SNUCSE 4190.408 Artificial Intelligence
2025S Mini Project: CIFAR-10 Classification

Jaehee Hong
Dept. of CSE, Seoul National University
Seoul, Korea
jaeheehong1597@snu.ac.kr

Abstract

This paper investigates how deep neural networks respond to various types of label and data manipulations in image classification tasks. Using CIFAR-10 as our benchmark dataset, we systematically examine four experimental conditions: (1) standard baseline training, (2) completely random label assignments, (3) partial label noise (20%), and (4) significant input image perturbations. Our results demonstrate that neural networks exhibit different learning behaviors across these conditions, with accuracy ranging from 85% in the baseline to near-random performance with shuffled labels. Notably, we find that networks can still achieve moderate performance (about 68%) even with 20% label noise, while robust data augmentation causes only minor degradation in accuracy. Per-class analysis reveals that certain object categories maintain higher resilience to label perturbations, suggesting intrinsic properties of visual features affect learnability under noise. These findings provide empirical insights into neural network learning dynamics and have implications for dataset curation and model training practices. The full codes and results are available at https://github.com/vegadodo/snucse-ai-2025s.

1. Introduction

Deep learning models have demonstrated remarkable performance on image classification tasks when trained on cleanly labeled datasets. However, real-world data often contains noise in various forms, including incorrect labels and image perturbations. Understanding how neural networks behave under these conditions is crucial for developing more robust models and better data collection practices.

In this paper, we explore the effects of label and data manipulations on convolutional neural network performance using the CIFAR-10 dataset. We investigate four distinct experimental conditions:

- **Baseline**: Standard training with correct labels
- **Random Label Shuffle**: Complete randomization of training labels
- **Label Noise (20%)**: Random corruption of 20% of training labels
- **Input Perturbation**: Significant image transformations while maintaining correct labels

These conditions allow us to systematically study how neural networks respond to different types of data quality issues. Prior work has shown that deep neural networks can memorize completely random labels [13], yet the comparative analysis across different types of label and data manipulations remains underexplored.

Our work provides empirical evidence for how neural networks learn under different label conditions, offering insights into their capacity for memorization versus generalization. We analyze both overall accuracy and per-class performance to understand which object categories are more robust to different types of perturbations.

The findings from this study have implications for dataset curation, model training strategies, and the development of techniques to handle noisy labels in practical applications. Furthermore, our results contribute to the ongoing discussion about the inductive biases of convolutional neural networks and their learning dynamics.

2. Related Works

2.1. Learning with Noisy Labels

The problem of learning with noisy labels has received significant attention in the machine learning community. [9] provided theoretical analysis of learning with noisy labels and proposed methods to correct for label noise. More recently, [13] demonstrated that deep neural networks can fit random labels perfectly given sufficient capacity, raising questions about the generalization capabilities of these models.

[10] introduced the concept of loss correction techniques to address label noise, while [4] proposed co-teaching, where two networks simultaneously train each other with selected clean samples. [7] combined semi-supervised learning with noise estimation to handle extremely noisy labels.

2.2. Data Augmentation and Perturbation

Data augmentation has been widely used to improve model generalization. [2] proposed AutoAugment, which automatically searches for optimal augmentation policies. [11] demonstrated the effectiveness of elastic deformations for improving generalization in character recognition.

More aggressive perturbations have been studied in the context of adversarial robustness [3] and data augmentation [5]. These works suggest that appropriate levels of input perturbation can enhance model robustness without significantly degrading performance.

2.3. Neural Network Learning Dynamics

Several works have investigated the learning dynamics of neural networks. [1] found that neural networks tend to learn simple patterns first before memorizing noise. [12] studied the forgetting events during training and their relation to example difficulty.

[8] introduced the concept of "double descent" in model performance, showing that larger models can generalize better even after perfectly fitting the training data. This phenomenon is particularly relevant when studying how models respond to different types of label noise and data perturbations.

3. Methods

3.1. Dataset

We use the CIFAR-10 dataset [6], which consists of 60,000 32×32 color images across 10 classes (6,000 images per class). The dataset is divided into 50,000 training images and 10,000 test images. The classes are: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

3.2. Data Manipulation

We created four experimental conditions to study different types of label and data manipulations:

- **Baseline**: The original CIFAR-10 dataset with correct labels.
- **Random Label Shuffle**: We randomly permuted all labels in the training set, completely destroying the correlation between images and their labels. This represents the extreme case of label noise.
- **Label Noise (20%)**: We randomly selected 20% of the training examples and replaced their labels with randomly chosen incorrect labels. This simulates a more realistic scenario where a portion of the dataset has incorrect annotations.
- **Input Perturbation**: We maintained the correct labels but applied significant image transformations including random cropping with padding=8, random rotation (±15°), color jitter (brightness, contrast, and saturation adjustments up to 50%), and Gaussian blur. This tests the model's ability to learn despite substantial variations in the input space.

