LongReMix: Robust Learning with High Confidence Samples in a Noisy Label Environment

Filipe R. Cordeiro [1] Ragav Sachdeva [2] Vasileios Belagiannis [3] Ian Reid [2] Gustavo Carneiro [2]

Abstract

Deep neural network models are robust to a limited amount of label noise, but their ability to memorise noisy labels in high noise rate problems is still an open issue. The most competitive noisy-label learning algorithms rely on a 2-stage process comprising an unsupervised learning to classify training samples as clean or noisy, followed by a semi-supervised learning that minimises the empirical vicinal risk (EVR) using a labelled set formed by samples classified as clean, and an unlabelled set with samples classified as noisy. In this paper, we hypothesise that the generalisation of such 2-stage noisy-label learning methods depends on the precision of the unsupervised classifier and the size of the training set to minimise the EVR. We empirically validate these two hypotheses and propose the new 2-stage noisy-label training algorithm LongReMix. We test LongReMix on the noisy-label benchmarks CIFAR-10, CIFAR-100, WebVision, Clothing1M, and Food101-N. The results show that our LongReMix generalises better than competing approaches, particularly in high label noise problems. Furthermore, our approach achieves state-of-the-art performance in most datasets. The code will be available upon paper acceptance.

1. Introduction

Training Deep Neural Networks (DNNs) often requires large data sets to perform well on challenging problems such as image classification (Litjens et al., 2017). However, the larger the data set, the greater the likelihood for it to be contaminated with noisy labels due to reasons such as low-quality data, human failure, or challenging labelling tasks (Frénay & Verleysen, 2013). The main issue is that DNNs can easily fit noisy labels, particularly for large rates of label noise, reducing their accuracy, as shown by Zhang et al. (Zhang et al., 2016).

In the literature, several methods have been proposed to deal with noisy labels (Kim et al., 2019; Wang et al., 2019b; Ren et al., 2018; Wang et al., 2019b; Nguyen et al., 2019; Li et al., 2020), where the most successful methods explore a 2-stage process formed by an unsupervised learning method to classify training samples as clean or noisy, followed by a semi-supervised learning (SSL) to minimise the empirical vicinal risk (EVR) with a labelled set formed by the samples classified as clean, and an unlabelled set with the samples classified as noisy. The unsupervised learning stage is generally based on the small-loss strategy (Yu et al., 2019), where at every epoch, samples with small loss are classified as clean, and large loss as noisy. This strategy can lead to a low classification precision of clean samples, particularly in high noise rate scenarios, where the loss values can be unstable at different training epochs. The SSL stage (Arazo Sanchez et al., 2019; Nguyen et al., 2019; Li et al., 2020) is usually based on MixMatch (Berthelot et al., 2019) that minimises the empirical vicinal risk (EVR) (Zhang et al., 2017), where a robust estimation of the vicinal distribution is critical for an effective optimisation that generalises well. Such robust estimation depends on a large training set to minimise the EVR (Berthelot et al., 2019; Zhang et al., 2018), but problems with high noise rate usually cause the unsupervised learning stage to build a small training set to be used by this optimisation, affecting the generalisation of the SSL stage.

In this paper, we hypothesise that the generalisation of 2-stage noisy-label learning methods depends on the precision of the unsupervised learning stage to classify clean and noisy samples and a large training set to minimise the EVR at the SSL stage. We empirically validate these two hypotheses and propose a new 2-stage noisy-label training algorithm, called LongReMix. LongReMix is based on a theoretically sound unsupervised learning method to maximise the precision of the clean sample classification by considering the small-loss strategy over a range of epochs instead of a single one. Then, we artificially increase the training set size to improve the generalisation of MixMatch for the minimi-

[1] Universidade Federal Rural de Pernambuco, Recife, Brazil [2] University of Adelaide, Adelaide, Australia [3] Universität Ulm, Ulm, Germany. Correspondence to: Filipe R. Cordeiro <filipe.rolim@ufrpe.br>.

Supported by Australian Research Council through grants DP180103232 and FT190100525.

sation of the EVR during the SSL stage (Berthelot et al., 2019). We evaluate our approach on the noisy-label learning benchmarks of CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), WebVision (Li et al., 2017), Clothing1M (Xiao et al., 2015), and Food101-N (Lee et al., 2018), where LongReMix shows the best performance in the field in almost all of those data sets, particularly in problems with extremely large noise rates. We also show that LongReMix finds a set of clean samples with higher precision than the competing methods, and is robust to over-fitting in problems with high label noise.

2. Prior Work

Several methods have been proposed for the noisy-label problem, and they explore different strategies, such as robust loss functions (Wang et al., 2019a;b), label cleansing (Jaehwan et al., 2019; Yuan et al., 2018), sample weighting (Ren et al., 2018), meta-learning (Han et al., 2018a), ensemble learning (Miao et al., 2015), and others (Yu et al., 2018; Kim et al., 2019; Zhang et al., 2019). Below, we focus on the prior work that is close to our approach and that show competitive results on the main benchmarks. It is important to mention that we do not consider methods that need a clean validation set, such as (Zhang et al., 2020), because we believe this forms a less general experimental setup.

Several approaches explore the sample noise characterisation. Xue et al. (2019) present a probabilistic Local Outlier Factor algorithm (pLOF) to estimate the probability that a sample is an outlier, which is assumed to have label noise. The idea explored by pLOF is that the density around a noisy sample is significantly different from the density around its (clean) neighbors. However, in high noise rate problems, the effectiveness of pLOF is reduced because it cannot find significant differences between the densities of noisy and clean samples. Wang et al. (2018) also use pLOF combined with a Siamese network to increase the dissimilarities between clean and noisy samples. Nevertheless, the incorrect classification of clean samples by pLOF can induce the learning of wrong feature representations. Arazo Sanchez et al. (2019) propose the use of a Beta Mixture Model (BMM) to separate the clean and noise samples during training, based on the loss value of each sample. Similarly, Li et al. (2020) use Gaussian Mixture Model (GMM) for the same goal. Although the use of BMM and GMM applied on the loss values works well for low noise rate, for high noise regimes it becomes less precise. One of the issues affecting the precision of the classification of clean samples from the training set is that they usually rely on an estimation of clean and noisy sets using the loss from the latest training epoch and do not consider the stability of the classification of clean samples over several epochs.

