

Is your noise correction noisy?

PLS: Robustness to label noise with two stage detection

Paul Albert, Eric Arazo, Tarun Krishna, Noel E. O’Connor, Kevin McGuinness
 School of Electronic Engineering,
 Insight SFI Centre for Data Analytics, Dublin City University (DCU)
 paul.albert@insight-centre.org

Abstract

Designing robust algorithms capable of training accurate neural networks on uncurated datasets from the web has been the subject of much research as it reduces the need for time consuming human labor. The focus of many previous research contributions has been on the detection of different types of label noise; however, this paper proposes to improve the correction accuracy of noisy samples once they have been detected. In many state-of-the-art contributions, a two phase approach is adopted where the noisy samples are detected before guessing a corrected pseudo-label in a semi-supervised fashion. The guessed pseudo-labels are then used in the supervised objective without ensuring that the label guess is likely to be correct. This can lead to confirmation bias, which reduces the noise robustness. Here we propose the pseudo-loss, a simple metric that we find to be strongly correlated with pseudo-label correctness on noisy samples. Using the pseudo-loss, we dynamically down weight under-confident pseudo-labels throughout training to avoid confirmation bias and improve the network accuracy. We additionally propose to use a confidence guided contrastive objective that learns robust representation on an interpolated objective between class bound (supervised) for confidently corrected samples and unsupervised representation for under-confident label corrections. Experiments demonstrate the state-of-the-art performance of our Pseudo-Loss Selection (PLS) algorithm on a variety of benchmark datasets including curated data synthetically corrupted with in-distribution and out-of-distribution noise, and two real world web noise datasets. Our experiments are fully reproducible github.com/PaulAlbert31/PLS.

1. Introduction

Standard supervised datasets for image classification using deep learning [15, 7, 20, 14] are constituted by large amounts of images gathered from the web which have been

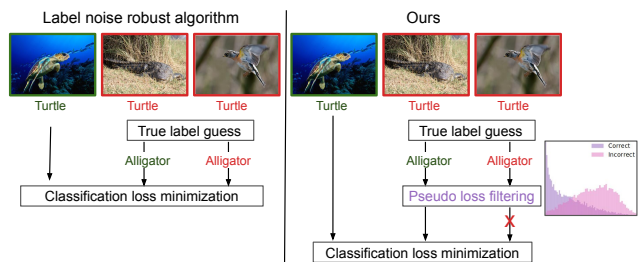


Figure 1. Two stage label noise mitigation on detected noisy samples. Contrary to state-of-the-art label noise robust algorithms, we filter out incorrect pseudo-labels using the pseudo-loss to avoid confirmation bias on incorrect corrections.

heavily curated by multiple human annotators. In this paper, we propose to devise an algorithm which aims to train an accurate classification network on a web crawled dataset [19, 32] where the human curation process was skipped. By doing so, the dataset creation time is greatly reduced but label noise becomes an issue [2] and can greatly degrade the classification accuracy [42]. To counter the effect of noisy annotations, previous contributions have focused on detecting the noisy samples using the natural robustness of deep learning architectures to noise in early training stages [3, 4]. These algorithms will identify noisy samples because they tend to be learned slower than their clean counterpart [17], because of incoherences with the labels of close neighbors in the feature space [23, 18], a confident prediction from the neural net in a different class than the target class [38, 21], inconsistent predictions across iterations [22, 34], and more. Once the noisy samples are identified, a corrected label is produced, yet ensuring that labels are correctly guessed is less studied in the label noise literature. Some propositions inspired by semi-supervised learning [28, 41] have been made recently by Li *et al.* [18] where only pseudo-labels whose value in the max softmax bin (confidence) is superior to a hyper-parameter threshold value are kept or by Song *et al.* [29] where low entropy predictions indicate a confident pseudo-label. This paper proposes to

focus on the correction of noisy samples once they have been detected. We specifically propose a novel metric, the pseudo-loss, which is able to retrieve correctly guessed pseudo-labels and that we show to be superior to the pseudo-label confidence previously used in the semi-supervised literature. We find that incorrectly guessed pseudo-labels are especially damaging to the supervised contrastive objectives that have been used in recent contributions [23, 1, 18]. We propose an interpolated contrastive objective between class-conditional (supervised) for the clean or correctly corrected samples, where we encourage the network to learn similar representation for images belonging to the same class; and an unsupervised objective for the incorrectly corrected noise. This results in `Pseudo-Loss Selection (PLS)` a two-stage noise detection algorithm where the first stage detects all noisy samples in the dataset while the second stage removes incorrect corrections. We then train a neural network to jointly minimize a classification and a supervised contrastive objective. We design PLS on synthetically corrupted datasets and validate our findings on two real world noisy web crawled datasets. Figure 1 illustrates our proposed improvement to label noise robust algorithms. Our contributions are:

- A two-stage noise detection using a novel metric where we ensure that the corrected targets for noisy samples are accurate;
- A novel softly interpolated confidence guided contrastive loss term between supervised and unsupervised objective to learn robust features from all images;
- Extensive experiments of synthetically corrupted and web-crawled noisy datasets to demonstrate the performance of our algorithm.

2. Related work

Label noise robust algorithms

Label noise in web crawled datasets has been evidenced to be a mixture between in-distribution (ID) noise and out-of-distribution (OOD) noise [2]. In-distribution noise denotes an image that was assigned an incorrect label but can be corrected to another label in the label distribution. Out-of-distribution noise are images whose true label lie outside of the label distribution and cannot be directly corrected. While some algorithms have been designed to detect ID and OOD separately, others reach good results by assuming all noise is ID. The rest of this section will introduce state-of-the-art approaches to detect and correct noisy samples.

