Supplementary Materials: Reinforced Attention for Few-Shot Learning and Beyond

Jie Hong[*1,2], Pengfei Fang[1,2], Weihao Li[2], Tong Zhang[*3], Christian Simon[1,2],
Mehrtash Harandi[4], Lars Petersson[2]
[1]Australian National University, [2]Data61-CSIRO, [3]EPFL, [4]Monash University
jie.hong@anu.edu.au, pengfei.fang@anu.edu.au, weihao.li@data61.csiro.au, tong.zhang@epfl.ch,
christian.simon@anu.edu.au, mehrtash.harandi@monash.edu, lars.petersson@data61.csiro.au

1. Experiment

1.1. Datasets

CIFAR-FS. CIFAR-FS [5] is a few-shot learning dataset including 64 train classes, 16 validation classes and 20 test classes of images. It originates from the split of CIFAR-100 dataset [4].

1.2. Few-shot Learning

1.2.1 *mini*ImageNet

Comparisons with few-shot learning baselines which exploit attention are provided in Table 1. Some visualization examples of attention map are shown in Fig. 1.

1.2.2 CIFAR-FS

The few-shot classification results on CIFAR-FS are shown in Table 2. We apply RAP on the baseline, MetaOptNet-SVM [5]. From the table, RAP model surpasses the corresponding baseline model with a healthy margin.

1.3. Analysis

1.3.1 Validation Set

In our work, ℓ_{train} equals to r_T (if α=1) on the training set while ℓ_{rein} is built by $\{r_1, r_2, ..., r_T\}$ on the validation set. Therefore, we claim that our policy network is jointly trained on training and validation data. If ℓ_{rein} uses training data as ℓ_{train} does, the reuse of training data has limited potential in boosting performance. Thus, we make RAP receive feedback from validation set when building ℓ_{rein}. We accordingly include Table 3 to support this statement.

Comparisons between RAP models and baseline models on *mini*ImageNet/CUB-200-2011 dataset are provided in Table 4. As reported in Table 4, even under the same data settings, RAP models still beat baseline models.

[*]corresponding author

1.3.2 Recurrent Process

The proposed RAP models with different time step T are evaluated and results are shown in Table 5. In this experiment, we test 2-step model, 5-step model and 8-step model on *mini*ImageNet/CUB-200-2011 dataset.

1.3.3 Model Complexity

Table 6 compares the model complexity and memory footprint against baselines in the case of 5-way 1-shot classification on CUB-200-2011. It can be seen from the table that RAP requires more compute, but the inference time is not drastically different from that of other models. Though RAP has a higher complexity, the performance gain almost justifies the additional cost.

1.3.4 Reinforcement Learning Stability

In our work, we have made use of a supervision signal in designing the reward. This in turn improves the stability significantly. In practice, the RAP model can successfully converge despite small fluctuations and such characteristics can be observed in Fig. 4 of the paper. Our experiments show that RAP is not sensitive to varying random seeds. As an indicator, we re-trained RAP-ProtoNet in 5-way 1-shot classification on CUB-200-2011 with 10 different random seeds and obtained mean accuracy 55.94%/75.22% with a standard deviation of 0.83/0.47 (Conv-4/ResNet-10).

1.3.5 Image Variance and Conv Block

Policy gradient methods are known to be sensitive to high variance among training samples ($\mathbf{a}$, $\mathbf{s}$, r). We expect, one of states, $\mathbf{s}^I = \mathbf{I}_o$ can be better observed via attending its more low-level information. Hence, we separately use a shallower and narrower conv block to process $\mathbf{I}_o$ (see Fig. 3 of the paper). Table 7 shows that without conv block, the performance degrades.

