NeurIPS 2025

CovMatch: Cross-Covariance Guided Multimodal Dataset Distillation with Trainable Text Encoder

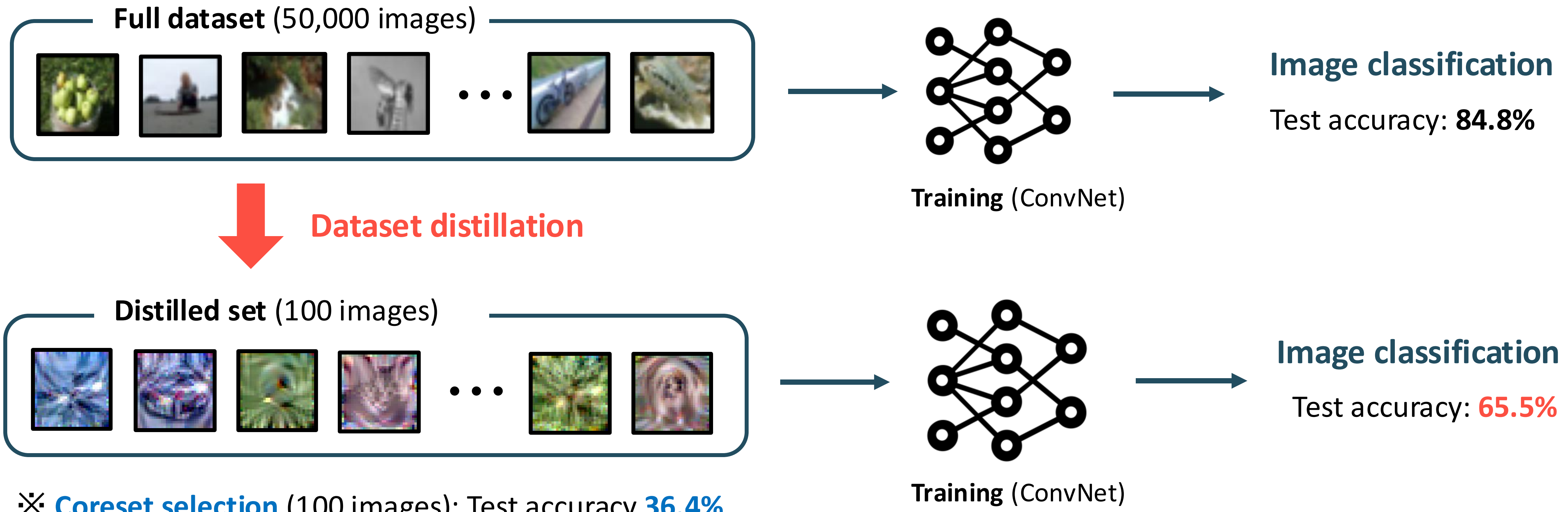
Yongmin Lee, Hye Won Chung

KAIST

Dataset Distillation

Synthesize a tiny dataset that captures the rich information encoded in the original dataset

