Mind the Gap Between Prototypes and Images in Cross-domain Finetuning

Hongduan Tian, Feng Liu, Zhanke Zhou, Tongliang Liu, Chengqi Zhang, Bo Han

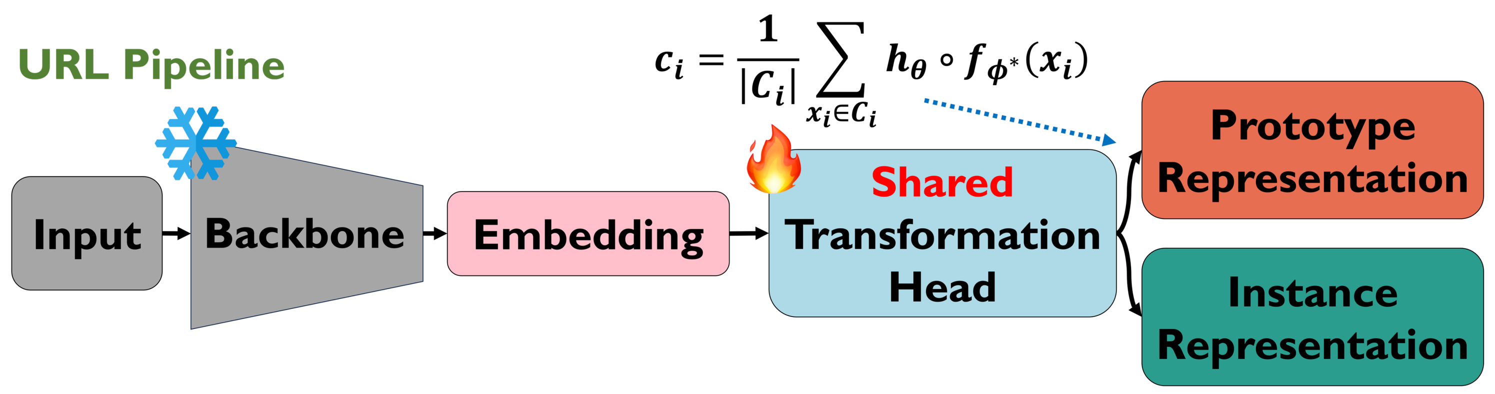
NEURAL INFORMATION PROCESSING SYSTEMS · TMLR · 香港浸會大學 HONG KONG BAPTIST UNIVERSITY · THE UNIVERSITY OF MELBOURNE · THE UNIVERSITY OF SYDNEY · THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

Paper Code Slides

Assumption: Prototypes and Images share the same transformation

URL[1] framework:

- The prototype is the average of all available images In a class;
- URL implicitly assumes that the same transformation are shared.



$$c_i = \frac{1}{|C|}\sum_{x \in C}\left(f_{\phi^*}(x)\Theta\right) = \left(\frac{1}{|C|}\sum_{x \in C} f_{\phi^*}(x)\right)\Theta$$

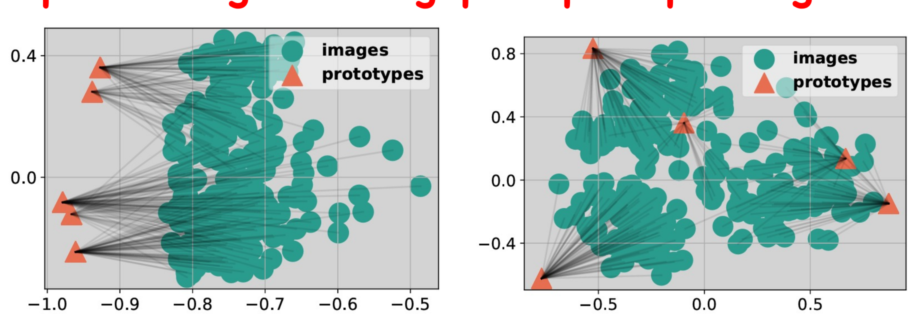
image instance representations — prototype embeddings

1. Li et al., Universal representation learning from multiple domains for few-shot classification, ICCV 2021.

Observation: The gap between prototypes & images

Motivation:

- Similar to texts in multi-modal framework, prototypes describe higher level information compared to image instances.
- According to [2], there exists a gap between text and image data, and preserving such a gap helps improve generalization performance.

