Tracking representational formats of intrusive memories using deep neural networks

Rebekka Heinen (rebekka.heinen@rub.de)
Department of Neuropsychology, Institute of Cognitive Neuroscience, Ruhr University Bochum
Universitätsstraße 100, 44801 Bochum, Germany

Malte Kobelt (malte.kobelt@rub.de)
Department of Neuropsychology, Institute of Cognitive Neuroscience, Ruhr University Bochum
Universitätsstraße 100, 44801 Bochum, Germany

Nikolai Axmacher (nikolai.axmacher@rub.de)
Department of Neuropsychology, Institute of Cognitive Neuroscience, Ruhr University Bochum
Universitätsstraße 100, 44801 Bochum, Germany

Abstract

Intrusive memories of traumatic events differ in their content and their quality (i.e., representational format) from voluntary memories. We investigated these representational formats in 22 participants using a trauma film paradigm with a subsequent resting period to collect memory intrusions during functional magnetic resonance imaging (fMRI). We employed a convolutional neural network (DNN) re-trained to identify emotions, and large language model to quantify visual and semantic format. Using representational similarity analysis on DNN features we observed higher similarities between trauma than between neutral clips in both the visual and the semantic model, indicating generalization across content. However, on a neural level, encoding of trauma-analog clips revealed more pronounced visual formats. Our next steps will be to employ the semantic model and to analyze the resting period containing memory intrusions.

Keywords: representational format; deep neural networks; representational similarity analysis; large language models; fMRI; memory intrusions

Introduction

Memory traces consist of multiple representational formats which can be dynamically transformed from a visual format to a more abstract, semantic format (Xue, 2022). This transformation can be influenced by complex factors such as the valence of a stimulus which can improve later memory in case of negative stimuli (Kensinger, Garoff-Eaton, & Schacter, 2007). By contrast, extremely negative emotional content may also lead to detrimental effects on memory such as involuntary retrieval in the form of intrusions, with current theories suggesting a disrupted integration of visual content into long-term memory (Clark, Holmes, Woolrich, & Mackay, 2016; Brewin, 2014). Convolutional deep neural networks (cDNNs) and large language models (LLMs) have been shown to reflect visual and semantic formats, respectively, during neutral stimulus processing (Heinen, Bierbrauer, Wolf, & Axmacher, 2023), and recent studies demonstrated their applicability to complex emotional stimuli (Kragel, Reddan, LaBar, & Wager, 2019; Horikawa, Cowen, Keltner, & Kamitani, 2020). To shed light on the representational formats of traumatic memories we thus combined our data from a previous study (Kobelt et al., 2024) with deep neural network approaches. Here, we present preliminary results on the representational format of trauma-analog stimuli during encoding. Our next analysis steps are described in the outlook section.

Methods

The study was approved by the ethical committee of the Faculty of Psychology at Ruhr University Bochum, Germany. We tested 22 (all female) participants (age *M* = 24.5, *SD* = 3.9) using 3 Tesla fMRI (Philips Achieva, Philips Healthcare, Best, Netherlands; TR = 2.5 sec., 2 mm isotropic). We presented 21 trauma-analog and 21 matched control clips. In a subsequent resting period of 12 minutes participants indicated via button press when they experienced an intrusion and reported its content. DNN similarities were extracted from video frames using a cDNN pre-trained on ImageNet (emo-cDNN; implemented as in Kragel et al. (2019)) which we fine-tuned to identify 27 different emotions (Cowen & Keltner, 2017). Semantic similarities were extracted using a LLM (Cer et al., 2018) on the image labels provided by 7 independent raters. Using representational similarity analysis (RSA; Kriegeskorte, Mur, and Bandettini (2008)) we computed video-by-video similarity matrices for the network models and correlated them with neural similarity matrices in a whole-brain searchlight, separate for trauma and for control clips. All t-statistics are adjusted for multiple comparisons using Bonferroni, fMRI results are corrected using FDR.

Results

More negative emotions in trauma clips

We first tested the performance of the emo-cDNN on our video clips, revealing a significant difference in predictions for negative and positive emotions ($F_{(1,40)}$ = 4.67, p = 0.04; Figure 1B) as well as between video types ($F_{(1,40)}$ = 5.34, p = 0.03). Specifically, we find an interaction of video type and valence indicating more negative compared to positive emotions for trauma-analog (p = 0.02) but not for control clips ($F_{(1,40)}$ = 5.33, p = 0.03). Analyzing the similarity of trauma-analog clips to each other (trauma within - control within) emo-cDNN features revealed a significantly higher similarity among trauma clips. This effect was found across almost all layers of the network (conv1-5: all $p<0.001$, fc7: p = 0.004; Figure 1C), except for the first and last fully connected layer, suggesting more similar visual features among trauma clips than among control clips.

