Middle-level Distillation Method Based on Multi-source Information Matching Guided by Auxiliary Classification

Yifan Zhang*

Department of computer, North China Electric Power University (Baoding), Baoding, Hebei, China

**Corresponding Author.*

Abstract

In traditional KD, the output of the large model (pseudo-label) is usually used to supervise the small model, and the small model conforms to the real label and the teacher's pseudo-label. Later work proposed feature loss as a middle-level supervisor to further mine teacher information. However, the performance of feature loss is not as strong as the output KD, and the definition and implementation of feature loss are more complicated, and it is not robust to different T model structures. A simple, novel and effective auxiliary middle-level distillation (AID) middle-level supervision method is raised, effectively enhancing the performance of the learner/S model (S model for short) under the lecturer/ T model (T model for short). Specifically, we use the auxiliary branch as the transformation of the student network, and use the feature distance, logits distance, sample pair relationship and other multi-source information matching to shrink the difference between the S model features and the T model's advanced features. Use orthogonalization and logical normalization techniques to make auxiliary branches better transfer feature knowledge. Our very novel method is the first in the KD research field to use multi-source information to match middle-level supervision. We have achieved excellent results on common benchmarks. In CIFAR100 and CIFAR10, the accuracy of 11 models increased by 5.46% and 2.49% respectively on average. In ImageNet, AID achieves 1.57 times compression and 1.81 times acceleration without loss of accuracy. As teachers for many models that perform well, we can improve student performance more effectively under the training of our methods.

Keywords: *knowledge distillation, middle-level supervision, multi-source information matching*

I. Introduction

AlexNet mentioned in the study of Krizhevsky et al. (2012) brought about extraordinary success of deep convolutional NNs (DCNN) in a variety of tasks. Nevertheless, the prerequisite of its high performance is the adequate depth or width. Deep & extensive ones require a lot of computing and memory storage, and are not fit for environments with depletable resources, including portable or embedded systems. To address this problem, numerous studies were done to improve more micro but more precise neural networks (NNs). Some of the technologies that have been established in this research field are parameter quantification or binarization (Rastegari et al., 2016), pruning (Li et al., 2016) and knowledge distillation (KD) (Hinton et al., 2015). KD has always been a hot topic to study on, as a workaround, compressing huge models or integrating models into smaller models. In [1], Hinton et al. initially transferred the knowledge from the teachers with advanced ability model to the intensively-informative-scholar model by correcting the smooth output prediction between teachers and students. This model is called "distillation" to improve students' Performance. Since then, people have invented many promising methods of knowledge distillation, using various "knowledge" to promote the optimization of the distillation system, such as intermediate representation [2], inter-layer flow [3], attention mapping [4], structure Relationship [5] and activation similarity [6].

In the traditional KD algorithm, the output of the large model (pseudo-label) is usually used to supervise the small model. This small model matches the real label and the teacher's fake label. In the follow-up work, we propose to use feature loss as a middle-level manager to further mine the information of the teacher network. However, the

overall performance of the function loss is not as robust as the output KD (refer to the result of Contrastive Representation Distillation (CRD)), and the definition and implementation are complicated. Recently, online KD has achieved good performance through mutual learning teacher and student models (TS) or self-meeting era, but at the same time the cost of guiding TS is not small, it is no longer suitable for large teacher networks, and is as special as some training methods, semi-supervision, etc., so it is only suitable use the trained model. At the same time, many models with excellent performance (85%+ on ImageNet), such as visual deformers, are not suitable for lightweight scenes due to large parameters and delays, but they can be used in the process of lightweight and small models. Own pre-training model serves as a teacher to improve student performance. Similar scenarios are also common in natural language processing (NLP) tasks. Therefore, how to use these effective and complex pre-training models to cultivate S models and maximize the use of information has become an urgent and important issue.

In this method, we add anti-branch in the middle layer of students, support teachers' pseudo-labels on corresponding potential features, and provide auxiliary intermediate supervision. Compared with self-distillation auxiliary classifiers, this method can optimize the distribution of classifiers and provide correct monitoring information more effectively. At the same time, we use Norm KD to make teachers' pseudo-labels smoother and more stable, and solve various problems that previously caused better teachers to not improve their scores. We have achieved excellent results in the general CRD benchmark tests, and for many models that perform well on ImageNet (more than 85%, such as visual converters for teachers), we can more effectively conduct methodological training to let learner show better achievements. The whole work is the first comprehensive introduction to the new method of inverse branch coding teacher marking the middle distillation, and compared with the existing method. We have provided proof and analysis for applications that provide models (such as visual converters) in lightweight scenarios.

