BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions

https://arxiv.org/abs/2308.09936

0. Introduction

- 기존 Vision-Language Models(VLMs)은 이미지 이해에는 강하지만, 텍스트가 풍부한 이미지 처리에는 성능이 크게 제한됨.
- 이미지 내 텍스트 정보를 담기 위해 사용되는 query embeddings는 토큰 수 제약 때문에 충분한 정보 전달이 어려움.
- 이러한 한계로 인해 OCR-VQA, TextVQA 등 텍스트 중심 과제에서 기존 모델들은 질문에 대한 정확한 답변 생성에 어려움을 겪음.
- BLIVA는 간단한 구조적 개선을 통해 이러한 제약을 극복하고, 텍스트 풍부 이미지에서의 VQA 성능을 크게 향상시키는 것을 목표로 함.

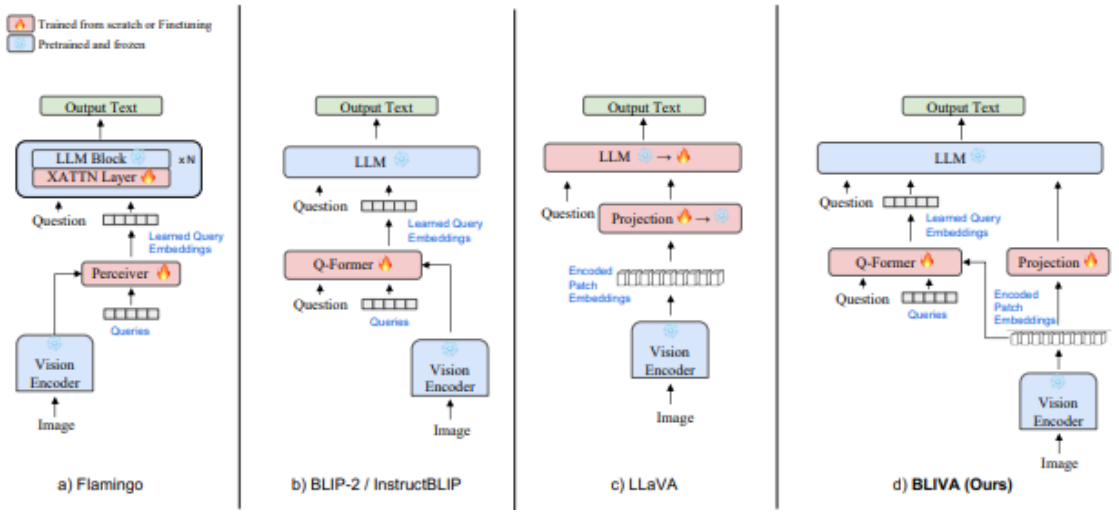
1. Overview

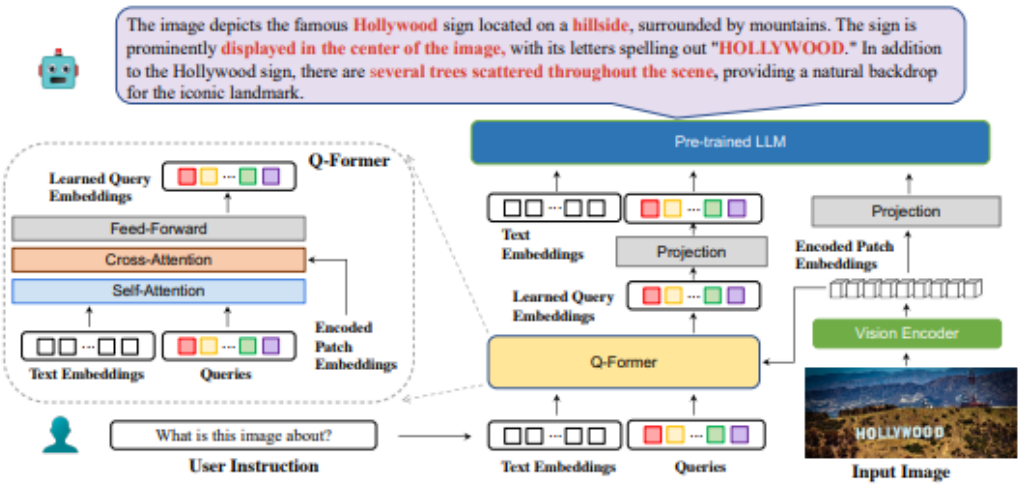
- BLIVA는 InstructBLIP을 기반으로 하며, 이미지 정보를 단순 query embedding에만 의존하지 않고, 인코딩된 patch embeddings를 LLM에 직접 전달하는 구조를 채택.
- 이 방식은 LLaVA의 접근법을 참고하여, 이미지 내 텍스트나 세부 요소가 손실되지 않도록 보완.
- 모델 학습은 두 단계로 진행되며, 사전 학습(pretraining) 후 instruction tuning을 통해 멀티모달 QA 성능을 강화.
- 설계 목표는 단순성과 효율성을 유지하면서도 텍스트 중심 VQA 과제에서 기존 VLM 대비 큰 성능 향상을 달성하는 것임.

