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

Hongduan Tian[1, 2], Feng Liu[3], Tongliang Liu[4], Bo Du[5], Yiu-ming Cheung[2], 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]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University

Outline

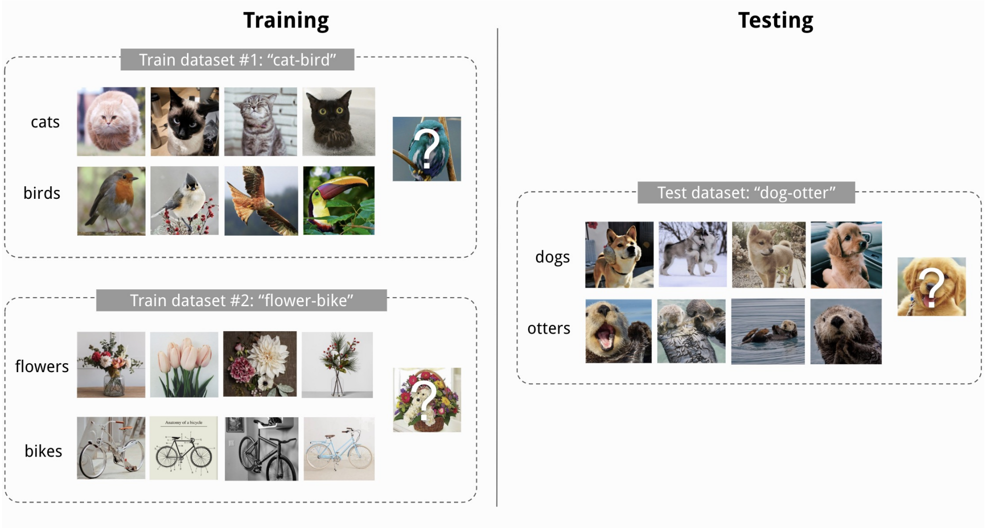
- Background
- Understanding of NCC-based loss via HSIC
- Maximizing Optimized Kernel Dependence (MOKD)
- Summary

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

Hongduan Tian [1 2] **Feng Liu** [3] **Tongliang Liu** [4] **Bo Du** [5] **Yiu-ming Cheung** [2] **Bo Han** [1 2]

Preliminary: Cross-domain Few-shot Classification

Few-shot classification with prototypes



An example of conventional few-shot classification tasks

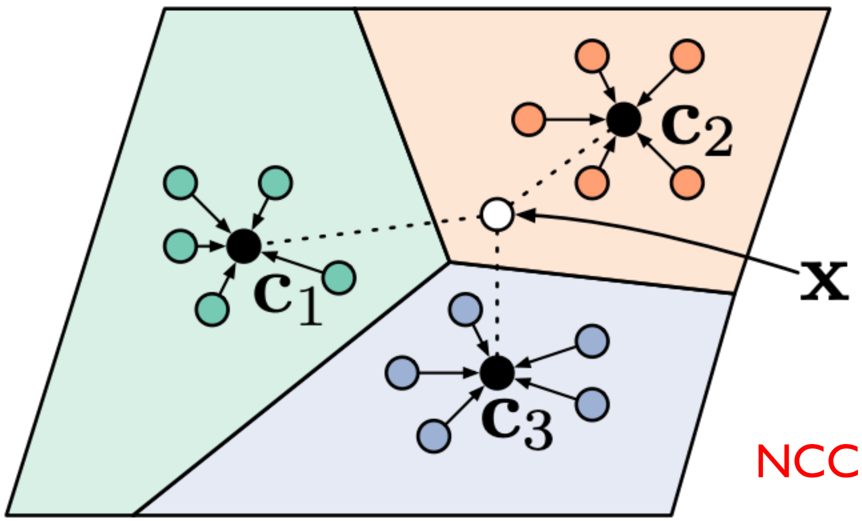
Challenges in CFC:

- Numbers of ways & shots vary among tasks;
- Discrepancies between source and target domains

