A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception

William Lotter[1], **Gabriel Kreiman**[1], **David Cox**[1,2,3]
[1]Harvard University, [2]MIT-IBM Watson AI Lab, [3]IBM Research
lotter.bill1@gmail.com, gabriel.kreiman@tch.harvard.edu, david.d.cox@ibm.com

Abstract

While deep neural networks take loose inspiration from neuroscience, it is an open question how seriously to take the analogies between artificial deep networks and biological neuronal systems. Interestingly, recent work has shown that deep convolutional neural networks (CNNs) trained on large-scale image recognition tasks can serve as strikingly good models for predicting the responses of neurons in visual cortex to visual stimuli, suggesting that analogies between artificial and biological neural networks may be more than superficial. However, while CNNs capture key properties of the average responses of cortical neurons, they fail to explain other properties of these neurons. For one, CNNs typically require large quantities of labeled input data for training. Our own brains, in contrast, rarely have access to this kind of supervision, so to the extent that representations are similar between CNNs and brains, this similarity must arise via different training paths. In addition, neurons in visual cortex produce complex time-varying responses even to static inputs, and they dynamically tune themselves to temporal regularities in the visual environment. We argue that these differences are clues to fundamental differences between the computations performed in the brain and in deep networks. To begin to close the gap, here we study the emergent properties of a previously-described recurrent generative network that is trained to predict future video frames in a self-supervised manner. Remarkably, the model is able to capture a wide variety of seemingly disparate phenomena observed in visual cortex, ranging from single unit response dynamics to complex perceptual motion illusions. These results suggest potentially deep connections between recurrent predictive neural network models and the brain, providing new leads that can enrich both fields.

1 Introduction

The fields of neuroscience and machine learning have long enjoyed productive dialogue, with neuroscience offering inspiration for how artificial systems can be constructed, and machine learning providing tools for modeling and understanding biological neural systems. Recently, as deep convolutional neural networks (CNNs) have emerged as leading systems for visual recognition tasks, these same models have emerged—without any modification or tailoring to purpose—as leading models for explaining the population responses of neurons in primate visual cortex [54, 53, 16]. These results suggest that the connections between artificial deep networks and brains may be more than skin deep.

However, while deep CNNs capture some important details of the responses of visual cortical neurons, they fail to explain other key properties of the brain. Notably, the level of strong supervision used to train state-of-the-art CNNs is much greater than that available to our brain. To the extent that representations in the brain are similar to those in CNNs trained on e.g. ImageNet, the brain must be

Preprint. Work in progress.

arriving at these representations by different, largely unsupervised routes. Another key difference is that CNNs are fundamentally static and lack a notion of time, whereas neuronal systems are highly dynamic, producing responses that vary dramatically in time, even in response to static inputs. Figure 2a shows a typical response profile of a visual cortical neuron to a static input [37]. The neuron produces a brief transient response to the onset of the visual stimulus, followed by near total suppression of that response. When the stimulus is removed, the neuron responds again with a transient burst of activity (known as an "off" response). Neurons throughout visual cortex show a variety of dynamic response profiles, and the computational purpose of these dynamics is currently not well understood.

To further complicate matters, the responses of neurons in the primate visual cortex are also sensitive to long range temporal structure in the visual world. For instance, Meyer and Olson [26] showed that neurons in inferior temporal cortex (IT) could be strongly modulated by prior experience with sequences of presented images. After repeated presentations of arbitrary images with predictable transition statistics (e.g. "image B always follows image A"), neurons appeared to learn the sequence statistics, responding robustly only to sequence transitions that were unexpected. The importance of temporal context in perception is further illustrated in various motion illusions, such as the flash-lag effect [29, 24, 7] and static motion illusions [50], where the motion of objects is incorrectly perceived by humans in predictable ways. Again, standard feedforward CNNs are insufficient to explain these temporal phenomena.

Here, inspired by past success in using "out-of-the-box" artificial deep neural networks as models of visual cortex, we explore whether modern predictive recurrent neural networks built for unsupervised learning can also explain dynamic phenomena in the brain. In particular, we consider a deep predictive coding network ("PredNet"; [23]), a network that learns to perform next-frame prediction in video sequences [32, 4, 8, 25, 2, 13, 48, 46, 47]. The PredNet is motivated by the principle of "predictive coding" [33, 9, 42, 5, 51]; the network continually generates predictions of future sensory data via a top-down path, and it sends prediction errors in its feedforward path (Fig. 1). At its lowest layer, the network predicts the input pixels at the next time-step, and it has been shown to make successful predictions in real-world settings (e.g. the KITTI car-mounted camera dataset [10]). The internal representations learned from video prediction also proved to be useful for subsequent decoding of underlying latent parameters of the video sequence, consistent with the suggestion of prediction as a good loss function for unsupervised learning [41, 31, 22, 49, 25, 44, 30, 6, 8].

Predictive coding has a rich history in neuroscience literature [34, 45, 3, 14, 43, 1]. Rao and Ballard helped popularize the notion of predictive coding in neuroscience in 1999, proposing that spatial predictive coding could explain a key property of neurons in primary visual cortex (V1) known as end-stopping ([33]; see Section 3). Predictive coding has also been proposed as an explanatory framework for a variety of sensory systems in neuroscience [19, 55, 27]. The PredNet formulates predictive coding principles in a deep learning framework to work on natural sequences, providing an opportunity to test a wide range of neuroscience phenomena using a single model. Below, we show that despite being trained only to predict next frames in video sequences, the PredNet naturally captures a wide array of seemingly unrelated fundamental properties of neuronal responses and perception, including on/off dynamics, length suppression, sequence learning effects in visual cortex, norm-based coding of faces, illusory contours, and the flash-lag illusion.

2 Deep Predictive Coding Networks

The deep predictive coding network proposed in [23] ("PredNet") consists of repeated, stacked modules where each module generates a prediction of its own feedforward inputs, computes errors between these predictions and the observed inputs, and then forwards these error signals to subsequent layers. The model consists of four components: targets to be predicted (A_l), predictions ($\hat{A}_l$), errors between predictions and targets (E_l), and a recurrent representation from which predictions are made (R_l). On an initial time step, the feedforward pass can be viewed as a standard CNN, consisting of alternating convolutional and pooling layers. Predictions are made in a top-down pass via convolutions over the representational units, which are first updated using the representational units from the layer above and errors from the previous time step as inputs. The R_l units are implemented as convolutional LSTMs [11, 38]. Here, for the sake of biological interpretability, we replace the $tanh$ output activation function for the LSTMs with a $relu$ activation, enforcing positive "firing rates". On the KITTI dataset this leads to a marginally (8%) worse prediction mean-squared error (MSE)

than the standard formulation, but it is still 2.6 times better than the MSE that would be obtained by simply copying the last frame seen (compared to 2.8 for $tanh$).

The error layers in the model, E_l, are calculated as a simple difference between the targets and predictions, followed by splitting into positive and negative error populations with $relu$ rectification. The loss function for the network is set as the (weighted) sum of the error activations across each layer. We utilize the L_{all} formulation presented in the original paper, which places a non-zero loss on the error unit activity in every level in the network. Except where stated otherwise, results presented here use a model trained on the KITTI car-mounted camera dataset [10]. The same model hyperparameters were used as presented in the paper (besides the $relu$ activation in the LSTM units). Particularly, the model consists of four layers. With 0-indexing used here, Layer 1 would be analogous to V1 in visual cortex.

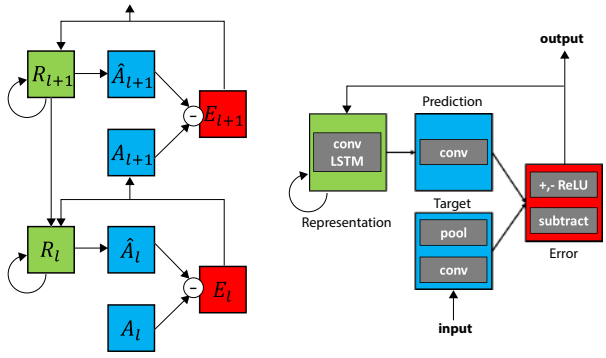

Figure 1: Deep Predictive Coding Networks (PredNets) [23]. Left: Each layer consists of representation neurons (R_l), which output a layer-specific prediction at each time step ($\hat{A}_l$), which is compared against a target (A_l) to produce an error term (E_l), which is then propagated laterally and vertically. Right: Module operations for case of video sequences. The target at the lowest layer of the network, $\hat{A}_0$, is set to the current input image.

