Combating Noisy Labels by Agreement:
A Joint Training Method with Co-Regularization

Hongxin Wei[1]   Lei Feng[1*]   Xiangyu Chen[2]   Bo An[1]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Open FIESTA Center, Tsinghua University, China

{owenwei,boan}@ntu.edu.sg   feng0093@e.ntu.edu.sg   chenxian18@mails.tsinghua.edu.cn

Abstract

Deep Learning with noisy labels is a practically challenging problem in weakly supervised learning. The state-of-the-art approaches "Decoupling" and "Co-teaching+" claim that the "disagreement" strategy is crucial for alleviating the problem of learning with noisy labels. In this paper, we start from a different perspective and propose a robust learning paradigm called JoCoR, which aims to reduce the diversity of two networks during training. Specifically, we first use two networks to make predictions on the same mini-batch data and calculate a joint loss with Co-Regularization for each training example. Then we select small-loss examples to update the parameters of both two networks simultaneously. Trained by the joint loss, these two networks would be more and more similar due to the effect of Co-Regularization. Extensive experimental results on corrupted data from benchmark datasets including MNIST, CIFAR-10, CIFAR-100 and Clothing1M demonstrate that JoCoR is superior to many state-of-the-art approaches for learning with noisy labels.

1. Introduction

Deep Neural Networks (DNNs) achieve remarkable success on various tasks, and most of them are trained in a supervised manner, which heavily relies on a large number of training instances with accurate labels [14]. However, collecting large-scale datasets with fully precise annotations is expensive and time-consuming. To alleviate this problem, data annotation companies choose some alternating methods such as crowdsourcing [39, 43] and online queries [3] to improve labelling efficiency. Unfortunately, these methods usually suffer from unavoidable noisy labels, which have been proven to lead to noticeable decrease in performance of DNNs [1, 44].

As this problem has severely limited the expansion of neural network applications, a large number of algorithms

_____
*Corresponding author.

have been developed for learning with noisy labels, which belongs to the family of weakly supervised learning frameworks [2, 5, 6, 7, 8, 9, 11]. Some of them focus on improving the methods to estimate the latent noisy transition matrix [21, 24, 32]. However, it is challenging to estimate the noise transition matrix accurately. An alternative approach is training on selected or weighted samples, e.g., Mentornet [16], gradient-based reweight [30] and Co-teaching [12]. Furthermore, the state-of-the-art methods including Co-teaching+ [41] and Decoupling [23] have shown excellent performance in learning with noisy labels by introducing the "Disagreement" strategy, where "when to update" depends on a disagreement between two different networks. However, there are only a part of training examples that can be selected by the "Disagreement" strategy, and these examples cannot be guaranteed to have ground-truth labels [12]. Therefore, there arises a question to be answered: Is "Disagreement" necessary for training two networks to deal with noisy labels?

Motivated by Co-training for multi-view learning and semi-supervised learning that aims to maximize the agreement on multiple distinct views [4, 19, 34, 45], a straightforward method for handling noisy labels is to apply the regularization from peer networks when training each single network. However, although the regularization may improve the generalization ability of networks by encouraging agreement between them, it still suffers from memorization effects on noisy labels [44]. To address this problem, we propose a novel approach named JoCoR (**Jo**int Training with **Co-R**egularization). Specifically, we train two networks with a joint loss, including the conventional supervised loss and the Co-Regularization loss. Furthermore, we use the joint loss to select small-loss examples, thereby ensuring the error flow from the biased selection would not be accumulated in a single network.

To show that JoCoR significantly improves the robustness of deep learning on noisy labels, we conduct extensive experiments on both simulated and real-world noisy datasets, including MNIST, CIFAR-10, CIFAR-100 and

Clothing1M datasets. Empirical results demonstrate that the robustness of deep models trained by our proposed approach is superior to many state-of-the-art approaches. Furthermore, the ablation studies clearly demonstrate the effectiveness of Co-Regularization and Joint Training.

2. Related work

In this section, we briefly review existing works on learning with noisy labels.

Noise rate estimation. The early methods focus on estimating the label transition matrix [24, 25, 28, 37]. For example, F-correction [28] uses a two-step solution to heuristically estimate the noise transition matrix. An additional softmax layer is introduced to model the noise transition matrix [10]. In these approaches, the quality of noise rate estimation is a critical factor for improving robustness. However, noise rate estimation is challenging, especially on datasets with a large number of classes.

Small-loss selection. Recently, a promising method of handling noisy labels is to train models on small-loss instances [30]. Intuitively, the performance of DNNs will be better if the training data become less noisy. Previous work showed that during training, DNNs tend to learn simple patterns first, then gradually memorize all samples [1], which justifies the widely used *small-loss criterion*: treating samples with small training loss as clean ones. In particular, MentorNet [16] firstly trains a teacher network, then uses it to select clean instances for guiding the training of the student network. As for Co-teaching [12], in each mini-batch of data, each network chooses its small-loss instances and exchanges them with its peer network for updating the parameters. The authors argued that these two networks could filter different types of errors brought by noisy labels since they have different learning abilities. When the error from noisy data flows into the peer network, it will attenuate this error due to its robustness.

