Supplementary Materials of From Knowledge Distillation to Self-Knowledge Distillation: A Unified Approach with Normalized Loss and Customized Soft Labels

Zhendong Yang[1,2*] Ailing Zeng[2] Zhe Li[3] Tianke Zhang[1] Chun Yuan[1†] Yu Li[2†]
[1]Tsinghua Shenzhen International Graduate School [2]International Digital Economy Academy (IDEA)
[3]Institute of Automation, Chinese Academy of Sciences
{yangzd21,ztk21}@mails.tsinghua.edu.cn axel.li@outlook.com
yuanc@sz.tsinghua.edu.cn {zengailing, liyu}@idea.edu.cn

Appendix

A. Effects of the Temperature in NKD

The temperature λ in NKD is a hyper-parameter used to adjust the distribution of the teacher's soft non-target labels. NKD always applies $\lambda > 1$ to make the logit become smoother, which causes the logit contains more non-target distribution knowledge. The target output probability of the same model will get a higher value on an easy dataset, such as CIFAR-100. This causes T_i^{λ} to contain less knowledge, which may bring adverse effects to the distillation. In this subsection, we explore the impact of using different temperatures to distill the student ResNet18 on CIFAR-100, shown in Tab. 1. The results show that temperature is an important hyper-parameter, especially for an easy dataset. For all the experiments, we adopt $\lambda = 1$ on ImageNet. While for CIFAR-100, we follow the training setting from DKD for a fair comparison.

λ	*Baseline*	1.0	2.0	3.0	4.0
Top-1	*78.58*	80.55	**80.76**	80.72	80.54
Top-5	*94.10*	**95.14**	**95.14**	95.11	95.05

Table 1. NKD's results on the CIFAR-100 with different temperature. We use ResNet-34 as the teacher to distill ResNet-18.

B. NKD for ViT-liked Models

To further evaluate the effectiveness of our NKD, we apply them to a vision transformer model DeiT, as shown in Tab. 2. We conduct experiments on the tiny and base DeiT with NKD, bringing them excellent improvements. DeiT-Base achieves 84.96% Top-1 accuracy with NKD, which is 3.20% higher than the baseline. Besides, NKD also outperforms the classical KD for ViT's distillation, which proves the effectiveness of our modification of KD's formula again.

C. Square for Soft Target Label

For the target loss L_{target}, we first square the student's target output and then smooth it as the soft target label. Here we explore the effects of the operation. We conduct experiments by training MobileNet and RegNetX-1.6GF on the ImageNet dataset, which is shown in Tab. 3. The square enlarges the difference between different samples' S_t in a training batch and brings more improvements to the model with self-knowledge distillation. Specifically, the model MobileNet achieves 70.18% top-1 accuracy with our target distillation loss L_{target}. While the model's top-1 accuracy without square is 70.04%. The results demonstrate the effectiveness of the square for the soft target label.

D. Sensitivity Study of USKD's Parameters

In our proposed method USKD, we use two hyper-parameters α and β to balance the target loss $\mathcal{L}_{target}$ and the non-target loss $\mathcal{L}_{non}$, respectively. To explore the effects of the two hyper-parameters for self-KD, we conduct experiments by training ResNet-18 with our method on the ImageNet dataset. As shown in Fig. 1, α is used to adjust the target loss scale. Our method is not sensitive to it. When α varies from 0.6 to 1.4, the student's worst accuracy is 70.68%, which is just 0.13% lower than the highest accuracy. Besides, it is still 0.78% higher than training the model directly. As for the β for the non-target loss in Fig. 2, our method is not sensitive to it when $\beta < 0.1$. However, when it is too large, *e.g.* 0.14, the performance improvement may be affected.

*This work was done when Zhendong was an intern at IDEA.
†Corresponding authors.

Teacher	Student	Top-1 Acc. (%)
DeiT III-Small (82.76)	DeiT-Tiny	*74.42*
	KD	76.01 (+1.59)
	NKD (Ours)	**76.68** (+2.26)
DeiT III-Large (86.81)	DeiT-Base	*81.76*
	KD	84.06 (+2.30)
	NKD (Ours)	**84.96** (+3.20)

Table 2. Comparison of NKD and KD on DeiT on ImageNet-1k. The teacher is pre-trained on ImageNet-21K.

	MobileNet	RegNetX-1.6GF
Baseline	*69.21*	*76.84*
w/o square	70.04	76.79
with square	70.18	76.87

Table 3. Comparison of training the models with different distillation methods on the target class. The experiments are conducted on the ImageNet dataset.

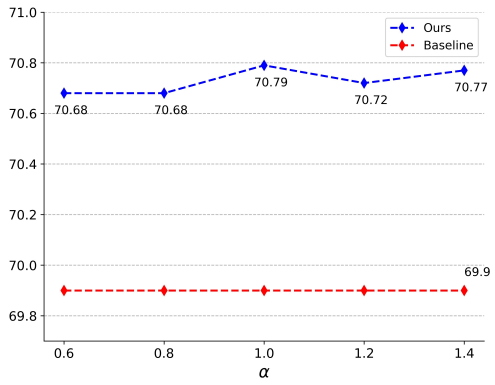

Figure 1. The hyper-parameter α for target loss $\mathcal{L}_{target}$ with ResNet-18 on ImageNet dataset.

