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

Hongduan Tian[1,2], Feng Liu[3], Zhanke Zhou[1,2], Tongliang Liu[4], Chengqi Zhang[5], Bo Han[1,2]

[1]TMLR Group, Hong Kong Baptist University

[2]Department of Computer Science, Hong Kong Baptist University

[3]TMLR Group, University of Melbourne, [4]Sydney AI Centre, The University of Sydney

[5]Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University

Outline

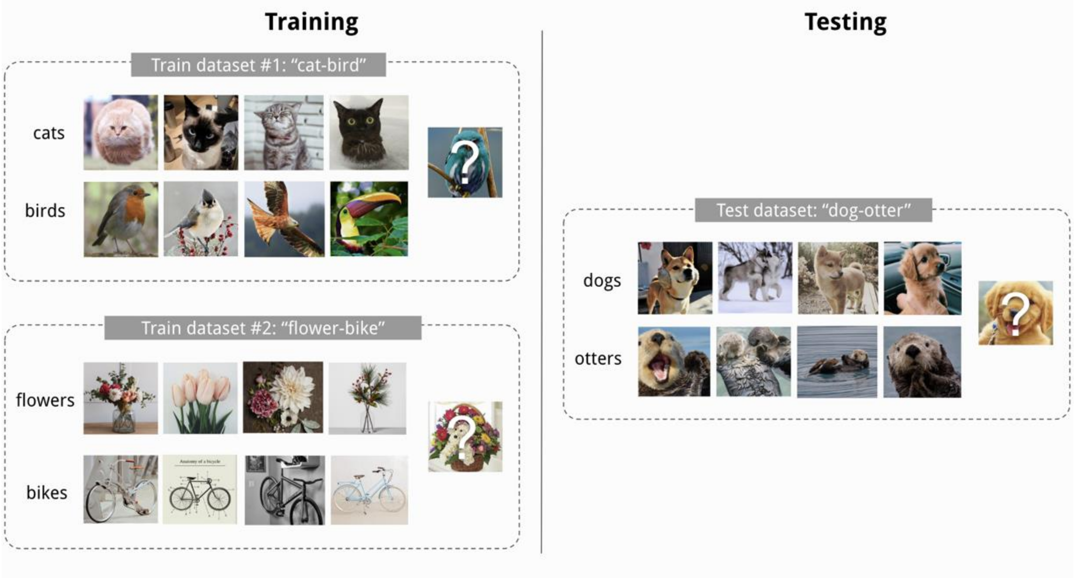
- Background
- Revisit Previous Adaptation Strategy
- Contrastive Prototype-Image Adaptation (CoPA)
- Summary

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

Hongduan Tian[1], Feng Liu[2], Zhanke Zhou[1], Tongliang Liu[3], Chengqi Zhang[4], Bo Han[1†]
[1]TMLR Group, Department of Computer Science, Hong Kong Baptist University
[2]TMLR Group, University of Melbourne [3]Sydney AI Center, The University of Sydney
[4]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

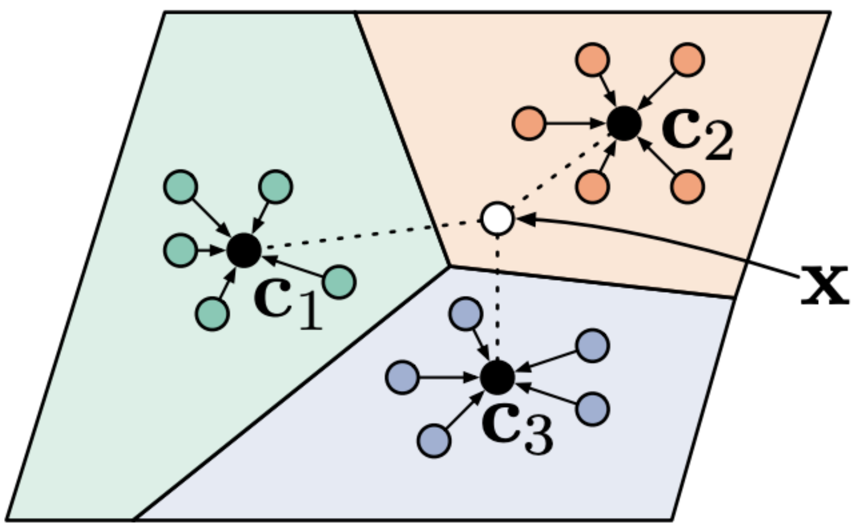
Challenges in CFC:

- **Varied** numbers of classes & shots in CFC tasks;
- **Distribution discrepancy** among datasets.

Phillip Lippe, Tutorial 16: Meta-Learning - Learning to Learn, UvA DL Notebooks v1.2 Documentation.

