Rethinking Class Relations: Absolute-relative Supervised and Unsupervised Few-shot Learning (Supplementary Material)

Hongguang Zhang[1,2] Piotr Koniusz[3,2] Songlei Jian[5] Hongdong Li[2] Philip H. S. Torr[4]
[1]Systems Engineering Institute, AMS [2]Australian National University [3]Data61/CSIRO
[4]University of Oxford [5]National University of Defense Technology
firstname.lastname@{anu.edu.au[2], data61.csiro.au[3],eng.ox.ac.uk[4]}

Table 1: Top-1 accuracy on the novel test classes of the *tiered*–Imagenet dataset (5-way acc. given). Note that 'U-' variants do not use class labels during learning at all.

Model	1-shot	5-shot
MAML	51.67 ± 1.81	70.30 ± 0.08
Prototypical Net	53.31 ± 0.89	72.69 ± 0.74
Relation Net	54.48 ± 0.93	71.32 ± 0.78
SoSN	58.62 ± 0.92	75.19 ± 0.79
Pixel (Cosine)	27.13 ± 0.94	32.35 ± 0.76
BiGAN(k_{nn})	29.65 ± 0.92	34.08 ± 0.75
U-RN	37.23 ± 0.94	49.54 ± 0.83
U-PN	38.83 ± 0.92	50.64 ± 0.81
U-SoSN	42.07 ± 0.92	56.21 ± 0.76
U-SoSN $+ ArL$	$\mathbf{43.68 \pm 0.91}$	$\mathbf{58.56 \pm 0.74}$

1. Additional results in unsupervised setting.

Below we supplement additional results in the unsupervised setting on two popular datasets, *tiered*–Imagenet and OpenMIC.

***tiered*–Imagenet** consists of 608 classes from ImageNet. We follow the protocol that uses 351 base classes, 96 validation classes and 160 novel test classes.

Open MIC is the Open Museum Identification Challenge (Open MIC) [1], a recent dataset with photos of various museum exhibits, *e.g.* paintings, timepieces, sculptures, glassware, relics, science exhibits, natural history pieces, ceramics, pottery, tools and indigenous crafts, captured from 10 museum spaces according to which this dataset is divided into 10 subproblems. In total, it has 866 diverse classes and 1–20 images per class. Following the setup in SoSN, we combine (*shn+hon+clv*), (*clk+gls+scl*), (*sci+nat*) and (*shx+rlc*) into subproblems *p1*,..., *p4*, and form 12 possible pairs in which subproblem x is used for training and y for testing (x→y).

Results on tiered-Imagenet. Table 1 shows that our proposed unsupervised few-shot learning strategy achieves strong results of 42.31% and 57.21% accuracy for 1- and 5-shot learning protocols. Though it does not outperform the recent supervised works, the performance of many prior works is not provided for this recent dataset. In general, we believe that our ArL approach boosts unsupervised learning and our unsupervised learning yields reasonable accuracy given no training labels being used in this process at all.

Table 2: Ablation studies re. the impact of absolute and relative learning modules given *mini*Imagenet dataset (5-way acc. with Conv-4 backbone given). We denote the same/different class relation as (*bin.*), attribute-based labels (relative and absolute) as (*att.*), *word2wec* embedding (relative and absolute) as (*w2v.*) and absolute class labeling as (*cls.*) RL and AL are Absolute and Relative Learners.

Model	1-shot	5-shot
	Relative Learning	
RelationNet-RL(w2v.)	**53.20**	66.21
RelationNet-RL(att.)	52.38	**66.74**
RelationNet-RL(bin. + att. + w2v.)	52.38	66.73
SoSN-RL(w2v.)	54.31	69.64
SoSN-RL(att.)	54.49	70.21
SoSN-RL(bin. + att. + w2v.)	**55.49**	**70.86**
SalNet-RL(w2v.)	58.15	72.45
SalNet-RL(att.)	58.43	72.91
SalNet-RL(bin. + att. + w2v.)	**58.67**	**73.01**
	Absolute Learning	
Relation Net-AL(cls.)	51.41	66.01
Relation Net-AL(att.)	52.35	66.53
Relation Net-AL(w2v.)	**52.67**	**66.91**
Relation Net-AL(cls.+att.+w2v.)	52.30	66.51
SoSN-AL(cls.)	55.12	70.91
SoSN-AL(att.)	**55.61**	**71.03**
SoSN-AL(w2v.)	54.78	70.85
SoSN-AL(cls.+att.+w2v.)	55.40	71.02
Salnet-AL(cls.)	57.98	72.56
SalNet-AL(att.)	**58.94**	**73.12**
SalNet-AL(w2v.)	58.36	72.96
SalNet-AL(cls.+att.+w2v.)	58.41	73.05

