Image Memorability on Vision Transformers (ViT) and Class Activation Mapped (CAM) Residual Networks (Res-Net50)

Yagiz Devre (yagiz.devre@princeton.edu)
Department of Computer Science, Princeton University
Princeton, NJ 08540 USA

Sevan Harootonian (skh@princeton.edu)
Departments of Psychology and Computer Science, Princeton University
Princeton, NJ 08540 USA

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science, Princeton University
Princeton, NJ 08540 USA

Abstract

What makes certain images more memorable than others? In this paper, we are utilizing advanced computational models, including Vision Transformers and CAM ResNets to answer this question. We propose a novel approach to predict and visualize what makes an image memorable. Furthermore, our proposed model utilizing Vision Transformers achieve a 0.0066 MSE value and a 0.72 Spearman correlation coefficient, marking a significant advancement in this field.

Keywords: Memorability, LaMem Dataset, Visual Stimuli, Machine Learning Vision Transformers, Class Action Maps, Residual Networks, Cognitive Psychology

Introduction

Memory has always been one of the most prominent areas of research in the field of cognitive psychology. Especially, remembrance of visual inputs, and the memorability of images that observed has been one of the fascinating and popular topics to explore.

At first glance, the idea of memory seams to rely strongly on previous experiences and encounters that are observed. However multiple behavioral studies have effectively showed a more complementary perspective on the memorability of visual stimuli. Isola et al. discusses the level of subjectivity of visual stimuli. They argue that memorability is an intrinsic and stable feature of a visual stimuli that is shared across different viewers and their background. (Isola, Xiao, Parikh, Torralba, & Oliva, 2014) Therefore the understanding of memorability of an image is highly objective and standardized in a large population. These findings additionally suggest that memorability of certain images are very distinctive and occasionally produces a common ground. Some images are more memorable than others, they "stich" to our memory while others are indeed forgotten (Isola, Xiao, Torralba, & Oliva, 2011).

Furthermore, Lionel Standing, in the article "Learning 10,000 Pictures" argues the fact of standardized visual memory and how the human memory is capable of holding abstract information such as letters, words, and numbers along with more concrete stimuli such as objects scenes and sounds (Standing, 1973). Standing discusses the idea that humans are able to memorize and remember a certain image that they observed for a short period even after thousands of unrelated randomized images (Isola, Xiao, Torralba, & Oliva, 2011). These results presented show that the capacity of memorization follows a degrading "power law model" dependent on the number of images observed which shows superiority over verbal stimuli (Standing, 1973).

Since the memorability of a visual input is highly dependent to the intrinsic features, as mentioned, rather than personal experiences, the investigation and analysis of the aspects that makes an image memorable have become increasingly applicable in the field of cognitive psychology.

One key development and implication of this pivot towards understanding image memorability is the analysis of low-level, and high-level features and their effect over memorability of images (Jaegle, et al., 2019) (Hagen & Espeseth, 2023). According to the recent behavioral studies, a weak-correlation between the memorability of an image and the low-level aspects of the visual stimuli such as contrast, and brightness has been found (Dubey, Peterson, Khosla, Yang, & Ghanem, 2015). Yet, these low-level features are not sufficient to accurately effect and predict the memorability of an image stimuli alone. Therefore, other features such as the category of the object presented in the input image and visual salience tend to be more dominant on the memorability of an image. Images that contain these attributes, therefore, effectively represents a *high utility of information* (Bylinskii, Goetschalckx, Newman, & Oliva, 2022) suggesting the idea that memorability of an image is also correlated to how "different", "surprising", "dangerous" or "untrustworthy" it is. This idea of *high utility of information* depends mostly on the context of the image shown. Furthermore, *utility of information* of a visual input is found to be highly universal, supporting the idea that what makes an image memorable is based more on the intrinsic aspects of the stimuli rather than observer's personal experience. In their research, Blentski et al. argues there is a universal effect, a universal trend, that creates a common *high utility of information* on emotional/affective stimuli, "unexpected actions, social aspects, animated objects (human faces, gestures, interactions, etc)" (Bylinskii, Goetschalckx,

Newman, & Oliva, 2022). Likewise, in their research Blentski et al. concluded that there is a psychological reasoning behind such a universal common phenomenon, supporting the idea that remembering what is "different" and "dangerous" is more crucial in evolution, since remembering these types of visual stimuli provides a preparation and an adjustment to the world view of the observers, which is encoded in the dynamics of human evolution and survival.

