Sparse predictive-coding networks account for Bayesian and 'anti-Bayesian' effects in human orientation perception

Stefan P Brugger (BruggerSP@Cardiff.ac.uk)
Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University
Maindy Rd, Cardiff, Wales, UK CF24 4HQ

Christoph Teufel (TeufelC@Cardiff.ac.uk)
Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University
Maindy Rd, Cardiff, Wales, UK CF24 4HQ

Abstract

Natural scenes are dominated by horizontal and vertical local orientations. Bayesian models of vision therefore suggest that the visual system implements a prior biasing orientation perception towards cardinal orientations. The existing evidence, however, suggests that this view may be too simplistic: while neuroimaging studies report neural representations biased towards cardinal orientations, psychophysical work suggests that perceived orientation is biased away from cardinal orientations. Here, we reconcile these findings using neural-network modelling combined with psychophysical testing. We implemented a sparse predictive-coding network as a biologically-plausible model of perception and learning in the visual system. Following training on natural scenes, orientation processing was tested with orientated gratings of varying signal-to-noise ratio. In line with previous work, the network developed orientation-tuned receptive fields. Anisotropy emerged spontaneously, with greater preponderance of units tuned to cardinal than oblique orientations. This non-homogeneity acted as a structural constraint, reproducing the oblique effect seen in human vision, as well as generating attractive biases towards cardinal orientations in neural representations. Importantly, due to lateral inhibition, biases increased with stimulus signal-to-noise ratio. Consequently, in simulated psychophysical experiments, the network reproduced the pattern of apparent repulsive biases seen in human observers. These results are able to reconcile apparently contradictory findings in human psychophysics and visual neuroscience.

Keywords: predictive coding; sparse coding; orientation perception; decoding; vision

Background

Biases in perception are typically thought to reflect the influence of prior expectations on sensory processing. Natural scenes are dominated by horizontal and vertical local orientations, and it has therefore been argued that the visual system implements a prior biasing orientation perception towards cardinal orientations (Girshick et al., 2011).

However, the complex nature of the existing evidence suggests that this view may be too simplistic. Behavioural studies report perceived orientation to be biased away from cardinal orientations, with greater signal-to-noise ratio leading associated with lower bias (de Gardelle et al., 2010). Recent accounts explain this apparently 'anti-Bayesian' effect via the phenomenon of likelihood repulsion: a bias away from regions of high encoding precision, which, in the case of orientation correspond to cardinal orientations (Wei & Stocker, 2015; Hahn & Wei, 2024).

These behaviorally-measured biases contrast sharply with the findings of neuroimaging studies, which, in line with more conventional Bayesian models, suggest that neural representations are biased towards cardinal orientations (Harrison et al., 2023). Furthermore, one weakness of behavioural approaches is that it is only possible to measure a relative perceptual bias: the bias for one stimulus relative to bias in another (reference) stimulus.

We attempt to reconcile these divergent findings. We draw upon sparse predictive coding as a biologically plausible framework for how the brain performs inference and learning: by minimizing mismatch between current sensory inputs and those predicted by a probabilistic generative model of the environment. Predictive coding is able to account for a wealth of phenomena in early visual cortex (Rao & Ballard, 1999; Spratling, 2010), and comes with its own claims to optimally (Friston & Kiebel, 2009).

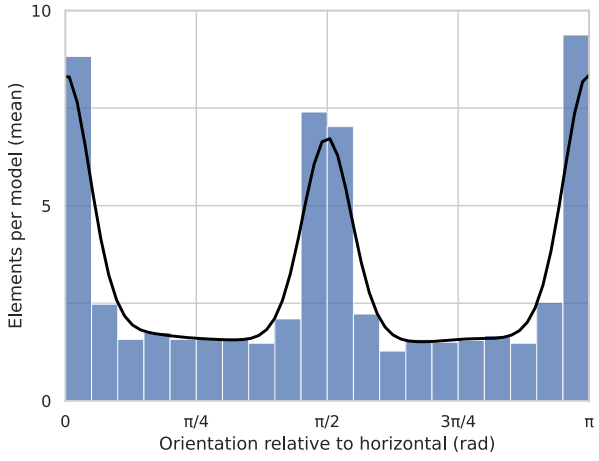

Figure 1: Distribution of learned receptive field orientations for 40 models trained on COCO images.

