# Small Datasets – CNNs vs. Transformers

## Computer Vision


## Authors

Damian Krzyżelewski
Michał Kulawiak
Jakub Wiśniowski



## Abstract

Convolutional Neural Networks (CNNs) have long been the standard for image recognition tasks due to their inductive biases such as locality and translation invariance. Recently, Visual Transformers (ViTs) have gained popularity for vision tasks, outperforming CNNs on large-scale benchmarks. However, for small datasets, ViTs often underperform due to their lack of built-in priors and higher data requirements. This poster presents a comparative analysis demonstrating that CNNs retain superior generalization and efficiency when data is limited, highlighting practical considerations for model selection in resource-constrained scenarios.


## Introduction

Arguably, the most important revolution of computer vision was the introduction of deep learning, especially Convolutional Neural Networks (CNNs). The emergence of Visual Transformers (ViTs) challenged this dominance by achieving excellent results on large datasets like ImageNet-21k. ViTs are unfortunately heavily reliant on huge training data sets. The work below tries to show why ViTs are not always as reliable as standard CNNs.

## Related Work

There have been numerous studies that have compared CNNs and ViTs. It's been shown that pre-trained transformers are more effective than standard convolutional networks (Dosovitskiy et al., 2020). In order to reduce the dependency on pre-trained models, a data-efficient transformer has been proposed (Touvron et al., 2021). Still, it is a matter of dispute whether ViTs are really more effective than CNNs (Bai et al., 2021).

## Data

In order to demonstrate the issue with ViTs, the EuroSAT dataset has been chosen for this experiment. EuroSAT is a labeled benchmark dataset for land use and land cover classification, created from Sentinel-2 satellite images provided by the European Space Agency (ESA). It was introduced to support research in remote sensing and satellite image classification.

## Methods

The aim of the experiment was to compare a representative Convolutional Neural Network with a Visual Transforer. The CNN model that was used is the ResNet-18 model. It is, the lightest model in the family of ResNet. (residual networks). The transformer model that has been used is the baseline ViT-B/16 model. Both models have been trained from scratch. Afterwards, the training time, loss and accuracy of both models is compared in order to show the differences between those models.

## Experiments

The first notable difference is the training time of both models. It is the sole reason why it is highly recommended to use pre-trained Visual Transformer. It turns out the training time of ViT-B/16 is almost 3 times higher than the training time of ResNet-18, as shown on Fig. 1. One might suspect that the longer training time of the ViT might yield better accuracy. It turns out that relatively small training image pool, it is not the case. The achieved loss of ViT-B/16 both on training and validation sets is lower than that of ResNet-18. Fig. 2 shows the achived loss during 10 epochs of training of the two models.

