Uncertainty-aware Learning Against Label Noise on Imbalanced Datasets

Yingsong Huang[1]*, **Bing Bai**[1]*, **Shengwei Zhao**[1], **Kun Bai**[1], **Fei Wang**[2]

[1]Tencent Security Big Data Lab, Tencent Inc., China
[2]Department of Population Health Sciences, Weill Cornell Medicine, USA
{hudsonhuang,icebai,swzhao,kunbai}@tencent.com
few2001@med.cornell.edu

Abstract

Learning against label noise is a vital topic to guarantee a reliable performance for deep neural networks. Recent research usually refers to dynamic noise modeling with model output probabilities and loss values, and then separates clean and noisy samples. These methods have gained notable success. However, unlike cherry-picked data, existing approaches often cannot perform well when facing imbalanced datasets, a common scenario in the real world. We thoroughly investigate this phenomenon and point out two major issues that hinder the performance, i.e., *inter-class loss distribution discrepancy* and *misleading predictions due to uncertainty*. The first issue is that existing methods often perform class-agnostic noise modeling. However, loss distributions show a significant discrepancy among classes under class imbalance, and class-agnostic noise modeling can easily get confused with noisy samples and samples in minority classes. The second issue refers to that models may output misleading predictions due to epistemic uncertainty and aleatoric uncertainty, thus existing methods that rely solely on the output probabilities may fail to distinguish confident samples. Inspired by our observations, we propose an Uncertainty-aware Label Correction framework (ULC) to handle label noise on imbalanced datasets. First, we perform epistemic uncertainty-aware class-specific noise modeling to identify trustworthy clean samples and refine/discard highly confident true/corrupted labels. Then, we introduce aleatoric uncertainty in the subsequent learning process to prevent noise accumulation in the label noise modeling process. We conduct experiments on several synthetic and real-world datasets. The results demonstrate the effectiveness of the proposed method, especially on imbalanced datasets.

Introduction

Widely accessible curated datasets, such as ImageNet (Deng et al. 2009), is one of the crucial factors that deep neural networks (DNNs) flourished. These datasets are usually artificially balanced among classes and carefully annotated by humans. Nevertheless, high-quality human annotation is often expensive and time-consuming. To obtain large-scale annotated data with limited budgets, practical solutions, such as crowd-sourcing (Welinder et al. 2010),

search engines (Fergus et al. 2010), or machine generation (Kuznetsova et al. 2020) are available. However, these methods inevitably introduce mistakes or label noise, which in turn deteriorates the performance of DNNs (Frénay and Verleysen 2013). Another common problem in real-life applications is that some classes have a higher volume of examples in the training set than other classes, which we refer to as the class imbalance problem (Buda, Maki, and Mazurowski 2018). Japkowicz and Stephen (2002) had well studied that the class imbalance problem can significantly affect training machine-learning models. When encountering class imbalance, existing label noise learning methods face new challenges, and how to deal with label noise on imbalanced datasets is essential in practice.

The effectiveness of recent works on noise-robust learning is often attributed to the *memorization effect* (Arpit et al. 2017; Li, Soltanolkotabi, and Oymak 2020), which states that DNNs learn patterns of correctly-labeled data first and then eventually memorize incorrectly-labeled data. On these grounds, based on the given noisy labels and the model's output probabilities, existing methods usually treat samples with small loss values in the early stage of model training as clean ones and then perform loss correction (Reed et al. 2015; Konstantinov and Lampert 2019) or label correction (Arazo et al. 2019; Tanaka et al. 2018).

However, despite their immense success, these methods still face the following issues, especially in class imbalance situations. Firstly, on imbalanced datasets, the loss distributions among classes show a significant discrepancy. As pointed out by Kang et al. (2020), the norms of the classifier layer's weights correlate with the cardinality of the classes. Thus, the predicted probability of majority classes may naturally get more significant than that of minority classes, resulting in the inter-class loss distribution difference. Existing methods that do not consider this difference may easily confuse samples in minority classes and samples with corrupted labels as their loss values all tend to be relatively large, thus getting bad performance for minority classes.

Secondly, models may be uncertain with their predictions, even if they give high output probabilities (Gal 2016; Gal and Ghahramani 2016). Specifically, the wide range of possibilities in model structures and parameters leads to *epistemic uncertainty*, while the uncertainty caused by noisy data is referred to as *aleatoric uncertainty* (Kendall and Gal

---

2017). For learning against label noise, on the one hand, misleading predictions due to *epistemic uncertainty* may affect the identification of clean samples from noisy ones, especially for the minority classes because models see much fewer samples in the minority classes and may get underfit during training. On the other hand, the label noise can hardly be eliminated. The residual noise may accumulate and introduce the *aleatoric uncertainty* to the model, especially for the majority classes and during the late learning phase, where overfitting is more likely to occur. These two kinds of uncertainty may impair the effectiveness of learning against label noise.

To address the shortcomings of existing methods on imbalanced datasets, we introduce a novel Uncertainty-aware Label Correction (ULC) framework. To better distinguish clean samples and correct labels in imbalanced noisy datasets, we improve the common class-agnostic practice by considering the inter-class loss distribution discrepancy problem and modeling samples' noise individually for each class. Furthermore, we estimate the model's epistemic uncertainty and incorporate the epistemic uncertainty in addition to the output probability, which establishes the proposed Epistemic Uncertainty-aware Class-Specific noise modeling method (EUCS). On the other side, as samples filtering and label correction can hardly be perfect, there is always residual noise that may bring negative impacts. We propose the Aleatoric Uncertainty-aware Learning (AUL) against label noise by estimating the sample-wise aleatoric uncertainty induced by the residual noise and adding corresponding logit corruption in the loss to prevent overfitting.

The key contributions of this work are summarized as follows.

