# Ensemble Co-Teaching for Robust Learning of Deep Neural Networks under Label Noise

**Renato O. Miyaji**[1], **Pedro L. P. Corrêa**[1]

[1]Escola Politécnica – Universidade de São Paulo (USP)

{re.miyaji,pedro.correa}@usp.br

***Abstract.*** *Training Deep Neural Networks under Label Noise is challenging due to their memorization ability. To address this issue, various methods have been developed to facilitate robust learning under such conditions. Methods based on multiple networks, such as Stochastic Co-Teaching, have demonstrated superior performance in identifying correctly labeled instances compared to state-of-the-art approaches. In this paper we propose a new method, Ensemble Co-Teaching, which introduces the concept of ensemble learning into robust learning techniques by incorporating perturbations in the network weights. This ensures diversity between the two networks and enhances their ability to detect clean label samples. The proposed Ensemble Co-Teaching method achieved an accuracy improvement, with 91.0% compared to 88.9% from the Co-Teaching method.*

## 1. Introduction

Numerous real-world datasets are created through processes that often result in the production of unreliable labels. Furthermore, even domain experts may struggle with the complexities of accurate labeling, and labels can be intentionally manipulated through label-flipping. These unreliable labels, known as noisy labels, stem from deviations from the ground-truth labels [Song et al. 2022]. Training machine learning models for classification tasks in the presence of Label Noise can be particularly challenging, especially when using Deep Neural Networks.

In recent years, various methods based on robust architectures, robust regularization, robust loss design and sample selection have been developed to enable robust training of these models under such conditions [Song et al. 2022].

Multiple networks have been employed in various methods to tackle this issue. The underlying hypothesis of these techniques is that different networks can develop distinct learning capabilities to filter out different types of errors. However, in methods like Stochastic Co-Teaching, as training progresses and the number of epochs increases, the weights of the networks tend to converge, rendering this hypothesis invalid [Yu et al. 2019].

Ensemble methods have been successfully employed in the literature to enable robust learning of Deep Neural Networks, as demonstrated by Learning with Ensemble Consensus [Lee and Chung 2020]. To overcome the limitation of network convergence, we propose Ensemble Co-Teaching. This approach introduces ensemble concepts into Stochastic Co-Teaching [Vos et al. 2023].

This paper is structured into five main sections. Section 1 presents the motivation behind the proposed method. Section 2 introduces different methods that were used in

the literature to train Deep Neural Networks under Label Noise. Section 3 details the methodology, presenting Ensemble Co-Teaching, along with the case study design. Section 4 presents the results from the case study and offers a discussion of the findings. Finally, the Conclusion summarizes the key findings, addresses the study's limitations, and suggests directions for future research.

## 2. Related Works

In the literature, various methods have been developed to enable the robust training of Neural Networks under Label Noise. Some of them focus on modifying the architecture of the Network by adding a Noisy Adaptation Layer, others on improving regularization techniques or on using more robust loss functions. Techniques that focus on identifying correctly labeled instances to enhance the training process fall under the Sample Selection approach [Song et al. 2022].

To achieve this, some methods iteratively repeat the training rounds using a single network in a Multi-Round learning process. An example of this approach is the Iterative Trimmed Loss Minimization (ITLM) [Shen and Sanghavi 2019] method that repeats a cycle of selecting instances with correct labels and using them to retrain a Neural Network model. However, this approach carries risks related to accumulated errors and misclassified labels.

To mitigate these limitations, other methods employ multiple networks to identify instances with noisy labels. One of these Multi-Network methods is Decoupling [Malach and Shalev-Shwartz 2017], in which two Neural Networks are trained simultaneously using the same data. The instances where the predictions of the two Networks do not coincide are used to update their weights. However, there remains a challenge in addressing uncertain labels more explicitly.

Other method that also employs two Neural Networks is the MentorNet [Jiang et al. 2018]. Nevertheless, unlike Decoupling, one of the Networks is trained first and used for selecting instances with correct labels, referred to as the Mentor. Based on these selected instances, the second Network is trained. However, there are limitations and risks associated with sample-selection bias.

