BLIP (2022)

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

우다연

dyeonwu@gmail.com

TNT Vision

2024/03/12

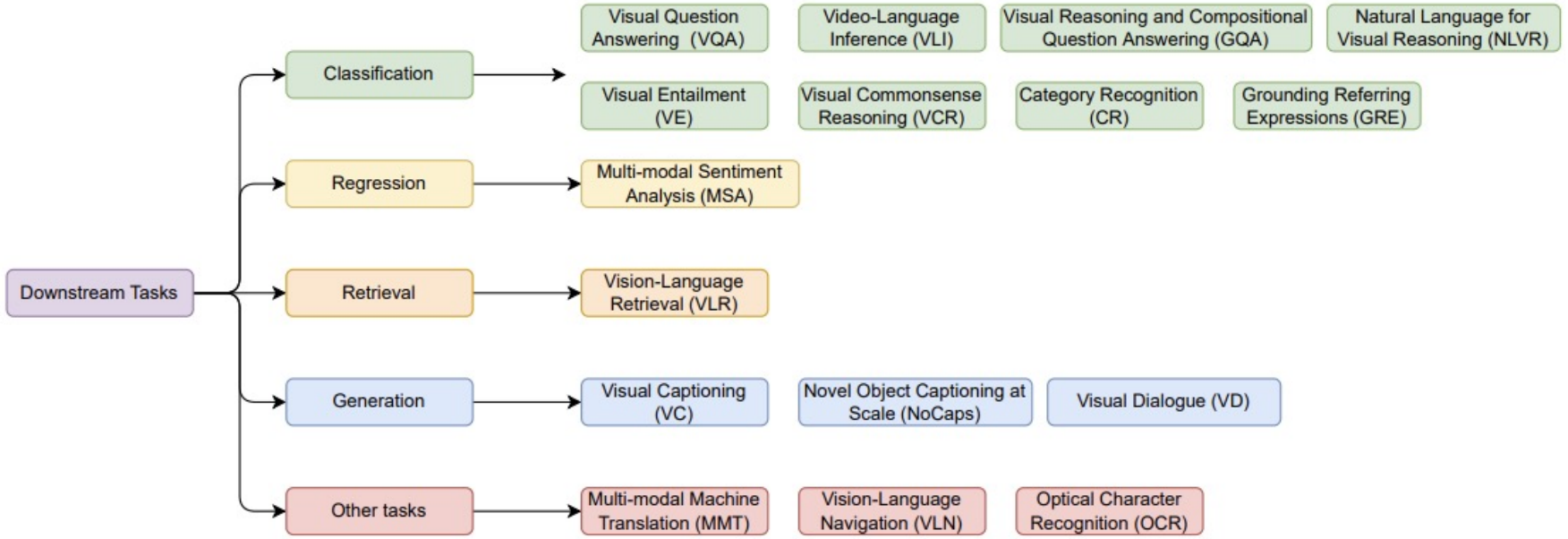
TRAIN AND TEST

Contents

TRAIN AND TEST

Introduction

VLP(Vision-Language Pre-training)



- 모델이 이미지와 관련된 텍스트 정보를 동시에 이해할 수 있도록 만드는 방법론
- Transformer 아키텍처를 기반으로 하는 경우가 일반적
- 대표적인 모델: CLIP, VisualBERT, DALL-E

TRAIN AND TEST

Introduction

기존 VLP 연구의 한계

1) Model Perspectives

- 대부분 Encoder-based / Encoder-Decoder 구조의 모델
- Encoder-based model: Text generation tasks (예: image captioning)에 약함
- Encoder-Decoder model: Understanding-based tasks (예: Image-text retrieval tasks)에 약함

2) Data Perspectives

- 당시 SOTA 모델들은 web에서 수집된 image-text pairs로 Pre-train 되었음 (CLIP, ALBEF 등)
- 정제 작업으로 성능 향상을 이루어내었지만, 노이즈의 양을 고려하면 Optimal한 데이터셋 X