Given that the problem of memorability of an image is highly dependent on the context represented, it requires capturing of the higher-level aspects such as composition and semantics that lead a direct gateway to the utilization of computational models. Since the high-level features, along with minor effects of aesthetics(low-level features) changes how memorable an image really is, deep learning model techniques and Convolutional Neural Networks (CNNs) has provided new methodologies and models to predict image memorability. (Squalli-Houssaini, Duong, Gwenaëlle, & Demarty, 2018)

Furthermore, the exploration of image memorability through advanced computational models like and Class Activation Mapped(CAM) Residual Networks(ResNets) and Vision Transformers(ViTs) represents a convergence of human cognition and artificial intelligence. In this paper, a comprehensive analysis of the factors influencing image memorability is analyzed through advanced computational models and is presented. Initially, we will discuss an overview of existing research in the field of image memorability, specifically Large Scale Image Memorability (LaMem) (Khosla, Raju, Torralba, & Oliva, 2015) dataset containing 60,000 images with memorability scores, representing the benchmarking methods and previous results obtained through different algorithmic structures. Following this, we will present a new model that integrates a transfer learning integrated structure of Vision Transformer developed by Google Research in the article "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (Dosovitskiy, et al., 2021) and Class Activation Mapped Residual Network (CAM ResNet) to analyze, visualize and predict image memorability. This model will serve as a cornerstone for the paper, enabling the investigation of various features that contribute to an image's memorability.

Literature Review

Previous research in this area of cognitive psychology of utilizing machine learning algorithms to identify how memorable an image is, and what makes an image more memorable has often focused on image recognition, object detection, and various other applications for understanding the image content.

Since their introduction with AlexNet in 2012, CNNs and their adapted models showed a high popularity in the area of image memorability, specifically used for Large Scale Image Memorability(LaMem) Dataset (Khosla, Raju, Torralba, & Oliva, 2015). The dataset represents a significant milestone in the study of image memorability. Developed by researchers at the Massachusetts Institute of Technology, the LaMem dataset is specifically designed to standardize and act as a benchmark of what makes certain images more memorable than others. The dataset contains of 60,000 non categorized images, each annotated with memorability scores derived from human subject evaluations through Amazon Mechanical Turk(AMT). These scores offer an objective measure of how likely an image is to be remembered.

LaMem Dataset Benchmarks

Before presenting and analyzing the features of the model proposed in this paper, a clear analysis of the pervious works is needed to be mentioned. LaMem dataset had been utilized in multiple models varying from CNN architectures to Residual Networks. Specifically, there were several pivotal papers that shaped and set the benchmark for this dataset.

MemNet (2015)

MemNet marked a breakthrough advancement in the field of image memorability prediction using Convolutional Neural Networks (CNNs). Introduced in the same article as LaMem (Khosla, Raju, Torralba, & Oliva, 2015), MemNet pioneered the application of CNNs, specifically utilizing the already implemented architecture of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) trained on 1.3 million high-resolution images in the ImageNet training set with 1000 different classes to establish an initial benchmark in this area.

Squalli-Houssaini et al. (2018)

In their research titled "Deep learning for predicting image memorability" Sequalli Houssaini et al. utilized the CNN-based approach a step further by integrating image captioning (IC). They have generated an IC model that builds an encoder with a CNN and LSTM (Long-Short Term Memory) Recurrent Network (Squalli-Houssaini, Duong, Gwenaëlle, & Demarty, 2018) for learning a image-text embedding. The integration of textual descriptions generated by IC with the visual analysis offered by CNNs allowed them for a more diverse feature understanding of the images, leading to enhanced and accurate memorability predictions.

