Exploring Brain Responses to Memorability-Controlled Generated Images

Hyewon Willow Han (hhan228@uwo.ca)
Neuroscience Graduate Program, Western University, London, ON N6A 3K7, Canada
Vector Institute for Artificial Intelligence, Toronto, ON M5G 0C6, Canada

Mansoure Jahanian (mjahani@uwo.ca)
Neuroscience Graduate Program, Western University, London, ON N6A 3K7, Canada

Johann Cardenas (jcarden4@uwo.ca)
Department of Computer Science, Western University, London, ON N6A 3K7, Canada

Yalda Mohsenzadeh (ymohsenz@uwo.ca)
Department of Computer Science, Western University, London, ON N6A 3K7, Canada
Vector Institute for Artificial Intelligence, Toronto, ON M5G 0C6, Canada

Abstract

Some images are more memorable than others, but the underlying neural mechanisms of memorability are not fully understood. This study investigates how the human brain processes images based on their memorability using fMRI responses to natural images by employing advanced deep neural network (DNN) and generative adversarial network (GAN) models. Our work successfully manipulated the memorability of images in both increasing and decreasing directions and discovered that brain regions associated with the face and body exhibited differential activity based on the memorability alterations, with specific activity patterns emerging in early visual areas. The amygdala responded to changes in both directions, whereas the hippocampus was primarily responsive when memorability decreased. Notably, place-associated areas showed reverse responses, being less active when images with increased memorability were presented. This investigation contributes to our understanding of the cognitive processes involved in visual memory, demonstrating the potential of integrating generative neural network models with neuroimaging to study brain function, and paving the way for formulation of experimentally testable hypotheses.

Keywords: memorability; generative adversarial networks (GANs); visual memory

Introduction

It is still not fully understood what features make an image more memorable, and the neural mechanisms shaping this behavioral phenomenon. Previous studies using human fMRI have demonstrated that variations in response magnitude within the high-level visual cortex correlate well with the memorability of faces and scene images (Bainbridge, Dilks, & Oliva, 2017; Bainbridge & Rissman, 2018; Lahner, Mohsenzadeh, Mullin, & Oliva, 2024). In this work, we aimed to explore how the brain reacts differently to variations in the memorability of images by using a diverse set of images beyond faces and scenes. Inspired by the work of (Gu et al., 2022), our research leverages artificial deep neural network (DNN) models as brain encoders and generative adversarial network (GAN) models for cognitive neuroscientific discovery, integrating these with human fMRI data to analyze responses in the face, body, place, and early visual areas, as well as memory-related regions such as the hippocampus and amygdala.

Materials and Methods

The Natural Scenes Dataset (NSD) contains high-resolution fMRI responses from 8 subjects to images of everyday scenes featuring common objects in their natural contexts (Allen et al., 2022). We trained the encoding model for each subject, using responses from 1,000 images that were shared across all subjects as the validation set and responses from the remaining images as the training set.

First, we manipulated the memorability of images in both increasing and decreasing directions (Fig 1A). As the NSD lacked memorability scores, we used MemNet for scoring. Inspired by (Younesi & Mohsenzadeh, 2022), which employs logistic regression to differentiate high- and low-memorable images, we utilized a GAN inversion approach for conversion to latent vectors. Specifically, StyleGAN-XL, pre-trained on ImageNet, was used to invert the images to their latent spaces (Sauer, Schwarz, & Geiger, 2022).