=> BLIP: 기존의 한계를 극복하고 더 다양한 Downstream task를 수행하는 프레임워크

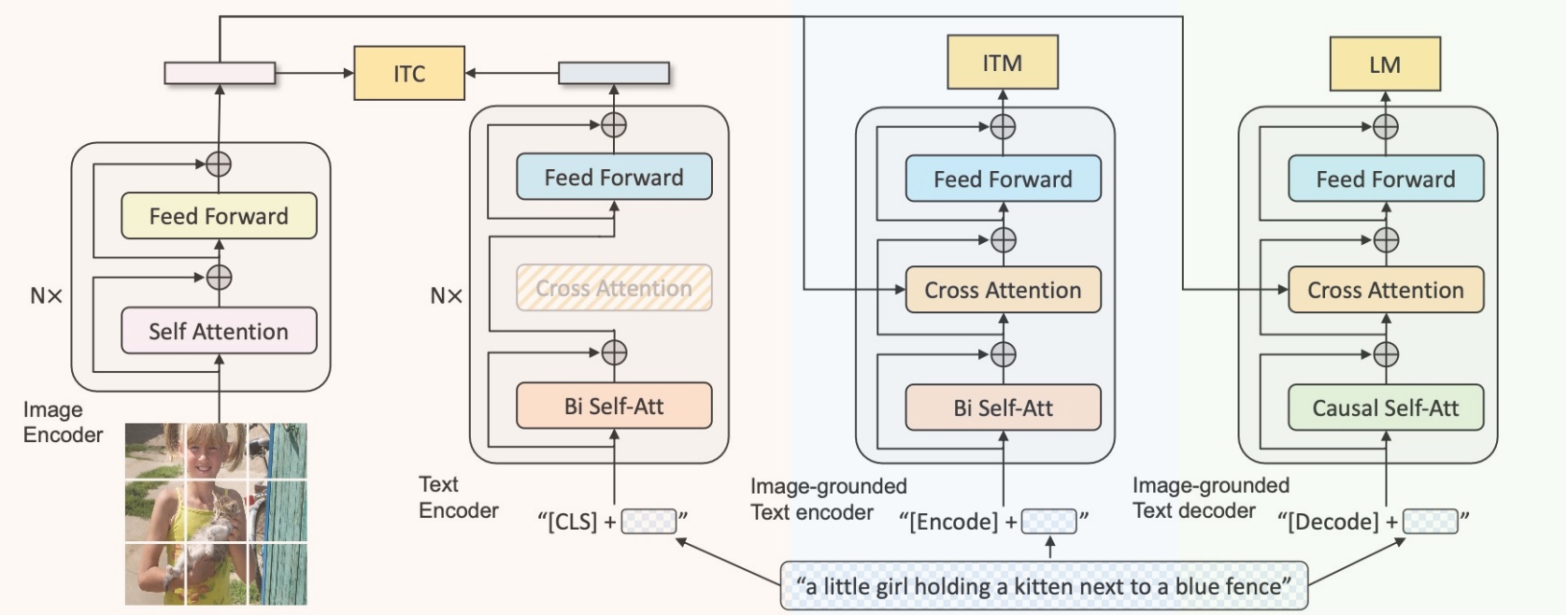
TRAIN AND TEST

Introduction

BLIP

1) Multimodal mixture of Encoder-Decoder (MED)

- 원활한 Pre-training & Transfer learning을 위한 모델 아키텍처
- 3가지 vision-language objectives에 따라 Pre-train 됨

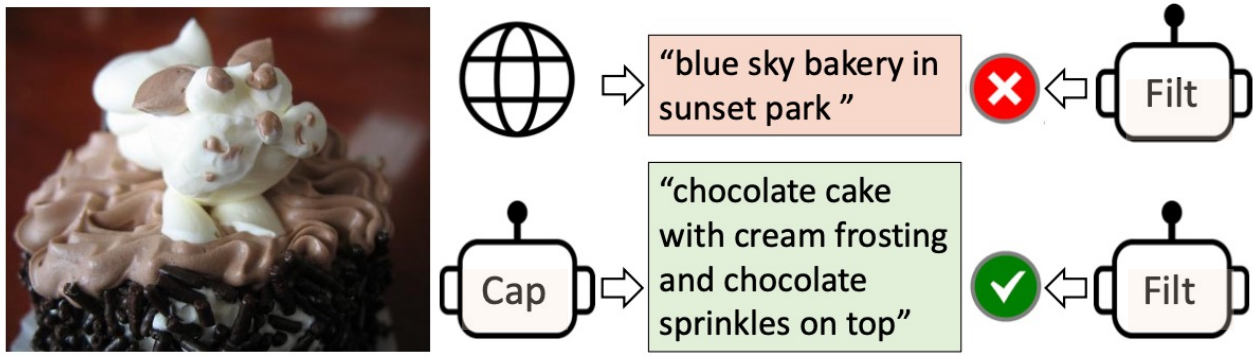


Introduction

BLIP

2) Captioning and Filtering (CapFilt)

- Noisy image-text pair 데이터셋을 학습시키기 위한 bootstrapping 방법
- Pre-trained MED를 2가지 모듈로 Fine-tuning



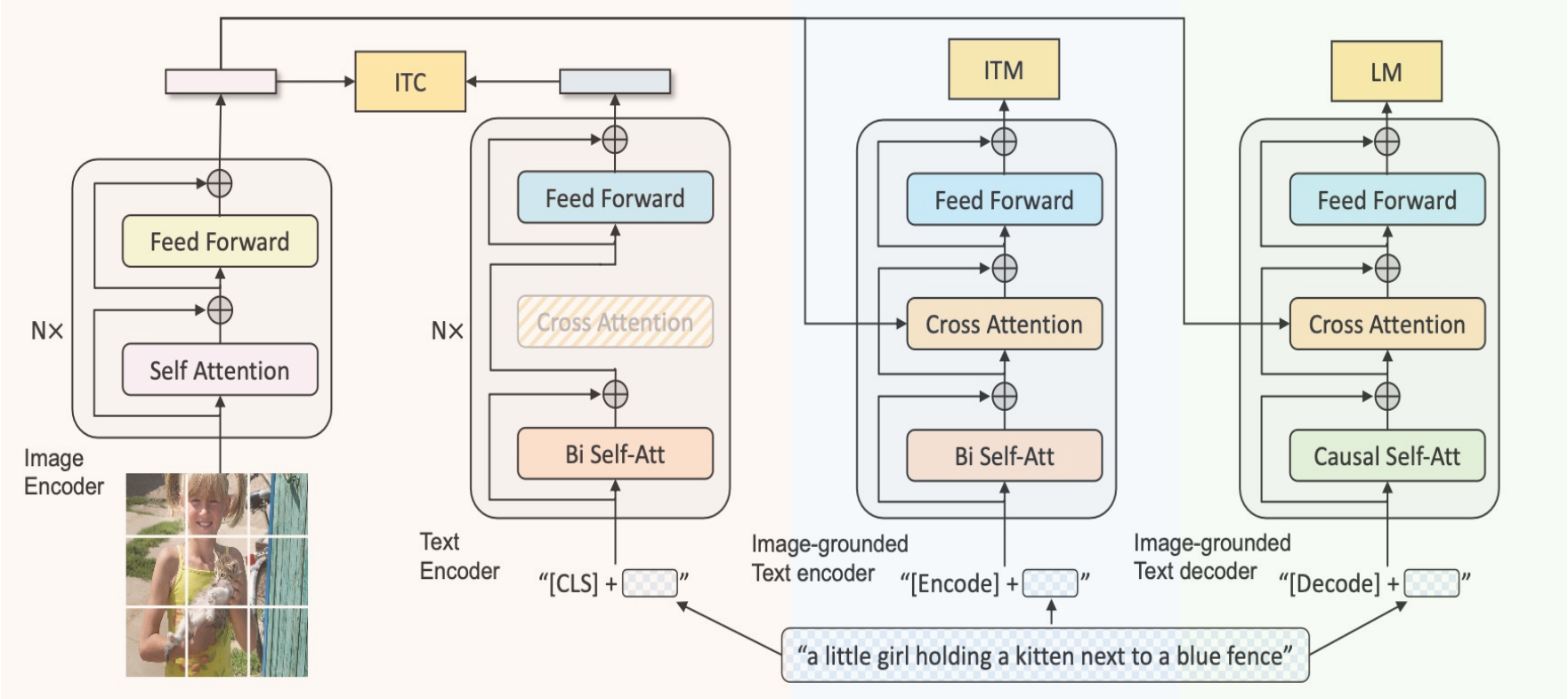
- Captioner: 주어진 web 이미지에 대한 synthetic caption 생성
- Filter: 생성한 caption과 수집된 web text 중 적절하지 않은 것 필터링

