Do Biologically-Realistic Recurrent Architectures Produce Biologically-Realistic Models?

Grace W. Lindsay (gracewlindsay@gmail.com)
Theodore H. Moskovitz (thm2118@columbia.edu)
Guangyu Robert Yang (gyyang.neuro@gmail.com)
and Kenneth D. Miller (kendmiller@gmail.com)
Center for Theoretical Neuroscience, Columbia University
New York, NY USA

Abstract

Many details are known about microcircuitry in visual cortices. For example, neurons have supralinear activation functions, they're either excitatory (E) or inhibitory (I), connection strengths fall off with distance, and the output cells of an area are excitatory. This circuitry is important as it's believed to support core functions such as normalization and surround suppression. Yet, multi-area models of the visual processing stream don't usually include these details. Here, we introduce known-features of recurrent processing into the architecture of a convolutional neural network and observe how connectivity and activity change as a result. We find that certain E-I differences found in data emerge in the models, though the details depend on which architectural elements are included. We also compare the representations learned by these models to data, and perform analyses on the learned weight structures to assess the nature of the neural interactions.

Keywords: convolutional neural networks; visual cortex; excitation; inhibition; recurrence

Background

The visual processing stream is a hierarchy composed of brain regions each with their own recurrent connectivity. This recurrent connectivity is believed to implement many important functions such as normalization and surround suppression. In previous work (Rubin, Van Hooser, & Miller, 2015), a model was built based on the architecture of primary visual cortex that can implement these functions. This model (the stabilized supralinear network, or SSN) includes several features found in the visual cortex of mammals, such as neurons with firing rates that are a supralinear function of their input and connection strengths that depend on the similarity between preferred stimuli. Of particular relevance to this study is the "ring" version of the SSN, which is a series of excitatory-inhibitory cell pairs which all represent the same spatial location but have different preferred features. This network implements "cross-feature" normalization: the response to two simultaneously presented stimuli is less than the sum of the responses to each stimulus presented individually.

Convolutional neural networks are currently some of the best models available for capturing the transformations performed by the visual processing stream (Yamins et al., 2014). Yet these models do not usually include any recurrent processing, and lack many of the features found in biology.

Here we successively add different biological details of recurrent processing to a standard convolutional architecture, and measure the extent to which these additions make the model a better match to data, evaluated along several different axes of variation.

Methods

Network Architectures

The base architecture was inspired by AlexNet and contained 5 convolutional layers and 3-fully connected layers. Recurrent connections were included at the fifth convolutional layer. The recurrence was run for 11 time steps, with the last time step used to calculate classification performance.

The recurrence was convolutional with filter size 3x3 (for reference, feature maps at layer 5 are 8x8). In the "notEI" network there were no constraints on recurrent weights. In EI networks, half of the channels were excitatory and half inhibitory. The recurrent filters applied to the E channels were restricted to contain only non-negative values and those of I channels could contain only non-positive values. This was enforced by applying an absolute value function to the weight matrices (followed by multiplication by -1 for inhibitory channels).

Each cell in the recurrent layers had a response (r) that evolved according to this equation, adapted from (Rubin et al., 2015):

$$\tau_i \frac{dr_i}{dt} = -r_i + ([I_i]_+)^n \quad (1)$$

where I_i is the sum of the feedforward input coming to neuron i from the layer below, which is constrained to be non-negative, and its recurrent input. In supralinear networks, $n = 1.8$. In "linear" networks, $n = 1$ (note that the function is still nonlinear, as it is rectified). In all models, $dt = 2.0$ms and τ_i was 20ms for E cells and 10 ms for I cells.

In a standard EI network, the depth of the recurrent convolutional filters is equal to the number of channels (256). In other words, the recurrent connectivity is all-to-all in feature space. We also studied networks with distance-dependent connections in feature space. In this case, we took the channels to be arranged on a ring of 128 nodes, with an E channel and an I channel at each node (or simply two channels in the 'notEI' networks). The weights from a given channel to chan-

nels more than x nodes away (here $x = 10$ or 20) were set to 0.

In some networks, only the excitatory cells served as output to the next layer of the network. In most, all did.

In most but not all networks, the contrast of the input images (the range of positive/negative values about the mean of each color channel) was scaled by a random value between .5 and 5. This was done to encourage the network to learn to do normalization.

For the "SSN" model, the recurrent weights were not learned, but held constant at the values used in (Rubin et al., 2015). Specifically, the recurrent connections were such that a ring network was placed at each spatial location (recurrent spatial filters in this network were 1x1) to implement cross-feature normalization. Only excitatory neurons served as output in this model, and the image contrasts were varied during training as above.

Networks were trained via stochastic gradient descent on the ImageNet dataset.

