KUBIG 24-WINTER

COMPUTER VISION STUDY

BLIP Review

KUBIG 17기 임청수

01. Abstract

■ 전체 요약

- vision language pre-training (이하 VLP)가 등장하면서 vision language 분야가 빠르게 성장, 하지만 두 가지 한계점을 지님

한계1) 하지만 기존 모델들은 이해 기반 task나 생성 기반 task에 약한 모습을 보임
한계2) image-text pair dataset은 노이즈가 많이 포함됨

따라서 BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation에서는 두 가지 개선 방안을 제시함

개선1) 언어 이해 Task와 언어 생성 task에 유연하게 적용할 수 있는 새로운 VLP 프레임워크인 BLIP
개선2) 합성 캡션을 생성하고 필터가 노이즈가 있는 캡션을 제거하는 부트스트랩 방식의 CapFilt

02. Introduction

■ 한계1) 모델 관점

대부분의 모델들이 인코더 기반 모델 또는 인코더-디코더 모델인데, 각 유형 별 모델은 Understanding과 Generation 중 하나의 Task만 잘 해결하는 경향이 있다.

- 인코더 기반 모델(CLIP, ALBEF)은 텍스트 생성 task에 적용이 어렵다(ex. image captioning)
- 인코더-디코더 모델(VL-T5, SimVLM)은 이미지-텍스트 검색 task에 적용이 어렵다.(ex. image-text Retrieval)

BLIP 모델은 세 가지로 이루어져 있다.

unimodal encoder : contrastive learning 수행. Image-text understanding improving.

image-grounded text encoder : image-text matching 수행

image-grounded text decoder : image-conditioned language modeling 수행. Generation에 도움

02. Introduction

■ 한계2) 데이터 관점

- 대부분의 모델들은 web에서 크롤링한 image - alt text pair를 사용하여 사전학습한다.
- 많은 데이터를 사용하더라도 사람이 annotation 하지 않는다면 노이즈가 많이 포함되어 있는 web text로 인해 학습에 한계가 존재한다.

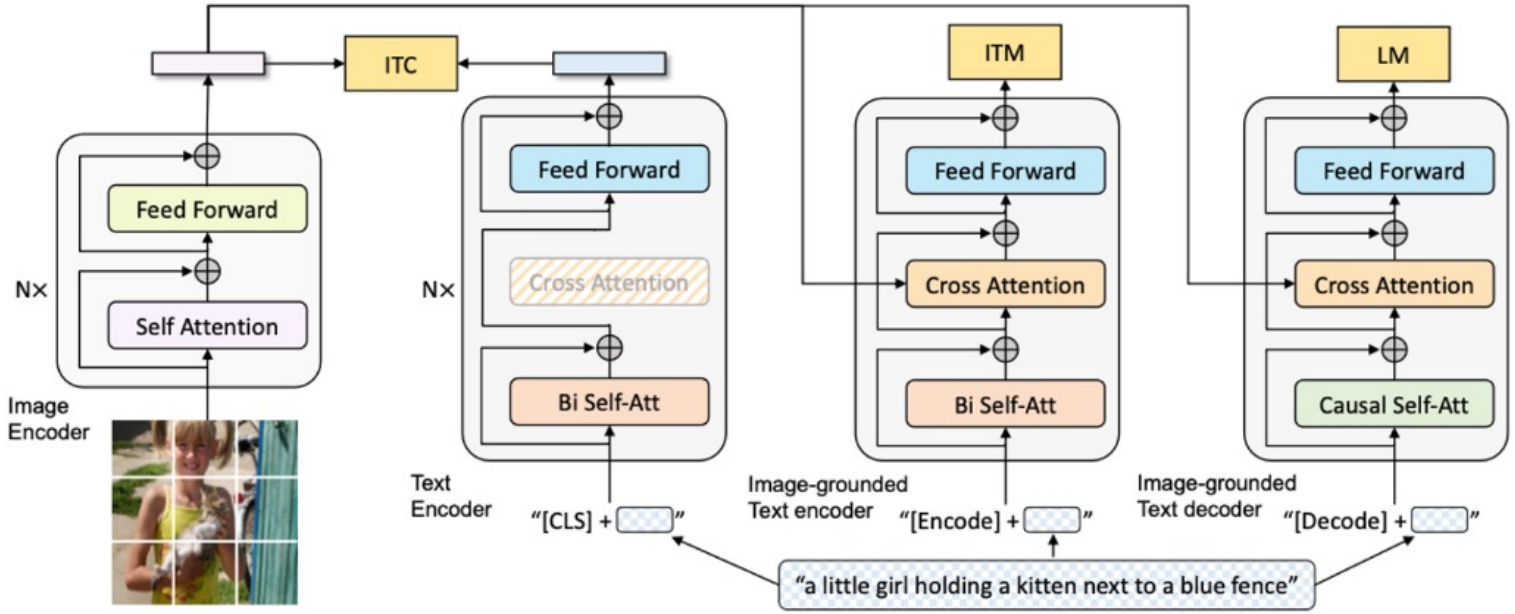
CapFilt는 synthetic model(decoder)인 captioner와 understanding model(encoder)인 filter를 사용하여 original web caption과 synthetic caption 중 noisy한 샘플을 없애는 작업을 수행한다.

Decoder로 caption을 생성하고 encoder로 noisy한 caption을 필터링하여 데이터셋의 퀄리티를 높일 수 있다.

