

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

**Hongduan Tian**<sup>1,2</sup>, Feng Liu<sup>3</sup>, Zhanke Zhou<sup>1,2</sup>, Tongliang Liu<sup>4</sup>,  
Chengqi Zhang<sup>5</sup>, Bo Han<sup>1,2</sup>

<sup>1</sup>TMLR Group, Hong Kong Baptist University

<sup>2</sup>Department of Computer Science, Hong Kong Baptist University

<sup>3</sup>TMLR Group, University of Melbourne, <sup>4</sup>Sydney AI Centre, The University of Sydney

<sup>5</sup>Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University



## Outline

- Background
- Revisit Previous Adaptation Strategy
- Contrastive Prototype-Image Adaptation (CoPA)
- Summary

---

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

---

**Hongduan Tian<sup>1</sup>, Feng Liu<sup>2</sup>, Zhanke Zhou<sup>1</sup>, Tongliang Liu<sup>3</sup>, Chengqi Zhang<sup>4</sup>, Bo Han<sup>1†</sup>**

<sup>1</sup>TMLR Group, Department of Computer Science, Hong Kong Baptist University

<sup>2</sup>TMLR Group, University of Melbourne    <sup>3</sup>Sydney AI Center, The University of Sydney

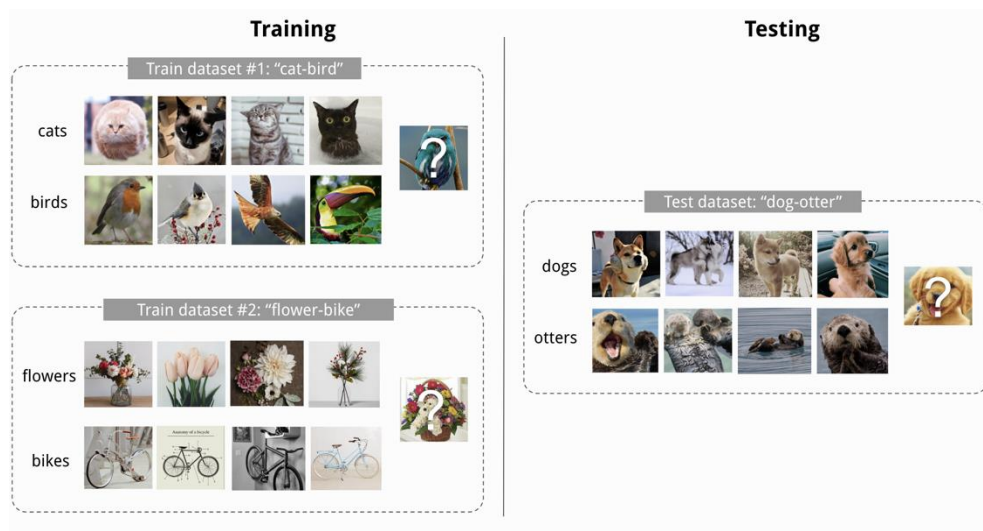
<sup>4</sup>Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University

{cshdtian, cszkzhou, bhanml}@comp.hkbu.edu.hk,

fengliu.ml@gmail.com, tongliang.liu@sydney.edu.au, chengqi.zhang@polyu.edu.hk

# Preliminary: Cross-domain Few-shot Classification

## Few-shot classification



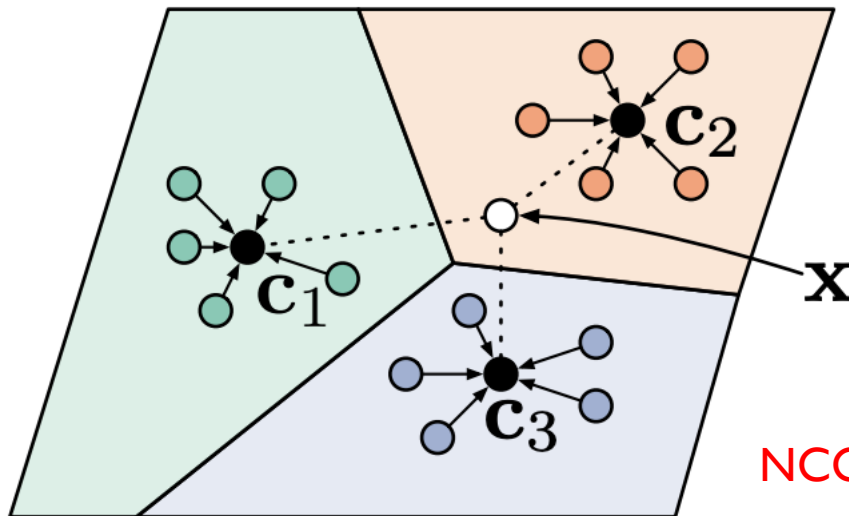
An example of conventional few-shot classification tasks

## Challenges in CFC:

- Numbers of ways & shots vary among tasks;
- Discrepancies between source and target domains

## Preliminary: Prototypical Networks

### Few-shot classification with prototypes



NCC-based loss

- Construct prototypes:

$$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

- Calculate similarities/distances:

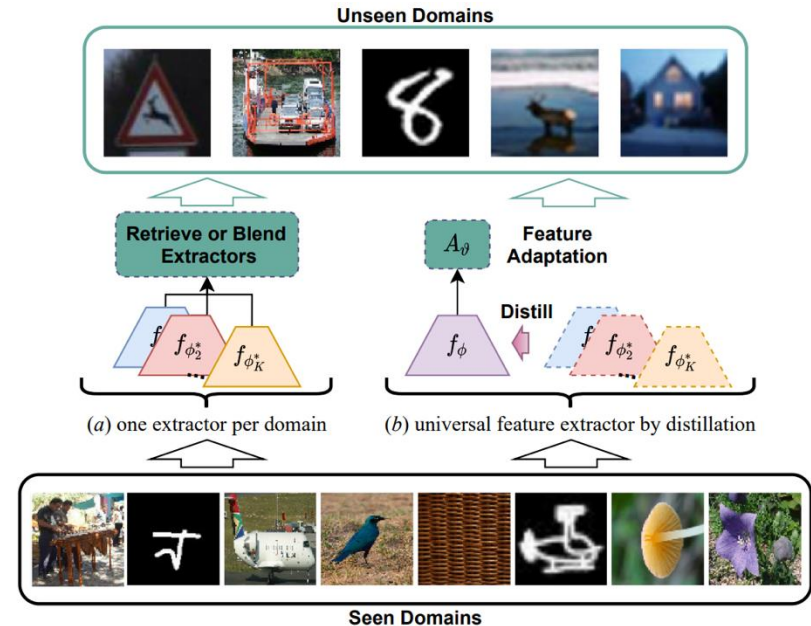
$$L = \frac{1}{|D_T|} \sum_{i=1}^{|D_T|} \log(p(\hat{y} = y_i | x_i))$$

