Data Filtering in Vision-Language Pre-training

Pu Yang
School of Mathematical Sciences, Peking University

2023.10.18

Table of contents

1. Introduction
2. Data filtering
   - Momentum Distillation
   - Caption-Filtering
3. Summary

Vision-Language Pre-training

- Goal: develop AI systems that can understand and reason about visual concepts and language in an interconnected way.
- Various downstream vision-language tasks:
  - text generation (i.e. image captioning, visual question answering)
  - image generation (i.e. style transfer)
  - image analysis (i.e. segmentation)

Main Research Directions

- Model and Architecture
- Task and Objective Function
- **Data**
- Training Strategy

Model and Architecture

- Architecture: Transformer, ViT
- Model

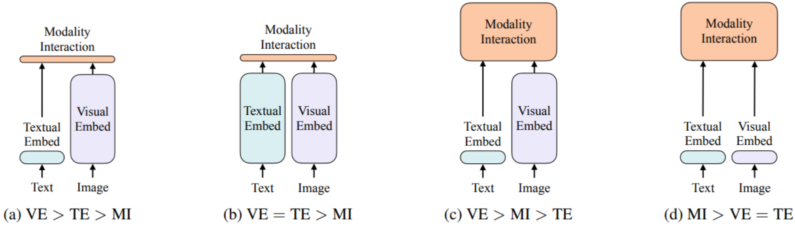


Figure 1: Four categories of vision-and-language models.[1]

It turns out that (c) is the best!

- Other tricks: Mixture of Experts (MoE)[2]

[1] Wonjae Kim, Bokyung Son, and Ildoo Kim. "Vilt: Vision-and-language transformer without convolution or region supervision". In: International Conference on Machine Learning. PMLR. 2021, pp. 5583–5594.

[2] Robert A Jacobs et al. "Adaptive mixtures of local experts". In: Neural computation 3.1 (1991), pp. 79–87.

Model and Architecture

An example:

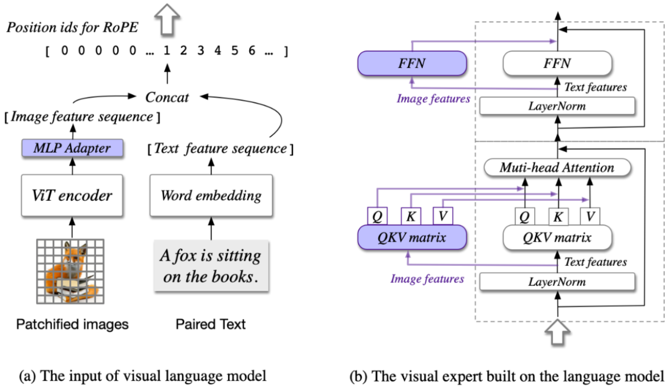


Figure 2: The architecture of CogVLM. `https://github.com/THUDM/CogVLM`

Task and Objective Function

- Image-text contrastive (ITC): CLIP
- Object detection (OD): ViLBERT, UNITER
- Image-text matching (ITM): ViLBERT, UNITER, ViLT
- Mask language modeling (MLM): BERT
- Predict the next ~~word~~ token: GPT

Recent methods prefer the task of predicting the next token!

Task and Objective Function

An example:

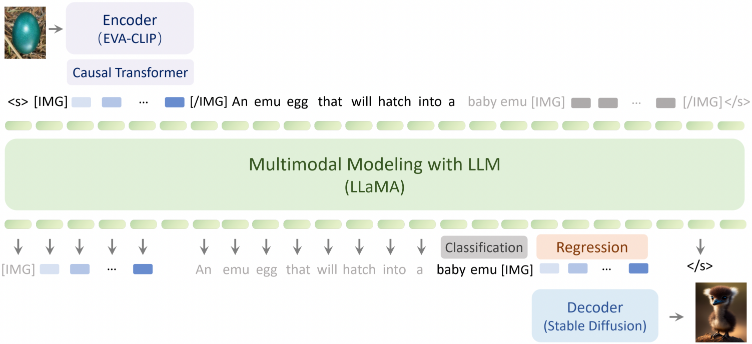


Figure 3: Emu unifies the modeling of different modalities in an auto-regressive manner.[3]

[3] Quan Sun et al. "Generative pretraining in multimodality". In: arXiv preprint arXiv:2307.05222 (2023).

