Transductive Episodic-Wise Adaptive Metric for Few-Shot Learning

SUPPLEMENTARY MATERIAL

1. The Proof for Lemma 1

Lemma 1 *Let $\mathcal{X}, \mathcal{Y}$ be two symmetric and positive-define matrices of the same size, then the function:*

$$f(\mathcal{X}) = tr(\mathcal{X}\mathcal{Y}) - logdet(\mathcal{X})$$

is minimized uniquely by:

$$\mathcal{X}^* = \mathcal{Y}^{-1}$$

PROOF. First of all, by introducing an auxiliary variable $\mathcal{Z} = \mathcal{Y}^{\frac{1}{2}}\mathcal{X}\mathcal{Y}^{\frac{1}{2}}$ (note that $\mathcal{Z}$ is symmetric and positive-define iff $\mathcal{X}$ is), and the conclusion $log\ det(A) = tr(log(A))$ which has been proved in [15], then we have:

$$\begin{aligned} f(\mathcal{X}) &= tr(\mathcal{X}\mathcal{Y}) - log\ det(\mathcal{X}) \\ &= tr(\mathcal{Y}^{-\frac{1}{2}}\mathcal{Z}\mathcal{Y}^{-\frac{1}{2}}\mathcal{Y}) - log\ det(\mathcal{Y}^{-\frac{1}{2}}\mathcal{Z}\mathcal{Y}^{-\frac{1}{2}}) \\ &= tr(\mathcal{Z}\mathcal{Y}^{-\frac{1}{2}}\mathcal{Y}\mathcal{Y}^{-\frac{1}{2}}) - tr(log(\mathcal{Y}^{-\frac{1}{2}}\mathcal{Z}\mathcal{Y}^{-\frac{1}{2}})) \\ &= tr(\mathcal{Z}) - tr(log(\mathcal{Z})) + tr(log(\mathcal{Y})) \\ &= tr(\mathcal{Z}) - log\ det(\mathcal{Z}) + log\ det(\mathcal{Y}) \end{aligned} \tag{1}$$

According to Eq. (1), minimizing $f(\mathcal{X})$ is equivalent to optimizing the following equations $g(\mathcal{Z})$:

$$g(\mathcal{Z}) = tr(\mathcal{Z}) - log\ det(\mathcal{Z}) \tag{2}$$

If $\mathcal{Z}$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then we rewrite Eq. (2) as:

$$g(\mathcal{Z}) = \sum_{i=1}^{n}\lambda_i - log\prod_{i=1}^{n}\lambda_i = \sum_{i=1}^{n}(\lambda_i - log(\lambda_i))$$

Now $h(\lambda) = \lambda - log(\lambda)$ is minimized uniquely at $\lambda = 1$, so $g(\mathcal{Z})$ is minimized uniquely when $\mathcal{Z} = \mathcal{I}$. Finally, combining this equation with $\mathcal{Z} = \mathcal{Y}^{\frac{1}{2}}\mathcal{X}\mathcal{Y}^{\frac{1}{2}}$, we thus have the solution as: $\mathcal{X}^* = \mathcal{Y}^{-1}$

2. Detailed Experimental Results

In this section, we provide further experimental results over three few-shot benchmark datasets, the ablation study and more visualization in detail.

2.1. Accuracy with 95% Confidence Interval

To verify the effectiveness of our approach for few-shot classification, we compare the TEAM framework with our re-implemented baseline (ProtoNet [12]) and many start-of-the-art methods in various setting on three benchmark datasets (*mini*-ImageNet, Cifar-100 and CUB). All results are shown in Table 1-3. Note that each accuracy is averaged over 1000 test tasks which are randomly selected from the testing set and reported with 95% confidence intervals for comparison.

