Convolutions inspired by the human retina enable learning of more robust features

Tristan Thrush
tristant@mit.edu

Abstract

In this paper, I present the retinal convolution, which is a convolution that is regularized in a way that mimics the distribution of sensory cells in the human retina. I provide evidence that this convolution can learn more robust features than traditional convolutions. I also argue that the retinal convolution is different from similar attempts, because it clearly has less variance and, in a reasonable sense, lower complexity than a standard convolution with the same receptive field size.

1 Introduction

It is known that the sensory cells in a human retina are distributed densely at the center and sparsely around the periphery. Yena Han et al. [2] found that the decrease in resolution with eccentricity may impose a computational structure that is important for robust feature detection and particularly for scale invariance. A computational model for the human visual cortex has been proposed by Poggio et al. [6] that is consistent with these results. In this paper, I provide a convolution [3] that is inspired by this framework and evidence that it indeed achieves superior performance across some image transformations. There have been other recent scale-invariant learning attempts [7]. However, I argue that mine is different because it can be reduced to a constrained minimization problem of a standard convolution with the same kernel size and has provably lower variance, while also retaining important features for image understanding.

2 Related work

A variety of convolutional architectures have been proposed that provide some amount of invariance to scale. But it is unclear if the problem of learning with these models is "easier" in the sense that they have less variance (which implies better worst-case generalization) than a standard convolution of the same kernel size. In this paper, the goal is to present a more basic and fundamental scale-invariant neural building block that has these properties.

A dilated convolution can be seen as a regularized non-dilated convolution of the same kernel size because it is essentially equivalent to the standard convolution with a subset of its parameter weights multiplied by zeros [8]. The problem of updating a dilated convolution's parameters to minimize training error is therefore equivalent to constrained minimization over a non-dilated convolution:

$$\inf_{f \in \mathcal{H}_{c'}} \hat{\mathcal{E}}(f)$$

Where $\mathcal{H}_c$ is the hypothesis space of a non-dilated convolution, $\mathcal{H}_{c'}$ is the hypothesis space of a dilated convolution with the same kernel size, and $\mathcal{H}_{c'} \subseteq \mathcal{H}_c$. Consequently, dilated convolutions impose strictly more structure on the learning problem and are, in a sense, easier to train. But these convolutions are not invariant to scale on their own: a single dilated convolution only considers a single scale, and the outputs of several such convolutions need to be aggregated to achieve invariance

MIT 9.520/6.860: Statistical Learning Theory and Applications (Fall 2018).

at different scales. Once several layers of dilated convolutions are used, it becomes unclear if the learning problem will result in lower complexity or variance compared to a single standard convolution with the same receptive field size.

Additionally, the SiCNN [7] model has been proposed as a way to learn scale invariance without drastically increasing the number of parameters (through weight sharing at different scales), but this property alone says little about the actual complexity of the SiCNN. The SiCNN undoubtedly takes longer to train in practice than its standard CNN counterpart [7].

3 Model

Here, I present the retinal convolution. Similarly to a dilated convolution, the retinal convolution can be seen as a regularization of a standard convolution by posing a constrained minimization over a subset of the hypotheses of a standard convolution of the same kernel size. But added complexity though the aggregation of several convolutions to learn scale invariance is not necessary because a single retinal convolution can already learn features at several different scales. As a consequence, the retinal convolution actually takes less time to train and is still capable of learning more robust features than the standard convolution counterpart (which I show in section 4).

A retinal convolution has weights for only a subset of its kernel patch in such a way that the density of the weight distribution decreases with eccentricity (this is similar to the distribution of sensory cells in the human retina). Pseudocode for exactly how the retinal convolution decides which pixels receive corresponding weights in a patch (for every channel) is given in Algorithm 1. The actual implementation of the retinal convolution is more complex than this because it was necessary to vectorize it and write an efficient operator to actually perform the convolution.

Algorithm 1 Decide Weights

```
procedure DecideWeights(patch, $k1$, $k2$, $r1$, $r2$)
    if $k1 \leq 1$ and $k2 \leq 1$ then
        include all weights in this patch
    else
        for subpatch in subpatches of dimension $k1$ by $k2$ around the edges of this patch do
            include weight in the middle of this subpatch (or closest possible if $k1$ or $k2$ are even)
        DecideWeights(patch with subpatches from loop removed, $k1 - r1$, $k2 - r2$, $r1$, $r2$)
```

Note that $k1$ and $k2$ are hyperparameters that specify the initial resolution (I refer to these as "starting sparsity" parameters) around the edges of the convolution's patch and $r1$ and $r2$ are hyperparameters that specify how this resolution increases as the center of the patch is reached (I refer to these as "density increase rate" parameters). Note that the algorithm assumes that the convolution's patch is large enough to accommodate the requested subpatches at the desired resolutions.

