Osprey: Pixel Understanding with Visual Instruction Tuning

https://arxiv.org/pdf/2312.10032

0. Introduction

- 기존 시각 언어 모델(MLLM)은 이미지 수준이나 박스 수준에서의 이해에 집중함. 그러나 세밀한 영역 수준의 이해에는 한계가 있음.
- 현재의 모델들은 마스크 기반의 지시문 데이터를 활용하지 않아 발전이 제한적임.
- Osprey는 마스크-텍스트 지시 튜닝 접근 방식을 제안하여, 마스크 영역을 언어 지시문에 통합함으로써 픽셀 수준의 시각 이해를 목표로 함.
- 이를 위해 724K 샘플로 구성된 마스크 기반 영역-텍스트 데이터셋인 Osprey-724K를 구축함.
- Osprey는 고해상도 입력을 처리할 수 있는 컨볼루션 기반 CLIP 백본을 사용하며, 마스크 인식 기능을 갖춘 시각 추출기를 채택함.
- 이러한 설계를 통해 Osprey는 객체 수준 및 부분 수준의 세밀한 의미 이해를 달성함.
- 또한, Segment Anything Model(SAM)과의 통합을 통해 다중 해상도의 의미를 추출할 수 있음.
- Osprey는 개체 분류, 개방형 어휘 인식, 지역 수준 캡셔닝 및 세부 지역 설명 작업에서 뛰어난 성능을 보임

1. Overview

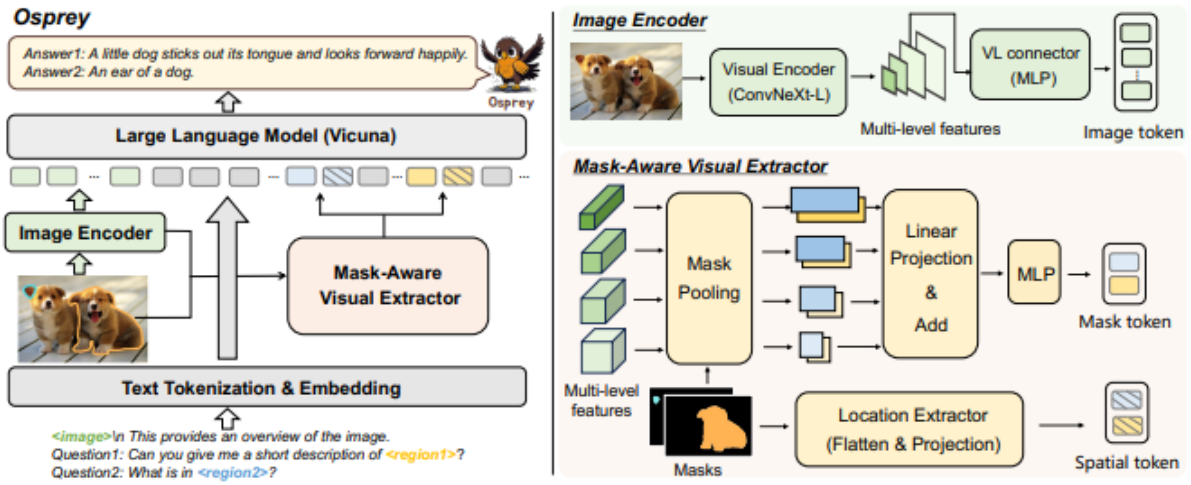
- Osprey는 시각 언어 모델(MLLM)의 한계를 극복하기 위해 마스크-텍스트 지시 튜닝 방식을 도입함.
- 기존 MLLM은 이미지 수준이나 박스 수준의 이해에 집중하여 픽셀 수준의 세밀한 이해에는 한계가 있었음.
- Osprey는 724K 샘플로 구성된 Osprey-724K 데이터셋을 활용하여 마스크 기반의 영역-텍스트 지시를 학습함.

- 이 모델은 고해상도 입력을 처리할 수 있는 컨볼루션 기반 CLIP 비전 인코더와 마스크 인식 기능을 갖춘 시각 추출기를 채택함.
- Osprey는 Segment Anything Model(SAM)과 통합되어 다중 해상도의 의미를 추출할 수 있음.
- 이러한 설계를 통해 Osprey는 객체 수준 및 부분 수준의 세밀한 의미 이해를 달성함.

