Rapid detection of social interactions in the human brain

Leyla Isik (lisik@mit.edu)

Anna Mynick (amynick@mit.edu)

Dimitrios Pantazis (pantazis@mit.edu)

Nancy Kanwisher (ngk@mit.edu)

McGovern Institute for Brain Research, Center for Brains, Minds and Machines, MIT
77 Massachusetts Avenue, Cambridge, MA 02139

Abstract:

Social interaction perception is a crucial part of the human visual experience that develops early in infancy and is shared with other primates. However, it remains largely unknown how humans compute information about others' social interactions from visual input. Is social interaction detection a rapid perceptual process, or a slower post-perceptual inference? To answer this question, we used magnetoencephalography (MEG) decoding and computational modeling to ask how fast the human brain detects third-party social interactions. In particular, subjects in the MEG viewed snapshots of visually matched real-world scenes containing a pair of people who were either engaged in a social interaction or acting independently. We could read out the presence versus absence of a social interaction from subjects' MEG data extremely quickly, as early as 150 ms after stimulus onset. We next showed that late, but not early, layers of a purely feedforward convolutional neural network (CNN) model could detect social interactions in the same images and contained representations that matched those in the MEG data. Taken together, these results suggest that the detection of social interactions is a rapid feedforward perceptual process, rather than a slow post-perceptual inference.

Keywords: Vision; Social interaction perception; MEG decoding; Neural networks

Introduction

Humans are extremely skilled at recognizing social interactions. This ability develops early in infancy (Hamlin, Wynn, and Bloom 2007) and is shared with other primates (Sliwa and Freiwald 2017). We recently identified a region of the human posterior superior temporal sulcus (pSTS) that is selectively engaged when people view third-party social interactions (Isik et al. 2017; Walbrin, Downing, and Koldewyn 2017). These findings underscore the importance of perceiving social interactions, but leave unanswered the question of how quickly and automatically it occurs. Here we ask if and when we can read out information about others' social interactions from MEG data, and if these neural representations are well modeled by a feedforward convolutional neural network.

Results

Experiment

Eleven subjects viewed snapshots of visually matched real-world scenes containing a pair of people who were either engaged in a social interaction or acting independently (Figure 1) in the MEG. These images were captured across 13 different scene/actor pairings to provide visual variability across the images. Subjects viewed each image 25 times. To separate decoding from task demands, subjects performed an orthogonal task: reporting if the pair of people were the same or different gender.

Figure 1: Social interaction dataset depicting pairs of individuals either engaged in a social interaction (left) or acting independently (right). Dataset contained 13 different actor and scene combinations (three shown).

Rapid readout of social interactions

We used MEG decoding to ask if and when we could read out properties of the images. We trained and tested a linear correlation coefficient classifier on each subject's MEG data at each 25 ms non-overlapping time bin, following the pre-processing and decoding methods used in (Isik et al. 2014).

As expected, we could read out image identity (52-way image decoding, training on 80% of the data and testing on 20% held out data) as early as 75 ms after image onset, peaking at 125 ms (Figure 2a). This finding is consistent with previous reports of image decoding (Carlson et al. 2013; Cichy, Pantazis, and Oliva 2014; Isik et al. 2014)

Our key finding is that we could read out whether the subject was viewing an image depicting the presence vs. absence of a social interaction within 150 ms of image onset, and the decoding accuracy peaked at 200 ms (Figure 2b). This latency is similar to those of previously reported visual processes that are thought to be primarily feedforward, such as invariant object recognition (Isik et al., 2014). While early, this decoding is likely not based on very low-level visual features. It is substantially later than the low-level image identity decoding and importantly, social interactions in our dataset cannot be decoded based on low-level image information such as pixels or the output of a V1-like model (Figure 3b).

Social interaction decoding generalizes

To further un-confound the social interaction decoding from low-level image properties, we asked if the neural representations for social interaction generalizes across our different image scenarios. We trained a classifier on 11 of the 13 scenarios in our dataset (each row of Figure 1 depicts a separate scenario), and tested the classifier with data from the two held-out scenes. Although the overall accuracy is lower, we can still decode whether or not the subject viewed a social interaction with the same onset and peak latencies reported above (Figure 2c).

Detecting social interactions with a feedforward CNN

The rapid nature of our MEG decoding suggests that computations are carried out in a primarily feedforward manner. We therefore asked whether a purely feedforward convolutional neural network (CNN) model

