Supplementary Material for Understanding the Effective Receptive Field in Deep Convolutional Neural Networks

Wenjie Luo* **Yujia Li*** **Raquel Urtasun** **Richard Zemel**
Department of Computer Science
University of Toronto
{wenjie, yujiali, urtasun, zemel} @ cs.toronto.edu

1 Extra Analysis of Effective Receptive Field

1.1 Dropout

Dropout [2] is a technique that sets each unit in a neural network randomly to zero during training, which has found great success as a regularizer to prevent deep networks from over-fitting. Assume the network has a dropout probability of r uniformly across all units. Here we do the variance analysis, Eq.6 in the original paper becomes

$$g(i,j,p-1) = \sum_{a=0}^{k-1}\sum_{b=0}^{k-1} w_{a,b}^p z_{i+a,j+b}^p g(i+a, j+b, p) \tag{1}$$

where $z_{i+a,i+b}^p$ are independent Bernoulli variables with probability $1-r$ to be 1. In this case it is easy to see the expectation of gradient is still 0, and $\mathbb{E}[z_{i+a,j+b}^p] = 1-r$, Eq.8 in the main paper becomes

$$\mathrm{Var}[g(i,j,p-1)] = C(1-r)^2 \sum_{a=0}^{k-1}\sum_{b=0}^{k-1} \mathrm{Var}[g(i+a, j+b, p)]. \tag{2}$$

Note the variance now does not factorize into a product of individual variances, as the Bernoulli variables has a non-zero mean. Nevertheless, this case again reduces to the uniform weight case.

1.2 Subsampling and dilated convolutions

Subsampling reduces the resolution of the convolutional feature maps, and makes each of the following convolutional layers operate on a larger scale. It is therefore a great way to increase the receptive field.

Subsampling followed by convolutional layers can be equivalently implemented as changing all the convolutional layers after subsampling from dense convolutions to dilated convolutions[3]. Thus we can apply the same theory we developed above to understand networks with subsampling layers. However, with exponentially growing receptive field introduced by the subsampling or exponentially dilated convolutions, many more layers are needed to see the Gaussian shape clearly.

1.3 Skip connections

Skip connections are another type of popular architecture designs for deep neural networks in general. Recent state-of-the-art models for image recognition, in particular the Residual Networks (ResNets)

*Denotes equal contribution

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

[1] make extensive use of skip connections. The ResNet architecture is composed of residual blocks, each residual block has two pathways, one is a path of q (usually 2) convolutional layers plus nonlinearity and batch-normalization, the other one is a path of a skip connection that goes directly from the input to the output. The output is simply a sum of the results of the two pathways.

In a ResNet of D residual blocks, there are 2^D possible paths to go from input to the output. On a path that selects d skip connections in all D residual blocks, we get an effective CNN with $(D-d)q$ layers which has an ERF that is Gaussian. The overall ERF is a sum of all these Gaussians, weighted by binomial coefficients $\binom{D}{d}$. We don't have an explicit expression for the ERF size yet, but it is clearly smaller than the biggest receptive field possible, which is achieved when the pathway that goes through the convolutional layers are chosen in all residual blocks, this translates to an ERF size proportional to $\sqrt{Dq}$. On the other hand, the term with the biggest binomial coefficient has approximately $D/2 \cdot q$ convolutional layers and the ERF size is proportional to $\sqrt{Dq/2}$, which is $1/\sqrt{2}$ of the biggest receptive field, and a even smaller fraction of the theoretical receptive field size.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[2] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[3] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.