Self-Supervised Pretraining for Improved Downstream Decoding of Audio-Evoked fMRI Sequences

Sean Paulsen
Department of Computer Science
Dartmouth College
Hanover, NH 03755
paulsen.sean@gmail.com

Michael Casey
Departments of CS and Music
Dartmouth College
Hanover, NH 03755
mcasey@dartmouth.edu

Abstract

We present a sequential transfer learning framework for transformers on functional Magnetic Resonance Imaging (fMRI) data and demonstrate its significant benefits for decoding instrumental timbre. In the first of two phases, we pretrain our stacked-encoder transformer architecture on Next Thought Prediction, a self-supervised task of predicting whether or not one sequence of fMRI data follows another. This phase imparts a general understanding of the temporal and spatial dynamics of neural activity, and can be applied to any fMRI dataset. In the second phase, we finetune the pretrained models and train additional randomly initialized models on the supervised task of predicting whether or not two sequences of fMRI data were obtained while listening to the same musical timbre. The finetuned models achieve significantly higher accuracy on heldout participants than the randomly initialized models, demonstrating the efficacy of our framework for facilitating transfer learning on fMRI data. This work contributes to the growing literature on transformer architectures for sequential transfer learning on fMRI data.

1 Introduction

A functional MRI (**fMRI**) scan measures blood-oxygen-level-dependent (**BOLD**) responses that reflect changes in metabolic demand consequent to neural activity [1, 4, 13]. Researchers have adopted machine learning techniques to analyze the complex relationship between BOLD signal and the underlying task, stimulus, disease, or biological information. More specifically, training a model to predict such information given the BOLD data as input is known as **task-state decoding**, or **brain decoding**. Toward the goal of more powerful brain decoding models, many advances in modern deep machine learning have been applied to fMRI research. These include convolution-based models [15, 6, 12], recurrent neural networks [2], and graph neural networks [7]. Most recently, transformer [14] based models have achieved state of the art results on several brain decoding tasks [8, 1, 10], having already grown to dominate most other areas of deep learning research, in particular natural language processing Devlin et al. [3].

Further, the transformer architecture has emerged as a superior alternative to recurrent methods for fMRI timeseries modeling with the strategy of **transfer learning** [8, 10, 11]. The first phase of this strategy "pretrains" a model on an unsupervised or self-supervised task to acquire general knowledge inherent in the dataset. In the second phase, this knowledge is "transferred" to the target or "finetuning" task for improved performance and reduced data burden [5]. In this paper we introduce a sequential transfer learning framework for transformers on *pairs* of sequences of audio-evoked fMRI data. We then demonstrate our transformer architecture's ability to learn a novel self-supervised pretraining task and transfer that knowledge to significantly improve performance on a supervised auditory brain decoding task. We further demonstrate a positive relationship between pretraining

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

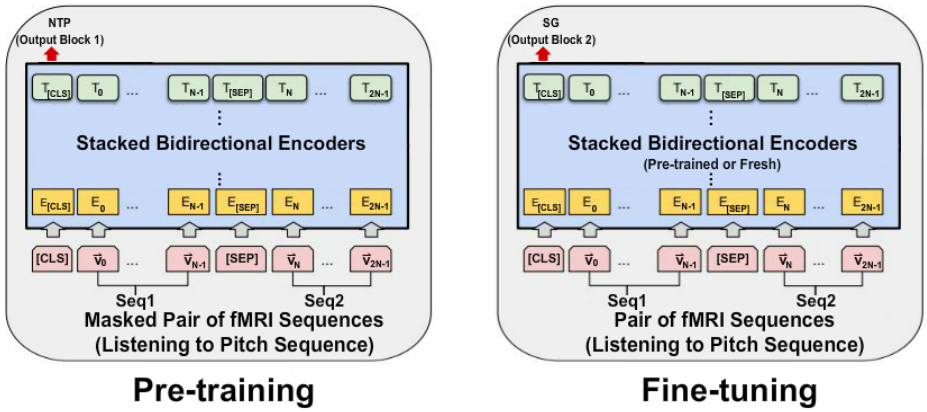

Figure 1: Architecture for pretraining and finetuning phases

performance and finetuning performance, which provides further evidence that our transfer learning is genuine. These are, to the best of our knowledge, the first models to successfully distinguish instrumental timbre in fMRI data.

