MIT Libraries | DSpace@MIT

MIT Open Access Articles

Understanding and Predicting Image Memorability at a Large Scale

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Khosla, Aditya, et al. "Understanding and Predicting Image Memorability at a Large Scale." 2015 IEEE International Conference on Computer Vision (ICCV), 7-13 December 2015, Santiago, Chile, IEEE, 2015, pp. 2390–98.

As Published: http://dx.doi.org/10.1109/ICCV.2015.275

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: http://hdl.handle.net/1721.1/112993

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike

Massachusetts Institute of Technology

Understanding and Predicting Image Memorability at a Large Scale

Aditya Khosla
MIT
khosla@mit.edu

Akhil S. Raju
MIT
araju@mit.edu

Antonio Torralba
MIT
torralba@mit.edu

Aude Oliva
MIT
oliva@mit.edu

Abstract

Progress in estimating visual memorability has been limited by the small scale and lack of variety of benchmark data. Here, we introduce a novel experimental procedure to objectively measure human memory, allowing us to build LaMem, the largest annotated image memorability dataset to date (containing 60,000 images from diverse sources). Using Convolutional Neural Networks (CNNs), we show that fine-tuned deep features outperform all other features by a large margin, reaching a rank correlation of 0.64, near human consistency (0.68). Analysis of the responses of the high-level CNN layers shows which objects and regions are positively, and negatively, correlated with memorability, allowing us to create memorability maps for each image and provide a concrete method to perform image memorability manipulation. This work demonstrates that one can now robustly estimate the memorability of images from many different classes, positioning memorability and deep memorability features as prime candidates to estimate the utility of information for cognitive systems. Our model and data are available at: http://memorability.csail.mit.edu

1. Introduction

One hallmark of human cognition is our massive capacity for remembering lots of different images [2, 20], many in great detail, and after only a single view. Interestingly, we also tend to remember and forget the same pictures and faces as each other [1, 13]. This suggests that despite different personal experiences, people naturally encode and discard the same types of information. For example, pictures with people, salient actions and events, or central objects are more memorable to all of us than natural landscapes. Images that are consistently forgotten seem to lack distinctiveness and a fine-grained representation in human memory [2, 20]. These results suggest that memorable and forgettable images have different intrinsic visual features, making some information easier to remember than others. Indeed, computer vision works [12, 18, 15, 7] have been able to reliably estimate the memorability ranks of novel pictures, or faces, accounting for half of the variance in human consistency. However, to date, experiments and models for predicting visual memorability have been limited to very small datasets and specific image domains.

Intuitively, the question of an artificial system successfully predicting human visual memory seems out of reach. Unlike visual classification, images that are memorable, or forgettable, do not even look alike: an elephant, a kitchen, an abstract painting, a face and a billboard can all share the same level of memorability, but no visual recognition algorithms would cluster these images together. What are the common visual features of memorable, or forgettable, images? How far we can we go in predicting with high accuracy which images people will remember, or not?

In this work, we demonstrate that a deep network trained to represent the diversity of human visual experience can reach astonishing performance in predicting visual memorability, at a near-human level, and for a large variety of images. Combining the versatility of many benchmarks and a novel experimental method for efficiently collecting human memory scores (about one-tenth the cost of [13]), we introduce the LaMem dataset, containing 60,000 images with memorability scores from human observers (about 27 times larger than the previous dataset [13]).

By fine-tuning Hybrid-CNN [37], a convolutional neural network (CNN) [23, 21] trained to classify more than a thousand categories of objects and scenes, we show that our model, MemNet, achieves a rank correlation of 0.64 on novel images, reaching near human consistency rank correlation (0.68) for memorability. By visualizing the learned representation of the layers of MemNet, we discover the emergent representations, or diagnostic objects, that explain what makes an image memorable or forgettable. We then apply MemNet to overlapping image regions to produce a memorability map. We propose a simple technique based on non-photorealistic rendering to evaluate these memorability maps. We find a causal effect of this manipulation on human memory performance, demonstrating that our deep memorability network has been able to isolate the correct components of visual memorability.

