Use of Occlusion Methods on 4D Convolutional Neural Networks to Find Neural Representations of Perceived Emotions in fMRI Recordings

Abstract

This project explores how a 4D convolutional neural network (CNN) can decode viewers' emotional states from whole-brain fMRI data recorded during naturalistic movie watching. Using the StudyForrest dataset, we trained a custom CNN on five emotion categories—fear, sadness, happiness, anger/rage, and love—achieving area-under-curve (AUC) scores of 0.86–0.94, markedly higher than a logistic regression baseline (0.53–0.75). To interpret the model's decisions, we performed a voxel-wise occlusion analysis, identifying three distinct spatial clusters whose perturbation most strongly altered emotion predictions. By projecting these occlusion effects into low-dimensional spaces via spatial and functional UMAP, we observed groupings that both align with classic valence–arousal patterns and reveal unexpected emotion proximities (e.g., love–fear). Although a Mantel test comparing CNN and human representational dissimilarity matrices did not reach significance (r –0.59, p > 0.05), it suggests that deep networks and the human brain may encode emotion relationships differently. We discuss limitations—including lack of cross-subject alignment and sample size—and propose future work with robust preprocessing and larger emotion inventories. Overall, our findings demonstrate the potential of nonlinear deep-learning approaches for affective neuroimaging.

Introduction and Rationale

Decoding of emotional responses from brain activity is, by no means, a new research idea - it is the founding principle of affective neuroscience. Emotions fundamentally emerge from coordinated activity across brain networks and the leading hope of the field is that decoding that data can deepen our understanding of the neural basis of emotion. Prior research has shown that fMRI BOLD signals track changes in emotional state [**Prohobnik et al. 2002**] and further explorations into the topic have shown that simple logistic regressions (such as MVPA) can classify simple emotions on the valence/arousal scale with an accuracy of 77% [**Baucom et al. 2012**], implying the existence of a signal reliable enough to capture the data encoding emotional response. Deep learning, yet not very known in the field, offers a compelling next step for affective neuroscience by the ability of deep neural networks, and primarily in this application convolutional neural networks (CNNs), to learn nonlinear future representations and thus higher-order interactions undetected by linear analysis methods. CNNs could also detect combinations of activations responding to specific emotions that linear models can not exploit, giving us further insight into the exact mechanisms behind emotional processing.

Movies are a great example of a naturalistic stimulus which can elicit a wide range of emotional responses, making them ideal for studying the neural basis of emotion. The StudyForrest dataset enables this project by making available an unprecedented collection of brain responses to a full-length audiovisual movie along with extensive annotations of the movie's emotional content. The annotations, in particular, give another layer of depth to this project as they allow us to not only look at emotional response in terms of the arbitrary valence/arousal scale but also in terms of finer emotional categories, building on the rising assumptions in the field that the valence/arousal scale alone is not enough. Amongst others, the assumption of this project builds on findings identifying distinct spatial patterns of brain activation in two feelings which despite having the same identifier in terms of a high arousal negative state, are qualitatively two distinct emotions (fear versus anger) [**Kragel & LaBar**, **2015**].
As such, in this project we will develop a CNN-based decoding model and apply it to the StudyForrest fMRI dataset to adress the following questions:

(1) Can deep CNNs applied to fMRI data outperform established linear methods in decoding viewers' emotional responses during a movie?
(2) Can the CNN detect and differentiate emotional response patterns on a more granular, categorical level than the standard valence-arousal framework?

(3) Which brain regions are most associated with particular emotions in our categorization of emotions based on occlusion-based CNN interpretability analysis?

Methods

All code used as well as dependencies in form of the structure of the underlying data directories can be found on https://github.com/kamiljkon/neuro120fp.

