Fine-Tuning Pretrained Image Transformer Models on the iNaturalist Dataset: A Comparative Study

Yuhang Song
Boston University
yuhangs@bu.edu

Zhengrong Gu
Boston University
gzr@bu.edu

Abstract

Image classification, a significant application of computer vision, is now widely used in industries such as transportation, manufacturing, and agriculture. Various computer vision models, including Residual Neural Networks (ResNets), Convolutional Neural Networks (CNNs), and Vision Transformers (ViTs), enable such applications. This project focuses on recently published vision transformer models, specifically, the Neighborhood Attention Transformer (NAT), the Dilated Neighborhood Attention Transformer (Di-NAT), the Dual Attention Vision Transformers (DaViT), and the Multi-Axis Vision Transformer (MaxViT) for image classification tasks. Our research objectives will be achieved by first demonstrating and comparing the structures of these four models from a theoretical perspective. Subsequently, we will fine-tune these models on the iNaturalist dataset for fungi classification tasks and compare their performance. Lastly, we will attempt to optimize the models to achieve superior performance.

1. Introduction

In late 2020, the Vision Transformer (ViT) was introduced as an image classifier that utilizes a Transformer Encoder operating on an embedded space of image patches, primarily for large-scale training [2]. This approach ignited a trend in creating Transformer-like models with high data efficiency. For instance, the Neighborhood Attention Transformer (NAT) was proposed in 2022 [4], building upon the foundations of the Vision Transformer. Subsequent research led to the emergence of the Dilated Neighborhood Attention Transformer (DiNAT) [3]. The Dual Attention Vision Transformers and Multi-Axis Vision Transformer are other examples of ViT models that were recently published [1, 6].

In this work, we revisit the concepts of Neighborhood Attention (NA) and the Neighborhood Attention Transformer (NAT), a scalable hierarchical transformer based on NA. We explore the DiNAT, a simple yet potent sparse global attention pattern, allowing for exponential receptive field growth and capturing longer-range context without any additional computational burden, all while maintaining the symmetry in the neighborhood introduced by NA. We compare these two models from both theoretical and experimental perspectives. Additionally, we study the DaViT, another simple, yet effective, transformer architecture designed to capture the global context while maintaining computational efficiency. Lastly, we discuss the MaxViT, an efficient and scalable attention model that can "see" globally throughout the entire network, even at earlier, high-resolution stages.

2. Related Works

All four models we have chosen come with available code. In this project, we analyze and compare the original code published with the papers. Then, we use the pretrained model provided by Hugging Face to fine-tune and optimize the models.

3. Dataset

The project utilizes the iNaturalist dataset [5] to train and evaluate the performance of different models. Due to time and device constraints, we plan to use the Fungi subcategories as our image classification target. This Fungi subset contains 121 categories with 5,826 training images and 1,780 validation images. For demonstration, Figure 1 shows three examples of categories in the Fungi dataset: Phaeolus schweinitzii, Geastrum saccatum, and Lactarius alnicola, from left to right. However, this dataset is highly biased, as shown in Figure 2, which displays the number of images in each category.

4. Model Comparisons

The NAT and DiNAT are both variations of the transformer model that have been proposed to improve the modeling of long-range dependencies in sequences. While both models share similarities with the standard transformer ar-

Figure 1. Illustration of the iNaturalist Fungi Images

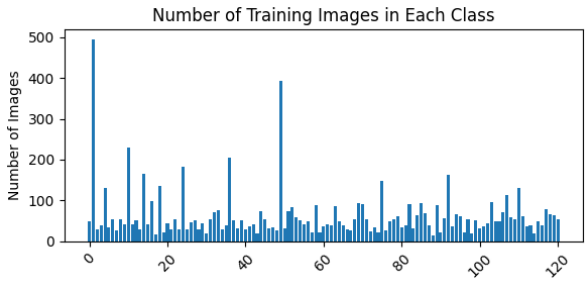


Figure 2. Biased iNaturalist Fungi Dataset

chitecture, they possess key differences in terms of their attention mechanism and receptive field.

The NAT model was introduced as an extension of the original transformer to address the computational inefficiency of the standard self-attention mechanism. In the standard transformer, self-attention is applied to all positions in the input sequence, resulting in quadratic complexity relative to the sequence length. NAT aims to reduce this complexity by limiting attention to a neighborhood of nearby positions. Rather than attending to all positions, NAT restricts attention to a fixed-size neighborhood around each position. This significantly reduces computational complexity and makes it more efficient for longer sequences. The receptive field of each position is limited to a fixed neighborhood size, regardless of the distance between the positions. This permits the effective capture of local dependencies but may struggle when modeling long-range dependencies.

