Going Deeper with Convolutions (GoogLeNet) Summary

Background

The basic problem that motivated this paper was the need of increase in depth of a neural network. Deeper Networks become obstructive mainly because of two reasons- large parameter count and higher computational cost. According to the paper, the most intuitive way of combating these issues is to move to sparsely connected networks.

Along with this realization, the paper provides insights on implementation of sparse networks, including: -

- motivations from Hebbian Principle.
- Networks in Networks.
- Use of dense submatrices after clustering of sparse ones.

Architecture

The constructive features of GoogLeNet Arcitecture are as follows:

Inception Modules

To quote the paper,

'*The main idea of the Inception architecture is based on finding out how an optimal local sparse structure in a convolutional vision network can be approximated and covered by readily available dense components*.'

The Inception Modules use 1x1, 3x3 and 5x5 filters to avoid alignment of correlated patches, but mostly this use stems from convenience regarding finite computation. Inception Modules, in base sense, refer to the stacking up or concatenation of output filter banks from the above-mentioned layers to form a single input vector to the next layer. The ratio of 3x3 and 5x5 are suggested to increase because higher layers capture more abstract features causing decrease in spatial concatenation.

The major problem arising in the naïve Inception Module is basically that of computation expense of larger filters, used even in modest number. Understanding the toll of computation in addition to the layers of the CNN and pooling considerations brings us to the next feature- 1x1 filters.

Dimension Reduction with 1X1 filters

The paper proposes a solution to the mentioned inefficiency-

'-judiciously applying dimension reductions and projections wherever the computational requirements would increase too much otherwise.'

This solution thrives in the fact that even low embeddings have valuable information of large patches, though in a condensed state. Due to regressive trainability of such condensed information, the paper limits the use of dimension reduction to a reasonable extent.

1x1 convolutions are used to compute reductions before the expensive 3x3 and 5x5 convolutions.

Besides being used as reductions, they also include the use of rectified linear activation which makes them dual-purpose.

These features culminate in the final Inception Module architecture provided by the paper-

'An Inception network is a network consisting of modules of the above type stacked upon each other, with occasional max-pooling layers with stride 2 to halve the resolution of the grid. For technical reasons (memory efficiency during training), it seemed beneficial to start using Inception modules only at higher layers while keeping the lower layers in traditional convolutional fashion.'

Major benefits include: -

- Increasing the number of units at each stage significantly without an uncontrolled blow-up in computational complexity.
- Use of dimension reduction allows for shielding the large number of input filters of the last stage to the next layer, first reducing their dimension before convolving over them with a large patch size.
- It aligns with the intuition that visual information should be processed at various scales and then aggregated so that the next stage can abstract features from different scales simultaneously.

Completed with 22 layers, GoogLeNet architecture technicalities include use ReLU non-linearity in all layers, 70% dropout implementation and use softmax classifier with auxiliary output layers. Trained using asynchronous stochastic gradient descent with 0.9% momentum with a fixed learning rate schedule (4% decrease every 8 epochs).