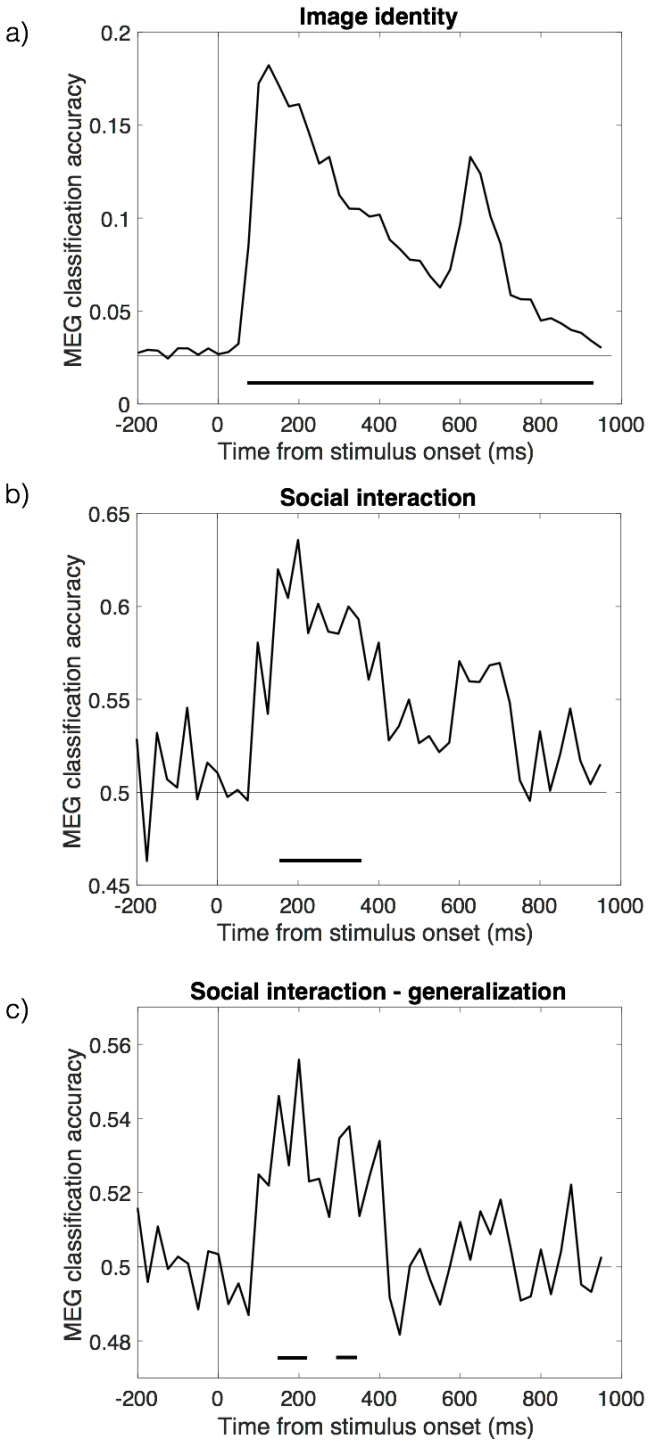

Figure 2: Decoding image identity (a), presence vs. absence of a social interaction (b), and presence vs. absence of social interaction, generalizing across different scenes (c). Black bars at bottom of plots indicate when decoding is $p<0.05$ significant based on a permutation test, cluster corrected to show significant time periods of at least 50 ms.

could identify images depicting the presence or absence of a social interaction. Specifically, we used a version of VGG-16 pre-trained on the ImageNet object classification challenge (Simonyan and Zisserman 2014). We trained a linear classifier using the output of each network layer as features to a linear SVM using the same train/test scheme as the above MEG decoding.

