Which Strategies Matter for Label Noise? Insight into Loss and Uncertainty

Wonyoung Shin[1], **Jung-Woo Ha**[1], **Shengzhe Li**[1], **Yongwoo Cho**[1], **Hoyean Song**[2], **Sunyoung Kwon**[3*]
[1]NAVER Corp. [2]Riiid! [3]Pusan National University
{wonyoung.shin,jungwoo.ha,s.li,yongwoo.cho}@navercorp.com,
sjhshy@gmail.com, skwon@pusan.ac.kr

Abstract

Label noise is a critical factor that degrades the performance of deep neural networks, thus leading to severe issues in real-world problems. Existing studies have employed strategies based on either loss or uncertainty to address noisy labels, and ironically some strategies contradict each other; emphasizing or discarding uncertain samples, or concentrating on high or low loss samples. To elucidate how opposing strategies can enhance performance and offer insights into training with noisy labels, we present analytical results on how loss and uncertainty of samples change throughout training. From the in-depth analysis, we design a new training method that emphasizes clean and informative samples, while minimizing the influence of noise using both loss and uncertainty. We demonstrate the effectiveness of our method with extensive experiments on synthetic and real-world datasets.

1 Introduction

Recent advances in deep learning have significantly improved performance in numerous tasks due to large quantities of human-annotated data. While large-scale benchmark datasets used for research are generally clean and error-free, most real-world datasets contain noisy labels. The ubiquity of noise is a critical issue because learning with noisy labels severely degrades model performance [1].

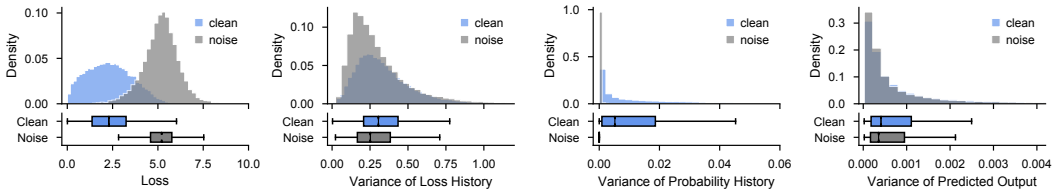
One way of addressing label noise is to focus on the uncertainty of samples during the training phase. Some methods emphasize uncertain samples, the predictions of which are inconsistent during training [2], while some methods reduce the importance or exclude uncertain samples so that only highly certain samples remain in the training data [3].

Another approach deals with noisy labels by managing samples based on their loss. Loss can signify the difficulty and the confidence of predictions, so giving precedence to samples with low loss or samples with high loss can work well depending on the amount of noise in the data or the complexity of the problem [4]. There have been studies that increase the weights of high loss samples [5] or take the opposite approach by emphasizing easy samples [6, 7].

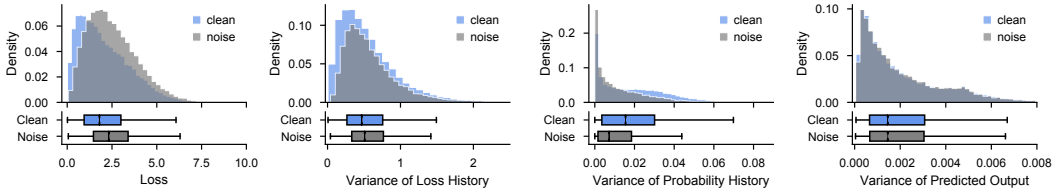
Motivated to understand how all approaches can enhance accuracy, we analyze the changing loss and uncertainty of samples in the course of training with different types of noise. Data show that symmetric noise is easy to identify using either loss or uncertainty, whereas asymmetric noise is challenging to distinguish from clean samples. Inspired by the finding that only a minority of samples with low loss and high uncertainty have noisy labels, we design FOCI (*Focus On Clean and Informative samples*), a novel training method that emphasize samples likely to be clean and informative. To validate our method, we conducted extensive experiments on CIFAR-10, CIFAR-100, and Tiny ImageNet with diverse noise types from 40% to 70% of noise levels, as well as on a real-world dataset Clothing1M.

*Correspondence should be addressed to Sunyoung Kwon.

