Revisit Knowledge Distillation: a Teacher-free framework

Li Yuan[1], Francis E.H.Tay[1], Guilin Li[2], Tao Wang[1], Jiashi Feng[1]
[1]National University of Singapore   [2]Huawei Noah's Ark Lab
{ylustcnus,twangnh}@gmail.com, {mpetayeh,elefjia}@nus.edu.sg, liguilin2@huawei.com

Abstract

Knowledge Distillation (KD) aims to distill the knowledge of a cumbersome teacher model into a lightweight student model. Its success is generally attributed to the privileged information on similarities among categories provided by the teacher model, and in this sense, only strong teacher models are deployed to teach weaker students in practice. In this work, we challenge this common belief by following experimental observations: 1) beyond the acknowledgment that the teacher can improve the student, the student can also enhance the teacher significantly by reversing the KD procedure; 2) a poorly-trained teacher with much lower accuracy than the student can still improve the latter significantly. To explain these observations, we provide a theoretical analysis of the relationships between KD and label smoothing regularization. We prove that 1) KD is a type of learned label smoothing regularization and 2) label smoothing regularization provides a virtual teacher model for KD. From these results, we argue that the success of KD is not fully due to the similarity information between categories, but also to the regularization of soft targets, which is equally or even more important.

Based on these analyses, we further propose a novel Teacher-free Knowledge Distillation (Tf-KD) framework, where a student model learns from itself or manually-designed regularization distribution. The Tf-KD achieves comparable performance with normal KD from a superior teacher, which is well applied when teacher model is unavailable. Meanwhile, Tf-KD is generic and can be directly deployed for training deep neural networks. Without any extra computation cost, Tf-KD achieves up to 0.65% improvement on ImageNet over well-established baseline models, which is superior to label smoothing regularization. The codes are in: `https://github.com/yuanli2333/Teacher-free-Knowledge-Distillation`

1 Introduction

Hinton et al. (2015) proposed Knowledge Distillation (KD) to transfer knowledge from one neural network (teacher) to another (student). Usually, the teacher model has strong learning capacity with higher performance, which teaches a lower-capacity student model through providing "soft targets". It is commonly believed that the soft targets of a teacher model can transfer "dark knowledge" containing privileged information on similarity among different categories (Hinton et al., 2015) to enhance the student model.

In this work, we first examine such a common belief through following exploratory experiments: 1) let student models teach teacher models by transferring soft targets of the students; (2) let poorly-trained teacher models with worse performance teach students. Based on the common belief, it is expected that the teacher model would not be enhanced significantly via training from the students and poorly-trained teachers would not enhance the students, as the weak student and poorly-trained teacher models cannot provide reliable similarity information between categories. However, after extensive experiments on various models and datasets, we surprisingly observe contradictory results: the weak student can improve the teacher and the poorly-trained teacher can also enhance the student remarkably. Such intriguing results motivate us to interpret KD as a regularization term, and we re-examine knowledge distillation from the perspective of Label Smoothing Regularization

(LSR) (Szegedy et al., 2016) that regularizes model training by replacing the one-hot labels with smoothed ones.

We then analyze theoretically the relationships between KD and LSR. For LSR, by splitting the smoothed label into two parts and examining the corresponding losses, we find the first part is the ordinary cross-entropy for ground-truth distribution (one-hot label) and outputs of model, and the second part, surprisingly, corresponds to a virtual teacher model which provides a uniform distribution to teach the model. For KD, by combining the teacher's soft targets with the one-hot ground-truth label, we find that KD is a learned LSR where the smoothing distribution of KD is from a teacher model but the smoothing distribution of LSR is manually designed. In a nutshell, we find *KD is a learned LSR and LSR is an ad-hoc KD*. Such relationships can explain the above counterintuitive results—the soft targets from weak student and poorly-trained teacher models can effectively regularize the model training, even though they lack strong similarity information between categories. We therefore argue that the similarity information between categories cannot fully explain the dark knowledge in KD, and the soft targets from the teacher model indeed provide effective regularization for the student model, which are equally or even more important.

Based on the analyses, we conjecture that with non-reliable or even zero similarity information between categories from the teacher model, KD may still well improve the student models. We thus propose a novel Teacher-free Knowledge Distillation (Tf-KD) framework with two implementations. The first one is to train the student model by itself (i.e., self-training), and the second is to manually design a targets distribution as a virtual teacher model which has 100% accuracy. The first method is motivated by replacing the dark knowledge with predictions from the model itself, and the second method is inspired by the relationships between KD and LSR. We validate through extensive experiments that the two implementations of Tf-KD are both simple yet effective. Particularly, in the second implementation without similarity information in the virtual teacher, Tf-KD still achieves comparable performance with normal KD, which clearly justifies:

> *dark knowledge does not just include the similarity between categories, but also imposes regularization on the student training*.

Tf-KD well applies to scenarios where the student model is too strong to find teacher models or computational resource is limited for training teacher models. For example, if we take a cumbersome single model ResNeXt101-32×8d (Xie et al., 2017) as the student model (with 88.79M parameters and 16.51G FLOPs on ImageNet), it is hard or computationally expensive to train a stronger teacher model. We deploy our virtual teacher to teach this powerful student and achieve 0.48% improvement on ImageNet without any extra computation cost. Similarly, when taking a powerful single model ResNeXt29-8×64d with 34.53M parameters as a student model, our self-training implementation achieves more than 1.0% improvement on CIFAR100 (from 81.03% to 82.08%).

Our contributions are summarized as follows:

- By designing two exploratory experiments, we find knowledge distillation can be interpreted as a regularization method. We then provide theoretical analysis to reveal the relation between knowledge distillation and label smoothing regularization.
- We propose Teacher-free Knowledge Distillation (Tf-KD) by self-training or using a well-designed virtual teacher model. Tf-KD achieves comparable performance with normal knowledge distillation that has a superior teacher.
- Tf-KD can be deployed as a generic regularization to train deep neural networks, which achieves superior performance to label smoothing regularization on ImageNet-2012.

2 EXPLORATORY EXPERIMENTS AND COUNTERINTUITIVE OBSERVATIONS

To examine the common belief on dark knowledge in KD, we conduct two exploratory experiments:

1) The standard knowledge distillation is to adopt a teacher to teach a weaker student. What if we reverse the operation? Based on the common belief, the teacher should not be improved significantly because the student is too weak to transfer effective knowledge.

