Going deeper with convolutions

1 Introduction

In our case, the word “deep” is used in two different meanings: first of all, in the sense that we introduce a new level of organization in the form of the “Inception module” and also in the more direct sense of increased network depth.

2 Related Work

Starting with LeNet-5 [10], convolutional neural networks (CNN) have typically had a standard

structure – stacked convolutional layers (optionally followed by contrast normalization and maxpooling) are followed by one or more fully-connected layers.

When applied to convolutional layers, the method could be viewed as additional 1X1 convolutional layers followed typically by the rectified linear activation. This enables it to be easily integrated in the current CNN pipelines. We use this approach heavily in our architecture. However, in our setting, 1X1 convolutions have dual purpose: most critically, they are used mainly as dimension reduction modules to remove computational bottlenecks, that would otherwise limit the size of our networks. This allows for not just increasing the depth, but also the width of our networks without significant performance penalty.

R-CNN decomposes the overall detection problem into two subproblems: to first utilize low-level cues such as color and superpixel consistency for potential object proposals in a category-agnostic fashion, and to then use CNN classifiers to identify object categories at those locations.

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 depth – the number of levels – of the network and its

width: the number of units at each level. 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.

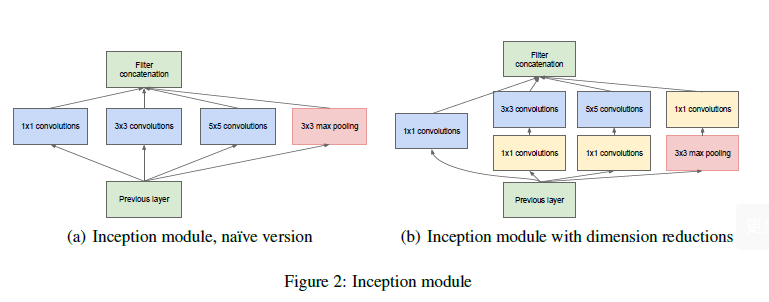
Another drawback of uniformly increased network size is the dramatically increased use of computational resources.

4 Architectural Details

The main idea of the Inception architecture is based on finding out how an optimal local sparse

structure in a convolutional vision network can be approximated and covered by readily available

dense components.



This leads to the second idea of the proposed architecture: judiciously applying dimension reductions

and projections wherever the computational requirements would increase too much otherwise.

This is based on the success of embeddings: even low dimensional embeddings might contain a lot

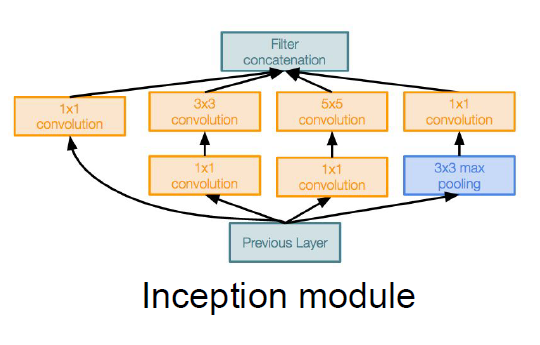
of information about a relatively large image patch.

