Information Sciences 649 (2023) 119647

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Label correction using contrastive prototypical classifier for noisy label learning

Chaoyang Xu [a], Renjie Lin [b], Jinyu Cai [b], Shiping Wang [b,*]

[a] School of Information Engineering, Putian University, Putian 351100, China
[b] College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China

ARTICLE INFO

Keywords:
Deep neural networks
Noisy label learning
Prototypical classification
Contrastive learning
Label correction

ABSTRACT

Deep neural networks typically require a large number of accurately labeled images for training with cross-entropy loss, and often overfit noisy labels. Contrastive learning has proven impressive in noisy label learning because it can learn discrimination representations. However, the weak correlation between the samples and their semantic class, which ignores the correlation between the instances and labels, as well as the instance semantic divergence with the same label, may inevitably lead to class collisions, hampering the label correction. To address these problems, this study proposes a noisy label learning framework that performs label correction and constructs a contrastive prototypical classifier cooperatively. In particular, the prototypical classifier maximizes the distance between the instances and class prototypes to improve the intraclass compactness using contrastive prototypical loss. Furthermore, we provide a theoretical guarantee that the contrastive prototypical loss has a smaller Lipschitz constant and boosts the robustness. Motivated by the theoretical analysis, this framework performs label correction using the prediction of a contrastive prototypical classifier. Extensive experiments demonstrate that the proposed framework achieves superior classification accuracy on synthetic datasets with various noise patterns and levels.

1. Introduction

Noisy label learning in deep neural networks has attracted increasing attention in recent years. For instance, noisy labels are inevitable in web-scale image classification when search engines or crowd-sourced workers are employed [1–3]. Similarly, medical image analysis presents a challenge owing to the high level of noise in the data, which results in the need for the domain expertise to label, and high inter- and intra-observer variability [4]. Service call analysis is another domain where speech data often suffer from personal mood swings and understanding biases [1]. In classification tasks, the cross-entropy loss is the most commonly used function because of its fast convergence and high generalization capability. However, recent studies indicate that training deep neural networks with cross-entropy can cause a model to fit and memorize arbitrary labels, ultimately leading to decreased performance and generalization [5]. Thus, the accurate training of deep neural networks on noisy label datasets poses a significant challenge.

Classical work on mitigating the effects of model overfitting has mainly focused on using explicit regularization techniques [6], avoiding noisy samples [7], reweighting the loss of noisy samples [8], identifying noisy samples and dividing data [9,10], correcting label noise [11–13], and designing robust loss functions [14,15]. Robust loss functions aim to modify the standard cross-entropy

* Corresponding author.
E-mail address: shipingwangphd@163.com (S. Wang).

https://doi.org/10.1016/j.ins.2023.119647
Received 9 March 2023; Received in revised form 18 August 2023; Accepted 28 August 2023

Available online 1 September 2023

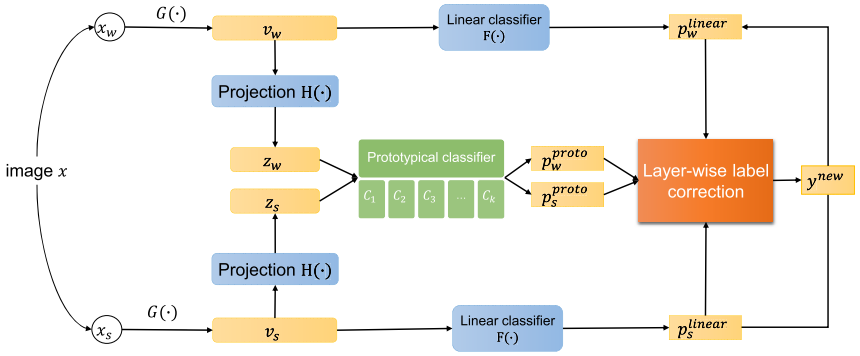


Fig. 1. The input image augmentations x_w and x_s are transmitted to the encoder $G(\cdot)$ to obtain feature representations v_w and v_s, respectively. The feature representations v_w and v_s are fed into the projection layer to produce the contrasting representations z_w and z_s. Formally, a prototypical classifier predicts the p_w^{proto} and p_s^{proto} by comparing the contrastive representations with a set of class prototypes. The outputs p_w^{linear} of the linear classifier and p_w^{proto} of the prototypical classifier are supervised with a pseudo-label generated from the noisy label correction module. We used a model with linear classifier-based predictions at the test time, and the prototypical classifier was discarded after training.

to achieve a better generalization under noisy labels. Ghosh et al. [16] proved a sufficient condition for robust loss functions, such that the risk minimization with the function becomes noise-tolerant for multiclass classification using deep learning. The generalized cross-entropy [14] combines the mean absolute error and standard cross-entropy. The symmetric cross-entropy proposes [15] a reverse cross-entropy loss inspired by the symmetry of the Kullback-Leibler divergence. However, the performance of such a robust loss function is affected significantly by the number of classes and heterogeneous noise patterns [17]. Correcting the label noise generally relies on the linear predictions of the deep model alone [11,13], which are error-prone and not scalable. The neural collapse [18] provides a mathematically elegant formalization, in which deep neural networks can be considered as linear classifiers on top of last-layer features. During training, the linear classifier is optimized by minimizing the cross-entropy loss. Previous studies [19,20] have empirically demonstrated training feature collapse owing to the noisy labels and their effect on model generalization.

Recent studies [21–23] have demonstrated that unsupervised contrastive learning exerts some form of implicit regularization during optimization. Huang et al. [12] utilized the statistics of the unsupervised contrastive loss of each class to achieve dynamic balance during the label correction procedure. Supervised contrastive learning, which involves training a linear classifier on top of a contrastive loss, can learn robust and semantic representations of underlying samples from noisy datasets and has shown better generalization performance for noisy labels [24,25]. Moreover, supervised representation learning can learn robust semantic representations, and many studies have used the similarity between the semantic features of noise samples for label correction. For instance, Li et al. [26] identified confident examples from noisy samples by counting the original labels of the top-K neighbors in terms of their representations, whereas Ortego et al. [10] leveraged the similarity between the noisy samples to detect clean examples from noisy samples and measured the agreement between the pseudo and noisy labels. Huang et al. [27] demonstrated that label correction and contrastive learning are mutually beneficial for noisy label learning. Furthermore, Li et al. [28] showed that learning more semantic representations by utilizing prototypical contrastive loss forces an instance to be more similar to its corresponding class prototype as opposed to the other class prototypes. However, despite its empirical success, the theoretical understanding of the effect of prototypical contrastive learning on improving the robustness of deep networks against noisy labels is limited.

