간단 논세

Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models

Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." International conference on machine learning. PMLR, 2023.

25.01.15

유하영

Introduction

VL 모델의 pre training 비용이 end-to-end 훈련으로 인해 부담이 높다.

본 논문에서는, pretrained image Encoder와 frozen LLM을 통해 VL을 bootstrap하는 전략인 BLIP-2 제안

- Query Transformer를 통해 modality gap 해소

    1) frozen image encoder로부터 vision-language representation 학습을 bootstrap

    2) frozen language encoder로부터 vision-language generative 학습을 bootstrap

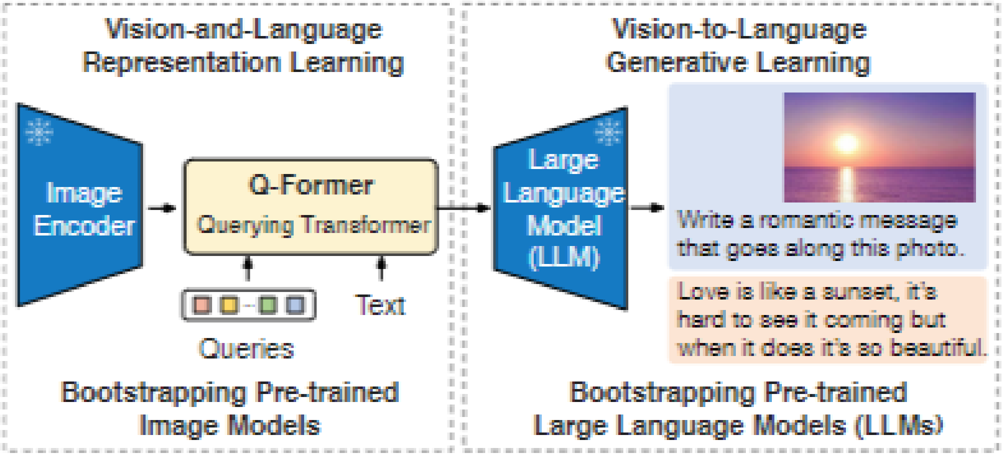
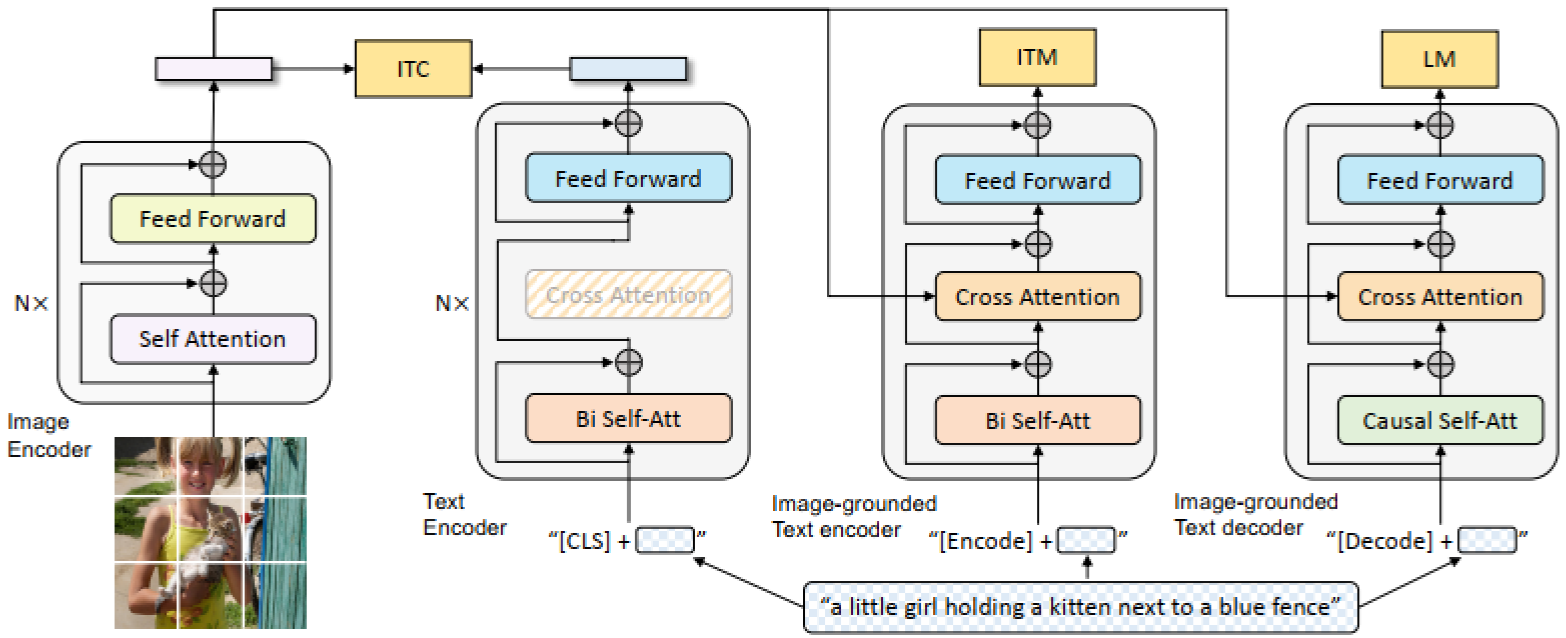