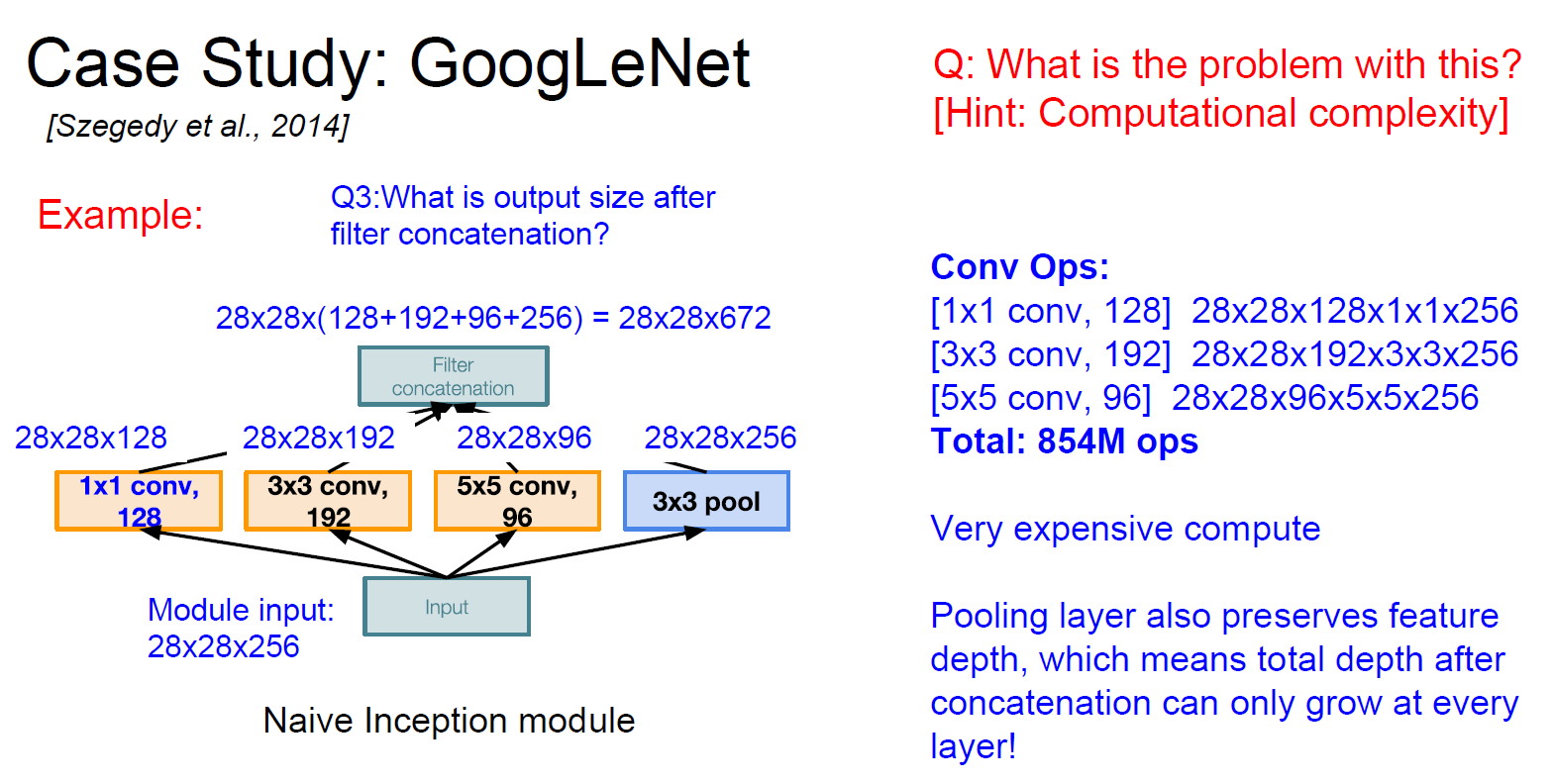
That is, 1X1 convolutions are used to compute reductions before the expensive 3X3 and 5X5 convolutions.

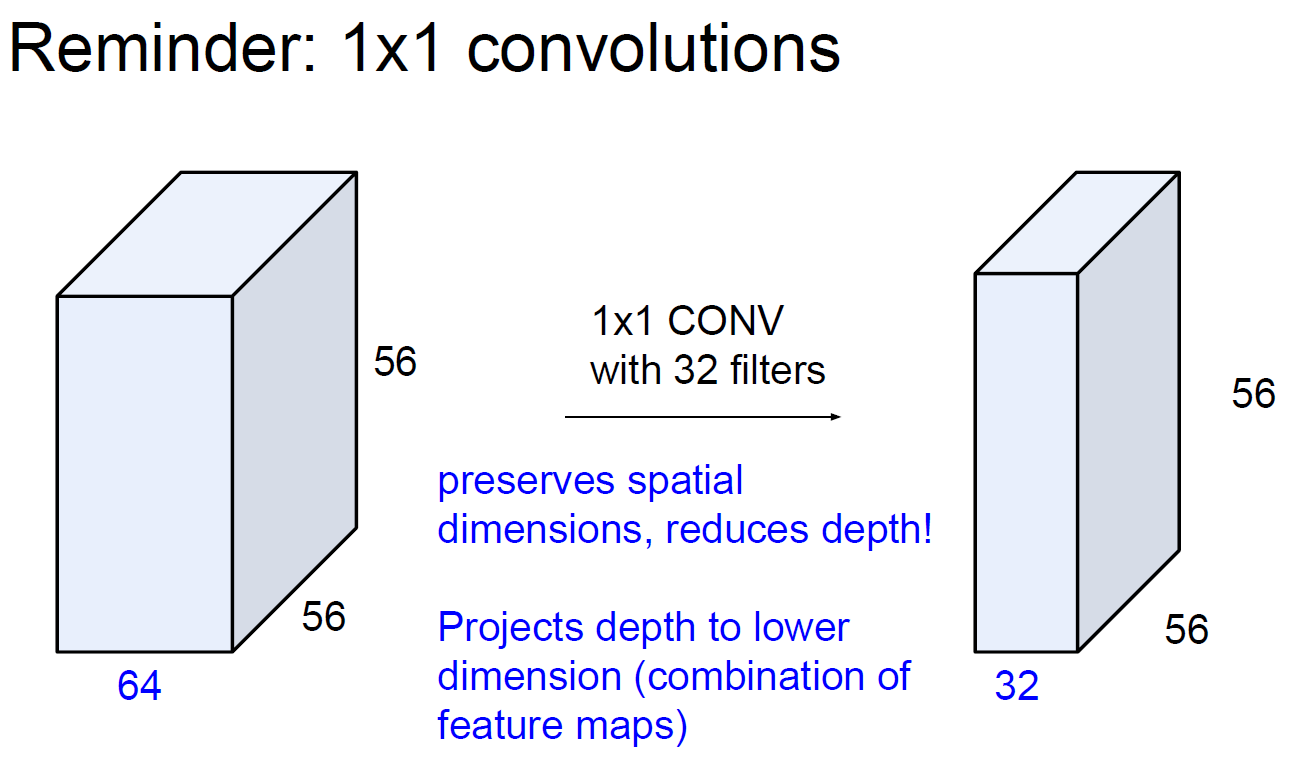
One of the main beneficial aspects of this architecture is that it allows for increasing the number of

