Ubiquitous visual representations during neural processing of a naturalistic movie

Hannah Small (hsmall2@jhu.edu)
Johns Hopkins University, Department of Cognitive Science
Baltimore, MD 21218, US

Haemy Lee Masson (haemy.lee-masson@durham.ac.uk)
Durham University, Department of Psychology
Durham, DH13LE, UK

Ericka Wodka (wodka@kennedykrieger.org)
Kennedy Krieger Institute, Center for Autism Services, Science, and Innovation
Baltimore, MD 21211, US
Johns Hopkins School of Medicine, Department of Psychiatry and Behavioral Sciences
Baltimore, MD 21205, US

Stewart H. Mostofsky (mostofsky@kennedykrieger.org)
Kennedy Krieger Institute, Center for Neurodevelopmental and Imaging Research
Baltimore, MD 21205, US
Johns Hopkins School of Medicine, Department of Neurology, Department of Psychiatry and Behavioral Sciences
Baltimore, MD 21205, US

Leyla Isik (lisik@jhu.edu)
Johns Hopkins University, Department of Cognitive Science
Baltimore, MD 21218, US

Abstract

Social cognition depends on integrating information from both vision and language. However, prior work has mostly studied vision and language separately, not accounting for the rich social visual and verbal semantic signals that occur simultaneously in natural settings. To understand how this information is integrated during natural movie viewing, we fit a voxel-wise encoding model that included low- and mid-level visual and auditory features, as well as higher-level social and language features, including the presence of a social interaction and language model embeddings of the spoken language in the movie. We find distinct voxels supporting visual social processing and language. However, surprisingly, we also find that both social and language voxels across cortex are best predicted by visual features extracted from a convolutional neural network (CNN), suggesting that when vision and language are combined in naturalistic settings, visual features dominate neural processing.

Keywords: social processing; naturalistic stimuli; fMRI encoding; vision; multi-modal processing

Introduction

Social processing involves integrating visual and linguistic input, however, the processing of these two inputs are often studied separately. Previous work using controlled stimuli has separately mapped the responses to diverse social signals, including visual (biological motion, faces) and linguistic (voices, theory of mind, and language) input in bilateral superior temporal sulcus (STS). This work revealed regions that were highly selective for specific types of social stimuli but also regions that responded to multiple types and modalities of social information (Deen, Koldewyn, Kanwisher, & Saxe, 2015). Recently there has been a push to understand if the insights gained from simple, controlled experiments can generalize to more naturalistic stimuli. Looking at brain responses to more ecological stimuli often reveals broader activations in the brain and novel findings in social cognition (Redcay & Moraczewski, 2020). Naturalistic stimuli provide a straightforward way to study simultaneous vision and language.

To date, encoding models have been used to analyze either language responses from listening to stories or podcasts (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Schrimpf et al., 2021; Goldstein et al., 2022), or visual responses from watching movies (Huth, Nishimoto, Vu, & Gallant, 2012; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017). Some have analyzed both vision and language in the same subjects, but using different stimuli for each modality (e.g., silent movies and podcasts) (Popham et al., 2021). Previous work has shown that an encoding model approach to naturalistic stimuli can reveal voxels with unique variance explained by regions that support social interaction perception in adult bilateral STS (Lee Masson & Isik, 2021). However, none have analyzed the neural processing of both vision and language from the same naturalistic input. This could be due to the difficulties in interpreting responses to movies since many social features of interest co-vary with perceptual features (Grall & Finn, 2022).

Here, we study the neural responses to simultaneous vision and language during natural movie viewing using an encoding model approach that controls for co-varying features. We model a combination of annotated and automatically extracted visual, social, and linguistic features in the movie. We find distinct voxels supporting social and linguistic representations, yet they are both best predicted by visual features.

