Shared visual illusions between humans and artificial neural networks

Ari S. Benjamin[1*], **Cheng Qiu**[2*], **Ling-Qi Zhang**[2*], **Konrad P. Kording**[1], **Alan A. Stocker**[2]
[1]Department of Bioengineering
[2]Department of Psychology
University of Pennsylvania
Philadelphia, PA, 19104, USA

Abstract

Any information processing system should allocate resources where it matters: it should process frequent variable values with higher accuracy than less frequent ones. While this strategy minimizes average error, it also introduces an estimation bias. For example, human subjects perceive local visual orientation with a bias away from the orientations that occur most frequently in the natural world. Here, using an information theoretic measure, we show that pretrained neural networks, like humans, have internal representations that overrepresent frequent variable values at the expense of certainty for less common values. Furthermore, we demonstrate that optimized readouts of local visual orientation from these networks' internal representations show similar orientation biases and geometric illusions as human subjects. This surprising similarity illustrates that when performing the same perceptual task, similar characteristic illusions and biases emerge for any optimal information processing system that is resource limited.

Keywords: neural networks; visual illusions; perceptual biases; psychophysics; uncertainty

Introduction

When inferring certain features about the world from sensory inputs, one inevitably must deal with any uncertainty introduced by sensory noise or external ambiguity. If one has a fixed budget of resources with which to minimize uncertainty, it is clear that one should preferentially minimize uncertainty about important aspects of the world, such as those that occur very frequently or that are important for behavior. This general principle, often referred to as Barlows efficiency principle (Barlow et al., 1961), should apply to any efficient system, either biological or artificial.

In the study of human perception, our uncertainty about the world is often quantified by discrimination thresholds and by perceptual errors. This vein of psychophysical studies has revealed a great deal about the processes underlying human perception (Fechner, 1860; Körding & Wolpert, 2004). Many of these findings conform to the predictions of efficiency given physical constraints and behavioral needs. Recently, such ideas have helped to explain a class of illusions in which the perceived value of a feature can be counterintuitively biased away from more probable values under the prior distribution (Wei & Stocker, 2015, 2017). This effect is explained as deriving, ultimately, from having greater uncertainty on less probable values of a feature, as expected from efficiently allocating resources with which to minimize uncertainty. This means that many potential low-probability stimuli may be compatible with the same activity pattern.

An open debate in neuroscience is the precise sense in which (artificial) deep neural networks (DNNs) reproduce aspects of human vision. While many surprising similarities have been noted (Glaser, Benjamin, Farhoodi, & Kording, 2019; Hassabis, Kumaran, Summerfield, & Botvinick, 2017), it is not obvious that DNNs trained on image classification tasks should share perceptual biases with humans. On the one hand, neural networks may be different from brains, e.g. by having no internal noise. On the other hand, neural networks do have noise related to learning, must represent countless potentially useful variables with a finite number of nodes, and must resolve unavoidable ambiguity. As such, their behavior may be similar.

Here we ask whether deep networks have greater uncertainty for rarer variables, as in humans. We also test whether this results in human-like perceptual biases and illusions when

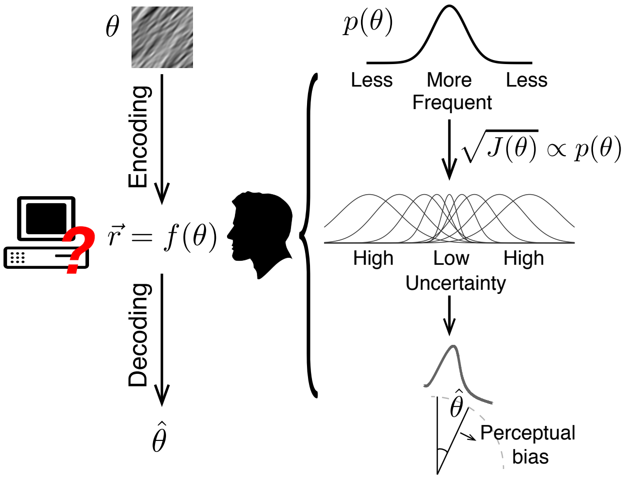

Figure 1: Perception is often described as a process of encoding and decoding variables in the world. Perceptual biases emerge when internal representations are more accurate for frequent values of variables and concomitantly less accurate for less frequent values. This has been well-established empirically for human observers, but not yet for deep neural networks.

