An accurate detection is not all you need to combat label noise in web-noisy datasets

Paul Albert[1], Jack Valmadre[1], Eric Arazo[2], Tarun Krishna[3], Noel E. O'Connor[3], Kevin McGuinness[3]

[1] Australian Institute for Machine Learning, University of Adelaide
[2] CeADAR: Ireland's Centre for Applied Artificial Intelligence
[3] Insight Centre for Data Analytics, Dublin City University
paul.albert@adelaide.edu.au

Abstract. Training a classifier on web-crawled data demands learning algorithms that are robust to annotation errors and irrelevant examples. This paper builds upon the recent empirical observation that applying unsupervised contrastive learning to noisy, web-crawled datasets yields a feature representation under which the in-distribution (ID) and out-of-distribution (OOD) samples are linearly separable [2]. We show that direct estimation of the separating hyperplane can indeed offer an accurate detection of OOD samples, and yet, surprisingly, this detection does not translate into gains in classification accuracy. Digging deeper into this phenomenon, we discover that the near-perfect detection misses a type of clean examples that are valuable for supervised learning. These examples often represent visually simple images, which are relatively easy to identify as clean examples using standard loss- or distance-based methods despite being poorly separated from the OOD distribution using unsupervised learning. Because we further observe a low correlation with SOTA metrics, this urges us to propose a hybrid solution that alternates between noise detection using linear separation and a state-of-the-art (SOTA) small-loss approach. When combined with the SOTA algorithm PLS, we substantially improve SOTA results for real-world image classification in the presence of web noise
https://github.com/PaulAlbert31/LSA

1 Introduction

Developing learning algorithms that are robust to label noise promises to enable the use of deep learning for a variety of tasks where automatic but imperfect annotation is available. This paper studies the specific case of web-noisy datasets for image classification. A web-noisy dataset [25, 33] is in fact the starting point for most generic image classification datasets, before human curation and label correction is conducted. To create a web-noisy dataset, the only required human intervention is the definition of a set of classes to be learned. Once the classes are defined, examples are recovered by text-to-image search engines, sometimes aided by query expansion and image-to-image search. Since the text surrounding an image on a web-page may not be an accurate description of the semantic content

of the image, some training examples will incorrectly represent the target class, leading to a degradation of both the model's internal representation and its final decision. Research has identified that, in the case of web-noisy datasets, out-of-distribution (OOD) images are by far the most dominant form of noise [3].

We propose to build upon the observations made in SNCF [2], who observed that representations learned by unsupervised contrastive algorithms on OOD-noisy datasets displayed linear separability between in-distribution (ID) and OOD images. We extend this observation to web-noisy datasets containing OOD images where we notice that the separation is not as good as SNCF observed on synthetically corrupted datasets. Upon further investigation, we however notice that the separation is recovered when evaluating intermediate representations, computed earlier in the network. Another limitation of SNCF we aim to address is the reliance on clustering to retrieve the noisy samples so we propose to directly estimate the linear separator. We compute an approximated linear separation using SOTA noise-robust algorithms [1, 24] to obtain an imperfect clean/noisy detection, which we then use to train a logistic regression on the unsupervised contrastive features. This produces an accurate web-noise detection.

Interestingly, when substituting our more accurate noise detection for the original detection metric in naive ignore-the-noise algorithms and subsequently the noise robust algorithm PLS [1], we observe a decrease in classification accuracy. In fact, we identify that few simple yet important clean examples are missed by our linear separation although they are correctly retrieved by SOTA noise detection [1, 24]. Because we find that our linear separation is decorrelated these SOTA noise detectors, we propose a detection strategy that combines linear separation (which achieves high specificity and sensitivity) and SOTA noise-detection approaches (which correctly retreive those few important samples) by alternating each every epochs. We combine this noise detection with PLS to create PLS-LSA which we find to be superior to existing noise-robust algorithms on a variety of classification tasks in the presence of web-noise. We contribute:

1. A novel noise detection approach that extends the work of SNCF [2] to web-noisy datasets where we improve the detection of OOD samples present in web-noise datasets by explicitly estimating the linear separation between ID and OOD samples. We demonstrate that this detection strategy is weakly correlated to existing small-loss and distance-based approaches.
2. An investigation into the disparity between noise retrieval performance and classification accuracy of noise-robust algorithms.
3. A novel noise correction approach, Linear Separation Alternating (LSA), that combines linear separation with uncorrelated SOTA noise detection.
4. A series of experiments and ablation studies, including a voting co-training strategy PLS-LSA+ that concurrently trains two models. We conduct these experiments on controlled and real-world web-noisy datasets to demonstrate the efficacy of our algorithm PLS-LSA.

2 Related work

Detection and correction of incorrect labels The most popular approach to tackle label noise is to explicitly detect ID samples with incorrect labels, either because they are harder to learn than their clean counterparts or because they are distant from same-class training samples in the feature space. Noise detection strategies include evaluating the training loss [5, 10, 23, 28], the Kullback-Leibler or Jensen-Shannon divergence between prediction and label [40], the entropy or the confidence of the prediction [3, 18] or the consistency of the prediction across epochs [35]. An alternative is to measure the distance between noisy and clean samples in the feature space: RRL [24] detects noisy samples as having many neighbors from different classes and NCR [16] regularizes training samples with similar feature representations to have similar predictions, reducing noisy label overfitting. We also note here that all recent label noise algorithms utilize the mixup [43] regularization which has proven to be highly robust to label noise [5].

While many noise-robust algorithms have proposed loss-based or distance-based noise detection metrics, the distinct advantages and biases of each strategy remain unexplored. Furthermore, considering that loss-based and distance-based detections are sufficiently decorrelated, combining the strengths of these distinct metrics is appealing, yet has not been previously explored. This paper observes the decorrelation of some noise detection metrics and proposes a non-trivial combination that improves generalization noise-robust algorithms over either metric taken independently.