$$p(\hat{y} = y_i | x_i) = \frac{\exp(-d(x, c_i))}{\sum_j \exp(-d(x, c_j))}$$

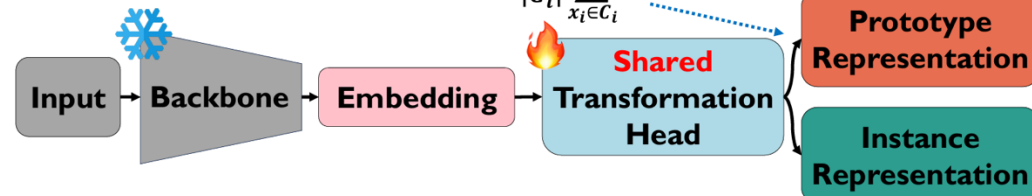
## Previous Works

Finetuning a transformation on top of a universal pretrained backbone

- **Pretraining:** Distill a universal backbone from several task-specific backbones.
- **Meta-Test:** Adapting a light transformation module on top of the pretrained backbone.

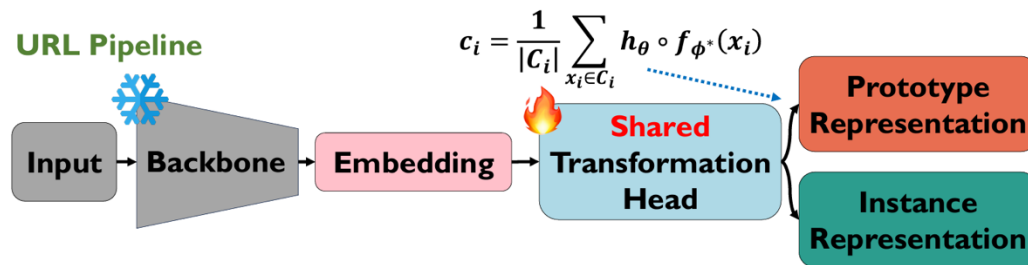


URL Pipeline



## Motivation

### An Implicit Assumption

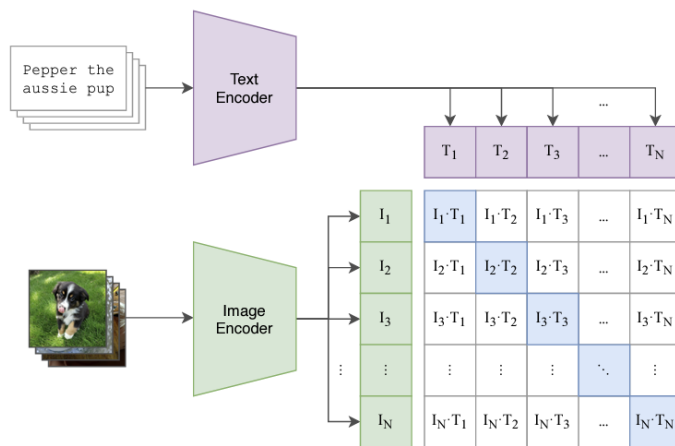


$$c = \frac{1}{|C|} \sum_{x \in C} \underbrace{(f_{\phi^*}(x)\Theta)}_{\text{Instance Representations}} = \underbrace{\left( \frac{1}{|C|} \sum_{x \in C} f_{\phi^*}(x) \right)}_{\text{Prototype Embeddings}} \Theta$$

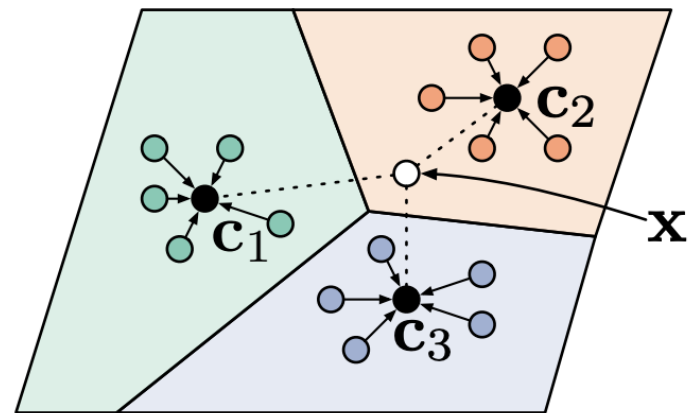
Instance-level and prototype-level embeddings share the same representation transformation.

