RETHINKING SOFT LABELS FOR KNOWLEDGE DISTILLATION: A BIAS–VARIANCE TRADEOFF PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation is an effective approach to leverage a well-trained network or an ensemble of them, named as the teacher, to guide the training of a student network. The outputs from the teacher network are used as soft labels for supervising the training of a new network. Recent studies (Müller et al., 2019; Yuan et al., 2020) revealed an intriguing property of the soft labels that making labels soft serves as a good regularization to the student network. From the perspective of statistical learning, regularization aims to reduce the variance while the original hard labels are for bias error reduction. Therefore, during training soft labels are used for both bias error and variance reduction. In this paper, we investigate the bias-variance tradeoff brought by distillation with soft labels. Specifically, we observe that bias-variance tradeoffs are sample-wise and the distillation performance is negatively associated with the number of samples facing the tradeoff. Nevertheless, we empirically find that samples facing the tradeoff are still beneficial for distillation. Our discoveries inspired us to propose the novel weighted soft labels to help the network focus on the sample not facing the tradeoff. Experiments on standard evaluation benchmarks validate the effectiveness of our method. Our code is available at *.

1 INTRODUCTION

For deep neural networks (Goodfellow et al., 2016), knowledge distillation (Ba & Caruana, 2014; Hinton et al., 2015) refers to the technique that uses well-trained networks to guide the training of another network. Typically, the well-trained network is named as the teacher network while the network to be trained is named as the student network. For distillation, the predictions from the teacher network are leveraged and referred to as the soft labels (Balan et al., 2015; Müller et al., 2019). Soft labels generated by the teacher network have been proven effective in large-scale empirical studies (Liang et al., 2019; Tian et al., 2020; Zagoruyko & Komodakis, 2017; Romero et al., 2015) as well as recent theoretical studies (Phuong & Lampert, 2019).

However, the reason why soft labels are beneficial to the student network is still not well explained. Giving a clear theoretical explanation is challenging: The optimization details of a deep network with the common one-hot labels are still not well-studied (Nagarajan & Kolter, 2019), not to mention training with the soft labels. Nevertheless, two recent studies (Müller et al., 2019; Yuan et al., 2020) shed light on the intuitions about how the soft labels work. Specifically, label smoothing, which is a special case of soft labels based training, is shown to regularize the activations of the penultimate layer to the network (Müller et al., 2019). The regularization property of soft labels is further explored in (Yuan et al., 2020). They hypothesize that in knowledge distillation, one main reason why the soft labels work is the regularization introduced by soft labels. Based on the assumption, the authors design a teacher-free distillation method by turning the predictions of the student network into soft labels.

The evidence of the regularization brought by soft labels drives us to rethink soft labels for knowledge distillation: Soft labels are essentially targets in supervised learning, indicating that soft labels are both supervisory signals and regularizers. Meanwhile, it is known that there is a tradeoff between fitting the data and regularizing the model complexity, i.e., the bias-variance dilemma, according to statistical learning theories (Kohavi & Wolpert, 1996; Bishop, 2006). Recall that soft labels are supervision and regularizers, it can be inferred that soft labels can help reduce the bias error along

with the variance. Since the bias-variance tradeoff is an important issue in statistical learning, we investigate bias-variance tradeoffs for soft labels.

We first present an intuitive understanding of bias-variance tradeoff for soft labels, by exploring the semantics behind the regularization brought by soft labels. Then, through inspecting the gradient from bias reduction and variance reduction, we notice that for soft labels, bias-variance tradeoffs are sample-wise. Moreover, by looking into a conclusion from (Müller et al., 2019), we observe that the number of samples facing bias-variance tradeoff is negatively associated with the distillation performance. As for soft labels the bias-variance tradeoff is sample-wise, we propose to avoid sample-wise tradeoff, instead of tuning the tradeoff. To achieve this, we first examine the effectiveness of several straightforward solutions, leading us to think that samples facing the tradeoff are still useful for distillation but not fully exploited. In the light of these findings, we propose weighted soft labels for distillation to adaptively assign a lower weight to the samples facing the tradeoff. To sum up, our contributions are:

- For knowledge distillation, we analyze how the soft labels work from a perspective of bias-variance tradeoff.
- We discover that bias-variance tradeoffs are sample-wise and the number of samples facing the tradeoff is negatively associated with the distillation performance.
- We design straightforward schemes to avoid the tradeoffs and then propose the novel weighted soft labels for distillation. Experiments on large scale datasets validate the effectiveness of the proposed weighted soft labels.

