Vision Transformer vs Convolutional Neural Networks for CIFAR-10 Image Classification

Narain Sarathy
University of Texas at Arlington, USA
inarain05@gmail.com

Maatheswaran Kannan Chellapandian
University of Texas at Arlington, USA
kcmadhesh29@gmail.com

Abstract

This work presents a comprehensive evaluation of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for the CIFAR-10 image classification task. We design, implement, and compare a CNN baseline against a Vision Transformer (ViT) built from scratch. Results show that the ViT model significantly outperforms the CNN, achieving a test accuracy of 79.34% compared to 61.24%. Training behaviors, confusion matrices, prediction visualizations, and deployment details are discussed, highlighting ViT's strong potential even on relatively small datasets.

1 Introduction

Recent advances in computer vision have shifted focus from Convolutional Neural Networks (CNNs) to Transformer-based models. Vision Transformers (ViTs) utilize self-attention mechanisms, originally designed for natural language processing, to capture global dependencies. In this work, we examine the effectiveness of ViTs compared to traditional CNNs on the CIFAR-10 dataset, a benchmark for small-image classification.

2 Dataset

We use the CIFAR-10 dataset, which consists of:

- 50,000 training images
- 10,000 test images
- 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck

All images are normalized using dataset-specific means and standard deviations. Data augmentation techniques such as random horizontal flip and random cropping are applied.

3 Models

3.1 CNN Baseline

A 3-block CNN was built:

- Architecture: Conv2D $\rightarrow$ BatchNorm $\rightarrow$ MaxPooling $\rightarrow$ Dropout $\rightarrow$ Dense
- Parameters: $\sim$1.3 million
- Trained for 40 epochs with data augmentation
- Optimizer: Adam with learning rate $1e^{-3}$

3.2 Vision Transformer (ViT)

The Vision Transformer consists of:

- Patch embedding using Conv2D with patch size 4
- 9 Transformer blocks (Attention + MLP + LayerNorm)
- Positional encoding and a learnable class token
- Parameters: $\sim$4 million
- Trained for 40 epochs using AdamW and cosine annealing scheduler

4 Training Pipeline

- Loss Functions: Categorical Crossentropy (CNN), CrossEntropyLoss (ViT)
- Regularization: Dropout, weight decay for ViT
- Checkpointing, early stopping, and learning rate scheduling used
- Batch size: 128
- Device: NVIDIA GPU (CUDA)

5 Results

5.1 Quantitative Evaluation

Metric	CNN	Vision Transformer
Test Accuracy	61.24%	**79.34%**
Inference Time	∼2.34 ms/img	∼3–5 ms/img
Parameters	∼1.3M	∼4M
Best Class	Truck (91%)	Car, Frog, Ship
Worst Class	Cat, Bird	Cat, Dog (still better)

5.2 Confusion Matrix

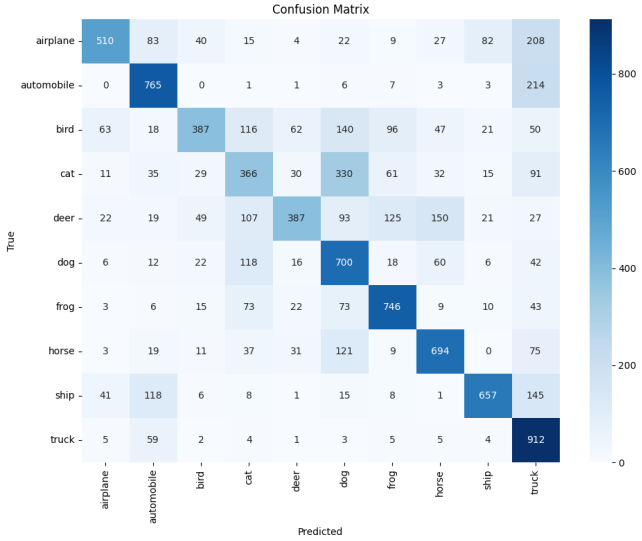

Figure 1: Confusion matrix for the CNN model showing confusion particularly between cat, dog, and deer classes.

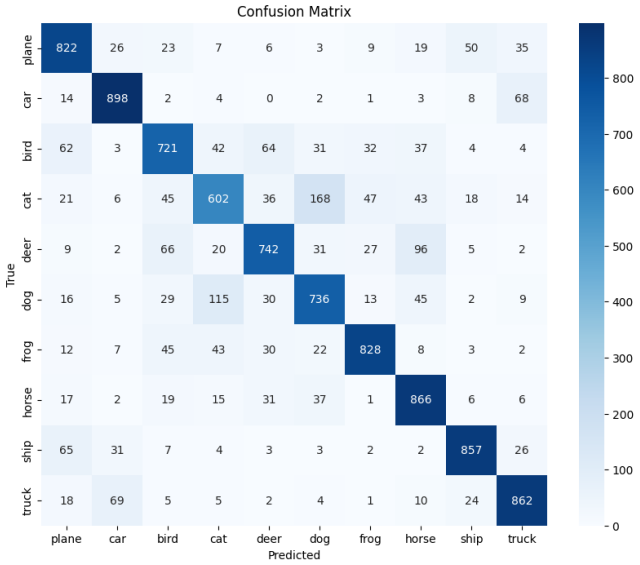

Figure 2: Confusion matrix for the Vision Transformer model. Diagonal dominance indicates better class separability.