Another technique being studied for noisy-label learning is the use of multiple models to improve the robustness of sample noise characterisation. Han et al. (2018b) propose Co-teaching, which trains two models simultaneously, where each model estimates the clean sample set to be used by the other model. However, with an increase in the number of epochs, both networks converge to a consensus and show little difference between their estimated clean sets. Co-teaching+ (Yu et al., 2019) relies on small loss samples that disagree on the predictions to select the data for the other model. Although this multiple model strategy shows better results for filtering clean samples, noisy samples are usually ignored during training, decreasing the effectiveness of the approach.

After distinguishing between clean and noisy samples, methods either disregard the noisy samples during training (Thulasidasan et al., 2019; Han et al., 2018b), or use both the clean and noisy samples in a semi-supervised learning (SSL) approach (Li et al., 2020; Arazo Sanchez et al., 2019; Sachdeva et al., 2021), where SSL-based methods tend to show better results on benchmarks. One particularly successful technique that relies on SSL is DivideMix (Li et al., 2020) that relies on MixMatch (Berthelot et al., 2019) to linearly combine training samples classified as clean or noisy for the EVR minimisation (Zhang et al., 2017). The generalisation of the EVR minimisation has been theoretically shown to depend on a large training set (Zhang et al., 2018). However, recent methods, such as DivideMix (Li et al., 2020), constrain this training set to be of the same size as the clean set, which tends to be small in large noise rate scenarios. Our approach removes this constraint, allowing a better generalisation of the EVR minimisation.

3. Preliminaries

3.1. Problem Definition

Consider the training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathcal{S}$ is the i^{th} image and $\mathbf{y}_i \in \{0,1\}^{|\mathcal{Y}|}$ is a one-hot vector representing the noisy label, with $\mathcal{Y} \in \{1, ..., |\mathcal{Y}|\}$ denoting the set of labels, and $\sum_{c \in \mathcal{Y}} \mathbf{y}_i(c) = 1$. The label $\mathbf{y}_i$ may differ from the unknown true label $\hat{\mathbf{y}}_i$ as a result of a noise process represented by $\mathbf{y}_i \sim p(\mathbf{y}|\mathbf{x}_i, \mathcal{Y}, \hat{\mathbf{y}}_i)$, with $p(\mathbf{y}(j)|\mathbf{x}_i, \mathcal{Y}, \hat{\mathbf{y}}_i(c)) = \eta_{jc}(\mathbf{x}_i)$, where the $j, c \in \mathcal{Y}$ are the classes, $\eta_{jc}(\mathbf{x}_i) \in [0, 1]$ the probability of flipping the class c to j, and $\sum_{j \in \mathcal{Y}} \eta_{jc}(\mathbf{x}_i) = 1$. We assume that this noise process can be of three types, namely symmetric (Kim et al., 2019), asymmetric (Patrini et al., 2017), and semantic (Lee et al., 2019). The symmetric noise, also called uniform noise, refers to a noise type that the hidden label flips to a random class with a fixed probability η, where the true label is included into the label flipping options, which means that in $\eta_{jc}(\mathbf{x}_i) = \frac{\eta}{|\mathcal{Y}|-1}, \forall j \in \mathcal{Y}$, such that $j \neq c$, and $\eta_{cc}(\mathbf{x}_i) = 1 - \eta$. The asymmetric noise is based on flipping labels between similar classes (Patrini et al., 2017), where

$\eta_{jc}(\mathbf{x}_i)$ depends only on the classes $j, c \in \mathcal{Y}$, but not on $\mathbf{x}_i$. For example, using CIFAR-10 data set (Krizhevsky et al., 2009), the asymmetric noise maps *truck* → *automobile*, *bird* → *plane*, *deer* → *horse*, as mapped by (Zhang & Sabuncu, 2018). The semantic noise (Lee et al., 2019) depends on both the classes $j, c \in \mathcal{Y}$ and the image $\mathbf{x}_i$.

3.2. Background

We only consider 2-stage noisy-label learning approaches (Li et al., 2020; Ding et al., 2018; Kong et al., 2019) that hold state-of-the-art (SOTA) results on all benchmarks – these approaches are based on: 1) an unsupervised learning classifier that characterises training samples as clean or noisy; and 2) a semi-supervised learning classifier that assumes that the training samples classified as clean are labelled, and the samples classified as noisy are unlabelled. The SOTA noise-robust classifier (Li et al., 2020; Nguyen et al., 2019) is formed by an ensemble of two classifiers, each represented by $f : \mathcal{S} \times \Theta \rightarrow [0,1]^{|\mathcal{Y}|}$, where the classifier structure is the same, but their parameters are denoted by $\theta(1), \theta(2) \in \Theta$. The training for $\theta(1)$ influences $\theta(2)$ and vice-versa, where this can be achieved by co-training (Li et al., 2020) or student-teacher (Nguyen et al., 2019) approaches. Our training relies on co-training.

The unsupervised learning classifier predicts the clean and noisy samples based on their loss values (Arazo Sanchez et al., 2019; Li et al., 2020; Lee et al., 2019; Jiang et al., 2020). Formally, assuming that the training is minimising the empirical risk $\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} \ell(f(\mathbf{x}_i;\theta),\mathbf{y}_i)$, the set of clean and noisy samples are respectively defined by

$$\begin{aligned}\mathcal{X} &= \left\{(\mathbf{x}_i,\mathbf{y}_i) : p\left(\text{clean}|\ell_i,\gamma\right) \geq \tau\right\},\\ \mathcal{U} &= \left\{(\mathbf{x}_i,\mathbf{y}_i^*) : p\left(\text{clean}|\ell_i,\gamma\right) < \tau\right\},\end{aligned} \tag{1}$$

where $\mathbf{y}_i^* = f(\mathbf{x}_i;\theta)$, $\ell_i = \ell(f(\mathbf{x}_i;\theta),\mathbf{y}_i)$ represents a classification loss (e.g., cross entropy), and $p\left(\text{clean}|\ell(f(\mathbf{x}_i;\theta),\mathbf{y}_i),\gamma\right)$ is a function that computes the probability that the training sample $(\mathbf{x}_i,\mathbf{y}_i)$ is clean based on its loss ℓ_i (Jiang et al., 2020; Li et al., 2020; Zhang et al., 2020; Nguyen et al., 2019) and parameterised by γ (in this paper, this probability function computes the posterior of the smaller-mean component of a bi-modal GMM, where this smaller mean represents the clean GMM component (Li et al., 2020)). To learn $\theta(1)$ and $\theta(2)$, co-training uses the clean and noisy sets from model $\theta(1)$ to train $\theta(2)$, and vice-versa.