Methods

fMRI experiment

Naturalistic stimuli and data analysis Participants (n=17, neurotypical, ages 19-34, 10 female) watched a 45 minute episode of the BBC series Sherlock, split into 2 runs. See experimental details in Chen et al. (2017). For each subject, the fMRI BOLD series for each voxel within a previously computed intersubject correlation (ISC) mask (Lee Masson & Isik, 2021) was predicted with a banded ridge regression model (Dupré la Tour, Eickenberg, Nunez-Elizalde, & Gallant, 2022) using previously annotated and automatically extracted features (Chen et al., 2017). This included the first 147 PCs of features extracted from the fifth layer of Alexnet (Lee Masson & Isik, 2021), which predict visual responses in high-level visual cortex (Eickenberg et al., 2017). We also extracted motion energy features using pymoten (Nunez-Elizalde, Deniz, Dupré la Tour, Visconti di Oleggio Castello, & Gallant, 2021). Additionally, we extracted language features from the episode transcript on both the word and sentence level using a word2vec model (Mikolov, Chen, Corrado, & Dean, 2013) and a sentence transformer model (sbert; all-MiniLM-L6-v2, huggingface.co), respectively. Banded ridge regression allows each feature space to learn a separate ridge penalty to better account for different sizes in the feature spaces (i.e., the uni-dimensional annotated features versus high-dimensional language and visual features). To account for temporal autocorrelation in the movie and fMRI data, we grouped the signal into blocks of 17 TRs (25.5 s) before splitting into train/test. We examined the individual product measure, a measure of the predictive contribution of each feature space that takes the correlation between feature spaces into account (Dupré la Tour et al., 2022).

Controlled stimuli and data analysis All participants also watched videos of point light walkers that were engaged in social actions and point light walkers perform-

ing independent actions (Isik, Koldewyn, Beeler, & Kanwisher, 2017). A subset of participants (n=7) completed a language localizer, listening to audio of intact and degraded speech (Scott, Gallée, & Fedorenko, 2017). We identified the top 100 social interaction and language selective voxels within temporal and frontal regions.

Results

The full encoding model explains significant group-level variance ($p<0.001$, FDR corrected) in all voxels in the ISC mask (Figure 1).

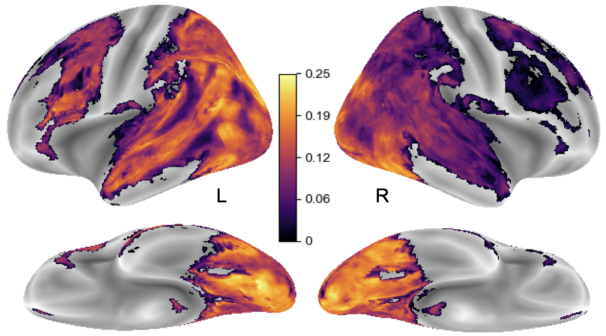

Figure 1: Group map of the explained variance, averaged across subjects per voxel in MNI space.

The majority of voxels across the whole brain are best predicted by visual features extracted from Alexnet and next by motion energy in both group analyses and individual subjects (Figure 2). However, in individual subjects, there are voxels that are best predicted by other features, including the social features (notably valence and to a lesser extent social interaction) and language features.

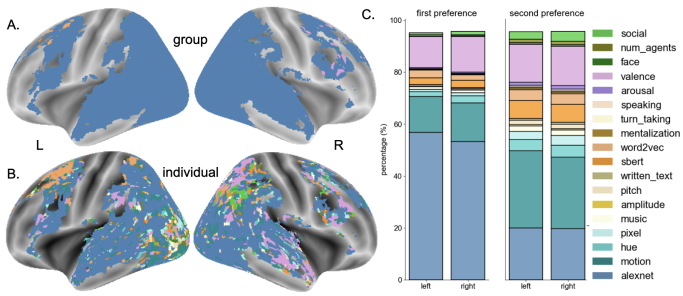

Figure 2: A. Group and B. representative preference map showing the feature that explains the most variance in each voxel. C. Percentage of voxels with the most variance explained by each feature in left and right hemispheres, averaged across subjects.

Surprisingly, we find strong visual feature predictivity in social interaction and even language selective regions. In temporal social perception regions and temporal and frontal language regions, Alexnet embeddings are significantly more predictive than either social or language features (Figure 3).

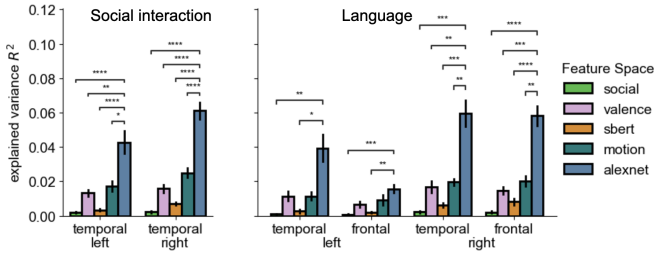

Figure 3: Average explained variance (measured as individual product measure) in regions localized using controlled stimuli. Bars represent mean and error bars represent the standard error of the mean. Asterisks indicate statistical significance.

