Comparative Analysis of Vision Transformers and Convolutional Neural Networks for Image Classification

Raasikh Naveed, y979j497

School of Computing

College of Engineering

raasikhahmed@gmail.com

*Abstract -* **This study compares the performance of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) on CIFAR-10, CIFAR-100, Oxford-IIIT Pets, and Oxford Flowers-102 datasets across accuracy, training time, and memory efficiency. Results demonstrate ViT’s superior accuracy and generalization but at a cost of higher computational overhead, while CNNs provide a more resource-efficient alternative for simpler tasks.**

1. INTRODUCTION

Convolutional Neural Networks (CNNs) have been the dominant architecture in computer vision for decades, excelling in effectively processing spatial hierarchies within image data. They leverage convolutional layers to extract features, starting with simple patterns such as edges and progressing to more intricate structures like shapes and textures as the network deepens. This hierarchical feature extraction makes CNNs highly effective for tasks like image classification, object detection, and segmentation. Vision Transformers (ViTs), a more recent innovation in computer vision, adopt a fundamentally different approach. Unlike CNNs, ViTs employ a patch-based tokenization strategy where images are divided into fixed-size patches. These patches are then treated as sequences and processed through transformer layers. Borrowing the self-attention mechanisms originally developed for natural language processing (NLP), ViTs model global relationships between image patches, allowing them to capture long-range dependencies more effectively than CNNs. By focusing on global context rather than local patterns, ViTs offer a distinct advantage in handling complex image datasets with diverse spatial relationships. This innovation challenges the traditional dominance of CNNs, marking a paradigm shift in how image data is processed in deep learning.

1. RELATED WORK

Inspired by the seminal paper *"An Image is Worth 16x16 Words,"* Vision Transformers (ViTs) revolutionize computer vision by tokenizing images into patches and processing them as sequences, achieving state-of-the-art results on large datasets like ImageNet. The key insight is that images can be treated similarly to sequences of words in a sentence. ViTs divide images into fixed-size patches (e.g., 16x16 pixels), which are flattened and embedded into vectors. These patch embeddings are then passed through transformer layers, where self-attention mechanisms enable the model to capture global relationships between patches. This approach stands in stark contrast to Convolutional Neural Networks (CNNs), which are designed to exploit local features through convolutional operations. ViTs, by modeling global context from the outset, are particularly adept at capturing long-range dependencies in image data.

ViTs have demonstrated remarkable performance on large datasets like ImageNet, where the availability of extensive training data reduces their reliance on inductive biases such as locality and translation invariance, which CNNs inherently leverage. However, this strength comes with a trade-off: ViTs often require pretraining on massive datasets or rely on transfer learning to generalize effectively, particularly for smaller datasets where limited data can pose a challenge.

1. PROPOSED APPROACH
   1. Motivation

CNNs have been the go-to architecture for image classification tasks, while ViTs have shown potential for better generalization, especially on complex datasets. This study aims to evaluate their comparative strengths and weaknesses.

* 1. Model Architecture
* **CNN:** ResNet18 pretrained on ImageNet, with the fully connected layer replaced for task-specific classes.
* **ViT:** Vision Transformer Base with a patch size of 16, pretrained on ImageNet, and fine-tuned on target datasets.

C. Dataset

* **CIFAR-10:** 10 classes, small images (32x32 pixels).
* **CIFAR-100:** 100 classes, more complex small images.
* **Oxford-IIIT Pets:** 37 classes, medium complexity.
* **Oxford Flowers-102:** 102 classes, high-class diversity.

D. Training Procedure

* **CNN:** Trained using SGD optimizer, learning rate of 0.01, and CrossEntropy loss.
* **ViT:** Trained using AdamW optimizer with a learning rate of 0.0001.
* Layers frozen for computational efficiency, with only the fully connected (or head) layers fine-tuned.

1. Results

A. Evaluation Metrics

* **Accuracy:** Validation accuracy at specific thresholds (e.g., 95%).
* **Training Time:** Seconds per epoch.
* **Memory Usage:** GPU memory allocated during training.