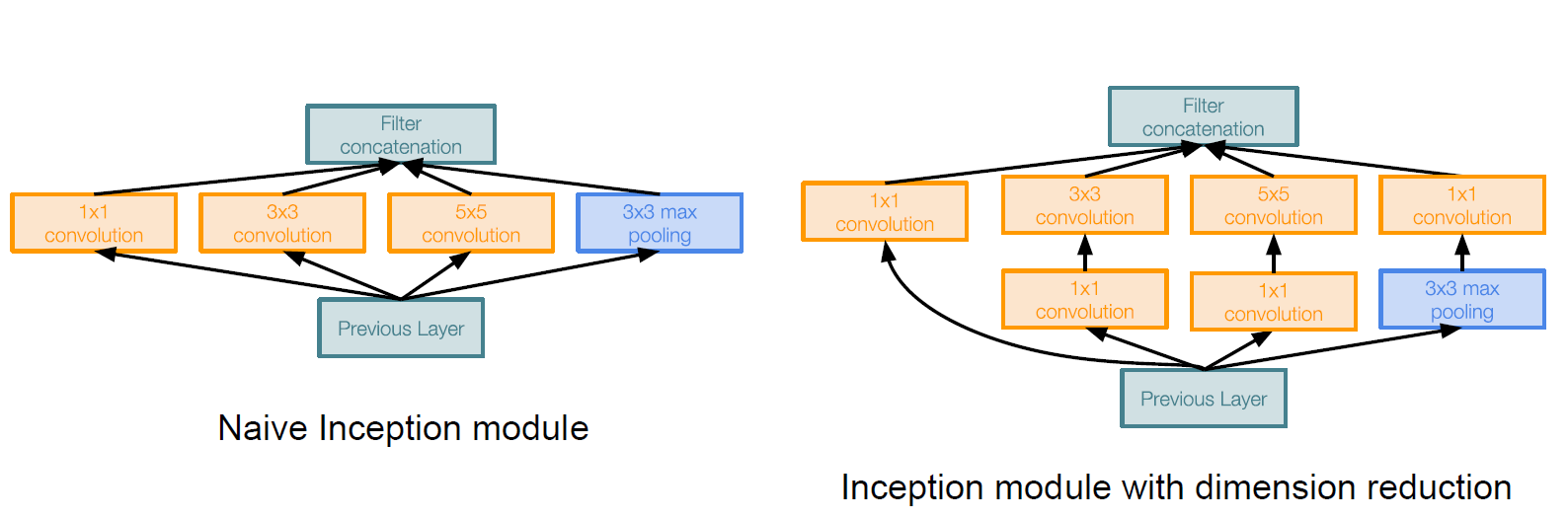
units at each stage significantly without an uncontrolled blow-up in computational complexity. Another practically useful aspect of this design is that it aligns with the intuition that visual information should be processed at various scales and then aggregated so that the next stage

