Going Deeper with Convolutions

Christian Szegedy[1], Wei Liu[2], Yangqing Jia[1], Pierre Sermanet[1], Scott Reed[3], Dragomir Anguelov[1], Dumitru Erhan[1], Vincent Vanhoucke[1], Andrew Rabinovich[4]

[1]Google Inc. [2]UNC Chapel Hill. [3]University of Michigan, Ann Arbor. [4]Magic Leap Inc.

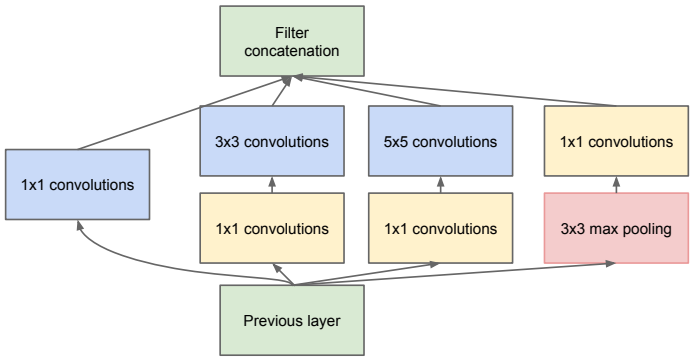

Figure 1: Schematic description of the Inception module replacing the traditional convolutional layers in our network.

We propose an efficient deep neural network architecture for computer vision, codenamed "Inception", which derives its name from the "Network in network" paper by Lin et al [5] in conjunction with the "we need to go deeper" internet meme [1]. 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. Its design took inspiration and guidance from the theoretical work by Arora et al [2]. The benefits of the architecture are experimentally verified on the ILSVRC 2014 classification and detection challenges, where it significantly outperforms the current state of the art while using only 1.5 billion multiply-adds for each network evaluation.

The Inception architecture aims at approximating and covering the optimal local sparse structure of a convolutional vision network by readily available locally dense components. Assuming translation invariance suggests the use of convolutional building blocks: finding the optimal local model and to repeat it spatially. [2] suggests a layer-by layer construction where one should analyze the correlation statistics of the preceding layer and cluster them into groups of units with high correlation. In the layers close to the input correlated units would concentrate in local regions. This way, we end up with a lot of clusters concentrated in a single region which be covered by a layer of 1×1 convolutions, as suggested in [5]. However, one can also expect that there will be a smaller number of more spatially spread out clusters that can be covered by convolutions over larger patches, and there will be a decreasing number of patches over larger and larger regions.

Also we need to apply dimension reduction wherever the computational requirements would increase too much otherwise. The representation should be kept sparse at most places by keeping the dimension of the activation outputs as high as possible, according to the conditions of [2]) and compress the signals only whenever they have to be aggregated en masse. In our GoogLeNet submission, this was achieved by significantly increasing the overall number of filters in the convolutational layers to 1024 in the 7×7 grid. This is enabled by a liberal use of 1×1 convolutions for dimension reduction purposes before the expensive 3×3 and 5×5 convolutions. These reduction layers also include the use of rectified linear activation which makes them dual-purpose by increasing their representation power. The final result is depicted in Figure 1.

The approach taken by GoogLeNet for detection is similar to the R-CNN by [4], but is augmented with the Inception model as the region classifier. Additionally, the region proposal step is improved by combining the selective search [6] approach with multi-box [3] predictions for higher object bounding box recall. In order to reduce the number of false positives, the superpixel size was increased by $2\times$. This halves the proposals coming from the selective search algorithm. We added back 200 region proposals coming from multi-box [3] resulting, in total, in about 60% of the proposals used by [4], while increasing the coverage from 92% to 93%. The overall effect of cutting the number of proposals with increased coverage is a 1% improvement of the mean average precision for the single model case. Finally, we use an ensemble of 6 GoogLeNets when classifying each region. This leads to an increase in accuracy from 40% to 43.9%. Note that contrary to R-CNN, we could get away without bounding box regression which simplified our pipeline.

Team	Year	Place	Error (top-5)	External data
SuperVision	2012	1st	16.4%	none
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	none
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	none
VGG	2014	2nd	7.32%	none
GoogLeNet	2014	1st	6.67%	none

Table 1: ILSVRC classification performance.

Team	Year	Place	mAP %	External data	Ensemble size
UvA-Euvision	2013	1st	22.6	none	?
Deep Insight	2014	3rd	40.5	ImageNet 1k	3
CUHK DeepID-Net	2014	2nd	40.7	ImageNet 1k	?
GoogLeNet	2014	1st	43.9	ImageNet 1k	6

Table 2: ILSVRC detection performance. Unreported values are noted with question marks.

[1] Know your meme: We need to go deeper. http://knowyourmeme.com/memes/we-need-to-go-deeper.

[2] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. *CoRR*, abs/1310.6343, 2013. URL http://arxiv.org/abs/1310.6343.

[3] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *CVPR*, 2014.

[4] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on*, 2014.

[5] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*, abs/1312.4400, 2013. URL http://arxiv.org/abs/1312.4400.

[6] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders. Segmentation as selective search for object recognition. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1879–1886, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-1101-5. doi: 10.1109/ICCV.2011.6126456. URL http://dx.doi.org/10.1109/ICCV.2011.6126456.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.