2) If we use a poorly-trained teacher which has much worse performance than the student to teach the student, it is assumed to bring no improvement to the latter. For example, if a

poorly-trained teacher with only 10% accuracy is adopted in an image classification task, the student would learn from its soft targets with 90% error, thus the student should not be improved or even suffer worse performance.
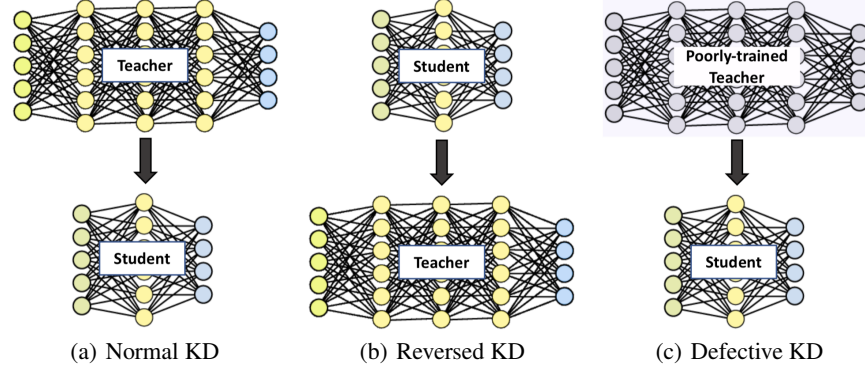

Figure 1: (a) Normal KD framework. (b)(c) Diagrams of exploratory experiments we conduct.

We name the "student teach teacher" as Reversed Knowledge Distillation (Re-KD), and the "poorly-trained teacher teach student" as Defective Knowledge Distillation (De-KD) (Fig. 1). We conduct Re-KD and De-KD experiments on CIFAR10, CIFAR100 and Tiny-ImageNet datasets with a variety of neural networks. For fair comparisons, all experiments are conducted with the same settings. Detailed implementation and experiment settings are given in Appendix A.1.

2.1 REVERSED KNOWLEDGE DISTILLATION

We conduct Re-KD experiments on the three datasets respectively. CIFAR10 and CI-FAR100 (Krizhevsky et al., 2009) contain natural RGB images of 32x32 pixels with 10 and 100 classes, respectively, and Tiny-ImageNet[1] is a subset of ImageNet (Deng et al., 2009) with 200 classes, where each image is down-sized to 64x64 pixels. For generality of the experiments, we adopt 5-layer plain CNN, MobilenetV2 (Sandler et al., 2018) and ShufflenetV2 (Ma et al., 2018) as student models and ResNet18, ResNet50 (He et al., 2016), DenseNet121 (Huang et al., 2017) and ResNeXt29-8×64d as teachers. The results of Re-KD on the three datasets are given in Tabs. 1 to 3.

In Tab. 1, the teacher models are improved significantly by learning from students, especially for teacher models ResNet18 and ResNet50. The two teachers obtain more than 1.1% improvement when taught by MobileNetV2 and ShuffleNetV2. We can also observe similar results on CIFAR10 and Tiny-ImageNet. When comparing Re-KD (S→T) with Normal KD (T→S), we can see in most cases, Normal KD achieves better results. It should be noted that Re-KD takes the teacher' accuracy as the baseline accuracy, which is much higher than that of Normal KD. However, in some cases, we can find Re-KD outperforms Normal KD. For instance, in Tab. 2 (3rd row), the student model (plain CNN) can only be improved by 0.31% when taught by MobileNetV2, but the teacher (MobileNetV2) can be improved by 0.92% by learning from the student. We have similar observations for ResNeXt29 and ResNet18 (4th row in Tab. 2).

We claim that while the standard knowledge distillation can improve the performance of students on all datasets, the superior teacher can also be enhanced significantly by learning from a weak student, as suggested through the Re-KD experiments.

2.2 DEFECTIVE KNOWLEDGE DISTILLATION

We conduct De-KD on CIFAR100 and Tiny-ImageNet. We adopt MobileNetV2 and ShuffleNetV2 as student models and ResNet18, ResNet50 and ResNeXt29 (8×64d) as teacher models. The poorly-trained teachers are trained by 1 epoch (ResNet18) or 50 epochs (ResNet50 and ResNeXt29), with

---

[1]https://tiny-imagenet.herokuapp.com/

3

Table 1: Normal KD and Re-KD experiment results on CIFAR100. We report mean±std (in %) over 3 runs. The number in parenthesis means increased accuracy over baseline (T: teacher, S: student).

Teacher: baseline	Student: baseline	Normal KD (T→S)	Re-KD (S→T)
ResNet18: 75.87	MobileNetV2: 68.38	71.05±0.16 (**+2.67**)	77.28±0.28 (**+1.41**)
	ShuffleNetV2: 70.34	72.05±0.13 (**+1.71**)	77.35±0.32 (**+1.48**)
ResNet50: 78.16	MobileNetV2: 68.38	71.04±0.20 (**+2.66**)	79.30±0.11 (**+1.14**)
	ShuffleNetV2: 70.34	72.15±0.18 (**+1.81**)	79.43±0.39 (**+1.27**)
DenseNet121: 79.04	MobileNetV2: 68.38	71.29±0.23 (**+2.91**)	79.55±0.11 (**+0.51**)
	ShuffleNetV2: 70.34	72.32±0.25 (**+1.98**)	79.83±0.05 (**+0.79**)
ResNeXt29: 81.03	MobileNetV2: 68.38	71.65±0.41 (**+3.27**)	81.53±0.14 (**+0.50**)
	ResNet18: 75.87	77.84±0.15 (**+1.97**)	81.62±0.22 (**+0.59**)

Table 2: Re-KD experiment results (accuracy, mean±std over 3 runs in %) on CIFAR10.

Teacher: baseline	Student: baseline	Normal KD (T→S)	Re-KD (S→T)
ResNet18: 95.12	Plain CNN: 87.14	87.67±0.17 (**+0.53**)	95.33±0.12 (**+0.21**)
	MobileNetV2: 90.98	91.69±0.14 (**+0.71**)	95.71±0.11 (**+0.59**)
MobileNetV2: 90.98	Plain CNN: 87.14	87.45±0.18 (**+0.31**)	91.81±0.23 (**+0.92**)
ResNeXt29: 95.76	ResNet18: 95.12	95.80±0.13 (**+0.68**)	96.49±0.15 (**+0.73**)

Table 3: Re-KD experiment results (accuracy, in %) on Tiny-ImageNet.