TRAIN AND TEST

Method

Model Architecture

Understanding, Generation 모두 우수하게 수행하는 Multi-task 구조



Unimodal Encoder

Image-grounded Text Encoder

Image-grounded Text Decoder

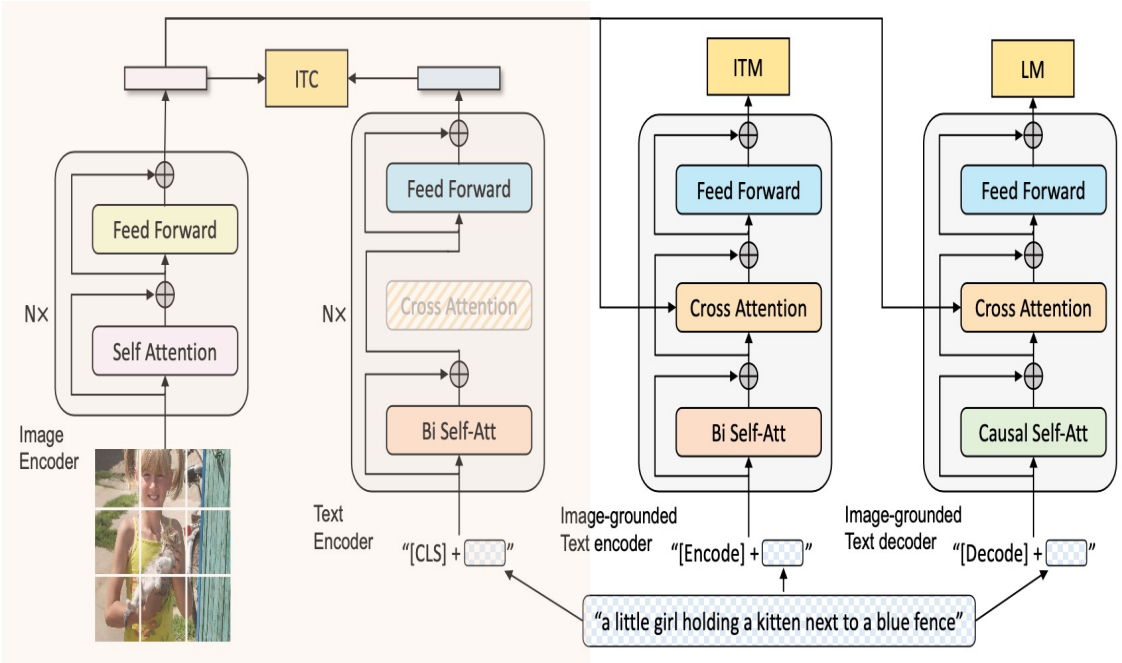
TRAIN AND TEST

Method

Model Architecture

1) Unimodal Encoder

- Image와 Text를 개별적으로 Encoding
- Image encoder
  - ViT 사용
  - Input image를 patch들로 나누어 임베딩
  - [CLS] token: global image feature 나타냄
- Text encoder
  - BERT와 동일
  - [CLS] token: Input text를 요약하는 역할

