Modeling Visual Memorability Assessment: A Computational Approach Using Autoencoders

Elham Bagheri (Elham.bagheri@vectorinstitute.ai)
Vector Institute for Artificial Intelligence, Toronto, ON M5G 0C6, Canada
Department of Computer Science, Western University, London, ON N6A 3K7, Canada

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

Abstract

Certain images stick in our mind while others vanish quickly. This study explores the computational aspects of image memorability by employing a pretrained autoencoder, specifically a VGG-based model trained on the ImageNet dataset. The research investigates the relationship between the memorability of images, quantified as the likelihood of their remembrance after a single exposure, and their reconstruction errors and distinctiveness in the latent space of an autoencoder, finetuned on the MemCat dataset comprising 10,000 images in diverse categories. The predictive power of latent representations in determining memorability is also evaluated. The findings suggest that images with unique features that challenge the autoencoder's capacity are inherently more memorable. This correlation indicates a new pathway for evaluating image memorability, potentially impacting industries reliant on visual content and fostering advancements in the fields of artificial intelligence and cognitive science. It also demonstrates how machine learning models can emulate human cognitive processes to assess memorability, leading to improvements in algorithmic performance.

Keywords: memorability; autoencoders; reconstruction error; memorable images; MemCat dataset; VGG16 deep neural network; latent code representation

Introduction

Theories on memory processing suggest that deeper levels of cognitive processing lead to stronger memories (Craik & Lockhart, 1972). Image memorability, an intrinsic characteristic of the image, is the likelihood of an image being remembered upon a single exposure. The exploration of image memorability aids in understanding how different image attributes influence memorability, thus bridging the gap between visual perception and memory retention (Bainbridge, 2019). Advancements in artificial intelligence (AI) are deepening our understanding of human memory. Machine learning algorithms, which mimic the brain's pattern recognition (Gregor & LeCun, 2010), can identify features that enhance image memorability (Isola, Xiao, Parikh, Torralba, & Oliva, 2013; Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017). This study aims to further the research by using neural networks to mimic human brain's function in processing images. For this purpose, we employ a vgg16-based autoencoder, trained on the ImageNet to explore how image reconstruction correlates with memorability. The VGG16 architecture, developed by Simonyan & Zisserman Simonyan and Zisserman (2014), is a well-known deep learning model for image processing. ImageNet (Deng et al., 2009) is a diverse dataset of over 14 million images in over 20,000 classes. It enables the development of models with a rich base of features, capable of recognizing numerous visual patterns and objects.

Analysis and Results

Modeling Memorability Assessment Experiment

To model memorability assessment experiment using neural networks, we fine-tuned a VGG16-based pretrained autoencoder on a single epoch with a batch size of one, on the MemCat dataset, intending to replicate the human memorability experiment conditions where humans are exposed only once to each image. MemCat data includes 10,000 images from five broad categories: animals, sports, food, landscapes, and vehicles (Goetschalckx & Wagemans, 2019). Figure 1 (A) demonstrates the Schematic representation of the autoencoder's architecture we adapted for this experiment, along with representative images from the MemCat dataset and their corresponding reconstructions. After finetuning, the autoencoder was tested on the same images, and the reconstruction error for each image was calculated as a measure of the finetuned autoencoder's ability to replicate the input to which it was exposed once. We finetuned models with various loss functions, including perceptual and structural losses, to best capture the nuances of image fidelity and perceptual similarity. Varying values of learning rate were also considered. The best model was selected based on the reconstruction error. The following results were obtained after finetuning and testing on a subset of 1,000 images which is close to the number of images shown to human participants in a single session (Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015).

Spearman's rank correlation was used for quantifying correlation in this study. A significant correlation of 0.46 was observed between the reconstruction error and memorability, suggesting images that are more challenging for the autoencoder to reconstruct tend to be more memorable. Figure 1 (B) shows the overall distribution of reconstruction errors relative to the scores, with a fit regression line showing the trend. Such correlation also holds across image categories