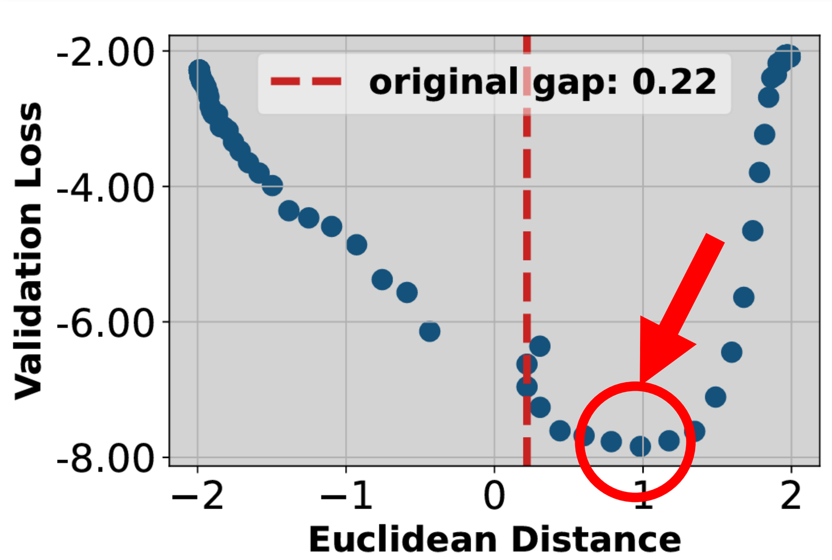


(a) $\|\vec{\Delta}\| = 0.22$ (b) $\|\vec{\Delta}\| = 0.12$

Two Observations:
Left: The existence of a gap between image and prototype embeddings.
Right: The application of the shared transformation head undermines the gap.

2. Liang et al., Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning, NeurIPS 2022.

The property of the gap



Conjecture of reasons:

- Enlarging the prototype and image embedding gap potentially helps alleviate the overfitting;
- Enlarging the gap helps align the representations.

Slightly enlarging the gap between the prototype and image embeddings improves the generalization performance!

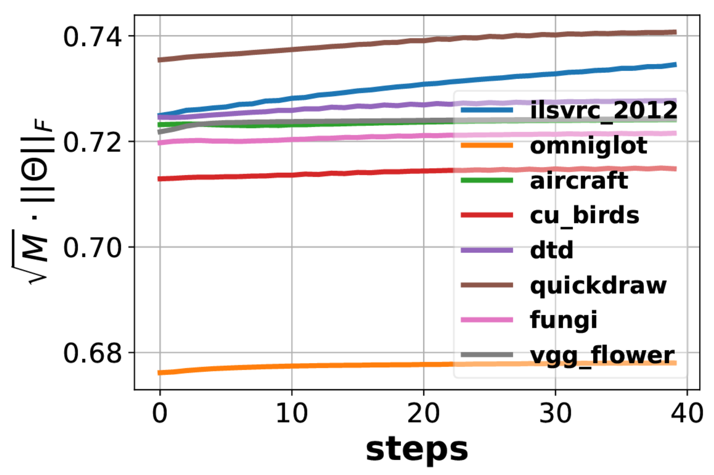
Theoretical Analyses of the Gap

The upper bound of the prototype and image representation gap in the context of the shared transformation head:

$$\left\|\frac{1}{|\mathcal{Z}|}\sum_{z \in \mathcal{Z}} z - \frac{1}{|\mathcal{C}|}\sum_{c \in \mathcal{C}} c\right\|_2 \leq \max_{1 \leq j \leq d}\cos\left(\vec{\Delta}, \Theta^j\right)\|\Theta\|_F\left\|\vec{\Delta}\right\|_2$$

