Gap-KD: Bridging the Significant Capacity Gap Between Teacher and Student Model

Shan Huang
Yunnan University
Kunming, China
huangshan@stu.ynu.edu.cn

Wenhua Qian*
Yunnan University
Kunming, China
whqian@ynu.edu.cn

Abstract

Deploying efficient deep learning models in resource-constrained environments is challenging due to their growing scale and complexity. Knowledge Distillation (KD) offers a practical solution by transferring knowledge from a large teacher model to a compact student model. However, traditional KD methods often fall short when significant capacity gaps exist. To address this issue, we introduce Gap-KD, which utilizes dynamic temperature scaling and a double decoupling technique to bridge these gaps. The distillation temperature is adjusted dynamically, progressively increasing the student model's learning difficulty. A Teacher Assistant (TA) model is introduced as an intermediary layer, initially reducing the gap. Building on this, the outputs of the teacher, TA, and student models are doubly decoupled, further reducing information loss and error accumulation. Extensive experiments on CIFAR-10 and CIFAR-100 datasets with ResNet and CNN architectures demonstrate that Gap-KD achieves state-of-the-art performance specifically in scenarios with significant capacity gaps, highlighting its effectiveness for these challenging conditions. The code is available at https://anonymous.4open.science/r/Gap-KD-C411

Keywords: Knowledge Distillation, Deep Learning, Computer Vision, Neural Networks

1. Introduction

Over the past few decades, deep learning-based methods[40, 4, 16] has made significant breakthroughs in a number of areas, including computer vision[40, 19, 34, 48], natural language processing[12, 31], and target detection[15, 36]. However, most of these successes have been due to the complexity of the models and the increase in computational power, which limits precisely the adoption and deployment to mobile devices[6, 11].

*Corresponding author

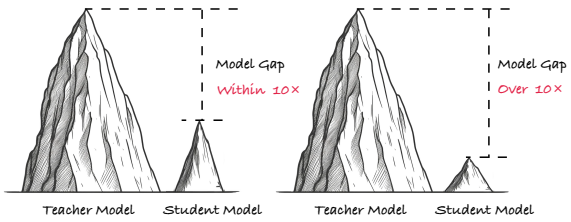


Figure 1: Problem definition of the large gap between a teacher and a student model. While the majority of knowledge distillation research concentrates on model disparities within a tenfold range, our aim is to tackle scenarios where the gap extends beyond this threshold.

A case in point is large language model(LLM)[53] such as the GPT-4[1], LaMDA[43] and LLaMA[45], any one of them contains billions of parameters. Obviously, deploying these models to devices with limited computing power is nearly impossible. In this regard, knowledge distillation(KD)[18], as a prominent method of model compression, has been widely adopted in the deep learning field.

KD is the method to transfer the knowledge of a pre-trained teacher model to a lightweight student model. It works by softening the teacher's output logits with temperature to provide additional supervision for the student model. Specifically, the soft logits contain more inter-class information than the hard class target of the student model itself.

Despite the fact that numerous studies[24, 51, 46, 5, 9, 8, 23] on the KD method have demonstrated impressive achievements across a wide array of tasks[13, 30, 37, 35], it is subjected to certain specific constraints as well. As illustrated in Figure 2, with the increasing model disparity, the improvement of traditional knowledge distillation methods over the models' training outcomes diminishes. When the model gap exceeds tenfold, the effectiveness of knowledge distillation falls below the results obtained through the models' self-training, marking the ineffectiveness of conventional

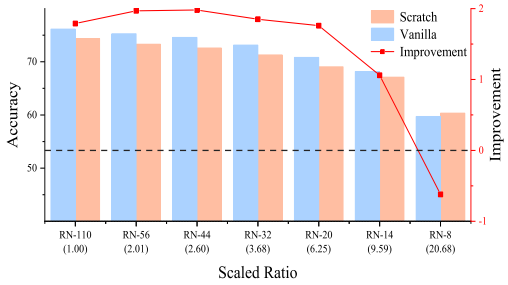


Figure 2: This graph displays the efficacy of KD versus training from scratch for ResNet models. The horizontal axis categorizes models by their scale ratios, defined as the parameter count of ResNet-110 relative to the model in question. The left vertical axis measures accuracy, while the right vertical axis shows the accuracy gains from KD compared to scratch training. Blue and yellow bars indicate the accuracies from KD and scratch training, respectively. The red line graph shows the extent of improvement due to KD, and the black dotted line marks no improvement (improvement = 0). Notably, KD's performance falls short when the scale ratio exceeds 10.

knowledge distillation approaches at this point. Previous efforts in the field primarily focused on teacher-student model pairs with relatively minor disparities, such as ResNet56 and ResNet20, or VGG13 and VGG8, to circumvent this issue. Nonetheless, the challenge posed by significant gaps holds substantial relevance in practical applications. Often, the computational capabilities of the devices designated for deploying distilled models are fixed, and the disparity between these devices and the larger models intended for distillation can be quite pronounced. This underscores the critical importance of addressing how to effectively conduct knowledge distillation when faced with substantial disparities. In this paper, we define the issue as the capacity gap problem .

To address the capacity gap problem, [33] proposed the TAKD method to bridge the gap via an extra auxiliary model(teacher assistant) whose size is between the size of the teacher and the student. There are two phases of the whole training process. Firstly, the TA model is trained by the teacher model. Then, the TA plays the role of the teacher to teach the student. In this way, the huge gap between the teacher and the student is split into two parts: the gap between teacher and TA and the gap between student and TA. While TAKD achieved a significant performance improvement in addressing the capacity gap problem , there are still a number of issues. Firstly, the errors in the TA training process can be easily propagated to the student, which is referred to as the error avalanche problem. [41] proposed the DGKD to alleviate the error avalanche problem by densely guided multiple TAs. Besides, there is a distillation paths problem with both TAKD and DGKD. The more TAs used as intermediate layers, the better the distillation effect achieved. Specifically, when using CNN-2 as the student model and CNN-10 as the teacher model on the CIFAR-100 dataset, the accuracy of the complete distillation path ($10 \rightarrow 8 \rightarrow 6 \rightarrow 4 \rightarrow 2$) is 0.86% [33] higher than using a single TA ($10 \rightarrow 8 \rightarrow 2$). However, this approach also increases the cost of training. Therefore, determining how to select the appropriate TA and the distillation path to achieve a relative balance between computational cost and effectiveness presents an intractable challenge.