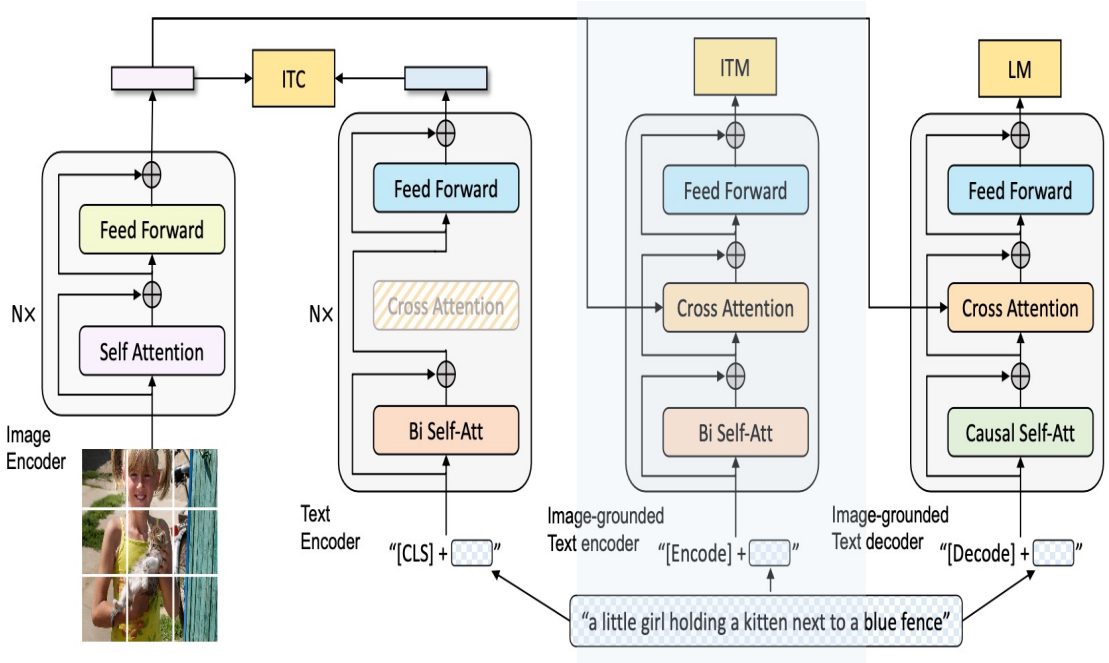


Method

Model Architecture

2) Image-grounded Text Encoder

- Cross-Attention layer: visual information 주입
- [Encode] token: 이미지와 텍스트 간의 상호작용으로부터 얻은 중요한 정보를 집약 representation 생성에 활용



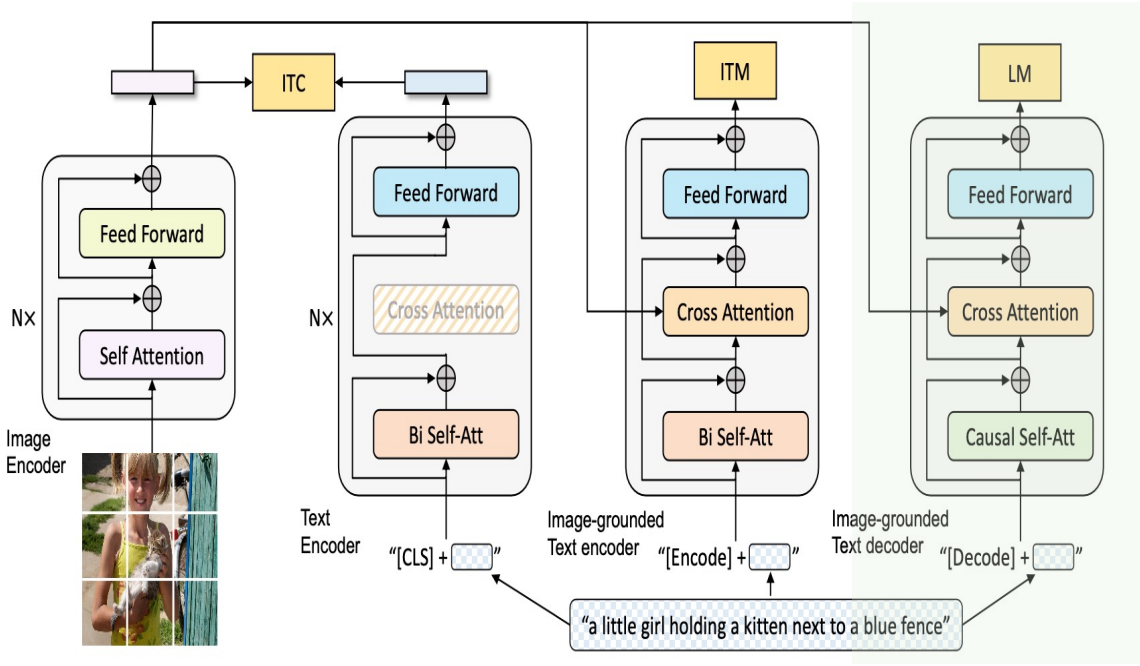
TRAIN AND TEST

Method

Model Architecture

3) Image-grounded Text Decoder

- Bi Self-Attention 대신 Casual Self-Attention 수행
- [Decode] token: Sequence의 시작을 의미
- [EOS] token: Sequence의 끝을 의미

