Report: Image Classification using Vision Transformers and Transfer Learning

1. Introduction:

Image classification is a fundamental task in computer vision, aiming to assign a label to an input image from a predefined set of categories. This report explores the use of Vision Transformers (ViTs) and transfer learning techniques in the context of image classification. ViTs, a relatively recent innovation, have shown remarkable performance in image classification tasks, and transfer learning further enhances their capabilities by leveraging pre-trained models on large datasets.

2. Vision Transformers (ViTs):

Vision Transformers are a class of deep neural network architectures that depart from the traditional Convolutional Neural Networks (CNNs) commonly used for image analysis. ViTs divide the input image into fixed-size non-overlapping patches and linearly embed these patches into sequences. The model then employs transformer architecture, originally designed for natural language processing, to capture global and local dependencies within the image.

3. Transfer Learning:

Transfer learning involves using a pre-trained model on a large dataset as a starting point for a specific task, such as image classification. It has become a cornerstone in modern deep learning, allowing models to generalize better with limited labeled data. In the context of ViTs, transfer learning involves fine-tuning pre-trained ViT models on the target image classification task.

4. Benefits of Using Vision Transformers and Transfer Learning:

Attention Mechanism: Vision Transformers capture both local and global features of an image through self-attention mechanisms. This enables them to understand relationships between different parts of the image effectively.

Scalability: ViTs can handle images of various sizes without requiring modifications to the model architecture. This scalability is advantageous when dealing with diverse datasets.

Transfer Learning Efficiency: Transfer learning with pre-trained ViT models helps in initializing the network with well-learned features from a massive dataset, leading to faster convergence and improved performance with a smaller dataset.

Reduced Overfitting: Pre-trained models have already learned generic features, which can help in reducing overfitting on smaller datasets.

5. Workflow:

The process of image classification using ViTs and transfer learning typically involves the following steps:

Data Preparation: Collect and preprocess the dataset, including resizing images, data augmentation, and splitting into training, validation, and test sets.

Pre-trained Model Selection: Choose a suitable pre-trained Vision Transformer model based on architecture and performance on related tasks.

Transfer Learning: Initialize the selected ViT model with pre-trained weights and fine-tune it on the target image classification task using the training set.

Hyperparameter Tuning: Adjust hyperparameters such as learning rate, batch size, and optimizer settings for optimal convergence.

Validation and Testing: Evaluate the fine-tuned model on the validation set to select the best model and fine-tuning strategy. Finally, assess the model's performance on the test set for a fair evaluation.

6. Challenges and Considerations:

Computational Resources: Training large Vision Transformer models can be resource-intensive, requiring access to powerful GPUs or TPUs.

Data Augmentation: Data augmentation techniques are crucial to enhance model generalization and mitigate overfitting, but finding the right augmentation strategies can be a trial-and-error process.

Domain Adaptation: Pre-trained models might not be directly applicable to certain domains. Domain adaptation techniques may be required to improve performance on domain-specific tasks.

7. Results:

The model is trained on a dataset of 2 types of flowers:

From scratch: The model is on an average 61% confident on prediction it made.

Using transfer learning: The model is on an average 96% confident on its prediction

8. Conclusion:

Image classification using Vision Transformers and transfer learning is a potent combination that brings the advantages of attention mechanisms and pre-trained representations to the field of computer vision. By leveraging large pre-trained models and transfer learning techniques, practitioners can achieve robust and accurate image classification models, even with limited labeled data. As the field continues to evolve, the integration of ViTs and transfer learning holds promise for addressing various challenges and pushing the boundaries of image classification performance.