## Motivation

### The role of prototypes



**Text:** Abstract information of a set of images.

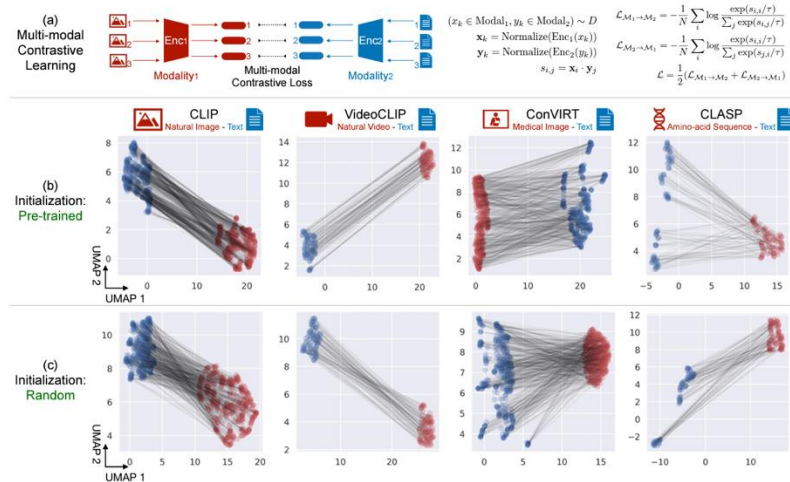


**Prototype:** Information commonly shared across of images in a class.

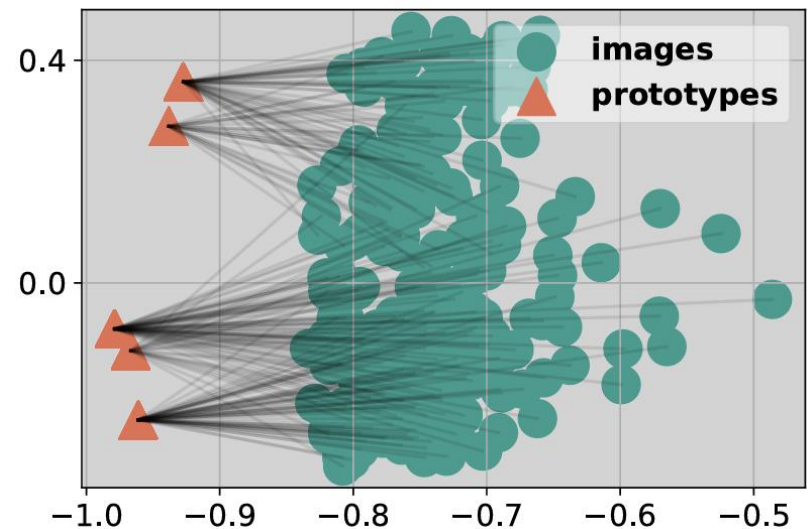
Prototypes play the similar role (Higher level information) to the texts in multimodal frameworks.

## Revisit Previous Adaptation Strategy

### Gap between prototypes and images



$$\vec{\Delta} := \frac{1}{|\mathcal{D}_{\mathcal{T}}|} \sum_{i=1}^{|\mathcal{D}_{\mathcal{T}}|} \mathbf{z}_i - \frac{1}{N_C} \sum_{j=1}^{N_C} \mathbf{c}_j$$



There exist modality gaps between different modalities of data

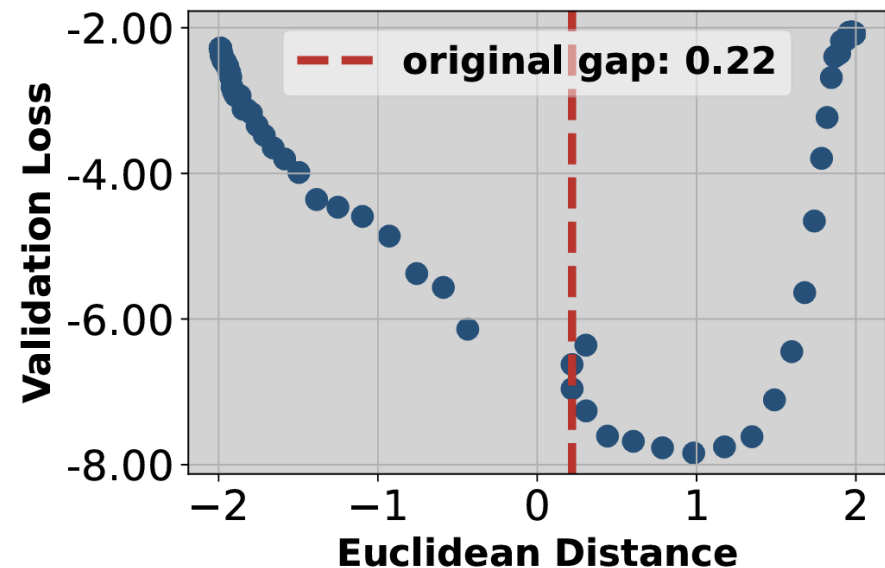
There also exist a gap between prototype and image embeddings!



## Revisit Previous Adaptation Strategy

Larger gap facilitates better generalization performance

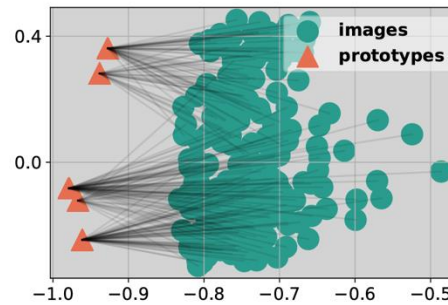
- In original method, the prototype representations are generated directly from image instance representations → **Overfitting**;
- The representations of prototypes and image instances are not well aligned.



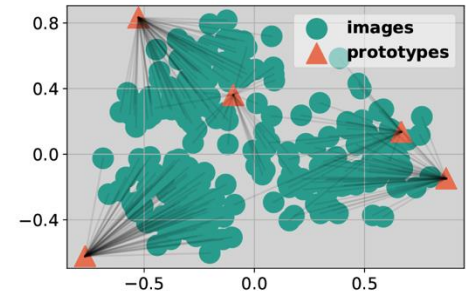
## Revisit Previous Adaptation Strategy

### The effect of the shared transformation

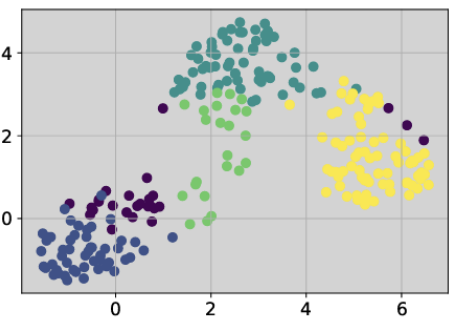
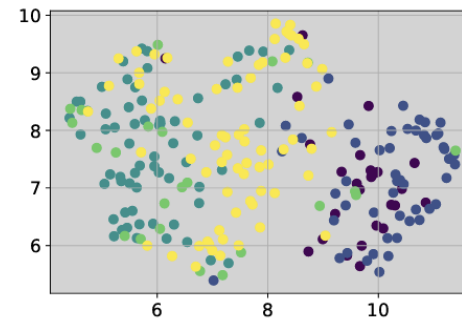
- The shared transformation narrows the gap between prototype and image instance representations;
- The shared transformation fails to learn image instance representations that are well clustered for each class.



