

Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability

Anelise Newman*, Camilo Fosco*, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva

Massachusetts Institute of Technology

Abstract. A key capability of an intelligent system is deciding when events from past experience must be remembered and when they can be forgotten. Towards this goal, we develop a predictive model of human visual event memory and how those memories decay over time. We introduce *Memento10k*, a new, dynamic video memorability dataset containing human annotations at different viewing delays. Based on our findings we propose a new mathematical formulation of memorability decay, resulting in a model that is able to produce the first quantitative estimation of how a video decays in memory over time. In contrast with previous work, our model can predict the probability that a video will be remembered at an arbitrary delay. Importantly, our approach combines visual and semantic information (in the form of textual captions) to fully represent the meaning of events. Our experiments on two video memorability benchmarks, including Memento10k, show that our model significantly improves upon the best prior approach (by 12% on average).

Keywords: Memorability estimation, memorability decay, multimodal video understanding.

1 Introduction

Deciding which moments from past experience to remember and which ones to discard is a key capability of an intelligent system. The human brain is optimized to remember what it deems to be important and forget what is uninteresting or redundant. Thus, human memorability is a useful measure of what content is interesting and likely to be retained by a human viewer. If a system can predict which information will be highly memorable, it can evaluate the utility of incoming data and compress or discard what is deemed to be irrelevant. It can also filter to select the content that will be most memorable to humans, which has potential applications in design and education.

However, memorability of dynamic events is challenging to predict because it depends on many factors. First, different visual representations are forgotten at different rates: while some events persist in memory even over long periods, others are forgotten within minutes [3,22,25,34]. This means that the probability

* Equal contribution.

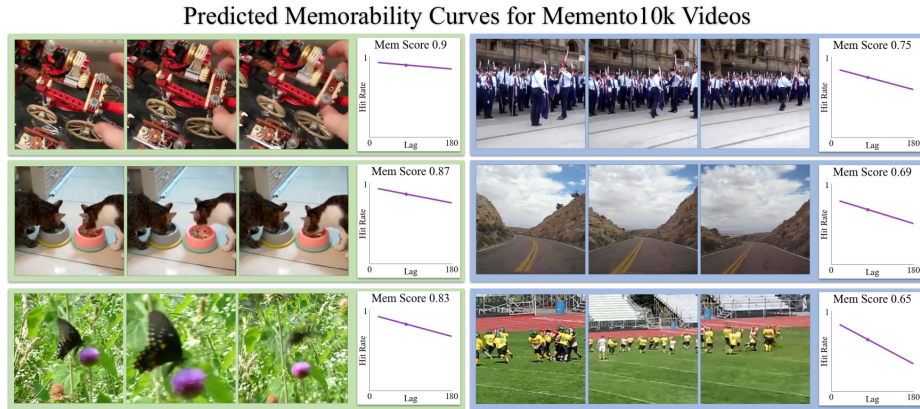


Fig.1. How do visual and semantic features impact memory decay over time? We introduce a multimodal model, SemanticMemNet, that leverages visual and textual information to predict the memorability decay curve of a video clip. We show predictions on videos from Memento10k, our new video memorability dataset. SemanticMemNet is the first model to predict a decay curve that represents how quickly a video falls off in memory over time.

that someone will remember a certain event varies dramatically as a function of time, introducing challenges in terms of how memorability is represented and measured. Second, memorability depends on both visual and semantic factors. In human cognition, language and vision often act in concert for remembering an event. Events described with richer and more distinctive concepts are remembered for longer than events attached to shallower descriptions, and certain semantic categories of objects or places are more memorable than others [25,33].

In this paper, we introduce a new dataset and a model for video memorability prediction that address these challenges¹. Memento10k, the most dynamic video memorability dataset to date, contains both human annotations at different viewing delays and human-written captions, making it ideal for studying the effects of delay and semantics on memorability. Based on this data, we propose a mathematical formulation of memorability decay that allows us to estimate the probability that an event will be remembered at any point over the first ten minutes after viewing. We introduce SemanticMemNet, a multimodal model that relies on visual and semantic features to predict the decay in memory of a short video clip (Fig. 1). SemanticMemNet is the first model that predicts the entire memorability decay curve, which allows us to estimate the probability that a person will recall a given video after a certain delay. We also enhance our model’s features to include information about video semantics by jointly predicting verbal captions. SemanticMemNet achieves state-of-the-art performance on video memorability prediction on two different video memorability baselines.

