

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

ICML 2022 (1460 Citations)

Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi

Presented by:
Michael Cruz, Christopher Lee, Saurabh Aggarwal, Robert Martin, Taylor Tiedge

Introduction

Problem: Task specific models, and noisy data

Solution: ensemble of task specific models for overall task generalization and CapFilt for improved data



T_w : "a week spent at our rented beach house in Sandbridge"

T_s : "an outdoor walkway on a grass covered hill"



T_w : "that's what a sign says over the door"

T_s : "the car is driving past a small old building"

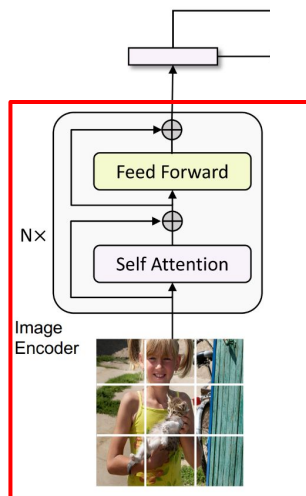


T_w : "hand held through the glass in my front bedroom window"

T_s : "a moon against the night sky with a black background"

Architecture

VIT Encoder



BERT Encoder

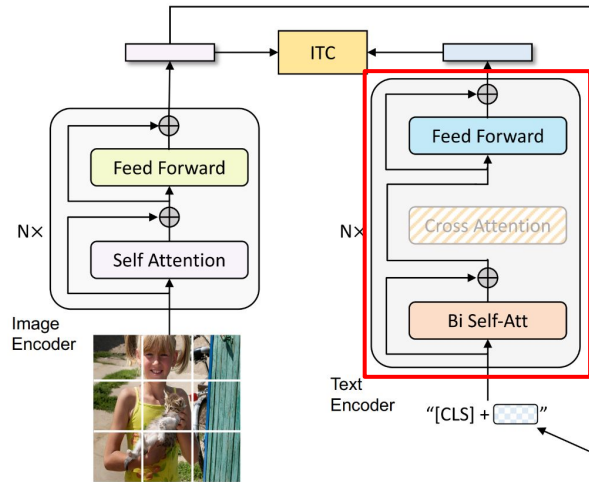
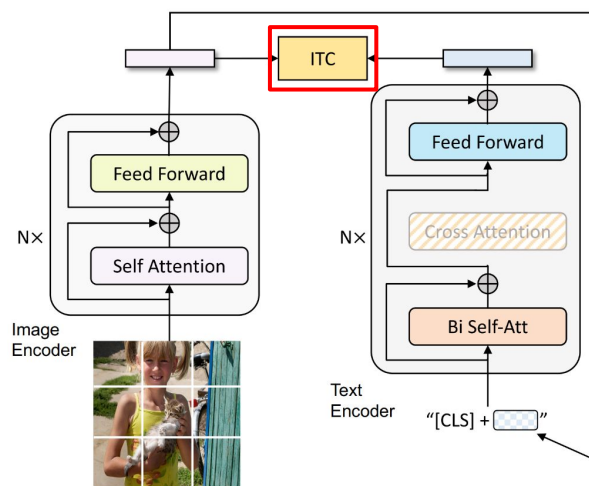


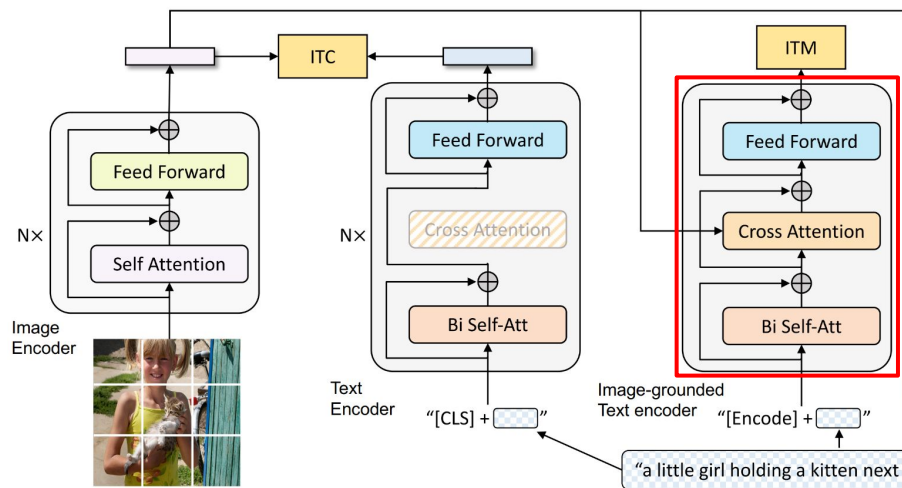
Image-Text Contrastive Loss



ITC:

$$\mathcal{L}_{\text{itc}} = \frac{1}{2} \mathbb{E}_{(I,T) \sim D} [\mathbb{H}(\mathbf{y}^{\text{i2t}}(I), \mathbf{p}^{\text{i2t}}(I)) + \mathbb{H}(\mathbf{y}^{\text{t2i}}(T), \mathbf{p}^{\text{t2i}}(T))]$$

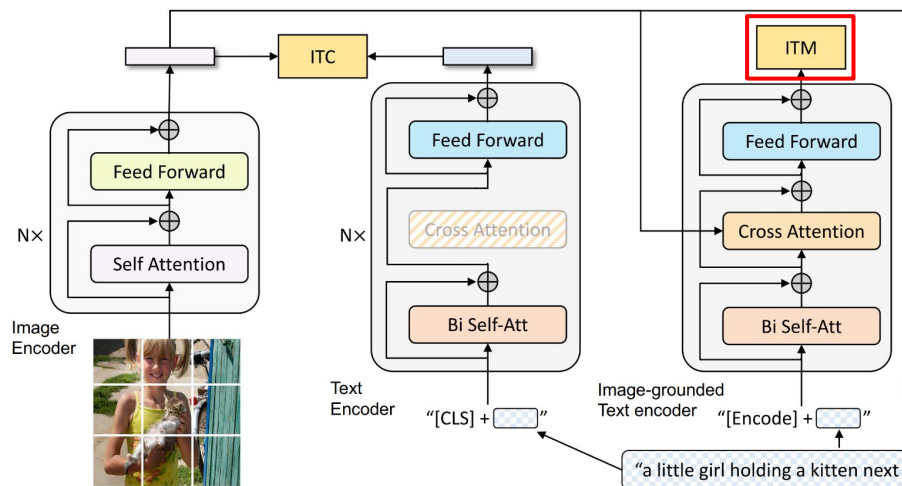
Image-Ground Text encoder



ITC:

$$\mathcal{L}_{\text{itc}} = \frac{1}{2} \mathbb{E}_{(I,T) \sim D} [\mathbb{H}(\mathbf{y}^{\text{i2t}}(I), \mathbf{p}^{\text{i2t}}(I)) + \mathbb{H}(\mathbf{y}^{\text{t2i}}(T), \mathbf{p}^{\text{t2i}}(T))]$$

Image-Text Matching Loss



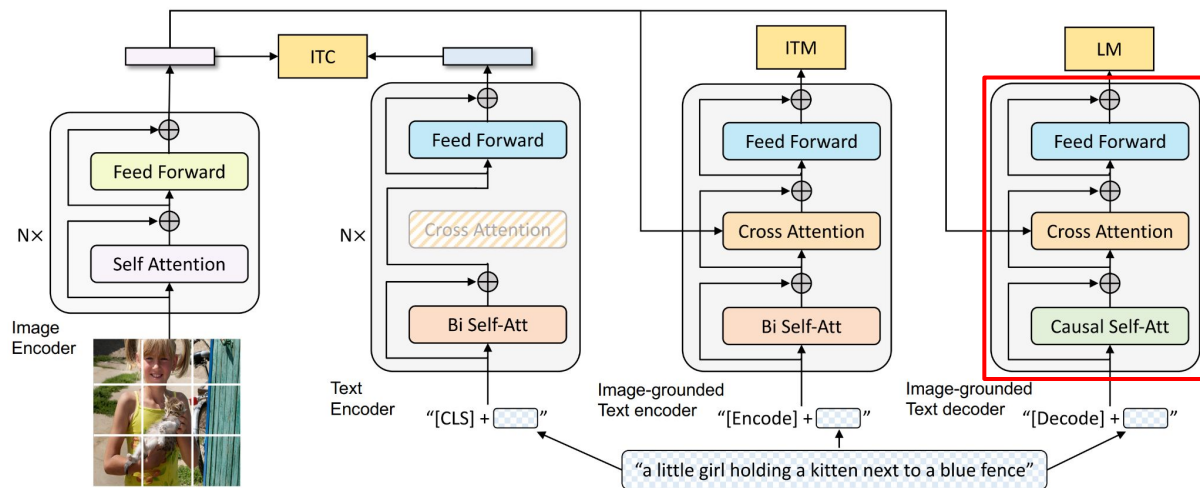
ITC:

$$\mathcal{L}_{\text{itc}} = \frac{1}{2} \mathbb{E}_{(I,T) \sim D} [\mathbb{H}(\mathbf{y}^{\text{i2t}}(I), \mathbf{p}^{\text{i2t}}(I)) + \mathbb{H}(\mathbf{y}^{\text{t2i}}(T), \mathbf{p}^{\text{t2i}}(T))]$$

ITM:

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T) \sim D} \mathbb{H}(\mathbf{y}^{\text{itm}}, \mathbf{p}^{\text{itm}}(I, T))$$

Image-Ground Text Decoder



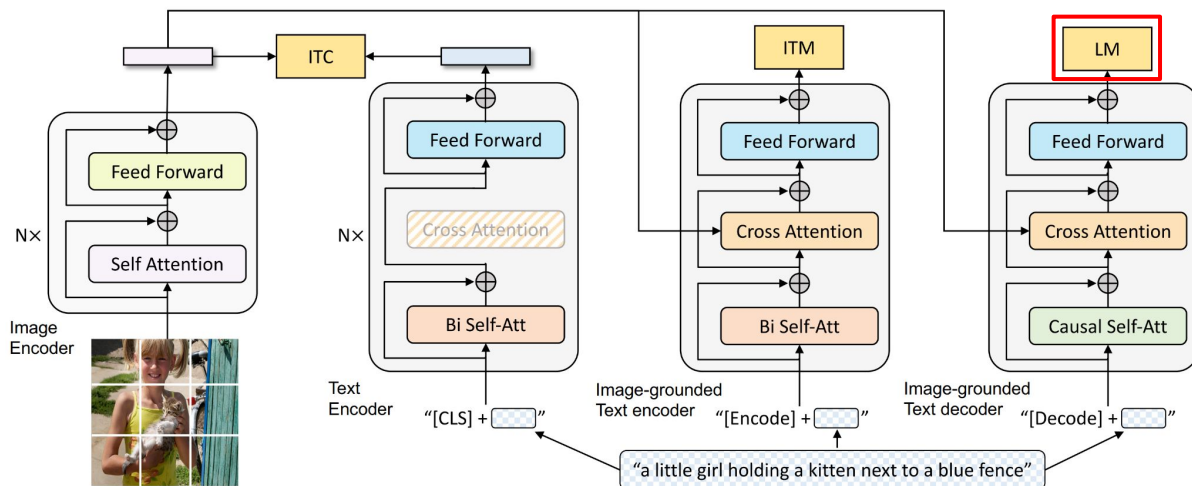
ITC:

$$\mathcal{L}_{\text{itc}} = \frac{1}{2} \mathbb{E}_{(I,T) \sim D} [\mathbb{H}(\mathbf{y}^{\text{itc}}(I), \mathbf{p}^{\text{itc}}(I)) + \mathbb{H}(\mathbf{y}^{\text{itc}}(T), \mathbf{p}^{\text{itc}}(T))]$$

ITM:

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T) \sim D} \mathbb{H}(\mathbf{y}^{\text{itm}}, \mathbf{p}^{\text{itm}}(I, T))$$

