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

Zhengrong Gu, Yuhang Song

Phaeolus schweinitzii

Geastrum saccatum

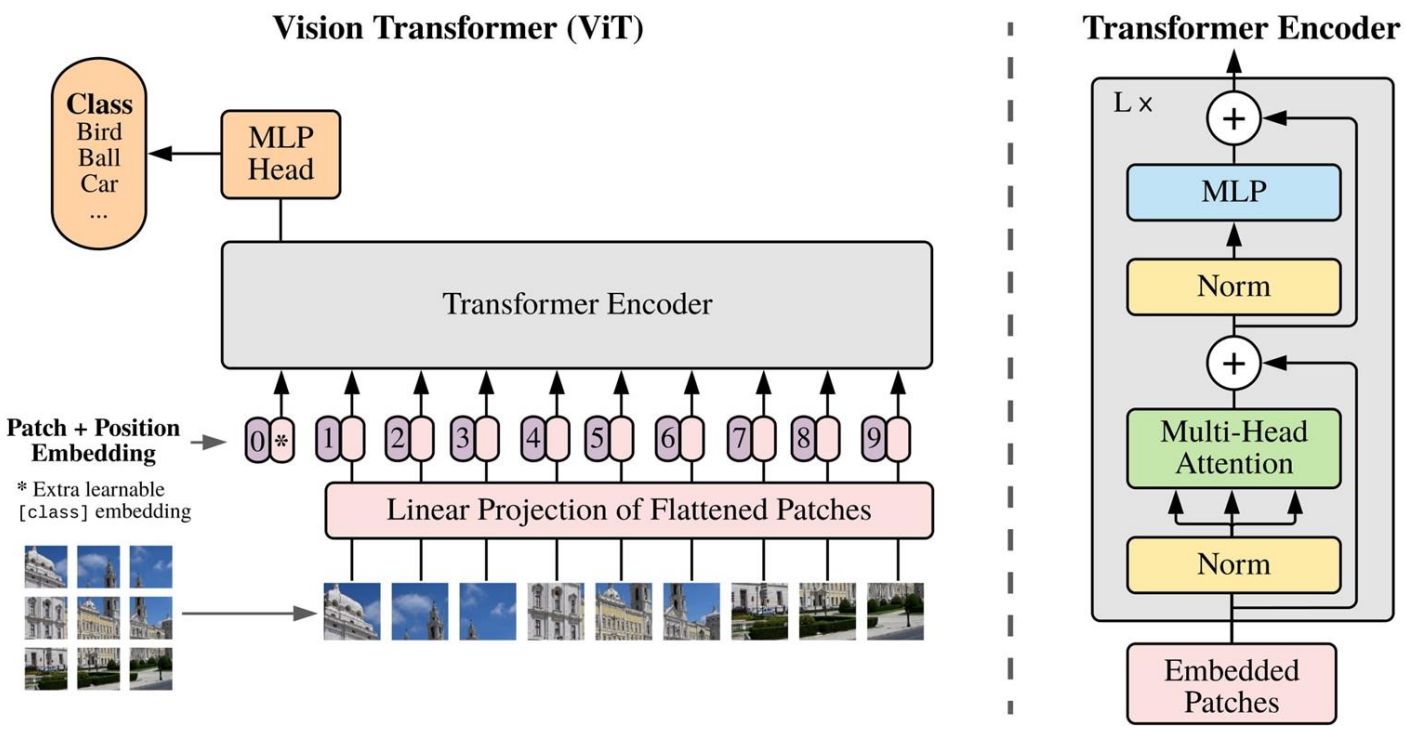
Lactarius alnicola

Introduction & Goal

- Task: image classification using transformers
- Models:
  - Neighborhood Attention Transformer (NAT)
  - Dilated Neighborhood Attention Transformer (Di-NAT)
  - Dual Attention Vision Transformers (DaViT)
  - Multi-Axis Vision Transformer (MaxViT)
- Goal:
  - demonstrate how these four models work
  - compare the differences
  - evaluate these models using the dataset from iNaturalist 2017

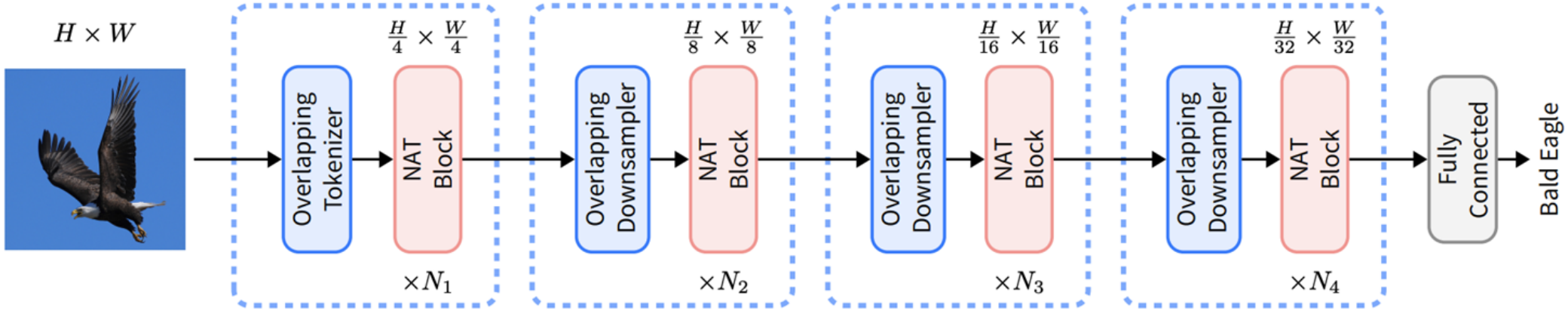
Model Comparison and Performance Insights

- All the models are based on Vision Transformer, introduced by “ An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”



Model Comparison and Performance Insights

Neighborhood Attention Transformer(NAT)

