Orientation representations in convolutional neural networks are more discriminable around the cardinal axes

Margaret Henderson (mmhender@ucsd.edu)
University of California, San Diego, 9500 Gilman Drive
La Jolla, CA 92092 USA

John T. Serences (jserences@ucsd.edu)
University of California, San Diego, 9500 Gilman Drive
La Jolla, CA 92092 USA

Abstract:

Convolutional neural networks (CNNs) share some similarity in representational structure to the primate ventral visual stream, however less is known about whether low-level visual features are represented in the same way by CNNs and the brain. Here, we focus on orientation perception, a well-understood aspect of the primate visual system. We asked whether convolutional neural networks trained to perform object recognition on a natural image database would exhibit an "oblique effect" such that cardinal (vertical and horizontal) orientations are represented with higher precision than oblique (diagonal) orientations, as has been measured in the primate brain. We obtained activation patterns from two networks (NASnet and Inception-V3) presented with oriented grating stimuli, and used a Euclidean distance metric to measure the discriminability between patterns corresponding to different pairs of orientations. In agreement with human perception, we find that the discriminability of representations generally peaks around the cardinal axes. This finding suggests that cardinality effects in human visual perception are not dependent on a hard-wired anatomical bias, but can instead emerge through experience with the statistics of natural images.

Keywords: convolutional neural network; orientation; vision; primates; perceptual bias

Introduction

Hierarchically organized neural network models trained to perform object categorization have been shown to provide a reasonable approximation of the features represented by neurons in the ventral visual cortex of primates (Kubilius et al., 2018; Yamins et al., 2014). This general correspondence is present even at the earliest layers of convolutional neural networks (CNNs), which are often found to learn Gabor-wavelet-like filters (Yamins & DiCarlo, 2016). However, the organization of low-level feature representations by CNNs has not been extensively characterized. Understanding whether CNNs develop idiosyncrasies that mimic the properties of the primate visual system is important for developing models that can inform our understanding of the brain. Additionally, because the majority of neural network properties are acquired through training, examining feature representations of CNNs is a useful tool for determining which properties of the primate brain might be innate and which are likely to be acquired through experience.

In this paper we focus on the well-known "oblique effect", in which human and non-human primate observers tend to show higher acuity around cardinal orientations (horizontal and vertical) compared to oblique orientations (Bauer, Owens, Thomas, & Held, 1979; Higgins & Stultz, 1950). This effect is thought to originate from an over-representation of neurons tuned to horizontal and vertical orientations, which has been measured in primary visual cortex of mice and cats, as well as primate V2 (Li, Peterson, & Freeman, 2003; Salinas, Velez, Zeitoun, Kim, & Gandhi, 2017; Shen et al., 2014). According to an efficient coding framework, this anisotropy is adaptive because it allows for optimal processing of natural scenes, in which horizontal and vertical edges are common (Girshick, Landy, & Simoncelli, 2011).

Based on this framework, we hypothesized that if a CNN is trained on a dataset of natural images, it may develop similar properties. To test this, we took neural networks that were pre-trained on a database of natural images, and obtained activations after presenting them with circular grating stimuli of varying orientations. We then measured the discriminability of activation patterns at each layer corresponding to neighboring orientations. We then evaluated how discriminability changed as a function of position in orientation space. Our results suggest that, similar to the primate brain, CNNs exhibit an anisotropic representation of orientation.

Methods

Visual stimuli

Each CNN was presented with visual grating stimuli (square images 140 x 140 pixels) at a range of orientations, spatial frequencies, and noise levels. Stimuli were circular, sinusoidal gratings with smoothed edges (kernel size = 10 pixels, sd = 5 pixels), presented against a mid-gray background. After smoothing, each grating had a radius of 65 pixels. Orientations ranged between 1-180 in 1 degree steps, and spatial frequencies ranged from 0.04-0.22 cycles per pixel, in 4 logarithmically spaced steps (0.04, 0.07, 0.12,

0.22). We also superimposed three levels of Gaussian noise onto gratings. The first level had zero noise, the second level had Gaussian noise with a standard deviation equal to the grating amplitude/8, and the highest noise level had a standard deviation equal to the grating amplitude/4. We generated 4 gratings at each orientation, noise level, and spatial frequency, for a total of 8640 images. The phase of each grating was randomly selected within the range of 1-180 degrees.