- We analyze why existing methods for label correction may not work well with imbalanced datasets and ascribe the reasons for the discrepancy of loss distribution among classes and the ignorance of uncertainty.
- We propose a novel uncertainty-aware label correction framework for imbalanced datasets. The framework is based on a combination of epistemic uncertainty-aware class-specific noise modeling and aleatoric uncertainty-aware learning against label noise.
- Extensive experiments on synthetic and real-world noisy datasets show the effectiveness of the proposed framework, especially on imbalanced datasets.

Related work

In this section, we review important related work, including learning with noisy labels, and the methods for uncertainty estimation.

Learning with noisy labels   Recent studies on noise-robust learning with DNNs primarily fall into two categories, namely *loss correction* and *label correction*.

Loss correction methods either explicitly rectify the cost function considering the noise distribution (Sukhbaatar and Fergus 2015; Patrini et al. 2017; Hendrycks et al. 2018; Han et al. 2018a), or propose loss functions specifically designed to weaken the impact of noisy samples (Ghosh, Kumar, and Sastry 2017; Zhang and Sabuncu 2018; Ma et al.

2018). On the other side, recent research focuses more on label correction which deals with label noise by identifying and correcting noisy labels. For examples, Tanaka et al. (2018) correct labels as model predictions (soft labels) or one-hot vectors (hard labels) during training with an alternating optimization approach. Arazo et al. (2019) model label noise with a two-component mixture model on the loss to discriminate noise from clean samples, correct labels as a convex combination of noisy labels and predicted soft or hard labels, and apply MixUp (Zhang et al. 2018) to enhance performance. DivideMix (Li, Socher, and Hoi 2020) applies MixMatch (Berthelot et al. 2019) in a semi-supervised manner and exploits Co-teaching (Han et al. 2018b) to avoid confirmation bias. Nishi et al. (2021) explores different augmentations for noise modeling and learning in the label correction framework, improving results on the state of the art. SMP (Han, Luo, and Wang 2019) generates corrected labels by feature clustering and multi-prototypes selection, which is an elaborate process.

Our proposed approach falls into the *label correction* category. Although previous label correction approaches have achieved success on nearly class-balanced noisy datasets, the performance on class-imbalanced datasets has not been deeply studied. Apart from methods that focus on learning against label noise, Koziarski, Woźniak, and Krawczyk (2020) and Chen et al. (2021) propose to improve the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. 2002) under label noise. However, their primary goal is to perform better over-sampling against the class imbalance problem, so their focus is different from ours.

Uncertainty estimation   Uncertainty estimation to establish prediction confidence is vital for deep learning and has been studied thoroughly (Lakshminarayanan, Pritzel, and Blundell 2017; Gal 2016; Malinin and Gales 2018; Maddox et al. 2019). Bayesian Neural Networks (BNNs) (Gal and Ghahramani 2016) capture epistemic uncertainty in neural networks by placing a prior distribution over the model's parameters. Kendall and Gal (2017) present a framework combining input-dependent aleatoric uncertainty together with epistemic uncertainty. In addition, for applications, uncertainty has been applied to improve the performance of semi-supervised learning (SSL) and self-training, and achieves desired effects (Rizve et al. 2021; Mukherjee and Awadallah 2020).

However, learning against noisy labels on imbalanced datasets is more challenging, as it requires disambiguation of epistemic uncertainty from aleatoric uncertainty (Northcutt, Jiang, and Chuang 2021). In this paper, we address this problem and propose the ULC framework.

Perspective

As the proposed ULC falls into the label correction category, this section presents our perspective regarding the issues of existing label correction frameworks when dealing with class-imbalanced datasets. We briefly introduce the label correction framework in deep learning, discuss the issues, and present case studies.

Label Correction Frameworks Against Label Noise

Let $\mathcal{D} = (\mathcal{X}, \widetilde{\mathcal{Y}}) = \{(x_i, \widetilde{y}_i)\}_{i=1}^{N}$ be a noisy dataset with N samples, where x_i is the input and $\widetilde{y}_i$ is the associated one-hot label over C classes with possible noise. We denote the latent true label corresponding to $\widetilde{y}_i$ by y_i^* such that $\widetilde{y}_i, y_i^* \in \{0, 1\}^C$ and $\mathbf{1}^\top \widetilde{y}_i = 1, \mathbf{1}^\top y_i^* = 1$. In a more general case, there is $p(\widetilde{y}_{ij} = 1 | y_{ik}^* = 1, x_i) = \eta_{jk}^{x_i}$, where $\widetilde{y}_{ij}$ corresponds to the j-th element of $\widetilde{y}_i$ and y_{ik}^* corresponds to k-th element of y_i^*. Let $f_W(x) = v$ denote the logits vector for x, where f_W is a DNN model with trainable parameters W. The final class prediction $\widehat{y}$ is squashed by a softmax function, such that $\widehat{y} = \text{softmax}(v)$. For cross-entropy loss ℓ, which is the most commonly used loss for classification, the empirical risk is as follows:

$$\ell(W) = \frac{1}{N} \sum_{i=0}^{N} \ell_i = -\frac{1}{N} \sum_{i=0}^{N} \widetilde{y}_i^\top \log(\widehat{y}_i), \qquad (1)$$

where $\ell_i = \ell_i(W)$, and $\widehat{y}_i = \widehat{y}_i(x_i; W)$.

On account of noisy labels $\widetilde{y}_i$, the empirical risk minimization under loss ℓ leads to an ill-suited solution. Label correction methods refine/discard the observed true/corrupted labels $\widetilde{y}_i$ and learn on the corrected labels. Generally, we denote the clean samples by $x_i \in \widehat{\mathcal{X}}$ with corrected labels denoted by $y_i \in \mathcal{Y}$, and denote the rejected noisy samples that we cannot assign any labels confidently by $x_i \in \widehat{\mathcal{U}}$. Accordingly, the optimization problem concerning learning on noisy labels with label correction is formulated as semi-supervised learning (SSL), which is as follows:

$$\min_{W, \mathcal{Y}} \ell_c = \ell_x(W) + \lambda_u \ell_u(W) + \lambda_r \ell_{\text{reg}}(W)$$

$$\ell_x(W) = -\frac{1}{|\widehat{\mathcal{X}}|} \sum_{x_i \in \widehat{\mathcal{X}}} y_i^\top \log(\widehat{y}_i), \qquad (2)$$

where y_i can be initialized with $\widetilde{y}_i$, and update during model training, ℓ_u denotes any unsupervised learning loss, and ℓ_{reg} denotes any regularization term.