2 Related Works

Knowledge distillation. Hinton et al. (2015) proposed to distill outputs from large and cumbersome models into smaller and faster models, which is named as knowledge distillation. The outputs for large networks are averaged and formulated as soft labels. Also, other kinds of soft labels have been widely used for training deep neural networks (Szegedy et al., 2016; Pereyra et al., 2017). Treating soft labels as regularizers were pointed out in (Hinton et al., 2015) since a lot of helpful information can be carried in soft labels. More recently, Müller et al. (2019) showed the adverse effect of label smoothing upon distillation. It is a thought-provoking discovery for the reason that both label smoothing and distillation are exploiting the regularization property behind soft labels. Yuan et al. (2020) further investigated the regularization property of soft labels and then proposed a teacher free distillation scheme.

Distillation loss. One of our main contributions is that we improve the distillation loss. For adaptively adjusting the distillation loss, TODO: Zhou: add here.

Bias–variance tradeoff. Bias–variance tradeoff is a well-studied topic in machine learning (Kohavi & Wolpert, 1996; Domingos, 2000; Bishop, 2006) and for neural networks (Geman et al., 1992; Neal et al., 2018; Belkin et al., 2019; Yang et al., 2020). Existing methods are mainly concerned with the variance brought by the choice of network models. Our perspective is different from the previous methods since we focus on the behavior of samples during training. Besides, our main contribution is not to study how to tune the tradeoff, but how to avoid sample-wise tradeoff.

3 Bias–Variance Tradeoff for Soft Labels

Soft labels play the role of ground-truth labels and regularizations at the same time. We begin our analysis with some mathematical descriptions. For a sample $\boldsymbol{X}$ labeled as i-th class, let the ground-truth label be a one-hot vector $\boldsymbol{y}$ where $y_i = 1$ and other entries are 0. Then for $\boldsymbol{X}$ and softmax output temperature τ, the soft label predicted by the teacher network is denoted as $T(\boldsymbol{X}, \tau)$ and the output from the student is denoted as $S(\boldsymbol{X}, \tau)$. The soft label $T(\boldsymbol{X}, \tau)$ is then used for training the student by the distillation

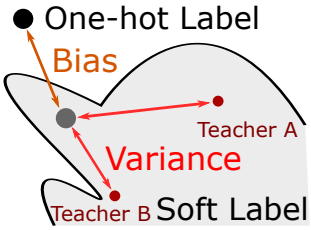
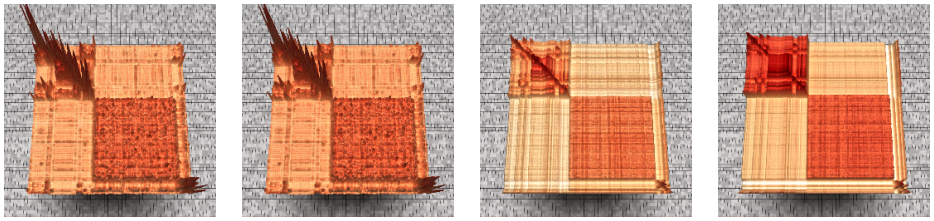


Figure 1: Bias and variance.



(a) (b) (c) (d)

Figure 2: Visualization of the resemblances introduced by soft label regularizers: (a) VGG-19 $\rightarrow$ VGG-16, (b) ResNet-50 $\rightarrow$ ResNet-18. And semantic similarity between label names: (c) LCH similarity (Pedersen et al., 2004), (d) WUP similarity (Pedersen et al., 2004). Darker areas means larger values.

loss, i.e. $L_{\mathrm{kd}} = \tau^2 \mathrm{KL}(S(\boldsymbol{X}, \tau), T(\boldsymbol{X}, \tau)) = \tau^2 \sum_k \hat{y}^t_{k,\tau} \log \hat{y}^s_{k,\tau}$, where KL means the Kullback-Leibler divergence and $\hat{y}^s_{k,\tau}, \hat{y}^t_{k,\tau}$ means the k-th element of the student's output $S(\boldsymbol{X}, \tau)$ and the teacher's output $T(\boldsymbol{X}, \tau)$, respectively. Following statistical learning theories (Hastie et al., 2009), we assume that there exists a function f such that $\boldsymbol{y} = f(\boldsymbol{X}) + \epsilon$, where the noise ϵ has zero mean and variance σ^2. Also, denote the model as $\hat{f}$, then for sample $\boldsymbol{X}$ and the squared-error loss is

$$\mathrm{Err}(\boldsymbol{X}) = \mathbb{E}[(\boldsymbol{y} - \hat{f}(\boldsymbol{X}))^2] = \sigma^2 + (\mathbb{E}[\hat{f}(\boldsymbol{X}) - f(\boldsymbol{X})])^2 + \mathbb{E}[\hat{f}(\boldsymbol{X}) - \mathbb{E}[\hat{f}(\boldsymbol{X})]]^2. \tag{1}$$

