INTERPRETING MULTIMODAL VIDEO TRANSFORMERS USING BRAIN RECORDINGS

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

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

ABSTRACT

Integrating information from multiple modalities is arguably one of the essential prerequisites for grounding artificial intelligence systems with an understanding of the real world. Recent advances in video transformers that jointly learn from vision, text, and sound over time have made some progress toward this goal, but the degree to which these models integrate information from the input modalities still remains unclear. In this work, we present a promising approach for probing a multimodal video transformer model by leveraging neuroscientific evidence of multimodal information processing in the brain. We use the brain recordings of subjects watching a popular TV show to interpret the integration of multiple modalities in a video transformer, before and after it is trained to perform a question-answering task that requires vision and language information. For the early and middle layers, we show that fine-tuning on the vision-language task does not improve the alignment in brain regions that are thought to support the integration of multimodal information over their pre-trained counterparts. We further show that the top layers of the fine-tuned model align substantially less with the brain representations, and yield better task performances than other layers, which indicates that the task may require additional information from the one available in the brain recordings.

1 INTRODUCTION

The advent of transformers has facilitated the development of video models that are capable of representing videos over time through multiple modalities such as language, sound, and vision (Selva et al., 2022). While empirical results show that these models seem to learn strong multimodal representations(Xu et al., 2021; Zellers et al., 2022), it remains unclear how they integrate information across modalities to perform a particular task. To investigate this question, we contrast two versions of the same multimodal model – one that was only pre-trained to predict masked audio or text sequences by self-supervision, and a second one that was further fine-tuned to perform a vision-language question-answering task. We expect that the fine-tuned model would further learn how to integrate visual and language information, which is required by the task, and the comparisons with the pre-trained version would be revealing about the integration in the multimodal transformer.

Most of the work probing internal multimodal representations has been devoted to vision-language transformers trained on static images and text (Frank et al., 2021; Salin et al., 2022; Hendricks & Nematzadeh, 2021), with significantly less work focusing on the multimodal ability of a video transformer. This is in part due to the difficulty of analyzing exactly how a model integrates multimodal information over time. We thereby propose to look at the brain – the only processing system capable of understanding video stimuli that contains complex events and situations that evolve over time. We can use fMRI (functional Magnetic Resonance Imaging) recordings to trace how the brain represents video stimuli, and then relate them with models' representations. With the alignment between the brain and models, the processing of multimodal information in the brain can be seen as a proxy for probing these representations.

In this work, we focus on MERLOT Reserve (Zellers et al., 2022), a state-of-the-art multimodal transformer that learns script knowledge of videos over time, jointly through vision, language, and sound. We use three complementary probing techniques to interpret the contrasts between pre-trained and fine-tuned MERLOT Reserve that are further trained on TVQA dataset (Lei et al., 2018), namely behavior measures (i.e. task accuracy), representational similarity measures (i.e. centered kernel alignment (CKA) (Kornblith et al., 2019)) and brain alignment of model representations with fMRI recordings of 5 participants watching the Friends TV show (Boyle et al., 2020). Further details on CKA and task accuracy are provided in Appendix A.2.

We make the following contributions:

- We present an approach for interpreting the internal representation of multimodal video transformers using multimodal brain activity; this relies on the relations between the properties of multimodal video stimuli and brain responses.
- We observe that the middle layers of a multimodal video transformer are better at predicting multimodal brain activity than other layers, indicating that the middle layers encode the most brain-related properties of the video stimuli.
- We show that the early and middle layers of a multimodal transformer that is fine-tuned on a vision-language task are similar to their pre-trained counterparts when predicting the activity in brain regions that are thought to integrate multimodal information. This suggests that fine-tuning for a vision-language task may not necessarily lead to additional integration of modalities. We further show that the top layers of the fine-tuned model have the worst alignment with brain activity but obtain the best vision-language task accuracies. This suggests that performing the vision-language task requires at least some information that is different from the one available in the brain recordings.