Inspired by the recent success of prototypical and contrastive learning frameworks in improving the robustness and learning semantic representations, we propose a framework with noisy labels that performs label correction and cooperatively learns a contrastive prototypical classifier. The prototypical classifier maximizes the distance between the instances and class prototypes to improve the intraclass compactness using contrastive prototypical loss. This classifier is easy to implement and expand, and it relies on a set of class prototypes, in which examples are clustered around a single prototype representation for each class. Class prototypes are essentially local mean vectors calculated using running averages. The prototypical classifier is optimized via the contrastive prototypical loss, which ensures that instances from the same class are close to the corresponding class prototypes, while remaining far from the prototypes from other classes. Furthermore, we theoretically guarantee that the contrastive prototypical loss has a smaller Lipschitz constant. A smaller Lipschitz constant enables a linear layer trained on such representations to learn clean labels effectively without overfitting noise. Finally, cooperatively performing label correction and learning a contrastive prototypical classifier can be further leveraged by the framework to achieve state-of-the-art performance under label noise.

The proposed framework is illustrated in Fig. 1. The input image augmentations x_w and x_s are transmitted to the encoder $G(\cdot)$ to obtain feature representations v_w and v_s, respectively. The feature representations v_w and v_s are fed into the projection layer to produce the contrasting representations z_w and z_s. Formally, a prototypical classifier predicts the p_w^{proto} and p_s^{proto} by comparing the contrastive representations with a set of class prototypes. The outputs p_w^{linear} of the linear classifier and p_w^{proto} of the prototypical classifier are supervised with a pseudo-label generated from the noisy label correction module. We used a model with linear classifier-based predictions at the test time, and the prototypical classifier was discarded after training. The main contributions of this study are summarized as follows.

1) A prototypical classifier, which is easy to implement and extend, is proposed to learn the semantic representations using contrastive prototype learning.
2) A theoretical guarantee is provided to demonstrate that the contrastive prototypical loss has a smaller Lipschitz constant and boosts robustness.
3) The proposed framework, which performs label correction and cooperatively learns a contrastive prototypical classifier, outperforms the existing label correction methods.

The remainder of this paper is organized as follows: Recent studies on prototypical classifiers and contrastive label correction are presented in Section 2. The proposed noisy label learning framework comprising the implementation details and theoretical analysis is presented in Section 3. An ablation study conducted on the framework and comparison of its performance on synthetic datasets are presented in Section 4. Finally, a summary of this study is presented in Section 5.

2. Related studies

In this section, we first review the recent studies on prototypical classifiers. We then review the existing literature on noisy label correction based on contrastive learning.

2.1. Prototype-based learning

Prototype-based learning, which classifies images by comparing them to a set of predefined prototypes or exemplars, has been extensively studied in the field of machine learning. One of the most influential prototype-based learning methods is the Prototypical Networks (ProtoNets) [29], in which each class prototype is the mean vector of the feature representation from a support set belonging to its class. Class prototypes are used to classify new and unseen instances. Prototype-based learning has also been applied to other computer vision tasks, such as unsupervised representation learning [30–32], class-imbalanced learning [33], imperfect annotations [34], partial label learning [35], noisy label learning [28], and semi-supervised learning [36]. The ProtoNCE [30] and ProPos [31] employ the K-means algorithm, whereas the SwAV [32] utilizes the Sinkhorn-Knopp algorithm to generate the class prototypes from contrastive representations.

Our approach is comparable to that of the RRL [28]. However, RRL relies on the normalized mean vector of contrastive representations corresponding to its class, which may not be scalable. Moreover, the robustness of the class prediction for new samples lacks theoretical analysis.

2.2. Contrastive label correction

Label correction is a technique used to improve the robustness of deep neural networks in labeling the noise [37,38]. To this end, several approaches have been proposed, including the model prediction, co-training, curriculum learning, feature representation, and clustering the characteristics of samples. However, label correction using model predictions can produce ambiguous pseudo-labels, leading to a confirmation bias. Various methods have been proposed for addressing this problem. For instance, some studies [11,13] constrained the consistency between the past model predictions at different epochs to generate more reliable pseudo-labels. Other methods include co-training [21], which leverages the diversity of multiple models to identify and correct errors and curriculum learning [28], which gradually introduces more complex examples after exposure to simpler ones. Xue et al. [39] found that representations learned via contrastive learning have one prominent singular value corresponding to each subclass in the data and significantly smaller remaining singular values. The soft-label scores are the same as the label distribution [40], which covers a certain number of labels and represents the degree, to which each label describes an instance. The process of generating soft label scores is similar to that of recovering the label distribution, which is called the label enhancement [40–42].

Recent studies [43,44] have exploited the structure of a low-dimensional subspace to alleviate the confirmation bias problem by aggregating information from the top-k neighboring samples. In addition, with the development of supervised contrastive learning, it has become possible to learn robust semantic representations. Several studies [10,26,27] have leveraged these representations and employed the clustering characteristics of samples to identify the confident examples of noisy samples. The prototypical contrastive loss [28] has been used to enforce instances that are more similar to their corresponding class prototypes, resulting in a more accurate semantic representation. However, there are still challenges to be addressed, such as the complexity of finding the top-k neighboring samples and ensuring the heterogeneity and homogeneity of the class prototypes.

3. Proposed framework

This section introduces a method for training a prototypical classifier, and provides a theoretical guarantee that the classifier outputs are robust. The proposed LC-CPC framework, called the LC-CPC, learns a contrastive prototypical classifier using label correction. The LC-CPC is designed for a noisy dataset with N samples of K classes, denoted as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{1, \cdots, K\}$ representing the noisy label of the sample.

Fig. 1 illustrates the LC-CPC framework, which consists of several component networks: a CNN backbone $G(\cdot)$ that extracts the representations v from the augmented images $v = G(x)$; projection network $H(\cdot)$ that maps the feature vectors v into a low-dimensional contrastive representation $z = H(v)$; linear classifier $F(\cdot)$ that predicts the model output $p^{linear} = F(v)$; and prototypical

classifier that generates the prototypical prediction p^{proto}. Linear and prototypical classifiers are jointly trained by performing label corrections. Following subsections provide the detailed descriptions of each component of the proposed approach.

3.1. Prototypical classifier learning

The prototypical classifier comprises K learnable class prototypes $\mathbf{c}_k$, where k is the index of classes K. The class prototypes $\mathbf{c}_k$ are the normalized vectors, as is the contrastive representation z. The prototypical classifier produces a prototypical prediction p_i^{proto} over the classes for a given representation z_i, based on a softmax function that measures the distances between the representation and class prototypes in the contrastive representation space. Specifically, when using the cosine similarity as a distance measure, we obtain:

$$p_i^{proto} = \frac{\exp\left(\mathbf{c}_k^\top \cdot z_i/\tau\right)}{\sum_{j=1}^{k} \exp\left(\mathbf{c}_j^\top \cdot z_i/\tau\right)}, \tag{1}$$

where τ is a temperature parameter.

