Decoupling Representation and Classifier for Noisy Label Learning

Hui Zhang[1],   Quanming Yao[1,2]

[1]4Paradigm Inc.   [1,2]Department of Electronic Engineering, Tsinghua University

{zhanghui, yaoquanming}@4paradigm.com

Abstract

Since convolutional neural networks (ConvNets) can easily memorize noisy labels, which are ubiquitous in visual classification tasks, it has been a great challenge to train ConvNets against them robustly. Various solutions, e.g., sample selection, label correction, and robustifying loss functions, have been proposed for this challenge, and most of them stick to the end-to-end training of the representation (feature extractor) and classifier. In this paper, by a deep re-thinking and careful re-examining on learning behaviors of the representation and classifier, we discover that the representation is much more fragile in the presence of noisy labels than the classifier. Thus, we are motivated to design a new method, i.e., REED, to leverage above discoveries to learn from noisy labels robustly. The proposed method contains three stages, i.e., obtaining the representation by self-supervised learning without any labels, transferring the noisy label learning problem into a semi-supervised one by the classifier directly and reliably trained with noisy labels, and joint semi-supervised retraining of both the representation and classifier. Extensive experiments are performed on both synthetic and real benchmark datasets. Results demonstrate that the proposed method can beat the state-of-the-art ones by a large margin, especially under high noise level.

1. Introduction

Convolutional neural networks (ConvNets) [17, 33] have achieved remarkable success in many computer vision tasks, e.g., image classification [26, 53] and object detection [16, 25], because of their ability to model complex patterns. To fully exploit the learning ability of ConvNets, large-scale and well-annotated datasets, e.g., ImageNet [50] and COCO [38], are needed. However, noisy labels are ubiquitous and inevitable, since such large and accurate datasets are expensive and time-consuming to acquire. Besides, modern ConvNets can easily overfit and memorize these noisy labels due to over-parameterization, which subsequently leads to very poor generalization [2, 43, 68]. Thus, how to train ConvNets robustly against noisy labels

Table 1. A comparison between the proposed REED with state-of-the-art methods. Manner: how a method learn from noisy labels; Semi: can semi-supervised learn from samples with wrong labels; HN: can deal with high noise level; NCln: no need for extra clean training data.

Method	Learning manner	Properties		
		Semi	HN	NCln
Co-Teaching [21]	end2end	✗	✗	✓
F-correction [44]	end2end	✗	✗	✓
Ren et al. [49]	end2end	✗	✗	✗
PENCIL [66]	end2end	✗	✗	✓
M-correction [1]	end2end	✗	✗	✓
NLNL [32]	end2end	✓	✗	✓
Zhang et al. [72]	end2end	✗	✓	✗
DivideMix [35]	end2end	✓	✗	✓
REED (ours)	decoupled	✓	✓	✓

becomes a problem of great importance.

Recently, many approaches have been proposed to robustly learn from noisy labels [20], and they generally follow three directions, i.e., sample selection [21, 29, 35, 51, 60, 67], label correction [22, 54, 56, 62, 66], and robustifying loss functions [8, 14, 32, 39, 42, 55, 71]. Specifically, sample selection methods construct some criterion attempting to pick up samples with clean labels for training. These criterion, e.g., small-loss [21, 29] and disagreement [41, 67], usually rely on the memorization effect [2, 68] of deep networks. Label correction methods attempt to directly correct the possibly noisy labels. They achieve this purpose by, e.g., pseudo labeling technologies [34], using class prototype [22], or even treating label as learnable and latent variables [54, 66]. Finally, since the existing loss functions for classification, e.g., categorical cross entropy (CE) [14] and focal loss [37], can be skewed by noisy labels [71], more robust loss functions are proposed. Examples are generalized CE loss [71] and curriculum loss [39]. They are less biased on noisy labels and can be learned together with rep-

resentation.

Indeed, due to the superior performance resulting from the end-to-end training of deep networks [17], most aforementioned methods also jointly learn the representation and classifier in an end-to-end manner. However, such a joint learning scheme neglects an important issue - *is there any difference in the learning behaviors between the representation and classifier with noisy labels?* Intuitively, when the representation is good enough, [1]the decision boundary of the classifier can be easy to find even there is strong noise [23, 58]. Thus, we decouple the training scheme with noisy labels into representation and classifier learning, and then look inside their learning behaviors. Interestingly, we discover that noisy labels will damage the representation learning much more significantly than classifier learning, and the classifier itself indeed can exhibit strong robustness w.r.t. noisy labels with a good representation. These discoveries are not noticed previously and new to the literature of noisy label learning.

