Contrastive Learning Improves Model Robustness Under Label Noise

Aritra Ghosh Andrew Lan
University of Massachusetts Amherst
{arighosh,andrewlan}@cs.umass.edu

Abstract

Deep neural network-based classifiers trained with the categorical cross-entropy (CCE) loss are sensitive to label noise in the training data. One common type of method that can mitigate the impact of label noise can be viewed as supervised robust methods; one can simply replace the CCE loss with a loss that is robust to label noise, or re-weight training samples and down-weight those with higher loss values. Recently, another type of method using semi-supervised learning (SSL) has been proposed, which augments these supervised robust methods to exploit (possibly) noisy samples more effectively. Although supervised robust methods perform well across different data types, they have been shown to be inferior to the SSL methods on image classification tasks under label noise. Therefore, it remains to be seen that whether these supervised robust methods can also perform well if they can utilize the unlabeled samples more effectively. In this paper, we show that by initializing supervised robust methods using representations learned through contrastive learning leads to significantly improved performance under label noise. Surprisingly, even the simplest method (training a classifier with the CCE loss) can outperform the state-of-the-art SSL method by more than 50% under high label noise when initialized with contrastive learning. Our implementation will be publicly available at https://github.com/arghosh/noisy_label_pretrain.

1. Learning under Label Noise

In standard classification tasks, we are given a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ where $\mathbf{x}_i$ is the feature vector of the i^{th} sample (image) and $\mathbf{y}_i \in \{0,1\}^K$ is the class label vector with K total classes. We minimize the following empirical risk minimization (ERM) objective,

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{CCE}}(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})), \tag{1}$$

where $f(\cdot; \mathbf{w})$ is a deep neural network (DNN)-based classifier with parameters $\mathbf{w}$ and ℓ_{CCE} is the categorical cross-entropy (CCE) loss. However, real-world datasets often contain noisy labels; i.e., $\mathbf{y}_i$ can be corrupted. DNNs are sensitive to label noise when they are trained with the CCE loss, which reduces their ability to generalize to clean dataset.

Most of the early works on learning under label noise can be called as *supervised robust methods* and they are equally applicable to image, text, or any other data types. A general trick to mitigate the impact of label noise is to replace the CCE loss function ℓ_{CCE} with a loss that is more robust to label noise in Eq. 1 [39, 25, 26, 13, 14, 11, 32, 25, 35]. In [11], the authors show that a loss function ℓ is robust to uniform label noise if it satisfies the condition $\sum_{k=1}^{K} \ell(k, f(\mathbf{x}_i; \mathbf{w})) = C$ for some constant C. The mean absolute error (MAE) loss satisfies this symmetric condition; however, the MAE loss is difficult to optimize under the ERM objective with DNNs. Several loss functions have been proposed that offer model robustness under label noise and they are easier to optimize compared to the MAE loss [25, 35, 39, 26]. For example, the generalized cross-entropy loss L_q ($q \in (0,1]$ is a hyper-parameter) is defined as [39] $L_q(\mathbf{y}, f(\mathbf{x}; \mathbf{w})) = \frac{1-\mathbf{y}^\intercal f(\mathbf{x};\mathbf{w})^q}{q}$. The L_q loss is equivalent to the CCE loss when $q \to 0$ and is equivalent to the MAE loss when $q = 1$. However, these robust loss functions do not perform well on large image datasets.

Another common strategy for learning under label noise is to separate out the noisy samples from the clean samples or re-weight the training samples and stick with the CCE loss; we can simply change the objective as

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{W}(\mathbf{x}_i, \mathbf{y}_i) \ell_{\text{CCE}}(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})), \tag{2}$$

where $\mathcal{W}(\mathbf{x}_i, \mathbf{y}_i) \in [0,1]$ is the assigned weight for the training sample $(\mathbf{x}_i, \mathbf{y}_i)$. A common heuristic, studied in earlier research, is that noisy samples have higher loss values compared to the clean samples [4, 3]. Many recent methods apply this idea to filter out or lower weights to possibly noisy samples [15, 17, 31, 7, 17, 20, 31, 33, 37, 23, 28, 16]. Instead of filtering samples based on the loss value, a more principled way is to *learn* a weighting *function* $\mathcal{W}(\ell(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})); \theta)$ in a data-driven manner where the function takes the loss value as the input. Meta-learning-based methods have been particularly useful to learn a weighting function [30, 12, 29, 22]. As an example, Meta-Weight Network (MWNet) [30] learns a weighting function