Analysis Methods

Comparing to Data Features We found in the literature several measures of excitatory and inhibitory activity, along with other findings relating to recurrent processing, that we wished to compare to our models. We tested whether each of our models is a qualitative match to each of these findings. The statistical significance of these findings were tested via a 2-sample t-test. For analyses of the "notEI" network, one feature map at each node was treated as excitatory and the other as inhibitory, despite these distinctions not existing in this network.

Comparing Representations We use the pre-established method of representational similarity analysis (RSA, kriegeskorte2008representational) to compare model representations to each other and to V4 data.

Non-normality and E-I Interaction We used several methods to characterize the connectivity patterns of the excitatory and inhibitory recurrent weights. First, we summed over the spatial dimensions of the recurrent weight tensor to obtain a $C \times C$ (where C is the number of feature maps) block matrix

$$W = \begin{pmatrix} W_{EE} & W_{EI} \\ W_{IE} & W_{II} \end{pmatrix}, \tag{2}$$

where W_{XY} contains connections from neuron type Y to X. We studied their interactions via the Schur decomposition, which allows us to write

$$W = U \Lambda U^{-1}, \tag{3}$$

where U is a unitary transformation matrix and Λ is an upper triangular matrix whose diagonal elements are the eigenvalues of W. Significant strength in the interaction terms W_{EI} and W_{IE} renders W strongly non-normal (Murphy & Miller, 2009), i.e. its eigenvectors deviate strongly from orthogonality, and this produces the effective feedforward weights between activity patterns in the upper triangular part of Λ. Because the summed absolute square of the matrix entries, as well as the eigenvalues, are preserved under unitary transformations, the summed absolute square of the feedforward weights of Λ, relative to the summed absolute square of all the weights, is a unitary invariant representing the relative strength of the feedforward weights, and is one scalar measure of the degree of non-normality of W. This is computed as

$$\rho = \frac{||W||_F^2 - \sum_i |\lambda_i|^2}{||W||_F^2}, \tag{4}$$

where $||\cdot||_F$ denotes the Frobenius norm and $\{\lambda_i\}_{i=1}^C$ are the eigenvalues of W (Trefethen & Embree, 2005). If W is normal, then Λ is exactly diagonal and $\rho = 0$.

The second measure we computed is derived from the Schur vectors stored in the unitary transformation matrix U. Due to the block structure of W, the first $C/2$ entries of each Schur vector correspond to excitatory weights, while the second $C/2$ entries correspond to inhibitory weights. A weight matrix with evenly mixed interactions both within and between cell types would therefore be expected to have relatively equal weightings in each half of its Schur vectors. This balance can be quantified by the *Schur ratio* r_S:

$$r_S = \frac{1}{C} \sum_{j=1}^{C} \min \left\{ \frac{\sum_{i=1}^{C/2} |U_{ij}|^2}{\sum_{k=C/2+1}^{C} |U_{kj}|^2}, \frac{\sum_{k=C/2+1}^{C} |U_{kj}|^2}{\sum_{i=1}^{C/2} |U_{ij}|^2} \right\}, \tag{5}$$

with higher values of r_S associated with a greater degree of E-I balance.

Results

Once trained, the networks all displayed similar top-1 performance accuracy (ranging from 46.5-47.1%), with the exception of the SSN network, which only reached 39.5%. Though enhancing performance is not our immediate goal, performance is correlated with ability of a model to match data representations (Yamins et al., 2014).

Testing the Model Recurrence for Data Features

Inspired by findings in the literature, we asked whether certain features regarding the activity of the neurons were present in our recurrent circuit. These included: Is mean I firing higher than mean E firing ("FR: $I > E$")? Are I↔I correlations higher than E↔E correlations ("Cor: $II > EE$")? Do correlations decrease over time during the response ("Cor: $Ear > Late$", (Maor, Shalev, & Mizrahi, 2016))? Are the outputs of a cell more strongly tuned than the inputs ("Tune: $O > I$", (Liu, Wu, Arbuckle, Tao, & Zhang, 2007))? Are E cells more strongly tuned than I cells? ("Tune: $E > I$", (Kerlin, Andermann, Berezovskii, & Reid, 2010)) Do the cells perform normalization (measured as the percent of cells performing sublinear summation, "Sublin: %")? Is the strength of E projections to the spatial surround stronger than I ("Sur: $E > I$", (Hirsch & Gilbert, 1991))? Whether or not each of our models matched the data on these measures can be found in Table 1. The measures of correlations were particularly helpful in discriminating between models. The networks which aligned with the

Table 1: Model Comparisons to Data Features. Model Key: l = linear, nl = supralinear, VC = contrast varied, H10 or H20 = distance-dependent connections restricted to distance 10 or 20, EO = output is excitatory cells. Data features described in Results. If the comparison was statistically significant ($p < .05$) and matched the data, the p-value is listed, otherwise an 'N' is present.