The semi-supervised learning based on MixMatch (Berthelot et al., 2019) mixes the elements of $\mathcal{X}$ and $\mathcal{U}$ to minimise the empirical vicinal risk (EVR) (Zhang et al., 2017):

$$\begin{aligned}\ell_{EVR} = &\frac{1}{|\mathcal{X}'|}\sum_{\substack{(\tilde{\mathbf{x}}_i,\tilde{\mathbf{y}}_i)\\ \in\mathcal{X}'}} \ell^{(\mathcal{X}')}(f(\tilde{\mathbf{x}}_i;\theta),\tilde{\mathbf{y}}_i)+\\ &\frac{\lambda^{(\mathcal{U}')}}{|\mathcal{U}'|}\sum_{\substack{(\tilde{\mathbf{x}}_i,\tilde{\mathbf{y}}_i)\\ \in\mathcal{U}'}} \ell^{(\mathcal{U}')}(f(\tilde{\mathbf{x}}_i;\theta),\tilde{\mathbf{y}}_i),\end{aligned} \tag{2}$$

where $\lambda^{(\mathcal{U}')}$ weights the noisy set loss, $\ell^{(\mathcal{X}')}(.)$ and $\ell^{(\mathcal{U}')}(.)$ denote the losses in the clean and noisy sets, respectively defined as

$$\begin{aligned}\mathcal{X}' &= \{(\tilde{\mathbf{x}}_i,\tilde{\mathbf{y}}_i) : (\tilde{\mathbf{x}}_i,\tilde{\mathbf{y}}_i) \sim v(\tilde{\mathbf{x}},\tilde{\mathbf{y}}|\mathbf{x}_i,\mathbf{y}_i), (\mathbf{x}_i,\mathbf{y}_i) \in \mathcal{X}\}\\ \mathcal{U}' &= \{(\tilde{\mathbf{x}}_i,\tilde{\mathbf{y}}_i) : (\tilde{\mathbf{x}}_i,\tilde{\mathbf{y}}_i) \sim v(\tilde{\mathbf{x}},\tilde{\mathbf{y}}|\mathbf{x}_i,\mathbf{y}_i), (\mathbf{x}_i,\mathbf{y}_i) \in \mathcal{U}\},\end{aligned} \tag{3}$$

with

$$\begin{aligned}v(\tilde{\mathbf{x}},\tilde{\mathbf{y}}|\mathbf{x}_i,\mathbf{y}_i) = &\frac{1}{|\mathcal{X}\cup\mathcal{U}|}\\ &\sum_{\substack{(\mathbf{x}_j,\mathbf{y}_j)\\ \in\mathcal{X}\cup\mathcal{U}}} \mathbb{E}_\lambda\left[\delta\left(\tilde{\mathbf{x}} = \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \tilde{\mathbf{y}} = \lambda\mathbf{y}_i + (1-\lambda)\mathbf{y}_j\right)\right],\end{aligned} \tag{4}$$

where δ is a Dirac mass centered at $(\tilde{x},\tilde{y})$, $\lambda \sim \text{Beta}(\alpha,\alpha)$, and $\alpha \in (0,\infty)$. In (Li et al., 2020), the noisy set size $|\mathcal{U}'|$ and clean set size $|\mathcal{X}'|$ are constrained to be equal to $|\mathcal{X}|$, which means that $|\mathcal{X}'| = |\mathcal{U}'| = |\mathcal{X}|$.

3.3. Our Hypothesis

We hypothesise that the generalisation of 2-stage noisy-label learning methods depend on: 1) the precision of the classification of clean samples to be included in $\mathcal{X}$ in (1), and 2) the size of the clean set denoted by $|\mathcal{X}|$. In particular, a large $|\mathcal{X}|$ with a high proportion of positives will reduce the bound of the difference between the estimated and vicinal risks (Zhang et al., 2018), improving the semi-supervised classification accuracy.

Let us begin with our proposed method to increase the precision in the classification of clean samples in $\mathcal{X}$. Our idea is to classify as clean, the samples that consistently show $p(\text{clean}|\ell_i,\gamma) \geq \tau$ for ζ epochs. Assuming that P_{cc} denotes the probability of classifying a clean sample as clean, P_{nc} the probability of classifying a clean sample as noisy (i.e., $P_{cc} + P_{nc} = 1$). Similarly, P_{nn} represents the probability of classifying a noisy sample as noisy, P_{cn} the probability of classifying a noisy sample as clean (i.e., $P_{nn} + P_{cn} = 1$). Also, P_c and P_n denote the proportion of clean and noisy samples in the training set, with $P_n + P_c = 1$. The probability of a clean sample being in the clean set $\mathcal{X}$ after ζ epochs is P_{cc}^{ζ}, and the probability of a noisy sample being in the clean set after ζ epochs is $P_{cn}^{\zeta} = (1-P_{nn})^{\zeta}$.

Lemma 3.1. *Assuming that* $P_{cc} \in (0.5, 1.0)$ *(so* $P_{nc} \in (0.0, 0.5)$*) and* $P_{nn} \in (0.5, 1.0)$ *(so* $P_{cn} \in (0.0, 0.5)$*), the* *classification precision of clean samples in $\mathcal{X}$ tends to 1 and recall tends to 0, as ζ increases.*

Proof. The precision and recall are calculated with:

$$\begin{aligned} \text{Precision} &= \frac{P^{\zeta}_{cc} \times P_c}{P^{\zeta}_{cc} \times P_c + P^{\zeta}_{cn} \times P_n}, \\ \text{Recall} &= \frac{P^{\zeta}_{cc} \times P_c}{P^{\zeta}_{cc} \times P_c + (1 - P^{\zeta}_{cc}) \times P_c}. \end{aligned} \tag{5}$$

Given the assumption that $P_{cn} \in (0.0, 0.5)$ and $P_{cc} \in (0.5, 1.0)$ and that $\lim_{\zeta \to \infty}(P^{\zeta}_{cn}/P^{\zeta}_{cc}) \to 0$, Precision tends to 1, and similarly, given that $\lim_{\zeta \to \infty}((1 - P^{\zeta}_{cc})/P^{\zeta}_{cc}) \to \infty$, Recall tends to 0. □

In low noise rate problems, P_{cc} tends to be large and P_{cn}, small, so even for small values of ζ, Precision will be close to one with a relatively high Recall in (5), allowing for a large $|\mathcal{X}|$. According to Theorem 8 in (Zhang et al., 2018), a large $|\mathcal{X}|$ will decrease the bound for vicinal risk minimisation. On the other hand, P_{cc} tends to be small and P_{cn} large, in high noise rate scenarios, which means that ζ needs to increase to push the Precision to be close to one, but that can reduce the Recall to very low values, resulting in a potentially small $|\mathcal{X}|$, which will increase the vicinal risk minimisation bound (Zhang et al., 2018). Therefore, ζ is a hyper-parameter that needs to be estimated to achieve a good trade-off between Precision and Recall. Nevertheless, for high noise rate scenarios, even with a careful estimation of ζ, $|\mathcal{X}|$ can still be small. Hence, we propose that $\mathcal{X}$ must be sampled with replacement when mixing up $\mathcal{X}'$ and $\mathcal{U}'$ in (3), such that $|\mathcal{X}'| = |\mathcal{U}'| = |\mathcal{D}|$.