For further clarification, Figure 1 shows a retinal convolution's weights (in red) overlayed with weights from a standard convolution with the same kernel size. Figure 2 shows an example of what a retinal convolution "sees" if it is looking directly at an image of a dog with the same size as its kernel.

At each recursion of Algorithm 1, the number of weights in a retinal convolution are given by:

$$c1 \cdot c2 \cdot \left(2\left\lceil\frac{k1}{p1}\right\rceil + 2\left\lceil\frac{k2}{p2}\right\rceil - 4\right)$$

Or they are given by the size of the patch times the channels if the base step is reached. The ceiling functions are due to the fact that the loop over the edges of a patch can use subpatches with lower dimension than $k1$ by $k2$ to complete the loop if necessary. Note that $p1$ and $p2$ are the dimensions of the patch at that stage of the recursion and $c1$ and $c2$ are the number of input and output channels. The -4 is to prevent corners from being counted twice. It is easy to see that this amounts to far less than the number of weights from a standard convolution of the same kernel size (or in the very worst case, the same). Further, one can see that the retinal convolution is essentially equivalent to a standard convolution after a subset of its weights have been multiplied by zero. This fact means that the retinal convolution indeed has lower variance, and in a sense, lower complexity than the typical convolution counterpart.

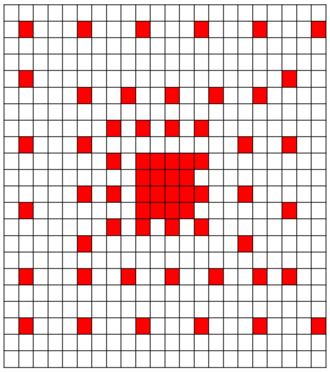

Figure 1: An example of a retinal convolution with kernel size of (22,22), starting sparsity of (4,4), and density increase rate of (1,1).

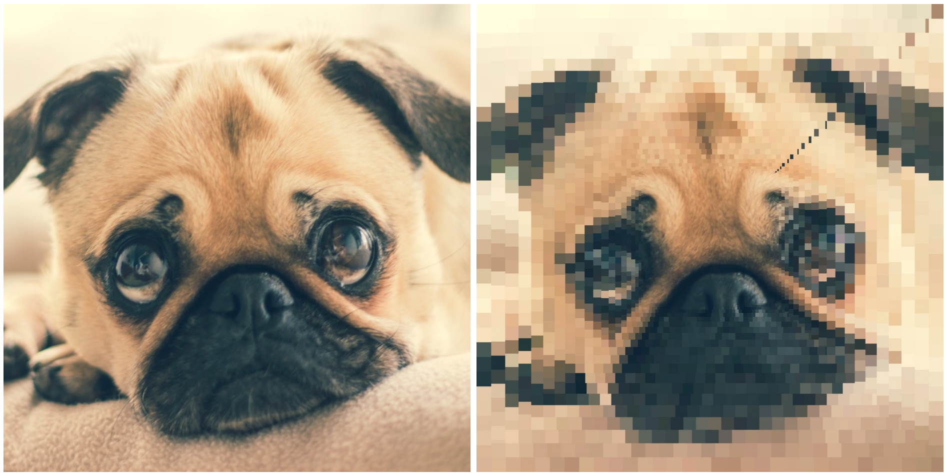

Figure 2: An example of what a retinal convolution "sees." The image patch on the left is passed into a retinal convolution with kernel size of (700, 700), starting sparsity of (22,22), and density increase rate of (1,1). The image on the right contains only the pixels that are used by the convolution, with the pixels around the periphery scaled up for visualization purposes.

4 Results

Here, I compare the retinal convolution to a dilated convolution and a standard convolution on three distortions of the MNIST dataset [4]. I only compare single convolutions as opposed to more complex structures that aggregate convolutions at several scales, because the point of this paper is to present a simple regularization technique of a single convolution that enables robustness and less variance.

The overall architecture of each network is provided in Table 1; after MNIST images are fed through the architecture, there is a log softmax layer that classifies digits, and the nets are trained with stochastic gradient descent [1]. I tried a variety of architectures to increase certainty that the retinal convolution is generally helpful for learning robust features. The retinal convolution that I use has a starting sparsity of (4,4), and a density increase rate of (1,1); these parameters were chosen arbitrarily and not fine-tuned. The dilated convolution that I use has a dilation of 2, so a standard convolution with the same number of parameters would have a kernel size of (11,11).