2. Challenges

- Osprey는 시각 언어 모델(MLLM)의 한계를 극복하기 위해 마스크-텍스트 지시 튜닝 방식을 도입함.
- 기존 MLLM은 이미지 수준이나 박스 수준의 이해에 집중하여 픽셀 수준의 세밀한 이해에는 한계가 있었음.
- Osprey는 724K 샘플로 구성된 Osprey-724K 데이터셋을 활용하여 마스크 기반의 영역-텍스트 지시를 학습함.
- 이 모델은 고해상도 입력을 처리할 수 있는 컨볼루션 기반 CLIP 비전 인코더와 마스크 인식 기능을 갖춘 시각 추출기를 채택함.
- Osprey는 Segment Anything Model(SAM)과 통합되어 다중 해상도의 의미를 추출할 수 있음.
- 이러한 설계를 통해 Osprey는 객체 수준 및 부분 수준의 세밀한 의미 이해를 달성함.

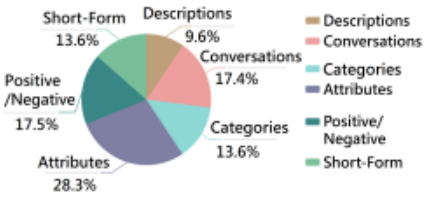
3. Method



- Osprey는 시각-언어 모델에 픽셀 수준의 이해를 추가하기 위해 마스크-텍스트 지시 튜닝 방식을 도입함
- 724K 샘플로 구성된 마스크 기반 영역-텍스트 데이터셋인 Osprey-724K를 구축하여 학습에 활용함
- 고해상도 입력을 처리할 수 있는 컨볼루션 기반 CLIP 비전 인코더를 사용함
- 마스크 인식 기능을 갖춘 시각 추출기를 채택하여 정확한 시각 마스크 피처를 추출함
- *Segment Anything Model(SAM)**과 통합하여 다중 해상도의 의미를 추출할 수 있음
- 언어 지시문과 시각 마스크 피처를 결합하여 모델 입력 시퀀스를 생성함
- 이러한 설계를 통해 객체 수준 및 부분 수준의 세밀한 의미 이해를 달성함

4. Experiments

Type	Form	Raw Data	GPT-4	#Samples
Object-level	Descriptions	COCO/RefCOCO/RefCOCO+/	✓	70K
	Conversations	RefCOCOg/LLaVA-115K	✓	127K
Part-level	Categories	PACO-LVIS	✓	99K
	Attributes		✓	207K
Robustness &Flexibility	Positive/Negative	COCO/RefCOCO/RefCOCO+/	✗	64K/64K
	Short-Form	RefCOCOg/LLaVA-115K/LVIS	✓	99k



- Osprey를 여러 시각-언어 태스크에서 평가함
- 데이터셋은 개체 분류, 지역 수준 캡셔닝, 개방형 어휘 인식, 세부 지역 설명 등 포함
- Ablation study 진행. 마스크-텍스트 지시 제거, SAM 통합 제거, 데이터셋 샘플링 변화 시 성능 확인
- 다양한 해상도와 마스크 피처 활용이 성능에 미치는 영향 분석
- 결과, 마스크 기반 영역-텍스트 지시 튜닝과 SAM 통합이 세밀한 픽셀 이해 성능 향상에 크게 기여함
- 실험을 통해 Osprey의 픽셀 수준 의미 이해 능력과 일반화 가능성 확인

5. Results

Method	Type	Cityscapes			ADE20K-150		
		PQ	AP	mIoU	PQ	AP	mIoU
CLIP-ConvNeXt-L [43]	Mask	22.53	12.07	23.06	36.86	39.38	28.74
CLIP-Surgery-ViT-L [30]	Mask	27.24	28.35	21.92	26.55	29.70	21.42
Kosmos-2 [40]	Box	12.09	9.81	13.71	6.53	4.33	5.40
Shikra-7B [5]	Box	17.80	11.53	17.77	27.52	20.35	18.24
GPT4RoI-7B [58]	Box	34.70	21.93	36.73	36.32	26.08	25.82
Ferret-7B [54]	Mask	35.57	26.94	38.40	39.46	29.93	**31.77**
Osprey-7B (Ours)	Mask	**50.64**	**29.17**	**49.78**	**41.89**	**41.24**	29.63

