Research on the Theory and Application of Deep Interactive Learning

Ziyuan Wang *, Fan Guo
China Agricultural University, College of Information and Electrical Engineering, Beijing, China.
* 2018308160237@cau.edu.com

Abstract. Knowledge distillation (KD), in which a small network (students) is trained to mimic a larger one(teachers), with high precision, has been widely used in various fields. However, the interaction between teachers and students is still weak. It is found in this study that most existing methods, such as Deep Mutual Learning (DML), mainly construct loss function through soft weight indexes. Few researchers pay attention to the sharing of hard and heavy ones. As an improvement of DML, a new online learning distillation method, namely, Deep Interactive Learning (hereinafter DIL), was proposed in this research, which has deeper interaction than DML. We not only output the features of layers, but also disclose the features of hidden layers. We transfer the features to other models to obtain the corresponding softer distribution or features for distillation. Extensive experiments on various data sets show that the accuracy of our method is improved by almost 3% in CIFAR and 2% in ImageNet, which proves the validity of our method.

Keywords: knowledge distillation, deep mutual learning, deep interactive learning, online distillation

1. Introduction
In recent years, the deep learning model has been widely used in many fields, and has a good performance in solving complex problems. However, these models usually need huge memory space and powerful computing power support, so it is difficult to lay out the models on edge devices with limited resources. Knowledge distillation is considered as an effective method to solve this problem.

Knowledge distillation [1] is a famous technology for learning compact deep neural network model with competitive accuracy, in which smaller networks (students) are trained to mimic the larger networks (teachers) with higher accuracy. The popularity of knowledge distillation is mainly due to its simplicity and universality. The student model based on teacher learning is straightforward, and there are no restrictions on the network architecture of these two models. Now most scholars mainly study how to transfer knowledge from a trained teacher model to a student model more effectively. Such as Parameter quantization or binarization [2], pruning [3] [4] [5] and knowledge distillation [1] are representative methods in this research field. As a solution, KD has always been an active research field, which improves the capacity of lightweight networks by taking the knowledge of large pre-training networks (or small network sets) as teachers' networks. Knowledge distillation is usually divided into
two different methods: online distillation and offline distillation. Offline KD method usually requires a high-capacity teacher model has been well trained to perform unidirectional transmission [6] [7] [8] [9] [10]. However, training a "good" teacher has become the main problem of offline distillation. The performance of student network will decrease when the difference between teachers' and students' capacity is huge. Besides, many parameters and the high calculation cost have been required in the two-stage KD. In order to overcome these difficulties, some works pay more attention to online KD, which trains a group of students at the same time by learning their peers' predictions. In the field of image classification, Deep mutual learning (DML) [11] and on-the-fly native ensemble (one) [12] are typical on-line distillation methods and have achieved good model results. However, in online distillation, the specific roles of teachers and students are diluted. Just like students learn from each other in class, instead of just getting knowledge through teachers. Training from the model, all neural networks learn from each other online through mutual transmission of knowledge. It uses the traditional cross entropy as loss function to improve the network performance. And it also imitates the loss from the companions. The network trained by this online distillation method is not only better than the cross entropy loss network trained by itself, but also better than the network trained by the traditional offline distillation method from the pre-trained teacher network.

Although these methods are helpful to improve the generalization ability of the target network, they can only mine limited knowledge like output results, and cannot provide teachers with stronger ability/build stronger network with teachers being more capable to further improve students. The output results are like formulas in textbooks, which are difficult for students to understand. By analogy to our real world, in order to make students have a better understanding of what they have learned, teachers often teach students the gradual derivation of formulas. The characteristics of the middle layer of the teacher network are compared with the gradual derivation of the formula. Therefore, we try to transfer the middle layer characteristics of the teacher network to the student network. In order to realize our idea, we propose an improved online distillation method called deep interactive learning.

Compared with DML, DIL method enhances the depth of interaction, not only in the output layer, but also in the middle layer, as shown in Figure 1. DML can't correct the accumulated errors of network stratification in training middle school students, but DIL can solve this problem well, thus forming a stronger and more accurate student network. Specifically, the purpose of guiding the hidden layer of teachers to the student network is to correct the errors accumulated in the previous layers of the student network. At the same time, in order to make teachers aware of the errors accumulated on the students' network, we reverse the strategy by guiding the hidden layer of students to the teachers' network.