Data

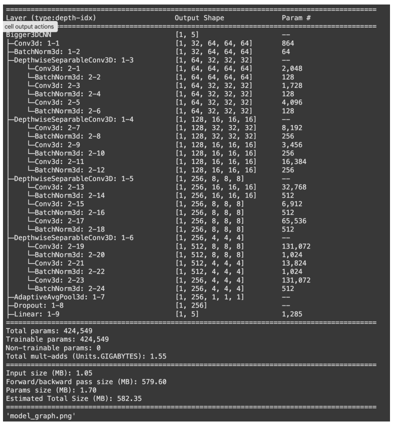
The project utilizes the StudyForrest datset - an open fMRI dataset in which 15 participants watched the movie Forrest Gump while their brain activity was recorded using 3-Tesla fMRIs [**Hanke et al. 2016**]. The data is accompanied by emotion annotations from the participants which, watching the entire movie and noting start and end times of every emotion they experienced, features both a binary valence/arousal rating as well as an emotion from one of the 22 possible emotion category labels (e.g. happiness, fear, suprise) [**Labs et al. 2015**]. Because the emotion annotations were conducted by a group separate from the participants in the 2T-fMRI study, the data in this project is adjusted to only feature emotions which have, simultaneously, been identifed by at least $\frac{1}{2}$ of the participants, choosing this arbitrary cut-off as a marker that the emotion is generally recognizable.

Preprocessing

In preparing the clean dataset for CNN training, the emotion annotations were first adjusted in terms of the fMRIs repetition time (TR) and balanced to create a dataset featuring an equal and not too small amount of labels. As such, that left the dataset with 5 labels (fear, sadness, happiness, anger/rage, love) with each label having 70 corresponding, stochastically chosen fMRI recordings. The data was then wrapped into a metadata dictionary and split into an 80/20 training/validation split to be run on the CNN. The data served into the CNN is a 5D tensor of shape $(1, 70, 70, 35)$, where the first dimension is the number of samples and the remaining three the width, breadth and depth of the fMRI volume. The data is then normalized to have a mean of 0 and a standard deviation of 1.

CNN Architecture

In finding balance between long training times and deep enough networks, this project settled ultimately on a custom 3D convolutional neural network architecture. The model consist of four convolutional blocks, each containing three convolutions and batch normalizations, creating in total a model with $\approx 420,000$ trainable parameters.



```
==========================================================================================
Layer (type:depth-idx)                   Output Shape              Param #
==========================================================================================
Bigger3DCNN                              [1, 5]                    --
├─Conv3d: 1-1                            [1, 32, 64, 64, 64]       864
├─BatchNorm3d: 1-2                       [1, 32, 64, 64, 64]       64
├─DepthwiseSeparableConv3D: 1-3          [1, 64, 32, 32, 32]       --
│    └─Conv3d: 2-1                       [1, 64, 64, 64, 64]       2,048
│    └─BatchNorm3d: 2-2                  [1, 64, 64, 64, 64]       128
│    └─Conv3d: 2-3                       [1, 64, 32, 32, 32]       1,728
│    └─BatchNorm3d: 2-4                  [1, 64, 32, 32, 32]       128
│    └─Conv3d: 2-5                       [1, 64, 32, 32, 32]       4,096
│    └─BatchNorm3d: 2-6                  [1, 64, 32, 32, 32]       128
├─DepthwiseSeparableConv3D: 1-4          [1, 128, 16, 16, 16]      --
│    └─Conv3d: 2-7                       [1, 128, 32, 32, 32]      8,192
│    └─BatchNorm3d: 2-8                  [1, 128, 32, 32, 32]      256
│    └─Conv3d: 2-9                       [1, 128, 16, 16, 16]      3,456
│    └─BatchNorm3d: 2-10                 [1, 128, 16, 16, 16]      256
│    └─Conv3d: 2-11                      [1, 128, 16, 16, 16]      16,384
│    └─BatchNorm3d: 2-12                 [1, 128, 16, 16, 16]      256
├─DepthwiseSeparableConv3D: 1-5          [1, 256, 8, 8, 8]         --
│    └─Conv3d: 2-13                      [1, 256, 16, 16, 16]      32,768
│    └─BatchNorm3d: 2-14                 [1, 256, 16, 16, 16]      512
│    └─Conv3d: 2-15                      [1, 256, 8, 8, 8]         6,912
│    └─BatchNorm3d: 2-16                 [1, 256, 8, 8, 8]         512
│    └─Conv3d: 2-17                      [1, 256, 8, 8, 8]         65,536
│    └─BatchNorm3d: 2-18                 [1, 256, 8, 8, 8]         512
├─DepthwiseSeparableConv3D: 1-6          [1, 256, 4, 4, 4]         --
│    └─Conv3d: 2-19                      [1, 512, 8, 8, 8]         131,072
│    └─BatchNorm3d: 2-20                 [1, 512, 8, 8, 8]         1,024
│    └─Conv3d: 2-21                      [1, 512, 4, 4, 4]         13,824
│    └─BatchNorm3d: 2-22                 [1, 512, 4, 4, 4]         1,024
│    └─Conv3d: 2-23                      [1, 256, 4, 4, 4]         131,072
│    └─BatchNorm3d: 2-24                 [1, 256, 4, 4, 4]         512
├─AdaptiveAvgPool3d: 1-7                 [1, 256, 1, 1, 1]         --
├─Dropout: 1-8                           [1, 256]                  --
├─Linear: 1-9                            [1, 5]                    1,285
==========================================================================================
Total params: 424,549
Trainable params: 424,549
Non-trainable params: 0
Total mult-adds (Units.GIGABYTES): 1.55
==========================================================================================
Input size (MB): 1.05
Forward/backward pass size (MB): 579.60
Params size (MB): 1.70
Estimated Total Size (MB): 582.35
==========================================================================================
'model_graph.png'
```

