EQUAL LOSS: A SIMPLE LOSS FUNCTION FOR NOISE ROBUST LEARNING

Lei Cui[⋆]*, Huan Peng*[‡]*, Yangguang Li*[†]*, Chuming Li*[†]*, Xingrun Xing*[∗]

[⋆] Tsinghua University
[†] Sensetime Group
[‡] Huazhong University of Science and Technology
[∗] Beihang University
cuil19@mails.tsinghua.edu.cn, huanpeng@hust.edu.cn,
{liyangguang, lichuming, xingxingrun}@sensetime.com

ABSTRACT

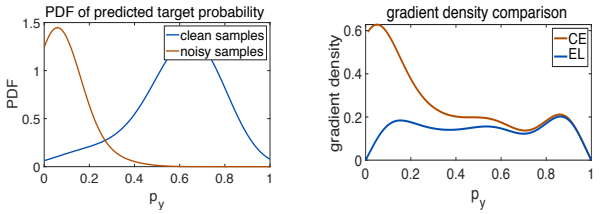
Training accurate deep neural networks in the presence of noisy labels is an important task. Though a number of approaches have been proposed for learning with noisy labels, many open issues remain. In this paper, we show that DNN learning with Cross Entropy is not robust to label noise and exhibits imbalance between the gradient of clean and noisy samples. We propose a new loss function, Equal Loss (EL), boosting DNN with a relaxed target probability and balanced gradient density. Both theoretical analysis and experiments on a range of benchmarks and real-world datasets show that EL outperforms state-of-the-art methods.

***Index Terms*—** neural networks, image classification, loss function, noisy learning, label correction

1. INTRODUCTION

Modern deep learning requires large-scale datasets with high quality annotations for proper training. However, the process of labeling large-scale datasets is error-prone due to the limitation on the human resources cost. As a result, most large datasets contain considerate numbers of noisy labels. Therefore, improving the performance of deep neural networks trained on noisy datasets has become a significant topic in deep learning.

Many works have studied the improvement of DNN learning with noisy labels. Label correction methods [1–5] predict the true labels of the raw noisy data and use them to retrain the target model. Another category of methods [6–12] are based on refining the training process. MentorNet [6, 7] uses clean data to learn a sample re-weighting scheme, which guides the training of the StudentNet. Decoupling [8] and Co-teaching [9] simultaneously train two networks and combine their predictions to guide the sampling weight or parameter update. D2L [10], iterative learning [11] and joint optimization [12] introduce complex intervention into the training strategy. Loss modification methods [13–20] propose loss functions more robust to noisy labels. Our method belongs to loss modification, which is complementary to the first two categories of methods.



(a) PDF at epoch 30 under CE (b) gradient density at epoch 30

Fig. 1. Visualization of target probability p_y and gradient of CIFAR-10 images with noise rate 40% at the 30-th training epoch of ResNet-18: (a) the probability distribution of p_y of clean samples and noisy samples, (b) gradient density of samples with different levels of p_y under CE and EL.

The existing loss functions do not delve into the dynamics of gradients of clean and noisy samples during training, which directly influences the performance of model. In this paper, we provide further insights into the dynamics of gradients. While Cross Entropy (CE) loss is the most common loss used in DNN training, we find that CE is not robust to noise because of the imbalance property of its gradient, which assigns much more total gradient for noisy data. To begin with, we define p_y as the predicted probability of a image belonging to its noisily labelled category. The correct estimation should make p_y small for the noisy sample, because the label for this sample is wrong. In contrast, the correct estimation should make p_y large for the clean sample, because the label for this sample is correct. As shown in Figure 1(a), at the middle of training, the probability density function (PDF) of clean and noisy samples' target probability p_y have two separate peaks far away from each other. The peak is around 0 for noisy samples but around 0.7 for clean samples. When the distribution of predicted probability and the corresponding magnitude of gradients are considered for the CE loss, the gradient density (defined as the PDF p_y for the sample multiplied by the

gradient norm for the sample) is used. As shown in Figure 1(b), the gradient density has high peak around zero in Figure 1(b), where most of the samples are noisy (as shown in Figure 1(a)). In this case, most of the large gradients are from noisy samples, which mislead the learning of DNN. As a result, when the training finishes, the network produces wrong estimation to predict a large value of p_y for many noisy samples.

