Asymmetric Temperature Scaling Makes Larger Networks Teach Well Again

Xin-Chun Li¹, Wen-Shu Fan¹, Shaoming Song², Yinchuan Li², Bingshuai Li², Yunfeng Shao², De-Chuan Zhan¹
¹ State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
² Huawei Noah’s Ark Lab, Beijing, China
{lixc, fanws}@lamda.nju.edu.cn, zhandc@nju.edu.cn
{shaoming.song, liyinchuan, libingshuai, shaoyunfeng}@huawei.com

Abstract

Knowledge Distillation (KD) aims at transferring the knowledge of a well-performed neural network (the teacher) to a weaker one (the student). A peculiar phenomenon is that a more accurate model doesn’t necessarily teach better, and temperature adjustment can neither alleviate the mismatched capacity. To explain this, we decompose the efficacy of KD into three parts: correct guidance, smooth regularization, and class discriminability. The last term describes the distinctness of wrong class probabilities that the teacher provides in KD. Complex teachers tend to be over-confident and traditional temperature scaling limits the efficacy of class discriminability, resulting in less discriminative wrong class probabilities. Therefore, we propose Asymmetric Temperature Scaling (ATS), which separately applies a higher/lower temperature to the correct/wrong class. ATS enlarges the variance of wrong class probabilities in the teacher’s label and makes the students grasp the absolute affinities of wrong classes to the target class as discriminative as possible. Both theoretical analysis and extensive experimental results demonstrate the effectiveness of ATS. The demo developed in Mindspore is available at https://gitee.com/lxcnju/ats-mindspore and will be available at https://gitee.com/mindspore/models/tree/master/research/cv/ats.

1 Introduction

Although large-scale deep neural networks have achieved overwhelming successes in many real-world applications [22, 11, 60], the vast capacity hinders them from being deployed on portable devices with limited computation and storage resources [3]. Some efficient architectures, e.g., MobileNets [14, 37] and ShuffleNets [59, 29], have been proposed for lightweight deployment, while their performances are usually constrained. Fortunately, knowledge distillation (KD) [46, 13] could transfer the knowledge of a more complex and well-performed network (i.e., the teacher) to them.

The original KD [13] forces the student to mimic the teacher’s behavior via minimizing the Kullback-Leibler (KL) divergence between their output probabilities. Recent studies generalize KD to various types of knowledge [56, 57, 17, 12, 33, 1, 34, 44, 52, 54, 53, 50, 20, 23] or various distillation schemes [61, 20, 23]. An intuitive sense after the proposal of KD [13] is that larger teachers could teach students better because their accuracies are higher. A recent work [6] first points out that the teacher accuracy is a poor predictor of the student’s performance. That is, more accurate neural networks don’t necessarily teach better. Until now, this phenomenon is still counter-intuitive [51], surprising [31], and unexplored [24]. Different from some existing empirical studies and theoretical analysis [40, 18, 30, 34, 55, 53, 6, 28, 15], we investigate the miraculous phenomenon in detail and aim to answer the following questions: What’s the real reason that more complex teachers can’t teach...
Teacher Student
Decomposition of KD
Teacher's Label
Correct Guidance
Smooth Regularization
Class Discriminability
Temperature
Lower
Higher

Figure 1: Left: Decomposition of a teacher’s label. The first class is the target. As temperature increases, correct guidance is weaker, smooth regularization is stronger, while class discriminability (measured by the variance of wrong class probabilities) will first increase and then decrease. Right: Larger/Smaller teachers’ logits are consistent in relative class affinities, i.e., logit values of the four wrong classes are in the same order of magnitude. However, larger teachers are over-confident and give a larger target logit or smaller inherent variance, leading to a smaller derived variance under traditional temperature scaling, i.e., less distinct wrong class probabilities after softmax.

well? Is it really impossible to make larger teachers teach better through simple operations, such as temperature scaling?

To answer the first question, we focus on analyzing the distinctness of wrong class probabilities that a teacher provides in KD. We decompose the teacher’s label into three parts (see Sect. 4.1): (I) Correct Guidance: the correct class’s probability; (II) Smooth Regularization: the average probability of wrong classes; (III) Class Discriminability: the variance of wrong class probabilities (defined as derived variance). The commonly utilized temperature scaling could control the efficacy of these three terms (the left of Fig. [1]). More complex teachers are over-confident and assign a larger score for the correct class or less varied scores for the wrong classes. If we use a uniform temperature to scale their logits, the class discriminability of the larger teacher is less effective (theoretically analyzed in Sect. 4.2), i.e., the probabilities of wrong classes are less distinct (the right of Fig. [1]).

As to the second question, we focus on enlarging the variance of wrong class probabilities (i.e., derived variance) that a teacher provides to make the distillation process more discriminative. To specifically enhance the distinctness of wrong class probabilities, we separately apply a higher/lower temperature to the correct/wrong class’s logit instead of a uniform temperature (see Sect. 4.3). We name our method Asymmetric Temperature Scaling (ATS), and abundant experimental studies verify that utilizing this simple operation could make larger teachers teach well again.

2 Related Works

KD with Larger Teacher: Although KD has been a general technique for knowledge transfer in various applications [13, 61, 42, 25], could any student learn from any teacher? [6] first studies the KD’s dependence on student and teacher architectures. They find that larger models do not often make better teachers and propose the early-stopped teacher as a solution. [31] introduces a multi-step KD process, employing an intermediate-sized network (the teacher assistant) to bridge the capacity gap. [51] formulates KD as a multi-task learning problem with several knowledge transfer losses. The transfer loss will be utilized only when its gradient direction is consistent with the cross-entropy loss. [10, 24] define the knowledge gap as residual, which is utilized to teach the residual student, and then they take the ensemble of the student and residual student for inference. These works attribute the worse teaching performance to capacity mismatch, i.e., weaker students can’t completely mimic the excellent teachers. However, they don’t explain this peculiar phenomenon in detail.

Understanding of KD: Quite a few works focus on understanding the advantages of KD from a principled perspective. [28] unifies KD and privileged information into generalized distillation. [35, 18] utilize gradient flow and neural tangent kernel to analyze the convergence property of KD under deep linear networks and infinitely wide networks. [5] explains KD via quantifying the task-
We consider a $C$-class classification problem with $\mathcal{Y} = [C] = \{1, 2, \ldots, C\}$. Given a neural network and a sample pair $(x, y)$, we could obtain the “logits” as $f(x) \in \mathbb{R}^C$. We denote the softmax function with temperature $\tau$ as $\text{SF}(\cdot; \tau)$, i.e., $p_c(\tau) = \exp(f_c(x)/\tau)/Z(\tau)$ and $Z(\tau) = \sum_{c=1}^{C} \exp(f_c(x)/\tau)$, where $p(\tau)$ is the softened probability vector that a network outputs and $c$ is the index of class. Later, we may omit the dependence on $x$ and $\tau$ if without any ambiguity. We use $f_c$ and $p_c$ to denote the correct class’s logit and probability, while we use $g$ and $q$ to represent the vector of wrong classes’ logits and probabilities, i.e., $g = [f_c]_{c \neq y}$ and $q = [p_c]_{c \neq y}$. The notations could be found in Tab. 1.

The most standard KD [13] contains two stages of training. The first stage trains complex teachers, and then the second stage transfers the knowledge from teachers to a smaller student via minimizing the KL divergence between softened probabilities. Usually, the loss function during the second stage (i.e., the student’s learning objective) is a combination of cross-entropy loss and distillation loss:

$$
\ell = -(1 - \lambda) \log p_y^S(1) - \lambda \tau^2 \sum_{c=1}^{C} p_c^T(\tau) \log p_c^S(\tau),
$$

where the upper script “$T$”/“$S$” denotes “Teacher”/“Student” respectively. Commonly, a default temperature of 1 is utilized for the CE loss, and the student could also take a temperature of 1 for the KD loss, e.g., $p^S(\tau = 1)$ [13, 31, 45, 44].

Suppose we have two teachers, denoted as $T_{\text{large}}$ and $T_{\text{small}}$, and the larger teacher performs better on both training and test data. If we use them to teach the same student $S$, we could find that the student’s performance is worse when mimicking the larger teacher’s outputs. Adjusting the temperature could neither make the larger teacher teach well. The details of this phenomenon could be found in [6, 31] and Fig. 9. Obviously, $p^{T_{\text{large}}}$ could differ a lot from $p^{T_{\text{small}}}$, which is the only difference in the loss function when teaching the student. Hence, we focus on analyzing what probability distributions are tended to be provided by teachers with different capacities.