3 Single Neuron Response Properties

We begin by comparing the response properties of units in the PredNet to established single unit response properties of neurons in the primate visual system, which have been studied extensively using microelectrode recordings. Here, we primarily compare response properties in the PredNet's error ("E") units, the output units of each layer, to neuronal recordings in the superficial layers of cortex. Response properties of other units in the PredNet (e.g. the "R" units) are included in the Supplemental Materials, and would likely map onto other parts of the cortical circuit.

On/Off Temporal Dynamics As mentioned in the introduction, a conspicuous feature of visual cortical neuron responses is that they are highly dynamic, even when a static, unchanging image is presented to the subject. As an example of the commonly seen pattern of image on/off dynamics, Fig. 2a shows a raster plot and peri-stimulus-time histogram of a recorded neuron in the secondary visual cortex (V2) of a macaque monkey [37]. Peaks in firing rate shortly after image display and removal are prominent. Fig. 2b shows the average response of PredNet E units in different layers over a set of 25 naturalistic objects appearing on a gray background. The on/off dynamics are apparent on the population average level, for all four layers of the network. These dynamics are also evident at the individual unit level, as illustrated in the Supplement, though there is variability. While on/off dynamics have an an "error"-like quality—an object unpredictably appears and disappears—these dynamics manifest in the A and R layers as well (see Supplement).

End-Stopping and Length Suppression Prediction in time and prediction in space are inextricably intertwined. As Rao and Ballard [33] illustrated, end-stopping in V1 can be explained by spatial predictive coding. End-stopping, or length suppression, is the phenomenon where a neuron tuned for a particular orientation becomes less responsive to a bar at this orientation, when the bar extends beyond its classical receptive field [12]. The predictive coding explanation is that lines/edges tend

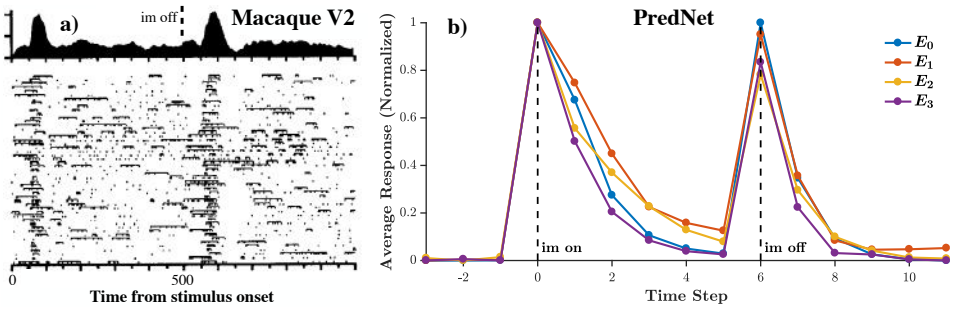

Figure 2: On/off temporal dynamics. Left: Exemplar macaque V2 neuron responding to a static image. Reproduced with permission from Schmolesky et al. [37]. Right: PredNet response to a set of naturalistic objects appearing on a gray background, after training on KITTI. Responses are grouped by layer for the E units, and averaged across all units and all stimuli, per layer.

to be continuous in nature, and thus the center of a long bar can be predicted from its flanks. A short, discontinuous bar, however, deviates from natural statistics, and responding neurons signal the deviation. One potential source for conveying the long range predictions in the case of an extended bar could be feedback from higher visual areas with larger receptive fields. This hypothesis was elegantly tested in Nassi et al. [28] using reversible inactivation of V2 paired with V1 recordings in the macaque. As illustrated in the left side of Fig. 3, cryoloop cooling of V2 led to a significant reduction in length suppression, indicating that feedback from V2 to V1 is essential for the effect.

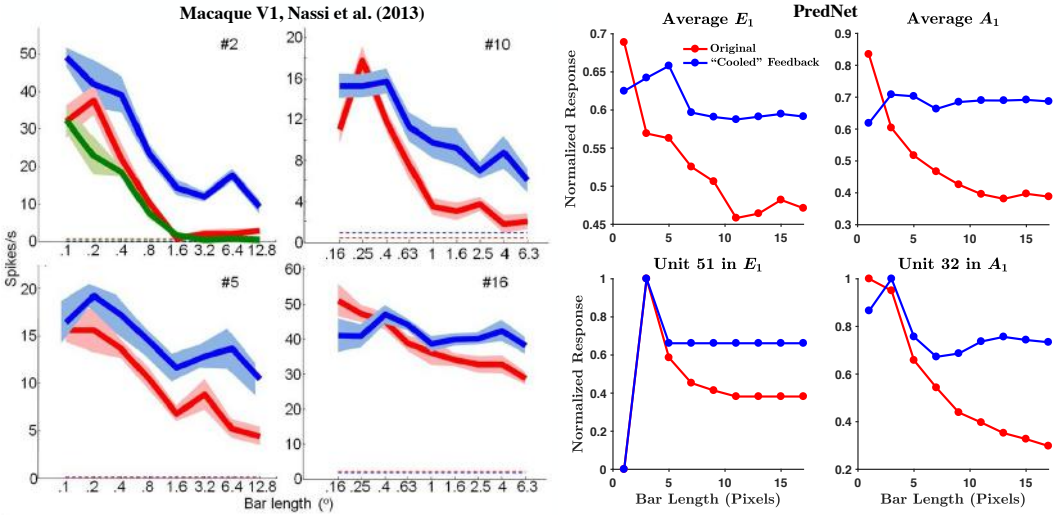

Figure 3: Length suppression. Left: Responses of example macaque V1 units to bars of different lengths before (red), during (blue), and after (green) inactivation of V2 via cryoloop cooling. Reproduced with permission from [28]. Right: PredNet after training on the KITTI dataset – average E_1 and A_1, and examples. Red: Original network. Blue: Feedback weights from R_2 to R_1 set to zero.

The right side of Fig. 3 demonstrates that length suppression, and its mediation through top-down feedback, are also present in the PredNet. The upper left and right panels contain the mean normalized response for units in the E_1 and A_1 layers, respectively, to bars of different lengths. The red curves correspond to the original network (trained on KITTI) and the blue curves correspond to zero-ing the feedback from R_2 to R_1. For each filter channel, a set of 2D Gabor wavelet stimuli was first used to determine the optimal orientation. Responses to bars of different length at this orientation were then measured, as a sum of the activations over stimulus duration (10 time steps). The bottom row contains exemplar E_1 and A_1 units. Quantifying percent length suppression (%LS) as $100 * \frac{R_{max} - R_{longest\ bar}}{R_{max}}$, the median decrease in %LS upon removing top-down signaling was 16% for E_1 units ($p < 0.05$, Wilcoxon signed rank test) and 33% for A_1 units ($p < 0.0005$). For R_1 units, the median %LS decrease was 2% ($p = 0.18$). Indeed, the average R_1 response did not exhibit much length suppression (see Supplement), though, there were particular examples with a strong effect.

Sequence Learning Effects in Visual Cortex Predictions are often informed by recent experience, and violations of these predictions can be highly salient. Meyer and Olson [26] provided a striking example of this in visual cortex via image sequence learning. The authors exposed monkeys to image pairs in a fixed order for over 800 trials for each pair. The left panel of Fig. 4 shows the mean response of 81 IT neurons in a subsequent testing period, for predicted and unpredicted pairs. When the second image differs from expectations, the response is much stronger than when the expected image is presented.

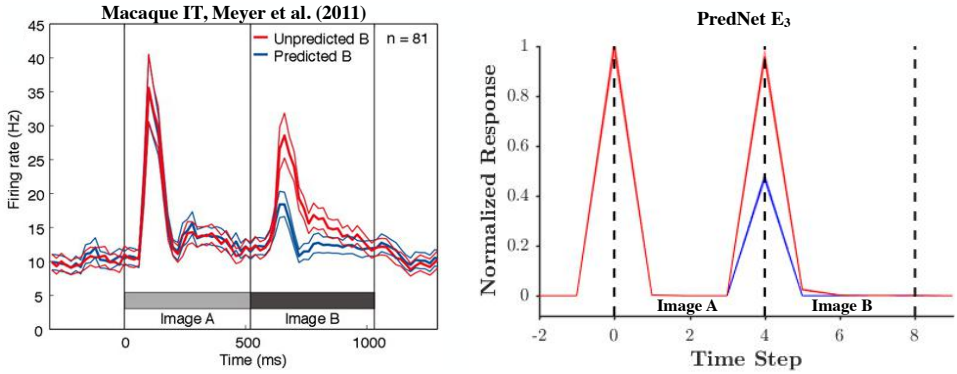

Figure 4: Predicted vs. unpredicted image transitions. Left: Mean of 81 neurons recorded in macaque (IT). Reproduced with permission from [26]. Right: Mean ($\pm$ SE) across PredNet E_3 units.

