Is Image Memorability Prediction Solved?

Shay Perera
Technion
shaymush@gmail.com

Ayellet Tal
Technion
ayellet@ee.technion.ac.il

Lihi Zelnik-Manor
Technion
lihi@ee.technion.ac.il

Abstract

This paper deals with the prediction of the memorability of a given image. We start by proposing an algorithm that reaches human-level performance on the LaMem dataset—the only large scale benchmark for memorability prediction. The suggested algorithm is based on three observations we make regarding convolutional neural networks (CNNs) that affect memorability prediction. Having reached human-level performance we were humbled, and asked ourselves whether indeed we have resolved memorability prediction—and answered this question in the negative. We studied a few factors and made some recommendations that should be taken into account when designing the next benchmark.

1. Introduction

Our lives wouldn't be the same if we were unable to store visual memories. The vast majority of the population relies heavily on visuals to identify people, places and objects. Interestingly, despite different personal experiences, people tend to remember and forget the same pictures [1, 12]. This paper deals with the ability of algorithms to assess the memorability of a given image. Good memorability prediction could be useful for many applications, such as improving education material, storing for us things we tend to forget, producing unforgettable ads, or even presenting images in a way that is easier to consume.

Image memorability is commonly measured as the probability that an observer will detect a repetition of a photograph a few minutes after exposition, when presented amidst a stream of images [1, 12, 13, 14, 2, 8], as illustrated in Figure 1. According to cognitive psychological studies, this measurement determines which images left a trace in our long-term memory [13, 3, 18, 23].

Several methods for memorability prediction have been proposed over the years [13, 11, 20, 16]. A key observation is that both the type of scene and the type of objects in the image are highly related to its memorability [12, 14, 8]. Based on this observation, Khosla *et al*. [14] collected a large-scale dataset, called *LaMem*, and proposed *MemNet* for memorability prediction.

In this paper we describe a system that achieves what seem to be astonishing results—reaching the limit of human performance on *LaMem*. Does this mean that image memorability prediction is a solved problem? To answer this question, we look deeper into the factors that impact human performance. We discuss some factors that have been overlooked when building the existing datasets for image memorability. Our observations may lead to further studies not only in the design of meaningful datasets and effective algorithms for memorability prediction, but also in finding additional factors influencing memorability.

In the first part of the paper (Section 3), we propose a framework, *MemBoost*, for predicting image memorability, which achieves state-of-the-art results on all existing datasets. This is done by delving into the relation between networks for image classification and memorability prediction. Our study gives rise to three main insights on which we base the design of *MemBoost*: (i) As object classification CNNs improve, so does memorability prediction. (ii) Scene classification plays a bigger role in memorability prediction than object classification. This resolves conflicting opinions on the matter. (iii) It suffices to train a regression layer on top of a CNN, which is designed and trained for object & scene recognition, to achieve on par results with those attained by re-training the entire CNN for memorability prediction. This insight contradicts previous observations.

Since our prediction results are surprising, in the second part of the paper (Section 4) we re-visit some aspects that influence human performance in the memory game. Via empirical analysis we show that changing some of the design decisions in the data collection could lead to data that better represents human memorability. The main conclusion from this study is that reaching human performance on LaMem does *not* mean that memorability prediction has been solved. We further provide guidelines for building future datasets.

In summary, this paper makes three major contributions. First, it presents insights that should be the basis for mem-

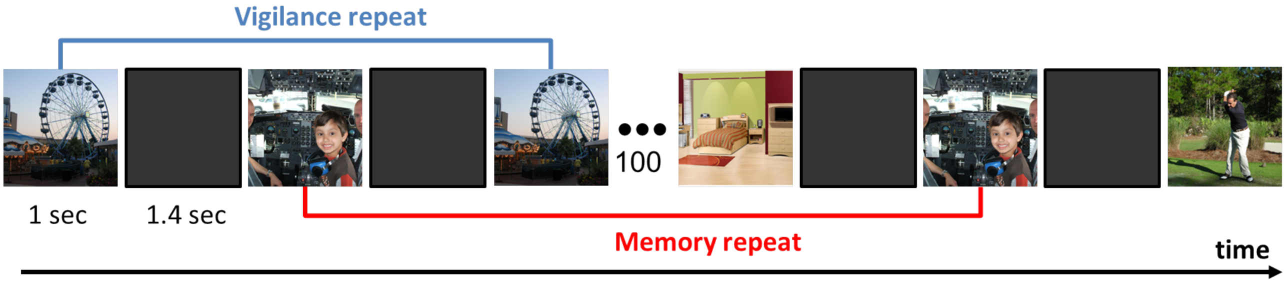


Figure 1: **Visual Memorability Game.** Participants watch for repeats in a long stream of images.

orability prediction algorithms. A couple of these insights have already been demonstrated for other tasks in computer vision. Second, the paper suggests a new framework that achieves state-of-the-art results, reaching the limit of human performance on LaMem. Third, we put in question the portrayal of human memorability by the current datasets and give some recommendations towards the creation of the next large dataset.

