Article

Not peer-reviewed version

An Empirical Investigation into Fine-Tuning Methodologies: A Comparative Benchmark of LoRA for Vision Transformers

Satwik Sai Prakash *

Posted Date: 3 November 2025

doi: 10.20944/preprints202510.2514.v1

Keywords: vision transformer; LoRA; fine-tuning; CNN; ResNet; EfficientNet

Article

An Empirical Investigation into Fine-Tuning Methodologies: A Comparative Benchmark of LoRA for Vision Transformers

Satwik Sai Prakash

School of Electrical Sciences, Odisha University of Technology and Research, Ghatikia, Bhubaneswar 751029, India; sahoospsatwik@gmail.com

Abstract

In the field of modern computer vision, the standard practice for solving new problems is not to train a model from scratch but to adapt a massive, pre-trained model. This process, known as fine-tuning, is a cornerstone of deep learning. This paper investigates two fundamentally different philosophies for fine-tuning. The first is the traditional, widely used approach where the core of a pre-trained Convolutional Neural Network (CNN) is kept frozen, and only a small, new classification layer is trained. This method is fast and computationally cheap. The second is a more modern, parameter-efficient fine-tuning (PEFT) technique called Low-Rank Adaptation (LoRA), which allows for a more profound adaptation of a model's internal workings without the immense cost of full retraining. To test these competing methods, we designed a benchmark using three powerful, pre-trained models. For the traditional approach, we used ResNet50 and EfficientNet-B0, two highly influential CNNs. For the modern approach, we used a Vision Transformer (ViT), an architecture that processes images using a self-attention mechanism, and adapted it with LoRA. We then evaluated these models on three datasets of increasing complexity: the simple MNIST (handwritten digits), the moderately complex Fashion-MNIST (clothing items), and the significantly more challenging CIFAR-10 (color photos of objects like cars, dogs, and ships). This ladder of complexity was designed to reveal under which conditions each fine-tuning strategy excels or fails. The outcomes were striking on the intricate CIFAR-10 dataset: the ViT-LoRA model performed exceptionally well, achieving 97.3% validation accuracy, while the ResNet50 and EfficientNet-B0 models only managed 83.4% and 80.5%, respectively. The task was not difficult enough to distinguish between the models on the much easier MNIST dataset, though, and all of them received nearly flawless scores of 99%. Critically, the superior performance of the ViT-LoRA model was achieved with incredible efficiency. The LoRA method required training only 0.58% of the total model parameters, a tiny fraction of the 8-11% of parameters that needed to be trained for the traditional CNN approach. This leads to the central conclusion of our work: LoRA is not just a more efficient method, but a more effective one. For complex, real-world tasks, the ability to adapt a model's internal representations, as LoRA does for the ViT's attention layers, provides a decisive performance advantage that the rigid, classifier-only fine-tuning method cannot match.

Keywords: vision transformer; LoRA; fine-tuning; CNN; ResNet; EfficientNet

1. Introduction

The advent of deep learning has revolutionized the field of computer vision, with pre-trained models serving as powerful feature extractors for a wide array of tasks. The dominant paradigm involves taking a model pre-trained on a large-scale dataset, such as ImageNet, and fine-tuning it for a specific downstream task. This process, known as transfer learning, has democratized access to state-of-the-art performance. However, the optimal strategy for fine-tuning remains an active area of research, particularly as models grow in size and complexity.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 November 2025 doi:10.20944/preprints202510.2514.v1

This paper explores two contrasting methodologies. The first is the traditional approach, commonly applied to Convolutional Neural Networks (CNNs), where the pre-trained feature extraction layers are frozen, and only a newly added, task-specific classifier head is trained. This method is computationally efficient but limited in its adaptive capacity. The second is a modern Parameter-Efficient Fine-Tuning (PEFT) technique called Low-Rank Adaptation (LoRA), which allows for the adaptation of a model's core components by injecting small, trainable matrices into its layers. We apply LoRA to a Vision Transformer (ViT), a powerful architecture that has recently challenged the dominance of CNNs.

1.1. Convolutional Neural Networks (CNNs)

CNNs have been the cornerstone of computer vision for over a decade. Two prominent examples are ResNet and EfficientNet.

- **ResNet (Residual Network)**: Introduced to solve the vanishing gradient problem in very deep networks, ResNet [1] employs "residual connections" or shortcuts that allow gradients to bypass layers, enabling the training of networks with hundreds or even thousands of layers. This architecture proved that depth is a critical component of model performance.
- **EfficientNet**: This model family introduced a more principled approach to model scaling. Instead of scaling only one dimension (depth, width, or resolution), EfficientNet uses a compound scaling method to uniformly scale all three dimensions, achieving a better balance between accuracy and computational resources.

1.2. Vision Transformers (ViT)

Vision Transformers adapt the Transformer architecture, originally designed for natural language processing, to image data [1]. Instead of processing tokens of words, a ViT processes a sequence of image patches.

The process begins by resizing an input image to a fixed resolution (e.g., 224x224 pixels) and splitting it into a grid of non-overlapping patches (e.g., 16x16 pixels each), as illustrated in Figure 1. These 2D patches are then flattened into 1D vectors and linearly projected into 'patch embeddings.'

