On Learning Contrastive Representations for Learning with Noisy Labels

Li Yi[1] Sheng Liu[2] Qi She[3] A. Ian McLeod[1] Boyu Wang*[1,4]
[1]University of Western Ontario, [2]NYU Center for Data Science
[3]ByteDance Inc. [4]Vector Institute
lyi7@uwo.ca shengliu@nyu.edu sheqi1991@gmail.com
aimcleod@uwo.ca bwang@csd.uwo.ca

Abstract

Deep neural networks are able to memorize noisy labels easily with a softmax cross entropy (CE) loss. Previous studies attempted to address this issue focus on incorporating a noise-robust loss function to the CE loss. However, the memorization issue is alleviated but still remains due to the non-robust CE loss. To address this issue, we focus on learning robust contrastive representations of data on which the classifier is hard to memorize the label noise under the CE loss. We propose a novel contrastive regularization function to learn such representations over noisy data where label noise does not dominate the representation learning. By theoretically investigating the representations induced by the proposed regularization function, we reveal that the learned representations keep information related to true labels and discard information related to corrupted labels. Moreover, our theoretical results also indicate that the learned representations are robust to the label noise. The effectiveness of this method is demonstrated with experiments on benchmark datasets.

1. Introduction

The successes of deep neural networks [19, 34] largely rely on availability of correctly labeled large-scale datasets that are prohibitively expensive and time-consuming to collect [46]. Approaches to addressing this issue includes: acquiring labels from crowdsourcing-like platforms or non-expert labelers or other unreliable sources [49,55] but while theses methods can reduce the labeling cost, label noise is inevitable. Due to the over-parameterization of deep networks [19], examples with noisy labels can ultimately be memorized with a cross entropy loss [3, 27, 32], which is known as the *memorization effect* [30, 53], leading to poor performance [53]. Therefore, it is important to develop methods that are robust to the label noise.

*Corresponding author

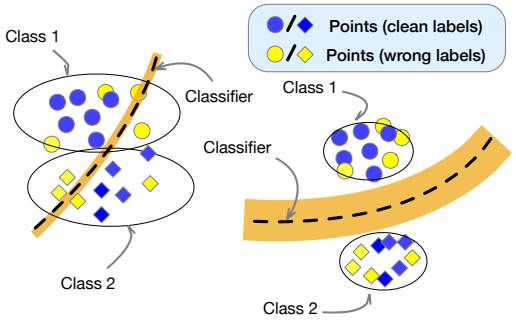

Figure 1. Illustration of the proposed method with noisy labels. Black curves are the best classifiers that are learned during training. **Left:** Deep networks without contrastive regularization. **Right:** Deep networks with contrastive regularization. Two classes are better separated by deep networks that points with the same class are pulled into a tight cluster and clusters are pushed away from each other.

Cross entropy (CE) loss is widely used as a loss function for image classification tasks due to its strong performance on clean training data [37] but it is not robust to label noise. When labels in training data are corrupted, the performance drops [4, 5]. Given the memorization effect of deep networks, training on noisy data with the CE loss results in the representations of the data clustered in terms of their noisy labels instead of the ground truth. Thus, the final layer of the deep networks cannot find a good decision boundary from these noisy representations.

To overcome the memorization effect, noise-robust loss functions have been actively studied in the literature [12,31, 42, 55]. They aim to design noise-robust loss functions in a way such that they achieve small loss on clean data and large loss on wrongly labeled data. However, it has been empirically shown that being robust alone is not sufficient for a good performance as it also suffers from the *underfitting* problem [29]. To address this issue, these noise-robust loss functions have to be explicitly or implicitly jointly used

with the CE loss, which brings a trade-off between non-robust loss and robust loss. As a result, the memorization effect is alleviated but still remains due to the non-robust CE loss.

In this paper, we tackle this problem from a different perspective. Specifically, we investigate contrastive learning and the effect of the clustering structure for learning with noisy labels. Owing to the power of contrastive representation learning methods [7–9, 16, 20], learning contrastive representations has been extensively applied on various tasks [21, 28, 48]. The key component of contrastive learning is positive contrastive pair (x_1, x_2). Training a contrastive objective encourages the representations of x_1, x_2 to be closer. In supervised classification tasks, correct positive contrastive pairs are formed by examples from the same class. When label noise exists, defining contrastive pairs in terms of their noisy labels results in adverse effects. Encouraging representations from different classes to be closer makes it even more difficult to separate images of different classes. Similar to our attempt to learn contrastive representations from noisy data, previous work has focused on reducing the adverse effects by re-defining contrastive pairs according to their pseudo labels [10, 14, 24, 25]. However, pseudo labels can be unreliable, and then wrong contrastive pairs are inevitable and can dominate the representation learning.

To address this issue, we propose a new contrastive regularization function that does not suffer from the adverse effects. We theoretically investigate benefits of representations induced by the proposed contrastive regularization function from two aspects. First, the representations of images keep information related to true labels and discard information related to corrupted labels. Second, we theoretically show that the classifier is hard to memorize corrupted labels given the learned representations, which demonstrates that our representations are robust to label noise. Intuitively, learning such contrastive representations of data helps combat the label noise. If data points are clustered tightly in terms of their true labels, then it makes the classifier hard to draw a decision boundary to separate the data in terms of their corrupted labels. We illustrate this intuition in Figure 1. Our main contributions are as follows.

- We theoretically analyze the representations induced by the contrastive regularization function, showing that the representations keep information related to true labels and discard information related to corrupted labels. Moreover, we formally show that representations with insufficient corrupted label-related information are robust to label noise.
- We propose a novel algorithm over data with noisy labels to learn contrastive representations, and provide gradient analysis to show that correct contrastive pairs can dominate the representation learning.
- We empirically show that our method can be applied with existing label correction techniques and noise-robust loss functions to further boost the performance. We conduct extensive experiments to demonstrate the efficacy of our method.

