# Improving Label Noise Robustness with Data Augmentation and Semi-Supervised Learning (Student Abstract)

**Kento Nishi,**[1] **Yi Ding,**[2] **Alex Rich,**[2] **Tobias Höllerer**[2]
[1] Lynbrook High School, San Jose, CA 95129
[2] Four Eyes Lab, Department of Computer Science, University of California, Santa Barbara, CA 93106
kento24gs@outlook.com

## Abstract

Modern machine learning algorithms typically require large amounts of labeled training data to fit a reliable model. To minimize the cost of data collection, researchers often employ techniques such as crowdsourcing and web scraping. However, web data and human annotations are known to exhibit high margins of error, resulting in sizable amounts of incorrect labels. Poorly labeled training data can cause models to overfit to the noise distribution, crippling performance in real-world applications. In this work, we investigate the viability of using data augmentation in conjunction with semi-supervised learning to improve the label noise robustness of image classification models. We conduct several experiments using noisy variants of the CIFAR-10 image classification dataset to benchmark our method against existing algorithms. Experimental results show that our augmentative SSL approach improves upon the state-of-the-art.

## Introduction

Often times, large-scale datasets are collected using methods that sacrifice quality in pursuit of quantity – for instance, the CIFAR-10/100 datasets contain internet photos tagged by paid student volunteers at the University of Toronto (Krizhevsky and Hinton 2009). Although cost efficient, internet data and human labels are often inaccurate, inevitably resulting in many noisy labels. Noisy datasets can be problematic for machine learning models, as they can easily overfit to the noise distribution of the training set or fail to converge at all (Zhang et al. 2016).

To combat these issues, several methods have been proposed. The most common approach to learning with noisy labels (LNL) is loss correction (LC), where the loss function is modified to better handle noisy labels (Song et al. 2020). Another active field of research that shows potential for LNL is semi-supervised learning (SSL), where noisy samples are detected using LC and repurposed as unlabeled data for pseudo-labeling (Song et al. 2020).

In this work, we analyze the effect of data augmentation on noise tolerant models and propose an approach to take advantage of augmented samples for semi-supervised LNL. Benchmarks on noisy versions of the CIFAR-10 dataset demonstrate up to an 11.5% increase in absolute accuracy.



## Related Work

### Filtering Noisy Samples

When training models on noisy data, incorrectly labeled samples must first be filtered out. Previous works have shown that neural networks tend to learn correctly labeled samples more quickly than incorrect ones (Arpit et al. 2017) – using this empirical property, many methods treat samples with low loss as clean data. Among these methods, Co-teaching (Han et al. 2018) and Co-teaching+ (Yu et al. 2019) train two divergent networks, where each network curates samples with low loss for the other to train on. Training two models that supervise each other effectively prevents confirmation bias, where a model accumulates mistakes over time (Tarvainen and Valpola 2017).

### Semi-Supervised Learning with Label Noise

One major drawback of filtering noisy samples is that the amount of data available to the model is dramatically reduced. To combat this issue, more recent methods such as DivideMix (Li, Socher, and Hoi 2020) merge SSL with the two-model supervision architecture. Similar to Co-teaching and Co-teaching+, DivideMix trains two models that select clean data pools for each other. At each epoch, a Gaussian Mixture Model (GMM) is fit to classify between clean and noisy samples based on their loss. Then, the detected noisy samples are used to train the two networks with the state-of-the-art MixMatch SSL algorithm (Berthelot et al. 2019). Over time, most, if not all unlabeled samples are converted into accurately labeled data, increasing the dataset size and improving overall performance (Li, Socher, and Hoi 2020).

## Methods

MixMatch, the SSL algorithm used in DivideMix, employs a regularization technique called consistency regularization (CR). CR essentially encourages the model to output the same predictions for several augmented variations of a given sample. More formally, basic CR is implemented by adding the following term to the loss:

$$\|p_{model}(y \mid Aug(x); \theta) - p_{model}(y \mid Aug(x); \theta)\|_2^2 \quad (1)$$

where $Aug(x)$ randomly augments sample $x$ (Berthelot et al. 2019). Note that $Aug(x)$ is a stochastic transformation, resulting in unique values for $p_{model}(y \mid Aug(x); \theta)$.

| | 50% Noise | 90% Noise |
|---|---|---|
| **Momentum** | *0.9* | |
| **Weight Decay** | *0.0005* | |
| **Batch Size** | *64* | |
| **Epochs** | *250* | |
| **Learning Rate** | $\begin{cases} 0.02 & n < 150 \\ 0.002 & n \geq 150 \end{cases}$ | |
| **Augmentations** | $P_{horizontal\,flip} = 0.5$<br>$P_{random\,contrast} = 0.5$<br>$P_{shift\,scale\,rotate} = 0.8$<br>$P_{random\,brightness} = 0.5$<br>$P_{hue\,saturation\,value} = 0.9$ | |
| **Other** | $\alpha = 4, T = 0.5, M = 2, n_{warm\,up} = 10$ | |
| | $\lambda_u = 25, \tau = 0.5$ | $\lambda_u = 50, \tau = 0.5$ |

Table 1: Parameter values used for evaluating our method.

