Augmenting goal-driven CNN models of visual cortex with a data-driven CNN model of the retina

Elias Wang
Department of Electrical Engineering
Stanford University
elias.wang@stanford.edu

Jonathon Walters
Department of Psychology
Stanford University
waltersj@stanford.edu

Akshay Jagadeesh
Department of Psychology
Stanford University
akshayj@stanford.edu

Tyler Bonnen
Department of Psychology
Stanford University
bonnen@stanford.edu

Abstract

There is a striking correspondence between goal-driven models of sensory processing and primate sensory cortex. For example, convolutional neural networks (CNNs) optimized for visual object recognition produce state-of-the-art results in predicting neural responses to natural images in macaque V1, V4, and IT ([8], [1]). Unlike the primate visual systems, however, these models receive raw pixels as input, while primary visual cortex receives input already highly transformed by earlier processing stages. In the retina alone, raw inputs undergo a 10-fold dimensionality reduction before leaving the retina via retinal ganglion cells (RGCs) [7]. To incorporate these representational constraints into a hierarchical model of visual cortex, we replaced the early layers of a goal-driven model (VGG16) with a model of the Salamander retina. This retinal front-end was a 3-layer CNN trained to predict the firing rates of Salamander RGC responses to natural images. We fixed the weights of this retinal front-end and fine-tuned the weights of the remaining cortical VGG16 [6] on ImageNet [2]. To evaluate our model, we compared the ability of the original VGG16 and the retinal-VGG16 to predict neural responses in macaque V4 and IT. Our results indicate that our current methods are not able to predict retinal responses with pretrained VGG16 weights better than random weights and that adding a retinal frond-end does not improve neural fits to IT.

1. Introduction

There is a striking correspondence between goal-driven models of sensory processing and primate sensory cortex. Even beyond formal model comparison, qualitative analysis reveals that edge-detection emerges in primary visual cortex as well as the principal layers of CNNs, that face detection emerges in downstream layers of both networks just as spatial information is preserved across these networks, etc.. Unlike these convolutional models of vision, however, the cortex does not receive pixel inputs. Sensory information is transformed, principally, within the retina, where visual inputs undergo a 10-fold dimensionality reduction. Can we incorporate these representational constraints into a hierarchical model of vision? And might this biologically constrained model better account for variance in the neural data? We propose to build a goal-driven model of vision trained not from the raw pixels from images, but from the outputs of a retinal front end. This front end will be a model pre-trained on natural images to predict retinal ganglion cells. The downstream cortical regions will be a convolutional neural network constrained to loosely approximate the architecture of visual cortex (e.g. not 15 layers). After pre-training the retinal front end, its weights will be fixed and the cortical layers will be trained on Imagenet. Validation will occur using the neural data and visual stimuli that has been available to us throughout the course.

2. Methods

2.1. Neural Data

Here we describe the neural data used in our experiments. Retinal data is obtained from salamander RGCs and higher visual cortex (V4 and IT) recordings are from macaques.

2.1.1 Salamander Retinal Data

The activity of 5 OFF-type tiger salamander RGCs were recorded in response to two kinds of stimuli: white noise (spatiotemporal patterns of binary checkers) and natural scenes (natural images sampled from a natural image database that were jittered to match statistics of eye movements). These data were acquired by the Baccus Lab at Stanford University, and additional details can be found in [5]. As the RGC window of temporal integration is approximately 40 frames under a 100 Hz frame rate, each stimulus represented a short movie consisting of 50x50 images across 40 frames (i.e., 50x50x40). Each stimulus was labeled with the firing rates of the 5 RGCs, which were computed from 10ms bins and a 10ms Gaussian smoothing filter. The white noise dataset had 323,762 training images and 5,957 test images, while the natural scene dataset had 323,756 training images and 5,956 test images. The datasets were represented in TFRecords format.

2.1.2 Macaque V4 and IT data

The activity of neurons in macaque V4 and IT were recorded in response to a stimulus set of 6,000 natural images. These images comprised 64 3-dimensional objects across 8 categories, superimposed upon a randomly selected photographic background. Images were presented according to three levels of variation (low, medium, and high) in terms of object position, pose, and size. Additional details can be found in [8].

