Visual Instruction Tuning

https://arxiv.org/pdf/2304.08485

0. Introduction

- 비전과 언어를 통합해 모델이 시각적 정보를 이해하고 자연어로 소통하는 능력이 중요해짐
- 기존 멀티모달 모델들은 특정 태스크에 한정되거나, 복잡한 파인튜닝 절차가 필요함
- 본 논문은 다양한 시각-언어 명령을 학습하는 Visual Instruction Tuning (VIT) 프레임워크를 제안
- VIT는 광범위한 시각-언어 태스크를 단일 모델에 통합하고, 사용자 친화적인 인터페이스 제공
- 공개된 대규모 시각-언어 데이터셋을 활용해 범용성과 확장성을 강화
- 실험 결과, 여러 벤치마크에서 기존 모델 대비 뛰어난 성능과 유연성을 입증함

1. Overview

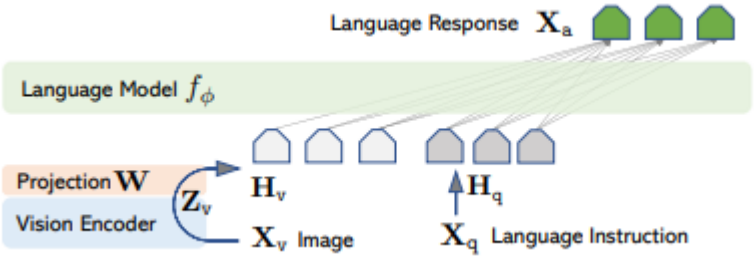
- Visual Instruction Tuning(VIT)은 다양한 시각-언어 명령을 하나의 모델로 통합하는 접근법임
- 모델은 이미지와 자연어 명령을 함께 입력받아, 다양한 시각적 질문과 작업에 대응 가능
- 대규모 시각-언어 데이터셋을 사용해 범용적인 능력을 학습함
- 파인튜닝 과정이 간단하며, 추가 태스크에 빠르게 적응할 수 있도록 설계됨
- VIT는 기존 멀티모달 모델 대비 유연성과 성능 면에서 우수함을 보임

2. Challenges

- 시각-언어 통합 모델은 다양한 태스크와 명령을 모두 효과적으로 처리하기 어려움
- 기존 모델들은 특정 작업에 최적화되어 범용성이 부족함
- 대규모 멀티태스크 학습 시 데이터 간 불균형과 잡음 문제 존재
- 시각 정보와 언어 정보 간 복잡한 상호작용을 효율적으로 학습하는 것이 어려움

- 빠른 적응성과 사용자 친화적 인터페이스 제공이 요구됨

3. Method



- VIT는 이미지와 자연어 명령을 함께 입력으로 받는 단일 모델 구조를 사용함
- 다양한 시각-언어 태스크 데이터를 통합해 멀티태스크 학습 수행
- 모델은 Transformer 기반으로 설계되어 시각 및 언어 정보를 효과적으로 융합함
- 사용자 명령을 반영해 다양한 시각적 질문과 작업에 대응할 수 있도록 튜닝함
- 파인튜닝 과정은 단순하며, 새로운 태스크 추가 시 빠른 적응 가능

4. Experiments

- 다양한 시각-언어 벤치마크(예: VQA, COCO Caption, Visual Reasoning 등)에서 평가함
- 대규모 시각-언어 데이터셋을 활용해 멀티태스크 학습 진행
- 기존 멀티모달 모델들과 성능 비교를 통해 우수성 검증
- 다양한 명령 유형과 태스크에 대해 모델의 적응력과 범용성 평가
- 평가 지표로 정확도, CIDEr 등 다양한 메트릭 사용

5. Results

	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)

	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

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**

- 다양한 시각-언어 태스크에서 기존 모델 대비 우수한 성능 달성
- VQA, 이미지 캡셔닝, 시각 추론 등에서 정확도와 언어 생성 품질 향상
- 멀티태스크 학습을 통해 모델의 범용성과 적응력 강화됨
- 사용자 명령 변화에 유연하게 대응 가능함을 실험으로 확인
- 파인튜닝 비용이 낮아 효율적인 실시간 적용 가능

6. Insight

- 시각-언어 명령을 통합하는 멀티태스크 학습이 모델의 범용성과 적응력을 크게 향상시킴
- Transformer 기반 구조를 활용해 시각과 언어 정보를 효과적으로 융합함
- 사용자 친화적인 인터페이스를 제공해 다양한 실제 응용에 적합함
- 파인튜닝 비용이 낮아 실시간 환경에 적용 가능함
- 다만, 멀티태스크 학습 중 데이터 불균형과 잡음 문제는 해결 과제로 남아 있음

- 특정 복잡한 태스크나 드문 명령에 대해서는 추가 개선이 필요함