Vision Transformer for CIFAR-10 Image Classification

Shamail Aamir khan
Department of Data Science
FAST University

***Abstract*—This report presents the implementation and evaluation of three deep learning architectures for CIFAR-10 image classification: a Vision Transformer (ViT), a hybrid Convolutional Neural Network (CNN) and Multi-Layer Perceptron (MLP) model, and a transfer learning-based ResNet-18 model.**

The ViT model is trained from scratch and achieves competitive results in terms of accuracy and F1-score, showcasing its potential for image classification tasks. The CNN + MLP hybrid architecture, which uses a CNN for feature extraction followed by an MLP for classification, demonstrates strong performance with an accuracy of 78.54For comparison, transfer learning using a pretrained ResNet-18 model was applied, fine-tuning it on the CIFAR-10 dataset. This approach achieves an accuracy of 82.81, outperforming both the ViT and CNN + MLP models. The models were evaluated using standard metrics such as accuracy, precision, recall, and F1-score, highlighting the strengths and trade-offs of each approach.

I. Introduction

Image classification is a core task in computer vision, and the CIFAR-10 dataset, consisting of 60,000 32x32 color images in 10 different classes, is widely used to benchmark image classification models due to its diversity and complexity. In this project, we explore multiple deep learning architectures, including a pretrained ResNet-18 model with transfer learning, a Vision Transformer (ViT), and a hybrid Convolutional Neural Network (CNN) + Multi-Layer Perceptron (MLP) architecture, to classify CIFAR-10 images.

While Convolutional Neural Networks (CNNs) have long been the go-to models for image classification tasks, the Vision Transformer (ViT) has emerged as a strong contender, applying transformer architecture traditionally used in natural language processing to image classification. The project demonstrates the application of both ViT and the CNN + MLP hybrid architecture on CIFAR-10, with results compared to the performance of the pretrained ResNet-18 model fine-tuned for CIFAR-10, leveraging learned features from large-scale datasets like ImageNet.

II. Methodology and Results

A. Dataset and Preprocessing

The CIFAR-10 dataset consists of 60,000 images, divided into 50,000 training images and 10,000 test images. Each image is 32x32 pixels and belongs to one of 10 classes. The images were resized to 224x224 pixels to match the input size of the pretrained ResNet-18 model. Data augmentation techniques, such as random crop, horizontal flip, and color jitter, were applied to the training set. The images were normalized to have pixel values in the range [-1, 1] using the mean and standard deviation of the CIFAR-10 dataset.

B. Model Architectures

1) Vision Transformer (ViT): The Vision Transformer model is based on transformer encoder blocks. The image is first divided into non-overlapping patches, which are linearly embedded into a fixed-dimensional space. These patch embeddings are supplemented with positional encodings and passed through multiple transformer layers. The output is passed through a fully connected layer to classify the image into one of the 10 CIFAR-10 classes.

ViT Model Hyperparameters:

- Patch size: 4
- Embedding dimension: 128
- Number of transformer layers: 6
- Number of attention heads: 4
- MLP dimension: 256
- Dropout rate: 0.1
- Learning rate: 3e-4
- Batch size: 128
- Epochs: 5

2) CNN + MLP Hybrid Architecture: The hybrid model consists of two parts: a CNN for feature extraction and an MLP for classification. The CNN includes three convolutional layers followed by max-pooling, batch normalization, and ReLU activations. The output of the CNN is passed through an adaptive average pooling layer, flattened, and fed into the MLP. The MLP contains two fully connected layers with ReLU activations and dropout regularization.

3) ResNet-18 with Transfer Learning: The ResNet-18 model is used for transfer learning, with its final fully connected layer modified to output predictions for the 10 classes of CIFAR-10. The model is pretrained on ImageNet, and only the final layer is fine-tuned.

C. Training Process

Both the ViT and CNN + MLP models were trained using the Adam optimizer with a learning rate of 0.001 for the CNN + MLP model and 3e-4 for the ViT model. Cross-entropy loss was used for classification. The ResNet-18 model was trained for 15 epochs using the same optimizer settings as the CNN + MLP model,but with a learning rate of 1×103 and weight decay for regularization.