Disagreement. The "Disagreement" strategy is also applied to this problem. For instance, Decoupling [23] updates the model only using instances on which the predictions of two different networks are different. The idea of disagreement-update is similar to hard example mining [33], which trains model with examples that are misclassified and expects these examples to help steer the classifier away from its current mistakes. For the "Disagreement" strategy, the decision of "when to update" depends on a disagreement between two networks instead of depending on the label. As a result, it would help decrease the divergence between these networks. However, as noisy labels are spread across the whole space of examples, there may be many noisy labels in the disagreement area, where the Decoupling approach cannot handle noisy labels explicitly. Combining the "Disagreement" strategy with cross-update in Co-teaching, Co-teaching+ [41] achieves excel-
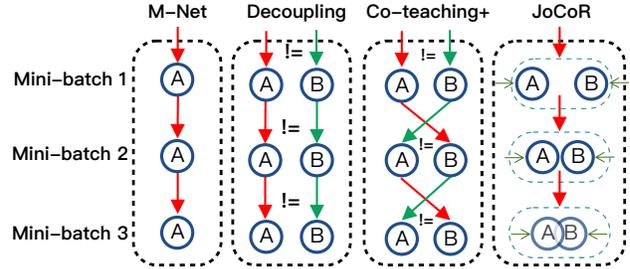

Figure 1. Comparison of error flow among MentorNet (M-Net) [16], Decoupling [23], Co-teaching+ [41] and JoCoR. Assume that the error flow comes from the biased selection of training instances, and error flow from network A or B is denoted by red arrows or green arrows, respectively. **First panel**: M-Net maintains only one network (A). **Second panel**: Decoupling maintains two networks (A&B). The parameters of two networks are updated, when the predictions of them disagree (!=). **Third panel**: In Co-teaching+, each network teaches its small-loss instances with prediction disagreement (!=) to its peer network. **Fourth panel**: JoCoR also maintains two networks (A&B) but trains them as a whole with a joint loss, which makes predictions of each network closer to ground true labels and peer network's.

lent performance in improving the robustness of DNNs against noisy labels. In spite of that, Co-teaching+ only selects small-loss instances with different predictions from two models so very few examples are utilized for training in each mini-batch when datasets are with extremely high noise rate. It would prevent the training process from efficient use of training examples. This phenomenon will be explicitly shown in our experiments in the symmetric-80% label noise case.

Other deep learning methods. In addition to the aforementioned approaches, there are some other deep learning solutions [13, 17] to deal with noisy labels, including pseudo-label based [35, 40] and robust loss based approaches [28, 46]. For pseudo-label based approaches, Joint optimization [35] learns network parameters and infers the ground-true labels simultaneously. PENCIL [40] adopts label probability distributions to supervise network learning and to update these distributions through back-propagation end-to-end in each epoch. For robust loss based approaches, F-correct[28] proposes a robust risk minimization method to learn neural networks for multi-class classification by estimating label corruption probabilities. GCE [46] combines the advantages of the mean absolute loss and the cross entropy loss to obtain a better loss function and presents a theoretical analysis of the proposed loss functions in the context of noisy labels.

Semi-supervised learning. Semi-supervised learning also belongs to the family of weakly supervised learning frameworks [15, 18, 22, 26, 27, 31, 47]. There are some interesting works from semi-supervised learning that are highly relevant to our approach. In contrast to "Disagreement" strategy, many of them are based on a agree-

ment maximization algorithm. Co-RLS [34] extends standard regularization methods like Support Vector Machines (SVM) and Regularized Least squares (RLS) to multi-view semi-supervised learning by optimizing measures of agreement and smoothness over labelled and unlabelled examples. EA++ [19] is a co-regularization based approach for semi-supervised domain adaptation, which builds on the notion of augmented space and harnesses unlabeled data in the target domain to further assist the transfer of information from source to target. The intuition is that different models in each view would agree on the labels of most examples, and it is unlikely for compatible classifiers trained on independent views to agree on an incorrect label. This intuition also motivates us to deal with noisy labels based on the agreement maximization principle.

3. The Proposed Approach

As mentioned before, we suggest to apply the agreement maximization principle to tackle the problem of noisy labels. In our approach, we encourage two different classifiers to make predictions closer to each other by explicit regularization method instead of hard sampling employed by the "Disagreement" strategy. This method could be considered as a meta-algorithm that trains two base classifiers by one loss function, which includes a regularization term to reduce divergence between the two classifiers.

For multi-class classification with M classes, we suppose the dataset with N samples is given as $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, where $\boldsymbol{x}_i$ is the i-th instance with its observed label as $y_i \in \{1, \ldots, M\}$. Similar to Decoupling and Co-teaching+, we formulate the proposed JoCoR approach with two deep neural networks denoted by $f(\boldsymbol{x}, \boldsymbol{\Theta}_1)$ and $f(\boldsymbol{x}, \boldsymbol{\Theta}_2)$, while $\boldsymbol{p}_1 = [p_1^1, p_1^2, \ldots, p_1^M]$ and $\boldsymbol{p}_2 = [p_2^1, p_2^2, \ldots, p_2^M]$ denote their prediction probabilities of instance $\boldsymbol{x}_i$, respectively. In other words, $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are the outputs of the "softmax" layer in $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$.

Network. For JoCoR, each network can be used to predict labels alone, but during the training stage these two networks are trained with a pseudo-siamese paradigm, which means their parameters are different but updated simultaneously by a joint loss (see Figure 2). In this work, we call this paradigm as "Joint Training".

Specifically, our proposed loss function ℓ on $\boldsymbol{x}_i$ is constructed as follows:

$$\ell(x_i) = (1 - \lambda) * \ell_{\text{sup}}(\boldsymbol{x}_i, y_i) + \lambda * \ell_{\text{con}}(\boldsymbol{x}_i) \quad (1)$$

In the loss function, the first part ℓ_{sup} is conventional supervised learning loss of the two networks, the second part ℓ_{con} is the contrastive loss between predictions of the two networks for achieving Co-Regularization.

