Vision-Language Dataset Distillation

Xindi Wu[1], Byron Zhang[1], Zhiwei Deng[2], Olga Russakovsky[1]
[1]Princeton University, [2]Google Research
xindiw@princeton.edu @cindy_x_wu

Princeton Visual AI Lab

Website / Arxiv / Code

Data is the cornerstone in multimodal ML

Flickr30K (2014) Visual Genome (2016) LAION 400M/5B (2021)

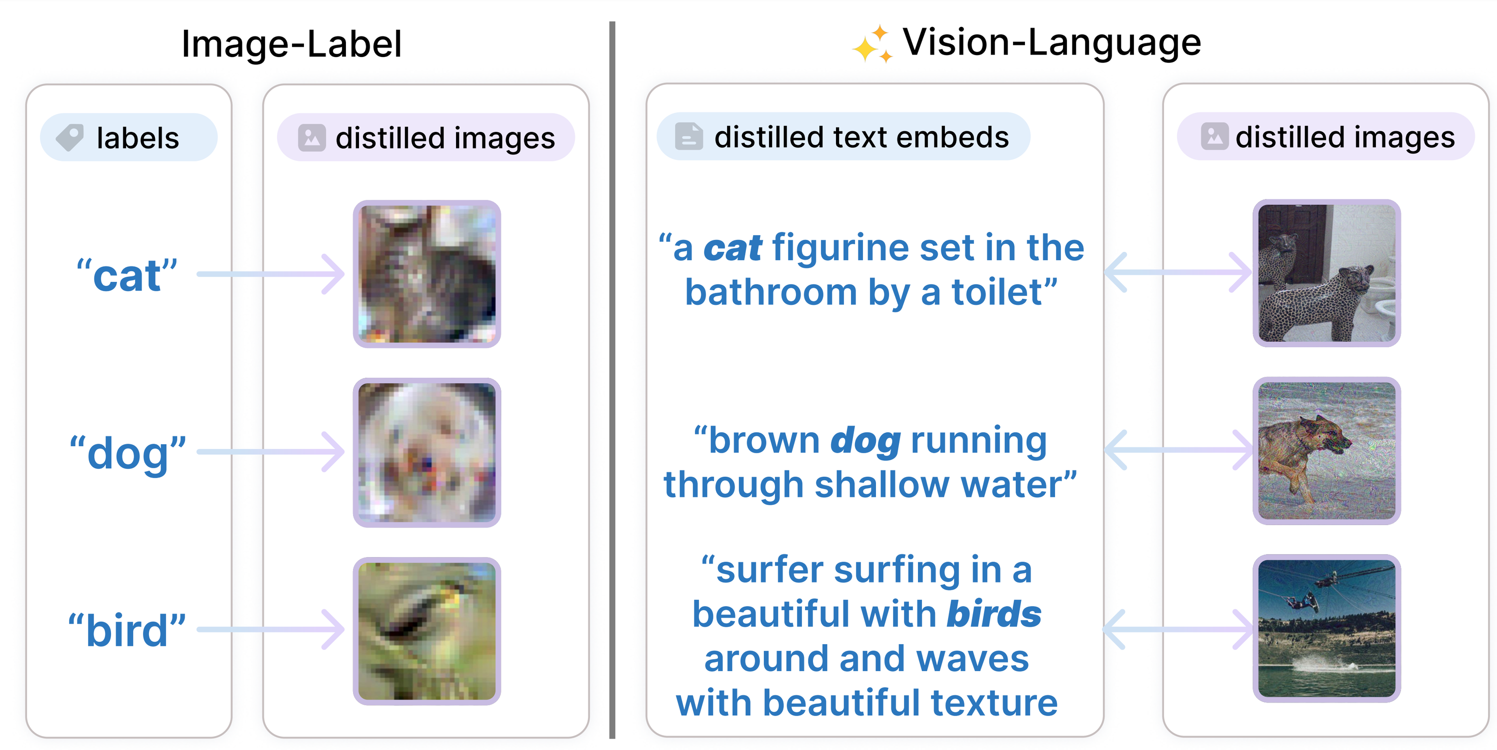
COCO (2014) Conceptual Captions (2018) DataComp (2023)

- Vision-language datasets have been growing increasingly large, reaching millions or even billions of samples.
- The vision-language pairs are often excessively noisy and complex.