2 Architecture and Training Tasks

Our architecture is a modified stacked bi-directional encoder [3], depicted in Figure 1. It has two separate output blocks, one for pretraining and one for finetuning. All training data in this work were built from the Auditory Imagery fMRI Dataset [9], which has been anonymized and is available upon request. We refer to that work for the full details of the data collection, but the relevant information is that the scan was performed while the participants listened to various pitches of clarinet and trumpet.

Each input to the model is constructed by extracting a contiguous sequence of fMRI images, denoted (**Seq1**), limited to Superior Temporal Gyrus (**STG**), and then pairing it with another such sequence (**Seq2**), as seen in Figure 1. The length of each of the two sequences in an input is 5.

The separator token (**SEP**) and classification token (**CLS**) [3] are used in the familiar way [8, 10, 11]. SEP is inserted between the two sequences, and CLS is inserted at the front to serve as a pooling token which extracts information from the rest of the sequence.

Our novel self-supervised pre-training task is Next Thought Prediction (**NTP**). The goal of this task is binary classification, predicting whether or not Seq2 follows immediately after Seq1 in the original data. We hypothesized that this task would teach the model a general understanding of the temporal and spatial dynamics of neural activity, which could then be transferred to a target brain decoding task. Our novel supervised brain decoding task used for finetuning is the Same Timbre (**ST**) task. The goal of this task is binary classification, predicting whether or not Seq1 corresponds to listening to the same instrumental timbre (clarinet or trumpet) as Seq2.

3 Experiments and Results

A summary of experiment parameters is given in Table 1. The format of the Pretraining and Finetuning Samples rows is "{training samples} and {validation samples}." Additional experiment information can be found in the Appendix below.

We performed 8-fold cross validation for both phases. Sixteen of the seventeen participants were partitioned uniformly at random in groups of 2 to be held out as validation data for each of the 8 folds. Results for Left STG and Right STG are reported separately. Baseline chance for both tasks is 50%. The "Best Val. Acc." column reports the highest accuracy obtained during training on the heldout participants. The epoch in which that accuracy was obtained is given in the "Best Epoch" column, from 0 to 9 inclusive. The average of the best validation accuracies across the eight folds is given at the bottom of the corresponding columns with $\pm$ standard deviation. For each fold, we saved the model's state after the Best Epoch to be used for transfer learning. During finetuning, for each

Table 1: Summary of auditory imagery dataset

Auditory Imagery Dataset	
# of Participants	17
Training Regimen	8-fold cross-val, heldout participants
Regions of Interest	Left and Right STG
Pretraining Task	Next Thought Prediction
Pretraining Samples	26,640 and 3,552
Finetuning Task	Same-Timbre
Finetuning Samples	2,520 and 336

Table 2: Results of 8-fold cross-validation NTP pretraining

	Left STG		**Right STG**	
Fold	Best Val. Acc.	**Best Epoch**	Best Val. Acc.	**Best Epoch**
0	87.6	5	91.4	6
1	84.8	7	87.5	6
2	84.7	4	86.1	9
3	88.3	6	86.7	4
4	90.9	9	92.3	8
5	83.7	4	89.6	8
6	87.9	9	87.2	7
7	88.2	8	85.3	9
Average	87.0 ± 2.4	6.5	88.3 ± 2.5	7.1

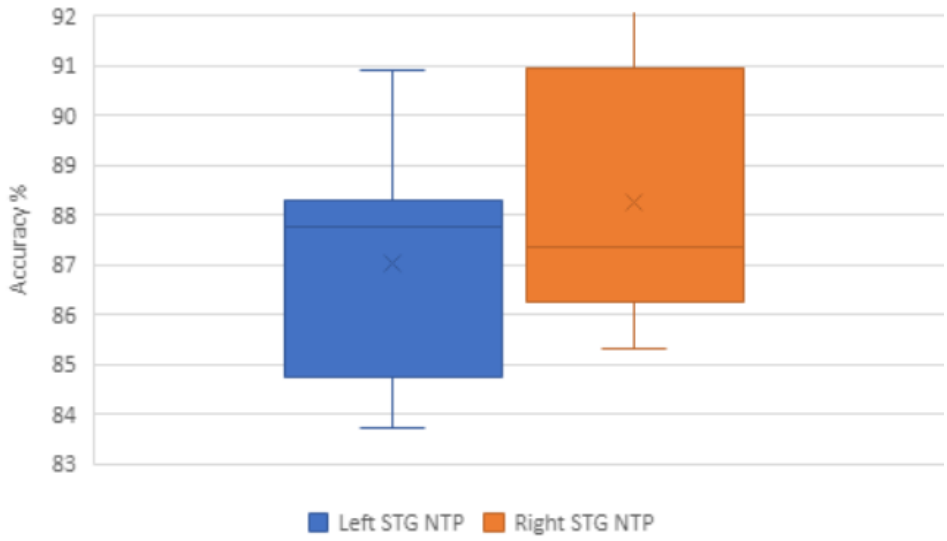

Figure 2: Average best validation accuracies for 8-fold cross-validation pretraining on NTP

