Multi-view Contrastive Learning for Online Knowledge Distillation

Chuanguang Yang[1,2], **Zhulin An**[1]*, **Xiaolong Hu**[1,2], **Hui Zhu**[1,2], **Kaiqiang Xu**[1,2], **Yongjun Xu**[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China
{yangchuanguang, anzhulin, huxiaolong18g}@ict.ac.cn
{zhuhui, xukaiqiang, xyj}@ict.ac.cn

Abstract

Existing Online Knowledge Distillation (OKD) aims to perform collaborative and mutual learning among multiple peer networks in terms of probabilistic outputs, but ignores the representational knowledge. We thus introduce Multi-view Contrastive Learning (MCL) for OKD to implicitly capture correlations of representations encoded by multiple peer networks, which provide various views for understanding the input data samples. Contrastive loss is applied for maximizing the consensus of positive data pairs, while pushing negative data pairs apart in embedding space among various views. Benefit from MCL, we can learn a more discriminative representation for classification than previous OKD methods. Experimental results on image classification and few-shot learning demonstrate that our MCL-OKD outperforms other state-of-the-art methods of both OKD and KD by large margins without sacrificing additional inference cost.

1 Introduction

Modern deep neural networks [20, 7] achieve predominant performance on many tasks but suffer prohibitive costs of computation and memory footprint. Knowledge Distillation (KD) [11] is an effective technique to improve the performance of a light-weight student model that is trained by the outputs derived from a high capacity teacher model, because probabilistic outputs contain richer information of classification evidences compared to hard labels. Classical KD always performs two-stage pipeline, in which we need to obtain a pre-trained teacher model with high capacity at first, so the training cost is significantly increased. Online Knowledge Distillation (OKD) [30, 30] solves the above regrets that it simultaneously performs collaborative and mutual learning among multiple peer student networks, where ensemble of the probabilistic outputs derived by some students always plays the role of an online teacher to further conduct knowledge transfer. OKD trains the student model by one-stage and end-to-end optimization without an additional teacher, which is more applicable than KD in practice. Our paper aims to further explore the effectiveness of OKD.

Previous methods of OKD [30, 31, 22] always perform transfer learning by the form of instance-level probabilistic outputs of independent data samples, but highly ignore the representational information holden in multiple student models. It has been observed that semantically correlated inputs always lead to the similar activations towards a trained network and vice versa, so transferring explicit relations of data samples from teacher to student has been extensively exploited [17, 25, 18, 24]. In this paper, we implement Multi-view Contrastive Learning (MCL) to implicitly capture correlations of encoded representations of data samples among multiple peer networks, where one peer network represents one view for understanding the input. In MCL, we try to maximize the agreement for

*Corresponding author.

Preprint. Under review.

the representations of the same input sample from various views, while pushing the representations of input samples with different labels from various views apart. Our motivation is inspired by the mechanism that various people always view a same objective in the real world with individual understandings, and the consensus is always quite robust for discriminating this objective. While each person may also carry an additional inductive bias as the noise information in their understandings. Here, all peer networks build a group of people and try to model the person-invariant representations.

Similar with the previous methods of OKD [31, 22, 3], our training graph contains multiple same networks, except that additional fully-connected layers for linearly transforming representations to the contrastive embedding space, in which we perform pair-wise contrastive loss among all peer networks. Importantly, all peer networks are supervised with the ground-truth labels to learn their individual understandings of classification evidences, which is the prerequisite for the success of MCL and an inborn advantage compared to unsupervised representational learning. Moreover, we also build an ensemble teacher from all peer networks similar with ONE [31] but is different in that the teacher only transfers probabilistic knowledge to the specific student network, which is used for final deployment. This practice alleviates the homogenization problem among multiple views and reduces the negative impact on MCL. Based on the above techniques, we name our framework as MCL-OKD. We conduct experiments on image classification tasks of CIFAR-100 [14] and ImageNet [5] across widely used networks to compare MCL-OKD against other State-Of-The-Art (SOTA) methods of both KD and OKD, the results show that our MCL-OKD achieves the best performance in terms of individual network and ensemble. Moreover, we observe that MCL-OKD combined with KD can further improve the performance gain. Extensive experiments on few-shot classification show the superiority of MCL-OKD in metric learning.

Overall, our contributions lie in three folds: (1) We systematically analyse the capacity of performance gain and limitations of existing OKD methods. (2) We propose MCL to exploit representational knowledge holden in peer networks. (3) We establish new SOTA results among existing KD and OKD methods across widely used networks.