Discussion

To our knowledge, this is the first attempt at modeling both visual and linguistic signals in one naturalistic context at the same time. Previous encoding model approaches have analyzed these modalities in separate stimuli (Huth et al., 2016; Schrimpf et al., 2021; Goldstein et al., 2022; Huth et al., 2012; Eickenberg et al., 2017; Popham et al., 2021; Tang, Du, Vo, Lal, & Huth, 2023) or focus on analysis of only one type of only one modality (Lee Masson & Isik, 2021). Here, we model both the visual and linguistic signals in one naturalistic context with one encoding model. We find that visual features extracted from Alexnet are the best predictor of neural activity across the brain. This is somewhat surprising as Alexnet is now far from the best vision model available in terms of either performance or match to neural data (Conwell, Prince, Kay, Alvarez, & Konkle, 2022). Further, this is not driven by the dimensionality of the feature space, as Alexnet is reduced to 147 dimensions by PCA, while the word2vec feature space is 300D and the sbert feature space is 384D.

The high performance of Alexnet could be due to a shared semantic space that is well captured by vision model embeddings. This is also in line with other work finding semantic alignment between vision and language (Popham et al., 2021; Tang et al., 2023). Our results may also support a recent proposal that visuospatial coding is ubiquitous in the brain, even in areas beyond visual cortex, serving to ground human cognition in a common reference space (Groen, Dekker, Knapen, & Silson, 2022). Future work will look at other non-visual models (including larger language models) and the shared and unique variance of semantic representations across vision and language. Overall, this work highlights the need for multi-modal studies of social perception in natural contexts.

Acknowledgments

We are grateful to Elizabeth Im for help with data collection and members of the Isik lab for helpful discussion of this work. We thank Alyssa DeRonda, Natalie Alessi, and Beatrice Ojuri for help with subject recruitment and testing. This work was supported by NIMH R21MH129899 and NSF GRFP DGE2139757.

References

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/nn.4450

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022, March). *What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?* doi: 10.1101/2022.03.28.485868

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(11), 4596–4609. doi: 10.1093/cercor/bhv111

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728. doi: 10.1016/j.neuroimage.2022.119728

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. doi: 10.1016/j.neuroimage.2016.10.001

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. doi: 10.1038/s41593-022-01026-4

Grall, C., & Finn, E. S. (2022). Leveraging the power of media to drive cognition: a media-informed approach to naturalistic neuroscience. *Social Cognitive and Affective Neuroscience*, *17*(6), 598–608. doi: 10.1093/scan/nsac019

Groen, I. I. A., Dekker, T. M., Knapen, T., & Silson, E. H. (2022). Visuospatial coding as ubiquitous scaffolding for human cognition. *Trends in Cognitive Sciences*, *26*(1), 81–96. doi: 10.1016/j.tics.2021.10.011

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. (Publisher: Nature Publishing Group) doi: 10.1038/nature17637

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1714471114

Lee Masson, H., & Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage*, 118741. doi: 10.1016/j.neuroimage.2021.118741

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* arXiv. (arXiv:1301.3781 [cs])

Nunez-Elizalde, A. O., Deniz, F., Dupré la Tour, T., Visconti di Oleggio Castello, M., & Gallant, J. L. (2021). *pymoten: scientific python package for computing motion energy features from video.* doi: 10.5281/zenodo.6349625

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, *24*(11), 1628–1636. (Number: 11 Publisher: Nature Publishing Group) doi: 10.1038/s41593-021-00921-6

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, 116392. doi: 10.1016/j.neuroimage.2019.116392

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. doi: 10.1073/pnas.2105646118

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. doi: 10.1080/17588928.2016.1201466

Tang, J., Du, M., Vo, V., Lal, V., & Huth, A. (2023). Brain encoding models based on multimodal transformers can transfer across language and vision. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 29654–29666). Curran Associates, Inc.