Figure 1. Overview of BLIP-2's framework. We pre-train a

성과
- 기존 method에 비해 train 매개변수가 적었음에도 불구하고 SOTA 달성
- Zero-shot 이미지 텍스트 생성능력 향상

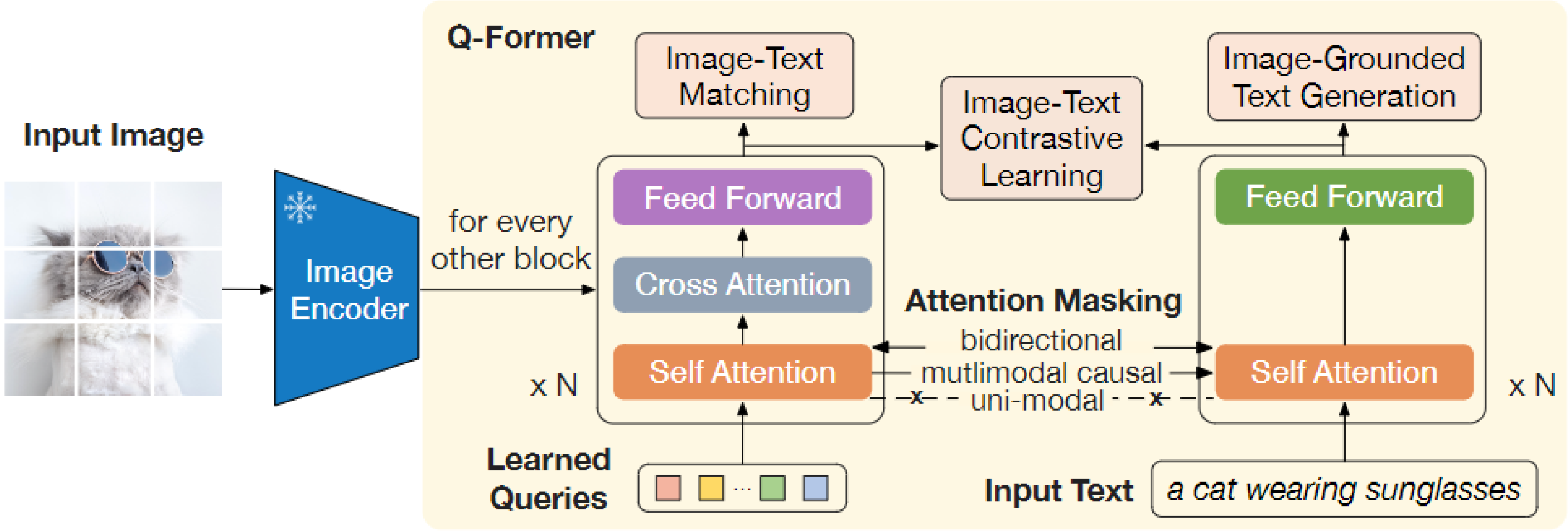
