

BLIP



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

Arzhantsev Andrei, 202

Vision-language tasks

- Image-Text Retrieval
- Image Captioning
- Visual Question Answering
- Visual Dialog
- Natural Language Visual Reasoning
- Memes explanation


NLVR2

 <p><i>Kropsoq (CC BY-SA 3.0); subhv150 (Pixabay)</i></p>	<p><i>Two hot air balloons are predominantly red and have baskets for passengers.</i></p>	<p>True</p>
 <p><i>babasteve (CC BY 2.0); Yathin S Krishnappa (CC BY-SA 3.0)</i></p>	<p><i>All elephants have ivory tusks.</i></p>	<p>False</p>



VQA

VisDial



Caption: A man and woman on bicycles are looking at a map.

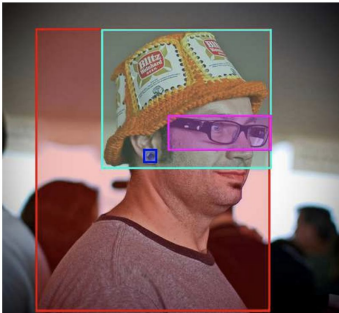
Person A (1): where are they located
 Person B (1): in city
 Person A (2): are they on road
 Person B (2): sidewalk next to 1
 Person A (3): any vehicles
 Person B (3): 1 in background
 Person A (4): any other people
 Person B (4): no
 Person A (5): what color bikes
 Person B (5): 1 silver and 1 yellow
 Person A (6): do they look old or new
 Person B (6): new bikes
 Person A (7): any buildings
 Person B (7): yes
 Person A (8): what color
 Person B (8): brick
 Person A (9): are they tall or short
 Person B (9): i can't see enough of them to tell
 Person A (10): do they look like couple
 Person B (10): they are

NoCaps



1. Two hardcover **books** are on the **table**
2. Two magazines are sitting on a **coffee table**.
3. Two **books** and many crafting supplies are on this **table**.
4. a recipe **book** and sewing **book** on a craft **table**
5. Two hardcover **books** are laying on a **table**.

Flickr30k



A man with **pierced ears** is wearing **glasses** and an **orange hat**.
 A man with **glasses** is wearing a beer can **crocheted hat**.
 A man with **gauges** and **glasses** is wearing a **Blitz hat**.
 A man in an **orange hat** starring at **something**.
 A man wears an **orange hat** and **glasses**.

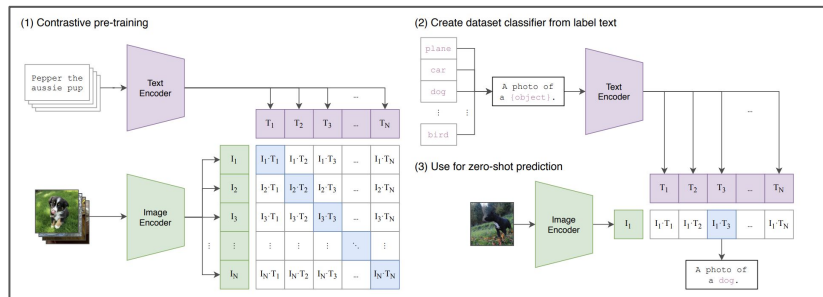
Model perspective

- Encoder-based models

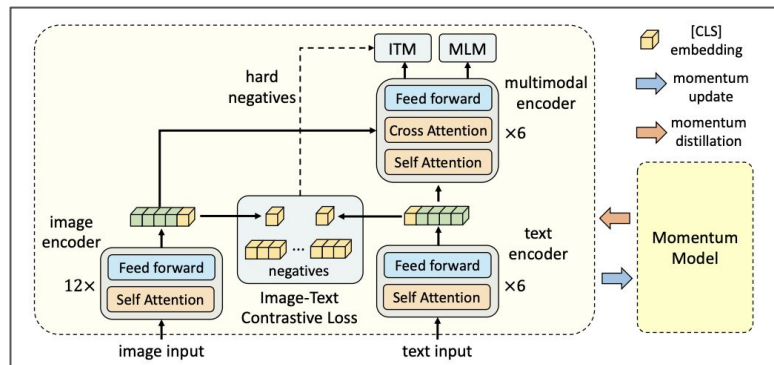
- CLIP: <https://arxiv.org/pdf/2103.00020.pdf>
- ALBEF: <https://arxiv.org/pdf/2107.07651.pdf>
- not optimal for text generation tasks