The term $(\mathbb{E}[\hat{f}(\boldsymbol{X}) - f(\boldsymbol{X})])^2$ is known as bias error, while the term $\mathbb{E}[\hat{f}(\boldsymbol{X}) - \mathbb{E}[\hat{f}(\boldsymbol{X})]]^2$ is the variance. Intuitively, if soft labels is used, the bias error during training will become larger since $\hat{f}$ no longer targets at learning the one-hot label. Meanwhile, the variance will become smaller because the output becomes smoothed. An illustration about the bias error and variance is shown in Fig. 1. As the original label $\boldsymbol{y}$ only contains the ground-truth information, we can view $\boldsymbol{y}$ as the part for reducing bias error. If we rewrite L_{kd} as $L_{\mathrm{kd}} = L_{\mathrm{kd}} - L_{\mathrm{ce}} + L_{\mathrm{ce}}$, then the cross-entropy loss $L_{\mathrm{ce}} = -y_k \log \hat{y}^s_{k,1}$ is for bias reduction. Therefore, the variance reduction realized by the distillation comes from the difference of the two losses, i.e., $L_{\mathrm{kd}} - L_{\mathrm{ce}}$.

In this section, we first give an intuitive analysis of the variance reduction by visualizing $L_{\mathrm{kd}} - L_{\mathrm{ce}}$. Then we analyze the changes in the relative importance between bias reduction and variance reduction during training with soft labels. Next, we conduct experiments to analyze the relative importance of bias reduction and variance reduction.

3.1 Intuitive understanding of bias-variance tradeoff

First of all, we dive into intuitively understanding bias–variance tradeoff of soft labels. Bias reduction plainly helps the model fit the training data, while variance reduction of soft labels is complicated. As analyzed above, the variance reduction is achieved by the loss term $L_{\mathrm{kd}} - L_{\mathrm{ce}}$. In other words, the term $L_{\mathrm{kd}} - L_{\mathrm{ce}}$ represents the regularization brought by training with soft labels. To have an intuitive understanding of how the regularization works, we propose to visualize its gradient on the logits of the student network. That is, let $\boldsymbol{z}$ be the logits output of the student and z_i is i-th element of it, then we are interested in visualizing $\frac{\partial(L_{\mathrm{kd}} - L_{\mathrm{ce}})}{\partial z_i}$. Mathematically we have

$$\frac{\partial(L_{\mathrm{kd}} - L_{\mathrm{ce}})}{\partial z_i} = \tau^3(\hat{y}^s_{i,\tau} - \hat{y}^t_{i,\tau}) - (\hat{y}^s_{i,1} - y_i) = \tau^3 \left(\frac{e^{z_i/\tau}}{\sum_k e^{z_k/\tau}} - \hat{y}^t_{i,\tau} \right) - \left(\frac{e^{z_i}}{\sum_k e^{z_k}} - y_i \right), \tag{2}$$

where $y^t_{i,\tau}$ denotes the i-th element of the teacher's prediction, i.e., $T(\boldsymbol{X}, \tau)$. The term $L_{\mathrm{kd}} - L_{\mathrm{ce}}$ is easy to understand when $\tau = 1$ since the gradient now becomes $y_i - \hat{y}^t_{i,1}$. As $y^t_{i,\tau}$ is the output from the teacher network and computed by a linear mapping of the activations in the teacher's penultimate layer, the regularization indicates that the student need to follow the learnt the resemblances between classes (Hinton et al., 2015; Müller et al., 2019). Still, two questions are unclear: 1) what the resemblances are and 2) whether the regularization still indicates resemblances if τ is set higher than 1, a widely adopted hyperparameter (Tian et al., 2020).

Table 1: We count the number of samples facing the bias-variance tradeoff (denoted as $\#\{\boldsymbol{X}|\operatorname{sign}(ab) = -1\}$) with different distillation settings on CIFAR-100. The teacher-student network pair is WRN-40-2 (Zagoruyko & Komodakis, 2017) and WRN-16-2 and results are averaged over 5 repeated runs. The temperature column means the value of temperature for distillation and the label smoothing column means whether the teacher network is trained with label smoothing trick.