The Equal Loss (EL) we propose, smooths the gradient density at different levels of target probability and treats samples with different target probability equally, thus improves the robustness of DNNs to noisy labels.

Our main contributions are summarized:

- We provide insights to the dynamics of gradients of DNNs under CE loss and find that the imbalance problem of CE is the key issue for learning with noisy labels.
- We propose Equal Loss (EL), which mitigates the gradient imbalance problem of CE. Both theoretical analysis and empirical understanding of EL are provided.
- We empirically demonstrate that EL is more robust to noisy labels than state-of-the-art approaches on both synthetic and real world noisy datasets.

2. EQUAL LOSS

In this section, as the previous work [15] has found that Cross-Entropy Loss is not sufficient for learning with noisy samples due to its gradient imbalance problem on noisy learning, so we propose Equal Loss (EL), a simple loss function that is robust to noisy labels. We also provide theoretical analysis about the formulation and gradient of EL.

2.1. Definition

Formulation As demonstrated above, CE lacks robustness to noisy labels because it does not balance the gradients between samples with different levels of p_y. Therefore, we consider a relax function t which maps p_y from a small value to a larger one $\tilde{p}_y = t(p_y)$, where $\tilde{p}_y \geq p_y$. Here t should meet three requirements: (1) $t(x) \geq x$, (2) $0 \leq t(x) \leq 1$ and (3) t is continuously differentiable and $t'(x) > 0$. We choose a quite simple family of t, which is a linear function:

$$t(x) = \frac{x+\epsilon}{1+\epsilon}. \tag{1}$$

It maps p_y from $[0, 1]$ to $[\frac{\epsilon}{1+\epsilon}, 1]$. By cascading t with CE, we propose Equal Loss:

$$l_{el}(f(x), y) = -(1+\epsilon)log(t(p_y)). \tag{2}$$

We add a multiplier $1+\epsilon$ to compensate the reduction of the gradients of all samples. As shown in Figure 2, the EL for p_y close to zero will be much smaller than CE, while the EL for p_y close to 1 is almost the same as CE. Note that although a linear function is applied to the predicted result, this function has non-linear effect on the gradient to the logits, with details given in Section 2.2. **Comparison of CE, EL, GCE, and MAE** We show an intuitive comparison among EL, CE and MAE in Figure 2. When $\epsilon = 0$, EL degenerates to CE. Further, we show that when $\epsilon = \infty$, EL will converge to MAE. The proof is as the following:

$$\begin{aligned}\lim_{\epsilon\to\infty} -(1+\epsilon)log(\frac{p_y+\epsilon}{1+\epsilon}) &= 1 - p_y \\ &= \frac{1}{2}\sum_{k=1}^{K}\|q_k - p_k\|_1 \\ &= \frac{1}{2}l_{mae}.\end{aligned} \tag{3}$$

2.2. Theoretical Analysis

Deriving the Gradient Consider the case of a single true label, the gradient of the sample-wise EL with respect to the logits z_j can be derived as:

$$\frac{\partial l_{el}}{\partial z_j} = \begin{cases} -(1+\epsilon)\frac{p_y}{p_y+\epsilon}(1-p_j), & \text{if } j = y, \\ (1+\epsilon)\frac{p_y}{p_y+\epsilon}p_j, & \text{if } j \neq y. \end{cases} \tag{4}$$

Figure 3(a) shows the gradient of CE, EL, and MAE as the functions of p_y.

