뛰어난 시각 채팅 능력을 확인할 수 있었음.