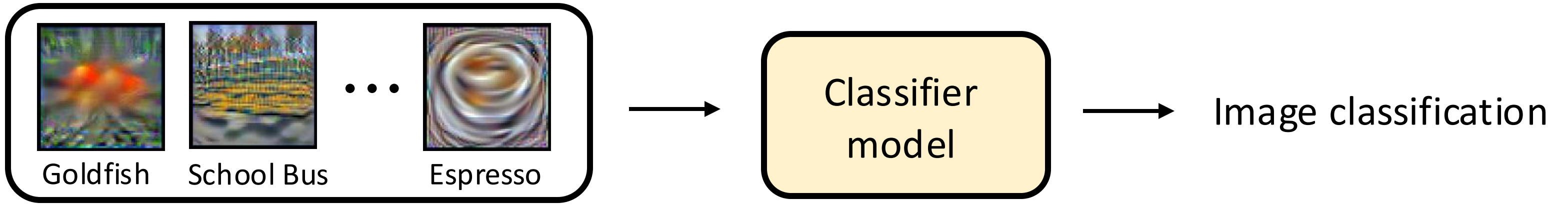
Example) CIFAR-10



※ **Coreset selection** (100 images): Test accuracy **36.4%**

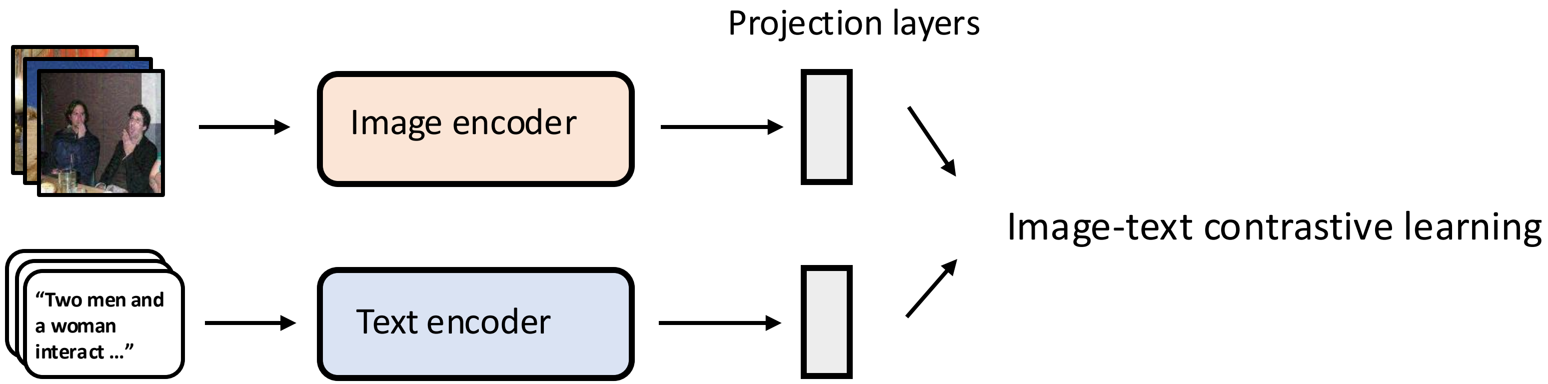
Extension to Multi-modal Dataset Distillation

Uni-modal dataset distillation



- Synthetic images that **best represent each class**.

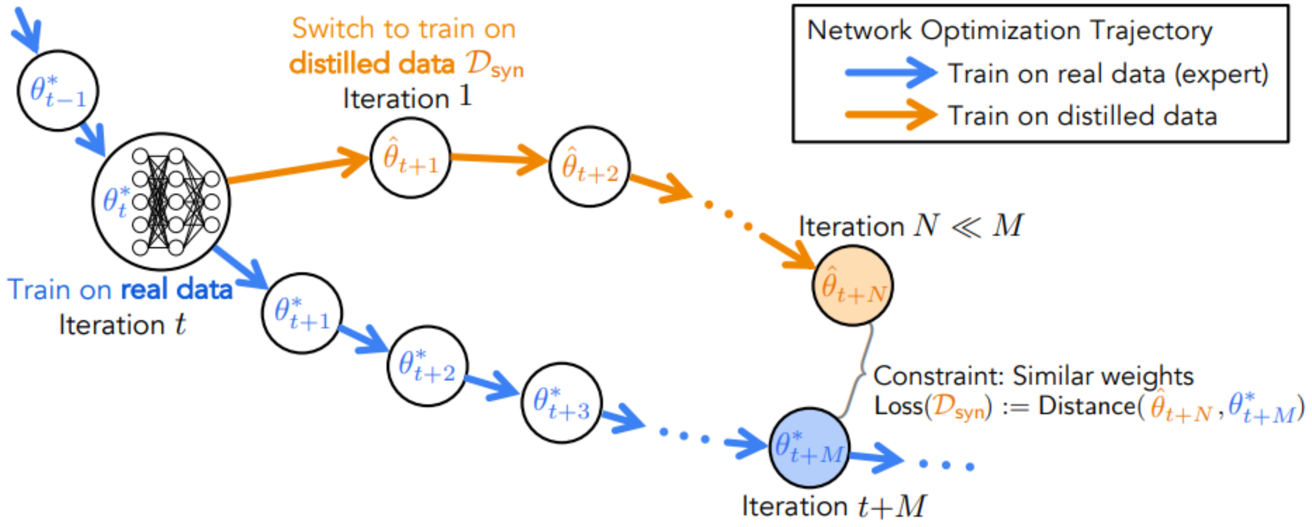
Vision-Language dataset distillation



- No class information.
- It is important to **learn cross-modal correspondences** between image and text data.