Model	FR: $I > E$	Cor: $II > EE$	Cor: $Ear > Late$	Tune: $O > I$	Tune: $E > I$	Sublin: %	Sur: $E > I$
l_notEI_VC	N	N	$2.7e-9$	$2.3e-104$	N	19.3	N
l_VC	$1.4e-61$	N	$1.8e-5$	$3.4e-104$	.0090	46.3	N
lH10_VC	N	.027	N	.026	$8.1e-6$	26.4	.0036
lH20	$2.7e-12$	N	N	$2.6e-56$	$1.3e-10$	20.7	.022
lH20_VC	$4.6e-29$	.0015	N	$1.1e-58$	$2.6e-4$	19.0	.024
lH20EO_VC	0	$2.1e-55$	$1.1e-47$	$3.5e-72$	$8.1e-29$	58.9	N
nl_VC	$6.3e-133$	N	.011	$7.5e-167$	$1.3e-5$	79.4	N
nlH10_VC	0	N	N	$6.6e-151$	$2.9e-83$	18.0	N
nlH20	0	N	.044	$8.4e-177$	$4.8e-32$	23.4	N
nlH20_VC	0	N	$2.6e-4$	$4.6e-171$	$4.2e-25$	37.3	N
nlH20EO_VC	$4.3e-263$	$1.4e-13$	$1.9e-4$	$1.4e-146$	$6.2e-07$	50.5	N
SSN	0	0	$3.8e-308$	$2.0e-42$	N	82.0	-

most features of biological data were the two models wherein only the excitatory cells formed the output cells of the layer. The network with the SSN explicitly included also performs well, and has the highest percent of cells displaying sublinear summation.

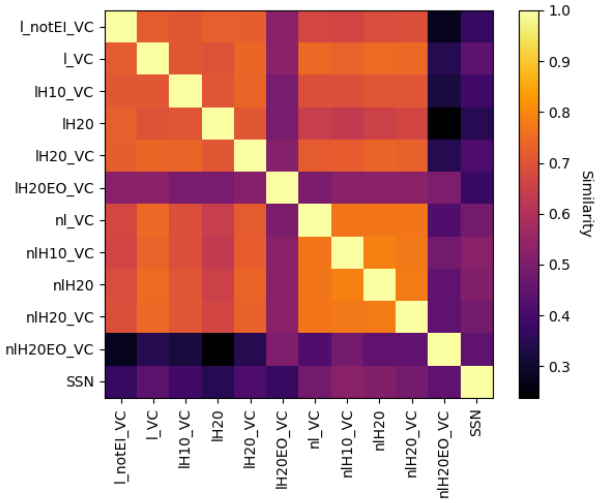

Figure 1: RSA similarity between models (correlation of RDMs) at the final time point. Model name key in Table 1

Comparing Representations

In addition to the data features discussed above, we also explored how architectural features impact the representations these networks learn. We used RSA to compare representations elicited in response to 256 ImageNet images (Figure 1). Two of the most biologically-realistic models (nlH20EO_VC and SSN) have representations that are very different from the other models (and also each other). extent, as to a lesser extent did the linear excitatory-output network (lH20EO_VC). We also compared representations in the model to cortical data. We used RSA to compare model responses to the response of a population of V4 neurons to the same 64 object images (Yamins et al., 2014). Because the recurrence in our model introduces temporal dynamics, we are able to compare point-for-point the model, which had 11 time steps, to the data binned into 11 20ms bins. (Figure 2). The SSN model has one of the worst matches to the data, which may be a result of its low performance, rather than any aspect of the SSN per se. It will also need to be explored if a different mapping between the temporal dynamics of the model and that of the data would result in better matches. Interestingly, the linear models, which are less biologically accurate, have some of the best fits to the data.

Analyzing Connectivity

Results for the weight matrix non-normality and Schur ratios were plotted against each other in Figure 3. There is a relatively strong correlation between the two measures and several trends are evident.