Baseline Top-1 Acc	Teacher: 76.55 w/ label smoothing, 75.61 w/o label smoothing Student: 73.26		
Temperature	Label smoothing?	Student Top-1 Acc	$\#\{\boldsymbol{X}\|\operatorname{sign}(ab) = -1\}$
$\tau = 2$	✗	**74.89**	15379
	✓	74.76	25235
$\tau = 4$	✗	**75.06**	17709
	✓	74.59	24775
$\tau = 6$	✗	**75.10**	17408
	✓	74.56	24538

Towards answering the questions we now visualize the value of gradient vector $\frac{\partial(L_{\mathrm{kd}}-L_{\mathrm{ce}})}{\partial \boldsymbol{z}}$ concerning each class. Specifically, on ImageNet (Deng et al., 2009) training set with $\tau = 3$, we calculate the average value of $\frac{\partial(L_{\mathrm{kd}}-L_{\mathrm{ce}})}{\partial \boldsymbol{z}}$ for each class. Let $\boldsymbol{M}$ be the matrix of values with the ij-th entry M_{ij} means averaged $\frac{\partial(L_{\mathrm{kd}}-L_{\mathrm{ce}})}{\partial \boldsymbol{z}}$ for class j. Since $\sum_i M_{ij} = 0$, diagonal elements are ignored for visualization. The results are visualized in Fig. 2. We find that plotting the common correlation matrix heatmap is ambiguous, because the matrix to be visualized is of large size (1000×1000) with a large variance. By treating each entry M_{ij} as a vertex and then constructing a mesh for the matrix, we apply subdivision (Loop, 1987) to the mesh for smoothing the extreme points and finally rendering the mesh by ray-tracing package PlotOptiX. We can observe the several facts from the figures: 1) Comparing the sub-figure (a) and (b), we can see that for distillation resemblances implied by regularizers are similar across different teacher-student pairs. 2) The resemblances are consistent with the semantic similarity of image class names.

In a word, the regularization brought by soft labels implies resemblances among labels, which are consistent with the semantic distance of class names. In the next section, we will analyze how bias-variance tradeoff changes when training with soft labels.

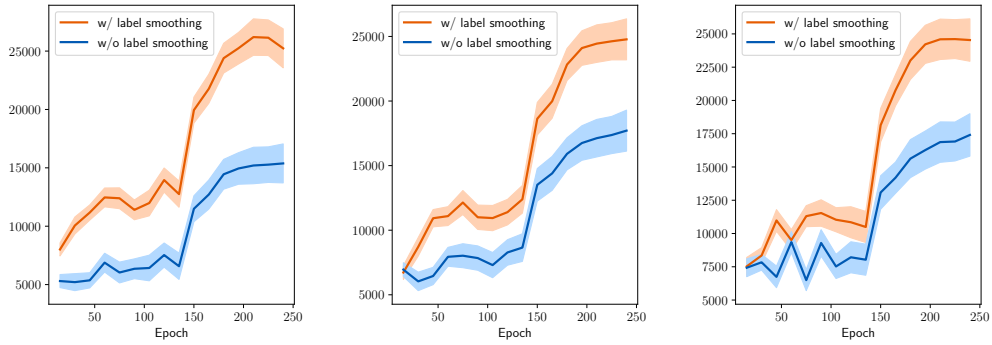
3.2 Sample-wise bias-variance tradeoff

Now that we are aware of what the variance reduction brought by soft labels is, a question then arises: How soft labels affect the final prediction error in terms of bias-variance tradeoff? We discover that the sample-wise bias-variance tradeoff affects the final prediction error.

Strictly speaking, if bias reduction and variance reduction lead to different gradient for the network weights, we need to take care of the bias-variance tradeoff. Also, it is known that bias reduction and variance reduction are often in conflict and we need to balance bias and variance. However, during the early phase of training, if we simply choose only bias reduction, both bias and variance can decrease since the model just begins to fit into the data. As training goes on, some samples start facing the bias-variance tradeoff if bias and variance cannot decrease at the same time. In other words, training with soft labels requires us to handle sample-wise bias-variance tradeoff.