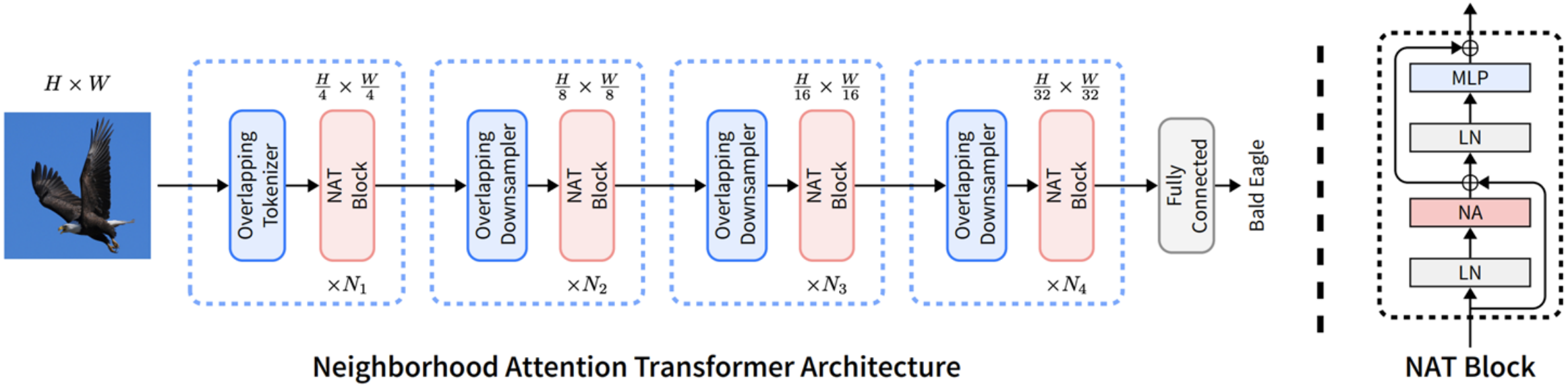


Neighborhood Attention Transformer Architecture

- *NAT is transformer based on NA*
- *a convolutional downsampler + 4 sequential levels, each consisting of multiple NAT Blocks, which are transformer-like encoder layers.*
- *Between the levels, feature maps are downsampled to half their spatial size, while their depth is doubled.*

Model Comparison and Performance Insights

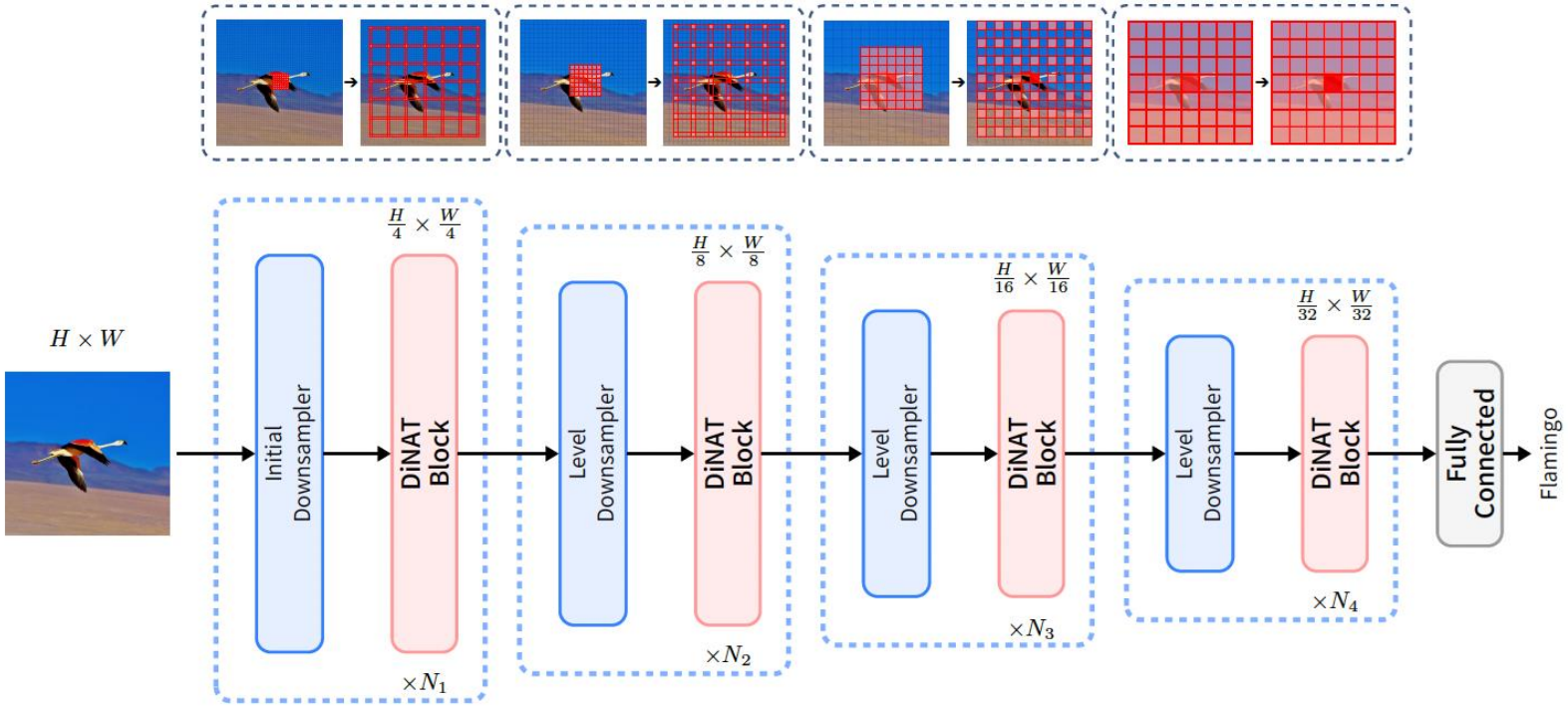
Neighborhood Attention Transformer(NAT)



- *Each layer is comprised of a multi-headed neighborhood attention, a multi-layered perceptron, Layer Norm before each module, and skip connections.*

Model Comparison and Performance Insights

Dilated Neighborhood Attention Transformer (DiNAT)

