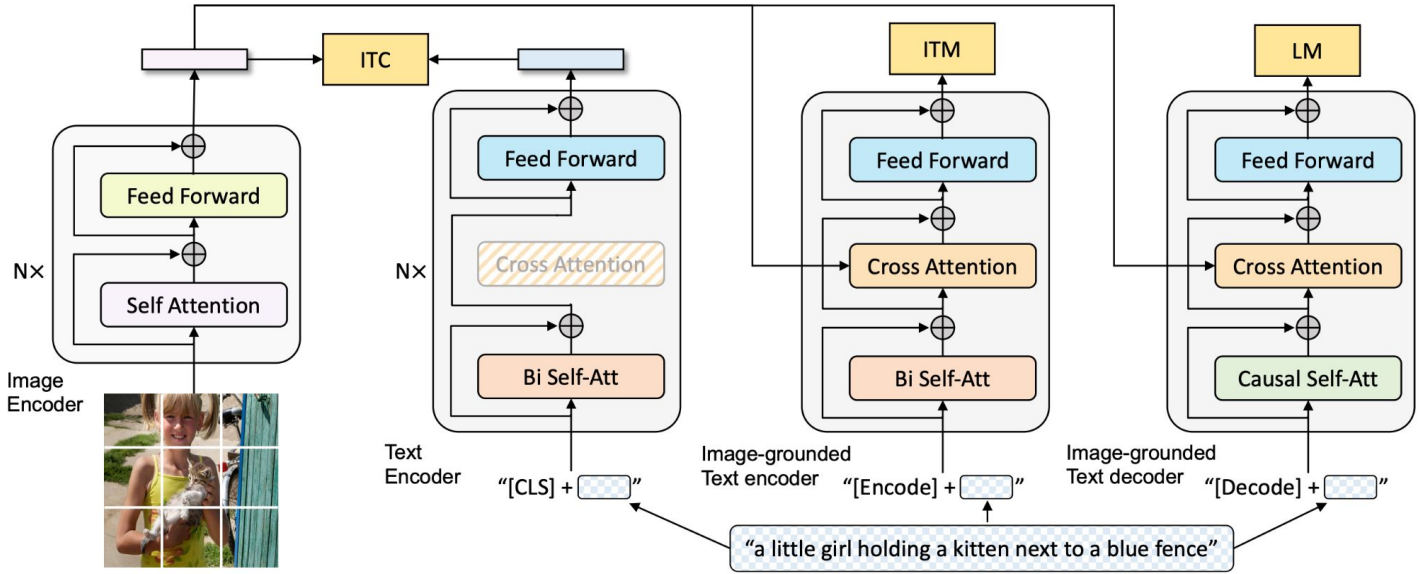
BLIP: **Bootstrapping** Language-Image Pre-training for Unified Vision-Language **Understanding and Generation**



Authors: Junnan Li Dongxu Li Caiming Xiong Steven Hoi @ Salesforce Research

Presenter: Filippos Bellos @ UofM

Motivation and Previous Work

- Model perspective: existing vision-language pretrained models lack flexibility when transferred to downstream tasks
  Encoder based models: Less straightforward to directly transfer to text generation tasks (CLIP, ALBEF)
  Encoder-Decoder models: Have not been successfully adopted for image-text retrieval tasks (SimVLM)

- Data perspective: most models pre-train on image and alt-text pairs that are automatically collected from the web
  The web texts often do not accurately describe the visual content of the images, making them a noisy source of supervision

Caption: from bridge near my house
Wrong image-text pair from the web

BLIP's solutions:

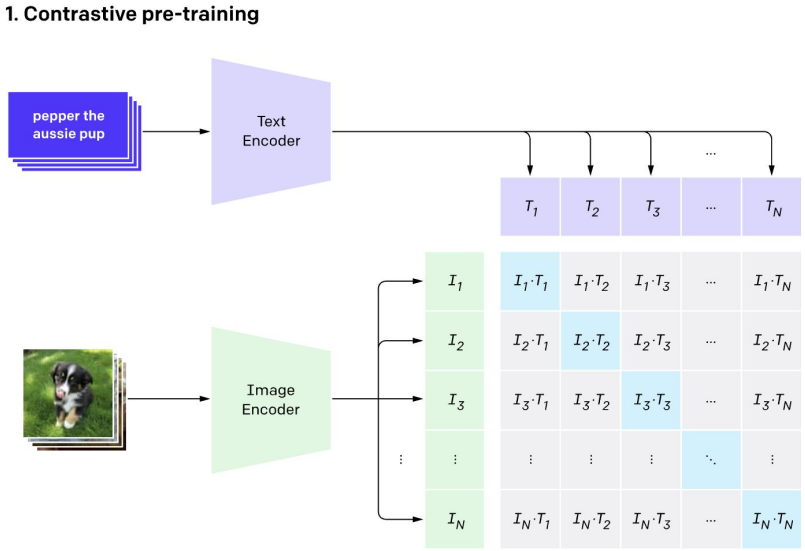
- a **new model architecture** that enables a wider range of downstream tasks than existing methods
- a **new dataset bootstrapping method** for learning from noisy web data.
  - https://arxiv.org/pdf/2103.00020.pdf
  - https://arxiv.org/pdf/2107.07651.pdf
  - https://arxiv.org/pdf/2108.10904.pdf

Vision Language Models background - Limitation 1

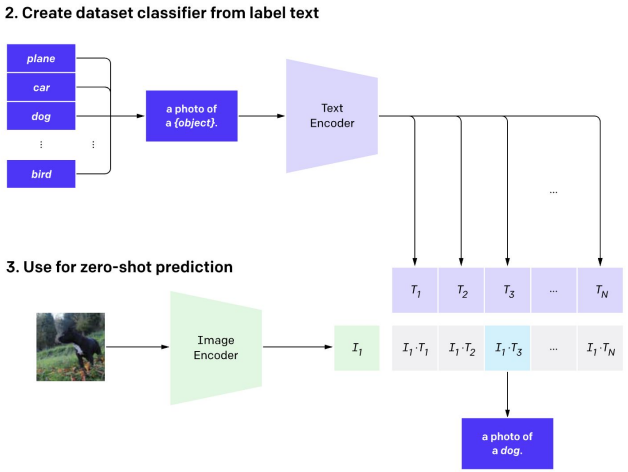
Most existing work on vision-language representation learning fall into two categories.

1. The first category focuses on learning separate unimodal encoders for image and text.
2. The second category focuses on modelling the interactions between image and text features with transformer-based multimodal encoders.