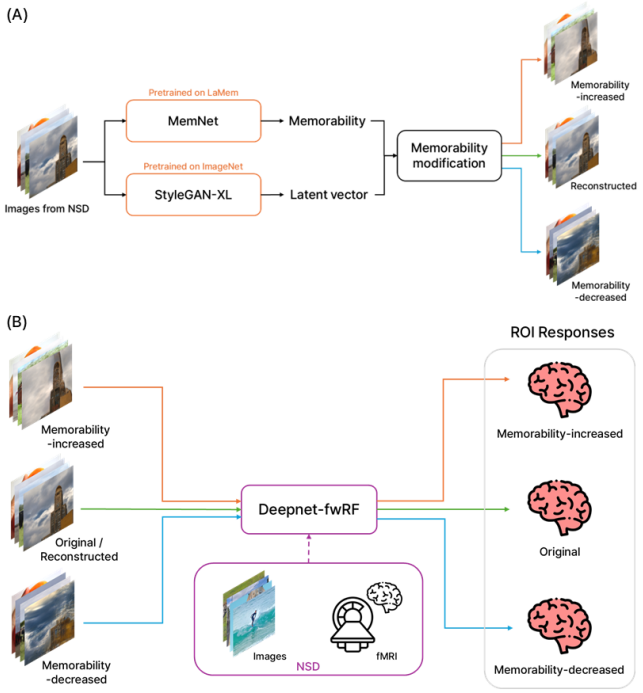

Figure 1: Schematic diagrams of the workflow. (A) Controlling memorability of real images with generative adversarial networks inspired by (Younesi & Mohsenzadeh, 2022). (B) Predicting brain responses of memorability-modified images using the encoding model.

Peak signal-to-noise ratio (PSNR) and learned perceptual image patch similarity (LPIPS) were used to screen overly disrupted images and exclude those altered beyond recognition after memorability control. We also excluded images not shown to all subjects, resulting in 152 images from the shared 1,000 in NSD used for the experiment. The feature-weighted receptive field (fwRF) model with AlexNet as the feature extractor was used as the brain encoder (Deepnet-fwRF), as described by (St-Yves & Naselaris, 2018). The workflow of predicting brain responses is depicted in Fig 1B.

Results

Controlling Memorability of Images

To validate that the memorability of the images was effectively manipulated, we fitted mixed linear regression models to reconstructed images (Goetschalckx, Andonian, Oliva, & Isola, 2019). The fitted model confirmed that for every twenty-unit increase, memorability increased by 0.012 for the reconstructed

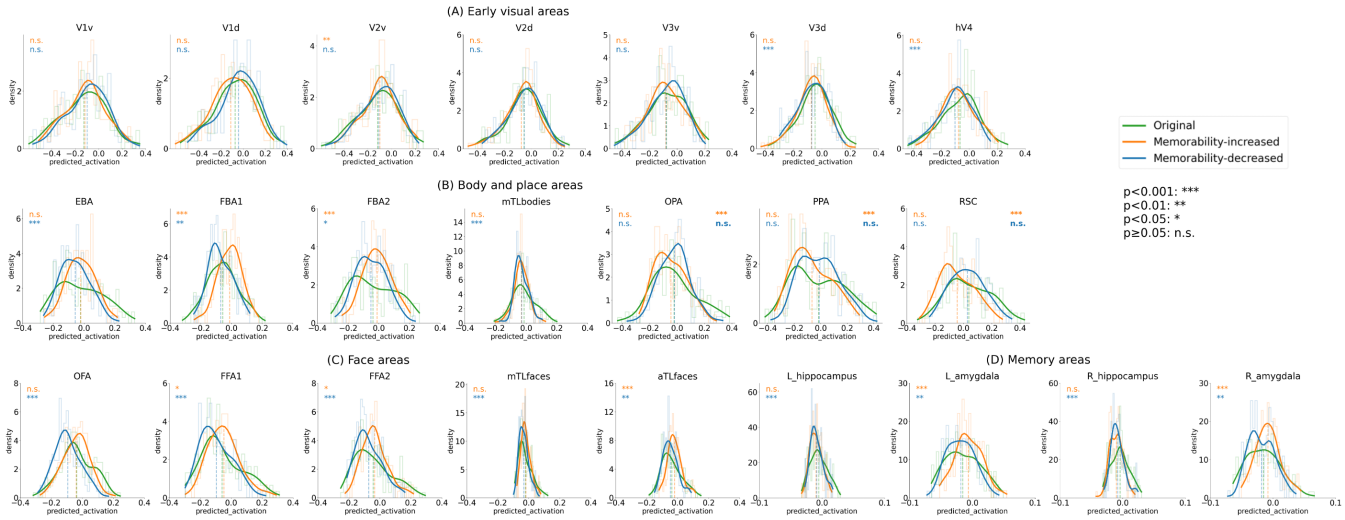

Figure 3: Histograms and kernel density estimate (KDE) plots of predicted ROI responses for memorability-modified images. Orange texts indicate paired t-test results between original and memorability-increased images and blue texts are for memorability-decreased images. Test results of the reverse hypothesis for place area are shown on the right side of each plot with bold text.