- Encoder-Decoder models

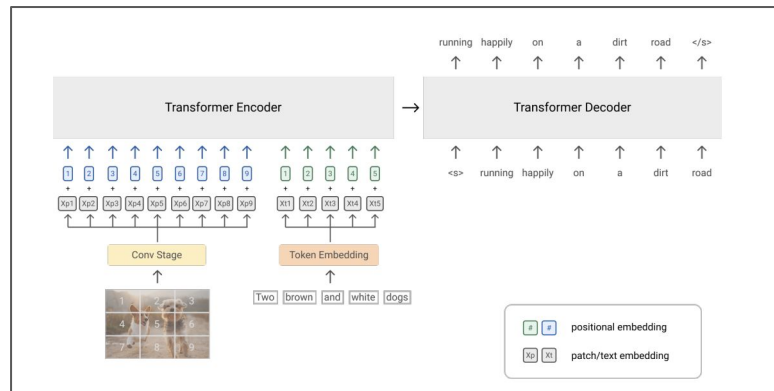
- VL-T5: <https://arxiv.org/pdf/2102.02779.pdf>
- SimVLM: <https://arxiv.org/pdf/2108.10904.pdf>
- not optimal for image-text retrieval tasks



CLIP



ALBEF



SimVLM

Data perspective

- Pre-train on relevant image-text pairs
 - few data, noisy data is bad
- Knowledge Distillation
 - <https://arxiv.org/pdf/1503.02531.pdf>
 - (ALBEF) Momentum Distillation
 - performs worse than CapFILT used in BLIP
- Data Augmentation
 - well known for CV tasks
 - recently used for NLP tasks
 - no analogue for VL tasks

[MASK] and the fiancée at their engagement party



GT: person
Top-5 pseudo-targets:
1. husband
2. person
3. me
4. actor
5. boyfriend

kitten playing with a [MASK]



GT: dog
Top-5 pseudo-targets:
1. toy
2. blanket
3. ball
4. mouse
5. bone

[MASK] at the guesthouse or nearby



GT: animal
Top-5 pseudo-targets:
1. fish
2. animal
3. animals
4. wildlife
5. food

[MASK] clouds in the sky



GT: red
Top-5 pseudo-targets:
1. pink
2. colorful
3. sunset
4. red
5. dramatic

Momentum distillation in ALBEF

BLIP: Captioning and Filtering (CapFilt)

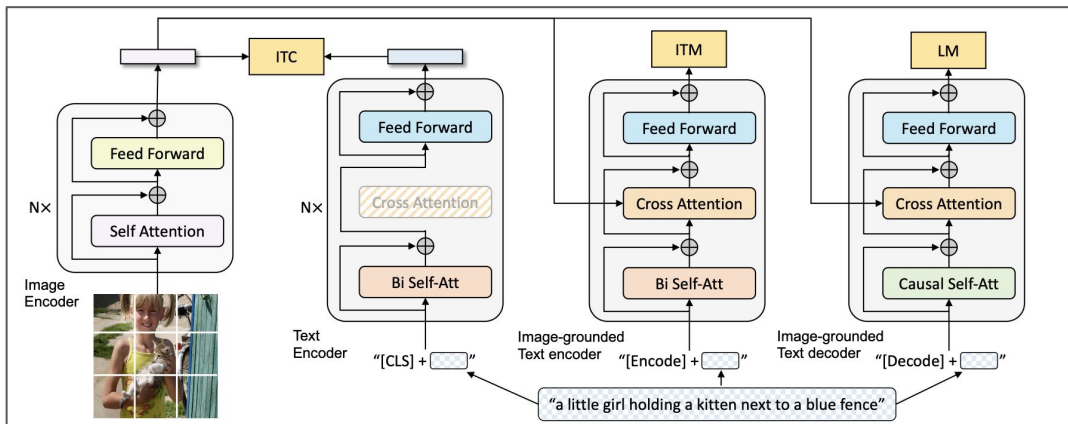
- Human-annotated Image-Text pairs (Ih-Th)
- Large dataset of web Image-Text pairs (Iw-Tw)
- Captioner
 - Generates captions (T_s) given web images (Iw)
 - Initialized from pre-trained MED model
 - Image-grounded text decoder
 - Finetuned with the LM objective (from Ih-Th)
- Filter
 - Removes noisy image-text pairs
 - Initialized from pre-trained MED model
 - Image-grounded text encoder
 - finetuned with the ITC (contrastive) and ITM (matching) objectives to learn whether a text matches an image
- With CapFilt we have
 - Ih-Th and new filtered Iw-Tw or Iw-Ts