The networks were trained on 60000 MNIST examples with a padding of 10 pixels around the perimeter. They were then evaluated on 10000 unseen test images with a variety of distortions. In the first test, the images were undistorted. In the second test, the images were scaled up by a factor of two. In the third test, the images were scaled up by a factor of two and rotated randomly by up to 45 degrees in either direction.

{retinal, dilated, standard} convolution
kernel size of (22, 22), 1 in channel, 10 out channels
{standard convolution, no convolution}
kernel size of (6,6), dropout p=0.5, 10 in channels, 20 out channels
fully connected
{320 in, 810 in}, 50 out
fully connected
50 out, 10 in

Table 1: The architecture of the different networks. The goal is not to achieve the best accuracy on MNIST with a complicated architecture; instead, it is to compare a single retinal convolution with alternatives. Note that the convolutions are passed through a max pool of 2, and the layers are separated by ReLU [5], but for brevity these are not part of the table. The "{}"'s represent different options that were tested.

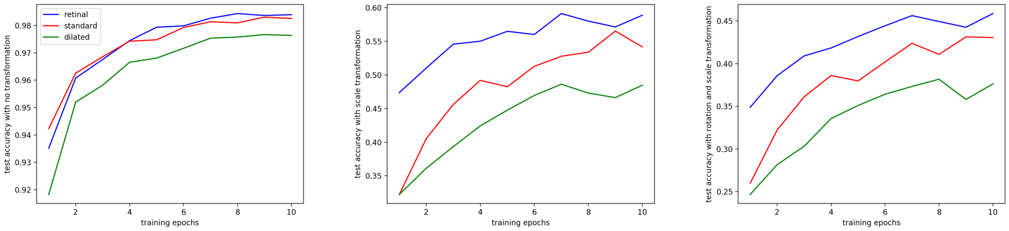

Figure 3: A comparison of the retinal convolution's invariance to transformations with other convolutions on distorted MNIST test data. These are the results with the second convolution from Table 1 in the net architecture.

The retinal convolution has fewer parameters, is faster to train, and has less variance than both the dilated convolution and the standard convolution that I compare with. However, there is evidence that it can learn more robust features (particularly scale invariance) without training with data that explicitly encourages the learning of these features. Furthermore, the retinal convolution can accomplish this task without sacrificing any significant performance on an undistorted test set. Figures 3 and 4 support my claims.

5 Discussion and conclusion

The retinal convolution can be seen as a regularization of its standard convolution counterpart because adjusting its parameters to minimize a loss function amounts to constrained minimization over a subset of the hypotheses that a standard convolution has access to. As such, the retinal convolution is qualitatively different than previous attempts to capture scale invariance, which, to the best of my knowledge, typically aggregate the results of several convolutions. The retinal convolution has lower variance, training time, and parameters than a standard convolution with

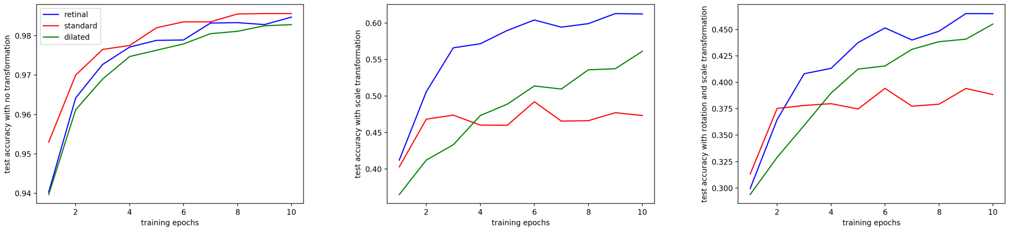

Figure 4: A comparison of the retinal convolution's invariance to transformations with other convolutions on distorted MNIST test data. These are the results without the second convolution from Table 1 in the net architecture.

the same kernel size, yet, despite being less complex in this sense, I have shown that the retinal convolution is capable of achieving superior performance (particularly on scale invariance learning) without loosing performance on problems where scale is fixed. The success in this paper warrants more investigation to further test the abilities of this new neural building block that is the result of a principled regularization of a standard convolution inspired by findings regarding the human retina.

Acknowledgments

I thank the 9.520 staff for suggesting the excellent related work that was incorporated into this paper.

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *in COMPSTAT*, 2010.

[2] Yena Han, Gemma Roig, Gadi Geiger, and Tomaso Poggio. On the human visual system invariance to translation and scale. *Vision Sciences Society*, 2017.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[4] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[5] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress.

[6] Tomaso A. Poggio, Jim Mutch, and Leyla Isik. Computational role of eccentricity dependent cortical magnification. *CoRR*, abs/1406.1770, 2014.

[7] Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *CoRR*, abs/1411.6369, 2014.

[8] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015.