FC-LoRA: A Framework for Efficient Fine-Tuning Guided by Fisher Score and Covariance

Abstract

Fine-tuning large-scale foundation models, such as the Data-efficient Image Transformer (DeiT), has become a standard for achieving state-of-the-art performance on specialized tasks. However, updating all model parameters during this process is computationally prohibitive, posing a significant barrier to research and deployment. We introduce Fisher score and covariance LoRA (FC-LoRA), a parameter-efficient fine-tuning (PEFT) framework that intelligently allocates a limited budget of trainable parameters to the most critical components of the model. By creating a composite importance score from Fisher information, gradient magnitude, and activation covariance, FC-LoRA dynamically assigns LoRA ranks to attention layers. Our experiments on CIFAR-100 and CIFAR-10 show that by training as few as 0.23% of the model's parameters, FC-LoRA achieves over 98% of the accuracy of full fine-tuning while requiring less than half the training time and producing a significantly more reliable and well-calibrated final model.

1 Introduction

The advent of large, pre-trained Vision Transformers (ViTs) has revolutionized the field of computer vision. The dominant paradigm involves adapting these models to downstream tasks via fine-tuning. However, the "full fine-tuning" approach, which updates every parameter in the model, is exceptionally resource-intensive, demanding significant computational power, time, and energy. This cost limits the accessibility of state-of-the-art AI.

Parameter-Efficient Fine-Tuning (PEFT) methods have emerged as a powerful alternative. Among these, Low-Rank Adaptation (LoRA) is prominent. LoRA freezes the original model weights and injects small, trainable low-rank matrices into its layers. While effective, standard LoRA implementations typically use a uniform rank across all adapted layers. This one-size-fits-all approach is suboptimal, as it fails to account for the heterogeneous importance of different layers in adapting to a new task.

To address this limitation, we propose FC-LoRA, a framework that enhances LoRA by introducing a principled, data-driven method for allocating the parameter budget. Instead of a fixed rank, FC-LoRA assigns ranks dynamically based on a layer's measured importance, ensuring that computational resources are focused where they will have the greatest impact.

2 Methodology: The FC-LoRA Framework

The core principle of FC-LoRA is the intelligent allocation of a fixed parameter budget. This is achieved through a two-phase process: first, scoring the importance of each target layer, and second, distributing the available LoRA ranks according to those scores. For our experiments, we targeted the query, key, and value projection layers within the self-attention blocks of the DeiT model.

2.1 Importance Scoring

Before commencing full training, FC-LoRA performs a brief analysis using a small number of training batches to calculate a composite importance score for each layer. This score synthesizes three distinct, complementary metrics:

- **Fisher Score (Parameter Sensitivity)**: Approximated from the squared gradients, the Fisher score quantifies how sensitive the model's output is to changes in a layer's parameters. A high Fisher score indicates that a layer's parameters are highly influential and thus critical for the task.
- **Gradient Magnitude (Learning Momentum)**: The average magnitude of the gradients for a layer's parameters indicates how rapidly that layer is adapting during the initial stages of training. Layers with high-magnitude gradients are considered to be actively and significantly engaged in the learning process.
- **Activation Covariance (Feature Diversity)**: Using forward hooks, we measure the trace of the covariance matrix of a layer's output activations. A higher trace suggests that the layer is producing a more diverse and information-rich set of features, making it essential for capturing the nuances of the new dataset.

These three metrics are individually normalized and then combined into a single, robust importance score, with the Fisher score being the most heavily weighted component.

2.2 Adaptive Rank Allocation and Training

With importance scores established, FC-LoRA distributes a total, fixed budget of ranks (e.g., 100 total ranks) among the target layers. This is performed using a "water-filling" algorithm: layers with higher importance scores receive proportionally more ranks, subject to a predefined maximum rank per layer. This ensures that the most critical layers are given the greatest learning capacity.

Following rank allocation, the model is trained using a state-of-the-art regimen. Only the newly injected LoRA adapter matrices and the final classification head are trainable, freezing over 99% of the original model. To maximize performance and generalization, we employ CutMix data augmentation and a cosine annealing learning rate schedule.

Finally, to enhance the model's reliability, we perform temperature scaling. This post-hoc calibration step tunes a single temperature parameter on the validation set to adjust the model's output probabilities, ensuring its confidence scores accurately reflect its likelihood of being correct.

3 Experimental Results

We evaluated FC-LoRA against a full fine-tuning baseline on the CIFAR-100 and CIFAR-10 datasets. The results demonstrate a compelling case for our adaptive approach.

4 Analysis and Discussion

Unprecedented Efficiency: On the complex CIFAR-100 task, FC-LoRA achieved 98.9% of the baseline's accuracy while using only 42% of the training time and 0.23% of the

Table 1: CIFAR-100 Results (15 Epochs)

Metric	FC-LoRA (Our Method)	Full Fine-Tuning
Trainable Parameters	198,144 (0.23%)	86,075,236 (100%)
Test Accuracy	89.19%	90.18%
Test Loss	1.1284	1.1345
ECE / Scaled ECE	0.0920 / 0.0146	0.0560 / 0.0266
Class Accuracy $\mu \pm \sigma$	0.89 ± 0.08	0.90 ± 0.07
Training Time	$\sim$48 minutes	$\sim$114 minutes

Table 2: CIFAR-10 Results (10 Epochs)

Metric	FC-LoRA (Our Method)	Full Fine-Tuning
Trainable Parameters	125,952 (0.15%)	85,933,834 (100%)
Test Accuracy	98.27%	98.51%
ECE / Scaled ECE	0.0886 / 0.0045	0.0840 / 0.0051
Training Time	$\sim$32 minutes	$\sim$38 minutes

trainable parameters. This demonstrates a dramatic improvement in computational efficiency with a negligible impact on raw predictive performance.

Superior Reliability: A key advantage of FC-LoRA is its improved model calibration. As measured by Scaled Expected Calibration Error (ECE), where lower is better, the FC-LoRA model was significantly more reliable (0.0146) than the fully fine-tuned model (0.0266) on CIFAR-100. This means its confidence scores are a more accurate reflection of its true correctness, a critical feature for trustworthy AI systems.

Robust Performance: The framework's success on both CIFAR-100 and CIFAR-10 validates its robustness. Even on the simpler CIFAR-10 task, FC-LoRA achieves near-identical accuracy and better calibration in less time.

5 Conclusion

The FC-LoRA framework provides a robust and highly efficient method for adapting large-scale Vision Transformers. By moving beyond a uniform parameter allocation and instead using a data-driven importance score to guide the distribution of a parameter budget, FC-LoRA achieves performance nearly identical to full fine-tuning at a fraction of the computational cost. Furthermore, our method yields models that are better calibrated and more reliable. This work represents a significant step towards making the power of state-of-the-art AI more accessible, sustainable, and practical for a wider range of applications.