Understanding neural representations in early visual areas using convolutional neural networks

Yimeng Zhang*, Corentin Massot*, Tiancheng Zhi, George Papandreou, Alan Yuille, Tai Sing Lee

Summary We compared the neural representations of population of macaque V1 and V2 neurons in response to visual stimuli of different degrees of complexity with model representations of AlexNet (a convolutional neural network) and several models of V1 and V2. We found that AlexNet matched the neural data better than others, particularly for complex stimuli. Our analysis showed that it matched better because of (1) normalization and nonlinear pooling mechanism, and (2) more complex and diverse stimulus preference. Among all layers of AlexNet, the first pooling layer matched V1 neurons the best in terms of individual tuning as well as population representation, and the second pooling layer matched V2 neurons the best. We visualized the preferred features of the units corresponding to real neurons using deconvolutional networks and found that some V1 neurons could not be described by simple Gabor filters, and that some V2 neurons appeared to encode more complex surface structures such as textures. These findings suggest that V1 and V2 neurons might have more complex codes than previously thought.

Methods We recorded the responses of 286 V1 and 390 V2 neurons in 2 monkeys (K, F) to 150 stimuli using multi-electrode arrays. The 150 stimuli are divided into 3 subsets of 50, denoted Edge (E), Appearance (A), and Exemplar (EX). These stimulus subsets are designed to represent the same edge shapes (prototypes) in levels of increasing complexity. Edge stimuli are the most simple and artificial while Exemplar stimuli are the most complex and natural, and Appearance stimuli are in between. Fig. 1 shows the stimulus set, and the average spike count responses of an example neuron.

The neural representation $\phi(x_i)$ of one stimulus x_i is a vector in a high dimensional space, where each dimension is the average response of a neuron to this stimulus. Given a neural representation $\phi(x)$ of a stimulus set x, we computed the neural representational dissimilarity matrix (RDM, Kriegeskorte et al. (2008)) RDM($\phi(x)$) as RDM$(\phi(x))_{ij} = 1 - \rho(\phi(x_i), \phi(x_j))$, where $\phi(x_i)$, $\phi(x_j)$ are the neural representations of stimuli i and j, and $\rho(\cdot, \cdot)$ is the Pearson's correlation coefficient between them. The neural RDM captures the distance between every pair of stimuli in the neural response space. Similarly, we computed a model RDM for each model based on its units' responses (feature vectors) to the same stimulus set. For each stimulus subset we evaluated the similarity between neural and model representations using the Spearman rank correlation between the corresponding upper-triangular, non-diagonal elements of their RDMs. For CNN models of different layers, we performed statistical tests with bootstrap resampling of the stimulus set as in Khaligh-Razavi & Kriegeskorte (2014) to identify the layers that were not significantly different from the best matched layer.

We tested a number of models, including the CNN model "AlexNet" (Krizhevsky et al. 2012), and several models for V1 and V2, including Gabor filter-based models (V1like) and overcomplete sparse coding models (LCA) (Olshausen 2013; Pinto et al. 2008; Rozell et al. 2008) as well as hierarchical sparse coding models (Sparse DBN) (Lee et al. 2008) for comparison. We optimized the hyperparameters such as stimulus size and number of units to get the best results for different models. Our goal of comparing models was not to show that AlexNet (or CNNs in general) is the best model, but to find out key features in representations and mechanisms which might contribute to a better match with neural data.

To better understand a real neuron's code we visualized its best matching AlexNet units. To examine the preferred stimuli of the real neuron, for each stimulus subset, we picked out the AlexNet unit having highest response correlation with the neuron on that subset, in a specific layer, and then visualized the three best matching units (one for each subset) in this layer using deconvolutional networks (Zeiler & Fergus 2013) for comparison. Due to the limited size of our stimulus set, there is no guarantee that we have identified the correct AlexNet unit(s) for real neurons. But the consistency of visualizations across the three subsets could provide confidence on whether the AlexNet units give us a glimpse of the true code of that neuron.

Results Fig. 2 shows the performance of different models in terms of their RDM correlations with the neural data. The dashed and solid lines are lower and upper estimates of the performance that an ideal model can achieve given the variation between monkeys, using the method in Khaligh-Razavi & Kriegeskorte (2014). AlexNet reached or crossed the lower bounds, showing that it matched well with neural data. The non-AlexNet models were optimized over a large number of hyperparameters separately for each of 2x3 panels and their correlations were lower, particularly for the EX stimuli. For AlexNet, we show both the best performance among 12 models for all layers (purple), as well as the pool1 and pool2 models for V1 and V2 (red), using a single set of hyperparameters for all panels. We single out pool1 and pool2 models because for V1, only pool1 was identified to be statistically the same as the best layers on all subsets, and for V2, pool2 and conv3 were identified using the same criteria. Comparing the models provides insight into the key features in the models contributing to the match with neural data. The observation that