Out-of-distribution noise in web-noisy datasets In web-noisy datasets, OOD (or *open-world*) noise is the dominant type [3]. Since ID noise is still present in small amounts in web-noisy datasets, algorithms propose concurrently detect ID and OOD noise. EvidentialMix [29] and DSOS [3] use specialised losses that exhibit three modes when evaluated over all training samples. Each of the modes are observed to mostly contain clean, OOD and ID noisy samples. A mixture of gaussians is then used to retrieve each noise type. SNCF [2] observed that unsupervised contrastive learning trained on a web-noisy dataset learns representations that are linearly separated between ID and OOD samples and use a clustering strategy based on OPTICS [4] to retrieve each noise type. The linear separability of in-distribution (ID) and out-of-distribution (OOD) representations noted in SNCF holds promise but has yet to be transitioned from synthetically corrupted to web-noisy data. This paper aims to address this gap.

Unsupervised learning and label noise Optimizing a noise robust (un)-supervised contrastive objective together with the classification loss can help improve the representation quality as well as detect OOD samples in the feature space. ScanMix learns SimCLR representations as a starting point for noise-robust contrastive clustering [30] and SNCF [2] observes linear separability on unsupervised iMix features [22]. Unsupervised contrastive features have also been used to initialize networks prior to noise-robust training in PropMix [10] and C2D [46] or used as a regularization to the supervised objective in RRL [24]. While unsupervised initialization or regularization has been employed to enhance

the generalization accuracy of networks trained under label noise, our primary focus lies in its ability to detect noisy samples before starting the supervised learning phase. We aim to enhance the linear separation observed in SNCF, particularly by extending it from synthetically corrupted to web-noisy datasets and by eliminating the requirement for clustering.

In this review, we find that unsupervised learning shows promise in identifying OOD images even before noise-robust supervised training begins. While many algorithms demonstrate an effective identification of OOD images in synthetically corrupted datasets, their generalization to web-noisy datasets is non-evident. Furthermore, although we observe a high correlation between loss-based and distance-based metrics, neither correlates directly with the linear detection observed in SNCF. We propose to evaluate the disparities between these metrics and to explore potential combinations to enhance noise detection, surpassing the capabilities of each metric taken independently.

3 Linear Separation Alternating (LSA)

This section details the contributions of this paper and the alternating noise detection strategy we use to combat label noise. We consider in this paper the case of a noisy web-noisy image dataset $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ of size N where the images $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ are associated with a classification label $\{\mathbf{y}_i\}_{i=1}^{N} \in \{1, \dots C\}$. We denote vectors with bold letters. The classification labels are expected to be possibly mis-assigned, i.e. incorrectly characterize the target object in the image they are assigned to (label noise). The clean or noisy nature of the training samples is unknown. Our goal is to learn an accurate classifier $\Phi(\mathbf{x})$ that performs an accurate classification despite the label noise present in $\mathcal{X}$. In our case, we consider that Φ is a neural network.

3.1 Identifying OOD images in web-noisy datasets

This section proposes to detect OOD images in web-noisy datasets by building on the detection of SNCF [2]. SNCF observes that an unsupervised algorithm trained on a web-noisy dataset containing OOD images will learn linearly separable representations for the ID and OOD samples in the dataset. While this is primarily an empirical observation, it was hypothesized to be a consequence of the uniformity and alignment principles of contrastive learning [38]. The alignment principle in contrastive learning encourages samples with similar visual features to cluster together while the uniformity principle encourages training samples to be uniformly distributed in the feature space. OOD images cannot satisfy the alignment principle since they are visually different from all other images in the dataset and are pushed by all ID samples on one side on hypersphere, becoming linearly separable from the ID samples [2]. As an aside, this hypothesis implies that the linear separation may not occur when training on visually similar out-of-sample images, a problem which we will revisit in the following sections. Importantly, the separability of ID and OOD samples only occurs for

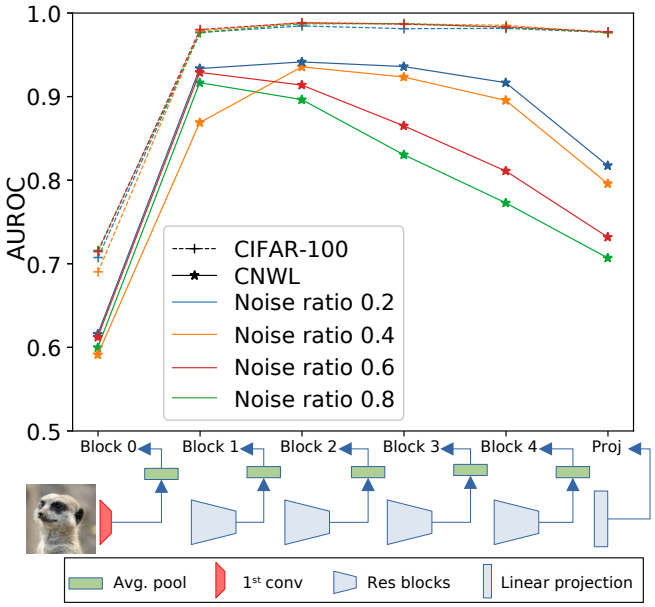

Fig. 1: Extending the work of [2] we observe that for web noise (CNWL), ID and OOD samples become more separable in earlier representations in the network

samples the unsupervised algorithm is trained on and cannot generalize to new unseen OOD images.