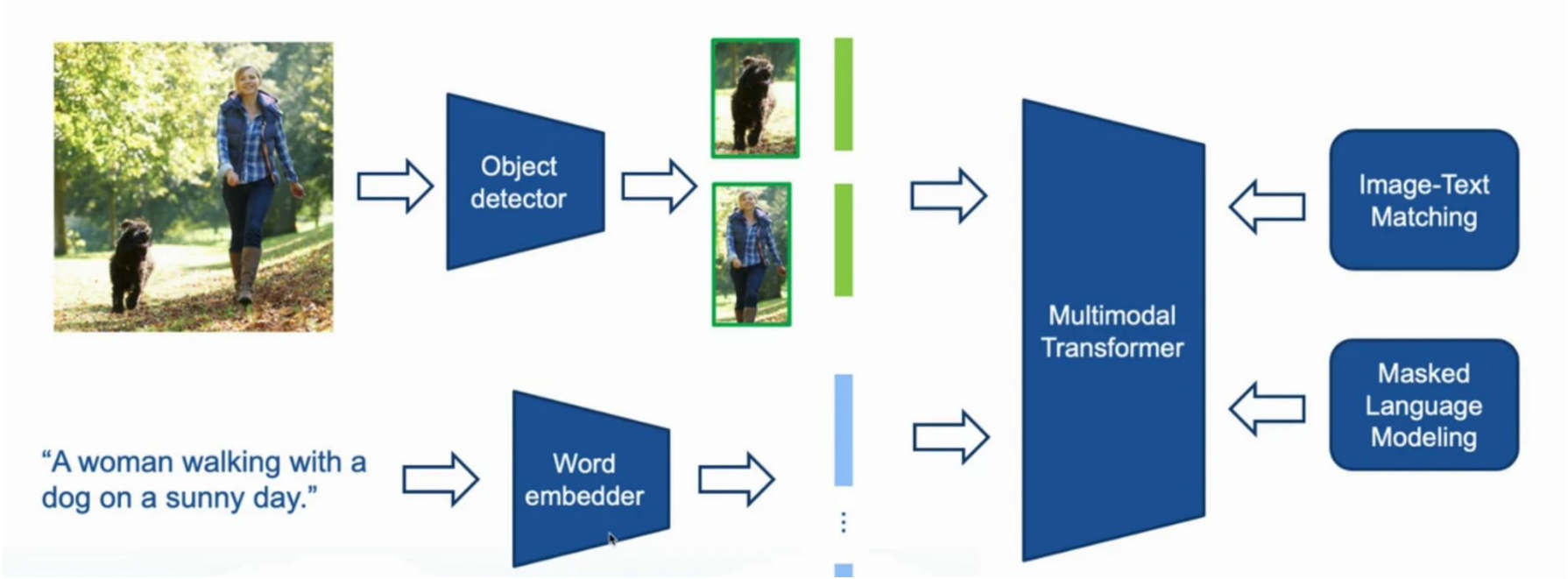
Unimodal Encoders



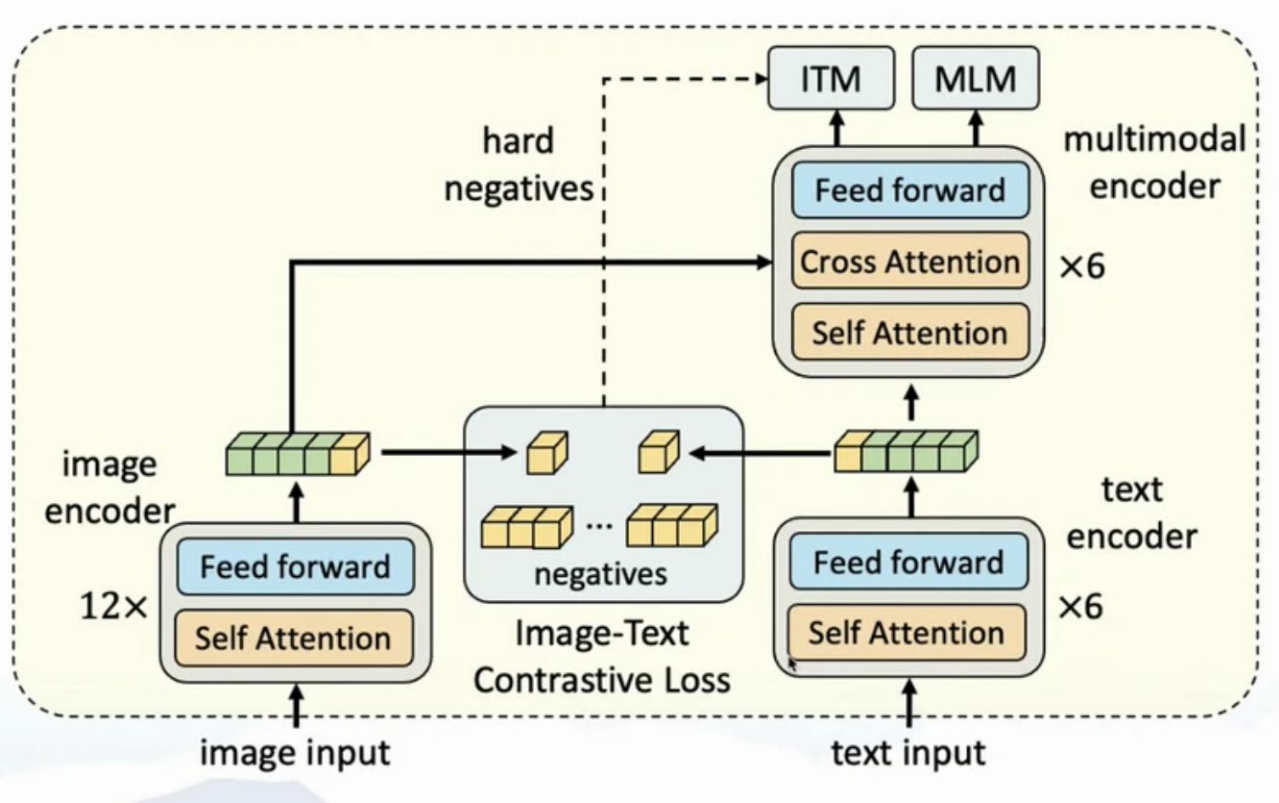
- Perform pre-training on massive noisy web data using a contrastive loss
- Remarkable performance on image-text retrieval tasks, but lack the ability to model more complex interactions between image and text for other V+L tasks.



Multimodal encoders



Multimodal encoders



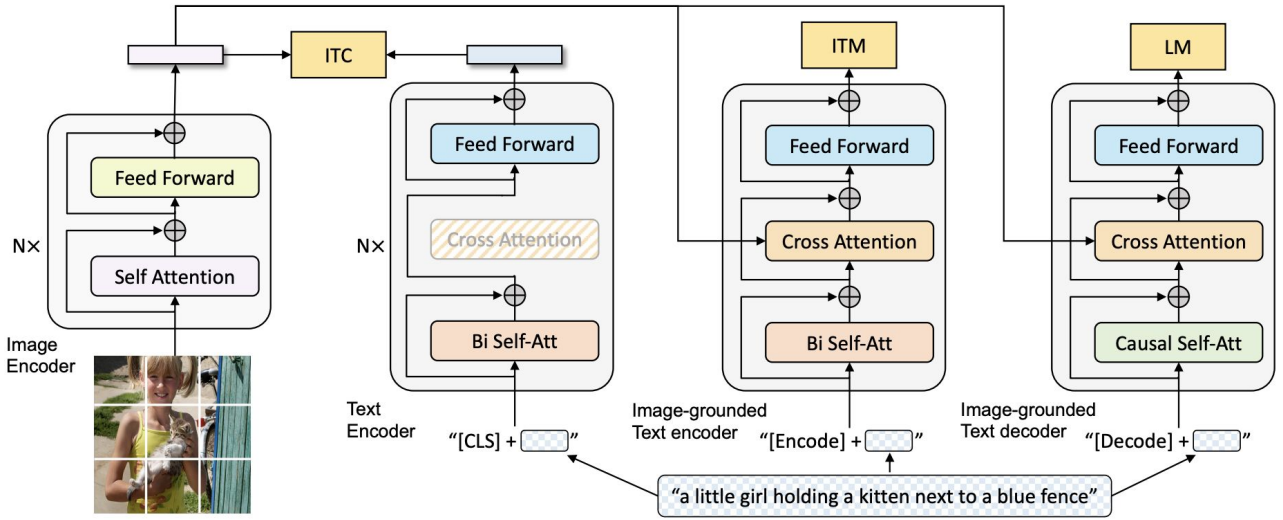
- Image encoder: ViT
- Text encoder: Bert
- Multimodal encoder for fusion, using cross attention layer.
- ITCL for better alignment

