Summarizing and Analyzing Vision Transformer

Overview of the Vision Transformer Architecture

The Vision Transformer (ViT) is a deep learning model that adapts the transformer architecture—originally developed for natural language processing—to computer vision tasks. Unlike traditional convolutional neural networks (CNNs), ViT processes images by dividing them into fixed-size patches, treating each patch as a token (analogous to words in NLP), and feeding these tokens into a transformer encoder. This approach enables ViT to model long-range dependencies and global relationships within images more effectively than CNNs, which primarily focus on local features.

Core Components

Patch Embedding

The input image is split into non-overlapping patches (e.g., 16x16 pixels). Each patch is flattened and linearly projected into a higher-dimensional embedding space, forming a sequence of patch embeddings.

Positional Encoding

Since transformers lack inherent spatial awareness, positional encodings are added to patch embeddings to retain information about the spatial arrangement of patches within the image.

Transformer Encoder

The sequence of patch embeddings (with positional encodings) is processed by multiple transformer encoder layers. Each encoder layer consists of:

- Multi-head self-attention: Allows the model to focus on different parts of the image simultaneously.
- Feed-forward network: Applies non-linear transformations to the outputs of the attention mechanism.
- Layer normalization and residual connections are used to stabilize and enhance training.

Classification Head

A special classification token (CLS token) is prepended to the sequence. After passing through the transformer encoders, the output corresponding to the CLS token is used for final classification via a multi-layer perceptron (MLP) head.

Comparison with CNN-Based Models (ResNet/SEResNet)

Aspect	Vision Transformer	CNN (SEResNet/ResNet)
Feature Extraction	Global (via self-attention)	Local (via convolution)
Inductive Bias	Minimal	Strong (translation equivariance, locality)
Data Efficiency	Requires large datasets	Performs well on smaller datasets
Scalability	Highly scalable	Scalability limited by local receptive fields
Performance	Outperforms CNNs on large datasets	Often superior on small/medium datasets

- ViTs outperform CNNs in accuracy and robustness on large datasets and are more consistent in handling occlusions and shape variations.
- CNNs, with their inductive biases, are easier to train and generalize better on smaller datasets.

ViT Struggles and Advantage in Vision Tasks

- **Why performs well?**: Self-attention enables ViT to capture relationships across the entire image, beneficial for complex tasks. ViT models scale well with increased data and computational resources, achieving state-of-the-art results on large benchmarks. The architecture can be adapted for various vision tasks beyond classification, such as segmentation and detection.
- **Why it struggles?**: ViTs require massive datasets for effective training due to the lack of strong inductive biases, struggling on small or medium-sized datasets without pretraining. More sensitive to hyperparameters and optimizer choices compared to CNNs, making training less straightforward. Understanding what ViTs learn is more challenging, as their attention patterns are less intuitive than CNN feature maps.

Insights from the Original Paper ("An Image is Worth 16x16 Words")

The original ViT paper demonstrated that, when pre-trained on very large datasets (e.g., JFT-300M), ViT models can match or exceed the performance of CNNs like ResNet on image classification tasks. The lack of inductive bias (e.g., translation equivariance) is compensated by massive data and model scale, showing that large-scale training can overcome architectural limitations. ViT achieves high accuracy on ImageNet and other benchmarks, but only when pre-trained at scale and fine-tuned for specific tasks.

Challenges in Training ViTs vs. CNNs

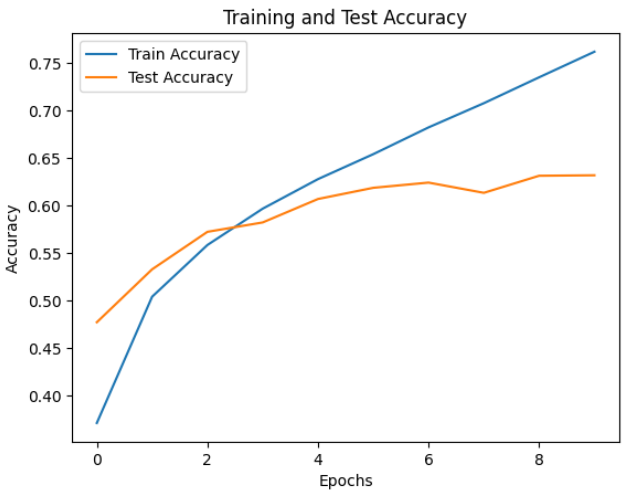
- **Data Requirements**: ViTs need much more data to generalize well, whereas CNNs can perform strongly with smaller datasets due to their built-in biases.
- **Optimization**: ViTs benefit from adaptive optimizers (like AdamW), while CNNs are typically trained with SGD. ViTs are prone to overfitting without strong data augmentation.
- **Computational Cost**: Training ViTs from scratch is computationally intensive, often requiring more resources than CNNs.
- **Generalization**: ViTs can be less robust to distribution shifts if not properly regularized or pre-trained

Practical Use Cases and Applications

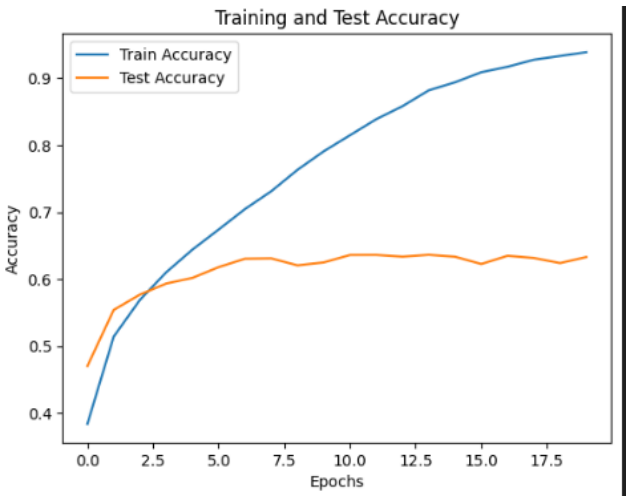
ViTs have found applications in a wide range of vision tasks, including: Image Classification, Object Detection & Segmentation, Medical Imaging, Remote Sensing, Robotics, Generative Modeling, etc.

Observations and Opinions on Practical Use Cases

ViTs are best suited for scenarios where large-scale labeled data is available and global context is crucial. For resource-constrained or data-limited environments, hybrid models or CNNs may still be preferable. As research progresses, techniques like masked autoencoders and data-efficient training are making ViTs more accessible for smaller datasets and broader applications. The flexibility and scalability of ViTs suggest they will play an increasingly important role in the future of computer vision, especially as data and computational resources continue to grow.



(a) Train Accuracy $> 76\%$ and Test Accuracy $> 63\%$



(b) Train Accuracy $> 93\%$ and Test Accuracy $> 63\%$

Figure 1: Comparison between two training setup