Scene Recognition Project

October 15, 2023

Executive Summary

Context: With the rise of image classification in the digital era, scene recognition stands out, aiming to understand the broader context of images beyond just identifying individual elements.

Significance: Scene recognition can revolutionize online platforms by enhancing their ability to interpret diverse scenes from user-contributed images.

Research Focus: Investigate the performance of different models in scene recognition tasks.

Benchmarking: Compare the Swin Transformer's performance against models like ResNet-50 and Vision Transformer (ViT).

Key Achievement: The Swin-b model, trained on Places365, achieved a Top 1 accuracy of ~58.679%, surpassing the WaveMix model's 56.45%.

Conclusion: Swin Transformer sets a promising benchmark in scene recognition, paving the way for future research in this field.

Problem statement

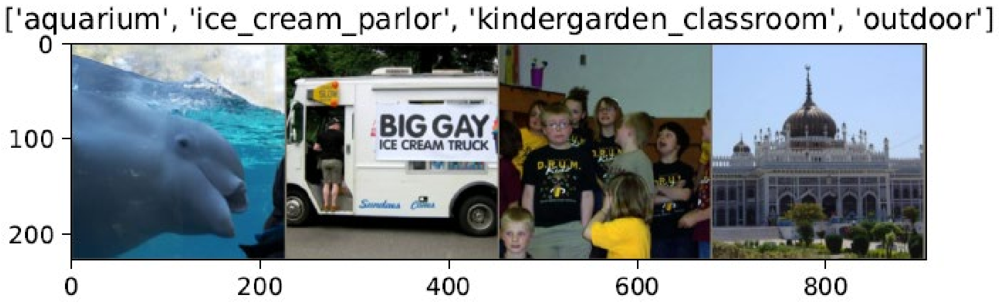
The rapid development of digital imagery in today's digital age has necessitated the discovery of advanced scene recognition techniques.

Unlike traditional image classification, scene recognition goes deeper into understanding the context and environment depicted in an image. This is crucial for applications ranging from augmented reality to surveillance with online and offline scenarios.

Image Recognition Vs Scene Recognition

ImageNet Dataset
- One object type, fails to identify others

Places365 Dataset
- Consider the whole picture as a scene



Related Work

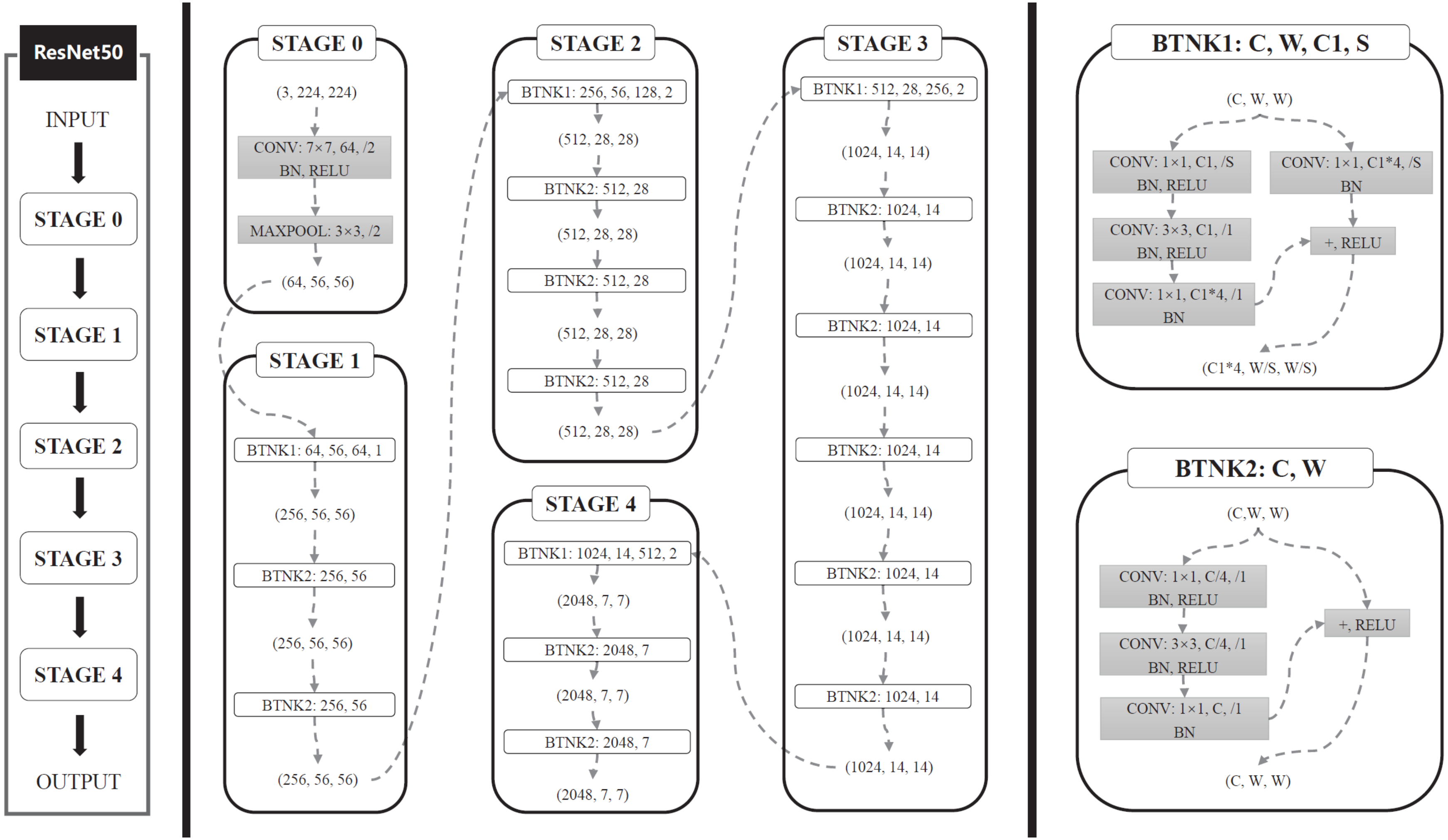
Places365 dataset:

- 1803460 training images
- 365000 validation images
- 365 classes

Convolutional Neural Networks (CNNs), especially Resnet, became the de facto standard for image-related tasks.

WaveMix achieved 56.45% on Places365-standard using this CNN approach.

ResNet-50 Architecture

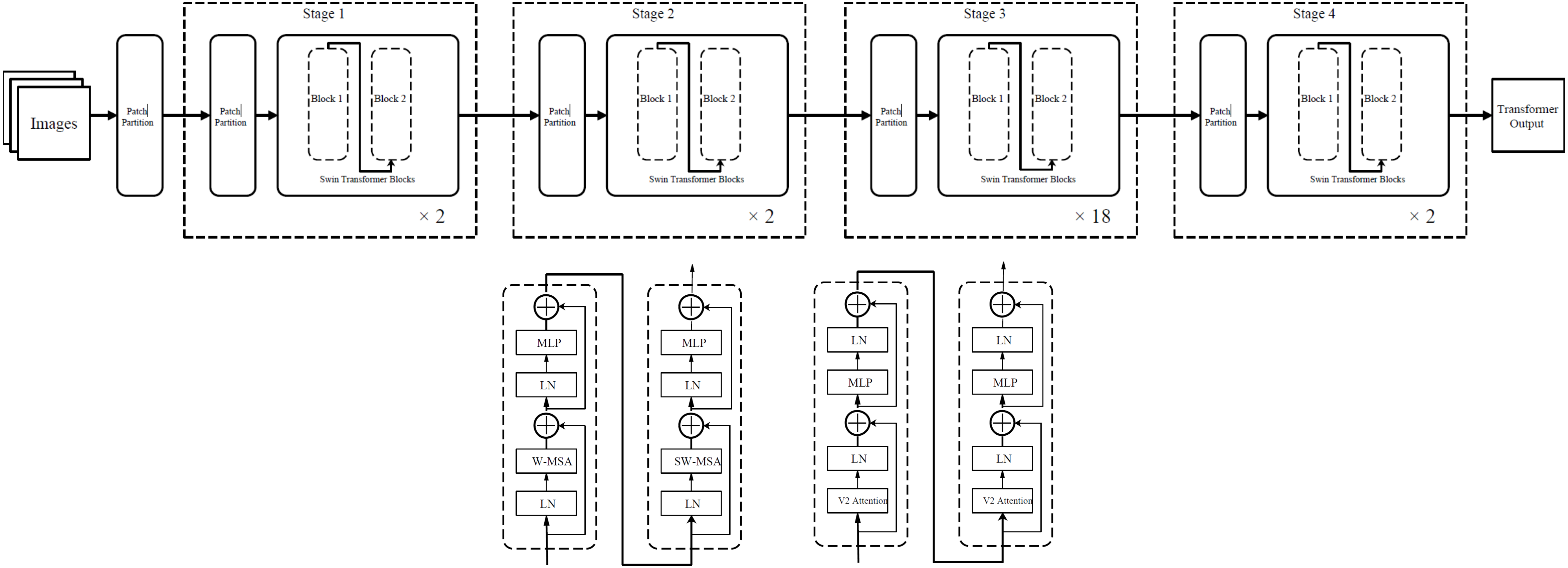


Related Work

Recently, transformer architectures, initially designed for natural language processing tasks, have shown promise in computer vision tasks, leading to the development of models like the Vision Transformer(ViT) or Swin Transformer.

InternImage achieved 61.2% on Places365 using ViT with deformable convolutions and extra training data.

Swin Transformer Architecture



(a) Swin Transformer Blocks

(b) Swin V2 Blocks

Proposed Work

- Utilizing the Swin Transformer architecture for its self-attention feature to capture global image patterns for scene recognition scenarios.
- Leveraging transfer learning: Starting with Swin Transformer v2 weights trained on ImageNet for improved performance on Places365.
- Comprehensive evaluation against models like ResNet-50, Swin-b, Swin-v2-b, and ViT to benchmark performance in scene recognition.
- Extend the training for Swin-b and try to reach maximum accuracy.

Evaluation Metrics

- **Top-1 Accuracy:**

The proportion of correctly predicted labels as the most probable of all the labels in a dataset.

- **Top-5 Accuracy:**

The percentage of correctly predicted labels within the top 5 most probable labels out of all the labels in a dataset.

- **Training Loss:**

A measure of the model's error on the training dataset; lower values indicate better model performance.

- **Validation Loss:**

A measure of the model's error on a separate validation dataset; lower values suggest better generalization ability.

- **Training Time:**

The duration taken to train the model on a specified dataset

Experiment Setup

- GPU: Nvidia RTX 4090 24GB memory
- Training Epochs: 20
- Models:

  Swin Transformer base*

  Swin Transformer v2 base**

  Resnet 50**

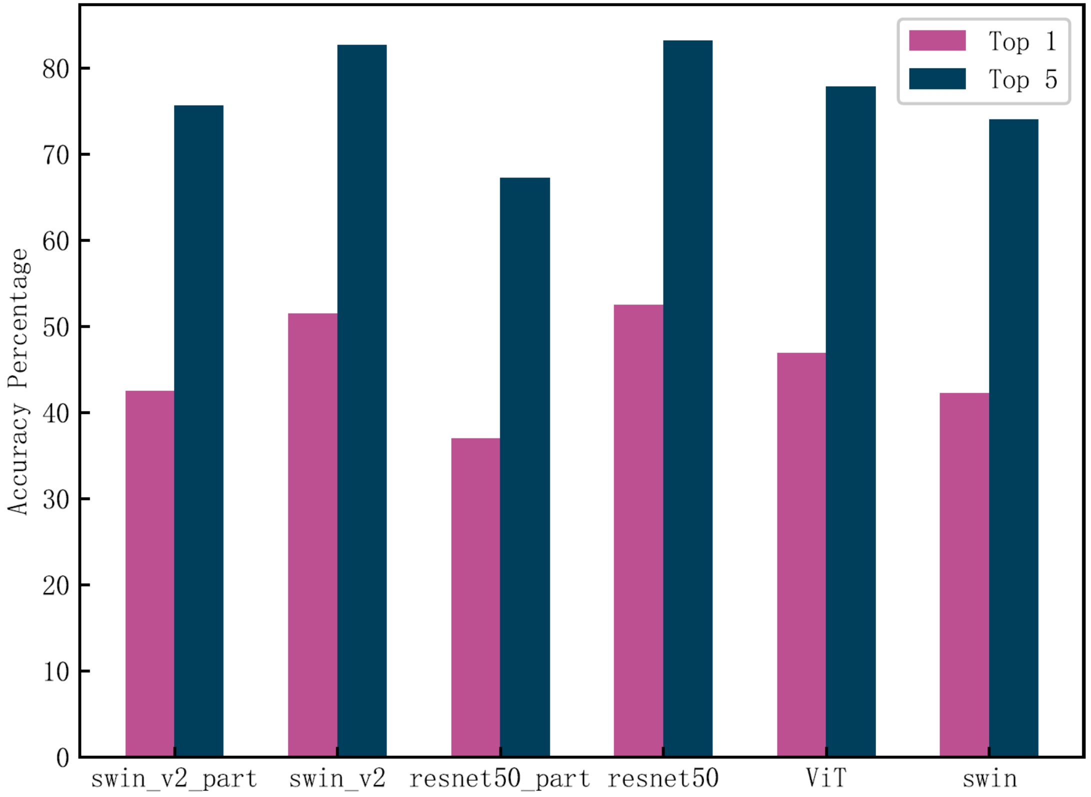
  Vision Transformer

Note: * Swin_b model has a unique training session for extended epochs (266 epochs)

**Some experiments use only partial data for training (10%) or less than 20 epochs

Accuracy across different models

- Resnet-50 outperforms every other model
- Partially trained models are underperformed
- Swin_v2 maintains a similar level of accuracy as Resnet-50