03. Methodolgies

■ 모델 구조

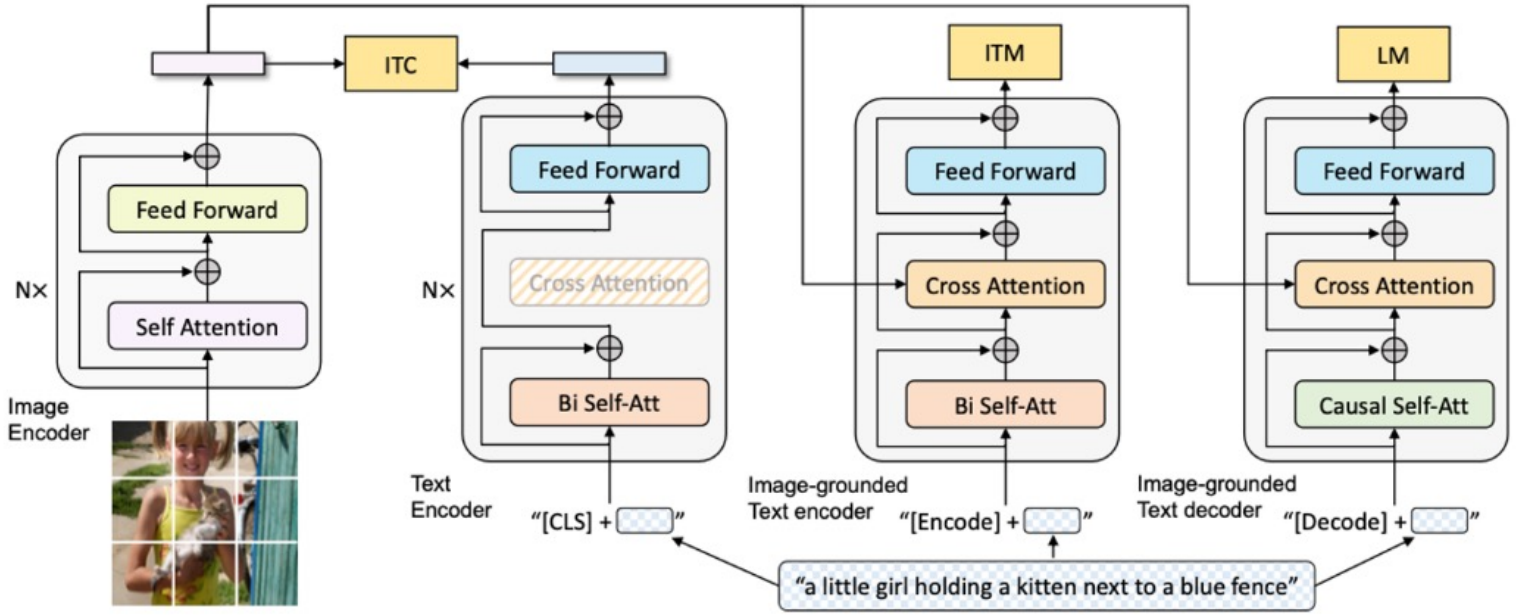
- unimodal encoder, image-grounded text encoder, image-grounded text decoder로 구성
- understanding, generation 기능을 모두 갖춘 unified model을 사전 학습하기 위해 Multi-task 모델인 MED를 제안



03. Methodolgies

■ 모델 구조

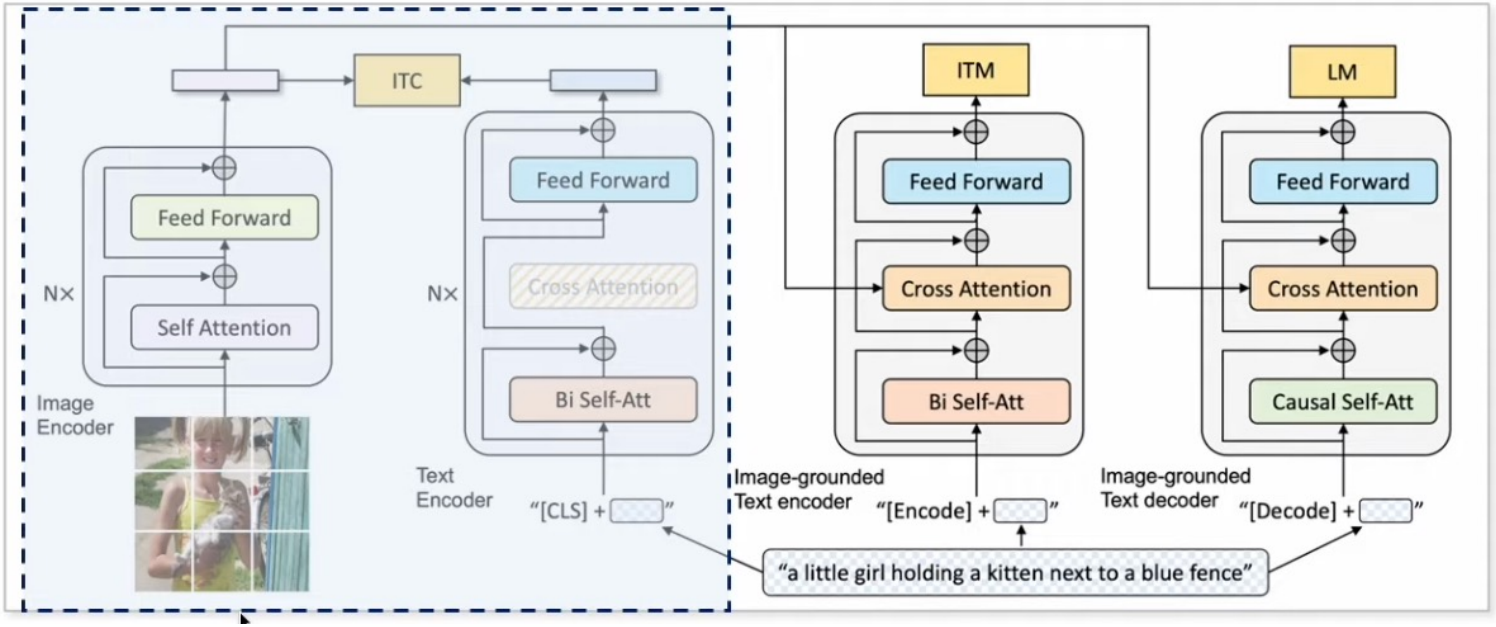
- unimodal encoder, image-grounded text encoder, image-grounded text decoder로 구성
- understanding, generation 기능을 모두 갖춘 unified model을 사전 학습하기 위해 Multi-task 모델인 MED를 제안



