Very Deep Convolutional Networks for Large Scale Image Recognition

Karen Simonyan, Andrew Zisserman

Summary:

This paper talks about a list of networks and explains the various tests and procedures that the authors followed to get good classification on a large variety of datasets.The performance of the deep networks were improved by changing the window size and the stride length. They were also improved by training the network on various scales. This paper deals with another important aspect of deep neural networks which is the depth of the networks.

The authors start off by explaining the configuration of the network. The input for the network was 224 x 224 images. The mean of the RGB values were subtracted from each pixel. The convolutional layers had very small receptive fields of 3 x 3. The stride was 1 pixel. There were five max pooling layers with a window size of 2 x 2 and stride of 2 pixels. There was 3 fully connected layers after the convolution and max pooling layers. The first two of these fully connected layers had 4096 channels and the last one had 1000 channels, one for each class. The author's networks do not have Local Response Normalization(LRN) except for one. These increase memory consumption and offer no significant improvement.

There were six configurations that were investigated in the paper labelled A-E.All the configurations differ only in depth and have almost similar architecture otherwise. The author's networks have a small receptive field with 3 x 3 convolutional layers and 1 pixel stride. Two such layers have an effective receptive field of 5 x 5 . The advantage of using three 3 x 3 convolutional layers is that there are 3 times the rectified units which makes the network more discriminative. This approach also decreases the number of parameters. 1 x1 convolutional layers were introduced to increase the non-linearity without affecting receptive fields.

Then the authors talk about the training of the network.The training was carried out by optimising a multinomial logistic regression objective using mini batch gradient descent. The authors also used momentum to set the learning rate.The batch size was set at 256 and momentum to 0.9.weight decay regularization was used also dropout regularization was used in 2 of the fully connected layers.

For deep networks, initialization of the weights improperly can lead to unstable networks. To overcome this issue the authors start off from a shallow network and gradually build a deeper network. The 224 x 224 image that will be used for training is cropped from the original image. The image will be cropped at various scales.

Two approaches were followed for setting the scale . The first approach was to use a fixed scale and the second approach was to sample the scale from a range of scales.

While testing the test images are not cropped and the entire network is applied on the full images. Crops can however lead to improved accuracies. The network is evaluated under various input forms.In single scale evaluation the invresed number of layers improved the accuracy in general. The use of 1 x 1 filters improved performance ,but the 3 x 3 filters are needed to improve accuracy and capture features.

Then the model was tested using multiscale evaluation. The scale is changed during test time and the posterior was averaged over all the outputs.scale jittering at test time leads to better outputs. Multi crop evaluation along with dense evaluation outperforms each of them. The authors also combined the outputs of several models and obtained good results , they outperformed individual models. The authors then compare their approach with the state of the art. The authors have demonstrated the effect the depth of networks have on classification performance.

The authors also tested the network on tasks other than classification like localization. In localization the model outputs the vertices of the bounding boxes. The authors also try generalization of very deep features. To do this the last fully connected layer is removed and the previous layer's 4096 channel output is used as a feature. The models seem to perform well in these tasks as well.

Positives:

The main positive aspect of this paper is that the authors go through the entire process of training the networks and also discuss the results of each experiment. The authors also explain any irregularities in the observations. The authors also discuss the various configurations of the network and explain the function of each layer which is very educational for beginners.

Negatives:

The main disadvantage might be the lack of graphical representation of the performance. Though the authors provide the numbers, It would have been better to use some visual aids. Deep networks are hard to understand, seeing what the network sees in each convolutional layers might have helped the user.

Novelty:

I learned how new networks are actually evaluated in research. I also learned how deep networks affect the performance.I learned how a network can be used for various tasks other than classification.

Difference between this and other papers:

The important thing that I noticed is that this paper studies the performance of a particular type of configuration. All the previous papers were general papers taht talked about general topics in deep learning and did not give much insight into how a model is created and tested.

Report on the output Visualization of VGG16 and VGG19

Aim:

The aim of this experiment is to visually verify the output of the convolutional layers to find out what features were seen by the neural network.

Summary:

Pretrained VGG16 and VGG 19 models were used to conduct the experiment. The weights were downloaded from the github repo. The image that was given as input to the models is displayed below:

The image is very challenging to classify,as only the face of the cat is visible. The VGG 16 model incorrectly classified the image as “great grey owl”. The VGG 19 model however classified the picture as “lynx”. The input image was taken from a google search for cats. Lynx or wild cats are medium sized cats with long whiskers.

The initial convolutional layers output from both models were more of the coarse outlines of the faces and as the layers go deeper and deeper the network finds more features like its eyes and nose,which is very interesting to see. The most predominant feature that was visible on most of the layers for this image was the stripes on the forehead and the black and white fine patterns on the chin .

The resultant images are available in the “outputimages” folder.

Result:

The intermediate outputs of the network was analysed and interesting and intriguing facts were found.