InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning

https://arxiv.org/abs/2305.06500

0. Introduction

- 시각 언어 모델은 이미지와 텍스트를 함께 처리하며 캡셔닝이나 VQA 같은 다양한 태스크를 해결함
- 기존 모델은 멀티태스크 학습으로 성능을 올리지만 새로운 태스크로 넘어가면 일반화가 잘 안 됨
- 이유는 시각 인코더가 instruction과 무관하게 항상 같은 방식으로 이미지를 처리하기 때문임
- 예를 들어 이미지를 설명하는 태스크와 참 거짓을 판별하는 태스크가 달라도 똑같은 피처가 나와서 LLM이 태스크 의도에 맞게 반응하기 어려움
- InstructBLIP은 이런 한계를 해결하려고 제안된 모델임
- BLIP-2 구조를 바탕으로 Q-Former를 instruction-aware하게 바꿔서 시각 표현을 태스크에 맞게 뽑도록 설계함
- 핵심은 instruction tuning을 시각 언어 모델에도 적용해 unseen 태스크에서 제로샷 성능을 크게 개선하는 것임
- 목표는 범용적인 시각 언어 모델을 만드는 것임

1. Overview

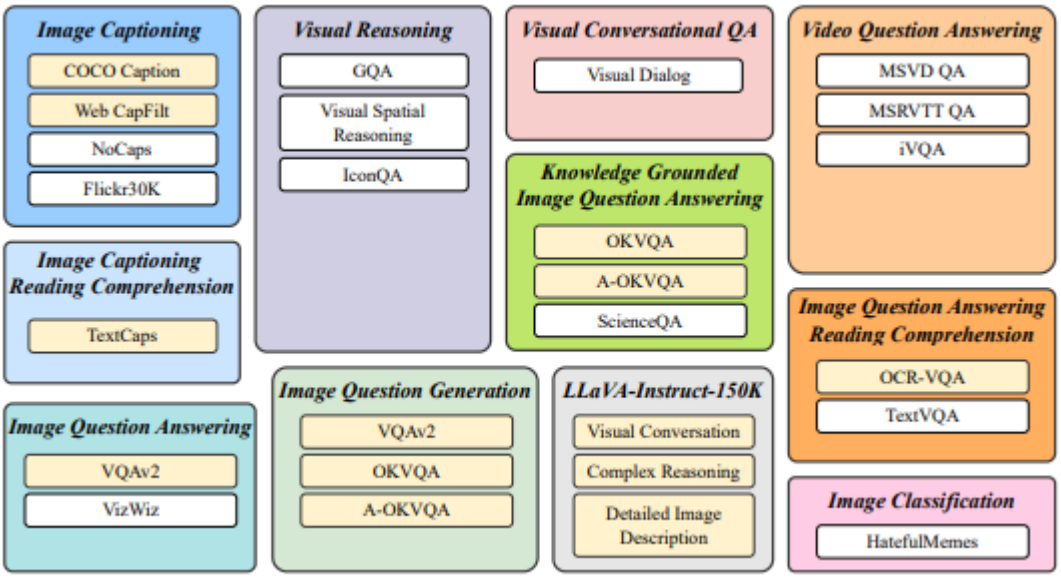
- InstructBLIP은 BLIP-2 구조를 기반으로 함. 이미지 인코더와 LLM은 그대로 두고 Q-Former만 학습함. 효율성과 확장성을 동시에 확보한 방식임.
- Q-Former는 단순히 이미지 피처만 뽑는 게 아니라 instruction 토큰도 같이 받아서 태스크에 맞는 시각 표현을 생성함. 같은 이미지를 보더라도 지시문이 다르면 다른 피처가 나오도록 설계된 거임.