Results on Open MIC. This dataset has very limited (3-15) images for both base and novel classes, which highlights its difference to *mini*Imagenet and *tiered*-Imagenet whose base classes consist of hundreds of images. Table 3 shows that our unsupervised variant of Second-order Similarity Network, U-SoSN with 224×224 res. images outperforms the supervised SoSN on all evaluation protocols. Even without high-resolution training images, our U-SoSN outperforms

Table 3: Evaluations on the Open MIC dataset (Protocol I) (given 5-way 1-shot learning accuracies). Note that the ' U-' variants do not use class labels during learning at all.

Model	$p1 \rightarrow p2$	$p1 \rightarrow p3$	$p1 \rightarrow p4$	$p2 \rightarrow p1$	$p2 \rightarrow p3$	$p2 \rightarrow p4$	$p3 \rightarrow p1$	$p3 \rightarrow p2$	$p3 \rightarrow p4$	$p4 \rightarrow p1$	$p4 \rightarrow p2$	$p4 \rightarrow p3$
Relation Net	71.1	53.6	63.5	47.2	50.6	68.5	48.5	49.7	68.4	45.5	70.3	50.8
SoSN	81.4	65.2	75.1	60.3	62.1	77.7	61.5	82.0	78.0	59.0	80.8	62.5
Pixle (Cosine)	56.8	40.4	57.5	33.3	35.1	46.1	32.3	44.6	45.9	33.5	50.1	34.6
BiGAN(k_{nn})	59.9	43.2	60.3	37.1	38.6	50.2	37.6	48.2	47.5	38.1	55.0	37.8
U-RN	70.3	50.3	64.1	42.9	48.2	61.1	53.2	59.1	55.7	48.5	68.3	45.2
U-PN	70.1	49.7	64.4	43.3	47.9	60.8	52.8	59.4	56.2	49.1	68.8	44.9
U-SoSN	78.6	58.8	74.3	61.1	57.9	72.4	62.3	75.6	73.7	58.5	76.5	54.6
U-SoSN + ArL	**80.2**	**59.7**	**76.1**	**62.8**	**59.6**	**74.4**	**64.2**	**78.4**	**75.2**	**60.1**	**79.2**	**57.3**

the supervised SoSN on many data splits. This observation demonstrates that our unsupervised relation learning is beneficial and practical in case of very limited numbers of training images where the few-shot learning task is closer to the retrieval setting (in Open MIC, images of each exhibit constitute on one class). Most importantly, combining ArL with unsupervised SoSN boosts results further by up to 4%.

2. Ablation study on absolute and relative learners.

Table 2 (*mini*Imagenet as example) illustrates that the semantic relation learner enhanced performance on Relation Net[2], SoSN[3] and SalNet[4]. The results in the table indicate that the performance of few-shot similarity learning can be improved by employing the semantic relation labels at the training stage. For instance, SoSN with attribute soft label (*att.*) achieves 0.6% and 1.7% improvements for 1- and 5-shot compared to the baseline (*SoSN*). Table 2 also demonstrates the ablation studies for absolute learning. It can be seen from the table that the attribute predictor works the best among all options except for SoSN, and applying multiple Absolute Learning modules does not further improve the accuracy. We expect that attributes are a clean form of labels in contrast to *word2vec* and very complementary to class labels *cls.*

3. Remaining experimental details.

For augmentations, we randomly apply resized crop (scale 0.6–1.0, ratio 0.75–1.33), horizontal+vertical flips, rotations (0–360°), and color jitter. Annotated per class attribute vectors (*mini*Imagenet) have 31 attributes (5 environments, 10 colors, 7 shapes, 9 materials). For augmentation keys, taking rotation as example, we set a 4-bit degree to annotate random rotations, '0001' refers to rotations with $0 \sim 90°$, '0010' refers to rotations with $90 \sim 180°$.

References

[1] Piotr Koniusz, Yusuf Tas, Hongguang Zhang, Mehrtash Harandi, Fatih Porikli, and Rui Zhang. Museum exhibit identification challenge for the supervised domain adaptation and beyond. *ECCV*, pages 788–804, 2018. 1

[2] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. *CoRR:1711.06025*, 2017. 2

[3] Hongguang Zhang and Piotr Koniusz. Power normalizing second-order similarity network for few-shot learning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1185–1193. IEEE, 2019. 2

[4] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2