Noise-Robust Learning from Multiple Unsupervised Sources of Inferred Labels

Amila Silva, Ling Luo, Shanika Karunasekera, Christopher Leckie

School of Computing and Information Systems
The University of Melbourne
Parkville, Victoria, Australia
{amila.silva@student., ling.luo@, karus@, caleckie@}unimelb.edu.au

Abstract

Deep Neural Networks (DNNs) generally require large-scale datasets for training. Since manually obtaining clean labels for large datasets is extremely expensive, unsupervised models based on domain-specific heuristics can be used to efficiently infer the labels for such datasets. However, the labels from such inferred sources are typically noisy, which could easily mislead and lessen the generalizability of DNNs. Most approaches proposed in the literature to address this problem assume the label noise depends only on the true class of an instance (i.e., class-conditional noise). However, this assumption is not realistic for the inferred labels as they are typically inferred based on the features of the instances. The few recent attempts to model such instance-dependent (i.e., feature-dependent) noise require auxiliary information about the label noise (e.g., noise rates or clean samples). This work proposes a theoretically motivated framework to correct label noise in the presence of multiple labels inferred from unsupervised models. The framework consists of two modules: (1) MULTI-IDNC, a novel approach to correct label noise that is instance-dependent yet not class-conditional; (2) MULTI-CCNC, which extends an existing class-conditional noise-robust approach to yield improved class-conditional noise correction using multiple noisy label sources. We conduct experiments using nine real-world datasets for three different classification tasks (images, text and graph nodes). Our results show that our approach achieves notable improvements (e.g., 6.4% in accuracy) against state-of-the-art baselines while dealing with both instance-dependent and class-conditional noise in inferred label sources.

Introduction

Motivation. DNNs have achieved remarkable success in a wide range of applications. However, their performance largely relies on the availability of large-scale labeled datasets. Getting manual labels for large datasets is extremely expensive and time-consuming. As a solution, previous works (Niu et al. 2021; Silva et al. 2020; Yang et al. 2019; Veličković et al. 2019) propose various unsupervised models based on domain-specific heuristics to label large datasets in a time- and cost-effective manner. Such inferred labels generated without manual effort could be subsequently used to learn supervised DNNs. However, the labels from such unsupervised models are typically noisy. As found by recent studies (Zhang et al. 2021; Reed et al. 2015), DNNs can overfit the noise in such noisy labels, which substantially degrades their performance. As a result, learning robust DNNs using noisy labels has recently become a critical problem attracting considerable research effort (Berthelot et al. 2019; Liu et al. 2020; Jiang et al. 2018; Lyu et al. 2020; Yu et al. 2019). Nevertheless, almost all the existing noisy-robust learning strategies do not focus on datasets with multiple noisy inferred labels that are produced using multiple unsupervised models. The availability of multiple noisy label sources could be useful to address several limitations of the existing robust learning techniques based on one noisy label source. ***This work aims to address this research gap by proposing a noise-robust learning paradigm for DNNs using multiple sources of noisy inferred labels.*** Such a learning technique could be practically significant for ensembling of multiple unsupervised label sources with a proper understanding of their noise, while alleviating the following limitations in existing noise-robust learning techniques.

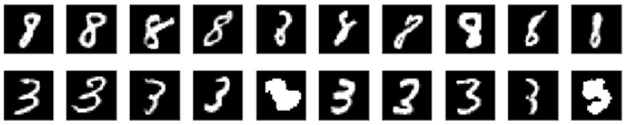

Figure 1: Examples of 8 (first row) and 3 (second row) in MNIST. The label noise could depend on its actual class labels (i.e., CCN) as there are semantically similar classes (e.g., 8 and 3). It is not the only factor to determine label noise as label noise could depend on the features of the instances (i.e., IDN) too – e.g., the last instance in each row.

First, most existing works (Liu et al. 2020; Menon et al. 2020; Ma et al. 2020; Patrini et al. 2017) assume that the noise in the labels (i.e., the probability of flipping to a different label from actual label) are independent of input feature given the actual class labels. However, the study in (Chen et al. 2021) shows that noisy labels in real-world datasets do not follow this assumption. This is because the instances of the same class could have different label noise based on their features as can be seen in Fig. 1. In this work, we decompose the noise in inferred labels into two components: (1) class-conditional noise (CCN), the noise component that is independent from input features given the actual class labels; (2)

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

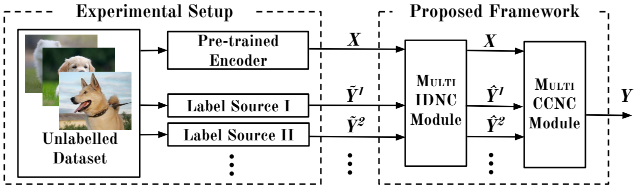

Figure 2: Overview of our experimental framework and the proposed approach for noise-robust learning, which consists of two modules: (1) MULTI-IDNC and (2) MULTI-CCNC, to model IDN and CCN in the labels, respectively.

instance-dependent noise (IDN) that depends on the instance features, but not modelled under CCN. Although there are a few previous attempts at modelling IDN (Cheng et al. 2020; Chen et al. 2021; Zhu et al. 2021; Berthon et al. 2021; Xia et al. 2020), they either require a small dataset with clean labels to train their models or adopt other assumptions that may not be realistic in all practical applications (see Related Work). Also, almost all these previous works do not distinguish IDN and CCN (i.e., define IDN as the combination of IDN and CCN noise components), which makes it difficult to analyse these two types of noise separately. As a solution, ***this work proposes a novel instance-dependent noise correcter based on multiple sources of noisy inferred labels (*MULTI-IDNC*) that is capable of filtering out IDN in noisy labels without affecting CCN.*** MULTI-IDNC is designed such that it can be combined with any CCN correction technique to jointly model both IDN and CCN.