To mitigate the risks of the Decoupling and MentorNet, [Han et al. 2018] propose the Co-Teaching technique, which utilizes two neural networks trained simultaneously. During training on each mini-batch of the dataset, the networks communicate about the instances with the lowest loss. Based on this information, each network updates its weights. By exchanging information during training, Co-Teaching is able to mutually reduce the error rates in both networks [Han et al. 2018].

A limitation of the Co-Teaching is that the level of uncertainty in the data must be initially known. To overcome this, [Vos et al. 2023] propose Stochastic Co-Teaching, which employs stochasticity to select or reject training instances.

As the number of epochs in Co-Teaching increases, the two neural networks may converge to a consensus, potentially leading to sample-selection bias, as it happens in MentorNet. To mitigate this risk, [Yu et al. 2019] propose Co-Teaching+. Unlike the original method, Co-Teaching+ updates the network weights using instances where there is disagreement between the predictions of the two networks.

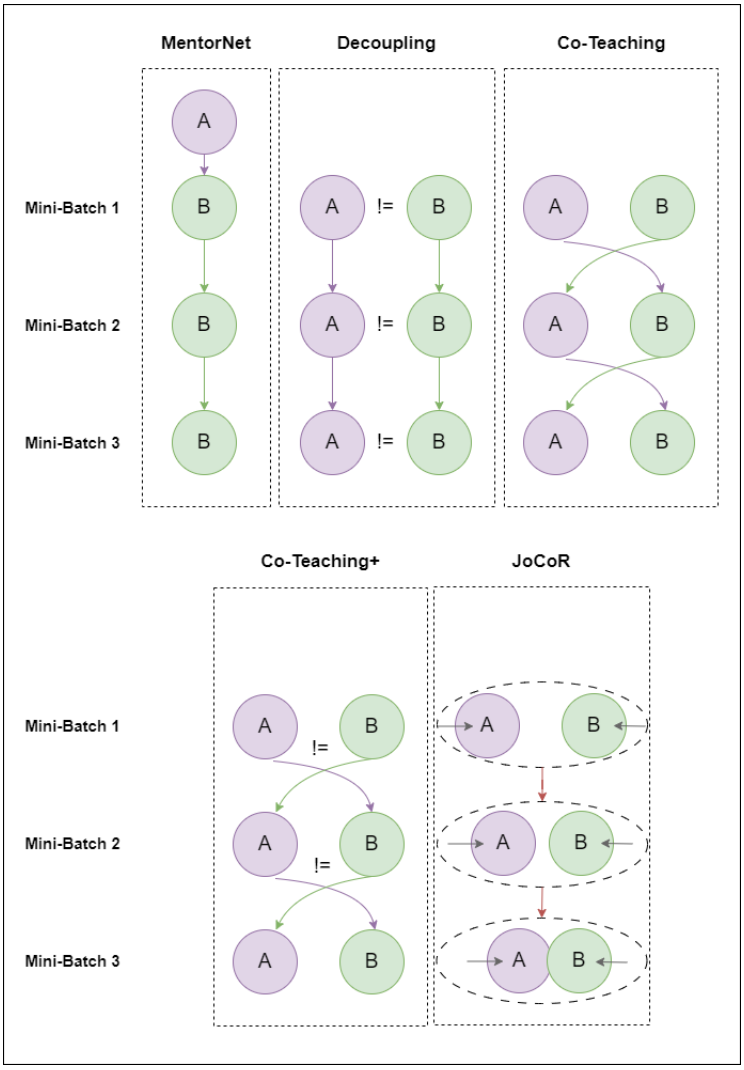


**Figure 1. Comparison between MentorNet, Decoupling, Co-Teaching, Co-Teaching+ and Joint Training Method with Co-Regularization (JoCoR) [Han et al. 2018] [Yu et al. 2019] [Wei et al. 2020]**

Co-Teaching based methods differ from other techniques, since the two Neural Networks communicate during their training process, as shown in Figure 1. In MentorNet, one Neural Network acts as a Mentor by initially selecting the instances with correct labels to guide the training of the second Neural Network [Jiang et al. 2018]. In Decoupling, both Neural Networks are trained simultaneously in mini-batches of the dataset, and there is communication to determine the instances where there is disagreement between the predictions. When this occurs, these instances are used to update the weights of the Networks [Malach and Shalev-Shwartz 2017].