Data

- Data mining - Discovering and extracting new image-text data from multimodal sources like the web, books, social media etc.
- **Data filtering** - Developing robust methods to clean noisy web data and retain useful training examples.
- Data augmentation - Techniques like text and image augmentation and synthesis to increase diversity and generalizability.
- Balanced sampling - Strategies to ensure models see diverse, representative data and avoid biases.

Training Strategy

- Multi-stage
- End-to-end

An example:

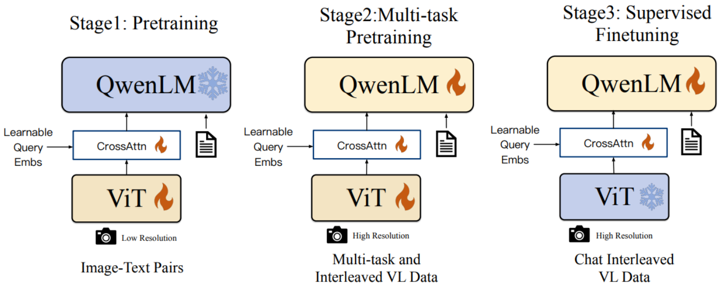


Figure 4: The training pipeline of the Qwen-VL series.[4]

[4] Jinze Bai et al. "Qwen-vl: A frontier large vision-language model with versatile abilities". In: arXiv preprint arXiv:2308.12966 (2023).

Align before Fuse (ALBEF)

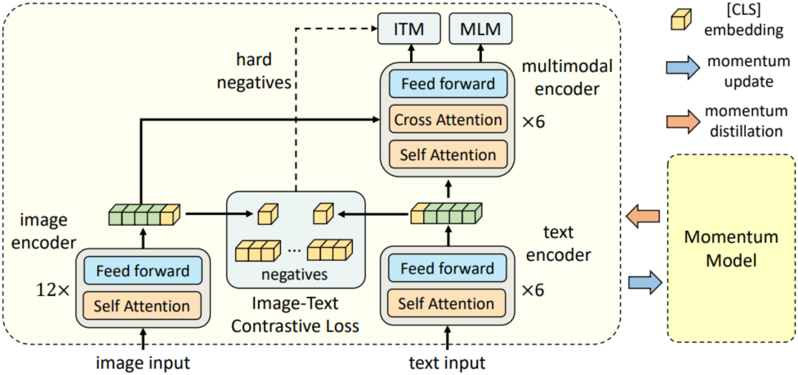


Figure 5: Illustration of ALBEF.[5]

[5] Junnan Li et al. "Align before fuse: Vision and language representation learning with momentum distillation". In: Advances in neural information processing systems 34 (2021), pp. 9694–9705.

Pre-training Objectives

- Image-Text Contrastive Learning (the same as Moco)
  Let g_I and g_T are linear transformations that map the [CLS] embeddings to normalized representations, and g_I' and g_T' are representations from the momentum encoders. We define the similarity

$$s(I,T) = g_I^\top \cdot g_T', \quad s(T,I) = g_T^\top \cdot g_I'$$

  and the softmax-normalized similarity

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

  Thus, the loss function is

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

  where $\boldsymbol{y}$ is the ground-truth one-hot similarity, and H denotes the cross-entropy.

Pre-training Objectives

- Masked Language Modeling (same as BERT)

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{(I,\hat{T})\sim D} H(\boldsymbol{y}^{\text{msk}}, \boldsymbol{p}^{\text{msk}}(I,\hat{T}))$$

  where $\hat{T}$ denotes a masked text, and $\boldsymbol{p}^{\text{msk}}(I,\hat{T})$ denotes the model's predicted probability for a masked token.