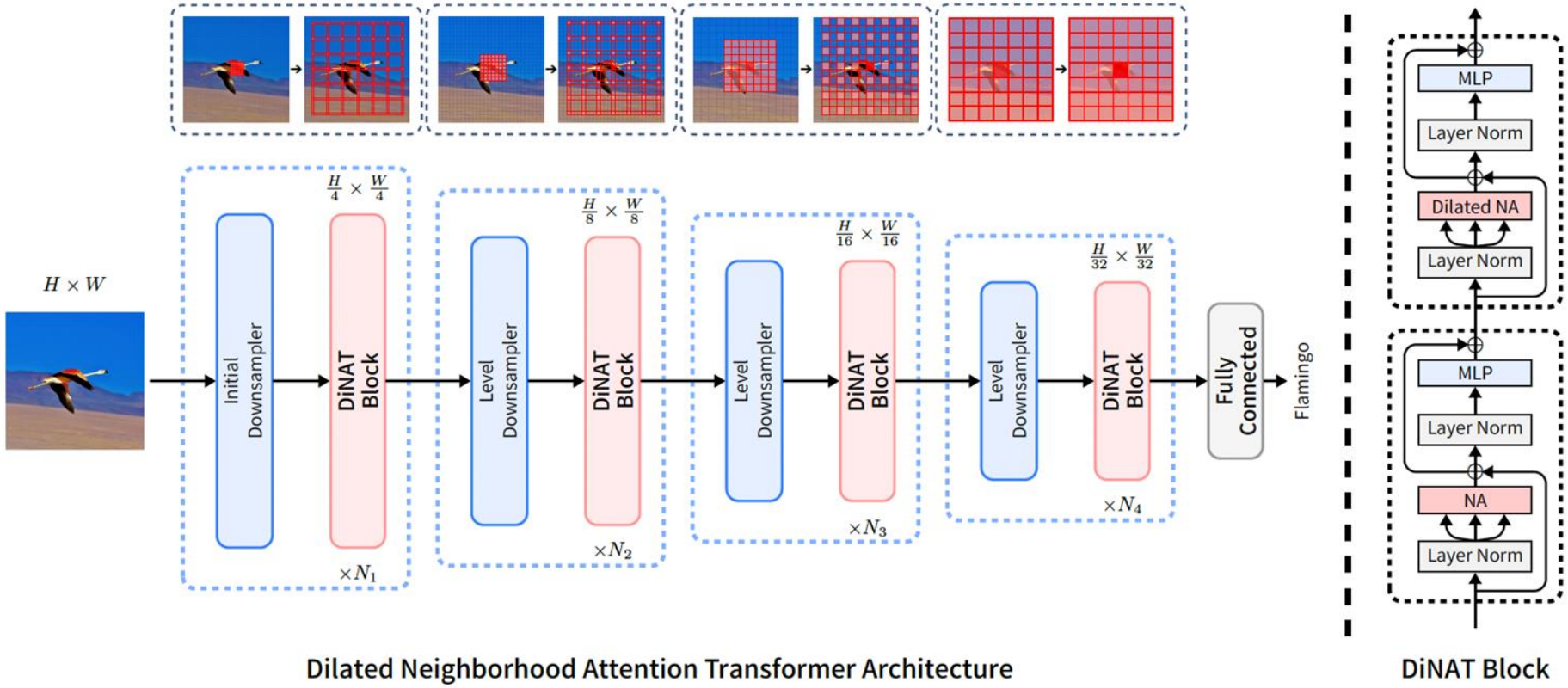


Dilated Neighborhood Attention Transformer Architecture

- *Downsamples inputs to a quarter of their original spatial resolution*
- *Sends them through 4 layers of DiNA Transformer encoders*

Model Comparison and Performance Insights

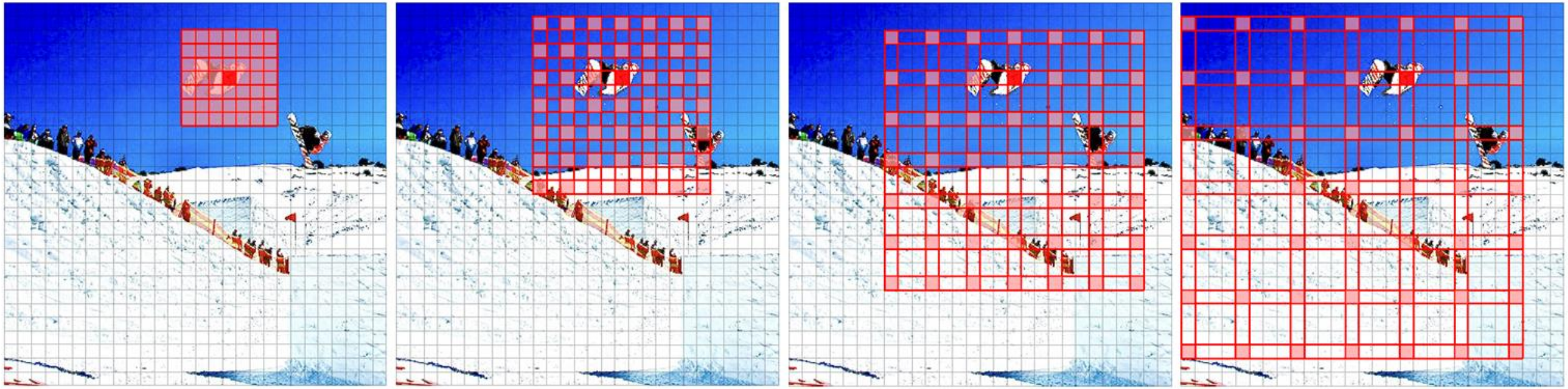
Dilated Neighborhood Attention Transformer (DiNAT)



- *DiNAT layers are similar to NAT, but it switches between local NA and sparse global DiNA at every other layer*

Model Comparison and Performance Insights

Dilated Neighborhood Attention Transformer (DiNAT)



NA / DiNA (dilation = 1) DiNA (dilation = 2) DiNA (dilation = 3) DiNA (dilation = 4)

