Convolutional Neural Networks: Architectures and Evolution

Idant Srivastava

December 2025

Contents

1 LeNet-5 (1998) 3

1.1 Architecture . . . 3

1.2 Key Innovations . . . 3

1.3 Intuition . . . 3

2 AlexNet (2012) 4

2.1 Architecture . . . 4

2.2 Key Innovations . . . 4

2.3 Intuition . . . 4

3 InceptionNet (GoogLeNet, 2014) 5

3.1 Architecture . . . 5

3.2 Key Innovations . . . 5

3.3 Intuition . . . 5

4 Residual Networks (ResNet, 2015) 6

4.1 Residual Block . . . 6

4.2 Architecture . . . 6

4.3 Key Innovations . . . 6

4.4 Intuition . . . 6

5 MobileNets (2017-2019) 8

5.1 Depthwise Separable Convolution . . . 8

5.2 Key Innovations . . . 8

5.3 Intuition . . . 8

6 EfficientNet (2019) 9

6.1 Key Innovations and Features . . . 9

6.2 Intuition . . . 9

Introduction

Convolutional Neural Networks (CNNs) have revolutionized computer vision and image processing tasks since their inception. From their beginnings in the 1980s to their dominance in modern deep learning applications, CNNs have undergone significant architectural evolution, driven by the need for better performance, efficiency, and generalization.

The fundamental principle behind CNNs is the use of spatial hierarchies in data. Unlike fully connected networks that treat input features independently, CNNs preserve spatial relationships through local connectivity and weight sharing. This makes them particularly effective for image data where nearby pixels are strongly correlated.

This report examines the theoretical foundations of CNNs, traces their architectural evolution, and provides intuitive explanations for the design choices that have shaped modern computer vision systems.

Chapter 1

LeNet-5 (1998)

LeNet-5, developed by Yann LeCun, was one of the first successful CNNs, designed for handwritten digit recognition.

1.1 Architecture

- Input: 32×32 grayscale images
- Conv1: 6 filters $(5 \times 5) \rightarrow$ Average pooling
- Conv2: 16 filters $(5 \times 5) \rightarrow$ Average pooling
- Fully connected layers: $120 \rightarrow 84 \rightarrow 10$ (output)

1.2 Key Innovations

- Introduced the convolutional-pooling pattern
- Demonstrated end-to-end learning from raw pixels
- Used sparse connections between layers

1.3 Intuition

LeNet established the fundamental CNN pattern: alternating convolution and pooling layers to progressively extract hierarchical features, followed by fully connected layers for classification.

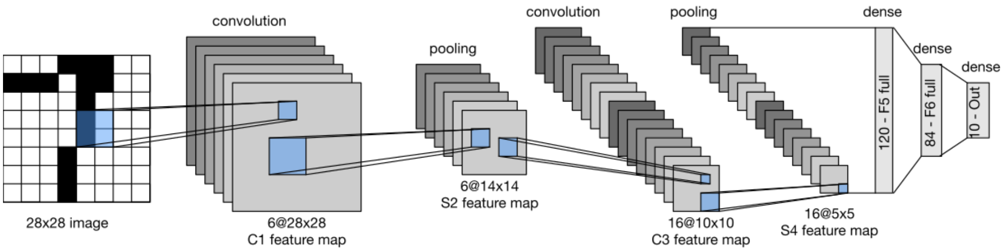

Figure 1.1: LeNet-5

Chapter 2

AlexNet (2012)

AlexNet, developed by Krizhevsky et al., marked the beginning of the deep learning revolution by winning ImageNet 2012 with a significant margin.

2.1 Architecture

- 5 convolutional layers with ReLU activation
- 3 max-pooling layers
- 3 fully connected layers
- Dropout for regularization

2.2 Key Innovations

- First large-scale use of ReLU instead of tanh/sigmoid
- Introduced dropout to prevent overfitting
- Used data augmentation extensively
- Leveraged GPU training for parallel computation

2.3 Intuition

AlexNet showed that deeper networks with more parameters could learn better representations when trained on large datasets with powerful hardware. The use of ReLU activation was crucial for training such deep networks.

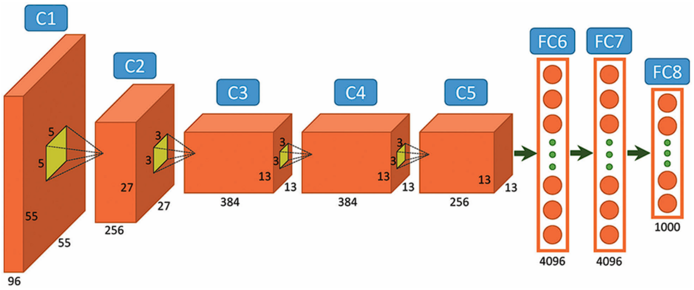

Figure 2.1: AlexNet

Chapter 3

InceptionNet (GoogLeNet, 2014)

The Inception architecture introduced by Szegedy et al. focused on computational efficiency and multi-scale feature extraction.