Alternating optimization methods have turned to be adequate to progressively search for a proper solution for $\mathcal{Y}$ and W, in which $\mathcal{Y}$ can be estimated based on the predicted probability $\widehat{y}_i$, the loss value ℓ_i or the feature representations (Tanaka et al. 2018; Li, Socher, and Hoi 2020; Han, Luo, and Wang 2019). It appears to be more robust to establish $\mathcal{Y}$ in terms of the feature representations. However, it is too complicated to select valid prototypes. Alternatively, existing methods often update $\mathcal{Y}$ based on loss values.

$$y_i^{n+1} = h(\mathcal{L}^n, \widetilde{y}_i), \qquad (3)$$

where $\mathcal{L}^n = \{\ell_i^n\}$ is the set of loss values in the n-th update and $\mathcal{Y}^{n+1} = \{y_i^{n+1}\}$ is corrected labels in the $(n+1)$-th update.

For instance, recent works (Arazo et al. 2019; Li, Socher, and Hoi 2020) conceive of the distribution of loss values $\mathcal{L}$ as a two-component Beta Mixture Model (BMM) or Gaussian Mixture Model (GMM), and then refine the labels with:

$$y_i^{n+1} = h(\mathcal{L}, \widetilde{y}_i) = p(y_i^* = \widetilde{y}_i | \mathcal{L}) \widetilde{y}_i + (1 - p(y_i^* = \widetilde{y}_i) | \mathcal{L})) \widehat{y}_i, \quad (4)$$
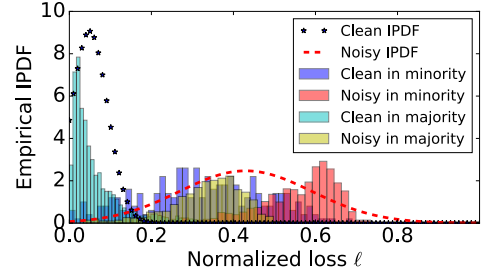

Figure 1: A binary example for empirical IPDF and estimated with class-agnostic two-component GMM over loss values on noisy CIFAR-10, warm-up of 30 epochs. Under 50% symmetric noise and 1:10 class imbalance, most of the samples in minority classes would be handled as noise. Class-agnostic label noise modeling fails due to the discrepancy between inter-class loss distributions.

where $p(y_i^* = \widetilde{y}_i | \mathcal{L})$ is the estimated probability of the observed label being true, corresponding to the posterior probability of the BMM or GMM component with the smaller mean (smaller loss).

Rethinking Label Correction Frameworks

In this section, we analyze why existing methods may fail to perform well under class imbalance. Existing noise modeling methods ignore that 1) the loss distribution for each class may vary significantly, especially on class-imbalanced data; 2) due to the existence of uncertainty, the predictive probability may fail to indicate the confidence accurately.

Inter-class loss distribution discrepancy Kang et al. (2020) decouple representations and classifiers in the learning procedure to reveal the effect of class imbalance. The work shows that the norms of the classifier layer's weights are correlated with the cardinality of the classes. When trained on imbalanced data, the inter-class loss distribution inevitably disagrees with each other. Samples in majority classes get larger logits and smaller losses, while samples in minority classes get smaller logits and larger losses.

Figure 1 illustrates the problem mentioned above by presenting a binary example, which shows the individual probability density function (IPDF) of clean and noise components modeled by GMM. The case is under 50% symmetric noise, we randomly choose two classes for demonstration, i.e., class #0 for the minority and class #3 for the majority, and they are under 1:10 imbalance. As we can see, the loss distributions of noisy samples in the majority class overlaps with clean samples in the minority class. With a two-component GMM, most of the samples with the observed label in the minority class are handled as noise. These results suggest that class-agnostic noise modeling may not work well when the loss distributions of the different classes vary significantly, which can become a typical case with imbalanced datasets.

Misleading predictions due to uncertainty Misleading predictions exist, accompanied by a high level of uncer-

(a) Distribution of loss values   (b) Joint distribution of loss values and epistemic uncertainty
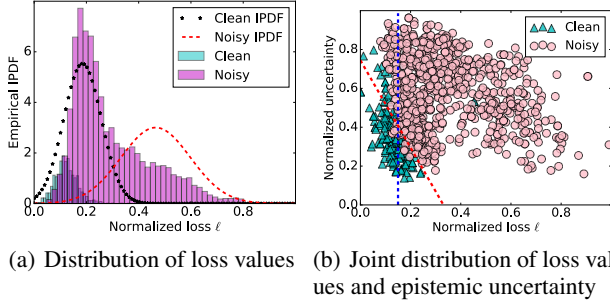
Figure 2: Loss values and epistemic uncertainty distribution on CIFAR-10 with 90% symmetric noise, a warm-up of 30 epochs. (a) Loss distribution of clean samples overlaps with noisy samples due to misleading predictions induced by epistemic uncertainty. (b) Label noise modeling considering epistemic uncertainty distinguishes clean samples from noisy ones more accurately.

tainty (Gal 2016). Learning with noisy labels needs to overcome epistemic and aleatoric uncertainty simultaneously. Moreover, epistemic uncertainty and aleatoric uncertainty are dominant in different classes and phases during the learning procedure.

On the one hand, epistemic uncertainty due to the possibilities of model structures and parameters could be reduced by observing more data (Gal and Ghahramani 2016). However, models may get underfit with minority classes, especially in the early learning phase which is critical for the effectiveness of the memorization effect. Thus epistemic uncertainty may have a more significant influence on minority classes. On the other hand, aleatoric uncertainty induced by residual label noise may lead to overfitting to noisy samples (Gal 2016), especially for the majority classes and during the late learning phase. Moreover, aleatoric uncertainty cannot be explained away simply with observing more data (Kendall and Gal 2017).

