Convolutional recurrent neural network models of dynamics in higher visual cortex

Aran Nayebi*, Jonas Kubilius*, Daniel Bear, Surya Ganguli, James J. DiCarlo, Daniel L. K. Yamins

Neurons in the ventral visual pathway exhibit behaviorally relevant temporal dynamics during image viewing. However, the most accurate existing computational models of this system are feedforward hierarchical convolutional neural networks (HCNNs), which capture neurons' time-averaged responses, but do not account well for their complex temporal trajectories. Here we show that HCNNs augmented with both local and global recurrent connections are quantitatively accurate models of dynamics in higher visual cortex.

We began with a five-layer HCNN that achieved state-of-the-art predictions of temporally-averaged visual responses in macaque V4 and IT neurons. To model within-area dynamics, we replaced units in each layer with one of several local recurrent circuit motifs, including simple Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory (LSTM) units. We also included combinations of global feedback connections, in which outputs of later convolutional layers were added to inputs of earlier layers. Using backpropagation through time, these new parameters were optimized to predict V4 and IT neural response patterns. Finally, we tested these networks' ability to predict responses on held-out images and neurons not used for model optimization.

We found that the best network structure led to substantial improvements over the feedforward baseline, explaining close to 100% of the explainable variance in V4 neurons and above 75% in IT neurons on average across time points. This network made use of gated local recurrence, with LSTMs and GRUs proving superior to simple RNNs. Furthermore, the presence of specific global feedback connections in this network was critical for best predicting V4 neuron dynamics. In summary, we have developed a deep recurrent neural network architecture that accurately captures temporal dynamics in several ventral cortical areas, opening the door to more detailed computational study of the circuit structures underlying complex visual behaviors.