Teacher: baseline	Student: baseline	Normal KD (T→S)	Re-KD (S→T)
ResNet18: 63.44	MobileNetV2: 55.06	56.70 (**+1.64**)	64.12 (**+0.68**)
	ShuffleNetV2: 60.51	61.19 (**+0.68**)	64.35 (**+0.91**)
ResNet50: 67.47	MobileNetV2: 55.06	56.02 (**+0.96**)	67.68 (**+0.21**)
	ShuffleNetV2: 60.51	60.79 (**+0.28**)	67.62 (**+0.15**)
	ResNet18: 63.44	64.23 (**+0.79**)	67.89 (**+0.42**)

very poor performance. For example, ResNet18 only obtains 15.48% accuracy on CIFAR100 and 9.41% accuracy on Tiny-ImageNet after trained with 1 epoch, and ResNet50 obtains 45.82% and 31.01% on CIFAR100 and Tiny-ImageNet, after trained with 50 epochs (200 epochs in total).

From De-KD experiment results on CIFAR100 in Tab. 4, we observe that the student can be greatly promoted even when distilled by a poorly-trained teacher. For instance, the MobileNetV2 and ShuffleNetV2 can be promoted by 2.27% and 1.48% when taught by the one-epoch-trained ResNet18 with only 15.48% accuracy (2nd row). For poorly-trained ResNeXt29 with 51.94% accuracy (4th row), we find ResNet18 can still be improved by 1.41%, and MobileNetV2 obtains 3.14% improvement. From the De-KD experiment results on Tiny-ImageNet in Tab. 4, we find ResNet18 with 9.14% accuracy can still enhance the teacher model MobileNetV2 by 1.16%. Other poorly-trained teachers are all able to enhance the students to some degree.

To better demonstrate the distillation accuracy of a student when taught by poorly-trained teachers with different levels of accuracy, we save 9 checkpoints of ResNet18 and ResNeXt29 in the normal training process. Taking these checkpoints as teacher models to teach MobileNetV2, we observe that MobileNetV2 can always be improved by poorly-trained ResNet18 or poorly-trained ResNeXt29 with different levels of accuracy (Fig. 2). So we can say while a poorly-trained teacher provides much more noisy logits to the student, the student can still be enhanced. The De-KD experiment results are also conflicted with the common belief.

The counterintuitive results of Re-KD and De-KD make us rethink the "dark knowledge" in KD, and we argue that it does not just contain the similarity information. Lacking enough similarity

Table 4: De-KD accuracy (in %) on two datasets. Pt-Teacher is "Poorly-trained Teacher". Refer to the "Normal KD" in Tabs. 1 to 3 for the accuracy of students taught by "fully-trained teacher".

Dataset	Pt-Teacher: baseline	Student: baseline	De-KD
CIFAR100	ResNet18: 15.48	MobileNetV2: 68.38 ShuffleNetV2: 70.34	70.65±0.35 (**+2.27**) 71.82±0.11 (**+1.48**)
	ResNet50: 45.82	MobileNetV2: 68.38 ShuffleNetV2: 70.34 ResNet18: 75.87	71.45±0.23 (**+3.09**) 72.11±0.09 (**+1.77**) 77.23±0.11 (**+1.23**)
	ResNeXt29: 51.94	MobileNetV2: 68.38 ShuffleNetV2:70.34 ResNet18: 75.87	71.52±0.27 (**+3.14**) 72.26±0.36 (**+1.92**) 77.28±0.17 (**+1.41**)
Tiny-ImageNet	ResNet18: 9.41	MobileNetV2: 55.06 ShuffleNetV2: 60.51	56.22 (**+1.16**) 60.66 (**+0.15**)
	ResNet50: 31.01	MobileNetV2:55.06 ShuffleNetV2: 60.51	56.02 (**+0.96**) 61.09 (**+0.58**)
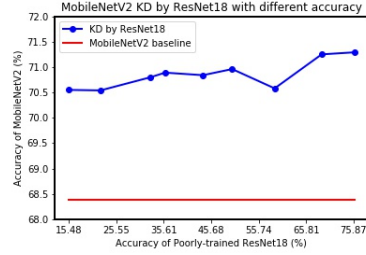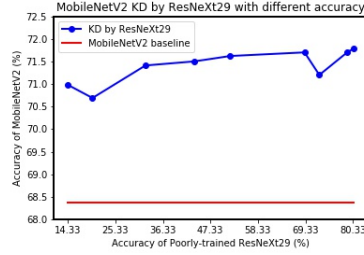
(a) ResNet18   (b) ResNeXt29

Figure 2: MobileNetV2 taught by ResNet18 and ResNeXt29 with different accuracy on CIFAR100. MobileNetV2 is enhanced by different poorly-trained teachers compared with baseline (the red line). The final point of two blue lines is the result taught by "fully-trained teacher".

information, a model can still provide "dark knowledge" to enhance other models. To explain this, we make a reasonable assumption and view knowledge distillation as a model regularization, and investigate what is the additional information in the "dark knowledge" of a model. In the next, we will analyze the relationships between knowledge distillation and label smoothing regularization to explain the experiment results of Re-KD and De-KD.

3 KNOWLEDGE DISTILLATION AND LABEL SMOOTHING REGULARIZATION

We mathematically analyze the relationships between Knowledge Distillation (KD) and Label Smoothing Regularization (LSR), hoping to explain the intriguing results of exploratory experiments in Sec. 2. Given a neural network S to train, we first give loss function of LSR for S. For each training example x, S outputs the probability of each label $k \in \{1...K\}$: $p(k|x) = \text{softmax}(z_k) = \frac{\exp(z_k)}{\sum_{i=1}^{K} \exp(z_i)}$, where z_i is the logit of the neural network S. The ground truth distribution over the labels is $q(k|x)$. We write $p(k|x)$ as $p(k)$ and $q(k|x)$ as $q(k)$ for simplicity. The model S can be trained by minimizing the cross-entropy loss: $H(q,p) = -\sum_{k=1}^{K} q(k)\log(p(k))$. For a single ground-truth label y, the $q(y|x) = 1$ and $q(k|x) = 0$ for all $k \neq y$.

