Leveraging Adaptive Loss Scaling for Noisy Label Filtering

Kirill Pupkov Igor Ignashin Aleksandr Beznosikov

Moscow Institute of Physics and Technology

2025

Goals

Goal

Develop a more effective method for identifying mislabeled examples in classification task.

Problem

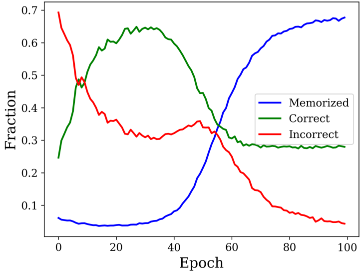
Machine learning models easily overfit noisy labels in training datasets.

Proposed Solution

- Utilize Adaptive Loss Scaling (ALSO) optimizer to learn sample weights.
- Apply a Beta Mixture Model (BMM) to these learned weights to separate clean and noisy examples.

Problem

Training data $D = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$ often contains labels $\tilde{y}_i$ that are corrupted versions of the true labels y_i (with probability η).



- **Memorization Effect:** Deep neural networks have the capacity to perfectly fit (memorize) even random labels

Figure: Memorization on CIFAR10 with 80% symmetric noise (ResNet18)

Literature

- Zhang *et al.* (2017): Deep networks can fit random/noisy training labels
- (2025): ALSO optimizer for dynamic loss scaling
- Arazo *et al.* (2019): Beta Mixture Models for loss modeling
- Patrini *et al.* (2017): Bootstrapping methods to make DNNs robust to label noise

ALSO

Original ALSO Optimizer

- Authors reformulate Empirical Risk Minimization as a min-max problem:

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

  where $f_i(\theta) = \ell(y_i, q_\theta(x_i))$, π are sample weights.
- This encourages higher weights for samples that are more difficult to classify.

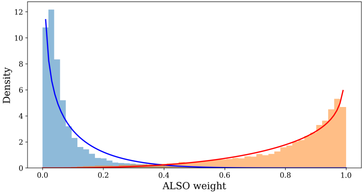
Modification

- We replace maximum over π with minimum to get smaller weights for noisy samples and prevent overfitting

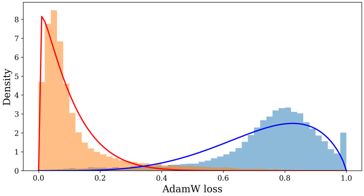
Beta Mixture Model

Model loss or weight distribution as a mixture of two Beta destributions. Fit the model using EM algorithm.

$$p(l) = \sum_{k=1}^{2} w_k \cdot p(l|k); \quad p(l|k) = \frac{l^{\alpha_k - 1}(1-l)^{\beta_k - 1}}{B(\alpha_k, \beta_k)}$$



(a) ALSO weights



(b) AdamW losses

Figure: BMM fitted to modified ALSO weights (left) and AdamW losses (right) under 0.4 symmetric noise

Experiment

- Hypothesis: Our method separates noisy and clean labels better and is less prone to overfitting
- Data: CIFAR10 with symmetric or asymmetric noise at various levels
- Network Architectures: ResNet-18, ResNet-34.
- Setup: Our method (ALSO + BMM) vs. Baseline (AdamW + BMM).
- Metric: ROC-AUC to measure separation of clean and noisy labels.

Results and Conclusion

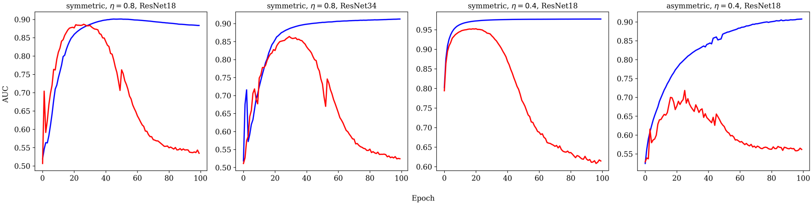


Figure: Blue: Our method (ALSO), Red: Baseline (AdamW)

- Our method consistently outperforms the baseline, achieving AUC > 0.9.
- Our method is more resistant to overfitting
- **Future work:** incorporate the approach into existing semi-supervised learning pipelines to correct noisy labels