A Comparative Analysis of Vision Transformers, CNN-MLP, and ResNet Models on CIFAR-10

Manahil Sarwar
Department of Artificial Intelligence
University Name
Section : AI-K

***Abstract*—This report presents a comprehensive analysis of three different deep learning architectures—Vision Transformer (VIT), Convolutional Neural Network with Multilayer Perceptron (CNN-MLP), and ResNet—on the CIFAR-10 dataset. We summarize the methodology, present detailed results, and discuss the strengths and weaknesses of each model architecture.**

I. Introduction

This report investigates the performance of three distinct architectures on CIFAR-10: Vision Transformer (VIT), CNN-MLP hybrid, and a pretrained ResNet50 model. The goal is to analyze their effectiveness in image classification by comparing their training accuracy, inference time, and performance metrics.

II. Methodology

A. Dataset and Preprocessing

We use the CIFAR-10 dataset, consisting of 60,000 32x32 color images across 10 classes. The data preprocessing included scaling the images to a range of [-1, 1] and applying data augmentation techniques like random flipping and cropping.

B. Model Architectures

1. **Vision Transformer (VIT)**: Divides images into patches and applies transformer layers for feature extraction. 2. **CNN-MLP Hybrid**: Combines a CNN layer for feature extraction with an MLP layer for classification. 3. **ResNet50**: A pretrained ResNet50 model with frozen layers and custom classification layers added.

III. Results

Training and validation accuracy and loss are shown for each model. Visualized plots for each metric are included.

A. Metrics

- **VIT Model**: Accuracy: 69.93%, Inference Time: 1.17s
- **CNN-MLP Model**: Accuracy: 55.49%, Inference Time: 1.01s
- **ResNet Model**: Accuracy: 33.8%, Inference Time: 8.33s

