Peer Collaborative Learning for Online Knowledge Distillation

Guile Wu and Shaogang Gong
Queen Mary University of London
{guile.wu, s.gong}@qmul.ac.uk

Abstract

Traditional knowledge distillation uses a two-stage training strategy to transfer knowledge from a high-capacity teacher model to a smaller student model, which relies heavily on the pre-trained teacher. Recent online knowledge distillation alleviates this limitation by collaborative learning, mutual learning and online ensembling, following a one-stage end-to-end training strategy. However, collaborative learning and mutual learning fail to construct an online high-capacity teacher, whilst online ensembling ignores the collaboration among branches and its logit summation impedes the further optimisation of the ensemble teacher. In this work, we propose a novel Peer Collaborative Learning method for online knowledge distillation. Specifically, we employ a multi-branch network (each branch is a peer) and assemble the features from peers with an additional classifier as the peer ensemble teacher *to transfer knowledge from the high-capacity teacher to peers and to further optimise the ensemble teacher. Meanwhile, we employ the temporal mean model of each peer as the* peer mean teacher *to collaboratively transfer knowledge among peers, which facilitates to optimise a more stable model and alleviate the accumulation of training error among peers. Integrating them into a unified framework takes full advantage of online ensembling and network collaboration for improving the quality of online distillation. Extensive experiments on CIFAR-10, CIFAR-100 and ImageNet show that the proposed method not only significantly improves the generalisation capability of various backbone networks, but also outperforms the state-of-the-art alternative methods.*

1. Introduction

Deep learning has achieved incredible success in many computer vision tasks in recent years. Whilst many studies focus on developing deeper and/or wider networks for improving the performance [5, 26, 24], these cumbersome networks require more computational resources hindering their deployments in resource-limited scenarios. To alleviate this problem, knowledge distillation is developed to transfer knowledge from a stronger teacher [6] or an online ensemble [13] to a student model, which is more suitable for deployment.

Traditionally, knowledge distillation requires to pre-train a high-capacity teacher model in the first stage, and then transfer the knowledge of the teacher to a smaller student model in the second stage [6, 18, 17]. Via aligning the soft prediction [6] or the feature representation [18] between the teacher and the student, the student model usually obtains approximate accuracy as the teacher, but significantly reduces the model complexity for deployment. However, this traditional strategy usually requires more training time and computational cost, since the teacher and the student are trained in two separate stages.

On the other hand, recent online knowledge distillation [13, 28, 1] proposes to directly optimise the target network, following a one-stage end-to-end training strategy. Instead of pre-training a high-capacity teacher, online distillation typically integrates the teacher into the student model using a hierarchical network with shared intermediate-level representations [21] (Fig. 1(a)), multiple parallel networks for mutual distillation [28](Fig. 1(b)), or a multi-branch network with online ensembling [13] (Fig. 1(c)). Although these methods have shown their superiority over their traditional counterparts, collaborative learning and mutual learning fail to construct a stronger ensemble teacher to transfer knowledge from a high-capacity teacher to a student, whilst online ensembling ignores the collaboration among branches and its logit summation impedes the further optimisation of the ensemble teacher.

In this work, we propose a novel Peer Collaborative Learning (PCL) method for online knowledge distillation. As shown in Fig. 1(d), we integrate online ensembling and network collaboration into a unified framework to take full advantage of them for improving the quality of online distillation without pre-training a high-capacity teacher model. Specifically, we construct a multi-branch network (each branch is a peer), in which the low-level layers are shared and the high-level layers are separated. To facilitate the online distillation, wo employ two type of peer collaborations:

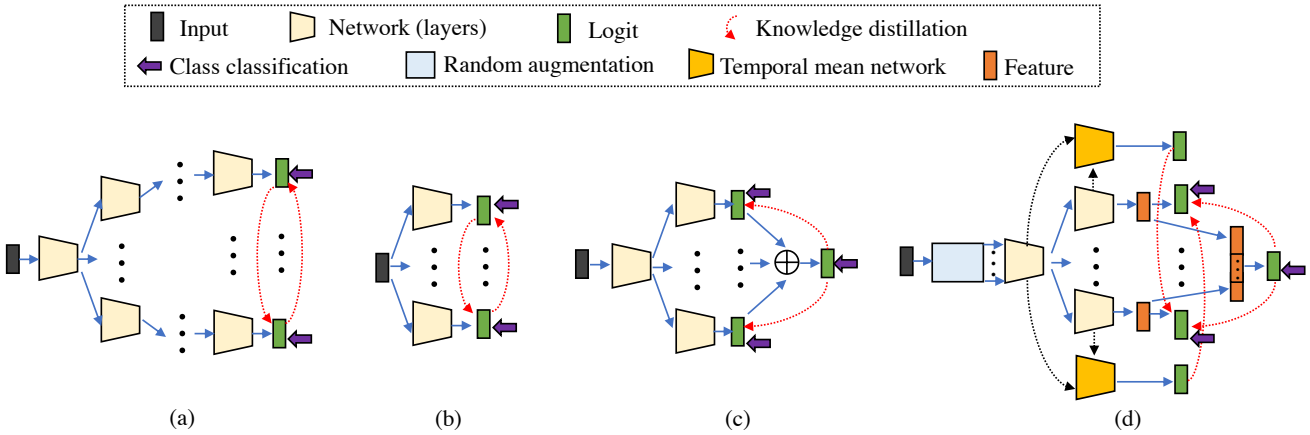

Figure 1. The diagrams of four online knowledge distillation mechanisms. (a) Collaborative learning. (b) Mutual learning. (c) Online ensembling. (d) Peer collaborative learning (Proposed). Our method integrates two types of peer collaborations (*i.e.* peer ensemble teacher and peer mean teacher) into a unified framework to improve the quality of online distillation.