Comparison of CE and EL for Noisy Samples Combining Eq. 4 with gradient of sample-wise CE loss, we have

$$\frac{\partial l_{el}}{\partial z_j} = (1+\epsilon)\frac{p_y}{p_y+\epsilon}\frac{\partial l_{ce}}{\partial z_j}, \tag{5}$$

and

$$\|\frac{\partial l_{el}}{\partial \boldsymbol{z}}\|_1 = (1+\epsilon)\frac{p_y}{p_y+\epsilon}\|\frac{\partial l_{ce}}{\partial \boldsymbol{z}}\|_1. \tag{6}$$

Eq. (6) shows that the ratio between the L_1 norms of gradients for EL l_{el} and and CE l_{ce} is $(1+\epsilon)\frac{p_y}{p_y+\epsilon}$, which increases from 0 to 1 as p_y increases from 0 to 1, as shown in Figure 3(b). Thus, at the early stage of training, when noisy samples have lower values of p_y and CE loss assigns large gradient for these samples, while EL uses the weight $\frac{p_y}{p_y+\epsilon}$ to reduce the difference between the gradient of samples with different levels of p_y. Therefore, EL makes use of the p_y predicted from the model to distinguish noisy and correct labels. So, our loss form is much more flexible in balancing between the gradients at different p_y values.

Robustness analysis Using the modeling of noisy samples in [14], we will show that EL achieves better solution than CE. By default, we assume all samples x are in a category k. We denote the noisy label of x as $\hat{y}$, in contrast to its true label $y = k$. Given any classifier f and loss function l_{el}, we define the risk (the expectation of loss) under label noise rate $\boldsymbol{\eta}$ as $R^{\boldsymbol{\eta}}_{el}(f) = E_{x,\hat{y}}(l_{el})$ and the risk under clean labels as $R_{el}(f) = E_{x,y}(l_{el})$. Here $\boldsymbol{\eta}_i$ means the probability of samples in k being labeled with i. Specially, $\boldsymbol{\eta}_k$ is the probability of being correctly labeled. Similarly, we define the risk $R^{\boldsymbol{\eta}}_{ce}(f) = E_{x,\hat{y}}(l_{ce})$ and $R_{ce}(f) = E_{x,y}(l_{ce})$ for CE.

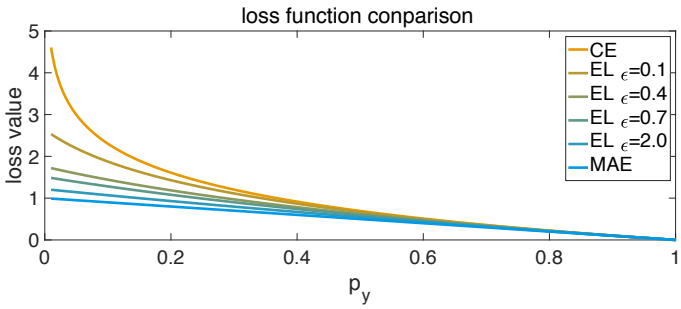


Fig. 2. Comparison of EL and CE, MAE. When $\epsilon = 0$, EL=CE, and when $\epsilon = \infty$, EL=MAE

Let $\widetilde{f}_{el}^{\eta}$ and $\widetilde{f}_{ce}^{\eta}$ be the minimizers of $R_{el}^{\boldsymbol{\eta}}(f)$ and $R_{ce}^{\boldsymbol{\eta}}(f)$ respectively. We prove that $\widetilde{f}_{el}^{\eta}$ has a lower loss under CE on clean dataset than $\widetilde{f}_{ce}^{\eta}$, which means $R_{ce}(\widetilde{f}_{el}^{\eta}) < R_{ce}(\widetilde{f}_{ce}^{\eta})$, under an acceptable noise rate. *Proof.* For noise rate $\boldsymbol{\eta}$:

$$R_{el}^{\boldsymbol{\eta}(f)} = \sum_{x \in \mathrm{X}_k} \sum_{i=1}^{K} -\boldsymbol{\eta}_i log(t(p(\hat{y} = i|x))) \tag{7}$$