### Table 1: The used notations in this paper. The definitions of Derived Average, Derived Variance and Inherent Variance are only for wrong classes (Sect. 4.1 and Sect. 4.2).

<table>
<thead>
<tr>
<th>Notations</th>
<th>All Classes</th>
<th>Wrong Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit</td>
<td>$f$</td>
<td>$g = [f_c]_{c \neq y}$</td>
</tr>
<tr>
<td>Probability</td>
<td>$p = \text{SF}(f)$</td>
<td>$q = [p_c]_{c \neq y}$</td>
</tr>
<tr>
<td>Derived Average of Probabilities</td>
<td>$\epsilon(q) = \sum_j q_j/(C - 1)$</td>
<td>$v(q) = \sum_j (q_j - \epsilon(q))^2/(C - 1)$</td>
</tr>
<tr>
<td>Derived Variance of Probabilities</td>
<td>$\bar{q} = \text{SF}(g), v(q) = \sum_j (\bar{q}_j - \epsilon(q))^2/(C - 1)$</td>
<td></td>
</tr>
<tr>
<td>Inherent Variance of Probabilities</td>
<td>$\bar{q} = \text{SF}(g), v(q) = \sum_j (\bar{q}_j - \epsilon(q))^2/(C - 1)$</td>
<td></td>
</tr>
</tbody>
</table>
4 Proposed Methods

This section first decomposes KD into three parts and defines several quantitative metrics. Then, we present theoretical analysis to demonstrate why larger networks can’t teach well. Finally, we propose a more appropriate temperature scaling approach as an alternative.

4.1 KD Decomposition

We omit the coefficient of $\lambda \tau^2$ in Eq. (1) and define $e(q^T(\tau)) = \frac{1}{C-1} \sum_{j:y \neq c} p^T_j(\tau)$, where $q^T(\tau) = [p^T_c(\tau)]_{c \neq y}$. Then, we have the following decomposition:

\[
\ell_{kd} = -p^T_y(\tau) \log p^S_y(\tau) - \sum_{c \neq y} e(q^T(\tau)) \log p^S_c(\tau) - \sum_{c \neq y} (p^T_c(\tau) - e(q^T(\tau))) \log p^S_c(\tau). \tag{2}
\]

(I) Correct Guidance: this term guarantees correctness during teaching. The decomposition in (9) also contains this term, which is explained as importance weighting. This term works similarly to the cross-entropy loss, which could be dealt with separately when applying temperature scaling.

(II) Smooth Regularization: some previous works [55, 62, 63] attribute the success of KD to the efficacy of regularizing and studying its relation to label smoothing (LS). The combination of this term with correct guidance works similarly to LS. Notably, $e(q^T(\tau))$ differs across samples, implying that the strength of smoothing is instance-specific, which is similar to the analysis in [62].

(III) Class Discriminability: this term tells the student the affinity of wrong classes to the correct class. Transferring the knowledge of class similarities to students has been the mainstream guess of the “dark knowledge” in KD [13, 38]. Ideally, a good teacher should be as discriminating as possible in telling students which classes are more related to the correct class.

Illustrations of the decomposition are presented in the left of Fig. (1). Obviously, an appropriate temperature should simultaneously contain the efficacy of the three terms, e.g., the shaded row in Fig. (1). A too high or too low temperature could lead to smaller class discriminability, making the guidance less different among wrong classes, which weakens the distillation performance in practical. Among these three terms, we advocate that class discriminability is more fundamental in KD and present more discussions in Appendix A (verified in Fig. 2 and Fig. 3).

To measure these three terms quantitatively, we use the target class probability (i.e., $p_y$), the average of wrong class probabilities (i.e., $\bar{e}(q) = \frac{1}{C-1} \sum_{j:y \neq c} p_j$), and the variance of wrong class probabilities (i.e., $\sigma(q) = \frac{1}{C-1} \sum_{j:y \neq c} (p_j - \bar{e}(q))^2$) as estimators. $\bar{e}(\cdot)$ and $\sigma(\cdot)$ calculates the mean and variance of the elements in a vector. In some cases, we use the standard deviation as an estimator for the third term, i.e., $\sigma(q) = \sigma^{1/2}(q)$. Because the latter two terms are calculated after applying softmax to the complete logit vector, we define them as Derived Average (DA) and Derived Variance (DV), respectively. In experiments, we calculate these metrics for all training samples and sometimes report the average or standard deviation across these samples.

4.2 Theoretical Analysis

We analyze the mean and variance of the softened probability vector, i.e., the teacher’s label $p^T(\tau)$ used in KD. We defer the proofs of Lemma 4.1 and Proposition 4.3 to Appendix B.

**Lemma 4.1 (Variance of Softened Probabilities)**. Given a logit vector $f \in \mathbb{R}^C$ and the softened probability vector $p = SF(f; \tau)$, $\tau \in (0, \infty)$, $\sigma(p)$ monotonically decreases as $\tau$ increases.

As $\tau$ increases, $p(\tau)$ becomes more uniform, i.e., its entropy increases. However, we especially focus on the wrong classes, where the mean and variance are more intuitive to calculate and analyze.

**Assumption 4.2**. The target logit is higher than other classes’ logits, i.e., $f_y \geq f_c, \forall c \neq y$.

Assumption 4.2 is rational because well-performed teachers could almost achieve a higher accuracy (e.g., >95%) on the training set, and most training samples meet this requirement.

**Proposition 4.3**. Under Assumption 4.2, $p_y$ monotonically decreases as $\tau$ increases, and $\bar{e}(q)$ monotonically increases as $\tau$ increases. As $\tau \to \infty$, $\bar{e}(q) \to 1/C$. 

\[
\bar{e}(q) = \frac{1}{C-1} \sum_{j:y \neq c} p_j = e(q^T(\tau)) \geq e(q^T(\infty)) = \frac{1}{C-1} \sum_{j:y \neq c} e(q^T(\infty)) = \frac{1}{C-1} C \ln \frac{1}{C}.
\]
Proposition 4.4 implies that increasing temperature could lead to a higher derived average (empirically see Fig. 7) and strengthen the smooth regularization term in Eq. 2.

Before we analyze the class discriminability term, we define \( \bar{q}(\tau) \) as the result of applying softmax only to the wrong logits with temperature \( \tau \), i.e., \( q(\tau) = \text{SF}(g; \tau) \). For the element index \( c' \) of \( q \), we have

\[
\bar{q}_{c'}(\tau) = \exp(g_{c'}/\tau) / \sum_j \exp(g_j/\tau).
\]

Notably, \( \bar{q} \) differs from \( q \) a lot. Specifically, the former satisfies \( \sum_{c'} \bar{q}_{c'} = 1 \), while the summation of the latter is \( \sum_{c} q_c = 1 \). The former does not depend on the correct class’s logit while the latter does. We name \( v(\bar{q}) \) Inherent Variance (IV) because it only depends on wrong classes’ logits.

**Proposition 4.4 (Derived Variance vs. Inherent Variance).** The derived variance is determined by the square of derived average and the inherent variance via:

\[
v(q) = (C-1)^2 \frac{\epsilon^2(q)}{\epsilon^2(q)} \cdot v(\bar{q}).
\]

With \( \tau \) increases, \( \epsilon(q) \) increases (Proposition 4.3) while \( v(q) \) decreases (Lemma 4.1), and hence, it is not so easy to judge the specific monotonicity of \( v(q) \) w.r.t. \( \tau \). Empirically, we observe that the derived variance first increases and then decreases (see Fig. 7), which conforms to the change of the class discriminability as illustrated in Fig. 1.

We could use Proposition 4.4 to clearly analyze why larger teacher networks can’t teach well. Before this, we present another two properties and a corollary without detailed proof.

**Remark 4.5.** Fixing \( g \) and \( \tau \), a higher target logit \( f_y \) leads to a higher \( p_y \), i.e., a smaller derived average \( \epsilon(q) \).

**Remark 4.6.** Fixing \( \tau \), less varied wrong logits \( g \) leads to less varied \( q \), i.e., a smaller inherent variance \( v(q) \).

**Corollary 4.7.** Suppose we have two teachers \( T_1 \) and \( T_2 \), and their logit vectors for a same sample are \( f^{T_1} \) and \( f^{T_2} \).