*equal contribution

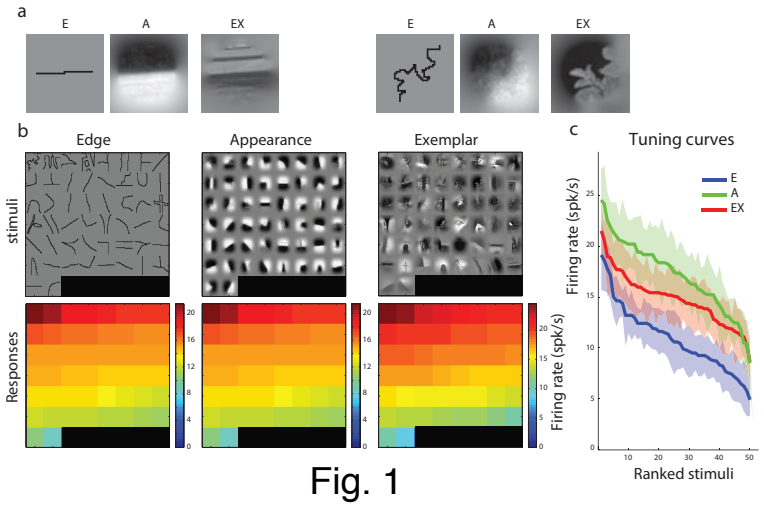


Fig. 1

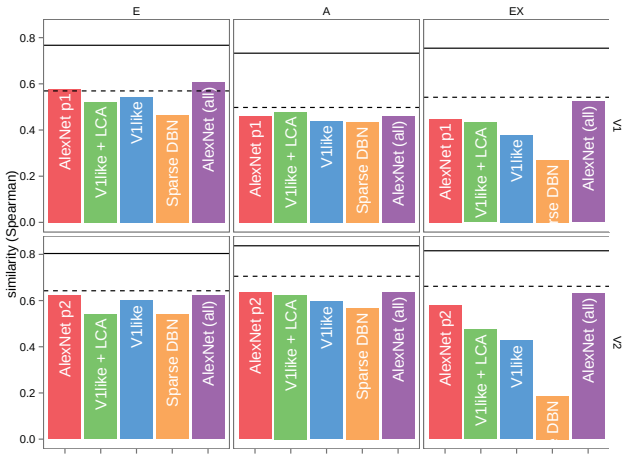


Fig. 2

Fig. 3

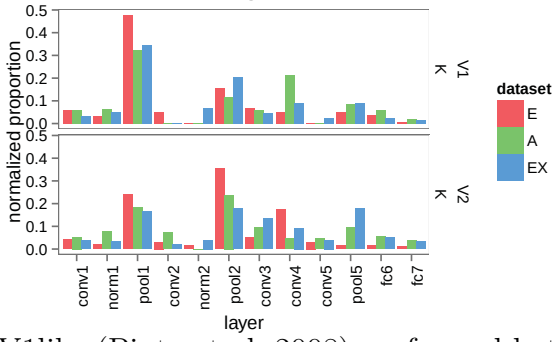
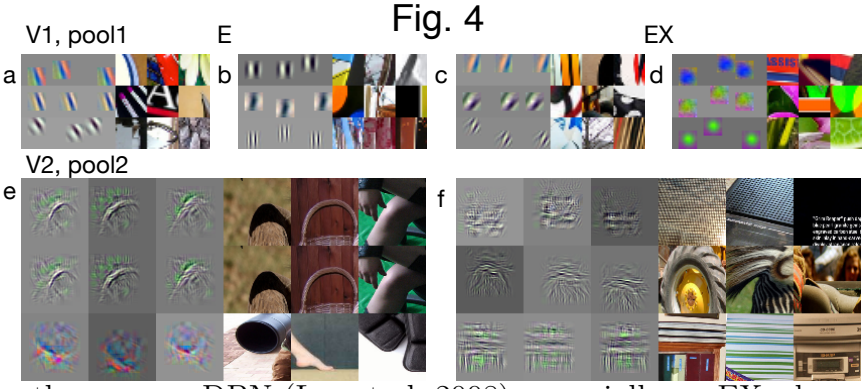


Fig. 4



V1like (Pinto et al. 2008) performed better than sparse DBN (Lee et al. 2008), especially on EX, shows that normalization and pooling (which are in V1like but not in Sparse DBN) are important, and in fact more important than having more complex features such as curves and corners (which are encoded in the second layer of the sparse DBN). The observation that V1like + LCA (V1like with Gabor filters replaced by highly over-complete sparse coding dictionaries learned by a Locally Competitive Algorithm (Rozell et al. 2008) in Olshausen (2013)) outperformed V1like implies that V1 neurons have a more diverse set of filter shapes than just Gabor. Finally, the observation that Alexnet's first pooling layer (p1) was better than LCA models indicates that some V1 neurons are likely more complex than filters learned by sparse coding.

Fig. 3 shows the distribution of the layers for AlexNet units having highest correlations with individual V1 and V2 neurons for monkey K (result similar for F, not shown), supporting the findings obtained with the RDM analysis: V1 matched best to AlexNet pool1 and V2 matched best to pool2. Complex stimuli (EX in green) tended to shift the distribution to higher layers compared to simple stimuli (E in red). V2 neurons were also more correlated to higher layer AlexNet units than V1 neurons.

Fig. 4 shows the visualization results of 4 V1 neurons (a-d) and 2 V2 neurons (e,f), using pool1 and pool2, respectively. Each panel has three rows, showing the visualized features of the AlexNet units having highest response correlations with the neuron on the three subsets (E, A, EX). On the left are shown the neurons that exhibited higher correlations with AlexNet units for simpler (E) stimuli. The preferred features showed a strong consistency as they were similar for units selected by three different stimulus subsets and were mostly like Gabor filters for V1 (a,b), and displayed large curves and contours for V2 (e). On the right are shown the neurons that exhibited higher correlations with AlexNet units for complex (EX) stimuli. The preferred features were not well described by Gabor filters for some V1 neurons (d), and appeared to encode texture or surface stimuli for V2 neurons (f). Hence the visualization results provide additional evidence for the complexity of the codes in V1 and V2.

Conclusion AlexNet provides a reasonably good match to V1 and V2 neural data. We show that pooling/normalization and the more diverse nature of codes of AlexNet units are key factors for this good match. Our findings suggest that V1 might be coding more complex features and V2 might encode more texture and surface structures as suggested by others. On the other hand, AlexNet is still far from the upper-bound of an "ideal model" (Fig. 2), suggesting it can be refined with additional biological constraints to produce a better match to V1 and V2.