Comparison of Foveated Downsampling Techniques in Image Recognition

Parsa Torabian — Department of Systems Design Engineering[1]
Ronak Pradeep — David R. Cheriton School of Computer Science[1]
Jeff Orchard — David R. Cheriton School of Computer Science[1]
Bryan Tripp — Department of Systems Design Engineering[1]
Email: {p2torabi, rpradeep, jorchard, bptripp}@uwaterloo.ca
1 : University of Waterloo

Abstract

Foveation is an important part of human vision, and a number of deep networks have also used foveation. However, there have been few systematic comparisons between foveating and non-foveating deep networks, and between different variable-resolution downsampling methods. Here we define several such methods, and compare their performance on ImageNet recognition with a custom DenseNet network. The best variable-resolution method performed similarly to uniform downsampling. Thus in our experiments, foveation did not substantially help or hinder object recognition in deep networks. However, a naturalistic foveation method with continuously varied resolution performed better than a widely used method with several discrete resolutions.

1 Introduction

The retinas of humans, monkeys, and many other animals have a high-resolution fovea. In humans, this disproportionate representation of the central visual field carries through the whole visual cortex, and eye movements to foveate task-relevant features are an essential part of vision. Deep convolutional networks are inspired by the primate visual system, but they usually lack foveation, which may be a limitation in some contexts. In humans, foveation allows both the wide field of view needed for tasks like visual navigation, and the high resolution needed for tasks like reading, without impractical brain size or metabolic cost. Foveation could potentially provide similar benefits in artificial systems, particularly mobile and edge-computing systems. Some previous studies have used a rough approximation of natural foveation made up of several distinct images at different resolutions, e.g. as in [1]. In contrast, resolution changes gradually in natural systems. This may have benefits, but it is not clear how to arrange such a representation for input to a convolutional network. A circular image with high magnification at the centre wastes pixels at the corners. A polar representation does not, but it sacrifices translational equivariance. In summary, while foveation could potentially have benefits for deep networks, it is not clear when, or how best to implement foveation.

To help fill this gap, we compare several foveated downsampling approaches to uniform downsampling in object recognition. In this context, the different foveated methods perform fairly similarly to each other, and the best performs roughly the same as uniform downsampling (top-1 validation accuracy 62.7% vs. 62.6%; Table 1). Therefore, foveation does not seem to be important for object recognition, which is unsurprising given the good performance of standard deep networks, but it does not substantially interfere either. This suggests that foveation could be incorporated into more general vision systems that perform multiple tasks, such as in robots that must recognize objects and also read text in the environment.

2 Methods

2.1 Network architecture and training

We trained deep networks on the ImageNet recognition task, with various kinds of downsampled images as input. In each case we used a version of the DenseNet [2] architecture network. It is similiar to the DenseNet-121 architecture with some modifications to account for the fact that the images are downsampled as a preprocessing step. The kernal size of the initial features is changed to a dimension-preserving value of three (down from seven) and the subsequent max pooling layers are removed. The first DenseBlock of size 6 is also removed. The original hyperparameters and training procedure are used, including random horizontal flips, batch size etc. We trained each network for 90 epochs, using SGD (initial learning rate 0.1, reduced by 10x every 30 epochs).

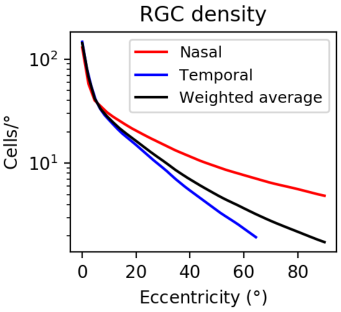


Fig. 1: Estimate of retinal ganglion cell (RGC) density as a function of degrees from the fovea. We use estimates from [3], which provides data along the nasal-temporal axis. [4] shows that density is similar in temporal, dorsal, and ventral directions, but higher in the nasal direction. To calculate radially symmetric mean values, we sum nasal and temporal fits from [3] with weights 0.25 and 0.75 (to account for the fact that nasal density is atypical).

2.2 Downsampling techniques

Uniform downsampling: As a baseline method, ImageNet images were uniformly box-downsampled to a 32x32 resolution.

Multi-resolution downsampling: We produced a simple foveated representation composed of four 16×16 downsampled images with different magnifications. The first spanned the whole image, the second spanned the central half of the width and height of the image, the third a quarter the width and height, and the fourth an eighth. Several past papers have used a similar approach, e.g. [1].

Polar retinal downsampling: We sampled the image in polar coordinates, creating a rectangular image (44×23 pixels) in which the long edge corresponded to the angle and the short edge the radial distance from the fovea. The density of samples in the radial direction declined with greater distance from the centre. We based the sampling density on retinal ganglion cell (RGC) density (see Figure 1). We used gaussian filters with radially increasing widths to reduce artifacts. See example in Figure 2.

