Feature Normalized Knowledge Distillation for Image Classification

Kunran Xu[1], Lai Rui[1,*], Yishi Li[1], and Lin Gu[2,3]

[1] School of Microelectronics, Xidian University, Xi'an Shaanxi 710071, China
aazzttcc@gmail.com; Corresponding author:rlai@mail.xidian.edu.cn;
yshlee1994@outlook.com
[2] RIKEN AIP, Tokyo103-0027, Japan lin.gu@riken.jp
[3] The University of Tokyo, Japan

Abstract. Knowledge Distillation (KD) transfers the knowledge from a cumbersome teacher model to a lightweight student network. Since a single image may reasonably relate to several categories, the one-hot label would inevitably introduce the encoding noise. From this perspective, we systematically analyze the distillation mechanism and demonstrate that the L_2-norm of the feature in penultimate layer would be too large under the influence of label noise, and the temperature T in KD could be regarded as a correction factor for L_2-norm to suppress the impact of noise. Noticing different samples suffer from varying intensities of label noise, we further propose a simple yet effective feature normalized knowledge distillation which introduces the sample specific correction factor to replace the unified temperature T for better reducing the impact of noise. Extensive experiments show that the proposed method surpasses standard KD as well as self-distillation significantly on Cifar-100, CUB-200-2011 and Stanford Cars datasets. The codes are in https://github.com/aztc/FNKD

Keywords: Label noise; Knowledge Distillation; Image Classification

1 Introduction

Convolutional Neural Network (CNN) has achieved great success in the field of artificial intelligence in recent years, especially in computer vision [10, 39, 2, 8, 7, 16, 1]. However, this success is accompanied by high inference cost in computation and memory. Many works have devoted to reduce computational complexity, such as model pruning [9, 5], lightweight network structure design [12, 28, 20] and automated architecture search [31, 32]. One promising and widely used method for model lightweight is Knowledge Distillation (KD) proposed by Hinton et al. [11], which transfers 'dark knowledge' from an ensemble or full model to a single compact model via soft-target cross entropy loss function. Through distillation, student model not only inherits better quality from the teacher, but also be more efficient for inference due to its compactness.

Recently, KD has made huge success, some works have extended this effective idea to other application domains [13, 39, 18], others [37, 17, 24] improve

the standard KD by kinds of techniques such as feature based distillation and achieve better results. Researchers also attempt to seek for better understanding about KD. For example, Lopez-Paz et.al [19] established the connection between KD and privileged information, Yuan et al. [38] regarded KD as an special case of Label Smoothing Regularization (LSD) which imposes constraint on student training and Tang et al. [33] broke down the effects of KD into label smoothing, example re-weighting and prior knowledge of optimal output layer geometry three aspects. In this paper, we systematically analyze the mechanism of temperature in KD from the perspective of label noise and introduce the L_2-norm of the feature in penultimate layer into soft-target to make a further improvement.

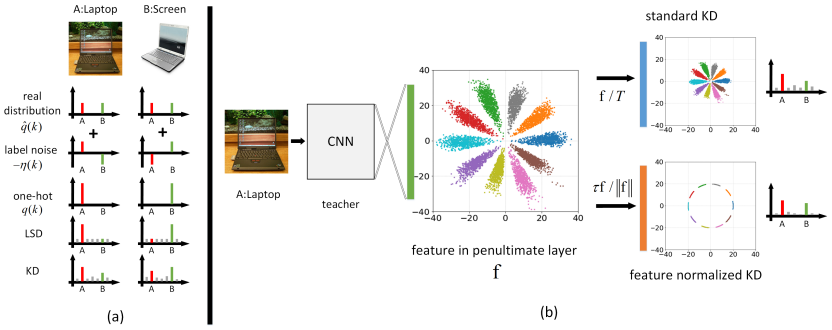

Fig. 1. The Overview for this paper. (a) Since there are visual similarities between images, one-hot label which assumes classes are independent can't always accurately describe image's real distribution over classes. The difference between one-hot label and real distribution is a kind of label noise. LSD and KD can provide better supervisions. (b) KD softens the label by reducing the L_2-norm of the feature in penultimate layer with a unified T while our method with a unique $\|\mathbf{f}\|$ for each sample.

Since there are visual similarities between images, one-hot label which assumes classes are independent can't always accurately describe image's real distribution over classes [23], as shown in Fig 1.(a). The difference between one-hot label and real distribution is a kind of label noise and harms model accuracy [6]. From the perspective of label noise, the methods like LSD [30] are actually introducing the priors to reduce this noise. We show that KD could be regarded that the teacher network learns noise information and produces better denoised labels (Fig 1.(a)). Then, we demonstrate that the L_2-norm of the feature in penultimate layer would be too large under the influence of label noise, and the temperature in KD could be regarded as a correction factor for L_2-norm to suppress the impact of noise. Since the L_2-norm also indicates the intensity of label noise, we introduce the L_2-norm into soft-target as the sample specific correction factor to replace the unified temperature T for better reducing the impact of noise (Fig 1.(b)). Finally, we empirically show that our proposed method

combines the advantages of both KD and Hypersphere Embedding (HE) [26, 35] which is another effective regularization.

In summary, we make the following contributions:

1) We systematically analyzed the mechanism of temperature in Knowledge Distillation from the novel perspective of one-hot label noise and verified that the higher temperature is actually a correction factor for L_2-norm of penultimate layer's feature which represents the intensity of label noise.
2) Based on our theoretical analysis and empirical findings, we proposed a simple yet effective innovative feature normalized KD for further refinement of temperature mechanism. Extensive experiments on image classification datasets support that our proposed method could improve over standard KD.
3) We has shown the relationship between KD, LSD and HE and verified empirically that the proposed feature normalized knowledge distillation benefits from both KD and HE.