Next we use p_i to replace $p(\hat{y} = i|x)$ for simplicity. To analyze the minimizer of $\sum_{i=1}^{K} -\boldsymbol{\eta}_i log(t(p_i))$, we details the Karush-Kuhn-Tucker (KKT) condition of the following problem:

$$\begin{aligned} &\min \sum_{i=1}^{K} -\boldsymbol{\eta}_i log(t(p_i)) \\ &\text{s.t. } p_i \geq 0, i = 1, ..., K, \quad \sum_{i=1}^{K} p_i = 1. \end{aligned} \tag{8}$$

We define λ and u_i as the Lagrange multiplier respectively for constraint $\sum_{i=1}^{K} p_i = 1$ and $p_i \geq 0$ and achieve the following KKT conditions for the minimizer of $R_{el}^{\boldsymbol{\eta}(f)}$:

$$\begin{aligned} L_{el} &= \sum_{i=1}^{K} -\boldsymbol{\eta}_i log(t(p_i)) + \lambda(\sum_{i=1}^{K} p_i - 1) - \sum_{i=1}^{K} u_i p_i \\ &\nabla L_{el} = 0, \\ &\sum_{i=1}^{K} p_i = 1, \\ &u_i p_i = 0, i = 1, ..., K. \end{aligned} \tag{9}$$

By expanding the condition $\nabla L_{el} = 0$ and knowing $u_i p_i = 0$, we have:

$$\boldsymbol{\eta}_i = \lambda(p_i + \epsilon) - u_i \epsilon. \tag{10}$$

Eq. (10) provides a clue to get the value of p_i, which will consequently reach the value of $R_{ce}(\widetilde{f}_{el}^{\eta})$. Thus, we start by seeking for the value of u_i in Eq. (10). To begin with, we repermutate the index of $\boldsymbol{\eta}_i$ and p_i via descending orders o and $\hat{o}$, which satisfies $\boldsymbol{\eta}_{o_1} \geq \boldsymbol{\eta}_{o_2} \geq ... \geq \boldsymbol{\eta}_{o_K}$ and $p_{\hat{o}_1} \geq p_{\hat{o}_2} \geq ... \geq p_{\hat{o}_K}$. Further, we prove that $o = \hat{o}$ when p_i is the minimizer.

Proof. If p_i is the minimizer and there exist i and j, s.t. $\boldsymbol{\eta}_i > \boldsymbol{\eta}_j$ and $p_i < p_j$, we can exchange the value of p_i and p_j to get a lower weighted sum. With the proof above, we use o as the common order of $\boldsymbol{\eta}$ and p. The condition $u_i p_i = 0$ implies that $u_i = 0$ for $p_i > 0$. We assume $p_{o_i} > 0$ for $i \leq Z$ and $p_{o_i} = 0$ for $i > Z$. By summarizing Eq. (10) from 1 to Z, we have:

$$\lambda = \frac{\sum_{i=1}^{Z} \boldsymbol{\eta}_{o_i}}{1 + Z\epsilon} \tag{11}$$

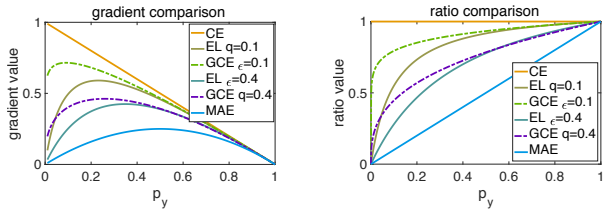
By introducing the expression of λ into Eq. (10), we obtain p_k:

$$p_k^{el} = \frac{(1 + Z\epsilon)\boldsymbol{\eta}_k}{\sum_{i=1}^{Z} \boldsymbol{\eta}_{o_i}} - \epsilon \tag{12}$$