- If \( f_y^{T_1} \geq f_y^{T_2} \) while \( g^{T_1} = g^{T_2} \), then \( p_y^{T_1} \geq p_y^{T_2} \) (Remark 4.5) while \( v(q^{T_1}) \approx v(q^{T_2}) \). Hence, \( v(q^{T_1}) \leq v(q^{T_2}) \).

- If \( f_y^{T_1} \approx f_y^{T_2} \) while \( v(g^{T_1}) \leq v(g^{T_2}) \), then \( p_y^{T_1} \approx p_y^{T_2} \) while \( v(q^{T_1}) \leq v(q^{T_2}) \) (Remark 4.6). Hence, \( v(q^{T_1}) \leq v(q^{T_2}) \).

This corollary explains why a larger teacher can’t teach better. Because the larger teacher tends to be over-confident, the target logit \( f_y \) may be larger or the variance of wrong logits \( v(g) \) may be smaller. These are illustrated in Fig. 1 and empirically verified in Fig. 4. Then the derived variance \( v(q) \) may be smaller, limiting the efficacy of class discriminability in Eq. 2. Empirical results are in Fig. 7.

Notably, we focus on analyzing the variance of wrong class probabilities instead of all classes. Maximizing the variance of all classes’ probabilities does not mean maximizing the variance of wrong classes’. For example, although a very low temperature can maximize the variance of all classes’ probabilities, the generated teacher’s label is one-hot that shows no distinctness between wrong classes. In other words, the effectiveness of KD should be more related to the distinctness between wrong classes rather than all classes. However, traditional temperature scaling applies a uniform temperature for all classes, which cannot separately handle the wrong classes.

### 4.3 Asymmetric Temperature Scaling

We conclude the above analysis: if a larger teacher makes an over-confident prediction, the wrong class probabilities it provides could be not discriminative enough. Utilizing a uniform temperature could not enlarge the derived variance as much as possible with the interference of the target class’s logit (see the middle of Fig. 7). Thanks to the decomposition in Eq. 2, the correct guidance term works similarly to the cross-entropy loss and allows us to deal with it separately. Hence, we propose a novel temperature scaling approach:

\[
p_c(\tau_1, \tau_2) = \exp(f_i/\tau_1) / \sum_{j \in [C]} \exp(f_j/\tau_j), \quad \tau_i = I\{i = y\} \tau_1 + I\{i \neq y\} \tau_2, \forall i \in \mathbb{C},
\]

As illustrated in Fig. 5, for a same sample, we have

\[
\|f^{T_1}\|_2 \leq \|f^{T_2}\|_2 \Rightarrow \|f^{T_1} - f^{T_2}\|_2 \leq \|g^{T_1} - g^{T_2}\|_2 \Rightarrow \|v(q^{T_1}) - v(q^{T_2})\|_2 \leq \|v(g^{T_1}) - v(g^{T_2})\|_2.
\]
where we take $\tau_1 > \tau_2 > 0$. This approach is named Asymmetric Temperature Scaling (ATS) because we apply different temperatures to the logits of correct and wrong classes. According to Eq. 4, ATS could bring such benefits when the teacher is over-confident:

- If the teacher outputs a larger logit $f_q$ for the correct class, a relatively larger $\tau_1$ could decrease it to a reasonable magnitude, i.e., decreasing $p_q$ and increasing $e(q)$, and finally increasing the derived variance $v(q)$;
- If the teacher outputs less varied logits $g$ for wrong classes, a relatively smaller temperature $\tau_2$ could make them more diverse, i.e., increasing $v(\tilde{q})$, finally increasing the derived variance $v(q)$.

ATS is more flexible in enlarging the derived variance (see the right of Fig. 7), i.e., it could generate more discriminative distillation guidance during teaching. Take the demo in Fig. 1 as an example, the smaller and larger teacher captures the same relative class affinities, e.g., they both know that the fourth/fifth class (shaded cells) is the most/least relevant to the target class. However, with a uniform temperature 4.0, the smaller teacher provides probabilities $(0.073, 0.060)$ for these two classes, while over-confident larger teachers provide $(0.040, 0.033)$ or $(0.074, 0.067)$. Clearly, the absolute affinities of the larger teachers are not so discriminative as the smaller teacher’s. Utilizing ATS, we could respectively apply $(\tau_1 = 4.67, \tau_2 = 4.0)$ or $(\tau_1 = 4.0, \tau_2 = 2.0)$ to the over-confident teacher’s logits, generating the same probability vector as the smaller teacher’s. ATS utilizes two temperatures and creates the wiggle room to make the distribution over wrong classes more discriminative.

5 Experiments

We use CIFAR-10/CIFAR-100 [21], TinyImageNet [43], CUB [47], Stanford Dogs [19], and Google Speech Commands [48] as the datasets. For teacher networks, we use different versions of ResNet [11], WideResNet [56], ResNeXt [49]. For student networks, we use VGG [39], ShuffleNetV1/V2 [59, 29], AlexNet [22], MobileNetV2 [37], and DSCNN [60].

We majorly follow the training settings in [44]. Except that the Google Speech Commands takes 50 epochs, we train networks on other datasets with 240 epochs. We use the SGD optimizer with 0.9 momentum. For VGG, AlexNet, ResNet, WideResNet, and ResNeXt, we set the learning rate as 0.05 (recommended by [44]). For ShuffleNet and MobileNet, we use a smaller learning rate of 0.01 (recommended by [44]). We use the pre-trained models provided in PyTorch for CUB and Stanford Dogs, and correspondingly, their learning rates are scaled by $0.1 \times$. During training, we decay the learning rate by 0.1 every 30 epochs after the first 150 epochs (recommended by [44]). For Google Speech Commands, we decay the learning rate via the cosine annealing. We set the batch size as 128 for CIFAR data, 64 for other datasets. Other dataset, network and training details are in Appendix C.

5.1 Observations

Class discriminability matters a lot in KD and correlates with the KD improvement. To show the importance of class discriminability in KD, we omit the distinctness of wrong class probabilities during distillation. Specifically, we only keep the first two terms in Eq. 2 which works similarly to

\[ \text{Derived Average} \]

\[ \text{Derived Variance} \]

\[ \text{Class Discriminability} \]

\footnotesize{https://github.com/HobbitLong/RepDistiller}
Figure 4: The distributions of the \textit{target logit} ($f_y$) and the \textit{standard deviation of wrong logits} ($\sigma(g)$) of the 50K training samples on CIFAR-10/100. Rows show networks with various capacity.

Figure 5: The \textit{derived variance} and \textit{inherent variance} on CIFAR-100 using ResNeXt. Each bar shows the mean and standard deviation across 50K training samples. The x-axis shows networks with various capacity.

Figure 6: Several metrics among teachers with various capacities on CIFAR-10. The first three metrics are related to the relative magnitudes of teachers’ label, while distillation distance is related to the absolute magnitudes.

Figure 7: The change of \textit{derived average} ($e(q)$) and \textit{derived variance} ($\sigma(q)$) as $\tau$ increases from 0.1 to 10.0 on CIFAR-10. The third one shows the results of ResNet110 with the proposed ATS. \textit{DV} under TS is limited while ATS enlarges it.

the instance-specific label smoothing (abbreviated as “ILS”). Fig. 2 shows students’ performances on CIFAR-100. Without the third term in Eq. 2 both small teachers (“ST”) and larger teachers (“LT”) teach worse significantly. Then, we investigate the correlations of KD performance improvement (i.e., the test accuracy change ratio with KD w.r.t. without KD) with smooth regularization and \textit{class discriminability} under 270 pairs of \textit{“(teacher, student, temperature)”}. Details are in Appendix C.4.

Fig. 3 plots the scatters, where dots show pairs whose improvement is higher than 2%. Clearly, teachers with a larger \textit{derived variance} tend to guide better. These observations show the rationality of the proposed KD decomposition and imply that enhancing derived variance is beneficial.

