Attention weights accurately predict language representations in the brain

Mathis Lamarre,[1,2] Catherine Chen,[3] Fatma Deniz[1,2,3]

[1]Technische Universität Berlin, [2]Bernstein Center for Computational Neuroscience, [3]University of California, Berkeley

m.lamarre@tu-berlin.de

Introduction

During language comprehension, humans extract the meaning of individual words and integrate the meaning from nearby words to form contextual representations.

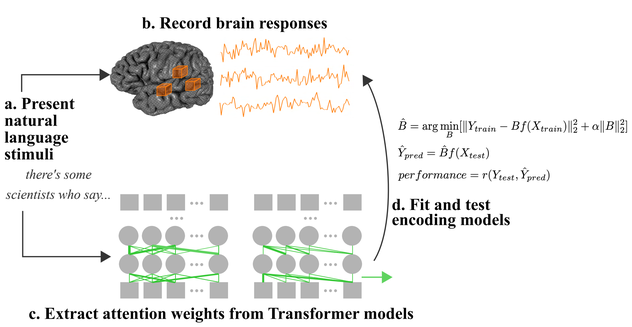
Using word embeddings, previous work showed where these lexical and contextual semantic information are located across the cerebral cortex (Huth et al. 2016, Deniz et al. 2019, Schrimpf et al. 2021). However, it is unclear whether the mechanism that builds the embeddings can inform on brain representations of language.

To study where context integration occurs, we use voxelwise encoding models with attention weights: vectors that reflect the process of context integration in state-of-the-art Transformer language models (LMs).