Second, some existing techniques on correcting CCN assume that the noise rates of the label sources are known. Although there are recent attempts (Liu et al. 2020) to correct CCN in label sources without knowing their noise rates, the unavailability of noise rates could impact different aspects of these models (e.g., convergence rate). By selecting peer loss (Liu et al. 2020), one of the strongest baselines in this category, as the base model, ***this work proposes* MULTI-CCNC*, which is capable of revealing the latent noise rates with the help of multiple label sources and improving the correction of CCN in noisy labels.***

In addition, most existing works in this field are not tested using a wide range of practical applications. For example, the works in (Zhu et al. 2021; Cheng et al. 2021) are evaluated only using image classification tasks. To bridge this gap, this work evaluates our approach using various downstream tasks such as image, node and text classification to show the generalizability of the proposed approach for a wide range of practical applications. Also, some works add synthetic noise (e.g., symmetric, uniform) to ground truth labels to generate synthetic noisy labels, which may not reflect the noise in inferred labels from unsupervised models, especially instance-dependent noise (Chen et al. 2021). As a solution, this work adopts multiple unsupervised models to provide realistic noisy inferred labels (see Fig. 2).

Contribution. Our contributions are as follows:

- We propose two theoretical models: (1) MULTI-IDNC and (2) MULTI-CCNC, for modelling IDN and CCN using multiple sources of noisy labels inferred based on unsupervised models. To the best of our knowledge, this is the first attempt to explicitly model IDN and CCN in the labels from unsupervised models.
- We evaluate our approach using three classification tasks (i.e., images, text and nodes) to quantitatively show the potential of our approach to correct realistic noisy labels. Our results show that the proposed model outperforms state-of-the-art methods by as much as 6.4% in accuracy.

Related Work

Learning with Noisy Labels

Our work is primarily categorized under the field of research on learning with noisy labels. The earliest works in this field focus on the *random classification noise* (RCN) model, where observed noisy labels are flipped independently with probability $\in (0, \frac{1}{2}]$ (Bylander 1994; Cesa-Bianchi et al. 1999). However, label noise typically depend on their true labels or features. Hence, how to learn DNNs with label-dependent and feature-dependent noisy labels has attracted considerable attention recently.

Learning with Class-conditional Noisy Labels: Most recent works on this topic are explicitly designed with the CCN assumption, where the label noise can be determined only using actual class labels. With this assumption, the noise transition process can be fully specified by a matrix $T \in \mathbb{R}^{c \times c}$, where c is the number of classes in the particular task. Almost all the works belonging to this subcategory attempt to mitigate the effect of CCN by modelling T either using the prior knowledge of noise rates/types (e.g., uniform/symmetric noise (Ren et al. 2018; Arazo et al. 2019; Lukasik et al. 2020) and tri/column/block-diagonal noise (Han et al. 2018a)) or without relying on such prior knowledge (Xu et al. 2019; Liu et al. 2020). Although the CCN assumption in these works simplifies the noise model, this assumption does not always hold in practical applications and the label noise could depend on the features of instances (Chen et al. 2021). Our model addresses this limitation by modelling both IDN and CCN, which makes our work different from the aforementioned works.

Learning with Instance-dependent Noisy Labels: To the best of our knowledge, there are few previous works that model IDN. Some pioneering works on learning with IDN are restricted to binary classification (Menon et al. 2018; Bootkrajang et al. 2020; Cheng et al. 2020), which considerably restricts their applications in practice. In contrast, our approach can be applied for any general c-class classification task. In addition, most of the works in this category make various assumptions to simplify their IDN model. The work in (Xia et al. 2020) assumes that IDN is parts-dependent, where the instance-dependent transition matrix is modelled as a weighted combination of parts-dependent matrices. Several works assume that only the samples closer to the decision boundary of the Bayes-optimal classifier have strong noise or could be mislabelled (Menon et al. 2018; Wang et al. 2021). The work in (Berthon et al. 2021) assumes that the likelihood of an instance to have a correct label is known. Some works (Cheng et al. 2021) assume that the noise is bounded. Although such assumptions simplify the modelling of IDN, they may not always be realistic. Our model differs from these works as our model does not make such assumptions. In addition, some works (Cheng et al. 2021; Zhu et al. 2021) rely on a clean/nearly clean dataset to learn their models, which have been sampled from the original noisy datasets. These works may not effectively utilize the knowledge in the whole dataset to train their models, and thus, may not be applicable to applications with limited data. In contrast, our model adopts the whole dataset to train the model. In addition, almost all these previous works adopt a single noisy label source to model IDN. Our work explores how we can jointly exploit the knowledge available in multiple labels sources to model IDN, which is another clear distinction between our work and the rest.

Multi-source Learning

Since our work exploits multiple noisy label sources to learn the model, this section discusses the related literature in multi-source learning. Most of the literature on multi-source learning exploit multiple sources (i.e., views or modalities) to extract features (Baltrušaitis et al. 2018; Mogadala et al. 2019; Guo et al. 2019). In contrast, our work exploits multiple label sources. There are a few previous works (Raykar et al. 2009; Yan et al. 2014; Tanno et al. 2019; Li et al. 2020) that propose strategies to learn machine learning models with multiple noisy annotations. The works in (Raykar et al. 2009; Yan et al. 2014; Tanno et al. 2019; Li et al. 2020) propose end-to-end approaches to evaluate the reliability of each noisy source and estimate actual labels from noisy labels. Some of these works (Yan et al. 2014) require a clean dataset to estimate the reliability of label sources. Also, they (Raykar et al. 2009; Yan et al. 2014) mostly fail to outperform majority voting when the number of annotators (i.e., sources) are small ($\leq$ 10). To the best of our knowledge, none of these works are explicitly designed to model IDN with a strong theoretical background. Also, all of these efforts focus on aggregating annotations from multiple annotators in crowd-sourcing platforms, thus, they are not evaluated on a wide range of applications.

Problem Statement

Consider a c-class classification problem with a dataset D that consists of labels inferred from M different noisy label sources inferred by unsupervised models, $D = \{(x_1, \tilde{y}_1^1, ..., \tilde{y}_1^M), ..., (x_N, \tilde{y}_N^1, ..., \tilde{y}_N^M)\}$, where $x_i \in \mathbb{R}^d$ denotes the features of the i^{th} data point, and $\tilde{y}_i^m \in [0,1]^c$ is the noisy label assigned to the i^{th} data point by the m^{th} noisy label source. Let X and $\tilde{Y}^m$ be the corresponding random variable of x_i and $\tilde{y}_i^m$. We decompose the label noise in $\tilde{Y}^m$ into two independent components: (1) CCN that can be determined by only using the actual class labels Y – $P(y_i - \hat{y}_i^m|y_i, x_i) = P(y_i - \hat{y}_i^m|y_i)$; (2) IDN that depends on X, but not modelled under CCN – $P(\hat{y}_i^m - \tilde{y}_i^m|y_i, x_i) = P(\hat{y}_i^m - \tilde{y}_i^m|x_i)$. Here, $\hat{y}_i^m$ is the label after correcting IDN in $\tilde{y}_i^m$, and $\hat{Y}^m$ be the corresponding random variable of $\hat{y}_i^m$.