In this work, we proposed our Gap-KD method to tackle all of the capacity gap, error avalanche and distillation paths problems. In Gap-KD, we also adopt two-stage distillation. In the first stage, the TA model is trained by the teacher model. In the second stage, the student model is trained using both the TA and the teacher models. Inspired by [52], we double decouple target class knowledge from teacher and non-target class knowledge from TA separately, which could avoid error avalanche problem and alleviate capacity gap problem .In addition to the above, the temperature factor is incorporated at both stages of distillation. This methodology does not merely bridge the performance gap between the models, but also facilitates dynamic adjustment of the temperature, thereby uncovering more latent details within the models. Consequently, this enhances the model's competence in addressing the distillation paths problem to a certain extent. As a result, our approach can significantly surpass the performance of both TAKD and DGKD along complete distillation paths, even when utilizing only a single Teaching Assistant. Overall, the contributions of this paper are summarized as follows:

- We propose the Dynamic Temperature Module (DTM) which effectively organizes the distillation task from easy to hard by dynamically adjusting the temperature parameter in a rate decay fashion.

- We introduce the Double Decoupling Module (D2M) which decouples both sources (teacher, assistant, student) and categories (target and non-target) for distillation. This approach not only enhances distillation efficiency but also mitigates error propagation.

- We apply the Gap-KD method to ResNet8 and plain CNN models on both CIFAR-10 and CIFAR-

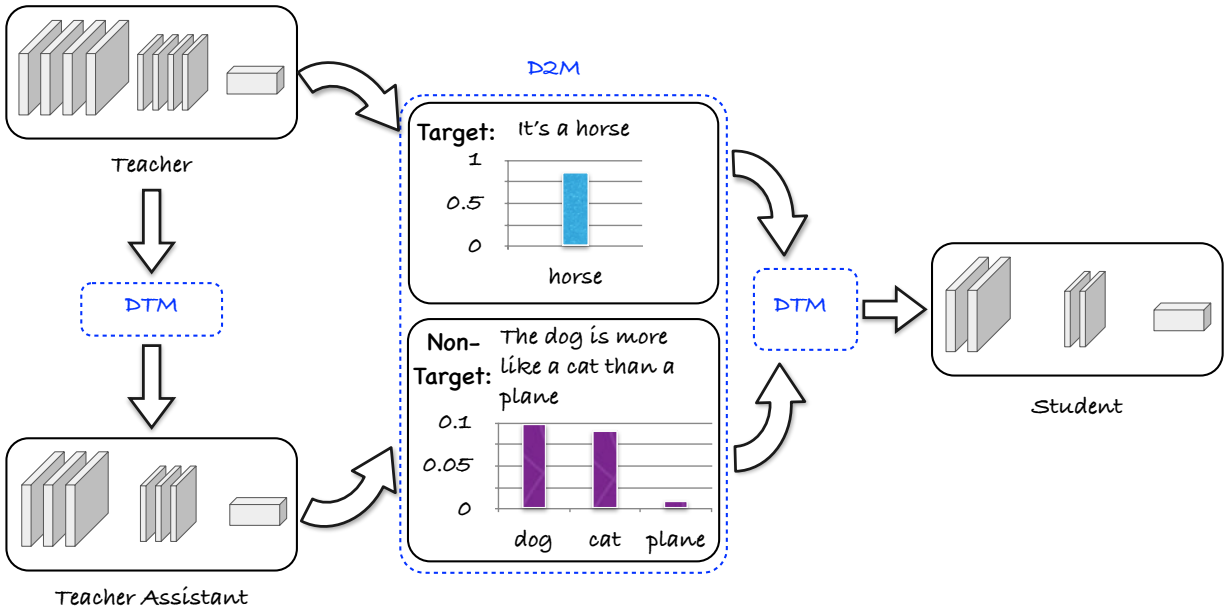


Figure 3: The overall architecture of our proposed method. To address the capacity gap problem, error avalanche problem, and distillation paths problem in previous approaches, we introduce the Dynamic Temperature Module (DTM) and the Double Decoupling Module (D2M).

100 image classification tasks and achieved the state-of-the art results.

2. Related Work

2.1. Knowledge Distillation

Knowledge Distillation(KD), as a popular model compression method [7], was originally proposed in [18]. In KD, the computational expensive model(teacher model) transfers the "dark knowledge" to the single computational efficient neural network (student model) via soft labels from teachers [14, 49]. Since then, KD has been widely adopted across various learning tasks, especially in vision tasks [20, 26, 3, 29, 49, 27]. Original knowledge distillation starts with a pre-trained cumbersome teacher model. Then a light-weight student model is trained under the supervision of the teacher model. To better convey the information, [18] uses fixed temperature to soften the logits provided by the teacher, which could provide more dark knowledge than hard labels and be easier to be learned by the student. Specifically, given the labeled classification dataset $D = \{(x_i, y_i)\}_{i=1}^{I}$, the student mimics the teacher by minimizing the Kullback-Leibler(KL) divergence loss between their soft output probabilities:

$$\mathcal{L}_{KD}(q^t, q^s, \tau) = \sum_{i=1}^{I} \tau^2 KL(\sigma(q_i^t/\tau), \sigma(q_i^s/\tau)) \quad (1)$$

where q^t and q^s denote the logits of the teacher and the student respectively, τ is the temperature for softening the distribution, and $\sigma(\cdot)$ means the softmax function.

2.2. Large Capacity Gap Between Teacher and Student

Quite a bit of past literature [32, 33, 10] has observed a phenomenon: good teachers do not necessarily teach students well, just as a college professor who teaches elementary school knowledge is not necessarily as good as an elementary school teacher. When the teacher's capacity and the student's are in a certain range, the better the teacher's capacity, the better supervision it provides to the student. However, when the gap between teacher and student is too large, the student does not have enough ability to mimic the teacher's behavior despite receiving hints.