$\mathcal{W}$ with parameter θ using a small number of clean validation samples in a bilevel setup [10]. The objective is to learn the optimal weighting function $\mathcal{W}(\ell(\cdot,\cdot);\theta^*)$ such that the optimal classifier parameters $\mathbf{w}^*(\theta^*)$ on the training samples (train), obtained from Eq. 2, optimizes the ERM objective on the clean validation samples (val). The bilevel optimization problem can be written as

$$\min_{\theta} \sum_{j \in \text{val}} \ell\Big(\mathbf{y}_j, f(\mathbf{x}_j; \mathbf{w}^*(\theta))\Big)$$
$$\text{s.t. } \mathbf{w}^*(\theta) = \underset{\mathbf{w}}{\text{argmin}} \sum_{i \in \text{train}} \mathcal{W}\Big(\ell(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})); \theta\Big) \ell(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})).$$

These *supervised robust methods* perform well across many data types.

Recently, many semi-supervised learning (SSL) methods have been proposed for image datasets to mitigate the impact of label noise. SSL methods aim to improve the performance of a DNN classifier by exploiting unlabeled data [2]. Common tricks in SSL methods include using a consistency regularization loss to encourage the classifier to have similar predictions for an image $\mathbf{x}_i$ and the augmented view of the image $\text{Aug}(\mathbf{x}_i)$, an entropy minimization objective to promote high confidence predictions, and a label guessing method to produce a good guess from many augmentations of the same image [2, 38]. DivideMix, an SSL method, divides the training dataset into the clean (labeled) and noisy (unlabeled) parts using the observation that noisy samples tend to have a higher loss value [21]. These SSL methods have been shown to be superior to the supervised robust methods on image datasets [24, 21].

1.1. Contributions

We observe that SSL methods for label noise can use unlabeled (noisy) samples effectively to improve their representation learning capability. Consequently, prior supervised robust learning methods suffer a significant drop in performance compared to the SSL methods on image datasets. Hence, we ask the following question:

- Is the performance drop of supervised robust methods caused by label noise or the impaired representation learned using fewer clean samples?

Thus, we study the effect of fine-tuning these supervised robust methods after initializing them with *good* representations learned by a self-supervised method. Contrastive learning has emerged as a key method for self-supervised learning from visual data; the general idea is to learn good visual representations of images through comparing and contrasting different views of an original image under various data augmentation operations [5, 6]. We find that the supervised robust methods work remarkably well when they are initialized with the contrastive representation learning model. Surprisingly, we notice that even using the (most sensitive) CCE loss can outperform state-of-the-art SSL methods under high label noise. Moreover, we observe that the generalized cross-entropy loss [39] can retain good performance even under 95% uniform label noise on the CIFAR-100 dataset whereas training with a random initializer does not outperform a random model. These observations suggest that the drop in performance for the supervised robust methods is due to the lack of good visual representations. We use one representative method from each of the two major paradigms we described for the supervised robust methods (the L_q loss for the loss correction approach and the MWNet method for the sample re-weighting strategy) to illustrate the benefits of fine-tuning representations learned through contrastive learning with a classification task under label noise.

1.2. Related Works

The idea of using a pre-trained model initializer or self-supervised learning is not new in label noise research. In [19], the authors use auxiliary tasks, such as rotation prediction, to improve model robustness under label noise. In [18], the authors propose to use a pre-trained Imagenet classifier to improve model robustness. These methods lead to improved performance under high label noise, adversarial perturbation, class imbalance conditions, and on out-of-distribution detection tasks. However, they require a larger similar dataset where label noise is not present or need to use an auxiliary loss from the self-supervised tasks in addition to the classification task. In contrast, our work does not propose any additional auxiliary tasks or require any larger datasets. We learn the contrastive model for visual representations from the same dataset as the classification task. This is helpful when the classification task uses datasets (e.g., in medical imaging datasets) that are very different than the commonly used large-scale image datasets (e.g., the ImageNet dataset). The most related work is [9], which uses a contrastive model to improve the DivideMix algorithm. However, we show that a self-supervised contrastive learning model initializer can improve model robustness under label noise for many supervised robust methods.

