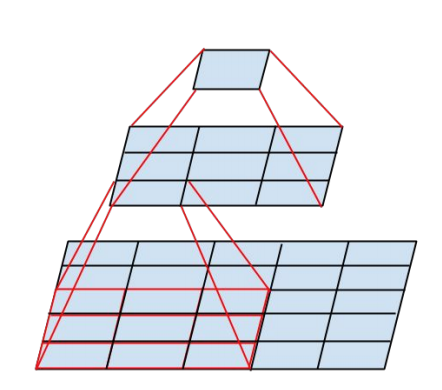
Very Deep Convolutional Networks

for Large Scale Image Recognition

Jo MinKi

1. Introduction   
   These days (around 2014) CNN is astonishing the computer vision field and many researchers are working on this. They found that one of important aspect of CNN architecture design is depth. Based on this, they made deeper CNN which has 16-19 layer and secured first and second places in the localization and classification tacks in ImageNet Challenge. And this work is former version or base of VGG net.   
     
   + Background  
   This research seems start from the curiosity that how and what will happen when the depth gets deeper. So they just tried to keep the dimension and set the 6 different architectures   
     
   + Extra information around here (this is irrelevant to this paper so you can skip this)  
   VGG net is developed by VGGNet team in Oxford at 2014. If Google Net didn’t appear at that time, VGG would get all of glory of the year, but unfortunately it got the second prize at the ImageNet challenge. But even though Google Net performs better than VGG net, VGG has been used much more than Google Net because the gap between both of them are not huge and most of all, the structure of VGG Network is much simpler than Google net so people can easily analyze and fix it. And actually, GoogLe Net seems great but the truth is, it is hard to figure out where to fix it.
2. ConvNet Configuration   
   Network designed using same principles, inspired by Alex Net and the other DNN research about the image classification.   
   
   1. Architecture   
      Input image size is 224224RGB and the only pre-processing they did is subtracting the mean RGB value for each cell. They used really small filter which size is 33, and they also tried with 11 filter which make the layer works like linear transformation. The stride is fixed to 1 pixel and the padding size is set to almost half of filter size so the spatial resolution is preserved after convolution. (For example, padding is 3 when the size of filter is 7)   
      Interesting part of this is that when you use the 33 filter, it factorizes the 55 filter or 77 filter so you can make it deeper while keeping less number of parameter. For example, you can factorize 55 filter with 33 filter, so you can reduce the number of free parameter from 25 to 9+9. This concept is explained precisely in [1].   
      Some of the conv. Layer which has a 22 window perform the max pooling. (And I don’t get this. Is this counted as a layer?) All layers use ReLU. But they did not use LRN (or LCN) and they claim that LRN is not helping to improve the performance on ILSVRC dataset. … is it?
   2. Configurations   
      They have 6 different configurations that named A to E. All of them follow the generic design presented above, but only one difference between of them is depth: from 11 to 19. Width of layer (maybe means number of filter) is start from 64, increase by a factor by 2 after each max-pooling layer until it reaches 512.
   3. Discussion  
      This model has quite different structure from the winner of 2012 and 2013. (It is true) VGG use the smallest size of filter that can contain meaningful spatial information, (Q. how about 22?) so they got some good properties. First, when it comes to stack of smaller filter rather than one large filter, the decision function of classifier gets more discriminative. (Q. Why the stack of smaller filters makes the decision function wiser?) Second, as described above, this way can decrease the number of parameters. For thanks to [1], we can convert one large filter to stack of 33 filters.   
      Meanwhile, 11 filter does not have the spatial information, but still have meaningful function that increase the non-linearity of the decision function without affecting the receptive fields.  
      Anyway the main concept of this model is that increase the depth with using small filter. And even though the GoogLeNet has slightly better accuracy (0.6% better), but it has pretty complex structure with inception, and they reduced the resolution of the feature maps aggressively to decrease the amount of computation.
3. Classification framework
   1. Training

Question.

1. At the Introduction, what does the SVM classifier without fine-tuning part?
2. The term ‘receptive field’ means? Again this
3. At the 2.1 second paragraph, figure out why it is.
4. Is max pooling process counted as a layer? If so, (and it seems so) is this process does same that i know?
5. At the 2.1, They claim that LRN does not improve the performance on the ILSVRC dataset but just
6. Why do they use odd number for size of filter? Can’t we use 22 filter instead of 33
7. Why the stack of smaller filters make the decision function wiser?

Ref.

[1] Rethinking the Inception Architecture for Computer Vision / 2015 / Christian Szegedy