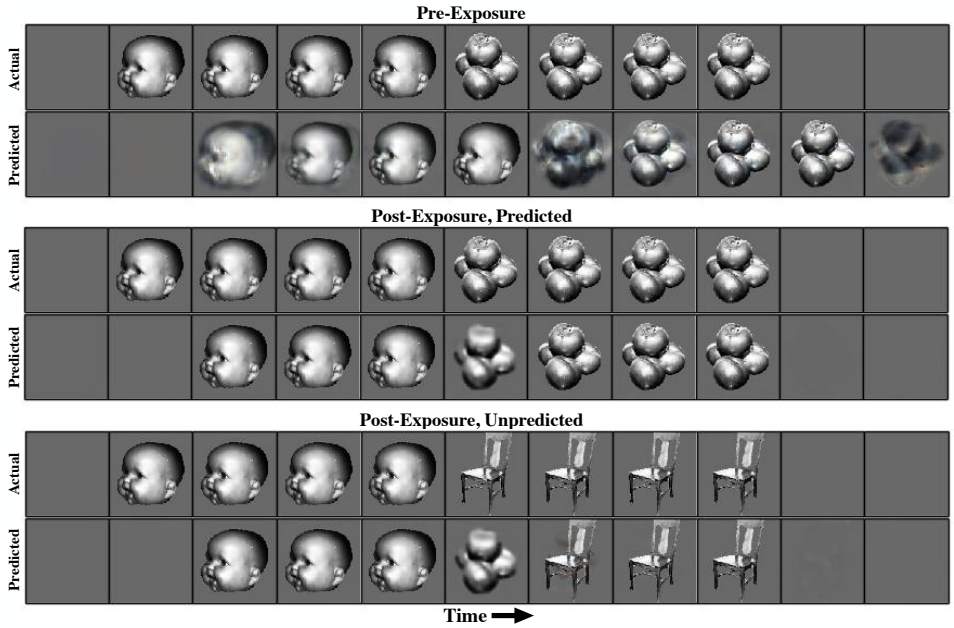

Figure 5: Learned image transitions. Top: Predictions of a KITTI-trained PredNet model on an example sequence. Middle: PredNet predictions after repeated "training" on the sequence. Bottom: PredNet predictions for an unpredicted image transition.

The right panel of Fig. 4 demonstrates a similar effect in the PredNet after an analogous experiment. The model was trained on five image pairs for 800 epochs. Fig. 5 contains an example sequence and the corresponding next-frame predictions before and after the training. The model, prior to exposure to the images in this experiment (trained only on KITTI), settles into a noisy, copy-last-frame prediction mode. After exposure, the model is able to successfully make predictions for the expected image pair (row 2). Since the chosen image pair is unknown *a priori* and the model is fully convolutional, the initial prediction is the constant gray background when the first image appears. The model then rapidly copies this image for the ensuing three frames. Next, the model successfully predicts the transition to the second object (a stack of tomatoes in this case). In row 3, a sequence that differs from the training pair is presented. The model still makes the prediction of a transition

to tomatoes, even though a chair is presented, but then copies the chair into subsequent predictions. Fig. 4 shows that the unexpected transitions result in a larger average response in the final E layer of the network. In fact, in all levels and all unit types (E, A, R), there is a larger response to the unpredicted images (Supp. Table 1). The overall magnitude of the difference is similar for A and E, and is lower for R.

Norm-Based Coding of Faces The representation of deviations from expectations can also explain observed neural embeddings of familiar stimuli such as faces. The norm-based coding theory suggests that faces are encoded with respect to a mean face [35]. Leopold et al. [21] demonstrated that face-responsive neurons in macaque anterior IT are frequently tuned monotonically (often positively) to directions away from an average face (Fig. 6). This was tested by using synthetic faces with continuously varying levels of caricature. Training for next-frame prediction of rotating, computer-generated faces [23], we see similar effects in the PredNet. The faces were created using software implementing a principal component analysis of a corpus of human faces (FaceGen [40]). For training, $16K$ sequences were generated with a random face, initial orientation, and rotation velocity. The blue curve in Fig. 6b illustrates the post-training response of E units in the network as a function of caricaturization level for 200 previously unseen faces. The response is calculated by first averaging the response of all units in a given layer, then averaging the layer responses. The PredNet E units become significantly more responsive to increasing levels of caricature after predictive training compared to the random initialization (red curve). This effect is diminished when the same network is trained on the same set of images, but in a static, autoencoder fashion (yellow curve). Note that the randomly initialized network already responds more highly to caricature, likely because the caricature faces tend to have higher contrast, sharper edges, etc., which even random CNNs can be tuned for [36]. When training on an unrelated dataset (e.g. the KITTI dataset), this effect is reduced (purple curve). Training the network on rotating faces that had been generated using half the standard deviation for each principal component results in an even larger caricature response (green curve). All of these effects are consistent in the A units as well, though the results are mixed for R (Supplement).

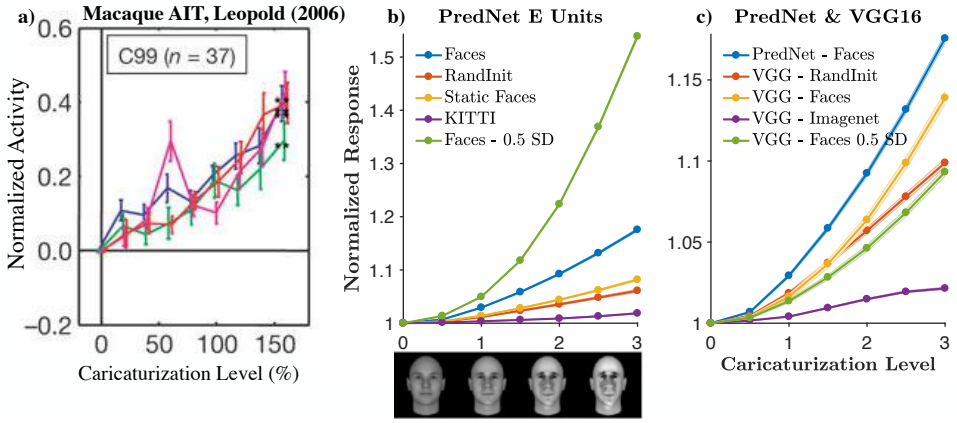

Figure 6: Norm-based coding of faces. a): Population response of 37 neurons recorded in macaque IT for four different faces, as a function of caricaturization. Reproduced with permission from [21]. b): Mean PredNet E unit responses for models trained on different stimuli. Faces - Rotating synthetic faces [23]. RandInit - Random initial weights. Static Faces - Same collection of images used in Rotating Faces except presented statically. Faces 0.5 SD - Rotating faces except each face generated from a distribution with half of the original standard deviation. c): Comparing PredNet with VGG16. VGG Faces - VGG trained in a siamese manner on a same/different task using the static face images.

Fig. 6c compares the PredNet responses to a popular CNN, VGG16 [39]. VGG responses were quantified by averaging over the outputs of its five convolutional blocks (after the max-pooling). Similar to the PredNet, the randomly initialized VGG16 displayed higher activity for more caricatured faces. ImageNet training decreased this effect, akin to the PredNet KITTI training. To train VGG on the synthetic faces, a same/different task was performed using the network in a siamese fashion. Using the same images as the PredNet, a training example consisted of a pair of images at different orientations with a binary cross-entropy objective for same/different identity classification. This training procedure resulted in an increased response to higher caricature levels (at least on the original dataset – yellow curve in Fig. 6c), although somewhat less than the PredNet E units. While analogous

norm-based face encoding effects can be seen with discriminatively-trained (VGG) models, the PredNet architecture is able to capture these same effects (even more strongly) in a fully unsupervised way.

4 Visual Illusions

Visual illusions can provide powerful insight into the underpinnings of perception. Here we demonstrate that the PredNet exhibits correlates of two illusions: illusory contours and the flash-lag effect, both of which have aspects of spatial and temporal prediction. The PredNet has also recently been shown to predict the illusory motion perceived in the rotating snakes illusion [50].