Background. Our work relates to previous work that uses brain recordings to interpret the representations derived from neural networks (Toneva & Wehbe, 2019; Aw & Toneva, 2022). Brain recordings capture a meaningful and observable spatio-temporal structure of how a natural stimulus is processed, which current highly-distributed deep learning systems fail to provide (Toneva & Wehbe, 2019; Kar et al., 2022). When the activity of a brain region is significantly predicted by the model's representations given the same stimulus, then the brain-related properties of that stimulus are thought to be encoded in the model's representations. The responses from different brain regions can therefore decompose a model's representations into interpretable brain-related properties. Because this approach requires no intervention in models and provides a human-like reference for representing stimuli, it is a promising framework for probing multimodal video transformers' ability to encode multimodal and temporal information.

Related work. Previous efforts exploring model-brain alignment have been centered around the models trained on unimodal data, such as text (Caucheteux et al., 2021; Gauthier & Levy, 2019), audio (Vaidya et al., 2022; Millet et al., 2022), or images (He et al., 2016; Schrimpf et al., 2018). Recent work has begun to align representations of vision-language models with brain recordings of subjects viewing static real images and comparing them to representations from vision-only models (Wang et al., 2022; Reddy Oota et al., 2022). These works show that the information learned from one modality can greatly enhance brain alignment in unimodal regions that support the other modality. In contrast with these previous works that use unimodal brain recordings, we investigate brain alignment of model representations with fMRI recordings in a fully multimodal task setting, namely watching videos.

2 METHODS

Model. We use a pre-trained MERLOT Reserve model provided by Zellers et al. (2022) that consists of 12 joint encoder layers, and its fine-tuned equivalent on the TVQA dataset Lei et al. (2018). Pre-trained MERLOT Reserve is trained on 20 million YouTube videos, through a learning objective by choosing the correct snippet of text (and audio) based on a contextualized representation of a video. Fine-tuned MERLOT Reserve is further trained on the TVQA dataset, which consists of 152,545 QA pairs (with 5 options) from 21,793 video clips of the TV show. At each layer, we extract representations from the joint encoder (for all modalities and timestamps) of the pre-trained and fine-tuned MERLOT Reserve when answering 400 questions from the TVQA dataset. These questions were selected such that they appear in the first two seasons of the Friends TV show, which

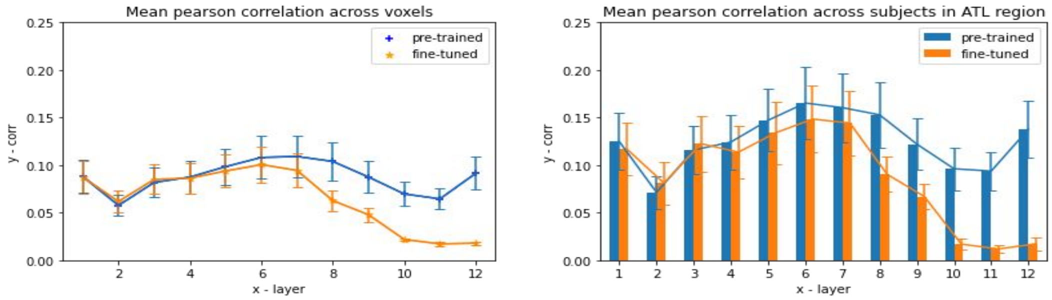

Figure 1: Mean Pearson correlation brain alignment over the significantly predicted voxels in the whole brain between pre-trained and fine-tuned MERLOT Reserve (Left); Mean Pearson correlation brain alignment over the significantly predicted voxels in bilateral anterior temporal lobes (ATL) region between pre-trained and fine-tuned MERLOT Reserve (Right). See Appendix A.2 for similar plots for mean Pearson correlation in other language processing regions (Fedorenko et al., 2010).

correspond to the stimulus of the brain recordings (see below). To extract the representation corresponding to each question video clip, we feed each model a 35-second video centered around the time region for each question The models then contextualize video frames with ten sequences that contain a question, one of five multiple-choice answers, and a masked text (or audio) token followed by subtitles (or audio). For each model, we thereby concatenate the representations of 10 masked tokens across the 12-layer joint encoder with a hidden size of 768. Hence, the model representation of 400 questions for each layer is a matrix of $d_h = 768 * 4000$.

