THE EFFECTS OF NEURAL RESOURCE CONSTRAINTS ON EARLY VISUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The visual system of vertebrates is organized in a hierarchical structure that processes visual information in successive stages. Neural representations vary drastically across the first stages of visual processing. At the output of the retina, receptive fields (RFs) exhibit a clear antagonistic center-surround structure, whereas in the primary visual cortex (V1), typical RFs are sharply tuned to a precise orientation. Moreover, there are striking differences in early visual representations across species. Specifically, the retinas of small vertebrates (e.g. salamander, mouse) perform sophisticated non-linear computations extracting features directly relevant to behavior. In contrast, the dominant retinal cell types in primates encode the visual scene quasi-linearly and respond to a much broader range of stimuli. There is currently no theory explaining these differences in representation across layers and across species. Here, using a deep convolutional neural network trained on image recognition as a model of the visual system, we show that such differences in representation emerge as a direct consequence of different neural resource constraints on the retinal and brain networks. Whereas previous studies of efficient coding invoked inconsistent sets of metabolic constraints to account for the differences in RF shapes found in the retina and V1, we find a unique model from which both geometries spontaneously emerge at the appropriate stages of visual processing. The key constraint of this model is a reduced number of neurons at the retinal output, a constraint consistent with the anatomy of the optic nerve. Furthermore, we find that retinal neurons in our model emerge as non-linear and lossy feature detectors for small brain networks, whereas they emerge as linear and faithful encoders of the visual scene for large brains. Our results thus reconcile the two seemingly incompatible views of the retina as either performing feature extraction or efficient compression of natural scenes, by suggesting that all vertebrates lie on a spectrum between these two objectives, depending on the neural resources allocated to their visual system.

1 INTRODUCTION

Why did natural selection shape our visual representations to be the way they are? Traditionally, the properties of the early visual system have been explained with theories of *efficient coding*, which are based on the premise that the neural representations are optimal at preserving information about the visual scene, under a set of metabolic constraints such as total firing rate or total number of synapses. These theories can successfully account for the antagonistic center-surround structure of receptive fields (RFs) found in the retina (Atick & Redlich, 1990; 1992; Vincent & Baddeley, 2003; Karklin & Simoncelli, 2011; Doi et al., 2012), as well as for the oriented structure of RFs found in the primary visual cortex V1 (Olshausen & Field, 1996; 1997; Bell & Sejnowski, 1997).

However, a number of properties of the early visual system remain unexplained. First, it is unclear why RF geometries would be so different in the retina and V1. A study (Vincent et al., 2005) has proposed that both representations are optimal at preserving visual information under different metabolic constraints: a constraint on total number of synapses for the retina, and one on total firing rate in V1. However, it is unclear why the two systems would be optimized for these two different objectives. Second, retinal representations differ dramatically across species (Roska & Meister, 2014). In small vertebrates (e.g. frog, salamander), the output of the retina is dominated by ganglion cell types that perform highly non-linear computations and extract behaviorally relevant

features, such as potential prey or predators (Gollisch & Meister, 2010). For example, cell types in the frog are tuned to detect moving flies or looming predators (Lettvin et al., 1959). In the mouse, the most numerous cell type is a non-linear feature detector thought to detect overhead predators (Zhang et al., 2012). These observations seem to indicate that each cell type acts as a feature map of the visual scene, dedicated to detecting one important behavioral cue in the environment (Gollisch & Meister, 2010; Roska & Meister, 2014). In contrast, in large vertebrates such as primates, the retina is dominated by cell types that process visual information quasi-linearly (midget and parasol cells, comprising 65% of all retinal ganglion cells and >95% in the fovea, Dacey (2004); Chichilnisky & Kalmar (2002)). In intermediate-brain-size species such as rodents or cats, the retina seems to combine cell types that do quasi-linear processing of visual information (X-cells, Fukuda & Stone (1974); Johnson et al. (2018)), and cell types that do highly non-linear processing (Y-cells and W-cells, Sur & Sherman (1982); Carcieri et al. (2003); Jacoby & Schwartz (2017)).

The limitations of current efficient coding theories might reside in the simplistic assumption that the objective is to simply relay indiscriminately all visual information to the next stages of processing. Indeed, the ultimate goal of the visual system is to extract *meaningful features* from the visual scene in order to produce an adequate behavioral response, not necessarily to faithfully encode it. A recent line of work has proposed using the information bottleneck framework as a way to move beyond the simplistic objective of information preservation towards more realistic objectives (Chalk et al., 2016; 2018). Another study has shown that by changing the objective from efficiently encoding the present to efficiently encoding the future (predictive coding), one could better account for the spatio-temporal RFs of V1 cells (Singer et al., 2018). Although promising, these approaches were limited to the study of a single layer of neurons, and they did not answer the aforementioned questions about cross-layer or cross-species differences. Deep convolutional networks have proven to be accurate models of the visual system, whether they are trained directly on reproducing neural activity (McIntosh et al., 2016; Cadena et al., 2017), or on a behaviorally relevant task (Yamins et al., 2014; Cadena et al., 2017), but they have not yet been used in the context of efficient coding theories.

