Generative perceptual inference in deep neural network models of object recognition induces illusory contours and shapes

Tahereh Toosi (tahereh.toosi@columbia.edu)
Kenneth D. Miller (ken@neurotheory.columbia.edu)

Center for Theoretical Neuroscience
Zuckerman Mind Brain Behavior Institute, Columbia University, 3227 Broadway
New York, New York 10027

Abstract

Illusory contours and shapes highlight the striking gap between how natural and artificial vision perceive the world. In this study, we show that a pattern recognition model embodies a generative model that integrates perceptual priors and the sensory processing. We introduce a novel perceptual algorithm, Generative Perceptual Inference (GPI), which iteratively updates the activations by accumulating propagated error in the early layers. Given a Kanizsa square as input to a deep neural network (DNN) optimized for robust object classification, our results show that running GPI led to the emergence of edge-like patterns in the area of the perceived 'white square'. Moreover, when GPI is applied to the same DNN with Rubin's vase image as input, it creates a vase-like pattern, while GPI in a DNN with the same architecture but optimized for face recognition creates face-like patterns. Thus, we found the direct link between natural image prior and perception of illusory contours and shapes, through an image-computable algorithm that captures experimental findings regarding processing of illusions in animals and humans. More broadly, this work reconciles the views of the visual cortex as both a pattern recognition and a generative model in a unified framework.

Keywords: visual illusions; object recognition; deep neural networks; perceptual inference

Introduction

Integrating perceptual priors with sensory inputs, known as perceptual inference, facilitates the brain's interpretation of ambiguous or complex stimuli by leveraging previously acquired knowledge, stored as internal models, to enhance current sensory processing. However, the neural mechanisms that implement such a generative internal model remain elusive. In contrast, the view of the brain as a pattern recognition system has been successful in predicting the neural activity patterns using DNNs and offers a (crude but well-defined) mechanistic mapping to the stage of processing along the ventral pathway (Yamins et al., 2014). However, this account starts to fail in the face of challenging and degraded stimuli (Geirhos et al., 2018), or even completely fails to explain visual illusions (Baker, Erlikhman, Kellman, & Lu, 2018).

We hypothesized that the network during pattern recognition training constructs an implicit internal model about the distribution of the data it was trained on (e.g. natural image prior in the case of training on ImageNet). When faced with noisy, degraded, or unusual stimuli, such as high-saturated images, and using a proper inference algorithm, this internal model can be queried for priors regarding this image to aid perception. The ability to access implicit priors imposes constraints on network architecture, objective function, and the learning rule (Kadkhodaie & Simoncelli, 2021; Toosi & Issa, 2023), which could limit the functionality of the network. For instance, denoiser autoencoders or other generative models fall short on object recognition, and they do not provide a good explanation for the pattern of neural activity in the brain (Schrimpf et al., 2018). We aim to devise an inference algorithm that can extract priors from a pattern recognition network. It is important to note that, in contrast to previous attempts to model illusions, our model aims to adhere the main function of ventral stream, i.e. object recognition, and shows direct image-computable connection between illusory contours and shapes and the priors. In particular, the DNNs used in this study were neither specifically enhanced with feedback capabilities nor optimized to detect or decode these illusory contours (Pang, O'May, Choksi, & VanRullen, 2021); parameters were only trained for robust object recognition.

