Very Deep Convolutional Networks for Large-Scale Image Recognition

Cheng Guan

August 9, 2018

Abstract

Today, after I know about deep learning deeper, I want to learn some classic convolutional network. This is a paper on VGG16. In this work the authors investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Their main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3 × 3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 1619 weight layers. These findings were the basis of their ImageNet Challenge 2014 submission, where their team secured the first and the second places in the localisation and classification tracks respectively. They also show that their representations generalise well to other datasets, where they achieve state-of-the-art results. They have made two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.

1. Introduction

Convolutional networks (ConvNets) have recently enjoyed a great success in large-scale image and video recognition [2, 6, 4, 5] which has become possible due to the large public image repositories, such as ImageNet, and high-performance computing systems, such as GPUs or large-scale distributed clusters. In particular, an important role in the advance of deep visual recognition architectures has been played by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [3], which has served as a testbed for a few generations of large-scale image classification systems, from high-dimensional shallow feature encodings (the winner of ILSVRC-2011) to deep ConvNets (the winner of ILSVRC-2012).

With ConvNets becoming more of a commodity in the computer vision field, a number of attempts have been made to improve the original architecture of [2] in a bid to achieve better accuracy. For instance, the best-performing submissions to the ILSVRC- 2013 utilized 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.

The authors come up with significantly more accurate ConvNet architectures, which not only achieve the state-of-the-art accuracy on ILSVRC classification and localisation tasks, but are also applicable to other image recognition datasets, where they achieve excellent performance even when used as a part of a relatively simple pipelines

2. Convnet Configuration

To measure the improvement brought by the increased ConvNet depth in a fair setting, all their ConvNet layer configurations are designed using the same principles, inspired by [1, 2]. In this section, the authors first describe a generic layout of their ConvNet configurations.

2.1. Architecture

During training, the input to their ConvNets is a fixed-size 224 × 224 RGB image. The only pre-processing they do is subtracting the mean RGB value, computed on the training set, from each pixel. The image is passed through a stack of convolutional (conv.) layers, where they use filters with a very small receptive field: 3 × 3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations they also utilize 1 × 1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1 pixel for 3 × 3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2 × 2 pixel window, with stride 2.

A stack of convolutional layers (which has a different depth in different architectures) is followed by three Fully-Connected (FC) layers: the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the softmax layer. The configuration of the fully connected layers is the same in all networks.

References

[1] D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *IJCAI*, 2011. 1

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet large scale visual recognition challenge. *IJCV*, 2015. 1

[4] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 1

[5] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[6] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1