Analysis of Convolutional Neural Networks and Vision Transformers

1 Introduction

Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are two prominent architectures used in computer vision. CNNs have been the backbone of image recognition tasks for many years, while ViTs represent a newer approach leveraging the Transformer architecture, initially designed for natural language processing, for visual tasks.

2 Convolutional Neural Networks (CNNs)

2.1 What is a CNN?

CNNs are deep learning models specifically designed for processing structured grid data such as images. They utilize convolutional layers to extract features from the input image, enabling tasks like image classification, object detection, and segmentation.

2.2 Architecture

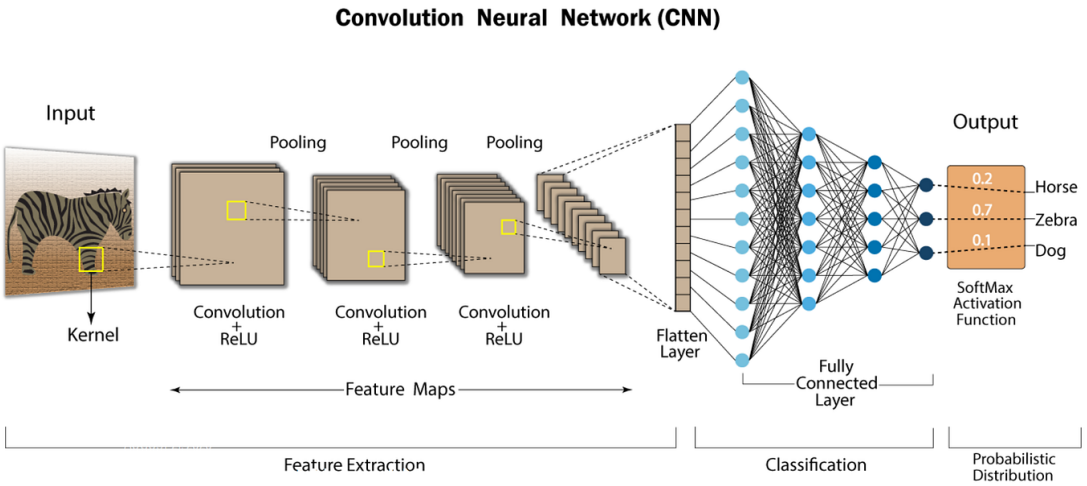

Figure 1: CNN Architecture

- **Convolutional Layers:** These layers apply convolution operations to the input image using filters/kernels, detecting various features such as edges and textures.
- **Pooling Layers:** These layers reduce the dimensionality of the feature maps, typically using operations like max pooling or average pooling.
- **Fully Connected Layers:** After several convolutional and pooling layers, fully connected layers are used to classify the extracted features into specific categories.

2.3 Activation Functions

Commonly used activation functions in CNNs include ReLU (Rectified Linear Unit) which introduces non-linearity to the model.

2.4 Loss Function

The loss function used in CNNs depends on the specific task. For classification tasks, Cross-Entropy Loss is typically used, while Mean Squared Error (MSE) is used for regression tasks.

2.5 Applications

CNNs are widely used in various applications including image and video recognition, medical image analysis, and autonomous driving.

3 Vision Transformers (ViTs)

3.1 What is a ViT?

Vision Transformers adapt the Transformer architecture for image recognition tasks. They treat images as sequences of patches and use self-attention mechanisms to process these sequences, capturing global dependencies more effectively than CNNs.

3.2 Architecture

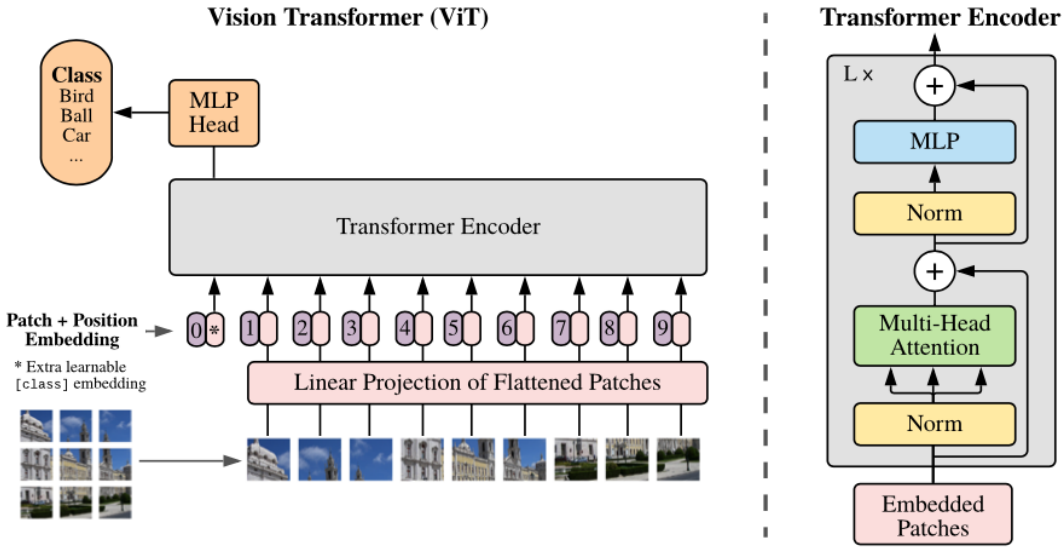

Figure 2: ViT Architecture

- **Patch Embedding:** The input image is divided into fixed-size patches, each of which is flattened and linearly embedded.
- **Transformer Encoder:** The embedded patches are fed into a standard Transformer encoder, which consists of multi-head self-attention layers and feed-forward neural networks.
- **Classification Head:** A classification token is added to the sequence of embedded patches, and the output corresponding to this token is used for the final classification.

3.3 Activation Functions

ViTs use activation functions like GELU (Gaussian Error Linear Unit) in their feed-forward networks within the Transformer encoder.

3.4 Loss Function

Similar to CNNs, ViTs often use Cross-Entropy Loss for classification tasks.

3.5 Applications

ViTs have shown great promise in various image classification tasks and are being explored in areas like image segmentation and object detection.

4 Differences Between CNNs and ViTs

- **Architecture:** CNNs use convolutional and pooling layers to process images, focusing on local features, whereas ViTs use self-attention mechanisms to capture global relationships in the image patches.
- **Data Requirements:** ViTs generally require more data for training compared to CNNs due to their complexity and need for extensive feature learning.
- **Performance:** ViTs have shown superior performance in tasks where capturing global context is crucial, while CNNs excel in scenarios where local feature extraction is more important.

5 Applications Comparison

5.1 CNNs

Well-suited for tasks where local feature extraction is critical, such as edge detection, and are highly efficient in terms of computational resources.

5.2 ViTs

Ideal for tasks requiring global context understanding, showing robustness in handling complex image classification tasks with sufficient training data.

6 Conclusion

Both CNNs and ViTs have their unique strengths and are suitable for different types of computer vision tasks. While CNNs remain a powerful tool for various applications, ViTs offer an innovative approach that leverages the power of Transformers to capture complex patterns in images. Understanding their differences and applications helps in choosing the right model architecture for specific computer vision challenges.

Nithish Chouti [ENGS3700]