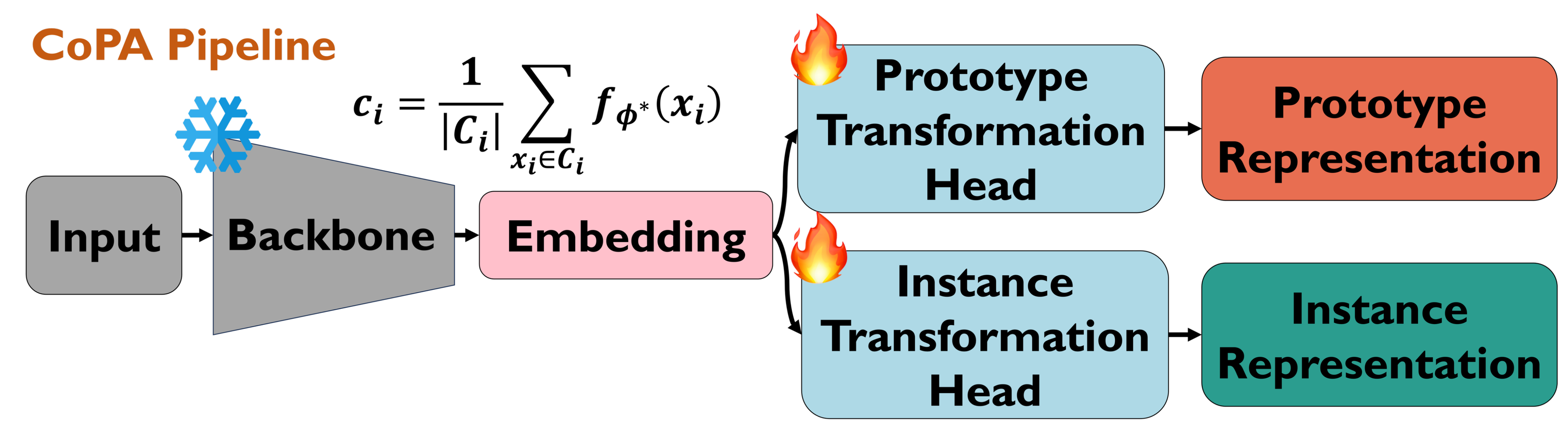
Two aspects in the bound:

- The Frobenius norm of the transformation matrix Θ;
- The maximal similarity between the embedding gap and column vectors of transformation matrix.



The coefficient of the upper bound is consistently smaller than 1.0, which indicate that the gap will be narrowed.

Method: Contrastive Prototype-image Adaptation



Details of CoPA:

1. Randomly sample a support set $\{X, Y\}$, Y is one-hot label;
2. Generate pseudo labels $Y_{\text{pseudo}} = \{0, 1, 2, \ldots\}$;
3. Generate instance representations $Z_{\mathrm{I}} = f_{\phi^*}(X)\Theta_{\mathrm{I}}$ and prototype embeddings, $Z_{\mathrm{P}} = YY^T f_{\phi^*}(X)\Theta_{\mathrm{P}}$;
4. Iteratively optimize Θ_{I} and Θ_{P} with the objective: $\mathcal{L}_{\mathrm{CE}}\left(\frac{1}{\tau}Z_{\mathrm{I}}Z_{\mathrm{P}}^T, Y_{\text{pseudo}}\right) + \mathcal{L}_{\mathrm{CE}}\left(\frac{1}{\tau}Z_{\mathrm{P}}Z_{\mathrm{I}}^T, Y_{\text{pseudo}}\right)$

- The different transformation heads in CoPA pipeline help preserve discriminative information in gradients.
- $YY^T f_{\phi^*}(X)$ explicitly leverage the cluster structure of data.