In Figure 2, we present the loss values and epistemic uncertainty distribution on CIFAR-10 with 90% symmetric noise after warming up 30 epochs. Figure 2(a) shows that if we only use the loss to distinguish clean samples, the separability will be poor, as the loss distribution of clean samples overlaps with noisy samples due to misleading predictions induced by epistemic uncertainty. While if we turn to the joint distribution of loss values and epistemic uncertainty, we can find that label noise modeling considering epistemic uncertainty distinguishes clean samples from noisy ones much more accurately, as shown in Figure 2(b).

Method

In this section, to address the issues discussed above, we propose the epistemic uncertainty-aware class-specific noise modeling method to fit the inter-class discrepancy on loss distribution. Then we formulate aleatoric uncertainty as class-dependent and instance-dependent Gaussian noise for logit corruption to prevent noise accumulation.

Epistemic Uncertainty-aware Class-specific Noise Modeling Adapted for Class-imbalanced Data

Inspired by these observations of inter-class loss distribution discrepancy and misleading predictions induced by epistemic uncertainty, we propose a robust epistemic uncertainty-aware class-specific label noise modeling method. We obtain the predicted values and epistemic uncertainty, then we integrate the results to assess whether the labels are true or corrupted.

We first estimate the integrated prediction and the epistemic uncertainty. The basic idea is that instead of taking the model's parameters as fixed values, we place a prior distribution over the parameters $p(W)$, infer the posterior $p(W|\mathcal{X}, \widetilde{\mathcal{Y}})$, and finally obtain the marginal probability $p(\widetilde{\mathcal{Y}}|\mathcal{X})$. We then can estimate the epistemic uncertainty as the entropy of $p(\widetilde{\mathcal{Y}}|\mathcal{X})$. This procedure can be processed efficiently by MC-Dropout (Gal and Ghahramani 2016). Considering $q_\theta(W)$ to be the surrogate Dropout distribution, we perform T stochastic forward passes through the network with dropout enabled for each instance x_i. Thus, with sampled model weights $\{W_t\}_{t=1}^T \sim q_\theta(W)$, the approximate integrated prediction can be obtained as $\widehat{y}_i = \frac{1}{T}\sum_{t=1}^T \text{softmax}(f(x_i, W_t))$ and the epistemic uncertainty $\epsilon_i = \epsilon(\widehat{y}_i)$ can be estimated using the entropy of predicted probability vector and then normalized[1].

Then, we try to estimate the probability of the observed label being true, i.e., the clean probability $\omega_i = p(y_i^* = \widetilde{y}_i|\ell_i, \epsilon_i)$. The posterior probability conditional on class-specific loss values and epistemic uncertainty cannot be evaluated analytically because the inference of joint distribution of loss values and epistemic uncertainty is intractable. However, we find that the posterior probability can be approximately negatively correlated to epistemic uncertainty (see Figure 2(b)). We assume that the effects of epistemic uncertainty and loss values are independent of each other, and samples with lower-level epistemic uncertainty and smaller loss values are more presumably to be clean. Thus, we empirically approximate the posterior probability as the weighted geometric mean of $1 - \epsilon_i$ and $p(y_i^* = \widetilde{y}_i|\ell_i)$:

$$\omega_i = p(y_i^* = \widetilde{y}_i|\ell_i, \epsilon_i) = (1 - \epsilon_i)^r \, p(y_i^* = \widetilde{y}_i|\ell_i)^{1-r}, \quad (5)$$

where r is the hyper-parameter to balance the influence of epistemic uncertainty and loss values.

With the uncertainty ϵ_i estimated, the problem left is to obtain $p(y_i^* = \widetilde{y}_i|\ell_i)$. Similar to Li, Socher, and Hoi (2020), this probability can be approximated by fitting a two-component mixture model on loss values. However, considering the inter-class loss distribution discrepancy (see Figure 1), we model the loss distribution separately for each class j. Let $\mathcal{L}_j = \{\ell_i|\ell_i \in \mathcal{L}, \widetilde{y}_{ij} = 1\}$ be the set of loss values of instance with observed label j. We fit a GMM on $\mathcal{L}_j$, and the mean of the two components are denoted by μ_{j0}, μ_{j1}[2]. Generally, we assume $\mu_{j0} \leq \mu_{j1}$. Hence, the probabil-

---

[1] Readers may refer to the appendix and Gal and Ghahramani (2016) for more information about uncertainty estimation.

[2] We find that BMM easily degenerates under a high-level epistemic uncertainty.
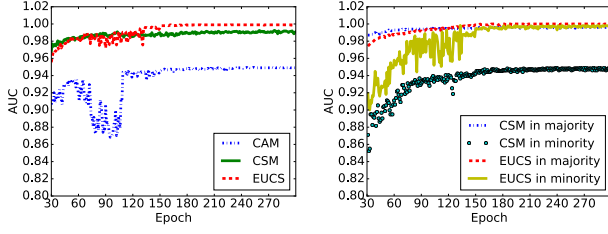
(a) Comparison of different label noise modeling methods

(b) Effects of different label noise modeling methods on minority and majority classes

Figure 3: AUC for clean/noisy sample classification on synthetic CIFAR-10 under 1:10 imbalance with 50% symmetric noise. Train for 300 epochs with different label noise modeling strategies. (a) In contrast to Class-agnostic modeling (CAM), Class-specific modeling (CSM) is more robust to class imbalance and separates clean and noisy samples more accurately under class imabalance. (b) EUCS mainly contributes to identifying noise in minority classes compared to CSM.

ity $p(y_i^* = \widetilde{y}_i | \ell_i)$ can be approximated as $p(\mu_{j0} | \ell_i)$.