Preliminary: Prototypical Networks

Few-shot classification with prototypes



- **Construct prototypes:**

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

- **Calculate similarities/distances:**

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

NCC-based loss

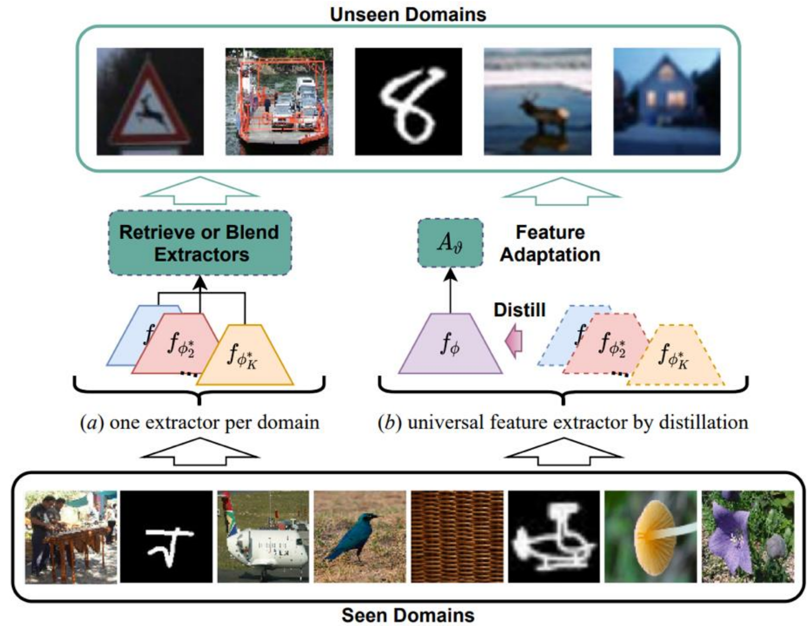
$$p(\hat{y} = y_i | x_i) = \frac{exp\left(-d(x, c_i)\right)}{\sum_j exp\left(-d(x, c_j)\right)}$$

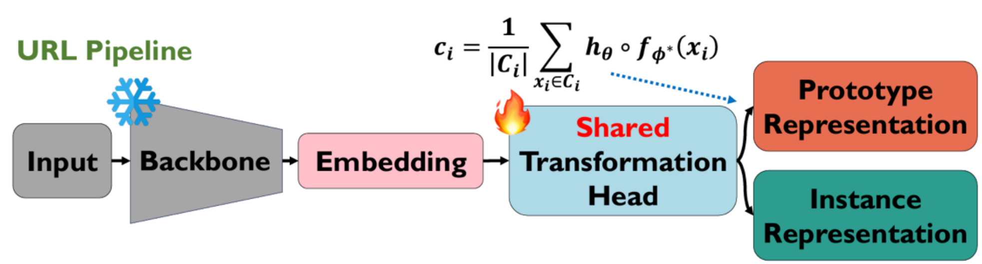
Snell et al., Prototypical networks for few-shot learning, NIPS 2017.

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 simple transformation module on top of the pretrained backbone.

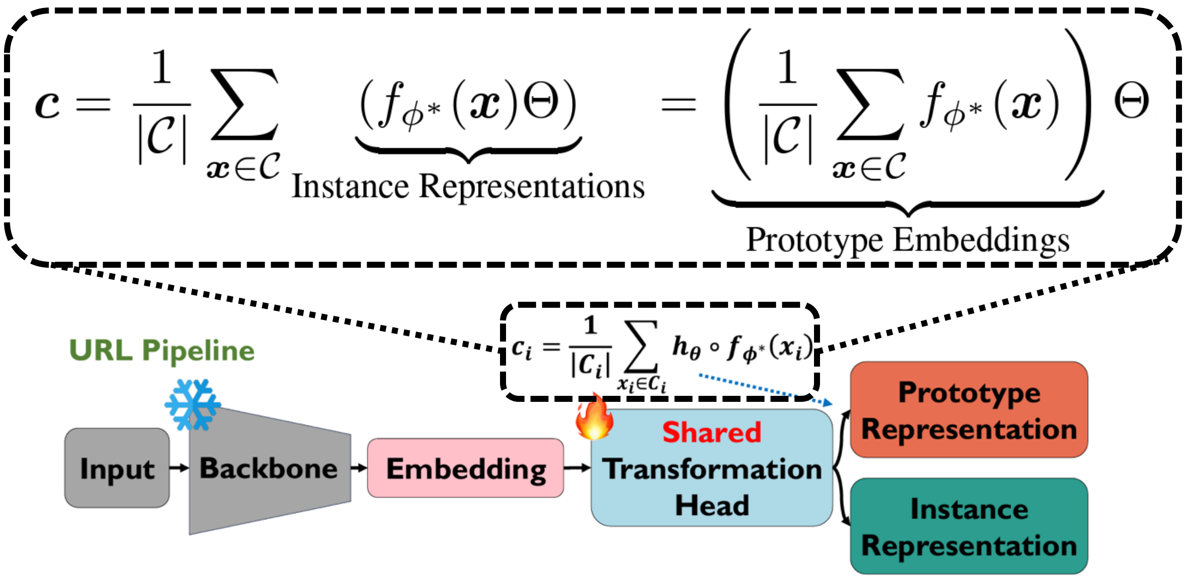




Li et al., Universal representation learning from multiple domains for few-shot classification, ICCV 2021.