2. Theoretical Analysis

In this section, we first introduce some notations and we then investigate the benefits of representations learned by the contrastive regularization function.

2.1. Preliminaries

We use uppercases $X, Y, \dots$ to represent random variables, calligraphic letters $\mathcal{X}, \mathcal{Y}, \dots$ to represent sample spaces, and lowercases $x, y, \dots$ to represent their realizations. Let X be input random variable and Y be its true label. We use $\tilde{Y}$ to denote the wrongly-labeled random variable that is not equal to Y. The entropy of the random variable Y is denoted by $H(Y)$ and the mutual information of X and Y is $I(X, Y)$.

Contrastive learning aims to learn representations of data that only the data from the same class have similar representations. In this paper, we propose to learn the representations by introducing the following contrastive regularization function over all examples $\{(x_i, y_i)\}$ from $\mathcal{X} \times \mathcal{Y}$ and y_i is the ground truth.

$$\mathcal{L}_{\text{ctr}}(x_i, x_j) = -\big(\langle \tilde{q}_i, \tilde{z}_j \rangle + \langle \tilde{q}_j, \tilde{z}_i \rangle\big) \mathbb{1}\{y_i = y_j\}, \quad (1)$$

where $\tilde{q}_k = \frac{q_k}{\|q_k\|_2}$ and $\tilde{z}_k = \frac{z_k}{\|z_k\|_2}$. Following SimSiam [9], we define $q = h(f(x))$, $z = \texttt{stopgrad}(f(x))$, f is an encoder network consisting of a backbone network and a projection MLP, and h is a prediction MLP. Minimizing Eq. (1) on $\{(x_i, y_i), (x_j, y_j)\}$ pulls representations of x_i and x_j closer if $y_i = y_j$. The designs of the stop-gradient operation and h applied on representations are mainly to avoid trivial constant solutions.

2.2. The Benefits of Representations Induced by Contrastive Regularization

We first relate the solutions that minimize Eq. (1) to a mutual information $I(Z; X^+) = \iint p(z, x^+) \log \frac{p(z|x^+)}{p(z)} \, \mathrm{d}x^+ \, \mathrm{d}z$, where $z = f(x)$ and x^+ is from the same class as x.

Theorem 1. *Representations Z learned by minimizing Eq. (1) maximizes the mutual information $I(Z; X^+)$.*

Theorem 1 reveals the equivalence between the contrastive learning and mutual information maximization. Intuitively, Eq. (1) encourages to pull representations from the same class together and push those from different classes

apart. The estimate of z conditioned on x^+ is more accurate than random guessing because the representation z of x is similar to the representation of x^+. Thus the pointwise mutual information $\log \frac{p(z|x^+)}{p(z)}$ increases by minimizing Eq. (1).

We denote $Z^\star = \arg\max_{Z_\theta} I(Z_\theta, X^+)$ by the representation that maximizes the mutual information, where Z_θ is a representation of X parameterized by the neural network f with parameters θ. To understand what $Z^\star$ is learned from inputs and to show that $Z^\star$ is noise-robust, we introduce the notion of (ϵ, γ)-*distribution*:

Definition 1 ((ϵ, γ)-distribution)**.** *A distribution $D(X, Y, \tilde{Y})$ is called (ϵ, γ)-Distribution if there exists $\gamma \gg \epsilon > 0$ such that*

$$I(X; Y|X^+) \leq \epsilon, \tag{2}$$

and

$$I(X; \tilde{Y}|X^+) > \gamma. \tag{3}$$

Eq. (2) characterizes the connection between images and their true labels. If we already know an image X^+, then there is the limited extra information related to the true label by additionally knowing X. We use a small number ϵ to restrict this additional information gain. Eq. (3) characterizes the connection between those images and their corrupted labels. By knowing an additional image X^+, the information X contains about its corrupted label $\tilde{Y}$ is still larger than γ. The above condition $\gamma \gg \epsilon > 0$ states that images from the same class are much more similar with respect to the true label than the corrupted label. As it is mentioned in [38], if there is a perfect prediction of Y given X^+, then $\epsilon = 0$.

We illustrate the intuitions behind Definition 1 in Figure 2. We use the Grad-CAM [35] to highlight the important regions in the images for predictions. The highlighted regions captured by the model are most related to labels. For images with the same clean labels, their information related to true labels are similar. For example, when Cat 1 and Cat 2 in Figure 2 are labeled as "cat", cat faces are captured as the true label-related information and they all look alike. For images with corrupted labels, their information related to corrupted labels are quite different. When Cat 1 and Cat 2 in Figure 2 are labeled as "dog", the windows bars captured as the corrupted label-related information for Cat 1 is different from the floor and wall for Cat 2.

With the notion of (ϵ, γ)-distribution, the following theorem help us understand the benefits of representations $Z^\star$ in depth.