2.1. Label noise detection

Label noise in datasets can be detected by exacerbating the natural resistance of neural networks to noise. Small loss algorithms [3, 17, 22] observe that noisy samples tend

to be learned slower than their clean counterparts and that a bi-modal distribution can be observed in the training loss where noisy samples belong to the high loss mode. A mixture model is then fit to the loss distribution to retrieve the two modes in an unsupervised manner. Other approaches evaluate the neighbor coherence in the network feature space where images are expected to have many neighbors from the same class [23, 18, 25] and a hyper-parameter threshold is used on the number of neighbors from the same class to allow to identify the noisy samples. In some cases, a separate OOD detection can be performed to differentiate between correctable ID noise and uncorrectable OOD samples. OOD samples are detected by evaluating the uncertainty of the current neural network prediction. EvidentialMix [24] uses the evidential loss [26], JoSRC evaluates the Jensen-Shannon divergence between predictions [38], and DSOS [2] computes the collision entropy. An alternative approach is to use a clean subset to learn to detect label noise in a meta-learning fashion [36, 10, 35, 37] but we will assume in this paper that a trusted set is unavailable.

2.2. Noise correction

Once the noisy samples have been detected, state-of-the-art approaches guess true labels using current knowledge learned by the network. Options include guessing using the prediction of the network on unaugmented samples [3, 21], semi-supervised learning [17, 23], or neighboring samples in the feature space [18]. Some approaches also simply discard the detected noisy examples to train on the clean data alone [11, 12, 27, 40]. In the case where a separate out-of-distribution detection is performed, the samples can either be removed from the dataset [24], assigned a uniform label distribution over the classes to promote rejection by the network [38, 2], or used in an unsupervised objective [1].

2.3. Noise regularization

Another strategy when training on label noise datasets is to use strong regularization either in the form of data augmentation such as mixup [43] or using a dedicated loss term [21]. Unsupervised regularization has also shown to help improve the classification accuracy of neural networks trained on label noise datasets [18, 30].

3. PLS

We consider an image dataset $\mathcal{X} = \{x_i\}_{i=1}^N$ associated with one-hot encoded classification labels \mathcal{Y} over C classes. An unknown percentage of labels in $\mathcal{Y} = \{y_i\}_{i=1}^N$ are noisy, *i.e.* y_i is different from the true label of x_i . We aim to train a neural network ϕ on the imperfect label noise dataset to perform accurate classification on a held out test set.

3.1. Detecting the noisy samples

Our contributions do not include detecting the noisy labels but we propose to focus here on improving the correction of the noisy samples once they have been detected. We use a known phenomenon in previous research for label noise classification [3, 17, 22] where in early stages of training, the cross-entropy loss between ϕ 's prediction on an unaugmented view on an image $\phi(x_i)$ and the associated (possibly noisy) ground-truth label y_i is observed to separate into a low loss clean mode and high loss noisy mode. We therefore propose to fit a Gaussian Mixture Model (GMM) to the training loss to retrieve each mode in an unsupervised fashion. Clean samples are finally identified as belonging to the low loss mode with a probability superior to a threshold $t = 0.95$. Alternative metrics have been proposed to retrieve noisy labels but we find that while approaches retrieve noisy samples very similarly for synthetic noise, the training loss is more accurate in the case of real world noise. We justify this statement in Section 4.2.

3.2. Confident correction of noisy labels

3.2.1 Guessing labels for detected noisy samples

To guess the true label of detected noisy samples, we propose to use a consistency regularization approach. Given an image x_i associated to a noisy label, we produce two weakly augmented views x_{i1} and x_{i2} . Weak augmentations are random cropping after zero-padding and random horizontal flipping. Using the current state of ϕ , we guess the pseudo-label \hat{y}_i as

$$\hat{y}_i = \left(\frac{\phi(x_{i1}) + \phi(x_{i2})}{2} \right)^\gamma, \quad (1)$$

with $\gamma = 2$ being a temperature hyper-parameter. We then apply a max normalization over \hat{y}_i to ensure that the values of the pseudo-label are between 0 and 1.

3.2.2 Correcting only confident pseudo-labels

We propose to only correct those pseudo-labels that are likely to be correctly guessed by ϕ . This solution has already been explored in the semi-supervised literature [28, 41] where pseudo-labels are only kept if the value of the maximum probability is superior to an hyperparameter threshold. Both prediction confidence measured by highest probability bin [18] or prediction entropy [29] has also been successfully applied in the label noise literature. We propose to identify correct pseudo-labels by evaluating a different metric, which we name the pseudo-loss. The pseudo-loss evaluates the cross-entropy loss between the pseudo-label \hat{y}_i and the prediction of the model on an unaugmented view $\phi(x_i)$:

$$l_{\text{pseudo}} = -\hat{y}_i \log \phi(x_i). \quad (2)$$

We observe that much like the noise detection loss in Section 3.1, the pseudo-loss is bi-modal (see Figure 1 and Section 4.3). We propose to fit a second GMM to the pseudo-loss and to use the posterior probability of a sample to belong to the low l_{pseudo} mode (correct pseudo-label, left-most gaussian) as w , a weight in the classification loss l_{classif} that reduces the impact of incorrect pseudo-labels. Underconfident, high pseudo-loss samples are weighed with values close to 0 (low probability of belonging to the low pseudo-loss mode) while confident pseudo-labels are weighed with values close to 1 (high probability of belonging to the low pseudo-loss mode). The classification loss we use is a weighed cross-entropy with mixup:

$$l_{\text{classif}} = \frac{1}{\sum_{i=1}^N w_{\text{mix},i}} \sum_{i=1}^N -w_{\text{mix},i} \hat{y}_{\text{mix},i} \log \phi(x_{\text{mix},i}), \quad (3)$$

where w_{mix} , x_{mix} and y_{mix} are linearly interpolated with another random sample in the mini-batch using parameter $\lambda \sim \mathcal{U}(0, 1)$, sampled for every mini-batch (mixup [43]). We evaluate how the pseudo-loss compares to pseudo-label confidence in Section 4.3.