Thus, instead of the classical end-to-end training method, which can lead to sub-optimal performance, we are motivated to decouple the representation and classifier in noisy label learning. The proposed method, named REED, contains three stages and can take good care of both representation and classifier by leveraging the above discoveries (see Table 1). Specifically, in the first stage, inspired by the recent advances of self-supervised representation learning technologies [10, 24], we learn the representation through a contrastive pretext task. Then, in the second stage, we utilize the intrinsic robustness in classifier learning to obtain a reliable classifier with noisy labels, which helps to transfer the noisy label learning into a semi-supervised learning problem. Finally, to fully explore the information in the transferred labels, we construct a class-balanced sampler and graph-structured regularizer to jointly fine-tune the representation and classifier in the third stage. Contributions of this paper are as follows:

- We decouple the classical end-to-end training procedures into representation learning and classifier learning and systematically explore their robustness in the presence of noisy labels respectively. We discover that representation matters much more than the classifier, since the representation is very fragile while classifier can exhibit strong robustness in the presence of noisy labels.

- We propose an effective and three-stage learning manner, i.e., REED, which leverages self-supervised representation learning to solve the fragility of representation learning and make full use of the robustness of the classifier. Also by assigning credibility to samples and two improvements for semi-supervised learning (i.e., graph-structured regularization and class-balanced sampler), we

---

[1]We offer a detailed example in Appendix **??**.

further improve the ability of classification.

- We perform extensive experiments on both synthetic, i.e., noisy CIFAR-10 and CIFAR-100 datasets, and real benchmark, i.e., Clothing-1M dataset. Results demonstrate that the proposed method can beat the state-of-the-art ones by a large margin, especially under high noise level. Effectiveness of each stage is also elaborated in ablation studies.

Notations. Let x_i denote the image, y_i be the clean label, and $\bar{y}_i$ be the noisy version of y_i. A ConvNet typically contains two parts, one is the representation learning part, i.e., $z_i = h(x_i; \theta)$, where z_i is the representation for x_i and h is implemented by the feature extractor (i.e., deep stack of convolutional and pooling layers) with parameter θ; another part is the classifier, i.e., $f = g(z_i; \mathbf{W})$, where $\mathbf{W}$ is the parameters of a multilayer perceptron (MLP). Its prediction probability (or confidence) for the ith class is given by a softmax function as $p_i = e^{f_i} / \sum_{k=1}^{C} e^{f_k}$, where C is the number of classes. Cross-entropy loss is used for training, i.e., $\mathcal{L}_{ce}(p, \bar{y}) = -\sum_{k=1}^{C} \bar{y}_k \log p_k$, and the class with the highest confidence, i.e., $y' = \arg\max_i(p)$, is taken as the prediction for the sample.
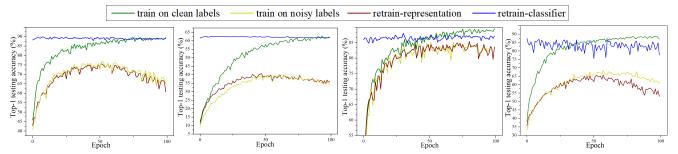
2. Related work

2.1. Noisy Label Learning

Many methods have been developed to make deep classification model robust w.r.t. noisy labels [20], as mentioned in Section 1, they generally follow three directions.

- **Sample selection.** A straightforward approach to handle noisy labels is to find and make them less important in training. Relying on the memorization effect [2, 68], many methods separate clean and noisy samples by using small-loss [1, 21, 29, 35, 51, 64] and disagreement [41, 67]. Among them, MentorNet [29] introduces an auxiliary network to select clean samples; Co-teaching [21] maintains two networks to avoid confirmation bias. And many approaches [1, 35, 64, 67] attempt to improve the sampling criterion or better reuse the noisy labels. As for re-weighting methods [49, 51], optimal weights are learned for different samples by using meta-learning [28]. However, these methods heavily rely on the training strategies that discard or reuse the noisy samples [64].

- **Label correction.** This direction attempts to directly correct possibly noisy labels into correct ones. One common approach to achieve this is leveraging pseudo labeling technologies, which reuses confident predictions from the trained model [32, 34]. Joint optimization [54] updates noisy labels with network parameters in an alternating strategy. Meta re-labeing [72] incorporates con-

(a) CIFAR-10-Sym40%-ResNet18.	(b) CIFAR-100-Sym40%-ResNet18.	(c) CIFAR-10-Asym20%-ResNet18.	(d) CIFAR-10-Sym50%-ResNet50.

Figure 1. Top 1 testing accuracy v.s. epochs for the representation and classifier learning with different datasets (i.e., Figure 1(b)), different noisy type (i.e., Figure 1(c)), and different architectures (i.e., Figure 1(d)). "Sym" and "Asym" denote symmetric and asymmetric noisy type respectively as in Section 5.1. Implementation details are in Appendix **??**.

ventional pseudo labeling into meta optimization. PEN-CIL [66] uses label distribution to replace noisy labels and updated it during training. And class prototype is used in [22] to replace noisy samples. However, these methods either need extra clean data [72], or fail with high noise, since pseudo-label predicted by the networks is not accurate enough [32].