More pronounced visual format for trauma-analog clips

On a neural level, emo-cDNN similarities reveal a difference between trauma-analog and control clips: convolutional layers (conv1,conv5) indicate a more pronounced visual format for trauma-analog clips in areas such as the precuneus, the postcentral and the angular gyrus (Figure 1D) but no difference for fully-connected layers. In contrast, fully connected (fc6-8) but not convolutional layers show a better similarity fit for regions such as the medial and lateral occipital cortex, and the lingual gyrus suggesting an altered visual processing of trauma-analog content while control clips are processed along the normal hierarchy along the ventral visual pathway (VVS).

Generalization of semantic formats of trauma clips

We next investigated the semantic content of trauma-analog and control clips using word embeddings from descriptions of the videos from 7 independent raters. We find that the semantic format of trauma clips is highly similar (within vs between: $t_{(40)}$ = 18.59, $p<0.001$; Figure 1F) while at the same time it is distinct from control clips (trauma vs control: $t_{(40)}$ = 16.75, $p<0.001$). Although trauma clips share more common semantic content than controls, we find that semantic

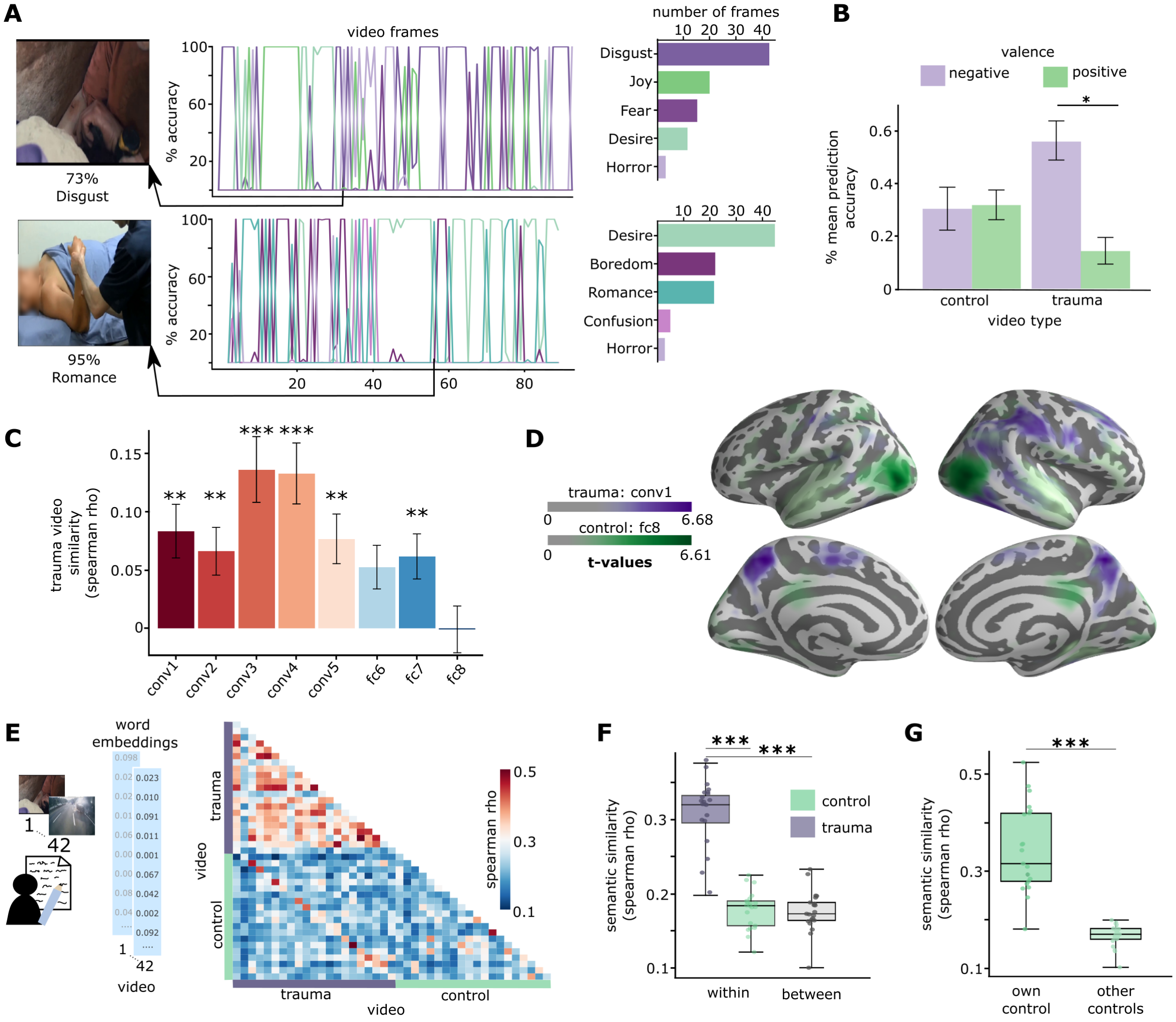

Figure 1: A) Prediction accuracy of the emo-cDNN on video frames of a trauma-analog (amputation) and a control (massage) video. Right bars: Top five predicted emotions (purple: negative, green: positive). B) More negative than positive emotion ratings in trauma-analog videos. No difference in control clips. C) Trauma-analog clips are more similar to each other (trauma-trauma similarity - control-control similarity) across almost all emo-cDNN layers. D) Better emo-cDNN to neural fit for trauma-analog clips in convolutional layers (conv1, purple) compared to a better fit for fully-connected layers for controls (fc8, green). E) We extracted semantic similarities from rater descriptions and computed a video by video similarity matrix. F) Higher semantic similarity within trauma videos compared to the similarity within control and between trauma and control. G) Higher similarity to the semantic content of the matched control clip compared to other control clips. Searchlight results are reported with $p < 0.05$, FDR-corrected for multiple comparisons

content between trauma clips and their matched controls significantly differs from the content of other controls ($t_{(40)} = 8.77$, $p < 0.001$; Figure 1G). These results suggest higher generalization of trauma clips based on shared semantic features, which is not found for control clips. Thus, the trauma-analog content used in this study seems to share both visual and semantic features.

Outlook

First, we will analyze the whole-brain searchlight using the semantic model to test whether semantic processing is altered for trauma-analog content, compared to the stronger involvement of visual format for these clips. In a next step, we will further investigate the link between semantic formats and neural data using topic modeling and Hidden Markov Models (Heusser, Fitzpatrick, & Manning, 2021; Lee & Chen, 2022; Perl et al., 2023) on the trauma clip descriptions but also the intrusion reports to understand the underlying representational geometry. Finally, we will test whether visual or semantic formats can be linked to the neural data during intrusions from the resting period.

