Addressing out-of-distribution label noise in webly-labelled data

Paul Albert, Diego Ortego, Eric Arazo, 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

A recurring focus of the deep learning community is towards reducing the labeling effort. Data gathering and annotation using a search engine is a simple alternative to generating a fully human-annotated and human-gathered dataset. Although web crawling is very time efficient, some of the retrieved images are unavoidably noisy, i.e. incorrectly labeled. Designing robust algorithms for training on noisy data gathered from the web is an important research perspective that would render the building of datasets easier. In this paper we conduct a study to understand the type of label noise to expect when building a dataset using a search engine. We review the current limitations of state-of-the-art methods for dealing with noisy labels for image classification tasks in the case of web noise distribution. We propose a simple solution to bridge the gap with a fully clean dataset using Dynamic Softening of Out-of-distribution Samples (DSOS), which we design on corrupted versions of the CIFAR-100 dataset, and compare against state-of-the-art algorithms on the web noise perturbated MiniImageNet and Stanford datasets and on real label noise datasets: WebVision 1.0 and Clothing1M. Our work is fully reproducible https://git.io/JKGcj.

1. Introduction

Deep neural networks (DNNs) are now the standard approach for accurately solving image classification tasks [26, 48]. However, their principal drawback is the large amount of labeled examples required for training. There exist numerous alternatives to deal with the limited availability of labels, such as but not limited to, semi-supervised learning [1, 3, 4], self-supervised learning [8, 5] and robust training on automatically annotated datasets [23, 12]. This paper focuses on the latter.

Designing robust algorithms to train image classification DNNs in the presence of label noise is an important focus for the community [36]; these enable better adaptation of current DNN solutions to real-world problems where extensive curated datasets are unavailable or too expensive to build. Controlled label noise datasets are then often created by synthetically introducing label corruptions in the CIFAR-100 [17] comparison benchmark. Although good noise robustness is shown on these artificial datasets, web label noise has proven that these solutions generalize poorly to more realistic scenarios and can, in specific cases, be outperformed by robust data augmentation strategies such as mixup [12, 27].

We hypothesize that the main limitation for the correction of label noise in web crawled datasets comes from a common assumption made by most label noise robust algorithms [21, 30, 29, 41] where the true labels for noisy samples lie inside the label set, i.e. the label noise is *in-distribution* (ID). Conversely, we hypothesize that the label noise present in web crawled datasets is predominantly *out-of-distribution* (OOD), meaning the real labels for noisy samples cannot be inferred from the distribution. To confirm our hypothesis, we conduct a small but representative survey on the WebVision 1.0 dataset [23] to identify the type of noise one can expect in automatically annotated datasets crawled from the web. We then build and validate the DSOS method on controlled corrupted versions of the CIFAR-100 dataset [17] where ID noise is introduced using symmetric label flipping and where we use the ImageNet32 [6] dataset to introduce OOD noise. We compare with state-of-the-art label noise algorithms on multiple real-world open-source web-crawled datatsets including corrupted versions of the miniImageNet [39] and Stanford Cars [16] datasets provided by Jiang *et al*. [12], the mini-WebVision dataset [23], and the Clothing1M [44] dataset. We observe that noisy OOD samples can be leveraged to improve network generalization by enforcing dynamically softening of labels tending to a uniform distribution [20] rather than discarding them.

This paper's contributions are:

1. We conduct a representative survey over the type of noise to be expected when constructing a dataset using web queries.

2. We motivate and propose a novel noise detection metric, entropy of the interpolation of the network prediction and the ground-truth label, that is capable to accurately differentiate between clean, ID and OOD noise.

3. We propose DSOS, a simple solution to combat ID and OOD noise in web-crawled datasets and conduct controlled experiments and ablation studies on corrupted versions of the CIFAR-100 dataset.

4. We compare DSOS against state-of-the-art, noise-robust algorithms on real-world web-crawled datasets, demonstrating the validity of our findings for real-world applications.

2. Related work

2.1. Label noise detection

Label noise detection aims at distinguishing between clean and noisy samples in an unsupervised manner. The commonly used method is the small loss trick [2, 35, 37], which is based on the assumption that when training a neural network with a high learning rate, noisy samples will have a higher loss than their clean counterpart. The small loss observation extends to other metrics such as forgetting event count [38], pre-trained mentor network scoring [13], uncertainty [15], prediction consistency [35], accuracy, or entropy. The small loss can also be applied in multiple network settings to improve the detection [21, 11]. Other noise detection algorithms include using the Local Outlier Factor [42] to identify isolated samples in the feature space or meta-learning [41]. Training neural networks to detect OOD examples could also be considered relevant to creating a label noise robust algorithm, but this approach systematically requires at least a trusted, exclusively ID dataset [31, 47] and sometimes an additional exclusively OOD dataset [40]. This constraint is too limiting in scenarios where the nature of the noise is unknown.