Altogether, this work stands as the first near-human per-

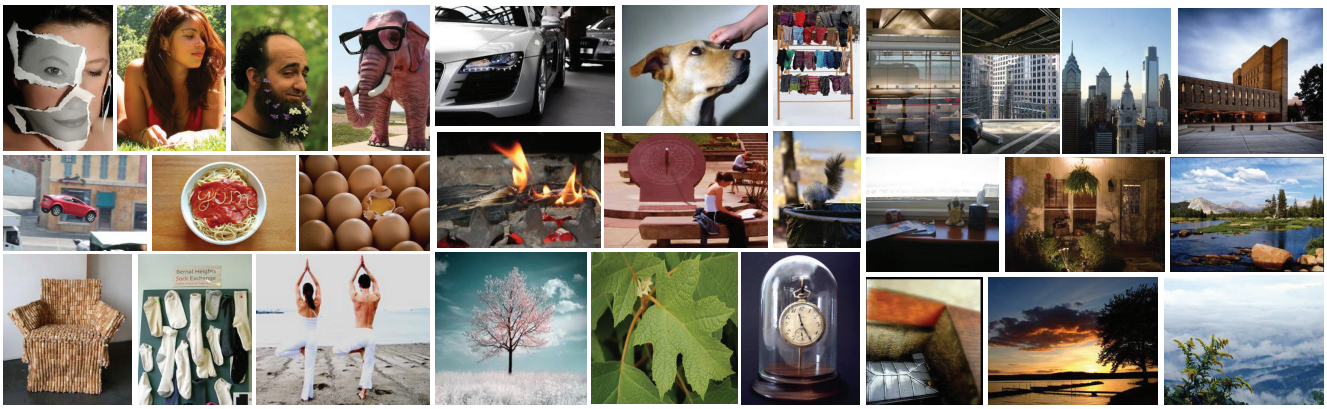


Figure 1. Sample images from LaMem arranged by their memorability score (decreasing from left to right). LaMem contains a very large variety of images ranging from object-centric to scene-centric images, and objects from unconventional viewpoints.

formance benchmark of human visual memory, offering an understanding and a concrete algorithm for predicting the memorability of an image and its regions. We envision that many applications can be developed out of deep memorability features, akin to the recent astonishing impact that deep networks have had on object and scene classification tasks. Our work shows that predicting human cognitive abilities is within reach for the field of computer vision.

2. LaMem: Large-scale Memorability Dataset

Here, we introduce an optimized protocol of the memory game introduced by [13] to collect human memory scores. In this game, images are presented successively, and some are repeated. Observers must press a key when they recognize an image seen before. This allows us to collect ground truth scores on how memorable images are. The basic idea of our novel procedure is to allow the second occurrence of an image to occur at variable time intervals. This procedure is based on the finding that the memorability ranks of images are time-independent [13]. We propose an algorithm to account for this varied time interval allowing us to obtain high consistency with the existing benchmark [13]. Furthermore, using this new experimental setting, we build a novel massive memorability dataset, with scores on $60,000$ images (~ 27 times the previous largest benchmark), while keeping a low cost. Our dataset contains significantly more variety in the types of images (see Fig. 1), while still maintaining a high human consistency on memorability.

First, in Sec. 2.1, we briefly describe the sources of images used for building the dataset to demonstrate its variety as compared to existing datasets. Then, in Sec. 2.2, we describe the efficient visual memory game for obtaining large-scale memorability annotations. Last, in Sec. 2.3, we provide experimental validation of the proposed method.

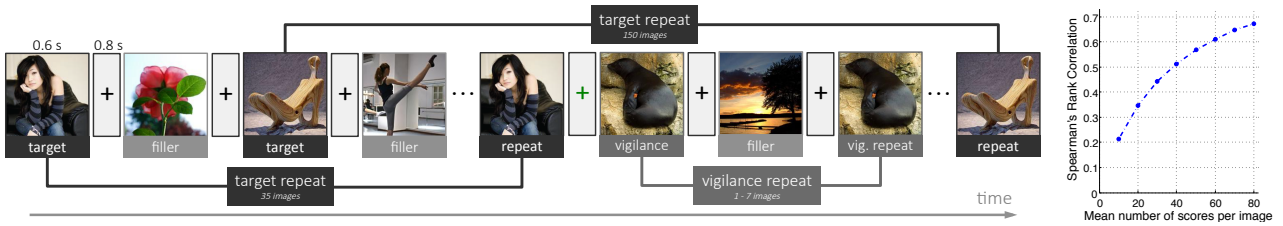
2.1. Collecting images

To create a varied dataset, we sampled images from a number of existing datasets such as MIR Flickr [11], AVA dataset [27], affective images dataset [25] (consisting of Art and Abstract datasets), image saliency datasets (MIT1003 [14] and NUSEF [28]), SUN [34], image popularity dataset [16], Abnormal Objects dataset [31] and aPascal dataset [9]. Thus, our dataset contains scene-centric images, object-centric images and other types such as images of art, images evoking certain emotions, and other user-generated images such as 'selfies'. We explore the correlation between a variety of these attributes and memorability in Sec. 3.2.

2.2. Efficient Visual Memory Game

Our experimental procedure consists of showing target repeats (the second occurrence of an image) at variable time intervals. For example, some targets may be repeated after just 30 images, while others are repeated after 100. As shown in [13], memorability scores change predictably as a function of the time interval between repeats, while memorability ranks are largely conserved i.e., if the time between the showing of a target and its repeat is increased, the memorability scores of all images decrease by a similar amount, thereby preserving the rank ordering. In our method, we use this information to propose a method based on coordinate descent that explicitly accounts for the difference in interval lengths. This allowed us to collect ground truth memorability scores for a large number of images (here $60,000$), in a short amount of time, and at a very reasonable cost.