3.2.3 Supervised contrastive learning

To improve the quality of representations learned by ϕ , we propose to train a supervised contrastive objective jointly with the classification loss. We compute the contrastive features as a linear projection g from the classification features to the L_2 normalized contrastive space. A contrastive objective aims to learn similar contrastive features for images belonging to the same class. Given a training mini-batch of images X_b with associated classification labels Y_b , we produce a weakly augmented view X_{b1} and a strongly augmented view X'_b . The strong augmentations are the SimCLR augs [5]: random resized crop, color jitter, random grayscale, and random horizontal flipping. We compute the label similarity matrix $L = Y_b Y_b^t$ and the feature similarity matrix:

$$P = \frac{g(\phi(X_{i1}))g(\phi(X'_i))^T}{\mu}, \quad (4)$$

with $\mu = 0.2$ being a temperature scaling parameter. Both P and L are $B \times B$ matrices with B the mini-batch size. The contrastive loss is the row-wise cross-entropy loss:

$$l_{\text{naivecont}} = \frac{1}{B} \sum_{i=1}^B -\frac{L_i \log P_i}{\sum_{c=1}^C L_{i,c}}, \quad (5)$$

where L_i and P_i denote the row i of the corresponding matrix. Because label noise is present in the datasets we train on, minimizing $l_{\text{naivecont}}$ directly is detrimental since similarities will be enforced between samples whose pseudo-label cannot be trusted. We propose instead to account for

pseudo-label incorrectness and train a confidence guided contrastive objective.

3.2.4 Confidence guided contrastive learning in the presence of label noise

So far the proposed confidence guided contrastive objective does not account for label noise in the dataset which conflicts with the classification objective and will be harmful to the learned representation. The first step to account for label noise is to replace the correctly guessed labels in Section 3.2.1 for the detected noisy samples to produce \hat{Y}_b . As we do for the classification loss, we also want to prevent incorrect pseudo-labels from interfering with the contrastive algorithm. Rather than simply weighting the contrastive loss for a low confidence pseudo-labeled noisy sample, we propose to use the unsupervised capabilities of contrastive losses. Depending on the confidence in a pseudo-label for a noisy sample, the pseudo-label will be used to enforce similar features with other samples from the same guessed class (high pseudo-label confidence) or will only be encouraged to learn similar features between augmented views of the same image (low confidence pseudo-label). To do so in a continuous manner, without the need for a threshold on w , we modify the initial classification labels \hat{Y}_b by concatenating a weighted one-hot positional encoding of samples in the mini-batch with \hat{Y}_b using w . The label for sample i in the mini-batch becomes:

$$y_{\text{cont},i} = \text{concat}(w_i \times \hat{y}_i, (1 - w_i) \times \mathcal{O}(i, B)), \quad (6)$$

where $\mathcal{O}(i, B)$, the one-hot positional encoding of the sample i in the mini-batch is a zero-vector of size B with value 1 at position i with B the mini-batch size. An illustration for computing y_{cont} is available in the supplementary material. Repeating the process, we create $Y_{\text{cont},b}$ of size $B \times (C+B)$. Finally, to benefit from the noise robustness of mixup in the contrastive objective, we adopt a similar setting as in iMix [16] and linearly interpolate X_{i1} among samples in the mini-batch to create $X_{\text{mix},i}$ (InputMix) before extracting the features as well as the corresponding label $Y_{\text{cont},b}$ to create $Y_{\text{mix},b}$ using $\beta \sim \mathcal{U}(0, 1)$. To compute the confidence guided contrastive objective, we use $L' = Y_{\text{mix},b} Y_{\text{mix},b}^t$ and $P' = g(\phi(X_{\text{mix},i}))g(\phi(X'_i))^T / \mu$. The confidence guided, noise robust contrastive learning loss we minimize is

$$l_{\text{cont}} = \frac{1}{B} \sum_{i=1}^B - \frac{L'_i \log P'_i}{\sum_{c=1}^C L'_{i,c}}, \quad (7)$$