Method	LVIS		PACO	
	SS	S-IoU	SS	S-IoU
LLaVA-1.5 [32]	48.95	19.81	42.20	14.56
Kosmos-2 [40]	38.95	8.67	32.09	4.79
Shikra-7B [5]	49.65	19.82	43.64	11.42
GPT4RoI-7B [58]	51.32	11.99	48.04	12.08
Ferret-7B [54]	63.78	36.57	58.68	25.96
Osprey-7B (Ours)	**65.24**	**38.19**	**73.06**	**52.72**

Method	Detailed Description
LLaVA-1.5 [32]	71.11
Kosmos-2 [40]	40.89
Shikra-7B [5]	40.97
GPT4RoI-7B [58]	49.97
Osprey-7B (Ours)	77.54
Osprey-7B* (Ours)	**83.78**

Sampling	Metrics	Osprey-7B*	Ferret-7B	Shikra-7B	LLaVA-1.5	InstructBLIP	MiniGPT4	MM-GPT	mPLUG-Owl
Random	Accuracy	89.47	**90.24**	86.90	88.73	88.57	79.67	50.10	53.97
	Precision	93.40	97.72	94.40	88.89	84.09	78.24	50.05	52.07
	Recall	84.93	83.00	79.26	88.53	95.13	82.20	100.00	99.60
	F1 Score	88.97	89.76	86.19	88.71	89.27	80.17	66.71	68.39
	Yes (%)	45.47	43.78	43.26	49.80	56.57	52.53	99.90	95.63
Popular	Accuracy	**87.83**	84.90	83.97	85.83	82.77	69.73	50.00	50.90
	Precision	89.94	88.24	87.55	83.91	76.27	65.86	50.00	50.46
	Recall	85.20	80.53	79.20	88.67	95.13	81.93	100.00	99.40
	F1 Score	87.50	84.21	83.16	86.22	84.66	73.02	66.67	66.94
	Yes (%)	47.37	45.63	45.23	52.83	62.37	62.20	100.00	98.57
Adversarial	Accuracy	**85.33**	82.36	83.10	72.10	65.17	79.20	50.00	50.67
	Precision	85.43	83.60	85.60	74.69	65.13	61.19	50.00	50.34
	Recall	85.20	80.53	79.60	88.34	95.13	82.93	100.00	99.33
	F1 Score	85.31	82.00	82.49	80.94	77.32	70.42	66.67	66.82
	Yes (%)	49.87	48.18	46.50	59.14	73.03	67.77	100.00	98.67

Method	Type	METEOR	CIDEr
GRIT [50]	Box	15.2	71.6
Kosmos-2 [40]	Box	14.1	62.3
GLaMM [45]	Box	16.2	105.0
Osprey-7B (Ours)	Mask	**16.6**	**108.3**

Input	#Image Tokens	Speed	SS	S-IoU
224	196	**6.0**	53.20	26.12
336	441	5.8	56.70	28.90
512	1024	3.5	65.24	38.19
800	2500	1.9	**68.29**	**42.66**

- Osprey는 개체 분류, 지역 수준 캡셔닝, 개방형 어휘 인식, 세부 지역 설명 등 다양한 태스크에서 우수한 성능을 보임
- 마스크 기반 영역-텍스트 지시 튜닝과 SAM 통합이 성능 향상에 크게 기여함
- Ablation study에서 마스크 지시 제거 시 정확도와 세밀도 모두 하락, 마스크 기반 학습 중요성 확인
- 다양한 해상도 입력 처리와 마스크 피처 활용으로 픽셀 수준 이해 능력이 향상됨

- 결과적으로 Osprey는 기존 시각-언어 모델 대비 세밀한 영역 이해와 일반화 성능이 개선됨

6. Insight

- Osprey는 마스크-텍스트 지시 튜닝과 SAM 통합 덕분에 픽셀 수준 세밀한 의미 이해 능력이 뛰어남
- 다양한 해상도와 영역 단위를 동시에 처리할 수 있는 구조가 일반화 성능 향상에 기여함
- 마스크 기반 데이터셋 구축과 instruction 포맷 학습이 unseen 태스크에서도 성능 안정성을 높임
- 발전 가능성으로는 영상, 3D, 다중 모달 환경 등으로 확장, 더 다양한 픽셀 수준 태스크 적용 가능
- 학습 효율화, 모델 경량화, 실시간 응용 가능성 확보, 다양한 LLM과 결합해 범용 모델 개발이 향후 과제임