Leveraging Adaptive Loss Scaling and Embeddings for Label Correction

Kirill Pupkov, Igor Ignashin, Aleksandr Beznosikov

2025

Abstract

In machine learning, numerous challenges can degrade model performance, including noisy features and incorrect labeling in the training data. Various approaches exist to mitigate these issues, such as Adaptive Loss Scaling. In this paper, we propose an improvement to this approach by incorporating a label correction mechanism. Our method replaces the labels of high-loss samples with those of their nearest neighbors in the embedding space. We demonstrate the effectiveness of our approach on the MNIST dataset with a fraction of noisy labels.

1 Introduction

Machine learning models are often applied under the assumption that training data are correctly labeled and accurately represent the true data distribution. However, in practice, training sets often contain noisy labels, negatively affecting model performance. Various approaches have been proposed in the literature to address this issue.

One common strategy involves filtering out objects whose labels do not match the labels of their neighbors. For instance, *Deep k-Nearest Neighbors filtering* (Bahri et al., 2020) uses a preliminary trained model to identify incorrectly labeled examples by comparing an object's label with those of its neighbors. This approach effectively removes mislabeled samples, improving the accuracy of the final model and often outperforming more complex robust training methods.

Another family of approaches focuses on reweighting training examples. Some methods utilize meta-learning to learn these weights (Ren et al., 2018), while others use specialized networks to determine them (Jiang et al., 2018). Recent work (Beznosikov, 2025) discusses how the task of assigning weights can be formulated as a min-max optimization problem and proposes a method for solving it using the ALSO optimizer.

In this paper, we propose a method to leverage the learned weights from the ALSO optimizer to progressively identify and relabel mislabeled training examples, improving model accuracy. Following the approach of Bahri et al. (2020), we utilize object embeddings to determine the correct labels.

The main contributions of our paper include:

- **A new method for label correction.** We introduce a novel approach leveraging recent advances in robust learning and embedding-based methods.
- **Empirical evaluation.** We extensively evaluate our approach and compare it against state-of-the-art methods, demonstrating superior performance.

Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a clean dataset where y_i is the true label for the object x_i. We observe a noisy training set $\tilde{\mathcal{D}} = \{(x_i, \tilde{y}_i)\}_{i=1}^n$, where $\tilde{y}_i$ may differ from y_i.

We typically optimize the following objective:

$$\min_{\theta\in\Theta} \left\{ \frac{1}{n}\sum_{i=1}^n f_i(\theta) + \frac{\tau}{2}\|\theta\|_2^2 \right\}, \tag{1}$$

where $f_i(\theta) := L(q(\theta, x_i), \tilde{y}_i)$ is the loss function computed using the noisy labels, $\tau > 0$ is a regularization parameter, and $\frac{\tau}{2}\|\theta\|_2^2$ is the regularization term to prevent overfitting. However, as previously mentioned, optimizing this objective directly might lead to memorization of incorrect labels and therefore poor performance.

To address this, Beznosikov (2025) propose the following reformulation:

$$\max_{\pi\in\Delta_{n-1}} \min_{\theta\in\Theta} \left\{ \sum_{i=1}^n \pi_i f_i(\theta) + \frac{\tau}{2}\|\theta\|_2^2 - \tau \mathrm{KL}[\pi\|\hat{\pi}] \right\}, \tag{2}$$

where $\tau > 0$ is the regularization parameter (temperature), $\mathrm{KL}[\cdot\|\cdot]$ denotes the KL-divergence between two distributions, and $\hat{\pi}$ is a prior distribution of weights from Δ_{n-1}. A simple baseline choice for $\hat{\pi}$ is the uniform distribution $\hat{\pi} = \mathcal{U}(\mathbf{1}, n)$. This formulation encourages higher weights for samples that are more difficult to classify correctly, helping identify potentially mislabeled data.

In later sections, we demonstrate how to leverage iteratively computed weights π_i to effectively identify and correct mislabeled data points.

2 Experiments

2.1 Data

To evaluate the effectiveness of our algorithm, we conducted experiments using the MNIST dataset (Deng, 2012), which consists of 60,000 28×28 grayscale images of handwritten digits. To simulate label noise we randomly selected a proportion p of samples from the training set and assigned them incorrect labels. We then compared our algorithm against baseline methods across various noise levels to assess its performance.

