Cortical Semantic Encoding Varies with Attention and Habituation

Isaac Ray Christian (isaacrc@princeton.edu)
Department of Psychology, Washington Rd
Princeton, New Jersey 08540 USA

Michael Graziano (graziano@princeton.edu)
Princeton Neuroscience Institute, Washington Rd
Princeton, New Jersey 08540 USA

Rachel Metzgar (rm9561@princeton.edu)
Department of Psychology, Washington Rd
Princeton, New Jersey 08540 USA

Sam Nastase (snastase@princeton.edu)
Princeton Neuroscience Institute, Washington Rd
Princeton, New Jersey 08540 USA

Abstract:

How do attentional goals bias how the brain encodes semantic content in the external world? We studied brain activity as participants performed two tasks: (1) focusing attention internally and de-emphasizing task-irrelevant external information, and (2) focusing attention externally to enhance task-relevant external information. We presented movie clips to participants while collecting fMRI data. Each clip was repeated four times in sequence to explore how attentional goals interact with novelty. We measured semantic encoding performance as subjects ignored or paid attention to the videos during four repetitions. We used the multimodal transformer model CLIP to determine the extent to which semantic content from the video stimulus was encoded in different regions of the cortex. We found that semantic networks were less sensitive to manipulations of attention while fronto-parietal attention networks and visual cortices encoded the video stimuli in a manner modulated by task goals. More broadly, we show that representation of external content can diminish due to interference from task goals and habituation to the stimulus.

Keywords: multimodal; CLIP; fronto-parietal; default; encoding; control; habituation.

Introduction

There has been considerable progress in using the internal representations of artificial neural networks to better understand the internal representations of the brain. Neuroimaging work has found that representations learned in artificial neural networks can predict brain responses, suggesting similarity in the way semantic representations emerge (Naselaris, Kay, Nishimoto, & Gallant, 2011). Recent work further identifies that using artificial networks to predict brain responses improves when network architectures more plausibly mimic cognition. For example, networks that incorporate multimodal input streams better predict brain activity than do unimodal counterparts (Lu et al., 2022). Moreover, higher order networks distributed across fronto-parietal cortices are better explained by multimodal representations than lower- order visual and linguistic areas (Wang, Kay, Naselaris, Tarr, & Wehbe, 2023).

While neural network models may help elucidate semantic information encoded in the brain, they do not capture the internal attention processes that orient and modify these representations. Such attentional processes bias information based on task goals by selectively enhancing or inhibiting the appropriate internal representations. These control representations are also thought to be distributed across frontal parietal networks (Corbetta & Shulman, 2002; Vincent, Kahn, Snyder, Raichle, & Buckner, 2011), overlapping with higher order areas that show representation of semantic content according to encoding models (Huth, Nishimoto, Vu, & Gallant, 2012). While some work has investigated the effects of attending to different stimulus features (e.g. Çukur, Nishimoto, Huth, & Gallant, 2013), very little work has explored how internal attention can disrupt or enhance the encoding of external content. In this work, we tested how internal versus external attention affect semantic encoding, and how this relates to stimulus novelty and habituation.

Methods

Thirty-nine subjects participated in an attention task during fMRI acquisition. In the external attention condition, a 2-minute video clip was presented. Participants were instructed to pay attention to the video and press a button if they noticed that their mind had wandered from the video content. Participants repeated this task four times, repeatedly viewing the same video. In the internal attention condition, participants viewed another 2-minute video clip, this time tasked with ignoring the video and instead focusing on the rhythmic sensation of their breathing while keeping their eyes open and on screen (a task commonly associated with meditation). Once again subjects were instructed to press the button if they noticed that their mind had wandered from their breathing. The internal attention task was also completed four times in a row with the same external video. A total of 6 movies were presented to participants, three of which are reported in the current analysis. Each movie was used in the external condition for half of subjects and in the internal condition for the other half of subjects.

For every TR (1.5 seconds), the list of spoken words transcribed from the video during the 1.5 second interval and words from the previous 4 TRs (6 seconds) were collated and supplied as input to the multimodal transformer model CLIP (Radford et al., 2021). Embeddings for each TR were extracted from the last layer of CLIP's text encoder and used to predict brain activity of 400 parcels selected using the Schaeffer atlas (Schaefer et al., 2018) using leave-one-subject-out cross-validation and banded ridge regression. The model-predicted brain activity was correlated with the actual activity of each parcel, producing 400 correlation values for each subject. For each analysis, rank sign Wilcoxon tests were conducted across subjects and then corrected for multiple comparisons using FDR. Significant parcels were averaged across movies and participants and projected back onto the cortical surface of the brain for plotting.

Results

When participants attended to the movie (external condition), we observed significant encoding accuracy across fronto-parietal, sensory, and semantic networks (Figure 1A).