(1) We assemble the feature representations of peers with an additional classifier as the *peer ensemble teacher*; (2) We use the temporal mean model of each peer as the *peer mean teacher*. The first teacher is a high-capacity model, which helps to distil ensembled knowledge from a stronger teacher to each peer, and in turn further improves the ensemble teacher model. Instead of using peer logit summation as the ensemble teacher [13], we assemble the features of peers with an additional classifier as the ensemble teacher to learn discriminative information among the peer feature representations and to further optimise the teacher. Since peers are separated in the multi-branch network, the training error among peers will be accumulated during online ensembling, which impairs the collaboration among peers. To alleviate this limitation, we use the second teacher to collaboratively distil knowledge among peers. Instead of using mutual learning among peers [28], we utilise the temporal mean of shared low-level layers and separated high-level layers to construct the peer mean teacher to distil knowledge among peers, which helps to alleviate the accumulation of training error among peers and to facilitate the optimisation of a more stable model. Besides, we perform random augmentation multiple times on the input to each peer to further enhance the diversity of knowledge learned in the peers. In test, we use the temporal mean network of a peer for deployment, which has the identical number of parameters as the backbone network, so there is no extra inference cost for deployment. Furthermore, the output feature representations plus the additional classifier from the peer mean teachers forms a high-capacity ensemble for deployment to get better performance in the scenarios where computational cost is less constrained.

Extensive experiments on CIFAR-10 [11], CIFAR-100 [11] and ImageNet [19] using various backbone networks (ResNet [5], VGG [20], DenseNet [9], WRN [26] and ResNeXt [24]) show that the proposed method significantly improves the generalisation capability of the backbone networks and outperforms the state-of-the-art alternatives.

2. Related Work

Traditional Knowledge Distillation [6] is one of the most effective solutions to compress a cumbersome model or an ensemble of models into a smaller model. In [6], Hinton *et al*. firstly introduce the process of transferring the knowledge from a high-capacity teacher model to a compact student model as "distillation", which is accomplished by aligning the soft output prediction between the teacher and the student. After that, many promising knowledge distillation methods have been designed to facilitate the optimisation process of distillation via exploiting various "knowledge", such as intermediate representations [18], flow between layers [25], attention maps [27], structural relations [16] and activation similarity [23]. Although these methods have shown promising performance in compressing the model for deployment, they typically follow a two-stage training solution by pre-training a high-capacity teacher model for transferring knowledge to a compact student model, which requires more training time and computational cost.

Online Knowledge Distillation [13, 10, 1, 28] follows a one-stage end-to-end training strategy to optimise the target network for deployment with knowledge distillation among multiple networks or branches without pre-training a high-capacity teacher. In [21], Song *et al*. propose to distil knowledge among multiple classifier heads of a hierarchical network for improving the generalisation capability of the network without extra inference cost. In [28], Zhang *et*

al. introduce a mutual learning solution to distil knowledge among multiple parallel networks with the same input. Although these methods help to improve the generalisation of the target network, they only distil limited knowledge among parallel networks or heads and fail to construct stronger online teachers to further improve the students. More similar to our work, Lan *et al*. [13] use a multi-branch network and assemble logits from multiple branches (students) as the teacher to improve the generalisation of each student network. However, the logit summation impedes the further optimisation of the ensemble teacher, and online ensembling ignores the collaboration among branches, resulting in suboptimal performance. In [10], Kim *et al*. integrate feature representations of multiple branches into the online ensembling, but their method requires more convolutional operations for the feature fusion and also fails to exploit the collaboration among branches. To address these limitations, in our work: (2) we assemble the features from peers with an additional classifier as the *peer ensemble teacher* to distil knowledge from a stronger teacher to each peer and to further optimise the teacher; (1) we exploit the temporal mean models of each peer as the *peer mean teacher* to distil knowledge among peers, which helps to optimise a more stable model and alleviate the accumulation of training error. The integration of these two teachers into a unified framework enables to significantly improve the generalisation capability of each peer and the ensemble, resulting in better performance.

