BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi

2026.07.20

MuliLAB Journal Club
인공지능응용학과 장승호

Overview

01 Author & Journal

Junnan Li

Salesforce AI Research
Verified email at salesforce.com

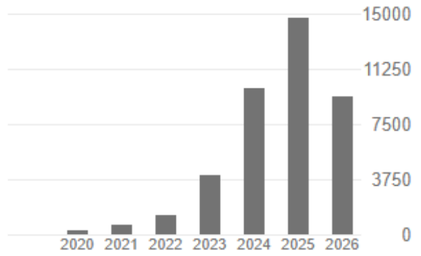
FOLLOW

GET MY OWN PROFILE

TITLE	CITED BY	YEAR
Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models J Li, D Li, S Savarese, S Hoi International conference on machine learning, 19730-19742	12863	2023
Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation J Li, D Li, C Xiong, S Hoi International conference on machine learning, 12888-12900	9811	2022
Align before fuse: Vision and language representation learning with momentum distillation J Li, R Selvaraju, A Gotmare, S Joty, C Xiong, SCH Hoi Advances in neural information processing systems 34, 9694-9705	3568	2021

Cited by

	All	Since 2021
Citations	40925	40477
h-index	45	42
i10-index	64	64



- Salesforce AI Research
- Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation
- Align before fuse: Vision and language representation learning with momentum distillation

1) Study Purpose

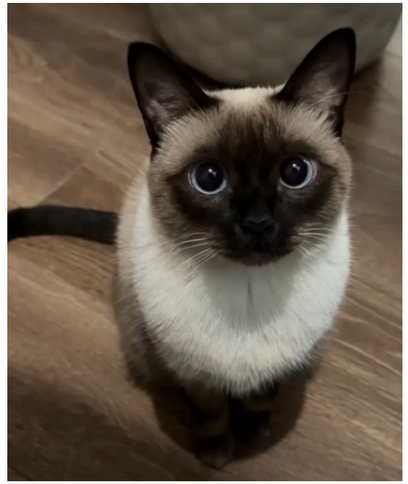
- Multimodal의 학습 메커니즘(ITC, ITM, ITG)에 대한 이해.
- image-text에서 Representaion Gap을 해결하는 원리에 대한 이해.
- Q-Former에 대한 이해.

03 Overview

멀티모달이 어려운 이유,

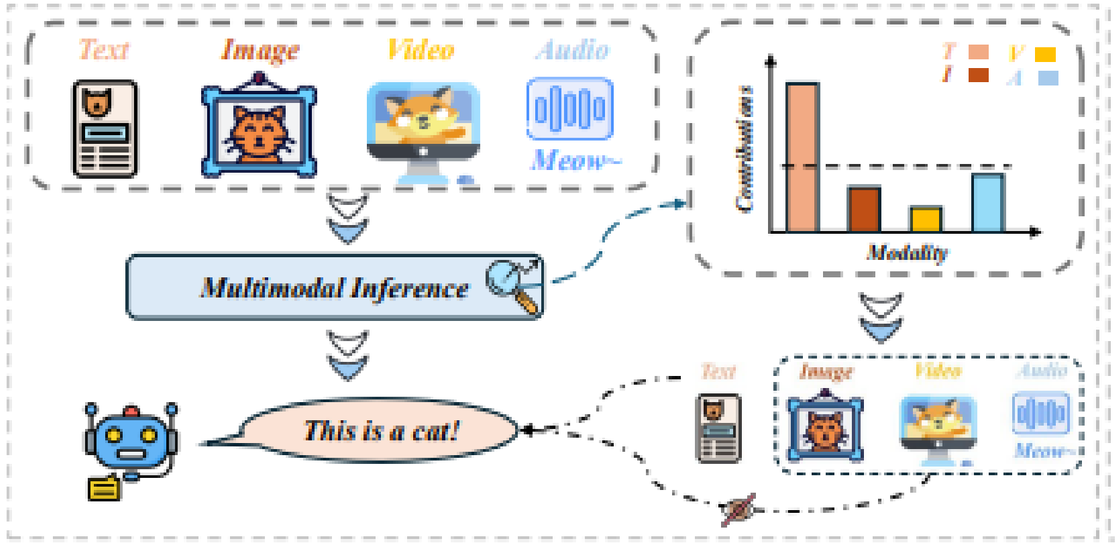
1. modality 간 representation이 다름.
2. feature space가 불균형함. (텍스트는 하나인데 이미지는 여러장이 나오는 경우)
3. 모델이 특정 modality에 편향된 경우.

CAT



CAT

*

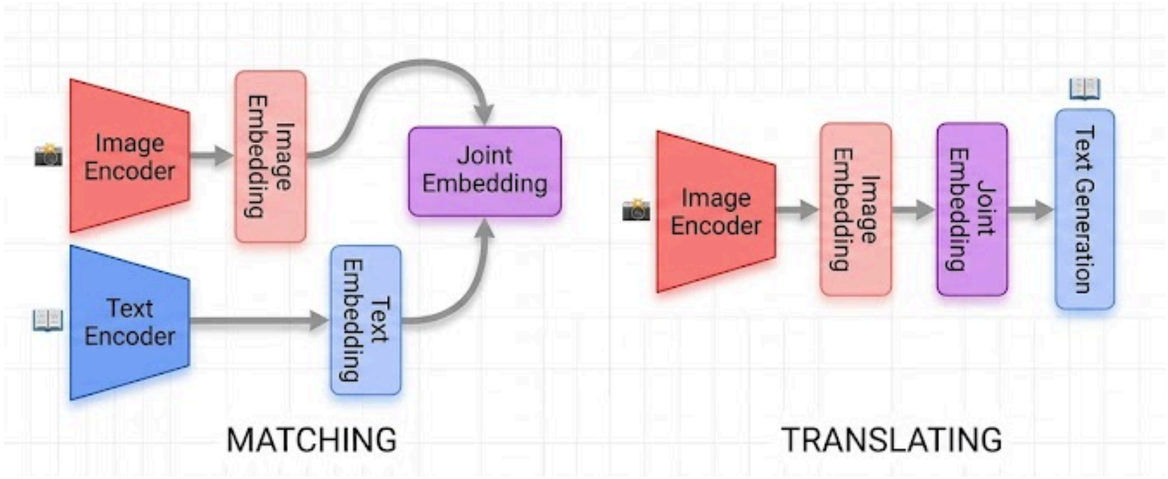


* Zheng, X., et al. (2025). "MLLMs are Deeply Affected by Modality Bias." arXiv preprint arXiv:2505.18657