03. Methodolgies

■ unimodal encoder

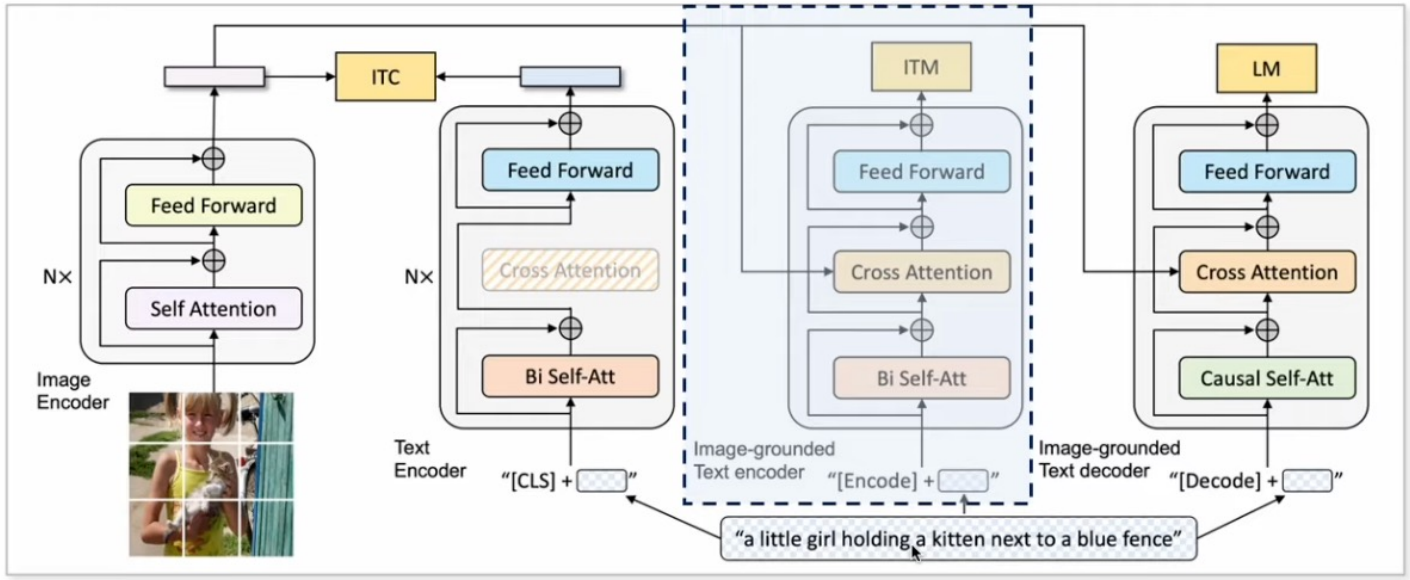
- image encoder는 ViT를 사용. 입력 이미지를 패치로 나누어 임베딩하고, global image feature를 표현하기 위해 [CLS] token 추가
- 사전 학습된 object detector보다 ViT를 사용하는 것이 더 효율적
- text encoder는 BERT 사용. 문장을 임베딩하기 위해 텍스트 입력 시작 부분에 [CLS] 토큰 추가



03. Methodolgies

■ image-grounded text encoder

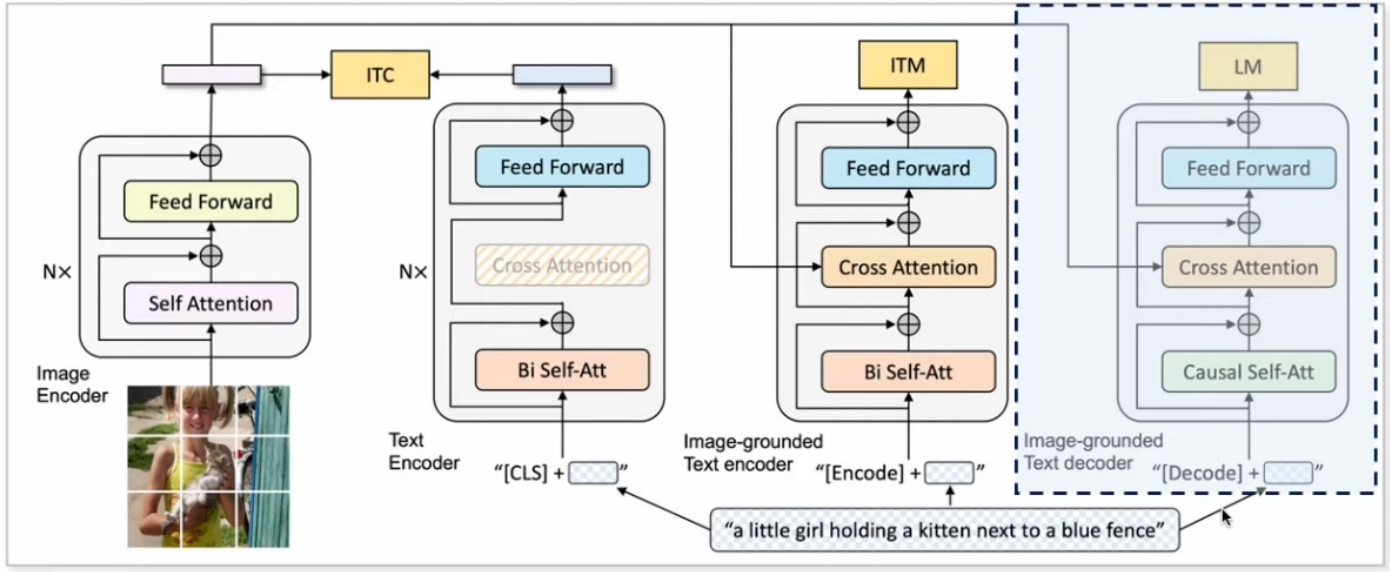
- text encoder와 유사한 구조를 가짐
- 중간에 cross-attention layer를 통해 image embedding 정보를 text encoder에서 활용함으로써 이미지와 **텍스트의 matching 정도를 임베딩으로 출력**