Neural Network Ensembling is a simple and effective solution for improving the generalisation performance of a model [4, 29, 15]. Although this can usually bring better performance, training multiple neural networks to create an ensemble requires significantly more training time and computational cost. Recent trend in neural network ensembling focuses on training a single model and exploiting different phases of the model as an ensemble for performance improvement. In [8], Huang *et al*. force the model to visit multiple local minima and use the corresponding models as the snapshots for neural network ensembling. In [12], Laine *et al*. propose to use temporal ensembling of network predictions over multiple training epochs as the teacher to facilitate the optimisation of the current model for semi-supervised learning. Our work differs from these works in that we use the feature representations of peers from a multi-branch network as the ensemble teacher for online knowledge distillation, instead of using the network predictions from different phases or generating multiple networks for ensembling. The peer collaborative distillation by peer mean teachers in our method shares the merit of mean teacher [22]. In [22], network weights over previous training epochs are averaged as a teacher to minimise the distance of predictions between the student and the teacher as the consistency regularisation for semi-supervised learning. In contrast, our method uses the shared layers and multiple separated layers to form multiple peer mean teachers as the more accurate peer mean teachers to alleviate the accumulation of training error and to generate a more stable teacher for online distillation.

3. Peer Collaborative Learning

3.1. Approach Overview

The overview of the proposed Peer Collaborative Learning (PCL) is depicted in Fig. 2. We employ a m-branch network for model training and define each branch as a peer. Since the low-level layers across different branches usually contain similar low-level features regarding minor details of images, sharing them enables to reduce the training cost and improve the collaboration among peers. We therefore share the low-level layers and separate the high-level layers in the m-branch network.

As shown in Fig. 2, to facilitate online knowledge distillation, we use the feature concatenation of peers as the *peer ensemble teacher* and use the temporal mean model of each peer as the *peer mean teacher*. The training optimisation objective of PCL contains three components: The first component is the standard cross-entropy loss for multi-class classification of the peers ($\mathcal{L}_{ce}^{p}$) and the peer ensemble teacher ($\mathcal{L}_{ce}^{t}$); The second component is the peer ensemble distillation loss $\mathcal{L}_{pe}$ for transferring knowledge from a stronger teacher to the student, which in turn improves the ensemble teacher; The third component is the peer mean distillation loss $\mathcal{L}_{pm}$ for collaboratively distilling knowledge among peers. The overall training objective $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_{ce}^{p} + \mathcal{L}_{ce}^{t} + \mathcal{L}_{pe} + \mathcal{L}_{pm} \tag{1}$$

In test, we use a temporal mean network of a peer for deployment, which has the identical number of parameters as the backbone network, so there is no extra inference cost for deployment. In the scenarios where computational cost is less constrained, the output feature representations from the peer mean teachers can form a high-capacity ensemble teacher model to get better accuracy for deployment.

3.2. Peer Ensemble Teacher

Input Augmentation for Peers. Suppose there are n samples in a training dataset $\{(x_i, y_i)\}_{i=1}^{n}$, where x_i is the i-th input sample, $y_i \in \{1, 2, ..., C\}$ is the corresponding label, and C is the number of classes in the dataset ($C \leq n$). Existing multi-branch online distillation methods [13, 1] directly use x_i (applying random augmentation once) as the input to all the branches, which causes the homogenisation among peers and decreases the generalisation of the network. To alleviate this problem, we apply random augmentation m times to x_i to produce m counterparts of x_i

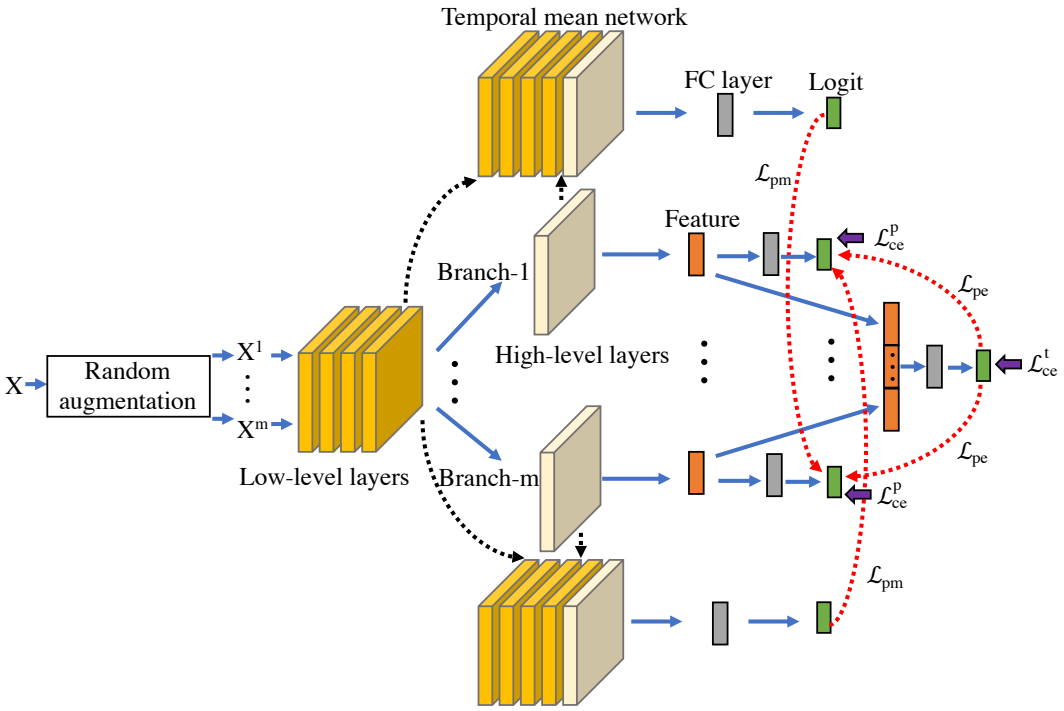

Figure 2. The overview of Peer Collaborative Learning (PCL) for online knowledge distillation. In the multi-branch network, the low-level layers are shared and the high-level layers are separated. The input is randomly transformed multiple times to generate the individual input to each peer. The output features of each peer form the stronger peer ensemble teacher, while the temporal mean models of peers form the peer mean teachers, which together form a unified framework for online knowledge distillation.