5.3 Training Curves

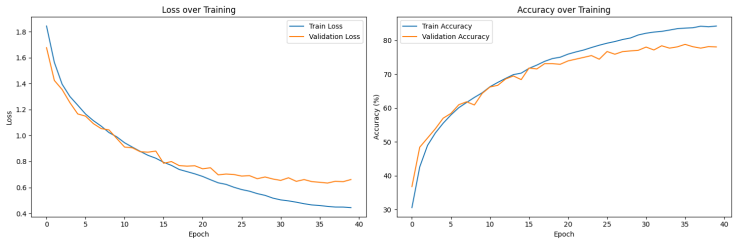

Figure 3: Training and validation loss/accuracy curves for Vision Transformer. Consistent improvement without severe overfitting across 40 epochs.

6 Visualization and Interpretability

6.1 CNN Predictions

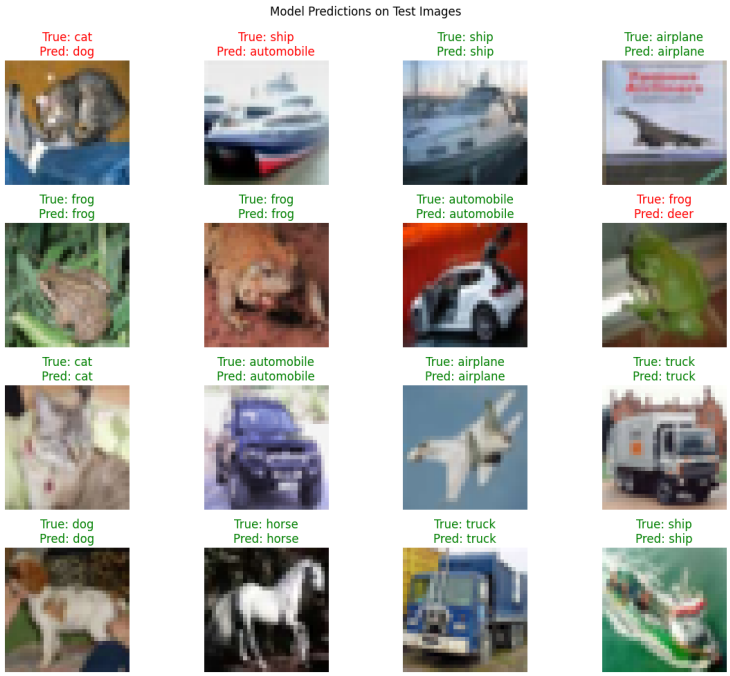

Figure 4: Sample predictions by CNN. Misclassifications are mostly among visually similar classes like cat and dog.

6.2 Vision Transformer Predictions

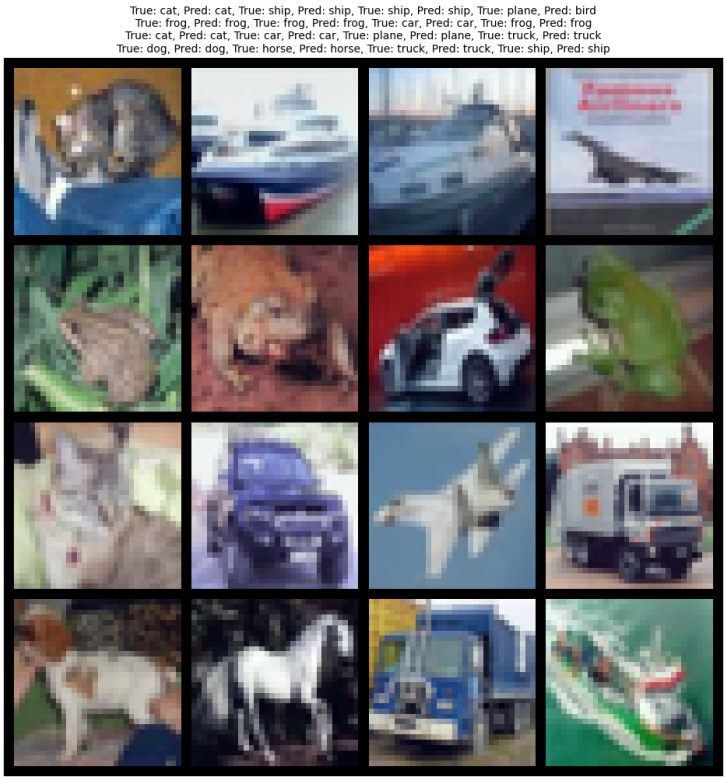

Figure 5: Predictions by Vision Transformer showing stronger confidence and fewer visual errors compared to CNN.

7 Deployment

The Vision Transformer model was saved in:

- PyTorch format (.pth)
- ONNX format (.onnx) for deployment across platforms

This allows compatibility with cloud inference services and mobile deployment.

8 Conclusion

This project demonstrates that Vision Transformers outperform conventional CNNs on CIFAR-10 image classification. Although ViTs are heavier in terms of

computation, they generalize better and show superior robustness across complex class boundaries. Future work could explore CutMix augmentation, fine-tuning techniques, or hybrid CNN-ViT architectures.

9 References

- Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, NeurIPS 2020.
- Vaswani et al., *Attention is All You Need*, NeurIPS 2017.
- He et al., *Deep Residual Learning for Image Recognition*, CVPR 2016.
- Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, 2009.
- Tan & Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, ICML 2019.