03 Overview

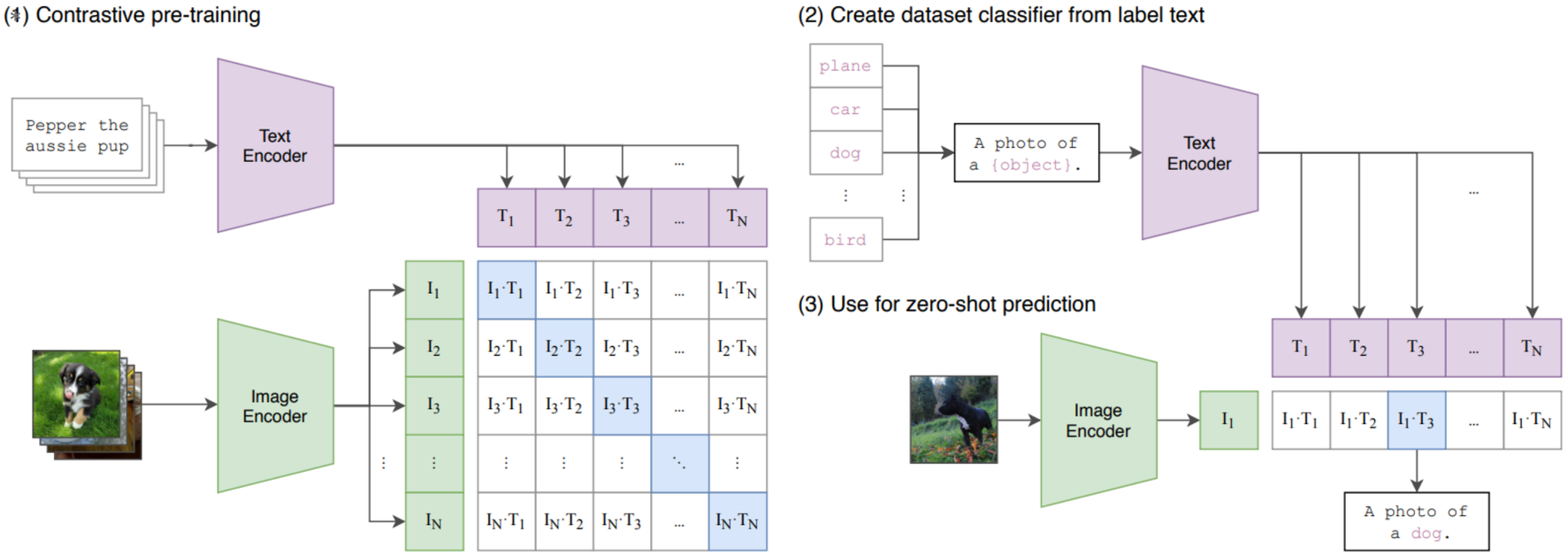
학습하기 어려운 문제를 해결하기 위해 많은 multi-modal embedding 기법이 존재.

1. matching 서로 다른 데이터를 joint해서 공통된 공간으로 보내 매칭.
2. translating 하나의 데이터를 다른 데이터로 변환. (captioning)
3. referencing 하나의 modality에서 같은 모달리티로 출력을 내보낼 때 다른 모달리티를 참조하여 상호작용.



03 Overview

CLIP : Encoder-based



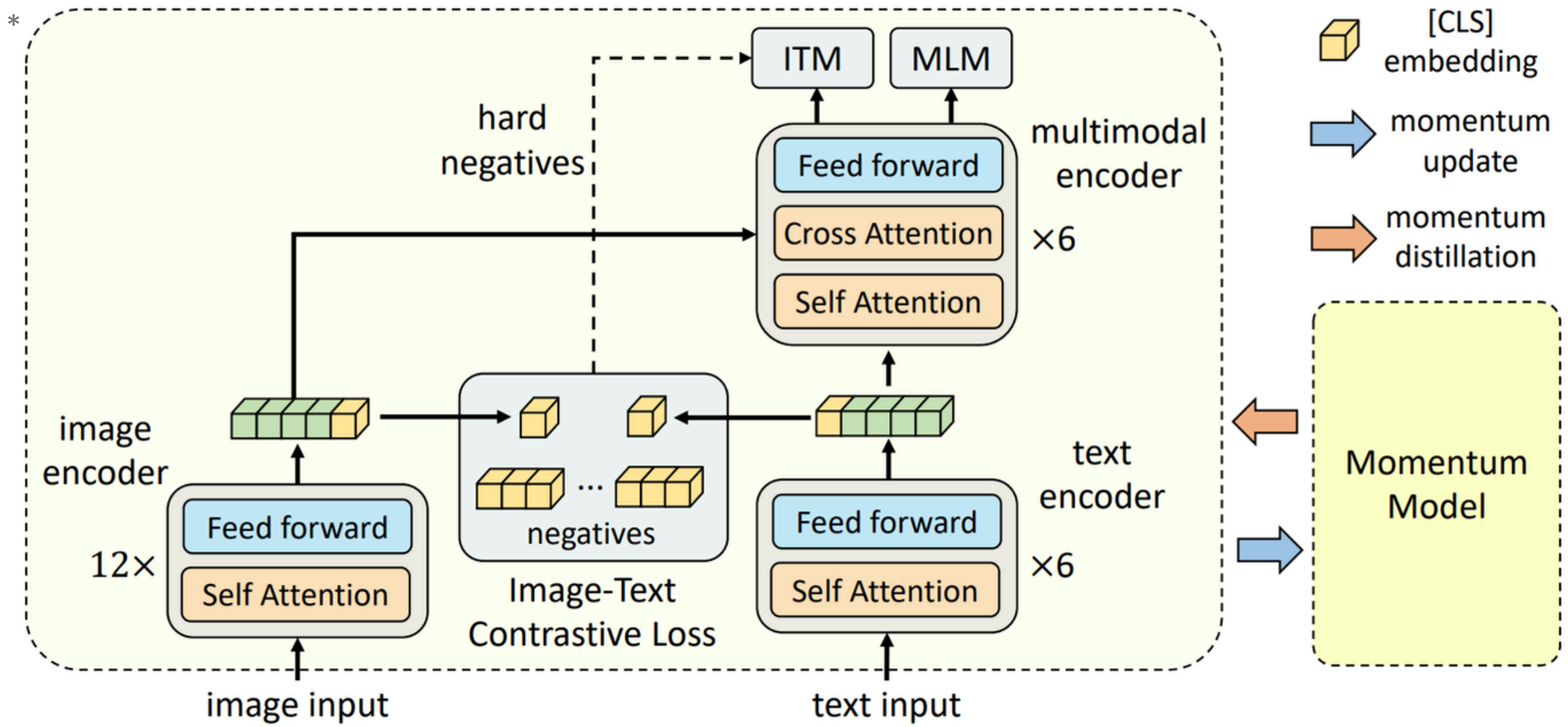
- Contrastive learning을 통해 text/image 입력에 대해 각각 encoder를 적용하여 나온 embedding의 유사도 계산.
- Diagonal term이 가장 유사도가 커지고, 나머지는 없어지도록 infoNCE loss 설정.