(*i.e.* $\{x_i^1, x_i^2, ..., x_i^m\}$, and use each counterpart as the input to each peer. This provides additional richer knowledge of the inputs, which improves the generalisation among peers in addition to the random model initialisation.

Online Ensembling. To construct a stronger teacher online for improving online distillation, logits from multiple networks/branches are usually summed (w/ or w/o attention gates) [13]. However, this hinders the ensemble teacher from further optimisation and ignores the discriminative information among peer representations, which might lead to a suboptimal solution since the summation is not further learned. In our work, we concatenate the features from peers and use an additional fully connected layer for classification to produce a stronger peer ensemble teacher. Therefore, the multi-class classification is performed for both the peer and the teacher as:

$$\mathcal{L}_{ce}^{p} = -\sum_{j=1}^{m}\sum_{c=1}^{C} y_c log \frac{exp(z_{j,c}^{p})}{\sum_{k=1}^{C} exp(z_{j,k}^{p})} \tag{2}$$

$$\mathcal{L}_{ce}^{t} = -\sum_{c=1}^{C} y_c log \frac{exp(z_{c}^{t})}{\sum_{k=1}^{C} exp(z_{k}^{t})} \tag{3}$$

where $z_{j,c}^{p}$ is the output logit from the last fully connected layer of the j-th peer over a class c, y_c is the ground-truth label indicator, z_c^t is the output logit from the fully connected layer of the peer ensemble teacher over a class c.

To transfer knowledge from the ensemble teacher to each peer, we compute the soft prediction of the j-th peer and the ensemble teacher as:

$$p_{j,c}^{p} = \frac{exp(z_{j,c}^{p}/T)}{\sum_{k=1}^{C} exp(z_{j,k}^{p}/T)}, \; p_{c}^{t} = \frac{exp(z_{c}^{t}/T)}{\sum_{k=1}^{C} exp(z_{k}^{t}/T)} \tag{4}$$

where T is a temperature parameter [6], $p_{j,c}^{p}$ is the soft prediction of the j-th peer over a class c, and p_c^t is the soft prediction of the ensemble teacher over a class c. Using Kullback Leibler (KL) divergence, the peer ensemble distillation loss $\mathcal{L}_{pe}$ is formulated as:

$$\mathcal{L}_{pe} = \omega(e) \cdot T^2 \sum_{j=1}^{m}\sum_{c=1}^{C} p_c^t \cdot log \frac{p_c^t}{p_{j,c}^p} \tag{5}$$

where e is the current training epoch and $\omega(\cdot)$ is a weight ramp-up function [12] defined as:

$$\omega(e) = \begin{cases} \lambda \cdot exp(-5 * (1 - \frac{e}{\alpha})^2) & , e \leq \alpha \\ \lambda & , e > \alpha \end{cases} \tag{6}$$

where α is the epoch threshold for the ramp-up function and λ is a parameter weighting the gradient magnitude.

Remarks. The proposed peer ensemble teacher differs from existing feature fusion [7, 10, 2] in that we concatenate features from a multi-branch network and use a fully connected layer for classification without using additional convolutional operation or multiple networks. More importantly, as the input to each peer is performed with random augmentation, richer peer knowledge is exploited in the online ensembling resulting in the further improvement of online knowledge distillation.

3.3. Peer Mean Teacher

Online ensembling is capable of constructing a stronger teacher for online distillation, but it ignores the collaboration among peers. On the other hand, mutual learning [28] and collaborative learning [21] benefit from mutual distillation among networks/heads, but they fail to construct a stronger teacher to further facilitate the optimisation among peers. In our work, we further use peer mutual distillation for improving the collaboration among peers. Instead of directly transferring knowledge among peers, which might accumulate the training error, we use temporal mean models [22] of each peer as the peer mean teacher for peer collaborative distillation.