Motivation

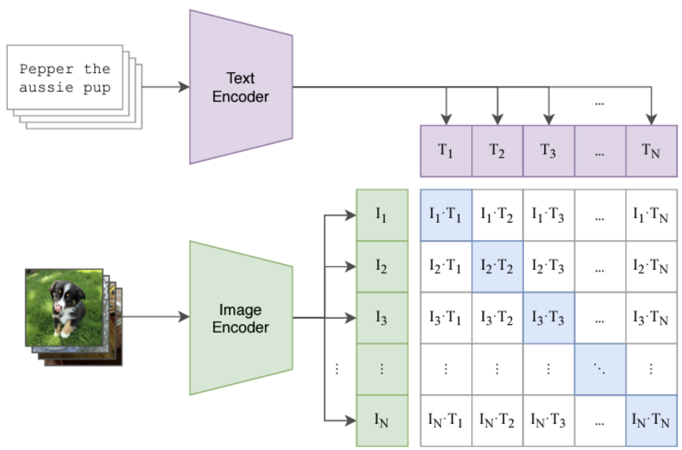
An Implicit Assumption

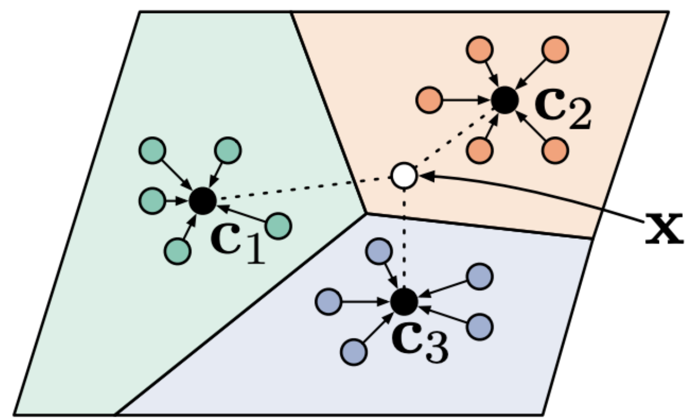


Instance-level and prototype-level embeddings share the same representation transformation.

Motivation

An intuition of prototypes





Text: Abstract information of a set of image instances.

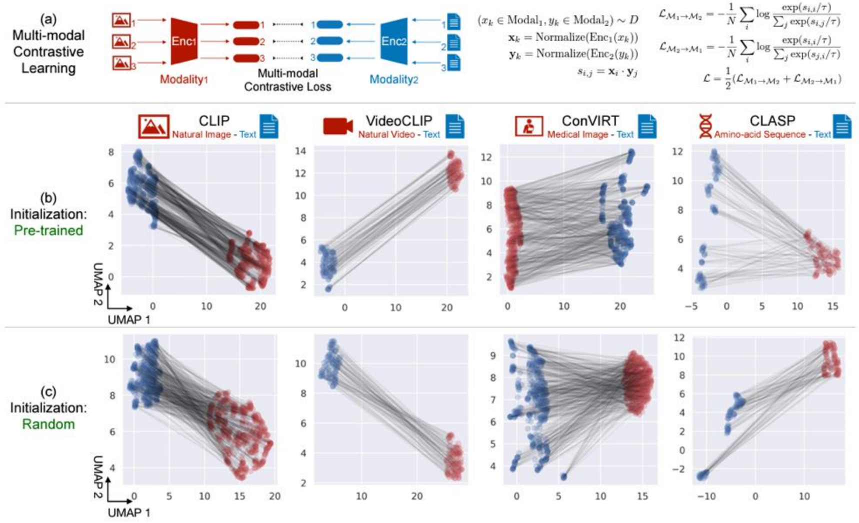
Prototype: Information commonly shared across of images in a class.