Similar to Decoupling, in Co-Teaching, the two Neural Networks are also trained simultaneously. The main difference between the methods is that the communication between the Networks from each one to its peer for each mini-batch of the dataset. Moreover, the instances that will be used to update the weights of the Networks are selected based on their loss [Han et al. 2018].

In Co-Teaching+, the communication between the Networks happens in the same way as in Co-teaching, but a disagreement between the Networks predictions must occur to an instance to be selected. Then, they are ranked based on their loss [Yu et al. 2019].

Unlike Decoupling [Malach and Shalev-Shwartz 2017] and Co-Teaching+ [Yu et al. 2019], where the disagreement between the Networks are used to update their weights, in Joint Training Method with Co-Regularization (JoCoR) [Wei et al. 2020] a joint loss with Co-Regularization is calculated to each instance. Then, the ones with lowest losses are used to update the Network weights. Differently from the other methods that aim to generate two different Networks with distinct learning abilities, such as Decoupling and Co-Teaching+, in JoCoR the goal is to make two different classifiers converge.

Thus, considering the results of the literature review, it is evident that Ensemble techniques have not yet been employed with Co-Teaching methods to enable robust learning of Deep Neural Networks under Label Noise.

## 3. Method

**Problem Definition.** Consider a $K$-class classification task in a dataset $D$ with $X$ features and $Y = \{1, \ldots, K\}$ labels. The training dataset contains asymmetric label noise introduced through pair flipping, where similar classes are flipped at a specified noise rate $\epsilon$.

A Deep Neural Network with a Softmax output layer is used to determine a function that maps the feature space to the label space $h(X) = argmax_i f(X)_i$, where $f(X)$ represents the probability $p(K|X)$ that features $X$ belong to the class $K$.

The primary challenge that Ensemble Co-Teaching seeks to address is enabling robust learning of Deep Neural Networks under asymmetric Label Noise using a Sample Selection method combined with Ensemble concepts. Specifically, the method aims to determine the weights of two neural networks with similar architectures $W_1$ and $W_2$, such that they minimize the categorical cross-entropy loss $L(f(X), K)$, which is the loss of $f(X)$ with respect to label $K$. To prevent sample-selection bias, it ensures that $W_1 \neq W_2$ [Yu et al. 2019].

**Method.** In Ensemble Co-Teaching, the two Neural Networks are trained simultaneously. At each epoch $T_N$ to $T_{max}$ at iteration $N_{max}$, a mini-batch $\tilde{D}$ from the dataset $D$ is constructed. For each mini-batch, forward propagation is performed in both Networks, obtaining the instances with the lowest losses for each Network $\tilde{D}_1$ and $\tilde{D}_2$. The primary hypothesis for selecting samples with clean labels is the small-loss trick. According to this approach, samples with small loss values are considered to be correctly labeled, as clean label samples are often not explicitly known in real-world datasets [Song et al. 2022].

The number of instances selected with the lowest losses is defined by the Forget Rate $p$ that considers the level of uncertainties present in the mini-batch $\tilde{D}$. However, this parameter is not always known. Thus, the stochastic process proposed by [Vos et al. 2023] is used to determine $p$. The Forget Rate is defined randomly at each iteration from a Beta probability distribution given by Equation 1, in which $B$ is defined by Equation 2 [Vos et al. 2023].

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{1}$$

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt \tag{2}$$

The Beta distribution has values between 0 and 1, controlled by the parameters $\alpha$ and $\beta$. If the parameters are equal, the distribution is symmetric, but it can become bimodal, uniform, or exhibit positive or negative skewness as the parameters vary [Vos et al. 2023].