Methods

Network architecture. We implemented the predictive coding model as a simple convolutional neural network. All work described here employed a single $n_c = 64$ channel transposed convolutional layer. Early work on predictive coding incorporated a sparsity-inducing prior on neuronal activations (Rao & Ballard, 1999; Boutin et al., 2021), in line with sparse coding approaches (Olshausen & Field, 1996). Here, we introduce a biologically inspired sparsity-inducing mechanism into the generative model itself. This mechanism takes the form of a spatially localised Softmax Linear Unit (SoLU) non-linearity (Elhage, 2022), applied to neuronal activations v:

$$SoLU(v_i) = v_i \sigma(v_i) = \frac{e^{v_i}}{\sum_{i=k_0}^{k_1} e^{v_i}} v_i$$

Neurons thus compete with others within their local neighbourhood V_K, defined by a square-shaped region centred on the neuron's location, i.e. comprising those neurons V_K, $\{v_i \in V_K | D_{Ch}(v_i, v_k) \leq n_p\}$, with a Chebyshev (chessboard/King's move) distance D_{Ch} of at most n_p pixels. The output of the non-linearity is passed to the transposed convolutional layer $\mathbf{C}^T$ to reconstruct the image. We do not employ precision-weighting: the objective function is simply a sum of squared prediction errors:

$$F = \frac{1}{2}(x - \mathbf{C}^T SoLU(v))^2$$

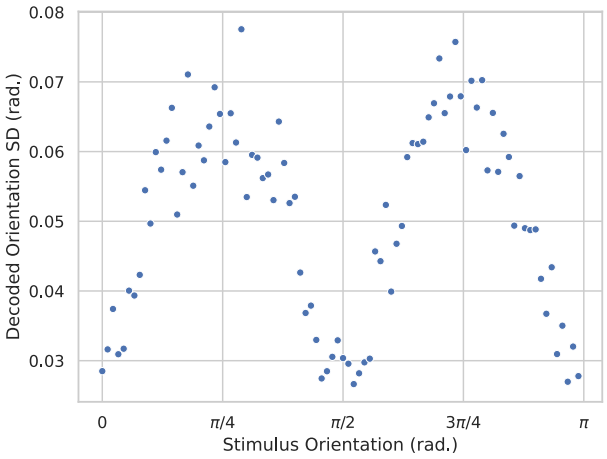

Figure 2: Orientation sensitivity, measured by the standard deviation of the vector average readout over 20 trials.

Training & Testing Models were trained on the COCO dataset (Lin et al., 2014). Images were prefiltered with a Laplacian of Gaussian filter to simulate sub-cortical processing. We trained a population of 40 single-layer models, all with 64 channels and a convolution kernel size of 11x11 pixels. Values of n_p ranged from 1 to 5 pixels. Orientation of learned elements was estimated using the structure tensor method (Bigün, 1988). Models were tested on orientated grating stimuli of different signal-to-noise ratio, manipulated by varying contrast. Model estimates of orientation were decoded from neuronal activations using a standard population vector average decoding approach (Georgopoulos et al., 1983).

Results

Models learned Gabor-like receptive field structures, with spontaneously emerging anisotropy in orientation distribution strongly favoring cardinal orientations (Figure 1). This induced a marked oblique effect (Figure 2). In line with neuroimaging studies, the network's representation of orientation was biased towards cardinal orientations. These biases increased in line with increasing signal-to-noise-ratio, contrasting to the predictions of standard Bayesian observer models (figure 3; solid lines). Furthermore, when tested in a 2-interval-forced-choice paradigm characteristic of human psychophysics, employing a test stimulus with low or moderate signal-to-noise-ratio and a reference with high signal-to-noise-ratio, the network reproduced the pattern of apparent repulsive biases reported in human observers (de Gardelle et al., 2010).