Sufficient experiments show that this method can achieve consistent and significant accuracy improvement on various NNs and data sets. Experiments conducted on 10 NNs in 5 data sets show that auxiliary middle-level distillation (AID) has greater advantages in both image classifications. The average accuracy of CIFAR100, CIFAR10 and ImageNet datasets increased by 5.46%, 1.71%, 1.18%, 1.25% and 0.82%, respectively. In addition, ablation studies and hyperparameter sensitivity studies were also carried out to prove the effectiveness and stability of AID.

II. Work Related

In this part, we make a concise summary and summary of the existing methods of knowledge distillation.

2.1 Knowledge Distillation (KD)

Being a commonly-adopted technological ways in model compression, KD is defined by Hinton et al. as addition of the loss function (hereinafter LF) of KD to the training, and consequently, the S model can effectively simulate the output of the teachers' one. Many methods were implemented improve the learning efficiency of the S model. Romero et al. first raised up FitNet [7], which involves clue learning into minimizing the difference of the feature maps between learners and lecturers. Agoruyko considers this issue in terms of attention mechanism, trying to unify the characteristics of attention area. Knowledge distillation has also shown promise in other fields. Furlanello interactively assimilates the refined S model into a T model set, and obtains better generalized test data from the model set. Bagherinezhad applies knowledge refinement to data authentication, adding up the label to an entropy with larger number. Paper not used knowledge refinement to resist confrontational attacks, and Gupta did exactly the identical to in transferring knowledge between different modes of data [8]. Among the above methods, generally speaking, both S model and T model operate independently, and the knowledge transfers from one model to the other. In comparison, in our self-refinement method, both the S model and the T model come from the same CNN, which is a true self-distillation framework.

2.2 Auxiliary supervision

In order to urge the pace of convergence and cope with the gradient disappearance, the monitoring signal is more directly transmitted to the hidden layer through the assistance established on the intermediate layer. GoogleNET [9] and DSN [10] are two parallel projects that use this advanced monitoring method. They build primitive deep NNs on basic image classification tasks. Recently, it has been used in some extra visual identity (VI) tasks, including edge detection [11], human pose estimation [12], scene analysis [13], semantic segmentation [14], key point positioning [15], automatic contour [16] , Travel time estimate [17]. Despite these recent advances in new applications, auxiliary classifiers are rarely used in modern CNN classification models. To address the problem of defining and measuring the knowledge that will be imparted from the lecturer, the existing models usually take into account the lecturers' class probability [18] and intermediate features [19]. What merits the attention is that the auxiliary learning strategy put forward in this study takes re-sharing to aid optimization, and its motivation and the KD method are unlike. In this method, we use an auxiliary classifier to transform the middle-level features of the network and transform these features into probability distributions, making it easier to match with KL.

III. Method

3.1 Overview

We use auxiliary branches, which support orthogonalization and logical normalization, better conversion of student networks, and use KL distance to narrow the gap between the differences between the characteristic S model and the teacher's advanced characteristic model. The part of the supervision of the characteristics of the lecturer network on the characteristics of the learner network is removed, and the output of the former is directly used to supervise the characteristics of the latter.