[*]Denotes equal contribution

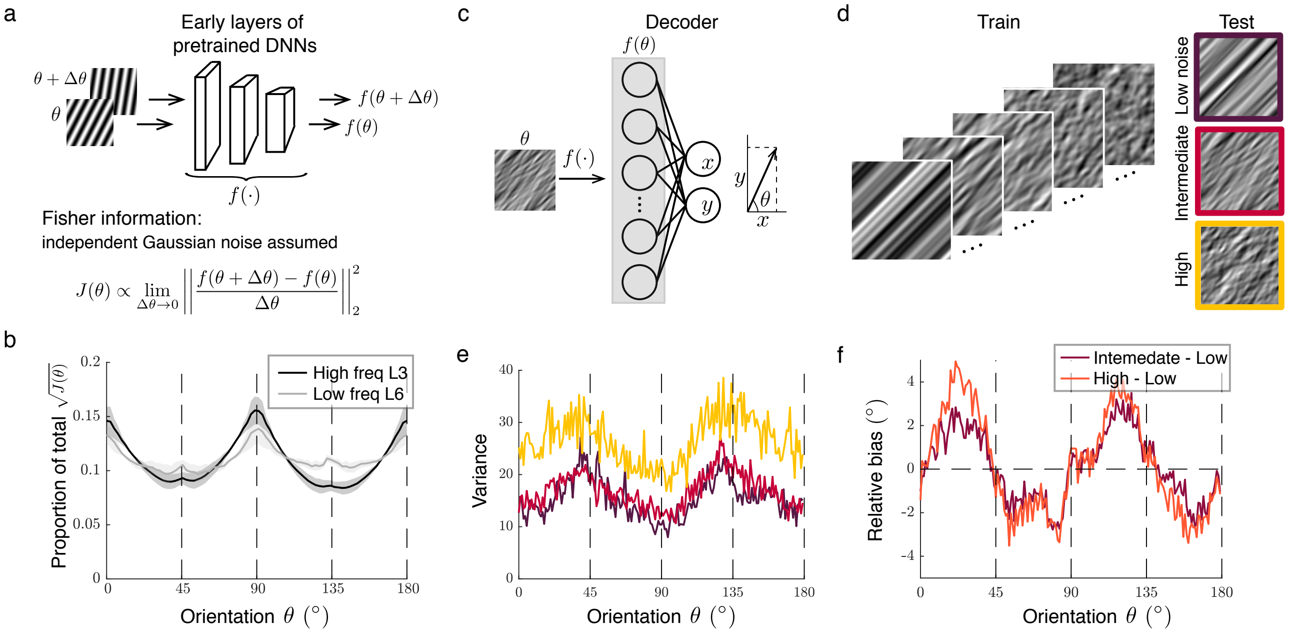

Figure 2: Representation uncertainty for orientations and perceived biases in neural networks. (a) Fisher information as a measure of representation uncertainty in pretrained convolutional neural networks. Image patches of oriented sinusoidal gratings are forwarded through early layers of a pretrained neural network to calculate Fisher information in network representations. (b) The normalized square root of Fisher information measured in early layers of the pretrained AlexNet (Krizhevsky et al., 2012). The patterns indicate higher sensitivity (lower uncertainty) at cardinal orientations, and are consistent with the prior distribution of orientations in natural scene statistics, and such pattern for low spatial frequency patches emerges later in the network. (c) Fully connected decoding layers fine-tuned for network reporting perceived orientation. (d) Training samples are lowpass-filtered spatial frequency noise textures each with various level of bandpass orientation noise. Image patches with low, intermediate, and high noise in orientations are tested. (e) Variance of the network perceived orientation nicely tracks (the inverse of) Fisher information. (f) Larger repulsive biases (away from the nearest cardinal orientation) are shown for larger noise magnitudes.

such variables are decoded from network representations.

Results

We examined the uncertainty of the internal representations and decoding biases in pretrained deep neural networks (DNNs) for two simple tasks: estimating the absolute orientation of image patches with sinusoidal gratings or noise textures, and estimating the relative orientation between two lines. Humans are known to exhibit characteristic, nonuniform patterns of discrimination thresholds and biases indicating higher sensitivity to a certain range of orientations in both tasks.

We first quantified the representation uncertainty using Fisher information. We focused on early layers of the pretrained network, which show selectivity to the low-level visual variables we investigate here. A pretrained neural network defines a deterministic mapping from the stimulus to a population response vector, $\vec{r} = f(\theta)$ where θ is a scalar variable that controls some aspect of the stimulus. We can use Fisher information, $J(\theta)$ to measure the uncertainty of $\vec{r}$ as one changes θ (Seriès, Stocker, & Simoncelli, 2009; Berardino, Laparra, Ball, & Simoncelli, 2017). Specifically, assuming independent Gaussian noise, Fisher information simplifies to $J(\theta) = \|\partial f / \partial\theta\|_2^2$. Thus, in the presence of either noise or nuisance variability, the sensitivity of the internal representation determines how accurately information can be read out.