We denote the weights of the shared low-level layers as θ_l and the weights of the separated high-level layers of the j-th peer as $\theta_{h,j}$. At the g-th global training step [1], the j-th peer mean teacher $\{\theta^t_{l,g}, \theta^t_{h,j,g}\}$ is formulated as:

$$\begin{cases} \theta^t_{l,g} = \phi(g)\cdot\theta^t_{l,g-1} + (1-\phi(g))\cdot\theta_{l,g} \\ \theta^t_{h,j,g} = \phi(g)\cdot\theta^t_{h,j,g-1} + (1-\phi(g))\cdot\theta_{h,j,g} \end{cases} \tag{7}$$

where $\theta^t_{l,g}$ are the weights of the shared low-level layers of the peer mean teachers, $\theta^t_{h,j,g}$ are the weights of the separated high-level layers of the j-th peer mean teacher, $\phi(g)$ is a smoothing coefficient function defined as:

$$\phi(g) = min(1-\frac{1}{g}, \beta) \tag{8}$$

where β is the smoothing coefficient hyper-parameter. Note that, the additional classifier of the peer ensemble teacher is also aggregated for the ensemble deployment. We compute the soft predictions $p^{mt}_{j,c}$ of the j-th mean teacher over a class c using Eq. (4) with the output logit $z^{mt}_{l,c}$ of this mean teacher. Thus, the peer mean teacher distillation loss $\mathcal{L}_{pm}$ is formulated as:

$$\mathcal{L}_{pm} = \omega(e)\cdot T^2 \frac{1}{m-1}\sum_{j=1}^{m}\sum_{l=1,l\neq j}^{m}\sum_{c=1}^{C} p^{mt}_{l,c}\cdot log\frac{p^{mt}_{l,c}}{p^{p}_{j,c}} \tag{9}$$

Remarks. Traditional mean teacher is used for semi-supervised/unsupervised learning [22, 14, 3], which mainly enforces the distance between the model predictions to be close. In our work, we integrate it into a multi-branch network for the peer collaborative distillation during online knowledge distillation by aligning the soft distributions between the peer and its counterpart's mean teacher. Compared with mutual distillation among peers [28] which might accumulate the training error, averaging model weights temporally over training epochs enables the peer mean teacher to stabilise soft predictions for improving peer collaboration (see Experiment for further analysis). Besides, as the input to each peer/mean teacher is randomly transformed, $\mathcal{L}_{pm}$ helps each peer to learn richer knowledge, resulting in the improvement of online distillation.

[1] In mini-batch training, $g = e\cdot Batch_{num} + Batch_{inx}$, where $Batch_{num}$ is the total number of training batches and $Batch_{inx}$ is the index of the current batch.

Algorithm 1 Peer Collaborative Learning for Online Knowledge Distillation.

Input: Training data $\{(x_i, y_i)\}_{i=1}^n$.
Output: A trained target model $\{\theta^t_l, \theta^t_{h,1}\}$, and a trained ensemble model $\{\theta^t_l, \theta^t_{h,j}\}_{j=1}^m$.

```
/* Training */
Initialisation: Randomly initialise model parameters
for e = 0→Epoch_max do    /* Mini-Batch SGD */
    Randomly transform x_i to get counterparts {x_i}_{j=1}^m
    Compute features and logits ({z_j^p}_{j=1}^m) of each peer
    Concatenate features as the peer ensemble teacher
    Compute the logit z^t of the ensemble teacher
    Compute peer classification loss L_ce^p (Eq.(2))
    Compute ensemble classification loss L_ce^t (Eq.(3))
    Compute peer ensemble loss L_pe(Eq.(5))
    Compute mean teacher loss L_pm(Eq.(9))
    Update peer models with Eq.(1)
    Update peer mean teachers with Eq.(7)
end for
/* Testing */
Deploy with a single target model {θ_l^t, θ_{h,1}^t}
Deploy with an ensemble model {θ_l^t, θ_{h,j}^t}_{j=1}^m
```

Summary. As summarised in Algorithm 1, the proposed PCL follows a one-stage end-to-end training strategy without pre-training a high-capacity teacher model. With the peer ensemble teacher and peer mean teachers in a multi-branch network, peers collaborate to improve the quality of online knowledge distillation. In test, we use a single peer mean teacher model as the target model (PCL) without adding extra inference cost. Besides, the ensemble peer mean teachers can also form a high-capacity teacher model (PCL-E) for deployment.

Table 1. Comparisons with the state-of-the-arts on CIFAR-10. Top-1 error rates (%).

Network	DML [28]	CL [21]	ONE [13]	FFL-S [10]	OKDDip [1]	Baseline	PCL(ours)
ResNet-32	6.06±0.07	5.98±0.28	5.80±0.12	5.99±0.11	5.83±0.15	6.74±0.15	**5.67±0.12**
ResNet-110	5.47±0.25	4.81±0.11	4.84±0.30	5.28±0.06	4.86±0.10	5.01±0.10	**4.47±0.16**
VGG-16	5.87±0.07	5.86±0.15	5.86±0.23	6.78±0.08	6.02±0.06	6.04±0.13	**5.26±0.02**
DenseNet-40-12	6.41±0.26	6.95±0.25	6.92±0.21	6.72±0.16	7.36±0.22	6.81±0.02	**5.87±0.13**
WRN-20-8	4.80±0.13	5.41±0.08	5.30±0.14	5.28±0.13	5.17±0.15	5.32±0.01	**4.58±0.04**
ResNeXt-29-2×64d	4.46±0.16	4.45±0.18	4.27±0.10	4.67±0.04	4.34±0.02	4.72±0.03	**3.93±0.09**

Table 2. Comparisons with the state-of-the-arts on CIFAR-100. Top-1 error rates (%).

