Article │ Received 23 December 2024; Accepted 7 March 2025; Published 27 March 2025
https://doi.org/10.55092/aias20250002

Bridging human emotion processing and deep neural networks: insights from representational similarity analysis

Lu Nie[1], Ke Chen[2,3], Yue Li[1], Yonghong Tian[3,4] and Yixuan Ku[1,3,]*

[1] Guangdong Provincial Key Laboratory of Brain Function and Disease, Center for Brain and Mental Well-being, Department of Psychology, Sun Yat-sen University, Guangzhou, China
[2] School of Artificial Intelligence, Sun Yat-sen University, Guangzhou, China
[3] Peng Cheng Laboratory, Shenzhen, China
[4] School of AI for Science, Shenzhen Graduate School, Peking University, Beijing, China

* Correspondence author; E-mail: kuyixuan@mail.sysu.edu.cn.

Highlights:

- Unifying human and AI emotion recognition (ER) via RSA.
- EmoNet mimics human emotional processing hierarchically.
- RSA enhances AI model interpretability in ER tasks.

Abstract: Emotion is a complex psychophysiological response to external stimuli, essential for human survival, social interaction, and human-computer interaction. Emotion recognition plays a critical role in both biological systems and artificial agents. However, existing research often treats these systems independently, limiting opportunities for interaction and hindering the development of more advanced models. This study employs representational similarity analysis (RSA) to bridge this gap by comparing emotional representations between the human brain and neural networks, aiming to improve understanding of emotion recognition in deep learning models. By correlating the emotion recognition model EmoNet with EEG signals from the human brain during emotional image processing and introducing AlexNet for comparison, we reveal EmoNet's human-like representation for emotional images and its hierarchical structure for emotion recognition. The results show that RSA effectively aligns human emotional processing with deep neural networks, offering new avenues for improving the interpretability and performance of emotional AI models. Moreover, they underscore EmoNet's potential to simulate human emotional responses, paving the way for future research to enhance emotion recognition models by incorporating human emotional evaluations into their training processes, thereby improving efficiency and specificity.

Keywords: emotion recognition; EEG; EmoNet; ANN

1. Introduction

Emotion is a complex psychophysiological response to external stimuli, intricately linked to human survival, development, and social interaction [1]. For individuals, emotion permeates every aspect of human life, influencing perception, decision-making [2,3], and even mental health, with emotional abnormalities often imposing significant burdens on daily functioning [4]. In the age of artificial intelligence (AI), emotion plays an increasingly pivotal role in human-computer interaction [5–7]. Interestingly, while emotions were traditionally seen as biological phenomena, recent studies have begun exploring whether intelligent systems can perceive emotions as well [5,8,9], potentially enabling better integration with human society. Understanding how emotions are processed and recognized remains a fundamental and challenging question in both affective neuroscience, which focuses on humans, and affective artificial intelligence, which focuses on computers [6,9].

Research about affective neuroscience and AI has provided valuable insights into the mechanisms of emotion processing. However, a significant gap exists between these two fields, limiting our ability to fully understand and advance the science of emotion recognition. Researchers in affective neuroscience have long debated whether emotion processing is primarily governed by subcortical structures such as the amygdala or by sensory cortices. Although brain damage and neuroimaging studies have emphasized the amygdala's critical role in emotion detection [10,11], recent research has highlighted the sensory cortices' involvement in emotion processing [12,13]. Indeed, comprehensive evidence remains elusive, as fully isolating the interactions between these regions in human studies is challenging. The constructionist theory of emotion [2,14], which posits a hierarchical process beginning with low-level features generated by subcortical structures and culminating in higher cortical regions that categorize emotions, provides a potential framework for resolving this debate. Yet, empirically uncovering the dynamic and interrelated sub-stages of emotion processing remains an ongoing challenge.

In parallel, research leveraging artificial neural networks (ANNs) has further contributed to our understanding of emotion recognition. For example, EmoNet, a convolutional neural network (CNN) trained on emotional stimuli, has demonstrated its ability to classify emotional images into 20 distinct categories [15]. Notably, EmoNet is designed to mimic the human sensory cortex, deliberately avoiding input from subcortical structures. EmoNet employs a multilayer convolutional structure to extract image features, which simulates the human visual system's gradual processing of low-level features (such as edges and colors) to high-level features (such as shapes and objects). This approach allows the model to not only classify complex images but also to accurately predict human participants' BOLD signal responses when exposed to the same stimuli [15]. Interestingly, emotional detection was observed in object recognition networks that had not been specifically trained for emotion classification [8]. However, like the human brain, ANNs operate as "black boxes", complicating efforts to interpret the neural unit organization across layers. The optimization of these networks through hyperparameters further obfuscates the specific mechanisms behind emotion classification [16,17]. As a result, the lack of interpretability in ANN models severely hampers the development of more effective and innovative emotion recognition systems. Taken together, although both affective neuroscience and computational models using ANNs have advanced in emotion processing, the lack of communication and cross-pollination between these fields hinders progress. This gap is particularly evident with the growing prominence of brain-inspired neural networks [17–21], and the increasing emphasis on AI as a tool to understand the human brain [22,23]. Strengthening the link between these fields is thus crucial for advancing the science of affective computation.