- **Robust loss function.** Since commonly used loss functions for classification, e.g., categorical cross-entropy (CE) and Focal loss (FL), are not robust in the context of noisy labels [14, 40]. Many approaches are proposed to modify or redesign them [18, 40, 42, 45, 69, 71]. Generalized cross-entropy loss (GCE) [71] proposes to mix CE with robust mean absolute error (MAE) loss by Box-Cox transformation. Normalized loss [40] shows that a simple normalization can make any loss function robust to noisy labels. However, since the deep network is over-parameterized, these methods can still memorize the noisy labels given sufficient training time [68].

The proposed method and above existing ones are compared in Table 1. Our insights in Section 3 motivate us to decouple the representation and classifier in noisy label learning, which further lead to a new direction in training a robust ConvNet.

2.2. Self-Supervised Representation Learning

Self-supervised representation learning (SSRL) [6, 15] is a form of unsupervised learning where the data itself provides supervision without external human annotations. In general, it frames a pretext task, e.g., rotations [15] and relative patch position prediction [12], and the network is forced to learn what a task really cares about, e.g., the semantic representation, in order to learn well [73].

Such an idea is widely used in natural language processing (NLP) field [11, 47]. Recently, it also makes great progress in the vision domain, which has shown to be a promising alternative to supervised representation learning [10, 19, 24, 74]. And various heuristics pretext tasks

can be unified into some forms of contrastive losses [24], which measure the similarity of sample pairs in representation space. There are many contrastive learning based methods, e.g., SimCLR [10] and MoCo [24], which have shown a great promise and achieve the state-of-the-art in SSRL.

3. Revisiting Noisy Representation Learning

Intuitively, representation can matter more than classifier, since the decision boundary is obvious when representation is good enough (see Appendix **??**). However, most aforementioned approaches train representation and classifier jointly in an end-to-end manner without care about their difference, which might be suboptimal. In this section, to justify this, we attempt to look deeply into the representation and classifier's learning behavior by decoupling the classical end-to-end training manner.

As shown in Figure 1(a), we first train the whole ConvNet with clean labels (see the curve *train on clean labels*), which can be regarded as the ground-truth model. Then, we follow the control variate method to fairly and separately compare the robustness of the representation and classifier. Specifically, we retrain the representation with noisy labels while fixed the classifier inherited from the ground-truth model (i.e., *retrain-representation*); and retrain the classifier with noisy labels while fixed the representation inherited from the ground-truth model (i.e., *retrain-classifier*). Finally, we include the baseline training the whole network with noisy labels in an end-to-end manner (i.e., *train on noisy labels*), which is adopted by most current noisy label learning approaches.

First, we can see that the curve of *train on clean labels* rise steadily, while the curve of *train on noisy labels* first increases and then slowly drops. This is known as the memorization effect in the literature [2], which is also the crux of sample selection methods [21, 29, 35, 67, 65]. More importantly, we have the following three new observations:

- The curve of *retrain-representation* is very similar to that

of *train on noisy labels*, which demonstrates that representation learning is very sensitive to noisy labels;

- It is surprised to see that *retrain-classifier* can easily get accuracy near to the ground-truth model in a few epochs and then keep steadily without overfitting. This shows the classifier itself can be intrinsically robustness w.r.t noisy labels with a good representation;

- Finally, we can see that above phenomenons do not change w.r.t. different dataset (i.e., Figure 1(b)), different noisy type (i.e., Figure 1(c)), and different architecture (i.e., Figure 1(d)). This further supports our claim that noisy labels will damage the representation learning significantly than the classifier learning.

These observations subsequently motivate us to propose a new method which can take good care of both representation learning and classifier for noisy label learning.

4. The Proposed Method

Next, we elaborate our approach, i.e., REED, which contains three stages to fully utilizes our observations. Specifically, in the first stage (Section 4.1), we show how to solve the fragility of representation learning by leveraging recent SSRL methods. Next, in the second stage (Section 4.2), we utilize the intrinsic robustness in classifier learning to obtain a reliable classifier with noisy labels. Then, we measure the label credibility for each sample using the obtained classifier, which helps transfer the noisy label learning into a semi-supervised learning problem in the last stage (Section 4.3). An overview of the proposed method is in Figure 2.
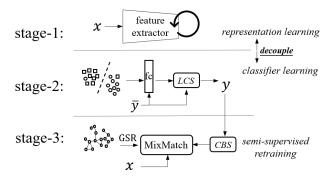

Figure 2. The framework of REED. "LCS", "GSR" and "CBS" are short for label credibility assessment, graph-structured regularization and class-balanced sampler respectively.

4.1. Stage-1: Representation Learning

Previously, both the representation h and classifier g are trained in an end-to-end manner in noisy label learning (see Table 1). However, as observed in Section 3, such a learning manner for h may not be proper. Thus, can we learn the representation without using noisy labels? Fortunately, recent advances in SSRL have shown promising results in this direction [24, 10, 19]. As a result, we adopt such a learning paradigm in Stage-1 to learn the representation without using noisy labels. See more implementation details in Appendix **??**.