Introduction



Blip architecture

Method

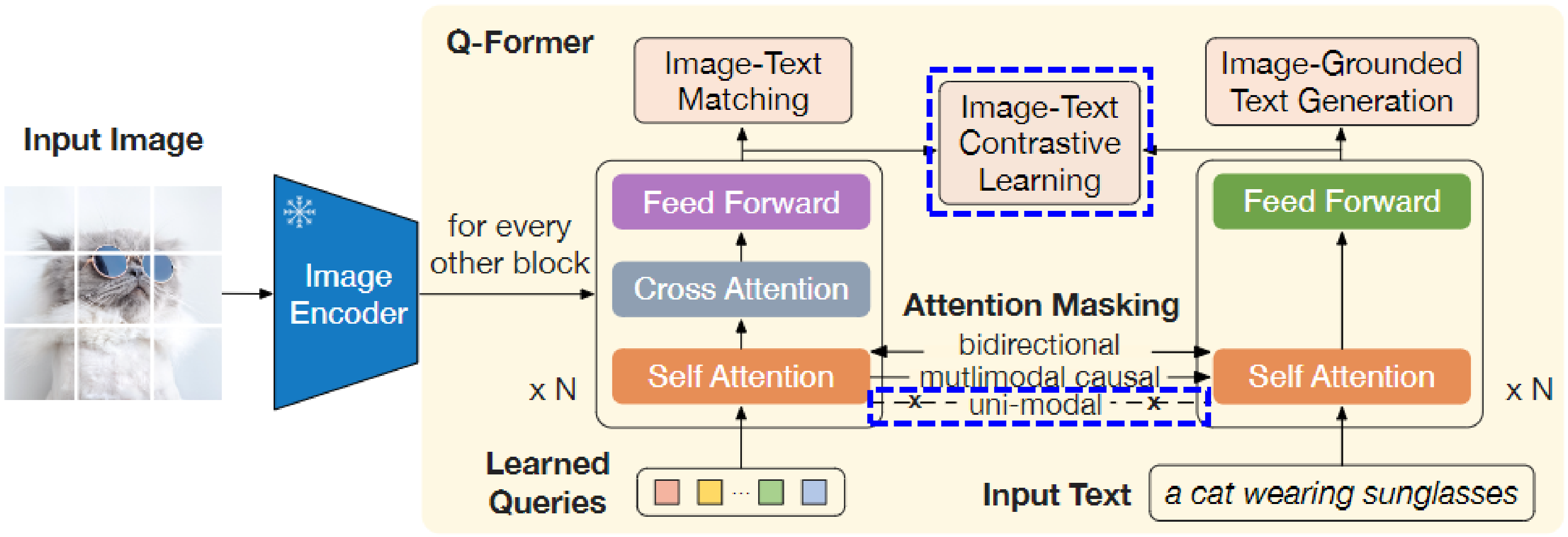
1. Frozen Image Encoder   2. Q- Former

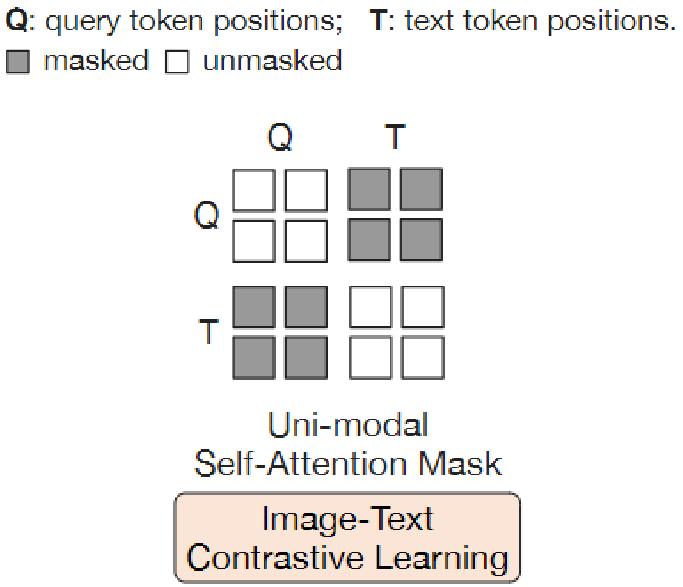


Method

1. Frozen Image Encoder