The rest of the paper is organized as follows. We review related works in Section 2. The systematic demonstration and the proposed innovative method is introduced in Section 3. Experimental results on image classification datasets are presented and analyzed in Section 4 and followed by conclusions in Section 5.

2 Related Works

Since Hinton et al. [11] proposed Knowledge Distillation to implement knowledge transfer, many works have extended this approach to other domains. For example, by introducing KD, Zheng et al. [13] achieved fast and accurate super-resolution with a compact CNN. Zhang et al. [39] distilled the knowledge from multiple image parts into a single model to improve the performance of fine-grained visual classification. Li et al. [18] attempted to learn from noisy data with distillation which transfers the knowledge learned in small clean dataset.

In addition to exploring the application of KD in other fields, another important direction works on further improving the performance of knowledge distillation. Romero et al. [27] suggested that a student model could also imitate intermediate representations (feature maps) learned by teacher model and proposed to distill the knowledge by MSE loss and soft-label loss jointly. Guided by this work, feature based distillation has made impressive developments. Yim et al. [37] demonstrated that learning transform direction between two layers of the teacher model instead of just mimicking the features would be more efficient. Park et al. [24] introduced a novel approach that transfers relational properties of different sample's feature. Sun et al. [29] extended the FITNET [27] to minimise the mean-squared error between each individual layer of the student and teacher, which is effective for BERT model compression.

Except for feature based distillation, there are methods to improve the training procedure. Since KD requires a two-phase training procedure, Lan et al. [17]

presented an On-the-fly Native Ensemble model learning strategy for one-stage online distillation to reduce the complexity of training phase. Yang et al. [36] extracted information from earlier epochs in the same generation to enable teacher-student optimization in one generation, which could further improve the efficiency of training. Mirzadeh et al. [21] found the student network performance degrades when the gap between student and teacher is large and introduced an assistant mechanism to bridge the gap.

Despite the large success of knowledge distillation, surprisingly few theoretical research has been done to better understand the mechanism of how it works. Phuong et al. [25] provided the first insights into the working mechanisms of distillation by studying the special case of deep linear classifiers and found three key factors that determine the success of distillation, which is data geometry, optimization bias and strong monotonicity. Müller et al. [22] found the conflict between KD and LSD and argued that the resemblances between instances of different classes is crucial for distillation. Moreover, Tang et al. [33] broke down the effects of KD into label smoothing, example re-weighting and prior knowledge of optimal output layer geometry three aspects which is similar to [25]. The most relevant work to ours is Yuan et al. [38], who also realized the relation ship between KD and LSD, however they did not further reveal the underlying causes of both methods. In addition, we are working on an improved KD while they aimed for a teacher-free framework.

3 Method

In this section, we provide a systematic analyses for the mechanisms behind our proposed method based on theoretical and empirical results. We start by introducing the ubiquitous label noise in one-hot label and analyzing the standard KD from the perspective of label denoising. After that, we propose our feature normalized knowledge distillation based on the finding that the L_2-norm of the feature in penultimate layer is a good estimation for label noise.

3.1 Noise in One-Hot Label

Label noise is a common problem in image classification. As suggested by Benoît [6], label noise has four main classes of potential sources. In this paper, we focus on label encoding noise.

Considering K-way image classification, for a given $\mathbf{x}$, the goal is to learn a parametric mapping function $\sigma(\psi(\mathbf{x};\theta))$ which generates a class distribution $\mathbf{p}(k|\mathbf{x})$ to estimate the real distribution $\hat{\mathbf{q}}(k|\mathbf{x})$, where $k \in \{1...K\}$, σ is the *softmax* normalized function and $\psi(\mathbf{x};\theta)$ denotes the CNN consisting of stacked convolution, pooling, relu, etc. In most cases, we don't know $\hat{\mathbf{q}}(k)$ exactly and use the ont-hot encoded label $\mathbf{q}(k) = \{^{1}_{0} \quad ^{k=t}_{k\neq t}$ to approximate it, where t is the ground-truth class of $\mathbf{x}$ and we omit $\mathbf{x}$ in conditional distribution for simplicity. However, the approximation by $\mathbf{q}(k)$ would introduce label noise. Here, we show some examples from datasets CUB-200-2011 [34] and ImageNet [4] in Figure 2.

From the top row, we can see that although images in "Black footed Albatross" are very similar to those in "Sooty Albatross", the probabilities that they are in class "Sooty Albatross" are still assigned 0 by one-hot label. In addition, the class "Black footed Albatross" looks much closer to "Sooty Albatross" than "Pileated Woodpecker", but one-hot label makes these two classes have the same "distance" to "Sooty Albatross". These phenomenons are more salient in ImageNet which contains 1000 categories with more visually similar subclasses. For instance, the images of the middle bottom row in Figure 2 could be classified as either "Laptop Computer" or "Screen CRT", however, they are eventually assigned a hundred percent probability of being in category "Laptop Computer" out of some subjective reasons. This phenomenon is serious and common in this dataset.