4. LongReMix

Our proposed LongReMix algorithm is divided into two stages (Figure 1). The first stage, comprising the High Confidence Training (HCT), trains the model to find a high confidence set of clean samples with high precision. Next, in the second stage, we build a core set of clean samples using the largest high confidence set obtained from the first stage. With this core set, we retrain the model. Moreover, we propose a new way to build the data sets $\mathcal{X}'$ and $\mathcal{U}'$ in (3), called LongMix, which enables the number of MixUp operations to be proportional to $|\mathcal{D}|$ instead of $|\mathcal{X}|$, as described in Sec. 3.3.

4.1. First Stage: High Confidence Training

The **high confidence training (HCT)** aims to increase the precision of the unsupervised classification of clean and noisy training samples. Following the idea presented in Sec. 3.3, we re-define how to form the sets of clean and noisy samples, originally defined in (1), as follows:

$$\begin{aligned} \mathcal{X}^{(e)}_1 &= \left\{(\mathbf{x}_i, \mathbf{y}_i, w_i) : w_i = p(\text{clean}|\ell^{(e)}_i, \gamma) \geq \tau, \forall e \in \mathcal{E}\right\}, \\ \mathcal{U}^{(e)}_1 &= \left\{(\mathbf{x}_i, \mathbf{y}^*_i, w_i) : w_i = p(\text{clean}|\ell^{(e)}_i, \gamma) < \tau, \exists e \in \mathcal{E}\right\}, \end{aligned} \tag{6}$$

where $\ell^{(e)}_i$ represents the loss of sample $(\mathbf{x}_i, \mathbf{y}_i)$ at training epoch e and $\mathcal{E}$ denotes the confidence window comprising the current and the previous $(\zeta - 1)$ epochs – this is represented by the block "filter" that produces the high confidence samples in Fig. 1. Hence, a sample to be in the clean set $\mathcal{X}^{(e)}_1$ must be classified as clean for ζ epochs in a row, resulting in a more consistent, but smaller, set of clean samples, containing fewer noisy samples than the set in (1).

4.2. Second Stage: Guided Training

The second stage of the training depends on the core set of clean samples estimated from the first training stage with

$$\mathcal{H} = \arg \max_{\mathcal{X}^{(e)}: e \in \{\frac{E}{2}, \ldots, E\}} |\mathcal{X}^{(e)}_1|, \tag{7}$$

where E is the total number of training epochs for the first stage of training. In the second stage of training, we define the labelled and unlabelled sets as in (1), but we use $\mathcal{H}$ to update these sets as follows:

$$\begin{aligned} \mathcal{X}^{(e)}_2 =& \left\{(\mathbf{x}_i, \mathbf{y}_i, w_i) : \begin{cases} w_i = 1, \text{if } (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{H}; \text{ or} \\ w_i = p\left(\text{clean}|\ell^{(e)}_i, \gamma\right) \geq \tau, \text{otherwise} \end{cases}\right\}, \\ \mathcal{U}^{(e)}_2 =& \left\{(\mathbf{x}_i, \mathbf{y}^*_i, w_i) : w_i = p\left(\text{clean}|\ell^{(e)}_i, \gamma\right) < \tau, \right. \\ & \left. \text{and } (\mathbf{x}_i, \mathbf{y}_i) \notin \mathcal{H}\right\}. \end{aligned} \tag{8}$$

During the second stage of LongReMix, we retrain the model from scratch[1] using the core set of clean samples $\mathcal{H}$ from (7) included in the predicted clean and with the original labels from $\mathcal{D}$.

As explained in Sec. 3.3, we hypothesise that by sampling the clean set with replacement, we increase the number of MixUp operations in the EVR loss in (2), resulting in a smaller bound of the difference be-tween estimated and vicinal risks (Zhang et al., 2018). Therefore, we propose **LongMix** that increases the number of MixUp operations to be $|\mathcal{D}|$, instead of the number of predicted clean samples. A criticism faced by LongMix is that adding more MixUp iterations per epoch may be equivalent to a simple increase in the number of epochs, but we show in the experiments that this is not true.

[1] We compared if we should fine-tune the model trained from the first stage or train from scratch, and the latter approach showed the best results.

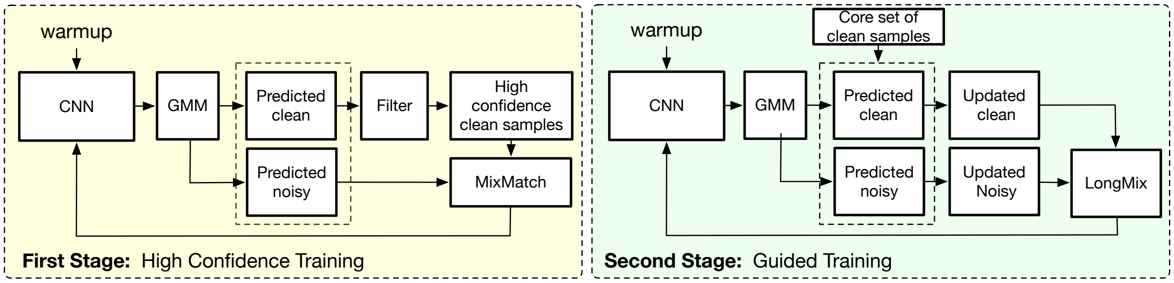


Figure 1. Our **proposed LongReMix method** is composed of two stages: the High Confidence Training and the Guided Training. The High Confidence Training is responsible for finding the high confidence samples. The Guided Training uses the high confidence samples to update the predicted clean set and train the model.

