BLIP-2

Abstract

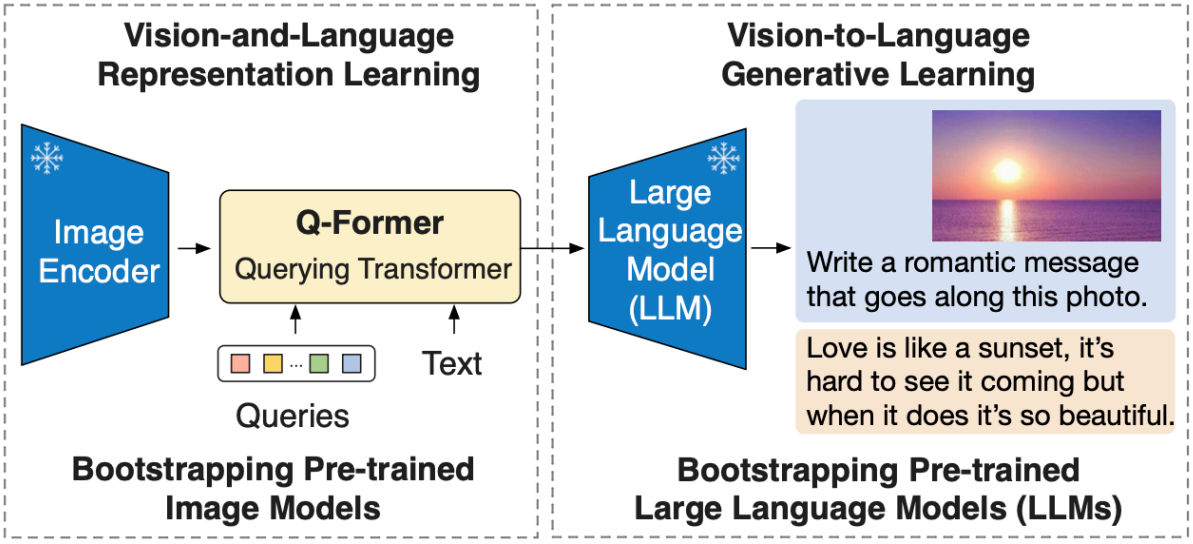


Figure 1. Overview of BLIP-2's framework. We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen LLM, which enables zero-shot instructed image-to-text generation (see Figure 4 for more examples).

BLIP-2의 경우 상용화 되어 있는 Image Encoder와 LLM을 활용한 VLM을 제안합니다.

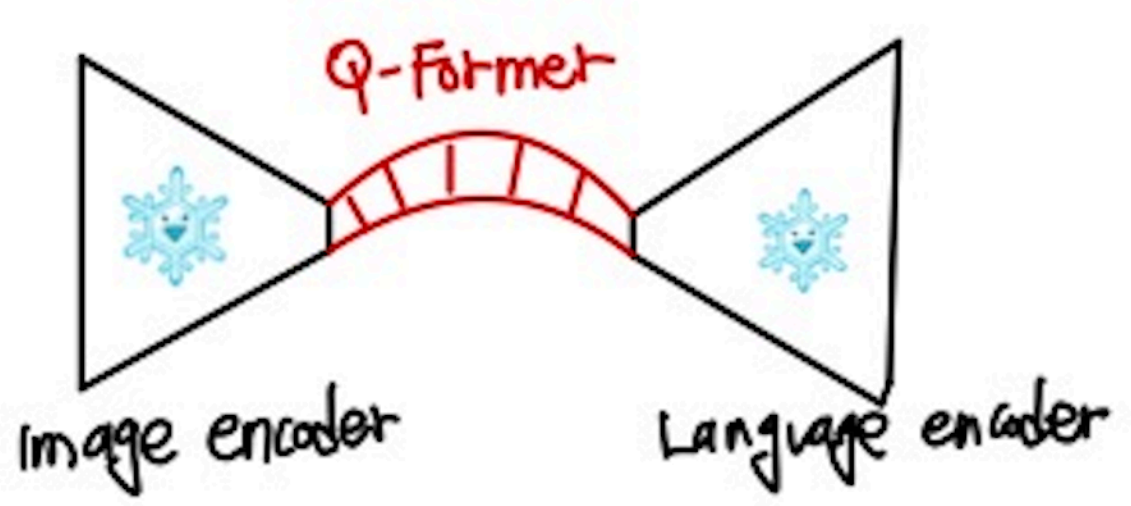
효율적인 학습을 위해서 사용화된 Image Encoder와 LLM의 경우 학습하지 않고 Frozen 된 상태로 사용합니다. 그리고 이미지와 텍스트의 모달 차이를 최소화 하기 위해서 Q-Fromer라는 작은 transformer를 활용합니다. 즉, Q-Former는 두 모달리티의 차이를 줄여주는 다리 역할을 하게 됩니다. 그래서 BLIP-2의 경우 메모리 효율적인 학습을 통해서 VQA (Vision Question & anwering) task에서 SOTA를 달성했다고 합니다.

