Student Customized Knowledge Distillation: Bridging the Gap Between Student and Teacher

Yichen Zhu, Yi Wang
Midea Group
zhuyc25, wangyi55@midea.com

Abstract

Knowledge distillation (KD) transfers the dark knowledge from cumbersome networks (teacher) to lightweight (student) networks and expects the student to achieve more promising performance than training without the teacher's knowledge. However, a counter-intuitive argument is that better teachers do not make better students due to the capacity mismatch. To this end, we present a novel adaptive knowledge distillation method to complement traditional approaches. The proposed method, named as Student Customized Knowledge Distillation (SCKD), examines the capacity mismatch between teacher and student from the perspective of gradient similarity. We formulate the knowledge distillation as a multi-task learning problem so that the teacher transfers knowledge to the student only if the student can benefit from learning such knowledge. We validate our methods on multiple datasets with various teacher-student configurations on image classification, object detection, and semantic segmentation.

1. Introduction

Deep neural networks have achieved state-of-the-art results in a variety of applications such as computer vision [20], speech recognition [1], and natural language processing [6, 30]. Although it is established that introducing more computational costs often improves the performance of the models, big models are computationally too expensive to be deployed on devices, which only limited computational resources are available such as mobile devices and embedded devices. Model compression techniques have emerged to address such issues, and knowledge distillation [12] has proven to be a promising way to obtain a small model without significant performance loss among those techniques. It works by encouraging a lightweight student model to mimic the behavior learns by a cumbersome teacher model.

For the success of knowledge distillation, some researchers have focused on the relation between students and teachers. These works challenge a common intuition that better teachers make better students. Mirzadeh *et.al.* [24] found out that students distilled from a bigger teacher, one with more parameters and higher accuracy,

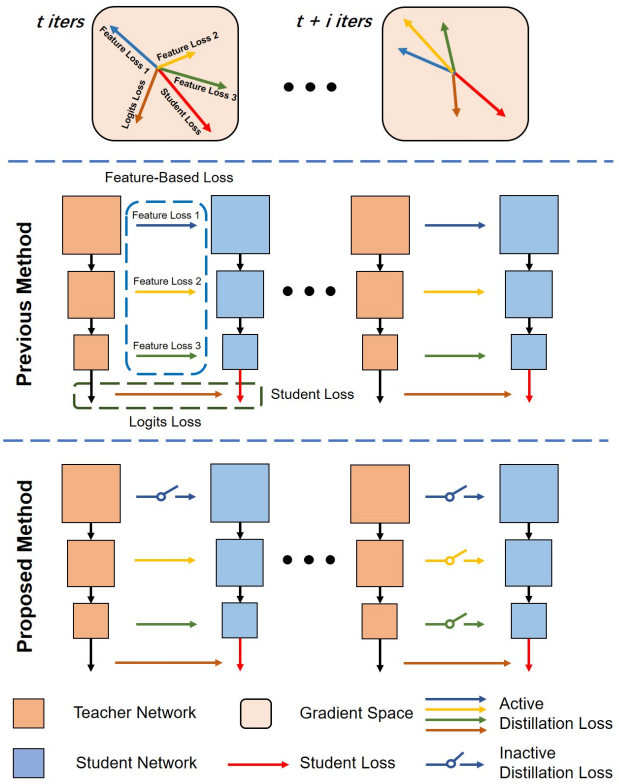

Figure 1: Best view in color. **Top**: The gradient similarity of knowledge distillation and student loss at different iterations in the gradient space. **Middle**: Prior approaches. The knowledge distillation process between two networks is stationary in different iterations. **Bottom**: Our approach automatically decides to switch on or switch off the knowledge distillation loss based on their corresponding relative gradient direction to student loss.

can perform worse than the same students distilled from a smaller teacher. Cho *et.al.* [2] also discovered the same phenomenon, and it is even more severe when training on a large-scale, challenging dataset such as ImageNet. Both works conclude that the student and the teacher's capacity mismatch is the reason for the negative correlation between teachers' accuracy and students' performance. Meanwhile, other works also show that the same teacher model [9, 36] or even teacher with lower performance [34, 25, 35] than the student model can be used as the teacher model to perform knowledge distillation, which gives rise to self-distillation methods [5, 38, 39].

Though the "better teacher, worse student" contradiction has been discovered, how to resolve this issue is still rarely explored. TAKD [24] present to use teacher assistant, which is a smaller teacher, as a media to smooth the knowledge transfer procedure between large teacher and small student. ESKD [2] propose an early stop strategy during the knowledge distillation process, which reduces the negative effect of KD. However, these methods require manual tuning. When the student model changes, these methods need to carefully choose either a teacher assistant model or an appropriate early stopping criterion to balance the trade-off between the positive and negative effects that are brought by the teacher's knowledge.

