Image Classification Models: Vision Transformers vs. ConvNeXt

In the fast-changing world of computer vision, two powerful ways to build AI have become popular, making image classification better than ever. This guide compares the groundbreaking Vision Transformers with the carefully updated ConvNeXt, looking at their main ideas and how they compete.

Vision Transformers (ViT)

This new approach changes how we classify images by breaking them into small pieces, like words in a sentence. This helps the AI understand how different parts of an image connect, using an "**attention**" feature that's common in understanding text, but now applied to image recognition.

ConvNeXt (Modernized CNNs)

ConvNeXt is a careful **update** of the well-known **ResNet**, a traditional AI method. It uses important ideas from Transformers, like using **bigger processing areas** and certain **decision-making steps**. This shows that even traditional AI methods (Convolutional Neural Networks, or CNNs) are still very good compared to newer approaches.

Made By: Mohamed Hussein Ragab

Vision Transformers (ViT)

Paper: *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (Dosovitskiy et al.)

📊 Key Performance Highlights

88.55%

ImageNet Accuracy

Achieved on a huge dataset, demonstrating ViT's strong ability to handle more data and complexity.

94.55%

CIFAR-100 Accuracy

Shows impressive results across various image recognition challenges.

77.63%

VTAB Score

Performs better than older image recognition models on many different tasks, proving its versatility.

💡 Vision Transformer's Core Innovations

Breaking Images into "Words"

Images are divided into small, flat squares, much like words in a sentence. This allows the model to understand the whole picture.

Focusing on Important Parts

Instead of traditional convolutional layers, ViT uses a "self-attention" mechanism. This lets it intelligently focus on key sections of an image, catching connections across long distances.

Learning from Lots of Data

ViT thrives on massive amounts of data for training. This helps it overcome the lack of built-in assumptions about how images work (which older models have) and achieve top results on very large datasets.

📐 How it Works

- **Image Preparation:** Images are chopped into fixed-size squares (e.g., 16×16 pixels). Each square is then flattened into a single line of data, converted into a format the model understands, and fed into the Transformer's main processing unit.
- **Classification:** A special "class token" is added to the start of the sequence. Its final processed output is then used to decide what the image contains.
- **Location Information:** Standard learnable pieces of information are added to each square to tell the model where it is in the image.
- **Variations:** There are different sizes of ViT models, like ViT-Base, ViT-Large, and ViT-Huge, similar to how other large language models are scaled.

📊 Test Results

- **Datasets Used:** Models were trained on ImageNet-21k (14 million images) and JFT-300M (303 million images).
- **Comparison:** ViT performed better than other advanced image recognition models like ResNet (Big Transfer) and EfficientNet, all while needing less computing power.
- **Scaling:** Larger ViT models only show their strength with very large datasets. For smaller datasets, older models that have built-in assumptions about images still work better.

🔍 In-depth Look

- **How it Focuses:** Even in the early stages, some parts of the model look at the entire image, while others focus on smaller, localized areas—similar to how older models "see" an image.
- **Efficiency:** ViT models use less computer memory and get better results more efficiently as you add more computing power compared to ResNet models.
- **Combined Models:** Mixing traditional image models with ViT can help when dealing with smaller image sizes, but this doesn't offer much benefit when working with very large datasets.

🚀 What's Next

- **Beyond Simple Recognition:** Applying ViT to more complex vision tasks like finding objects within an image or outlining their shapes.
- **Learning Without Labels:** Early tests of predicting missing parts of an image (without being told what they are) show promise, though it's not yet as good as training with explicit labels.
- **Growth:** Even larger models and more training data are likely to lead to even better results.

ConvNeXt (Modernized CNNs)

Paper: *A ConvNet for the 2020s* (Liu et al.) / **Update:** *ConvNeXt V2* (Woo et al., 2023)

📊 Key Performance Highlights

87.8%

ImageNet-1K Accuracy

Achieved top performance in recognizing images, showing it can match powerful Transformer models.

49% Faster

GPU Throughput

Processes information much quicker than Swin Transformers on powerful graphics cards.

Competitive

Benchmark Leadership

Performs as well as, or better than, Swin Transformers in tasks like finding objects and understanding image regions.

🔑 Core Idea: Traditional Image Networks (CNNs) Fight Back

- Newer Vision Transformers (ViTs) became very popular for classifying images, but they sometimes struggle with tasks like finding specific objects or understanding different parts of an image without some built-in assumptions about how images work.
- Swin Transformers showed that core ideas from traditional image networks, like looking at local areas and building up a layered view, are still important.
- The researchers improved a well-known traditional image network called ResNet, step-by-step, by borrowing successful design ideas from Transformer models. They wanted to see how powerful a network based purely on traditional image processing could become.
- The result is **ConvNeXt**, a family of traditional image networks that performs just as well as, or even better than, Transformer models in terms of accuracy, how well they can grow with more data, and how efficiently they run.

🛠️ Modernization Roadmap: Turning ResNet into ConvNeXt

01

Advanced Training

Adopted sophisticated training methods, similar to those used for Transformers (e.g., AdamW, Mixup, CutMix, RandAugment, and training for a longer time).

02

Overall Design Changes

Adjusted how different parts of the network were scaled and replaced the initial processing layer with one that breaks images into smaller blocks, similar to how Transformers handle data.

03

Efficient Layer Design

Used an efficient type of convolutional layer, common in ResNeXt models, and made the network wider to handle more information.

04

Smarter Bottleneck

Used a design trick, similar to MobileNetV2, that makes the network more efficient by processing information in a clever way.

05

Large Processing Areas (7×7)

Used larger filters that can look at a bigger part of the image, helping the network to understand global patterns, much like how "self-attention" works in Transformers.

06

Small Adjustments

Switched some internal functions (ReLU to GELU), simplified how information flows, and changed how data is standardized within layers (BatchNorm to LayerNorm).

07

Separate Size Reduction

Improved stability and accuracy by using dedicated layers to reduce the image size at certain points in the network.

📌 Key Takeaways

Traditional Networks Are Still Powerful

ConvNeXt shows that networks relying on traditional image processing methods can compete with Transformers, even without their "attention" components.

Simple Yet Modern Design

Achieves top performance by using a straightforward structure combined with the latest training techniques.

Grows With Data

It can effectively improve its performance when trained on more and more data, challenging the idea that only Transformers excel in this area.