There are numerous approaches have been proposed to address the problem. TAKD [33] introduced an intermediate teacher assistant model whose capacity is

greater than the student but smaller than the teacher to bridge the huge gap. Building on this foundation, DGKD [41] employs a densely connected network of assistant models to further enhance the effectiveness of the distillation process. RCO [22] devised a sequential learning pathway that progressively mimics the teacher, guiding the student model through the teacher's optimization trajectory.

3. Method

3.1. Motivation

In knowledge distillation, a key challenge is the significant capacity gap between the teacher and student models. The teacher, being more complex, often makes it difficult for the student model to effectively learn, leading to the failure of traditional distillation methods. This situation can be compared to a university professor teaching advanced concepts directly to an elementary school student, which is highly ineffective due to the knowledge gap.

The previously proposed approach[33] introduced a teaching assistant to help bridge this gap. Although this approach alleviates some of the issues, relying solely on the assistant can lead to potential errors and limit the student's learning potential.

Our approach, inspired by real-world education, addresses these issues by allowing the student to learn simultaneously from both the teacher and the assistant. In the first phase, the teacher distills its knowledge into the assistant model. In the second phase, the student learns from both the teacher and the assistant—the assistant provides foundational knowledge (basic concepts), while the teacher provides advanced knowledge (correct answers). The entire process is designed to mimic the gradual progression of real-world education, with learning difficulty increasing from easy to hard by adjusting the distillation temperature from high to low, making the process more intuitive and effective.

3.2. Overall Architecture

The overall architecture of our proposed method is illustrated in Figure 3. It consists of two main stages: Stage I - Training the Teacher Assistant with a Teacher, and Stage II - Training the Student with both the Teacher and the Teacher Assistant. Each stage is described in detail in section 3.3 and section 3.4 , respectively. To aid understanding of the complete workflow, the corresponding algorithm is provided in Algorithm 1.

Algorithm 1 Gap-KD

Input: Training dataset $D = \{(x_i, y_i)\}_{i=1}^{I}$; Pre-trained Teacher $\mathcal{T}$; $\tau_{\max}$; $\tau_{\min}$; Training epochs in stage I : N_1 and stage II : N_2
Output: Well-trained Student $\mathcal{S}$
Initialize: Epoch $i = 1$; Temperature value $\tau = \tau_{\max}$

▷stage I
while $i \leq N_1$ do
  for each data batch x in D do
    Forward propagation through $\mathcal{T}$ and $\mathcal{A}$ to obtain outputs;
    Calculate τ_i based on the Eqn.6 and Eqn.7;
    Calculate the vanilla KD loss soften by τ_i and update $\mathcal{A}$ through backward propagation;
  end for
  $i = i + 1$;
end while
▷stage II
while $i \leq N_2$ do
  for each data batch x in D do
    Forward propagation to obtain the outputs from $\mathcal{T}$, $\mathcal{S}$, and $\mathcal{A}$ to obtain outputs respectively;
    Calculate τ_i based on the Eqn.6 and Eqn.7;
    Calculate the $\mathcal{L}_{\text{total}}$ based on Eqn.11, Eqn.12, Eqn.13, and Eqn.14, update $\mathcal{S}$ through backward propagation;
  end for
  $i = i + 1$;
end while
return $\mathcal{S}$

3.3. Stage I :Training Teacher Assistant with a Teacher

In this stage, the TA is being trained under the supervision of both the teacher model and the ground truth. Without any loss of generality, we assume that $\theta_{\mathcal{A}}(i)$ represents the TA's optimized parameters at the start of i-th epoch, will be updated during the i-th epoch of training to obtain $\theta_{\mathcal{A}}(i+1)$:

$$\mathcal{L}_{CE}(\theta_{\mathcal{A}}(i)) = CE\big(y, \mathcal{A}(x; \theta_{\mathcal{A}}(i))\big), \tag{2}$$

$$\mathcal{L}_{KL}(\theta_{\mathcal{A}}(i), \tau_i) = \tau_i^2 KL(T(x, \theta_T)||\mathcal{A}(x, \theta_{\mathcal{A}}(i)); \tau_i), \tag{3}$$

$$\mathcal{L}(\theta_{\mathcal{A}}(i), \tau_i) = \alpha\mathcal{L}_{CE}(\theta_{\mathcal{A}}(i)) + \beta\mathcal{L}_{KL}(\theta_{\mathcal{A}}(i), \tau_i), \tag{4}$$

$$\theta_{\mathcal{A}}(i+1) \leftarrow \min_{\theta_{\mathcal{A}}} \mathcal{L}(\theta_{\mathcal{A}}(i), \tau_i) \tag{5}$$

Existing literature [18, 28] indicates that the temperature parameter in the distillation process softens the logits output by the teacher model. This transforms the distribution from one-hot encoding to a richer distribution containing inter-class relationships,

aiding the student model in learning more dark knowledge. A higher temperature results in a smoother distribution, facilitating quicker and easier knowledge transfer to the student model while lowering confidence in the target class. Conversely, a lower temperature produces a steeper distribution, focusing the distillation on the teacher's maximum logits, which increases the learning difficulty for the student model but enhances confidence in the target class.

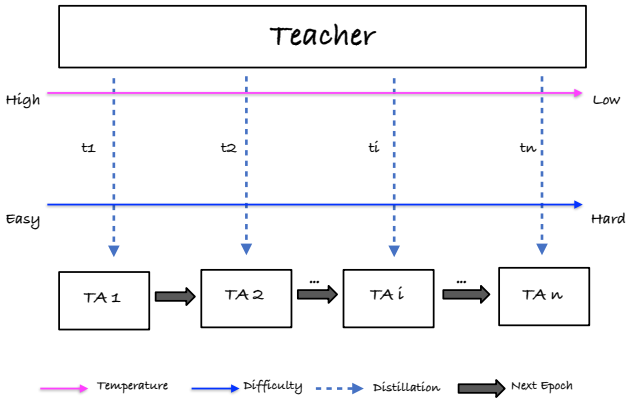


Figure 4: Overview of the stage I. By designing a gradient of temperature settings that decrease from high to low in each epoch, the learning difficulty for TA progresses from easy to hard, thereby optimizing the learning efficiency of the TA model.