In this paper, we tackle this issue from the perspective of gradient similarity between the teacher and the student during the KD training process. We first analyze that the capacity mismatch does not continuously happen in the training stage by checking the network representation similarity. We then formulate the knowledge distillation as multi-task learning and present an adaptive knowledge distillation approach that can be adapted based on target student model, named as *Student Customized Knowledge Distillation (SCKD)*. As a result, the SCKD performs different knowledge distillation strategies for different student models and ideally allocates the optimal knowledge transfer process. The Figure 1 shows a comparison of SCKD and prior approaches. Our framework makes no restrictions on the number of knowledge, knowledge types (i.e., single teacher, multi-teachers, or self-distillation), and place to perform distillations (i.e., on intermediate representation or output space). It can plugin into any existing knowledge distillation framework and improve the student performance immediately. Additionally, our approach is applicable to a variety of vision tasks, including image recognition, object detection, and semantic segmentation. Our contributions are summarized as follows:

- We propose an adaptive knowledge distillation, named as Student Customized Knowledge Distillation (SCKD) method. The SCKD can automatically adjust the KD process based on the target student model, which is achieved by calculating the gradient similarity between the teacher's distillation loss and student loss during training.

- The proposed SCKD shows evident advantages over conventional knowledge distillation approaches on various visual tasks, and show immediate performance improvement by inserting into existing knowledge distillation framework.

2. Related Work

Various knowledge distillation methods [29, 10] were introduced in recent years, including transferring output probability distribution (output-based knowledge) [12, 33], intermediate feature representation (feature-based knowledge) [26], and their variants [32]. These works pay attention to what knowledge to distill. Another line of researches focuses on how to distill the knowledge (i.e., self-distillation). Mobahi *et.al.* [25] theoretically prove that under the infinite width setting, the self-distillation amplify the regularization [15]. Meanwhile, Yuan *et.al.* [34] empirically show that weak teacher can improve student performance, and its behavior is similar to label smooth. Ji *et al.* [17] and Jang *et al.* [16] argues that fixed linked for intermediate feature distillation is suboptimal. They present an alternative approach, which adopt meta-learning and attention feature to finding the optimal linkage and feature matching strategy for feature-based knowledge distillation, respectively. Some works study the knowledge transfer process on multi-teacher KD approaches, such as multiple heterogeneous teachers [28] or homogeneous teachers [7, 23].

The most correlated works to us are the researches on discussing the relation between teacher and student. Better teachers do not make better students is a counter-intuitive argument, which was first observed by Cho *et al.* [2] and Mirzadeh *et al.* [24], who assume and prove that it happens because of the mismatched capacity between teachers and students through a series of empirical analysis. Based on the assumption, they provide solutions to fix this issue: Cho *et al.* [2] argue that the knowledge distillation process can be trained with an early stopping strategy. Mirzadeh *et al.* [24] present a teacher assistant knowledge distillation (TAKD), which is a smaller vision of the teachers to distill the student. This is inspired by BAN [24] where the teacher is required to distill knowledge to students progressively through multiple generations. These prior works have shed light on the teacher-student relation; however, their solutions require manually tuning the teacher assistant model or early stopping criterion—these settings need to refresh when the student model changes. Contrary, our approach can automatically adjust the training strategy regarding different student models.

