Dynamic Refining Knowledge Distillation Based on Attention Mechanism

Xuan Peng and Fang Liu(✉)

School of Electronic Science, National University of Defense Technology,
Changsha, China
smartlf@sina.com

Abstract. Knowledge distillation is an effective strategy to compress large pre-trained Convolutional Neural Networks (CNNs) into models suitable for mobile and embedded devices. In order to transfer better quality knowledge to students, several recent approaches have demonstrated the benefits of introducing attention mechanisms. However, the existing methods suffer from the problems that the teachers are very rigid in their teaching and the application scenarios are limited. In face of such problems, a dynamic refining knowledge distillation is proposed in this paper based on attention mechanism guided by the knowledge extraction (KE) block whose parameters can be updated. With the help of the KE block, the teacher can gradually guide students to achieve the optimal performance through a question-and-answer format, which is a dynamic selection process. Furthermore, we are able to select teacher networks and student networks more flexibly with the help of channel aggregation and refining factor r. Experimental results on the CIFAR dataset show the advantages of our method for training small models and having richer application scenarios compared to other knowledge distillation methods.

Keywords: Network compression · Knowledge distillation · Dynamic refining · Attention mechanism

1 Introduction

Convolutional neural networks (CNNs) have achieved impressive success in computer vision tasks such as image classification [4,23], object detection [14,16], and semantic segmentation [21,24]. However, the advantages of performance are driven at the cost of training and deploying resource intensive networks with millions of parameters. As application scenarios shift toward mobile and embedded devices, the computational cost, memory consumption, and power consumption of large CNNs prevent them from being deployed to these devices, which drives research on model compression. Several directions such as model pruning [10,11,20], model quantization [12], and knowledge distillation [5,9,15,17,22] are proposed to enable the model to be deployed in resource-constrained scenarios.

S. Khanna et al. (Eds.): PRICAI 2022, LNCS 13630, pp. 45–58, 2022.
https://doi.org/10.1007/978-3-031-20865-2_4

Among them, knowledge distillation aims to compress a network by using the knowledge of a larger network or its ensemble (teacher) as a supervision to train a compact network (student) [19]. Different from other compression methods, it can compress the network regardless of the structural differences between teachers and students.

Attention plays a critical role in the human visual experience. In computer vision, methods of focusing attention on the most important regions of an image and ignoring irrelevant parts are called attention mechanisms [3]. In a vision system, the attention mechanism can be considered as a dynamic selection process, which is implemented by adaptively weighting the features according to the importance of the input. [22] first introduced spatial attention in knowledge distillation (AT), which transfers spatial attention maps to students as knowledge. [17] introduced channel attention in knowledge distillation (KDPA) through borrowing the squeezing operation of Squeeze-and-Excitation Networks proposed by [6]. These methods have yielded good results, but there are still some problems.

For example, firstly, teachers are too rigid in teaching students as they only give steps on how to solve a problem, which is more like students learning on their own through reference answers. However, this is not enough, because a real teacher usually guides his students' learning through a question-and-answer format. More interaction should be generated between the teacher and the students. Secondly, the choice of teacher-student combinations is restricted. AT must ensure that the spatial dimensions $W \times H$ of the blocks corresponding to the teacher and student networks are equal, while KDPA needs to ensure that the channel dimension C of the blocks corresponding to the teacher and student networks is equal.

In order to address these issues, we propose a dynamic refining knowledge distillation based on attention mechanism named DRKD, which introduces the KE block whose parameters can be updated. During training, a complete question and answer session is composed of one forward and one backward propagation. The forward propagation means that the teacher and the student give their answers separately to the same problem. During the back propagation, the parameters of both the KE block and the student are updated. The process of the student's parameters being updated indicates that the student is correcting the answer based on the teacher's response, and the parameters of the KE block being updated means that the teacher is recalibrating the answer based on the student's feedback. After many question and answer sessions, the teacher gradually guides the students to find the best answer. Moreover, with the help of the channel encoding and the channel refining, the choice of teacher-student combinations can be more flexible regardless of the dimensional differences in the feature maps of the corresponding blocks between teachers and students. In short, the contributions of this paper can be summarized as follows:

1) We propose a novel knowledge distillation method named DRKD. By introducing the KE block with parameters that can be updated, our approach is able to dynamically adjust the knowledge transferred to students based on

their feedback. The approach emulates the human knowledge transfer approach driven by questions.

