Model and Data Agreement for Learning with Noisy Labels

Anonymous submission

Abstract

Learning with noisy labels is a vital topic for practical deep learning as models should be robust to noisy open-world datasets in the wild. The state-of-the-art noisy label learning approach JoCoR fails when faced with a large ratio of noisy labels. We speculate the reason lies in that letting the two models teach each other might cause them to learn errors from each other as the performance of both of them is low. Moreover, selecting small-loss samples can also cause error accumulation as once the noisy samples are mistakenly selected as small-loss samples, they are more likely to be selected again. In this paper, we try to deal with error accumulation in noisy label learning from both model and data perspectives. We introduce mean point ensemble to guide the two models to imitate a stronger model instead of teaching each other to reduce error accumulation from the model perspective. Furthermore, as the flip images have the same semantic meaning as the original images, we select small-loss samples according to the loss values of flip images instead of the original ones to reduce error accumulation from the data perspective. Extensive experiments on CIFAR-10, CIFAR-100, and large-scale Clothing1M show that our method outperforms state-of-the-art noisy label learning methods with different levels of label noise. Our method can also be seamlessly combined with other noisy label learning methods to further improve their performance and generalize well to other tasks. The source code to implement all the experiments will be released.

Introduction

The performance improvement of Deep Neural Networks (DNNs) depends largely on the large-scale training datasets. However, collecting large-scale datasets with fully precise annotations is usually expensive and time-consuming. There are lots of label noise in the open-world datasets, which degrades the performance of deep learning models in practical applications. At the same time, it is widely known that deep neural networks can easily memorize large-scale data with even completely random labels, making unreliable predictions when generalizing to other tasks (Zhang et al. 2017; Arpit et al. 2017; Jiang et al. 2018). Thus, learning with noisy labels has drawn lots of attention in recent years (Patrini et al. 2017; Han et al. 2018a; Zhang and Sabuncu 2018; Wang et al. 2019; Junnan et al. 2019; Ren et al. 2018; Li, Socher, and Hoi 2020; Jiang et al. 2018; Malach and Shalev-Shwartz 2017; Han et al. 2018b; Yu et al. 2019a; Wei et al. 2020).

Some previous works try to seek theoretically guaranteed methods to solve the noisy label learning problem (Patrini et al. 2017; Han et al. 2018a; Zhang and Sabuncu 2018; Wang et al. 2019). However, they are usually not suitable for real-world label noise as the transition probability between the noisy label and ground truth in real-world datasets is hard to estimate. Others follow the small-loss selection path (Jiang et al. 2018; Malach and Shalev-Shwartz 2017; Han et al. 2018b; Yu et al. 2019a; Wei et al. 2020). They treat the small-loss samples as clean samples and only use them for training, while there exists the debate that whether the model should utilize "agreement" or "disagreement". (Malach and Shalev-Shwartz 2017; Yu et al. 2019a) claim that only using the samples that two models have different predictions as training samples can keep the models distinct and avoid fitting noisy labels. However, (Wei et al. 2020) argues that disagreement is not necessary and proposes a training paradigm that pushes two models together. We observe that (Wei et al. 2020) achieves excellent performance when learning with a small ratio of noisy labels while it fails with a large ratio of noisy labels. We argue that this might be because the two models imitating each other can cause them to learn errors from each other when facing a large ratio of noisy labels as the accuracy of the two models are both low learning with noisy labels. In the meantime, selecting small-loss samples as clean samples for training can also lead to error accumulation from the data perspective. Some noisy samples might get small loss values during training and thus be selected for training. After updating the gradients, the selected noisy samples might be remembered by the models and are more likely to be selected again in the following training epochs, which leads to error accumulation.

Motivated by the above two problems, in this paper, we propose to deal with the error accumulation in noisy label learning from both model and data perspectives. For the model perspective, instead of simply pushing the two models together, we introduce mean point ensemble to guide the two models to imitate a stronger model, which improves their performance under a large ratio of label noise significantly. From the data perspective, we find that the model fits noisy samples through memorization, while it cannot remember the flipped noisy samples to the noisy labels though they are with the same semantic meaning. Thus, we propose to utilize the flipped images to better detect noisy labels for the first time to

reduce error accumulation from the data perspective. Though flip is a common augmentation method, using the loss values of flipped images to detect noisy labels is rarely explored. We also show that using flip to detect noisy samples can bring more benefits than simply using flip as a basic augmentation under label noise.

We evaluate our proposed method Model and Data Agreement (MDA) on image classification datasets CIFAR-10, CIFAR-100, and large-scale Clothing1M. Extensive experiments validate the effectiveness of each of the two proposed modules. MDA is also compatible with other state-of-the-art noisy label learning methods to further improve their performance and can generalize well to other noisy label learning tasks.

The main contributions of our work are as follows:

1. We introduce mean point ensemble to guide the two models to imitate a stronger model instead of imitating each other. We validate mean point ensemble is superior to agreement learning through extensive experiments.
2. We find that the flipped samples can help the model to better detect noisy samples. We also validate that using flip to detect noisy labels achieves better results than simply using flip as a data augmentation method under the task of learning with label noise.
3. Extensive experiments show that the proposed method advances state-of-the-art noisy label learning methods on noisy label CIFAR-10, CIFAR-100 and large-scale noisy dataset Clothing1M.
4. The proposed method can be easily combined with other noisy label learning methods to further improve their performance and generalize well to other noisy label learning tasks.

