Self-Knowledge Distillation with Progressive Refinement of Targets

Kyungyul Kim1  ByeongMoon Ji1  Doyoung Yoon1  Sangheum Hwang2*

1LG CNS AI Research, Seoul, South Korea
2Seoul National University of Science and Technology, Seoul, South Korea
{kyungyul.kim, jibm, dy0916}@lgcns.com, shwang@seoultech.ac.kr

Abstract

The generalization capability of deep neural networks has been substantially improved by applying a wide spectrum of regularization methods, e.g., restricting function space, injecting randomness during training, augmenting data, etc. In this work, we propose a simple yet effective regularization method named progressive self-knowledge distillation (PS-KD), which progressively distills a model’s own knowledge to soften hard targets (i.e., one-hot vectors) during training. Hence, it can be interpreted within a framework of knowledge distillation as a student becomes a teacher itself. Specifically, targets are adjusted adaptively by combining the ground-truth and past predictions from the model itself. We show that PS-KD provides an effect of hard example mining by rescaling gradients according to difficulty in classifying examples. The proposed method is applicable to any supervised learning tasks with hard targets and can be easily combined with existing regularization methods to further enhance the generalization performance. Furthermore, it is confirmed that PS-KD achieves not only better accuracy, but also provides high quality of confidence estimates in terms of calibration as well as ordinal ranking. Extensive experimental results on three different tasks, image classification, object detection, and machine translation, demonstrate that our method consistently improves the performance of the state-of-the-art baselines. The code is available at https://github.com/lgcnsai/PS-KD-Pytorch.

1. Introduction

The recent progress made in deep neural networks (DNNs) has significantly improved performance in various tasks related to computer vision as well as natural language processing, e.g., image classification [14, 18, 22, 37], object detection / segmentation [13, 32], machine translation [41] and language modeling [20]. Scaling up of DNN is widely adopted as a promising strategy to achieve higher performances [12, 14, 39]. However, deeper networks require a large number of model parameters that need to be learned, which could make models more prone to overfitting. Thus, DNNs typically produce overconfident predictions even for incorrect predictions, and this is because the predictions are highly miscalibrated [10, 26].

To improve generalization performance and training efficiency of DNNs, a number of regularization methods have been proposed. The widely employed methods in practice include: $L_1$- and $L_2$-weight decay [23, 27] to restrict the function space, dropout [36] to inject randomness during training, batch normalization [19, 33] to accelerate training speed by normalizing internal activations in every layer. There also have been several methods that are specifically designed for a particular task. For example, advanced data augmentation techniques that are specific to computer vision tasks such as Cutout [5], Mixup [51], AugMix [16] and CutMix [49] have shown to boost classification accuracy and also improve robustness and uncertainty of a model. Another effective regularization method is to adjust the targets when they are given in the form of one-hot coded vectors (i.e., hard targets), including label smoothing (LS) [38], label perturbation [43], etc.

Among those methods about adjusting targets, LS [38] has been widely applied to many applications [31, 41, 53] and has shown to improve generalization performance as well as the quality of confidence estimates (in terms of calibration) on image classification and machine translation tasks [25]. LS softens a hard target as a smoothed distribution by assigning a small amount of probability mass to non-target classes. However, it is also empirically confirmed that it is not complementary to current advanced regularization techniques. For example, if we utilize LS and CutMix simultaneously for image classification, the performance on both classification and confidence estimation is substantially degraded [3].

One natural question raised on LS could be: is there a more effective strategy to soften hard targets so as to obtain...
more informative labels? To answer this question, we propose a simple regularization technique named progressive self-knowledge distillation (PS-KD) that distills the knowledge in a model itself and uses it for training the model. It means that a student model becomes a teacher model itself, which gradually utilizes its own knowledge for softening hard targets to be more informative during training. Specifically, the model is trained with the soft targets which are computed as a linear combination of hard targets and the past predictions at a certain epoch, which are adjusted adaptively as training proceeds.

To justify our proposed method, we have shown that PS-KD gives more weights to hard-to-learn examples by a gradient rescaling scheme during training, which clearly reveals that a student model can be enhanced even if trained with a teacher worse than the student (e.g., past predictions). The proposed method is easy to implement, can be applied to any supervised learning tasks where the hard targets are given as the ground-truth labels. Moreover, it can be easily combined with current advanced regularization techniques, thereby enhancing further their generalization performance. With this simple method, the generalization ability of DNNs can be greatly improved regarding the target metrics (e.g., accuracy) as well as the quality of confidence estimates in terms of both calibration and ordinal ranking (i.e., misclassification detection) [9, 24].