2 Related Work

Contrastive learning. Contrastive learning [27, 16, 12, 8, 23, 4] has been extensively exploited for unsupervised visual representation learning. The main idea of contrastive learning is to perform contrastive loss on positive pairs against negative pairs, such that different data samples can be separated in the embedding space. Many prior works define positive and negative pairs from two views. CPC [16] summarizes all past features for generating a context representation to contrast future representations along with sequential data. Deep InfoMax [12] learns to match the input and its output from a deep neural network encoder. Instance Discrimination [27] performs instance-level contrast to learning a discriminative embedding space. SimCLR [4] considers the two views of the same data sample with different augmentation techniques, and maximize the consistency between them. Besides two views, CMC [23] and AMDIM [1] propose contrastive multi-view coding across the multiple sensory channels or independently-augmented copies of the input image, respectively. In this vein, we implement MCL by leveraging multiple peer networks to encode the same data sample, which is different from creating views in term of the data itself compared to previous contrastive learning, because our method is more advantageous to the scenario of supervised learning.

Knowledge distillation. The initial idea of KD is to transfer the output from a powerful teacher to a student for improving its generalization capacity proposed by Buciluǎ *et al.* [2] and Hinton *et al.* [11]. Some subsequent distillation methods aim to transfer the intermediate information, such as feature maps [19], attention maps [29], FSP matrices [28], activation boundaries [10, 9] and so on. Recently proposed KD methods aim to transfer the relations of data samples [17, 25, 18, 24], where the concurrent CRD [24] implements contrastive learning between teacher and student, which is related to us but different in several folds: we introduce multiple views for on-the-fly contrastive representations learning, while CRD performs only two views with a freezing teacher that implements contrastive-unaware learning, resulting in a suboptimal performance. Moreover, the reasoning process and optimized objective of MCL-OKD is clearly distinct compared to CRD.

OKD is applicable to the teacher-absent scenario. DML [30] shows that a group of students learn collaboratively and mutually that can improve the performance of individual networks. ONE [31] and CL-ILR [22] propose the frameworks which share the low-level layers to reduce the training

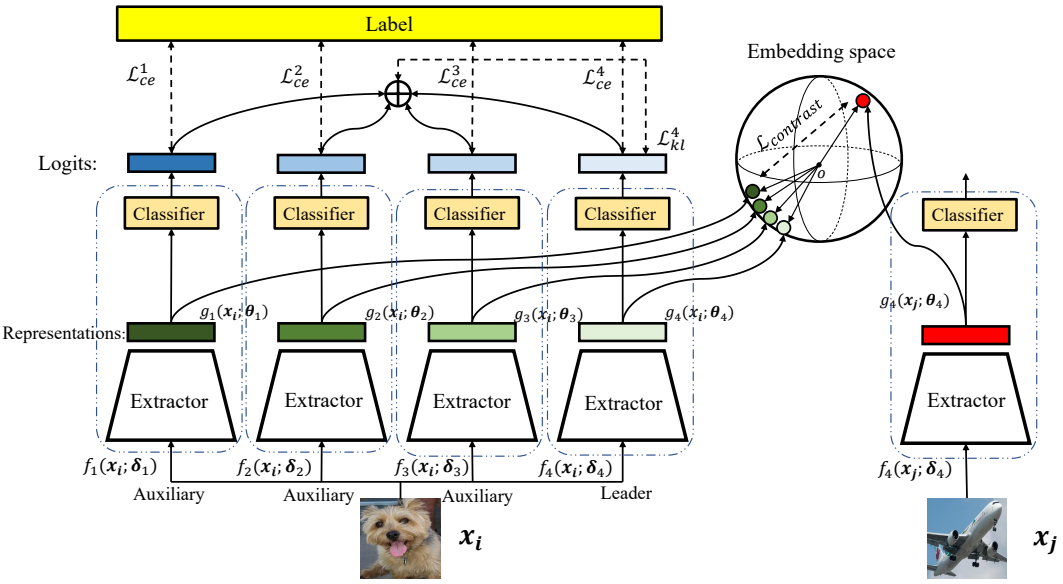

Figure 1: An overview of the proposed framework of multi-view contrastive learning for online knowledge distillation. Given a dog, 4 peer networks $\{f_m(\boldsymbol{x};\boldsymbol{\delta}_m)\}_{m=1}^4$ provide 4 views for generating representations, contrastive learning aims to push them closed, meanwhile pushing the representations derived from the samples with different labels apart in the sphere embedding space.

complexity and perform knowledge transfer among various branches of high-level layers. OKDDip [3] alleviates the homogenization problem in previous OKD methods by introducing two-level distillation and self-attention mechanism. All previous methods handle the probabilistic output to perform OKD, but differ in the ways of supervision. Our MCL-OKD further improves the performance of OKD from the perspective of representation learning.