D. Results

1) Vision Transformer (ViT): The ViT model was trained for 5 epochs and achieved the following evaluation metrics on the CIFAR-10 test set:

- Accuracy: 0.5181
- Precision: 0.5235
- Recall: 0.5181
- F1-Score: 0.5166

2) CNN + MLP Hybrid Architecture: The CNN + MLP model was trained for 10 epochs, and the evaluation metrics on the test set are as follows:

- Accuracy: 78.54
- Precision: 78.78
- Recall: 78.55
- F1-Score: 78.04

3) ResNet model: The ResNet model was trained for 15 epochs, and the evaluation metrics on the test set are as follows:

- Accuracy: 0.8281
- Precision: 0.8513
- Recall: 0.8500
- F1-Score: 0.825

4) ResNet-18 Transfer Learning Results: The ResNet-18 model achieved an accuracy of 82.81 on the CIFAR-10 validation set after 5 epochs of training.

5) Classification Reports: **ViT Classification Report:**

TABLE I: Classification Report for CIFAR-10 using ViT

Class	Precision	Recall	F1-Score
Airplane	0.62	0.54	0.58
Automobile	0.62	0.55	0.59
Bird	0.34	0.45	0.39
Cat	0.40	0.39	0.39
Deer	0.48	0.43	0.46
Dog	0.50	0.34	0.40
Frog	0.59	0.51	0.55
Horse	0.54	0.58	0.56
Ship	0.64	0.72	0.68
Truck	0.51	0.65	0.57
Accuracy	0.5181		

CNN + MLP Classification Report:

TABLE II: Classification Report for CIFAR-10 using CNN + MLP

Class	Precision	Recall	F1-Score
Airplane	0.83	0.84	0.83
Automobile	0.84	0.92	0.88
Bird	0.82	0.62	0.70
Cat	0.70	0.53	0.60
Deer	0.73	0.76	0.75
Dog	0.78	0.62	0.69
Frog	0.72	0.93	0.81
Horse	0.71	0.90	0.79
Ship	0.90	0.86	0.88
Truck	0.86	0.87	0.87
Accuracy	78.54		

ResNet-18 Classification Report:

TABLE III: Classification Report for CIFAR-10 using ResNet-18

Class	Precision	Recall	F1-Score
Airplane	1.00	0.75	0.86
Automobile	0.88	1.00	0.93
Bird	0.75	0.75	0.75
Cat	0.56	1.00	0.71
Deer	1.00	0.60	0.75
Dog	0.88	1.00	0.93
Frog	1.00	0.60	0.75
Horse	0.60	1.00	0.75
Ship	1.00	0.80	0.89
Truck	0.86	1.00	0.92
Accuracy	82.81		

III. Discussion

All models performed reasonably well on the CIFAR-10 dataset. The Vision Transformer (ViT) achieved a modest accuracy of 51.81, while the hybrid CNN + MLP architecture achieved an accuracy of 78.54. The ViT model faced challenges in distinguishing between certain visually similar classes, while the CNN + MLP model showed robustness across most categories. Both models demonstrated strengths and weaknesses in specific classes, indicating that further improvements such as hyperparameter tuning and data augmentation could enhance their performance.

The ResNet-18 model, trained using transfer learning, outperformed the other models, achieving an accuracy of 82.81. This demonstrates the effectiveness of transfer learning, which allowed the model to leverage pretrained features from ImageNet, resulting in faster convergence and stronger performance on CIFAR-10. Fine-tuning the last residual block further improved class-specific precision and recall, especially for vehicle categories. However, some confusion remained between visually similar classes, like cats and dogs. The training was also affected by challenges such as data imbalance and finding an optimal learning rate schedule.

IV. Conclusion

In this report, we successfully implemented three image classification models: the Vision Transformer (ViT), a hybrid CNN + MLP architecture, and a pretrained ResNet-18 model using transfer learning. Among the three, the ResNet-18 model achieved the highest accuracy of 82.81, showcasing the strong potential of transfer learning for classification tasks on small datasets like CIFAR-10. The ViT model achieved 51.81 accuracy, while the CNN + MLP hybrid architecture achieved 78.54.

Future work could focus on improving the ViT model's performance through hyperparameter tuning and data augmentation. Additionally, further fine-tuning and exploring more advanced models like ResNet-34 or ResNet-50 could lead to better results.

V. Prompts

The following prompts were used for this study:

- Fine-tuning a pretrained ResNet model for CIFAR-10 classification.
- Training the model on the CIFAR-10 dataset with transfer learning.
- Visualizing model performance using metrics like confusion matrix, accuracy, and F1-score.
- Implementing Vision Transformer (ViT) and hybrid CNN + MLP models for comparison.

VI. References

References

[1] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical Report, University of Toronto.

[3] Pan, S. J., Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.

[4] Dosovitskiy, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Machine Learning (ICML)*.