2. Previous Work

We start by describing how the ground-truth data has been collected. We then review image attributes that have been studied with regard to memorability.

The memory game. Image memorability is commonly measured using a memory-game approach, which was originally proposed by Isola *et al*. [13]. Briefly, the participants view a sequence of images, each of which is displayed for a predefined period of time, with some gap in between image presentations, as illustrated in Figure 1. The task of the participants is to press a button whenever they see an identical repeat of an image at any time in the sequence [3, 18]. The participants receive feedback (correct or incorrect) whenever they press a key.

Unbeknown to the participants, the sequence of images is composed of targets and fillers, both are randomly sampled from the dataset. The role of the fillers is two-fold. First, they provide spacing between the first and the second repetition of a target. Second, responses on repeated fillers constitute a vigilance task that allowed us to continuously check that the participants are attentive to the task [11, 12]. Each target is repeated exactly once and each filler is presented at most once (vigilance-task fillers are sequenced to repeat exactly once).

The memorability score assigned to each target image is the percentage of correct detections by the participants. Throughout this paper, we refer to the memorability scores collected through the memory game as ground truth. It is believed that by randomizing the sequence each participant sees, the measurements depend only on factors that are intrinsic to the image, independent of extrinsic variables, such as display order, time delay, and local visual context.

Observations. Since our study deals with the relation between scenes, objects, and memorability, we next provide a brief review of previous observations regarding which image attributes affect (or do not affect) image memorability.
1. *Object category and scene category attributes [12, 13, 8, 11, 16, 15, 4].* A bunch of studies on this topic concluded that some object categories, such as people, vehicles, and animals, and some scene categories, such as indoor scenes, are more memorable than others.
2. *Semantic attributes.* Scene semantics go beyond just content and scene category [12, 13, 11]. Features such as spatial layout or actions are highly correlated with memorability and are an efficient way of characterizing memorability.
3. *Saliency.* While Khosla *et al*. [15] and Celikkale *et al*. [5] show a reasonable correlation between memorability and attention, Mancas *et al*. [20] show almost no correlation between the two. Furthermore, Khosla *et al*. [14] found a reasonable correlation between fixation duration to memorability. When considering object memorability, rather than image memorability, Dubey *et al*. [8] found high correlation with the number of unique fixation points within the object.
4. *Object statistics [12, 13, 11].* The number of objects and the object area have low correlation with memorability and are ineffective at predicting memorability.
5. *Aesthetics, interestingness, emotion and popularity.* Both image aesthetics and interestingness show no correlation to memorability [12, 14], while they are correlated with each other. Popularity is correlated to memorability only for the most memorable images, but not for others [14]. Negative emotions, such as anger and fear, tend to be more memorable than those portraying positive ones [14].
6. *Colors [12, 13, 11, 17, 9].* Colors have only weak correlation with memorability.
7. *What people think is memorable [12].* Asking people to guess which images are the most memorable in a collection reveals low correlation to the actually memorable ones. Participants have wrong intuition, erroneously as-

Setup	**Memorability Prediction**			**Classification**	
Vary Network	**LaMem**	**SUN-Mem**	**Figrim**	**SUN**	**Places**
ResNet152-ImageNet+XGBoost	**0.64**	0.64	**0.56**	-	54.74
VGG16-ImageNet+XGBoost	**0.64**	**0.65**	**0.56**	48.29	55.24
GoogLeNet-ImageNet+XGBoost	0.62	0.63	0.55	43.88	53.63
AlexNet-ImageNet+XGBoost	0.61	0.6	0.51	42.61	53.17

Table 1: **Memorability prediction improves with classification.** The table shows results of two tasks: image memorability prediction and image classification. For classification, we show results on two datasets: SUN [25] and Places [27]. The accuracy is computed for four architectures, all trained on ImageNet. Consistently, as classification accuracy improves, so does memorability prediction. We adopt the following naming and color-code convention for the setups: Network-Training dataset+Regression type.

suming that beautiful and interesting images will produce a lasting memory.

3. MemBoost: A System for Predicting Image Memorability

Motivated by the strong evidence for the correlation between scenes, objects and image memorability, we explored the utility of this correlation for training CNNs. We start our study with short descriptions of three insights we make on memorability prediction with CNNs. Then, we propose the *MemBoost* algorithm, which is based on these insights, and provides state-of-the-art results.