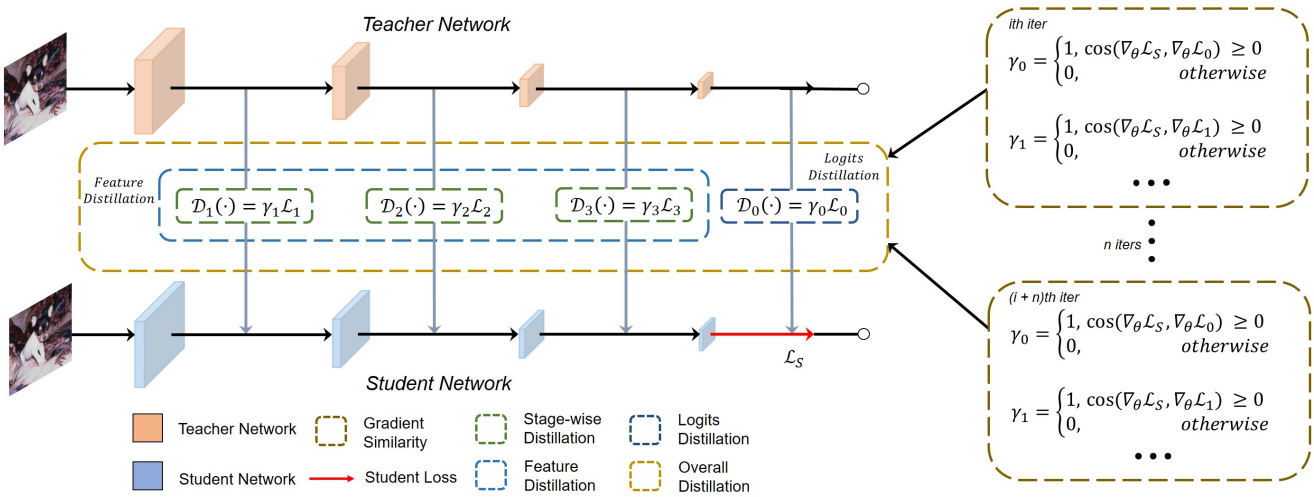

Figure 2: Best view in color. The overview of SCKD. At every iteration, the connection for the distillation loss, both feature distillation loss and logits distillation loss, are determined by the gradient similarly.

3. Method

3.1. Preliminaries

We formally introduce symbols and notations in this subsection. We illustrate two types of knowledge distillation (KD) method: output knowledge KD and feature knowledge KD.

Given a teacher model T and a student model S, we denote the output of two networks as p^T and p^S. Then KD encourages that the output of the student model mimics the output of the teacher model by minimizing the following objectives:

$$\mathcal{L}_{out} = \alpha_0 \mathcal{H}(p^S, p^T) \tag{1}$$

where p is the output, $\mathcal{H}(\cdot)$ is the loss to measure the discrepancy of output distribution between teachers and students, a commonly used loss is Kullback–Leibler divergence [12]. The α_0 is the hyper-parameter to control the output KD loss; for symbol consistency, we add a subscript here to denote this is the 0-th hyper-parameters in the knowledge distillation framework.

Other than output knowledge distillation, many approaches have investigated distill feature knowledge on the intermediate representations [26]. Let f^T and f^S denote the feature maps of the teacher and student model, respectively. Therefore, the objective of feature knowledge distillation can be written as:

$$\mathcal{L}_{feat} = \mathcal{D}(F^T, F^S) = \mathcal{D}(r^T(f^T), r^S(f^S)) \tag{2}$$

where F is the feature knowledge, r^T and r^S are the mapping functions to align the sizes of feature maps for two models, f^T and f^S are the feature maps of the teacher and student, and $\mathcal{D}(\cdot)$ is the distance metric measuring the similarity of two learned features. In principle, the mapping function and the distance metric can be arbitrary.

Generally, the feature knowledge is leveraged in multiple stages; for example, there can be four intermediate features in a ResNet [11] and five intermediate features in a MobileNet [13, 27] to be distilled. Previous studies [17, 16] have shown that the importance of knowledge on different stages varies. Therefore we decompose the feature knowledge into more fine-grained stage-wise feature knowledge, which can be written as:

$$\mathcal{L}_{feat} = \mathcal{D}(\sum_{l=1}^{L} r_l^T(f_l^T), \sum_{l=1}^{L} r^S(f_l^S)) \tag{3}$$

where the L is the total number of stages that are used to transfer teachers' knowledge, r_l^T and r_l^T are mapping functions at stage l, and f_l^T and f_l^S are the feature maps of the teacher and student model at stage l, respectively. We write the mapping function and feature map in a separate format to emphasize that the weights of these mapping functions are not shared most of the time.

Moreover, it is common to have multiple features knowledge that need to be distilled, especially in downstream tasks [21, 36]. Therefore, we extend the loss function to multi-feature knowledge optimization objective:

$$\mathcal{L}_{feat} = \sum_{n=1}^{N} \alpha_n \mathcal{D}_n(F_n^T, F_n^S) \tag{4}$$

where N is the total number of feature knowledge, α_n is the hyper-parameters to control the contribution of n-th loss function to the gradient, and $\mathcal{D}_n$ is the distance metric for $n-$th feature knowledge.

As a result, the final optimization objective of a whole knowledge distillation framework can be written as the following:

$$\begin{aligned} L_{KD} &= \mathcal{L}_{out} + \mathcal{L}_{feat} + \mathcal{L}_S \\ &= \alpha_0 \mathcal{H}(p^S, p^T) \\ &+ \sum_{n=1}^{N} \alpha_n \mathcal{D}_n(F_n^T, F_n^S) + \mathcal{H}(p^S, y) \end{aligned} \tag{5}$$

where $\mathcal{L}_S$ is the student loss supervised by the ground truth label y. In the rest of the paper, we consider $\mathcal{L}_S$ as the student's primary loss. Note that we can make the final objective contain either output knowledge or feature knowledge only by setting the corresponding hyper-parameters to 0.