Obtaining neural network activations

We used two pre-trained CNNs for this study, NASnet (Zoph, Vasudevan, Shlens, & Le, 2017), and Inception-V3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2015), both of which were trained on the ImageNet classification dataset (Deng et al., 2009). We evaluated each model using Tensorflow 1.12.0 (Abadi et al., 2016) using the TF-slim library and Python 3.6 (Python Software Foundation).

The full image set was passed through each CNN in 96 batches of 90 images each, and the resulting activation patterns at each layer were recorded. To reduce the size of the activations, we performed principal components analysis (PCA) across all 8640 images, and saved a maximum of 500 components for each layer.

Measuring discriminability

To evaluate how orientation discriminability varied at different points in orientation space, we calculated the Euclidean distance between activation patterns corresponding to each pair of neighboring orientations (1 degree apart). We performed this calculation within each spatial frequency and noise level separately. Since there were 4 gratings presented at each orientation (with randomized phase), this gave a total of 32 comparisons between each orientation and its leftward and rightward neighbors. We report the mean and standard deviation across these 32 comparisons. Note that though the absolute values of Euclidean distance reported here are not particularly meaningful, the relative values are interpretable.

Results and Discussion

To visualize the organization of orientation representations at each layer of each CNN, we first plotted the first two principal components corresponding to each orientation (example shown in Figure 1). This revealed that as orientation was varied, representations tended to follow either a circular or linear trajectory. Similar patterns were found in both networks, with some variation across layers. Clustering by spatial frequency was also apparent. More importantly, the spacing between points on these plots reveals that pairs of stimuli close to the cardinal axes tended to be more dissimilar than pairs spaced an equal number of degrees apart but located near an oblique.

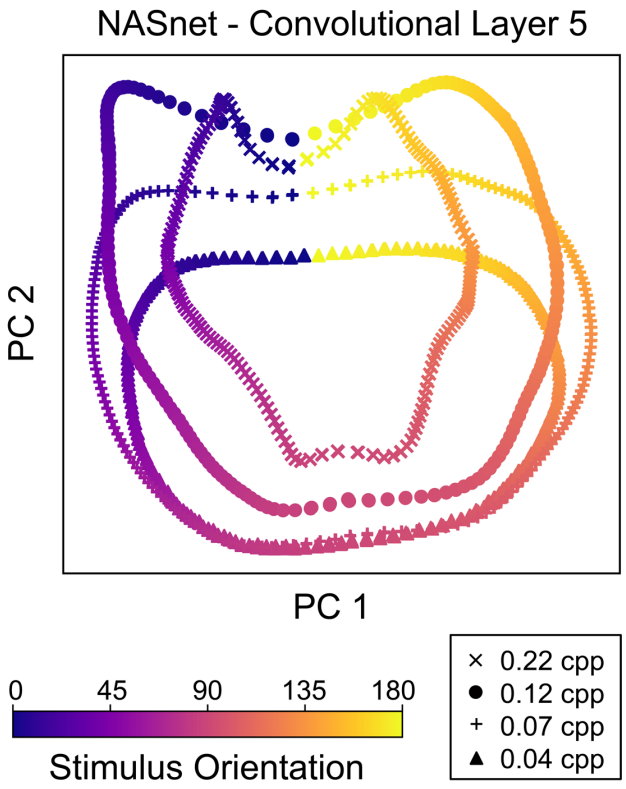

Figure 1. Orientation representations around cardinals are more spaced out than those around obliques. The first two principal components corresponding to each stimulus are plotted for an example layer, with colors indicating orientation and shapes indicating spatial frequency.

Next, we quantified this differential spacing effect by calculating the discriminability at each point in orientation space as described above. We focused first on the noise-free gratings. As shown in Figures 2 and 3, we found that the discriminability between neighboring orientations varied substantially with position in orientation space. Across the middle and late layers of both networks, discriminability was highest at the cardinals and was lowest at the obliques. Interestingly, many layers also showed an additional, smaller, peak centered over the oblique orientations (45 and 135 degrees). This secondary peak is consistent with human psychophysics studies that have found a small boost in performance for orientations centered directly on an oblique. This finding also raises the possibility that the ImageNet dataset might not precisely match the orientation distribution of the natural environment, but may instead have an over-representation of obliques as well as cardinals. Measuring the empirical distribution of orientations in this image database will be an important avenue for future work.