Method overview

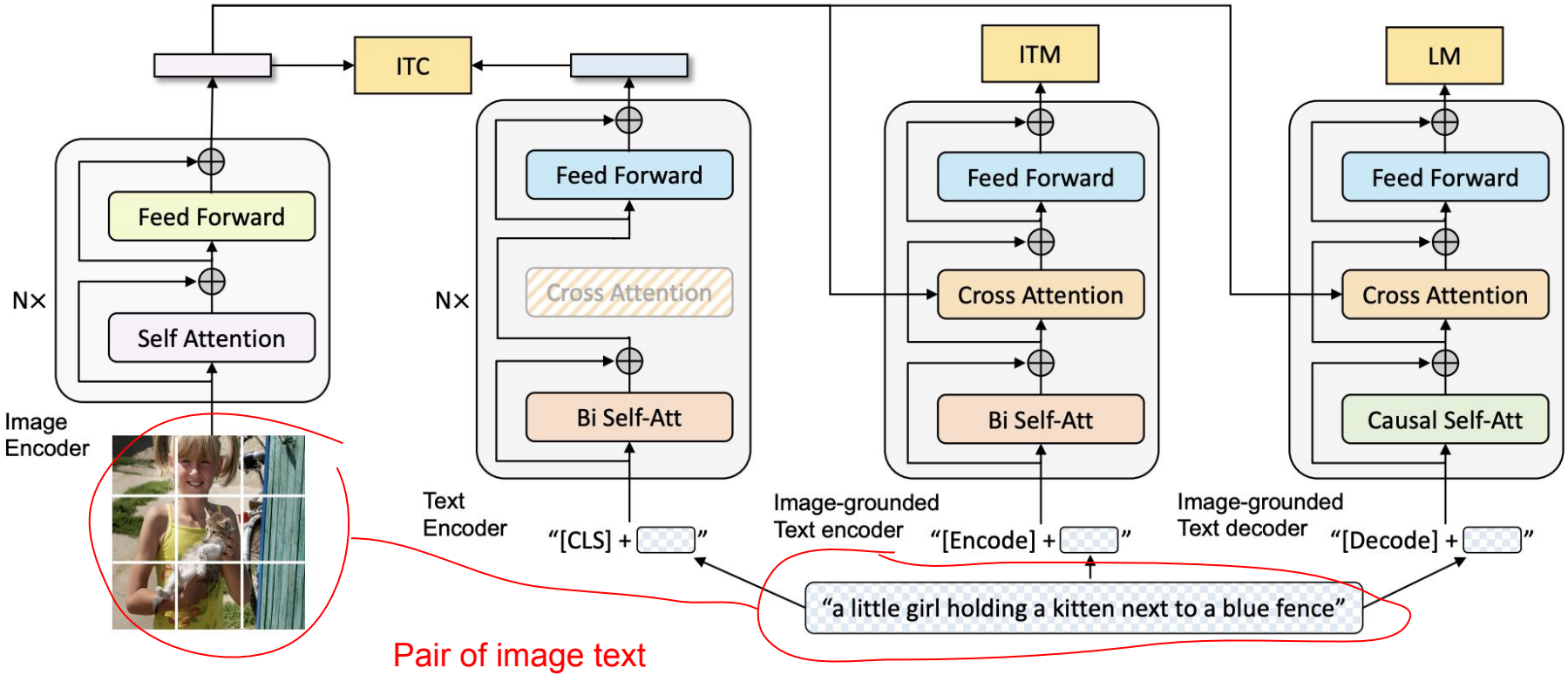
Unified model for vision-language understanding and generation called Multimodal mixture of Encoder-Decoder (MED):

3 architectural modes:
(1) Unimodal encoders (2) Image-grounded text encoder (3) Image-grounded text decoder

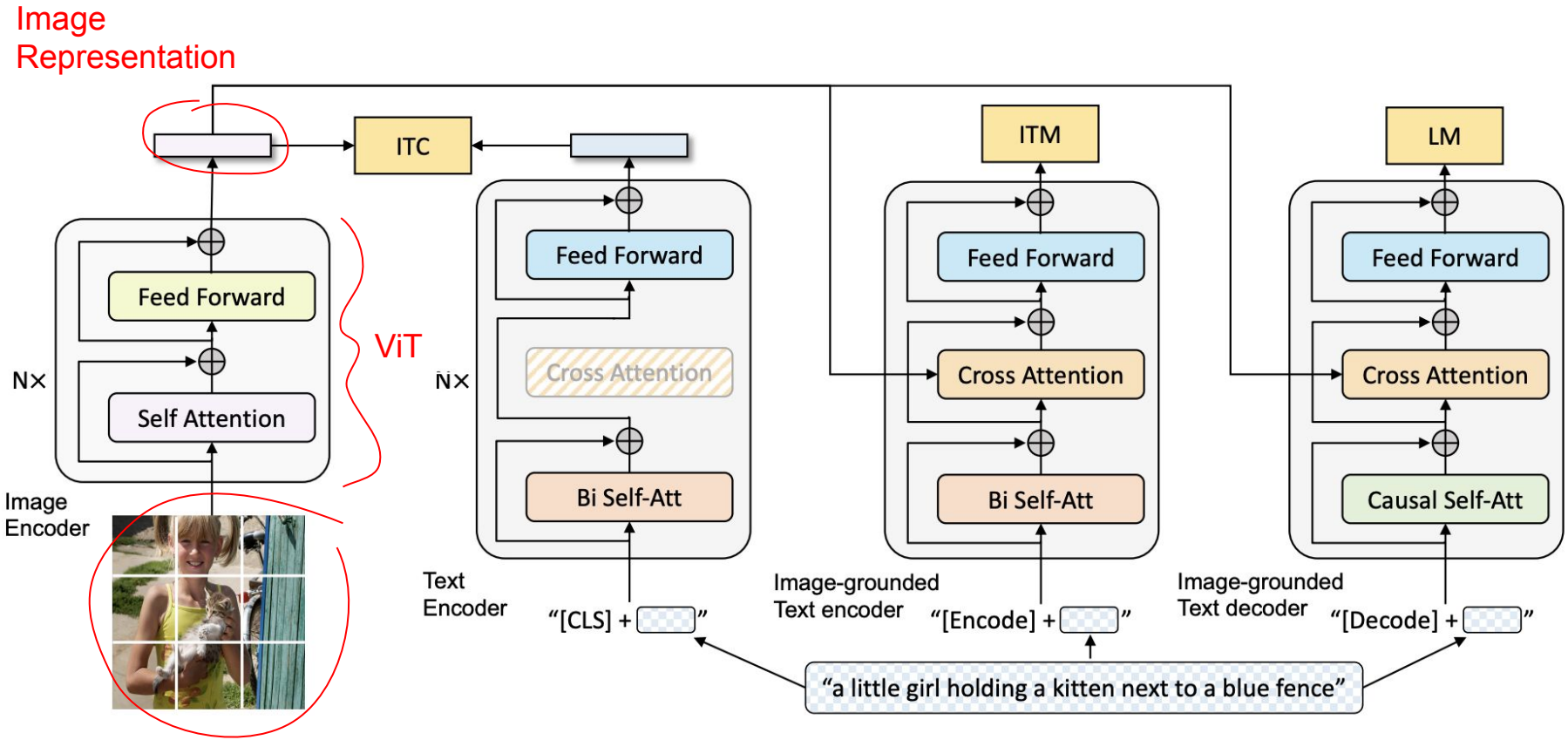
3 objectives:
(1) Image-text contrastive learning (2) Image-text matching (3) Image-conditioned language modeling



