Brain-Agnostic 3DCNNs Learn Naturalistic Emotion from 7t fMRI

Joshua Lunger[1], Mason Hu[1], Aditya Rajeev[1], Tohya Tanemura[1], Samuel Kostousov[1], Jurgen Germann[1,2]
[1]*University of Toronto*, [2]*Krembil Brain Institute*

***Abstract*—Understanding emotions through neural activity is a key challenge in affective computing and neuroscience. In this work, we leverage brain-agnostic 3D convolutional neural networks (3DCNN) to learn functional representations of emotions from large-scale naturalistic 7T fMRI data. Our learned representations are consistent with neurobiological principles, highlighting the potential of deep learning for neural emotion inference. Code and data are available at [https://github.com/lungerjo/DeepEmotion].**

I. Introduction

Understanding emotions through large-scale naturalistic data offers a pathway to more effective and scalable emotion recognition. These systems could support care workers in identifying individuals with atypical emotional processing while also aiding disorder treatment through objective, immediate feedback. Advancements in scalable inference on neural data could further bridge human cognition and machine learning, enabling more personalized and adaptive interactions. This work paves the way for integrating neural preferences into machine learning, expanding brain-aware AI applications in healthcare and beyond.

II. Methodology

A. Data Collection and fMRI Preprocessing

We utilized the publicly available high-resolution 7T fMRI dataset [1] from the StudyForrest project consisting of whole-brain fMRI recordings collected while participants listened to an audio-described version of the movie Forrest Gump. The dataset scans are acquired at a spatial resolution of 1.4 mm isotropic and a temporal resolution of 2 seconds.

For our study, we leveraged non-linear anatomically aligned fMRI data mapped to a common group template using iterative affine and non-linear transformations included with the dataset. This approach minimizes inter-subject anatomical variability. The alignment procedure followed an iterative group-based registration process, where each participant's motion-corrected and distortion-corrected EPI images were first aligned using an affine transformation and then refined using a high-resolution non-linear warp field.

B. Annotation Preprocessing

Emotion annotations were collected from eight external observers who rated perceived emotions in the film. To ensure balanced training and a strong emotional signal, we applied heuristic clustering to group annotations by mapping to the 5 most frequently observed emotion categories. Each fMRI sample was assigned an emotion label based on the majority vote among observers, with the condition that at least half of them agreed on the emotion.

Fig. 1: Covariance scores for emotion annotations across observers. Heuristic clustering was used to map annotations to the five most common emotions.

C. Training

We trained 20 brain-agnostic 3D Convolutional Neural Networks (3DCNN) [2] to classify emotion states from fMRI data. 19 models were trained on one subject each and one model was trained on all 19 subjects. The models were trained using stochastic minibatch gradient descent with categorical cross-entropy loss and optimized with the Adam optimizer at a learning rate of 0.001. Training was conducted for 50 epochs on a NVIDIA Quadro4000.

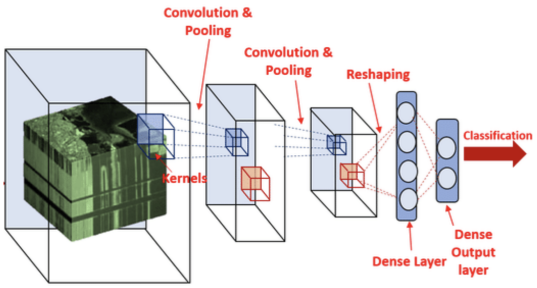

Fig. 2: The network consists of 3 convolutional layers with 3D kernels, batch normalization, and ReLU activations. A series of max-pooling operations were applied to downsample spatial dimensions while preserving feature representations. The final convolutional features were flattened and passed through 2 fully connected layers before a softmax classification head. [3]

