Learning with Noisy Labels Using Transfer Learning and Loss-Based Noise Mitigation on CIFAR-100N

Aliciuc Alexandru-Gabriel, Țiplea Matei

January 8, 2025

Abstract

As the presence of noisy labels in real-world datasets poses significant challenges for machine learning models, we attempt to address the problem of learning from noisy labels using the CIFAR-100N dataset, containing 40% noisy labels. We propose a framework that combines transfer learning and fine-tuning of a pretrained model on another dataset, effective but simple small loss-based noise-handling strategies, and advanced data augmentation techniques. The ConvNeXt Base model, pretrained on large-scale datasets, serves as the backbone for our approach. The proposed methodology achieves a final validation accuracy of 70.4% on CIFAR-100N, demonstrating its effectiveness in handling noisy labels. This work highlights the importance of model architecture, noise mitigation, and augmentation techniques in noisy label scenarios and provides a practical blueprint for future research and applications, attempting to balance performance with good results.

1 Introduction

In machine learning, high-quality labeled data is a cornerstone for training accurate and reliable models. However, in many real-world scenarios, achieving clean and precise annotations is challenging, leading to datasets contaminated with noisy labels. Noisy labels arise from human errors, ambiguous categories, or inconsistencies in labeling processes, and they can significantly hinder a model's ability to generalize effectively. This issue becomes particularly pronounced in large-scale datasets where manual verification of every annotation is infeasible.

The CIFAR-100N dataset exemplifies this challenge by incorporating real-world annotation noise into the widely used CIFAR-100 dataset. With 100 fine-grained classes, CIFAR-100N introduces complexities that closely mimic real-world noisy labeling conditions, providing a valuable benchmark for exploring robust learning methodologies. Unlike synthetic noise often used in research, the human-induced noise in CIFAR-100N mirrors the challenges faced in practical machine learning applications, such as crowd-sourced labeling or subjective annotation tasks.

Addressing noisy labels requires innovative solutions that combine robust model architectures, data processing techniques, and noise-mitigation strategies. Traditional machine learning algorithms, designed under the assumption of accurate labels, often fail to perform well in these conditions. As a result, there has been a growing body of research focused on designing models and training strategies that can effectively learn even in the presence of noise.

In this study, we aim to contribute to this growing area by exploring the intersection of advanced neural architectures and noise-handling strategies, specifically within the context of the CIFAR-100N dataset. By investigating both foundational and experimental approaches to mitigate the impact of noise, we not only aim to improve performance on this challenging benchmark but also derive insights into broader applications of these techniques in noisy real-world environments. This investigation is a step toward bridging the gap between academic research and practical deployment of machine learning models in imperfect data settings.

2 Proposed Methods

2.1 Overview

To address the challenges posed by noisy labels in the CIFAR-100N dataset, our methodology integrates transfer learning and fine tuning of robust pretrained model architectures, noise-handling strategies based on small loss selection, and augmentation techniques, each contributing to a holistic solution for learning in noisy label environments. Our final solution uses the **ConvNeXt Base** model as the backbone of our training, which is pretrained on ImageNet, leveraging the power of transfer learning for the initial part of the training. Afterwards, we unfreeze the rest of the layers in order to further more fine tune our model. After the model has "warmed up" and learned certain characteristics of the dataset, we use the so far trained model for filtering the potential noise in the data by employing loss-based filtering (a smaller loss on a particular input means that the model is more certain of a prediction, and, as such, the probability that the labeled data is noise is smaller), and continue training this model on the filtered data. Finally, we use test-time augmentation in order to improve the final results.

2.2 Transfer Learning & Fine-Tuning

At the heart of our framework lies the ConvNeXt Base model, a cutting-edge convolutional neural network (CNN) designed to leverage the latest advancements in deep learning for computer vision tasks. ConvNeXt combines the computational efficiency of modern convolutional architectures with the performance benefits traditionally associated with transformer-based models, making it a strong candidate for tasks involving noisy datasets like CIFAR-100N. Its use of pretraining on large-scale datasets provides a solid foundation for transfer learning, enabling the model to adapt effectively to the challenges of a noisy label environment.