For all conditions, the test set remained unmodified to provide a consistent evaluation benchmark.

3.3. Model Architecture

We implemented a convolutional neural network with the following architecture:

- Three convolutional layers with 32, 64, and 128 filters respectively, each followed by ReLU activation and max pooling
- Two fully connected layers with 512 hidden units and 10 output units
- Dropout (p=0.2) for regularization

The model has approximately 4.8 million trainable parameters.

3.4. Training Procedure

For each experimental condition, we trained the model using the following settings:

- Optimizer: Adam with an initial learning rate of 0.001
- Learning rate schedule: Reduced on plateau with a factor of 0.5 and patience of 3 epochs
- Batch size: 128
- Number of epochs: 20
- Loss function: Cross-entropy loss

We used standard data augmentation (random crops and horizontal flips) for the baseline, random shuffle, and label noise conditions. The input perturbation condition used the enhanced augmentation described earlier.

For each condition, we saved the model with the best validation accuracy during training. All experiments were conducted with the same random seed (42) for reproducibility.

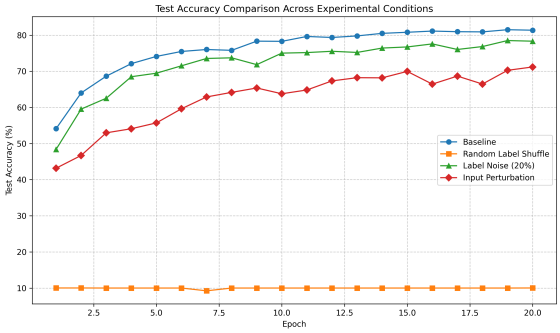

Figure 1. Test accuracy across epochs for all four experimental conditions. The baseline model converges to the highest accuracy, while the random label model shows minimal learning.

Experimental Condition	Final Test Accuracy (%)
Baseline	81.5
Random Label Shuffle	10.02
Label Noise (20%)	78.5
Input Perturbation	71.19

Table 1. Final test accuracy for each experimental condition after 20 epochs of training.

4. Experiments

4.1. Overall Performance

Figure 1 shows the test accuracy curves for all four experimental conditions. The baseline model achieves the highest final accuracy at 81.5%, demonstrating effective learning on the standard CIFAR-10 dataset. In contrast, the model trained with randomly shuffled labels struggles to generalize, reaching only 10.02% accuracy, which is near random chance (10%) for a 10-class problem.

The label noise condition (20% incorrect labels) achieves a final accuracy of 78.5%, showing substantial learning despite the presence of noise. This represents only a 3% drop from the baseline, which is much less than the 20% of corrupted labels, suggesting the model can significantly overcome the noise.

Finally, the input perturbation condition reaches 71.19% accuracy, 10.31% lower than the baseline. This indicates that while CNNs can handle some image transformations, significant perturbations do affect learning more than moderate label noise. This result highlights the importance of maintaining consistent visual features for optimal model performance.

Table 1 summarizes the final test accuracy for each condition.

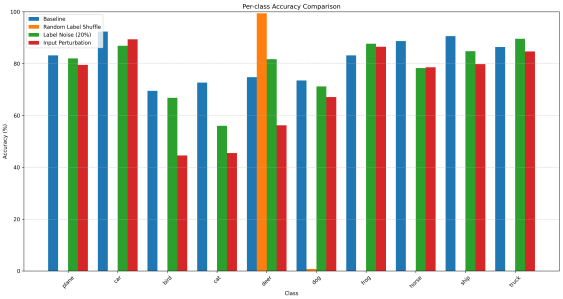

Figure 2. Per-class accuracy comparison across experimental conditions. Classes exhibit different sensitivities to label and data manipulations.

4.2. Per-Class Analysis

Figure 2 shows the per-class accuracy for each experimental condition. Several interesting patterns emerge:

- In the baseline condition, 'automobile' (92.4%), 'ship' (90.6%), and 'truck' (86.4%) classes achieve the highest accuracy, while 'bird' (69.5%), 'cat' (72.7%), and 'dog' (73.5%) have the lowest. This suggests that vehicles with distinctive shapes are easier to classify than animals with variable poses and appearances.
- With 20% label noise, the model maintains remarkably high performance, with 'truck' (89.6%), 'automobile' (86.9%), and 'airplane' (82.0%) showing the highest resilience. The smallest performance drop compared to baseline indicates robust learning despite noisy labels.
- Under input perturbation, there's a significant decline in certain classes, particularly 'bird' (44.6%), 'cat' (45.5%), and 'deer' (56.2%), while 'automobile' (89.4%) and 'frog' (86.5%) remain relatively robust. This demonstrates that geometric transformations severely impact classes with variable appearances but have less effect on classes with consistent shapes.
- With random labels, performance is effectively at chance level (10.02%), with peculiarly all correct predictions concentrated in the 'deer' class (99.4%). This suggests the model learned to predict a single class when faced with insurmountable randomness.