Network	DML [28]	CL [21]	ONE [13]	FFL-S [10]	OKDDip [1]	Baseline	PCL(ours)
ResNet-32	26.32±0.14	27.67±0.46	26.21±0.41	27.82±0.11	26.75±0.38	28.72±0.19	**25.86±0.16**
ResNet-110	22.14±0.50	21.17±0.58	21.60±0.36	22.78±0.41	21.46±0.26	23.79±0.57	**20.02±0.55**
VGG-16	24.48±0.10	25.67±0.08	25.63±0.39	29.13±0.99	25.32±0.05	25.68±0.19	**23.11±0.25**
DenseNet-40-12	26.94±0.31	28.55±0.34	28.40±0.38	28.75±0.35	28.77±0.14	28.97±0.15	**26.91±0.16**
WRN-20-8	20.23±0.07	20.60±0.12	20.90±0.39	21.78±0.14	21.17±0.06	21.97±0.40	**19.49±0.49**
ResNeXt-29-2×64d	18.94±0.01	18.41±0.07	18.60±0.25	20.18±0.33	18.50±0.11	20.57±0.43	**17.38±0.23**

4. Experiment

4.1. Datasets and Settings

Datasets. We used three image classification benchmarks for evaluation: (1) *CIFAR-10* [11] contains 60000 images in 10 classes, with 5000 training images and 1000 test images per class. (2) *CIFAR-100* [11] consists of 60000 images in 100 classes, with 500 training images and 100 test images per class. (3) *ImageNet ILSVRC 2012* [19] contains 1.2 million training images and 50000 validation images in 1000 classes.

Implementation Details. We implemented the proposed PCL with a variety of backbone network architectures, including ResNet [5], VGG [20], DenseNet [9], WRN [26], and ResNeXt [24]. Following [13], the last block and the classifier layer of each backbone network were separated (on ImageNet, the last two blocks were separated), while the other low-level layers were shared. We set $m = 3$, so there are three peers in the multi-branch network. For fair comparison with the alternative methods, we applied standard random crop and horizontal flip for the random augmentation to generate counterparts of inputs, but other augmentation approaches [12] are applicable. We used SGD as the optimiser with Nesterov momentum 0.9 and weight decay $5e$-4. We trained the network for $Epoch_{max} = 300$ epochs on CIFAR-10/100 and 90 epochs on ImageNet. The initial learning rate was set to 0.1 and dropped to $\{0.01, 0.001\}$ at $\{150, 225\}$ epochs on CIFAR-10/100 and at $\{30, 60\}$ epochs on ImageNet. We empirically set the mini-batch size as 128, $T = 3$ to generate soft predictions, $\alpha = 80$ epochs for ramp-up weighting, $\beta = 0.999$ to learn temporal mean models, $\lambda = 1.0$ for CIFAR-10/100 and $\lambda = 0.1$ for ImageNet. We reported the average results with standard deviation over 3 runs.

4.2. Comparison with the State-of-the-Arts

Competitors. We compared the proposed PCL with the backbone network (Baseline) and five state-of-the-art online knowledge distillation methods (DML [28], CL [21], ONE [13], FFL-S [10], OKDDip [1]). For fair comparison, we employed three-branch networks (the low-level layers are shared) in ONE, CL, FFL-S, OKDDip and PCL, and used three parallel sub-networks in DML.

Results. As shown in Table 1 and Table 2, the proposed PCL improves the performance of various backbone networks (baseline) by approximately 1% and 2% on CIFAR-10 and CIFAR-100, respectively. This shows the effectiveness of PCL for improving the generalisation performance of various backbone networks in online distillation. On CIFAR-10 and CIFAR-100, PCL achieves the best top-1 error rates compared with the state-of-the-art online distillation methods. For example, on CIFAR-10, PCL improves the state-of-the-arts by approximately 0.13% and 0.34% with ResNet-32 and ResNet-110, respectively; Whilst on CIFAR-100, PCL improves the state-of-the-arts by about 0.65% and 1.15% with ResNet-32 and ResNet-110, respectively. These improvements attribute to the integration of the peer mean teacher and the online peer ensemble teacher into a unified framework. When extended to the large-scale ImageNet benchmark, as shown in Table 3, PCL improves the baseline by approximately 0.9% with ResNet-18. Compared with the state-of-the-art alternative methods, PCL still achieves the best top-1 error rate (29.58%±0.13% with ResNet-18), which verifies the effectiveness of PCL on

Table 3. Comparisons with the state-of-the-arts on ImageNet. Top-1 error rates (%).

Network	DML [28]	CL [21]	ONE [13]	FFL-S [10]	OKDDip [1]	Baseline	PCL(ours)
ResNet-18	30.18±0.08	29.96±0.05	29.82±0.13	31.15±0.07	30.07±0.06	30.49±0.14	**29.58±0.13**

Table 4. Component effectiveness evaluation with ResNet-110 on CIFAR-100. Top-1 error rates (%). P.E.: Peer Ensemble teacher. P.M.: Peer Mean teacher.

