Knowledge Distillation via Information Matching

Honglin Zhu[1], Ning Jiang[1,3]✉, Jialiang Tang[2], and Xinlei Huang[1,3]

[1] School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, Sichuan, 621000, China
✉Corresponding author:jiangning@swust.edu.cn
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094, China
[3] Jiangxi Qiushi Academy for Advanced Studies, Nanchang, Jiangxi, 330036, China

Abstract. Knowledge distillation can enhance network generalization by guiding a smaller student network to learn from a more complex teacher network. The challenge lies in maximizing the performance of the student network under the supervision of the teacher network. Currently, the feature-based distillation approach utilizes the middle-layer features of the teacher network to improve the performance of the student network. However, this approach lacks a measure to evaluate the content of the information present in the intermediate layers of both the teacher and student networks, which leads to a distillation mismatch of features and damages the student's performance. In this study, we propose a new feature distillation method to solve this problem. We measure the information content in the intermediate layers of the teacher and student networks based on the receptive fields of corresponding features. Subsequently, the suitable number and locations of transmission features are decided based on information content, effectively alleviating the risk of information mismatch during distillation. Our experimental results demonstrate that the proposed method significantly improves the performance of the student network.

Keywords: Model Compression · Knowledge Distillation · Receptive Field· Feature Distillation.

1 Introduction

Deep neural networks (DNNs) have been widely used in computer vision for tasks such as object detection [16] and semantic segmentation [7]. However, these excellent DNNs with a large number of parameters often require huge computing consumption and memory occupation, which make it difficult to deploy the model on resource-constrained devices. Many model compression methods have been proposed to solve this problem, including network quantization [9], network pruning [6, 11, 12, 20], lightweight network design [14, 28], and knowledge distillation (KD) [1,8,17,26,29]. Among them, the knowledge distillation method has attracted much attention due to its effectiveness and simplicity.

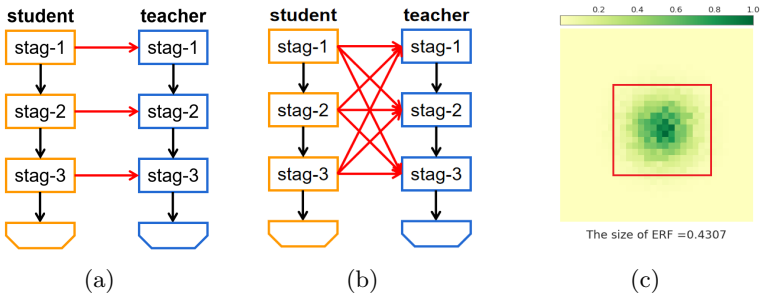
Knowledge distillation refers to using the pre-training teacher network to transfer knowledge to a small student network to improve the performance of

the student network. Generally speaking, there are different types of knowledge, such as soft labels [8] and middle hidden layer information [17]. Compared with the category dependency information in soft labels, there is rich texture information contained in the features of the middle hidden layer, which can be taken as useful supervision signals for the student network. In general, as shown in Fig. 1a and 1b, most of the existing methods artificially select feature transfer locations and lack measurement methods to quantify the information content in different feature locations between the teacher and the student. For example, a series of methods [2,3,17,25,26] propose to match the middle-layer features of the teacher network and the student network one-to-one. In fact, due to the structural differences between the teacher and the student, the student network needs to mimic information suitable for itself. Furthermore, with the change of network parameters during training, the conventional one-to-one feature matching may cause an information mismatch, which inevitably damages the performance of the student network.

In deep learning, the receptive field [13] is an important tool for understanding how the network works, which is defined as a region where a specific feature is affected by the input space. The receptive field can quantify the size of an image captured by a feature. Intuitively, if a feature can capture a large size for an image, it contains much information content about the image. Therefore, we calculate the receptive fields of the features of the teacher and student to quantify their information content, which can be utilized to precisely match the transferring features from teacher to student.