Sufficient experiments show that the proposed DIL has achieved consistent and significant accuracy improvement in various neural networks and data sets. Experiments on ten neural networks on five data sets show that DIL is much better than the most advanced distillation method in both image classifications. On average, the accuracy improvement of 5.46%, 1.71%, 1.18%, 1.25% and 0.82% can be observed on CIFAR100, CIFAR10 and ImageNet datasets. In addition, ablation studies and hyperparametric sensitivity studies were conducted to show the effectiveness and stability of DIL.

![Fig. 1 Comparison between DML and DIL.](image-url)
2. Theoretical Basis of the Research

2.1. Knowledge Distillation (KD)
Knowledge distillation is defined as refining a larger neural network model into a smaller model [1]. The larger neural network is like a teacher network, while the smaller neural network can be regarded as a student network. In addition to probability distribution, other studies also try to extract various features to students. In training the model efficiency of a small student network, KD can achieve the same effect as some model compression methods, such as pruning [2] and quantification. Knowledge distillation is like a teacher instilling knowledge into students without getting feedback from students, which is obviously not conducive to students' learning. Therefore, some scholars have proposed a method of Deep Mutual Learning (DML) [11]. By interacting the characteristics of the output layer of the teacher network with the student network, kullback-leibler is used to measure the degree of interaction. In this framework, the boundary between each student network and the teacher network is no longer obvious, and students are more inclined to learn from each other. One advantage of this method is that it can flexibly apply any different network architecture. But the fly in the ointment is that this method can only exchange limited information, because it does not make full use of the rich information of teacher model. Another classical online distillation method is the on-the-fly native ensemble (ONE). It focus on improving the performance of student network using the gate of branch logic. But trying to transfer the knowledge from teacher model to student model is focal point of this method. The shortcoming of the ONE is that it can only train a single architecture because of the branch gate logic. Our method is based on DML, and some improvements have been made.

2.2. Parameters Sharing
Parameters Sharing can be regarded as a method to regularize parameters by requiring a set of parameters to be shared among multiple networks [17]. It is a common practice to improve the accuracy while keeping the model size, which is widely used in neural architecture search (NAS) and multi-task learning (MTL). For MTL with convolutional network, several related tasks are jointly optimized, and some feature representations (early layers in deep ConvNet) are shared among these tasks; For example Action recognition and attitude estimation, target detection and instance segmentation, segmentation and surface normal estimation. Therefore, the network can be divided into two parts according to its parameters: the shared network with general parameters (generally the bottom layer) and the task-specific network with individual parameters (generally the top layer). By using the training examples of several tasks, the shared parameters are more regularized, which often leads to better generalization. Similar to multi-task learning, the underlying parameters are shared among multiple head classifiers integrated in our group. However, these classifiers focus on the same task instead of MTL, and they focus on different tasks. Although trained on the same task, parameter sharing adds an extra regularization for sharing basic parameters. The method in this study extends the idea of parameter sharing, and we share parameters between teachers' network and students. [21]

3. Research Method

3.1. KD and DML
Review is first done in this part on the formulas of KD [1] and Deep Mutual Learning DML [11]. To simplify the instructions, we only consider the very basic situation of a single teacher model and a single student model. We selected n images as our training data \(X = \{x_n\}_{n=1}^N\), and then we used \(Y = \{y_n\}_{n=1}^N\) to represent the ground truth labels of these samples. Then, in the process of knowledge distillation, we need a strong teacher network \(W_t\) that has been trained and a student network \(W_s\) that is lighter than the teacher network. In KD, the loss function of student network is defined as
Figure 2. Deep Interactive learning: The hidden layers of each student and teacher network are interconnected. DIL loss represents the loss function we put forward in Equation 7. DML loss represents the loss function of deep mutual learning in formula 5 [11].

\[ L_S = L_c(W_S, X, Y) + \lambda L_{kd}(\hat{P}_t, \hat{P}_s) \]  

(1)

\( L_s \) request to be trained as small as possible. And \( L_c \) is the cross entropy loss of output values from the student model and the correct answer we given in advance. We use \( L_{kd} \) to represent the distillation loss, and \( \lambda \) is a weight parameters to balance these two loss items. In [14], \( L_{kd} \) is defined as

\[ L_{kd}(\hat{P}_t, \hat{P}_s) = \frac{1}{N} \sum_n \sum_{m=1}^M \hat{P}_t^m(x_n) \log \hat{P}_s^m(x_n) \]  

(2)

As an improved distillation loss, \( L_{kd} \) is no longer the cross entropy loss between the ground truth label and the student network, but the former ground truth label is replaced by the trained teacher network output.