Larger teachers provide a larger target logit or less varied wrong logits. We first compare the logit distributions provided by larger and smaller teachers on the training set. Fig. 4 plots the histograms (200 bins) of the target logit (i.e., $f_y$) and the standard deviation of wrong logits (i.e., $\sigma(g)$). The left and right, respectively, show the results on CIFAR-100 and CIFAR-10. Clearly, the first column shows that ResNet110 tends to generate a larger target logit (i.e., $E_x[f_y] \approx 15.0$) than ResNet14 (i.e., $E_x[f_y] \approx 10.0$). On CIFAR-10, the smallest $f_y$ given by WRN28-8 is larger than WRN28-1. Furthermore, WRN28-8 gives smaller variance (the fourth column), i.e., smaller \textit{inherent variance}. These reveal specific manifestations of complex networks’ over-confidence.

Larger and smaller teachers have similar inherent variance while different derived variance under \textit{traditional temperature scaling}. We use $\tau = 1.0$ to soften the complete logits and only the wrong class logits, respectively, and then show the mean and standard deviation of \textit{derived variance} and \textit{inherent variance} across training samples in Fig. 5. Although the larger models’ \textit{inherent variance} is smaller, the difference between RNX29-4-4d and RNX29-64-4d is only up to 1.3×. However, the \textit{derived variance} differs a lot, where the smaller teacher’s variance is approximately 9.7× as the larger teacher’s. These observations imply that \textit{traditional temperature scaling} could seriously decrease the \textit{derived variance} of larger teachers though they have appreciable inherent variance.

Complex teachers know approximately the same as smaller teachers on relative class affinities. Only if the relative magnitudes of the wrong class probabilities are correct, it is valid to enlarge the discrimination between them. Otherwise, if a teacher himself misunderstands the knowledge’s principles, it will be counterproductive to reinforce this knowledge to students. Given two teachers $T_1$ and $T_2$, we first calculate the \textit{Spearman correlation} between $p^{T_1}$ and $p^{T_2}$. Second, the set overlap...
The proposed ATS could enlarge the derived variance of larger teachers. The above analysis verifies that the over-confident teachers experience lower derived variance under traditional temperature scaling although they grasp the relative class affinities well. For an intuitive visualization, we plot the probabilities for a randomly sampled instance whose correct class is $y = 1$ from CIFAR-100. The top two rows in Fig. 8 show the results of RNX29-4-4d and RNX29-64-4d under traditional temperature scaling (TS), where the latter really experiences a smaller derived variance. Using ATS could enhance the derived variance as shown at the bottom row (the bars are more jagged). Then, we study the change of derived average (DA) and derived variance (DV) as temperature increases. Given a $\tau$, we obtain the DA and DV for all training samples via softmax and then calculate the average. For ATS, we use $\tau_1 = 1.25\tau$ and $\tau_2 = 0.75\tau$. The curves are shown in Fig. 7. According to Proposition 4.3, $\sigma(q)$ increases as $\tau$ increases, which enhances the efficacy of smooth regularization.
Table 3: Comparisons with SOTA methods on TinyImageNet, CUB, and Stanford Dogs. WRN50-2 and RNX101-32-8d are teachers. AlexNet, SFV2, and MV2 are students.

<table>
<thead>
<tr>
<th>Teacher</th>
<th>TinyImageNet</th>
<th>CUB</th>
<th>Stanford Dogs</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NoKD</td>
<td>WRN50-2 (66.28)</td>
<td>RNX101-32-8d (79.30)</td>
<td>RNX101-32-8d (73.98)</td>
<td></td>
</tr>
<tr>
<td>ANet</td>
<td>34.62</td>
<td>45.79</td>
<td>52.03</td>
<td>50.20</td>
</tr>
<tr>
<td>SFV2</td>
<td>34.62</td>
<td>45.79</td>
<td>52.03</td>
<td>50.20</td>
</tr>
<tr>
<td>MV2</td>
<td>34.62</td>
<td>45.79</td>
<td>52.03</td>
<td>50.20</td>
</tr>
<tr>
<td>KD</td>
<td>55.69</td>
<td>69.92</td>
<td>72.85</td>
<td>68.48</td>
</tr>
<tr>
<td>ESKD</td>
<td>55.69</td>
<td>69.92</td>
<td>72.85</td>
<td>68.48</td>
</tr>
<tr>
<td>TAKD</td>
<td>55.69</td>
<td>69.92</td>
<td>72.85</td>
<td>68.48</td>
</tr>
<tr>
<td>SCKD</td>
<td>55.69</td>
<td>69.92</td>
<td>72.85</td>
<td>68.48</td>
</tr>
<tr>
<td>KD+ATS</td>
<td>55.69</td>
<td>69.92</td>
<td>72.85</td>
<td>68.48</td>
</tr>
<tr>
<td>Ens</td>
<td>55.69</td>
<td>69.92</td>
<td>72.85</td>
<td>68.48</td>
</tr>
<tr>
<td>ResKD</td>
<td>55.69</td>
<td>69.92</td>
<td>72.85</td>
<td>68.48</td>
</tr>
<tr>
<td>KD+ATS+Ens</td>
<td>55.69</td>
<td>69.92</td>
<td>72.85</td>
<td>68.48</td>
</tr>
</tbody>
</table>

Notably, this term changes nearly the same between teachers with various capacities. However, the derived variance \(v(q)\) differs a lot. Empirically, \(v(q)\) first increases and then decreases, and the maximal of the larger teacher’s DV is smaller, which verifies the Corollary 4.7. Because derived variance corresponds to the efficacy of class discriminability, this shows why larger teachers can’t teach well. Using the proposed ATS could enhance the derived variance, which equivalently improves the efficacy of class discriminability in KD. We conclude that traditional temperature scaling leads to distillation labels with less discriminative information among wrong class probabilities; our proposed ATS could enhance the discrimination among wrong classes and benefit the distillation process.

5.2 Performances

**ATS makes larger teachers teach well again.** Previous studies find that more accurate teachers can’t necessarily teach well [6, 31]. As shown in Fig. 9, although we tune temperatures in \(\{1.0, 2.0, 4.0, 8.0, 12.0, 16.0\}\), larger teachers still teach worse under traditional temperature scaling (the solid curves). However, using ATS (the dashed curves) could make larger teachers teach well or better again. The details are in Appendix C.4.

**ATS surpasses previous methods with advanced techniques.** We compare with SOTA methods and list the results on CIFAR-100, TinyImageNet, CUB, and Dogs in Tab. 2 and Tab. 3. The sota compared methods include ESKD [6], TAKD [31], SCKD [51], and ResKD [10, 24]. NoKD trains students without the teacher’s supervision. ST-KD trains students under the guidance of a smaller teacher. KD trains students under the guidance of the larger teacher. More details of compared methods are in Appendix C.4. The last column of these tables shows the average performance of corresponding rows. The larger teacher could slightly improve the students’ performances via traditional KD, i.e., 68.24% \(\rightarrow\) 69.21% and 57.94 \(\rightarrow\) 58.89%. However, using a smaller teacher (the row of “ST-KD”) could obtain about 2% improvement on average, i.e., 68.24% \(\rightarrow\) 70.32% and 57.94% \(\rightarrow\) 59.77%. This again verifies that larger teachers really teach worse on various datasets. Taking advantage of the early-stopped teacher (ESKD), the teacher assistant (TAKD), and the student customized teacher (SCKD) only improves the larger teacher slightly, e.g., 69.21% \(\rightarrow\) 69.86% and 58.89% \(\rightarrow\) 59.04%. These results do not even surpass the small teacher. ResKD improves the students’ performances a lot via introducing the residual student and taking the two students’ ensemble, which surpasses the ensemble performances of two separately trained students under different initializations (the “Ens” row). For a fair comparison, we also test the performances of our methods via repeating the “KD+ATS” two times and making predictions via the ensemble. The results are almost 1% higher than ResKD. Results on CIFAR-10 and speech data are in Appendix D.

**Ablation Studies** One might argue that the validity of ATS is due to the tuning from a larger hyper-parameter space. We vary \(\tau_1\) from 1.0 to 6.0, \(\tau_2\) from 1.0 to 5.0, and record the performances in Fig. 10. Obviously, setting \(\tau_1 > \tau_2\) could be better, and especially, we recommend the setting of \(\tau_2 \in [\tau_1 - 2, \tau_1 - 1]\). Although we introduce one more hyper-parameter, ATS is simple to implement and surpasses SOTA methods that take advanced techniques. We achieve the goal of only utilizing simple operations to make larger teachers teach well again. Then, we study the trade off of the KD.
loss and CE loss under different $\lambda$. The results of “WRN28-8 → MV2” on CIFAR-100 (Fig. 11) verifies that ATS could improve the performances under various $\lambda$. We also compare ATS with other types of KD under various scenes, and the results are in Appendix D.3.