Leonardi et al. (2019)

In their research titled "Image Memorability Using Diverse Visual Features and Soft Attention", Leonardi et al. utilized a dual Convolutional Neural Network architecture pre-trained for object recognition and memorability estimation both incorporating a soft attention mechanism (Leonardi, Celona, Napoletano, & Bianco, 2019). More specifically, they have used a pretrained the Residual Network (Res-Net50) developed by Microsoft Research (He, Zhang, Ren, & Sun, 2016) as their CNN model and an LSTM model for their memorability score regression.

Zhu et al. (2020)

In the paper "Aesthetics-Assisted Multi-task Learning with Attention for Image Memorability Prediction", Zhu et al. utilizes a learning network trained on a combination of datasets: the LaMem dataset and the AADB dataset, a dataset compiled for 10,000 images and their respective aesthetic scores (Brady , Konkle, Alvarez, & Oliva, 2008), for predicting image memorability and aesthetic scores. As for

the model, Pixel-wise Contextual Attention network (PiCANet) was utilized to generate an attention map at each pixel and construct a feature to capture the memorability and aesthetic information regarding to the image (Zhu, Zhu, Zhu, & Li, 2020).

ResMem-Net (2021)

In the article "ResMem-Net: memory based deep CNN for image memorability estimation", Praven et al. presents a ResNet-50 based architecture that incorporates a combination of Convolutional Neural Network and LSTM layers (Praveen, et al., 2021). Their model utilizes information from the hidden layers since hidden layers in their hybrid model contains the intrinsic features of the images. The proposed architecture learns visual emotions and saliency, using GradRAM technique for heatmaps generation.

VitMem(2023)

Finally, in the paper "Image Memorability Prediction with Vision Transformers" Hagen et al. discusses an idea of shift by applying Vision Transformers (ViT) to memorability prediction. It uses ViTs for capturing global image contexts, aiming to provide more accurate and human-like memorability predictions (Hagen & Espeseth, 2023). VitMem is by far the most similar architecture to the model that is proposed in this paper since both architectures suggest a potential improvement over previous CNN-based models. Furthermore, both models utilize the unique capabilities of Vision Transformers, due to the unique capability of transformers called "linear embeddability of images" (Hagen & Espeseth, 2023) just like NLP transformers.

To conclude, the evolution of models predicting image memorability has been marked by significant advancements over the years. These reports have all been benchmarked via two significant attributes: Mean Squared Error Loss and Spearman's Rank Correlation Coefficient (ρ).

Mean Squared Error Loss

The MSE Loss measures the average of the squares of the errors, i.e., the average squared difference between the estimated memorability predicted by the models and the actual value. It's given by the following equation where Y_i is the ground truth and $\hat{Y}_i$ is the predicted value by the model.

$$MSE\ Loss = \frac{1}{N}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Spearman's Rank Correlation Coefficient (ρ)

Spearman's rank correlation coefficient is a measure of rank correlation. Fundamentally, this coefficient assesses how well the relationship between two variables(actual and predicted score) can be described using a monotonic function which is described in the equation where *d* represents the difference between the predicted memorability and the ground truth memorability. (Mukaka, 2012)

$$\rho = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

To conclude the resulting benchmarks for these baseline models can be summarized as in *Table 1* which shows an increasing correlation and a decreasing throughout the years.

Table 1: Summary of Previous Benchmarks.

Model Name	Year	MSE	Spearman ρ
MemNet	2015	N/A	0.640
Squalli-Houssaini et al.	2018	0.0079	0.720
Leonardi et al.	2019	0.0092	0.687
Zhu et al.	2020	N/A	0.670
ResMem-Net	2021	0.011	0.679
ViTMem	2023	0.0076	0.711

Data

The data used in this research paper, as mentioned earlier, is the Large-Scale Image Memorability (LaMem) dataset, a comprehensive image dataset curated by researchers at Massachusetts Institute of Technology(MIT) that contains a diverse array of images, each annotated with memorability scores (Khosla, Raju, Torralba, & Oliva, 2015). This section will detail the composition, characteristics, and analysis of the dataset, including its overview, variety, and the methodology employed in scoring the images. By examining the LaMem dataset's structure, this section will lay the foundation for understanding the dataset's role in the development of the proposed model.