The final training objective we optimize is:

$$l = l_{\text{classif}} + l_{\text{cont}}. \quad (8)$$

4. Experiments

4.1. Setup

We conduct noise robustness experiments on four image datasets. For synthetically corrupted datasets, we train on CIFAR-100 and miniImageNet. CIFAR-100 is corrupted with symmetric or asymmetric in-distribution noise where we randomly flip a the labels of a fixed percentage of the dataset to another from the same distribution. For out-of-distribution noise, we replace a fixed percentage of samples with images from ImageNet32 or Places365 as in Albert *et al.* [1] where r_{in} and r_{out} respectively denotes the in-distribution and out-of-distribution noise ratios. For mini-ImageNet, we use the web noise corruption from Jiang *et al.* [11]. We train on both of these datasets at a resolution of 32×32 with a pre-activation ResNet18 [13]. We train for 200 epochs, starting with a learning rate of 0.1. We use a batch size of 256, stochastic gradient descent with a weight decay of 5×10^{-5} . We perform the warmup phase with the supervised objective only for 30 epochs on CIFAR-100 and for 1 epoch on MiniImageNet. We evaluate our approach on real world data by conducting experiments on the webly fined grained datasets [32]. We follow the setup of Zeren *et al.* [30] and use a ResNet50 [13] pretrained on ImageNet [15] at a resolution of 448×448 . We train with a learning rate of 0.003 and use a batch size of 16, stochastic gradient descent with a weight decay of 10^{-3} and warmup for 10 epochs. We find that the class-balanced regularization (class reg) commonly used in the label noise litterature [17, 23] helps improve the validation accuracy so we minimize it jointly to the classification objective. For all experiments, we employ a cosine learning rate decay after the end of the warmup phase.

4.2. Detecting and correcting label noise

We propose to evaluate how commonly used metrics fair at retrieving noisy samples in both synthetic and controlled web noisy data. Figure 2 plots AUC retrieval scores for noisy samples for different metrics proposed in the literature. We study neighbor agreement and disagreement in the contrastive feature space as in [23] which is also used in [18]; training (small) loss based methods as in [3, 17]; Kullback–Leibler (kl) divergence as in [38, 31, 30]. Area under the curve (AUC) retrieval score for the noisy samples are reported at every epoch in Figure 2. We observe that metrics behave similarly in the presence of ID noise but greater differences are observed when retrieving controlled web noise from the CNWL dataset. We observe in that case that the cross-entropy loss (small loss) is the most accurate at retrieving noisy web samples. Note also that the retrieval accuracy of the different metrics greatly drops when compared to the synthetic in-distribution noise which prompts further research to improve the detection of web noise when

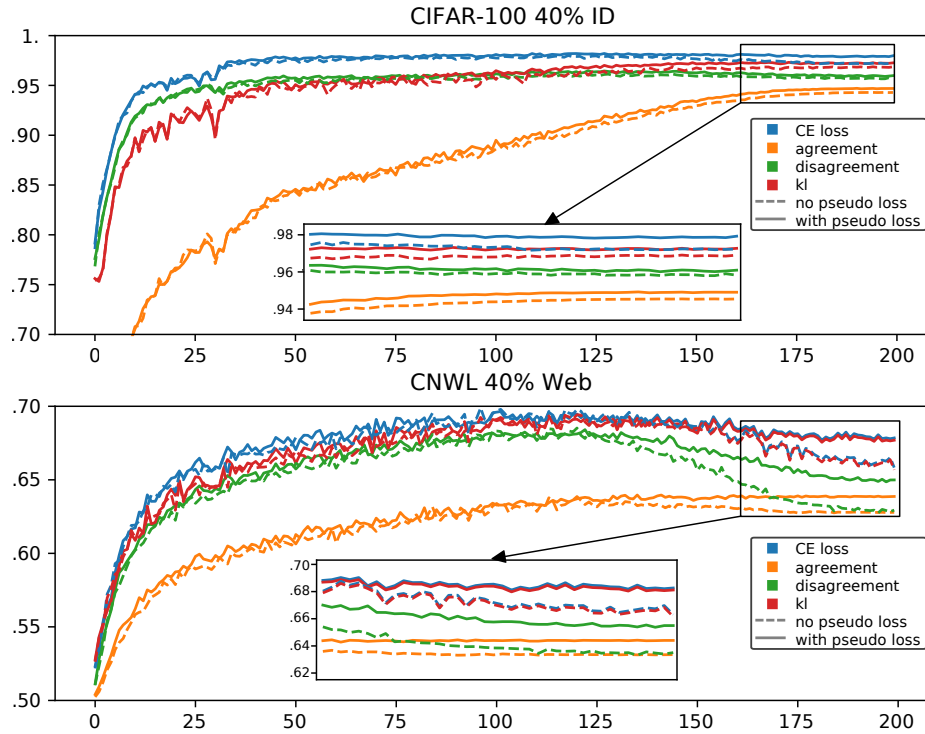


Figure 2. AUC for commonly used metrics in the literature when retrieving noisy samples. Full lines indicate that we remove incorrect pseudo-labels with the pseudo-loss. Dashed lines indicate that all pseudo-labels are used. Accounting for incorrect pseudo-labels improves the detection of noisy samples.

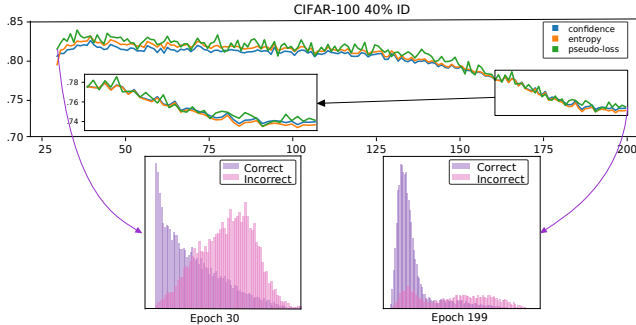


Figure 3. Pseudo-loss bi-modality and capacity to retrieve correctly guessed pseudo-labels. The top graph represents the AUC when retrieving correctly guessed pseudo-labels using the prediction confidence, entropy or the pseudo-loss. The bottom part of the figure shows the bi-modality of the pseudo-loss at two points during training and the confirmation bias at the end of the training.

no held-out labeled set is present.