Language Modeling Loss



ITC:

$$\mathcal{L}_{\text{itc}} = \frac{1}{2} \mathbb{E}_{(I,T) \sim D} [\mathbb{H}(\mathbf{y}^{\text{i2t}}(I), \mathbf{p}^{\text{i2t}}(I)) + \mathbb{H}(\mathbf{y}^{\text{t2i}}(T), \mathbf{p}^{\text{t2i}}(T))]$$

ITM:

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T) \sim D} \mathbb{H}(\mathbf{y}^{\text{itm}}, \mathbf{p}^{\text{itm}}(I, T))$$

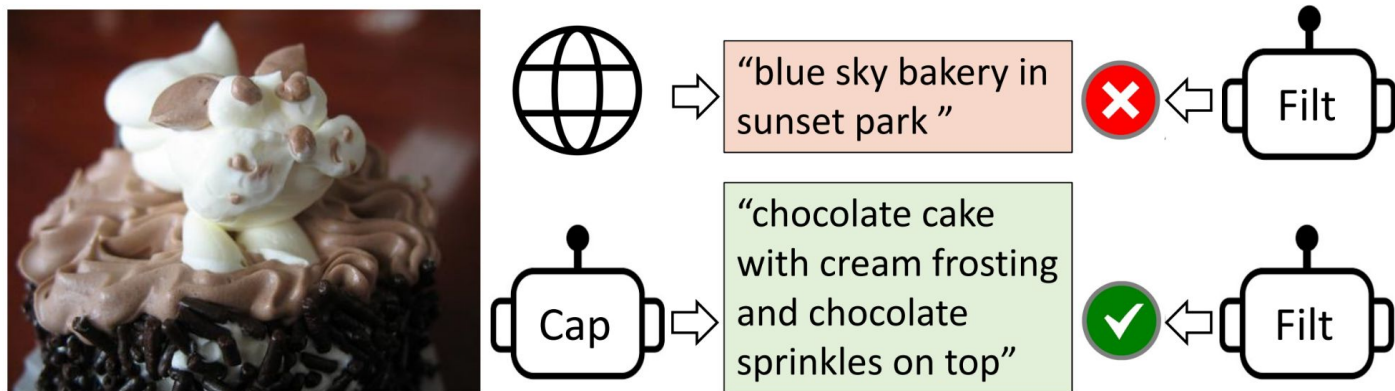
LM:

$$H(t, p) = - \sum_{s \in S} t(s) \cdot \log(p(s))$$

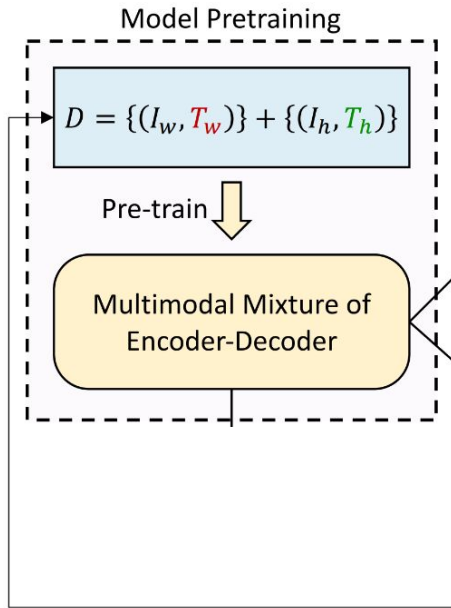
CapFilt System

Captioning and Filtering

The goal of CapFilt is to take noisy web data, filter out unusable captions, and provide higher quality synthetic captions