Based on the above analysis, as shown in Fig. 2, we propose a distillation framework called **I**nformation **M**atching **K**nowledge **D**istillation (IMKD) to efficiently align the teacher and the student. Firstly, the receptive field computation is introduced to automatically compute the information difference between student and teacher in the middle layer. Secondly, we propose a simple feature-matching method to obtain suitable teacher-student layer pairs according to the information difference. The student network mimics the information of the suitable feature locations to obtain higher performance and stronger generalization. Our method has mainly two advantages for distillation over previous feature-based distillation: 1) By measuring the information difference between the features of the teacher and the student, we can properly assign features of the teacher network to transfer to the student network; 2) we can adaptively match the target features according to the performance improvement of the student network. We conduct experiments on different combinations of teacher-student networks. The experimental results in Section 4 show that our method significantly improves different neural networks. For example, the proposed method improves the performance of ShuffleNetV2 [14] from 71.82% to 78.13% on CIFAR100. To sum up, the main contributions of this paper are as follows:

- We propose a novel information-matching knowledge distillation (IMKD) to adaptively match target features, which can effectively alleviate the problem of information mismatch in feature distillation.



(a) (b) (c)

Fig. 1: (a) One-to-one feature matching. (b) Fully connected feature matching. (c) Effective receptive field. The area in the red box is a high contribution area. The effective receptive field size is obtained by calculating the ratio of the high contribution area to the entire area. Visualization method derived from work [5]

- We analyze the importance of the receptive field for feature distillation and introduce it as a new metric to quantify the difference between teacher-student layer pairs.
- Our proposed method can adjust the learning features for the student network according to the performance change of the student network in distillation to further improve the performance of the student network.

2 Related Work

2.1 Knowledge Distillation

Knowledge distillation aims to train an excellent, compact student network by learning the predictions of a cumbersome teacher network. Hinton *et al.* [8] propose the pioneering method, where the student network learns the soften output probabilities of the teacher network. Subsequent works explore rich knowledge, *e.g.*, intermediate layer responses [17], attention maps [26], sample relations composed of sample similarity matrices [15], and representation learning [21]. In addition, cross-layer feature distillation methods [2, 3, 31] propose to make the student network fully mimic the multi-layer features of the teacher network and achieve surprising results.

2.2 Receptive Field

The receptive field is a concept that helps to understand and analyze DNNs, which can measure the degree of association between the feature output of the network and the input. In 2016, Luo *et al.* [13] proposed the concept of the effective receptive field (ERF) and visualized the ERF of the network. The ERF of the network is much smaller than the theoretical receptive field, which presents a Gaussian distribution in the center of the image. Recently, Ding *et al.* [5] greatly improved the performance of the network by increasing the ERF. In this study,

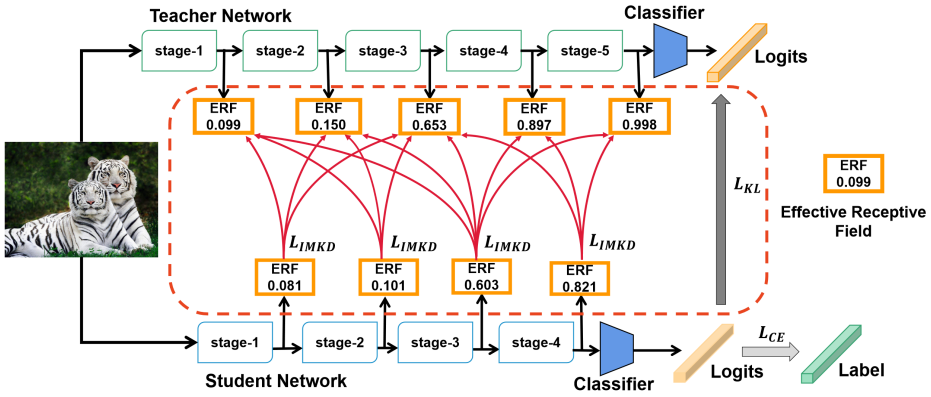


Fig. 2: Overview of our proposed distillation method (IMKD). First, the features and effective receptive fields of each layer are obtained through forward propagation. Second, the difference between each teacher-student layer pair is computed, and the proposed IMKD automatically matches suitable teacher-student layer pairs. Finally, the student mimics the teacher's predictions via L_{total}.

we measure the information difference between the teacher and the student by introducing the ERF into distillation training. We propose a matching algorithm to enable the student network to obtain better feature representation capabilities by matching appropriate feature information.