fold of training data, a model was initialized with the saved weights from that fold of pretraining and a separate baseline model was randomly initialized (**RI**). Each fold held out the same two participants in both phases. Pretraining results are presented in Table 2 and a bar graph with error bars for pretraining on NTP is given in Figure 2. Finetuning results are reported in Table 3 and a bar graph with error bars for finetuning on ST is given in Figure 3. Four conditions are shown in that graph: transferring from Left NTP to Left ST, learning Left ST with RI, transferring from Right NTP to Right ST, and learning Right ST with RI. Baseline chance on this task is 50%.

Table 3: Results of 8-fold cross-validation ST transfer learning

	Same-Timbre in Left STG				Same-Timbre in Right STG			
	Transfer from NTP		RI		Transfer from NTP		RI	
Fold	**Best Val. Acc.**	**Best Epoch**	**Best Val. Acc.**	**Best Epoch**	**Best Val. Acc.**	**Best Epoch**	**Best Val. Acc.**	**Best Epoch**
0	76.8	2	63.4	8	78.3	1	70.5	8
1	68.2	9	61.0	6	70.2	7	67.9	3
2	75.0	6	75.3	9	75.0	1	66.4	7
3	75.9	5	64.3	9	69.0	8	61.0	4
4	71.4	1	68.2	9	72.9	1	71.4	9
5	67.3	5	69.0	7	73.2	3	68.8	9
6	81.5	9	73.5	9	75.6	6	65.5	7
7	72.3	7	71.1	8	64.9	5	66.7	8
Avg.	73.5 ± 4.7	5.5	68.2 ± 5.0	8.1	72.4 ± 4.2	6.9	67.3 ± 3.2	6.9

Figure 3: Average best validation accuracies for 8-fold cross-validation finetuning on ST

For pretraining, the folds on both hemispheres significantly outperformed the baseline chance of 50% when generalizing to heldout participants (one sample t-test for both hemispheres against hypothetical mean of 50%, $p < .001$). However, there is no significant difference between the ability of Left and Right STG to generalize to heldout participants (paired t-test between the two sets of accuracies, p=.2643). We contribute this result as significant evidence that the Next Thought Prediction task is well-defined and our novel paired-sequence architecture is capable of learning it.

For finetuning, the folds on both hemispheres significantly outperformed the baseline chance of 50% (one-sample t-tests, p<.0001 in all cases). The Best Val. Acc. values obtained via transfer learning in Left STG are significantly higher than those obtained on the RI models (paired t-test between two sets of values, p=.0306). Similar for Right STG (paired t-test between two sets of values, p=.0105). There is no significant difference between Left and Right STG for the transfer learning models' performance (paired t-test, p=.5256). We contribute this result as significant evidence of our framework's ability to perform sequential transfer learning to improve performance on a brain decoding task. Further, we contribute our success on the Same-Timbre task as the first significant evidence of distinguishability of instrumental timbre in STG.

References

[1] Hasan Atakan Bedel, Irmak Şıvgın, Onat Dalmaz, Salman Ul Hassan Dar, and Tolga Çukur. BolT: Fused Window Transformers for fMRI Time Series Analysis, February 2023. URL http://arxiv.org/abs/2205.11578. arXiv:2205.11578 [cs, eess].

[2] Jumana Dakka, Pouya Bashivan, Mina Gheiratmand, Irina Rish, Shantenu Jha, and Russell Greiner. Learning Neural Markers of Schizophrenia Disorder Using Recurrent Neural Networks, December 2017. URL http://arxiv.org/abs/1712.00512. arXiv:1712.00512 [cs].

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

[4] Elizabeth M. C. Hillman. Coupling mechanism and significance of the BOLD signal: a status report. *Annual Review of Neuroscience*, 37:161–181, 2014. ISSN 1545-4126. doi: 10.1146/annurev-neuro-071013-014111.

[5] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing, August 2021. URL http://arxiv.org/abs/2108.05542. arXiv:2108.05542 [cs].

