CAN LABEL-NOISE TRANSITION MATRIX HELP TO IMPROVE SAMPLE SELECTION AND LABEL CORRECTION?

Anonymous authors
Paper under double-blind review

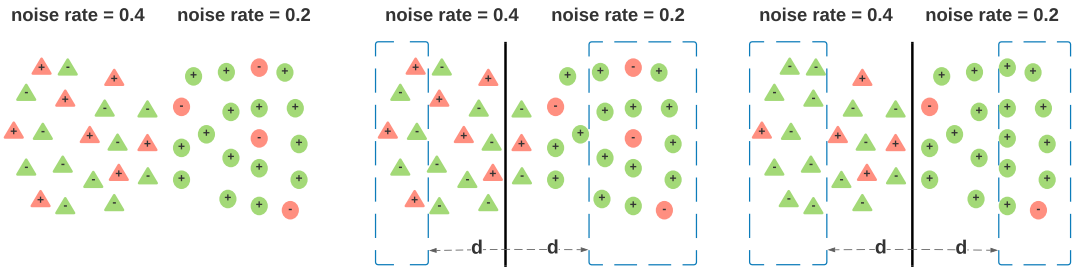
ABSTRACT

Existing methods for learning with noisy labels can be generally divided into two categories: (1) sample selection and label correction based on the memorization effect of neural networks; (2) loss correction with the transition matrix. So far, the two categories of methods have been studied independently because they are designed according to different philosophies, i.e., the memorization effect is a property of the neural networks independent of label noise while the transition matrix is exploited to model the distribution of label noise. In this paper, we take a first step in unifying these two paradigms by showing that modelling the distribution of label noise with the transition matrix can also help sample selection and label correction, which leads to better robustness against different types of noise. More specifically, we first train a network with the loss corrected by the transition matrix and then use the confidence of the estimated clean class posterior from the network to select and re-label instances. Our proposed method demonstrates strong robustness on multiple benchmark datasets under various types of noise.

1 INTRODUCTION

While deep learning has achieved remarkable success in various tasks, it often heavily relies on large-scale human-annotated data. Due to the expensiveness of accurately annotating large datasets, alternative and inexpensive annotating methods have been widely used, e.g., querying search engines with a keyword (Fergus et al., 2010; Schroff et al., 2010), harvesting social media images (Mahajan et al., 2018), etc. However, as a trade-off, these alternative methods have sacrificed the accuracy of annotations for the scale of the dataset. As it has been shown that deep neural networks can easily memorize noisy labels which lead to degenerated classification performance (Zhang et al., 2017), how to robustly learn with noisy labels has attracted a lot of attention in recent years (Li et al., 2019; Nguyen et al., 2019; Liu & Guo, 2020).

To make neural networks robust to label noise, one stream of methods focuses on designing heuristics for sample selection and label correction to reduce the side-effect of noisy labels. Most of these heuristics are designed based on the *memorization effect* of deep neural networks (Arpit et al., 2017), i.e., they would memorize easy instances first, and gradually adapt to hard instances with the increasing amount of training. Inspired by this, many methods use the classification loss on noisy data as the measure of the cleanliness of examples (Jiang et al., 2018; Han et al., 2018; Nguyen et al., 2019; Li et al., 2019; Bai et al., 2021), i.e., an example is likely to be clean if it has a small loss on noisy data. While these methods have shown promising results being combining with different techniques such as warm-up (Xu et al., 2019), co-training (Han et al., 2018), and mixup (Li et al., 2019), they are not guaranteed to be statistically consistent and often need extensive hyperparameter tuning on clean data. Moreover, to achieve high classification accuracy on clean data, some methods need different regularization terms for different types of label noise (Li et al., 2019; Nguyen et al., 2019).



(a) (b) (c)

Figure 1: Circles denote instances with clean positive labels, and triangles denote instances with clean negative labels. Different signs represent different noisy labels. Black lines denote decision boundaries. The example which is far away from the black line is more confident. The confident examples are in the blue dashed box. (a) A binary training dataset contains asymmetric label noise. (b) An illustration of confident examples selected by current sample selection methods based on the small loss on noisy data. The instances (circles) in the class with a smaller noise rate are easier to learn based on the memorization effect. As a result, those instances are more confident and far away from the decision boundary. (c) An illustration of confident examples selected by our method, which are more robust to label noise. By exploiting the transition matrix, the estimated clean class posteriors can be employed to select and relabel confident examples.