2.2. Algorithms robust to label noise

DNNs have been shown to easily overfit noisy labels, leading to generalization degradation [49]. We categorize the first class of label noise robust algorithms as label correction algorithms. The goal for label correction algorithms is to denoise the dataset by guessing the true label for noisy samples. We include here approaches that perform this correction online when computing the final loss. Label guessing strategies include: label transition matrices [29], current network predictions [2, 30], semi-supervised learning [7, 27], and meta-learning inspired backpropagation [41]. The second correction strategy is centered around limiting the contributions of the noisy labels to the network's parameters by either using a curriculum [13, 10] or contribution weights in the final loss [34] that diminish the contribution of noisy samples in the gradient update. Other strategies include pushing apart the representations of clean and noisy samples in the feature space [42] or noise robust data augmentation [50]. Two algorithms have recently been proposed to tackle separate ID and OOD retrieval. EvidentialMix [32] proposes a separate detection of ID and OOD samples using the evidiential loss [33] but chooses to ignore the OOD samples to increase accuracy on the ID noise correction using the DivideMix [21] algorithm. JoSRC [45] proposes to differentiate between OOD and ID samples using a contrastive evaluation using multiple views of the same noisy sample. JoSRC additionally proposes a fixed smoothing for the labels of detected OOD samples using a fixed temperature hyperparameter. Both of these algorithms additionally require two networks. Recent studies [12, 27] show that the improvements noise robust algorithm observe on synthetic datasets do not always translate to realistic label noise scenarios.

Table 1: Analysis on the noise types and ratios found in mini-WebVision. We randomly sample three subsets (S) of 2000 images and report correctly-labeled samples and in-distribution (ID) and out-of-distribution (OOD) noisy samples. Image examples are available in the supplementary material.

	S1	S2	S3	Average (%)
Correct	1441	1440	1335	1405.33 (70.30)
OOD	460	429	573	487.33 (24.38)
ID	98	130	91	106.33 (5.32)

3. Web datasets and out-of-distribution noise

Recent state-of-the-art for label noise detection and correction rely on strong assumptions verified on synthetically generated noise. Recent contributions [12, 27] demonstrated that many algorithms developed on synthetic datasets do not generalize well to real-world label noise and that improvements are often inferior to using data augmentation (mixup [50]). We suggest that this limitation is a consequence of a strong assumption made by noise-robust algorithms where noisy samples have their labels corrected by assigning another label from the known label distribution, i.e. the noise is in-distribution. We conversely hypothesize that most of the noise in web labeled datasets is out-of-distribution, meaning the real unknown label lies outside of the known label set. To verify this hypothesis we randomly sample images from the real-world label noise dataset mini-WebVision (first 50 classes subset of the WebVision 1.0 dataset [23]) and manually categorize their label in three categories: clean, in-distribution noise, and out-of-distribution noise. We separate the labeling process in two steps: a first round on the images to detect any image whose label does

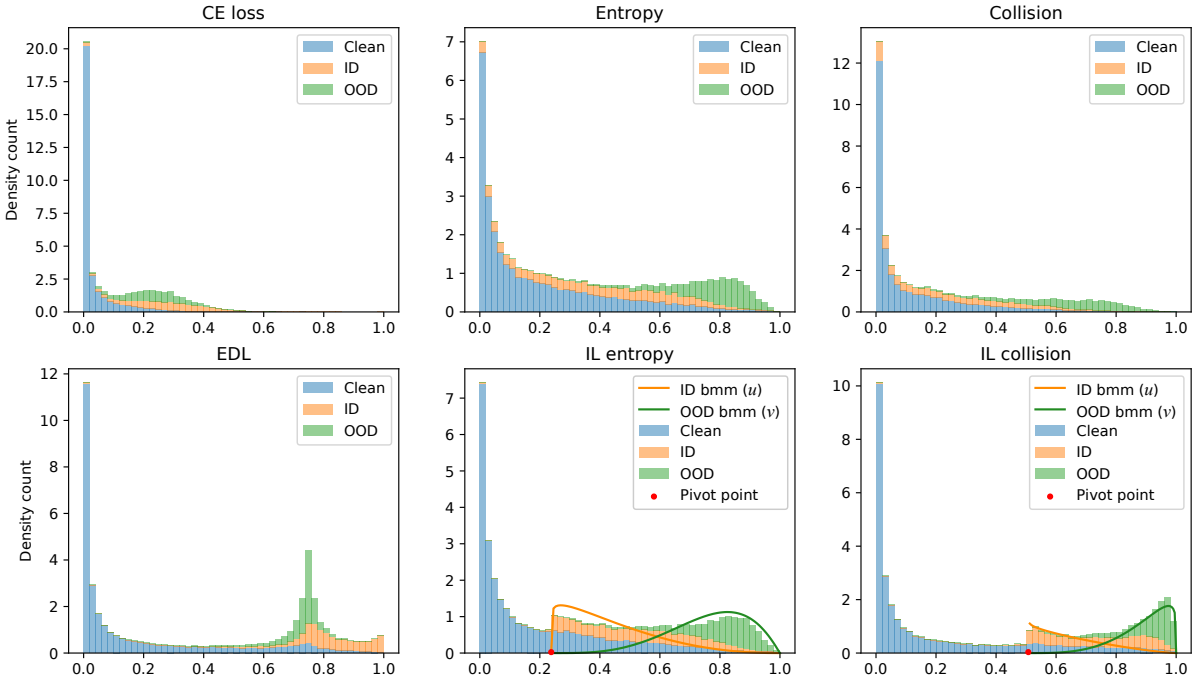

Figure 1: Stacked density histograms for multiple noisy sample retrieval measures on CIFAR-100 with $\rho = \psi = 0.2$. All metrics are min-max normalized. For the entropy of the intermediate label (IL) we also draw the decision function (BMM) that we fit to the data. The pivot point in red separates clean from noisy samples.