2. Q- Former

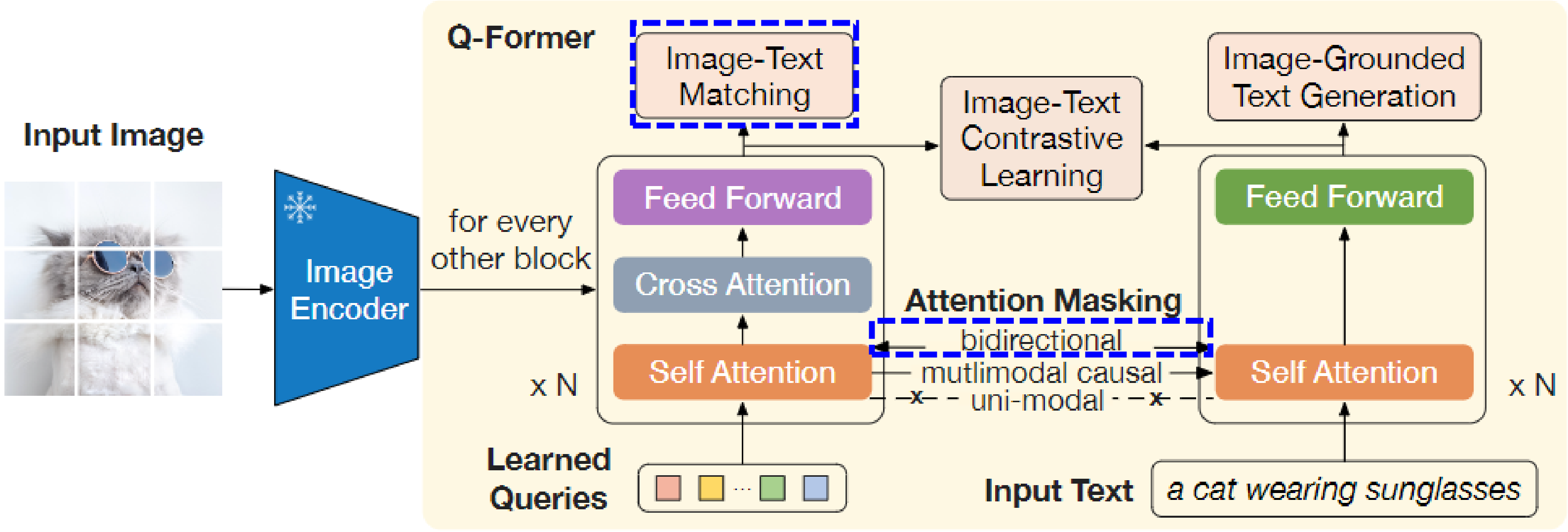


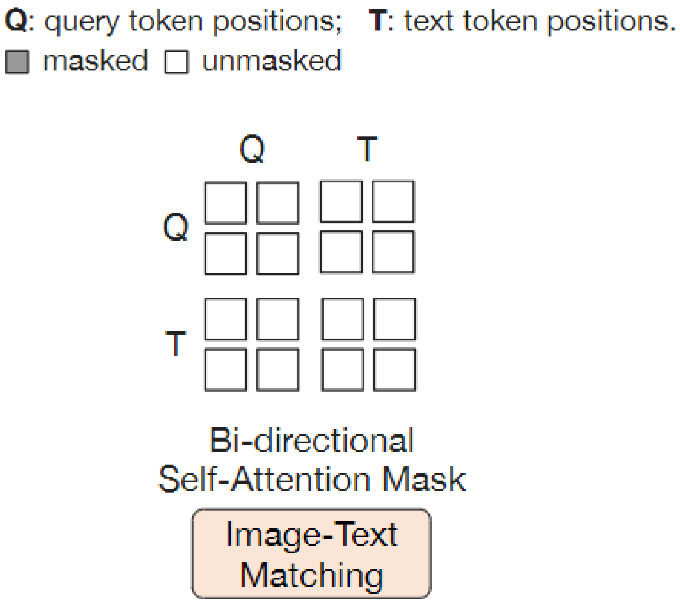


Method

1. Frozen Image Encoder

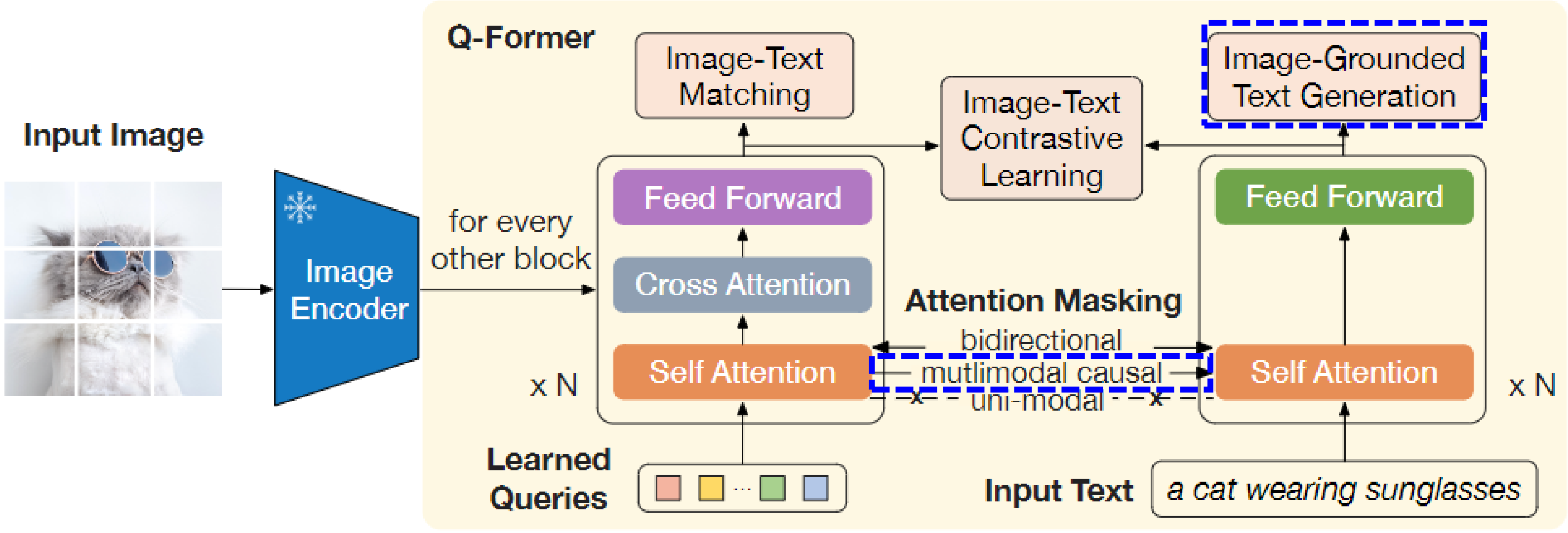