One promising avenue for bridging this gap is the concept of convergent evolution, borrowed from comparative biology, which describes how distant biological systems evolve similar functions to address shared challenges [17,19]. Recently, this idea has been applied to brain-machine research, including studies on orientation tuning in the primary visual cortex and CNN neurons for object recognition [17], as well as the analogy between grid cell structures in the entorhinal cortex and JPEG image compression [17]. In this context, representational alignment—a technique used to align brain and model representations—offers a practical method for mapping convergent evolutionary traits across both domains [7,24]. This approach aligns biological and artificial intelligence systems through task alignment and cross-modal representational similarity analysis [25,26], improving both model performance and interpretability. For instance, brain-inspired networks have outperformed traditional models, with representations that closely resemble the activity of primate IT neurons [21,27]. Furthermore, neural networks have provided insights into brain information processing, such as the fatigue mechanism of face repetition suppression [23] and the transition from visual to semantic information processing [22]. These frameworks highlight the value of improving communication between biological and artificial intelligence systems. However, the lack of cross-talk in affective computation between the human brain and machines leaves us uncertain about how closely machine emotion recognition mirrors the brain's emotional processing. Addressing this gap is crucial for advancing emotion computation models. Moreover, steady state visual evoked potential is utilized to detect the response of the visual cortex to emotional information, which aligns with the neural network's processing of emotional stimuli [28]. This parallel suggests a potential similarity between human and artificial systems in emotion representation.

This study seeks to address this gap by employing representational similarity analysis (RSA) to compare the representational geometry of emotional stimuli in the human brain and deep neural network models. Our goal is to enhance understanding of how deep network models perform emotion recognition. Specifically, we compare EmoNet, a neural network model validated to distinguish emotion categories, with the dynamic processing of emotional images in human brain EEG signals. Additionally, by introducing AlexNet, a model trained for object recognition, we aim to uncover potential reasons behind EmoNet's ability to generate fine-grained emotion classifications. This study makes the following key contributions:

- Proposes a representational similarity analysis method to understand affective computation in models, inspiring model development in emotion recognition.
- Demonstrates the hierarchical structure of neural networks in combination with human EEG data.
- Compares the representational structures of emotional images in both humans and models, revealing emotion-specific processing.

2. Methods

2.1. Participants

This study involved twenty-five healthy college students from Sun Yat-sen University, including 15 females, with a mean age of 20.8 years (SD = 2.0), among whom one participant is left-handed. The sample size was determined based on a previous similar EEG decoding study [29]. All participants had normal or corrected-to-normal vision and reported no history of psychiatric or neurological disorders. Written informed consent was obtained from all participants before the experiment, and they were compensated ¥80 for their participation. Additionally, participants completed the Beck Depression Inventory (BDI-II; Steer, 1996) to report their level of depression (M = 7.04, SD = 6.52). The Institutional Review Board of Sun Yat-sen University approved the current study and adhered to the Declaration of Helsinki.

2.2. Experimental paradigm and stimuli

The nine images used in this study were selected from the International Affective Picture Systems (IAPS) [30]. Prior to the experiment, all participants provided ratings for each image on emotional valence (pleasantness *vs.* unpleasantness) and arousal (calm *vs.* tension), using a scale from 1 to 9 (1 = 'extremely unpleasant' or 'calm' and 9 = 'extremely pleasant' or 'tense').

The experimental procedure was programmed using the Psychophysics Toolbox [31] based on Matlab 2020a (Mathworks, Natick, MA, USA). Stimuli were displayed on an LED monitor (AOC G2460P) with a 120 Hz refresh rate and 24-inch screen resolution of 1024 × 768, set against a gray background. All participants were seated 60 cm from the monitor, viewing the stimuli at a visual angle of approximately 12 × 9°.

A scene working memory task adapted from a previous study [29] was employed, with EEG signals recorded throughout. The experiment consisted of 1080 trials, including 864 main trials (96 repetitions per picture) and 72 test trials. Only the main trials were used for the subsequent analysis, while the test trials and the trials immediately before and after them were excluded from analysis. This exclusion helped prevent memory reports from influencing the EEG signal during the processing of subsequent stimuli and reduced trial-by-trial variability [31]. Detailed experimental procedures can be found in a prior publication. Each trial began with a white fixation cross for 1000 ms, followed by a memory image target for 1000 ms. A white fixation dot then appeared for 1000 ms (working memory delay), during which participants were instructed to remember the affective valence and semantic content of the image for occasional memory tests (7% of trials). In the test trials, a test screen presented three image options for participants to report either the affective valence or semantic category of the prior image. Stimulus sequences were pseudo-randomized, and high-frequency flickering was applied during stimulus encoding, which did not result in visible flickering for the participants [28,32]. This aspect is not central to the focus of current study.

2.3. EEG acquisition and preprocessing

EEG signals were continuously recorded using a 64-electrode Ag/AgCl electrode cap arranged according to the extended International 10/20 system. A Neuroscan SynAmp2 amplifier was used for signal acquisition in DC mode, with data recorded via Curry 7 software at a sampling rate of 1000 Hz. Electrode impedances were kept below 10 kΩ throughout the experiment. The electrooculogram (EOG) was monitored to track eye movements, with vertical EOG recorded via electrodes placed above and below the left eye, and horizontal EOG recorded via electrodes at the lateral canthi of both eyes. All signals were referenced online to a REF electrode placed at the top of the head.