Classification loss. For multi-class classification, we use Cross-Entropy Loss as the supervised part to minimize the distance between predictions and labels.
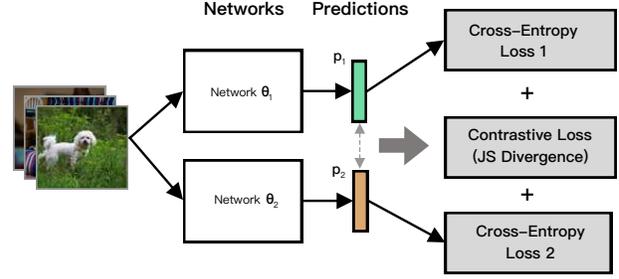

Figure 2. JoCoR schematic.

$$\begin{aligned} \ell_{\text{sup}}(\boldsymbol{x}_i, y_i) &= \ell_{\text{C1}}(\boldsymbol{x}_i, y_i) + \ell_{\text{C2}}(\boldsymbol{x}_i, y_i) \\ &= -\sum\nolimits_{i=1}^N \sum\nolimits_{m=1}^M y_i \log(p_1^m(\boldsymbol{x}_i)) \\ &\quad - \sum\nolimits_{i=1}^N \sum\nolimits_{m=1}^M y_i \log(p_2^m(\boldsymbol{x}_i)) \end{aligned} \quad (2)$$

Intuitively, two networks can filter different types of errors brought by noisy labels since they have different learning abilities. In Co-teaching [12], when the two networks exchange the selected small-loss instances in each mini-batch data, the error flows can be reduced by peer networks mutually. By virtue of the joint-training paradigm, our JoCoR would consider the classification losses from both two networks during the "small-loss" selection stage. In this way, JoCoR can share the same advantage of the cross-update strategy in Co-teaching. This argument will be clearly supported by the ablation study in the later section.

Contrastive loss. From the view of *agreement maximization principles* [4, 34], different models would agree on labels of most examples, and they are unlikely to agree on incorrect labels. Based on this observation, we apply the Co-Regularization method to maximize the agreement between two classifiers. On one hand, the Co-Regularization term could help our algorithm select examples with clean labels since an example with small Co-Regularization loss means that two networks reach an agreement on its predictions. On the other hand, the regularization from peer networks helps the model find a much wider minimum, which is expected to provide better generalization performance [45].

In JoCoR, we utilize the contrastive term as Co-Regularization to make the networks guide each other. To measure the match of the two networks' predictions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, we adopt the Jensen-Shannon (JS) Divergence. To simplify implementation, we could use the symmetric Kullback-Leibler(KL) Divergence to surrogate this term.

$$\ell_{\text{con}} = D_{\text{KL}}(\boldsymbol{p}_1 \| \boldsymbol{p}_2) + D_{\text{KL}}(\boldsymbol{p}_2 \| \boldsymbol{p}_1) \quad (3)$$

where

$$D_{\text{KL}}(\boldsymbol{p}_1 \| \boldsymbol{p}_2) = \sum\nolimits_{i=1}^N \sum\nolimits_{m=1}^M p_1^m(\boldsymbol{x}_i) \log \frac{p_1^m(\boldsymbol{x}_i)}{p_2^m(\boldsymbol{x}_i)}$$

$$D_{\text{KL}}(\boldsymbol{p}_2 \| \boldsymbol{p}_1) = \sum\nolimits_{i=1}^N \sum\nolimits_{m=1}^M p_2^m(\boldsymbol{x}_i) \log \frac{p_2^m(\boldsymbol{x}_i)}{p_1^m(\boldsymbol{x}_i)}$$

Algorithm 1 JoCoR

Input: Network f with $\Theta = \{\Theta_1, \Theta_2\}$, learning rate η, fixed τ, epoch T_k and $T_{\max}$, iteration $I_{\max}$;

  1: **for** $t = 1,2,\ldots,T_{\max}$ **do**
  2:     **Shuffle** training set D;
  3:     **for** $n = 1,\ldots,I_{\max}$ **do**
  4:         **Fetch** mini-batch D_n from D;
  5:         $\boldsymbol{p}_1 = f(\boldsymbol{x}, \Theta_1), \forall \boldsymbol{x} \in D_n$;
  6:         $\boldsymbol{p}_2 = f(\boldsymbol{x}, \Theta_2), \forall \boldsymbol{x} \in D_n$;
  7:         **Calculate** the joint loss ℓ by (1) using $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$;
  8:         **Obtain** small-loss sets $\tilde{D}_n$ by (4) from D_n;
  9:         **Obtain** L by (5) on $\tilde{D}_n$;
10:         **Update** $\Theta = \Theta - \eta \nabla L$;
11:     **end for**
12:     **Update** $R(t) = 1 - \min\left\{\frac{t}{T_k}\tau, \tau\right\}$
13: **end for**

Output: Θ_1 and Θ_2

Small-loss Selection Before introducing the details, we first clarify the connection between small losses and clean instances. Intuitively, small-loss examples are likely to be the ones that are correctly labelled [12, 30]. Thus, if we train our classifier only using small-loss instances in each mini-batch data, it would be resistant to noisy labels.

To handle noisy labels, we apply the "small-loss" criterion to select "clean" instances (step 8 in Algorithm 1). Following the setting of Co-teaching+, we update $R(t)$ (step 12), which controls how many small-loss data should be selected in each training epoch. At the beginning of training, we keep more small-loss data (with a large $R(t)$) in each mini-natch since deep networks would fit clean data first [1, 44]. With the increase of epochs, we reduce $R(t)$ gradually until reaching $1 - \tau$, keeping fewer examples in each mini-batch. Such operation will prevent deep networks from over-fitting noisy data [12].