To learn the parameters of the classifier, contrastive prototypical loss is used, which encourages samples from the same class to be close to their corresponding class prototypes, while keeping them far from the prototypes from other classes. The contrastive prototypical loss is defined as the negative log-probability of noisy label y_i:

$$\mathcal{L}^{proto} = -\sum_{i=1}^{N} y_i \log p_i^{proto}. \tag{2}$$

This is different from the supervised contrastive approach, which aims to keep the samples from the same class together, while pushing the samples from different classes far apart.

Although the class prototypes used in the prototypical classifier and classifier weights implemented with a linear layer in the linear classifier have the same classification effect, their methods of allocating learnable parameters differ. Specifically, the linear classifier is capable of assigning learnable parameters to each class, whereas the prototypical classifier relies on a good feature representation to ensure that the samples from the same class are close to the corresponding class prototypes while being far from the prototypes of other classes. Consequently, the linear classifier can leverage learnable parameters to focus more on the discriminative feature dimensions and suppress the irrelevant feature dimensions by assigning higher or lower weights to different dimensions. In contrast, the prototypical classifier fails to leverage such learnable parameters, and as a result, requires more discriminative feature representations to achieve a good performance. Kornblith et al. [19] found that prototypical and linear classifiers resulted in significantly different levels of class separation. Representations learned from the linear classifier with a higher class separation achieved higher accuracy on the test dataset.

Contrastive prototypical loss can be viewed as an enhanced and extended version of the previous class prototypes that depend on the normalized mean vector of the corresponding contrastive representations. Contrastive prototypical loss consists of two parts: the tightness and contrastiveness

$$\begin{aligned}\mathcal{L}^{proto} &= -\sum_{i=1}^{N} y_i \log p_i^{proto} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp \mathbf{c}_{y_i}^\top z_i}{\sum_{k=1}^{K} \exp \mathbf{c}_k^\top z_i} \\ &= \underbrace{\left(-\frac{1}{N}\sum_{i=1}^{N} \mathbf{c}_{y_i}^\top z_i\right)}_{\text{tightness part}} + \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N} \log \sum_{k=1}^{K} \exp \mathbf{c}_k^\top z_i\right)}_{\text{contrastive part}}.\end{aligned} \tag{3}$$

The tightness part $\mathcal{L}^{tight}$ optimizes the class prototypes to be closer to the samples they represent, and is defined as:

$$\mathcal{L}^{tight} = \frac{1}{N}\sum_{i=1}^{N} -z_i^\top \mathbf{c}_{y_i}. \tag{4}$$

The gradient of the tightness part $\mathcal{L}^{tight}$ with respect to the class prototypes can be obtained directly from

$$\frac{\partial \mathcal{L}^{tight}}{\partial \mathbf{c}_k} = -\frac{1}{N}\sum_{i:\, y_i=k} z_i. \tag{5}$$

By minimizing the tightness loss, we update the class prototypes using the following iterative formula:

$$\mathbf{c}_k^0 = \eta \frac{1}{N}\sum_{i:\, y_i=k} z_i^0, \quad \mathbf{c}_k^{t+1} = \mathbf{c}_k^t + \eta \frac{1}{N}\sum_{i:\, y_i=k} z_i^{t+1}, \tag{6}$$

where t is the iteration index and η is the learning rate. This is equivalent to setting the class prototypes to the mean of the normalized class contrastive representations with momentum updates, where the new prototype is a combination of the new iteration representation mean and previous iteration mean. In the ablation study, we conducted experiments to demonstrate that minimizing

the contrastive prototypical loss yields superior results compared with simply setting the class prototypes to the hard mean of each class.

3.2. Theoretical analysis

This subsection presents a theoretical analysis that guarantees the contrastive risk bound of the prototypical loss and its ability to enhance robustness. We are inspired by work [33] and consider a multi-class classification problem, where $\mathbf{c}_1, \ldots, \mathbf{c}_K$ are given, and z is norm-bounded by B, such that $\|z\|_2 \leq B$. We assume that any loss $L(z,i)$ satisfies the condition $\mathcal{L}_C(z) = \sum_{i=1}^{K} L\left(C^\top z, i\right)$, where $\mathcal{L}_C(z)$ is the λ-Lipschitz. Under the symmetric label noise with $\eta < \frac{K-1}{K}$, we derive the following risk bound:

$$R_L(\hat{f}) - R_L\left(f^*\right) \leq \frac{2\eta\lambda B}{(1-\eta)K-1}, \tag{7}$$

where $\hat{f}$ and f^* denote the global minimizers of $R_L^\eta(f)$ and $R_L(f)$, respectively.

For the contrastive prototypical loss $\mathcal{L}^{proto}$, the risk bound depends on the Lipschitz constant of $\mathcal{L}_C(z)$ when the B and noise rate η are fixed. The Lipschitz constant relies on the selection of C, as shown in Equation (3), where

$$\mathcal{L}^{proto} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp \mathbf{c}_{y_i}^\top z_i}{\sum_{k=1}^{K} \exp \mathbf{c}_k^\top z_i}. \tag{8}$$

The derivative of $\mathcal{L}_C(z)$ with respect to z is

$$\frac{\partial \mathcal{L}_C(z)}{\partial z} = \sum_{i=1}^{K}\left[\mathbf{c}_i - \sum_{j=1}^{K} \frac{\exp\left(\mathbf{c}_j^\top z\right)}{\sum_{t=1}^{K}\exp\left(\mathbf{c}_t^\top z\right)}\mathbf{c}_j\right]. \tag{9}$$

The Lipschitz constant of $\mathcal{L}_C(z)$ depends on the upper bound of $\left\|\frac{\partial \mathcal{L}_C(z)}{\partial z}\right\|_2$. This bound exists when both $\mathbf{c}_i$ and z are normalized, such that $\|\mathbf{c}_i\|_2 = 1$, and $\|z\|_2 = B$. However, cross-entropy is usually not Lipschitz continuous in the case when $\mathbf{c}$ is not normalized, resulting in a Lipschitz constant of $\mathcal{L}_W$ that may be infinitely large, as $\mathbf{c}_i = tz(t \to \infty)$, and $\mathbf{c}_j (j \neq i)$ is fixed.