Illusory Contours Illusory contours, as in the Kanizsa figures [15], elicit perceptions of edges and shapes, despite the lack of enclosing lines. Lee et al. [20] found that neurons in monkey V1 can be responsive to illusory contours, albeit at a reduced response and increased latency to physical contours. Fig. 7a contains an example of such a neuron. The stimuli in the experiment consisted of sequences starting with an image of four circles, which then abruptly transitioned to one of numerous test images, including the illusion. Illustrated in Fig. 7b, the population average of 49 superficial V1 neurons responded more strongly to the illusion than similar, but non-illusory stimuli. This preference was also apparent in V2, with a response that was, interestingly, of a shorter latency compared to V1.

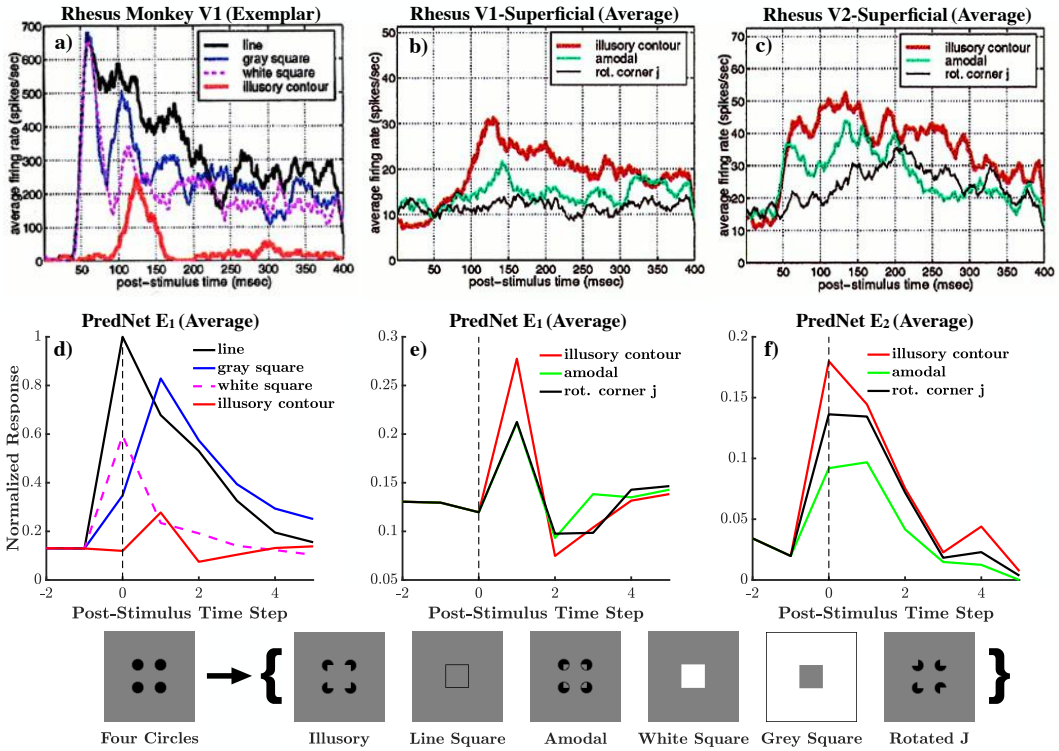

Figure 7: Illusory contours. Top: Reproduced with permission from Lee et al. [20]. A given trial consisted of the four circles image abruptly transitioning to one of the displayed test images.

Fig. 7c-e demonstrate that the effects discovered by [20] are also present in the PredNet. In the population average of E_1 units, there is indeed a response to the illusory contour, and at an increased latency compared to a physical line square (Fig. 7c). The response was calculated separately for each filter channel, by first finding the optimal orientation using short bar segments. The responses were then normalized (division by max response over all stimuli) and averaged. Fig. 7d illustrates that the average E_1 response was higher for the illusory contour than the similar control images. This was also the case for E_2 units, with a peak response one time step before E_1. Indeed, the size of the stimuli was such that it was larger than the feedforward receptive field of the layer 1 neurons, but smaller than that of the layer 2 neurons (matching the protocol of [20]). Quantifying the preference

of the illusion to the amodal and rotated "J" images for each individual unit [20]), we find that the average is positive (more preference to the illusion) for all tested layers (E, A, R, layers 1,2).

The Flash-Lag Effect Another illusion for which prediction has been proposed as having a role is the flash-lag effect. Fundamentally, the flash-lag effect describes illusions where an unpredictably appearing stimulus (e.g. a line or dot) is perceived as "lagging" a predictably moving stimulus nearby, even when the stimuli are, in fact, precisely aligned in space. These illusions are sometimes interpreted as evidence that the brain is performing inference to predict the likely true current position of a stimulus, even in spite of substantial latency (up to hundreds of milliseconds) in the visual system [17]. The version of the illusion tested here consists of an inner, continuously rotating bar and an outer bar that periodically flashes on. Fig. 8 contains an example prediction by the PredNet on a sample sequence within a flash-lag stimulus. The rotation speed of the inner bar in the clip was set to 6 degrees per time step. The first feature of note is that the PredNet is indeed able to make reasonable predictions for the inner rotating bar. Quantifying this, the average angle of the bar in the outputted predictions is 1.4±1.2°(s.d.) behind the actual bar (see Supp. Methods), which is significantly less than a 6°difference, which would result from simply copying the last seen frame. Again, the model was trained on real-world videos, so the generalization to this impoverished stimulus is non-trivial. Secondly, the post-flash predictions made by the model tend to resemble the perceived illusion. The average angular difference between the predicted outer bar and inner bar is 6.8±2.0°. Considering that the model was trained for next frame prediction on a corpus of natural videos, this suggests that our percept matches the statistically predicted next frame (as estimated by the PredNet) more than the actual next frame. This natural statistics interpretation of the flash-lag illusion has, in fact, been similarly suggested by Wojtach et al. [52].

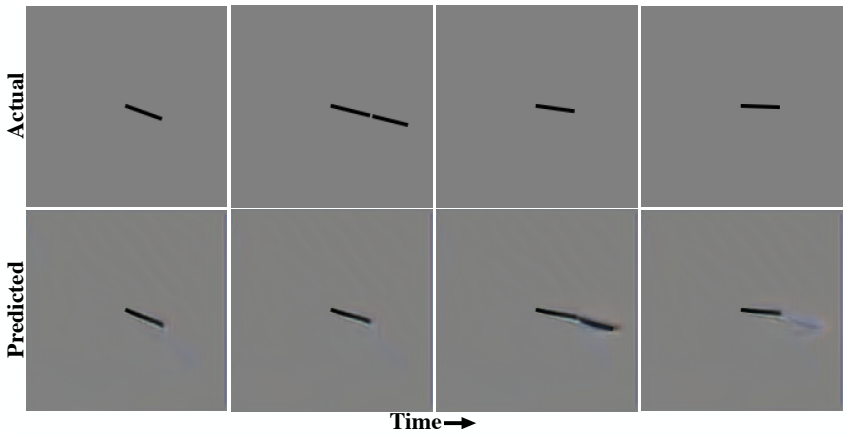

Figure 8: Flash-lag effect. Top: Stimulus clip. Bottom: PredNet predictions after KITTI training.

5 Discussion

We have shown that an off-the-shelf recurrent neural network trained to predict future video frames can explain a wide variety of seemingly unrelated phenomena observed in visual cortex and visual perception. These phenomena range from the details of responses of individual neurons, to complex visual illusions. Importantly, throughout, we used a base model trained on natural videos. Our work adds to a growing body of literature showing that deep neural networks trained to perform relevant tasks can serve as surprisingly good models of biological neural networks, often even outperforming models designed to explain neuroscience phenomena.

While we have shown that the PredNet architecture demonstrates a wide range of phenomena reminiscent of biology, we do not claim that the PredNet architecture *per se* is required to explain these phenomena. Rather, we argue that the network is *sufficient* to produce these phenomena, and we note that explicit representation of prediction errors in units within the feedforward path of the PredNet provides a straightforward explanation for the transient nature of responses in visual cortex in response to static images. That a single, simple objective—prediction—can produce such a wide variety of observed neural phenomena underscores the idea that prediction may be a central organizing principle in the brain [33], and points toward fruitful directions for future study in both neuroscience and machine learning.