Although SNCF [2] observed the ID/OOD separation on synthetically corrupted datasets, i.e CIFAR-100 [20] corrupted by ImageNet32 [9], Figure 1 further described below shows that we do not observe as good a separation when moving to the web-noisy CNWL [18] but that the separability improves when looking at earlier representations. The weaker separability of OOD/ID images in web-noisy datasets compared to artificially corrupted datasets is explained by OOD images in web-noise datasets retaining weak semantic similarities with ID images. This is particularly true at the text level, which exhibited relevant similarities for the search engine during dataset creation. We propose that stronger separation occurs in low-level representations because they are more generic. Earlier representations easily align ID images of the same class due to shared low-level semantics, while OOD samples only become overfit in deeper layers, thus making separation increasly more difficult. A lesser corruption of earlier representations by label noise has for example been observed in [28].

Linear separation improves in deeper layers Our first contribution is then to observe that although the linear separation between ID and OOD is less evident in web-noisy datasets, it improves again when using earlier representations. Figure 1 gives an overview of the linear separability of ID and OOD images in the CNWL dataset [18] (web-noise) compared to the CIFAR-100 dataset [20] artificially corrupted with OOD images from ImageNet32 [9] using the unsu-

pervised algorithm SimCLR [8] to pre-train a PreActivation ResNet18 [15]. To compute the lower level features, we average-pool then L2 normalize representations at the end of each ResNet block. To compute the linear separability, we utilize the clean-noisy oracle to train a non-penalized logistic regressor to predict noise from the unsupervised features. The linear regressor is then evaluated on a held-out noisy test set previously unseen by the regressor. We report the area under the ROC curve (AUROC) for the logistic regressor to identify correct/incorrect training samples. Our train/test split for evaluation of the linear classifier comprises $45,000$ training and $5,000$ testing images, and is constructed from the full $50,000$ training images available for the overall classification task, all of which were used in unsupervised representation learning.

Estimating the linear separator A straight-forward approach to estimating the linear separation is to task human annotators to label randomly selected samples as ID or OOD, thus fulfilling the oracle role in Section 3.1. This strategy is usually referred to as learning to combat label noise with a trusted subset $\hat{\mathcal{T}} = \{\mathbf{x}_i, \hat{z}_i\}_{i=1}^{K}$ where $\hat{z}_i = 1$ means that the image $\mathbf{x}_i$ is OOD ($\hat{z}_i = 0$ for ID). Although we will show in the supplementary material that a good approximation for the linear separator can be achieved even given a small human-labeled subset of 100 images, most state-of-the-art noise robust algorithms do not rely on ID/OOD human annotations. We thus propose to estimate $\hat{\mathcal{T}}$ in an unsupervised manner. The unsupervised strategy of SNCF [2] is to use a clustering approach based on OPTICS [4]. We propose in this paper to avoid clustering and instead to train the linear separation using an unsupervised ID/OOD subset $\hat{\mathcal{T}}$.

We propose to build $\hat{\mathcal{T}}$ using unsupervised noise detection metrics $z(\mathbf{x}_i, \mathbf{y}_i) = \hat{z}_i$ [1,23,24,26,27]. We will examine recent examples of loss-based and distance-based noise detections later in the paper. Given $\hat{\mathcal{T}}$ estimated from $Z = \{z_i\}_{i=1}^{N}$, we can train the linear regressor to effectively refine the estimated noisiness of a training sample ($\mathbf{x}_i$) to $\mathcal{L}_r(\mathbf{x}_i, z_i) = w_i$ where $\mathcal{L}_r$ is a linear classification algorithm. Effectively, given the initial noise detection Z and the unsupervised contrastive features, we produce an improved one $W = \{w_i\}_{i=1}^{N}$, the linear-separation detection.

We find that, although an unsupervised $\hat{\mathcal{T}}$ contains detection errors, we still accurately estimate the linear separation due to the natural outlier robustness of linear classifiers. We additionally attempted to construct $\hat{\mathcal{T}}$ by selecting only the M most confidently clean/incorrect samples according to the metric z but found that it lead to a less accurate W.

3.2 Does better noise detection imply better classification?

We aim to quantify the accuracy benefits of W over loss-based or distance-based noise detection strategies. We select PLS [1] for the loss-based approach (small loss strategy as in [5, 10, 23, 40]) and RRL [24] for the distance-based approach (similar to [16,27,30]). To avoid interacting with complex noise-robust mechanisms, we employ an ignore-the-noise algorithm whereby we train on the detected clean samples only using a cross-entropy loss. To obtain the RRL and

Table 1: Using multiple noise detection metrics to train a naive noise-ignoring algorithm on the CNWL datset. We report noise retreival performance and classification accuracy. None signifies training without noise removal. We bold the **best results** and underline the worst, higher is better. Results averaged over 3 random seeds ± std

Noise ratio ↓	Metric →	None	PLS	RRL	W_{PLS}	W_{RRL}	Oracle
0.2	AUROC	–	66.4±0.2	58.4±0.2	**84.9±0.2**	**84.9±0.0**	100
	Clean recall	–	81.9±0.1	94.6±0.4	**89.3±0.1**	**89.3±0.1**	100
	Noise recall	–	50.9±0.4	22.3±0.4	**80.5±0.4**	**80.5±0.1**	100
	Accuracy	56.9±0.1	**59.5±0.1**	58.9±0.2	58.6±0.2	58.0±0.1	60.2±0.2
0.8	AUROC	–	58.1±0.1	54.8±0.2	**63.4±0.2**	62.3±0.1	100
	Clean recall	–	86.7±0.2	90.8±0.1	**95.1±0.4**	93.4±0.0	100
	Noise recall	–	29.5±0.1	18.8±0.3	**31.7±0.1**	31.2±0.1	100
	Accuracy	38.5±0.1	**45.1±0.2**	43.1±0.2	41.3±0.2	41.7±0.1	46.2±0.2

PLS detection, we train the algorithms using the official code and utilize the noise detection at the end of training. We then estimate W_{PLS} and W_{RRL} as detailed in Section 3.1 using unsupervised SimCLR features. Table 1 reports the results on the CNWL under 20% and 80% web noise where we report noise detection performance by computing an AUROC curve and the clean or noisy recall as well as the classification accuracy of the ignore-the-noise algorithm.