6 Conclusion

We study the miraculous phenomenon in KD that a more accurate model doesn’t necessarily teach better. The proposed KD decomposition attributes the success of a better teacher to three factors, including correct guidance, smooth regularization, and class discriminability. Through theoretical analysis, over-confident teachers could not release their potential abilities of the class discriminability under traditional temperature scaling. As a simple yet effective solution, we propose Asymmetric Temperature Scaling (ATS) to enhance the derived variance of larger teachers, making their distillation labels more discriminative when teaching students. Extensive experimental results verify the superiorities of our proposed methods.

7 Broader Impact

We focus on the variance of wrong class probabilities to analyze why larger teachers cannot teach well and hope that our research could bring a new perspective to the KD field. Our work has no potential negative societal impacts.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (Grant No. 61921006, 41901270), and Natural Science Foundation of Jiangsu Province (Grant No. BK20190296). Thanks to Huawei Noah’s Ark Lab NetMIND Research Team. We gratefully acknowledge the support of Mindspore used for this research. Professor De-Chuan Zhan is the corresponding author.

Figure 10: Ablation studies on asymmetric temperatures on CIFAR-100 and TinyImageNet ($\tau_1, \tau_2$ in Eq. [5]).

Figure 11: Ablation studies on the weighting of KD loss and CE loss on CIFAR-100 ($\lambda$ in Eq. [1]).
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Introducing one more hyperparameter.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] No potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
A More Discussions About KD Decomposition (Eq. 2)

We decompose KD into three terms as shown in Eq. 2, i.e., the correct guidance, the smooth regularization, and the class discriminability. We take a reverse thinking that how can we construct these three terms if without KD? First, to ensure the learning process of the student is correct, the target class’s label should be the highest. This is simple because we have one-hot labels of training samples. That is, the cross-entropy loss could guarantee the student learns correctly. Second, how can we introduce smooth regularization? This is also simple because we could utilize label smoothing (LS) as regularizations. However, using cross-entropy and LS is still not as effective as KD. Hence, the class discriminability also matters a lot. This term is not easy to construct because we do not have the prior class similarities or the instance-level class similarities. Fortunately, a teacher may contain the information of class similarities in KD. We could find that the class discriminability may matter more through the above analysis. Hence, we focus on its efficacy when analyzing why larger models can’t teach well.

Furthermore, the correct guidance is closely related to the second term, i.e., smooth regularization, as \( e(q) = \frac{1}{e^q - 1} (1 - p_q) \). That is, the strength of correct guidance is negatively correlated with the strength of smooth regularization. As shown in Fig. 7 after applying ATS, the smooth regularization curves in these three subfigures are nearly the same, which indicates that the correct guidance term is slightly affected when utilizing ATS.

B Detailed Proofs of Lemma 4.1, Proposition 4.3 and Proposition 4.4

We first present some basic results about softmax operations. The softmax operation is \( p_c = \frac{\exp(f_c / \tau)}{Z} \), where \( Z = \sum_{j=1}^{C} \exp(f_j / \tau) \). We then present some basic derivatives:

\[
\frac{\partial Z}{\partial \tau} = - \sum_{c=1}^{C} f_c \exp(f_c / \tau), \quad \frac{\partial p_c}{\partial \tau} = \frac{p_c}{\tau^2} \left( \sum_{j=1}^{C} p_j f_j - f_c \right)
\]  
(6)

Lemma B.1 (Variance of Softened Probabilities (Lemma 4.1 in the Body)). Given a logit vector \( f \in \mathbb{R}^C \) and the softened probability vector via softmax function, i.e., \( p = SF(f; \tau), \tau \in (0, \infty) \), \( \nu(p) \) monotonically decreases as \( \tau \) increases.

Proof. Obviously, \( e(p) = \frac{1}{C} \). Then, we have:

\[
\nu(p) = \frac{1}{C} \sum_{c=1}^{C} (p_c - e(p))^2 = \frac{1}{C} \sum_{c=1}^{C} p_c^2 - \frac{1}{C^2}.
\]
(7)

We take the derivative of \( \nu(p) \) w.r.t. \( \tau \), and obtain:

\[
\frac{\partial \nu(p)}{\partial \tau} = \frac{2}{C} \sum_{c=1}^{C} \frac{\partial p_c}{\partial \tau} \sum_{c=1}^{C} p_c \left( \sum_{j=1}^{C} p_j f_j - f_c \right)
= \frac{2}{C \tau^2} \left( \sum_{c=1}^{C} p_c^2 \left( \sum_{j=1}^{C} p_j f_j \right) - \sum_{c=1}^{C} p_c \sum_{j=1}^{C} p_j^2 f_j \right).
\]
(8)

Substituting \( f_c = \tau \log p_c + \tau \log Z \), we get:

\[
\frac{\partial \nu(p)}{\partial \tau} = \frac{2}{C \tau} \left( \sum_{c=1}^{C} p_c^2 \left( \sum_{j=1}^{C} p_j \log p_j \right) - \sum_{c=1}^{C} p_c^2 \log p_c \right).
\]
(9)

Then, we define \( \hat{p}_c = p_c^2 / \sum_{c=1}^{C} p_c^2 \), and we could derive the following equation:

\[
\frac{\partial \nu(p)}{\partial \tau} = \frac{2}{C \tau} \left( \sum_{c=1}^{C} \hat{p}_c \log p_c - \sum_{c=1}^{C} \hat{p}_c \log \hat{p}_c \right).
\]
(10)

We then calculate the KL divergence of \( \hat{p} \) and \( p \):

\[
KL(\hat{p}||p) = \sum_{c=1}^{C} \hat{p}_c \log \frac{\hat{p}_c}{p_c} = \frac{C}{C} \hat{p}_c \log \frac{\hat{p}_c}{\sum_{j=1}^{C} p_j^2} = \sum_{c=1}^{C} \hat{p}_c \log p_c - \sum_{c=1}^{C} \hat{p}_c \log p_c - \sum_{c=1}^{C} \hat{p}_c \log \left( \sum_{j=1}^{C} p_j^2 \right).
\]
(11)
Due to the non-negativity of KL divergence, we have:

\[
\sum_{c=1}^{C} \tilde{p}_c \log \frac{p_c}{\tilde{p}_c} \geq \sum_{c=1}^{C} p_c \log \left( \sum_{j=1}^{C} p_j^2 \right) - \log \left( \sum_{c=1}^{C} p_c^2 \right). \tag{12}
\]

Hence, we have:

\[
\sum_{c=1}^{C} p_c \log \frac{p_c}{\tilde{p}_c} - \sum_{c=1}^{C} \tilde{p}_c \log \frac{p_c}{\tilde{p}_c} \leq \sum_{c=1}^{C} p_c \log p_c - \log \sum_{c=1}^{C} p_c^2 \leq 0, \tag{13}
\]

where the first inequality is according to Eq. \((12)\) and the second inequality is according to Jensen’s inequality. This proves that \(\frac{\partial v}{\partial \tau} \leq 0\).

**Proposition B.2** (Proposition 4.3 in the Body). Under the Assumption 4.2, \(p_y\) monotonically decreases as \(\tau\) increases, and \(c(q)\) monotonically increases as \(\tau\) increases. As \(\tau \to \infty\), \(c(q) \to 1/C\).

**Proof.** We take the derivative of \(p_y\) w.r.t. \(\tau\), and obtain:

\[
\frac{\partial p_y}{\partial \tau} = \frac{p_y}{\tau^2} \left( \sum_{c=1}^{C} p_c f_c - f_y \right) \leq 0, \tag{14}
\]

which shows the monotonicity of \(p_y\) w.r.t. \(\tau\). Then the later is obvious because \(c(q) = \frac{1}{C-1} \sum_{c \neq y} p_c = \frac{1}{C-1} (1 - p_y)\). Finally, according to the properties of limitation, \(\lim_{\tau \to \infty} p_y = \lim_{\tau \to \infty} \frac{\sum_{c=1}^{C} \exp(f_c/\tau)}{\sum_{j=1}^{C} \exp(f_j/\tau)} = 1/C\).