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:

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

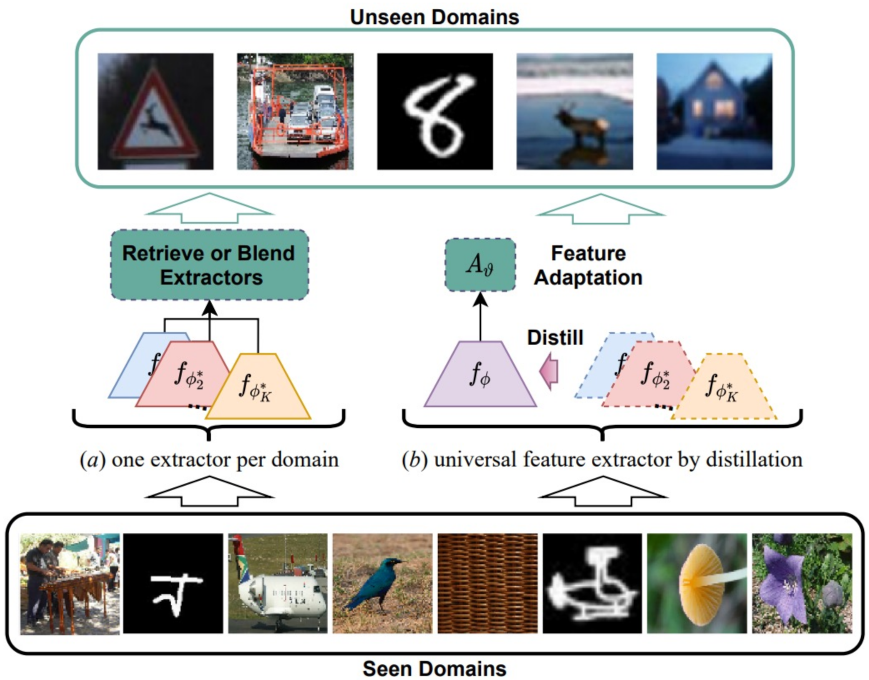
NCC-based loss

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

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

Previous Works

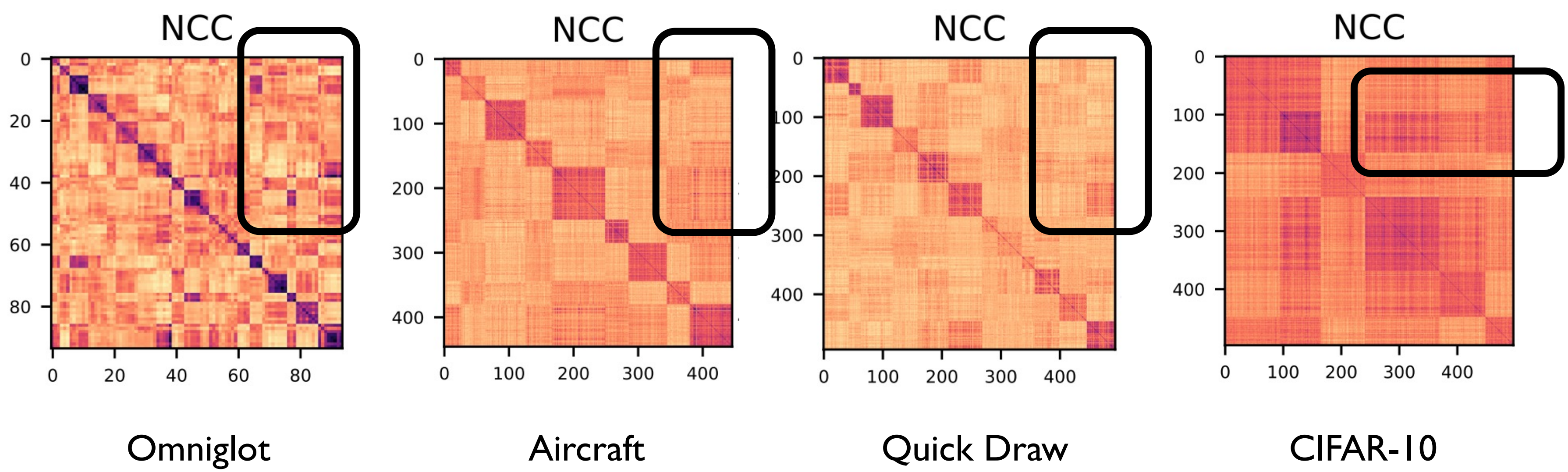
Finetuning a transformation on top of a universal pretrained backbone