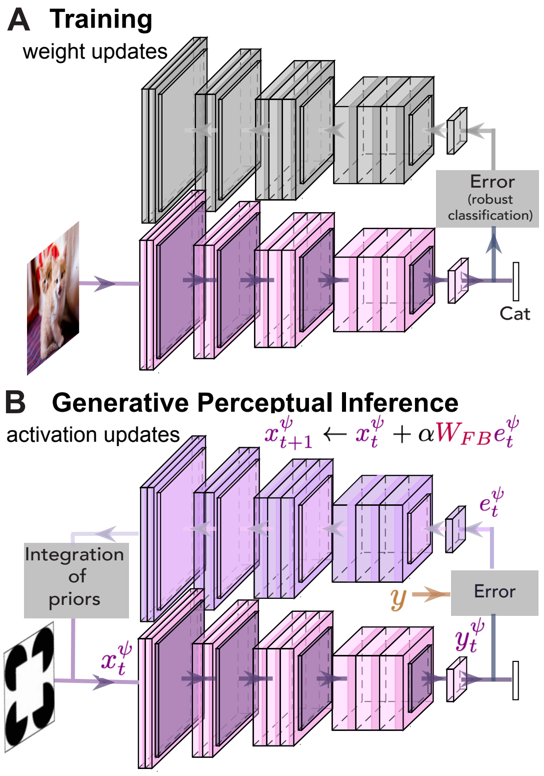

Figure 1: Generative perceptual inference **A** Training network: Parameters of the feedforward DNN are tuned for robust object classification, e.g. 1000-way classification on Imagenet. Feedback network shown in gray is the error propagation network used in BackPropagation (BP) or bio-plausible variants of BP. **B** Neural activation updates according to GPI in the same DNN. Activations x_t^{Ψ} in the network are iteratively updated by the integrating the adjusted propagated error e_t^{Ψ} through the feedback W_{FB}. Error can be computed as mean square difference to the pure sensory activation, or a target value in last layer both denoted by y. For simplicity, the normalizations for adjustment of the values are not included. α is the learning rate for activation updates

Illusions such as Kanizsa's square or Rubin's vase (AKA face-or-vase) have been studied for decades in animals, humans and in both healthy and neural disorders. Although some viewed visual illusions as perceptual errors and others linked them to perceptual priors, there is no precise mechanistic explanation for why we perceive illusions. Research in animals and humans suggests that 1) activity in early visual areas represents the perceptual state when viewing a visual illusion, and these activities build up over time (Lee & Nguyen, 2001; Parkkonen, Andersson, Hämäläinen, & Hari, 2008) 2) Illusory counters only appear in superficial layers (layer 2/3) but not in the input layer (layer 4) of early visual areas (Lee & Nguyen, 2001; Shin et al., 2023) 3) feedback connections play a causal role in inducing these activities in early visual areas (Pak, Ryu, Li, & Chubykin, 2019). We hypothesized that these mechanistic clues that are shared between species point to a general framework for perceptual inference. In this work, we introduce "Generative Perceptual Inference" in which we postulate that in face of degraded, noisy or stimuli with unusual statistics (e.g. high saturated in case of Kanizsa square and Rubin's face, inference prolongs as the result of accumulation of perceptual priors stored in synaptic weights. We show that a simple realization of GPI in an off-the-shelf feedforward neural network trained for robust object recognition induces the illusory contours and shapes in neural networks.

Generative perceptual inference in pattern recognition networks

Architecture and training. Our objective is to design an inference algorithm to integrate the priors learned during training. Usually, implicit priors are obtained through generative models that are explicitly trained to estimate the prior, but previous work showed that a feedforward neural network can give access to implicit priors using the intrinsic feedback structure normally used for backpropagation of error (Toosi & Issa, 2023). Here, we show that in a feedforward architecture trained for pattern recognition (robust object classification), we can estimate the implicit prior by backpropagated error to the early layers (results not included), thereby enabling the pattern recognition model to exhibit inferential properties often pertained to generative models (Figure 1 **A**).

Inference. In early layers, the backpropagated errors are adjusted and added to the current neural activation in early layers to obtain the updated inferred neural activation, and this iteration continues until convergence (Figure 1 **B**). The error function here need not be the same error function used during training (object classification); rather, it can compute the error to the representation of the pure sensory information.

Results

We took an off-the-shelf model of robust object recognition, this model is capable of 1000-way image classification on ImageNet and has already been shown to be a good predictor of pattern of neural activity (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2018). Without any additional training, we feed it the illusion-induced images (Kanisza square and Rubin's vase) and implement the GPI algorithm as explained above. Figure 2 shows examples of activation patterns generated by GPI. This induced activity in early layers captures the experimental findings indicating the representation of induced contours and shapes in ealy visual areas (Lee & Nguyen, 2001; Pak et al., 2019). Moreover, GPI confirms the causal role of feedback in integrating priors and inducing illusions is pivotal, as found in optogenetic study in mice (Pak et al., 2019).