Overview of Data

The LaMem dataset, is effectively one of the largest image dataset on the field of memorability, containing 60,000 images sourced from diverse backgrounds, and is designed to provide an objective measure of human memory in relation to visual stimuli. The dataset is organized, with images and the respective memory scores, split into 5 sections splits.

Within each split, a training, validation, and test set is included (for instance, train_1.txt, val_1.txt, and test_1.txt) with a separation of 45k, 4k, and 10k images for training validation and testing respectively.

Every image in each split has a unique floating-point value ranging from 0 to 1, quantifying the image's memorability as shown in Figure 1. The memorability score is achieved through a study called *efficient visual memory game* (Khosla, Raju, Torralba, & Oliva, 2015). Every individual completes the study though Amazon Mechanical Turk. The process involves varying the time interval between the first and repeated showings of images. The method also check for different time interval lengths, allowing for efficient collection of memorability scores for a large number of images.

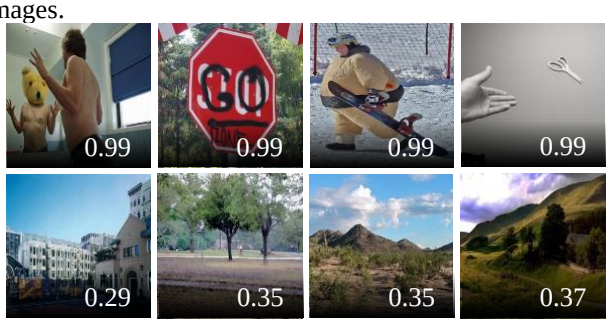

Figure 1: The LaMem dataset, four high(above) and low(bellow) memorability scores and their mapped images.

Pre-processing

The dataset contains 5 splits, which occasionally contains duplicate versions of the images in other splits. This means that an {Image, Memory Score} pair is present in both the validation of split 5, training of split 2 and testing of split 3. Additionally, it's important to note that a specific image may have varying memorability scores across different splits. This variation is due to the methodology employed in scoring.

Due to these variations in the dataset, it is needed to perform a preprocessing to effectively separate the dataset into 3 sections without the need of 5 splits via taking the average of every image's memorability score represented in any of the 5 splits using the Algorithm 1.

Algorithm 1 Data Pre-processing with Averaging.

initialize empty dictionary of lists *image_scores*
for *split_id* ∈{1, 2, 3, 4, 5} **do**
  for *split_type* ∈{*training*, *validation*, *testing*} **do**
    open file *'splits/{split_type}_{split_id}.txt'*
    for *line* ∈ *file* **do**
      image_name, *score* = *line*
      add *score* to array with the *image_name* key
initialize empty dictionary *average_scores*
for *image_score_pair* ∈ *image_scores* **do**
  image_name, *score_list* = *image_score_pair*
  sum_of_scores = **sum** *score_list*
  length_of_scores = **length** *score_list*
  average_score = *sum_of_scores / length_of_scores*
  add *average_score* with the *image_name* key

With the use of this idea represented as a pseudo code in Algorithm 1, a new data table was constructed with all of the 60,000 images which will be randomly shuffled and separated into %10 Testing, %10 Validation and %80 Training in the model preparation for the analysis of the bias and variance of the model via testing it with unseen inputs. Further analysis over the averaged data highlights the following properties about the dataset shown in *Table 2*, which was not presented in the original paper in 2015 due to the splitting method.

Table 2: Features of the averaged-out memorability scores

Mean(μ)	Median	Mode	Standard Deviation(σ)
0.7557	0.7693	1.0	0.1155

Likewise, the minimum and maximum average memorability value was found as 0.2657 and 1.0 with four different images therefore setting the mode to 1.0. As shown in Table 2, the mean value of memorability is lower than the median, indicating a negatively skewed graph distributed evenly, as shown in Figure 2 that also visualizes the Cumulative Distribution Function(CDF) of the dataset.

