Learning from Noisy Labels with Complementary Loss Functions

Deng-Bao Wang,[1,2] **Yong Wen,**[3] **Lujia Pan,**[3,4] **Min-Ling Zhang**[1,2,5*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
[3]Noah's Ark Lab, Huawei Technologies
[4]NSKEYLAB, Xi'an Jiaotong University
[5]Collaborative Innovation Center of Wireless Communications Technology, China
wangdb@seu.edu.cn, {wenyong4, panlujia}@huawei.com, zhangml@seu.edu.cn

Abstract

Recent researches reveal that deep neural networks are sensitive to label noises hence leading to poor generalization performance in some tasks. Although different robust loss functions have been proposed to remedy this issue, they suffer from an underfitting problem, thus are not sufficient to learn accurate models. On the other hand, the commonly used Cross Entropy (CE) loss, which shows high performance in standard supervised learning (with clean supervision), is non-robust to label noise. In this paper, we propose a general framework to learn robust deep neural networks with complementary loss functions. In our framework, CE and robust loss play complementary roles in a joint learning objective as per their learning sufficiency and robustness properties respectively. Specifically, we find that by exploiting the memorization effect of neural networks, we can easily filter out a proportion of hard samples and generate reliable pseudo labels for easy samples, and thus reduce the label noise to a quite low level. Then, we simply learn with CE on pseudo supervision and robust loss on original noisy supervision. In this procedure, CE can guarantee the sufficiency of optimization while the robust loss can be regarded as the supplement. Experimental results on benchmark classification datasets indicate that the proposed method helps achieve robust and sufficient deep neural network training simultaneously.

Introduction

With highly efficient stochastic optimization methods and loss functions, deep neural networks (DNNs) have been shown to be very powerful modeling tools for many real-world learning tasks involving complex input patterns. However, current deep neural networks have been shown to be sensitive to label noises and can easily overfit noisy labels in training, leading to poor performance in generalization (Zhang et al. 2017). Moreover, labeling large-scale datasets is costly in terms of expense and time, and stands as a critical bottleneck in many tasks. For this reason, learning from less expensive labeled data has been extensively studied in the last decades (Zhou 2017).

In recent years, learning from noisy labels has been extensively studied and a number of methods have been proposed. They can be grouped into four main categories. The first one is based on estimating the label transition matrix, which reflects the probabilities that most probable true labels flip into other noise ones (Patrini et al. 2017; Hendrycks et al. 2018; Han et al. 2018a). The second type is based on importance reweighting which tries to assign small weights to the possibly mislabeled samples and large weights to the potentially clean samples (Jiang et al. 2018; Ren et al. 2018; Shu et al. 2019). The third one is based on self/co-training strategy which tries to learn a classifier from the supervision generated by the classifier itself or its peer classifier (Han et al. 2018b; Yu et al. 2019; Li, Socher, and Hoi 2020; Li, Huang, and Chen 2021).

The fourth type of approaches is based on the robust loss functions. In the last several years, different researches have been studied to leverage proper loss function for robust DNN learning. Compared to the other three types of methods, using robust loss functions is a simpler and arguably more generic solution. Mean Absolute Error (MAE), as a symmetric loss function, has been theoretically proved robust to label noise (Ghosh, Kumar, and Sastry 2017). In (Ma et al. 2020), the authors provide theoretical insights that a simple normalization can make any loss function robust to noisy labels. However, it has been found empirically that the robust losses usually lead to underfitting problem and hence are not able to achieve good performance (Zhang and Sabuncu 2018; Ma et al. 2020). On the other hand, the commonly used Cross Entropy (CE) has the superiority of learning sufficiency while it is not noise-tolerant. Motivated by this, several researches have been studied for seeking a generalized mixture of MAE and CE to balance their weakness and superiority (Zhang and Sabuncu 2018; Feng et al. 2020). Whilst these loss functions have demonstrated improved robustness and learning sufficiency at same time, they are only partially robust to noisy labels and unsatisfactory when dealing with complex datasets.

In this paper, we show that the dilemma between overfitting and underfitting in learning from noisy labels can be addressed with complementary loss functions. Our main idea is to learn DNNs with both CE and a noise-tolerant loss (like MAE) from mixed supervision. The mixed supervision consists of two part: one is the original noisy

*Corresponding author

supervision which guides the robust learning with noise-tolerant loss, and another one is the pseudo supervision which is generated by the model itself. Specifically, we find that by exploiting the memorization effect of DNNs, we can easily filter out a proportion of hard samples and generate reliable pseudo labels for easy samples, and thus reduce the label noise to a quite low level. Then, we learn the classifier with CE on the pseudo supervision to guarantee the sufficiency of optimization, while regard the original noisy supervision as the supplement and learn from this noisy supervision with robust loss functions. In our framework, non-robust loss and robust loss play complementary roles in a joint learning objective with their learning sufficiency and robustness properties respectively. We evaluate our method on benchmarks and empirically demonstrate that the complementary loss outperforms other loss functions by considerable margins. Furthermore, we push the state-of-the-art on noise-label learning one step forward by combining our method with *mixup* (Zhang et al. 2018) data augmentation.

Related Work

In this section, we briefly review existing approaches for robust learning with noisy labels.

Learning with Noise Transition Goldberger and Ben-Reuven (2017) proposed to model the noise transition by adding an additional linear layer on top of the neural network that connects the correct labels to the noisy ones. Patrini et al. (2017) proposed a loss correction method based on pre-calculated *Backward* or *Forward* noise transition matrix, which are obtained by exploiting anchor points (i.e., data points that belong to a specific class almost surely). The method proposed in (Hendrycks et al. 2018) assumes that the model has access to a small set of clean samples to estimate the noise transition matrix and proposes *Gold Loss Correction* (GLC) which corrects the loss function with the estimated transition matrix. Xia et al. (2019) proposed a transition revision method to effectively learn transition matrices from noisy data without employing anchor points.