To rigorously evaluate the advantages of PS-KD, we conduct extensive experiments on diverse tasks with popular benchmark datasets: image classification on CIFAR-100 and ImageNet, object detection on PASCAL VOC, and machine translation on IWSLT15 and Multi30k. The experimental results demonstrate that training with PS-KD further enhances the generalization performance of the state-of-the-art baselines. For image classification on CIFAR-100 and ImageNet, our results show that the model trained with PS-KD provides not only better predictive performance, but also high quality of confidence estimates. We further confirm that the advanced image augmentation techniques such as Cutout and CutMix also benefit from PS-KD. From the evaluation on object detection with PASCAL VOC, it is observed that PS-KD makes a model learn better representations compared to the existing approaches. To show the wide applicability of PS-KD, we also conduct the experiments on machine translation, which improve BLEU scores of baselines significantly.

2. Related Work

Conventional Knowledge Distillation (KD) [17] methods use knowledge from a larger and better performing teacher model to generate soft targets for a student network. Recently, several works have tried to use the student network itself as a teacher, named self-knowledge distillation (self-KD) [11, 45, 50].

One approach to self-KD is to reduce the feature distance between similar inputs from a single network. Xu and Liu [45] propose a mechanism based on image distortion. Given an image, it generates two separate distorted images by using random mirroring and cropping. Then, model is trained to decrease the distance between features extracted from those two images. Yun et al. [50] propose a method called class-wise self-KD (CS-KD) which focuses on distilling knowledge between samples in the same class. For an input $x$, another data $x'$ with the same label is randomly sampled, and the KL divergence between predictive distributions from those is minimized during training. However, this approach may cause overfitting since it forces all samples in a specific class to have a high density in the learned representation space.

Another approach is to directly use outputs from a teacher whose architecture is exactly the same as a student. This approach is closer to the original notion of KD than the aforementioned approach. Born-Again Networks (BANs) [8] first train a network and use this pretrained network as a teacher for next generation. Then, it repeats this process, and thereby, performs multiple generations of KD where the $k$-th generation model is trained with knowledge distilled from the $(k - 1)$-th model. Similar to BANs, Yuan et al. [48] empirically examine the common belief on KD and suggest the teacher-free KD (TF-KD) method, which uses a pretrained student as a teacher for a single generation. Zhang et al. [52] propose a method which di-
vides a network into several blocks and attaches auxiliary classifiers to them independently. These auxiliary classifiers are trained using outputs of the final classifier as a teacher’s knowledge. Yang et al. [47] introduce utilizing a snapshot model at a certain epoch as a teacher: the whole training process is split into multiple mini-generations and the last snapshot model of each mini-generation is used as a teacher for the next mini-generation. These methods require augmenting a network’s architecture [52] or several hyperparameters to be carefully tuned, e.g., a learning rate policy and the number of mini-generation [47].

Along with previous studies, our proposed PS-KD method uses a model’s own predictions as a teacher’s knowledge to enhance the generalization performance of DNNs. However, our method provides distinct advantages over them from a practical viewpoint: it does not require a pretraining phase unlike BANs [8] and TF-KD [48], and can be easily applied to any supervised learning tasks due to its generality and simplicity compared to [45, 11, 52, 47]. More importantly, it is shown that we can enjoy the advantages of KD even if a student is taught by a poor teacher (i.e., lower predictive performance than a student), which is distinguishable from previous works relying on a well-performing teacher.

3. Self-Knowledge Distillation

3.1. Knowledge Distillation as Softening Targets

KD is a technique to transfer knowledge from one model (i.e., a teacher) to another (i.e., a student), usually from a larger model to a smaller one. The student learns from more informative sources, the predictive probabilities from the teacher, besides one-hot labels. Hence, it can attain a similar performance compared to the teacher although it is usually a much smaller model, and show comparable or even better performance when the student has the same capacity as the teacher [8, 48].

For an input $x$ and a $K$-dimensional one-hot target $y$, a model produces the logit vector $z(x) = [z_1(x), \ldots, z_K(x)]$, and then outputs the predicted probabilities $P(x) = [p_1(x), \ldots, p_K(x)]$ by a softmax function. Hinton et al. [17] suggest to utilize temperature scaling to soften these probabilities for better distillation:

$$\bar{p}_i(x, \tau) = \frac{\exp(z_i(x)/\tau)}{\sum_j \exp(z_j(x)/\tau)} \quad (1)$$

where $\tau$ denotes a temperature parameter. By scaling the softmax output $P^T(x)$ of the teacher as well as $P^S(x)$ of the student, the student is trained with the loss function $L_{KD}$, given by:

$$L_{KD}(x, y) = (1 - \alpha)H(y, P^S(x)) + \alpha \tau^2 H(\frac{P^T(x, \tau)}{\tau}, \frac{P^S(x, \tau)}{\tau}) \quad (2)$$

where $H$ is a cross-entropy loss and $\alpha$ is a hyperparameter. Note that when the temperature $\tau$ is set to 1, Eq. (2) is equivalent to the cross-entropy of $P^S(x)$ to the soft target, a linear combination of $y$ and $P^T(x)$:

$$L_{KD}(x, y) = H((1 - \alpha)y + \alpha P^T(x), P^S(x)) \quad (3)$$

Therefore, the existing methods that use the soft targets for regularization can be interpreted within the framework of knowledge distillation. For example, LS [25] is equivalent to distilling the knowledge from the teacher which produces uniformly distributed probabilities on any inputs.

