COMPUTER VISION

Unveiling Vision Transformers: Revolutionizing Computer Vision Beyond Convolution

Hansa hettiarachchi · Follow
4 min read · Aug 12, 2023

--

. . .

[img src:https://aisuperior.com/blog/how-computer-vision-is-transforming-the-healthcare-industry/]

What is a Vision Transformer?

Vision Transformer (ViT) is a groundbreaking neural network architecture that reimagines how we process and understand images. The Vision Transformer (ViT) model was introduced in 2021 in a conference research paper titled “An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale,” published at ICLR 2021. Inspired by the success of Transformers in natural language processing, ViT introduces a new way to analyze images by dividing them into smaller patches and leveraging self-attention mechanisms. This allows the model to capture both local and global relationships within images, leading to impressive performance in various computer vision tasks.

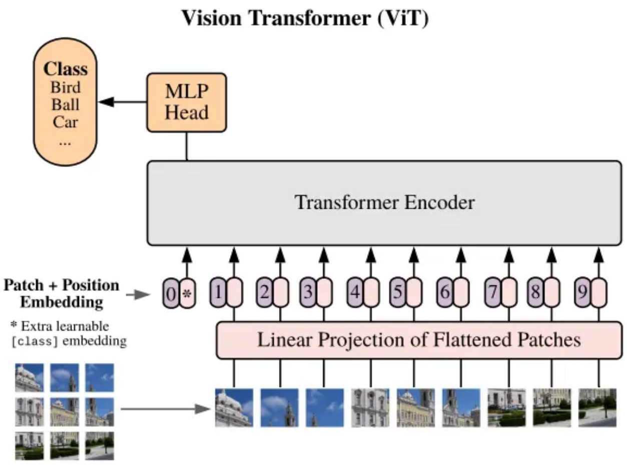
> *We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train [1]*

Vision Transformers vs. CNN?

ViT differs from Convolutional Neural Networks (CNNs) in several key aspects:

- Input Representation: While CNNs process raw pixel values directly, ViT divides the input image into patches and transforms them into tokens.
- Processing Mechanism: CNNs use convolutional and pooling layers to hierarchically capture features at different spatial scales. ViT employs self-attention mechanisms to consider relationships among all patches.
- Global Context: ViT inherently captures global context through self-attention, which helps in recognizing relationships between distant patches. CNNs rely on pooling layers for coarse global information.
- Data Efficiency: CNNs often require large amounts of labeled data for training, whereas ViT can benefit from pre-training on large datasets and then fine-tuning on specific tasks.

How does the Vision Transformer work?

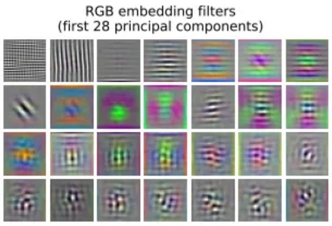


https://arxiv.org/abs/2010.11929

The Vision Transformer's operation can be broken down into several steps, each of which plays a crucial role in its overall functioning:

1. Patch Embedding:

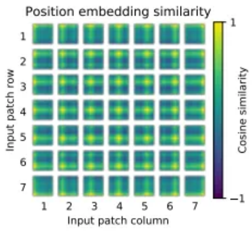
- The input image is divided into fixed-size square patches. Each patch is then linearly transformed into a vector using a learnable linear projection. This results in a sequence of patch embeddings, which serve as the input tokens for the subsequent layers.



Filters of the initial linear embedding of RGB values of ViT-L/32 model [https://arxiv.org/abs/2010.11929]

2. Positional Embedding:

- Since the Vision Transformer lacks any inherent understanding of spatial relationships, positional information needs to be explicitly provided. This is done by adding positional encodings to the patch embeddings.
- Positional encodings help the model differentiate between different positions in the image and capture spatial relationships. They are usually learned and added to the patch embeddings at the input stage.

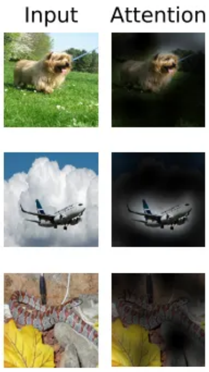


The similarity of position embeddings of ViT-L/32 model [https://arxiv.org/abs/2010.11929]

3. Encoder Layers:

- The core of the Vision Transformer consists of multiple encoder layers, each containing two primary sub-layers: multi-head self-attention and feedforward neural networks.

4. Multi-Head Self-Attention:



https://arxiv.org/abs/2010.11929

- The self-attention mechanism captures the relationships between different patches in the input sequence.
- For each patch embedding, self-attention computes a weighted sum of all patch embeddings, where the weights are determined by the relevance of each patch to the current one.
- This mechanism allows the model to focus on important patches while considering both local and global contexts.
- Multi-head attention employs multiple sets of learnable parameters (attention heads) to capture different types of relationships.

5. Feedforward Neural Networks:

- After self-attention, the output from each patch’s self-attention mechanism is passed through a feedforward neural network.
- This network typically consists of a fully connected layer followed by an activation function like ReLU (Rectified Linear Unit).
- The purpose of the feedforward network is to introduce non-linearity and allow the model to learn complex relationships between patches.

6. Layer Normalization and Residual Connections:

- Both the self-attention mechanism and the feedforward network outputs are followed by layer normalization and residual connections.
- Layer normalization helps stabilize and speed up training by normalizing the inputs to each sub-layer.
- Residual connections, also known as skip connections, add the original input embeddings to the output of each sub-layer. This helps in the flow of gradients during training and prevents the vanishing gradient problem.

Applications of Vision Transformers

- Image Classification: Classifying images into predefined categories
- Object Detection: Identifying and localizing objects within images
- Semantic Segmentation: Assigning a label to each pixel in an image to identify object boundaries
- Image Generation: Generating new images based on a given context or description

Limitations of Vision Transformers

- Large Datasets: Training Vision Transformers effectively often requires large datasets, which might not be available for all domains.
- Computational Demands: Training ViT can be computationally intensive due to the self-attention mechanisms.
- Spatial Information: ViT’s sequential processing might not capture fine-grained spatial patterns as effectively as CNNs for tasks like segmentation.

References

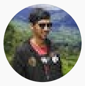
[1] https://arxiv.org/abs/2010.11929

[2] Vision Transformers (ViT) in Image Recognition—2023 Guide
Read more at: https://viso.ai/deep-learning/vision-transformer-vit/

[3] https://www.v7labs.com/blog/vision-transformer-guide

Computer Vision
Vision Transformer
Machine Learning
Artificial Intelligence
Convolutional Neural Net

--



Written by Hansa hettiarachchi

44 Followers

Follow