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



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



## Revisit Previous Adaptation Strategy

### Further Analyses

- The shared transformation tends to drop the discriminative information in gradients respectively for prototypes and images

**Theorem 3.1.** Let the measure  $d(\cdot, \cdot)$  be the cosine similarity function. Given a set of normalized finite support data representation  $\mathcal{Z} = \{(\mathbf{z}_i, y_i)\}_{i=1}^n$ , where  $\|\mathbf{z}\|_2 = 1$  for  $\forall \mathbf{z} \in \mathcal{Z}$  and  $N_C$  classes are included, then we have a lower bound of the NCC-based loss in Eq. (1):

$$\mathcal{L}(\theta) \geq -\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^\top \mathbf{c}_c + \frac{\alpha}{n} \sum_{i=1}^n \sum_{\mathbf{z}' \in \mathcal{Z}} \mathbf{z}_i^\top \mathbf{z}',$$

where  $\mathbf{z}'$  is an independent copy of samples in  $\mathcal{Z}$ ,  $\mathbf{c}_c$  denotes sets of sample representations  $\mathcal{C}_c = \{\mathbf{z}_i | y_i = c\}$ , and  $\alpha$  is a constant that satisfies  $0 \leq \alpha < 1/(N_C |\mathcal{C}_j|)$  for  $\forall j$ .

$$\mathcal{L}(\Theta_P, \Theta_I) = -\frac{1}{|\mathcal{D}_T|} \text{Tr} (f_{\phi^*}(\mathbf{X}) \Theta_I (Y Y^\top f_{\phi^*}(\mathbf{X}) \Theta_P)^\top) + \frac{\alpha}{|\mathcal{D}_T|} \text{Tr} (f_{\phi^*}(\mathbf{X}) \Theta_I \Theta_I^\top f_{\phi^*}(\mathbf{X})^\top),$$

where  $\mathbf{X} \in \mathbb{R}^{|\mathcal{D}_T| \times d_{\text{out}}}$  and  $Y \in \mathbb{R}^{|\mathcal{D}_T| \times N_C}$  respectively denote the support image instances and the corresponding one-hot labels,  $\Theta_P \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$  and  $\Theta_I \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$  denote the model parameters of linear transformation heads respectively for prototype and image instance embeddings,  $\text{Tr}(\cdot)$  denotes the matrix trace operation.  $Y Y^\top f_{\phi^*}(\mathbf{X}) \in \mathbb{R}^{|\mathcal{D}_T| \times d_{\text{out}}}$  denotes the prototypes which are expanded to the same size of instance embeddings. In this way, the gradients w.r.t.  $\Theta_P$  and  $\Theta_I$  are:

$$\nabla_{\Theta_P} \mathcal{L}(\Theta_P, \Theta_I) = -\frac{1}{|\mathcal{D}_T|} \Theta_I^\top f_{\phi^*}(\mathbf{X})^\top Y Y^\top f_{\phi^*}(\mathbf{X}),$$

$$\nabla_{\Theta_I} \mathcal{L}(\Theta_P, \Theta_I) = -\frac{1}{|\mathcal{D}_T|} \Theta_P^\top f_{\phi^*}(\mathbf{X})^\top Y Y^\top f_{\phi^*}(\mathbf{X}) + \frac{2\alpha}{|\mathcal{D}_T|} \Theta_I^\top f_{\phi^*}(\mathbf{X})^\top f_{\phi^*}(\mathbf{X}).$$

## Revisit Previous Adaptation Strategy

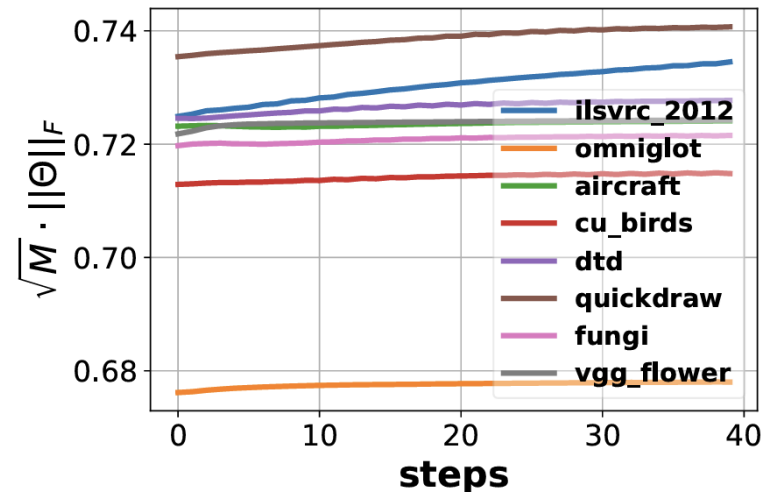
### Further Analyses

- The empirical results regarding the coefficient of the upper bound indicate that the shared transformation will shrink the gap since the coefficient is consistently smaller than 1.0.

**Theorem 3.2 (The shared transformation).** Consider a support data set  $\mathcal{D}_T = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_T|}$  composed of  $N_C$  classes and a frozen pretrained backbone  $f_{\phi^*} : \mathbb{R}^{d_{in}} \rightarrow \mathbb{R}^d$  parameterized with the optimal parameters  $\phi^*$ . Let  $\Theta \in \mathbb{R}^{d \times d}$  be a shared linear transformation across the prototype and image instance embeddings. Then, we can obtain the image instance representations  $\mathcal{Z} = \{z_i\}_{i=1}^{|\mathcal{D}_T|} = \{f_{\phi^*}(x_i)\Theta\}_{i=1}^{|\mathcal{D}_T|}$ , and the prototype representations  $\mathcal{C} = \{c_i\}_{i=1}^{N_C}$ , where  $c_i = \frac{1}{|\mathcal{C}_i|} \sum_{z' \in \mathcal{C}_i} z' = \frac{1}{|\mathcal{C}_i|} \sum_{x' \in \mathcal{C}_i} f_{\phi^*}(x')\Theta$ . Then we can obtain the bounds of the representation gap:

$$m \|\Theta\|_F^2 \|\vec{\Delta}_{\text{emb}}\|_2^2 \leq \left\| \frac{1}{|\mathcal{D}_T|} \sum_{z \in \mathcal{Z}} z - \frac{1}{N_C} \sum_{c \in \mathcal{C}} c \right\|_2^2 \leq M \|\Theta\|_F^2 \|\vec{\Delta}_{\text{emb}}\|_2^2,$$