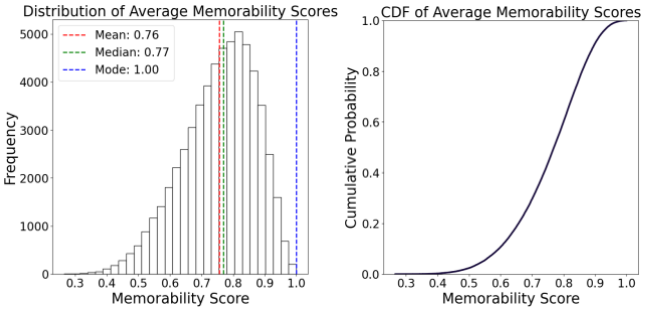

Figure 2: The visual representation of averaged data

Data Classification

As mentioned earlier in this paper, LaMem dataset does not provide any categorical information regarding to the images themselves since the dataset contains only the images and their respective memorability score. Therefore, as the final step of the data visualization and understanding, a ResNet-50 algorithm pretrained on the ImageNet data was used to classify and categorize each image. Since the residual connections enable the model to bypass one or more layers, it prevents overfitting and preserves a good accuracy. Therefore, a ResNet-50 model was selected for this type of task because of its high accuracy and efficient categorization of diverse images based on learned features from ImageNet.

After classification, every image that is mapped to the same category was averaged based on the memorability score. This progress generates valuable information on how diverse the dataset is, and which categories represent more memorable images. Finally, a visualization of the dataset in a categorical format was processed, identifying which categories are more memorable to the observers than others as shown in Figure 3.

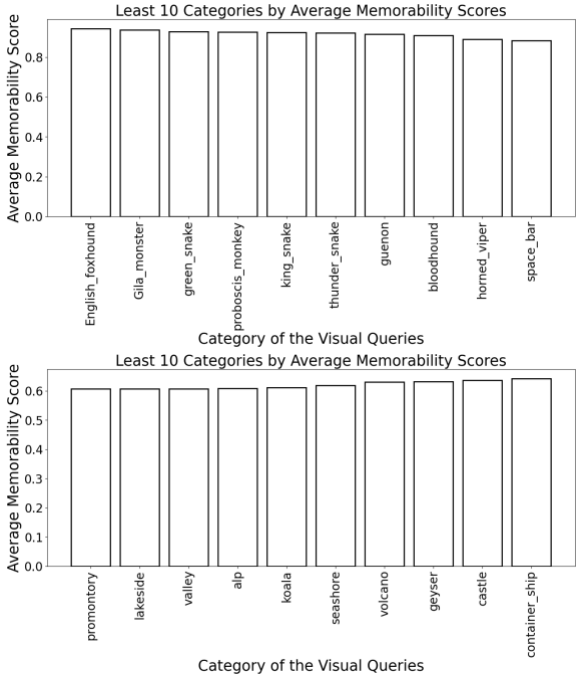

Figure 3: The ResNet-50 categorized LaMem dataset with top memorable and least memorable categories of inquiries.

As seen from the Figure 3, the top and the least memorable categories represent a similarity between each other, since the top memorable categories are more "different" and "dangerous" than the less memorable images such as "valley", "promontory", and "lakeside" which all represent a geographical landscape.

These results perfectly align with the previously made statement on by Bylinskii et al. remembering what is "different" and "dangerous" is more crucial in evolution, since remembering these types of visual stimuli provides a preparation for the future encounters (ie. Encounter with an English foxhound, or a snake). Finally, since ResNet-50 managed to identify 993 out of 1,000 categories of ImageNet, this showed that the dataset is diverse with categories that reduces bias towards one specific category.

Methodology

As described in the Literature Review section, deep neural networks, particularly CNNs, have been instrumental in mirroring aspects of human perception, showing a high potential for the use as scientific model regarding visual stimuli and its remembrance. Their effectiveness in mimicking human cognitive strategies (Cichy & Kaiser, 2019) and in deriving psychological representations of images (Peterson, Abbott, & Griffiths, 2018) further highlights their potential in identifying features that makes an image memorable.

