# CNN Architectures and Evolution

## Introduction

Convolutional Neural Networks (CNNs) have evolved rapidly, with each major architecture introducing a key idea that improved performance, training stability, or computational efficiency. This evolution reflects a shift from simple convolution–pooling pipelines to deeper, more optimized, and more efficient designs. This report discusses the progression of CNN architectures from early models such as LeNet-5 to modern efficient architectures like MobileNet, highlighting their core innovations and impact on computer vision.

## LeNet-5: Early CNN Foundations

LeNet-5 is one of the earliest successful CNN architectures, originally developed for handwritten digit recognition. It established the basic structure still used in CNNs today: convolution layers for feature extraction, pooling layers for spatial reduction, and fully connected layers for classification.

A typical LeNet-style pipeline follows:

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

LeNet-5 demonstrated that learned local filters could outperform handcrafted features in vision tasks.

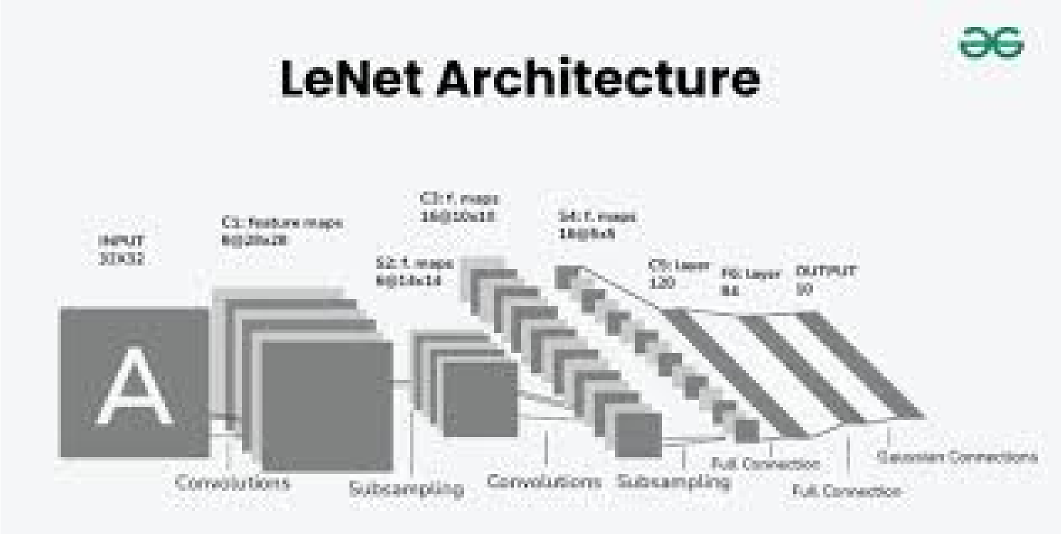

Figure 1: LeNet-5 architecture overview

## AlexNet: Impact on Computer Vision

AlexNet marked a breakthrough in computer vision by winning the ImageNet Large Scale Visual Recognition Challenge in 2012 with a significant margin. It demonstrated that deep CNNs trained on large datasets using GPUs could vastly outperform traditional methods.

Key contributions of AlexNet include:

- Increased network depth compared to earlier CNNs
- Use of ReLU activation functions for faster convergence
- Dropout for regularization to reduce overfitting
- Extensive data augmentation and GPU-based training

AlexNet's success triggered widespread adoption of deep learning in computer vision.