In this study, we trained deep convolutional neural networks on image recognition (CIFAR-10, Krizhevsky (2009)) and varied their architectures to explore the sets of constraints that could have shaped vertebrates' early visual representations through natural selection. We modeled the visual system with a series of two convolutional networks, one corresponding to the retina and one downstream network corresponding to the ventral visual system in the brain. By varying the architecture of these networks, we first found that a reduction in the number of neurons at the retinal output – corresponding to a realistic physical constraint on the number of fibers in the optic nerve – accounts simultaneously for the emergence of center-surround RFs in our model of the retina, and for the emergence of oriented receptive fields in the primary visual relay of the brain. Furthermore, we show that the neural resources allocated to the brain network in our model shapes the retinal representation in a way consistent with biology. Given a deep visual network in our brain model, the retinal processing emerged as quasi-linear and retained most information about the visual scene. In contrast, for a shallow brain model, the retinal processing emerged as non-linear and more information-lossy, but was better at extracting relevant features for the object classification task. These observations are in striking agreement with the qualitative differences found in retinal representations across species, suggesting that we have successfully identified some of the key constraints shaping these representations.

2 Framework: a deep Convolutional Neural network model of the visual system

The retinal architecture is strongly conserved across species (Masland, 2001), and consists of three layers of feed-forward convolutional neurons (photoreceptors, bipolar cells, ganglion cells) and two layers of inhibitory interneurons (horizontal, amacrine cells). However, we chose to model the retina as a convolutional neural network (LeCun et al., 2015) with only two layers (fig. 1A). Indeed the retinal response of many species to complex stimuli has been modeled successfully with only one or two-layer models (Deny et al., 2017; Maheswaranathan et al., 2018; Gollisch & Meister, 2010), with some rare exceptions of models requiring more layers (McIntosh et al., 2016). We refer to this network as the retina-net. In our simulations, we varied the number of neurons in the second layer

of the retina-net, which is the output of the retina, corresponding to the physical bottleneck of the optic nerve conveying all the visual information to the brain (fig. 1B).

We modeled the ventral visual system – defined here as the system associated with object recognition in the brain across all vertebrates (Hubel, 1995) – as a convolutional neural network taking its inputs from the retina-net (fig. 1A). We varied the neural resources allocated to the ventral visual system network (VVS-net) by changing the number of layers in it (fig. 1B).

We trained the neural network composed of the retina-net and VVS-net end-to-end on an object classification task (CIFAR-10, fig. 1A-B-C). Even though the visual system does much more than just classify objects in natural images, this objective is already much more complex and biologically realistic than the one used in previous studies of efficient coding, namely preserving all information about the visual scene. Moreover, we are encouraged by the fact that previous studies using this objective have found a good agreement between neural activity in artificial and biological visual networks (Yamins et al., 2014; Cadena et al., 2017).

More specifically, we trained a convolutional neural network on a grayscale version of the standard CIFAR-10 dataset for image classification. The retina-net consisted of two convolutional layers of 32 channels and N_{BN} channels, respectively, with ReLU nonlinearities at each layer. The VVS-net consisted of a varying number D_{VVS} of convolutional layers with 32 channels followed by two fully connected layers, with ReLU nonlinearities at each layer. The full system encompassing the retina-net and VVS-net thus had $32 \rightarrow N_{BN} \rightarrow 32 \rightarrow 32 \rightarrow ...$ channels respectively, where we varied the retinal bottleneck N_{BN}, as well as the number D_{VVS} of convolutional brain layers (not counting the fully connected layers). We used 9x9 convolutional filters with a stride of 1 at each step. The large filter size was chosen to give the network flexibility in determining the optimal filter arrangement. We trained our network with the RMSProp optimizer for 20 epochs on the training set with batches of size 32. All optimizations were performed using Keras and TensorFlow. For all results presented, we tested statistical significance by training 10 identical networks with different random initializations of weights and biases taken from a Glorot-uniform distribution.

After training, we determined the linear approximation of RFs of each convolutional channel of the network in each layer. This was achieved by computing the gradient of the activation of that channel