Related Work

Noisy Label Learning

How to achieve good performance learning with noisy labels has drawn lots of attention in recent years (Patrini et al. 2017; Han et al. 2018a; Zhang and Sabuncu 2018; Thulasidasan et al. 2019; Xu et al. 2019; Jiang et al. 2018; Ren et al. 2018; Arazo et al. 2019; Han et al. 2018b; Malach and Shalev-Shwartz 2017; Wei et al. 2020; Xie and Huang 2021; Yi and Wu 2019; Kim et al. 2019; Huang et al. 2019; Han, Luo, and Wang 2019; Li, Socher, and Hoi 2020; Ye and Yuen 2020; Nguyen et al. 2019; Li, Xiong, and Hoi 2021). The noisy label learning methods in recent years can be mainly categorized into two types.

Theory Guaranteed Noisy Label Learning Methods The first type of noisy label learning method estimates the noise transition matrix to model the label transition probability or propose generalized cross-entropy loss. Patrini *et al.* (Patrini et al. 2017) propose loss correction methods that estimate the label noise transition matrix through training with noisy datasets. Hendrycks *et al.* (Hendrycks et al. 2018) estimate the noise transition matrix by using a small set of trusted data. Han *et al.* (Han et al. 2018a) incorporate human cognition of invalid class transitions to help estimate the noise transition matrix. Zhang *et al.* (Zhang and Sabuncu 2018) propose noise-robust loss functions that can be seen as a generalization of mean absolute error (MAE) loss and cross entropy (CCE) loss, which can utilize both the merits of MAE loss and CCE loss. Xu *et al.* (Xu et al. 2019) design a novel loss function based on the mutual information theory, which is provably robust to instance-independent label noise. The common merit of these methods is that they are theoretically guaranteed. However, they might be unsuitable for real-world datasets with noisy labels as they usually do not conform to the theoretical assumptions of these methods.

Sample Selection Label Learning Methods The other type of noisy label learning approach is based on the observation that DNNs learn simple patterns before memorizing the noisy labels (Arpit et al. 2017). They treat the small-loss samples as clean samples to train the DNNs and filter out large-loss samples as they are likely to be noisy samples. Jiang *et al.* (Jiang et al. 2018) train a mentor net to imitate a human teacher. The mentor net selects clean samples to teach the student network to avoid remembering noisy samples. Han *et al.* (Han et al. 2018b) train two differently initialized models to teach each other as the two different models can mitigate different types of errors caused by noisy labels. Malach *et al.* (Malach and Shalev-Shwartz 2017) update the parameters only on the instances that the two models have different predictions to maintain divergence. Yu *et al.* (Yu et al. 2019b) combine co-teaching (Han et al. 2018b) with decoupling (Malach and Shalev-Shwartz 2017) to further improve the performance of co-teaching as updating only using the samples with different predictions can keep the two models distinct. Wei *et al.* (Wei et al. 2020) argue that keeping two models distinct is not necessary and they propose a method named JoCoR to push two models closer during training and use their agreement degree to select small-loss samples. Though the agreement of two models can help to detect noisy samples, we argue that JoCoR fails when facing lots of noisy labels as pushing two models closer might cause them to learn errors from each other. Instead of simply pushing two models close to each other, we propose to use mean point ensemble to guide the two models to imitate a stronger model, which stables the training and improves performance under a large ratio of label noise.

Model Ensemble

Model ensemble is a very effective technique that can significantly improve the performance of deep learning models. By ensembling multiple models together, we can reduce the bias and overfitting of a single model, which can make deep learning models more robust to noisy labels. Deep model ensemble methods in supervised learning (Ganaie, Hu et al. 2021) can be mainly categorised into bagging (Breiman 1996), boosting (Zhang and Zhang 2008), negative correlation learning (Liu and Xin 1999), explicit/implicit ensembles (Srivastava et al. 2014; Wan et al. 2013; Huang et al. 2016; Singh, Hoiem, and Forsyth 2016), homogeneous/heterogeneous ensemble (Breiman 2001; Li et al. 2018), decision fusion strategies (Ju, Bibaut, and van der Laan 2018). Several aforementioned noisy label learning methods can be viewed as utilizing model ensemble methods. Co-teaching (Han et al. 2018b) can be viewed as two different initialized models that

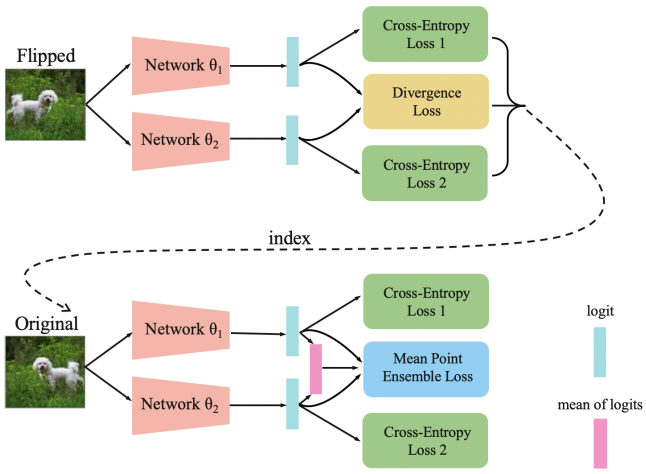

Figure 1: The framework of our MDA.

