Time-Resolved Correspondences Between a Deep Feed-Forward Neural Network and Human Object Processing: EEG Measurements

Nathan C. L. Kong (nclkong@stanford.edu)
Department of Psychology, Stanford University

Blair Kaneshiro (blairbo@stanford.edu)
Center for Computer Research in Music and Acoustics, Stanford University

Anthony M. Norcia (amnorcia@stanford.edu)
Department of Psychology, Stanford University

Abstract

The ventral visual system is known to exhibit hierarchical structure, where early and higher visual areas respond to simple and relatively complex features respectively. A clear, quantitative explanation for the image computations performed in each visual area is, however, lacking. Feed-forward hierarchical convolutional neural networks have been a step forward in attempting to model these computations. Here we model the temporal evolution of EEG responses recorded during passive viewing of multiple object categories using layers of a convolutional neural network trained to perform image categorization. We found a modest hierarchical correspondence between the depth of the layer in the neural network and the neural response time at which model and neural representations are maximally correlated. However, we show that shallow layer and deep layer representations start to correlate with neural representations at similar time bins. A reliability analysis indicated that the modest correspondences are far from the limit imposed by variability of the data, but are largely due to the inadequacies of the model. These results provide suggestive evidence that early visual areas perform more than just simple feature detection and that strictly feed-forward convolutional neural network models are insufficient to model human object processing dynamics.

Keywords: Object recognition; EEG; representational similarity analysis; convolutional neural network

Introduction

Object recognition is a complex task that humans perform effortlessly. In order to understand the mechanisms underlying this function, we need to be able to model the neural computations performed on an image. Recent success in leveraging the capabilities of large-scale compute power have enabled the training of machine learning models with millions of parameters, resulting in the development of hierarchical convolutional neural networks (CNNs) that when trained on a large image set surpass human object recognition capabilities (Simonyan & Zisserman, 2014).

In hierarchical CNNs, features extracted in the shallow parts of the neural network qualitatively resemble edge detectors and features extracted deeper in the network are much more complex. Artificial neurons in these models have been shown to correspond well to time-averaged single-unit recordings in visual areas of macaques (Cadena et al., 2019; Yamins et al., 2014). Furthermore, a human neuroimaging study has reported that there is a hierarchical correspondence in both space and time between these neural networks and the human ventral visual pathway (Cichy et al., 2016). In that study, the feature representation of shallow layers corresponded to early time points and that of deeper layers corresponded to later time points in the temporal dynamics of human object processing, measured using magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) data fusion. Furthermore, two fMRI studies have reported that shallow layers corresponded to early visual areas and deeper layers corresponded to higher visual areas (Cichy et al., 2016; Güçlü & van Gerven, 2015).

Here we seek to understand how well purely feed-forward CNNs map on to time-evolving neural data acquired using electroencephalography (EEG), which, like MEG, is known for its high temporal resolution. As in the previous MEG study (Cichy et al., 2016), we used representational similarity analysis (RSA, Kriegeskorte, Mur, and Bandettini (2008)) to compare how well representative layers of a CNN map on to different time points in the EEG data, hypothesizing that layers early in the CNN would map on to early time points in the neural data and that deeper layers in the CNN would map on to later time points in the neural data. We found that, depending on the similarity metric used in RSA, there was a positive relationship between the depth of the layer in the CNN and the time at which peak correlation occurred between the model and the data. On the other hand, we found no relationship between the onset time of correlations and the depth of the layer in the CNN for any of the metrics. These results only partially support the hypothesis of a hierarchical correspondence between CNNs and time-resolved object representations in humans.

Methods

The hierarchical nature of CNNs suggests that the following features should be observed in neural data. Firstly, the onset of correlations between model layer representations and neural representations should shift progressively as a function of layer depth. Secondly, the time of peak correlation should shift similarly.

In order to map stimulus representations from different layers of the CNN to EEG data, we used RSA as it provides a method to compare data from different modalities. In order to compare representations as a function of neural response latencies, we have used an open data set of high temporal resolution EEG data (Kaneshiro, Arnardóttir, Norcia, & Suppes, 2015) that was analyzed in a previous study (Kaneshiro, Perreau Guimaraes, et al., 2015). In that study, EEG responses were recorded from $N = 10$ humans passively viewing 72 images of objects from six different categories: human body (HB), human face (HF), animal body (AB), animal face (AF), fruit and vegetable (FV) and inanimate object (IO). After preprocessing and cleaning, the temporal resolution of the data was 62.5 Hz (16 ms). In order to reduce the effects of different levels of noise inherent to each electrode, multivariate noise normalization was performed on the data prior to downstream analyses (Guggenmos, Sterzer, & Cichy, 2018).

A key ingredient in RSA is the similarity metric that is used to create the representational dissimilarity matrix (RDM). Three different similarity metrics were used to compute the EEG RDMs: pairwise decoding accuracy (using linear discriminant analysis), cross-validated Euclidean distance and cross-validated Pearson correlation. These metrics have been described previously in Guggenmos et al. (2018). EEG RDMs were computed at each time bin, resulting in time-resolved RDMs that show the temporal evolution of similarities between stimuli.