Previous Works

- Previous multimodal dataset distillation methods [1,2] rely on **training trajectories matching** approach.

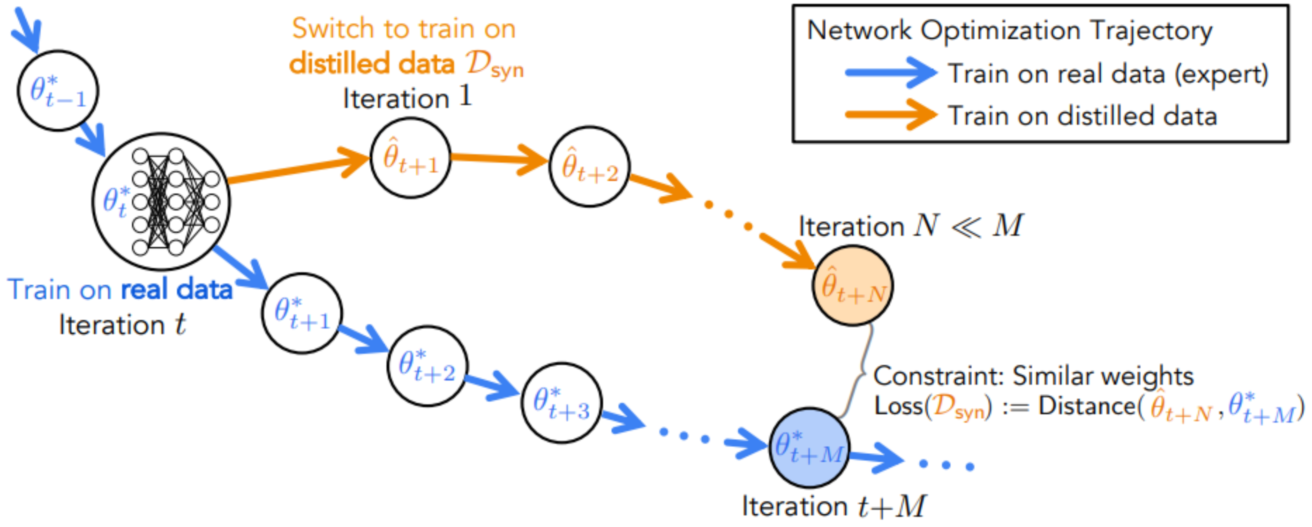


Matching Training Trajectories (MTT) [3]

1) Wu, Xindi, et al. "Vision-language dataset distillation." TMLR 2024.
2) Xu, Yue, et al. "Low-Rank Similarity Mining for Multimodal Dataset Distillation." ICML 2024.
3) George Cazenavette et al., Dataset Distillation by Matching Training Trajectories, CVPR 2022

Previous Works

- Previous multimodal dataset distillation methods [1,2] rely on **training trajectories matching** approach.



Matching Training Trajectories (MTT) [3]

Problem:

- It requires to precompute lots of expert training trajectories before distillation process.

1) Wu, Xindi, et al. "Vision-language dataset distillation." TMLR 2024.
2) Xu, Yue, et al. "Low-Rank Similarity Mining for Multimodal Dataset Distillation." ICML 2024.
3) George Cazenavette et al., Dataset Distillation by Matching Training Trajectories, CVPR 2022

Limitation of Previous Works

Uni-modal MTT

- ConvNet (1.24 MB)
- 1.9 hours and 120 GB for expert trajectories.
- Distillation time: 1.0 sec per iteration.