teach each other, which is a kind of ensemble. JoCoR (Wei et al. 2020) also uses model ensemble, which is named Deep Mutual Learning (Zhang et al. 2018). Deep Mutual Learning makes several student models learn collaboratively and teach each other throughout the whole training process. JoCoR incorporates Deep Mutual Learning to let the two models teach each other to achieve high performance through model ensemble. However, when facing a large ratio of label noise, the performance of both models is low. Thus, letting the two models teach each other brings the drawback that the two models might learn the errors from each other. Thus, we propose to introduce mean point ensemble to guide the two models to imitate a stronger model.

Proposed Method

In this section, we illustrate the implementation details of our proposed Model and Data Agreement (MDA) method. We propose to deal with the noisy label learning problem from both model and data perspectives. Specifically, we introduce mean point ensemble to guide two models to imitate a stronger model. We further utilize the flipped images which are unseen during the training process to select clean samples, which is superior to just use flip as a data augmentation.

Model and Data Agreement

Given the dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ with N samples, we first carry out flip augmentation to all the samples and get the flipped dataset denoted as $\widetilde{D}$. In the training stage, we use $\widetilde{D} = \{\widetilde{\mathbf{x}}_i, y_i\}_{i=1}^{N}$ to compute the classification loss and utilize the divergence loss to select clean samples for training, while we only backpropagate gradients on the original dataset D.

Specifically, we also utilize the agreement maximization principles (Sindhwani, Niyogi, and Belkin 2005; Wei et al. 2020) to detect noisy samples. We use the joint loss (1) of classification and agreement to select clean samples.

$$l_{sel}(\widetilde{\mathbf{x}}_i) = l_{cls}(\widetilde{\mathbf{x}}_i, y_i) + \lambda * l_{ag}(\widetilde{\mathbf{x}}_i), \tag{1}$$

where $l_{cls}(\widetilde{\mathbf{x}}_i, y_i)$ means the sum of the classification loss of the two models of sample $\widetilde{\mathbf{x}}_i$ and $l_{ag}(\widetilde{\mathbf{x}}_i)$ represents the agreement level of the two models towards the prediction results of the sample $\widetilde{\mathbf{x}}_i$. They are computed following (2) and (5) respectively. λ is the weight of the agreement loss when selecting the clean samples.

$$l_{cls}(\widetilde{\mathbf{x}}_i, y_i) = l_{C1}(\widetilde{\mathbf{x}}_i, y_i) + l_{C2}(\widetilde{\mathbf{x}}_i, y_i), \tag{2}$$

where

$$l_{C1}(\widetilde{\mathbf{x}}_i, y_i) = -\frac{1}{N}\sum_{i=1}^{N}(\log \frac{e^{f^1_{y_i}(\widetilde{\mathbf{x}}_i)}}{\sum_j^M e^{f^1_j(\widetilde{\mathbf{x}}_i)}}), \tag{3}$$

$$l_{C2}(\widetilde{\mathbf{x}}_i, y_i) = -\frac{1}{N}\sum_{i=1}^{N}(\log \frac{e^{f^2_{y_i}(\widetilde{\mathbf{x}}_i)}}{\sum_j^M e^{f^2_j(\widetilde{\mathbf{x}}_i)}}), \tag{4}$$

$f^1(\widetilde{\mathbf{x}}_i)$ and $f^2(\widetilde{\mathbf{x}}_i)$ are the outputs of input image $\widetilde{\mathbf{x}}_i$ from the two models. N and M represent the number of samples and classes.

$$l_{ag}(\widetilde{\mathbf{x}}_i) = D_{KL}(f^1(\widetilde{\mathbf{x}}_i)||f^2(\widetilde{\mathbf{x}}_i) + D_{KL}(f^2(\widetilde{\mathbf{x}}_i)||f^1(\widetilde{\mathbf{x}}_i)), \tag{5}$$

where

$$D_{KL}(f^1(\widetilde{\mathbf{x}}_i)||f^2(\widetilde{\mathbf{x}}_i)) = \sum_{i=1}^{N}\sum_{j=1}^{M} f^1_j(\widetilde{\mathbf{x}}_i)\log\frac{f^1_j(\widetilde{\mathbf{x}}_i)}{f^2_j(\widetilde{\mathbf{x}}_i)}, \tag{6}$$

$$D_{KL}(f^2(\widetilde{\mathbf{x}}_i)||f^1(\widetilde{\mathbf{x}}_i)) = \sum_{i=1}^{N}\sum_{j=1}^{M} f^2_j(\widetilde{\mathbf{x}}_i)\log\frac{f^2_j(\widetilde{\mathbf{x}}_i)}{f^1_j(\widetilde{\mathbf{x}}_i)}. \tag{7}$$

When selecting clean samples according to the loss values. Batch size is a very important factor to be considered. Selecting small-loss samples in the mini-batch scale is less effective than selecting them from the whole training set. To reduce the randomness of the sample selection process which might cause performance degradation, we first compute the loss values on the whole dataset $\widetilde{D}$. We then select clean samples according to the selection loss l_{sel} values of all the training samples. We mark the indexes of the small-loss samples and then backpropagate gradients only on the corresponding samples from the original dataset D. The details of the sample selection are shown in Algorithm 1.