However, the high-dimensionality and redundancy in CNN models in visual queries such as images raise several challenges. These challenges lead to inquiries about the actual relevance of extensive features that CNNs generate which might not be useful and redundant in memorability regression. Therefore, a method thar reduces the dimensionality of CNN representations to better align with human psychological representations is needed. This suggests a more compact model with smaller dimensionality in psychological feature representations than the full network of CNN features.

Model

As a solution to the redundant dimensionality in the CNN models, this paper proposes the use of Vision Transformer (ViT) models, specifically vit-base-patch16-224. The reasoning behind the use of ViT model by Google is because unlike traditional CNNs, a pure transformer can be applied directly to sequences of image patches through an encoded transformer (Dosovitskiy, et al., 2021). ViTs utilize a transformer embedding, a type of self-attention mechanism, enabling the model to focus on different parts of an image and understand contextual relationships more effectively just like how transformers work for language generation and identification. More formally, a ViT model breaks a 2D image into a 1D sequence of token embeddings when it traverses over batches. These 1D sequence of token embeddings are then used for a Transformer through a technique called linear projection to obtain the embedding z_0.

These embedding are then used within a transformer, that consists of layers of multiheaded self-attention(MSA) and Multi Layered Perceptron(MLP) blocks, with a Layernorm(LN) layer. The MSA blocks allows the model to focus on different parts of the image at the same time whereas MLP blocks act like a network of perceptron, creating a neural network that process the information obtained from the MSA blocks. Finally, LN ensures stability throughout learning process in neural networks via normalization of the inputs across the features for each layer. (Dosovitskiy, et al., 2021)This structure can be modeled mathematically as:

$$z_0 = [x_{class};\ x_p^1\mathrm{E};\ x_p^2\mathrm{E};\ x_p^3\mathrm{E}\ldots x_p^N\mathrm{E}]$$
$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1}$$
$$z_\ell = MLP(LN(z'_\ell)) + z'_\ell$$
$$y = LN(z_\ell)$$

(Dosovitskiy, et al., 2021)

where ℓ is the batch number in the sequence of batches, E is the linear projection matrix and E_{pos}represents the positional embeddings of the batch over the liner projection matrix. Finally, y represents the predicted image score, x_{class} represents the "learnable embeddings" and x_p^k are all the input embeddings of the 2D image turned into 1D which are linearly projected using the matrix E. (Dosovitskiy, et al., 2021)

As seen from the mathematical formal model, at each step the encoded queries are normalized through LN. Furthermore, the model first focusses on features of the image and separates them before making a decision using MLP perceptron network. This method of "focusing different features just like words in a sentence" of a visual query mirrors aspects of human visual perception, allowing us to capture both local and global image features. In this research, the pre-trained ViT model, vit-base-patch16-224, is utilized to analyze the averaged out LaMem dataset that is split into %10 validation, %10 Testing and %80 training just like the previous benchmarking models. Additionally, the proposed model uses and Adam optimizer with a learning rate of 0.001 for 50 epochs along with a learning rate scheduler and early stopping feature to prevent overfitting as well as Google Tesla A100 GPU cores with high RAM options for 200 Credit hours to reduce the training and testing time.

Results and Analysis

In this section, the results and the analysis of the trained model will be performed. After training the model with the specified configurations, a model with 0.0066 MSE Loss with a Spearman's coefficient of 0.7184 is achieved. These results show that the proposed modified ViT model is the most accurate and the most correlated model that was ever trained on LaMem dataset, also shown in Table 3.

More interestingly, the model did not need the entirety of 50 epochs mentioned in the model statement. Due to the fact that learning rate scheduler implemented managed to

Table 3: Proposed Model Benchmarks.