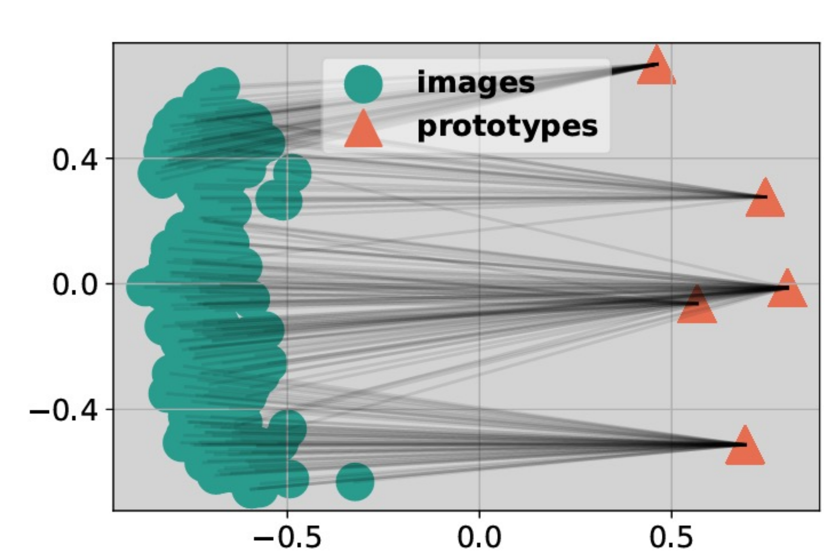
Experiments

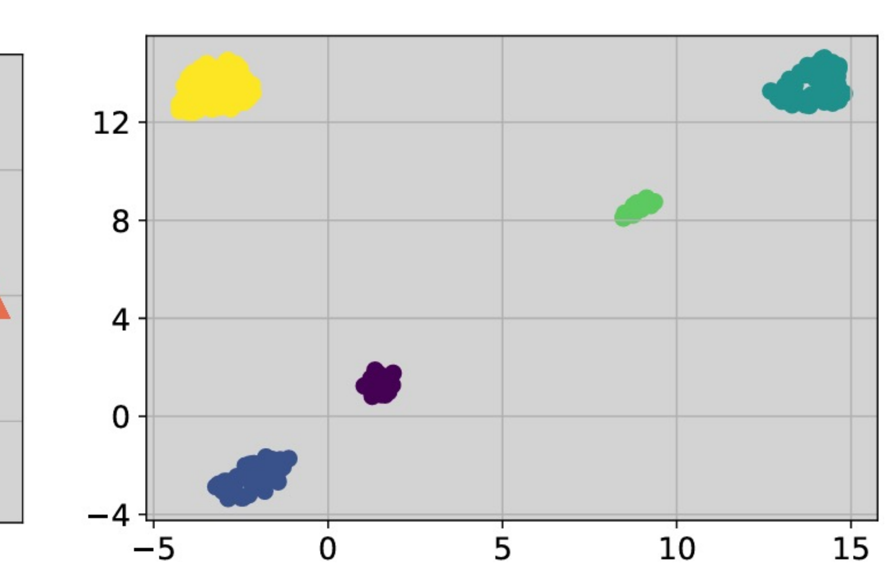
- **SOTA quantitative results:**

Datasets	CNAPS	S-CNAPS	SUR	URT	Tri-M	FLUTE	URL	CoPA	TSA	TA²-Net	CoPA + TSA
	Main Results								More Learning Modules		
ImageNet	50.8±1.1	58.4 ±1.1	56.2 ± 1.0	56.8 ± 1.1	**58.6 ± 1.0**	51.8 ± 1.1	57.3 ± 1.1	**57.8 ± 1.1**	57.4 ± 1.1	57.5 ± 1.1	**57.8 ± 1.1**
Omniglot	91.7±0.5	91.6 ± 0.6	94.1 ± 0.4	94.2 ± 0.4	92.0 ± 0.6	93.2 ± 0.5	94.1 ± 0.4	**94.3 ± 0.5**	**94.7 ± 0.4**	94.6 ± 0.4	94.6 ± 0.4
Aircraft	83.7±0.6	82.0 ± 0.7	85.5 ± 0.5	85.8 ± 0.5	82.8 ± 0.7	87.2 ± 0.5	88.2 ± 0.5	**88.8 ± 0.5**	88.9 ± 0.5	89.0 ± 0.5	**89.3 ± 0.5**
Birds	73.6±0.9	74.8 ± 0.9	71.0 ± 1.0	76.2 ± 0.8	75.3 ± 0.8	79.2 ± 0.8	80.2 ± 0.7	**80.8 ± 0.8**	80.8 ± 0.8	80.7 ± 0.8	**81.2 ± 0.8**
Textures	59.5±0.7	68.8 ± 0.9	71.0 ± 0.8	71.6 ± 0.7	71.2 ± 0.8	68.8 ± 0.8	76.2 ± 0.7	**77.8 ± 0.7**	77.1 ± 0.7	76.9 ± 0.7	**77.8 ± 0.7**
Quick Draw	74.7±0.8	76.5 ±0.8	81.8 ± 0.6	82.4 ± 0.6	77.3 ± 0.7	79.5 ± 0.7	82.2 ± 0.6	**82.8 ± 0.6**	82.2 ± 0.6	82.2 ± 0.6	**82.7 ± 0.6**
Fungi	50.2±1.1	46.6 ± 1.0	64.3 ± 0.9	64.0 ± 1.0	48.5 ± 1.0	58.1 ± 1.1	68.7 ± 1.0	**69.5 ± 1.0**	67.4 ± 1.0	68.1 ± 1.0	**69.0 ± 1.0**