Acknowledgments

R.H. received funding from the German Mercator Research Center Ruhr (MERCUR) - project no. Ex-2021-001. M.K. was funded by the German Academic Scholarship Foundation. N.A. received funding from the ERC project no. 864164 Grid Representations. We would like to thank all volunteers for their participation in this study.

References

Brewin, C. R. (2014). Episodic memory, perceptual memory, and their interaction: foundations for a theory of posttraumatic stress disorder. *Psychological bulletin*, *140*(1), 69. doi: 10.1037/a0033722

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., ... Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*. doi: 10.48550/arXiv.1803.11175

Clark, I. A., Holmes, E. A., Woolrich, M. W., & Mackay, C. E. (2016). Intrusive memories to traumatic footage: the neural basis of their encoding and involuntary recall. *Psychological Medicine*, *46*(3), 505–518. doi: 10.1017/S0033291715002007

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, *114*(38), E7900–E7909. doi: 10.1073/pnas.1702247114

Heinen, R., Bierbrauer, A., Wolf, O. T., & Axmacher, N. (2023). Representational formats of human memory traces. *Brain structure & function*. doi: 10.1007/s00429-023-02636-9

Heusser, A. C., Fitzpatrick, P. C., & Manning, J. R. (2021). Geometric models reveal behavioural and neural signatures of transforming experiences into memories. *Nature Human Behaviour*, *5*(7), 905–919. doi: 10.1038/s41562-021-01051-6

Horikawa, T., Cowen, A. S., Keltner, D., & Kamitani, Y. (2020). The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. *Iscience*, *23*(5), 101060. doi: 10.1016/j.isci.2020.101060

Kensinger, E. A., Garoff-Eaton, R. J., & Schacter, D. L. (2007). How negative emotion enhances the visual specificity of a memory. *Journal of Cognitive Neuroscience*, *19*(11), 1872–1887. doi: 10.1162/jocn.2007.19.11.1872

Kobelt, M., Waldhauser, G., Rupietta, A., Heinen, R., Rau, E., Kessler, H., & Axmacher, N. (2024). The memory trace of an intrusive trauma-analog episode. *Current Biology*. doi: 10.1016/j.cub.2024.03.005

Kragel, P. A., Reddan, M. C., LaBar, K. S., & Wager, T. D. (2019). Emotion schemas are embedded in the human visual system. *Science advances*, *5*(7), eaaw4358. doi: 10.1126/sciadv.aaw4358

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. doi: 10.3389/neuro.06.004.2008

Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, *13*(1), 4235. doi: 10.1038/s41467-022-31965-2

Perl, O., Duek, O., Kulkarni, K. R., Gordon, C., Krystal, J. H., Levy, I., ... Schiller, D. (2023). Neural patterns differentiate traumatic from sad autobiographical memories in ptsd. *Nature neuroscience*, *26*(12), 2226–2236. doi: 10.1038/s41593-023-01483-5

Xue, G. (2022). From remembering to reconstruction: The transformative neural representation of episodic memory. *Progress in Neurobiology*, *219*, 102351. doi: 10.1016/j.pneurobio.2022.102351