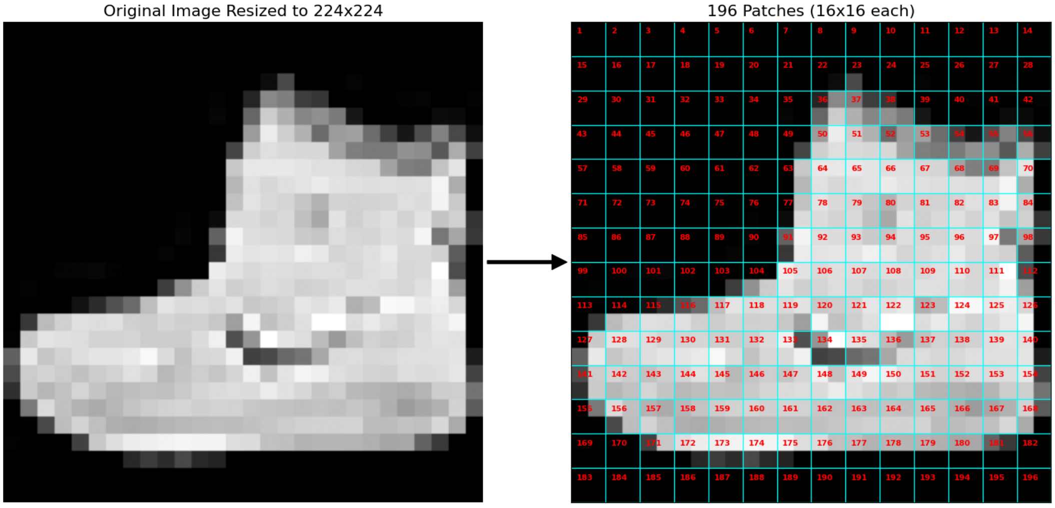


Figure 1. The ViT input pipeline. An original image (an ankle boot from Fashion-MNIST) is resized to 224x224 and then divided into a 14x14 grid of 16x16 patches, resulting in a sequence of 196 patches.

These patch embeddings, along with positional embeddings to retain spatial information, are fed into a series of Transformer Encoder blocks. The core of each encoder is a self-attention mechanism, which allows the model to weigh the importance of every patch relative to every other patch, capturing global context and long-range dependencies within the image. The output from the encoders is then passed to an MLP head for final classification, as shown in the architectural diagram in Figure 2. This study investigates how this architecture can be efficiently fine-tuned using LoRA.

2. Related Work

The comparison between Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) has been a central theme in recent computer vision literature. Raghu et al. explored whether ViTs "see" like CNNs, finding that ViTs tend to learn more global representations early on, whereas CNNs build up from local features [1]. This fundamental difference in representation learning may explain their varying performance on different tasks and their response to fine-tuning strategies. Further comparative studies, such as the one by Dingeto and Kim on adversarial defenses, continue to probe the distinct properties of these two architectural families [2].

The Fashion-MNIST dataset has emerged as a popular benchmark for such comparisons. Bbouzidi et al. provided a literature review on the application of both CNNs and ViTs to this specific dataset [3]. Numerous studies have sought to achieve state-of-the-art results on Fashion-MNIST [4], often by employing ensembles of CNNs or novel architectures [5]. The introduction of ViTs offered a new avenue for enhancing fashion classification, as explored by Abd Alaziz et al., who used ViT for classification and developed recommendation systems [6]. Similarly, the MNIST and CIFAR-10 datasets have served as foundational benchmarks for decades. Chen et al. assessed various neural networks on MNIST [7], while the work by Krizhevsky and Hinton on CIFAR-10 was seminal in demonstrating the power of deep convolutional networks [8].

As models like ViTs grow, the need for efficient adaptation methods has become critical [9]. Low-Rank Adaptation (LoRA) has recently gained prominence as a powerful Parameter-Efficient Fine-Tuning (PEFT) technique. Ulku et al. successfully applied LoRA to adapt ViTs for remote sensing tasks, demonstrating its effectiveness in a specialized domain [10]. This study builds on that work by applying LoRA in a comparative benchmark setting against traditional fine-tuning methods on general-purpose vision datasets.

3. Methodology

This study employs a rigorous benchmarking methodology to compare the performance of a LoRA-adapted Vision Transformer against traditionally fine-tuned CNNs.

3.1. Datasets

We utilize three standard image classification datasets:

- **MNIST**: A dataset of 60,000 training and 10,000 test images of handwritten digits (0-9). The images are grayscale and 28x28 pixels.
- **Fashion-MNIST**: A dataset with the same structure as MNIST, but containing images of 10 classes of clothing items (e.g., T-shirt, Trouser, Ankle boot).
- **CIFAR-10**: A dataset of 50,000 training and 10,000 test color images (32x32 pixels) across 10 classes (e.g., Airplane, Automobile, Dog, Cat).

For compatibility with the pre-trained models, all images were converted to 3-channel RGB and resized to 224 x 224 pixels.

3.2. Models and Fine-Tuning Strategies

- **ResNet50 & EfficientNet-B0**: For these CNNs, we adopt the traditional classifier-only fine-tuning approach. We load models pre-trained on ImageNet, freeze all convolutional backbone layers, and replace the final fully-connected layer with a new classifier head tailored to the 10 classes of each dataset. Only the parameters of this new head are trained.
- **ViT-B/16 with LoRA**: We use a pre-trained Vision Transformer Base model with 16x16 patches. Instead of full fine-tuning, we apply Low-Rank Adaptation (LoRA). The original model weights are frozen. We inject small, trainable low-rank matrices into the query and value projection

matrices of the self-attention mechanism in each Transformer Encoder layer. The update to a weight matrix W_0 is represented by a low-rank decomposition $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are the trainable matrices, and the rank $r \ll d, k$. For this study, we use a rank $r = 16$ and an alpha scaling factor of 32. Similar to the CNNs, the final classifier head is also replaced and trained.