$$\mathcal{L}_{\text{NCE}} = -\mathbb{E}_{p(a,b)}\left[\log \frac{\exp(s(a,b))}{\sum_{\hat{b}\in\hat{B}}\exp(s(a,\hat{b}))}\right]$$

* clip:learning transferable visual models from natural language supervision 2021

03 Overview

ALBEF : Fusion-encoder-based



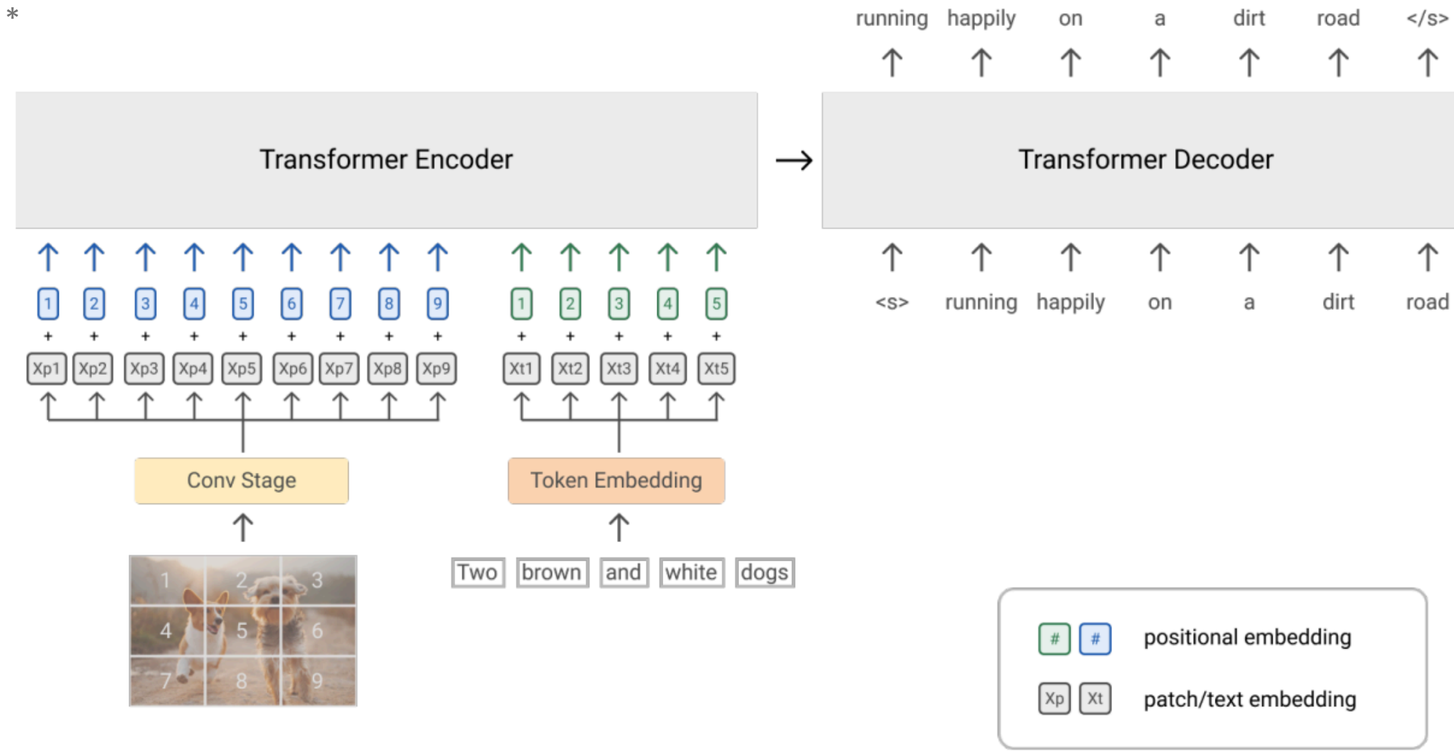
- Clip은 정렬만 하므로, 이미지와 텍스트의 세부 요소 간 상호작용을 나타낼 수 없음.
- Unimodal image representation, text representation을 ITC로 Align을 한 후, multimodal encoder에서 Cross Attention을 거쳐 Fuse.
- 한계: Generation 능력이 부족함.

* Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

03 Overview

SimVLM : Encoder-decoder-based

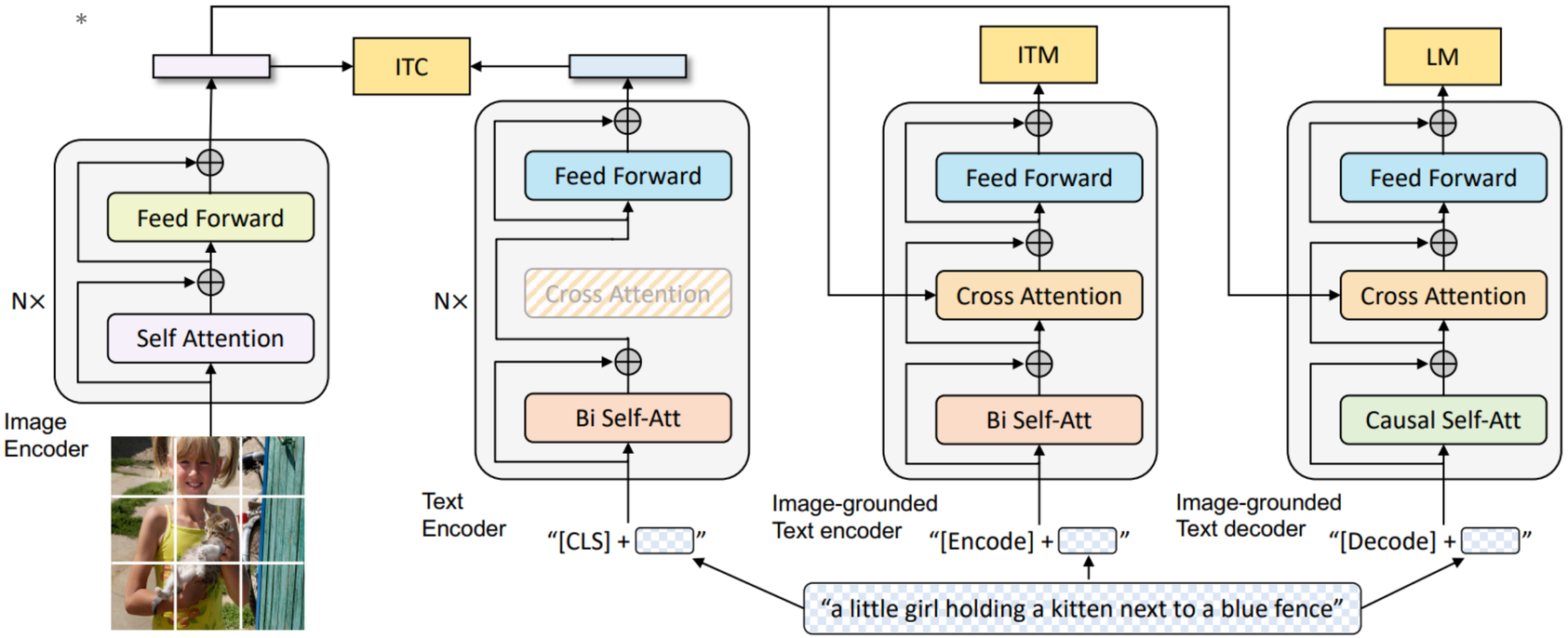
*



- Prefix를 concat한 언어 모델 학습으로 텍스트 생성 task까지 통합.
- 한계: 이미지와 텍스트 정보가 cross-attention에서 복잡하게 엮여 Retrieval 시, 검색 비용이 높음.