	Component	CIFAR-100
	Backbone	23.79±0.57
Proposed	Backbone+L_{ce}^{p}	23.56±0.50
Proposed	Backbone+L_{ce}^{p}+L_{ce}^{t}	23.48±0.99
Proposed	Backbone+L_{ce}^{p}+L_{ce}^{t}+L_{pe}	21.19±0.62
Proposed	Backbone+L_{ce}^{p}+L_{ce}^{t}+L_{pe}+L_{pm} (full model)	**20.02±0.55**
Variant	P.E. + Mutual Distillation [28]	21.09±0.18
Variant	P.E. + Traditional Mean Model [22]	26.80±0.35
Variant	ONE [13] + P.M.	20.43±0.71
Variant	P.E. + P.M. (full model)	**20.02±0.55**

the large-scale benchmark.

Discussion. These results validate the performance advantages of PCL for online knowledge distillation over the state-of-the-art alternatives. Besides, since we only use a temporal mean model of a peer for deployment, which has the identical number of parameters to the backbone network, our method doesn't require extra inference cost for deployment.

4.3. Component Effectiveness Evaluation

Table 4 shows the evaluation on the component effectiveness of PCL. We can observe that: (1) With all components, PCL (full model) achieves the best performance, which shows the effectiveness of integrating of peer ensemble teacher and peer mean teacher into a unified framework for online distillation. (2) Backbone+L_{ce}^{p}+L_{ce}^{t}+L_{pe} significantly improves the performance of the Backbone by approximately 2.6%, which verifies the effectiveness of the peer ensemble teacher. (3) PCL (full model) improves Backbone+L_{ce}^{p}+L_{ce}^{t}+L_{pe} by about 1.1%, which indicates the effectiveness of the peer mean teacher. (4) Replacing P.E. or P.M. with some contemporary variants leads to performance degradation, which demonstrates the superiority of the proposed PCL. Furthermore, from Fig. 3, we can observe that PCL with all components (red line) gets better generalisation capability. Interestingly, the test top-1 error rate (red line) of PCL (full model) drops rapidly during 0 to 50 epochs, and then gradually reaches to the optimal value; In contrast, other test lines fluctuate dramatically, especially during 0 to 225 epochs. This shows the importance of the peer mean teacher for alleviating the accumulation of error among peers to stabilise online knowledge distillation.

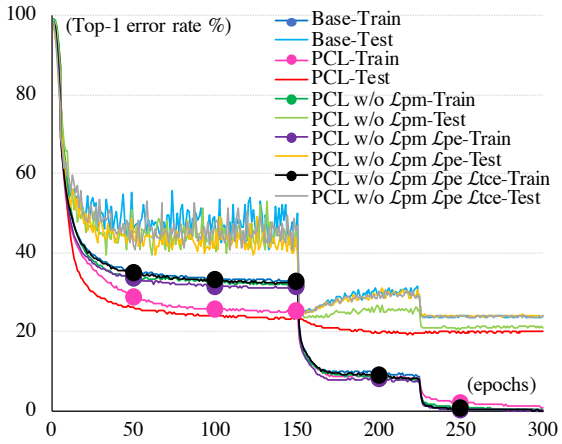

Figure 3. Component comparison with ResNet-110 during training and testing on CIFAR-100.

Table 5. Ensemble effectiveness evaluation with ResNet-110 on CIFAR-10/100. Top-1 error rates (%) and number of model parameters are reported.

Method	CIFAR-10		CIFAR-100	
	Top-1	Param.	Top-1	Param.
ONE-E	4.75±0.27	**2.89M**	20.10±0.24	**2.96M**
FFL (fused)	4.99±0.07	3.10M	21.78±0.28	3.19M
OKDDip-E	4.79±0.12	2.91M	20.93±0.57	2.98M
PCL-E(ours)	**4.42±0.12**	2.90M	**19.49±0.49**	3.04M

4.4. Ensemble Effectiveness Evaluation

We compare the proposed ensemble PCL-E with three alternative online ensembles: ONE-E [13], FFL (FFL-S with fused ensembles) [10], and OKDDip-E [1]. As shown in Table 5, PCL-E improves the state-of-the-arts by about 0.35% and 0.61% on CIFAR-10 and CIFAR-100, respectively. This validates the effectiveness of the proposed peer online ensembling and collaboration. Besides, from Table 5, we can see that compared with ONE-E, the alternative method with the fewest model parameters, although PCL-E achieves significantly better performance, it only increases the number of model parameters by 0.01M and 0.08M with ResNet-110 on CIFAR-10 and CIFAR-100, respectively.

4.5. Peer Variance for Online Ensembling Analysis

In Fig. 4, we further analyse the peer (branch) variance for online ensembling over the training epochs. Here, we compute the average Euclidean distance between the pre-

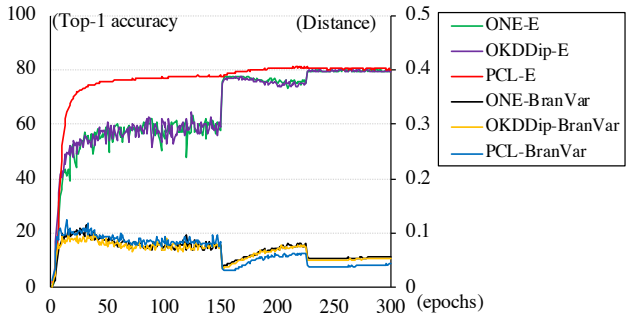

Figure 4. Peer variance for online ensembling analysis with ResNet-110 on CIFAR-100. '-BranVar': the branch variance. Here, we use top-1 accuracy for better visualisation.