NeurIPS 2020 Workshop on Dataset Curation and Security

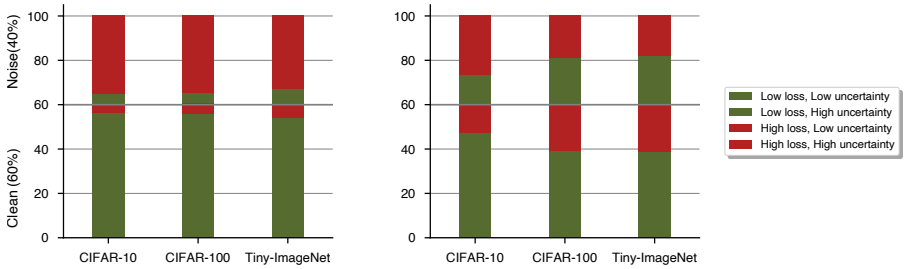


(a) Symmetric noise.



(b) Asymmetric noise.

Figure 1: Normalized distribution of loss and uncertainty on CIFAR-100 with 40% noise at epoch 50.



(a) Symmetric noise (b) Asymmetric noise

Figure 2: Combinations of loss and uncertainty with 40% noise rate at epoch 50. Clean (60%) and noisy (40%) samples are divided vertically. The green and red colors represent low and high loss samples, respectively. The solid and stripe patterns represent low and high uncertainty, respectively.

2 Loss and Uncertainty in Noisy Datasets

We explore how loss and uncertainty differ for various label noise by training DenseNet (L=25, k=12) on CIFAR-10, CIFAR-100, and Tiny ImageNet. We corrupted data by following typical protocols [6, 8]. For k classes, noise is given by swapping labels with a constant probability, namely, noise rate τ. While labels are swapped between two classes for asymmetric noise, labels are swapped to classes other than the true class label with probability of $\frac{\tau}{k-1}$ for symmetric noise [9, 10].

As shown in Figure 1, we explore various representations of uncertainty. Uncertainty can be quantified by the variance of loss or predicted probabilities for a given class in a q-sized history ($q = 15$) [11]. Another definition is by the variance of the predicted probabilities over all the classes at one step [12]. The difference of distributions between noisy and clean samples is more pronounced for variance based on the history of predicted probabilities, so we use this definition for uncertainty. We check samples based on their loss and uncertainty (prediction variance) by dividing them into four groups (see Figure 2). Samples are split into low and high loss with a ratio of $1 - \tau : \tau$ as suggested by [6]. The same applies for uncertainty in which the ratio of low and high uncertainty samples is $1 - \tau : \tau$.

As shown in Figures 1a and 2a, clean and noisy samples have distinct characteristics for symmetric noise, seemingly due to the easiness of symmetric noise and deep networks' capability of generalizing on data with symmetric noise [13]. The majority of noisy samples have higher loss and lower uncertainty. The loss of noisy samples tends to be higher than those of clean samples because predictions are different from the given labels, and the uncertainty of noisy samples is close to zero because the predicted probabilities for the given noisy labels are minute. Approaches that emphasize easy (low loss) or uncertain samples can thus benefit from this fact. These findings not only support our idea of emphasizing low loss and high uncertainty samples, but also confirm that symmetric noise is an easy problem and less practical as stated in prior works [6, 14].

Table 1: Classification accuracy (%) on benchmark datasets with 40% noise. Asymmetric and symmetric noise are denoted by A and S respectively.

		Default	Active Bias	Coteaching	SELFIE	FOCI
Asymmetric noise	CIFAR-10	71.8±1.5	78.0±1.5	83.7±1.4	84.9±0.1	**86.2±0.4**
	CIFAR-100	45.3±1.4	50.3±0.6	47.3±1.4	52.8±0.5	**59.5±0.9**
	Tiny ImageNet	30.8±0.1	33.2±0.9	30.3±0.5	36.1±0.3	**37.6±0.9**
Mixed noise (A-30, S-10)	CIFAR-10	79.6±0.8	84.9±0.5	80.6±1.7	84.9±0.9	**85.7±0.4**
	CIFAR-100	49.9±0.7	56.2±0.5	50.9±0.8	58.5±0.3	**61.5±0.5**
	Tiny ImageNet	35.0±0.8	36.2±0.4	34.0±0.6	39.0±0.6	**39.7±0.7**
Mixed noise (A-20, S-20)	CIFAR-10	81.2±0.8	84.8±0.3	82.1±0.3	84.8±0.4	**86.1±0.9**
	CIFAR-100	53.0±0.4	58.2±1.1	54.2±1.0	59.1±0.5	**60.6±0.5**
	Tiny ImageNet	36.8±0.9	37.6±0.6	37.1±1.5	37.9±0.2	**37.9±0.2**
Nearest noise	CIFAR-100	45.8±0.8	54.8±0.7	55.9±0.8	57.8±0.3	**57.9±0.5**

Table 2: Classification accuracy (%) on CIFAR-100 with high-level noise. Asymmetric and symmetric noise are denoted by A and S respectively.