When both $\mathbf{c}_i$ and z are normalized, we have $\mathbf{c}_1 = \frac{z}{\|z\|_2}$ and $\mathbf{c}_2 = \ldots \mathbf{c}_k = -\frac{z}{\|z\|_2}$. As a result, we obtain

$$\left\|\frac{\partial \mathcal{L}_W(z)}{\partial z}\right\|_2 = \frac{2(\exp(2B)-1)}{\frac{\exp(2B)}{k-1}+1}. \tag{10}$$

In other words, the Lipschitz constant of $\mathcal{L}_C(z)$ is $\frac{2(\exp(2B)-1)}{\frac{\exp(2B)}{K-1}+1}$. We derived the following risk bound based on Equation (7):

$$R_L(\hat{f}) - R_L\left(f^*\right) \leq \frac{2(\exp(2B)-1)}{\frac{\exp(2B)}{K-1}+1}. \tag{11}$$

Therefore, the theoretical analysis demonstrates that by performing ℓ_2 normalization on the features and prototypes, we can provide a risk bound for the loss function $\mathcal{L}^{proto}$, that satisfies the λ-Lipschitz and enhances its ability to tolerate the noise in learning.

3.3. Noise label correction

The key to successful learning with noisy labels lies in the ability to identify and correct them accurately. In particular, label correction and contrastive learning are mutually beneficial for noisy label learning [27]. Furthermore, our theoretical analysis of the comparative prototype loss demonstrates the robustness of the output of the prototype classifier. To achieve this goal, we propose a simple noisy label correction procedure that utilizes the softmax output probability of linear and prototypical classifiers. Specifically, for a given training instance $\mathbf{x}_i$, the predicted pseudo-label probability is obtained as:

$$P_i = \frac{1}{2}(p_i^{linear} + p_i^{proto}). \tag{12}$$

We define a set of clean samples that satisfy two criteria: (1) the soft label score for the original noisy label y_i, $P_i(y_i)$, exceeds a low threshold $\mathcal{T}_{low}$, indicating high confidence in the prediction, and (2) maximum score for any label exceeds a high threshold $\mathcal{T}_{high}$, indicating the potential for correct classification. This set of clean samples was constructed to reduce the effect of noisy labels on the learning algorithm and improve the accuracy of the model. This set of clean samples D_{clean} is defined as:

$$D_{clean} = \left\{x_i, y_i \mid P_i\left(y_i\right) > \mathcal{T}_{low}\right\} \cup \left\{x_i, \hat{y}_i = \arg\max_k P_i(k) \mid \forall \max_k P_i(k) > \mathcal{T}_{high}, k \in \{1, \ldots, K\}\right\}. \tag{13}$$

3.4. Implementation detail

The structure of the training network is illustrated in Fig. 1. The LC-CPC framework consists of a linear classifier and prototypical classifier. Interpolation training strategies have demonstrated excellent performance in classification frameworks and promising results in preventing label noise memorization [6]. Inspired by this success, the linear classifier employs a mixed cross-entropy objective with clean examples, $\mathcal{D}_{clean}$. Virtual examples $\{\bar{x}_i, \bar{y}_i\}$ are generated by linearly interpolating the sample x_i with another randomly selected sample x_j from the same mini-batch as:

$$\bar{x}_i = \lambda_s x_i + (1-\lambda_s)x_j \tag{14}$$

$$\bar{y}_i = \lambda_s y_i + (1-\lambda_s)y_j, \tag{15}$$

where $\lambda_s \sim \text{Beta}(\alpha, \alpha)$ and $\alpha = 5.0$. The loss is defined as:

$$\mathcal{L}_{sup} = -\sum_{i \in D_{clean}} \left[\lambda_s \bar{y}_i \log\left(p(\bar{x}_i)\right) + (1-\lambda_s)\bar{y}_i \log\left(p(\bar{x}_i)\right)\right]. \tag{16}$$

During the early stages of training, a linear classifier with mixed cross-entropy may not generate an effective representation. To overcome this challenge, unsupervised contrastive losses, such as the SimCLR [45], have been used to improve representation learning. Unsupervised contrastive learning aims to learn the discrimination representations of samples without label supervision by mapping two weakly augmented views of the same image in a minibatch to neighboring embeddings. Specifically, M images $\{x_i\}_{i=1}^{M}$ are randomly sampled and a pair of augmentations is generated for each image in a mini-batch, producing an augmented batch $\mathcal{B}^a$ with size M, denoted as $\mathcal{B}^a = \{x_i, x_i^a\}_{i=1}^{M}$. Unsupervised contrastive loss is defined as:

$$\mathcal{L}_{SimCLR} = \sum_{i=1}^{B} -\log \frac{\exp(z_i \cdot z_i^a/\tau)}{\sum_{j=1}^{M} \mathbf{1}_{j \neq i} \cdot \exp(z_i \cdot z_j/\tau) + \sum_{j=1}^{M} \exp(z_i \cdot z_j^a/\tau)}, \tag{17}$$

where τ is a temperature parameter as defined in Equation (1).

The overall loss function is expressed as:

$$\mathcal{L}_{overall} = \mathcal{L}_{sup} + \omega \mathcal{L}_{proto} + \gamma \mathcal{L}_{SimCLR}. \tag{18}$$

The contrasting prototypical loss weights must be reduced to synchronize the training and test processes. To balance the mixup cross-entropy, unsupervised contrastive, and contrastive prototypical losses, the linear parameters ω and γ are introduced, which both are initialized as $\frac{\text{max.epoch}-\text{epoch}}{\text{max.epoch}}$, where epoch and max.epoch represent the current training and total number of epochs, respectively. Finally, the model employs the linear classifier-based predictions at test time, and the prototypical classifier is discarded after training.

Algorithm 1 summarized the proposed algorithm.

Algorithm 1 Training procedure for the proposed LC-CPC framework.

Input: Noisy dataset: $D = \{(x, y)\}$, low threshold: P_{low}, high threshold: P_{high}, number of epochs: T.
Output: LC-CPC model.

for epoch = 1 to T **do**:
  Obtain p^{linear}, p^{proto}, z_w, z_s from the model;
  Obtain predicted pseudo-label probability P using Equation (12);
  Obtain clean subset D^t_{clean} using Equation (13);
  for each batch $\mathcal{X}_b$ from $\mathcal{X}$ **do**:
    Update mixup cross-entropy $\mathcal{L}_{sup}$ by Equation (16);
    Update unsupervised contrastive loss $\mathcal{L}_{SimCLR}$ by Equation (17);
    Update contrastive prototypical loss $\mathcal{L}_{proto}$ by Equation (3);
    Train the model using Equation (18);
  end for;
end for;
Return LC-CPC model.

4. Experiments

This section presents comparative evaluations of the proposed framework with previously published frameworks on the CIFAR-10/100 datasets that contain noise from varying patterns and levels. In addition, ablation experiments were conducted to analyze the independent effects of the prototype classifier and noisy label correction module on the framework performance.