Introduction

이전 VLP의 연구들의 경우 SOTA를 달성하기 위해서 수많은 컴퓨팅 자원을 필요로 하였습니다. 그리고 자연스럽게 Image와 Text 각각의 분야에서 나온 방법들로 부터 많은 정보를 얻으며 발전하였습니다. 이미지와 텍스트의 정보를 활용 시 “치명적 망각 (catastrophic foregetting) “ 을 막기 위해서 각각의 모달리티들을 Frozen 했어야 합니다. (그렇지 않는 경우 LLM 이 가지고 있는 본연의 zero-shot 의 능력이 저하되며, 다양한 정보가 섞여서 성능 저하로 이어지게 된다). 하지만 각 모델을 Forzen 하고 cross-modal alignment를 하는 것은 어려운 과제로 여겨져왔습니다. (cross-modal alignment : 서로 다른 모달들을 의미상으로 공통된 공간으로 매핑 시키는 것)

이러한 방법으로 예전에는 주로 이미지 데이터를 활용해서 텍스틀 생성하는 방식을 학습함으로 써 이미지와 텍스트 사이의 표현력을 모델이 학습하도록 하였습니다. 하지만 본 논문에서는 이미지 → 텍스트 학습이라는 단 방향이 학습이 이미지와 텍스트의 모달리티 차이를 줄이는데 충분하지 않음을 지적하고 있습니다.

본 논문은 Q-Former (Querying Transformer)로 불리는 작은 transformer를 사용하여 두 모달 사이의 information bottleneck (정보 병목 : 모든 정보를 다 활용하기 보다는 각 모달리티에서 중요한 정보만 사용하도록 유도하는 방법) 을 만들어 모달리티의 차이를 줄이는 방법을 제안하였습니다.



Q-Former를 학습하기 위해서 2단계 (Representation & Generative Learning)를 활용합니다.

(1) Vision-language representation learning : 텍스트와 관련된 이미지의 특징을 추출하도록 Q-Former를 학습합니다.

(2) Vision-to-language generative learing : Image로 학습된 Q-Former의 결과를 LLM에 넣어 텍스트를 생성한 후 손실 값을 활용하여 Q-Former를 학습합니다.

Pre-trained Image Enocder와 Pre-trained LLM의 추가 학습없이, 적은 파라미터를 갖는 Q-Former만을 학습하여 거대 이미지 모델과 거대 언어 모델의 표현력의 차이를 줄여 보다 VQA에서 SOTA급 성능을 보여줄 수 있다고 주장합니다.

