MOKD: Cross-domain Finetuning for Few-shot Classification via Maximizing Optimized Kernel Dependence

Hongduan Tian, Feng Liu, Tongliang Liu, Bo Du, Yiu-ming Cheung, Bo Han

TMLR, Hong Kong Baptist University, The University of Melbourne, The University of Sydney, Wuhan University

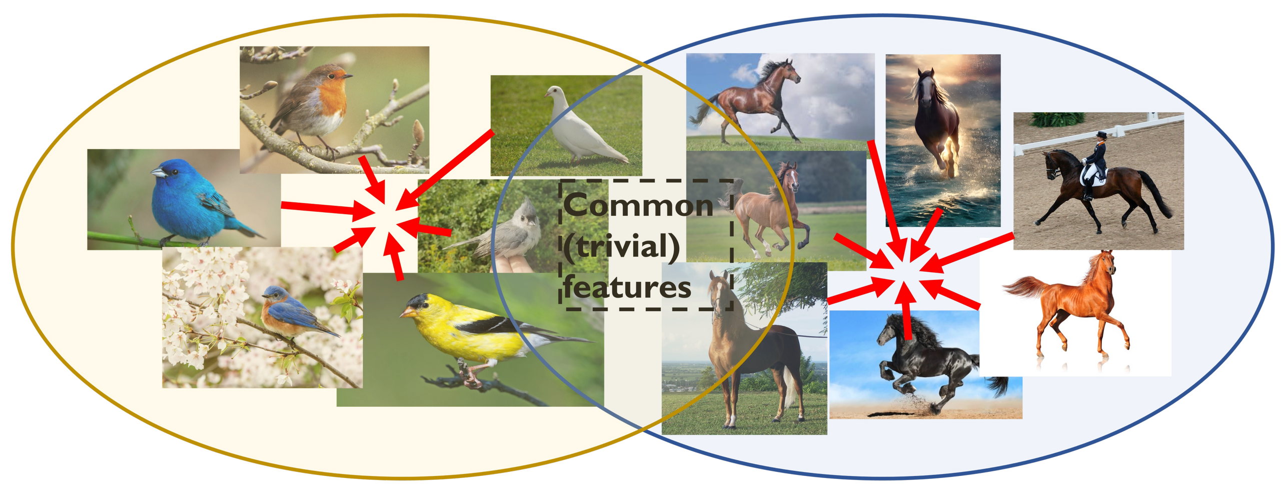
Paper Code Slides

Problem: Representation Learning in FSL

Two intuitions of NCC-based loss[1, 2] (a.k.a., prototypical loss):

- **Pull** samples belonging to the class closer to the class centroid;
- **Push** samples from other classes away from the class centroid

$$\mathcal{L}_{\mathrm{NCC}} = -\frac{1}{|\mathcal{D}_{\mathcal{T}}|}\sum_{z\in\mathcal{D}_{\mathcal{T}}}\log\frac{\exp\big(k(z,c_c)\big)}{\sum_{i=1}^{N_C}\exp\big(k(z,c_i)\big)}$$

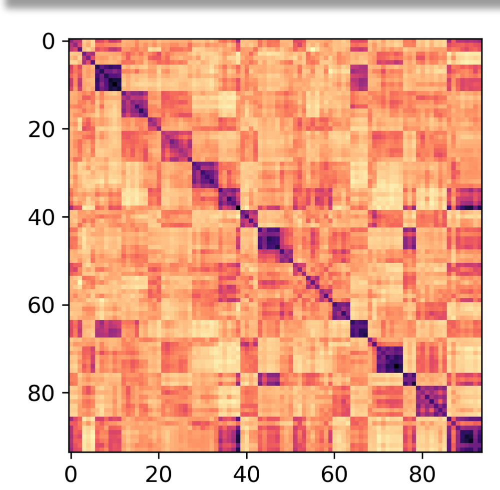


Representation learning perspective:

- **Maximize** similarities among samples within the same class;
- **Minimize** similarities between samples from different classes

1. Snell et al., Prototypical networks for few-shot learning, NIPS 2017.
2. Li et al., Universal representation learning from multiple domains for few-shot classification, ICCV 2021.

