A Simple Untrained Recurrent Attention Architecture Aligns to the Human Language Network

Badr AlKhamissi (badr.alkhamissi@epfl.ch)
School of Computer and Communication Sciences, School of Life Sciences, NeuroX Institute
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Antoine Bosselut[†] (antoine.bosselut@epfl.ch)
School of Computer and Communication Sciences
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

Martin Schrimpf[†] (martin.schrimpf@epfl.ch)
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL (Swiss Federal Institute of Technology), Lausanne, 1015 Switzerland

[†] Equal Supervision.

Abstract

Certain Large Language Models (LLMs) are effective models of the Human Language Network, predicting most explainable variance of brain activity in current datasets. Even with architectural priors alone and no training, model representations remain highly aligned to brain data. In this work, we investigate the key architectural components driving this surprising alignment of untrained models. To estimate LLM-to-brain similarity, we first select language-selective units within an LLM, similar to how neuroscientists identify the language network in the human brain. We then benchmark the brain alignment of these LLM units across three neural datasets and three metrics. Building a model architecture from the ground up, we identify that token aggregation is a key component driving the similarity of untrained models to brain data. Increased aggregation via multi-headed attention significantly increases brain alignment, and, for longer contexts in particular, adaptive aggregation via recurrence further boosts model similarity to neural activity. We summarize our findings in a simple untrained recurrent transformer model that achieves near-perfect brain alignment.

Keywords: Brain Alignment; Large Language Models; Human Language Network; Functional Localization; Recurrence

Introduction

Unraveling the neural mechanisms underlying language processing in the human brain has been a longstanding challenge in neuroscience. Driven by recent advances in machine learning, large language models (LLMs) trained via next-word prediction, are now a particularly promising model family to capture the internal processing of the human language network. When exposed to the same text stimuli (e.g., words and sentences) as human participants during neuroimaging and electrophysiology sessions, the most brain-like LLMs predict most of the variance of neural responses relative to the estimated noise ceiling (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022). A peculiar observation with LLMs as models of the brain is that untrained models can exhibit internal representations that are nearly as brain-like as those of their trained counterparts (Schrimpf et al., 2021). We here explore the model components underlying the high alignment of untrained models and identify the aggregation of input tokens as the primary factor driving this model-to-brain similarity. Codifying our findings, we propose a simple untrained recurrent Transformer model with near-perfect alignment to the human language network under current benchmarks.

Localization of the Language Network

Selecting the appropriate layer or set of units in a language model for comparison against the human language network is a crucial step that significantly impacts the final alignment. Ideally, each model should possess a fixed set of neurons designated as its language network, independent of the dataset or metric used. Previous studies have assessed brain alignment by analyzing the output of each block in a Transformer model and selecting the maximum alignment as the final score (Schrimpf et al., 2021). We here propose an approach that more closely follows the methodology employed by neuroscientists to localize the language network in the brain – the human language network is defined as the set of units (e.g. voxels/electrodes) that are more selective to sentences over perceptually-matched controls Fedorenko et al. (2010), and we characterize the model language network in the same way.