EEG data were processed and analyzed using the EEGLAB toolbox v14.1.2 [33] and the Fieldtrip toolbox [34]. First, continuous signals were detrended to remove the linear shift. The signals were then band-pass filtered between 0.1 and 90 Hz (roll-off 6dB/octave) and down-sampled to 500 Hz. A 50 Hz notch filter was applied to remove power line interference. The EEG signals were re-referenced offline to the average of the left and right mastoid electrodes and then segmented into epochs from −200 to 2000 ms relative to stimulus onset with the prestimulus 200 ms activity as the baseline. These 2200 ms epochs were manually screened to remove artifacts and then entered into an infomax independent component analysis (runica) [33]. Blink-related components were identified and removed. Epochs with voltage differences exceeding ± 120 μV were automatically rejected to further reduce artifacts.

2.4. Representational similarity analysis

RSA is a multivariate pattern analysis method that provides insights into how the brain represents information. By comparing neural responses to specific stimuli, RSA reveals the geometric structural relationship between them in a high-order space. Using distance measures, RSA enables the comparison of information across different modalities.

Human-rating RDM construction: the IAPS provides 9-point valence and arousal ratings for each image based on human ratings. To construct the human emotional experience RDMs, we calculated the Euclidean distance between pairs of images based on their valence, arousal, and combined ratings (2D vector). We also extracted RDMs for positive, negative, and neutral images for further analysis. The images were categorized according to normative IAPS ratings, with random downsampling used to match the smallest category size. This resulted in a general RDM (1187 × 1187) and three category-specific RDMs (118 × 118).

EmoNet RDM construction: EmoNet, a convolutional neural network designed to classify images into emotion categories [15], was used to extract emotion-related features. The model consists of eight layers, five convolutional layers, and three fully connected layers, which maps the hierarchical structure of the ventral visual stream. We extracted the simulated activations of EmoNet for each IAPS image and reduced the feature dimensions using principal component analysis (PCA), retaining components that explained over 95% of the variance. For each layer of EmoNet, we computed the cosine dissimilarity between the activation vectors corresponding to any two images as the dissimilarity measure, yielding the 1181 × 1181 general EmoNet RDM. Similarly, we extracted specific EmoNet RDMs for positive, negative, and neutral images. To perform similarity analysis with human brain EEG signals, we also constructed an RDM for the 9 images used in the EEG experiment.

AlexNet RDM construction: To examine whether EmoNet exhibits emotion-specific representation, we performed the same analysis on AlexNet activations [35]. AlexNet shares the same model architecture as EmoNet but was originally trained for object recognition rather than emotion classification. We extracted AlexNet's activation patterns for all IAPS images and calculated the cosine dissimilarity between pairs of images to construct a series of RDMs, including a general AlexNet RDM (1187 × 1187) and specific RDMs for different emotional categories (118 × 118). This comparison with EmoNet allows us to test whether EmoNet's emotional specificity differs from a general object-recognition model like AlexNet.

EEG RDM construction: The preprocessed EEG data were used to construct the EEG representation dissimilarity matrix (RDM) using the rdm_cal module in NeuroRA [36] with the default parameters. For each participant, we computed the Pearson correlation between the EEG signals from all channels evoked by any two images at each time point, resulting in a 9 × 9 correlation matrix. The correlation distance (1-correlation) was then calculated to convert the 9 × 9 correlation matrix into an EEG RDM. Each cell in the resulting RDM represented the dissimilarity in scalp distribution between the EEG responses elicited by two images at a particular time point. We used a 5-time point sliding window with a step size to compute the time course of the EEG RDMs.

RSA: We performed RSA to assess the relationship between the different RDMs constructed above (Figure 1). Spearman's correlation was used to evaluate the similarity between the human-derived RDMs (both EEG and behavioral ratings) and those generated by the CNN models (EmoNet and AlexNet). For the EmoNet RDMs, we used partial correlation to control for the influence of other layers and isolate the unique contribution of each layer to the representation of emotional images. Since valence and arousal often show correlations, we also performed partial correlation to examine the unique similarity between the neural network models and valence/arousal while controlling for the other dimension.