To simplify our analysis, we consider an important tradeoff case that bias reduction and variance reduction come into conflict in the direction of optimizing the ground-truth related logit. More precisely, for a sample labeled as i-th class, if $\frac{\partial L_{\mathrm{ce}}}{\partial z_i}$ and $\frac{\partial(L_{\mathrm{kd}}-L_{\mathrm{ce}})}{\partial z_i}$ have a different sign, the bias reduction and variance reduction come into conflict and bias-variance dilemma arises. We discover that the final prediction error is higher when more training samples are facing the bias-variance tradeoff.

Our analysis starts with a conclusion from Müller et al. (2019): *if a teacher network is trained with label smoothing, knowledge distillation into a student network is much less effective.* Inspired by the



(a) Temperature $\tau = 2$ (b) Temperature $\tau = 4$ (c) Temperature $\tau = 6$

Figure 3: The number of samples facing bias-variance tradeoff, i.e., $\#\{\boldsymbol{X} | \operatorname{sign}(ab) = -1\}$ with respect to training epochs. The distillation settings are the same as the settings in Tab. 1.

phenomenon, we calculate the number of conflicting examples with different settings. That is, let $a = \frac{\partial L_{\text{ce}}}{\partial z_i}$ and $b = \frac{\partial (L_{\text{kd}} - L_{\text{ce}})}{\partial z_i}$, then we are interested in the number of instances with $\operatorname{sign}(ab) = -1$, when distilling from the teacher with or without label smoothing. In Tab. 1, we present the count of samples facing bias-variance tradeoff for a student network trained by knowledge distillation. For distillation with a temperature higher than 1, which is the common setting, we observe that if the teacher network is trained with label smoothing, more samples will face the bias-variance tradeoff. Also, distillation from a teacher trained with label smoothing performs worse, which is consistent with Müller et al. (2019). Considering that the performance of the student is negatively associated with the number of conflicting samples, our answer to the question asked at the beginning of this section is: For soft labels, samples facing bias-variance tradeoff affect the final distillation performance.

Moreover, we plot the number of samples facing bias-variance tradeoff with respect to different training epochs in Fig. 3. As demonstrated in the plots, the number of such samples increases much faster when using the teacher trained with label smoothing for distillation. The gap of the number between with and without smoothing becomes larger for more training epochs. These observations verify our motivation that bias-variance tradeoffs are sample-wise and evolve during the training process.

From the above results, we conclude that bias-variance tradeoff for soft labels is sample-wise, which makes tuning the tradeoff harder than those loss based bias-variance tradeoffs. In the next section, we investigate whether we can artificially tune the soft labels such that bias reduction does not conflict with variance reduction.

3.3 Samples facing bias-variance tradeoff are useful

The results presented in the last section suggest that we should avoid training with samples facing bias-variance tradeoff. Hence, we design two straightforward solutions and then find that training with samples facing bias-variance tradeoff is beneficial.

The first experiment we conduct is to manually resolve the conflicting gradient on the label related logit, as defined in section 3.2. Specifically, we apply a mask to the distillation loss L_{kd} such that $\frac{\partial L_{\text{kd}}}{\partial z_i} = 0$ if $y_i = 1$. Consequently, the loss for this sample now becomes $L^*_{\text{kd}} = \sum_{k \neq i} \hat{y}^t_{k,\tau} \log \hat{y}^s_{k,\tau}$. For knowledge transfer, the motivation behind the mask distillation loss is that only knowledge contained in the soft part, which reflects resemblances, is distilled from the teacher network. Another experiment is to figure out what role in distillation those conflicting samples will play. To investigate this, we carry out knowledge distillation on two subsets of samples: 1) L_{kd} is valid when $\operatorname{sign}(ab) = -1$ and 2) L_{kd} is valid when $\operatorname{sign}(ab) = 1$. In other words, knowledge distillation is performed on only conflicting samples or only not conflicting samples.

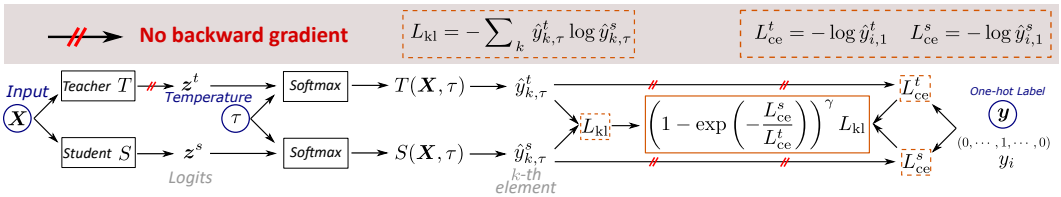


Figure 4: Computational graph of knowledge distillation with our proposed weighted soft labels.