Surprisingly, we observe that although $W_{RRL/PLS}$ improves the noise metrics in terms of AUROC and noise recall, using it to detect the noise decreases the classification accuracy of Φ. This implies that W mis-identifies important samples needed to achieve high accuracy classification.

3.3 Clean samples missed by the linear separation

Following the observation of missing important samples in the previous section, we take a look at the clean images missed by W. We observe that missed clean samples predominantly represent the target object on a uniform background (typically black or white). In the context of ID and OOD separation through the alignment and uniformity of unsupervised contrastive learning presented in Section 3.1, this observation suggests that the unsupervised contrastive algorithm aligns uniformly colored background using this simple visual cue and independently of the ID or OOD class depicted. This problem is similar to the case where the OOD noise is structured, i.e. contains subsets of highly similar OOD images (humans holding OOD objects would be a common example).

Although W misses important examples, because these depict the target object with no distractors in the background, we suggest that they will easily be detected by the original SOTA noise detection metrics, biased toward detecting simple to fit or highly representative samples. In fact, we show in Figure 2 randomly selected clean examples missed by W_{PLS} most of which are correctly retrieved by PLS. Examples for the opposite scenario can be found in the supplementary material.

Fig. 2: Examples of clean samples missed by our linear separation W_{PLS} but correctly recovered (green) by a small loss approach, here PLS. 20% noise CNWL.

3.4 Linear Separation Alternating

Because PLS and RRL retreive samples that W misses, we aim to quantify the correlation between these noise detection metrics to justify their complementarity. We observe in Figure 3 that while the noise detection of RRL and PLS remain correlated during training (> 0.8 Pearson correlation) our linear separation W is much more decorrelated with either RRL or PLS (< 0.5). This low correlation further motivates the complementarity of W with SOTA noise detection approaches. We also notice that using either PLS or RRL for the trusted subset $\hat{\mathcal{T}}$ leads to very similar linear separation as W_{PLS} and W_{RRL} are highly correlated, explained by RRL and PLS being highly correlated to begin with.

To combine W and Z (PLS or RRL), we experiment with multiple combination strategies including voting or successive use (see Section 4.2). We find that alternating every epoch between W and PLS or RRL to be the better strategy. One dominant advantage of the alternating strategy is that it prevents forgetting one noise-detection over the other, effectively avoiding a form of confirmation bias [6] where mis-detections become hard to correct. We name this alternating noise detection strategy Linear Separation Alternating or LSA. Results comparing combination strategies are available in the experiments, Section 4.2.

3.5 PLS-LSA

LSA is independent from the noise-robust algorithm used whether it performs distance-based or loss-based noise detection. We choose to build on PLS [1], a

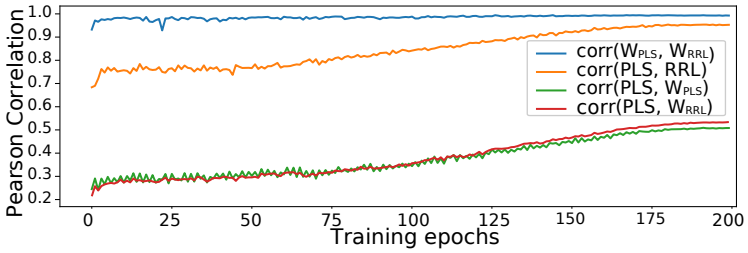

Fig. 3: Low correlation of our linear separation with the PLS and RRL metrics trained on CNWL with 20% web noise. $W_{PLS/RRL}$ denotes using PLS or RRL for $\hat{\mathcal{T}}$.

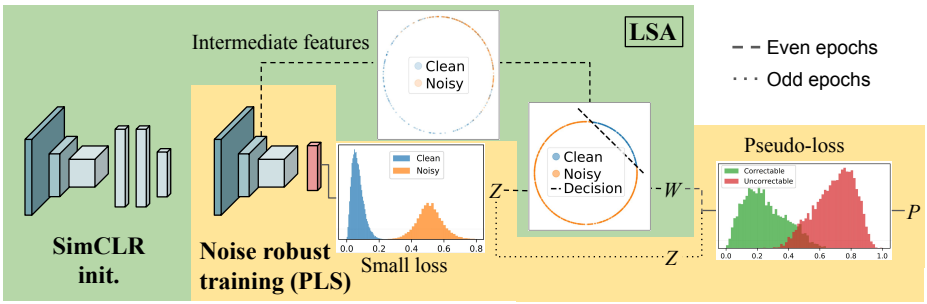

Fig. 4: Illustration of the noise detection of PLS with LSA (PLS-LSA). We use Z to estimate the linear separation W on even epochs.

