Summary of Image Classification Architectures

Paper 1: EfficientNet – Rethinking Model Scaling for Convolutional Neural Networks (2019)

Problem Addressed

Before EfficientNet, improving image classification models mainly relied on increasing network depth, width, or input image resolution individually. These approaches were inefficient and often resulted in unnecessary computational cost without optimal performance improvements.

Proposed Solution

EfficientNet introduced a method called Compound Scaling. Instead of scaling only one dimension of a neural network, the model scales depth, width, and image resolution simultaneously using a balanced mathematical formula.

Architecture

EfficientNet is built on a baseline model called EfficientNet-B0. Larger models, ranging from B1 to B7, are created by applying compound scaling. The architecture uses MBConv blocks, squeeze-and-excitation attention mechanisms, and depthwise separable convolutions to improve efficiency and performance.

Key Contributions

- Improved classification accuracy while reducing the number of model parameters.
- Balanced scaling strategy for better efficiency.
- Provided multiple model sizes suitable for different computational resources.

Results

EfficientNet-B7 achieved state-of-the-art performance on the ImageNet dataset while using fewer computational resources compared to older models such as ResNet and Inception.

Importance

EfficientNet is widely used in transfer learning and real-world applications due to its strong balance between accuracy and computational efficiency.

Paper 2: Vision Transformers for Image Classification – A Comparative Survey

Background

Most traditional image classification models rely on Convolutional Neural Networks (CNNs). This paper explores a modern approach called Vision Transformers (ViT), which adapts transformer architectures originally designed for natural language processing.

Main Idea

Vision Transformers divide images into smaller patches, convert each patch into a vector representation, and process them using self-attention mechanisms instead of convolution operations.

Architecture

1. Divide the image into fixed-size patches.
2. Convert each patch into an embedding vector.
3. Add positional encoding to preserve spatial information.
4. Process embeddings using a Transformer encoder.
5. Use the output for image classification.

Advantages

- Captures global relationships between image regions.
- Provides strong performance with large datasets.
- Highly scalable and flexible architecture.

Limitations

- Requires very large training datasets.
- High computational cost.
- CNNs may perform better on smaller datasets.

Comparison with CNNs

CNNs are highly effective at capturing local features and typically perform well on smaller datasets. Vision Transformers, on the other hand, excel at capturing global dependencies but generally require more data and computational power.

Importance

Vision Transformers represent a major advancement in computer vision and are widely used in modern classification, detection, and segmentation tasks.