Logit Standardization in Knowledge Distillation

Shangquan Sun[1,2], Wenqi Ren[3†], Jingzhi Li[1], Rui Wang[1,2], Xiaochun Cao[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
{sunshangquan,lijingzhi,wangrui}@iie.ac.cn, {renwq3,caoxiaochun}@mail.sysu.edu.cn

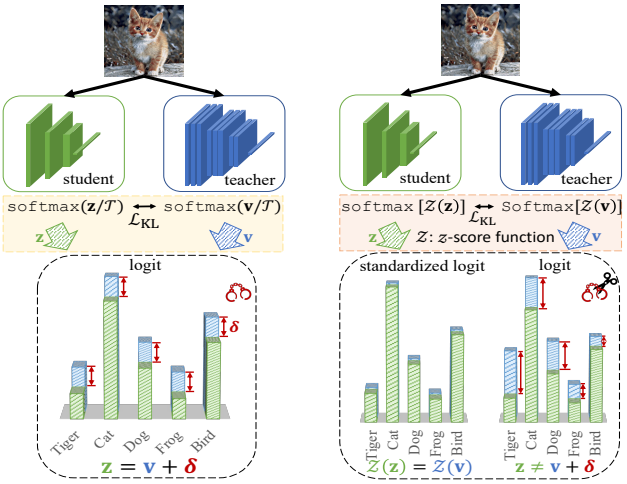
2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.01489

Abstract

Knowledge distillation involves transferring soft labels from a teacher to a student using a shared temperature-based `softmax` *function. However, the assumption of a shared temperature between teacher and student implies a mandatory exact match between their logits in terms of logit range and variance. This side-effect limits the performance of student, considering the capacity discrepancy between them and the finding that the innate logit relations of teacher are sufficient for student to learn. To address this issue, we propose setting the temperature as the weighted standard deviation of logit and performing a plug-and-play $\mathcal{Z}$-score pre-process of logit standardization before applying* `softmax` *and Kullback-Leibler divergence. Our pre-process enables student to focus on essential logit relations from teacher rather than requiring a magnitude match, and can improve the performance of existing logit-based distillation methods. We also show a typical case where the conventional setting of sharing temperature between teacher and student cannot reliably yield the authentic distillation evaluation; nonetheless, this challenge is successfully alleviated by our $\mathcal{Z}$-score. We extensively evaluate our method for various student and teacher models on CIFAR-100 and ImageNet, showing its significant superiority. The vanilla knowledge distillation powered by our pre-process can achieve favorable performance against state-of-the-art methods, and other distillation variants can obtain considerable gain with the assistance of our pre-process. The codes, pre-trained models and logs are released on Github.*

1. Introduction

The development of deep neural networks (DNN) has revolutionized the field of computer vision in the past decade.



(a) Vanilla KD (b) KD w/ our logit standardization

Figure 1. Vanilla knowledge distillation implicitly enforces an exact match between the magnitudes of teacher and student logits. It is an unnecessary side-effect because it is found sufficient to preserve the innate relations between their logits. Given the capacity gap between them, it is also challenging for a lightweight student to produce logits with the same magnitude as a cumbersome teacher. In contrast, the proposed $\mathcal{Z}$-score logit standardization pre-process mitigates the side-effect. The standardized student logits have arbitrary magnitude suitable for the student's capacity while preserving the essential relations learned from the teacher.

However, with increasing performance and capacity, the model size and computational cost of DNN have also been expanding. Despite of a tendency that larger models have greater capacity, many efforts have been made by researchers to cut down model size without sacrificing much accuracy. In addition to designing lightweight models, knowledge distillation (KD) has emerged as a new approach to achieve this goal. It involves transferring the knowledge of a pre-trained heavy model, known as the teacher network, to a small target model, known as the student network.

Hinton *et al*. [13] firstly proposes distilling a teacher's knowledge into a student by minimizing a Kullback-Leibler (KL) divergence between their predictions. A scaling factor

† Corresponding author.
This work has been supported in part by National Natural Science Foundation of China (No. 62322216, 62025604, 62306308), in part by Shenzhen Science and Technology Program (Grant No. JCYJ20220818102012025, KQTD20221101093559018).

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.01489

of the softmax function in this context, called temperature $\mathcal{T}$, is introduced to soften the predicted probabilities. Traditionally, the temperature is set globally beforehand as a hyper-parameter and remains fixed throughout training. CTKD [24] adopts an adversarial learning module to predict sample-wise temperatures, adapting to varying sample difficulties. However, existing logit-based KD approaches still assume that the teacher and student should share temperatures, neglecting the possibility of distinct temperature values in the KL divergence. In this work, we demonstrate that the general softmax expression in both classification and KD is derived from the principle of entropy maximization in information theory. During this derivation, Lagrangian multipliers appear and take the form of temperatures, based on which we establish the irrelevance between the temperatures of teacher and student, as well as the irrelevance among temperatures for different samples. The proof supports our motivation to allocate distinct temperatures between teacher and student and across samples.

Compared to an exact match of logit prediction, it is found that the inter-class relations of predictions are sufficient for student to achieve performance similar to teacher [15]. A lightweight student faces challenges in predicting logits with a comparable range and variance as a cumbersome teacher, given the capacity gap between them [7, 30, 37]. However, we demonstrate that the conventional practice of sharing temperatures in KL divergence still implicitly enforces an exact match between the student and teacher's logits. Existing logit-based KD methods, unaware of this issue, commonly fall in the pitfall, resulting in a general performance drop. To address this, we propose using weighted logit standard deviation as an adaptive temperature and present a $\mathcal{Z}$-score logit standardization as a pre-processing step before applying softmax . This pre-processing maps arbitrary range of logits into a bounded range, allowing student logits to possess arbitrary ranges and variances while efficiently learning and preserving only the innate relationships of teacher logits. We present a typical case where the KL divergence loss under the setting of sharing temperatures in softmax may be misleading and cannot reliably measure the performance of distilled students. In contrast, with our $\mathcal{Z}$-score pre-process, the issue of the shared temperatures in the case is eliminated.

In summary, our contributions are in three folds

- Based on the entropy maximization principle in information theory, we use Lagrangian multipliers to derive the general expression of softmax in logit-based KD. We show that the temperature comes from the derived multipliers, allowing it to be selected differently for various samples and distinctly for student and teacher.
- To address the issues of the conventional logit-based KD pipeline caused by shared temperatures, including an implicit mandatory logit match and an inauthentic indication of student performance, we propose a pre-process of logit distillation to adaptively allocate temperatures between teacher and student and across samples, capable of facilitating existing logit-based KD approaches.
- We conduct extensive experiments with various teacher and student models on CIFAR-100 [18] and ImageNet [34] and demonstrate the superior advantages of our method as a plug-and-play pre-process.