Table 2: Study of the impact on distillation for the samples facing bias-variance tradeoff. Loss function presented here is for the loss on a specific sample. Results are classification Top-1 accuracy. We follow the settings used in Tab. 1 and set $\tau = 4$. Conflicting samples refer to the samples facing bias-variance tradeoff as defined in Sec 3.2.

Teacher: 75.61; Student with direct training: 73.26; Student with standard KD: 74.92				
Setting	Loss function	Student performance	Performance gap	
			to direct training	to KD
Mask KL loss on the label related logit	$L_{\mathrm{ce}} + L^*_{\mathrm{kd}}$	73.51	+0.25	-1.41
Only on conflicting samples	$\begin{cases} L_{\mathrm{ce}}, \text{if } \operatorname{sign}(ab) \geqslant 0 \\ L_{\mathrm{kd}}, \text{if } \operatorname{sign}(ab) = -1 \end{cases}$	73.86	+0.60	-1.06
Only on not conflicting samples	$\begin{cases} L_{\mathrm{ce}}, \text{if } \operatorname{sign}(ab) = -1 \\ L_{\mathrm{kd}}, \text{if } \operatorname{sign}(ab) \geqslant 0 \end{cases}$	74.59	+1.33	-0.33

The results of the two experiments are presented in Tab. 2. We can observe that all of the three approaches are not as good as the baseline knowledge distillation performance, but are better than the direct training baseline. First, since masking L_{kd} loss result in worse performance, $\operatorname{sign}(ab)$ is just an indicator, rather than a sufficient condition, for judging conflicting samples. So we cannot directly resolve the tradeoff unless a sufficient condition is theoretically derived. Moreover, from the other two experiments, we see that performing on the subsets are less effective. Thus, the conflicting samples are still valuable for distillation.

4 Weighted Soft Labels

From the last section, we realize that there exist sample-wise bias-variance tradeoffs and the number of these samples is negatively associated with the final distillation performance. Yet, such samples are still useful for distillation. The above evidence inspires us to lower the weight of the conflicting samples. However, the metric $\operatorname{sign}(ab)$ used previously is binary, which brings difficulty to adaptively adjusting the weight. Recall that $\operatorname{sign}(ab)$ is designed to see if the distillation loss is in conflict with the cross-entropy loss, we now need a soft metric achieving the same goal. Specifically, we use $\frac{L^t_{\mathrm{ce}}}{L^s_{\mathrm{ce}}} = \frac{\log \hat{y}^s_{i,1}}{\log \hat{y}^t_{i,1}}$ as a substitution to $\operatorname{sign}(ab)$. In this way, if a student network is poorly trained on this sample and produces a large ratio, it is likely that this sample is not facing the bias-variance tradeoff, thus we can increase the weight of this sample. Otherwise, if the student performs better than the teacher on this sample, resulting in a small ratio, this sample will be assigned a small weight.

In the light of previous works that assign sample-wise weights (Lin et al., 2017) TODO: more paper?, we propose weighted soft labels for knowledge distillation, which is formally defined as

$$L_{\mathrm{wsl}} = \alpha \left(1 - \exp\left(-\frac{L^s_{\mathrm{ce}}}{L^t_{\mathrm{ce}}}\right)\right)^{\beta} L_{\mathrm{kd}} = \alpha \left(1 - \exp\left(-\frac{\log \hat{y}^s_{i,1}}{\log \hat{y}^t_{i,1}}\right)\right)^{\beta} L_{\mathrm{kd}}, \tag{3}$$

where i is the ground truth class of the sample. The above equation means that a weight is assigned to each sample's L_{kd}. In Fig. 4, the whole computational graph of knowledge distillation with the proposed weighted soft labels is demonstrated. Finally, we add L_{wsl} and L_{ce} together to supervise the network, i.e., $L_{\mathrm{total}} = L_{\mathrm{ce}} + L_{\mathrm{wsl}}$.

5 EXPERIMENTS

To evaluate our weighted soft labels comprehensively, we first conduct experiments with various teacher-student pair settings on CIFAR-100 (Krizhevsky et al., 2009). Next, we compare our method with current state-of-the-art distillation methods on ImageNet (Deng et al., 2009). To validate the effectiveness of our method in terms of handling bias-variance tradeoffs, we conduct ablation experiments by applying weighted soft labels on different subsets. Finally, we analyze the impact of hyperparameter defined for the weighted soft labels in Eq. (3).