Another stream of methods aims to design *classifier-consistent* algorithms, where classifiers learned by exploiting noisy data will asymptotically converge to the optimal classifiers defined on the clean domain (Natarajan et al., 2013; Liu & Tao, 2016; Patrini et al., 2017). To build such algorithms, the noise *transition matrix* $T(\boldsymbol{x})$ has been exploited. Specifically, the transition matrix captures the probability of a clean label flips into a noisy label, i.e., $T_{ij}(\boldsymbol{x}) = P(\tilde{Y} = j|Y = i, X = \boldsymbol{x})$, where X, $\tilde{Y}$ and Y are the random variables of instances/features, noisy labels and clean labels, respectively. Suppose the transition matrix is independent of the instance when conditioning on the clean label, i.e., $P(\tilde{Y} = j|Y = i, X = \boldsymbol{x}) = P(\tilde{Y} = j|Y = i)$, it can be learned under mild conditions (Goldberger & Ben-Reuven, 2017; Scott, 2015; Xia et al., 2019; Li et al., 2021). However, they are not able to achieve satisfactory classification performance compared with the methods leveraging semi-supervised learning techniques (Li et al., 2019; Nguyen et al., 2019).

Currently, these two streams of methods are studied independently according to different philosophies. Sample selection and label correction methods exploited the memorization effect which is a property of the neural network, while loss correction methods focused on the transition between the noisy and clean class distributions. A natural question that arises here is that if one stream of methods can help to improve the other one. The answer is Yes. Intuitively, the first stream of methods employs the classification loss on noisy labels as a measure of the cleanliness. However, this measure is entangled with the noisiness of training data. For example, in Figure 1(a), we illustrate a training dataset that contains asymmetric label noise. Specifically, the noise rate is 0.2 for the clean positive class (circle) and 0.4 for the clean negative class (triangle). Under such circumstances, existing small-loss based methods could select more instances in the class with a lower noise rate as confident examples, which is proved in Section 2. Additionally, the labels of these examples may contain noise and can not be fully trusted. These phenomenons are shown in Figure 1(b), i.e., confident examples selected with the small loss can be class imbalanced and inaccurate if the noise is asymmetric.

To solve these issues, we train a model with the loss corrected by the transition matrix and use the confidence of the estimated clean class posterior as the selection measure instead of the classification loss with noisy labels. With this calibrated measure, we could select some high-confident examples and then relabel them

according to their estimated clean class posteriors. In such a way, the selection measure is disentangled with the noisiness of training data, i.e., examples will be selected solely based on the confidence of the estimated clean class posteriors while the noise is handled by the transition matrix. As shown in Figure 1(c), with the help of the transition matrix, the quality of selected examples can be improved.

The major contribution of this paper includes that 1) we have analysed the property of the sample selection methods based on the small loss on noisy data from a theoretical point of view, which shows that the selected examples could be class imbalanced and inaccurate; 2) We have proposed a calibrated sample selection and label correction method by exploiting the transition matrix; 3) Empirical results on both synthetic and real-world noisy datasets show that our method significantly improves the quality of selected confident examples and classification performance.

The rest of this paper is organized as follows. In Section 2, we introduce the current sample selection methods and analyze the limitation. In Section 3, we introduce our calibrated sample selection and label correction method. Experimental results on both synthetic and real-world datasets are provided in Section 4. Finally, we conclude the paper in Section 5.

2 SAMPLE SELECTION WITHOUT THE TRANSITION MATRIX

Let $\mathcal{D}$ be the distribution of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where the feature space $\mathcal{X} \subseteq \mathbb{R}^d$, the label space $\mathcal{Y} = \{1, 2, \ldots, C\}$ and C is the number of classes. Instead of drawing samples from the underlying distribution $\mathcal{D}$, in the problem of learning with noisy labels, we only have samples $\{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^n$ drawn from the noisy distribution $\tilde{\mathcal{D}}$, i.e., the distribution of the noisy random pair $(X, \tilde{Y}) \in \mathcal{X} \times \mathcal{Y}$.

