Technical Note: FC-LoRA, an Adaptive Rank Allocation Method for Parameter-Efficient Fine-Tuning of Vision Transformers

Subhasish Bandyopadhyay

May 2025

Abstract

Parameter-Efficient Fine-Tuning (PEFT) methods, particularly Low-Rank Adaptation (LoRA), have become standard for adapting large pre-trained models. This work introduces FC-LoRA, an adaptive rank allocation strategy that uses a composite importance metric—combining Fisher information, Covariance of activations, and gradient magnitude—to intelligently distribute a rank budget across a model's layers. We evaluate this method by fine-tuning a DeiT transformer on the CIFAR-10 and CIFAR-100 datasets and compare its performance directly against a traditional full fine-tuning baseline. The results are compelling. On CIFAR-10, FC-LoRA achieves 98.27% accuracy, reaching 99.75% of the performance of full fine-tuning (98.51%) while training only 0.15% of the parameters and completing 16% faster. On the more complex CIFAR-100 dataset, FC-LoRA achieves 89.19% accuracy—98.9% of the full fine-tuning performance (90.18%)—using just 0.23% of the parameters and training more than twice as fast. Furthermore, while the fully fine-tuned models exhibit slightly better raw calibration, the FC-LoRA models achieve superior calibration after temperature scaling. This demonstrates that an adaptive, parameter-efficient approach can offer a highly effective trade-off, delivering near-identical performance with substantial gains in computational and memory efficiency.

1 Introduction

The advent of the Transformer architecture has revolutionized not only Natural Language Processing but also the field of computer vision. Vision Transformers (ViTs) and their derivatives, such as the Data-efficient Image Transformer (DeiT), have set new benchmarks in image classification tasks. These models are typically pre-trained on massive datasets and then fine-tuned for specific downstream applications. However, the full fine-tuning of these multi-billion parameter models is computationally expensive, memory-intensive, and requires vast amounts of task-specific data.

To address these challenges, Parameter-Efficient Fine-Tuning (PEFT) techniques have emerged. PEFT methods aim to adapt pre-trained models by training only a small fraction of their parameters while keeping the vast majority of the original weights frozen. Among these techniques, Low-Rank Adaptation (LoRA) has gained significant popularity due to its effectiveness and simplicity. LoRA injects trainable, low-rank decomposition matrices into the layers of the Transformer and trains only these new matrices.

A crucial hyperparameter in LoRA is the rank r, which is conventionally set to a uniform value across all targeted layers. This work challenges that convention, hypothesizing that an adaptive rank allocation strategy can yield better results by assigning more capacity to more important layers. We introduce FC-LoRA, a method for calculating layer importance to guide this allocation. We then benchmark this approach against the ”gold standard” of full fine-tuning to rigorously evaluate its performance-to-efficiency trade-off on two distinct image classification tasks.

2 Methodology

The core of our methodology is a two-stage process: first, an importance-scoring phase to determine the significance of each target layer, and second, a fine-tuning phase using the adaptively configured LoRA model.

2.1 Baseline Model and Dataset

- **Model**: We use the facebook/deit-base-distilled-patch16-224 model. Its final classification head is replaced with a new torch.nn.Linear layer appropriate for the target dataset (10 classes for CIFAR-10, 100 for CIFAR-100).
- **Datasets**: CIFAR-10 and CIFAR-100. For both, we use an 80%/20% split of the official training data for our own training and validation sets, respectively. All images are resized to 224x224 and normalized.

2.2 Low-Rank Adaptation (LoRA)

LoRA operates by freezing the pre-trained weight matrix $W_0 \in R^{d\times k}$ and modifying the forward pass as:

$$h = W_0x + \Delta Wx = W_0x + BAx$$

where $A \in R^{r\times k}$ and $B \in R^{d\times r}$ are the trainable low-rank matrices, and the rank $r \ll \min(d, k)$. We target the query, key, and value projection matrices within the self-attention blocks of the DeiT encoder.

2.3 FC-LoRA: Adaptive Rank Allocation

Instead of using a fixed rank r, we devise a scheme to allocate ranks from a total budget, R_{total}, based on layer importance.

1. **Importance Metrics**: We perform an initial forward and backward pass on a small subset of the training data (100 batches) to collect statistics for three metrics:
    - **Fisher Information (Approximation)**: The empirical Fisher, approximated by the average squared gradients for each layer, indicates parameter sensitivity.
    - **Gradient Magnitude**: The average absolute gradient magnitude, $|\nabla L|$, reflects how much a parameter affects the loss.
    - **Activation Covariance Trace**: The trace of the activation covariance matrix captures the informational richness of a layer's output.
2. **Composite Score and Budget Allocation**: The scores for each metric are min-max normalized to a [0, 1] range. The final composite importance score, S, is a weighted sum:

    $$S = 0.6 \cdot S_{Fisher_{norm}} + 0.2 \cdot S_{Grad_{norm}} + 0.2 \cdot S_{Cov_{norm}}$$

    Layers are sorted by this score, and ranks are allocated proportionally from the total budget R_{total}, subject to a minimum rank of 1 and a maximum of 16.