4.3. Identifying incorrect pseudo-labels

We aim to demonstrate that not accounting for pseudo-label correctness in the detected noise is detrimental to both noise detection in the subsequent epochs as well as the overall true label recovery and validation accuracy throughout the training. We compare the pseudo-loss against the prediction confidence or entropy of the pseudo-label which commonly

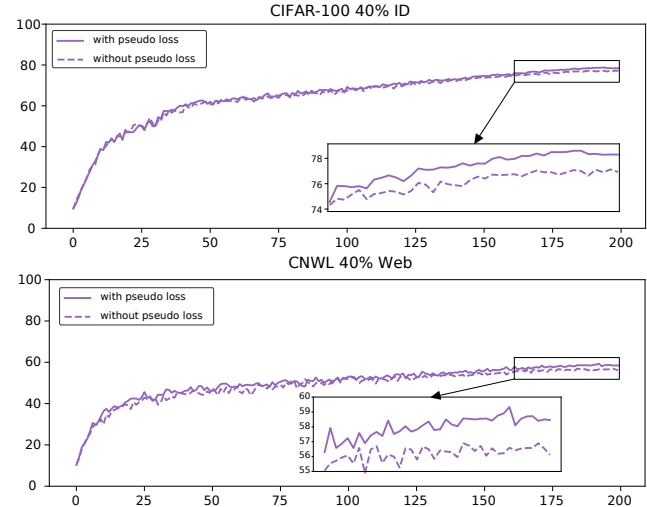


Figure 4. Validation accuracy per epoch, with or without removing incorrect pseudo-labels using the pseudo-loss

in the semi-supervised literature in Figure 3 where we train on CIFAR-100 corrupted with 40% of ID noise with no pseudo-label filtering. We observe that the pseudo-loss is on par with other metrics when retrieving correctly guessed pseudo-labels and that the detection on incorrect pseudo-labels becomes more challenging throughout the training as

Table 1. Ablation study and importance of the pseudo-loss selection w for the contrastive loss l_{cont} . Experiments conducted on CIFAR-100 corrupted with $r_{in} = 0.4, r_{in} = r_{out} = 0.2$ and on the CNWL corrupted with 40% web noise. Top-1 accuracy.

| | correct | cont | w | Noise level | | CNWL 40% |
|--|---------|------|---------------|----------------|-----------------------------------|----------|
| | | | | $r_{in} = 0.4$ | $r_{in} = 0.2$ $r_{out} = 0.4$ | |
| | ✗ | ✗ | ✗ | 66.31 | 59.54 | 46.40 |
| | ✓ | ✗ | ✗ | 74.53 | 69.17 | 55.08 |
| | ✓ | ✗ | ✓ | 75.57 | 69.53 | 55.72 |
| | ✓ | ✓ | ✗ | 76.21 | 70.10 | 58.62 |
| | ✓ | ✓ | ✓ | 77.43 | 72.21 | 59.90 |
| | ✓ | ✓ | $l_{classif}$ | 77.84 | 69.09 | 57.76 |

the learning rate is reduced and confirmation bias increases. More importantly, we find the pseudo-loss distribution over detected noisy samples to be bi-modal much like the small loss for noise detection. Consequently, we apply the same methodology as for the first stage detection where we fit a two mode gaussian mixture to the pseudo-loss and use the probability of a sample to belong to the low loss (correct pseudo-label) mode as w . This allows us to remove the need for a hyper-parameter threshold on the pseudo-label confidence/entropy as is always the case in the semi-supervised literature. PLS dynamically adapts as the training progresses and the network becomes naturally very confident in its predictions.

4.4. Pseudo-loss based selection of correct pseudo-labels

After assigning a probability of being correct for all guessed pseudo-labels on detected noisy samples, we evaluate how correct pseudo-label selection using the pseudo-loss influences label noise detection when the correction starts in Figure 2. We train a neural network with or without our proposed pseudo-label selection using the pseudo-loss weights from equations 3 and 6 (full line). We observe that both the noise retrieval and validation accuracy (Figure 4) are improved when incorrect pseudo-labels are removed. By avoiding to compute weight updates over incorrect pseudo-labels, our pseudo-loss selection reduces confirmation bias and improves the retrieval accuracy of noisy samples. Treating un-guessable samples as unlabeled in equation 6 helps us to further improve the classification accuracy (see ablation study in Table 1).

4.5. Ablation study

To understand better the benefit of each component to the final classification accuracy, we run an ablation study in Table 1. We study the case of in-distribution noise only ($r_{in} = 0.4$), when out-of-distribution noise is present ($r_{in} = 0.2, r_{out} = 0.4$) and in the case of web noise (CNWL with 40% web noise). We note that the selection of pseudo-labels using the pseudo-loss significantly improves the clas-

sification accuracy when training on with datasets presenting both ID, OOD or Web noise. We also evaluate how important pseudo-label selection is when optimizing the confidence guided contrastive objective l_{cont} . We run PLS and apply label selection in the classification objective $l_{classif}$ but not in the contrastive objective l_{cont} . Noisy samples are corrected using the current consistency regularization guess but incorrect pseudo-labels are not filtered for l_{cont} (they are for $l_{classif}$). Table 1 reports best accuracy results for CIFAR-100 corrupted with 40% in-distribution (ID) or 20% ID together with 40% out-of-distribution (OOD) from ImageNet32 and the CNWL dataset with 40% web noise. While we observe no major change for the ID corruption, a significant decrease in classification accuracy can be observed when keeping incorrect pseudo-labels in the contrastive objective when OOD or web noise is present (last row). For CIFAR-100 corrupted with 40% OOD and 20% ID noise, the accuracy benefits of training the contrastive objective are negated when compared to our noise correction baseline (row 2). We believe this motivates further research on the harmful impact OOD noise and incorrect pseudo-labels have when training on a web noisy dataset using a supervised contrastive objective.