5.1 DATASET AND HYPERPARAMETER SETTINGS

The datasets used in our experiments are CIFAR-100 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009). CIFAR-100 contains 50K training and 10K test images of size 32×32. ImageNet contains 1.2 million training and 50K validation images. Except the loss function, training settings like learning rate or training epochs are the same with Tian et al. (2020) for CIFAR-100 and Heo et al. (2019) for ImageNet. For distillation, we set the temperature $\tau = 2$. For loss function, we set TODO: $\alpha = 2, \beta = 1$ for distillation on CIFAR and TODO: $\alpha = 2, \beta = 1$ for ImageNet via grid search.

For comparison, the following recent state-of-the-art methods are chosen: FitNet (Romero et al., 2015), AT (Zagoruyko & Komodakis, 2017), SP (Tung & Mori, 2019), CC (Peng et al., 2019), VID (Ahn et al., 2019), RKD (Park et al., 2019), PKT (Passalis & Tefas, 2018), AB (Heo et al., 2019), FT (Kim et al., 2018), FSP (Yim et al., 2017), NST (Huang & Wang, 2017).

5.2 MODEL COMPRESSION

Results on CIFAR-100 In Tab. 3, we present the Top-1 classification accuracy of our method and comparison methods. The results of comparison methods are quoted from Tian et al. (2020). Teacher-student pairs of the same and different architecture styles are considered, which again follows the setting in Tian et al. (2020). For pairs of same architecture style, we use wide residual networks (Zagoruyko & Komodakis, 2017) and residual networks (He et al., 2016). For pairs of different architecture style, residual networks and ShuffleNet (Zhang et al., 2018) pairs are chosen for experiments. As shown in the table, for distillation with both same and different architecture style, our method reached new state-of-the-art results.

Results on ImageNet In Tab. 4, we compare our method with current SOTA methods on ImageNet. Note that for the ResNet34 $\rightarrow$ ResNet-18 distillation setting, the result of CRD+KD is trained 10 more extra epochs while ours is the same as other methods. Our results outperform all the existing methods, verifying the practical value of our method.

5.3 ABLATION STUDIES

Weighted soft labels on different subsets. Recall that we propose weighted soft labels for sample-wise bias-variance tradeoff, it is still unclear whether the improvements come from a well-handled sample-wise bias-variance tradeoff. To investigate this issue, we compare the performance gain of weighted soft labels on different training subsets. Similar to the settings used in Tab. 2, we apply weighted soft labels on two different subsets: the samples facing tradeoff and the samples not facing tradeoff. In Tab. 5, the conflicting subset means the subset of samples facing tradeoff, which is judged by the metric $\text{sign}(ab) = -1$. From the significant improvements, we conclude that weighted soft labels can tackle the sample-wise bias-variance tradeoff.

Table 5: Performance on different subsets with soft labels and our weighted soft labels.

Subsets	Normal	Weighted
Conflicting	73.86	74.46
Not conflicting	74.59	75.15

Distillation with label smoothing trained teacher. Our exploration of bias-variance tradeoff starts with the conclusion made in Müller et al. (2019): a teacher network trained with the label smoothing trick is less effective for distillation. It is worthwhile to study whether the

Table 6: Distillation using weighted soft labels and teacher trained with label smoothing (denoted as *LS?*).

τ	*LS?*	Normal	Weighted
4	✗	75.06	75.11
4	✓	74.59	74.90
6	✗	75.10	75.28
6	✓	74.56	75.03

Table 3: Top-1 classification accuracy results on CIFAR-100. Comparison results are quoted from Tian et al. (2020). We report our results over 5 repeated runs.