3.2. Rethinking Capacity Mismatch Between Student and Teacher

Existing literatures [2, 24] have suggested that better teachers do not make better student because students are unable to mimic the teacher, which is caused by the capacity mismatch between teacher and student. We hypothesize that the capacity mismatch happens intermittently instead of continuously in the KD training stage. In other words, at some iterations, the student fails to mimic the teacher due to the capacity gap. Consequently, it brings neutral or even negative effects on students. On the other hand, students do benefit from knowledge distillation most of the time. We evaluate our hypothesis on CIFAR100. First, we develop a criterion to measure the "capacity mismatch." In KD, the students are encouraged to mimic the behavior of the teacher. Therefore, a perfect KD method should produce the same representation given the same input on both teacher and student. As a result, the similarity of neural network representation could reflect the level of capacity mismatch to some extent. We use Center Kernel Alignment (CKA) [18], a technique that has been shown to be effective in measuring neural network similarity. We choose ResNet34 as our pre-trained teacher model and ResNet18 as the student model. We perform standard feature knowledge distillation [26] on stage three and stage four. There are three layers on stage 4 and six layers on stage three. Figure 3 shows that the representation similarity in six layers on the third stage between teacher and student. We choose to compare the network representation at ten different iterations at epoch eighty. For the CKA score figure, we primarily focus score on the diagonal, which indicate the representation similarity of convolution layer at the same position in the network. We observe that the representation of some pair of convolution layers between teacher and student in the receive high CKA score, where CKA ≤ 50, while some pair convolution layers obtain very low CKA score (CKA ≤ 30). This indicates that the capacity mismatch 1) is not consistent across different layers. At the same iteration, the

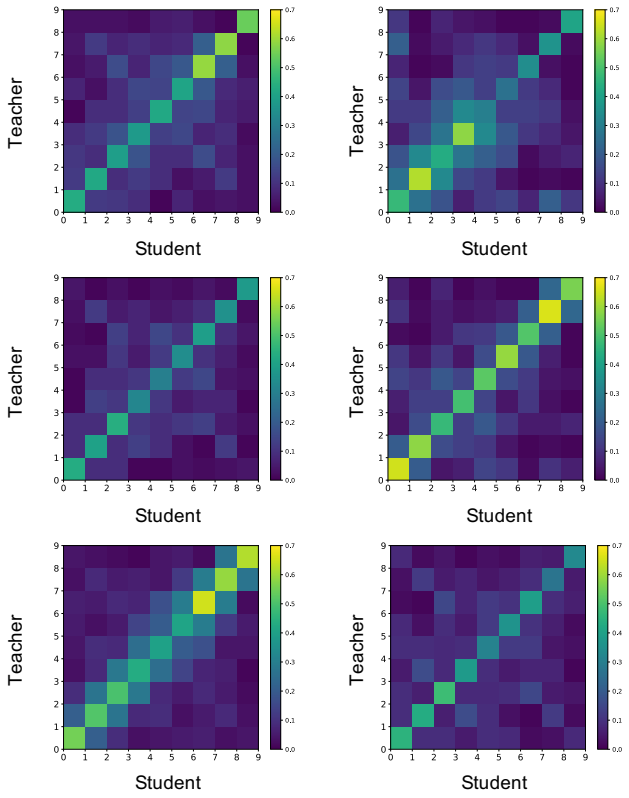

Figure 3: The CKA score for ResNet18 and ResNet34 on CIFAR100.

CKA score at stage four is higher yet the score is low at stage three. 2) different at every training iteration. At some iteration the CKA score are high while some iteration the CKA score is very low. Therefore, we conclude that it is important to control the KD process, by determining which knowledge at what stage is the student can benefit from the teacher.

3.3. Knowledge Distillation As Multi-Task Learning

To prevent distilling the teacher's knowledge to the student when their capacity is mismatched, our intuition is that we can measure the direction of the gradient for each part of knowledge distillation loss versus the primary supervision loss $\mathcal{L}_S$ in Equation 5. Motivated by the success of multi-task learning (MTL), we formulate the knowledge distillation framework as an MTL problem with each task corresponding to each distillation loss. Different from the conventional MTL objective where the model is encouraged to do their best on all types of tasks, in our setting, we only require the "main" task, which is $\mathcal{L}_S$, to achieve the best result. Therefore, we can compare the gradient similarity between the primary loss and other distillation loss and suspend any negative transfer for the student during training.

Concretely, Equation 5 presents a naive approach when

distilling knowledge from the teacher to student, where the optimization objective is unchanged during the entire training process. Contrary to the naive approach, we present the gradient-based adaptive knowledge distillation based on the student model. Instead of fixing the KD in the student training stage, SCKD adaptively changes the KD connection based on the gradient behavior of student loss and individual KD loss. As a result, our approach controls the knowledge distillation process during KD, such that any negative transfer caused by mismatch capacity is prevented.