This problem aims to learn the mapping function $f_\psi : X \rightarrow Y$ that reveals the actual latent label y_i of a given data point i using its features x_i and weak labels $\{\tilde{y}_i^1, ..., \tilde{y}_i^M\}$.

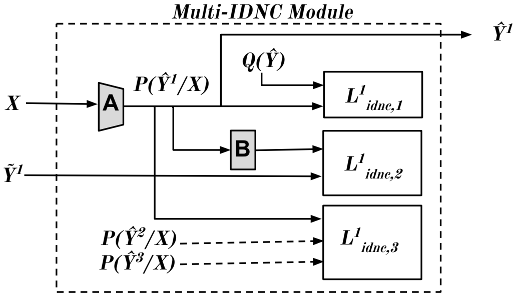

Figure 3: MULTI-IDNC module for a single label source, where **A** and **B** represents the categorical encoder learned with categorical reparameterization trick ($p_\theta(\hat{Y}^m|X)$) and the neural decoder to model $p_\alpha(\tilde{Y}^m|\hat{Y}^m)$ respectively.

Our Approach

We propose a two-step approach to solve the aforementioned learning problem, which initially adopts MULTI-IDNC to correct IDN and subsequently filters out CCN using MULTI-CCNC. MULTI-IDNC learns a mapping function $p_{\theta^m} : X \rightarrow \hat{Y}^m$ for each label source such that $\hat{Y}^m$ satisfies the following conditions: (a) $\hat{Y}^m$ only includes CCN; (b) motivated by the information bottleneck principle (Tishby et al. 1999) as $\hat{Y}^m$ removes the superfluous information in $\tilde{Y}^m$ to predict Y. Consequently, MULTI-IDNC filters out IDN of $\tilde{Y}^m$ while retaining CCN.

In our MULTI-CCNC module, we learn $f_\psi : X \rightarrow Y$ to generate the actual label of an instance from its features. We propose a technique to improve the learning with peer loss (Liu et al. 2020), a CCN-invariant loss function, using multiple IDN-corrected label sources. We show that peer loss learns poorly if we learn it using a noisy label source with a high noise rate. Even if we jointly learn f_ψ using multiple noisy sources, we show that the label sources with high noise rates are under-exploited. To address that, MULTI-CCNC proposes a meta-learning framework to assign weights to the label sources to exploit label sources equally, which improved the quantitative performance and the convergence speed of MULTI-CCNC.

MULTI-IDNC

MULTI-IDNC aims to generate the IDN-corrected label $\hat{Y}^m$ for a given label source m from its original noisy label $\tilde{Y}^m$ (Fig 3). Without loss of generality, the aforementioned objective should satisfy the following conditions:

- $I(X; \hat{Y}^m|Y) = 0$ as $\hat{Y}^m$ should ultimately include only CCN – i.e., $P(\hat{Y}^m|Y, X) = P(\hat{Y}^m|Y)$; and
- $I(X; \hat{Y}^m|\tilde{Y}^m) = 0$ as $\hat{Y}^m$ should only remove IDN from the corresponding $\tilde{Y}^m$ – i.e., $I(X; \hat{Y}^m) \leq I(X; \tilde{Y}^m)$.

where $I(.)$ stands for Mutual Information. Thus, we define the loss function of MULTI-IDNC as follows:

$$L_{idnc} = \sum_{m=1}^{M} \beta * I(X; \hat{Y}^m|Y) + (1 - \beta) * I(X; \hat{Y}^m|\tilde{Y}^m) \tag{1}$$

where $\beta \in [0, 1]$ controls the importance assigned to each loss term in Eq. 1. Instead of optimizing the loss in Eq. 1, we optimize the following upper bound of L_{idnc} (see Fig. 3)[1]:

Theorem 1 *[Upper bound for L_{idnc}]*

$$L_{idnc} \leq \sum_{m=1}^{M} L^m_{idnc,1} + (1-\beta) * L^m_{idnc,2} + \beta * L^m_{idnc,3} \quad (2)$$

where: $L^m_{idnc,1} = I(X; \hat{Y}^m)$, $L^m_{idnc,2} = -I(\hat{Y}^m; \tilde{Y}^m)$ *and* $L^m_{idnc,3} = -I(\hat{Y}^0; \hat{Y}^1; ...; \hat{Y}^M)$*, the negative of the interaction information between* $\{\hat{Y}^0; \hat{Y}^1; ...; \hat{Y}^M\}$.

Optimization of L_{idnc}. The major challenge of solving Eq. 2 is the mutual information (or interaction information) terms that are computationally intractable. Recently, various variational bounds of mutual information (Poole et al. 2019) have been proposed to deal with this problem. Using such bounds, we jointly optimize the three terms–i.e., $L^m_{idnc,1}$; $L^m_{idnc,2}$; and $L^m_{idnc,3}$ in Eq. 2 as follows.

To optimize $L^m_{idnc,1}$, we adopt the following upper bound (Agakov 2004):

Lemma 1 *[Upper bound for $L^m_{idnc,1}$]*

$$L^m_{idnc,1} \leq \mathbb{E}_X KL(p_{\theta^m}(\hat{Y}^m|X)||q(\hat{Y}^m)) \quad (3)$$

where KL(.) denotes the Kullback–Leibler divergence, $q(\hat{Y}^m)$ is the approximated prior of $\hat{Y}^m$, and $p_{\theta^m}(\hat{Y}^m|X)$ is the latent posterior distribution of $\hat{Y}^m$ given X.