Acknowledgments

This work was supported by IARPA (contract D16PC00002), the National Science Foundation (NSF IIS 1409097), and the Center for Brains, Minds and Machines (CBMM, NSF STC award CCF-1231216).

References

[1] J. J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 1992.
[2] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. *ICLR*, 2018.
[3] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston. Canonical microcircuits for predictive coding. *Neuron*, 2012.
[4] B. D. Brabandere, X. Jia, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. *NIPS*, 2016.
[5] R. Chalasani and J. C. Principe. Deep predictive coding networks. *arXiv*, 2013.
[6] A. Dosovitskiy and V. Koltun. Learning to act by predicting the future. *ICLR*, 2017.
[7] D. M. Eagleman and T. J. Sejnowski. Motion integration and postdiction in visual awareness. *Science*, 2000.
[8] C. Finn, I. J. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. *NIPS*, 2016.
[9] K. Friston. A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci*, 2005.
[10] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.
[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.
[12] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 1968.
[13] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. *arXiv*, 2016.
[14] R. Kanai, Y. Komura, S. Shipp, and K. Friston. Cerebral hierarchies : predictive processing , precision and the pulvinar. *Philos Trans R Soc Lond B Biol Sci*, 2015.
[15] G. Kaniza. *Organization in Vision: Essays on Gestalt Perception*. Praeger, 1979.
[16] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 2014.
[17] M. A. Khoei, G. S. Masson, and L. U. Perrinet. The flash-lag effect as a motion-based predictive shift. *PLOS Computational Biology*, 2017.
[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, 2014.
[19] S. Kumar, W. Sedley, K. V. Nourski, H. Kawasaki, H. Oya, R. D. Patterson, M. A. H. III, K. J. Friston, and T. D. Griffiths. Predictive coding and pitch processing in the auditory cortex. *Journal of Cognitive Neuroscience*, 2011.
[20] T. S. Lee and M. Nguyen. Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences*, 2001.
[21] D. A. Leopold, I. V. Bondar, and M. A. Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 2006.
[22] W. Lotter, G. Kreiman, and D. Cox. Unsupervised learning of visual structure using predictive generative networks. *ICLR (Workshop Track)*, 2016.
[23] W. Lotter, G. Kreiman, and D. D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *ICLR*, 2017.
[24] D. M. Mackay. Perceptual stability of a stroboscopically lit visual field containing self-luminous objects. *Nature*, 1958.
[25] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *ICLR*, 2016.
[26] T. Meyer and C. R. Olson. Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences*, 2011.
[27] D. Mumford. On the computational architecture of the neocortex. *Biological Cybernetics*, 1992.
[28] J. J. Nassi, S. G. Lomber, and R. T. Born. Corticocortical feedback contributes to surround suppression in v1 of the alert primate. *Journal of Neuroscience*, 2013.
[29] R. Nijhawan. Motion extrapolation in catching. *Nature*, 1994.
[30] R. C. O'Reilly, D. Wyatte, and J. Rohrlich. Learning through time in the thalamocortical loops. *arXiv*, 2014.
[31] R. B. Palm. Prediction as a candidate for learning deep hierarchical models of data. *Master's thesis, Technical University of Denmark*, 2012.

[32] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv*, 2014.
[33] R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 1999.
[34] R. P. N. Rao and T. J. Sejnowski. Predictive sequence learning in recurrent neocortical circuits. *NIPS*, 2000.
[35] G. Rhodes and L. Jeffery. Adaptive norm-based coding of facial identity. *Vision Research*, 2006.
[36] A. Saxe, M. Bhand, Z. Chen, P. W. Koh, B. Suresh, and A. Y. Ng. On random weights and unsupervised feature learning. *Workshop: Deep Learning and Unsupervised Feature Learning (NIPS)*, 2010.
[37] M. T. Schmolesky, Y. Wang, D. P. Hanes, K. G. Thompson, S. Leutgeb, J. D. Schall, and A. G. Leventhal. Signal timing across the macaque visual system. *Journal of Neurophysiology*, 1998.
[38] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *NIPS*, 2015.
[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.
[40] Singular Inversions, Inc. FaceGen. `http://facegen.com`.
[41] W. R. Softky. Unsupervised pixel-prediction. *NIPS*, 1996.
[42] M. W. Spratling. Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Computation*, 2012.
[43] M. V. Srinivasan, S. B. Laughlin, and A. Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London B: Biological Sciences*, 1982.
[44] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *arXiv*, 2015.
[45] C. Summerfield, T. Egner, M. Greene, E. Koechlin, J. Mangels, and J. Hirsch. Predictive codes for forthcoming perception in the frontal cortex. *Science*, 314, 2006.
[46] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. *ICLR*, 2017.
[47] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. *ICML*, 2017.
[48] C. Vondrick and A. Torralba. Generating the future with adversarial transformers. *CVPR*, 2017.
[49] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. *arXiv*, 2015.
[50] E. Watanabe, A. Kitaoka, K. Sakamoto, M. Yasugi, and K. Tanaka. Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology*, 2018.
[51] H. Wen, K. Han, J. Shi, Y. Zhang, E. Culurciello, and Z. Liu. Deep predictive coding network for object recognition. *arXiv*, 2018.
[52] W. T. Wojtach, K. Sung, S. Truong, and D. Purves. An empirical explanation of the flash-lag effect. *Proceedings of the National Academy of Sciences*, 2008.
[53] D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 2016.
[54] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 2014.
[55] C. Zelano, A. Mohanty, and J. A. Gottfried. Olfactory predictive codes and stimulus templates in piriform cortex. *Neuron*, 2011.

6 Supplementary Material

6.1 On/Off Temporal Dynamics

Temporal dynamics were tested with a set of 25 objects. Examples of the objects can be seen in Fig. 5. Testing sequences consisted of a gray background for 7 time steps, followed by an object on the background for 6 time steps. As a general theme, we see some diversity in the response profiles of all units, but especially those in the "R" layers. We have focused in the main text on the "E" units, which most naturally map onto Layer 2/3 cortical pyramidal neurons (which are the output units in a putative cortical microcircuit). Diversity of responses is also observed throughout the neuroscience literature, and we hypothesize that to the extent that units in the PredNet map in a direct way onto cortical circuits [3], the less-often experimentally sampled deep neurons might be reasonable analogs to the "R" units. We present representative units from all parts of the PredNet here for completeness.

Summary responses for the A and R units are contained in Fig. 9. The responses are grouped per layer and consist of an average across all the units (all filters and spatial locations) in a layer. The mean responses were then normalized between 0 and 1. Responses for layer 0, the pixel layer, are omitted in Fig. 9 because of their heavy dependence on the input pixels for the A and R layers. Note that, by notation in the network's update rules, the input image reaches the R layers at a time step after the E and A layers.

As illustrated in Fig. 9, the A and R layers seem to generally exhibit on/off dynamics, similar to the E layers. R_1 also seems to have another mode in its response, specifically a ramp up between time steps 3 and 5 post image onset. As will be illustrated below, this results from a few strongly firing neurons that exhibit this pattern.

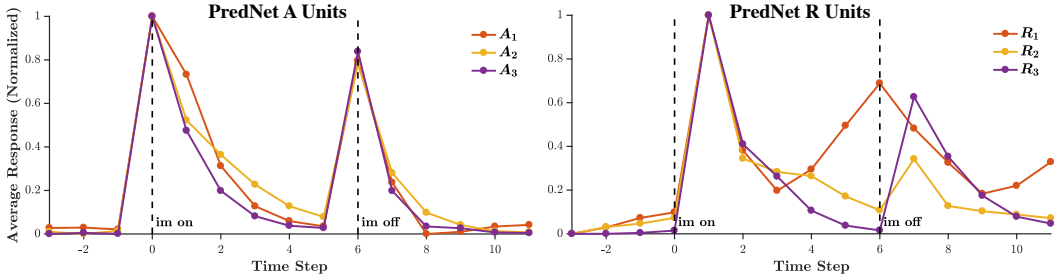

Figure 9: Average temporal dynamics in A and R units in response to set of naturalistic objects on a gray background, after training on the KITTI dataset.

Fig. 10-12 illustrate the variety of responses present among the units in the model. Each plot for a given layer shows all the active units in the layer at the central receptive field. The average response for each unit over the 25 images is shown, and each row is normalized to have a max of one. In each of the plots, it is apparent that a large proportion of the neurons have a peak response closely following image onset and/or offset. However, there are a good number of neurons that have peaks at different times. Overall, the on and off responses for individual neurons tend to be asymmetric – some neurons have stronger "on" responses, some have stronger "off" responses.