Sample Reweighting By assigning small weights to the possibly mislabeled samples and large weights to the potentially clean samples, importance reweighting strategy can filter out noisy labels and hence guarantee robust model training. Self-paced learning (Kumar, Packer, and Koller 2010) and its extensions (Jiang et al. 2014) specify the weighting function as monotonically decreasing so the classifier can focus on the easy samples first and then fit the hard samples. This weighting scheme has been shown to be helpful to address the overfitting problem in noise-label learning. Jiang et al. (2018) proposed to use a learned Mentor-Net to output the weights of the examples to teach the Student-Net based on the training loss. Ren et al. (2018) proposed a meta-learning algorithm that learns to assign weights to training examples based on their gradient directions. Then, Shu et al. (2019) proposed to automatically learn an explicit loss-weight function, which is parameterized by an MLP, from an additional clean dataset in a meta-learning manner.

Self/Co-Training Self-training (Scudder 1965) and Co-training (Blum and Mitchell 1998) are two of the earliest and simplest strategies in weakly supervised learning, and have been widely revisited recently (He et al. 2020; Xie et al. 2020). Han, Luo, and Wang (2019) proposed a self-training framework to train the network in an end-to-end manner, which iteratively generates the corrected labels by selecting multiple prototypes for each class. Laine and Aila (2017) introduced *self-ensembling* for dealing with semi-supervised learning. In *self-ensembling*, a consensus prediction is formed as the pseudo supervision for each unlabeled sample using the outputs of the network-in-training over different training epochs. Nguyen et al. (2020) extended this idea to noise-label learning and proposed *self-ensemble label filtering* (SELF) to progressively filter out the wrong labels during training. Han et al. (2018b) and Yu et al. (2019) proposed a paradigm called *Co-teaching* which trains two networks simultaneously and let them teach each other. Recent study in (Li, Socher, and Hoi 2020) proposed a novel method named *DivideMix* by leveraging semi-supervised learning techniques in the *Co-teaching* framework. By adapting *mixup* augmentation (Zhang et al. 2018), they achieved state-of-the-art performance on several benchmarks.

Robust Loss Functions Besides the above three kinds of methods, designing robust loss functions that are inherently tolerant to label noise has received increasing attention recently since their simplicity and generality in DNNs training. The pioneering study (Ghosh, Manwani, and Sastry 2015) proved that binary loss functions that satisfy the symmetric condition (i.e., for some positive constant K, $\ell(\hat{y}, 1) + \ell(\hat{y}, -1) = K$), are robust to uniform label noise. In (Ghosh, Kumar, and Sastry 2017), the authors proved that for multi-class classification, the loss functions which satisfy the symmetric condition, would also be inherently tolerant to both uniform and class conditional label noise. For example, Mean Absolute Error (MAE), as a symmetric loss function, has been theoretically proved robust to label noise. However, a recent study (Zhang and Sabuncu 2018) showed that it is not able to achieve good performance by learning DNNs with MAE due to slow convergence caused by gradient saturation. Ma et al. (2020) showed that any loss can be made robust to noisy labels by applying a simple normalization. Unfortunately, the authors also empirically found that simply being robust is not sufficient for a loss function to train accurate DNNs in practice. The Generalized Cross Entropy (GCE) (Zhang and Sabuncu 2018) applies a Box-Cox transformation strategy, which can behave like a generalized mixture of MAE and CE. Wang et al. (2019) proposed the Symmetric Cross Entropy (SCE) by combining Reverse Cross Entropy (RCE) (which satisfies the symmetric condition) together with the CE. By applying Taylor Series, Feng et al. (2020) derived an alternative representation of CE, in which they can flexibly adjust the order of Taylor Series to balance between MAE and CE. Recently, Ma et al. (2020) proposed Active Passive Loss (APL) which combines two robust loss functions namely active loss and passive loss that mutually boost each other.

Methodology

Preliminaries

We firstly consider the problem of ordinary multi-class classification. Let $p(X, Y)$ be the distribution of a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where X denotes the variable of instances, Y the variable of instance, $\mathcal{X}$ the feature space, $\mathcal{Y} = \{1, 2, ..., c\}$ the label space and c the size of labels. The goal of learning is to learn a multi-class classifier f, which is a function that maps the feature space to the label space ($\mathcal{X} \to \mathbb{R}^c$), that minimizes the classification risk: $\mathcal{R}(f) = \mathbb{E}_{p(X,Y)} \ell\left(f(X), Y\right)$, where $\ell : \mathbb{R}^c \times \mathcal{Y} \to \mathbb{R}$ is a loss function that measures the learned classifier. In label-noise learning, the samples are independently drawn from noisy distribution $\bar{p}(X, \bar{Y})$. The goal of label-noise learning is still to find a classifier that minimizes the classification risk: $\bar{\mathcal{R}}(f) = \mathbb{E}_{p(X,\bar{Y})} \bar{\ell}\left(f(X), \bar{Y}\right)$, where $\bar{\ell} : \mathbb{R}^c \times \mathcal{Y} \to \mathbb{R}$ is a proper loss function for learning from noisy labels. In this paper, we consider the common case where the function f is a DNN with the softmax output layer.