3.2. Distilling Knowledge from the Past Predictions

We propose a new way of KD, called progressive self-knowledge distillation (PS-KD), which distills the knowledge of itself to enhance the generalization capability. In other words, the student becomes the teacher itself, and utilizes its past predictions to have more informative supervisions during training as can be seen in Fig. 1. Let $P^S_t(x)$ be the soft target, $P^{S_t}_{<t}(x)$ be the prediction about $x$ from the student at $t$-th epoch. Then, our objective at $t$-th epoch can be written as:

$$L_{KD,t}(x, y) = H((1 - \alpha)y + \alpha P^{S_{<t}}(x), P^S_t(x)) \quad (4)$$

Note that using the predictions from $t$-th epoch as the teacher’s knowledge is trivial since it will not incur any loss.

The main difference from the conventional KD is that the teacher is not a static model, but dynamically evolves as training proceeds. Among all past models that are candidates for the teacher, we use the model at $(t - 1)$-th epoch as the teacher since it can provide most valuable information among the candidates. Concretely, in $t$-th epoch of training, the target for the input $x$ is softened as $(1 - \alpha)y + \alpha P^{S_{t-1}}(x)$. It is empirically observed that this approach utilizing the past model as a teacher regularizes the model effectively.

One more thing we have to consider is how to determine $\alpha$ in Eq. (4). The $\alpha$ controls how much we are going to trust the knowledge from the teacher. In the conventional KD, the teacher remains unchanged so the $\alpha$ is usually set to a fixed value during training. However, in PS-KD, the reliability of the teacher should be considered since the model generally does not have enough knowledge about data at the early stage of training. To this end, we increase the value of $\alpha$ gradually. Like the learning rate scheduling, there are several strategies to increase the $\alpha$ as a function of the epoch, e.g., step-wise, exponential, linear growth, etc. To minimize the number of hyperparameters involved in the scheduling, we apply the linear growth approach. Therefore, the $\alpha$ at $t$-th epoch is computed as follows:

$$\alpha_t = \alpha_T \times \frac{t}{T} \quad (5)$$

where $T$ is the total epoch for training and $\alpha_T$ is the $\alpha$ at the last epoch, which is a single hyperparameter to be determined via validation process. Surprisingly, this simple
strategy combined with past predictions improves the generalization performance significantly across a wide range of tasks. To summarize, our objective function at $t$-th epoch can be written as:

$$L_{KD,t}(x, y) = H \left( (1 - \alpha_t)y + \alpha_t P^S_{t-1}(x), P^S_t(x) \right).$$

(6)

**Theoretical support.** We show that PS-KD pays more attention to hard examples during training when a hyperparameter $\alpha_t$ is properly set, and the $\alpha_t$ value should be gradually increased to preserve such an effect of hard example mining. This effect is realized by example re-weighting, motivated by Proposition 2 in Tang et al. [40]. The gradient of $L_{KD,t}$ in Eq. 6 with respect to a logit value $z_i$, $i = 1, ..., K$ for a fixed $\alpha^t$ is given by

$$\frac{\partial L_{KD,t}}{\partial z_i} = \partial_i^{KD,t} = (1 - \alpha_t)(p_t,i - y_i) + \alpha_t(p_t,i - p_{t-1,i}).$$

(7)

Therefore, for $z_{GT}$ where $GT$ denotes the target class,

$$\partial_{t}^{KD,t} = (1 - \alpha_t)(p_t,GT - 1) + \alpha_t(p_t,GT - p_{t-1,GT})$$

(8)

...and for $z_i$, $i \neq GT$,

$$\partial_i^{KD,t} = (1 - \alpha_t)p_t,i + \alpha_t(p_t,i - p_{t-1,i}).$$

(9)

If $\alpha$ is set to be $p_t,i - \alpha p_{t-1,i} \geq 0$ for all $i \neq GT$, i.e., $\alpha \leq \min_i (p_t,i/p_{t-1,i})$, then $(p_t,GT - 1) - \alpha(p_t,GT - p_{t-1,GT}) < 0$ and $\sum_{i \neq GT} |p_t,i - \alpha p_{t-1,i}| = (1 - p_t,GT) - \alpha(1 - p_{t-1,GT})$ holds. Therefore, $L_1$ norm $\sum_i |\partial_i^{KD,t}|$ of the gradient can be written as:

$$\sum_i |\partial_i^{KD,t}| = 2(1 - p_t,GT) - 2\alpha(1 - p_{t-1,GT}).$$

(10)