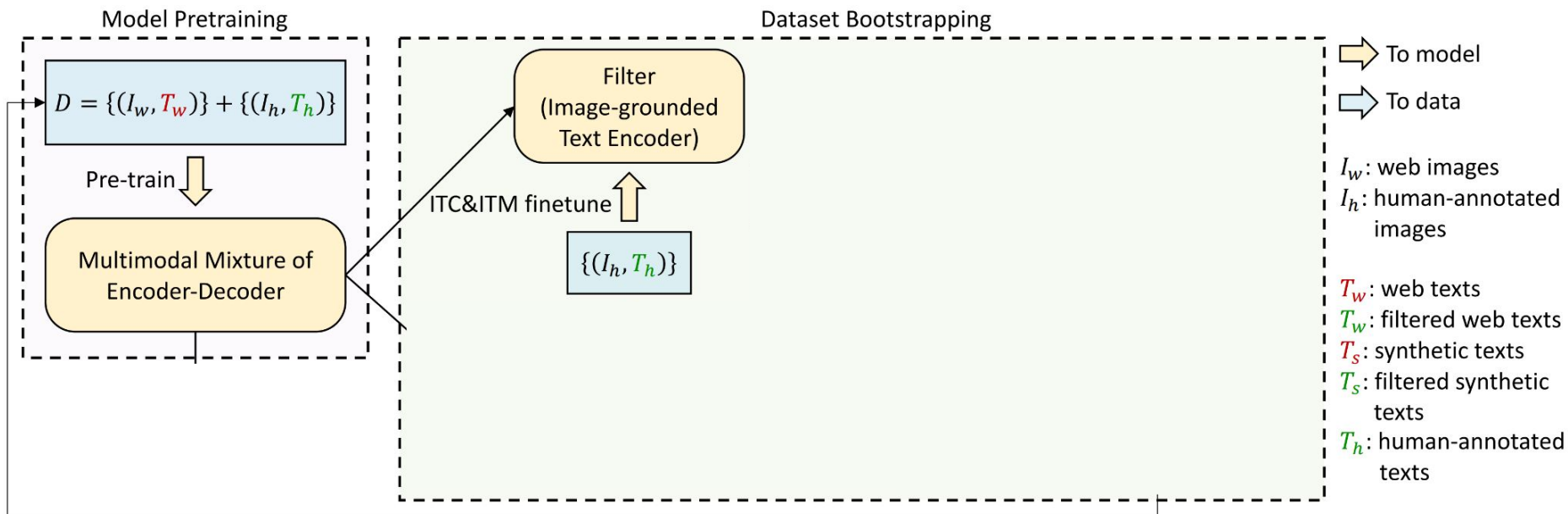


Model Pretraining

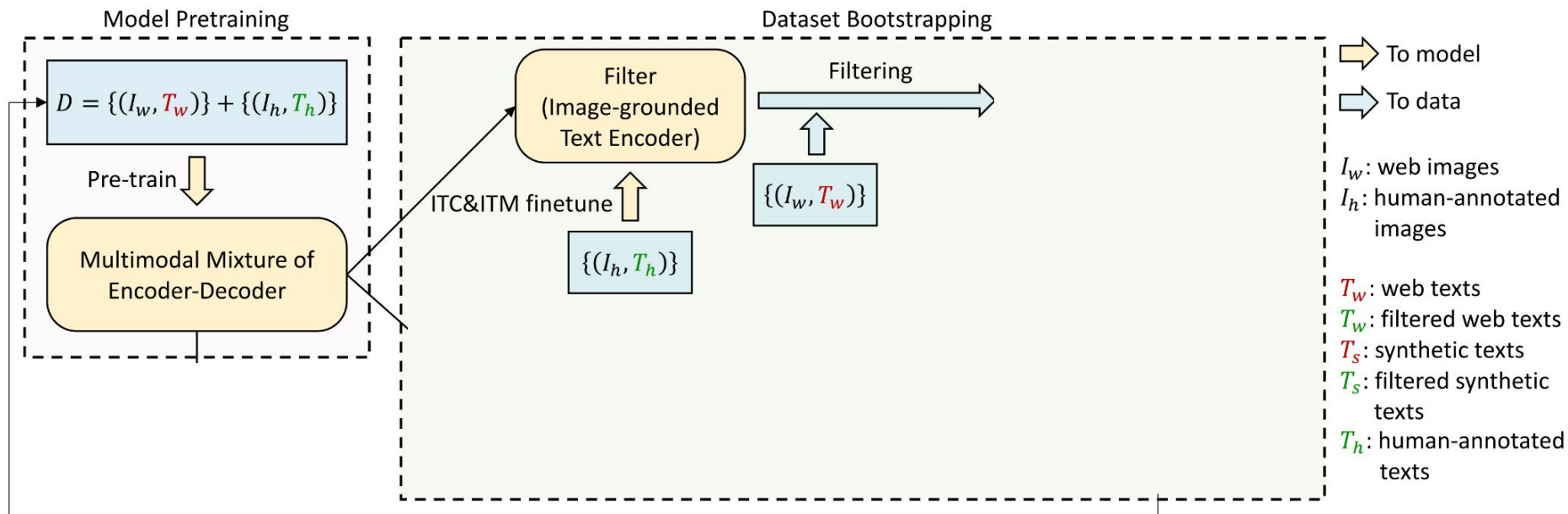


Starts with unfiltered web data
and human annotated data

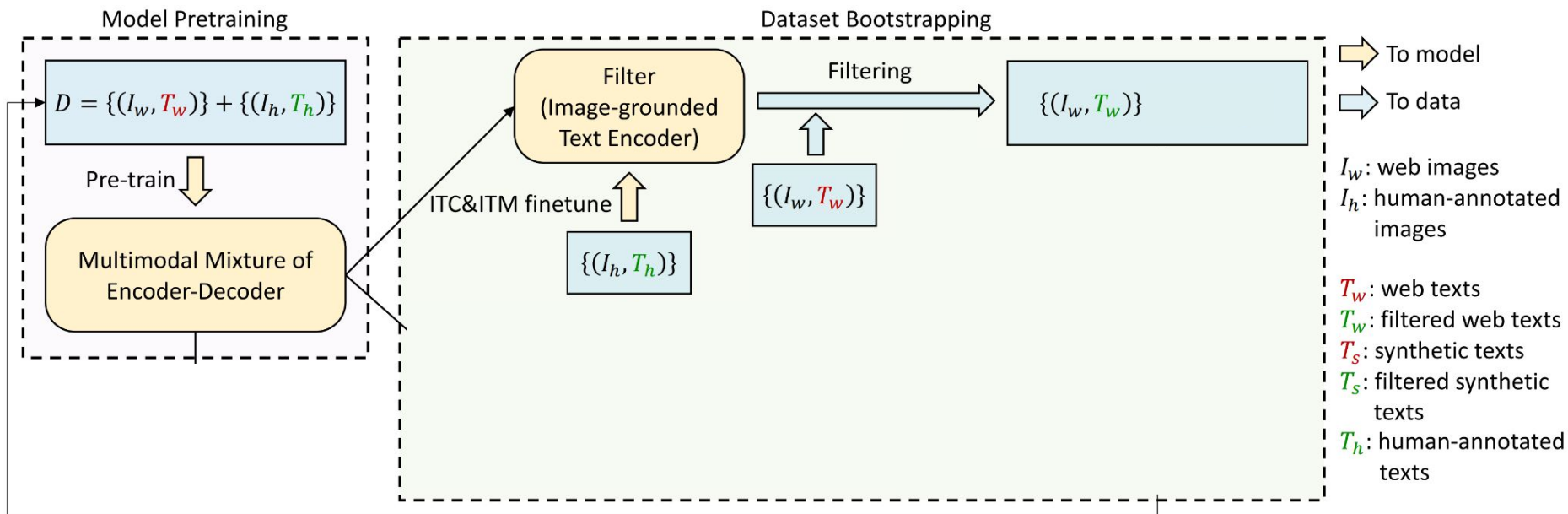
Filter Finetuning



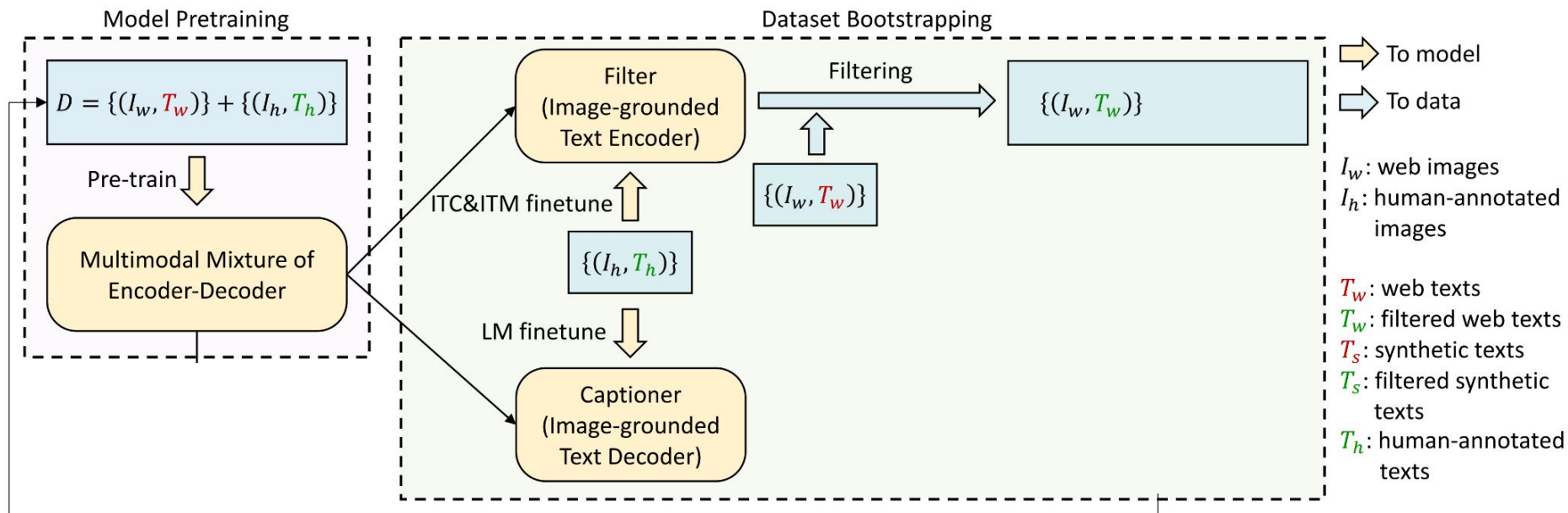
Filtering Web Data



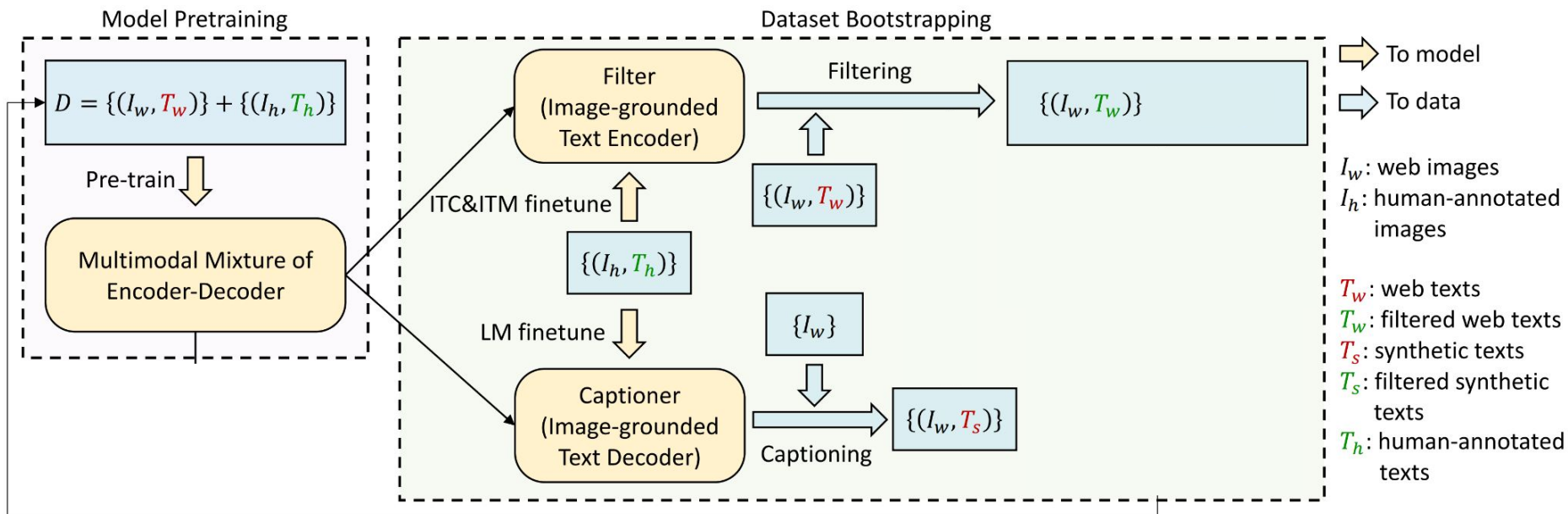
Filtered Web Data



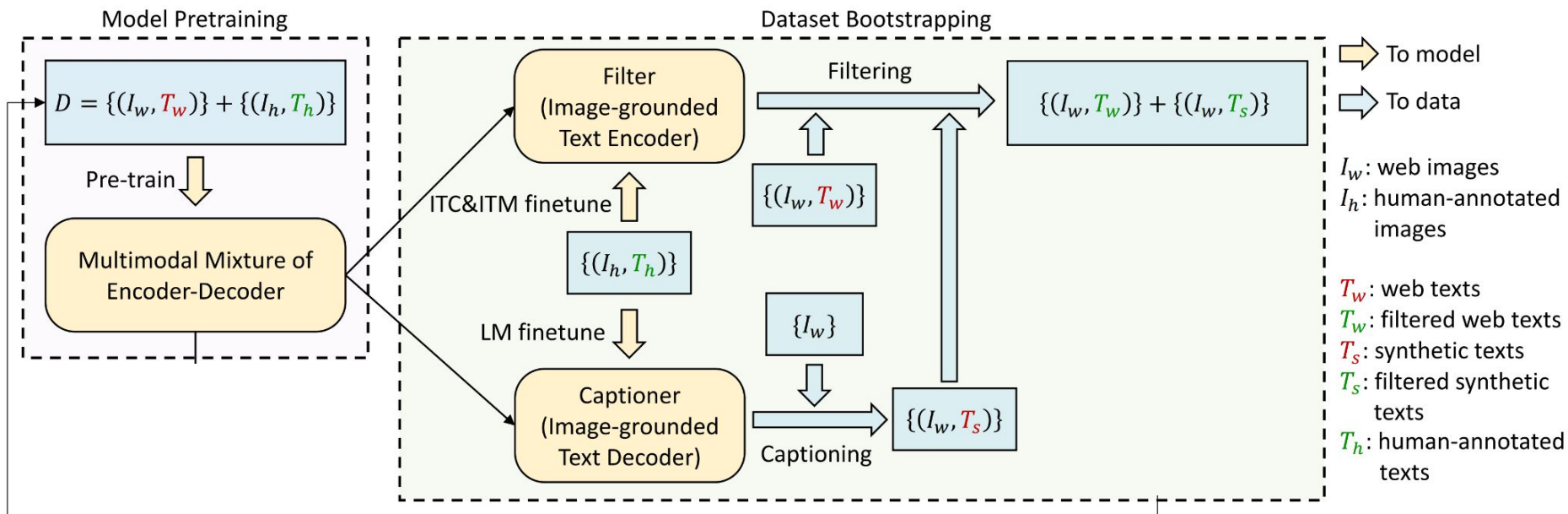
Captioner Finetuning



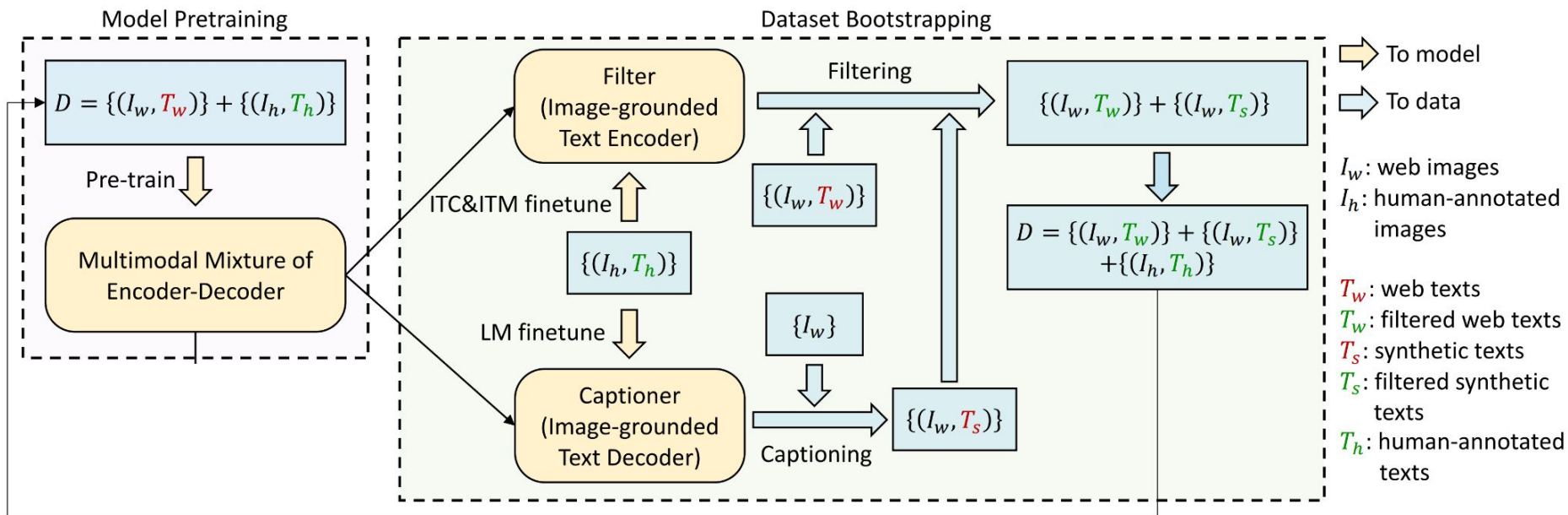
Caption Generation



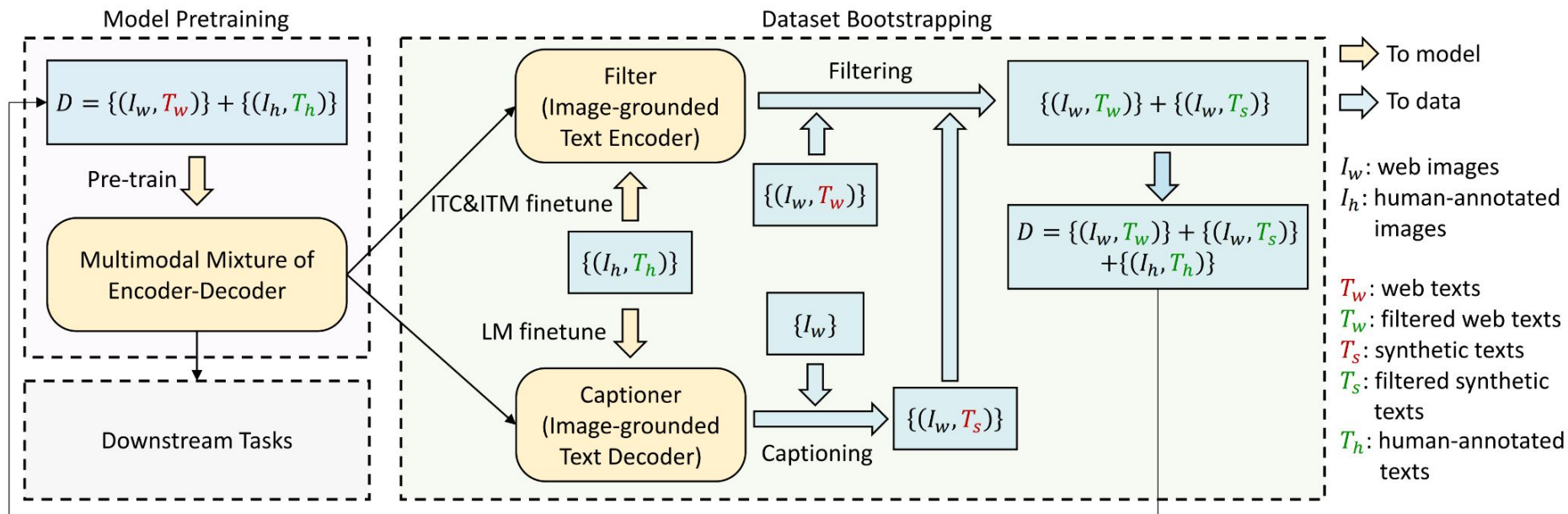
Filtering Synthetic Captions



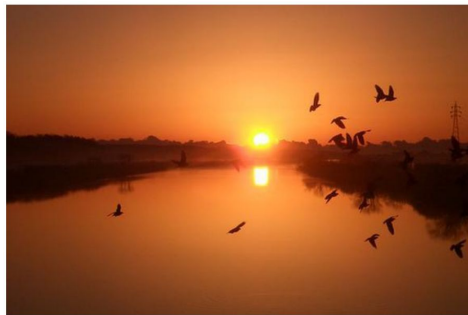
Bootstrapped Data



Improved Downstream Task Performance



Examples



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 : “a week spent at our rented beach house in Sandbridge”

T_S : “an outdoor walkway on a grass covered hill”



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”

Datasets

- Pre-training
 - 3 web based ~14M
 - 2 human annotated ~450K
 - An additional noisy web based ~115M
- Finetuning
 - COCO
- Additional experiment datasets
 - VQA, NLVR², MSRVT/MSVD - QA, NoCap

Results