B. Key Results

* **CIFAR-10**

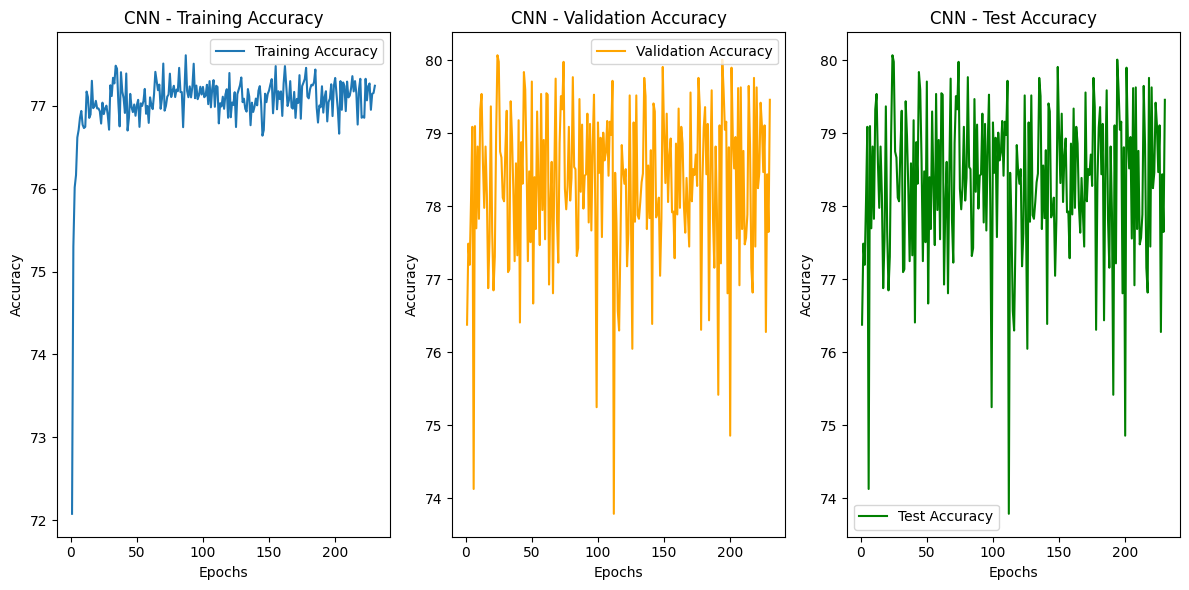


Fig-1

The training accuracy shows a steady increase, stabilizing at approximately 77% to 78%. This suggests that the model learns progressively, although its capacity might be somewhat constrained for CIFAR-10 due to the frozen layers in the pre-trained ResNet18. The reliance on the pre-trained architecture with limited fine-tuning may cap its performance on this dataset.

Validation accuracy exhibits more significant fluctuations, ranging between 78% and 80%. While the model demonstrates good generalization, the variations in validation accuracy might indicate sensitivity to the composition of validation batches, which could be influenced by factors like class imbalance or noise in the dataset. Similarly, test accuracy fluctuates within a comparable range of 78% to 80%, aligning closely with the validation performance. This consistency suggests that the model's ability to generalize extends to unseen data, despite the observed fluctuations.

