Going deeper with convolutions

Authors: Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich.

Publisher: IEEE Conference on Computer Vision and Pattern Recognition
Year: 2015

Keywords: CNN, Inception, ILSVRC, GoogLeNet

1. **Abstract:**

"We propose a deep convolutional neural network architecture codenamed Inception, which was responsible for setting the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14). The main hallmark of this architecture is the improved utilization of the computing resources inside the network. This was achieved by a carefully crafted design that allows for increasing the depth and width of the network while keeping the computational budget constant. To optimize quality,the architectural decisions were based on the Hebbian principle and the intuition of multi-scale processing. One particular incarnation used in our submission for ILSVRC14 is called GoogLeNet, a 22 layers deep network, the quality of which is assessed in the context of classification and detection."

2. **Introduction:**
    - Normally, in the field of computer vision we observe that to increase the efficiency of object recognition, the authors increase the depth of the neural networks or the size of the training dataset.
    - Both the above processes cause the computational expense and the time expense. But in contrast some of the authors had concentrated in optimizing the algorithms and the approaches irrespective of the hardware and data available.
    - These authors also considered this problem and created a state of art to increase the efficiency of object recognition by a deep convolutional architecture named as Inception by the former researchers.
    - With the help of this architecture they participated in one of the biggest competitions ILSVRC in 2014 with some phenomenal improvements.

3. **Related Works:**
    - Many of the authors for the training of the larger datasets had used high number of layers in the network and increase in layer size which would end up in overfitting of the model.

- In Spite of the fact that the max pooling concludes in the loss of information, it had been implemented and seen some satisfactory results in pose detection, object recognition and localization.
- Some more approaches like network in network lead to the introduction of Inception by Line et al., increased the accuracy of the neural networks. This is the approach used by the authors with deep CNNs.
- One of the dominant approaches was that RNN decomposing the object detection problem into two auxiliary problems introduced by Girschick et al.,
- This division of single problem into two would result in the accuracy of bounding boxes even with the low level inputs. Thus similar architecture is also adapted by the authors.

4. Content and Approach:

- When the training set is limited, the increase in number of layer and parameters would result in the overfitting of the model. This process would be unsuitable for the high level images and their object detection.
- To be true, if two layers of CNN had been increased and chained to the neural networks, it would increase the computation in quadruple. So this would result in computational expense which is said to be limited in practice.
- So, for this purpose, instead of connecting the neural networks fully, they were connected sparsely inside the convolutional neural networks too.
- As done in one of the previous works by Arora et al., the sparse networks can be connected by taking the correlations layer by layer and grouping the neurons with high correlations.
- The correlated units would be concentrating on the local regions for the low level layers. This procedure is very likely to the Hebbian's principle.
- The computation of the non uniform sparse data structures are very complex and inefficient to be computed with present hardware. So, it would exploit the hardware due to the computations of dense matrices.
- This can be overcome by converting into the relatively dense matrices by grouping the sparse matrices for the sparse matrix multiplication.
- The authors had claimed that the inception architecture had been very convenient for the localization, object recognition and pose detection after checking out the improved training methodologies, learning rate and hyper parameters.
- Despite of the fact that it had been very useful for the purpose, there had been ambiguity regarding the principles lead to the construction of the architecture.
- The motivation of the Inception Architecture is to approximate the convolutional networks in a optimal sparse structure with available dense components.
- For the purpose of evading patching issues, the sizes of the patching were restricted to 1x1, 3x3, 5x5 by the inception architecture.

- GoogLeNet had been the name of the team of the authors participated in the competition of ILSVRC14. It is the name given in tribute to LeNet5 network introduced by Yann LeCuns.
- The network used for the training which consists the parameters were around 22 and 27 by adding the pooling layers.
- Some classifiers were regarded as smaller CNNs and placed above the Inception layers. Their losses would be resulted in the sum of total losses of the network multiplied by their corresponding weights.
- For training, the distributive system DistBelief was used for the computational and time considerations. In this process the training had been done on different GPUs parallelly.
- The authors used the asynchronous gradient descent algorithm for optimization with momentum of 0.9 and 4% increase in learning rate for every 8 epochs.
- During the competition the team has not used any other data for the training except that provided by the competition hosts.
- There were two types of accuracy rates in evaluating the results named as top-1 accuracy rate and top-5 error rate based on the comparision of ground truth to the certain number of top answers achieved by the authors.

5. Results:

- The detection is correct if the bounding box overlaps with more than 50% around the object. And if the external objects were detected, it would be penalized.
- With the definite rules defined as follows, the team has shown an significant results of 43.9% mean average precision which had been more that any other team in the competition.
- The top-5 error rate also had been just 6.67% of the test case which had been lower than any other team and even from the previous years. The other challenge to be noted was that the team had not used any external data for the training.
- In the detection results, the accuracy had almost doubled compared to the past year finest results.

6. Conclusion and Future Works:

- To be concluded, the approach had shown the significant results which had been best so far without considering increase in the hardware or dataset.
- This approach also proved that the application of sparse network architectures would be reasonable method in improving the accuracy in the computer vision.
- Even without performing any bounding box regression, the model had provided better results, which shows the efficiency of Inception architecture.
- In future the works could be looked into creating automatic sparse networks for implementation of Inception architectures which had shown productive results.