E. Weak Supervision for the Weak Logit

For soft non-target labels' rank, we use a hyper-parameter $\mu < 1$ to achieve weak supervision for the weak logit. Here we conduct experiments to investigate the influence of weak supervision on self-KD. For normal supervision, μ should be set to 1. However, the rank of the weak logit is similar to that of the final logit when $\mu = 1$. With a smaller μ, the supervision becomes weaker, and the difference between the two logits becomes larger. As shown in Fig. 3, the model's top-1 accuracy is 70.18% when $\mu = 0.02$. The RegNetX-1.6GF model achieves better performance with weaker supervision when $\mu > 0.005$. However, when μ is too small, for example, 0.002, the supervision is too weak, resulting in the model's weak logits being the same for all non-target classes, which negatively affects performance improvement.

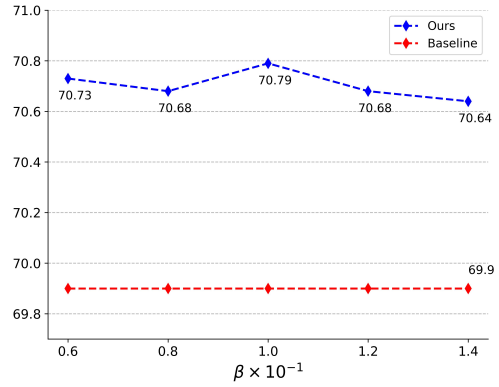

Figure 2. The hyper-parameter β for non-target loss $\mathcal{L}_{non}$ with ResNet-18 on ImageNet dataset.

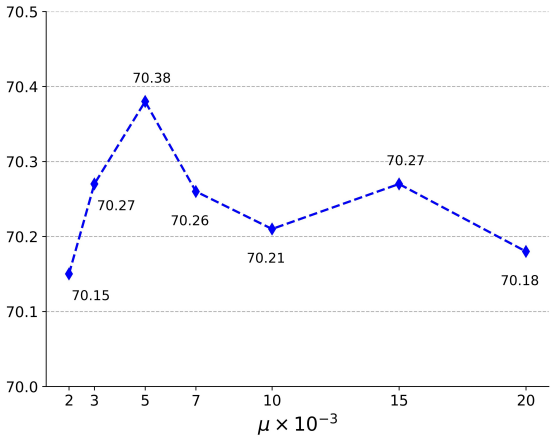

Figure 3. The hyper-parameter μ for weak supervision with RegNetX-1.6GF on ImageNet dataset.

$\mu \times 10^{-3}$	2	3	5	7	10
Acc.(%)	70.61	70.69	70.79	70.77	70.63

Table 4. The hyper-parameter μ with ResNet-18 on ImageNet.

F. Extension to Regression Tasks

We apply NKD to object detection, surpassing KD and DKD in Tab. 5. NKD shows significant mAP gains of 1.3 for Faster RCNN, indicating its potential for other tasks.

COCO	*baseline*	+KD	+DKD	+NKD$_{ours}$
mAP	*37.4*	37.8	38.2	**38.7**

Table 5. The detection results on COCO. Teacher: Faster RCNN ResNet-101 (2x). Student: Faster RCNN ResNet-50 (1x).

G. Analysis on the Coefficient.

The difference between the formula of NKD and DKD is $-(1-T_t)log(1-S_t)$. While DKD incorporates this term, NKD excludes it. DKD combines it with $T_tlog(S_t)$ as a whole to analyze its effects, and we only utilize $T_tlog(S_t)$ for distillation. In other word, DKD's coefficient for $log(1-S_t)$ is $-(1-T_t)$ and NKD's coefficient is 0. We conduct experiments to analyze the effects in Tab. 6 for KD and Tab. 7 for self-KD. The settings are: **T1:** $-T_tlog(S_t)$, **T2:** $-(1-T_t)log(1-S_t)$. Our NKD approach employs **T1**, while DKD utilizes **T1+T2**. In Tab. 7, we also try to get the soft labels with Exponential Moving Average (EMA) and compare with our USKD. The results of KD and self-KD both prove the superiority of our NKD and USKD.

Baseline	T1 $(NKD)_{ours}$	T2	T1+T2 (DKD)
69.90	**71.96**	70.91	71.70

Table 6. KD results of Res18 on ImageNet. **T1:** $-T_tlog(S_t)$, **T2:** $-(1-T_t)log(1-S_t)$. NKD uses **T1** and DKD uses **T1+T2**.

Baseline	T1 $(S1)_{ours}$	T2 (S1)	T1+T2 (S1)	T1 (S2)
69.90	**70.79**	69.50	70.22	70.36

Table 7. Self-KD results of Res18 on ImageNet. **S1** (USKD) and **S2** (EMA) denote different soft labels.