not correspond to the class to which it is assigned, and a second pass on the noisy images alone to classify them as ID when the true label lies in the known set of classes, or OOD when it does not. We repeat the process for three random subsets of the mini-WebVision dataset. Table 1 shows the results of the study, demonstrating the clear domination of out-of-distribution noise over in-distribution noise (A visualization of the noise categorization of the images is available in the supplementary material). This observation sheds light on the limited improvements of in-distribution label correction techniques when applied to web crawled datasets, while explaining the benefits of undersampling algorithms, which sample noisy data less often to reduce their contribution [10, 12, 32] at the cost of ignoring a part of the data.

4. DSOS

Taking into consideration the results observed in Section 3, we propose Dynamic Softening of Out-of-distribution Samples (DSOS), a label correction algorithm for robust learning on web label noise distributions. We aim to solve an image classification task over C classes as learning a DNN model h_ψ given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N samples where $x_i \in \mathcal{X}$. More specifically, we tackle the case where the dataset consists of a correctly labeled set $\mathcal{D}_c = \{(x_i, y_i)\}_{i=1}^{N_c}$ with corresponding one-hot encoded labels $y_i \in \{0, 1\}^C$, an incorrectly labeled in-distribution noisy set $\mathcal{D}_{in} = \{(x_i, y_i)\}_{i=1}^{N_{in}}$ and of an out-of-distribution noisy set $\mathcal{D}_{out} = \{(x_i, y_i)\}_{i=1}^{N_{out}}$. We denote $N = N_c + N_{in} + N_{out}$ the total number of available samples. We consider unknown the distribution of the samples between $\mathcal{D}_c$, $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$. We note $h : \mathcal{X} \to [0, 1]^C$ the deep neural network (DNN) we train to classify the images as belonging to a class $c \in \{1, \dots, C\}$.

4.1. Separate detection of ID and OOD noise

4.1.1 Motivation

We motivate here the need for a new metric for the dual detection of ID and OOD noise in web crawled datasets by considering the ideal case where a network has been trained on a web-crawled dataset and did not overfit the noise. Samples would then be characterized by either a confident correct prediction (clean samples), a confident incorrect prediction (ID noise), or an un-confident prediction (OOD noise). Using a DNN to detect noisy samples, metrics from in the label noise literature propose to either quantify the accuracy of the prediction [2, 21, 11] (cross-entropy loss, accuracy, Kullback-Leibler divergence) or the uncertainty of the prediction [38, 45] (forgetting events, entropy of the prediction, contrastive predictions). Relying on one characterization of the network prediction alone is problematic when presented with the duality of the noise present in web-crawled datasets

Table 2: AUC retreival score for different types of metrics after warm-up on CIFAR-100 with $\rho = \psi = 0.2$

	Clean	ID	OOD
Small loss	95	87	81
EDL	93	90	75
IL entropy	91	81	94
IL collision	93	85	92

as ID and OOD noise cannot be independently retrieved. While accuracy approaches indistinguishably retrieve incorrectly predicted OOD and ID noise (both having low agreement with their noisy label), certainty-based approaches only retrieve under-confident OOD noise. EvidentialMix [32] proposes an independent retreival of ID and OOD noise, where a mean square error + variance loss [33] (evidential loss, EDL) is shown to separate ID and OOD noise on artificial corrupted noisy datasets (CIFAR-10 [17]. We argue that the limitation of the evidential loss for web-crawled datasets lies in the absence of separation between OOD noise and lower-confidence predictions in general, resulting in a sub-optimal OOD retrieval, the dominant noise type for web-crawled datasets. This limitation is evidenced in Figure 1 (described in Section 4.1.2) and in Table 2 where we compare retrieval scores for Clean/ID/OOD samples (one versus all) for an accuracy (CE loss) or confidence metric (entropy) against using the EDL loss fitted with a 3 components Gaussian mixture model [32], and two variations of our proposed metric (see Section 4.1.2). The table highlights the trade-off we make for better OOD detection at the cost of less accurate ID retrieval when compared with EDL.

4.1.2 Metric

We propose a novel noise detection metric that allows the separate detection of confident clean samples, confident ID noisy samples, and OOD noisy samples. To do so, we propose to compute the intermediate label between the current network prediction $\tilde{Y}$ and the target label Y: $y_{int} = \frac{y_i + \tilde{y_i}}{2}$ and to study its collision entropy:

$$l_{detect} = -\log\left(\sum_{c=1}^{C} y_{int,c}^2\right). \tag{1}$$

We aim to detect three different events for y_{int}: the clean event where prediction and ground truth agree, resulting in a low entropy; the ID event where prediction and ground truth are both confident but disagree (medium entropy); and the OOD event where the prediction is under-confident (high entropy). Studying the entropy of the intermediate label l_{detect} allows us to reverse the detection hierarchy observed in the EDL from clean-OOD-ID to clean-ID-OOD since confident incorrect predictions are now observed in y_{int} as a bimodal distribution that has a lower entropy than an interpolation of the ground truth with an un-confident uniform prediction. A fundamental property of l_{detect} is that it differentiates between low confidence but correct predictions (clean samples) and confident incorrect predictions (ID noise), which is evidenced by the pivot point. The pivot point is defined for y_{int} being a perfect bi-model distribution, i.e. two high probability modes with values 0.5 with all other bins to 0, resulting in $l_{detect} = -\log 0.5$, the pivot point. Detecting these events of high probability motivate our choice of using the collision entropy, which is more sensitive to high probability events than the Shannon entropy. Using the pivot point together with the observed bimodality of the noisy samples, we classify the samples in three distinct categories where every sample whose l_{detect} value is inferior to the pivot point is considered clean and where we fit a two components Beta Mixture Model (BMM) to the noisy samples. By computing the posterior probability of a sample to belong to each component, we evaluate the ID and OOD nature of every noisy sample.

Figure 1 illustrates the clean/ID/OOD separation observed for accuracy based and uncertainty based metrics on the CIFAR-100 dataset corrupted with 20% symetric ID noise and 20% OOD noise from ImageNet32 [6] at the end of the warm-up phase (see Section 5.1 for training details). The figure illustrates how the collision entropy improves the separation between clean and ID noise over the Shannon entropy and how we trade off improved OOD detection for a decreased ID detection over the evidential loss (EDL) [32] (see Table 2). The pivot point is indicated in red. An additional illustration explaining the behavior of l_{detect} for intermediate configurations of y_{int} is available in the supplementary material.

4.2. DSOS

We build DSOS as a single network based, single training cycle algorithm which aims to first discover ID and OOD samples in a corrupted dataset before separately addressing ID and OOD noise using dynamic label correction strategies. Figure 2 illustrates the DSOS algorithm. We aim to correct ID samples using confident predicted label assignments and to promote high entropy prediction for OOD samples which cannot be corrected. DSOS aims to minimize the following empirical risk over the noisy dataset:

$$R_e = \frac{1}{N}\sum_{i=1}^{N} -{y_i^t}^T \log h(x_i), \tag{2}$$

where the logarithm is applied element-wise and y_i^t denotes the, possibly unknown, true label for sample x_i. Although it is possible to directly minimize R_e for ID noisy samples by correcting the noisy label y_i to the true label y_i^t, this is not the

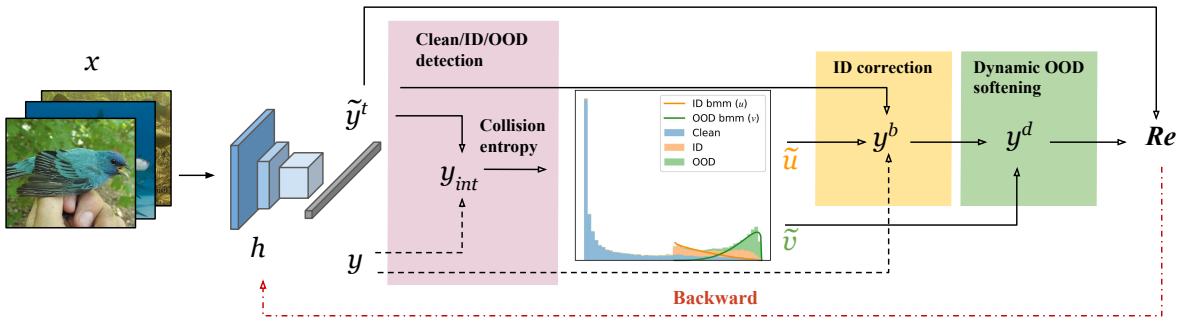

Figure 2: Visualization of the DSOS algorithm. DSOS identifies and corrects the ID and OOD noise from the training distribution before applying targeted label correction.

case for OOD label noise. We propose then not to attempt to approximate the true label of OOD samples using a label from the known distribution but instead to promote better network calibration by encouraging high-entropy predictions, i.e. a uniform prediction over ID classes. We then rewrite empirical risk as:

$$R_e = -\frac{1}{N_c+N_{in}}\sum_{i=1}^{N_c+N_{in}} {y_i^t}^T \log h(x_i) - \frac{1}{N_{out}}\sum_{j=1}^{N_{out}} {y_s}^T \log h(x_j), \tag{3}$$

where y_s is the softened label, i.e. a perfect uniform prediction over all the classes C. To obtain a dynamic softening from y_i^t to y_s and given a OOD classifier $\mathcal{V} = \{v_i\}_{i=1}^N, v_i \in [0,1]$ where $v_i = 0$ means sample x_i is OOD, we minimize:

$$R_e = -\frac{1}{N}\sum_{i=1}^{N} f(y_i^t, v_i)^T \log h(x_i), \tag{4}$$

with $f(y_i^t, v_i)$ the smoothing function where $f(y_i^t, 0) = y_s$ and $f(y_i^t, 1) = y_i^t$.