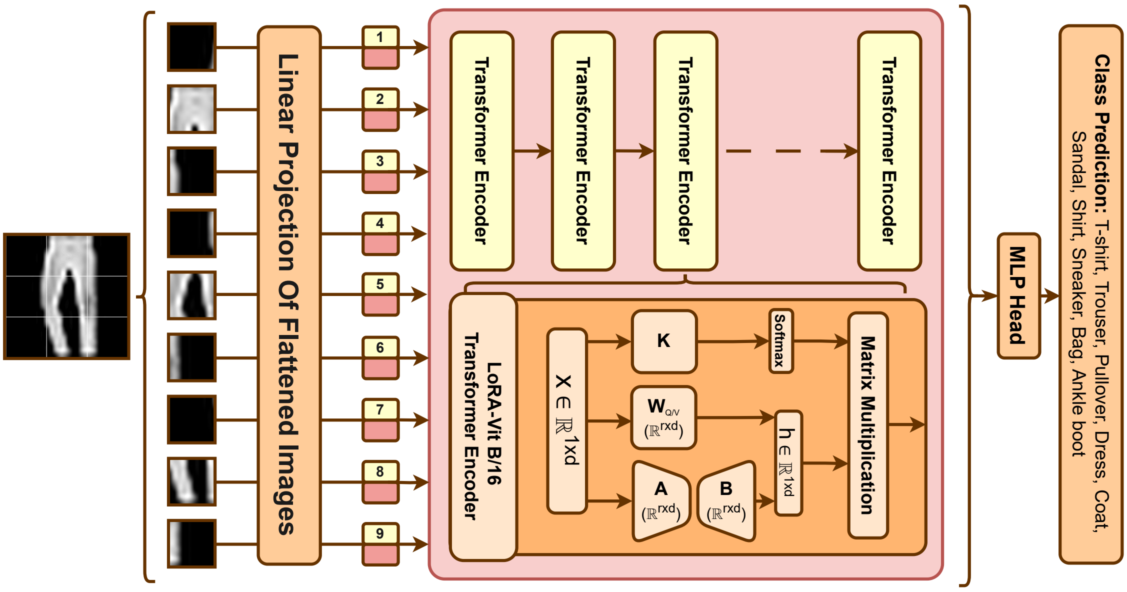


Figure 2. Architecture of the LoRA-adapted Vision Transformer. The input image is decomposed into patches, which are linearly projected and fed through a series of Transformer Encoders. The breakout box shows a detailed view of a LoRA-adapted encoder, where the original weights (W_{qv}) are frozen and the update is learned via two low-rank matrices, A and B.

3.3. Training and Evaluation

All models were trained for 5 epochs using the AdamW optimizer with a Cross-Entropy Loss function. The batch size was set to 64. Standard data augmentations, including random horizontal flips, affine transformations, and brightness/contrast adjustments, were applied to the training sets. Performance was evaluated on the test sets using Validation Accuracy, weighted F1-Score, micro/macro AUC, and detailed confusion matrices.

4. Results

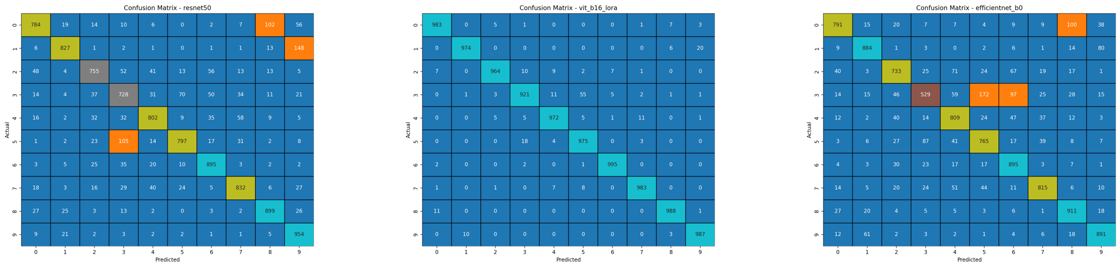
4.1. Benchmark 1: MNIST

On the MNIST handwritten digit dataset, all models achieved near-perfect performance, quickly reaching a point of saturation.

As shown in Table 1, all models surpassed 98.8% validation accuracy. The ViT-LoRA model technically achieved the highest score at 99.12% and a perfect AUC of 1.0000, but the performance difference is negligible. The most significant differentiator is the parameter efficiency: ViT-LoRA achieved this score by training only 0.58% of its parameters, over 14 times fewer than ResNet50.

Table 1. Benchmark Results for MNIST Dataset.

Model	Val Accuracy	F1-Score	AUC-Micro	AUC-Macro	Trainable Pct
vit_b16_lora	0.9912	0.9912	1.0000	1.0000	0.58%
resnet50	0.9902	0.9902	0.9999	0.9999	8.53%
efficientnet_b0	0.9882	0.9882	0.9999	0.9999	10.96%



Confusion Matrix - resnet50

	0	1	2	3	4	5	6	7	8	9
0	784	19	14	10	6	0	2	7	102	56
1	6	827	1	2	1	0	1	1	13	148
2	48	4	755	52	41	13	56	13	13	5
3	14	4	37	728	31	70	50	34	11	21
4	16	2	32	32	802	9	35	58	9	5
5	1	2	23	105	14	797	17	31	2	8
6	3	5	25	35	20	10	895	3	2	2
7	18	3	16	29	40	24	5	832	6	27
8	27	25	3	13	2	0	3	2	899	26
9	9	21	2	3	2	2	1	1	5	954

Confusion Matrix - vit_b16_lora

	0	1	2	3	4	5	6	7	8	9
0	983	0	5	1	0	0	0	1	7	3
1	0	974	0	0	0	0	0	0	6	20
2	7	0	964	10	9	2	7	1	0	0
3	0	1	3	921	11	55	5	2	1	1
4	0	0	5	5	972	5	1	11	0	1
5	0	0	0	18	4	975	0	3	0	0
6	2	0	0	2	0	1	995	0	0	0
7	1	0	1	0	7	8	0	983	0	0
8	11	0	0	0	0	0	0	0	988	1
9	0	10	0	0	0	0	0	0	3	987