Model: We first describe one possible interpretation of the memorability score computation proposed by [13], and extend that to our setting. Let us define $m^{(i)}$ as the memorability of image i. For image i, we have some $n^{(i)}$ observations given by $x_j^{(i)} \in \{0, 1\}$ and $t_j^{(i)}$ where $x_j = 1$ implies that the image repeat was correctly detected when



(a) The efficient visual memory game. Each image is shown for 600ms, separated by a blank fixation of 800ms. The worker can press a key anytime during this 1.4s.

(b) Human consistency.

Figure 2. Illustration of the efficient visual memory game (left), and the resulting human consistency averaged over 25 random splits (right) obtained using the proposed method on the LaMem dataset.

it was shown after time t_j. The memorability score proposed by [13] is the average hit rate per image, which can also be seen as the value that minimizes the ℓ_2 error $\sum_j ||x_j^{(i)} - m^{(i)}||_2^2 \implies m^{(i)} = \frac{1}{n^i}\sum_j x_j^{(i)}$. In this case, the different times of repeat presentation, t_j, are not taken into account explicitly as all repeats are shown at about the same delay to all participants. Next, we modify the above model to suit our new scenario with variable delays.

Memorability follows a log-linear relationship with time delay between images [13]. Let us assume that the memorability of image i is $m_T^{(i)}$ when the time interval between repeated displays is T. Thus, we can write the memorability of image i as $m_T^{(i)} = \alpha \log(T) + c^{(i)}$, where $c^{(i)}$ is the *base memorability* for the given image and α is the decay factor of memorability over time. Similarly, for some other time t, we can write the memorability of the same image as $m_t^{(i)} = \alpha \log(t) + c^{(i)}$. Thus, we obtain the relationship:

$$m_t^{(i)} - m_T^{(i)} = \alpha \log(t) - \alpha \log(T) \tag{1}$$

$$\implies m_t^{(i)} = m_T^{(i)} + \alpha \log\left(\frac{t^{(i)}}{T}\right) \tag{2}$$

As before, we have some n observations for image i given by $x_j^{(i)} \in \{0, 1\}$ and $t_j^{(i)}$ where $x_j = 1$ implies that the image repeat was correctly detected when it was shown after time t_j. For N images, we can now write the overall ℓ_2 error, E, as:

$$E(\alpha, m_T^{(i)}) = \sum_{i=1}^{N}\sum_{j=1}^{n^{(i)}} ||x_j^{(i)} - m_{t_j}^{(i)}||_2^2 \tag{3}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{n^{(i)}} \left|\left| x_j^{(i)} - \left[m_T^{(i)} + \alpha \log\left(\frac{t_j^{(i)}}{T}\right)\right]\right|\right|_2^2 \tag{4}$$

Note that we write the combined error (as compared to individual errors per image) as the decay factor α is shared across all images. Our goal is to find $m_T^{(i)}$ and α that minimize E. By adjusting the value of T, we can adjust the time delay at which we want to find the memorability score. Also, by finding all scores at a fixed delay T, the scores for all images become comparable, as is the case in the model proposed by [13].

Optimization: We observe that we can find the global minima of E with respect to $m_T^{(i)}$ if we fix the value of α, and similarly, we can find α if we fix the value of $m_T^{(i)}$. Thus, we can minimize E by iteratively updating α, followed by $m_T^{(i)}$ and so on. By differentiating E with respect to each of the variables, and setting it to 0, we can find the update equations:

$$\alpha \leftarrow \frac{\sum_{i=1}^{N}\frac{1}{n^{(i)}}\sum_{j=1}^{n^{(i)}} \log(t_j^{(i)}/T)\left[x_j^{(i)} - m_T^{(i)}\right]}{\sum_{i=1}^{N}\frac{1}{n^{(i)}}\sum_{j=1}^{n^{(i)}}\left[\log(t_j^{(i)}/T)\right]^2} \tag{5}$$

and

$$m_T^{(i)} \leftarrow \frac{1}{n^{(i)}}\sum_{j=1}^{n^{(i)}}\left[x_j^{(i)} - \alpha \log\left(\frac{t_j^{(i)}}{T}\right)\right] \tag{6}$$

As the update equations find the global optima of E when keeping the other fixed, we ensure that the error is always decreasing, guaranteeing convergence. In practice, we initialize $m_T^{(i)}$ to the mean hit rate ignoring time delay, and find that approximately 10 iterations are enough for convergence. Note that our model has no hyperparameters.

2.3. Dataset experiments

In this section, we describe in detail the experimental setup of our efficient visual memory game, and conduct several experiments comparing the results of the proposed methodology with [13]. Further, we demonstrate that the proposed model can increase human consistency by accounting for variable time delays between repeats and it results in a consistent decay factor, α, across splits.