Issue: High similarity across classes



Two phenomena:

- The similarities among samples within the same class are high
- High similarities between samples from different classes

A lower bound of NCC-based loss:

$$\mathcal{L}(\theta) \geq -\frac{1}{|\mathcal{D}_{\mathcal{T}}|}\sum_{i=1}^{|\mathcal{D}_{\mathcal{T}}|}\frac{1}{|\mathbb{C}|}\sum_{z^{+}\in\mathbb{C}}k(z_i,z^{+}) + \frac{1}{|\mathcal{D}_{\mathcal{T}}|}\sum_{i=1}^{|\mathcal{D}_{\mathcal{T}}|}\sum_{z'\in\mathbb{Z}}\frac{k(z_i,z')}{|\mathcal{D}_{\mathcal{T}}|} + \mathcal{O}\big(k(z,z')\big) + \mathrm{const}$$

An interpretation of NCC-based loss from the kernel: With a linear kernel (e.g., cosine similarity), under few-shot settings, we know that NCC-based loss tends to:

- **Maximize** similarities among samples within the **same class**;
- **Minimize** similarities between samples from **different classes.**

More powerful kernels are required to learn desirable representations.

Method: maximizing optimized kernel dependence

Connection between NCC-based loss & HSIC[3]:

$$\mathrm{HSIC}(Z,Y) \geq \frac{\lambda\Delta l}{|\mathcal{D}_{\mathcal{T}}|^2}\sum_{i=1}^{|\mathcal{D}_{\mathcal{T}}|}\sum_{z^{+}\in\mathbb{C}}k(z_i,z^{+}) - \frac{\lambda\Delta l}{|\mathcal{D}_{\mathcal{T}}|^2}\sum_{i=1}^{|\mathcal{D}_{\mathcal{T}}|}\sum_{z'\in\mathbb{Z}}\frac{1}{|\mathcal{D}_{\mathcal{T}}|}k(z_i,z')$$

$$\Rightarrow \mathcal{L}_{\mathrm{NCC}} \geq -HSIC(Z,Y) + \gamma HSIC(Z,Z)$$

Core ideas:

- Introduce more powerful kernel. linear → Gaussian/IMQ
- Test power maximization. sensitivity to dependence of the kernel UP!

Test power is used to measure the probability that, for particular two dependent distributions and the number of samples m, the null hypothesis that the two distributions are independent is correctly rejected.

With unbiased HSIC estimator $\widehat{\mathrm{HSIC}}_{\mathrm{u}}$, under the hypothesis that the two distributions are independent, the CLT implies that test power can be formulated as:

$$\Pr\big(m\widehat{\mathrm{HSIC}}_{\mathrm{u}} > r\big) \to \Phi\left(\frac{\sqrt{m}\mathrm{HSIC}}{v} - \frac{r}{\sqrt{m}v}\right)$$

Test power maximization: select an optimal parameter (e.g., bandwidth of Gaussian kernel) to maximize HSIC/v, where v can be estimated[4].

3. Gretton et al., Measuring statistical dependence with Hilbert-Schmidt norms, ALT 2005.
4. Song et al., Feature selection via dependence maximization, JMLR 2012.

Implementation: Bi-level optimization framework

Main objective:

$$\min_{\theta} -\mathrm{HSIC}(Z,Y;\,\sigma_{ZY}^{*},\theta) + \gamma\mathrm{HSIC}(Z,Z;\,\sigma_{ZZ}^{*},\theta),$$

$$s.t.\ \max_{\sigma_{ZY}}\frac{\mathrm{HSIC}(Z,Y;\,\sigma_{ZY},\theta)}{\sqrt{v_{ZY}+\epsilon}},\ \max_{\sigma_{ZZ}}\frac{\mathrm{HSIC}(Z,Z;\,\sigma_{ZZ},\theta)}{\sqrt{v_{ZZ}+\epsilon}}$$

Details of MOKD:

Given a list of candidate bandwidth: $\Sigma = [\sigma_1, \sigma_2, \ldots, \sigma_T]$.