Confident Examples and Sample Selection. Let $P_{\hat{\theta}}(\tilde{Y}|X)$ denote the estimated noisy class posteriors parameterized by $\hat{\theta}$ learned from noisy training data. Typically, the objective of existing methods based on small-loss sample selection is formulated as follows (Jiang et al., 2018; Han et al., 2018):

$$\mathcal{L}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} v_i \ell(P_{\hat{\theta}}(\tilde{Y}|\boldsymbol{x}_i), \tilde{y}_i) = \frac{1}{n}\sum_{i=1}^{n} -v_i \log(P_{\hat{\theta}}(\tilde{Y} = \tilde{y}_i|\boldsymbol{x}_i)), \tag{1}$$

where ℓ is the cross-entropy loss and $v_i \in [0, 1]$ is the per-instance weight. The idea is that if the given label $\tilde{y}_i$ from a training data pair $(\boldsymbol{x}_i, \tilde{y}_i)$ is likely to be clean, then v_i should be equal or close to 1, so that it contributes more than those data pairs whose labels are likely to be incorrect.

To find the weight v_i for each instance, based on the memorization effect, one popular criterion is using the classification loss on noisy data:

$$v_i = \mathbb{1}(\ell_i \leq \lambda) = \mathbb{1}(\ell(P_{\hat{\theta}}(\tilde{Y}|\boldsymbol{x}_i), \tilde{y}_i) \leq \lambda), \tag{2}$$

where $\mathbb{1}$ is the indicator function, ℓ_i is the loss for instance $\boldsymbol{x}_i$ and λ is the loss threshold. Specifically, if a data pair $(\boldsymbol{x}_i, \tilde{y}_i)$ has a loss smaller than the threshold λ, then it is treated as a "clean" data, and will be selected in training $(v_i^* = 1)$ as a confident example. Otherwise, it will not be selected $(v_i^* = 0)$. For example, Jiang et al. (2018) used a mentor network to select confident examples. Han et al. (2018) maintained two networks that select small-loss instances, where the loss threshold is continuously increased during training so that more instances are dropped when the number of epochs gets large. Except for selecting small-loss instances, some methods reweighted examples so that mislabeled samples contribute less to the loss, e.g., Ren et al. (2018) reweighted instances according to their gradient directions. Arazo et al. (2019); Li et al. (2019) calculated per-instance weights by modelling the classification loss distribution with a mixture model.

Bias of Using Small-loss Criterion. Existing methods based on the small loss on noisy data mainly rely on the memorization effect of the deep neural network to select samples. We show that confident examples selected with the small loss on noisy data can be class imbalanced and inaccurate when the clean data is balanced while the noise is asymmetric. This is because, based on the memorization effect, empirically, the instances from a class with a small noise rate tend to be learned "faster" and have smaller losses than examples from a class with a relatively large noise rate. Thus, instances from the class with a small noise rate or low complexity will be too frequently selected and examples from the class with a relatively large noise rate will not be learned well.

We further show that, theoretically, even an optimal hypothesis f^* which perfectly learns the noisy class posterior distribution can be obtained, the small-loss selection criteria still have the bias issue mentioned above. Let loss function ℓ be the widely used cross-entropy loss. Intuitively, the examples with smaller losses are those which have higher confidence on noisy class posteriors (Mohri et al., 2018). Furthermore, the examples from a class with a lower noise rate averagely have higher confidence than examples from other classes. Therefore, the examples in the class with a lower noise rate are more likely to be selected as confident examples than other classes, i.e., the selected examples could be class-imbalanced. Moreover, the selected confident examples should not be treated as "clean" data, because the noisy labels can be different from *Bayes labels* on the clean class-posterior distribution[1]. As a result, the classification accuracy can be degenerated if a model is directly trained with those selected examples. We analyze these problems in Theorem 1 and Theorem 2. To clearly illustrate the relationship between noise type and selection bias, we focus on the binary classification. However, the results can also be extended to a multi-class classification problem, as it can be reduced to several binary classification problems by using the one-vs-rest strategy (Anzai, 2012). We leave all proofs in Appendix A.

Theorem 1. *Let $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ be two examples such that* $\arg\max_{i\in\{0,1\}} P(Y=i|\boldsymbol{x}_1) = \arg\max_{j\in\{0,1\}} P(\tilde{Y}=j|\boldsymbol{x}_1) = 1$, $\arg\max_{i\in\{0,1\}} P(Y=i|\boldsymbol{x}_2) = \arg\max_{j\in\{0,1\}} P(\tilde{Y}=j|\boldsymbol{x}_2) = 0$, *and* $P(Y=0|\boldsymbol{x}_2) = P(Y=1|\boldsymbol{x}_1)$. *If* $P(\tilde{Y}=1|Y=0) - P(\tilde{Y}=0|Y=1) > 0$, *then* $\min_{i\in\{0,1\}} \ell(f^*(\boldsymbol{x}_2), i) > \min_{i\in\{0,1\}} \ell(f^*(\boldsymbol{x}_1), i)$.