4.2. Stage-2: Robust Classifier Training

After learning the representation without labels, inspired by observations in Section 3, we fix the representation and train a robust classifier g on noisy labels. Moreover, based on the reliable classifier, we can accurately assign the credibility for each sample's label, which helps transfer the hard noisy label learning problem to an easier semi-supervised learning one.
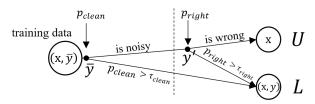

Figure 3. Label credibility assessment module: transferring noisy labels into clean or unknown labels.

To fully leverage $\bar{y}$ in the training data, we transfer it into clean labeled data (L) and unknown labeled data (U) (Figure 3). Firstly, the original label $\bar{y}$ is divided into clean or noisy. Next, for a noisy label, we correct it if classifier g predicts right y'. To achieve this, we assign credibility p_{clean} and p_{right} to corresponding labels, i.e., $\bar{y}$ and y', and divide clean or noisy by a threshold τ_{clean}, judge right or wrong y' by another threshold τ_{right}. More Specifically, due to memorization effect [2], loss can be regarded as a good criterion for dividing clean and noisy [21, 35]. We estimate p_{clean} by a two-component Gaussian mixture models (GMM) [46], which is fitted on $\mathcal{L}_{\text{ce}}$ distribution following DivideMix [35]. Besides this, since samples with highly confident predictions tend to be correctly classified [57], our paper equips another GMM to estimate p_{right}, which is optimized on the distribution of p_i.

4.3. Stage-3: Semi-supervised Retraining

Since the representation and classifier can still be biased by decoupling in the previous two stages (see Table 2), in this final stage, we re-train the whole network on the transferred labels by semi-supervised learning. However, comparing with standard semi-supervised learning methods [4, 52, 63], we propose two unique designs here, i.e., graph-structured regularization to fully make use of self-learned representation in Stage-1 and class-balanced sampler to solve the class-imbalanced problem caused by relabeling in Stage-2.

4.3.1 Graph-structured regularization

As representation self-learned in Stage-1 narrows the distance of similar samples, and samples with stronger similarity are more likely to share the same label. Inspired by neural graph machines [5], to make use of this information, we treat image samples in a mini-batch as nodes in a graph $\mathcal{G}$, and the edges in $\mathcal{G}$ correspond to the similarity between pairs of samples. Such a neighbor graph-structured can be constructed by Algorithm 1, where sim denotes cosine similarity and τ_c is a threshold for determining the connection of nodes. We used it with existing semi-supervised learning methods. Specifically, we introduce the following regularization:

$$\mathcal{R} = \lambda_{\text{LU}} \sum_{u \in U, v \in L} A_{uv} \|\hat{p}(x_u) - y_v\|_2^2 + \\ \lambda_{\text{UU}} \sum_{u,v \in U} A_{uv} \|\hat{p}(x_u) - \hat{p}(x_v)\|_2^2,$$

where $\hat{p} = p_i^{1/T} / \sum_{k=1}^{C} p_k^{1/T}$ is a sharpening operation that reduce the entropy of predicted label distribution, T is the "temperature", $\lambda_{\text{LU}}, \lambda_{\text{UU}} \geq 0$ are hyperparameters. Formally, the final objective $\mathcal{L}$ for Stage-3 is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_u \mathcal{L}_{\text{unsup}} + \mathcal{R},$$

where $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{unsup}}$ denote supervised and unsupervised loss term respectively. Note that many existing semi-supervised learning methods can be used, e.g., Mix-Match [4], FixMatch [52] and ReMixMatch [3]. However, for a fair comparison with existing works in noisy label learning, we use MixMatch [4] here following DivideMix [35].

Algorithm 1 Construct neighbor graph structure.

Requires: Representation matrix $Z^{N \times d}$, threshold τ_c.
Ensure: Adjacency Matrix $A^{N \times N}$ of neighbor graph $\mathcal{G}$.

 1: **for** i from 0 to $N-1$ and j from 0 to $N-1$ **do**
 2:     $A_{ij} \leftarrow ReLU(sim(Z_i, Z_j) - \tau_c)$
 3: **end for**

4.3.2 Class-balanced sampler

Recall that we have transferred the noisy dataset into L and U in Figure 3, such a division could cause class-imbalanced issue (see Figure 5(c)) since the fitting ability for different categories is not the same. However, existing semi-supervised learning methods [3, 4, 52], typically require that given L and U been class-balanced, and they do not generalize well in the imbalanced scenarios, since those methods make use of the unlabeled data by pseudo-label. Thus, they are very sensitive to label quality and can be easily biased toward majority classes [31]. To alleviate this issue, we formulate a class-balanced sampler for L, which

uses sampling probability $P_i = 1/C$ for the ith class. Finally, to make the U balanced, we take all x in the original dataset as candidates for sampling unlabeled data.