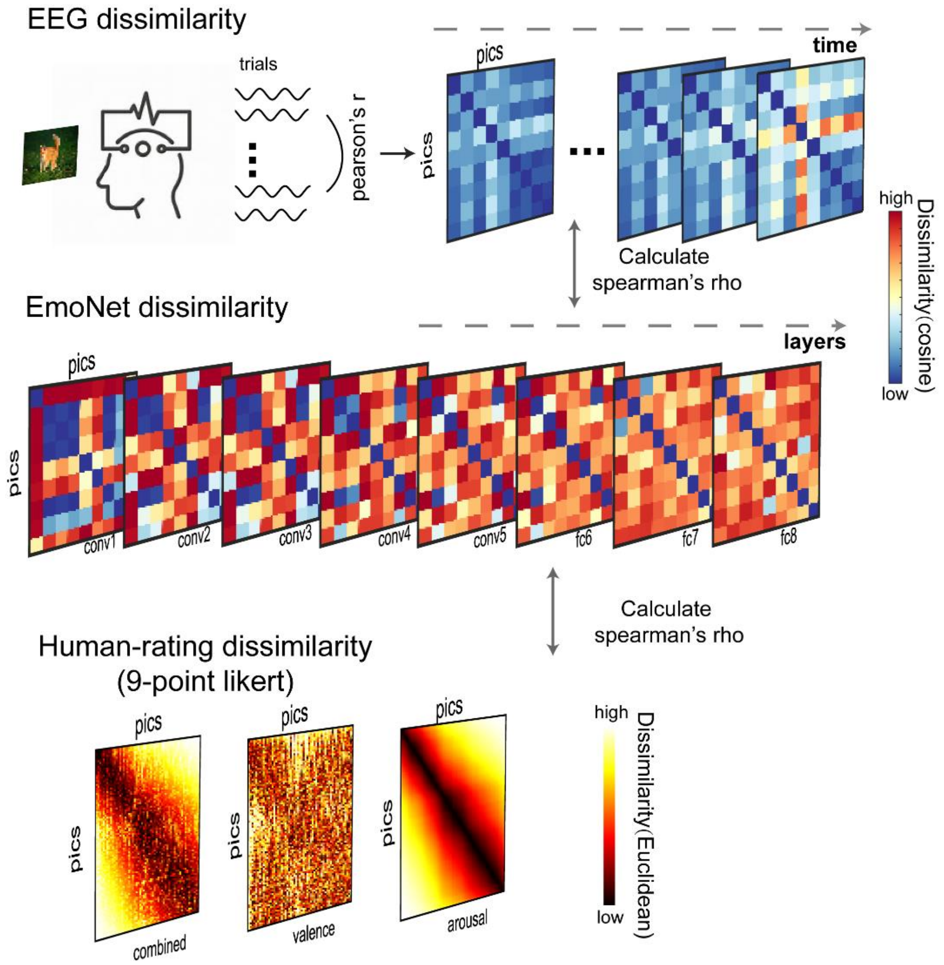

Figure 1. Representational similarity analysis pipeline. The representational dissimilarity matrix (RDM) for the DNN model (e.g., EmoNet) was computed using cosine dissimilarity between activations for each pair of images. The EEG RDM was constructed by calculating the Pearson correlation between EEG signals from all channels for each pair of images at each time point. The human-rating-based RDM was obtained by calculating the Euclidean distance between image ratings. Finally, Spearman's correlation was used to assess the representational similarity between the different RDMs, linking the representations from human brain data and models.

2.5. Statistical inference

For the RSA between the human-rating RDM and the EmoNet /AlexNet's representations, we performed a permutation test by randomly shuffling the matrices 1000 times to generate a null distribution of r-values. The observed correlation values were then compared to this null distribution to calculate the p-value. These p-values were corrected for multiple comparisons using the False Discovery Rate (FDR) at a threshold of 0.05.

For the RSA between the EEG RDMs and the EmoNet RDMs, non-parametric permutation tests were used to solve the multiple comparisons correction and statistical distribution assumptions [37]. These tests were performed using the function ft_timelockstatistic in Feildtrip [34]. Specifically, we used cluster-based inference to determine whether the Spearman r calculated from the RSA was significantly greater than zero (one-tailed test) at a given time point within the analysis window. For each time point, we computed the t-values and defined consecutively significant t-values as clusters. The size of the clusters was calculated by the sum of t-values. This procedure was repeated 1000 times, each time disrupting the label of the r value with the chance (0) and then calculating the t-value. This gives a permutation-based null distribution, and comparing this to the actual observed clusters gives a statistical p-value (0.05).

3. Results

3.1. EmoNet reflects the hierarchical processing structure of emotion perception in the human brain

To characterize emotion representations at the image-specific level, we used RSA to compare the neural representations in the human brain with those in EmoNet (Figure 2 and Methods). Figure 2 shows the representational similarity between EEG RDMs and EmoNet RDMs within the analysis window. The comparison of representational similarity values with zero for individual layers of EmoNet reveals significant correlations between neural representations during perceptual encoding and working memory delay and EmoNet representations in both early and deep layers (see Figure 2). Specifically, for layer 3, correlations are observed from 210 to 700 ms and 1100 to 1260 ms; for layer 4, from 260 to 550 ms; for layer 5, from 330 to 520 ms; for layer 6, from 1860 to 1980 ms; for layer 7, from 1880 to 1980 ms; and for layer 8, from 960 to 1060 ms and 1860 to 1960 ms. These results were validated through a cluster-based permutation test, yielding $p < .05$ in a one-tailed test. These results reveal a hierarchical processing pattern: early layers of EmoNet, which handle lower-level information, are significantly correlated with the early stages of emotion processing in the human brain, while deeper layers, responsible for generating and outputting emotional concepts, are significantly correlated with the later stages of emotion processing. Notably, the highest correlation coefficients are observed around 1 second after the image disappeared, which coincides with the time when participants were preparing to judge the emotional category (positive, negative, or neutral) of the image.