Specifically, the above theorem shows that given two examples having the same confidence on clean class posterior distribution and asymmetric noise, the instance $\boldsymbol{x}_1$ from the class with a lower noise rate $P(\tilde{Y}=0|Y=1)$ could have a smaller loss than the instance $\boldsymbol{x}_2$ from the other class with higher noise rate $P(\tilde{Y}=1|Y=0)$. Therefore, the examples in the class with lower noise rate are more likely to be selected as confident examples than the other class which could cause the class-imbalanced issue.

Theorem 2. *When $P(\tilde{Y}=1|Y=0) - P(\tilde{Y}=0|Y=1) > 0$, if an example x_1 such that* $0.5 < P(Y=0|\boldsymbol{x}_1) < \frac{(1-2P(\tilde{Y}=0|Y=1))}{(1-2P(\tilde{Y}=1|Y=0))} P(Y=1|\boldsymbol{x}_1)$, *then* $P(\tilde{Y}=1|\boldsymbol{x}_1) > 0.5$.

Theorem 2 shows that the largest clean and noisy class posteriors of an instance may not be identical if the noise is asymmetric. Under such circumstances, the training examples could have different Bayes labels on the clean and noisy class posteriors, respectively. As a result, the confident examples selected by using the small-loss criterion could be inaccurate, because the examples have been treated as "clean" data directly (Jiang et al., 2018; Han et al., 2018).

[1]The Bayes label is the label with the largest class posterior. For example, the Bayes label on the clean class-posterior distribution Y^* of an instance $\boldsymbol{x}$ is defined as $Y^* = \arg\max_{i\in\{0,1\}} P(Y=i|x)$ (Mohri et al., 2018)

3 SAMPLE SELECTION AND LABEL CORRECTION WITH THE TRANSITION MATRIX

In this section, we propose our method named T-SSLC (sample selection and label correction with the transition matrix), a calibrated sample selection and label correction method by exploiting the transition matrix for learning with noisy labels.

3.1 MOTIVATION

To the best of our knowledge, most of existing label-noise learning methods select confident examples based on noisy class posteriors Eq. (2) (Han et al., 2018; Nguyen et al., 2019; Li et al., 2019). As aforementioned, the selected confident examples can be class imbalanced and have a low clean ratio when the training set contains asymmetric noise. As a result, the classification accuracy will be degenerated by using the select confident examples. Additionally, loss correction methods focused on the statistical property of label noise, and it has been shown that the transition matrix can be accurately estimated with anchor points (Patrini et al., 2017) or other similar assumptions (Li et al., 2021). However, they often can not achieve satisfactory classification accuracy on test data without using semi-supervised techniques such as co-training (Han et al., 2018) and mixup (Zhang et al., 2018). To this end, we propose a method that uses the advantage of the loss correction methods to help the sample selection. In such a way, confident examples are selected based on the estimated clean class posteriors.

Specifically, we first train a network with the loss corrected by the transition matrix to mitigate the effect of different types of label noise. Then we use the confidence of the estimated clean class-posterior to select examples. In this way, the selection measure is disentangled with the label noise, and only those examples with confident clean class posterior will be selected with corrected labels. Next, we describe each part of our proposed method.

3.2 METHODOLOGY

Loss Correction. Our method selects confident examples based on the estimated clean class posterior which can be obtained by exploiting the noisy posterior and the transition matrix. Let $P_\theta(\tilde{Y}|X)$ be the noisy class posterior parameterized by θ, and $P_\phi(Y|X)$ be the clean class posterior parameterized by ϕ. We first learn ϕ with the loss corrected by the transition matrix $\boldsymbol{T}$:

$$\mathcal{L}(\phi) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{T}P_\phi(Y|X=\boldsymbol{x}_i), \tilde{y}_i). \tag{3}$$

where ℓ is the cross-entropy loss and the transition matrix $\boldsymbol{T}$ can be estimated beforehand (Patrini et al., 2017; Xia et al., 2019; Yao et al., 2020) or jointly learned with the network (Goldberger & Ben-Reuven, 2017; Li et al., 2021).