4.6. State-of-the-art label noise robust algorithms

We propose to compare against the following state-of-the-art label noise robust algorithms: mixup (M) [43] has shown to be a strong regularization naturally robust to label noise; MentorMix [11] (MM) uses a student-teacher architecture to detect noisy samples before ignoring them; FaMUS [37] (FaMUS) is a meta-learning algorithm to detect label noise; Dynamic Bootstrapping [3] (DB) fits a beta mixture to the loss of training samples to detect noisy ones; S-model [8] (SM) uses a noise adaptation layer optimized using an EM algorithm; DivideMix [17] (DM) uses an ensemble of networks to detect noisy samples; PropMix [6] (PM) only corrects the simplest of the noisy samples according to their training loss; ScanMix [25] (SM) corrects samples using a semantic clustering approach; Robust Representation Learning [18] cluster class prototypes and use a weighed average of neighbors labels to correct noisy samples; Multi Objective Interpolation Training [23] (MOIT) train an interpolated contrastive objective and use neighbor label agreement to detect noisy samples. Regarding algorithms performing explicit in- and out-of-distribution noise detection, EvidentialMix [24] (EDM) fits a three component GMM to the evidential-loss [26]; JoSRC [38] (JoSRC) uses the Jensen–Shannon divergence; Dynamic Softening for Out-of-distribution Samples [2] (DSOS) uses the collision entropy and Spectral Noise clustering from Contrastive Features [1] (SNCF) clusters unsupervised features using OPTICS; Progressive Label Correction [44] (PLC) iteratively refine their noise detection under Bayesian guaranties,

Table 2. Mitigating ID noise on CIFAR-100. Accuracy numbers from respective papers or run using publicly available code. We bold the highest best accuracy and report standard deviation over 3 random noisy datasets and network initializations.

| Noise type | r_{in} | CE | M | DB | DM | ELR+ | MOIT+ | Sel-CL+ | RRL | PLS |
|------------|----------|-------|-------|-------|-------|-------|-------|---------|----------------|-------------------------|
| Symetric | 0.0 | 76.99 | 79.29 | 64.79 | 72.75 | 83.14 | 77.07 | 79.90 | 80.70 | 78.85 \pm 0.21 |
| | 0.2 | 62.60 | 71.55 | 73.9 | 77.3 | 77.6 | 75.89 | 76.5 | 79.4 \pm 0.1 | 80.03 \pm 0.15 |
| | 0.5 | 46.59 | 61.12 | 66.1 | 74.6 | 73.6 | 67.54 | 72.4 | 75.0 \pm 0.4 | 76.48 \pm 0.25 |
| | 0.8 | 23.46 | 37.66 | 45.67 | 60.2 | 60.08 | 51.36 | 59.6 | 32.21 | 63.33 \pm 0.38 |

Table 3. Mitigating ID noise and OOD noise on CIFAR-100 corrupted with ImageNet32 or Places365 images. Accuracy numbers from [1]. We bold the highest best accuracy and report standard deviation over 3 random noisy corruptions and network initialization.

| Corruption | r_{out} | r_{in} | CE | M | DB | JoSRC | ELR | EDM | DSOS | RRL | SNCF | PLS |
|------------|-----------|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------------------------|
| INet32 | 0.2 | 0.2 | 63.68 | 66.71 | 65.61 | 67.37 | 68.71 | 71.03 | 70.54 | 72.64 | 72.95 | 76.29 \pm 0.28 |
| | 0.4 | 0.2 | 58.94 | 59.54 | 54.79 | 61.70 | 63.21 | 61.89 | 62.49 | 66.04 | 67.62 | 72.06 \pm 0.19 |
| | 0.6 | 0.2 | 46.02 | 42.87 | 42.50 | 37.95 | 44.79 | 21.88 | 49.98 | 26.76 | 53.26 | 57.78 \pm 0.26 |
| | 0.4 | 0.4 | 41.39 | 38.37 | 35.90 | 41.53 | 34.82 | 24.15 | 43.69 | 31.29 | 54.04 | 56.92 \pm 0.49 |
| Places365 | 0.2 | 0.2 | 59.88 | 66.31 | 65.85 | 67.06 | 68.58 | 70.46 | 69.72 | 72.62 | 71.25 | 76.35 \pm 0.05 |
| | 0.4 | 0.2 | 53.46 | 59.75 | 55.81 | 60.83 | 62.66 | 61.80 | 59.47 | 65.82 | 64.03 | 71.65 \pm 0.61 |
| | 0.6 | 0.2 | 39.55 | 39.17 | 40.75 | 39.83 | 37.10 | 23.67 | 35.48 | 49.27 | 49.83 | 57.31 \pm 0.31 |
| | 0.4 | 0.4 | 32.06 | 34.36 | 35.05 | 33.23 | 34.71 | 20.33 | 29.54 | 26.67 | 50.95 | 55.61 \pm 0.55 |

Table 4. Comparison against state-of-the-art algorithms on the fine grained web datasets. We bold the best results. Top-1 best accuracy.

| Algorithm | Web-Aircraft | Web-bird | Web-car |
|---------------|--------------|--------------|--------------|
| CE | 60.80 | 64.40 | 60.60 |
| Co-teaching | 79.54 | 76.68 | 84.95 |
| PENCIL | 78.82 | 75.09 | 81.68 |
| SELFIE | 79.27 | 77.20 | 82.90 |
| DivideMix | 82.48 | 74.40 | 84.27 |
| Peer-learning | 78.64 | 75.37 | 82.48 |
| PLC | 79.24 | 76.22 | 81.87 |
| PLS | 87.58 | 79.00 | 86.27 |

Peer-Learning [32], Co-teaching [9], PENCIL [39] co-train two networks and identify clean samples by voting agreement; SELFIE [29] select low entropy noisy samples to be relabeled while discarding the rest.