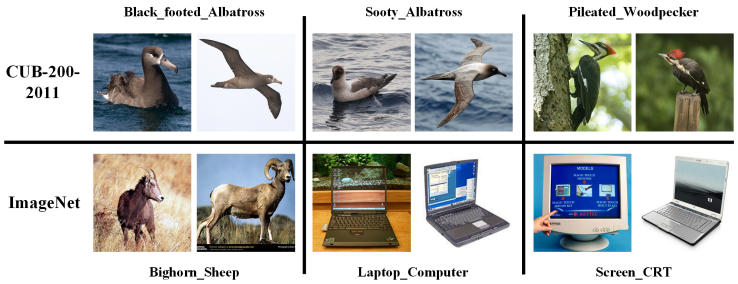

Fig. 2. The visual examples for label noise. From the top row, we could see that although images belonging to "Black footed Albatross" are very similar to images of "Sooty Albatross", the probability that they are in class "Sooty Albatross" is still assigned 0 by one-hot label. The class "Black footed Albatross" looks much closer to "Sooty Albatross" than "Pileated Woodpecker", but these three classes have the same "distance" to each other hypothesized by one-hot label. Similar situation would be observed in the ImageNet.

According to the above analysis, we could see that one-hot label assumes that categories are independent of each other and every $\mathbf{x}$ has no correlation to non ground-truth classes. However, images belonging to different classes often have much visual similarities even though they are semantically independent. Therefore, the strong hypothesis of one-hot label will bring about noise between $\hat{\mathbf{q}}(k)$ and $\mathbf{q}(k)$ (Fig 1.(a)). In view of this, we introduce a compensation distribution $\eta(k)$ and let

$$\hat{\mathbf{q}}(k) = \mathbf{q}(k) + \eta(k). \tag{1}$$

The $-\eta(k)$ could be used to represent the noise caused by one-hot label, as shown in Fig 1.(a). Although datasets with one-hot label often contain noise, it is still difficult to find a more precise estimation for $\hat{\mathbf{q}}(k)$ than $\mathbf{q}(k)$ manually. So, the noise $-\eta(k)$ would be present widely to varying degrees and causes some troubles such as over-fitting.

In addition to qualitative analysis above, we perform a simple experiment on CUB-200-2011 to quantitatively measure the impact of label noise on model accuracy. Although we don't know the true $\hat{\mathbf{q}}(k)$ so the real noise $-\eta(k)$ can't be obtained, we can still introduce some priors to estimate it just like Label Smoothing Regularization (LSD) [30]. LSD is an effective regularization that softens one-hot label by mixing $\mathbf{q}(k)$ with an uniform distribution $u(k)$, which is mathematically formulated as

$$\mathbf{q}_{lsd}(k) = (1-\epsilon_1)\mathbf{q}(k) + \epsilon_1 u(k), \tag{2}$$

where $\mathbf{q}_{lsd}(k)$ is the modified label and ϵ_1 controls its smoothness. Comparing Equation 2 and Equation 1, we can regard $\frac{\epsilon_1}{1-\epsilon_1}u(k)$ as an estimation for $\eta(k)$, so the LSD is actually a label denoising approach by using isotropic filtering to produce a distribution approximating $\hat{\mathbf{q}}(k)$. Considering that most images exhibit varying degrees of visual similarities within each category instead of uniform diversities hypothesized by LSD, we present an anisotropic LSD which introduces another non-uniform distribution $\rho(k)$ to estimate $\hat{\mathbf{q}}(k)$ better. The anisotropic LSD label $\mathbf{q}_{alsd}(k)$ can be expressed as:

$$\begin{aligned} \mathbf{q}_{alsd}(k) &= (1-\epsilon_1-\epsilon_2)\mathbf{q}(k) + \epsilon_1 u(k) + \epsilon_2 \rho(k) \\ \rho(k) &= \begin{cases} 1/|\mathcal{A}(\mathbf{x})| & k \in \mathcal{A}(\mathbf{x}), \\ 0 & k \notin \mathcal{A}(\mathbf{x}), \end{cases} \end{aligned} \tag{3}$$

where $\mathcal{A}(\mathbf{x})$ denotes the set containing classes more 'closer' to the ground-truth class of $\mathbf{x}$. $\mathcal{A}(\mathbf{x})$ is decided by category's name provided by the dataset. For instance, CUB-200-2011 contains 5 different species of "Woodpecker", so the probability for each "Woodpecker" subclass will be $\epsilon_1/5$ more than other non ground-truth classes. Due to the introduction of meta-information about categories, we could presume that $\mathbf{q}_{alsd}(k)$ is a better estimation for $\hat{\mathbf{q}}(k)$ than $\mathbf{q}_{lsd}(k)$ and will result in higher model accuracy.

To verify this, we use LSD and anisotropic LSD with $\epsilon_1 = 0.2$ and $\epsilon_2 = 0.02$ to train Resnet [10] and then compare their performance as shown in Table 1. As expected, LSD outperforms one-hot label in terms of accuracy significantly which is consistent with that reporterd by Szegedy et.al [30], our anisotropic LSD further improves by introducing more information about $\hat{\mathbf{q}}(k)$. This experiment indicates that the noise in one-hot label indeed causes a decline in model accuracy, as the estimation for $\eta(k)$ becomes more accurate, we would obtain a more 'clean' target to alleviate the problem caused by one-hot label. Seeing that LSD suppresses label noise via a pre-defined prior, we argue that learning from data is another way to estimate the real distribution $\hat{\mathbf{q}}(k)$. In the following section, we will explain that KD is actually one of such methods.