		Default	Active Bias	Coteaching	SELFIE	FOCI
Mixed noise	50% (A-40, S-10)	38.0±1.2	41.2±1.0	37.3±0.9	42.1±2.7	**48.8±2.1**
Mixed noise	60% (A-30, S-30)	35.9±0.4	40.8±1.5	37.0±0.7	43.8±0.4	**48.5±1.4**
Mixed noise	70% (A-20, S-50)	32.9±0.7	35.8±0.5	32.2±0.2	41.5±0.9	**42.0±2.2**
Nearest noise	60%	36.3±0.9	43.2±0.1	44.0±1.6	46.6±1.1	**47.1±0.9**

Inspection of Figure 2b indicates that noisy samples can have high or low loss and uncertainty, providing evidence for enhanced accuracy of strategies that contradicted each other. However, according to Figure 1b, it is not effective to distinguish clean and noisy samples solely based on loss or uncertainty; the loss of clean and noisy samples are alike, and the difference between the uncertainty of clean and noisy samples is subtle. Considering both loss and uncertainty seems more effective and plausible when training with asymmetric noise, which is problematic and similar to real-world noise [14, 15].

3 Method

Let the training set be $\mathcal{D} = \{(x, y)\}$ of size N, and the dataset for a mini-batch be $\mathcal{D}_b$ of size N_b. The algorithm starts by updating the network in the conventional way, because deep networks can learn simple and common patterns, even with the presence of noise during the early warm-up phase [16]. Training parameters θ in the warm-up phase with learning rate α can be formulated as:

$$\theta \leftarrow \theta - \alpha \nabla \left(\frac{1}{N_b} \sum_{x \in \mathcal{D}_b} \mathcal{L}(x, y; \theta)\right), \tag{1}$$

However, since real-world datasets are bound to have noise, our method pursues robust training after the warm-up phase by reweighting samples so clean and informative samples are emphasized and the impact of noisy samples are minimized:

$$\theta \leftarrow \theta - \alpha \nabla \left(\frac{1}{N_b} \sum_{x \in D_b} W(x, q) \mathcal{L}(x, y; \theta)\right), \tag{2}$$

where $W(x, q)$ is the reweighting function. To favor samples with low loss and high uncertainty, we compute $W(x, q)$ using $P_t(y|x)$, the predicted probability of the given label, and $\mathrm{Var}(P_{t-q+1:t}(y|x))$, the variance of predicted probabilities in the history queue for epochs from $t - q + 1$ to t.

$$W(x, q) = \mathrm{normalize} \sqrt{P_t(y|x) \cdot \mathrm{Var}(P_{t-q+1:t}(y|x))} \tag{3}$$

The weights are subsequently standardized, bounded with the sigmoid function to give a clipping effect, and further divided by a normalizing factor to have unit mean. We also reduce the impact of samples that are likely to be noisy using methods partially based on *SELFIE*. We screen samples with inconsistent predictions and high loss or samples with consistent predictions but the predicted label disagrees from the given label, and set their weights to zero.

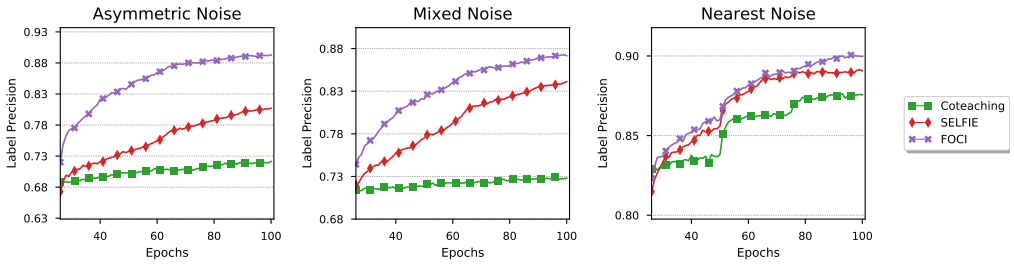


Figure 3: Label precision comparison on CIFAR-100 with 40% noise after the warmup phase (epoch 25). Mixed noise comprises 30% asymmetric noise and 10% symmetric noise. *Default* and *Active Bias* do not distinguish clean samples, so they are not included for comparison.