It defined that the probabilities of a sample \( n \) belongs to the classified \( m \) as \( \hat{P}^m(x_n) \)

\[ \hat{P}^m(x_n) = \frac{\exp(z_n^m/T)}{\sum_{m=1}^M \exp(z_n^m/T)} \]  

(3)

In this formula, \( z_n^m \) is the logarithm of the output result of the neural network model in the last layer of the model, and the function of \( T \) is the probability that the output will soften to 0 to 1.

The knowledge distillation (KD) introduced above is obviously an off-line distillation method. It uses the pre-trained teacher network model to train a more lightweight student network model. This is a one-way knowledge transfer. Generally speaking, there are two keys to KD, one is the representation of knowledge, and the other is the transfer strategy of knowledge. In DML, there is no difference between teacher model and student model, and all models are untrained student models. In order to improve the latter, DML proposed a bidirectional knowledge transfer method. It is no longer one-way instilling a teacher model into student models, but optimizing the network by encouraging students to learn from each other. In the following formula, we use subscripts \( s_1 \) and \( s_2 \) to represent the first and second student networks respectively:

\[ L_{s_1} = L_c(W_{s_1}, X, Y) + \lambda L_{dml}(\hat{P}_{s_1}, \hat{P}_{s_2}) \]  

(4)
Similar to KD, the Loss function of two student networks is defined in the form of Formula 4. In order to simplify the operation, we fixed $\lambda$ as 1. Compared with Formula 1, we redefined the distillation loss function $L_{kd}$ and changed it to $L_{dml}$. In DML, Kullback-Leibler(KL) divergence is used to define the output matching of two student networks.

$\hat{P}_i = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} \hat{P}_i^m(x_n) \log \frac{\hat{P}_i^m(x_n)}{\tilde{P}_i^m(x_n)}$ (5)

$\hat{P}_i^m(x_n)$ is also defined by formula 3. KL divergence has the same effect as cross entropy from the angle of gradient calculation. DML is an online knowledge distillation method. Compared with KD, the biggest difference of DML is that DML trains two student networks together, while KD needs to separate the two networks into two different training stages. DML only interacts with the output layer of the network, while ignoring a large amount of information contained in the hidden layer. The DIL method proposed by us has been improved.

### 3.2. Deep Interactive Learning(DIL)

The goal of this study is to cultivate two untrained networks so as to further realize a compact student network. In order to distill features in the hidden layers of teacher model, we propose deep interactive learning, as shown in Figure 2. This model can transfer the hidden features of teacher-student model to other models, so that it can obtain corresponding softer distribution or characteristics for distillation. To facilitate the comparison with DML, two untrained models, still, are defined as two student network $F^{s1}$ and $F^{s2}$. The l-convolution feature map of their corresponding network is expressed $F^{s1}$ as $h_i^{s1} = \sigma(W_i^{s1} * h_{\cdot i}^{s1})$, where $\sigma(\cdot)$ is the activation function and $*$ is the convolution operation. Here, a new approach, to balance $L_c$ and $L_i$, and to make the hard connection between $h_{s1}$ and $h_{s2}$ softer has been presented by us. We draw a diagrammatic sketch which showed in Figure 1(b).In this paper, the new feature graphs $\hat{h}_{s1}$ and $\hat{h}_{s2}$ whose $h_{s1}$ and $h_{s2}$ are redefined by cross connection (convex connection) are defined as,

$$\begin{bmatrix} \hat{h}_{s1} \\ \hat{h}_{s2} \end{bmatrix} = \begin{bmatrix} \alpha \\ 1-\alpha \end{bmatrix} \begin{bmatrix} h_{s1} \\ h_{s2} \end{bmatrix} + \begin{bmatrix} (1-\beta) \\ \beta \end{bmatrix} \begin{bmatrix} \hat{h}_{s1} \\ \hat{h}_{s2} \end{bmatrix}$$ (6)