The CNN model that was used in this study is VGG19 (Simonyan & Zisserman, 2014), which is a 19-layer neural network consisting of five convolutional and max pooling blocks, succeeded by three fully connected layers. It was trained to perform object categorization using the ImageNet (Deng et al., 2009) data set which consists of over one million images and 1000 image categories. The representative layers that we used in the model-data comparison were `pool1`, `pool2`, `pool3`, `pool4`, `pool5` and `fc2`, corresponding to each block in VGG19. Using the Pearson correlation similarity metric, an RDM was constructed at each of these layers, resulting in what we call “layer RDMs”. As the next step in RSA analysis, each of these layer RDMs was rank correlated with each time-resolved EEG RDM, allowing us to observe when EEG RDMs *start to correlate* with layer RDMs and when EEG RDMs are *most correlated* with particular layer RDMs.

In order to compute the time at which the onset of non-zero correlations occurs, a two segment piece-wise linear function was fit to the correlations from pre-stimulus onset time to 112 ms. The correlation onset time is defined to be the time bin at which the two line segments intersect. The time of peak correlation is defined to be the time bin at which the maximal correlation between the CNN layer representations and neural representations occurs. The average time bins and error on the mean for each of these cases were computed using 1000 bootstrap samples across subjects.

Results

Representational Dissimilarity Matrices

RDMs are measurements that show us the similarity of internal representations for pairs of stimuli. The RDM can be highly structured reflecting object category-level representation (Kriegeskorte et al., 2008). The category-level structure for our present analyses can be seen in Figure 1, which shows EEG RDMs for each of our three similarity metrics. Common to all three RDMs, averaged across time, is a degree of high-level categorical structure. This can be seen from the blocks along the diagonal of each matrix, indicating that objects within a particular category can be differentiated from objects in other categories, but not from objects within the same category. In particular, the EEG responses to human faces (HF) can be robustly differentiated from objects of other categories.

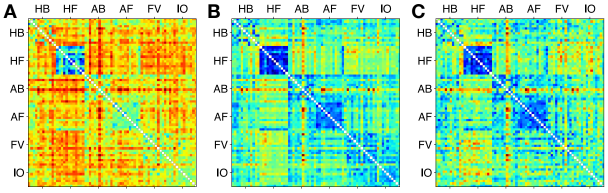

Figure 1: RDMs computed from the EEG data, averaged across all time bins. **A**: Decoding accuracy; **B**: Euclidean distance; **C**: Pearson correlation. Axis labels are the stimulus categories. Within each category, there are 12 images.

Correlation Time Courses

We obtained correlation time courses by computing the rank correlation between each layer RDM with the EEG RDMs computed at each time bin of the neural response. These time courses show us how the similarity between model representations and EEG representations evolves over time. Figure 2 shows the correlation time courses for the representative layers in the CNN where the EEG RDMs were computed using the decoding accuracy similarity metric. Qualitatively, we first observe that correlations start to increase at approximately the same time for all the layers ($60-70$ ms). Secondly, we note that the time at which the peak correlation occurs for each layer increases as the depth of the layer increases, reaching a maximum of approximately $150-160$ ms for deeper layers versus 120 ms for shallow layers. Finally, the noise ceiling (black line in Figure 2) computed using the EEG data is much higher than the correlations computed between VGG19 and the data. The first two observations will be quantified in the next section.

Comparison of Onset and Peak Correlation Times

We quantify the relationship between the depth of the layer in the CNN with the neural response time at which correlation onset occurs by fitting a two segment linear function on

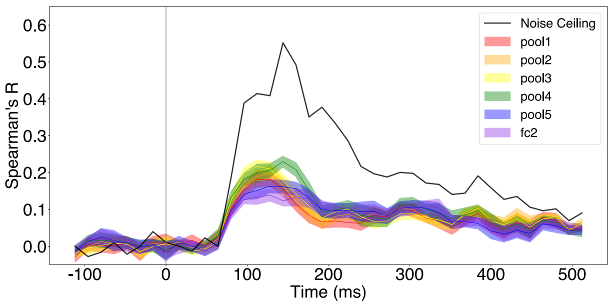

Figure 2: Correlation time courses computed for all the representative layers in the CNN. Solid lines indicate the mean correlation across participants and shaded regions indicate the standard error of the mean across participants. The grey vertical bar is the time of stimulus onset.

the correlation time course and determining the time of peak correlation, as described in the methods. Figure 3 visualizes how onset and peak time vary with the depth of the layer in the CNN. Each subplot corresponds to the specific similarity metric used to compute the EEG RDMs. Across each similarity metric, the slope of the line of best fit for the mean onset times as a function of layer depth is not significantly different from zero ($p > 0.05$), indicating that there is no relationship between onset time and layer depth. When decoding accuracy and Pearson correlation are used for the similarity metric, there is a positive relationship between the time of peak correlation and the depth of the layer in the CNN (Decoding: slope$= 6.3$, $p < 0.05$; Pearson: slope$= 11.3$, $p < 0.05$). However, when the Euclidean distance metric is used, the slope is not significantly different from zero.