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

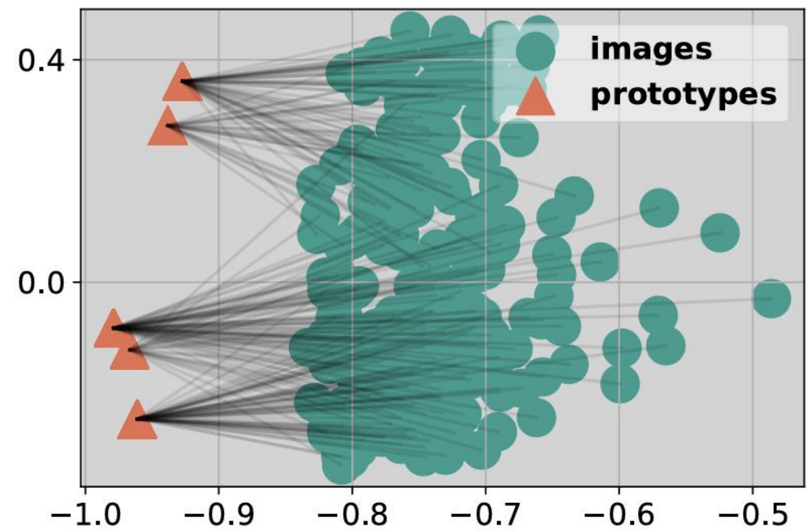
Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021.

Revisit Previous Adaptation Strategy

Gap between prototypes and images



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



There exist modality gaps between different modalities of data

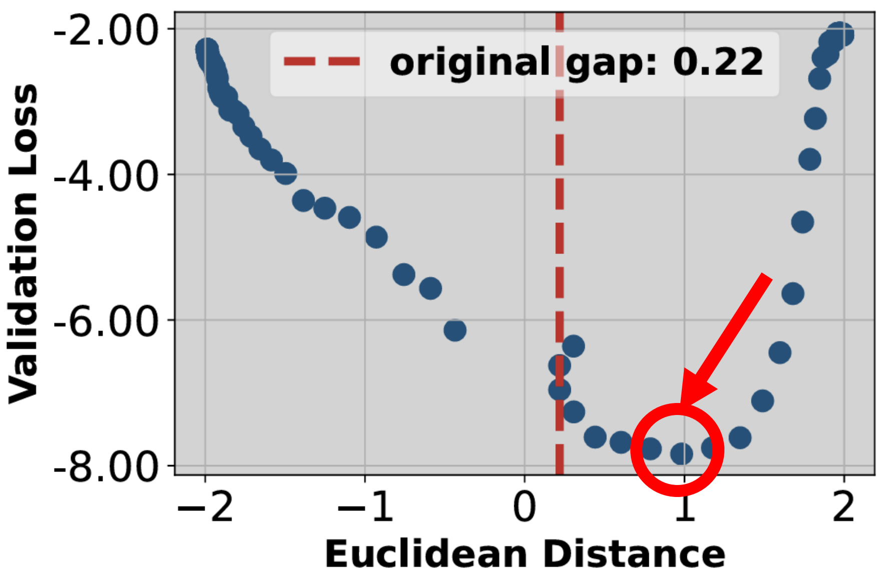
There also exist a gap between prototype and image embeddings!

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

Revisit Previous Adaptation Strategy

Larger gap facilitates better generalization performance

Slightly enlarging the gap improves the generalization performance!



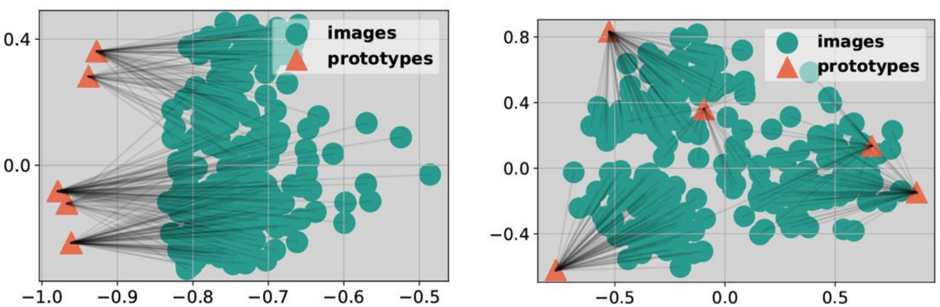
Potential Reasons:

- **[Overfitting]** The prototype representations are generated directly from image instance representations. Thus, instances are naturally similar to prototypes;
- **[Alignment]** Representations of prototypes and image instances are not well aligned.

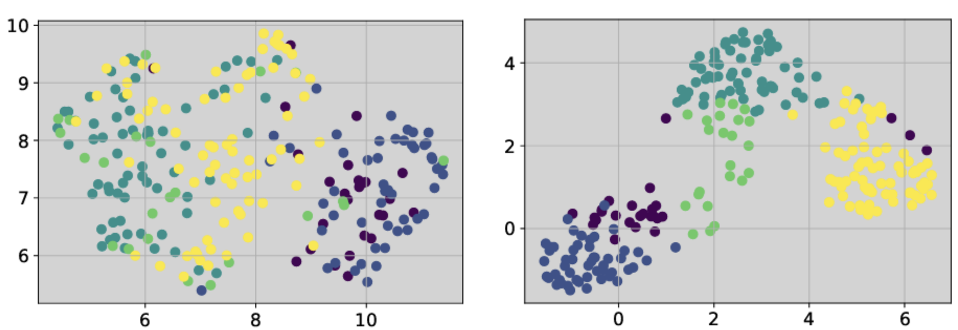
Revisit Previous Adaptation Strategy

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 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 $\mathcal{Z} = \{(\boldsymbol{z}_i, y_i)\}_{i=1}^{n}$, where $||\boldsymbol{z}||_2 = 1$ for $\forall \boldsymbol{z} \in \mathcal{Z}$ and N_C classes are included, then we have a lower bound of the NCC-based loss in Eq. (1):*