[6] Jeremy Kawahara, Colin J. Brown, Steven P. Miller, Brian G. Booth, Vann Chau, Ruth E. Grunau, Jill G. Zwicker, and Ghassan Hamarneh. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, February 2017. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2016.09.046.

[7] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H. Staib, Pamela Ventola, and James S. Duncan. BrainGNN: Interpretable Brain Graph Neural Network for fMRI Analysis. *Medical Image Analysis*, 74:102233, December 2021. ISSN 1361-8423. doi: 10.1016/j.media.2021.102233.

[8] Itzik Malkiel, Gony Rosenman, Lior Wolf, and Talma Hendler. Self-Supervised Transformers for fMRI representation, August 2022. URL http://arxiv.org/abs/2112.05761. arXiv:2112.05761 [cs, eess].

[9] Lloyd May, Andrea R. Halpern, Sean D. Paulsen, and Michael A. Casey. Imagined Musical Scale Relationships Decoded from Auditory Cortex. *Journal of Cognitive Neuroscience*, 34(8):1326–1339, 07 2022. ISSN 0898-929X. doi: 10.1162/jocn_a_01858. URL https://doi.org/10.1162/jocn_a_01858.

[10] Sam Nguyen, Brenda Ng, Alan D. Kaplan, and Priyadip Ray. Attend and Decode: 4D fMRI Task State Decoding Using Attention Models, January 2021. URL http://arxiv.org/abs/2004.05234. arXiv:2004.05234 [cs].

[11] Sean Paulsen and Michael Casey. Self-supervised pretraining on paired sequences of fmri data for transfer learning to brain decoding tasks, 2023.

[12] Sean Paulsen, Lloyd May, and Michael Casey. Decoding imagined auditory pitch phenomena with an autoencoder based temporal convolutional architecture. In *BRAININFO*, Nice, France, July 2021. IARIA.

[13] J. C. Rajapakse, F. Kruggel, J. M. Maisog, and D. Y. von Cramon. Modeling hemodynamic response for analysis of functional MRI time-series. *Human Brain Mapping*, 6(4):283–300, 1998. ISSN 1065-9471. doi: 10.1002/(sici)1097-0193(1998)6:4<283::aid-hbm7>3.0.co;2-#.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

[15] Liang Zou, Jiannan Zheng, Chunyan Miao, Martin J. Mckeown, and Z. Jane Wang. 3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI. *IEEE Access*, 5:23626–23636, 2017. ISSN 2169-3536. doi: 10.1109/ACCESS.2017.2762703. URL http://ieeexplore.ieee.org/document/8067637/.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Results and significance testing are clear.
   (b) Did you describe the limitations of your work? [Yes] Mention that data is limited to Superior Temporal Gyrus, no other limitations.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Read and confirmed.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Data is available on request, code is linked in appendix. Currently lacking explicit instructions but we are available to help interested parties.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Briefly mentioned above, further information in appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Standard deviation given for averages across our 8-fold cross-validation in both phases above, error bars included on graphs.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Cited original paper of the dataset used.
   (b) Did you mention the license of the assets? [No] Dataset not currently licensed, only available on request for now.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] Code linked in appendix.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Data belongs to our lab.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Mentioned that data has been anonymized.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Potential Negative Societal Impact

The results of this work do not carry any potential for negative societal impact.

A.2 Additional Experiment Information

All reported models contained three transformer-encoder layers, and each layer had 2 attention heads with a forward expansion factor of 4. These hyperparameters were obtained via grid search. Inputs to the model are sequences of 420-dimensional vectors. In both the pretraining and finetuning phases, the final transformed state of the CLS token is fed to an output layer which yields the probabilities for the "No" and "Yes" labels. In both phases the output layer consists of a linear layer which projects down from 420 dimensions to 210, then a second linear layer projecting from 210 dimensions to 2, and finally a softmax is applied to obtain the probabilities. The loss for each task is calculated as the Cross-Entropy between the result of the output block and a one-hot encoding of the ground truth. All models were trained for ten epochs via backpropagation with the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and weight_decay $= 0.0001$. Each model was trained with a different RNG seed for reproducibility. Our models and inputs are relatively small and thus only require less than 10 hours' training on CPU on our internal high performance cluster. Due to the low resource demand, we did not record exact training times.

A.3 Code

All code used in this work can be found at https://github.com/paulsens/fmriBERT. At time of writing we have not written step by step instructions but we are available by email to assist any interested parties.