Fine-tuning is a critical component of our approach, allowing us to adapt the pretrained ConvNeXt Base model to the specific characteristics of the CIFAR-100N dataset. Initially, we freeze the majority of the model's layers to preserve the rich features learned during pretraining. Only the final layers, which are responsible for the task-specific classification, are allowed to train in the early stages. This ensures the model focuses on adapting to the dataset-specific nuances without disrupting the pretrained representations.

As training progresses, we introduce an unfreezing phase—a technique designed to gradually unlock more layers for training after a designated epoch. This approach balances the benefits of leveraging pretrained features with the flexibility needed to refine deeper layers for the CIFAR-100N task. The unfreezing process also aligns with the warmup phase of our noise-handling strategy, providing stability during the early stages of training and gradually increasing the model's capacity for fine-tuning as it becomes more robust to label noise.

By integrating transfer learning and fine-tuning through strategic layer freezing and unfreezing, we maximize the potential of the ConvNeXt Base model, enabling it to generalize effectively despite the challenges posed by noisy labels. This approach not only accelerates training convergence but also reduces the risk of overfitting, particularly in the presence of mislabeled samples.

2.3 Noise Handling

Dealing with noisy labels is one of the most challenging aspects of training robust machine learning models. In our framework, we address this issue using a combination of loss-based filtering, warmup strategies, and a dynamic loss-threshold decay mechanism. These techniques collectively enable the model to learn effectively from noisy data while minimizing the adverse impact of mislabeled samples.

2.3.1 Loss-Based Filtering

Loss-based filtering is a strategy that dynamically identifies and mitigates the influence of noisy labels during training. The core idea is to monitor the loss associated with individual samples: high-loss samples are more likely to be mislabeled. This loss is calculated after an initial warmup phase, allowing the model to learn ceartain features of the dataset. We then discard these potentially mislabeled samples, allowing the model to train on more accurate data.

2.3.2 Warmup Phase

To ensure a stable start to training, we incorporate a warmup phase, during which the model is trained without applying loss-based filtering. This initial phase allows the model to learn basic patterns and features from the data without being prematurely influenced by noisy labels. The warmup phase also helps stabilize gradients, laying a strong foundation for subsequent noise-handling steps.

2.3.3 Loss-Threshold Decay

After the warmup phase, we gradually introduce a dynamic loss-threshold decay mechanism. This involves setting an initial loss threshold for filtering noisy samples and progressively relaxing it over time. The decay allows the model to accommodate a broader range of samples as it becomes more robust to noise, balancing the need to filter out highly noisy samples with the potential value of moderately noisy data.

2.4 Data Augmentation

Augmentation is especially valuable in the context of noisy labels, as it helps mitigate overfitting to mislabeled samples by providing a diverse and enriched set of training examples. In our approach, we employed a carefully curated set of augmentation techniques tailored to the CIFAR-100N dataset.

2.4.1 Random Crop and Horizontal Flip

We applied a random cropping operation with padding to simulate slight spatial variations in the input images. This is complemented by random horizontal flipping, which introduces further variability by altering the orientation of the images. These transformations mimic natural variations that might occur in real-world scenarios, improving the model's robustness to spatial and geometric changes.

2.4.2 RandAugment

To further enhance the diversity of the training data, we incorporated RandAugment, an efficient augmentation technique that applies a random combination of transformations from a predefined set. RandAugment allows for the systematic exploration of augmented variants without requiring extensive manual tuning of parameters. By introducing transformations such as rotation, color adjustment, and contrast changes, RandAugment enriches the dataset and forces the model to focus on meaningful features rather than overfitting to noise or specific patterns in the original data.

2.4.3 Random Erasing

Another augmentation technique we utilized is random erasing, which randomly occludes a portion of an image by replacing it with a solid color or noise. This operation simulates real-world scenarios where parts of objects might be obscured and encourages the model to focus on broader contextual information rather than specific, potentially noisy, details in the image.

2.4.4 Test-Time Normalization

For the test set, we applied normalization using dataset-specific mean and standard deviation values. This ensures consistent scaling and centering of pixel values, aligning the test data with the preprocessed training data for robust inference.