We can optimize the RHS of Eq. 3 (i.e., learning θ^m) using backpropagation with the reparameterization trick proposed in (Kingma et al. 2013), which is commonly used to learn variational autoencoders with continuous latent variables (Kingma et al. 2013; Rezende et al. 2014). In our case, the latent variable $\hat{Y}^m$ is discrete with a categorical distribution. Thus, assuming its latent posterior distribution follows a continuous distribution as in (Kingma et al. 2013) it may not be able to preserve the discrete structure in $\hat{Y}^m$. To address this problem, we adopt the categorical reparameterization trick proposed in (Jang et al. 2017) with Gumbel-Softmax that allows learning categorical distribution for the latent posterior distribution using backpropagation.

We can optimize the second loss term $L^m_{idnc,2}$ using the following upper bound:

Lemma 2 *[Upper bound for $L^m_{idnc,2}$]*

$$L^m_{idnc,2} \leq H_{\tilde{Y}^m}(p_{\alpha^m}(\tilde{Y}^m|\hat{Y}^m))$$

where $H_{\tilde{Y}^m}(p_{\alpha^m}(\tilde{Y}^m|\hat{Y}^m))$ is the cross-entropy loss between $\tilde{Y}^m$ and $p_{\alpha^m}(\tilde{Y}^m|\hat{Y}^m)$ – i.e., the latent posterior distribution between $\tilde{Y}^m$ given $\hat{Y}^m$ parameterized by α^m.

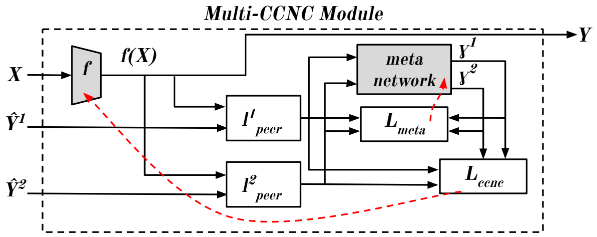

Figure 4: MULTI-CCNC Module, where f is the mapping function that yields Y from X. The $meta\,network$ learns γ^m of each label source. The red arrows show how the loss functions are back-propagated to learn the parameters

The minimization of the third loss term $L^m_{idnc,3}$ is analogous to the maximization of $I(\hat{Y}^0; \hat{Y}^1; ...; \hat{Y}^M)$. Without loss of generality, we can prove that the interaction information between IDN corrected labels from unsupervised models is always positive (see Lemma 3 in (Silva et al. 2021)). Thus, we can maximize $I(\hat{Y}^0; \hat{Y}^1; ...; \hat{Y}^M)$ using the sample-based differentiable mutual information lower bound (i.e., Jensen-Shannon I_{JS}) proposed in (Hjelm et al. 2019). This method requires introducing an auxiliary parameteric model $g_\zeta(\hat{Y}^0; \hat{Y}^1; ...; \hat{Y}^M)$ to approximate the mutual information, which is jointly optimized during the training procedure using the samples from $\bigcup_{\forall m} p_{\theta^m}(\hat{Y}^m|X)$.

MULTI-IDNC adopts the aformentioned optimization procedure to get IDN-free labels that include only CCN. Hence, this module can be universally applied with any learning approach that reveals actual labels from class-conditional noisy labels (e.g., Peer Loss (Liu et al. 2020), L_{DMI} (Xu et al. 2019) and Forward/Backward Correction (Patrini et al. 2017)) to correct both IDN and CCN in the noisy labels. In this work, we propose MULTI-CCNC with multiple class-conditional noisy label sources, which is superior compared to the existing robust learning approaches with a single class-conditional noisy source.

MULTI-CCNC

MULTI-CCNC is motivated by a recently proposed robust loss function called *peer loss* (Liu et al. 2020), which enables learning from a CCN label source without knowing its noise rates. To the best of our knowledge, this is the strongest model at the time of writing this manuscript to learn from CCN labels sources.

Preliminaries on Peer Loss For a given class-conditional noisy label $\hat{Y}^m$, peer loss is defined as follows:

$$l_{PL}(f_\psi(X), \hat{Y}^m) = l(f_\psi(X), \hat{Y}^m) - l(f_\psi(X_{n_1}), \hat{Y}^m_{n2})$$

where l is a surrogate loss function (e.g., cross-entropy loss) f_ψ to $1-0$ loss, $f_\psi(.)$ with ψ parameters maps input features X to the actual labels Y. X_{n_1} and $\hat{Y}^m_{n2}$ corresponding to the peer samples from X and $\hat{Y}^m$ respectively.

Corollary 1 *[Peer loss is invariant to class-conditional label noise (Liu et al. 2020)]*

$$\mathbb{E}[l_{PL}(f_\psi(X), \hat{Y}^m)] = \gamma^m * \mathbb{E}[l_{PL}(f_\psi(X), Y)] \quad (4)$$

[1] Due to space limitations, the Supplementary Material (Silva et al. 2021) provides the proofs and more details about the implementation of the loss terms in MULTI-IDNC and MULTI-CCNC.

where $\gamma^m \in (0, 1]$ is a constant that is monotonically decreasing with the class-conditional noise rates of the corresponding label source m[2].

Therefore, minimizing peer loss using class-conditional noisy labels minimizes peer loss over the true clean distribution. However, there are two limitations in the peer loss function: (1) there is a weak relationship between the peer loss terms with respect to the noisy labels and the corresponding true labels if γ^m of the corresponding noisy source is low due to high noise rates. Consequently, when jointly learning peer loss using multiple noisy sources, the label sources with high noise rates are under-exploited – i.e., the down-weighting issue in peer loss (See Lemma 1 in (Zhu et al. 2021)). Consequently, low γ^m values could also reduce the convergence rate of peer loss; and (2) peer loss does not guarantee to induce $f_\psi(.)$ that minimizes 0-1 loss on the clean dataset when the class distribution in the clean dataset is imbalanced (see Theorem 3 in (Liu et al. 2020)).