4.3. Training Loss

The training loss for our proposed LongReMix is (Li et al., 2020):

$$\ell = \ell_{EVR} + \lambda^{(reg)}\ell^{(reg)}, \quad (9)$$

where ℓ_{EVR} denotes the empirical vicinal error defined in (2), with $\ell^{(\mathcal{X}')}(f(\tilde{\mathbf{x}}_i;\theta),\tilde{\mathbf{y}}_i) = -\tilde{\mathbf{y}}_i^\top \log(f(\tilde{\mathbf{x}}_i;\theta))$, $\ell^{(\mathcal{U}')}(f(\tilde{\mathbf{x}}_i;\theta),\tilde{\mathbf{y}}_i) = \|\tilde{\mathbf{y}}_i - f(\tilde{\mathbf{x}}_i;\theta)\|_2^2$, $\lambda^{(reg)}$ weights the regularisation loss, and

$$\ell^{(reg)} = KL\left[\pi_{|\mathcal{Y}|}\left|\left|\frac{1}{|\mathcal{X}'|+|\mathcal{U}'|}\sum_{\mathbf{x}\in(\mathcal{X}'\bigcup\mathcal{U}')} f(\mathbf{x};\theta)\right.\right.\right], \quad (10)$$

with $\pi_{|\mathcal{Y}|}$ denoting a vector of $|\mathcal{Y}|$ dimensions with values equal to $1/|\mathcal{Y}|$, and $KL[a||b]$ representing the Kullback Leibler divergence between a and b. The pseudo-code for the training of LongReMix is shown in Algorithm 1 in the supplementary material.

5. Experiments

We compare LongReMix with related approaches on five noisy-label learning benchmarks. We also analyze the performance of LongReMix on a number of ablation studies. All comparisons are performed with the same network architecture and trained for the same number of epochs as the compared methods.

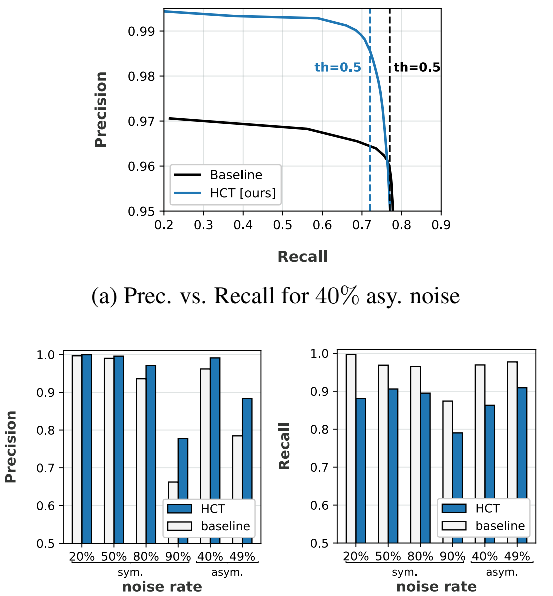
5.1. Data Sets

We conduct our experiments on the data sets CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), Clothing1M (Xiao et al., 2015), WebVision (Li et al., 2017) and Food101-N (Lee et al., 2018). CIFAR-10 and CIFAR-100 have 50000 training and 10000 testing images of size 32 × 32 pixels, where CIFAR-10 has 10 classes and CIFAR-100 has 100 classes and all training and testing sets have a perfectly balanced number of images per classes. As CIFAR-10 and CIFAR-100 data sets originally do not contain label noise, a common approach is to add synthetic noise to evaluate the models. For CIFAR-10/CIFAR-100 we investigated three noise types: symmetric, asymmetric and semantic, as defined in Sec. 3.1. The symmetric noise is generated using $\eta \in \{0.2, 0.5, 0.8, 0.9\}$, with η defined in Sec. 3.1. The asymmetric noise is produced following the mapping used in (Li et al., 2020; Patrini et al., 2017), with $\eta_{jc} \in \{0.4, 0.49\}$ (note that we study $\eta_{jc} = 49\%$ because it is close to the theoretical limit of 50% for this type of noise). We also evaluate the semantic noise scenario, where we follow the setup from (Lee et al., 2019) to generate semantically noisy labels based on a trained VGG (Simonyan & Zisserman, 2015), DenseNet (DN), and ResNet (RN) on CIFAR-10 and CIFAR-100.

Clothing1M consists of 1 million training images acquired from online shopping websites and it is composed of 14 classes.As the images from the data set vary in size, we resized the images to 256 × 256 for training, as used in (Li et al., 2020; Han et al., 2019). The data set is heavily imbalanced and most of the noise is asymmetric (Yi & Wu, 2019), with noise rate estimated to be around 40% (Xiao et al., 2015). The data set provide additional clean sets for training, validation, and test of 50k, 14k and 10k images, respectively. For our experiments we do not use any of the clean training or validation sets, but we use the test set for evaluation.

WebVision contains 2.4 million images collected from the internet, with the same 1000 classes from ILSVRC12 (Deng et al., 2009) and images resized to 256 × 256 pixels. It provides a clean test set of 50k images, with 50 images per class. We compare our model using the first 50 classes of the Google image subset, as used in (Li et al., 2020; Chen et al., 2019).

Food101-N (Lee et al., 2018) contains 310,009 training images of food recipes classified in 101 classes and 25,000 images for the testing set. The images from this data set were resized to 256 × 256. This data set is based on the Food101 data set (Bossard et al., 2014), but it has more images with noisy labels. The test set is the same provided



(a) Prec. vs. Recall for 40% asy. noise

(b) Precision vs. Noise

(c) Recall vs. Noise

Figure 2. (a) Precision versus Recall for our proposed $\mathcal{X}_1^{(e)}$ from (6) (denoted by HCT) and $\mathcal{X}$ from (1) (Baseline), for 40% asymmetric noise on CIFAR-10, where $\tau \in [0, 1]$ (denoted by *th*) for $p(\text{clean}|\ell, \gamma)$. (b) Precision and (c) Recall for different noise rates, for CIFAR-10, using $\tau = 0.5$.