Mean point ensemble

We introduce mean point ensemble to let the two models imitate a stronger model. Instead of letting two different models learn from each other, we let the two models imitate a stronger model. The mean prediction results can be viewed as the ensemble of two models. It is common sense that an ensemble achieves good performance than individual models. Thus, instead of letting the two models teach each other which might cause them to learn different types of errors, we let them imitate a stronger model which has higher performance to reduce the error accumulation during training. The mean

Algorithm 1: Model and Data Agreement

Require: original training set D and the flipped counterpart $\widetilde{D}$, Network f with $\Theta = \{\Theta_1, \Theta_2\}$, learning rate η, noise rate τ, warm up selection epochs T_k and total epochs $T_{\max}$, data loader iteration I_{train};

for t = 1,2,...,$T_{\max}$ **do**
  $\boldsymbol{p}_1 = f(\widetilde{\boldsymbol{x}}, \Theta_1), \forall \widetilde{\boldsymbol{x}} \in \widetilde{D}$;
  $\boldsymbol{p}_2 = f(\widetilde{\boldsymbol{x}}, \Theta_2), \forall \widetilde{\boldsymbol{x}} \in \widetilde{D}$;
  Calculate the selection loss ℓ_{sel} by (1);
  Obtain the indexes of small-loss set on $\widetilde{D}$ according to the ratio $R(t)$;
  for $n = 1, \ldots, I_{\text{train}}$ **do**
    Fetch mini-batch D_n from D;
    Obtain ℓ_{cls} by (2) on D_n samples with indexes fall into the small-loss set and ℓ_{ens} by (8) on all the samples;
    Obtain ℓ_{train} by (10);
    Update $\Theta = \Theta - \eta \nabla \ell_{train}$;
  end for
  Update $R(t) = 1 - \min\left\{\frac{t}{T_k}\tau, \tau\right\}$
end for

Ensure: Θ_1 and Θ_2

point ensemble regularization loss can be simplified to the following

$$l_{ens} = \frac{1}{N}\sum_{i=1}^{N}((f^1(\mathbf{x}_i) - \bar{f}(\mathbf{x}_i))^2 + (f^2(\mathbf{x}_i) - \bar{f}(\mathbf{x}_i))^2), \quad (8)$$

where

$$\bar{f}(\mathbf{x}_i) = \frac{1}{2}(f^1(\mathbf{x}_i) + f^2(\mathbf{x}_i)). \quad (9)$$

Note that we use a symmetric mean squared error loss, which is shown to be more robust to noisy labels (Ghosh, Kumar, and Sastry 2017). We also calculate l_{ens} on all training samples rather than only the selected samples, making the two models learn useful information from all the images.

Overall Training Loss

Having acquired the indexes of the clean samples from the flipped images, we use them to index the clean samples from the original training images. We then compute the joint loss of classification on selected samples and the loss of mean point ensemble on all samples to train the two models.

$$l_{train}(\mathbf{x}_i) = l_{cls}(\mathbf{x}_i, y_i) + \gamma * l_{ens}(\mathbf{x}_i), \quad (10)$$

γ is the weight of the mean point ensemble loss of the two models, which we will study in the ablation study.

Experiments

In this section, we illustrate the implementation details of MDA. We then verify MDA on the CIFAR-10, CIFAR-100 (Krizhevsky, Hinton et al. 2009) and Clothing1M (Xiao et al. 2015) with different levels of label noise. The ablation studies are carried out to study the effectiveness of the flip noise detection and mean point ensemble separately. We further display some visualization results to provide an intuitive understanding of flip noise detection. We also show that MDA is compatible with other noisy label learning methods and can further improve their performance.

Implementation Details

To make a fair comparison with JoCoR (Wei et al. 2020), we use a 7-layer CNN network architecture for CIFAR-10, CIFAR-100 (Krizhevsky, Hinton et al. 2009) and ResNet-18 (He et al. 2016) for Clothing1M (Xiao et al. 2015) as the backbone network. We *do not* use any data augmentation tricks. For CIFAR-10 and CIFAR-100, the batch size is 512. The initial learning rate is 0.001. We use Adam (Kingma and Ba 2014) optimizer with the weight decay as 0.0001. We run 200 epochs in total and linearly decay the learning rate to zero from 80 to 200 epochs. We select small-loss samples following the ratio $R(t)$ as: $R(t) = 1 - min(\frac{t}{T_k}\tau, \tau)$. Following Co-teaching (Han et al. 2018b) and JoCoR (Wei et al. 2020), we assume that the noise rate τ is known. t is the current training epoch and we set the T_k as 10 for CIFAR-10 and CIFAR-100 to make a fair comparison with other methods. For experiments on Clothing1M, we train the models for 15 epochs in total. During the training stage, we set the learning rate to 0.0008 , 0.0005 and 0.00005 for 5 epochs each. We also use Adam optimizer (momentum=0.9) and set the batch size to 64 following (Wei et al. 2020). Experiments are conducted on 4 NVIDIA RTX 2080Ti GPUs.

