Going Deeper with Convolutions

Cheng Guan

June 14, 2018

1 Introduction

In the last three years, our object classification and detection capabilities have dramatically improved due to advances in deep learning and convolutional networks . One encouraging news is that most of this progress is not just the result of more powerful hardware, larger datasets and bigger models, but mainly a consequence of new ideas, algorithms and improved network architectures. No new data sources were used, for example, by the top entries in the ILSVRC 2014 competition besides the classification data set of the same competition for detection purposes. Our GoogLeNet submission to ILSVRC 2014 actually uses 12 times fewer parameters than the winning architecture of Krizhevsky *et al* [2] from two years ago, while being significantly more accurate. On the object detection front, the biggest gains have not come from naive application of bigger and bigger deep networks ,but from the synergy of deep architectures and classical computer vision,like the R-CNN algorithm by Girshick *et al* [3].

Another notable factor is that with the ongoing traction of mobile and embedded computing, the efficiency of our algorithmsCespecially their power and memory useCgains importance. It is noteworthy that the considerations leading to the design of the deep architecture presented in this paper included this factor rather than having a sheer fixation on accuracy numbers. For most of the experiments,the models were designed to keep a computational budget of 1.5 billion multiply-adds at inference time, so that the they do not end up to be a purely academic curiosity,but could be put to real world use, even on large datasets, at a reasonable cost.

2 Related Work

Starting with LeNet-5 [3], convolutional neural networks (CNN) have typically had a standard structure C 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 layers and layer size ,while using dropout to address the problem of overfitting.

Despite concerns that max-pooling layers result in loss of accurate spatial information, the same convolutional network architecture as [3] has also been successfully employed for localization,object detection [1] and human pose estimation [4].

Figure 1: Two distinct classes from the 1000 classes of the ILSVRC 2014 classification challenge. Domain knowledge is required to distinguish between these classes.

Inspired by a neuroscience model of the primate visual cortex, Serre used a series of fixed Gabor filters of different sizes to handle multiple scales. We use a similar strategy here. However, contrary to the fixed 2-layer deep model , all filters in the Inception architecture are learned. Furthermore, Inception layers are repeated many times, leading to a 22-layer deep model in the case of the GoogLeNet model.

3 Motivation and High Level Considerations

The most straightforward way of improving the performance of deep neural networks is by increasing their size. This includes both increasing the depthCthe number of net work levels C as well as its width: the number of units at each level. This is an easy and safe way of training higher quality models, especially given the availability of a large amount of labeled training data. However, this simple solution comes with two major drawbacks.

Bigger size typically means a larger number of parameters,which makes the enlarged network more prone to overfitting, especially if the number of labeled examples in the training set is limited. This is a major bottleneck as strongly labeled datasets are laborious and expensive to obtain,often requiring expert human raters to distinguish between various fine-grained visual categories such as those in ImageNet (even in the 1000-class ILSVRC subset) as shown in Figure 1.

References

[1] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *CVPR*, 2014. 1

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

[3] R.B.Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[4] T.Serre, L.Wolf, S.M.Bileschi, M.Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE TPAMI*, 2007. 1