2) Our proposed method effectively mitigates the problem that many excellent knowledge distillation methods cannot be used in most teacher-student combinations, which greatly enriches the application scenarios of the algorithm.
3) We experimentally demonstrate that our approach provides significant improvements in the training of small models and shows flexibility in the selection of teacher-student combinations.

2 Related Work

Knowledge Distillation. Many studies have been conducted since [5] proposed the first knowledge distillation based on the soften class probabilities. [15] first introduced the knowledge of the hidden layer to improve knowledge distillation, which suggests that the knowledge of the hidden layer also has an important impact on students during the process of knowledge transfer. Inspired by this, various other methods have been proposed to indirectly match the feature activation values of teacher and student networks. [9] proposed knowledge distillation combined with singular value decomposition (SVD) to effectively remove the spatial redundancy in the feature map by reducing the spatial dimension of the feature maps. [8] introduced the so-called "factors", which uses convolutional operations to paraphrase teacher's knowledge and to translate it for the student. [7] utilized the outputs of the hint layer of teacher to supervise student, which reduces the performance gap between teacher and student. [22] proposed to use the sum of absolute values of a feature as the attention map to implement knowledge distillation. [17] used the channel attention mechanism to highlight the expressive feature in the middle layer.

Channel Attention Mechanism. In deep neural networks, different channels in different feature maps usually possess different features [1]. Channel attention adaptively adjusts the weights of each channel, which can be seen as a feature selection process to determine what should be paid attention to [3]. [6] first proposed the concept of channel attention and presented SENet, which can capture channel-wise relationships and improve representation ability. Inspired by this, many SENet-based channel attention studies began to emerge. [2] proposed a global second-order pooling block to solve the problem of SENet's difficulty in capturing higher-order statistics. [18] proposed the efficient channel attention block which uses a 1D convolution to determine the interaction between channels. It tackles the issue that SENet cannot directly model the correspondence between weight vectors and inputs. Only using the global average pooling in the squeeze module limits representation ability. To obtain a more powerful representation ability, [13] rethought global information captured from the viewpoint of compression and analysed global average pooling in the frequency domain.

3 Methodology

The core idea of our proposed approach is how to dynamically extract the knowledge transferred to students. This section is divided into three parts to present our proposed method. Section 3.1 presents the general structure of DRKD. Sect. 3.2 introduces the specific details of implementing the KE block. Finally, we define the loss terms in Sect. 3.3 based on the carefully designed distilled knowledge.

3.1 Overall Structure of DRKD

The structure of the DRKD is shown in Fig. 1. Most existing neural networks are composed of several blocks. For example, WideResNet (WRN) consists of three blocks and ResNet consists of four blocks. Each block contains many convolutional layers, batch normalization layers and activation layers. In this paper, the dynamic refining process is implemented by introducing a pair of the KE blocks at the output of the corresponding blocks in the teacher and student networks. The refining process does not mean to extract specific knowledge, but rather to dynamically adjust the knowledge transferred to students based on their feedback. And this process is more similar to the dynamic selection process of the attention mechanism.