4.1. Synthetic noise dataset

The CIFAR-10 and CIFAR-100 datasets, originally introduced by Krizhevsky et al. [46], comprise $50,000$ training images and $10,000$ test images with of size $32 * 32 * 3$ pixels. The CIFAR-10 dataset includes 10 categories, namely airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks, whereas the CIFAR-100 dataset includes 10 super-classes and 100 fine-grained

Table 1
Test accuracy (%) on CIFAR-10 and CIFAR-100 corresponding to different levels of RCN and CCN noise. The best results are indicated in bold.

		RCN				CCN		
Method/Noise ratio		20%	40%	60%	80%	20%	30%	40%
CIFAR10	Standard	88.51	82.73	76.26	59.25	86.12	81.70	83.23
	Mixup	87.30	76.60	71.29	46.70	88.00	83.34	77.27
	LRT	90.36	87.42	82.12	51.15	91.48	90.63	88.72
	PLC	86.74	83.16	69.65	33.56	88.95	86.76	81.98
	BLC	90.88	88.16	84.26	81.86	91.32	90.66	88.45
	CLC	92.96	88.68	88.37	84.34	91.44	90.82	89.61
	CTRR	93.05	92.16	87.34	83.66	-	-	89.00
	Colearning	92.21	87.34	83.41	61.20	91.07	86.89	81.42
	LC-CPC	**94.09**	**93.97**	**92.04**	**88.19**	**92.84**	**92.16**	**91.31**
CIFAR100	Standard	60.57	52.48	43.20	22.96	63.60	53.28	44.45
	Mixup	66.34	52.18	34.52	17.60	65.10	57.36	48.02
	LRT	66.62	59.68	45.79	26.32	67.25	60.15	48.44
	PLC	65.43	54.88	42.28	23.56	64.43	57.32	42.68
	BLC	72.22	66.86	55.96	34.36	73.68	**72.36**	**69.42**
	CLC	72.86	67.46	56.88	36.16	74.24	**73.10**	**70.26**
	CTRR	70.09	65.32	54.20	43.69	-	-	54.47
	Colearning	66.58	56.57	50.11	35.45	65.26	56.97	47.62
	LC-CPC	**74.25**	**71.53**	**67.26**	**46.21**	**74.59**	69.12	60.63

classes. Our study focused on three types of synthetic label noise: the random classification label noise (RCN), class conditional label noise (CCN), and instance dependent label noise (IDN), each designed to be controlled and reproducible. To generate noise labels, we adopted criteria similar to those used in the previous studies, such as the study by Huang et al. [27]. RCN is introduced by randomly altering the ground-truth labels to incorrect labels, while CCN changes the ground-truth labels to specific incorrect labels, such as $cat <-> dog$, $bird -> airplane$, $deer -> horse$, and $truck -> automobile$ in CIFAR-10, and to the 'next' labels within the super-classes in the CIFAR-100. The IDN was generated according to the methodology described by Zhang et al. [47]. Specifically, the IDN can be categorized as Type-I, Type-II, and Type-III. We set the noise level of the IDN to 35%, of the RCN to $[20\%, 40\%, 60\%, 80\%]$, and of the CCN to $[20\%, 30\%, 40\%]$. Each noise pattern and level was saved in a file to ensure the reproducibility of the experimental results.

4.2. Experimental environment

Experiments involving random classification label noise (RCN) and class conditional label noise (CCN) [10] utilize the PreAct ResNet-18 [48] as the CNN backbone, whereas the ResNet-34 [48] is used in the experiments involving instance-dependent label noise (IDN) that we follow [27]. The projection layer had a dimension of 128, and stochastic gradient descent (SGD) is employed to train the framework with a momentum of 0.9, weight decay of 0.0005, and batch size of 256 for 300 epochs. The initial learning rate was 0.03, with a cosine decay schedule adopted, and the model was warmed for 10 epochs before label correction in all the experiments. For all the experiments, P_high was set to 0.9. For CIFAR-10, P_low was set to 0.1 for the RCN and IDN noises, and 0.4 for the CCN noise, while for CIFAR-100, P_low was set to 0.02 for the RCN and CCN noises and 0.1 for the IDN noise.

4.3. Baselines

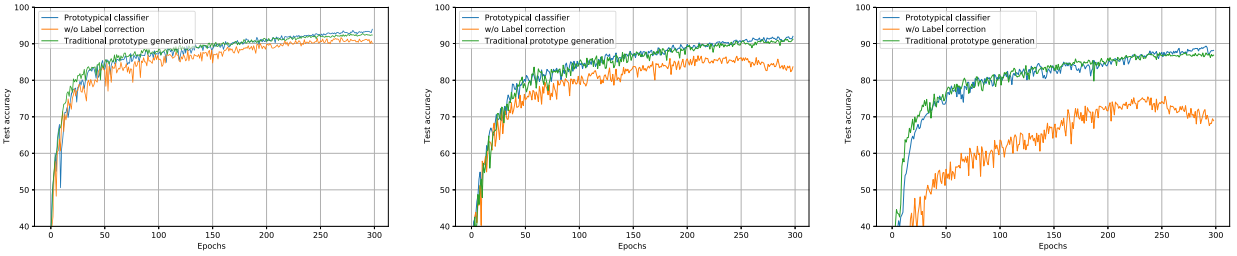
In this study, we compared the proposed method with the following state-of-the-art approaches for learning from noisy labels: (1) standard, which employs the standard cross-entropy for training; (2) mixup [6], which incorporates the linear combinations of the inputs and their corresponding noisy labels; (3) likelihood ratio test (LRT) [13]; (4) progressive label correction (PLC) [47]; and (5) balance label correction (BLC) [12], which employs only model predictions to correct the noisy labels; (6) contrastive label correction (CLC) [27], which performs label correction via expectation-maximization; (7) CTRR model [23], which learns from the noisy labels via high-confidence contrastive pairs; (8) co-learning model [21], which imposes a structure-preserving constraint on pairs of contrastive representations and corresponding model predictions.