3 Methodology

3.1 Distillation framework

Overall architecture. Similar with pervious frameworks of OKD [22, 31], adding additional network architectures to the object network for auxiliary training and distillation is an effective technique for improving its performance. Importantly, only the object network is kept and other additional architectures are discarded at inference time, resulting in no additional computational cost and memory footprint for practical deployment. In our distillation framework, we construct multiple peer networks to perform online MCL, where all peer networks have the same architecture so as to conform the protocol of pervious frameworks of OKD [30, 22, 31], as illustrated in Figure 1, where one peer network represents one view for the input data in our MCL.

Specifically, M peer networks $\{f_m(\boldsymbol{x};\boldsymbol{\delta}_m)\}_{m=1}^M$ participate in the process of distillation during training, where $\boldsymbol{x}$ denotes the input sample and $\boldsymbol{\delta}_m$ denotes the parameters of the m-th network, which includes a CNN feature extractor and a linear classifier. For contrastive learning, we aim to optimize the representations after the penultimate layer (before logits) among peer networks. To reduce the complexity, we linearly transform these representations into a relative lower-dim (128-d in this paper) embedding space, where we perform contrastive learning among multiple views. For ease of notation, we use $g_m(\boldsymbol{x};\boldsymbol{\theta}_m)$ to denote the composite of CNN feature extractor and linear transformation matrix of the m-th network, which connects the input samples to the contrastive embedding space. Similar with the previous branch-based OKD[22, 31], low-level layers across the M peer networks can also be shared to reduce the complexity and regularize the training networks.

Training and deployment. At the training stage, each peer network is supervised by hard labels, and only the M-th network is supervised by an additional soft probability distribution from the naive ensemble for the probabilistic outputs of all networks. Meanwhile, we perform contrastive learning

across each pair of the views to learn more discriminative representations. At the test stage, only the M-th network $f_M(\boldsymbol{x};\boldsymbol{\delta}_M)$ can be kept, which is named as the leader network, and other $M-1$ networks $\{f_m(\boldsymbol{x};\boldsymbol{\delta}_m)\}_{m=1}^{M-1}$ are discarded, which are named as auxiliary networks.

3.2 Learning objectives

Overall, we design three learning objectives: (1) Cross-Entropy (CE) between the probabilistic outputs and ground-truth labels for individual peer networks, (2) Kullback-Leibler (KL) divergence between the probability distribution of leader network and that of ensemble, (3) contrastive loss among the representations derived from peer networks.

Learning from labels. Given a training set $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$ including N samples with label space $\mathcal{Y}=\{1,2,\cdots,C\}$, where $\boldsymbol{x}_i$ is the i-th data with its label y_i. As same as the widely used practice on classification optimization, we use CE objective function between the probabilistic output of each peer network and one-hot ground-truth label distribution. Specifically, given a sample $\boldsymbol{x}_i$ with label y_i, probabilistic class posterior of the m-th peer network is $p_m(y_i|\boldsymbol{x}_i;\boldsymbol{\delta}_m)$, the CE objective across the training set of the m-th peer network can be expressed as:

$$\mathcal{L}_{ce}^m=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\eta_{c,y_i}\log p_m(y_i|\boldsymbol{x}_i;\boldsymbol{\delta}_m) \tag{1}$$

Where η_{c,y_i} is indicator that return 1 if $c=y_i$ else 0. $p_m(y_i|\boldsymbol{x}_i;\boldsymbol{\delta}_m)$ is calculated from the logits distribution $f_m(\boldsymbol{x}_i;\boldsymbol{\delta}_m)$ over C classes of the m-th peer network by the softmax normalization as:

$$p_m(y_i|\boldsymbol{x}_i;\boldsymbol{\delta}_m)=\frac{\exp\left(f_m^{y_i}(\boldsymbol{x}_i;\boldsymbol{\delta}_m)\right)}{\sum_{c=1}^{C}\exp\left(f_m^{c}(\boldsymbol{x}_i;\boldsymbol{\delta}_m)\right)} \tag{2}$$

Where $f_m^c(\boldsymbol{x}_i;\boldsymbol{\delta}_m)$ denotes the c-th value of the logits. It is noteworthy that supervision by the ground-truth label to each peer networks is crucial for the success of contrastive learning, for the account that multiple views from various peer networks on the same data could understand various classification evidences for the ground-truth label, motivating us to further model view-invariant information, which is quite robust and valuable for accurate classification.