Brain data. We use the fMRI recordings of 5 subjects watching the first two seasons of the Friends TV show. 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. For each of the 400 questions, we only select the brain representations at the TR at which the question-related contents are presented. We then concatenate the brain's representation of 400 questions and obtain a matrix $Y \in R^{400*V}$, where V is the number of voxels. One of the 5 participants has 382 data points, instead of 400, due to missing brain data.

Model-brain alignment. Following the prior work that learns model-brain alignment Toneva & Wehbe (2019); Reddy Oota et al. (2022); Aw & Toneva (2022), we construct an encoding model from the models' representations when answering a question, and then predict the brain matrix of a participant viewing a video clip related to the same question. Each voxel value in the brain matrix is estimated from the inputs using a linear function regularized by the ridge penalty. We then train the encoding model through four-fold cross-validation. The parameters are selected with nested cross-validation. We perform a permutation test on fMRI predictions, where the elements in fMRI predictions are randomly shuffled and then formulate 1000 permuted sets. For each voxel, we calculate the chance of the Pearson correlation score of permuted sets as or more extreme (at the rate of 0.05) than unpermuted fMRI predictions. For the evaluation, we calculate the mean Pearson correlation over the voxels that are significantly predicted across 5 participants.

3 RESULTS

We first investigate the brain alignment of the pre-trained MERLOT Reserve model across its intermediate layers. We present the average brain alignment across the significantly predicted voxels in the whole brain in Figure 1(Left, blue). We observe that the brain alignment peaks in the middle intermediate layers of the model, suggesting that the middle layers of these models encode the most brain-related properties of video stimuli. This finding is consistent with the results of brain alignment in large language models, such as BERT (Toneva & Wehbe, 2019) and GPT-2 (Caucheteux & King, 2022).

We then focus on comparing the brain alignment of MERLOT Reserve when pre-trained versus when fine-tuned on the TVQA task that involves vision-language information. We present this

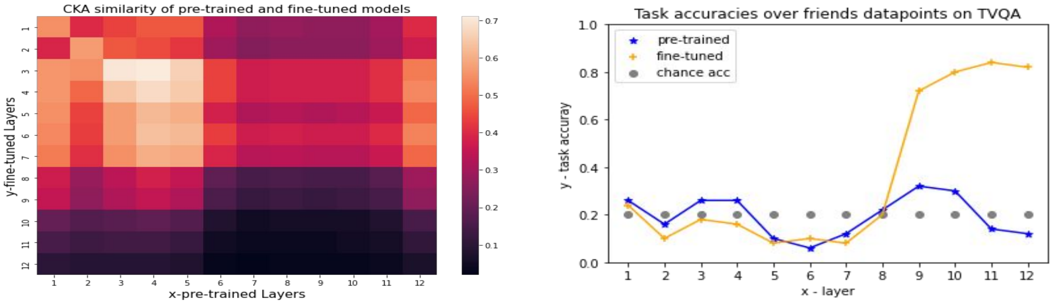

Figure 2: CKA similarity scores of two representations across different layers from pre-trained and fine-tuned MERLOT Reserve (Left); TVQA Task accuracy when feeding the representation from an intermediate layer to the prediction layer of pre-trained and fine-tuned MERLOT Reserve (Right)

contrast for the significantly predicted voxels in the whole brain in Figure 1(Left, blue vs. orange). We observe that the early and middle layers of the fine-tuned model lead to better alignment with brain representations compared to the top layers of the fine-tuned model. These layers of fine-tuned models also lead to similar brain alignment as their pre-trained counterparts. Comparing the alignment of the pre-trained and fine-tuned models across language processing regions (Fedorenko et al., 2010), we find that their early and middle layers exhibit similar alignment in the regions that are thought to support the integration of multimodal information. For instance in Figure 1 (Right), we show that no significant differences are observed between the early and middle layers of pre-trained and fine-tuned models in the bilateral anterior temporal lobes (ATL) region, which is thought to be a hub for multimodal integration (Farahibozorg et al., 2022). The results suggest that fine-tuning the video transformer for the vision-language question-answering task may not necessarily lead to better integration of multimodal information in these layers over pre-trained models.