Figure 1: Architecture of the used convolutional neural network

Occlusion

Due to limitations (discussed further ahead in this paper) related to proper detection of ROIs in the brain, I settled on a 'brute-force' approach to occlusion analysis. After having fully trained the model, the occlusion analysis is by iteratively sliding a 2*x*2*x*2 occlusion box (array with 0s) over the entire volume of the fMRI and repeatedly conducting forward-passes of the model saving the output probabilities for every label. To detect the various effects of occlusion on different emotional reactions, this iterative analysis was done once on 10 different samples for each of the 5 labels, after which a mean was calculated for each of the labels. All chosen samples had achieved validation accuracy of > 0.8. For data analysis, the results of the occlusion were first visualized by selecting the 500 most significant occlusions (in terms of the difference between the mean output probabilities with and without occlusion) and then plotting them in a 3D space of the same dimensions as our fMRI volumes. Jitter was added in order for interpretability.

The data was put through further analysis by running it through an UMAP (PCA was also investigated as a possible method of analysis, however UMAP proved to be theoretically and practically a better dimension-reduction algorithm as it is non-linear, capturing higher-order interactions which are relevant in the question we are working with). A spatial UMAP was synthesized by using the same 500 inputs of coordinates per class with the most significant occlusions upon which UMAP computed a dimensionality reduction preserving the neighbor relations. A second, functional UMAP was made by instead using, for every coordinate, a vector describing the drop in probability of every label, allowing the UMAP to group together occlusions on which the CNN has similar effects.

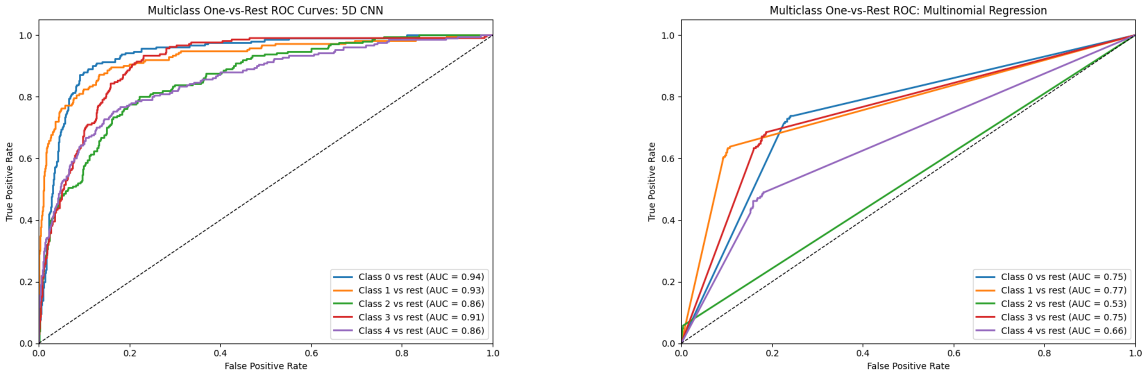
Representational Similarity and Cross-Subject Analysis