2. Challenges

- 기존 VLM은 query embeddings 수가 제한적이라 이미지 내 텍스트 정보를 충분히 전달하지 못함.
- 이미지 속 텍스트는 위치·맥락에 따라 의미가 달라지지만, 기존 모델은 이를 세밀하게 반영하기 어려움.
- LLM에 시각적 정보를 연결하는 과정에서 중요한 텍스트 단서가 손실되거나 희석되는 문제가 발생.
- 텍스트가 많은 벤치마크(OCR-VQA, TextVQA 등)에서 기존 접근법은 질문에 필요한 핵심 정보를 놓치기 쉬움.

3. Method





- BLIVA는 이미지 처리 과정에서 patch embeddings를 LLM에 직접 전달해, query embeddings만 사용하는 기존 방식의 정보 손실 문제를 해결.
- 모델 구조는 InstructBLIP 기반으로, 이미지 특징 추출과 텍스트 인코딩을 병렬로 수행할 수 있도록 설계됨.
- 학습은 2단계로 이루어짐
  1. Pretraining – 대규모 멀티모달 데이터로 기본 시각-언어 이해 능력 학습
  2. Instruction tuning – 다양한 VQA 과제에 맞춰 LLM이 구체적 질문에 답하도록 조정
- 이 구조 덕분에 BLIVA는 텍스트가 풍부한 이미지에서도 질문과 답변 간 의미 전달을 효율적으로 수행.

4. Experiments

- BLIVA는 텍스트 중심 VQA 데이터셋인 STVQA, OCR-VQA, TextVQA를 활용하여 이미지 속 텍스트 기반 질문에 답하도록 평가함.
- 일반 VQA 및 멀티모달 데이터셋으로는 VSR, IconQA, Visual Dialog, Flickr30K, Hateful Memes, MSRVTT를 사용하여 시각적 추론과 이미지-텍스트 이해 능력을 검증함.
- 실험은 pretraining 후 instruction tuning 단계로 수행되었으며, 모델 입력으로 query embeddings와 patch embeddings를 모두 사용함.
- 성능 평가는 VQA 정확도와 멀티모달 LLM 벤치마크(MME)를 기준으로 진행함.