**Proposition B.3** (Derived Variance vs. Inherent Variance (Proposition 4.4 in the Body)). The derived variance is determined by the square of derived average and the inherent variance via the following equation:

\[
\frac{v(q)}{DV} = (C-1)^2 \frac{c^2(q)}{DA^2} \frac{v(\tilde{q})}{IV}. \tag{15}
\]

**Proof.** With the property in Eq. \((3)\) we have: \(\tilde{q}_{c'} = \frac{\exp(g_{c'}/\tau)}{\sum_{j} \exp(g_{j}/\tau)} = \frac{p_c}{\sum_{j} p_j} = \frac{p_c}{1-p_y} \). \(c'\) is the corresponding index of \(p_c\) in \(q\) after removing the correct class probability \(p_y\), i.e., \(c' = c\) when \(c < y\), and \(c' = c - 1\) when \(c > y\).

\[
v(q) = \frac{1}{C-1} \sum_{c \neq y} p_c^2 - e^2(q) = \frac{1}{C-1} \sum_{c \neq y} p_c^2 - \left( \frac{1-p_y}{C-1} \right)^2
\]

\[
= \left( 1 - p_y \right)^2 \left( \frac{1}{C-1} \sum_{c \neq y} \left( \frac{p_c}{1-p_y} \right)^2 - \frac{1}{(C-1)^2} \right)
\]

\[
= \left( 1 - p_y \right)^2 \left( \frac{1}{C-1} \sum_{c' \neq y} q_{c'}^2 - \frac{1}{(C-1)^2} \right)
\]

\[
= \left( 1 - p_y \right)^2 v(\tilde{q}) = (C-1)^2 e^2(q)v(\tilde{q}). \tag{16}
\]

**C** Details of Datasets, Networks and Training

**C.1** Dataset Details

The datasets used in our experiments are CIFAR-10/CIFAR-100 (C10/C100) \([21]\), TinyImageNet (TIN) \([43]\), CUB \([47]\), Stanford Dogs (Dogs) \([19]\), and Google Speech Commands (GSC) \([48]\). CIFAR-10/CIFAR-100 are image classification datasets, and each contains 50K training and 10K test samples of size 32 \times 32. TinyImageNet contains 200 classes, 100K training samples, and 10K
Table 4: Statistics of datasets. **Task:** Image Classification (IC), Fine-Grained Recognition (FGR), Keyword Spotting (KWS). **Size:** the input size of a single sample, we omit the dimension of channel. **Networks:** ResNet (RES), WideResNet (WRN), ResNeXt (RNX), ShuffleNetV1/2 (SFV1/2), MobileNetV2 (MV2), AlexNet (ANet).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Size</th>
<th>C</th>
<th>Num.Tr</th>
<th>Num.Te</th>
<th>Teacher Networks</th>
<th>Student Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>C10</td>
<td>IC</td>
<td>(32, 32)</td>
<td>10</td>
<td>50K</td>
<td>10K</td>
<td>RES, WRN, RNX</td>
<td>VGG8, SFV1, MV2</td>
</tr>
<tr>
<td>C100</td>
<td>IC</td>
<td>(32, 32)</td>
<td>100</td>
<td>50K</td>
<td>10K</td>
<td>RES, WRN, RNX</td>
<td>VGG8, SFV1, MV2</td>
</tr>
<tr>
<td>TIN</td>
<td>IC</td>
<td>(64, 64)</td>
<td>200</td>
<td>100K</td>
<td>10K</td>
<td>WRN50-2</td>
<td>Anet, SFV2, MV2</td>
</tr>
<tr>
<td>CUB</td>
<td>FGR</td>
<td>(224, 224)</td>
<td>200</td>
<td>6K</td>
<td>5.8K</td>
<td>RNX101-32-8d</td>
<td>Anet, SFV2, MV2</td>
</tr>
<tr>
<td>Dogs</td>
<td>FGR</td>
<td>(224, 224)</td>
<td>120</td>
<td>12K</td>
<td>8.6K</td>
<td>RNX101-32-8d</td>
<td>Anet, SFV2, MV2</td>
</tr>
<tr>
<td>GSC</td>
<td>KWS</td>
<td>(101, 40)</td>
<td>35</td>
<td>106K</td>
<td>11K</td>
<td>RES</td>
<td>DSCNN</td>
</tr>
</tbody>
</table>

Table 5: Details of networks and their performances on training and test set (Accs). “RNX” is abbreviated as “R” to save space. Student networks are shaded while teachers are not.

<table>
<thead>
<tr>
<th>Network</th>
<th>N.P</th>
<th>Accs</th>
<th>Network</th>
<th>N.P</th>
<th>Accs</th>
<th>Network</th>
<th>N.P</th>
<th>Accs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10 Teachers</td>
<td></td>
<td></td>
<td>CIFAR-10 Students</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet</td>
<td></td>
<td></td>
<td>ResNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RES14</td>
<td>0.18M</td>
<td>98.4, 91.5</td>
<td>WRN28-1</td>
<td>0.37M</td>
<td>99.7, 92.8</td>
<td>VGG8</td>
<td>3.9M</td>
<td>99.9, 91.7</td>
</tr>
<tr>
<td>RES20</td>
<td>0.27M</td>
<td>99.5, 92.3</td>
<td>WRN28-2</td>
<td>1.5M</td>
<td>100, 94.9</td>
<td>SFV1</td>
<td>0.86M</td>
<td>99.9, 92.1</td>
</tr>
<tr>
<td>RES32</td>
<td>0.47M</td>
<td>99.9, 93.5</td>
<td>WRN28-3</td>
<td>3.3M</td>
<td>100, 95.3</td>
<td>VGG8</td>
<td>3.9M</td>
<td>99.9, 91.7</td>
</tr>
<tr>
<td>RES44</td>
<td>0.66M</td>
<td>99.9, 93.8</td>
<td>WRN28-4</td>
<td>5.8M</td>
<td>100, 95.6</td>
<td>SFV1</td>
<td>0.86M</td>
<td>99.9, 92.1</td>
</tr>
<tr>
<td>RES56</td>
<td>0.86M</td>
<td>100, 93.9</td>
<td>WRN28-6</td>
<td>13M</td>
<td>100, 95.9</td>
<td>SFV1</td>
<td>0.86M</td>
<td>99.9, 92.1</td>
</tr>
<tr>
<td>RES110</td>
<td>1.7M</td>
<td>100, 94.3</td>
<td>WRN28-8</td>
<td>23M</td>
<td>100, 96.1</td>
<td>VGG8</td>
<td>3.9M</td>
<td>99.9, 91.7</td>
</tr>
<tr>
<td>CIFAR-10 Students</td>
<td></td>
<td></td>
<td>CIFAR-100 Teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet</td>
<td></td>
<td></td>
<td>ResNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RES14</td>
<td>0.18M</td>
<td>81.1, 67.0</td>
<td>WRN28-1</td>
<td>0.38M</td>
<td>91.8, 70.0</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES20</td>
<td>0.28M</td>
<td>88.1, 69.1</td>
<td>WRN28-2</td>
<td>1.5M</td>
<td>99.8, 74.2</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES32</td>
<td>0.47M</td>
<td>94.3, 71.1</td>
<td>WRN28-3</td>
<td>3.3M</td>
<td>100, 76.7</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES44</td>
<td>0.67M</td>
<td>97.3, 72.2</td>
<td>WRN28-4</td>
<td>5.9M</td>
<td>100, 77.9</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES56</td>
<td>0.86M</td>
<td>98.7, 72.9</td>
<td>WRN28-6</td>
<td>13M</td>
<td>100, 79.1</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES110</td>
<td>1.7M</td>
<td>99.8, 74.1</td>
<td>WRN28-8</td>
<td>23M</td>
<td>100, 79.7</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>CIFAR-100 Students</td>
<td></td>
<td></td>
<td>CIFAR-100 Students</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet</td>
<td></td>
<td></td>
<td>ResNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RES14</td>
<td>0.18M</td>
<td>81.1, 67.0</td>
<td>WRN28-1</td>
<td>0.38M</td>
<td>91.8, 70.0</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES20</td>
<td>0.28M</td>
<td>88.1, 69.1</td>
<td>WRN28-2</td>
<td>1.5M</td>
<td>99.8, 74.2</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES32</td>
<td>0.47M</td>
<td>94.3, 71.1</td>
<td>WRN28-3</td>
<td>3.3M</td>
<td>100, 76.7</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES44</td>
<td>0.67M</td>
<td>97.3, 72.2</td>
<td>WRN28-4</td>
<td>5.9M</td>
<td>100, 77.9</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES56</td>
<td>0.86M</td>
<td>98.7, 72.9</td>
<td>WRN28-6</td>
<td>13M</td>
<td>100, 79.1</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
<tr>
<td>RES110</td>
<td>1.7M</td>
<td>99.8, 74.1</td>
<td>WRN28-8</td>
<td>23M</td>
<td>100, 79.7</td>
<td>R29-4-4d</td>
<td>1.2M</td>
<td>100, 93.9</td>
</tr>
</tbody>
</table>

