A Comparison on Performance of Deep Convolutional Neural Networks: AlexNet and VGG on Image Classification

Duc-Cuong Dao
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
duccuong.hust@gmail.com

Olivier Morére
Agency for Science, Technology and Research (A*STAR)
Singapore
olivier.morere@gmail.com

ABSTRACT

We describe the state-of-the-art architecture of Convolutional Neural Networks for image classification tasks.

Keywords

Deep learning, convolutional neural networks, image classification

1. INTRODUCTION

Since their introduction by in the early 1990's, Convolutional Neural Networks (ConvNets) have demonstrated excellent performance at task such as hand-written digit classification and face detection. Recently, serveral papers have shown that they can also deliver outstanding performance on more challenging visual classfication tasks. [Ciresan et al., 2012] demonstrate state-of-the-art performance on NORB and CIFAR-10 datasets. Most notably, [Krizhevsky et al., 2012] show record beating performance on the ImageNet 2012 classification benchmark, with their ConvNet model achieving an error rate of 16.4%, compared to the 2nd place result of 26.1%. Several factors are responsible for this renewed interest in ConvNet models: (i) the availibility of much larger training sets, with millions of labeled examples; (ii) powerful GPU implementations, making the traininPOKg of very large models practical and (iii) better model regularization, such as Dropout [Hinton et al., 2012]

2. RELATED WORK

Starting with LeNet-5 [1], ConvNets have typically had a standard structure - stacked convolutional layers (optionally followed by contrast normalization and max-pooling) are followed by one or more fully-connected layers. Variants of this basic design are prevalent in the image classification literature and have yielded the best results to-date on MNIST, CIFAR and most notably on the ImageNet classification challenge. For larger datasets such as Imagenet, the recent trend has been to increase the number of layer and layer size, while using dropout [7] to address the problem of overfitting.

Despite the concerns that max-pooling layers result in loss of accurate spatial information, the same convolutional network archiecture as [9] has aslo been sucessfully employed for localization, object detection and human post estimation. Inspire by a neuroscience model of the primate visual cortex, .. use a series of fixed Gabor filters of different sizes in order to handle multiple scales, similarly to the Inceptition model. However, contrary to the fixed 2-layer model of [15], all filters in the Inception model are learned. Furthermore, Inception layers are repeated many time, leading to a 22-layer deep model in the case of the GoogLeNet model.

Network-in-Network is an approach proposed by Lin et al. in order to increase the respresentational power of neural newtworks. When appliead to convolutional layers, the method could be viewd as addition $1x1$ convolutional layers follwed typically by the rectified linear activation. This enables it to be easily integrated in the current ConvNet pipelines. We use this approach heavil in our architecture.

3. BACKGROUND

We begin by providing a rough overview of deep learning and its application to various types of problem in machine learning and describe the architecture of typical ConvNet. After that, we describe the traditional learning strategies of ConvNets for a particular problem. Although ConvNets can be trained for unspervised learning tasks, we concentrate on the supervised learning as our target problem (image classification) falls into that family (Need to paraphase)

3.1 Deep Learning

Deep Learning is a branch of machine learning that allows computational models to learn complicated and abstract representation of features.

3.2 Convolutional Neural Networks

ConvNets are hierarchical neural networks whose convolutional layers alternate with subsampling layers, reminiscent of simple and complex ceels in the primary visual cortex. ConvNets vary in how convolutional and subsample layer are realized and how they are trained.

SoICT 2015 Hue, Vietnam

3.2.1 Convolutional Layer

A convolutional layer is parametrized by the size and the number of maps, kernel sizes, skipping factors and the connection table. Each layer has M maps of equal size (M_x, M_y). A kernel of size (K_x, K_y) is shifted over the valid region of the input image (i.e. the kernel has be completely inside the image). The skipping factors S_x and S_y define how many pixels the filter/kernel skips in x- and y-direction between subsequent convolutions. The size of the output maps is then defined as:

$$M_x^n = \frac{M_x^{n-1} - K_x^n}{S_x^n + 1} + 1; M_y^n = \frac{M_y^{n-1} - K_y^n}{S_y^n + 1} + 1 \qquad (1)$$

where index n indicates the layer. Each map in layer L^n is connected to at most M^{n-1}. Neurons of a given map share their weights but have different receptive fields.

3.2.2 Pooling Layer

The biggest architectural difference between our implementation and the ConvNet of [leCun et al] is the use of a max-pooling layer instead of a sub-smapling layer. No such layer is used by [Simard et al..] who simply skips nearby pixels prior to convolution, instead of pooling or averaging.

3.2.3 Classification Layer

4. EXPERIMENTS

4.1 The Dataset

We used a crawler to collect ship images from different sources along with their meta information (name, category, etc.).

4.2 Training CNNs for Image Classification and Deep Features

4.3 Experimental Results

5. CONCLUSIONS

6. ACKNOWLEDGMENTS

We would like to thank Assoc. Prof., Stéphane Bressan at School of Computing, National University of Singapore for insightful supervision during the work of this paper and to Dr. Antoine Veillard at Image Persuasive Lab (IPAL), Insitute of Infocomm Research (I2R), A*STAR for valuable discussions.

References

[1] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan-Kaufmann, 1990.