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

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

$$\mathcal{L}(\Theta_{\mathrm{P}}, \Theta_{\mathrm{I}}) = -\frac{1}{|\mathcal{D}_\mathcal{T}|}\mathrm{Tr}\left(f_{\phi^*}(\boldsymbol{X})\Theta_{\mathrm{I}}(YY^\top f_{\phi^*}(\boldsymbol{X})\Theta_{\mathrm{P}})^\top\right) + \frac{\alpha}{|\mathcal{D}_\mathcal{T}|}\mathrm{Tr}\left(f_{\phi^*}(\boldsymbol{X})\Theta_{\mathrm{I}}\Theta_{\mathrm{I}}^\top f_{\phi^*}(\boldsymbol{X})^\top\right),$$

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

$$\nabla_{\Theta_{\mathrm{P}}}\mathcal{L}(\Theta_{\mathrm{P}}, \Theta_{\mathrm{I}}) = -\frac{1}{|\mathcal{D}_\mathcal{T}|}\Theta_{\mathrm{I}}^\top f_{\phi^*}(\boldsymbol{X})^\top YY^\top f_{\phi^*}(\boldsymbol{X}),$$

$$\nabla_{\Theta_{\mathrm{I}}}\mathcal{L}(\Theta_{\mathrm{P}}, \Theta_{\mathrm{I}}) = -\frac{1}{|\mathcal{D}_\mathcal{T}|}\Theta_{\mathrm{P}}^\top f_{\phi^*}(\boldsymbol{X})^\top YY^\top f_{\phi^*}(\boldsymbol{X}) + \frac{2\alpha}{|\mathcal{D}_\mathcal{T}|}\Theta_{\mathrm{I}}^\top f_{\phi^*}(\boldsymbol{X})^\top f_{\phi^*}(\boldsymbol{X}).$$

Revisit Previous Adaptation Strategy

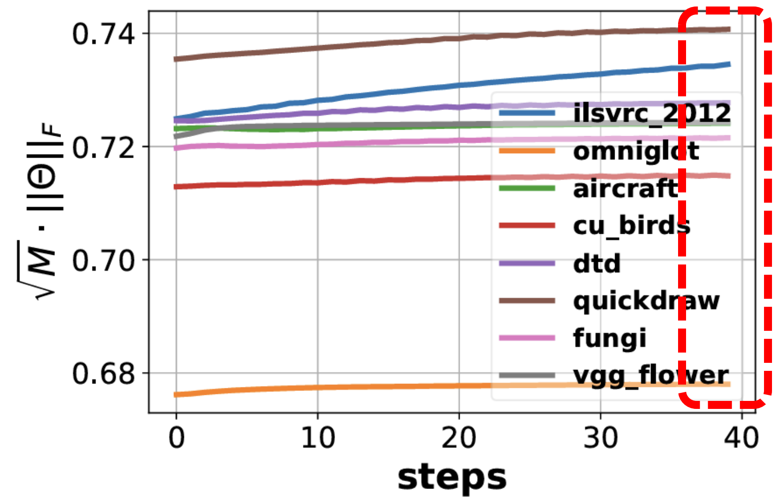
Further Analyses

The empirical results regarding the coefficient of the upper bound indicate that the shared transformation will shrink the gap since **<u>the coefficient is consistently smaller than 1.0</u>**.

Theorem 3.2 (The shared transformation). *Consider a support data set $\mathcal{D}_\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{|\mathcal{D}_\mathcal{T}|}$ composed of N_C classes and a frozen pretrained backbone $f_{\phi^*} : \mathbb{R}^{d_{\text{in}}} \rightarrow \mathbb{R}^d$ parameterized with the optimal parameters ϕ^*. 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 $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{|\mathcal{D}_\mathcal{T}|} = \{f_{\phi^*}(\boldsymbol{x}_i)\Theta\}_{i=1}^{|\mathcal{D}_\mathcal{T}|}$, and the prototype representations $\mathcal{C} = \{\boldsymbol{c}_i\}_{i=1}^{N_C}$, where $\boldsymbol{c}_i = \frac{1}{|\mathcal{C}_i|}\sum_{\boldsymbol{z}' \in \mathcal{C}_i} \boldsymbol{z}' = \frac{1}{|\mathcal{C}_i|}\sum_{\boldsymbol{x}' \in \mathcal{C}_i} f_{\phi^*}(\boldsymbol{x}')\Theta$. Then we can obtain the bounds of the representation gap:*

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

where $\vec{\Delta}_{\text{emb}} = \frac{1}{|\mathcal{D}_\mathcal{T}|}\sum_{\boldsymbol{x} \in \mathcal{D}_\mathcal{T}} f_{\phi^*}(\boldsymbol{x}) - \frac{1}{N_C}\sum_{b=1}^{N_C}\left(\frac{1}{|\mathcal{C}_b|}\sum_{\boldsymbol{x}' \in \mathcal{C}_b} f_{\phi^*}(\boldsymbol{x}')\right)$ 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 $\cos^2(\vec{\Delta}_{\text{emb}}, \Theta^j)$.