1.3. Methodology

We will use the SimCLR framework for contrastive learning [5, 6]; however, other visual representation learning methods (including other contrastive learning methods) can also potentially improve model robustness under label noise. We use a base encoder $\hat{f}(\cdot)$ (ResNet-50 in this paper) to encode each image $\mathbf{x}_i$ to $\mathbf{h}_i = \hat{f}(\mathbf{x}_i)$, and a two-layer multilayer perceptron $g(\cdot)$ as the projection head to project into a fixed dimension embedding $\mathbf{z}_i = g(\mathbf{h}_i)$. Using M images and two augmentations for each image, we construct a dataset of $2M$ images $\{\mathbf{x}_{i,0}, \mathbf{x}_{i,1}\}_{i=1}^{M}$ and project them into $\{\mathbf{z}_{i,0}, \mathbf{z}_{i,1}\}_{i=1}^{M}$ using the base encoder and the projection head. The final objective in the SimCLR framework is defined as

$$\sum_{i=1}^{M}\sum_{j=0}^{1}-\log\frac{\exp\left(\text{sim}(\mathbf{z}_{i,j},\mathbf{z}_{i,j+1\%2})/\tau\right)}{-\exp\left(1/\tau\right)+\sum_{k=1,l=0}^{k=M,l=1}\exp(\text{sim}(\mathbf{z}_{i,j},\mathbf{z}_{k,l})/\tau)},$$

where τ is the temperature parameter, and $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ is the normalized cosine similarity $\frac{\mathbf{z}_i^\intercal \mathbf{z}_j}{||\mathbf{z}_i||||\mathbf{z}_j||}$. We use the same dataset $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ to learn the SimCLR encoder $\hat{f}(\cdot)$. The base encoder $\hat{f}$ does not contain the classification head (last output layer). For supervised robust methods, we use this encoder $\hat{f}$ to initialize the DNN classifier $f(\cdot;\mathbf{w})$ and we set the weights and biases of the classification head of $f(\cdot;\mathbf{w})$ to zero at initialization. Note that we fine-tune the final classifier $f(\cdot)$ for each method and do not keep the base encoder $\hat{f}(\cdot)$ fixed.

2. Experimental Results

Datasets and Experimental Setup: We demonstrate the efficacy of our proposed approach on CIFAR-10, CIFAR-100, and Clothing1M datasets. Unless otherwise specified, we use ResNet-50 (RN-50) as the classifier; for CIFAR datasets, we adopt the common practice of replacing the first convolutional layer of kernel size 7, stride 2 with a convolutional layer of kernel size 3 and stride 1 and removing the first max-pool operation in RN-50 [5].

CIFAR-10 and CIFAR-100 datasets contain 50k training samples and 10k test samples; label noise is introduced synthetically on the training samples. We keep 1000 clean training samples for validation purposes. We experiment with symmetric noise and asymmetric noise. Under symmetric noise, the true class label is changed to any of the class labels (including the true label) whereas, under asymmetric noise, the true class label is changed to a similar class label. We use the exact same setup of [39, 27] for introducing asymmetric noise. For CIFAR-10, the class mappings are TRUCK $\rightarrow$ AUTOMOBILE, BIRD $\rightarrow$ AIRPLANE, DEER $\rightarrow$ HORSE, CAT $\leftrightarrow$ DOG. For CIFAR-100, the class mappings are generated from the next class in that group (where 100 classes are categorized into 20 groups of 5 classes).

Clothing1M dataset is real-world datasets consisting of 1M training samples; labels are generated from surrounding text in an online shopping website [34]. Clothing1M dataset contains around 38% noisy samples [9] and we do not introduce any additional noise on this dataset.

