Toward Real-World Emotion Decoding: A Transformer-Based Approach Using Movie fMRI

Bo-Gyeom Kim [1], Patrick Styll [2], Jiook Cha [1]

Department of Psychology, Seoul National University

Institute of Computer Engineering, TU Wien (Vienna University of Technology), Vienna, Austria

Abstract

Real-world emotion recognition arises through continuous interactions among multiple sensory cues—dynamics often missed by standard laboratory paradigms. To investigate these dynamics, we applied a Transformer-based deep-learning model (SwiFT combined with Perceiver IO) to functional MRI data from 512 youths (ages 5–21) watching a 10-minute movie. By modeling neural signals as continuous time-series, we tracked short-term (~40s, 50 TRs) changes in seven emotions (e.g., positive, fear). Longer sequence windows and explicit hemodynamic modeling (double-gamma HRF) improved decoding accuracy, highlighting the importance of extended temporal context and precise BOLD-delay modeling. The prominent contribution of the visual cortex suggests reliance on low-level visual features within rich audiovisual stimuli. These findings demonstrate that flexible sequence-to-sequence methods effectively capture the temporal dynamics of emotion recognition under realistic conditions, deepening our understanding of real-world emotional processing.

Keywords: emotion perception; emotion decoding; predictive coding; constructivist theory; fMRI; deep learning; naturalistic stimuli

Introduction

Real-world emotion recognition arises from continuous interactions among multiple sensory cues, which standard laboratory paradigms often fail to capture. Conventional neuroscience methods, typically relying on averaged or segmented neural signals, also struggle with time-series complexity. More flexible approaches are thus needed to track moment-to-moment changes under naturalistic conditions.

Several theoretical frameworks highlight the need to study emotions in continuously evolving contexts. Predictive coding proposes that the brain updates its predictions based on incoming signals, making emotional states dynamic inferences shaped by internal and external factors (Friston & Kiebel, 2009). Constructivist theory (Barrett, 2013) emphasizes personal context and past experiences, leading individuals to interpret the same stimulus differently. Together, these views suggest examining emotion perception as a process developing over time, rather than through discrete, controlled snapshots.

In this study, we investigate how 512 children and adolescents perceive emotion in a naturalistic setting by applying a Transformer-based model to functional MRI data collected during movie viewing. By modeling neural signals as a continuous time-series rather than averaging across discrete blocks, we aim to capture the moment-to-moment unfolding of emotion perception in an environment that more closely mirrors real-world experiences. This approach provides insight into how participants interpret evolving stimuli, bridging the gap between tightly controlled laboratory tasks and the complexities of everyday emotional encounters.

Methods

We analyzed fMRI data from 512 youths (5–21 years, HBN dataset) including ADHD (n=301), ASD (n=63), and comorbidities (n=59). Participants watched a 10-minute *Despicable Me* (750 TRs at 0.8 s), with per-TR ratings of seven emotions obtained from human raters (Camacho et al., 2023).

To assess the feasibility of decoding emotional states, we developed a hybrid deep-learning model by combining SwiFT (Swin Transformer for fMRI; Kim et al., 2023) and Perceiver IO (Jaegle et al., 2021). SwiFT encodes spatiotemporal patches via multi-head self-attention, while Perceiver IO performs sequence-to-sequence regression of emotion at each TR. Multiple sequence lengths (5, 10, 30, 50 TRs) and input offsets (0, 3, 5, 10, 20, 40 TRs) were explored to account for different stimulus-response lags, and a double-gamma hemodynamic response function (peak=5 s, undershoot=12 s) was applied to model BOLD delays. Integrated Gradients (IG) was computed for the top 20% of test participants (n=21) whose predictions were most accurate, enabling voxel-level contribution analysis for each emotion category.

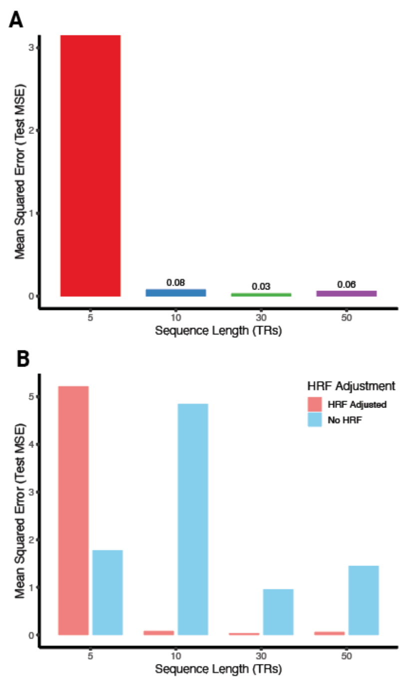

Figure 1: (A) Effect of sequence length on decoding performance. We tested windows of 5, 10, 30, and 50 TRs at a fixed learning rate, finding that longer sequences consistently yielded lower MSE. **(B)** Comparison of MSE between non-HRF (fixed 6 s delay) and a double-gamma HRF. Modeling the full hemodynamic response consistently lowered MSE, highlighting the importance of accurately capturing hemodynamic delay.