To better understand why the later layers of the fine-tuned model drastically decrease their brain alignment, we further investigate the similarity of its representations to those of the pre-trained model using CKA (Kornblith et al., 2019). We present the results in Figure 2 (Left), which show that the later layers of the fine-tuned model are dissimilar to both the early and later layers of the pre-trained model. In contrast, the early layers of the fine-tuned model are not only approximately similar to the early layers but even to the later layers of the pre-trained model. One hypothesis for why the early layers of the fine-tuned model are similar to all layers of the pre-trained model is that the fine-tuning may be acting to “compress” some of the information from the pre-trained model in order to increase the capacity for encoding more task-specific information in the later layers. This hypothesis may be further explored by future work.

Furthermore, while the later layers of the fine-tuned model show a significant decrease in predicting brain activity compared to the pre-trained model, as seen in Figure 1 (Left), we observe that these later layers of the fine-tuned model also show a sharp increase in accuracy on the question-answering task (see Figure 2 (Right)). In contrast, the earlier layers of the fine-tuned model and all layers of the pre-trained model perform at chance accuracy (0.20). This result, combined with previous findings that later layers appear to encode more task-specific information (Merchant et al., 2020; Zhou & Srikumar, 2021; Durrani et al., 2021; Mosbach et al., 2020), suggests that not all features encoded in brain representation are task-relevant. One possible reason could be that the brain activity was recorded while the participants were simply watching the show, rather than answering questions about it.

4 Future work

This work expands the exciting line of work that aligns brain activity with neural networks to a fully multimodal setting. We hope to further understand the precise role of brain regions when engaging in complex multimodal reasoning, such as video question answering. One next step that follows from this work is to investigate the alignment between the models and brains when the human participants engage in an active task that requires multimodal information, rather than passive viewing.

5 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

Khai Loong Aw and Mariya Toneva. Training language models for deeper understanding improves brain alignment. *arXiv preprint arXiv:2212.10898*, 2022.

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

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Charlotte Caucheteux, Alexandre Gramfort, and J. R. King. Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning*, 2021.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. How transfer learning impacts linguistic knowledge in deep nlp models? *arXiv preprint arXiv:2105.15179*, 2021.

Seyedeh-Rezvan Farahibozorg, Richard N Henson, Anna M Woollams, and Olaf Hauk. Distinct roles for the anterior temporal lobe and angular gyrus in the spatiotemporal cortical semantic network. *Cerebral Cortex*, 32(20):4549–4564, 2022.

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

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. *arXiv preprint arXiv:1910.01244*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

Kohitij Kar, Simon Kornblith, and Evelina Fedorenko. Interpretability of artificial neural network models in artificial intelligence vs. neuroscience. *arXiv preprint arXiv:2206.03951*, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*, 2020.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning. *arXiv preprint arXiv:2206.01685*, 2022.

Marius Mosbach, Anna Khokhlova, Michael A Hedderich, and Dietrich Klakow. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. *arXiv preprint arXiv:2010.02616*, 2020.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. On the origins of the block structure phenomenon in neural network representations. *arXiv preprint arXiv:2202.07184*, 2022.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Visio-linguistic brain encoding. *arXiv e-prints*, pp. arXiv–2204, 2022.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. In *AAAI 2022*, 2022.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *arXiv preprint arXiv:2201.05991*, 2022.

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

Aditya R Vaidya, Shailee Jain, and Alexander Huth. Self-supervised models of audio effectively explain human cortical responses to speech. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21927–21944. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/vaidya22a.html`.

Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *bioRxiv*, 2022.

John M Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Similarity analysis of contextual word representation models. *arXiv preprint arXiv:2005.01172*, 2020.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 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, 2022.

Yichu Zhou and Vivek Srikumar. A closer look at how fine-tuning changes bert. *arXiv preprint arXiv:2106.14282*, 2021.