Improved Peer Loss using Multiple Label Sources MULTI-CCNC proposes a meta-learning based novel framework to optimize peer loss using multiple noisy labels while alleviating the aforementioned limitations of peer loss. This module aims to minimize the following loss function:

$$L_{ccnc} = \frac{1}{M}\sum_{m=1}^{M}\frac{\mathbb{E}[l_{PL}(f_\psi(X), \hat{Y}^m)]}{\hat{\gamma}^m} + \mathbb{E}[l(f_\psi(X), Y^{f_\psi})]$$

such that $\forall m \in M$,

$$\mathbb{E}[l_{PL}(f_\psi(X), \hat{Y}^m)]/\hat{\gamma}^m = \mathbb{E}[l_{PL}(f_\psi(X), Y^{f_\psi})] \qquad (5)$$

where $Y^{f_\psi} = \text{argmax}(f_\psi(X))$ is the predicted label for each instance from f_ψ and $\{f_\psi, \hat{\gamma}^1, \hat{\gamma}^2, ..., \hat{\gamma}^M\}$ are trainable parameters.

To solve the problem above, we propose a bi-level optimization strategy. In this approach, we learn a feed-forward neural network followed by Sigmoid activation $g_\omega : \{l^1_{PL}, l^2_{PL}, ..., l^M_{PL}, l^{f_\psi}_{PL}\} \rightarrow \{\hat{\gamma}^1, \hat{\gamma}^2, ..., \hat{\gamma}^M\}$ to generate the weighting of the peer loss terms in Eq. 5, where $l^m_{PL} = \mathbb{E}[l_{PL}(f_\psi(X), \hat{Y}^m)]$ and $l^{f_\psi}_{PL} = \mathbb{E}[l_{PL}(f_\psi(X), Y^{f_\psi})]$. g_ω is learned using the following meta-learning loss function:

$$L_{meta} = \sum_{m=1}^{M} p_{\hat{\gamma}}(m) \log p_{\hat{\gamma}}(m) \qquad (6)$$

where $p_{\hat{\gamma}}(m) = \frac{\exp(-l^m_{PL}/\hat{\gamma}^m)}{[\sum_{i=1}^{M}\exp(-l^i_{PL}/\hat{\gamma}^i)] + exp(-l^{f_\psi}_{PL})}$. Then, the optimization procedure can be summarized as follows:

$$\min_\omega L_{meta}(\psi^*, \omega) \;\text{ s.t. }\; \psi^* = argmin_\psi L_{ccnc}(\psi, \omega) \qquad (7)$$

We can prove that the proposed weighted peer loss in Eq. 5 alleviates the aforementioned down-weighting issue in peer loss – i.e., makes the relationship between the peer loss with respect to the noisy labels and the clean labels invariant to the noise rates in the particular source.

Theorem 2 *[Resolving down-weighting issue in peer loss] If f_ψ and γ^m are learned to jointly minimize L_{ccnc} and L_{meta},*

$$\mathbb{E}[l_{PL}(f_\psi(X), \hat{Y}^m)]/\hat{\gamma}^{m_1} = \mathbb{E}[l_{PL}(f_\psi(X), Y)] \qquad (8)$$

Also, it can be proved that the proposed variant of peer loss provides optimality guarantee even with unequal prior.

Theorem 3 *[Optimality guarantee] If f^*_ψ is induced from Eq. 7, then $f^*_\psi \in \text{argmin}(l(f_\psi(X), Y))$.*

Optimization of L_{ccnc} and L_{meta}. We adopt the following one-step SGD update (Shu et al. 2020) to approximate the optimal solution for Eq. 7, from which the gradient for the meta-parameters ω can be estimated as:

$$\nabla_\omega L_{meta}(\psi - \eta\nabla_\psi L_{ccnc}(\psi, \omega), \omega) \qquad (9)$$

where η is the learning rate of SGD. Since both L_{ccnc} and L_{meta} do not require ground truth labels, we jointly optimize both loss functions using the same training dataset.

After learning MULTI-IDNC (θ^m, α^m) and MULTI-CCNC (ψ, ω) as proposed, the noisy labels of unseen data points can be corrected by passing them through these two modules sequentially.

Experiments

Experimental Setup

Dataset Construction We evaluate our approach using three general classification tasks: (1) image; (2) text; and (3) node classification. We select three widely-used datasets for each classification task (see Table 1). We randomly choose 75% of each dataset for training and the remaining 25% for testing. We adopt Accuracy and Normalized Mutual Information Score (NMI) as evaluation metrics.

To have a consistent experimental framework across different tasks, we adopt pretrained embedding techniques to represent the inputs (i.e., text, image, or node) of each task as d-dimensional vectors $x_i \in \mathbb{R}^d$. To generate multiple weak noisy labels $\{y^1_i, y^2_i, ..., y^M_i\}$ for each task, we adopt the state-of-the-art unsupervised clustering techniques in the corresponding domain. The selected embedding techniques and noisy label sources for each task are listed in Table 1 and the default hyper-parameters reported in the original papers of these models are used.

Baselines In Table 2, we compare our approach with two unsupervised ensembling techniques: (1) Majority Voting; (2) DeepCCA (Andrew et al. 2013), and six widely used baselines on learning with noisy labels: (1) Bootstrapping (Reed et al. 2015); (2) Co-teaching (Han et al. 2018b); (3) L_{DMI} (Xu et al. 2019); (4) Peer Loss (Liu et al. 2020); (5) CORES (Cheng et al. 2021); (6) CAL (Zhu et al. 2021).

Parameter Settings We designed the learnable mapping functions (p_{θ^m}, p_{α^m}, f_ψ, g_ω) in our model as feed-forward neural networks[3]. After performing a grid search, we set β

[2] For a binary classification task with a symmetric noisy source, $\gamma^m = 1 - 2 \times noise\ rate$. See (Liu et al. 2020) for more details.

[3] Due to space limitations, we present detailed information about the selected baselines, datasets and the mapping functions in our model in (Silva et al. 2021).

Task	Image Classification	Text Classification	Node Classification
Datasets	CIFAR10, CIFAR100, STL-10	Amazon MR, 20NG, SearchSnippet	Cora, Citeseer, Pubmed
Encoding Method	SimCLR(Chen et al. 2020)	Sentence-BERT (Reimers et al. 2019)	DGI (Veličković et al. 2019)
Noisy Label Source I	CC (Li et al. 2021)	LDA (Blei et al. 2003)	DGI (Veličković et al. 2019)
Noisy Label Source II	SCAN (Gansbeke et al. 2020)	NVDM-GSM (Miao et al. 2017)	SSGC (Zhu et al. 2020)
Noisy Label Source III	SPICE (Niu et al. 2021)	BATM (Wang et al. 2020)	InfoClust (Costas et al. 2021)

Table 1: The selected datasets, encoding techniques, and unsupervised noisy label sources in the experimental setup.