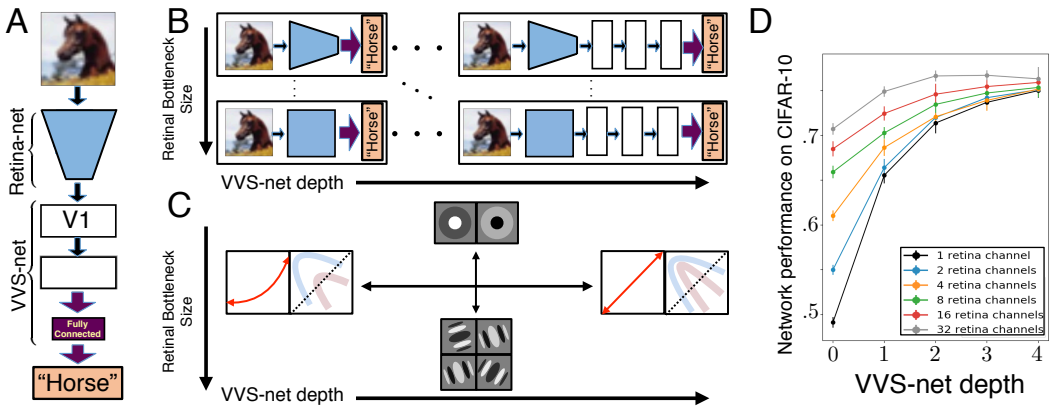

Figure 1: Illustration of the framework we used to model early visual representations. A: We trained convolutional neural networks on an image recognition task (CIFAR-10). The networks were composed of two parts, a retina-net and a ventral-visual-system-net (VVS-net), which receives input from the retina-net. B: We varied the number of layers in the VVS-net (white boxes) and the number of channels at the output of the retina-net (blue box). C: Key results (1) A bottlenck at the output of the retina yielded center-surround retinal RFs. (2) A shallow VVS-net yielded more non-linear retinal responses (linearity is schematized by the red arrow), which better disentangled image classes (represented as bent manifolds). D: Test-set accuracy of all model architectures on CIFAR-10, averaged over ten trials (i.e. random initial weights) each. Performance increases with VVS-net depth and retinal channel, indicating that both factors are meaningful constraints on the network in the regime tested.

with respect to a blank image. This gradient map gives a first-order approximation of the image pattern that maximally activates the cells in the channel of interest. This computation is mathematically equivalent to measuring the cell's spike-triggered average in response to a perturbative white-noise stimulus (Koelling & Nykamp, 2008; Schwartz et al., 2006), a commonly used method for determining receptive fields in experimental biology (Chichilnisky, 2001). This equivalence allowed us to compare directly the geometries of RFs experimentally measured in biological networks with the ones found in our models.

The performance of our neural network model of the visual system at the recognition task increased both with the number of channels in the retinal bottleneck, and with the number of layers in the VVS (fig. 1D), confirming that we were in a regime where the restrictions on neural resources in the VVS and at the output of the retina were critical to the ability of the network to perform the task.

3 A UNIFIED MODEL FOR CENTER-SURROUND RFS IN THE RETINA AND ORIENTED RFS IN V1

3.1 A DIMENSIONALITY BOTTLENECK AT THE RETINAL OUTPUT YIELDS THE EXPECTED REPRESENTATIONS IN RETINA AND V1

When reducing the number of neurons at the output of the retina we found that RFs with antagonistic center and surround emerged. For $N_{BN} = 32$, our control setting with no bottleneck at the retinal output, we observed mostly oriented receptive fields in the second layer of the network (fig. 2A). For $N_{BN} = 4, 2$, and 1, we observed center-surround receptive fields in the second layer of the network and mostly oriented receptive fields in the third layer, which is the first layer of the ventral visual system in our model (fig. 2B). The RF geometries did not depend qualitatively on the VVS-net depth D_{VVS} (results shown for $D_{VVS} = 2$). These results are in good agreement with the organization of the biological visual system, where retinal RFs are center-surround and most downstream RFs in primary visual cortex (V1) are sharply oriented (Hubel, 1995). Moreover, these results suggest that the key constraint explaining these differences of representations might be the dimensionality bottleneck at the output of the retina. It is worth noting that for both conditions (bottleneck or no bottleneck), the RFs of downstream layers in the VVS-net after the first layer exhibited complex shapes that were neither clearly oriented, nor circular, and the RFs in the first layer of the retina did not appear to have any well-defined structure (data not shown).

We then tested in our model the hypothesis of Hubel and Wiesel concerning how center-surround cells are pooled to give rise to oriented RFs in V1 (Hubel, 1995). We found that orientation-selective neurons in the VVS-net typically draw primarily from center-surround neurons in the retina-net that are aligned with the direction of the edge, with positive or negative weights corresponding to whether the polarity (light-selective / dark-selective) of the two neurons are consistent or inconsistent (fig. 2C). These qualitative results are in good agreement with Hubel and Wiesel's hypothesis (Hubel, 1995). Of course, this hypothesis remains to be tested in the real brain, since there is no evidence that the micro-circuitry of the brain matches that of our simulation.

In the visual system of mammals, the main relay of visual information taking its input from the retina is the LGN (thalamus), which has center-surround RFs and a similar total number of neurons as the retinal output (Hubel, 1995). We created a network reflecting this architecture by having two low-dimensionality layers in a row instead of just one (fig. 2C). After training, we found center-surround RFs in the two layers with a bottleneck (retinal output and LGN), and oriented RFs in the next layer, corresponding to the primary visual cortex (V1). These results suggest that center-surround representations remain advantageous as long as the dimensionality of the representation remains low, and hence dimensionality expansion seems to be the crucial factor explaining the qualitative change of RFs found between LGN and V1.