Theorem 2. *Given a distribution $D(X, Y, \tilde{Y})$ that is (ϵ, γ)-Distribution, we have*

$$I(X; Y) - \epsilon \leq I(Z^\star; Y) \leq I(X; Y), \tag{4}$$

$$I(Z^\star; \tilde{Y}) \leq I(X; \tilde{Y}) - \gamma + \epsilon. \tag{5}$$

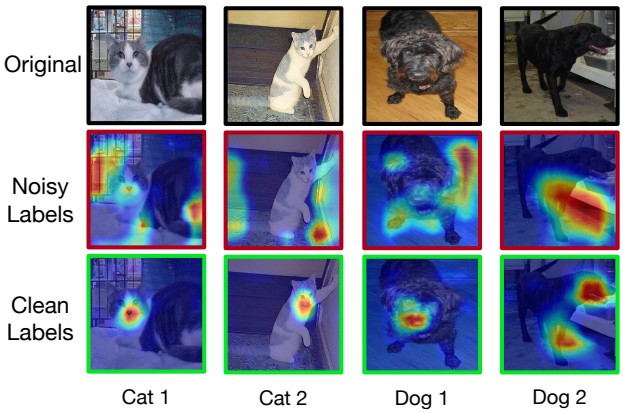

Figure 2. An example of Grad-CAM [35] results of Resnet34 trained on noisy dataset with 40% symmetric label noise and clean dataset, separately. When there is label noise, information related to corrupted labels captured by the model varies from image to image (e.g. window bars in Cat 1 v.s. floor and wall in Cat 2). When there is no label noise, information related to true labels are similar for images from the same class (e.g. cat face in Cat 1 v.s. cat face in Cat 2).

Given images X and their labels Y, the mutual information $I(X; Y)$ is fixed. The theorem states that the learned representations $Z^\star$ keep as much true label-related information as possible and discard much corrupted label-related information. Since the corrupted label-related information is discarded from the representations $Z^\star$, memorizing the corrupted labels based on $Z^\star$ is diminished. Lemma 1 establishes the lower bound on the expected error on *wrongly-labeled* data.

Lemma 1. *Consider a pair of random variables $(X, \tilde{Y})$. Let $\hat{Y}$ be outputs of any classifier based on inputs Z_θ, and $\tilde{e} = \mathbb{1}\{\hat{Y} \neq \tilde{Y}\}$, where $\mathbb{1}\{A\}$ be the indicator function of event A. Then, we have*

$$\mathbb{E}[\tilde{e}] \geq \frac{H(\tilde{Y}) - I(Z_\theta; \tilde{Y}) - H(\tilde{e})}{\log\Big(|\tilde{\mathcal{Y}}|\Big) - 1}.$$

Lemma 1 provides a necessary condition on the success of learning with noisy labels based on representation learning and sheds new light on this problem by highlighting the role of minimizing $I(Z_\theta; \tilde{Y})$. To see this, note that small $I(Z_\theta; \tilde{Y})$ implies robustness to label noise since $\mathbb{E}[\tilde{e}]$ is the expected error over the corrupted labels. On the other hand, when minimizing Eq. (1), small $I(Z^\star; \tilde{Y})$ can be achieved as indicated by the upper bound in Eq. (12). In the meanwhile, the lower bound on $I(Z^\star; Y)$ in Eq. (11) also shows that $Z^\star$ can retain the discriminative information of the data to avoid a trivial solution to $I(Z_\theta; \tilde{Y})$ minimization (i.e., Z_θ is a constant representation).

While Lemma 1 combined with Theorem 2 indicates that $Z^\star$ is robust to label noise, the following Lemma shows that

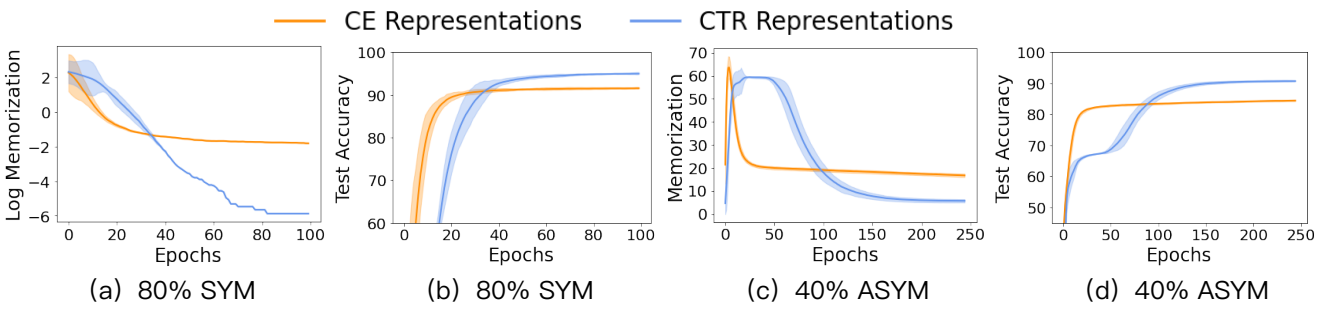

Figure 3. Results of memorization of label noise and performance on test data on CIFAR-10 with 80% symmetric label noise (SYM) and 40% asymmetric label noise (ASYM). The memorization is defined by the fraction of wrongly labeled examples whose predictions are equal to their labels.

$Z^\star$ can also avoid underfitting. Specifically, it implies that that a good classifier achieved under the clean distribution can also be achieved based on our representations $Z^\star$.

Lemma 2. *Let* $R(X) = \inf_g \mathbb{E}_{X,Y}[\mathcal{L}(g(X), Y)]$ *be the minimum risk over the joint distribution* $X \times Y$*, where* $\mathcal{L}(p, y) = \sum_{i=1}^{\mathcal{Y}} y^{(i)} \log p^{(i)}$ *is a CE loss and* g *is a function mapping from input space to label space. Let* $R(Z^\star) = \inf_{g'} \mathbb{E}_{Z^\star,Y}[\mathcal{L}(g'(Z^\star), Y)]$ *be the minimum risk over the joint distribution* $Z^\star \times Y$ *and* g' *maps from representation space to label space. Then,*

$$R(Z^\star) \leq R(X) + \epsilon.$$

To show the robustness and performance of the contrastive (CTR) representation $Z^\star$, we empirically compare it to the representation learned by the CE loss. We first use clean labels to train neural networks with different loss functions. Then we initialize the parameters of the final linear classifier and fine-tune the linear layer with noisy labels. We denote the memorization by the fraction of corrupted examples whose predictions are equal to their labels. Figure 3 illustrates the improved performance and robustness in terms of test accuracy and reduced memorization with the CTR representation.

