Revisiting Knowledge Distillation via Label Smoothing Regularization — Supplementary Material

1. Implementation details for exploratory experiments

In this section, we provide implementation details and experiment settings for the exploratory experiments. We conduct experiments based on the standard implementation of knowledge distillation [1]. The loss function for standard knowledge distillation is:

$$\mathcal{L}_{KD} = (1-\alpha)H(q,p) + \alpha D_{KL}(p^t_\tau, p_\tau) \qquad \text{(S.1)}$$

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^t_τ 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. 1; τ and α for De-KD are given in Tab. 2. For fair comparisons, the hyper-parameters of Normal KD, Re-KD and De-KD are grid searched from 70 training epochs (200 epochs in total), and 8 GPUs are used to search these hyper-paramwters (4 NVIDIA V, 4 NVIDIA X). All experiments 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) $\rightarrow$ $batchnorm$ $\rightarrow$ $fc(128)$ $\rightarrow$ $fc(10)$.

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. 3, 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 $3 \times 224 \times 224$.

2. Comparison between Tf$_{reg}$ with LSR

We first recall some important equantions for LSR and Tf$_{reg}$. 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

$$q'(k) = (1-\alpha)q(k) + \alpha u(k), \qquad \text{(S.2)}$$

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 loss function of LSR can be written as

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

For Tf$_{reg}$, we define a virtual teacher model which will output distribution for classes as the following:

$$p^d(k) = \begin{cases} a & \text{if } k = c, \\ (1-a)/(K-1) & \text{if } k \neq c, \end{cases} \qquad \text{(S.4)}$$

where K is the total number of classes, c is the correct label and a is the correct probability for the correct class. And the loss function is

$$L_{reg} = (1-\alpha)H(q,p) + \alpha D_{KL}(p^d_\tau, p_\tau), \qquad \text{(S.5)}$$

where τ is the temperature to soften the manually-designed distribution p^d (as p^d_τ after softening). We set a high temperature $\tau \geq 20$ to make this virtual teacher output a soft

Table 1. Parameters for Normal KD and Re-KD experiments (Temperature τ and α)

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

Table 2. Parameters for De-KD (Temperature τ and α)

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

Table 3. 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

probability, in which way it gains the smoothing property as LSR.

We compare Tf$_{reg}$ with LSR in two aspects.

(1). Some special parameters will make Tf$_{reg}$ be identical to LSR? For example, let $a = (1 - \alpha) + \alpha/K$, then Eq S.4 will be

$$p^d(k) = \begin{cases} 1 - a + \alpha/K & \text{if } k = c, \\ \alpha/k & \text{if } k \neq c, \end{cases} \tag{S.6}$$

which seems that this $p^d(k)$ is identical to the soft distribution of original LSR. However, it is still not a uniform distribution as LSR. After we split the loss function of LSR as two part, the distribution of LSR is a uniform distribution, so it is not the same as $p^d(k)$ even we let $a = (1-\alpha)+\alpha/K$.

(2). Check if Tf$_{reg}$ is an over-parameters version of LSR. If Tf$_{reg}$ is an over-parameters version of LSR, we can tune the parameters α of LSR to make it reach the same/similar performance as Tf$_{reg}$. When we tune α in LSR, we find that in most of the cases, the LSR will obtain the optimal performance as α is around 0.1, but which is still lower than our Tf$_{reg}$. We search α between [0.01, 1] with an interval of 0.02, and the search results of LSR on CIFAR100 are given in Tab 4. It can be observed that the best results of LSR are still worse than Tf$_{reg}$.

Table 4. Comparison between LSR and Tf-KD$_{reg}$ on CIFAR100.

Model	Baseline	Tf-KD$_{reg}$	+ LSR (optimal α)
MobileNetV2	68.38	70.88 (**+2.50**)	69.54 ($\alpha = 0.19$)
ShuffleNetV2	70.34	72.09 (**+1.75**)	70.98 ($\alpha = 0.23$)
ResNet18	75.87	77.36 (**+1.49**)	77.49 ($\alpha = 0.15$)

On the other hand, we cannot let Eq. S.5 be the same as Eq. S.3 because the output of student model (p_τ) also soften by τ. Same as normal KD, the temperature τ is always very high ($\tau \gg 1$), thus the distribution of Tf$_{reg}$ obtain the smoothing property. The temperature $\tau \neq 1$, so LSR and Tf-KD$_{reg}$ will not the same when we tune the α in LSR. Actually, in our analyze, even the original KD has a similar format LSR, so the Tf-KD will have a similar format with LSR because our Tf-KD is proposed based on the KD and LSR.

We summary that even Tf$_{reg}$ has a similar format with LSR, but it is not identical to LSR or an over-parameters version of LSR.

3. Visualization of the output distribution of teacher

To better comparing the $p^t_\tau(k)$ (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. S.1, we can observe that with the temperature τ increasing, the two distributions become closer. In the common experiments of knowledge distillation, we always adopt temperature as 20 [1].

4. Experiment settings for Tf$_{self}$ and Tf$_{reg}$

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. 5.

Table 5. 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

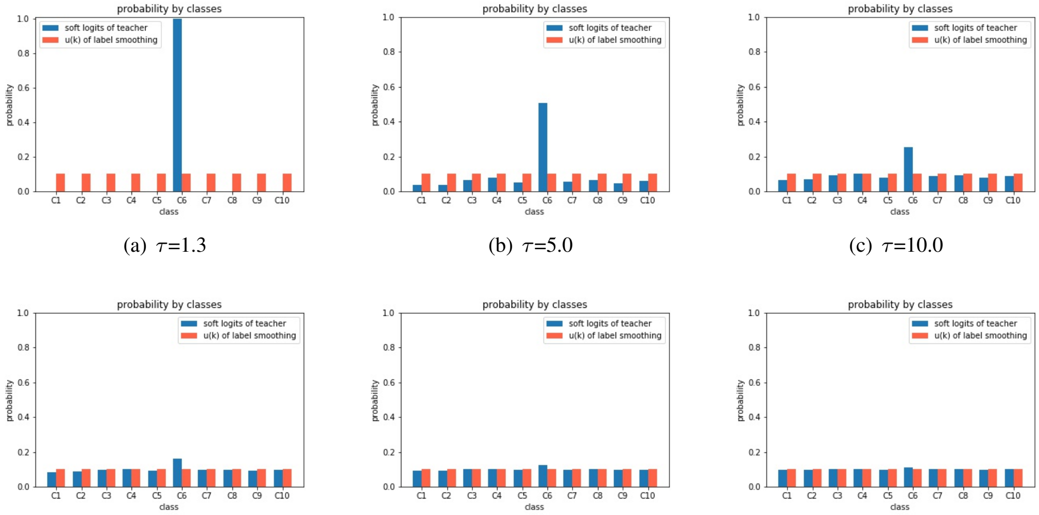
For all TF$_{reg}$ experiments on ImageNet, we set temperature τ=20, and weight α=0.10. The hyper-parameters on CIFAR100 and Tiny-ImageNet are given in Tab. 6.

Table 6. 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

References

[1] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 3



(a) τ=1.3

(b) τ=5.0

(c) τ=10.0

(d) τ=20.0

(e) τ=50.0

(f) τ=100.0

Figure S.1. Comparison between label smoothing and soft targets of KD in different temperature τ. C6 is the correct label.