Evaluation Metrics

- Recall at K (R@K) $recall@k = \frac{\text{number of recommended relevant items among top k}}{\text{number of all relevant items in the system}}$

- Mean Rank (MR) $MR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \text{rank}_i$

- Mean Reciprocal Rank (MRR) $MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$

- Bilingual Evaluation Understudy (BLEU or B@4) $BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$ $BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right)$

- Consensus-Based Image Description Evaluation (CIDEr) $CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\| \|\mathbf{g}^n(s_{ij})\|}$ $CIDEr(c_i, S_i) = \sum_{n=1}^N w_n CIDEr_n(c_i, S_i)$

- Semantic Propositional Image Caption Evaluation (SPICE) $P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|}$ $R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$ $SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)}$

Comparative Study: Image-Text Retrieval (Finetuned)

Method	Pre-train # Images	COCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
UNITER (Chen et al., 2020)	4M	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VILLA (Gan 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
OSCAR (Li et al., 2020)	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
UNIMO (Li et al., 2021b)	5.7M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	1.8B	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
ALBEF (Li et al., 2021a)	14M	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
		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
BLIP _{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
BLIP _{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

Comparative Study: Image-Text Retrieval (Zero-shot)

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
BLIP _{CapFilt-L}	129M	96.0	99.9	100.0	85.5	96.8	98.7
BLIP _{ViT-L}	129M	96.7	100.0	100.0	86.7	97.3	98.7

Comparative Study: Image Captioning (Finetuned)

Method	Pre-train #Images	NoCaps validation								COCO Caption Karpathy test	
		in-domain		near-domain		out-domain		overall		B@4	C
		C	S	C	S	C	S	C	S		
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