Following the gradual learning principles in human education, it is intuitive to progressively decrease the temperature from high to low, making the student's learning difficulty transition from easy to hard. This approach aligns with the stepwise complexity in human learning, enabling the student model to quickly absorb initial knowledge and subsequently focus on details and accuracy, ultimately achieving optimal learning outcomes. Inspired by [50], we propose the Dynamic Temperature Module(DTM), take rate decay technique to change the temperature value τ w.r.t current epoch, which is empirically observed to help both optimization and generalization. Let the decay rate δ, be defined by the formula:

$$\delta = \left(\frac{\tau_{\min}}{\tau_{\max}}\right)^{\frac{1}{\mathcal{N}_1 - 1}} \tag{6}$$

where $\tau_{\min}$ and $\tau_{\max}$ represent the minimum and maximum temperatures, respectively, and $\mathcal{N}_1$ denotes the total number of epochs in stage I. Then, the temperature τ_i at the i-th epoch is given by:

$$\tau_i = \tau_{\max} \cdot \delta^{i-1} \tag{7}$$

Subsequently, we soften the teacher's output for each epoch utilizing the derived τ_i, facilitating a graduated distillation process from simple to hard by progressively decreasing τ_i from high to low. This approach effectively mitigates the Capacity Gap Problem. Moreover, in the subsequent stage stage II, it enables the attainment of higher accuracy compared to other methods employing multiple TAs, despite utilizing only a single TA, thereby alleviating the Distillation Paths Problem to a certain extent.

3.4. Stage II :Training Student with Teacher and Teacher Assistant

In this stage, the student model will be jointly trained with the TA trained in stage I , the teacher model, and the true labels.

Inspired by [52], we propose a Double Decoupling Module(D2M): not only decoupling both the teacher model and the assistant teacher model, but also decoupling the target and non-target class outputs. In this way, the D2M significantly improves the accuracy and efficiency of classification. The operational mechanism of this module will be detailed next.

For a given training instance belonging to the t-th class, the vector of classification probabilities may be represented as $p = [p_1, p_2, ..., p_t, ..., p_C] \in \mathbf{R}^{1\times C}$, where p_i signifies the probability assigned to the i-th class, with C denoting the total number of class categories. For any class i, p_i can be obtained by:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \tag{8}$$

where z_i is the logit of i-th class.

The binary probability distribution for a given instance w.r.t. the target class and the aggregate of all other non-target classes are encapsulated by: $\mathtt{b} = (p_t, \overline{p})$, where p_t is the probability of the target class and $\overline{p}$ represents the cumulative probability of all classes excluding the target, which can be derived as follows:

$$\begin{aligned} p_t &= \frac{\exp(z_t)}{\sum_{j=1}^{C} \exp(z_j)}, \\ \overline{p}_t &= \frac{\sum_{k=1,k\neq t}^{C} \exp(z_k)}{\sum_{j=1}^{C} \exp(z_j)} \end{aligned} \tag{9}$$

Concurrently, we define $\hat{\mathbf{p}} = [\hat{p}_1, ..., \hat{p}_{t-1}, \hat{p}_{t+1}, ..., \hat{p}_C] \in \mathbf{R}^{1\times(C-1)}$ to exclusively represent the probability distribution across non-target classes, explicitly excluding the t-th class. Each element is calculated by:

$$\hat{p}_i = \frac{\exp(z_i)}{\sum_{j=1,j\neq t}^{C} \exp(z_j)} \tag{10}$$

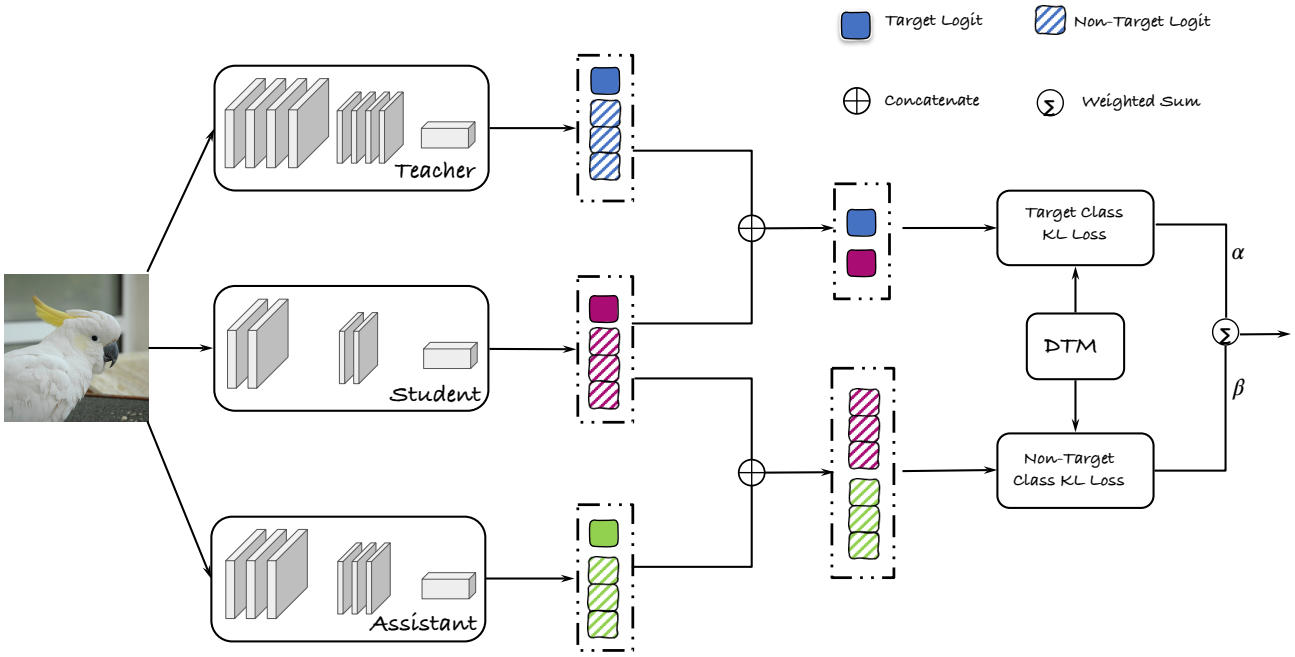


Figure 5: Overview of the stage II. The target class outputs of the teacher and the student are used to compute the Target Class KL Loss (TCKL), while the non-target class outputs of the TA and the student are utilized to calculate the Non-Target Class KL Loss (NCKL). Dynamic Temperature Modulation (DTM) is introduced in the calculation of both TCKL and NCKL. Finally, a weighted sum of these two components is computed.