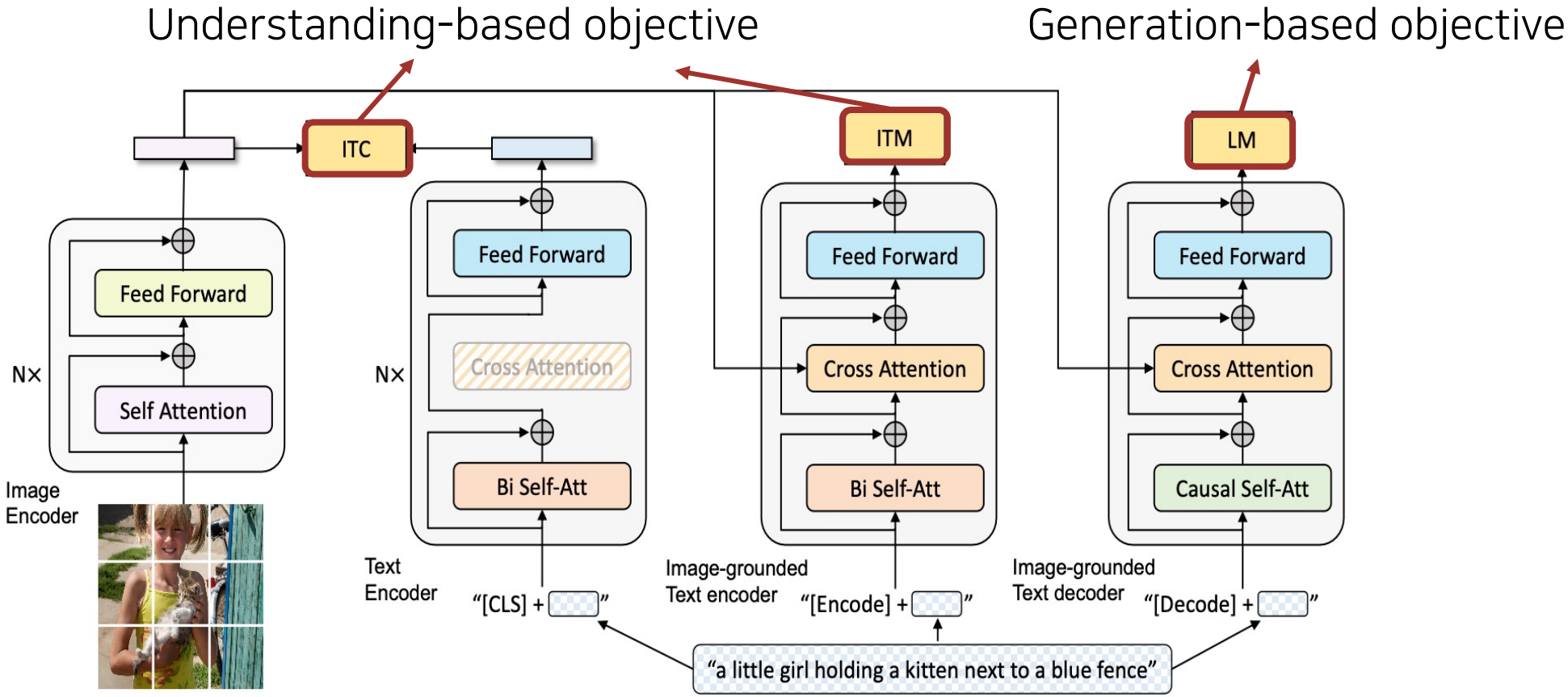


TRAIN AND TEST

Method

Pre-training Objectives

$$L = L_{itc} + L_{itm} + L_{lm}$$



TRAIN AND TEST

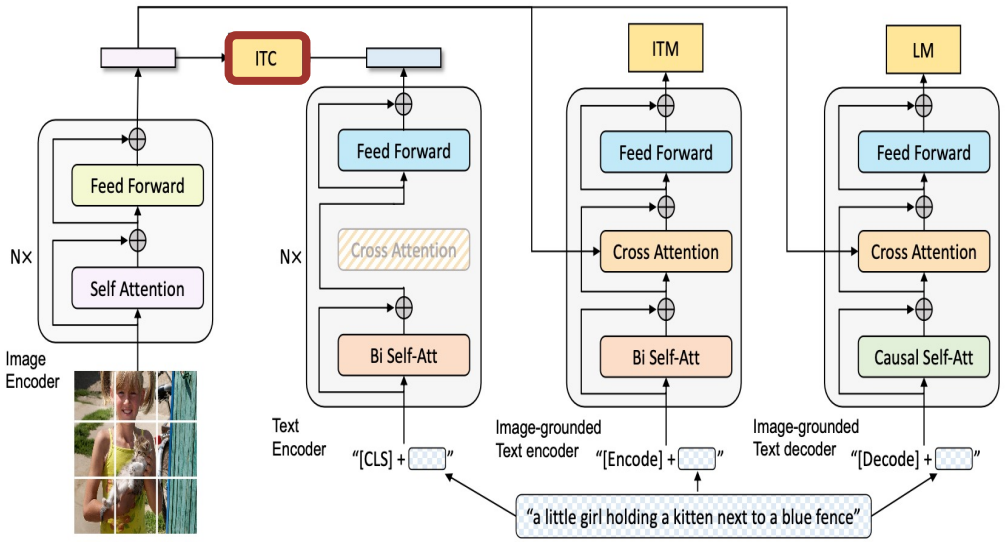
Method

Pre-training Objectives

1) Image-Text Contrastive Loss (ITC)

- Unimodal Encoder 활성화
- 매치되는 image-text pair가 유사한 representation 값을 갖도록 함으로써 Visual transformer & Text transformer의 feature space를 align 시킴
- ALBEF의 ITC Loss 활용
  - Momentum Encoder 도입: Feature 생성

=> Vision & Language 이해 성능 높임



TRAIN AND TEST

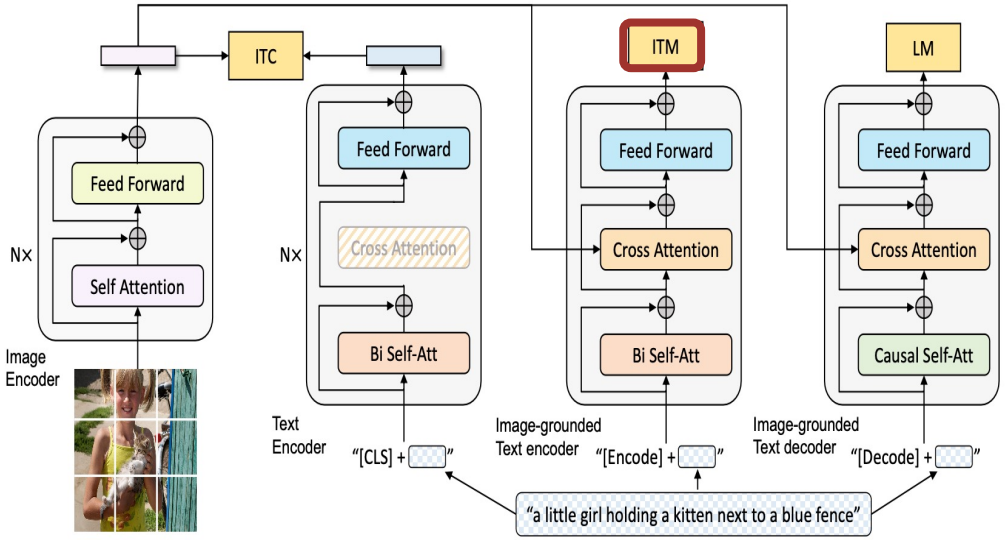
Method

Pre-training Objectives

2) Image-Text Matching Loss (ITM)

- Image-grounded Text Encoder 활성화
- ITM head를 사용하여 image-text pair가 매치하는지 아닌지 예측하는 Binary classification task

