Assignment 11

Very Deep Convolutional Networks for Large-scale Image Recognition - VGG Net

Nisim Hurst

Sunday 8 April 2018

Abstract

The paper on [1] by Karen Simonyan and Andrew Zisserman, describes and proves more efficient ways of using small receptive field convolutional filters on ConvNets to produce more deep networks. First, to augment non-linearities between layers, it uses cheap 1x1 convolutional filters. Then, it uses 3x3 filters to produce an equivalent 7x7 filter that also preserve information better. It also gradually test increasing network depths and channels to derive an optimal architecture.

Hypothesis

A reduction in the perception field produces more discriminative functions by summarizing less contiguous pixels. However, a trade-off to keep in mind is the increased complexity of the output layer from the convolutional layers chunk. So, reducing perceptive field will be possible only by reducing the ensued complexity to a tractable level through other strategies. Increased channels, i.e. more filters, will produce equally capable non-linear nets but with less parameters by a factor of $O(n^2)$.

Also, increased depth in the convolutional layers at the beginning of the network will be beneficial to accuracy because it will capture more non-linearities in the data.

Evidence and Results

Training

Training was made using multinomial logistic regression and mini batch gradient descent backpropagation with momentum:

1. 256 batch size
2. 0.9 momentum
3. Weight decay of $5 * 10^{-4}$
4. Dropout on the first FCL of 0.5
5. 10^{-2} learning rate decreased by a factor of 10.
6. Weight initialization using a shallower network.
7. Crop and rescaling to [256, 384, 512], RGB color shift and flipping data augmentation.

Testing

The fully connected layers are converted to convolutional layers. The the net is applied to a rescaled image, most of the cases to 256 but some of 384 and 512. For accuracy improvement reference, 150 crops over 3 scales were used. Padding with zeros over the overall feature map is used to fill the 224x244 required dimensions.

The best results were obtained at the net E (*VGG-19*) using multi-crop with 24.4% on the top-1 error and 7.1% on the top-5 error, followed by D (*VGG-16*) who is extremely close with 24.4% and 7.2%.

By combining those two nets (E and D, averaging their softmax posteriors) an impressive 23.7%/6.8%.

Experiments

The ImageNet dataset has 1000 classes. The dataset is divided as follows: 1.3M for training, 50K for validation and 100K for testing. Top-1 and top-5 error measure is used.

6 nets from A to E were tested, first without scale jittering. Here are some of the important observations:

1. There is one, *A-LRN*, with local response normalization but it was found it doesn't improve accuracy significantly.
2. The fact that D is better than C evince that having a 3x3 receptive field captures better non-linearities than 1x1 filters.
3. The architecture saturates upon reaching 19 layers for the ImageNet dataset, but other dataset could use more layers.

Without scale jittering the results were obtained at the net E (*VGG-19*) using a cropping size of 384 with 25.5% on the top-1 error and 8% on the top-5 error. However, D (*VGG-16*) is extremely close with 25.6% and 8.1%.

With scale jittering at test time a significant accuracy improvement was found. Net E obtained 24.8%/7.5% and net D 24.8%/7.5% (the same) .

With multi-crop a slight improvement was found. Net E obtained 24.4%/%7.1 while net D obtained 24.4%/7.2%.

A final experiment was made by averaging the softmax layer posteriors of net E and D and obtained an impressive 23.7%/6.8%. Other datasets were tested at the appendix B.

Contributions

The main contribution of this paper is that it proposes an evolution of AlexNet architecture that allows a more efficient use of more depth in convolutional layers . The authors found a correspondence between a network that has a single 7x7 perception fields with 3 convolutional layers of 3x3 perception field. This correspondence is less complex and better capable of capturing non-linearities. Namely it has two advantages: 1. it incorporates more non-linear rectification layers, and 2. the number of parameters to train decreases at a rate of f^2 where f is the size of the filter.

By using this architecture, the proposed set of networks won the first and second place on the ImageNet Challenge 2014 (ILSVRC).

Another contribution of the paper is Appendix A that shows how to extend the architecture to the *localization* task of the challenge.

Architecture

1. 224x224 input layer (normalized around zero).
2. 6 distinct nets with:
    1. 3x3 stack of convolutional layers on different lengths.
    2. *Stride* fixed at 1.
    3. *Same* padding.
    4. Max-pooling layers of 2x2 and *stride* of 2.
    5. The number of channels is doubled on each of the five nets.
    6. There are increments of depth in the convolutional layer following the series: [11, 11, 13, 16, 16, 19]
    7. One of them (*net C*) has 1x1 convolutional filters.
3. Three fully connected layers, the first two of 4096 units and the last one is a softmax of 1000 units (classes).
4. Each hidden layers uses ReLU and one of the networks (*A-LRN*) Local Response Normalization (LRN).

Implementation

1. Caffe C++ based.
2. 4 GTX Titan Black GPU's, 2-3 weeks of training (for parallelizing batch training, gradient is calculated asynchronously).

Weaknesses

Height and weight of the image at each layer goes down as you go deeper in the network, until reaching 7x7. Conversely, the number of channel keeps doubling until reaching 512. It is really hard to tell whether the nets are improving because the increase of the number of channels or because the increase depth. The leaps are not made uniformly enough to tell.

Future Work

Depth in convolutional networks is beneficial for classification accuracy. Future work could include to relate certain database properties that impose and upper mathematical bound of how much layers we could have until the net saturates itself (on ImageNet it was 19).

References

[1] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," pp. 1–14, 2014.