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

Remaining Issue

High similarities between samples from different classes



High similarities between samples from different classes may induce uncertainty and result in misclassification.

Theoretical Understanding of NCC-loss from HSIC

Insights behind NCC-based loss

Theorem 3.2 (Lower bound of NCC-based loss). *Given a set of normalized support representations $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{|\mathcal{D}_\mathcal{T}|} = \{h_\theta \circ f_{\phi^*}(\boldsymbol{x}_i)\}_{i=1}^{|\mathcal{D}_\mathcal{T}|}$ and the corresponding labels $\{y_i\}_{i=1}^{|\mathcal{D}_\mathcal{T}|}$ that includes N_C classes from a support set $\mathcal{D}_\mathcal{T}$. Let $k(\cdot,\cdot)$ be the cosine similarity function. Then, with Assumption 3.1, the NCC-based loss (Eq. (1)) owns a lower bound:*

$$
\begin{aligned}
\mathcal{L}(\theta) \geq & -\frac{1}{|\mathcal{D}_\mathcal{T}|}\sum_{i=1}^{|\mathcal{D}_\mathcal{T}|}\frac{1}{|\mathcal{C}|}\sum_{\boldsymbol{z}^+\in\mathcal{C}} k(\boldsymbol{z}_i,\boldsymbol{z}^+) \\
& + \frac{1}{|\mathcal{D}_\mathcal{T}|}\sum_{i=1}^{|\mathcal{D}_\mathcal{T}|}\sum_{\boldsymbol{z}'\in\mathcal{Z}}\frac{k(\boldsymbol{z}_i,\boldsymbol{z}')}{|\mathcal{D}_\mathcal{T}|} + \mathcal{O}\left(k(\boldsymbol{z},\boldsymbol{z}')\right) + const,
\end{aligned}
$$

where $\boldsymbol{z}^+$ denotes the data samples belonging to the same class as $\boldsymbol{z}_i$, $\mathcal{C}$ denotes the class that $\boldsymbol{z}_i$ belongs to, $\mathcal{O}\left(k(\boldsymbol{z},\boldsymbol{z}')\right)$ denotes a high-order moment term. In addition, $const = \log \alpha_e N_C$, where N_C denotes the number of classes in task, α_e is a constant.

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

Theoretical Understanding of NCC-loss from HSIC

Hilbert-Schmidt Independence Criterion

$$\mathrm{HSIC}(X,Y) = ||\mathbb{E}[\varphi(X)\psi(Y)^\top] - \mathbb{E}[\varphi(X)]\mathbb{E}[\psi(Y)]^\top||^2_{HS}$$

Theorem 3.2 (Lower bound of NCC-based loss). *Given a set of normalized support representations $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{|\mathcal{D}_\mathcal{T}|} = \{h_\theta \circ f_{\phi^*}(\boldsymbol{x}_i)\}_{i=1}^{|\mathcal{D}_\mathcal{T}|}$ and the corresponding labels $\{y_i\}_{i=1}^{|\mathcal{D}_\mathcal{T}|}$ that includes N_C classes from a support set $\mathcal{D}_\mathcal{T}$. Let $k(\cdot,\cdot)$ be the cosine similarity function. Then, with Assumption 3.1, the NCC-based loss (Eq. (1)) owns a lower bound:*