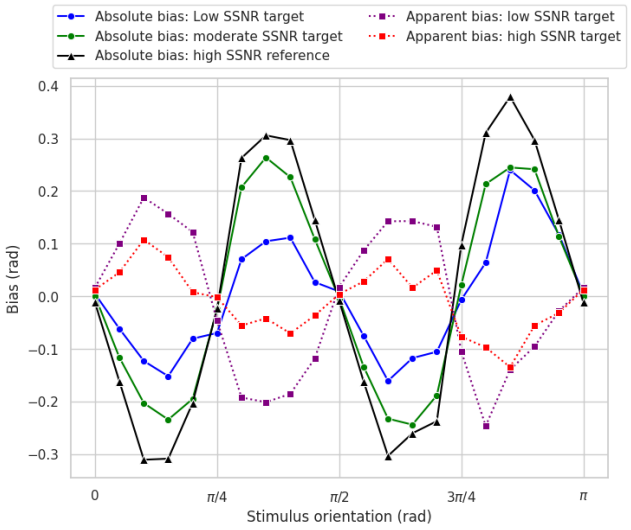

Figure 3: Decoded absolute biases (attractive; solid lines) and relative or apparent biases from simulated psychophysical study (repulsive; dotted lines). Apparent repulsive biases follow reorted findings in humans: greater repulsion is seen for lower (purple line) than for higher (red line) signal-to-noise ratio in the test stimulus.

Conclusion

Our findings demonstrate that, when optimised for natural images, sparse predictive-coding networks spontaneously learn structural constraints leading to counter-intuitive effects in tests with artificial stimuli. These effects mirror those seen in human observers. Overall, our network is able to reconcile apparently contradictory results from psychophysical and neuroimaging studies on human orientation perception.

Acknowledgments

This work was supported in part by grant MR/N0137941/1 for the GW4 BIOMED MRC DTP, awarded to the universities of Bath, Bristol, Cardiff, and Exeter from the Medical Research Council (MRC) and UK Research and Innovation (UKRI).

References

Bigün, J. (1988). Pattern recognition by detection of local symmetries. In *Machine intelligence and pattern recognition* (p. 75–90). Elsevier. doi: 10.1016/b978-0-444-87137-4.50012-6

Boutin, V., Franciosini, A., Chavane, F., Ruffier, F., & Perrinet, L. (2021). Sparse deep predictive coding captures contour integration capabilities of the early visual system. *PLOS Computational Biology*, *17*(1), e1008629.

de Gardelle, V., Kouider, S., & Sackur, J. (2010). An oblique illusion modulated by visibility: Non-monotonic sensory integration in orientation processing. *Journal of Vision*, *10*(10), 6–6. Retrieved from http://dx.doi.org/10.1167/10.10.6 doi: 10.1167/10.10.6

Elhage, N. (2022). *Softmax linear units.* https://transformer-circuits.pub/2022/solu/index.html. Anthropic AI.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. Retrieved from http://dx.doi.org/10.1098/rstb.2008.0300 doi: 10.1098/rstb.2008.0300

Georgopoulos, A. P., Caminiti, R., Kalaska, J. F., & Massey, J. T. (1983). Spatial coding of movement: A hypothesis concerning the coding of movement direction by motor cortical populations. In *Experimental brain research supplementum* (p. 327–336). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-68915-4$_3$4

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. doi: 10.1038/nn.2831

Hahn, M., & Wei, X.-X. (2024). A unifying theory explains seemingly contradictory biases in perceptual estimation. *Nature Neuroscience*, *27*(4), 793–804. doi: 10.1038/s41593-024-01574-x

Harrison, W. J., Bays, P. M., & Rideaux, R. (2023). Neural tuning instantiates prior expectations in the human visual system. *Nature Communications*, *14*(1). doi: 10.1038/s41467-023-41027-w

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2014). Microsoft coco: Common objects in context.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. Retrieved from http://dx.doi.org/10.1038/381607a0 doi: 10.1038/381607a0

Rao, R. P. N., & Ballard, D. H. (1999, jan). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. Retrieved from http://dx.doi.org/10.1038/4580 doi: 10.1038/4580

Spratling, M. W. (2010, mar). Predictive coding as a model of response properties in cortical area v1. *The Journal of Neuroscience*, *30*(9), 3531–3543. Retrieved from http://dx.doi.org/10.1523/JNEUROSCI.4911-09.2010 doi: 10.1523/jneurosci.4911-09.2010

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. Retrieved from http://dx.doi.org/10.1038/nn.4105 doi: 10.1038/nn.4105