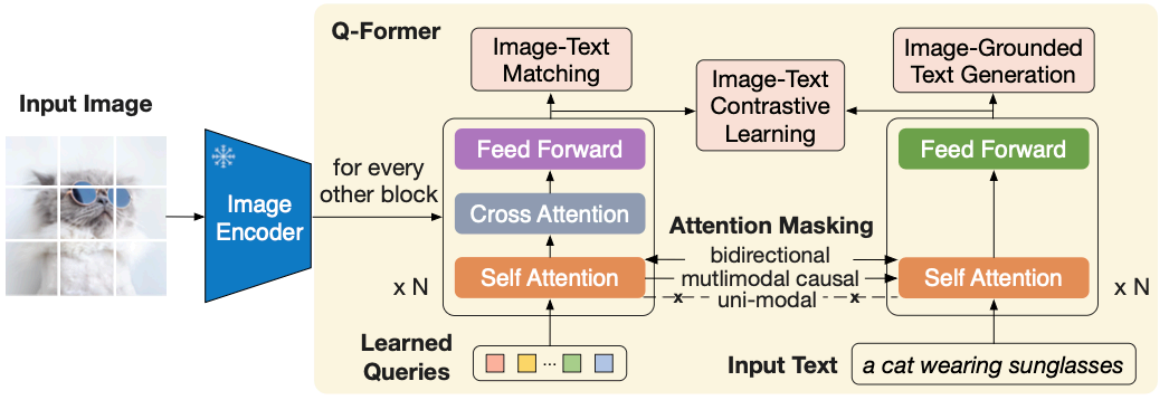
+) VLP 의 경우 주로 추가의 학습 가능한 모듈을 두어서 두 사이의 모달리티 차이를 줄이는 방법이 주로 사용되는 듯 하다. 결국 두 모델은 각각 독립적으로 잘 발전되어 왔기에, 두 사이의 모달리티만 줄이면 높은 성능을 보여주는 모델이 탄생한다는 개념

Method

본 논문에서는 BLIP-2라는 Unimodal (각 독립된 모달리티로 학습된 모델) 을 Q-Former라는 다리역할을 수 행하는 모델로 두 모달리이 사이의 차이를 줄이도록 2단계를 걸쳐 학습하게 됩니다.

우선 (1) 텍스트와 연관된 이미지의 특징을 잡아내고 (2) 이미지정보를 활용하여 텍스트를 생성함으로써 두 모달리티의 차이를 줄입니다.

Modal Architecture



Q-Former를 자세히 보면 다음과 같이 2개의 sub-transformer로 구성되어있습니다.

(1) 번 Trasnformer의 경우 Image Encoder와 연결이 되어 있어있습니다. 그래서 학습가능한 Learned Queries와 가중치가 고정이 된 Image Enocder와 Attention을 통해서 이미지의 특징을 학습하게 됩니다. Learned Quereis의 경우 (경험적으로) 32개를 사용하는데, 이는 Information bottle Neck을 제공합니다. Information bottle Neck을 통해서 이미지의 모든 특징을 사용하기 보다, 정말로 중요한 이미지 32개의 특징만을 활용하게 유도함으로 써 보다 효율적으로 이미지의 특징을 학습하게 됩니다.

(2) 번 Transformer의 경우 text transformer로 사용되게 됩니다. 해당 transformer의 경우 encoder와 decoder 로 사용됩니다.

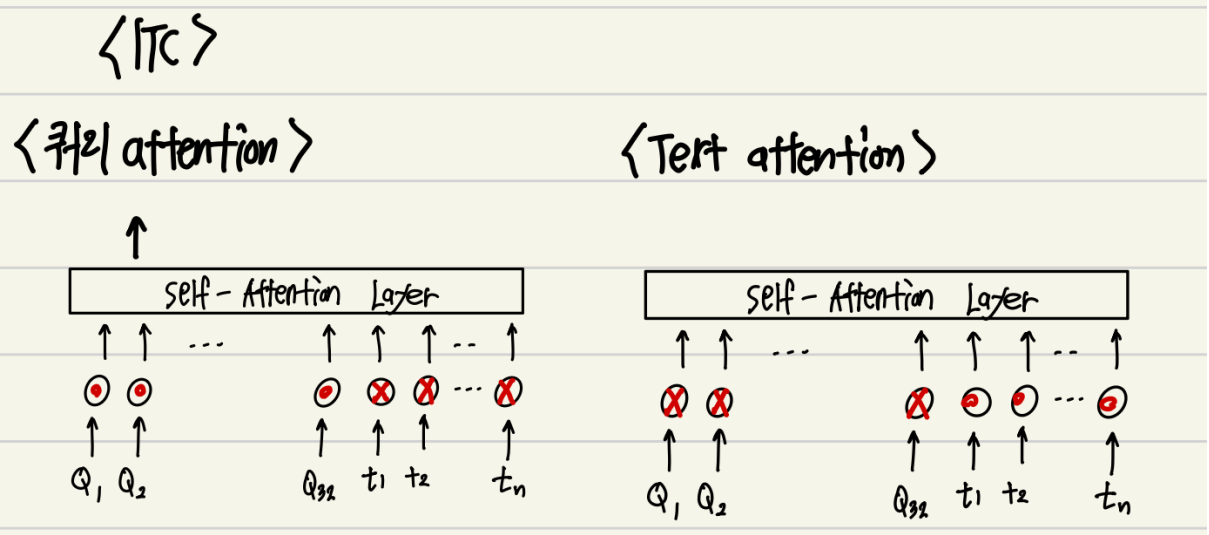
Self-Attention의 경우 두 sub-transformer가 서로 공유하는 Layer입니다. 그래서 각각의 하위 테스크에 대해서 서로 다른 Maksing 전략을 활용해서 학습의 유연하을 추가해줄 수 있습니다. 기본적으로 Q-Former 는 BERT의 사전학습 가중치를 사용하며, Cross Attention의 경우 랜덤 가중치로 초기화 한다고 합니다.