Differently from Co-Teaching [Han et al. 2018] and Co-Teaching+ [Yu et al. 2019], in Ensemble Co-Teaching the instances used to update the weights of the two Neural Networks $W_1$ and $W_2$ are those classified as low loss by at least one of the Networks. This approach leverages the different error patterns learned by each Network [Yu et al. 2019] to enhance the learning capabilities of both Networks. Then, the weights $W_1$ and $W_2$ of the Networks are updated using the Update Rule $U$ and the selected instances $\tilde{D}_{12} = \tilde{D}_1 + \tilde{D}_2$.

To ensure that the two Neural Networks do not converge ($W_1 = W_2$), in Ensemble Co-Teaching, at each epoch $T$ a small perturbation $\delta W$ is introduced in one of the Network's weights ($W_1 = W_1 + \delta W$). The perturbation $\delta W$ is sampled from a Normal distribution ($\mathcal{N}(0, \sigma^2)$) with a small standard deviation $\sigma$. The primary hypothesis behind introducing perturbations is that clean samples are learned through patterns, whereas noisy samples are learned through memorization. By introducing a small perturbation in $W_1$, predictions for samples learned via memorization will fluctuate, resulting in increased training losses [Lee and Chung 2020].

Figure 2 compares Ensemble Co-Teaching with Co-Teaching and Co-Teaching+. Algorithm 1 presents the proposed method.

**Experiments.** The Ensemble Co-Teaching method was compared to other state-of-the-art methods (Co-Teaching [Han et al. 2018], Co-Teaching+ [Yu et al. 2019] and Stochastic Co-Teaching [Vos et al. 2023]) on a benchmark dataset with an asymmetric Label Noise with $\epsilon = 45\%$.

The benchmark dataset used was MNIST (Modified National Institute of Standards and Technology database) that is composed of handwritten digits [LeCun et al. 1998]. The training set has 60,000 samples, while test set has 10,000 samples.

Since it is a clean dataset, a manual corruption was added so that Label Noise could be introduced [Patrini et al. 2017]. The noise transition matrix $Q$ was added, where $Q_{ij} = P(\tilde{y} = j|y = i)$ where the clean label $y$ is flipped to the noisy label $\tilde{y}$. The probability function that determines the noise transition matrix can be symmetric, in which each noisy label $\tilde{y}$ has the same probability, or asymmetric [Rooyen et al. 2015]. In asymmetric flipping, a pair flipping strategy is adopted, where there are mistakes only between similar classes (for instance, between classes 5 and 6 for the MNIST dataset) [Han et al. 2018].

An asymmetric Label Noise model with $\epsilon = 45\%$ was selected, as it represents a

**Algorithm 1:** Ensemble Co-Teaching

**Input:** Model Weights $W_1$ and $W_2$, Forget Rate $p$, Epoch $T_N$ to $T_{max}$, Iteration $N$, Perturbation $\delta W$ and Update Rule $U$

**Data :** Training Dataset $D$ with $X$ features and $Y = \{1, \ldots, K\}$ labels

**for** $T = 1, 2$ *until* $T_{max}$ **do**
  Randomize Training Dataset $D$;
  **for** $N = 1$ *until* $N_{max}$ **do**
    Obtain Mini-Batch $\tilde{D}$ from $D$;
    Sample Forget Rate $p \in Beta(\alpha, \beta)$;
    Obtain $\tilde{D}_1$;
    Obtain $\tilde{D}_2$;
    Generate $\tilde{D}_{12} = \tilde{D}_1 + \tilde{D}_2$;
    Update $W_1 \leftarrow U(W_1, \tilde{D}_{12})$;
    Update $W_2 \leftarrow U(W_2, \tilde{D}_{12})$;
    Sample Perturbation $\delta W \in \mathcal{N}(0, \sigma^2)$;
    Add Perturbation $W_1 = W_1 + \delta W$;
  **end**
**end**

