# Review on Going Deeper with Convolutions

## 1 Abstract

Researchers proposed a 22-layered deep network, GoogLeNet which is able to gain quality in the context of classification and detection. This model allows us to increase the depth and width of the network while keeping the **computational budget constant.**

## 2 Introduction

GoogLeNet used 12x fewer computational power than Krizhevsky [3] (the winner of ILSVRC2012) and it was significantly more accurate. The model is able to keep the computational budget of 1.5 billion multiply-adds at an inference time, which could be proved to real-world use.

## 3 Dataset Description

They used a dataset from ***ImageNet*** *Large Scale Visual Recognition Challenge 2014*, where 1.2 million images were used for training, 50,000 for validation, and 100,000 for testing.

1. The task of classifying the image into one of 1000 leaf-node categories in the Imagenet hierarchy.
2. The ILSVRC detection task is to produce bounding boxes around objects in images among 200 possible classes.

## 4 Related Work

1. To remove the computational bottleneck, GoogLeNet uses an additional 1x1 convolutional layers, followed by rectified linear activation. [4]
2. GoogLeNet uses a similar pipeline used in R-CNN proposed by Girshick [2], but explored but have explored enhancements in both stages, such as multi-box [1] prediction for higher object bounding box recall, and ensemble approaches for better categorization of bounding box proposals.

## 5 Motivation and High-Level Consideration

### 5.1 Two Drawbacks

1. 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.
2. Another drawback of uniformly increased network size is the dramatically increased use of computational resources.

### 5.2 Proposed Ideas

1. The fundamental way of solving both issues would be by ultimately moving from fully connected to sparsely connected architectures, even inside the convolutions.

2. Using non-uniform sparse data structures is not a good approach, it's better to use uniformity of the structure and a large number of filters and greater batch size allow for utilizing efficient dense computation.

## 6 Architectural Details

1. In this architecture, 11 convolutions are used to compute reductions before the expensive 33 and 55 convolutions. Besides being used as reductions, they also include the use of rectified linear activation which makes them dual-purpose.

2. For technical reasons (memory efficiency during training), it seemed beneficial to start using Inception modules only at higher layers while keeping the lower layers in traditional convolutional fashion.

## 7 GoogLeNet

1. The network is 22 layers deep when counting only layers with parameters (or 27 layers if we also count pooling). The overall number of layers (independent building blocks) used for the construction of the network is about 100.

2. A move from fully connected layers to average pooling improved the top-1 accuracy by about 0.6%.

3. A 11 convolution with 128 filters for dimension reduction and rectified linear activation.

4. A fully connected layer with 1024 units and rectified linear activation.

5. A dropout layer with 70% ratio of dropped outputs.

6. A linear layer with softmax loss as the classifier (predicting the same 1000 classes as the main classifier, but removed at inference time).

## 8 Training Methodology

1. The model's training used asynchronous stochastic gradient descent with 0.9 momentum [6], fixed learning rate schedule (decreasing the learning rate by 4% every 8 epochs). Polyak averaging [5] was used to create the final model used at inference time.

2. Some of the models were mainly trained on smaller relative crops, others on larger ones.

## 9 ILSVRC 2014 Challenge Setup and Result

### 9.1 Classification Challange

The challenge uses the top-5 error rate for ranking purposes. GoogLeNet gained first place in that challenge where the error rate was 6.67% and they did not use any external dataset.

### 9.2 Detection Challange

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).
***Results are reported using the mean average precision (mAP).***
GoogLeNet achieved first place in that challenge where mAP was 43.9% and GoogLeNet used external data named ImageNet 1k and was used for 6 out of the 7 models in their ensemble.

## 10 Conclusion

This model able to create expected optimal sparse structure by readily available dense building blocks is a viable method for improving neural networks. Besides it, a modest increase in computational requirements helps this model to gain a significant quality.

## References


[1] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[3] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[4] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[5] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[6] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.