Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 502] Very Deep Convolutional Network for Large-Scale image recognition

1. Paper Summary

Large public image repositories and high-performance computing systems have boosted the research in large scale image and video recognition via ConvNet. Since the AlexNet took the first place in the ImageNet Large-Scale Visual Recognition Challenge(ILSVRC), many variants have been attempted and excessed its performance. This paper is mainly dealing with the depth as an important aspect of ConvNet architecture. They submitted their model named as "VGGNet" which has won the first and second places in the localization and classification tracks respectively in the ImageNet Challenge 2014.

The author set the generic layout of an architecture with 8 convolution layers and 3 fully connected layers. Then, several more architectures with additional convolutional layers were considered as configurations so that the depth of architecture now varies from 11 to 19. Comparing to the previous works, 3 x 3 receptive fields were solely used throughout the networks except one which also utilized 1 x 1 receptive fields. Rather than using bigger size of filter, they stick to use the smallest size filter which is just enough to capture the notion of left/right, up/down, center to incorporate more non-linear activations. This can decrease the number of parameters so that the network can avoid an overfitting issue even when using a deeper model. Furthermore, it is a reasonable strategy as the effective receptive field of the larger filters is the same as that of the stacked smaller filters. Then, the representation power can be maintained or even better as the non-linear activations can be injected in between so that the network becomes more discriminative.

When training VGGNet, they initialized the first four convolutional layers and the last three fully connected layers with the baseline model. Then, the multinomial logistic regression objective was optimized using mini-batch gradient descent with momentum. In addition, it is regularization by the L2 weight decay and dropout for the first two fully connected layers. As a ILSVRC submission, they ensembled several models by averaging their soft-max class posteriors, which finally achieved the error rate of 6.8% by further trials. One remarkable point to be referred is that comparing to AlexNet, the network required less epochs to converge in spite of larger number of parameters and greater depth.

As an experiment, the author conducted the image classification on the ILSVRC 2012 dataset which was evaluated by the top-1 and top-5 error. It shows that the error decreases with the increased ConvNet depth. Notably, with the same depth, the 1 x 1 receptive field worsens the performance. This indicates that even if a smaller filter size may be preferred, it should be able to at least capture the spatial context. The result that that a deep net with small filters outperforms a shallow net with larger filters further rationalize the use of 3 x 3 receptive fields.

Other important side remarks to refer from the paper are follows.

- Local Response Normalization(LRN) which used in AlexNet, does not improve the performance of their networks.
- Scale jittering at training leads to significantly better results than just rescaling to fixed smallest side, though the latter was used at test.
- Using multiple crops performs slightly better than dense evaluation and it can be further improved by combining the two.

2. Discussion

Here, I want to offer two discussion points. To begin with, what would be the smallest possible number of channels for each convolutional layer which do not worsen the performance in classification tasks? This seems to be a meaningful task to struggle, because too many parameters slow down the training process and requires a big data size. Then its use in real world application will be restricted as the recognition must be carried out in a timely fashion on a computationally limited platform. Next, what if the different filter shape is used? For example, the cross-shaped filter with 5 parameters can also capture the spatial context referred in the paper with 45%(=4/9) less parameters.