	Text Classification						Image Classification						Node Classification					
	Amazon MR		20NG		SearchSnip.		CIFAR10		CIFAR20		STL10		Cora		Citeseer		Pubmed	
	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
Label Source I	72.5	15.2	19.0	21.7	36.5	18.2	79.0	70.5	48.5	44.7	85.0	76.4	70.5	51.1	69.1	43.8	66.5	29.3
Label Source II	79.7	27.2	24.5	19.8	57.7	33.9	86.8	73.5	50.9	49.9	81.0	61.8	71.5	50.3	68.6	41.3	71.3	32.6
Label Source III	82.3	32.9	35.2	26.2	65.5	44.0	91.7	84.1	53.7	56.5	92.6	84.2	72.1	49.2	69.6	43.8	69.4	30.7
Majority Vote	80.1	28.1	36.5	29.8	62.1	38.9	90.9	83.4	53.5	56.2	91.4	82.2	71.8	52.9	70.5	45.2	71.0	35.2
DeepCCA	81.2	29.7	34.2	24.1	63.1	39.7	89.9	82.1	52.4	54.1	91.3	81.9	71.1	51.7	69.4	42.9	69.7	30.8
CVL	85.9	43.6	37.4	30.4	68.1	47.7	90.6	82.1	54.3	57.5	91.8	82.6	73.2	52.6	70.9	44.5	71.4	33.3
Bootstrapping	83.9	38.7	35.5	27.9	66.2	44.9	90.2	81.9	53.5	56.7	92.3	83.4	72.9	51.5	70.1	44.7	71.8	34.5
Co-teaching	85.3	42.7	36.8	30.2	67.3	47.1	90.7	82.3	53.9	57.2	92.0	83.0	73.9	53.2	71.0	45.8	71.5	33.1
L_DMI	85.2	42.0	36.2	29.6	66.9	45.8	90.8	82.7	54.8	57.5	93.1	82.8	74.1	53.7	70.8	44.9	71.3	32.6
Peer Loss	86.4	44.9	37.3	30.6	68.5	47.6	91.2	83.6	56.4	58.1	92.2	83.7	74.6	54.3	71.2	46.1	71.9	36.5
CORES	86.3	46.8	37.3	30.9	68.8	48.2	92.1	84.7	57.7	59.3	93.6	86.1	74.8	58.1	71.4	46.5	72.0	36.3
CAL	86.7	47.1	37.6	31.1	69.2	49.6	92.3	85.7	58.9	60.3	93.2	84.1	74.4	57.6	71.6	46.8	72.4	36.7
Our Approach	**89.4**	**49.5**	**38.9**	**33.2**	**73.6**	**53.5**	**92.9**	**86.9**	**60.8**	**61.2**	**94.4**	**87.7**	**76.5**	**61.3**	**72.6**	**48.5**	**73.8**	**37.6**
Ablation Study																		
- MULTI-IDNC	87.4	46.9	37.6	31.5	70.3	51.2	92.3	85.6	56.4	59.7	92.7	83.6	75.4	58.8	71.7	46.9	72.4	36.8
- MULTI-CCNC	88.7	48.1	38.3	32.7	72.4	53.0	92.7	86.1	59.3	60.6	93.1	84.2	75.8	60.7	71.9	47.6	72.1	36.6

Table 2: Results for image, text, and node classification tasks

to 0.5 (see Fig. 5 (a)). We (Silva et al. 2021) found that the performance of our model is not particularly sensitive to other hyper-parameters. For the specific parameters of the baselines, we use the default values mentioned in their original papers. We adopt the Adam optimizer and set the $learning\ rate$ and $batch\ size$ to 0.01 and 128 respectively.

Results

Quantitative Results for Classification Tasks As shown in Table 2, the proposed approach yields better results for all three tasks, outperforming the best baseline by as much as 6.4% in accuracy. The improvements are particularly significant for weak label sources. Out of the baselines, IDN correction models – i.e., CAL and CORES, generally achieve better results compared to the other baselines, which shows the importance of correcting IDN. In Table 2, most baselines (except Majority Vote, DeepCCA and our approach) rely on a single label source. For such baselines (e.g., CAL, CORES), we reported the results using the strongest label source (Label Source III). However, the noise rates of the label sources are typically unavailable. Since our approach jointly exploits all label sources, our approach does not require such knowledge about the strongest label source. Compared to the baselines that exploit multiple label sources – i.e., Majority Vote and DeepCCA, our approach achieves up to 16.6% improvement in accuracy.

If we compare the improvements of our approach across different tasks, the improvements are less statistically significant for image classification tasks. This could be because the selected label sources for this task are able to correct noise. For example, SPICE (Label Source III in image classification) model adopts a pseudo labelling approach to generate noisy labels and these labels are subsequently fine-tuned in this model using a confidence-based regularizer. Thus, the level of noise from such a source could be small. We empirically verify this reasoning by comparing the results for image classification tasks without Noisy Label Source III, which drops the accuracy of our approach and CAL, the strongest baseline, by 3% and 5.6% respectively.

Ablation Study Our ablation study in Table 2 shows that without the MULTI-IDNC module ((-) MULTI-IDNC) or MULTI-CCNC module ((-) MULTI-CCNC), the performance of the model could be reduced by as much as 4.7% in accuracy, which verifies the positive contributions of these two modules in the proposed framework. Our MULTI-IDNC module adopts a loss function with three different loss terms – i.e., $L_{idncs,1}$, $L_{idncs,2}$ and $L_{idncs,3}$. To check whether they are contributing towards the final performance of our model, Fig. 5 (a) presents the sensitivity of our model to β parameters on three datasets, which defines the relevant importance of $L_{idncs,2}$ and $L_{idncs,3}$ with respect to $L_{idncs,1}$. By setting $\beta = 0$ and $\beta = 1$, $L_{idncs,2}$ and $L_{idncs,3}$ terms can be removed from our loss function. Fig. 5 (a) shows that our model yields the best results for β values around 0.5, which verifies the importance of the loss terms in L_{idnc}.