=> Image와 Text간의 일치성을 판단함

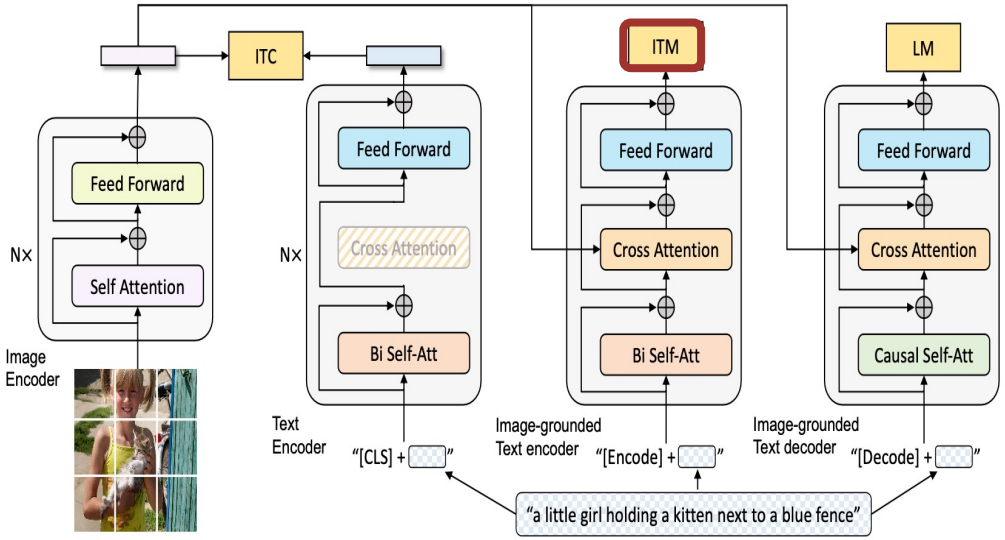


Method

Pre-training Objectives 3) Language Modeling Loss (LM)

- Image-grounded Text Decoder 활성화
- 주어진 image에 대한 text description 생성
- 텍스트의 Likelihood 최대화하는 방식으로 Cross Entropy Loss 최적화