In contrast, the dilated version of the NAT model, DiNAT, extends the neighborhood attention mechanism by incorporating dilation. Dilation introduces a parameter that controls the spacing between the attended positions, allowing for a larger receptive field without sacrificing computational efficiency. DiNAT introduces a dilation factor that determines the spacing between attended positions. By increasing the dilation factor, the receptive field can be expanded exponentially while maintaining a linear computational complexity. Unlike NAT, DiNAT has a variable receptive field that can capture both local and long-range dependencies. The receptive field increases with the dilation factor, enabling the model to effectively capture global context.

Providing a detailed comparison between NAT (and DiNAT) with DaViT and MaxViT is challenging. However, there are specific details that distinguish these models. The concept of dual attention in vision transformers typically refers to the incorporation of two different types of attention mechanisms to enhance the model's ability to capture both spatial and channel-wise dependencies. Spatial attention focuses on modeling the relationships between different spatial positions within an image. It helps the model attend to relevant regions and capture spatial dependencies, allowing for more effective feature extraction. Channel attention, on the other hand, focuses on modeling relationships between different channels or feature maps. It aims to assign importance to specific channels to capture relevant semantic information and enhance feature representation. By combining both spatial and channel attention mechanisms, the DaViT model can potentially improve the performance of vision transformers by capturing both spatial and channel-wise dependencies simultaneously.

On the other hand, the concept of multi-axis attention in vision transformers refers to the utilization of attention mechanisms along multiple axes or dimensions within an image. Instead of applying attention only in the spatial domain, MaxViT extends the attention mechanism to other axes, such as scales or resolutions. By incorporating attention along multiple axes, the MaxViT model aims to capture multi-scale information and hierarchical relationships in visual data. This allows the model to handle variations in object sizes, capture context at different levels of detail, and potentially improve performance on tasks that require understanding and processing images at multiple scales.

5. Experiments and Results

5.1. General Results and Accuracy

We fine-tuned the Neighborhood Attention Transformer based on nat-mini-in1k-224 and achieved an accuracy of 75.28%, as shown in Figure 3.

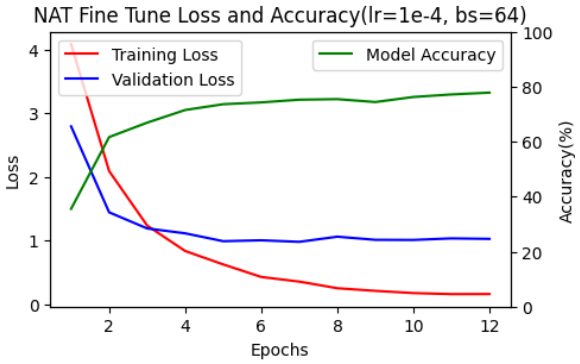


Figure 3. NAT Fine Tune Loss and Accuracy

Next, we fine-tuned the DiNAT with the base model nat-mini-in1k-224 and achieved an accuracy of 77.32%, as shown in Figure 4.

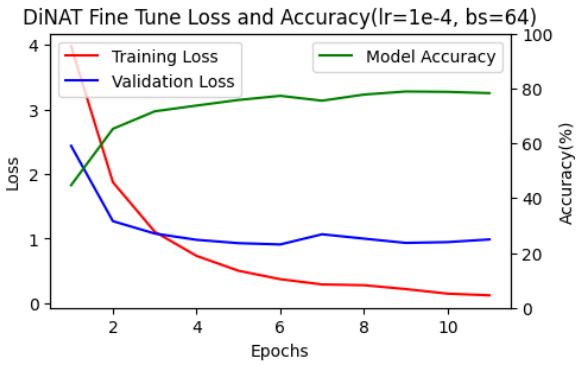


Figure 4. DiNAT Fine Tune Loss and Accuracy

We then fine-tuned the Multi-Axis ViT based on maxvit_tiny_rw_224 and achieved an accuracy of 78.56%, as shown in Figure 5.

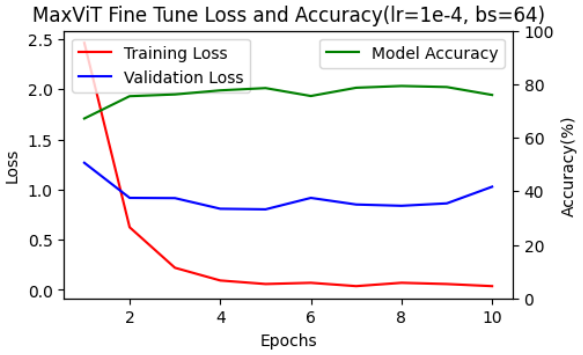


Figure 5. MaxVit Fine Tune Loss and Accuracy

Finally, we achieved the highest accuracy with a model fine-tuned based on davit_tiny.msft_in1k, which achieved an accuracy of 80.93%, as shown in Figure 6.