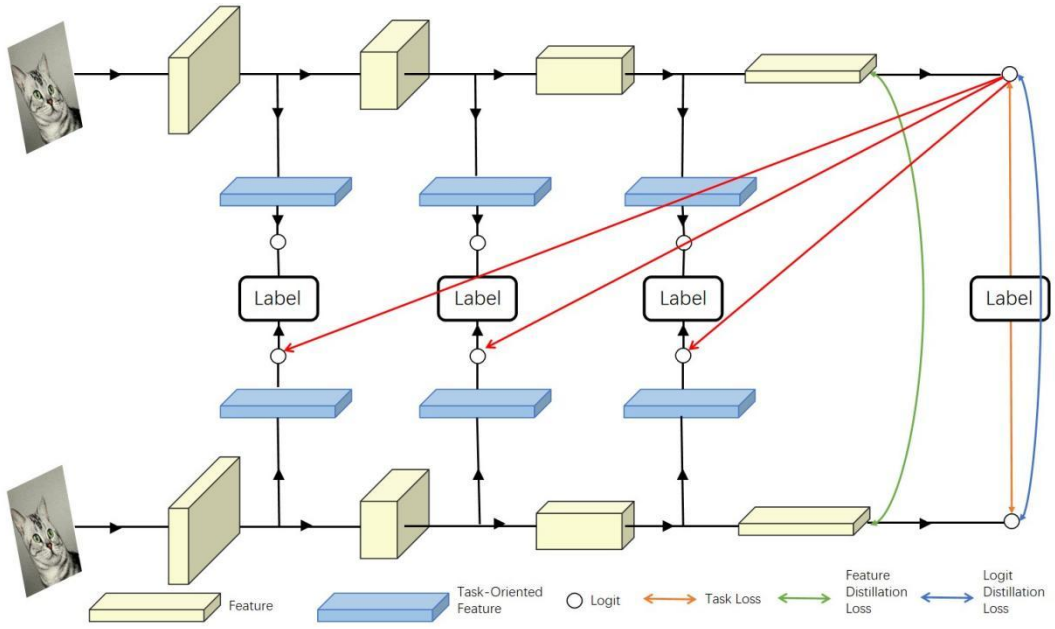

Fig. 1 overall framework of the methods.

3.2 Re-examine the Feature distillation method

Enlightened by the study of Fitnets et al. (2014), the L2 gap among computed characteristic mapping in front of the final Fully connected layer (FC) was calculated. For one thing, tips lost also provides knowledge of the inchoate classifier, which helps convergence classifier. For another, Mirzadeh et al. (2019) pointed out that when students do not have enough abilities or mechanisms to imitate how the teachers behave, knowledge distillation may be

ineffective. The implied loss compels the student to get closer to the weight distribution of the lecturers, that is, to describe the divide between the lecturer and the learner.

$$\text{loss}_{3} = \sum_{i=1}^{N-1} \|F_i - F_N\|_2^2 \quad (1)$$

In formula (1), F_i stands for the characteristic mapping before the FC layer.

This section examines the graph components of feature extraction techniques used to accomplish network compression and raise new insights to improve the method, which shows better result than the previous one. At the very beginning, the loss function (LF) of featured distillation is described. As Fig. 1 demonstrates, feature of teacher (FT) is the note for characteristics of teacher network, and that of the student is denoted by feature of student (FS). For realizing fit the feature dimensions T_t and T_s, the features F_t and F_s are to be invariant. The divide between the characteristics before and after the transformation, d, is taken to be the LF $L_{distill}$. Namely, its characteristic distillation is generalized as,

$$L_{\text{distill}} = d\big(T_t(F_t), T_s(F_s)\big) \quad (2)$$

Students are nurtured by minimizing the distillation loss (hereinafter DL) $L_{distill}$ in the distillation process. It is best to plot the DLs to convey all characteristic information except the lack of any necessary teacher information. For this reason, a new characteristic of DL is raised. In this, all necessary information is transmitted from the teacher to improve the distillation performance as much as possible. To this end, the layout factor of characteristic distillation loss (CDL) is analyzed. As the following table 1 tells, the sketch components of CDL are divided into four categories, namely teacher's transformation, students' transformation, feature position as well as distance function.

Table 1: Four types of sketch components of CDL.Distillation distinguishes itself from various characteristics. Most distillations use teacher switching with information loss.

Method	Teacher transform	Student transform	Distillation feature position	Distance	Missing information
FitNets	None	1×1 conv	Mid layer	L_2	None
AT	Attention	Attention	End of group	L_2	Channel dims
FSP	Correlation	Correlation	End of group	L_2	Spatial dims
Jacobian	Gradient	Gradient	End of group	L_2	Channel dims
FT	Auto-encoder	Auto-encoder	End of group	L_2	Auto-encoded
AB	Binarization	1×1 conv	Pre-ReLU	Marginal L_2	Feature values
Proposed	Auxiliary classifier	Auxiliary classifier	End of group	Characteristic distance+ Distribution distance+ Sample distance	None

Teachers transformation. Teacher transformation T_t transforms the hidden characteristics of teachers into a form that can be easily transformed. As one significant part of characteristic distillation, it accounts for information lost in the process. Since there are both favorable and unfavorable information in the features, we use marginal ReLU in the method proposed in this paper, that is, positive (favorable) information is utilized free from any conversion while the negative counterpart (unfavorable) is suppressed. The results show that this method can be calibrated without losing useful information.

Student transformation. On this occasion, the student's feature size does not present a trend of declining, but increases, so no information is lost. In the method given by this study, we use this asymmetric transformation form as a student transformation.

The distance functions. An applicable distance function is supposed to build in accordance with the teacher's conversion and distillation spot pre-ReLU position pre-ReLU information transfer from the lecturer to the learner, but it contains the negative value of pre-ReLU characteristic, as the activation of ReLU shielding bad information, which is not used by the teacher of the network. To solve this problem, all the value transfer may have a negative impact on student networks. A new distance function, called sectional L2 range, is proposed, out of the purpose of leaving out the distillation of information in negative regions.

