Flamingo: a Visual Language Model for Few-Shot Learning

https://arxiv.org/abs/2204.14198

0. Introduction

- 비전과 언어를 결합한 대규모 멀티모달 모델 연구함
- Flamingo는 few-shot 학습에 강한 비주얼 언어 모델임
- 이미지와 텍스트를 동시에 처리할 수 있는 새로운 구조 제안함
- 사전학습된 비전 백본과 언어 모델을 결합해 효과적 정보 전달함
- 적은 수의 예시만으로도 다양한 태스크에 빠르게 적응 가능함
- 자연어와 이미지 기반 태스크에서 높은 성능과 범용성 보임

1. Overview

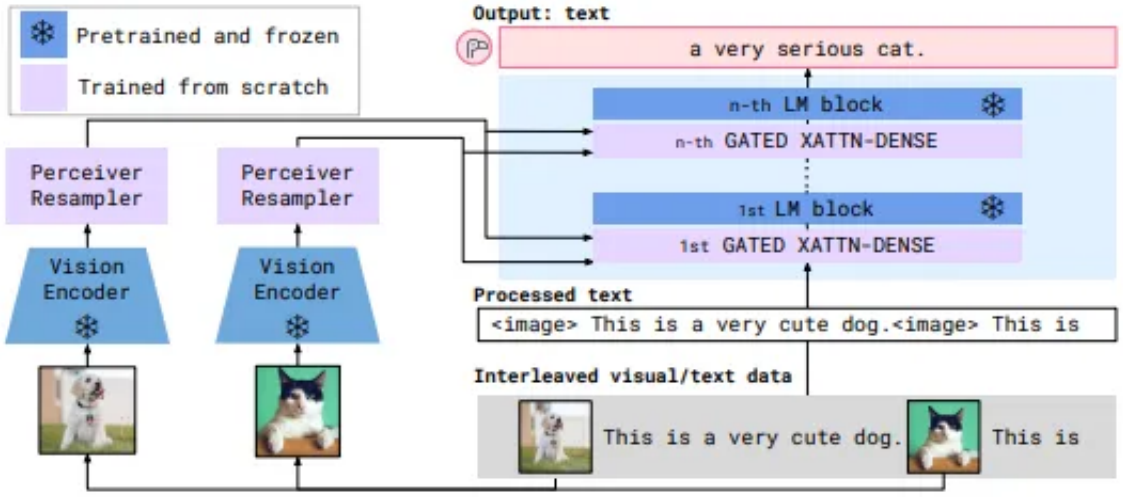
- Flamingo는 비전 인코더와 언어 모델을 결합한 구조임
- 이미지와 텍스트 입력을 순차적으로 처리하는 멀티모달 트랜스포머 기반임
- 비전 백본은 이미지 특징 추출에 사용됨
- 언어 모델은 텍스트 생성과 이해에 특화됨
- 두 모듈 간에 cross-attention으로 정보를 주고받음
- 적은 수의 예시(few-shot)로도 빠르게 새로운 태스크 학습 가능
- 다양한 비전-언어 태스크에 범용적으로 적용됨

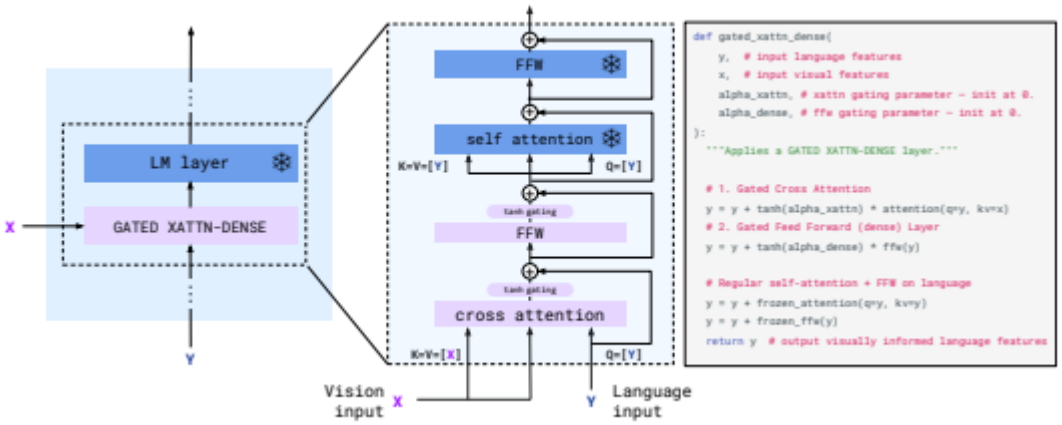
2. Challenges

- 비전과 언어 간 효과적 정보 융합이 어려움
- 적은 데이터로 빠르게 학습하는 few-shot 학습 자체가 난제임
- 대규모 사전학습 모델을 멀티모달 태스크에 효율적으로 적용하는 문제 존재

- 이미지와 텍스트 길이 및 구조 차이 조정 필요함
- 모델이 새로운 태스크에 적응할 때 과적합 위험 있음
- 계산 비용과 메모리 요구가 매우 높음

