Teasing apart the representational spaces of ANN language models to discover key axes of model-to-brain alignment

Eghbal A. Hosseini (ehoseini@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
McGovern Institute for Brain Research, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Noga Zaslavsky (nogazs@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
McGovern Institute for Brain Research, Massachusetts Institute of Technology
The MIT Quest for Intelligence Initiative, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Colton Casto (ccasto@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
McGovern Institute for Brain Research, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Evelina Fedorenko (evelina9@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
McGovern Institute for Brain Research, Massachusetts Institute of Technology
Speech and Hearing Bioscience and Technology (SHBT) Program, Harvard University
Cambridge, MA 02139, USA

Abstract:

A central goal of neuroscience is to uncover neural representations that underlie sensorimotor and cognitive processes. Artificial neural networks (ANN) can provide hypotheses about the nature of neural representations. However, in the domain of language, multiple ANN models provide a good match to human neural responses. To dissociate these models, we devised an optimization procedure to select stimuli for which model representations are maximally distinct. Surprisingly, we found that all models struggle to predict brain responses (fMRI) to such stimuli. We further a) confirmed that these sentences are not outliers in terms of linguistic properties and that neural responses to these sentences are as reliable as to random sentences, and b) replicated this finding in another, previously collected, dataset. Stimuli for which model representations differ can be used to uncover dimensions of ANN-to-brain alignment, and serve to build more brain-like computational models of language.

Keywords: ANNs; language; controversial stimuli; representations

Introduction

Successful behavior relies on the brain's ability to construct the statistical representations of the environment that can guide actions. Characterizing these representations is a critical component in understanding brain computations (Simoncelli & Olshausen, 2001). One approach to uncovering neural representations is to build computational models that perform some behavior of interest, measure the models' alignment with brain activity, and—in the presence of good alignment—use the models for generating hypotheses about the structure of neural representations (Bao et al., 2020). This approach has recently helped characterize perceptual and motor processes across domains (Kell et al., 2018; Sussillo et al., 2015; Yamins et al., 2014) and has promise for understanding higher-level cognitive processes. For example, for language comprehension, a number of ANN models have been shown to be able to capture human brain activity (e.g., (Caucheteux & King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021). However, what aspects of the *models' representations of linguistic strings* lead to good model-to-brain alignment remains poorly understood.

One way to make progress is to treat different ANN models as competing hypotheses and test their alignment with the brain (Golan et al., 2023; Schrimpf et al., 2020). To do so effectively requires stimuli that elicit distinct representations across models. Here, we devised an optimization procedure (Figure 1) to select a stimulus set that would separate several top-performing ANN language models at the level of their representations. We then tested these models for their ability to capture brain responses to these stimuli, and we used a similar approach on an existing dataset via stimulus sub-sampling.

Methods

Stimulus optimization: We selected 7 high-performing ANN language models with diverse architectures from a prior study (Schrimpf et al., 2021): roberta-B; xlnet-L-cased; bert-L-uncased-whole-word-masking; xlm-mlm-en-2048; gpt2-xl; albert-xxlarge-v2; and ctrl models. We extracted these models' representations for a set of 8,409 sentences (selected from the Universal Dependencies corpus (de Marneffe et al., 2021) constraining sentences to be between 6 and 19 words long. We used a representational dissimilarity matrix (RDM) to characterize each model's representation of these sentences, and measured model-to-model dissimilarity via second-order RDMs for each model pair. In the critical step, we iteratively sampled and substitute stimuli to find a subset of n=200 stimuli that maximized the distances among the models (S_{max}; Figure 1A-C). Prior to data collection, we examined the selected sentences' linguistic properties to ensure that they are not outliers / 'edge cases' in any of the key dimensions of linguistic variation (Figure 1D), nor model representations (not shown here).

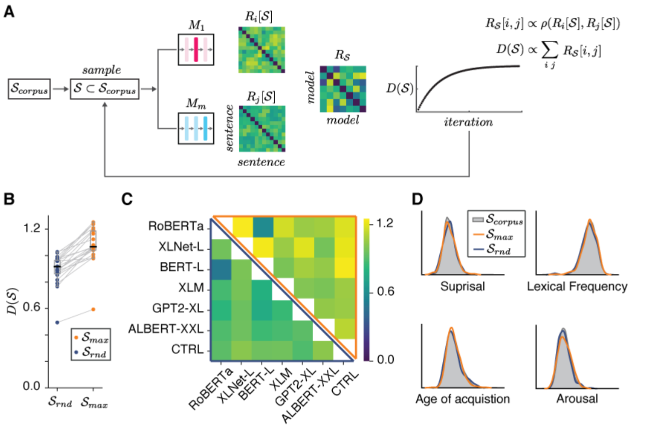

Figure 1: (A) Stimulus optimization procedure; (B) Inter-model distance for the optimized sentences (orange dots) vs. random sentences (blue dots); (C) Inter-model distance for optimized (top right triangle) vs. random (bottom left triangle) sentences; (D) Distributions of linguistic features for optimized and random sentences

(orange and blue curves, respectively) shown against the distribution for the entire corpus (grey).

Experimental design: We recorded auditory versions of the sentences and presented them in fMRI to 8 participants (2-6 seconds per sentence with 4 second ISI). We extracted responses to each sentence from the language areas of each participant (identified with an independent localizer; Fedorenko et al., 2010).