Conventionally, the memorization of label noise increases as the training progresses [27, 44]. We remark that previous memorization is observed and proved in *over-parameterized* models, where the ratio of the number parameters and the sample size is around 220. In their settings, the fraction of examples memorized by the model will increase. However, the memorization in our setting is measured on a linear classifier on top of frozen data representations, where the ratio of the number parameters and the sample size is around 0.1, which is *under-parameterized*. This explains why Figure 3 shows that the memorization decreases as the training progresses.

3. Algorithm

In practice, as we are only given a noisy data set, we do not know if a label is clean or not. Consequently, simply minimizing Eq. (1) can lead to deteriorated performances. To see this, note that Eq. (1) is activated only when $\mathbb{1}\{y_i = y_j\} = 1$. Thus, two representations from different classes will be pulled together when there are noisy labels.

Since deep networks first fit examples with clean labels and the probabilistic outputs of these examples are higher than examples with corrupted labels [3, 26], one straightforward approach to tackle this issue is to replace the indicator function with a more reliable criterion $\mathbb{1}\{p_i^\top p_j \geq \tau\}$:

$$\mathcal{L}'_{\text{ctr}}(x_i, x_j) = -\big(\langle \tilde{q}_i, \tilde{z}_j \rangle + \langle \tilde{q}_j, \tilde{z}_i \rangle\big) \mathbb{1}\{p_i^\top p_j \geq \tau\}, \quad (6)$$

where p_i is the probabilistic output produced by linear classifier on the representation of image x_i and τ is a confidence threshold. However, minimizing Eq. (6) only helps representation learning during the early stage. After that period, examples with corrupted labels will dominate the learning procedure since the magnitudes of gradient from correct contrastive pairs overwhelm that from wrong contrastive pairs. In particular, given two clean examples x_i, x_j with $y_i = y_j$ and a wrongly labeled example x_m with $\tilde{y}_m = y_i = y_j$, during the early stage, representations $\tilde{q}_i^\top \tilde{q}_j \to 1$ and $\tilde{q}_i^\top \tilde{q}_m \approx 0$. After the early stage, deep networks starts to fit wrongly labeled data. At this moment, the wrong contrastive pairs (x_i, x_m) and (x_j, x_m) are wrongly pulled together and they impair the representation learning instead of the correct pair (x_i, x_j):

$$\begin{aligned} \left\| \frac{\partial \mathcal{L}'_{\text{ctr}}(x_i, x_m)}{\partial q_i} \right\|_2^2 &= c_i \underbrace{(1 - \tilde{q}_i^\top \tilde{q}_m)}_{\approx 1} \\ \gg c_i \underbrace{(1 - \tilde{q}_i^\top \tilde{q}_j)}_{\approx 0} &= \left\| \frac{\partial \mathcal{L}'_{\text{ctr}}(x_i, x_j)}{\partial q_i} \right\|_2^2, \end{aligned} \quad (7)$$

where $c_i = 1/\|q_i\|_2^2$ and we take h as an identity function

for simplicity. The proof is shown in supplementary materials.

To address this issue, we propose the following regularization function to avoid the negative effects from wrong contrastive pairs:

$$\widetilde{\mathcal{L}}_{\text{ctr}}(x_i, x_j) = \left(\log\left(1 - \langle \tilde{q}_i, \tilde{z}_j \rangle\right) + \log\left(1 - \langle \tilde{q}_j, \tilde{z}_i \rangle\right) \right) \mathbb{1}\{p_i^\top p_j \geq \tau\} \tag{8}$$

Eq. (8) still aims to learn similar representations for data with the same true labels. Since the maximum of Eq. (8) is the same as that of Eq. (1), our theoretical results about $Z^\star$ still hold. Moreover, the gradient analysis of Eq. (8) is given by

$$\left\| \frac{\partial \widetilde{\mathcal{L}}_{\text{ctr}}(x_i, x_j)}{\partial q_i} \right\|_2^2 = c_i(1 + \tilde{q}_i^\top \tilde{q}_j), \tag{9}$$

which indicates that the gradient in L2 norm increases if $\tilde{q}_i$ and $\tilde{q}_j$ approach to each other. In other words, the gradient from the correct pair (x_i, x_j) is larger than the gradient from the wrong pair (x_i, x_m) $(1 + \tilde{q}_i^\top \tilde{q}_j > 1 + \tilde{q}_i^\top \tilde{q}_m \approx 1)$ during the learning procedure. Compared to the gradient given by Eq. (7), our proposed regularization function does not suffer from the gradient domination by wrong pairs. Meanwhile, the model does not overfit clean examples even though the gradients of Eq. (8) from correct pairs are larger than wrong pairs. As Eq. (7) describes the gradient with respect to the representation, its magnitude can be viewed as the strength of pulling clean examples from the same class closer, which is *not* directly related to overfitting to clean examples. Moreover, we use a separate linear layer on top of the representations as the classifier, thus as long as the gradients of the classification loss with respect to the parameters in the linear layer are not large on the clean examples, the model would not overfit to them.

Finally, the overall objective function is given by

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \widetilde{\mathcal{L}}_{\text{ctr}}, \tag{10}$$

where $\widetilde{\mathcal{L}}_{\text{ctr}}$ serves as a contrastive regularization (CTRR) on representations and λ controls the strength of the regularization.

4. Experiments

Datasets. We evaluate our method on two artificially corrupted datasets CIFAR-10 [15] and CIFAR-100 [15], and two real-world datasets ANIMAL-10N [36] and Clothing1M [47]. CIFAR-10 and CIFAR1-00 contain $50,000$ training images and $10,000$ test images with 10 and 100 classes, respectively. ANIMAL-10N has 10 animal classes and $50,000$ training images with confusing appearances and 5000 test images. Its estimated noise level is around 8%. Clothing1M has a million training images and $10,000$ test images with 14 classes. Its estimated noise level is around 40%.