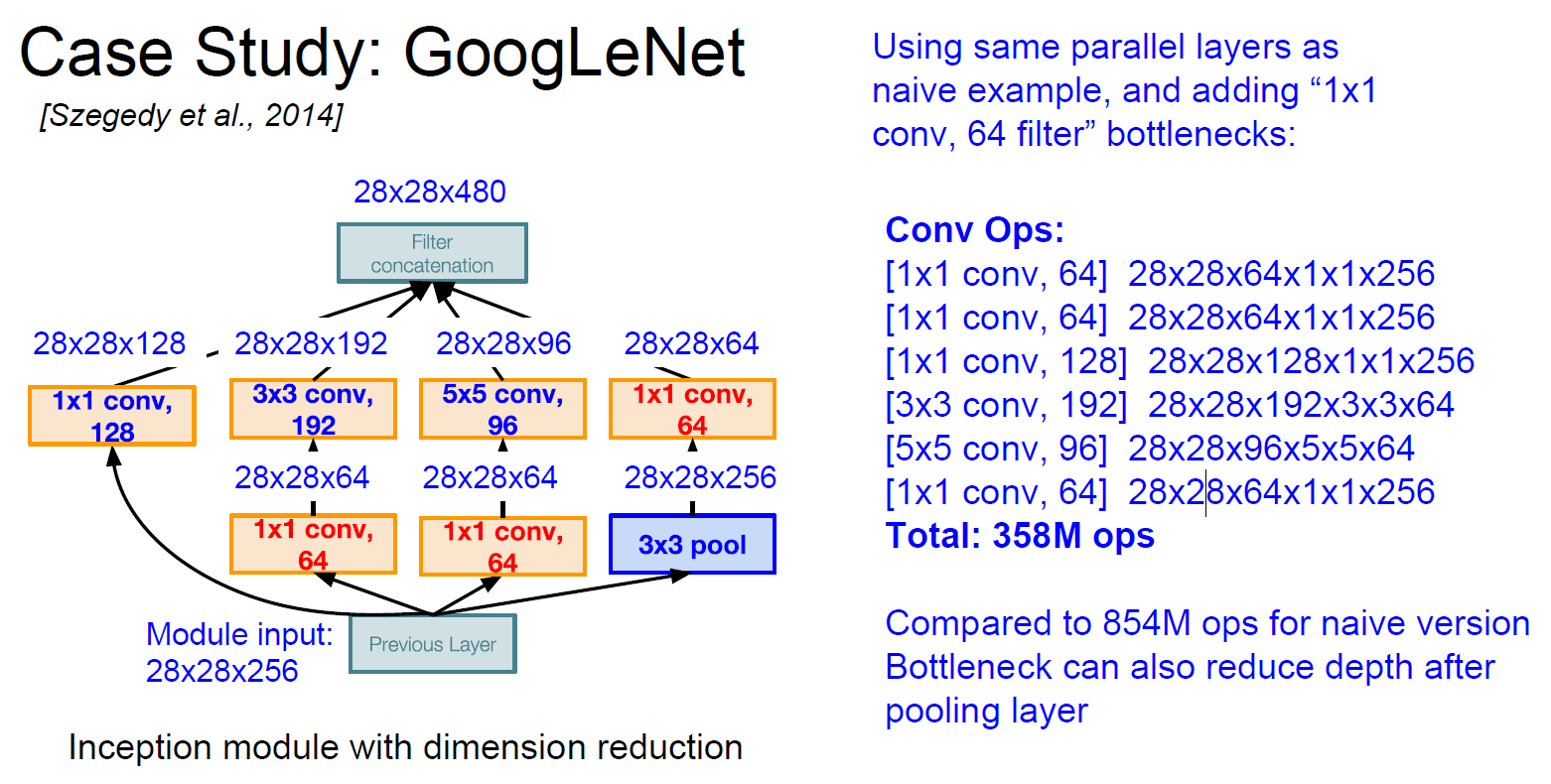
can abstract features from different scales simultaneously.