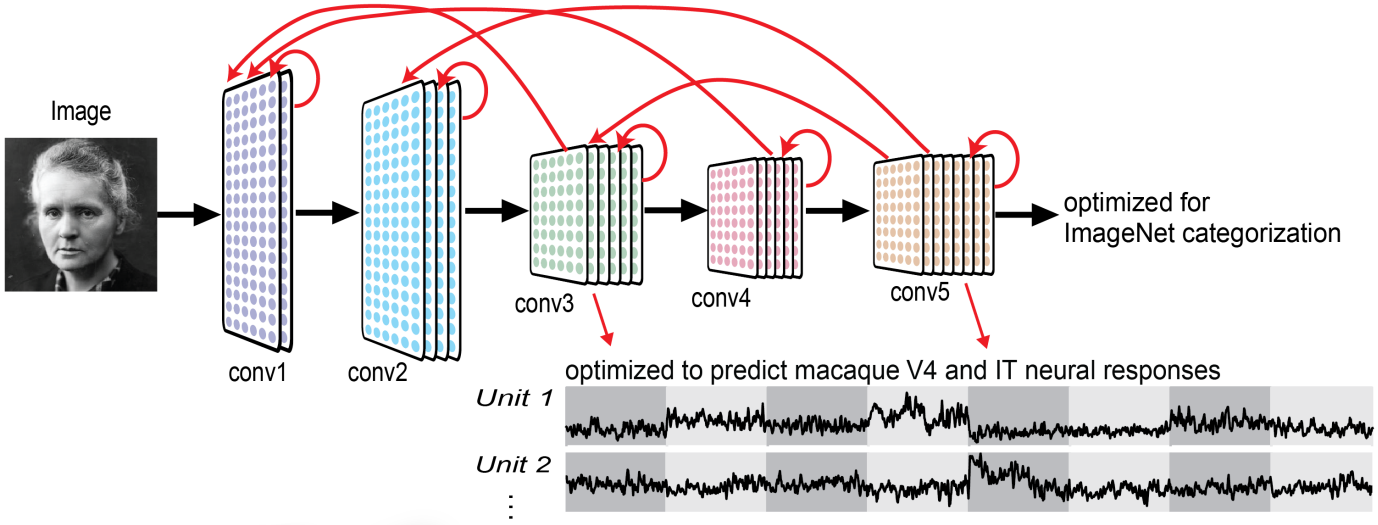

Figure 1. Training convolutional recurrent neural networks to predict temporal dynamics in the ventral visual pathway. First, feedforward weights (black arrows) of a fixed HCNN with five convolutional layers and a fully connected softmax readout layer were optimized to maximize object categorization on the ImageNet 2012 ILSVRC dataset. Then, local recurrent and global feedback connections were introduced. Local recurrent circuits are convolutional, i.e. share weights between spatial locations. For global feedback connections, outputs from source layers are spatially resized to align with the target layer and concatenated along the channel dimension. Recurrent and feedback parameters were then optimized to match dynamic trajectories of V4 and IT responses in 20 ms time bins (Majaj *et al.* 2015). V4 neurons were linearly regressed from the conv3 layer, while IT neurons were linearly regressed from the conv5 layer (see Fig. 2). A variety of model structures were tested, including different local recurrent motifs, e.g. simple RNNs, convolutional GRUs (cGRUs), and convolutional LSTMs (cLSTMs), as well as various patterns of global feedback. All models were evaluated on held-out images not seen during optimization. The highest performing model had cLSTM recurrences at every layer and a global feedback connection from conv5 to conv2.

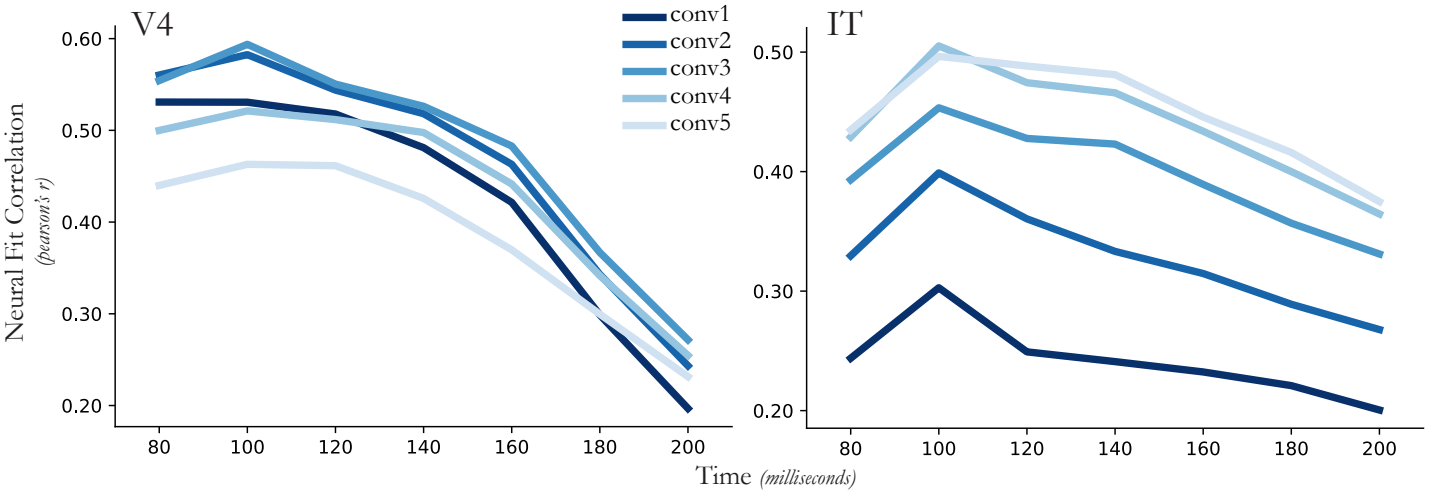

Figure 2. Finding the optimal layers for V4 and IT neural fitting. We first determined which model layers mapped best to V4 and IT neural data. Using Partial Least Squares (PLS) regression with 25 components, we separately fit each layer of the feedforward HCNN to each time bin of neural data. V4 and IT neurons were best fit by the conv3 and conv5 layers, respectively. All subsequent fitting used these optimal mappings.

