Dense Feature Fusion for Online Mutual Knowledge Distillation

Dongsheng Ni
School of Beihang University, Beijing 100000, China
18373225@buaa.edu.cn

Abstract. Feature maps contain rich information about image intensity and spatial correlation. However, previous online knowledge distillation methods only utilize the class probabilities, ignoring the middle-level supervision, resulting in low efficiency in training many models. Even if some methods have joined the middle-level supervision, the previous way is to define the characteristic loss, and the effect is general. We propose a new method of middle-level supervision, through the fusion of features between teacher academic network to enter the supervision. The specific method is to fuse the features of the teacher academic network, and establish an auxiliary fusion branch to process the fusion information, so that the feature fusion can effectively strengthen the feature interaction of the teacher academic network. At the same time, we added the normalized integrated distillation of the output, and our method reached the SOTA in online KD. We have done a lot of experiments on cifar-10, cifar-100 and ImageNet datasets, and proved that this method is more effective than other methods in the performance of sub network and fusion classifier, as well as generating meaningful feature maps.

Keywords: Online Mutual Knowledge Distillation, Feature Fusion.

1. Introduction
With the advent of Alexnet [1], deep convolution neural networks have achieved remarkable success in a variety of tasks. However, high-performance of deep neural network is often gained by increasing the depth or the width of a network. Deep and wide networks cost a large number of computation as well as memory storage which is not suitable for a resource-limited environment such as mobile or embedded systems. To overcome this issue, many researches have been conducted to develop smaller but more accurate neural networks. Some of the well-known methods in this line of research are parameter quantization or binarization [2], pruning [3] and knowledge distillation (KD) [4].

KD has been an active area of research as a solution to improve the performance of a light-weight network by transferring the knowledge of a large pre-trained network (or an ensemble of small networks) as a teacher network. KD sets the teacher network’s class probabilities as a target which a small student network tries to mimic. By aligning the student’s predictions to those of the teacher, the student can improve its performance. Although traditional knowledge distillation has shown good performance in compressing models for deployment, they usually follow a two-stage training solution, that is, pre training a high-quality teacher model to transfer knowledge to a compact student model,
which requires more training time and computational cost. Recently, some studies have shown that rather than using a pre-trained teacher, simultaneously training networks to learn from each other in a peer-teaching manner is also possible. This approach is called online distillation. Deep mutual learning (DML) [5] and on-the-fly native ensemble (ONE) [6] are the representative online distillation methods that show appealing results in the image classification tasks. Conventional distillation method requires pre-training a powerful teacher network and performs an one-way transfer to a relatively small and untrained student network. On the other hand, in online mutual distillation, there is no specific teacher-student role. All the networks learn simultaneously by teaching each other from the start of training. It trains with the conventional cross-entropy loss from the ground truth label along with the mimicry loss to learn from its peers. Networks trained in such an online distillation way achieves results superior not only to the networks trained with the cross-entropy loss alone but also to those trained in a conventional offline distillation manner from a pre-trained teacher network.

Online knowledge distillation: traditional offline method needs to train a teacher model in advance, while online method does not need any model in advance. Instead, these networks teach each other by sharing knowledge throughout the training process. Some recent examples of online distillation methods are DML [5] and one [6], which show promising results.

Feature maps contain rich information about image intensity and spatial correlation. However, previous online knowledge distillation methods only utilize the class probabilities. aforementioned online distillation methods make use of only the logit information. While the logit contains the probabilistic information over classes, the feature map, the output of convolution layer, has more meaningful and abundant feature information on image intensity and spatial correlation. In offline distillation which utilizes a pre-trained model as a teacher network, many methods such as FitNet [8], attention transfer (AT) [9] and factor transfer (FT) [10] make use of this intermediate feature representation as a target to learn for the student network. Unlike the offline methods that have a clear target to mimic, there is no static target to follow in an online method. At every training iteration, the feature maps of the co-trained network change, thus in online feature map-level distillation, the problem turns into mimicking the moving target properly. While each node of the logit is confined to represent its assigned class probability which does not change drastically over iterations, at the feature map-level, much more flexibility comes into play, which makes the problem more challenging. Therefore, the direct aligning method such as using L1 or L2 distance is not suitable for online mutual feature map distillation because it updates the network parameters to generate a feature map trying to mimic the current output feature map of the other network. In other words, the direct alignment method only tries to minimize the distance between the two feature map points (one for each network), hence it ignores the distributional difference between the two feature maps.