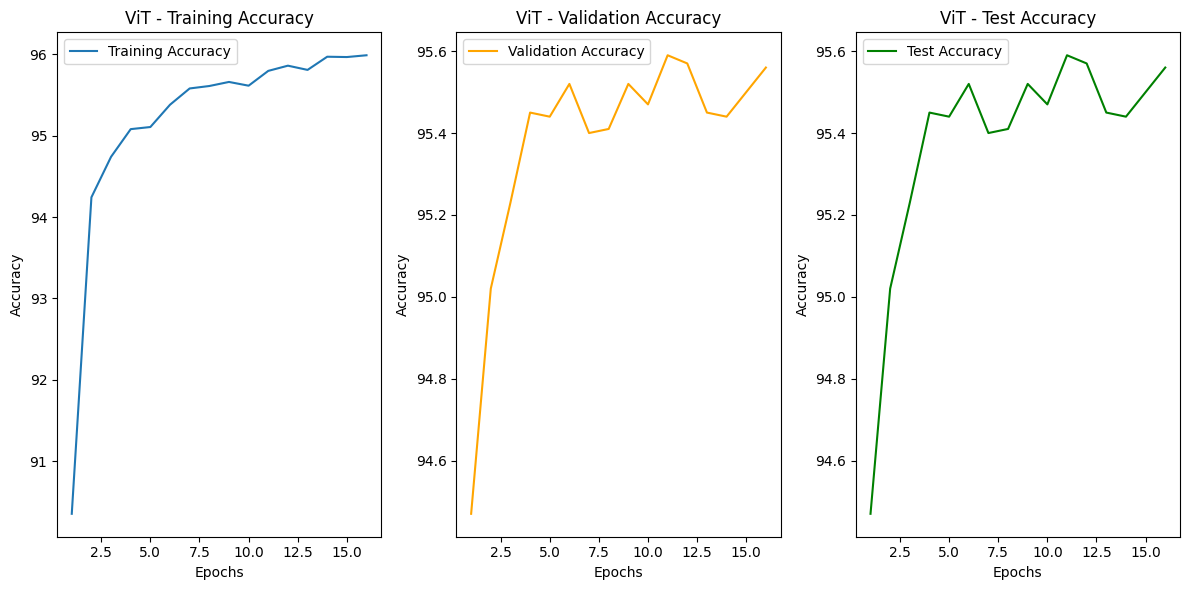


Fig-2

The training accuracy increases rapidly during the initial epochs, reaching approximately 96% by epoch 15. This highlights the ViT model's ability to efficiently learn from the training data, leveraging its patch embeddings and self-attention mechanisms to extract and model features effectively.

The validation accuracy shows a consistent upward trend, peaking at around 95.6% by epoch 15. This alignment with training accuracy indicates that the model not only learns the training data efficiently but also generalizes well to the validation set, reflecting robust performance.

Similarly, the test accuracy stabilizes around 95.5% in the final epochs, closely mirroring the validation accuracy. This close alignment between test and validation accuracies further confirms the model’s strong generalization capabilities, demonstrating its effectiveness on unseen data.

**CIFAR-100:**

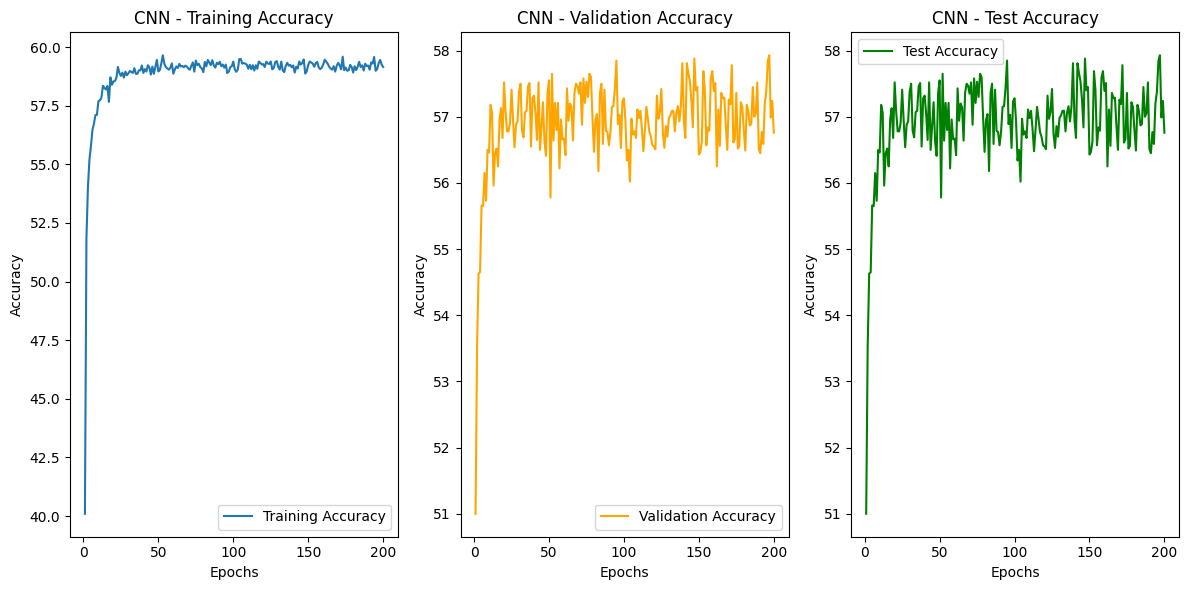


Fig-3

The training accuracy starts at around **40%** and gradually increases to approximately **60%** over 200 epochs. The slow and steady improvement in accuracy highlights that the model takes longer to learn the more complex and diverse features present in CIFAR-100 compared to CIFAR-10. The validation accuracy stabilizes at around **57%** with slight oscillations after the initial epochs. The plateau suggests that the model may not effectively capture the diverse features of the CIFAR-100 dataset, likely due to its complexity. The test accuracy aligns closely with the validation accuracy, peaking at **58%**. The close alignment between test and validation accuracy indicates consistent generalization, but the overall performance remains relatively modest.