Encoding models: Following (Schrimpf et al., 2021), we built a cross-validated regression for each model from unit activations to voxel-level sentence responses (Figure 2A). We then calculated the average correlation between predicted and actual voxel responses for a left-out set of stimuli and normalized it by a ceiling computed from inter-subject similarity.

Results

ANN models struggle to predict neural responses for the optimized stimuli. We first compared the models' performance on the optimized stimuli relative to the Pereira et al. (2018) benchmark (used in Schrimpf et al., 2021). Contra our expectation that these optimized stimuli would reveal that some ANN models are better able to capture human neural responses than others, we found that *all models* show poor ability to predict brain responses to these sentences (Figure 2B, orange bars; cf. blue bars, which correspond to performance on Pereira2018 sentences). Importantly, response reliability, as reflected in the ceiling value, was similar Pereira2018 dataset (Schrimpf et al., 2021), ruling out the possibility that the responses are simply less consistent / noisier (Figure 2A).

The results are robust and generalize to another dataset. To ensure the robustness of our findings, we used a similar approach on an existing fMRI dataset (Pereira et al., 2018), where we sub-sampled stimuli (n=100 from the full set of 243 and 384 sentences in the two experiments that comprise Pereira2018) in a similar way as we did from a large corpus for the main experiment. In addition, we also sub-sampled a set of 100 sentences such that their representations were maximally *similar* across models (S_{min}), and a set of 100 random sentences (S_{rnd}). This sub-sampling was successful (Figure 2C). If the optimization process leads to different levels of model-to-brain alignment, then we expect a gradient pattern with best model performance on S_{min}, intermediate performance on S_{rnd}, and worst performance on S_{max} sentences. This is the pattern we observed (Figure 2D). This finding strengthens the results from the main experiment and establishes inter-model distance optimization as a new method to control the amount of model-brain alignment.

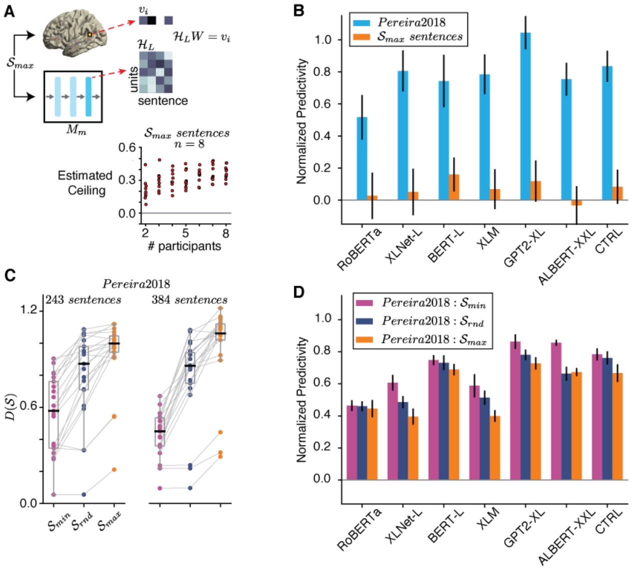

Figure 2: (A) The encoding analysis (top) and subject reliability measure (bottom); (B) Model performance on the optimized S_{max}sentences (orange) and the original Pereira2018 benchmark (blue); (C) Distribution of inter-model distances (each line is a model pair) for the sentences sub-sampled from Pereira2018 (S_{min}purple; S_{rnd}blue; S_{max}orange); (D) Model performance on subsets of sentences from Pereira2018 selected randomly (blue), or to maximize (orange) or minimize (purple) inter-model representational distances.

Discussion

We developed a method for differentiating ANN models' representational spaces as needed for generating and testing hypotheses about the representational structure of the human language network. This approach led to the discovery that a subset of natural sentences constitute a 'blind spot' for multiple ANN language models, such that they struggle to predict human neural responses to those sentences. This finding may therefore reveal features that are differentially represented by the models vs. the brain. We also found that sentences that are represented in a similar way by multiple models lead to stronger model-to-brain alignment, which may help capture aspects of sentence representation that are shared between the models and the brain. Understanding the features of stimuli that lead to strong vs. poor model-brain alignment will reveal dimensions that are used by language models to encode linguistic input and can lead to a better understanding of human neural computations that underlie language comprehension.

Acknowledgments

This work was supported by NIH awards R01-DC016607, R01-DC016950, and U01- NS121471 and research funds from the McGovern Institute for Brain Research, Brain and Cognitive Sciences department, and the Simons Center for the Social Brain to EF. EAH was supported by the Friends of McGovern graduate fellowships. NZ was supported by a K. Lisa Young ICoN Center postdoctoral fellowship

References

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics (Association for Computational Linguistics)*, 1–54.

Golan, T., Taylor, J., Schütt, H. H., Peters, B., Sommers, R. P., Seeliger, K., Doerig, A., Linton, P., Konkle, T., van Gerven, M., & al., E. (2023). *Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses*. https://doi.org/10.31234/osf.io/tr7gx

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16.

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

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*. https://doi.org/10.1016/j.neuron.2020.07.040

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, *18*(7), 1025–1033.

Yamins, D. L. K., 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 of the United States of America*, *111*(23), 8619–8624.