Experimental setup: The efficient visual memory game is summarized in Fig. 2(a). We conducted memorability experiments using Amazon's Mechanical Turk (AMT) on the $60,000$ target images obtained by sampling the various datasets mentioned in Sec. 2.1. Each task lasted about 4.5 minutes consisting of a total of 186 images divided into 66

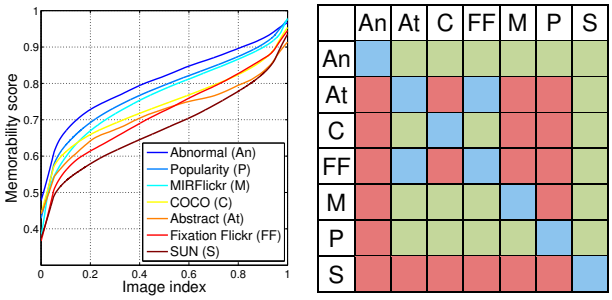
targets, 30 fillers, and 12 vigilance repeats. Targets were repeated after at least 35 images, and at most 150 images. Vigilance repeats were shown within 7 images from the first showing. The vigilance repeats ensured that workers were paying attention leading to a higher quality of results. Workers who failed more than 25% of the vigilance repeats were blocked, and all their results discarded. Further, we used a qualification test to ensure the workers understood the task well. We obtained 80 scores per image on average, resulting in a total of about 5 million data points. Similar to [13], we use rank correlation to measure consistency.

Comparison with [13]: Before describing the results on our new dataset, we first compare the performance of our method to the one proposed by Isola et al [13] on the SUN memorability dataset to ensure that our modifications are valid. We randomly selected 500 images from their dataset, and collected 80 scores per image. After applying our algorithm to *correct* the memorability scores, we obtained a within-dataset human rank correlation of 0.77 (averaged over 25 random splits), as compared to 0.75 using the data provided by [13]. Further, we obtain a rank correlation of 0.76 when comparing the independently obtained scores from the two methods. This shows that our method is well suited for collecting memorability scores.

Results on LaMem: Fig. 2(b) shows the human consistency as the number of number of human annotations per image increases. At 80 scores per image, we obtain a human rank correlation of 0.67 (averaged over 25 random splits) if we simply take the average of the correct responses ignoring the difference in time delays (i.e., same formula as [13]) which increases to 0.68 after applying our method. Note that the impact of using our method is small in this case as the range of average delays of each image is relatively small, ranging only from 62 to 101 intervening images. While our method can rectify the errors caused by variable delays, the error here is rather insignificant.

To further verify our algorithm, we created *adversarial* splits of the data where the responses for each image are divided based on the delays i.e., all the responses when delays are low go into one split, and all the responses when delays are high go into the other split. We randomly assign the low and high delay split of each image to different overall splits. Using the method of [13] (i.e., simple averaging) in this case significantly reduces the human rank correlation to 0.61, which can be restored to 0.67 using our method. This demonstrates the importance of applying our method when the interval distribution is more diverse.

Interestingly, we find that the decay factor, α, found by our method is largely consistent across various splits of data, having a standard deviation of less than 1% from the mean. This further verifies the finding made by [13] that memorability decays consistently over time, and our method provides a robust way to estimate this decay factor.



(a) Memorability scores (b) Statistical differences in mean

Figure 3. (a) Memorability scores of images from different datasets. For each dataset, the memorability scores of the images are independently sorted from low to high i.e., image index 0 to 1. Note that the image index ranges from 0 to 1 (instead of 1 to N) as each dataset has a different number of images. (b) Matrix indicating if the differences in the mean memorability scores of different datasets are statistically significant at the 5% level of significance. Blue indicates no difference, red indicates *col*>*row*, while green indicates *row*>*col*.

Overall, the high human consistency obtained on LaMem despite the large variety of images strengthens the importance of the concept of memorability and shows that it is a universal and intrinsic property of each image.

3. Understanding Memorability

As described in Sec. 2.1, LaMem is composed of a variety of other datasets that contain additional annotation such as aesthetics, popularity, image emotions, objects, and so on. In this section, we explore the relationship of some of these image attributes to memorability.

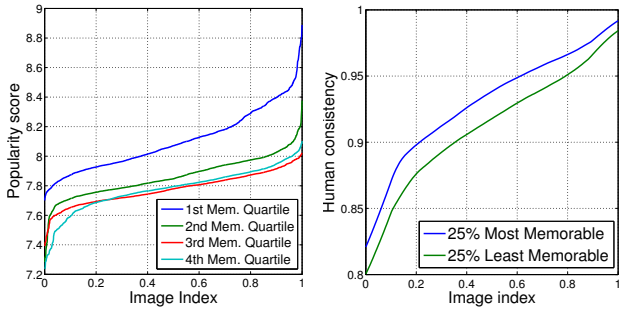
3.1. Differences across datasets

In Fig. 3(a), we plot the memorability scores of some of the datasets contained in LaMem[1]. We find that the distribution of memorability scores for the different datasets tends to look rather different. While images from the Abnormal Objects dataset [31] and image popularity dataset [16] tend to be extremely memorable, those from the SUN dataset [34] tend to be rather forgettable. In Fig. 3(b) we evaluate whether these perceived differences are statistically significant using a one-sided t-test. We find that most of the differences are significant at the 5% level.