Filtering from both web texts and synthetic texts

BLIP: Multimodal mixture of Encoder-Decoder (MED)

- Unimodal encoder
 - separately encodes image and text
 - Text encoder same as BERT
- Image-grounded text encoder
 - injects visual information
 - one additional CA between SA and FF
- Image-grounded text decoder
 - generates captions given images
 - uses casual SA



Pre-training model architecture

BLIP: Multimodal mixture of Encoder-Decoder (MED)

- Image-Text
Contrastive Loss
(ITC)

- aligns feature space of the visual transformer and text transformer
- learns a similarity function

$$s(I, T) = g_v(\mathbf{v}_{\text{cls}})^\top g'_w(\mathbf{w}'_{\text{cls}}) \text{ and } s(T, I) = g_w(\mathbf{w}_{\text{cls}})^\top g'_v(\mathbf{v}'_{\text{cls}}).$$

$$p_m^{\text{i2t}}(I) = \frac{\exp(s(I, T_m)/\tau)}{\sum_{m=1}^M \exp(s(I, T_m)/\tau)}, \quad p_m^{\text{t2i}}(T) = \frac{\exp(s(T, I_m)/\tau)}{\sum_{m=1}^M \exp(s(T, I_m)/\tau)}$$

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

- Image-Text
Matching Loss (ITM)

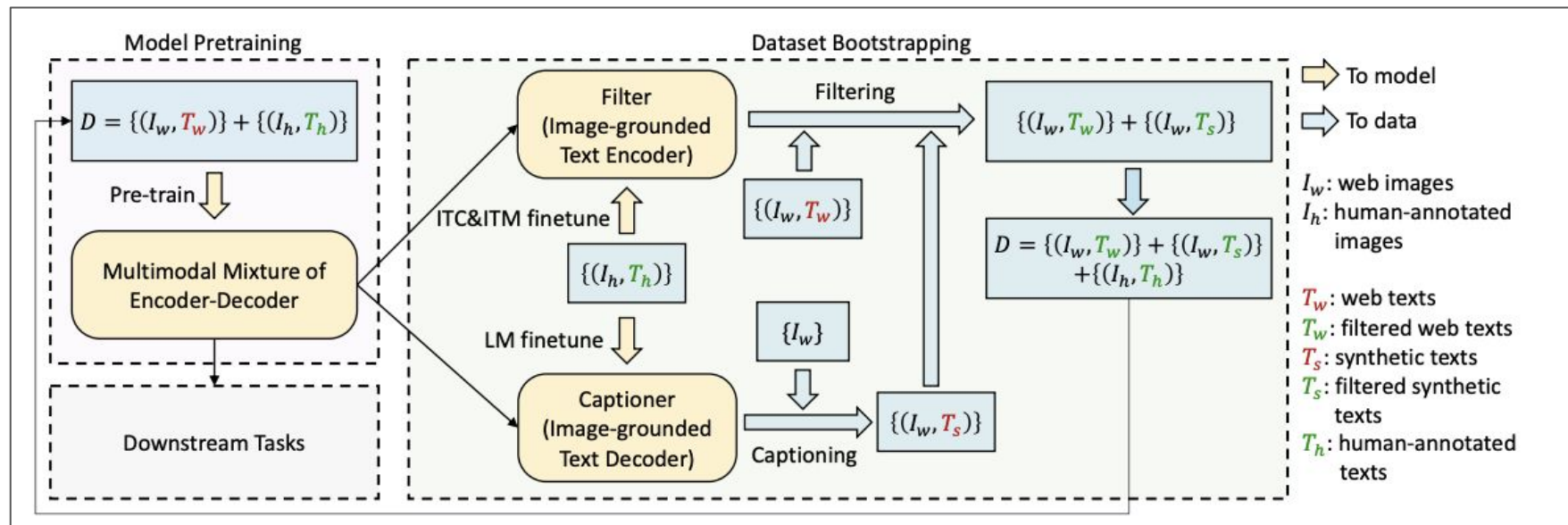
- learns image-text multimodal representation
- binary classification task