2.2. CNN model of the Salamander Retina

2.2.1 Fitting Early Layers of Pretrained VGG16 to Retinal Responses

As we sought to replace the most retina-like part of VGG16 with a CNN model of the Salamander retina, we aimed to test the ability of each early convolutional layer to predict retinal data, removing the part of VGG16 up to and including the layer that led to the best retinal fit. Due to time constraints we evaluated only the feature maps of conv2_1. To evaluate fit to RGC data, a fully connected layer was added after conv2_1 with a softplus activation function and Poisson loss function. As the retinal data input has a depth of 40 and the VGG16 expects an input depth of 3, a 1x1 convolutional layer was preprended to project 40 channels to 3. The FC layer we used a random normal initialization of mean 0 and standard deviation of 0.05 (with biases also initialized at 0). We used Adam optimizer with a learning rate of 1e4. Batch size was set to 5000.

For evaluation, we calculated the Pearson correlation coefficient for each of the 5 RGCs:

$$r = \frac{\sum_{i=1}^{n}(\hat{y}^{(i)} - s\left(\hat{y}^{(i)}\right))(y^{(i)} - s\left(y^{(i)}\right))}{std(\hat{y}^{(i)}) \cdot std(y^{(i)})}, \qquad (1)$$

where n is the number of the test images, $y(i)$ and $\hat{y}(i)$ are, respectively, the actual and predicted firing rates of the 5 RGCs, and $s(\cdot)$ and $std(\cdot)$ are batch means and standard deviations for each of the 5 RGCs.

We also experiment with adding batch normalization before the VGG conv layers and subsampling the output VGG conv features before the final fc layer. The intuition behind the batch normalization is alleviate the mismatch between the output statistics of the 1x1 conv and the inputs to the pretrained VGG weights. Since the output of the final conv layer has shape 25x25x128, we added subsampling to make the dimensions of the final fc more reasonable. Using a spatial subsampling of 5x, we reduce the total features from 80,000 to 3,200. The final architecture is shown in Fig. 1. Since we are investigating the fit of the pretrained VGG weights, these are kept frozen and only the layers outside the purple box in Fig. 1 are trained.

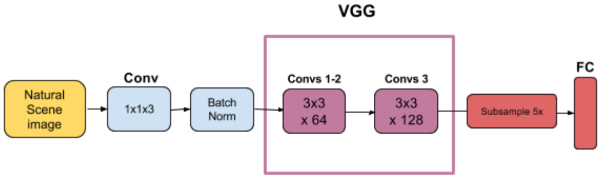

Figure 1. Diagram of architecture for fitting pretrained VGG to retinal responses.

2.2.2 Retinal Response Model (RRM): Training a CNN to Predict Retinal Responses to Natural Scene Stimuli

The first and second convolutional layers used 16 15x15 filters and 8 9x9 filters, respectively. Both used a stride of 1, VALID padding, L2 regularization of 1e-3, and a ReLU activation function. During training, Gaussian noise was added to the activations in each of these layers before being passed into the ReLU, with a mean of 0 and standard deviation of 0.1. The fully connected layer had an output dimension of 5, used a softplus nonlinearity and L2 regularization of 1e-3. We used Xavier initialization for the convolutional layers (with biases initialized at 0). The FC layer we used a random normal initialization of mean 0 and standard deviation of 0.05 (with biases also initialized at 0). We used Adam optimizer with a learning rate of 1e4. Batch size was set to 5000.

We used the following Poisson loss function:

$$\ell(y^{(i)}, \hat{y}^{(i)}) = \frac{1}{N}\left(\sum_{k=1}^{N}\left(\hat{y}_k^{(i)} - y_k^{(i)} \cdot \log(\hat{y}_k^{(i)} + \varepsilon)\right)\right) \qquad (2)$$

2.3. Adding the retina model to VGG16

2.3.1 Replacing early layers of VGG16 with CNN model of retina

In order to emulate the architectural and representational constraints that act on biological visual systems, we combined models that reflect both retinal and cortical information processing: A convolutional model (RRM) trained to predict the activity of RGCs and a goal-driven model (VGG19) of image classification that demonstrates a high correspondence between. In order to integrate these models, there are two architectural challenges this project has addressed. The first is firmly motivated by neuroscientific questions: How can we integrate these two computational models in a way that best reflects the functional relationship between the retina and the cortex? The second stems from more technical considerations: Given the data sets available, how can the the RMM take inputs of the appropriate dimensions and output tensors tailored to the expectations of VGG16?