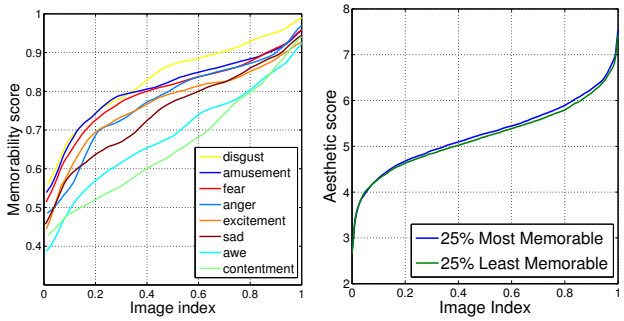
3.2. Image attributes

In this section, we explore how some of the image attributes, such as popularity, saliency, emotions and aesthetics, affect the memorability of images and vice-versa. We would like to highlight that the significant diversity of LaMem allows for this exploration at a large-scale.

[1]For clarity, we only show the plots for a subset of the datasets. The full set of plots are provided in the supplemental material.



(a) Memorability vs popularity (b) Memorability vs saliency



(c) Memorability vs emotions (d) Memorability vs aesthetics

Figure 4. Plots showing the relationship of memorability and various image attributes. For each curve, the images are sorted independently using ground-truth memorability scores. As each curve may contain a different number of images, the image index above has been normalized to be from 0 to 1.

Popularity: In [16], popularity is defined as the log-normalized view-count of an image on Flickr. Using the 5000 images from this dataset contained in LaMem, in Fig. 4(a), we plot the popularity scores of the images divided into quartiles based on their ground-truth memorability scores. We find that the popularity scores of the most memorable images (1st quartile) are statistically higher than those of the other quartiles[2]. On the other hand, when the memorability scores are low-medium, there is little difference in the popularity scores. This could be an insightful finding for people attempting to design images that become popular. Note that even though these images are popular on Flickr, we do not expect the AMT workers to have seen them before in general as the most popular image had fewer than 100k views. Furthermore, if they had seen the image before, they would have generated a false alarm on the first presentation of the image resulting in a lower memorability score for the image.

Saliency: Using images from the Fixation Flickr [14] dataset, we explore the relationship between human fixations and memorability. As shown in Fig. 4(b), we find that images that are more memorable tend to have more consistent human fixations. In fact, we find that human fixation consistency and memorability have a reasonable rank correlation of 0.24. A high human consistency on saliency often occurs when humans have one or a few specific objects to fixate on, which would tend to imply that the image contains more close-ups or larger objects. Essentially, this suggests that when humans have a specific point of focus in an image, they are better able to remember it and vice versa. These findings are similar to those of [26] and [3].

Emotions: In Fig. 4(c), we plot the memorability scores of images portraying various emotions from the affective images dataset [25]. We find that images that evoke *disgust* are statistically more memorable than images showing most other emotions, except for *amusement*. Further, images portraying emotions like *awe* and *contentment* tend to be the least memorable. This is similar to the findings in [12] where they show attributes like 'peaceful' are strongly negatively correlated with memorability. Overall, we find that images that evoke negative emotions such as *anger* and *fear* tend to be more memorable than those portraying positive ones. The analysis on the statistical differences between the memorability of emotions is in the supplemental material.

Aesthetics: Fig. 4(d) shows the aesthetic scores of the 25% most and least memorable images from the AVA dataset [27]. As in [1, 13], we find that the aesthetic score of an image and its memorability have little to no correlation.

4. Predicting Memorability

In this section, we focus on predicting image memorability using deep networks. In Sec. 4.1, we describe the experimental setup and our approach, MemNet, for predicting memorability. Then, in Sec. 4.2 and 4.3, we apply the proposed algorithms to the SUN memorability dataset and our new LaMem dataset respectively. Last, in Sec. 4.4 we provide additional analysis such as visualizing the internal representation learned by MemNet.

4.1. MemNet: CNN for Memorability

Given the recent success of convolutional neural networks (CNN) in various visual recognition tasks [10, 21, 29, 35, 32, 37], we use them here for memorability prediction. As memorability depends on both scenes and objects, we initialize the training using the pre-trained Hybrid-CNN from [37], trained on both ILSVRC 2012 [30] and Places dataset [37]. Memorability is a single real-valued output, so we use a Euclidean loss layer to fine-tune the Hybrid-CNN. We call our final network *MemNet*.

Setup and baseline: We followed the same experimental procedure as [13] where we distribute the data into random train and test splits: the train split is scored by one half of the workers, and the test split by the other half. For the SUN Memorability dataset, we repeat the experiment for

[2]The same holds if we plot the memorability scores of the 25% most and least popular images. This is included in the supplemental material.