2. Related Work

Knowledge distillation [13] is designed to transfer the "dark" knowledge from a cumbersome teacher model to a lightweight student model. By learning from the soft labels of teacher, student can achieve better performance than training on hard labels only. The traditional method trains a student by minimizing a difference such as KL divergence between its predicted probability and the teacher's. The prediction of probability is commonly approximated by the softmax of logit output. KD algorithms can be classified into three types, i.e., logit-based [3, 13, 17, 22, 30, 54, 56, 57], feature-based [1, 4, 5, 10, 12, 23, 25, 28, 33, 39, 40, 51], and relation-based [15, 19, 29, 31, 32, 43, 50] methods.

A temperature is introduced to flatten the probabilities in logit-based methods. Several works [2, 13, 27] explore its properties and effects. They reach an identical conclusion that temperature controls how much attention student pays on those logits more negative than average. A very low temperature makes student ignore other logits and instead mainly focus on the largest logit of teacher. However, they do not discuss why teacher and student share a globally predefined temperature. It was unknown whether temperature can be determined in an instance-wise level until CTKD [24] proposed predicting sample-wise temperatures by leveraging adversarial learning. However, it assumes that teacher and student should share temperatures. It was still undiscovered whether teacher and student can have divergent temperatures. ATKD [9] proposes a sharpness metric and chooses adaptive temperature by reducing the gap between teacher and student. However, their assumption of a zero logit mean relies on numerical approximation and limits its performance. Additionally, they do not thoroughly discuss where the temperature is derived from and whether distinct temperatures can be assigned. In this work, we provide an analytical derivation based on the entropy-maximization principle, demonstrating that students and teachers do not necessarily share a temperature. It is also found sufficient to preserve the innate relationship of prediction, instead of exact logit values of teacher [15]. However, the existing logit-based KD pipelines still implicitly mandate an exact match between teacher and student logits. We thus define the temperature to be the weighted standard deviation of logit to alleviate the issue and facilitate the existing logit-based KD approaches.

3. Background and Notation

Suppose we have a transfer dataset $\mathcal{D}$ containing totally N samples $\{\mathbf{x}_n, y_n\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^{H\times W}$ and $y_n \in [1, K]$ are the image and label respectively for the n-th sample. The notations of H, W and K are image height, width and the number of classes. Given an input $\{\mathbf{x}_n, y_n\}$, teacher f_T and student f_S respectively predict logit vectors $\mathbf{v}_n$ and $\mathbf{z}_n \in \mathbb{R}^{1\times K}$. Namely, $\mathbf{z}_n = f_S(\mathbf{x}_n)$ and $\mathbf{v}_n = f_T(\mathbf{x}_n)$.

It is widely accepted that a `softmax` function involving a temperature $\mathcal{T}$ is used to convert the logit to probability vectors $q(\mathbf{z}_n)$ or $q(\mathbf{v}_n)$ such that their k-th items have

$$q(\mathbf{z}_n)^{(k)} = \frac{\exp(\mathbf{z}_n^{(k)}/\mathcal{T})}{\sum_{m=1}^K \exp(\mathbf{z}_n^{(m)}/\mathcal{T})}, \tag{1}$$

$$q(\mathbf{v}_n)^{(k)} = \frac{\exp(\mathbf{v}_n^{(k)}/\mathcal{T})}{\sum_{m=1}^K \exp(\mathbf{v}_n^{(m)}/\mathcal{T})}, \tag{2}$$

where $\mathbf{z}_n^{(k)}$ and $\mathbf{v}_n^{(k)}$ are the k-th item of $\mathbf{z}_n$ and $\mathbf{v}_n$ respectively. A knowledge distillation process is essentially letting $q(\mathbf{z}_n)^{(k)}$ mimic $q(\mathbf{v}_n)^{(k)}$ for any class and all samples. The objective is realized by minimizing KL divergence

$$\mathcal{L}_{\mathrm{KL}}\left(q(\mathbf{v}_n)||q(\mathbf{z}_n)\right) = \sum_{k=1}^K q(\mathbf{v}_n)^{(k)} \log\left(\frac{q(\mathbf{v}_n)^{(k)}}{q(\mathbf{z}_n)^{(k)}}\right),$$

which is theoretically equivalent to a cross-entropy loss when optimizing solely on $\mathbf{z}$,

$$\mathcal{L}_{\mathrm{CE}}\left(q(\mathbf{v}_n), q(\mathbf{z}_n)\right) = -\sum_{k=1}^K q(\mathbf{v}_n)^{(k)} \log q(\mathbf{z}_n)^{(k)}. \tag{3}$$

Note that they are empirically nonequivalent as their gradients diverge due to the negative entropy term of $q(\mathbf{v}_n)$.

4. Methodology

It is widely accepted that $\mathcal{T}$ is shared for teacher and student in Eq. 1 and Eq. 2. In contrast, in Sec. 4.1, we show the irrelevance between the temperatures of teacher and student, as well as across different samples. Guaranteed that temperatures can be different between teacher and student and among sample, we further show two side-effect drawbacks of shared-temperatures setting in conventional KD pipelines in Sec. 4.2. In Sec. 4.3, we propose leveraging logit standard deviation as a factor in temperature and derive a preprocess of logit standardization.

4.1. Irrelevance between Temperatures

In Sec. 4.1.1 and 4.1.2, we first give a derivation of the temperature-involved `softmax` function in classification and KD based on the entropy-maximization principle in information theory. This implies the temperatures of student and teacher can be distinct and sample-wisely different.