5. Results

	STVQA ↑	OCRVQA ↑	TextVQA ↑	DocVQA ↑	InfoVQA ↑	ChartQA ↑	ESTVQA ↑	FUNSD ↑	SROIE ↑	POIE ↑	Average ↑
OpenFlamingo (Awadalla et al. 2023)	19.32	27.82	29.08	5.05	14.99	9.12	28.20	0.85	0.12	2.12	13.67
BLIP2-OPT$_{6.7b}$ (Li et al. 2023b)	13.36	10.58	21.18	0.82	8.82	7.44	27.02	0.00	0.00	0.02	8.92
BLIP2-FLanT5$_{XXL}$ (Li et al. 2023b)	21.38	30.28	30.62	4.00	10.17	7.20	42.46	1.19	0.20	2.52	15.00
MiniGPT4 (Zhu et al. 2023)	14.02	11.52	18.72	2.97	13.32	4.32	28.36	1.19	0.04	1.31	9.58
LLaVA (Liu et al. 2023a)	22.93	15.02	28.30	4.40	13.78	7.28	33.48	1.02	0.12	2.09	12.84
mPLUG-Owl (Ye et al. 2023)	26.32	35.00	37.44	6.17	**16.46**	**9.52**	**49.68**	1.02	0.64	**3.26**	18.56
InstructBLIP (FlanT5$_{XXL}$) (Dai et al. 2023)	26.22	55.04	36.86	4.94	10.14	8.16	43.84	1.36	0.50	1.91	18.90
InstructBLIP (Vicuna-7B) (Dai et al. 2023)	28.64	47.62	39.60	5.89	13.10	5.52	47.66	0.85	0.64	2.66	19.22
BLIVA (FlanT5$_{XXL}$)	28.24	61.34	39.36	5.22	10.82	9.28	45.66	**1.53**	0.50	2.39	20.43
BLIVA (Vicuna-7B)	**29.08**	**65.38**	**42.18**	**6.24**	13.50	8.16	48.14	1.02	**0.88**	2.91	**21.75**

Models	VSR ↑	IconQA ↑	TextVQA ↑	Visdial ↑	Flickr30K ↑	HM ↑ (val)	VizWiz ↑ (val-dev)	MSRVTT ↑ (val-dev)
Flamingo-3B (Alayrac et al. 2022)	-	-	30.1	-	60.6	-	-	-
Flamingo-9B (Alayrac et al. 2022)	-	-	31.8	-	61.5	-	-	-
Flamingo-80B (Alayrac et al. 2022)	-	-	35.0	-	67.2	-	-	-
MiniGPT-4 (Zhu et al. 2023)	50.65	-	18.56	-	-	29.0	34.78	-
LLaVA (Liu et al. 2023a)	56.3	-	37.98	-	-	9.2	36.74	-
BLIP-2 (Vicuna-7B) (Dai et al. 2023)	50.0	39.7	40.1	44.9	74.9	50.2	**49.34**	4.17
InstructBLIP (Vicuna-7B) (Dai et al. 2023)	54.3	43.1	50.1	45.2	82.4	54.8	43.3	18.7
InstructBLIP Baseline (Vicuna-7B)	58.67	44.34	37.58	40.58	84.61	50.6	44.10	20.97
BLIVA (Vicuna-7B)	**62.2**	**44.88**	**57.96**	**45.63**	**87.1**	**55.6**	42.9	**23.81**

Models	VSR ↑	IconQA ↑	TextVQA ↑	Visdial ↑	Flickr30K ↑	HM ↑ (val)	VizWiz ↑ (val-dev)	MSRVTT ↑ (val-dev)
BLIP-2 (FlanT5$_{XXL}$) (Li et al. 2023b)	68.2	45.4	44.1	46.9	73.7	52.0	29.4	17.4
InstructBLIP (FlanT5$_{XXL}$) (Dai et al. 2023)	65.6	51.2	46.6	**48.5**	83.5	**53.6**	41.35	20.79
BLIVA (FlanT5$_{XXL}$)	**68.82**	**52.42**	**57.2**	36.18	**87.66**	50.0	**43.97**	**23.78**