Our solution to these challenges has been inspired by recent work on the correspondence between early layers of VGG and primary sensory cortex. Cadena et al. have shown in [1] that it is not the earliest, but intermediate layers of VGG that best predict neural firing (e.g. conv3 and not 'conv1'). This leaves open the possibility that prior layers of VGG are a better approximation of processing in either the retina or lateral geniculate nucleus. Our response to the first challenge above, then, is to replace the first two layers of VGG16 with our RRM model. Because responses in the retina are thought to be stable [4], we fixed the weights in RRM model. In order to map the representations from the final layer of the RRM to the input space required by VGG, we use a single convolutional layer whose weights must be learned.

Learning to map between RRM and VGG, in this case, was accomplished by training on object categorization with Imagenet. This requires, however, that the RRM can take Imagenet as inputs–that is, reformatting 224 x 224 x 3 images as inputs into the 50 x 50 x 40 base layer of the RRM. To accomplished this, we first resized the images to fit the 50x50 input size of the first retinal convolutional kernel; then we used a 1x1 convolutional kernel with output depth of 40 to project the 3 channels from the ImageNet images to the 40 channels accepted by the retinal model. These modifications allow us to pass labeled images to a 3-layer CNN, which has been trained to predict the firing rates of RGC responses to natural images. In this case, however, we are passing the "retinal predictions" from the RRM to 'conv3_1' of VGG, which is thought to roughly reflect the representational structure of primary visual cortex.

We trained the retinal-VGG16 model on the ImageNet dataset to better fit the weights of our retinal-VGG16 model. Specifically, we froze the RRM weights while training the other layers, so as to fine tune the mapping between the RRM output and the VGG representation.

2.3.2 Evaluating fit of retinal-VGG16 to macaque V4 and IT data

We apply the methods for neural fitting used for Assignment 1. We trained a linear regression on the activations of each layer in the VGG component of the network to predict the neural activations of IT. We look at the correlations between the representational dissimilarity matrix (RDM)[3] of our retinal-VGG model and macaque neural data from higher visual cortex as well as the R^2 of the PLS regression. These are computed for various layers in our model.

3. Results

3.1. Fitting Early Layers of Pretrained VGG16 to Retinal Responses

We compare the final Pearson correlation values for the 5 RGCs when using pretrained VGG weights versus random weights. The two networks were trained for 5000 steps and the training and validation losses are shown in Fig. 3. We notice that both the pretrained VGG weights and the random weights reach about the same Poisson loss value of ~ -2.

Fig. 4 shows the mean Pearson correlation between our model outputs and the RGCs responses for the two models, pretrained VGG and random. We observe that the correlation for random weights is actually slightly higher but this is not statistically significant. This suggests that our current method of projecting the temporal dimension of the retinal inputs into the color channels may not be the optimal approach for leveraging the pretrained VGG weights. An alternative would be some sort of recurrent structure that would allow us to feed in the temporal stack of images to the pretrained weights sequentially. Another possibility is that batch normalization was not sufficient for replicating the effect of the specific preprocessing used for the pretrained VGG. The network architectures without batch normalization and subsampling yielded similar results.

3.2. Fine-tuning higher layers of retinal-VGG16 on ImageNet

We fixed the early retinal layers of the network and trained the remaining layers, both those involved in reshaping the images and those involved in VGG. We were first interested to see whether this model would learn at all, that is whether the model's loss function would decrease over time. Fig. 5 shows that the retinal-VGG model's loss does indeed steadily decrease over time, suggesting that the model is improving in its ability to classify images.