Test set:		Train set: SUN Memorability					Train set: LaMem				
		fc6	fc7	fc8	fine-tune	HOG2x2	fc6	fc7	fc8	MemNet	HOG2x2
SUN Mem	no FA	0.57	0.60	0.58	0.51	0.45	0.56	0.59	0.57	0.59	0.47
	with FA	0.61	**0.63**	0.62	0.53	0.48	0.57	0.59	0.58	**0.61**	0.48
LaMem	no FA	0.46	0.48	0.46	0.43	0.35	0.54	0.55	0.53	0.57	0.40
	with FA	0.52	0.54	**0.55**	0.47	0.43	0.61	0.61	0.60	**0.64**	0.47

Table 1. Rank correlation of training and testing on both LaMem and SUN Memorability datasets. The reported performance is averaged over various train/test splits of the data. For cross-dataset evaluation, we use the full dataset for training and evaluate on the same test splits to ensure results are comparable. 'fc6', 'fc7' and 'fc8' refer to the different layers of the Hybrid-CNN [37], and 'FA' refers to false alarms. Please refer to Sec. 4.1 for additional details.

25 splits, but for LaMem, we use 5 splits due to the computationally expensive fine-tuning step. As the baseline, we report performance when using HOG2x2 features that are extracted in a similar manner to [18] i.e., we densely sample HOG [4] in a regular grid and use locality-constrained linear coding [33] to assign descriptors to a dictionary of size 256. Then, we combine features in a spatial pyramid [22] resulting in a feature of dimension 5376. This is the best performing feature for predicting memorability as reported by various previous works [15, 18, 13]. For both HOG2x2 and features from CNNs, we train a linear Support Vector Regression machine [8, 6] to predict memorability. We used validation data to find the best B and C hyperparameters[3].

As proposed in [15], we evaluate two notions of memorability - one that does not account for false alarms (no FA), and one that does (with FA). It can be important to account for false alarms to reduce the *noise* in the signal as people may remember some images simply because they are *familiar*, but not *memorable*. Indeed, we find that this greatly improves the prediction rank correlation despite using the same features. In our experiments, we evaluate performance using both metrics. Note that the models for 'no FA' and 'with FA' as mentioned in Tbl. 1 are trained independently.

4.2. SUN Memorability dataset

Tbl. 1 (left) shows the results of training on the SUN Memorability dataset and testing on both datasets. We observe that deep features significantly outperform the existing state-of-the-art by about 0.15 (0.63 vs 0.48 with FA, and 0.60 vs 0.45 no FA). This demonstrates the strength of the deep features as shown by a variety of other works. Similar to [15], we observe that the performance increases significantly when accounting for false alarms. Apart from high performance on the SUN Memorability dataset, the features learned by CNNs generalize well to the larger LaMem dataset. Despite having significantly less variety in the type of images, the representational power of the features allow the model to perform well.

Fine-tuning has been shown to be important for improving performance [29], but we find that it reduces performance when using the SUN Memorability dataset. This is due to the limited size of the data, and the large number of network parameters, leading to severe overfitting of the training data. While the rank correlation of the training examples increases over backpropagation iterations, the validation performance remains constant or decreases slightly. This shows the importance of having a large-scale dataset for training a robust model of memorability.

Note that Tbl. 1 only compares against having the single best feature (HOG2x2), but even with multiple features the best reported performance [18] is 0.50 (no FA), which we outperform significantly. Interestingly, our method also outperforms [13] (0.54, no FA) and [19] (0.58, no FA) which use various ground truth annotations such as objects, scenes and attributes.

4.3. LaMem dataset

Tbl. 1 (right) shows the results of training on the LaMem dataset, and testing on both datasets. In this case, we split the data to 45k examples for training, 4k examples for validation and 10k examples for testing. We randomly split the data 5 times and average the results. Overall, we obtain the best rank correlation of 0.64 using MemNet. This is remarkably high given the human rank correlation of 0.68 for LaMem. Importantly, with a large-scale dataset, we are able to successfully fine-tune deep networks without overfitting severely to the training data, and preserving generalization ability in the process.

Additionally, we find that the learned models generalize well to the SUN Memorability dataset achieving a comparable performance to training on the original dataset (0.61 vs 0.63, with FA). Further, similar to the SUN Memorability dataset, we find that higher performances can be attained when accounting for the observed false alarms.

4.4. Analysis

In this section, we investigate the internal representation learned by MemNet. Fig. 5 shows the average of images that maximally activate the neurons in two layers near the output of MemNet, ordered by their correlation to memorability. We see that many units near the top of *conv5* look like close-ups of humans, faces and objects while units

[3]Note that, since Liblinear [8] regularizes the bias term, B, we found that it was important to vary it to maximize performance.

Figure 5. Visualizing the CNN features after fine-tuning, arranged in the order of their correlation to memorability from highest (top) to lowest (bottom). The visualization is obtained by computing a weighted average of the top 30 scoring image regions (for conv5, this corresponds to its theoretical receptive field size of $163 * 163$, while for fc7 it corresponds to the full image) for each neuron in the two layers. From top to bottom, we find the neurons could be specializing for the following: people, busy images (lots of gradients), specific objects, buildings, and finally open scenes. This matches our intuition of what objects might make an image memorable. Note that fc7 consists of 4096 units, and we only visualize a random subset of those here.

