Brain-like functional organization in Topographic Transformer models of language processing

Taha Binhuraib (taha@novuswriter.com)
Novus Technologies, 55 Court St, Boston, MA 02108 USA

Greta Tuckute (gretatu@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street Cambridge, MA 02139 USA

Nicholas M. Blauch (nblauch@fas.harvard.edu)
Department of Psychology, Harvard University, 33 Kirkland St, Cambridge MA 02138

Abstract

Topographic organization is a key feature of biological brains. However, representations within most machine learning models lack spatial biases, instead manifesting as disorganized vector spaces that are difficult to visualize and interpret. Here, we make two contributions. First, we introduce a new family of spatially-constrained topographic Transformer ("Topoformer") models. We train a 16-layer Topoformer model on a masked language modeling objective, and demonstrate significant topography in the learned responses. Second, we investigate an fMRI dataset of sentence-level responses to 1,000 sentences and demonstrate that human fronto-temporal language-responsive areas exhibit topographic response variability, variability which shows significant alignment with that of the Topoformer model. Our results motivate further examination of functional topography of language representations in brains and models, along with a task-optimized approach to topographic modeling more generally.

Keywords: Transformer; Topography; Neuroscience; Language

Introduction

Biological brains exhibit spatial organization, such as category-selective areas, broad feature maps, and large-scale networks. Recent computational neuroscience work has modeled the topographic organization of the ventral visual stream in vision DNNs by incorporating local smoothness or wiring cost minimization, resulting in easily visualized smooth functional organization (Margalit et al., 2023; Blauch, Behrmann, & Plaut, 2022). However, despite progress in natural language processing (NLP) and application of these models to study the neural basis of language (Fedorenko, Ivanova, & Regev, 2024), topographical priors have yet to be applied to language processing models. Our work aims to induce topographic organization within the Transformer architecture, using local-connectivity approaches, yielding models we term "Topoformers". Such models pave the way for a computational description of the topographic organization of language-responsive cortex.

Methods

Adding topographic priors to self-attention

To encourage topographic organization in Transformer models, we add spatial biases to two operations of the self-attention operation (Vaswani et al., 2017), as illustrated in Figure 1A. The first modification is spatial querying; here, each token's query is associated with a local pool of queries from other tokens, rather than individual keys. A binary intermediate matrix $M \in \mathbb{R}^{d \times d}$, where d is the embedding dimension and the columns of M determine the spatial pool of queries associating with a given key. This local pooling of queries promotes local smoothness in representations and ensures spatial correspondence between query and key representations.

The second modification is known as spatial reweighting; here, we convert the outer reweighting matrix W^O to a locally connected layer W^O_{local}, with strictly positive weights. This pressures the model towards topographic representations in the values and attention outputs. More detail on spatial querying and reweighting can be found in the equations of Figure 1A.

Model training and evaluation

We trained a single-head 16-layer Topoformer BERT model using the Masked Language Modeling objective (Devlin, Chang, Lee, & Toutanova, 2019). We followed the training paradigm introduced by Geiping and Goldstein (2022) on the Bookcorpus-Wikipedia dataset (Zhu et al., 2015). A standard, non-topographic single-head BERT model with an otherwise identical setup served as a control model. To evaluate the models' performance, we followed the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019) procedure as in (Geiping & Goldstein, 2022).

Neuroimaging experiment

We examined language organization in 5 native English speakers using fMRI during a sentence reading task (Tuckute et al., 2024). Participants read 1,000 diverse 6-word sentences. After preprocessing, we identified language-responsive voxels by comparing responses to sentences and nonwords (from an independent 'localizer' task) using a weak positive threshold ($t > 1$) within a set of 5 broad, anatomical parcels from prior studies (Lipkin et al., 2022), forming the "language network". Within this network, we extracted sentence-level responses in the form of beta coefficients derived from GLMsingle (Prince et al., 2022).