DML simply applies KD losses mutually, treating each other as teachers, and it achieves results that is even better than the offline KD method. The drawback of DML is that it lacks an appropriate teacher role, hence provides only limited information to each network. ONE pointed out this defect of DML. Rather than mutually distilling between the networks, ONE generates a gated ensemble logit of the training networks and uses it as a target to align for each network. ONE tries to create a powerful teacher logit that can provide more generalized information. The flaw of ONE is that it can not train different network architectures at the same time due to its architecture of sharing the low-level layers for the gating module. The common limitation of existing online methods is that they are dependent only on the logit and do not make any use of the feature map information. Considering that KD loss term is only applicable to the classification task, transferring knowledge at feature map-level can enlarge the applicability to other tasks. Therefore, our method proposes a distillation method that utilizes not only the logit but also the feature map via adversarial training, moreover, our method can be applied in case where the co-trained networks have different architectures.

To alleviate this problem, in this paper, we propose a novel online distillation method which is called Dense Feature Fusion for Online Mutual Knowledge Distillation. The specific method is to fuse the teacher student features, and establish an auxiliary fusion branch to process the fusion information. In this way, the feature fusion can effectively strengthen the feature interaction of the teacher
academic network. Coupled with the normalized integration distillation of the output, our method is called SOTA in online KD.

Extensive experiments on CIFAR-10 [11], CIFAR-100 [11] and ImageNet [12] show that the proposed method significantly improves the effect of online distillation, and a large amount of effective information contained in the middle layer plays an important role through the feature fusion of the middle layer.

2. Related work

2.1. Traditional Knowledge Distillation

Traditional Knowledge Distillation [4] is one of the most effective solutions to compress a cumbersome model or an ensemble of models into a smaller model. The rationale behind is taking advantage of extra supervision provided by the teacher model during training the target model, beyond a conventional supervised learning objective such as the cross-entropy loss subject to the training data labels. In [4], Hinton 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. Although knowledge distillation shows good performance in compressing models for deployment, it usually follows a two-stage training solution, that is, pre training a high-quality teacher model to transfer knowledge to a compact student model, which requires more training time and computational cost. Besides probability distribution, some other researches have tried to distill various features to the student [11, 13-17]. In terms of training a small student network for the model efficiency, KD is also considered as one of model compression methods such as pruning and quantization. However, the traditional knowledge distillation also has the following disadvantages: it needs to train a strong teacher model in advance, and only uses the information of the output.

2.2. Online knowledge distillation

Traditional offline method needs to train a teacher model in advance, while online method does not need any pre trained model. Through online KT, the networks share their own knowledge, and imitate the performance of peer-to-peer network in the training process to teach each other. Deep mutual learning (DML) [5] and dynamic native integration (one) [6] are two representative online knowledge distillation methods. They show very promising performance. The disadvantage of DML is that it lacks a proper teacher role, so it can only provide limited information to each network. This defect of DML has been pointed out. Instead of extracting each other between networks, it is better to generate the gating set logit of the training network and use it as the alignment target of each network. An attempt to create a strong teacher logic can provide a wider range of information. One is that the gating module shares the underlying structure and can not train different network structures at the same time. The common limitation of existing online methods is that they only rely on logit and do not use any feature map information. Considering that KD loss item is only applicable to classification task, knowledge transfer in feature mapping layer can expand the applicability to other tasks. Therefore, we propose an extraction method, which not only uses logit, but also extracts the network features of the middle layer for fusion to form new branches to improve the performance of each sub network.

3. Proposed method

The overall process of this method is called Dense Feature Fusion for Online Mutual Knowledge Distillation. The sub network fuses the two networks from the middle level through feature fusion and creates a fusion classifier. At the same time, the sub network integrates the classifiers of the two networks into an integrated classifier to train each sub network. Then, the ensemble classifier transfers its knowledge to each sub network. At the same time, the fusion module transfers its knowledge back to each sub network. This kind of dense feature fusion of online mutual knowledge distillation helps to
get better performance gain in the fusion classifier and sub network, and also makes effective use of the rich information of the middle network.

**Figure 1.** The overall process of our method is called Dense Feature Fusion for Online Mutual Knowledge Distillation (DFF). The sub-networks and the fusion classifier create an ensemble classifier for training the fusion module. Then, the ensemble classifier transfers its knowledge to the fusion module and sub-networks. This online mutual knowledge distillation helps to obtain better performance gain in the fused classifier as well as the sub-networks.