LEMON _{base} [†] (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
LEMON _{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
BLIP _{CapFilt-L}	129M	111.8	14.9	108.6	14.8	111.5	14.2	109.6	14.7	39.7	133.3
LEMON _{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
SimVLM _{huge} (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP _{ViT-L}	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

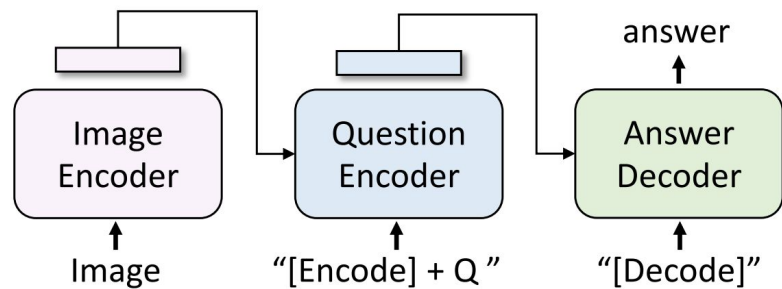
Visual Question Answering Task (VQA)

Goal

- Predict an answer given a image and a question

Setup

- Rearrange the pre-trained model
- Image-Question encoded into multimodal embeddings
- Embeddings are passed to an Answer Decoder
- VQA model is tuned using LM loss



Comparative Study: Visual Question Answering

Method	Pre-train #Images	VQA	
		test-dev	test-std
LXMERT	180K	72.42	72.54
UNITER	4M	72.70	72.91
VL-T5/BART	180K	-	71.3
OSCAR	4M	73.16	73.44
SOHO	219K	73.25	73.47
VILLA	4M	73.59	73.67
UNIMO	5.6M	75.06	75.27
ALBEF	14M	75.84	76.04
SimVLM _{base} [†]	1.8B	77.87	78.14
BLIP	14M	77.54	77.62
BLIP	129M	78.24	78.17
BLIP _{CapFilt-L}	129M	78.25	78.32