3.2 Standard Knowledge Distillation

As suggested by Li et al. [38], KD is a special case of LSR. In this section, we will further discuss their relationship from the perspective of label denoising and

Table 1. Training different architectures with different labels. LSD denotes the Label Smoothing Regularization and anisotropic LSD is its improved version. These two labels could be regarded as the denoised versions of original one-hot label and achieve much better model accuracy, which indicates that the noise in one-hot label have serious impact on model training. The last row shows the results of training by KD with a Resnet50 teacher and a Resnet152 teacher respectively.

Label	Resnet18 [10]	Resnet50 [10]
one-hot label	77.23	83.49
LSD [30]	77.85	84.05
anisotropy LSD	78.21	84.45
KD $T = 1$	79.46	85.78

argue that the teacher in KD provides a more accurate estimation for label noise, so KD could obtain a distribution approximating real distribution $\hat{\mathbf{q}}(k)$ better to help promote a student.

A CNN is usually trained by minimizing the cross-entropy loss $\mathcal{H}(\mathbf{q}, \mathbf{p}) = -\sum_{k=1}^{K} \mathbf{q}(k) log(\mathbf{p}(k))$ over the entire datasets. Instead of optimizing the $\mathcal{H}(\mathbf{q}, \mathbf{p})$, KD modified the objective by adding another regularization term. The optimizing objective is generally defined as

$$\mathcal{L}_{kd} = (1-\alpha)\mathcal{H}(\mathbf{q}, \sigma(\psi(\mathbf{x};\theta))) + \alpha\mathcal{H}(\mathbf{q}_{kd}, \sigma(\frac{\psi(\mathbf{x};\theta)}{T})) \tag{4}$$

$$\mathbf{q}_{kd}(k) = \frac{exp(\mathbf{v}_k/T)}{\sum_{i=1}^{K} exp(\mathbf{v}_i/T)}, \tag{5}$$

where $\mathbf{q}_{kd}(k)$ is another label generated by the teacher model based on its logits $\mathbf{v} \in \mathbb{R}^K$ and α controls the balance between two terms. The temperature T introduced by Hinton et al. [11] acts as a factor to smooth both the outputs of teacher and student.

When $T = 1$, the Equation 4 can be simplified to

$$\mathcal{L}_{kd}^{T=1} = \mathcal{H}(\mathbf{q}_t, \sigma(\psi(\mathbf{x};\theta))) \tag{6}$$

$$\mathbf{q}_t(k) = (1-\alpha)\mathbf{q}(k) + \alpha\mathbf{q}_{kd}(k). \tag{7}$$

Comparing Equation 7 with Equation 2 and Equation 3, it is easy to find that these equations have the similar form. This implies that $\mathbf{q}_t(k)$ here plays the same role as $\mathbf{q}_{lsd}(k)$ and $\mathbf{q}_{alsd}(k)$. Based on the discussion above, we argue that $\frac{\alpha}{1-\alpha}\mathbf{q}_{kd}(k)$ is also a estimation for compensation distribution $\eta(k)$ and the only difference is that $u(k)$ and $\rho(k)$ is pre-defined while $\mathbf{q}_{kd}(k)$ is learned by another CNN from data. Therefore, from the perspective of label denoising, both LSD and KD aim to remove noise in one-hot label and produce distributions closer to real distribution $\hat{\mathbf{q}}(k)$. To further compare KD and LSD quantitatively, we use

Resnet152 and Resnet50 as teachers to train Resnet50 and Resnet18 respectively using Equation 6 with $\alpha = 0.5$. As shown in the bottom row in Table 1, KD outperforms LSD and anisotropic LSD with a large margin in both settings, which demonstrates that teacher could learn about label noise from training data and thus provide a better compensation $\mathbf{q}_{kd}(k)$ than the uniform prior $u(k)$.

To further suppress the noise's influence, Hinton et al. [11] introduced an simple yet effective way by raising the temperature T. With T increasing, the distribution computed by Equation 5 is going to be softer. When $T \gg \mathbf{v}_k$, we can Taylor expand both numerator and denominator in Equation 5 to approximate $\mathbf{q}_{kd}(k)$ as

$$\mathbf{q}_{kd}(k) \approx \frac{1 + \mathbf{v}_k/T}{K + \sum_i \mathbf{v}_i/T} \approx \frac{1}{K} = u(k). \tag{8}$$

Equation 8 illustrates that as temperature raises, $\mathbf{q}_{kd}(k)$ becomes gradually smooth until perfectly even. Previous works [17, 11] have found empirically that a moderate T would yield a better student. From the perspective of label noise, we argue that this is because the $\mathbf{q}_{kd}(k)$ with small T is more noisy due to its less diversity from one-hot label while the higher temperature would filter out much noise's influence and help student better. To make this argument more clear, we will discuss what the logits $\mathbf{v}$ of teacher is and how it is affected by one-hot label noise in the next section.

3.3 Feature in Penultimate Layer

In this section, we will at first show that the L_2-norm of the feature in the penultimate layer could effectively indicate the one-hot label noise. We further demonstrate that the temperature T in KD could be regarded as an correction factor for this noise.