4.1.1 Derivation of `softmax` in Classification

The `softmax` function in classification can be proved to be the unique solution of maximizing entropy subject to the normalization condition of probability and a constraint on the expectation of states in information theory [16]. The derivation is also leveraged in confidence calibration to formulate temperature scaling [8]. Suppose we have the following constrained entropy-maximization optimization,

$$\max_q \mathcal{L}_1 = -\sum_{n=1}^N \sum_{k=1}^K q(\mathbf{v}_n)^{(k)} \log q(\mathbf{v}_n)^{(k)}$$
$$s.t. \begin{cases} \sum_{k=1}^K q(\mathbf{v}_n)^{(k)} = 1, \quad \forall n \\ \mathbb{E}_q[\mathbf{v}_n] = \sum_{k=1}^K \mathbf{v}_n^{(k)} q(\mathbf{v}_n)^{(k)} = \mathbf{v}_n^{(y_n)}, \quad \forall n. \end{cases} \tag{4}$$

The first constraint holds due to the requirement of discrete probability density, while the second constraint controls the scope of the distribution such that model accurately predicts the target class. Suppose $\hat{q}_n$ to be the one-hot hard probability distribution whose values are all zero except at the target index $\hat{q}_n^{(y_n)} = 1$. The second constraint is then actually $\mathbb{E}_q[\mathbf{v}_n] = \sum_{k=1}^K \mathbf{v}_n^{(k)} \hat{q}_n^{(k)} = \mathbf{v}_n^{(y_n)}$. This is equivalent to making model predict the correct label y_n. By applying Lagrangian multipliers $\{\alpha_{1,i}\}_{i=1}^N$ and $\{\alpha_{2,i}\}_{i=1}^N$, it gives

$$\mathcal{L}_T = \mathcal{L}_1 + \sum_{n=1}^N \alpha_{1,n}\left(\sum_{k=1}^K q(\mathbf{v}_n)^{(k)} - 1\right) + \sum_{n=1}^N \alpha_{2,n}\left(\sum_{k=1}^K \mathbf{v}_n^{(k)} q(\mathbf{v}_n)^{(k)} - \mathbf{v}_n^{(y_n)}\right).$$

Taking the partial derivative with respective to $\alpha_{1,n}$ and $\alpha_{2,n}$ yields back the constraints. In contrast, taking the derivative with respective to $q(\mathbf{v}_n)^{(k)}$ gives

$$\frac{\partial \mathcal{L}_T}{\partial q(\mathbf{v}_n)^{(k)}} = -1 - \log q(\mathbf{v}_n)^{(k)} + \alpha_{1,n} + \alpha_{2,n}\mathbf{v}_n^{(k)}, \tag{5}$$

which leads to a solution by making the derivative zero:

$$q(\mathbf{v}_n)^{(k)} = \exp\left(\alpha_{2,n}\mathbf{v}_n^{(k)}\right)/Z_T, \tag{6}$$

where $Z_T = \exp\left(1-\alpha_{1,n}\right) = \sum_{m=1}^K \exp\left(\alpha_{2,n}\mathbf{v}_n^{(m)}\right)$ is the partition function to fulfill the normalization condition.

4.1.2 Derivation of `softmax` in KD

Following the idea, we define a problem of entropy-maximization to formulate the `softmax` in KD. Given a well-trained teacher and its prediction $q(\mathbf{v}_n)$, we have the objective function for the prediction of student as follows,

$$\max_q \mathcal{L}_2 = -\sum_{n=1}^N \sum_{k=1}^K q(\mathbf{z}_n)^{(k)} \log q(\mathbf{z}_n)^{(k)}$$
$$s.t. \begin{cases} \sum_{k=1}^K q(\mathbf{z}_n)^{(k)} = 1, \quad \forall n \\ \sum_{k=1}^K \mathbf{z}_n^{(k)} q(\mathbf{z}_n)^{(k)} = \mathbf{z}_n^{(y_n)}, \quad \forall n \\ \sum_{k=1}^K \mathbf{z}_n^{(k)} q(\mathbf{z}_n)^{(k)} = \sum_{k=1}^K \mathbf{z}_n^{(k)} q(\mathbf{v}_n)^{(k)}, \quad \forall n. \end{cases} \tag{7}$$

By applying Lagrangian multipliers $\beta_{1,n}$, $\beta_{2,n}$ and $\beta_{3,n}$,

$$\mathcal{L}_S = \mathcal{L}_2 + \sum_{n=1}^{N} \beta_{1,n} \left(\sum_{k=1}^{K} q(\mathbf{z}_n)^{(k)} - 1 \right) + \sum_{n=1}^{N} \beta_{2,n} \left(\sum_{k=1}^{K} \mathbf{z}_n^{(k)} q(\mathbf{z}_n)^{(k)} - \mathbf{z}_n^{(y_n)} \right) + \sum_{n=1}^{N} \beta_{3,n} \sum_{k=1}^{K} \mathbf{z}_n^{(k)} \left(q(\mathbf{z}_n)^{(k)} - q(\mathbf{v}_n)^{(k)} \right).$$

Taking its derivative with respective to $q(\mathbf{z}_n)^{(k)}$ gives

$$\frac{\partial \mathcal{L}_S}{\partial q(\mathbf{z}_n)^{(k)}} = -1 - \log q(\mathbf{z}_n)^{(k)} + \beta_{1,n} + \beta_{2,n}\mathbf{z}_n^{(k)} + \beta_{3,n}\mathbf{z}_n^{(k)}.$$

Suppose $\beta_n = \beta_{2,n} + \beta_{3,n}$ for simplicity and it gives

$$q(\mathbf{z}_n)^{(k)} = \exp\left(\beta_n \mathbf{z}_n^{(k)}\right) / Z_S, \tag{8}$$

where $Z_S = \exp(1-\beta_{1,n}) = \sum_{k=1}^{K} \exp(\beta_n \mathbf{z}_n^{(k)})$ holds because of the normalization condition of probability density. The formulation in Eq. 8 has the same structure as Eq. 6.

Distinct Temperatures. Note that the partial derivatives of $\mathcal{L}_T$ with respective to $\alpha_{1,n}$ and $\alpha_{2,n}$ lead back to the two constraints respectively in Eq. 4 and the constraints are irrelevant to $\alpha_{1,n}$ and $\alpha_{2,n}$. A similar case holds for Eq. 7. As a result, no explicit expression of them can be given and thus their values can be manually defined. If we set $\beta_n = \alpha_{2,n} = 1/\mathcal{T}$, Eq. 6 and 8 turn to the expressions in KD involving a shared temperature for both student and teacher. When $\beta_n = \alpha_{2,n} = 1$, the formulations revert to the traditional `softmax` function commonly used in classification. Eventually, we can choose $\beta_n \neq \alpha_{2,n}$, indicating that students and teachers can have distinct temperatures.

Sample-wisely different Temperatures. It is common to define a global temperature for all samples. Namely for any n, $\alpha_{2,n}$ and β_n are defined as a constant value. In contrast, they could vary across different samples due to the lack of restrictions on them. It lacks a foundation to choose a global constant value as temperature. As a result, adopting sample-wise varying temperatures is allowed.