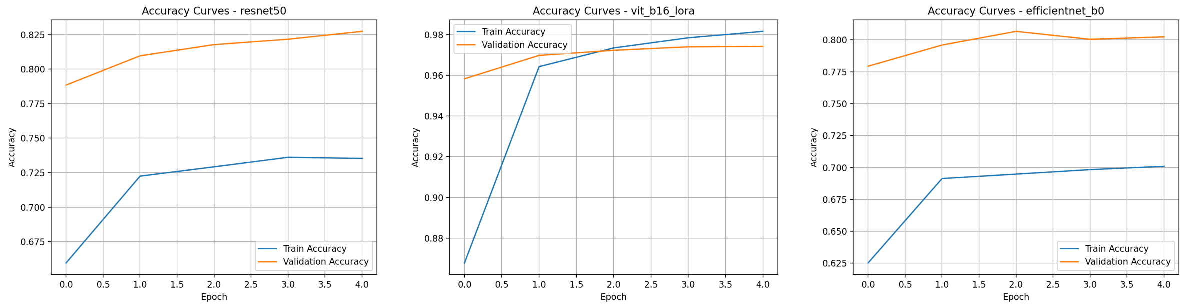
Confusion Matrix - efficientnet_b0

	0	1	2	3	4	5	6	7	8	9
0	791	15	20	7	7	4	9	9	100	38
1	9	884	1	3	0	2	6	1	14	80
2	40	3	733	25	71	24	67	19	17	1
3	14	15	46	529	59	172	97	25	28	15
4	12	2	40	14	809	24	47	37	12	3
5	3	6	27	87	41	765	17	39	8	7
6	4	3	30	23	17	17	895	3	7	1
7	14	5	20	24	51	44	11	815	6	10
8	27	20	4	5	5	3	6	1	911	18
9	12	61	2	3	2	1	4	6	18	891

(a) ResNet50 (b) ViT-B/16-LoRA (c) EfficientNet-B0

Figure 3. Confusion matrices for the MNIST dataset. All models show intensely dominant diagonals, indicating extremely few misclassifications and a near-perfect ability to distinguish between the 10 digits.

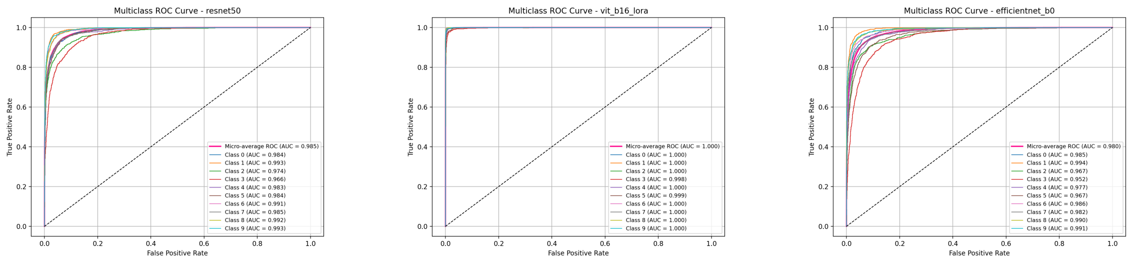
The confusion matrices in Figure 3 visually confirm this near-perfect performance. The off-diagonal errors are minimal, often in the single digits, for all three models. This indicates that the MNIST task is not sufficiently challenging to differentiate the capabilities of these advanced models.



(a) ResNet50 Accuracy (b) ViT-LoRA Accuracy (c) EfficientNet-B0 Accuracy

Figure 4. Accuracy curves for the MNIST dataset. All models rapidly achieve and maintain validation accuracies near 99%, demonstrating quick convergence on this simple task.

A collective analysis of results figures confirms near-perfect performance of all models. The confusion matrices in Figure 3 are virtually flawless, with sporadic, single-digit errors and no systematic confusion between any two classes. This decisive classification is a direct result of the ideal training process visible in the learning curves of Figures 4 and **??**, which show immediate convergence to high accuracy and minimal loss.



(a) ResNet50 ROC (b) ViT-B/16-LoRA ROC (c) EfficientNet-B0 ROC

Figure 5. Multiclass ROC curves for the MNIST dataset. The curves for all models are pushed into the top-left corner, corresponding to perfect (AUC=1.000) or near-perfect (AUC=0.999) classification ability across all classes.

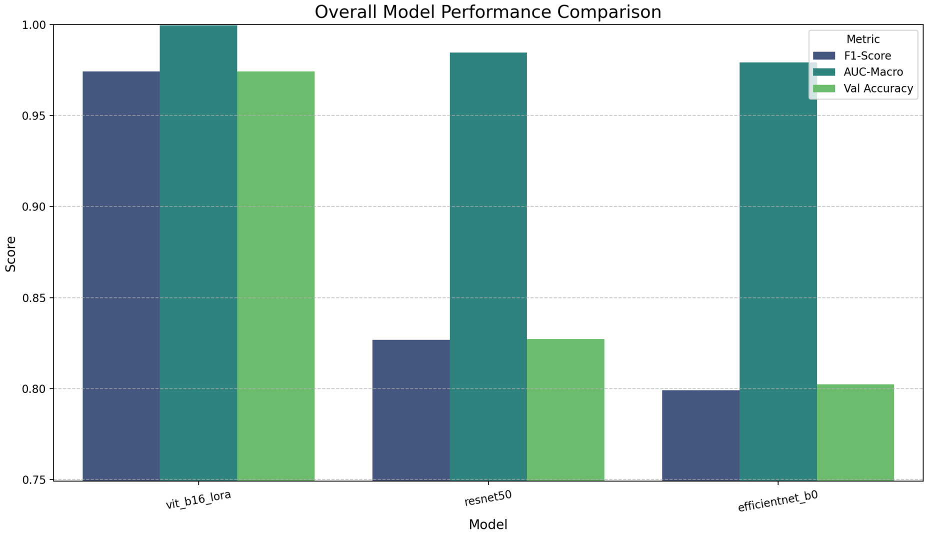


Figure 6. Final performance comparison on MNIST. The bars are all extremely high and close to the maximum score of 1.0, visually representing the performance saturation of all models on this dataset.