Suppose the γ^m is the gate to either switch on or switch off the target KD loss, regarding the *m*-th iteration and $\gamma^m \in \{0, 1\}$. At every iteration, a mini-batch is selected to train the model. Then we can formulate our SCKD loss as the following:

$$\begin{aligned}\mathcal{L}_{SCKD^m} &= \gamma^m_{out}\mathcal{L}_{out} + \gamma^m_{feat}\mathcal{L}_{feat} + \mathcal{L}_S \\ &= \gamma^m_{out}\alpha_0 H(p^S, p^T) \\ &+ \sum_{n=1}^{N} \gamma^{m,n}_{feat}\alpha_n \mathcal{D}_n(F^T_n, F^S_n) \\ &+ \mathcal{H}(p^S, y)\end{aligned} \tag{6}$$

where m represents the current training iteration. At every iteration, we check if there is a negative transfer between the teacher's certain knowledge, either stage-wise feature-based knowledge or output-based knowledge, to the student. If the negative transfer is detected, we eliminate this knowledge at this round; otherwise, we include this knowledge to contribute to the optimization process. The overview of SCKD can be found in Figure 2. We note that our approach encapsulates early stopping [2], that can be achieved by trivially set all KD gate to 1 at the initial stage and set all KD gate to 0 at the predefined stop point.

3.4. Adaptive Knowledge Distillation via Gradient Similarity

In general, one can exploit any algorithm that can reflect the behavior difference between teacher and student during the KD training process, to control the knowledge transfer from teacher to student. Our approach is inspired by [8], we introduce to use the gradient cosine similarly to measure the gradient direction between each KD loss and student primary loss.

Specifically, we calculate the gradient of student primarily supervision $\nabla_\theta \mathcal{L}_S$, and the gradient of each knowledge distillation loss $\nabla_\theta \mathcal{L}_{kd_i}$, either feature-based knowledge or output-based knowledge. The θ is the weights that are updated through optimization by certain loss function. We obtain the gradient cosine similarity by calculating $cos(\nabla_\theta \mathcal{L}_S, \nabla_\theta \mathcal{L}_{kd_i})$. If $cos(\cdot)$ is greater or equal than some threshold ϕ, we regard such KD loss have negative transfer on student, and thus remove this KD loss at the current optimization step. Otherwise, we will include this KD loss similar to the conventional knowledge distillation approach.

Further, we need to study if there is a positive correlation between the cosine similarity and the capacity mismatch. To verify our assumption, we perform Pearson correlation test on gradient cosine similarity and the CKA score to check if there is statistically significant linear correlation between these two factors. We collect the CKA score at the first layer of stage three between teacher and student, and record the gradient cosine similarity over the last ten epochs. Our result shows that Pearson's R is equal to 0.6, which indicates that linear correlation between CKA and gradient cosine similarly statistically significant. This provides us a good measurement to control the KD process during the training stage. It also can be explained intuitively. Heuristically, when a particular knowledge loss from the teacher is used in the optimization, the student chooses to follow the direction of this knowledge since the teacher is always pretrained on the target dataset, thus ahead of the student in the gradient space. Nevertheless, the direction of teachers' knowledge does not always validate for a certain student, especially when the student fails to grasp the teacher's informative knowledge. As a result, we can manipulate the KD process from the perspective of gradient similarity. A Pytorch-style pseudo code is shown in Algorithm 1.

Algorithm 1 Training of Student Customized Knowledge Distillation

Require: Define Teacher Model T, Student Model S, list of KD loss $[\mathcal{L}_{KD}] = [(\gamma_0, \mathcal{L}_{kd_0}), \ldots, (\gamma_n, \mathcal{L}_{kd_n})]$.
Initialize all $\gamma = 1, \gamma \in (\gamma_1, \ldots, \gamma_n)$.
Initialize δ of T and Pretrain T, fix δ after pretrain.
Initialize θ of S and all distillation modules.
for $t = 1, \ldots, T$ **do**
  Get data x and target y of current mini-batch.
  Clear gradients for all parameters, *optimizer.zero_grad()*
  Compute $\nabla_\theta L_S$.
  $\mathcal{L}_{total}$ = $\mathcal{L}_S$.
  for $\gamma_i, \mathcal{L}_{kd_i}$ in $[\mathcal{L}_{KD}]$ **do**
    Compute $\nabla_\theta \mathcal{L}_{kd_i}$
    if *cos*($\nabla_\theta \mathcal{L}_S$, $\nabla_\theta \mathcal{L}_{kd_i}$) $\geq \phi$ **then**
      $\gamma_i = 0$
    end
    $\mathcal{L}_{total}$ += γ_i $\mathcal{L}_{kd_i}$
  end
  Compute gradients, $\mathcal{L}_{total}.backward()$
  Update weights, $optimizer.step()$.
end

where $\nabla_\theta L_{kd_i}$ is the gradient of $i-$th KD loss. In all our experiments, we trivially set the $\phi = 0$ unless otherwise indicated. When ϕ is set to 0, any KD loss that is orthogonal to or deviates from the student loss will be removed from the