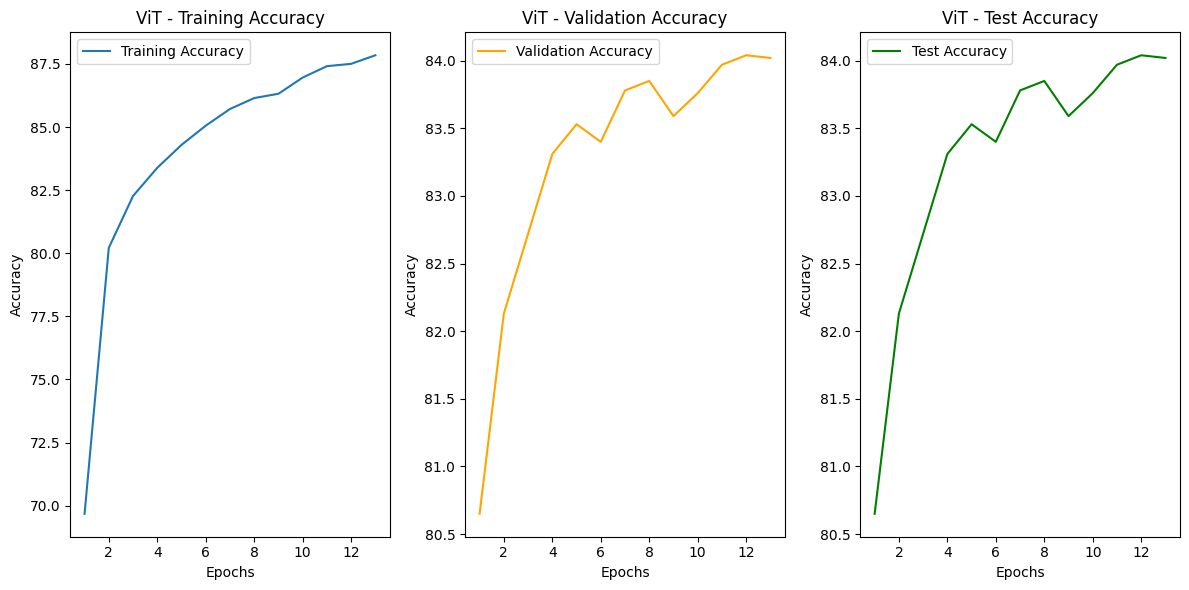


Fig-4

The training accuracy starts at **70%** and quickly rises to over **87%** within just 12 epochs..

Validation accuracy follows a similar trend, stabilizing at approximately **84%**. The high validation accuracy suggests strong generalization capabilities, indicating that the ViT is well-suited for the dataset's complexity. The test accuracy aligns closely with validation accuracy, peaking at around **84%**. The consistency between validation and test accuracy demonstrates reliable performance and minimal overfitting.