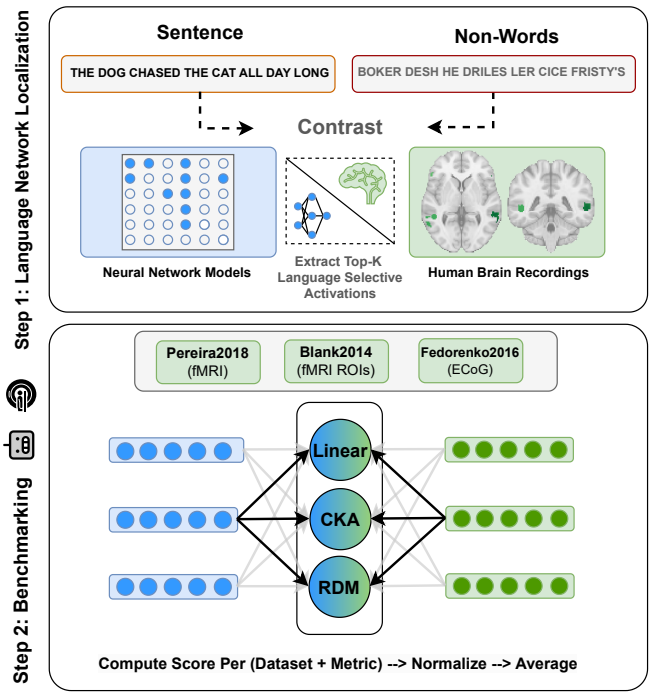

Figure 1: **Comparing models to the human language network.** (**Top**) We first select the top-k language selective units in models and brain recordings by contrasting the difference in unit activations between sentences and lists of non-words, following Fedorenko et al. (2010). (**Bottom**) We then measure the alignment between the language selective units in the model and the language network in the brain on three datasets and three metrics. Model scores are reported as the mean across all these nine benchmark scores after normalizing them relative to the estimated cross-subject consistency.

Specifically, we present a set of sentences and lists of non-words to each model, obtaining activations for each stimulus from all units. We then define the model language network as the top-k units (here $k = 4096$) that maximize the difference between sentence activations and non-word activations (Figure 1). This localization method selects a distributed set of units from across the entire network, rather than constraining the choice of representations to a single layer as in prior work.

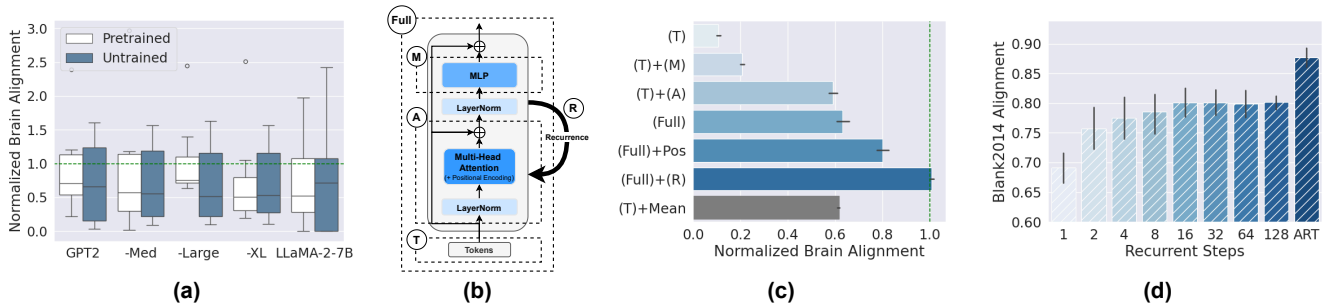

Figure 2: **Untrained models achieve high brain alignment via token aggregation.** Alignment is evaluated on each model's top $4,096$ language-selective units (Figure 1). The green dashed line indicates the cross-subject consistency estimate. **(a)** Brain alignment scores for various pretrained models and their untrained counterparts. Plots show the mean and distribution of scores on individual benchmarks. For untrained models, each benchmark score is averaged over 5 initializations sampled from the same distribution. **(b)** Building a transformer block from the ground up, components are labeled for the ablation study in (c). **(c)** Brain alignment of a single untrained transformer block with different ablations. We use each model's representations to the last token, except for `(T)+Mean`. Labels refer to (b). **(d)** Brain alignment of the `(Full)+(R)` model on the `Blank2014` dataset (story stimuli) as a function of the number of recurrent steps. `ART` refers to adaptive recurrence relative to the number of tokens.

Benchmarks

Datasets To evaluate model alignment to the human language network, we use three brain recording datasets: functional magnetic resonance imaging (fMRI) data from Pereira et al. (2018) in subjects reading short passages, electrocorticography (ECoG) data from Fedorenko et al. (2016) in subjects reading sentences one word at a time, and fMRI data aggregated into functional regions of interest from Blank et al. (2014) in subjects listening to ~ 5-minute long naturalistic stories. We use the datasets as packaged in **`Brain-Score`** (Schrimpf et al., 2018, 2020) which include only those voxels and electrodes that are selective to language (Figure 1).