Multi-modal MTT

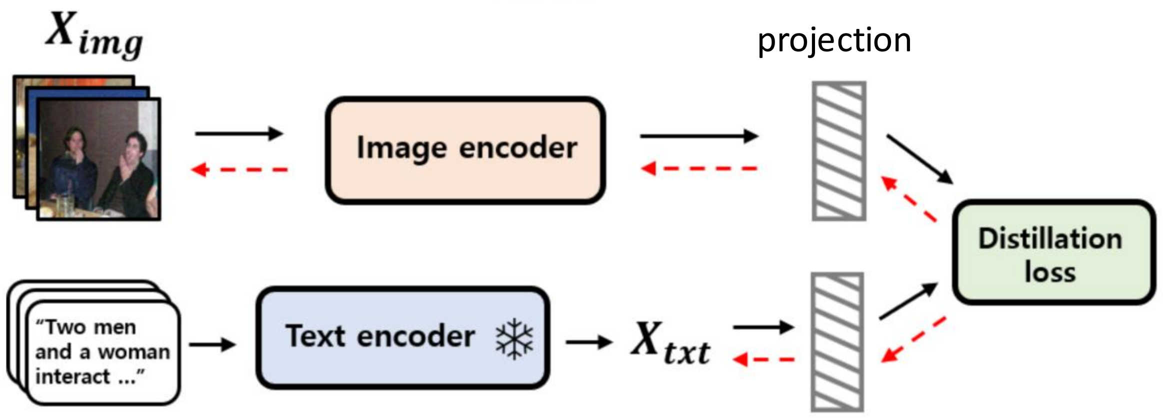
- Image encoder: NFNet (**140 MB**)
- Text encoder: BERT (**450 MB**)
- **132 hours and 120 GB** for expert trajectories.
- Distillation time: **16.9 sec** per iteration.

- More than 100x larger models.
- It takes **5 days and needs 120 GB** for preparing expert trajectories !
- Distillation time significantly increases.

Too much cost !!

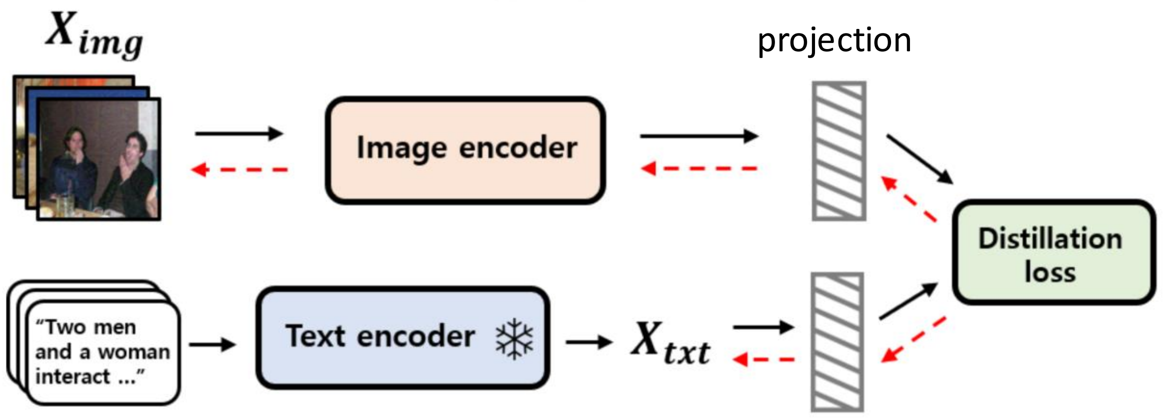
Limitation of Previous Works

- Previous works freeze the text encoder .

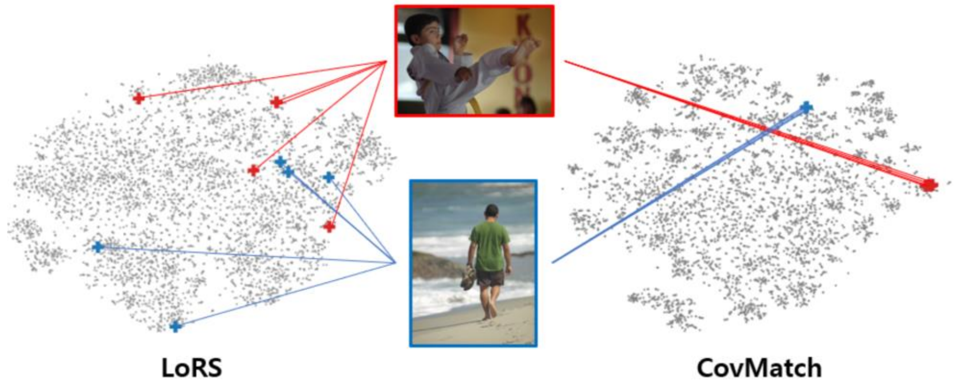


Limitation of Previous Works

- Previous works freeze the text encoder .