4.2. Drawbacks of Shared Temperatures

In this section, we show two shortcomings of the shared temperatures setting in conventional KD pipeline. We first rewrite the `softmax` in Eq. 8 in a general formulation by introducing two hyper-parameters a_S and b_S,

$$q\left(\mathbf{z}_n; a_S, b_S\right)^{(k)} = \frac{\exp\left[(\mathbf{z}_n^{(k)} - a_S)/b_S\right]}{\sum_{m=1}^{K} \exp\left[(\mathbf{z}_n^{(m)} - a_S)/b_S\right]},$$

where a_S can be cancelled out and does not violate the equality. When $a_S = 0, b_S = 1/\beta_n$, it yields back the special case in Eq. 8. The similar equation for the case of teacher can be written by introducing a_T and b_T.

For a finally well-distilled student, we assume the KL divergence loss reaches minimum and its predicted probability density matches that of teacher, i.e., $\forall k \in [1, K]$, $q(\mathbf{z}_n; a_S, b_S)^{(k)} = q(\mathbf{v}_n; a_T, b_T)^{(k)}$. Then for arbitrary pair of indices $i, j \in [1, K]$, it can easily lead to

$$\frac{\exp\left[(\mathbf{z}_n^{(i)} - a_S)/b_S\right]}{\exp\left[(\mathbf{z}_n^{(j)} - a_S)/b_S\right]} = \frac{\exp\left[(\mathbf{v}_n^{(i)} - a_T)/b_T\right]}{\exp\left[(\mathbf{v}_n^{(j)} - a_T)/b_T\right]}$$

$$\Rightarrow \left(\mathbf{z}_n^{(i)} - \mathbf{z}_n^{(j)}\right)/b_S = \left(\mathbf{v}_n^{(i)} - \mathbf{v}_n^{(j)}\right)/b_T.$$

By taking a summation across j from 1 to K, we have

$$\left(\mathbf{z}_n^{(i)} - \overline{\mathbf{z}}_n\right)/b_S = \left(\mathbf{v}_n^{(i)} - \overline{\mathbf{v}}_n\right)/b_T, \tag{9}$$

where $\overline{\mathbf{z}}_n$ and $\overline{\mathbf{v}}_n$ are the mean of the student and teacher logit vectors respectively, i.e., $\overline{\mathbf{z}}_n = \frac{1}{K}\sum_{m=1}^{K} \mathbf{z}_n^{(m)}$ (similar and omitted for $\overline{\mathbf{v}}_n$). By taking the summation of the squared Eq. 9 across i from 1 to K, we can obtain

$$\frac{\sigma(\mathbf{z}_n)^2}{\sigma(\mathbf{v}_n)^2} = \frac{\frac{1}{K}\sum_{i=1}^{K}\left(\mathbf{z}_n^{(i)} - \overline{\mathbf{z}}_n\right)^2}{\frac{1}{K}\sum_{i=1}^{K}\left(\mathbf{v}_n^{(i)} - \overline{\mathbf{v}}_n\right)^2} = \frac{b_S^2}{b_T^2}, \tag{10}$$

where σ is the function of standard deviation for an input vector. From Eq.9 and 10, we can describe two properties of a well-distilled student in terms of the logit shift and variance matching.

Logit shift. From Eq.9, it can be found that a constant shift exists between the logits of student and teacher in arbitrary index under the traditional setting of shared temperature ($b_S = b_T$), i.e.,

$$\mathbf{z}_n^{(i)} = \mathbf{v}_n^{(i)} + \Delta_n, \tag{11}$$

where $\Delta_n = \overline{\mathbf{z}}_n - \overline{\mathbf{v}}_n$ can be considered as constant for the n-th sample. This implies in the traditional KD approach, student is forced to strictly mimic the shifted logit of teachers. Considering the gap of their model size and capacity, student may be unable to produce as wide logit range as teacher [7, 30, 37]. In contrast, a student can be considered excellent enough when its logit rank matches teacher [15], i.e., given the indices that sorts the teacher logits $t_1, ..., t_K \in [1, K]$ such that $\mathbf{v}_n^{(t_1)} \leq \cdots \leq \mathbf{v}_n^{(t_K)}$, then $\mathbf{z}_n^{(t_1)} \leq \cdots \leq \mathbf{z}_n^{(t_K)}$ holds. The logit relation is the essential knowledge that makes student predict as excellently as teacher. Such a logit shift is thus a side-effect in conventional KD pipeline and a shackle compelling student to generate unnecessarily difficult results.

Variance match. From Eq. 10, we come into a conclusion that the ratio between the temperatures of student and teacher equals the ratio between the standard deviations of their predicted logits for a well-distilled student, i.e.,

$$\sigma(\mathbf{z}_n)/\sigma(\mathbf{v}_n) = b_S/b_T. \tag{12}$$

In the setting of temperature sharing in vanilla KD, the student is forced to predict logit such that $\sigma(\mathbf{z}_n) = \sigma(\mathbf{v}_n)$. This is another shackle applied to student restricting the standard deviation of its predicted logits. In contrast, since the hyper-parameter comes from Lagrangian multiplier and is flexible to tune, we can define $b_S^* \propto \sigma(\mathbf{z}_n)$ and $b_T^* \propto$

Algorithm 1: Weighted $\mathcal{Z}$-score function.

Input: Input vector $\mathbf{x}$ and Base temperature τ
Output: Standardized vector $\mathcal{Z}(\mathbf{x};\tau)$

$\overline{\mathbf{x}} \leftarrow \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^{(k)}$
$\sigma(\mathbf{x}) \leftarrow \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{x}^{(k)}-\overline{\mathbf{x}}\right)^2}$
return $(\mathbf{x}-\overline{\mathbf{x}})/\sigma(\mathbf{x})/\tau$

Algorithm 2: $\mathcal{Z}$-score logit standardization pre-process in knowledge distillation.

Input: Transfer set $\mathcal{D}$ with image-label sample pair $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$, Base Temperature τ, Teacher f_T, Student f_S, Loss $\mathcal{L}_{\mathrm{KD}}$ (e.g., $\mathcal{L}_{\mathrm{KL}}$), loss weight λ, and $\mathcal{Z}$-score function $\mathcal{Z}$ in Algo. 1
Output: Trained student model f_S

foreach $(\mathbf{x}_n, y_n)$ *in* $\mathcal{D}$ **do**
  $\mathbf{v}_n \leftarrow f_T(\mathbf{x}_n)$, $\mathbf{z}_n \leftarrow f_S(\mathbf{x}_n)$
  $q(\mathbf{v}_n) \leftarrow \texttt{softmax}\left[\mathcal{Z}(\mathbf{v}_n;\tau)\right]$
  $q(\mathbf{z}_n) \leftarrow \texttt{softmax}\left[\mathcal{Z}(\mathbf{z}_n;\tau)\right]$
  $q'(\mathbf{z}_n) \leftarrow \texttt{softmax}\left(\mathbf{z}_n\right)$
  Update f_S towards minimizing
  $\lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}}\left(y_n, q'(\mathbf{z}_n)\right) + \lambda_{\mathrm{KD}}\tau^2\mathcal{L}\left(q(\mathbf{v}_n), q(\mathbf{z}_n)\right)$
end

$\sigma(\mathbf{v}_n)$. In that way, the equation in Eq. 12 always holds.