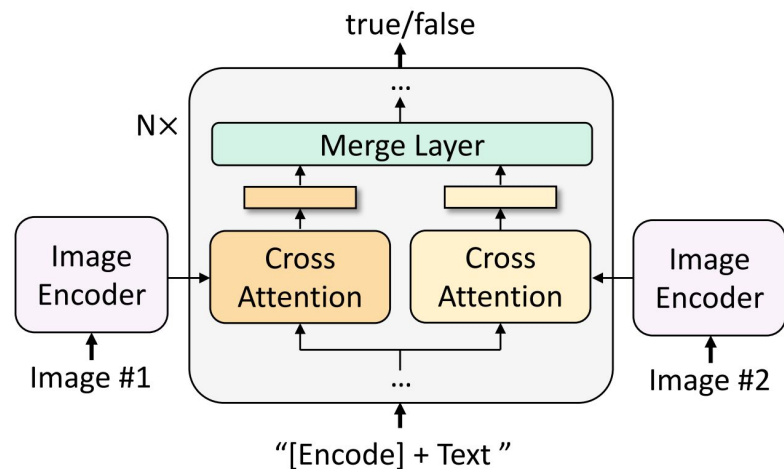
Natural Language Visual Reasoning Task (NLVR²)

Goal

- Predict if a sentence describes an image pair

Setup

- Rearrange the pre-trained model
- Pass both images to Image Encoders
- Pass embeddings to two cross-attention layers
- Outputs are merged
- Outputs fed to each blocks feed forward network



Comparative Study: Natural Language Visual Reasoning

Method	Pre-train #Images	NLVR ²	
		dev	test-P
LXMERT	180K	74.90	74.50
UNITER	4M	77.18	77.85
VL-T5/BART	180K	-	73.6
OSCAR	4M	78.07	78.36
SOHO	219K	76.37	77.32
VILLA	4M	78.39	79.30
UNIMO	5.6M	-	-
ALBEF	14M	82.55	83.14
SimVLM _{base} [†]	1.8B	81.72	81.77
BLIP	14M	82.67	82.30
BLIP	129M	82.48	83.08
BLIP _{CapFilt-L}	129M	82.15	82.24

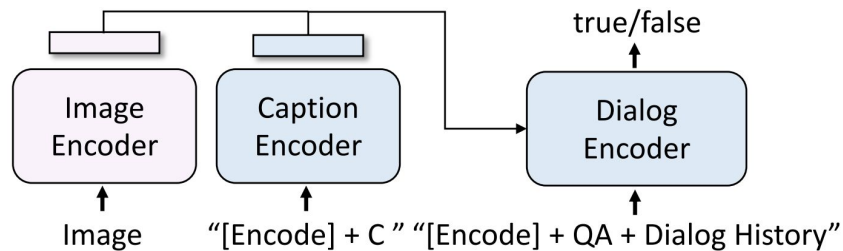
Visual Dialog Task (VisDial)

Goal

- Predict an answer using question-answer pair, dialog history, and the image's caption

Setup

- Rearrange the pre-trained model
- Concatenate image and caption embeddings
- Pass embeddings to a Dialog Encoder using CA
- Train the dialog encoder with IRM loss



Comparative Study: Visual Dialog

Method	MRR \uparrow	R@1 \uparrow	R@5 \uparrow	R@10 \uparrow	MR \downarrow
VD-BERT	67.44	54.02	83.96	92.33	3.53
VD-ViLBERT \dagger	69.10	55.88	85.50	93.29	3.25
BLIP	69.41	56.44	85.90	93.30	3.20

Comparative Study: Text-To-Video Retrieval (Zero-shot)

Method	R1↑	R5↑	R10↑	MdR↓
<i>zero-shot</i>				
ActBERT (Zhu & Yang, 2020)	8.6	23.4	33.1	36
SupportSet (Patrick et al., 2021)	8.7	23.0	31.1	31
MIL-NCE (Miech et al., 2020)	9.9	24.0	32.4	29.5
VideoCLIP (Xu et al., 2021)	10.4	22.2	30.0	-
FiT (Bain et al., 2021)	18.7	39.5	51.6	10
ALPRO (Li et al., 2022)	24.1	44.7	55.4	8
BLIP	43.3	65.6	74.7	2
<i>finetuning</i>				
ClipBERT (Lei et al., 2021)	22.0	46.8	59.9	6
VideoCLIP (Xu et al., 2021)	30.9	55.4	66.8	-
ALPRO (Li et al., 2022)	33.9	60.7	73.2	3

Comparative Study: Video Question Answering (Zero-Shot)

Method	MSRVTT-QA	MSVD-QA
<i>zero-shot</i>		
VQA-T (Yang et al., 2021)	2.9	7.5
BLIP	19.2	35.2
<i>finetuning</i>		
HME (Fan et al., 2019)	33.0	33.7
HCRN (Le et al., 2020)	35.6	36.1
VQA-T (Yang et al., 2021)	41.5	46.3
ALPRO (Li et al., 2022)	42.1	45.9

Data Bootstrapping with CapFilt

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

Pre-train dataset	Bootstrap		Vision backbone	Retrieval-FT (COCO)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
	C	F		TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	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
(129M imgs)	✗	✗	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

Ablation Study: Beam Search vs Nucleus Sampling

Generation method	Noise ratio	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
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

Ablation Study: Encoder/Decoder 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

Ablation Study: Captioner/Filter Parameter Sharing

Captioner & Filter	Noise ratio	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
Share parameters	8%	79.8	62.2	94.3	83.7	38.4	129.0	103.5	14.2
Decoupled	25%	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

Limitations

- Higher Computational Costs
- Residual noise limiting the downstream models
- Annotation limitations

Conclusion & Future Work

- CapFilt improves dataset quality, boosting downstream task performance
- Versatile model, excels in visual-language understanding and generation

- Explore multiple bootstrapping rounds to refine CapFilt
- Generate multiple captions per image for richer data
- Create a larger ensemble of models to boost CapFilt's performance

References

- Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation. In NeurIPS, 2021a. <https://arxiv.org/abs/2107.07651>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.
- Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. SimVLM: Simple visual language model pre training with weak supervision. arXiv preprint arXiv:2108.10904, 2021.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. Training data-efficient image trans-formers & distillation through attention. arXiv preprint arXiv:2012.12877, 2020.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.
- Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
- Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer International Publishing, 2016.
- Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
- Efimov, V. (2023, December 13). Comprehensive Guide to Ranking Evaluation Metrics - towards Data Science. Medium. <https://towardsdatascience.com/comprehensive-guide-to-ranking-evaluation-metrics-7d10382c1025>