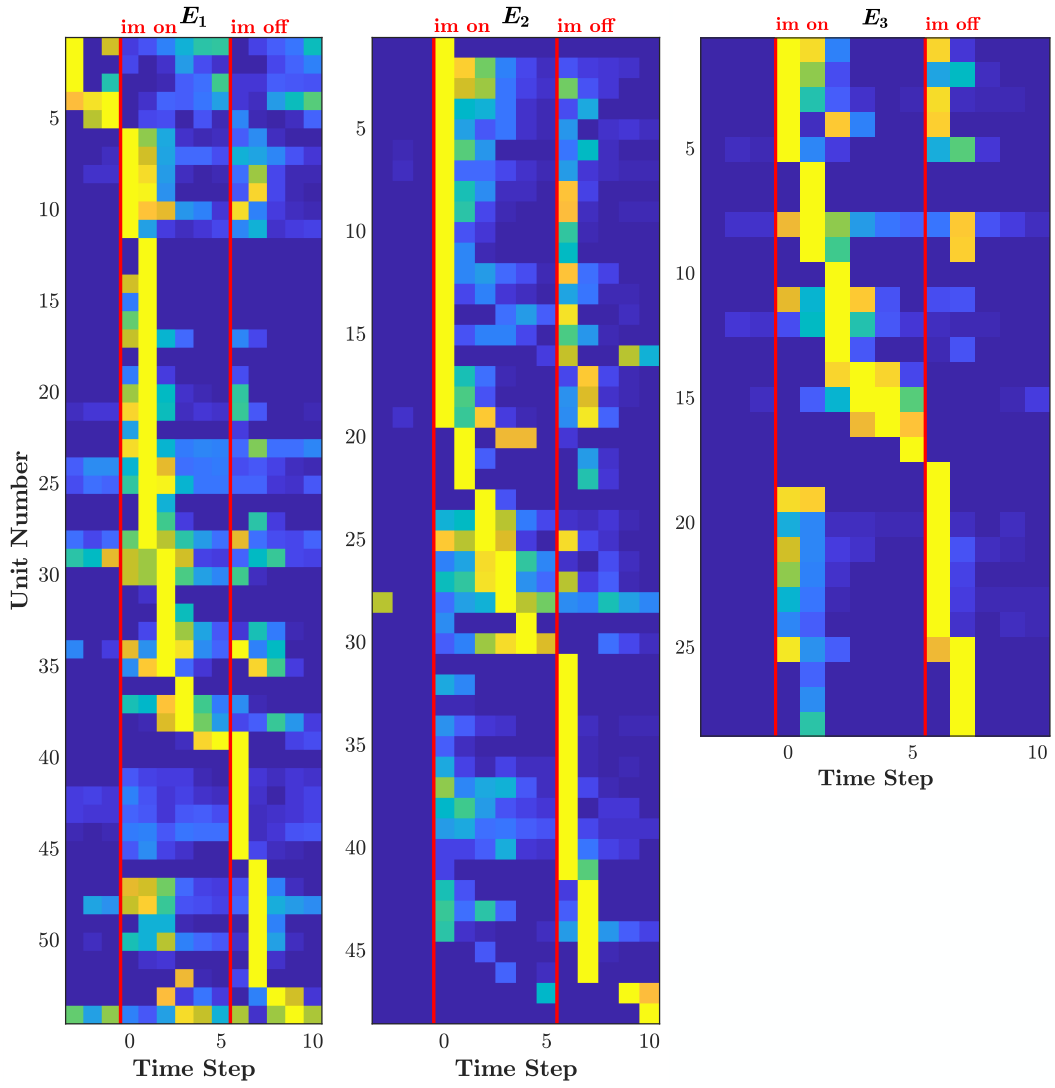

Figure 10: Mean response of each active unit in the E layer at the central receptive field. Each row (unit) is normalized by its max response.

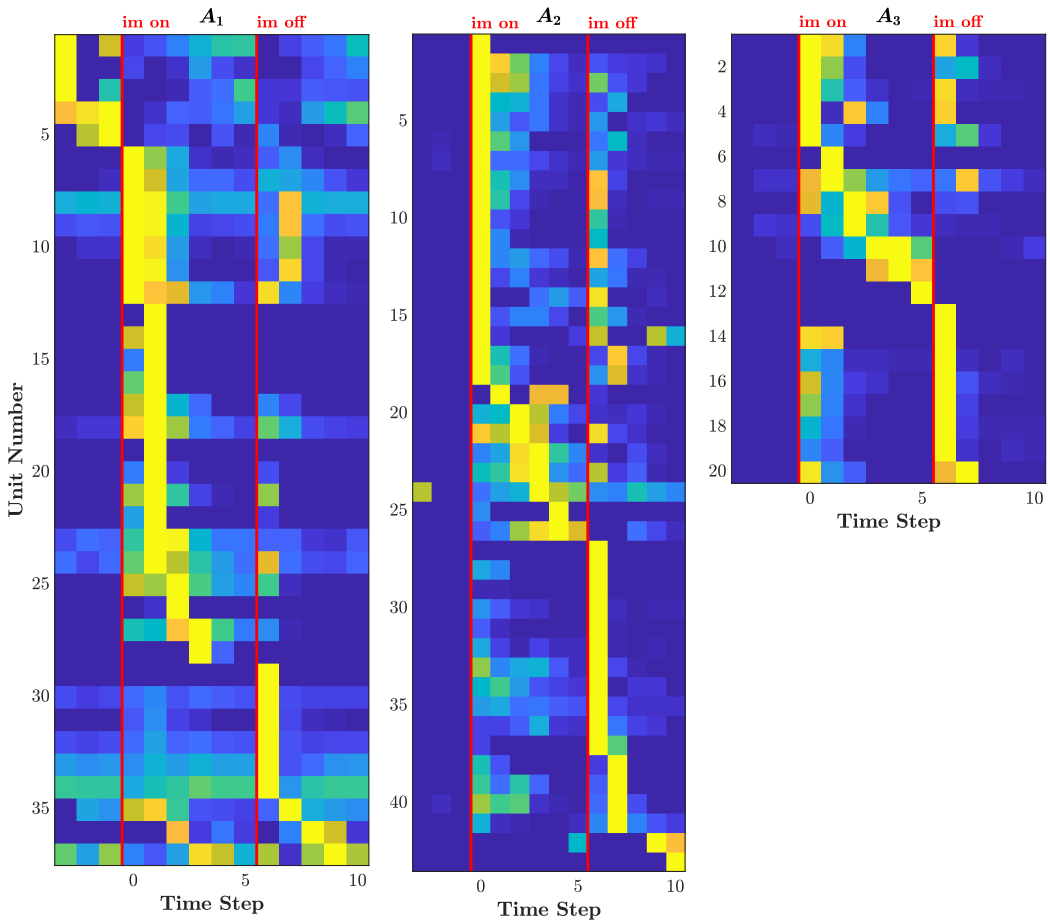

Figure 11: Mean response of each active unit in the A layer at the central receptive field. Each row (unit) is normalized by its max response.