4.2.1 Label softening of out-of-distribution samples

We minimize the risk in Eq. 4 using a label correction approach where we aim to first correct the labels for noisy ID samples to their true label using a bootstrapping inspired approach [2, 30, 35]. For the OOD samples, we propose a dynamic softening strategy by computing the cross-entropy loss with regards to a dynamically smoothed label (the more likely a sample is detected to be OOD, the more uniform the target) and avoid using an additional regularization term (Kullback-Leibler divergence minimization between the prediction and a uniform target would be a common solution [20]). To correct ID label noise, we consider a first estimated metric $\tilde{U} = \{\tilde{u}_i\}_{i=0}^N$, where $\tilde{u}_i \in \{0,1\}$, evaluating whether a sample is noisy but in-distribution, i.e. the label can be corrected to another from the distribution. $\tilde{u}_i = 1$ denotes sample x_i is noisy but ID. We denote $\tilde{y}_i^t$ the current true label guess for sample x_i an correct it with,

$$y_i^b = (1-\tilde{u}_i)y_i + \tilde{u}_i\tilde{y}_i^t. \tag{5}$$

Regarding OOD label noise, we consider a second metric $\tilde{V} = \{\tilde{v}_i\}_{i=0}^N$ estimating $\mathcal{V}$ and evaluating whether a sample is noisy and OOD ($\tilde{v}_i \in (0,1]$) with $v_i = 0$ meaning a sample is considered OOD. We re-normalize the possibly bootstrapped label y_i^b for a sample x_i assigned to an OOD noisiness metric estimation $\tilde{v}_i$ as

$$y_i^d = \frac{\exp\frac{\tilde{v}_i y_i^b}{\alpha}}{\sum_{c=1}^{C}\exp\frac{\tilde{v}_i y_{i,c}^b}{\alpha}}. \tag{6}$$

with $\alpha \in [0,1]$ a hyperparameter. y_i^d is a dynamically smoothed correction of the corrected label y_i^b where $\frac{\tilde{v}_i}{\alpha}$ serves as a dynamic temperature depending on the out-of-distribution noisiness of the sample. In Figure 1, $\tilde{U}$ corresponds to the posterior probability given l_{detect} for the left-most beta mixture being superior to 0.5 and $\tilde{V}$ is the posterior probability of the right-most beta mixture given l_{detect} (no threshold). We evaluate $\tilde{U}$ and $\tilde{V}$ every epoch starting at the end of the warm-up phase where the network is trained without correction on the noisy dataset. We end the warm-up phase one epoch after the first learning rate reduction. In summary, OOD noisy labels will be dynamically replaced by a uniform distribution hence promoting their rejection by the network and the clean and corrected ID noisy samples will be assigned a moderately smoothed label, which has been proven to be beneficial for robust DNN training in the presence of label noise [19, 25]. Both $\tilde{U}$ and $\tilde{V}$ are cut of from the computation graph and neither is backpropageted in equation 4.

4.2.2 Additional regularization

In order to be competitive with the state-of-the-art, we pair DSOS with two different regularization strategies commonly

Table 3: DSOS for mitigating ID and OOD noise on CIFAR-100 corrupted with ImageNet32 images. We run each algorithm with the exact same noise corruption. We report best and last accuracy (best/last).

ρ	ψ	CE	M	DB	ELR	EDM	JoSRC	DSOS		
								ID	OOD	both
0.2	0.2	63.68/55.52	66.71/62.52	65.61/65.61	63.90/63.72	65.11/64.49	67.37/64.17	68.09/67.78	69.37/69.37	70.54/70.54
0.4	0.2	58.94/44.31	59.54/53.16	54.79/54.42	57.16/56.91	55.65/54.49	61.70/61.37	60.12/59.32	62.34/61.03	62.49/62.05
0.6	0.2	46.02/26.03	42.87/40.39	42.50/42.50	31.20/29.55	28.51/10.47	37.95/37.11	46.10/42.93	46.54/40.23	49.98/49.14
0.4	0.4	41.39/18.45	38.37/33.85	35.90/35.90	22.85/21.63	24.15/01.62	41.53/41.44	40.94/35.89	42.53/39.76	43.69/42.88

Table 4: Ablation study for DSOS. We report best and last accuracy.

	Best	Last
CE	63.68	55.52
+ mixup	66.71	62.52
+ Entropy regularization	67.27	63.04
+ Batch normalization tuning	67.56	65.69
+ In-distribution bootstrapping	68.09	67.78
+ Out-of-distribution softening	70.54	70.54

used to combat label noise. The first regularization we add to the loss promotes high-entropy predictions on ID samples:

$$l_e = -\frac{1}{N}\sum_{i=1}^{N}\tilde{v}_i\sum_{i=1}^{N}h(x_i)\log(h(x_i)). \tag{7}$$

We find l_e to be especially important in the warm-up phase as it promotes confident predictions for both the clean samples and the ID samples, which enables better detection. During the label correction phase of DSOS, the regularization is proportionally weighted according to the clean and noisy ID samples detection $\tilde{V}$ so as to not to go against the label softening strategy for OOD samples. We additionally pair DSOS with mixup [50] data augmentation, which has shown to be robust to label noise and that is commonly used in related state-of-the-art noise robust approaches. An ablation study for the different components of DSOS including the effect of the regularizations is given in Section 5.3. The final loss DSOS minimizes is:

$$l = -\frac{1}{N}\sum_{i=1}^{N}y^{d^T}\log(h(x_i)) + \gamma l_e \tag{8}$$

with $\gamma = 0.4$.