To adapt MixMatch for the two-model architecture, DivideMix replaces the $p_{model}(y \mid Aug(x); \theta)$ terms in (1) with predictions on unlabeled samples from each model:

$$\|p - p_{model}(x; \theta)\|_2^2 \tag{2}$$

where $p_{model}(x; \theta)$ is the mean output of one network given light augmentations of an unlabeled sample $x$ (flip/shift), and $p$ is the other network's mean sharpened output given the same augmentations of $x$ (Li, Socher, and Hoi 2020).

Although the CR term used by DivideMix shown in (2) effectively leverages mislabeled samples as unlabeled data, the two models are only encouraged to output similar predictions given an identical set of images, neglecting consistency across separate augmentations. To solve this problem, our proposed method generates two unique sets of augmented images given a common input image, each of which are separately fed to the two divergent networks. More concretely, we replace (2) with the following:

$$\|p_{model_1}(Aug(x); \theta_1) - p_{model_2}(Aug(x); \theta_2)\|_2^2 \tag{3}$$

With this change, (3) encourages the two models to output similar predictions given two separately augmented sets of $x$. In summary, our modified loss rewards consistency across augmentations of a given sample, improving generalization.

## Preliminary Evaluation and Results

To compare our method against existing techniques, we trained an 18-layer PreAct ResNet CNN (He et al. 2016) on the CIFAR-10 dataset (Krizhevsky and Hinton 2009) at 50% and 90% symmetric noise. We used nearly identical hyperparameters as DivideMix (Li, Socher, and Hoi 2020). Values are shown in Table 1.

Our preliminary experimental results are shown in Table 2. At both 50% and 90% noise, our training method improves the accuracy of the model; however, our method evidently has a more significant impact on accuracy at 90% noise. Our approach decreases the error rate by 9.3% and 47.9% at 50% and 90% noise, respectively. The improvement at 90% noise is substantial, outperforming the state-of-the-art by a margin of 11.5% in absolute accuracy.

| | 50% Noise | 90% Noise |
|---|---|---|
| Standard Cross Entropy | *79.4%* | *42.7%* |
| Co-teaching+ (Yu et al.) | *85.7%* | *47.9%* |
| DivideMix (Li, Socher, and Hoi) | *94.6%* | *76.0%* |
| **Our Method** | ***95.1%*** | ***87.5%*** |

Table 2: Results on CIFAR-10 with 50% and 90% noise.

## Conclusions and Future Work

In this paper, we introduced a novel semi-supervised training procedure that implements data augmentation to improve performance when learning with noisy labels. Results from benchmarks on the CIFAR-10 dataset show that our approach improves upon the state-of-the-art for LNL at varying levels of symmetric label noise.

For future work, we are interested in using additional datasets such as CIFAR-100 and Clothing-1M, experimenting with asymmetric noise, and exploring more advanced augmentation techniques including generative and learned augmentations. We hope that our contributions can pave the way for further improvements in label noise robustness.

## Acknowledgements

We would like to thank Aiwen Xu, Lina Kim, and the Research Mentorship Program at the University of California, Santa Barbara for their generous support.

## References

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. *arXiv e-prints* arXiv:1706.05394.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. *arXiv e-prints* arXiv:1905.02249.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. *arXiv e-prints* arXiv:1804.06872.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. *arXiv e-prints* arXiv:1603.05027.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto* .

Li, J.; Socher, R.; and Hoi, S. C. H. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. *arXiv e-prints* arXiv:2002.07394.

Song, H.; Kim, M.; Park, D.; and Lee, J.-G. 2020. Learning from Noisy Labels with Deep Neural Networks: A Survey. *arXiv e-prints* arXiv:2007.08199.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv e-prints* arXiv:1703.01780.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I. W.; and Sugiyama, M. 2019. How does Disagreement Help Generalization against Label Corruption? *arXiv e-prints* arXiv:1901.04215.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2016. Understanding deep learning requires rethinking generalization. *arXiv e-prints* arXiv:1611.03530.