Model	Method	Top-1 Acc (%)
ResNet18	NOKD	69.56
	BLKD	71.02
	TAKD	71.10
	ESKD	71.21
	SCKD	**71.73**
ResNet50	NOKD	72.79
	BLKD	74.95
	TAKD	75.25
	ESKD	75.09
	SCKD	**75.64**

Table 1: Comparison of model performance on CIFAR100 with various knowledge distillation frameworks to minimize capacity gap between teacher and student.

optimization for the current mini-batch.

4. Experiments

In this section, we evaluate our proposed method on three visual tasks: image recognition, object detection, and semantic segmentation.

4.1. Experiments on Image Recognition

4.1.1 Output-Based Knowledge

The experiments of image classification are conducted with three kinds of convolutional neural networks, including ResNet [11], MobileNetV2[27], ShuffleNetV2[22] on CIFAR100 [19] and ImageNet [4]. In CIFAR experiment, each model is trained with 300 epochs by SGD optimizer and the batch size is 128. In ImageNet experiments, each model is trained with 90 epochs by SGD optimizer and the batch size is 256. We investigate several teacher-student configurations, including the same network architecture (ResNet101-ResNet18) and different network architecture (ResNet101-MobileNetV2, ResNet101-ShuffleNetV2). Moreover, for the same teacher-student pair, we also conduct experiment on different student capacity (ResNet101-ResNet50, ResNet101-ResNet18). The temperature for all experiments are set to 1 follows Hinton *et.al.* [12]. We set the hyper-parameters for 0.9 for all experiments in this section.

Since our method is designed to resolve the capacity gap between teacher and student, we compare our method to 1) NOKD, which denotes no knowledge distillation and is trained from scratch. 2) BLKD, which denotes the baseline knowledge distillation. It is naively trained with a knowledge distillation method based on Hinton *et al.* [12]. 3) TAKD [24] utilized teacher assistant to ease the learning curve of the student and enable the student to achieve better performance than naively training by KD. 4) ESKD [2], represent early stopping knowledge distillation. From the results on CIFAR100 in Table 1, we can observe that our method outperforms the baseline KD. Additionally, our method consistently achieves superior performance than TAKD and ESKD on two kinds of teacher-student configurations. Our method seems to enjoy the large capacity gap between teacher and student.

Model	Method	Top-1 Acc (%)
ResNet18	NOKD	70.3
	BLKD	70.7
	TAKD	70.9
	ESKD	70.7
	SCKD	**71.3**
MobileNetV2	NOKD	70.9
	BLKD	71.8
	TAKD	71.9
	ESKD	72.0
	SCKD	**72.4**
ShuffleNetV2	NOKD	69.4
	BLKD	70.2
	TAKD	70.4
	ESKD	70.5
	SCKD	**71.3**

Table 2: Comparison of model performance on ImageNet.

We further conduct experiments on the large-scale dataset ImageNet. Table 2 shows the experimental results on ImageNet are consistent as on CIFAR100, where our method outperforms all three methods on various teacher-student configurations. Besides, the performance gap between our method and baseline is even bigger on the same network architecture (ResNet101-ResNet18), which indicates the robustness of our proposed method. We assume that our performance gain comes from choosing to transfer knowledge smartly instead of brutally stop all the knowledge transfer given a stopping point (e.g., ESKD). We also think building an intermediate network with a smaller size is a compromise, which didn't solve the root's capacity gap.

4.1.2 Feature-Based Knowledge

As the above experiments are conducted on a simple knowledge distillation scenario, where only output distillation is involved, we further validate our method on a more complex and state-of-the-art knowledge distillation framework. Since our method does not present any new distillation loss, and it is easy to insert into any existing framework, we choose TOFD [37] as our KD framework and apply our method based on TOFD. TOFD is a task-oriented knowledge distillation method, which is a state-of-the-art KD method that is composed of three knowledge distilla-

Model	Backbone	Method	FPS	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN	ResNet18	Baseline	30.57	34.6	55.0	37.1	19.3	36.9	45.9
		FBOD	30.57	37.0	57.2	39.7	19.9	39.7	50.3
		SCKD	30.57	37.5	57.6	40.2	20.9	42.6	50.8
RetinaNet	ResNet18	Baseline	23.30	33.4	51.8	35.1	16.9	35.6	44.9
		FBOD	23.30	35.9	54.4	38.0	17.9	39.1	49.4
		SCKD	23.30	36.5	56.1	38.9	18.2	39.6	49.8

Table 3: Comparison of model performance on MS COCO object detection. We evaluate on both two-stage detector (Faster RCNN) and one-stage detector (RetinaNet) based on FBOD framework.