Model Name	Year	MSE	Spearman ρ
MemNet	2015	N/A	0.640
Squalli-Houssaini et al.	2018	0.0079	0.720
Leonardi et al.	2019	0.0092	0.687
Zhu et al.	2020	N/A	0.670
ResMem-Net	2021	0.011	0.679
ViTMem	2023	0.0076	0.711
Proposed Model	**2023**	**0.0066**	**0.7184**

reduce overfitting, both training and validation losses were already minimized, and early stopping was called after the loss was significantly low after 5th epoch as seen in Figure 4. This shows how powerful the pre-trained ViTs are since just after 5 epochs, a state-of-the-art model was produced. Finally, as seen from Figure 4, there is an optimal balance of bias and variance in this model since both the validation and training losses decrease at each epoch, showcasing the reliable nature of the proposed model.

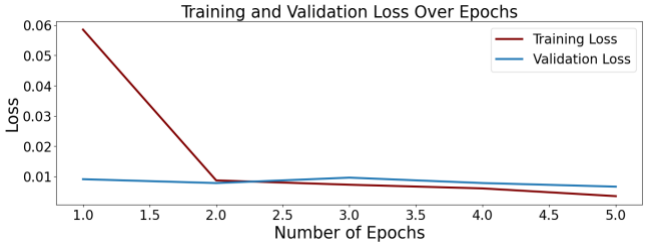

Figure 4: Training-validation loss of the proposed model

As the following step of the evaluation of the model, Figure 5 was generated, which includes two randomly selected images from the test set and their actual and predicted memorability scores respectively, showcasing the accuracy of the model in random selection.

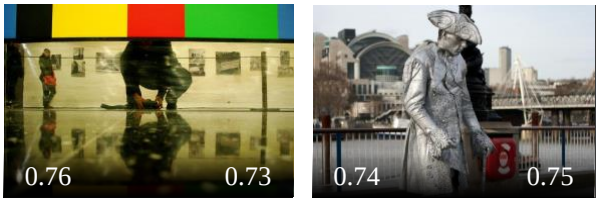

Figure 5: Randomly selected images and their scores. Left corner is the estimated score right corner is actual the score.

As the final step of the evaluation of the model, a ResNet-50 based model was also trained utilizing this model to generate a GradCam heatmap of the image. A key note is that, this model is used just to generate an activation map, not memory scores and is for purely visual purposes. The reason for this decision was to detect which features do the AI models pick, specifically amongst the Top 10 categories described in Figure 3. A separate model was needed to be embedded for this visualization task since ViT models(unlike ResNet and CNN models) don't have a functional Class Activated Map(CAM) generation since they operate more like a transformer than a gradient descent algorithm with convolutions. Based on this, Figure 6 was generated.

One key detail that Figure 6 suggests is that the ViT model was able to identify the memorability score with a 0.03

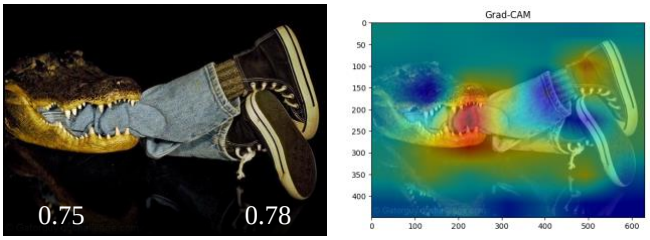

Figure 6: Predicted (left corner) and actual(right corner) memorability of a crocodile category with GradCam Heatmap

difference, furthermore, as the Grad-Cam results show that AI models are able to extract the key features, such as the crocodile in Figure 6 and base their judgements on these features, such as the teeth of the crocodile, highlighted red.

General Discussion and Conclusion

The research presented in this paper provides a comprehensive analysis on image memorability, a subject at the intersection of cognitive psychology and computer science. The findings build on the foundational work of Isola et al., Standing, and others that suggests that image memorability is defined at the intrinsic features of visual stimuli. With the use of advanced computational models, particularly Vision Transformers (ViT) and Class Activation Mapped Residual Networks (CAM ResNet), this research represents a novel approach in predicting and visualizing image memorability. This methodology not only aligns with the human cognitive process of prioritizing images that are "different" and "dangerous" as highlighted by Blentski (Bylinskii, Goetschalckx, Newman, & Oliva, 2022), but also allows for a deeper understanding of which intrinsic qualities make an image memorable.