4.4. Comparison with Existing Works

Pre-training [13] is a popular approach to obtain a general representation in many visual tasks [25, 74], which can improve robustness w.r.t. noise labels by providing a good initialization [27]. However, the representation obtained from ImageNet is not enough to train a accurate classifier to divide noisy labels (see Section 5.3.2), due to the domain gap [30, 59, 61]. Additionally, collecting well-annotated datasets like ImageNet is very challenging and not easy to scale up, but SSRL can help leverage large-scale noisy labeled target dataset, e.g., Clothing-1M.

DivideMix [35] is the state-of-the-art and the most related work with our method, which also leverages semi-supervised techniques in noisy label learning task. However, there are three important differences. First, our method makes full use of the different robustness in representation learning and classifier learning, which is not concerned in DivideMix and other literature. Second, DivideMix needs warmup on the noisy dataset firstly, which can easily lead to over-fitting especially on the extreme noise, while our method does not have such an issue. Finally, we introduce label correction by label credibility assessment module (Section 4.2) and proposed effective improvements to semi-supervised learning (Section 4.3).

5. Experiments

All codes are implemented in the PyTorch framework run on a server with 8 NVIDIA RTX 2080 GPUs.

5.1. Datasets

Noisy CIFAR-10 and CIFAR-100. Both CIFAR-10 and CIFAR-100 consist of 60000 color images of size 32×32, 50000 for training, and 10000 for testing. Following [21, 35, 67], we apply two types of label noise, i.e., *symmetric* and *asymmetric* (see elaborations in Appendix **??**), on the original CIFAR dataset. Symmetric noise is generated by replacing the clean labels with uniformly selected labels among the classes; asymmetric noise is designed to simulate the fine-grained classification task, where clean labels are replaced by similar classes.

Clothing-1M [62]. This is a large dataset containing noisy labels from real scenarios. It contains 1 million training cloth images from the shopping website, which are categorized into 14 classes by parsing surrounding texts. And it also provides 50k, 14k, 10k clean data for training, validation, and testing respectively. Following previous methods [35, 36, 54], we do not use the 50k clean training data. The shorter size of the image is first resized to 256

Table 2. Comparison with state-of-the-art methods on CIFAR-10/100 datasets with symmetric noise. For fair comparison under the same network architecture, i.e., Pre-ResNet18, we reuse the re-implemented baseline results from [35]. † means without model ensemble. "CE" is the standard ConvNet trained with Cross-Entropy loss in an end-to-end manner.

| Dataset | | CIFAR-10 | | | | CIFAR-100 | | | |
Method/Noise ratio		20%	50%	80%	90%	20%	50%	80%	90%
Cross-Entropy (CE)	Best	86.8	79.4	62.9	42.7	62.0	46.7	19.9	10.1
	Last	82.7	57.9	26.1	16.8	61.8	37.3	8.8	3.5
Bootstrap [48]	Best	86.8	79.8	63.3	42.9	62.1	46.6	19.9	10.2
	Last	82.9	58.4	26.8	17.0	62.0	37.9	8.9	3.8
F-correction [44]	Best	86.8	79.8	63.3	42.9	61.5	46.6	19.9	10.2
	Last	83.1	59.4	26.2	18.8	61.4	37.3	9.0	3.4
Co-teaching+ [67]	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 [69]	Best	95.6	87.1	71.6	52.2	67.8	57.3	30.8	14.6
	Last	92.3	77.6	46.7	43.9	66.0	46.6	17.6	8.1
PENCIL [66]	Best	92.4	89.1	77.5	58.9	69.4	57.5	31.1	15.3
	Last	92.0	88.7	76.5	58.2	68.1	56.4	20.7	8.8
Meta-Learning [36]	Best	92.9	89.3	77.4	58.7	68.5	59.2	42.4	19.5
	Last	92.0	88.8	76.1	58.3	67.7	58.0	40.1	14.3
M-correction [1]	Best	94.0	92.0	86.8	69.1	73.9	66.1	48.2	24.3
	Last	93.8	91.9	86.6	68.7	73.4	65.4	47.6	20.5
DivideMix† [35]	Best	<u>95.2</u>	<u>94.2</u>	<u>93.0</u>	75.5	<u>75.2</u>	<u>72.8</u>	58.3	29.9
	Last	<u>95.0</u>	<u>93.7</u>	<u>92.4</u>	74.2	<u>74.8</u>	<u>72.1</u>	57.6	29.2
REED (no Stage-3)	Best	91.8	91.7	90.8	<u>89.1</u>	65.2	64.1	<u>60.2</u>	<u>54.8</u>
	Last	91.8	91.6	90.4	<u>88.4</u>	65.0	63.9	<u>59.9</u>	<u>54.1</u>
REED	Best	**95.8**	**95.6**	**94.3**	**93.6**	**76.7**	**73.0**	**66.9**	**59.6**
	Last	**95.7**	**95.4**	**94.1**	**93.5**	**76.5**	**72.2**	**66.5**	**59.4**

pixels and then takes a 224×224 crop from it for fitting the network.