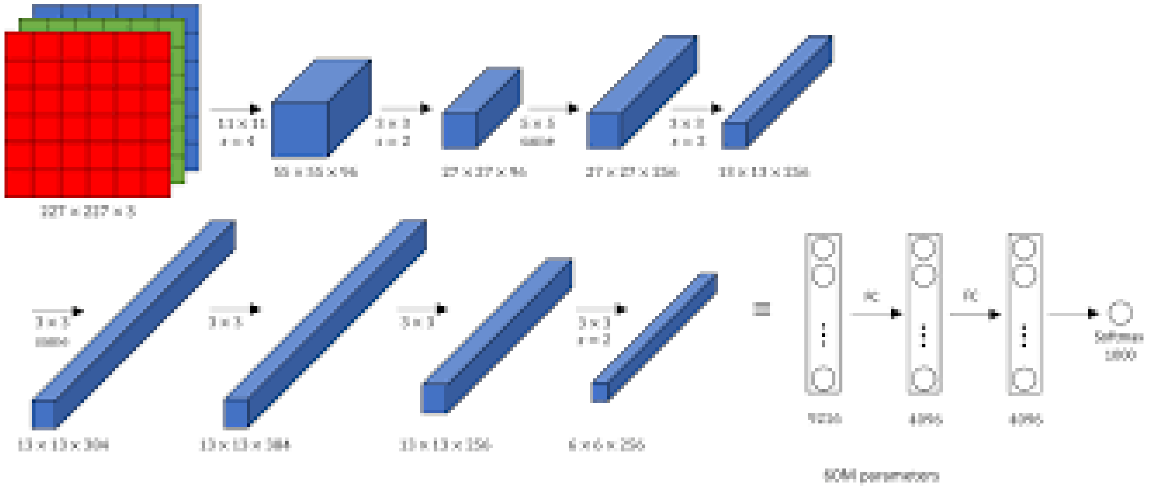

Figure 2: AlexNet architecture (high-level representation)

# Inception/GoogLeNet: Multi-Scale Feature Learning

GoogLeNet introduced the Inception architecture, which captures visual features at multiple spatial scales within the same layer. Instead of committing to a single kernel size, an Inception module applies several convolutions in parallel and concatenates their outputs.

This design allows the network to efficiently learn both fine and coarse features while keeping computational cost manageable.

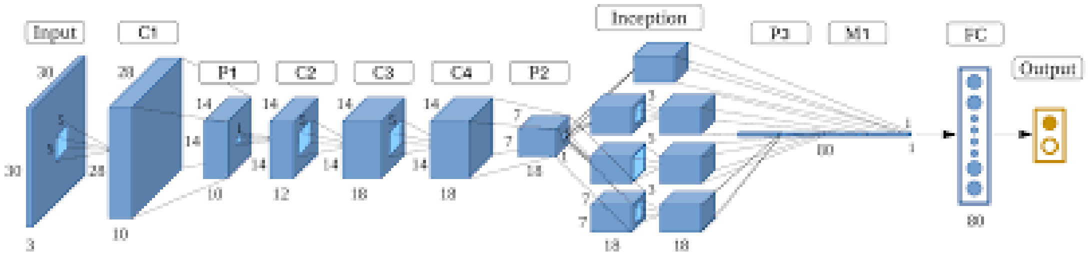

Figure 3: Inception module with parallel convolution branches

## Role of $1 \times 1$ Convolutions

A key component of Inception modules is the use of $1 \times 1$ convolutions for dimensionality reduction. These layers mix information across channels while reducing the number of feature maps before applying expensive convolutions.

Mathematically, a $1 \times 1$ convolution performs:

$$Y(h, w, c) = \sum_{k=1}^{C_{in}} W(c, k)\, X(h, w, k) + b(c)$$

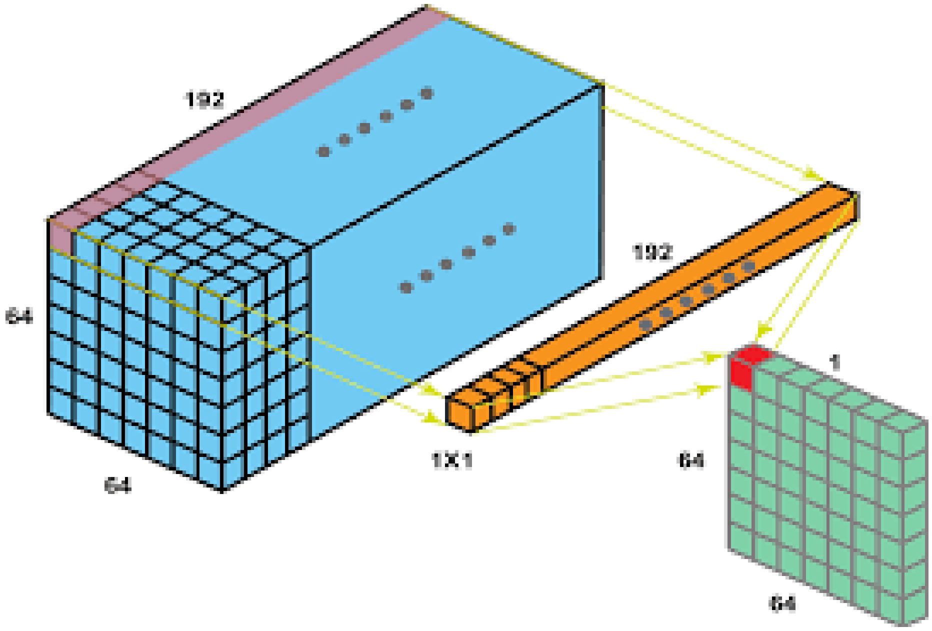

Figure 4: $1 \times 1$ convolution used for channel reduction

## ResNet: Skip Connections and the Degradation Problem

As CNNs became deeper, training accuracy sometimes degraded despite increased model capacity. This degradation problem arises from optimization difficulties rather than overfitting. ResNet addressed this issue by introducing residual connections.

A residual block learns a residual function $F(x)$ and adds it to the input:

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

These skip connections allow gradients to flow more easily through deep networks, enabling the successful training of very deep CNNs.