2. Q- Former

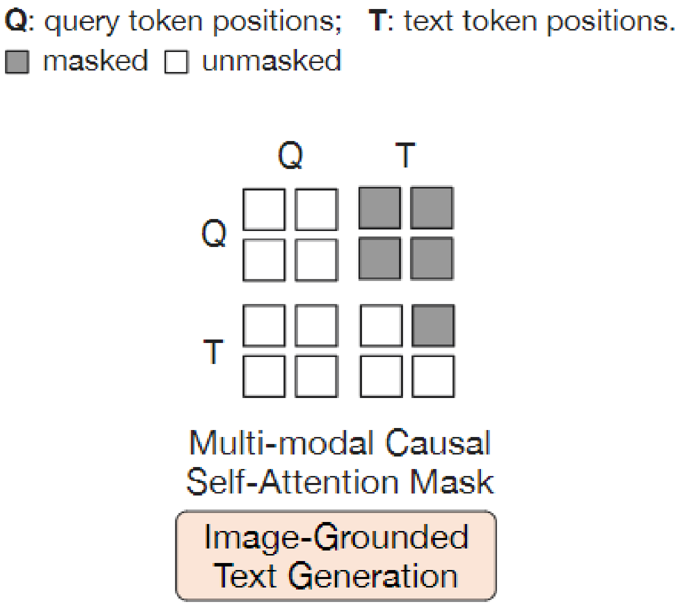




Method

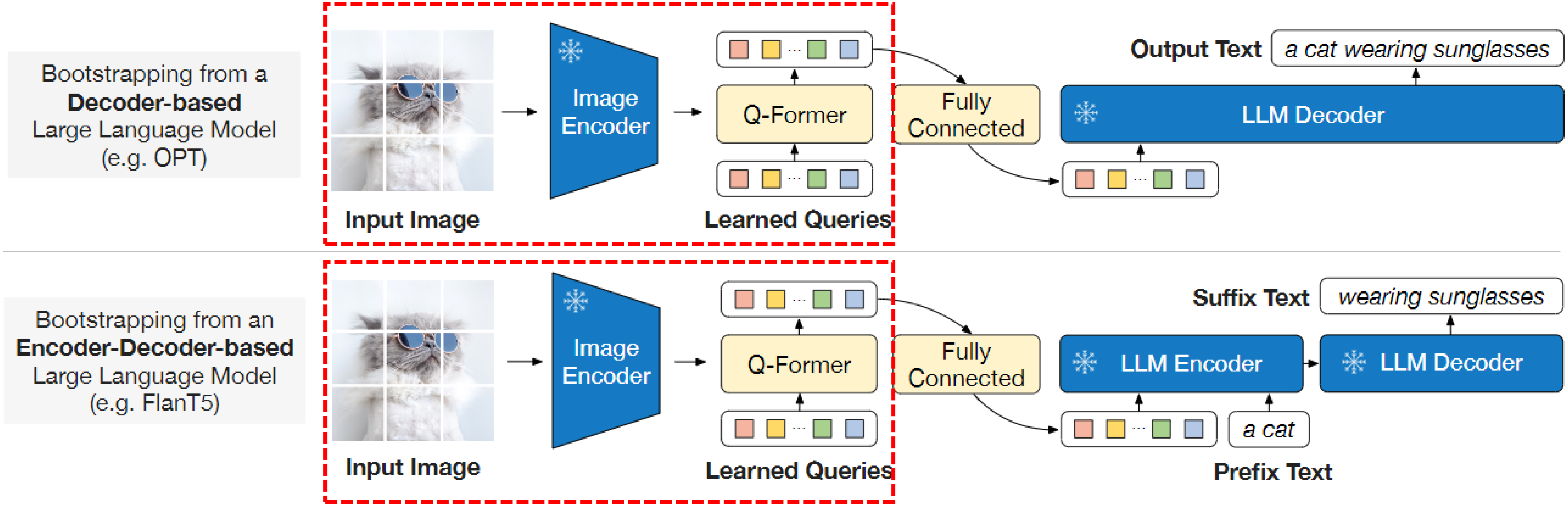
1. Frozen Image Encoder   2. Q- Former





Method

3. Frozen LLM



Experiments

Models	#Trainable Params	#Total Params	VQAv2 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
VL-T5$_{no\text{-}vqa}$	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.