In our algorithm, we use the joint loss (1) to select small-loss examples. Intuitively, an instance with small joint loss means that both two networks could be easy to reach a consensus and make correct predictions on it. As two networks have different learning abilities based on different initial conditions, the selected small-loss instances are more likely to be with clean labels than those chosen by a single model. Specifically, we conduct small-loss selection as follows:

$$\tilde{D}_n = \arg\min_{D'_n : |D'_n| \geq R(t)|D_n|} \ell\left(D'_n\right) \qquad (4)$$

After obtaining the small-loss instances, we calculate the average loss on these examples for further backpropagation:

$$L = \frac{1}{|\tilde{D}|} \sum_{\boldsymbol{x} \in \tilde{D}} \ell(\boldsymbol{x}) \qquad (5)$$

Table 1. Comparison of state-of-the-art and related techniques with our JoCoR approach. In the first column, "small loss": regarding small-loss samples as "clean" samples, which is based on the memorization effects of deep neural networks; "double classifiers": training two classifiers simultaneously; "cross update": updating parameters in a cross manner instead of a parallel manner; "joint training": training two classifiers with a joint loss; "disagreement": updating two classifiers on disagreed examples during the entire training epochs; "agreement": maximizing the agreement of two classifiers by regularization during the whole training epochs.

	Decoupling	Co-teaching	Co-teaching+	JoCoR
small loss	✗	✓	✓	✓
cross update	✗	✓	✓	✗
joint training	✗	✗	✗	✓
disagreement	✓	✗	✓	✗
agreement	✗	✗	✗	✓

Relations to other approaches. We compare JoCoR with other related approaches in Table 1. Specifically, Decoupling applies the "disagreement" strategy to select instances while Co-teaching use small-loss criterion. Besides, Co-teaching updates parameters of networks by the "cross-update" strategy to reduce the accumulated error flow. Combining the "disagreement" strategy and the "cross-update" strategy, Co-teaching+ achieves excellent performance. As for our JoCoR, we also select small-loss examples but update the networks by Joint Training. Furthermore, we use the Co-Regularization to maximize agreement between the two networks. Note that Co-Regularization in our proposed method and "disagreement" strategy in Decoupling are both essentially to reduce the divergence between the two classifiers. The difference between them lies in that the former uses an explicit regularization methods with all training examples while the latter employs hard sampling that reduces the effective number of training examples. It is especially important in the case of small-loss selection, because the selection would further decrease the effective number of training examples.

4. Experiments

In this section we first compare JoCoR with some state-of-the-art approaches, then analyze the impact of Joint Training and Co-Regularization by ablation study. We also analyze the effect of λ in (1) by sensitivity analysis and put it in supplementary materials.

4.1. Experiment setup

Datasets. We verify the effectiveness of our proposed algorithm on four benchmark datasets: *MNIST*, *CIFAR-10*, *CIFAR-100* and *Clothing1M* [38], and the detailed characteristics of these datasets can be found in supplementary materials. These datasets are popularly used for the evaluation of learning with noisy labels in previous literatures [10, 18, 29]. Especially, *Clothing1M* is a large-scale real-
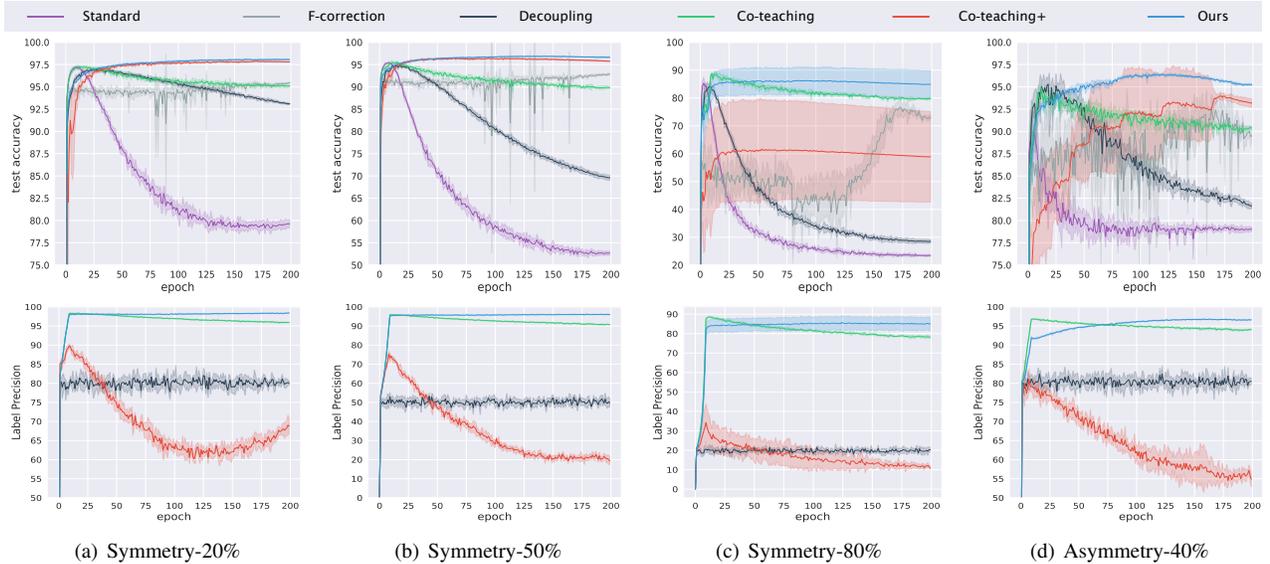
Figure 3. Results on MNIST dataset. Top: test accuracy(%) vs. epochs; bottom: label precision(%) vs. epochs.

(a) Symmetry-20%　　(b) Symmetry-50%　　(c) Symmetry-80%　　(d) Asymmetry-40%

Table 2. Average test accuracy (%) on *MNIST* over the last 10 epochs.