Pre-Training: We compare the supervised robust methods using two initialization, namely the SimCLR initializer and the ImageNet pre-trained initializer [18]. To train the SimCLR encoder $\hat{f}(\cdot)$ and the projection head $g(\cdot)$, we use a batch size of 1024 (1200) and run for 1000 (300) epochs with the LARS optimizer [36] on a single NVIDIA RTX8000 (12 NVIDIA M40) GPU(s) on the CIFAR-10/100 (Clothing1M) datasets. For the Clothing1M dataset, we use the standard pre-trained ImageNet RN-50 initialization. For the CIFAR datasets, we train a RN-50 classifier from scratch (with CIFAR changes in the first convolutional layer) on the ImageNet-32 $\times$ 32 (INet32) dataset [8] that achieves 43.67% Top-1 validation accuracy.

Methods: We use the SimCLR RN-50 initializer for three methods: standard ERM training with the CCE loss, ERM training with the generalized cross-entropy loss L_q (q=0.5 or 0.66), and MWNet. The value of q in L_q loss is important only for a high rate of noise (≥ 0.8), where the L_q loss with a large q (0.66) is difficult to optimize; however, for all other noise rates, a higher value of q leads to better performance. We fine-tune for 120 epochs on the CIFAR datasets with the SGD optimizer, learning rate of 0.01, momentum of 0.9, weight-decay of 0.0001, and a batch size of 100. For other baseline methods, we use the results listed in their respective paper (or their public implementation). Note that different prior works use different architectures; thus, we also list the test accuracy from training on clean samples with that architecture (and initialization). For CIFAR datasets, we list the average accuracy from five runs of noisy label generation.

We use the CCE loss, the L_q loss (q=0.66), and the MAE loss with the SimCLR initializer on the Clothing1M dataset. We use the SGD optimizer, a batch size of 32, momentum of 0.9, and an initial learning rate of 0.001. Following [24, 21], we randomly sample 4000x32 training samples in each epoch such that the total number of samples from each of the classes are equal. We fine-tune for 60 epochs and reduce the learning rate by a factor of 2 after every 10 epochs.

2.1. Results and Discussion

Table 1 lists classification performance on the test set under symmetric noise on the CIFAR datasets. The SimCLR initializer significantly improves performance for the CCE loss, the L_q loss, and the MWNet method. Under 90% label noise, the CCE loss has an accuracy of 42.7% (10.1%) with a random initializer and DivideMix has an accuracy of 93.2% (31.5%) on the CIFAR-10 (CIFAR-100) dataset. Under the same noise rate, the CCE loss with the SimCLR initializer has an accuracy of 82.9% (52.11%) on the CIFAR-10 (100) dataset which translates to a 9% (65%) gain compared to the state-of-the-art method DivideMix. Moreover, the SimCLR initializer beats these performances even further with the MWNet method and the L_q loss. Under very high levels of label noise, MWNet and the L_q loss are not able to learn anything useful with the standard random initializer. However, with the SimCLR initializer, these methods perform significantly better than the state-of-the-art method.

Table 2 lists the classification performance on the test set under asymmetric label noise on the CIFAR datasets. Similarly, we observe that the SimCLR initializer improves model robustness under asymmetric label noise. However, supervised robust methods do not beat the prior-state-of-the-art method for the asymmetric noise case.

Table 3 lists the test performance on the Clothing1M dataset. Although the CCE loss, the L_q loss, and the MAE loss do not outperform state-of-the-art methods with the