Perceived orientation in neural networks

It has been shown that human observers are better in detecting small changes around cardinal (i.e. vertical and horizontal) rather than oblique orientations. Since cardinal orientations are more prevalent than oblique ones, this indicates that representations of orientation in the human visual system are adapted to the orientation statistics of our natural environment (Girshick, Landy, & Simoncelli, 2011). In particular, the visual system minimizes uncertainty for the most frequent feature values, Assuming efficient encoding, this leads to systematic bias in perceived orientations (Wei & Stocker, 2015, 2017). Here, we probed whether a convolutional neural network pretrained on image classification shares the same encoding-decoding properties.

Representation sensitivity To calculate Fisher information $J(\theta)$, we used sinusoidal grating patches where θ controls the overall orientation in the patch (Fig. 2a). We found that

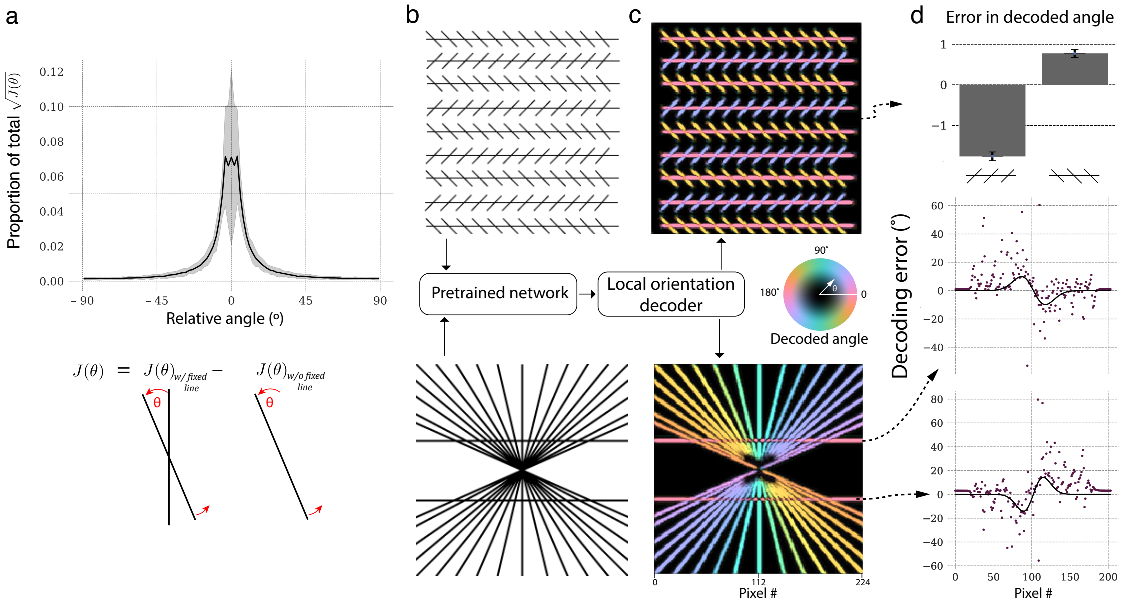

Figure 3: Local orientations decoded from pretrained DNNs show biases for geometric illusions that are consistent with human observers. a) The Fisher information of the relative angle between two lines is higher for very acute angles. To ensure that this does not reflect the Fisher of the absolute angle, we subtracted the Fisher with respect to the same rotating line but without the fixed crossed line. b) We trained a new decoder to output the local orientation at each pixel of an image fed to the VGG16 network. We then tested it with various geometric illusions (the Zllner illusion, top, and the Hering illusion, bottom). c) The angle of the decoded local orientation at each pixel is represented here as hue. The value (lightness) of each pixel corresponds to the magnitude of the orientation vector. d) The error of the decoder is in the direction expected given the illusory percept. For the Zllner illusion, top, we plot the average error over a left-hatched and right-hatched horizontal line (with error bars showing the S.E.M. of pixel errors). As expected, right-hatched lines are estimated as too clockwise-rotated (negative error) and left-hatched lines are too counterclockwise-rotated (positive). For the Hering illusion, we show the error along the top horizontal line (middle plot), which first bends upwards (positive error) before erroneously bending downwards (negative). The error along the lower horizontal line shows the opposite trend.

Fisher information measured in early layers of the pretrained AlexNet clearly reflects the distribution of local visual orientation in natural images (Fig. 2b). In particular it is proportional to the square of the prior distribution, which is expected when the representation is learned with the constraint of finite resources, e.g. in efficient coding (Wei & Stocker, 2015).

Perceived orientation bias For the network to “report” its perceived orientation, we added an additional decoding layer with two fully connected output units that we trained to output orientation in the form of a vector with coordinates x and y. The training and test images are bandpass-filtered, noisy orientations textures (Fig. 2d). Figures 2e,f show variance and relative bias between high noise and low noise orientation patches of the network. Smaller variance corresponds to larger Fisher information. The network also shows bias patterns similar to those of human observers where, with larger repulsive biases for larger noise levels (de Gardelle, Kouider, & Sackur, 2010; Wei & Stocker, 2015).