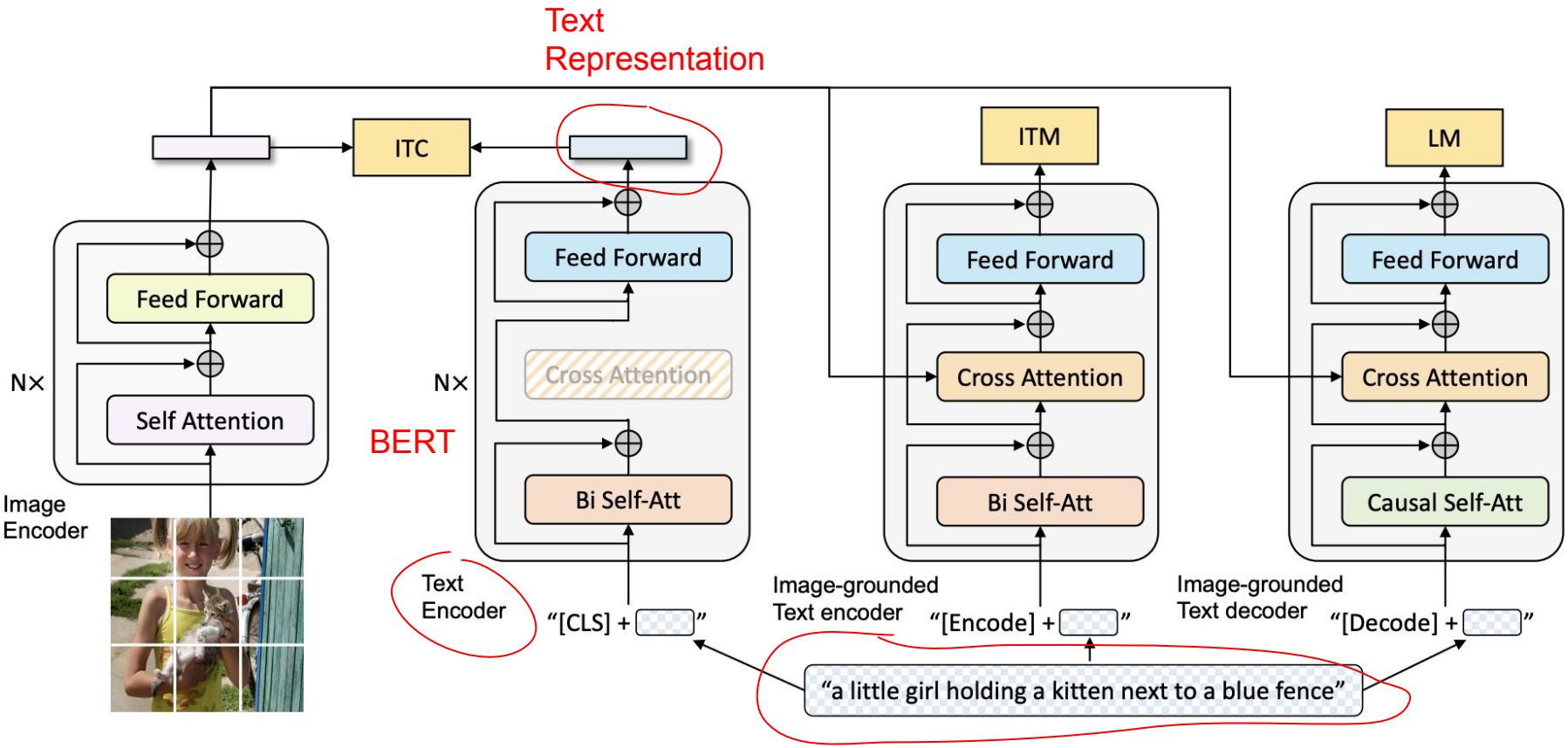
Method analysis



Method analysis

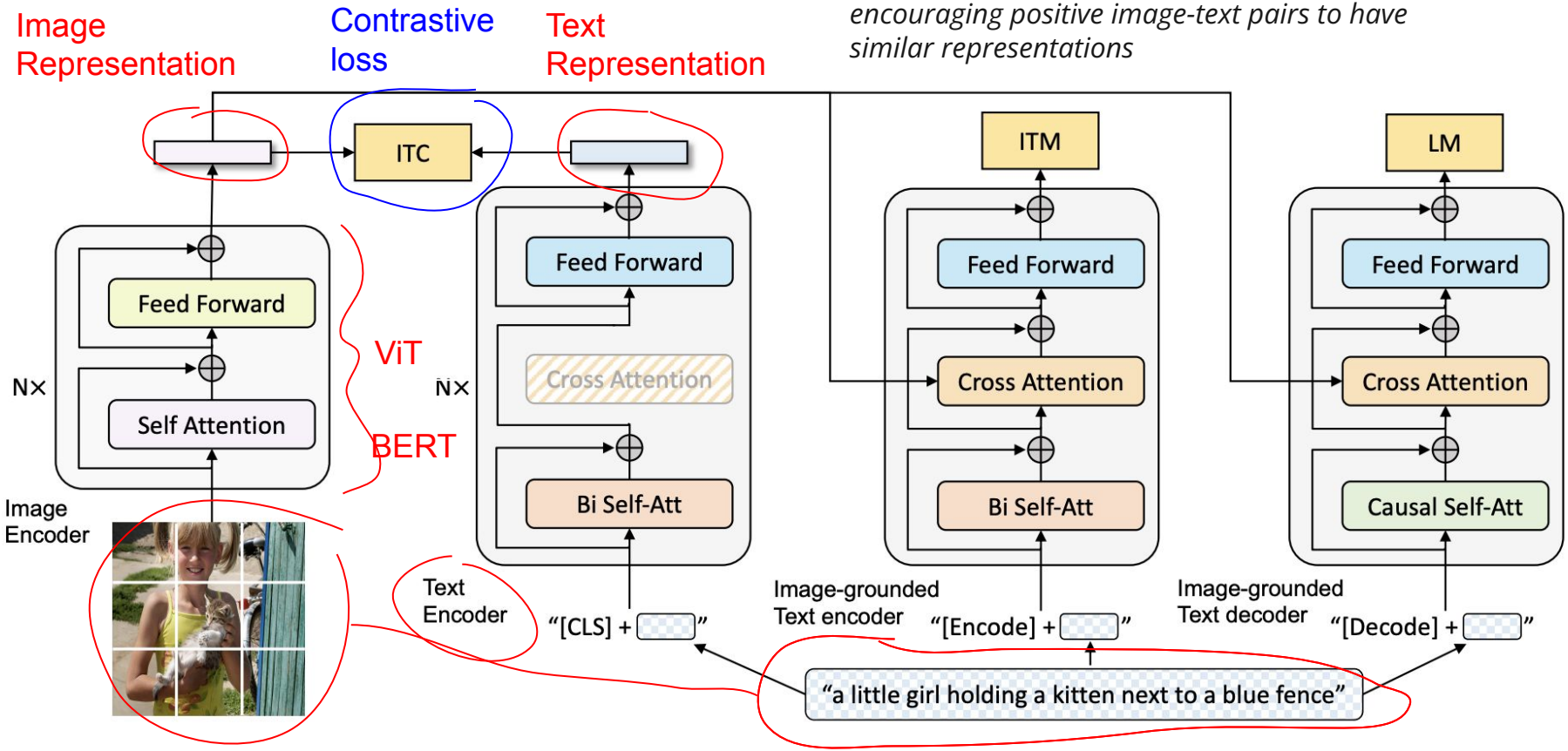


Method analysis



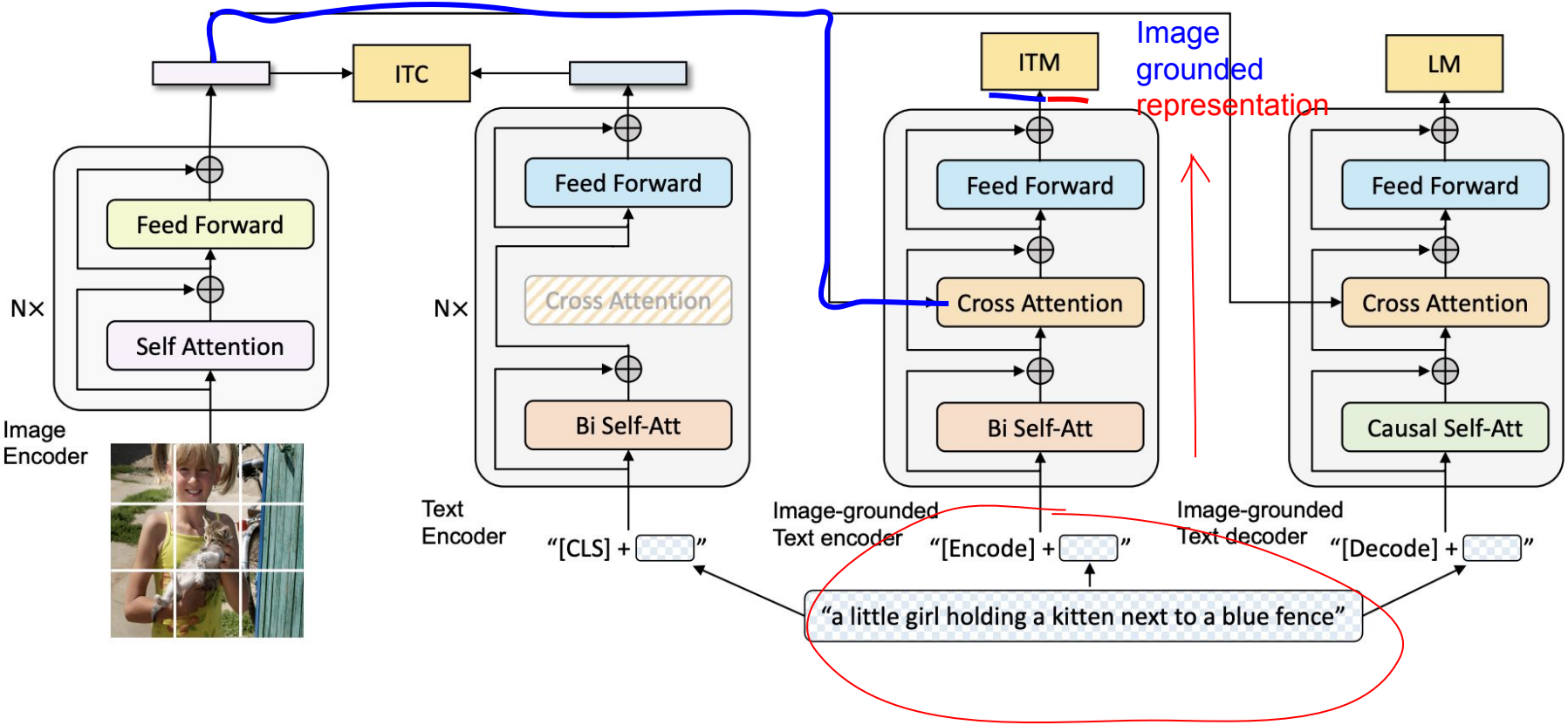
Method analysis

ITC: aims to *align the feature space of the visual transformer and the text transformer by encouraging positive image-text pairs to have similar representations*