Finally, we prove that $p_k^{el} > p_k^{ce}$, which is equivalent with $R_{ce}(\widetilde{f}_{el}^{\eta}) < R_{ce}(\widetilde{f}_{ce}^{\eta})$:

$$\begin{aligned} p_k^{el} - p_k^{ce} = \boldsymbol{\eta}_k &(\frac{1}{\sum_{i=1}^{Z} \boldsymbol{\eta}_{o_i}} - 1) \\ &+ \epsilon \frac{Z}{\sum_{i=1}^{Z} \boldsymbol{\eta}_{o_i}} (\boldsymbol{\eta}_k - \frac{\sum_{i=1}^{Z} \boldsymbol{\eta}_{o_i}}{Z}) \end{aligned} \tag{13}$$

The first term $\boldsymbol{\eta}_k(\frac{1}{\sum_{i=1}^{Z} \boldsymbol{\eta}_{o_i}} - 1) > 0$, $\boldsymbol{\eta}_k - \frac{\sum_{i=1}^{Z} \boldsymbol{\eta}_{o_i}}{Z} \geq 0$. Thus, we have $p_k^{el} - p_k^{ce} > 0$, which completes the proof.



(a) (b)

Fig. 3. Visualization of (a) gradient L1 norm of different loss functions and (b) gradient norm ratio between different loss functions and CE.

3. EXPERIMENT

We first evaluate EL's robustness against noisy labels on CIFAR-10 and CIFAR-100 [22], and a large-scale real-world dataset ImageNet[23], then provide some empirical understanding of our EL.

Noise setting We test two types of label noise: symmetric (uniform) noise and asymmetric (class-dependent) noise. Symmetric noisy labels are generated by flipping the labels of a given proportion of training samples to one of the other class labels uniformly. While for asymmetric noisy labels, flipping labels only occurs within a specific set of classes [15, 24]. For CIFAR-10, we flip TRUCK $\leftrightarrow$ AUTOMOBILE, BIRD $\leftrightarrow$ AIRPLANE, DEER $\leftrightarrow$ HORSE, CAT $\leftrightarrow$ DOG; for CIFAR-100, the 100 classes are grouped into 20 super-classes with each has 5 sub-classes, then we randomly select two sub-classes within each super-class and flip between them.

Table 1. Test accuracy (%) of different models on benchmark datasets with various rates of noisy labels. The average accuracy and standard deviation of 4 random runs are reported and the best results are in bold.