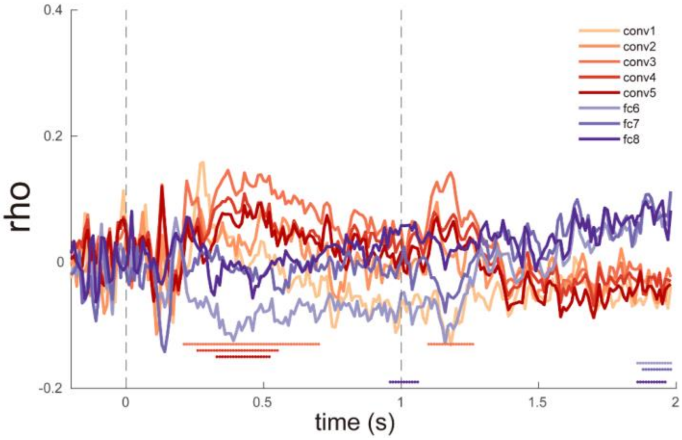

Figure 2. Temporal dynamic representational similarity between human brain EEG signals and EmoNet. Significant correlations are detected between neural signals and various layers within the DNN, including early, middle, and late layers. Representations seem to emerge earlier in the early convolutional layers and later in the deeper, fully connected layers (layer 3: 210 to 700 ms and 1100 to 1260 ms; layer 4: 260 to 550 ms; layer 5: 330 to 520 ms; layer 6: 1860 to 1980 ms; layer 7: 1880 to 1980 ms; layer 8: 960 to 1106 ms and 1860 to 1960 ms; cluster-based permutation test, p < .05, one-tailed test).

3.2. Representational similarity between EmoNet and human emotional ratings

Figure 3 shows the representational similarity between each layer of EmoNet and human ratings (valence, arousal, and combined dimensions) of emotional images. For both the valence (all *p*s < .015) and combined dimensions (all *p*s < .002), each layer of the model shows significant correlations with the geometric structure of human emotional ratings. However, for the arousal dimension, no significant correlations are observed (all *p*s > .1).

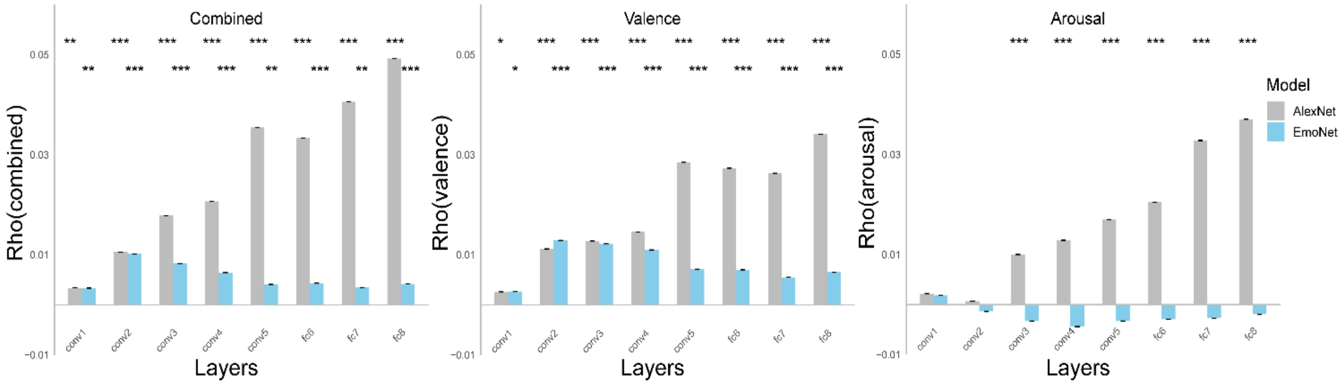

Figure 3. Representational similarity between each layer of the DNN and human ratings in valence, arousal, and combined dimensions. Error bars represent the SEM derived from permutation tests. The asterisks above the bars indicate significance: the first row corresponds to AlexNet, and the second row to EmoNet. $^{*}p < .050$; $^{**}p < .010$; $^{***}p < .001$.

When images were further categorized based on emotional valence (with positive images rated from 1–3, neutral images from 4.5–5.5, and negative images from 7–9), the results reveal (Figure 4, left) that for the combined dimension, EmoNet exhibits a significant similarity with human ratings for positive (fc7: $p = .032$) and negative images (conv1: $p < .001$). Regarding the valence dimension, the majority of EmoNet layers demonstrate a significant similarity with human ratings (conv1: $p < .001$; conv2: $p = .002$; conv3: $p = .008$; conv4: $p = .006$; conv5: p < .001; fc8: $p = .009$). For the arousal dimension, the fc6 ($p = .020$) and fc7 ($p < .001$) layers of EmoNet shows a significant correlation with human experience for positive images. These results indicate that EmoNet selectively aligns with human emotional experience for emotional images, but not for neutral images.