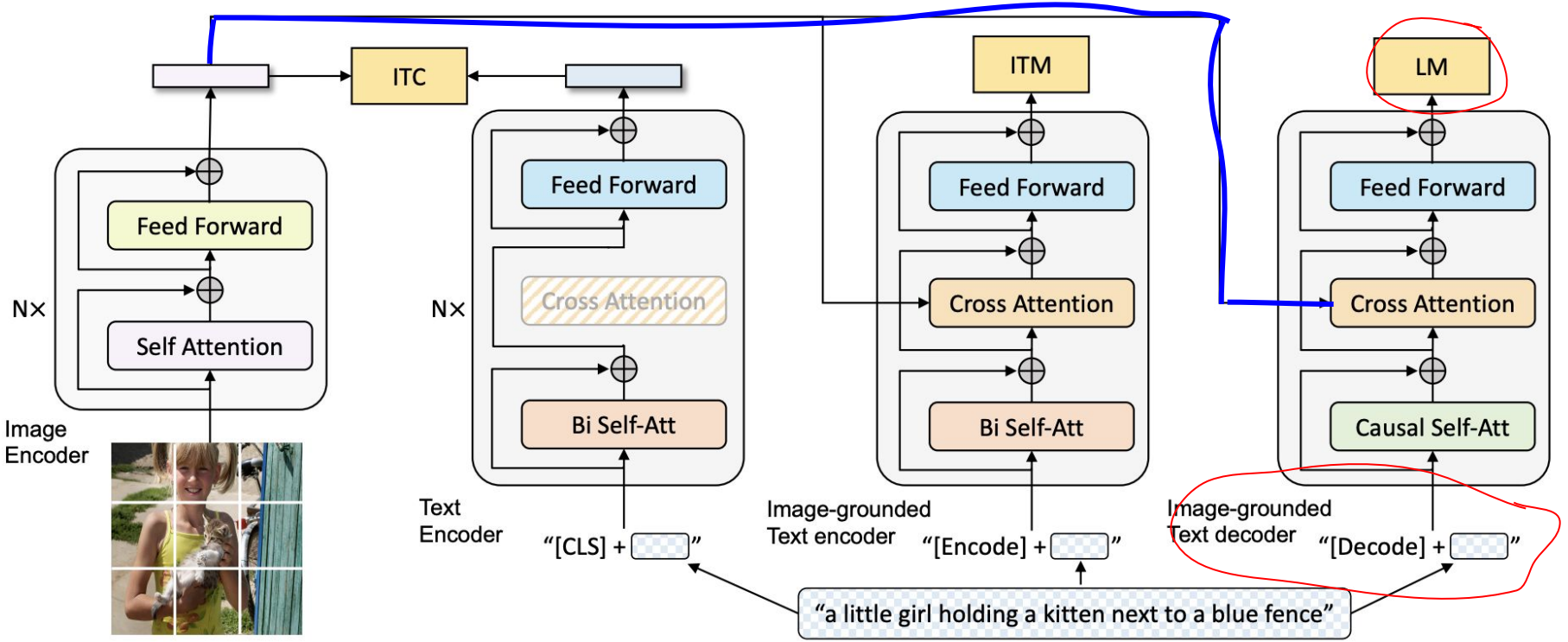
Method analysis

ITM is a binary classification task, where *the model is asked to predict whether an image-text pair is positive (matched) or negative (unmatched)*



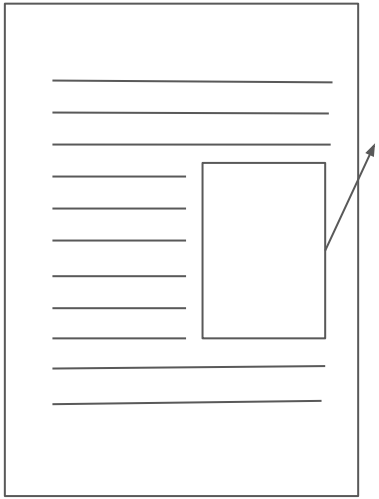
Method analysis

LM: aims to *generate textual descriptions conditioned on the images*.