Distillation from an online teacher. Inspired by the ONE [31], we also simply construct an online teacher by implementing the naive ensemble for the probabilistic outputs of all peer networks. The most difference of our distillation compared to ONE is that only the leader network $f_M(\boldsymbol{x};\boldsymbol{\delta}_M)$ can be optimized by soft probability distribution derived from online ensemble prediction. Because we empirically observed that transferring ensemble knowledge to each peer network in ONE gives rise to serious homogenization problem, which damages the diversity among individual peer networks meanwhile the performance of ensemble, thus further leading to ineffective modeling on MCL. We compute soft probability distribution of leader network $f_M(\boldsymbol{x};\boldsymbol{\delta}_M)$ as equ.3 and that of ensemble as equ.4 with a temperature T, where a higher T leads to softer distribution:

$$\tilde{p}_L(c|\boldsymbol{x}_i;\boldsymbol{\delta}_M)=\frac{\exp\left(f_M^c(\boldsymbol{x}_i;\boldsymbol{\delta}_M)/T\right)}{\sum_{d=1}^{C}\exp\left(f_M^d(\boldsymbol{x}_i;\boldsymbol{\delta}_M)/T\right)},c\in\{1,2,\cdots,C\} \tag{3}$$

$$\tilde{p}_E(c|\boldsymbol{x}_i;\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_M)=\frac{\exp\left(\frac{1}{M}\sum_{m=1}^{M}f_m^c(\boldsymbol{x}_i;\boldsymbol{\delta}_m)/T\right)}{\sum_{d=1}^{C}\exp\left(\frac{1}{M}\sum_{m=1}^{M}f_m^d(\boldsymbol{x}_i;\boldsymbol{\delta}_m)/T\right)},c\in\{1,2,\cdots,C\} \tag{4}$$

KL divergence is used for aligning the soft predictions between leader network and teacher as:

$$\mathcal{L}_{kl}^M=\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\tilde{p}_E(c|\boldsymbol{x}_i;\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_M)\log\frac{\tilde{p}_E(c|\boldsymbol{x}_i;\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_M)}{\tilde{p}_L(c|\boldsymbol{x}_i;\boldsymbol{\delta}_M)} \tag{5}$$

Multi-view contrastive learning. Besides transferring probability distribution, we also concentrate on leaning a discriminative embedding space, in which the representations from various views are mutually closed for the same data sample meanwhile far away between negative sample pairs. Given two different peer networks $g_a(\boldsymbol{x};\boldsymbol{\theta}_a)$ and $g_b(\boldsymbol{x};\boldsymbol{\theta}_b), a\neq b$, collections of generated representations across the training set are $\{g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)\}_{i=1}^N$ and $\{g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b)\}_{i=1}^N$, respectively. We define the positive

pair as $(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b))$, and the negative pair as $(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)), y_i \neq y_j$. Then we define a critic $h(\cdot,\cdot)$ with a constant scale factor τ to measure the cosine similarity for a sample pair:

$$h(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)) = \exp\left(\frac{g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)\cdot g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)}{\|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)\|\cdot\|g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)\|}\cdot\frac{1}{\tau}\right), i,j\in\{1,2,\cdots,N\} \tag{6}$$

We would like to make positive and negative pairs achieve high and low scores, respectively. Inspired by the recent settings [16], we consider a given representation $g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)$ of sample $\boldsymbol{x}_i$ from the fixing view g_a and enumerating the corresponding positive representation $g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b)$ and K negative representations $\{g_b(\boldsymbol{x}^{i,k};\boldsymbol{\theta}_b)\}_{k=1}^K$ from g_b view, and the probability of $g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b)$ matching $g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)$ as the positive pair is $p(g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a))$:

$$p(g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)) = \frac{h(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b))}{h(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b)) + \sum_{k=1}^K h(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}^{i,k};\boldsymbol{\theta}_b))} \tag{7}$$

We can perform contrastive loss by minimizing the negative log-likelihood:

$$\mathcal{L}_{contrast}^{g_a\to g_b} = -\frac{1}{N}\sum_{i=1}^N \log p(g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)) \tag{8}$$

In practice, straightforward optimization for equ. 8 is quite intractable, because the hyperparameter K can be very huge for large-scale dataset, e.g. around 1.2 million samples in ImageNet. To circumvent this problem, we adapt Noise-Contrastive Estimation (NCE) [6] to approximate the full softmax distribution, such that we no longer need to calculate the similarities with all samples in the dataset.