Finally, we identify the clean and noisy samples by placing a threshold τ on the probabilities ω_i. We discard corrupted labels and mark the samples as unlabeled. True labels are also refined as:

$$y_i = \omega_i \widetilde{y}_i + (1 - \omega_i)\widehat{y}_i . \quad (6)$$

Figure 3 shows the effectiveness of the proposed epistemic uncertainty-aware class-specific label noise modeling (EUCS) method. We evaluate the performance on synthetic CIFAR-10 and present the area under the receiver operating characteristic curve (AUC) metrics for clean/noisy sample classification based on Eq. (5). We can find that EUCS consistently outperforms traditional class-agnostic modeling (CAM) and class-specific modeling (CSM) without considering epistemic uncertainty. Moreover, label noise modeling with consideration of epistemic uncertainty can significantly facilitate the identification of noise in minority classes.

Aleatoric Uncertainty-aware Learning Against Label Noise

Noise modeling can filter out most of the label noise. However, residual noise still exists and contributes to overfitting. To relieve misleading predictions induced by residual noise, we model latent aleatoric uncertainty in the learning procedure and introduce it as logit corruption to prevent overfitting.

Firstly, we model aleatoric uncertainty. Kendall and Gal (2017) shows that aleatoric uncertainty can be modeled by logit corruption with Gaussian noise, and the formulation leads to new loss functions, which can be interpreted as learned loss attenuation – attenuating the effect from corrupted labels and making the loss more robust to noisy data. Inspired by this, we inject instance-dependent and class-dependent noise $\delta_{\mathcal{Y}}^{x_i}$ into the logit vector. We assume that

the class-dependent and instance-dependent noisy factors in $\delta_{\mathcal{Y}}^{x_i}$ are independent of each other. Let δ^{x_i} be a diagonal matrix, we refer to it as the instance-dependent noise factor. Let $\delta^{\mathcal{Y}}$ be a stochastic matrix, such that $\sum_i \delta_{ij}^{\mathcal{Y}} = 1$, we refer it as the class-dependent noise factor. The induced logit corruption process can be formulated as:

$$\widehat{v}_i(W) = \delta^{x_i}(W) + \delta^{\mathcal{Y}} f_W(x_i) . \quad (7)$$

We approximate $\delta^{\mathcal{Y}}$ by $I + \delta$. In addition, we assume a Gaussian distribution over δ^{x_i} and δ with variance σ^{x_i} and σ, i.e., $\delta^{x_i} \sim \mathcal{N}(0, \sigma^{x_i})$, $\delta \sim \mathcal{N}(0, \sigma)$. σ^{x_i} and σ can be viewed as a measure of aleatoric uncertainty. The corrupted logits vector is squashed with the softmax function to obtain $\widehat{y}_i$:

$$\widehat{y}_i = \text{softmax}((I + \delta)f_W(W) + \delta^{x_i}(W)) \quad (8)$$

Then we estimate σ^{x_i} and σ, and learn with loss attenuation. We attach an affiliated branch to the head of the network to predict σ^{x_i}. The reparametrization trick (Blum, Haghtalab, and Procaccia 2015) is used to obtain a minibatch-based Monte Carlo estimator of ℓ_x. The numerically stable stochastic loss for cleaned samples is given by:

$$\ell_x(W) = -\frac{1}{|\widehat{\mathcal{X}}|} \sum_{x_i \in \widehat{\mathcal{X}}} (y_i)^\top \log \frac{1}{T} \sum_{t=1}^{T} (\widehat{y}_{it}(x_i; W, \sigma^{x_i}, \sigma)) \quad (9)$$

with $\widehat{y}_i^{(t)}$ the softmax ouput of the t-th sampling over logits conditional on σ^{x_i} and σ. The extra computation cost is marginal since that σ is a $C \times C$ matrix and σ^{x_i} is actually a $C \times 1$ vector and sampling on them is extremely fast.

Implementation Details

The algorithm procedure follows label correction frameworks similar to DivideMix (Li, Socher, and Hoi 2020), and we add the proposed epistemic uncertainty-aware class-specific noise modeling and aleatoric uncertainty-aware learning.

We conventionally warm up the model for a few epochs using cross-entropy loss with an entropy term, then the learning procedure alternates between two phases. First, as introduced before, we estimate the clean probabilities ω_i and the corrected labels y_i by Eq. (5) and (6). With a threshold τ on $\{\omega_i\}$, we divide the training data into clean samples with corrected labels $\widehat{\mathcal{X}}$ and noisy samples without labels $\widehat{\mathcal{U}}$. Following DivideMix, Co-teaching is used to alleviate the sample selection bias. Next, semi-supervised learning (SSL) is performed based on $\widehat{\mathcal{X}}$ and $\widehat{\mathcal{U}}$. Also following DivideMix, we use MixMatch (Zhang et al. 2018) to augment $\widehat{\mathcal{X}} \cup \widehat{\mathcal{U}}$ into $\mathcal{X}' \cup \mathcal{U}'$. During the SSL phase, as discussed before, we sample on logits conditional on aleatoric uncertainty δ^x and δ, and introduce the aleatoric uncertainty as logit corruption to prevent overfitting by Eq. (9). Accordingly, the total stochastic loss in Eq. (2) can be rewritten as:

$$\ell_c = \ell_x + \lambda_u \ell_u$$

$$\ell_x = -\frac{1}{|\mathcal{X}'|} \sum_{x_i, y_i \in \mathcal{X}'} (y_i)^\top \log \frac{1}{T} \sum_{t=1}^{T} (\widehat{y}_{it}(x_i; W, \sigma^{x_i}, \sigma))$$

$$\ell_u = \frac{1}{|\mathcal{U}'|} \sum_{x_i, y_i \in \mathcal{U}'} \left\| y_i - \frac{1}{T} \sum_{t=1}^{T} (\widehat{y}_{it}(x_i; W, \sigma^{x_i}, \sigma)) \right\|_2^2 .$$

(10)

The additional regularization term ℓ_{reg} in Eq. (2) used in prior label correction works (Tanaka et al. 2018; Arazo et al. 2019; Li, Socher, and Hoi 2020) imposing a uniform prior distribution assumption on the prior probabilities is proven to be imperative to prevent the assignment of all labels to a single class. However, we find it inessential in our method.