Limitation 2: Noisy web data

Html page



<img src:"...."
alt:"...">

Not aligned image - text pairs

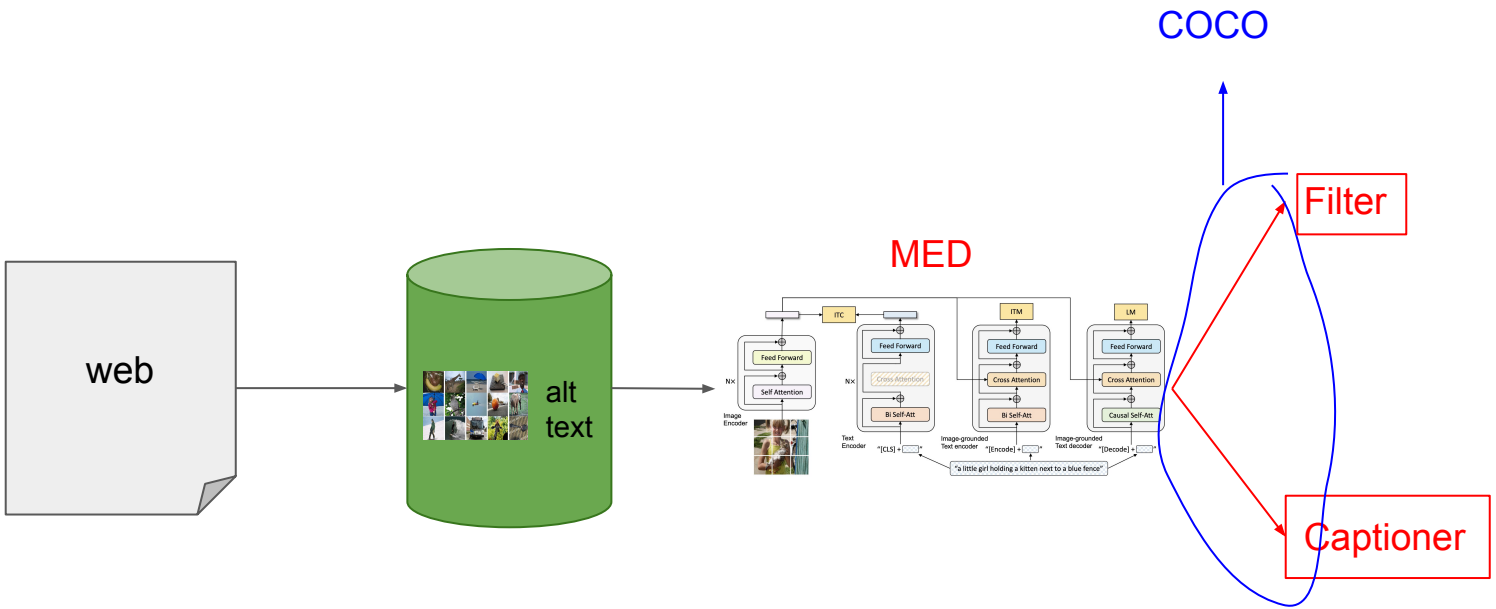
T_w: "from bridge near my house"

T_s: "a flock of birds flying over a lake at sunset"

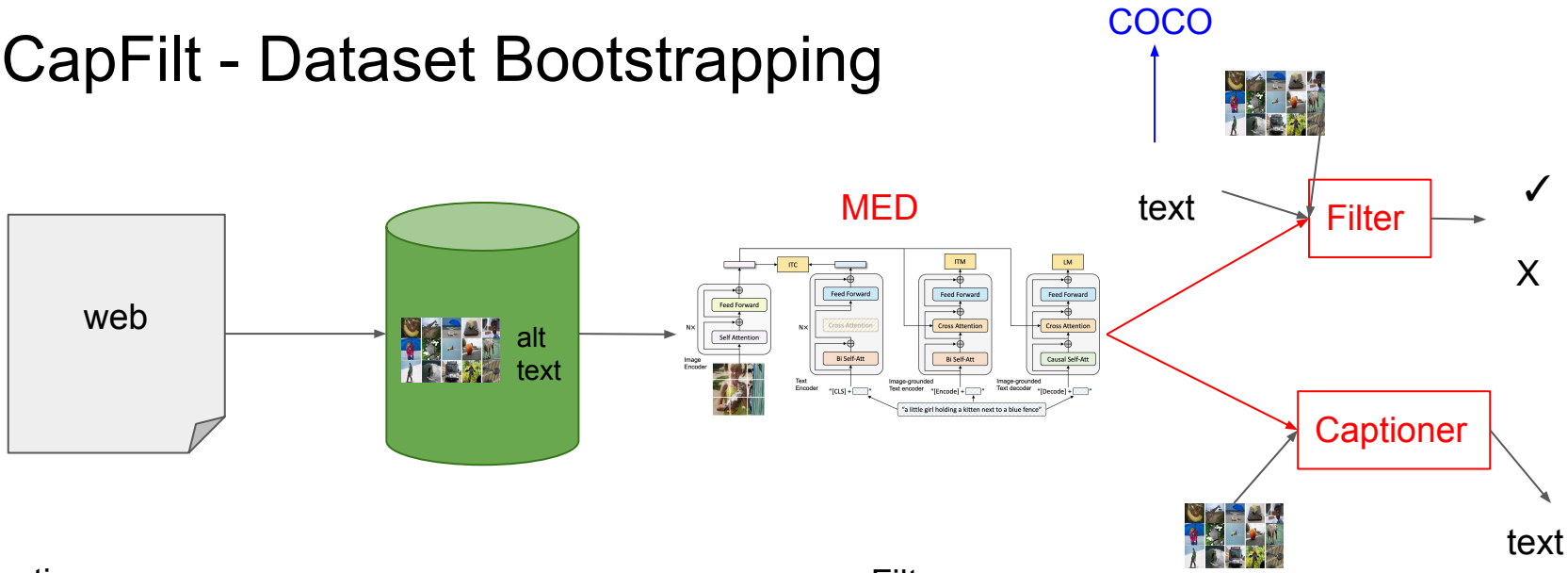
T_w: "in front of a house door in Reichenfels, Austria"

T_s: "a potted plant sitting on top of a pile of rocks"