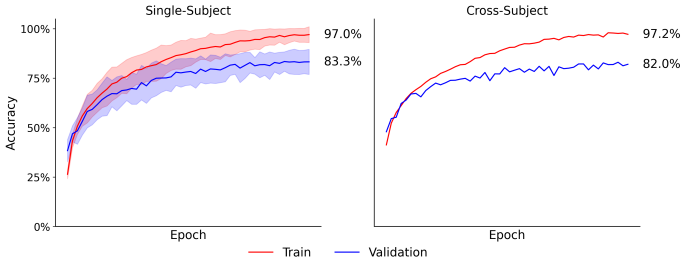

Fig. 3: Train and validation accuracy over epochs. Single-subject accuracy is averaged over 19 models trained on 1 subject each. Cross-subject accuracy is 1 model trained on all 19 subjects.

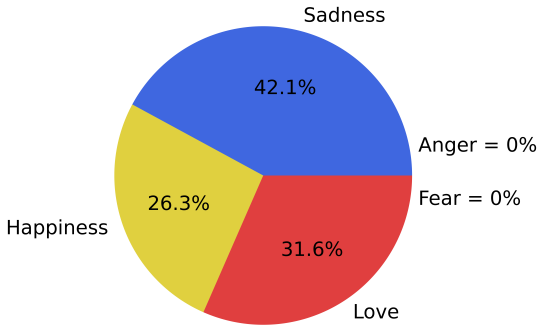

Fig. 4: Model predictions on held-out data from subject 18 during love scenes.

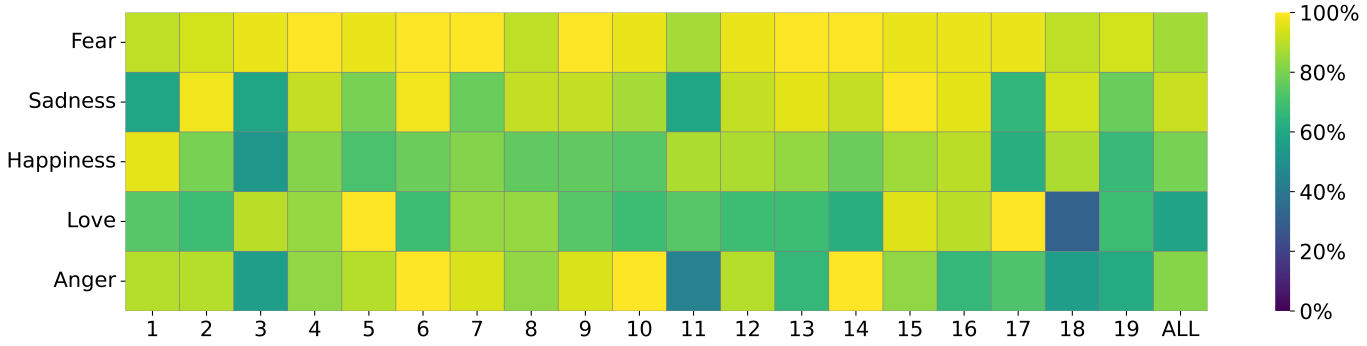

Fig. 5: Emotion-wise accuracy on held-out data within subject. The column label indicates the subject trained on by this model.

III. Results

We observe impressive performance on held-out data from both single-subject and cross subject models with limited compute and hyperparameter tuning demonstrating the power of this technique when applied to large-scale naturalistic fMRI data. In particular, our single-subject models perform similarly on held out data to the cross-subject model, obtaining an average classification accuracy around $\%80$.

IV. Discussion

Our study is, to our knowledge, one of the first to successfully apply a generic deep 3D-CNN to naturalistic fMRI for emotion decoding at 7T resolution. Prior fMRI-based emotion decoders have either focused on region-of-interest features [4], non-naturalistic emotional tasks [5], or statistical models with inductive biases [6]. This data-driven strategy allows the model to discover relevant spatiotemporal patterns of emotion across the brain in a naturalistic setting unbounded by anatomical assumptions. By demonstrating that a 3D-CNN can be trained on whole-brain 7T fMRI responses to a complex movie and decode emotional states above chance, we establish a new benchmark for large-scale neural decoding in the emotion domain and highlights the promise of modern deep learning in mapping between brain activity and rich emotional experiences.