* **Oxford-IIIT Pets:**

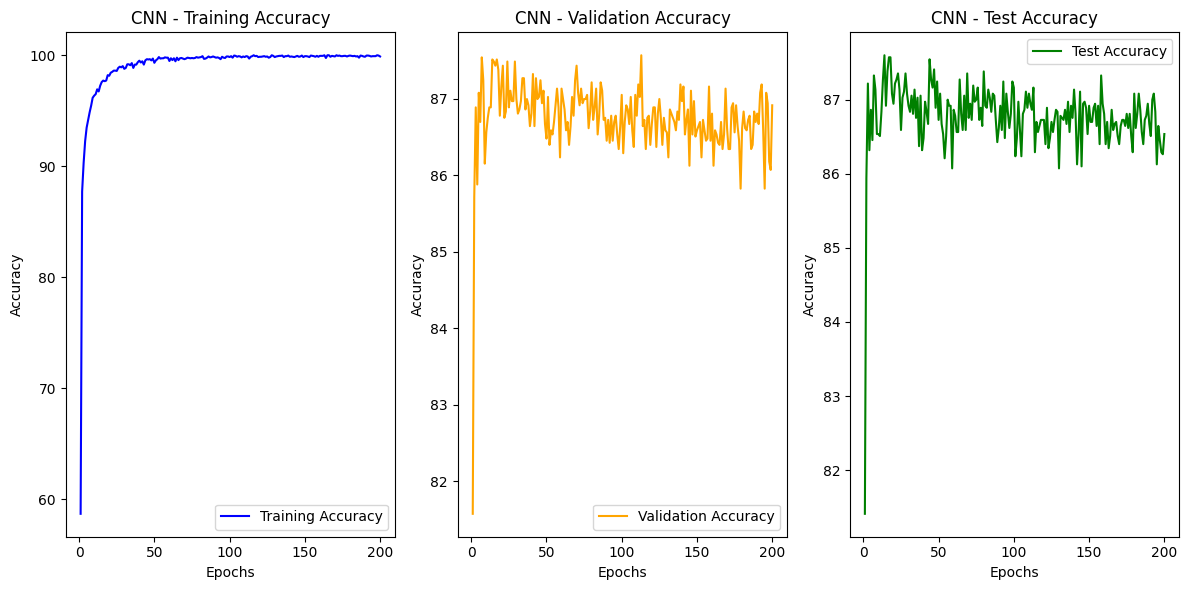


Fig-5

The training accuracy starts around **60%** and converges rapidly, reaching **near-perfect accuracy** (~100%) after approximately 50 epochs. The steep initial learning curve indicates that the CNN quickly learns the hierarchical features of the Oxford Pets dataset. However, the high accuracy of training raises potential concerns about overfitting. The validation accuracy stabilizes around **87%**, showing small fluctuations throughout the 200 epochs. The test accuracy is consistent with the validation accuracy, stabilizing around **87%**. The minimal gap between validation and test accuracy indicates that the model generalizes well to unseen data from the same distribution.

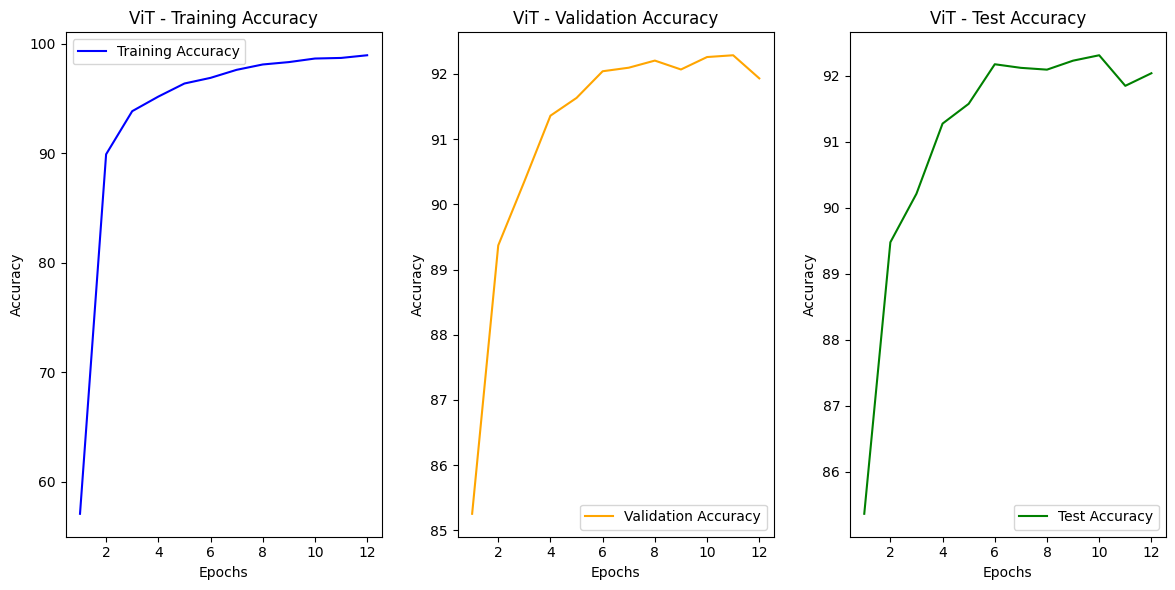


Fig-6

The training accuracy starts around **60%** and increases steadily, reaching nearly **100%** by the 10th epoch. ViT demonstrates a rapid learning curve, indicating that the model effectively captures global patterns using its self-attention mechanism. Validation accuracy starts at **85%** and stabilizes around **92%** after the 6th epoch. Test accuracy closely follows the validation accuracy trend, stabilizing around **92%** after the 6th epoch. The consistent test accuracy implies strong generalization, with ViT effectively capturing dataset-specific features while avoiding overfitting.

