



HAL
open science

Probing brain context-sensitivity with masked-attention generation

Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, Christophe Pallier

► **To cite this version:**

Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, Christophe Pallier. Probing brain context-sensitivity with masked-attention generation. CCN 2023 - Conference on Cognitive Computational Neuroscience, Aug 2023, Oxford, United Kingdom. 10.48550/arXiv.2305.13863 . hal-04285201

HAL Id: hal-04285201

<https://hal.science/hal-04285201>

Submitted on 14 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Probing Brain Context-Sensitivity with Masked-Attention Generation

Alexandre Pasquiou

(alexandre.pasquiou@inria.fr)

UNICOG, Cognitive Neuroimaging Unit, INSERM, CEA, Neurospin
Gif-sur-Yvette, France

MIND, INRIA, CEA, Neurospin
Gif-sur-Yvette, France

Yair Lakretz

UNICOG, Cognitive Neuroimaging Unit, INSERM, CEA, Neurospin
Gif-sur-Yvette, France

Bertrand Thirion

MIND, INRIA, CEA, Neurospin
Gif-sur-Yvette, France

Christophe Pallier

UNICOG, Cognitive Neuroimaging Unit, INSERM, CEA, Neurospin
Gif-sur-Yvette, France

Abstract

Two fundamental questions in neurolinguistics concerns the brain regions that integrate information beyond the lexical level, and the size of their window of integration. To address these questions we introduce a new approach named *masked-attention generation*. It uses GPT-2 transformers to generate word embeddings that capture a fixed amount of contextual information. We then tested whether these embeddings could predict fMRI brain activity in humans listening to naturalistic text. The results showed that most of the cortex within the language network is sensitive to contextual information, and that the right hemisphere is more sensitive to longer contexts than the left. Masked-attention generation supports previous analyses of context-sensitivity in the brain, and complements them by quantifying the window size of context integration per voxel.

Keywords: fmri;transformers; context;brain;encoding

Introduction

Following the works of Bemis & Pykkänen (2011, 2013), a few studies have tried to leverage computational models to identify the neural bases of compositionality and quantify brain regions' sensitivity to increasing sizes of context. Some of them, using ecological paradigms, have found a hierarchy of brain regions that are sensitive to different types of contextual information and different temporal receptive fields (e.g., Jain & Huth, 2018; Toneva et al., 2022; Wehbe et al., 2014). A notable investigation (Jain & Huth, 2018) used pre-trained LSTM (Hochreiter & Schmidhuber, 1997) models to study context integration. They varied the amount of context used to generate word embeddings, and obtained maps indicating brain regions' sensitivity to different sizes of context. In this work, we study context-sensitivity using the attention mechanisms of GPT-2 which better integrate context than LSTMs.

Methods

fMRI Brain data. The brain data consisted of the functional Magnetic Resonance Imaging (fMRI)

scans from the English participants of *The Little Prince* fMRI Corpus (Li et al., 2022)¹.

Modelling Context-limited Features with GPT-2 using attention masks. Contextual information was controlled by playing with the attention mechanisms of the GPT-2 (Radford et al., 2019)² transformer. The method involves providing the model with an input sequence and attention mask pair for each word in the text, and retrieving the target word's embedding for each pair. An example is given in Fig. 1 for a context-window size of 4.

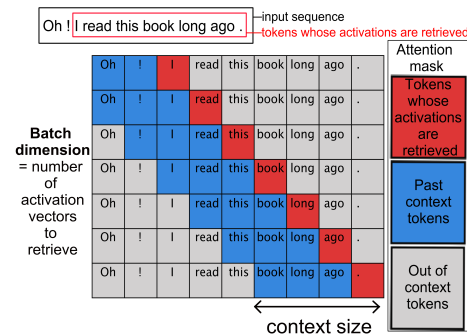


Figure 1: **Controlling tokens' interaction using attention masks.** Examples of (input sequence, attention mask) pairs to retrieve the embedding of each word of the target sentence (framed in red above). An input sequence is represented by a row, the target token is colored in red, tokens in the attention mask are blue or red (context size = 4), and out-of-context tokens are grey.

The attention mask removed interactions with words outside the window, while preserving interactions within the context window (see Fig. 1). The mask was a binary vector containing 0 except for the target word and the previous $n-1$ words, where it equaled 1. It preserved the positional encoding of words in the sentence and the right use of the special tokens, while using complete sentences. The attention mask is the same for all the tokens in the input sequence, modulo the incrementality. Otherwise, information could propagate outside the context window because of model's depth.

Encoding models. The same encoding approach as Pasquiou et al. (2022, 2023) was used.