Noise Generation. For CIFAR-10, we consider two different types of synthetic noise with various noise levels. For symmetric noise, each label has the same probability of flipping to any other classes, and we randomly choose r training data with their labels to be flipped for $r \in \{20\%, 40\%, 60\%, 80\%, 90\%\}$. For asymmetric noise, following [6], we flip labels between TRUCK→AUTOMOBILE, BIRD→AIRPLANE, DEER→HORSE, and CAT↔DOG. we randomly choose 40% training data with their labels to be flipped according to the asymmetric labeling rule. For CIFAR-100, we also consider two different types of synthetic noise with various noise levels. The generation for symmetric label noise is the same as that for CIFAR-10 with the noise level $r \in \{20\%, 40\%, 60\%, 80\%\}$. To generate asymmetric label noise, we randomly sample 40% data and flip their labels to the next classes.

Baseline methods. To evaluate our method, we mainly compare our robust loss function to other robust loss function methods: 1) CE loss. 2) Forward correction [33], which corrects loss values by a estimated noise transition matrix. 3) GCE [55], which takes advantages of both MAE loss and CE loss and designs a robust loss function. 4) Co-teaching [17], which maintains two networks and uses small-loss examples to update. 5) LIMIT [18], which introduces noise to gradients to avoid memorization. 6) SLN [6], which adds Gaussian noise to noisy labels to combat label noise. 7) SL [42], which uses CE loss and a reverse cross entropy loss (RCE) as a robust loss function. 8) APL (NCE+RCE) [29], which combines two mutually boosted robust loss functions for training.

Implementation details. We use a PreAct Resnet18 as the encoder for CIFAR datasets, and Resnet18 as the encoder for the two real-world datasets. The project MLP and the prediction MLP are the same for all encoders. Following SimSiam [9], the projection MLP consists of 3 layers which have 2048 hidden dimensions and output 2048-dimensional embeddings. The prediction MLP consists of 2 layers which have 512 hidden dimensions and output 2048-dimensional embeddings. Following [7], we apply strong augmentations to learn data representations, where the strong augmentation includes Gaussian blur, color distortion, random flipping and random cropping. We use weak augmentations to optimize the cross-entropy loss, which includes random flipping and random cropping. More implementation details can be found in supplementary materials.

Method	CIFAR-10						
	Sym.						Asym.
	0%	20%	40%	60%	80%	90%	40%
CE	$93.97_{\pm 0.22}$	$88.51_{\pm 0.17}$	$82.73_{\pm 0.16}$	$76.26_{\pm 0.29}$	$59.25_{\pm 1.01}$	$39.43_{\pm 1.17}$	$83.23_{\pm 0.59}$
Forward	$93.47_{\pm 0.19}$	$88.87_{\pm 0.21}$	$83.28_{\pm 0.37}$	$75.15_{\pm 0.73}$	$58.58_{\pm 1.05}$	$38.49_{\pm 1.02}$	$82.93_{\pm 0.74}$
GCE	$92.38_{\pm 0.32}$	$91.22_{\pm 0.25}$	$89.26_{\pm 0.34}$	$85.76_{\pm 0.58}$	$70.57_{\pm 0.83}$	$31.25_{\pm 1.04}$	$82.23_{\pm 0.61}$
Co-teaching	$93.37_{\pm 0.12}$	$92.05_{\pm 0.15}$	$87.73_{\pm 0.17}$	$85.10_{\pm 0.49}$	$44.16_{\pm 0.71}$	$30.39_{\pm 1.08}$	$77.78_{\pm 0.59}$
LIMIT	$93.47_{\pm 0.56}$	$89.63_{\pm 0.42}$	$85.39_{\pm 0.63}$	$78.05_{\pm 0.85}$	$58.71_{\pm 0.83}$	$40.46_{\pm 0.97}$	$83.56_{\pm 0.70}$
SLN	$93.21_{\pm 0.21}$	$88.77_{\pm 0.23}$	$87.03_{\pm 0.70}$	$80.57_{\pm 0.50}$	$63.99_{\pm 0.79}$	$36.64_{\pm 1.77}$	$81.02_{\pm 0.25}$
SL	$94.21_{\pm 0.13}$	$92.45_{\pm 0.08}$	$89.22_{\pm 0.08}$	$84.63_{\pm 0.21}$	$72.59_{\pm 0.23}$	$51.13_{\pm 0.27}$	$83.58_{\pm 0.60}$
APL	$93.97_{\pm 0.25}$	$92.51_{\pm 0.39}$	$89.34_{\pm 0.33}$	$85.01_{\pm 0.17}$	$70.52_{\pm 2.36}$	$49.38_{\pm 2.86}$	$84.06_{\pm 0.20}$
CTRR	$\mathbf{94.29_{\pm 0.21}}$	$\mathbf{93.05_{\pm 0.32}}$	$\mathbf{92.16_{\pm 0.31}}$	$\mathbf{87.34_{\pm 0.84}}$	$\mathbf{83.66_{\pm 0.52}}$	$\mathbf{81.65_{\pm 2.46}}$	$\mathbf{89.00_{\pm 0.56}}$

Table 1. Test accuracy on CIFAR-10 with different noise types and noise levels. All method use the same model PreAct Resnet18 [19] and their best results are reported over three runs.