Current state-of-the-art methods assume that the label noise is conditionally *independent* to the inputs, i.e., $P(\bar{Y} = j \mid Y = i) = P(\bar{Y} = j \mid Y = i, X = \boldsymbol{x})$. Under the *instance-independent* assumption, label noise is either *uniform* or *class-conditional*. Denote the overall noise rate by $\eta \in [0, 1]$ and the class-wise noise rate from class i to class j by η_{ij}. If $\eta_{ij} = \frac{\eta}{c-1}$ for all $j \neq i$, then the noise is said to be *uniform*, otherwise, the noise is said to be *class-conditional*.

The Dilemma of Choosing Between Loss Functions

Denote $f_j(\boldsymbol{x})$ as the j-th element of $f(\boldsymbol{x})$, and $\boldsymbol{e}_j$ as a one-hot vector with value 1 at the j-th element and 0 otherwise, then CE and MAE can be represented as

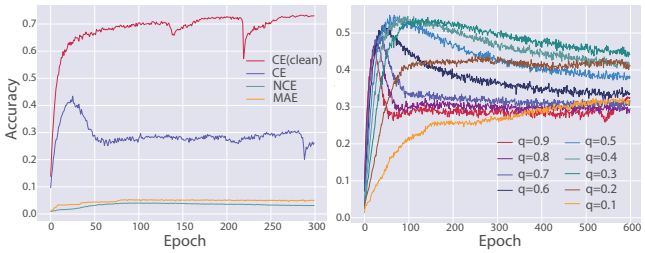
$$\ell_{CE}(f(\boldsymbol{x}), j) = -\boldsymbol{e}_j \log f(\boldsymbol{x}) = -\log f_j(\boldsymbol{x})$$
$$\ell_{MAE}(f(\boldsymbol{x}), j) = \left\|\boldsymbol{e}_j - f(\boldsymbol{x})\right\|_1 = 2 - 2 f_j(\boldsymbol{x})$$

Next, we analyse them from two perspectives: *robustness* and *learning sufficiency*.

On the Robustness Previous work has theoretically proved that for multi-class classification, the loss functions which satisfy the symmetric condition $\sum_{j=1}^{c} \ell(f(\boldsymbol{x}), j) = K$ for some positive constant K, would be inherently tolerant to label noise (Ghosh, Kumar, and Sastry 2017). For CE and MAE, we have

$$\sum_{j=1}^{c} \ell_{CE}(f(\boldsymbol{x}), j) = \sum_{j=1}^{c} \log \frac{1}{f_j(\boldsymbol{x})} \tag{1}$$
$$\sum_{j=1}^{c} \ell_{MAE}(f(\boldsymbol{x}), j) = \sum_{j=1}^{c} (2 - 2 f_j(\boldsymbol{x})) = 2c - 2 \tag{2}$$

Obviously, MAE satisfies symmetry condition while CE does not satisfy. Then based on the theoretical results of (Ghosh, Kumar, and Sastry 2017), MAE is noise-tolerant for both uniform and class-conditional label noise within certain range of noise level. Let $\bar{f}^*$ be the global minimizers of $\bar{\mathcal{R}}(f)$. Robust loss functions like MAE ensure that $\bar{f}^*$ is also the global minimizer of $\mathcal{R}(f)$ under some constraints. Specifically, MAE is robust (1) under uniform label noise with $\eta < \frac{c-1}{c}$, or (2) under class-conditional noise with $\eta_{ij} < \eta_{ii}, \forall i \neq j$ and $\mathcal{R}(f^*) = 0$.



(a) (b)

Figure 1: Test accuracies on CIFAR-100 under 0.5 uniform noise. The accuracy curve on clean data with CE is also plotted in (a).

On the contrary, loss function without the robustness property is sensitive to noisy labels. In this case, as deep networks have large learning capacities, they will eventually overfit the noisy labels (Zhang et al. 2017). In practice, when training DNNs with CE on noisy labeled data, the models usually learn easy patterns and ignore the noisy labels in the beginning epochs of training. After the warmup stage, neural networks tend to memorize the noisy samples to minimize the global risk, and hence result in poor generalization performance.

On the Sufficiency of Learning Although MAE is demonstrated to be theoretically robust to label noise, it has some drawbacks as a classification loss function for training DNNs on large scale datasets with stochastic gradient based techniques. In practice, CE is preferred when training DNNs with clean data since its optimization advantages. The study (Zhang and Sabuncu 2018) suggests that this can be explained in terms of their gradient forms. The gradients of CE and MAE w.r.t. model parameters $\boldsymbol{\theta}$ can be shown as:

$$\frac{\partial \ell_{CE}(f(\boldsymbol{x}), j)}{\partial \boldsymbol{\theta}} = -\frac{1}{f_j(\boldsymbol{x})} \nabla_{\boldsymbol{\theta}} f_j(\boldsymbol{x}) \tag{3}$$
$$\frac{\partial \ell_{MAE}(f(\boldsymbol{x}), j)}{\partial \boldsymbol{\theta}} = -2 \nabla_{\boldsymbol{\theta}} f_j(\boldsymbol{x}) \tag{4}$$

In CE, samples with smaller prediction confidences are weighted more than those with larger ones for gradient update. This means that, when training with CE, the optimizer will pay more attention to the ambiguous samples. This implicit weighting scheme is desirable for training DNNs on clean data. On the contrary, MAE treats all the samples equally. Therefore, there is no sufficient contribution made by those ambiguous samples to the optimization. As a result, this leads to significantly longer training time before convergence and the underfitting problem. In (Ma et al. 2020), the authors provide new theoretical insights that a simple normalization can make any loss function robust to noisy labels. For example, we can easily derive normalized CE loss:

$$\ell_{NCE}(f(\boldsymbol{x}), j) = \frac{\ell_{CE}(f(\boldsymbol{x}), j)}{\sum_{k=1}^{c} \ell_{CE}(f(\boldsymbol{x}), k)} = \frac{\log f_j(\boldsymbol{x})}{\sum_{k=1}^{c} \log f_k(\boldsymbol{x})} \tag{5}$$

However, all the normalized loss functions suffer from similar underfitting issue with MAE, and thus are not sufficient by themselves to train accurate DNNs.