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

- Language Modeling
Loss (LM)

- cross entropy loss
- aims to generate textual descriptions given an image
- trains the model to maximize the likelihood of the text in an autoregressive manner

BLIP: Full schema



Effect of CapFilter

The efficacy of CapFilter on downstream tasks,
including image-text retrieval and image captioning

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
	✗	✗	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

Experiments

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

Image captioning

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
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

Image-text retrieval

Method	Pre-train #Images	VQA		NLVR ²	
		test-dev	test-std	dev	test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	-
ALBEF	14M	75.84	76.04	82.55	83.14
SimVLM _{base} †	1.8B	77.87	78.14	81.72	81.77
BLIP	14M	77.54	77.62	82.67	82.30
BLIP	129M	78.24	78.17	82.48	83.08
BLIP _{CapFilt-L}	129M	78.25	78.32	82.15	82.24

VQA and NLVR

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

Visual Dialog

and even more



BLIP: potential directions for future

1. Multiple rounds of dataset bootstrapping
2. Generate multiple synthetic captions per image to further enlarge the pre-training corpus
3. Model ensemble by training multiple different captioners and filters and combining their forces in CapFilt

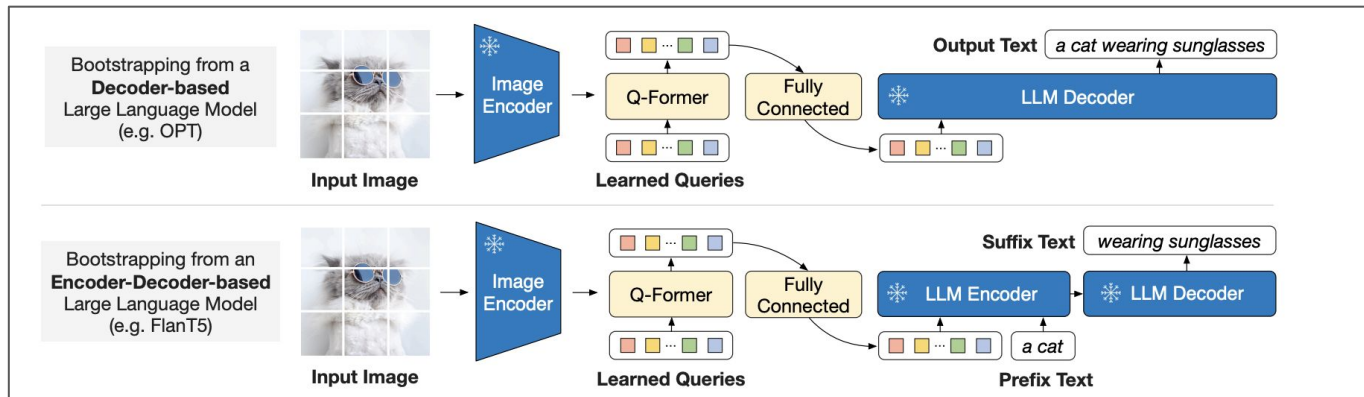
BLIP-2 (June 2023)

- Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

<https://arxiv.org/pdf/2301.12597.pdf>

- Novelties:

- Q-Former - trainable module to connect a frozen image encoder and a frozen LLM
- new objective in model pretraining: Image-grounded Text Generation
- ...



BLIP-2's second-stage vision-to-language generative pre-training

That's all



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”



T_w : “stunning sky over walney island, lake district, july 2009”

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



T_w : “living in my little white house”

T_s : “a tiny white flower with a bee in it”



T_w : “the pink rock from below”

T_s : “some colorful trees that are on a hill in the mountains”

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
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	-

Table 10. Comparisons with state-of-the-art methods for text-to-video retrieval on the 1k test split of the MSRVT dataset.

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

Table 11. Comparisons with state-of-the-art methods for video question answering. We report top-1 test accuracy on two datasets.

Not contained examples and experiment tables