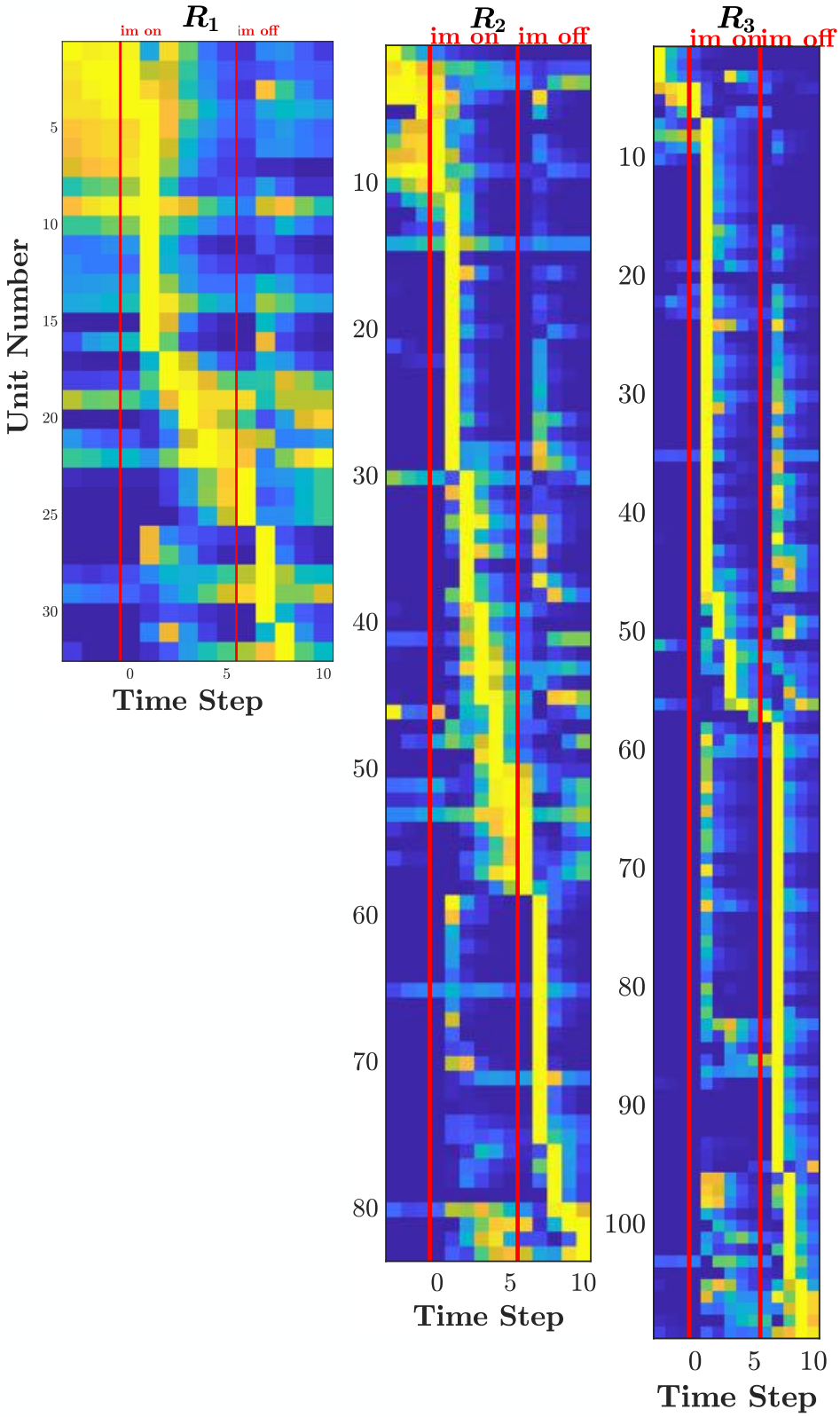

Figure 12: Mean response of each active unit in the R layer at the central receptive field. Each row (unit) is normalized by its max response.

Figure 13 is similar to the previous plots, except a global normalization is used instead of row normalization. There are subsets of neurons that are particularly more active than others. For R_1, rows 23-25 contain examples of units that contribute to the ramping behavior from time-steps 3 to 5 in Fig. 9.

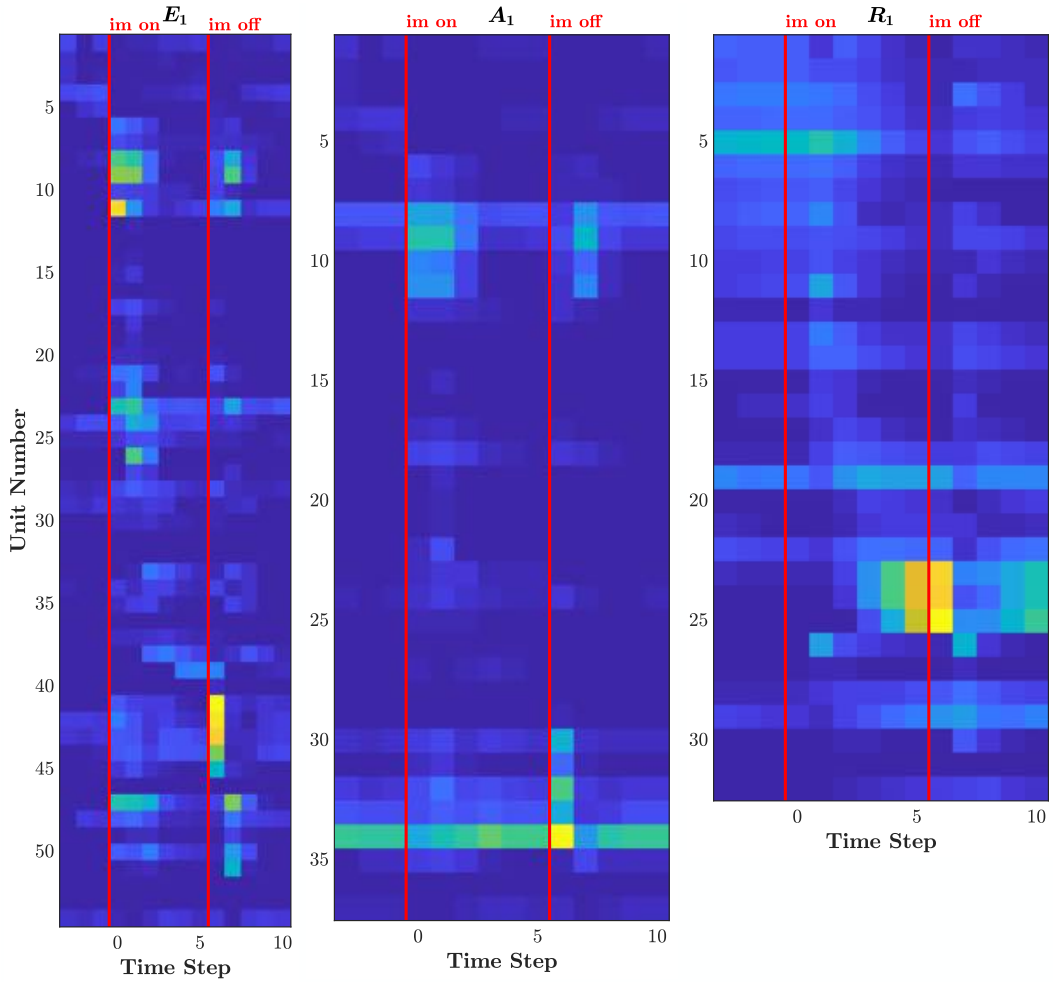

Figure 13: Mean response of each active unit with a central receptive field for all three unit types at Layer 1. Responses are normalized globally per each unit type.

6.2 End-Stopping and Length Suppression

Responses to bars of different length were quantified for the length suppression experiment as a sum over the stimulus duration of 10 time steps. The bars appeared on a gray background, which was first presented to the network for 5 time steps, to allow the network to settle to a steady state before stimulus presentation. For each filter channel, the response at the central receptive field was quantified and normalized to unit maximum before averaging. The average R_1 response and two exemplar units are displayed in Fig. 14. As mentioned in the main text, R_1 did not have a significant length suppression effect, with some neurons showing length suppression (right panel) and others showing an opposite effect (middle panel).

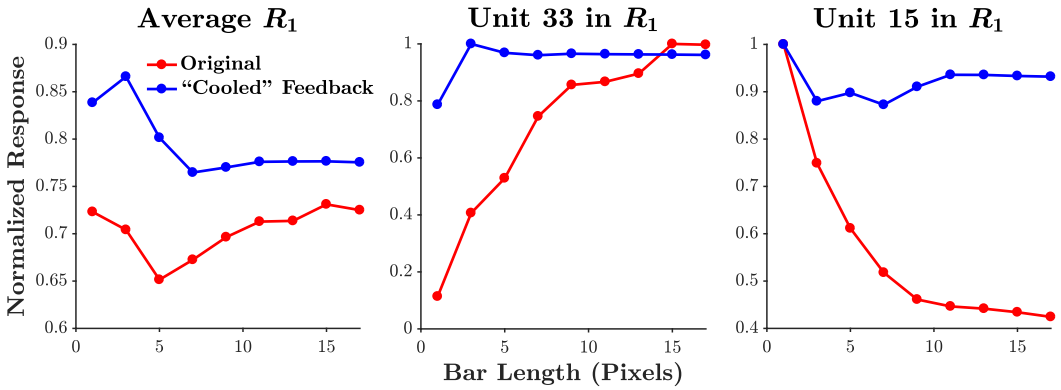

Figure 14: Length suppression analysis for R_1 units.