**Oxford Flowers-102**

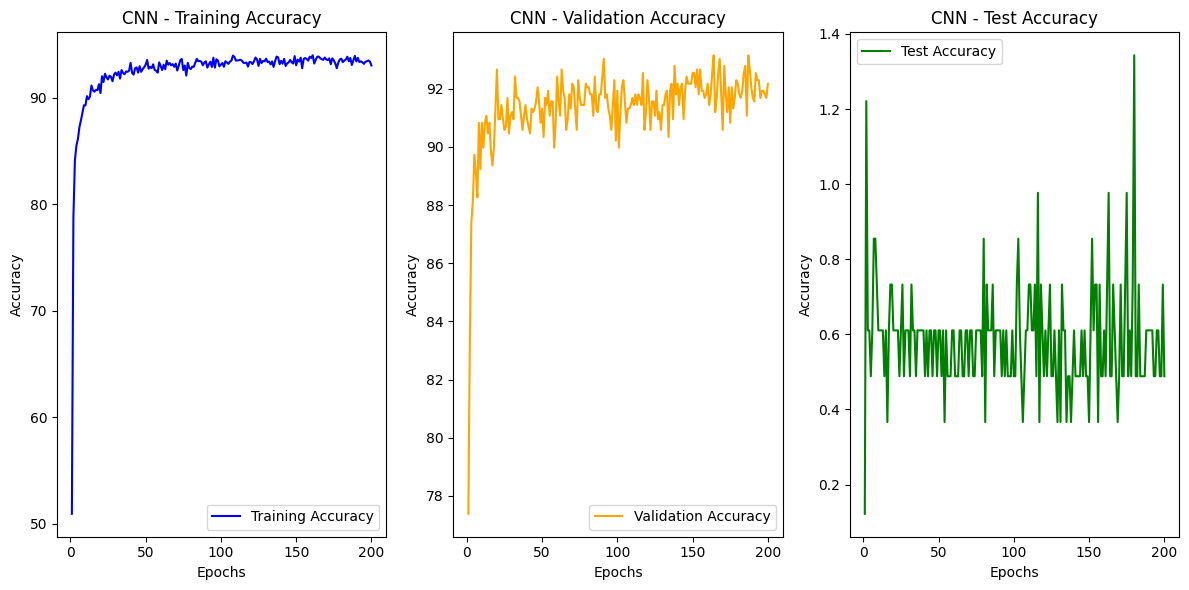


Fig- 7

The **training accuracy** starts at around 50% and steadily climbs to above 90% by approximately 50 epochs. After that, it plateaus with small fluctuations.

This indicates that the model is effectively learning during training.

The **validation accuracy** begins near 78% and increases rapidly to around 90% in the early epoch. It stabilizes above 92% with minor oscillations over the remaining epochs.

This close alignment between validation and training accuracy suggests that the model generalizes well during training and does not significantly overfit the training data.

The **test accuracy graph**, in contrast, is quite erratic, fluctuating significantly across epochs. It ranges between very low values (close to 0.2) and peaks near 1.4 (interpreted as percentages here).