Relative Orientation and Geometric Illusions

Humans consistently perceive acute angles as wider than they are (Carpenter & Blakemore, 1973; Heywood & Chessell, 1977). This effect is closely related to various geometric illusions consisting of intersecting lines, such as the Zollner and Hering illusions, in which measurably straight and parallel lines appear to bend, converge, or diverge, always in manners consistent with acute angles appearing erroneously wide. It has been noted that the distribution of angles in typical scenes is peaked at small angles, which means that the overestimation of an angle is in the direction away from more probable values (Nundy, Lotto, Coppola, Shimpi, & Purves, 2000; Howe & Purves, 2005). This raises the possibility that DNNs would also overrepresent small angles at the expense of larger ones, leading to decoded orientations consistent with common geometric illusions.

Representation sensitivity We first examined if DNNs trained on ImageNet are sensitive to relative angle in a manner consistent with the natural distribution of angles. We built stimuli consisting of two crossed lines, fed them to a pretrained

VGG16 network (Simonyan & Zisserman, 2015), and examined the activations of an early convolutional layer (before the 1st maxpool). We calculated the derivative of the activations with respect to relative angle with a finite-difference method; by holding one line fixed and perturbing the orientation of the other. Since here we were interested in the Fisher information with respect to the relative angle, and not the absolute angle of the line (which also affects the Fisher information), we subtracted the Fisher information with respect to the same rotating line but without the fixed crossed line, and furthermore marginalized over the orientation of the fixed line. The result is the sensitivity to the relative angle. We found that the Fisher information was larger for smaller relative angles, consistent with the distribution previously found to exist in the natural world.

Geometric illusions We then trained a decoder to estimate the local orientation of every pixel in an input image from the DNN activations, and asked if this would show effects consistent with Zollner and Hering illusions. In order to train this decoder, we generated tens of thousands of images of many crossed, black lines of various curvatures, and then minimized the mean-squared error of the pixel-wise orientation angle and magnitude (as output by a convolved quadrature filter with kernel of approximately 10 pixels). In Fig. 3c, we show the output for the tested illusions, and Fig. 3d shows the error on the angle outputs along the key slices. The errors were in the direction consistent with the visual percept.

Conclusion

We have found that when DNNs are trained to classify images in ImageNet, they learn representations that reflect the statistics of the natural world in a similar way as humans. Uncertainty about low-level visual features (here, orientation and angle) is decreased in proportion to that feature's occurrence in the world. When decoding this information either with a readout network, as shown here, or perhaps internally by downstream layers this results in perceptual biases.

It will be interesting to investigate the degree to which this result reflects the distribution of the labels in ImageNet. One can imagine a task (i.e. an allocation of labels) for which a DNN requires only a subset of all low-level features of the world. If, say, the task required knowledge of which locations had angles near 45°, the DNN would learn to overrepresent that angle despite its relative under-occurence in the data. Thus, the extent to which the Fisher information is proportional to a prior distribution of a low-level feature indicates how equally the information about the label is distributed across that feature. For complex features closer to object identities, we expect the Fisher to diverge from the prior distribution.

Acknowledgments

A.B. and K.K. acknowledge NIH grant number MH103910.

References

Barlow, H. B., et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*, 217–234.

Berardino, A., Laparra, V., Ball, J., & Simoncelli, E. (2017). Eigen-Distortions of Hierarchical Representations. In *31st Conference on Neural Information Processing Systems* (p. 10). Long Beach, CA, USA.

Carpenter, R., & Blakemore, C. (1973). Interactions between orientations in human vision. *Experimental Brain Research*, *18*.

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.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig, Germany: Breitkopf und Haertel.

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.

Glaser, J. I., Benjamin, A. S., Farhoodi, R., & Kording, K. P. (2019). The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology*, *175*, 126–137.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258.

Heywood, S., & Chessell, K. (1977). Expanding Angles? Systematic Distortions of Space Involving Angular Figures. *Perception*, *6*(5), 571–582.

Howe, C. Q., & Purves, D. (2005). Natural-scene geometry predicts the perception of angles and line orientation. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1228–1233.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Nundy, S., Lotto, B., Coppola, D., Shimpi, A., & Purves, D. (2000). Why are angles misperceived? *Proceedings of the National Academy of Sciences*, *97*(10), 5592–5597.

Seriès, P., Stocker, A. A., & Simoncelli, E. P. (2009). Is the homunculus aware of sensory adaptation? *Neural Computation*, *21*(12), 3271–3304.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015.*

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, *18*(10).

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244–10249.