Representation-First Emotion Decoding from Naturalistic 7T fMRI

Lunger, J.*
University of Toronto
Toronto, CA
j.lunger@mail.utoronto.ca

Germann, J.
University Health Network
Toronto, CA
Jurgen.Germann@uhn.ca

Abstract

We present a scalable signal processing framework for measuring naturalistic emotional responses in 7T fMRI using 3D convolutional neural networks. Our model learns low-dimensional representations of affective activity from whole-brain recordings during narrative-driven auditory stimulation, recovering structure consistent with valence–arousal dimensions in affective neuroscience. By prioritizing emotional representation learning over anatomical interpretability, the model maps neural activity across individuals into a shared latent space aligned with canonical affective geometry, enabling scalable cross-subject analysis without region-specific assumptions. We also observe subject-specific deviations in these representations that may capture individual differences in emotion processing, suggesting opportunities for downstream interpretation and personalized analysis. This work establishes a scalable deep learning framework for emotion-aware representation learning in fMRI.

1 Introduction

Decoding emotional responses from neural data has broad applications in affective computing [19], mental health [8], and brain-computer interfaces [17]. While deep learning has shown success in modeling emotion from text, audio, and video [4], decoding affect directly from neural signals remains challenging, particularly in high-dimensional modalities like fMRI. Compared to electroencephalography (EEG), fMRI offers higher spatial resolution and access to deep cortical and subcortical regions implicated in affective processing, but presents unique signal processing challenges due to its low temporal resolution, subject variability, and noise [14, 13].

Affective neuroscience has shown that emotional states are structured along low-dimensional axes such as valence and arousal [20, 7], and that fMRI signals reflect these dimensions in distributed activation patterns [21, 22]. Much prior work in emotion decoding relies on task-based stimuli or small, controlled datasets [2, 6]. In contrast, naturalistic emotion processing such as experiencing a narrative engages broad and dynamic networks, making it an important yet underexplored setting for emotion decoding [12].

In this work, we present a scalable neural decoding framework for learning emotional response representations from 7T fMRI using 3D convolutional neural networks. Leveraging the StudyForrest dataset [9], which provides whole-brain recordings from 19 participants listening to an audio-description of the film *Forrest Gump*, we obtain a shared latent space projection. Our learned representations are consistent with neuro-psychological theory of affect while capturing subject-specific variation that may reflect differences in emotional response.

*This work was completed through the University of Toronto Machine Intelligence Student Team (UTMIST).

2 Related work

Cross-subject alignment methods such as the Shared Response Model (SRM) demonstrate that fMRI responses can be mapped into common low-dimensional spaces [5], though these have primarily been used for perceptual or semantic representations rather than affective dimensions [10]. Deep learning has been applied to emotion decoding with convolutional and residual networks trained on fMRI responses to static faces [11], which achieve cross-subject generalization but emphasize classification accuracy over the geometry of affective representations. More recently Borriero et. al [1] applied deep learning to the StudyForrest dataset while focusing on neuroanatomical attribution with subject-specific models. In contrast to these approaches, our contribution is to use deep neural networks to learn a shared, interpretable valence–arousal latent space, prioritizing representation learning of emotional responses over neuroanatomical interpretability and enabling cross-subject analysis of both shared structure and individual variability.

3 Methodology

3.1 Data collection and fMRI preprocessing

We use the publicly available 7T fMRI dataset from the StudyForrest project [9], which includes whole-brain recordings from 19 participants listening to an audio-description of the film *Forrest Gump*. Scans were acquired at a spatial resolution of 1.4 mm isotropic with a temporal resolution of 2 seconds. We use the preprocessed and anatomically aligned data provided by the StudyForrest authors. These volumes were motion- and distortion-corrected, then mapped to a common group template using iterative affine and nonlinear registration. This alignment enables voxel-wise correspondence across subjects and supports learning subject-invariant neural representations.