Noise Rate (%)			0	20	40	50	60	80	90	95	0	20	40	50	60	80	90	95
Method	Arch	Initializer	CIFAR-10								CIFAR-100							
CCE [21]	PRN-18	No	95.11	86.8	-	79.4	-	62.9	42.7	-	74.4	62.0	-	46.7	-	19.9	10.1	-
MLNT [22, 21]				92.9	-	89.3	-	77.4	58.7	-		68.5	-	59.2	-	42.4	19.5	-
F-Correction [27, 21]				86.8	-	79.8	-	63.3	42.9	-		61.5	-	46.6	-	19.9	10.2	
M-Correction [1, 21]				94.0	-	92.0	-	86.8	69.1	-		73.9	-	66.1	-	48.2	24.3	-
Divide-Mix [21]	PRN-18	No	95.8*	96.1	-	94.6	-	**93.2**	76	-	78.9*	77.3	-	74.6	-	60.2	31.5	-
MWNet [30]	WRN-28-10	No	95.60	92.45	89.27	87.49	84.07	69.65	25.8	18.49	79.95	73.99	67.73	66.88	58.75	30.55	5.25	3.05
L_q [39]	RN-34	No	93.34	89.83	87.13	-	82.54	64.07	-	-	76.76	66.81	61.77	-	53.16	29.16	-	-
ELR [24]				92.12	91.43		88.87	80.69	-	-		74.68	68.43	-	60.05	30.27	-	-
CCE	RN-50	INet32	**96.52**	92.37	-	91.56	-	83.34	62.66	39.09	**81.51**	70.26	-	65.76	-	54.33	38.9	20.59
L_q				94.07	-	93.75	-	90.56	85.89	73.14		77.22	-	69.87	-	60.5	54.83	44.3
MWNet				**97.33**	-	**96.17**	-	93.12	**90.88**	85.27		**82.85**	-	**80.28**	-	**71.29**	**58.21**	44.62
CCE	RN-50	SimCLR	94.59	93.29	-	91.96	-	88.75	82.9	66.07	75.36	71.98	-	67.89	-	59.84	52.11	39.57
L_q				94.02	-	92.94	-	90.85	88.45	83.76		73.33	-	70.14	-	63.26	55.93	**45.7**
MWNet				93.88	-	92.92	-	91.51	90.19	**87.23**		73.2	-	69.88	-	64.05	57.6	44.91

Table 1. Test accuracy (%) for various methods on the CIFAR datasets under symmetric label noise. We re-use results for the 'No' initializer cases from their respective papers. We re-run public implementation of MWNet [30] for some (missing) noise levels. The test accuracy under 0% label noise refers to the accuracy obtained from minimizing the ERM objective with the CCE loss except for DivideMix (*) for which test accuracy is obtained from training the MixUp objective with Preactivated ResNet-18 (PRN-18) [38]. We bold performance for the best method under each noise settings.

Noise Rate (%)			0	20	30	40	0	20	30	40
Method	Arch	Initializer	CIFAR-10				CIFAR-100			
CCE [39]	RN-34	No	93.34	88.59	86.14	80.11	76.76	59.20	51.40	42.74
F-correction[27, 39]				90.35	89.25	88.12		71.08	70.76	70.82
L_q[39]				89.33	85.45	76.74		66.59	61.45	47.22
ELR [24]				93.28	92.70	90.35		74.20	74.02	**73.73**
MWNet [30]	WRN-28-10	No	95.60	93.14	91.45	89.71	79.95	71.55	66.07	56.05
CCE	RN-50	INet32	**96.52**	92.93	91.78	90.22	**81.51**	69.76	62.41	52.4
L_q				93.77	93.23	90.24		71.63	67.29	59.29
MWNet				**96.85**	**95.9**	**94.99**		**80.74**	**77.36**	72.93
CCE	RN-50	SimCLR	94.59	93.3	92.13	88.38	75.36	69.63	63.91	54.28
L_q				93.54	92.69	90.27		71.26	68.04	59.26
MWNet				93.67	93.18	92.59		72.17	69.86	64.92

Table 2. Test accuracy (%) for various methods on the CIFAR datasets under asymmetric label noise. We re-use results for the 'No' initializer cases from their respective papers except for MWNet for which we run their public implementation with asymmetric noise.

Method	Initializer	Accuracy
CCE [21]	ImageNet	68.94
F-Correction [27]		69.84
MLNT [22]		73.47
DivideMix [21]		74.76
ELR [24]		72.87
ELR+ [24]		74.81
CCE	SimCLR	73.27
L_q		73.35
MAE		73.36

Table 3. Test accuracy (%) for various methods on Clothing1M. We use results for the ImageNet initializer from their respective papers.