5. Experiments

5.1. Experimental setup

We conduct controlled experiments on corrupted versions of the CIFAR-100 dataset [17] using ImageNet32 [6] images for the OOD noise. The CIFAR-100 dataset is a 32×32 image dataset composed of 50.000 training images and 10.000 test images, equally distributed over 100 classes. The ImageNet32 dataset is a 32×32 downsized version of the ILSVRC12 [18] dataset (1.000 classes and 1.2M images). In order to corrupt CIFAR-100, we consider the OOD noise ratio ρ and the ID noise ratio ψ. We first replace a random fraction ρ of the CIFAR-100 images by randomly selected ImageNet32 [6] images and randomly flip a ψ fraction of the clean samples to a random label assignment. The total noise ratio is $\psi + \rho$. We train for 100 epochs, using a PreActivation ResNet18 [14], SGD with momentum 0.9 and weight decay 5×10^{-4}, starting from a learning rate of 0.03 and reducing it by 10 at epochs 50 and 80, batch size 32 (64 for the warm-up).

For controlled web-crawled datasets, we consider different noise levels ($0\%, 30\%, 50\%, 80\%$) for the web label noise corruption released for the MiniImageNet ($50k$ training images, 10.000 test images) and StanfordCars ($8k$ training images, $8k$ test images) datasets [12], adopting the 299×299 image resolution for training and the Inception-ResNetV2 network architecture. We train for 200 epochs, using SGD with momentum 0.9 and weight decay 5×10^{-4}, starting from a learning rate of 0.01 and reducing it by 10 at epochs 100 and 160, batch size 32. For real-world web-crawled datasets, we report results training on the mini-Webvision [23] dataset (first 50 classes of WebVision) ($66k$ training images, $2.5k$ test images) at resolution 224×224. We train for 100 epochs, using an InceptionResNetV2, SGD with momentum 0.9 and weight decay 5×10^{-4}, starting from a learning rate of 0.01 and reducing it by 10 at epochs 50 and 80, batch size 32. We use the mini-WebVision validation set for early stopping and the ILSVRC12 dataset [18] as a test set. For Clothing1M [44] ($1M$ training images, $15k$ test images) we sample 1000 random batches every epoch, resolution 227×227. We train for 100 epochs using a ResNet50 pretrained on ImageNet, SGD with momentum 0.9 and weight decay 1×10^{-3}, starting from a learning rate of 0.002 and reducing it by 10 at epochs 50 and 80, batch size 32. The dataset configurations and networks used follows the state-of-the-art we compare with [12, 21, 24]. A summary of the training details is available in the supplementary material.

Table 5: Comparison of DSOS with state-of-the-art algorithms on MiniImageNet and Stanford Cars corrupted with web label noise gathered by [12] (red noise). We bold best and underline last accuracy for the best performing algorithm.

Dataset	Noise level	CE	D	SM	B	M	MN	MM	DSOS
MiniImageNet	0	70.9/68.5	71.8/65.7	71.4/68.4	71.8/68.4	72.8/72.3	71.2/68.9	74.3/73.7	**74.52**/<u>74.10</u>
	30	66.1/56.5	66.6/55.0	65.2/56.3	66.6/56.7	66.8/61.8	66.2/64.0	68.3/67.2	**69.84**/<u>67.86</u>
	50	60.9/51.7	62.1/50.01	61.3/51.3	62.6/52..5	63.2/58.4	61.7/58.0	63.3/61.8	**66.14**/<u>65.18</u>
	80	48.8/39.8	49.5/37.6	49.0/40.6	50.1/40.1	50.7/45.5	49.3/43.4	50.2/48.4	**55.26**/<u>52.24</u>
Stanford Cars	0	90.8/90.8	**92.2**/<u>92.2</u>	90.1/90.1	90.3/90.0	91.9/91.9	90.2/90.1	91.8/91.6	91.38/91.27
	30	80.4/80.2	87.6/87.6	82.2/81.9	83.4/83.0	85.6/85.2	81.1/80.9	87.8/87.7	**88.36**/<u>88.14</u>
	50	70.6/70.3	79.3/79.2	70.1/70.1	73.6/73.5	79.1/78.9	72.0/72.0	80.4/79.8	**82.04**/<u>81.72</u>
	80	43.3/43.0	61.8/61.8	46.4/46.4	47.4/46.7	55.7/55.4	51.0/50.9	58.6/58.6	**62.36**/<u>62.36</u>

Table 6: Classification accuracy for DSOS and state-of-the-art methods against methods using a unique network vs an ensemble. We train the network on the mini-Webvision dataset and test on the Imagenet 1k test set (ILSVRC12). All results except our own (DSOS) are from [24]. We bold the best results.