3.2 Annotation preprocessing

Emotion annotations were collected from external observers who rated the perceived emotional content of both the audio-description and film. There were 3 external observer annotations for the audio-description and 9 for the film. The audio-description and film annotations were time aligned and the audio content differed only in narrative elements present in the audio-description during film scenes without spoken word. These labels presented two key challenges: (1) most timepoints were rated as emotionally neutral, and (2) inter-observer agreement was often low, with no clear consensus on the dominant emotion. To address these issues, we applied heuristic clustering and a majority-vote consensus threshold to derive high-confidence labels.

Heuristic clustering was applied to group annotations by mapping to the five most frequently observed emotion categories, supported by affective science literature that emphasizes the reliability of coarse-grained emotional dimensions in evoking consistent neural responses [15] [22]. Collapsing fine-grained emotion labels into higher-level clusters mitigates label sparsity and reduces the risk of overfitting to individual annotator idiosyncrasies [7].

Each fMRI sample was assigned an emotion label based on majority vote conditioned on agreement of at least half of the observers, rounded up. This approach is consistent with prior work using crowd-annotated emotion labels, where requiring a minimum consensus threshold has been shown to improve inter-rater reliability and downstream model performance [24], [16]. By focusing on events with higher inter-observer agreement, we ensured that the assigned labels reflect a consistent percept across observers rather than isolated or ambiguous interpretations. After this process, volumes with no emotion label were discarded. Though this reduced the overall dataset size, it increased the signal-to-noise ratio by removing low-certainty samples. This aligns with evidence from neuroimaging studies indicating that strongly perceived emotions yield more robust and discriminable neural representations [14] [13].

These conditions prioritized strong emotional signal over data quantity, allowing the model to learn reliable mappings between emotionally salient perceptual events and their corresponding neural representations. After preprocessing, there were roughly 150 images for each emotion label and subject resulting in around 16,000 images for the full dataset.

Figure 1: Covariance scores for emotion annotations across observers. The majority of annotation mass was allocated to five most common emotions. Heuristic clustering was used to map annotations to these five emotions.

3.3 Training

We trained 3D convolutional neural networks to classify emotion annotations from 7t fMRI images. A separate model was trained on each subject and a single 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. Notably, our model makes no further neuroanatomical assumptions beyond the preprocessing adopted from the original dataset.

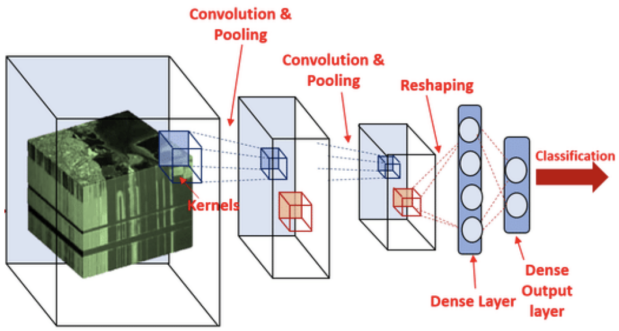

Figure 2: Reproduced from [11]. The network consists of 3 convolutional layers with 3D kernels and ReLU activations. A series of max-pooling operations were applied to downsample spatial dimensions while preserving feature representations. The final convolutional features were flattened and passed through 2 fully connected layers before a softmax classification head.

4 Results

4.1 Classification performance

We observe impressive performance on classification of held-out data from both single-subject and cross-subject models on audio-only and audio-visual annotations with limited compute and hyperparameter tuning. All models obtain an average classification accuracy of roughly 80% on held-out data. We found low variability in model performance across individuals and recording sessions. The single-subject and cross-subject models performed similarly.

4.2 Neuro-psychological consistency in performance

There are several key characteristics of emotion-wise classification performance on held-out data directly consistent with previous neuro-psychological findings. First, our model's performance varies significantly across emotions by individual aligning with evidence of naturalistic emotional responses exhibiting high inter-individual variability [26]. Second, 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 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 [18]. Finally, the cross-subject model showed poor generalization for love, aligning with evidence that love is a highly variable, individually learned response rather than a universal emotion [3], and with neuroimaging findings highlighting its heterogeneous neural representation across individuals [25]. Taken together, the alignment between these performance patterns and established neuro-psychological findings suggests that the model is capturing meaningful neural representations of specific emotions rather than relying on incidental stimulus features.