To demonstrate the above discussed dilemma, we train a PreAct-ResNet-18 network (Krause et al. 2016) using CE, NCE and MAE as loss function respectively on CIFAR-100 dataset. As can be observed in Figure 1a: (1) due to the non-robust property, CE starts overfitting to noisy labels at about epoch 30; (2) Although MAE and NCE are robust to label noise, they suffer from severe underfitting problem and thus can not lead to more accurate models. The performance discrepancy of CE and MAE is introduced by the weighting term in Eq. (3). In view of this fact, we conduct another experiments with more fine-grained weights on the gradients. Specifically, we using the following gradient form in model updating:

$$\frac{\partial \ell(f(\boldsymbol{x}), j)}{\partial \boldsymbol{\theta}} = -\frac{1}{(f_j(\boldsymbol{x}))^q} \nabla_{\boldsymbol{\theta}} f_j(\boldsymbol{x}) \tag{6}$$

This is consistent with the gradient form of GCE (Zhang and Sabuncu 2018). With $q \in [0, 1]$, the above form is a generalization of the gradients of CCE and MAE. In Figure 1b, we show different q values applied to the gradient calculation. We find that larger q leads to more severe overfitting, and smaller q leads to more severe underfitting. Therefore, the dilemma of loss functions is actually the dilemma between overfitting and underfitting.

Learning with Complementary Loss Functions

Our main idea is to learn with complementary loss functions (i.e. CE and robust losses like MAE) on mixed supervision. The mixed supervision consists of two parts: one is the original noisy supervision which guides the robust learning with noise-tolerant loss, and another one is the pseudo supervision which is generated by the model itself. To obtain this pseudo supervision, we use the self-ensembling strategy which leverages the outputs of a DNN over different learning iterations. For a specific instance $\boldsymbol{x}$, we simply calculate its average prediction in epoch k as

$$\boldsymbol{p}(\boldsymbol{x}) = \frac{1}{w} \sum_{i=k-w}^{k-1} f^{(i)}(\boldsymbol{x}) \tag{7}$$

where w denotes the window size of ensemble. Then, for the instance of current mini-batch $\mathcal{B}$, we can obtain a pseudo labeled mini-batch $\widetilde{\mathcal{B}}$:

$$\widetilde{\mathcal{B}} = \{(\boldsymbol{x}, \widetilde{y}) | (\boldsymbol{x}, \bar{y}) \in \mathcal{B}\} \tag{8}$$

where the pseudo label $\widetilde{y}$ is generated by assigning the index of the max element of $\boldsymbol{p}$: $\widetilde{y} = \arg\max_{j \in [c]} p_j$.

There will be some wrong ones among all the pseudo labels. Therefore, using all the generated labels as supervision may cause error accumulation problem and hence hurt the model. We remedy this issue by filtering out the samples with low output confidences. The previous study of *Confidence-Aware Learning* suggests that model confidence can be estimated by *True Class Probability* strategy (Corbière et al. 2019). However, the true class is unavailable during training in noise-label learning. In our work, we use entropy-based uncertainty as the measure of output confidence. We argue that the easy samples tend to have stable outputs over different learning iterations and the hard samples tend to have unstable predictions. Therefore, we form a pseudo labeled mini-batch $\widetilde{\mathcal{B}}'$ by filtering out a proportion of hard samples:

$$\widetilde{\mathcal{B}}' = \arg\min_{\mathcal{B}': |\mathcal{B}'| \geq \gamma_1 |\mathcal{B}|} \sum_{(\boldsymbol{x}, y) \in \widetilde{\mathcal{B}}'} H(\boldsymbol{p}(\boldsymbol{x})) \tag{9}$$

where $H(\boldsymbol{p})$ denotes the entropy of distribution $\boldsymbol{p}$, and γ_1 denotes the proportion of easy samples which are selected for guiding the training with CE. Then, we can select a proportion of hard samples in mini-batch $\mathcal{B}$ as the supplement:

$$\bar{\mathcal{B}} = \arg\max_{\mathcal{B}': |\mathcal{B}'| \geq \gamma_2 |\mathcal{B}|} \sum_{(\boldsymbol{x}, y) \in \mathcal{B}'} E(\boldsymbol{p}(\boldsymbol{x})) \tag{10}$$

where γ_2 denotes the proportion of hard samples which are selected for guiding the training with robust loss.