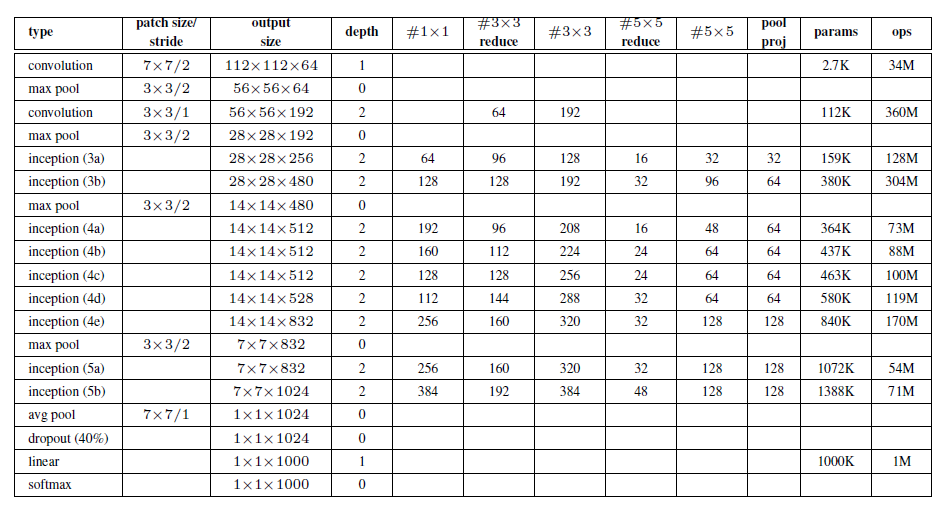


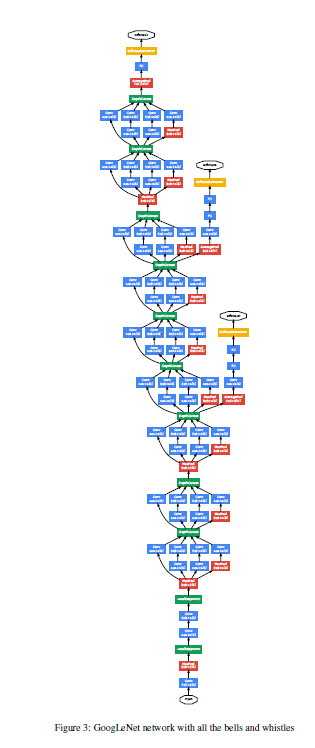






5 GoogLeNet





6 Training Methodology

7 ILSVRC 2014 Classification Challenge Setup and Results

The ILSVRC 2014 classification challenge involves the task of classifying the image into one of

1000 leaf-node categories in the Imagenet hierarchy. There are about 1.2 million images for training,

50,000 for validation and 100,000 images for testing. Each image is associated with one ground

truth category, and performance is measured based on the highest scoring classifier predictions.

Two numbers are usually reported: the top-1 accuracy rate, which compares the ground truth against

the first predicted class, and the top-5 error rate, which compares the ground truth against the first

5 predicted classes: an image is deemed correctly classified if the ground truth is among the top-5,

regardless of its rank in them. The challenge uses the top-5 error rate for ranking purposes.

1. We independently trained 7 versions of the same GoogLeNet model (including one wider

version), and performed ensemble prediction with them. These models were trained with

the same initialization (even with the same initial weights, mainly because of an oversight)

and learning rate policies, and they only differ in sampling methodologies and the random

order in which they see input images.

2. During testing, we adopted a more aggressive cropping approach.

3. The softmax probabilities are averaged over multiple crops and over all the individual classifiers

to obtain the final prediction. In our experiments we analyzed alternative approaches

on the validation data, such as max pooling over crops and averaging over classifiers, but

they lead to inferior performance than the simple averaging.

8 ILSVRC 2014 Detection Challenge Setup and Results

The ILSVRC detection task is to produce bounding boxes around objects in images among 200

possible classes. Detected objects count as correct if they match the class of the ground truth and

their bounding boxes overlap by at least 50% (using the Jaccard index). Extraneous detections count

as false positives and are penalized. Contrary to the classification task, each image may contain many objects or none, and their scale may vary from large to tiny. Results are reported using the mean average precision (mAP).

9 Conclusions

Our results seem to yield a solid evidence that approximating the expected optimal sparse structure

by readily available dense building blocks is a viable method for improving neural networks for

computer vision. The main advantage of this method is a significant quality gain at a modest increase

of computational requirements compared to shallower and less wide networks. Also note that

our detection work was competitive despite of neither utilizing context nor performing bounding box

regression and this fact provides further evidence of the strength of the Inception architecture. Although it is expected that similar quality of result can be achieved by much more expensive networks of similar depth and width, our approach yields solid evidence that moving to sparser architectures is feasible and useful idea in general. This suggest promising future work towards creating sparser and more refined structures in automated ways on the basis of [2].