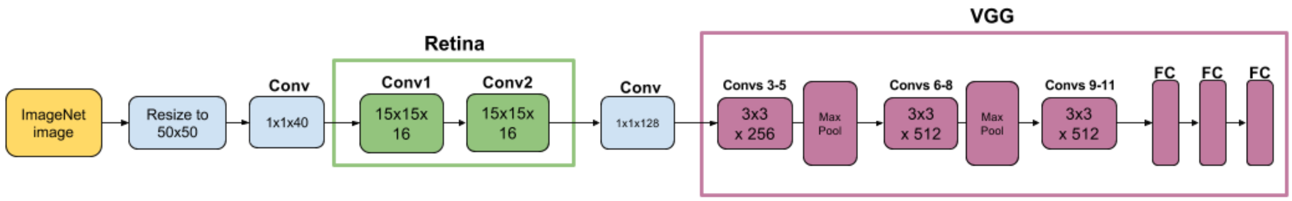

Figure 2. Diagram of architecture with early VGG16 layers replaced with CNN retina model

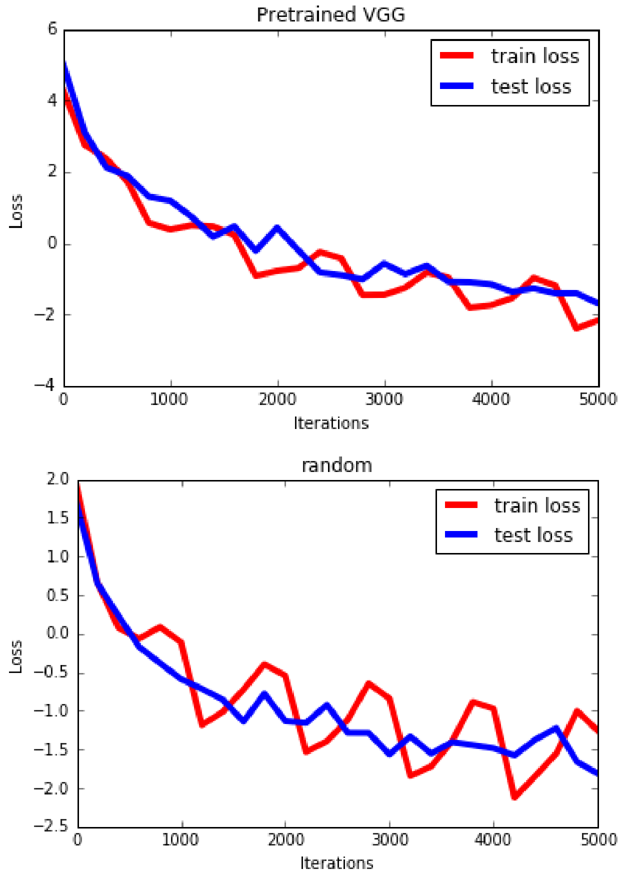

Figure 3. Training and Validation loss curves for pretrained VGG weights (**top) and random weights (bottom).**

3.3. Evaluating fit of retinal-VGG16 to macaque V4 and IT data

We evaluated the fit of the retina-VGG16 model to macaque IT data by computing a representational dissimilarity matrix (RDM) for each layer and comparing it to that of IT. Fig 6 shows the RDMs for four different layers in the retinal-VGG model: conv4_1, conv4_2, conv5_1, and fc8. The layer with the highest spearman correlation coefficient is conv5_1 at 0.57.

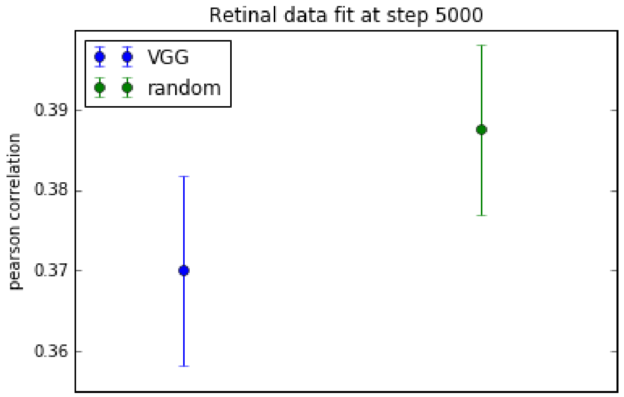

Figure 4. Comparison of mean correlation for 5 RGCS for pretrained VGG weights and random weights.

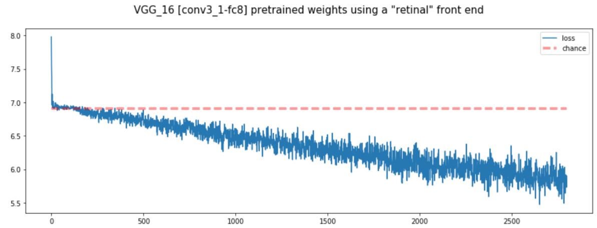

Figure 5. Loss function over time for the Retinal-VGG model.

