Enhancing Vision Transformers for Image Classification: A Study on Infused Adapters, Inhibition, and Amplification of Internal Mechanisms

Sai Krishna Sangeetha
sa447901@ucf.edu

Sricharan Maddena
sr537872@ucf.edu

Abstract

In the field of computer vision, significant advancements have been made by adapting large pre-trained models, such as Vision Transformers, to specific tasks through fine-tuning. Typically, fine-tuning involves either updating the entire model's parameters or using linear probes. This report is dedicated to exploring and evaluating fine-tuning techniques that are more efficient in terms of parameter usage, specifically for Vision Transformers applied to image classification tasks. We approached the concept of efficient fine-tuning as an issue of subspace training and conduct a thorough comparison of various efficient fine-tuning techniques. My empirical analysis focuses on the balance between each method's performance and its parameter efficiency. In particular, we compare the performance of the Infused Adapter by Inhibiting and Amplifying Inner Activations (IA^3) [1] against standard linear probe techniques and complete model fine-tuning [2] in both a standard dataset fine-tuning scenario and a few-shot learning context [3].

1. INTRODUCTION

In recent years, there has been a marked increase in interest in large-scale vision and language models pre-trained on extensive datasets, demonstrating notable performance improvements. Concurrently, with advancements in computational hardware, the size of these models is expanding rapidly. For instance, Vision Transformer (ViT) models like ViT-Large [4] possess billions of parameters and continue to grow. It is anticipated that we will soon see pre-trained vision models that are even larger. While these models excel when applied to downstream vision tasks, the storage and deployment costs of multiple fine-tuned model instances can be considerable, potentially limiting the practical use of large-scale Vision Transformers.

Driven by the necessity for parameter-efficient learning, this study focuses on investigating parameter-efficient fine-tuning methods for vision transformers. Traditionally in the computer vision community, transfer learning has involved fine-tuning all parameters or utilizing linear probes. Yet, fully fine-tuning pre-trained Vision Transformers is becoming less viable due to its financial and environmental impact and the high computational demands, which only escalate as models grow in size. An alternative, linear probing, involves adding a trainable multi-layer perceptron (MLP) at the model's end, which is more parameter-efficient but typically yields less optimal performance.

The goal is to develop fine-tuning strategies that strike an optimal balance between efficiency and effectiveness. This involves enhancing the parameter efficiency of fine-tuning processes while preserving the model's ability to effectively transfer learning to downstream vision tasks, particularly image classification.

Model Name	Total Parameters	Trainable Parameters	% of weights
ViT (Full Model Finetuning)	88302197	88302197	100
ViT (Linear Probe)	88302197	5005	0.00005668
ViT (IA^3)	88340597	43405	0.00049134

Fig. 1: Comparison of techniques based on the number of trainable parameters in a few-shot learning setup.

To assess the Infused Adapter by Inhibiting and Amplifying Inner Activations (IA^3) technique, we have conducted a comparison of its effectiveness against both full parameter training and training just the task-specific multi-layer perceptron (MLP) layer in both standard fine-tuning and few-shot learning scenarios. Notably, the IA^3 technique has already been established as the leading method for few-shot learning when applied to a large-scale language model (T5). This report contributes in the following ways:

1) It provides an evaluation of the IA^3 technique applied to a pre-trained ImageNet Vision Transformer (ViT) in a fine-tuning context.

2) It offers an assessment of the IA^3 technique when applied to a pre-trained ImageNet ViT in a few-shot learning context.

We have not only implemented the IA^3 technique but also conducted evaluations for image classification tasks within both fine-tuning and few-shot learning frameworks.

2. MATERIAL AND METHODS

2.1 Dataset Description

To appraise the IA^3 technique, we executed fine-tuning and few-shot learning trials using a Vision Transformer (ViT) model pre-trained on the ImageNet dataset, specifically targeting the Caltech-UCSD Birds dataset (CUB) [5]. This dataset comprises 11,788 images spanning 200 bird species classes.

