CNN ARCHITECTURES REVISION NOTES

Prepared By Sumit Shukla

Date: September 15, 2025

1. BASICS OF CNN (CONVOLUTIONAL NEURAL NETWORKS)

Quick Intro:

CNNs are deep learning models specifically designed to process grid-like data such as images. They automatically learn hierarchical features from raw pixel data without manual feature extraction.

Key Explanation:

- Architecture: CNNs consist of convolutional layers, pooling layers, and fully connected layers
- Feature Learning: Early layers detect simple features (edges, textures), deeper layers identify complex patterns (shapes, objects)
- Spatial Hierarchy: Maintains spatial relationships between pixels while progressively reducing spatial dimensions
- Translation Invariance: Can recognize patterns regardless of their position in the input
- Parameter Sharing: Same filter weights are used across different spatial locations, reducing parameters
- Applications: Image classification, object detection, image segmentation, medical imaging

Key Components:

- Convolutional layers (feature extraction)
- Pooling layers (dimensionality reduction)
- Activation functions (non-linearity)
- Fully connected layers (classification)

2. CONVOLUTION OPERATION

Quick Intro:

Convolution is the core mathematical operation in CNNs where a filter (kernel) slides across the input image to detect specific features through element-wise multiplication and summation.

Key Explanation:

- Process: A small matrix (kernel/filter) moves across the input image, computing dot products at each position
- Feature Detection: Different kernels detect different features (edges, corners, textures)
- Output: Creates a feature map highlighting where specific patterns are found
- Kernel Size: Typically 3×3, 5×5, or 7×7
- Stride: Number of pixels the kernel moves (usually 1 or 2)
- Padding: Adding borders to maintain spatial dimensions

Types of Convolutions:

- Standard Convolution
- Depthwise Convolution
- Pointwise Convolution (1×1)
- Dilated Convolution

3. CONCEPT OF FEATURE MAP

Quick Intro:

A feature map is the output of a convolutional layer, representing the activation of specific filters across different spatial locations in the input image.

Key Explanation:

- Definition: 2D array showing where specific features are detected in the input
- Generation: Created when a filter is convolved with the input image
- Interpretation: High values indicate strong presence of the detected feature
- Multiple Maps: Each filter produces one feature map; multiple filters create multiple feature maps
- Depth: Number of feature maps equals the number of filters in the layer
- Spatial Information: Preserves spatial relationships of detected features
- Progressive Abstraction: Feature maps become more abstract as depth increases

Hierarchy:

- Early layers: Low-level features (edges, lines, corners)
- Middle layers: Mid-level features (shapes, textures)
- Deep layers: High-level features (objects, complex patterns)

4. IMAGENET

Quick Intro:

ImageNet is a large-scale image database containing over 14 million high-resolution images across thousands of categories, serving as the standard benchmark for computer vision research.

Key Explanation:

- Scale: 14+ million images, 20,000+ categories based on WordNet hierarchy
- ImageNet Challenge (ILSVRC): Annual competition that drove major breakthroughs in deep learning

- Dataset Structure: Organized hierarchically using WordNet taxonomy
- Applications: Training and benchmarking CNN models, Transfer learning source dataset, Pre-trained model development
- Annotation: Human-annotated with bounding boxes for object detection
- Significance: Most successful CNN architectures (AlexNet, VGG, ResNet, Inception) were developed and tested on ImageNet
- Legacy: Established the foundation for modern computer vision and deep learning

Key Statistics:

- 1000 classes in ILSVRC challenge
- 1.2 million training images
- Standard input size: 224×224 pixels

Historical Impact:

- 2012: AlexNet breakthrough (15.3% top-5 error)
- Catalyst for deep learning revolution in computer vision

5. TRANSFER LEARNING

Quick Intro:

Transfer learning is a machine learning technique where a pre-trained model developed for one task is adapted and reused as the starting point for a related task, significantly reducing training time and data requirements.