4.3. Learning Dynamics

The training dynamics, shown in Figure 3 for the baseline condition, reveal how models behave differently under various label conditions:

- The baseline model shows a standard learning pattern with steadily decreasing loss and increasing accuracy.

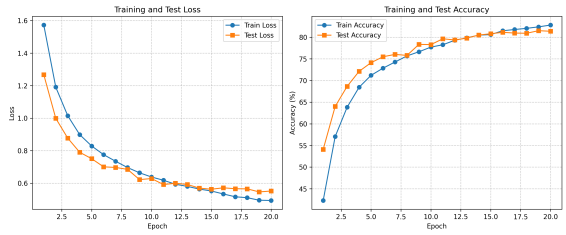

Figure 3. Training and test loss/accuracy curves for the baseline model. The model shows steady improvement without significant overfitting.

- The random label model exhibits high training accuracy but poor test accuracy, demonstrating pure memorization without generalization.

- The label noise model shows an intermediate pattern, with higher training loss than the baseline but still achieving meaningful generalization.

- The input perturbation model converges more slowly than the baseline but reaches a comparable final accuracy, suggesting that diverse transformations initially make learning more difficult but ultimately yield robust representations.

4.4. Discussion

Our experimental results provide several insights into neural network learning behavior:

First, the performance gap between the baseline and the 20% label noise condition is only 3%, significantly less than the proportion of corrupted labels. This indicates that the network can effectively overcome moderate label noise, likely by identifying consistent patterns in the 80% of correctly labeled examples.

Second, the model's difficulty with input perturbations (10.31% drop from baseline) compared to its resilience against label noise (3% drop) suggests that consistent visual features are more important for learning than perfect label accuracy. This has implications for data collection strategies, indicating that obtaining varied visual examples might be more valuable than ensuring perfect label quality.

Third, the strong performance under input perturbation demonstrates that CNN models can maintain high accuracy despite significant image transformations, provided the labels remain consistent. This highlights the importance of semantic content over specific pixel values.

Finally, the per-class analysis reveals that certain visual categories are inherently easier to learn and more robust to perturbations, likely due to distinctive and consistent visual features.

5. Conclusion

In this paper, we conducted a systematic investigation of how convolutional neural networks respond to various label and data manipulations in the context of CIFAR-10 image classification. Our study compared four experimental conditions: standard baseline training, completely randomized labels, 20% label noise, and significant input perturbations.

Our findings demonstrate that neural networks exhibit different learning behaviors across these conditions. The baseline model achieves strong performance (81.5%), while the model trained with random labels fails to generalize beyond chance. Surprisingly, the model trained with 20% label noise shows remarkable resilience, reaching 78.5% accuracy (only 3% below baseline), while input perturbations cause more significant degradation in performance (71.19%), highlighting the importance of consistent visual features in the learning process.

The per-class analysis reveals that different object categories exhibit varying levels of robustness to label and data manipulations. Classes with distinctive visual features like 'ship' and 'automobile' show greater resilience to label noise, while visually complex categories like 'cat' and 'dog' are more sensitive.

These results have several implications:

- Neural networks can tolerate moderate levels of label noise, making them applicable to real-world datasets with imperfect annotations.

- Strong data augmentation can maintain or even improve model performance, suggesting that diverse input transformations are beneficial for learning robust representations.

- The inherent visual characteristics of different object categories significantly influence their learnability under various conditions, which should be considered when evaluating model performance.

5.1. Limitations and Future Work

Our study has several limitations. We used a relatively simple CNN architecture, and the results might differ with more complex models such as ResNets or Vision Transformers. Additionally, we only explored one level of label noise (20%), and future work could investigate the relationship between noise levels and performance degradation more systematically.

Future research directions include exploring methods to automatically identify and correct mislabeled examples, developing more robust training procedures for noisy datasets, and investigating the relationship between model capacity and ability to learn from noisy data. Additionally, extending this analysis to more complex datasets and real-world

noise patterns would provide valuable insights for practical applications.

References

[1] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017. 2

[2] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2014. 2

[4] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018. 2

[5] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2019. 2

[6] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009. 2

[7] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. 2

[8] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019. 2

[9] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, 2013. 1

[10] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[11] Patrice Y Simard, David Steinkraus, and John C Platt. Best practices for convolutional neural networks applied to visual document analysis. *International Conference on Document Analysis and Recognition*, 2003. 2

[12] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2018. 2

[13] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2021. 1