3 Method

In this section, we first introduce the necessary concepts of knowledge distillation and basic notations in Section 3.1. Then, effective receptive field calculation is described in Section 3.2. Finally, we propose the novel matching mechanism and main formulations in Section 3.3.

3.1 Preliminaries

Formally, we define the training dataset as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where n denotes the number of examples in $\mathcal{D}$ and $\mathcal{D}$ contains c categories. The fixed pre-trained teacher network and the randomly initialized student network are denoted as T and S, respectively. For an image $\mathbf{x} \in \mathcal{D}$ inputting into the teacher network T and the student network S, the logits are obtained as of the student network S and the teacher network T as $\mathbf{g}_S = S(\mathbf{x})$ and $\mathbf{g}_T = T(\mathbf{x})$, respectively. The features of the student network and the teacher network are denoted as $F_{S_l} \in \mathbb{R}^{C_{S_l} \times H_{S_l} \times W_{S_l}}$ and $F_{T_l} \in \mathbb{R}^{C_{T_l} \times H_{T_l} \times W_{T_l}}$, where $S_l \in \{S_1, ..., S_L\}$ and $T_l \in \{T_1, ..., T_L\}$. C, H, and W represent the channel number, height, and width of the corresponding feature, respectively. S_L and T_L are the total number of features of the teacher and the student, where S_L and T_L are not correlated. $\sigma(\cdot)$ denotes as the softmax function, and l represents the index of layers in the corresponding model. The

symbol ω is a balance parameter that is used to reasonably control the size of the difference.

The core idea of knowledge distillation is to improve network performance by transferring the knowledge of the teacher network to the student network. In classical KD [8], distillation training is achieved by a minidiscrepancy between the soft labels of the teacher network and student network:

$$L_{KD} = L_{CE}(y, \sigma(\mathbf{g}_S)) + \tau^2 L_{KL}(\sigma(\frac{\mathbf{g}_S}{\tau}), \sigma(\frac{\mathbf{g}_T}{\tau})), \tag{1}$$

where $L_{CE}(\cdot, \cdot)$ represents the cross-entropy loss, which is used to calculate the difference between the truth label y and the prediction $\sigma(g_S)$, L_{KL} is the KL divergence, τ is a temperature parameter to control the soften degree of the soft labels.

3.2 Effective Receptive Field Calculation

In our IMKD, we introduce ERF to quantify the gap of features between the teacher and the student in preparation for selecting the appropriate feature pairs.

To obtain the effective receptive field of a feature, we first compute the gradient of the feature. Suppose $\mathbf{x}^{j,u,v}$ is denoted as a pixel of the input $\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^{3 \times H_{\mathbf{x}} \times W_{\mathbf{x}}}$. First, the feature $F_{S_l} \in \mathbb{R}^{C_{S_l} \times H_{S_l} \times W_{S_l}}$ are converted to $\tilde{F}_{S_l} \in \mathbb{R}^{H_{S_l} \times W_{S_l}}$ by summation operation. Following the work [13], the partial derivative $\partial \tilde{F}^{1,1}_{S_l} / \partial \mathbf{x}^{j,u,v}$ can measure the importance of $\mathbf{x}^{j,u,v}$ with respect to $\tilde{F}^{1,1}_{S_l}$. Similarly, the importance of an input $\mathbf{x}$ with respect to $\partial \tilde{F}^{1,1}_{S_l}$ can be expressed as $G^{1,1}_{S_l} = \sum_{j=1}^{3} \sum_{u=1}^{H_{\mathbf{x}}} \sum_{v=1}^{W_{\mathbf{x}}} \frac{\partial \tilde{F}^{1,1}_{S_l}}{\partial \mathbf{x}^{j,u,v}}$. Then, we can obtain the gradient G_{S_l} of the feature F_{S_l}, where the gradient $G_{S_l} = [G^{1,1}_{S_l}, ..., G^{H_{S_l}, W_{S_l}}_{S_l}] \in \mathbb{R}^{H_{S_l} \times W_{S_l}}$. We show the effective receptive field by visualizing this gradient in Fig. 1c. Moreover, the high contribution region containing 99% of non-zero gradient values is selected, and its height and width are both K_{S_l}. We obtain the size R_{S_l} of the ERF through the area ratio of the high contribution area to the entire area:

$$R_{S_l} = \frac{K_{S_l} \times K_{S_l}}{H_{S_l} \times W_{S_l}}. \tag{2}$$

In this study, the size R of the ERF at each layer of the teacher network and the student network can be expressed as $C = \{(R_{S_l}, R_{T_l}) | \forall R_{S_l} \in [R_{S_1}, \cdots, R_{S_L}], \forall R_{T_l} \in [R_{T_1}, \cdots, R_{T_L}]\}$.

3.3 Feature Matching Distillation

The key point of our method is to measure the information difference and match target features suitable for the student network to learn. Overall, our method can be divided into three steps: 1) The same examples are input into the teacher and the student to obtain effective receptive fields; 2) through the calculated effective receptive fields, the student network selects suitable target features;

and 3) knowledge from the teacher network is transferred to the student network based on the selected features.

First, we can use the following formula to get the difference in the receptive field of each layer:

$$\alpha_{(T_l,S_l)} = |R_{T_l} - R_{S_l}|. \tag{3}$$

We need to obtain target feature maps matching the student network to avoid negative effects. Therefore, we classify the difference factor α as follows: 1) If α is smaller than ω, we keep such a teacher-student layer pair; 2) otherwise, we regard such a feature pair as a semantic mismatch and discard it. The specific formula is as follows:

$$W_{(T_l,S_l)} = \begin{cases} 1, & if \quad \alpha_{(T_l,S_l)} < \omega \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where the reasonable value range of ω is shown in Sec 4.4. Then, the mean square error (MSE) is used to minimize the distance between selected target features from the teacher network and those from the student network. Distillation loss L_{IMKD} is as follows:

$$L_{IMKD} = \sum_{S_l=1}^{S_L} \sum_{T_l=1}^{T_L} W_{(T_l,S_l)} MSE(r(F_{S_l}), r(F_{T_l})), \tag{5}$$

where $r(\cdot)$ is an adaptive layer composed of 1x1 convolution and an adaptive pooling layer, which is used to align the dimension of features.

Finally, we use the following total loss to train the student network:

$$L_{TOTAL} = L_{CE} + \tau^2 L_{KL} + \beta L_{IMKD}, \tag{6}$$

where β is a hyperparameter used to balance the loss function.

4 Experiments

The training details and datasets for all experiments are in Sec. 4.1. The effectiveness of the proposed distillation method has been verified on different datasets. In Sec. 4.2, we show experimental results of teacher-student networks with the same structures and different structures for distillation on CIFAR-100, STL10 and Tiny-ImageNet. To further verify the effectiveness of the method, we provide the results of the comparison of different distillation methods. Moreover, Section 4.3 provides some visualizations of our approach. Finally, the results of ablation experiments are presented in Sec. 4.4.

4.1 Datasets and Experiments Configuration

Three basic image classification datasets, including CIFAR100 [10], Tiny-ImageNet [22], and STL10 [4], are used to verify the effectiveness of the proposed method.

Following previous work [2, 21], we employ a stochastic gradient descent optimizer with momentum set to 0.9. For all experiments, the weight decay is set to 5×10^{-4}, and the batch size is set to 64. For CIFAR100, the initial learning rate is set to 0.05, and the learning rate decreases by 10 at 150, 180, and 210 epochs until 240 epochs. The configuration of the Tiny-ImageNet dataset is consistent with that of CIFAR100. For STL10, the initial learning rate is set to 0.05 and is decayed by 0.1 every 30 epochs. We set the L_{CE} weight and temperature parameters τ as 1 and 4, respectively.