- It severely **limits the capacity for semantic alignment** in multimodal contrastive learning.

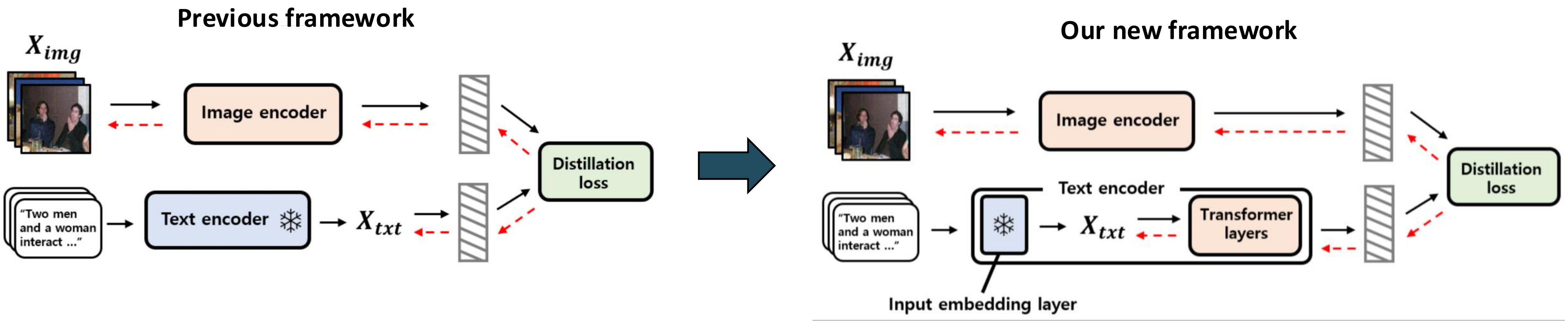


➢ Text embeddings corresponding to the same image.

➢ LoRS (previous SOTA): scattered

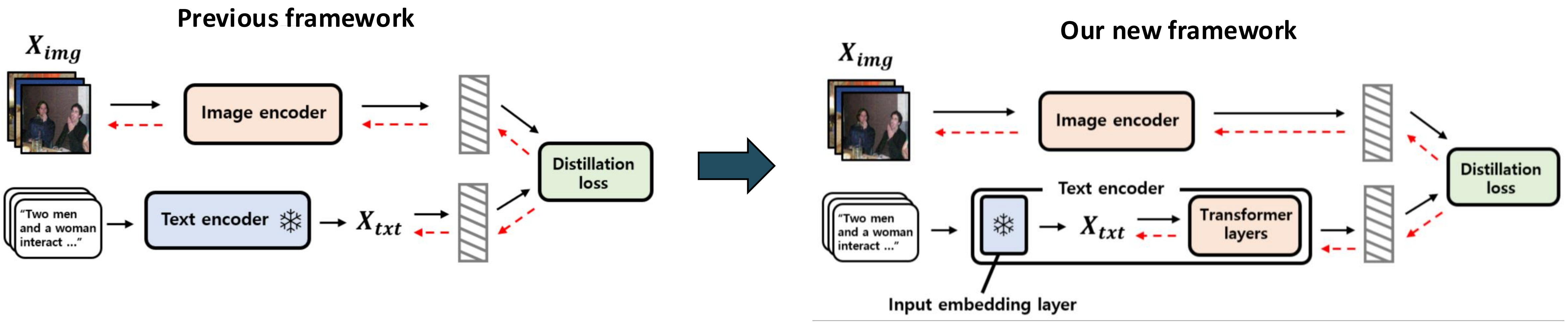
➢ CovMatch (ours): clustered

New Framework



- We freeze only input embedding layer and include the transformer layers in the distillation process.

New Framework



- We freeze only input embedding layer and include the transformer layers in the distillation process.

How can we design efficient algorithm that remains computationally lightweight even with the text encoder included in the distillation process?