3.3 Auxiliary intermediate distillation

In formula (1), $F_t - F_s$ of layer to layer in traditional networks is as follows,

$$(T_1 - S_1) + (T_2 - S_2) + (T_3 - S_3) \tag{3}$$

Recently there has been plenty of work that has found the top of the T model has more semantic information, so monitor of T_3 as cross-level is as follows,

$$F(T_3 - S_1) + F(T_3 - S_2) + F(T_3 - S_3) \tag{4}$$

But there are large gaps in semantics and dimensions between such cross-level supervision, so we use the auxiliary classifier as the transformation of S_1, S_2, S_3, which is better than the probability score of T_3,

$$KL(T_3 - S_1) + KL(T_3 - S_2) + KL(T_3 - S_3) \tag{5}$$

Let $\chi = \{x_i\}_{i=1}^{m}$ be a group of training images and $y = \{y_i\}_{i=1}^{m}$ be the corresponding labels. $F_i(\cdot)$ is the feature graph of i_{th} convolution block, and $c_i(\cdot)$ is the fully connected classifier in i_{th} convolution block. Superscripts T and S represent the lecturers' model and learners' one, respectively. In a neural network with N convolution blocks, the Logit distillation [20] loss can be written as,

$$\frac{1}{m}\sum_{i=1}^{m} \cdot L_{KL}\left(c_N^8\left(F_N^8(x_i)\right), c_N^t\left(F_N^t(x_i)\right)\right) \tag{6}$$

where L_{KL} is the scattering loss of KL. The LF of characteristic distillation can be got by the following formula (6),

$$\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{N} L_2\left(T_j\left(F_j^8(x_i)\right), T_j\left(F_j^t(x_i)\right)\right) \tag{7}$$

where L_2 is the norm loss of L_2 and T is the transformation function of their characteristics. In the characteristics extraction techniques applied in the current literature, T represents non-parametric transformation, for instance, pooling and low-rank integration. In comparison, the T in the presented AID contains numerous CONV layers whose parameters are drilled by task losses and DL. The raised task-aimed CDL can be expressed as,

$$\frac{1}{m}\sum_{i=1}^{m}\{\sum_{j=1}^{N}\alpha \cdot \underbrace{L_2\left(T_j\left(F_j^8(x_i)\right), T_j\left(F_j^t(x_i)\right)\right)}_{\mathcal{L}_{\text{feature}}} + \underbrace{L_{CE}\left(c_j\left(T_j\left(F_N^8(x_i)\right)\right), y_i\right)}_{\mathcal{L}_{\text{task}}}\} \tag{8}$$

where α is the super parameter to balance the two losses. In addition, the Logit DL is brought in, serving as an assistance trainer for T and all-connected layer C. The relation is shown in the following formula (8).

$$\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{N} \underbrace{L_{KL}\left(c_j^8\left(T_j^8\left(F_N^8(x_i)\right)\right), c_j^t\left(T_j^t\left(F_N^t(x_i)\right)\right)\right)}_{\mathcal{L}_{logit}} \tag{9}$$

3.4 Sample pair relationship matching

A group in small scale containing N data points $\{x_i\}_{i=1:N}$ is preset, and each data point is independently transformed $t(\cdot)$ (sampled from the same distribution τ) to obtain $\{\tilde{x}_i\}_{i=1:N}$. Both x_i and $\tilde{x}_i$ are input into the lecturer or learner network to extract $\phi_i = f(x_i)$. $\tilde{\phi}_i = f(\tilde{x}_i)$ denotes inserting a projection head (a perceptron with two layers or more) at the top of the network. It will represent the potential space mapped to an application contrast loss, that is, $z_i = \mathrm{MLP}(\phi_i), \tilde{z}_i = \mathrm{MLP}(\tilde{\phi}_i)$.