03. Methodolgies

■ image-grounded text decoder

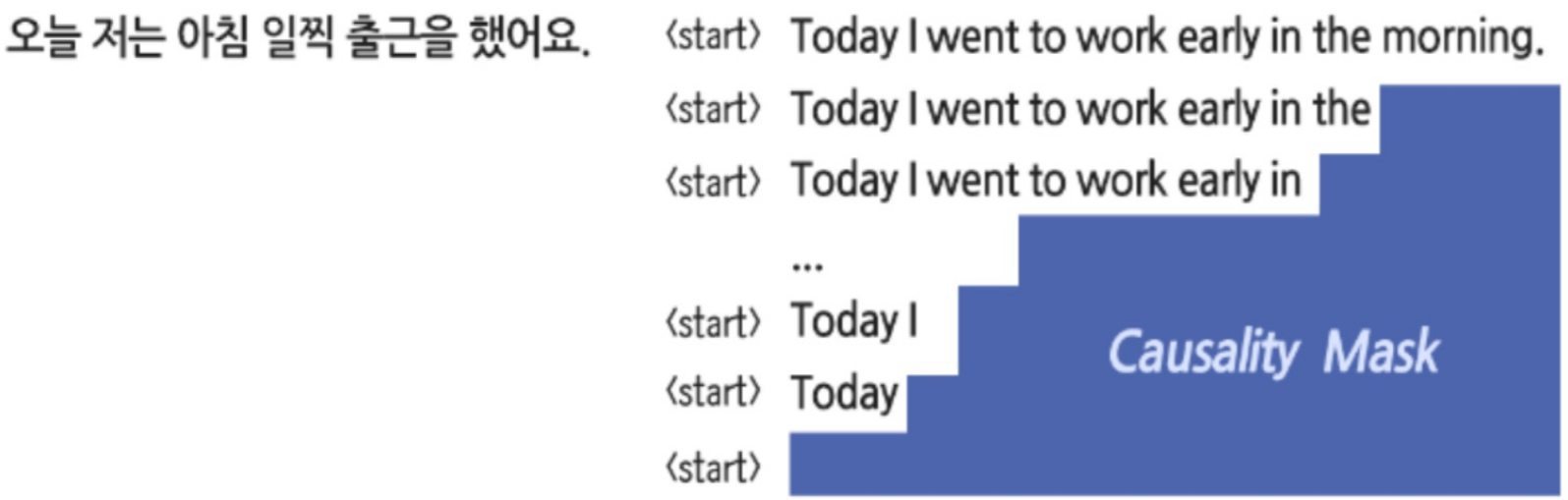
- image-grounded text encoder의 bidirectional self-attention layer를 casual self-attention layer로 대체한 구조
- image를 기반으로 synthesized caption을 augmentation(=dataset bootstrapping) 방법으로 사용하기 위해 설계
- [Decode] 토큰은 시퀀스의 시작을 알리며 [EOS] 토큰은 시퀀스의 종료를 알리는데 사용