Simplifying Bi-level optimization problem

Original Bi-level optimization problem

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathcal{L}_{\text{NCE}}(\hat{\Theta}; \mathcal{T}) \quad \text{where} \quad \hat{\Theta} = \arg\min_{\Theta} \mathcal{L}_{\text{NCE}}(\Theta; \mathcal{S})$$

- InfoNCE loss: $\mathcal{L}_{\text{NCE}} = -\frac{1}{M}\sum_{i=1}^{M}\left[\log\frac{\exp(s_{ii}/\tau)}{\sum_{j\neq i}\exp(s_{ij}/\tau)} + \log\frac{\exp(s_{ii}/\tau)}{\sum_{j\neq i}\exp(s_{ji}/\tau)}\right]$

Simplifying Bi-level optimization problem

Original Bi-level optimization problem

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathcal{L}_{\text{NCE}}(\hat{\Theta}; \mathcal{T}) \quad \text{where} \quad \hat{\Theta} = \arg\min_{\Theta} \mathcal{L}_{\text{NCE}}(\Theta; \mathcal{S})$$

- InfoNCE loss: $\mathcal{L}_{\text{NCE}} = -\frac{1}{M}\sum_{i=1}^{M}\left[\log \frac{\exp(s_{ii}/\tau)}{\sum_{j\neq i}\exp(s_{ij}/\tau)} + \log \frac{\exp(s_{ii}/\tau)}{\sum_{j\neq i}\exp(s_{ji}/\tau)}\right]$

Assumption: linearized contrastive learning

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} L_{\text{lin}}(\hat{G}_v, \hat{G}_l; \mathcal{T}) \text{ where } \hat{G}_v, \hat{G}_l = \arg\min_{G_v, G_l} L_{\text{lin}}(G_v, G_l; S)$$

- Only optimize linear projection layers G_v, G_l
- Linear contrastive loss: $\mathcal{L}_{\text{lin}}(G_v, G_l; \mathcal{D}) = \frac{1}{2|\mathcal{D}|(|\mathcal{D}|-1)}\sum_{i=1}^{|\mathcal{D}|}\sum_{j\neq i}(s_{ij} - s_{ii}) + \frac{1}{2|\mathcal{D}|(|\mathcal{D}|-1)}\sum_{i=1}^{|\mathcal{D}|}\sum_{j\neq i}(s_{ji} - s_{ii}) + \frac{\rho}{2}\|G_v^\top G_l\|_F^2$