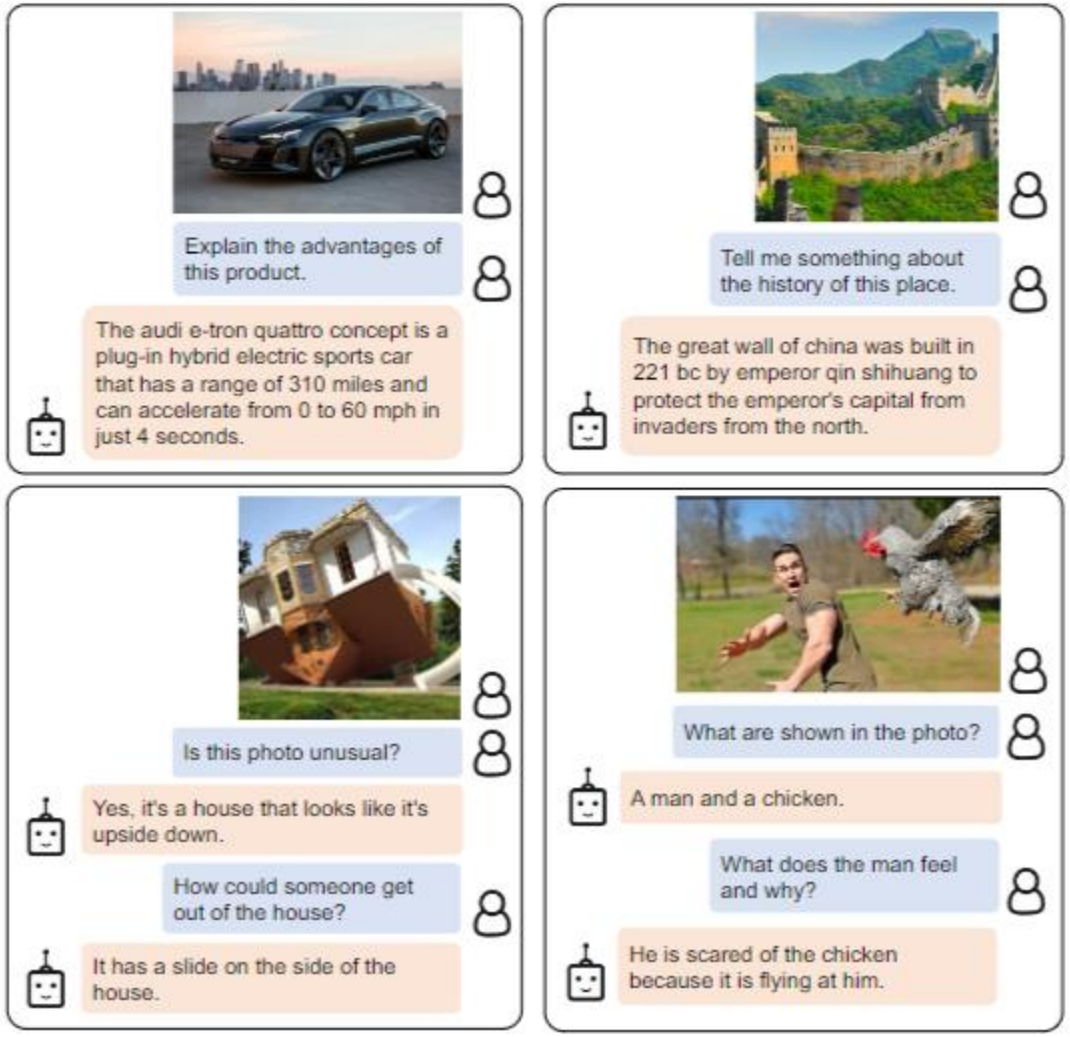
Models	#Trainable Params	VQAv2 test-dev	VQAv2 test-std
Open-ended generation models			
ALBEF (Li et al., 2021)	314M	75.84	76.04
BLIP (Li et al., 2022)	385M	78.25	78.32
OFA (Wang et al., 2022a)	930M	82.00	82.00
Flamingo80B (Alayrac et al., 2022)	10.6B	82.00	82.10
BLIP-2 ViT-g FlanT5$_{XL}$	1.2B	81.55	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**

Table 4. Comparison with state-of-the-art models fine-tuned for visual question answering.