3.2 EMERGENCE OF ON AND OFF POPULATIONS OF CENTER-SURROUND CELLS IN THE RETINA

We then verified that the emergence of center-surround RFs in the retina-net is a consequence of reducing the number of neurons at the retinal output, not of reducing the number of channels, our model's equivalent of biological cell types. In order to test this, we employed locally connected

layers – equivalent to convolutional layers, but without parameter-tying between artificial neurons within a channel at different spatial locations. Such a network contains too many parameters to be trained from scratch by gradient descent; to work around this, we trained the model stage-wise by first training our convolutional control network ($N_{BN} = 32$ with parameter tying) and then substituting the untied retina-net in place of the first layer of the control model for the rest of training. Even in the untied retina-net, in which each neuron is effectively its own channel, we found that center-surround RFs emerge (fig. 2E). Moreover, the neurons clustered in two populations of ON and OFF cells. We clustered the cells by measuring their activations in response to 10000 natural images, computing the first 20 principle components of this 10000-dimensional space, and running t-SNE to visualize the clustering of neuron types. We found that two distinct clusters emerged, that corresponded visually to ON and OFF center-surround RFs (fig. 2F). These results indicate that center-surround RFs are the network's preferred strategy for passing information through a dimensionality bottleneck even when no constraint on the number of cell types is imposed.

4 Retinal representations are function of the neural resources allocated to the brain

4.1 The retina becomes more linear as brain complexity increases

We then showed that the representation at the output of the retina-net varies considerably as we vary the total number of layers in the VVS-net in a way that is consistent with our knowledge on inter-species differences in retinal representations (fig. 1C). We characterize the differences in retinal representations below.

As we increased the number of layers in the VVS-net, the retinal computation became more linear (fig. 3A), as measured by the ability of the raw image to linearly map on to the neural representation of the retinal output (see methods). This observation is in good agreement with the retinal representations found experimentally across vertebrate species with different brain sizes (see intro). The linearization of the retinal response with increased brain complexity was true for different values of bottleneck N_{BN}. However, when we did not use any bottleneck ($N_{BN} = 32$), the trend became non-monotonic, with a peak of linearity of the response when the VVS-net had 1 conv layer (data not

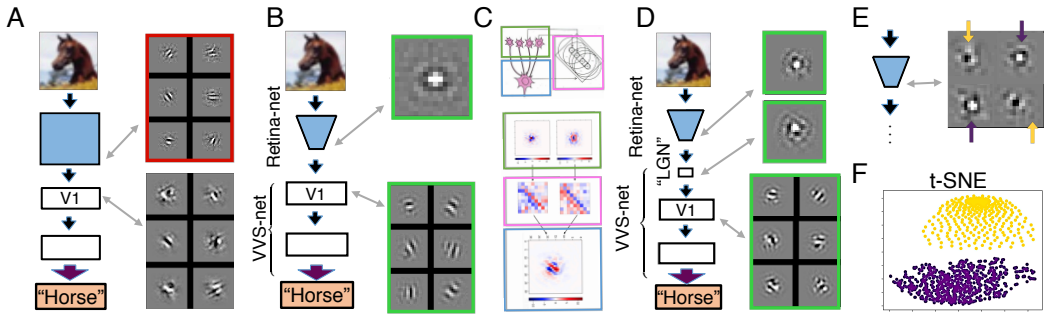

Figure 2: Effects of a bottleneck constraint on receptive fields (RFs). All results are shown for $D_{VVS} = 2$. A: Examples of RFs of cells at selected layers of a control network (layers 2 and 3). No center-surround RFs appear. B: Center-surround RFs emerge at the output of the retina-net (layer 2) and oriented RFs emerge in the first layer of the VVS-net when we impose a bottleneck constraint at the output of the retina ($N_{BN} = 1$) C: Top: Hubel and Wiesel's hypothesis on oriented cell formation in V1 (Hubel, 1995). Bottom: A representative example of an orientation-selective neuron (bottom RF) drawing from center-surround channels (top RFs) in the previous layer with weight matrices (center) according to their polarity. Light / dark-selective regions of a receptive field, and positive / negative weights, are represented with red / blue, respectively. D: Examples of RFs in a network with an extra bottleneck layer corresponding to mammalian LGN. Center-surround RFs appear at both the retinal output and LGN layer. E: Examples of ON and OFF center-surround RFs in the untied network ($N_{BN} = 4$). F: t-SNE clustering of the retinal neurons of the untied network (see text). Two distinct cell type clusters form corresponding to ON and OFF center-surround receptive fields.

shown). Another interesting phenomenon to note is that linearity of the retinal response decreased as we increased the number of channels in the bottleneck, regardless of brain depth (fig. 3A).