After obtaining $\widetilde{\mathcal{B}}'$ and $\bar{\mathcal{B}}$ for current mini-batch, we propose to combine CE loss function and a robust loss function into an complementary loss framework for both sufficient learning and robustness. Formally, we have

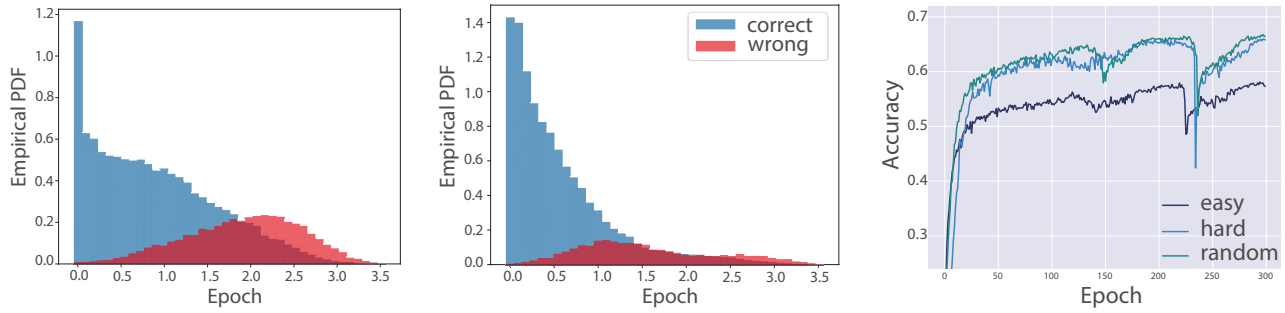
$$\mathcal{L}_{CL} = \sum_{(\boldsymbol{x}, \widetilde{y}) \in \widetilde{\mathcal{B}}'} \ell_{CE}(f(\boldsymbol{x}), \widetilde{y}) + \alpha \sum_{(\boldsymbol{x}, \bar{y}) \in \bar{\mathcal{B}}} \ell_{ROB}(f(\boldsymbol{x}), \bar{y}) \tag{11}$$

where α is used to balance the two terms. Our method is summarized in Algorithm 1. In the beginning epochs of our method, we need to warm up the model by training on original noisy data using CE loss. The warm-up phase is essential for our method, since the model outputs can be significantly unreliable in the beginning epochs of training.

Algorithm 1: Learning with Complementary Losses.

Input: Training set $\mathcal{D}$, robust loss ℓ_{ROB}, T, T_{warm}, α, γ_1, γ_2, batch size B and Optimizer $\mathcal{O}$.
Output: Model parameters $\boldsymbol{\theta}$.

for $t = 1; t \leq T$ **do**
  for $b = 1; b \leq \frac{|\mathcal{D}|}{B}$ **do**
    Fetch mini-batch $\mathcal{B}$ from $\mathcal{D}$;
    if $t < T_{warm}$ **then** ▷ Warm up
      $\mathcal{L} = \sum_{(\boldsymbol{x}, \bar{y}) \in \mathcal{B}} \ell_{CE}(f(\boldsymbol{x}), \bar{y})$;
      Update $\boldsymbol{\theta} = \mathcal{O}(\mathcal{L}, \boldsymbol{\theta})$;
    end
    else ▷ Main train
      Calculate the average prediction for each sample using Eq. (7);
      Obtain $\widetilde{\mathcal{B}}$ using Eq. (8);
      Obtain $\widetilde{\mathcal{B}}'$ using Eq. (9);
      Obtain $\bar{\mathcal{B}}$ using Eq. (10);
      Calculate $\mathcal{L}_{CE} = \sum_{(\boldsymbol{x}, \widetilde{y}) \in \widetilde{\mathcal{B}}'} \ell_{CE}(f(\boldsymbol{x}), \widetilde{y})$;
      Calculate $\mathcal{L}_{ROB} = \sum_{(\boldsymbol{x}, \bar{y}) \in \bar{\mathcal{B}}} \ell_{ROB}(f(\boldsymbol{x}), \bar{y})$;
      $\mathcal{L}_{CL} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{ROB}$;
      Update $\boldsymbol{\theta} = \mathcal{O}(\mathcal{L}_{CL}, \boldsymbol{\theta})$
    end
  end
  Preserve the model outputs of current epoch;
end
return $\boldsymbol{\theta}$.



(a) Epoch 60 (b) Epoch 300 (c) Epoch 300

Figure 2: (a),(b): Entropy distributions of the average predictions of correct and incorrect pseudo labeled samples (trained on CIFAR-100 under 0.5 uniform noise). (c): Test accuracy curves on different subset of CIFAR-100.

Why Complementary Loss Works? Here, we provide some insights for explaining why complementary-loss can address learning sufficiency and robustness simultaneously. We can understand it from the following two aspects.

1. Memorization effect benefits noise reduction.
In our method, CE loss on pseudo labeled samples guarantees the sufficient optimization, hence generating quality pseudo labels is crucial. Fortunately, the memorization effect of deep networks can help us achieve this goal. The memorization effect implies that DNNs learn easy patterns before overfitting the noisy samples. Therefore, in the early training epochs, the outputs of easy samples are usually sharper than those of harder samples. On the other hand, the easy samples tend to have stable outputs over different learning epochs and the outputs of hard samples tend to be unstable. We use an illustrative experiment to demonstrate this point. Figure 2a shows the density of correct pseudo labels and incorrect ones after warm up stage (at epoch 60) according to the entropy of mean outputs. We can see that the correct and incorrect samples are separable according to the entropy value. In particular, after filtering out 30% hard samples with large entropy, the noise rate of the retaining dataset drops to 0.14. Furthermore, in Figure 2b we show the distributions when the model is trained with complementary loss functions after 300 epochs. The proposed strategy can significantly reduce the mislabeling rate and leads to a smaller overlap between the entropy distributions of correct and incorrect samples. In this case, after filtering out 30% hard samples with large entropy, the noise rate of the retained dataset drops to 0.06.