We take $(\tilde{x}_i, x_i)$ as a positive pair and $(\tilde{x}_i, x_k)_{k\neq i}$ as a negative counterpart. Given a certain $\tilde{x}_i$, the task for comparison and anticipation is to pick out the corresponding x_i from the set of $\{x_i\}_{i=1:N}$. In order to realize it, the similarity between positive pairs should be maximized and vice versa. The similarities in cos value is taken in this research. If the similarities between $\{\tilde{x}_i\}$ and $\{x_i\}$ are organized into matrix form $\mathcal{A}$, it is supposed to get that

$$\mathcal{A}_{i,j} = \mathrm{cosine}(\tilde{z}_i, z_j) = \frac{\mathrm{dot}(\tilde{z}_i, z_j)}{\|\tilde{z}_i\|_2 \|z_j\|_2} \tag{10}$$

where $\mathcal{A}_{i,j}$ stands for the resemblance between $\tilde{x}_i$ and x_j.The lack of comparison anticipation is

$$L = -\sum_i \log\left(\frac{\exp(\mathrm{cosine}(\tilde{z}_i, z_i)/\tau)}{\sum_k \exp(\mathrm{cosine}(\tilde{z}_i, z_k)/\tau)}\right) = -\sum_i \log\left(\frac{\exp(\mathcal{A}_{i,i}/\tau)}{\sum_k \exp(\mathcal{A}_{i,k}/\tau)}\right) \tag{11}$$

In this equation, τ is another temperature parameter. Being softmax-like loss, it may be comprehended as probability maximum of $\tilde{z}_i$ and z_i in a positive pair. When matching $\{\tilde{x}_i\}$ and $\{x_i\}$, the network learns to transform invariants. However, in SSKD, the prime objective does not lie in the transformation of invariant representations. Instead, regarding contrast anticipation as an auxiliary to mine more abundant knowledge from the T model is.

3.5. Orthogonal loss

In most distillation cases, the teacher and student feature sizes are different and their distance cannot be directly minimized. For the sake of alleviating this, the CONV layer or the all-connected layer is given up because either it or the fully connected one would trigger the loss of characteristic information of lecturer. Represents the distilled feature of the T model as vector x, the weight of the characteristic adjustment layer is W, and the adjusted characteristic will be noted as Wx. In order to retain characteristic information during adjustment, based on Bansal et al. [21], an orthogonality loss is drawn in, which concurrently penalizes the orthogonality of row space and column space spanned by W in feature adjustment layer, i.e., the loss is set as,

$$\beta \cdot \underbrace{(\|W^T W - I\| + \|WW^T - I\|)}_{\mathcal{L}_{\mathrm{othogonal}}} \tag{12}$$

In this formula, β is a hyperparameter to maintain the equilibrium between its size and losses in various forms. If a CONV layer rather than an all-connected layer is used to adjust its characteristics, the weight will first be changed from $S \times H \times C \times M$ to $SHC \times M$, S, H, C, M are width, height, number of channels of input tunnel and output one respectively. In brief, the overall function for loss can be summarized as,

$$\mathcal{L}_{\mathrm{overall}} = \mathcal{L}_{\mathrm{featute}} + \mathcal{L}_{\mathrm{logit}} + \mathcal{L}_{\mathrm{task}} + \mathcal{L}_{\mathrm{orthogonal}} \tag{13}$$

The total LF includes DL, Logit distillation loss, task loss, orthogonal loss and two hyperparameters. Section 6 will describe ablation studies and sensitivity studies to demonstrate their effectiveness and sensitivity.

IV. Experiment

The raised method in the paper on CIFAR-100 is assessed in this part and its performance with existing knowledge distillation methods is compared. In order to compare objectively, the public codes of different KD methods were adopted, and the same training and data preprocessing settings throughout the experiment was also used. All

experiments are implemented with PyTorch [22]. In our DIL, we added Conv1×1 to each connection path because we found that it slightly improves the distillation performance (see the Ablation Research section). Based on DIL, we also apply another integrated distillation to the output log (logit) of the two heads of the merged model to further improve the performance of each head, which is called DIL*.

4.1 Implementation

Dataset. CIFAR-100 [23], containing 50,000 training images and 10,000 test images in 100 classes, is the most popular classification data set for evaluating the performance of knowledge distillation methods.

Image classification experiments use ResNet [24], PreActResNet [25], SENet [26], ResNet [27], MobileNetV1 [28], MobileNetV2 [13], ShufflfleNetV1 [29], ShufflfleNetV2 [30], WideResNet [31], 9 convolutional neural networks and CIFAR100, CIFAR10 [32], ImageNet [18], three data sets. In the CIFAR experiment, the SGD optimizer was used to train each model for 300 epochs, and the batch size was 128. In the ImageNet experiment, the SGD optimizer is employed to drill each model for 90 epochs, and the batch size is 256. Comparative experiment. Four knowledge distillation methods, KD [33], FitNet [34], DML [35] and self-distillation [36] are used for comparison. All these experiments are done by ourselves.

Table 2: Experimental results of CIFAR100 and CIFAR10 (Top-1 Accuracy /%). Numbers in bold are the highest.