Let us consider the ratio of $L_1$ norms $\sum_i |\partial_i^{KD,t}|/\sum_i |\partial_i|$, which represents the gradient rescaling factor induced by PS-KD. By combining Eq. 10 and the fact that $\sum_i |\partial_i| = 2(1 - p_t,GT)$, the rescaling factor is given by:

$$\frac{\sum_i |\partial_i^{KD,t}|}{\sum_i |\partial_i|} = 1 - \alpha \left( \frac{1 - p_{t-1,GT}}{1 - p_t,GT} \right) \equiv 1 - \alpha \left( \frac{\gamma_{t-1}}{\gamma_t} \right).$$

(11)

Note that $\gamma$ represents the probability of being incorrect. Without loss of generality, it can be assumed that $p_t,GT \geq p_{t-1,GT}$ and $p_t,i \leq p_{t-1,i}$ for all $i \neq GT$ since $P^S_t$ provides better prediction than $P^S_{t-1}$ on training data. Therefore, $\gamma_{t-1} \geq \gamma_t$ always holds. A large $\gamma_{t-1}/\gamma_t$ means that predictions on that example are greatly improved during iterations (i.e., easy-to-learn). Conversely, hard-to-learn examples will have a small value. Consequently, the gradient rescaling factors for hard-to-learn examples are greater than those for easy-to-learn examples, which implies that PS-KD gives more weights on hard-to-learn examples during training, and it is empirically confirmed as shown in Fig. 2. Fig. 2 shows that other methods are more overconfident on incorrect predictions than PS-KD, and clearly demonstrates that PS-KD focuses on hard examples implicitly by the gradient rescaling scheme.

To expect such hard example mining effects, $\alpha$ should satisfy the condition described above for all examples. From this, $\alpha$ should be set to a sufficiently small value during an early training phase. As training proceeds, the differences of $\gamma_{t-1}/\gamma_t$ for all examples become smaller. Therefore, $\alpha$ should be gradually increased to preserve the effect.

**Implementation.** For PS-KD, the predictions from the model at $(t - 1)$-th epoch are necessary for training at $t$-th epoch. There are two ways to obtain these past predictions. One is to load the model at $(t - 1)$-th epoch on memory when $t$-th epoch is started so that the past predictions for softening targets are also computed in forward passes. The other is to save the past predictions on disk in advance during $(t - 1)$-th epoch, and read this information to compute the soft targets at $t$-th epoch. These two approaches have pros and cons. The former way may need more GPU memory. On the other hand, the latter way does not need additional GPU memory but requires more space to store past predictions.

The choice of how to obtain the past predictions depends on the task we are dealing with. For example, on machine translation with a large-scale corpus, it is nearly impossible to store the predicted probabilities over all tokens. For this, we can choose the former strategy. Note that softening targets via a moving average [1] or distilling knowledge with task-specific operations [45, 50] is not applicable to this task. In our experiments, we employ an efficient way according to the task, e.g., the past predictions from the model on GPU memory is utilized for the tasks on ImageNet classification and IWSLT15 machine translation.
4. Experimental Results

In this section, we show the effectiveness of PS-KD across a variety of tasks including image classification, object detection, and machine translation. More details on datasets, evaluation metrics and experimental settings are available in the supplementary material. All experiments were performed on NVIDIA DGX-1 system with PyTorch [30].

4.1. CIFAR-100 Classification

On CIFAR-100 classification, we consider four CNN models: ResNet [14], ResNeXt [44], DenseNet [18], and PyramidNet [12]. First, we compare PS-KD with LS and recent self-KD methods, CS-KD [50] and TF-KD [48], on the architectures we consider. Then, we demonstrate that PS-KD is complementary to the existing regularization methods including data augmentation (e.g., Cutout [5], CutMix [49], etc.), and ensembles.

Experimental settings. The detailed architectures we consider are PreAct ResNet-18 [15], ResNet-101 [14], ResNeXt-29 (cardinality=8, width=64) [44], DenseNet-121 (growth rate=32) [18] and PyramidNet-200 (widening factor=240) [12]. We follow standard data augmentation schemes: 32×32 random crop after padding with 4 pixels and random horizontal flip. All CNNs are trained using SGD with a momentum of 0.9 for 300 epochs, and the learning rate is decayed by a factor of 10 at 150 and 225 epochs. For ResNet, ResNeXt, and DenseNet, we set the mini-batch size, a weight decay, and an initial learning rate to 128, 0.0005, and 0.1, respectively. For PyramidNet, the mini-batch size, a weight decay, and an initial learning rate are set to 64, 0.0001, and 0.25, respectively, following [12, 49].

To compare the performance of PS-KD with LS, CS-KD and TF-KD, the hyperparameters are set according to those reported in the corresponding studies. Our PS-KD has only a single hyperparameter $\alpha_T$. To determine the optimal $\alpha_T$, we use randomly sampled 10% of training data as a validation dataset. In this experiment, we set the $\alpha_T$ to 0.8 which shows the optimal validation performance in terms of accuracy and calibration on ResNet-18. With this parameter value, we then train a model on the entire training dataset for a fair comparison. Note that the value tuned for ResNet-18 is also used for other architectures since we expect that PS-KD is fairly robust to the hyperparameter $\alpha_T$, and it is confirmed from the experimental results.