Data = Information + Irrelevant Data [1]

Research Question

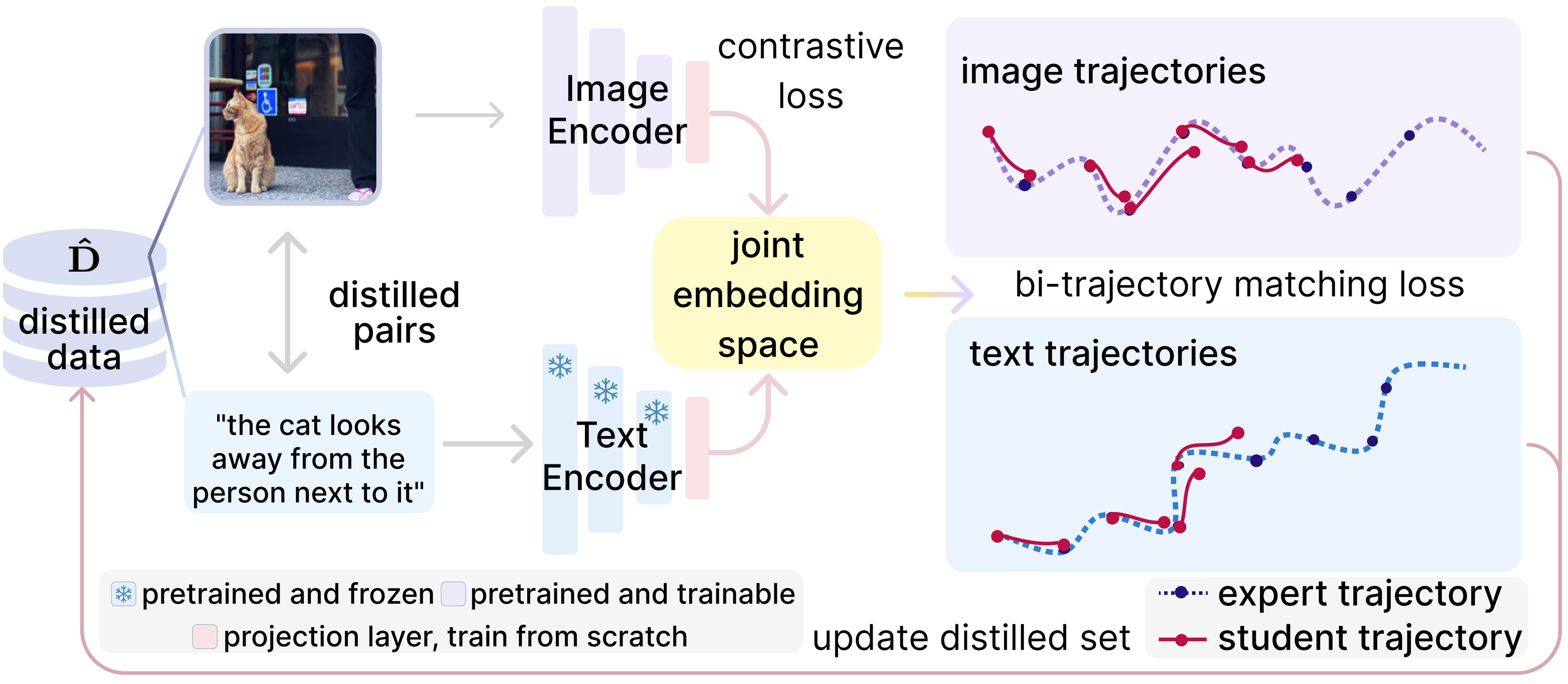
How can we distill the most critical information from vision-language datasets?



- Prior works distill each class separately [2, 3].
- We distill vision-language datasets that lack discrete classes.