4.4. Results

This study reports the average test accuracies of a learning framework under RCN and CCN noise, as shown in Table 1. The results indicate that the CLC, CTRR, and LC-CPC outperform the standard, Mixup, LRT, and PLC methods across all the noise patterns and levels. These findings are consistent with those of the related studies, suggesting the benefits of contrastive learning for noisy label learning [27]. Among all the methods, the CLC and LC-CPC exhibited superior performance in most cases, highlighting the mutual benefits of label correction and contrastive learning during noisy label learning. LC-CPC outperformed all the other methods because prototype learning proved to be more effective in learning the robust representations under noisy labels than the supervised contrastive learning. However, when tested on CIFAR-100 with CCN values of 30% and 40%, LC-CPC's performance was found

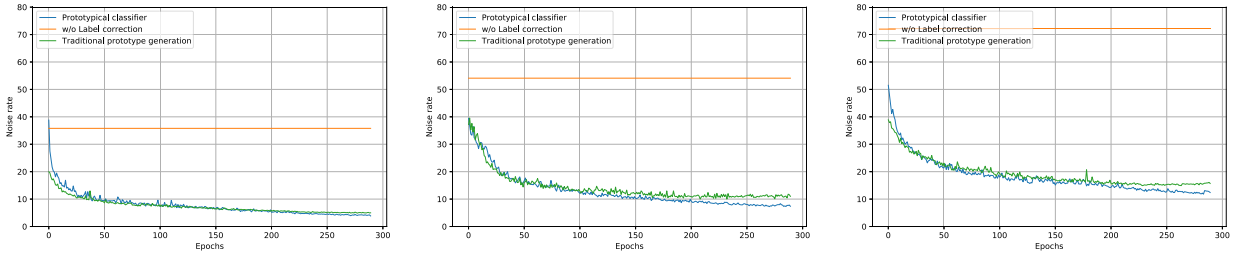
Table 2
Test accuracy (%) on CIFAR-10 and CIFAR-100 in the presence of three types of IDNs with a noise ratio of 35%. The best results are indicated in bold.

Method/Noise ratio		Type-I	Type-II	Type-III
CIFAR100	Standard	78.11 ± 0.74	76.65 ± 0.57	76.89 ± 0.79
	LRT	80.98 ± 0.80	80.74 ± 0.25	81.08 ± 0.35
	PLC	82.80 ± 0.27	81.54 ± 0.47	81.50 ± 0.50
	BLC	83.80 ± 0.12	82.50 ± 0.20	83.60 ± 0.30
	CLC	86.20 ± 0.30	85.70 ± 0.30	86.90 ± 0.20
	LC-CPC	$\mathbf{89.71 \pm 0.30}$	$\mathbf{87.24 \pm 0.20}$	$\mathbf{88.33 \pm 0.40}$
CIFAR100	Standard	57.68 ± 0.29	57.83 ± 0.25	56.07 ± 0.79
	LRT	56.74 ± 0.34	57.25 ± 0.68	56.57 ± 0.30
	PLC	60.01 ± 0.43	63.68 ± 0.29	63.68 ± 0.29
	BLC	63.40 ± 0.20	64.30 ± 0.20	64.10 ± 0.15
	CLC	64.50 ± 0.30	65.10 ± 0.10	65.30 ± 0.26
	LC-CPC	$\mathbf{67.74 \pm 0.12}$	$\mathbf{68.41 \pm 0.25}$	$\mathbf{67.96 \pm 0.23}$



(a) 40% RCN (b) 60% RCN (c) 80% RCN

Fig. 2. Test accuracy of CIFAR-10 dataset with 40%, 60%, and 80% RCN.



(a) 40% RCN (b) 60% RCN (c) 80% RCN

Fig. 3. Noise rate of CIFAR-10 dataset with 40%, 60%, and 80% RCN.

to be inferior to those of BLC and CLC. One possible explanation for this discrepancy is that the contrast prototypical classifier is influenced by severely imbalanced datasets. Further research is necessary to explore effective label-correction methods for these scenarios. The average test accuracies under IDN noise are presented in Table 2, which further confirms the effectiveness of the proposed framework in achieving improvements across various noise types. In summary, the results indicate that the proposed noisy label learning framework is highly effective.

4.5. Ablation study

Influence of prototype generation. We performed experiments on the CIFAR-10 dataset to investigate the effectiveness of the prototypical classifier as opposed to the traditional prototype-generation approach, which involves simply setting the class prototypes to the hard mean for each class. We conducted an ablation study with 40%, 60%, and 80% RCN noise, and the experiments demonstrated that the prototypical classifier outperformed the traditional method. Our objective was to establish and highlight the superior performance of the prototypical classifier in comparison with that of the traditional prototype generation.

The resulting test accuracy and noise rates were recorded over 300 epochs, as shown in Figs. 2 and 3. The results demonstrate that both the prototypical classifier and traditional prototype generation methods can acquire semantic feature representations that facilitate label correction. In addition, these figures reveal that the prototypical classifier achieves higher testing accuracy and better

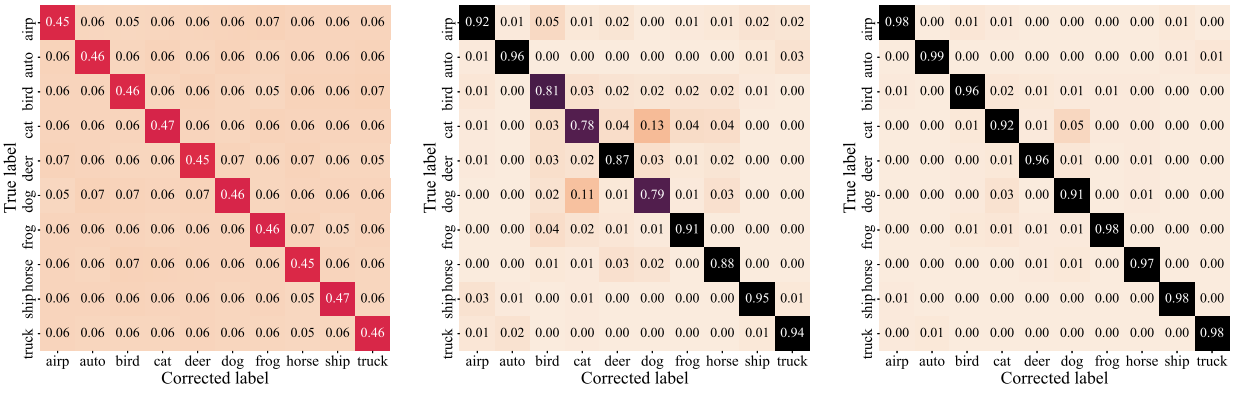


Fig. 4. Confusion matrix of CIFAR-10 dataset with 60% RCN. Correction accuracy is reported in parentheses.