2.2 Fine-Tuning Configuration

For the fine-tuning configuration, a Vision Transformer (ViT) model, initially set with weights from ImageNet, undergoes fine-tuning specifically for the Caltech-UCSD Birds (CUB) dataset. The effectiveness of this fine-tuning is determined by the accuracy on the CUB dataset's test set. As previously stated, the IA^3 technique's performance is benchmarked against both the comprehensive fine-tuning of the entire model and the fine-tuning of only the linear probe layers.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Fig. 2: Accuracy formula

Fig. 3: Fine-tuning flow

2.3 Few-Shot Learning Configuration

In the context of this report, the few-shot learning setup involves partitioning the dataset, denoted as D (which refers to the CUB dataset), into three separate subsets: Dtr for training, Dval for validation, and Dtest for testing. The validation subset, Dval, is utilized for model selection purposes, while the test subset, Dtest, is reserved for the final assessment of the model's performance.

This setup is designed in alignment with the few-shot classification frameworks established by seminal works of Vinyals et al. (2016), Sung et al. (2018), and Snell, Swersky, and Zemel (2017), which propose episodic training. In this methodology, training occurs across various tasks, each derived from a distinct probability distribution, p(T). For any given task Ti, a selection of K samples is randomly chosen from N distinct classes, forming an (N-way, K-shot) classification task.

Specifically, each task is comprised of a support set S, which includes K examples per class, and a query set Q, with Q examples per class. These sets are designed to be exclusive; the NQ query samples and NK support samples do not overlap, ensuring that the model's generalization capabilities can be accurately gauged. The structure and process of the few-shot learning setup are depicted in figure 3.

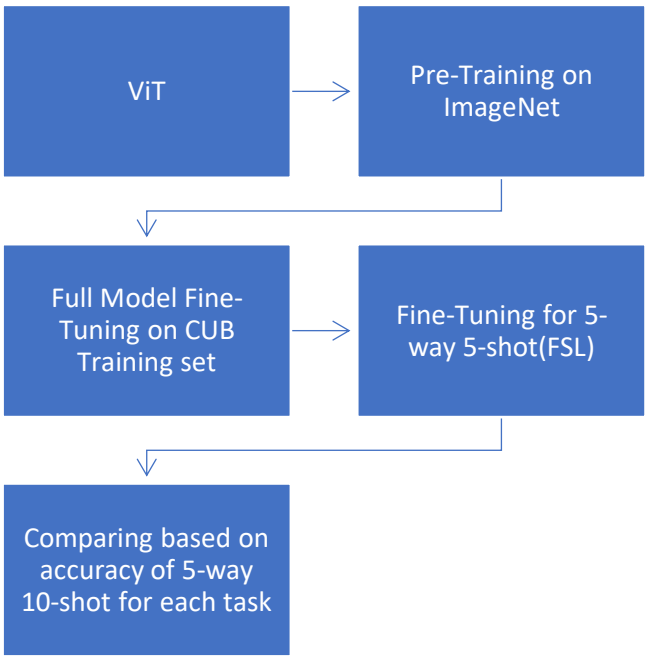

Fig. 4: Few-Shot Learning flow in Computer Vision

2.4 Vision Transformer (ViT)

A Vision Transformer (ViT) is a specialized deep learning framework designed for computer vision applications, such as image classification, object detection, and image segmentation. It is adapted from the transformer model, originally developed for natural language processing, where it has set new benchmarks.

The architecture of a ViT is fundamentally akin to transformer models used in NLP, but with adaptations tailored to image data. One significant modification is the employment of a two-dimensional self-attention mechanism, which interprets images as grids of pixels instead of sequences of text.

ViTs may also integrate unique components catered to image processing, such as spatial transformer layers that enable the model to discern the spatial relationships among objects in an image. Another example is the multi-scale transformer layer, which equips the model to handle images at various resolutions and scales.

In essence, the ViT architecture leverages the strengths of the transformer model, aligning it with the demands of computer vision, and has demonstrated considerable potential across numerous computer vision tasks.

