Going deeper with convolutions

Summary written by Sahana Srihari
April 16, 2020

Summary: The paper describes one of the state of the art 22-layer network used for classification and detection in the ILSVRC14 challenge which achieved a top-5 error rate of 6.67%, which challenged even human performance. The highlight of the proposed method is the increase of depth and width of the network without great compromise on computational complexity(almost 12x lesser parameters). They are strong proponents on the feasibility of sparser architectures and provide experimental proof on the functionality using the ImageNet dataset.

Related work: The main principle of the network is based on CNN and models using this[5]. The inception model draws inspiration from the network of Lin et al [6]. The findings based on using sparser networks can be attributed to the work accomplished by Arora et al. [1]. During training, inspiration was drawn from the works of [8] and [7] during inference time and [3] for overfitting. This model outperformed the previously well performing networks such as krizhevsky [4] by a large margin and comparable to ResNet [2].

Approach: ***Architecture Details:*** The main idea of the model is the inception model which consists of concatenation of parallel filters for convolution ranging from high detail 1x1 filter to 5x5 filters and the optimal local sparse structure in the convolutional vision network. These filters are learn-able & being Gabor filters of different sizes, able to handle multiple object scales. In the inception block, multiple features are simultaneously extracted by the (1x1), (3x3), (5x5), (3x3 max pooling) and concatenated into a single vector which serves as the input to the next layer. The 1x1 convolutional layers is used as a bottleneck for reduction in computation as well as dimensionality reduction. The network consists of many of these "inceptions modules" stacked on top of each other. Features of higher abstraction captured by higher 3x3 and 5x5 layers have 1x1 layer to compute reductions as needed with the added benefits of ReLU units. The lower stages of the network have traditional convolutions whereas the higher staged employ the stacked inception modules. Overall 27 layers deep with the pooling considered. Auxiliary classifiers are connected to intermediate layers for an increase in discrimination & gradient signal propagated as additional regularization. The network is as follows:-

Input$\rightarrow$ $Conv(7x7/2) \rightarrow maxpool(3x3/2) \rightarrow conv(3x3/1) \rightarrow maxpool(3x3/2) \rightarrow inception \rightarrow inception \rightarrow maxpool(3x3/2) \rightarrow inception \rightarrow inception \rightarrow inception \rightarrow inception \rightarrow inception \rightarrow maxpool(3x3/2) \rightarrow inception \rightarrow inception \rightarrow Avgpooling \rightarrow Dropout \rightarrow Linear \rightarrow softmax$ With all the convolutional units accompanied by ReLu activations. The FC layer has 1024 units, dropout with 70% dropped units and a linear layer with softmax is utilized as a classifier.

Datasets, Experiments and Results: Training and validation was done on the ImageNet dataset. Training - 1.2 million images,validation-50,000 and testing - 100,000. Performance metric used was top-1 and top-5 error rates. 7 versions of the network was trained and ensembled for prediction. Images were resized to scales of 256,288, 320 and 352.Left, center & right square, 4 corners and the center crop encompassed the testing set. The softmax predictions averaged across multiple crops. This achieved a top-5 error of 6.67%. Performance on detection challenge- 38.02% on mean average precision single model .

Strengths: The main strength is the ability to go deeper with the network construction without an increase in computational stress. Given the depth and the width of the network, there are still far fewer parameters to deal with due to the sparse architecture.The performance on the detection challenge was impressive given that bounding box regression was not used. The inference for this model can be run on limited resourced computers as well and each stage had increased number of units which do not blow up in computational complexity. The top-5 error rate of 6.67% was a great improvement compared to SuperVision (56% reduction) and Clarifai(40% reduction).

Weaknesses: Although the complexity is drastically reduced in terms of the number of parameters to consider, in the naive form the 5x5 filter size excessively increase the complexity. The paper provides the performance of the model in computer vision domain but has not given more information on other domain data. The outputs of the pooling layers and convolution layers increase the number of outputs between stages. These very large networks are relatively prone to overfitting. Choice of the right kernel size also poses as an issue due to the variation in location of information.

Reflections: The paper made great strides in improvement in the image classification and detection challenges and blew the previously contending models out of the water. The model was constructed to enable greater performance at a lesser cost and also providing evidence in the benefits of using sparser architectures which could be explored further. It would be interesting to see the application of the network in other domains.

References

[1] S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. In *International Conference on Machine Learning*, pages 584–592, 2014.

[2] R. Girshick, J. Donahue, T. Darrell, and J. 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] A. G. Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] 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*, 1(4):541–551, 1989.

[6] M. Lin, Q. Chen, and S. Yan. Network in network. corr abs/1312.4400 (2013). *arXiv preprint arXiv:1312.4400*, 2013.

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

[8] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.