Is the nonlinearity of retinal response due to the first or second stage of non-linearity in our retina-net? To test this, we represented the real response against the response predicted by a purely linear model for the most shallow and for the deepest VVS-nets (fig. 3B). If the linear prediction were inaccurate because of the first stage of non-linear processing in the retina-net, we would expect the points on the scatter plot to be scattered around the unit line. If the prediction error were due to the second-stage of non-linearity, we would expect the linear approximation to make incorrect negative predictions for inactive neurons. In practice, we found that the prediction error of the linear model was partly explained by both stages of non-linearity in the retina-net model.

4.2 THE RETINAL REPRESENTATION IS THE RESULT OF A TRADE-OFF BETWEEN INFORMATION TRANSMISSION AND FEATURE EXTRACTION

Why would retinal representation be more linear when the brain has more resources? One hypothesis is that with a restricted number of neurons, the retina must trade-off between the two incentives of (1) compressing visual information in order to transmit it to downstream layers and (2) extracting non-linear features of the image to start disentangling the manifolds corresponding to different classes of objects (Chung et al., 2018a;b). According to this hypothesis, when the VVS is shallow, the priority of the retina should be to work toward extracting relevant features. When the VVS is deep, the

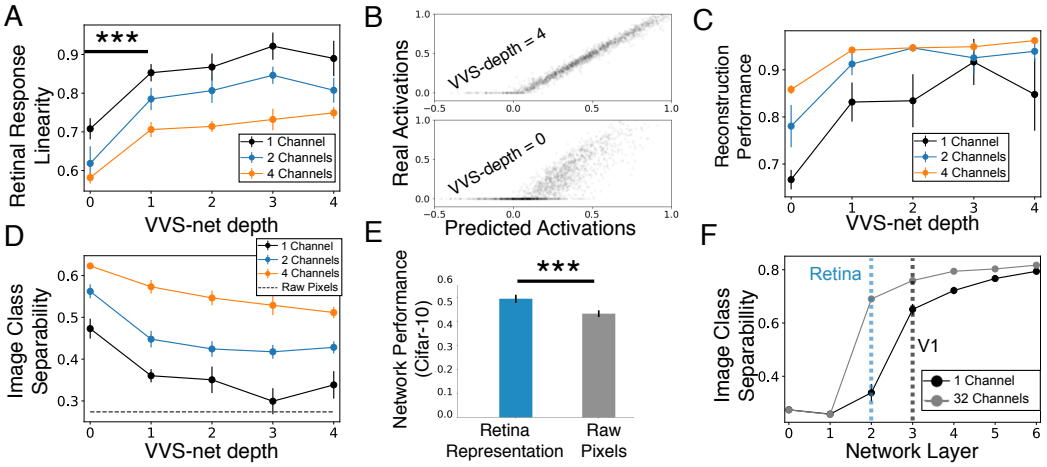

Figure 3: Retinal representations as a function of the depth of the VVS-net. All error bars represent the 95% confidence interval about the mean (all simulations were repeated over 10 networks trained from random initial conditions). Three stars indicate t-test significance ($p<0.001$). A: Linearity of the retinal response increases with the number of layers in the VVS-net. Note that it also decreases with the number of cells at the retinal output (different lines). B: Responses of example retina-net output cell to natural images, vs. best linear fit prediction from raw image, for most (top) and least (bottom) deep VVS-nets. Nonlinearity arises from two sources: rectification within the retina-net (corresponds to the spread of the bulk of the point cloud) and rectification at the retina-net output (corresponds to inactive neurons being incorrectly predicted to have negative activations). C: Quality of image reconstruction from the retinal representation as a function of VVS-net depth. The retinal representation retains more information about the raw image for deep VVS-nets. D: Linear separability of classes of objects at the retinal output, as a function of the VVS-net depth. Dashed line indicates separability of classes of images from the raw image pixels. Classes are less separable at the retinal output for deeper VVS-nets. E: Performance on CIFAR-10 for a two-layer densely connected network taking its input from the retina-net or from a raw image. Class information is more accessible from retinal representation. F: Class separability at all layers of network for a deep VVS-net ($D_{VVS} = 4$) with and without bottleneck ($N_{BN} = 1$ and $N_{BN} = 32$). Retinal representation of bottleneck network has low separability. However, the first layer of the VVS-net has high separability (see text).

priority of the retina should be to transmit as much visual information as possible for downstream processing. We validated this hypothesis in two ways in our model.

First we showed that the retinal representation retained more information about the image as the VVS-net complexity increased (fig. 3C). To quantify information retention, we trained a linear decoder (see methods) from the output of the retina to reconstruct the image and we measured the reconstruction error. The reconstruction error provided a lower bound on the information that the retina retained about the stimulus. This result corroborated our hypothesis that, as the VVS-net becomes more complex, the retinal representation gets better at retaining visual information for further processing by the VVS-net.