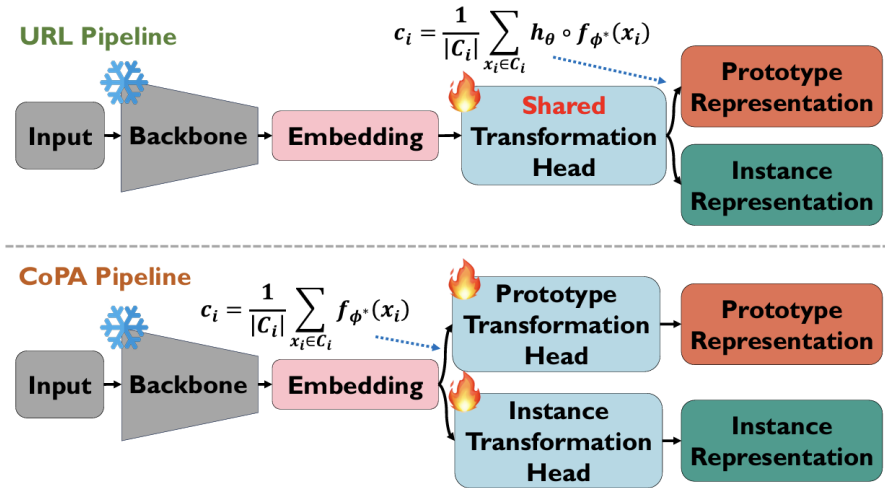
where  $\vec{\Delta}_{\text{emb}} = \frac{1}{|\mathcal{D}_T|} \sum_{x \in \mathcal{D}_T} f_{\phi^*}(x) - \frac{1}{N_C} \sum_{b=1}^{N_C} \left( \frac{1}{|\mathcal{C}_b|} \sum_{x' \in \mathcal{C}_b} f_{\phi^*}(x') \right)$  denotes the gap between prototype and image embeddings,  $m = \min_{1 \leq i \leq d} \cos^2(\vec{\Delta}_{\text{emb}}, \Theta^i)$  denotes the minimum value of  $\cos^2(\vec{\Delta}_{\text{emb}}, \Theta^i)$ , and  $M = \max_{1 \leq j \leq d} \cos^2(\vec{\Delta}_{\text{emb}}, \Theta^j)$  denotes the maximum of  $\cos^2(\vec{\Delta}_{\text{emb}}, \Theta^j)$ .



## CoPA: Contrastive Prototype and Image Adaptation

$$\min_{\theta_P, \theta_I} \mathcal{L}(\theta_P, \theta_I) := \mathcal{L}_{\text{CE}}\left(\frac{1}{\tau} \mathbf{Z}_I \mathbf{Z}_P^\top, Y_{\text{pseudo}}\right) + \mathcal{L}_{\text{CE}}\left(\frac{1}{\tau} \mathbf{Z}_P \mathbf{Z}_I^\top, Y_{\text{pseudo}}\right)$$

- The discriminative information in gradients are preserved in different sets of parameters;
- The expanded prototypes indicate the cluster structure of the given support data set.



### Algorithm 1 CoPA Algorithm.

**Input:** pre-trained backbone  $f_{\phi^*}$ , number of inner iterations  $n$ , learning rate  $\eta$ , linear transformation heads  $h_{\theta_P}$  and  $h_{\theta_I}$ , temperature coefficient  $\tau$ .