Sample Selection and Label Correction. After training, the estimated clean class posterior of an instance $\boldsymbol{x}_i$ can be calculated by $P_{\hat{\phi}}(Y|X=\boldsymbol{x}_i)$. Then, to select instances, instead of using the classification loss $\ell(P_{\hat{\theta}}(\tilde{Y}|X=\boldsymbol{x}_i), \tilde{y}_i)$ on the noisy class-posterior, we use $H(P_{\hat{\phi}}(Y|X=\boldsymbol{x}_i))$ as the selection measure where $H(\cdot)$ denote a function for measuring the confidence on the clean class posterior, i.e., we select an instance if we are confident that the estimated clean class posterior of the instance is correct. The problem remains how to design an appropriate measure of confidence. For classification problems, it is obvious that easy examples are ones whose correct output can be predicted easily (they lie far from the decision boundary or they are close to anchor points). To this end, we use the entropy of the estimated clean class posterior as the confidence measure and our selection criterion can be formulated as follows:

$$v_i = \mathbb{1}(H(P_{\hat{\phi}}(Y|X = \boldsymbol{x}_i)) \leq \beta), \forall i \in [1, n]. \tag{4}$$

where $H(\cdot)$ is the entropy function and β is the selection threshold. Intuitively, an instance whose estimated clean class posterior has entropy smaller than the threshold β will be selected ($v_i = 1$). Otherwise, it will not be selected ($v_i = 0$).

With the proposed criterion, we divide the training data into a labeled set and an unlabeled set. However, since the network is trained with the corrected loss, confident prediction of an instance does not necessarily mean that the label of the instance is clean. Thus, we re-label those selected instances as follows:

$$\hat{y}_i = \arg\max_c P_{\hat{\phi}}(Y = c|X = \boldsymbol{x}_i). \tag{5}$$

3.3 Implementation

Empirically, the clean class-posterior distribution $P_\phi(Y|X)$ can be modeled by a mapping (e.g., neural network) $g_\phi : \mathcal{X} \to \Delta^{C-1}$, where Δ^{C-1} denotes a probability simplex. Given the transition matrix, the model parameter ϕ can be directly estimated from noisy data as follows:

$$\hat{\phi} = \arg\min_\phi \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{T} g_\phi(\boldsymbol{x}_i), \tilde{y}_i). \tag{6}$$

However, the transition matrix $\boldsymbol{T}$ can be unknown and needed to be estimated. In experiments, we assume the transition matrix $\boldsymbol{T}$ is not given, and the state-of-the-art method VolMinNet (Li et al., 2021) is used to estimate $\boldsymbol{T}$. The reasons we use this method are that 1) it is general and can identify the transition matrix under the mildest assumption by far; 2) it is a computationally efficient method that allows us to learn the transition matrix and the noisy class posterior simultaneously. After having the estimated transition matrix $\hat{\boldsymbol{T}}$ and model parameter $\hat{\phi}$, we could re-label the training data to get a confident labeled set S_l as follows:

$$S_{\mathrm{l}} = \{(\boldsymbol{x}_i, \hat{y}_i)|H(g_{\hat{\phi}}(\mathbf{x}_i)) \leq \beta \, , \boldsymbol{x}_i \in S\}. \tag{7}$$

In Section 5, we show that our method significantly improves the quality of selected examples, and therefore, the classification accuracy of existing label-noise learning methods based on sample selection can also be improved by employing our method.

4 Related Works

In this section, we review existing methods in label-noise learning. We divided existing methods for label-noise learning into two categories: heuristic algorithms and statistically consistent algorithms.

Methods in the first category focus on employing heuristics to reduce the side-effect of noisy labels. For example, many methods use a specially designed strategy to select reliable samples (Yu et al., 2019; Han et al., 2018; Malach & Shalev-Shwartz, 2017; Ren et al., 2018; Jiang et al., 2018; Bai et al., 2021) or correct labels (Ma et al., 2018; Kremer et al., 2018; Tanaka et al., 2018; Reed et al., 2015). Although those methods empirically work well, there is not any theoretical guarantee on the consistency of the learned classifier. It is also worth mentioning that label correction used by these methods are based on estimated noisy class posteriors, which is also entangled with the label noise and can be biased under asymmetric noise. Recently, some methods exploiting semi-supervised learning techniques have been proposed to solve the label-noise learning problem like SELF (Nguyen et al., 2019) and DivideMix (Li et al., 2019). These methods are aggregations of multiple techniques such as augmentations, sample selection and multiple networks. Noise

