Title: A ConvNet for the 2020s

Authors: Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie

Publication: CVPR 2022

Link: https://arxiv.org/abs/2201.03545

Summary:

In the paper, the authors seek to modernize Convolutional Neural Networks (ConvNets) by borrowing training techniques and architectural modifications from the Transformer models that have dominated the field in recent years. Through a series of six stages, they transform a ResNet50 model into a ConvNeXt architecture, which can compete with Transformers in various tasks, including image classification, object detection, and semantic segmentation.

The modernization stages include:

- Adopting similar training techniques as used in Transformers, such as the AdamW optimizer, heavy data augmentation, and regularization.
- Adjusting the macro design of ResNet by modifying the number of blocks in each stage and altering the sliding windows to resemble patch-based Transformers.
- Incorporating depthwise convolutions and widening the network.
- Implementing inverted bottlenecks with an expansion ratio of 4.
- Introducing larger kernel sizes to mimic the global receptive field of Transformers.
- Changing micro design choices, such as replacing ReLU with GELU, using fewer normalization layers, and replacing Batch Normalization with Layer Normalization.

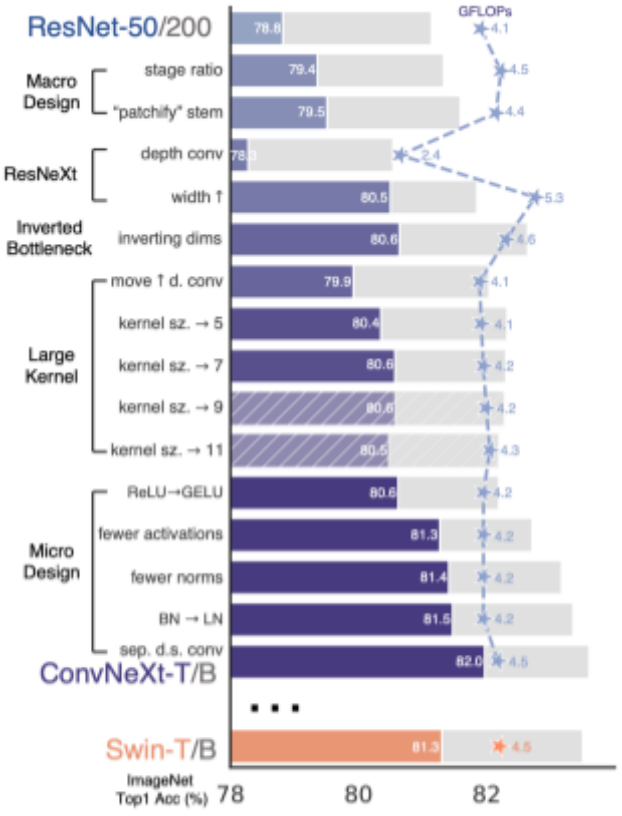

Fig 1: Modernization stages leading to ConvNeXt

The experiments demonstrate that modernized ConvNets can achieve competitive performance with Swin Transformers while offering potential lessons for other architectures. The paper suggests that optimizing hyperparameters and small architectural choices can significantly impact the performance of a model, rather than solely relying on a specific architecture.

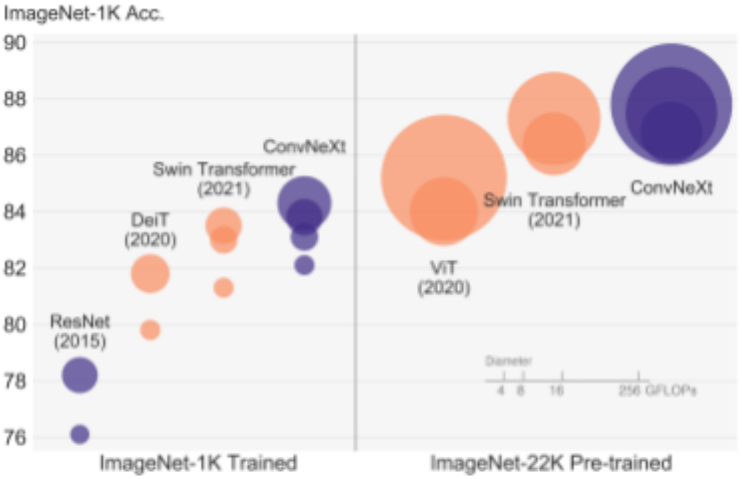

Fig 2: Plot of accuracies and GFLOPS of different transformers and convolution architectures

Why this publication was interesting:

- Bridging the gap between ConvNets and Transformers: The study showcases that by adopting techniques and architectural modifications from Transformers, it is possible to significantly improve the performance of ConvNets, narrowing the gap between the two types of models in various computer vision tasks.

- Potential for hybrid architectures: The successful integration of Transformer-inspired techniques into ConvNets could lead to the development of more powerful hybrid architectures that benefit from the strengths of both approaches.

- Significance of hyperparameters and design decisions: The paper emphasizes the value of adhering to conventional best practices like fine-tuning hyperparameters and implementing minor architectural adjustments to boost a model's performance, contesting the idea that success solely relies on a particular architecture.

- Renewed interest in ConvNets: This research could potentially spark renewed interest in ConvNets as a viable competitor to Transformers for computer vision tasks, leading to further exploration and innovation in the field.