Simplifying Bi-level optimization problem

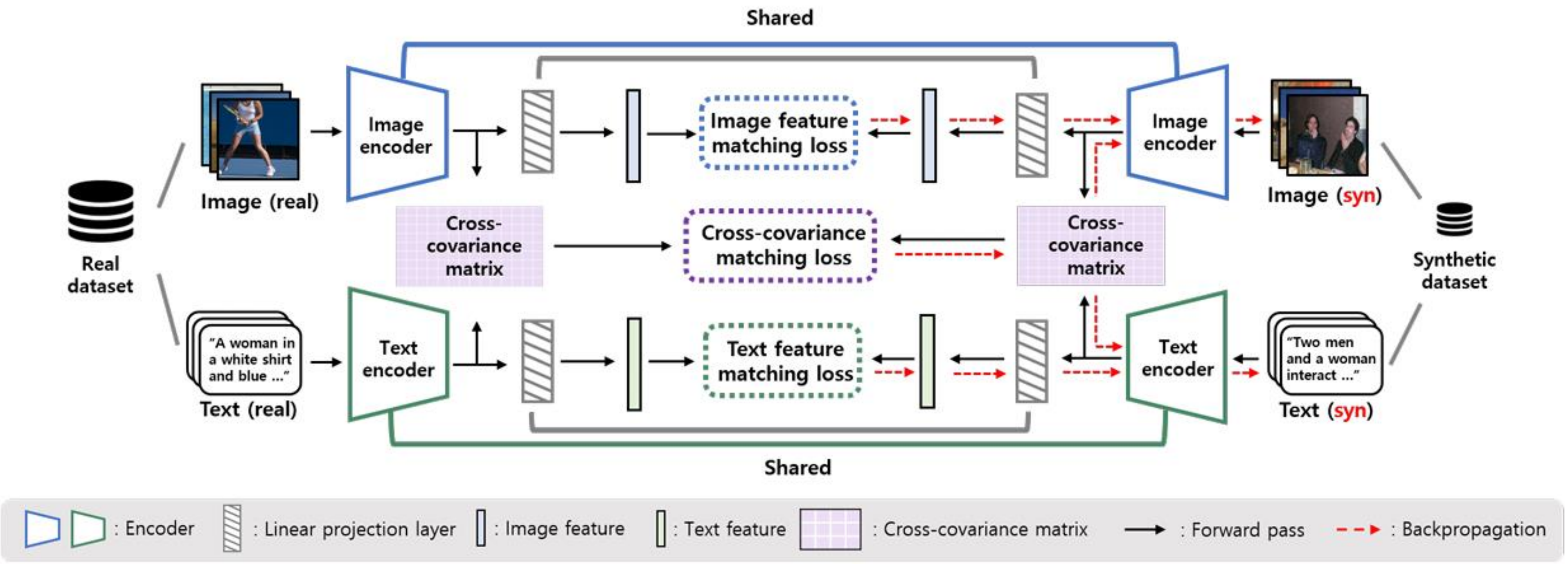
Closed-form solution

$$\mathcal{S}^* = \arg\max_{\mathcal{S}} \mathrm{Tr}({C^{\mathcal{T}}}^\top C^{\mathcal{S}})$$

- Cross-covariance matrix: $C^{\mathcal{D}} = \frac{1}{|\mathcal{D}|-1}\sum_{i=1}^{|\mathcal{D}|}(h_v^i - \mu_{h_v})(h_l^i - \mu_{h_l})^\top$

- Multimodal dataset distillation can be simplified as **aligning the cross-covariance matrices of real and synthetic dataset !**
- From this insight, we develop cross-covariance matching algorithm (**CovMatch**).

CovMatch: Cross-Covariance Matching Algorithm



- **Cross-covariance matching loss**: $\mathcal{L}^{\text{cov}}(\mathcal{T}, \mathcal{S}) = \|\rho \cdot C^{\mathcal{T}} - C^{\mathcal{S}}\|_F^2$
- **Feature matching loss**: Regularization to prevent trivial solutions.
- **Online model update**: To prevent overfitting to specific model state, we update network at every distillation step.

Main Results

- IR@K: Image retrieval given text (Top-K accuracy)
- TR@K: Text retrieval given image (Top-K accuracy)

Pairs	Method	Flickr30k							COCO						
		IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Avg	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Avg
100	Random	2.0	7.5	12.6	3.3	10.4	16.0	8.6	0.7	2.8	5.1	1.0	4.0	6.9	3.4
	Herding	2.2	8.0	13.4	3.0	9.9	15.6	8.7	0.7	2.9	5.3	1.1	4.1	6.8	3.5
	K-Center	2.0	7.6	13.0	2.8	9.7	16.4	8.6	0.7	3.2	6.0	0.9	4.2	7.6	3.8
	MTT-VL	4.7	15.7	24.6	9.9	28.3	39.1	20.4	1.3	5.4	9.5	2.5	10.0	15.7	7.4
	LoRS	8.3	24.1	35.1	11.8	35.8	49.2	27.4	1.8	7.1	12.2	3.3	12.2	19.6	9.4
	CovMatch	**10.1**	**28.6**	**40.9**	**14.8**	**38.0**	**50.6**	**30.5**	**2.8**	**10.5**	**17.7**	**3.8**	**13.1**	**21.1**	**11.5**
200	Random	3.3	11.5	18.4	5.7	15.8	23.9	13.1	1.1	4.6	8.3	1.7	6.5	11.1	5.6
	Herding	3.0	11.3	18.3	4.7	15.4	22.9	12.6	1.2	4.7	8.5	1.6	6.6	11.2	5.6
	K-Center	3.2	11.1	17.7	5.3	15.2	23.2	12.6	1.2	5.1	8.9	1.9	6.7	11.6	5.9
	MTT-VL	4.6	16.0	25.5	10.2	28.7	41.9	21.2	1.7	6.5	12.3	3.3	11.9	19.4	9.2
	LoRS	8.6	25.3	36.6	14.5	38.7	53.4	29.5	2.4	9.3	15.5	4.3	14.2	22.6	11.4
	CovMatch	**12.3**	**33.6**	**45.8**	**17.4**	**41.7**	**55.8**	**34.4**	**3.8**	**13.4**	**21.8**	**5.3**	**17.3**	**27.0**	**14.8**
500	Random	6.9	21.0	31.2	10.0	28.0	38.7	22.6	2.2	8.8	14.9	3.5	11.9	19.2	10.1
	Herding	6.8	20.8	30.9	9.3	26.4	36.8	21.8	2.3	8.8	14.8	2.9	11.2	18.9	9.8
	K-Center	6.9	22.1	32.2	10.6	29.5	40.6	23.7	2.4	9.0	15.4	3.6	12.4	20.0	10.5
	MTT-VL	6.6	20.2	30.0	13.3	32.8	46.8	25.0	2.5	8.9	15.8	5.0	17.2	26.0	12.6
	LoRS	10.0	28.9	41.6	15.5	39.8	53.7	31.6	2.8	9.9	16.5	5.3	18.3	27.9	13.5
	CovMatch	**14.7**	**38.4**	**51.4**	**19.9**	**46.7**	**59.5**	**38.4**	**5.4**	**18.0**	**28.2**	**8.1**	**23.5**	**34.6**	**19.6**