3.1 Architecture

The core idea is to apply multiple filter sizes in parallel and concatenate the results across dimensions.

- 1×1, 3×3, and 5×5 convolutions
- 3×3 max pooling

3.2 Key Innovations

- Multi-scale feature extraction in single module
- 1×1 convolutions for dimensionality reduction
- Auxiliary classifiers for training deeper networks
- Efficient use of parameters (only 5M vs AlexNet's 60M)

3.3 Intuition

Instead of choosing a single filter size, Inception modules capture patterns at multiple scales simultaneously. The 1×1 convolutions act as "network-in-network" operations, reducing computational cost while adding non-linearity. This makes the network more efficient.

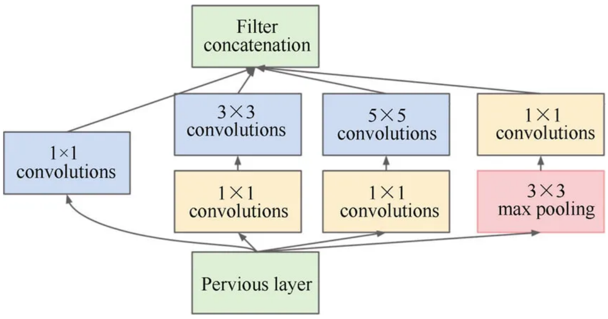

Figure 3.1: Inception Layer

Chapter 4

Residual Networks (ResNet, 2015)

ResNet, introduced by He et al., revolutionized deep learning by enabling training of extremely deep networks (up to 152 layers and beyond).

4.1 Residual Block

The fundamental innovation is the skip connection:

$$\mathbf{y} = F(\mathbf{x}) + \mathbf{x} \tag{4.1}$$

where $F(\mathbf{x})$ represents the learned residual mapping.

4.2 Architecture

- Multiple stages with residual blocks
- Increasing channels: 64 → 128 → 256 → 512
- Bottleneck blocks for efficiency (1×1 → 3×3 → 1×1)

4.3 Key Innovations

- Skip connections enable gradient flow through deep networks
- Solves vanishing gradient problem
- Networks can learn identity mappings easily
- Batch normalization after each convolution

4.4 Intuition

Traditional deep networks suffer from degradation—adding more layers can actually hurt performance. ResNet's skip connections allow gradients to flow directly backward, enabling effective training. If a layer is unnecessary, the network can learn to set $F(\mathbf{x}) \approx 0$, making the block an identity function. This makes optimization easier and allows networks to be much deeper.