### 3.1. Fusion Feature

Different from DualNet [18], our method does not use simple sum or average operation in feature fusion. Instead, we connect the features of the sub-networks and then perform convolution operations through the fusion module. In order to reduce the number of parameters, we use a simple depth convolution and a 1 × 1 convolution, called pointwise convolution, which has been used in mobilenet [19]. We start with the feature map of the middle layer, because the feature map of the middle layer contains a lot of effective information, and the size is smaller than the initial layer, so the calculation is less. Let C1 and C2 be the channel number of the feature map in the last layer of network 1 and network 2 respectively, then the channel number m from the cascaded feature map will be C1 + C2.

**Figure 2.** The number of output channels n from the fusion module can be operated as needed as shown in Figure 2, we first perform 3 × 3 deep direction convolution, apply a filter to each input channel, and then apply point direction convolution to create a linear combination of feature map slices, so as to combine them well.

In this work, we deal with the multi class classification task suppose there are m classes, then the logarithm of the kth network forwarding is defined as

\[
z_k = \left\{ z_k^1, z_k^2, ..., z_k^m \right\}
\]
In the process of training, we use softening probability to generalize the model. Given $z_k$, the softening probability is defined as

$$\sigma_l(z_k; T) = \frac{e^{z_k/T}}{\sum_i^m e^{z_i/T}}$$

(2)

When $t = 1$, it is the same as the original softmax. If the true value is given as

$$y = \{y^1, y^2, \ldots, y^m\}$$

(3)

The cross entropy loss of the $N$ network is defined as

$$g_{ce}^k = -\sum_{i=1}^m y^{(i)} \log(\sigma_l(z_k; 1))$$

(4)

In order to make full use of the given knowledge, we make a powerful classification of sets. A set classifier is created by a logit set in the subnet to train the fusion module. Assuming that there are $n$ subnets, the set of Logit is calculated as follows:

$$z_e = \frac{1}{n} \sum_{k=1}^n z_k$$

(5)

In order to train the fusion module, the ensemble classifier extracts its knowledge into the subnet classifier. This is called integrated knowledge distillation (EKD). EKD loss is defined as the KL divergence between the softening distribution of the whole classifier and that of the fusion classifier. If the logit in the fusion classifier is expressed as $Z_F$, the loss of EKD is as follows:

$$g_{kl}^e = \sum_{l=1}^m \sigma_l(z_e; T) \log(\frac{\sigma_l(z_e; T)}{\sigma_l(z_f; T)})$$

(6)

Compared with other methods, this EKD loss can help the training fusion module to generate meaningful feature maps. The feature graph from the last layer of the subnet is connected and put into the fusion module. In order to train each subnet, the fusion classifier in the fusion module extracts its knowledge into each subnet. This is called fusion knowledge distillation (FKD). The FKD loss is defined by distilling the softening distribution of the fusion classifier into each subnet

$$g_{kl}^f = \sum_{k=1}^n \sum_{l=1}^m \sigma_l(z_f; T) \log(\frac{\sigma_l(z_f; T)}{\sigma_l(z_k; T)})$$

(7)

In addition to distillation loss, each sub network and fusion classifier also learn the real label through cross entropy, and the total loss becomes:

$$g_{total} = \sum_{k=1}^n \left( g_{ce}^k + g_{ce}^f + T^2 \times (g_{kl}^e + g_{kl}^f) \right)$$

(8)

In our DFF, each sub-network and the fused classifier learns through ground-truth with cross-entropy loss. At the same time, the ensemble classifier distills its knowledge to the fused classifier and the subnet classifier with $g_{kl}^e$ and in return. Through such mutual knowledge distillation (MKD), the fusion module generates meaningful features for classification. Since the scale of the gradient produced by the softened distribution is $1/T^2$, we multiply $T^2$ according to the recommendations of [4]. Subnetworks and the fusion module in FFL are trained simultaneously.
4. Experiments

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 [12] contains 1.2 million training images and 50000 validation images in 1000 classes.