The idea behind NCE-based approximation is to transform the sample-level multi-classification into binary classification for discriminating positive pairs and negative pairs. Given the representation $g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)$ of $\boldsymbol{x}_i$ from the fixing view g_a, the probability of one representation $g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b), j\in\{1,2,\cdots N\}$ from the view g_b matching the $g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)$ is $p_p(g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a))$:

$$p_p(g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)) = \frac{h(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b))}{Z_i}, Z_i = \sum_{n=1}^N h(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_n;\boldsymbol{\theta}_b)) \tag{9}$$

$p_p(g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a))$ is a conditional distribution for being regarded as the positive pair. Following the prior practice [27], we define a uniform distribution for negative pairs, i.e. $p_n(g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)) = 1/N$, and we assume that frequency of sample pair lies in every S negative pairs along with 1 positive pair concurrently. Then the posterior probability of a given $g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)$ drawn from the actual data distribution of positive pair (denoted as $D=1$) is:

$$\begin{aligned} p(D=1|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)) &= \frac{p_p(g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a))}{p_p(g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a)) + Sp_n(g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b)|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a))} \\ &= \frac{h(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b))/Z_i}{h(g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_j;\boldsymbol{\theta}_b))/Z_i + S/N} \end{aligned} \tag{10}$$

We minimize negative log-likelihood of both positive and negative pairs, which approximates the contrastive loss from view g_a to view g_b as equ.11. $\{\boldsymbol{x}^{i,s}\}_{s=1}^S$ denotes a collection of negative samples of $\boldsymbol{x}_i$ randomly drawn from the training set and is updated along with each iterative batch in practice. Z_i is a normalizing constant in equ.10, we dynamically compute it from $\{\boldsymbol{x}^i, \{\boldsymbol{x}^{i,s}\}_{s=1}^S\}$ to reduce computation, and we empirically found this approximation performs well in practice.

$$\begin{aligned} \mathcal{L}_{contrast}^{g_a\to g_b} = -\frac{1}{N}\sum_{i=1}^N \mathop{\mathbb{E}}_{\{\boldsymbol{x}^i,\{\boldsymbol{x}^{i,s}\}_{s=1}^S\}} \{&\log(p(D=1|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}_i;\boldsymbol{\theta}_b))) \\ &+ \sum_{s=1}^S [\log(1-p(D=1|g_a(\boldsymbol{x}_i;\boldsymbol{\theta}_a), g_b(\boldsymbol{x}^{i,s};\boldsymbol{\theta}_b)))]\} \end{aligned} \tag{11}$$

Symmetrically, we can also fix view g_b and enumerate over view g_a, resulting in the mutual contrastive loss as $\mathcal{L}(g_a, g_b) = \mathcal{L}_{contrast}^{g_a\to g_b} + \mathcal{L}_{contrast}^{g_b\to g_a}$. We further extend contrastive learning from two views to multiple views, which includes two patterns: core view and full graph. For core view pattern, we focus on optimizing the view of leader network g_M, and contrast with the other $M-1$ views

Table 1: Comparison of error rate (Top-1, %) among network-based OKD methods on CIFAR-100 and ImageNet.

Dataset	Network	Baseline	DML	OKDDip	MCL-OKD
CIFAR-100	DenseNet-40-12	28.54 ± 0.34	26.64 ± 0.17	26.10 ± 0.03	$\mathbf{25.06 \pm 0.11}$
	ResNet-32	28.65 ± 0.03	26.47 ± 0.26	25.40 ± 0.08	$\mathbf{24.38 \pm 0.09}$
	VGG-16	26.08 ± 0.05	24.73 ± 0.23	24.88 ± 0.06	$\mathbf{23.51 \pm 0.05}$
	ResNet-110	23.82 ± 0.20	22.50 ± 0.11	21.09 ± 0.17	$\mathbf{20.23 \pm 0.14}$
ImageNet	ResNet-34	26.70	26.03	25.42	**24.60**

Table 2: Comparison of error rate (Top-1, %) among branch-based OKD methods on CIFAR-100 and ImageNet.