Method	CIFAR-100					
	Sym.					Asym.
	0%	20%	40%	60%	80%	
CE	$73.21_{\pm 0.14}$	$60.57_{\pm 0.53}$	$52.48_{\pm 0.34}$	$43.20_{\pm 0.21}$	$22.96_{\pm 0.84}$	$44.45_{\pm 0.37}$
Forward	$73.01_{\pm 0.33}$	$58.72_{\pm 0.54}$	$50.10_{\pm 0.84}$	$39.35_{\pm 0.82}$	$17.15_{\pm 1.81}$	-
GCE	$72.27_{\pm 0.27}$	$68.31_{\pm 0.34}$	$62.25_{\pm 0.48}$	$53.86_{\pm 0.95}$	$19.31_{\pm 1.14}$	$46.50_{\pm 0.71}$
Co-teaching	$73.39_{\pm 0.27}$	$65.71_{\pm 0.20}$	$57.64_{\pm 0.71}$	$31.59_{\pm 0.88}$	$15.28_{\pm 1.94}$	-
LIMIT	$65.53_{\pm 0.91}$	$58.02_{\pm 1.93}$	$49.71_{\pm 1.81}$	$37.05_{\pm 1.39}$	$20.01_{\pm 0.11}$	-
SLN	$63.13_{\pm 0.21}$	$55.35_{\pm 1.26}$	$51.39_{\pm 0.48}$	$35.53_{\pm 0.58}$	$11.96_{\pm 2.03}$	-
SL	$72.44_{\pm 0.44}$	$66.46_{\pm 0.26}$	$61.44_{\pm 0.23}$	$54.17_{\pm 1.32}$	$34.22_{\pm 1.06}$	$46.12_{\pm 0.47}$
APL	$73.88_{\pm 0.99}$	$68.09_{\pm 0.15}$	$63.46_{\pm 0.17}$	$53.63_{\pm 0.45}$	$20.00_{\pm 2.02}$	$52.80_{\pm 0.52}$
CTRR	$\mathbf{74.36_{\pm 0.41}}$	$\mathbf{70.09_{\pm 0.45}}$	$\mathbf{65.32_{\pm 0.20}}$	$\mathbf{54.20_{\pm 0.34}}$	$\mathbf{43.69_{\pm 0.28}}$	$\mathbf{54.47_{\pm 0.37}}$

Table 2. Test accuracy on CIFAR-100 with different noise levels. All method use the same model PreAct Resnet18 [19] and their best results are reported over three runs.

Method	ANIMAL-10N	Clothing1M
CE	$83.18_{\pm 0.15}$	$70.88_{\pm 0.45}$
Forward	$83.67_{\pm 0.31}$	$71.23_{\pm 0.39}$
GCE	$84.42_{\pm 0.39}$	$71.34_{\pm 0.12}$
Co-teaching	$85.73_{\pm 0.27}$	$71.68_{\pm 0.21}$
SLN	$83.17_{\pm 0.08}$	$71.17_{\pm 0.12}$
SL	$83.92_{\pm 0.28}$	$72.03_{\pm 0.13}$
APL	$84.25_{\pm 0.11}$	$72.18_{\pm 0.21}$
CTRR	$\mathbf{86.71_{\pm 0.15}}$	$\mathbf{72.71_{\pm 0.19}}$

Table 3. Test accuracy on the real-world datasets ANIMAL-10N and Clothing1M. The results are obtained based on three different runs.

4.1. CIFAR Results

Table 1 and Table 2 show the results on CIFAR-10 and CIFAR-100 with various label noise settings. We use PreAct Resnet18 [19] for all methods and report the best test accuracy for them based on three runs. Our method achieves the best performance on all tested noise settings. The improvement is more substantial when the noise level is higher. Especially when noise levels reach to 80% or even 90%, our method significantly outperforms other methods. For example, on CIFAR-10 with $r = 90\%$, CTRR maintains a high accuracy of 81.65%, whereas the second best one is 49.65%.

4.2. ANIMAL-10N & Clothing1M Results

Table 3 shows the results on the real-world datasets ANIMAL-10N and Clothing1M. All methods use the same model and the best results are reported over three runs. In order to be consistent with previous works for a fair comparison, we use a random initialized Resnet18 and an ImageNet pre-trained Resnet18 on ANIMAL-10N and Clothing1M, respectively, and the best results are reported over three runs. For Clothing1M, following [6,23], we randomly sample a balanced subset of 20.48K images from the noisy

Regularization Functions	CIFAR-10					
	0%	20%	40%	60%	80%	90%
$\mathcal{L}'_{\text{ctr}}(6)$	$93.58_{\pm 0.11}$	$86.05_{\pm 0.33}$	$82.34_{\pm 0.25}$	$74.35_{\pm 0.54}$	$54.83_{\pm 1.00}$	$40.96_{\pm 0.99}$
$\widetilde{\mathcal{L}}_{\text{ctr}}(8)$	$\mathbf{94.29_{\pm 0.21}}$	$\mathbf{93.05_{\pm 0.32}}$	$\mathbf{92.16_{\pm 0.31}}$	$\mathbf{87.34_{\pm 0.84}}$	$\mathbf{83.66_{\pm 0.52}}$	$\mathbf{81.65_{\pm 2.46}}$

Table 4. The performance of the model with respect to different regularization functions.

Contrastive Frameworks	CIFAR-10				
	20%	40%	60%	80%	90%
CTRR (SimSiam)	$93.05_{\pm 0.32}$	$92.16_{\pm 0.31}$	$87.34_{\pm 0.84}$	$83.66_{\pm 0.52}$	$81.65_{\pm 2.46}$
CTRR (SimCLR)	$92.50_{\pm 0.35}$	$90.12_{\pm 0.43}$	$87.41_{\pm 0.83}$	$84.96_{\pm 0.44}$	$79.57_{\pm 1.32}$
CTRR (BYOL)	$93.31_{\pm 0.16}$	$92.12_{\pm 0.16}$	$88.71_{\pm 0.52}$	$86.99_{\pm 0.59}$	$84.31_{\pm 0.66}$

Table 5. Extending our method to other contrasitve learning frameworks.