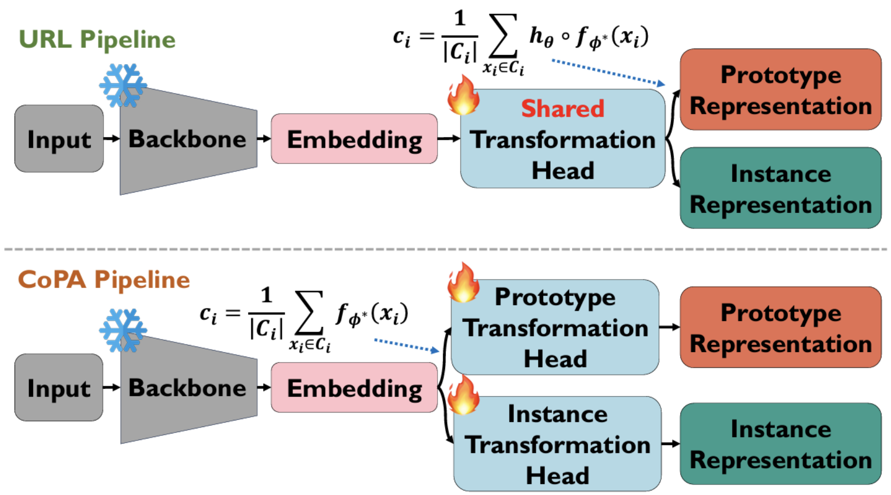


The coefficient is consistently smaller than 1.0.

CoPA: Contrastive Prototype and Image Adaptation

$$\min_{\theta_\mathrm{P},\theta_\mathrm{I}} \mathcal{L}(\theta_\mathrm{P},\theta_\mathrm{I}) := \mathcal{L}_\mathrm{CE}(\frac{1}{\tau}\boldsymbol{Z}_\mathrm{I}\boldsymbol{Z}_\mathrm{P}^\top, Y_\mathrm{pseudo}) + \mathcal{L}_\mathrm{CE}(\frac{1}{\tau}\boldsymbol{Z}_\mathrm{P}\boldsymbol{Z}_\mathrm{I}^\top, Y_\mathrm{pseudo})$$