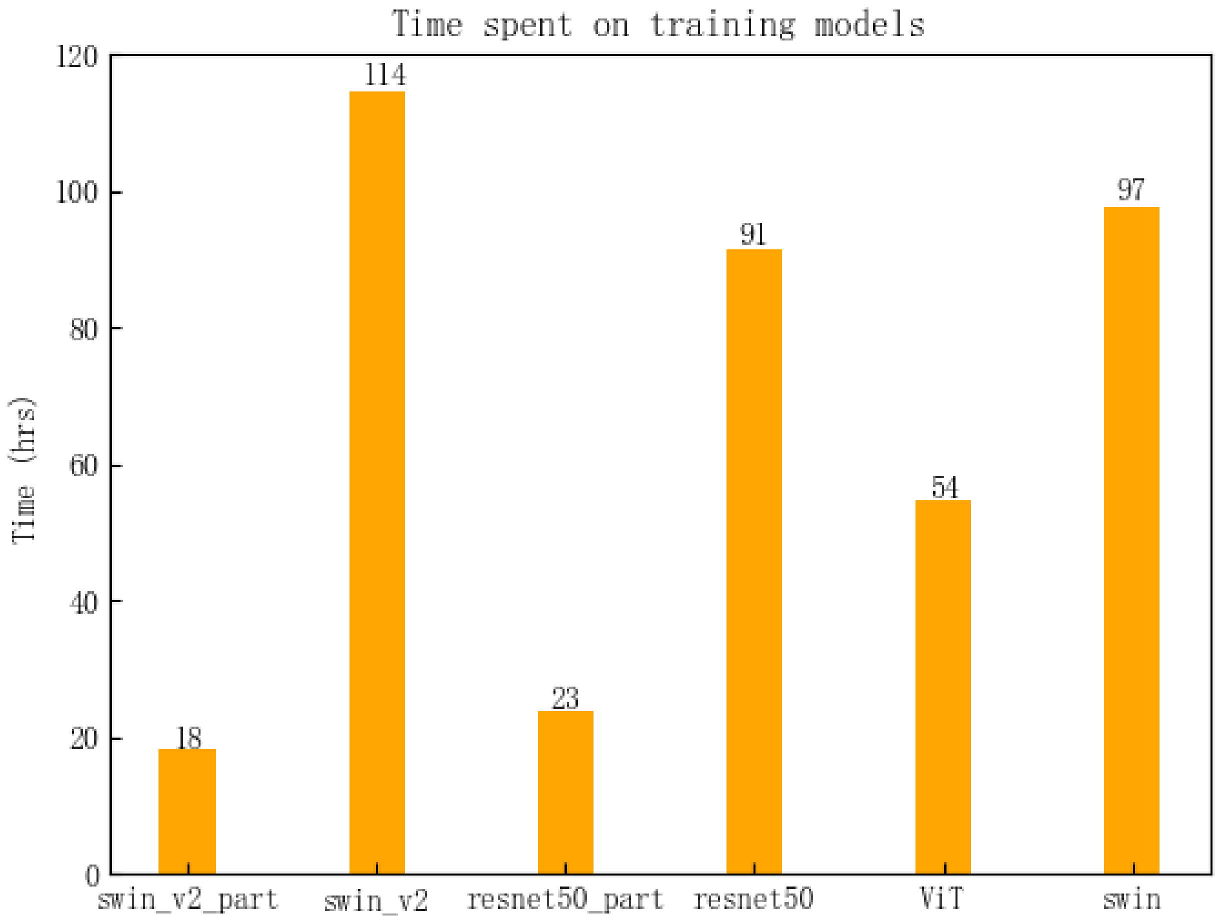


Accuracy in 20 Epochs

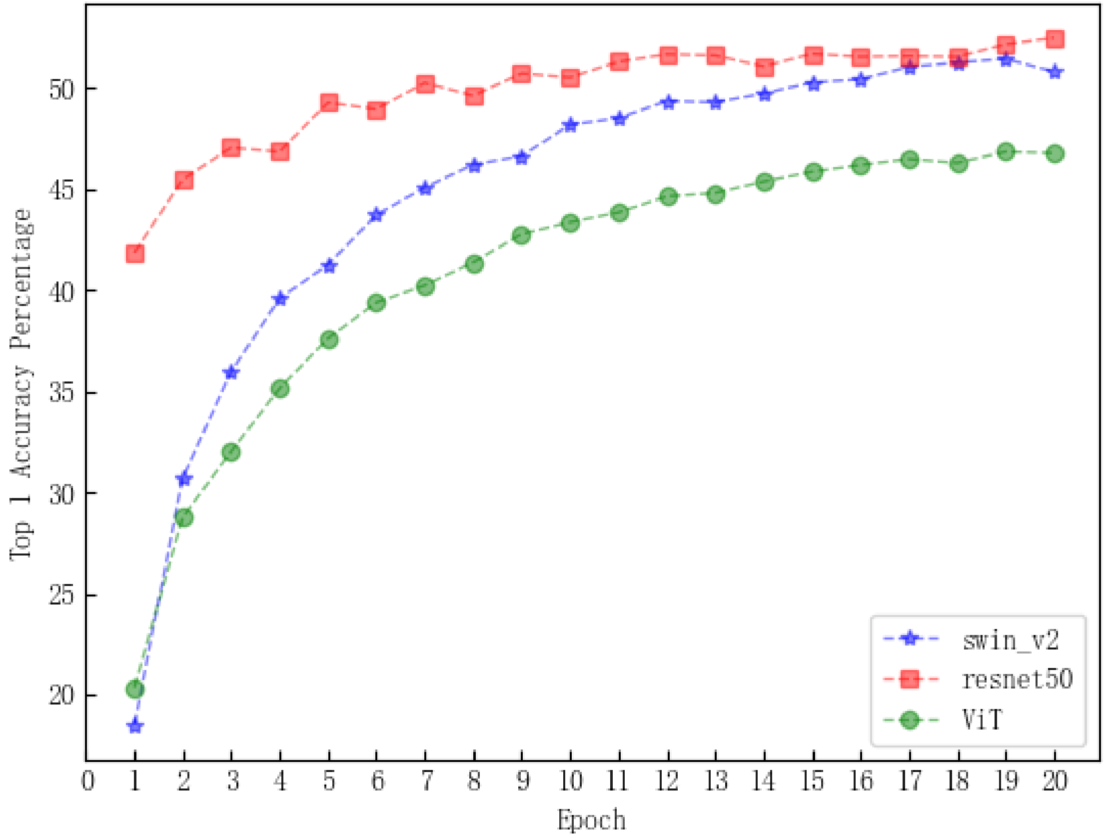
Models	Top 1 Accuracy	Top 5 Accuracy
swin_v2_part	0.42493	0.75589
swin_v2	0.51460	0.82625
resnet50_part	0.36951	0.67241
resnet50	0.52499	0.83159
ViT	0.46860	0.77825
swin	0.42203	0.74022

Training Time

- Swin v2 has the longest training hours for 20 epochs
- Partially trained models are quicker but less than 20 epochs, will be removed for further analysis
- Resnet-50 outperforms other transformer models again



