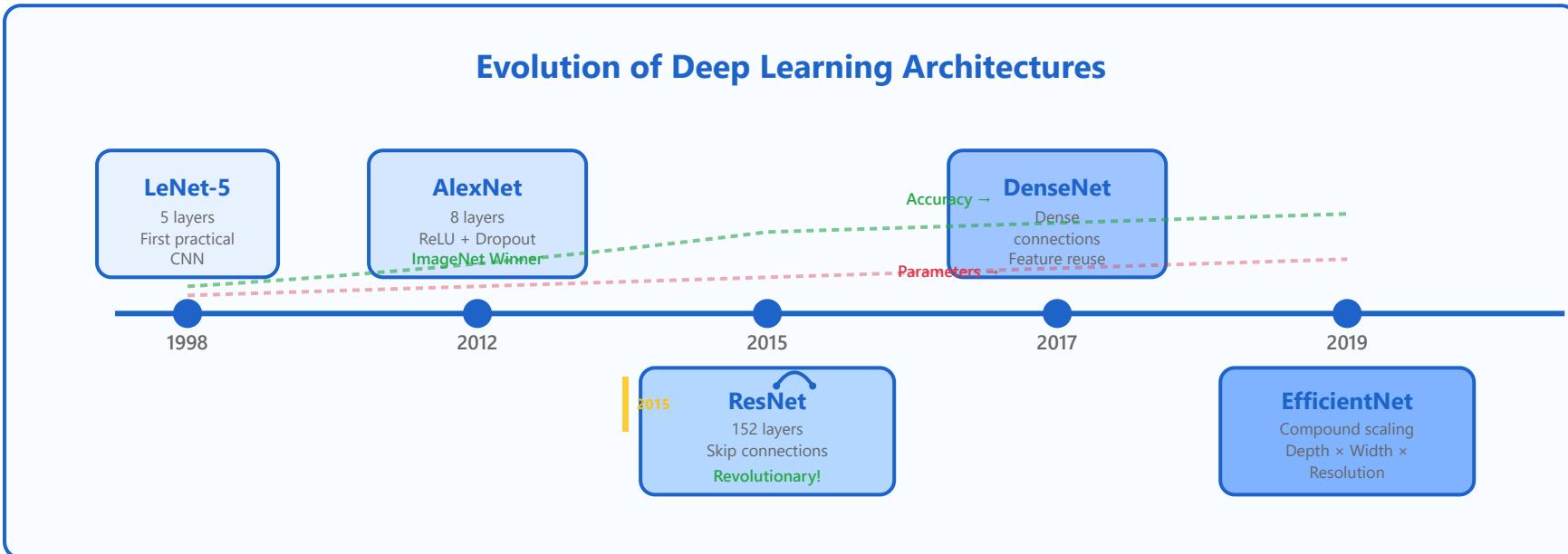


CNN Architectures Evolution



LeNet to AlexNet

Early pioneers (1998-2012): Basic convolution layers, introduced ReLU and dropout for deeper networks

ResNet & Skip Connections

Revolutionary residual connections enable 100+ layer networks. Solves vanishing gradient problem

DenseNet

Dense connections between all layers. Enhanced feature propagation and parameter efficiency

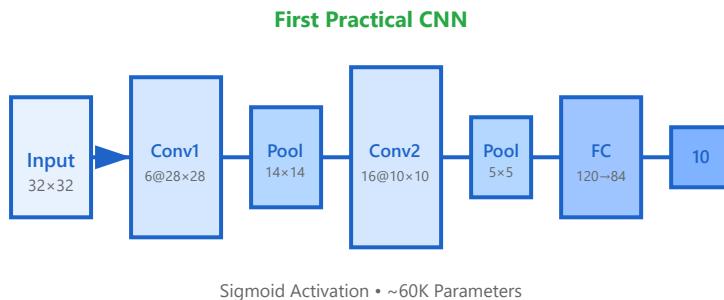
EfficientNet

Compound scaling of depth, width, and resolution. Optimal accuracy-efficiency tradeoff

Modern Trend: **Neural Architecture Search (NAS)** - Automated discovery of optimal architectures for specific tasks

1. LeNet-5 & AlexNet: The Pioneers

LeNet-5 Architecture (1998)



LeNet-5 Overview

Developed by Yann LeCun in 1998, LeNet-5 was the first successful application of CNNs for handwritten digit recognition. Originally designed for reading zip codes and bank checks.