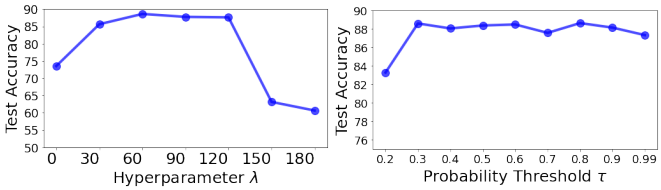

Figure 4. Analysis of λ and τ on CIFAR-10 with 60% symmetric label noise.

training data and report performance on 10K test images. Our method is superior to other baselines on the two real-world datasets.

5. Ablation Studies and Discussions

5.1. Effects of hyperparameters

The hyperparameter λ controls the strength of the regularization to representations of data. A weak regularization is not able to address the memorization issue, while a strong regularization makes the neural network mainly focus on optimizing the regularization term and ignoring optimizing the linear classifier. Figure 4 (left) shows the test accuracy with different λ. The results are in line with the expectation that too strong or too weak regularizations leads to poor performance.

The τ is the confidence threshold for choosing two examples from the same classes. When the score for the two examples exceeds the threshold, the two examples are considered as the correct pair. Many wrong pairs are selected if τ is set too low. Figure 4 (right) shows the test accuracy with different τ. When we are always confident about any pairs (τ=0), the model performance is reduced significantly ($\sim 20\%$).

5.2. Effects of regularization functions

To study the effect of the proposed regularization function, we compare the performance of Eq. (8) to Eq. (6). Empirical results are consistent with the previous gradient analysis and they are shown in Table 4. Our proposed regularization function Eq. (8) outperforms Eq. (6) by a large margin across all noise levels. Thus, learning data representations with Eq. (8) can avoid wrong pairs dominating the representation learning.

5.3. Other contrasitve learning frameworks

Since the InfoMax principle [40] of contrastive learning and the gradient analysis can apply to other contrastive learning frameworks, we apply CTRR to other contrastive learning frameworks. Table 5 shows that our principle is not limited to the SimSiam framework but can also be applied on other contrastive learning frameworks. Since BYOL leverages an additional exponential moving average model to learn representations, the performance of CTRR with BYOL performs better, compared with SimSiam. CTRR works slightly worse under SimCLR than the other two frameworks. For its implementation, we simply replace the inner product of positive representations in SimCLR with our regularization function Eq. (8) and keep the SimCLR objective function from negative pairs the same. A study on how negative pairs from SimCLR affects representation learning in presence of the label noise is beyond the scope of this paper.

5.4. Combination with other methods

Furthermore, CTRR is orthogonal to label correction techniques [27, 54]. In other words, our method can be integrated with these techniques to further boost learning performances. Specifically, we use the basic label correction strategy following [6] that labels are replaced by weighted averaged of both model predictions and original

Label Correction Technique	CIFAR-10			
	20%	40%	60%	80%
✗	$93.05_{\pm 0.32}$	$92.16_{\pm 0.31}$	$87.34_{\pm 0.84}$	$83.66_{\pm 0.52}$
✓	$\mathbf{93.32}_{\pm \mathbf{0.11}}$	$\mathbf{92.76}_{\pm \mathbf{0.67}}$	$\mathbf{89.23}_{\pm \mathbf{0.18}}$	$\mathbf{85.40}_{\pm \mathbf{0.93}}$

Table 6. ✓/✗ indicates the label correction technique is enabled/disabled.

Method	CIFAR-10			
	20%	40%	60%	80%
GCE	$91.22_{\pm 0.25}$	$89.26_{\pm 0.34}$	$85.76_{\pm 0.58}$	$70.57_{\pm 0.83}$
CTRR	$93.05_{\pm 0.32}$	$92.16_{\pm 0.31}$	$87.34_{\pm 0.84}$	$83.66_{\pm 0.52}$
CTRR+GCE	$\mathbf{93.94}_{\pm \mathbf{0.09}}$	$\mathbf{93.06}_{\pm \mathbf{0.29}}$	$\mathbf{92.79}_{\pm \mathbf{0.06}}$	$\mathbf{90.25}_{\pm \mathbf{0.40}}$

Table 7. The performance of the model with respect to GCE, CTRR and CTRR+GCE.

labels, where weights are scaled sample losses. In Table 6, we show that the performance is improved after enabling a simple label correction technique.

Note that GCE [55] is a partial noise-robust loss function implicitly combined with CE and MAE. It is of interest to re-validate the loss function GCE along with our proposed regularization function. We show the performance of a combination of our method and GCE [55] in Table 7. With representations induced by our proposed method, there is a significant improvement on GCE, which demonstrates the effectiveness of the learned representations. Meanwhile, the success of this combination implies that our proposed method is beneficial to other partial noise-robust loss functions.

6. Related Work

In this section, we briefly review existing approaches for learning with label noise.

Noise-robust loss functions are designed to achieve a small error on clean data instead of corrupted data while training on the noisy training data [1, 29, 37]. Mean absolute error (MAE) is robust to label noise [13] but it is not able to solve complicated classification tasks. The determinant based mutual information loss L_{DMI} is proved to be robust to label noise [49] but it only works on the instance-independent label noise. The generalized cross entropy (GCE) [55] takes advantages of MAE and implicitly combined it with CE. The symmetric cross entropy (SL) [42] designs a noise-robust reverse cross entropy loss and explicitly combines it with CE. However, they have not completely addressed the issue as CE is prone to memorizing corrupted labels. The LIMIT [18] proposes to add noise to gradients to address the memorization issue. SLN [6] proposes to combat label noise by adding noise to labels of data. However, they may suffer from underfitting problems.