- Image-Text Matching (same as binary classification)

$$\mathcal{L}_{\text{itm}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D} H(\boldsymbol{y}^{\text{itm}}, \boldsymbol{p}^{\text{itm}}(I,T))$$

  where $p^{\text{itm}}(I,T)$ is the predicted two-class probability.

The image-text pairs are noisy

Positive pairs are usually weakly-correlated

- For ITC: negative texts for an image may also match the image's content.
- For MLM, there may exist other words different from the annotation that describes the image equally well (or better).

Momentum Distillation (MoD)

Learn from pseudo-targets generated by the momentum model.

- For ITC, We use the similarity from the momentum model

$$s'(I,T) = g_I'^{\top} \cdot g_T', \quad s'(T,I) = g_T'^{\top} \cdot g_I'$$

and the momentum model's softmax-normalized similarity

$$q_m^{\text{i2t}}(I) = \frac{\exp\left(s'\left(I, T_m\right)/\tau\right)}{\sum_{m=1}^{M} \exp\left(s'\left(I, T_m\right)/\tau\right)}, \quad q_m^{\text{t2i}}(T) = \frac{\exp\left(s'\left(T, I_m\right)/\tau\right)}{\sum_{m=1}^{M} \exp\left(s'\left(T, I_m\right)/\tau\right)}$$

Thus, the loss function is

$$\mathcal{L}_{\text{itc}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{itc}} + \frac{\alpha}{2}\mathbb{E}_{(I,T)\sim D}[\mathrm{KL}(\boldsymbol{q}^{\text{i2t}}(I)) \,||\, \boldsymbol{p}^{\text{i2t}}(I) + \mathrm{KL}(\boldsymbol{q}^{\text{t2i}}(T)) \,||\, \boldsymbol{p}^{\text{t2i}}(T)]$$

- For MLM

$$\mathcal{L}_{\text{mlm}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{mlm}} + \alpha \mathbb{E}_{(I,\hat{T})\sim D} \text{KL}(\boldsymbol{q}^{\text{msk}}(I,\hat{T}) \mid\mid \boldsymbol{p}^{\text{msk}}(I,\hat{T}))$$

where $\boldsymbol{q}^{\text{msk}}(I,\hat{T})$ denotes the momentum model's prediction probability for the masked token.

The author sets $\alpha = 0.4$.

Illustration

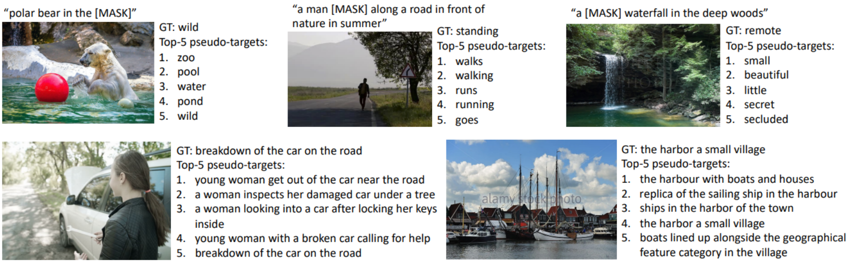


Figure 6: : Examples of the pseudo-targets for MLM (1st row) and ITC (2nd row). The pseudo-targets can capture visual concepts that are not described by the ground-truth text (e.g. "beautiful waterfall", "young woman").