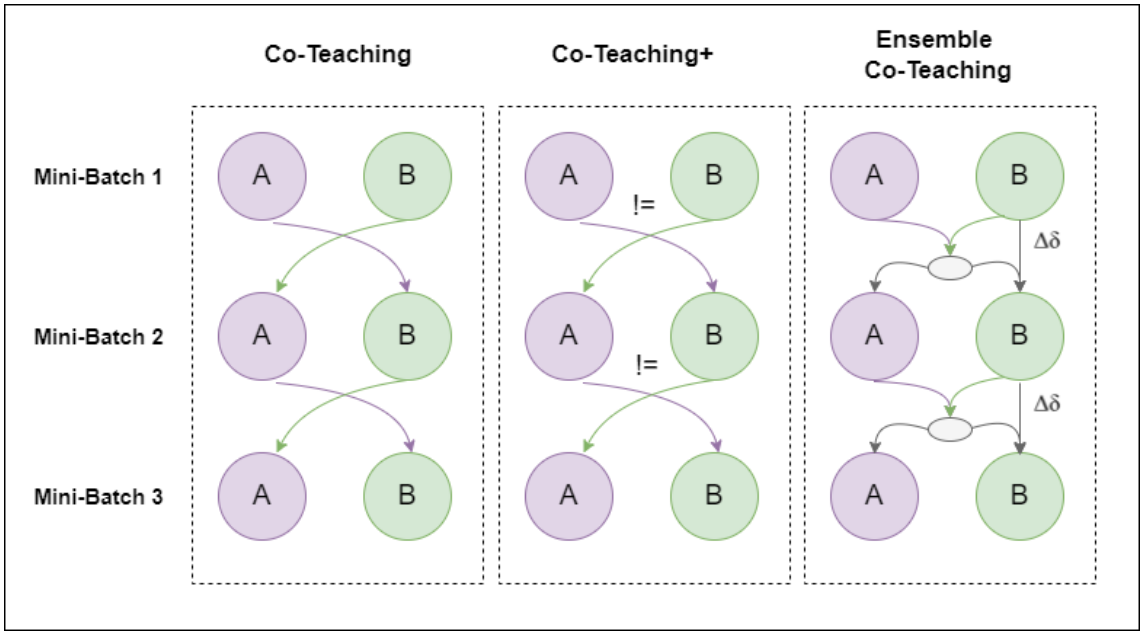


**Figure 2. Comparison between Co-Teaching, Co-Teaching+ and Ensemble Co-Teaching [Han et al. 2018] [Yu et al. 2019]**

more realistic scenario found in most real-world datasets and a case with extremely noisy labels [Han et al. 2018]. The Noise Transition Matrix is presented in Figure 3.

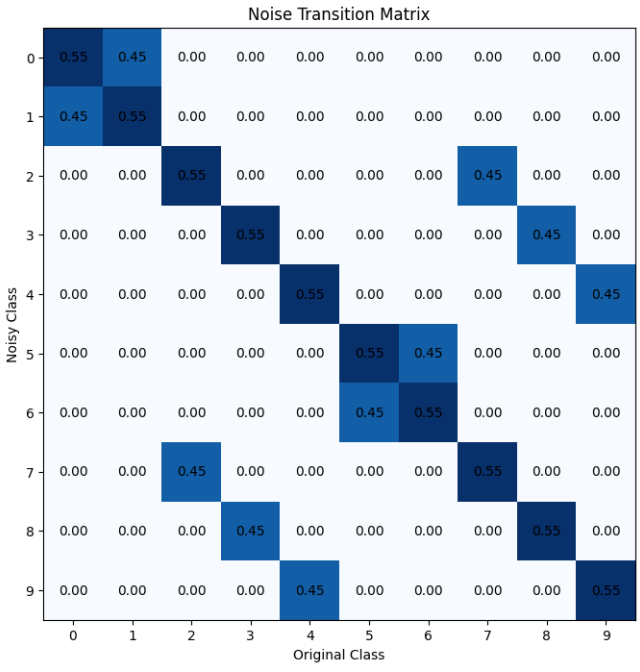


**Figure 3. Noise Transition Matrix for MNIST dataset with asymmetric Label Noise model with $\epsilon = 45\%$**

A 3-layer Convolutional Neural Network (CNN) architecture with ReLU activation functions was used for the compared methods, since it presented the best results in the literature for this dataset [Han et al. 2018]. The Network training was performed with a

Stochastic Gradient Descent (SGD) optimizer with a learning rate scheduler (initially set as 0.01), a batch size of 200, 30 epochs and Glorot normal initialization. The architecture is presented in Table 1. The experiments were conducted using Keras and TensorFlow frameworks.