$$\mathcal{L}(\theta) \geq - \frac{1}{|\mathcal{D}_\mathcal{T}|}\sum_{i=1}^{|\mathcal{D}_\mathcal{T}|}\frac{1}{|\mathcal{C}|}\sum_{\boldsymbol{z}^+\in\mathcal{C}} k(\boldsymbol{z}_i,\boldsymbol{z}^+) + \frac{1}{|\mathcal{D}_\mathcal{T}|}\sum_{i=1}^{|\mathcal{D}_\mathcal{T}|}\sum_{\boldsymbol{z}'\in\mathcal{Z}}\frac{k(\boldsymbol{z}_i,\boldsymbol{z}')}{|\mathcal{D}_\mathcal{T}|} + \mathcal{O}\left(k(\boldsymbol{z},\boldsymbol{z}')\right) + const,$$

where $\boldsymbol{z}^+$ denotes the data samples belonging to the same class as $\boldsymbol{z}_i$, $\mathcal{C}$ denotes the class that $\boldsymbol{z}_i$ belongs to, $\mathcal{O}\left(k(\boldsymbol{z},\boldsymbol{z}')\right)$ denotes a high-order moment term. In addition, $const = \log \alpha_e N_C$, where N_C denotes the number of classes in task, α_e is a constant.

Theorem 3.4. *Given a support representation set $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{|\mathcal{D}_\mathcal{T}|} = \{h_\theta \circ f_{\phi^*}(\boldsymbol{x}_i)\}_{i=1}^{|\mathcal{D}_\mathcal{T}|}$ where N_C classes are included, let $k(\cdot,\cdot)$ be a linear kernel function on data representations and $l(\cdot,\cdot)$ be a label kernel defined in Eq. (4), then $\mathrm{HSIC}(Z,Y)$ owns a lower bound:*

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

where $\boldsymbol{z}^+$ denotes the data samples belonging to the same class as $\boldsymbol{z}_i$, $\mathcal{C}$ denotes the class that $\boldsymbol{z}_i$ belongs to, $\boldsymbol{z}'$ is an independent copy of $\boldsymbol{z}$, λ is a scale constant.

Theoretical Understanding of NCC-loss from HSIC

Hilbert-Schmidt Independence Criterion

Since the constant scaling does not affect the optimization and it is easy to obtain that the high-order moment term satisfies $\mathcal{O}(k(\boldsymbol{z},\boldsymbol{z}')) \geq \gamma\mathrm{HSIC}(Z,Z)$, where $\gamma = \frac{|\mathcal{D}_\mathcal{T}|}{2N_C C_{\max}}$, $C_{\max}$ is a constant that satisfies $C_{\max} \geq |\mathcal{C}_c|$ for $\forall c \in \{1,2,...,N_C\}$ (see Appendix B.4 for more details), we then can build a connection between NCC-based loss and HSIC measure via omitting the scale constants as following:

$$\boxed{\mathcal{L}(\theta) \geq -\mathrm{HSIC}(Z,Y) + \gamma\mathrm{HSIC}(Z,Z) + const.}$$

Understandings:

- Under CFC settings, both HSIC and NCC-based loss play the same role in representation learning;
- NCC-based loss is a special case of HSIC when the kernel is specialized as a linear kernel.

Why high similarities? Kernel.

Test power maximization

Two drawbacks of the linear kernel:

- An undesirable case that HSIC value is zero yet the two variables are dependent may happen [1].
- We cannot further optimize a linear kernel to increase its capability in dependence measure.