3. Method





- Flamingo는 이미지 인코더와 대형 언어 모델을 결합함
- 이미지 인코더는 비전 백본으로 이미지 특징 추출함
- 언어 모델은 텍스트 생성 및 이해 담당함
- 두 모듈 사이에 cross-attention 레이어를 삽입해 정보 교환함
- cross-attention이 이미지 정보를 언어 모델에 주입하는 역할 수행함

- 적은 수의 예시를 받아 태스크 적응을 위한 few-shot 학습 가능하게 설계함
- 사전학습된 언어 모델과 비전 백본을 효율적으로 결합함
- 모델 효율성과 확장성을 고려해 구조 최적화함

4. Experiments

- 다양한 비전-언어 태스크에서 Flamingo 성능 평가함
- 사용한 데이터셋은 COCO, VQAv2, WebQA 등임
- few-shot 환경에서 태스크 적응력 집중 측정함
- 기존 SOTA 모델과 성능 비교함
- 다양한 크기 모델과 학습 설정에 따른 ablation study 진행함
- 이미지와 텍스트 간 cross-attention 효과 분석함
- 인간 평가와 자동 평가 지표 모두 활용해 결과 검증함

5. Results

Method	VQAV2		COCO	VATEX	VizWiz		MSRVTTQA	VisDial		YouCook2	TextVQA		HatefulMemes
	test-dev	test-std	test	test	test-dev	test-std	test	valid	test-std	valid	valid	test-std	test seen
32 shots	67.6	-	113.8	65.1	49.8	-	31.0	56.8	-	86.8	36.0	-	70.0
Fine-tuned	**82.0**	**82.1**	138.1	**84.2**	**65.7**	**65.4**	**47.4**	61.8	59.7	118.6	**57.1**	54.1	**86.6**
SotA	81.3† [133]	81.3† [133]	**149.6**† [119]	81.4† [153]	57.2† [65]	60.6† [65]	46.8 [51]	**75.2** [79]	**75.4**† [123]	**138.7** [132]	54.7 [137]	**73.7** [84]	84.6† [152]

	Ablated setting	*Flamingo*-3B original value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	MSVDQA top1↑	VATEX CIDEr↑	Overall score↑
	***Flamingo*-3B model**			3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Training data	All data	w/o Video-Text pairs	3.2B	1.42s	84.2	43.0	53.9	34.5	46.0	67.3
			w/o Image-Text pairs	3.2B	0.95s	66.3	39.2	51.6	32.0	41.6	60.9
			Image-Text pairs→ LAION	3.2B	1.74s	79.5	41.4	53.5	33.9	47.6	66.4
			w/o M3W	3.2B	1.02s	54.1	36.5	52.7	31.4	23.5	53.4
(ii)	Optimisation	Accumulation	Round Robin	3.2B	1.68s	76.1	39.8	52.1	33.2	40.8	62.9
(iii)	Tanh gating	✓	✗	3.2B	1.74s	78.4	40.5	52.9	35.9	47.5	66.5
(iv)	Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN	2.4B	1.16s	80.6	41.5	53.4	32.9	50.7	66.9
			GRAFTING	3.3B	1.74s	79.2	36.1	50.8	32.2	47.8	63.1
(v)	Cross-attention frequency	Every	Single in middle	2.0B	0.87s	71.5	38.1	50.2	29.1	42.3	59.8
			Every 4th	2.3B	1.02s	82.3	42.7	55.1	34.6	50.8	68.8
			Every 2nd	2.6B	1.24s	83.7	41.0	55.8	34.5	49.7	68.2
(vi)	Resampler	Perceiver	MLP	3.2B	1.85s	78.6	42.2	54.7	35.2	44.7	66.6
			Transformer	3.2B	1.81s	83.2	41.7	55.6	31.5	48.3	66.7
(vii)	Vision encoder	NFNet-F6	CLIP ViT-L/14	3.1B	1.58s	76.5	41.6	53.4	33.2	44.5	64.9
			NFNet-F0	2.9B	1.45s	73.8	40.5	52.8	31.1	42.9	62.7
(viii)	Freezing LM	✓	✗ (random init)	3.2B	2.42s	74.8	31.5	45.6	26.9	50.1	57.8
			✗ (pretrained)	3.2B	2.42s	81.2	33.7	47.4	31.0	53.9	62.7

- Flamingo가 few-shot 학습 환경에서 기존 모델 대비 뛰어난 성능 보임

- 이미지 캡셔닝, VQA, 이미지-텍스트 매칭 등 태스크에서 우수함
- cross-attention 레이어가 정보 융합에 효과적임 확인됨
- ablation study로 구성 요소들의 기여도 명확히 검증됨
- 대규모 사전학습이 전반적 성능 향상에 크게 기여함
- 인간 평가에서도 자연스러운 생성 결과 보여줌

6. Insight

- 비전과 언어를 결합한 few-shot 학습 모델로서 Flamingo가 효과적임
- cross-attention을 통한 모달리티 간 정보 교환이 핵심임
- 사전학습과 멀티태스크 학습이 성능과 범용성에 큰 기여를 함
- 적은 데이터로도 다양한 태스크에 빠르게 적응 가능함
- 다만 계산 비용과 메모리 요구가 높아 실무 적용에 부담 될 수 있음
- 새로운 태스크 적응 시 과적합 위험 존재함