2.2. The performance with various training/testing shots.

In order to verify the nature of transduction [5, 6], where more training data are available, the less performance improvement will be, we further perform 5-way k-shot (k=1, 3, 5, 7, 9) experiments on *mini*-ImageNet and all results are shown in Table. 5. As the number of shots increases, TEAM consistently outperforms our baseline with a large margin, but the performance improvement from TEAM decreases slightly, which further verifies the above analysis about transductive inference.

Table 1. Few-shot classification accuracy on *mini*ImageNet. Tran: The different type of transduction. Top results are highlighted.

Model	Tran.	5-Way 1-Shot		5-Way 5-Shot	
		ConvNet	ResNet	ConvNet	ResNet
MatchNet [13]	No	43.56 ± 0.84	-	55.31 ± 0.73	-
MAML [2]	BN	48.70 ± 1.84	-	63.10 ± 0.92	-
MAML+ [6]	Yes	50.83 ± 1.85	-	66.19 ± 1.85	-
Reptile [8]	BN	49.97 ± 0.32	-	65.99 ± 0.58	-
ProtoNet [12]	No	49.42 ± 0.78	-	68.20 ± 0.66	-
GNN [3]	No	50.33 ± 0.36	-	64.02 ± 0.51	-
RelationNet [16]	BN	50.44 ± 0.82	-	65.32 ± 0.70	-
PFA [10]	No	54.53 ± 0.40	59.60 ± 0.41	67.87 ± 0.20	73.74 ± 0.19
TADAM [9]	No	-	58.50 ± 0.30	-	$\mathbf{76.70 \pm 0.30}$
adaResNet [7]	No	-	56.88	-	71.94
LEO [11]	No	-	60.06 ± 0.08	-	75.72 ± 0.12
TPN [6]	Yes	55.51 ± 0.86	59.46	69.86 ± 0.65	75.65
Baseline (Ours)	No	51.68 ± 0.31	55.25 ± 0.20	68.71 ± 0.20	70.58 ± 0.40
TEAM (Ours)	Yes	$\mathbf{56.57 \pm 0.21}$	$\mathbf{60.07 \pm 0.32}$	$\mathbf{72.04 \pm 0.12}$	75.90 ± 0.20

Table 2. Few-shot classification performance on Cifar-100. Tran: The different type of transduction. Top results are highlighted.

Model	Tran.	5-Way 1-Shot		5-Way 5-Shot	
		ConvNet	ResNet	ConvNet	ResNet
MatchNet [13]	No	50.53 ± 0.87	-	60.30 ± 0.82	-
MAML [2]	BN	49.28 ± 0.90	-	58.30 ± 0.80	-
ProtoNet [12]	No	56.66 ± 0.53	-	76.29 ± 0.14	-
DEML [4]	No	-	61.62 ± 1.01	-	77.94 ± 0.74
Baseline (Ours)	No	57.83 ± 0.27	66.30 ± 0.40	76.40 ± 0.33	80.46 ± 0.27
TEAM (Ours)	Yes	$\mathbf{64.07 \pm 0.30}$	$\mathbf{70.43 \pm 0.24}$	$\mathbf{79.05 \pm 0.38}$	$\mathbf{81.25 \pm 0.22}$

Table 3. Few-shot classification performance on CUB. Tran: The different type of transduction. Top results are highlighted.

Model	Tran.	5-Way 1-Shot		5-Way 5-Shot	
		ConvNet	ResNet	ConvNet	ResNet
MatchNet [13]	No	56.53 ± 0.99	-	63.54 ± 0.85	-
MAML [2]	BN	50.45 ± 0.97	-	59.60 ± 0.84	-
ProtoNet [12]	No	58.43 ± 0.30	-	75.22 ± 0.36	-
RelationNet [16]	BN	62.45 ± 0.98	-	76.11 ± 0.69	-
DEML [4]	No	-	66.95 ± 1.06	-	77.11 ± 0.78
TriNet [1]	No	-	69.61 ± 0.46	-	84.10 ± 0.35
Baseline (Ours)	No	69.39 ± 0.20	74.55 ± 0.45	82.78 ± 0.24	85.98 ± 0.17
TEAM (Ours)	Yes	$\mathbf{75.71 \pm 0.18}$	$\mathbf{80.16 \pm 0.48}$	$\mathbf{86.04 \pm 0.14}$	$\mathbf{87.17 \pm 0.45}$

Table 4. Few-shot classification performance for ablation study. Proto (Ours): the baseline. TEAM‡: baseline+*TIM*. TEAM†: baseline+*TIM*+*EAM*. TEAM: baseline+*TIM*+*EAM*+*Bi-SIM*.

