Mapping and decoding language representations in fMRI

Alexander G. HUTH[1]

1. Departments of Neuroscience & Statistics, University of California, Berkeley, United States

E-mail: *alex.huth@berkeley.edu*

Abstract: We study language processing by recording fMRI responses to hours of natural speech. Encoding models from neural network features reveal how information maps across cortex, and can also decode perceived or imagined language, enabling new brain-computer interface approaches.

Keywords: Neuroimaging, language, encoding models, semantics, language models, decoding

Introduction

Language enables humans to compress complex thoughts, ideas, and memories into linear sequences of tokens, and to decode those sequences back into thought. These processes involve a network of brain areas that range from low level—processing sounds to identify words—to high level—representing ideas and thoughts. We map these processes using a natural language experiment, in which research participants listen to hours of natural narrative stories while their brain responses are recorded using functional magnetic resonance imaging (fMRI). To model these data, we drawn on neural network language models (LMs) or self-supervised learning (SSL) speech models. These are large-scale neural networks, typically based on the transformer architecture, that learn to predict one part of a stimulus, such as a word in a piece of text or an audio frame in a recording, from the other parts. Through this process, the neural networks learn to form highly compressed and generalizable representations of the stimuli. We then use those representations as the basis for encoding models, which are machine learning models that predict brain responses to a stimulus based on features extracted from that stimulus. After fitting the encoding models using large training datasets, we can then use them to predict responses to new stimuli that were not used for model fitting. Model prediction performance can then be assessed by comparing predicted to actual brain responses, providing a sensitive metric for how well the model can explain the brain responses. Separate encoding models are fit for each voxel in each participant's brain, providing extremely detailed and high-resolution models of cortical representations.

Once they have been fit, encoding models can be used in several ways to answer scientific questions about the brain. One technique is to compare prediction performance between encoding models that use different types of stimulus features to make their predictions. If encoding models employing one type of feature consistently outperforms encoding models using another type, then we can conclude that the first type of features is a better match to brain representations. For example, this technique might be employed to determine where information about sound frequencies, phonemes, parts-of-speech, or high-level semantic concepts are represented in cortex.

Another technique uses the parameters of the estimated encoding models to make determinations about which exact features within a particular feature space are represented in each brain area. Here we combine this technique with a novel way to construct encoding models, which we call question-answering (QA) encoding models [1]. Most recent encoding models for language use representations drawn from the state vectors of neural network language models, which are highly compressed and not interpretable. To create a fundamentally interpretable encoding model, we replaced these LM-based embeddings with embeddings that consist of answers to various questions about a piece of text.

Finally, we can reverse encoding models to create decoding models, which predict language from brain activity [2]. Here we use a technique called Bayesian decoding [3] to try to find the most likely sequence of words to have caused a particular spatiotemporal pattern of brain activity. Successful decoding of continuous language using fMRI would enable a new type of brain-computer interface that we call a semantic decoder, as it reads out the meaning of a piece of text rather than its exact words.

Together, these techniques demonstrate how the power of modern machine learning methods and large brain imaging datasets using natural stimuli for mapping and understanding how our brains process language.

Methods

These experiments use fMRI data collected while 3 participants listen to up to 20 hours of natural, narrative language stimuli drawn from *The Moth Radio Hour* and other sources [4]. These data are publicly shared and may be used by anyone to construct encoding and decoding models.

QA encoding models were constructed by using GPT-4 to answer a series of questions about each 10-word span in the narrative language stimuli [1]. Questions were designed partially by hand and partially by prompting GPT-4. Each question is answered by yes or no, e.g. "Does this piece of text mention a place or location?". The answers are encoded as 1 or 0 and then concatenated to form vectors that represent each piece of text. Encoding models are fit using these vectors. Semantic decoders were constructed using encoding models based on GPT combined with multivariate normal noise models [3]. Decoding was

performed using a beam search procedure [2]. These methods were applied to data collected while participants listened to stories, imagined telling stories, or watched silent videos.

Results

QA encoding models were found to predict brain responses to new stimuli with a high degree of accuracy, comparable to that of encoding models that are fit using features drawn from the latest Llama LLM. Examining the weights of these models shows which questions best map to each brain area.

Semantic decoding on perceived speech is highly successful at reproducing the overall semantic content of the stimuli, but not its exact wording or form. Decoding on imagined speech is less precise, but still well above chance level. Semantic decoding applied to data collected while participants view silent videos produces a coarse description of the events of the video.

Discussion

Our findings demonstrate that large-scale neural network language models can successfully capture the computational principles underlying human language processing across cortical areas. The comparable performance of our interpretable QA encoding models to state-of-the-art LLM-based models suggests that explicit semantic features can effectively characterize brain representations without sacrificing predictive power. This interpretability advantage allows us to identify which specific linguistic and semantic properties are encoded in different cortical regions, providing mechanistic insights into the brain's language network.

The success of semantic decoding across multiple conditions—perceived speech, imagined speech, and silent video viewing—reveals the robustness of cortical language representations. While decoded content captures semantic meaning rather than exact wording, this finding supports theories that the brain primarily encodes conceptual rather than surface-level linguistic information.

These results have important implications for brain-computer interfaces. Unlike traditional BCIs that focus on motor control, our semantic decoder demonstrates the feasibility of directly accessing high-level cognitive content from cortical activity. However, the current temporal resolution of fMRI limits real-time applications, suggesting that future work should explore similar approaches with higher temporal resolution recording methods.

Conclusions

We present a comprehensive framework for mapping language processing in the human brain using natural stimuli and modern machine learning techniques. Our QA encoding models provide interpretable insights into cortical language representations, while our semantic decoder demonstrates the potential for meaning-based brain-computer interfaces. The successful decoding of semantic content across perceived speech, imagined speech, and visual narrative understanding suggests that cortical language networks encode abstract conceptual information that transcends specific input modalities. This work establishes a foundation for both advancing our scientific understanding of human language processing and developing next-generation neural interfaces that can access the semantic content of human thought.

Acknowledgements

This work was supported by NIH/NSF CRCNS award, the Burroughs-Wellcome Fund Career Award at the Scientific Interface (CASI), and generous gifts from Intel and Microsoft.

References

[1] Benara, Vinamra, et al. "Crafting interpretable embeddings for language neuroscience by asking LLMs questions." *Advances in neural information processing systems* 37 (2024): 124137.

[2] Tang, Jerry, et al. "Semantic reconstruction of continuous language from non-invasive brain recordings." *Nature Neuroscience* 26.5 (2023): 858-866.

[3] Nishimoto, Shinji, et al. "Reconstructing visual experiences from brain activity evoked by natural movies." *Current biology* 21.19 (2011): 1641-1646.

[4] LeBel, Amanda, et al. "A natural language fMRI dataset for voxelwise encoding models." *Scientific Data* 10.1 (2023): 555.