- *An illustration of a single pixel's attention span in NA and DiNA.*
- *NA localizes attention to the pixel's nearest neighbors.*
- *DiNA extend NA's local attention to a less constrained sparse global attention.*

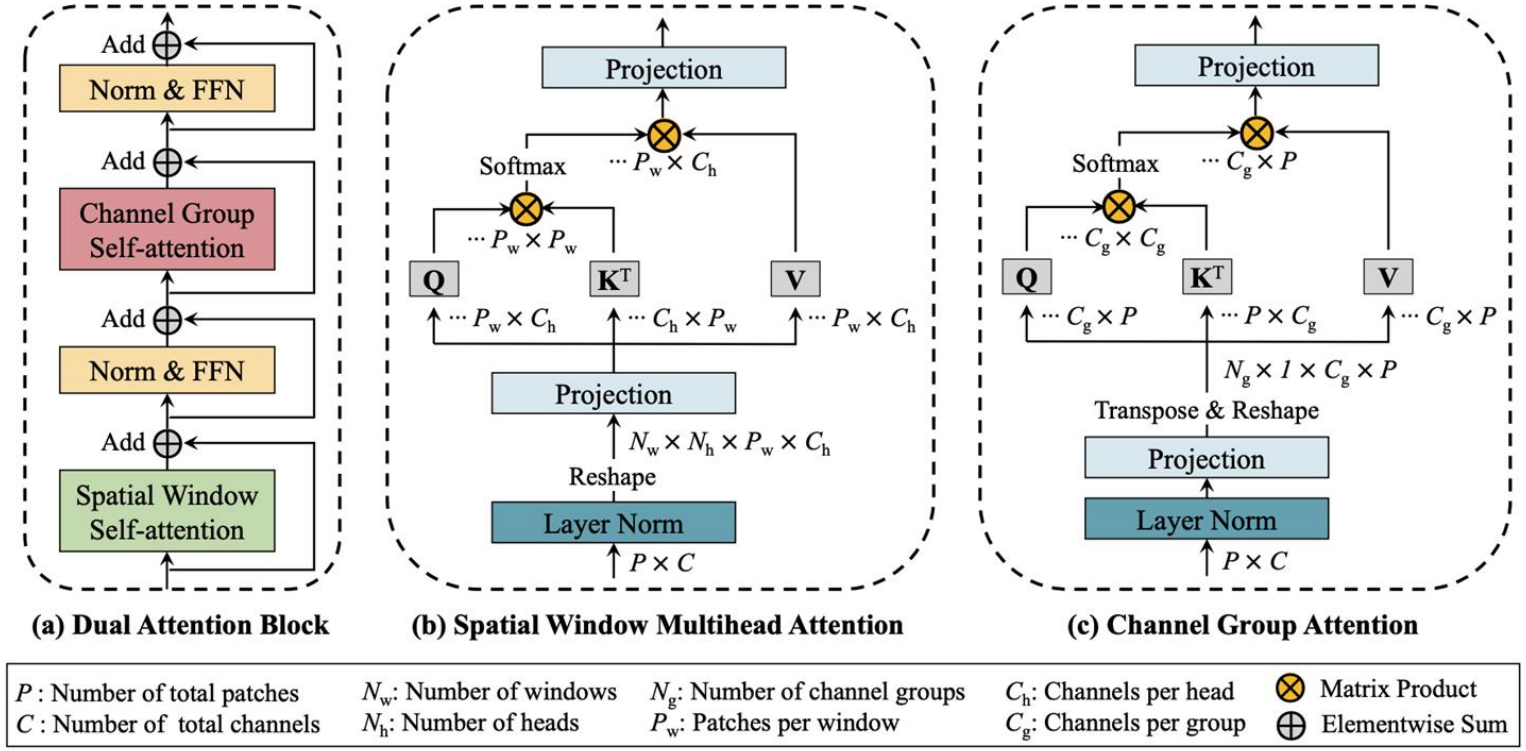
NAT vs. DiNAT

- DiNATextends NAE by introducing a dilated attention mechanism attention is applied not only to neighboring regions within a fixed spatial range but also to regions that are further away.
- The dilation factor controls the distance between regions, with larger values allowing for a wider range of attention.

Model Comparison and Performance Insights

Dual Attention Vision Transformers(DaViT)