1. Randomly sample a new task and get the representation-label pairs: Z, Y;
2. Perform test power maximization to select the optimal bandwidth:

$$\sigma_{ZY}^{*} = \max_{\Sigma}\frac{\widehat{\mathrm{HSIC}}(Z,Y;\,\sigma_{ZY},\theta)}{\sqrt{v_{ZY}+\epsilon}},\ \sigma_{ZZ}^{*} = \max_{\Sigma}\frac{\widehat{\mathrm{HSIC}}(Z,Z;\,\sigma_{ZZ},\theta)}{\sqrt{v_{ZZ}+\epsilon}}$$

3. Iteratively minimizing the objective and update the parameters:

$$\theta \leftarrow \theta - \eta\nabla_{\theta}[-\mathrm{HSIC}(Z,Y;\,\sigma_{ZY}^{*},\theta) + \gamma\mathrm{HSIC}(Z,Z;\,\sigma_{ZZ}^{*},\theta)]$$

Experiments

Table 1. **Results on Meta-Dataset (Trained on ImageNet Only).** Mean accuracy and 95% confidence interval are reported.

Datasets	Finetune	ProtoNets	ProtoNets(large)	BOHB	FP-MAML	ALFA+FP-MAML	FLUTE	SSL-HSIC	URL	**MOKD(Ours)**
ImageNet	45.8±1.1	50.5±1.1	53.7±1.1	51.9±1.1	49.5±1.1	52.8±1.1	46.9±1.1	55.5±1.1	**57.3±1.1**	**57.3±1.1**
Omniglot	60.9±1.6	60.0±1.4	68.5±1.3	67.6±1.2	63.4±1.3	61.9±1.5	61.6±1.4	66.4±1.2	69.4±1.2	**70.9±1.3**
Aircraft	**68.7±1.3**	53.1±1.0	58.0±1.0	54.1±0.9	56.0±1.0	63.4±1.1	48.5±1.0	49.5±0.9	57.6±1.0	**59.8±1.0**
Birds	57.3±1.3	68.8±1.0	**74.1±0.9**	70.7±0.9	68.7±1.0	69.8±1.1	47.9±1.0	71.6±0.9	72.9±0.9	**73.6±0.9**
Textures	69.0±0.9	66.6±0.8	68.8±0.8	68.3±0.8	66.5±0.8	70.8±0.9	63.8±0.8	72.2±0.7	75.2±0.7	**76.1±0.7**
Quick Draw	42.6±1.2	49.0±1.1	53.3±1.0	50.3±1.0	51.5±1.0	59.2±1.2	57.5±1.0	54.2±1.0	57.9±1.0	**61.2±1.0**
Fungi	38.2±1.0	39.7±1.1	40.7±1.2	41.4±1.1	40.0±1.1	41.5±1.2	31.8±1.0	43.4±1.1	46.2±1.0	**47.0±1.1**
VGG Flower	85.5±0.7	85.3±0.8	87.0±0.7	87.3±0.6	87.2±0.7	86.0±0.8	80.1±0.9	85.5±0.7	86.9±0.6	**88.5±0.6**
Traffic Sign	**66.8±1.3**	47.1±1.1	58.1±1.1	51.8±1.0	48.8±1.1	60.8±1.3	46.5±1.1	50.5±1.1	61.2±1.2	**61.6±1.1**
MSCOCO	34.9±1.0	41.0±1.1	41.7±1.1	48.0±1.0	43.7±1.1	48.1±1.1	41.4±1.0	51.4±1.0	53.0±1.0	**55.3±1.0**
MNIST	-	-	-	-	-	-	80.8±0.8	77.0±0.7	86.2±0.7	**88.3±0.7**
CIFAR-10	-	-	-	-	-	-	65.4±0.8	71.0±0.8	69.5±0.8	**72.2±0.8**
CIFAR-100	-	-	-	-	-	-	52.7±1.1	59.0±1.0	62.0±1.0	**63.1±1.0**
Average Seen	45.8	50.5	53.7	51.9	49.5	52.8	46.9	55.5	**57.3**	**57.3**
Average Unseen	-	-	-	-	-	-	56.5	62.5	66.6	**68.1**
Average All	-	-	-	-	-	-	55.8	62.0	65.9	**67.3**
Average Rank	7.1	8.4	4.6	5.5	6.8	4.4	8.9	4.9	2.8	**1.4**

[1] The results on URL and MOKD are the average of 5 reproductions with different random seeds.

Table 2. **Results on Meta-Dataset (Trained on All Datasets).** Mean accuracy and 95% confidence interval are reported.

