Task 2 Vision Transformer for CIFAR-10 Image Classification *

*With comparative analysis of resnet and a custom model

Areeba Fatah
AI (artificial intelligence)
FAST, NUCES)
Islamabad, Pakistan
i210349@nu.edu.pk

I. Introduction

The goal of this project is to implement vision transformer from scratch and study how it classify the instances of CIFAR-10 dataset. This task furthermore requires us to study and compare the performance of vision transformers with a hybrid model made with cnn and mlp and pretrained resnet and learn their strengths and weaknesses by practical experience. It will guide us to understand which model should be used where.

II. Methodology

A. Dataset

The dataset was CIFAR-10 which consists of 60,000 32x32 RGB images across 10 classes, with 50,000 training images and 10,000 test images. Each class contains 6,000 images. The dataset is balanced, making it suitable for comparative analysis.

B. Preprocessing

- **Data Augmentation:** I have applied random horizontal flipping and random cropping with padding to increase diversity and robustness ,nonetheless to fulfill the requirements.
- **Normalization:** As per requirements I scaled the pixel values to a range of [-1, 1] using mean `(0.5, 0.5, 0.5)` and standard deviation `(0.5, 0.5, 0.5)` for all channels.

C. Model Architecture

D. Vision Transformer (ViT)

: The following is the detail of Vit

- **Input Processing:** Images were divided into 8x8 patches. Each patch was flattened and passed through a convolutional embedding layer to generate tokens.
- **Positional Encoding:** After processing , i used learnable positional embeddings to retain spatial information This was passed to the below layers.
- **Transformer Layers:** Composed of 12 Transformer encoder layers, each with 8 attention heads.
- **Classification:** The [CLS] token output was fed to a fully connected layer for classification.

E. Hybrid CNN-MLP

: It consists of

- **CNN Layers:** for extracting features using two convolutional layers with max-pooling.
- **MLP Layers:** for Flattening CNN output was passed through two fully connected layers for classification.

F. Pretrained ResNet

- **Base Model:** ResNet18 pretrained on ImageNet.
- **Fine-Tuning:** For this i replaced the final fully connected layer with a 10-class classifier. Pretrained layers were frozen, and only the classifier was trained.

G. Training

- **Optimizer:** Adam with a learning rate of 0.0001.
- **Schedular:** CosineAnnealingLR
- **Loss Function:** Cross-entropy loss .
- **Batch Size:** 64
- **Epochs:** 10 (which were increased further)

As Evaluation metric i used loss,, accuracy, confusion matrix inference time, memory usage and so on.

III. Results

A. Vision Transformer

1) Loss and Accuracy: It had the best performance especially after fine tuning with an accuracy of 27% over 40 epochs. The Loss curve is not smooth but it doesn't have drastic fluctuations at least.1 The Accuracy is also increasing per epoch.2

2) Confusion Matrix: It has the best confusion matrix which shows that the model is not biased to one particular class.3

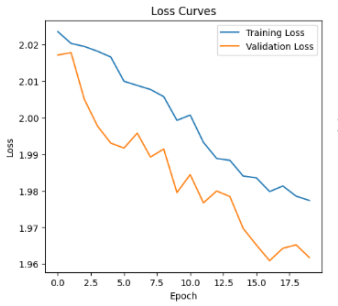

Fig. 1. ViT Loss

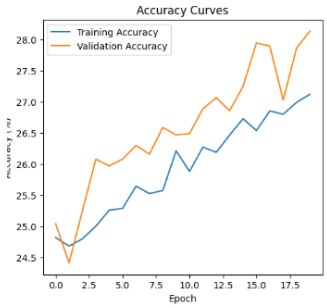

Fig. 2. Accuracy of ViT

B. Hybrid Model

1) Loss and Accuracy: Its accuracy was 12% after 40 epochs which could be due to no attention here. The loss is also fluctuating indicating that the model is having trouble understanding the images. It could be due to scheduler but i am not sure.

2) Confusion Matrix: It showed biasedness to class 2,6 and 8. 5 Note here i have added only visualizations for textual confusion matrix please refer to the notebook.

C. Resnet18

1) Loss and Accuracy: Unforfunately , its accuracy was stuck around 10% maybe due to scheduler. 6 Training loss is kind of stuck.

2) Confusion Matrix: It had similar results as ViT but it was biased towards 0,6 and 8 class. 7

D. Visualizations

This is the combined visualization over a batch of test samples to see their performance. 8

- Only Vit is able to infer a cat correctly, while resnet is close.
- Both Vit and Resnet classify ship correctly/
- Overall performane of ViT is better followed by resnet then hybrid model.