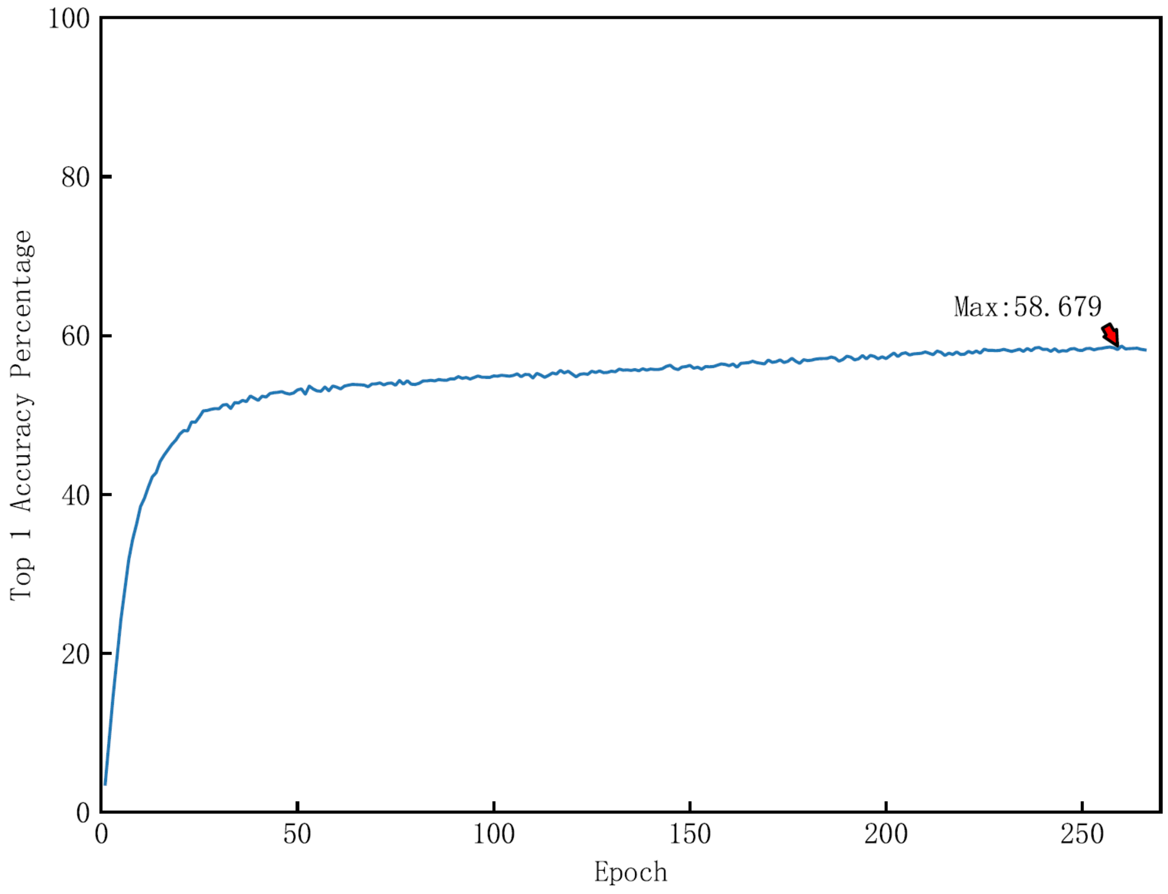
Trend Analysis



- Resnet has a faster convergence rate
- Swin_v2 needs more training epochs
- ViT is in between but has a lower accuracy

Swin Result Analysis

- Swin_b model reaches an exciting 58.679% Top 1 accuracy
- Extremely long training time, more than 14 days
- Converges in about 40 epochs, then finetune itself to over 58%



Compare with Other Researches

Models	Top 1 Accuracy	Extra training data
InternImage	61.2%	YES
Swin-b	58.679%	NO
WaveMix	56.45%	NO

Timeline

To ensure the timely completion of this research, I propose the following one-month timeline. All the timelines are finished accordingly:

- Week 1: Data Preprocessing and Initial Model Training
- Week 2: Model Tuning and Optimization
- Week 3: Evaluation and Comparison with Baseline Models
- Week 3 Update: Further training for Swin-b model
- Week 4: Final Analysis, Writing, and Submission

Thank You!

References

1. Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba.2017. Places: A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).
2. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]
3. Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, XiaogangWang, and Yu Qiao. 2023.InternImage: Exploring Large-Scale Vision Foundation
4. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. CoRR abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762
5. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. CoRR abs/2103.14030 (2021).arXiv:2103.14030 https://arxiv.org/abs/2103.14030
6. Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2021. Swin Transformer V2: Scaling Up Capacity and Resolution. CoRR abs/2111.09883 (2021). arXiv:2111.09883 https://arxiv.org/abs/2111.09883
7. 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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]
8. Pranav Jeevan, Kavitha Viswanathan, Anandu A S, and Amit Sethi. 2023. WaveMix: A Resource-efficient Neural Network for Image Analysis. arXiv:2205.14375 [cs.CV]
9. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791