Metrics Alignment between model predictions and brain data can be tested in different ways and since each metric might focus on different aspects of the data, we employ three common metrics in the field. Specifically, we use `Linear Predictivity` with an ordinary least-squares linear regression following Schrimpf et al. (2021), Centered Kernel Alignment (`CKA`) (Kornblith et al., 2019), and Representational Dissimilarity Matrices (`RDM`) (Kriegeskorte et al., 2008).

Estimation of Cross-Subject Consistency To estimate the similarity and potential noise of brain recordings, we compute each benchmark's *cross-subject consistency*—referred to as noise ceiling in previous work. For benchmarks with a `Linear Predictivity` metric we estimate the consistency by predicting the brain activity of one held-out subject from all other subjects. For non-parametric metrics (`CKA` and `RDM`), we compute the similarity between two halves of subjects over all possible combinations. The model score on each benchmark, i.e. each dataset-metric pair, is normalized with the cross-subject consistency estimate ($\frac{\text{raw score}}{\text{consistency}}$) and we report the final score for each model as the average across all nine benchmarks (three datasets $\times$ three metrics).

Results & Discussion

Untrained Models Can Exhibit High Brain Alignment Evaluating the brain alignment for pretrained models from the `GPT-2` family (Radford et al., 2019) and the `LLaMA-2-7B` model (Touvron et al., 2023) as well as their untrained counterparts as initialized by the HuggingFace library (Wolf et al., 2019), we find that scores of untrained models are consistently similar to trained models (Figure 2(a); Welch's t-test $t = 0.97$, $p = 0.33$).

Token Aggregation is Driving Brain Alignment Ablating the components of a single untrained causal transformer block and measuring the resulting model's brain alignment, we find that increased aggregation over untrained token embeddings improves brain alignment (Figure 2(b,c)). For instance, the increased alignment from the attention mechanism (`(T)+(A)`) is comparable to a simple mean over input tokens (`(T)+Mean`).

Adaptive Recurrent Transformer (ART) To further aggregate over input tokens, we add recurrence to the untrained single transformer block (Figure 2(b) `R`). This model adapts its computational depth according to the sequence length, allocating more computation for longer sequences ($\lceil \frac{\#\text{tokens}}{8} \rceil$). Adaptively aggregating token embeddings by iterative application of the attention mechanism leads to very high brain alignment with large gains over previous state-of-the-art on especially the `Blank2014` stories dataset (Figure 2(c,d)).

Conclusion

Our results suggest that the right inductive biases in an architecture alone yield representations that are highly aligned to a suite of brain recordings of the human language network. We synthesize our findings into a simple untrained recurrent attention architecture with high brain alignment.

References

Blank, I., Kanwisher, N., & Fedorenko, E. (2014, September). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, *112*(5), 1105–1118. Retrieved 2023-09-19, from https://www.physiology.org/doi/10.1152/jn.00884.2013 doi: 10.1152/jn.00884.2013

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

Fedorenko, E., Hsieh, P.-J., Nieto-Castanon, A., Whitfield-Gabrieli, S. L., & Kanwisher, N. G. (2010). New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, *104 2*, 1177-94. Retrieved from https://api.semanticscholar.org/CorpusID:740913

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256-E6262. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1612132113 doi: 10.1073/pnas.1612132113

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... Hasson, U. (2022, March). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. Retrieved 2023-09-27, from https://www.nature.com/articles/s41593-022-01026-4 doi: 10.1038/s41593-022-01026-4

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. Retrieved from https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008 doi: 10.3389/neuro.06.004.2008

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., ... Fedorenko, E. (2018, March). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963. Retrieved 2023-06-19, from https://www.nature.com/articles/s41467-018-03068-4 doi: 10.1038/s41467-018-03068-4

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2105646118 doi: 10.1073/pnas.2105646118

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018, September). *Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?* (preprint). Neuroscience. Retrieved 2023-06-20, from http://biorxiv.org/lookup/doi/10.1101/407007 doi: 10.1101/407007

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*, *108*(3), 413-423. Retrieved from https://www.sciencedirect.com/science/article/pii/S089662732030605X doi: https://doi.org/10.1016/j.neuron.2020.07.040

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, *abs/2307.09288*. Retrieved from https://api.semanticscholar.org/CorpusID:259950998

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, *abs/1910.03771*. Retrieved from https://api.semanticscholar.org/CorpusID:208117506