Evaluation of MDA on CIFAR-10 and CIFAR-100 with Noisy Labels

We quantitatively evaluate the improvement of our proposed MDA against other state-of-the-art methods. We train our model on the same noisy dataset as other methods to make fair comparisons. We report the mean test accuracy of the last 10 epochs. We follow the tradition to run each experiment 5 times and report the mean and the standard deviation of the accuracy. We explore the robustness of MDA with three levels of label noise including the ratio of 20%, 50%, 80% on CIFAR-10 and CIFAR-100.

As shown in Table 1, our method outperforms all other state-of-the-art label noise learning methods by a non-trivial margin. For example, MDA outperforms JoCoR under 20%, 50%, 80% label noise by 0.35%, 1.39%, 12.98% respectively on CIFAR-10. MDA outperforms JoCoR under 20%, 50%, 80% label noise by 3.43%, 6.47%, 8.31% on CIFAR-100. MDA improves state-of-the-art methods more on large noise ratio, which implies that our method can deal with harder noisy datasets. We owe the large improvements on 80% label noise to the introduction of mean point ensemble and flip noise detection, which reduces error accumulation from both model and data perspectives.

Evaluation of MDA on Large-Scale Noisy Data Clothing1M

To further evaluate the effectiveness of our proposed MDA, we carry out experiments on the large-scale real-world noisy dataset Clothing1M. The experiment results are shown in Table 3. MDA outperforms JoCoR by 0.94% and 1.28% on

Table 1: Average test accuracy (%) on *CIFAR-10* over the last 10 epochs. Each experiment is run five times, shown with the mean and standard deviation.

Noise Rate	Baseline	Decoupling	Co-teaching	Co-teaching+	JoCoR	MDA
20%	69.18 ± 0.52	69.32 ± 0.40	78.23 ± 0.27	78.71 ± 0.34	85.73 ± 0.19	86.08 ± 0.14
50%	42.71 ± 0.42	40.22 ± 0.30	71.30 ± 0.13	57.05 ± 0.54	79.41 ± 0.25	80.80 ± 1.96
80%	16.24 ± 0.39	15.31 ± 0.43	26.58 ± 2.22	24.19 ± 2.74	27.78 ± 3.06	40.76 ± 5.41

Table 2: Average test accuracy (%) on *CIFAR-100* over the last 10 epochs. Each experiment is run five times, shown with the mean and standard deviation.

Noise Rate	Baseline	Decoupling	Co-teaching	Co-teaching+	JoCoR	MDA
20%	35.14 ± 0.44	33.10 ± 0.12	43.73 ± 0.16	49.27 ± 0.03	53.01 ± 0.04	56.44 ± 0.13
50%	16.97 ± 0.40	15.25 ± 0.20	34.96 ± 0.50	40.04 ± 0.70	43.49 ± 0.46	49.96 ± 0.38
80%	4.41 ± 0.14	3.89 ± 0.16	15.15 ± 0.46	13.44 ± 0.37	15.49 ± 0.98	23.80 ± 1.19

Table 3: Best test accuracy (%) and Last epoch test accuracy (%) on *Clothing1M*.

Accuracy	Baseline	F-correction	Decoupling	Co-teaching	JoCoR	MDA
Best epoch acc.	67.22	68.93	68.48	69.21	70.30	71.24
Last epoch acc.	64.68	65.36	67.32	68.51	69.79	71.07

the best test accuracy and last epoch test accuracy respectively. As Clothing1M is a very large dataset containing 1 million samples, our proposed MDA can be considered as outperforming JoCoR by a non-trivial margin. More importantly, compared with JoCoR, the best test accuracy during the training process and the last epoch test accuracy of MDA is more similar, which means our method can be more robust to the label noise and less likely to overfit the noisy labels for the last several epochs.

The Effectiveness of Flip Noise Detection

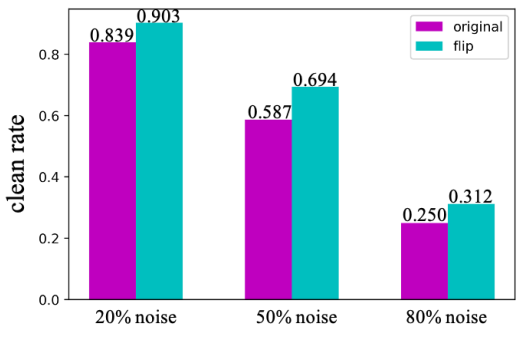

Figure 2: Effectiveness of flip noise detection.