Where $\alpha, \beta \in [0,1]$ are hyper-parameters, and different amount of percentage is adjusted for cross connection. Convex combination ensures that the input norm is almost the same after cross connection (assuming $\|h_{s1}\| \approx \|h_{s2}\|$), so the parameter size remains unchanged. We define the loss of deep interactive learning as

$\hat{L}(W^{s2}) = \|\sigma(W^{s1} * \hat{h}^{s1}) - \sigma(W^{s2} * \hat{h}^{s2})\|_F^2$ (7)

$\hat{L}(W^{s1}) = \|\sigma(W^{s2} * \hat{h}^{s2}) - \sigma(W^{s1} * \hat{h}^{s1})\|_F^2$ (8)
Algorithm 1: The DIL algorithm
Input :Training data \{X,Y\}, two CNN models \(F^{s_1}\) and \(F^{s_2}\), learning rate \(\gamma_i\)
Initialise \(F^{s_1}\) and \(F^{s_2}\)
Repeat
\(i \leftarrow i + 1\), update \(\gamma_i\);
1. Randomly sample a batch of data from \{X,Y\};
2. Get the feature map \(\hat{h}_{s_1}\) and \(\hat{h}_{s_2}\) by Eq.6
3. Compute loss \(L_{DIL}\) by Eq.7
4. Calculate gradients and update parameters
Until Converge;

4. Experiments

4.1. Experiment Settings
Image classification uses ResNet, PreActResNet, SENET, ResNeXt, MobileNetV1, MobileNetV2, ShuffleNetV1, ShuffleNetV2, WideResNet and two data sets, including CIFAR10 and ImageNet. In CIFAR10 experiment, each model is trained by SGD optimizer for 300 epochs, and the batch size is 128. In ImageNet experiment, each model is trained by SGD optimizer for 90 epochs, and the batch size is 256.

4.2. Comparative experiment
In order to explain our experimental results more powerfully, we compare four knowledge distillation methods with DIL, including KD[1], FitNet, DML[11] and self-distillation (SD). We finished all these experiments by ourselves.

Table 1. Experimental results of 1: CIFAR10 (TOP-1 precision/%). Numbers in bold are the highest.

| Model                 | Baseline | KD     | FitNet | DML    | SD    | DIL    |
|-----------------------|----------|--------|--------|--------|-------|--------|
| ResNet18              | 94.25    | 94.67  | 95.57  | 95.19  | 95.87 | **96.92** |
| ResNet50              | 94.69    | 94.56  | 95.83  | 95.73  | 96.01 | **96.84** |
| PreActResNet18        | 94.20    | 93.74  | 95.22  | 94.80  | 95.08 | **96.49** |
| PreActResNet50        | 94.39    | 93.53  | 94.98  | 95.87  | 95.82 | **96.93** |
| SEResNet18            | 94.78    | 94.53  | 95.64  | 95.37  | 95.51 | **96.80** |
| SEResNet50            | 94.83    | 94.80  | 95.31  | 94.83  | 95.45 | **97.02** |
| ResNeXt50-4           | 94.49    | 95.41  | 95.78  | 95.41  | 96.01 | **97.09** |
| MobileNetV1           | 90.16    | 91.70  | 90.53  | 91.65  | 91.98 | **93.93** |
| MobileNetV2           | 90.43    | 92.86  | 90.49  | 90.49  | 91.02 | **93.34** |
| ShuffleNetV1          | 91.33    | 92.57  | 92.23  | 91.40  | 92.47 | **92.73** |
| ShuffleNetV2          | 90.88    | 92.42  | 91.83  | 91.87  | 92.51 | **93.47** |

Table 2. Experimental results on 2: ImageNet (TOP-1 precision/%).