03. Methodolgies

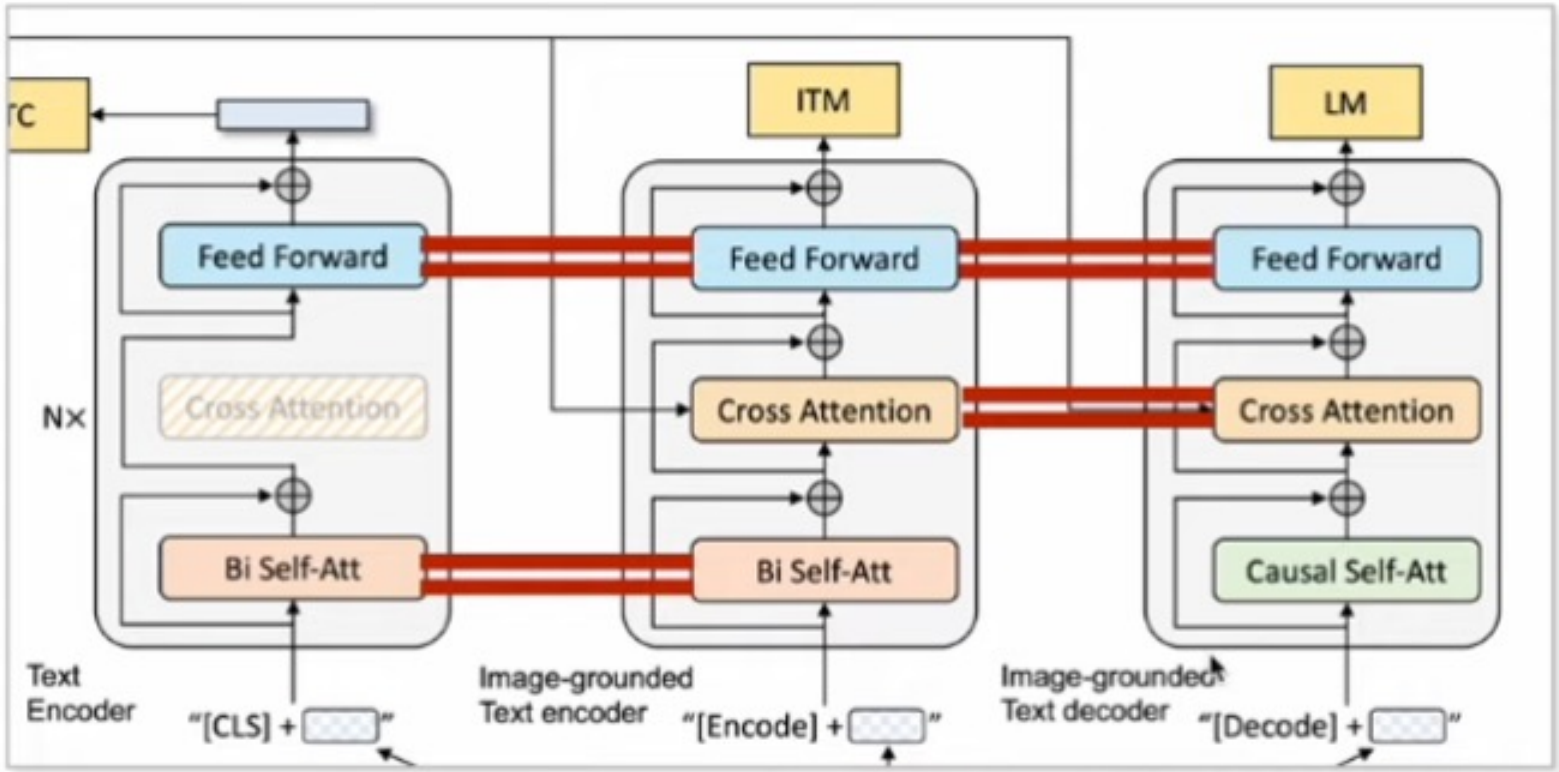
■ casual self-attention

목표하는 문장의 일부를 가려서 인위적으로 연속성을 학습하게 함 → Causality Masking → Autoregressive 문장 생성



03. Methodolgies

■ Parameter sharing



< Parameter Sharing Example >

03. Methodolgies

■ pre training objectives

- 세 가지 objective를 공동으로 최적화. $L = L_{itc} + L_{itm} + L_{lm}$
- 두 가지 understanding based objective (L_{itc}, L_{itm})
- 한 가지 generation based objective (L_{lm})

03. Methodolgies

■ image text contrastive learning loss

- unimodal encoder를 활성화
- visual transformer와 text transformer가 서로 positive pair(image와 text description이 일치하는 경우)라면 가깝게, negative pair(일치하지 않는 모든 경우)라면 멀게 encoding하게끔 unimodal encoder를 학습
- Vision과 text에 대한 understanding을 improving할 수 있는 방법으로, 앞서 설명했던 바와 같이 대부분의 VLP에서 진행하는 contrastive learning 과정
- 이를 통해 ViT와 text transformer의 feature space를 align하는 것을 목표로 함

03. Methodolgies

■ image text matching loss

- image grounded text encoder를 활성화
- binary classification task를 수행
- vision, language 사이의 fine grained alignment를 포착하는 image multimodal representation을 학습
- 이진 분류 작업으로 모델이 ITM Head(linear layer)를 사용하여 image text 쌍이 multimodal 특징을 고려할 때 positive(matched)인지, negative(unmatched)인지 예측
- 단순히 BCE task로 접근하게 되면 contrastive loss와는 다르게 negative sample의 어려움 정도에 따라 loss optimization이 진행될 수 없기 때문에 hard negative mining strategy를 통해 분류가 어려운 negative sample이 학습에 더 많은 관여를 할 수 있게끔 유도
- Hard Negative : 학습 데이터셋에서 서로 연관이 없는 것처럼 보이지만 실제로는 미묘한 관련성을 갖는 이미지 텍스트 쌍을 의미(contrastive similarity가 높은 negative 쌍을 선택해 손실을 계산)

03. Methodolgies

■ Langauge model loss

- image grounded text decoder를 활성화하여 이미지가 주어진 텍스트 설명을 생성하는 것을 목표로 함(기존 VLP에서 MLM과 같이 text 출력을 위해 일반화하는 역할)
- auto regressive 방식으로 텍스트의 likelihood를 최대화하도록 모델을 학습시키는 cross entropy loss 최적화
- 효율적인 pre-training을 위해 text encoder와 decoder는 self-attention을 제외한 모든 매개 변수 공유
- 인코더는 현재 입력된 토큰에 대한 representation을 구축하기 위해 bi-directional self-attention을 사용하는 반면, 디코더는 다음 토큰을 예측하기 위해 causal self-attention 사용(autoregressive한 특성 반영)
    - causal self-attention : mask를 사용하여 현재까지의 시퀀스 텍스트로만 self-attention을 진행.