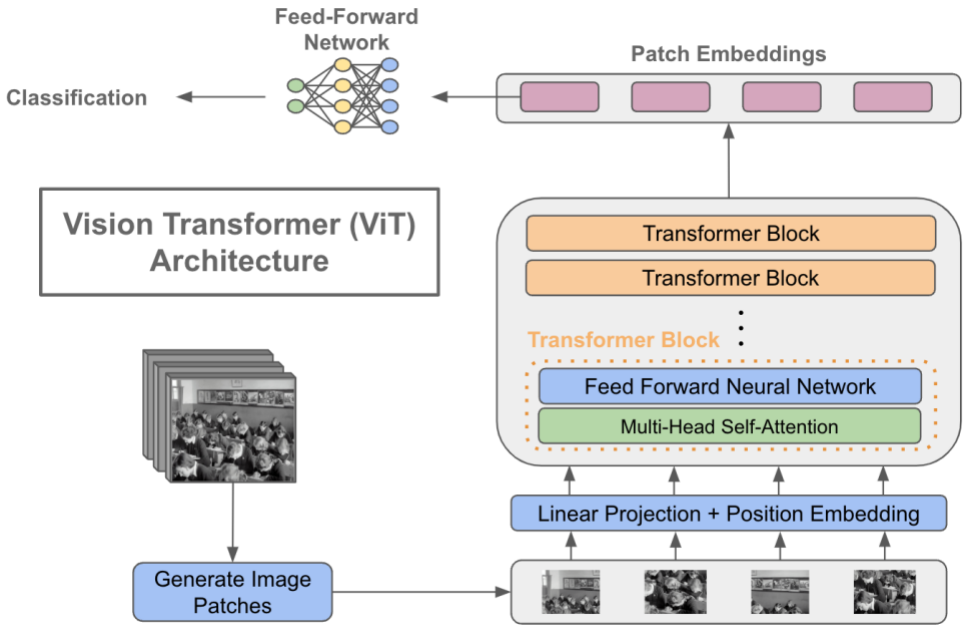

Fig. 5: Vision Transformer Architecture

2.5 Parameter Efficient Fine-tuning

Optimizing a Vision Transformer for a new dataset can be resource-intensive because it requires recalibrating the model's weights to align with the new information. Employing parameter-efficient tactics is one strategy to streamline the fine-tuning procedure. These methods focus on minimizing the model's parameter count without sacrificing performance. By utilizing parameter-efficient approaches, fine-tuning becomes more resource-efficient, enabling the adaptation of a pre-trained model to new tasks with reduced computational demands.

2.6 Infused Adapter by Inhibiting and Amplifying Inner Activations (IA^3)

The IA^3 method is a performance enhancement technique for transformers, a class of deep learning models widely employed in diverse domains like natural language processing (NLP) and computer vision. IA^3 works by modulating the internal activity of the model's layers, suppressing or boosting the activations where needed to infuse additional information into the model's processing.

Originally devised for NLP tasks, the IA^3 approach also holds potential for application in vision transformers. Within the vision domain, IA^3 could refine the model's processing of visual data, which may enhance its capabilities in tasks such as image classification or object recognition.

The practical application of IA^3 in a Vision Transformer involves the insertion of tunable vectors that modify the key and value vectors across each attention head within the model. Similarly, adaptable vectors are introduced to each of the positional feed-forward layers. The specific mechanics of this integration are detailed in figures 4 and 5, which elucidate the architectural adjustments within the model that facilitate the IA^3 technique.

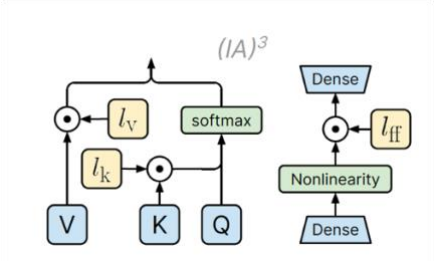

Fig. 6: Adapter layer architecture

$$\text{softmax}\left(\frac{Q(l_{\text{k}} \odot K^T)}{\sqrt{d_k}}\right)(l_{\text{v}} \odot V)$$

Fig. 7: Key and Value computation

3. RESULTS

The IA^3 technique is evaluated in fine-tuning and few-shot learning setup and the results are as follows:

A. Fine-Tuning Setup

For the fine-tuning process, the model starts with ImageNet pre-trained weights and is then adapted for the Caltech-UCSD Birds (CUB) dataset. Considering that the CUB dataset encompasses 200 distinct classes, a multi-layer perceptron (MLP) and a SoftMax layer, each with 200 neurons, are appended to the tail end of the pre-trained Vision Transformer (ViT) to calculate the probability for each class.