In this section, we illustrate the effectiveness of flip noise detection. We train a classification model for 40 epochs using CIFAR-100 with different levels of label noise. We then load the trained model to select small-loss samples using the loss values of original images and flip images separately. The experiment result is shown in Figure 2. When training with 20% label noise, we select 80% small-loss samples as clean samples using original images or flip images. The result shows that 83.9% of the small-loss original samples are without noisy labels while 90.3% of the small-loss flip images are samples without noisy labels. This experiment illustrates the reason why we select small-loss samples using the flipped ones during training can improve the classification accuracy. We also carry out experiments to compare with the JoCoR method which simply uses flip augmentation for noisy label training. Under 30% label noise in CIFAR-100, using flip augmentation only gets 19.40% accuracy, while using flip to detect noise gets 23.80%, which illustrates that simply using flip augmentation has relatively small improvement on training with noisy labels.

Evaluation of Different Modules

To show the influence of the flip noise detection module and the mean point ensemble module separately, we carry out an ablation study on CIFAR-100 with different levels of label noise. The results are shown in Table 4. Without the flip noise detection and mean point ensemble modules, the model degrades to the baseline method. When we detect the label noise through the flipped images, the test accuracy on the test set increases compared with the baseline, which illustrates that using the unseen flipped images can help the model to better filter out noisy samples. When we only use mean point ensemble, the two models are imitating a stronger model, which reduces error accumulation from the model perspective. The results show that using flip noise detection or mean point ensemble alone can both help the model to achieve higher performance than baseline while using the two modules together achieves the best performance. We conclude the two modules can be used separately to deal with noisy labels while they only achieve the best performance when they cooperate with each other.

Visualization of Accuracy and Clean rate

We plot the test accuracy versus training epochs under different noisy labels in Figure 3. When the noise rate is 20% or 50%, JoCoR can achieve high performance and the improvement from our proposed method is limited. However, when it comes to 80% noise, we can clearly view that JoCoR

Table 4: Evaluation of the different modules of MDA on CIFAR-100 with different levels of label noise

flip noise detection	mean point ensemble	20%	50%	80%
x	x	35.14%	16.97%	4.41%
✓	x	54.84%	45.41%	23.73%
x	✓	55.12%	47.39%	23.22%
✓	✓	56.34%	49.58%	24.69%

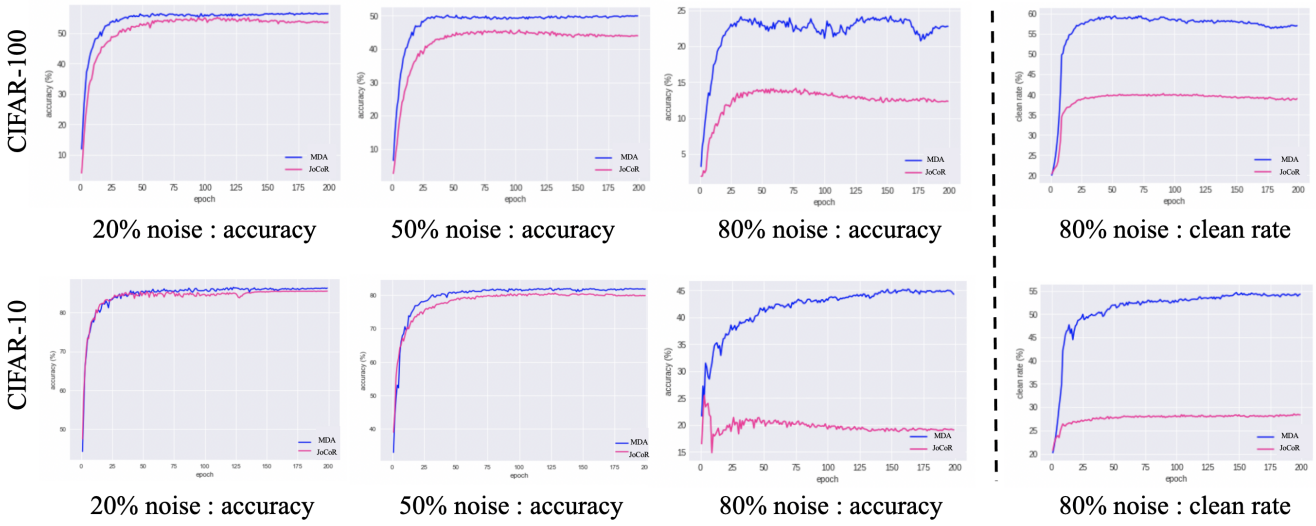

Figure 3: The test accuracy of JoCoR and MDA on CIFAR-10 and CIFAR-100. MDA outperforms JoCoR under all different noise levels. We also plot the clean rate of the two methods learning with 80% label noise, MDA can get a cleaner training set because of the flip noise detection and mean point ensemble.

fails in both CIFAR-10 and CIFAR-100 and achieves very low performance. The failure of JoCoR lies in pulling two models closer as when the noise rate is larger than 50%, the accuracies of the two models are both low, which means imitating each other might cause the two models to accumulate different errors. The results show that our method outperforms JoCoR by a large margin when the noise rate is large. We also plot the clean rate of JoCoR and our method under a large noise rate, it aligns with the test accuracy curves, which conforms to our speculation.

Ablation Studies

We also carry out ablation studies to explore the effect of the mean point ensemble loss weight γ and the agreement loss weight λ.