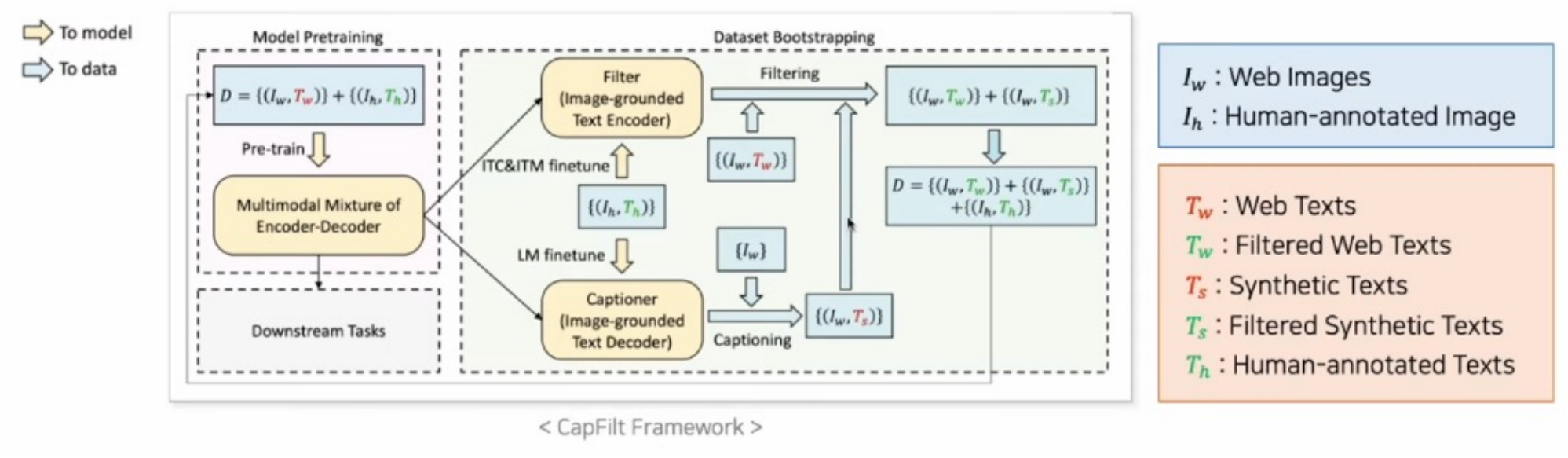
03. Methodolgies

■ CapFilt

- web에서 자동으로 수집된 image-Alt-text pair는 noise가 다수 포함됨

synthetic model(decoder)인 captioner와 understanding model(encoder)인 filter를 사용하여 original web caption과 synthetic caption 중 noisy한 샘플을 없애는 작업을 수행

- captioner, fliter, pretraining으로 나뉨



< CapFilt Framework >

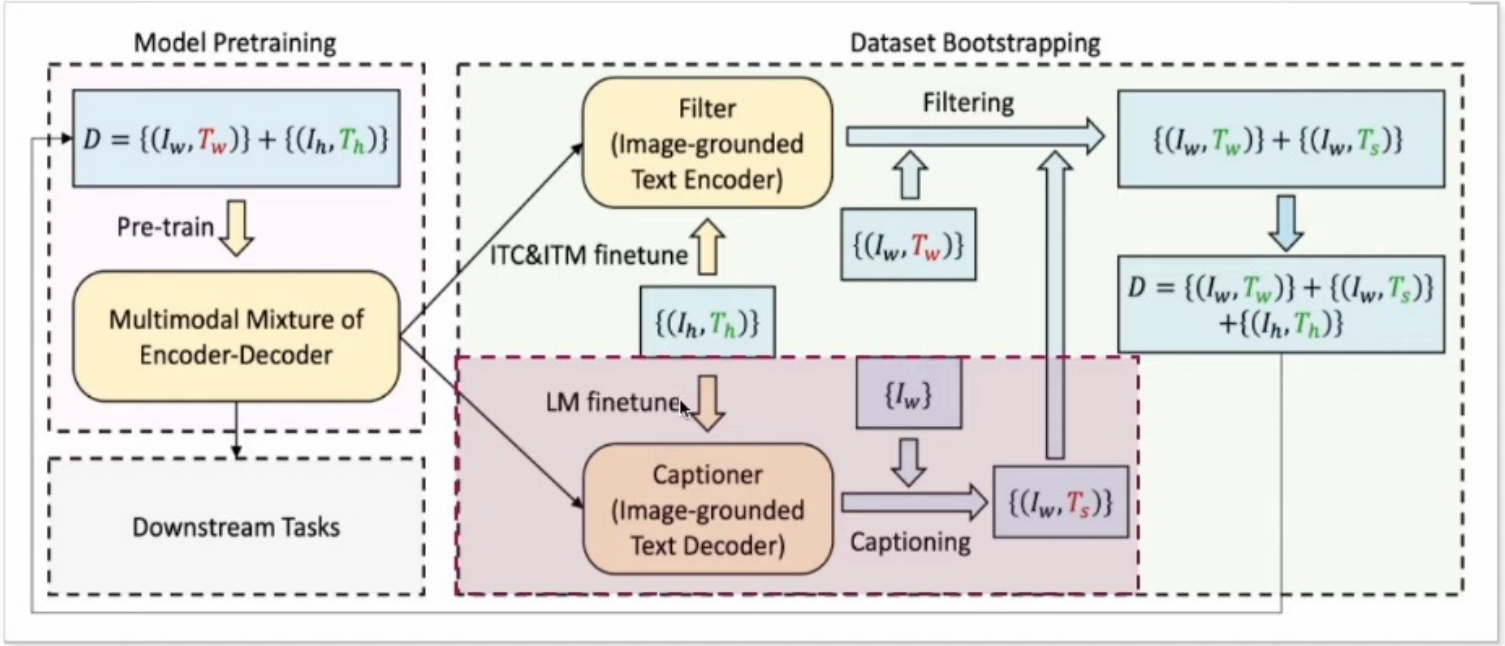
03. Methodolgies

■ captioner (=image-grounded text decoder)

- web에서 자동으로 수집된 image-Alt-text pair는 noise가 다수 포함됨

synthetic model(decoder)인 captioner와 understanding model(encoder)인 filter를 사용하여 original web caption과 synthetic caption 중 noisy한 샘플을 없애는 작업을 수행

