AlexNet Paper Summary (Sections 3, 4, and 5)

Section 3: Architecture

Familiar Concepts:

- **Convolutional Layers:** The model uses multiple convolutional layers to extract features from images.
- **Fully Connected Layers:** The last layers of the network perform classification.
- **Activation Functions:** The network uses activation functions to introduce non-linearity.
- **Pooling Layers:** Pooling reduces the spatial size of feature maps.

New Ideas Introduced:

1.

Use of ReLU Activation Function:

2.

   - ReLU (Rectified Linear Unit) speeds up training by avoiding the vanishing gradient problem.
   - It performs significantly better than traditional activation functions like tanh and sigmoid.

3.

Training on Multiple GPUs:

4.

   - The network is split across **two GPUs** to improve efficiency.
   - GPUs communicate only at specific layers, reducing overhead.

5.

Local Response Normalization (LRN):

6.

   - Normalizes neuron responses within a local neighborhood.
   - Encourages competition among neurons to improve generalization.

7.

Overlapping Pooling:

8.

- Instead of standard pooling, overlapping pooling is used to reduce overfitting.
- It decreases the error rate compared to non-overlapping pooling.

Section 4: Reducing Overfitting

Familiar Concepts:

- **Overfitting:** When a model memorizes training data instead of generalizing.
- **Data Augmentation:** A technique to artificially expand the training dataset.

New Ideas Introduced:

1.

Data Augmentation Techniques:

2.

- **Image Cropping & Flipping:** Random **224×224** patches are extracted from **256×256** images.
- **Color Intensity Modifications:** Principal Component Analysis (PCA) is used to alter brightness and color variations.

3.

Dropout Regularization:

4.

- Randomly disables neurons during training with a 50% probability.
- Prevents co-adaptation of neurons, forcing them to learn independent features.
- At test time, all neurons are used, but their outputs are scaled by 0.5.

Section 5: Details of Learning

Familiar Concepts:

- **Stochastic Gradient Descent (SGD):** Optimization algorithm for training deep networks.
- **Learning Rate:** Controls how much weights are updated in each step.
- **Momentum:** Helps accelerate convergence by smoothing updates.

New Ideas Introduced:

1.

Specific Learning Parameters:

2.

- **Mini-batch size:** 128 images per batch.
- **Momentum:** Set to 0.9 for stable updates.
- **Weight Decay (Regularization):** 0.0005 to prevent overfitting.

3.

Learning Rate Scheduling:

4.

- Starts at **0.01** and is reduced manually when validation accuracy plateaus.
- This ensures stable convergence.

5.

Weight Initialization:

6.

- Weights are initialized from a Gaussian distribution with zero mean.
- Biases in some layers are set to **1** instead of **0** to speed up early training.

Conclusion

AlexNet introduced several innovations, such as **ReLU activation, multiple GPUs, LRN, overlapping pooling, advanced data augmentation, dropout, and learning rate scheduling**. These techniques significantly improved deep learning performance and paved the way for modern CNN architectures.