Next, we will introduce our loss function of D2M, $\mathcal{L}_{\text{total}}$, is defined as:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{KL}_t} + \gamma\mathcal{L}_{\text{KL}_{\bar{t}}} \tag{11}$$

where α, β and γ are weighting coefficients that balance the contributions of the respective loss components. $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss. $\mathcal{L}_{\text{KL}_t}$ and $\mathcal{L}_{\text{KL}_{\bar{t}}}$ refer to the Kullback–Leibler divergence loss for the target class and non-target class respectively.

$$\mathcal{L}_{\text{CE}} = -\sum_i y_i \log(p_i) \tag{12}$$

$$\mathcal{L}_{\text{KL}_t} = \text{KL}(\mathbf{b}^{\mathcal{T}} || \mathbf{b}^{\mathcal{S}}) = p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + \overline{p}_i^{\mathcal{T}} \log(\frac{\overline{p}_t^{\mathcal{T}}}{\overline{p}_t^{\mathcal{S}}}) \tag{13}$$

$$\begin{aligned}\mathcal{L}_{\text{KL}_{\bar{t}}} &= \text{KL}(\hat{\mathbf{p}}^{\mathcal{A}} || \hat{\mathbf{p}}^{\mathcal{S}}) \\ &= \overline{p}_t^{\mathcal{A}} \sum_{i=1, i\neq t}^{C} \hat{p}_i^{\mathcal{A}} \log(\frac{\hat{p}_i^{\mathcal{A}}}{\hat{p}_i^{\mathcal{S}}})\end{aligned} \tag{14}$$

where y_i represents the ground truth label for the i-th class in a one-hot encoded vector. We formulate the loss function with the binary probabilities $\mathbf{b}$ and the probabilities among non-target classes $\hat{\mathbf{p}}$, and we define $\mathcal{T}$, $\mathcal{A}$ and $\mathcal{S}$ as the teacher, the teacher assistant and the student respectively.

By employing a D2M approach, we can achieve two main advantages that make the knowledge distillation process more effective and precise. Firstly, traditional KD loss functions highly couple the Target Class KL Loss (TCKL) with the Non-Target Class KL Loss(NCKL), overlooking their distinct roles in the distillation process. Specifically, TCKL conveys knowledge about the difficulty of each training sample, reflecting the ease or challenge of identifying each sample; whereas NCKL contains what's known as 'dark knowledge', contributed through coefficients negatively correlated with the teacher model's prediction confidence for the target category. This means higher prediction scores result in smaller weights. However, the coupling in traditional KD methods significantly reduces the influence of NCKL on well-predicted training samples, which is undesirable since a teacher model's high confidence in a training sample indicates the reliability and

value of the transferred knowledge. Moreover, excessive coupling restricts the possibility of independently adjusting the weights for both parts, thus greatly limiting the distillation effect. Secondly, D2M ensures that the student model's learning of the target class relies solely on the teacher model and the true labels. This effectively avoids the potential for the teaching assistant model to transmit incorrect information, fundamentally solving the error avalanche problem. Through this method, we can ensure that the student model acquires more accurate and reliable knowledge during the learning process, thereby improving learning efficiency and model performance.

4. Experiments

In this section, we evaluate the performance of Gap-KD on the image classification task. To ensure a comprehensive and fair evaluation, we design our experiments in two parts. The first part benchmarks our approach against the most recent conventional KD SOTA methods from the past three years. The second part involves a comparison with state-of-the-art methods specifically designed to address significant model capacity gaps.

4.1. Datasets

To maintain consistency with previous work [38, 33, 21, 22, 18] focused on knowledge distillation in the context of significant model capacity gaps and to facilitate comparison, we assess Gap-KD on CIFAR-10 and CIFAR-100 datasets [25], which are widely used as benchmarks for image classification. Both datasets contain 60,000 of 32 32 RGB images, with 50,000 designated for training and 10,000 for testing. There are 10 classes for CIFAR-10 and 100 classes for CIFAR-100, respectively.

Furthermore, we acknowledge that compressing models to extremely small architectures, such as ResNet-8 or CNN-2, has limited practicality when applied to large-scale datasets like ImageNet. The dramatic reduction in capacity can lead to a significant drop in model performance, making the application of such compact models less meaningful for large and complex datasets. Thus, we primarily focus on CIFAR-10 and CIFAR-100, where the feasibility and impact of our approach can be more effectively demonstrated and compared against existing methods.

4.2. Part 1: Comparision with state-of-the-art Methods

The first part of our experiments focuses on comparing our approach with recent state-of-the-art KD methods from the past three years, such as DKD[52], SimKD[9], CTKD[28], LSKD[42]. Additionally, we compare several common feature-based and logit-based knowledge distillation methods. We utilize the CIFAR-100 datasets for this comparison. Both homogeneous and heterogeneous teacher-student networks are evaluated. For homogeneous networks, we use configurations such as ResNet-110 to ResNet-8, and for heterogeneous networks, we employ configurations such as ResNet-32×4 to MobileNet-V2[39]. The hyperparameters for all baseline methods are set according to their default configurations as provided in their respective published codes. For LSKD, the MLKD[23]+logit standardization method is specifically used.

Table 1: Results on the CIFAR-100 dataset. All results are the average of three independent experiments, bold indicates the best result, underline represents the second best, and Δ denotes the improvement compared to the KD method.