semi-supervised strong baseline in web-noise robust algorithms. The following is a quick overview of the PLS algorithm [1]. In PLS, the network predicts two noisiness estimation metrics: a general noisiness $z(\mathbf{x}_i, \mathbf{y}_i) = z_i$ that estimates if a sample is clean $z_i = 1$ or noisy $z_i = 0$ (small loss based, using a two mode gaussian mixture [5, 23]) and the pseudo-loss prediction $p(\mathbf{x}_i, \tilde{\mathbf{y}}_i, z_i) = p_i$ that estimates whether a semi-supervised imputation $\tilde{\mathbf{y}}_i$ is a trustworthy correction for a noisy sample ($p_i = 1$). PLS optimizes 3 losses :

$$\mathrm{L}_{\mathtt{sup}}(\mathbf{x}_i, \mathbf{y}_i, z_i) = -z_i \times \mathbf{y}_i \times \log(\mathtt{softmax}(\Phi(\mathbf{x}_i))), \tag{1}$$

$$\mathrm{L}_{\mathtt{ssl}}(\mathbf{x}_i, \tilde{\mathbf{y}}_i, p_i) = -p_i \times \tilde{\mathbf{y}}_i \times \log(\mathtt{softmax}(\Phi(\mathbf{x}_i))) \tag{2}$$

and L_{cont} a supervised contrastive objective [22] that uses a SimCLR augmented view $\mathbf{x}'_i$ and is sensitive to p_i whose definition we refer to the original paper [1]. The final training loss in PLS without subscripts is

$$\mathrm{L}_{\mathtt{PLS}}(\mathbf{x}, \mathbf{x}', \mathbf{y}, z, p) = \mathrm{L}_{\mathtt{sup}}(\mathbf{x}, \mathbf{y}, z) + \mathrm{L}_{\mathtt{ssl}}(\mathbf{x}, \tilde{\mathbf{y}}, z, p) + \mathrm{L}_{\mathtt{cont}}(\mathbf{x}, \mathbf{x}', \mathbf{y}, p) \tag{3}$$

We call our version of PLS using LSA, PLS-LSA where we pretrain Φ using SimCLR and replace Z with W on even epochs, i.e. on even epochs we compute $p_i = p(\mathbf{x}_i, \tilde{\mathbf{y}}_i, w_i)$ and $L_{sup}(\mathbf{x}_i, \mathbf{y}_i, w_i)$. For W we use features extracted in the second ResNet block, an ablation can be found in the supplementary material.

3.6 Semi-supervised imputation and Co-training

Because PLS-LSA lacks some common additions to the recent noise-robust literature, we propose to use a stronger data augmentation for semi-supervised imputation and introduce an optional voting co-training strategy. We modify the PLS label imputation strategy as follows: given $\mathbf{x}_i''$ augmented using RandAugment [12], we modify the semi-supervised loss of PLS to

$$\mathrm{L}_{\mathtt{ssl}}(\mathbf{x}_i, \mathbf{x}_i'', \tilde{\mathbf{y}}_i, p_i) = -p_i \times \mathtt{sg}(\mathtt{softmax}(\Phi(\mathbf{x}_i))) \times \log(\mathtt{softmax}(\Phi(\mathbf{x}_i''))) \quad (4)$$

where $\mathtt{sg}(.)$ is the stop gradient operation. This imputation strategy is in line with recent semi-supervised classification research [31, 42].

PLS-LSA+ is a co-training strategy for PLS-LSA that uses two co-trained networks. We use a voting approach where the two networks vote for noisy samples detection z_i, w_i and p_i as well as for classification at test time. Our voting noise detection is different from previous approaches [14, 23, 30] where networks predict noisiness for each other. Additionally, before voting on p_i we introduce a co-guessing strategy where a semi-supervised prediction $\tilde{\mathbf{y}}_i$ of network 1 is evaluated as correct by network 2. The naive strategy would be each network evaluating if their own guess is correct which introduces more confirmation bias.

4 Experiments

4.1 Structure of the experiments section

We structure the experiment section as follows. First we study different combination strategies for PLS and W_{PLS}. We then conduct an ablation study of PLS-LSA to highlight the importance of each of our proposed addition over PLS. We finally compare PLS-LSA with SOTA algorithms on the Controlled Noisy Web Labels (CNWL) dataset [17] and real world datasets mini-Webvision [25] and Webly-fg [33]. The CNWL dataset corrupts miniImageNet [37] with human curated web-noisy examples. The dataset proposes noise ratios ranging from 20 to 80%. Following previous research, we train on the CNWL at a resolution of 32^2 using a PreActivation ResNet18 [15]. mini-WebVision is a subset of the first 50 classes of Webvision [25] which mimic ImageNet [21] classes. We train on mini-WebVision at a resolution of 224^2 using an InceptionResNetV2 [34]. The Webly-fg datasets are noisy datasets that target fined-grained classification of aircrafts, birds or cars. We train at a resolution of 448^2 using a ResNet50 initialized either on ImageNet as done in previous research or using SimCLR. All datasets contain unidentified web-noisy samples, which are either OOD or ID noisy (mislabeled). We compare the performance of noise-robust algorithms trained on web-noisy datasets by their ability to accurately classify a clean validation set. More detailed experimental settings are available in the supplementary material.

Our experimental settings are the same as used in PLS [1] and comparable to evaluation settings used in the algorithms we compare with. Unless otherwise specified, we initialize our networks using SimCLR [8] and solo-learn [36].

Table 2: Best strategy to combine PLS and W_{PLS} on the CNWL.

Noise ratio	0.2	0.4	0.6	0.8
PLS	62.30±0.24	59.11±0.28	54.26±0.20	48.71±0.36
W_{PLS}	62.97±0.23	60.41±0.05	52.18±0.12	47.11±0.08
AND	62.66±0.34	58.80±0.41	54.82±0.15	48.19±0.11
OR	58.79±0.40	58.27±0.34	50.57±1.11	45.93±0.69
W_{PLS} →PLS	62.88±0.46	59.89±0.14	52.34±0.12	48.88±0.23
PLS→ W_{PLS}	63.49±0.12	60.21±0.12	55.36±0.84	49.23±0.21
LSA	**64.20**±0.16	**60.98**±0.24	**55.64**±0.30	**49.73**±0.13
Oracle	64.10±0.10	61.45±0.22	56.04±0.39	50.19±0.48