For existing regularization methods to be combined with PS-KD, we also follow the hyperparameter values reported

<table>
<thead>
<tr>
<th>Model + Method</th>
<th>Top-1 Err (%)</th>
<th>Top-5 Err (%)</th>
<th>NLL</th>
<th>ECE</th>
<th>AURC ($\times 10^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-18</td>
<td>24.18</td>
<td>6.90</td>
<td>1.10</td>
<td>11.84</td>
<td>67.65</td>
</tr>
<tr>
<td>+ LS</td>
<td>20.94</td>
<td>6.02</td>
<td>0.98</td>
<td>10.79</td>
<td>57.74</td>
</tr>
<tr>
<td>+ CS-KD</td>
<td>21.30</td>
<td>5.70</td>
<td>0.88</td>
<td>6.24</td>
<td>56.56</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>22.68</td>
<td>6.01</td>
<td>1.05</td>
<td>11.96</td>
<td>61.77</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>20.82</td>
<td>5.10</td>
<td>0.76</td>
<td>1.77</td>
<td>52.10</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>20.57</td>
<td>5.28</td>
<td>0.89</td>
<td>10.02</td>
<td>55.45</td>
</tr>
<tr>
<td>+ LS</td>
<td>19.84</td>
<td>5.07</td>
<td>0.93</td>
<td>3.43</td>
<td>95.76</td>
</tr>
<tr>
<td>+ CS-KD</td>
<td>20.76</td>
<td>5.62</td>
<td>1.02</td>
<td>12.18</td>
<td>64.44</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>20.10</td>
<td>5.10</td>
<td>0.84</td>
<td>6.14</td>
<td>58.8</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>19.43</td>
<td>4.30</td>
<td>0.74</td>
<td>6.92</td>
<td>49.01</td>
</tr>
<tr>
<td>DenseNet-121</td>
<td>20.05</td>
<td>4.99</td>
<td>0.82</td>
<td>7.34</td>
<td>52.21</td>
</tr>
<tr>
<td>+ LS</td>
<td>19.80</td>
<td>5.46</td>
<td>0.92</td>
<td>3.76</td>
<td>91.06</td>
</tr>
<tr>
<td>+ CS-KD</td>
<td>20.47</td>
<td>6.21</td>
<td>1.07</td>
<td>13.80</td>
<td>73.37</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>19.88</td>
<td>5.10</td>
<td>0.85</td>
<td>7.33</td>
<td>69.23</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>18.73</td>
<td>3.90</td>
<td>0.69</td>
<td>3.71</td>
<td>45.55</td>
</tr>
<tr>
<td>ResNeXt-29</td>
<td>18.65</td>
<td>4.47</td>
<td>0.74</td>
<td>4.17</td>
<td>44.27</td>
</tr>
<tr>
<td>+ LS</td>
<td>17.60</td>
<td>4.23</td>
<td>1.05</td>
<td>22.14</td>
<td>41.92</td>
</tr>
<tr>
<td>+ CS-KD</td>
<td>18.26</td>
<td>4.37</td>
<td>0.80</td>
<td>5.95</td>
<td>42.11</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>17.33</td>
<td>3.87</td>
<td>0.74</td>
<td>6.73</td>
<td>40.34</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>17.28</td>
<td>3.60</td>
<td>0.71</td>
<td>9.15</td>
<td>39.78</td>
</tr>
<tr>
<td>PyramidNet</td>
<td>16.80</td>
<td>3.69</td>
<td>0.73</td>
<td>8.04</td>
<td>36.95</td>
</tr>
<tr>
<td>+ LS</td>
<td>17.82</td>
<td>4.72</td>
<td>0.89</td>
<td>3.46</td>
<td>105.02</td>
</tr>
<tr>
<td>+ CS-KD</td>
<td>18.31</td>
<td>5.70</td>
<td>1.17</td>
<td>14.70</td>
<td>70.05</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>16.48</td>
<td>3.37</td>
<td>0.79</td>
<td>10.48</td>
<td>37.04</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>15.49</td>
<td>3.08</td>
<td>0.56</td>
<td>1.83</td>
<td>32.14</td>
</tr>
</tbody>
</table>

Table 1. Evaluation results on CIFAR-100 compared to other methods with popular architectures, averaged over three runs. The best result is shown in boldface.

in the literature, for example, the hole size in Cutout is set to 8 and the parameter $\alpha$ of Beta distribution in CutMix is set to 1. Cutout and CutMix produce randomly synthesized images from two inputs at every iteration. In this case, applying PS-KD with them at the same time is not straightforward. Therefore, for the experiments where PS-KD is combined with Cutout or CutMix, each data selects the regularization method with a probability of 0.5. Simply, PS-KD is applied to the half of the data in a randomly shuffled mini-batch, and Cutout or CutMix is performed on another half of the data.