4.3. Logit Standardization

To break the two shackles, we therefore propose setting the hyper-parameters a_S, b_S, a_T and b_S to be the mean and the weighted standard deviation of their logits respectively, i.e.,

$$q\left(\mathbf{z}_n;\overline{\mathbf{z}}_n,\sigma(\mathbf{z}_n)\right)^{(k)} = \frac{\exp(\mathcal{Z}(\mathbf{z}_n;\tau)^{(k)})}{\sum_{m=1}^{K}\exp(\mathcal{Z}(\mathbf{z}_n;\tau)^{(m)})},$$

where $\mathcal{Z}$ is the $\mathcal{Z}$-score function in Algo. 1. The case for teacher logit is similar and omitted. A base temperature τ is introduced and shared for both teacher and student models. The $\mathcal{Z}$-score standardization has at least four advantageous properties, i.e., zero mean, finite standard deviation, monotonicity and boundedness.

Zero mean. The mean of standardized vector can be easily shown to be zero. In previous works [9, 13], a zero mean is assumed and usually empirically violated. In contrast, the $\mathcal{Z}$-score function intrinsically guarantees a zero mean.

Finite standard deviation. The standard deviation of the weighted $\mathcal{Z}$-score output $\mathcal{Z}(\mathbf{z}_n;\tau)$ can be shown equal to $1/\tau$. The property makes the standardized student and teacher logit map to an identical Gaussian-like distribution with zero mean and definite standard deviation. The mapping of standardization is many-to-one, meaning that its reverse is indefinite. The variance and value range of original student logit vector $\mathbf{z}_n$ is thereby free of restriction.

Monotonicity. It is easy to show that $\mathcal{Z}$-score is a linear transformation function and thus lies in monotonic functions. This property ensures that the transformed student logit remains the same rank as the original one. The necessary innate relation within teacher logit can thus be preserved and transferred to student.

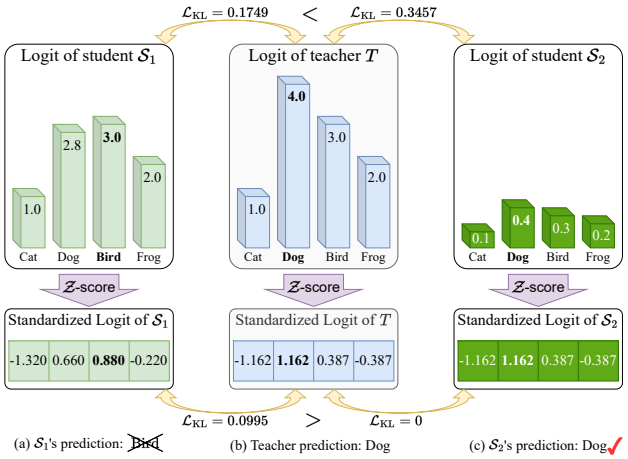


Figure 2. A toy case where two students, $\mathcal{S}_1$ and $\mathcal{S}_2$, learning from the same teacher with an identical temperature (assumed 1 for simplicity). Student $\mathcal{S}_1$ generates the logits much closer to the teacher's in terms of magnitude and thus has lower loss of 0.1749, but it returns a wrong prediction of "bird". In contrast, Student $\mathcal{S}_2$ outputs the logits far from the teacher's and yields greater loss value of 0.3457, but it returns the correct prediction of "dog". After the proposed logit standardization, the issue is addressed.

Boundedness. The standardized logit can be shown bounded within $[-\sqrt{K-1}/\tau, \sqrt{K-1}/\tau]$. Compared to traditional KD, it is feasible to control the logit range and avoid extremely large exponential value. To this end, we define a base temperature to control the range.

The pseudo-code of the proposed logit standardization pre-process is illustrated in Algo. 2.

4.3.1 Toy Case

Fig. 2 shows a typical case where the conventional logit-based KD setting of shared temperature may lead to an inauthentic evaluation of student performance. The first student $\mathcal{S}_1$ predicts logits closer to the teacher T in terms of magnitude, while the second student $\mathcal{S}_2$ preserves the same innate logit relations as the teacher. As a result, $\mathcal{S}_1$ obtains lower KL divergence loss of 0.1749, much better than the second student $\mathcal{S}_2$ (0.3457). However, $\mathcal{S}_1$ has a wrong prediction of "Bird" while $\mathcal{S}_2$ predicts "Dog" correctly, which contradicts the loss comparison. By applying our $\mathcal{Z}$-score, all logits are re-scaled and the relations among logits instead of their magnitudes are emphasized in the evaluation. Namely, $\mathcal{S}_2$ gets a loss of 0, much better than $\mathcal{S}_1$ of 0.0995, which is in line with the observed predictions.

5. Experiments

Datasets. We conduct experiments on CIFAR-100 [18] and ImageNet [34]. CIFAR-100 [18] is a common dataset

Table 1. The Top-1 Accuracy (%) of different knowledge distillation methods on the validation set of CIFAR-100 [18]. The teacher and student have distinct architectures. The KD methods are sorted by the types, i.e., feature-based and logit-based. We apply our logit standardization to the existing logit-based methods and use Δ to show its performance gain. The values in blue denote slight enhancement and those in red non-trivial enhancement no less than 0.15. The best and second best results are emphasized in **bold** and underlined cases.