The augmentation techniques are designed to complement our noise-handling strategies. By introducing controlled randomness and variability, the augmented data reduces the likelihood of overfitting to noisy labels. This synergy ensures that the model learns to generalize from diverse training scenarios while mitigating the adverse effects of label noise.

2.5 Test Time Augmentation

Test-Time Augmentation (TTA) is a technique employed during inference to enhance model robustness and improve prediction accuracy. The core idea behind TTA is to apply various transformations

to each test image, generate predictions for the augmented variants, and then aggregate these predictions (typically through averaging). This reduces the variance in predictions and improves the model's ability to generalize to unseen data.

In our framework, TTA is implemented by applying transformations such as horizontal flips to the test images. By leveraging multiple augmented views of the same input, the model's predictions become less sensitive to specific artifacts or noise present in individual samples. This approach is particularly beneficial in noisy datasets like CIFAR-100N, where label inconsistencies or input variations can otherwise mislead the model during inference. TTA complements our overall strategy, contributing to more stable and reliable predictions without requiring additional training or model complexity.

3 Experimental Results

Our framework for tackling noisy labels in the CIFAR-100N dataset (with 40% noise) was implemented using PyTorch and leverages a carefully designed training pipeline to ensure optimal performance. Below, we detail the hyperparameter choices, model architecture, optimizer configurations, and training strategies employed in our experiments.

3.1 Implementation Details

3.1.1 Model Architecture

We utilized the ConvNeXt Base model as the backbone for our experiments, selected for its advanced feature extraction capabilities and scalability. This model, pretrained on large-scale datasets, served as a robust starting point for transfer learning. To adapt the model to the CIFAR-100N task, we replaced its classification head with a custom layer stack, including an adaptive average pooling layer, a dropout layer (0.5 dropout rate), and a fully connected layer to output predictions across 100 classes.

During the initial training epochs, we froze the backbone layers to preserve the pretrained representations, fine-tuning only the classification head. Starting at epoch 15, we unfroze the backbone layers to allow the entire model to adapt to the dataset, effectively balancing computational efficiency with the need for fine-tuning.

3.1.2 Data Preprocessing and Augmentation

The training pipeline included various data preprocessing and augmentation strategies. Input images were normalized using the CIFAR-100 dataset's mean and standard deviation values to ensure consistency. Augmentation techniques such as random cropping with padding and random horizontal flipping introduced spatial variability, while RandAugment and random erasing enhanced the dataset's diversity by applying randomized transformations and occluding regions of the image, respectively. For the test set, we applied normalization to align it with the preprocessed training data.

3.1.3 Optimizer and Learning Rate Schedule

The model was trained using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 5e-4, balancing efficient convergence with regularization. The learning rate started at 0.001 and was linearly warmed up over the first five epochs, which stabilized training and improved early performance metrics. After the warmup phase, a cosine annealing scheduler was applied to decay the learning rate progressively, tailoring it to the evolving needs of the training process.

3.1.4 Noise Handling and Training Strategy

To mitigate the impact of noisy labels, we incorporated a loss-based filtering (based on Cross Entropy) mechanism starting after a warmup phase of 25 epochs. During this phase, the model trained without filtering, allowing it to establish a strong baseline for identifying noisy samples. Once filtering was introduced, a loss threshold of 2.0 was used to filter samples with higher loss values, which are more likely to be mislabeled.

The filtering process was dynamically refined using a loss-threshold decay factor of 0.995, progressively lowering the threshold as training progressed. By epoch 50, for example, the threshold had

decreased to approximately 1.77, enabling the model to tolerate moderately noisy samples while still focusing on cleaner examples.

3.1.5 Mixed-Precision Training and Device Configuration

The training pipeline leveraged mixed-precision training to improve computational efficiency without compromising accuracy. This was achieved using PyTorch's torch.cuda.amp module and the GradScaler for dynamic scaling of gradients. All experiments were conducted on a GPU-enabled environment with CUDA optimizations and pinned memory to maximize data transfer efficiency.

