Terms:

Benchmarks: standardized datasets and evaluation metrics used to measure & compare the performance of different algorithms/models in a fair way

Visual Geometry Group nets: a series of deep convolutional neural networks (CNNs) known for their simple and uniform architecture (introduced in 2014)

Convolutional neural networks: A type of deep learning algorithm designed to analyze visual data like images & videos by automatically detecting spatial hierarchies of features (from low-level patterns to high-level representations). Inspired by the human visual cortes

Filters: Also known as kernels, are small matrices of weights that extract specific features like edges, corners, or textures. They work by the process of feature maps

Feature Maps: The output of a convolutional layer that highlights specific features (edges, shapes, or textures, in the input image). Each

Degradation Problem: Occurs when more layers to a neural network are added beyond a certain point that leads to higher training and testing errors. This results in decreased performance. The issue is not caused by overfitting, as it also affects training error.

Experiment Setup

Datasets:

1. ImageNet (ILSVRC 2012)
    - One of the largest & most influential benchmarks in computer vision
    - Contains over **1.2 million training images**, **50k validation images**, **100k test images**
    - 1,000 object categories (from dogs & cats to airplanes and keyboards)

Why it matters: If a method works on ImageNet, it's powerful enough for real world applications

2. CIFAR-10
    - A much smaller dataset used for testing idea quickly
    - Contains **50k training images, 10k test images, 10 classes**
    - Images are tiny (32*32 pixels)
    - It revealed fundamental problems with training deep networks

Why it matters: great for testing very deep networks to see how training behaves without the cost of ImageNet

Networks Tested:

1. Plain Networks (Baseline)
    - Inspired by Visual geometry groups (VGG) nets that contains a Stack of 3X3 convolution layers, with rules:
        - Same number of filters for same feature map size
        - When reducing the image size by half, double the filters (to keep computation balanced)

2. Residual Networks (ResNets)
    - Exactly the same structure, but with **shortcuts** every 2-3 layers
    - They also tested deeper ResNets: 50,101, & 152 layers (using bottleneck blocks to keep complexity manageable)

Metrics:

Top-1 error (ImageNet): The percentage of test images where the model's top guess is wrong
Top-5 -error (ImageNet: the percentage of test images where the correct answer is *not in the top 5 guesses*

Main Findings

ImageNet Results:

1. Plain Networks

- The 34-layer net had higher training & validation error than the 18-layer net
- UNEXPECTED - more layers should mean more power, but there made optimization worse (degradation problem)

2. Residual networks
    - 34-layer ResNet outperformed the 18-layer ResNet by %2.8
    - Training error was much lower = easier to optimize

CIFAR-10 Results

1. Plain Networks
    - As depth increased (20 to 50), training error got worse
    - Same degradation problem as ImageNet

2. Residual Networks:
    - Trained successfully at depths up to 110 layers
    - Accuracy improved with depth:
        - ResNet-20: **8.75% error**
        - ResNet-56: **6.97% error**
        - ResNet-110: **6.43% error**
    - Resnet-1202 (over 1000 layers!)
    - Training error <0.1% (model fit the data perfectly)
    - Test error - 7.93% = overfitting because CIFAR-10 is too small for such a massive net