Experiments on the proposed methods

#Pre-train Images	Training tasks	TR (flickr test)	IR (flickr test)	SNLI-VE (test)	NLVR^2 (test-P)	VQA (test-dev)
4M	MLM + ITM	93.96	88.55	77.06	77.51	71.40
	ITC + MLM + ITM	96.55	91.69	79.15	79.88	73.29
	ITC + MLM + ITM_{hard}	97.01	92.16	79.77	80.35	73.81
	ITC_{MoD} + MLM + ITM_{hard}	97.33	92.43	79.99	80.34	74.06
	Full (ITC_{MoD} + MLM_{MoD} + ITM_{hard})	97.47	92.58	80.12	80.44	74.42
	ALBEF (Full + $\text{MoD}_{\text{Downstream}}$)	97.83	92.65	80.30	80.50	74.54
14M	ALBEF	98.70	94.07	80.91	83.14	75.84

Three main improvements:

- Objective function
- Larger dataset
- MoD

Bootstrapping Language-Image Pre-training (BLIP)

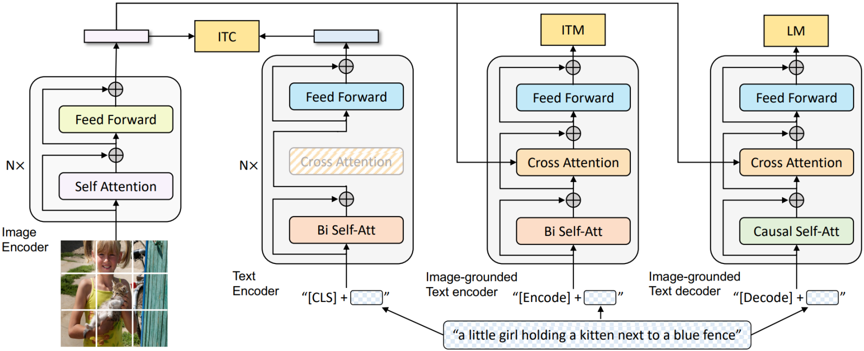


Figure 7: Illustration of BLIP.[6]

[6] Junnan Li et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: International Conference on Machine Learning. PMLR. 2022, pp. 12888–12900.

Dataset

- a limited number of high-quality human-annotated image-text pairs $\{(I_h, T_h)\}$, e.g., COCO 200K
- a much larger number of image and alt-text pairs collected from the web $\{(I_w, T_w)\}$, e.g. Conceptual Captions 12M, LAION 115M

Bootstrapping: Captioning and Filtering (CapFilt)

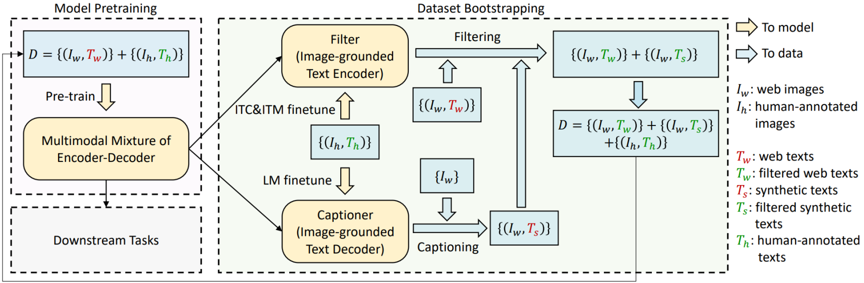


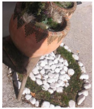
Figure 8: Learning framework of BLIP.

- Captioner is an image-grounded text decoder which generates synthetic captions given web images.
- Filter is an image-grounded text encoder which removes noisy image-text pairs.

Illustration

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"

Figure 9: 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.

Experiments

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
	✗	$\checkmark_B$		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	$\checkmark_B$	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	$\checkmark_B$	$\checkmark_B$		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
	$\checkmark_B$	$\checkmark_B$		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	$\checkmark_L$	$\checkmark_L$		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
	$\checkmark_L$	$\checkmark_L$		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. $\checkmark_{B/L}$: captioner or filter uses ViT-B / ViT-L as vision backbone.

Other Applications

- Generate synthetic caption for image data without text
  https://lambdalabs.com/blog/
  how-to-fine-tune-stable-diffusion-how-we-made-the-text-to-poke

Summary

Key idea: leverage **self-supervision signals** or **contrastive learning** to identify low quality or noisy samples and filter them out or reduce their impact during pre-training.