Table 1: Top-1 test accuracy of different distillation approaches on CIFAR100.

Teacher	ResNet32x4	ResNet32x4	ResNet-110x2	ResNet32x4	WRN-40-2
Student	ShuffleNetV1	ShuffleNetV2	ResNet-116	resnet8x4	ShuffleNetV1
Teacher	79.42%	79.42%	78.18%	79.42%	75.61%
Student	70.50%	71.82%	74.46%	72.50%	70.50%
KD [8]	74.07%	74.45%	76.14%	73.33%	74.83%
FitNet [17]	73.59%	73.54%	76.20%	73.44%	73.73%
AT [26]	71.73%	72.73%	76.84%	72.94%	73.32%
CRD [21]	75.11%	75.65%	76.83%	75.51%	76.05%
SRRL [24]	75.18%	76.19%	77.19%	75.39%	76.30%
SemCKD [2]	76.31%	77.62%	76.69%	76.23%	76.83%
DKD [30]	76.45%	77.07%	77.08%	76.32%	76.70%
Ours	**77.17%**	**78.13%**	**78.05%**	**76.89%**	**77.31%**

4.2 Compared to Different Distillation Methods

Table 1 show the top-1 test accuracy of multiple methods with different network combinations on CIFAR100. We compare eight different distillation methods, among which FitNet [17], AT [26], and SemCKD [2] are feature-based distillation methods. From Table 1, we can see that the proposed method significantly improves network accuracy compared to other methods on CIFAR100. For example, in the network combination "ResNet32x4 [23]/ShuffleNetV1 [28]", the test accuracy of our method achieves a performance improvement of 6.51% compared to the baseline. The higher performance of the network reflects that the proposed method alleviates the information mismatch problem of feature distillation.

Table 2 shows the experimental results of different teacher-student architectures on Tiny-ImageNet. From Table 2, we can see that IMKD outperforms other feature distillation methods. Especially, the best example is in the setting of "ResNet32x4 [23]/ShuffleNetV2 [14]" where the student's performance is even better than the teacher's performance. The reason for the excellent performance of our proposed approach is to select connections suitable for training.

Table 3 shows the comparison results of different distillation methods on STL10. Our method significantly improves the test accuracy of different networks on the STL10 dataset. For example, for the network combinations "WRN-

40-2 [27]/ShuffleNetV1 [28]" and "WRN-40-2 [27]/MobileNetV2 [18]", IMKD outperforms semckd by 2.05% and 3.03%. Furthermore, IMKD significantly outperforms KD methods in multiple network combinations. These excellent test results show that our method can select the target features suitable for distillation, and the training of IMKD is stable and robust.

Table 2: Top-1 test accuracy of different distillation approaches on Tiny-ImageNet.

Teacher Student	WRN-40-2 WRN-16-2	ResNet110 ResNet20	ResNet32x4 ShuffleNetV2
Teacher	62.53%	60.36%	65.28%
Student	58.60%	54.03%	62.40%
KD [8]	60.92%	55.75%	67.95%
FitNet [17]	58.63%	53.20%	67.82%
AT [26]	60.20%	55.86%	66.90%
RKD [15]	58.72%	54.22%	68.28%
SemCKD [2]	60.82%	56.52%	68.13%
Ours	**61.67%**	**57.28%**	**68.87%**

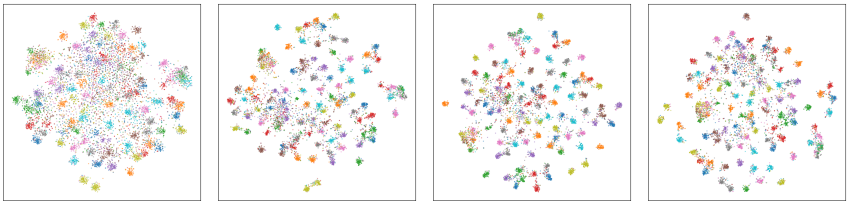
Table 3: Top-1 test accuracy of different distillation approaches on STL10.