Dataset	Network	Baseline	CL-ILR	ONE	OKDDip	MCL-OKD
CIFAR-100	DenseNet-40-12	28.36 ± 0.17	27.38 ± 0.03	28.18 ± 0.13	28.21 ± 0.04	$\mathbf{25.87 \pm 0.14}$
	ResNet-32	27.45 ± 0.09	25.92 ± 0.10	25.97 ± 0.09	25.63 ± 0.14	$\mathbf{24.45 \pm 0.17}$
	VGG-16	25.95 ± 0.27	25.18 ± 0.21	25.63 ± 0.03	25.15 ± 0.19	$\mathbf{23.31 \pm 0.10}$
	ResNet-110	22.49 ± 0.04	21.04 ± 0.07	21.64 ± 0.27	21.00 ± 0.14	$\mathbf{19.31 \pm 0.42}$
ImageNet	ResNet-34	26.62	25.83	25.77	25.56	**24.71**

$\{g_m\}_{m=1}^{M-1}$, resulting in the contrastive loss $\mathcal{L}_{contrast} = \sum_{m=1}^{M-1} \mathcal{L}(g_M, g_m)$. For full graph pattern, we contrast each view pair and therefore model $\binom{M}{2}$ relationships, resulting in the contrastive loss $\mathcal{L}_{contrast} = \sum_{1 \le a < b \le M} \mathcal{L}(g_a, g_b)$. Compared to core view pattern, full graph pattern can capture more shared information among various peer networks, but resulting in more complexity. In our implementation, we use full graph pattern to pursue a high performance for downstream classification. Inspired by Wu *et al.* [27], we create a memory bank to store the feature vectors of all data samples in training set computed from previous batches, which allows us efficiently obtaining abundant negative samples to finish the contrastive computing.

Overall learning objective. We combine above three objectives to construct our final objective:

$$\mathcal{L}_{MCL-OKD} = \sum_{m=1}^{M} \mathcal{L}_{ce}^{m} + T^2 \mathcal{L}_{kl}^{M} + \beta \mathcal{L}_{contrast} \tag{12}$$

Where T^2 is used for balancing contributions between hard and soft labels, β is a constant factor for rescaling the magnitude of contrastive loss.

4 Experiments

We evaluate the effectiveness of our MCL-OKD framework for image classification and few-shot learning. And we provide analysis for learned capability and representational knowledge of OKD.

4.1 Image Classification

Dataset and setup. We use CIFAR-100 [14] and ImageNet [5] benchmark datasets for evaluations. We use $T = 3$ and $\beta = 0.025$ in equ.12, where we empirically found $\beta \in [0.01, 0.04]$ works reasonably well. Following the prior practice on contrastive learning [27, 3], we use $\tau = 0.1$ and $\tau = 0.07$ in equ.6 for CIFAR-100 and ImageNet respectively, and $S = 16384$ in equ.11. We use 4 peer networks in all OKD methods, i.e. $M = 4$, unless otherwise specified. On CIFAR-100, we report the mean accuracy with standard deviation over 3 runs. More introduction about datasets, settings of OKD methods, training details are provided in Section 1 of Supplementary Material.

Results of OKD methods. Table 1 and Table 2 show the comparisons of the performance among SOTA OKD methods across the widely used architectures of VGG [20], ResNet [7] and DenseNet [13], where Table 1 aims to the network-based setting, and Table 2 investigates the branch-based setting. It can be obviously observed that our MCL-OKD consistently outperforms all other methods such as DML [30], CL-ILR [22], ONE [31] and OKDDip [3] by large margins, which

Table 3: Comparison of ensemble error rate (Top-1, %) among OKD methods including branch-based (B) and network-based (N) methods on CIFAR-100. Blue/Red: Best and second best results.

Network	Baseline		DML	CL-ILR	ONE	OKDDip		MCL-OKD	
	B	N	N	B	B	B	N	B	N
DenseNet-40-12	**26.32**	**23.59**	25.70	26.36	27.80	27.43	23.74	**23.76**	**22.82**
ResNet-32	**23.29**	**22.67**	26.47	24.84	24.87	23.45	22.77	**22.24**	**22.02**
VGG-16	25.64	**22.04**	22.12	24.99	25.54	**24.95**	22.33	**22.97**	**21.22**
ResNet-110	**19.45**	**19.53**	20.81	19.55	20.26	19.54	19.55	**18.76**	**18.58**

Table 4: Comparison of error rate (Top-1, %) among SOTA methods of KD and OKD for ResNet-32 with an additional teacher ResNet-110 on CIFAR-100. Extensive comparison with other KD methods across various architectures is shown in Section 2 of Supplementary Material.