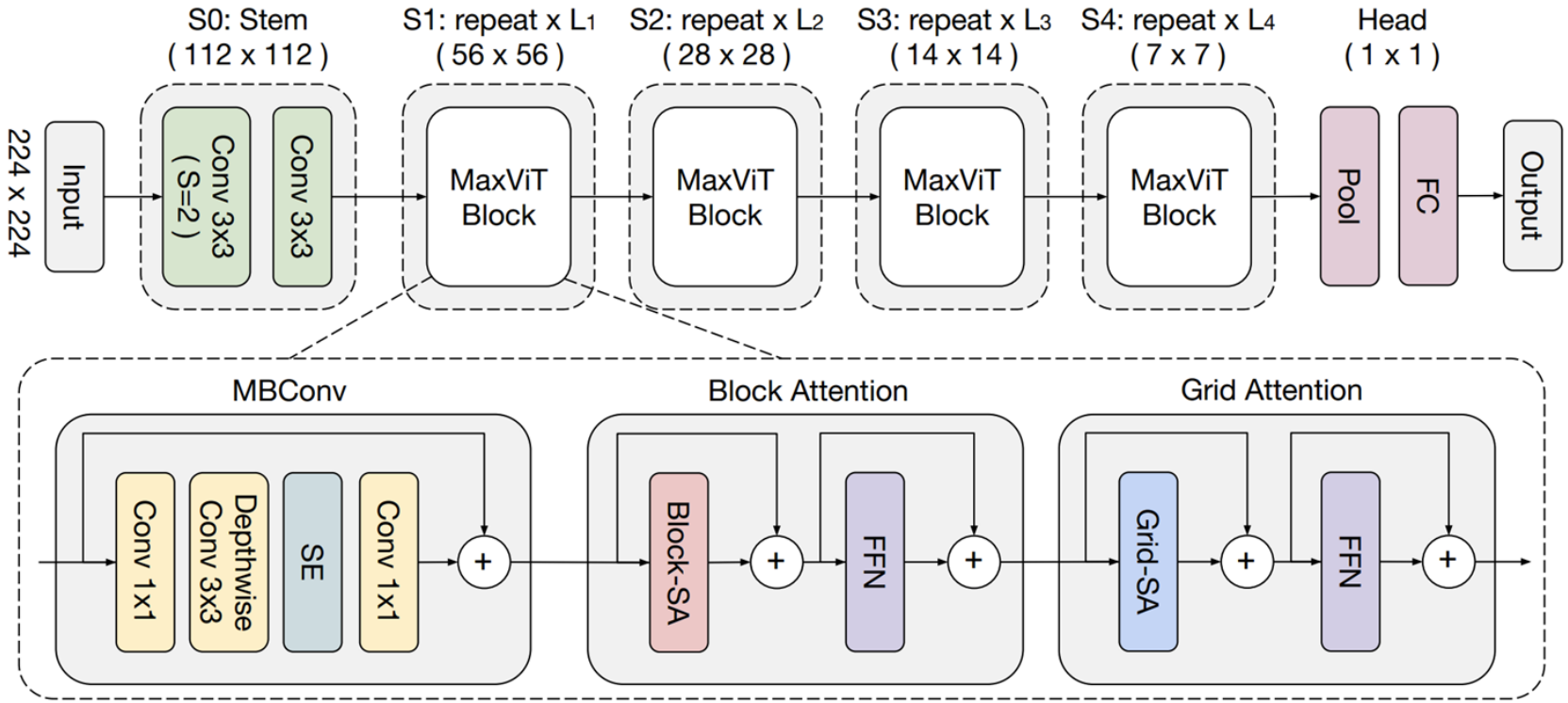
- Compared with regular ViT, DaViT has dual attention layer: Channel Group Attention and Spatial Window Multihead Attention



Model Comparison and Performance Insights

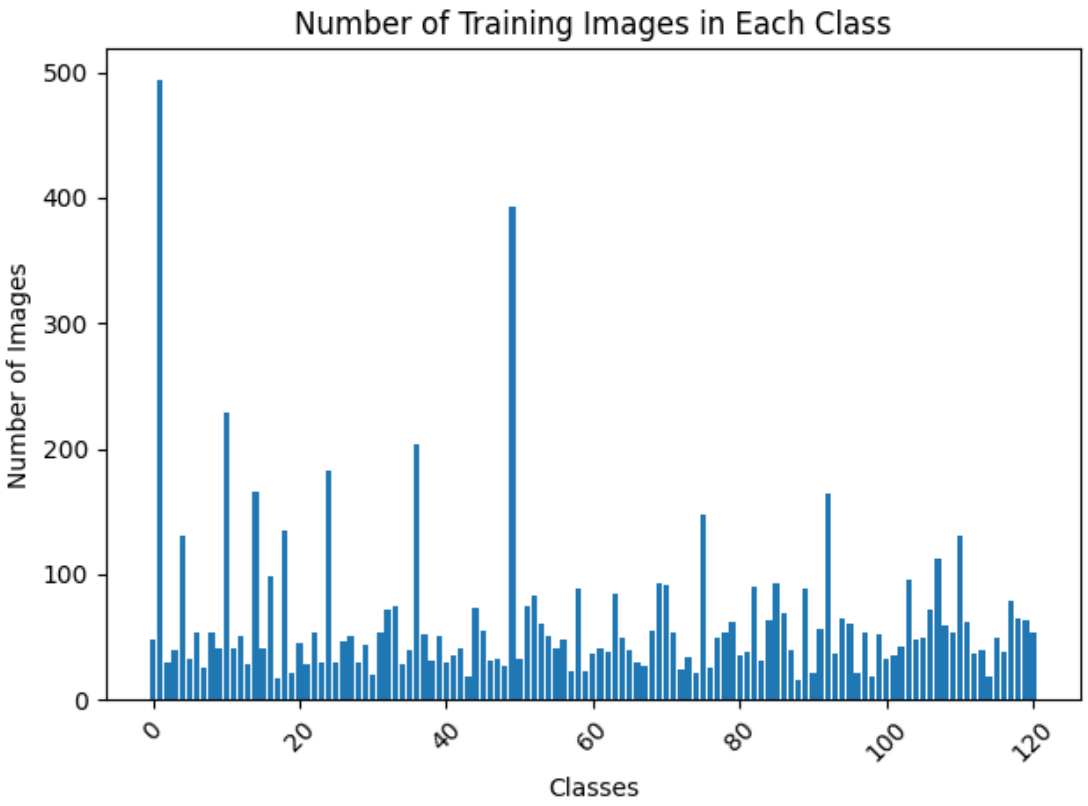
Multi-Axis Vision Transformer(MaxViT)

- MaxViT: simply repeating the basic building block over multiple stages.



iNaturalist 2017 dataset

- We utilize the fungi subset of the iNaturalist 2017 dataset
- 7,606 images, distributed across 121 classes
- Highly unbalanced data