Model	Baseline	KD	FitNet	DML	SD	AID
ResNet18	94.25	94.67	95.57	95.19	95.87	96.92
ResNet50	94.69	94.56	95.83	95.73	96.01	96.84
PreactResNet18	94.20	93.74	95.22	94.80	95.08	96.49
PreactResNet50	94.39	93.53	94.98	95.87	95.82	96.93
SEResNet18	94.78	94.53	95.64	95.37	95.51	96.80
SEResNet50	94.83	94.80	95.31	94.83	95.45	97.02
ResNeXt50-4	94.49	95.41	95.78	95.41	96.01	97.09
MobileNetV1	90.16	91.70	90.53	91.65	91.98	93.93
MobileNetV2	90.43	92.86	90.49	90.49	91.02	93.34
ShuffleNetV1	91.33	92.57	92.23	91.40	92.47	92.73
ShuffleNetV2	90.88	92.42	91.83	91.87	92.51	93.74

Table 3: The results of the experiment on ImageNet(Top-1 Accuracy /%).

Model	Baseline	AID	MAC(G)	Param(M)
ResNet18	69.76	70.92	1.82	11.69
ResNet50	76.13	77.52	4.11	25.56
ResNet101	77.37	78.64	7.83	44.55
ResNet152	78.31	79.21	11.56	60.19
ResNeXt50-32-4	77.62	78.93	4.26	25.03
WideResNet50-2	78.47	79.52	11.43	68.88

Main results of CIFAR10 and CIFAR100. The table 2 indicates the precision of the student network on CIFAR100 and CIFAR10. It can be concluded that: (a) In comparison to the baseline model, the newly-constructed AID can significantly improve the accuracy. In CIFAR100, the accuracy of 11 models increased by 5.46% on average, the highest accuracy of SENet50 was 6.75%, and the lowest accuracy of ShufflfleNetV1 was 3.78%. In

CIFAR10, the accuracy of 11 models increased by 2.49% on average, the largest being 3.77% of MobileNetV1, and the smallest being 1.40% of ShufflfleNetV1. (b) In all models, the proposed AID is much better than the sub-optimal distillation method. The precisions of CIFAR100 and CIFAR10 were 3.13% and 1.28% higher than that of the sub-optimal distillation method, respectively. (c) The proposed AID is not only applicable to hyperparameter models such as ResNet and SEnet, but also in ShufflfleNet such as MobileNet. On the CIFAR100 and CIFAR10 datasets, the accuracy of the lightweight model increased by 4.40% and 2.74% on average.

Main results of ImageNet. The table 3 tells the experimental outcomes of AID on ImageNet. In all these experiments, the ResNet152 model was used as the T model. We observe (a) AID improves accuracy by 1.18% on average among 6 NNs. (b) The distilled ResNet50 and ResNet101 present more precise result than the baseline ResNet101 and ResNet152, respectively. By replacing the compressed ResNet50 and ResNet101 with ResNet101 and ResNet152, AID achieves 1.57 times compression and 1.81 times acceleration without loss of accuracy.

4.2 Ablation study

This part includes isolating the influence of per element in the proposed method and compare with possible variants. All experiments are operated on the CIFAR-100. AID instead of AID* is adopted, and will not use learning rate warm-up in all experiments for better ablation studies. For each set of experiments, the method was operated 3 times and the precision of the first "average standard" was reported.

Auxiliary classifier position. We explored the influence of the position of the auxiliary classifier. This is very important because different locations have different semantics and robustness. We consider each setting by adding up to three blocks (including blocks Conv2_x, Conv3_x, and Conv4_x, denoted as C2, C3, and C4, respectively) on CIFAR100. The specific research outcomes are demonstrated in Table 4. It is discovered that adding connections in Conv2_x and Conv3_x at the same time can bring the best performance improvement, followed by only adding connections in Conv2_x. These results show that adding connections in shallow layers such as Conv2_x and Conv3_x can transfer information well, while Conv4_x extracts higher-level semantics and therefore may have lower robustness, which limits the effect of connection supervision. To some extent, this is consistent with multi-task learning and multi-branch network design [19], where the shallow network is shared, and the high-level network is divided into separate branches.

Table 4: The connection position of the auxiliary classifier.Conv2_x, Conv3_x, and Conv4_x, denoted as C2, C3, and C4, respectively).