SimCLR initializer, they perform remarkably well.

We also observe that RN-50 pre-trained on the INet32 dataset also improves model robustness under label noise on the CIFAR datasets. Note that there is significant overlap between the classes of the INet32 dataset and the classes of the CIFAR datasets. On the CIFAR-100 dataset, the RN-50 pre-trained model (on INet32) significantly improves test accuracy to 81.51% from fine-tuning compared to $\sim 75\%$ with a random or the SimCLR initializer. However, the SimCLR initializer does not require such a considerable knowledge transfer from another larger dataset. Moreover, we observe that the drop in performance (w.r.t. the accuracy from the clean training samples) is significantly lower for the SimCLR initializer compared to the pre-trained ImageNet initializer. The pre-trained INet32 initializer seems to help more in the case of the asymmetric noise case; class overlap and label corruptions to a similar class might be a reason behind the improvement. In contrast, in the Clothing1M dataset, only two of the classes (out of 14) are present in the ImageNet dataset. Consequently, the CCE loss with the SimCLR initializer improves the test performance by 6% compared to the ImageNet pre-trained RN-50 initializer.

3. Conclusion

In this paper, we have shown that many of the supervised robust methods do not learn anything useful under high label noise rates. However, they perform significantly better with the SimCLR initializer on image datasets and can even outperform previous state-of-the-art methods for learning under label noise. Even the typical method, i.e., training a deep neural network-based classifier under the categorical cross-entropy loss, can outperform previous state-of-the-art methods under some noise conditions. These observations suggest that lack of good visual representations is a possible reason that many supervised robust methods perform poorly on image classification tasks. We believe that our findings can serve as a new baseline for learning under label noise on image datasets. Moreover, we believe that decoupling the representation learning problem from learning under label noise would lead to new methods that can do well on either of these tasks with complementary strengths without the need for methods targeting both of these tasks together.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, pages 312–321. PMLR, 2019.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

[3] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

[4] Carla E Brodley, Mark A Friedl, et al. Identifying and eliminating mislabeled training instances. In *AAAI*, pages 799–805, 1996.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[7] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *ICCV*, pages 1431–1439, 2015.

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

[9] Zheltonozhskii Evgenii, Baskin Chaim, Mendelson Avi, M. Bronstein Alex, and Or Litany. Contrast to divide: self-supervised pre-training for learning with noisy labels. 2021. under review.

[10] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, pages 1568–1577, 2018.

[11] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, pages 1919–1925, 2017.

[12] Aritra Ghosh and Andrew Lan. Do we really need gold samples for sample weighting under label noise? In *WACV*, pages 3922–3931, January 2021.

[13] Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

[14] Aritra Ghosh, Naresh Manwani, and PS Sastry. On the robustness of decision tree learning under label noise. In *PAKDD*, pages 685–697. Springer, Cham, 2017.

[15] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

[16] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. In *NeurIPS*, pages 5836–5846, 2018.

[17] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8527–8537, 2018.

[18] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, pages 2712–2721. PMLR, 2019.

[19] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

[20] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2304–2313, 2018.

[21] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

[22] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *CVPR*, pages 5051–5059, 2019.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[24] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *arXiv preprint arXiv:2007.00151*, 2020.

[25] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *ICML*, 2020.

[26] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, 2020.

[27] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017.

[28] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *ICLR Workshops*, 2015.

[29] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, pages 4334–4343, 2018.

[30] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, pages 1919–1930, 2019.

[31] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pages 5552–5560, 2018.

[32] Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. In *NeurIPS*, pages 10–18, 2015.

[33] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, pages 839–847, 2017.

[34] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015.

[35] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: An information-theoretic noise-robust loss function. 32:6225–6236, 2019.

[36] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[37] Bodi Yuan, Jianyu Chen, Weidong Zhang, Hung-Shuo Tai, and Sara McMains. Iterative cross learning on noisy labels. In *WACV*, pages 757–765. IEEE, 2018.

[38] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[39] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pages 8778–8788, 2018.