Teacher Student	VGG13 VGG8	WRN-40-2 MobileNetV2	WRN-40-2 ShuffleNetV1	ResNet32x4 MobileNetV2
Teacher	71.58%	71.26%	71.26%	69.67%
Student	69.30%	53.30%	59.60%	53.30%
KD [8]	71.65%	58.07%	61.33%	58.06%
FitNet [17]	71.81%	63.79%	68.13%	63.26%
AT [26]	70.41%	58.43%	69.86%	60.51%
SRRL [24]	69.76%	53.83%	65.86%	57.09%
SemCKD [2]	72.61%	66.58%	70.33%	61.99%
DKD [30]	69.73%	58.04%	63.58%	58.89%
Ours	**73.30%**	**69.61%**	**72.38%**	**65.01%**

4.3 Visualization

To verify the advantage of our proposed IMKD, we provide the visualization of the penultimate layer of the student network. In Fig. 3, IMKD has stronger representation ability compared to the student network. Moreover, our approach has more scattered embeddings than KD.

To provide visual interpretations of IMKD, we use Grad-CAM [19] to highlight regions that the model values in Table 5. Three images labeled "Espresso", "Labrador" and "Orange" are randomly selected on Tiny-ImageNet. In Table 5,



(a) Student (b) KD (c) Teacher (d) IMKD

Fig. 3: The penultimate layer visualization of ShuffleNetV1 (student) with KD, IMKD and the teacher (ResNet32x4) on CIFAR-100.

Table 4: The ablation experiments with different feature matching approachs.

Teacher Student	ResNet32x4 ShuffleNetV1	WRN-40-2 ShuffleNetV1	ResNet32x4 ResNet8x4
Teacher	79.42%	75.61%	79.42%
Student	70.50%	70.50%	72.50%
one-to-one feature matching	76.09%	75.85%	75.18%
Fully connected feature matching	76.10%	76.20%	75.54%
Ours(IMKD)	**77.17%**	**77.31%**	**76.68%**

the Grad-CAM maps of IMKD pays more attention to important regions than KD and baseline. As shown in the visualization in the fourth row of Table 5, our method can focus on the location of the orange.

4.4 Ablation

We set up three feature matching methods to verify the effectiveness of the proposed information matching method in Table 4. One-to-one feature matching is set as the same stage of information transfer, as shown in Fig. 1a. The fully connected feature matching approach is shown in Fig. 1b. Different from other feature distillation frameworks, our IMKD performs automatic feature matching based on information differences. According to the data in Table 4, suitable information transfer can significantly improve the performance of the student network. In particular, the other feature matching methods are significantly lower than the information matching distillation method, which illustrates the suboptimal performance of hand-crafted feature matching methods. The information matching distillation can effectively obtain excellent target features.

Table 6 shows the test accuracy of the network under different settings of β and ω, where β and ω belong to Equation 6 and Equation 4, respectively. The network combination is set to "ResNet32x4 [23]/ShuffleNetV2 [14]". From Table 6, the experimental results show that matching suitable target features can effectively improve performance ($\omega = 0.7$). Furthermore, boost is best when β is equal to 60. Combining the information in Table 6 and Figure 2, when the locations of multiple target features are set to be similar to the stage of

Table 5: Grad-CAM visualization of WRN-16-2 (student) with KD, IMKD and the teacher (WRN-40-2) on Tiny-ImageNet. The labels of the three randomly selected photos are “Espresso”, “Labrador” and “Orange”.