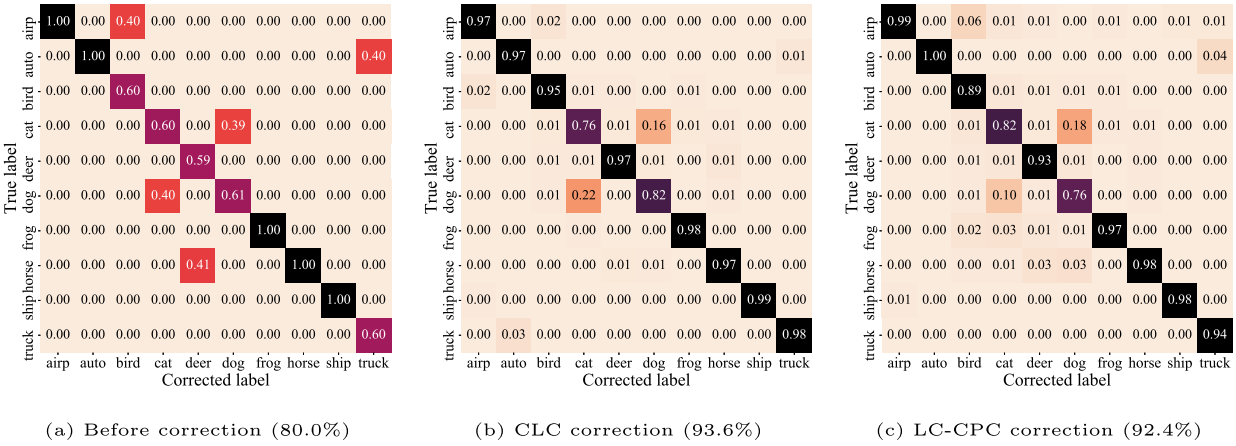


Fig. 5. Confusion matrix of CIFAR-10 dataset with 40% CCN. Correction accuracy is reported in parentheses.

label correction effects than the traditional prototype generation approach. This finding suggests that the prototypical classifier is more proficient at learning semantic feature representations than the traditional prototype generation methods. In addition, Fig. 2 reveals that with a slight amount of noise (40% RCN), the removal of the label correction module has an insignificant impact on the test accuracy. Furthermore, the classifier exhibits strong robustness as it did not fluctuate drastically over the 300 epochs. Conversely, the test accuracy decreased by approximately 10% under severe noise conditions (60% and 80% RCN). These results demonstrate that the label correction module is more effective at facilitating learning with noisy labels.

Results on label correction. In this subsection, we conduct an ablation study on the CIFAR-10 dataset using the LC-CPC and CLC methods to investigate the label correction effectiveness, following the work of Huang et al. [27]. Figs. 4, 5, and 6 show the confusion matrices of the CIFAR-10 dataset with 60%, 40%, and 35% Type I IDN, respectively. The results demonstrate that the LC-CPC achieved accuracies of 96.1%, 92.4%, and 85.2%, respectively, compared with the pre-correction accuracies of 54.1%, 80.0%, and 65.0%, respectively. These findings highlight the label correction effectiveness of the LC-CPC, which learns semantic representations using a contrastive prototypical loss.

Furthermore, we observed that the LC-CPC significantly outperformed the CLC in the case of 60% RCN, indicating the effectiveness of the proposed prototypical classifier. As discussed in Subsection 3.2, the LC-CPC is robust under uniform noise owing to its smaller Lipschitz constant, and its output is helpful in label correction. Interestingly, the LC-CPC demonstrated lower label correction accuracy than that of the CLC in terms of 40% CCN and 35% Type-I noise types. One possible explanation for this is that CLC uses the nearest-neighbor aggregation method for label correction, whereas the LC-CPC performs only simple threshold processing of the output. Therefore, further investigation into effective label correction methods is required.

5. Conclusions

This study presents a novel framework for learning using noisy labels. This framework combined label correction with a contrastive prototypical classifier and achieved significantly improved classification accuracy on synthetic datasets with different types and levels of noise. We demonstrated the effectiveness of the proposed framework through a series of experiments and provide a

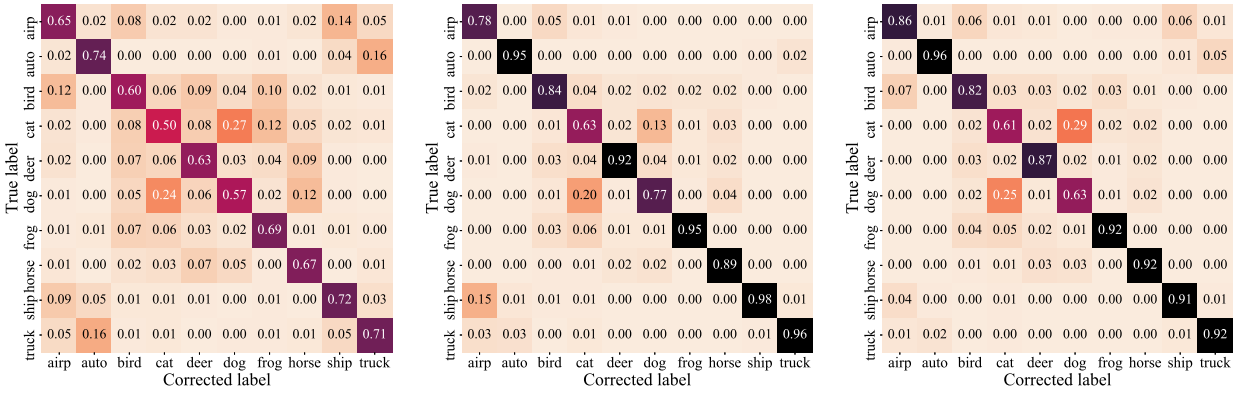


Fig. 6. Confusion matrix of CIFAR-10 dataset with 35% IDN. Correction accuracy is reported in parentheses.

theoretical analysis of contrastive prototypical loss. These results suggest that the proposed framework is effective for various types of noise. An ablation study further confirmed the effectiveness of the label correction component of the framework, which leveraged the contrastive prototypical loss to learn semantic representations. However, further investigation is required to develop effective label correction methods for imbalanced noise data, such as the CCN and IDN.

CRediT authorship contribution statement

Chaoyang Xu: Data curation, Methodology, Software, Validation. **Renjie Lin:** Writing – review & editing. **Jinyu Cai:** Writing – review & editing. **Shiping Wang:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work is in part supported by the National Natural Science Foundation of China under Grants U21A20472 and 62276065, the National Key Research and Development Plan of China under Grant 2021YFB3600503, and the Science and Technology Projects of Fujian Province under Grant No. 2020J01922.