**Output:** the optimal parameters for linear transformation heads  $\theta_P^*$  and  $\theta_I^*$ .

# Sample a task

**Sample** a new support data set  $\mathcal{D}_{\mathcal{T}} = \{\mathbf{X}, \mathbf{Y}\}$ ;

**Generate** pseudo labels  $Y_{\text{pseudo}} = \{0, 1, \dots, |\mathcal{D}_{\mathcal{T}}| - 1\}$ ;

# Performing contrastive prototype-image adaptation

**for**  $i = 1$  to  $n$  **do**

**Obtain** the prototype and instance representations:

$$\mathbf{Z}_P = h_{\theta_P}(Y Y^\top f_{\phi^*}(\mathbf{X}));$$

$$\mathbf{Z}_I = h_{\theta_I}(f_{\phi^*}(\mathbf{X}));$$

**Compute** SCE loss  $\mathcal{L}(\theta_P, \theta_I)$  in Eq. (3);

**Update** parameters:

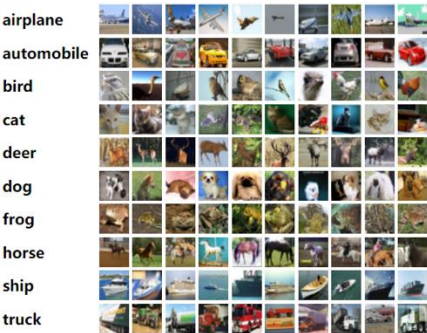
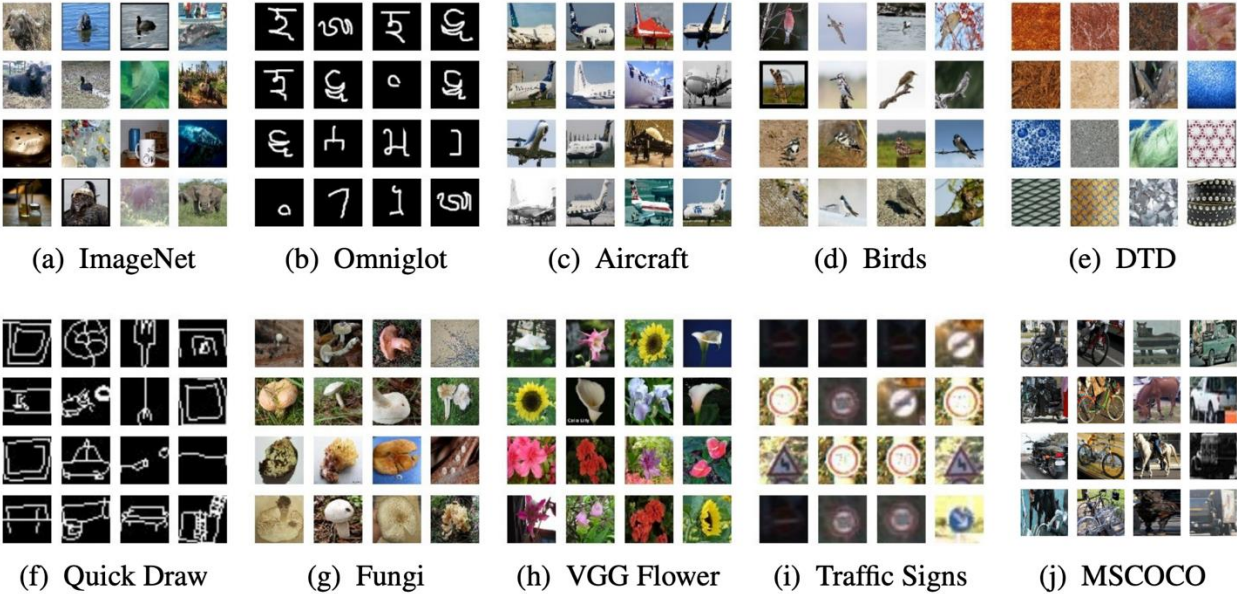
$$\theta_P \leftarrow \theta_P - \eta \nabla_{\theta_P} \mathcal{L}(\theta_P, \theta_I);$$

$$\theta_I \leftarrow \theta_I - \eta \nabla_{\theta_I} \mathcal{L}(\theta_P, \theta_I);$$

**end for**

# Experiments

## Meta-Dataset



Triantafillou et al., Meta-dataset: A dataset of datasets for learning to learn from few examples, ICLR 2020.  
Requeima et al. Fast and flexible multi-task classification using conditional neural adaptive processes. NeurIPS 2019.



## Experiments

### Main results: train on ImageNet only

Table 2: Results on Meta-Dataset under the “train on ImageNet only” setting. Under the “train on ImageNet only” setting, only ImageNet is treated as “seen domain” while the remaining as “unseen domains”. Mean accuracy and 95% confidence interval are reported.

Datasets	Main Results								More Learning Modules		
	Finetune	ProtoNets(large)	BOHB	FP-MAML	AFP-MAML	FLUTE	URL	CoPA	TSA	TA <sup>2</sup> -Net	CoPA+TSA
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	<b>57.7±1.1</b>	<b>57.7 ± 1.1</b>	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	<b>70.9±1.2</b>	<b>73.5 ± 1.2</b>	72.8 ± 1.2	73.3 ± 1.2
Aircraft	<b>68.7±1.3</b>	58.0±1.0	54.1±0.9	56.0±1.0	63.4±1.1	48.5±1.0	57.6±1.0	<b>61.6±1.0</b>	<b>65.1 ± 1.1</b>	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	<b>74.2±0.9</b>	74.0 ± 0.9	73.8 ± 0.9	<b>74.7 ± 0.9</b>
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	<b>77.0±0.7</b>	76.8 ± 0.7	76.6 ± 0.7	<b>77.6 ± 0.7</b>
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	<b>61.3±1.0</b>	64.6 ± 1.0	63.9 ± 1.0	<b>64.7 ± 1.0</b>
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	<b>48.0±1.1</b>	46.8 ± 1.1	47.6 ± 1.1	<b>48.3 ± 1.1</b>
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	<b>88.9±0.6</b>	89.8 ± 0.6	89.6 ± 0.6	<b>90.6 ± 0.6</b>
Traffic Sign	<b>66.8±1.3</b>	58.1±1.1	51.8±1.0	48.8±1.1	60.8±1.3	46.5±1.1	61.2±1.2	<b>63.8±1.1</b>	82.2 ± 0.9	<b>87.7 ± 0.8</b>	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	<b>56.1±1.0</b>	55.8 ± 1.0	51.3 ± 1.2	<b>57.4 ± 1.0</b>
MNIST	-	-	-	-	-	80.8±0.8	86.2±0.7	<b>87.3±0.7</b>	93.6 ± 0.6	94.7 ± 0.5	<b>95.1 ± 0.6</b>
CIFAR-10	-	-	-	-	-	65.4±0.8	69.5±0.8	<b>72.4±0.8</b>	<b>79.6 ± 0.8</b>	76.1 ± 0.9	76.8 ± 0.8
CIFAR-100	-	-	-	-	-	52.7±1.1	62.0±1.0	<b>62.7±1.0</b>	<b>70.6 ± 1.0</b>	65.7 ± 1.1	68.9 ± 0.9
Average Seen	45.8	53.7	51.9	49.5	52.8	46.9	57.3	<b>57.7</b>	<b>57.7</b>	57.5	57.5
Average Unseen	-	-	-	-	-	56.5	66.6	<b>68.7</b>	72.7	71.9	<b>73.2</b>
Average All	-	-	-	-	-	55.8	65.9	<b>67.7</b>	71.6	70.8	<b>72.0</b>
Average Rank	9.3	7.2	8.0	9.0	7.1	10.1	5.3	<b>4.1</b>	2.5	3.2	<b>2.2</b>

<sup>1</sup> The results on URL, TSA, TA<sup>2</sup>-Net and our proposed methods are reproduced with 5 random seeds and reported as the average of the 5 reproduction. The ranks only consider the first 10 datasets and are calculated only with the methods in the table.

## Experiments

### Main results: train on all datasets

Table 1: **Results on Meta-Dataset under the “train on all datasets” setting.** Under the “train on all datasets” setting, the first 8 datasets are treated as “seen domains” while the last 5 are treated as “unseen domains”. Mean accuracy and 95% confidence interval are reported.