* SIMVLM: SIMPLE VISUAL LANGUAGE MODEL PRETRAINING WITH WEAK SUPERVISION

Blip



- MED 구조: 필요에 따라 인코더로도 쓰고, 디코더로도 쓸 수 있는 유연한 모델.
- Unimodal Encoder, Image-grounded Text Encoder/Decoder로 작동할 수 있음.

- 한계: End-to-End 학습의 문제. 비용이 높고, 기존 모델의 성능 저하가 발생.

* BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

ITC, ITM, ITG loss

- **ITC loss - ALBEF와 동일**

Clip과 같은 Contrastive learning으로 visual transformer와 text transformer의 feature space를 align 하는 것이 목표.

$$s(I,T) = g_v(\boldsymbol{v}_{\text{cls}})^\top g'_w(\boldsymbol{w}'_{\text{cls}}) \text{ and } s(T,I) = g_w(\boldsymbol{w}_{\text{cls}})^\top g'_v(\boldsymbol{v}'_{\text{cls}}).$$

유사도 함수는 각각의 이미지와 텍스트에 대한 cls 토큰을 내적하여 사용.

$$p_m^{\text{i2t}}(I) = \frac{\exp(s(I,T_m)/\tau)}{\sum_{m=1}^{M}\exp(s(I,T_m)/\tau)}, \quad p_m^{\text{t2i}}(T) = \frac{\exp(s(T,I_m)/\tau)}{\sum_{m=1}^{M}\exp(s(T,I_m)/\tau)}$$

Softmax로 유사도를 확률로 변환한 후, negative는 0, positive pair는 1로 두고 cross-entropy 계산.

$$\mathcal{L}_{\text{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\left[\text{H}(\boldsymbol{y}^{\text{i2t}}(I), \boldsymbol{p}^{\text{i2t}}(I)) + \text{H}(\boldsymbol{y}^{\text{t2i}}(T), \boldsymbol{p}^{\text{t2i}}(T))\right]$$

어째서 유사도를 확률로 변환하는가?

- 단순 유사도는 상대성을 고려하지 않기 때문에, 유사도끼리 비교할 수 없으므로, 정답과 오답의 유사도를 상대적으로 비교하기 위해.

* Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

ITC, ITM, ITG loss

- **ITM loss - ALBEF와 동일**

ITC 점수를 기반으로, 유사도가 높은 데이터를 필터링하여 itm으로 보내어,
Imformative negative를 찾기 위해, 유사도가 높은 negative를 학습하는 hard negative 전략.

positive를 맞추거나, 잘못된 쌍을 주고 negative인 것을 맞추게 함.

ground truth label을 나타내는 2차원 one-hot vector와 유사도로 도출한 확률을 cross entropy하면 base itm loss가 됨.

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T)\sim D}\mathrm{H}(\boldsymbol{y}^{\text{itm}}, \boldsymbol{p}^{\text{itm}}(I,T))$$

ITC, ITM, ITG는 동시에 병렬로 진행되나, 이 탓에 ITC가 ITM보다 선행함.

* Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

ITC, ITM, ITG loss

- **ITG loss**

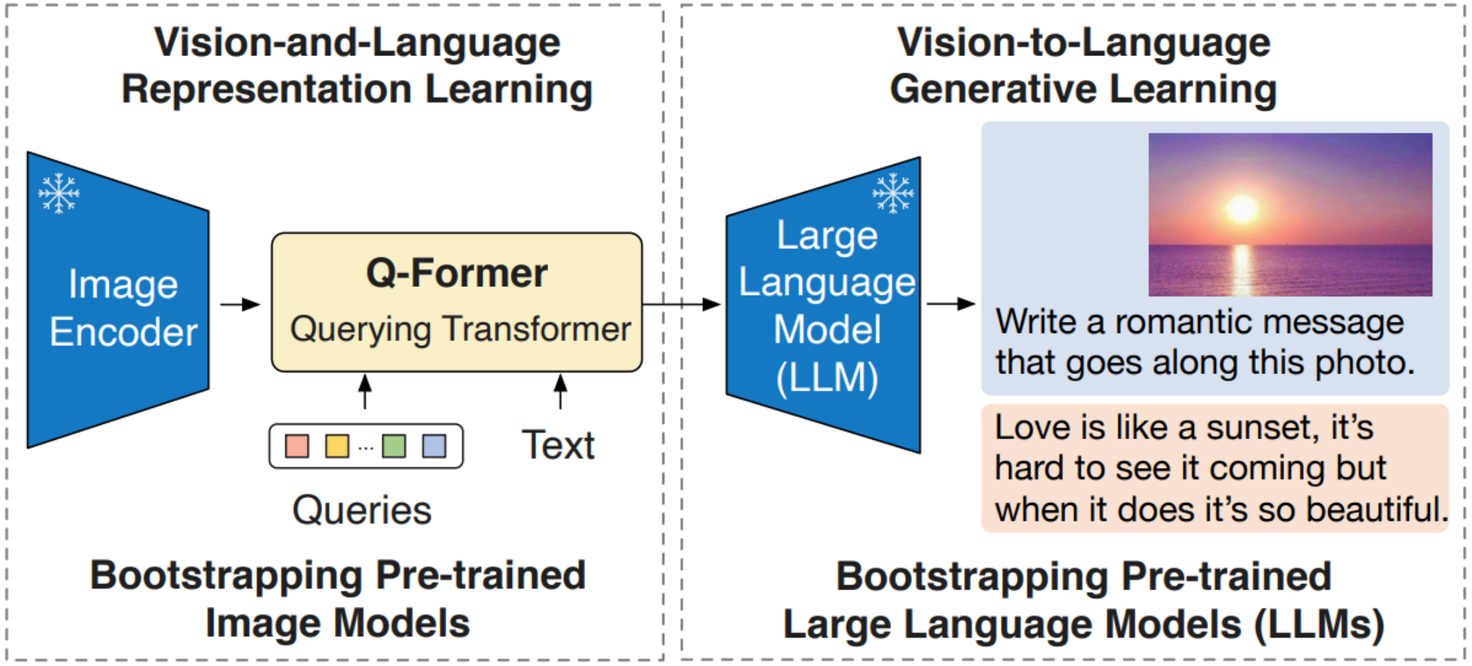
  Auto-regressive 방식으로, Cross entropy loss를 최적화.