A detailed analysis of the learning curves in Figures 4 and **??** provides a clear narrative of performance saturation on the MNIST dataset. All three models—ViT-B/16-LoRA, ResNet50, and EfficientNet-B0 demonstrate an exceptionally rapid mastery of the task. Within the very first epoch, the validation accuracies for all models surge to over 98%, indicating an almost immediate convergence to a near-optimal solution. This initial, steep ascent is mirrored by a dramatic drop in validation loss, which quickly settles into a flat, stable trajectory close to zero. The learning process is so efficient that the curves offer little differentiation between the models after the first epoch.

The generalization gap, which measures the difference between training and validation performance, is virtually non-existent for all three architectures. The training and validation curves for both accuracy and loss remain tightly coupled throughout the five epochs of training. This signifies that the models are not overfitting the training data; rather, they have learned a robust representation of the handwritten digits that generalizes perfectly to the unseen test data. The stability of the curves, which are smooth and devoid of significant fluctuations, further points to a very stable and well-behaved training process.

Ultimately, the learning dynamics confirm that the MNIST dataset is not sufficiently complex to act as a meaningful differentiator for these powerful, pre-trained models. All three fine-tuning approaches are highly effective, pushing the models to their performance ceiling almost instantly. While the final metrics show ViT-B/16-LoRA with a minuscule lead, the learning curves demonstrate that all three models have effectively "solved" the dataset. The primary takeaway from these curves is not a performance distinction but a confirmation of the models' high capacity and the dataset's fundamental simplicity.

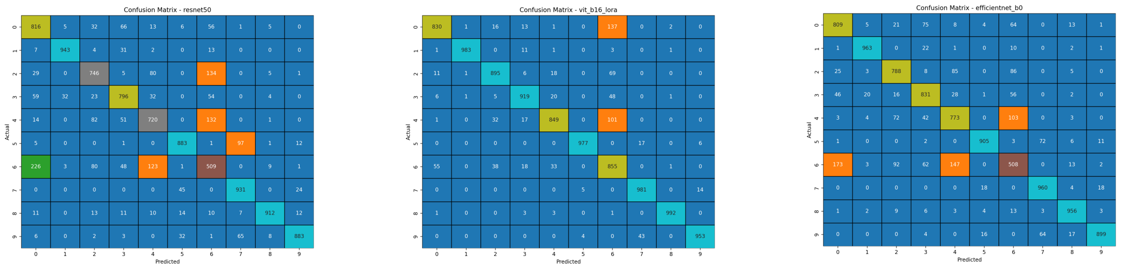
4.2. Benchmark 2: Fashion-MNIST

The Fashion-MNIST benchmark introduced more complexity, allowing for greater performance differentiation. ViT-LoRA established a clear lead over the CNNs.

Table 2 shows the ViT-LoRA model achieving 95.01% accuracy, a significant margin over EfficientNet-B0 (92.31%) and ResNet50 (92.02%). The parameter efficiency of LoRA remains a key advantage.

Table 2. Benchmark Results for Fashion-MNIST Dataset.

Model	Val Accuracy	F1-Score	AUC-Micro	AUC-Macro	Trainable Pct
vit_b16_lora	0.9501	0.9500	0.9982	0.9982	0.58%
efficientnet_b0	0.9231	0.9229	0.9960	0.9960	10.96%
resnet50	0.9202	0.9201	0.9953	0.9953	8.53%

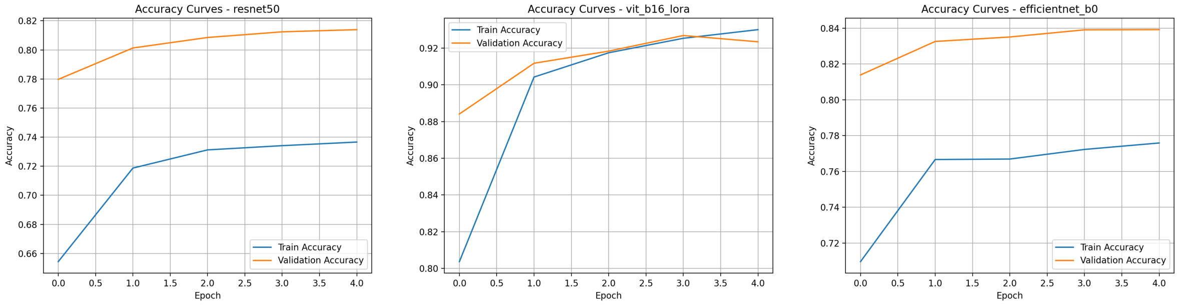


(a) ResNet50 *(b) ViT-B/16-LoRA* *(c) EfficientNet-B0*

Figure 7. Confusion matrices for the Fashion-MNIST dataset. While all models perform well, the ViT-LoRA matrix shows a cleaner diagonal with fewer off-diagonal errors. All models found Class 6 (Shirt) to be the most challenging, often confusing it with other tops.

The results figures for Fashion-MNIST collectively illustrate superior performance and robustness of the ViT-B/16-LoRA model. The confusion matrices in Figure 7 reveal that while all models struggled to distinguish "Shirt" (Class 6) from other tops, the ViT-LoRA matrix shows a cleaner diagonal with fewer misclassifications, demonstrating better fine-grained recognition.