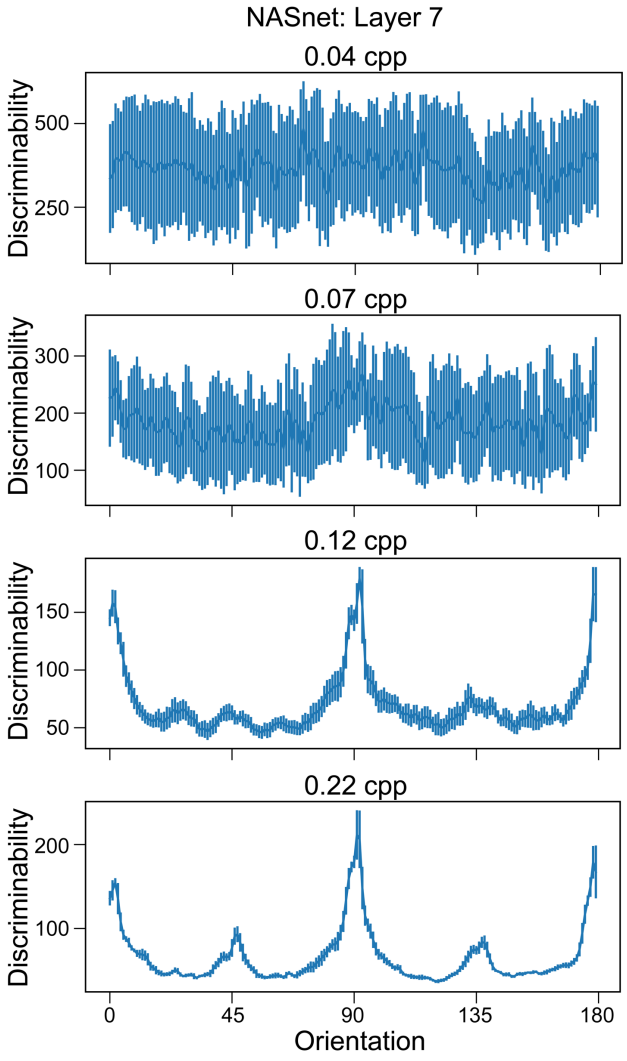

Figure 2. Discriminability is highest at cardinal orientations, especially at high spatial frequencies. Discriminability (Euclidean distance) is plotted versus orientation for an example NASnet layer (Layer 7). Error bars reflect standard deviation over 32 pairwise comparisons.

This discriminability effect was least pronounced at the earliest layers of each network, and became markedly more robust at higher layers. This may suggest that the effect was enhanced by feedforward connections between the earliest layers, as has been suggested to occur between macaque V1 and V2 (Shen et al., 2014). Furthermore, the effect was most pronounced at the highest spatial frequencies (Figure 2), consistent with previous findings that orientation anisotropy in the tuning of single neurons is most robust for neurons preferring higher spatial frequencies (Li et al., 2003; Salinas et al., 2017; Shen et al., 2014).

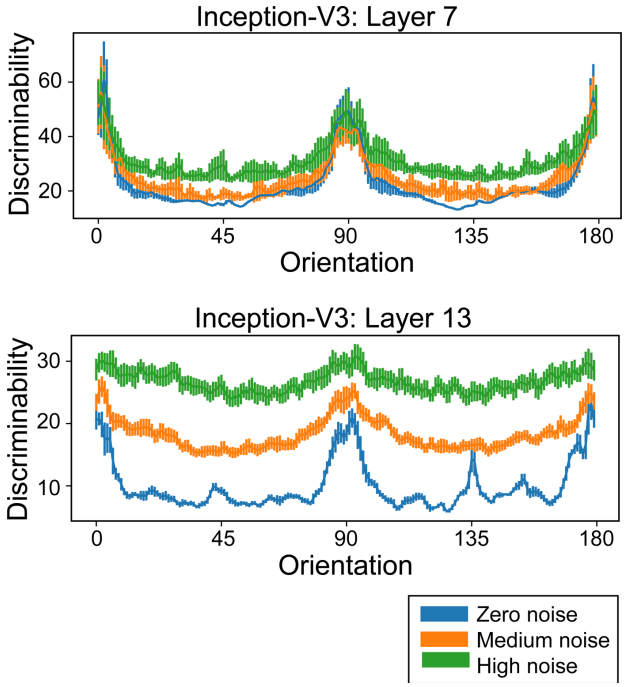

Figure 3. Changes in discriminability across orientation space are less pronounced when Gaussian noise is added to the stimuli. Error bars are as in Figure 2, colors indicate amount of noise.

