Rahul Thapa
Paper: Very Deep Convolutional Networks for Large-Scale Image Recognition

Summary

In this paper, the authors experiment on how the depth of the convolutional neural network affects the accuracy of the large-scale image recognition task. Through their experiment, they found out that by using very small (3 * 3) convolutional filters, they are able to design a deep convolutional network which is a significant improvement on the prior-art configurations.

During training, the authors keep the input to the ConvNets a fixed-size image. The only preprocessing that they do is subtract the mean RGB value, computed on the training set from each pixel. They used a filter of the small receptive field: 3*3. On one of the configurations, they even utilize a 1*1 convolutional filter which can be seen as a linear transformation of the input channels. The convolutional stride is fixed to 1 pixel. The spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolution. All hidden layers are equipped with the ReLU non-linearity. The authors also do not use any Local Response Normalization because they realized that it does not improve the performance on the ILSVRC dataset by that much for increased memory consumption and computation time.

The authors create a model that contains 11 layers to 19 layers, and they compare the result of those networks on the ImageNet dataset. The last three-layer in all these models are fully connected layers. The width of the layer and the receptive field that they use is relatively smaller and therefore, despite their depth, the number of parameters is not greater than the number of weights in a shallower net with larger convolutional layer widths and receptive fields. By using a stack of 3*3 convolutional layer instead of a single larger layer such as 7*7, they incorporated 3 non-linear rectification layers instead of a single one, which makes the decision function more discriminative. It also helps decrease the number of parameters. The incorporation of a 1*1 convolutional layer is a way to increase the non-linearity of the decision function without affecting the receptive fields of the convolutional layers. The authors hypothesize that depth large number of parameters and greater depth, their nets required fewer epochs to converge due to implicit regularization imposed by greater depth, smaller convolutional sizes and pre-initialization of certain layers.

The authors found out that local response normalization indeed does not improve on the performance of their model. They observed that the classification error decreases with the increased ConvNet depth. Finally, they also noted that scale jittering at training time leads to significantly better results than training on an image with a fixed smallest side.

On multi-scale evolution, they found out that using multiple crops performs slightly better than dense evaluation, and the two approaches are indeed complimentary, as their combination outperforms each of them.

The authors also talk about ConvNet fusion, where they basically combine the outputs of several models by averaging their soft-max class posteriors. This improves performance due to the complementarity of the models.

Strengths

- The paper is very nicely structured. There are multiple segments each explaining well a specific part of the architecture/experiment.
- The introduction is concise and clear.
- There are multiple tables in the paper each including a comprehensive result of their experiments as well as results of other comparable models.

- The paper has the appendix section which provides extra information on how the model can be used for other purposes. Therefore, for those who want to play around with this model, it's a great resource.

Weaknesses

- The paper could have included a few mathematical equations and images to help better understand their architecture.
- They do not do a very good job of explaining some of the procedures such as testing such that a wider group of the audience will understand it.

Confusions

- What exactly is 1 * 1 convolutional layer doing in the model that this paper proposes?
- I did not fully understand how they were modifying the training image size.
- I also had trouble understanding their testing procedure. What is the score map? What is the purpose of converting a fully connected layer to a convolutional layer and when exactly are they doing it?

Discussion

- As the paper explains, how exactly increasing the depth results in the better performance of the network?
- What are some things that we need to be careful when increasing the depth of the neural network?
- What is the purpose of using a small receptive field (3 * 3) as opposed to larger fields (7 * 7)?