  학습의 효율화를 위해,

  세 개의 loss를 합산해서 한번에 역전파하여, q-former와 쿼리를 동시에 업데이트.

04 Method

blip2



blip에서의 bootstrap이 capfilt를 통해 데이터를 정제해 질을 높이는 것이었다면 blip2에서는 파운데이션 모델의 능력을 해치지 않고 내가 원하는 작업에 잘 쓴다는 의미.

- **blip의 문제를 해결하기 위해, 거대 모델을 frozen하되 Q-former 구조를 통해 vision-text 멀티모달을 달성.**

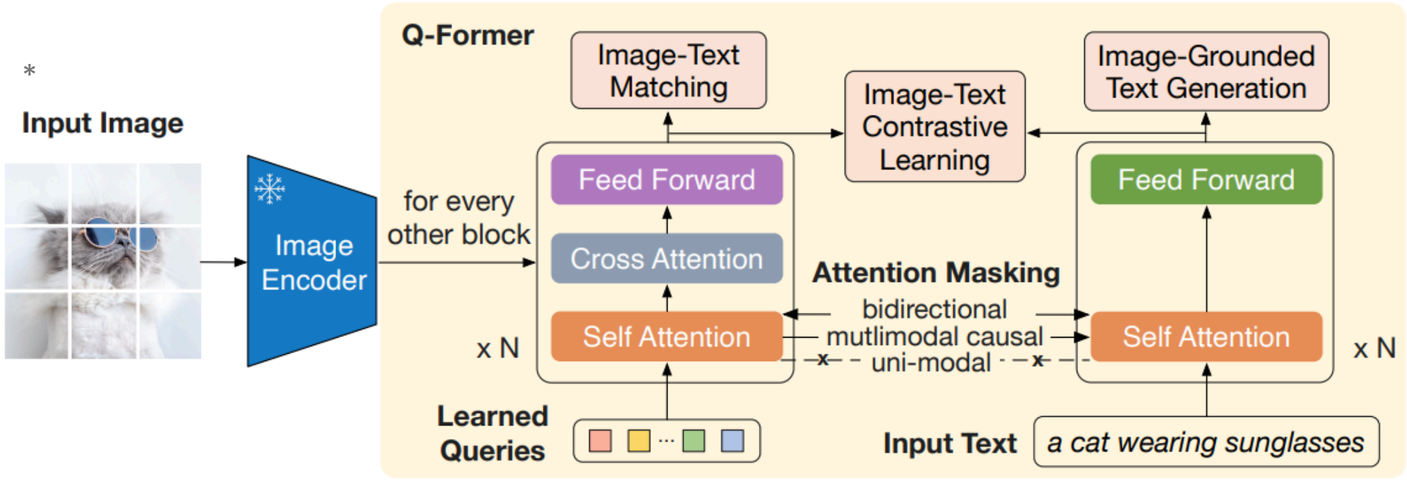
✓ Stage 1, vision-and-language representation learning
✓ Stage 2, vision-to-language generative learning

* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

04 Method

Q-former

✓ Image에 대한 정보를 LLM이 이해할 수 있게 만들어 보내준다
✓ Query를 사용하여 정보를 압축해 적은 파라미터로 높은 효율을 보여준다.



Cross-Attention (시각 정보 추출) - Q:qurey. K,V : image

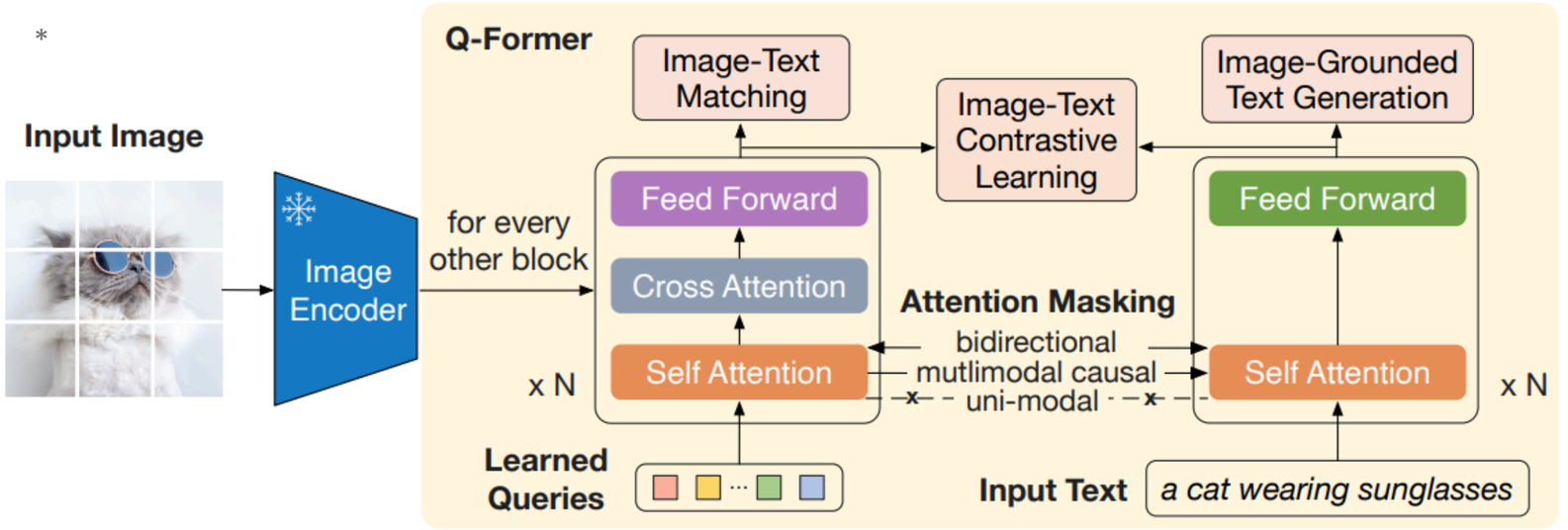
- Learned Queries (32개)만 참여. Query는 초기에 무작위 값을 가짐.
- 텍스트는 Image Encoder의 출력에 대해 무지하며. 32개의 쿼리만이 이미지를 훑어보며 LLM에게 전달할 핵심 시각 정보를 압축.

Self-Attention

- Learned Queries(32개)와 Input Text 토큰들이 Concatenate 동일한 Self-Attention 레이어를 통과.
- 쿼리는 텍스트를 보고, 텍스트는 쿼리를 보면서 서로 정보를 교환. 목적에 따라 masking을 하여 제어.

* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Q-former

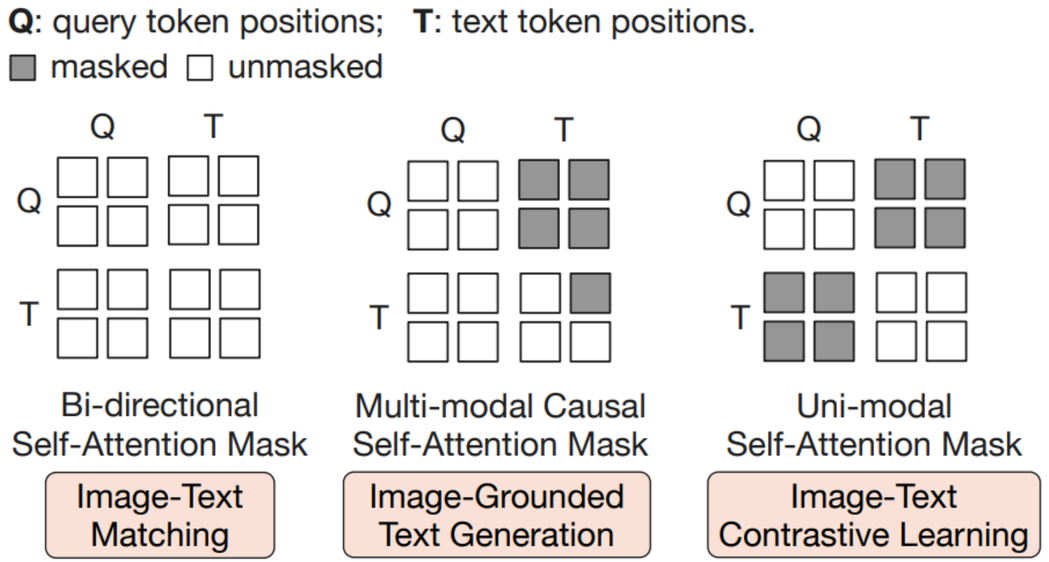


- ITC (대조 학습): 현재 배치의 이미지와 원래 짝인 텍스트 쌍은 Positive, 배치 내의 다른 텍스트들은 오답 Negative으로 두고 Loss를 계산.
- ITM (이진 분류): 올바른 짝이 들어오면 true, 다른 짝은 false임을 맞추도록 학습.
- ITG (생성): 이미지를 보고 다음 단어를 예측할 때, 실제 데이터셋에 있는 텍스트의 다음 단어가 정답 레이블.

* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Q-former

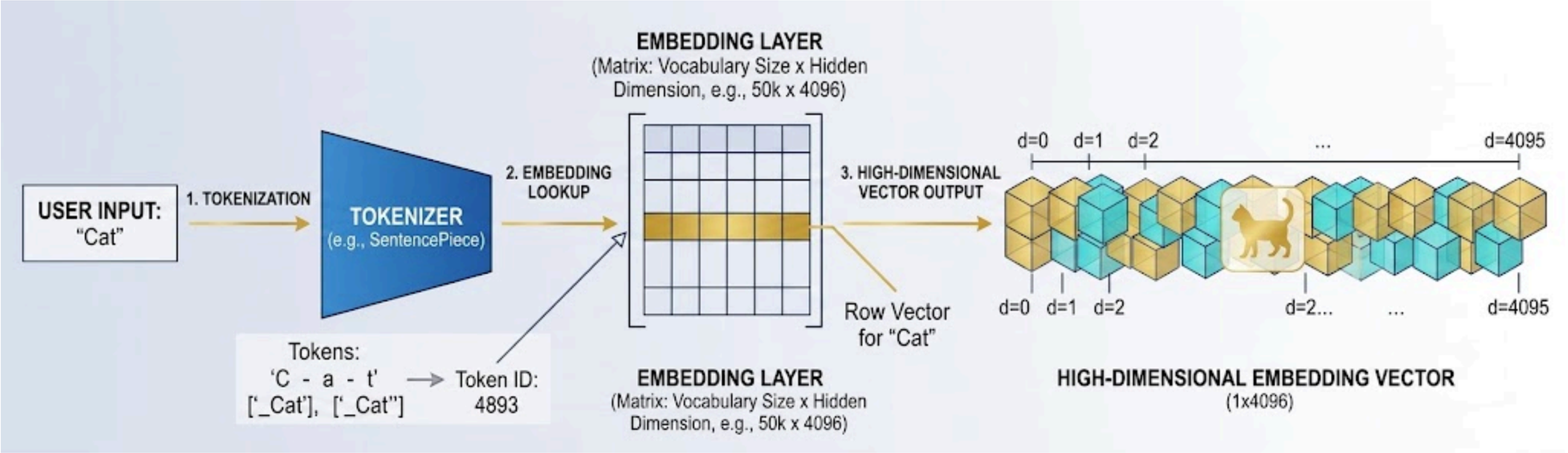
*



- ITC (대조 학습): Text와 Query를 Align하기 때문에 date leakage를 막기 위해서 자기 자신만을 볼 수 있도록 마스킹.
- ITM (이진 분류): Text와 Query가 쌍을 찾는 과정이 필요하므로 마스킹 X.
- ITG (생성): t-1 시점까지의 Text와 Query를 바탕으로 t시점의 text를 맞춰야하므로 미래 text에 대해 마스킹.

* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Stage 2

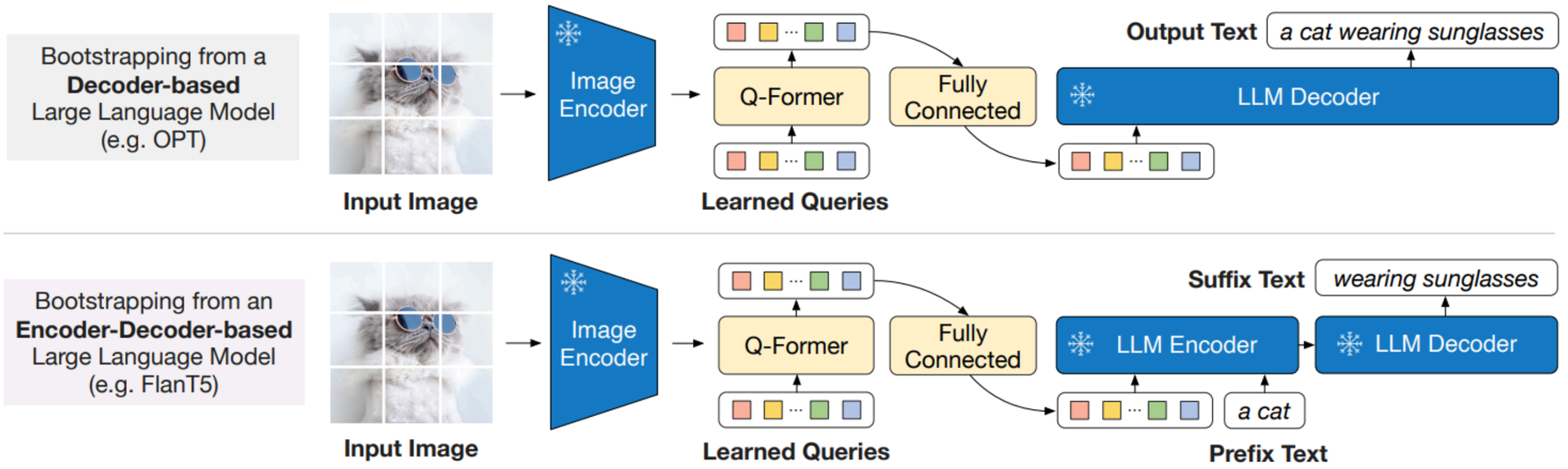


- 일반적인 LLM: 사용자가 "Cat"이라는 단어를 입력. Tokenizer가 이를 토큰으로 변환 → Embedding Layer가 토큰을 고차원 벡터로 변환 1x4096.
- BLIP2: Q-Former는 이미지를 보고 32x768의 벡터를 출력, Fully Connected Layer로 LLM 입력과 차원 맞춰줌. 생성된 32x4096 크기의 벡터를 LLM의 원래 텍스트 임베딩 앞부분에 Concat.

04 Method

Stage 2

*



- Decoder based는 text와 label를 전부 decoder에 입력하여 학습.
- Encoder-Decoder-based는 분리하여 학습.

하지만 모두 공통적으로 auto-regressive의 cross-entropy loss로 fully connected, q-former와 query를 최적화한다.

* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Model Pre-training

- **BLIP과 동일한 데이터셋**

  COCO, Visual Genome, CC3M, CC12M, SBU, LAION400M

- **CapFilt 적용**

  10개 캡션을 생성 -> 유사도 기반으로 순위를 매김 -> 상위 2개의 캡션 중 무작위 하나를 사용.

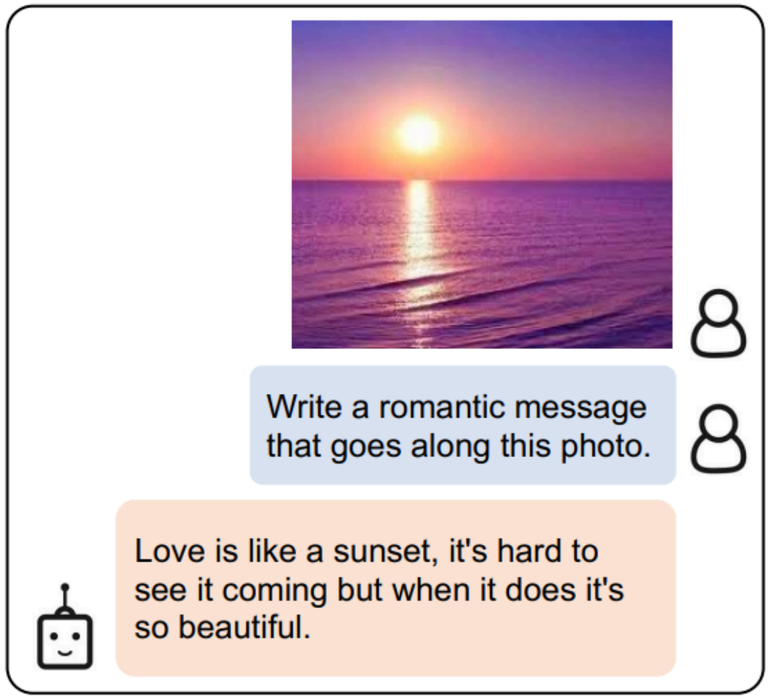
- **Foundation model**

  Image - CLIP : ViT-L/14, EVA-CLIP : ViT-g/14
  Language - OPT, FlanT5

05 Conclusion

Zero-shot Image-to-Text Generation

이미지만 보고 풀 수 없는 문제.



Models	#Trainable Params	#Total Params	VQAv2 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
VL-T5$_{\text{no-vqa}}$	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	**50.6**	-
BLIP-2 ViT-L OPT$_{2.7B}$	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT$_{2.7B}$	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT$_{6.7B}$	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5$_{XL}$	103M	3.4B	62.6	62.3	39.4	44.4
BLIP-2 ViT-g FlanT5$_{XL}$	107M	4.1B	63.1	63.0	40.7	44.2
BLIP-2 ViT-g FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**	45.9	**44.7**

Table 2. Comparison with state-of-the-art methods on zero-shot visual question answering.

Image + text를 LLM의 input으로 사용.
모델 답변 후보 중 최적의 후보를 찾기 위해 Beam search 활용.
Length-penalty를 -1로 설정하여 짧은 답변을 내도록 유도.

대체적으로 더 성능이 좋은
image encoder, LLM 사용시 성능이 좋음.

* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

05 Conclusion

Image captioning, Visual Question Answering

Image captioning- ‘a photo of’ 프롬프트로 텍스트 생성을 유도.

Models	#Trainable Params	VQAv2 test-dev	VQAv2 test-std
Open-ended generation models			
ALBEF (Li et al., 2021)	314M	75.84	76.04
BLIP (Li et al., 2022)	385M	78.25	78.32
OFA (Wang et al., 2022a)	930M	82.00	82.00
Flamingo80B (Alayrac et al., 2022)	10.6B	82.00	82.10
BLIP-2 ViT-g FlanT5$_{XL}$	1.2B	81.55	81.66
BLIP-2 ViT-g OPT$_{2.7B}$	1.2B	81.59	81.74
BLIP-2 ViT-g OPT$_{6.7B}$	1.2B	**82.19**	**82.30**
Closed-ended classification models			
VinVL	345M	76.52	76.60
SimVLM (Wang et al., 2021b)	∼1.4B	80.03	80.34
CoCa (Yu et al., 2022)	2.1B	82.30	82.30
BEIT-3 (Wang et al., 2022b)	1.9B	**84.19**	**84.03**

VQA- 이미지를 바탕으로 주관식 질문에 대한 정답 생성
질문 text를 Q-former의 입력으로 넣어
질문에 대한 영역에 focus하게 유도.

최종적으로 Query와 질문 text가
LLM에 input으로 입력됨.

* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

05 Conclusion

Image-Text Retrieval

Model	#Trainable Params	Flickr30K Zero-shot (1K test set)						COCO Fine-tuned (5K test set)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Dual-encoder models													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	**100.0**	81.5	95.6	97.8	84.8	96.5	98.3	67.2	**87.7**	**92.8**
Fusion-encoder models													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
Dual encoder + Fusion encoder reranking													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	**100.0**	**100.0**	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 ViT-L	474M	96.9	**100.0**	**100.0**	88.6	97.6	**98.9**	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 ViT-g	1.2B	**97.6**	**100.0**	**100.0**	**89.7**	**98.1**	**98.9**	**85.4**	**97.0**	**98.5**	**68.3**	**87.7**	92.6

텍스트 생성 작업이 아니므로, LLM 없이 ViT + Q-Former만 사용.
zero-shot에서 아주 뛰어난 성능.

* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

05 Conclusion

Limitation

- **Q-former의 query vector 32개는 과연 충분한가?**

  128x128과 2048x2048 이미지를 같은 식으로 처리한다는 것이 의문점.

- **image-text 1개씩을 쌍으로 학습하는 것의 한계**

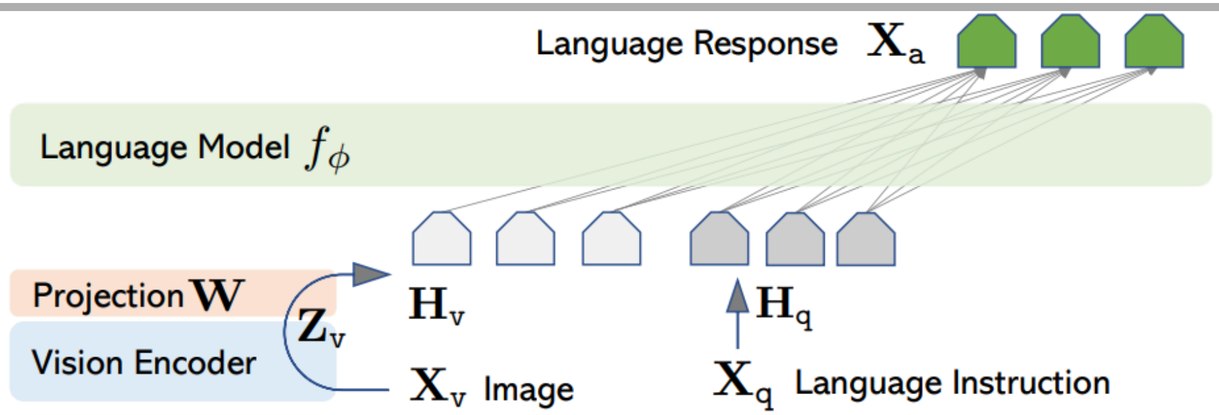
  이미지를 여러 개 입력받거나 하는 방식은 불가능하며, few-shot 성능이 상승하지 않음.

- **Foundation model에 의존성이 높음**

  ViT 사용시, 이미지의 형태가 강제되어 왜곡이 발생.
  LLM의 환각 효과에서 자유롭지 못함.

05 Conclusion

추가 탐색



	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

blip2와 다른 방향으로 q-former과 같은 병목을 버리고
이미지 인코더에서 나온 토큰들을 다 llm에 넣거나
이미지 인코더조차 없이 patch로 나누어
linear layer를 통해 차원을 일치시킨 후 llm에 넣는 시도.

- Visual Instruction Tuning

05 Conclusion

추가 탐색

	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

BLIP2의 후속 논문인 InstructBLIP에서 Q-Former를 버리지 않고
발전시켜 LLaVA 같은 모델들보다 더 높은 성능을 보임.

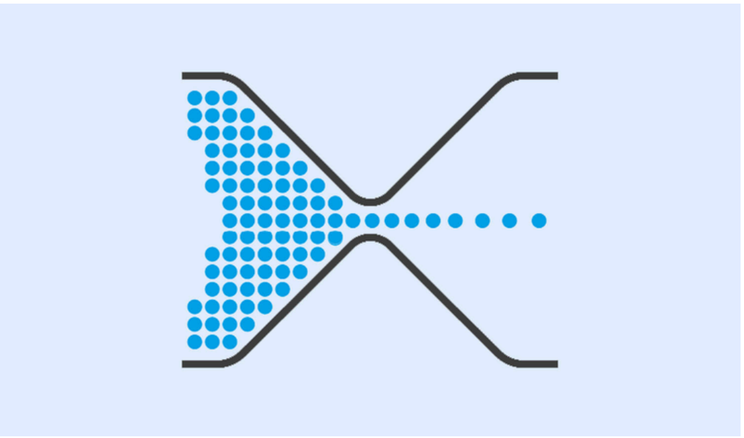
- InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning

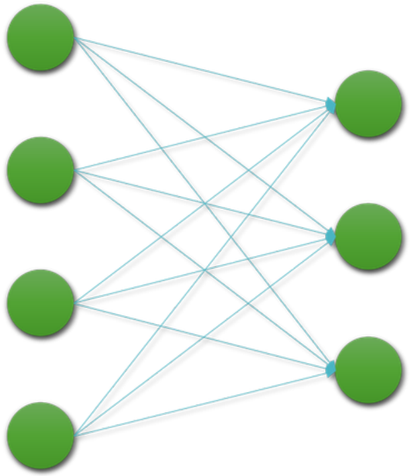
05 Conclusion

LLaVA VS QWEN, InternVL

차이점

Image information
-vision 정보를 그대로 LLM에 주입, LLaVA.
-vision 정보를 압축하여 LLM에 주입, QWEN.

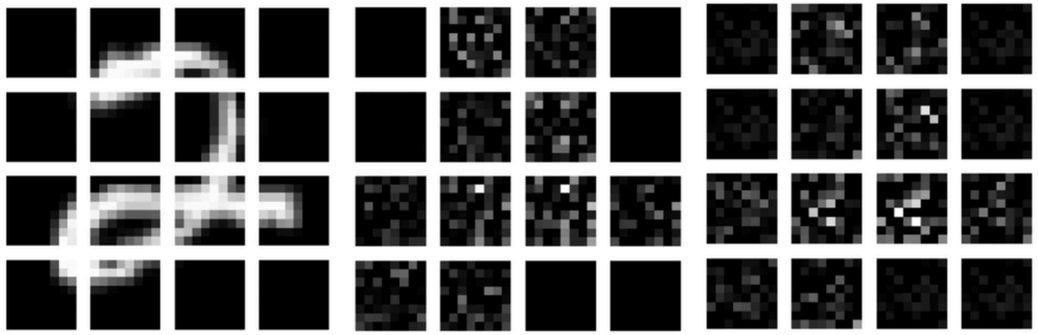




LLaVA, InternVL VS QWEN

차이점

Dynamic High-resolution Slicing, LLaVA
-patch로 쪼갠 뒤, ViT 통과 후 LLM에서 병합.
Native Dynamic Resolution, QWEN
-어떤 비율과 크기의 이미지도 처리할 수 있는 ViT.



동향

InternVL - LLM 뿐만 아니라, Vision Encoder 역시 Large하게 가져감. Visual Reasoning 능력 확보.

Gemini, GPT-4o, Qwen2.5-VL - Native Early Fusion
이미지, 텍스트, 비디오, 오디오를 따로 쪼개지 않고, pre-training 단계에서 모든 모달리티의 토큰을 섞어 하나의 거대한 트랜스포머로 한 번에 학습. 모달리티 경계의 소멸.

Thank You

2026.07.20

MuliLAB Journal Club
인공지능응용학과 장승호