The Forget Rate in both Ensemble Co-Teaching and Stochastic Co-Teaching was sampled from a symmetric Beta Distribution with $\alpha = 2$ and $\beta = 5$. For both Co-Teaching and Co-Teaching+, it was set as $p = \epsilon$. In Ensemble Co-Teaching, the perturbation was sampled from $\mathcal{N}(0, 0.01)$.

| Input | 28 X 28 Gray Image |
|---|---|
| Layer 1 | 3x3 conv, 32 ReLU |
| Layer 1 | 2x2 max-pool, stride 2 |
| Layer 2 | 3x3 conv, 64 ReLU |
| Layer 2 | 2x2 max-pool, stride 2 |
| Layer 3 | 3x3 conv, 128 ReLU |
| Layer 3 | avg-pool |
| Output | dense $128 \rightarrow 10$ |

**Table 1. CNN model used for the experiments with MNIST dataset.**

To measure the performance, the Test Accuracy was used in a Test dataset without the asymmetric Label Noise. For each method, the experiment was conducted 5 times with different initialization seeds.

## 4. Results

The average results for the Standard CNN without Robust Learning techniques, with Co-Teaching, Co-Teaching+, Stochastic Co-Teaching and Ensemble Co-Teaching are presented in Table 2 and Figure 4.

| Method | Average Test Accuracy |
|---|---|
| Standard | $82.7\% \pm 0.1\%$ |
| Co-Teaching | $88.9\% \pm 0.2\%$ |
| Co-Teaching+ | $86.5\% \pm 1.2\%$ |
| Stochastic Co-Teaching | $87.8\% \pm 0.7\%$ |
| Ensemble Co-Teaching | $91.0\% \pm 0.9\%$ |

**Table 2. Average test accuracy in 5 experiments.**

The standard CNN, without any robust learning method, achieved an average Test Accuracy of 82.7%, significantly lower than the average Test Accuracy in the dataset without Label Noise.

State-of-the-art baselines demonstrated similar performance, with no statistically significant difference between Co-Teaching and Stochastic Co-Teaching as determined by a Paired t-Test with a statistical significance of 95%. Since the noise rate is known for this dataset, there is no significant gain with the use of the stochastic process proposed in Stochastic Co-Teaching method.