4.7. Synthetic corruption

We first evaluate the capacity of PLS to mitigate in-distribution synthetic corruption. We run experiments on CIFAR-100 and compare against state-of-the-art algorithms in Table 2. To evaluate how much of the compared algorithms improvements come from an improve baseline accuracy over a superior noise correction, we also run the scenario where no noise is present. Because we do not use tricks such as unsupervised regularization and network ensembling, our algorithm presents a lower baseline when no noise is present but we achieve state-of-the-art results as soon as noise is introduced. This demonstrates the superiority of our approach when label noise is present in datasets.

4.8. Out-of-distribution corruption

Real world noisy data is often out-of-distribution [2]. We propose here to evaluate the performance of label noise robust algorithms on CIFAR-100 corrupted by a mixture between out-of-distribution images from ImageNet32 or

Places365 and symmetric in-distribution noise. Table 3 reports our results and compare with state-of-the-art algorithms. We observe here that the pseudo-loss allows us to effectively deal with out-of-distribution images which are filtered out in the pseudo-loss since no corrected label can be guessed by the neural network.

4.9. Controlled web noise

We validate our approach using the controlled web corruptions proposed in the CNWL datasets trained at a resolution of 32×32 . We report results in Table 5 and observe improvements over state-of-the-art algorithms including SNCF [1] which uses self-supervised pre-training to detect label noise.

4.10. Real world noise

We conduct experiments on real world noisy datasets that have been directly crawled from the web with no human curation. Table 4 reports results for the fine grained web datasets with results for other algorithms from Zeren *et al.* [31]. Note that even if we use a single network to train and make a prediction, we get comparable or better results than ensemble or co-learning methods. We report results for algorithms that do not use the label softening strategy (LSR [33]) as using this regularization technique is show in [32] to provide a strong baseline performance boost independently of the noise correction capabilities of the algorithms that use it.

4.11. Hyper-parameters

Table 6 reports the hyper-parameters used in all experiments.

5. Conclusion

This paper proposes a novel way to detect incorrect pseudo-label correction when dealing with label noise cor-

Table 5. Web-corrupted miniImageNet from the CNWL [11] (32×32). We run our algorithm; other results are from [6]. We report top-1 best accuracy and bold the best results.

| Noise level | CE | M | DM | MM | FaMUS | SM | PM | SNCF | PLS |
|-------------|-------|-------|-------|-------|-------|-------|-------|--------------|-------------------------|
| 20 | 47.36 | 49.10 | 50.96 | 51.02 | 51.42 | 59.06 | 61.24 | 61.56 | 63.10 \pm 0.14 |
| 40 | 42.70 | 46.40 | 46.72 | 47.14 | 48.03 | 54.54 | 56.22 | 59.94 | 60.02 \pm 0.15 |
| 60 | 37.30 | 40.58 | 43.14 | 43.80 | 45.10 | 52.36 | 52.84 | 54.92 | 54.41 \pm 0.49 |
| 80 | 29.76 | 33.58 | 34.50 | 33.46 | 35.50 | 40.00 | 43.42 | 45.62 | 46.51 \pm 0.20 |

Table 6. Hyperparameters used in the experiments

| Dataset | r_{in} | r_{out} | lr | epochs | Res | Net | GMM thresh | wd | warmup |
|--------------|----------|-----------|-------|--------|-----|----------|------------|--------------------|--------|
| CIFAR-100 | 0.0 | 0.0 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 30 |
| | 0.2 | 0.0 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 30 |
| | 0.5 | 0.0 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 30 |
| | 0.8 | 0.0 | 0.1 | 200 | 32 | PreRes18 | 0.5 | 5×10^{-5} | 30 |
| CIFAR-100 | 0.2 | 0.2 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 30 |
| | 0.2 | 0.4 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 30 |
| | 0.2 | 0.6 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 30 |
| | 0.4 | 0.4 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 30 |
| CNWL | 0.0 | 0.2 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 1 |
| | 0.0 | 0.4 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 1 |
| | 0.0 | 0.6 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 1 |
| | 0.0 | 0.8 | 0.1 | 200 | 32 | PreRes18 | 0.95 | 5×10^{-5} | 1 |
| Web-aircraft | - | - | 0.003 | 110 | 448 | Res50 | 0.95 | 10^{-3} | 10 |
| Web-bird | - | - | 0.003 | 110 | 448 | Res50 | 0.95 | 10^{-3} | 10 |
| Web-car | - | - | 0.003 | 110 | 448 | Res50 | 0.95 | 10^{-3} | 10 |

ruption. We use a state-of-the-art noise detection metric to detect noisy samples and guess their true label using a consistency regularization approach. The validity of the guessed true label is then evaluated using the pseudo-loss, which we show to be strongly correlated with pseudo-label correctness. Weight updates computed on pseudo-labels with a low probability of being correct are removed during training. We additionally propose to use an interpolated contrastive objective where correct pseudo-labels are used to learn inter class semantics while images with incorrect pseudo-labels are used in an unsupervised objective. We achieve state-of-the-art results on both synthetic and real world data.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number 16/RC/3835 - Vistamilk and 12/RC/2289_P2 - Insight as well as the support of the Irish Centre for High End Computing (ICHEC).