In LSR, it minimizes the cross-entropy between modified label distribution $q'(k)$ and the network output $p(k)$, where $q'(k)$ is the smoothed label distribution formulated as

$$q'(k) = (1-\alpha)q(k) + \alpha u(k), \tag{1}$$

5

which is a mixture of $q(k)$ and a fixed distribution $u(k)$, with weight α. Usually, the $u(k)$ is uniform distribution as $u(k) = 1/K$. The cross-entropy loss $H(q', p)$ defined over the smoothed labels is

$$H(q', p) = -\sum_{k=1}^{K} q'(k) \log p(k) = (1 - \alpha)H(q, p) + \alpha H(u, p) \tag{2}$$
$$= (1 - \alpha)H(q, p) + \alpha(D_{KL}(u, p) + H(u)),$$

where D_{KL} is the Kullback-Leibler divergence (KL divergence) and $H(u)$ denotes the entropy of u and is a constant for the fixed uniform distribution $u(k)$. Thus, the loss function of label smoothing to model S can be written as

$$\mathcal{L}_{LS} = (1 - \alpha)H(q, p) + \alpha D_{KL}(u, p). \tag{3}$$

For knowledge distillation, the teacher-student learning mechanism is applied to improve the performance of the student. We assume the student is the model S with output prediction $p(k)$, and the output prediction of the teacher network is $p_\tau^t(k) = \text{softmax}(z_k^t) = \frac{\exp(z_k^t/\tau)}{\sum_{i=1}^{K} \exp(z_i^t/\tau)}$, where z^t is the output logits of the teacher network and τ is the temperature to soften $p^t(k)$ (written as $p_\tau^t(k)$ after softened). The idea behind knowledge distillation is to let the student (the model S) mimic the teacher by minimizing the cross-entropy loss and KL divergence between the predictions of student and teacher as

$$\mathcal{L}_{KD} = (1 - \alpha)H(q, p) + \alpha D_{KL}(p_\tau^t, p_\tau). \tag{4}$$

Comparing Eq. (3) and Eq. (4), we find the two loss functions have a similar form. The only difference is that the $p_\tau^t(k)$ in $D_{KL}(p_\tau^t, p_\tau)$ is a distribution from a teacher model and $u(k)$ in $D_{KL}(u, p)$ is the pre-defined uniform distribution. From this view, we can consider KD as a special case of LSR where the smoothing distribution is learned but not pre-defined. On the other hand, if we view the regularization term $D_{KL}(u, p)$ as a virtual teacher model of knowledge distillation, this teacher model will give a uniform probability to all classes, meaning it has a random accuracy (1% accuracy for CIFAR100, 0.1% accuracy for ImageNet).

Since $D_{KL}(p_\tau^t, p_\tau) = H(p_\tau^t, p_\tau) - H(p_\tau^t)$, where the entropy $H(p_\tau^t)$ is constant for a fixed teacher model, we can reformulate Eq. (4) to

$$L_{KD} = (1 - \alpha)H(q, p) + \alpha(D_{KL}(p_\tau^t, p_\tau) + H(p_\tau^t)) = (1 - \alpha)H(q, p) + \alpha H(p_\tau^t, p_\tau). \tag{5}$$

If we set the temperature $\tau = 1$, we have $L_{KD} = H(\tilde{q}^t, p)$, where $\tilde{q}^t$ is

$$\tilde{q}^t(k) = (1 - \alpha)q(k) + \alpha p^t(k). \tag{6}$$

If we compare Eq. (6) with Eq. (1), it is more clearly seen that KD is a special case of LSR. Moreover, the distribution $p^t(k)$ is a learned distribution (from a trained teacher) instead of a uniform distribution $u(k)$. We visualize the output probability $p^t(k)$ of a teacher and compare it with label smoothing in Appendix A.2, and find with higher temperature τ, the $p^t(k)$ is more similar with the uniform distribution $u(k)$ of label smoothing.

Based on the comparison of the two loss functions, we summarize the relationships between knowledge distillation and label smoothing regularization as follows:

- Knowledge distillation is a learned label smoothing regularization, which has a similar function with the latter, i.e. regularizing the classifier layer of the model.

- Label smoothing is an ad-hoc knowledge distillation, which can be revisited as a teacher model with random accuracy and temperature $\tau = 1$.

- With higher temperature, the distribution of teacher's soft targets in knowledge distillation is more similar to the uniform distribution of label smoothing.

Therefore, the experiment results of Re-KD and De-KD can be explained as the soft targets of the model in high temperature are closer to a uniform distribution of label smoothing, where the learned soft targets can provide model regularization for the teacher model. That is why a student can enhance the teacher and a poorly-trained teacher can still improve the student model.

4 TEACHER-FREE KNOWLEDGE DISTILLATION

As we above analyzed, the "dark knowledge" in a teacher model is more of a regularization term than the similarity information between categories. Intuitively, we consider replacing the output distribution of the teacher model with a simple one. We therefore propose a novel Teacher-free Knowledge Distillation (Tf-KD) framework, which uses no teacher model. Tf-KD is especially applicable to cases where the teacher model is not available, or only limited computation resources are provided. We propose two implementations of our Tf-KD.

The first Tf-KD method is self-training knowledge distillation, denoted as Tf-KD$_{self}$. As afore-mentioned, the teacher can be taught by a student and a poorly-trained teacher can also enhance the student. Hence when the teacher model is not available, we propose to deploy "self-training". Specifically, we first train the student model in the normal way to obtain a pre-trained model, which is then used as the teacher to train itself by transferring soft targets as in Eq. (4). Formally, given a model S, we denote its pre-trained model as teacher S^t; then we try to minimize the KL divergence of the logits between S and S^t by Tf-KD$_{self}$. The loss function of Tf-KD$_{self}$ to train model S is

$$L_{self} = (1 - \alpha)H(q, p) + \alpha D_{KL}(p_\tau^t, p_\tau), \tag{7}$$

where p, p_τ^t are the output probability of S and S^t respectively, τ is the temperature and α is the weight to balance the two terms.

The second implementation of our Tf-KD method is to manually design a teacher with 100% accuracy. In Sec. 3, we reveal LSR is a virtual teacher model with random accuracy. So, if we design a teacher with higher accuracy, we can assume it would bring more improvement to the student. We propose to combine KD and LSR to build a simple teacher model which will output distribution for classes as the following:

$$p^d(k) = \begin{cases} a & \text{if } k = c, \\ (1 - a)/(K - 1) & \text{if } k \neq c, \end{cases} \tag{8}$$

where K is the total number of classes, c is the correct label and a is the correct probability for the correct class. We always set $a \geq 0.9$, so the probability of a correct class is much higher than that of an incorrect one, and the manually-designed teacher model has 100% accuracy for any dataset. We name this method as Teacher-free KD by manually-designed regularization, denoted as Tf-KD$_{reg}$. The loss function is

$$L_{reg} = (1 - \alpha)H(q, p) + \alpha D_{KL}(p_\tau^d, p_\tau), \tag{9}$$

where τ is the temperature to soften the manually-designed distribution p^d (as p_τ^d after softening). We set a high temperature $\tau \geq 20$ to make this virtual teacher output a soft probability, in which way it gains the smoothing property as LSR. We visualize the distribution of the manually designed teacher in Fig. 3. As Fig. 3 shows, this manually designed teacher model outputs soft targets with 100% classification accuracy, and also has the smoothing property of label smoothing.