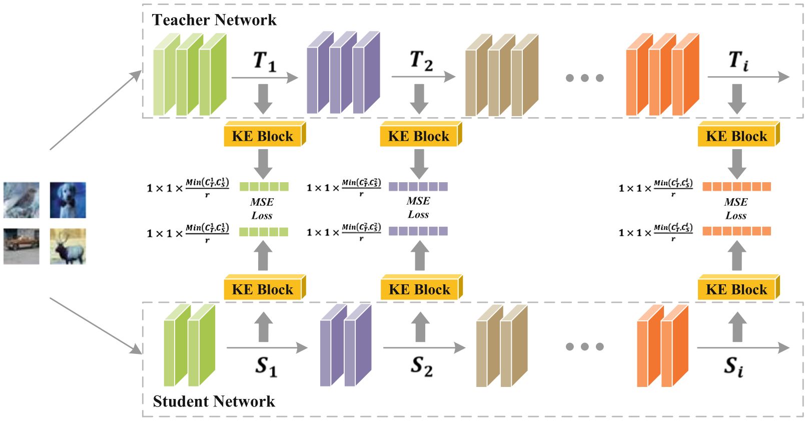

Teacher Network
T_1
T_2
T_i
KE Block
KE Block
KE Block
$1\times1\times\frac{Min(C_T^1,C_S^1)}{r}$
$1\times1\times\frac{Min(C_T^2,C_S^2)}{r}$
$1\times1\times\frac{Min(C_T^i,C_S^i)}{r}$
MSE Loss
MSE Loss
MSE Loss
$1\times1\times\frac{Min(C_T^1,C_S^1)}{r}$
$1\times1\times\frac{Min(C_T^2,C_S^2)}{r}$
$1\times1\times\frac{Min(C_T^i,C_S^i)}{r}$
KE Block
KE Block
KE Block
S_1
S_2
S_i
Student Network

Fig. 1. Schematic diagram of the overall structure of the algorithm. T_i and S_i denote the output feature maps of the i-th block of the teacher and student networks, respectively. C_T^i and C_S^i denote the number of channels of the feature map of the i-th block of the teacher and student networks, respectively.

In details, the feature map of i-th block of the teacher network is written as $T_i = \left\{ f_{T_i}^1, f_{T_i}^2, \cdots, f_{T_i}^{C_T^i} \right\}$, C_T^i denotes the number of channels of the T_i, and the feature map of all blocks of the teacher network can be described as $T =$

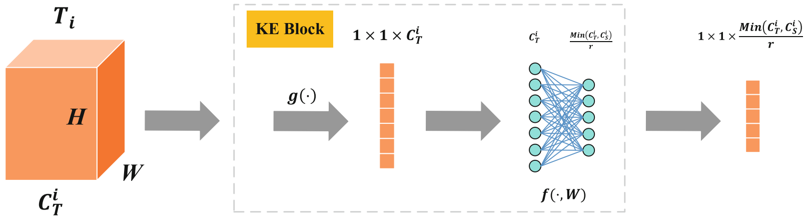

T_i
H
W
C_T^i
KE Block
$g(\cdot)$
$1 \times 1 \times C_T^i$
C_T^i
$\frac{Min(C_T^i, C_S^i)}{r}$
$f(\cdot, W)$
$1 \times 1 \times \frac{Min(C_T^i, C_S^i)}{r}$

Fig. 2. Schematic diagram of the structure of the KE block, which is divided into two steps: $g(\cdot)$ and $f(\cdot, W)$. Where $W \in R^{C_T^i \times \frac{Min(C_T^i, C_S^i)}{r}}$, which is determined by the number of channels of the i-th block feature map of the teacher and student networks.

$\{T_1, T_2, \cdots, T_N\}$, N denotes the number of the entire network block. The feature map of i-th block of the student network is written as $S_i = \left\{ f_{S_i}^1, f_{S_i}^2, \cdots, f_{S_i}^{C_S^i} \right\}$, C_S^i denotes the number of channels of the S_i, and the feature map of all blocks of the student network can be described as $S = \{S_1, S_2, \cdots, S_N\}$, N denotes the number of the entire network block.

3.2 KE Blocks

Figure 2 shows the specific structure of the KE block with T_i as the input example. The KE block is implemented on the features of each block through two steps: channel encoding and channel refining.

Channel Encoding. In order to tackle the issue of spatial dimension mismatch between corresponding blocks of the teacher and student networks, it is a feasible approach that encodes the global spatial information of each channel into a channel descriptor. The study by [6] also showed that features T_i or S_i in the hidden layer can be interpreted as a collection of the local descriptors whose statistics are expressive for the whole image. Many aggregation strategies can be used to achieve channel aggregation. Considering the computational complexity, the simplest global average pooling is chosen. The statistics $Z_{T_i} \in \mathbb{R}^{C_T^i}$ and $Z_{S_i} \in \mathbb{R}^{C_S^i}$ are generated by shrinking T_i and S_i through spatial dimensions, respectively. The k-th element of Z_{T_i} and the m-th element of Z_{S_i} are calculated by:

$$Z_{T_i}^k = g\left(f_{T_i}^k\right) = \frac{1}{H_{T_i} \times W_{T_i}} \sum_{x=1}^{H_{T_i}} \sum_{y=1}^{W_{T_i}} f_{T_i}^k(x, y) \tag{1}$$

$$Z_{S_i}^m = g\left(f_{S_i}^m\right) = \frac{1}{H_{S_i} \times W_{S_i}} \sum_{x=1}^{H_{S_i}} \sum_{y=1}^{W_{S_i}} f_{S_i}^m(x, y) \tag{2}$$

where $Z_{T_i}^k$ denotes the k-th element of the channel descriptor vector of the i-th block of the teacher network, $f_{T_i}^k$ denotes the k-th channel feature map of the

i-th block of the teacher network, and $H_{T_i} \times W_{T_i}$ denotes the spatial dimension of the i-th block of the teacher network. The student network is as above. Where $1 \leq k \leq C_T^i$, $1 \leq m \leq C_S^i$.

Channel Refining. Related studies [10,11] have shown that there is a certain degree of redundancy in the numerous channels in the convolutional neural networks. Therefore, in order to take advantage of the information aggregated in the channel encoding operation, we follow it with a second operation which aims to dynamically refine the knowledge transferred to the students based on their feedback. To fulfil this objective, the function must satisfy two criteria: first, its parameters must be updatable since we need to ensure that knowledge transfer is a dynamic selection process, and second, its input must be 1D tensor as the output of the channel encoding is 1D tensor. Besides, the function must also act as a dimensionality reduction, considering the problem of channel dimension mismatch between teachers and students. Therefore, the fully connected layer is the only choice:

$$V_{T_i} = f\left(Z_{T_i}, W_{T_i}\right) = \sigma\left(W_{T_i} Z_{T_i}\right) \tag{3}$$

$$V_{S_i} = f\left(Z_{S_i}, W_{S_i}\right) = \sigma\left(W_{S_i} Z_{S_i}\right) \tag{4}$$

where $W_{T_i} \in \mathbb{R}^{\frac{Min\left(C_T^i, C_S^i\right)}{r} \times C_T^i}$, $W_{S_i} \in \mathbb{R}^{\frac{Min\left(C_T^i, C_S^i\right)}{r} \times C_S^i}$ and σ refers to sigmoid activation function. The r is a hyperparameter which plays a crucial role in our proposed algorithm. With the help of r, the problem of mismatching the number of channels in the corresponding blocks of the teacher and student networks can be solved. And r is usually set to an integer value, $1 \leq r \leq Min\left(C_T^i, C_S^i\right)$. As r increases, the total amount of knowledge transferred from the teacher to the students is decreasing, with a greater tendency to filter for high-priority features. The balance between quality and quantity is very important in the knowledge transfer process. The degree of dynamic refining can be adjusted according to the actual situation with the help of r (the choice of this hyperparameter is discussed in Sect. 4.4).

3.3 Loss Function

The loss function of our proposed method consists of two components. One is a cross-entropy loss based on the ground-truth labels and the predicted labels of the student network, and the other is a dynamic refining (DR) loss based on the middle layer features of the network.

At the beginning of training, the ground-truth loss plays an important role in improving the convergence speed of the student network. The loss is calculated by:

$$\mathcal{L}_{cross} = \mathcal{H}_{cross}(y, \hat{y}) \tag{5}$$

where y denotes the ground-truth label, $\hat{y}$ denotes the predicted label of the student network, and $\mathcal{H}_{\text{cross}}$ denotes the cross-entropy function.

During network training, the DR loss acts as a regularization term and helps improve robust. The loss is calculated by:

$$\mathcal{L}_{\mathcal{DR}} = \sum_{i=1}^{N} \frac{1}{\lambda_i} \left\| V_{T_i} - V_{S_i} \right\|_2^2, \lambda_i = \frac{Min\left(C_T^i, C_S^i\right)}{r} \tag{6}$$

where $V_T = \{V_{T_1}, V_{T_2}, \cdots, V_{T_N}\}$ represents the knowledge transferred from the teacher to the student. The r is usually set to an integer value and $1 \leq r \leq Min\left(C_T^i, C_S^i\right)$.