Table 3: Ablation study CNWL

Dataset	mini 20%	mini 80%
Baselines		
mixup	57.27±0.39	38.48±0.24
mixup + SimCLR	57.03±0.10	39.62±0.28
PLS	62.83±0.39	45.80±0.72
PLS + SimCLR	62.39±0.14	47.21±0.63
PLS ours	63.25±0.24	48.03±0.38
PLS-LSA ablation		
PLS-LSA	64.61±0.51	48.20±0.16
PLS-LSA no SimCLR	64.04±0.27	47.29±0.27
PLS-LSA no DA	63.25±0.21	43.55±0.45
PLS-LSA no SimCLR DA	61.83±0.55	43.75±0.35
PLS-LSA+ ablation		
PLS-LSA+	66.52±0.10	52.03±0.32
PLS-LSA+ no SimCLR DA	66.34±0.27	47.68±0.59

Table 4: Ablation study Webvision

Algorithm	Webvision
mixup	77.99
mixup + SimCLR	78.88
PLS	79.01
PLS-LSA	81.36
PLS-LSA no SimCLR	78.68
PLS-LSA no DA	79.00
PLS-LSA no SimCLR DA	76.80

4.2 Combining PLS and W_{PLS}

We investigate here mulitple strategies for PLS-LSA combining the decorrelated W and PLS so that we can maximize classification accuracy on the held out validation set. We propose to use AND or OR logic operators (clean is false and noisy true), sucessive noise detection where we train using either metric for the first half of training and then switch to the other for the remainder (W_{PLS} →PLS and PLS→ W_{PLS}) or our alternating approach (LSA) where either metric is alternatively used every epoch. Table 2 displays our results.

We find that two strategies are superior to the PLS baseline: the second best strategy is PLS→ W_{PLS}, explained because the simple samples W_{PLS} misses are less important to get right in later training steps when the network has already learned strong base features for each class. The LSA strategy is the best approach overall, we believe this is because training the algorithm on both detection regularly allows to learn from the clean training examples provided by both metrics while avoiding over-fitting either metric's defects. These results solidify LSA as the better alternative for combining W and PLS.

Table 5: CNWL [18] (32×32). We run PLS and PLS-LSA; other results are from [1]. We report top-1 best accuracy and bold the best results with and without co-training. Accuracy results averaged over 3 random seeds ± one std.

	No co-training						Co-training					
Noise level	M	MM	FaMUS	SNCF	PLS	PLS-LSA	DM	SM	PM	LRM	MDM	PLS-LSA+
20	49.10	51.02	51.42	61.56	63.25±0.24	**64.43**±0.21	50.96	59.06	61.24	56.03±0.5	64.40	**66.52**±0.10
40	46.40	47.14	48.03	59.94	60.42±0.23	**61.14**±0.35	46.72	54.54	56.22	50.69±0.3	61.40	**63.42**±0.42
60	40.58	43.80	45.10	54.92	55.34±0.38	**57.18**±0.30	43.14	52.36	52.84	46.81±0.3	56.20	**59.41**±0.30
80	33.58	33.46	35.50	45.62	48.03±0.20	**49.53**±0.46	34.50	40.00	43.42	38.24±0.2	47.80	**52.03**±0.32

4.3 Ablation study

We conduct an ablation study to evaluate the importance of each of our design choices in Table 3. We first ablate on the CNWL dataset under 20% and 80% noise. We evaluate the improvements of SimCLR when added to a simple noise robust training using Mixup or the original PLS algorithm. Interestingly we observe that unsupervised initialization has little effect on validation accuracy for lower noise ratios. We also report PLS (ours) which denotes our improved version of PLS (PLS-LSA without LSA) which uses SimCLR initialization and improved data augmentations. Our version performs slightly better when compared to the original PLS. The second part of the table ablates elements from PLS-LSA: SimCLR initialization, stronger data augmentation (DA) or both (nothing). Strong data augmentations appears to be an important element of PLS-LSA. This is explained by our semi-supervised imputation strategy being largely dependent on stronger data augmentations whereas another SSL imputation strategie (i.e. MixMatch [7] used in PLS) would be better suited when not having access to stronger DA. Interestingly, PLS-LSA does not catastrophically fail when we remove the SimCLR initialization. This hints towards observing the linear separation without self-supervised pre-training and shows that the alternating strategy provides stability though to the original PLS detection. We finally observe that PLS-LSA+ nothing manages to use co-training to maintain a high accuracy in the lower noise scenario even if we remove SimCLR initialization and strong DA.

We additionally run ablations experiments on mini-Webvision to measure impacts in the real world. Results are available in Table 4. In this context, SimCLR initialization appears to play a more important role than on the CNWL and is important to maintain a good classification accuracy with PLS-LSA.

4.4 SOTA comparison on the CNWL dataset

We compare with related SOTA on the CNWL dataset corrupted with $20, 40, 60, 80\%$ web noise in Table 5. We report the accuracy results of both PLS-LSA and PLS-LSA+. The noise-robust algorithms we compare with are Mixup [43] a noise robust regularization, FaMUS [39] a meta learning approach and sample correction algorithms: DivideMix (DM) [23], MentorMix (MM) [18], ScanMix [30], PropMix (PM) [10], SNCF [2], LongReMix (LRM) [11], Manifold DivideMix

Table 6: Classification accuracy for training on mini-Webvision using InceptionResNetV2. We denote with † algorithms using unsupervised initialization. We test on the mini-Webvision valset and ImageNet 1k test set (ILSVRC12). We run PLS and PLS-LSA, other results are from SNCF [2]. We bold the best results. Accuracy results averaged over 3 random seeds ± one std.

