# CNN vs Vision Transformer (ViT) Performance Comparison

This report compares the performance of a Convolutional Neural Network (CNN) and a Vision Transformer (ViT) model on the CIFAR-10 dataset. The analysis covers the performance gap, training efficiency, and learning dynamics observed during training.

|  |  |  |
| --- | --- | --- |
| Metric | CNN Model | Vision Transformer (ViT) |
| Best Epoch | 11 | 10 |
| Train Accuracy | 94.63% | 67.15% |
| Train Loss | 0.1480 | 1.2415 |
| Val Accuracy | 86.00% | 66.69% |
| Val Loss | 0.4882 | 1.2398 |
| Test Accuracy | 85.42% | 65.77% |
| Test Loss | 0.4862 | 1.2463 |
| Performance Gap (Train–Test) | ~9.21% drop | ~1.38% drop |

## Analysis

1. Performance Gap  
- CNN shows a larger train–test gap (~9.2%), indicating some overfitting but still strong generalization.  
- ViT has a very small train–test gap (~1.4%), suggesting balanced training but overall lower accuracy, likely due to insufficient data or model complexity.

2. Training Efficiency  
- CNN converged faster to high accuracy, reaching above 80% validation accuracy within the first 10 epochs.  
- ViT had a more gradual improvement, plateauing around 66–67% accuracy, suggesting it needs more epochs, larger datasets, or stronger regularization.

3. Learning Patterns/Dynamics  
- CNN shows clear overfitting signs after its best epoch (accuracy rises but validation stagnates), common in small-to-medium datasets.  
- ViT maintained close tracking between training and validation accuracy, indicating stable learning but possible underfitting due to limited inductive bias.