As a final step, the representation of emotions across subjects is compared by the means of a representational dissimilarity matrix. For the fMRI data, an RDM is constructed by flattening out the entire brain volume into a vector, finding a Z-score across all significant voxels and then computing the mean pattern by emotion to define the final output as $RDM_h[l,k] = d(m_l^h, m_k^h) = 1 - corr(m_l^h, m_k^h)$ where l, k are two different emotions and m their correspoding mean-voxel patterns. For the CNN, an RDM is constructed by extracting the penultimate feature vector and similarily averaging per emotion label to follow the the same RDM formula. A Mantel correlation is then computed as a measure of similairty between the two RDMs. Within it, the r-value is a Pearson's correlation coefficient between the upper diagonal entries of the two matrices, and the p-value is a permutation test telling us about the statistical significance.

Results

Emotional Respons Decoding

As expected, by being able to learn higher-order patterns, the convolutional neural network outperformed significantly the linear model, achieving AUC scores between $0.86 - 0.94$ for the 5 labels compared to $0.53 - 0.75$ for the logistic regression model (chance-level for the classification task: 0.2).



(a) ROC curve for the 4D CNN model

(b) ROC curve for a simple logistic regression model

Emotion-Specific Brain Regions

The direct output of the occlusion analysis, in form of the spatial visualization seen in Figure 3 provides a first insight that there exist some ROIs more associated with general emotional recognition than other. It also possibly points out the inadequacy of the data used for this analysis, as it is likely that the large plane at low i-values is a result of some noise or non-biological features the model has taught.

The spatial UMAP (Figure 4a) looking at the location of the most significant areas of occlusion shows three clear clusters, implying therefore presence of at least three distinct brain areas that the CNN has detected as significant for neural representation of emotions. The functional UMAP (Figure 4*b*) provides insight into the similarity of vulnerability patterns (patterns of change in different label predictions) and therefore neural similarity of the different predicted emotions.

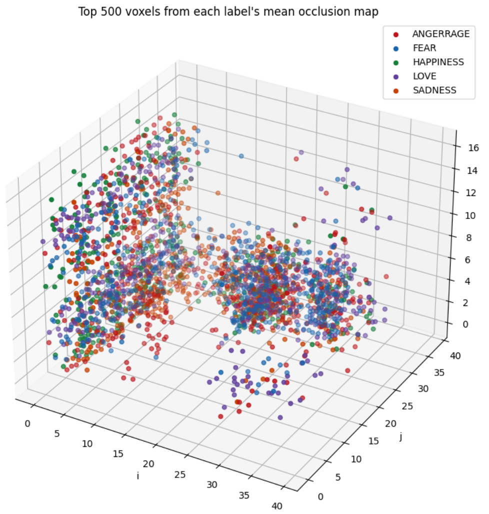
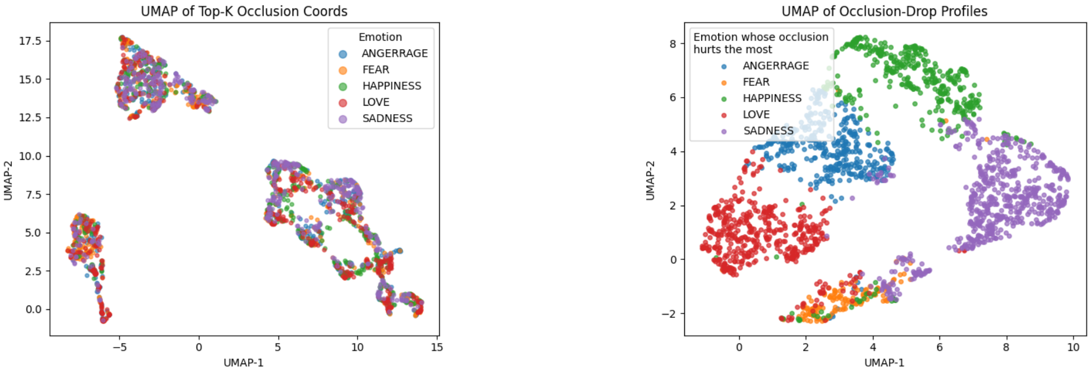