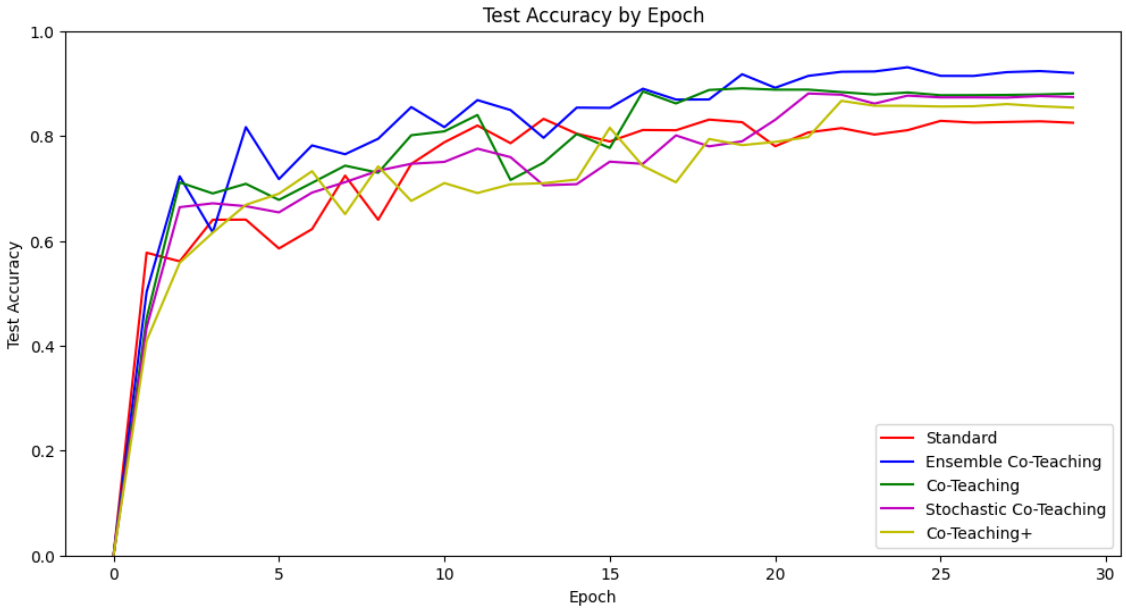


**Figure 4. Test accuracy vs. number of epochs on MNIST dataset.**

The Co-Teaching+ method showed lower performance compared to the others, as the limited number of samples used for weight updates (based on disagreement) hindered its learning capability. The average Test Accuracy for Co-Teaching+ was 86.5% that is more than 1 p.p. lower than the Test Accuracy achieved by Stochastic Co-Teaching.

Finally, the Ensemble Co-Teaching method achieved an average Test Accuracy of 91.0% significantly higher than other techniques (Co-Teaching and Stochastic Co-Teaching) with a statistical significance of 95%. This result demonstrates that perturbing the network weights ensures greater diversity among the networks and leads to a more robust selection of clean samples. Additionally, the instances $\tilde{D}_{12} = \tilde{D}_1 + \tilde{D}_2$ contribute to learning more diverse error patterns when updating the network weights.

## 5. Conclusions

In this paper, we proposed a new method for Robust Training of Deep Neural Networks under Label Noise: the Ensemble Co-Teaching. The proposed method combines concepts of Ensemble Learning into state-of-the-art robust learning techniques, such as Stochastic Co-Teaching.

The results indicate that introducing perturbations in the Network weights and utilizing a more diverse set of instances for weight updates effectively enhance the learning ability of Neural Networks through Ensemble Confident Learning. The proposed method achieved an average Test Accuracy significantly higher than the baseline techniques (Co-Teaching and Stochastic Co-Teaching) with a statistical significance of 95%. Experiments were developed on a benchmark dataset (MNIST) with an asymmetric Label Noise with $\epsilon = 45\%$.

Future works could involve evaluating this method on other benchmark datasets, such as CIFAR-10 and CIFAR-100, and with other types of label noise (symmetric and asymmetric with different noise rates $\epsilon$). Moreover, its performance could be assessed

with different architectures, such as Transformer networks.

## Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. It was made possible by the Thematic Projects of FAPESP "Life cycles and aerosol clouds in the Amazon" (2017/17047-0) and "Research Centre for Greenhouse Gas Innovation - RCG2I" (2020/15230-5).

## References

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. In *Proceeding of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.

Jiang, L., Zhou, Z., Leung, T., Li, L., and Fei-Fei, L. (2018). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceeding of the International Conference on Machine Learning (ICML 2018)*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE 1998*.

Lee, J. and Chung, S. (2020). Robust training with ensemble consensus. In *Proceeding of the 8th International Conference on Learning Representations (ICLR 2020)*.

Malach, E. and Shalev-Shwartz, S. (2017). Decoupling "when to update" from "how to update". In *Proceeding of the Conference on Neural Information Processing Systems (NIPS 2017)*.

Patrini, H., Rozza, A., Menon, A., Nock, R., and Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *Proceeding of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*.

Rooyen, B., Menon, A., and Williamson, R. (2015). Learning with symmetric label noise: The importance of being unhinged. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*.

Shen, Y. and Sanghavi, S. (2019). Learning with bad training data via iterative trimmed loss minimization. In *Proceeding of the International Conference on Machine Learning (ICML 2019)*.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):135–8153.

Vos, B., Jansen, G., and Isgum, I. (2023). Stochastic co-teaching for training neural networks with unknown levels of label noise. *Scientifc Reports*, 13(16875).

Wei, H., Feng, L., Chen, X., and An, B. (2020). Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceeding of the Conference on Computer Vision and Pattern Recognition (CVPR 2020)*.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., and Sugiyama, M. (2019). How does disagreement help generalization against label corruption? In *Proceeding of the International Conference on Machine Learning (ICML 2019)*.