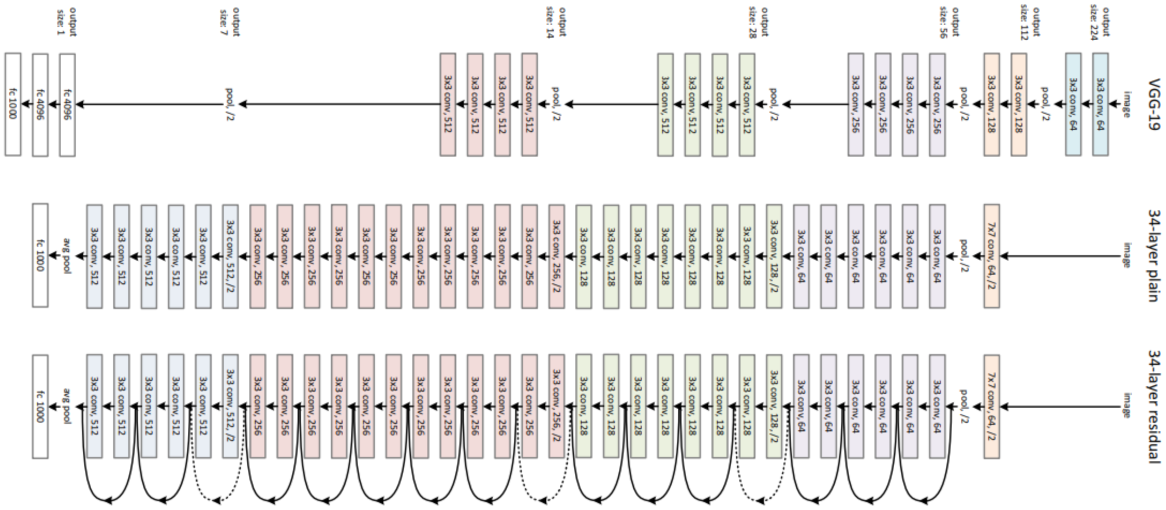

Figure 4.1: ResNet

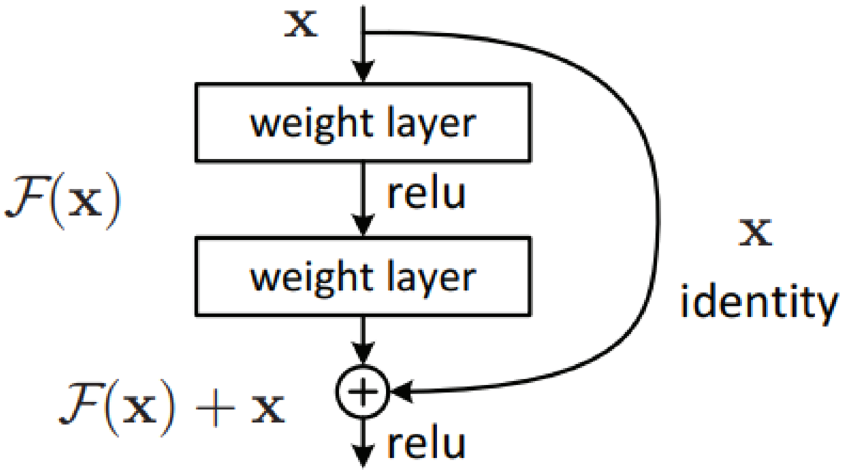
x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 4.2: Residual Block

Chapter 5

MobileNets (2017-2019)

MobileNet is a family of lightweight, power-efficient CNN architectures designed for mobile and embedded vision applications. It uses depthwise separable convolutions to significantly reduce parameter count and computational cost—often featuring 4.2 million parameters compared to 138 million in VGG16—while maintaining competitive accuracy.

5.1 Depthwise Separable Convolution

Standard convolution is factorized into two operations:

1. **Depthwise convolution:** Applies a single filter per input channel
2. **Pointwise convolution:** 1×1 convolution to combine channels

5.2 Key Innovations

- Depthwise separable convolutions
- Width multiplier for scaling network size
- Resolution multiplier for computational trade-offs
- Inverted residuals (MobileNetV2)

5.3 Intuition

Standard convolutions are computationally expensive because they mix spatial and channel-wise information together. By separating these operations, MobileNets achieve comparable accuracy with dramatically fewer operations. The inverted residual blocks in V2 expand channels before depthwise convolution, allowing more expressive transformations in the efficient depthwise step.

Chapter 6

EfficientNet (2019)

EfficientNet is a family of powerful and efficient CNNs, known for achieving high accuracy in computer vision tasks with fewer parameters and computations, primarily through its innovative compound scaling method that uniformly scales depth, width, and resolution using a single coefficient.

6.1 Key Innovations and Features

- Instead of scaling just layers (depth) or neurons (width), EfficientNet scales depth, width, and input resolution together, finding an optimal balance for better performance.
- The baseline EfficientNet-B0 was found using Neural Architecture Search (NAS) and then scaled up to create larger, more powerful versions (B1 to B7).
- It provides high accuracy with significantly fewer parameters which is crucial for low level devices.

6.2 Intuition

Scaling only one dimension (depth, width, or resolution) leads to diminishing returns. EfficientNet scales all three dimensions in a balanced way. This achieves better performance than scaling any single dimension, while maintaining computational efficiency.

Comparative Analysis

Table 6.1: Comparison of Major CNN Architectures

Architecture	**Year**	**Depth**	**Parameters**	**Key Innovation**
LeNet-5	1998	7	60K	Conv-Pool pattern
AlexNet	2012	8	60M	ReLU, Dropout, GPU
VGG-16	2014	16	138M	Small filters, Depth
GoogLeNet	2014	22	5M	Inception modules
ResNet-50	2015	50	25M	Skip connections
MobileNetV2	2018	53	3.4M	Depthwise separable
EfficientNet-B0	2019	-	5.3M	Compound scaling

Performance Trade-offs:

- VGG: Simple but memory-intensive
- Inception: Efficient but complex architecture
- ResNet: Excellent depth scalability
- MobileNet: Best for resource-constrained devices
- EfficientNet: Optimal accuracy-efficiency balance

Conclusion

The evolution of CNN architectures reflects a continuous pursuit of better performance, efficiency, and understanding. From LeNet's proof-of-concept to modern efficient networks, each innovation has built upon previous work while addressing specific limitations.

Key lessons from this evolution include:

1. **Depth matters:** Deeper networks can learn more complex hierarchical representations.
2. **Skip connections enable depth:** Residual connections solve optimization challenges in very deep networks.
3. **Efficiency through design:** Clever architectural choices can dramatically reduce computation without sacrificing accuracy.
4. **Multi-scale is powerful:** Processing information at multiple scales improves robustness and performance.
5. **Balanced scaling:** Uniformly scaling all dimensions is more effective than arbitrary increases.

As we move forward, the integration of CNNs with attention mechanisms, the development of more efficient architectures through neural architecture search, and the exploration of hybrid designs will continue to push the boundaries of what's possible in computer vision.