Model	Overall ↑	Perception ↑										Cognition ↑				Avg. ↑
		Exist.	Count	Pos.	Color	OCR	Poster	Cele.	Scene	Land.	Art.	Comm.	NumCal.	Trans.	Code	
LLaVA(Liu et al. 2023a)	712.5	50.0	50.0	50.0	50.0	50.0	50.0	48.8	50.0	50.0	49.0	57.1	50.0	57.5	50.0	50.9
MiniGPT-4(Zhu et al. 2023)	694.3	68.3	55.0	43.3	43.3	57.5	41.8	54.4	71.8	54.0	60.5	59.3	45.0	0.0	40.0	49.6
mPLUG-Owl(Ye et al. 2023)	1238.4	120.0	50.0	50.0	50.0	65.0	136.1	100.3	135.5	159.3	96.3	78.6	60.0	80.0	57.5	88.5
InstructBLIP(Dai et al. 2023)	1417.9	185.0	143.3	66.7	66.7	72.5	123.8	101.2	153.0	79.8	134.3	129.3	40.0	65.0	57.5	101.3
BLIP-2(Li et al. 2023b)	1508.8	160.0	135.0	73.3	73.3	110.0	141.8	105.6	145.3	138.0	136.5	110.0	40.0	65.0	75.0	107.8
BLIVA	**1669.2**	180.0	138.3	81.7	180.0	87.5	155.1	140.9	151.5	89.5	133.3	136.4	57.5	77.5	60.0	**119.2**

InstructBLIP (Dai et al. 2023)	Baseline (Instruction Tuning Qformer)	Patch Embedding	Pre-Training	Finetuning LLM	ST-VQA	OCR-VQA	Text-VQA	Doc-VQA	Info-VQA	Chart-QA	EST-VQA	FUNSD	SROIE	POIE	Improvement
✓					28.64	47.62	39.60	5.89	13.10	5.52	47.66	0.85	0.64	2.66	+ 0 %
✓	✓				**30.08**	65.8	40.5	6.13	12.03	8.08	47.02	0.85	0.57	2.62	+ 7.40%
✓	✓	✓			28.86	65.04	40.7	**6.65**	**14.28**	**8.24**	47.72	**1.19**	**1.66**	2.83	+ **31.72%**
✓	✓	✓	✓		29.08	65.38	**42.18**	6.24	13.50	8.16	**48.14**	1.02	0.88	**2.91**	+ 17.01%
✓	✓	✓	✓	✓	29.94	**66.48**	41.9	6.47	12.51	7.52	46.76	1.02	0.51	2.85	+ 9.65%

- BLIVA는 텍스트 중심 VQA에서 OCR-VQA와 TextVQA를 포함한 과제에서 기존 InstructBLIP 대비 최대 +17.76%p 성능 향상을 달성함.
- 일반 VQA 과제에서는 VSR과 IconQA 등에서 기존 모델 대비 최대 +7.9%p 향상을 기록함.
- 멀티모달 LLM 벤치마크(MME) 평가에서는 Perception과 Cognition 분야 평균 성능이 +17.72%p 상승하였고, 특히 Color, Poster, Commonsense Reasoning 항목에서 최고 점수를 달성함.
- Ablation Study 결과, patch embeddings 도입만으로도 대부분 VQA 과제에서 성능이 향상되었으며, 2단계 학습(pretraining + instruction tuning)이 전반적인 성능 개선에 핵심적인 역할을 함.
- 반면, LLM 동시 LoRA 파인튜닝은 일부 과제에서 미미한 개선 효과만을 보여, 모든 과제에서 효율적이지는 않았음.

6. Insight

- BLIVA는 단순 구조 개선만으로 텍스트가 풍부한 이미지에서 기존 VLM 대비 성능을 크게 향상시킴.
- Patch embeddings를 직접 LLM에 전달하는 접근은 query embeddings의 정보 제한 문제를 효과적으로 해결함.
- 2단계 학습(pretraining + instruction tuning)이 멀티모달 이해 능력 향상에 결정적 역할을 함.
- OCR 기반 QA, 이미지-텍스트 분석, 시각적 질문 응답 등 다양한 실제 응용에서 활용 가능함.
- 계산량과 메모리 부담이 증가할 수 있으며, 일부 과제에서는 LLM 동시 LoRA 파인튜닝 효과가 미미함.
- 실험이 특정 VQA와 일부 멀티모달 과제에 한정되어 있어, 다른 환경에서의 일반화 가능성은 추가 검증 필요함.
- 텍스트와 이미지 복합적 맥락에는 강하지만, 동영상이나 시퀀스 기반 시각 정보 처리에는 제한적일 수 있음.