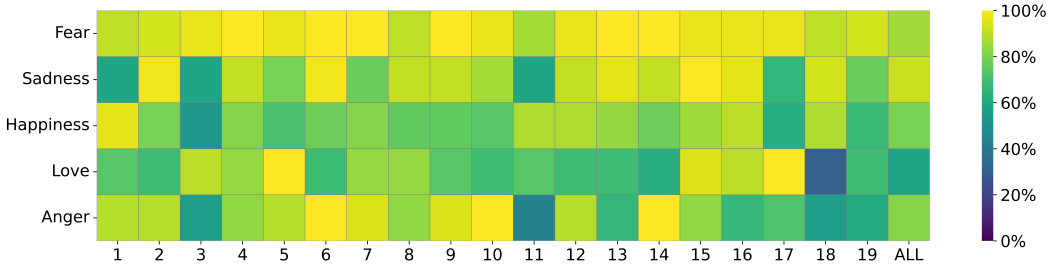

Figure 3: Emotion-wise classification accuracy for cross-subject and single-subject models on held-out data. The x-axis denotes what subject the model was trained on, while the y-axis denotes emotion label and the color-magnitude denotes classification accuracy.

4.3 Representational analysis

We reduce the last hidden layer to two dimensions with PCA, a method well suited for emotion representation since affective states are widely modeled along orthogonal arousal and valence dimensions [20]. Neuroimaging and behavioral studies show that dimensionality reduction often recovers this two-dimensional latent space [14, 21, 7, 4], suggesting that PCA should extract these components from distributed emotion representations.

We use Procrustes alignment to rotate the learned latent space into the canonical valence–arousal basis [10, 23]. For each emotion, we compute centroids from the dimension-reduced cross-subject matrix X_{all}, forming C_{all}. A target matrix Y specifies canonical valence–arousal coordinates. Solving $\min_R ||RC_{all} - Y||^2$ yields the rotation R, which we apply to X_{all} to obtain latents expressed in the valence–arousal basis.

Once aligned, the dimension reduced representations reveal important structural consistency and variability across subjects. First, the fMRI embeddings associated with segments labeled for particular emotions are arranged in a pattern broadly consistent with valence–arousal geometry. This further suggests that our model has uncovered the correct psychological mapping. Second, the relative in-cluster structures of emotions and cluster-wise distances capture subject-specific variation that may reflect differences in emotional response. For many subjects, sadness and love overlapped in valence–arousal space, suggesting shared psychological underpinnings. Cluster centroid positions varied widely: Subject 2 showed tightly grouped centroids with low variability; Subjects 9 and 14 displayed out-of-distribution fear responses, with exceptionally low valence and high arousal, respectively. Subject 11's clusters closely matched the population mean, potentially reflecting a normative emotional profile. Such variability, particularly in fear and in the overlap of sadness and love, underscores the need for future work linking these neural patterns to psychological profiles or clinical histories of individual subjects.

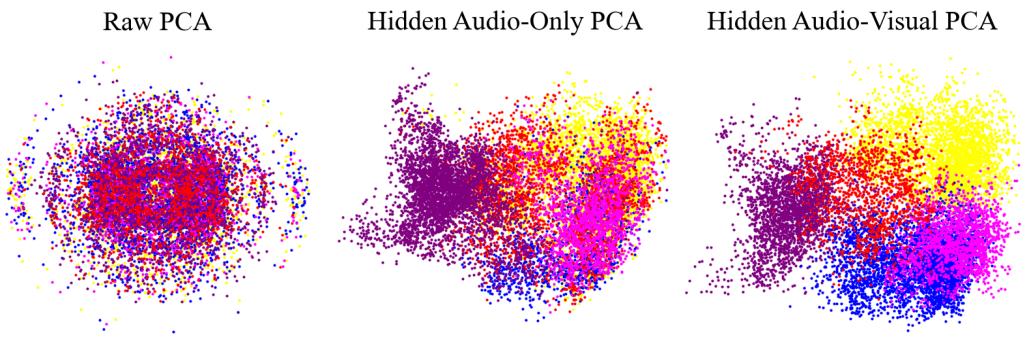

Figure 4: PCA plots of raw 7t fMRI vs learned representations from the last hidden layer before network classification head. Emotion colors are consistent with legend below. Images trained on the audio-visual annotations showed improved emotion coupling, perhaps due to the greater number of external observers.