Connection	KD	FitNet	DML	SD	AID
Baseline	67.84	68.47	68.82	68.98	71.36
C2	86.12	86.48	87.76	88.34	90.38
C3	86.36	86.87	88.38	89.26	90.53
C4	75.79	76.44	78.61	79.36	80.78
C2&C3	87.68	88.42	88.85	89.42	93.76
C2&C4	83.26	83.98	84.79	86.10	86.83
C2&C3&C4	84.20	84.67	86.16	86.79	87.46

Auxiliary classifier transformation. We compared the performance when using direct connections and adding a Conv1×1 layer to each connection to improve the alignment in feature semantics. We observe that adding Conv1×1 can get slightly higher results than direct connection. This means that without any direct connection of Conv1×1, promising knowledge distillation performance can still be obtained. This is mainly due to our first auxiliary teacher-generated module. By using dynamic additive convolution instead of normal convolution, the teacher has a good alignment with the students in the depth of the network, and the I-O feature dimensions of every convolutional layer in teacher-student network are exactly the same.

4.3 Deeper Analysis

Few-Shot scene. The sample available to be trained is restricted in substantial world. For figuring out how SSKD performs in a few scenes, experiments on a subset of CIFAR100 were carried out. Images from each class were picked out in a random way to create a new set for operation and training. While keeping the same test set, the newly created training set is taken to drill the S model. Vgg13 and vgg8 were taken as T model and S model respectively. We compare the performance of our students with KD [22], in FT [37] and CRD [38]. The proportions of reserved samples are 25%, 50%, 75% and 100% respectively. To offset other intervention factor so as to realize a rather just comparison, the identical data in different methods were used. The precision of the CIFAR100 test set in a scene with few shots and noisy tags. (a) Train students on a subset of CIFAR100. SSKD turned out to be the best in all situations. This advantage is particularly obvious when only 25% of the training data is available. Learners receive training on data with interference labels. The precision of FT and CRD decreases sharply with the increase of noise tags, while SSKD is steadier and better-behaved in all cases. Among all the data ratios, SSKD achieved the best results. With the reduction of training samples, the advantages of our method become more and more obvious. For example, when the proportion of retained samples is 25%, the accuracy rate is definitely improved by 7% compared with all competing methods. The previous methods are mainly to learn a variety of intermediate characteristics of teachers or explore the relationship between samples. Over-imitation will lead to over-fitting of the training set. In SSKD, the varied images and self-supervised tasks enable the S model to have structured knowledge and strong regularization, so that the model shows more excellent generalization of the test set.

Noisy-Label scene. This course requires students to imitate the teacher in category classification tasks and self-supervision tasks. Students can learn more comprehensive knowledge from the T model than counting solely on annotation tags. This tacit enhances the capacity of the S model to deal with label noise. This part studied the how KD [22], FT [37], CRD [38] and SSKD performs with operation under the scene of noisy label data. Vgg13 and vgg8 were selected as T and S models. It is assumed that the teacher has received pure data training and all learners share the same teacher, and this hypothesis will cause little error on the robustness test on KD. When training the S model, the labels of certain parts of the training data are randomly perturbed and the original test data were selected for evaluation. The identical interference to all methods were drawn in. Considering the loss weight of the cross-entropy on the annotated label affects the ability of the model to resist label noise, one loss weight is adopted for all methods. The percent interference tags are valued to 0%, 10%, 30%, and 50%. SSKD is superior to competing methods in all noise ratios. As the noise data increases, the outcome of FT and CRD drops sharply. KD and SSKD were seen steadier trend. Specifically, when the percentage of noisy data increases from 0% to 50%, the accuracy of SSKD only drops by 0.45%, indicating the robustness of SSKD to noisy data labels. The robustness can account for the structured knowledge provided by self-monitoring tasks.

V. Different from existing methods

In this paper, a novel middle-level supervision method based on multi-source information matching is proposed, which is different from the previous methods as follows:

1. Compared with the feature distillation method: The auxiliary classification we introduced, mapping the corresponding feature distillation to the supervised matching of logits and the sample-to-relationship, resulted in a better effect, which is currently not available in all such methods.
2. Compared with the auxiliary distillation method, the current auxiliary distillation method can only be used in the Online KD scenario. For the first time, we have realized the introduction of middle-level supervision under two-stage distillation. Compared with the previous method, the method is more novel and requires less training costs. At the same time, we are also the first work to introduce multi-source information supervision to obtain the best results so far.

In short, as far as we know, our very novel method is the first in the KD research field to use multi-source information to match the middle-level supervision work, and the effect exceeds the existing methods.

VI. Conclusion

This study raises a new KD method to assist intermediate distillation. It designs and develops a framework for knowledge distillation from a new perspective. Additive convolution is used to automatically generate auxiliary classifiers, and a dense feature connection surface is established to improve network performance through the back propagation of gradient flow. This method does not need to define the loss of distillation and to drill a lecturer network in advance is undesired. We hope that our work will stimulate future research on the design of knowledge distillation.