Objective function:

$$\mathcal{L}_{Total} = \mathcal{L}_{cross} + \alpha \mathcal{L}_{\mathcal{DR}} \tag{7}$$

where the α is a hyperparameter that adjusts the proportion of the DR loss term in the final objective function.

4 Experiments

In this section, WideResNet (WRN) and ResNet will be used as our deep neural network models and experimented on the CIFAR datasets. The CIFAR dataset contains CIFAR-10 and CIFAR-100, consisting of 60,000 RGB images of 32×32 pixels. The ratio of both training set and test set is 5 : 1.

4.1 Experiments on Benchmark Datasets

The performance of the algorithm will be proved in two aspects: different network architectures and different number of channels. Therefore, three teacher-student combinations will be chosen, which are the [ResNet34, ResNet18], the [WRN-28-2, WRN-16-2] and the [WRN-10-5, WRN-16-1]. In WRN-n-k, n denotes the depth of the network, and k denotes that the number of channels of the network is k times the number of base channels. During training, the teacher network is untrainable and the student network is used with stochastic gradient descent (SGD) as the optimizer, with momentum set to 0.9 and weight decay set to 5e-4. The initial value of the learning rate is set to 1e-1 and all learning rates are multiplied by 0.7 every 10 epochs. When the [ResNet34, ResNet18] is trained, the best received results are at $\alpha = 1.0$, $r = 1$. When the [WRN-28-2, WRN-16-2] is trained, the best received results are at $\alpha = 1.0$, $r = 2$. When the [WRN-10-5, WRN-16-1] is trained, the best received results are at $\alpha = 1.0$, $r = 16$.

Table 1 and Table 2 show the performance of DR on the CIFAR-10 and CIFAR-100, respectively. In the tables, the compression ratio is calculated as $\frac{T_{params} - S_{params}}{T_{params}}$. Among them, T_{params} denotes the parameters of the teacher, and S_{params} denotes the parameters of the student. When the experiment is conducted on the ResNet, ResNet34 is selected as the teacher network and ResNet18 is chosen as the student network. Compared with the student baseline, the accuracy of the student trained by DR on the CIFAR-10 and CIFAR-100 is improved by 1.99% and 3.29%, separately. When the experiment is carried out on the

Table 1. The performance of DR algorithm on CIFAR-10

Teacher	Student	Compression Ratio	FLOPs	Student Baseline	DR	Teacher Baseline
ResNet34, 21.28M	ResNet18, 11.17M	47.51%	0.56 G	94.10%	**96.09%**	94.21%
WRN-28-2, 1.47M	WRN-16-2, 0.69M	53.06%	0.13 G	92.80%	**95.70%**	93.89%
WRS-10-5, 1.90M	WRN-16-1, 0.08M	95.68%	0.01 G	90.19%	**93.36%**	92.05%

Table 2. The performance of DR algorithm on CIFAR-100

Teacher	Student	Compression Ratio	FLOPs	Student Baseline	DR	Teacher Baseline
ResNet34, 21.28M	ResNet18, 11.17M	47.51%	0.56G	76.01%	**79.30%**	76.71%
WRN-28-2, 1.47M	WRN-16-2, 0.69M	53.06%	0.13G	70.79%	**75.39%**	72.50%
WRS-10-5, 1.90M	WRN-16-1, 0.08M	95.68%	0.01G	64.92%	**71.09%**	70.09%

WideResNet, two teacher-student combinations are selected in terms of whether the number of channels matches. One is the [WRN-28-2, WRN-16-2], in which the accuracy of the student trained by DR on the CIFAR-10 and CIFAR-100 is improved by 2.90% and 4.60%, respectively, compared with the student baseline. The other is the [WRN-10-5, WRN-16-1], in which the accuracy of the student trained by DR on the CIFAR-10 and CIFAR-100 is improved by 3.17% and 6.17%, separately, compared with the student baseline.