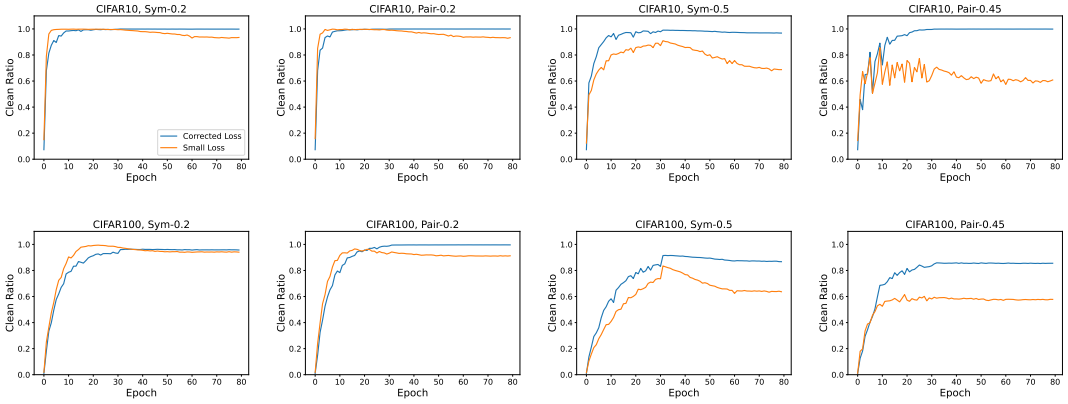


Figure 2: Sample selection on MNIST, CIFAR-10 and CIFAR-100 with different settings of label noise. When the noise rate is small (sym-0.2 and pair-0.2) or symmetric (sym-0.2 and sym-0.5), both methods can effectively select clean labels. With the help of the transition matrix, the proposed method (blue) shows better robustness against asymmetric label noise and high noise rate (pair-0.45 and sym-0.5) compared with existing small-loss sample selection method (orange).

robustness is significantly improved with these methods. Additionally, these methods are sensitive to the choice of hyperparameters.

Statistically consistent algorithms are primarily developed based on a loss correction procedure (Liu & Tao, 2016; Patrini et al., 2017; Zhang & Sabuncu, 2018). For these methods, the noise transition matrix plays a key role in building consistent classifiers. For example, Patrini et al. (2017) leveraged a two-stage training procedure of first estimating the noise transition matrix and then use it to modify the loss to ensure risk consistency. These works rely on anchor points or instances belonging to a specific class with probability one or approximately one. When there are no anchor points, all the aforementioned methods cannot guarantee the statistical consistency. Another approach is to jointly learn the noise transition matrix and classifier. For instance, on top of the softmax layer of the classification network (Goldberger & Ben-Reuven, 2017), a constrained linear layer or a nonlinear softmax layer is added to model the noise transition matrix (Sukhbaatar et al., 2015). Zhang et al. (2021) propose a end-to-end method for estimating the transition matrix and learning a classifier. Specifically, a total variation regularization term is used to prevent the overconfidence problem of the neural network. Li et al. (2021) propose another end-to-end method based on *sufficiently scattered* assumption, which by far the mildest assumption under which the transition matrix is identifiable.

5 Experiments

Datasets. We verify the effectiveness of our approach on the manually corrupted version of two datasets, i.e., *CIFAR10*, *CIFAR100* (Krizhevsky et al., 2009), and one real-world noisy dataset, i.e., *Clothing1M* (Xiao et al., 2015). *CIFAR10* contains 50,000 training images and 10,000 test images. *CIFAR10* and *CIFAR100* both contain 50,000 training images and 10,000 test images but the former have 10 classes of images, and later have 10 classes of images. The two dataset contain clean data, and different types of instance-independent label noise are manually added to the training sets. *Clothing1M* has 1M images with real-world noisy labels and 10k images with clean labels for testing. It also has an additional 50k clean training data

and 14k clean validation data. Note that we only exploit the $1\mathrm{M}$ data for the training and validate our model on the 14k clean validation data. For all the synthetic noisy datasets, the experiments are repeated 5 times.

Noise Types. Following prior works (Nguyen et al., 2019; Li et al., 2021), we conduct experiments with two commonly used types of noise: (1) symmetry flipping (Patrini et al., 2017) which randomly replaces a percentage of labels in the training data with all possible labels; (2) pair flipping (Han et al., 2018) which is a specific type of asymmetric noise, where labels are only replaced by similar classes. It is worth to mention that the noise rate is calculated differently compared with the original paper of DivideMix (Li et al., 2019) because the noise generative process is different. We use the same noise generative process proposed by Han et al. (2018). As a result, for example, pair flipping with 45% noise (pair-45%) in our paper is equivalent to asymmetric noise 50% (Asym-50%) in the paper of DivideMix (Li et al., 2019).