2. Hard samples are important.
From above analysis we see that after filtering out a proportion of hard samples, the label noise can be reduced to a quite low level. However, these discarded hard samples are usually carrying much useful knowledge about the decision boundary, and hence important for improving a classifier. We demonstrate this point with another illustrative experiment. We divide the CIFAR-100 dataset into two subsets, i.e., an easy one and a hard one, with same size according to the entropy of each sample's average prediction, and then retrain a classifier with CE on each subset respectively. We also train another classifier on a subset which consists of randomly selected samples. As is shown in Figure 2c, the model trained on the hard subset and random subset are significantly more accurate than the model trained on the easy subset. Therefore, only using the easy samples is not enough to learn a good classifier. Motivated by this phenomenon, our method treats the original noisy data as the supplement and learns from this supplementary supervision with a robust loss function.

Experiments

Experimental Setup

Datasets To verify the superiority of our approach, we conduct experiments on two commonly used image classification datasets in the literature of noise-label learning: CIFAR-10 and CIFAR-100, consisting of 32x32 color images arranged in 10 and 100 classes, respectively. Both datasets contain 50,000 training and 10,000 test images. We further use TinyImageNet (subset of ImageNet (Deng et al. 2009)) to test the generality of our approach. TinyImageNet contains 200 classes with 100K training images, 10K validation ones with resolution 64x64. Following previous works (Li, Socher, and Hoi 2020; Zhang and Sabuncu 2018), we experiment with two types of label noise: uniform and class-conditional noise. In addition, we also conduct experiment on Clothing1M, which contains 14 classes with 1M real-world noisy training samples.

Comparison Methods We compare our method with multiple baselines. First, we consider 3 noise-tolerant loss functions: GCE, SCE and APL. These losses are introduced in Related Work. We also reported the results of CE. Second, we consider 2 recently proposed state-of-the-art methods M-correction (Arazo et al. 2019) and DivideMix (Li, Socher, and Hoi 2020). M-correction is a loss correction approach that optimizes networks with a dynamically weighted loss by fitting a two-component beta mixture model (BMM) on the loss values. Note that M-correction is specifically designed for uniform noise, thus we only report its results under uniform noise setting. DivideMix is introduced in Related Work. These two methods both adapt *mixup* augmentation.

Implementation The implementation is based on PyTorch

Noise type	Uniform			Class-conditional		
Noise ratio	0.2	0.4	0.6	0.2	0.3	0.4
Standard CE	83.78±0.32	67.73±0.72	47.79±0.82	86.54±0.52	81.86±0.70	76.02±0.93
GCE (2018)	90.44±0.08	88.08±0.13	81.13±0.21	90.30±0.16	88.68±0.10	84.77±0.87
SCE (2019)	91.68±0.17	87.54±0.47	78.88±0.69	89.91±0.58	84.86±0.77	76.52±1.33
APL (2020)	87.79±0.48	79.13±0.75	66.51±1.38	90.14±0.34	83.70±1.08	76.02±1.20
Ours (CE+MAE)	**93.49±0.14**	**92.04±0.20**	**89.37±0.14**	**94.10±0.17**	93.01±0.20	91.52±0.21
Ours (CE+APL)	92.20±0.54	90.47±0.87	88.01±0.55	94.08±0.14	**93.39±0.17**	**91.89±0.19**
M-correction (2019)	93.69±0.32	93.18±0.10	90.59±0.33	——	——	——
DivideMix (2020)	95.23±0.22	93.62±0.15	92.84±0.19	93.20±0.28	92.38±0.24	91.15±0.69
Ours* (CE)	93.62±0.21	93.21±0.17	92.33±0.24	93.83±0.18	92.82±0.22	91.39±0.26
Ours* (CE+MAE)	**95.37±0.09•**	**94.79±0.11•**	**93.59±0.19•**	**94.98±0.14•**	**94.33±0.24•**	**92.18±0.42•**
Ours* (CE+APL)	94.70±0.13	93.80±0.14	92.68±0.27	94.34±0.14	93.38±0.28	91.10±0.40

Table 1: Classification accuracy (%) of each comparing algorithm on CIFAR-10 with different noise types and levels. The results (mean±std) are reported over 3 random runs and we use the last 10 epochs of each run. The best results of first 6 rows (without *mixup*) and last 5 rows (with *mixup*) are boldfaced respectively. In addition, • indicates the best results among all methods.

Noise ratio	Uniform (0.5)	Class-cond. (0.3)
Standard CE	23.16±0.53	43.67±0.22
SCE	23.89±0.32	38.92±0.44
APL	5.43±0.14	Fail
Ours (CE+MAE)	50.59±0.33	51.99±0.29
Ours (CE+APL)	50.67±0.29	51.31±0.19
Ours* (CE)	48.98±0.32	45.88±0.44
Ours* (CE+MAE)	**56.73±0.35**	**57.27±0.40**
Ours* (CE+APL)	55.24±0.39	56.64±0.29

Table 2: Comparison of accuracy (%) on TinyImageNet with both Uniform and Class-conditional noise. The results (mean±std) are reported over the last 10 epochs.