Second, we found that different classes of objects of CIFAR-10 (e.g. trucks, frogs) were more linearly separable from the retina-net representation when the VVS-net was shallow than when it was deep (fig. 3D). To measure linear separability of manifolds, we trained a linear SVM decoder to separate all pairs of classes and evaluated the performance of the SVM classifier on held-out images (see methods). Moreover, we showed that a VVS-net consisting of two fully connected layers only (no convolutional layers) equipped with a retina with a tight bottleneck $N_{BN} = 1$ (dimensionality of retinal output matches dimensionality of the input image) performed better at image recognition than the same VVS-net trained without a retina-net, taking raw images as input (fig. 3E). Both these results corroborate our hypothesis that the retina of a simple animal performs meaningful feature extraction whereas the retinas of animals with more complex visual cortices prioritize non-lossy encoding, postponing feature extraction to downstream layers that are better equipped to do it.

In the case of the deepest VVS-nets tested, the retinal processing was quasi-linear for the tightest bottleneck (cc = 0.9, $N_{BN} = 1$, fig. 3A). One might take this result to suggest that the retina-net in such models does little more than copy image information. However the very first layer of the VVS-net after the retina disentangled classes (as measured by linear separability) almost as well as the second layer of a VVS-net without retina (fig. 3F), suggesting that the retinal representation, while only moderately linearly separable itself, is especially transformable in a representation with a high linear separability. This result suggests that even when the retina-net is quasi-linear, it can still participate in extracting relevant features for downstream processing by the brain.

5 METHODS

To estimate the linearity of the response of retinal neurons, we fitted a linear model to predict the neural response from the image on 8,000 images. In order to prevent overfitting, we regularized the linear weights with an L2 penalty and optimized the weights using ridge regression. The value of the penalty term was chosen by 10-fold cross-validation on the training set. We then measured the Pearson correlation between the predicted and real responses on a testing set of 2,000 images.

To estimate the information about the input image retained by the retinal output representation, we fitted a linear model to reconstruct the image from the (fixed) outputs of the trained retina-net of interest. All numerical figures given are variance-explained results on the held-out test set.

To estimate the linear separability of classes of objects from the neural representation, we trained an SVM classifier between all pairs of classes on half of the testing set of CIFAR-10 (1,000 images that were not used to train the network), and we tested the performance of the SVM classifier on 1,000 held-out images from the testing set, as measured by the percentage of images classified correctly. We then averaged the performance of the SVM across all pairs of classes to obtain the linear separability score.

6 DISCUSSION

The structural differences between the receptive field shapes of retinal neurons and V1 neurons have until now eluded efficient coding theories. Karklin & Simoncelli (2011) found that efficient encoding of images with added noise and a cost on firing rate produce center-surround RFs, whereas the same task without noise produces edge detectors. However, this observation (as they note) does not explain the discrepancy between retinal and cortical representations. Vincent et al. (2005) propose a different set of constraints for the retina and V1, in which the retina optimizes for a metabolic constraint on total number of synapses, whereas V1 optimizes for a constraint on total firing rate.

It is not clear why each of these constraints would predominate in each respective system. Here we propose that these two representations emerge from the requirement to perform a biologically relevant task (extracting object identity from an image) with a bottleneck constraint on the dimensionality of the retinal output. Interestingly, this constraint differs from the ones used previously to account for center-surround RFs (number of synapses or total firing rate). It is worth noting that we unsuccessfully tried to reproduce the result of Karklin & Simoncelli (2011) in our network, by adding noise to the image and applying an L1 regularization to the retina-net activations. In our framework (different than the one of Karklin & Simoncelli (2011) in many ways), the receptive fields of the retina-net remained oriented across the full range of orders of magnitude of noise and L1 regularization that permitted successful task performance.

Another mystery in neuroscience is the incredible diversity of computations done by the retina across species. Neuroscientists have hypothesized that the difference of retinal processing found across species could reflect the different statistics of the environment of these species (Matsumoto & Yonehara, 2018), as well as the statistics of the cues relevant to each animal (e.g prey or predators, (Ishikane et al., 2005; Semmelhack et al., 2014; Temizer et al., 2015; Yilmaz & Meister, 2013)). Here we show that our model of the visual system, trained on the same task and with the same input statistics, can exhibit very different retinal representations depending on the amount of neural resource allocated to downstream processing by the brain. Of course, this possibility does not exclude the further specialization of the retinal circuit for specific environment statistics or specific behavioral requirements.

There is a long-standing controversy on whether the role of the retina is to extract relevant features from the environment (Lettvin et al., 1959; Gollisch & Meister, 2010; Roska & Meister, 2014), or to efficiently encode visual information (Barlow, 1961; Atick & Redlich, 1990; 1992). Here we propose that, despite its conserved structure across evolution, the retina serves different purposes in different species. In species with fewer brain resources devoted to visual processing, it extracts relevant features from the environment, and in species with complex visual cortices, it prioritizes an efficient transmission of visual information for further processing by the brain. This hypothesis could be directly tested experimentally by systematically measuring the linear separability between object classes from the retinal output of different species, and by estimating the information that the retinal responses carry about visual scenes, as we do with our model. However, this cross-species study would require a dataset that does not currently exist, consisting of recordings of retinal responses of different species to a common bank of visual stimuli. Mathematically, it would also be interesting to understand why center-surround RFs are well-suited to compressing natural statistics, whereas oriented RFs are well-suited to disentangling manifolds corresponding to different classes of objects.