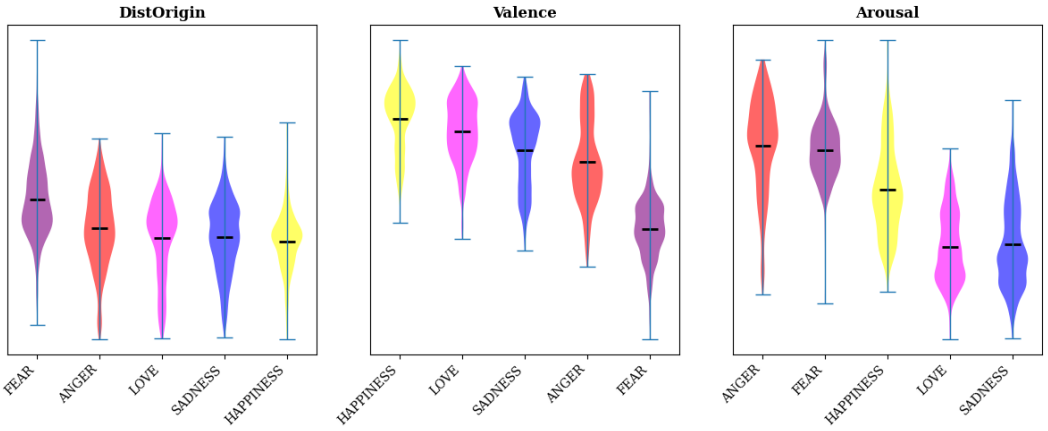

Figure 5: Centroid statistics for Procrustes aligned PCA. The relative orderings are consistent with affective theory.

5 Discussion

Our results demonstrate that deep neural networks trained on 7T naturalistic fMRI recover latent representations consistent with the canonical valence–arousal framework. By projecting into this basis, we identified both population-level structure and subject-specific deviations, suggesting that affective representations are at once shared and individualized. This duality highlights the promise of latent-space approaches for linking neural signals to psychological constructs without requiring neuroanatomical-interpretability.

Several limitations point to directions for future work. First, our consensus-based annotation strategy improved robustness but reduced data volume; richer labeling protocols and larger annotator pools could strengthen supervision. Second, we did not address appraisal theories of emotion, which conceptualize affect as arising from multi-dimensional evaluations of events and may require richer annotation schemes or model architectures to capture. Third, while we observed meaningful inter-subject variability in the learned latent spaces, we did not incorporate clinical or psychological profiles to validate whether these differences reflect stable individual traits or affective styles. Finally, extending beyond 3D CNNs to sequence models that explicitly capture temporal dependencies could better reflect the dynamic nature of emotion processing in narratives. We hope future work can address these limitations. This work is likely to have a positive social impact by enabling affect-aware technologies that align with human emotional experience while underscoring the need for responsible deployment that safeguards privacy, mitigates bias, and promotes equitable access.