The last layer of a CNN for classification is usually a K-way fully-connected operation. It takes feature $\mathbf{f} \in \mathbb{R}^D$ as input and produces the logits $\mathbf{v}$ by a linear transform $\mathbf{v} = \mathbf{W}\mathbf{f}$, where $\mathbf{W} \in \mathbb{R}^{C \times D}$ is a parameter matrix. Note that

$$\mathbf{v}_i = \mathbf{W}_i \cdot \mathbf{f} = \|\mathbf{f}\| \|\mathbf{W}_i\| \cos(\theta_i), \tag{9}$$

where $\|\cdot\|$ denotes L_2-norm, $\mathbf{v}_i$ is the i^{th} element of $\mathbf{v}$, $\mathbf{W}_i$ is the i^{th} row of matrix $\mathbf{W}$ and θ_i is the angle between vector $\mathbf{f}$ and $\mathbf{W}_i$. Since $\mathbf{f}$ has the same influence on each $\mathbf{v}_i$, if we just want to know the class of a given sample, $\|\mathbf{W}_i\| \cos(\theta_i)$ is sufficient, so what is the role of $\|\mathbf{f}\|$? Wang et al. [35] has proven that a feature $\mathbf{f}$ with bigger L_2-norm could produce a harder distribution and fit one-hot label better. Since a *softmax* loss always encourages examples classified correctly to have higher probability, so the L_2-norm of the feature in penultimate layer would be larger and larger in the training. In brief, the larger the $\|\mathbf{f}\|$ is, the closer the output distribution will be to one-hot label. However, considering the fact that there exists label noise as illustrated above, the real distribution $\hat{\mathbf{q}}(k)$ is actually softer than one-hot label, so large $\|\mathbf{f}\|$ is partially caused by noise and a shorter $\mathbf{f}$ would be more appropriate. Therefore, we can see that why there is a $T > 1$

in KD, in fact, the temperature T could be regarded as an correction factor for L_2-norm, which abates $\|\mathbf{f}\|$ to weaken the impact of label noise.

Since $\|\mathbf{f}\|$ is partially decided by label noise, we argue that the $\|\mathbf{f}\|$ could be used to indicate the noise intensity. To study this empirically, we conduct a experiment on CUB-200-2011 to select three categories and exhibit the images with first two biggest and smallest $\|\mathbf{f}\|$ of each class. As shown in Fig 3, the images with lower feature L_2-norm have similar angles, illuminations, backgrounds and very alike looking birds. Comparatively speaking, the images with larger $\|\mathbf{f}\|$ looks more characteristic and easier to tell apart. Rajeev [26] performed a similar experiment to divide the images from IJB-A [14] dataset into 3 sets based on the example's $\|\mathbf{f}\|$ and found that the images with smaller $\|\mathbf{f}\|$ have poor quality and are hard for model to classify correctly. These results demonstrate that although one-hot label encourages $\mathbf{f}$ to have large L_2-norm, these hard examples which contain more label noise would still remain relatively small L_2-norm on account of some model priors. Noticing different samples suffer from varying intensities of label noise, we propose a novel feature normalized KD by suppressing noise for each sample according to their L_2-norm instead of an identical T for all samples.

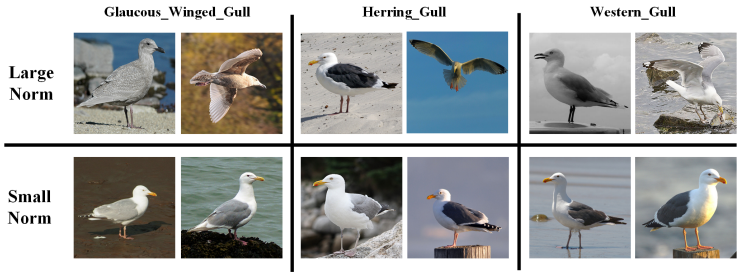

Fig. 3. Images with different feature L_2-norm. The bottom row shows these examples with lower feature L_2-norm, it could be seen that these images have similar angle, illumination, backgrounds and very similar looking birds. Comparatively speaking, the images with larger L_2-norm looks easier to tell apart. It indicates that the $\|\mathbf{f}\|$ could be used to represent the noise intensity in one-hot label.

3.4 Feature Normalized Knowledge Distillation

As previously mentioned, the L_2-norm of examples's feature represents the noise intensity in one-hot label, where the lower L_2-norm indicates stronger noise intensity. Therefore, we propose to weight every sample by the inverse of L_2-norm. With that in mind, we introduce a novel teacher's supervision distribution as

$$\mathbf{q}_{fn}(k) = \frac{exp(\frac{\tau \mathbf{v}_k}{\|\mathbf{f}\|})}{\sum_{i=1}^{K} exp(\frac{\tau \mathbf{v}_i}{\|\mathbf{f}\|})}, \tag{10}$$

where τ is a parameter controlling the smoothness of distribution $\mathbf{q}_{fn}(k)$ like the original temperature T. For students, we let them to compute a similar output following Equation 10 to mimic the teacher, so the final feature normalized KD learning objective is

$$\mathcal{L}_{fn} = \mathcal{H}(\mathbf{q}, \sigma(\psi(\mathbf{x}; \theta))) + \lambda^2 \mathcal{H}(\mathbf{q}_{fn}, \sigma(\frac{\tau \psi(\mathbf{x}; \theta)}{\|\mathbf{f}\|})), \tag{11}$$

in which we take away the parameter α and add a new weight λ compared to original soft-target cross-entropy loss expressed in Equation 4. To further illustrate, we compute the loss gradient, $\partial \mathcal{L}_{fn}/\partial \mathbf{z}_k$, with respect to each logit $\mathbf{z}_k$ of the distilled student model. This gradient is given by

$$\begin{aligned} \frac{\partial \mathcal{L}_{fn}}{\partial \mathbf{z}_k} &= \mathbf{p}(k) - \mathbf{q}(k) + \frac{\lambda^2 \tau}{\|\mathbf{f}_s\|}(\mathbf{p}_{fn}(k) - \mathbf{q}_{fn}(k)) \\ &= \mathbf{p}(k) - \mathbf{q}(k) + \frac{\lambda^2 \tau}{\|\mathbf{f}_s\|}(\frac{exp(\frac{\tau \mathbf{z}_k}{\|\mathbf{f}_s\|})}{\sum_i exp(\frac{\tau \mathbf{z}_i}{\|\mathbf{f}_s\|})} - \frac{exp(\frac{\tau \mathbf{v}_k}{\|\mathbf{f}_t\|})}{\sum_i exp(\frac{\tau \mathbf{v}_i}{\|\mathbf{f}_t\|})}). \end{aligned} \tag{12}$$