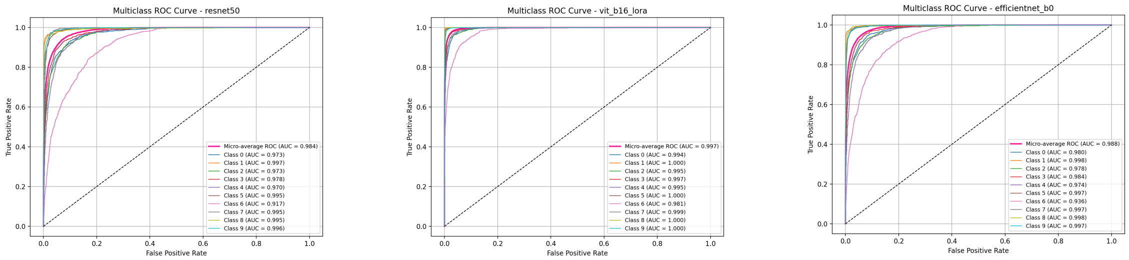
This visual evidence of superiority is supported by the learning curves in Figures 8 and **??**, which show the ViT-LoRA converging faster and to a higher final accuracy.



(a) ResNet50 Accuracy *(b) ViT-LoRA Accuracy* *(c) EfficientNet-B0 Accuracy*

Figure 8. Accuracy curves for the Fashion-MNIST dataset. The ViT-LoRA model converges faster and to a higher final validation accuracy (∼95%) compared to the CNNs (∼92%).

The learning curves (Figures 8 and **??**) and ROC curves (Figure 9) confirm the ViT-LoRA's superior performance, showing faster convergence to a better final result.



(a) ResNet50 ROC *(b) ViT-LoRA ROC* *(c) EfficientNet-B0 ROC*

Figure 9. Multiclass ROC curves for the Fashion-MNIST dataset. The ViT-LoRA model achieves the highest micro-average AUC of 0.998, indicating superior overall classification capability compared to the CNN models.

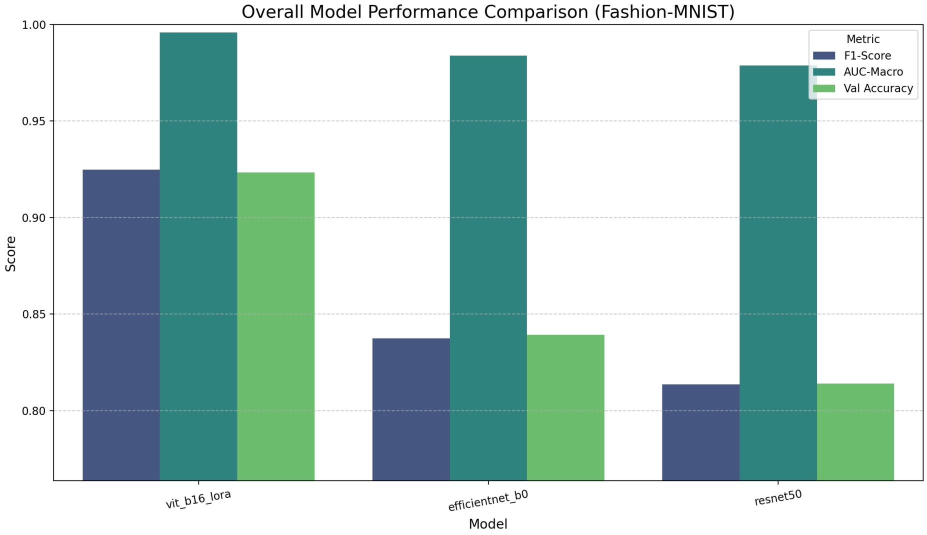


Figure 10. Final performance comparison on Fashion-MNIST. The ViT-LoRA model shows a clear performance advantage across all three key metrics compared to the highly competitive but lower-scoring CNNs.

The learning curves for the Fashion-MNIST dataset, as illustrated in Figures 8 and **??**, offer a more nuanced story and begin to reveal the distinct advantages of the ViT-B/16-LoRA approach. The LoRA-adapted transformer exhibits a demonstrably faster learning rate in the initial phase of training. Its accuracy curve shows a steeper incline, and its loss curve a more pronounced drop within the first epoch. This suggests that the ViT's self-attention mechanism, adapted via LoRA, is more efficient at capturing the essential features of the clothing items from the outset.

Throughout the training process, the ViT-B/16-LoRA model consistently maintains a performance advantage over its CNN counterparts. It converges to a higher final validation accuracy of approximately 95%, while ResNet50 and EfficientNet-B0 plateau around 92%. This sustained performance gap suggests a superior representational power. It is plausible that the ViT's ability to model long-range, global dependencies is better suited for distinguishing between clothing items based on holistic features like silhouette and overall shape, whereas the CNNs focus more on local textures and patterns. The generalization gap remains small for all models, indicating effective regularization and a lack of significant overfitting.

In conclusion, the learning dynamics on Fashion-MNIST highlight the ViT-B/16-LoRA as a more effective and efficient learner for this moderately complex task. While the CNNs perform admirably, displaying stable and consistent learning, their shallower learning curve and lower final accuracy indicate that the traditional classifier-only fine-tuning approach is less potent than the more flexible adaptation provided by LoRA. The ViT-LoRA not only learns faster but also achieves a higher level of mastery over the dataset.

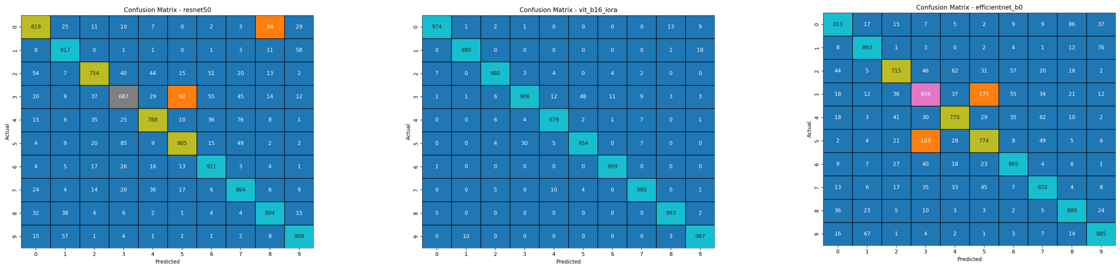
4.3. Benchmark 3: CIFAR-10

The CIFAR-10 dataset, with its color images and significant intra-class variation, presented the greatest challenge and revealed the most dramatic performance gap.