Bi-trajectory Guided Vision-Language Co-Distillation

- Heavy computational cost



- **Bi-trajectory matching**: Separately considers two trajectories to capture complex vision-text interactions via contrastive loss.

$$\ell_{trajectory} = \frac{\|\hat{\theta}_{img,s+\hat{R}} - \theta^*_{img,s+R}\|_2^2}{\|\theta^*_{img,s} - \theta^*_{img,s+R}\|_2^2} + \frac{\|\hat{\theta}_{txt,s+\hat{R}} - \theta^*_{txt,s+R}\|_2^2}{\|\theta^*_{txt,s} - \theta^*_{txt,s+R}\|_2^2}$$

- **Low-rank adaptation matching**: makes it computationally feasible for training with more complex models (e.g., ViTs).
- **Text distillation**: use continuous sentence embeddings to overcome the difficulties of optimizing discrete text directly.

Stage 1 Expert training

Training multiple models for T epochs on the full dataset D. Obtaining expert training trajectories $\tau^* = \{\theta^*_t\}_{t=0}^{T}$.

Stage 2 Distillation

- Training student models on current distilled dataset $\hat{\mathbf{D}} = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{M}$ with contrastive loss.
- Update the current distilled dataset based on the **bi-trajectory matching loss** of the student models' parameter trajectories and the expert trajectories.

Results

- Baseline comparisons

(Here we only report R@1)

		TR				
		Coreset Selection				
Dataset	#pairs	R	H	K	F	Dist (ours)
Flickr30K	100	1.3	1.1	0.6	1.2	**9.9** ± 0.3
COCO	100	0.8	0.8	1.4	0.7	**2.5** ± 0.3

Random (R), Herding (H), K-center (K) Forgetting (F)

- Different vision encoders

Vision Model	TR	IR
NFNet	9.9	4.7
VIT_LoRA	**10.4**	**5.4**
NF_ResNet50	6.5	3.46
NF_RegNet	7.8	3.28

- With and without LoRA on ViT

		Without LoRA		With LoRA	
Dataset	#Pairs	TR	IR	TR	IR
Flickr30K	100	1.5	0.6	**10.4**	**5.4**
	1000	3.3	1.5	**15.8**	**8.1**

- Cross-architecture generalization

Distill	Evaluate	TR	IR
NFNet	NFNet	9.9	4.7
	NF-ResNet50	5.2	4.5
	NF-RegNet	3.6	2.5
	ViT	3.1	2.3

- Different language encoders

Language Model	TR	IR
BERT	9.9	4.7
CLIP	**31.4**	**17.1**

Distilled Examples & Ablations

Distilled examples:

a newly married couple sharing a kiss in front of a convertible | a couple kissed in front of a beautiful three-tiered cake with blue ribbon and pink accents | a man in a black wet suite is surfing a huge wave in the beautiful blue water | a man surfs over a huge wave in the ocean

Increasing learning rate will change images more noticeably in distilled datasets but doesn't lead to performance improvement.

- Single-modality vs. multi-modality

T: text-only, I: image-only

	TR	IR
T	1.3	0.5
I	3.5	1.6
Ours	9.9	4.7

Takeaway: Distillation would be impossible if we solely optimize one modality. image component plays a more critical role in the distilled dataset.

- Image-Text Pair Initialization

Real Image	Real Text	TR	IR
		0.4	0.1
	✓	1.1	0.1
✓		9	3.9
✓	✓	**9.9**	**4.7**

Takeaway:
✅ Initializing texts from scratch
❌ Initializing images from scratch

[1] Wright, John, and Yi Ma. High-dimensional data analysis with low-dimensional models: Principles, computation, and applications. Cambridge University Press, 2022.
[2] Cazenavette, George, et al. "Dataset distillation by matching training trajectories." CVPR 2022.
[3] Deng, Zhiwei, and Olga Russakovsky. "Remember the past: Distilling datasets into addressable memories for neural networks." NeurIPS 2022.

This material is based upon work supported by the National Science Foundation under Grant No. 2107048 and No. 2112562.