Using a similar proof technique as Hinton et al. [11], assuming that $\sum_i \mathbf{z}_i = \sum_i \mathbf{v}_i = 0$ and $\|\mathbf{f}_s\| = \|\mathbf{f}_t\| \gg \tau \mathbf{z}_k$, we Taylor expand both numerator and denominator in the last term and get

$$\frac{\partial \mathcal{L}_{fn}}{\partial \mathbf{z}_k} \approx \mathbf{p}(k) - \mathbf{q}(k) + \frac{\lambda^2 \tau^2}{K \|\mathbf{f}_t\|^2}(\mathbf{z}_k - \mathbf{v}_k). \tag{13}$$

It is clear to see that Equation 13 introduces a teacher's supervision $\mathbf{v}_k$ to balance the influence from the one-hot label and $\|\mathbf{f}_t\|$ helps to control the contribution of $\mathbf{z}_k - \mathbf{v}_k$ which represents how different the student is from teacher.

Comparing Equation 13 and its counterpart in KD [11], we could find the biggest difference is the presence of $\|\mathbf{f}_t\|$ which assigns each sample a different weights between one-hot label and teacher instead of a same weight given by standard KD. When the noise in $\mathbf{q}(k)$ is strong, $\|\mathbf{f}_t\|$ is going to be small and the impact of $\mathbf{z}_k - \mathbf{v}_k$ will be relatively higher. In turn, if the one-hot label has less noise, $\|\mathbf{f}_t\|^2$ would reduce the effect of teacher to negligible and $\mathbf{p}(k) - \mathbf{q}(k)$ dominates. In brief, the feature normalized KD could determine adaptively how much a student needs to trust the teacher according to the intensity of label noise in examples.

As suggested by Hinton et al. [11], it is important to ensure that the relative contributions of the hard and soft targets remain roughly the same order of magnitude. Therefore, we keep $\lambda^2 \tau^2 / (\|\bar{\mathbf{f}}_t\|)^2 \approx 1$, where $\|\bar{\mathbf{f}}_t\|$ is the mean over the training set. To fairly compare with standard KD, we then let $\|\bar{\mathbf{f}}_t\|/\tau \approx T$. The specific parameter setting will be introduced in the following experiments section.

4 Experiments

In this section, we will conduct extensive experiments on Cifar-100, Cifar-10, CUB-200-2011 and Stanford Cars datasets to verify the effectiveness of our proposed feature normalized knowledge distillation (KD-fn) for varying image classification tasks. On this basis, we will further compare KD-fn with standard Knowledge Distillation (KD) and Hypersphere Embedding (HE) and discuss the relationship between them.

4.1 Results on Cifar

Cifar-100 and Cifar-10 are two widely studied datasets for image classification, both of which consist of 60000 32×32 color images. Cifar-100 involves 20 superclasses, each of which contains 5 subclasses, for a total of 100 classes while Cifar-10 only contains 10 distinct categories. To compare our approach with standard KD, we respectively utilize Resnet110 and Resnet56 as the teacher models to supervise the student models of Resnet56 and Resnet20. All models are trained on a single NVIDIA GeForce 1080Ti GPU using MXNet [3] for 200 epochs on the training set with 128 examples per minibatch, and evaluated on the test set. Learning rates start at 0.1 and are divided by 10 after 100 and 150 epochs for all models. We use SGD optimizer with momentum of 0.9, and weight decay is set to 1e-4. In addition to mean subtraction and standard deviation normalization, no more data augmentation or training strategies are used. We set $\alpha = 0.5$, $T = 3$ for KD and $\tau = 2$, $\lambda = 3$ for our method.

Table 2. Results on Cifar-100 and Ciar-10

Model	Method	Cifar-100	Cifar-10
Resnet56	Teacher	81.73	93.63
Resnet20	Student	78.30	92.11
	KD	80.34	**92.86**
	KD-fn	**81.19**	92.67
Resnet110	Teacher	82.01	94.29
Resnet56	Student	81.73	93.63
	KD	81.82	94.10
	KD-fn	**82.23**	**94.14**

As represented by the scores in Table 2, our proposed KD-fn achieves 81.19% high accuracy with Resnet20 student which surpasses KD with remarkable 0.85% in Cifar-100 set. When Resnet110 is utilized as the teacher, KD only brings student tiny improvement, by contrast, the student trained with KD-fn still achieves a promotion of 0.5% accuracy and even outperforms the teacher surprisingly. In Cifar-10, KD and KD-fn have very similar performances and both of them bring

relatively less improvement when compared with results of Cifar-100. According to the description of this dataset, the classes in Cifar-10 are completely mutually exclusive. There is no overlap and less visual similarity between these 10 classes, which means one-hot label is already a good approximation for real distribution $\hat{\mathbf{q}}(k)$ and introduces less noise. Since both KD and KD-fn could be regarded as the method to suppress the noise in label based on our previous discussion, they perform less effectively in Cifar-10 relative to Cifar-100.