Key Explanation:

- Core Concept: Leverage knowledge from pre-trained models on large datasets (like ImageNet)
- Advantages: Reduced training time, Better performance with limited data, Lower computational requirements, Faster convergence
- Frozen Layers: Retain general features (edges, textures)
- Trainable Layers: Adapt to task-specific features
- Applications: Medical imaging, satellite imagery, specialized domains with limited data

Process:

- 1. Start with pre-trained model
- 2. Identify transferable layers (usually early feature extraction layers)
- 3. Fine-tune or freeze layers based on target task
- 4. Train task-specific layers

Approaches:

- Feature Extraction: Freeze pre-trained layers, train only new classifier

- Fine-tuning: Update pre-trained weights with lower learning rate

Decision Framework:

- Small similar dataset → Freeze most layers
- Large similar dataset → Fine-tune more layers
- Different domain → Fine-tune extensively

6. ALEXNET

Quick Intro:

AlexNet is the groundbreaking 8-layer CNN architecture that won the 2012 ImageNet challenge, demonstrating the power of deep learning for image classification and sparking the deep learning revolution.

Key Explanation:

- Historical Significance: First CNN to win ImageNet (2012), reduced top-5 error from 26.2% to 15.3%
- Architecture: 8 layers total (5 convolutional + 3 fully connected)
- Parameters: ~60 million parameters
- Input Size: 227×227×3 RGB images
- Max Pooling: 3×3 with stride 2 after conv1, conv2, conv5

Key Innovations:

- ReLU Activation: First to use ReLU instead of sigmoid/tanh
- GPU Training: Utilized parallel GPU processing
- Dropout: Used dropout for regularization
- Data Augmentation: Extensive augmentation techniques

Layer Details:

- Conv1: 11×11 kernels, stride 4 (96 filters)
- Conv2: 5×5 kernels (256 filters)
- Conv3-5: 3×3 kernels (384, 384, 256 filters)
- FC layers: 4096, 4096, 1000 neurons

Legacy:

- Proved deep learning's effectiveness for computer vision
- Established CNN as standard for image classification
- Influenced all subsequent CNN architectures

7. VGG-16

Quick Intro:

VGG-16 is a 16-layer CNN architecture known for its simplicity and uniform structure, using only 3×3 convolutional filters throughout the network while achieving excellent performance on ImageNet.

Key Explanation:

- Architecture Philosophy: Simple, homogeneous design with small 3×3 filters
- Structure: 13 convolutional layers + 3 fully connected layers = 16 weight layers
- Parameters: ~138 million parameters
- Performance: 92.7% top-5 accuracy on ImageNet
- Model Size: 528 MB (large for deployment)

Key Features:

- Small Filters: Exclusively uses 3×3 convolutions
- Deep Network: 16 layers deep
- Uniform Design: Consistent architecture pattern
- Same Padding: Maintains spatial dimensions in conv layers

Block Structure:

- Block 1: 2×(3×3 conv, 64 filters) → MaxPool
- Block 2: 2×(3×3 conv, 128 filters) → MaxPool
- Block 3: 3×(3×3 conv, 256 filters) → MaxPool
- Block 4: 3×(3×3 conv, 512 filters) → MaxPool
- Block 5: 3×(3×3 conv, 512 filters) → MaxPool
- FC layers: 4096 → 4096 → 1000

Advantages:

- Simple, interpretable architecture
- Good feature extraction capabilities
- Excellent transfer learning performance

Limitations:

- Large model size and memory requirements
- Slow training due to many parameters
- Computational inefficiency

8. INCEPTION NET (GOOGLENET)

Quick Intro:

Inception Net (GoogLeNet) introduces the innovative Inception module that applies multiple convolution operations in parallel, enabling efficient multi-scale feature extraction while reducing computational cost.

Key Explanation:

- Core Innovation: Inception modules with parallel convolutions
- Architecture: 22 layers deep with 9 Inception modules
- Efficiency: Fewer parameters than AlexNet (4M vs 60M)
- Performance: Won ILSVRC 2014 with 6.67% top-5 error

Inception Module Components:

- 1×1 convolutions (dimensionality reduction)
- 3×3 convolutions (spatial features)
- 5×5 convolutions (larger spatial context)
- 3×3 max pooling (feature aggregation)
- Concatenation of all parallel outputs

Key Features:

- Multi-scale Processing: Captures features at different scales simultaneously
- 1×1 Convolutions: Reduces computational complexity
- Global Average Pooling: Replaces fully connected layers
- Auxiliary Classifiers: Help with gradient flow during training

Network Architecture:

- Initial convolutions and pooling
- 9 Inception modules (3a, 3b, 4a-4e, 5a, 5b)
- Global average pooling
- Final classification layer

Advantages:

- Computational efficiency
- Multi-scale feature detection
- Reduced overfitting
- Better gradient flow

Inception Versions:

- Inception v1 (GoogLeNet)
- Inception v2/v3 (batch normalization, factorized convolutions)
- Inception v4 (pure Inception)
- Inception-ResNet (with residual connections)

9. RESNET (RESIDUAL NETWORKS)

Quick Intro:

ResNet introduces skip connections (residual connections) that allow training of very deep networks by addressing the vanishing gradient problem, enabling networks with 50, 101, or even 152 layers.