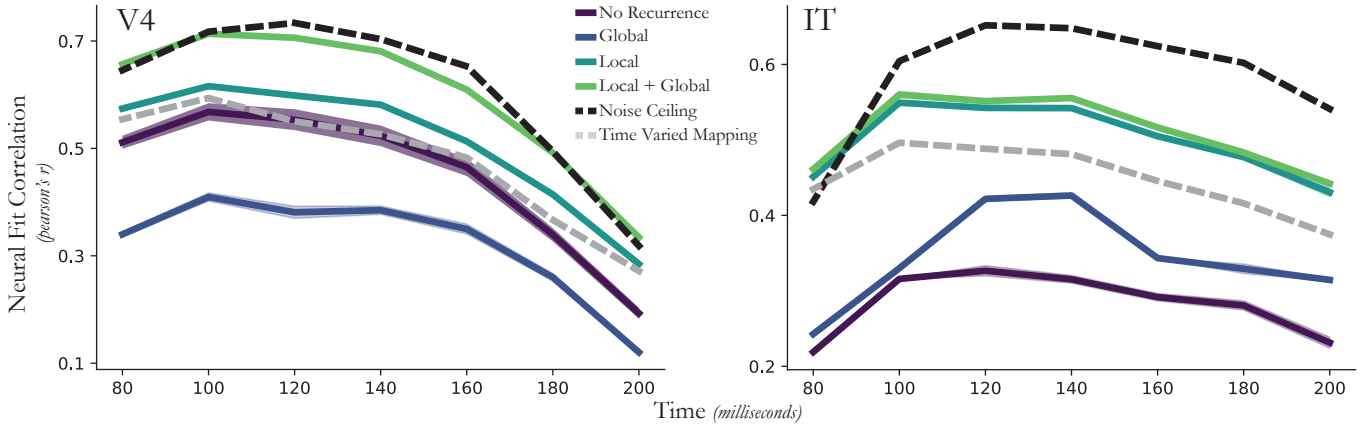

Figure 3. Both local recurrence and global feedbacks are needed to best fit neural data. Among a wide range of architectures with different local recurrent motifs and global feedback patterns, the best architecture was one with both gated local recurrence and a global feedback from conv5 to conv2. Local recurrent circuits were particularly useful for improving fits to IT neurons, whereas both local recurrence and global feedback were critical for improving fits to V4 neurons. All architectures were fit to V4 and IT neurons simultaneously. Except for "time-varied mapping," fixed model-unit-to-neuron linear mappings were shared across all time bins, constraining trajectories to be produced by actual dynamics of the network. In contrast, "time-varied mapping" indicates an independent PLS regression for each time bin. The fact that models with local recurrence and global feedbacks are better than "time-varied mapping" suggests that some nonlinear dynamics at earlier layers contributed meaningfully to network fits. SEM across four splits of held-out test images.

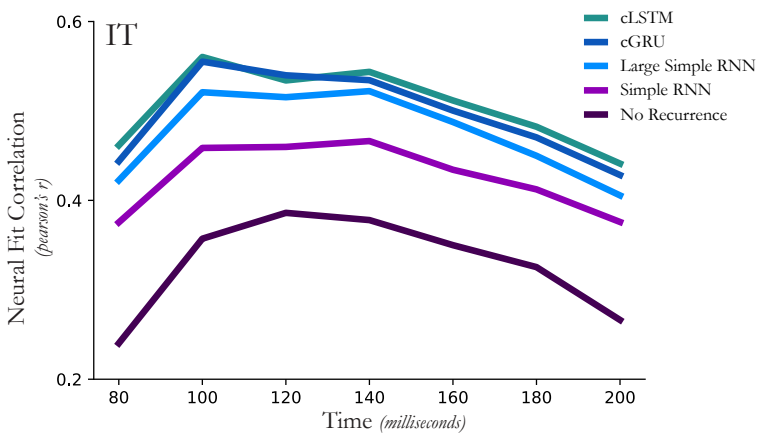

Figure 4. Local gated recurrence best fits IT responses. Models with gated local recurrent motifs (cLSTM, cGRU) matched neural data better than simpler RNN motifs. Large Simple RNN matched number of parameters to cLSTM, while Simple RNN matched the number of units in each layer. Consistent with its more elaborate gating structure, the cLSTM motif slightly outperformed the cGRU motif.

Majaj N, Hong H, Solomon EA, & DiCarlo JJ (2015). *Journal of Neuroscience* 35(39):13402-13418.