Evaluation metrics. We use top-1 and top-5 error as standard performance measures for multi-class classification. We also employ the negative log-likelihood (NLL), expected calibration error (ECE) [29] and the area under the risk-coverage curve (AURC) [9] to evaluate the quality of predictive probabilities in terms of confidence estimation. ECE is a widely used metric to determine whether a model’s predictions are well-calibrated, approximating the difference in expectation between classification accuracy and confidence estimates. AURC measures the area under the curve from plotting the risk (i.e., error rate) according to confidence estimates. A low AURC implies that correct and incorrect predictions can be well-separable based on confidence esti-
mates. In these experiments, the maximum class probability is used as a confidence estimator.

Result. The comparison results are summarized in Table 1. Note that we report average values over three runs. First, we observe that training with PS-KD performs better than baseline and LS in terms of classification accuracy across all architectures, e.g., an improvement of 1.37% and 0.32% from baseline and LS on ResNeXt, respectively. Compared with CS-KD and TF-KD, PS-KD also shows better accuracy while significantly improving the performance on confidence estimation, for example, it improves accuracy by 1.74% and 1.15% from CS-KD and TF-KD on DenseNet-121, respectively, and it reduces ECE by 12.87% and 8.65% from CS-KD and TF-KD on PyramidNet, respectively. The performance improvement on confidence estimation is consistently observed across all metrics (i.e., NLL, ECE, and AURC) except for the two cases, ECE on ResNet-101 and ResNeXt. Nevertheless, PS-KD provides robust calibration performance on all architectures, e.g., less than 10% in ECE, as can be seen from the calibration plots in Fig. 3. Interestingly, ResNeXt provides well-calibrated predictions in terms of ECE, and it would be worth investigating which architectural factors of CNNs affect calibration performance.

From the results, it is confirmed that PS-KD is the only method which shows consistent and robust performance improvement across all metrics. Importantly, CS-KD performs worse than LS or even baseline in some cases from our experiments. We suspect that these results are caused by the implicit property of CS-KD, which pulls all samples in the same class each other, and it may accelerate overfitting when a model has high capacity or hyperparameters are not properly tuned.

To show that PS-KD can be used in conjunction with other advanced regularization methods, we present the detailed experimental results on PyramidNet in Table 2. Compared with Cutout and CutMix on PyramidNet, PS-KD shows slightly higher accuracy while significantly improving performances on confidence estimation. We observe the top-1 error of 14.82% when Cutout is combined with PS-KD, which is 1.23% improvement of Cutout. When PS-KD, CutMix, and SD are utilized simultaneously, the top-1 error from the combination of CutMix and SD is reduced by 0.48%. In this setting, it is confirmed again that PS-KD provides a positive effect on confidence estimation: all metrics, NLL, ECE, and AURC, are improved by PS-KD. More experimental results on other self-KD methods are presented in the supplementary material. These results demonstrate that current state-of-the-art regularization methods benefit from PS-KD in terms of not only classification accuracy, but also confidence estimation. From the previous study [3], it is known that LS might be harmful to generalization performance when applied concurrently with the advanced methods. Our empirical findings reveal that how to soften the hard targets is important and the distilled knowledge from a model itself can be a good source to create more informative targets.

We also examine the ensemble effects of PS-KD. For ensembling, three trained models in Table 1 are used. The best result is shown in boldface.

---

**Table 2. Performance enhancement of data augmentation methods, Cutout and CutMix, by PS-KD, averaged over three runs. The best result is shown in boldface.**

<table>
<thead>
<tr>
<th>Model + Method</th>
<th>Top-1 Err (%)</th>
<th>Top-5 Err (%)</th>
<th>NLL (%)</th>
<th>ECE (%)</th>
<th>AURC (×10^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyramidNet + PS-KD</td>
<td>16.80</td>
<td>3.69</td>
<td>0.73</td>
<td>8.04</td>
<td>36.95</td>
</tr>
<tr>
<td>+ Cutout [5]</td>
<td>16.95</td>
<td>3.08</td>
<td>0.56</td>
<td>1.83</td>
<td>32.14</td>
</tr>
<tr>
<td>+ Cutout + PS-KD</td>
<td>16.05</td>
<td>3.42</td>
<td>0.67</td>
<td>7.15</td>
<td>33.20</td>
</tr>
<tr>
<td>+ CutMix [49]</td>
<td>14.82</td>
<td>2.86</td>
<td>0.54</td>
<td>3.69</td>
<td>29.77</td>
</tr>
<tr>
<td>+ CutMix + PS-KD</td>
<td>15.62</td>
<td>3.38</td>
<td>0.68</td>
<td>8.16</td>
<td>34.60</td>
</tr>
<tr>
<td>+ CutMix + SD [46]</td>
<td>15.03</td>
<td>2.91</td>
<td>0.58</td>
<td>5.81</td>
<td>30.22</td>
</tr>
<tr>
<td>+ CutMix + SD + PS-KD</td>
<td>14.07</td>
<td>2.38</td>
<td>0.51</td>
<td>3.96</td>
<td>28.65</td>
</tr>
</tbody>
</table>