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

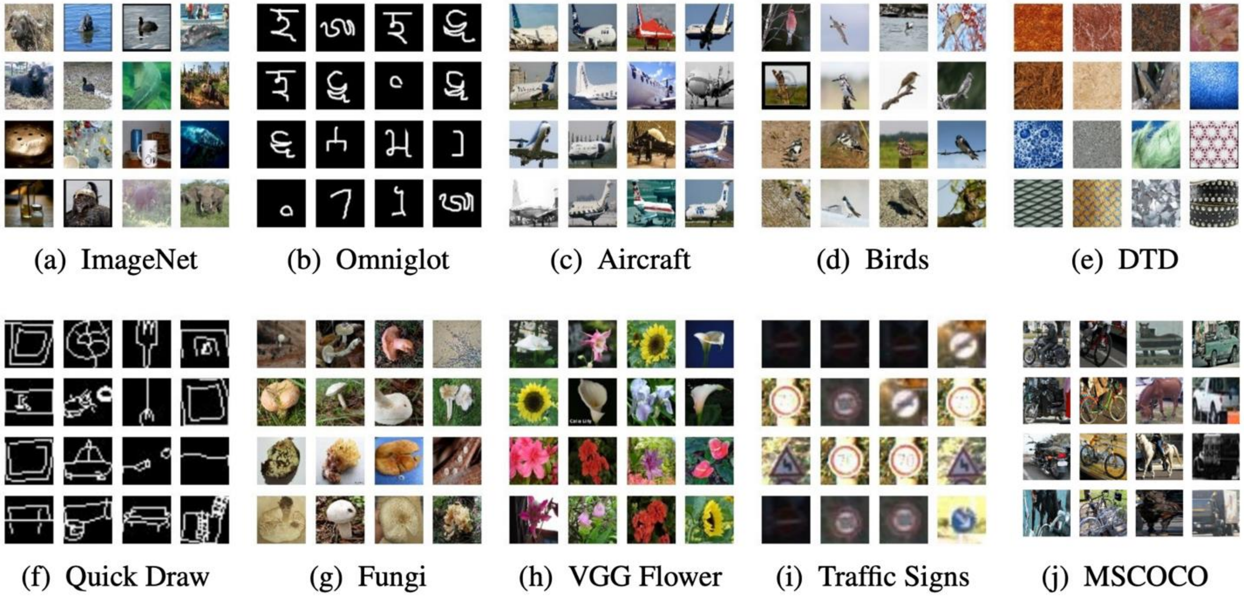


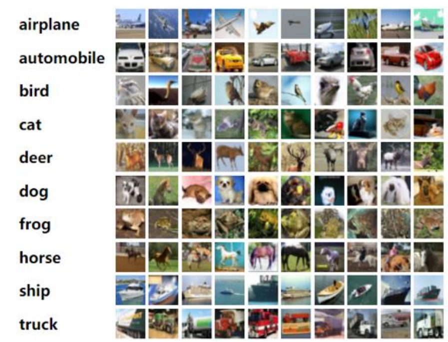
Algorithm 1 **CoPA Algorithm.**

Input: pre-trained backbone f_{ϕ^*}, number of inner iterations n, learning rate η, linear transformation heads h_{θ_P} and h_{θ_I}, temperature coefficient τ.
Output: the optimal parameters for linear transformation heads θ_P^* and θ_I^*.
 # Sample a task
Sample a new support data set $\mathcal{D}_\mathcal{T} = \{\boldsymbol{X}, Y\}$;
Generate pseudo labels $Y_\mathrm{pseudo} = \{0, 1, ..., |\mathcal{D}_\mathcal{T}| - 1\}$;
 # Performing contrastive prototype-image adaptation
for $i = 1$ **to** n **do**
  Obtain the prototype and instance representations:
    $\boldsymbol{Z}_\mathrm{P} = h_{\theta_\mathrm{P}}(YY^\top f_{\phi^*}(\boldsymbol{X}))$;
    $\boldsymbol{Z}_\mathrm{I} = h_{\theta_\mathrm{I}}(f_{\phi^*}(\boldsymbol{X}))$;
  Compute SCE loss $\mathcal{L}(\theta_\mathrm{P}, \theta_\mathrm{I})$ in Eq. (3);
  Update parameters:
    $\theta_\mathrm{P} \leftarrow \theta_\mathrm{P} - \eta\nabla_{\theta_\mathrm{P}}\mathcal{L}(\theta_\mathrm{P}, \theta_\mathrm{I})$;
    $\theta_\mathrm{I} \leftarrow \theta_\mathrm{I} - \eta\nabla_{\theta_\mathrm{I}}\mathcal{L}(\theta_\mathrm{P}, \theta_\mathrm{I})$;
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^2-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	**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**

[1] The results on URL, TSA, TA^2-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 domians" while the last 5 are treated as "unseen domains". Mean accuracy and 95% confidence interval are reported.

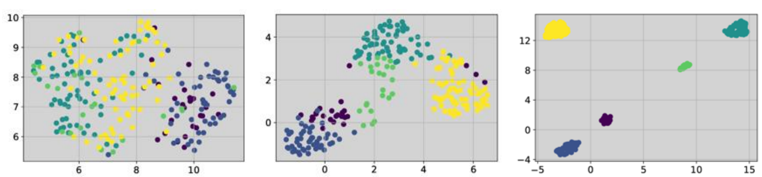
Datasets	Main Results								More Learning Modules		
	CNAPS	S-CNAPS	SUR	URT	Tri-M	FLUTE	URL	**CoPA**	TSA	TA^2-Net	**CoPA + TSA**
ImageNet	50.8±1.1	58.4 ±1.1	56.2 ± 1.0	56.8 ± 1.1	**58.6 ± 1.0**	51.8 ± 1.1	57.3 ± 1.1	**57.8 ± 1.1**	57.4 ± 1.1	57.5 ± 1.1	**57.8 ± 1.1**
Omniglot	91.7±0.5	91.6 ± 0.6	94.1 ± 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**

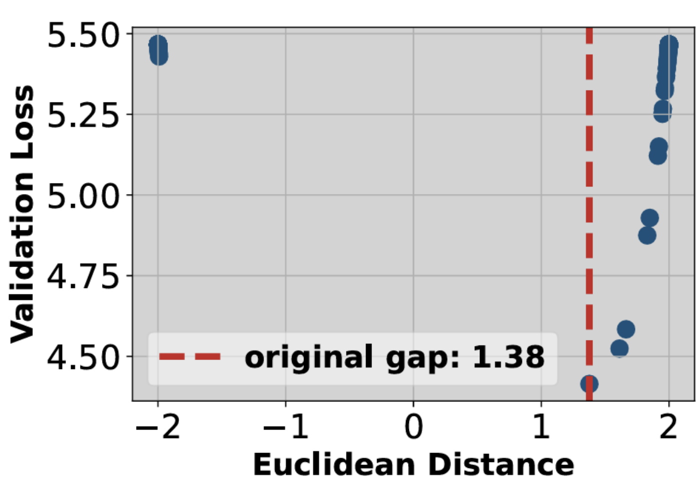
[1] For fairness, the results of URL, TSA, TA^2-Net, and our proposed CoPA methods are reproduced with 5 random seeds, and we report the average of the 5 reproductions in the table. Particularly, although the reported performance of URL is lower than that in the original paper, the reproduction results 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.