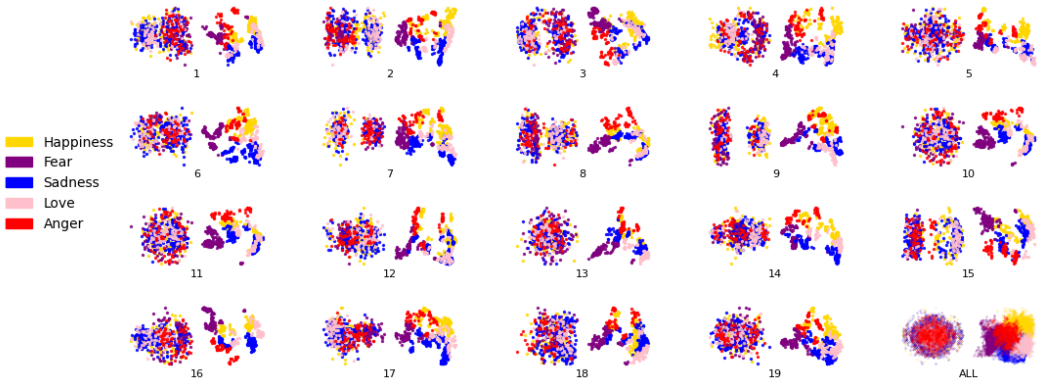

Figure 6: Subject-wise PCA plots of raw fMRI vs learned representations from the last hidden layer before the network classification head with audio-visual annotations. We observe the correct relative arrangement of emotion clusters for each subject under Procrustes alignment together with high structural variability within clusters.

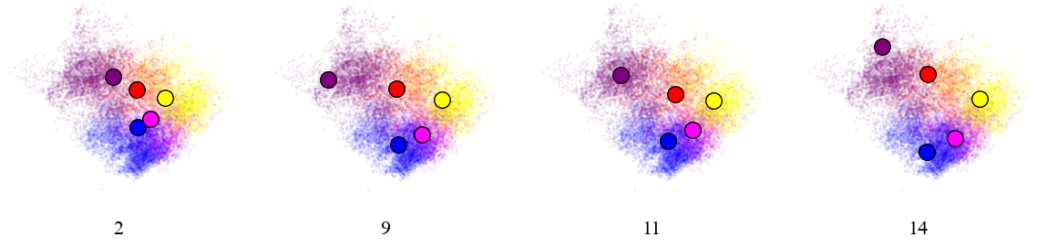

Figure 7: Subject-wise emotion centroid positions. Inter-subject variability in this space is psychologically interpretable and may reflect individual differences in emotion processing.

References

[1] Alessio Borriero, Martina Milazzo, Matteo Diano, Davide Orsenigo, Maria-Chiara Villa, Chiara Di Fazio, Marco Tamietto, and Alan Perotti. Explainable emotion decoding for human and computer vision. In *xAI (2)*, pages 178–201, 2024. URL `https://doi.org/10.1007/978-3-031-63797-1_10`.

[2] K. H. Brodersen, C. R. Ong, M. E. Stephan, and J. M. Buhmann. Generative embedding for model-based classification of fmri data. *PLOS Computational Biology*, 7(6):e1002079, 2011. doi: 10.1371/journal.pcbi.1002079.

[3] Elvira Burunat. Love is not an emotion. *Psychology*, 7:1883–1910, 2016. doi: 10.4236/psych.2016.714173. URL `https://doi.org/10.4236/psych.2016.714173`.

[4] L. J. Chang, P. J. Gianaros, A. E. Manuck, A. Krishnan, and T. D. Wager. A map of subjective feelings. *Trends in Cognitive Sciences*, 19(11):758–770, 2015. doi: 10.1016/j.tics.2015.09.002.

[5] Po-Hsuan Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James V. Haxby, and Peter J. Ramadge. A reduced-dimension fmri shared response model. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 460–468, Cambridge, MA, USA, 2015. MIT Press.

[6] J. Chikazoe, S. Lee, R. C. Kriegeskorte, K. S. Anderson, J. W. DeYoung, and A. L. Chiu. Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience*, 17:1114–1122, 2014. doi: 10.1038/nn.3749.

[7] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences (PNAS)*, 114(38):E7900–E7909, 2017. doi: 10.1073/pnas.1702247114.

[8] Amit Etkin and Tor D Wager. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in ptsd, social anxiety disorder, and specific phobia. *American Journal of Psychiatry*, 164(10):1476–1488, 2007. doi: 10.1176/appi.ajp.2007.07030504.

[9] Michael Hanke, Falko Baumgartner, Peter Ibe, and et al. A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie. *Scientific Data*, 1:140003, 2014. doi: 10.1038/sdata.2014.3. URL https://doi.org/10.1038/sdata.2014.3.

[10] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011. doi: 10.1016/j.neuron.2011.08.026.