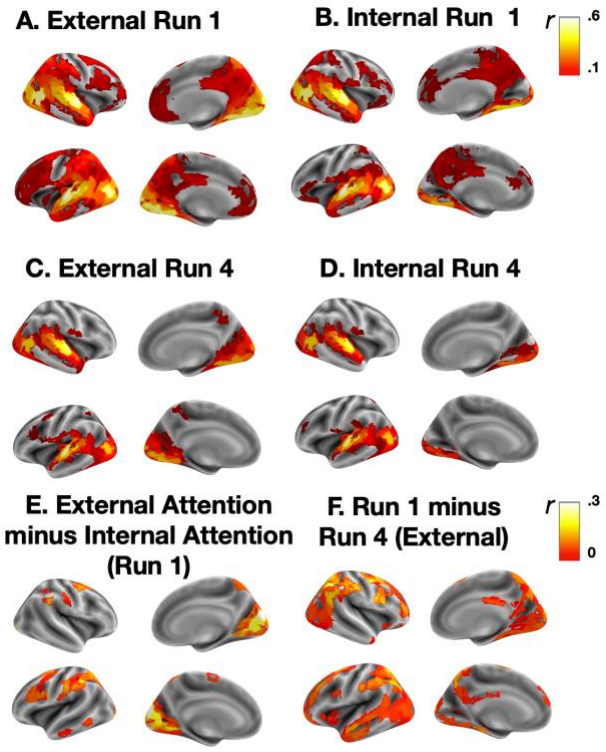

Figure 1: Significant encoding accuracy by condition.

When participants were tasked with ignoring the video and attending internally, significant encoding of the video was still observed in similar regions across the brain (Figure 1B). The largest differences between the two conditions were observed in visual cortex and attention networks ($p < .05$, FDR corrected), while regions in temporal semantic networks and the Default Mode Network (DMN) encoded information similarly (Figure 1E).

On the fourth repetition of the movie stimulus, when participants were tasked with paying attention to the movie (Figure 1C), we observed significant encoding performance in occipital and temporal regions, but not higher-level semantic regions (e.g. angular gyrus, precuneus). A left lateralized dorsal attention network encoded semantic information better when attention was directed externally but not internally, but only on the fourth repetition (Figure 1D). Effects of habituation were most prominently observed in fronto-parietal cortices, but also in visual regions and temporal semantic networks (Figure 1F).

Discussion

Our results show that cortical semantic encoding during naturalistic movie-viewing can be decreased or enhanced depending on the attentional task goal. Habituation, occurring after multiple repetitions of the same movie stimulus, produced worse encoding across the brain, but most notably in fronto-parietal cortices. We replicate findings that text embeddings trained

jointly with visual images predict widespread cortical semantic encoding, with strong semantic encoding in both visual and language areas, when paying attention to a stimulus (Popham et al., 2021).

We found that widespread semantic encoding across cortex declined when attending to internal, endogenous signals, as expected. Fronto-parietal attention networks, along with regions in visual cortices, were most sensitive to the attention manipulation. In contrast, bilateral temporal and DMN areas implicated in semantic representation were less sensitive to the attention manipulation. This suggests that ostensibly high-level semantic encoding nonetheless persists even when attention is directed inward. High-level semantic encoding (e.g. in DMN areas) does decline, however, with habituation. Encoding differences in attention networks potentially suggest functionality that can be repurposed by context: when the task goal is to attend externally, such networks aid in perception, and when the task goal is to ignore content, their role flexibly shifts to the internal task.

The observed habituation effect over multiple viewings of the same naturalistic stimulus has important implications for estimating within-subject noise ceilings in encoding analysis. Multiple repetitions of the same stimulus are sometimes used to estimate the reliability of neural data for evaluating model-based predictions (e.g. Huth et al., 2012). Our findings demonstrate that, even when participants make an effort to attend to the repeated stimulus, semantic encoding performance declines with diminishing novelty (habituation may also include memory and prediction processes; Aly & Turk-Browne, 2018; Michelmann et al., 2021). This will tend to underestimate the noise ceiling and therefore overestimate model performance relative to the ceiling.

Our results also show a left lateralized dorsal attention network that encodes information after four repetitions when the task is to attend externally, but not internally. This result perhaps indicates an attentional 'boost' of stimulus related content, aligning with previous work finding that attention can counter the effects of habituation (Pestilli, Viera, & Carrasco, 2007). Other fronto-parietal control networks that did not show encoding in either condition may subserve other attention functions, such as reorienting and maintenance.

References

Aly, M., & Turk-Browne, N. B. (2018). Flexible weighting of diverse inputs makes hippocampal function malleable. *Neuroscience Letters, 680*, 13-22. doi:10.1016/j.neulet.2017.05.063

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201-215. doi:10.1038/nrn755

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience, 16*(6), 763-+. doi:10.1038/nn.3381

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

Lu, H., Zhou, Q., Fei, N., Lu, Z., Ding, M., Wen, J., . . . He, H. (2022). Multimodal foundation models are better simulators of the human brain. *arXiv preprint arXiv:2208.08263*.

Michelmann, S., Price, A. R., Aubrey, B., Strauss, C. K., Doyle, W. K., Friedman, D., . . . Norman, K. A. (2021). Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nature Communications, 12*(1). doi:ARTN 5394
10.1038/s41467-021-25376-y

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage, 56*(2), 400-410. doi:10.1016/j.neuroimage.2010.07.073

Pestilli, F., Viera, G., & Carrasco, M. (2007). How do attention and adaptation affect contrast sensitivity? *Journal of Vision, 7*(7). doi:Artn 9
10.1167/7.7.9

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. doi:10.1038/s41593-021-00921-6

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning, Vol 139, 139*. Retrieved from <Go to ISI>://WOS:000768182704084

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., . . . Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex, 28*(9), 3095-3114. doi:10.1093/cercor/bhx179

Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E., & Buckner, R. L. (2011). Evidence for a Frontoparietal Control System Revealed by Intrinsic Functional Connectivity (vol 100, pg 3328, 2008). *Journal of Neurophysiology, 105*(3), 1427-1427. doi:10.1152/jn.z9k-0231-corr.2010

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature*

Machine Intelligence, 5(12), 1415-1426. doi:10.1038/s42256-023-00753-y