Implementation Details. We implemented the proposed DFF with a variety of backbone network architectures, including ResNet [21], VGG [22], DenseNet [23], WRN [24], and ResNeXt [20]. Following [25], 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 [26] are applicable. We used SGD as the optimiser with Nesterov momentum 0.9 and weight decay $5 \times 10^{-4}$. We trained the network for $\text{Epochmax} = 300$ 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 DFF with the backbone network (Baseline) and five state-of-the-art online knowledge distillation methods (DML [5], CL [27], ONE [6], DFF-S [14], OKDDip [28]). For fair comparison, we employed three-branch networks (the low-level layers are shared) in ONE, CL, DFF-S, OKDDip and DFF, and used three parallel sub-networks in DML. Results. As shown in Table 1 and Table 2, the proposed DFF improves the performance of various backbone networks (baseline) by approximately 1% and 2% on CIFAR10 and CIFAR-100, respectively. This shows the effectiveness of DFF for improving the generalisation performance of various backbone networks in online distillation. On CIFAR-10 and CIFAR-100, DFF achieves the best top-1 error rates compared with the state-of-the-art online distillation methods. For example, on CIFAR-10, DFF improves the state-of-the-arts by approximately 0.13% and 0.34% with ResNet-32 and ResNet-110, respectively; Whilst on CIFAR-100, DFF 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, DFF improves the baseline by approximately 0.9% with ResNet18. Compared with the state-of-the-art alternative methods, DFF still achieves the best top-1 error rate (29.58%±0.13% with ResNet-18), which verifies the effectiveness of DFF on the large-scale benchmark. Discussion. These results validate the performance advantages of DFF 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.
Table 1. Comparisons with the state-of-the-arts on CI FAR-100. Top-1 error rates (%).

| Network      | DML    | CL      | ONE     | FFL-S   | OKDDip  | Baseline | DFF     |
|--------------|--------|---------|---------|---------|---------|----------|---------|
| 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-29-2x64d | 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 |

4.3. Visualization

Figure 3. Using the latest method DFF and one and the original model of peer-to-peer network using the same architecture resnet32, we compare the grad cam [29] visualization effect of the proposed DFF. The label under each heat map is the corresponding prediction class with the highest prediction probability in brackets.

5. Conclusion

In this work, a new online knowledge extraction method, called Dense Feature Fusion for Online Mutual Knowledge Distillation, is proposed for online knowledge distillation. The specific method is to fuse the characteristics of teachers and students, and establish an auxiliary fusion branch to process the fusion information, which can improve the quality of online knowledge extraction. In this way, feature fusion can effectively enhance the feature interaction of teachers' academic network. Plus the normalized integral distillation of the output. Through a large number of experiments, the superiority of this method in cifar-10, cifar-100 and ImageNet is proved.

References

[1] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks [C] // NIPS. Curran Associates Inc. 2012.
[2] Rastegari M, Ordonez V, Redmon J, et al. Xnor-net: Imagenet classification using binary convolutional neural networks [C] // European conference on computer vision. Springer,
Cham, 2016: 525-542.

[3] Li, H., Kadav, A., Durandovcic, I., Samet, H., and Graf, H. P. Pruning filters for efficient convnets. arXiv preprint arXiv: 1608.08710, 2016.

[4] Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503.02531, 2015.

[5] Zhang, Y., Xiang, T., Hospedales, T. M., and Lu, H. Deep mutual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4320 – 4328, 2018.

[6] Lan, X., Zhu, X., and Gong, S. Knowledge distillation by on-the-fly native ensemble. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 7528–7538. Curran Associates Inc., 2018.

[7] S. Hou, X. Liu, and Z. Wang, “Dualnet: Learn complementary features for image recognition,” in 2017 IEEE International Conference on Computer Vision (ICCV), Oct 2017, pp. 502–510.

[8] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. arXiv preprint arXiv: 1412.6550, 2014.

[9] Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv: 1612.03928, 2016a.

[10] Kim, J., Park, S., and Kwak, N. Paraphrasing complex network: Network compression via factor transfer. In Advances in Neural Information Processing Systems, pp. 2760–2769, 2018.

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

[12] 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.

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

[15] Samuli Laine and Timo Aila. Temporal ensembling for semisupervised learning. In International Conference on Learning Representations, 2017.

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

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

[18] S. Hou, X. Liu, and Z. Wang, “Dualnet: Learn complementary features for image recognition,” in 2017 IEEE International Conference on Computer Vision (ICCV), Oct 2017, pp. 502–510.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv: 1704.04861, 2017.

[20] Saining Xie, Ross Girshick, Piotr Dollar, 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.

[21] 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.

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

[23] 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.

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

[25] 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.

[26] Samuli Laine and Timo Aila. Temporal ensembling for semisupervised learning. In International Conference on Learning Representations, 2017.

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

[28] 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.

[29] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In Proceedings of IEEE International Conference on Computer Vision, p. 618626. (2019).