Figure 6. The segmentations produced by neurons in conv5 that are strongly correlated, either positively or negatively, with memorability. Each row corresponds to a different neuron. The segmentations are obtained using the data-driven receptive field method proposed in [36].

near the bottom (so associated with more forgettable objects) look more like open and natural scenes, landscapes and textured surfaces. A similar trend has been observed in previous studies [13]. Additionally, to better understand the internal representations of the units, in Fig. 6, we apply the methodology from [36] to visualize the segmentation produced by five neurons from *conv5* that are strongly correlated with memorability (both positively and negatively). We observe that the neurons with the highest positive correlation correspond to body parts and faces, while those with a strong negative correlation correspond to snapshots of natural scenes. Interestingly, these units emerge automatically in MemNet without any explicit training to identify these particular categories.

5. Applications

In this section, we investigate whether our model can be applied to understanding the contribution of image regions to memorability [18]. Predicting the memorability of image regions could allow us to build tools for automatically modifying the memorability of images [17], which could have far-reaching applications in various domains ranging from advertising and gaming to education and social networking. First, we describe the method of obtaining memorability maps, and then propose a method to evaluate them using human experiments. Overall, using MemNet, we can accurately predict the memorability of image regions.

To generate memorability maps, we simply scale up the image and apply MemNet to overlapping regions of the image. We do this for multiple scales of the image and average the resulting memorability maps. To make this process computationally efficient, we use an approach similar to [24]: we convert the fully-connected layers, fc6 and fc7 to convolutional layers of size $1 * 1$, making the network fully-convolutional. This fully-convolutional network can now be applied to images of arbitrary sizes to generate different sized memorability maps e.g., an image of size 451×451 would generate an output of size 8×8. We do this for several different image sizes and average the outputs to generate the final memorability map (takes ~1s on a typical GPU). The second column of Fig. 7 shows some of the resulting memorability maps. As expected, the memorability maps tend to capture cognitively salient regions that contain meaningful objects such as people, animals or text.

While the maps appear semantically meaningful, we still need to evaluate whether the highlighted regions are truly the ones leading to the high/low memorability scores of the images. Given the difficulty of generating photorealistic renderings of varying details, we use non-realistic photo-renderings or cartoonization [5] to emphasize/de-emphasize different parts of an image based on the memorability maps, and evaluate its impact on the memorability of an image. Specifically, given an image and a heatmap, we investigate the difference in human memory for the following scenarios: (1) *high* – emphasizing regions of high memorability and de-emphasizing regions of low memorability (Fig. 7 col 3), (2) *medium* – having an *average* emphasis across the entire image (Fig. 7 col 4), and (3) *low* – emphasizing regions of low memorability and de-emphasizing regions of high memorability (Fig. 7 col 5). If our algorithm is identi-

fying the *correct* memorability of image regions, we would expect the memorability of the images from the above three scenarios to rank as *high* > *medium* > *low*.

Following the above procedure, we generate three cartoonized versions of 250 randomly sampled images based on the memorability maps generated by our algorithm. We use our efficient visual memory game (Sec. 2) to collect memorability scores of the cartoonized images on AMT. We ensure that a specific worker can see exactly one modification of each image. Further, we also cartoonize the filler and vigilance images to ensure that our target images do not stand out. We collect 80 scores per image, on average. The results of this experiment are summarized in Fig. 8. Interestingly, we find that our algorithm is able to reliably identify the memorability of image regions. All pairwise relationships, *low*<*medium*, *low*<*high* and *medium*<*high* are statistically significant (5% level). This shows that the memorability maps produced with our method are reliable estimates of what makes an image memorable or forgettable, serving as a building block for future applications. We also observe that the memorability of all cartoonized versions of an image tends to be lower than the original image, even though the *high* version emphasizes the more memorable regions. We expect that this is because even the *high* version of the image loses significant details of objects as compared to the original photograph. This might make it harder for people to distinguish between images and/or identify the objects.

6. Conclusion

Using deep learning and LaMem, a novel diverse dataset, we show unprecedented performance at estimating the memorability ranks of images, and introduce a novel method to evaluate memorability maps. We envision that many applications can be developed out of this work. For instance, for visual understanding systems, leveraging memorability would be an efficient way to concisely represent or alter information while skipping over irrelevant (forgettable) information. Understanding why certain things are memorable could lead to making systems and devices that preferentially encode or seek out this kind of information, or that store the important information that humans will certainly forget. For learning and education, new visual materials could be enhanced using the memorability maps approach, to reinforce forgettable aspects of an image while also maintaining memorable ones. In general, consistently identifying which images and which parts of an image are memorable or forgettable could be used as a proxy for identifying visual data useful for people, concisely representing information, and allowing people to consume information more efficiently.