In the appendix, Algorithm 1 outlines the uncertainty-aware label correction process, and we summarize the notations in Table 5.

Experiments

In this section, we report the experimental results. We introduce the experimental setup, report the performance of ULC and baselines, and also analyze the results of ablation experiments.

Experimental Setup

Datasets We perform extensive evaluations on five datasets: CIFAR-10, CIFAR-100, class-imbalanced CIFAR-10, class-imbalanced CIFAR-100, and Clothing1M. The noise injection methods for CIFAR-10 and CIFAR-100 follow the previous work (Li, Socher, and Hoi 2020). We experiment with two types of label noise: symmetric and asymmetric. To test the robustness of our approach to class imbalance, we resample the training set to construct class-imbalanced CIFAR-10 and CIFAR-100. Specifically, we randomly choose half the classes and randomly sub-sample 1/5 and 1/10 examples in these classes while other classes remain the same. Note that we only resample the training sets for class imbalance settings, i.e., the test sets remain the same as class balance settings. Then, labels in class-imbalanced CIFAR get flipped to the rest of the categories with the same probability. Clothing1M is a large-scale dataset with real-world noisy labels. In our experiment, clean training data is not used.

Experimental settings For CIFAR experiments, we use the PreAct ResNet-18 (He et al. 2016) which is commonly used to benchmark label noise learning methods (Li, Socher, and Hoi 2020). We train the network using SGD with a momentum of 0.9 for 300 epochs; warm-up 30 epochs for CIFAR-10 and 40 epochs for CIFAR-100. In the Clothing1M experiments, we use ResNet-50 with ImageNet pretrained weights, following the previous work (Li, Socher, and Hoi 2020). The warm-up period is 1 epoch for Clothing1M. τ is set as 0.6 for 90% noise ratio and 0.5 for others. λ_u is validated from $\{0, 25, 50, 150\}$. Generally, the hyperparameters setting for MixMatch is inherited from DivideMix without heavily tuning, because the SSL part is not our focus and can be replaced by other alternatives. We leverage MC-dropout (Gal and Ghahramani 2016) to estimate uncertainty, setting T to 10 and the dropout rate to 0.3. The uncertainty ratio r is set as 0.1 to obtain the final clean probability. We model aleatoric uncertainty for class-dependent and instance-dependent noise by 10 Monte Carlo samples on logits.

Baselines We compare our method (ULC) with multiple state-of-art approaches as follows: Co-teaching+ (CoT+) (Yu et al. 2019), PENCIL (Yi and

Dataset		CIFAR-10 Sym.				Asym.	CIFAR-100 Sym.			
Noise Rate		20%	50%	80%	90%	40%	20%	50%	80%	90%
CE	best	86.8	79.4	62.9	42.7	85.0	62.0	46.7	19.9	19.9
	last	82.7	57.9	26.1	16.8	72.3	61.8	37.3	8.8	3.5
CoT+	best	89.5	85.7	67.4	47.9	-	65.6	51.8	27.9	13.7
	last	88.2	84.1	45.5	30.1	-	64.1	45.3	15.5	8.8
PENCIL	best	92.4	89.1	77.5	58.9	88.5	69.4	57.5	31.1	15.3
	last	92.0	88.7	76.5	58.2	88.1	68.1	56.4	20.7	8.8
ML	best	92.9	89.3	77.4	58.7	89.2	68.5	59.2	42.4	19.5
	last	92.0	88.8	76.1	58.3	88.6	67.7	58.0	40.1	14.3
M-correction	best	94.0	92.0	86.8	69.1	87.4	73.9	66.1	48.2	24.3
	last	93.8	91.9	86.6	68.7	86.3	73.4	65.4	47.6	20.5
DivideMix	best	**96.1**	94.6	93.2	76.0	93.4	77.3	74.6	60.2	31.5
	last	95.7	94.4	92.9	75.4	92.1	76.9	74.2	59.6	31.0
ELR+	best	95.8	94.8	93.3	78.7	93.0	**77.6**	73.6	60.8	33.4
	last	94.6	93.8	91.1	75.2	92.7	**77.5**	72.4	60.2	30.8
ULC	best	**96.1**	**95.2**	**94.0**	**86.4**	**94.6**	77.3	**74.9**	**61.2**	**34.5**
	last	**95.9**	**94.7**	**93.2**	**85.8**	**94.1**	77.1	**74.3**	**60.8**	**34.1**

Table 1: Comparison with state-of-the-art methods in test accuracy on balanced CIFAR (%).

Wu 2019), Meta-Learning (ML) (Li et al. 2019), M-correction (Arazo et al. 2019), DivideMix (Li, Socher, and Hoi 2020), and ELR+ (Liu et al. 2020). All methods use the same architecture (PreAct ResNet-18 for CIFAR and ResNet-50 with ImageNet pretrained weights for Clothing1M). A brief introduction to the baselines is in the appendix.

Experimental Results

We conduct experiments to verify the effectiveness of our approach and the baselines on the general class-balanced and class-imbalanced noisy data. The results of baselines on class-balanced CIFAR and Clothing1M are taken from Li, Socher, and Hoi (2020) and Liu et al. (2020), and results on class-imbalanced CIFAR are obtained using publicly available code. We put more experimental results, including the results for robustness to hyperparameters and the choice of SSL methods, to the appendix.

Class-balanced synthetic CIFAR Table 1 shows a comparison among different methods on class-balanced CIFAR-10 and CIFAR-100. The noise rate settings are the same with Li, Socher, and Hoi (2020). We report both the best (indicating how well these methods learn from noisy samples) and the last (indicating whether finally overfit) test accuracy following Li, Socher, and Hoi (2020). ULC achieves comparable performance for low-level symmetric label noise, and outperforms state-of-the-art methods for high-level and asymmetric label noise.