5.2. Implementation Details

Noisy CIFAR-10 and CIFAR-100. In Stage-1 of representation learning, we adopt the SimCLR [10] implementation for the self-supervised representation learning and ResNet50 as encoder with most hyperparameters as introduced by the original paper. In the Stage-2 of robust classifier training, we fix the representation and only train classifier with SGD optimizer for 200 epoch on CIFAR-10 and 100 epoch on CIFAR-100 dataset, and the learning rate is set as 0.001 and 0.005 respectively. And the hyperparameters of two GMM are same as [35], τ_{clean} and τ_{right} are both set to 0.5. For Stage-3 retraining, following DivideMix [35], we adopt the PreAct-ResNet18 architecture. As for MixMatch [4], we choose T from $\{0.5, 0.8\}$, α from $\{4, 0.75\}$, the unsupervised loss weight λ_u from $\{0, 20, 50, 100, 150, 200\}$ and learning rate from $\{0.001, 0.005\}$ and the whole parameters are trained with

Adam optimizer and cosine learning rate scheduler for 500 epoch with batch size of 128 and eta_min of 0.0002, and parameter exponential moving average (EMA) is used as in [4]. For neighbor graph, λ_{LU}, λ_{UU}, and τ_c are set to 0.01, 0.005 and 0.5 respectively.

Clothing-1M. In Stage-1, we adopt the implementation from MoCo for representation learning to save memory, and ResNet50 as encoder with most hyperparameters as introduced by the original paper. In State-2, we train the classifier with an SGD optimizer for 5 epochs and decay learning rate at the 3rd epoch by 0.1, the learning rate and batch size are set to 5 and 512 respectively. In Stage-3, we adopt the ResNet50 architectures without ImageNet pre-trained weight. For saving training time, we reuse the weights trained in the previous two-stages, EMA is used as in [4], and fine-tune the weights by SGD optimizer with learning rate of 0.005 for classifier and 0.0005 for feature extractor.

6

5.3. Benchmark Comparison

5.3.1 Noisy CIFAR-10 and CIFAR-100 Datasets

Symmetric case. Following [35], we conduct experiments on noisy CIFAR datasets with four different noisy ratios, i.e., 20%, 50%, 80%, 90%. The best testing accuracy across all epochs and the average testing accuracy over the last 10 epochs are reported. Table 2 shows the top-1 testing accuracy on CIFAR-10 and CIFAR-100 with symmetric noisy type. As can be seen, compared with other competing methods, our method obtains the best performance. Specifically, on extreme noisy ratio (e.g., 90%), we get 26% relative improvement in CIFAR-10, and 103% relative improvement in CIFAR-100 compared with the state-of-the-art method respectively.

Asymmetric case. Following [35], we test our method on the asymmetric noisy CIFAR-10 dataset with noisy ratio of 40% and provide extra 20% noisy ratio results for better comparison. We keep the same noise generalization process as previous works [35, 36], and the top-1 testing accuracy is shown in the Table 3. Compared with DivideMix [35], REED gets 1.4% and 1.2% improvements in the noisy ratio of 20% and 40% respectively, and outperforms many methods that use deeper networks, e.g., Pre-ResNet32. Besides, our method do not leverages the extra clean data, while it is needed in [72] for meta-relabeling.

Table 3. Asymmetric noise on CIFAR-10. † means the reproduced result from the public code without model ensemble.

Method	Backbone	Noisy ratio	
		20%	40%
Joint-Optim [54]	Pre-ResNet32	92.8	91.7
PENCIL [66]	Pre-ResNet32	92.4	91.2
F-correction [44]	ResNet32	89.9	-
Zhang, et al. [72]	ResNet29	92.7	90.2
Meta-Learning [36]	Pre-ResNet18	-	88.6
M-correction [1]	Pre-ResNet18	-	86.3
Iterative-CV [9]	Pre-ResNet18	-	88.0
DivideMix† [35]	Pre-ResNet18	93.6	91.1
REED	Pre-ResNet18	**95.0**	**92.3**

Learning curves comparison. Finally, to further demonstrate the benefit of decoupling the representation and classifier learning in REED, we plot its learning curves in Stage-3 in Figure 4. Specifically, the best existing method in Table 2, i.e., DivdeMix, and the baseline CE which directly trains the whole model on noisy labels, are compared. As can be seen, the curves of REED grow much steadily compared with DivideMix and without the memorization effect compared with CE.
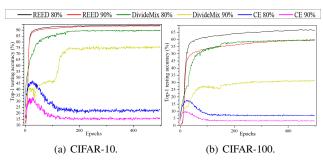

(a) CIFAR-10.   (b) CIFAR-100.