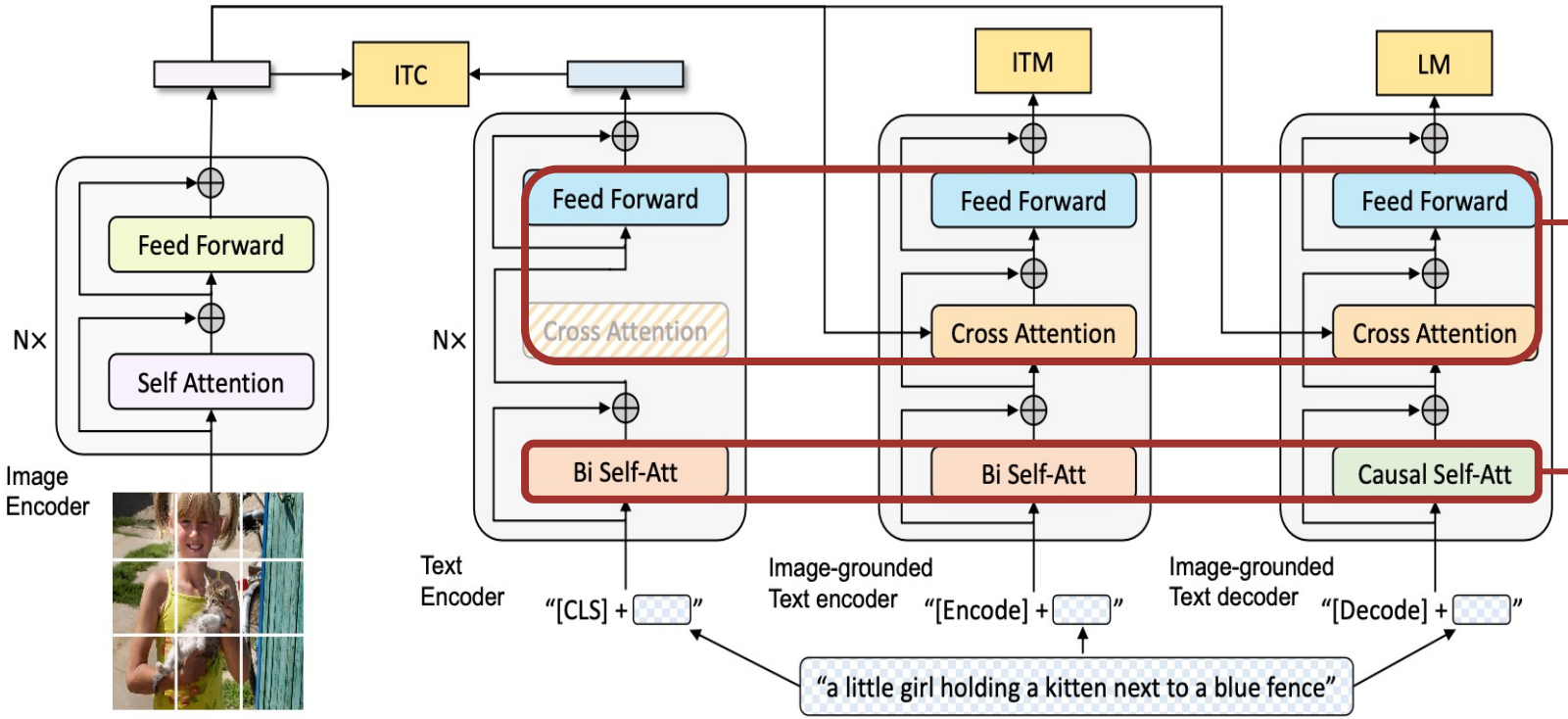
=> 주어진 image에 대한 caption 생성



Method

Parameter sharing

4개의 아키텍처가 융합된 복잡한 구조 -> 효율을 위해 parameter 공유함



Embedding,
CA, FFN:
파라미터 공유

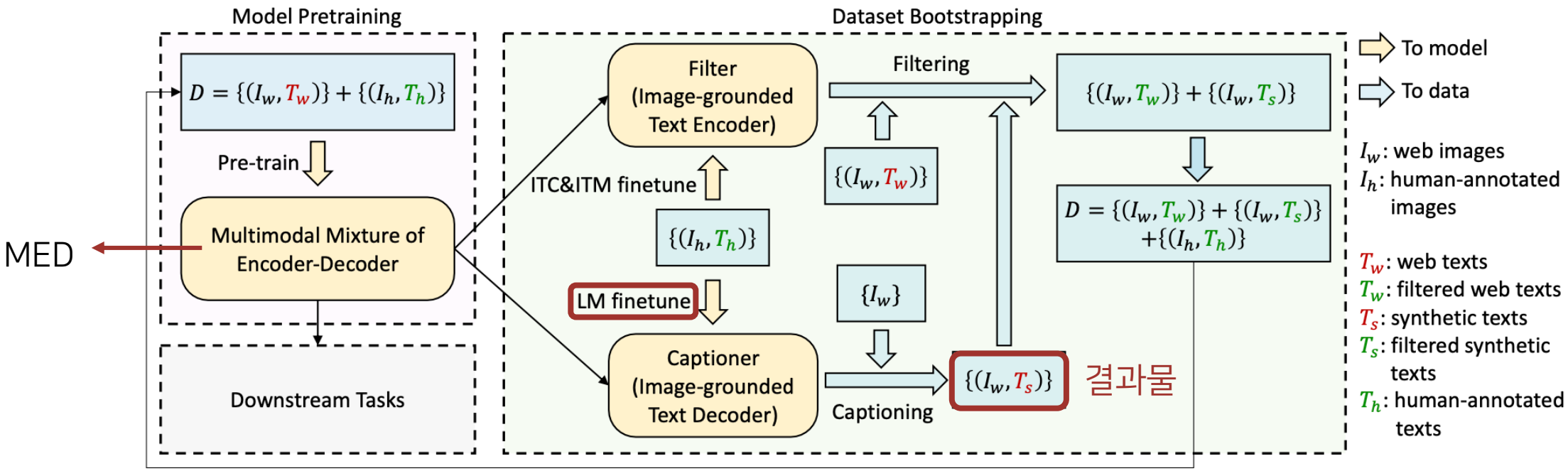
SA layers:
개별 파라미터 사용
(역할과 목적이 상이함,
이해 vs 생성)

TRAIN AND TEST

Method

CapFlit

1) Captioner



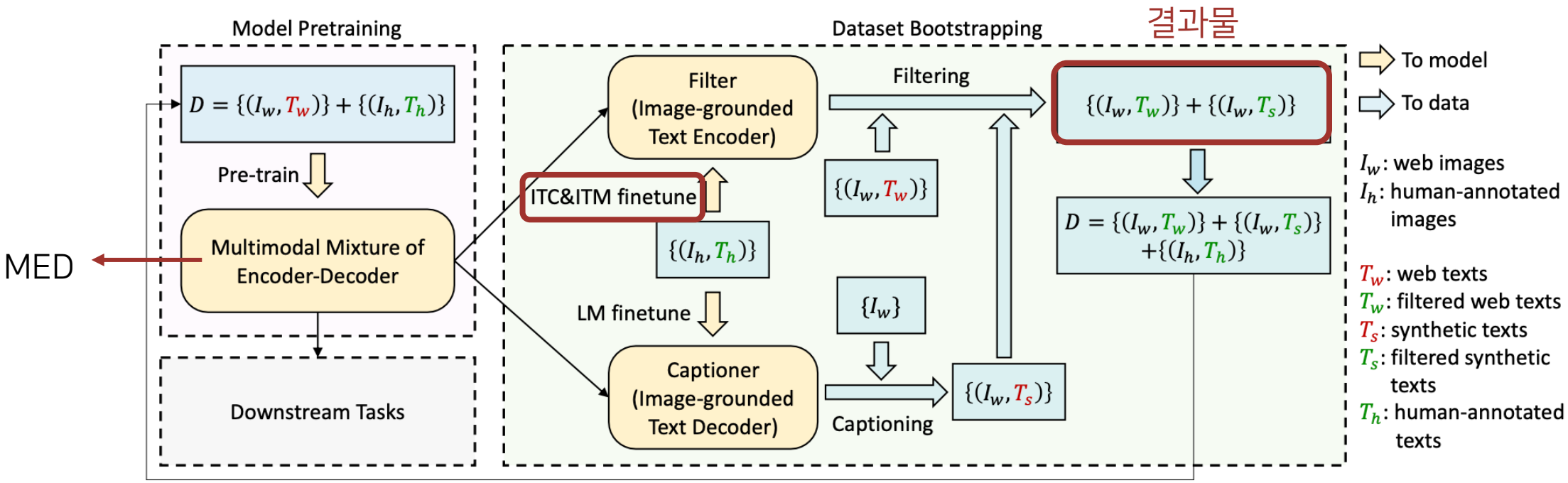
- Image-grounded Text Decoder & LM objective에 따라 Fine-tuning
- Human annotated (High quality dataset): noise가 없어 Fine-tuning 과정에 활용됨