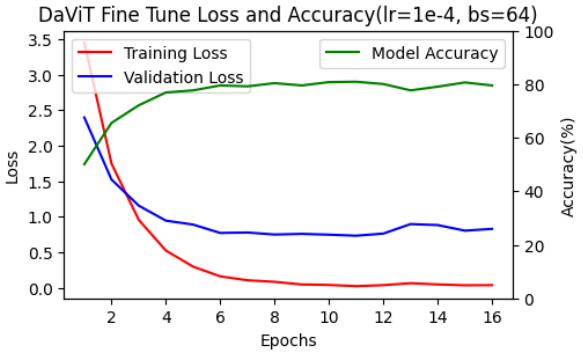
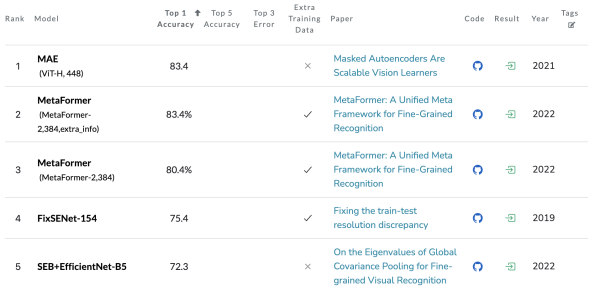


Figure 6. DaViT Fine Tune Loss and Accuracy

Compared to the most accurate algorithms on this dataset, as shown in Figure 7, our model achieved a comparable accuracy with state-of-the-art performance.

Since DaViT achieved the highest accuracy in our tests, we conducted analysis and optimization on that model.



Rank	Model	Top 1 Accuracy	Top 5 Accuracy	Top 3 Error	Extra Training Data	Paper	Code	Result	Year	Tags
1	**MAE** (ViT-H, 448)	83.4			×	Masked Autoencoders Are Scalable Vision Learners			2021	
2	**MetaFormer** (MetaFormer-2,384,extra_info)	83.4%			✓	MetaFormer: A Unified Meta Framework for Fine-Grained Recognition			2022	
3	**MetaFormer** (MetaFormer-2,384)	80.4%			✓	MetaFormer: A Unified Meta Framework for Fine-Grained Recognition			2022	
4	**FixSENet-154**	75.4			✓	Fixing the train-test resolution discrepancy			2019	
5	**SEB+EfficientNet-B5**	72.3			×	On the Eigenvalues of Global Covariance Pooling for Fine-grained Visual Recognition			2022	

Figure 7. Rank Board of iNaturalist 2017 Dataset

5.2. Model Debiasing

As we mentioned earlier, the dataset itself is biased. Some categories are overrepresented due to excess data, while others are underrepresented due to a lack of data. As per feedback from the professor during the presentation, we decided to try to debias the training data. Here, we manually set the number of images we wanted for each category, oversampling underrepresented categories and undersampling overrepresented ones. We analyzed how the number of images in each category affected model accuracy.

Shown in Figure 8, we found that when we set the number of images to 400 for each category, the model achieved its highest accuracy.

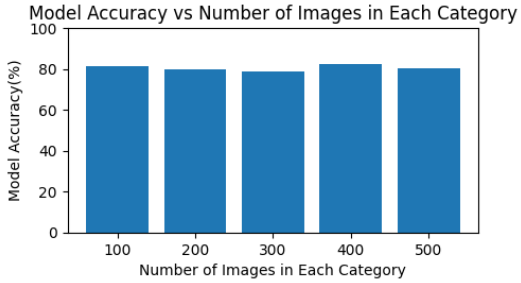


Figure 8. Model Accuracy vs Number of images in Each Category

5.3. Bias Analysis

After debiasing the training data, we plotted the confusion matrix in Figure 9, with a close-up view in Figure 10. Diagonal elements show the correct prediction. There is no single category that appears more likely to be classified incorrectly. This indicates our model is unbiased.

5.4. Learning Rate

As shown in Figure 11, we found that fine-tuning the model requires an extremely small learning rate since the model is already well-trained, and a large learning rate will prevent it from learning.