그리고 본 논문의 실험에서는 Learned Queries를 Q,K,V로 사용하는 경우 768차원으로 임베딩을 하게 되고, Image Encoder의 경우 ViT 기반의 모델이라 1024차원으로 임베딩 되어있지만, 간단한 Projection Layer를 활용하여 1024차원을 768차원으로 만들어 사용하게 됩니다. 따라서 Z (Learned Quereis) 의 경우 [32 X 768] 크기를 갖게 됩니다.

Bootstrap Vision-Language Representation Learning form a Frozen Image Encoder [Stage1]

해당 학습에서는 텍스트와 가장 연관이 높은 이미지 특징을 추출하는데 목적을 두고 있습니다. BLIP 논문에서 영감을 받아 동일한 입력과 가중치를 공유하는 ITC, ITG, ITM 총 3가지 학습을 진행합니다. 그리고 각각의 하위 task에 맞게 maksing 전략을 도입하여 보다 효과적으로 3가지 하위 task가 가능하도록 합니다.

(1) Image-Text Contrastive Learning (ITC)



해당 테스크의 경우 이미지 표현과 텍스트 포현이 서로 맞춰지도록 학습하여 양쪽의 상호정보를 최대화하기 위한 테스크입니다. 이를 위해 Positive-pair 와 Negative-pair를 구성하여 비슷한 의미를 갖는 경우 가까워지고, 그 외의 경우 멀어지도록 하는 대조학습을 진행하게 됩니다.

이미지 표현의 경우 (1) sub transformer에서 Learned Queries 와 Imag encoder와 Attention을 해서 나온 32개의 Z 값을 활용하고, 텍스트 표현의 경우 이미지와 매칭되는 텍스트를 Transformer의 입력으로 하여 얻은 최종 [CLS] 토큰을 활용합니다.

이때 [CLS] 토큰은 1개이고, Z 값은 32개 이기에 32개의 Z 중에서 [CLS]와 가장 유사도가 높은 Z만을 활용합니다. 그리고 해당 값이 positive pair인 경우 가깝게 유도하고, 그게 아닌 경우 멀게 유도합니다.

하지만 공유된 Self-Attention Layer를 사용하고 있기에, 만일 이미지와 토큰 정보를 둘다 사용하는 경우, 텍스트와 유사한 이미지 정보를 가져오기 보다, 단순히 텍스트 정보를 복사해서 가져오는 상황을 방지하기 위해서, Queriese를 Z로 만드는 과정 속에서는 텍스트와 관련된 정보들을 모두 Maksing 처리하고, 텍스트의 입력으로 [CLS]를 생성하는 과정에서는 Queriese를 모두 Masking (Unimodal attention mask) 하게 됩니다. 이를 통해서 각각의 모달리티는 독립적으로 학습하고, 학습된 결과를 가지고 대조학습하여, 이미지 특징과 텍스트 특징을 보다 효과적으로 학습할 수 있게 됩니다.