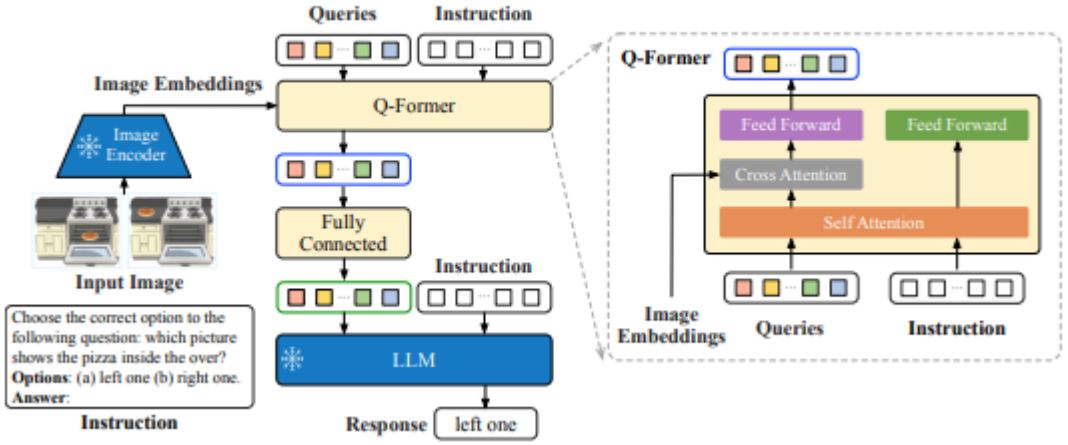
- 학습 데이터는 총 26개 공개 데이터셋으로 구성됨. 11개 카테고리로 묶어서 절반은 학습에 쓰고 나머지는 제로샷 평가용으로 활용함.
- 모든 데이터셋은 instruction 형태로 변환함. 캡션은 "이 이미지를 설명하라" 식으로 바꾸고 QA는 "이 질문에 답하라" 형태로 바꿔서 모델이 지시문 자체를 이해하게 함.
- 데이터셋 크기 차이가 크기 때문에 balanced sampling을 적용해 특정 대규모 데이터셋에 치우치지 않도록 조정함.
- 여러 버전의 모델을 만들 수 있도록 FlanT5와 Vicuna 같은 다양한 LLM을 붙여서 실험함.

2. Challenges

- 시각 언어 데이터셋은 태스크와 포맷이 제각각이라 하나의 모델로 통합하기 어려움
- 멀티태스크 학습을 해도 학습에 포함되지 않은 태스크로 가면 성능이 급격히 떨어짐
- 이미지 인코더가 instruction을 고려하지 않고 항상 같은 피처를 내놓기 때문에 태스크 차이에 대응하기 어려움
- 데이터셋 크기 차이가 커서 그대로 학습하면 큰 데이터셋 위주로만 학습됨
- LLM 전체를 학습시키면 파라미터가 너무 많아 계산 비용이 비싸고 효율성이 떨어짐

3. Method





- InstructBLIP은 BLIP-2 기반 모델임. 이미지 인코더와 LLM은 동결하고 Q-Former만 instruction-aware하게 학습함
- Q-Former는 instruction 토큰을 받아 이미지 피처와 상호작용함. 같은 이미지라도 instruction에 따라 다른 시각 표현을 생성함
- 학습 데이터는 26개 공개 시각-언어 데이터셋으로 구성됨
- 모든 데이터셋을 instruction 포맷으로 변환함. 캡션은 “이 이미지를 설명하라”, QA는 “질문에 답하라” 식으로 바꿔 모델이 지시문을 이해하도록 함
- 각 데이터셋마다 10개 이상의 템플릿을 만들어 같은 태스크도 다양한 지시문으로 학습시킴
- 데이터셋 크기 차이를 줄이기 위해 balanced sampling 적용함
- 모델은 FlanT5, Vicuna 등 다양한 LLM과 결합 가능하게 설계됨
- 이렇게 모델 구조와 학습 전략을 결합해 unseen 태스크에서도 instruction을 따라 제로샷 성능을 낼 수 있음

4. Experiments

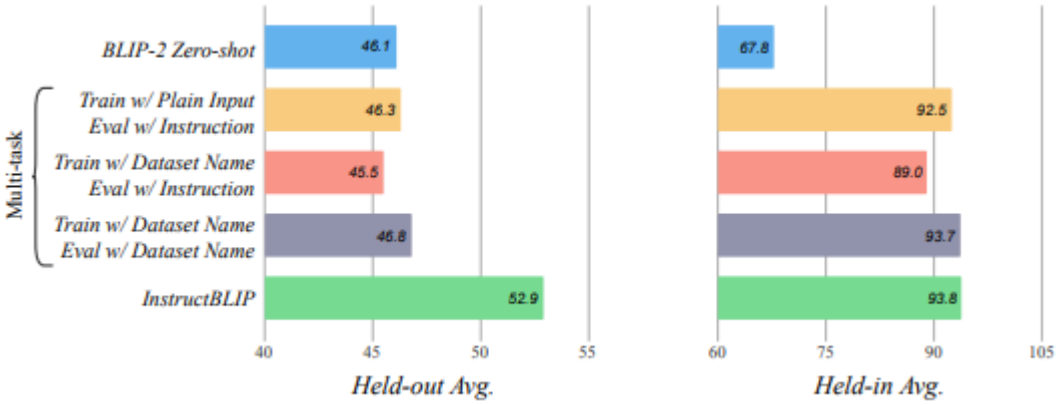
- 학습된 InstructBLIP 모델을 13개 held-out 데이터셋에서 제로샷 평가함
- 데이터셋 종류는 캡션, VQA, OCR, 추론 등 다양한 시각-언어 태스크 포함
- Ablation study 진행. instruction-aware Q-Former 제거, instruction 템플릿 단순화, balanced sampling 미적용 시 성능 변화 관찰
- 다양한 LLM(FlanT5, Vicuna)과 조합하여 실험. 모델 구조와 학습 전략이 다른 LLM에서도 안정적으로 작동하는지 확인