| TinyImageNet Teachers |     |                   | TinyImageNet Students |     |                   |
| Anet    | 2.7M  | 98.5, 89.0       | SFV2    | 1.5M  | 94.3, 45.8       | MV2     | 2.5M  | 85.9, 52.0       |
| CIFU Students |     |                   | CIFU Students |     |                   |
| Anet    | 2.7M  | 92.8, 55.7       | SFV2    | 1.5M  | 93.7, 71.2       | MV2     | 2.5M  | 95.8, 74.5       |
| Stanford Dogs Teachers |     |                   | Stanford Dogs Students |     |                   |
| Anet    | 2.6M  | 88.0, 50.2       | SFV2    | 1.4M  | 92.6, 68.7       | MV2     | 2.4M  | 94.5, 68.7       |
| Google Speech Commands Teachers |     |                   | Google Speech Commands Students |     |                   |
| Res     | 1.9M  | 98.8, 99.2       | -       | -     | -                 | -       | -     | -               |
| DSCNN   | 59.8K | 93.2, 94.5       | -       | -     | -                 | -       | -     | -               |
test samples of size $64 \times 64$. CUB and Stanford Dogs are fine-grained recognition datasets, which correspondingly contain 200 and 120 classes. Google Speech Commands is a benchmark for keyword spotting \cite{41,60}, which usually needs efficient model deployment. It aims to identify whether a 1s-long speech recording is a word, silence, or unknown. There are total 35 classes (35 words) in this task. We extract 40 MFCC features for each 30ms window frame with a stride of 10ms. We also follow the settings in Google Speech Commands (GSC) \cite{48}: performing random time-shift of $Y \in [-100, 100]$ milliseconds and adding 0.1 volume background noise with a probability of 0.8. The details are listed in Tab. 4. $C$ denotes the number of classes. “Num.Tr” and “Num.Te” denote the number of training and test samples.

C.2 Network Details

For teacher networks, we use different versions of ResNet \cite{11}, WideResNet \cite{56}, ResNeXt \cite{49}. For student networks, we use VGG \cite{39}, ShuffleNetV1/V2 \cite{29,59}, AlexNet \cite{22}, MobileNetV2 \cite{37}, and DSCNN \cite{60}.

- **ResNet (RES):** We utilize ResNet for CIFAR-10/100 and Google Speech Commands as teachers. For CIFAR-10/100, we use the original ResNet versions with “6n+2” layers in \cite{11}. These ResNets only use the “basic block” for CIFAR data. Specifically, we vary the number of layers in $\{14, 20, 32, 44, 56, 110\}$. For Google Speech Commands, we utilize the ResNet version proposed in \cite{41}. We modify the number of layer as 15 and the channel as 128.

- **WideResNet (WRN):** We utilize WideResNet for CIFAR-10/100 and TinyImageNet as teachers. For CIFAR-10/100, we use the WideResNet proposed in \cite{56}. We keep the depth as 28 and vary the widen factors in $\{1, 2, 3, 4, 6, 8\}$. For TinyImageNet, we use the WideResNet50-2 provided in PyTorch.

- **ResNeXt (RNX):** We utilize ResNeXt for CIFAR10-/100, CUB and Stanford Dogs as teachers. For CIFAR10-/100, we utilize the original ResNeXt versions for CIFAR data as proposed in \cite{49}. We take 29 layers and set the base width as 4, varying the number of groups in $\{4, 8, 16, 24, 32, 64\}$. For CUB and Stanford Dogs, we use the ResNeXt101-32-8d provided in PyTorch.

- **VGG8:** We utilize VGG8 \cite{39} for CIFAR-10/100 as the student. Although it contains more parameters, the plain architecture makes it weaker.

- **ShuffleNetV1/2 (SFV1/2):** We utilize ShuffleNet for CIFAR-10/100, TinyImageNet, CUB, and Dogs as students. For CIFAR-10/100, we use ShuffleNetV1 \cite{59} provided in RepDistiller repository. For other datasets, we use ShuffleNetV2 \cite{29} provided in PyTorch.

- **MobileNetV2 (MV2):** We utilize MobileNet for CIFAR-10/100, TinyImageNet, CUB, and Dogs as students. For CIFAR-10/100, we use MobileNetV2 \cite{37} in RepDistiller repository. For other datasets, we use ShuffleNetV2 \cite{37} provided in PyTorch.

- **AlexNet (ANet):** We utilize AlexNet provided in PyTorch for TinyImageNet, CUB, and Dogs as students.

- **DSCNN:** We use DSCNN \cite{60} for Google Speech Commands as the student. It utilizes depth-separable convolution to accelerate training and inference. We set the basic number of channels as 96.

Additionally, for CUB and Stanford Dogs, we use the corresponding pre-trained models provided in PyTorch as initialization. The number of parameters (“N.P”) and the training and test accuracies are in Tab. 5.

C.3 Training Details

**Train Teachers** We first train teacher networks on corresponding datasets and then save the checkpoints of them.

**Train Students** We train students via the loss function in Eq. 1. We set $\lambda = 0.5$ by default and tune $\tau \in \{1.0, 2.0, 4.0, 8.0, 12.0, 16.0\}$. Additionally, we vary the student’s $\tau$ in two settings: (1) the same...
Table 6: Comparisons with SOTA methods on CIFAR-10. ResNet110, WRN28-8, and RNX29-64-4d are used as teachers. VGG8, SFV1, and MV2 are students.

<table>
<thead>
<tr>
<th>Teacher</th>
<th>ResNet110 (94.3)</th>
<th>WRN28-8 (96.1)</th>
<th>RNX29-64-4d (95.8)</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoKD</td>
<td>91.68</td>
<td>92.11</td>
<td>88.36</td>
<td>90.72</td>
</tr>
<tr>
<td>ST-KD</td>
<td>92.47</td>
<td>93.20</td>
<td>88.83</td>
<td>92.56</td>
</tr>
<tr>
<td>KD</td>
<td>92.29</td>
<td>92.80</td>
<td>88.60</td>
<td>92.44</td>
</tr>
<tr>
<td>ESKD</td>
<td>92.07</td>
<td>92.84</td>
<td>88.18</td>
<td>92.32</td>
</tr>
<tr>
<td>TAKD</td>
<td>92.10</td>
<td>93.13</td>
<td>88.36</td>
<td>92.31</td>
</tr>
<tr>
<td>SCKD</td>
<td>91.75</td>
<td>92.79</td>
<td>88.70</td>
<td>91.91</td>
</tr>
<tr>
<td>KD+ATS</td>
<td>92.88</td>
<td>92.99</td>
<td>88.63</td>
<td>92.73</td>
</tr>
</tbody>
</table>

as the teacher’s; (2) 1.0. For our proposed KD method, the pair of \((\tau_1, \tau_2)\) in Eq. 5 is searched in \{(2.0, 1.0), (3.0, 1.0), (3.0, 2.0), (4.0, 2.0), (4.0, 3.0), (5.0, 2.0)\}. For our method, we keep the student’s temperature as 1.0 by default.

**Compared Methods** We then explain the compared methods in Tab. 2 and Tab. 3. For CIFAR-100/CIFAR-10, we use ResNet110, WRN28-8, and RNX29-64-4d as teachers. For TinyImageNet, CUB, and Stanford Dogs, we use WRN50-2 and ResNeXt101-32-8d as teachers. These models are complex networks with a larger depth, or a larger widen factor, or more groups. We have also explored the standard deviation of these KD learning methods. KD performances are relatively stable among several training random seeds with the same hyper-parameter settings. The standard deviations are also nearly the same across different KD methods, which are about 0.1%. Hence, we do not report the standard deviations due to space limitation.