Discussion

We provide partial support for the underlying hypothesis of a hierarchical correspondence between the CNN layer activations and the neural responses. For two of the three similarity metrics used to compute EEG RDMs, we found a positive relationship between the time of peak correlation with CNN layer depth. However, we did not find a relationship between the correlation onset times with layer depth for any of the similarity metrics.

Our results regarding the positive relationship between time of peak correlation and depth of CNN layer corroborate previous findings that there is some hierarchical correspondence in time between the complexity of features obtained from feedforward CNNs and neural responses (Cichy et al., 2016; Seeliger et al., 2017). Seeliger et al. (2017) show that with source reconstructed MEG responses, shallow layer features of the CNN predict sources at early time bins and deep layer features of the CNN predict more sources in later time bins. Their onset time analyses were based on a different criterion than we used: they used the time at which a highly significant correlation was present, whereas we measured time when correlation started to increase. Since correlation time courses were not shown, it is unclear whether their metric for onset time is more relevant to our metric for peak correlation or onset time.

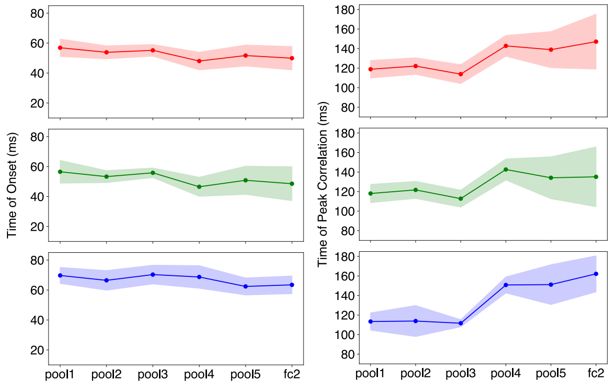

Figure 3: **Left**: Time at which correlation onset occurs. **Right**: Time at which peak correlation occurs. **Red**: Decoding accuracy. **Green**: Euclidean distance. **Blue**: Pearson correlation. These times are computed for each layer for each similarity metric used. Solid lines indicate the mean computed from 1000 bootstrap samples across participants and the shaded regions indicate the standard error of the mean computed from 1000 bootstrap samples across participants.

Cichy et al. (2016) reported that peak correlation time between MEG RDMs and layer RDMs occured at early time bins for shallow layers and at later time bins for deeper layers, similar to the observations from our analyses. A similar finding regarding onset times was reported in the supplementary material of Cichy et al. (2016) (cf. Suppl. Table 2), but not commented on. Fixed correlation onset times across layers is not compatible with the hypothesis of hierarchical correspondence since one expects an accumulation of synaptic delays as one ascends area by area over the ventral stream hierarchy. Taken together, the available analyses show that there is only partial support for the hierarchical correspondence between the CNN and the data.

Our observation that the RDMs obtained from the penultimate layer and from a shallow layer of the CNN start to correlate with EEG RDMs at a relatively early time bin is consistent with results of Yang, Tarr, Kass, and Aminoff (2018) and of Seeliger et al. (2017). In particular, Yang et al. (2018) show that features common between the shallowest and deepest layers of the CNN have some predictive power on source reconstructed MEG responses at early time points. Furthermore, Seeliger et al. (2017) show that mid to deep layer features of the CNN can also predict source reconstructed MEG responses at early time bins. These results and our observations suggest that aspects of the relatively complex features of deeper layers in CNNs may be computed early in time and presumably in early cortical areas. This suggests that feature processing in early visual cortex may be more complex than that suggested by shallow layers of CNNs.

Future work needs to address two issues that our work raises. Firstly, we note that the noise ceiling shown in Figure 2 is much higher than the achieved rank correlation values, indicating that the deficiency is not in the neural data, but in the model. This suggests that the use of a model incorporating temporal dynamics in object processing may correlate better with these data. Secondly, building models that make more complex features accessible early in the architecture may provide an improved match to onset time data presented here, in Cichy et al. (2016) and in Seeliger et al. (2017). Finally, future work should also include implementing other methods to compare model representations to neural representations such as building linear mappings from model features to neural responses to see if findings in those settings corroborate findings obtained while using other comparison methods and quantitatively improve the quality of the comparisons.

References

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, *15*(4), 1-27.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

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

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, *173*, 434–447.

Kaneshiro, B., Arnardóttir, S., Norcia, A. M., & Suppes, P. (2015). Object Category EEG Dataset (OCED). In *Stanford Digital Repository.* Retrieved from `http://purl.stanford.edu/tc919dd5388`

Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., & Suppes, P. (2015). A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *PloS One*, *10*(8), e0135697.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S., & van Gerven, M. (2017). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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.

Yang, Y., Tarr, M. J., Kass, R. E., & Aminoff, E. M. (2018). Exploring spatio-temporal neural dynamics of the human visual cortex. *bioRxiv*, 422576.