Key Features:

- ▶ Simple sequential architecture with 5 layers
- ▶ Used sigmoid and tanh activation functions
- ▶ Average pooling for downsampling
- ▶ ~60,000 parameters - very small by today's standards
- ▶ Proved that CNNs could learn hierarchical features

Foundation of Modern CNNs

AlexNet Overview

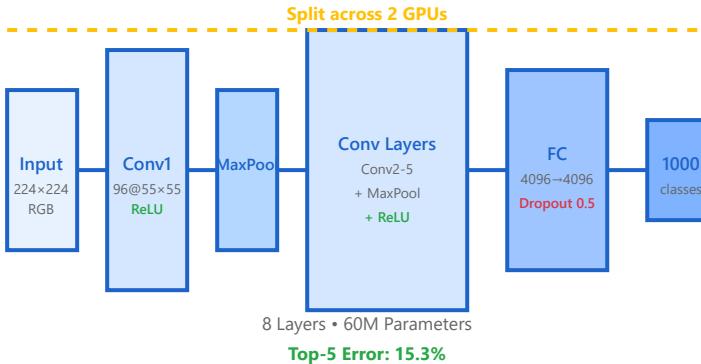
AlexNet won the ImageNet ILSVRC 2012 competition with a top-5 error of 15.3%, significantly better than the second place (26.2%). This breakthrough revived interest in deep learning.

AlexNet Architecture (2012)

Revolutionary Innovations:

- ▶ **ReLU Activation:** Replaced sigmoid/tanh, enabling faster training
- ▶ **Dropout:** First use of dropout (0.5) to prevent overfitting
- ▶ **Data Augmentation:** Random crops, flips, color jittering
- ▶ **GPU Training:** Used 2 GPUs for parallel training
- ▶ **Local Response Normalization** for better generalization
- ▶ 60 million parameters, 650,000 neurons

ImageNet Winner 2012



2. ResNet: Skip Connections Revolution

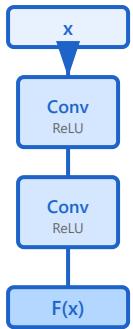
Residual Block Architecture

ResNet Innovation

ResNet (Residual Network) introduced skip connections that revolutionized deep learning. Instead of learning the desired mapping $H(x)$, layers learn the residual $F(x) = H(x) - x$, making optimization much easier.

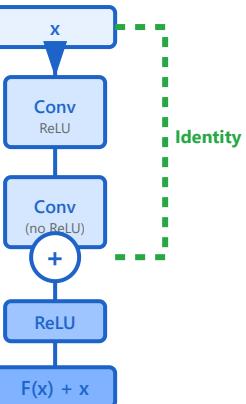
The Key Insight: If the optimal mapping is close to an identity function, it's easier to learn the

Traditional Block



Vanishing Gradient
Hard to train deep

Residual Block



✓ Easy to train
100+ layers possible

residual (small deviations) than to learn the complete transformation from scratch.

Key Advantages:

- ▶ **Solves Vanishing Gradient:** Gradients flow directly through skip connections
- ▶ **Deep Networks:** Enables training of 100+ layer networks (ResNet-152)
- ▶ **Better Optimization:** Easier to optimize than plain networks
- ▶ **No Extra Parameters:** Identity shortcuts add no complexity
- ▶ Won ImageNet 2015 with 3.57% top-5 error

ImageNet Winner 2015

ResNet Architecture Variants

ResNet-18

- 18 layers
- 11.7M parameters
- Basic blocks
- Fast inference

Good for edge devices

ResNet-34

- 34 layers
- 21.8M parameters
- Basic blocks
- Balanced

Popular baseline

ResNet-50 ⭐

- 50 layers
- 25.6M parameters
- Bottleneck blocks
- 1×1 convolutions

Most widely used

ResNet-152

- 152 layers
- 60.2M parameters
- Bottleneck blocks
- Highest accuracy

Competition winner

Increasing Depth & Accuracy →

3. DenseNet: Dense Connectivity Pattern

DenseNet Architecture

DenseNet (Densely Connected Convolutional Networks) takes the idea of skip connections further by connecting each layer to every other layer in a feed-forward fashion. For a network with L layers, there are $L(L+1)/2$ direct connections.

Core Principle: Each layer receives feature maps from all preceding layers and passes its own feature maps to all subsequent layers. This creates maximum information flow between layers.