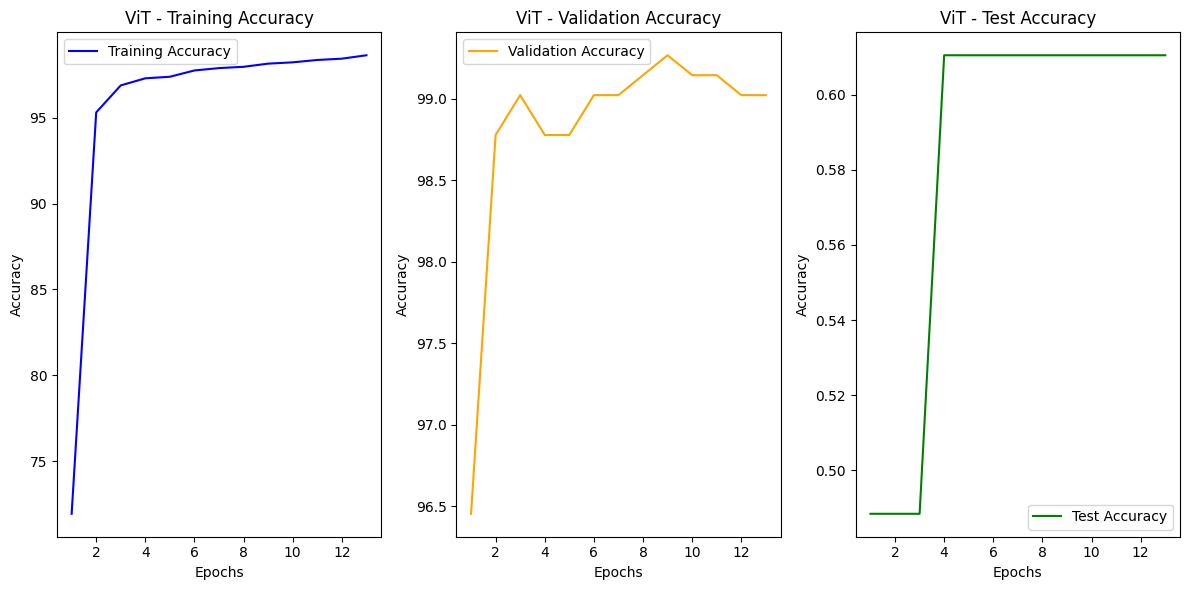


Fig- 8

The **training accuracy** shows a sharp increase, starting from around 70% and climbing above 95% within the first few epochs. By the 12th epoch, the accuracy is approaching 99%.

The **validation accuracy** also starts near 96% and stabilizes above 99% by the 6th epoch. This close alignment with the training accuracy indicates that the model is generalizing well to unseen data. The high validation accuracy confirms the effectiveness of ViTs on relatively complex datasets like Oxford Flowers. The **test accuracy** exhibits an unusual trend, showing a sudden jump from 0.5 (or 50%) to 0.6 (or 60%) and remaining constant across epochs.