VGG Flower	88.9±0.5	90.5 ± 0.5	82.9 ± 0.8	87.9 ± 0.6	90.5 ± 0.5	91.6 ± 0.6	91.9 ± 0.5	**92.7 ± 0.5**	92.5 ± 0.5	92.4 ± 0.5	**93.0 ± 0.5**
Traffic Sign	56.5 ±1.1	57.2 ± 1.0	51.0 ± 1.1	48.2 ± 1.1	63.0 ± 1.0	58.4 ± 1.1	63.3 ± 1.2	**66.6 ± 1.1**	83.5 ± 0.9	88.3 ± 0.8	**88.5 ± 0.9**
MSCOCO	39.4 ±1.0	48.9 ± 1.1	52.0 ± 1.1	51.5 ± 1.1	52.8 ± 1.1	50.0 ± 1.0	54.2 ± 1.0	**56.3 ± 1.0**	55.3 ± 1.1	49.9 ± 1.2	**57.9 ± 1.0**
MNIST	-	94.6 ± 0.4	94.3 ± 0.4	90.6 ± 0.5	**96.2 ± 0.3**	**95.6 ± 0.5**	94.7 ± 0.4	**95.2 ± 0.4**	96.7 ± 0.4	97.0 ± 0.4	**97.5 ± 0.4**
CIFAR-10	-	74.9 ± 0.7	66.5 ± 0.9	67.0 ± 0.8	75.4 ± 0.8	**78.6 ± 0.7**	71.9 ± 0.8	**73.0 ± 0.8**	**80.3 ± 0.8**	76.6 ± 0.9	78.7 ± 0.8
CIFAR-100	-	61.3 ± 1.1	56.9 ± 1.1	57.3 ± 1.0	62.0 ± 1.0	**67.1 ± 1.0**	62.9 ± 1.0	**63.4 ±1.0**	70.6 ± 1.0	64.5 ± 1.2	**70.9 ± 0.9**
Average Seen	71.6	73.7	75.9	77.4	76.2	76.2	79.9	**80.6**	80.1	80.2	**80.7**
Average Unseen	-	67.4	64.1	62.9	69.9	69.9	69.4	**70.9**	77.3	75.2	**78.7**
Average All	-	71.2	71.3	71.8	73.8	73.8	75.8	**76.8**	79.2	78.3	**79.9**
Average Rank	10.3	8.7	8.7	7.1	7.9	7.8	4.5	**3.0**	3.1	3.3	**2.6**