Finally, we evaluated whether the changes in discriminability across orientation space varied as Gaussian noise was added to the stimuli. This revealed that the overall discriminability between pairs of stimuli increased with noise (seen as an additive shift of the curves in Figure 3). However, increasing noise also decreased the magnitude of the cardinal bias. One interpretation of this is that adding noise masked the oblique effect, similar to the effect of decreasing spatial frequency. Interestingly, recent work has suggested that different types of noise may have opposing effects on cardinal biases in orientation perception (Wei & Stocker, 2015). Future work may focus on comparing the effects of different types of noise, such as bandpass-filtered noise, with the effect of Gaussian noise seen here.

Conclusion

Our results suggest that CNNs, like biological observers, represent stimulus orientations in an anisotropic manner, such that cardinal orientations are more discriminable than obliques. Since this bias was not built into the architecture of the networks, this suggests that cardinal biases can emerge solely as a consequence of experience with natural image statistics. This finding contrasts with results from mice and ferrets, in which cardinal over-representation decreases with experience (Coppola & White, 2004; Hoy & Niell, 2015), but is consistent with findings from primate V2, in which

cardinal biases become stronger with age (Shen et al., 2014). More generally, these findings highlight an example of convergence between CNNs and primate brains, and may inform the future development of more biologically-plausible computer vision models.

Acknowledgments

Funding provided by NEI R01-EY025872 to J.T.S. and a UCSD Institute for Neural Computation predoctoral fellowship awarded to M.H.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *ArXiv*. Retrieved from http://arxiv.org/abs/1605.08695

Bauer, J. A., Owens, D. A., Thomas, J., & Held, R. (1979). Monkeys Show an Oblique Effect. *Perception*, *8*(3), 247–253. http://doi.org/10.1068/p080247

Coppola, D. M., & White, L. E. (2004). Visual experience promotes the isotropic representation of orientation preference. *Visual Neuroscience*, *21*(1), 39–51. http://doi.org/10.1017/s0952523804041045

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). *ImageNet: A large-scale hierarchical image database*. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. http://doi.org/10.1109/CVPRW.2009.5206848

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. http://doi.org/10.1038/nn.2831

Higgins, G. C., & Stultz, K. (1950). Variation of Visual Acuity with Various Test-Object Orientations and Viewing Conditions*. *Journal of the Optical Society of America*, *40*(3), 135. http://doi.org/10.1364/JOSA.40.000135

Hoy, J. L., & Niell, C. M. (2015). Layer-specific refinement of visual cortex function after eye opening in the awake mouse. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(8), 3370–83. http://doi.org/10.1523/JNEUROSCI.3174-14.2015

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *BioRxiv*. http://doi.org/10.1101/408385

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique Effect: A Neural Basis in the Visual Cortex. *Journal of Neurophysiology*, *90*(1), 204–217. http://doi.org/10.1152/jn.00954.2002

Salinas, K. J., Velez, D. X. F., Zeitoun, J. H., Kim, H., & Gandhi, S. P. (2017). Contralateral Bias of High Spatial Frequency Tuning and Cardinal Direction Selectivity in Mouse Visual Cortex. *Journal of Neuroscience*, *37*(42), 10125–10138. http://doi.org/10.1523/JNEUROSCI.1484-17.2017

Shen, G., Tao, X., Zhang, B., Smith, E. L., Chino, Y. M., & Chino, Y. M. (2014). Oblique effect in visual area 2 of macaque monkeys. *Journal of Vision*, *14*(2). http://doi.org/10.1167/14.2.3

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the Inception Architecture for Computer Vision*. http://doi.org/10.1109/CVPR.2016.308

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience*, *18*(10), 1509–1517. http://doi.org/10.1038/nn.4105

Yamins, D., & DiCarlo, J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. http://doi.org/10.1038/nn.4244

Yamins, D., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. http://doi.org/10.1073/pnas.1403112111

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2017). Learning Transferable Architectures for Scalable Image Recognition. Retrieved from https://arxiv.org/abs/1707.07012