Model	*mini*ImageNet		Cifar-100		CUB	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Proto [13]	49.42 ± 0.78	68.20 ± 0.66	56.66 ± 0.53	76.29 ± 0.14	58.43 ± 0.30	75.22 ± 0.36
Proto (Ours)	51.68 ± 0.31	68.71 ± 0.20	57.83 ± 0.27	76.40 ± 0.33	69.39 ± 0.20	82.78 ± 0.24
TEAM‡	52.97 ± 0.21	70.45 ± 0.14	59.56 ± 0.42	77.65 ± 0.43	70.27 ± 0.24	84.68 ± 0.05
TEAM†	55.35 ± 0.25	71.59 ± 0.12	62.76 ± 0.41	78.80 ± 0.40	75.06 ± 0.25	$\mathbf{86.06 \pm 0.09}$
TEAM	$\mathbf{56.57 \pm 0.21}$	$\mathbf{72.04 \pm 0.12}$	$\mathbf{64.07 \pm 0.30}$	$\mathbf{79.05 \pm 0.38}$	$\mathbf{75.71 \pm 0.18}$	86.04 ± 0.14

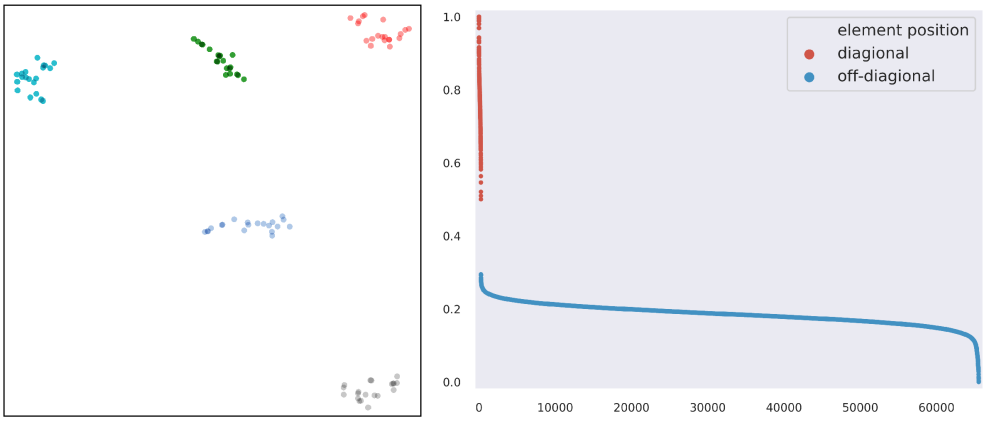
2.3. Sparsity Nature of Episodic-wise Adaptive Metric.

For the sake of illustration, we firstly exploit the classic LMNN algorithm [14] with all support and query samples to optimize an oracle metric, which ensures all examples in this task can be completely distinguished, see Fig. 2 (left). Then we scale all elements of the metric into region [0, 1] and reorganize all values with numerical descending order in Fig. 2 (right).

Table 5. 5-way performance with various training/testing shots.

Methods	1-shot	3-shot	5-shot	7-shot	9-shot
Baseline (Ours)	51.68 ± 0.31	63.87 ± 0.26	68.71 ± 0.20	71.28 ± 0.15	73.35 ± 0.44
TEAM (Ours)	56.57 ± 0.21	67.64 ± 0.21	72.04 ± 0.12	73.47 ± 0.16	75.04 ± 0.19
Accuracy (+)	4.89	3.77	3.33	2.19	1.69

Obviously, there is a large value gap between the diagonal elements and the off-diagonal elements. Moreover, we further visualize the heatmap of the oracle metric in Fig. 1 in great detail.