From these experiments, it can be seen that DR can significantly improve the performance of the student. As the capacity of the student network gradually decreases, the performance improvement of the students trained by DR gradually becomes larger and the value of the refining factor r increases. Among them, the improvement is more obvious on the CIFAR-100. These show that our method works very well when small models are trained since the lower capacity student network is transferred with higher quality knowledge, reflecting the adjustment effect of r on the balance between quantity and quality. Furthermore, the students even outperform the teachers due to the added ground-truth loss.

4.2 Comparison with Other Methods

In order to demonstrate the effectiveness of our proposed DR more extensively, it is used to compare with other typical knowledge distillation methods. WideResNet is widely used in various knowledge distillation methods for training on the CIFAR. Therefore, WRN-28-2 is chosen as the teacher and WRN-16-2 is selected as the student to perform experiments on the CIFAR. Table 3 shows the performance of DR compared with other typical knowledge distillation algorithms. The accuracy's improvement in the table refers to the comparison with the student baseline, which is obtained by training with a standard back-propagation algorithm. Here the teacher baseline corresponds to the last column of Table 1 and Table 2 and the student baseline corresponds to column 5 of Table 1 and Table 2.

Table 3. Comparison of DR and other typical algorithms on CIFAR

Algorithm	Parameters	FLOPs	CIFAR-10	CIFAR-100
Teacher	1.47 M	0.21 G	+1.09%	+1.71%
KD	0.69 M	0.13 G	+0.74%	+1.52%
AT	0.69 M	0.13 G	+1.17%	+1.66%
KDPA	0.69 M	0.13 G	+1.75%	+2.32%
DRKD (our)	0.69 M	0.13 G	**+2.90%**	**+4.60%**

From these experiments, it can be found that the accuracy of the student trained by DR is improved by 2.90% and 4.60% on the CIFAR-10 and CIFAR-100, respectively, compared to the student baseline. Compared to the teacher baseline, the improvement is 1.81% and 2.89% separately. It performs the best of all methods. And this improvement is even more evident on the CIFAR-100, which again demonstrates the advantage of our approach to train small models. This is because the capacity of the student network is small compared to the task complexity of the CIFAR-100. Our method is more advantageous in dealing with the problem that small capacity networks are difficult to train.

4.3 Ablation Experiments

In this section, a series of experiments based on the teacher-student combination [WRN-28-2, WRN-16-2] are employed to investigate the effect of the hyperparameter α and each block in the network on the algorithm.

First, the effect of each block in the network on the algorithm is studied. WideResNet has three blocks. θ is used to indicate that some blocks of the network are not involved in the loss calculation. For example, $\theta = 001$ means to only calculate the loss for the third block, and so on. When studying the importance of each block, hold other hyperparameters constant and let $\alpha = 1$,

$r = 2$. As shown in Table 4, the student obtains the optimal result when all blocks of the teacher and the student are involved in the loss calculation. From these experiments, it can be found that distillation is not very effective when the knowledge is transferred for only one block, suggesting that the shallow information of the network also plays an important role in guiding students.

Table 4. The effect of each block in the network on the performance of the DR algorithm

θ	CIFAR-10	CIFAR-100
001	93.75%	74.22%
010	95.31%	74.61%
011	94.14%	74.61%
100	94.53%	75.00%
101	94.14%	74.22%
110	94.14%	74.22%
111	**95.70%**	**75.39%**

Second, the importance of the DR loss term is investigated. The α is used to adjust the DR loss term as a percentage of the total loss. When exploring the effect of α on the algorithm, keep other hyperparameters unchanged and let $r = 2$, $\theta = 111$. Table 5 shows how the accuracy of the student network on the CIFAR changes when the DR loss term increases as a percentage of the total loss, and the student obtains the best results when $\alpha = 1$. At first, the accuracy of the student network increases as α becomes larger, but starts to decrease after reaching a certain threshold. Moreover, the algorithm is more sensitive to the value of α before reaching the threshold, because its small changes can lead to large fluctuations in accuracy. From these experiments, it can be found that a balance should be maintained between the DR loss term and the ground-truth loss term. It still helps to improve student's performance when the DR loss is small. But when the DR loss is too large, the degradation of the students' performance is very dramatic as the ground-truth loss term hardly works.