Figure 4. Learning curves on CIFAR-10/100 datasets with 80% and 90% symmetric noise. "CE" is short for CrossEntropy.

5.3.2 Clothing-1M Dataset

Performance comparison. Table 4 shows the top-1 accuracy on the clean testing set of Clothing-1M. It can be seen that our method performs the best. Note that, competing baseline approaches, i.e., F-correction, M-correction, Joint-Optim, Meta-Learning, PENCIL, DivideMix and Self-Learning, are pre-trained on ImageNet, while our method does not leverage information from other data. Besides, we find that DivideMix is very sensitive to ImageNet pre-training. When we train DivideMix from scratch, its accuracy drops 5.03%. This phenomenon further confirms our claim that representation is very important and fragile in the noisy label learning.

Table 4. Top-1 testing accuracy on the Clothing-1M. ‡ means without ImageNet pre-training.

Method	Top-1 testing accuracy
CrossEntropy	69.21
F-correction [44]	69.84
M-correction [1]	71.00
Joint-Optim [54]	72.16
Meta-Cleaner [70]	72.50
Meta-Learning [36]	73.47
PENCIL [66]	73.49
Self-Learning [22]	74.45
DivideMix [35]	74.76
DivideMix‡ [35]	69.73
REED	**75.81**

Distinctiveness of the representation. To understand the benefit of SSRL over pre-training better, we visualize the obtained representations by each method on the clean Clothing-1M testing set by t-SNE [7]. Five exemplar classes (i.e., Downcoat, Suit, Shawl, Dress, Underwear) are plotted in Figure 6, As can be seen, pre-training can provide coarse-grained clustering representation by transferable common pattern filters, but not as good as self-supervised, which direct learns representation on the Clothing-1M dataset. Finally, there is a clear gap between the representation of different classes in Figure 6 (a). As

(a) Histogram of $\mathcal{L}_{ce}$ on CIFAR-10. (b) Histogram of p_i on CIFAR-10. (c) Imbalanced class distribution on CIFAR-100.
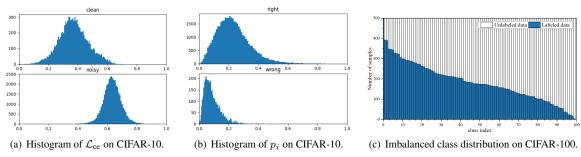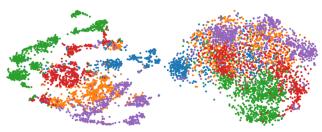
Figure 5. Histograms show the label imbalanced problem under 90% symmetric noise.

discussed in Section 1, such a distinctiveness on the representation further helps us find a robust classifier even with noisy labels.

(a) Self-supervised representation. (b) Pre-trained representation.

Figure 6. t-SNE visualization of representations for Clothing-1M obtained by SSRL (left) and ImageNet pre-training (right).

5.4. Ablation studies

5.4.1 Stage-1: Decoupled v.s. End-to-end training

For a better understanding the advantages of the decoupled learning strategy in Stage-1, Figure 7 plots a comparison of two different learning manners, i.e., Decoupled and End-to-end. Specifically, the curves of CE are trained end-to-end on noisy datasets. For the curves of Decoupled, we first train representation through SSRL, then train the classifier on noisy datasets. As can be seen, the curves of Decoupled grow much steadily and without the memorization effect comapred with CE. Besides, the curves of Decoupled are less sensitive to noisy ratio compared with CE, which further illustrates the robustness of the classifier to noise.
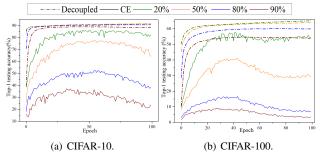

(a) CIFAR-10. (b) CIFAR-100.

Figure 7. Comparison of two different learning manners on the symmetric noisy CIFAR datasets. The line patterns (dash-dotted line and solid line) represent "Decoupled" and "CE" respectively.

5.4.2 Stage-2: Label transformation

Figure 5(a) shows that the optimized reliable classifier g can separate the clean and noisy labels with a large margin by using the loss distribution. Besides, the right predicted pseudo-labels can be judged by the confidence distribution (see Figure 5(b)). As analyzed in Section 4.3, the divided labels are class-imbalanced (see Figure 5(c)).

5.4.3 Stage-3: Improved semi-supervised learning

We propose two improvements, i.e., class-balanced sampler (denoted as "CBS") and graph-structured regularization (denoted as "GSR"), for the semi-supervised learning on the transferred labels. As shown in Table 5, they both help improve the performance of noisy label learning.

Table 5. Ablation study for semi-supervised retraining in Stage-3 on CIFAR-100.