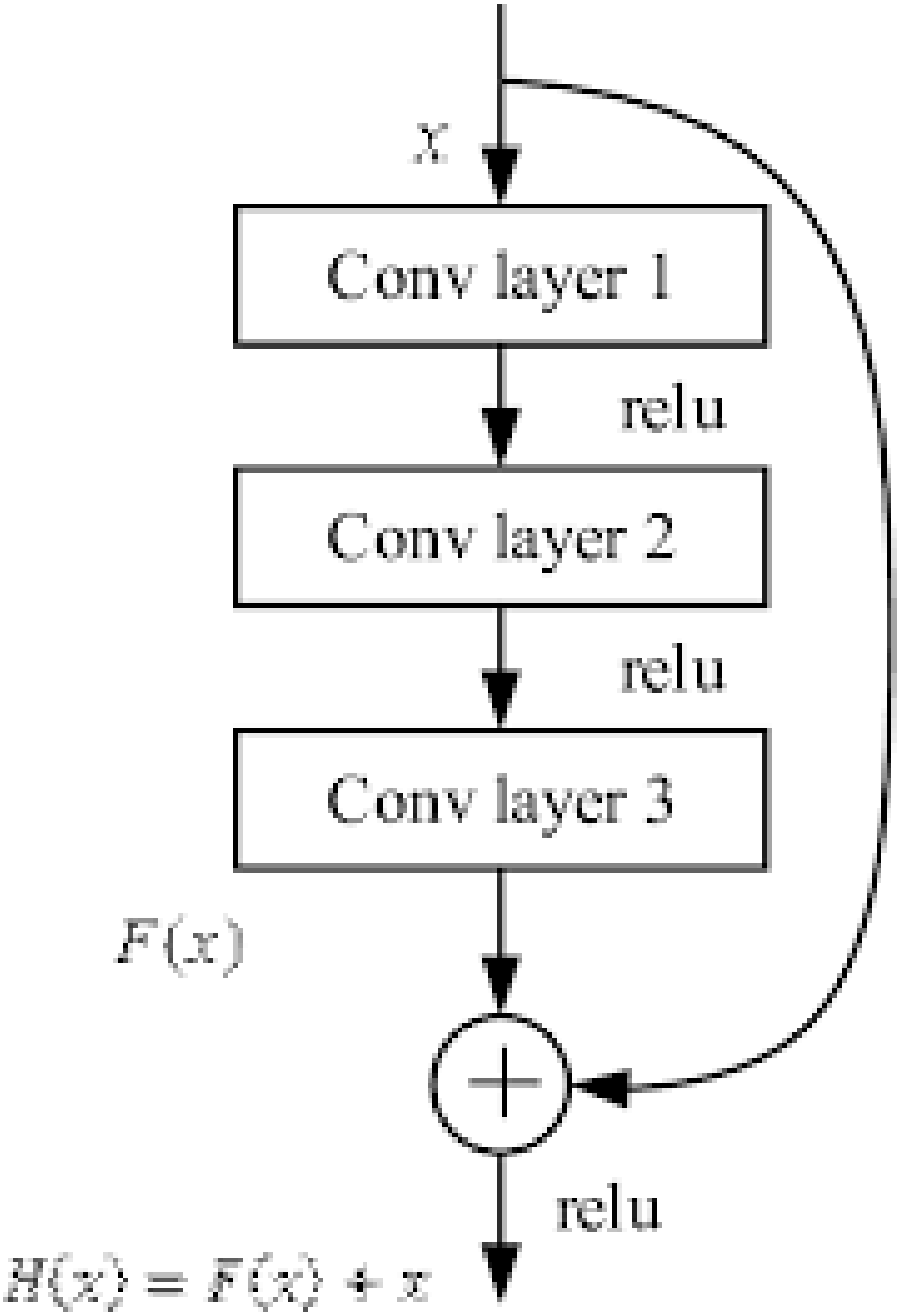

Figure 5: Residual block with skip connection

## MobileNet: Efficient CNNs for Edge Devices

MobileNet was designed to make CNNs computationally efficient for mobile and embedded devices. Its key innovation is depthwise separable convolution, which factorizes standard convolution into two simpler operations.

A standard convolution has computational cost:

$$K^2 \cdot C_{in} \cdot C_{out}$$

Depthwise separable convolution splits this into:

- **Depthwise convolution:** one filter per input channel
- **Pointwise convolution:** a $1 \times 1$ convolution to combine channels

This significantly reduces computation while maintaining strong performance.

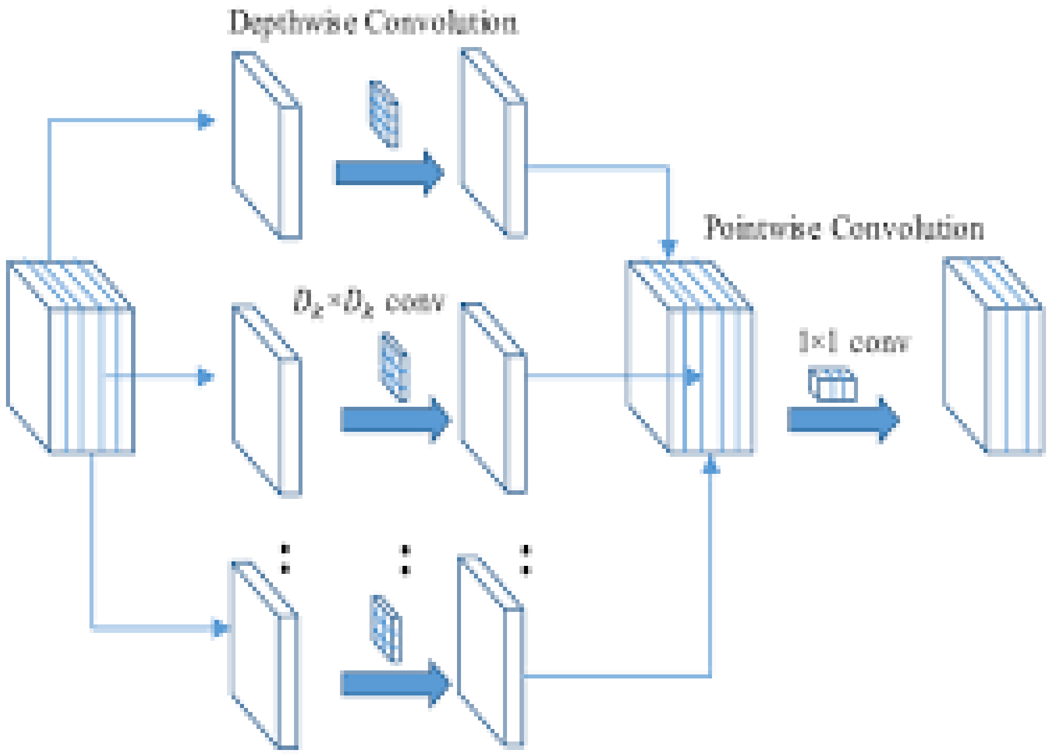

Figure 6: Depthwise separable convolution used in MobileNet

# Conclusion

The evolution of CNN architectures reflects a progression of key ideas: LeNet-5 introduced the convolution–pooling paradigm, AlexNet demonstrated large-scale deep learning for vision, Inception improved multi-scale feature extraction, ResNet enabled extremely deep networks through skip connections, and MobileNet brought CNNs to resource-constrained devices. These architectures collectively form the foundation of modern computer vision systems and continue to influence current CNN design.