Our experimental setup follows that of [14] for a variety of datasets and for a variety of networks. Briefly, we randomly split the images in each memorability dataset into a train set and a test set. We performed experiments on three datasets: LaMem [14], SUN-Mem [13], and Figrim [4]. We note that LaMem is a relatively large dataset, consisting of $58,741$ images, whereas the other datasets consist of only $2,222$ and $1,754$ images, respectively; see the appendix for details on these dataset (Table 5). For SUN-Mem and Figrim, splitting was repeated for 25 times, while for LaMem, we used 5 splits due to its size.

In order to measure the prediction performance, we follow the common practice in the evaluation of memorability prediction. That is to say, rather than comparing memorability scores directly, we compare ranks as follows. The images in the test set are ranked both according to their ground-truth memorability scores and according to the algorithm predictions. The Spearman's rank correlation (ρ) is computed between the two rankings.

3.1. Insights

In this section we suggest three insights and explore their validity, one by one, via thorough experiments.

1. Use a strong base CNN. It has recently become common knowledge that the stronger your backbone CNN is, the better results you'd get, even if the base CNN was not trained for your specific task. In accordance with this, we compare in Table 1 four architectures: ResNet152 [10], VGG16 [22], GoogLeNet [24] and AlexNet [19]. All of them were trained on ImageNet [7] and XGBoost was used for the regression layer. The table shows that indeed, as classification networks improve, so do the corresponding memorability prediction networks. These results are consistent across all three datasets. This implies that it might be unnecessary to develop special architectures for memorability prediction. Instead, it suffices to update the relevant layers of the best-performing classification network.

2. Training on scene classification is more important than training on object classification. Our second insight regards the essence of memorability: What makes an image memorable, the objects in it or the scene it describes? Answering this question is not only interesting theoretically, but it also has very practical implications, since it will enable a better selection of the training dataset for memorability prediction.

Previous works have found that scene category and object presence are, together, highly correlated with the memorability of an image (Spearman rank correlation $\rho = 0.43$) [13, 11, 15]. In [12] it is claimed that this correlation is mostly due to the scene category itself, which appears to summarize much of what makes an image memorable. However, this observation has not been used for memorability prediction.

We empirically verify this claim and show, in Table 2, that CNNs that are trained on datasets of scenes (Places205 [27]) outperform CNNs that are trained on datasets of objects (ImageNet [7]). However, the accuracy is not as good as that obtained when training on both objects and scenes (Hybrid1205 [27] or Hybrid1365 [26]). This behavior is persistent across datasets and networks (results are shown for both AlexNet and ResNet152).

We note that these datasets do not provide statistics regarding the balance between scenes and objects. Section 4

Setup	Memorability Prediction		
Vary training data	LaMem	SUN-Mem	Figrim
AlexNet-ImageNet+XGBoost	0.61	0.6	0.51
AlexNet-Places205+XGBoost	0.61	0.64	0.55
AlexNet-Hybrid1205+XGBoost	**0.64**	**0.65**	**0.57**
ResNet152-ImageNet+XGBoost	0.64	0.64	0.56
ResNet152-Places365+XGBoost	0.65	**0.66**	0.56
ResNet152-Hybrid1365+XGBoost	**0.67**	**0.66**	**0.57**

Table 2: **Scenes are more important than objects.** Memorability was predicted using AlexNet and ResNet152, trained on objects (ImageNet), on scenes (Places205 & Places365), or on their combination (Hybrid1205 & Hybrid1365). While scenes are more important than objects, their combination slightly improves the prediction. (The training datasets' details are given in the appendix).

	Training approach	LaMem	SUN	Figrim
	Human consistency	0.68	0.75	0.74
[14]	AlexNet-Hybrid1205+SVR	0.61	0.63	-
[MemNet]	AlexNet-Hybrid1205+Tune	**0.64**	0.53	-
Ours	AlexNet-Hybrid1205+XGBoost	**0.64**	**0.65**	**0.57**

Table 3: **Fine-tuning may be unnecessary.** Modifying only the regression layer provides comparable memorability prediction to re-training the entire network. The setup used by [14] is utilized.

discusses the validity of the training datasets. One such consideration is already clear—the right balance between objects and scenes should be sought after.

3. Re-training may be unnecessary. This insight regards the training of classification networks with memorability data. Is it really necessary to fine-tune the entire network or is it sufficient to train just the last regression layer? Answering this question in the affirmative means that we can achieve good results even when we have neither a lot of memorability data nor much computational resources for training. This is important since such data is not widely available and is difficult to collect, whereas classification data is more widespread.

