What aspects of NLP models and brain datasets affect brain-NLP alignment?

Subba Reddy Oota, Mariya Toneva

What aspects of NLP models and brain datasets affect brain-NLP alignment?

Subba Reddy Oota (subba-reddy.oota@inria.fr)
Inria Bordeaux, France

Mariya Toneva (mtoneva@mpi-sws.org)
MPI for Software Systems, Saarbrücken, Germany

Abstract

Recent brain encoding studies highlight the potential for natural language processing models to improve our understanding of language processing in the brain. Simultaneously, naturalistic fMRI datasets are becoming increasingly available and present even further avenues for understanding the alignment between brains and models. However, with the multitude of available models and datasets, it can be difficult to know what aspects of the models and datasets are important to consider. In this work, we present a systematic study of the brain alignment across five naturalistic fMRI datasets, two stimulus modalities (reading vs. listening), and different Transformer text and speech models. We find that all text-based language models are significantly better at predicting brain responses than all speech models for both modalities. Further, bidirectional language models better predict fMRI responses and generalize across datasets and modalities.

Keywords: Transformers, language models, speech models, fMRI, reading, listening, encoding models

Introduction

The increasing availability of naturalistic fMRI datasets and the use of large-scale neural models can enable a better understanding of the brain's response to natural stimuli. Just in the last few years, researchers have shown that brain responses of people comprehending language can be predicted well by text-based language models (Wehbe et al., 2014; Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2020; Schrimpf et al., 2021), as well as speech-based models (Nishida et al., 2020; Millet et al., 2022; Vaidya, Jain, & Huth, 2022; Tuckute, Feather, Boebinger, & McDermott, 2022). However, with the multitude of available models and datasets, it can be difficult to know what characteristics of the models and datasets are important to account for. Is the choice of stimulus modality (reading vs. listening) important for the study of brain alignment? Are all naturalistic fMRI datasets equally good for brain encoding? How does the type of model (text vs. speech and encoder vs. decoder) affect the resulting alignment?

In this work, we present a systematic study of the brain alignment across five popular naturalistic stories fMRI datasets (2-reading, 3-listening) and different natural language processing models (text vs. speech, unidirectional vs. bidirectional). We find that text-based pretrained language models outperform speech models in predicting brain responses elicited by stimuli in both modalities (i.e. reading vs. listening). Furthermore, both language and speech models have similar brain prediction performance in the auditory cortex, but language models outperform speech models in the remaining language regions. Further, we find that the average estimated noise ceiling across participants during listening and reading for the same stimuli is similar, and the normalized predictivity (i.e. the fraction of ceiling the neural model can predict) is similar in the two modalities.

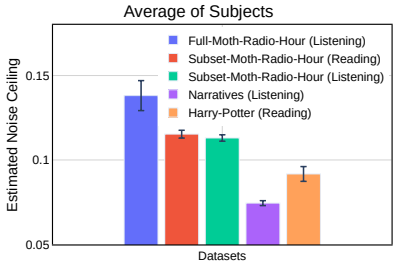

Figure 1: The estimated noise ceiling was computed across all subjects for each of the 5 naturalistic datasets. The average noise ceiling is shown across significantly predicted voxels (i.e. p-value < 0.05).

Table 1: Naturalistic Stories Datasets

Dataset	Modality	Subj	1-TR	# TRs
Full-Moth-Radio-Hour	Listening	8	2.0045s	9932
Subset-Moth-Radio-Hour	Reading	6	2.0045s	4028
Subset-Moth-Radio-Hour	Listening	6	2.0045s	4028
Narratives (21^{st}-Year)	Listening	18	1.5s	2250
Harry-Potter	Reading	8	2s	1211

Table 2: Neural Pretrained Transformer Models

Model Name	Pretraining	Type	Layers
BERT-base-uncased	Text	Encoder (Bidirectional)	12
GPT2-Small	Text	Decoder (Unidirectional)	12
BART-base	Text	Encoder-Decoder	12
FLAN-T5-base	Text	Encoder-Decoder	24
Wav2Vec2.0-base	Speech	Encoder	12
Whisper-small	Speech	Encoder-Decoder	24

Methodology

Datasets We use publicly available fMRI datasets which were recorded while human subjects *read* (Harry-Potter (Wehbe et al., 2014)), and *listened* to (Moth-Radio-Hour (LeBel et al., 2022) and Narratives (Nastase et al., 2021)). We include one more dataset as a comparison, which contains fMRI recordings elicited by reading and listening to the same stimuli (Deniz, Nunez-Elizalde, Huth, & Gallant, 2019). These datasets are summarized in Table 1.