Type				
	Teacher	ResNet110 74.31	ResNet110 74.31	ResNet32×4 79.42
	Student	ResNet8 61.37	MobileNet-V2 65.28	MobileNet-V2 65.28
Feature	FitNet[2]	57.67	61.66	60.83
	OFD[17]		61.48	
	CRD[44]		65.03	
	SimKD[9]	53.14	69.41	69.13
Logit	Vanilla KD[18]	61.41	68.95	66.83
	DKD[52]	58.91	69.25	68.79
	CTKD[28]	62.89	69.42	68.42
	LSKD[42]	60.97	68.92	68.84
	ours	63.99	69.68	69.55
	Δ	+2.58	+0.73	+2.72

Table 1 compares our method with recent state-of-the-art KD methods like DKD[52], SimKD[9], CTKD[28], and LSKD[42] from the past three years. Notably, in scenarios with significant model discrepancies, our method also achieved the best performance. This does not imply that our method universally surpasses all other methods. In situations with smaller model discrepancies, other methods tend to perform better. However, in cases of significant model discrepancies, other methods often perform poorly, sometimes even worse than vanilla KD. This highlights the necessity of researching and developing KD methods specifically tailored for scenarios involving large model discrepancies.

4.3. Part 2: Comparison with Previous Baseline for Addressing Significant Model Capacity Gaps

For CIFAR-10 and CIFAR-100, we adopt ResNet and plain CNN architectures as the foundational models to establish our baseline performance metrics. The experimental setups are similar to the TAKD method[33]. Specifically, for the ResNet-based evalua-

Table 2: Comparing the test accuracy of Pro-KD[38], TAKD[33], Annealing-KD[21], RCO[22], regular KD[18], and student without teacher on CIFAR-10 dataset with both ResNet and CNN models

Model	Type	Training Method	Accuracy
ResNet	Teacher(110)	from scratch	94.30
	TA(20)	Vanilla KD	93.35
	TA*(20)	DTM	93.68
	Student(8)	from scratch	87.44
	Student(8)	Vanilla KD	87.89
	Student(8)	TAKD	88.47
	Student(8)	RCO	88.90
	Student(8)	Annealing KD	89.44
	Student(8)	Pro-KD	90.01
	Student(8)	ours	90.18
CNN	Teacher(10)	from scratch	90.1
	TA(4)	Vanilla KD	82.39
	TA*(4)	DTM	83.86
	Student(2)	from scratch	72.75
	Student(2)	Vanilla KD	72.43
	Student(2)	TAKD	72.63
	Student(2)	Annealing KD	73.17
	Student(2)	ours	73.25

Table 3: Comparing the test accuracy of Pro-KD[38], TAKD[33], Annealing-KD[21], RCO[22], regular KD[18], and student without teacher on CIFAR-100 dataset with both ResNet and CNN models

ResNet	Teacher(110)	from scratch	74.31
	TA(20)	Vanilla KD	71.07
	TA*(20)	DTM	71.29
	Student(8)	from scratch	61.37
	Student(8)	Vanilla KD	61.41
	Student(8)	TAKD	61.82
	Student(8)	RCO	61.62
	Student(8)	Annealing KD	63.10
	Student(8)	Pro-KD	63.43
	Student(8)	ours	63.99
CNN	Teacher(10)	from scratch	64.89
	TA(4)	Vanilla KD	60.73
	TA*(4)	DTM	63.47
	Student(2)	from scratch	51.53
	Student(2)	Vanilla KD	51.62
	Student(2)	TAKD	51.85
	Student(2)	Annealing KD	53.35
	Student(2)	ours	53.79

tions, we designated ResNet-110 to serve as the teacher model, with ResNet-20 fulfilling the role of the Teaching Assistant (TA), and ResNet-8 positioned as the student model. In parallel, for experiments involving plain CNN structures[33], a 10-layer CNN was utilized as the teacher, a 4-layer CNN acted as the TA, and a 2-layer CNN was implemented as the student model. The comparative performance of our Gap-KD methodology against other established baselines across both CIFAR-10 and CIFAR-100 datasets is systematically presented in Table 2 and Table 3. For the ResNet baselines, the teacher ResNet-110 is trained from scratch. The TA is trained by the teacher via vanilla KD, while the TA* is trained using KD enhanced by DTM. Then we would like to train a ResNet-8 student via various methods and compare their accuracies against our Gap-KD method. Notably, our method utilizes TA* as the teacher assistant model while TAKD uses TA.

Table 4: Hyperparameter settings used in the implementation of our proposed method.

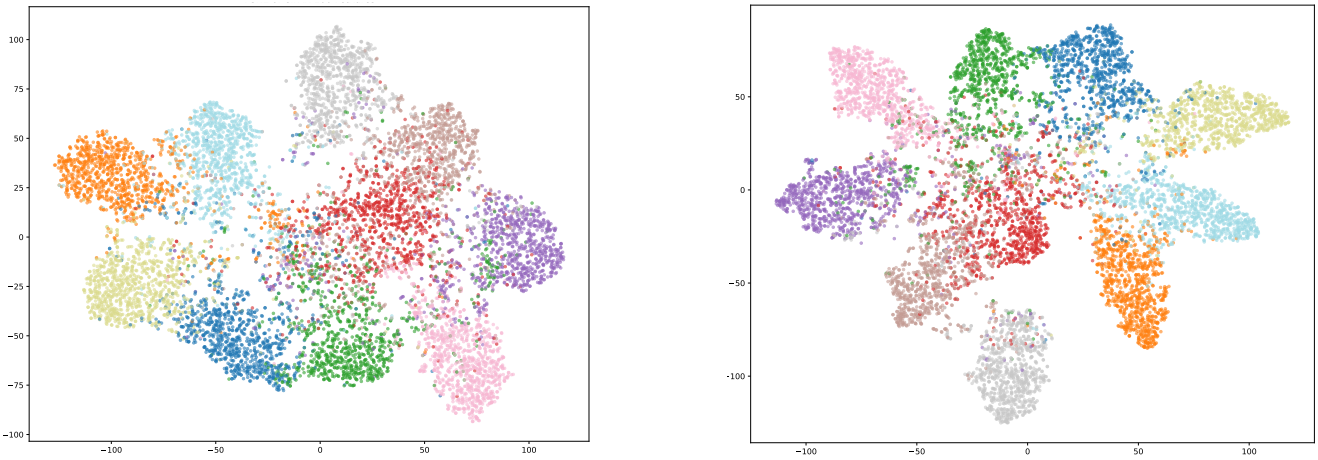
Hyper-parameter	ResNet(110→8)	CNN(10→2)
Batch Size	64	64
Learning Rate	0.05	0.01
Epochs	240	160
Decay Epochs	[150, 180, 210]	-
Weight Decay	5e-4	-
Momentum	0.9	0.9
Ce Weight	0.68	3.10
α	8.30	5.88
β	6.20	9.09
τ_{max}	24	20
τ_{min}	1	1
Warm Up	7	33