Table 5. The effect of α on the performance of the DR algorithm

α	CIFAR-10	CIFAR-100
0.1	94.53%	73.83%
0.5	94.92%	74.22%
0.7	95.31%	74.61%
1	**95.70%**	**75.39%**
2	94.92%	73.83%
4	95.31%	75.00%
6	94.53%	74.22%
8	94.53%	73.44%
30	93.75%	72.67%

4.4 Refining Factor

The refining factor r is an important hyperparameter that can be used to control the balance between the quantity and the quality of knowledge transferred to students. In this paper, the quantity of knowledge is simply measured by the number of channels. To investigate this relationship, the experiment has been conducted based on whether the number of teacher-student channels matches.

Table 6. Performance of the teacher-student combinations [ResNet34, ResNet18] and [WRN-28-2, WRN-16-2] on CIFAR when the refining factor r takes different values.

Teacher-Student	r	Compression ratio	FLOPs	CIFAR-10	CIFAR-100
[ResNet34, ResNet18]	1	47.51%	0.56 G	**96.09**%	**79.30**%
	2	47.51%	0.56 G	95.70%	78.91%
	4	47.51%	0.56 G	95.70%	78.91%
	8	47.51%	0.56 G	95.31%	78.91%
	16	47.51%	0.56 G	95.31%	77.73%
	32	47.51%	0.56 G	94.53%	75.39%
[WRN-28-2, WRN-16-2]	1	53.06%	0.13 G	93.75%	74.22%
	2	53.06%	0.13 G	**95.70**%	**75.39**%
	4	53.06%	0.13 G	94.92%	73.43%
	8	53.06%	0.13 G	94.53%	75.00%
	16	53.06%	0.13 G	94.53%	73.44%
	32	53.06%	0.13 G	94.14%	73.05%

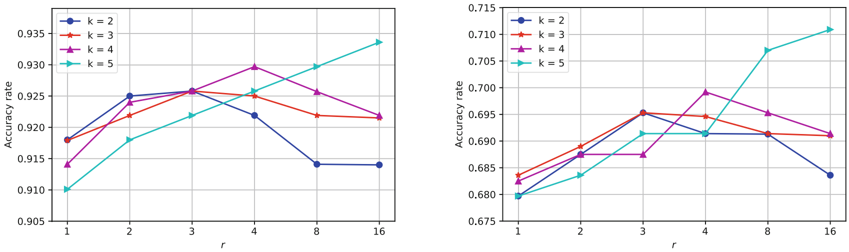
Table 7. Multiple teacher-student combinations with mismatched channel numbers

Student	Teacher	Compression Ratio	k
WRN-16-1	WRN-10-2	75.00%	2
	WRN-10-3	88.41%	3
	WRN-10-4	93.44%	4
	WRN-10-5	95.68%	5

When the Number of Channels in the Teacher-Student Combination Matches. Considering $C_T^i = C_S^i$, experiments have been performed based on teacher-student combinations [ResNet34, ResNet18] and [WRN-28-2, WRN-16-2] for a range of different r values. When studying the effect of r on the algorithm, hold the other hyperparameters constant and let $\alpha = 1$, $\theta = 111$.

The top half of Table 6 shows that the best results are obtained from the teacher-student combination [ResNet34, ResNet18] with the refining factor $r = 1$, meaning that the KE block is still beneficial for improving student's performance even without refining channel features. The bottom half of Table 6 shows that the teacher-student combination [WRN-28-2, WRN-16-2] obtains the optimal results with the refining factor $r = 2$. In summary, when the total amount of knowledge is equal, the knowledge of the high-capacity teacher network can be transferred to students without refining, on the contrary, the low-capacity teacher network needs to further improve the quality of knowledge.

When the Number of Channels in the Teacher-Student Combination Does Not Match. Considering $C_T^i \neq C_S^i$, experiments have been conducted based on Table 7 for a range of different r values, where k denotes the ratio of the number of teacher and student channels. When exploring the effect of r on the algorithm, keep the other hyperparameters constant and let $\alpha = 1$, $\theta = 111$.