Phaeolus schweinitzii

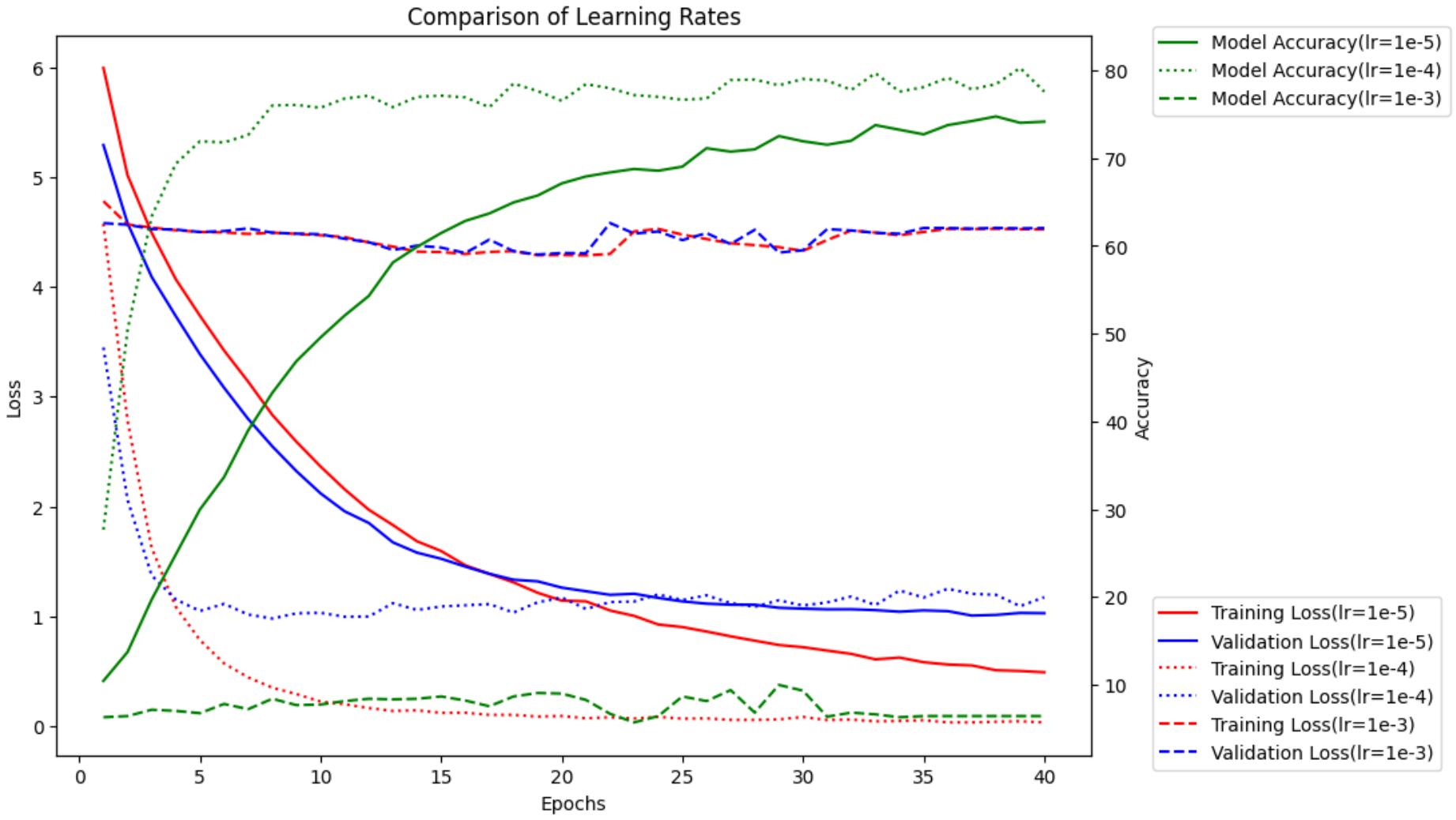
Geastrum saccatum

Lactarius alnicola

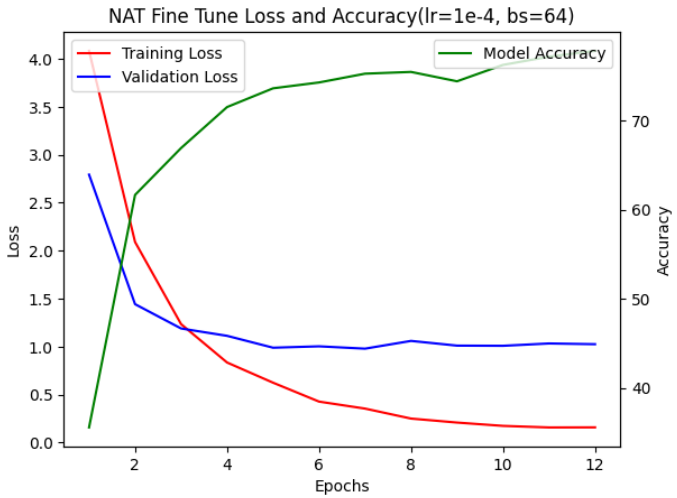
We are approaching State-of-the-Art Accuracy

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	
Our Model: Fine Tuned DaViT		**80.93%**			✓	**Fine Tune based on davit_tiny.msft_in1k**			**Today**	

Fine-Tuning Parameter Exploration



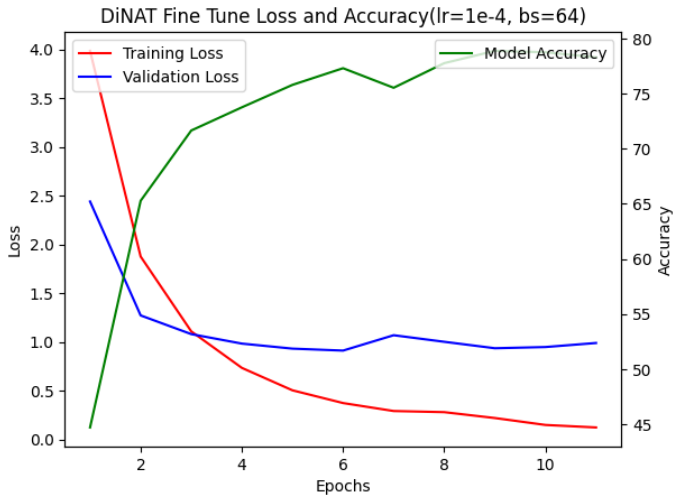
Fine-Tuning comparation on different model