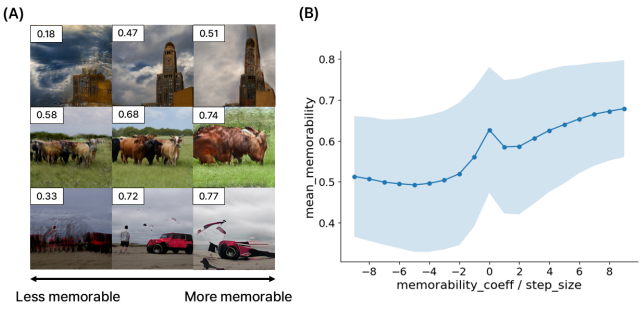

Figure 2: Changes in memorability as memorability coefficient increases. (A) Samples of memorability-decreased, reconstructed, and memorability-increased images, respectively. (B) Memorability changes in generated images.

images ($p < 0.001$). Changes in memorability as the coefficient increases are shown in Fig 2.

Brain Activity in Response to Changes in Memorability

To ensure that predicted activations for original and reconstructed images were similar, we calculated correlations between recorded and predicted activations for both image types across subjects. Subsequent two-sided paired t-tests of these correlations showed that 5 of 8 subjects had similar activations ($p > 0.05$), and we included these subjects (2, 3, 4, 7, and 8) in the analysis.

We predicted activations for memorability-increased, -decreased and original images across each region of interest (ROI), and conducted two paired t-tests to hypothesize: i) increased memorability leads to higher ROI activation compared to original images, and ii) decreased memorability leads to lower ROI activation. We checked 23 ROIs including early visual areas (V1, V2, V3, and V4), higher visual areas (occipital face area (OFA), fusiform face area (FFA), medial temporal lobe (mTL) face area, anterior temporal lobe (aTL) face area, extrastriate body area (EBA), fusiform body area (FBA), mTL body area, occipital place area (OPA), parahippocampal place area (PPA), retrosplenial cortex (RSC)), and components of medial temporal lobe (hippocampus and amygdala).

Predicted brain activity distributions for original images and their memorability-modified counterparts are illustrated in Fig 3. The results indicate statistically significant changes in both directions for FFA1, FFA2, aTL faces, FBA1, FBA2, and both amygdala ($p < 0.05$). Activity in V2v was statistically significant only with increased memorability, while OFA, mTL faces, EBA, mTL bodies, V3d, hV4, and left and right hippocampus were statistically significant when memorability decreased.

Conclusion

We explored how brain responses vary when the memorability of an image is either increased or decreased, excluding semantic influences. Notably, all face and body ROIs responded differently based on memorability changes. The V2 area activated more with increased memorability, while V3 and V4 were less active with decreased memorability. The amygdala responded in both directions of memorability adjustment, whereas the hippocampus only activated with decreased memorability. Significantly, none of the place areas showed any significant changes with our anticipated direction but exhibited the reverse, showing less active responses when images with increased memorability were given.

Acknowledgments

This study was generously supported by the Canada First Research Excellence Fund (CFREF) through a BrainsCAN grant to Y.M., a NSERC Discovery Grant to Y.M., and a Vector Institute Research Grant to Y.M. The authors also would like to acknowledge Vector Institute for providing computing resources.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage*, *149*, 141–152.

Bainbridge, W. A., & Rissman, J. (2018). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Scientific Reports*, *8*(1), 8679.

Goetschalckx, L., Andonian, A., Oliva, A., & Isola, P. (2019). Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5744–5753).

Gu, Z., Jamison, K. W., Khosla, M., Allen, E. J., Wu, Y., St-Yves, G., ... Kuceyeski, A. (2022). Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, *247*, 118812.

Lahner, B., Mohsenzadeh, Y., Mullin, C., & Oliva, A. (2024). Visual perception of highly memorable images is mediated by a distributed network of ventral visual regions that enable a late memorability response. *Plos Biology*, *22*(4), e3002564.

Sauer, A., Schwarz, K., & Geiger, A. (2022). Stylegan-xl: Scaling stylegan to large diverse datasets. In *Acm siggraph 2022 conference proceedings* (pp. 1–10).

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202.

Younesi, M., & Mohsenzadeh, Y. (2022). Controlling memorability of face images. *arXiv preprint arXiv:2202.11896*.