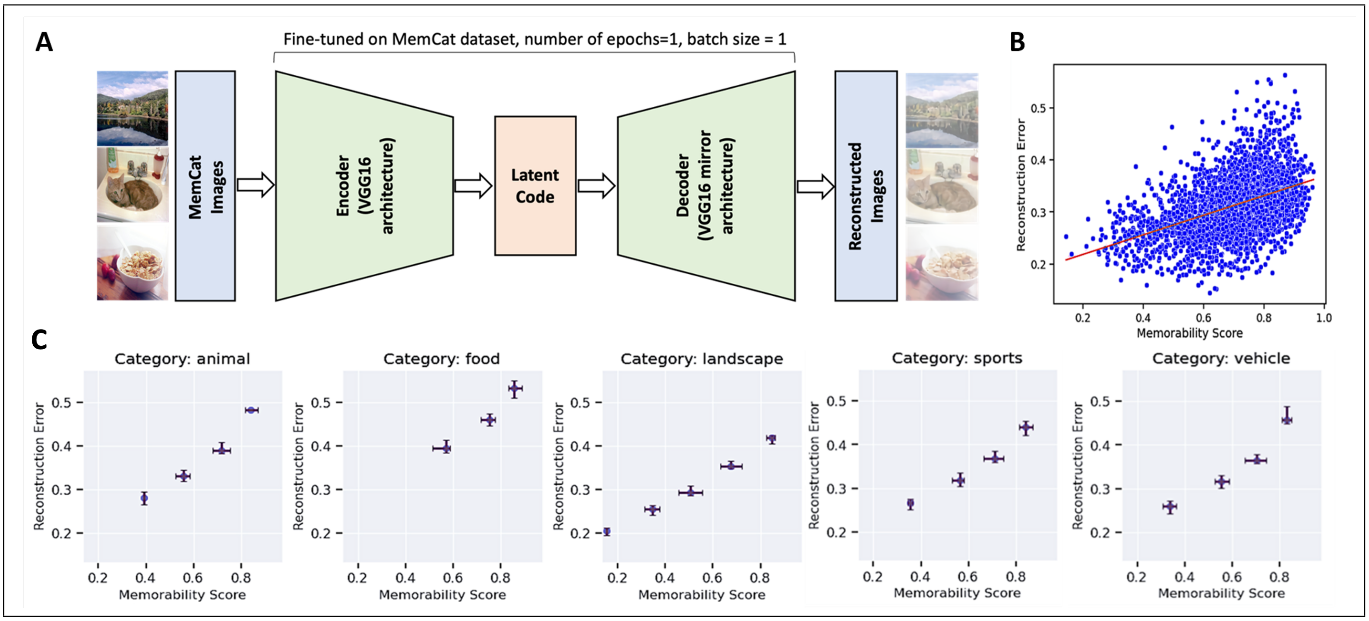

Figure 1: A) Schematic representation of the autoencoder model's architecture. B) Overall distribution of the reconstruction errors relative to memorability scores. C) Reconstruction errors plotted against memorability scores for various image categories within the MemCat dataset.

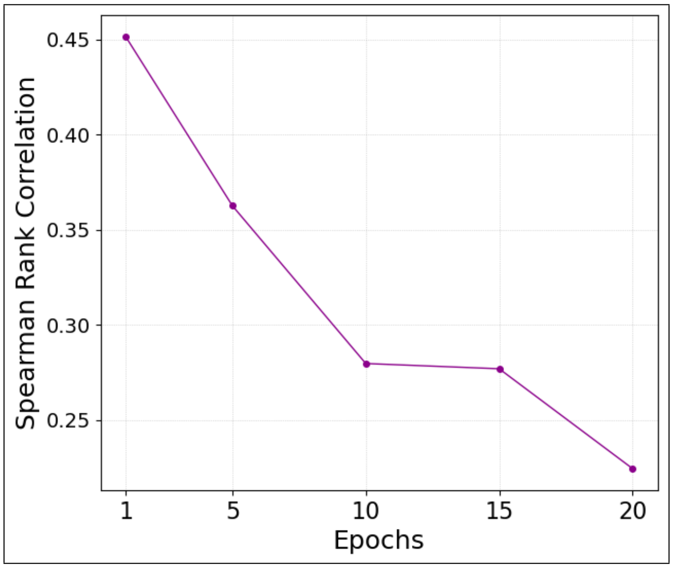

Figure 2: Correlation between the memorability scores and reconstruction error for varying number of epochs.

as seen in Figure 1 (C), where the scores and correlations are binned for visual clarity. Furthermore, inspired by (Lin, Li, Lafferty, & Yildirim, 2023), distinctiveness was quantified via the Euclidean distance between each image's latent code and its nearest neighbor. Distinctiveness was shown to have a positive correlation of 0.42 with memorability, suggesting that more unique and complex images are more memorable.

Analyzing the Effect of Epoch Size

By continuing to finetune the model for more epochs, the correlation between reconstruction error and memorability scores was reduced. This predicts that exposing individuals to images for more times may decrease the memorability effect. The trend is demonstrated in Figure 2 where the finetuning was carried on for 20 epochs.

Prediction of Memorability Scores by Latent Code

We trained a Gated Recurrent Unit (GRU) (Cho et al., 2014), an efficient type of Recurrent Neural Networks (RNN), to predict the memorability scores and to classify images into high and low memorable. The model development was solely using the finetuned autoencoder's latent code. The MemCat dataset was split into train, validation, and test sets by a ratio of 70:15:15. Hyperparameter optimization was performed by different learning rates and sequence lengths. The best-performing model demonstrated a robust predictive performance on the test set, with a receiver operating characteristic (ROC) curve area under the curve (AUC) of 0.72 and a Precision-Recall (PR) AUC of 0.65. Image classification into high or low memorable was based on a threshold set at the median value of the original scores. The model achieved an accuracy of 0.68, precision of 0.64, sensitivity of 0.64, and specificity of 0.71. Additionally, the model's predictions correlated significantly with the original memorability scores, as evidenced by Spearman's rank correlation of 0.47. This approach demonstrates that the encoder effectively encodes crucial attributes integral to memorability, and latent code contains attributes explaining memorability.

Conclusions

This study reveals that image memorability seems to be linked to the uniqueness of images, as evidenced by the positive correlation between an image's reconstruction error and its memorability score. Furthermore, it highlights the potential of using neural network models to simulate human-like memorability assessment and opens avenues for future research to explore image memorability using deep learning, with implications for cognitive neuroscience and fields reliant on visual content retention.

References

Bainbridge, W. A. (2019). Chapter one - memorability: How what we see influences what we remember. In K. D. Federmeier & D. M. Beck (Eds.), *Knowledge and vision* (Vol. 70, p. 1-27). Academic Press. doi: https://doi.org/10.1016/bs.plm.2019.02.001

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision research*, *116*, 165–178.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, *11*(6), 671–684.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Goetschalckx, L., & Wagemans, J. (2019). Memcat: a new category-based image set quantified on memorability. *PeerJ*, *7*, e8169.

Gregor, K., & LeCun, Y. (2010). Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning* (pp. 399–406).

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2013). What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, *36*(7), 1469–1482.

Lin, Q., Li, Z., Lafferty, J., & Yildirim, I. (2023). From seeing to remembering: Images with harder-to-reconstruct representations leave stronger memory traces. *arXiv preprint arXiv:2302.10392*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1452–1464.