Modeling Brain Responses to Video Stimuli Using Multimodal Video Transformers

Dota Tianai Dong (tianai.dong@mpi.nl)
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, the Netherlands.

Mariya Toneva (mtoneva@mpi-sws.org)
Max Planck Institute for Software Systems
Campus E1 5, 66123 Saarbrücken, Germany

Abstract

Prior work has shown that internal representations of artificial neural networks can significantly predict brain responses elicited by unimodal stimuli (i.e. reading a book chapter or viewing static images). However, the computational modeling of brain representations of naturalistic video stimuli, such as movies or TV shows, still remains underexplored. In this work, we present a promising approach for modeling vision-language brain representations of video stimuli by a transformer-based model that represents videos jointly through audio, text, and vision. We show that the joint representations of vision and text information are better aligned with brain representations of subjects watching a popular TV show. We further show that the incorporation of visual information improves brain alignment across several regions that support language processing.

Keywords: encoding models; multimodal artificial neural networks; fMRI; naturalistic video stimuli

Introduction

How do humans integrate the information of naturalistic video stimuli across multiple modalities? Understanding the complex vision-language representations of video stimuli requires appropriate computational modeling. The impressive performance of artificial neural networks in various tasks has encouraged neuroscientists to leverage their representations as a rich source of stimulus features to model human brain representations (Jain, Vo, Wehbe, & Huth, 2023). Previous work has focused on predicting brain responses using the models trained on unimodal data, such as text, audio, or static images when the same stimuli are presented (Toneva & Wehbe, 2019; Vaidya, Jain, & Huth, 2022; Schrimpf et al., 2018). Recently, researchers have begun aligning representations of vision-language models (trained on images with captions) with brain recordings of subjects viewing static real images, showing that the information learned from one modality enhances the alignment of the brain regions that support another modality (Wang, Kay, Naselaris, Tarr, & Wehbe, 2022; Reddy Oota, Arora, Rowtula, Gupta, & Bapi, 2022).

In contrast to previous work that typically models brain representations of participants observing unimodal stimuli, we investigate the alignment of model representations from a multimodal video transformer with brain recordings in a completely multimodal task setting, namely subjects watching a TV show. Perhaps closest to our work is the one by (Lahner et al., 2023) which studies the correspondences between brain activities and representations of deep neural networks when processing short video clips. However, their models lack multimodality, thus failing to capture the multimodal features needed to model the brain responses to video stimuli.

We instead propose to use MERLOT Reserve, one of the state-of-art multimodal video transformer models, which provides strong contextualized representations of a given video – jointly reasoning over video frames, text, and audio (Zellers et al., 2022). It was pre-trained on 20 million YouTube videos by learning to predict the correct snippet of text (and audio) based on contextualized representations of a video.

Methods

Model. We use the ”base” MERLOT Reserve, which consists of one 12-layer joint encoder with a hidden size of 768, which combines the outputs of 3 independent unimodal encoders: a 12-layer image encoder, a 12-layer audio encoder, a 4-layer text span encoder. In this work, we specifically focus on vision-language representations and leave audio information for future work. Thus we provide video frames and associated subtitles followed by a ’MASK’ token as the inputs. Each video is split into a sequence of non-overlapping 5-second segments in time, and each segment contains a video frame and associated text spans (due to the model limitation, at most three words are reserved for each frame). The model first encodes each modality in one segment independently using an image encoder or text span encoder and then fuses all representations for all modalities and segments using a joint encoder. We then extract the representations from the ’MASK’ token from every layer of the text span encoder and the joint encoder. To evaluate the effect of the language-vision integration, we include a second condition in which we provide correct subtitles but incorrect video frames from a video randomly sampled from the video dataset (see below). Finally, we include results from GPT-2 (Radford et al., 2019) encoding the same text span of each segment, as an additional baseline.

Video data. We construct a video dataset consisting of 1075 35-second video clips from three episodes of the TV show *Friends*. These videos are extracted from every three seconds of each episode so that they could be directly mapped to the timestamps when brain recordings are collected (see below). For each video, the subtitles (namely, the inputs for the text modality) were automatically transcribed using the Google Speech-to-Text API given the associated audio.