¹Available from <https://openneuro.org/datasets/ds003643/versions/1.0.2>

²Available from <https://huggingface.co/gpt2>

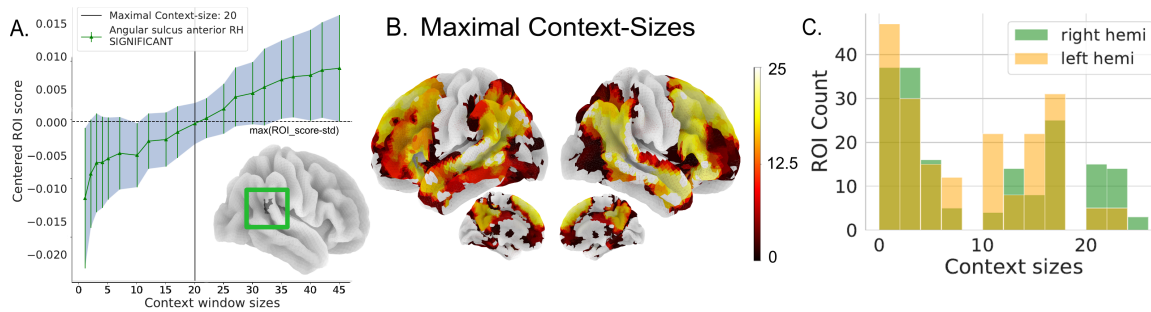


Figure 2: **Assessing the maximal context window size over which information is integrated.** A) Determination of the maximal context-size for each parcel of the Difumo atlas. The maximal context-size is defined as the last context-size whose ROI-score is inferior to the maximal averaged centered ROI-score minus its standard deviation. B) Surface projection of Difumo’s parcels maximal context-size in context-sensitive brain regions. C) Histograms representing the maximal context sizes distribution across context-sensitive ROIs, in the left hemisphere (orange), and the right hemisphere (green).

For each context-window size (21 values sampled between 1 and 45 tokens), the embeddings from layer 9 of the 12-layer model (dim=768) were used to fit each subject’s brain data (N=51). Then, we examined the impact of the context-window size on the models’ predictive performance (R scores). The motivation behind this approach is the following. If the model needs short-range information to build the embedding of a word, then the embedding won’t be affected when using a small context size. However, if the model needs long-range information, the embedding will be ‘damaged’ when using a small context size. Thus, increasing context size won’t improve R scores in the brain regions well-fitted by features using short-range information. However, brain regions well-fitted by features using long-range information will benefit from increasing context size.

DIFUMO atlas. We computed the median R score across voxels constituting 90% of the non-zero loadings of each parcel of the Difumo atlas (Dadi et al., 2020) (referred to as ROI -score).

Assessing brain regions sensitivity to context. For each participant and ROI, we fitted a Linear Regression on the (context_size, ROI-score) points to get the slope of increase of the ROI-score as a function of context-size. Brain regions’ context-sensitivity was estimated with a t-test on the slopes of increase across subjects, with a FDR correction of 0.01 to account for multiple comparisons.

Quantifying the window-size over which con-

text is integrated. For each context-sensitive parcel of the atlas, we estimated its *maximal context-size*, i.e. the last context-window size over which the ROI-score is less than one standard deviation away from its maximal value (Fig. 2A). Maximal context-sizes are reported in Fig. 2B.

Results & Discussion

First, **most of the language related brain regions are context-sensitive** (Fig. 2B). This network of context-sensitive brain regions is bilateral and mostly symmetrical. Notes that low-level regions such as the auditory, motor and visual cortices are not context-sensitive. These findings support the ones from Jain & Huth (2018).

Additionally, we found that **the right hemisphere shows sensitivity to longer contexts than the left** (Fig. 2C). The brain regions integrating longer-context revolves around the Temporo-Parietal Junction, Superior frontal regions and medial regions. This observation is consistent with other brain imaging studies that have supported the role of the right hemisphere in higher-level language tasks (see Beeman & Chiarello (2013); Jung-Beeman (2005)). Overall, our results show that modifications of language models’ architecture (e.g., unit ablation), or internal operations (e.g., modification of the attention mechanisms) can be used to probe precise linguistic processes.

Acknowledgments

This project/research has received funding from the American National Science Foundation under Grant Number 1607441 (USA), the French National Research Agency (ANR) under grant ANR-14-CERA-0001, the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3), and the KARAIB AI chair (ANR-20-CHIA-0025-01).

References

- Beeman, M. J., & Chiarello, C. (2013). *Right hemisphere language comprehension: Perspectives from cognitive neuroscience*. Psychology Press.
- Bemis, D. K., & Pykkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience*, *31*, 2801–2814.
- Bemis, D. K., & Pykkänen, L. (2013, August). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cereb. Cortex*, *23*(8), 1859–1873.
- Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K. J., Wassermann, D., Thirion, B., & Mensch, A. (2020). Fine-grain atlases of functional modes for fmri analysis. *NeuroImage*, *221*, 117126. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1053811920306121> doi: <https://doi.org/10.1016/j.neuroimage.2020.117126>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
- Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Retrieved from <https://proceedings.neurips.cc/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf>
- Jung-Beeman, M. (2005, November). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, *9*(11), 512–518. Retrieved 2015-02-12, from <http://linkinghub.elsevier.com/retrieve/pii/S1364661305002718> doi: 10.1016/j.tics.2005.09.009
- Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, N., ... Hale, J. (2022). Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, *9*. Retrieved from <https://doi.org/10.1038/s41597-022-01625-7>
- Pasquiou, A., Lakretz, Y., Hale, J. T., Thirion, B., & Pallier, C. (2022). Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps. In *Proceedings of the 39th international conference on machine learning (icml)* (Vol. 162, pp. 17499–17516). Retrieved from <https://arxiv.org/abs/2207.03380>
- Pasquiou, A., Lakretz, Y., Thirion, B., & Pallier, C. (2023). *Information-restricted neural language models reveal different brain regions' sensitivity to semantics, syntax and context*.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *arxiv*, 24.
- Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). Combining computational controls with natural text reveals new aspects of meaning composition. *BioRxiv*, 2020–09.
- Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014, October). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 233–243). Doha, Qatar: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/D14-1030> doi: 10.3115/v1/D14-1030