	Same architecture style					Different architecture style		
Teacher Student	WRN-40-2 WRN-40-1	resnet56 resnet20	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	resnet32x4 ShuffleNetV1	resnet32x4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher	75.61	72.34	74.31	74.31	79.42	79.42	79.42	75.61
Student	71.98	69.06	69.06	71.14	72.50	70.5	71.82	70.5
FitNet	72.24	69.21	68.99	71.06	73.50	73.59	73.54	73.73
AT	72.77	70.55	70.22	72.31	73.44	71.73	72.73	73.32
SP	72.43	69.67	70.04	72.69	72.94	73.48	74.56	74.52
CC	72.21	69.63	69.48	71.48	72.97	71.14	71.29	71.38
VID	73.30	70.38	70.16	72.61	73.09	73.38	73.40	73.61
RKD	72.22	69.61	69.25	71.82	71.90	72.28	73.21	72.21
PKT	73.45	70.34	70.25	72.61	73.64	74.10	74.69	73.89
AB	72.38	69.47	69.53	70.98	73.17	73.55	74.31	73.34
FT	71.59	69.84	70.22	72.37	72.86	71.75	72.50	72.03
FSP	n/a	69.95	70.11	71.89	72.62	n/a	n/a	n/a
NST	72.24	69.60	69.53	71.96	73.30	74.12	74.68	74.89
KD	73.54	70.66	70.67	73.08	73.33	74.07	74.45	74.83
CRD	74.14	71.16	71.46	73.48	75.51	75.11	75.65	76.05
CRD+KD	74.38	71.63	71.56	73.75	75.46	75.12	76.05	76.27
Ours	**74.48**	**72.15**	**72.19**	**74.12**	**76.05**	**75.46**	**75.93**	**76.21**

Table 4: Top-1 and Top-5 classification accuracy results on ImageNet validation set. All training hyperparameter like learning rate and training epochs are in accordance with (Heo et al., 2019).

Teacher: ResNet-34 → Student: ResNet-18		
Method	Top-1 Acc	Top-5 Acc
Teacher	73.31	91.42
Student	69.75	89.07
KD	70.67	
AT	71.03	
NST	70.29	
FSP	70.58	
RKD	70.40	
Overhual	71.03	
CRD+KD	71.38	90.49
Ours	**72.04**	

Teacher: ResNet-50 → Student: MobileNet-v1		
Method	Top-1 Acc	Top-5 Acc
Teacher	76.16	
Student	68.87	
KD	70.49	
AT	70.18	
NST	n	
FSP	n	
RKD	68.50	
Overhual	71.33	
CRD+KD	TODO:	TODO:
Ours	**71.52**	

conclusion remains true for distillation with our weighted soft labels. As discussed before, we hold the opinion that more samples facing tradeoff make the distillation less effective. Since our weighted soft label is proposed to mitigate the negative effects of the tradeoff, we conduct comparison experiments in Tab. 6 to see if the negative effects still exist. It is evident that weighted soft labels significantly improve the distillation performance, especially for distillation from the teacher trained with label smoothing. Although the performance gap between with and without label smoothing is considerably narrowed, using the teacher trained with label smoothing performs worse than that without label smoothing, which again verifying the conclusion drawn by Müller et al. (2019).

6 Conclusion

Recent studies (Müller et al., 2019; Yuan et al., 2020) point out that one important reason behind the effectiveness of distillation is the regularization effect brought by being soft. In this paper, we rethink the soft labels for distillation from a bias-variance tradeoff perspective. The tradeoff exists sample-wise and we propose weighted soft labels to handle the tradeoff.

References

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9163–9171, 2019.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, pp. 2654–2662, 2014.

Anoop Korattikara Balan, Vivek Rathod, Kevin P Murphy, and Max Welling. Bayesian dark knowledge. In *Advances in Neural Information Processing Systems*, pp. 3438–3446, 2015.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Pedro Domingos. A unified bias-variance decomposition. In *International Conference on Machine Learning*, pp. 231–238, 2000.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1921–1930, 2019.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in neural information processing systems*, pp. 2760–2769, 2018.

Ron Kohavi and David H Wolpert. Bias plus variance decomposition for zero-one loss functions. In *International Conference on Machine Learning*, volume 96, pp. 275–83, 1996.

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

Ruofan Liang, Tianlin Li, Longfei Li, Jing Wang, and Quanshi Zhang. Knowledge consistency between neural networks and beyond. In *International Conference on Learning Representations*, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Charles Loop. Smooth subdivision surfaces based on triangles. *Master's thesis, University of Utah, Department of Mathematics*, 1987.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pp. 4694–4703, 2019.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 11615–11626. Curran Associates, Inc., 2019.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision*, pp. 268–284, 2018.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet:: Similarity-measuring the relatedness of concepts. In *AAAI Conference on Artificial Intelligence*, volume 4, pp. 25–29, 2004.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5007–5016, 2019.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pp. 5142–5151, 2019.

PlotOptiX. `https://pypi.org/project/plotoptix/`. (Accessed on 09/08/2020).

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1365–1374, 2019.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. *arXiv preprint arXiv:2002.11328*, 2020.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.