[11] Saba Hesaraki. 3d cnn. https://medium.com/@saba99/3d-cnn-4ccfab119cc2, 2023. Accessed: 2025-08-10.

[12] Jiaxuan Ke, Huan Song, Zhenghan Bai, Monica D. Rosenberg, and Y. Chen Leong. Dynamic brain connectivity predicts emotional arousal during naturalistic movie-watching. *PLOS Computational Biology*, 21(4):e1012994, 2025. doi: 10.1371/journal.pcbi.1012994. URL https://doi.org/10.1371/journal.pcbi.1012994.

[13] J. Kim, D. Shin, B. Jeong, and S. Lee. Neural representation of emotional experience across emotion categories. *NeuroImage*, 226:117524, 2021. doi: 10.1016/j.neuroimage.2020.117524.

[14] P. A. Kragel and K. S. LaBar. Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. *Emotion*, 15(4):487–496, 2015. doi: 10.1037/emo0000048.

[15] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett. The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, 35(3):121–143, 2012. doi: 10.1017/S0140525X11000446.

[16] S. Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion Measurement*, pages 201–237. Elsevier, 2016.

[17] Md Rakibul Mowla, Rachael I Cano, Katie J Dhuyvetter, and David E Thompson. Affective brain-computer interfaces: Choosing a meaningful performance measuring metric. *Computers in Biology and Medicine*, 126:104001, 2020. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2020.104001. URL https://www.sciencedirect.com/science/article/pii/S0010482520303322.

[18] Lauri Nummenmaa, Heini Saarimäki, Enrico Glerean, Athanasios Gotsopoulos, Iiro P. Jääskeläinen, Riitta Hari, and Mikko Sams. Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks. *NeuroImage*, 102(Pt 2):498–509, November 2014. doi: 10.1016/j.neuroimage.2014.07.063. URL https://doi.org/10.1016/j.neuroimage.2014.07.063.

[19] Rosalind W. Picard. Affective computing. 1997. URL https://api.semanticscholar.org/CorpusID:262931595.

[20] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980. doi: 10.1037/h0077714.

[21] H. Saarimäki, M. G. Glerean, E. Hari, and L. Nummenmaa. Discrete neural signatures of basic emotions. *Cerebral Cortex*, 26(6):2563–2573, 2016. doi: 10.1093/cercor/bhv086.

[22] H. Saarimäki, M. G. Glerean, E. Hari, and L. Nummenmaa. Distributed affective space represents multiple emotion categories across the human brain. *Social Cognitive and Affective Neuroscience*, 13(5):471–482, 2018. doi: 10.1093/scan/nsy018.

[23] Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

[24] B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of EMNLP*, pages 1467–1478, 2011.

[25] Hsuan-Chih Shih, Mei-En Kuo, Chia-Wei Wu, Yung-Pin Chao, Hao-Wei Huang, and Ching-Mo Huang. The neurobiological basis of love: A meta-analysis of human functional neuroimaging studies of maternal and passionate love. *Brain Sciences*, 12(7):830, 2022. doi: 10.3390/brainsci12070830. URL `https://doi.org/10.3390/brainsci12070830`.

[26] Philip Tovote, Jonathan P. Fadok, and Andreas Lüthi. Neuronal circuits for fear and anxiety. *Nature Reviews Neuroscience*, 16(6):317–331, 2015. doi: 10.1038/nrn3945. URL `https://doi.org/10.1038/nrn3945`.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We tried to carefully ensure the strength of evidence matched the strength of claims in our work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A discussion of limitations is included in the discussion section of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our work does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Data, methodology and training specifications are disclosed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification:

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details and model architecture are included in the Training section of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

Justification: No statistical claims are made in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   All training was performed on a Quadro 4000 GPU as disclosed in the training section. Because the compute required is very minimal for deep learning research, we refrain from providing further compute requirements.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the code of ethics and believe our work is not in 911 violation of any rules.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: A discussion of the social impact of this work is included at the conclusion of the discussion section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original dataset is properly cited within the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

Justification: The paper uses human dataset from the StudyForrest dataset which received approval and followed ethical guidelines as necessary.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.