Reference

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.
[2] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In International Conference on Learning Representations, 2015.
[3] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4133–4141, 2017.
[4] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional NNs via attention transfer. In International Conference on Learning Representations, 2017.
[5] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3967–3976, 2019.
[6] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In Proceedings of the IEEE International Conference on Computer Vision, pages 1365–1374, 2019.
[7] Ko?Ak M . FITNET Fitness-for-Service Procedure: An Overview. Welding in the World, 2007, 51(5-6):94-105.
[8] Yu R, Li A, Morariu V I, et al. Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation. 2017.
[9] Wu C , Wen W , Afzal T , et al. A Compact DNN: Approaching GoogLeNet-Level Accuracy of Classification and Domain Adaptation// Computer Vision & Pattern Recognition. IEEE, 2017.
[10] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2014.
[11] Lopes R G, Fenu S, Starner T. Data-Free Knowledge Distillation for Deep NNs. 2017.
[12] Fukuda T , Suzuki M , Kurata G , et al. Efficient Knowledge Distillation from an Ensemble of Teachers// Interspeech 2017. 2017.
[13] Li Yaochen, Zhu Chao, Liu Yuehu, Hong Yuhui, Wang Jianji. Geometric and semantic analysis of road image sequences for traffic scene construction. Neurocomputing, 2021, 465.
[14] Wen Yang, Chen Leiting, Deng Yu, Ning Jin, Zhou Chuan. Towards better semantic consistency of 2D medical image segmentation. Journal of Visual Communication and Image Representation, 2021, 80.
[15] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806, 2014.
[16] Saining Xie, Ross Girshick, Piotr Dolla ŕ, Zhuowen Tu, Kaiming He. Aggregated residual transformations for deep NNs. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500, 2017.
[17] Ramin Saedi, Mohammadreza Saeedmanesh, Ali Zockaie, Meead Saberi, Nikolas Geroliminis, Hani S. Mahmassani. Estimating network travel time reliability with network partitioning. Transportation

Research Part C, 2020,112.
[18] Yim J, Joo D, Bae J, et al. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[19] Park Jeongsoo, Kim Jungrae, Ko Jong Hwan. Auto-Tiler: Variable-Dimension Autoencoder with Tiling for Compressing Intermediate Feature Space of Deep Neural Networks for Internet of Things. Sensors, 2021, 21(3).
[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In NeurIPS, 2014.
[21] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In Advances in Neural Information Processing Systems, pages 4261–4271, 2018.
[22] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 413–420. IEEE, 2009.
[23] Nicolas Pinto, Zak Stone, Todd Zickler, and David Cox. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In CVPR 2011 WORKSHOPS, pp. 35–42. IEEE, 2011.
[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1(8): 9, 2019.
[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.
[26] Dat Thanh Tran, Alexandros Iosifidis, and Moncef Gabbouj. Improving efficiency in convolutional NNs with multilinear filters. NNs, 105:328–339, 2018a.
[27] Geethu S, Vimina E R. Improved 3-D Protein Structure Predictions using Deep ResNet Model.. The protein journal, 2021.
[28] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In International Conference on Learning Representations, 2020.
[29] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856.
[30] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.
[31] Zagoruyko S, Komodakis N. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016.
[32] Li H, Liu H, Ji X, et al. Cifar10-dvs: an event-stream dataset for object classification. Frontiers in neuroscience, 2017, 11: 309.
[33] Kim Y, Rush A M. Sequence-level knowledge distillation. arXiv preprint arXiv:1606.07947, 2016.
[34] Nijhof S L, Bleijenberg G, Uiterwaal C S P M, et al. Effectiveness of internet-based cognitive behavioural treatment for adolescents with chronic fatigue syndrome (FITNET): a randomised controlled trial. The Lancet, 2012, 379(9824): 1412-1418.
[35] Grobauer B. Cost recurrences for DML programs. ACM SIGPLAN Notices, 2001, 36(10): 253-264.
[36] Zhang L, Song J, Gao A, et al. Be your own teacher: Improve the performance of convolutional neural networks via self distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3713-3722.
[37] Corbesier L, Vincent C, Jang S, et al. FT protein movement contributes to long-distance signaling in floral induction of Arabidopsis. science, 2007, 316(5827): 1030-1033.
[38] Morris J C. The clinical dementia rating (cdr): Current version and. Young, 1991, 41: 1588-1592.