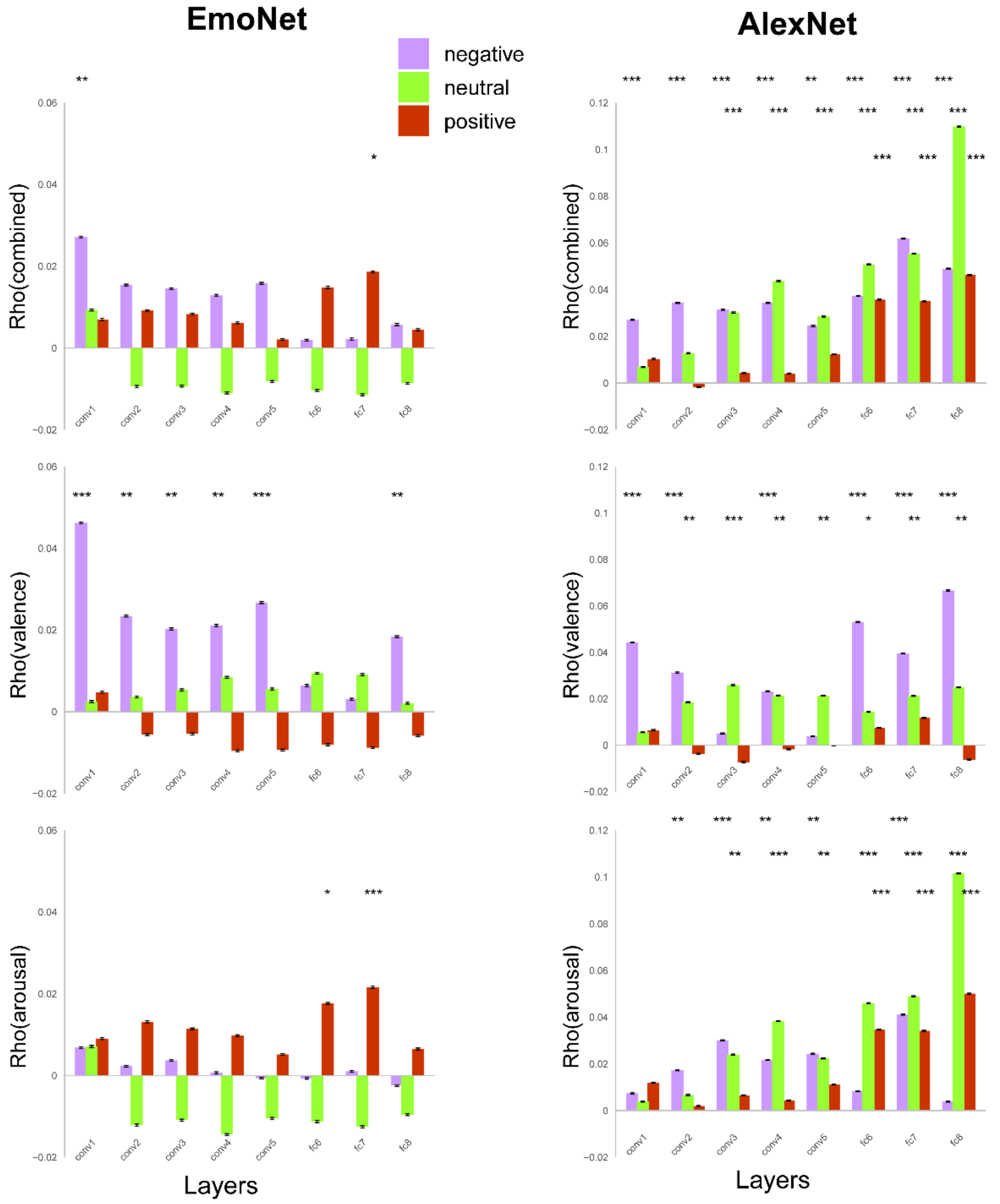

Figure 4. Similarity between each layer of the DNN and human ratings in valence, arousal, and combined dimensions across different emotion categories (negative, neutral, positive). Error bars represent the SEM derived from permutation tests. Asterisks indicate significance, with the rows representing the significance for negative, neutral, and positive image conditions, respectively. $^{*}p < .050$; $^{**}p < .010$; $^{***}p < .001$.

3.3. Absence of emotion-specific representational similarity in AlexNet

To test whether the representational similarity between EmoNet and human ratings on emotional images arises from its emotion-specific training, we conducted the same analysis on AlexNet. AlexNet shares the same model architecture as EmoNet, but it is trained for semantic classification (object recognition) rather than emotion classification. We extracted the activations of each layer of AlexNet's neurons for emotional images to construct the AlexNet RDM, and then correlated it with human emotional ratings.

The results reveal that at the general level, AlexNet exhibits significant representational similarity with human ratings across all dimensions—valence (all *p*s < .015), arousal (conv1: *p* = .075; conv2: *p* = .526; conv3-fc8: *p*s < .001), and the combined (all *p*s < .001) dimension. Furthermore, when images were categorized into positive, negative, and neutral categories, AlexNet's representational structure shows high similarity with human ratings for all images categories (positive images, fc6, fc7 and fc8: *p*s < 001; neutral images, conv3-fc8: *p*s < .003; negative images, conv2: *p* = .010; conv3: *p* < .001; conv4: *p* = .006; conv5: *p* = .005; fc7: *p* < .001). These findings highlight that AlexNet does not exhibit emotion-specific representational similarity, even though it performs similarly to human ratings in terms of overall representation.