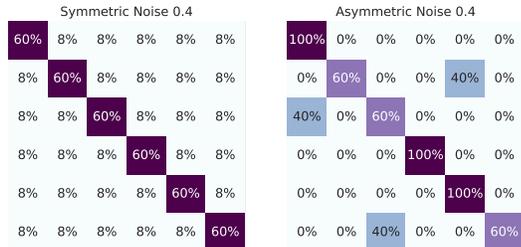| Flipping-Rate	Standard	F-correction	Decoupling	Co-teaching	Co-teaching+	JoCoR
Symmetry-20%	79.56 ± 0.44	95.38 ± 0.10	93.16 ± 0.11	95.10 ± 0.16	97.81 ± 0.03	$\mathbf{98.06} \pm 0.04$
Symmetry-50%	52.66 ± 0.43	92.74 ± 0.21	69.79 ± 0.52	89.82 ± 0.31	95.80 ± 0.09	$\mathbf{96.64} \pm 0.12$
Symmetry-80%	23.43 ± 0.31	72.96 ± 0.90	28.51 ± 0.65	79.73 ± 0.35	58.92 ± 14.73	$\mathbf{84.89} \pm 4.55$
Asymmetry-40%	79.00 ± 0.28	89.77 ± 0.96	81.84 ± 0.38	90.28 ± 0.27	93.28 ± 0.43	$\mathbf{95.24} \pm 0.10$

Figure 4. Example of noise transition matrix T (taking 6 classes and noise ratio 0.4 as an example)

world dataset with noisy labels, which is widely used in the related works[20, 28, 40, 38].

Since all datasets are clean except *Clothing1M*, following [28, 29], we need to corrupt these datasets manually by the label transition matrix Q, where $Q_{ij} = \Pr[\tilde{y} = j | y = i]$ given that noisy $\tilde{y}$ is flipped from clean y. Assume that the matrix Q has two representative structures: (1) Symmetry flipping [36]; (2) Asymmetry flipping [28]: simulation of fine-grained classification with noisy labels, where labellers may make mistakes only within very similar classes.

Following F-correction [28], only half of the classes in the dataset are with noisy labels in the setting of asymmetric noise, so the actual noise rate in the whole dataset τ is half of the noisy rate in the noisy classes. Specifically, when the asymmetric noise rate is 0.4, it means $\tau = 0.2$. Figure 4 shows an example of noise transition matrix.

For experiments on *Clothing1M*, we use the 1M images with noisy labels for training, the 14k and 10k clean data for validation and test, respectively. Note that we do not use the 50k clean training data in all the experiments because only noisy labels are required during the training process [20, 35]. For preprocessing, we resize the image to 256×256, crop the middle 224×224 as input, and perform normalization.

Baselines. We compare JoCoR (Algorithm 1) with the following state-of-the-art algorithms, and implement all methods with default parameters by PyTorch, and conduct all the experiments on NVIDIA Tesla V100 GPU.

(i) Co-teaching+ [41], which trains two deep neural networks and consists of disagreement-update step and cross-update step.

(ii) Co-teaching [12], which trains two networks simultaneously and cross-updates parameters of peer networks.

(iii) Decoupling [23], which updates the parameters only using instances which have different predictions from two classifiers.

(iv) F-correction [28], which corrects the prediction by the label transition matrix. As suggested by the authors, we first train a standard network to estimate the transition matrix Q.
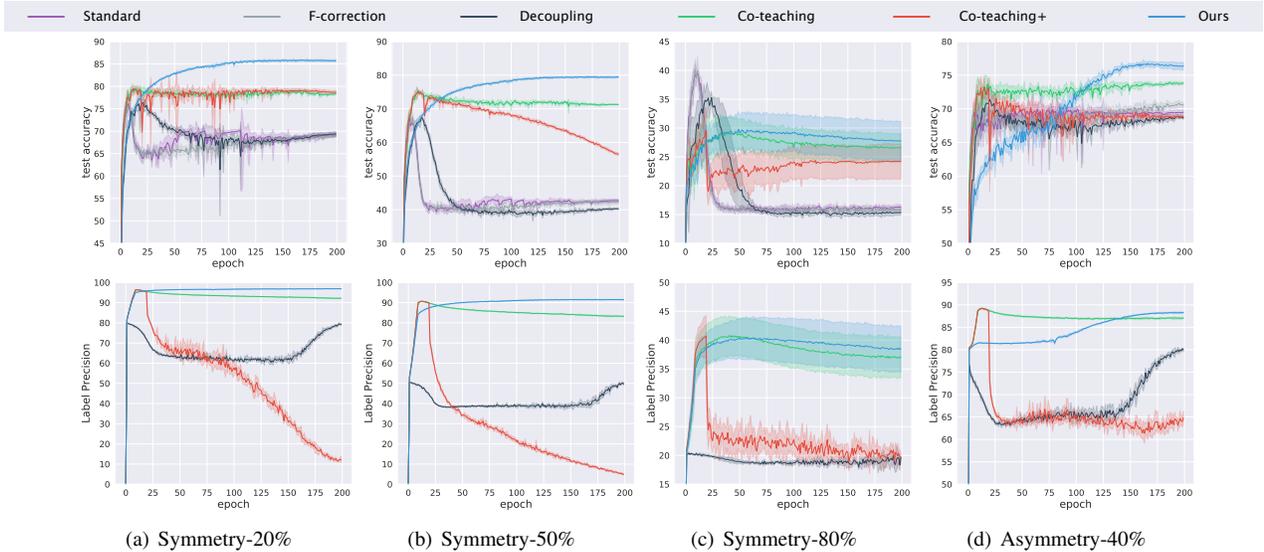
Figure 5. Results on CIFAR-10 dataset. Top: test accuracy(%) vs. epochs; bottom: label precision(%) vs. epochs.

Table 3. Average test accuracy (%) on *CIFAR-10* over the last 10 epochs.