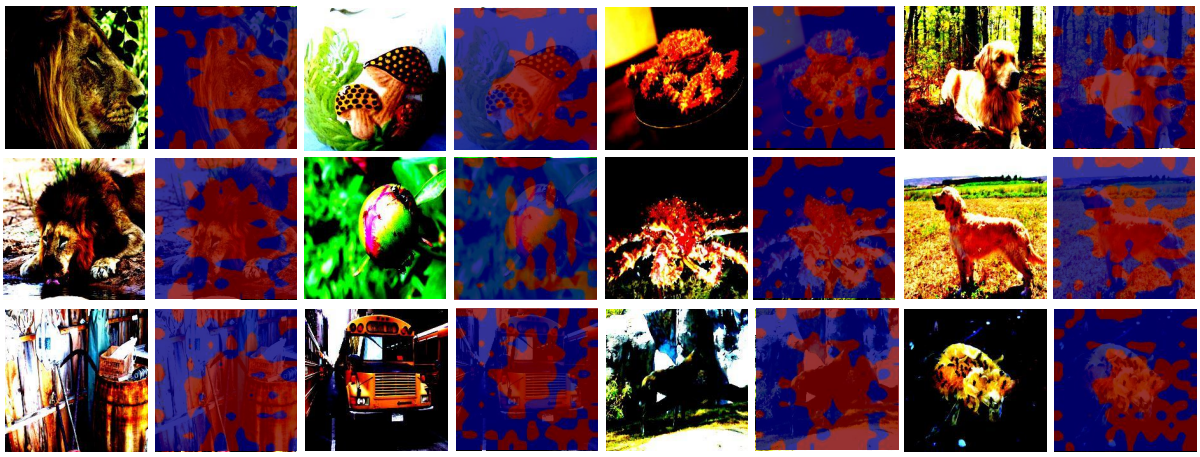

Figure 1. Visualization examples of attention action. These examples are provided from the inference on *mini*ImageNet using ResNet-10 backbone under 5-way 1-shot setting.

		5-way Acc.	
Model	**Backbone**	**1-shot**	**5-shot**
MatchingNet [8]	Conv-6	50.47±0.86	63.19±0.70
RAP-MAML	Conv-6	**52.57±0.61**	66.96±0.52
RAP-ProtoNet	Conv-6	51.72±0.72	**69.18±0.45**
MatchingNet [8]	ResNet-10	54.49±0.81	68.82±0.65
Attention Attractor [6]	ResNet-10	54.95±0.30	63.04±0.30
RAP-MAML	ResNet-10	56.13±0.62	68.74±0.54
RAP-ProtoNet	ResNet-10	53.64±0.60	74.54±0.45
RAP-LaplacianShot	ResNet-10	**71.34±0.19**	**81.98±0.14**
STANet [9]	ResNet-12	58.35±0.57	71.07±0.39
Cross-Attention [3]	ResNet-12	67.19±0.55	80.64±0.35
RAP-LaplacianShot	ResNet-12	**74.29±0.20**	**84.51±0.13**

Table 1. Comparison with attention based solutions in few-shot learning on *mini*ImageNet. We compare the other attention modules under the same backbone networks. The results of MatchingNet can be checked in [1].

		5-way Acc.	
Model	**Backbone**	**1-shot**	**5-shot**
ProtoNet [7]	ResNet-12	72.2±0.7	83.5±0.5
MetaOptNet-RR [5]	ResNet-12	72.6±0.7	84.3±0.5
MetaOptNet-SVM [5]	ResNet-12*	70.99±0.72	84.36±0.48
RAP-MetaOptNet-SVM	ResNet-12	**73.00±0.71**	**85.46±0.47**

Table 2. Few-shot classification on CIFAR-FS dataset. "*" indicates the result obtained from self-implemented networks.

	5-way 1-shot Classification on CUB-200-2011		
Model	**Backbone**	**On train set**	**On val set**
ProtoNet	Conv-4	50.46±0.88	51.61±0.65
	ResNet-10	73.22±0.92	74.48±0.65
RAP-ProtoNet	Conv-4	53.93±0.64	56.71±0.66
	ResNet-10	74.11±0.60	75.17±0.63

Table 3. The results of RAP-ProtoNet with the loss ℓ_{rein} based on the train set or the val set.

References

[1] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019. 2

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 3

[3] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems*, pages 4005–4016, 2019. 2

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[5] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019. 1, 2

[6] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard Zemel. Incremental few-shot learning with attention attractor networks. In *Advances in Neural Information Processing Systems*, pages 5276–5286, 2019. 2

[7] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 2, 3

[8] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 2

[9] Shipeng Yan, Songyang Zhang, Xuming He, et al. A dual attention network with semantic embedding for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9079–9086, 2019. 2

		*mini*ImageNet		CUB-200-2011	
		5-way Acc.		5-way Acc.	
Model	**Backbone**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
MAML [2]	Conv-4*	46.90±0.61	61.43±0.50	57.55±0.67	75.61±0.49
	Conv-6*	47.23±0.59	63.92±0.48	65.07±0.72	79.89±0.48
RAP-MAML	Conv-4	50.07±0.61	63.26±0.52	61.49±0.70	77.15±0.50
	Conv-6	52.57±0.61	66.96±0.52	69.95±0.68	81.48±0.44
ProtoNet [7]	Conv-4*	44.38±0.55	64.72±0.48	51.61±0.65	75.29±0.48
	Conv-6*	50.96±0.56	67.38±0.47	64.72±0.72	81.35±0.43
RAP-ProtoNet	Conv-4	48.51±0.57	65.32±0.48	56.71±0.66	78.70±0.44
	Conv-6	51.72±0.72	69.18±0.45	67.79±0.66	83.78±0.41

Table 4. Comparisons on *mini*ImageNet/CUB-200-2011 dataset. In this study, baseline models are trained on both the train set and validation set. "*" indicates the result obtained from self-implemented networks.