		No co-training							Co-training						
Testset		M	MM	RRL	FaMUS	PLS	FLY	†PLS-LSA	DM	ELR+	DSOS	†SM	RM	SNCF+	†PLS-LSA+
mini-WebVision	top-1	75.44	76.0	77.80	79.40	79.01$_{\pm 0.33}$	80.96	**81.28**$_{\pm 0.11}$	77.32	77.78	78.76	80.04	79.91	80.24	**82.08**$_{\pm 0.28}$
	top-5	90.12	90.2	91.30	92.80	92.05$_{\pm 0.46}$	93.56	**94.12**$_{\pm 0.09}$	91.64	91.68	92.32	93.04	93.61	93.44	**94.16**$_{\pm 0.10}$
ILSVRC12	top-1	71.44	72.9	74.40	77.00	76.15$_{\pm 0.35}$	– –	**78.32**$_{\pm 0.74}$	75.20	70.29	75.88	75.76	77.39	77.12	**79.19**$_{\pm 0.52}$
	top-5	89.40	91.10	90.90	92.76	92.53$_{\pm 0.23}$	– –	**94.64**$_{\pm 0.20}$	90.84	89.76	92.36	92.60	94.26	94.32	**94.84**$_{\pm 0.25}$

(MDM) [13] and PLS [1]. We find that PLS-LSA improves over existing approaches even when these use a co-training strategy (PLS-LSA only uses one network). PLS-LSA+ further improves the classification accuracy by 2 to 5 absolute points across noise levels.

4.5 Real world datasets

We now evaluate PLS-LSA on real world datasets. For mini-Webvision we add to the comparison a robust loss algorithm Early Learning Regularization (ELR) [26], as well as additional sample correction algorithms: Robust Representation Learning (RRL) [24], DSOS [3], RankMatch (RM) [45] and LNL-Flywheel (FLY) [19]. We also report our results on the webly fine-grained datasets as well as for Co-teaching [14], PENCIL [41], SELFIE [32], Peer-learning [33] and Progressive Label Correction (PLC) [44] which are all sample correction algorithms.

mini-Webvision We train PLS-LSA on mini-Webvision and report test results on the validation set of mini-Webvision and well as on the validation set of ImageNet2012 [21] in Table 6. We outperform co-training methods with PLS-LSA using only one network and PLS-LSA+ sets a new state-of-the-art by improving over PLS-LSA in terms of top-1 accuracy but we notice no significant improvements for top-5 accuracy. We report additional results when training a ResNet50 on mini-Webvision in the supplementary material where we observe similar improvements of PLS-LSA and PLS-LSA+ when compared to related works.

Webly-fg datasets We train PLS-LSA on the Webly-fg datasets [33] that present the added challenge of fine-grained classification over mini-Webvision. We report results on the bird, car and aircraft subsets in Table 7. Because other methods use ImageNet weights for pre-training, we report results using either ImageNet or SimCLR pretraining to exhibit the linear separation between ID and OOD noise. We find that PLS-LSA only marginally improves over PLS even in the case where we use self-superivsed features. We found that learning strong SimCLR features for Webly-fg datasets is challenging due to the fine grained nature of the dataset. It could be the case that using a different set of data augmentations or a different self-supervised algorithm would help improve our performance further. PLS-LSA+ improves 0.7 to 0.8 points over PLS-LSA.

Table 7: Comparison against state-of-the-art algorithms on the fine grained web datasets, we run PLS-LSA and bold the best results. Results for other algorithms from [1]. Top-1 best accuracy.

Initialization	Algorithm	Web-Aircraft	Web-bird	Web-car
ImageNet	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
	PLS-LSA	87.70	79.20	86.58
	PLS-LSA+	88.42	79.77	87.24
SimCLR	PLS-LSA	87.82	79.47	86.76
	PLS-LSA+	**88.51**	**80.03**	**87.50**

5 Conclusion

This paper builds on the previously observed linear separation of ID and OOD images in unsupervised contrastive feature spaces in the context of label noise datasets. We observe that the linear separation of ID and OOD features is not as evident as previously observed when moving to real-world data yet becomes apparent again when looking at lower level features. Instead of relying on clustering as done in previous research, we propose to compute the linear separation using an approximate ID/OOD detection using state-of-the-art noise-robust metrics. Although we find our noise detector to be highly accurate, we do not observe classification accuracy gains when compared to less accurate SOTA noise detectors. We evidence that the few samples we mis-identify are crucial to train a strong classifier. We combine our detection together with PLS by alternating the noise detection approach every epoch to create PLS-LSA. We further develop a co-train schedule using two networks to produce PLS-LSA+. Our results improve the SOTA classification accuracy on real-world web noise datasets. Because we only empirically observe the linear separation in earlier layers, we stress the need for further theoretical analysis of the phenomenon and encourage further research in this direction. Other future work we recommend is to study if intelligent alternating strategies could be developed to combine both detection approaches based on the current noise detection bias in the network. We also suggest that further attention be given to whether the linear separation can be enforced from a random initialization and as training progresses to remove the need for pretraining.

Acknowledgments This publication has emanated from research conducted with the joint financial support of the Center for Augmented Reasoning (CAR) and Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289_P2. The authors additionally acknowledge the Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support. The authors would like to issue special remembrance to our dearly missed friend and colleague Kevin McGuinness for his invaluable contributions to our research.