6.3 Sequence Learning Effects in Visual Cortex

For the exposure training phase in the learned sequence experiment, the Adam [18] optimizer was used with default parameters. Table 1 contains the percent increase in response between predicted and unpredicted sequences for each layer.

Table 1: Percent increase of response between predicted and unpredicted sequences

Unit Type	Layer 0	Layer 1	Layer 2	Layer 3
E	308	90	109	108
A	N/A	78	109	108
R	N/A	18	19	30

6.4 Norm-Based Coding of Faces

For the faces generated for the norm-based coding experiment, a caricature level of, say 2, corresponds to having all principal components with a of magnitude 2 (either positive or negative). The hyperparameters of the tested PredNet model were chosen to match those of the rotating faces model in the original paper [23]. Fig. 15 shows the responses of the A and R units to the caricature faces. Responses are calculated as an average per each layer, and then averaged across layers. Training on rotating faces led to a much higher caricature response in the A units, especially for training on faces generated with half of the original principal component standard deviation. The lower standard deviation had a similar effect in the R units, although training on the original rotating faces actually led to a smaller caricature response than the initial weights.

The siamese VGG network used for the same/diff face identification task was constructed by taking the squared element-wise difference between the flattened features at the last convolutional layer for the two inputs, followed by a fully-connected, softmax classification layer.

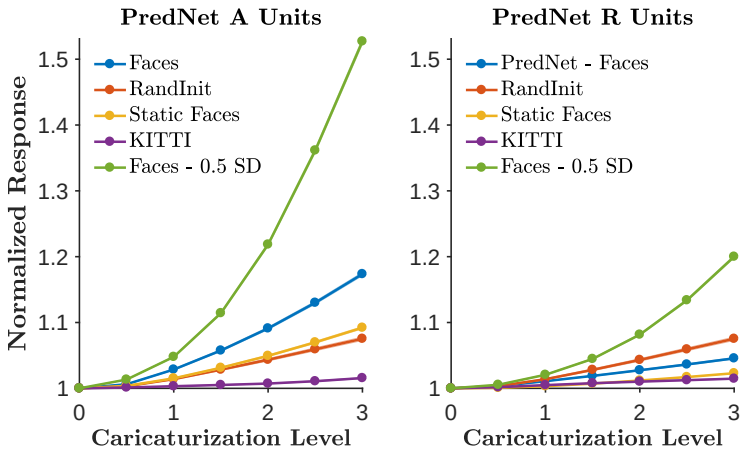

Figure 15: Responses of PredNet A and R units to varying levels of caricaturized faces, trained in different settings. Faces - Rotating synthetic faces. RandInit - Random initial weights. Static Faces - Same collection of images used in Rotating Faces except presented statically. Faces 0.5 SD - Rotating faces except each face generated from a distribution with half of the original standard deviation.

6.5 Illusory Contours

Fig. 16 contains the illusory contour response plots for the A and R layers. The stimuli sequences consisted of 10 time steps of the "four circles" image (see main text) followed by a test image for 10 time steps. The response to the illusory stimuli begins one time step after the response to the line square for all unit types in the first layer.

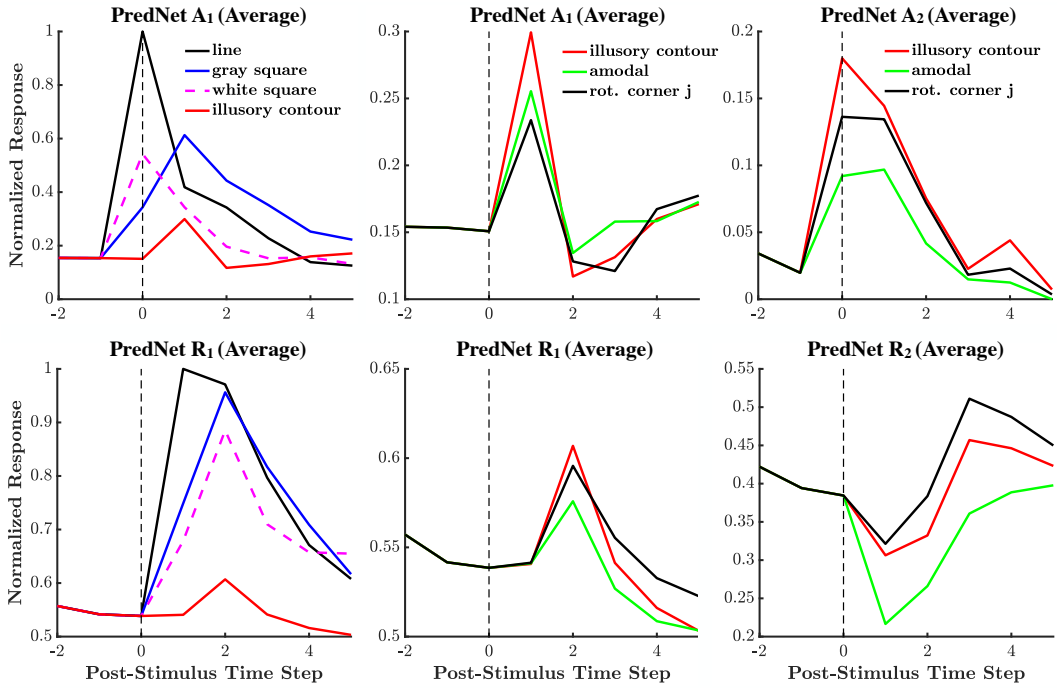

Figure 16: Illusory contours responses for A and R units.

To quantify illusory responsiveness, we follow Lee et al. [20] in calculating the following two measures: $IC_a = \frac{R_i - R_a}{R_i + R_a}$ and $IC_r = \frac{R_i - R_r}{R_i + R_r}$, where R_i is the response to the illusory contour (sum over stimulus duration), R_a is the response to amodal stimuli, and R_r is the response to the rotated J

Table 2: Illusory responsiveness measures for the units in Lee et al. [20] and the PredNet. IC_a and IC_r compare the response of the illusion to the amodal and rotated stimuli, respectively. Positive measures indicate a preference to the illusion. $*p < 0.05$ (T-test)

Source	Layer	IC_A	IC_R	Layer	IC_A	IC_R	Layer	IC_A	IC_R	Layer	IC_A	IC_R
Monkey A	V1S	0.19	0.31	V2S	0.21	0.11	V1D	0.09	0.11	V2D	0.28	0.24
Monkey B	V1S	0.10	0.16	V2S	0.08	0.12	V1D	0.04	0.13	V2D	0.07	0.20
PredNet	E_1	0.09	0.14*	E_2	0.15*	0.09	R_1	0.11*	0.04	R_2	0.12*	0.03
PredNet	A_1	0.03	0.10*	A_2	0.15*	0.09						

image. These indices were calculated separately for each unit with a non-uniform response. For both measures and all examined layers, the average across the layer was positive (significant in half of the calculations (Table 2)).

6.6 Flash-Lag Effect

The flash-lag stimulus was created with a rotation speed of $6°$per time step, with a flash every 6 time steps for 3 full rotations. Angles of the predictions were quantified over the last two rotations. The angles of the predicted bars were estimated by calculating the mean-squared error (MSE) between the prediction and a probe bar generated at $0.1°$increments and a range of centers, and taking the angle with the minimum mean-squared error. Fig. 17 contains additional predictions by the model after four consecutive flashes.

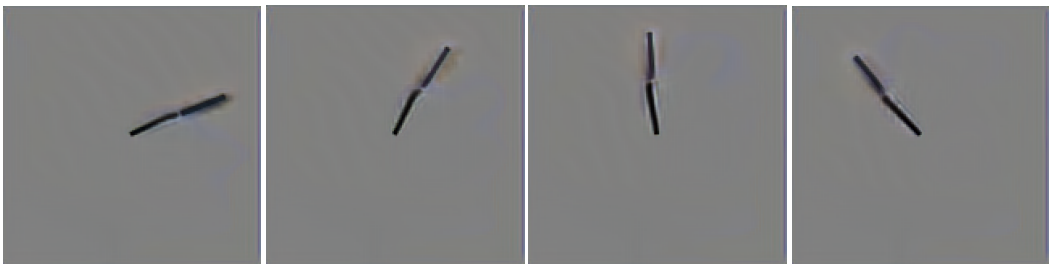

Figure 17: Four consecutive post-flash predictions by the PredNet model following training on the KITTI dataset.