			*mini*ImageNet		CUB-200-2011	
			5-way Acc.		5-way Acc.	
Model	T	**Backbone**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
RAP-MAML	2	Conv-4	49.05±0.62 (↑ 2.15)	62.92±0.50 (↑ 1.49)	60.15±0.71 (↑ 2.60)	75.76±0.50 (↑ 0.15)
	5		50.07±0.61 (↑ 3.17)	63.26±0.52 (↑ 1.83)	61.49±0.70 (↑ 3.94)	77.15±0.50 (↑ 1.54)
	8		49.64±0.61 (↑ 2.74)	62.60±0.50 (↑ 1.17)	63.16±0.70 (↑ 5.61)	77.36±0.48 (↑ 1.75)
	2	Conv-6	51.72±0.62 (↑ 4.49)	66.77±0.50 (↑ 2.85)	68.33±0.70 (↑ 3.26)	80.43±0.45 (↑ 0.54)
	5		52.57±0.61 (↑ 5.34)	66.96±0.52 (↑ 3.04)	69.95±0.68 (↑ 4.88)	81.48±0.44 (↑ 1.59)
	8		52.39±0.61 (↑ 5.16)	67.38±0.49 (↑ 3.46)	69.54±0.69 (↑ 4.47)	81.63±0.43 (↑ 1.74)
RAP-ProtoNet	2	Conv-4	47.41±0.57 (↑ 3.03)	64.81±0.48 (↑ 0.09)	54.01±0.65 (↑ 2.40)	78.12±0.45 (↑ 2.83)
	5		48.51±0.57 (↑ 4.13)	65.32±0.48 (↑ 0.60)	56.71±0.66 (↑ 5.10)	78.70±0.44 (↑ 3.41)
	8		48.48±0.54 (↑ 4.10)	64.89±0.50 (↑ 0.17)	54.83±0.66 (↑ 3.22)	78.69±0.47 (↑ 3.40)
	2	Conv-6	50.85±0.59 (↓ 0.11)	68.16±0.47 (↑ 0.78)	66.27±0.68 (↑ 1.55)	82.58±0.42 (↑ 1.23)
	5		51.72±0.72 (↑ 0.76)	69.18±0.45 (↑ 1.80)	67.79±0.66 (↑ 3.07)	83.78±0.41 (↑ 2.43)
	8		50.22±0.56 (↓ 0.74)	68.68±0.48 (↑ 1.30)	67.04±0.65 (↑ 2.32)	83.42±0.41 (↑ 2.07)

Table 5. Comparisons of RAP models with different time step T on *mini*ImageNet/CUB-200-2011 datasets. "↑" indicates the improvement from RAP model over the baseline network (see Table 4).

	5-way 1-shot Classification on CUB-200-2011					
Model	**Backbone**	**Params**	**Training epochs (100 batches per epoch)**	**Training time (per batch)**	**Inference time**	**GFLOPs**
ProtoNet	Conv-4	0.11M	800	0.47s	0.43s	0.2
SENet-ProtoNet		0.12M	800	0.51s	0.42s	0.2
CBAM-ProtoNet		0.12M	800	0.50s	0.44s	0.2
RAP-ProtoNet	Conv-4	0.16M	800	0.68s	0.46s	1.2

Table 6. Model complexity of different attention models with Conv-4 backbone.

	5-way 1-shot Classification on CUB-200-2011			
Model	**Backbone**	**Params**	**Dimension of $\mathbf{l}_t$**	**Acc.**
RAP-ProtoNet w/o conv	ResNet-10	21.1M	512	73.69±0.65
RAP-ProtoNet share conv with backbone		21.7M	768	74.51±0.62
RAP-ProtoNet	ResNet-10	21.3M	544	75.17±0.63

Table 7. The results of diverse image feature extraction modes of RAP-ProtoNet.