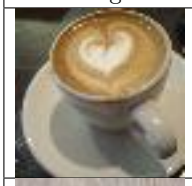
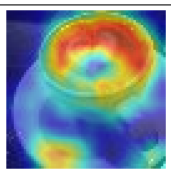
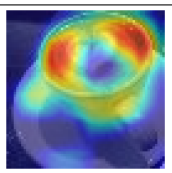
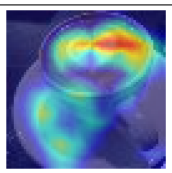
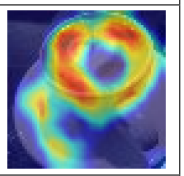
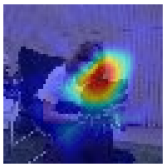
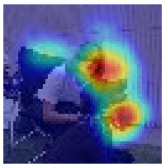
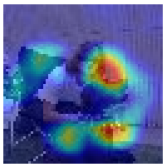
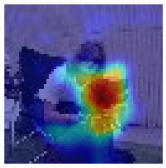
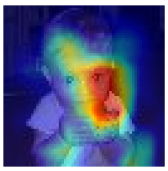
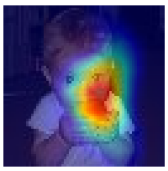
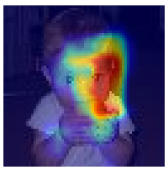
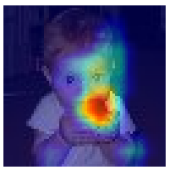
Image	Student	Teacher	KD	IMKD
				
				
				

Table 6: The ablation of β and ω on CIFAR100.

β	40	60	80	100
Top-1	77.02%	**78.13%**	77.54%	77.61%
ω	0.3	0.5	0.7	0.9
Top-1	77.46%	77.65%	**78.13%**	77.27%

the student network, the performance of the student network is significantly improved.

5 Conclusion

This academic paper highlighted a long-neglected aspect of feature map distillation-based methods: the absence of effective metrics and appropriate matching techniques for training student networks. To address this issue, we propose a novel approach called IMKD. Our IMKD utilizes the receptive field computation to calculate the difference in information between the teacher network and student network. To ensure that the features between teacher and student are appropriately matched during the distillation process, we introduce a feature matching distillation technique that adapts to the information differences, rather than relying on manual selection. This approach enables the student network to acquire the most optimal matching information continuously and iteratively throughout

the distillation training process. Our extensive experiments demonstrate that IMKD outperforms previous distillation methods.

Acknowledgement. This research is supported by Sichuan Science and Technology Program (No. 2022YFG0324), SWUST Doctoral Research Foundation under Grant 19zx7102.

References

1. Chen, D., Mei, J.P., Zhang, H., Wang, C., Feng, Y., Chen, C.: Knowledge distillation with the reused teacher classifier. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11933–11942 (2022)
2. Chen, D., Mei, J.P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., Chen, C.: Cross-layer distillation with semantic calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7028–7036 (2021)
3. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5008–5017 (2021)
4. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223. JMLR Workshop and Conference Proceedings (2011)
5. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11963–11975 (2022)
6. Ghosh, S., Srinivasa, S.K., Amon, P., Hutter, A., Kaup, A.: Deep network pruning for object detection. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 3915–3919. IEEE (2019)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
9. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2704–2713 (2018)
10. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
11. Liu, Z., Sun, M., Zhou, T., Huang, G., Darrell, T.: Rethinking the value of network pruning. arXiv preprint arXiv:1810.05270 (2018)
12. Luo, J.H., Wu, J., Lin, W.: Thinet: A filter level pruning method for deep neural network compression. In: Proceedings of the IEEE international conference on computer vision. pp. 5058–5066 (2017)
13. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems **29** (2016)
14. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018)

15. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
17. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
18. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
20. Tang, J., Liu, M., Jiang, N., Cai, H., Yu, W., Zhou, J.: Data-free network pruning for model compression. In: 2021 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–5. IEEE (2021)
21. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
22. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE transactions on pattern analysis and machine intelligence **30**(11), 1958–1970 (2008)
23. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
24. Yang, J., Martinez, B., Bulat, A., Tzimiropoulos, G., et al.: Knowledge distillation via softmax regression representation learning. International Conference on Learning Representations (ICLR) (2021)
25. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4133–4141 (2017)
26. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
27. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
28. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)
29. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4320–4328 (2018)
30. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022)
31. Zhu, H., Jiang, N., Tang, J., Huang, X., Qing, H., Wu, W., Zhang, P.: Cross-layer fusion for feature distillation. In: Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part IV. pp. 433–445. Springer (2023)