Datasets	Main Results								More Learning Modules		
	CNAPS	S-CNAPS	SUR	URT	Tri-M	FLUTE	URL	CoPA	TSA	TA <sup>2</sup> -Net	CoPA + TSA
<b>ImageNet</b>	50.8±1.1	58.4±1.1	56.2±1.0	56.8±1.1	<b>58.6±1.0</b>	51.8±1.1	57.3±1.1	<b>57.8±1.1</b>	57.4±1.1	57.5±1.1	<b>57.8±1.1</b>
<b>Omniglot</b>	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	<b>94.3±0.5</b>	<b>94.7±0.4</b>	94.6±0.4	94.6±0.4
<b>Aircraft</b>	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	<b>88.8±0.5</b>	88.9±0.5	89.0±0.5	<b>89.3±0.5</b>
<b>Birds</b>	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	<b>80.8±0.8</b>	80.8±0.8	80.7±0.8	<b>81.2±0.8</b>
<b>Textures</b>	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	<b>77.8±0.7</b>	77.1±0.7	76.9±0.7	<b>77.8±0.7</b>
<b>Quick Draw</b>	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	<b>82.8±0.6</b>	82.2±0.6	82.2±0.6	<b>82.7±0.6</b>
<b>Fungi</b>	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	<b>69.5±1.0</b>	67.4±1.0	68.1±1.0	<b>69.0±1.0</b>
<b>VGG Flower</b>	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	<b>92.7±0.5</b>	92.5±0.5	92.4±0.5	<b>93.0±0.5</b>
<b>Traffic Sign</b>	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	<b>66.6±1.1</b>	83.5±0.9	88.3±0.8	<b>88.5±0.9</b>
<b>MSCOCO</b>	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	<b>56.3±1.0</b>	55.3±1.1	49.9±1.2	<b>57.9±1.0</b>
<b>MNIST</b>	-	94.6±0.4	94.3±0.4	90.6±0.5	<b>96.2±0.3</b>	<b>95.6±0.5</b>	94.7±0.4	<b>95.2±0.4</b>	96.7±0.4	97.0±0.4	<b>97.5±0.4</b>
<b>CIFAR-10</b>	-	74.9±0.7	66.5±0.9	67.0±0.8	75.4±0.8	<b>78.6±0.7</b>	71.9±0.8	<b>73.0±0.8</b>	<b>80.3±0.8</b>	76.6±0.9	78.7±0.8
<b>CIFAR-100</b>	-	61.3±1.1	56.9±1.1	57.3±1.0	62.0±1.0	<b>67.1±1.0</b>	62.9±1.0	<b>63.4±1.0</b>	70.6±1.0	64.5±1.2	<b>70.9±0.9</b>
<b>Average Seen</b>	71.6	73.7	75.9	77.4	76.2	76.2	79.9	<b>80.6</b>	80.1	80.2	<b>80.7</b>
<b>Average Unseen</b>	-	67.4	64.1	62.9	69.9	69.9	69.4	<b>70.9</b>	77.3	75.2	<b>78.7</b>
<b>Average All</b>	-	71.2	71.3	71.8	73.8	73.8	75.8	<b>76.8</b>	79.2	78.3	<b>79.9</b>
<b>Average Rank</b>	10.3	8.7	8.7	7.1	7.9	7.8	4.5	<b>3.0</b>	3.1	3.3	<b>2.6</b>

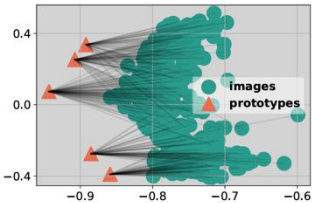
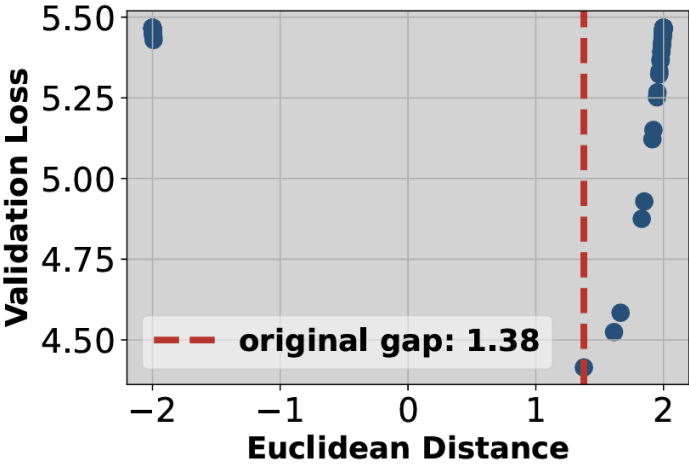
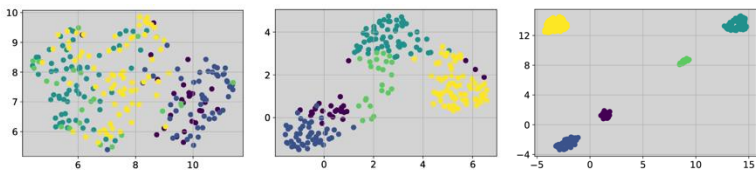
<sup>1</sup> For fairness, the results of URL, TSA, TA<sup>2</sup>-Net, and our proposed CoPA methods are reproduced with 5 random seeds, and we report the average of the 5 reproductions in the table. Particularly, although the reported performance of URL is lower than that in the original paper, the reproduction results are consistent with those reported on their [project website](#). The ranks are calculated only with the first 10 datasets and only with the methods mentioned above.