To further investigate the dependency between neural resources and retinal representations, it would be interesting to systematically compare the computations performed by retinal cell types of different species to the neurons of the retina-net. In this study, we allowed only a limited number of channels in the retinal bottleneck (1 to 4) to test the effect of high ratios between the representation in the brain (32 channels) and the retinal representation (a crucial condition to see the center-surround representation emerge). By using larger networks with more channels in the retina-net and the VVS-net, we could see the emergence of a greater diversity of neuron types in our retina-net and compare their properties to real retinal cell types. It would also be interesting to extend our approach to dynamical and recurrent neural networks applied to natural movies. Indeed, most feature detectors identified to date seem to process some form of image motion: wide-field, local or differential (Roska & Meister, 2014). Adding a temporal dimension to the model would be crucial to study them.

In summary, by studying the representations learned by a deep network trained on a biological task, we found that the learned retinal representations are compatible with the constraint of transmitting visual information through a low-dimensional communication channel, the optic nerve, and that the diversity of retinal representations found across species could be a consequence of the varying sophistication of the downstream processing performed by the brain. These insights illustrate how deep neural networks, whose creation was once inspired by the visual system, can shed light on the constraints and objectives that have shaped our visual systems through evolution.

REFERENCES

Joseph J. Atick and A. Norman Redlich. Towards a theory of early visual processing. *Neural Computation*, 2(3):308–320, 1990.

Joseph J. Atick and A. Norman Redlich. What does the retina know about natural scenes? *Neural computation*, 4(2):196–210, 1992.

HB Barlow. Possible principles underlying the transformations of sensory messages. In WA Rosenblith (ed.), *Sensory Communication*, pp. 217–234. MIT Press, 1961.

A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, December 1997. ISSN 0042-6989.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque V1 responses to natural images. *bioRxiv*, pp. 201764, October 2017. doi: 10.1101/201764.

Stephen M. Carcieri, Adam L. Jacobs, and Sheila Nirenberg. Classification of Retinal Ganglion Cells: A Statistical Approach. *Journal of Neurophysiology*, 90(3):1704–1713, September 2003. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.00127.2003.

Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1957–1965. Curran Associates, Inc., 2016.

Matthew Chalk, Olivier Marre, and Gaper Tkaik. Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences of the United States of America*, 115(1):186–191, 2018. ISSN 1091-6490. doi: 10.1073/pnas.1711114115.

E. J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network (Bristol, England)*, 12(2):199–213, May 2001. ISSN 0954-898X.

E. J. Chichilnisky and Rachel S. Kalmar. Functional asymmetries in ON and OFF ganglion cells of primate retina. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 22(7):2737–2747, April 2002. ISSN 1529-2401. doi: 20026215.

SueYeon Chung, Uri Cohen, Haim Sompolinsky, and Daniel D. Lee. Learning Data Manifolds with a Cutting Plane Method. *Neural Computation*, 30(10):2593–2615, October 2018a. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_01119.

SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and Geometry of General Perceptual Manifolds. *Physical Review X*, 8(3), July 2018b. ISSN 2160-3308. doi: 10.1103/PhysRevX.8.031003.

Dennis Dacey. Origins of perception: retinal ganglion cell diversity and the creation of parallel visual pathways. *In The cognitive neuroscience (2004)*, pp. 281–301, 2004.

Stephane Deny, Ulisse Ferrari, Emilie Mace, Pierre Yger, Romain Caplette, Serge Picaud, Gaper Tkaik, and Olivier Marre. Multiplexed computations in retinal ganglion cells of a single type. *Nature communications*, 8(1):1964, 2017.

Eizaburo Doi, Jeffrey L. Gauthier, Greg D. Field, Jonathon Shlens, Alexander Sher, Martin Greschner, Timothy A. Machado, Lauren H. Jepson, Keith Mathieson, Deborah E. Gunning, Alan M. Litke, Liam Paninski, E. J. Chichilnisky, and Eero P. Simoncelli. Efficient coding of spatial information in the primate retina. *The Journal of Neuroscience*, 32(46):16256–16264, November 2012.

Y Fukuda and J Stone. Retinal distribution and central projections of Y-, X-, and W-cells of the cat's retina. *Journal of Neurophysiology*, 37(4):749–772, July 1974. ISSN 0022-3077. doi: 10.1152/jn.1974.37.4.749.

Tim Gollisch and Markus Meister. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, 65(2):150–164, January 2010. ISSN 1097-4199. doi: 10.1016/j.neuron.2009.12.009.

David H. Hubel. *Eye, brain, and vision.* Scientific American Library/Scientific American Books, 1995.