Model	Baseline	KD	FitNet	DML	SD	TOFD	**SCKD**
ResNet18	77.09	78.34	78.57	78.72	78.72	82.92	**84.16**
SENet18	77.27	78.43	78.82	79.72	78.58	84.44	**85.49**
ShuffleNetV2	72.38	72.86	74.36	72.66	72.72	76.68	**77.58**

Table 4: Comparison of model performance on CIFAR100 with different knowledge distillation frameworks.

tion losses and an orthogonal loss which is applied on the feature resizing layer to prevent information loss resizing. The TOFD contains both output distillation loss and multiple feature distillation loss, and we adopt SCKD based on TOFD's framework. We follow the experimental setting in TOFD for fair comparisons, and we evaluate our method on CIFAR100 with multiple student network architectures (ResNet, SENet [14], and ShuffleNetV2).

Table 4 shows the experiment results of TOFD on ImageNet. We observe that our method consistently improves the TOFD framework. For example, with ResNet18 we improve the TOFD by 1.24%, with ShuffleNetV2 we boost TOFD's performance by 0.90%, and with SENet18 we increase the Top-1 accuracy by 1.05%. The improvement is significant. As a result, SCKD outperforms all state-of-the-art KD methods, which indicates the superiority of our proposed method.

4.2. Experiments on Object Detection

Our previous study shows that SCKD works on the image classification task. We further study the effectiveness of SCKD on downstream tasks. We evaluate SCKD on MS COCO object detection. We conduct experiments based on the Feature-Based Object Detector (FBOD) [36]. The FBOD is a state-of-the-art knowledge distillation method for object detection. It comprises three types of teacher feature knowledge modules: attention transfer module, attention mask module, non-local module. We follow the same experimental setting as in FBOD [36] and perform SCKD training algorithm based on FBOD's framework. In Table 3, we present the experimental results on RetinaNet-ResNet18 and Faster RCNN-ResNet18. As we can see, for both one-stage detector and two-stage detector, SCKD outperforms the FBOD, with over 0.5% AP on Faster RCNN and 0.6% AP on RetinaNet. Note that in the paper of FBOD, the author has empirically back up the argument of "better teacher makes better student" by experiments. However, our method still improves the performance over FBOD, which indicates the effectiveness of the proposed gradient similarity-based adaptive knowledge distillation.

4.3. Experiments on Semantic Segmentation

Besides applying SCKD to image classification and object detection, we also perform experiments on semantic segmentation, a challenging dense prediction vision task. Our model was built based on IFVD [31], which is a state-of-the-art KD method for semantic segmentation that consists of three knowledge distillation losses. The details can be found in the original paper [31]. Our experiments is conduct on CityScapes [3], a popular semantic segmentation benchmark. We test on various teacher-student configurations via same decoder architecture PSPNet, including ResNet101-ResNet18 (full width), ResNet101-ResNet18 (0.5 width), and ResNet101-EfficientNetB0. The training scheme follows the official implementation of IFVD, and we do not alter any hyper-parameters for fair comparisons. Table 5 summarizes the performance of the SCKD on CityScapes. For comparison, we report both validation mIoU and test mIoU. Our method had better performance in all settings than baseline methods and IFVD. This is expected as the task is more challenging than image classification. We also found that the SCKD performs better on teacher-student, which has a large model capacity gap, such as ResNet101-ResNet18(0.5). We assume that SCKD exhausts the IFVD framework's potential, where its performance was lower as it should be due to the capacity mismatch.

Method	*val* mIoU (%)	*test* mIoU (%)	Params (M)	FLOPs (G)
PSPNet-ResNet18‡(1.0)	57.50	56.00	13.07	125.8
SKD [21]	63.20	62.10		
IFVD [31]	66.63	65.72		
Ours	**67.25**	**66.30**		
PSPNet-ResNet18‡(0.5)	55.40	54.10	3.27	31.53
SKD [21]	61.60	60.50		
IFVD [31]	63.35	63.68		
Ours	**65.10**	**64.92**		
PSPNet-EfficientNetB0	58.37	58.06	4.19	7.97
SKD [21]	62.90	61.80		
IFVD [31]	64.73	62.52		
Ours	**65.17**	**63.08**		

Table 5: the performance on Cityscapes. The † indicates pretrained with ImageNet. The ‡ indicates train from scratch. (0.5) indicates half channel number compares to (1.0) which denotes the full channel numbers. The results in DeepLabV3-ResNet18 are re-implemented use the official codes release by the authors. The teacher network is PSPNet-ResNet101, and our method is built based on IFVD.