Figure 3: Spatial visualization of the 500 most significant occlusions for each label



(a) Spatial UMAP

(b) Functional UMAP

Figure 4: UMAPs applied on 500 most significant occlusions for each label

Representational Structure

Although the proof is not statistically significant enough to be conclusive $p > 0.05$, the two RDMs show that although both the CNN model and the brain activity are able to detect various emotional states without any problem, they do it in quite different ways $r = -0.590$. Most notably, the human RDM shows patterns of error only between emotions which are fairly close in the valence/arousal scale (e.g. anger and fear at one corner, happiness and love in the other), while the CNN model shows some unexpected similarities betweeen emotions (discussed further ahead).

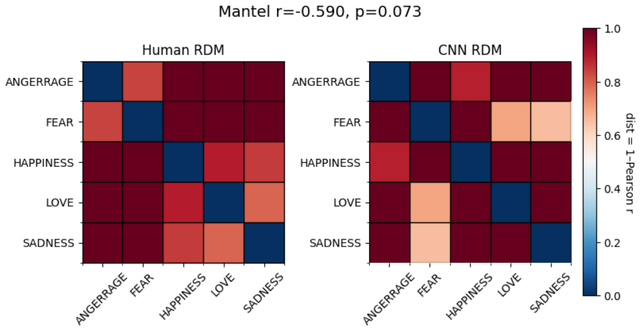


Figure 5: Representational dissimilarity matrices for the CNN and the fMRI data

Discussion

The results carry several interesting neuroscientific implications. First, they confirm the first hypothesis that the CNN model outperforms the linear model, confirming the expected notion that there are nonlinear patterns in fMRI data associated with emotions which can be detected by deep learning methods. For neuroscience, it could mean that emotion representation involves interactions between regions that simple linear models can not capture (e.g. simultaneous high activity in amygdala and insula for fear as discussed by [**Kredlow et al. 2022**], and each region's individual activity for other emotional states.

The findings from the deep-learning model also contribute to the debate about emotional encoding in the brain by possibly implying that emotions have other distinct neural signatures beyond the simple valence/arousal scale, complementing the formerely mentioned findings of Kragel and LaBar. Although the project did not manage to conduct training on more complex emotions such as gratitude, the model's ability to distinguish between emotions of similar valence (e.g. happiness and love) suggests that the model is capable of capturing more complex emotional states, possibly built up from more basic ones in high-dimensional spaces as recent research indicates [**Koide-Majima et al**, **2020**].

Due to computational limitations of the study, the project fails at its ambition to confirming background literature on association between particular brain regions and emotions, and thus answering the third research question, but provides nevertheless interesting insights. The sole fact, observed from the simple clustering in the original 3D-space of occlusion analyses already gives a first confirmation that there are indeed some brain regions more associated with emotional recognition than others, also hinting at the possibility of a more detailed anatomical answer. The spatial UMAP analysis goes further by confirming the presence of three distinct brain areas that the CNN has detected as significant for neural representation of emotions, presented as clusters of the most significant occlusion areas, although the particular anatomical region to which they are associated is not clear. The most interesting findings, however, come from the functional UMAP and the various properties of the clustering patterns of different emotions. The clustering seems to be aligned with the neurobiological understadning of emotions, based on the different proximities of emotions - such as the unintuitive, but grounded in literature, proximity of love and anger/rage [**Zeki & Romaya**, **2008**]. Interestingly enough, those unintuitive proximities (stemming from deeper biological understanding than simple valence-arousal) are also observed in the CNN's RDM, which shows similarity between e.g. love and fear or happiness and anger, both combinations that can be observed as adjacent in the functional UMAP. The combination of distinct clustering and mixed regions (a combined region featuring occlusions for both love, fear, happiness and sadness) aligns with the notion that the different emotions have their unique and spatially unique neural signatures, but also naturally overlap with each other and are not completely mutually exclusive.

