ImageNet Classification with Deep Convolutional Neural Networks
Authors: Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton

Introduction

Object Recognition is one of the primary tasks in computer vision, for the purpose of which majorly only labeled datasets containing close to thousands of images like NORB, CIFAR were being used. However, considering the variability of the setting in the outside world, datasets with much bigger sizes and variability are required. It was around the publication of this work that new larger datasets with high resolution images, such as LabelMe, ImageNet, etc. were introduced. However, aside from the size of the datasets, the neural networks used for the task of object recognition should also possess a lot of prior knowledge to compensate for the data not encompassed by the dataset. This further explains the need for convolutional neural networks, which have fewer connections compared to similarly-sized feedforward networks, but make stronger assumptions about the nature of images.

However, CNNs have proven to be computationally expensive to work with high-resolution images. The authors of this work propose a highly-optimized implementation of 2D convolutions which helps address this problem. The authors train a large network containing five convolutional and three fully-connected layers on the subsets of ImageNet used in ILSVRC-2010 and ILSVRC-2012 and achieve state-of-the-art results on them. This large size of the network also seemed to handle the problem of overfitting.

Network Architecture

The ImageNet dataset traditionally has over 15 million high-resolution images with over 22,000 labels. The authors in this work used around 1000 images in each of the 1000 categories, training on 1.2 million images, with 50,000 images for validation and 150,000 for testing. The authors also used this network in the ILSVRC-2012 competition and reported top-1 and top-5 error rates, wherein the top-5 error rate is when the label is not among the five labels considered most probable for the model. As a preprocessing step, the authors downsampled the images to a fixed resolution of 256 X 256, when the images were squared in shape; and cropped out the central 256 X 256 patch from the image resulted by rescaling the shorter side to a length of 256.

The network used by the authors consisted of five convolutional layers and three fully-connected layers. Training over gradient descent on non-linear activation functions such as sigmoid and tanh lead to higher training times as compared to Rectified Linear Units (ReLU). Thus, ReLU was used in this architecture. The authors observe a local normalization scheme on the inputs which helps them achieve improvement in the performance on CIFAR dataset. Futhermore, the authors use overlapped pooling which helped them improve the performance, and also prevent overfitting.

The most common method used to reduce overfitting is by artificially increasing the dataset using label-preserving transformations. The authors in this work employ two types of data augmentation. In one technique, the authors generate image translations and horizontal reflections, increasing the size of the training set by a factor of 2048. In another technique, the authors perform PCA on the set of RGB pixel values throughout the ImageNet training set, and add the multiples of the found principal components with the magnitudes corresponding to eigen values of the RGB pixel values. The authors also use dropout in the first two fully-connected layers. This helped prohibit overfitting substantially.

Results

In Table 1 in the paper, the authors summarize the performance of their architecture on the ILSVRC-2010 dataset, achieving 8.2% and 8.7% improvements on the Top-1 and Top-5 error rates. The authors entered the architecture in the ILSVRC-2012 competition, the dataset of which didn't have test labels, and for which they record the validation set errors for the Top-1 and Top-5 error rates. The authors also use this architecture on the ImageNet (Fall 2009 version) dataset, which doesn't have a train-test split, using the split commonly used by the previous authors, and achieve Top-1 and Top-5 error rates of 67.4% and 40.9% respectively, while the best published results prior to this were 78.1% and 60.9%.

Strengths & Weaknesses

The work in this paper is considered to be a milestone of CNN for image classification. Many methords introduced such as overlapped pooling, ReLU for CNNs etc are still the standard for computer vision. This paper proved that theoretically, the complexity of visual patterns can be extracted by adding more convolutional layers. However, AlexNet is not deep enough as compared to it's successors such as VGGNet, GoogleNet, ResNet, etc. Also, a large convolutional filter such as 5 X 5 was used in this network architecture, whereas, it has been proven to be non-beneficial in future research.

Further Work

Convolutional neural networks (CNNs) have accomplished astonishing achievements across a variety of domains. With the advent of better GPUs, several deep convolutional neural networks, as an extension to this and convolutional networks such as VGG16 [1], it's versions, ResNet50 and it's versions[2], Inception V3 [3], GoogleNet[4], etc. have been introduced for the purpose of object recognition.

References

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

[2] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[4] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.