Hiroshi Ishikane, Mie Gangi, Shoko Honda, and Masao Tachibana. Synchronized retinal oscillations encode essential information for escape behavior in frogs. *Nature Neuroscience*, 8(8):1087–1095, August 2005. ISSN 1097-6256. doi: 10.1038/nn1497.

Jason Jacoby and Gregory W. Schwartz. Three Small-Receptive-Field Ganglion Cells in the Mouse Retina Are Distinctly Tuned to Size, Speed, and Object Motion. *The Journal of Neuroscience*, 37 (3):610–625, January 2017. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2804-16.2016.

Keith P. Johnson, Lei Zhao, and Daniel Kerschensteiner. A Pixel-Encoder Retinal Ganglion Cell with Spatially Offset Excitatory and Inhibitory Receptive Fields. *Cell Reports*, 22(6):1462–1472, February 2018. ISSN 22111247. doi: 10.1016/j.celrep.2018.01.037.

Yan Karklin and Eero P. Simoncelli. Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24*, pp. 999–1007. Curran Associates, Inc., 2011.

Melinda E. Koelling and Duane Q. Nykamp. Computing linear approximations to nonlinear neuronal response. *Network (Bristol, England)*, 19(4):286–313, 2008. ISSN 1361-6536. doi: 10.1080/09548980802503139.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. pp. 60, 2009.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539.

J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts. What the Frog's Eye Tells the Frog's Brain. *Proceedings of the IRE*, 47(11):1940–1951, November 1959. ISSN 0096-8390. doi: 10.1109/JRPROC.1959.287207.

Niru Maheswaranathan, David B. Kastner, Stephen A. Baccus, and Surya Ganguli. Inferring hidden structure in multilayered neural circuits. *PLoS computational biology*, 14(8):e1006291, August 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006291.

Richard H. Masland. The fundamental plan of the retina. *Nature Neuroscience*, 4(9):877–886, September 2001. ISSN 1546-1726. doi: 10.1038/nn0901-877.

Akihiro Matsumoto and Keisuke Yonehara. Visual Circuits: Division of Labor Revealed. *Current Biology*, 28(5):R208–R210, 2018.

Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus. Deep Learning Models of the Retinal Response to Natural Scenes. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1369–1377. Curran Associates, Inc., 2016.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996. ISSN 0028-0836. doi: 10.1038/381607a0.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, 37(23):3311–3325, December 1997. ISSN 0042-6989.

Botond Roska and Markus Meister. The Retina Dissects the Visual Scene into Distinct Features. In *The New Visual Neurosciences (Werner, JS, Chalupa, LM, eds), pp 163182.*, pp. 20. Cambridge, MA: MIT Press, 2014.

Odelia Schwartz, Jonathan W. Pillow, Nicole C. Rust, and Eero P. Simoncelli. Spike-triggered neural characterization. *Journal of Vision*, 6(4):13, July 2006. ISSN 1534-7362. doi: 10.1167/6.4.13.

Julia L. Semmelhack, Joseph C. Donovan, Tod R. Thiele, Enrico Kuehn, Eva Laurell, and Herwig Baier. A dedicated visual pathway for prey detection in larval zebrafish. *eLife*, 3, December 2014. ISSN 2050-084X. doi: 10.7554/eLife.04878.

Yosef Singer, Yayoi Teramoto, Ben DB Willmore, Jan WH Schnupp, Andrew J. King, and Nicol S. Harper. Sensory cortex is optimized for prediction of future input, June 2018.

M. Sur and S. M. Sherman. Linear and nonlinear W-cells in C-laminae of the cat's lateral geniculate nucleus. *Journal of Neurophysiology*, 47(5):869–884, May 1982. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.1982.47.5.869.

Incinur Temizer, Joseph C. Donovan, Herwig Baier, and Julia L. Semmelhack. A Visual Pathway for Looming-Evoked Escape in Larval Zebrafish. *Current biology: CB*, 25(14):1823–1834, July 2015. ISSN 1879-0445. doi: 10.1016/j.cub.2015.06.002.

Benjamin T. Vincent and Roland J. Baddeley. Synaptic energy efficiency in retinal processing. *Vision Research*, 43(11):1283–1290, May 2003. ISSN 0042-6989.

Benjamin T. Vincent, Roland J. Baddeley, Tom Troscianko, and Iain D. Gilchrist. Is the early visual system optimised to be energy efficient? *Network: Computation in Neural Systems*, 16(2-3): 175–190, January 2005. ISSN 0954-898X. doi: 10.1080/09548980500290047.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111 (23):8619–8624, June 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111.

Melis Yilmaz and Markus Meister. Rapid innate defensive responses of mice to looming visual stimuli. *Current biology: CB*, 23(20):2011–2015, October 2013. ISSN 1879-0445. doi: 10.1016/j.cub.2013.08.015.

Yifeng Zhang, In-Jung Kim, Joshua R. Sanes, and Markus Meister. The most numerous ganglion cell type of the mouse retina is a selective feature detector. *Proceedings of the National Academy of Sciences*, pp. 201211547, August 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1211547109.