Datasets	ProtoMAML	CNAPS	S-CNAPS	SUR	URT	Tri-M	FLUTE	2LM	SSL-HSIC	URL	**MOKD**
ImageNet	46.5± 1.1	50.8±1.1	58.4 ±1.1	56.2 ± 1.0	56.8 ± 1.1	**58.6 ± 1.0**	51.8 ± 1.1	58.0 ± 3.6	56.5 ± 1.2	57.3 ± 1.1	57.3 ± 1.1
Omniglot	82.7± 1.0	91.7±0.5	91.6 ± 0.6	94.1 ± 0.4	94.2 ± 0.4	92.0 ± 0.6	93.2 ± 0.5	**95.3 ± 1.0**	92.0 ± 0.9	94.1 ± 0.4	**94.2 ± 0.5**
Aircraft	75.2± 0.8	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	87.3 ± 0.7	88.2 ± 0.5	**88.4 ± 0.5**
Birds	69.9± 1.0	73.6±0.9	74.8 ± 0.9	71.0 ± 1.0	76.2 ± 0.8	75.3 ± 0.8	79.2 ± 0.8	**81.8 ± 0.6**	78.1 ± 1.1	80.2 ± 0.7	**80.4 ± 0.8**
Textures	68.2± 1.0	59.5±0.7	68.8 ± 0.9	71.0 ± 0.8	71.6 ± 0.7	71.2 ± 0.8	68.8 ± 0.8	76.3 ± 2.4	75.2 ± 0.8	76.2 ± 0.7	**76.5 ± 0.7**
Quick Draw	66.8± 0.9	74.7±0.8	76.5 ±0.8	81.8 ± 0.6	**82.4 ± 0.6**	77.3 ± 0.7	79.5 ± 0.7	78.3 ± 0.7	81.4 ± 0.7	82.2 ± 0.6	82.2 ± 0.6
Fungi	42.0±1.2	50.2±1.1	46.6 ± 1.0	64.3 ± 0.9	64.0 ± 1.0	48.5 ± 1.0	58.1 ± 1.1	**69.6 ± 1.5**	63.5 ± 1.2	68.7 ± 1.0	68.6 ± 1.0
VGG Flower	88.7± 0.7	88.9±0.5	90.5 ± 0.5	82.9 ± 0.8	87.9 ± 0.6	90.5 ± 0.5	91.6 ± 0.6	90.3 ± 0.8	90.9 ± 0.8	91.9 ± 0.5	**92.5 ± 0.5**
Traffic Sign	52.4 ± 1.1	56.5 ±1.1	57.2 ± 1.0	51.0 ± 1.1	48.2 ± 1.1	63.0 ± 1.0	58.4 ± 1.1	63.6 ± 1.5	59.7 ± 1.3	63.3 ± 1.2	**64.5 ± 1.1**
MSCOCO	41.7 ± 1.1	39.4 ±1.0	48.9 ± 1.1	52.0 ± 1.1	51.5 ± 1.1	52.8 ± 1.1	50.0 ± 1.0	**57.0 ± 1.1**	51.4 ± 1.1	54.2 ± 1.0	**55.5 ± 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 ± 0.5	93.4 ± 0.6	94.7 ± 0.4	**95.1 ± 0.4**
CIFAR-10	-	-	74.9 ± 0.7	66.5 ± 0.9	67.0 ± 0.8	75.4 ± 0.8	**78.6 ± 0.7**	71.5 ± 0.9	70.0 ± 1.1	71.9 ± 0.8	**72.8 ± 0.8**
CIFAR-100	-	-	61.3 ± 1.1	56.9 ± 1.1	57.3 ± 1.0	62.0 ± 1.0	**67.1 ± 1.0**	60.0 ± 1.1	61.8 ± 1.1	62.9 ± 1.0	**63.9 ±1.0**
Average Seen	67.5	71.6	73.7	75.9	77.4	76.2	76.2	79.7	76.5	79.9	**80.0**
Average Unseen	-	-	67.4	64.1	62.9	69.9	69.9	69.4	68.2	69.4	**70.3**
Average All	-	-	71.2	71.3	71.8	73.8	73.8	75.7	74.6	75.8	**76.3**
Average Rank	-	-	7.2	7.3	6.4	5.2	5.2	3.4	5.5	3.1	**2.2**

[1] Results of URL are the average of 5 reproductions with different random seeds. The reproductions are consistent with the results reported on their website. The results of our method are the average of 5 random reproduction experiments. The ranks considers all 13 datasets and are calculated only with the methods in the table.