Figure 3: Distribution of manually designed teacher (softened by $\tau = 20$) on 10-class dataset. C6 is the correct label. As a comparison, the orange bar is the uniform distribution of LSR.

The two Teacher-free methods, Tf-KD$_{self}$ and Tf-KD$_{reg}$, are very simple yet effective, as validated via extensive experiments in the next section.

5 EXPERIMENTS ON TEACHER-FREE KNOWLEDGE DISTILLATION

In this section, we conduct experiments to evaluate Tf-KD$_{self}$ and Tf-KD$_{reg}$ on three datasets for image classification: CIFAR100, Tiny-ImageNet and ImageNet. For fair comparisons, all experiments are conducted with the same setting. Codes will be publicly available.

5.1 EXPERIMENTS FOR SELF-TRAINING KNOWLEDGE DISTILLATION

The hyper-parameters (temperature τ and α) for Tf-KD$_{self}$ are given in Appendix A.3.

**CIFAR100.** On CIFAR100, we use baseline models including MobileNetV2, ShuffleNetV2, GoogLeNet, ResNet18, DenseNet121 and ResNeXt29(8×64d). The baselines are trained for 200 epochs, with batch size 128. The initial learning rate is 0.1 and then divided by 5 at the 60th, 120th, 160th epoch. We use SGD optimizer with momentum of 0.9, and weight decay is set to 5e-4. For hyper-parameters, we use grid search to find the best value.

Fig. 4 (a) shows the test accuracy of the six models. It can be seen that our Tf-KD$_{self}$ consistently outperforms the baselines. For example, as a powerful model with 34.52M parameters, ResNeXt29 improves itself by 1.05% with self-training (Fig. 4(b)). Even when compared to Normal KD with a superior teacher in Fig. 5(a) and Fig. 5(b), our method achieves comparable performance (experiment settings for Tf-KD and Normal KD are the same). For example, with ResNet50 to teach ReseNet18, the student has a 1.19% improvement, but our method achieves 1.23% improvement without using any stronger teacher model.

Model	Baseline	Tf-KD$_{self}$
MobileNetV2	68.38	70.96 (**+2.58**)
ShuffleNetV2	70.34	72.23 (**+1.89**)
ResNet18	75.87	77.10 (**+1.23**)
GoogLeNet	78.72	80.17 (**+1.45**)
DenseNet121	79.04	80.26 (**+1.22**)
ResNeXt29 (8x64d)	81.03	82.08 (**+1.05**)

(a) Test accuracy (in %) on CIFAR100. We run 3 times and report the mean of the best results.
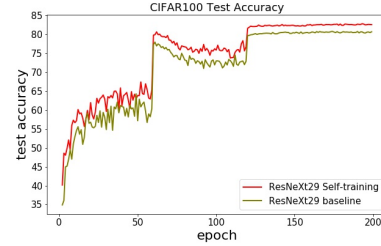

(b) ResNeXt29 on CIFAR100.

Figure 4: Classification accuracy by using Tf-KD$_{self}$. (a) Tf-KD$_{self}$ improves all six models. (b) Improvement on ResNeXt29 trained by itself (Tf-KD$_{self}$) is better than the normal training.

Model	Tf-KD$_{self}$	Normal KD [Teacher]
MobileNetV2	**+2.58**	**+2.67** [ResNet18]
ShuffleNetV2	**+1.89**	**+1.71** [ResNet18]
ResNet18	**+1.23**	**+1.19** [ResNet50]
GoogLeNet	**+1.45**	**+1.39** [ResNeXt29]
DenseNet121	**+1.22**	**+1.15** [ResNeXt29]
ResNeXt29	**+1.05**	**+1.12** [ResNeXt101]

(a) Accuracy improvement comparison (in %) on CIFAR100.
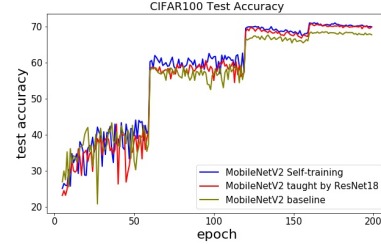
(b) MobileNetV2 on CIFAR100.

Figure 5: Comparison of our Tf-KD$_{self}$ with Normal KD. (a) Tf-KD$_{self}$ achieves comparable results to Normal KD. (b) MobileNetV2 obtains similar improvement by self-training or by ResNet18.

Tiny-ImageNet. On Tiny-ImageNet, we use baseline models including MobileNetV2, ShuffleNetV2, ResNet50, DenseNet121. They are trained for 200 epochs with batch size $bn = 128$ for MobileNetV2, ShuffleNetV2 and $bn = 64$ for ResNet50, DenseNet121. The initial learning rate is $\eta = 0.1 * \frac{bn}{128}$ and then divided by 10 at the 60th, 120th, 160th epoch. We use SGD optimizer with momentum of 0.9, and weight decay is set to 5e-4. Tab. 5 shows the results of Tf-KD$_{self}$ on Tiny-ImageNet. It can be seen that Tf-KD$_{self}$ consistently improves the baseline models and achieves comparable improvement with Normal KD.

ImageNet. ImageNet-2012 is one of the largest datasets for object classification, with over 1.3m hand-annotated images. The baseline models we use on this dataset include ResNet18, ResNet50,

Table 5: Tf-KD$_{self}$ experiment results on Tiny-ImageNet (in %).