Key Benefits:

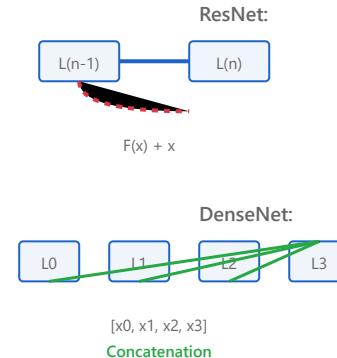
- **Feature Reuse:** All layers access features from previous layers
- **Gradient Flow:** Improved gradient propagation to early layers
- **Parameter Efficiency:** Fewer parameters than ResNet
- **Implicit Deep Supervision:** All layers receive gradients directly
- Better feature propagation and exploration

CVPR 2017 Best Paper

Dense Block Structure



Comparison



Feature Reuse

All previous features available to all layers

DenseNet Complete Architecture



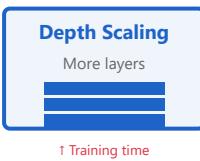
DenseNet-121: ~8M parameters • DenseNet-169: ~14M parameters

More efficient than ResNet with comparable accuracy

4. EfficientNet: Compound Scaling

Compound Scaling Method

Traditional Approaches



EfficientNet: Compound Scaling

Balanced Scaling with ϕ :

$$\text{depth: } d = \alpha^\phi$$

$$\text{width: } w = \beta^\phi$$

$$\text{resolution: } r = \gamma^\phi$$

$$\text{subject to: } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$

✓ Optimal accuracy-efficiency tradeoff

EfficientNet Innovation

EfficientNet systematically scales network depth, width, and resolution using a compound coefficient. Instead of arbitrarily scaling one dimension, it balances all three dimensions for optimal performance.

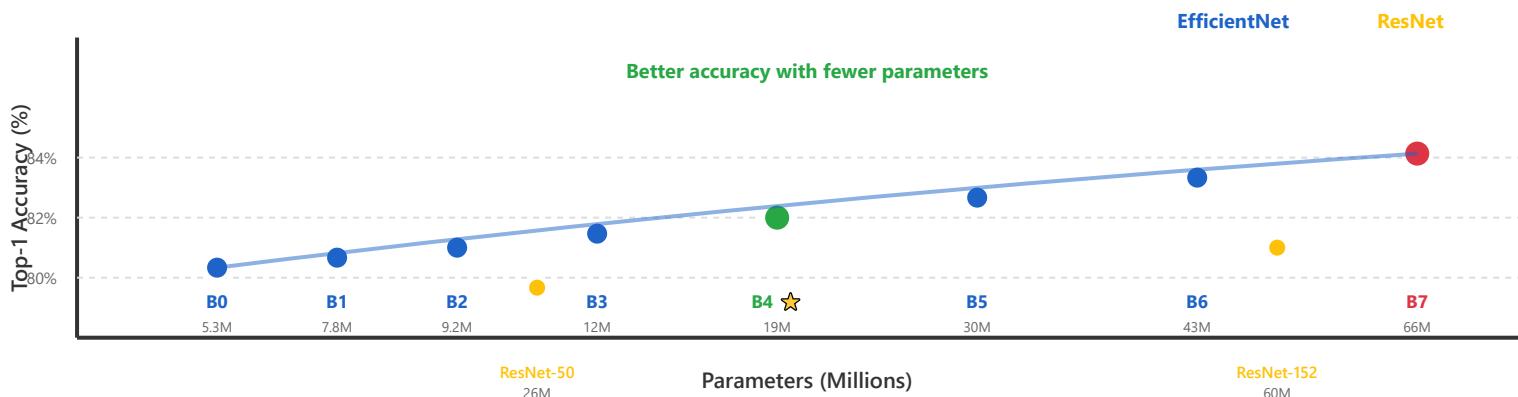
Key Insight: Network accuracy improves when depth, width, and resolution are scaled together in a principled way, rather than scaling dimensions independently.

Key Innovations:

- ▶ **Neural Architecture Search:** Base architecture (EfficientNet-B0) found via NAS
- ▶ **Compound Scaling:** Uniform scaling of all dimensions with fixed ratios
- ▶ **Mobile Inverted Bottleneck:** MBConv blocks with squeeze-excitation
- ▶ **Efficiency:** Up to 10x fewer parameters than ResNet
- ▶ State-of-the-art accuracy with better efficiency
- ▶ EfficientNet-B7: 84.4% top-1 accuracy on ImageNet

ICML 2019

EfficientNet Model Family (B0-B7)



Mobile Inverted Bottleneck Convolution (MBConv) Block

Inverted Residual: Expand → Filter → Compress

Efficient for mobile and resource-constrained environments

