Deep Learning for Large-Scale Image Recognition

Nemekhbayar Nomin[1],
[1]Computer Science Engineering Dep. of PNU.

1 Introduction

Deep convolutional neural networks (CNNs) have achieved significant breakthroughs in large-scale image recognition tasks. Two highly influential models: Very Deep Convolutional Networks (VGG) [3] and Deep Residual Networks (ResNet) [1] have set the standard for network architecture design and training deep models. Both models address the challenge of training deep networks while improving performance on key image datasets such as ImageNet and COCO. This paper presents a combined overview of these two models, highlighting their contributions to image recognition and addressing optimization challenges in training very deep architectures.

2 Deep Convolutional Networks

ConvNets have evolved with deeper architectures and smaller filter sizes, as demonstrated in the VGG models, which improve accuracy while maintaining computational efficiency. The best-performing submissions to the ILSVRC2013 [4] utilised smaller receptive window size and smaller stride of the first convolutional layer. Another line of improvements dealt with training and testing the networks densely over the whole image and over multiple scales [2]. Depth of the architecture design was addressed in this paper, and to increase the depth of the network by adding more convolutional layers, which is feasible due to the use of very small (3×3) convolution filters in all layers [3].

2.1 Network Architecture

During training, the input to our ConvNets is a fixed-size 224 × 224 RGB image with only preprocessing, the of mean RGB value, from each pixel. The image is passed through a stack of convolutional (conv.) layers, using 3 × 3 filters (which is the smallest size to capture the notion of left/right, up/down, center). The model includes 5 max-pooling layers, three fully connected (FC) layers: the first two with 4096 channels, and the third with 1000 channels for ILSVRC classification, followed by a softmax layer [3]. All hidden layers use ReLU as the activation function, and no Local Response Normalization (LRN) is applied, as it doesn't improve performance but increases memory and computation time.

2.2 Discussion

The ConvNet configurations, evaluated in [3] paper, are outlined in Table 1, one per column. In the following we will refer to the nets by their names (A–E). VGG configurations vary mainly in depth, with network A having 8 convolutional and 3 fully connected layers, while network E features 16 convolutional layers.

- VGG models demonstrated that increasing depth (up to 19 layers) leads to significant improvements in accuracy on image classification tasks. - Uniform design: Each convolutional layer uses 3x3 filters, and the network complexity is controlled by progressively doubling the number of filters as the feature map size reduces. - Improved generalization: VGG architectures generalize well to other visual recognition tasks such as object detection and localization. - The resulting ensemble of 7 networks has 7.3(configurations D and E), which reduced the test error to 7.0

3 Deep Residual Networks

Deep networks naturally integrate low/mid/high-level features and classifiers in an end-to-end multilayer fashion, and the "levels" of features can be enriched by the number of stacked layers (depth). Driven by the significance of depth, a question arises: *is learning better networks as easy as stacking more layers?* [1]

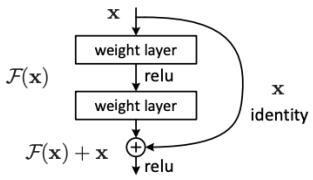

Figure 1: Figure 1. Residual learning: a building block

3.1 Residual Learning

In deep residual networks, instead of directly approximating complex functions (denoted as $H(x)$), the model learns a residual mapping, $F(x)$, defined as $F(x) := H(x) - x$. This reformulation means the model learns to approximate $F(x) + x$, which simplifies optimization. The reason behind this design is that learning residuals is often easier than approximating the original function $H(x)$ from scratch. This approach helps address the degradation problem, where adding more layers to deep networks increases training error, even though theoretically, deeper networks should not perform worse [1].

3.2 Identity Mapping by Shortcuts

In residual learning, shortcut connections are introduced that skip one or more layers and perform identity mappings. These shortcuts add the input x to the output $F(x)$ without adding extra parameters or computational complexity. The identity mapping helps the network maintain lower training errors, as it allows the network to use identity mappings when deeper layers do not improve the performance. For cases where input/output dimensions do not match, a linear projection is applied.

3.3 Network Architectures

Residual networks build upon the convolutional structure of VGG but introduce shortcut connections that bypass layers, which address the degradation problem and make deeper networks (up to 152 layers) easier to train. The residual networks use a series of convolutional layers (typically 3×3 filters) and include global average pooling and fully-connected layers for classification. The residual networks are easier to optimize and can achieve better performance with greater depth. In particular, a 34-layer residual network is less computationally expensive than a similarly deep VGG-19 network while still achieving better accuracy.

4 Comparison and Contributions to Image Recognition

Both VGG and ResNet models have significantly contributed to advancing the state of image recognition: - Depth: VGG networks pioneered the use of deep architectures (up to 19 layers), while ResNets extended this further to extreme depths (up to 152 layers and beyond) without degradation issues. - Performance: VGG networks achieved a top-5 error rate of 7.0- Generalization: Both models have shown strong generalization to tasks such as object detection, localization, and segmentation. ResNet, with its deeper architecture, demonstrated superior performance in these tasks, as seen in competitions like COCO 2015.

5 Conclusion

VGG and ResNet networks have been instrumental in shaping modern deep learning approaches for image recognition. VGG's simplicity and deep ar-

chitecture set a benchmark, while ResNet introduced residual learning to overcome optimization challenges in very deep networks. ResNets redefined the limits of network depth and inspired later architectures, such as DenseNet, which further exploit residual-like connections for efficient training of extremely deep models. The introduction of residual learning has become a cornerstone in modern deep network design.

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[2] Andrew G Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2014.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.