First, supralinear activation functions on average promote both stronger E→I and I→E interactivity and more evenly mixed interactions both within and between cell types. Second, restricting the output cells to only be excitatory has the strongest effect in increasing both the Schur ratio and the non-normality. Third, training with variable image contrasts appears to be a necessary but insufficient criterion to maximize these measures. Table 1 demonstrates that training with this transformation in the supralinear network does indeed increase the model's ability to perform sublinear summation. This suggests that connectivity between E and I cells in biological networks may be a function of the characteris-

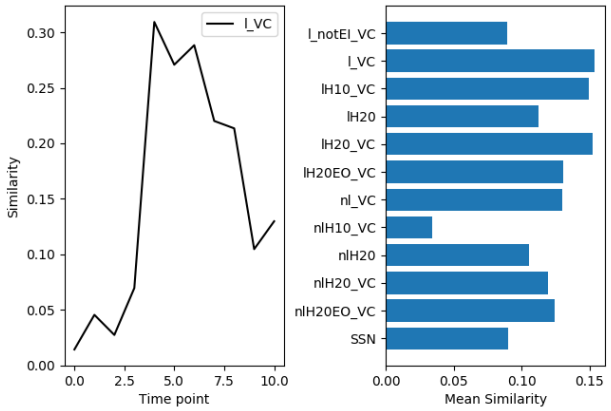

Figure 2: RSA between models and V4 data. Time point for time point comparison for the best performing model shown on the left; similarity averaged over time for all models on right.

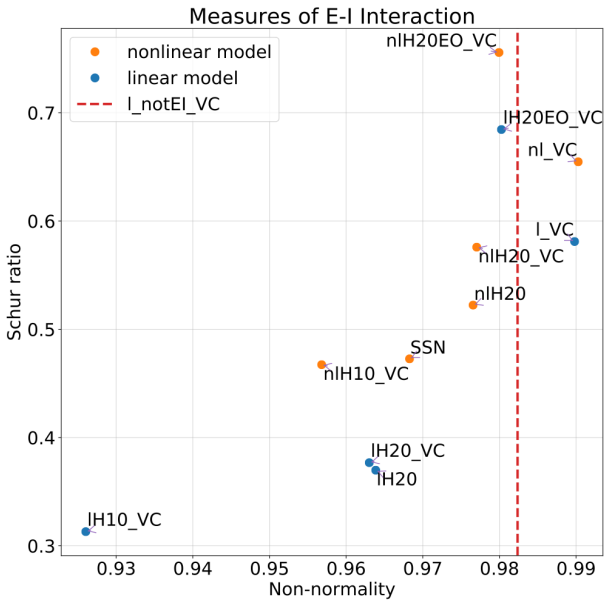

Figure 3: Scatter plot of two different measures of interaction in the weight matrices. For the "notEI" network, the Schur ratio value is not well-defined, as it has no restricted E-I cells.

tics of sensory inputs as well. Surprisingly, the SSN (the most biologically-plausible model) scores in the bottom half for both the Schur ratio and the non-normality. Conversely, the notEI model (the least biologically plausible) has one of the highest non-normality values.

Conclusions

Here we show that it is possible to incorporate more biologically-realistic details, in the form of recurrent connections, into a standard convolutional neural network architecture. This has the benefit of merging traditional single area computational models, which can replicate details of neural circuitry and activity statistics, with hierarchical multi-area models that can perform visual tasks and predict neural activity. In doing so, we show that certain architectural features—such as only allowing excitatory cells to be output cells—help replicate findings from the data and lead to different types of image representations. The architectural features that provide these benefits do not, however, necessarily make the image representations in the model more similar to that of V4 data. Reconciling these differences will be important.

Acknowledgments

We thank Dan Yamins and members of his lab, particularly Aran Nayebi and Daniel Bear, for assisting with model training code and providing data. We also thank Minni Sun, Li Ji-An, and Mario Dipoppa for helpful conversations. Funding sources: NSF DBI-1707398 and IIS-1704938, NIH T32 NS064929, and the Gatsby Charitable Foundation.

References

Hirsch, J. A., & Gilbert, C. D. (1991). Synaptic physiology of horizontal connections in the cat's visual cortex. *Journal of Neuroscience*, *11*(6), 1800–1809.

Kerlin, A. M., Andermann, M. L., Berezovskii, V. K., & Reid, R. C. (2010). Broadly tuned response properties of diverse inhibitory neuron subtypes in mouse visual cortex. *Neuron*, *67*(5), 858–871.

Liu, B.-h., Wu, G. K., Arbuckle, R., Tao, H. W., & Zhang, L. I. (2007). Defining cortical frequency tuning with recurrent excitatory circuitry. *Nature neuroscience*, *10*(12), 1594.

Maor, I., Shalev, A., & Mizrahi, A. (2016). Distinct spatiotemporal response properties of excitatory versus inhibitory neurons in the mouse auditory cortex. *Cerebral Cortex*, *26*(11), 4242–4252.

Murphy, B. K., & Miller, K. D. (2009). Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, *61*, 635-648.

Rubin, D. B., Van Hooser, S. D., & Miller, K. D. (2015). The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, *85*(2), 402–417.

Trefethen, L. N., & Embree, M. (2005). Scalar measures of nonnormality. In *Spectra and pseudospectra: the behavior of nonnormal matrices and operators* (p. 442-447). Princeton University Press.

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