3.1.6 Batch Processing and Data Loading

To streamline training and evaluation, we utilized custom data loaders with batched data processing. The training data was organized into batches of size 100, while the test set was processed in batches of size 500 to balance memory usage and throughput. Pin memory was enabled in the data loaders to accelerate GPU data transfers.

3.1.7 Checkpointing and Metrics Monitoring

To ensure reproducibility and facilitate model evaluation, we implemented a checkpointing mechanism that saved the model's state whenever it achieved a new best validation accuracy. Additionally, training metrics such as accuracy, loss, and learning rate were monitored and logged throughout the process, enabling real-time tracking of model performance and training dynamics.

By combining these carefully chosen design elements, our implementation provides a robust and efficient pipeline for tackling noisy label problems in CIFAR-100N, demonstrating the potential of advanced architectures and noise-mitigation strategies in challenging real-world datasets.

3.2 Results

The model was trained for 100 epochs, in a total time of approximately 90 minutes on an RTX 4070 Super graphics card, showcasing efficient utilization of hardware resources. Training results are showcased in Figure 1 and Figure 2, resulting in a validation accuracy of 70.1%, which is further improved to a **final validation accuracy of 70.4%** after employing the TTA (test-time augmentation) step.

We can observe the three training phases on these graphs: the transfer learning phase (pre-epoch 15), the fine-tuning phase (epochs 15-25), and the noise handling phase. The model reaches a validation accuracy of 63% at the end of the 25th epoch, after which it would continue to stagnate if left as it is. As such, we considered that this was the right time to introduce our noise handling mechanism. As the number of noisy labels greatly decreases after the 25th epoch, a high jump in training accuracy can be observed, resembling overfitting, but leading to a not so high but significant increase of approximately 7% in validation accuracy.

As such, although not reaching state-of-the-art performance, the results are very promising considering the relatively simple techniques used, demonstrating the capabilities of transfer learning and noise handling techniques.

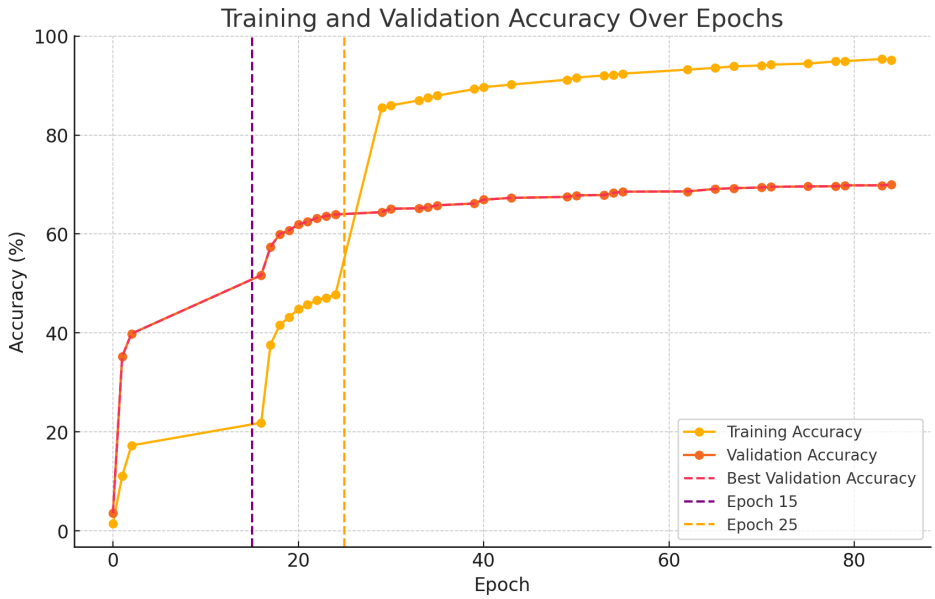

Figure 1: Training and Validation accuracy over epochs.

Figure 2: Validation Loss over epochs

3.3 Ablation Study

3.3.1 Effect of the backbone model

The backbone model played one of the most critical roles in our experiments as it influences feature extraction, dictating the performance of the entire framework. We compared the ConvNeXt Base model, selected for its modern design and robustness, against more traditional variations of the ResNet architecture. ConvNeXt, inspired by recent advancements in deep learning, utilizes a refined convolutional design that bridges the gap between convolutional and transformer-based approaches.