Evaluation of the mean point ensemble Loss Weight γ. We experiment with different values of γ under different levels of label noise on CIFAR-100. It is shown that we need to use larger γ to guide the two models to imitate a stronger model when the noise ratio increases, which helps to reduce the error accumulation from the model perspective. It is shown that we can choose γ from a large range randomly as the performance changes little. The optimal choice of γ is 0.1, 0.1, 0.05 under 20% and 50% noise and 80% noise. For simplicity, we set γ as 0.1 in all our experiments.

Evaluation of the Agreement Loss Weight λ. We further study the agreement loss weight λ on CIFAR-100 to understand the small-loss selection process to the model performance under noisy labels. It is shown that facing 20% or 50% label noise, MDA achieves the best performance when λ is around 0.7 while facing 80% label noise, MDA achieves the best performance when λ is 0.4. It is intuitive as when there is lots of label noise, the two models are less likely to agree with each other. The agreement loss becomes large. Thus, the agreement loss weight λ should be relatively small to ensure that the model can also consider the classification loss to select useful samples for training. We set λ as 0.7 when training with 20% or 50% label noise, while 0.4 when training with 80% label noise.

Combine with Other Noisy Label Learning Methods

MDA can also be combined with other state-of-the-art noisy label learning methods to further improve their performance. As MDA deals with noisy labels from model and data perspectives, we can improve other small-loss selection methods with MDA. Specifically, when selecting noisy samples, we could use the loss values of the flipped images instead of the original training images to reduce the error accumulation of sample selection from the data perspective. When training two models, we can combine the mean point ensemble with other methods to let the two models imitate a stronger model. Thus, reducing the error accumulation from the model

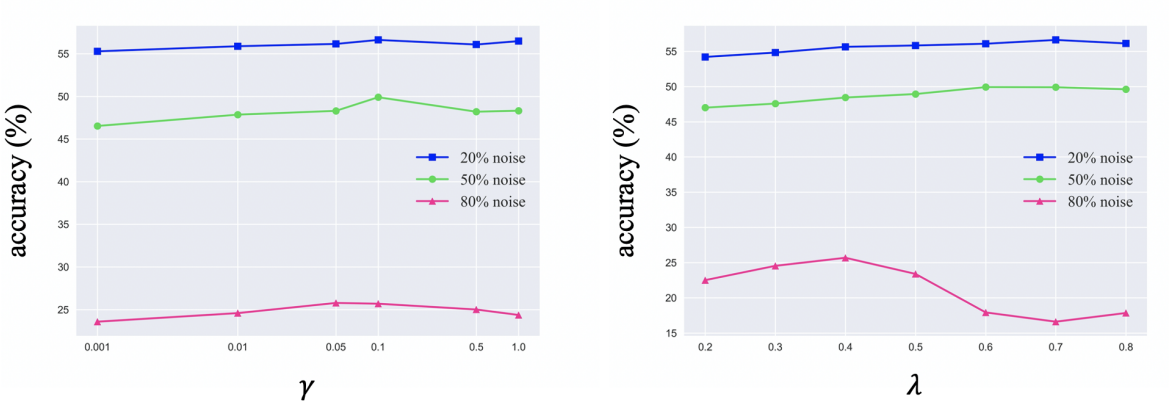

Figure 4: Ablation study of the mean point ensemble weight of γ and the agreement loss weight of λ

Table 5: Combine with Co-teaching

Methods	CIFAR10			CIFAR100		
	20% noise	50% noise	80% noise	20% noise	50% noise	80% noise
Baseline	69.38%	42.70%	16.55%	35.68%	17.38%	4.24%
Co-teaching	77.61%	71.16%	33.21%	44.10%	35.15%	14.81%
Co-teaching + MDA	84.76%	73.22%	33.24%	55.48%	43.91%	26.32%

perspective. Co-teaching maintains two different initialized models which select clean samples for each other to suppress the error accumulation. We add our proposed method to Co-teaching to improve its performance to show the plug-in and play characteristics of our proposed method. We select clean samples using flipped images and let the two different initialized models imitate a stronger model during the whole training process of Co-teaching. The results are shown in Table 5. Without bells and whistles, Co-teaching plus MDA achieves the best performance under all different noise levels on both CIFAR-10 and CIFAR-100 datasets, which illustrates that MDA can be seamlessly combined with sample selection noisy label learning methods to further improve their performance.

The Generalization Ability of MDA

MDA can also be utilized to solve other tasks. Facial Expression Recognition (FER) aims at helping computers to understand human behavior or even interact with a human by recognizing human expression. As we use flip noise detection to filter out the noisy samples, our method can generalize to other tasks with images of the same semantic meaning before and after the flip. The images of facial expression recognition are also symmetric. Thus, we carry out experiments on the FER dataset RAF-DB (Li, Deng, and Du 2017) and compare our method with several state-of-the-art FER noisy label learning methods.