Datasets	Finetune	ProtoNets(large)	BOHB	FP-MAML	AFP-MAML	FLUTE	URL	CoPA	TSA	TA²-Net	CoPA+TSA
	Main Results								More Learning Modules		
ImageNet	45.8±1.1	53.7±1.1	51.9±1.1	49.5±1.1	52.8±1.1	46.9±1.1	57.3±1.1	**57.7±1.1**	**57.7 ± 1.1**	57.4 ± 1.1	57.5 ± 1.1
Omniglot	60.9±1.6	68.5±1.3	67.6±1.2	63.4±1.3	61.9±1.5	61.6±1.4	69.4±1.2	**70.9±1.2**	**73.5 ± 1.2**	72.8 ± 1.2	73.3 ± 1.2
Aircraft	**68.7±1.3**	58.0±1.0	54.1±0.9	56.0±1.0	63.4±1.1	48.5±1.0	57.6±1.0	**61.6±1.0**	**65.1 ± 1.1**	63.5 ± 1.0	64.9 ± 1.1
Birds	57.3±1.3	74.1±0.9	70.7±0.9	68.7±1.0	69.8±1.1	47.9±1.0	72.9±0.9	**74.2±0.9**	74.0 ± 0.9	73.8 ± 0.9	**74.7 ± 0.9**
Textures	69.0±0.9	68.8±0.8	68.3±0.8	66.5±0.8	70.8±0.9	63.8±0.8	75.2±0.7	**77.0±0.7**	76.8 ± 0.7	76.6 ± 0.7	**77.6 ± 0.7**
Quick Draw	42.6±1.2	53.3±1.0	50.3±1.0	51.5±1.0	59.2±1.2	57.5±1.0	57.9±1.0	**61.3±1.0**	64.6 ± 1.0	63.9 ± 1.0	**64.7 ± 1.0**
Fungi	38.2±1.0	40.7±1.2	41.4±1.1	40.0±1.1	41.5±1.2	31.8±1.0	46.2±1.0	**48.0±1.1**	46.8 ± 1.1	47.6 ± 1.1	**48.3 ± 1.1**
VGG Flower	85.5±0.7	87.0±0.7	87.3±0.6	87.2±0.7	86.0±0.8	80.1±0.9	86.9±0.6	**88.9±0.6**	89.8 ± 0.6	89.6 ± 0.6	**90.6 ± 0.6**
Traffic Sign	**66.8±1.3**	58.1±1.1	51.8±1.0	48.8±1.1	60.8±1.3	46.5±1.1	61.2±1.2	**63.8±1.1**	82.2 ± 0.9	**87.7 ± 0.8**	86.7 ± 0.9
MSCOCO	34.9±1.0	41.7±1.1	48.0±1.0	43.7±1.1	48.1±1.1	41.4±1.0	53.0±1.0	**56.1±1.0**	55.8 ± 1.0	51.3 ± 1.2	**57.4 ± 1.0**
MNIST	-	-	-	-	-	80.8±0.8	86.2±0.7	**87.3±0.7**	93.6 ± 0.6	94.7 ± 0.5	**95.1 ± 0.6**
CIFAR-10	-	-	-	-	-	65.4±0.8	69.5±0.8	**72.4±0.8**	**79.6 ± 0.8**	76.1 ± 0.9	76.8 ± 0.8
CIFAR-100	-	-	-	-	-	52.7±1.1	62.0±1.0	**62.7±1.0**	**70.6 ± 1.0**	65.7 ± 1.1	68.9 ± 0.9
Average Seen	45.8	53.7	51.9	49.5	52.8	46.9	57.3	**57.7**	**57.7**	57.5	57.5
Average Unseen	-	-	-	-	-	56.5	66.6	**68.7**	72.7	71.9	**73.2**
Average All	-	-	-	-	-	55.8	65.9	**67.7**	71.6	70.8	**72.0**
Average Rank	9.3	7.2	8.0	9.0	7.1	10.1	5.3	**4.1**	2.5	3.2	**2.2**

- **Preserved representation gap & better data clusters:**





- **Better generalization performance & less time consumption:**