As previously discussed, the IA^3 method's performance is benchmarked against both full model fine-tuning and linear probe training. The specific details of the trainable parameters for each approach are outlined in figure 8. According to the data presented in figure 8, the accuracy levels of all models are relatively comparable. Specifically, full model fine-tuning exhibits a marginal advantage, scoring 0.01 higher than the other techniques. Given the minimal difference in performance, opting for either a linear probe or the IA^3 method could be more beneficial in terms of memory efficiency.

Model Name	Total Parameters	Trainable Parameters	% of weights
ViT (Full Model Finetuning)	88437332	88437332	0.86
ViT (Linear Probe)	88437332	140140	0.85
ViT (IA^3)	88465732	178540	0.85

Fig. 8: Number of trainable parameters and accuracy achieved in fine-tuning setup.

Train Batch Size	16
Learning Rate	1e-4
Loss Function	Categorical Cross Entropy
Epochs	100
Optimizer	Adam

Fig. 9: Hyperparameters for fine-tuning

B. Few-Shot Learning (FSL) Setup

The IA3 technique is assessed in a 5-way, 5-shot Few-Shot Learning (FSL) scenario. This entails fine-tuning the model on a set of five classes, with each class represented by five samples. The effectiveness of the FSL is determined based on the accuracy achieved with 10 shots per class. This fine-tuning is conducted on a Vision Transformer (ViT) model pre-trained with 140 training classes, and the mean accuracy across 30 repetitions of fine-tuning in 5-way, 5-shot tasks is presented in figure 10.

For the FSL task, the model undergoes fine-tuning for 300 epochs, while maintaining the same hyperparameters as those outlined in figure 9. Based on the accuracy metrics shown in figure 10, it can be deduced that full model fine-tuning yields the most effective performance in these FSL tasks.

Model Name	Trainable Parameters	Accuracy
ViT (Full Model Finetuning)	88302197	0.935
ViT (Linear Probe)	5005	0.924
ViT (IA^3)	43405	0.9119

Fig. 10: Number of trainable parameters and accuracy achieved in few-shot learning (5-way 5-shot) setup.

4. CONCLUSIONS AND FUTURE WORK

Based on the outcomes illustrated in figures 8 and 10, it can be concluded that the IA^3 method may not be the most effective Parameter-Efficient Fine-Tuning (PEFT) strategy for the Vision Transformer (ViT) in image classification tasks. A possible explanation for this could be the differing scales and contexts in which IA^3 was originally proven effective versus its application here. IA^3 showed success with T5, a substantial language model, whereas the ViT model is relatively smaller in scale. Interestingly, the linear probe approach not only outperforms IA^3 in this context but also requires fewer trainable parameters. This suggests that while IA^3 has its merits, its applicability and efficiency may vary significantly across different model architectures and scales. This scenario highlights an area ripe for further exploration in developing more effective PEFT techniques tailored specifically for vision models. The quest to refine these methods is an ongoing challenge, presenting ample opportunities for innovative research and development.

MY CONTRIBUTION

Myself, Sai Krishna Sangeetha, worked on the literature survey and data pre-processing, and contributed in writing code for the proposed model. We did extensive literature survey to find some of the publicly available codes that can be used as a reference. Reused the adapted codes as an extension to the proposed model.

REFERENCES

[1] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C. (2022). Few-Shot Parameter-Efficient Fine- Tuning is Better and Cheaper than In-Context Learning. arXiv. https://doi.org/10.48550/arXiv.2205.05638

[2] He, X., Li, C., Zhang, P., Yang, J., Wang, X. E. (2022). Parameter-efficient Model Adaptation for Vision Transformers. arXiv. https://doi.org/10.48550/arXiv.2203.16329

[3] Singh, A. (2022). Transductive Decoupled Varia- tional Inference for Few-Shot Classification. arXiv. https://doi.org/10.48550/arXiv.2208.10559

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weis- senborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain

Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[5] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.