Type								
	Teacher	ResNet32×4 79.42	ResNet32×4 79.42	ResNet32×4 79.42	WRN-40-2 75.61	WRN-40-2 75.61	VGG13 74.64	ResNet50 79.34
	Student	SHN-V2 71.82	WRN-16-2 73.26	WRN-40-2 75.61	ResNet8×4 72.50	MN-V2 64.60	MN-V2 64.60	MN-V2 64.60
Feature	FitNet [33]	73.54	74.70	77.69	74.61	68.64	64.16	63.16
	AT [53]	72.73	73.91	77.43	74.11	60.78	59.40	58.58
	RKD [31]	73.21	74.86	77.82	75.26	69.27	64.52	64.43
	CRD [39]	75.65	75.65	78.15	75.24	70.28	69.73	69.11
	OFD [12]	76.82	76.17	79.25	74.36	69.92	69.48	69.04
	ReviewKD [5]	77.78	76.11	78.96	74.34	71.28	70.37	69.89
	SimKD [4]	78.39	77.17	79.29	75.29	70.10	69.44	69.97
	CAT-KD [10]	78.41	76.97	78.59	75.38	70.24	69.13	**71.36**
Logit	KD [13]	74.45	74.90	77.70	73.97	68.36	67.37	67.35
	KD+Ours	75.56	75.26	77.92	77.11	69.23	68.61	69.02
	Δ	1.11	0.36	0.22	3.14	0.87	1.24	1.67
	CTKD [24]	75.37	74.57	77.66	74.61	68.34	68.50	68.67
	CTKD+Ours	76.18	75.16	77.99	77.03	69.53	68.98	69.36
	Δ	0.81	0.59	0.33	2.42	1.19	0.48	0.69
	DKD [57]	77.07	75.70	78.46	75.56	69.28	69.71	70.35
	DKD+Ours	77.37	76.19	78.95	76.75	70.01	69.98	70.45
	Δ	0.30	0.49	0.49	1.19	0.73	0.27	0.10
	MLKD [17]	78.44	76.52	79.26	77.33	70.78	70.57	71.04
	MLKD+Ours	**78.76**	**77.53**	**79.66**	**77.68**	**71.61**	**70.94**	71.19
	Δ	0.32	1.01	0.40	0.35	0.83	0.37	0.15

for image classification consisting of 50,000 training and 10,000 validation images. It contains 100 classes and its image size is 32×32. ImageNet [34] is a large-scale dataset for image classification containing 1,000 categories and around 1.28 million training and 50,000 validation images.

Baselines. We evaluate the effect of our logit standardization as pre-process for multiple logit-based KD approaches, including KD [13], CTKD [24], DKD [57] and MLKD[17]. We also compare with various feature-based KD methods including FitNet [33], RKD [31], CRD [32], OFD [12], ReviewKD [5], SimKD [4], and CAT-KD [10].

Implementation Details. We follow the same experimental settings as previous works [5, 17, 57]. For the experiments on CIFAR-100, the optimizer is SGD [38] and the epoch number is 240, except for MLKD being 480 [17]. The learning rate is set initially 0.01 for MobileNets[14, 35] and ShuffleNets [55] and 0.05 for other architectures consisting of ResNets [11], WRNs [52] and VGGs [36]. All results are reported by taking average over 4 trials. More detailed experimental settings are elaborated in the supplements.

5.1. Main Results

Results on CIFAR-100. We compare the KD results of different methods under various teacher/student settings in Tab. 1 and 2. Tab. 1 shows the cases that the teacher and student models have distinct structures, while Tab. 2 demonstrates the cases that they have the same architecture.

We evaluate our pre-process on four existing logit-based KD methods. As shown, our $\mathcal{Z}$-score standardization can constantly improve their performances. After applying our pre-process, the vanilla KD [13] achieves comparable performance to the state-of-the-art (SOTA) feature-based methods. As SOTA logit-based methods, DKD [57] and MLKD [17] can also be further boosted by our pre-process. CTKD [24] is a KD method that can determine sample-wise temperatures. We combine it with ours by leveraging it to predict the base temperature in Algo. 2. As illustrated in tables, student models distilled by CTKD benefits from our pre-process as well.

Results on ImageNet of different methods in terms of top-1 and top-5 accuracy are compared in Tab. 3. Our pre-process can achieve consistent improvement for all the three logit-based methods on the large-scale dataset as well.

Ablation Studies We conduct extensive ablation studies in terms of different configurations of base temperature and the weight of KD loss λ_{KD}. The part of results when base temperature is 2 are shown in Tab. 4. We can see as the weight of KD loss increases, the vanilla KD where `softmax` takes the original logit vectors as input cannot have a considerable performance gain. In contrast, our $\mathcal{Z}$-score pre-process achieves a significant enhancement. More ablation studies of other settings are in the supplements.

The weight of KD loss is set relatively larger than that of CE loss because our pre-process enables student to focus more on the dark knowledge from teacher instead of hard labels. Another reason of the large KD weight is to compensate the gradient involving $\mathcal{Z}$-score.

Table 2. The Top-1 Accuracy (%) of different knowledge distillation methods on the validation set of CIFAR-100 [18]. The teacher and student have identical architectures but different configurations. The KD methods are sorted by the types. We apply our logit standardization to the existing logit-based methods and use Δ to show its performance gain. The values in blue denote slight enhancement and those in red non-trivial enhancement no less than 0.15. The best and second best results are emphasized in **bold** and underlined cases.

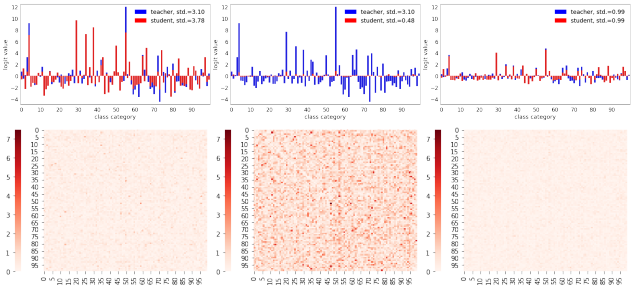
Type								
	Teacher	ResNet32×4 79.42	VGG13 74.64	WRN-40-2 75.61	WRN-40-2 75.61	ResNet56 72.34	ResNet110 74.31	ResNet110 74.31
	Student	ResNet8×4 72.50	VGG8 70.36	WRN-40-1 71.98	WRN-16-2 73.26	ResNet20 69.06	ResNet32 71.14	ResNet20 69.06
Feature	FitNet [33]	73.50	71.02	72.24	73.58	69.21	71.06	68.99
	AT [53]	73.44	71.43	72.77	74.08	70.55	72.31	70.65
	RKD [31]	71.90	71.48	72.22	73.35	69.61	71.82	69.25
	CRD [39]	75.51	73.94	74.14	75.48	71.16	73.48	71.46
	OFD [12]	74.95	73.95	74.33	75.24	70.98	73.23	71.29
	ReviewKD [5]	75.63	74.84	75.09	76.12	71.89	73.89	71.34
	SimKD [4]	78.08	74.89	74.53	75.53	71.05	73.92	71.06
	CAT-KD [10]	76.91	74.65	74.82	75.60	71.62	73.62	71.37
Logit	KD [13]	73.33	72.98	73.54	74.92	70.66	73.08	70.67
	KD+Ours	76.62	74.36	74.37	76.11	71.43	74.17	71.48
	Δ	3.29	1.38	0.83	1.19	0.77	1.09	0.81
	KD+CTKD [24]	73.39	73.52	73.93	75.45	71.19	73.52	70.99
	KD+CTKD+Ours	76.67	74.47	74.58	76.08	71.34	74.01	71.39
	Δ	3.28	0.95	0.65	0.63	0.15	0.49	0.40
	DKD [57]	76.32	74.68	74.81	76.24	71.97	74.11	71.06
	DKD+Ours	77.01	74.81	74.89	76.39	72.32	74.29	71.85
	Δ	0.69	0.13	0.08	0.15	0.35	0.18	0.79
	MLKD [57]	77.08	75.18	75.35	76.63	72.19	74.11	71.89
	MLKD+Ours	**78.28**	**75.22**	**75.56**	**76.95**	**72.33**	**74.32**	**72.27**
	Δ	1.20	0.04	0.21	0.32	0.14	0.21	0.38