Table 3: Results on Clothing1M.

LCCN [17]	Joint Optim. [18]	DMI [19]	MLNT [20]	PENCIL [15]	**FOCI**
71.63	72.16	72.46	73.47	73.49	**73.78**

4 Experiments

Experimental settings. We perform image classification on CIFAR-10, CIFAR-100, Tiny ImageNet, and a real-world dataset Clothing1M [21]. In this work, we are concerned with scenarios of data with poor but realistic label quality. We thus use three label corruption schemes: asymmetric noise, mixed noise, and nearest label transfer. Labelers make mistakes within very few and similar classes [14, 15], so asymmetric noise is injected and symmetric noise is mixed with asymmetric noise. To simulate confusions between visually similar classes, we employ nearest label transfer following [22].

We trained DenseNet (L=25, k=12) for 100 epochs using SGD with a momentum of 0.9. The initial learning rate is 0.1 and is divided by 5 at epoch 50 and 75. The batch size is 128, ϵ=0.1, q=15, and γ=25. We measure performance by the mean of the last accuracies over three runs [2, 3] and label precision by the fraction of true clean samples among all the samples selected for training or samples that have non-zero weights. We compare FOCI with a baseline algorithm (denoted by *Default*) trained without any strategies, uncertainty-based *Active Bias*, loss-based *Coteaching*, and hybrid *SELFIE*.

Benchmark datasets. According to Table 1, our method achieves the best performance for all types of noise. The accuracy of our method for CIFAR-100 with asymmetric noise differs with the second-best algorithm by 7%. We can also observe from the mixed noise results that the difference between *Default* and other baselines reduces with more symmetric noise, indicating that symmetric noise does not require developed algorithms. Furthermore, FOCI can identify noise with high precision as indicated in Figure 3; The label precision of our method surpasses other methods and continues to increase as training proceeds, while other methods converge towards the end of training.

As shown in Table 2, we conducted experiments on CIFAR-100 with larger noise rates for mixed and nearest noise. The results show that FOCI outperforms other methods and confirm that our method downplays noisy samples and emphasizes clean and informative samples even for extreme cases.

Clothing1M. We test on Clothing1M [21], which consists of 1M images with real-world noisy labels and additional clean data. We retrain ResNet50 pretrained on ImageNet for 20 epochs using the 1M noisy dataset without any clean data. The initial learning rate is 0.002 and is decreased by 10 every 5 epochs. We use SGD with momentum of 0.9, $\epsilon = 0.1$, $q = 5$, $\gamma = 5$, and $\tau = 0.4$ Our method achieves 73.8% accuracy, which is higher than recent state-of-the-art methods (see Table 3). For fair comparison, we do not include methods using clean data or different backbone models.

5 Conclusion

In this paper, we have shown that considering both loss and uncertainty is necessary for asymmetric noise—a scenario that commonly occurs in real-world datasets. Inspired by the findings, we have designed a method that downplays noisy samples while emphasizing clean and informative samples. Through series of experiments, we have demonstrated our method's robustness towards label noise.

References

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

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

[3] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915, 2019.

[4] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. In *International Conference on Learning Representations*, 2016.

[5] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[6] 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*, pages 8527–8537, 2018.

[7] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, pages 1062–1070, 2019.

[8] Brendan van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pages 10–18, 2015.

[9] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor W Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. International Conference on Machine Learning, 2020.

[10] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? International Joint Conference on Artificial Intelligence, 2020.

[11] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, pages 1002–1012, 2017.

[12] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *Advances in Neural Information Processing Systems*, pages 4225–4235, 2017.

[13] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

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

[15] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7017–7025, 2019.

[16] 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 *International Conference on Machine Learning*, pages 2309–2318, 2018.

[17] Jiangchao Yao, Hao Wu, Ya Zhang, Ivor W Tsang, and Jun Sun. Safeguarded dynamic label regression for noisy supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9103–9110, 2019.

[18] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.

[19] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. l_{DMI}: A novel information-theoretic loss function for training deep nets robust to label noise. In *Advances in Neural Information Processing Systems*, pages 6222–6233, 2019.

[20] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019.

[21] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015.

[22] Paul Hongsuck Seo, Geeho Kim, and Bohyung Han. Combinatorial inference against label noise. In *Advances in Neural Information Processing Systems*, pages 1171–1181, 2019.