A Appendix

A.1 Interpreting Metrics

Task Accuracy. We evaluate the task accuracy of pre-trained and fine-tuned MERLOT Reserve when answering TVQA questions following the initial design by Zellers et al. (2022). The model scores the representations for each masked token in 10 sequences through a linear projection layer and selects the option with the highest probability. We refer to the selected option as the model's prediction and compare it against the gold label.

CKA. We use CKA score (Kornblith et al., 2019) to compare the learned representations of two layers within or across pre-trained and fine-tuned models. Given two representations $X \in R^{N1*d}$ and $Y \in R^{N2*d}$, where N1 or N2 is the number of examples and d is the dimension of a representation, CKA score will be between 0 (dissimilar) and 1 (similar). CKA score has been widely used in measuring and analyzing layer-wise differences in models' representations (Wu et al., 2020; Kornblith et al., 2019; Nguyen et al., 2022).

A.2 Mean Pearson cross across language processing regions

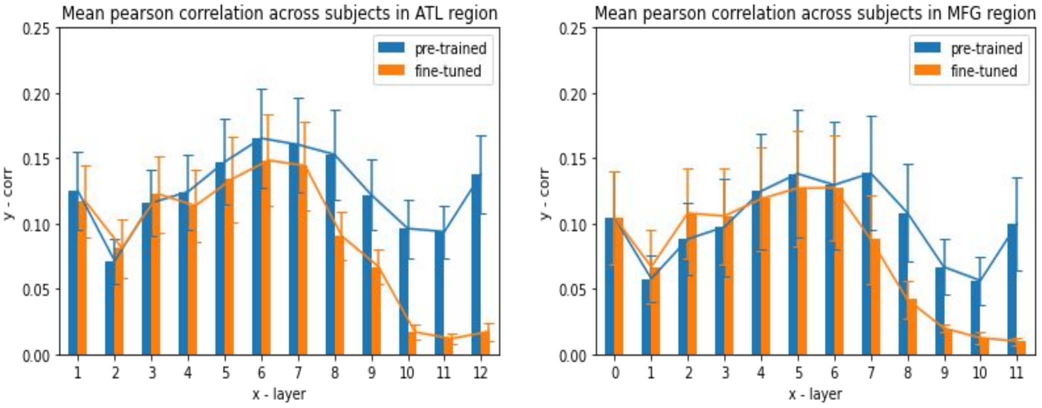

Figure 3: Mean Pearson correlation brain alignment over the significantly predicted voxels in bilateral anterior temporal lobes (ATL) region between pre-trained and fine-tuned MERLOT Reserve (Right); Mean Pearson correlation brain alignment over the significantly predicted voxels in middle frontal gyrus (MFG) region between pre-trained and fine-tuned MERLOT Reserve (Right)

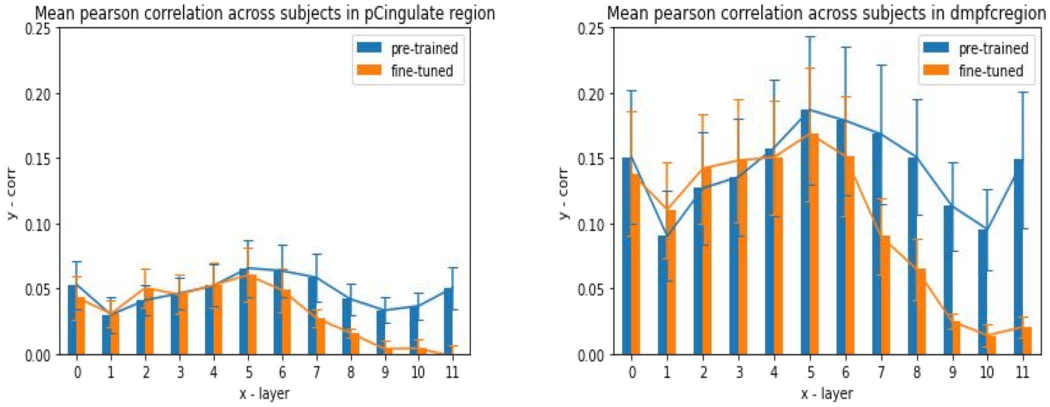

Figure 4: Mean Pearson correlation brain alignment over the significantly predicted voxels in posterior cingulated (pCingulate) region between pre-trained and fine-tuned MERLOT Reserve (Right); Mean Pearson correlation brain alignment over the significantly predicted voxels in dorsomedial prefrontal cortex (dmpfc) region between pre-trained and fine-tuned MERLOT Reserve (Right)