		Unique network						Ensemble of two networks		
		F	Co-T	M	MM	ELR	DSOS	DM	ELR+	DSOS
mini-WebVision	top-1	61.12	63.58	75.44	76.0	76.26	**77.76**	77.32	77.78	**78.76**
	top-5	82.68	85.20	90.12	90.2	91.26	**92.04**	91.64	91.68	**92.32**
ILSVRC12	top-1	57.36	61.48	71.44	72.9	68.71	**74.36**	75.20	70.29	**75.88**
	top-5	82.36	84.70	89.40	**91.10**	87.84	90.80	90.84	89.76	**92.36**

5.2. Experiments on CIFAR-100

We test DSOS in a controlled noise scenario on the CIFAR-100 dataset corrupted with ID symetric label noise and OOD images from the ImageNet32 dataset in Table 3. Contrary to previous works [45], the focus here is on OOD noise. We consider 4 different configurations for CIFAR-100 with $\rho \in [0.2, 0.4, 06]$ and $\psi \in [0.2, 0.4]$. We show the benefits of DSOS when performing ID label bootstrapping or OOD label softening alone as well as the combined benefits of the dual label correction (both in Table 3). We compare our approach with two simple baselines: CE, a simple cross-entropy training without any noise correction and mixup (M) [50] a data augmentation strategy robust to label noise. We additionally report results for state-of-the-art noise robust algorithms including Dynamic Bootstrapping (DB) [2] and Early Learning Regularization (ELR) [24]. Finally, we run algorithms focused on OOD and ID noise robustness: EvidentialMix (EDM) [32] and JoSRC [45]. We use the same hyperparameters and network as ours for training the algorithms we compare with except for JoSRC which uses the Adam optimizer by default. For DSOS, we perform a warm-up training up until after the learning rate reduction. One epoch after the learning rate reduction, we start performing ID and OOD noise detection and apply our label correction strategy with $\alpha = 0.05$. We find that performing warm-up with mixup (M) is better as long as the total noise is superior to 0.8 but use a simple CE warm-up for total noise levels of 0.8. We systematically use the entropy regularization term for the warm-up phase. We report running DSOS with ID or OOD correction alone as well as with both correction (both). If we notice that the BMM does not capture the ID mode (mode of the first beta distribution outside of the $[0, 1]$ interval) which we observe for total noise levels of 0.8, we fall back to using l_{detect} directly for detecting the ID noisy samples ($l_{detect} < 0.5$ means a samples is ID noisy). We draw the attention of the reader to the improvements DSOS brings when compared to other ID/OOD noise correction approaches even though we use a single network.

5.3. Ablation study

We conduct an ablation study to highlight the important elements of DSOS trained on CIFAR-100 with $\rho = 0.2$ and $\psi = 0.2$ (Table 4). We find entropy regularization [37] to be necessary to promote confident predictions and specifically study the case where the metrics tracking and the bootstrapped label predictions necessary to applying ID noise correction are computed with trainable batch normalization layers, i.e. the layers get tuned with unmixed samples before evaluation on the validation set. The ablation study highlights how the introduction of the dynamic label softening strategy improves accuracy results over applying ID label correction alone.

5.4. Comparison against the state-of-the-art

Table 5 reports results for DSOS when compared to state-of-the-art approaches on the web-corrupted versions of Stanford Cars and MiniImageNet [12]. Table 6 compares DSOS against state-of-the-art algorithms on the Web-

Table 7: Comparison of DSOS against state-of-the-art algorithms on Clothing1M. Top-1 best accuracy on the test set. We run ELR+ and DM using the code provided by the authors. All other results are from the specified works. We bold the best results.

	Unique network								Ensemble of two networks		
	CE	F	SL	JO	ELR	Me	P	DSOS	ELR+	DSOS	DM
Clothing1M	69.10	69.84	71.02	72.16	72.87	73.47	73.49	**73.63**	74.05	74.13	**74.76**

Table 8: Wall-clock training time comparison for state-of-the-art algorithms on the mini-Webvision dataset. All algorithms were run on an RTX 2080 Ti GPU using the PyTorch [28] framework.

	M	ELR	DSOS	ELR+	DM
Epoch	*9.5min*	*10.5min*	*11.25min*	*28min*	*50min*
Full training	*15.75h*	*17.5h*	*18.75h*	*46.75h*	*83h*

Vision 1.0 dataset [23] reduced to the 50 first classes (mini-WebVision, 66K images), a large scale dataset created using web queries. Table 7 reports results for Clothing1M. When necessary, we differentiate between methods using a unique network for inference and methods using an ensemble of two networks. In this case, we ensemble two networks trained using DSOS from different random initialization and show the direct benefits of using an ensemble in the web label noise scenario. We compare with loss or label correction algorithms: Forward correction (**F**) [29], Bootstrapping (**B**) [30], Probabilistic correction (**P**) [46], Joint Optimization (**JO**) [37], S-Model **(SM)** [9]; sample selection algorithms: Co-Teaching (**Co-T**) [11], MentorMix (**MM**) [12], MentorNet (**MN**) [13]; semi-supervised correction algorithm: DivideMix (**DM**) [21], Early Learning Regularization (**ELR** and **ELR+**) [24]; regularization algorithms: Mixup (**M**) [50], Symetric cross-entropy Loss (**SL**) [43]; meta-learning algorithms: Learning to learn (**Me**) [22]; standard cross-entropy training (**CE**), standard cross-entropy plus dropout (**D**).