Table 3. The top-1 and top-5 accuracy (%) on the ImageNet validation set [34]. The best and second best results are emphasized in **bold** and underlined.

Teacher/Student	ResNet34/ResNet18		ResNet50/MN-V1	
Accuracy	top-1	top-5	top-1	top-5
Teacher	73.31	91.42	76.16	92.86
Student	69.75	89.07	68.87	88.76
AT [53]	70.69	90.01	69.56	89.33
OFD [12]	70.81	89.98	71.25	90.34
CRD [39]	71.17	90.13	71.37	90.41
ReviewKD [5]	71.61	90.51	72.56	91.00
SimKD [4]	71.59	90.48	72.25	90.86
CAT-KD [10]	71.26	90.45	72.24	91.13
KD [13]	71.03	90.05	70.50	89.80
KD+Ours	71.42$_{+0.39}$	90.29$_{+0.24}$	72.18$_{+1.68}$	90.80$_{+1.00}$
KD+CTKD [24]	71.38	90.27	71.16	90.11
KD+CTKD+Ours	71.81$_{+0.43}$	90.46$_{+0.19}$	72.92$_{+1.76}$	91.25$_{+1.14}$
DKD [57]	71.70	90.41	72.05	91.05
DKD+Ours	71.88$_{+0.18}$	90.58$_{+0.17}$	72.85$_{+0.80}$	91.23$_{+0.18}$
MLKD [17]	71.90	90.55	73.01	91.42
MLKD+Ours	**72.08**$_{+0.18}$	**90.74**$_{+0.19}$	**73.22**$_{+0.21}$	**91.59**$_{+0.17}$

5.2. Extensions

Logit range. We plot a bar plot in the first row of Fig. 3 to show an example of logit range. The second row of Fig. 3 illustrates the extent of average logit difference between teacher and student. Without applying our pre-process, the student fails to produce as large logit as the teacher at the target label index (7.5 v.s. 12). The mean average distance between the teacher and student logits also reaches 0.27. The restriction of logit range impedes the student to predict correctly. In contrast, our pre-process breaks the restriction and enables student to generate the logits of appropriate range for itself. Its effective logit output after standardization on the contrary matches the teacher's nicely. The mean distance of standardized logits also shrinks to 0.18, implying its better mimicking the teacher.



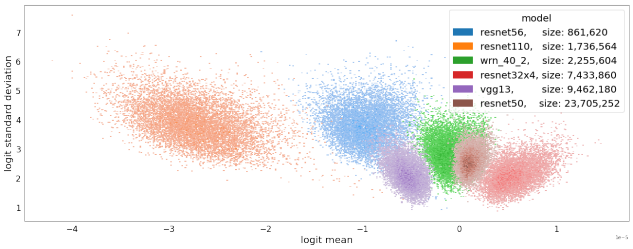
(a) Vanilla KD Mean: 0.27, Max: 3.03. (b) Ours w/o $\mathcal{Z}$-score Mean: 0.94, Max: 7.36. (c) Ours w/ $\mathcal{Z}$-score Mean: 0.18, Max:1.18.

Figure 3. **1st Row**: An example bar plot of logit output. **2nd Row**: The heatmap of the average logit difference between the teacher and student. Our pre-process indeed enables the student to generate the logits of divergent range from the teacher as shown in 3b, while its standardized logits (3c) are more closer to the teacher's than vanilla KD (3a).

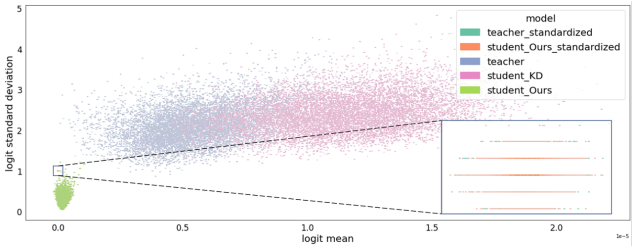
Logit variance. As illustrated in the first row of Fig. 3, the vanilla KD forces the variance of the student logits approaches the teacher's (3.78 v.s. 3.10). However, our pre-process breaks the shackle and the student logit can have flexible logit variance (0.48 v.s. 3.10), while its standard-

(a) Teacher (b) KD [13] (c) CTKD [24] (d) DKD [57] (e) MLKD [17]

Figure 4. The t-SNE [44] visualization of features. The teacher and student are ResNet32×4 and ResNet8×4.



(a) Teacher models of different sizes



(b) Teacher and student distilled by KD only and KD with our pre-process

Figure 5. The bivariate histogram of logit mean and logit standard deviation for multiple models on CIFAR-100.

ized logits have the same variance as the teacher (both 0.99).

Feature visualizations of deep representations are shown in Fig. 4. As implied, our pre-process improves the feature separability and discriminability of all the methods including KD [13], CTKD [24], DKD [57] and MLKD [17].

Improving distillation for large teacher. Existing works [7, 46, 57] observe that a larger teacher is unnecessarily a better one. The phenomenon implies that the transfer of knowledge from a large teacher is not always as smooth as a small one. They explain the observation with the existence of a capacity gap between cumbersome teachers and lightweight students. We interpret the issue as the difficulty of students in mimicking logits of comparable range and variance as teachers. As shown in the bivariate histogram of Fig. 5a, the bigger teachers like ResNet50 and VGG13 generates more condensed logits of mean closer to zero and smaller standard deviation. The tendency shows that smaller models intrinsically predicts the outputs of larger bias from zero mean and larger variance. Therefore, when students are small, the same tendency remains and they are difficult to produce as compact logits as large teachers.