**Accuracy**

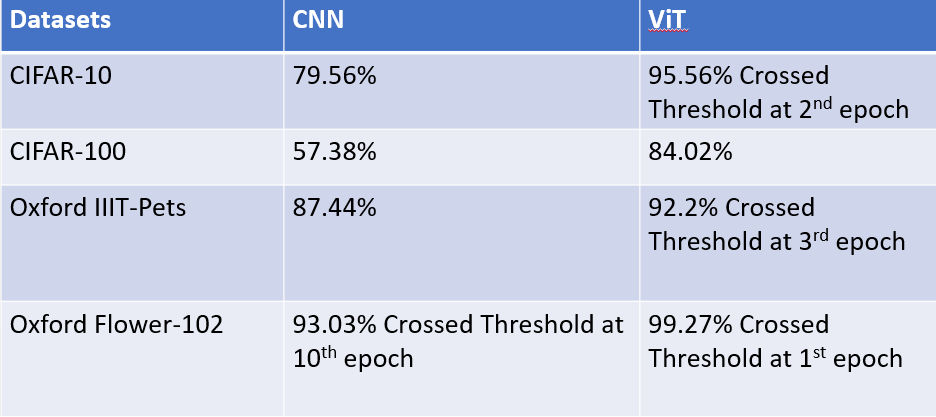


Fig-9

**Training Time**

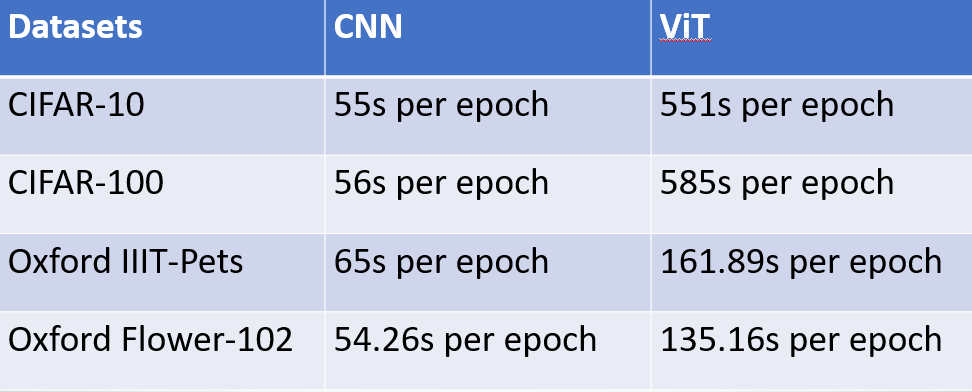


Fig-10

**Memory Used**

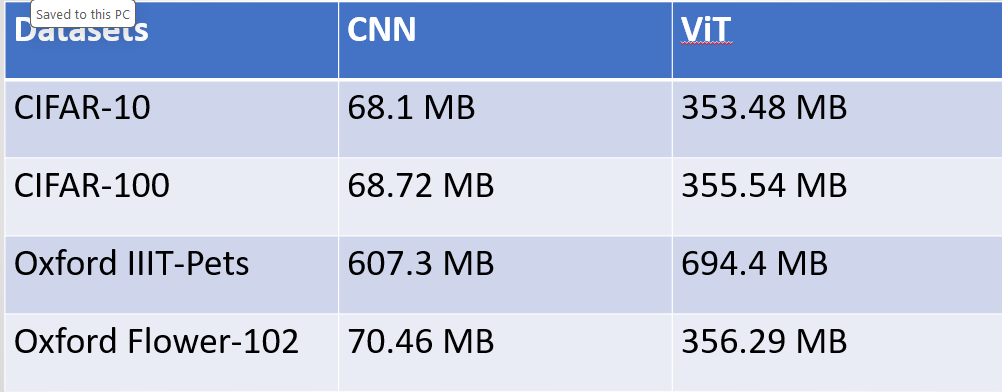


Fig-11

**C. Limitations**

* ViT’s training time per epoch is significantly higher (e.g., 585s for CIFAR-100 compared to CNN’s 56s).
* ViT’s memory usage is consistently greater, highlighting its computational demands.

1. CONCLUSION

The Vision Transformer (ViT) consistently outperformed Convolutional Neural Networks (CNNs) in terms of accuracy across all datasets. It demonstrated faster learning by achieving threshold accuracies in fewer epochs. However, ViT required longer training times per epoch compared to CNN due to its computational complexity. In terms of memory usage, ViT exhibited significantly higher memory allocation compared to CNN, particularly on smaller datasets like CIFAR-10. Interestingly, the gap in memory usage became less pronounced on larger datasets such as Oxford-IIIT Pets, highlighting ViT's scalability for more complex datasets. This adaptability across diverse datasets, from CIFAR-100 to Oxford Flowers, underscores ViT's strong generalization capabilities. While CNNs excelled in time and memory efficiency, they struggled to match the accuracy achieved by ViT, particularly on more challenging datasets like CIFAR-100. ViT's computational demands and resource requirements make it less suited for resource-constrained environments. However, its superior accuracy and ability to generalize across varied and complex datasets establish it as a compelling choice for applications where precision and performance are critical. Conversely, CNN offers a more resource-efficient alternative, making it an excellent choice for tasks where computational budgets are limited, and the trade-off for slightly lower accuracy is acceptable.

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to my professor, **Mr. Lokesh Das**, for his invaluable guidance and unwavering support throughout this semester. His assistance in meeting all the necessary requirements for this project has been truly instrumental. This project draws inspiration from the groundbreaking paper, *“An Image is Worth 16x16 Words.”* The paper was both fascinating and highly informative, offering fresh insights into the field of image classification in deep learning. It introduced innovative concepts with immense potential to revolutionize the domain. Many of the ideas and methodologies employed in this project are directly inspired by this work. I would also like to extend my appreciation to the various AI tools that played a crucial role in the successful completion of this project. These tools not only facilitated the technical aspects but also enabled me to understand complex concepts on a much deeper level.

REFERENCES

### Dobovitskiy, A. *et al.* (2021) *An image is worth 16x16 words: Transformers for image recognition at scale*, *arXiv.org*. Available at: https://arxiv.org/abs/2010.11929 (Accessed: 05 December 2024).

### <https://www.cs.toronto.edu/~kriz/cifar.html> CIFAR-10

### <https://www.cs.toronto.edu/~kriz/cifar.html> CIFAR-100

### [https](file:///C:\Users\HP\Downloads\https)[://paperswithcode.com/dataset/oxford-iiit-pets-1](https://paperswithcode.com/dataset/oxford-iiit-pets-1)

### [https](file:///C:\Users\HP\Downloads\https)[://www.robots.ox.ac.uk/~vgg/data/flowers/102/](https://www.robots.ox.ac.uk/~vgg/data/flowers/102/)