References

1. Albert, P., Arazo, E., Krishna, T., O'Connor, N.E., McGuinness, K.: Is your noise correction noisy? PLS: Robustness to label noise with two stage detection. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023)
2. Albert, P., Arazo, E., O'Connor, N.E., McGuinness, K.: Embedding contrastive unsupervised features to cluster in-and out-of-distribution noise in corrupted image datasets. In: European Conference on Computer Vision (ECCV) (2022)
3. Albert, P., Ortego, D., Arazo, E., O'Connor, N., McGuinness, K.: Addressing out-of-distribution label noise in webly-labelled data. In: Winter Conference on Applications of Computer Vision (WACV) (2022)
4. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: Optics: Ordering points to identify the clustering structure. ACM Sigmod record **28**(2), 49–60 (1999)
5. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Unsupervised Label Noise Modeling and Loss Correction. In: International Conference on Machine Learning (ICML) (2019)
6. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In: International Joint Conference on Neural Networks (IJCNN) (2020)
7. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: MixMatch: A Holistic Approach to Semi-Supervised Learning. In: Advances in Neural Information Processing Systems (NeuRIPS) (2019)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: International Conference on Machine Learning (ICML) (2020)
9. Chrabaszcz, P., Loshchilov, I., Hutter, F.: A downsampled variant of imagenet as an alternative to the cifar datasets. arXiv: 1707.08819 (2017)
10. Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: PropMix: Hard Sample Filtering and Proportional MixUp for Learning with Noisy Labels. arXiv: 2110.11809 (2021)
11. Cordeiro, F.R., Sachdeva, R., Belagiannis, V., Reid, I., Carneiro, G.: Longremix: Robust learning with high confidence samples in a noisy label environment. Pattern Recognition (2023)
12. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020)
13. Fooladgar, F., To, M.N.N., Mousavi, P., Abolmaesumi, P.: Manifold DivideMix: A Semi-Supervised Contrastive Learning Framework for Severe Label Noise. arXiv:2308.06861 (2023)
14. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision (ECCV) (2016)
16. Iscen, A., Valmadre, J., Arnab, A., Schmid, C.: Learning With Neighbor Consistency for Noisy Labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
17. Jiang, L., Zhou, Z., Leung, T., Li, L., Fei-Fei, L.: MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In: International Conference on Machine Learning (ICML) (2018)

18. Jiang, L., Huang, D., Liu, M., Yang, W.: Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels. In: International Conference on Machine Learning (ICML) (2020)
19. Kim, H., Chang, H.S., Cho, K., Lee, J., Han, B.: Learning with Noisy Labels: Interconnection of Two Expectation-Maximizations. arXiv: 2401.04390 (2024)
20. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
21. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (NeurIPS) (2012)
22. Lee, K., Zhu, Y., Sohn, K., Li, C.L., Shin, J., Lee, H.: i-Mix: A Strategy for Regularizing Contrastive Representation Learning. In: International Conference on Learning Representations (ICLR) (2021)
23. Li, J., Socher, R., Hoi, S.: DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In: International Conference on Learning Representations (ICLR) (2020)
24. Li, J., Xiong, C., Hoi, S.C.: Learning from noisy data with robust representation learning. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
25. Li, W., Wang, L., Li, W., Agustsson, E., Van Gool, L.: WebVision Database: Visual Learning and Understanding from Web Data. arXiv: 1708.02862 (2017)
26. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-Learning Regularization Prevents Memorization of Noisy Labels. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
27. Ortego, D., Arazo, E., Albert, P., O'Connor, N.E., McGuinness, K.: Multi-Objective Interpolation Training for Robustness to Label Noise. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
28. Ortego, D., Arazo, E., Albert, P., O'Connor, N.E., McGuinness, K.: Towards robust learning with different label noise distributions. In: International Conference on Pattern Recognition (ICPR) (2021)
29. Sachdeva, R., Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: EvidentialMix: Learning with Combined Open-set and Closed-set Noisy Labels. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020)
30. Sachdeva, R., Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: ScanMix: learning from severe label noise via semantic clustering and semi-supervised learning. Pattern Recognition (2023)
31. Sohn, K., Berthelot, D., L, C.L., Zhang, Z., Carlini, N., Cubuk, E., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. arXiv: 2001.07685 (2020)
32. Song, H., Kim, M., Lee, J.G.: SELFIE: Refurbishing Unclean Samples for Robust Deep Learning. In: International Conference on Machine Learning (ICML) (2019)
33. Sun, Z., Yao, Y., Wei, X.S., Zhang, Y., Shen, F., Wu, J., Zhang, J., Shen, H.T.: Webly Supervised Fine-Grained Recognition: Benchmark Datasets and An Approach. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
34. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Association for the Advancement of Artificial Intelligence (AAAI) (2016)
35. Toneva, M., Sordoni, A., Combes, R., Trischler, A., Bengio, Y., Gordon, G.: An empirical study of example forgetting during deep neural network learning. In: International Conference on Learning Representations (ICLR) (2019)

36. Victor Guilherme Turrisi da Costa and Enrico Fini and Moin Nabi and Nicu Sebe and Elisa Ricci: solo-learn: A library of self-supervised methods for visual representation learning. Journal of Machine Learning Research (2022)
37. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching Networks for One Shot Learning. In: Advances in Neural Information Processing Systems (NeuRIPS) (2016)
38. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning (ICLR) (2020)
39. Xu, Y., Zhu, L., Jiang, L., Yang, Y.: Faster meta update strategy for noise-robust deep learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
40. Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., Tang, Z.: Jo-SRC: A Contrastive Approach for Combating Noisy Labels. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
41. Yi, K., Wu, J.: Probabilistic End-to-end Noise Correction for Learning with Noisy Labels. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
42. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
43. Zhang, H., Cisse, M., Dauphin, Y., Lopez-Paz, D.: mixup: Beyond Empirical Risk Minimization. In: International Conference on Learning Representations (ICLR) (2018)
44. Zhang, Y., Zheng, S., Wu, P., Goswami, M., Chen, C.: Learning with feature-dependent label noise: A progressive approach. In: International Conference on Learning Representations (ICLR) (2021)
45. Zhang, Z., Chen, W., Fang, C., Li, Z., Chen, L., Lin, L., Li, G.: RankMatch: Fostering Confidence and Consistency in Learning with Noisy Labels. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
46. Zheltonozhskii, E., Baskin, C., Mendelson, A., Bronstein, A.M., Litany, O.: Contrast to divide: Self-supervised pre-training for learning with noisy labels. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2022)