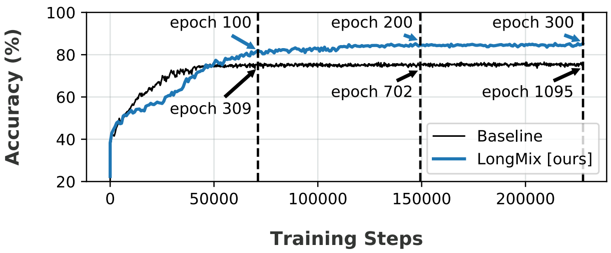


Figure 3. Test accuracy versus number of training steps (or iterations) for CIFAR-10 at 90% symmetric noise for our proposed LongMix (where sizes of $\mathcal{X}'$ and $\mathcal{U}'$ in (3) are $|\mathcal{D}|$) and the baseline (with sizes of $\mathcal{X}'$ and $\mathcal{U}'$ being $|\mathcal{X}|$ (Li et al., 2020)).

by the original Food101 (Bossard et al., 2014), which is a clean test set of 25K images.

5.2. Implementation

The model $f(\mathbf{x}; \theta)$ is represented by a 18-layer PreAct ResNet (PRN18) (He et al., 2016b) for CIFAR-10 and CIFAR-100, InceptionV2 (Szegedy et al., 2017) for WebVision (this is the model used by competing approaches), and ResNet-50 (He et al., 2016a) for Clothing1M and Food-101N. The models are trained with stochastic gradient descent with momentum of 0.8, weight decay of 0.0005 and batch size of 64. The learning rate is 0.02 which is reduced to 0.002 in the middle of the training. The WarmUp and total number of epochs is defined according to each data set, as defined in (Li et al., 2020). For CIFAR-10 and CIFAR-100, PRN18 is based on a WarmUp stage of 30 epochs, with 300 epochs of total training. For WebVision, the InceptionV2 is trained for 100 epochs, with a WarmUp stage of 1 epoch. For Clothing1M, ResNet-50 is trained for 80 epochs with WarmUp stage of 1 epoch. For Food-101N, we also use ResNet-50 and rely on the same training protocol as in (Han et al., 2019), consisting of training for 30 epochs, WarmUp stage of 1 epoch and reducing the learning rate by a factor of 10 every 10 epochs. The MixMatch parameter is $\alpha = 4$ in (4), and the regularisation weight for the loss in (9) is $\lambda^{(reg)} = 1$ for symmetric noise and $\lambda^{(reg)} = 0$ for asymmetric noise–these two parameters are as defined in (Li et al., 2020). We used a confidence window of $\zeta = 5$ in (6), which was defined empirically for all the experiments. In Table 1 of supplementary material we show that, in general, Precision increases and Recall decreases with larger ζ values. Also, classification accuracy reaches a peak for large noise rates (symmetric at 90% and asymmetric $\geq 40\%$) at $\zeta = 5$, and for lower noise rates, accuracy does not change much with different values of $\zeta \in \{1, ..., 10\}$.

5.3. Precision and Recall of the Clean Set

We evaluate the precision and recall of the clean set $\mathcal{X}_1^{(e)}$ from (6) in the last epoch of the first stage of training (HCT), compared to the clean set $\mathcal{X}$ from (1) that relies on the small loss result from the last epoch (Baseline). We assess that by computing Precision $= \frac{\text{TP}}{\text{TP+FP}}$ and Recall $= \frac{\text{TP}}{\text{TP+FN}}$ of the sets $\mathcal{X}_1^{(e)}$ from (6) and $\mathcal{X}$ from (1), where TP refers to the samples correctly predicted as clean, FP denotes the noisy samples incorrectly predicted as clean, and FN denotes the clean samples incorrectly predicted as noisy. Figure 2-(a) shows the Precision vs Recall of predicted clean set for CIFAR-10 with 40% asymmetric noise, where results are obtained by varying the threshold τ applied to $p(\text{clean}|\ell, \gamma)$ to form $\mathcal{X}_1^{(e)}$ and $\mathcal{X}$. We highlight the value of $\tau = 0.5$, which is the default value (Li et al., 2020) that we use to split the clean and noisy samples. Notice that in this highly asymmetric noise scenario, the curve from HCT shows a better trade-off than the Baseline. Figure 2-(b,c) shows that $\mathcal{X}_1^{(e)}$ from HCT trades off a higher precision for a lower recall, compared with $\mathcal{X}$ from the Baseline for several types of noise, As shown below, this has a large influence on the training efficacy of LongReMix.

5.4. LongMix Analysis

Figure 3 shows the test accuracy versus training steps (iterations) for LongMix compared to the baseline (Li et al., 2020), for CIFAR-10 at 90% symmetric noise – this figure shows that adding more MixUp iterations per epoch, as in

Data set		CIFAR-10						CIFAR-100			
Noise type		sym.				asym.		sym.			
Method/ noise ratio		20%	50%	80%	90%	40%	49%	20%	50%	80%	90%
Baseline (Li et al., 2020)	Best	**96.22**	94.93	93.33	76.49	93.24	82.90	78.03	74.87	**62.74**	29.79
	Last	**96.01**	94.68	92.99	75.45	91.79	75.57	77.43	74.23	**62.01**	29.37
LongMix [ours]	Best	96.18	**95.19**	**94.09**	**85.33**	**93.38**	**83.23**	**78.03**	**75.84**	62,24	**33.54**
	Last	95.98	**94.79**	**93.73**	**84.71**	**91.87**	**77.18**	**77.56**	**74.87**	61.60	**33.00**

Table 1. Comparison of the test accuracy between LongMix (where sizes of $\mathcal{X}'$ and $\mathcal{U}'$ in (3) are $|\mathcal{D}|$) and the baseline in (Li et al., 2020) (with sizes of $\mathcal{X}'$ and $\mathcal{U}'$ being $|\mathcal{X}|$), using the same number of iterations on CIFAR-10 and CIFAR-100 under symmetric (ranging from 20% to 90% and asymmetric (ranging from 40% and 49%) noise.