Brain data. We use the brain recordings of 6 subjects when watching the same episodes from the Courtois *Friends* TV show fMRI dataset (Boyle, Pinsard, & et al., 2020). The recordings are sampled at a repetition time (TR) of 1.49 seconds for one session, and at every TR, the activity level of each voxel in a subject's brain is recorded. Because the videos are extracted from every three seconds (roughly equal to $2 * 1.49$) of an episode, we can map the offset of videos to the TR at which the last segment of a video is presented. We concatenate the brain's representation of all videos, which results in a matrix $Y \in R^{1075 \times V_i}$, where V_i is the number of voxels in the fMRI recordings for participant i.

Model-brain alignment. We build an encoding model from the models' representations of the ’MASK’ token and then predict the brain matrix of each participant viewing a video segment. Each voxel value in the brain matrix is estimated from the inputs using a linear function regularized by the ridge penalty. We train the encoding model through six-fold cross-validation, where the parameters are selected with nested cross-validation. We then evaluate the encoding model based on the voxel-based mean Pearson correlation scores between fMRI predictions for held-out data and actual values. We finally perform a permutation test on fMRI predictions and report the mean correlations of significantly predicted voxels.

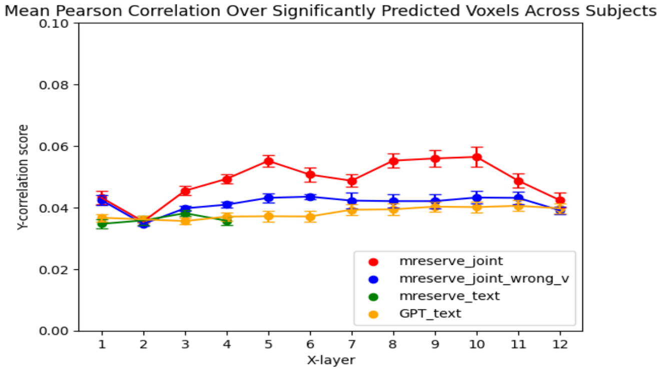

Figure 1: Pearson correlation of brain alignment over significantly predicted voxels across the whole brain.

Results

We first investigate whether the joint representation of text and vision information from the MERLOT Reserve model can better align with the brain responses of subjects viewing video stimuli. We present this contrast (red vs. green) for the significantly predicted voxels throughout the brain in Figure 1. We observe that the vision-language representations from the joint encoder are better aligned with brain representations throughout the early to late layers (layers 3-10), suggesting that these layers encode brain-related properties of video stimuli beyond that provided by the modality-specific encoder. We further show that these vision-language joint representations also outperform a text-only representation (Fig1, red vs. orange) obtained from another popular text-only language model–GPT-2–that has been shown to significantly predict fMRI recordings that contain language information (Schrimpf et al., 2021; Goldstein et al., 2022).

What is the reason for the superior performance of the joint encoder over the text encoder in predicting brain representations of video stimuli? We hypothesize that some improvements arise from a better understanding of the language information enriched by the vision modality. Nonetheless, it is plausible that these gains are the result of the in-depth processing of text information alone, since the text encoder has only 4 layers, while the joint encoder has 12 layers. We thus construct experiments in two conditions to study this possibility a) when text and correct video frames are provided; b) when text and incorrect video frames are provided. We present the contrast for the significantly predicted voxels across the language regions, computed over all layers. In Figure 2 (red vs. blue), we observe that the incorporation of correct video frames greatly improves the prediction of brain activity across the regions in the language network (Fedorenko, Hsieh, Nieto-Castanon, Whitfield-Gabrieli, & Kanwisher, 2010). The improvement cannot be due to the further processing of text-specific information in the joint encoder since the depth of text input processing is the same in both conditions and is unlikely to be due to vision-only information since these regions are known to support language processing.