(a) Embeddings with Using Oracle Metric

(b) All Values in the Oracle Metric

Figure 1. This figure illustrates the sparsity nature of the metric in few-shot learning. Left: The image embedding visualization using the oracle metric learned by LMNN. Right: The values distribution in different position of the matrix (sorted by descending order).

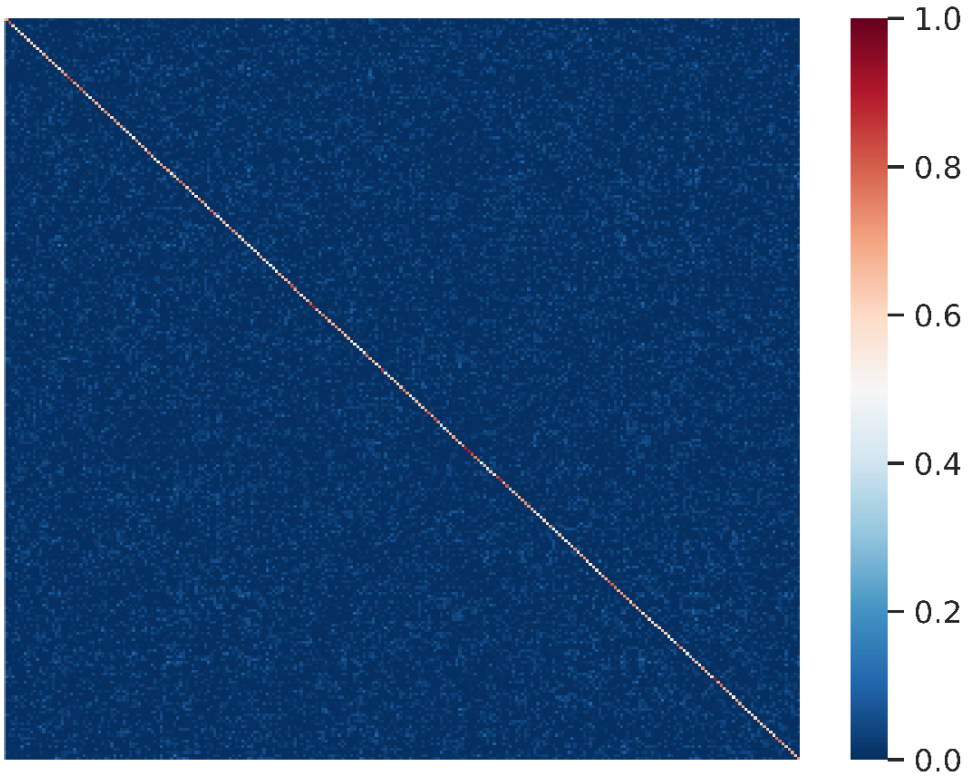


Figure 2. The heatmap of the oracle metric.

References

[1] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Semantic feature augmentation in few-shot learning. *arXiv:1804.05298*, 2018.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[3] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *ICLR*, 2017.

[4] Bharath Hariharan and Ross B Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, pages 3037–3046, 2017.

[5] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.

[6] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *ICLR*, 2018.

[7] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *ICML*, pages 3661–3670, 2018.

[8] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR, abs/1803.02999*, 2, 2018.

[9] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NIPS*, pages 719–729, 2018.

[10] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, volume 2, 2017.

[11] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019.

[12] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087, 2017.

[13] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016.

[14] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2006.

[15] Christopher S Withers and Saralees Nadarajah. log det a= tr log a. *International Journal of Mathematical Education in Science and Technology*, 41(8):1121–1124, 2010.

[16] Flood Sung Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.