Flipping-Rate	Standard	F-correction	Decoupling	Co-teaching	Co-teaching+	JoCoR
Symmetry-20%	69.18 ± 0.52	68.74 ± 0.20	69.32 ± 0.40	78.23 ± 0.27	78.71 ± 0.34	$\mathbf{85.73} \pm 0.19$
Symmetry-50%	42.71 ± 0.42	42.19 ± 0.60	40.22 ± 0.30	71.30 ± 0.13	57.05 ± 0.54	$\mathbf{79.41} \pm 0.25$
Symmetry-80%	16.24 ± 0.39	15.88 ± 0.42	15.31 ± 0.43	26.58 ± 2.22	24.19 ± 2.74	$\mathbf{27.78} \pm 3.06$
Asymmetry-40%	69.43 ± 0.33	70.60 ± 0.40	68.72 ± 0.30	73.78 ± 0.22	68.84 ± 0.20	$\mathbf{76.36} \pm 0.49$

(v) As a simple baseline, we compare JoCoR with the standard deep network that directly trains on noisy datasets (abbreviated as Standard).

Network Structure and Optimizer. We use a 2-layer MLP for *MNIST*, a 7-layer CNN network architecture for *CIFAR-10* and *CIFAR-100*. The detailed information can be found in supplementary materials. For *Clothing1M*, we use ResNet with 18 layers.

For experiments on *MNIST*, *CIFAR-10* and *CIFAR-100*, Adam optimizer (momentum=0.9) is used with an initial learning rate of 0.001, and the batch size is set to 128. We run 200 epochs in total and linearly decay learning rate to zero from 80 to 200 epochs.

For experiments on *Clothing1M*, we also use Adam optimizer (momentum=0.9) and set batch size to 64. During the training stage, we run 15 epochs in total and set learning rate to 8×10^{-4}, 5×10^{-4} and 5×10^{-5} for 5 epochs each.

As for λ in our loss function (1), we search it in [0.05, 0.10, 0.15,...,0.95] with a clean validation set for best performance. When validation set is also with noisy labels, we use the small-loss selection to choose a clean subset for validation. As deep networks are highly nonconvex, even with the same network and optimization method, different initializations can lead to different local optimum. Thus, following Decoupling [23], we also take two networks with the same architecture but different initializations as two classi-

fiers.

Measurement. To measure the performance, we use the test accuracy, i.e., *test accuracy = (# of correct predictions) / (# of test)*. Besides, we also use the label precision in each mini-batch, i.e., *label precision = (# of clean labels) / (# of all selected labels)*. Specifically, we sample $R(t)$ of small-loss instances in each mini-batch and then calculate the ratio of clean labels in the small-loss instances. Intuitively, higher label precision means less noisy instances in the mini-batch after sample selection, so the algorithm with higher label precision is also more robust to the label noise. All experiments are repeated five times. The error bar for STD in each figure has been highlighted as a shade.

Selection setting. Following Co-teaching, we assume that the noise rate τ is known. To conduct a fair comparison in benchmark datasets, we set the ratio of small-loss samples $R(t)$ as identical: $R(t) = 1 - \min\left\{\frac{t}{T_k}\tau, \tau\right\}$, where $T_k = 10$ for MNIST, CIFAR-10 and CIFAR100, $T_k = 5$ for Clothing1M. If τ is not known in advance, τ can be inferred using validation sets [21, 42].

4.2. Comparison with the State-of-the-Arts

Results on *MNIST*. At the top of Figure 3, it shows test accuracy vs. epochs on *MNIST*. In all four plots, we can see the memorization effect of networks, i.e., test accuracy of Standard first reaches a very high level and then gradu-
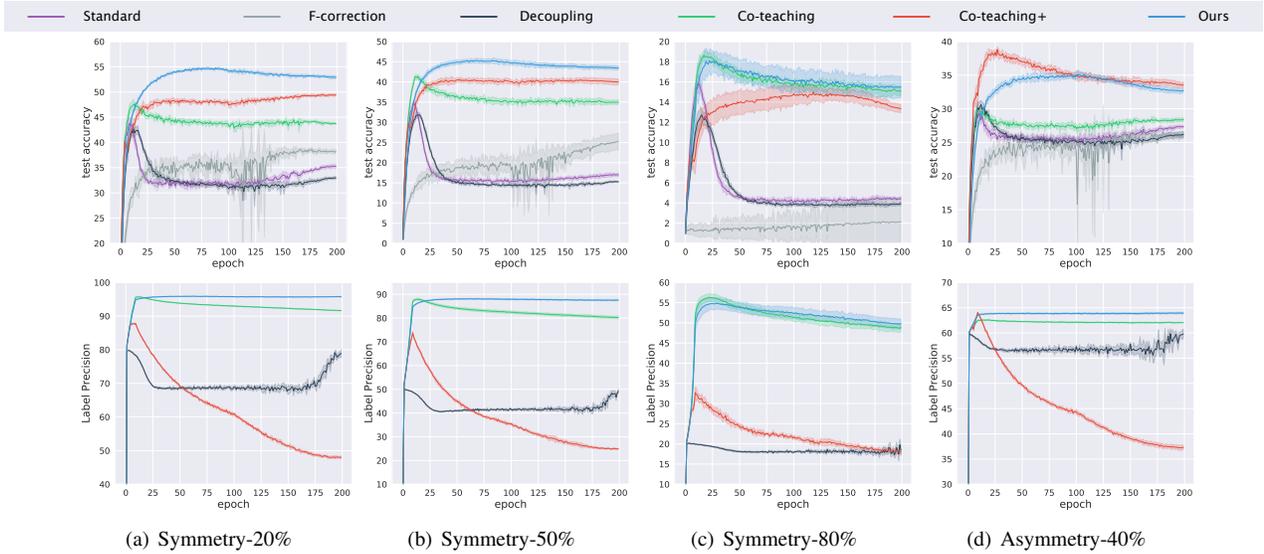
Figure 6. Results on CIFAR-100 dataset. Top: test accuracy(%) vs. epochs; bottom: label precision(%) vs. epochs.

Table 4. Average test accuracy (%) on *CIFAR-100* over the last 10 epochs.

