LEARNING RETINAL TILING IN A MODEL OF VISUAL ATTENTION

Brian Cheung, Eric Weiss & Bruno Olshausen
Redwood Center for Theoretical Neuroscience
University of California, Berkeley
Berkeley, CA 94720, USA
{bcheung,eaweiss,baolshausen}@berkeley.edu

ABSTRACT

We describe a neural network model in which the tiling of the input array is learned by performing a joint localization and classification task. After training, the optimal tiling that emerges resembles the eccentricity dependent tiling of the human retina.

1 INTRODUCTION

Attention has been applied successfully to a variety of different applications including natural language processing Bahdanau et al. (2014), vision Jaderberg et al. (2015); Mnih et al. (2014); Xu et al. (2015), and memory Graves et al. (2014). But to our knowledge, there has been little work investigating the properties and features necessary for a neural network to optimize the properties of its input window. In this work, we propose a learnable retinal sampling lattice to explore what properties are best suited for an attention mechanism in performing classification and localization. We find in the absence of the ability to rescale or 'zoom', our model learns to create an eccentricity dependent layout where a distinct region of high acuity emerges surrounded by a low acuity periphery.

2 MODEL

Attention in neural networks may be formulated in terms of a differentiable feedforward function. This allows these models to be trained with backpropagation. Many visual attention models today can be reformulated into a generic equation written as

$$V_i = \sum_{n}^{H} \sum_{m}^{W} U_{nm} k(m, n; \Phi_i) \tag{1}$$

$$\forall i \in [1, ..., H'W'] \tag{2}$$

The pixels in the input image U are mapped to a smaller output V. This can be interpreted as a form of routing where a select number of pixels from the input are connected to the output. The routing is defined by a kernel filter $k()$. The kernel defines which pixels in the input will contribute to a particular output. In Xu et al. (2015), the soft attention mechanism is computed with a softmax kernel. For visual attention calculated directly over an input image, this formulation can be prohibitively expensive because it creates a unique weight for every pixel in the input image. Most formulations of visual attention over the input image assume a factorization between the m and n dimensions of the input shown in 3. This includes the attention models proposed by Jaderberg et al. (2015); Mnih et al. (2014); Gregor et al. (2015) The parameters Φ_i define parameters specific to the kernel in each spatial dimension. For example, the parameters can specify the centers of the kernel as well as the shape (ex: gaussian, bilinear, etc).

$$k(m, n; \Phi_i) = k(m, \Phi_{xi}) k(n, \Phi_{yi}) \tag{3}$$

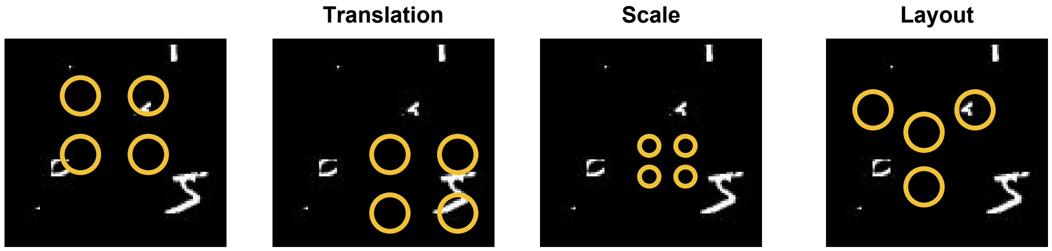

Figure 1: Examples of possible routing configurations for the attention mechanism formulated in 1. Each yellow circle corresponds to a single kernel filter.

While this factored formulation reduces the space of possible transformations from input to output, it can still can form many different mappings from an input U to output V. Figure 2 shows the possible windows which an input image can be mapped to an output V. The yellow circles denote the center location of a particular kernel, each kernel maps to one of the outputs V_i. By adjusting the parameters Φ, it is possible to translate, scale and even change the layout of the pixels which are mapped from U to V.

We develop a model of overt attention inspired by Mnih et al. (2014). An image U is reduced by a glimpse generator using 4. This glimpse V_t is processed by a fully-connected recurrent network $f_{recurrent}$.

$$V_{i,t} = \sum_{n}^{H}\sum_{m}^{W} U_{nm}k_{i,t-1} \tag{4}$$

$$h_t = f_{recurrent}(V_t, h_{t-1}) \tag{5}$$

$$g_c, g_s = f_{localization}(h_t) \tag{6}$$

$$\mu_i = g_s(g_c - x_i) \tag{7}$$

$$Z_i = \sum_{m} e^{\frac{-(m-\mu_i)^2}{2\sigma_i^2}} + \epsilon \tag{8}$$

$$k_{i,t} = \frac{1}{Z_i} e^{\frac{-(m-\mu_i)^2}{2\sigma_i^2}} \tag{9}$$

$$\tag{10}$$

Unlike Jaderberg et al. (2015) and Mnih et al. (2014), our filters are not constrained to lie on a regular rectangular grid. Our kernels are modeled by a set of gaussian filters each with a learnable mean μ_i and variance σ_i^2. The centers μ_i are calculated with respect to a global center g_c and scale g_s and learnable offset x_i specific to each kernel. The global center and scale are predicted by the localization network $f_{localization}$ which is parameterized by a fully-connected neural network. In this work, we investigate two variants of the proposed model.