(2) Image-grounded Text Generation (ITG)

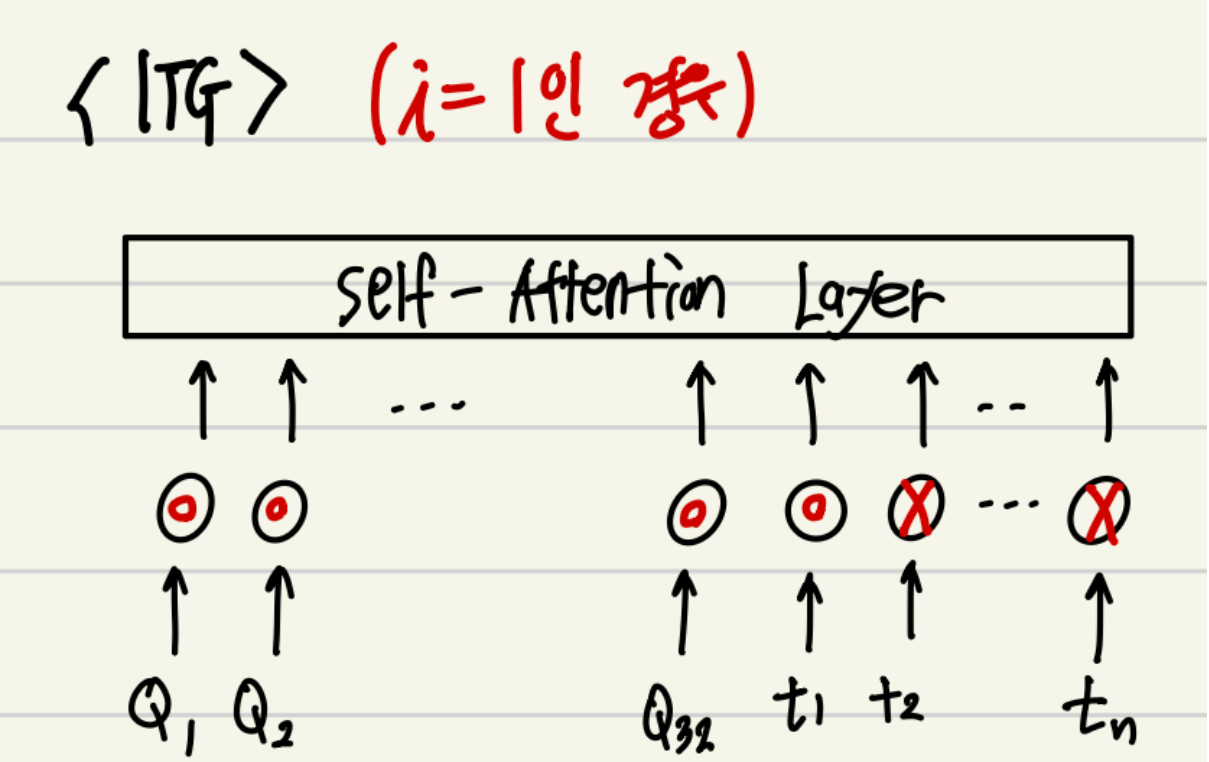
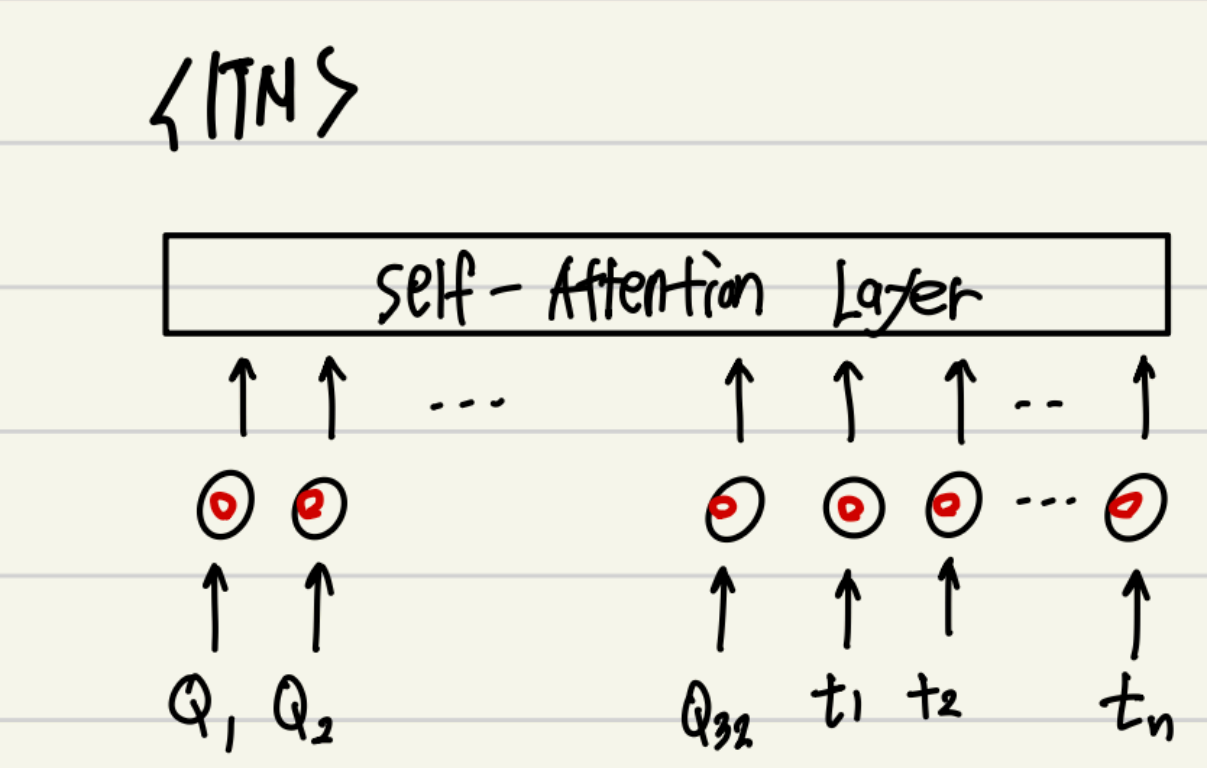


