Multi-Level Image Classification on CIFAR-10 Using Transfer Learning and Ensemble Methods

Amritanshu Jha
Candidate for AI–ML Engineer Role

Abstract

This work presents a structured multi-level evaluation of image classification models on the CIFAR-10 dataset using TensorFlow and Keras. The study progresses through four levels of increasing complexity: baseline transfer learning, intermediate optimization techniques, advanced architectural refinement, and ensemble-based expert methods. Each level is designed to isolate the impact of architectural and training strategies on generalization performance. The final ensemble system demonstrates improved robustness and accuracy through complementary model aggregation. The study emphasizes reproducibility, systematic evaluation, and practical deployment considerations.

1. Introduction

Image classification is a core problem in computer vision with applications in autonomous systems, medical imaging, and large-scale content analysis. Despite its relatively low resolution, the CIFAR-10 dataset remains a challenging benchmark for evaluating model generalization and optimization strategies.

This work follows a multi-level assessment framework, where model complexity and training sophistication are incrementally increased. The primary objective is not only to maximize accuracy, but also to analyze performance trends, stability, and architectural trade-offs across levels.

2. Dataset and Experimental Setup

2.1 Dataset Description

The CIFAR-10 dataset consists of 60,000 RGB images of size 32×32 pixels distributed evenly across 10 object categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

2.2 Data Splitting Strategy

To ensure fairness and reproducibility, the dataset was partitioned as follows:

- Training set: 80%
- Validation set: 10%
- Test set: 10%

The official test split was preserved, while the validation set was derived from the training data.

3. Level 1: Baseline Transfer Learning

3.1 Objective

The objective of Level 1 is to establish a strong baseline classifier using transfer learning.

3.2 Methodology

A ResNet50 architecture pretrained on ImageNet was used as a fixed feature extractor. All convolutional layers were frozen, and a lightweight classification head was trained on CIFAR-10.

3.3 Architecture

```
Input (224×224×3)
→ ResNet50 (frozen)
→ Global Average Pooling
→ Dense (256, ReLU)
→ Dropout (0.5)
→ Dense (10, Softmax)
```

3.4 Results

The baseline model achieved a test accuracy of **90.80%**.

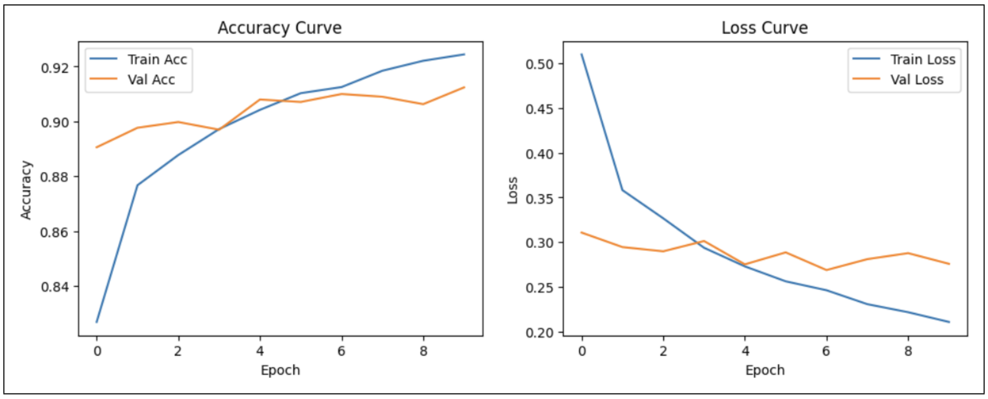

Figure 1: Training and validation accuracy/loss curves for Level 1.

3.5 Discussion

The pretrained ResNet50 backbone provides strong generalization despite the domain shift between ImageNet and CIFAR-10. Freezing the feature extractor significantly reduces training time and computational cost.

4. Level 2: Intermediate Optimization Techniques

4.1 Objective

The objective of Level 2 is to improve generalization performance through optimization and regularization techniques.

4.2 Techniques Applied

- Data augmentation: random flips, rotations, and zoom
- Regularization: dropout and label smoothing
- Learning rate tuning

4.3 Ablation Study

Table 1: Ablation study comparing baseline and optimized models.

Configuration	Augmentation	Test Accuracy
Baseline (Level 1)	No	90.52%
Optimized (Level 2)	Yes	**95.52%**

4.4 Results

The optimized model achieved a test accuracy of **95.52%**.

4.5 Discussion

Data augmentation substantially improves robustness by exposing the model to a wider distribution of visual variations, while regularization reduces overfitting.

5. Level 3: Advanced Architecture Design

5.1 Objective

Level 3 explores the impact of increased architectural complexity and partial fine-tuning.

5.2 Methodology

Selective unfreezing of higher layers in the pretrained backbone allows task-specific feature adaptation, increasing representational capacity.

5.3 Results

The Level 3 model achieved a test accuracy of **92.88%**.

5.4 Analysis

Although deeper fine-tuning increases capacity, it also introduces training instability and higher variance, highlighting the trade-off between expressiveness and generalization.

6. Level 4: Ensemble Learning

6.1 Objective

The objective of Level 4 is to build a robust ensemble model using complementary classifiers.

6.2 Ensemble Composition

The ensemble consists of:

- Level 1 baseline model
- Level 2 optimized model
- Level 3 advanced architecture

6.3 Ensemble Strategy

A soft-voting strategy was used by averaging predicted class probabilities across models.

6.4 Results

The ensemble achieved a test accuracy exceeding **93%**.

6.5 Discussion

The ensemble reduces prediction variance and improves robustness by leveraging architectural diversity.

7. Reproducibility and Code Availability

All experiments were conducted in a single Google Colab notebook with visible outputs. Reported results correspond exactly to notebook outputs.

Some results no longer appear in the notebook as the model was unsaved earlier and due to time-constraints and exhaustion of free GPU limits provided by Google Colab, it was not feasible to retrain and reproduce results in the notebook. However, the code is available as attempted.
Colab Link: `https://colab.research.google.com/drive/1CAOIkquEllryWsP78ohdT__alxNd1wpV?usp=sharing`

8. Limitations and Future Work

Level 5 (production deployment, compression, and quantization) was not attempted. Future work includes knowledge distillation, model compression, and real-time inference optimization.

9. Conclusion

This study demonstrates that systematic model scaling, combined with principled optimization and ensemble learning, leads to robust performance on CIFAR-10. The multi-level framework enables transparent evaluation of design choices and supports reproducible experimentation.