4. Discussion

This study used representational similarity analysis (RSA) to explore the alignment between human brain emotion processing and deep neural networks, providing new insights into emotion processing. RSA assesses how similar or dissimilar these patterns are across different stimulus response. Two main findings emerged: EmoNet shares a hierarchical structure with the human brain, and it demonstrates emotion-specific representations that align with human responses to emotional, but not neutral, images.

First, by correlating EmoNet with EEG signals from the human brain during emotional processing, we observe a similarity in hierarchical pattern. This alignment suggests that both the model and the brain share a similar processing structure for emotions. Specifically, EmoNet's early layers correlate with early emotion processing stages in the brain, while later layers align with more advanced stages. This hierarchical pattern reflects both the temporal dynamics and the progression of emotion recognition from low-level visual features to high-level semantic concepts [14]. Notably, the correlation coefficient peaks approximately one second after the image disappears (see Figure 2), the moment when participants are preparing to assess the emotional valence (positive, negative, or neutral) of the image. Moreover, this finding reveals the dynamic changes in encoding and transient storage of different types of information (visual and emotion concepts) during emotional image processing in the human brain, as shown by the DCNN model, and underscores the critical role of the sensory cortex in emotional processing.

Second, EmoNet demonstrated a unique correlation with human ratings of emotional images across both arousal and valence dimensions, as well as their combined effects, but did not show such correlation for neutral images (Figure 4). In contrast, AlexNet, which was trained solely for object recognition, showed similar responses to human for all categories of images (positive, negative, and neutral). EmoNet's ability to capture emotional specificity likely stems from its refinement through exposure to emotional stimuli, enabling efficient emotion recognition. For the emotionally neutral categories, the model's classification accuracy also decreased, similar to the confusion that the human brain may experience when processing ambiguous emotions [15]. This specificity in EmoNet's representation contrasts with AlexNet's broader classification, suggesting that EmoNet discards non-emotional information, analogous to the brain's cellular pruning and network refinement processes [38,39].

These findings highlight the potential of emotion AI models like EmoNet to more accurately model human emotional processing, opening up new directions for improving emotion recognition systems. Our attempt to represent the similarity between human neural data and deep learning model can pave the way for the performance enhancement of emotional recognition models. Future work could involve using human ratings and neural representation similarities as training constraints to enhance model efficiency and specificity.

5. Conclusion

In this study, we bridge the gap between human brain and neural networks by comparing the emotional representations within these systems, with the goal of enhancing the understanding of emotion recognition in deep learning models. By correlating the emotion recognition model EmoNet and the object recognition model AlexNet with EEG signals from the human brain during the processing of emotional images, we uncover that EmoNet exhibits a hierarchical structure that mirrors the human brain's emotion processing stages and demonstrates emotion-specific representations that closely align with human responses to emotional images. This selective alignment underscores EmoNet's ability to prioritize emotionally salient information, a capability absent in general-purpose models like AlexNet. This study contributes to a deeper understanding of the neural mechanisms underlying emotion processing and lays a theoretical foundation for more human-like AI systems in affective computing.

Acknowledgments

We thank Xiaoqing Huang for helping collecting data. This work was supported by the National Natural Science Foundation of China (32171082, 32471136, 62425101, 62088102), the National Social Science Foundation of China (17ZDA323), the Science and Technology Planning Project of Guangdong Province (2023B1212060018), and the Leading talent program (31620016) at Sun Yat-sen University.

Conflicts of interests

The authors declare no conflict of interest.

Ethical statement

The study was performed in accordance with the Declaration of Helsinki and approved by the name of the Ethics Committee or Institutional Review Board (approval date: 2020-08-23, and approval number: 2020-0515-0140).

Authors' contribution

Conceptualization, L.N. and Y.K.; investigation, L.N. and Y.L.; supervision, Y.K.; formal analysis, L.N., K.C., Y.L. and Y.K.; writing — original draft, L.N., K.C., Y.T. and Y.K.; writing — review & editing, L.N., K.C., Y.T. and Y.K. All authors reviewed the results and approved the final version of the manuscript.

References

[1] Al-Shawaf L, Conroy-Beam D, Asao K, Buss DM. Human Emotions: An Evolutionary Psychological Perspective. *Emot. Rev.* 2015, 8(2):173–186.

[2] Barrett LF. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* 2017, 12(1):1–23.

[3] Clore GL, Huntsinger JR. How emotions inform judgment and regulate thought. *Trends Cognit. Sci.* 2007, 11(9):393–399.

[4] Van den Bergh O, Brosschot J, Critchley H, Thayer JF, Ottaviani C. Better Safe Than Sorry: A Common Signature of General Vulnerability for Psychopathology. *Perspect. Psychol. Sci.* 2021, 16(2):225–246.

[5] Li C, Wang J, Zhang Y, Zhu K, Wang X, *et al.* The Good, The Bad, and Why: Unveiling Emotions in Generative AI. *ArXiv* 2023, ArXiv:2312.11111.

[6] Lee W, Norman MD. Affective Computing as Complex Systems Science. *Procedia Comput. Sci.* 2016, 95:18–23.