Results

To capture the continuous, real-world nature of emotion decoding, we systematically varied sequence length (5, 10, 30, 50 TRs), HRF modeling (fixed 6 s delay vs. double-gamma), and time offset (0–5 TRs), seeking the optimal configuration for a sequence-to-sequence approach. A 50-TR window, 3-TR offset, and double-gamma HRF yielded the best performance (MAE=0.058, MSE=0.040, r=0.894), while omitting HRF correction increased error rates (Figure 1).

Next, we assessed model performance across multiple emotion dimensions. Anger (*r* = 0748, MSE = 0.13), Fear (*r* = 0.548, MSE = 0.07) showed higher prediction accuracy, whereas Sad (*r* = 0.867, MSE = 2.35), Happy were more difficult to decode. These results suggest that dimensions with stronger arousal cues may exhibit more consistent neural signatures in this paradigm.

IG-based interpretation indicated that the visual cortex contributed prominently to predictions across all seven emotion categories, including positive and negative valence. In the top-performing group of participants, this region displayed consistently high voxel-wise attribution scores, suggesting a broad involvement of visual processing in fMRI-based emotion decoding under cinematic stimulation **(Figure 2).**

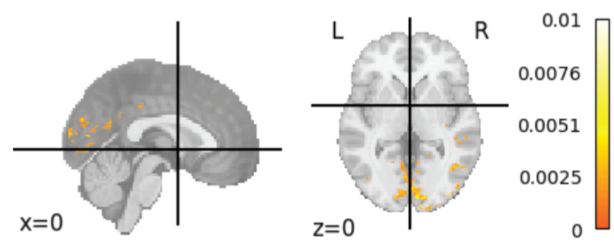

Figure 2: Integrated Gradients map for Positive emotion, displaying the top 5% of voxels contributing to predictions. Similar patterns emerged across all emotion categories, highlighting the visual cortex as a key region in continuous emotion decoding.

Discussion

Our findings show that longer temporal windows and explicit hemodynamic modeling significantly enhance continuous emotion decoding, with anger and fear predicted more accurately than sad—likely reflecting the stronger neural signals elicited by higher-arousal states. The visual cortex consistently emerged as a key contributor, aligning with prior findings on the modulation of visual processing by emotion (Vuilleumier & Driver, 2007). Future work will compare these methods to conventional approaches and investigate developmental/clinical subgroups to see how different populations encode emotion.

Acknowledgements

This work was supported by grants from the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (No. 2021R1C1C1006503, RS-2023-00266787, RS-2023-00265406, RS-2024-00421268, RS-2024-00342301, RS-2024-00435727, NRF-2021M3E5D2A01022515); the Creative-Pioneering Researchers Program through Seoul National University (No. 200-20240057, 200-20240135); the Semi-Supervised Learning Research Grant funded by Samsung (No. A0342-20220009); the "Identify the Network of Brain Preparation Steps for Concentration" Research Grant funded by Looxid Labs (No. 339-20230001); the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]; the Global Research Support Program in the Digital Field funded by MSIT, supervised by IITP (No. RS-2024-00421268); the National Supercomputing Center, with supercomputing resources including technical support (KSC-2023-CRE-0568); the Ministry of Education of the Republic of Korea and the NRF (No. NRF-2021S1A3A2A02090597); the Korea Health Industry Development Institute (KHIDI) and the Ministry of Health and Welfare, Republic of Korea (No. HR22C1605); the Artificial Intelligence Industrial Convergence Cluster Development Project funded by MSIT, Korea and Gwangju Metropolitan City; and the KBRI Basic Research Program funded by the Ministry of Science and ICT (No. 25-BR-05-01).

References

Barrett, L. F. (2013). Psychological Construction: The Darwinian Approach to the Science of Emotion. Emotion Review, 5(4), 379–389. https://doi.org/10.1177/1754073913489753

Camacho, M. C., Nielsen, A. N., Balser, D., Furtado, E., Steinberger, D. C., Fruchtman, L., Culver, J. P., Sylvester, C. M., & Barch, D. M. (2023). Large-scale encoding of emotion concepts becomes increasingly similar between individuals from childhood to adolescence. *Nature Neuroscience*, *26*(7), 1256–1266. https://doi.org/10.1038/s41593-023-01358-9

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Hénaff, O., Botvinick, M. M., Zisserman, A., Vinyals, O., & Carreira, J. (2021). Perceiver IO: A General Architecture for Structured Inputs & Outputs. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2107.14795

Kim, P. Y., Kwon, J., Joo, S., Bae, S., Lee, D., Jung, Y., Yoo, S., Cha, J., & Moon, T. (2023). SwiFT: Swin 4D fMRI Transformer. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.05916

Vuilleumier, P., & Driver, J. (2007). Modulation of visual processing by attention and emotion: windows on causal interactions between human brain regions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 837–855. https://doi.org/10.1098/rstb.2007.2092