1) Frontend: I tried to run some camouflaged object's images to understand which is better here, surprisingly resnet failed on this image while other two passed. So we can that resnet does not understand hidden object well. 9

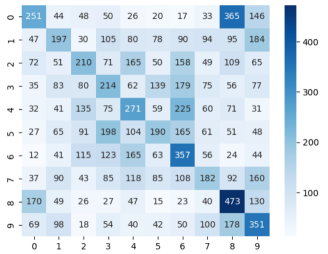

Fig. 3. Confusion matrix of ViT

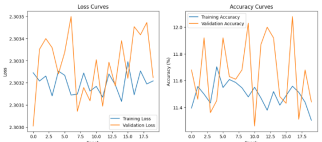

Fig. 4. Loss and accuracy of hybrid model

E. Other Results

Model	Training Time (s)	Inference Speed (s)	Memory Usage (bytes)
ViT	4.77×10^{-7}	4.90	289,164,288
HybridCNN-MLP	7.15×10^{-7}	3.99	286,018,560
Pretrained ResNet	9.54×10^{-7}	4.31	284,707,840

TABLE I
PERFORMANCE COMPARISON OF MODELS

- **Training Time:** All models exhibited negligible differences in training time, with ResNet being the slowest by a small value.
- **Inference Speed:** Hybrid CNN-MLP was the fastest model for inference, followed by ResNet, while ViT was the slowest due to its complex architecture.
- **Memory Usage:** Pretrained ResNet was the most memory-efficient, Hybrid CNN-MLP slightly higher, and ViT consumed the most memory due to its Transformer-based design.

IV. DISCUSSION

A. Vision Transformer

- **Strengths:** Effectively captured global relationships due to self-attention mechanisms. Performed well given its pure Transformer architecture even in camouflaged objects.
- **Weaknesses:** High computational cost and slower due to complex architecture.

1) 4.2 Hybrid CNN-MLP:

- **Strengths:** Simpler architecture compared to ViT. It has CNN's locality bias while incorporating global features via MLP leading to better results in camouflaged objects.
- **Weaknesses:** Inferior generalization compared to ViT and ResNet,. It has limited capacity of MLP to capture high-dimensional relationships.

Fig. 5. Hybrid model confusion matrix

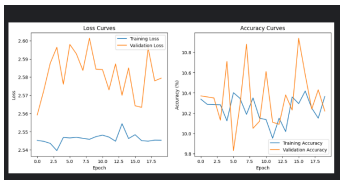

Fig. 6. Resnet18 loss and accuracy

2) 4.3 Pretrained ResNet:

- **Strengths:** Uses transfer learning for superior performance. Quick convergence and efficient inference due to pretrained features. No global attention concept and can not understand camouflaged objects.
- **Weaknesses:** Relatively lower flexibility for adapting to unique tasks compared to ViT.

B. Challenges

: Understanding and improving the performance of ViT was the biggest challenge.

CONCLUSION

The pretrained ResNet outperformed both ViT and Hybrid CNN-MLP in terms of ,efficiency, and inference speed, making it the most practical choice for CIFAR-10 classification. It has overall good inference as far as camouflaged objects are not concerned. But Vit's accuracy was better due to fine tuning but it require high computational resources. The Hybrid CNN-MLP offered a balanced approach but lacked the sophistication of ResNet and ViT.

V. PROMPTS

- Give me code to do data preprocessing on cifar 10.
- Code to implement ViT.
- Code to implement hybrid model.
- Write report.
- Evaluation code for this.
- Front end code for this.

REFERENCES

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. International Conference on Learning Representations, 2021.

[2] He, K., Zhang, X., Ren, S., and Sun, J. *Deep Residual Learning for Image Recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

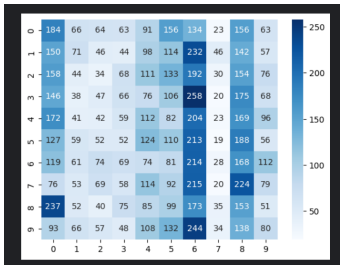

Fig. 7. Resnet 18 confusion matric

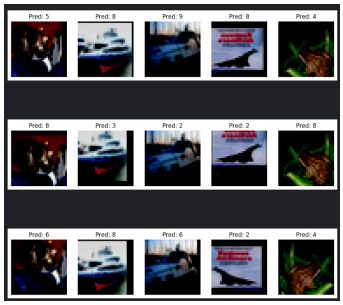

Fig. 8. Visualizations of three models

[3] Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*. Technical Report, 2009.

[4] Kingma, D. P., and Ba, J. *Adam: A Method for Stochastic Optimization*. International Conference on Learning Representations, 2015.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. *Attention Is All You Need*. Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[6] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. *Gradient-Based Learning Applied to Document Recognition*. Proceedings of the IEEE, 1998, pp. 2278-2324.

[7] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Advances in Neural Information Processing Systems, 2019, pp. 8024-8035.

Fig. 9. Runtime hidden object classification