4.2 Results on Fine-grained Visual Categorization

To verify the performance of our proposed KD-fn on fine-grained visual categorization (FGVC) tasks, we conduct experiments on two classical benchmarks of CUB-200-2011 [34] and Stanford Cars [15]. The CUB-200-2011 is a most widely studied bird's classification task with 5994 training images and 5794 test images from 200 wild bird species. It is one of the most competitive datasets since each category has only 30 images for training. The Stanford Cars dataset contains 8,144 training images and 8,041 test images over 196 classes.

During training, we set the batchsize to 72 and the initial learning rate as 0.05 with decay factor of 0.1 after every 30 epochs to train each model for 120 epoches. We use random cropping, brightness jitter and random flip data augmentations which are provided as standard training setting by MXNet [3], please refer to our code for more parameter settings. In the experiments, we set $\alpha = 0.5$, $T = 3$ for KD. Based on the principle of parameter setting discussed in Section 3 as well as the experimental result that $\|\mathbf{f}\|$ is roughly 24, we set $\tau = 8$ and $\lambda = 3$.

Considering that subordinate classes share most of the visual characteristics except for subtle differences in particular regions, the images from different classes always have more similarities than that in general image classification. Therefore, the difference between real distribution $\hat{\mathbf{q}}(k)$ and one-hot label is much larger, i.e. the noise in one-hot label is stronger. In view of this, we could speculate that both KD and KD-fn would perform more effectively on FGVC datasets. This is supported by our experimental results, Table 3 shows that teacher always brings about much promotion (at least 1.95%) in terms of accuracy, which is consistent with our above speculation. Furthermore, this is also an evidence that the effectiveness of KD is the suppression of one-hot label noise.

It is more noteworthy, the proposed KD-fn outperforms KD in all settings. On the CUB-200-2011 with Resnet18 student and Resnet50 teacher, KD-fn achieves 81.52% accuracy which surpasses the KD with remarkable 1.15%. Although it's a little bit lower increasing (0.79%) of accuracy on Stanford Cars, the student eventually surpasses its teacher significantly. Since distilling the knowledge to a lightweight network is common and important in application, we compare the performance of our proposed method with KD to distill the knowledge from Resnet50 to Mobilenetv2 which is a typical lightweight model. It is obvious that our approach also achieves good results and surpasses KD 0.92% and 0.61% respectively on two datasets, which indicates that our approach also applies across model family. Moreover, we have also tested the performance of our method

Table 3. Results on CUB-200-2011 and Stanford Cars

Model	Method	CUB-200-2011	Stanford Cars
Resnet50	Teacher	83.30	90.89
Resnet18	Student	77.23	88.10
	KD	80.37	90.74
	KD-fn	**81.52**	**91.53**
Resnet50	Teacher	83.30	90.89
Mobilenetv2	Student	79.53	87.09
	KD	81.48	90.49
	KD-fn	**82.40**	**91.10**
Resnet152	Teacher	83.76	92.13
Mobilenetv2	KD	81.57	90.71
	KD-fn	**82.80**	**91.42**

when there is a big gap between teacher and student. On CUB-200-2011, it can be seen that when changing the teacher model from Resnet50 to Resnet152, the student's accuracy obtained by KD only increases 0.09% while that trained with KD-fn improves 0.4%.

4.3 Self-distillation

Self-distillation refers to the special KD, where the student and teacher model share the same architecture. The idea is to feed in predictions of the trained model as teacher to provide new target values for retraining itself. When there are constraints on training resources or when it is hard to find a better teacher than the student, self-distillation is an effective option to obtain higher accuracy. From the perspective of label denosing, we argue that since student can also learn knowledge about the noise in one-hot label, it can use this knowledge to suppress the noise's influence and thus improve itself's learning. We compare self KD-fn and self KD on Cifar-100 and CUB-200-2011 with 5 different models. Table 4 shows that self KD-fn surpasses self KD in all settings, which further demonstrates the effectiveness of our proposed method.

Table 4. Results on self-distillation

Dataset	Model	Baseline	self KD	self KD-fn
Cifar-100	Resnet20	78.30	79.48	**80.16**
	Resnet56	81.73	83.33	**83.73**
CUB-200-2011	Resnet18	77.23	79.15	**79.61**
	Resnet50	83.30	83.92	**84.24**
	Mobilenetv2	79.53	80.32	**80.47**

4.4 The Relationship with Hypershpere Embedding

Hypershpere Embedding (HE) [26, 35] is a method to restrict the feature of penultimate layer to lie on a hypersphere of a fixed radius, which is prevalent in face verification. The method usually fine-tune a well-trained model using cross-entropy loss with multiplying the L_2-norm of the feature in penultimate layer by $r/\|\mathbf{f}\|, r \in \mathcal{R}$. We directly write the gradient with respect to logit of this method as

$$\frac{\partial \mathcal{L}_{he}}{\partial \mathbf{z}_k} = \frac{r}{\|\mathbf{f}\|}(\mathbf{p}_{he}(k) - \mathbf{q}(k)) \tag{14}$$

where $\mathbf{p}_{he}(k)$ is the output computed following Equation 10. We can find that this method is similar to KD-fn, which also weights each sample by $\|\mathbf{f}\|$ to pay more attention to those hard examples with much label noise. The difference with our method is that the $\mathbf{p}_{he}(k)$'s objective is still one-hot label while $\mathbf{p}_{fn}(k)$'s objective is another teacher. So, KD-fn combines the strengths of both teacher and $\|\mathbf{f}\|$ at the same time. Table 5 shows the comparison results. Due to the presence of teacher, KD-fn outperforms HE significantly.