Model	Baseline	Tf-KD$_{self}$	Normal KD [Teacher]
MobileNetV2	55.06	56.77 (**+1.71**)	56.70 (**+1.64**) [ResNet18]
ShuffleNetV2	60.51	61.36 (**+0.85**)	61.19 (**+0.68**) [ResNet18]
ResNet50	67.47	68.18 (**+0.71**)	68.23 (**+0.76**) [DenseNet121]
DenseNet121	68.15	68.29 (**+0.14**)	68.31 (**+0.16**) [ResNeXt29]

DenseNet121, RexNeXt101 (32x8d), and we adopt official implementation[2] of Pytorch to train them. We set batch size $bn = 512$ for ResNet18, ResNet50, DenseNet121, and $bn = 256$ for RexNeXt101. Following common experiment settings (Goyal et al., 2017), the initial learning rate is $\eta = 0.1 * \frac{bn}{256}$ which is then divided by 10 at the 30th, 60th, 80th epoch in total 90 epochs. We use SGD optimizer with momentum of 0.9, and weight decay is 1e-4. Results are reported in Tab. 6. We can see that the self-training can further improve the baseline performance on ImageNet-2012. As a comparison, we also use DenseNet121 to teach ResNet18 on ImageNet, and ResNet18 obtains 0.56% improvement, which is comparable with our self-training implementation (Tab. 7).

Table 6: Tf-KD$_{self}$ experiment results on ImageNet (Top1 accuracy, in %).

Model	Baseline	Tf-KD$_{self}$
ResNet18	69.84	70.42 (**+0.58**)
ResNet50	75.77	76.41 (**+0.64**)
DenseNet121	75.28	75.72 (**+0.44**)
ResNeXt101	79.28	79.56 (**+0.28**)

Table 7: Comparison between Tf-KD$_{self}$ and Normal KD on ImageNet (Top1 accuracy, in %).

Model	Baseline	Tf-KD$_{self}$	Normal KD [Teacher]
ResNet18	69.84	70.42 (**+0.58**)	70.40 (**+0.56**) [DenseNet121]

5.2 EXPERIMENTS FOR KNOWLEDGE-DISTILLATION WITH MANUALLY-DESIGNED TEACHER

For all experiments of Tf-KD$_{reg}$, we adopt the same implementation settings with Tf-KD$_{self}$, except for using a virtual output distribution as a virtual teacher (Eq. (9)) here. For fair comparisons, experiment settings for Normal KD and Tf-KD$_{reg}$ are the same. See Appendix A.4 for hyper-parameters of Tf-KD$_{reg}$.

CIFAR100 and Tiny-ImageNet. For Tf-KD$_{reg}$ experiments on CIFAR100 and Tiny-ImageNet, we set the probability for correct classes as $a = 0.99$ (Eq. (8)). The temperature τ and α in Eq. (9) are different for different baseline models (see Appendix A.4). From Tab. 8 and Tab. 9, we can observe with no teacher used and just a regularization term added, Tf-KD$_{reg}$ achieves comparable performance with Normal KD on both CIFAR100 and Tiny-ImageNet.

Table 8: Tf-KD$_{reg}$ achieves comparable results with Normal KD on CIFAR100.

Model	Baseline	Tf-KD$_{reg}$	Normal KD [Teacher]	+ LSR
MobileNetV2	68.38	70.88 (**+2.50**)	71.05 (**+2.67**) [ResNet18]	69.32 (**+0.94**)
ShuffleNetV2	70.34	72.09 (**+1.75**)	72.05 (**+1.71**) [ResNet18]	70.83 (**+0.49**)
ResNet18	75.87	77.36 (**+1.49**)	77.19 (**+1.32**) [ResNet50]	77.26 (**+1.39**)
GoogLeNet	78.15	79.22 (**+1.07**)	78.84 (**+0.99**) [ResNeXt29]	79.07 (**+0.92**)

---

[2]https://github.com/pytorch/examples/tree/master/imagenet

Table 9: Tf-KD$_{reg}$ experiment results on Tiny-ImageNet.

Model	Baseline	Tf-KD$_{reg}$	Normal KD [Teacher]	+ LSR
MobileNetV2	55.06	56.47 (**+1.41**)	56.53 (**+1.47**) [ResNet18]	56.24 (**+1.18**)
ShuffleNetV2	60.51	60.93 (**+0.42**)	61.19 (**+0.68**) [ResNet18]	60.66 (**+0.11**)
ResNet50	67.47	67.92 (**+0.45**)	68.15 (**+0.68**) [ResNeXt29]	67.63 (**+0.16**)
DenseNet121	68.15	68.37 (**+0.18**)	68.44 (**+0.26**) [ResNeXt29]	68.19 (**+0.04**)

ImageNet. For the Tf-KD$_{reg}$ on ImageNet, we adopt temperature $\tau = 20$ as normal knowledge distillation, and $\alpha = 0.1$ as label smoothing regularization. The probability for correct classes in the manually-designed teacher is $a = 0.99$ (Eq. (9)). We test our Tf-KD$_{reg}$ with four baseline models: ResNet18, ResNet50, DenseNet121 and ResNeXt101 (32x8d). As a regularization term, the manually designed teacher achieves consistent improvement compared with baselines (Fig. 6 (a)). For example, the proposed Tf-KD$_{reg}$ improves the top1 accuracy of ResNet50 by 0.65% on ImageNet-2012 (Fig. 6 (b)). Even for a huge single model ResNeXt101 (32x8d) with 88.79M parameters, our method achieves 0.48% improvement by using the manually designed teacher.

Model	Baseline	+Tf-KD$_{reg}$	+ LSR
ResNet18	69.84	70.24 (**+0.40**)	70.02 (**+0.18**)
ResNet50	75.77	76.42 (**+0.65**)	76.38 (**+0.61**)
DenseNet121	75.28	75.62 (**+0.34**)	75.24 (**-0.04**)
ResNeXt101	79.28	79.76 (**+0.48**)	79.67 (**+0.39**)

(a) Test accuracy improvement (in %) on ImageNet.
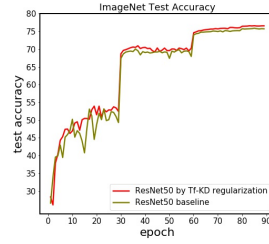
(b) ResNet50 on ImageNet.

Figure 6: Classification accuracy by using Tf-KD$_{reg}$ on ImageNet. (a) Top1 test accuracy on ImageNet. (b) Without extra computation, ResNet50 outperforms baseline model by 0.65%

Comparing our two methods Tf-KD$_{self}$ and Tf-KD$_{reg}$, we observe that Tf-KD$_{self}$ works better in small dataset (CIFAR100) while Tf-KD$_{reg}$ performs slightly better in large dataset (ImageNet).