- captioner, fliter, pretraining으로 나뉨



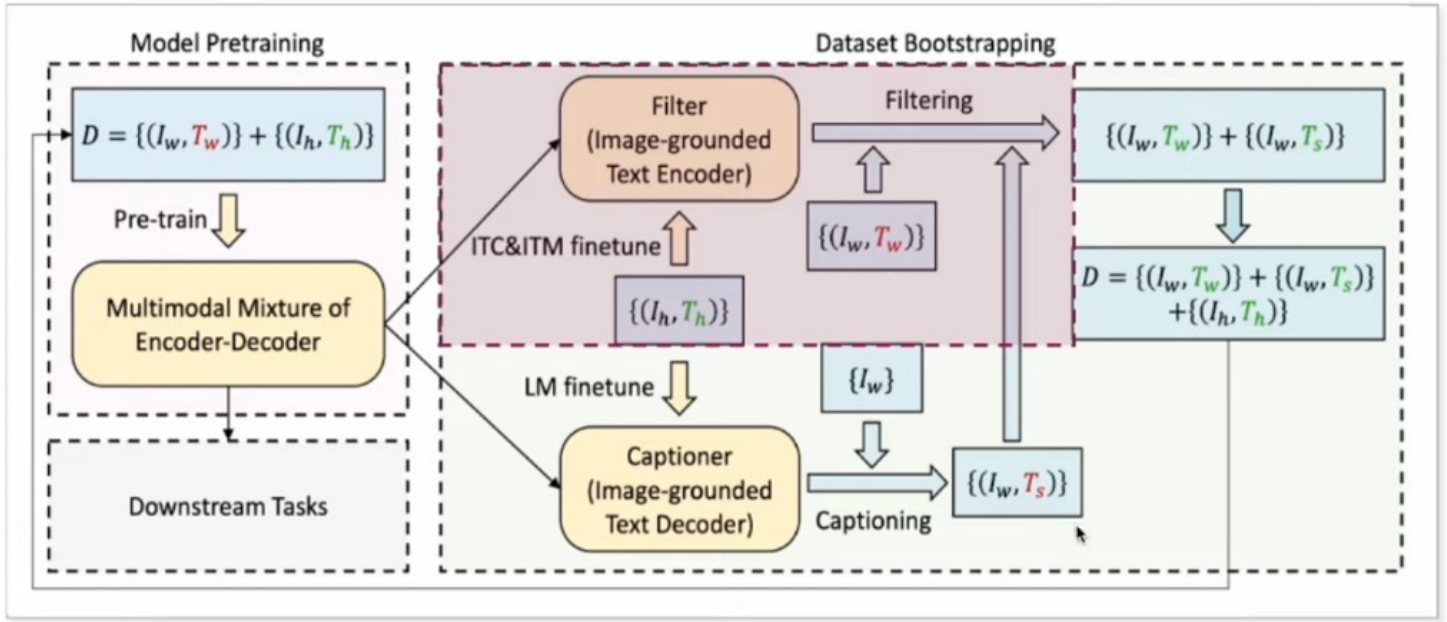
I_w : Web Images
I_h : Human-annotated Image

T_w : Web Texts
T_w : Filtered Web Texts
T_s : Synthetic Texts
T_s : Filtered Synthetic Texts
T_h : Human-annotated Texts

03. Methodolgies

■ fliter (=image grounded text encoder)

- text와 image가 일치하는지 여부를 학습하기 위해 ITC, ITM objective에 따라 세밀하게 fine tuning 진행
- filter는 원본 웹 텍스트 T_w와 합성 텍스트 T_s 모두에서 noise가 있는 텍스트르 제거하는덴 ITM head가 이미지와 일치하지 않는 것으로 예측하는 텍스트는 noise가 있는 것으로 간주됨



< CapFilt Framework >

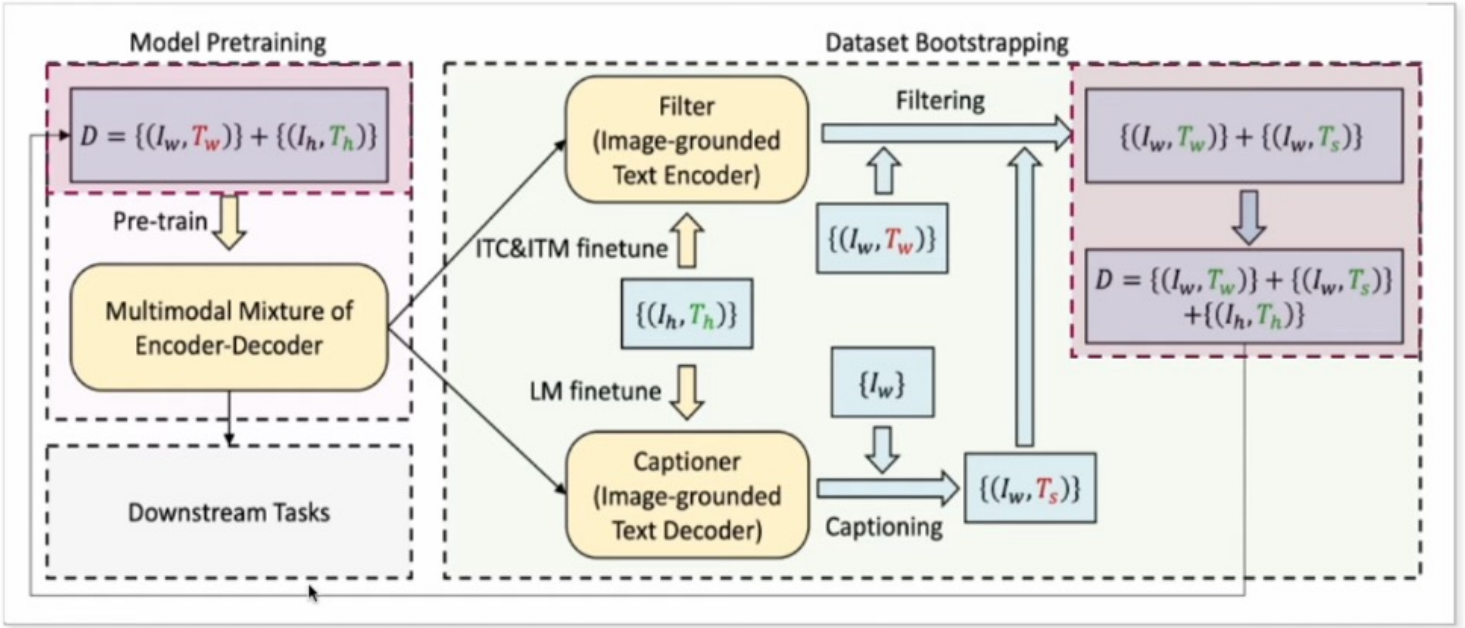
I_w : Web Images
I_h : Human-annotated Image