[7] Shen H, Knearem T, Ghosh R, Alkiek K, Krishna K, *et al.* Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. *ArXiv* 2024, ArXiv:2406.09264.

[8] Liu P, Bo K, Ding M, Fang R. Emergence of Emotion Selectivity in Deep Neural Networks Trained to Recognize Visual Objects. *PLoS Comput. Biol.* 2024, 20(3):e1011943.

[9] Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Inf. Fusion* 2024, 102:102019.

[10] Aleman A, Kahn R. Strange feelings: Do amygdala abnormalities dysregulate the emotional brain in schizophrenia? *Prog. Neurobiol.* 2005, 77(5):283–298.

[11] Adolphs R. What does the amygdala contribute to social cognition? *Ann. N. Y. Acad. Sci.* 2010, 1191(1):42–61.

[12] Li W, Keil A. Sensing fear: fast and precise threat evaluation in human sensory cortex. *Trends Cognit. Sci.* 2023, 27(4):341–352.

[13] Li Z, Yan A, Guo K, Li W. Fear-Related Signals in the Primary Visual Cortex. *Curr. Biol.* 2019, 29(23):4078–4083.

[14] Satpute AB, Lindquist KA. The Default Mode Network's Role in Discrete Emotion. *Trends Cognit. Sci.* 2019, 23(10):851–864.

[15] Kragel PA, Reddan MC, LaBar KS, Wager TD. Emotion schemas are embedded in the human visual system. *Sci. Adv.* 2019, 5(7):eaaw4358.

[16] Bowers JS, Malhotra G, Dujmović M, Llera Montero M, Tsvetkov C, *et al.* Deep problems with neural network models of human vision. *Behav. Brain Sci.* 2023, 46:e385.

[17] Simony E, Grossman S, Malach R. Brain–machine convergent evolution: Why finding parallels between brain and artificial systems is informative. *Proc. Natl. Acad. Sci.* 2024, 121(41):e2319709121.

[18] Celeghin A, Borriero A, Orsenigo D, Diano M, Méndez Guerrero CA, *et al.* Convolutional neural networks for vision neuroscience: significance, developments, and outstanding issues. *Front. Comput. Neurosci.* 2023, 17:1153572.

[19] Lonnqvist B, Bornet A, Doerig A, Herzog MH. A comparative biology approach to DNN modeling of vision: A focus on differences, not similarities. *J. Vis.* 2021, 21(10):17–17.

[20] Yamins DLK, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 2016, 19(3):356–365.

[21] Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* 2014, 111(23):8619–8624.

[22] Heinen R, Bierbrauer A, Wolf OT, Axmacher N. Representational formats of human memory traces. *Brain Struct. Funct.* 2024, 229(3):513–529.

[23] Lu Z, Ku Y. Bridging the gap between EEG and DCNNs reveals a fatigue mechanism of facial repetition suppression. *iScience* 2023, 26(12):108501.

[24] Schyns PG, Snoek L, Daube C. Degrees of algorithmic equivalence between the brain and its DNN models. *Trends Cognit. Sci.* 2022, 26(12):1090–1102.

[25] Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2008, 2:249.

[26] Ogg M, Bose R, Scharf J, Ratto CR, Wolmetz M. Turing Representational Similarity Analysis (RSA): A Flexible Method for Measuring Alignment Between Human and Artificial Intelligence. *arXiv* 2024, arXiv:2412.00577.

[27] Mehrer J, Spoerer CJ, Jones EC, Kriegeskorte N, Kietzmann TC. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proc. Natl. Acad. Sci.* 2021, 118(8):e2011417118.

[28] Nie L, Ku Y. Decoding emotion from high-frequency steady state visual evoked potential (SSVEP). *J. Neurosci. Methods* 2023, 395:109919.

[29] Bae GY. The Time Course of Face Representations during Perception and Working Memory Maintenance. *Cereb. Cortex Commun.* 2021, 2(1):tgaa093.

[30] Lang PJ, Bradley MM, Cuthbert BN. International Affective Picture System (IAPS): Instruction manual and affective ratings. Technical Report A-8, Gainesville: The Center for Research in Psychophysiology, University of Florida. 2008.

[31] Brainard DH. The Psychophysics Toolbox. *Spat. Vis.* 1997, 10:433–436.

[32] Pernet CR, Sajda P, Rousselet GA. Single-Trial Analyses: Why Bother? *Front. Psychol.* 2011, 2:322.

[33] Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 2004, 134(1):9–21.

[34] Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 2011:156869.

[35] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 2017, 60(6):84–90.

[36] Lu Z, Ku Y. NeuroRA: A Python Toolbox of Representational Analysis From Multi-Modal Neural Data. *Front. Neuroinform.* 2020, 14:563669.

[37] Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 2007, 164(1):177–190.

[38] Chen L, Chen Z, Jiang L, Liu XL, Xu L, *et al.* AI of Brain and Cognitive Sciences: From the Perspective of First Principles. *arXiv* 2023, arXiv:2301.08382.

[39] Dong HM, Zhang XH, Labache L, Zhang S, Ooi LQR, *et al.* Ventral attention network connectivity is linked to cortical maturation and cognitive ability in childhood. *Nat. Neurosci.* 2024, 27(10):2009–2020.