Table 6. Comparison with two-stage distillation with ResNet-32 on CIFAR-10/100. Top-1 error rates (%). †: Use ResNet-110 as the teacher model.

Dataset	Baseline	KD†	PCL
CIFAR-10	6.74±0.15	5.82±0.12	5.67±0.12
CIFAR-100	28.72±0.19	26.23±0.21	25.86±0.16

dictions of two branches as the branch diversity and use the average diversity of m branches as the branch variance. From Fig. 4, we can observe that: (1) From 0 to 150 epochs, the top-1 accuracy of PCL-E soars to a high level outperforming the alternative methods, and meanwhile, the branch variance of PCL (PCL-BranVar) is larger than the alternatives. This indicates that at the early stage, although generalisation capability of the model is poor, each branch in PCL collaborates to facilitate online ensembling with richer knowledge. (2) From 150 to 300 epochs, the top-1 accuracy of PCL-E is still better than the alternatives, whilst the branch variance of PCL becomes smaller than the alternatives. The main reason is that at this stage, the generalisation capabilities of temporal mean models of peers (branches) have been improved to a high level with accurate and similar prediction. In other words, since the accumulation of training error among peers has been significantly alleviated, each branch becomes stable and gets close generalisation capability, resulting in a stronger ensemble model (as shown in Table 5) and a more generalised target model (as shown in Table 2).

4.6. Further Analysis and Discussion

Comparison with Two-Stage Distillation. We compare the PCL with the traditional two-stage knowledge distillation (KD) [6] in Table 6. We can see that although PCL doesn't pretrain a high-capacity teacher model (*e.g.* ResNet-110), it still achieves better performance than the two-stage KD. This attributes to the integration of online ensembling and network collaboration into a unified framework for online distillation.

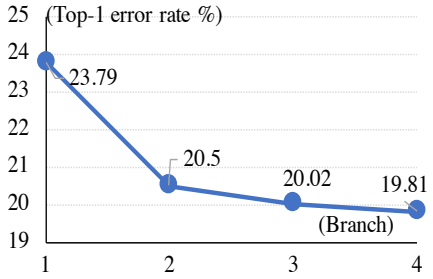

Figure 5. Evaluating PCL with different number of branches using ResNet-110 on CIFAR-100.

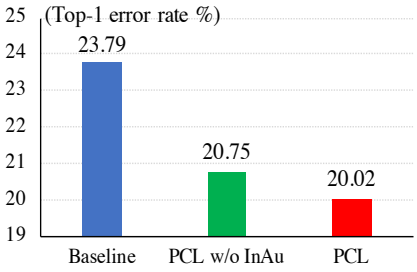

Figure 6. Evaluating the impact of multiple input augmentation for PCL with ResNet-110 on CIFAR-100.

Branch Number. As shown in Fig. 5, the performance of PCL improves when more peers/branches are exploited. Interestingly, the performance of PCL with two branches is already better than the state-of-the-art alternatives with three branches, which further shows the superiority of PCL.

Input Augmentation. As shown in Fig. 6, without using multiple input augmentation (PCL w/o InAu), the performance of PCL slightly decreases by about 0.5%, but it still achieves the state-of-the-art performance. This further verifies the effectiveness of the model formulation in PCL.

5. Conclusion

In this work, we propose a Peer Collaborative Learning (PCL) method for online knowledge distillation. It facilitates the collaboration among the peers in a multi-branch network by exploiting the peer feature concatenation as the high-capacity ensembling teacher and the peer temporal average models as the peer mean teachers. Doing so allows to improve the quality of online knowledge distillation in a one-stage end-to-end trainable fashion. Extensive experiments with a variety of backbone network architectures show the superiority of the proposed method over the state-of-the-art alternative methods on CIFAR-10, CIFAR-100 and ImageNet. In-depth ablation analyses further verify the effectiveness of the components in the proposed PCL.

References

[1] 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*, 2020.

[2] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE International Conference on Computer Vision Workshop*, pages 2590–2600, 2017.

[3] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020.

[4] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[5] 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.

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

[7] Saihui Hou, Xu Liu, and Zilei Wang. Dualnet: Learn complementary features for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 502–510, 2017.

[8] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. In *International Conference on Learning Representations*, 2017.

[9] 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.

[10] Jangho Kim, Minsung Hyun, Inseop Chung, and Nojun Kwak. Feature fusion for online mutual knowledge distillation. *arXiv preprint arXiv:1904.09058*, 2019.

[11] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

[12] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

[13] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems*, pages 7517–7527, 2018.

[14] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[15] Mohammad Moghimi, Serge J Belongie, Mohammad J Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li. Boosted convolutional neural networks. In *British Machine Vision Conference*, 2016.

[16] 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.

[17] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151, 2019.

[18] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

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

[22] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017.

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

[24] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[25] 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.

[26] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

[27] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

[28] 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.

[29] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 137(1-2):239–263, 2002.