Going Deeper with Convolutions

Lina Zhu

June 20 ,2018

Abstract

We propose a deep convolutional neural network architecture codenamed Inception, which implements the newest technology for classifying and detecting ImageNet in the 2014 Large-scale Visual Recognition Challenge(ILSVRC2014). The main sign of this building is to increase the utilization of computing resources in the network. With a well-designed design, we have increased the depth and width of the network while keeping the calculation budget unchanged. In order to optimize quality, these decisions of architecture are based on the intuition of the Hebrew principle and multi-scale processing. A 22-deep network whose quality is assessed in terms of classification and detection.

1. Introduction

In the past few years, our object classification and detection has improved significantly in deep learning and convolutional networks due to advancements[3]. An encouraging news is that most of this progress is not just for more powerful hardware, larger dataset results and larger models, but mostly as a result of new ideas, algorithms and improved network architecture. No new data source is used. In terms of object detection, the biggest benefits come not from larger nave applications and deeper networks, but from deeper synergetic architectures and classical computer vision. Another notable factor is the continuous traction movement and embedded computing, and our efficiency algorithm is particularly important for its power and memory usage benefits. It is worth noting that considering the design of the deep architecture leading up to this article includes this factor, rather than simply fixing the exact figures. For most experiments, the model aims to keep the 1.5 billion calculation budget multiplied by the inferred time so that they will not end purely academic curiosity but can be put into practice even on large datasets. The cost is used worldwide. In this article, we will focus on an efficient deep-neuron computer vision network architecture. The benefits of this architecture are to perform experimental verification and detection challenges. It is significantly superior to the state of the art.

2. Related Work

Starting from LeNet-5 [3], convolutional neural networks (CNNs) usually have a standard structure and a stacked convolutional layer followed by one or more fully connected layers. This basic design variation is common in image classification literature. For large data sets such as ImageNet, the recent trend has been to increase the number of layers and the number of layers while using dropout to solve the problem of excessive coordination. Although it is feared that the largest merging layer will result in loss of exact spatial information, the same convolutional network architecture[2] has also been successfully applied for positioning[2], object detection [1]and human posture estimation. Inspired by the primate visual neuroscience model cortex. In order to increase the representative strength of the neural network. In their model, an additional 1 x 1 convolutional layer is added to the network, increasing its depth. We use this method extensively in our architecture. However, in our setup, 11 convolution has two purposes: at most critical, they are mainly used as dimension reduction modules to eliminate computational bottlenecks, otherwise we limit the size of our network. This is not allowed just increase the depth, but also increase the width of our network without significant performance loss. Finally, the state of the art for object detection is the area with a convolutional neural network.

3. Motivation and High Level Factors

The most direct way to improve performance is to deepen neural networks by increasing their size. This includes increasing the depth and the number of network layers and the width: the number of units per level. This is a simple and safe training method quality model, especially considering the amount of training data available for usability annotation. But this simple solution has two major drawbacks. A larger size usually means more parameters, which makes the expanded network easier to overuse and fit, es-

Table 1. Classification performance.

Team	Year	Place	Error(top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

pecially if the number of marked instances is limited in the training set. This is a serious bottleneck. Tagged datasets are often laborious and costly. Expert human evaluators are required to distinguish between various evaluators' fine-grained visual categories. For example, the visual category in ImageNet is shown. Another disadvantage of uniformly increasing the network is the significant increase in the use of computing resources. For example, in a deep visual network, if there are two convolutional layers being linked, any number of filters that uniformly increase theirs results in a second increase in calculations. If the increased capacity is used inefficiently, most of the calculations are wasteful. As a computational budget is always limited, it is the effective allocation of resources for calculations that takes precedence over indiscriminate increases in scale, even if the primary goal is to improve the performance of the quality.

4. ILSVRC 2014 Classification Challenge Settings and Results

The ILSVRC 2014 classification challenge involves the task of classifying images into one of 1000 leaf node categories in the ImagNet hierarchy. There are approximately 1.2 million training images, validating images of 50,000 and 100,000 images for testing. Each image is associated with a ground truth class, and performance is based on the highest scored classifier prediction. Two numbers usually report: the accuracy of the first one, which compares the base facts with the first predicted category, and the error rate of the top five, which compares the basic facts to the top five predicted categories: images are viewed For if the ground truth is in the top-5, regardless of their rank. Challenge the use of the top-5 error rates for ranking purposes.

In the rest of this article, we analyzed a number of factors that contribute to the overall performance of the finals. Our final submission challenge received the top-5 error verification and test data of 6.67%, ranking first among other participants. This year's best method (Clarifai), both use external data for training classifiers. Table 1 shows some of the best performing method years in the statistics for the past three years.

5. Conclusion

Our findings yielded a solid piece of evidence showing that close to optimally sparse structural building blocks through easy-to-obtain denseness is a viable approach to improving neural networks for computer vision. The main advantage of this approach is that the computational requirements for significantly improved quality in the case of moderate increases are shallower and narrower architectures.

References

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

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 1

[3] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 2014. 1