Paper Review: ImageNet Classification with Deep

Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

The paper "ImageNet Classification with Deep Convolutional Neural Networks," commonly known as AlexNet, is a groundbreaking work in computer vision. It presents a deep convolutional neural network (CNN) architecture that achieved significant performance improvements on the ImageNet dataset, demonstrating the potential of deep learning. This paper's primary contribution lies in introducing a large-scale CNN and training it using powerful GPUs, which had a profound impact on subsequent research in deep learning.

Motivation

The motivation behind this work was to address the limitations of traditional computer vision approaches that heavily relied on handcrafted features. The authors recognized the potential of deep neural networks for learning hierarchical representations directly from raw pixel data. They aimed to leverage the power of GPUs to accelerate the training process and demonstrate the effectiveness of deep CNNs for large-scale image classification tasks.

Architecture

AlexNet introduced a deep CNN architecture comprising multiple convolutional and fully connected layers. It consisted of five convolutional layers followed by three fully connected layers. The first convolutional layer learned 96 filters of size 11x11 with a stride of 4 pixels. The subsequent convolutional layers learned 256 filters of size 5x5. The pooling layers were applied after the first and second convolutional layers. The network employed the Rectified Linear Unit

(ReLU) activation function, which addressed the vanishing gradient problem and improved training speed. The ReLU function is defined as:

$$f(x) = \max(0, x)$$

To enhance the network's ability to generalize, the authors applied Local Response Normalization (LRN) after the ReLU nonlinearity. LRN helps normalize the responses within a local neighborhood of the activation maps, promoting competition between different feature maps. This normalization is mathematically represented as:

$$b^i_{x,y} = a^i_{x,y} \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^{-\beta}$$

where $a^i_{x,y}$ denotes the activity of a neuron computed by applying the ReLU nonlinearity, $b^i_{x,y}$ represents the normalized activity, and n, k, α, and β are hyperparameters.

Overlapping pooling was employed after the LRN layer to reduce spatial dimensions. This approach helped improve translation invariance and extract more informative features from the data.

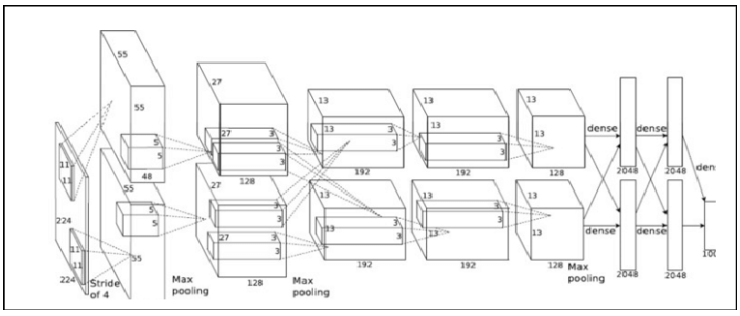

Figure 1: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000[1].

Algorithm and Learning Details

The learning algorithm used for training AlexNet was stochastic gradient descent (SGD) with a mini-batch size of 128. The network was trained on 1.2 million high-resolution images from the ImageNet dataset, labeled with 1,000

object categories. The training process was accelerated using two NVIDIA GTX 580 GPUs. The authors employed data parallelism, dividing the model across the GPUs, to speed up training. The network was trained for approximately six days, employing data augmentation techniques such as cropping and horizontal flipping to increase the size of the training set and reduce overfitting.

The objective function used during training was the softmax loss function, which measures the dissimilarity between predicted class probabilities and the ground truth labels. The softmax loss function is given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)$$

where N represents the number of training samples, y_i is the ground truth label for the ith sample, and f_j denotes the predicted score for class j.

Results

AlexNet achieved a top-5 error rate of 15.3 in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012, surpassing the previous state-of-the-art by a considerable margin. This breakthrough performance demonstrated the power of deep CNNs for image classification tasks and sparked a revolution in the field of computer vision. The success of AlexNet showcased the potential of deep learning in various application domains and inspired further research in neural network architectures and training methodologies.

Conclusion

In conclusion, the paper "ImageNet Classification with Deep Convolutional Neural Networks" (AlexNet) marked a significant milestone in the AI revolution. Its introduction of a deep CNN architecture and utilization of GPUs for training revolutionized the field of computer vision. AlexNet's success motivated further research, leading to the development of more sophisticated models that surpassed its performance on various visual recognition tasks.

In the years following AlexNet, numerous advancements have been made in deep learning for computer vision. Notable models include VGGNet, which introduced deeper architectures with smaller filter sizes, and GoogLeNet, which pioneered the use of inception modules for improved efficiency. The advent of residual connections in the ResNet architecture further improved the training of extremely deep networks, and DenseNet introduced densely connected layers to facilitate feature reuse. More recently, models such as EfficientNet and Vision Transformers (ViTs) have achieved impressive results with optimized model architectures and attention mechanisms[2],[3],[4].

Bibliography

1. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

2. Tan, M., Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105-6114).

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Hinton, G. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

4. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-9).