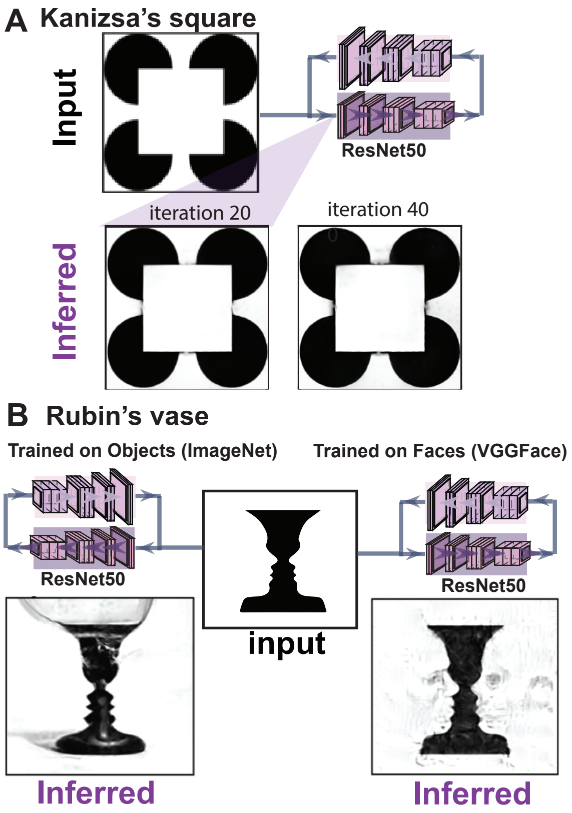

Figure 2: Generated inferred activity by GPI in response to illusory images. **A** When the input is Kanizsa square, the updated activity in the input to the inference channel is depicted at iterations 20 and 40. **B** When the input is Rubin's face, GPI in the network that was trained for robust face (object) classification generates face-like (vase-like) patterns in the input to the inference channel.

Conclusions

We show a pattern recognition model embodies a generative model which we could query by our proposed inference algorithm, accounting for experimental findings on processing illusory contours and shape in animals and humans. Although our model is not a dynamical model, it shows the principles behind updating the activations over time, which could be leveraged in designing dynamical models to exhibit prior integration over time. Our work is the first instance of directly showing how natural image priors induce illusory contours and shape in an image-computable model capable of object recognition.

Acknowledgments

T.T. is supported by NIH 1K99EY035357-01. This work was also supported by NIH RF1DA056397, NSF 1707398, Gatsby Charitable Foundation GAT3708, and was performed using the Columbia Zuckerman Axon GPU cluster.

References

Baker, N., Erlikhman, G., Kellman, P. J., & Lu, H. (2018). Deep convolutional networks do not perceive illusory contours. *Cognitive Science*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Proceedings of the 32nd international conference on neural information processing systems* (p. 7549–7561). Red Hook, NY, USA: Curran Associates Inc.

Kadkhodaie, Z., & Simoncelli, E. P. (2021). *Solving linear inverse problems using the prior implicit in a denoiser.*

Lee, T. S., & Nguyen, M. (2001). Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences*, *98*(4), 1907-1911. doi: 10.1073/pnas.98.4.1907

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International conference on learning representations.*

Pak, A., Ryu, E., Li, C., & Chubykin, A. A. (2019, December). Top-down feedback controls the cortical representation of illusory contours in mouse primary visual cortex. *The Journal of Neuroscience*, *40*(3), 648–660.

Pang, Z., O'May, C. B., Choksi, B., & VanRullen, R. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks*, *144*, 164-175. doi: https://doi.org/10.1016/j.neunet.2021.08.024

Parkkonen, L., Andersson, J., Hämäläinen, M., & Hari, R. (2008). Early visual brain areas reflect the percept of an ambiguous scene. *Proceedings of the National Academy of Sciences*, *105*(51), 20500-20504. doi: 10.1073/pnas.0810966105

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., . . . DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*.

Shin, H., Ogando, M. B., Abdeladim, L., Durand, S., Belski, H., Cabasco, H., . . . Adesnik, H. (2023, June). Recurrent pattern completion drives the neocortical representation of sensory inference.

Toosi, T., & Issa, E. (2023). Brain-like flexible visual inference by harnessing feedback feedforward alignment. In *Advances in neural information processing systems* (Vol. 36, pp. 56979–56997). Curran Associates, Inc.

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