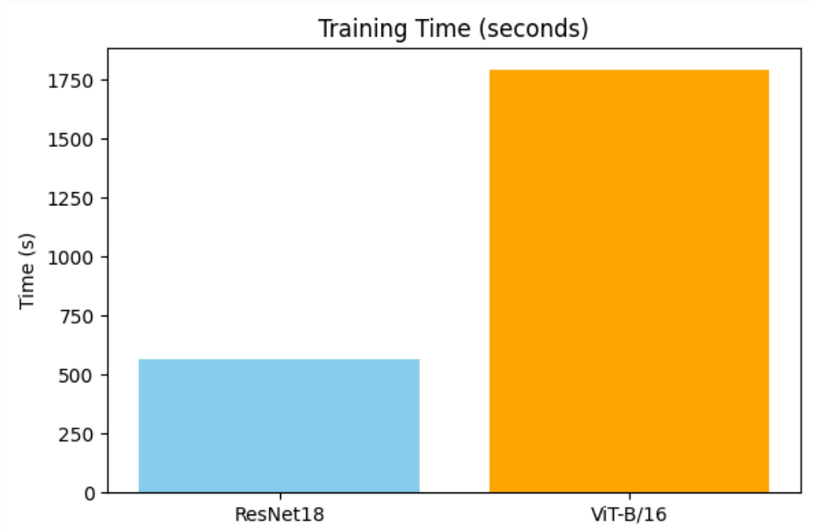


Fig. 1: Training time of ResNet-18 and ViT-B/16

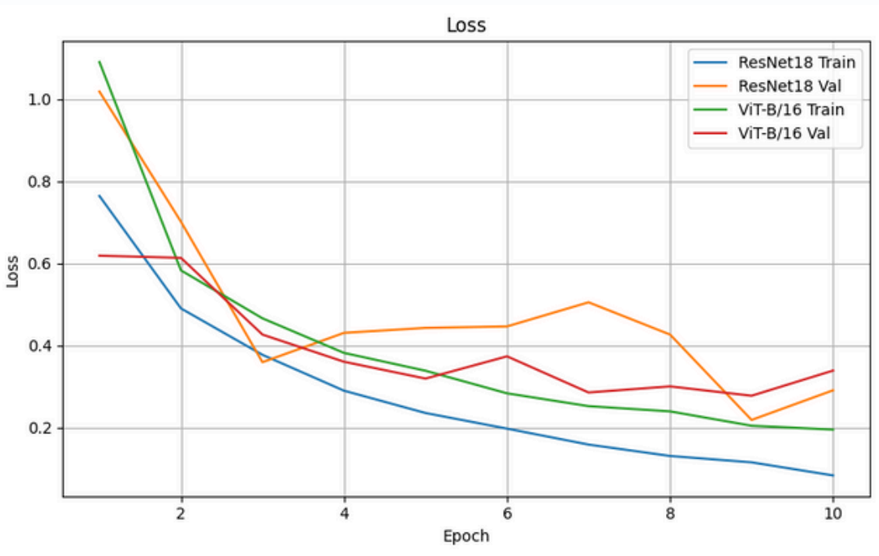


Fig. 2: Loss of ResNet-18 and ViT-B/16 during first 10 training epochs

Overall, to sum up the achieved result, the validation set accuracy of those two models is compared to their respective training times. Fig. 3. shows the achieved accuracy after specified time. As expected, the ResNet managed to achieved greater accuracy in shorter time.

The ResNet-18 model managed to achieve much greater accuracy almost immediately in the first few epochs. It already achieved 88% accuracy even before ViT finished its first training epoch. The transformer achieved comparable accuracy after being train for 3 times as long as the CNN model.

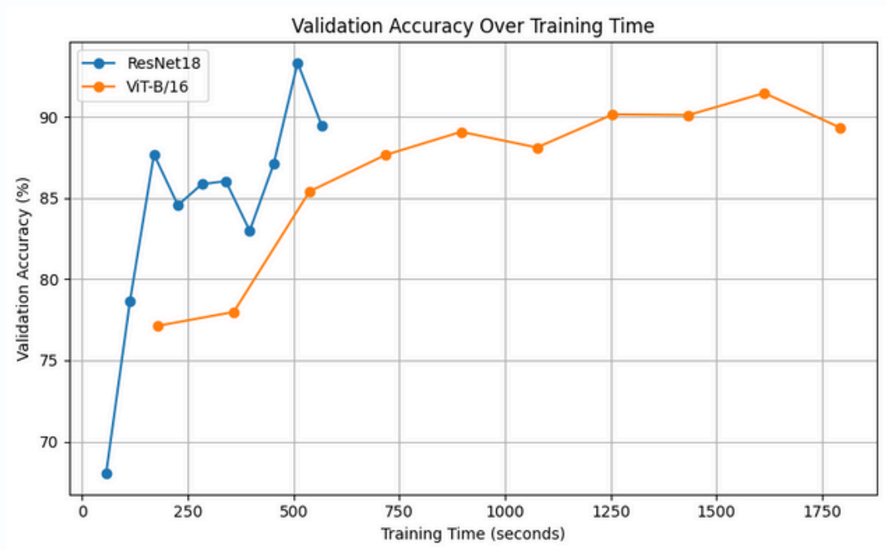


Fig. 3: Validation accuracy of ResNet-18 and ViT-B/16 over time

## Conclusion

To summarize, the ViT-B/16 model proved to be much less effective. Even after being trained for a longer period of time, it yielded comparable results to ResNet-18 (which is a much smaller CNN model). Without pre-training and with a small amount of training data, the Visual Transformer is just not worth the time to train as the CNN can achive a satisfying level of accuracy.

## Supplementary material

The code and additional plots are available at https://github.com/mixing01/Computer-Vision for further analysis. Additional classification tests have also been performed to compare the actual results. To sum those up, the confusion matrices for both models have been created.

## Literature


- **Dosovitskiy, A., et al. (2020).** *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.*
- **Touvron, H., et al. (2021).** *Training data-efficient image transformers & distillation through attention.*
- ***Seung Hoon Lee, Seunghyun Lee, Byung Cheol Song (2021).*** *Vision Transformer for Small-Size Datasets*
- **Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015)**. *Deep residual learning for image recognition.*
- *PyTorch documentation - ResNet [access 20.06.2025]* https://pytorch.org/hub/pytorch_vision_resnet/
- Vision Transformer documentation [access 20.06.2025] https://pprp.github.io/timm/models/vision-transformer/