There are several key consistencies with neurobiological findings directly observed from the emotion-wise inference results despite our brain-agnostic model. First, our model's performance varies significantly across emotions by individual. Indeed naturalistic emotional responses exhibit high inter-individual variability [7], [8]. Beyond differences in subjective emotional experience, there are also physiological and neural sources of variability. Each person's brain anatomy and functional organization is unique – the exact location and magnitude of emotion-related activations can shift from one brain to another, even if qualitatively the same networks (e.g. limbic system, TPJ, prefrontal cortex) are engaged. One subject might recruit a slightly different constellation of regions or have a different lateralization for a given emotion than another. This functional idiosyncrasy is well recognized as a hurdle in multi-subject fMRI analysis [9] . Consequently, a brain-agnostic CNN might misinterpret inter-individual differences as mere data variance, when in fact each subject has a distinct, reliable pattern for themselves that just doesn't match the group pattern well. Our model's difficulty in generalizing could thus be partly due to person-specific neural signatures of emotion.

Notably, our model consistently performs well detecting fear. Neuroimaging evidence suggests that fear triggers a particularly robust and stereotyped brain response across individuals, making it stand out from other emotions. In fMRI

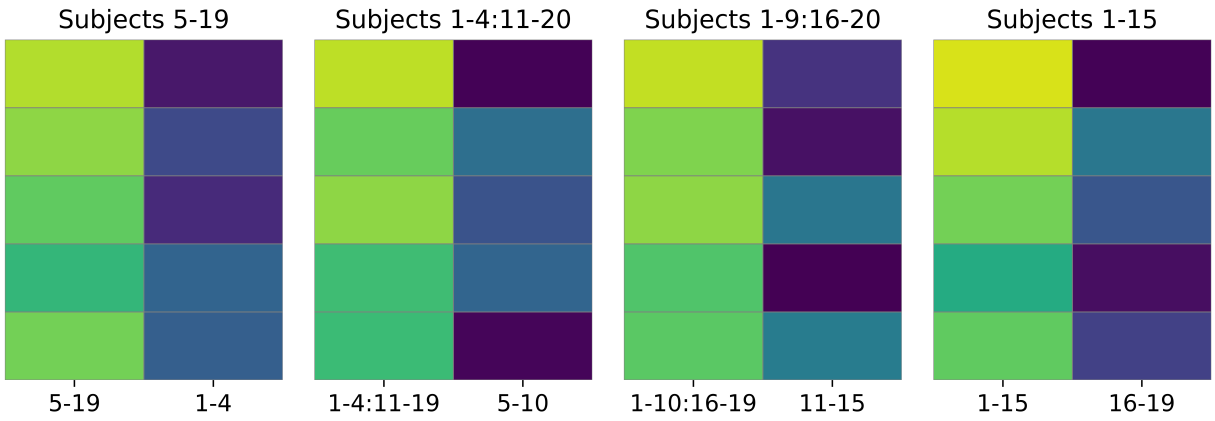

Fig. 6: Emotion-wise model accuracy on held-out subjects. The sub-figure headers are the subjects the model was trained on.

studies, negatively valenced, high-arousal stimuli (like fear-inducing scenes) drive highly synchronized activity in key emotion-processing regions (e.g. amygdala, insula, midcingulate), showing much greater inter-subject consistency than neutral or positive emotional content [10]. For example, a suspenseful horror film clip elicited nearly identical brain activation patterns across different viewers, indicating that fear evokes a shared neural signature that is easier for a general 3D-CNN model to detect compared to more variable emotional states [11].