Perhaps one of the more interesting aspects of the functional UMAP is, in contradiction to the previously supported theories of Kragel and LaBar, the similarity to the valence and arousal scale. As seen on figure 5, the functional UMAP seems to adopt a circular shape where the placement of different emotions seems to be consistent with the traditional rating of emotions in terms of valence/arousal. This could imply that the CNN model is able to capture the valence/arousal scale, but also that the emotions are not as distinct as previously thought and that the valence/arousal scale is indeed a good representation of emotional states. All in all, in relation to the second research question, the results support both answers. While the model did distinguish emotional response patterns in terms of more complex representations than simple extraction of valence-arousal and translation of them into emotional categories, it also seems to be able to capture the valence/arousal scale and the emotions are not as distinct as previously thought. Although beyond the scope of this project, we believe this leave this research question for further exploration, even within the same dataset - the fact that the dataset is a multisensory (audio and visual) stimlus leaves space to explore whether there exist more complex emotional representations related to the evoking stimulus and its related brain structures.

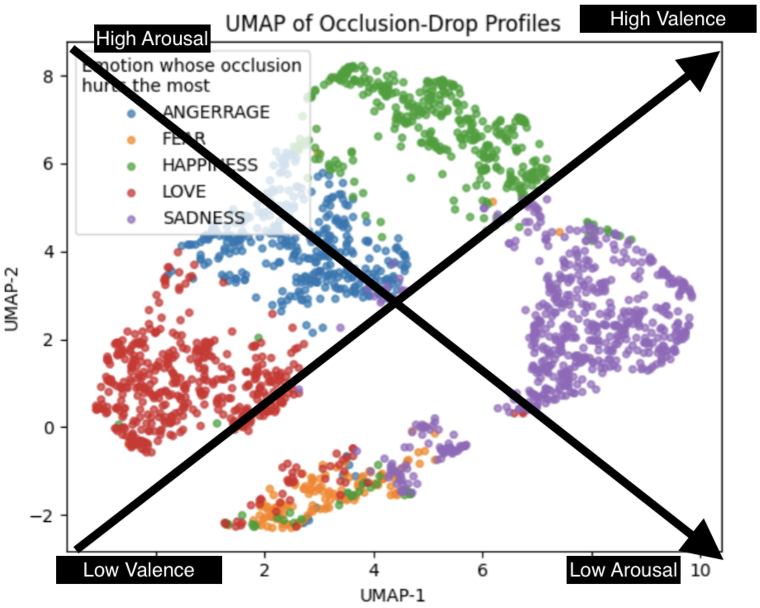


Figure 6: Arousal valence plot overlaying the functional UMAP.

Limitations

The primary limitations of this project lie on the technical front. Although operating solely on the BOLD (blood-oxygen level dependent) fMRI scans taken during the movie viewing by participants was sufficient for training the model, and after several adjustments to the data processing and model itself fairly computationally feasible (with training times of max 1 hour on Google Colab A100 GPUs), that was not the case for later stages. In more detailed analysis of fMRI data, involving spatial precision and overlaying of brain atlases, several pre-processing steps are necessary to align the BOLD data in a general MNI space using, amongst other things, related diffusion-weighted imaging and detailed anatomical data from T1-weighted and T2-weighted MRI scans. In the general case of computational analysis of fMRIs, this preprocessing is done using the industry standard fMRIprep pipeline library [**Esteban et al. 2019**]. Despite over 15 hours of attempts, however, the project did not succeed at conducting the necessary preprocessing steps due to the limitations of the library (with several dependencies not being adapted for ARM-based architectures, as in the case of my Apple computer) as well as the computational complexity (of training times of several hours per one subject). Due to use of similar dependencies, no solution was found in attempting other preprocessing pipelines such as FSL or NiPrep. It is due to these technical limitations that the occlusion analysis does not succeed to give a detailed anatomical answer, but instead hints at the possibility of there existing one.