Datasets	Methods	Symmetric Noise Rate				Asymmetric Noise Rate			Mean
		0.2	0.4	0.6	0.8	0.2	0.3	0.4	
CIFAR-10	CE	82.96 ± 0.05	78.70 ± 0.07	66.62 ± 0.15	34.80 ± 0.25	85.98 ± 0.03	83.53 ± 0.08	78.51 ± 0.05	73.01
	LSR[16]	83.49 ± 0.05	78.41 ± 0.03	67.38 ± 0.15	36.30 ± 0.16	85.38 ± 0.05	82.89 ± 0.12	77.88 ± 0.20	73.10
	Bootstrap[13]	83.95 ± 0.10	79.97 ± 0.07	71.65 ± 0.05	41.44 ± 0.49	86.57 ± 0.08	84.86 ± 0.05	79.76 ± 0.07	75.46
	Forward[21]	85.83 ± 0.05	81.37 ± 0.03	73.59 ± 0.08	47.10 ± 0.14	87.68 ± 0.01	86.86 ± 0.06	$\mathbf{85.73 \pm 0.04}$	78.31
	SL [17]	87.63 ± 0.06	$\mathbf{85.34 \pm 0.07}$	80.07 ± 0.02	53.81 ± 0.27	88.24 ± 0.05	85.36 ± 0.14	80.64 ± 0.10	80.16
	D2L[10]	81.13 ± 0.06	76.80 ± 0.12	60.67 ± 0.12	19.83 ± 0.05	82.72 ± 0.06	80.41 ± 0.05	73.33 ± 0.12	67.84
	GCE[15]	84.86 ± 0.06	81.35 ± 0.10	75.20 ± 0.09	40.81 ± 0.24	84.61 ± 0.09	82.11 ± 0.13	75.32 ± 0.10	74.89
	EL	$\mathbf{88.59 \pm 0.05}$	82.42 ± 0.05	$\mathbf{81.50 \pm 0.09}$	$\mathbf{54.89 \pm 0.20}$	$\mathbf{89.37 \pm 0.06}$	$\mathbf{86.89 \pm 0.05}$	82.36 ± 0.11	**80.86**
CIFAR-100	CE	59.26 ± 0.39	50.82 ± 0.19	25.39 ± 0.09	5.27 ± 0.06	62.97 ± 0.19	63.12 ± 0.16	61.85 ± 0.35	46.95
	LSR[16]	58.83 ± 0.40	50.05 ± 0.31	24.68 ± 0.43	5.22 ± 0.07	63.03 ± 0.48	62.32 ± 0.48	61.59 ± 0.41	46.53
	Bootstrap[13]	57.91 ± 0.42	48.17 ± 0.18	12.27 ± 0.11	1.00 ± 0.01	63.44 ± 0.35	63.18 ± 0.35	62.08 ± 0.22	44.01
	Forward[21]	59.75 ± 0.34	53.13 ± 0.28	24.70 ± 0.26	2.65 ± 0.03	64.09 ± 0.61	64.00 ± 0.32	60.91 ± 0.36	47.03
	SL [17]	60.01 ± 0.19	53.69 ± 0.07	41.47 ± 0.04	15.00 ± 0.04	65.58 ± 0.06	65.14 ± 0.05	63.10 ± 0.13	52.00
	D2L[10]	59.20 ± 0.43	52.01 ± 0.37	35.27 ± 0.28	5.33 ± 0.54	62.43 ± 0.28	63.20 ± 0.27	61.35 ± 0.66	48.40
	GCE[15]	59.06 ± 0.27	53.25 ± 0.65	36.16 ± 0.74	8.43 ± 0.80	63.03 ± 0.22	63.17 ± 0.26	61.69 ± 1.15	49.26
	EL	$\mathbf{61.56 \pm 0.12}$	$\mathbf{55.33 \pm 0.07}$	$\mathbf{44.65 \pm 0.10}$	$\mathbf{19.06 \pm 0.44}$	$\mathbf{66.69 \pm 0.13}$	$\mathbf{66.23 \pm 0.18}$	$\mathbf{64.29 \pm 0.27}$	**53.97**

3.1. Robustness to Noisy Labels

Experimental details: We use an 8-layer CNN with 6 convolutional layers for CIFAR-10 and a ResNet-44 [25] for CIFAR-100. For EL, we set ϵ as the value of noise rate. We test varying noise rates 20%, 40%, 60%, 80% for symmetric noise, and 20%, 30%, 40% for asymmetric noise. All networks are trained using nesterov SGD with momentum 0.9, weight decay $5e-3$ and an initial learning rate of 0.1. The learning rate decays via a cosine curve for CIFAR-10 and CIFAR-100. We use two simple data augmentation techniques, random crop and horizontal flip.

Robustness performance: The classification accuracies are reported in Table 1. As can be seen, EL improves the baselines via a large margin for almost all noise rates and all datasets. We also find that EL can be more effective than GCE and SL, particularly for high noise rates. The clear advantage of EL is the balanced learning between samples with different levels of target probability.