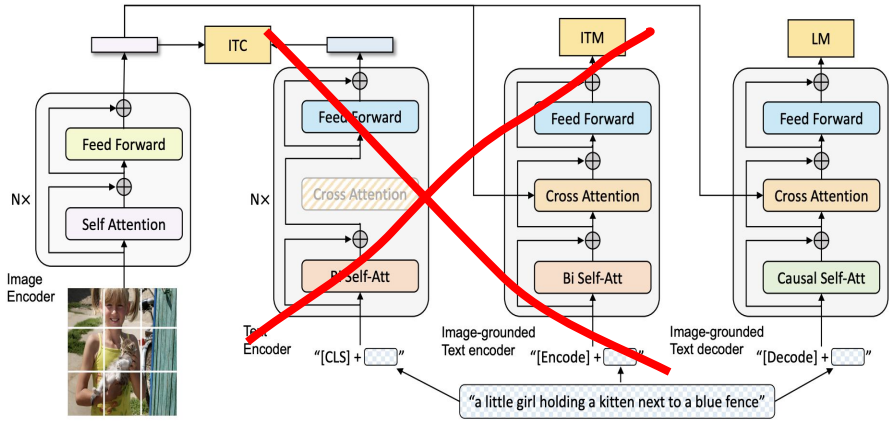
CapFilt - Dataset Bootstrapping



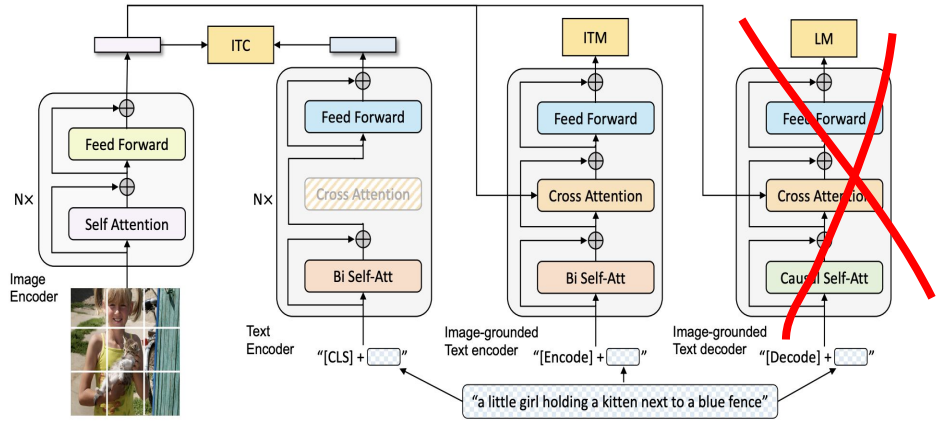
CapFilt - Dataset Bootstrapping



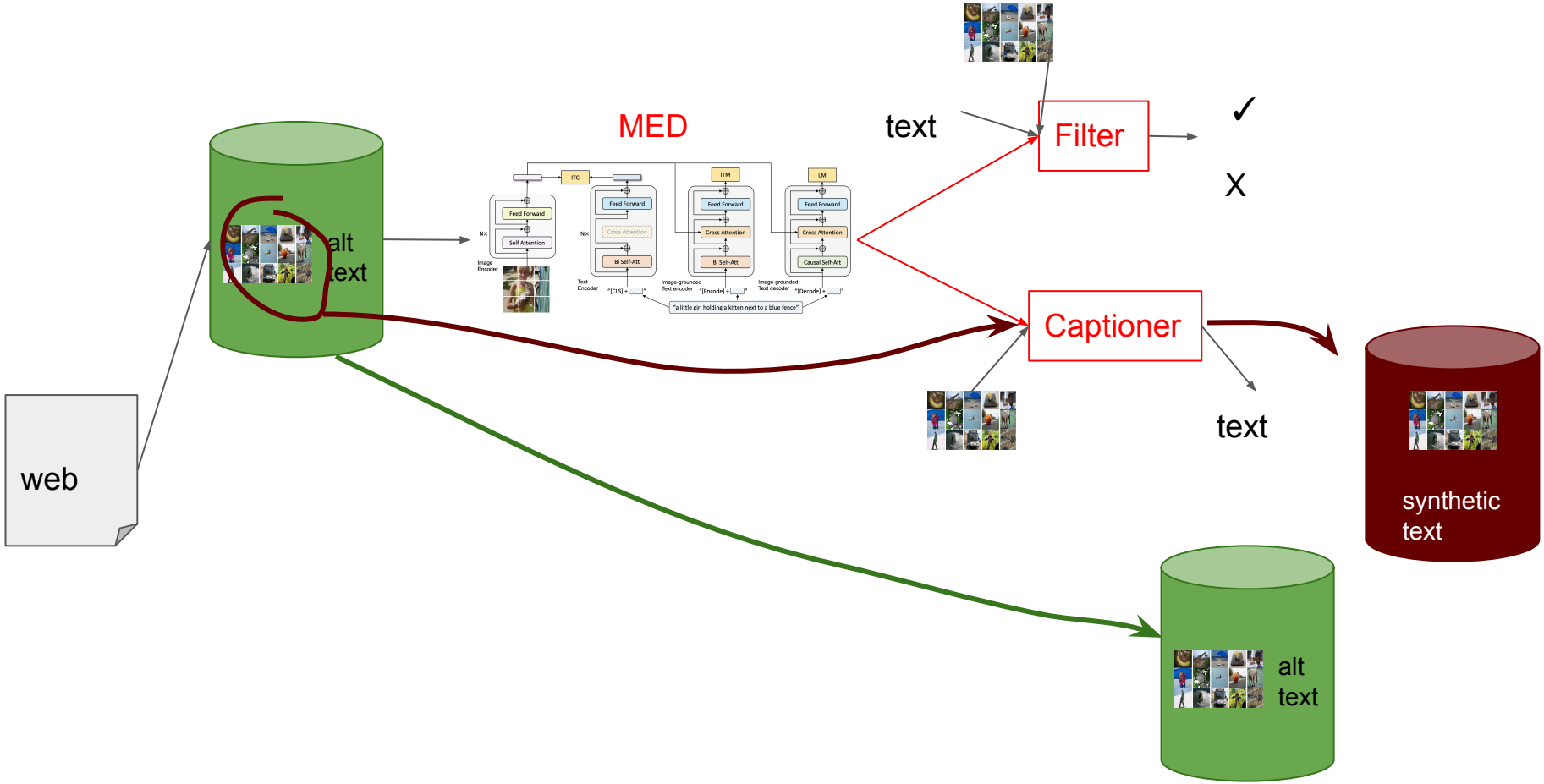
Captioner



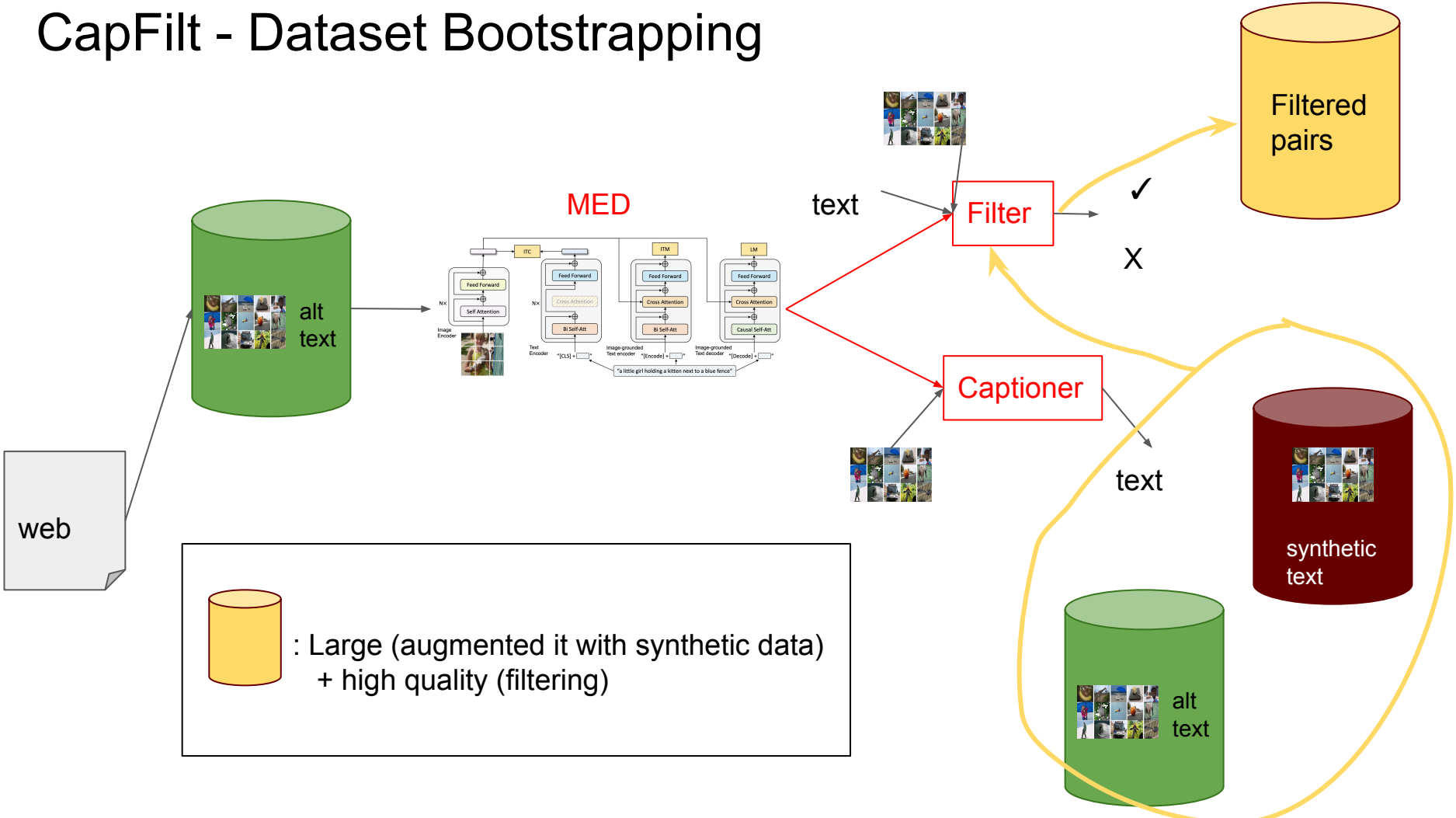
Filter



CapFilt - Dataset Bootstrapping



CapFilt - Dataset Bootstrapping



Experimental Setup

- **Model Architectures**
  - Image encoder: Two variants of ViTs: ViT-B/16 and ViT-L/16
  - Text encoder: Bert
- **Models Initialization**
  - ViT pretrained on ImageNet
  - Text transformer initialized from BERTbase
- **Pretraining**
  - Same pre-training dataset as Li et al. with 14M images in total
    - Two human-annotated datasets (COCO and Visual Genome)
    - Three web datasets (Conceptual Captions, Conceptual 12M, SBU captions).
  - Additional web dataset, LAION (more noisy texts)

	COCO	VG	SBU	CC3M	CC12M	LAION
# image	113K	100K	860K	3M	10M	115M
# text	567K	769K	860K	3M	10M	115M

Experiments

Effect of CapFilt

- Captioner and Filter can lead to performance improvement individually
- They lead to significant improvements collaboratively
- CapFilt works for different size of models and datasets
- Increase in data amount and model size as helpful as CapFilt

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

. Evaluation of the effect of the captioner (C) and filter (F) for dataset bootstrapping.

Parameter sharing

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

- During pre-training, the text encoder and decoder share all parameters except for the self-attention layers.
- The differences between the encoding and decoding tasks are best captured by the SA layers.
- On the other hand, the embedding layers, CA layers and FFN function similarly between encoding and decoding tasks, so they are being shared.

Experiments (Overview)

- Vision-Language (Understanding and Generation) Tasks and Datasets
    - Image-Text Retrieval: COCO and Flickr30K
    - Image Captioning: NoCaps and COCO
        - Increase performance by adding prompt at the beginning “A picture of”
    - Visual Question Answering (VQA)
    - VQA and NLVR
        - Generation instead of classification task
        - Outperform SOTAs with less parameters and data
    - Natural Language VIsual Reasoning
    - Visual Dialog: VisDial
    - Text-to-Video Retrieval and Video Question Answering: MSRVTT

Experiments (Image-Text retrieval)

Method	Pre-train # Images	COCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	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

BLIP achieves substantial performance improvement compared with existing methods. Using the same 14M pre-training images, BLIP outperforms the previous best model ALBEF by +2.7% in average recall@1 on COCO

Experiments (Zero-Shot Image-Text retrieval)

Method	Pre-train # Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	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

- Directly transferring the model finetuned on COCO to Flickr30K
