Neural Network Training Dynamics on MNIST: A Comprehensive Analysis

Author One Author Two

2025-04-15

Abstract

We present a comprehensive analysis of neural network training dynamics on the MNIST dataset. Our work systematically investigates the impact of various architectural choices, optimization strategies, and regularization techniques on model performance, convergence behavior, and generalization capabilities. Through extensive experimentation, we demonstrate that careful consideration of hyperparameters leads to significant improvements in model accuracy and robustness. We also provide visual insights into the learning process, examining weight distributions, activation patterns, and representation spaces across training epochs. Our findings contribute to a deeper understanding of the fundamental principles governing neural network learning on image classification tasks.

1 Introduction

The MNIST dataset of handwritten digits has become a canonical benchmark for evaluating machine learning algorithms since its introduction by LeCun et al. [1]. Despite its apparent simplicity, the dataset continues to provide valuable insights into the behavior of learning algorithms, particularly neural networks. In this paper, we conduct a thorough investigation of neural network training dynamics on MNIST, focusing on several key aspects:

1. The progression of weight distributions and activation patterns during training
2. The impact of network architecture on learning efficiency and final performance
3. The effectiveness of various optimization algorithms and learning rate schedules
4. The influence of regularization techniques on generalization performance

Our analysis combines quantitative metrics with qualitative visualizations to provide a comprehensive understanding of the learning process. We believe this work contributes to the field by systematically documenting patterns in neural network training that may inform both theoretical research and practical applications.

2 Related Work

Neural networks have a rich history dating back to the perceptron model of Rosenblatt [2]. The backpropagation algorithm, formalized by Rumelhart et al. [3], enabled efficient training of multi-layer networks. The MNIST dataset itself was introduced as a benchmark for comparing machine learning methods for digit recognition [1].

More recent work has explored visualization techniques for understanding neural network internals [4] and mathematical frameworks for analyzing optimization dynamics [5]. Our work builds upon these foundations by providing a unified analysis that combines multiple perspectives.

3 Methods

3.1 Dataset

The MNIST dataset consists of 70,000 grayscale images of handwritten digits (0-9), with a standard split of 60,000 training images and 10,000 test images. Each image is 28×28 pixels, resulting in 784 features per sample when flattened. We apply standard preprocessing by normalizing pixel values to the range [0,1].

3.2 Model Architecture

We experiment with a family of convolutional neural networks (CNNs) of varying depths and widths. Our baseline architecture consists of:

- Two convolutional layers with 32 and 64 filters respectively, each followed by 2×2 max pooling
- A fully connected layer with 128 units and ReLU activation
- A softmax output layer with 10 units (one per digit class)

For regularization, we employ dropout with probability 0.5 on the fully connected layer.

3.3 Training Procedure

Models are trained using mini-batch stochastic gradient descent (SGD) with a batch size of 128. We evaluate several optimization algorithms including vanilla SGD, SGD with momentum, RMSProp, and Adam. Learning rates are initially set to 0.01 for SGD and 0.001 for Adam, with decay schedules applied as training progresses.

Training continues for 20 epochs, with model checkpoints saved at each epoch for subsequent analysis.

4 Results

4.1 Performance Metrics

Our best model achieves 99.3% accuracy on the test set, which is comparable to state-of-the-art results for similarly sized networks on MNIST. Figure 1 shows the progression of training and validation accuracy across epochs for different optimization algorithms.

Source: Article Notebook

4.2 Weight Distribution Analysis

We track the distribution of weights in each layer throughout training. Figure 2 shows how these distributions evolve from their initial Gaussian form to more complex multimodal distributions as learning progresses.

Source: Article Notebook

Source: Article Notebook

4.3 Activation Pattern Analysis

By visualizing the activation patterns for different input samples, we gain insights into the representations learned by the network. Figure 3 presents t-SNE projections of the activations in the penultimate layer, showing clear separation of digit classes.

5 Discussion

Our results demonstrate several key principles in neural network training:

1. Early layers tend to learn general features (e.g., edge detectors) while later layers specialize in digit-specific features
2. The Adam optimizer consistently outperforms SGD in terms of convergence speed, though final accuracies are comparable
3. Dropout significantly improves generalization, particularly for deeper architectures
4. Weight distributions shift from initial Gaussian distributions toward more complex, task-specific distributions

These observations align with the broader literature on deep learning while providing specific insights into the MNIST use case.

6 Conclusion

This comprehensive analysis of neural network training dynamics on MNIST provides valuable insights into the learning process. Our findings illustrate how architectural choices, optimization strategies, and regularization techniques interact to determine model performance and training behavior.

Future work could extend this analysis to more complex datasets and architectures, potentially uncovering whether the patterns observed here generalize to more challenging domains.

7 References

Source: Article Notebook

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998

[2] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological review*, vol. 65, p. 386, 1958

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986

[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *European conference on computer vision*, pp. 818–833, 2014

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*MIT press, 2016