TRAIN AND TEST

Method

CapFlit

2) Filter

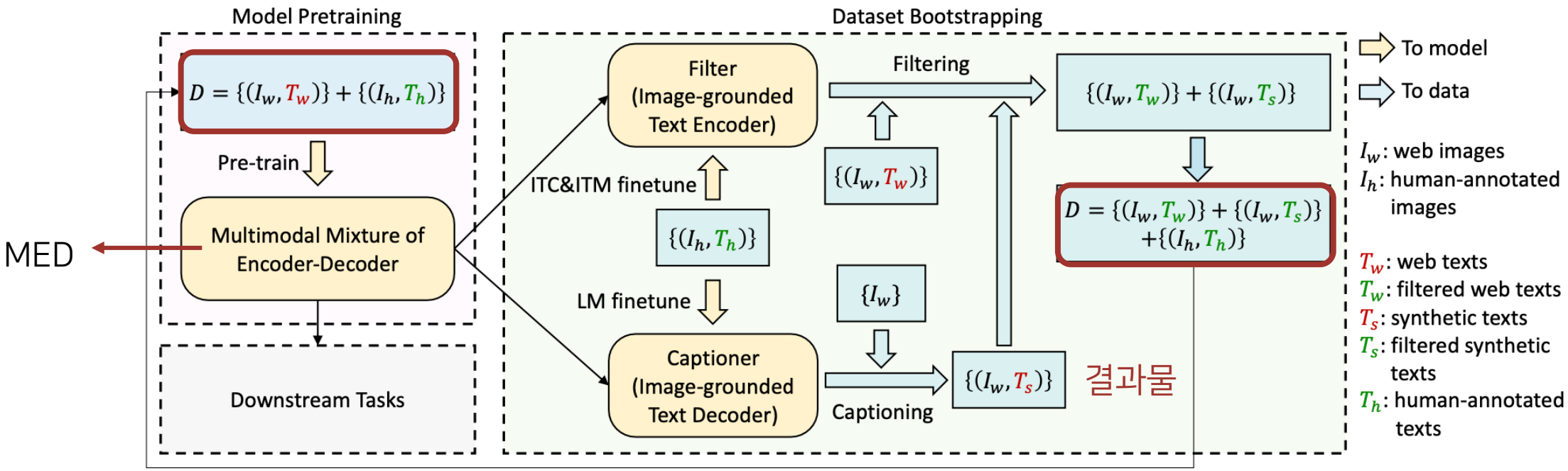


- Image-grounded Text Encoder & ITC, ITM objective에 따라 Fine-tuning
- Web에서 수집한 캡션 & 생성한 캡션 중 Noise가 있는 것 필터링

TRAIN AND TEST

Method

CapFlit

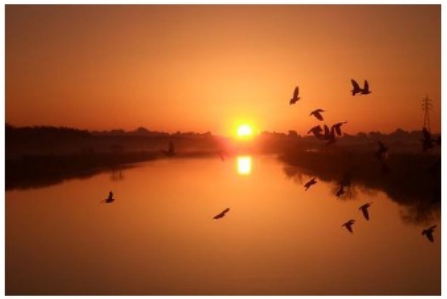


- 필터링된 데이터는 Human annotated 데이터와 결합되어 Pre-training에 활용
- 높은 퀄리티의 데이터셋 확보 => 성능 향상

TRAIN AND TEST

Method

CapFlit



T_w: “from bridge near my house”

T_s: “a flock of birds flying over a lake at sunset”

T_w: “in front of a house door in Reichenfels, Austria”

T_s: “a potted plant sitting on top of a pile of rocks”

T_w: “the current castle was built in 1180, replacing a 9th century wooden castle”

T_s: “a large building with a lot of windows on it”

TRAIN AND TEST

Experiments

CapFlit

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

TRAIN AND TEST

Experiments

Downstream tasks

다양한 vision-language tasks에 SOTA 성능

- image-text retrieval
- image captioning
- visual question answering
- visual reasoning
- visual dialog
- zero-shot performance (transfer to video-language tasks)
    - text-to-video retrieval
    - videoQA

TRAIN AND TEST

Experiments

Downstream tasks

Method	Pre-train # Images	COCO (5K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10
UNITER (Chen et al., 2020)	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA (Gan et al., 2020)	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR (Li et al., 2020)	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO (Li et al., 2021b)	5.7M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
ALIGN (Jia et al., 2021)	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF (Li et al., 2021a)	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
BLIP	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	**100.0**	87.2	97.5	98.8
BLIP	129M	**81.9**	95.4	97.8	**64.3**	85.7	91.5	**97.3**	**99.9**	**100.0**	87.3	97.6	**98.9**
$\text{BLIP}_{\text{CapFilt-L}}$	129M	81.2	**95.7**	**97.9**	64.1	**85.8**	**91.6**	97.2	**99.9**	**100.0**	**87.5**	**97.7**	**98.9**
$\text{BLIP}_{\text{ViT-L}}$	129M	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0

SOTA 모델과 Image-Text Retrieval 성능 비교

Method	Pre-train # Images	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.7	**100.0**	84.9	96.7	98.3
BLIP	129M	**96.0**	**99.9**	**100.0**	85.0	**96.8**	98.6
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**96.0**	**99.9**	**100.0**	**85.5**	**96.8**	**98.7**
$\text{BLIP}_{\text{ViT-L}}$	129M	96.7	100.0	100.0	86.7	97.3	98.7

Table 6. Zero-shot image-text retrieval results on Flickr30K.

Zero-shot 성능 비교

TRAIN AND TEST

Experiments

Downstream tasks

Method	Pre-train #Images	NoCaps validation								COCO Caption Karpathy test	
		in-domain		near-domain		out-domain		overall			
		C	S	C	S	C	S	C	S	B@4	C
Enc-Dec (Changpinyo et al., 2021)	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL† (Zhang et al., 2021)	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	**40.3**	**133.3**
BLIP	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
BLIP	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**111.8**	**14.9**	**108.6**	**14.8**	**111.5**	**14.2**	**109.6**	**14.7**	39.7	**133.3**
$\text{LEMON}_{\text{large}}$† (Hu et al., 2021)	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7
$\text{SimVLM}_{\text{huge}}$ (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
$\text{BLIP}_{\text{ViT-L}}$	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

SOTA 모델과 Image Captioning 성능 비교

TRAIN AND TEST

TRAIN AND TEST