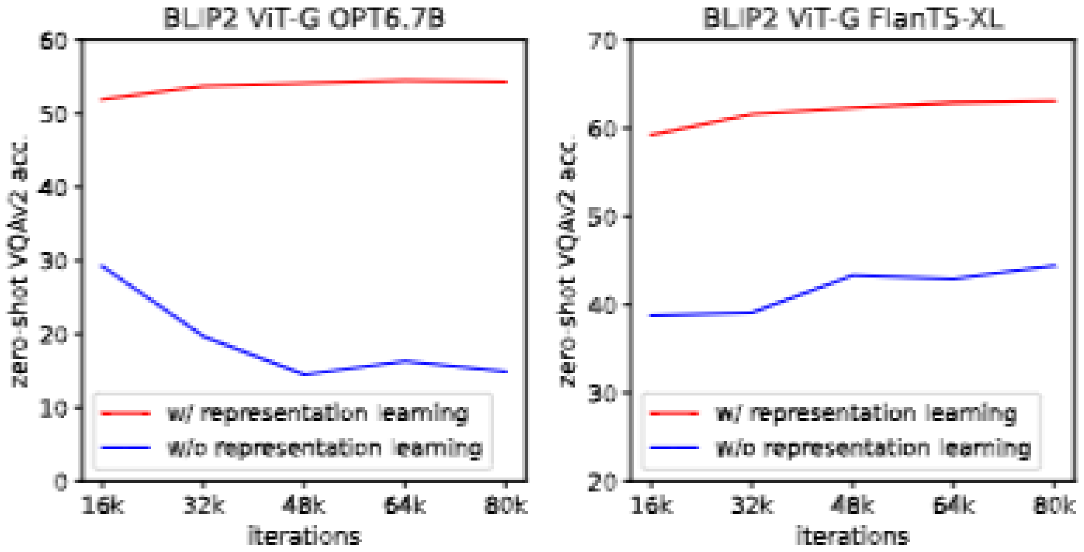
Models	#Trainable Params	NoCaps Zero-shot (validation set) in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Fine-tuned Karpathy test B@4	C
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	80.9	11.3	37.4	127.8
VinVL (Zhang et al., 2021)	345M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
BLIP (Li et al., 2022)	446M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7
OFA (Wang et al., 2022a)	930M	-	-	-	-	-	-	-	-	**43.9**	145.3
Flamingo (Alayrac et al., 2022)	10.6B	-	-	-	-	-	-	-	-	-	138.1
SimVLM (Wang et al., 2021b)	~1.4B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP-2 ViT-g OPT$_{2.7B}$	1.1B	123.0	15.8	117.8	15.4	123.4	**15.1**	119.7	15.4	43.7	**145.8**
BLIP-2 ViT-g OPT$_{6.7B}$	1.1B	**123.7**	15.8	119.2	15.3	124.4	14.8	121.0	15.3	43.5	145.2
BLIP-2 ViT-g FlanT5$_{XL}$	1.1B	**123.7**	**16.3**	**120.2**	**15.9**	**124.8**	**15.1**	**121.6**	**15.8**	42.4	144.5

Comparison with state-of-the-art image captioning methods on NoCaps and COCO Captio

Experiments



instructed zero-shot image-to-text generation using a BLIP-2 model



Effect of vision-language representation learning onvision-to-language generative learning.

Conclusions

LLM이 단일 Image-Text 대응관계를 학습했기 때문에 텍스트 표현 다양성이 부족.추후 다중 데이터셋을 개발할 예정이라고 함
또한, LLM 모델의 성능에 크게 의존할 수 있다는 한계점

Frozen model간의 modality gap을 메우기 위한 새로운 방법을 제안하고, 다양한 VL task에서 성능을 개선
또한 Pre-Training 과정에서의 Trainable Parameter 개수를 줄여 학습 효율성을 높임