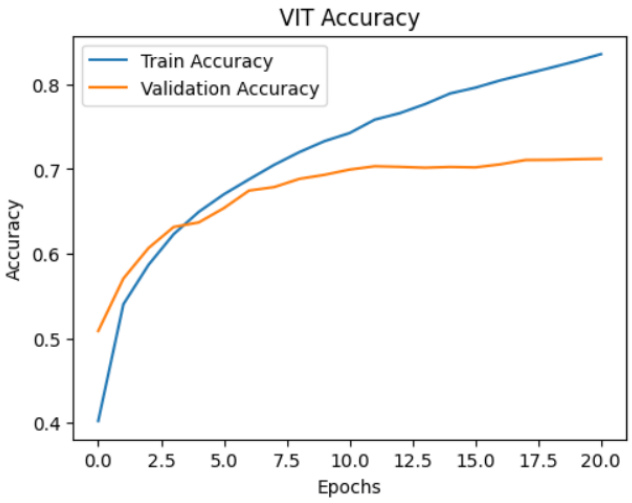

Fig. 1. Training Accuracy for VIT

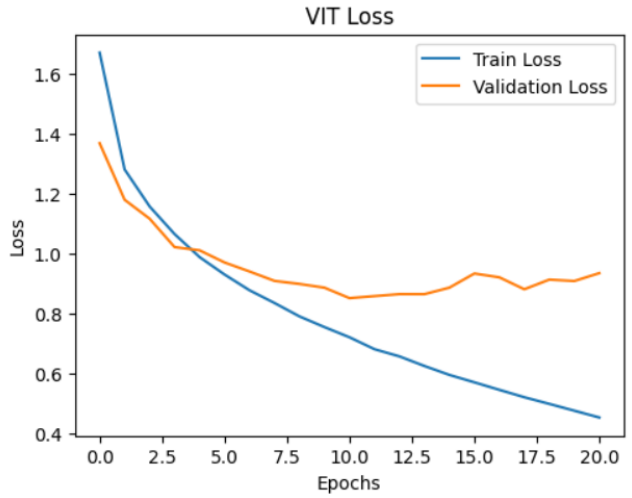

Fig. 2. Training Loss for VIT

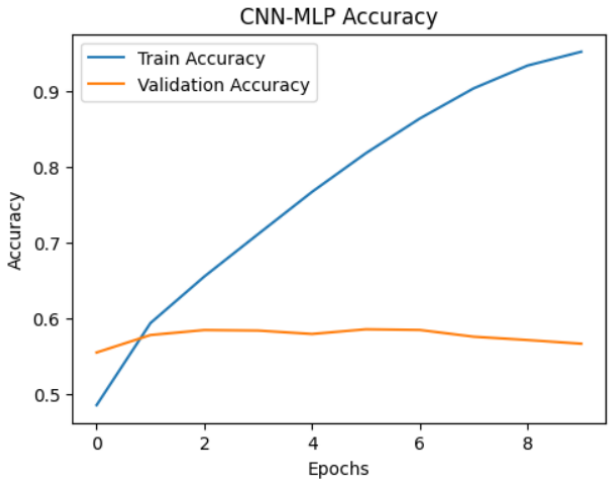

Fig. 3. Training Accuracy for CNN-MLP

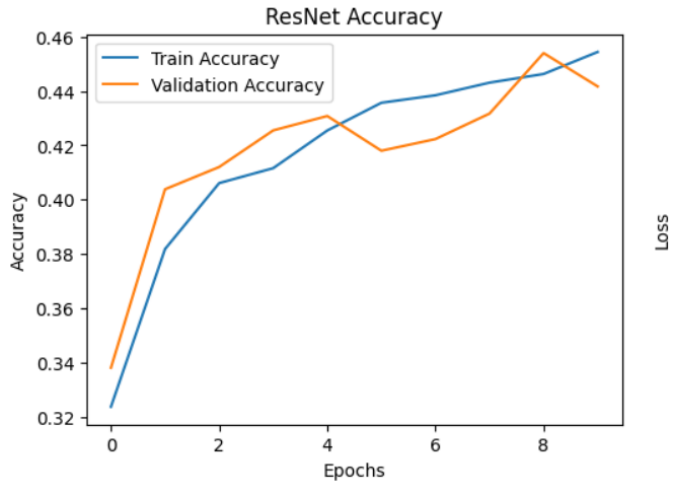

Fig. 5. Training Accuracy for ResNet

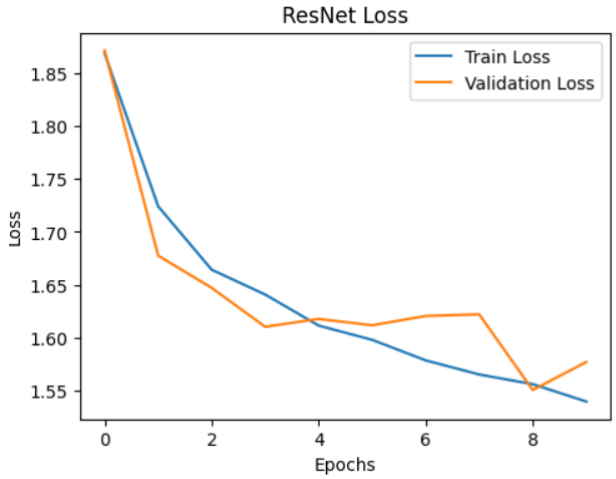

Fig. 6. Training loss for ResNet

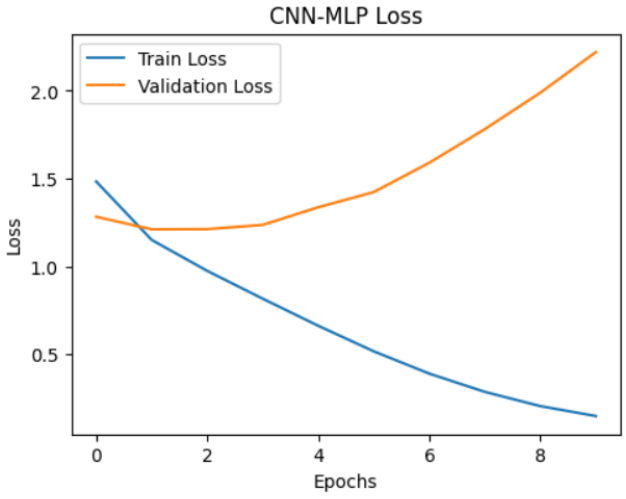

Fig. 4. Training loss for CNN-MLP

IV. Discussion

A. Model Analysis

- The **VIT** model achieved the highest accuracy with reasonable inference time, benefiting from the transformer architecture's attention mechanism. - The **CNN-MLP** model displayed moderate accuracy but had faster inference times due to fewer parameters. - The **ResNet** model, while known for robust performance, showed lower accuracy possibly due to the limited input size and frozen layers.

B. Challenges

Data augmentation and proper model tuning were essential. The pretrained ResNet model struggled with the 32x32 input, as the model is optimized for higher resolutions.

V. Conclusion

The Vision Transformer model proved most effective for CIFAR-10 with the highest accuracy and manageable inference time. Each model exhibited unique advantages, and the study demonstrates the potential trade-offs between accuracy and inference time across architectures.

VI. References

References

[1] Ashish Vaswani, et al., "Attention is All You Need," Advances in Neural Information Processing Systems, 2017.

[2] Kaiming He, et al., "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition, 2016.