References

- [1] Paul Albert, Eric Arazo, Noel O’Connor, and Kevin McGuinness. Embedding contrastive unsupervised features to cluster in-and out-of-distribution noise in corrupted image datasets. In *European Conference on Computer Vision (ECCV)*, 2022.
- [2] Paul Albert, Diego Ortego, Eric Arazo, Noel O’Connor, and Kevin McGuinness. Addressing out-of-distribution label noise in webly-labelled data. In *Winter Conference on Applications of Computer Vision (WACV)*, 2022.
- [3] E. Arazo, D. Ortego, P. Albert, N. O’Connor, and K. McGuinness. Unsupervised Label Noise Modeling and Loss Correction. In *International Conference on Machine Learning (ICML)*, 2019.
- [4] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M.S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien. A Closer Look at Memorization in Deep Networks. In *International Conference on Machine Learning (ICML)*, 2017.
- [5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv: 2006.10029*, 2020.
- [6] Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. PropMix: Hard Sample Filtering and Proportional MixUp for Learning with Noisy Labels. *arXiv: 2110.11809*, 2021.
- [7] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep.*, 2011.
- [8] J. Goldberger and E. Ben-Reuven. Training deep neural networks using a noise adaptation layer. In *International Conference on Learning Representations (ICLR)*, 2017.
- [9] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Realistic evaluation of deep semi-supervised

- learning algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [10] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems (NeurIPS)*, 2018.
- [11] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels. In *International Conference on Machine Learning (ICML)*, 2020.
- [12] L. Jiang, Z. Zhou, T. Leung, L.J. Li, and L. Fei-Fei. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *International Conference on Machine Learning (ICML)*, 2018.
- [13] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [14] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [15] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NeurIPS)*, 2012.
- [16] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-Mix: A Strategy for Regularizing Contrastive Representation Learning. In *International Conference on Learning Representations (ICLR)*, 2021.
- [17] J. Li, R. Socher, and S.C.H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations (ICLR)*, 2020.
- [18] Junnan Li, Caiming Xiong, and Steven CH Hoi. Learning from noisy data with robust representation learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [19] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool. Web-Vision Database: Visual Learning and Understanding from Web Data. *arXiv: 1708.02862*, 2017.
- [20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*, 2014.
- [21] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda. Early-Learning Regularization Prevents Memorization of Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [22] D. Ortego, E. Arazo, P. Albert, N. O’Connor, and K. McGuinness. Towards Robust Learning with Different Label Noise Distributions. In *International Conference on Pattern Recognition (ICPR)*, 2020.
- [23] Diego Ortego, Eric Arazo, Paul Albert, Noel E. O’Connor, and Kevin McGuinness. Multi-Objective Interpolation Training for Robustness to Label Noise. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [24] Ragav Sachdeva, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. EvidentialMix: Learning with Combined Open-set and Closed-set Noisy Labels. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.
- [25] Ragav Sachdeva, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. ScanMix: Learning from Severe Label Noise via Semantic Clustering and Semi-Supervised Learning. *arXiv: 2103.11395*, 2021.
- [26] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [27] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning (ICML)*, 2019.
- [28] K. Sohn, D. Berthelot, C.-L. L, Z. Zhang, N. Carlini, E. Cubuk, A Kurakin, H. Zhang, and C. Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv: 2001.07685*, 2020.
- [29] H. Song, M. Kim, and J.-G. Lee. SELFIE: Refurbishing Unclean Samples for Robust Deep Learning. In *International Conference on Machine Learning (ICML)*, 2019.
- [30] Zeren Sun, Huafeng Liu, Qiong Wang, Tianfei Zhou, Qi Wu, and Zhenmin Tang. Co-LDL: A Co-Training-Based Label Distribution Learning Method for Tackling Label Noise. *IEEE Transactions on Multimedia*, 2021.
- [31] Zeren Sun, Fumin Shen, Dan Huang, Qiong Wang, Xiangbo Shu, Yazhou Yao, and Jinhui Tang. PNP: Robust Learning From Noisy Labels by Probabilistic Noise Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [32] Zeren Sun, Yazhou Yao, Xiu-Shen Wei, Yongshun Zhang, Fumin Shen, Jianxin Wu, Jian Zhang, and Heng Tao Shen. Webly Supervised Fine-Grained Recognition: Benchmark Datasets and An Approach. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [34] M. Toneva, A. Sordoni, R. Combes, A. Trischler, Y. Bengio, and G. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*, 2019.
- [35] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. *Advances in neural information processing systems (NeurIPS)*, 2017.
- [36] N. Vyas, S. Saxena, and T. Voice. Learning Soft Labels via Meta Learning. *arXiv: 2009.09496*, 2020.
- [37] Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. Faster meta update strategy for noise-robust deep learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [38] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-SRC: A Contrastive Approach for Combating Noisy Labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

- [39] K. Yi and J. Wu. Probabilistic End-to-end Noise Correction for Learning with Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [40] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning (ICML)*, 2019.
- [41] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [42] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires re-thinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.
- [43] H. Zhang, M. Cisse, Y.N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018.
- [44] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *International Conference on Learning Representations (ICLR)*, 2021.