Network Structure and Optimization. For a fair comparison, we implement all methods with default parameters by PyTorch on Nvidia Geforce RTX 3090 GPUs. We use a PreResNet-18 network and PreResNet-32 network for CIFAR10 and CIFAR100, respectively. We use SGD to train the classification network with batch size 128, momentum 0.9, weight decay 10^{-3} and an initial learning rate 10^{-2}, the learning rate is divided by 10 after 40 epochs. The algorithm is run for 80 epochs for the sample selection and relabeling. For clothing1M, we use a ResNet-50 pre-trained on ImageNet. For each epoch, we also ensure the noisy labels for each class are balanced with undersampling.

Baselines. We compare the proposed method with the following methods: (i) Decoupling (Malach & Shalev-Shwartz, 2017), which trains two networks on samples whose predictions from the two networks are different. (ii) MentorNet (Jiang et al., 2018), Co-teaching (Han et al., 2018), which mainly handles noisy labels by training on instances with small loss values. (iii) Forward (Patrini et al., 2017), Reweight (Liu & Tao, 2016), and T-Revision (Xia et al., 2019). These approaches utilize a class-dependent transition matrix T to correct the loss function. (iv) DivideMix (Li et al., 2019) aggregates multiple techniques such as augmentations, multiple networks, and confident example selection. For all baselines, we follow the settings from their original papers.

5.0.1 Clean Ratio Comparison

To illustrate that our proposed method is more effective in selecting clean examples, we compare the clean ratio of the selected examples with the small-loss criteria. Specifically, we train a neural network for 80 epochs on CIFAR10 and CIFAR100 with different settings of label noise, at each epoch, we use our proposed method and small-loss criteria to select 50% examples in the training dataset as confident examples and compare their clean ratio, i.e., the number of selected clean labels divided by the size of the set.

We plot the clean ratio of the selected examples in Figure 2. The results validate that our method is disentangled with the label noise. Specifically, for different noise rates and different types of noise, our method has similar performance, i.e., clean ratios of the selected examples by using our method do not change a lot. However, clean ratios of the selected examples by the small-loss based method dramatically decrease with the increase of label noise.

5.1 Classification accuracy Evaluation

Classification Accuracy on Synthetic Noisy Datasets. To investigate how the sample selection of T-SSLC will affect the classification accuracy in label-noise learning, we embed our sample selection method T-SSLC into the state-of-the-art DividMix (Li et al., 2019) called T-SSLC-DM. We report average test accuracy over the last ten epochs of each model on the clean test set. Higher classification accuracy means that the algorithm is more robust to the label noise. In Table 1, we compare classification accuracies of T-SSLC-DM with dividmix other baseline methods on synthetic noisy datasets. T-SSLC-DM outperforms baseline

	CIFAR-10		CIFAR-100	
	Sym-20%	Sym-50%	Sym-20%	Sym-50%
Decoupling	77.32 ± 0.35	54.07 ± 0.46	41.92 ± 0.49	22.63 ± 0.44
MentorNet	77.42 ± 0.00	61.03 ± 0.20	39.22 ± 0.47	26.48 ± 0.37
Co-teaching	80.65 ± 0.20	73.02 ± 0.23	42.79 ± 0.79	27.97 ± 0.20
Forward	88.21 ± 0.48	77.44 ± 6.89	56.12 ± 0.54	36.88 ± 2.32
T-Revision	90.33 ± 0.52	78.94 ± 2.58	64.33 ± 0.49	41.55 ± 0.95
DMI	87.54 ± 0.20	82.68 ± 0.21	62.65 ± 0.39	52.42 ± 0.64
VolMinNet	$89.58 \pm \pm 0.26$	83.37 ± 0.25	64.94 ± 0.40	53.89 ± 1.26
DivideMix	95.13 ± 0.081	94.59 ± 0.33	74.72 ± 0.25	70.74 ± 0.36
T-SSLC-DM	$\mathbf{95.51 \pm 0.11}$	$\mathbf{94.97 \pm 0.29}$	$\mathbf{75.46 \pm 0.31}$	$\mathbf{72.92 \pm 0.42}$
	CIFAR-10		CIFAR-100	
	Pair-20%	Pair-45%	Pair-20%	Pair-45%
Decoupling	77.12 ± 0.30	53.71 ± 0.99	40.12 ± 0.26	27.97 ± 0.12
MentorNet	77.42 ± 0.00	61.03 ± 0.20	39.22 ± 0.47	26.48 ± 0.37
Co-teaching	80.65 ± 0.20	73.02 ± 0.23	42.79 ± 0.79	27.97 ± 0.20
Forward	88.21 ± 0.48	77.44 ± 6.89	56.12 ± 0.54	36.88 ± 2.32
T-Revision	90.33 ± 0.52	78.94 ± 2.58	64.33 ± 0.49	41.55 ± 0.95
DMI	89.89 ± 0.45	73.15 ± 7.31	59.56 ± 0.73	38.17 ± 2.02
VolMinNet	90.37 ± 0.30	88.54 ± 0.21	68.45 ± 0.69	58.90 ± 0.89
DivideMix	95.72 ± 0.04	87.02 ± 0.41	75.54 ± 0.43	45.20 ± 0.16
T-SSLC-DM	$\mathbf{95.80 \pm 0.05}$	$\mathbf{95.01 \pm 0.01}$	$\mathbf{76.68 \pm 0.25}$	$\mathbf{63.50 \pm 0.19}$