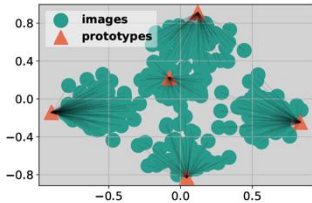
# Mind the Gap Between Prototypes and Images in Cross-domain Finetuning

## Experiments

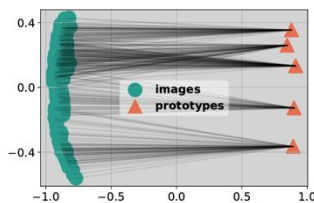
### Qualitative Analyses



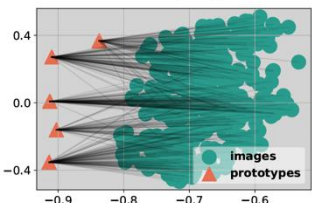
(a) Aircraft Emb:  $\|\vec{\Delta}\| = 0.14$



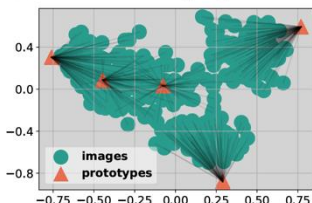
(b) Aircraft URL:  $\|\vec{\Delta}\| = 0.04$



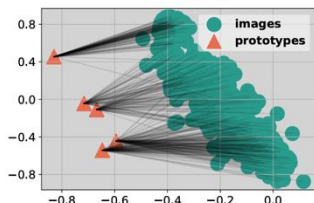
(c) Aircraft CoPA:  $\|\vec{\Delta}\| = 1.77$



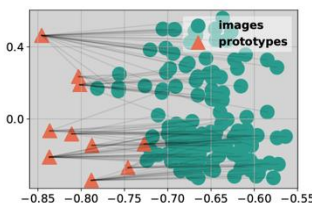
(d) DTD Emb:  $\|\vec{\Delta}\| = 0.22$



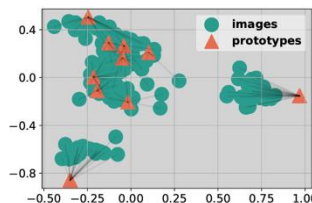
(e) DTD URL:  $\|\vec{\Delta}\| = 0.06$



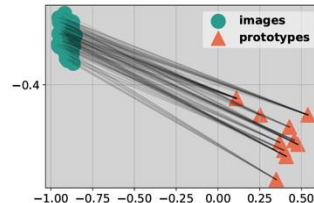
(f) DTD CoPA:  $\|\vec{\Delta}\| = 0.63$



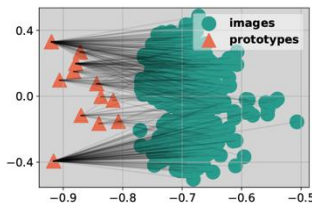
(g) QuickDraw Emb:  $\|\vec{\Delta}\| = 0.15$



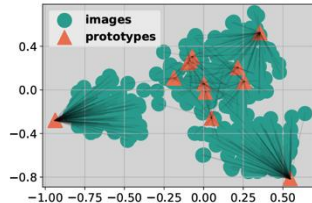
(h) QuickDraw URL:  $\|\vec{\Delta}\| = 0.04$



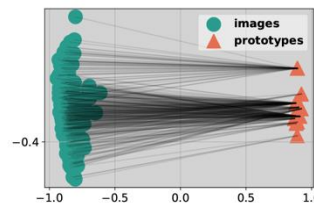
(i) QuickDraw CoPA:  $\|\vec{\Delta}\| = 1.33$



(j) Fungi Emb:  $\|\vec{\Delta}\| = 0.20$



(k) Fungi URL:  $\|\vec{\Delta}\| = 0.08$



(l) Fungi CoPA:  $\|\vec{\Delta}\| = 1.75$

## Experiments

### Further Analysis - Alignment

$$\mathcal{L}_{\text{SCE}} = -\frac{1}{|\mathcal{D}_{\mathcal{T}}|} \sum_{i=1}^{|\mathcal{D}_{\mathcal{T}}|} \log \frac{\exp(\mathbf{z}_i^{\top} \mathbf{c}_i)}{\sum_{j=1}^{|\mathcal{D}_{\mathcal{T}}|} \exp(\mathbf{z}_i^{\top} \mathbf{c}_j)} - \frac{1}{|\mathcal{D}_{\mathcal{T}}|} \sum_{i=1}^{|\mathcal{D}_{\mathcal{T}}|} \log \frac{\exp(\mathbf{c}_i^{\top} \mathbf{z}_i)}{\sum_{j=1}^{|\mathcal{D}_{\mathcal{T}}|} \exp(\mathbf{c}_i^{\top} \mathbf{z}_j)}. \quad (4)$$

**Theorem 5.1.** Given a set of normalized finite support data representation  $\mathcal{Z} = \{(\mathbf{z}_i, y_i)\}_{i=1}^n$  and a set of normalized prototype representations  $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^n$ , where  $\|\mathbf{z}\|_2 = 1$  for  $\forall \mathbf{z} \in \mathcal{Z}$  and  $\|\mathbf{c}\|_2 = 1$  for  $\forall \mathbf{c} \in \mathcal{C}$ , then we are able to obtain a lower bound of SCE loss in Eq. (4):

$$\mathcal{L}_{\text{SCE}} \geq -\frac{2}{n} \sum_{i=1}^n \mathbf{z}_i^{\top} \mathbf{c}_i + \frac{2}{n} \sum_{i=1}^n \sum_{k=1}^{N_C} \frac{|\mathcal{C}_k|}{n} \mathbf{z}_i^{\top} \mathbf{c}_k$$

where  $\mathcal{C}_k$  denotes the set of support data of the class  $k$  and  $N_C$  denotes the number of classes.

- The lower bound of SCE loss functions similarly to the NCC loss, which aims at maximizing the similarity between each sample and its corresponding prototype while minimizing the similarities between the sample and all prototypes. Minimizing the second term is equivalent to enlarging the gap.
- The similarities between images and prototypes are minimized with the weights calculated based on the size of the class set. Thus, the similarities between samples and the prototypes involving more samples will be significantly reduced.

# Summary

- ❑ **Empirically**, we find that there exists a gap, which resembles the modality gap, between prototype and image instance embeddings extracted from a frozen backbone. And the shared representation transformation tends to shrink the gap between prototype and image representations.
- ❑ **Theoretically**, we find that the shared transformation potentially drop the discriminative information in gradients and constrains learning representations where the gap is preserved.
- ❑ **Technically**, we propose a simple yet effective method, CoPA, to finetune two different transformations respectively for prototypes and image instances with SCE loss.
- ❑ **Empirically**, extensive experiments under several settings are conducted to verify the effectiveness of CoPA in improving generalization performance and demonstrate that CoPA can enlarge the gap between prototypes and image instances and learn a better image representation cluster for each class.

Thank You!

Paper



Code