The ViT-LoRA model's superiority is undeniable, as shown in Table 3. It achieved 97.29% accuracy, outperforming ResNet50 by nearly 14 percentage points and EfficientNet-B0 by nearly 17 percentage points.

Table 3. Benchmark Results for CIFAR-10 Dataset.

Model	Val Accuracy	F1-Score	AUC-Micro	AUC-Macro	Trainable Pct
vit_b16_lora	0.9729	0.9729	0.9992	0.9992	0.58%
resnet50	0.8341	0.8336	0.9859	0.9858	8.53%
efficientnet_b0	0.8048	0.8038	0.9806	0.9805	10.96%

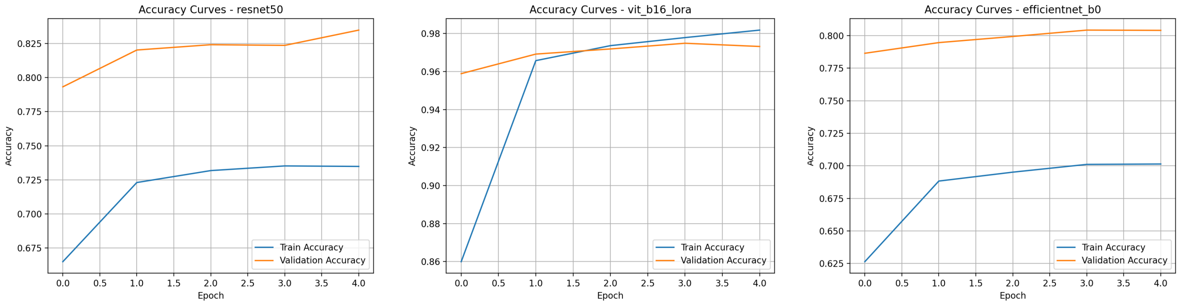


(a) ResNet50 (b) ViT-B/16-LoRA (c) EfficientNet-B0

Figure 11. Confusion matrices for the CIFAR-10 dataset. The ViT-LoRA matrix is exceptionally clean, indicating very few errors. The CNN matrices show significant confusion between related classes, such as cat (3) and dog (5).

The figures for the CIFAR-10 benchmark starkly visualize the immense performance gulf between the ViT-B/16-LoRA and the traditionally fine-tuned CNNs. The confusion matrices in Figure 11 show a remarkably clean diagonal for the ViT-LoRA, contrasting sharply with the significant off-diagonal noise in the CNN matrices, especially their confusion between 'cats' and 'dogs'.

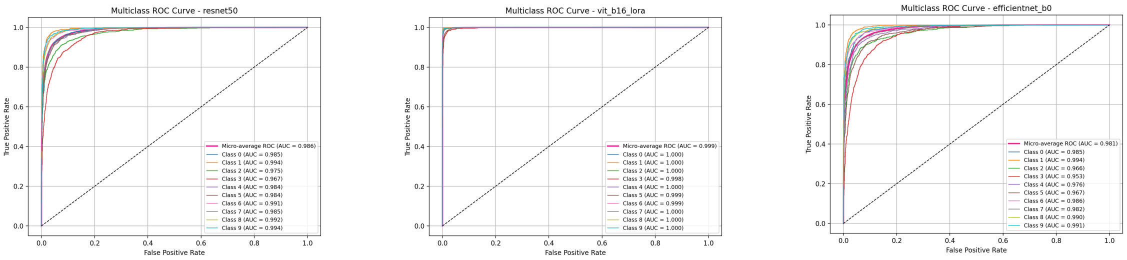
This classification failure is a direct outcome of the handicapped learning process depicted in the learning curves of Figures 12 and **??**, where the CNNs exhibit high loss and a low accuracy ceiling.



(a) ResNet50 Accuracy (b) ViT-LoRA Accuracy (c) EfficientNet-B0 Accuracy

Figure 12. Accuracy curves for the CIFAR-10 dataset. The ViT-LoRA model's validation accuracy starts high and ends at over 97%, whereas the CNNs struggle to surpass the mid-80s.

The learning curves (Figures 12 and **??**) and ROC curves (Figure 13) all tell the same story of ViT-LoRA's dominant performance.



(a) ResNet50 ROC

(b) ViT-LoRA ROC

(c) EfficientNet-B0 ROC

Figure 13. Multiclass ROC curves for CIFAR-10. The ViT-LoRA model's micro-average AUC of 0.999 is significantly higher than that of ResNet50 (0.986) and EfficientNet-B0 (0.981).

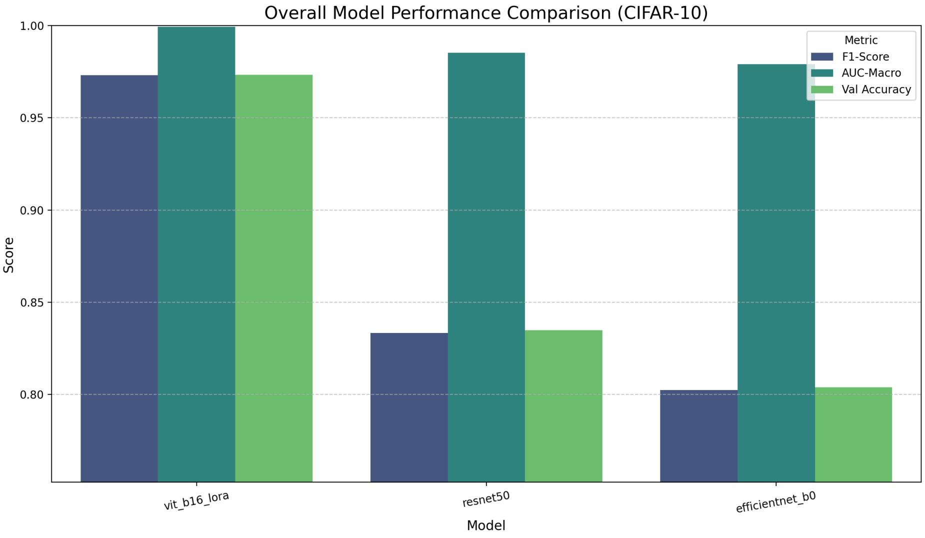


Figure 14. Final performance comparison on CIFAR-10. This plot starkly illustrates the massive performance gap between the ViT-LoRA model and the two CNNs across all metrics.

