CNN Architectures and Their Evolution

1 Introduction

Convolutional Neural Networks (CNNs) are a class of deep neural networks specifically designed to process grid-structured data such as images. Unlike traditional Artificial Neural Networks (ANNs), CNNs exploit spatial locality, parameter sharing, and hierarchical feature learning. Over time, CNN architectures have evolved to address challenges related to limited depth, computational inefficiency, training instability, and scalability. This report discusses the evolution of CNN architectures, focusing on LeNet-5, AlexNet, InceptionNet (GoogLeNet), ResNet, and modern CNN variants.

2 LeNet-5: The First Practical CNN

LeNet-5, proposed by Yann LeCun in 1998, was one of the earliest successful CNN architectures designed for handwritten digit recognition.

2.1 Architecture Pipeline

$$\text{Input} \rightarrow \text{Convolution} \rightarrow \text{Pooling} \rightarrow \text{Convolution} \rightarrow \text{Pooling} \rightarrow \text{Fully Connected} \rightarrow \text{Output}$$

2.2 Key Concepts

LeNet-5 introduced local receptive fields, weight sharing, and subsampling (average pooling). These ideas allowed the network to learn spatial hierarchies while keeping the number of parameters manageable.

2.3 Mathematical Insight

A convolution operation is defined as:

$$y_{i,j} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n} + b$$

2.4 Limitations

LeNet-5 is shallow and does not scale well to large datasets or complex image representations.

3 AlexNet: The Deep Learning Breakthrough

AlexNet, introduced in 2012 by Krizhevsky et al., demonstrated that deep CNNs trained on large datasets using GPUs could significantly outperform traditional computer vision methods.

3.1 Architecture Pipeline

Input $\rightarrow$ Conv + ReLU $\rightarrow$ Pooling $\rightarrow$ Conv + ReLU $\rightarrow$ Pooling $\rightarrow$ Fully Connected $\rightarrow$ Softmax

3.2 Key Innovations

AlexNet employed the ReLU activation function:

$$f(x) = \max(0, x)$$

along with dropout for regularization and extensive data augmentation.

3.3 Impact

AlexNet drastically reduced ImageNet classification error and replaced hand-crafted features with learned representations, marking the beginning of the deep learning era in computer vision.

4 InceptionNet (GoogLeNet): Multi-Scale Feature Learning

GoogLeNet introduced the Inception architecture to increase depth while maintaining computational efficiency.

4.1 Architecture Pipeline

$$\text{Input} \rightarrow \text{Inception Modules} \rightarrow \text{Global Average Pooling} \rightarrow \text{Softmax}$$

4.2 Inception Module

Each Inception module applies multiple convolutional operations in parallel and concatenates their outputs to capture multi-scale features.

4.3 Role of 1×1 Convolutions

$$y_{i,j} = \sum_{c} w_c \cdot x_{i,j,c}$$

These convolutions reduce channel dimensionality, lower computational cost, and introduce additional non-linearity.

4.4 Advantages

GoogLeNet achieved strong performance with significantly fewer parameters than AlexNet while being deeper.

5 ResNet: Deep Networks Made Trainable

5.1 The Degradation Problem

As network depth increases, training error can increase even without overfitting, a phenomenon known as the degradation problem.

5.2 Residual Learning

ResNet reformulates the learning objective by learning a residual function:

$$F(x) = H(x) - x$$

leading to:

$$y = F(x) + x$$

5.3 Architecture Pipeline

Input → Convolutional Layers → Residual Blocks → Global Average Pooling → Softmax

5.4 Gradient Flow

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial y} \left(I + \frac{\partial F(x)}{\partial x} \right)$$

The identity term ensures stable gradient propagation through deep networks.

5.5 Contributions

ResNet enabled the successful training of very deep networks (up to 150+ layers) and became a foundational architecture in modern computer vision.

6 Modern CNN Architectures

6.1 MobileNet

MobileNet uses depthwise separable convolutions to reduce computation:

$$\text{Depthwise Conv} \rightarrow \text{Pointwise Conv}$$

This significantly reduces computational cost compared to standard convolutions.

6.2 EfficientNet

EfficientNet introduces compound scaling that jointly scales network depth, width, and input resolution:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$

This leads to an optimal balance between accuracy and efficiency.

Architecture	Key Idea	Depth	Efficiency
LeNet-5	CNN Fundamentals	Shallow	Low
AlexNet	Deep CNN + GPU	Medium	Low
InceptionNet	Multi-scale Learning	Deep	High
ResNet	Skip Connections	Very Deep	High
Modern CNNs	Efficiency Scaling	Variable	Very High

7 Comparative Summary

8 Conclusion

The evolution of CNN architectures demonstrates a steady progression toward deeper, more efficient, and more stable models. From LeNet-5's foundational concepts to ResNet's residual learning and modern efficiency-driven designs, CNNs continue to serve as the backbone of state-of-the-art computer vision systems.