Test power of HSIC. In this paper, 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. Consider a $\widehat{\mathrm{HSIC}}_{\mathrm{u}}$ as an unbiased HSIC estimator (e.g., U-statistic estimator), under the hypothesis that the two distributions are dependent, the central limit theorem (Serfling, 2009) holds:

$$\sqrt{m}(\widehat{\mathrm{HSIC}}_{\mathrm{u}} - \mathrm{HSIC}) \xrightarrow{d} \mathcal{N}(0, v^2),$$

where v^2 denotes the variance, $\xrightarrow{d}$ denotes convergence in distribution. The CLT implies that test power can be formulated as:

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

where r denotes a rejection threshold and Φ denotes the standard normal CDF. Since the rejection threshold r will converge to a constant, and HSIC, v are constants, for reasonably large m, the test power is dominated by the first term. Thus, a feasible way to maximize the test power is to find a kernel function to maximize HSIC/v. The intuition of test power maximization is increasing the sensitivity of the estimated kernel to the dependence among data samples.

Gretton et al., A kernel statistical test of independence, NIPS 2017.

MOKD

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

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

Algorithm 1 Maximizing Optimized Kernel Dependence Algorithm

Input: pre-trained backbone f_{ϕ^*}, number of inner iterations n, learning rate η, linear transformation parameters h_θ, a list of bandwidths $\Sigma = \{\sigma_1, \sigma_2, ..., \sigma_T\}$, and $\epsilon = 1e-5$.
Output: the optimal parameters for linear transformation head θ^*.
 # Sample a task
Sample a new task $\mathcal{T} = \{\{\boldsymbol{X}^{\mathrm{s}}, Y^{\mathrm{s}}\}, \{\boldsymbol{X}^{\mathrm{q}}, Y^{\mathrm{q}}\}\}$;
Obtain the representations: $\mathcal{Z} = \{h_\theta \circ f_{\phi^*}(\boldsymbol{x}_i)\}_{i=1}^{|\boldsymbol{X}^{\mathrm{s}}|}$;
 # Inner optimization for test power maximization
Maximize the test power of $\widehat{\mathrm{HSIC}}(Z, Y; \sigma_{ZY}, \theta)$ and $\widehat{\mathrm{HSIC}}(Z, Z; \sigma_{ZZ}, \theta)$ with Eq. (6) and (7):

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

 # Outer optimization for dependence optimization
for $i = 1$ **to** n **do**
  Obtain the representations: $\mathcal{Z} = \{h_\theta \circ f_{\phi^*}(\boldsymbol{x}_i)\}_{i=1}^{|\boldsymbol{X}^{\mathrm{s}}|}$
  Compute $\widehat{\mathrm{HSIC}}(Z, Y, \sigma^*_{ZY}, \theta)$ and $\widehat{\mathrm{HSIC}}(Z, Z; \sigma^*_{ZZ}, \theta)$ with Eq. (6) for loss:

$$\mathcal{L}(Z, Y; \theta) = -\widehat{\mathrm{HSIC}}(Z, Y, \sigma^*_{ZY}, \theta) + \gamma \widehat{\mathrm{HSIC}}(Z, Z; \sigma^*_{ZZ}, \theta)$$

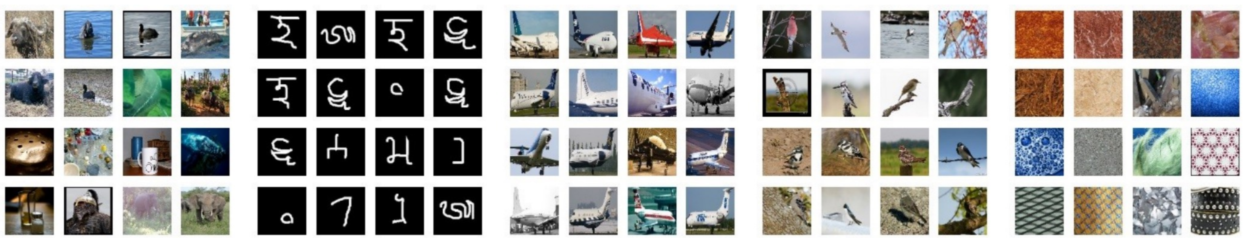
  Update parameters:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(Z, Y; \theta)$$