- BLIP also outperforms existing methods by a large margin

Experiments (Image-Captioning)

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

Comparison with state-of-the-art image captioning methods on NoCaps and COCO Caption. All methods optimize the crossentropy loss during finetuning. C: CIDEr, S: SPICE, B@4: BLEU@4. BLIPCapFilt-L is pre-trained on a dataset bootstrapped by captioner and filter with ViT-L. VinVL† and LEMON† require an object detector pre-trained on 2.5M images with human-annotated bounding boxes and high resolution (800×1333) input images. SimVLMhuge uses 13× more training data and a larger vision backbone than ViT-L.

Interesting Findings

Continue	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
Yes	80.6	63.0	94.5	84.6	38.5	129.9	104.5	14.2
No	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

- Continue training the original model does not help.
- This observation agrees with the common practice in knowledge distillation, where the student model cannot be initialized from the teacher.

Interesting Findings

CapFilt	#Texts	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
No	15.3M	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
No	24.7M	78.3	60.5	93.7	82.2	37.9	127.7	102.1	14.0
Yes	24.7M	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

- The original web texts are replicated to have the same number of samples per epoch as the bootstrapped dataset.
- Results verify that the improvement from CapFilt is not due to longer training time.

Quiz Questions

Which layers of the text encoder and decoder in MED share parameters? Is there a way to use Bi-self attention for the decoder as well, instead of causal attention?

Quiz Questions

Which layers of the text encoder and decoder in MED share parameters? Is there a way to use Bi-self attention for the decoder as well, instead of causal attention?

- They share all parameters except for the SA layers. The differences between the encoding and decoding tasks are best captured by the SA layers.
  On the other hand, the embedding layers, CA layers and FFN function similarly between encoding and decoding tasks, so they are being shared.

- We mask the tokens coming after the current token and we only attend to the previous ones, then it is possible to use a bi self-attention for the decoder as well.

Quiz Questions

The bootstrapped dataset is used to train a new model rather than continue training the original model. What could explain this?

Quiz Questions

The bootstrapped dataset is used to train a new model rather than continue training the original model. What could explain this?

- The experiments revealed that continuing the training of the pre-existing model with the bootstrapped dataset did not yield any benefits.
- Avoiding Overfitting: If the student model were initialized with the teacher model's parameters, it might inherit any overfitting or biases present in the teacher model. Starting from scratch helps to mitigate this risk.