(a) CIFAR-10

(b) CIFAR-100

Fig. 3. On the CIFAR-10, the student gets the highest accuracy rate of 92.58% at $r = 3$ when $k = 2$, 92.58% at $r = 3$ when $k = 3$, 92.97% at $r = 4$ when $k = 4$ and 93.36% at $r = 16$ when $k = 5$. On the CIFAR-100, the student gets the highest accuracy rate of 69.53% at $r = 3$ when $k = 2$, 69.53% at $r = 3$ when $k = 3$, 69.92% at $r = 4$ when $k = 4$ and 71.09% at $r = 16$ when $k = 5$.

Figure 3(a) and Fig. 3(b) show the variation in student's accuracy for different student-teacher combinations for different r values on the CIFAR-10 and CIFAR-100, respectively. From these experiments, it can be found that as the ratio of the number of channels between teachers and students increases, the refining factor r for obtaining the best performance increases as well. However, the refining factor cannot always be increased because of the limitations of the student network. In general, when the total amount of knowledge is not equal, the larger the total amount is, the more the teacher network needs to further improve the quality of the knowledge transferred to the students. This is also consistent with our conventional perception that the larger the total amount is, the more redundancy exists. But the refining factor r cannot be increased all the time and its maximum value is $Min\left(C_T^i, C_S^i\right)$ limited by the teacher and student network architecture.

5 Conclusion

In this paper, we propose a knowledge distillation method based on attention mechanism named DRKD, which aims to dynamically select the knowledge transferred to students. This provides a novel way of thinking, where the teacher gradually guides the students to get the best answer through a question-and-answer format as a real teacher teaches the students, rather than simply instilling them with knowledge. In addition, our proposed approach deeply explores the balanced relationship between the quantity and the quality of knowledge transferred from the teacher to the student, not only laying the theoretical foundation for achieving stronger compression for small model optimization, but also improving the versatility of knowledge distillation methods for multi-structural combination situations. Finally, we testify the effectiveness of this approach and the flexibility in selecting teacher-student combinations on the CIFAR-10 and CIFAR-100.

References

1. Chen, L., et al.: Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306 (2017)
2. Gao, Z., Xie, J., Wang, Q., Li, P.: Global second-order pooling convolutional networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3019–3028 (2019)
3. Guo, M.H., et al.: Attention mechanisms in computer vision: a survey. Computational Visual Media, pp. 1–38 (2022)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
5. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 2(7) (2015)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
7. Jin, X., et al.: Knowledge distillation via route constrained optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1345–1354 (2019)
8. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: network compression via factor transfer. In: Advances in Neural Information Processing Systems 31 (2018)
9. Lee, S.H., Kim, D.H., Song, B.C.: Self-supervised knowledge distillation using singular value decomposition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 339–354. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_21
10. Li, Y., et al.: Towards compact cnns via collaborative compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6438–6447 (2021)
11. Liu, L., et al.: Group fisher pruning for practical network compression. In: International Conference on Machine Learning, pp. 7021–7032. PMLR (2021)

12. Liu, Z., Wang, Y., Han, K., Zhang, W., Ma, S., Gao, W.: Post-training quantization for vision transformer. In: Advances in Neural Information Processing Systems 34 (2021)
13. Qin, Z., Zhang, P., Wu, F., Li, X.: Fcanet: frequency channel attention networks. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 763–772 (2021)
14. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. CoRR abs/1804.02767 (2018)
15. 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)
16. Tan, M., Pang, R., Le, Q.V.: Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)
17. Tang, J., Liu, M., Jiang, N., Yu, W., Yang, C., Zhou, J.: Knowledge distillation based on positive-unlabeled classification and attention mechanism. In: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5. IEEE (2021)
18. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–11539 (2020)
19. Wang, X., Fu, T., Liao, S., Wang, S., Lei, Z., Mei, T.: Exclusivity-consistency regularized knowledge distillation for face recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 325–342. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_20
20. Wang, Z., Li, C., Wang, X.: Convolutional neural network pruning with structural redundancy reduction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14913–14922 (2021)
21. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
22. 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)
23. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 87.1-87.12. BMVA Press (2016)
24. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)