Component			Noisy ratio			
CBS	GSR		20%	50%	80%	90%
✗	✗	Best	74.2	71.1	65.5	58.6
		Last	74.0	70.7	65.2	58.4
✗	✓	Best	74.8	71.9	65.6	59.0
		Last	74.7	71.5	65.2	58.4
✓	✗	Best	76.2	72.3	66.4	59.4
		Last	75.9	72.0	66.3	59.2
✓	✓	Best	**76.7**	**73.0**	**66.9**	**59.6**
		Last	**76.5**	**72.2**	**66.5**	**59.4**

6. Conclusion

In this paper, we look deeply into the robustness of representation and classifier learning in the presence of noisy labels. We find that noisy labels will damage the representation learning significantly than classifier learning, and the classifier itself can exhibit strong robustness w.r.t. noisy labels with a good representation. Motivated by this, we proposed an effective and three-stage learning manner to take care of both representation learning and classifier learning. Experiments demonstrate that

our method achieves state-of-the-art results on noisy label learning benchmarks.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In International Conference on Machine Learning, 2019.

[2] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In International Conference on Machine Learning, 2017.

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In International Conference on Learning Representations, 2020.

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In Advances in Neural Information Processing Systems, 2019.

[5] Thang D Bui, Sujith Ravi, and Vivek Ramavajjala. Neural graph machines: Learning neural networks using graphs. arXiv preprint arXiv:1703.04818, 2017.

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In European Conference on Computer Vision, pages 132–149, 2018.

[7] David M Chan, Roshan Rao, Forrest Huang, and John F Canny. Gpu accelerated t-distributed stochastic neighbor embedding. J. Parallel Distributed Comput., 2019.

[8] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In International Conference on Machine Learning, 2019.

[9] Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In International Conference on Machine Learning, 2019.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning, 2020.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In International Conference on Computer Vision, 2015.

[13] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In International Conference on Artificial Intelligence and Statistics, 2010.

[14] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In AAAI Conference on Artificial Intelligence, 2017.

[15] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In International Conference on Learning Representations, 2018.

[16] Ross B. Girshick. Fast r-cnn. In International Conference on Computer Vision, 2015.

[17] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning. 2016.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733, 2020.

[20] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. Technical report, arXiv preprint arXiv:2011.04406, 2020.

[21] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Coteaching: Robust training of deep neural networks with extremely noisy labels. In Advances in Neural Information Processing Systems, 2018.

[22] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In International Conference on Computer Vision, 2019.

[23] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In International Conference on Computer Vision, 2017.

[26] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[27] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In International Conference on Machine Learning, 2019.

[28] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. arXiv preprint arXiv:2004.05439, 2020.

[29] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In International Conference on Machine Learning, 2018.

[30] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[31] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In Advances in Neural Information Processing Systems, 2020.

[32] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In International Conference on Computer Vision, 2019.

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. 1998.

[34] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, International Conference on Machine Learning, 2013.

[35] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In International Conference on Learning Representations, 2020.

[36] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Learning to learn from noisy labeled data. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In International Conference on Computer Vision, 2017.

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In European Conference on Computer Vision, 2014.

[39] Yueming Lyu and Ivor W Tsang. Curriculum loss: Robust learning and generalization against label corruption. In International Conference on Learning Representations, 2020.

[40] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In International Conference on Machine Learning, 2020.

[41] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". In Advances in Neural Information Processing Systems, 2017.

[42] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In International Conference on Learning Representations, 2019.

[43] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In Advances in Neural Information Processing Systems, 2017.

[44] 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 IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[45] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In International Conference on Learning Representations, 2017.

[46] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. Pattern Recognition, 2006.

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 2019.

[48] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In International Conference on Learning Representations, 2015.

[49] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In International Conference on Machine Learning, 2018.

[50] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015.

[51] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In Advances in Neural Information Processing Systems, 2019.

[52] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685, 2020.

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[54] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[55] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. In International Conference on Machine Learning, 2019.

[56] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In Advances in Neural Information Processing Systems, 2017.

[57] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In European Conference on Computer Vision, 2020.

[58] Vladimir N Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.

[59] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. Neurocomputing, 2018.

[60] Yixin Wang, Alp Kucukelbir, and David M Blei. Robust probabilistic modeling with bayesian data reweighting. In International Conference on Machine Learning, 2017.

[61] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology, 2020.

[62] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[63] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In Advances in Neural Information Processing Systems, 2020.

[64] Hansi Yang, Quanming Yao, Bo Han, and Gang Niu. Searching to exploit memorization effect in learning from corrupted labels. In International Conference on Machine Learning, 2020.

[65] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Kwok. Searching to exploit memorization effect in learning from corrupted labels. In International Conference on Machine Learning, 2020.

[66] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[67] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In International Conference on Machine Learning, 2019.

[68] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In International Conference on Learning Representations, 2017.

[69] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In International Conference on Learning Representations, 2018.

[70] Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

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

[72] Z. Zhang, H. Zhang, S. Arik, H. Lee, and T. Pfister. Distilling effective supervision from severe label noise. In IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[73] Andrew Zisserman. zisserman-self-supervised. https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf, 2018.

[74] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. arXiv preprint arXiv:2006.06882, 2020.