Class-imbalanced synthetic CIFAR Table 2 shows the results on class-imbalanced CIFAR-10 and CIFAR-100 with different levels of label noise. Note that the test sets here remain class-balanced, so the results can be compared with the class-balanced setting. ULC consistently outperforms state-of-the-art methods across different imbalance ratios with different types of label noise. Class-imbalance challenges the task of learning with noisy labels and all these previous methods work cannot perform well in this scenario. However, ULC shows robustness to this challenging scenario and

Dataset		CIFAR-10				CIFAR-100			
		Sym. 20%		Sym. 50%		Sym.20%		Sym.50%	
Resampling Ratio		1:5	1:10	1:5	1:10	1:5	1:10	1:5	1:10
CE	best	87.1	85.4	72.0	68.4	64.9	61.4	48.2	43.0
	last	86.8	85.0	70.9	67.8	64.5	60.9	47.7	42.2
CoT+	best	82.5	76.7	71.9	53.2	49.2	39.9	30.6	29.2
	last	82.3	76.7	71.5	50.9	49.0	39.6	30.5	29.2
PENCIL	best	80.5	74.5	73.8	65.4	52.1	45.5	33.2	28.4
	last	80.0	73.0	73.5	65.2	51.1	43.3	31.9	25.2
ML	best	75.9	70.0	73.8	62.2	54.2	47.5	41.2	34.8
	last	74.6	68.0	61.6	54.0	51.1	44.4	35.2	30.3
M-correction	best	88.1	80.1	83.5	77.5	62.3	55.4	51.3	44.4
	last	87.0	77.3	83.1	76.8	62.2	54.7	50.0	44.1
DivideMix	best	93.9	74.8	85.5	66.9	65.0	51.2	56.9	44.2
	last	93.9	74.6	85.4	66.8	64.9	50.9	56.4	44.0
ELR+	best	88.2	79.8	82.7	65.5	59.6	49.9	53.4	44.7
	last	87.0	78.4	82.5	64.6	59.3	49.6	52.5	43.9
ULC	best	**95.0**	**93.8**	**94.9**	**92.5**	**75.5**	**72.8**	**71.6**	**57.2**
	last	**94.9**	**92.6**	**94.7**	**92.2**	**75.1**	**72.3**	**70.9**	**56.7**

Table 2: Comparison with state-of-the-art methods in test accuracy on imbalanced CIFAR (%).

Method	Test Accuracy
CE	69.2
M-correction	71.0
Meta-Learning	73.5
PENCIL	73.5
DivideMix	**74.8**
ELR+	**74.8**
ULC	**74.9±0.2**

Table 3: Comparison with state-of-the-art methods in test accuracy on Clothing1M (%).

works well even with only a few clean examples in minority classes (25 clean examples with 50% symmetric noise under 1:10 class imbalance for CIFAR-100).

Clothing1M Table 3 shows the results on the real-world noisy dataset Clothing1M. We report average accuracy as well as the standard deviation of five runs. ULC achieves comparable performance to state-of-the-art methods include DivideMix and ELR+. The insignificant improvement may be due to two reasons. Firstly, the estimated noise rate of Clothing1M is only 38.5% (Liu et al. 2020), secondly, the training set actually used for Clothing1M is nearly class-balanced (Li, Socher, and Hoi 2020).

Ablation Experiments

Here, we study the individual components and their influence. Table 4 summarizes the ablation study results. As we can see, each element provides an independent performance improvement.

Class-specific label noise modeling Based on the above analysis, inter-class loss distribution discrepancy is the key to the degeneration of previous methods. As expected, the performance training without class-specific modeling rapidly breaks down with an increasing resampling ratio. Besides, we find that the influence of inter-class loss dis-

Dataset		CIFAR-10				CIFAR-100			
		Sym. 20%		Sym. 50%		Sym.20%		Sym.50%	
Resampling Ratio		1:5	1:10	1:5	1:10	1:5	1:10	1:5	1:10
ULC	best	**95.0**	**93.8**	**94.9**	**92.5**	**75.5**	**72.8**	**71.6**	**57.2**
	last	**94.9**	**92.6**	**94.7**	**92.2**	**75.1**	**72.3**	**70.9**	**56.7**
ULC w/o CSM	best	94.2	82.5	88.5	70.5	67.1	59.4	61.3	50.2
	last	93.8	81.3	87.9	69.8	66.2	58.7	60.5	49.8
ULC w/o EUM	best	94.1	90.7	91.3	89.9	68.0	67.2	65.9	53.2
	last	93.8	90.1	90.8	88.7	67.5	66.9	65.4	52.6
ULC w/o AUL	best	94.2	91.8	92.5	91.2	70.2	66.5	68.3	52.2
	last	94.0	91.5	91.7	90.5	69.7	66.1	67.2	51.6

Table 4: Ablation study results on imbalanced CIFAR (%). CSM refers to class-specific noise modeling, EUM refers to epistemic uncertainty-aware noise modeling, and AUL refers to aleatoric uncertainty-aware learning.

tribution discrepancy correlates to noise ratio. The performance deterioration training without class-specific modeling under 1:5 resampling with 20% symmetric noise is modest, while it is very significant for 50% symmetric noise.

Epistemic uncertainty-aware label noise modeling The performance consistently decreases without consideration of the effect of epistemic uncertainty on loss modeling, especially when the noise rate is high, or when the class imbalance property gets worse, as epistemic uncertainty-aware label noise modeling can provide a more reliable criterion for noisy sample discovery under series class imbalance or label noise.

Aleatoric uncertainty-aware learning When noise ratio and resampling ratio are both high-level, learning with aleatoric uncertainty is significantly beneficial to performance improvements. As noise modeling cannot completely remove label noise, and the network may still overfit residual noise. This overfitting process is irreversible and may in turn affect the identification of label noise. Learning with aleatoric uncertainty can alleviate final overfitting.