The learning curves for the CIFAR-10 benchmark, presented in Figures 12 and **??**, starkly illustrate the profound impact of model architecture and fine-tuning strategy on a complex, real-world dataset. The ViT-B/16-LoRA model's performance is in a completely different class from the CNNs. Its validation accuracy begins at a remarkably high baseline of over 95% and rapidly converges to its final, exceptional score of over 97%. This is accompanied by an extremely low and stable validation loss, characteristic of a model that is well-suited to the dataset and is learning its features effectively.

In sharp contrast, the learning curves for ResNet50 and EfficientNet-B0 depict a handicapped and struggling learning process. Their validation accuracies start significantly lower (around 80%) and exhibit a much slower, more laborious climb, ultimately failing to surpass the mid-80s. Their validation loss, while decreasing, remains substantially higher than that of the ViT-LoRA model. This indicates a fundamental limitation of the classifier-only fine-tuning approach for this task. By keeping the convolutional backbone frozen, the models are unable to adapt their feature extractors to the nuanced and varied visual information present in the CIFAR-10 images, leading to significant underfitting and a low performance ceiling.

The enormous chasm in the learning dynamics provides a clear verdict. For a dataset as complex as CIFAR-10, merely training a new classifier head on a frozen feature extractor is an insufficient adaptation strategy. The LoRA methodology, by allowing for efficient yet impactful modifications to the Vision Transformer's core attention layers, enables the model to learn a much richer and more accurate representation of the data. The curves show that ViT-B/16-LoRA is not just quantitatively superior in its final metrics; its entire learning process is qualitatively more effective and powerful from the very first epoch.

5. Conclusions

This empirical investigation into fine-tuning methodologies has yielded clear and compelling results. Our comparative benchmark demonstrates that the choice of fine-tuning strategy is highly consequential and its effectiveness is strongly dependent on the complexity of the downstream task.

The primary conclusion of this study is that Low-Rank Adaptation (LoRA) of a Vision Transformer is a superior fine-tuning methodology compared to traditional classifier-only tuning of CNNs, especially for complex datasets. On CIFAR-10, the ViT-LoRA model did not just perform better; it operated in a different performance class entirely. This suggests that allowing the model to adapt its internal representations via the self-attention mechanism is crucial for successfully transferring knowledge to challenging tasks. The traditional method, by freezing the feature extractor, proves too rigid to adapt effectively.

On the simpler MNIST and Fashion-MNIST datasets, the performance gaps narrowed. For MNIST, all models reached the performance ceiling, rendering the choice of architecture less critical for accuracy. Here, the overwhelming advantage of LoRA is its parameter efficiency; achieving the same top-tier performance by training a fraction of the parameters is a significant practical benefit.

In summary, our findings strongly advocate for the adoption of PEFT methods like LoRA. They not only drastically reduce the computational burden of fine-tuning large models but can also unlock a higher level of performance that traditional, more restrictive methods cannot reach. Future work could explore applying LoRA to the CNN architectures themselves, investigating the impact of different LoRA ranks, and extending this benchmark to larger and more diverse datasets.

To ensure reproducibility and facilitate further investigation by the community, the complete source code for the experiments presented in this paper has been made publicly available on GitHub.

References

1. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems* **2021**, *34*, 12116–12128.
2. Dingeto, H.; Kim, J. Comparative Study of Adversarial Defenses: Adversarial Training and Regularization in Vision Transformers and CNNs. *Electronics* **2024**, *13*, 2534.
3. Bbouzidi, S.; Hcini, G.; Jdey, I.; Drira, F. Convolutional neural networks and vision transformers for fashion mnist classification: A literature review. *arXiv preprint arXiv:2406.03478* **2024**.
4. Mukhamediev, R.I. State-of-the-Art Results with the Fashion-MNIST Dataset. *Mathematics* **2024**, *12*, 3174.
5. Nocentini, O.; Kim, J.; Bashir, M.Z.; Cavallo, F. Image classification using multiple convolutional neural networks on the fashion-MNIST dataset. *Sensors* **2022**, *22*, 9544.
6. Abd Alaziz, H.M.; Elmannai, H.; Saleh, H.; Hadjouni, M.; Anter, A.M.; Koura, A.; Kayed, M. Enhancing fashion classification with vision transformer (ViT) and developing recommendation fashion systems using DINOVA2. *Electronics* **2023**, *12*, 4263.
7. Chen, F.; Chen, N.; Mao, H.; Hu, H. Assessing four neural networks on handwritten digit recognition dataset (MNIST). *arXiv preprint arXiv:1811.08278* **2018**.
8. Krizhevsky, A.; Hinton, G.; et al. Convolutional deep belief networks on cifar-10. *Unpublished manuscript* **2010**, *40*, 1–9.
9. Huynh, E. Vision transformers in 2022: An update on tiny imagenet. *arXiv preprint arXiv:2205.10660* **2022**.
10. Ulku, I.; Tanriover, O.O.; Akagündüz, E. LoRA-NIR: Low-Rank Adaptation of Vision Transformers for Remote Sensing with Near-Infrared Imagery. *IEEE Geoscience and Remote Sensing Letters* **2024**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.