Table 5. Comparison to Hypershpere Embedding

Dataset	Model (S+T)	HE (S)	KD-fn (S+T)
Cifar-100	Resnet20+Resnet56	78.96	**81.19**
	Resnet56+Resnet110	81.65	**82.23**
CUB-200-2011	Resnet18+Resnet50	80.25	**81.52**
	Mobilenetv2+Resnet50	80.55	**82.40**

5 Conclusions

In this paper, we propose a simple yet effective feature normalized distillation strategy for image classification. Based on the systematically analysis from the perspective of label denoising, we introduce the L_2-norm of the feature in penultimate layer into soft-target as the sample specific correction factor to replace the unified temperature of KD for better reducing the impact of noise in one-hot label. Comprehensive experiments show that the proposed method surpasses standard KD significantly on Cifar-100, CUB-200-2011 and Stanford Cars datasets.

6 Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant 2018YFE0202800, Natural Science Foundation of China (NSFC) (Grant Nos. 61674120, U1709218, 61672131), and JST, ACT-X Grant Number JPMJAX190D, Japan.

References

1. Chen, C., Liu, X., Qiu, T., Sangaiah, A.K.: A short-term traffic prediction model in the vehicular cyber-physical systems. Future Gener. Comput. Syst. **105**, 894–903 (2020)
2. Chen, C.C., J, H., and, Q.T.: Cvcg: Cooperative v2v-aided transmission scheme based on coalitional game for popular content distribution in vehicular ad-hoc networks. IEEE Transactions on Mobile Computing **18(12):2811-2828** (2019)
3. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z.: Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. CoRR **abs/1512.01274** (2015)
4. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 248–255 (2009)
5. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (2019)
6. Frénay, B., Verleysen, M.: Classification in the presence of label noise: A survey. IEEE Trans. Neural Netw. Learning Syst. **25**(5), 845–869 (2014)
7. Guan, J., Lai, R., Xiong, A.: Wavelet deep neural network for stripe noise removal. IEEE Access **7**, 44544–44554 (2019)
8. Guan, J., Lai, R., Xiong, A., Liu, Z., Gu, L.: Fixed pattern noise reduction for infrared images based on cascade residual attention CNN. Neurocomputing **377**, 301–313 (2020)
9. Han, S., Pool, J., Tran, J., Dally, W.J.: Learning both weights and connections for efficient neural network. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. pp. 1135–1143 (2015)
10. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV. vol. 9908, pp. 630–645 (2016)
11. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015)
12. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017)
13. Hui, Z., Wang, X., Gao, X.: Fast and accurate single image super-resolution via information distillation network. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 723–731 (2018)
14. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M.J., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 1931–1939 (2015)
15. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013)

16. Lai, R., Li, Y., Guan, J., Xiong, A.: Multi-scale visual attention deep convolutional neural network for multi-focus image fusion. IEEE Access **7**, 114385–114399 (2019)
17. Lan, X., Zhu, X., Gong, S.: Knowledge distillation by on-the-fly native ensemble. In: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada. pp. 7528–7538 (2018)
18. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.: Learning from noisy labels with distillation. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 1928–1936 (2017)
19. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. In: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)
20. Ma, N., Zhang, X., Zheng, H., Sun, J.: Shufflenet V2: practical guidelines for efficient CNN architecture design. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV. vol. 11218, pp. 122–138 (2018)
21. Mirzadeh, S., Farajtabar, M., Li, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. CoRR **abs/1902.03393** (2019)
22. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada. pp. 4696–4705 (2019)
23. Murphy, K.P.: Machine learning - a probabilistic perspective. MIT Press (2012)
24. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 3967–3976 (2019)
25. Phuong, M., Lampert, C.: Towards understanding knowledge distillation. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. vol. 97, pp. 5142–5151 (2019)
26. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. CoRR **abs/1703.09507** (2017)
27. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
28. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 4510–4520. IEEE Computer Society (2018)
29. Sun, S., Cheng, Y., Gan, Z., Liu, J.: Patient knowledge distillation for BERT model compression. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 4322–4331 (2019)
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2818–2826 (2016)

31. Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V.: Mnasnet: Platform-aware neural architecture search for mobile. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 2820–2828 (2019)
32. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. vol. 97, pp. 6105–6114 (2019)
33. Tang, J., Shivanna, R., Zhao, Z., Lin, D., Singh, A., Chi, E.H., Jain, S.: Understanding and improving knowledge distillation. CoRR **abs/2002.03532** (2020)
34. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
35. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L_2 hypersphere embedding for face verification. In: Liu, Q., Lienhart, R., Wang, H., Chen, S.K., Boll, S., Chen, Y.P., Friedland, G., Li, J., Yan, S. (eds.) Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017. pp. 1041–1049 (2017)
36. Yang, C., Xie, L., Su, C., Yuille, A.L.: Snapshot distillation: Teacher-student optimization in one generation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 2859–2868 (2019)
37. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 7130–7138 (2017)
38. Yuan, L., Tay, F.E.H., Li, G., Wang, T., Feng, J.: Revisit knowledge distillation: a teacher-free framework. CoRR **abs/1909.11723** (2019)
39. Zheng, H., Fu, J., Zha, Z., Luo, J.: Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 5012–5021 (2019)