2.2 Evaluation Metrics

We evaluated the performance of our approach using multiple metrics:

1. **Precision**: The fraction of samples identified as mislabeled that were actually mislabeled

2. **Recall**: The fraction of truly mislabeled samples that were successfully identified by the algorithm

3. **Test Accuracy**: The classification accuracy achieved on the clean, uncorrupted test set after training with our noise-handling approach

All experiments were run 5 times with different random seeds, and we report 95% confidence intervals in our plots to account for variability.

2.3 Baseline

2.4 Results

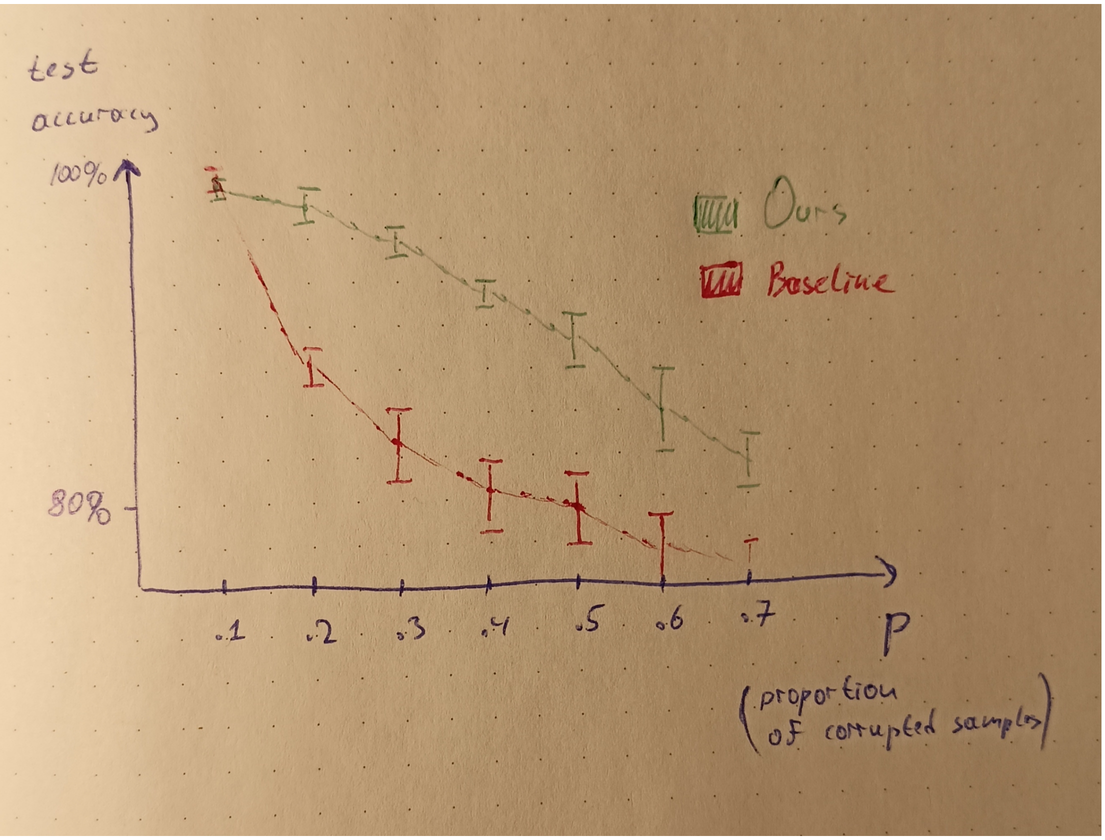

Figure 1: Performance comparison of our proposed method versus baseline approach across different noise levels. Bars represent 95% confidence intervals over 5 runs.

As shown in Figure 1, our method consistently outperforms the baseline approach in identifying mislabeled examples, particularly as the noise level increases.

References

Bahri, M., Ollion, C., Denoyer, L., and Gallinari, P. (2020). Deep k-nn filtering for noisy labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Beznosikov (2025). Mirror-prox algorithm with linear convergence rate and its application for dynamic loss scaling.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. (2018). Mentornet: Learning data-driven curriculum for very noisy supervision. In *ICML*.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. (2018). Learning to reweight examples for robust deep learning. In *ICML*.