---

**Table 3. Performance improvement of ensembles by PS-KD. For ensembling, three trained models in Table 1 are used. The best result is shown in boldface.**

<table>
<thead>
<tr>
<th>Model + Method</th>
<th>Top-1 Err (%)</th>
<th>Top-5 Err (%)</th>
<th>NLL (%)</th>
<th>ECE (%)</th>
<th>AURC (×10^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-18 + CS-KD</td>
<td>18.39</td>
<td>4.21</td>
<td>0.74</td>
<td>3.07</td>
<td>43.33</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>21.07</td>
<td>4.80</td>
<td>0.90</td>
<td>6.48</td>
<td>52.70</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>18.79</td>
<td>4.17</td>
<td>0.68</td>
<td>5.12</td>
<td>44.29</td>
</tr>
<tr>
<td>ResNet-101 + CS-KD</td>
<td>18.27</td>
<td>3.99</td>
<td>0.74</td>
<td>4.01</td>
<td>44.46</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>18.16</td>
<td>3.97</td>
<td>0.71</td>
<td>1.52</td>
<td>43.66</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>17.26</td>
<td>3.42</td>
<td>0.63</td>
<td>2.15</td>
<td>37.85</td>
</tr>
<tr>
<td>DenseNet-121 + CS-KD</td>
<td>17.09</td>
<td>3.51</td>
<td>0.65</td>
<td>1.96</td>
<td>39.01</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>16.96</td>
<td>3.32</td>
<td>0.71</td>
<td>5.36</td>
<td>39.31</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>16.25</td>
<td>2.85</td>
<td>0.57</td>
<td>1.92</td>
<td>34.64</td>
</tr>
<tr>
<td>ResNeXt-29 + CS-KD</td>
<td>16.72</td>
<td>3.44</td>
<td>0.65</td>
<td>3.78</td>
<td>36.98</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>16.86</td>
<td>3.51</td>
<td>0.75</td>
<td>8.58</td>
<td>37.06</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>15.99</td>
<td>3.10</td>
<td>0.68</td>
<td>11.1</td>
<td>21.79</td>
</tr>
<tr>
<td>PyramidNet + CS-KD</td>
<td>14.58</td>
<td>2.85</td>
<td>0.60</td>
<td>2.63</td>
<td>30.04</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>15.29</td>
<td>3.93</td>
<td>0.76</td>
<td>4.72</td>
<td>36.53</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>14.11</td>
<td>2.58</td>
<td>0.50</td>
<td>2.79</td>
<td>27.29</td>
</tr>
</tbody>
</table>

---

5When \(\alpha_T\) is tuned on ResNeXt-29, we observed that most metrics are further improved: for \(\alpha_T = 0.7\), Top-1=17.06%, Top-5=3.68%, NLL=0.69, ECE=6.3%, and AURC=38.64.

---
metrics used in the previous section. Our method shows

Table 4 shows performances evaluated by the

4.2. ImageNet Classification

In the case of a large-scale dataset like ImageNet [4],

Table 4. Top-1/top-5 error, NLL, ECE and AURC results on ImageNet validation dataset. “*” denotes results reported in the original papers. The best result is in bold.

Figure 4. Predicted probabilities for samples in the validation dataset from baseline and PS-KD. The ground-truth labels of these images are “bulletproof vest” (top), and “stove” (bottom). better accuracy than LS and other self-KD methods, achieving a top-1 error of 21.41%. Also, PS-KD achieves better performance on confidence estimation, i.e., it reduces ECE by 3.28% for CS-KD and 2.19% from TF-KD, respectively. From our validation results provided in the supplementary material, we observe that CS-KD is sensitive to the hyperparameters while PD-KD is much more robust to the hyperparameter $\alpha_T$. Additionally, PS-KD is further improved on all metrics when combined with CutMix, especially in terms of ECE. It is consistent with the results on CIFAR-100, which demonstrates that PS-KD provides additional benefits to the existing regularization methods. On the other hand, other self-KD methods in conjunction with CutMix are even worse than the vanilla CutMix. We expect that the performance improvement can be greater if the knowledge from the recent past model is utilized, for example, the predictions from the model at $(t - 0.5)$-th epoch.

Examples of how our PS-KD improves the quality of predicted probability are shown in Fig. 4 (see the supplementary material for more examples). For the top image whose label is “bulletproof vest”, both baseline and PS-
KD produce an incorrect prediction. However, PS-KD outputs the class probabilities distributed over the classes that have similar visual characteristics while the baseline outputs overconfident prediction on non-target class. The bottom image is labeled as “stove” while containing multiple objects including “coffee pot” and “stove”. Both baseline and PS-KD correctly classify this image, however, PS-KD also produces a high probability on “Dutch oven” that is visually similar to the objects in the image. These quantitative and qualitative results support the advantage of PS-KD which acts as an effective and strong regularizer.