Image 정보가 주어진 경우 단어를 생성하는 방식으로 Q-Former를 학습합니다. 즉, 텍스트를 추출하는 Trasnformer의 경우에는 [DEC] 라는 토큰을 처음 입력으로 받고 ([CLS] 토큰 대신 활용하여 디코딩 테스크의 시작임을 알린다고 합니다.), 모든 쿼리의 정보와, 이전 텍스트 토큰을 모두 활용할 수 있습니다. (입력 : [$Q_1, Q_2, ...Q_{32}, [DEC], t_1, ...t_n$]

하지만 쿼리의 경우 여전히 Image Encoder와 cross attention을 할 뿐, 텍스트와의 상호작용은 하지 않습니다.

그래서 해당 테스크에서는 텍스트가 추가로 이미지 정보를 활용해서 단어를 생성하는 방법을 학습하게 됩니다.

(3) Image-Text Matching (ITM)



Image와 Text 사이의 Fine-Grained Alignment를 학습한다고 합니다. 해당 테스크의 경우 2진 분류 문제를 수행하며, (이미지, 텍스트) 쌍이 서로 매칭이 되는지 아닌지를 예측하도록 합니다. 이때 모든 쿼리와 텍스트는 서로 상호작용 하게 됩니다 (서로 Attention 가능, 마스크 없다). 즉, 해당 과정을 통해서 쿼리로 부터 Z 를 얻게되는데, 이때 Z는 멀티모달의 정보를 갖게 됩니다. Igoti의 경우 32개의 Z 평균을 최종 pair score로 사용한다고 합니다. 그리고 학습시 보다 유용한 정보를 얻기 위해서 Hard Negative Mining (같은 배치 내에 구분이가장 어려운 Smaple을 pair로 삼는 방법) 으로 쌍을 매칭 시킨다고 합니다.

ITC, ITG, ITM 3가지 테스크로 Q-Former를 학습 시킴으로써 Q-Former가 텍스트에서 중요한 이미지 특징을 추출하는 방법을 학습할수 있게됩니다.