(Paszke et al. 2019) and experiments were carried out with NVIDIA Tesla V100 GPU. We use PreAct-ResNet-18 and train it using SGD with a momentum of 0.9, a weight decay of 0.0005, and a batch size 100 in our experiments. For both CIFAR-10 and CIFAR-100, the network is trained for 300 epochs in which the first 60 epochs are for warming up the networks. In warm-up stage, we use the weighted gradient as shown in Eq.(6) for parameter updating. We set the initial learning rate as 0.02 and reduce it by a factor of 10 after 200 epochs. For other hyperparameters of our method, we simply set $\alpha = 1, \gamma_1 = 0.9$ and $\gamma_2 = 1$ for all cases. For TinyImageNet, the total number of epochs is 200, and the initial learning rate is reduced by a factor of 10 after 100 epochs. We consider two robust loss functions: MAE and APL in our methods. APL consists of an *active* loss function and a *passive* loss function, and in our implementation, we use the combination of NCE (as *active* loss) and RCE (as *passive* loss). We re-implement the comparison methods using the same network architecture with our method. For the loss function methods, we use the same scheme for the learning rate policy and number of epochs. For M-correction and DivideMix, we maintain the schemes reported in their papers. We also implement another version of our method, in which we adapt the same *mixup* and data augmentation techniques used in (Li, Socher, and Hoi 2020), to compare with M-correction and DivideMix. DivideMix trains two networks simultaneously, thus we can choose to average the predictions from both networks in test phase. For fair comparison, we use the prediction from a single network, and we show that we can improve our method by using same averaging strategy at inference phase (See Appendix D[1]).

Experimental Results

Comparison on CIFAR-10/100 Tables 1 and 3 present the comparison results on benchmark datasets CIFAR-10 and CIFAR-100. We can see that: (1) Complementary loss outperforms existing loss functions with large margin, especially when the noise rate is high. (2) By adopting *mixup* technique, our method consistently outperforms current state-of-the-arts. (3) The last three rows of Table 1 and 3 reveal that the robust losses play important roles in complementary loss strategy.

Note that the improvement of complementary loss on CIFAR-10 is relatively smaller than that on CIFAR-100. We postulate this is because the pseudo labels generated by self-ensembling strategy are very accurate (See Appendix C) on CIFAR-10, thus we can obtain high performance by only training on these pseudo labels with CE. As reported in the tables, the combination of CE+MAE outperforms CE+APL, and CE+APL loses to self-ensembling under 0.4

[1]See appendices in supplementary file

Noise type	Uniform			Class-conditional		
Noise ratio	0.2	0.4	0.6	0.2	0.3	0.4
Standard CE	62.82±0.27	49.10±0.37	31.11±0.36	63.43±0.20	54.37±0.30	44.89±0.40
GCE (2018)	69.43±0.20	64.44±0.14	55.96±0.47	68.96±0.20	64.68±0.22	50.67±0.44
SCE (2019)	61.60±0.31	46.21±0.40	30.05±0.73	62.18±0.21	53.67±0.30	44.29±0.39
APL (2020)	70.74±0.33	61.92±0.52	48.64±0.83	63.11±0.45	52.91±0.46	42.48±0.68
Ours (CE+MAE)	**71.63±0.26**	**68.61±0.35**	**62.52±0.22**	**69.99±0.23**	**68.32±0.20**	**65.72±0.28**
Ours (CE+APL)	71.26±0.36	68.50±0.28	62.30±0.23	69.54±0.17	67.71±0.23	65.39±0.52
M-correction (2019)	68.95±0.53	65.43±0.30	59.43±0.36	——	——	——
DivideMix (2020)	74.80±0.28	72.92±0.20	69.38±0.26	74.90±0.41	72.14±0.45	50.79±0.62
Ours* (CE)	71.04±0.40	69.20±0.27	65.09±0.33	67.80±0.37	66.12±0.36	65.19±0.29
Ours* (CE+MAE)	**77.61±0.22•**	**75.81±0.37•**	**72.17±0.30•**	**76.74±0.53•**	**75.32±0.39•**	68.07±0.79
Ours* (CE+APL)	77.21±0.39	75.62±0.27	71.34±0.60	76.50±0.41	74.47±0.50	**68.18±1.16•**

Table 3: Classification accuracy (%) of each comparing algorithm on CIFAR-100 with different noise types and levels. The results (mean±std) are reported over 3 random runs and we use the last 10 epochs of each run. The best results of first 6 rows (without *mixup*) and last 5 rows (with *mixup*) are boldfaced respectively. In addition, • indicates the best results among all methods.

	CE	GCE	SCE	M-corr.	D-mix.	Ours
Acc.	68.80	69.75	71.02	71.00	74.76	73.59

Table 4: Accuracy (%) of different models on real-world noisy dataset Clothing1M. Results of other methods are from original literatures.

class-conditional noise on CIFAR-10. This is consistent with the empirical finding that although APL has been theoretically proved robust to label noise, it leads to overfitting under high noise rate settings in practice. In addition, the hyperparameters of our method are the same across all cases, while DivideMix needs to tune for different noise rates.

Generality of the Proposed Approach To demonstrate that our approach is effective on datasets other than CIFAR data, we report the comparison results on TinyImageNet in Table 2. The comparison results are similar to CIFAR-100: both CE+MAE and CE+APL can significantly outperform self-ensembling and other losses. Note that we use the same network, hyperparameters (except $\gamma_1 = 0.6$) and learning rate policy as with CIFAR. Furthermore, we conduct another experiment on real-world noisy dataset Clothing1M, with comparison results reported in Table 4. In this case, we use ResNet-50 with ImageNet pre-trained weights. Our method falls short of the state-of-the-art 74.76. We think this limitation is because the Clothing1M dataset contains instance-dependent noisy labels, and MAE and APL are not robust to this type of noise.