We alleviate the problem by our pre-process. As shown in Tab. 5, our pre-process consistently improves the distillation performance for various teachers of different sizes and capacities. We also show the mimicking goodness of students by a bivariate histogram plot in Fig. 5b. The student distilled by vanilla KD predicts logits apparently deviated from teachers' in terms of logit mean and standard deviation. In contrast, our pre-process enables the student to perfectly matching the teacher in terms of standardized logit mean and standard deviation (see the zoomed-in region at bottom right corner).

Table 4. The ablation studies under different settings in our $\mathcal{Z}$-score. The base temperature τ is set to be 2. By default $\lambda_{\text{CE}} = 0.1$. The logit vector of teacher $\mathbf{v}_n$ and student $\mathbf{z}_n$ are abbreviated as $\mathbf{z}$ for succinctness. The teacher and student are ResNet32×4 and ResNet8×4.

λ_{KD}	$\mathbf{z}$ (KD)	$\mathbf{z}-\overline{\mathbf{z}}$	$\frac{\mathbf{z}}{\sigma(\mathbf{z})}$	$\frac{(\mathbf{z}-\overline{\mathbf{z}})}{\sigma(\mathbf{z})}$ (Ours)
0.9	73.60	73.37	73.79	74.14
3.0	74.38	74.33	75.86	76.11
6.0	74.45	74.82	76.44	76.56
9.0	73.33	73.94	76.30	76.62
12.0	68.29	71.56	76.49	76.56
15.0	65.34	62.01	76.42	76.61
18.0	63.45	61.31	76.18	76.33

Table 5. The results of distillation of various teacher models on CIFAR-100. The student model is WRN-16-2.

Teacher	VGG13	W-28-2	W-40-2	W-16-4	W-28-4	ResNet50
	74.64	75.45	75.61	77.51	78.60	79.34
KD [13]	74.93	75.37	74.92	75.79	75.04	75.36
KD+Ours	75.03	76.32	76.11	76.72	75.77	76.24
DKD [57]	75.45	75.92	76.24	76.00	76.45	76.60
DKD+Ours	75.56	76.39	76.39	76.68	76.67	76.82

More experiments of distilling vision Transformers [6, 20, 21, 41, 42, 47–49], and two more datasets i.e., CUB200 [45] and COCO [26], are attached in the supplements.

6. Conclusion

In this work, we identify a lack of theoretical support for the global and shared temperature in conventional KD pipelines. Our analysis based on the principle of entropy maximization leads to the conclusion that the temperature is derived from a flexible Lagrangian multiplier, allowing for a flexible value assignment. We then highlight several drawbacks associated with the conventional practice of sharing temperatures between teachers and students. Additionally, we presented a toy case where the KD pipeline with shared temperature led to an inauthentic evaluation of student performance. To mitigate the concerns, we propose a logit $\mathcal{Z}$-score standardization as a pre-process to enable student to focus on the innate relations of teacher logits rather than logit magnitude. The extensive experiments demonstrate the effectiveness of our pre-process in enhancing the existing logit-based KD methods.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, 2019. 2

[2] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Yunqing Zhao, and Ngai-Man Cheung. Revisiting label smoothing and knowledge distillation compatibility: What was missing? In *ICML*, 2022. 2

[3] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *AAAI*, 2020. 2

[4] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *CVPR*, 2022. 2, 6, 7

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 2, 6, 7

[6] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: Data-efficient early knowledge distillation for vision transformers. In *CVPR*, 2022. 8

[7] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, 2019. 2, 4, 8

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 3

[9] Jia Guo. Reducing the teacher-student gap via adaptive temperatures, 2022. 2, 5

[10] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *CVPR*, 2023. 2, 6, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[12] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 2, 6, 7

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 5, 6, 7, 8

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 6

[15] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *NeurIPS*, 2022. 2, 4

[16] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957. 3

[17] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *CVPR*, 2023. 2, 6, 7, 8

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5, 6, 7

[19] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *AAAI*, 2022. 2

[20] Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *ECCV*, 2022. 8

[21] Lujun Li, Peijie Dong, Zimian Wei, and Ya Yang. Automated knowledge distillation via monte carlo tree search. In *ICCV*, 2023. 8

[22] Zheng Li, Ying Huang, Defang Chen, Tianren Luo, Ning Cai, and Zhigeng Pan. Online knowledge distillation via multi-branch diversity enhancement. In *ACCV*, 2020. 2

[23] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *ICCV*, 2021. 2

[24] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, 2023. 2, 6, 7, 8

[25] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *CVPR*, 2022. 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 8

[27] Jihao Liu, Boxiao Liu, Hongsheng Li, and Yu Liu. Meta knowledge distillation. *arXiv preprint arXiv:2202.07940*, 2022. 2

[28] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, 2019. 2

[29] Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M Patel, and Peyman Milanfar. Conditional diffusion distillation. *arXiv preprint arXiv:2310.01407*, 2023. 2

[30] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, 2020. 2, 4

[31] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 2, 6, 7

[32] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, 2019. 2, 6

[33] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *ICLR*, 2015. 2, 6, 7

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 2, 5, 6, 7

[35] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 6

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[37] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *ICCV*, 2021. 2, 4

[38] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013. 6

[39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2020. 2, 6, 7

[40] Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, and Nitesh V Chawla. Learning mlps on graphs: A unified view of effectiveness, robustness, and efficiency. In *ICLR*, 2023. 2

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 8

[42] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022. 8

[43] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019. 2

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[45] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 8

[46] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE TPAMI*, 2021. 8

[47] Ziyang Wang and Congying Ma. Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation. In *ICCV*, 2023. 8

[48] Ziyang Wang, Tianze Li, Jian-Qing Zheng, and Baoru Huang. When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation. In *ECCV*, 2022.

[49] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *ECCV*, 2022. 8

[50] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. In *ICLR*, 2021. 2

[51] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 2

[52] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6

[53] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *ICLR*, 2017. 6, 7

[54] Rongzhi Zhang, Jiaming Shen, Tianqi Liu, Jialu Liu, Michael Bendersky, Marc Najork, and Chao Zhang. Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. *arXiv preprint arXiv:2305.05010*, 2023. 2

[55] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 6

[56] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018. 2

[57] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022. 2, 6, 7, 8