Brain-model alignment

Representational alignment was conducted between the human language network and the final layer Topoformer unit activations using partial least squares singular value decomposition (PLS-SVD). PLS-SVD aligns brain responses X and unit activations Y by computing SVD on their cross-covariance matrix $X^TY = U\Sigma V$, yielding joint low-dimensional embeddings of brain responses and unit activations. Given brain and Topoformer responses, we can visualize the spatial organization of individual brain and model SVD components $U^{(i)}$ and $V^{T(i)}$. Moreover, the alignment of components can be computed as the correlation of $X_c^{(i)} = X_{test}U^{(i)}$ and $Y_c^{(i)} = Y_{test}V^{T(i)}$ using held-out data (20%).

Results

Following training of the Topoformer and a control model, we observed that the task performance of Topoformer on the GLUE benchmark (75.3) was similar to that of the non-topographic model counterpart (76.9), suggesting that our added spatial constraints do not significantly hinder task performance.

To interpret the resulting organization, we performed unit-level selectivity analyses in the Topoformer model, using 8

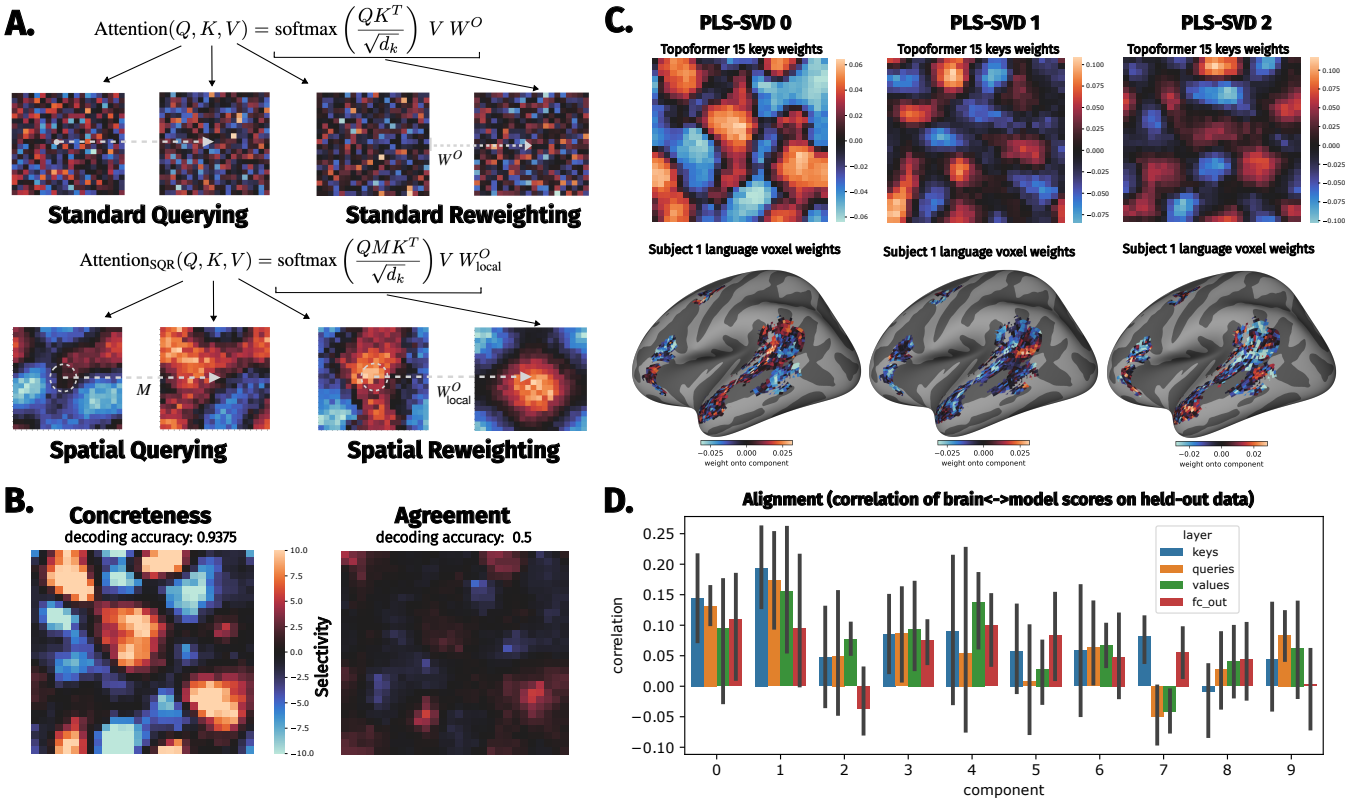

Figure 1: **Brain-like topographic organization in a Topoformer model. A.** Spatial querying and reweighting operations convert non-spatial self-attention operations to spatial versions, encouraging the development of topographic organization. The figure illustrates querying for a pair of tokens and reweighting for a single token, however when processing a full sequence, there is a 2D grid of the form shown here for each token. Each heatmap shows the loadings of the second PC of responses in a 1-layer toy model (top: control model, bottom: Topoformer model with spatial querying and reweighting). **B.** Selectivity of the Topoformer BERT model for two "test suites": "Concreteness" consists of concrete vs. abstract sentences, and "Agreement" consists of syntactically correct sentences vs. manipulating one word to violate the subject-verb agreement of the sentence (Marvin & Linzen, 2018). Selectivity was quantified using the t-statistic with significance p as $s = (-\text{sign}(t)\log_{10}(p))$. Condition decoding from unit activations is given above the heatmaps. **C.,D.** Low-dimensional topographic component alignment revealed with partial-least squares singular value decomposition (PLS-SVD). **C.** plots the weights of PLS-SVD components, visualizing their spatial structure in both the model and brain. **D.** plots the cross-validated alignment of PLS-SVD component scores, with error bars over 5 participants.

test suites, and performed binary condition decoding using population activity. We plot the results for 2 examples (Concreteness and Agreement) in Figure 1B. Whereas significant selectivity and multivariate decoding was seen for concreteness, less was seen for agreement. These results demonstrate a significant topographic sensitivity for concreteness but not syntactic agreement.