We can also compare the performance of the retinal-VGG to the unaltered VGG16 network. Fig 7 shows the comparison across layers of retinal-vgg to VGG16. VGG16, plotted in red, outperforms the retinal-vgg model at predicting IT data for all layers that we computed. However, an important caveat to this result is that the VGG16 model had converged in its loss function whereas the retinal-VGG model was still decreasing its loss at the time of writing this report, suggesting perhaps that it might have closed this gap had we trained it for longer.

4. Discussion

While previous work has shown correspondences between task optimized CNNs and regions along the ventral visual pathway, questions about where, if at all, retinal responses are best predicted in hierarchical deep neural networks. We address this problem by evaluating how well VGG16 weights pretrained on Imagenet capture RGC re-

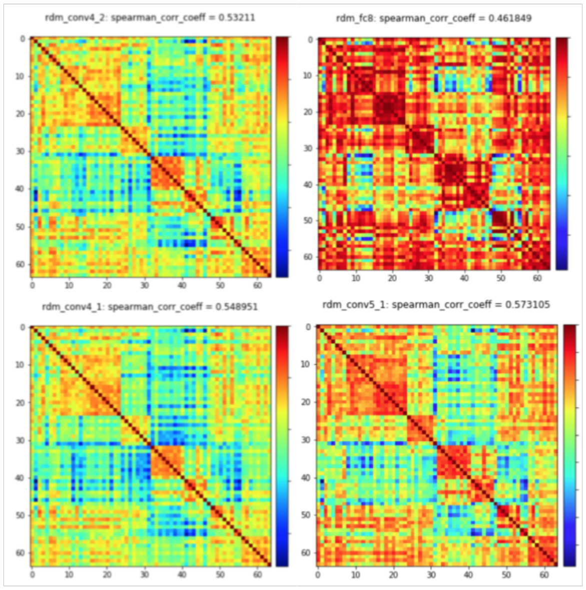

Figure 6. RDMs for different layers in the retinal-VGG model.

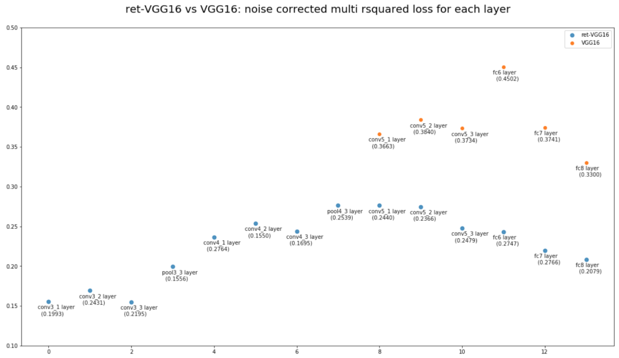

Figure 7. Comparison of the spearman correlation coefficient between VGG-16 (red) and retinal-VGG(blue) across multiple layers

sponses. Although we do not see an improvement compared to random weights, more sophisticated methods of predicting retinal responses with the pretrained weights may provide different results. It is also a possibility that a stronger task than image classification on Imagenet is required to constrain very early visual processing. Furthermore, we demonstrate that a convolutional neural network model that receives retinal representations as inputs rather than images can learn the representation of cortical area IT with high accuracy. However, this model does not outperform the unaltered VGG16 model at predicting IT responses, though more training may have helped to close the gap between these two models, as the retinal-VGG model had not yet converged at its maximum task performance at the time of writing this report. In conclusion, We demonstrate a novel approach for understanding the impact of early retinal processing in a hierarchical visual pathway.

References

[1] S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *bioRxiv*, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[3] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008.

[4] D. J. Margolis, G. Newkirk, T. Euler, and P. B. Detwiler. Functional stability of retinal ganglion cells after degeneration-induced changes in synaptic input. *Journal of Neuroscience*, 28(25):6526–6536, 2008.

[5] L. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. 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, editors, *Advances in Neural Information Processing Systems 29*, pages 1369–1377. Curran Associates, Inc., 2016.

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[7] 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*, 216(1205):427–459, 1982.

[8] 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*, 111(23):8619–8624, 2014.