5.3 TEACHER-FREE REGULARIZATION

For Tf-KD$_{reg}$, the manually designed teacher is a regularization term, meaning no extra computation added. So it can serve as a generic regularization method to normally train neural networks. We compare our Teacher-free regularization with label smoothing on CIFAR100, Tiny-ImageNet and ImageNet. For fair comparisons, experiment settings for Tf-KD$_{reg}$ and LSR are the same. The results are shown in Tab. 8, Tab. 9 and Figure 6(a). It can be seen that our Tf-KD$_{reg}$ regularization consistently outperforms label smoothing regularization, which is a better choice for model regularization compared with label smoothing.

6 RELATED WORK

Knowledge Distillation Since Hinton et al. (2015) proposed knowledge distillation, KD has been widely adopted or modified (Romero et al., 2014; Yim et al., 2017; Yu et al., 2017; Furlanello et al., 2018; Anil et al., 2018; Mirzadeh et al., 2019; Wang et al., 2019). Different from existing works, our work challenges the common belief of knowledge distillation based on our designed exploratory experiments. Another related work is deep mutual learning (Zhang et al., 2018), which proposes to let an ensemble of student models to learn with each other by minimizing the KL Divergence of predictions. Comparatively, our work reveals the relationship between KD and label smoothing, and our proposed Teacher-free regularization can serve as a general method for neural network training.

Label Smoothing Szegedy et al. (2016) proposed label smoothing regularization to replace the "hard labels" with smoothed labels, boosting performance of many tasks like image classification, language translation and speech recognition (Pereyra et al., 2017). Recently, Müller et al. (2019)

10

empirically showed label smoothing can also help improve model calibration. In our work, we adopt label smoothing regularization to understand the regularization function of knowledge distillation.

7 CONCLUSION AND FUTURE WORK

In this work, we find through experiments and analyses that the "dark knowledge" of a teacher model is more of a regularization term than similarity information of categories. Based on the relationship between KD and LSR, we propose Teacher-free KD. Experiment results show our Tf-KD can achieve comparable results with Normal KD in image classification. One limitation of this work is that we only test the proposed methods in image classification tasks. In the future, we will explore its application to video classification and natural language understanding. Our work also suggests that, when it is hard to find a good teacher for a powerful model or computation resource is limited for obtaining teacher models, the targeted model can still get enhanced by self-training or a manually-designed regularization term.

REFERENCES

Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*, 2018.

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

Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *arXiv preprint arXiv:1805.04770*, 2018.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

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

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, 2018.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. *arXiv preprint arXiv:1902.03393*, 2019.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 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.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

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.

Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1974–1982, 2017.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328, 2018.

A APPENDIX

A.1 IMPLEMENTATION DETAILS FOR EXPLORATORY EXPERIMENTS

In this appendix, we provide implementation details and experiment settings for the exploratory experiments. We conduct experiments based on the standard implementation of knowledge distillation (Hinton et al., 2015). The loss function for standard knowledge distillation is:

$$\mathcal{L}_{KD} = (1 - \alpha)H(q, p) + \alpha D_{KL}(p_\tau^t, p_\tau) \tag{10}$$

where q is the distribution of ground truth, p is the output distribution of student model, $H(,)$ is cross-entropy loss function and D_{KL} is KL divergence, and p_τ^t is the output distribution of teacher model soften by temperature τ. The temperature τ and weight α are hyper-parameters. The temperature τ and weight α for Reverse KD and Normal KD are given in Tab. 10; τ and α for De-KD are given in Tab. 11. For fair comparisons, Normal KD, Re-KD and De-KD are conducted with the same experiment settings.

CIFAR10 and CIFAR100 For exploratory experiments Re-KD and De-KD on CIFAR10 and CIFAR100, we train for 200 epochs, with batch size 128. For the Plain CNN, the The initial learning rate is 0.01 and then be divided by 5 at the 60'th, 120'th, 160'th epoch. For other used models (MobileNetV2, ShuffleNetV2, ResNet, ResNeXt, DenseNet), the initial learning rate is 0.1 and then be divided by 5 at the 60'th, 120'th, 160'th epoch. We use Adam optimizer for the Plain CNN and SGD optimizer with momentum 0.9 for other models, and the weight decay is set to be 5e-4. For hyper-parameters, τ (temperature) and α, we use grid search to find the best value.

The Plain CNN used in exploratory experiments is a 5-layer neural network with 3 convolutional layers and 2 fully-connected layers. On CIFAR10, the architecture of the Plain CNN is: $conv1$(3x3, 32, channels) $\rightarrow$ $batchnorm$ $\rightarrow$ $conv2$(3x3, 64, channels) $\rightarrow$ $batchnorm$ $\rightarrow$ $conv3$(3x3, 128, channels) $\rightarrow$ $batchnorm$ $\rightarrow$ fc(128) $\rightarrow$ fc(10).

Table 10: Parameters for Normal KD and Re-KD experiments (Temperature τ and α)

Dataset	Teacher	Student	Normal KD	Re-KD
CIFAR-10	ResNet18	Plain CNN MobileNetV2	τ=20, α=0.90 τ=20, α=0.90	τ=20, α=0.01 τ=20, α=0.05
	MobileNetV2	Plain CNN	τ=20, α=0.40	τ=20, α=0.10
	ResNeXt29	ResNet18	τ=6, α=0.95	τ=20, α=0.10
CIFAR100	ResNet18	MobileNetV2 ShuffleNetV2	τ=20, α=0.95 τ=20, α=0.95	τ=20, α=0.60 τ=20, α=0.60
	ResNet50:	MobileNetV2 ShuffleNetV2	τ=20, α=0.95 τ=20, α=0.95	τ=20, α=0.60 τ=20, α=0.60
	Densenet121	MobileNetV2 ShuffleNetV2	τ=20, α=0.95 τ=20, α=0.95	τ=20, α=0.60 τ=20, α=0.60
	ResNeXt29	MobileNetV2 ResNet18	τ=20, α=0.60 τ=20, α=0.60	τ=20, α=0.60 τ=20, α=0.60
T-ImageNet	ResNet18	MobileNetV2 ShuffleNetV2	τ=20, α=0.10 τ=20, α=0.10	τ=20, α=0.60 τ=20, α=0.60
	ResNet50	MobileNetV2 ShuffleNetV2 ResNet18	τ=20, α=0.10(bm) τ=20, α=0.10 τ=20, α=0.50	τ=20, α=0.10 τ=20, α=0.50 τ=20, α=0.10