Next, using PLS-SVD (see Methods) we examined the alignment of language representations in the human language network and the Topoformer model. Figure 1C shows example aligned component weights between the first three brain and model components, using the first participant and the Topoformer layer 15 keys representation. Figure 1D generalizes this analysis across all participants and sublayers, again using layer 15 of the model. In general, the first two components were significantly aligned for each sublayer, whereas later components were less aligned. This result demonstrates that the low-dimensional variability can be aligned in the topographic representations of the human language network and Topoformer language model.

Conclusion

Our work demonstrates that Transformer models can be trained to exhibit topographic organization with similarity to that of language cortex in the human brain, and highlights the presence of yet unexplored topographic organization within the language network. The results motivate further work both in understanding cortical organization of language, and in yielding interpretable topographic machine learning models for language as well as other domains.

References

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022, jan 18). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(3). Retrieved 2022-01-17, from http://www.pnas.org/lookup/doi/10.1073/pnas.2112566119 doi: 10.1073/pnas.2112566119

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 1–24.

Geiping, J., & Goldstein, T. (2022). *Cramming: Training a language model on a single gpu in one day.*

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H. H., ... Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fmri data from ¿800 individuals. *Scientific Data*, *9*.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. K. (2023). A unifying principle for the functional organization of visual cortex. *bioRxiv*, 2023.05.18.541361.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022, February). *GLMsingle: a toolbox for improving single-trial fMRI response estimates* (Tech. Rep.). bioRxiv. Retrieved 2022-03-24, from https://www.biorxiv.org/content/10.1101/2022.01.31.478431v1 (Section: New Results Type: article) doi: 10.1101/2022.01.31.478431

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., ... Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 1–18.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 2017-Decem, p. 5999–6009). Retrieved 2021-08-25, from https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=rJ4km2R5t7

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015, June). *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books.* arXiv. Retrieved 2023-04-16, from http://arxiv.org/abs/1506.06724 (arXiv:1506.06724 [cs]) doi: 10.48550/arXiv.1506.06724