Data set		CIFAR-10						CIFAR-100			
Noise type		sym.				asym.		sym.			
Method/ noise ratio		20%	50%	80%	90%	40%	49%	20%	50%	80%	90%
Cross-Entropy (Li et al., 2020)	Best	86.8	79.4	62.9	42.7	85.0	-	62.0	46.7	19.9	10.1
	Last	82.7	57.9	26.1	16.8	72.3	-	61.8	37.3	8.8	3.5
Coteaching+ (Yu et al., 2019)	Best	89.5	85.7	67.4	47.9	-	-	65.6	51.8	27.9	13.7
	Last	88.2	84.1	45.5	30.1	-	-	64.1	45.3	15.5	8.8
MixUp (Zhang et al., 2017)	Best	95.6	87.1	71.6	52.2	-	-	67.8	57.3	30.8	14.6
	Last	92.3	77.3	46.7	43.9	-	-	66.0	46.6	17.6	8.1
Meta-Learning (Li et al., 2019)	Best	92.9	89.3	77.4	58.7	89.2	-	68.5	59.2	42.4	19.5
	Last	92.0	88.8	76.1	58.3	88.6	-	67.7	58.0	40.1	14.3
M-correction (Arazo Sanchez et al., 2019)	Best	94.0	92.0	86.8	69.1	87.4	-	73.9	66.1	48.2	24.3
	Last	93.8	91.9	86.6	68.7	86.3	-	73.4	65.4	47.6	20.5
DivideMix (Li et al., 2020)	Best	96.1	94.6	93.2	76.0	93.4	83.7	77.3	74.6	60.2	31.5
	Last	95.7	94.4	92.9	75.4	92.1	**76.3**	76.9	74.2	59.6	31.0
LongReMix [ours]	Best	**96.2**	**95.0**	**93.9**	**82.0**	**94.7**	**84.7**	**77.8**	**75.6**	**62.9**	**33.8**
	Last	**96.0**	**94.7**	**93.4**	**81.3**	**94.3**	76.1	**77.5**	**75.1**	**62.3**	**33.2**

Table 2. Results using PRN18 on CIFAR-10 and CIFAR-100 under symmetric (ranging from 20% to 90% and asymmetric (ranging from 40% and 49%) noises. Results from related approaches are as presented in (Li et al., 2020).

LongMix, is not equivalent to adding more epochs, as in baseline (Li et al., 2020). This shows evidence for the claim in Sec. 4.2 that a simple increase in the number of epochs is not equivalent to adding more MixUp iterations, as we propose for LongMix. Table 1 shows further evidence for this claim by comparing LongMix and baseline (Li et al., 2020) using the same number of training iterations for different noisy rates on CIFAR-10 and CIFAR-100 – results show that LongMix is more accurate for most cases.

5.5. Comparison with the State-of-the-Art

For CIFAR-10 and CIFAR-100, we evaluate our model using different levels of symmetric label noise ranging from 20% to 90%. We also consider asymmetric noisy, with noise rates of 40% and 49%. We report both the best test accuracy across all epochs and the averaged test accuracy over the last 10 epochs of training, similar to (Li et al., 2020). Table 2 shows that for CIFAR-10 and CIFAR-100 data sets, our method obtains better results for all evaluated noisy rates. LongReMix displays a higher improvement for large symmetric noise and asymmetric noise scenarios, which can be considered as the most challenging cases. We believe that the improvement over higher noise rates is due to the LongMix approach, which runs a large number of MixUp operations proportional to the size of the training set. The retraining with high confidence samples also improves the results for asymmetric noise. The results for semantic noise (Lee et al., 2019) in Table 4 shows again the superiority of our approach compared to the related work.

Also, we evaluate our method on large-scale data sets. For WebVision, Table 5 shows the Top-1 and Top-5 accuracy, where LongReMix displays better results than competing methods. For the Clothing1M evaluation, the competing methods rely on a pre-trained ImageNet model for training on Clothing1M. In our experiments, we did not observe any improvement with pre-trained models, and therefore we trained from scratch with 128k images from Clothing1M. The results in Table 6 show that our model, trained from

Data set		CIFAR-10						CIFAR-100				Webv.	Cloth.	Food.	Mean Rank
Noise type		sym.				asym.		sym.				-	-	-	-
Method/ n. ratio		20%	50%	80%	90%	40%	49%	20%	50%	80%	90%	-	-	-	-
LongReMix	Best	***96.25***	**95.01**	**93.88**	**81.98**	***94.64***	**84.68**	**77.82**	**75.59**	***62.92***	***33.80***	***78.92***	**74.38**	***87.39***	1.46
	Last	***96.02***	**94.72**	**93.37**	**81.35**	***94.32***	76.08	**77.52**	**75.11**	***62.34***	***33.25***	***78.00***	73.00	***87.29***	1.69
LongMix	Best	96.18	***95.19***	***94.09***	***85.33***	93.38	83.23	***78.03***	***75.84***	**62.24**	**33.54**	**78.44**	74.05	**87.21**	1.92
	Last	**95.98**	***94.79***	***93.73***	***84.71***	91.87	**77.18**	***77.56***	***74.87***	**61.60**	**33.00**	77.72	***73.25***	**87.12**	1.69
Retrain	Best	**96.23**	94.85	92.86	78.47	**94.59**	***85.10***	77.20	74.41	60.29	30.61	77.84	**74.30**	87.16	2.61
	Last	95.89	94.60	92.54	77.51	**94.31**	***80.88***	76.89	73.89	59.88	30.37	**77.84**	**73.21**	86.98	2.61

Table 3. Ablation Study Results. The italic bold, bold, and regular numbers represent respectively the ranking of first, second and third results in accuracy. Last column shows the average rank of each approach (smaller is better).

Data set	CIFAR-10			CIFAR-100		
Method/ noise ratio	DN (32%)	RN (38%)	VGG (34%)	DN (34%)	RN (37%)	VGG (37%)
CE + RoG	68.33	64.15	70.04	61.14	53.09	53.64
Bootstrap + RoG	68.38	64.03	70.11	54.71	53.30	53.76
Forward + RoG	68.20	64.24	70.09	53.91	53.36	53.63
Backward + RoG	68.66	63.45	70.18	54.01	53.03	53.50
D2L + RoG	68.57	60.25	59.94	31.67	39.92	45.42
DivideMix*	84.57	81.61	85.71	68.40	66.28	66.84
LongReMix [ours]	**85.13**	**82.51**	**85.90**	**69.03**	**66.70**	**67.42**

Table 4. Results for Semantic Noise. Results from baseline methods are as presented in (Lee et al., 2019). Methods marked by * denote re-implementations based on public code.

Method	Top 1	Top 5
Decoupling (Malach & Shalev-Shwartz, 2017)	62.54	84.74
D2L (Ma et al., 2018)	62.68	84.00
MentorNet (Jiang et al., 2018)	63.00	81.40
Co-teaching (Han et al., 2018b)	63.58	85.20
Iterative-CV (Chen et al., 2019)	65.24	85.34
DivideMix (Li et al., 2020)	77.32	91.64
LongReMix [ours]	**78.92**	**92.32**

Table 5. Results for WebVision (Li et al., 2017). Results from baseline methods are as presented in (Li et al., 2020).