Stimulus Representations To simultaneously test both text and speech representations and their alignment with brain recordings, we use 6 popular pretrained Transformer models: 4 text-based language models (BERT (Devlin, Chang, Lee, & Toutanova, 2019), GPT2 (Radford et al., 2019), BART (Lewis et al., 2020), and FLAN-T5 (Chung et al., 2022)) and 2 speech models (Wav2Vec2.0 (Baevski, Zhou, Mohamed, & Auli, 2020) and Whisper (Radford et al., 2022)). The details of each model are reported in Table 2.

Encoding Model We trained ridge regression based encoding models to predict the fMRI brain activity associated with the stimulus representations obtained from both text and speech models. Each voxel value is predicted using a separate ridge regression model. Formally, at the time step (t), we encode the stimuli as $X_t \in \mathbb{R}^{N \times D}$ and brain region voxels $Y_t \in \mathbb{R}^{N \times V}$, where N denotes the number of training examples, D denotes the dimension of the concatenation of delayed 6 TRs, and V denotes the number of voxels.

Noise Ceiling To account for the intrinsic noise in biological measurements and obtain a more accurate estimate of the model's performance, we estimate the noise ceiling in all

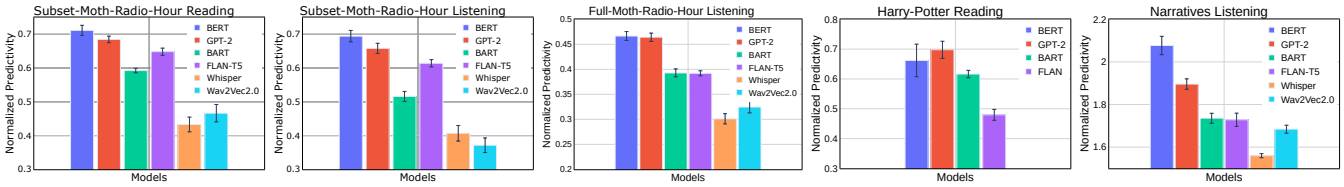

Figure 2: Average normalized brain predictivity was computed over the average of subjects for each model, across layers (5 datasets, 3 language models, 2 modalities (reading and listening), and 2 speech models).

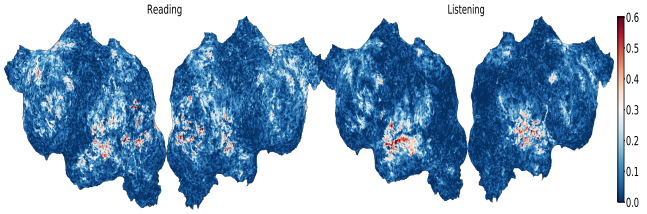

Figure 3: Noise Ceiling performance for subject-2 in reading (left) and listening (right). In reading, the voxels are distributed in visual and language brain regions. In listening, the voxels are distributed in auditory and language brain regions.

datasets (Schrimpf et al., 2020). This is achieved by estimating the amount of brain response in one subject that can be predicted using only data from other subjects, using an encoding model.