4.4. Visualization

To intuitively demonstrate the effectiveness of our method, we use t-SNE[47] visualizations to show the classification results on the CIFAR-10 dataset, where the teacher network is ResNet110 and the student network is ResNet8. As shown in Figure 6, compared to traditional KD, the model trained with Gap-KD exhibits more compact clusters within the same class and greater separation between different classes, demonstrating the effectiveness of our method.

5. Ablation Studies

To further validate the effectiveness of the different modules within our Gap-KD method and to determine the optimal values for various hyper-parameters, we designed a series of ablation experiments. These experiments systematically remove individual modules and adjust hyper-parameters to assess their impact on the overall performance of the model.



(a) Vanilla KD (b) Ours

Figure 6: t-SNE visualization of the student logits.

5.1. Ablation Study of the Modules

Table 5: Results of ablation studies in CIFAR-10

Model	DTM	D2M	Accuracy
T:ResNet110 S:ResNet8	×	×	87.89
	√	×	87.34
	×	√	89.20
	√	√	90.18
T:CNN10 S:CNN2	×	×	72.43
	√	×	72.52
	×	√	72.80
	√	√	73.25

Table 6: Results of ablation studies in CIFAR-100

Model	DTM	D2M	Accuracy
T:ResNet110 S:ResNet8	×	×	61.41
	√	×	62.55
	×	√	61.72
	√	√	63.99
T:CNN10 S:CNN2	×	×	51.62
	√	×	53.69
	×	√	53.51
	√	√	53.79

Table 5 and Table 6 present the results of ablation experiments conducted on the CIFAR-10 and CIFAR-100 datasets, respectively. Each table outlines two distinct setups: the first utilizes ResNet110 as the teacher model and ResNet8 as the student model, while the second employs CNN10 as the teacher model with CNN2 serving as the student model. It is evident from the data that both the DTM and D2M modules contribute positively to the enhancement of distillation outcomes. Notably, when these modules are combined, they achieve state-of-the-art performance levels.

5.2. Dynamic Temperature Adjustment Strategy

Table 7 reports presents the outcomes of employing various temperature adjustment strategies (Fixed, Linear, Exponential, Logarithmic, Sigmoid, Piecewise, DTM) in the context of knowledge distillation experiments. The experiments were conducted under two distinct configurations: the first row details results where a ResNet110 teacher model and a ResNet20 student model were utilized on the CIFAR-100 dataset, while the second row showcases outcomes using a Plane10 teacher model with a Plane4 student model on the CIFAR-10 dataset. Accuracy (Acc) percentages are reported for each strategy across both setups, highlighting the effectiveness of these methods in transferring knowledge from teacher to student models. Notably, the DTM strategy demonstrates superior performance in both configurations, underscoring its efficacy in temperature adjustment for knowledge distillation. Additionally, for all strtegies, $\tau_{max} = 20$ and $\tau_{min} = 1$.

5.3. DTM Parameters

Table 8 shows the ablation study results for varying values of the hyper-parameter τ_{max} of the DTM. The experiment evaluates the impact of τ_{max} on the accuracy (Acc) of knowledge distillation from a ResNet110(teacher network) to a ResNet32(student network). Values of τ_{max} ranging from 5 to 30 are

Table 7: Comparison of temperature adjustment strategies in knowledge distillation experiments across different architectures and datasets. And cross all experimental setups, DTM consistently achieves the best results.

Strategy	Fixed	Linear	Exponential	Logarithmic	Sigmoid	Piecewise	DTM
Acc1	70.85	70.71	70.80	71.12	71.23	69.84	71.29
Acc2	83.44	83.73	83.67	84.24	83.65	83.91	84.07

Table 8: Ablation study Results for the hyperparameter τ_{max} in the DTM, $\tau_{max} = 20$ works the best.

τ_{max}	5	10	15	20	25	30
Acc	73.11	72.70	72.86	73.25	73.02	73.12

considered, with the accuracy measured to assess the effectiveness of each setting. The experiment demonstrates that $\tau_{max} = 20$ yields the highest accuracy, indicating an optimal point for this hyperparameter in the context of transferring knowledge between these specific network architectures.

5.4. Distillation Paths

Table 9: Ablation study Results for distillation paths

PATH	TAKD	DGKD	ours
$10 \to 8 \to 6 \to 4 \to 2$	45.14	48.92	-
$10 \to 4 \to 2$	44.92	48.61	53.79

Table 9 demonstrates the performance of TAKD, DGKD, and our method on the CIFAR-100 dataset, using CNN10 as the teacher model and CNN2 as the student model across different distillation paths. The results indicate that for TAKD and DGKD, a more complete distillation path yields better performance. In contrast, our method outperforms both techniques even when utilizing only one teaching assistant, despite them employing three teaching assistants. This finding confirms that introducing a single assistant is sufficient to significantly alleviate the distillation path issue when using the Gap-KD approach.

6. Conclusion

In this paper, we present a two-stage distillation method incorporating DTM and D2M to effectively bridge the large gap between teacher and student models. This approach utilizes a single TA and transitions from high to low temperatures to ease learning complexity, markedly alleviating the Capacity Gap and Distillation Path problems. Additionally, our dual decoupling strategy prevents the Error Avalanche problem, enhancing model robustness. Employing these techniques, our proposed method achieves the state-of-the-art among the multiple distillation methods.