Our experiments revealed that ConvNeXt outperformed ResNeXt50 by approximately 15% in validation accuracy on CIFAR-100N and converged to good results much quicker, while trading off some performance. This improvement can be attributed to ConvNeXt's more efficient feature representation

and its ability to generalize better in the presence of noisy labels. Moreover, ConvNeXt's pretrained weights, fine-tuned on extensive datasets, provided a strong foundation for transfer learning, enabling better convergence. These findings solidified our choice of ConvNeXt as the backbone for this study.

3.3.2 Effect of the noise handling strategy

The integration of noise-handling strategies, including loss-based filtering and dynamic loss-threshold decay, also had a profound impact on the model's ability to generalize effectively. Without these strategies, the model achieved a validation accuracy of approximately 63%, struggling to overcome the challenges posed by noisy labels. When noise handling was introduced, validation accuracy improved significantly, reaching our final result of 70.4% accuracy — a relative improvement of approximately 7

3.3.3 Effect of the Test-Time Augmentation step

Test-time augmentation was applied during inference to enhance robustness and reduce prediction variance. This technique introduced minor transformations, such as horizontal flips, to the test images, aggregating predictions across these augmented variants.

While TTA's impact was more modest compared to noise handling, it still contributed a measurable improvement of 0.3% in validation accuracy. This gain underscores TTA's value in fine-tuning model predictions and achieving better alignment between training and inference conditions. Although the improvement may seem small, it is significant in scenarios where even minor gains can be critical for achieving state-of-the-art results.

4 Conclusions

In this study, we tackled the challenging problem of learning from noisy labels using the CIFAR-100N dataset. By leveraging a combination of advanced transfer learning techniques, robust noise-handling strategies, and effective data augmentation, we demonstrated the potential for achieving generalization in noisy label environments.

Our approach was built around the ConvNeXt Base model, which outperformed more traditional architectures such as ResNeXt50 by a significant margin. The use of fine-tuning, along with the strategic freezing and unfreezing of layers, allowed us to capitalize on the strengths of transfer learning, achieving rapid convergence and high accuracy. Additionally, noise-handling mechanisms, including loss-based filtering and dynamic loss-threshold decay, proved instrumental in mitigating the impact of these noisy labels.

These results, while not state-of-the-art, are highly encouraging given the simplicity and performance of the techniques employed. They demonstrate the value of transfer learning, noise-handling strategies, and test-time augmentation as practical tools for addressing noisy label challenges in real-world datasets.

Future work can explore the incorporation of more sophisticated noise detection algorithms, improvements in the backbone model, and more advanced augmentation and regularization techniques that can overcome issues such as overfitting. Also, improvements can be done regarding the filtered samples in the dataset, using them as part of more advanced semi-supervised learning approaches.

References

[1] Cristian Simionescu George Stoica. *FII ATNN 2024 - Project - Noisy CIFAR-100*. https://kaggle.com/competitions/fii-atnn-2024-project-noisy-cifar-100. Kaggle. 2024.

[2] Jiaheng Wei et al. *Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations*. 2022. arXiv: 2110.12088 [cs.LG]. URL: https://arxiv.org/abs/2110.12088.

[3] Zhuang Liu et al. *A ConvNet for the 2020s*. 2022. arXiv: 2201.03545 [cs.CV]. URL: https://arxiv.org/abs/2201.03545.

[4] Saining Xie et al. *Aggregated Residual Transformations for Deep Neural Networks*. 2017. arXiv: 1611.05431 [cs.CV]. URL: https://arxiv.org/abs/1611.05431.

[5] Ruixuan Xiao et al. *ProMix: Combating Label Noise via Maximizing Clean Sample Utility*. 2023. arXiv: 2207.10276 [cs.LG]. URL: https://arxiv.org/abs/2207.10276.

[6] Xian-Jin Gui, Wei Wang, and Zhang-Hao Tian. *Towards Understanding Deep Learning from Noisy Labels with Small-Loss Criterion*. 2021. arXiv: 2106.09291 [cs.LG]. URL: https://arxiv.org/abs/2106.09291.