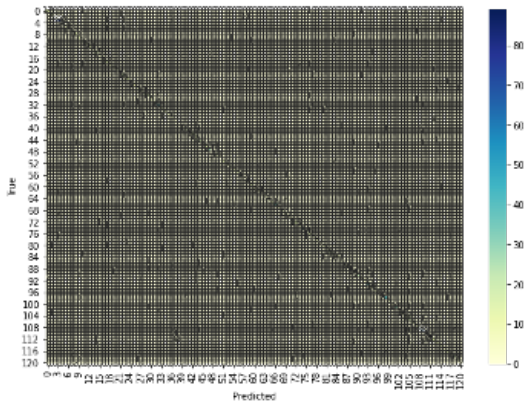


Figure 9. Confusion Matrix

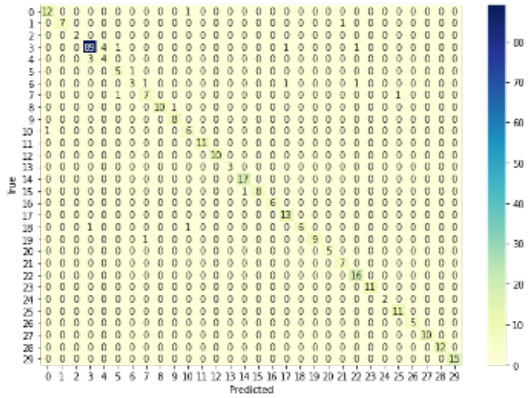


Figure 10. Close-Up View of Part of Confusion Matrix

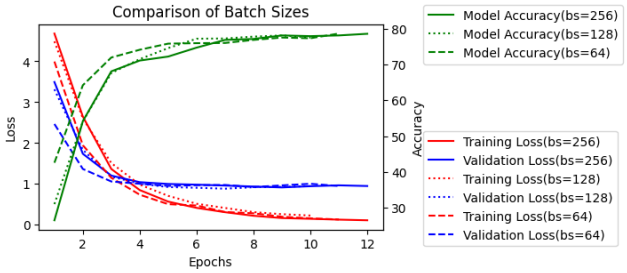


Figure 11. Close-Up View of Part of Confusion Matrix

5.5. Fine-Tuning with Extra Layers

In all previous experiments, we fine-tuned the existing parameters in the original model. However, another way to improve performance is by adding extra layers. The pre-trained model will be used as a feature extractor, and the extra layers will perform the classification.

We tried multiple combinations of layers and their sizes, and the best-performing combination was a single Linear layer added back to the model.

However, as shown in Figure 12, these extra layers did not improve the accuracy. We suspect this is because the dataset is too small, and the model is already overfitting. Adding extra layers increased the model complexity and worsened the overfitting. Hence, it did not improve the accuracy.

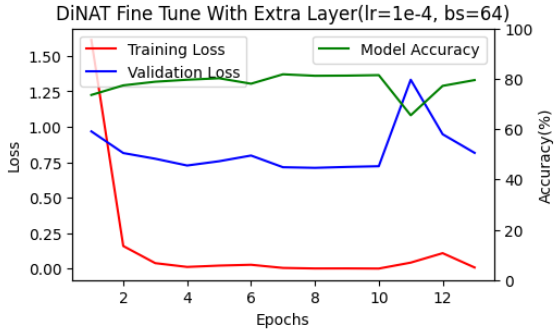


Figure 12. DiNAT Fine Tune With Extra Layer

5.6. Training a Lightweight DiNAT from Scratch

For comparison, we also trained a lightweight DiNAT from scratch. The loss is shown in Figure 13.

Training from scratch had the worst performance with an accuracy of 27.35%. It is not surprising that training from scratch has lower accuracy, as training a transformer requires a large amount of data, and here we only had about 5000 images for training.

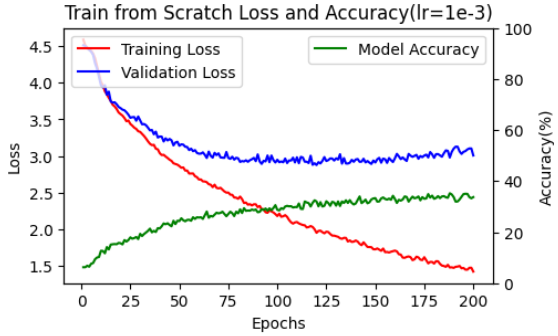


Figure 13. Training from Scratch Loss and Accuracy

6. Conclusion

This project offers a comparative analysis of the four aforementioned transformer-based models in image classification tasks, underlining the strengths and drawbacks of each model. By applying these models to the iNaturalist dataset, we aim to gain insights into their performance when faced with a different dataset. We hope that this project will prove valuable to researchers and practitioners in the fields of computer vision and machine learning, assisting them in selecting the most suitable model for their specific image classification tasks.

References

[1] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 74–92. Springer, 2022. 1

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[3] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer, 2023. 1

[4] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer, 2022. 1

[5] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist challenge 2017 dataset. *ArXiv*, abs/1707.06642, 2017. 1

[6] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer, 2022. 1

Contribution Specification

Our team members are Yuhang Song and Zhengrong Gu, both of whom made equal contributions to this project. In this project, Zhengrong primarily focused on the NAT and Di-NAT models, while Yuhang concentrated mainly on the DaViT and MaxViT models. However, many aspects of this project involved our collaborative work, including research idea generation, model performance comparison, understanding model mechanisms, and the preparation of presentations and write-ups.