Discussion Despite achieving superior performance across a wide range of classification tasks, the reported

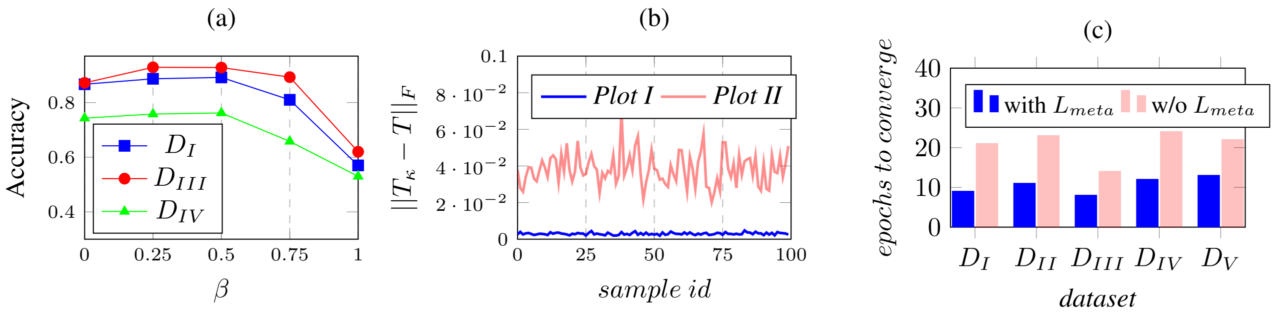

Figure 5: (a) Sensitivity analysis of β parameter; (b) Frobenius norm between the noise transition matrices of a stratified sample (κ) and the complete dataset of D_{IV} using BATM (Wang et al. 2020) as the label source – (1) *Plot I* - considering $\tilde{Y}^m$ as the noisy labels and (2) *Plot II* - considering $\hat{Y}^m$ as the noisy labels; (c) # epochs for convergence ($\Delta L_{peer}(f_\psi(X, Y)) < 0.001$) in MULTI-CCNC with and without the proposed meta-learning based loss weighting approach (D_I = AmazonMR, D_{II} = 20NG, D_{III} = CIFAR10, D_{IV} = Cora, D_V = Pubmed)

	n	λ	$\hat{\lambda}$	$\lambda/\hat{\lambda}$
Label Source I	0.1	0.8	0.811	0.99
Label Source II	0.2	0.6	0.608	0.99
Label Source III	0.3	0.4	0.394	1.02

Table 3: Comparison of the theoretical λ ($= 1 - 2 \times n$) and the learned $\hat{\lambda}$ using three synthetic and symmetric (with noise rate n) noisy label sources of Amazon MR

results in Table 2 do not answer the following two questions: (1) *Can* MULTI-IDNC *correct feature-dependent noise?* and (2) *Can* MULTI-CCNC *identify the class-conditional noise rates in label sources and exploit that to improve learning?* This section attempts to answer these questions.

To answer the first research question, we adopt the following procedure: Step 1 - for a given dataset, select a sample κ from the dataset (the size was set to 10% of the complete dataset) such that the class distribution of the sample is equal to the complete dataset; Step 2 - compute the noise transition matrices using κ (T_κ) and the complete dataset (T); Step 3 - compute the Frobenius norm between T_κ and T; and Step 4 - repeat Steps 1-3 for multiple iterations to compute the distribution of the Frobenius norms. We conduct this procedure considering noisy labels as $\tilde{Y}^m$ and $\hat{Y}^m$. Figure 5 (b) shows the result for this experiment using Amazon MR as the dataset and BATM (Wang et al. 2020) as the label source (m). Intuitively, if the instance-dependent noise are corrected from MULTI-IDNC, T_κ should not largely deviate from T. As can be seen in Fig 5 (b), Frobenius norms of the aforementioned experiments drop after correcting IDN in noisy labels using MULTI-IDNC. This verifies the potential of the proposed MULTI-IDNC module to correct IDN.

For the second research question, we construct three synthetic noisy label sources for the Amazon MR dataset by adding uniform noise for each class with three noise rates – i.e., 10%, 20%, and 30%. After correcting the noise in these sources using MULTI-CCNC, we check the weights $\hat{\gamma}^m$ assigned for each source from the proposed weighting approach in MULTI-CCNC. As shown in Table 3, MULTI-CCNC reveals the latent noise rates (from $\hat{\lambda}$s) in each source and assigns weights to the label sources accordingly. As a result, MULTI-CCNC alleviates the down-weighting issue in peer loss (see Section MULTI-CCNC), which helps to improve the convergence speed of MULTI-CCNC as shown in Fig. 5 (c). Our further experiments using this synthetic dataset showed that MULTI-CCNC outperforms our full model (MULTI-IDNC + MULTI-CCNC) by 0.8% in Accuracy. Thus, our full model may not always be effective for sources that only include CCN, though it is unrealistic to have such inferred sources.

Overall, our experiments verify that the proposed framework effectively exploits multiple noisy label sources to robustly learn deep learning models in the presence of both IDN and CCN, which ultimately helps to achieve superior performance for various practical applications.

Conclusion

In this work, we proposed a noise-robust learning framework for deep learning models in the presence of multiple unsupervised label sources, which consists of two individual modules: (1) MULTI-IDNC; and (2) MULTI-CCNC, to correct instance-dependent and class-conditional noise in the labels. The MULTI-IDNC module is motivated by information theoretic principles, which updates noisy labels to be feature-independent given ground truth labels. Our MULTI-CCNC extends peer loss, a CCN-robust loss function, by proposing a technique to identify class-conditional noise rates of the unsupervised label sources. We extensively evaluated our approach using 9 datasets from 3 domains. Our experiments show that the proposed framework outperforms existing baselines on learning with noisy labels by as much as 6.4% in accuracy with an improved convergence rate.

For future work, we intend to evaluate the potential of our model to correct IDN and CCN in other noisy label sources such as multiple non-expert annotators on crowd-sourcing platforms. Under this setting, the type and rate of noise could be different as they are significantly affected by the level of expertise of the annotators. Since the proposed framework is capable of identifying IDN and CCN in the labels separately, it can be used in future research to deeply analyse the bias of unsupervised machine learning models for each noise type and the properties of each noise type in large datasets.