T_w : Web Texts
T_w : Filtered Web Texts
T_s : Synthetic Texts
T_s : Filtered Synthetic Texts
T_h : Human-annotated Texts

03. Methodolgies

■ pretraining

filtering된 image-text pair를 사람이 주석을 단 쌍과 결합하여 새로운 데이터셋을 형성하고, 새로운 모델을 사전학습하는데 사용



< CapFilt Framework >

I_w : Web Images
I_h : Human-annotated Image

T_w : Web Texts
T_w : Filtered Web Texts
T_s : Synthetic Texts
T_s : Filtered Synthetic Texts
T_h : Human-annotated Texts

04. Experiments

■ Effect of CapFilt

- image text Retrieval, image captioning 등 understanding, generation Task에서 CapFilt의 따른 성능 향상을 보여줌
- 하이라이트처럼 captioner와 filter를 동시에 사용했을 때 서로 보완하는 효과가 있어 성능이 크게 향상됨
- pre train dataset size를 통해 더 큰 데이터셋을 사용할 때 성능이 향상됨을 보여줌

Pre-train dataset	Bootstrap C	Bootstrap F	Vision backbone	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
COCO+VG +CC+SBU (14M imgs)	✗	✗	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✗	✓$_B$		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	✓$_B$	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	✓$_B$	✓$_B$		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	✗	✗	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	✓$_B$	✓$_B$		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	✓$_L$	✓$_L$		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	ViT-L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	✓$_L$	✓$_L$		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

Table 1. Evaluation of the effect of the captioner (C) and filter (F) for dataset bootstrapping. Downstream tasks include image-text retrieval and image captioning with finetuning (FT) and zero-shot (ZS) settings. TR / IR@1: recall@1 for text retrieval / image retrieval. ✓$_{B/L}$: captioner or filter uses ViT-B / ViT-L as vision backbone.

04. Experiments

■ Effect of CapFilt

- captioner와 filter의 예시
  - green이 accept된 text를 의미함
  - red는 reject된 text를 의미함

Figure 4. Examples of the web text T_w and the synthetic text T_s. Green texts are accepted by the filter, whereas red texts are rejected.

04. Experiments

■ Diversity is Key for Synthetic Captions

- CapFilt에서 Synthetic Caption을 생성하기 위해 Nucleous Sampling을 사용함
- Nucleous Sampling : “Top p sampling”이라고도 불리며 단어의 확률분포에서 상위 p%에 해당하는 단어만 고려해 텍스트를 생성하는 기법(p=0.9로 사용)
- 단순히 가장 높은 확률로 캡션을 생성하는 Beam Search와 성능 비교 시 더 우수한 모습을 보임
- Nucleous Sampling가 노이즈는 더 많지만 더 나은 성능을 보이는 이유는 Diverse, Surprising Caption을 만들어주기 때문.

Generation method	Noise ratio	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
None	N.A.	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
Beam	19%	79.6	61.9	94.1	83.1	38.4	128.9	103.5	14.2
Nucleus	25%	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

Table 2. Comparison between beam search and nucleus sampling for synthetic caption generation. Models are pre-trained on 14M images.

04. Experiments

■ Parameter Sharing and Decoupling

- BLIP은 self-attention 제외하고 layer parameter를 공유했고 이때 성능도 향상되면서 파라미터 수가 줄어들어 효율성이 증대되었음
- SA까지 공유한 경우 인코더와 디코더의 Task가 충돌하여 성능이 하락함

Layers shared	#parameters	Retrieval-FT (COCO)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
		TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	SPICE
All	224M	77.3	59.5	93.1	81.0	37.2	125.9	100.9	13.1
All except CA	252M	77.5	59.9	93.1	81.3	37.4	126.1	101.2	13.1
All except SA	252M	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
None	361M	78.3	60.5	93.6	81.9	37.8	127.4	101.8	13.9

Table 3. Comparison between different parameter sharing strategies for the text encoder and decoder during pre-training.

04. Experiments

■ Comparison with State-of-the-arts

Image-Text Retrieval, Image Captioning, Visual Question Answering (VQA), Natural Language Visual Reasoning (NLVR2), Visual Dialog (VisDial), Zero-shot Transfer to Video-Language Tasks 등 sota 달성

05. Summary

BLIP

- BLIP의 MED 구조 - 멀티모달 수행을 위한 Mixture 구조
  - Unimodal Encoder는 L_{itc}를 사용해 image와 text 사이의 관계성 학습(contrastive learning)
  - Image grounded text encoder는 L_{itm}을 사용해 image-text pair가 positive인지 negative인지 예측
  - image grounded text Decoder는 L_{lm}을 사용해 이미지가 주어진 텍스트의 설명을 생성

- BLIP의 CapFilter
  - web에서 수집되는 데이터셋인 Alt—text의 Noise를 제거하기 위함
  - captioner는 이미지 당 하나의 synthetic caption 생성
  - filter는 원본 텍스트나 합성 텍스트 모두에서 noise가 있는 텍스트 제거
  - filtering된 데이터와 사람이 라벨링한 데이터를 합하여 새로운 모델 학습
- 한계
  - 다양한 Task에 높은 성능으르 보이는 하나의 End-to-End 구조를 만드는 과정이 매우 복잡함. 따라서 BLIP-2에서 개선