5.5. Training speed

Table 8 reports the wall-clock training time for state-of-the-art methods on the mini-Webvision subset. The first line reports average epoch time, warm-up included, and the second line reports the full training duration (100 epochs). Both of these metrics exclude evaluation on a validation set. We compare against state-of-the-art algorithms performing the best on mini-Webvision **DM** [21], **ELR** and **ELR+** [24], **M** [50]. DSOS improves accuracy results on mini-Webvision and trains significantly faster then the closest performing algorithms. Note that the training time for DivideMix [21] heavily depends on the training scenario as the algorithm oversamples the unlabeled data every epoch, i.e. the epoch length depends on clean/noisy detection.

5.6. Discussion

DSOS improves accuracy results on web crawled datasets such as mini-WebVision (Table 6) or web corrupted datasets: miniImageNet (large grained) and Stanford cars (fine grained) in Table 5. We explain the lower performance on Clothing1M by the specificity of the gathering process for the dataset which, according to the authors [44], contains very high levels of in-distribution noise because the dataset was crawled from a clothes database exclusively. This goes against our hypothesis in Section 3. Even then, our results are competitive and convergence is reached faster for DSOS, see Table 8.

6. Conclusion

This paper provides evidence of the nature of noise (dominantly out-of-distribution) in web-crawled datasets, which we believe to be the reason why improvements reported by recent state-of-the-art noise robust algorithms do not translate to real world noisy datasets. To train a noise-robust neural network on web crawled datasets, we propose DSOS, a simple algorithm using a novel noise detection metric capable of differentiating between clean, in-distribution noisy and out-of-distribution samples. We propose to detect and treat in-distribution and out-of-distribution noise differently to promote a dynamic rejection of unseen out-of-distribution samples, which in turn improves the generalization capabilities of the network. DSOS is a much simpler approach to label noise than the top state-of-the-art algorithms that we compare against as we use a one network and online correction strategy with a single training cycle. By properly identifying and correcting the two distinct label noise distributions, DSOS improves on the most competitive state-of-the-art algorithms. Other strategies could be used to improve network generalization by using out-of-distribution samples such as self-supervised learning, which can learn visual concepts without labels or data augmentation strategies using out-of-distribution samples to augment in-distribution samples. We leave this observation for future work.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/15/SIRG/3283 and SFI/12/RC/2289_P2.

References

[1] P. Albert, D. Ortego, E. Arazo, N.E. O'Connor, and K. McGuinness. ReLaB: Reliable Label Bootstrapping for Semi-Supervised Learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2021.

[2] 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.

[3] E. Arazo, D. Ortego, P. Albert, N.E. O'Connor, and K. McGuinness. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

[4] D. Berthelot, N. Carlini, I.J. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*, 2020.

[6] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv: 1707.08819*, 2017.

[7] Y. Ding, L. Wang, D. Fan, and B. Gong. A Semi-Supervised Two-Stage Approach to Learning from Noisy Labels. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[8] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[9] J. Goldberger and E. Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations (ICLR)*, 2017.

[10] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M.R. Scott, and D. Huang. CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images. In *European Conference on Computer Vision (ECCV)*, 2018.

[11] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[12] 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.

[13] 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.

[14] 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.

[15] M. Köhler, J.and Autenrieth and W. Beluch. Uncertainty Based Detection and Relabeling of Noisy Image Labels. In *IEEE Workshop on Computer Vision and Pattern Recognition (CVPRW)*, 2019.

[16] J Krause, M. Stark, J. Deng, and L. Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *Workshop on 3D Representation and Recognition (3dRR-13)*, 2013.

[17] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NeurIPS)*, 2012.

[19] D. Lee and Y. Cheon. Soft Labeling Affects Out-of-Distribution Detection of Deep Neural Networks. In *Workshop on International Conference on Machine Learning (ICMLW)*, 2020.

[20] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations (ICLR)*, 2018.

[21] 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.

[22] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool. WebVision Database: Visual Learning and Understanding from Web Data. *arXiv: 1708.02862*, 2017.

[24] 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.

[25] M. Lukasik, S. Bhojanapalli, A. K. Menon, and S. Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning (ICML)*, 2020.

[26] T. Mingxing and L. Quoc. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*, 2019.

[27] 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.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[29] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *International Conference on Learning Representations (ICLR)*, 2015.

[31] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[32] 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.

[33] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[34] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[35] H. Song, M. Kim, and J.-G. Lee. SELFIE: Refurbishing Unclean Samples for Robust Deep Learning. In *International Conference on Machine Learning (ICML)*, 2019.

[36] H. Song, M. Kim, D. Park, and J.-G. Lee. Learning from noisy labels with deep neural network: A survey. *arXiv: 2007.08199*, 2020.

[37] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint Optimization Framework for Learning with Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] 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.

[39] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D Wierstra. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2016.

[40] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *European Conference on Computer Vision (ECCV)*, 2018.

[41] N. Vyas, S. Saxena, and T. Voice. Learning Soft Labels via Meta Learning. *arXiv: 2009.09496*, 2020.

[42] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S.-T. Xia. Iterative Learning With Open-Set Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[43] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey. Symmetric cross entropy for robust learning with noisy labels. In *IEEE International Conference on Computer Vision (ECCV)*, 2019.

[44] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[45] 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.

[46] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[47] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[48] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv: 1605.07146*, 2016.

[49] 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.

[50] H. Zhang, M. Cisse, Y.N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018.