References

[1] G. Algan, I. Ulusoy, Image classification with deep learning in the presence of noisy labels: a survey, Knowl.-Based Syst. 215 (2021) 106771.
[2] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 5 (2014) 2019–2032.
[3] P. Huang, Z. Yang, W. Wang, F. Zhang, Denoising low-rank discrimination based least squares regression for image classification, Inf. Sci. (2022) 247–264.
[4] D. Karimi, H. Dou, S.K. Warfield, A. Gholipour, Deep learning with noisy labels: exploring techniques and remedies in medical image analysis, Med. Image Anal. 65 (2020) 101759.
[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM 64 (3) (2021) 107–115.
[6] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018, pp. 1–9.
[7] H. Wei, L. Feng, X. Chen, B. An, Combating noisy labels by agreement: a joint training method with co-regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13726–13735.
[8] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, D. Meng, Meta-weight-net: learning an explicit mapping for sample weighting, in: Advances in Neural Information Processing Systems, 2019, pp. 1–12.
[9] J. Li, R. Socher, S.C. Hoi, Dividemix: learning with noisy labels as semi-supervised learning, in: International Conference on Machine Learning, 2020, pp. 1–14.
[10] D. Ortego, E. Arazo, P. Albert, N.E. O'Connor, K. McGuinness, Multi-objective interpolation training for robustness to label noise, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6606–6615.
[11] P. Chen, J. Ye, G. Chen, J. Zhao, P.-A. Heng, Beyond class-conditional assumption: a primary attempt to combat instance-dependent label noise, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 1–10.
[12] B. Huang, A. Alhudhaif, F. Alenezi, S.A. Althubiti, C. Xu, Balance label correction using contrastive loss, Inf. Sci. 607 (2022) 1061–1073.

[13] S. Zheng, P. Wu, A. Goswami, M. Goswami, D. Metaxas, C. Chen, Error-bounded correction of noisy labels, in: International Conference on Machine Learning, 2020, pp. 11447–11457.
[14] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in: Advances in Neural Information Processing Systems, 2018, pp. 8778–8788.
[15] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, J. Bailey, Symmetric cross entropy for robust learning with noisy labels, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 322–330.
[16] A. Ghosh, H. Kumar, P.S. Sastry, Robust loss functions under label noise for deep neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 1–9.
[17] M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning, in: International Conference on Machine Learning, 2018, pp. 4334–4343.
[18] V. Papyan, X. Han, D.L. Donoho, Prevalence of neural collapse during the terminal phase of deep learning training, Proc. Natl. Acad. Sci. 117 (40) (2020) 24652–24663.
[19] S. Kornblith, T. Chen, H. Lee, M. Norouzi, Why do better loss functions lead to less transferable features?, in: Advances in Neural Information Processing Systems, 2021, pp. 28648–28662.
[20] D.A. Nguyen, R. Levie, J. Lienen, G. Kutyniok, E. Hüllermeier, Memorization-dilation: modeling neural collapse under noise, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1–32.
[21] C. Tan, J. Xia, L. Wu, S.Z. Li, Co-learning: learning from noisy labels with self-supervision, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1405–1413.
[22] A. Iscen, J. Valmadre, A. Arnab, C. Schmid, Learning with neighbor consistency for noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4672–4681.
[23] L. Yi, S. Liu, Q. She, A.I. McLeod, B. Wang, On learning contrastive representations for learning with noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16682–16691.
[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Advances in Neural Information Processing Systems, 2020, pp. 18661–18673.
[25] F. Graf, C. Hofer, M. Niethammer, R. Kwitt, Dissecting supervised contrastive learning, in: International Conference on Machine Learning, 2021, pp. 3821–3830.
[26] S. Li, X. Xia, S. Ge, T. Liu, Selective-supervised contrastive learning with noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 316–325.
[27] B. Huang, Y. Lin, C. Xu, Contrastive label correction for noisy label learning, Inf. Sci. 611 (2022) 173–184.
[28] J. Li, C. Xiong, S.C. Hoi, Learning from noisy data with robust representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9485–9494.
[29] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Advances in Neural Information Processing Systems, 2017, pp. 1–11.
[30] J. Li, P. Zhou, C. Xiong, S. Hoi, Prototypical contrastive learning of unsupervised representations, in: International Conference on Learning Representations, 2020, pp. 1–16.
[31] Z. Huang, J. Chen, J. Zhang, H. Shan, Learning representation for clustering via prototype scattering and positive sampling, IEEE Trans. Pattern Anal. Mach. Intell. (2022) 1–16.
[32] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. 33 (2020) 9912–9924.
[33] T. Wei, J.-X. Shi, Y.-F. Li, M.-L. Zhang, Prototypical classifier for robust class-imbalanced learning, in: Advances in Knowledge Discovery and Data Mining, Springer, 2022, pp. 44–57.
[34] X. Zhou, X. Liu, D. Zhai, J. Jiang, X. Gao, X. Ji, Prototype-anchored learning for learning with imperfect annotations, in: International Conference on Machine Learning, 2022, pp. 27245–27267.
[35] Y. Yan, Y. Guo, Mutual partial label learning with competitive label noise, in: International Conference on Learning Representations, 2023, pp. 1–14.
[36] H. Xu, L. Liu, Q. Bian, Z. Yang, Semi-supervised semantic segmentation with prototype-based consistency regularization, in: Advances in Neural Information Processing Systems, 2022, pp. 1–18.
[37] W. Li, C. Li, L. Jiang, Learning from crowds with robust logistic regression, Inf. Sci. (2023) 119010.
[38] X. Li, C. Li, L. Jiang, A multi-view-based noise correction algorithm for crowdsourcing learning, Inf. Fusion (2023) 529–541.
[39] Y. Xue, K. Whitecross, B. Mirzasoleiman, Investigating why contrastive learning benefits robustness against label noise, in: International Conference on Machine Learning, 2022, pp. 24851–24871.
[40] N. Xu, Y.-P. Liu, X. Geng, Label enhancement for label distribution learning, IEEE Trans. Knowl. Data Eng. (2019) 1632–1643.
[41] N. Xu, J. Shu, Y.-P. Liu, X. Geng, Variational label enhancement, in: International Conference on Machine Learning, 2020, pp. 10597–10606.
[42] H. Tang, Q. Zheng, J. Zhu, Label information bottleneck for label enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7497–7506.
[43] K. Sharma, P. Donmez, E. Luo, Y. Liu, I.Z. Yalniz, Noiserank: unsupervised label noise reduction with dependence models, in: European Conference on Computer Vision, 2020, pp. 737–753.
[44] P. Wu, S. Zheng, M. Goswami, D. Metaxas, C. Chen, A topological filter for learning with label noise, in: Advances in Neural Information Processing Systems, 2020, pp. 21382–21393.
[45] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, 2020, pp. 1597–1607.
[46] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Master's thesis, Department of Computer Science, University of Toronto.
[47] Y. Zhang, S. Zheng, P. Wu, M. Goswami, C. Chen, Learning with feature-dependent label noise: a progressive approach, in: International Conference on Machine Learning, 2021, pp. 1–13.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Computer Vision and Pattern Recognition, 2016, pp. 770–778.