### 4.3. Object Detection

We also examine that other visual recognition tasks can benefit from PS-KD. For this, we perform the experiment on the task of object detection using PASCAL VOC [7] dataset. We use the 5k VOC 2007 trainval and 15k VOC 2012 trainval as training sets, and use the PASCAL VOC 2007 test as a test set, following [32, 49]. As a baseline, Faster R-CNN [32] is considered, and the improvement of detection performance is examined by replacing the original VGG-16 [34] backbone network with a ResNet-152 trained on ImageNet. We utilize six different backbones trained for the previous section: ResNet-152, ResNet-152 with LS, PS-KD, CS-KD, TF-KD, and PS-KD+CutMix. We then fine-tune Faster R-CNN with each backbone network for 10 epochs with a mini-batch size of 1, an initial learning rate of 0.001 decayed by a factor of 10 at 5 epochs.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP (IoU &gt; 0.5) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-152</td>
<td>78.26</td>
</tr>
<tr>
<td>+ LS</td>
<td>78.44</td>
</tr>
<tr>
<td>+ CS-KD</td>
<td>78.33</td>
</tr>
<tr>
<td>+ TF-KD</td>
<td>78.28</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td>79.50</td>
</tr>
<tr>
<td>+ PS-KD + CutMix</td>
<td>79.72</td>
</tr>
</tbody>
</table>

Table 5. Effect of PS-KD as a pretrained backbone network for Faster R-CNN. The mAP value is computed by averaging APs over classes.

As shown in Table 5, ResNet-152 with PS-KD significantly improves the detection performance by 1.06%, 1.17%, and 1.22% of the mean average precision (mAP) compared to ResNet-152 with LS, CS-KD, and TF-KD, respectively. Furthermore, PS-KD shows better mAP when it is combined with CutMix. Note that this improvement is achieved by just replacing the backbone network. From this result, it is verified that training with PS-KD provides a strong backbone network, which provides generic representations that can be transferred to other visual recognition tasks.

### 4.4. Machine Translation

To verify the effectiveness of PS-KD on other tasks rather than multi-class classification, a machine translation task where classification is performed on a token-level, not an input-level is considered.

We use two benchmark datasets including IWSLT15 English to German (EN-DE) and German to English (DE-EN) [2], and Multi30k [6] from WMT16 [35]. The original purpose of Multi30k is for multimodal learning, consisting of images and descriptions associated with them. For the experiment, we extract only image descriptions written in English and German translations by professional translators. This dataset consists of 29K train data, 1K validation data, and 10K test data with 9,521 vocabularies.

We consider Transformer [41] as our baseline model. All hyperparameters involved in the architecture and training are set to those reported in [42]. In specific, we use the architecture with $N = 6$, $d_{model} = 512$, $h = 4$, $d_k = 64$, $d_{ff} = 1024$. We train the model for 150 epochs with a maximum of 4,096 tokens per a mini-batch, and employ Adam optimizer [21] with $\beta_1 = 0.9$, $\beta_2 = 0.98$. As a metric, BLEU, commonly used to evaluate the performance on machine translation, is used. The hyperparameter $\alpha_T = 0.7$ is determined through validation.

<table>
<thead>
<tr>
<th>Model</th>
<th>IWSLT15</th>
<th>Multi30k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN-DE</td>
<td>DE-EN</td>
</tr>
<tr>
<td>Transformer</td>
<td>28.5</td>
<td>34.6</td>
</tr>
<tr>
<td>+ LS</td>
<td>29.3</td>
<td>35.6</td>
</tr>
<tr>
<td>+ PS-KD</td>
<td><strong>30.0</strong></td>
<td><strong>36.2</strong></td>
</tr>
</tbody>
</table>

Table 6. BLEU scores on Transformer with LS or PS-KD

The results are summarized in Table 6. Our PS-KD achieves the best BLEU scores on all datasets. Consistent with the results from image classification and object detection, PS-KD shows better performance than the baseline Transformer and that with LS.

### 5. Conclusion

We propose a simple way to improve the generalization performance of DNNs, which distills the knowledge of a model itself to generate more informative targets for training. The targets are softened by using past predictions about data from the model at the previous epoch. We also provide theoretical justification, which shows that our method performs hard example mining implicitly during training. From the experimental results conducted across diverse tasks, we observe that the proposed method is effective to improve the generalization capability of DNNs.

---

7 APs over all classes are presented in the supplementary material.

---

8 Experiments were conducted using Fairseq (https://github.com/pytorch/fairseq) toolkit [28].
References


Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In The IEEE International Conference on Computer Vision, 2019.