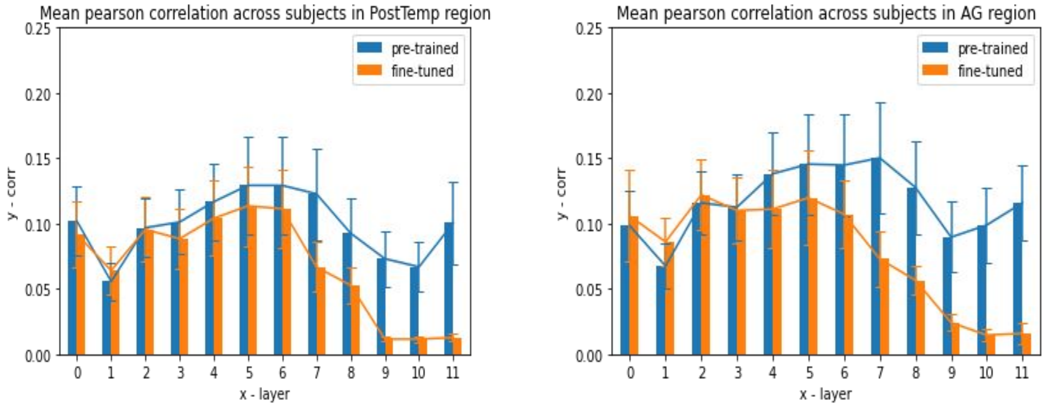

Figure 5: Mean Pearson correlation brain alignment over the significantly predicted voxels in posterolateral temporal (PostTemp) region between pre-trained and fine-tuned MERLOT Reserve (Right); Mean Pearson correlation brain alignment over the significantly predicted voxels in angular gyrus (AG) region between pre-trained and fine-tuned MERLOT Reserve (Right)

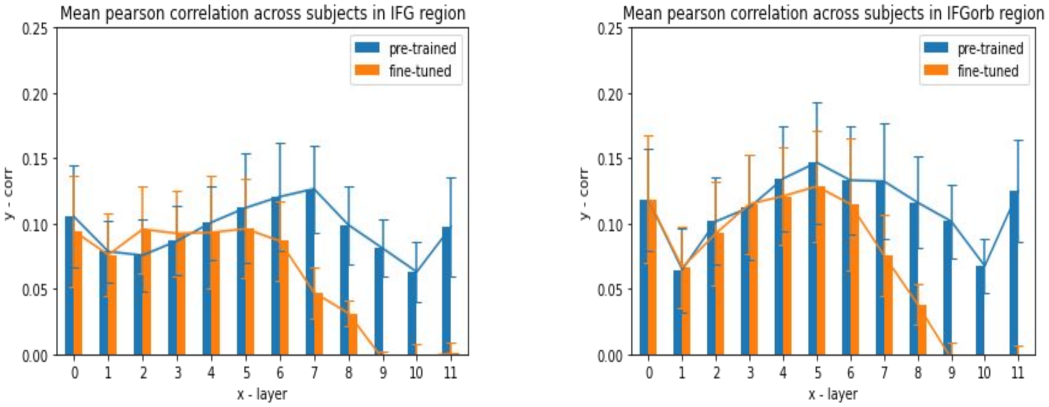

Figure 6: Mean Pearson correlation brain alignment over the significantly predicted voxels in inferior frontal gyrus (IFG) region between pre-trained and fine-tuned MERLOT Reserve (Right); Mean Pearson correlation brain alignment over the significantly predicted voxels in inferior frontal gyrus pars orbitalis (Right) between pre-trained and fine-tuned MERLOT Reserve (Right)