Key Explanation:

- Core Innovation: Skip connections that create residual blocks
- Problem Solved: Vanishing/exploding gradients in deep networks
- Residual Block: $H(x) = F(x) + x$ where $F(x)$ is learned residual function and x is identity shortcut connection
- Mathematical Insight: Instead of learning $H(x)$ directly, learn $F(x) = H(x) - x$

Architecture Variants:

- ResNet-50: 50 layers
- ResNet-101: 101 layers
- ResNet-152: 152 layers

Residual Block Types:

- Basic Block: 2 conv layers with skip connection
- Bottleneck Block: 1×1 → 3×3 → 1×1 with skip connection

Key Benefits:

- Gradient Flow: Skip connections provide direct paths for gradients
- Identity Function: Network can learn identity if needed
- Deep Training: Enables training of very deep networks
- Performance: Better accuracy with increased depth

Network Structure:

- Initial 7×7 conv + max pool
- 4 stages of residual blocks
- Global average pooling
- Final FC layer

Impact:

- Breakthrough in training ultra-deep networks
- Foundation for many subsequent architectures
- Won ImageNet 2015 with 3.57% error

10. MOBILENET

Quick Intro:

MobileNet is a family of efficient CNN architectures designed for mobile and embedded devices, using depthwise separable convolutions to dramatically reduce computational cost while maintaining accuracy.

Key Explanation:

- Design Goal: Efficient CNNs for resource-constrained devices
- Core Innovation: Depthwise Separable Convolutions
- Computational Savings: 8-9× reduction in operations vs standard convolution

Depthwise Separable Convolution:

- 1. Depthwise Convolution: Apply single filter per input channel
- 2. Pointwise Convolution: 1×1 conv to combine channel outputs

MobileNet V1 (2017):

- Pure depthwise separable convolutions
- Width multiplier (α): Controls model width
- Resolution multiplier (ρ): Controls input resolution
- 28 layers total, 4.2M parameters

MobileNet V2 (2018):

- Inverted Residuals: Expand → Depthwise → Project
- Linear Bottlenecks: Remove ReLU from narrow layers
- ReLU6 Activation: Better for quantization
- Residual connections between bottleneck layers
- 3.4M parameters

MobileNet V3 (2019):

- Neural Architecture Search (NAS): Automated architecture optimization
- Squeeze-and-Excitation: Attention mechanism
- Hard Swish Activation: Efficient activation function
- Platform-aware NAS for hardware optimization
- Two variants: Large and Small

Key Parameters:

- Width Multiplier (α): Thins network uniformly (0.25, 0.5, 0.75, 1.0)
- Resolution Multiplier (ρ): Reduces input image resolution

Applications:

- Mobile computer vision
- Edge computing
- Real-time applications

- Object detection (MobileNet-SSD)
- Semantic segmentation

Trade-offs:

- Accuracy vs Efficiency
- Model size vs Speed
- Power consumption optimization

QUICK COMPARISON SUMMARY

Architecture	Year	Key Innovation	Layers	Parameters	Top-5 Error
AlexNet	2012	ReLU, GPU training	8	60M	15.3%
VGG-16	2014	Small 3×3 filters	16	138M	7.3%
Inception	2014	Multi-scale modules	22	4M	6.67%
ResNet-50	2015	Skip connections	50	25M	3.57%
MobileNet	2017	Depthwise separable	28	4.2M	Mobile-focused

KEY REVISION POINTS

1. CNN Evolution: AlexNet → VGG → Inception → ResNet → MobileNet
2. Common Trends: Deeper networks, efficiency improvements, architectural innovations
3. Transfer Learning: All these architectures serve as pre-trained models for transfer learning
4. ImageNet Impact: Drove development of all major CNN architectures
5. Trade-offs: Accuracy vs Efficiency vs Model Size vs Speed

End of Revision Notes

Remember: Practice implementing these architectures and understand their practical applications!