end for

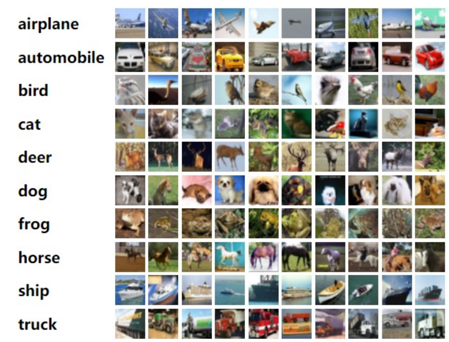
Experiments

Meta-Dataset



(a) ImageNet (b) Omniglot (c) Aircraft (d) Birds (e) DTD

(f) Quick Draw (g) Fungi (h) VGG Flower (i) Traffic Signs (j) MSCOCO



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 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.

Experiments

Main results: train on all datasets

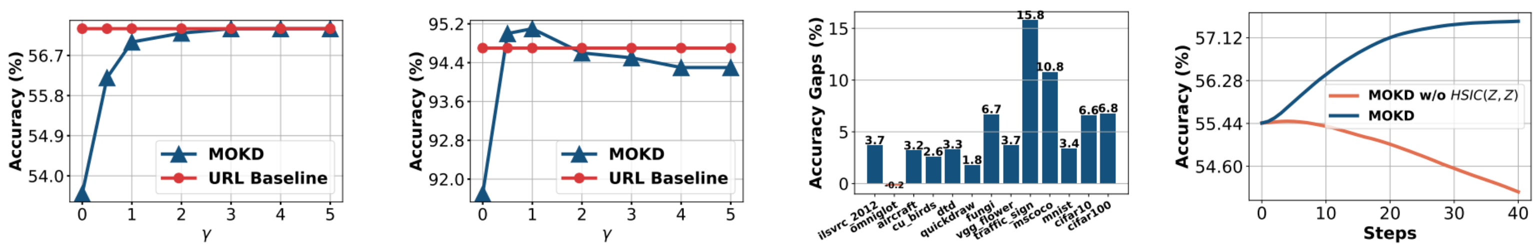
Table 2. **Results on Meta-Dataset (Trained on All Datasets).** Mean accuracy and 95% confidence interval are reported.

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.

Experiments

Analyses



(a) γ on ImageNet (b) γ on MNIST (c) Accuracy gaps (d) Accuracy curves

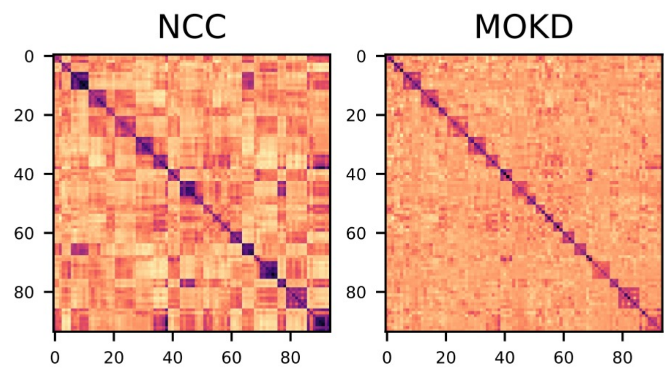
Figure 3. **Quantitative analysis of γ. (a).** Effect of γ on accuracy of ImageNet dataset; **(b).** Effect of γ on accuracy of MNIST dataset; **(c).** Performance gaps between MOKD w / w.o. $\mathrm{HSIC}(Z, Z)$; **(d).** Test accuracy curves of MOKD w. / w.o. $\mathrm{HSIC}(Z, Z)$ on ImageNet.