Results

Estimated Noise Ceiling Fig 1 displays the mean estimated noise ceiling across voxels for different naturalistic datasets. We observe that: (i) For the *Subset-moth-radio-hour* dataset where subjects are reading and listening to the same stimulus, the average estimated noise ceiling across voxels for the two modalities is not significantly different. However, Fig 3 shows that even though there is no significant difference on average across voxels, there are clearly regional differences. This finding agrees with previous work (Deniz et al., 2019), which found that the sensory regions process single modality information related to low-level processing of speech (early auditory areas) or reading words (early visual areas), and that the higher-level semantic representations are more invariant to the modality in high-level regions. (ii) For the listening datasets, the estimated noise ceiling is higher for Moth-Radio-Hour compared to Narratives 21^{st}-year. This difference in noise ceiling may be due to several subjects who have lower noise ceilings for the 21^{st}-year dataset.

Text vs Speech model prediction performance for different modalities To investigate whether text and speech Transformer models encode the brain activity in a modality-independent way, we compared the brain alignment of text and speech models during listening vs. reading. In Fig 2, we report the brain alignment of each model normalized by the noise ceiling for each dataset. We show the mean normalized brain alignment across subjects, layers, and voxels. We perform the *Wilcoxon signed-rank* test to test whether the differences between text and speech models are statistically significant. We found that all text models are statistically significantly better at predicting brain responses than all speech models in both modalities. Since we observed regional differences between the noise ceilings in the reading and listening conditions, there may also be an effect of where text and speech models predict brain activity best. Specifically, we observe that speech and text-based language models have similar normalized predictivity in the auditory cortex voxels. On the other hand, text-based language models align significantly better than speech models with the remaining language regions. It is possible that language models are better able to capture the complex and abstract linguistic structures that are processed in these specialized regions, while speech models may focus more on acoustic features of speech sounds that are processed in the more general auditory cortex.

Bidirectional vs. Unidirectional model differences To determine the performance differences between bidirectional (encoder) and unidirectional models (decoder), we compared the brain alignment across 5 different naturalistic datasets covering two modalities (reading and listening), as shown in Fig 2. We observe that BERT (bidirectional model) shows higher normalized predictivity than GPT-2 (unidirectional language model). We perform the *Wilcoxon signed-rank* test to test for significant differences between the BERT and GPT-2 brain alignment across participants and find that BERT outperforms GPT-2 in all datasets but Harry-potter, in which the differences are not statistically significant. One previous study that used different brain recordings (fMRI and ECoG) has reported that GPT-2 achieved higher normalized predictivity than BERT (Schrimpf et al., 2021). The discrepancy between our findings may be due to the use of different fMRI recordings and needs to be further investigated. Additionally, we find that models with both bidirectional and unidirectional architectures show lower normalized predictivity than only bidirectional or unidirectional models.

Discussion and Conclusion

We present a systematic study of a number of characteristics of natural language models and fMRI datasets that affect brain alignment. We found that text models predict fMRI recordings significantly better than speech models, irrespective of whether the fMRI recordings were recorded while subjects were listening or reading. We further show that bidirectional text-based language models yield higher normalized brain predictivity and generalize across datasets and modalities. We also find that the average estimated noise ceiling across participants during listening and reading of the same stimulus is similar on average across the brain, but shows regional differences. Further research & analysis are necessary to identify which brain regions are most important for different aspects of speech and language processing and how different neural models can capture the neural activity in these regions.

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*.

Caucheteux, C., & King, J.-R. (2020). Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... others (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*.

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

Jain, S., & Huth, A. G. (2018). Incorporating context into language encoding models for fmri. In *Nips* (pp. 6629–6638).

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., ... Huth, A. G. (2022). A natural language fmri dataset for voxelwise encoding models. *bioRxiv*, 2022–09.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., ... King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *arXiv:2206.01685*.

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., ... others (2021). The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*(1).

Nishida, S., Nakano, Y., Blanc, A., Maeda, N., Kado, M., & Nishimoto, S. (2020). Brain-mediated transfer learning of convolutional neural networks. In *Aaai* (Vol. 34, pp. 5281–5288).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

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

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2020). The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *BioRxiv*.

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*.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv preprint arXiv:1905.11833*.

Tuckute, G., Feather, J., Boebinger, D., & McDermott, J. H. (2022). Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence. *bioRxiv*.

Vaidya, A. R., Jain, S., & Huth, A. G. (2022). Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*(11).