Neighborhood Attention Transformer

Train based on nat-mini-in1k-224

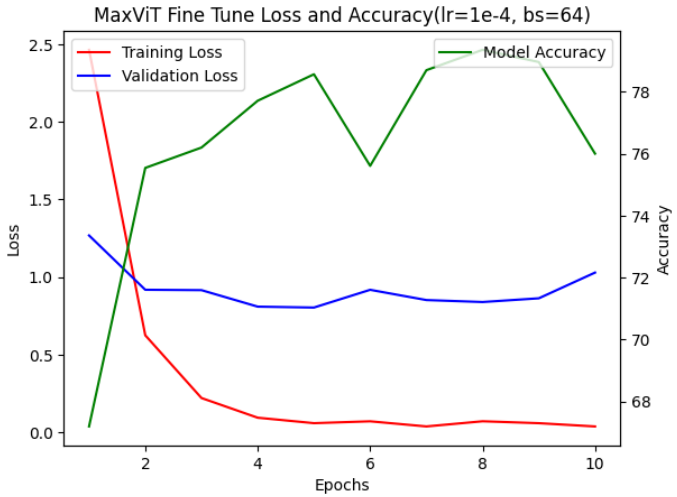
Accuracy: 75.28%



Neighborhood Attention Transformer

Train based on nat-mini-in1k-224

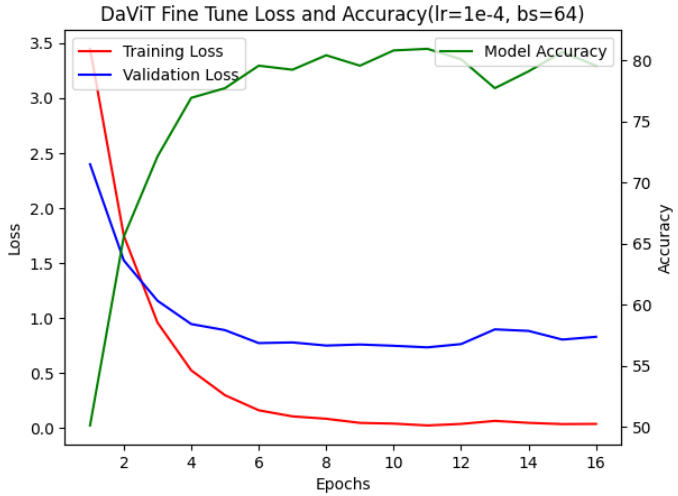
Accuracy: 77.32%



Multi-Axis ViT

Train based on maxvit_tiny_rw_224

Accuracy: 78.56%



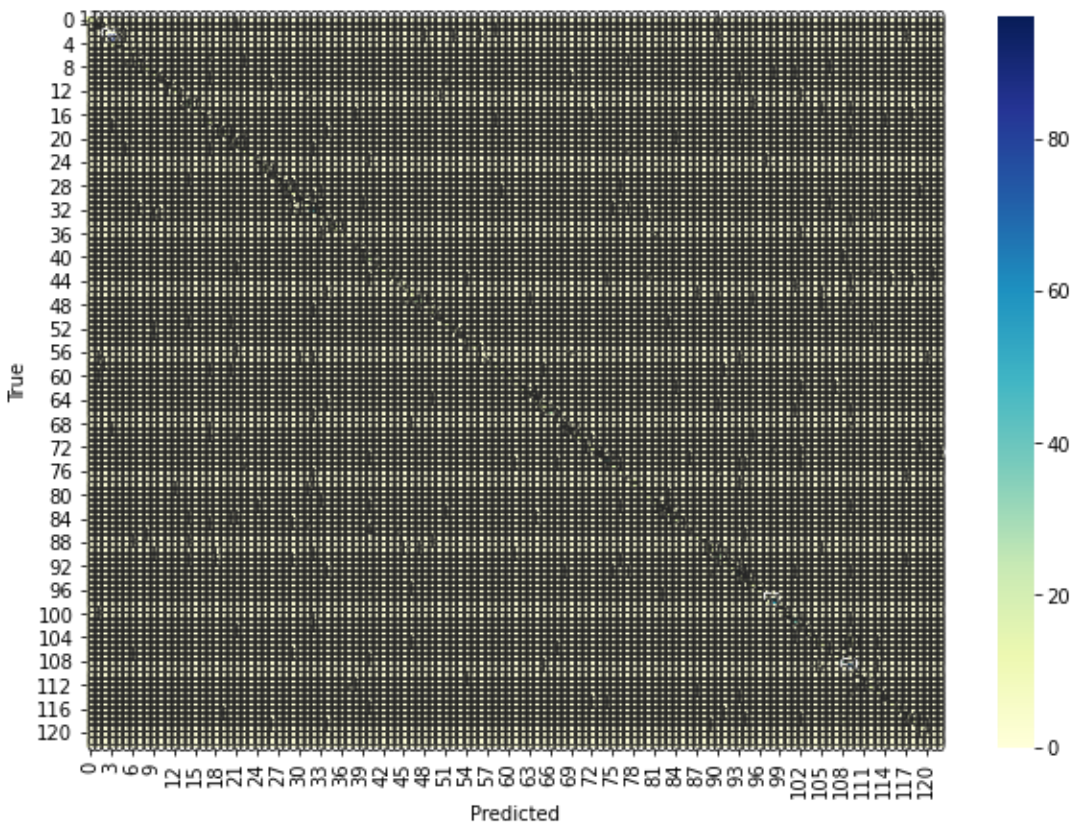
Dual Attention ViT

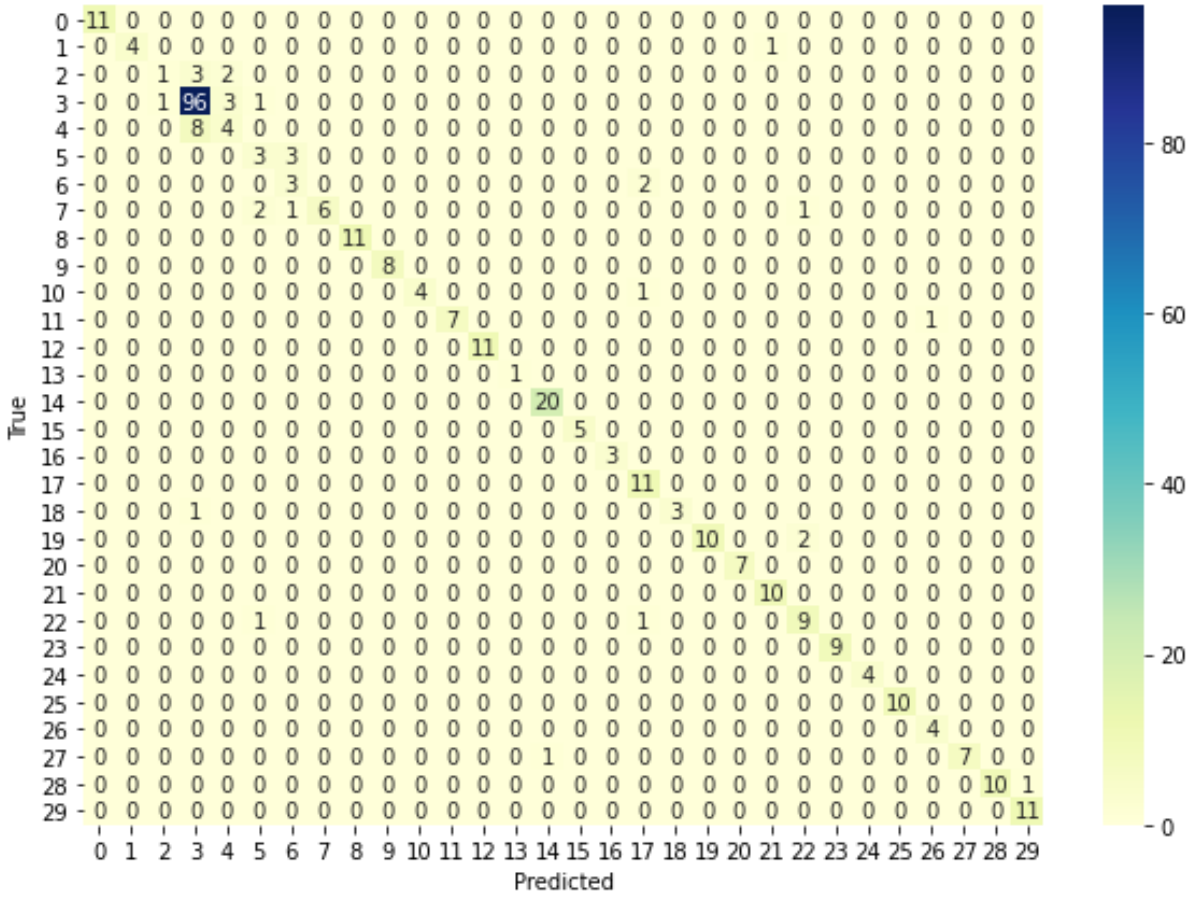
Train based on davit_tiny.msft_in1k

Accuracy: 80.93%

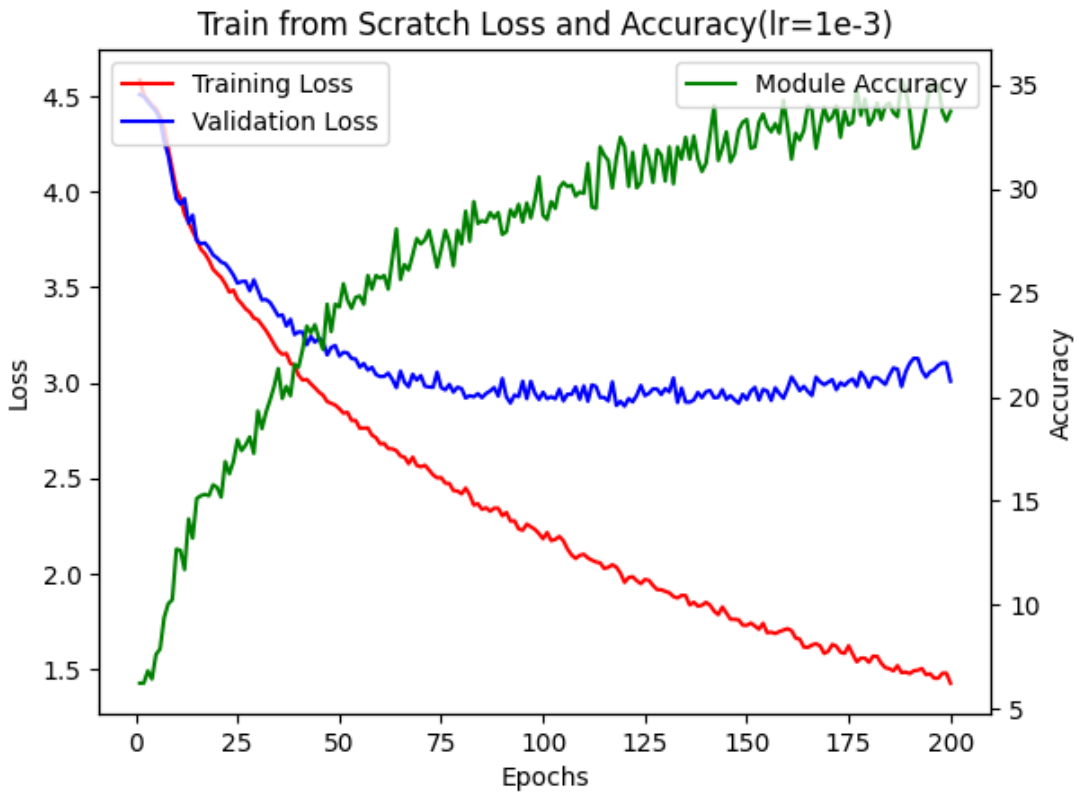
Fine-Tuned Model Accuracy

- The confusion matrix shows that there is no single category more likely to be classified incorrectly.





Training a Lightweight DiNAT from Sratch



- Training from sketch has worst performance (accuracy : 27.35%)

The plan for the remaining semester

- Optimizing the fine-tuned model by adding extra layers
- Conduct a more detailed analysis to compare the differences among four models
- Write a comprehensive report

Thank You

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

Zhengrong Gu, Yuhang Song