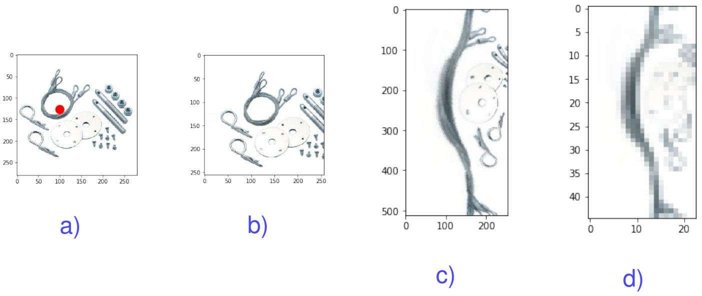


Fig. 2: An example of polar-retinal downsampling. (a) The focal point (highest saliency) is determined (red dot). (b) The image is then cropped so the center is at the focal point. (c) The image is then 'foveated' resulting in pixels closer to the center becoming over-represented while pixels close to the edge are under-represented. In this case, the white area on the left of the foveated image is representing the white pixels inside the loop of steel wire of the source image. (d) The result is downsampled uniformly.

Cartesian retinal downsampling: We sampled the image with the same radially-varying density as above, but created a circular image with strong barrel distortion (Figure 3), rather than a polar representation. This resulted in a transformation that better retains the translational equivariance property of convolutional networks, at the cost of wasting pixels in the corners.

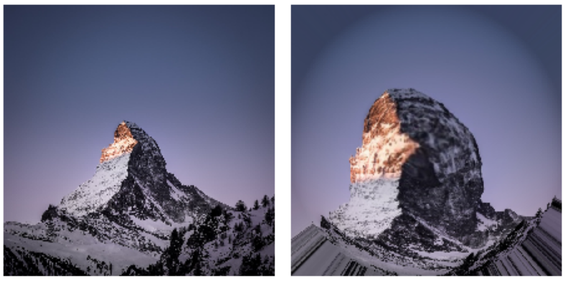
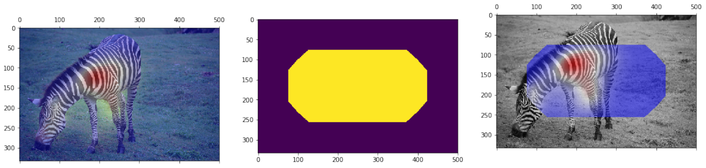


Fig. 3: An image before and after cartesian-retinal downsampling. Much like polar foveation, the center of the image is over-represented in the downsample while the extremities are under-represented, proportional to RGC density data.

2.3 Selection of image points to foveate

A saliency map was generated for each image with a DeepGaze II model [5]. This map estimated the likelihood of a human orienting to each pixel. Human gaze often orients to areas of interest such as faces and foreground objects, which often correspond to the target label. We selected the point of highest saliency, subject to a constraint that avoided points near image edges (as selecting a point near the edge would render much of the crop blank). Specifically, we only chose points around which at least 80% of a 256×256-pixel crop would fall within the image boundaries (Figure 4). If the resulting crop went outside the image boundaries, we extrapolated by copying edge pixels.

We chose three of the top saliency points in every image. The highest-saliency points are typically close together, and contain similar information. To avoid selecting multiple similar points, we modified the saliency maps after each selection. Specifically, we subtracted a square-gaussian function from the saliency map, with a peak equal to the saliency at the chosen point, and a width of 60 pixels.



a) Saliency output of DeepGaze II

b) The ineligible region in purple

c) Saliency map clipped to remove ineligible areas

Fig. 4: The process of finding a valid saliency map from which the point of highest saliency is chosen. (a) A DeepGaze II model determines a general saliency map. (b) An ineligible region is identified (in purple) where points would result in too much of the resultant crop (20% or more) falling outside the image. (c) The saliency map is clipped and normalized before points are chosen.

3 Results

Figure 5 shows training curves for each of the downsampling methods. During training, each crop surrounded one of the three most salient points (with sequential updating of the salience map, as described in the Methods) at random. Table 1 summarizes validation performance of the trained models. Predictions were based on all three foveations for each image (logits averaged across foveations).

4 Conclusion

The cartesian foveation method performed best in this study. Each of the foveated methods has a limitation that could potentially be improved in future work. The polar mapping sacrificed translational equivariance (e.g. the same edge detector could respond to a vertical edge at the bottom of the image and a horizontal edge at the side). This might be mitigated in the future by building rotational equivariance into the network, e.g. as in [6]. The cartesian representation wasted pixels at the corners of the image, which limits computational efficiency. Our version of the multi-resolution representation arranged resolutions side-by-side, which introduced edge effects. The resolutions could also be treated as separate input channels. We did not do this because we wanted to hold constant the numbers of parameters and sizes of the representations across models.

