Comparison of Vision Transformer (ViT) and Pre-Trained Models for Image Classification

Introduction

The study evaluates and compares the performance of Vision Transformer (ViT) models and pre-trained Convolutional Neural Networks (CNNs) on the CIFAR-10 dataset. This document presents the implementation details, performance metrics, and insights into the models' generalization capabilities. The goal is to analyze the effectiveness of transformers compared to CNNs in the domain of image classification.

Implementation Details

Model 1: Vision Transformer (ViT) Without Pre-training

Architecture: Custom-built Vision Transformer with the following hyperparameters:

- Number of Epochs: 20
- Batch Size: 256
- Patch Size: 16
- Embedding Dimension: 64
- Number of Attention Heads: 4
- Number of Layers: 4
- Optimizer: Adam
- Learning Rate: 0.001

Results:

- Final Loss: 1.0162
- Test Accuracy: **62.17%**

Model 2: Pre-Trained Vision Transformer (ViT)

Architecture: Pre-trained Vision Transformer ('google/vit-base-patch16-224-in21k'). Fine-tuned on CIFAR-10 with a custom classifier.

- Training Epochs: 3
- Optimizer: Adam
- Learning Rate: 0.001

Results:

- Final Loss: 0.1324
- Test Accuracy: **96.10%**

Model 3: Pre-Trained CNN (ResNet18)

Architecture: ResNet18 pre-trained on ImageNet with a modified final layer to classify CIFAR-10 data.

- Training Epochs: 3
- Optimizer: Adam
- Learning Rate: 0.001

Results:

- Final Loss: 0.0993
- Test Accuracy: **89.34%**

Key Insights

1. **Generalization Capability:**
    - The pre-trained ViT achieved the highest test accuracy of **96.10%**, outperforming both the custom ViT (**62.17%**) and ResNet18 (**89.34%**).
    - This highlights the ability of pre-trained transformers to generalize better, even when trained for a shorter duration.
2. **Performance Comparison:**
    - ResNet18 performed well, achieving **89.34%** accuracy, but fell short of the pre-trained ViT's performance.
    - The custom ViT's lower performance (**62.17%**) indicates the importance of pre-training for transformer models.
3. **Training Time:**
    - The pre-trained models required fewer epochs (3) to converge compared to the custom ViT (20 epochs), emphasizing the efficiency of transfer learning.
4. **Generalization vs Overfitting:**
    - The ResNet18 model showed strong training performance but slightly lower test accuracy, hinting at possible overfitting.
    - The pre-trained ViT's ability to achieve high test accuracy reflects its robustness across unseen data.

Visualization

The bar chart below compares the test accuracy of the three models:

- **ViT (Custom):** 62.17%
- **Pre-trained ViT:** 96.10%
- **ResNet18:** 89.34%

Conclusion

The results demonstrate the superior performance of the pre-trained Vision Transformer for image classification on CIFAR-10. Pre-trained models not only save training time but also exhibit better generalization capabilities compared to both custom transformers and CNNs. This underscores the importance of leveraging pre-trained models, especially for tasks with limited computational resources or datasets.