3.2. Experiments on Real-world Dataset

In the below experiments, we assess its applicability for a real-world large-scale dataset: ImageNet [23]. We use ResNet-50 and ResNet-101 [25] on ImageNet. For preprocessing, images are resized to 256×256, with mean value subtracted and cropped at the center of 224×224. We train the models with batch size 2048 and initial learning rate 0.8, which is decayed via a cosine surve. SGD with a momentum 0.9 and weight decay $1e-4$ are adopted as the optimizer. We set ϵ as 0.1. As shown in Table 2, EL obtains higher top-1 validation accuracy compared with the baseline CE. Additionally, EL can be incorporated with LSR and enhance the performance further.

Table 2. Top-1 test accuracy (%) of different models on ImageNet. The average accuracy and standard deviation of 4 random runs are reported and the best results are in bold.

Methods	CE	EL	LSR	LSR+EL
ResNet-50	76.80 ± 0.09	$\mathbf{77.65 \pm 0.09}$	77.06 ± 0.11	$\mathbf{77.74 \pm 0.15}$
ResNet-101	77.98 ± 0.09	$\mathbf{78.72 \pm 0.08}$	78.15 ± 0.10	$\mathbf{78.82 \pm 0.11}$

3.3. Empirical Understanding of EL

We conduct experiments on CIFAR-10 dataset towards a deeper understanding of EL.

Gradient smooth Under CE loss, the density of gradients has a peak around $p_y = 0$. However, this area has a high portion of noisy labels. When EL is applied, the difference of gradient density over different levels of p_y will be largely smoothed. The smoothness facilitates DNN to fairly treats different samples and map each image feature to the most likely category. Therefore, when the training finishes, The p_y distribution of noisy and clean samples under EL has two separate peaks locating at 0 and 1, while for CE, the two peaks are all near 1, which implies a severe overfitting to noisy labels.

4. CONCLUSIONS

In this paper, We propose Equal Loss (EL), boosting CE with a relaxed target probability to alleviate its imbalance problem. We provide both theoretical and empirical understanding on EL, and demonstrate its effectiveness against various types and rates of label noise on both benchmark and real-world datasets. Overall, due to its simplicity and ease of implementation, EL is a promising loss function for noise robust learning, and an attractive plug-in used with other denoising techniques.

5. REFERENCES

[1] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang, "Cleannet: Transfer learning for scalable image classifier training with label noise," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5447–5456.

[2] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1910–1918.

[3] Arash Vahdat, "Toward robustness against label noise in training deep discriminative neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 5596–5605.

[4] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 839–847.

[5] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2691–2699.

[6] 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," *arXiv preprint arXiv:1712.05055*, 2017.

[7] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama, "How does disagreement help generalization against label corruption?," *arXiv preprint arXiv:1901.04215*, 2019.

[8] Eran Malach and Shai Shalev-Shwartz, "Decoupling "when to update" from" how to update"," in *Advances in Neural Information Processing Systems*, 2017, pp. 960–970.

[9] 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 *Advances in neural information processing systems*, 2018, pp. 8527–8537.

[10] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey, "Dimensionality-driven learning with noisy labels," *arXiv preprint arXiv:1806.02612*, 2018.

[11] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia, "Iterative learning with open-set noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8688–8696.

[12] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Joint optimization framework for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5552–5560.

[13] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596*, 2014.

[14] Aritra Ghosh, Himanshu Kumar, and PS Sastry, "Robust loss functions under label noise for deep neural networks," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[15] Zhilu Zhang and Mert Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[17] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 322–330.

[18] Ehsan Amid, Manfred KK Warmuth, Rohan Anil, and Tomer Koren, "Robust bi-tempered logistic loss based on bregman divergences," in *Advances in Neural Information Processing Systems*, 2019, pp. 15013–15022.

[19] 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.

[20] Yueming Lyu and Ivor W Tsang, "Curriculum loss: Robust learning and generalization against label corruption," *arXiv preprint arXiv:1905.10045*, 2019.

[21] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1944–1952.

[22] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[24] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu, "Making neural networks robust to label noise: a loss correction approach," *stat*, vol. 1050, pp. 13, 2016.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.