4.4. Ablation Study and Sensitivity Study

Better teacher makes better student Although, we have seen performance enhancement of SCKD on the existing knowledge distillation framework, a further analysis on the argument "better teacher makes better student" is desired. We train a plain CNN with batch normalization, skip connection and ReLU activation as the student. It is distilled by large teachers of 4, 6, 8, and 10 layers on both CIFAR10 and CIFAR100. As expected and illustrated in Figure 4, by increase student number of layers, the student performance gain is initially increase and then decrease due to the capacity mismatch by conventional KD. On the other hand, the student performance gain is positive correlated to the student size by applying SCKD. This indicates that our method indeed make large model to be better teacher.

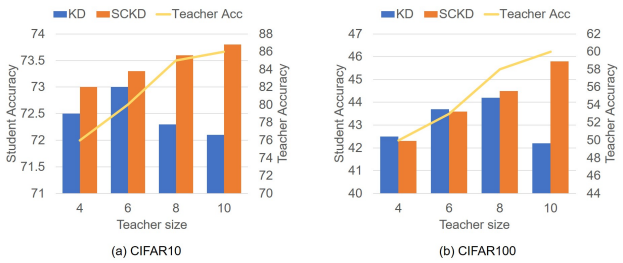

Figure 4: Best view in color. Ablation study on better teacher makes better student. With the teacher size increase (and accuracy increase correspondingly), convention KD makes worst student and SCKD makes better student.

Sensitivity Study on Cosine Similarity Threshold ϕ. The cosine similarity threshold ϕ determine when to eliminate the KD loss in the current iteration. Intuitively, any KD loss that have inverse (when $\phi > 0$) or orthogonal gradient direction (when $\phi = 0$) can be considered as a "bad" knowledge, thus cast during the training stage. We study a sensitivity study on this hyper-parameter. The results are shown in Table 6. We conclude that our method do sensitive to the hyper-parameters ϕ if it is set to the other value. Therefore we can heuristically set ϕ to zero and achieve satisfactory performance. Note that the accuracy of worst ϕ is still comparable to the baseline method as we shown in Table 1, we think this shows the necessity of controlling knowledge distillation process on-the-fly.

Threshold ϕ	1.2	1.1	1.0	0.9	0.8
Top-1 Acc (%)	75.2	75.1	75.4	75.0	74.9

Table 6: Comparison of model performance on CIFAR100 with different knowledge distillation frameworks.

5. Conclusion

In this paper, we present a adaptive knowledge distillation method to bridging the capacity gap between student and teacher. We examine the capacity mismatch from the perspective of gradient similarity between student loss and distillation loss. We then formulate knowledge distillation as multi-task learning problem. As such, our method can automatically find KD training strategy based on the target student model. We validate the effectiveness of our method on three visual tasks.

References

[1] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[2] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4794–4802, 2019.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Xiang Deng and Zhongfei Zhang. Learning with retrospection. *arXiv preprint arXiv:2012.13098*, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *Advances in Neural Information Processing Systems*, 33, 2020.

[8] Yunshu Du, Wojciech M Czarnecki, Siddhant M Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

[9] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.

[10] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[15] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[16] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. In *International Conference on Machine Learning*, pages 3030–3039. PMLR, 2019.

[17] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. *arXiv preprint arXiv:2102.02973*, 2021.

[18] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[21] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019.

[22] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[23] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

[24] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020.

[25] Hossein Mobahi, Mehrdad Farajtabar, and Peter L Bartlett. Self-distillation amplifies regularization in hilbert space. *arXiv preprint arXiv:2002.05715*, 2020.

[26] 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.

[27] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[28] Chengchao Shen, Mengqi Xue, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3504–3513, 2019.

[29] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[31] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *European Conference on Computer Vision*, pages 346–362. Springer, 2020.

[32] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2019.

[33] 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.

[34] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020.

[35] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13876–13885, 2020.

[36] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2021.

[37] Linfeng Zhang, Yukang Shi, Zuoqiang Shi, Kaisheng Ma, and Chenglong Bao. Task-oriented feature distillation. *Advances in Neural Information Processing Systems*, 33, 2020.

[38] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.

[39] Linfeng Zhang, Zhanhong Tan, Jiebo Song, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Scan: A scalable neural networks framework towards compact and efficient models. *arXiv preprint arXiv:1906.03951*, 2019.