Baseline	Teacher	KD	CRD+KD	OKDDip+KD	MCL-OKD+KD
28.68 ± 0.24	25.40	26.92 ± 0.18	26.25 ± 0.24	24.92 ± 0.08	$\mathbf{24.14 \pm 0.18}$

indicates that performing MCL on the representations among various peer networks is more effective than straightforward learning from probabilistic outputs that previous OKD methods always focus on. Compared to the previous SOTA OKDDip, MCL-OKD achieves $1.45\times$ and $2.65\times$ reductions of error rate on average for network-based and branch-based settings on CIFAR-100, respectively. Further experiments on the more challenge ImageNet validate that MCL-OKD significantly outperforms the OKDDip. Moreover, we observe that the performance of individual network of baseline in branch-based setting generally outperforms its network-based version, we conjecture that the shared low-level layers receive gradients from various branches, thus resulting in a more robust optimization for generating better representations.

In terms of training complexity, we adopt ResNet-34 on ImageNet for illustration. Compared to other OKD methods, MCL-OKD introduces additional 1.56 GFLOPS to perform contrastive computing for representations, which is around 11% of the original 14.7 GFLOPS. In practice, we experimentally observed no conspicuous increase of training time (e.g., 2.2 hours/epoch v.s. 2.4 hours/epoch on a single NVIDIA Tesla V100 GPU). The memory bank of each peer network needs about 600MB memory for storing all 128-d features, resulting in the total 2.4GB memory for 4 peer networks.

Ensemble results of OKD. As shown in Table 3, MCL-OKD can still perform the best among all other OKD methods in terms of the ensemble performance without discarding any auxiliary networks in both network-based and branch-based settings. Somewhat surprisingly, previous OKD methods always underperform the baseline, we conjecture that albeit the mutual supervisions of probabilistic outputs by other peer networks improve the performance of individual networks, they may lead to homogenization among peer networks and damage the learned capacity of ensemble knowledge. Only MCL-OKD can consistently outperform the baseline, which suggests that our method can learn a more powerful capacity of ensemble knowledge benefiting from representation learning.

Results of OKD with a teacher. We further exploit the performance gains for teacher-free OKD methods when a pre-trained teacher is available, and compared to SOTA KD method of CRD [24]. Here, the student model for OKD is in branch-based setting, and only the leader network is supervised by the teacher. As shown in Table 4, we can observe that MCL-OKD achieve better results than pervious SOTA CRD and OKDDip with $1.87\times$ and $1.21\times$ reductions of error rate, respectively. It also may reveal that combined OKD with KD can maximize the performance improvement.

4.2 Few-shot Learning

Dataset and setup. We use standard Omniglot [15] and *mini*ImageNet [26] benchmarks for few-shot classification to evaluate the performance of OKD methods. Prototypical network [21] is used as the baseline to perform few-shot learning, which also plays the role of peer network in OKD. We use the standard data split following Snell *et al.* [21]. At the test stage, we report average accuracy over 1000 randomly sampled episodes for Omniglot, and 600 randomly sampled episodes with 95% confidence intervals for *mini*ImageNet. For OKD methods, we add an auxiliary global classifier to the original

Table 5: Comparison of accuracy (Top-1, %) among KD and OKD methods on few-shot learning.

Model	Omniglot				*mini*ImageNet	
	5-way		20-way		5-Way	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
Baseline	98.55	99.56	95.11	98.68	49.10 ± 0.82	66.87 ± 0.66
CL_ILR	98.56	99.63	95.08	98.67	50.75 ± 0.40	67.75 ± 0.32
ONE	98.46	99.65	94.85	98.68	50.67 ± 0.41	67.58 ± 0.33
OKDDip	98.55	99.66	95.16	98.68	50.60 ± 0.42	67.41 ± 0.33
RKD-D	98.58	99.65	95.45	98.72	49.66 ± 0.84	67.07 ± 0.67
RKD-DA	98.64	99.64	**95.52**	98.67	50.02 ± 0.83	68.16 ± 0.67
MCL-OKD	**98.68**	**99.67**	95.40	**98.75**	$\mathbf{51.58 \pm 0.41}$	$\mathbf{69.49 \pm 0.33}$