To identify the language regions where the incorporation of visual information significantly improves alignment, we conduct an ROI-level Wilcoxon signed-rank test (p-value $<$ 0.05). We observe that significant differences are found between the two conditions in the Angular Gyrus (AG) region, which is thought to be responsible for multimodal integration (Farahibozorg, Henson, Woollams, & Hauk, 2022). We further observe significant differences in the bilateral Posterolateral Temporal Lobe (PTL), which are thought to support the processing of word-level (Hickok & Poeppel, 2007) and multi-word semantics (Toneva, Mitchell, & Wehbe, 2022). One possibility is that the incorporation of visual information enhances the representation of text-related information, but may not necessarily integrate with it in a brain-like way. For instance, the text provided to the model may not fully capture the brain's rich language representation of video stimuli, and vision information may complement it such that the resulting representation is more in line with the brain's language processing regions. This possibility should be further explored in future work.

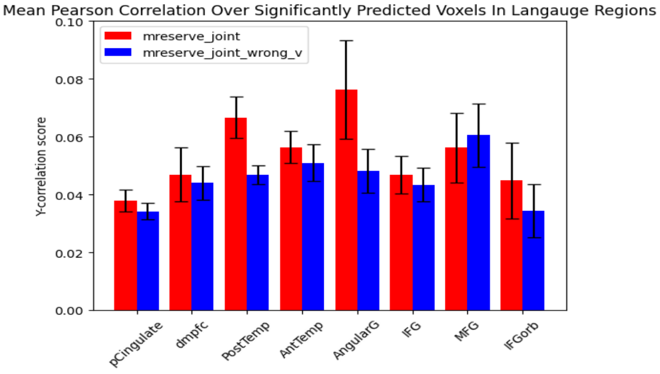

Figure 2: Pearson correlation brain alignment over the significantly predicted voxels across the language regions, computing over all 12 layers of the joint encoder.

Conclusion This work expands the exciting line of work that aligns brain activity with artificial neural networks in a fully multimodal setting. We hope to further understand the promise of joint representations obtained from the MERLOT Reserve model as valuable resources for studying the brain representations of video stimuli.

Acknowledgments

The Courtois project on neural modeling was made possible by a generous donation from the Courtois Foundation, administered by the Foundation Institut Gériatrie Montréal at CIUSSS du Centre-Sud-de-l'île-de-Montréal and University of Montreal. The Courtois NeuroMod team is based at "Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal", with several other institutions involved. See the cneuromod documentation for an up-to-date list of contributors (https://docs.cneuromod.ca). Dota is supported by an IMPRS fellowship.

References

Boyle, J., Pinsard, B., & et al. (2020, Jun). *The courtois project on neuronal modelling - 2020 data release.* (Poster 1939 was presented at the 2020 Annual Meeting of the Organization for Human Brain Mapping, held virtually.)

Farahibozorg, S.-R., Henson, R. N., Woollams, A. M., & Hauk, O. (2022). Distinct roles for the anterior temporal lobe and angular gyrus in the spatiotemporal cortical semantic network. *Cerebral Cortex*, *32*(20), 4549–4564.

Fedorenko, E., Hsieh, P.-J., Nieto-Castanon, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... others (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25*(3), 369–380.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Jain, S., Vo, V. A., Wehbe, L., & Huth, A. G. (2023). Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, 1–65.

Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., ... others (2023). Bold moments: modeling short visual events through a video fmri dataset and metadata. *bioRxiv*, 2023–03.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Reddy Oota, S., Arora, J., Rowtula, V., Gupta, M., & Bapi, R. S. (2022). Visio-linguistic brain encoding. *arXiv e-prints*, arXiv–2204.

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).

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, *2*(11), 745–757.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, *32*.

Vaidya, A. R., Jain, S., & Huth, A. (2022, 17–23 Jul). Self-supervised models of audio effectively explain human cortical responses to speech. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (Vol. 162, pp. 21927–21944). PMLR. Retrieved from `https://proceedings.mlr.press/v162/vaidya22a.html`

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2022). Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *bioRxiv*.

Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., ... Choi, Y. (2022). Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16375–16387).