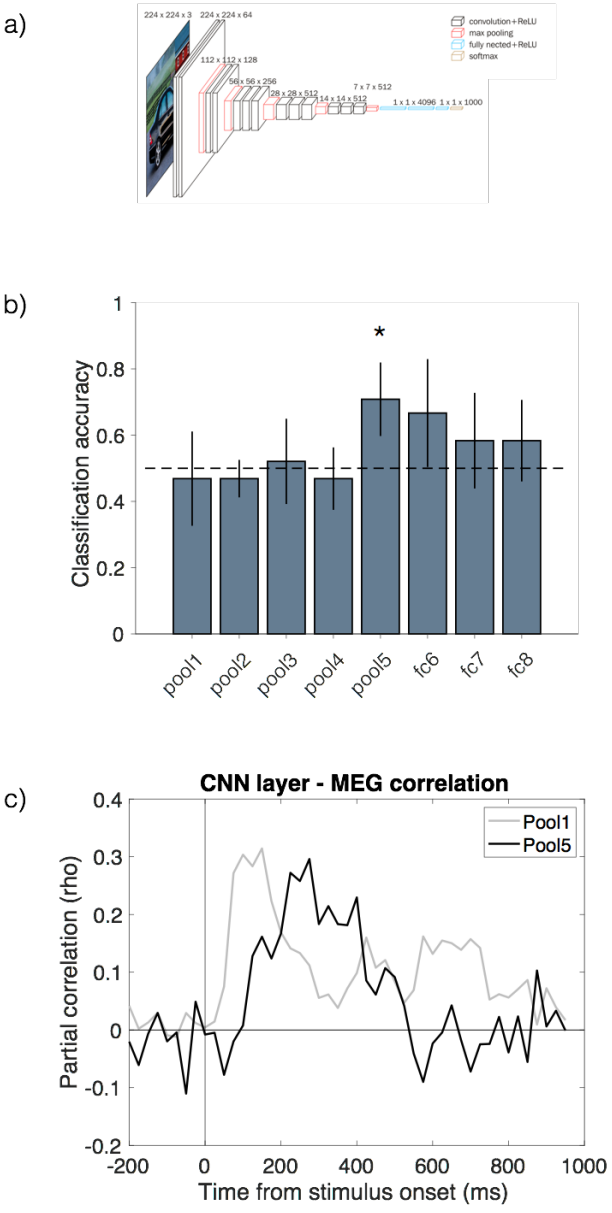

Figure 3: VGG-16 network architecture from (Simonyan and Zisserman 2014) (a), Classification accuracy of the presence vs. absence of social interactions based on each layer's output (b), Partial correlation between the MEG data and the first and last pooling layer of the CNN (c).

We could detect the presence of a social interaction using the network's final pooling layer, but none of the earlier layers (Figure 3a). This finding reinforces our conclusion that low-level image features cannot be used to detect social interactions in our dataset, but shows that later layers of a feedforward network can do so. We observe a drop off in accuracy in the final fully connected layers, likely due to the fact that these layers have been over-trained on an object recognition task.

While the CNN can classify presence vs. absence of a social interaction with above chance accuracy, this accuracy is substantially lower than a binary object categorization task that the network was trained on (e.g., cat versus dog). This finding suggests that detection of social interactions would be more efficient with templates specialized for this role, beyond those used for generic object recognition. Consistent with this idea, our prior neuroimaging results have identified a specialized cortical region for social interaction perception (Isik et al. 2017; Walbrin, Downing, and Koldewyn 2017).

Matching representations in CNN and MEG

We can further compare the representations in each model layer to our MEG data using representational similarity analysis (Kriegeskorte, Mur, and Bandettini 2008). We compute a dissimilarity matrix based on the pairwise distance between all images for each layer of our CNN as well as for our MEG data. In order to account for the high correlation between the different CNN layers, we compute the partial correlation between each CNN layer and the MEG data, factoring out the contribution from all other layers. We find that the representation in the last pooling layer not only discriminates the presence of an interaction, but also matches the MEG data at later time points when social interaction information is decodable (Figure 3c). Earlier CNN layers, on the other hand, match the MEG data at earlier time points when lower-level image identity is decodable. These CNN results suggest that social interactions can be identified based on a feedforward model and this model's representation matches that observed in our MEG data.

Conclusions

We report a rapid, spontaneous readout of the presence of social interactions from human MEG data. The latency of social interaction detection is similar to that observed for other primarily feedforward

processes, such as object recognition. These neural representations for social interactions generalize across different images and scenes, and are well matched by the last pooling layer of a feedforward CNN. Taken together, these results suggest that human brain can rapidly detect social interactions based on perceptual cues.

These results do not, however, indicate that all social interaction information is computed in a fast, feedforward manner. Social interaction perception has both challenges that are shared with other visual recognition problems, such as occlusion, as well as unique challenges, such as incorporating information about agents' goals and values, which likely require additional top-down or cognitive inputs. Our results suggest that like other vision problems, such as object and scene perception, social interaction recognition has a fast, perceptual component that may serve as input to later top-down and cognitive processing.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Carlson, Thomas, David A Tovar, Arjen Alink, and Nikolaus Kriegeskorte. 2013. "Representational Dynamics of Object Vision: The First 1000 Ms." *Journal of vision* 13(10): 1-.

Cichy, Radoslaw Martin, Dimitrios Pantazis, and Aude Oliva. 2014. "Resolving Human Object Recognition in Space and Time." *Nature neuroscience* 17(3): 455–62.

Hamlin, J Kiley, Karen Wynn, and Paul Bloom. 2007. "Social Evaluation by Preverbal Infants." *Nature* 450(7169): 557–59.

Isik, Leyla, Kami Koldewyn, David Beeler, and Nancy Kanwisher. 2017. "Perceiving Social Interactions in the Posterior Superior Temporal Sulcus." *Proceedings of the National Academy of Sciences of the United States of America* 114(43): E9145–52.

Isik, Leyla, Ethan M Meyers, Joel Z Leibo, and Tomaso Poggio. 2014. "The Dynamics of Invariant Object Recognition in the Human Visual System." *Journal of neurophysiology* 111(1): 91–102.

Kriegeskorte, Nikolaus, Marieke Mur, and Peter Bandettini. 2008. "Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience." *Frontiers in systems neuroscience* 2: 4.

Simonyan, Karen, and Andrew Zisserman. 2014. "Very Deep Convolutional Networks for Large-Scale Image Recognition." http://arxiv.org/abs/1409.1556.

Sliwa, J., and W. A. Freiwald. 2017. "A Dedicated Network for Social Interaction Processing in the Primate Brain." *Science* 356(6339).

Walbrin, Jon, Paul E. Downing, and Kami Koldewyn. 2017. "The Visual Perception of Interactive Behaviour in the Posterior Superior Temporal Cortex." *Journal of Vision* 17(10): 990.