Moreover, the clinical applications of inference are promising. Notably, the single-subject model trained exclusively on subject 19 performs poorly in classifying love scenes compared to other single-subject models. Indeed, our model misclassifies these scenes as exhibiting sadness in $\%42.1$ of fMRI labels, love in $\%31.6$ of labels and happiness in $\%26.3$ of labels. These misclassifications for subject 19, suggesting potentially atypical or "misaligned" neural responses during love scenes, are reminiscent of findings in clinical populations where aberrant emotional processing signals appear in fMRI data [12]. By analyzing these misclassifications, we gain insight into this subject's unique emotional responses, potentially identifying neural differences or deficits.

Finally, models that generalize to unseen subjects could enable the development of consumer applications leveraging neural data for human preferences. However, when training on a subset of subjects and inferring on held out subjects, our model accuracy collapses. These findings highlight the need for further techniques to enhance cross-subject generalization.

V. Conclusion

Our findings establish the promise of brain-agnostic 3D-CNNs in decoding emotional states from high-resolution, naturalistic 7T fMRI data. Our learned consistency with neurobiological theory confirms the quality of our learned representations. On the other hand, we also highlight the challenges in achieving robust cross-subject generalization and held-out subject inference. Future work can focus on more sophisticated alignment techniques, data augmentation, and larger, more diverse datasets to further improve the generality and reliability of deep learning–based emotion decoding.

References

[1] M. Hanke, F. Baumgartner, P. Ibe *et al.*, "A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie," *Sci Data*, vol. 1, p. 140003, 2014.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," 2015. [Online]. Available: https://arxiv.org/abs/1412.0767

[3] S. Hesaraki, "3dcnn," November 2023, accessed: 2025-03-16. [Online]. Available: https://medium.com/@saba99/3d-cnn-4ccfab119cc2

[4] G. Lettieri, G. Handjaras, E. Ricciardi, A. Leo, P. Papale, M. Betta, P. Pietrini, and L. Cecchetti, "Emotionotopy in the human right temporo-parietal cortex," *Nature Communications*, vol. 10, no. 1, p. 5568, 2019.

[5] M. Tchibozo, D. Kim, Z. Wang, and X. He, "Emotional brain state classification on fmri data using deep residual and convolutional networks," 2022. [Online]. Available: https://arxiv.org/abs/2210.17015

[6] H. Saarimäki, L. F. Ejtehadian, E. Glerean, I. P. Jääskeläinen, P. Vuilleumier, M. Sams, and L. Nummenmaa, "Distributed affective space represents multiple emotion categories across the human brain," *Social Cognitive and Affective Neuroscience*, vol. 13, no. 5, pp. 471–482, 2018.

[7] P. Tovote, J. Fadok, and A. Lüthi, "Neuronal circuits for fear and anxiety," *Nature Reviews Neuroscience*, vol. 16, pp. 317–331, 2015.

[8] T. Steimer, "The biology of fear- and anxiety-related behaviors," *Dialogues in Clinical Neuroscience*, vol. 4, no. 3, pp. 231–249, 2002.

[9] P.-H. Chen, J. Chen, Y. Yeshurun, U. Hasson, J. V. Haxby, and P. J. Ramadge, "A reduced-dimension fmri shared response model," in *Neural Information Processing Systems*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:10143260

[10] L. Nummenmaa, E. Glerean, M. Viinikainen, I. P. Jääskeläinen, R. Hari, and M. Sams, "Emotions promote social interaction by synchronizing brain activity across individuals," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 24, pp. 9599–9604, 2012. [Online]. Available: https://doi.org/10.1073/pnas.1206095109

[11] Y. Wang and Y. Wang, "A neurocinematic study of the suspense effects in hitchcock's psycho," *Frontiers in Communication*, vol. 5, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomm.2020.576840

[12] M. L. Phillips, W. C. Drevets, S. L. Rauch, and R. Lane, "Neurobiology of emotion perception ii: Implications for major psychiatric disorders," *Biological Psychiatry*, vol. 54, no. 5, pp. 515–528, 2003.