Khosla *et al*. [14] compare two approaches for re-training a CNN for memorability prediction. The first approach re-trains only the last regression layer using Support Vector Regression (SVR). In the second approach, called *MemNet*, the entire network is fine-tuned with memorability data. They achieve better results with MemNet and conclude that fine-tuning the entire CNN is essential.

We reach the opposite conclusion. We show that modifying only the regression layer can provide comparable memorability prediction to re-training the entire network. In particular, we took the same network setup as [14], using AlexNet trained on Hybrid1205. We eliminated the classification layer *(top softmax layer)* and considered the previous layer as features. We then replaced the classification layer by training a regressor model, which is based on boosted trees (using XGBoost library [6]) and maps the features to memorability scores.

As can be seen in Table 3, fine-tuning only the regression layer with XGBoost is a good idea. In particular, on the large-scale LaMem, training just the regression layer yields the same accuracy as MemNet (0.64 in both cases). On the smaller dataset SUN-Mem, training just the regression layer even gives better results than re-training the entire network (0.65 in comparison to 0.53).

3.2. The MemBoost algorithm

Our next step is to utilize the observations from Section 3.1 to design a novel algorithm, called *MemBoost*. As suggested by insight (1), we select ResNet152 as our base network. We follow insight (2) and use a version named ResNet152-Hybrid1365 that was trained both on an object dataset (ImageNet [21]) and on a scene dataset (Places365 [26]). Last, in sync with insight (3), we modify only the regression layer, using XGBoost [6], to map deep features to memorability scores. Figure 2 illustrates the pipeline of our approach for acquiring state-of-the-art memorability prediction.

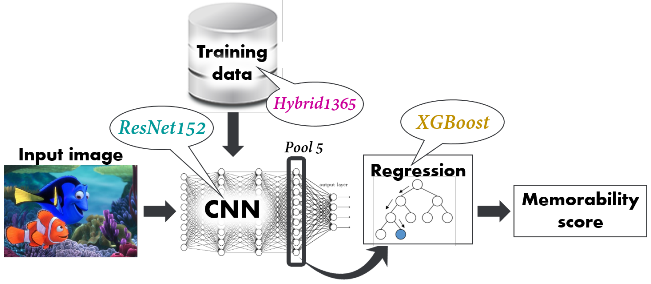
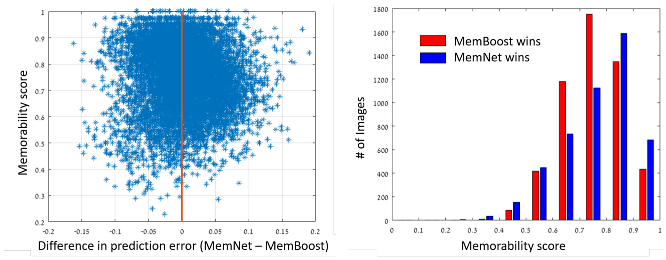


Figure 2: **The MemBoost pipeline.** Deep features are extracted from *pool5* layer of ResNet152, trained on Hybrid1365 dataset. A boosted-trees regression model is trained using these features to predict image memorability.

Table 4 summarizes our results. The top row of the table presents the memorability consistency across different groups of human observers. This serves as an upper bound. The next two rows show the previous best results, obtained by the two approaches of Khosla *et al*. [14]. The bottom-most row of the table shows the results of our MemBoost, which closes the gap with human prediction results on the LaMem dataset. MemBoost provides a significant improvement over MemNet on both LaMem and SUN-Mem ([14]

	Approach	LaMem	SUN-Mem	Figrim
	Human consistency	0.68	0.75	0.74
[14] [MemNet]	AlexNet-Hybrid1205+SVR	0.61	0.63	-
	AlexNet-Hybrid1205+Fine-tune	0.64	0.53	-
Our [MemBoost]	ResNet152-Hybrid1365+XGBoost	**0.67**	**0.66**	**0.57**

Table 4: **Memorability prediction results.** The table compares our MemBoost algorithm results with those of [14]. It shows that our insights lead to state-of-the-art memorability prediction on all three datasets.