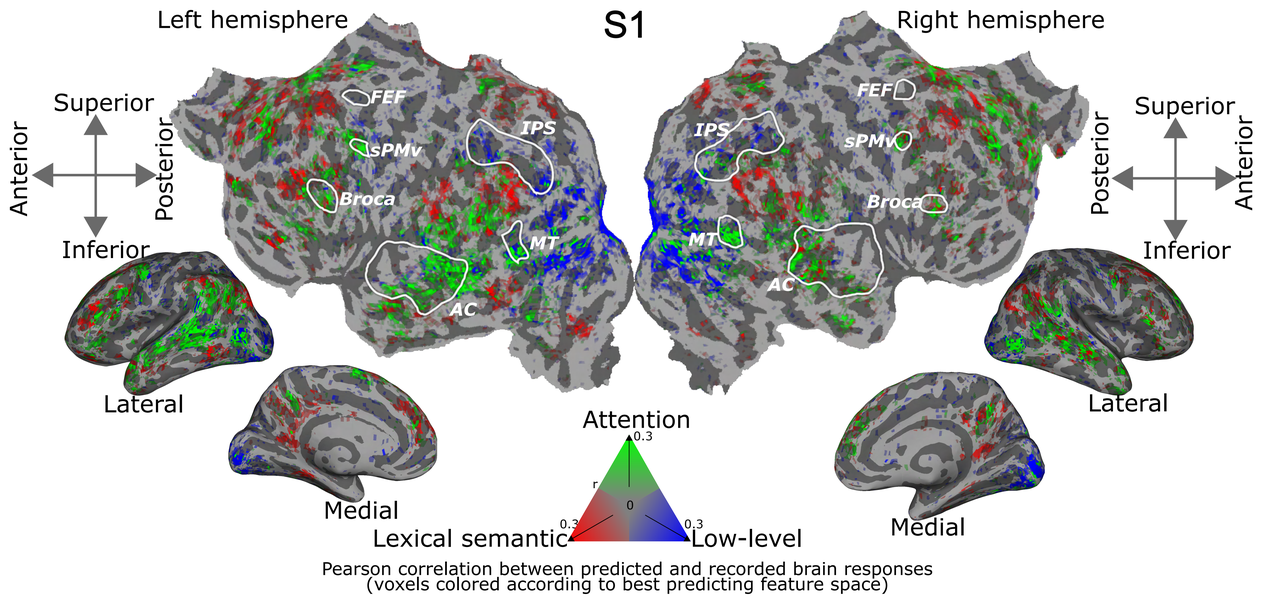
Methods

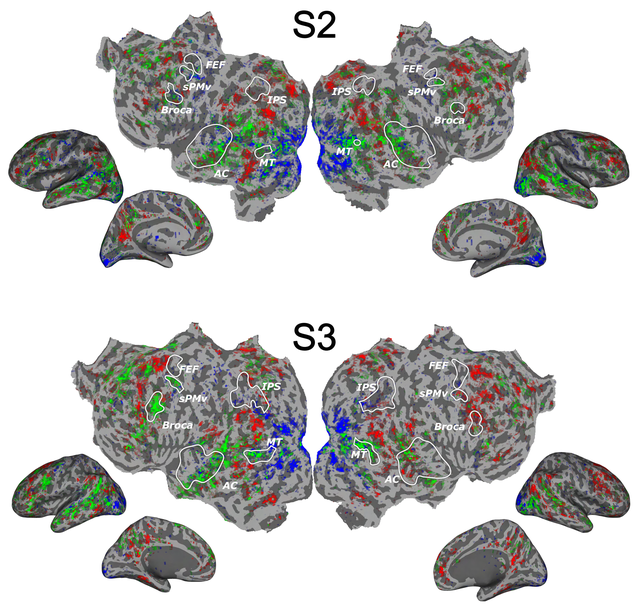
We analyzed fMRI recordings of six participants reading English language narratives. We used voxelwise modeling, where features extracted from the stimulus are used to predict the evoked brain response.

We provided the narrative text as input to two Transformer LMs ("BERT" and "GPT-2", Devlin et al. 2018, Radford et al. 2019) and extracted their attention weights. We then measured how well attention weights can predict the recorded brain responses in each voxel, and compared these predictions to those of lexical and contextual word embeddings. Low-level sensory features were included as nuisance regressors.



Attention weights predict brain responses more accurately than lexical embeddings

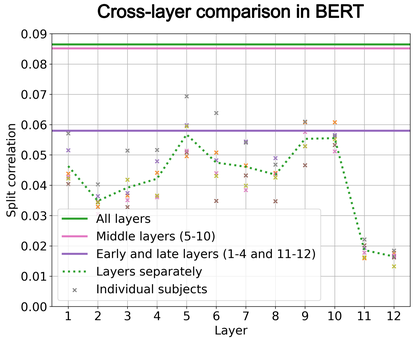




Pearson correlation between predicted and recorded brain responses
(voxels colored according to best predicting feature space)

Analysis of the language models

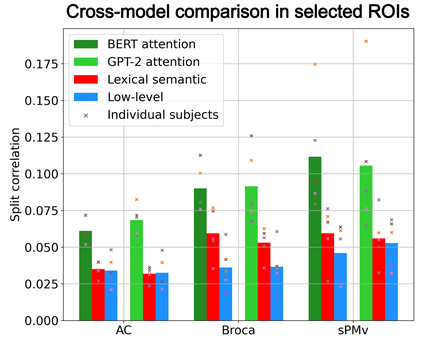
Cross-layer comparison in BERT



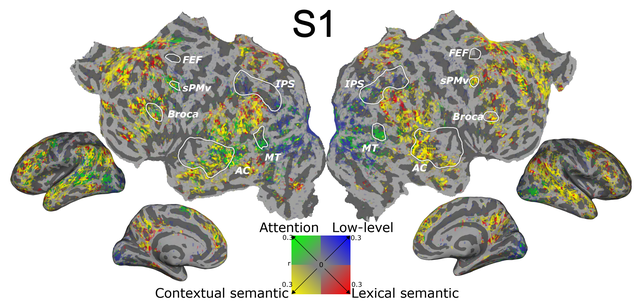
The best predictions come from attention weights in the middle layers of the LMs. Prior work has shown that these layers contain linguistically relevant attention activity (Clark et al. 2019).

Prediction performance is consistent across the two LMs we tested.

Cross-model comparison in selected ROIs



Attention can outperform contextual embeddings



Conclusions

The attention weights of Transformer language models accurately predict brain responses in much of the frontal and temporal cortices. Several of these areas have previously been associated with language processing, such as Broca's area, the high-level auditory cortex (AC), the superior temporal sulcus (STS) and the superior ventral premotor speech area (sPMv).

Attention weights outperform lexical embeddings in most of these areas, and even outperform contextual embeddings in portions of these areas. These portions of the cortex may be linked to context integration.

References

Alexander G. Huth, Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 2016.

Fatma Deniz, Anwar O. Nunez-Elizalde, Alexander G. Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. Journal of Neuroscience, 2019.

Martin Schrimpf, Idan A. Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. PNAS, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI blog, 2019.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. BlackboxNLP, 2019.

Acknowledgements

This work was funded by grants from the National Science Foundation (NAT-1912373) and the German Federal Ministry of Education and Research (BMBF 01GQ1906). CC was supported in part by an NSF GRFP and an IBM PhD fellowship.