Boostrap Vision-to-Language Generative Learning from a Frozen LLM [Stage2]

해당 테스크에서는 LLM 파라미터를 고정하여, 기존의 LLM이 가지고 있던 텍스트 처리에 대한 일반화 성능을 최대한 활용하도록 하였습니다. LLM의 입력으로는 Q-Former에서 나온 Z들이 들어가게 됩니다. 즉, 이미지 정보를 soft-visual-prompt로 활용하는 전략입니다.

Q-Former에서 32개의 쿼리를 할용한 Information bottleNeck을 통해서 불필요한 이미지 정보를 제거하여 LLM 입력시 매우 효율적이라고 주장합니다.

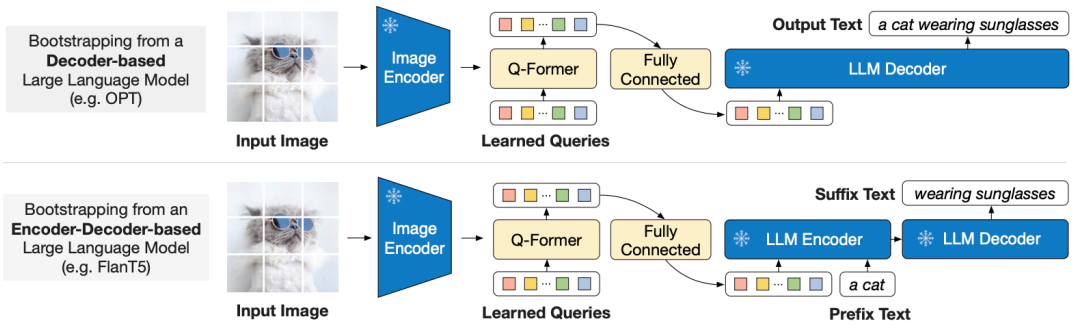


Figure 3. BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

총 2개의 하위 테스크를 통해서 Q-Former를 학습한다고 합니다.

(1) Decoder-based LLM

LLM의 Deocder만을 사용합니다. Q-Former에서 나온 이미지 정보를 1024차원으로 맞춰주고, 해당 이미지 정보를 활용하여 단어를 생성하도록 합니다.

+) 단순 Captioning 능력강화

(2) Encoder-Decoder based LLM

LLM의 Encoder와 Decoder를 모두 활용하며, 문장을 2개의 문장(Prefix + Suffix)으로 분해하고, Encoder의 입력으로 (이미지 정보 + Prefix) 를 활용하여 최종 Decoder가 Suffix 문장을 맞추도록 학습합니다.

+) 질의응답 및 혼합 테스크에 대한 능력 강화

Experiments

Models	#Trainable Params	Open-sourced?	**Visual Question Answering** VQAv2 (test-dev) VQA acc.	**Image Captioning** NoCaps (val) CIDEr	SPICE	**Image-Text Retrieval** Flickr (test) TR@1	IR@1
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-
BLIP-2	188M	✓	**65.0**	**121.6**	**15.8**	**97.6**	**89.7**

Table 1. Overview of BLIP-2 results on various **zero-shot** vision-language tasks. Compared with previous state-of-the-art models. BLIP-2 achieves the highest zero-shot performance while requiring the least number of trainable parameters during vision-language pre-training.

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.

기존의 모델들 보다 더 적은 학습 가능한 파라미터를 활용하여 더 높은 성능을 달성하였음을 보여준다. 추가로 Zero-shot Image, Language 테스크에서 모두 뛰어난 효과를 보여준다고 합니다.