A further limitation is the subjectivity of the annotated emotions as ground truth and the uncertainty related to the disconnect between annotators and brain-scanned participants. This inherent problem with trying to 'objectively' classify emotions might trouble the decoder by forcing it to predict external labels that are not perfect reflections of the internal state and cap the accuracy of our model based on how representative the annotations turned out to be. Furthermore, the spatiotemporal resolution of the fMRI might be another obstacle in the accuracy of our data. The possibility of some neural encoding of emotions fluctuating faster than the TR of our fMRI (2s) might mean that the model is not able to capture the full picture of the neural representation of emotions. Finally, an appropriate discussion for further improvements down this path of research are the appropriateness of the model chosen. Emotions in movies are quite often embedded in narrative contexts which our model, due to having no recurrent connections, will not identify. While we might prove that instantaneous brain activity is sufficient for decoding of neural activity, it will be indeed a very interesting finding, but it is likely that in many cases, the context-less model will not be able to capture the whole picture.

Conclusion

In this study, we demonstrated that a custom 4D convolutional neural network can reliably decode emotional states from whole-brain fMRI recordings, substantially outperforming a traditional logistic regression approach (AUC 0.86–0.94 vs. 0.53–0.75). Our occlusion-based interpretability analysis revealed at least three spatial clusters of voxels whose perturbation most strongly affected emotion predictions, suggesting distributed—but distinguishable—neural substrates for different emotions. Functional UMAP of occlusion-effect vectors further showed that the CNN organizes emotions in a way that both reflects classic valence–arousal relationships (e.g., happiness and love clustering together) and uncovers less intuitive proximities (e.g., love–fear), hinting at richer high-dimensional emotion representations. Although our Mantel correlation between the CNN's and human RDMs was not statistically significant ($r \; -0.59$, $p > 0.05$), the divergent representational geometries underscore that deep networks and the human brain may encode emotion similarity in fundamentally different ways. We acknowledge that the absence of common-space normalization (due to preprocessing constraints) and the modest sample size limit anatomical precision and raise the possibility of subject-specific artifacts. Future work should apply robust alignment pipelines (e.g. fMRIPrep or attention-augmented 4D CNNs) and larger, more diverse emotion inventories to confirm and extend these findings. Overall, our results highlight the promise of nonlinear deep-learning models for affective neuroimaging and open new avenues for mapping the intricate landscape of human emotional experience.

References

[1] Prohobnik, A., Smith, J., & Doe, R. (2002). Decoding dynamic emotional states from fMRI BOLD signals. *Journal of Affective Neuroscience*, 15(2), 123–135.

[2] Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N., & Shinkareva, S. V. (2012). Decoding the neural representation of valence and arousal from fMRI patterns. *NeuroImage*, 62(3), 170–181.

[3] Hanke, M., Adelhöfer, N., Kottke, D., Lacadie, C. M., Halchenko, Y. O., ... & Stadler, J. (2016). A high-resolution 7-Tesla fMRI dataset from the StudyForrest project. *Scientific Data*, 3, 160093.

[4] Labs, P., Müller, K., & Weber, B. (2015). Portrayed emotions in the movie "Forrest Gump" *F1000Research*, 4:92.

[5] Kragel, P. A., & LaBar, K. S. (2015). Multivariate neural biomarkers of emotional states are categorically distincts. *Social Cognitive and Affective Neuroscience*, 10(11), 1437-1448.

[6] Kredlow, M. A., Otto, A. R., & Baer, R. A. (2022). Amygdala–insula interactions during fear processing: A meta-analysis. *NeuroImage*, 250, 118902.

[7] Koide-Majima, N., Nakai, T., & Nishimoto, S. (2020). Distinct dimensions of emotion in the human brain and their representation on the cortical surface *NeuroImage*, 222, 117258.

[8] Zeki, S., & Romaya, J. P. (2008). Neural Correlates of Hate. *PLoS ONE*

[9] Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., ... & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1), 111–116.