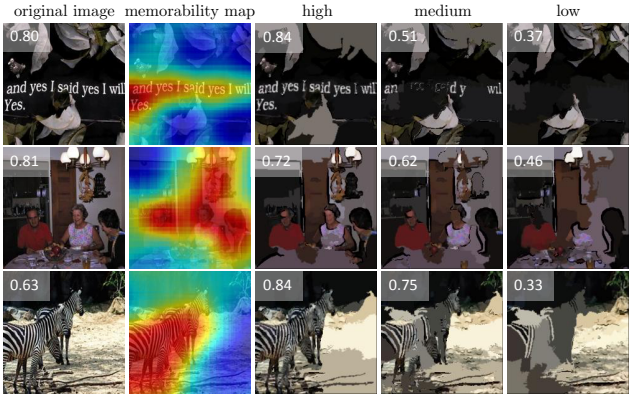


Figure 7. The memorability maps for several images. The memorability maps are shown in the jet color scheme where the color ranges from blue to red (lowest to highest). Note that the memorability maps are independently normalized to lie from 0 to 1. The last three columns show the same image modified using [5] based on the predicted memorability map: *high* image – regions of high memorability are emphasized while those of low memorability are de-emphasized e.g., in the first image text is visible but leaves are indistinguishable, *medium* image – half the image is emphasized at random while the other half is de-emphasized e.g., some text and some leaves are visible for the first image, and *low* image – regions of low memorability are emphasized while those of high memorability are de-emphasized e.g., text is not visible in first image but leaves have high detail. The numbers in white are the resulting memorability scores of the corresponding images.

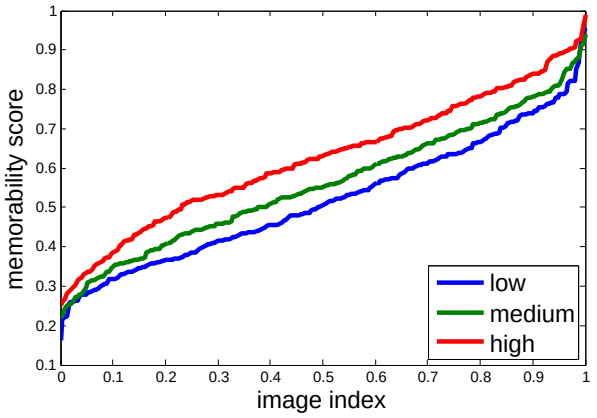


Figure 8. Memorability scores of the cartoonized images for the three settings shown in Fig. 7. Note that the scores for *low*, *medium* and *high* are independently sorted. Additional results are provided in the supplemental material.

Acknowledgements. We thank Wilma Bainbridge, Phillip Isola and Hamed Pirsiavash for helpful discussions. This work is supported by a National Science Foundation grant (1532591), the McGovern Institute Neurotechnology Program (MINT), MIT Big Data Initiative at CSAIL, research awards from Google and Xerox, and a hardware donation from Nvidia.

References

[1] W. A. Bainbridge, P. Isola, and A. Oliva. The intrinsic memorability of face photographs. *JEPG*, 2013. 1, 5

[2] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci, USA*, 105(38), 2008. 1

[3] B. Celikkale, A. T. Erdem, and E. Erdem. Visual attention-driven spatial pooling for image memorability. In *CVPR Workshop*. IEEE, 2013. 5

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6

[5] D. DeCarlo and A. Santella. Stylization and abstraction of photographs. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 769–776. ACM, 2002. 7, 8

[6] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *NIPS*, 1997. 6

[7] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem. What makes an object memorable? In *International Conference on Computer Vision (ICCV)*, 2015. 1

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 2008. 6

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2014. 5

[11] M. J. Huiskes and M. S. Lew. The MIR Flickr retrieval evaluation. In *ACM MIR*, 2008. 2

[12] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the instrinsic memorability of images. In *NIPS*, 2011. 1, 5

[13] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE PAMI*, 2014. 1, 2, 3, 4, 5, 6, 7

[14] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *CVPR*, 2009. 2, 5

[15] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva. Modifying the memorability of face photographs. In *ICCV*, 2013. 1, 6

[16] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *WWW*, 2014. 2, 4, 5

[17] A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva. Image memorability and visual inception. In *SIGGRAPH Asia 2012 Technical Briefs*, page 35. ACM, 2012. 7

[18] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *NIPS*, 2012. 1, 6, 7

[19] J. Kim, S. Yoon, and V. Pavlovic. Relative spatial features for image memorability. In *ACM MM*, 2013. 6

[20] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva. Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psych Science*, 21(11), 2010. 1

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 5

[22] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 6

[23] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361, 1995. 1

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015. 7

[25] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010. 2, 5

[26] M. Mancas and O. Le Meur. Memorability of natural scenes: The role of attention. In *ICIP*. IEEE, 2013. 5

[27] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. *CVPR*, 2012. 2, 5

[28] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In *ECCV*. 2010. 2

[29] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014. 5, 6

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[31] B. Saleh, A. Farhadi, and A. Elgammal. Object-centric anomaly detection by attribute-based reasoning. In *CVPR*, 2013. 2, 4

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deep-face: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 5

[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 6

[34] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2, 4

[35] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. *CVPR*, 2014. 5

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *ICLR*, 2015. 7

[37] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 1, 5, 6