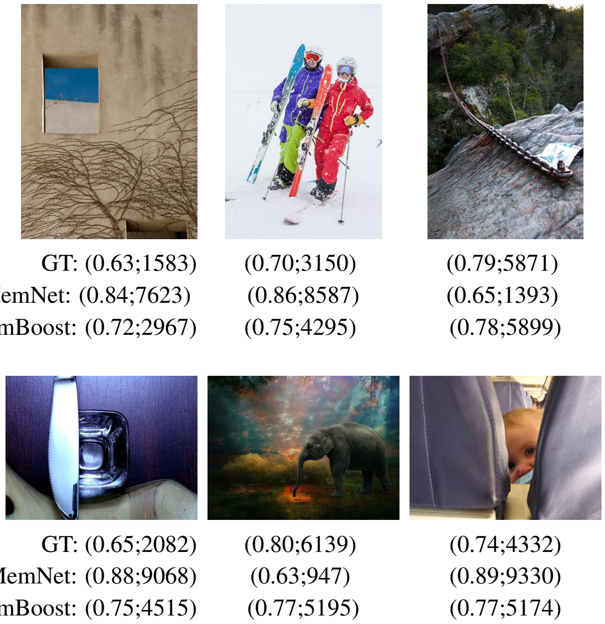


(a) (b)

Figure 3: **MemBoost vs. MemNet**. (a) Each point represents an image. The Y-axis shows memorability scores and the X-axis shows the difference in prediction error between MemBoost and MemNet. MemBoost is better when this difference is positive (right of the red line that represents equal-error). The distribution is not symmetric and shows that in the [0.6,0.8] range, i.e., when the images are not highly memorable or highly forgettable, MemBoost is more accurate. (b) Histograms of the number of images where MemBoost is more accurate vs. the number of images where MemNet is more accurate, as a function of the memorability score. It is evident that MemBoost is preferable in the mid-memorability range, while MemNet is more accurate for highly memorable images. Overall, MemBoost's results outperform those of MemNet.

did not test on Figrim and we failed to reproduce their training).

Result analysis. Figure 3 sheds light on where our predictions are more accurate than those of MemNet. Every point in the graph in Figure 3(a) corresponds to an image. The Y-axis shows the memorability score of the image. The X-axis shows the difference between MemBoost prediction error and MemNet's. When MemBoost outperforms MemNet, this difference is positive and vice versa. A symmetric distribution of points around the equal-error line (X=0) would mean both methods have similar distributions of errors. As can be seen, on LaMem the distribution looks more like a Pac-Man, with its mouth at medium memorability scores (0.6-0.8). In Figure 3(b) a different view of the same behavior is given. This stands to show that when the images are not highly memorable or highly forgettable, our algorithm wins. We believe that this is so since such images are more challenging as the scenes/objects are less recognizable. A more powerful prediction algorithm is hence essential in these cases.

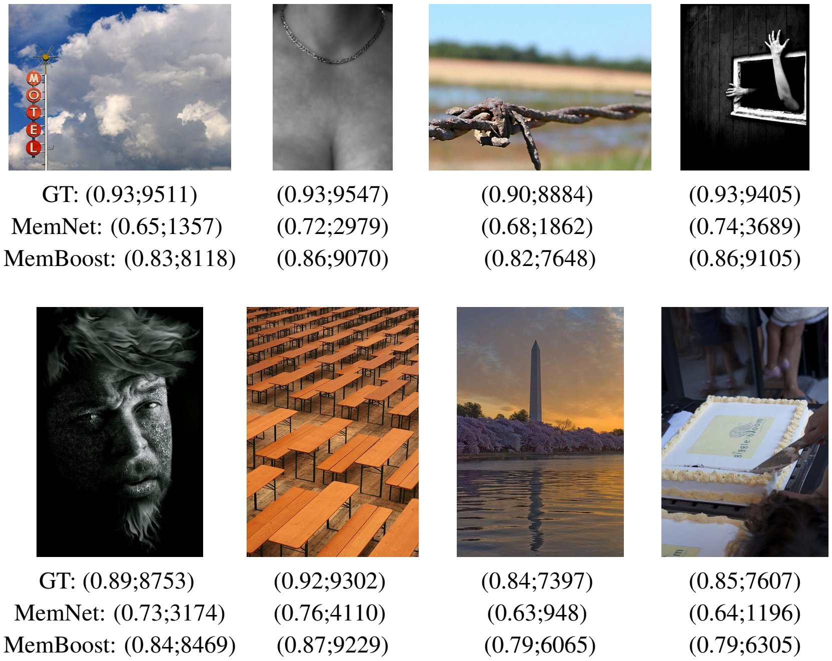


GT: (0.63;1583) (0.70;3150) (0.79;5871)
MemNet: (0.84;7623) (0.86;8587) (0.65;1393)
MemBoost: (0.72;2967) (0.75;4295) (0.78;5899)

GT: (0.65;2082) (0.80;6139) (0.74;4332)
MemNet: (0.88;9068) (0.63;947) (0.89;9330)
MemBoost: (0.75;4515) (0.77;5195) (0.77;5174)

Figure 4: **Qualitative results of images with medium memorability scores.** The memorability scores of these images are probably due to memorable objects and forgettable scenes (skiers, cats, baby, chain) or forgettable objects and memorable scenes (window, elephant, bathroom). Our algorithm, which combines objects and scenes, manages to do better than its competitors. For each image we present two values: its memorability score and its rank within the 10,000 images of Test_1 set of LaMem.