There are many different contrasitve regularization functions and architectures proposed to learn representations such as SimCLR [7], MoCo [8], BYOL [16], SimSiam [9] and SupCon [20], where SupCon is to learn supervised representations with clean labels while others focus on learning self-supervised representations without labels. We aim to learn representations with noisy labels. We mainly follow the SimSiam framework, but our method is not limited to the SimSiam framework. Recently, some methods existing methods [10, 14, 24, 25] leverage contrasitve representation learning to address noisy label problems. Compared to their methods, we theoretically analyze the benefits of learning such contrastive representations and we focus on addressing a fundamental issue of how to avoid wrong contrastive pairs dominating the representation learning.

There are many other methods for learning with noisy labels. Sample selection methods such as Co-teaching [17], Co-teaching+ [52], SELFIE [36], and JoCoR [43] are selecting small loss examples to update models where they treat small loss examples as clean ones. Loss correction methods such as Forward/Backward method [33] modify the sample loss based on a noise transition matrix. Some works propose to improve the estimation of the noise transition matrix such as T-Revision [45] and Dual T [50]. Label correction methods such as ELR [27], M-DYR-H [2] and PENCIL [51] replace noisy labels with pseudo-labels using different strategies. Methods like DivideMix [23] combine the sample selection, label correction and semi-supervised techniques and empirically demonstrate their success to combat noisy labels.

7. Conclusion

We present a simple but effective CTRR to address the memorization issue. Our theoretical analysis indicates that CTRR induces noise-robust representations without suffering from the underfitting problem. From algorithmic perspectives, we propose a novel regularization function to avoid adverse effects from wrong pairs. The empirical results also demonstrate the effectiveness of CTRR. On the one hand, we show the potential combinations of existing methods to improve the model performance. On the other hand, we evaluate our method under different contrastive learning frameworks. Both of them reveal the flexibility of our method and the importance of correctly regularizing data representations. We believe that CTRR can be jointly used with other existing methods to better solve machine learning tasks where there exists label noise.

Acknowledgement

Li Yi was supported by NSERC Discovery Grants Program. Sheng Liu was partially supported by NSF grant DMS 2009752, NSF NRT-HDR Award 1922658 and Alzheimer's Association grant AARG-NTF-21-848627. Boyu Wang was supported by NSERC Discovery Grants Program.

References

[1] Ehsan Amid, Manfred K. Warmuth, Rohan Anil, and Tomer Koren. Robust bi-tempered logistic loss based on bregman divergences. In *Advances in Neural Information Processing Systems*, 2019. 8

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 8

[3] 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 *Proceedings of the 34th International Conference on Machine Learning*, 2017. 1, 4

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 2010. 1

[5] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, 2006. 1

[6] Pengfei Chen, Guangyong Chen, Junjie Ye, Jingwei Zhao, and Pheng-Ann Heng. Noise against noise: stochastic label noise helps combat inherent label noise. In *9th International Conference on Learning Representations*, 2021. 5, 6, 7, 8

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, 2020. 2, 5, 8, 11

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 8

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 5, 8, 11, 12

[10] Madalina Ciortan, Romain Dupuis, and Thomas Peel. A framework using contrastive learning for classification with noisy labels. *Data*, 6(6):61, 2021. 2, 8

[11] Farzan Farnia and David Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems*, 2016. 13

[12] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020. 1

[13] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 8

[14] Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2703–2708, 2021. 2, 8

[15] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 5

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. 2, 8

[17] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018. 5, 8

[18] Hrayr Harutyunyan, Kyle Reing, Greg Ver Steeg, and Aram Galstyan. Improving generalization by controlling label-noise information in neural network weights. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 5, 8

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 6, 11

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020. 2, 8

[21] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, 2020. 2

[22] Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1104–1113. 13

[23] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations*, 2020. 6, 8

[24] Junnan Li, Caiming Xiong, and Steven C.H. Hoi. Learning from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2, 8

[25] Junnan Li, Caiming Xiong, and Steven C. H. Hoi. Mopro: Webly supervised learning with momentum prototypes. In *International Conference on Learning Representations*, 2021. 2, 8

[26] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020. 4

[27] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization pre-

vents memorization of noisy labels. In *Advances in Neural Information Processing Systems*, 2020. 1, 4, 7, 8

[28] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2021. 2

[29] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah M. Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 1, 5, 8

[30] Hartmut Maennel, Ibrahim M. Alabdulmohsin, Ilya O. Tolstikhin, Robert J. N. Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. What do neural networks learn when trained with random labels? In *Advances in Neural Information Processing Systems*, 2020. 1

[31] Naresh Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Trans. Cybern.*, 2013. 1

[32] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for robust training of deep neural networks against noisy labels. In *Advances in Neural Information Processing Systems*, 2020. 1

[33] 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 *Conference on Computer Vision and Pattern Recognition*, 2017. 5, 8

[34] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 1

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017. 3

[36] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: refurbishing unclean samples for robust deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 5, 8

[37] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020. 1, 8

[38] Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory*, 2008. 3

[39] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*, 2021. 12

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 7

[41] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 11

[42] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision*, 2019. 1, 5, 8

[43] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[44] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *9th International Conference on Learning Representations*, 2021. 4

[45] 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*, 2019. 8

[46] 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. 1

[47] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 5

[48] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[49] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *Advances in Neural Information Processing Systems*, 2019. 1, 8

[50] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual T: reducing estimation error for transition matrix in label-noise learning. In *Advances in Neural Information Processing Systems*, 2020. 8

[51] 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. 8

[52] 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. 8

[53] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations*, 2017. 1

[54] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *9th International Conference on Learning Representations*, 2021. 7

[55] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018. 1, 5, 8