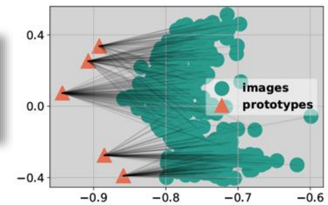
Mind the Gap Between Prototypes and Images in Cross-domain Finetuning

Experiments

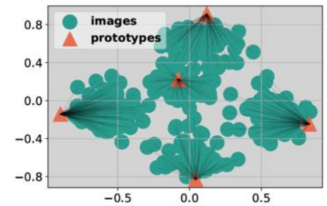
Qualitative Analyses



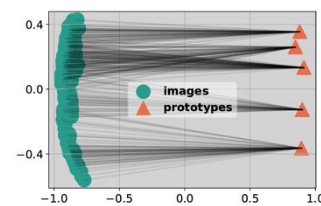




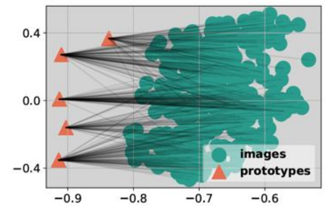
(a) Aircraft Emb: $||\vec{\Delta}|| = 0.14$



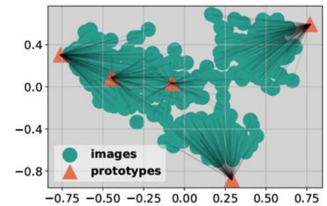
(b) Aircraft URL: $||\vec{\Delta}|| = 0.04$



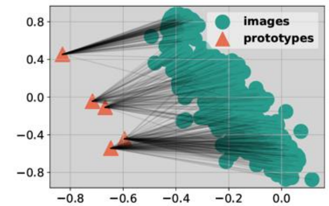
(c) Aircraft CoPA: $||\vec{\Delta}|| = 1.77$



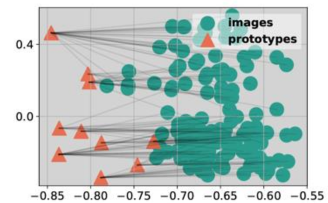
(d) DTD Emb: $||\vec{\Delta}|| = 0.22$



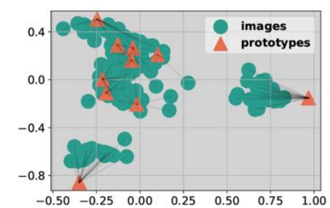
(e) DTD URL: $||\vec{\Delta}|| = 0.06$



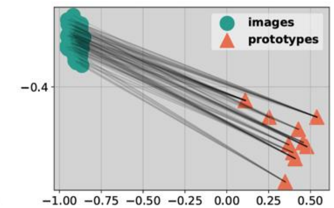
(f) DTD CoPA: $||\vec{\Delta}|| = 0.63$



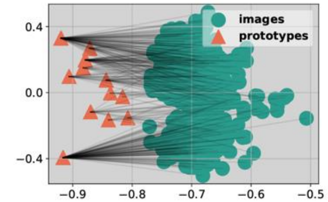
(g) QuickDraw Emb: $||\vec{\Delta}|| = 0.15$



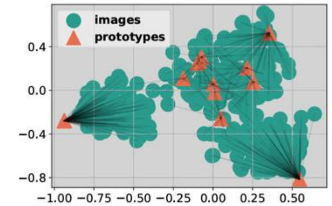
(h) QuickDraw URL: $||\vec{\Delta}|| = 0.04$



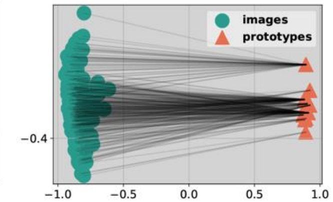
(i) QuickDraw CoPA: $||\vec{\Delta}|| = 1.33$



(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

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

Theorem 5.1. *Given a set of normalized finite support data representation $\mathcal{Z} = \{(\boldsymbol{z}_i, y_i)\}_{i=1}^n$ and a set of normalized prototype representations $\mathcal{C} = \{\boldsymbol{c}_i\}_{i=1}^n$, where $||\boldsymbol{z}||_2 = 1$ for $\forall \boldsymbol{z} \in \mathcal{Z}$ 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):*

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

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