To complement our argument, we present in Figures 4 & 5 several example images and their corresponding scores. Figure 4 shows images that have medium memorability scores, where MemBoost outperforms MemNet. The common content of these images is having either memorable objects within forgettable scenes or forgettable objects within memorable scenes.



GT: (0.93;9511)	(0.93;9547)	(0.90;8884)	(0.93;9405)
MemNet: (0.65;1357)	(0.72;2979)	(0.68;1862)	(0.74;3689)
MemBoost: (0.83;8118)	(0.86;9070)	(0.82;7648)	(0.86;9105)
GT: (0.89;8753)	(0.92;9302)	(0.84;7397)	(0.85;7607)
MemNet: (0.73;3174)	(0.76;4110)	(0.63;948)	(0.64;1196)
MemBoost: (0.84;8469)	(0.87;9229)	(0.79;6065)	(0.79;6305)

Figure 5: **Qualitative results of images for which MemBoost outperforms MemNet by the largest margin.** These images contain common objects in distinct scenes, for which MemNet failed to predict memorability accurately.

This observation is reinforced in Figure 5, which shows the images for which we got the largest gap in prediction accuracy between MemBoost and MemNet. Interestingly, these images are all highly memorable, as evident from their memorability scores. They all show common objects, but in unique scenes—text on a cake, hands coming out of a window, extreme repetition of objects, etc. Our algorithm, which combines scenes and objects, manages to predict that these are unique combinations, and hence memorable.

4. On the Validity of Current Memorability Scores

Recall that the upper bound on LaMem is the mean rank correlation between the memorability scores corresponding to different groups of people. As our solution reached this upper bound, we ask ourselves whether the memorability scores, obtained through the Memory Game described in Section 2, form a sufficient representation. We explore three key questions and answer them via experiments:

1. Does the number of observers per image suffice as a representative sample?
2. How consistent are the scores across observers? Should the mean score be utilized by itself, or should the variance be considered as well?
3. Should the order in which the images are displayed be taken into account?

1. What is a sufficient sample set size? Recall that the memory game takes as scores the mean memorability over multiple participants. A key question then is 'how many participants should attend the game for the mean scores to be meaningful?'

Isole *et al.* [13] show that as the number of participants increases, the mean scores become more stable. To prove stability they show that averaging memorability scores over groups of 40 participants yields Spearman rank correlation of $\rho = 0.75$ between different groups (on the SUN Memorability dataset). Similarly, Khosla *et al.* [14] measure rank correlation of $\rho = 0.68$ on the Lamem dataset, and Bylinskii *et al.* [4] measure rank correlation of $\rho = 0.74$ on the Figrim dataset.

One problem with these results is that they ignore the standard deviation of the correlation when computed over different splits into groups. That is, how stable are the rank correlations between groups? As it turns out, [4] report quite a large variability across splits, i.e., $\sigma = 0.2$ on Figrim [4]. This raises questions regarding the use of memorability scores for prediction, since falling within the variance should be considered as success.

We therefore aim at studying the best group size needed for consistent image memorability scores. To do it, we repeated the Memory Game, as described in Section 2, using target and filler images randomly selected from Figrim. We evaluate human consistency across different group sizes as follows. We measured memorability for 45 target images, each scored by 275 participants on average. The participants were split into two equal-size groups and the mean score per image was computed for each group. We then computed the Spearmans rank correlation between the scores of the two groups. This was done for 100 random splits into groups. We then computed both the mean and the variance of the correlation scores over all splits.

Figure 6 shows our results. For groups of 40 participants, the consistency is $\rho = 0.74$ with standard deviation of $\sigma = 0.12$ (compared to 0.74, 0.2 respectively, reported in [4]). For groups of 100 participants, the consistency significantly increases to 0.86 ($\sigma = 0.07$), while for groups of 135 participants, it only slightly increases further to 0.88 ($\sigma = 0.05$).

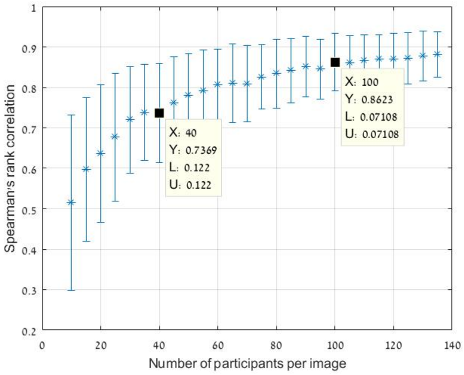


Figure 6: **Memorability consistency across groups of participans.** Using groups of 100 participants is a good compromise between the accuracy of human consistency and the complexity of collecting the data.