- **NoKD**: training the student without distillation.
- **ST-KD**: training the student under the guidance of a smaller teacher. For CIFAR-100, we select the best performance from all smaller networks as shown in Fig. 9. For TinyImageNet, CUB, and Stanford Dogs, we correspondingly use WRN26-2, RNX50-32-4d, RNX50-32-4d as smaller teachers.
- **KD**: training the student under the guidance of the larger teacher.
- **ESKD** [6]: training the student under the guidance of the *early-stopped teacher*. Specifically, we train the larger teachers only for \{60, 120, 150\} epochs using the cosine annealing learning rate and report the best accuracy of the student.
- **TAKD** [31]: training the student under the guidance of the *teacher assistant*. Specifically, we vary the TA in \{ResNet20, ResNet44\} for ResNet110, \{WRN28-2, WRN28-4\} for WRN28-8, \{RNX29-8-4d, RNX29-8-24d\} for RNX29-64-4d, \{WRN26-2\} for WRN50-2, and \{RNX50-32-4d\} for RNX101-32-8d.
- **SCKD** [51]: training the student under the *student customized teacher*. We use two KD losses, including the KL divergence of probabilities (Eq. 1) and the L2 feature distance loss as done in [36].
- **Ens**: training two students with different initializations two times and taking their ensemble.
- **ResKD**: training the student first and then training the *residual student* to fit the residual. The ensemble of these two students is used for inference.
- **KD+ATS**: training the student with the larger teacher’s guidance via the proposed ATS.
- **KD+ATS+Ens**: repeating the above two times and taking the ensemble results. When repeating “KD+ATS” two times, we use the same hyperparameters and keep the fairness of comparisons.

**C.4 Figure Details**

We explain some figures in detail.
Table 7: Comparisons with other KD methods on CIFAR-100. The results of first five columns are cited from [44]. The gray area shows the results of scenes with smaller “capacity gap”. ReKD denotes ReviewKD [4].

<table>
<thead>
<tr>
<th>Scene</th>
<th>Teacher</th>
<th>Student</th>
<th>KD</th>
<th>CC</th>
<th>NST</th>
<th>CRD</th>
<th>IE-AT</th>
<th>ReKD</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet110 → ResNet20</td>
<td>74.31</td>
<td>69.06</td>
<td>70.67</td>
<td>69.48</td>
<td>69.53</td>
<td>71.14</td>
<td>71.34</td>
<td>71.37</td>
<td>71.57</td>
</tr>
<tr>
<td>VGG13 → VGG8</td>
<td>74.64</td>
<td>70.36</td>
<td>72.98</td>
<td>70.71</td>
<td>71.53</td>
<td>73.94</td>
<td>74.05</td>
<td>73.60</td>
<td>73.65</td>
</tr>
<tr>
<td>ResNet56 → ResNet20</td>
<td>72.34</td>
<td>69.06</td>
<td>70.66</td>
<td>69.63</td>
<td>69.60</td>
<td>71.16</td>
<td>70.87</td>
<td>70.95</td>
<td>70.99</td>
</tr>
<tr>
<td>WRN40-2 → WRN16-2</td>
<td>75.61</td>
<td>73.26</td>
<td>74.92</td>
<td>73.56</td>
<td>73.68</td>
<td>75.48</td>
<td>74.80</td>
<td>75.67</td>
<td>75.03</td>
</tr>
</tbody>
</table>

Figure 12: Probability vector visualization of a randomly selected training sample in CIFAR-10. The networks are various ResNet. The target class is \( y = 1 \). The bottom row shows the efficacy of ATS for a larger teacher.

- Fig. 2 for “ResNet→SFV1”, we use ResNet14 as the small teacher (“ST”) and ResNet110 as the large teacher (“LT”); for “WRN→MV2”, we use WRN28-1 as the small teacher (“ST”) and WRN28-8 as the large teacher (“LT”).

- Fig. 3 on CIFAR-100, we vary 18 teachers and 3 students as listed in Tab. 5; we also vary temperatures in \( \{1.0, 2.0, 4.0, 8.0, 16.0\} \); the total pairs of distillation combinations are \( 18 \times 3 \times 5 = 270 \); for each pair, we calculate the teacher’s derived average and derived variance averaged across 50K training samples; the KD improvement ratio is calculated by \((\text{AccKD} - \text{AccNOKD})/\text{AccNOKD})\).

- Fig. 4 and Fig. 5, the former calculates \( f_y \) (i.e., target logit) and \( \sigma(y) \) (i.e., standard deviation of wrong logits) across training samples and plot the bins; the latter calculates \( \sigma(q) \) (i.e., derived variance) and \( \sigma(\tilde{q}) \) (i.e., inherent variance) and only report their mean and standard deviation across training samples.

- Fig. 6, for each pair of two teachers \( T_1 \) and \( T_2 \), we obtain their softened labels on all training samples; for each sample, we calculate the four metrics and report the average across training samples.

- Fig. 7, given a temperature \( \tau \), we could obtain softened probabilities of all training samples; then we calculate \( \sigma(q) \) (i.e., derived variance) and \( \sigma(\tilde{q}) \) (i.e., inherent variance) across all training samples and only report the average results.

- Fig. 8, For TS, we tune \( \tau \in \{1.0, 2.0, 4.0, 8.0, 12.0, 16.0\} \) and vary the student’s \( \tau \) in two settings: (1) the same as the teacher’s, (2) 1.0; For ATS, the pair of \( (\tau_1, \tau_2) \) in Eq. 5 is searched in \( \{(2.0, 1.0), (3.0, 1.0), (3.0, 2.0), (4.0, 2.0), (4.0, 3.0), (5.0, 2.0)\} \); through adjusting the hyper-parameters and selecting the best results, we plot the performance curves.

20
Figure 13: The change of derived average \( e(q) \) and derived variance \( v(q) \) as \( \tau \) increases from 0.1 to 10.0 on CIFAR-100 using various WRN.

Table 8: Comparisons on Google Speech Commands under various types of KD. \(+/−\) denotes multiplying \( \tau^2 \) or not in Eq. 1.

<table>
<thead>
<tr>
<th>( (\lambda, +/-) )</th>
<th>(0.5, +)</th>
<th>(0.9, +)</th>
<th>(0.5, -)</th>
<th>(0.9, -)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KD+TS</td>
<td>95.52</td>
<td>95.37</td>
<td>94.55</td>
<td>94.37</td>
</tr>
<tr>
<td>KD+ATS</td>
<td>96.79</td>
<td>97.21</td>
<td>96.23</td>
<td>95.84</td>
</tr>
</tbody>
</table>

D More Experimental Studies

D.1 CIFAR-10 Results

We also compare our proposed methods with SOTA on CIFAR-10, and the results are listed in Tab. 6. Because CIFAR-10 only contains 10 classes and is an slightly simpler benchmark, the performance improvement is not so obvious as other datasets.

D.2 Speech Data Results

ATS could also perform well on speech data, i.e., the Google Speech Commands benchmark as shown in Tab. 8 where we vary several types of KD loss. Our proposed ATS does not depend on multiplying \( \tau^2 \) or not in Eq. 1.

D.3 Comparisons With Other KD Methods on CIFAR-100

We also compare our proposed method with other KD methods, such as 44 [17] 34 [8] 16. The previous work 44 provides a comprehensive experimental study on many KD methods. For fair comparison, we directly cite their results and only report several scenes of “Larger Teacher → Smaller Student”, i.e., “ResNet110 (74.31%) → ResNet20 (69.06%)” and “VGG13 (74.64%) → VGG8 (70.36%)”. For IE-KD 16, we use the proposed variant of IE-AT. For IE-KD 16 and ReviewKD 4, we utilize the public code that the authors provide. We also compare with these two methods because they are the recently proposed SOTA KD methods. The results are listed in Tab. 7. Consistently, our method could improve the KD up to 1% and the results are comparable with CRD that utilizes more advanced techniques. We also apply ATS to the cases that students have almost the same capacity as teachers, e.g., “WRN40-2 (75.61%) → WRN16-2 (73.26%)” and “ResNet56 (72.34%) → ResNet20 (69.06%)”. Under these scenes, our method becomes not so effective and the performance improvement over KD seems more like the benefits of hyper-parameter tuning. Hence, our proposed ATS may be more effective when applying a more complex teacher to guide the learning process of a smaller student.

D.4 More Results for the Observations

We also present additional results for the observations in Fig. 8 and Fig. 7 to verify that the observations and conclusions are not accidental. Similar results are shown in Fig. 12 and Fig. 13.