The results are shown in Table 6, we compare MDA with Co-teaching (Han et al. 2018b). We also display some state-of-the-art FER noisy label learning methods, they do not need the exact noise rate to filter out the exact ratio of large-loss samples like Co-teaching and MDA. The results imply that

Table 6: MDA on noisy label FER tasks

Method	10% label noise	20% label noise	30% label noise
Baseline	81.01%	77.98%	75.50%
SCN	82.15%	79.79%	77.45%
RUL	86.17%	84.32%	82.06%
EAC	88.02%	86.05%	84.42%
Co-teaching	83.57%	81.75%	79.78%
MDA (Ours)	87.77%	87.18%	84.57%

our proposed MDA can acquire state-of-the-art performance knowing the exact noise rate compared with other FER noisy label learning methods. Furthermore, Co-teaching also needs to know the exact noise rate to filter out large-loss samples, while it does not generalize well to FER noisy label learning task as its performance is outperformed by MDA by a large margin under all the different noise levels.

Conclusion

In this paper, we deal with noisy label learning from both model and data perspectives. Instead of letting the two models teach each other, we introduce mean point ensemble to guide the models to imitate a better model during training. We also find that the flipped images can be utilized to better detect noisy samples and achieves better performance than just using flip as an augmentation method. Extensive experiments validate that our proposed MDA method outperforms other state-of-the-art noisy label learning methods and each of the modules improves the JoCoR method. Furthermore, MDA can be seamlessly combined with other noisy label learning methods to further improve their performance and generalize well to other noisy label learning tasks.

References

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

Arpit, D.; Jastrzbski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *ICML*.

Breiman, L. 1996. Bagging predictors. *Machine learning*, 24(2): 123–140.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1): 5–32.

Ganaie, M. A.; Hu, M.; et al. 2021. Ensemble deep learning: A review. *arXiv preprint arXiv:2104.02395*.

Ghosh, A.; Kumar, H.; and Sastry, P. S. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Han, B.; Yao, J.; Niu, G.; Zhou, M.; Tsang, I.; Zhang, Y.; and Sugiyama, M. 2018a. Masking: A new perspective of noisy supervision. In *NeurIPS*.

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 *NeurIPS*.

Han, J.; Luo, P.; and Wang, X. 2019. Deep self-learning from noisy labels. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hendrycks, D.; Mantas, M.; Wilson, D.; and Kevin, G. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*.

Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; and Weinberger, K. Q. 2016. Deep networks with stochastic depth. In *ECCV*.

Huang, J.; Qu, L.; Jia, R.; and Zhao, B. 2019. O2u-net: A simple noisy label detection approach for deep neural networks. In *ICCV*.

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 *ICML*.

Ju, C.; Bibaut, A.; and van der Laan, M. 2018. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15): 2800–2818.

Junnan, L.; Yongkang, W.; Qi, Z.; and Mohan, S. K.-h. 2019. Learning to learn from noisy labeled data. In *CVPR*.

Kim, Y.; Yim, J.; Yun, J.; and Kim, J. 2019. Nlnl: Negative learning for noisy labels. In *ICCV*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Learning from noisy data with robust representation learning. In *ICCV*.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*.

Li, W.; Ding, S.; Chen, Y.; and Yang, S. 2018. Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *Ieee Access*, 6: 54396–54406.

Liu, Y.; and Xin, Y. 1999. Ensemble learning via negative correlation. *Neural networks*.

Malach, E.; and Shalev-Shwartz, S. 2017. Decoupling" when to update" from" how to update". In *NeurIPS*.

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

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 *CVPR*.

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

Sindhwani, V.; Niyogi, P.; and Belkin, M. 2005. A co-regularization approach to semi-supervised learning with multiple views. *ICML Workshop on Learning With Multiple Views*.

Singh, S.; Hoiem, D.; and Forsyth, D. 2016. Swapout: Learning an ensemble of deep architectures. *NeurIPS*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Thulasidasan, S.; Bhattacharya, T.; Bilmes, J.; Chennupati, G.; and Mohd-Yusof, J. 2019. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*.

Wan, L.; Zeiler, M.; Zhang, S.; Le Cun, Y.; and Fergus, R. 2013. Regularization of neural networks using dropconnect. In *ICML*.

Wang, Y.; Xingjun, M.; Zaiyi, C.; Yuan, L.; Jinfeng, Y.; and James, B. 2019. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *CVPR*.

Xie, M.-K.; and Huang, S.-J. 2021. Partial multi-label learning with noisy label identification. *TPAMI*.

Xu, Y.; Cao, P.; Kong, Y.; and Wang, Y. 2019. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*.

Ye, M.; and Yuen, P. C. 2020. PurifyNet: A robust person re-identification model with noisy labels. *TIFS*.

Yi, K.; and Wu, J. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*.

Yu, X.; Bo, H.; Jiangchao, Y.; Gang, N.; Ivor, T.; and Masashi, S. 2019a. How does disagreement help generalization against label corruption? In *ICML*.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I. W.; and Sugiyama, M. 2019b. How does disagreement benefit co-teaching? *arXiv preprint arXiv:1901.04215*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning (still) requires rethinking generalization. In *ICLR*.

Zhang, C.-X.; and Zhang, J.-S. 2008. RotBoost: A technique for combining Rotation Forest and AdaBoost. *Pattern recognition letters*, 29(10): 1524–1536.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *CVPR*.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*.