We conclude that assigning memorability scores should be performed by averaging over groups bigger than 40 participants. Since collecting memorability measurements for large numbers of participants requires a great deal of work, we recommend using 100 observers, which seems like a good compromise between the consistency and the complexity of collecting the data.

2. How consistent are the scores? Should the variance be considered? Having noted that the scores consistency varies, we further ask ourselves whether the mean scores, which are used by the existing datasets, have the same meaning for all images. That is, we question whether representing an image by its mean memorability score suffices, as maybe the variance per image should also be considered.

To answer this, we measured the variance of the memorability scores given by different groups of people, per image. This was done for 45 images and for group sizes between 40 and 130. Figure 7 shows our results. Every point in the graph represents a different image. The main conclusion from this graph is that the variance is not fixed. Some images are highly memorable by most people, while others are memorable by some and not so much by others. This suggests that one may want to represent image memorability using two numbers, the mean memorability score and the variance of the scores.

However, a second conclusion from Figure 7, is that the larger the number of participants, the smaller the variance across groups. Therefore, if the number of observers is sufficiently large, it may suffice to maintain a single number—the memorability score—which is the common practice. This supports our previous claim that a larger number of observers per image could lead to more stable scores, with higher consistency.

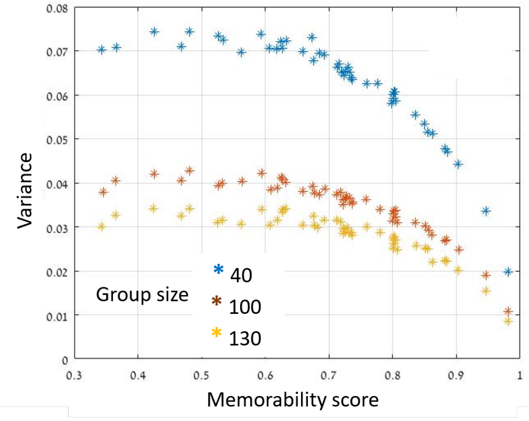


Figure 7: **Memorability variance varies across images.** Some images are highly memorable by all, while for other images there is high variance. The variance decreases when computing scores over larger participant group size.

3. Should the order in which the images are displayed be taken into account? In real life, images are always displayed in some context; for example, when looking at the newspaper or walking in a museum, the images intentionally appear in a certain order. In the street, the order of images is not intentional, but it definitely influences the scenes and objects we will remember or forget. Image memorability, however, is commonly measured across random sequences of images in order to isolate extrinsic effects such as the order of viewing. Is that the right practice? Should an image be assigned a single memorability score?

To assess the effect of image order on memorability, we designed a new version of the "Memory Game", in which the only difference from the original memory game is that rather than randomly creating image sequences, we used fixed sets of sequences. For a fixed set of target and filler images (120 altogether, randomly chosen from Figrim), we created 5 orders.

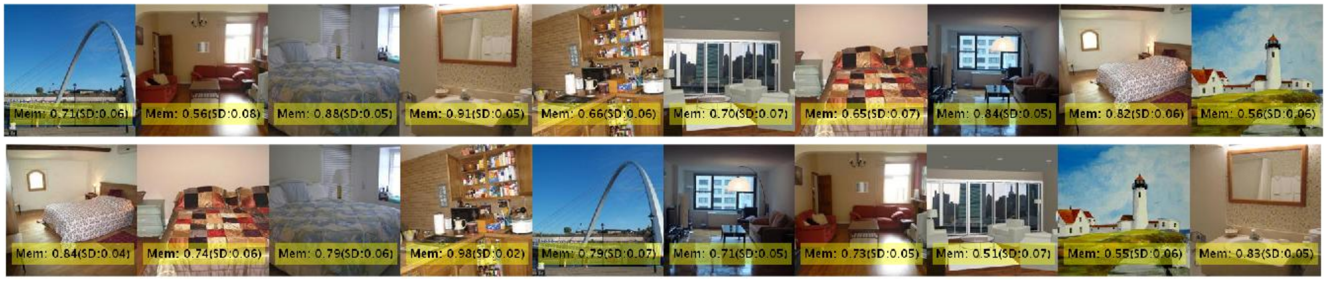


Figure 8: **Display order affects image memorability.** This figure shows the target images from two different orders of the same set of images (target and fillers), along with their memorability scores (Mem) and the standard deviation (SD). Depending on the context, the memorability score can dramatically change, e.g., the kitchen. Conversely, the lighthouse, which is unique in both sequences, gets the same memorability score.