- Translation and Scaling: This model follows equation 6 where it can both rescale and translate the kernels.
- Translation Only: The model can only translate the kernel filters $g_c = f_{localization}(h_t)$ and the global scale is fixed $g_s = 1$.

Before training, the kernel filters are initialized as a 12x12 grid (144 kernel filters), tiling uniformly over the input image creating a glimpse representation as shown in Figure 2B. Analogous to an eye movement, these kernels can be translated as a single group using g_c while x_i and σ_i^2 are learned parameters fixed after training. Our models are trained end-to-end using ADAM Kingma & Ba (2014) to simultaneously minimize both classification and localization error on a cluttered MNIST dataset. An example from the cluttered MNIST dataset is shown in Figure 2. Handwritten digits

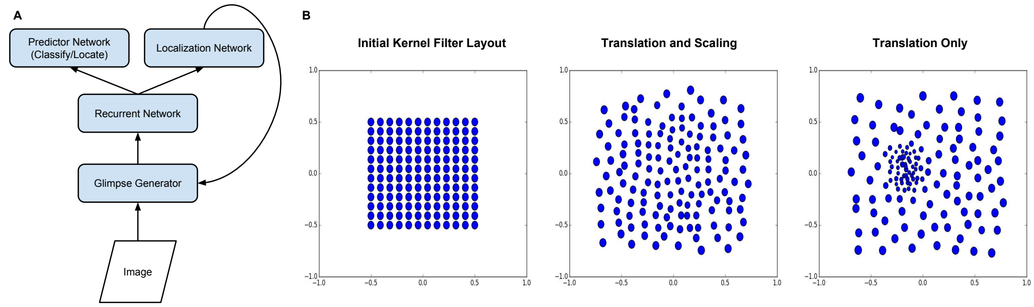

Figure 2: **A**: High level diagram of our attention model. **B**: Scatter plot of the centers of each kernel filter. Image boundaries correspond to [-1, +1]. The radius of each point corresponds to the variance σ_i^2 of the kernel. (Left) Initial layout of the kernel filters before training. (Middle) Layout learned after training a model which can translate and rescale. (Right) Layout learned after training a model which can only translate its kernel filters.

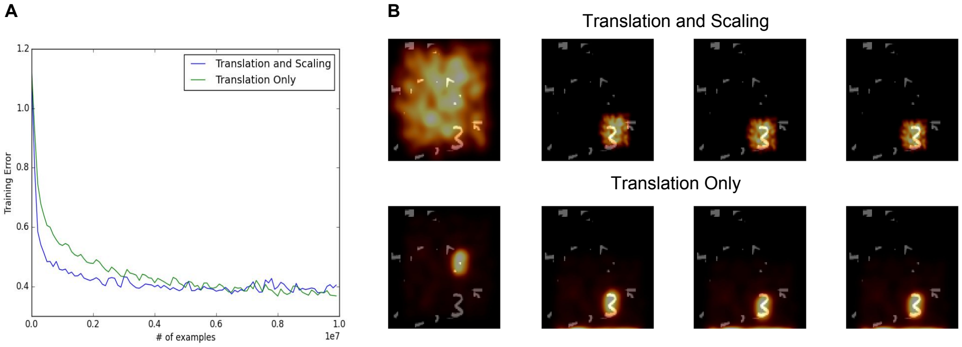

Figure 3: **A**: Comparison of training error during training. **B**: Attention window overlaid an example image for the Translation and Scaling model (top row) and Translation only model (bottom row) for 4 timepoints from left to right.

from the original MNIST dataset LeCun & Cortes (1998) are randomly scattered over a 100x100 image with varying amounts of clutter. In contrast to the cluttered MNIST dataset proposed in Mnih et al. (2014), the amount of clutter for each image varies randomly from 0 to 20 pieces. This prevents the attention model from learning a solution which depends on the number 'on' pixels in a given region. Our models are given T=4 glimpses before predicting the final location and class of the digit in the image.

3 Results and Conclusion

Figure 2B shows the layouts of the kernel filters before and after training. In both variants of our attention model, the kernel filters spread their centers to cover the full image. This is sensible as the digit can appear anywhere in the image with uniform probability.

Remarkably, the kernels in the translation only model tile in a similar fashion to those found in the human retina Curcio & Allen (1990). This layout is composed of a high acuity region at the center (low variance gaussians) surrounded by low acuity (high variance gaussians). Figure 3B shows this high acuity region is sized perfectly to fit the MNIST digit. Meanwhile, the translation and scaling model simply resizes its retina. These initial results indicate the possibility of using deep learning as a mechanism to discover the optimal tiling of retinal ganglion cells in a data driven manner. These results point the way towards models that could provide insight into the spatial sampling strategy in biological visual systems.

ACKNOWLEDGMENTS

We would like to acknowledge everyone at the Redwood Center for their helpful discussion and comments. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPUs used for this research.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Christine A Curcio and Kimberly A Allen. Topography of ganglion cells in human retina. *Journal of Comparative Neurology*, 300(1):5–25, 1990.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pp. 2008–2016, 2015.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yann LeCun and Corinna Cortes. The mnist database of handwritten digits, 1998.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pp. 2204–2212, 2014.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.