3 Experimental Setup and Results

We conducted two primary experiments, one on CIFAR-10 and one on CIFAR-100. In each case, we compare the results of our FC-LoRA method against a full finetuning baseline where all model parameters are updated. All experiments use AdamW, a cosine annealing learning rate schedule, label smoothing, and CutMix augmentation.

3.1 Experiment 1: CIFAR-10

- **Task**: 10-class image classification.
- **Epochs**: 10

Table 1: Results Summary (CIFAR-10)

Metric	FC-LoRA	Full Finetune
Trainable params	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
Wall-clock	1,903s (32 min)	2,279s (38 min)

3.2 Experiment 2: CIFAR-100

- **Task**: 100-class image classification.
- **Epochs**: 15

Table 2: Results Summary (CIFAR-100)

Metric	FC-LoRA	Full Finetune
Trainable params	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 Acc $\mu \pm \sigma$	0.89 ± 0.08	0.90 ± 0.07
Wall-clock	2,862s (48 min)	6,821s (1h 54m)

4 Comparative Analysis and Discussion

The direct comparison between FC-LoRA and full fine-tuning provides critical insights into the practical value of the adaptive PEFT approach.

4.1 The Accuracy vs. Efficiency Trade-off

The results clearly demonstrate that FC-LoRA offers an exceptional balance between performance and computational cost.

- On CIFAR-10, the performance gap is negligible. FC-LoRA achieves 99.75% of the accuracy of a fully fine-tuned model (98.27 / 98.51) with a minuscule parameter count (0.15%) and a 16% reduction in training time. For many applications, this trade-off is overwhelmingly favorable.
- On the more challenging CIFAR-100 dataset, the trade-off remains highly compelling. The accuracy gap widens slightly, with FC-LoRA reaching 98.9% of the full fine-tuning performance (89.19 / 90.18). However, the efficiency gains become even more pronounced: the training is completed in less than half the time (a 2.38x speedup) while updating fewer than 1 in 400 of the model's parameters. This is particularly crucial for scenarios with limited GPU availability or tight iteration cycles.

4.2 Model Calibration Insights

A surprising and valuable finding relates to model calibration. Calibration measures whether a model's confidence scores align with its actual accuracy.

- Before scaling, the fully fine-tuned models show slightly better raw calibration (lower ECE). This is especially true on CIFAR-100 (0.0560 vs. 0.0920).

- After post-hoc temperature scaling, the trend reverses. The FC-LoRA models achieve superior final calibration on both datasets (e.g., 0.0146 vs. 0.0266 on CIFAR-100).

This suggests that the constrained, targeted updates of FC-LoRA, while slightly harming initial calibration, produce a logit distribution that is more amenable to simple post-hoc correction. The PEFT approach acts as a regularizer, potentially leading to a ”cleaner” confidence landscape that can be more easily scaled to reflect true probabilities.

4.3 The Value of Adaptive Allocation

The ability of FC-LoRA to compete so closely with full fine-tuning stems directly from its intelligent allocation of parameters. The importance metrics consistently prioritized the value projection matrices within the self-attention blocks—layers known to be critical for adapting a Transformer's learned representations. By focusing the limited parameter budget on these high-impact areas, FC-LoRA maximizes its adaptive power, avoiding wasteful updates to less critical layers.

5 Conclusion and Future Work

This work successfully demonstrates that FC-LoRA, an adaptive rank allocation method, is a powerful and practical alternative to full fine-tuning. By benchmarking against this gold standard, we have shown that our method is not merely a ”cheaper” option but a highly competitive one.

Key Takeaways:

1. **Near-Parity Performance**: FC-LoRA achieves over 98.9% of the accuracy of full fine-tuning on both simple and complex classification tasks.
2. **Massive Efficiency Gains**: This performance is achieved while training fewer than 0.25% of the total parameters and with training speedups exceeding 2.3x.
3. **Superior Calibrated Confidence**: After post-hoc temperature scaling, FC-LoRA models provide more reliable confidence scores than their fully fine-tuned counterparts.

For practitioners, this means that adapting large vision models can be done faster, with significantly lower memory and compute requirements, and without a meaningful sacrifice in accuracy.

Future Work could explore dynamic rank allocation that adjusts during training, apply the FC-LoRA methodology to other domains like Large Language Models, and further investigate the relationship between PEFT techniques and model calibration.