Table 1: Classification accuracy (percentage) on CIFAR-10 and CIFAR-100.

Decoupling	MentorNet	Co-teaching	Forward	T-Revision
54.53	56.79	60.15	71.79	74.18
DMI	VolMinNet	DivideMix	T-SSLC-DM	
72.46	70.12	74.48	**74.92**	

Table 2: Classification accuracy (percentage) on Clothing1M.

methods on almost all settings of noise. This result is natural after we have shown that T-SSLC leads to a high clean ratio of selected examples. These results show the advantage of using the proposed T-SSLC.

Classification Accuracy on Clothing1M. Finally, we show the results on Clothing1M in Table 2. T-SSLC-DM outperforms previous transition matrix based methods and heuristic methods on the Clothing1M dataset. In addition, the performance on the Clothing1M dataset shows that the proposed method has certain robustness against instance-dependent noise as well.

6 Discussion and Conclusion

In this paper, we have proposed a calibrated sample selection and label correction method. We show that the confident examples selected with the small classification loss on noisy data could be class imbalanced and inaccurate. To solve these issues, we first use the transition matrix to estimate the clean class-posterior distribution, then the estimated clean class posterior for each instance is used for sample selection and label correction. Empirical results on both synthetic and real-world noisy datasets show that our method significantly improves the quality of selected confident examples and classification performance.

REPRODUCIBILITY STATEMENT

We have clearly explained of any assumptions for theoretical results. We have included a complete proof of claims in the appendix. The network structures and experiment settings are provided In Section 5. Our source code will be released upon acceptance.

REFERENCES

Yuichiro Anzai. *Pattern recognition and machine learning*. Elsevier, 2012.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pp. 312–321. PMLR, 2019.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, pp. 233–242, 2017.

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *arXiv preprint arXiv:2106.15853*, 2021.

Rob Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. Learning object categories from internet image searches. *Proceedings of the IEEE*, 98(8):1453–1466, 2010.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pp. 8527–8537, 2018.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pp. 2309–2318, 2018.

Jan Kremer, Fei Sha, and Christian Igel. Robust active label correction. In *AISTATS*, pp. 308–316, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2019.

Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. *arXiv preprint arXiv:2102.02400*, 2021.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *ICML*, pp. 6226–6236. PMLR, 2020.

Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, pp. 3361–3370, 2018.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.

Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". In *NeurIPS*, pp. 960–970, 2017.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NeurIPS*, pp. 1196–1204, 2013.

Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *ICLR*, 2019.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pp. 1944–1952, 2017.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR, Workshop Track Proceedings*, 2015.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343. PMLR, 2018.

Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Harvesting image databases from the web. *IEEE transactions on pattern analysis and machine intelligence*, 33(4):754–766, 2010.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*, pp. 838–846, 2015.

Sainbayar Sukhbaatar, Joan Bruna Estrach, Manohar Paluri, Lubomir Bourdev, and Robert Fergus. Training convolutional networks with noisy labels. In *ICLR*, 2015.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pp. 5552–5560, 2018.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, pp. 6835–6846, 2019.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pp. 2691–2699, 2015.

Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, pp. 6222–6233, 2019.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. In *NeurIPS*, 2020.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pp. 7164–7173, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. *arXiv preprint arXiv:2102.02414*, 2021.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pp. 8778–8788, 2018.