Conclusion

In this paper, we propose a simple yet powerful method, complementary loss, for learning from noisy labels. It is based on the observation that the theoretically proved robust loss functions suffer from an underfitting problem and CE is non-robust while shows high performance in network optimization. Our method learns DNN with a noise-tolerant loss (like MAE or APL) from the original noisy supervision as well as CE from the pseudo supervision generated by the model itself respectively. In our framework, CE and robust loss play complementary roles in a joint learning objective by exploiting their learning sufficiency and robustness properties. The experiments on CIFAR-10, CIFAR-100 and TinyImageNet show the strengths of our approach. By adopting *mixup*, our method pushes the state-of-the-art on instance-independent noise-label learning one step forward. The experiment on Clothing1M shows some limitations under instance-dependent noise which will be further investigated in future research.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Key R&D Program of China (2018YFB1004300), the China University S&T Innovation Plan Guided by the Ministry of Education, and partially supported by the Collaborative Innovation Center of Novel Software Technology and Industrialization. We thank the Big Data Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2019. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning*, 312–321. Long Beach, CA.

Blum, A.; and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, 92–100. Madison, WI.

Corbière, C.; Thome, N.; Bar-Hen, A.; Cord, M.; and Pérez, P. 2019. Addressing failure prediction by learning model confidence. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 2902–2913. Vancouver, BC: MIT Press.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of 22nd IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. Miami, FL.

Feng, L.; Shu, S.; Lin, Z.; Lv, F.; Li, L.; and An, B. 2020. Can cross entropy loss be robust to label noise? In *Proceedings of the 29th International Joint Conferences on Artificial Intelligence*, 2206–2212. Yokohama, Japan.

Ghosh, A.; Kumar, H.; and Sastry, P. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 1919–1925. San Francisco, CA.

Ghosh, A.; Manwani, N.; and Sastry, P. 2015. Making risk minimization tolerant to label noise. *Neurocomputing* 160: 93–107.

Goldberger, J.; and Ben-Reuven, E. 2017. Training deep neural-networks using a noise adaptation layer. In *Proceedings of 5th International Conference on Learning Representations*. Toulon, France.

Han, B.; Yao, J.; Niu, G.; Zhou, M.; Tsang, I.; Zhang, Y.; and Sugiyama, M. 2018a. Masking: A new perspective of noisy supervision. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 5836–5846. Montreal, QC: MIT Press.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018b. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 8527–8537. Montreal, QC: MIT Press.

Han, J.; Luo, P.; and Wang, X. 2019. Deep self-learning from noisy labels. In *Proceedings of 17th International Conference on Computer Vision*, 5138–5147. Seoul, Korea.

He, J.; Gu, J.; Shen, J.; and Ranzato, M. 2020. Revisiting self-training for neural sequence generation. In *Proceedings of 8th International Conference on Learning Representations*. Addis Ababa, Ethiopia.

Hendrycks, D.; Mazeika, M.; Wilson, D.; and Gimpel, K. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 10456–10465. Montreal, QC: MIT Press.

Jiang, L.; Meng, D.; Yu, S.-I.; Lan, Z.; Shan, S.; and Hauptmann, A. 2014. Self-paced learning with diversity. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, 2078–2086. Montreal, QC: MIT Press.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning*, 2304–2313. Stockholm, Sweden.

Krause, J.; Sapp, B.; Howard, A.; Zhou, H.; Toshev, A.; Duerig, T.; Philbin, J.; and Fei-Fei, L. 2016. The unreasonable effectiveness of noisy data for fine-grained recognition. In *Proceedings of 14th European Conference on Computer Vision*, 301–320. Munich, Germany.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In Lafferty, J. D.; Williams, C. K. I.; Shawe-Taylor, J.; Zemel, R. S.; and Culotta, A., eds., *Advances in Neural Information Processing Systems*, 1189–1197. Vancouver, BC: MIT Press.

Laine, S.; and Aila, T. 2017. Temporal ensembling for semi-supervised learning. In *Proceedings of 5th International Conference on Learning Representations*. Toulon, France.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *Proceedings of 8th International Conference on Learning Representations*. Addis Ababa, Ethiopia.

Li, S.-Y.; Huang, S.-J.; and Chen, S. 2021. Crowdsourcing aggregation with deep Bayesian learning. *Science China Information Sciences* 64(3): 130104.

Ma, X.; Huang, H.; Wang, Y.; Romano, S.; Erfani, S.; and Bailey, J. 2020. Normalized Loss Functions for Deep Learning with Noisy Labels. In *Proceedings of 37th International Conference on Machine Learning*, 6543–6553. Virtual Conference.

Nguyen, D. T.; Mummadi, C. K.; Ngo, T. P. N.; Nguyen, T. H. P.; Beggel, L.; and Brox, T. 2020. Self: Learning to filter noisy labels with self-ensembling. In *Proceedings of 8th International Conference on Learning Representations*. Addis Ababa, Ethiopia.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 8026–8037. Vancouver, BC: MIT Press.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 1944–1952. Honolulu, HI.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning.

In *Proceedings of the 35th International Conference on Machine Learning*, 4334–4343. Stockholm, Sweden.

Scudder, H. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory* 11(3): 363–371.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 1919–1930. Vancouver, BC: MIT Press.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of 17th International Conference on Computer Vision*, 322–330. Seoul, Korea.

Xia, X.; Liu, T.; Wang, N.; Han, B.; Gong, C.; Niu, G.; and Sugiyama, M. 2019. Are Anchor Points Really Indispensable in Label-Noise Learning? In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 6838–6849. Vancouver, BC: MIT Press.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of 33rd IEEE Conference on Computer Vision and Pattern Recognition*, 10687–10698. Seattle, WA.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning*, 7164–7173. Long Beach, CA.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *Proceedings of 5th International Conference on Learning Representations*. Toulon, France.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. Mixup: beyond empirical risk minimization. In *Proceedings of 6th International Conference on Learning Representations*. Vancouver, BC.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 8778–8788. Montreal, QC: MIT Press.

Zhou, Z.-H. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5(1): 44–53.