- CovMatch establishes new state-of-the-art performances across all settings.
- On Flickr30k with 500 synthetic pairs, CovMatch achieves a 6.8% absolute improvement over the strongest baseline.
- In contrast to CovMatch, baselines quickly saturate as the number of synthetic pairs increases.

Cross-Architecture Generalization

Cross-Architecture generalization experiment:

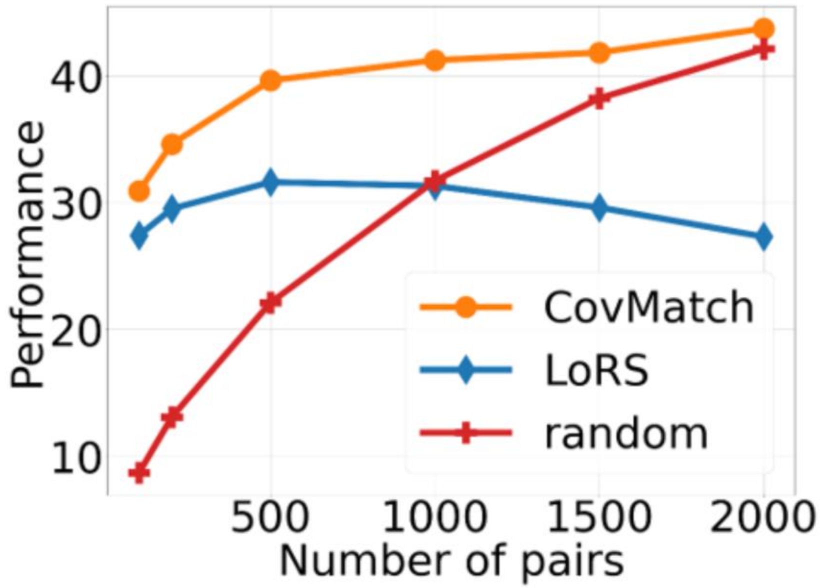
- Distill the dataset using NFNet (image encoder) and BERT (text encoder).
- Evaluate the distilled dataset using other unseen networks.

Text encoder	BERT				DistilBERT			
Image encoder	NFNet	NF-ResNet	NF-RegNet	ViT	NFNet	NF-ResNet	NF-RegNet	ViT
Random	8.4	8.9	8.3	10.8	9.4	10.2	8.7	11.5
MTT-VL	20.4	8.4	7.5	9.6	20.2	7.5	7.0	8.5
LoRS	28.1	8.8	8.4	9.3	23.5	8.9	8.3	8.9
CovMatch	**30.2**	**15.5**	**14.6**	**15.1**	**27.1**	**16.1**	**14.6**	**13.4**

* Flickr30k with 100 synthetic pairs.

- CovMatch shows much stronger cross-architecture generalization ability compared to baseline.

Scalability



- CovMatch scales well as we increase the number of synthetic pairs.
- **Strong scalability !**

Concluding Remarks

- We revisit the bi-level formulation of dataset distillation and, assuming linear contrastive learning, derive a simpler objective that allows us to also train the text encoder in multimodal distillation.
- Based on this, we propose **CovMatch**, a lightweight and scalable method that matches cross-covariance between real and synthetic image-text embeddings, with additional intra-modal feature regularization.
- CovMatch achieves strong gains over prior work, improving cross-modal retrieval, generalization to unseen architectures, and scalability.

Thank you for listening