Future Work

Currently, Gap-KD has only been evaluated on image classification tasks. As a logit-based knowledge distillation method, it faces challenges in outperforming state-of-the-art feature-based approaches in object detection due to the absence of positional information in the logits. In future work, we aim to address this limitation by incorporating mechanisms that can better capture positional knowledge, thereby enhancing the applicability of Gap-KD to object detection. Although Gap-KD achieves state-of-the-art performance in scenarios with significant model capacity differences, its effectiveness slightly lags behind the leading methods when the model gap is smaller, and vice versa. In future research, we plan to develop an adaptive approach that can dynamically adjust to varying model capacity differences, ensuring optimal performance regardless of the size of the gap.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 62162065), the Joint Special Project Research Foundation of Yunnan Province (Grant No.202401BF070001-023), the Yunnan Province Visual and Cultural Innovation Team (Grant No.202505AS350009), and the Yunnan Fundamental Research Projects (Grant No.202201AT070167). Additionally, we acknowledge the support from the Practical Innovation Project of Postgraduate Students in the Professional Degree Program of Yunnan University (Grant No.ZC-23234683). We sincerely appreciate the financial and institutional support that has contributed to the completion of this work.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 1

[2] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua. Fitnets: Hints for thin deep nets. Proc. ICLR, 2(3):1, 2015. 7

[3] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9163–9171, 2019. 3

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481–2495, 2017. 1

[5] J.-H. Bae, D. Yeo, J. Yim, N.-S. Kim, C.-S. Pyo, and J. Kim. Densely distilled flow-based knowledge transfer in teacher-student framework for image classification. IEEE Transactions on Image Processing, 29:5698–5710, 2020. 1

[6] A. Bie, B. Venkitesh, J. Monteiro, M. A. Haidar, M. Rezagholizadeh, et al. Fully quantizing a simplified transformer for end-to-end speech recognition. arXiv preprint arXiv:1911.03604, 2019. 1

[7] C. Bucila, R. Caruana, and A. Niculescu-Mizil. Model compression. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'06), 2006. 3

[8] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen. Online knowledge distillation with diverse peers. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 3430–3437, 2020. 1

[9] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen. Knowledge distillation with the reused teacher classifier. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11933–11942, 2022. 1, 7

[10] J. H. Cho and B. Hariharan. On the efficacy of knowledge distillation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4794–4802, 2019. 3

[11] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani. A comprehensive survey on model compression and acceleration. Artificial Intelligence Review, 53:5113–5155, 2020. 1

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[13] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Trans. Pattern Anal. Mach. Intell., 45(6):7157–7173, jun 2023. 1

[14] J. Gou, B. Yu, S. J. Maybank, and D. Tao. Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 2021. 3

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. 1

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1

[17] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi. A comprehensive overhaul of feature distillation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1921–1930, 2019. 7

[18] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. Computer Science, 14(7):38–39, 2015. 1, 3, 4, 7, 8

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017. 1

[20] T. Huang, S. You, F. Wang, C. Qian, and C. Xu. Knowledge distillation from a stronger teacher. Advances in Neural Information Processing Systems, 35:33716–33727, 2022. 3

[21] A. Jafari, M. Rezagholizadeh, P. Sharma, and A. Ghodsi. Annealing knowledge distillation. arXiv preprint arXiv:2104.07163, 2021. 7, 8

[22] X. Jin, B. Peng, Y. Wu, Y. Liu, J. Liu, D. Liang, J. Yan, and X. Hu. Knowledge distillation via route constrained optimization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1345–1354, 2019. 4, 7, 8

[23] Y. Jin, J. Wang, and D. Lin. Multi-level logit distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24276–24285, 2023. 1, 7

[24] N. Komodakis and S. Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In ICLR, 2017. 1

[25] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[26] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, and D. Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In Proceedings of the AAAI conference on artificial intelligence, volume 36, pages 1306–1313, 2022. 3

[27] Z. Li, Y. Huang, D. Chen, T. Luo, N. Cai, and Z. Pan. Online knowledge distillation via multi-branch diversity enhancement. In Proceedings of the Asian Conference on Computer Vision, 2020. 3

[28] Z. Li, X. Li, L. Yang, B. Zhao, R. Song, L. Luo, J. Li, and J. Yang. Curriculum temperature for knowledge distillation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 1504–1512, 2023. 4, 7

[29] S. Lin, H. Xie, B. Wang, K. Yu, X. Chang, X. Liang, and G. Wang. Knowledge distillation via the target-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10915–10924, 2022. 3

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. 1

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019. 1

[32] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643, 2015. 3

[33] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 5191–5198, 2020. 2, 3, 4, 7, 8

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016. 1

[35] N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4512–4525, Online, Nov. 2020. Association for Computational Linguistics. 1

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015. 1

[37] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In International Conference on Learning Representations, 2021. 1

[38] M. Rezagholizadeh, A. Jafari, P. Salad, P. Sharma, A. S. Pasand, and A. Ghodsi. Pro-kd: Progressive distillation by following the footsteps of the teacher. arXiv preprint arXiv:2110.08532, 2021. 7, 8

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520, 2018. 7

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Y. Bengio and Y. LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 1

[41] W. Son, J. Na, J. Choi, and W. Hwang. Densely guided knowledge distillation using multiple teacher assistants. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9395–9404, 2021. 2, 4

[42] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao. Logit standardization in knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15731–15740, 2024. 7

[43] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239, 2022. 1

[44] Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. In International Conference on Learning Representations, 2020. 7

[45] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. 1

[46] F. Tung and G. Mori. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1365–1374, 2019. 1

[47] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008. 8

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 1

[49] L. Wang and K. Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE transactions on pattern analysis and machine intelligence, PP, 2021. 3

[50] K. You, M. Long, J. Wang, and M. I. Jordan. How does learning rate decay help modern neural networks? arXiv preprint arXiv:1908.01878, 2019. 5

[51] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4320–4328, 2018. 1

[52] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang. Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pages 11953–11962, 2022. 2, 5, 7

[53] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023. 1