Acknowledgments

This research was financially supported by Melbourne Graduate Research Scholarship and Rowden White Scholarship. We are also grateful for the time and effort of the reviewers in providing valuable feedback on our manuscript.

References

Agakov, D. B. F. 2004. The IM algorithm: a variational approach to information maximization. In *Poc. of NIPS*.

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *Proc. of ICML*.

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

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. of NeurIPS*.

Berthon, A.; Han, B.; Niu, G.; Liu, T.; and Sugiyama, M. 2021. Confidence scores make instance-dependent label-noise learning possible. In *Proc. of ICML*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of Machine Learning Research*.

Bootkrajang, J.; and Chaijaruwanich, J. 2020. Towards instance-dependent label noise-tolerant classification: a probabilistic approach. *Pattern Analysis and Applications*.

Bylander, T. 1994. Learning linear threshold functions in the presence of classification noise. In *Proc. of COLT*.

Cesa-Bianchi, N.; Dichterman, E.; Fischer, P.; Shamir, E.; and Simon, H. U. 1999. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM (JACM)*.

Chen, P.; Ye, J.; Chen, G.; Zhao, J.; and Heng, P.-A. 2021. Beyond Class-Conditional Assumption: A Primary Attempt to Combat Instance-Dependent Label Noise. In *Proc. of AAAI*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *Proc. of ICML*.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2021. Learning with Instance-Dependent Label Noise: A Sample Sieve Approach. In *Proc. of ICLR*.

Cheng, J.; Liu, T.; Ramamohanarao, K.; and Tao, D. 2020. Learning with bounded instance and label-dependent label noise. In *Proc. of ICML*.

Costas, M.; and Karypis, G. 2021. Graph InfoClust: Leveraging cluster-level node information for unsupervised graph representation learning. In *Proc. of PAKDD*.

Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: Learning to classify images without labels. In *Proc. of ECCV*.

Guo, W.; Wang, J.; and Wang, S. 2019. Deep multimodal representation learning: A survey. *IEEE Access*.

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

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018b. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proc. of NIPS*.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *Proc. of ICLR*.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparameterization with gumbel-softmax. In *Proc. of ICLR*.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proc. of ICML*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, S.; Ge, S.; Hua, Y.; Zhang, C.; Wen, H.; Liu, T.; and Wang, W. 2020. Coupled-view deep classifier learning from multiple noisy annotators. In *Proc. of AAAI*.

Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021. Contrastive clustering. In *Proc. of AAAI*.

Liu, Y.; and Guo, H. 2020. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proc. of ICML*.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *Proc. of ICML*.

Lyu, Y.; and Tsang, I. W. 2020. Curriculum loss: Robust learning and generalization against label corruption. In *Proc. of ICLR*.

Ma, X.; Huang, H.; Wang, Y.; Romano, S.; Erfani, S.; and Bailey, J. 2020. Normalized loss functions for deep learning with noisy labels. In *Proc. of ICML*.

Menon, A. K.; Rawat, A. S.; Reddi, S. J.; and Kumar, S. 2020. Can gradient clipping mitigate label noise? In *Proc. of ICLR*.

Menon, A. K.; Van Rooyen, B.; and Natarajan, N. 2018. Learning from binary labels with instance-dependent noise. *Machine Learning*.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *Proc. of ICML*.

Mogadala, A.; Kalimuthu, M.; and Klakow, D. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*.

Niu, C.; and Wang, G. 2021. Spice: Semantic pseudo-labeling for image clustering. *arXiv:2103.09382*.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proc. of CVPR*.

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On variational bounds of mutual information. In *Poc. of ICML*.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Jerebko, A.; Florin, C.; Valadez, G. H.; Bogoni, L.; and Moy, L. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proc. of ICML*.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2015. Training deep neural networks on noisy labels with bootstrapping. In *Proc. of ICLR*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of EMNLP*.

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

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. of ICML*.

Shu, K.; Zheng, G.; Li, Y.; Mukherjee, S.; Awadallah, A. H.; Ruston, S.; and Liu, H. 2020. Leveraging multi-source weak social supervision for early detection of fake news. In *Proc. ECML-PKDD*.

Silva, A.; Lo, P.-C.; and Lim, E.-P. 2020. JPLink: on linking jobs to vocational interest types. In *Proc. of PAKDD*.

Silva, A.; Luo, L.; Karunasekera, S.; and Leckie, C. 2021. Supplementary Materials for Noise-robust Learning from Multiple Unsupervised Sources of Inferred Labels.

Tanno, R.; Saeedi, A.; Sankaranarayanan, S.; Alexander, D. C.; and Silberman, N. 2019. Learning from noisy labels by regularized estimation of annotator confusion. In *Proc. of CVPR*.

Tishby, N.; Pereira, F. C.; and Bialek, W. 1999. The information bottleneck method. In *Proc. of the Allerton Conference on Communication, Control, and Computing*.

Veličković, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2019. Deep graph infomax. In *Proc. of ICLR*.

Wang, Q.; Han, B.; Liu, T.; Niu, G.; Yang, J.; and Gong, C. 2021. Tackling instance-dependent label noise via a universal probabilistic model. In *Proc. of AAAI*.

Wang, R.; Hu, X.; Zhou, D.; He, Y.; Xiong, Y.; Ye, C.; and Xu, H. 2020. Neural topic modeling with bidirectional adversarial training. In *Proc. of ACL*.

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Part-dependent label noise: Towards instance-dependent label noise. In *Proc. of NeurIPS*.

Xu, Y.; Cao, P.; Kong, Y.; and Wang, Y. 2019. L_DMI: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise. In *Proc. of NeurIPS*.

Yan, Y.; Rosales, R.; Fung, G.; Subramanian, R.; and Dy, J. 2014. Learning from multiple annotators with varying expertise. *Machine learning*.

Yang, S.; Shu, K.; Wang, S.; Gu, R.; Wu, F.; and Liu, H. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proc. of AAAI*.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *Proc. of ICML*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*.

Zhu, H.; and Koniusz, P. 2020. Simple spectral graph convolution. In *Proc. of ICLR*.

Zhu, Z.; Liu, T.; and Liu, Y. 2021. A second-order approach to learning with instance-dependent label noise. In *Proc. of CVPR*.