Table 1: Performance on the validation set

Model	Top 1 Accuracy	Top 5 Accuracy
Uniform	**62.7**	**83.7**
Polar-Retinal	59.8	81.1
Cartesian-Retinal	*62.6*	*83.5*
Multi-Resolution	60.2	81.6

We did not find an advantage of foveated imaging in image recognition, but the results provide important information for several other lines of work. Foveation can be beneficial for certain other tasks in robotics, e.g. [7, 8]. However, image recognition is an important part of many tasks, so the effect of foveation on image recognition performance may affect the contexts in which these benefits can be realized. Relatedly, the good performance of cartesian downsampling supports our ongoing work to develop a foveated lens that can produce a higher-resolution, non-downsampled image with a similar pattern of distortion [9]. In the future, we also hope to incorporate cartesian foveation into convolutional-network models of the primate visual system, building on [10] (and see [11] for related work). The good performance of cartesian downsampling supports this type of input distortion as a means of modelling physiological foveation and cortical magnification within a convolutional network. Finally, recent work [12] has shown that small image patches are sufficient for state-of-the-art image recognition, if the patches are chosen via reinforcement learning rather than a saliency model. It would be worthwhile to compare the performance of this approach, as well as other visual attention methods, using foveated image patches.

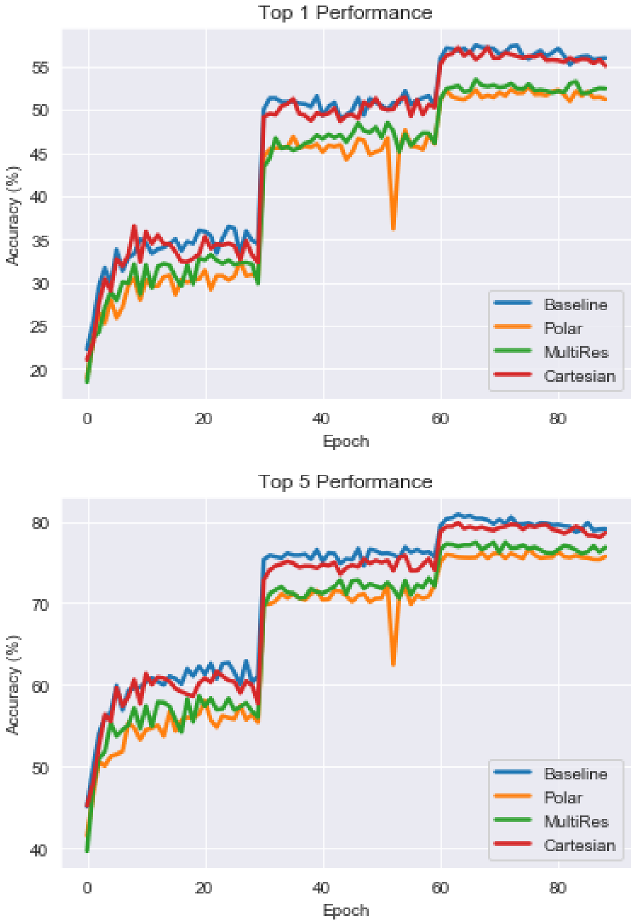


Fig. 5: Top 1 and Top 5 validation accuracy during training

Acknowledgements

We thank Shamak Dutta for helpful discussions.

References

[1] P. R. Medeiros, R. B. Gomes, E. W. Clua, and L. Gonçalves, “Dynamic multifoveated structure for real-time vision tasks in robotic systems,” *J Real-Time Image Processing*, pp. 1–17, 2019.

[2] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[3] H. Wassle, U. Grunert, J. Rohrenbeck, and B. Boycott, “Cortical magnification factor and the ganglion cell density of the primate retina,” *nature*, pp. 643–646, 1989.

[4] C. A. Perry VH, Oehler R, “Retinal ganglion cells that project to the dorsal lateral geniculate nucleus in the macaque monkey,” *Neuroscience*, pp. 1101–1123, 1984.

[5] M. Kümmerer, T. S. A. Wallis, and M. Bethge, “DeepGaze II: Reading fixations from deep features trained on object recognition,” 2016.

[6] M. Weiler, F. A. Hamprecht, and M. Storath, “Learning steerable filters for rotation equivariant cnns,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 849–858.

[7] V. J. Traver and A. Bernardino, “A review of log-polar imaging for visual perception in robotics,” *Robotics and Autonomous Systems*, vol. 58, no. 4, pp. 378–398, 2010.

[8] Y. Zaky, G. Paruthi, B. Tripp, and J. Bergstra, “Active perception and representation for robotic manipulation,” *arXiv preprint arXiv:2003.06734*, 2020.

[9] S. Huber, B. Selby, and B. Tripp, “OREO: An open-hardware robotic head that supports practical saccades and accommodation,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2640–2645, 2018.

[10] B. Tripp, “Approximating the architecture of visual cortex in a convolutional network,” *Neural computation*, vol. 31, no. 8, pp. 1551–1591, 2019.

[11] J. Kubilius, M. Schrimpf, K. Kar, R. Rajalingham, H. Hong, N. Majaj, E. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt *et al.*, “Brain-like object recognition with high-performing shallow recurrent ANNs,” in *Advances in Neural Information Processing Systems*, 2019, pp. 12 805–12 816.

[12] G. Elsayed, S. Kornblith, and Q. V. Le, “Saccader: Improving accuracy of hard attention models for vision,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 702–714. [Online]. Available: http://papers.nips.cc/paper/8359-saccader-improving-accuracy-of-hard-attention-models-for-vision.pdf