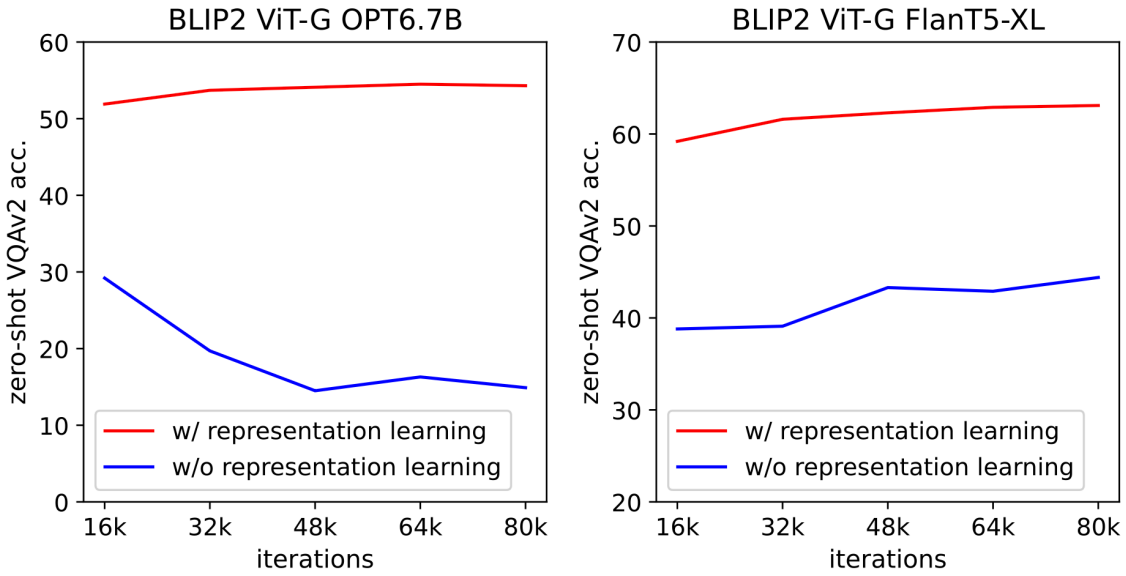


Figure 5. Effect of vision-language representation learning on vision-to-language generative learning. Without representation learning, the Q-Former fails the bridge the modality gap, leading to significantly lower performance on zero-shot VQA.

다음 실험에서는 이미지와 텍스트 모달리의 차이를 줄이기 위해서 텍스트와 유사한 이미지 특징을 추출하는 Representation Learning이 미치는 성능에 대해 시각적으로 보여주고 있습니다.

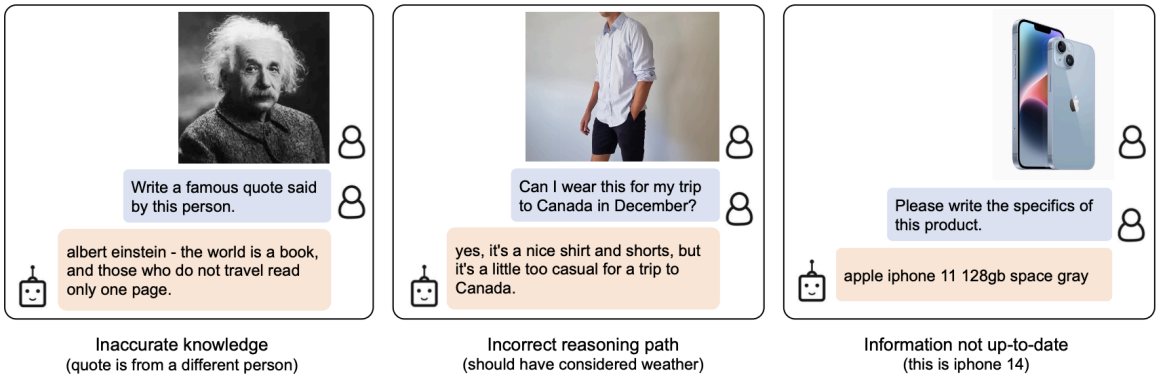


Figure 6. Incorrect output examples for instructed zero-shot image-to-text generation using a BLIP-2 model w/ ViT-g and FlanT5$_{XXL}$.

Conclusion

BLIP-2의 경우 Frozen 된 Image Encoder와 LLM을 활용하여 적은 파라미터를 갖는 Q-Former로 이미지와 텍스트 사이의 모달리티 차이를 줄여주는 모델을 제안하였습니다. 이 모델은 적은 파라미터를 가지고도 zero-shot 에도 충분히 높은 성능을 보여주며, AI 에이전트 발전의 중요한 이정표가 될것입니다.