Method	Test Accuracy
Cross-Entropy (Li et al., 2020)	69.21
M-correction (Arazo Sanchez et al., 2019)	71.00
PENCIL(Yi & Wu, 2019)	73.49
DeepSelf (Han et al., 2019)	74.45
CleanNet (Lee et al., 2018)	74.69
DivideMix (Li et al., 2020)	**74.76**
LongReMix † [ours]	74.38

Table 6. Results for Clothing1M (Xiao et al., 2015). Results from baseline methods are as presented in (Li et al., 2020). The marker † denotes the model is trained from scratch.

Method	from pre-trained	from scratch
Cross-Entropy	81.44	-
CleanNet	83.95	-
DeepSelf	85.10	-
DivideMix*	86.91	75.53
LongReMix [ours]	**87.39**	**78.57**

Table 7. Results for Food-101N (Lee et al., 2018). Methods marked by * denote re-implementations based on public code.

scratch and with a reduced training set, obtained comparable results to the competing approaches. Lastly, Table 7 summarizes the results for Food-101N. For this problem, we evaluate our approach with a pre-trained model and trained from scratch, and LongReMix outperforms all other approaches in both scenarios.

5.6. Ablation Study

We analyze the effect of the different components of our proposal in an ablation study, shown in Table 3. Below, ”Retrain” denotes the high-confidence training explained in Sec. 4.1 which increases the accuracy of the classifier that distinguishes between clean and noisy samples; and ”LongMix” represents the guided training from Sec. 4.2 that increases the number of MixUp operations. We first evaluate our approach without LongMix – this approach is referred to as ”Retrain”. Then we evaluate training only with the LongMix, without the second stage of re-training, and the whole model is denoted as LongReMix. In general, we can observe that the LongReMix is competitive for all noise scenarios (being best or second best for all cases), but it is generally better for the large-scale data sets. Considering different data sets and noise rates, LongReMix shows the best average rank.

6. Conclusion

We presented LongReMix, a new 2-stage noisy-label learning algorithm based on an unsupervised learning stage to classify clean and noisy training samples, followed by an SSL stage to minimise the EVR using a labelled set formed by samples classified as clean, and an unlabelled set with samples classified as noisy. Our LongReMix improves the precision of the unsupervised learning stage and improves the generalisation of the EVR minimisation. We show that LongReMix reaches state-of-the-art performance on several benchmarks, and is robust to over-fitting in high label noise problems.

References

Arazo Sanchez, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. Unsupervised label noise modeling and loss correction. 2019.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5049–5059, 2019.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Chen, P., Liao, B., Chen, G., and Zhang, S. Understanding and utilizing deep neural networks trained with noisy labels. *arXiv preprint arXiv:1905.05040*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ding, Y., Wang, L., Fan, D., and Gong, B. A semi-supervised two-stage approach to learning from noisy labels. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1215–1224. IEEE, 2018.

Frénay, B. and Verleysen, M. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Han, B., Niu, G., Yao, J., Yu, X., Xu, M., Tsang, I., and Sugiyama, M. Pumpout: A meta approach for robustly training deep neural networks with noisy labels. 2018a.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pp. 8527–8537, 2018b.

Han, J., Luo, P., and Wang, X. Deep self-learning from noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5138–5147, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.

Jaehwan, L., Donggeun, Y., and Hyo-Eun, K. Photometric transformer networks and label adjustment for breast density prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

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

Jiang, L., Huang, D., Liu, M., and Yang, W. Beyond synthetic noise: Deep learning on controlled noisy labels. ICML, 2020.

Kim, Y., Yim, J., Yun, J., and Kim, J. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 101–110, 2019.

Kong, K., Lee, J., Kwak, Y., Kang, M., Kim, S. G., and Song, W.-J. Recycling: Semi-supervised learning with noisy labels in deep neural networks. *IEEE Access*, 7: 66998–67005, 2019.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lee, K., Yun, S., Lee, K., Lee, H., Li, B., and Shin, J. Robust inference via generative classifiers for handling noisy labels. *arXiv preprint arXiv:1901.11300*, 2019.

Lee, K.-H., He, X., Zhang, L., and Yang, L. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456, 2018.

Li, J., Wong, Y., Zhao, Q., and Kankanhalli, M. S. Learning to learn from noisy labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5051–5059, 2019.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Li, W., Wang, L., Li, W., Agustsson, E., and Gool, L. V. Webvision database: Visual learning and understanding from web data. *CoRR*, 2017.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S., Xia, S., Wijewickrema, S., and Bailey, J. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pp. 3355–3364, 2018.

Malach, E. and Shalev-Shwartz, S. Decoupling" when to update" from" how to update". In *Advances in Neural Information Processing Systems*, pp. 960–970, 2017.

Miao, Q., Cao, Y., Xia, G., Gong, M., Liu, J., and Song, J. Rboost: Label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners. *IEEE transactions on neural networks and learning systems*, 27(11):2216–2228, 2015.

Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., and Brox, T. Self: Learning to filter noisy labels with self-ensembling. *arXiv preprint arXiv:1910.01842*, 2019.

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

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343, 2018.

Sachdeva, R., Cordeiro, F. R., Belagiannis, V., Reid, I., and Carneiro, G. Evidentialmix: Learning with combined open-set and closed-set noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3607–3615, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. Combating label noise in deep learning using abstention. In *International Conference on Machine Learning*, pp. 6234–6243. PMLR, 2019.

Wang, X., Hua, Y., Kodirov, E., and Robertson, N. M. Imae for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude's variance matters. *arXiv preprint arXiv:1903.12141*, 2019a.

Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., and Xia, S.-T. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8688–8696, 2018.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 322–330, 2019b.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Xue, C., Dou, Q., Shi, X., Chen, H., and Heng, P.-A. Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1280–1283. IEEE, 2019.

Yi, K. and Wu, J. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7017–7025, 2019.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 68–83, 2018.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I. W., and Sugiyama, M. How does disagreement help generalization against label corruption? *arXiv preprint arXiv:1901.04215*, 2019.

Yuan, B., Chen, J., Zhang, W., Tai, H.-S., and McMains, S. Iterative cross learning on noisy labels. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 757–765. IEEE, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, C., Hsieh, M.-H., and Tao, D. Generalization bounds for vicinal risk minimization principle. *arXiv preprint arXiv:1811.04351*, 2018.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, W., Wang, Y., and Qiao, Y. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7373–7382, 2019.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pp. 8778–8788, 2018.

Zhang, Z., Zhang, H., Arik, S. O., Lee, H., and Pfister, T. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9294–9303, 2020.