Conclusion

In this work, we propose an uncertainty-aware label correction framework (ULC) to adapt for learning with noisy labels on class-imbalanced data. We mainly focused on overcoming inter-class loss distribution discrepancy and alleviating the effect of epistemic uncertainty and aleatoric uncertainty. We model label noise incorporating class-specific loss distribution with estimated epistemic uncertainty to identify more confident clean samples from noisy samples. In the process of learning with corrected labels, we formulate aleatoric uncertainty induced by residual label noise as logit corruption into loss to alleviate overfitting in majority classes and during the late learning phase. Experiments show that ULC yields strong results on different benchmarks despite the increasing noise ratio and resampling ratio. In the future, we will investigate the facilitation effect of uncertainty on learning with open-set label noise, considering that out-of-domain data can be viewed as a reason leading to epistemic uncertainty.

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2019. Unsupervised Label Noise Modeling and Loss Correction. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 312–321.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A. C.; Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, 233–242.

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*, 5050–5060.

Blum, A.; Haghtalab, N.; and Procaccia, A. D. 2015. Variational Dropout and the Local Reparameterization Trick. In *Proceedings of the 29th International Conference on Neural Information Processing Systems (NeurIPS 2015)*, 2575–2583.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chen, B.; Xia, S.; Chen, Z.; Wang, B.; and Wang, G. 2021. RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise. *Information Sciences*, 553: 397–428.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. ImageNet: A Large-scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 248–255.

Fergus, R.; Fei-Fei, L.; Perona, P.; and Zisserman, A. 2010. Learning Object Categories from Internet Image Searches. *Proceedings of the IEEE*, 98(8): 1453–1466.

Frénay, B.; and Verleysen, M. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5): 845–869.

Gal, Y. 2016. Uncertainty in Deep Learning. *University of Cambridge*, 1(3): 4.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, 1050–1059.

Ghosh, A.; Kumar, H.; and Sastry, P. S. 2017. Robust Loss Functions under Label Noise for Deep Neural Networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI 2017)*, 1919–1925.

Han, B.; Yao, J.; Niu, G.; Zhou, M.; Tsang, I. W.; Zhang, Y.; and Sugiyama, M. 2018a. Masking: A New Perspective of Noisy Supervision. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, 5841–5851.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018b. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, 8536–8546.

Han, J.; Luo, P.; and Wang, X. 2019. Deep Self-Learning From Noisy Labels. In *Proceddings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, 5137–5146.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. In *Proceddings of the 14th European Conference on Computer Vision (ECCV 2016)*, 630–645.

Hendrycks, D.; Mazeika, M.; Wilson, D.; and Gimpel, K. 2018. Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, 10477–10486.

Japkowicz, N.; and Stephen, S. 2002. The Class Imbalance Problem: A Systematic Study. *Intelligent data analysis*, 6(5): 429–449.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *The 8th International Conference on Learning Representations (ICLR 2020)*.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, 5574–5584.

Konstantinov, N.; and Lampert, C. 2019. Robust Learning from Untrusted Sources. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 3488–3498.

Koziarski, M.; Woźniak, M.; and Krawczyk, B. 2020. Combined Cleaning and Resampling Algorithm for Multi-class Imbalanced Data with Label Noise. *Knowledge-Based Systems*, 204: 106223.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 128(7).

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, 6402–6413.

Li, J.; Socher, R.; and Hoi, S. C. H. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *The 8th International Conference on Learning Representations (ICLR 2020)*.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2019. Learning to Learn From Noisy Labeled Data. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, 5051–5059.

Li, M.; Soltanolkotabi, M.; and Oymak, S. 2020. Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks. In *Proceddings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 4313–4324.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-Learning Regularization Prevents Memorization of Noisy Labels. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, 20331–20342.

Ma, X.; Wang, Y.; Houle, M. E.; Zhou, S.; Erfani, S. M.; Xia, S.; Wijewickrema, S. N. R.; and Bailey, J. 2018. Dimensionality-Driven Learning with Noisy Labels. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 3361–3370.

Maddox, W. J.; Izmailov, P.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2019. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*, 13132–13143.

Malinin, A.; and Gales, M. J. F. 2018. Predictive Uncertainty Estimation via Prior Networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, 7047–7058.

Mukherjee, S.; and Awadallah, A. H. 2020. Uncertainty-aware Self-training for Few-shot Text Classification. In *Proceedings of the 24th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, 21199–21212.

Nishi, K.; Ding, Y.; Rich, A.; and Hollerer, T. 2021. Augmentation strategies for learning with noisy labels. In *Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, 8022–8031.

Northcutt, C.; Jiang, L.; and Chuang, I. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70: 1373–1411.

Patrini, G.; Rozza, A.; Menon, A. K.; Nock, R.; and Qu, L. 2017. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2233–2241.

Reed, S. E.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2015. Training Deep Neural Networks on Noisy Labels with Bootstrapping. In *The 3rd International Conference on Learning Representations (ICLR 2015)*.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *The 9th International Conference on Learning Representations (ICLR 2021)*.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the 24th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, 596–608.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1): 1929–1958.

Sukhbaatar, S.; and Fergus, R. 2015. Learning from Noisy Labels with Deep Neural Networks. In *The 3rd International Conference on Learning Representations (ICLR 2015)*.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint Optimization Framework for Learning With Noisy Labels. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 5552–5560.

Welinder, P.; Branson, S.; Belongie, S. J.; and Perona, P. 2010. The Multidimensional Wisdom of Crowds. In *Proceedings of the 24th International Conference on Neural Information Processing Systems (NeurIPS 2010)*, 2424–2432.

Yi, K.; and Wu, J. 2019. Probabilistic End-To-End Noise Correction for Learning with Noisy Labels. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, 7017–7025.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I. W.; and Sugiyama, M. 2019. How does Disagreement Help Generalization against Label Corruption? In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 7164–7173.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. Mixup: Beyond Empirical Risk Minimization. In *The 6th International Conference on Learning Representations (ICLR 2018)*.

Zhang, Z.; and Sabuncu, M. R. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, 8792–8802.