- Fine-tuning downstream 태스크도 진행. ScienceQA-IMG 등 이미지 포함 QA에서 성능 측정
- 실험 결과, instruction 포맷 데이터셋과 Q-Former 학습이 unseen 태스크 일반화와 fine-tuning 성능에 기여함

5. Results

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5$_{XL}$) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5$_{XXL}$) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5$_{XL}$)	119.9	**84.5**	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5$_{XXL}$)	120.0	83.5	47.9	**65.6**	**51.2**	46.6	**48.5**	54.1	30.9	**70.6**	**44.3**	**25.6**	**53.8**
InstructBLIP (Vicuna-7B)	**123.1**	82.4	49.2	54.3	43.1	50.1	45.2	**59.6**	**34.5**	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	**49.5**	52.1	44.8	**50.7**	45.4	57.5	33.4	63.1	41.2	24.8	51.0

Model	Held-in Avg.	GQA	ScienceQA (image-context)	IconQA	VizWiz	iVQA
InstructBLIP (FlanT5$_{XL}$)	94.1	48.4	70.4	50.0	32.7	53.1
w/o Instruction-aware Visual Features	89.8	45.9 (↓2.5)	63.4 (↓7.0)	45.8 (↓4.2)	25.1 (↓7.6)	47.5 (↓5.6)
w/o Data Balancing	92.6	46.8 (↓1.6)	66.0 (↓4.4)	49.9 (↓0.1)	31.8 (↓0.9)	51.1 (↓2.0)
InstructBLIP (Vicuna-7B)	100.8	49.2	60.5	43.1	34.5	52.2
w/o Instruction-aware Visual Features	98.9	48.2 (↓1.0)	55.2 (↓5.3)	41.2 (↓1.9)	32.4 (↓2.1)	36.8 (↓15.4)
w/o Data Balancing	98.8	47.8 (↓1.4)	59.4 (↓1.1)	43.5 (↑0.4)	32.3 (↓2.2)	50.3 (↓1.9)



	ScienceQA IMG	OCR-VQA	OKVQA	A-OKVQA Direct Answer Val	A-OKVQA Direct Answer Test	A-OKVQA Multi-choice Val	A-OKVQA Multi-choice Test
Previous SOTA	LLaVA [25] 89.0	GIT [43] 70.3	PaLM-E(562B) [9] **66.1**	[15] 56.3	[37] 61.6	[15] 73.2	[37] 73.6
BLIP-2 (FlanT5$_{XXL}$)	89.5	72.7	54.7	57.6	53.7	80.2	76.2
InstructBLIP (FlanT5$_{XXL}$)	**90.7**	**73.3**	55.5	57.1	54.8	**81.0**	**76.7**
BLIP-2 (Vicuna-7B)	77.3	69.1	59.3	60.0	58.7	72.1	69.0
InstructBLIP (Vicuna-7B)	79.5	72.8	62.1	**64.0**	**62.1**	75.7	73.4

- 제로샷 평가에서 held-out 데이터셋 전체에서 우수한 성능 보임
- 다양한 instruction 템플릿 덕분에 unseen 태스크에서도 안정적인 결과 나타남
- 학습에 포함되지 않은 태스크에서도 BLIP-2 대비 평균 성능 약 15% 이상 개선
- Fine-tuning downstream 태스크에서도 초기화로 활용 시 학습 안정성과 성능 향상 확인
- 예시로 ScienceQA-IMG에서는 90.7% 정확도 달성
- Ablation study에서 instruction-aware Q-Former 제거 시 성능 하락, instruction 템플릿과 balanced sampling 중요성 확인
- 전반적으로 instruction tuning과 Q-Former 설계가 결과에 큰 영향을 주는 것으로 확인됨

6. Insight

- instruction tuning과 instruction-aware Q-Former는 unseen 태스크에서 제로샷 성능을 크게 높임
- 이미지 인코더와 LLM을 동결하고 Q-Former만 학습해 효율성을 확보함
- 다양한 instruction 템플릿과 balanced sampling 덕분에 데이터셋 편향 문제를 완화함
- 다양한 시각-언어 태스크를 하나의 모델로 통합할 수 있는 가능성을 보여줌
- 발전할 점은 더 복잡하고 현실적인 멀티모달 태스크로 확장하는 것, 예를 들어 영상, 3D, 시뮬레이션 데이터까지 포함하는 범용 모델 개발
- instruction 포맷 생성 자동화와 LLM과의 더 긴밀한 통합도 향후 성능 개선 여지가 있음
- 실용적 측면에서는 모델 경량화, 실시간 응용 가능성 확보, 다양한 LLM과의 호환성 강화가 발전 방향임