Flipping-Rate	Standard	F-correction	Decoupling	Co-teaching	Co-teaching+	JoCoR
Symmetry-20%	35.14 ± 0.44	37.95 ± 0.10	33.10 ± 0.12	43.73 ± 0.16	49.27 ± 0.03	$\mathbf{53.01} \pm 0.04$
Symmetry-50%	16.97 ± 0.40	24.98 ± 1.82	15.25 ± 0.20	34.96 ± 0.50	40.04 ± 0.70	$\mathbf{43.49} \pm 0.46$
Symmetry-80%	4.41 ± 0.14	2.10 ± 2.23	3.89 ± 0.16	15.15 ± 0.46	13.44 ± 0.37	$\mathbf{15.49} \pm 0.98$
Asymmetry-40%	27.29 ± 0.25	25.94 ± 0.44	26.11 ± 0.39	28.35 ± 0.25	$\mathbf{33.62} \pm 0.39$	32.70 ± 0.35

ally decreases. Thus, a good robust training method should stop or alleviate the decreasing process. On this point, Jo-CoR consistently achieves higher accuracy than all the other baselines in all four cases.

We can compare the test accuracy of different algorithms in detail in Table 2. In the most natural Symmetry-20% case, all new approaches work better than Standard obviously, which demonstrates their robustness. Among them, JoCoR and Co-teaching+ work significantly better than other methods. When it goes to Symmetry-50% case and Asymmetry-40% case, Decoupling begins to fail while other methods still work fine, especially JoCoR and Co-teaching+. However, Co-teaching+ cannot combat with the hardest Symmetry-80% case, where it only achieves 58.92%. In this case, JoCoR achieves the best average classification accuracy (84.89%) again.

To explain such excellent performance, we plot label precision vs. epochs at the bottom of Figure 3. Only Decoupling, Co-teaching, Co-teaching+ and JoCoR are considered here, as they include example selection during training. First, we can see both JoCoR and Co-teaching can successfully pick clean instances out. Note that JoCoR not only reaches high label precision in all four cases but also performs better and better with the increase of epochs while Co-teaching declines gradually after reaching the top. This shows that our approach is better at finding clean instances. Then, Decoupling and Co-teaching+ fail in selecting clean

examples. As mentioned in Related Work, very few examples are utilized by Co-teaching+ in the training process when noise rate goes to be extremely high. In this way, we can understand why Co-teaching+ performs poorly on the hardest case.

Results on *CIFAR-10*. Table 3 shows test accuracy on *CIFAR-10*. As we can see, JoCoR performs the best in all four cases again. In the Symmetric-20% case, JoCoR works much better than all other baselines and Co-teaching+ performs better than Co-teaching and Decoupling. In the other three cases, JoCoR is still the best and Co-teaching+ cannot even achieve comparable performance with Co-teaching.

Figure 5 shows test accuracy and label precision vs. epochs. JoCoR outperforms all the other comparing approaches on both test accuracy and label precision. On label precision, while Decoupling and Co-teaching+ fail to find clean instances, both JoCoR and Co-teaching can do this. An interesting phenomenon is that in the Asymmetry-40% case, although Co-teaching can achieve better performance than JoCoR in the first 100 epochs, JoCoR consistently outperforms it in all the later epochs. The result shows that JoCoR has better generalization ability than Co-teaching.

Results on *CIFAR-100*. Then, we show our results on *CIFAR-100*. The test accuracy is shown in Table 4. Test accuracy and label precision vs. epochs are shown in Figure 6. Note that there are only 10 classes in *MNIST* and *CIFAR-10* datasets. Thus, overall the accuracy is much

Table 5. Classification accuracy (%) on the *Clothing1M* test set

Methods	*best*	*last*
Standard	67.22	64.68
F-correction	68.93	65.36
Decoupling	68.48	67.32
Co-teaching	69.21	68.51
Co-teaching+	59.32	58.79
JoCoR	**70.30**	**69.79**

lower than previous ones in Tables 2 and 3. But JoCoR still achieves high test accuracy on this datasets. In the easiest Symmetry-20% and Symmetry-50% cases, JoCoR works significantly better than Co-teaching+, Co-teaching and other methods. In the hardest Symmetry-80% case, JoCoR and Co-teaching tie together but JoCoR still gets higher testing accuracy. When it turns to Asymmetry-40% case, JoCoR and Co-teaching+ perform much better than other methods. On label precision, JoCoR keeps the best performance in all four cases.

Results on *Clothing1M*. Finally, we demonstrate the efficacy of the proposed method on the real-world noisy labels using the *Clothing1M* dataset. As shown in Table 5, *best* denotes the scores of the epoch where the validation accuracy is optimal, and *last* denotes the scores at the end of training. The proposed JoCoR method gets better result than state-of-the-art methods on *best*. After all epochs, JoCoR achieves a significant improvement in accuracy of +5.11 over Standard, and an improvement of +1.28 over the best baseline method.

4.3. Ablation Study

To conduct ablation study for analyzing the effect of Co-Regularization, we set up the experiments above *MNIST* and *CIFAR-10* with Symmetry-50% noise. For implementing Joint Training without Co-Regularization (*Joint-only*), we set the λ in (1) to 0. Besides, to verify the effect of the Joint Training paradigm, we introduce Co-teaching and Standard enhanced by "small-loss" selection (*Standard+*) to join the comparison. Recall that the joint-training method selects examples by the joint loss while Co-teaching uses cross-update method to reduce the error flow [12], these two methods should play a similar role during training according to the previous analysis.

The test accuracy and label precision vs. epochs on *MNIST* are shown in Figure 7. As we can see, JoCoR performs much better than the others on both test accuracy and label precision. The former keeps almost no decrease while the latter decline a lot after reaching the top. This observation indicates that Co-Regularization strongly hinders neural networks from memorizing noisy labels.