Table 11: Parameters for De-KD (Temperature τ and α)

Dataset	Poorly-trained Teacher	Student	τ and α
CIFAR100	ResNet18: 15.48%	MobileNetV2 ShuffleNetV2	τ=20, α=0.95 τ=20, α=0.95
	ResNet50: 45.82%	MobileNetV2 ShuffleNetV2 ResNet18	τ=20, α=0.95 τ=20, α=0.95 τ=20, α=0.60
	ResNeXt29: 51.94%	MobileNetV2 ShuffleNetV2 ResNet18	τ=20, α=0.95 τ=20, α=0.95 τ=20, α=0.60
Tiny-ImageNet	ResNet18: 9.41%	MobileNetV2 ShuffleNetV2	τ=20, α=0.10 τ=20, α=0.10
	ResNet50: 31.01%	MobileNetV2 ShuffleNetV2	τ=20, α=0.10 τ=20, α=0.10

Tiny-ImageNet For exploratory experiments on Tiny-ImageNet, all models are trained for 200 epochs, with batch size $bn = 128$ for MobileNetV2, ShuffleNetV2, ResNet18 and $bn = 64$ for ResNet50, DenseNet121. The initial learning rate is $\eta = 0.1 * \frac{bn}{128}$ and then be divided by 10 at the 60'th, 120'th, 160'th epoch. We use SGD optimizer with momentum of 0.9, and the weight decay is set to be 5e-4.

Model complexity We provide the model complexity (size and FLOPs) of all models we used in this work in Tab. 12, which is the reference to choose teacher and student model. The model size is measured by the total number of learnable parameters within each model. The FLOPs of model is tested with image size of $3 \times 224 \times 224$.

A.2 VISUALIZATION OF THE OUTPUT DISTRIBUTION OF TEACHER

To better comparing the $p_\tau^t(k)$ (the output distribution of teacher model) and $u(k)$ (the uniform distribution of label smoothing), we visualize the soft targets of ResNet18 (trained on CIFAR10 with 95.12% accuracy) and compare the soft targets in different temperature with $u(k)$. As shown in Fig. 7, we can observe that with the temperature τ increasing, the two distributions become closer.

Table 12: Model complexity of all used models in this work.

Model	MobileNetV2	ShuffleNetV2	ResNet18	ResNet50
# param	3.50×10^6	2.28×10^6	11.69×10^6	25.56×10^6
FLOPs	0.32×10^9	0.15×10^9	1.82×10^9	4.12×10^9
Model	GoogLeNet	DenseNet121	ResNeXt29 (8x64d)	ResNeXt101 (32×8d)
# param	13.0×10^6	7.98×10^6	34.52×10^6	88.79×10^6
FLOPs	1.51×10^9	2.88×10^9	4.40×10^9	16.51×10^9

In the common experiments of knowledge distillation, we always adopt temperature as 20 (Hinton et al., 2015).
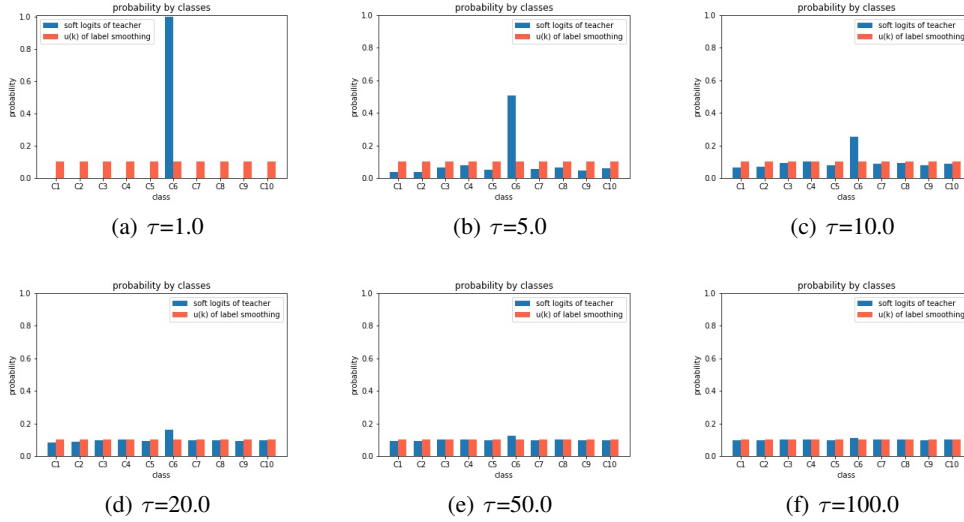

Figure 7: Comparison between label smoothing and soft targets of KD in different temperature τ. C6 is the correct label.

A.3 EXPERIMENT SETTINGS FOR TF_{self}

For all TF_{self} experiments on ImageNet, we set temperature τ=20, and weight α=0.10. The hyper-parameters on CIFAR100 and Tiny-ImageNet are given in Tab. 13.

Table 13: Hyper-parameters for TF_{self} experiments (Temperature τ and α)

Dataset	Model	TF_{self}
CIFAR100	MobileNetV2	τ=20, α=0.95
	ShuffleNetV2	τ=20, α=0.95
	ResNet18	τ=6, α=0.95
	GoogLeNet	τ=20, α=0.40
	DenseNet121	τ=20, α=0.95
	ResNeXt29 (8x64d)	τ=20, α=0.90
T-ImageNet	MobileNetV2	τ=20, α=0.10
	ShuffleNetV2	τ=20, α=0.10
	ResNet18	τ=6, α=0.10
	ResNet50	τ=20, α=0.10
	DenseNet121	τ=20, α=0.15

A.4 EXPERIMENT SETTINGS FOR TF$_{reg}$

For all TF$_{reg}$ experiments on ImageNet, we set temperature τ=20, and weight α=0.10. The hyper-parameters on CIFAR100 and Tiny-ImageNet are given in Tab. 14.

Table 14: Hyper-parameters for TF$_{reg}$ experiments (Temperature τ and α)

Dataset	Model	TF$_{reg}$
CIFAR100	MobileNetV2	τ=40, α=0.95
	ShuffleNetV2	τ=20, α=0.95
	ResNet18	τ=20, α=0.10
	GoogLeNet	τ=40, α=0.10
T-ImageNet	MobileNetV2	τ=20, α=0.10
	ShuffleNetV2	τ=20, α=0.10
	ResNet50	τ=20, α=0.10
	DenseNet121	τ=20, α=0.10