| Model                 | Baseline | DIL   | MAC(G) | Param(M) |
|-----------------------|----------|-------|--------|----------|
| ResNet18              | 69.76    | **70.92** | 1.82   | 11.69    |
| ResNet50              | 76.13    | **77.52** | 4.11   | 25.56    |
| ResNet101             | 77.37    | **78.64** | 7.83   | 44.55    |
| ResNet152             | 78.31    | **79.21** | 11.56  | 60.19    |
| ResNeXt50-32-4        | 77.62    | **78.93** | 4.26   | 25.03    |
| WideResNet50-2        | 78.47    | **79.52** | 11.43  | 68.88    |
4.3. Results of Cifar10
The accuracy of our training results on CIFAR10 is shown in table 1. We observe that: (a) Compared with the basic model, the proposed DIL leads to a significant improvement in accuracy. In CIFAR10, on average, 2.49% improvement in precision can be found on 11 models, ranging from the largest 3.77% to the smallest 1.40%. (b) In all models, the proposed DIL is superior to the second best distillation method to a great extent. On average, 3.13% and 1.28% improvement in precision can be observed on CIFAR10 compared with the second best distillation method. (c) The proposed DIL is not only applicable to models with too high parameters, such as ResNet and SENet, but also shows remarkable effectiveness in MobileNet and MobileNet in lightweight models. On average, 2.74% accuracy improvement of lightweight model can be observed on CIFAR10 dataset.

4.4. Results of ImageNet
Table 4 shows the experimental results of DIL on ImageNet. In all these experiments, ResNet152 model is used as the teacher model. We observe that (a) on average, the accuracy of DIL in six neural networks is improved by 1.18. (b) the accuracy of ResNet50 and ResNet101 after distillation is higher than the baseline of ResNet101 and ResNet152, respectively. By replacing the distilled ResNet50 and ResNet101 with ResNet101 and ResNet152, DIL achieved 1.57 times compression and 1.81 acceleration, without precision loss.

5. Conclusion
In this paper, we propose an online distillation method called Deep Interactive Learning (DIL) was proposed, which transfers the characteristics of output layer and hidden layer to another model. Through the stronger network interaction between teachers and students, our method has higher accuracy. The model put forward in this study was operated in CIFAR database and ImageNet database, and their accuracy is improved by 3% and 2% respectively. The experimental results show that this method is superior to other competitive basic models.

References
[1] Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
[2] Courbariaux, M., Bengio, Y. & David, J. P. (2015). Binaryconnect: Training deep neural networks with binary weights during propagations. In: NeurIPS.
[3] Wu, J., Leng, C., Wang, Y., Hu, Q. & Cheng, J. (2016). Quantized convolutional neural networks for mobile devices. In: CVPR.
[4] Chebotar, Y. & Waters, A. (2016). Distilling knowledge from ensembles of neural networks for speech recognition. In: Interspeech.
[5] Sindhwani, V., Sainath, T. & Kumar, S. (2015). Structured transforms for small-footprint deep learning. In: NeurIPS.
[6] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, and Neil D. Lawrence. Variational information distillation for knowledge transfer. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
[7] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In Interspeech, pages 3697–3701, 2017.
[8] Liang Gao, Xu Lan, Haibo Mi, Dawei Feng, Kele Xu, and Yuxing Peng. Multistructure-based collaborative online distillation. Entropy, 21(4):357, 2019.
[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.
[10] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In Advances in Neural Information Processing Systems, pages 2760–2769, 2018.
[11] Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR (2018)
[12] Lan, X., Zhu, X., and Gong, S. Knowledge Distillation by On-the-Fly Native Ensemble. In NeurIPS, 2018.
[13] Tuong Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, and Quang D Tran. Compact trilinear interaction for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, pages 392–401, 2019.
[14] Jong-Chyi Su and Subhransu Maji. Adapting models to signal degradation using distillation. arXiv preprint arXiv:1604.00433, 2016.
[15] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8427–8436, 2018.
[16] Wu, J., Leng, C., Wang, Y., Hu, Q. & Cheng, J. (2016). Quantized convolutional neural networks for mobile devices. In: CVPR.
[17] Zhang, X., Zhou, X., Lin, M. & Sun, J. (2018a). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: CVPR.
[18] Sindhwani, V., Sainath, T. & Kumar, S. (2015). Structured transforms for small-footprint deep learning. In: NeurIPS.
[19] Han, S., Pool, J., Tran, J. & Dally, W. (2015). Learning both weights and connections for efficient neural network. In: NeurIPS.
[20] Wang, Y., Xu, C., Xu, C. & Tao, D. (2019e). Packing convolutional neural networks in the frequency domain. IEEE TPAMI 41(10),2495–2510.
[21] [online] Available: ] https://arxiv.org/. Cornell University, the Simons Foundation.