We measured the correlation between different groups of observers. When shown sets of different order, the correlation was low $\rho = 0.4$, $\sigma = 0.2$, while when shown sets of the the same order, the correlation was high $\rho = 0.7$, $\sigma = 0.1$. An example of why this happens is displayed in Figure 8. Take for instance, the kitchen scene. When presented after different scenes (top), its memorability score is 0.66, however, when presented after a sequence of bedrooms (bottom), its memorability score jumped to 0.98.

This suggests that the order of images in the sequence should not be overlooked. The current practice of taking random orders and averaging might alleviates the influence of this extrinsic effect, but it does not consider it fully. In fact, it averages sequences where the image is memorable with sequences where it is not, making the mean scores an inaccurate representation. We conclude that a single score per image is not a sufficient representation.

5. Conclusion

This paper has studied the relation between convolutional neural networks for image classification and memorability prediction. It introduced *MemBoost*, a network that reaches the limit of human performance on the largest existing dataset for image memorability, *LaMem*.

MemBoost is based on three key observations: (i) As object classification CNNs improve, so does image memorability prediction. (ii) Scene classification plays a bigger role in memorability prediction than object classification, but their combination is preferable. (iii) It suffices to train a regression layer on top of a CNN for object & scene recognition to achieve on par results with those attained by re-training the entire CNN. These observations were examined one-by-one via extensive experiments and were thoroughly analyzed.

Since our network already achieves human performance on LaMem, the next stage in memorability prediction is to produce a larger, more challenging dataset. When doing so, various factors should be re-considered. We provide some guidelines for designing the next-generation memorability dataset. These guidelines regard the number of observers, the validity of maintaining one score per image, and the need to re-think the order of images. In the future, more factors may be studied.

acknowledgements: We gratefully acknowledge the support of the Israel Science Foundation (ISF) 1083/18 and the Ollendorff Foundation

A. Dataset details

Tables 5-6 provide the details on the datasets used throughout the paper.

	Memorability Datasets		
Properties	**LaMem [14]**	**SUN-Mem [13]**	**Figrim [4]**
Data type	Objects & scenes	397 Scenes	21 scenes
# Images	58,741	2,222	1,754
Mean memorability	75.6±12.4	67.5±13.6	66±13.9
Human consistency	0.68	0.75	0.74

Table 5: **Memorability datasets.** The memorability scores are the mean and standard deviation over the entire dataset. The consistency values are the average of the Spearman Rank Correlation between different groups of observers.

Training Datasets	**Data type**
ImageNet [7]	1000 objects
Places205 [27]	205 scenes
Places365 [26]	365 scenes
Hybrid1205 [27]	1000 objects + 205 scenes
Hybrid1365 [26]	1000 objects + 365 scenes

Table 6: **Training datasets of images** of objects, of scenes and of both. We note that Hybrid1205(365) contains ImageNet and Places205(365).

References

[1] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323, 2013. 1

[2] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013. 1

[3] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008. 1, 2

[4] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015. 2, 3, 6, 7, 8

[5] Bora Celikkale, Aykut Erdem, and Erkut Erdem. Predicting memorability of images using attention-driven spatial pooling and image semantics. *Image and Vision Computing*, 42:35–46, 2015. 2

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016. 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 3, 8

[8] Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. What makes an object memorable? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1089–1097, 2015. 1, 2

[9] Timothy F Brady Talia Konkle George and A Alvarez Aude Oliva. Are real-world objects represented as bound units? independent decay of object details from short-term to long-term memory. *Journal of Vision*, 10(12):1–11. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[11] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2011. 1, 2, 3

[12] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482, 2014. 1, 2, 3

[13] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 145–152. IEEE, 2011. 1, 2, 3, 6, 8

[14] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2390–2398, 2015. 1, 2, 3, 4, 5, 6, 8

[15] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Memorability of image regions. In *NIPS*, volume 2, page 4, 2012. 2, 3

[16] Jongpil Kim, Sejong Yoon, and Vladimir Pavlovic. Relative spatial features for image memorability. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 761–764. ACM, 2013. 1, 2

[17] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3):558, 2010. 2

[18] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. Scene memory is more detailed than you think the role of categories in visual long-term memory. *Psychological Science*, 21(11):1551–1556, 2010. 1, 2

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[20] Matei Mancas and Olivier Le Meur. Memorability of natural scenes: The role of attention. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 196–200. IEEE, 2013. 1, 2

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[23] Lionel Standing. Learning 10000 pictures. *The Quarterly journal of experimental psychology*, 25(2):207–222, 1973. 1

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 3

[25] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010. 3

[26] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3, 4, 8

[27] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 3, 8