The test accuracy and label precision vs. epochs on *CIFAR-10* are shown in Figure 8. In this figure, JoCoR still maintains a huge advantage over the other three meth-
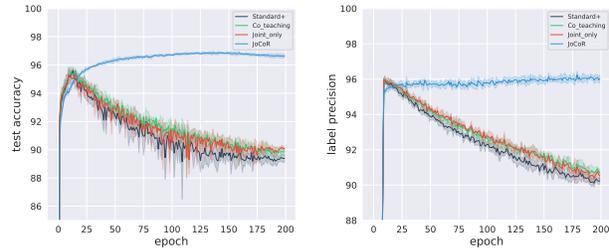

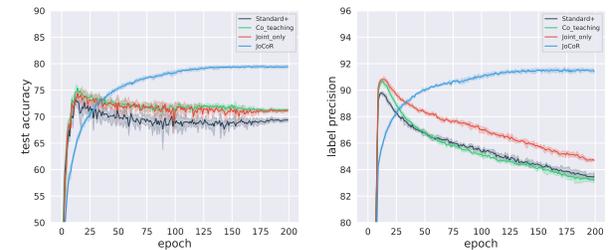Figure 7. Results of ablation study on *MNIST*

Figure 8. Results of ablation study on *CIFAR-10*

ods on both test accuracy and label precision while Joint-only, Co-teaching and Standard+ remain the same trend as these for *MNIST*, keeping a downward tendency after increasing to the highest point. These results show that Co-Regularization plays a vital role in handling noisy labels. Moreover, Joint-only achieves a comparable performance with Co-teaching on test accuracy and performs better than Co-teaching and Standard+ on label precision. It shows that Joint Training is a more efficient paradigm to help select clean examples than cross-update in Co-teaching.

5. Conclusion

The paper proposes an effective approach called JoCoR to improve the robustness of deep neural networks with noisy labels. The key idea of JoCoR is to train two classifiers simultaneously with one joint loss, which is composed of regular supervised part and Co-Regularized part. Similar to Co-teaching+, we also select small-loss instances to update networks in each mini-batch data by the joint loss. We conduct experiments on *MNIST*, *CIFAR-10*, *CIFAR-100* and *Clothing1M* to demonstrate that, JoCoR can train deep models robustly with the slightly and extremely noisy supervision. Furthermore, the ablation studies clearly demonstrate the effectiveness of Co-Regularization and Joint Training. In future work, we will explore the theoretical foundation of JoCoR based on the view of traditional Co-training algorithms [19, 34].

References

[1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 233–242, 2017.

[2] H. Bao, G. Niu, and M. Sugiyama. Classification from pairwise similarity and unlabeled data. In *International Conference on Machine Learning*, pages 452–461, 2018.

[3] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003.

[4] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[5] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[6] M. C. du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, pages 703–711, 2014.

[7] Lei Feng and Bo An. Leveraging latent label distributions for partial label learning. In *International Joint Conferences on Artificial Intelligence*, pages 2107–2113, 2018.

[8] Lei Feng and Bo An. Partial label learning by semantic difference maximization. In *International Joint Conferences on Artificial Intelligence*, pages 2294–2300, 2019.

[9] Lei Feng and Bo An. Partial label learning with self-guided retraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3542–3549, 2019.

[10] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *Proceedings of the 5th International Conference on Learning Representation*, 2016.

[11] Chen Gong, Hengmin Zhang, Jian Yang, and Dacheng Tao. Learning with inadequate and incorrect supervision. In *2017 IEEE International Conference on Data Mining)*, pages 889–894. IEEE, 2017.

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pages 8527–8537, 2018.

[13] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5138–5147, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] T. Ishida, G. Niu, and M. Sugiyama. Binary classification for positive-confidence data. In *Advances in Neural Information Processing Systems*, pages 5917–5928, 2018.

[16] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.

[17] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 101–110, 2019.

[18] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, pages 1675–1685, 2017.

[19] Abhishek Kumar, Avishek Saha, and Hal Daume. Co-regularization based semi-supervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 478–486, 2010.

[20] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019.

[21] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2015.

[22] Nan Lu, Gang Niu, Aditya K. Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. In *Proceedings of the International Conference on Learning Representation*, 2019.

[23] Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". In *Advances in Neural Information Processing Systems*, pages 960–970, 2017.

[24] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.

[25] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

[26] G. Niu, M. C. du Plessis, T. Sakai, Y. Ma, and M. Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in Neural Information Processing Systems*, pages 1199–1207, 2016.

[27] Gang Niu, Wittawat Jitkrittum, Bo Dai, Hirotaka Hachiya, and Masashi Sugiyama. Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning. In *International Conference on Machine Learning*, pages 10–18, 2013.

[28] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[29] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Train-

ing deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[30] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018.

[31] Tomoya Sakai, Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *International Conference on Machine Learning*, pages 2998–3006, 2017.

[32] Tyler Sanderson and Clayton Scott. Class proportion estimation with application to multiclass anomaly rejection. In *Artificial Intelligence and Statistics*, pages 850–858, 2014.

[33] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[34] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning With Multiple Views*, pages 74–79, 2005.

[35] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.

[36] Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pages 10–18, 2015.

[37] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing Systems*, pages 6835–6846, 2019.

[38] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015.

[39] Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327, 2014.

[40] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. *arXiv preprint arXiv:1903.07788*, 2019.

[41] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement benefit co-teaching? *arXiv preprint arXiv:1901.04215*, 2019.

[42] Xiyu Yu, Tongliang Liu, Mingming Gong, Kayhan Batmanghelich, and Dacheng Tao. An efficient and provable approach for mixture proportion estimation using linear independence assumption. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4489, 2018.

[43] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *Proceedings of the European Conference on Computer Vision*, pages 68–83, 2018.

[44] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[45] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[46] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, pages 8778–8788, 2018.

[47] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.