The results from the LaMem dataset analysis, particularly the categorization using ResNet-50, highlight the concept of *high utility information* in images. This classification between memorability and generation of categories in LaMem dataset is novel to this paper.

With the proposed novel and state-of-the-art prediction model, that has 0.0066 MSE and 0.72 Spearman coefficient, future research can delve into the implementation of GAN networks for producing more memorable images using the proposed model as a discriminator. A key limitation of the research was the dependency to a single sense: vision. A possible future work on the field can interpret how other domains like context, sound, and smell plays a role in enhancing these features and feature maps found using AI models. Likewise, the integration of neuroscientific insights into these models could lead to more accurate and nuanced predictions, allowing a very close interpretation of visual memory. In conclusion, this research contributes significantly to the understanding of image memorability, providing both theoretical insights and practical applications. As the research in this field continues to explore the intricate relationship between human cognition and computational models, we as humans will uncover even deeper truths and facts about how we perceive, remember, and interact with the visual world that surrounds us.

Acknowledgements

This work has been supported by the faculty and resources of the COS360: Computational Models of Cognition class at Princeton University. The guidance and insights provided by the faculty, Prof. Thomas L. Griffiths and Dr. Sevan Harootonian have been invaluable in the development of this research.

References

Jaegle, A., Mehrpour , V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *Elife*, 8:e47596.

Praveen, A., Noorwali, A., Samiayya, D., Khan, M. Z., Vincent, D. R., Bashir, A. K., . . . Tariq, M. (2021). ResMem-Net: memory based deep CNN for image memorability estimation. *PeerJ Comput Sci*, 5;7:e767.

Rust, N., & Jannuzi, B. (2022). Identifying Objects and Remembering Images: Insights From Deep Neural Networks. *Current Directions in Psychological Science*, 31(4), 316-323.

Rust, N., & Mehrpour, V. (2020). Understanding Image Memorability. *Trends in Cognitive Sciences*, Vol. 24, No. 7.

Hagen, T., & Espeseth, T. (2023). Image Memorability Prediction with Vision Transformers. *arXiv*, arxiv.org/pdf/2301.08647.pdf.

Jeong, S. K. (2023). Cross-cultural consistency of image memorability. *Sci Rep*, 13, 12737.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 145-152.

Khosla, A., Raju, A., Torralba, A., & Oliva, A. (2015). Understanding and Predicting Image Memorability at a Large Scale. *International Conference on Computer Vision (ICCV)*, DOI 10.1109/ICCV.2015.275.

Needell, C., & Bainbridge , W. (2022). Embracing New Techniques in Deep Learning for Estimating Image Memorability. *Comput Brain Behav*, 168–184.

Squalli-Houssaini, H., Duong, N., Gwenaëlle, M., & Demarty, C.-H. (2018). Deep learning for predicting image memorability. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2371–2375.

Leonardi, M., Celona, L., Napoletano, P., & Bianco, S. (2019). Image Memorability Using Diverse Visual Features and Soft Attention. *Image Analysis and Processing – ICIAP*, 171-180.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, 25:2, 207-222.

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1469-1482.

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What makes an object memorable? *IEEE International Conference on Computer Vision (ICCV)*, 1089-1097.

Bylinskii, Z., Goetschalckx, L., Newman, A., & Oliva, A. (2022). Memorability: An image-computable measure of information utility. *Human Perception of Visual Information. Springer*, 207–239.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*, 2010.11929.

Cichy, R., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends Cogn Sci*, 305-317.

Peterson, J., Abbott, J., & Griffiths, T. (2018). Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations. *Cogn Sci*, 42(8):2648-2669.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *arXiv*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*.

Brady , T., Konkle, T., Alvarez, G., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105 (38) 14325-14329.

Zhu, T., Zhu, F., Zhu, H., & Li, L. (2020). Aesthetics-Assisted Multi-task Learning with Attention for Image Memorability Prediction. *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China,*, 360-363.

Mukaka, M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J.*, 24(3):69-71. PMID: 23638278; PMCID: PMC3576830.