Table 3. **Comparisons of MOKD with different characteristic kernels.**

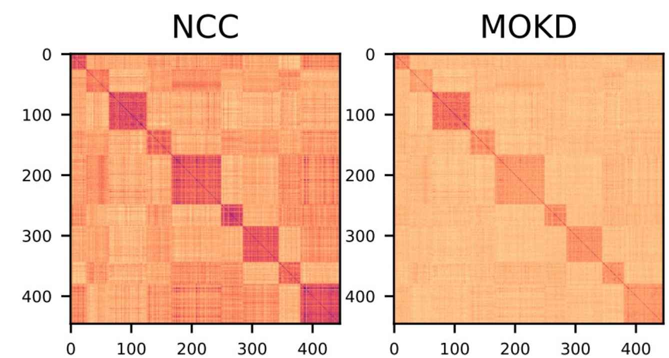
Datasets	ImageNet	Omniglot	Aircraft	Birds	DTD	QuickDraw	Fungi	VGG_Flower	Traffic Sign	MSCOCO	MNIST	CIFAR10	CIFAR100
Gaussian	57.3±1.1	94.2±0.5	**88.4±0.5**	80.4±0.8	**76.5±0.7**	82.2±0.6	**68.6±1.0**	**92.5±0.5**	**64.5±1.1**	**55.5±1.0**	95.1±0.4	72.8±0.8	**63.9±1.0**
IMQ	57.3±1.1	94.3±0.5	88.0±0.5	**80.5±0.8**	76.2±0.7	82.3±0.6	67.7±1.0	92.1±0.5	63.8±1.1	54.8±1.0	**95.4±0.4**	72.7±0.8	63.7±1.0

Experiments

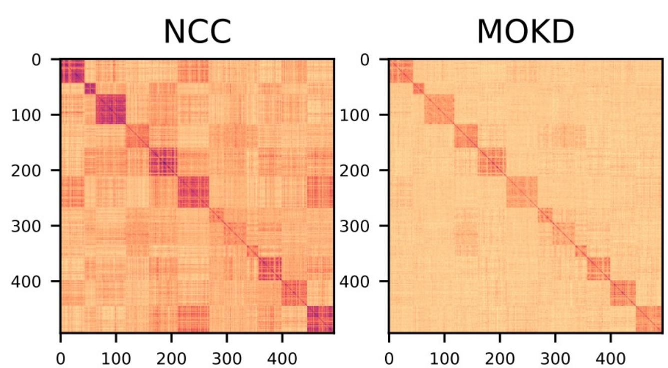
Visualization results



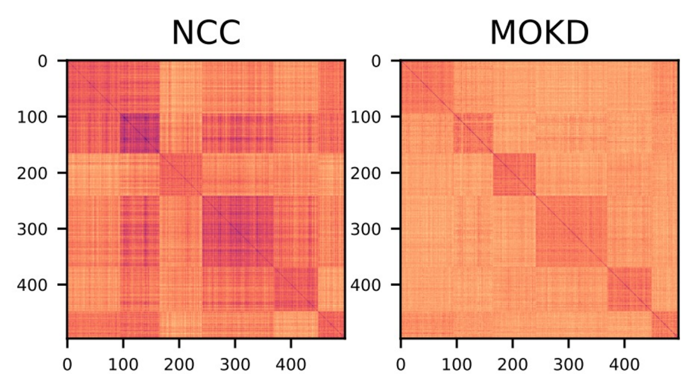
(a) Omniglot (21 classes)



(b) Aircraft (9 classes)



(e) Quick Draw (12 classes)



(d) CIFAR 10 (6 classes)

Summary

- **Empirically**, we find that there exist high similarities between NCC-learned representations of data from different classes, which may further induce uncertainty and result in misclassification of data.
- **Theoretically**, we build a connection between NCC-based loss and kernel HSIC measure and demonstrate that both of them maximize the similarities among samples within the same class while minimize the similarities between samples from different classes.
- **Technically**, we propose a bi-level framework, MOKD, to first maximize the test power of kernels adopted in kernel HSIC and then optimize the kernel HSIC to control the dependence respectively between representations and labels and among all representations.
- **Empirically**, extensive experiments under several settings are conducted to verify the effectiveness of MOKD in improving generalization performance and alleviating the high similarities between samples.

Thank You!

Paper

Code