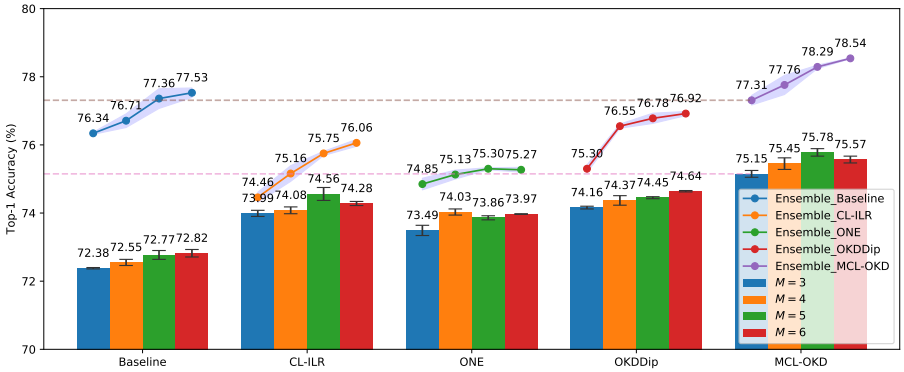

Figure 2: Comparison of accuracy among OKD methods with $3 \sim 6$ peer networks on CIFAR-100.

prototypical network over the class space of training set, and perform learning of probabilistic outputs among 4 peer networks. Moreover, we adopt $T = 4$, $\beta = 0.25$ and $\tau = 0.1$ in MCL-OKD.

Results. Table 5 compares accuracy among SOTA KD and OKD methods on few-shot learning. We can observe that MCL-OKD significantly outperforms other OKD methods across the several few-shot settings, which demonstrates that contrastive learning on representations is more effective than learning on probabilistic outputs especially for metric learning tasks, due to the superiority of generating discriminative feature embeddings for samples from newly unseen classes. Moreover, MCL-OKD achieves better results than RKD [17], which is the SOTA KD method for few-shot learning that needs a pre-trained teacher as the prerequisite.

4.3 Analysis of OKD methods

Number of peer networks. It is also valuable to exploit that whether the performance of OKD can be further improved as the number of peer networks increases. We use ResNet-32 as the base model and implement the branch-based OKD methods with $3 \sim 6$ peer networks, the results in terms of individual network and ensemble are shown in Figure 2. It is obvious that MCL-OKD achieves the best performance among OKD methods in all cases, even if under the minimum capacity with 3 peer networks, MCL-OKD can still significantly outperform all other methods which have 6 peer networks. Moreover, the results of ensembles trained by MCL-OKD consistently outperform the original baselines, which surprisingly perform the second best. This evidence further demonstrates that MCL-OKD learns the crucial representational knowledge that previous methods ignore.

Learned similarities among data samples. To demonstrate that MCL-OKD can indeed learn a more discriminative representation space, we provide the comparison for the cosine similarities of intra-class samples and inter-class samples among branch-based OKD methods. The detailed visualization is shown in Section 3 of Supplementary Material. It can be observed that MCL-OKD learns a larger intra-class similarity meanwhile a smaller inter-class similarity than other OKD methods, which suggests that the margins between data samples and classification boundary are the

farthest trained by MCL-OKD. We consider that MCL-OKD constructs a more robust classification boundary in representation space, which leads to the significant reduction of error rate.

5 Conclusion

We propose multi-view contrastive learning for OKD to learn a more powerful representation benefiting from the mutual communications among peer networks. Experimental evidences show that MCL-OKD significantly outperforms all existing OKD methods in terms of individual network and ensemble, which proves the better superiority of learning informative representations than probabilistic outputs alone. When a pre-trained teacher is available, the performance gain can be further improved, which makes our MCL-OKD become a prior choice for model deployment.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15509–15519, 2019.

[2] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[3] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3430–3437, 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[9] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1921–1930, 2019.

[10] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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

[15] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[17] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[18] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5007–5016, 2019.

[19] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[22] Guocong Song and Wei Chai. Collaborative learning for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 1832–1841, 2018.

[23] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[24] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[25] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019.

[26] 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.

[27] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[28] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[29] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[30] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[31] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. In *Advances in neural information processing systems*, pages 7517–7527, 2018.