¹ Dataset, code, and models can be found at our website: <http://memento.csail.mit.edu/>

To summarize, our key contributions are:

- We present a new **multi-temporal, multimodal memory dataset** of 10,000 video clips, *Memento10k*. With over 900,000 human memory annotations at different delay intervals and 50,000 captions describing the events, it is the largest repository of dynamic visual memory data.
- We propose a **new mathematical formulation for memorability decay** which estimates the decay curve of human memory for individual videos.
- We introduce **SemanticMemNet**, a model capitalizing on learning visual and semantic features to predict both the memorability strength and the memory decay curve of a video clip.

2 Related Work

Memorability in Cognitive Science. Four landmark results in cognitive science inspired our current approach. First, memorability is an *intrinsic* property of an image: people are remarkably consistent in which images they remember and forget [3,11,22,25,26,27,30,36]. Second, semantic features, such as a detailed conceptual representation or a verbal description, boost visual memory and aid in predicting memorability [33,35,44,45]. Third, most stimuli decay in memory, with memory performances falling off predictably over time [22,25,34,43]. Finally, the classical old-new recognition paradigm (i.e. people press a key when they recognize a repeated stimulus in a sequence) allows researchers to collect objective measurements of human memory at a large scale [9,26] and variable time scales, which we draw on to collect our dataset.

Memorability in Computer Vision. The intrinsic nature of visual memorability means that visual stimuli themselves contain visual and semantic features that can be captured by machine vision. For instance, earlier works [24,25,26] pointed to content of images that were predictive of their memorability (i.e. people, animals and manipulable objects are memorable, but landscapes are often forgettable). Later work replicated the initial findings and extended the memorability prediction task to many photo categories [1,5,17,19,31,38,48], faces [3,29,40], visualizations [7,8] and videos [13,15,39].

The development of large-scale image datasets augmented with memorability scores [30] allowed convolutional neural networks to predict image memorability at near human-level consistency [1,5,19,30,48] and even generate realistic memorable and forgettable photos [21,40]. However, similar large-scale work on video memorability prediction has been limited. In the past year, Cohendet et al. introduced a video-based memorability dataset [14] that is the only other large benchmark comparable to Memento10k, and made progress towards building a predictive model of video memorability [13,15]. Other works on video memorability have largely relied on smaller datasets collected using paradigms that are more challenging to scale. [23] collected memorability and fMRI data on 2400 video clips and aligned audio-visual features with brain data to improve prediction. [39] used a language-based recall task to collect memorability scores as a function of response time for 100 videos. They find that a combination of

semantic, spatio-temporal, saliency, and color features can be used to predict memorability scores. [15] collected a long-term memorability dataset using clips from popular movies that people have seen before.

Past work has confirmed the usefulness of semantic features for predicting memorability [13,41,39]. Work at the intersection of Computer Vision and NLP has aimed to bridge the gap between images and text by generating natural language descriptions of images using encoder-decoder architectures (e.g. [28,42,47]) or by creating aligned embeddings for visual and textual content [18,20,32]. Here, we experiment with both approaches in order to create a memorability model that jointly learns to understand visual features and textual labels.

3 Memento10k: A Multimodal Memorability Dataset

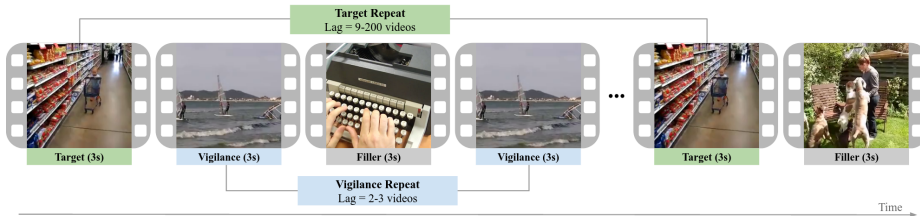


Fig. 2. Task flow diagram of *The Memento Video Memory Game*. Participants see a continuous stream of videos and press the space bar when they see a repeat.

Memento10k focuses on both the visual and semantic underpinnings of video memorability. The dataset contains 10,000 video clips augmented with memory scores, action labels, and textual descriptions (five human-generated captions per video). Importantly, our memorability annotations occur at presentation delays ranging from several seconds to ten minutes, which, for the first time, allows us to model how memorability falls off with time.

The Memento Video Memory Game. In our experiment, crowdworkers from Amazon’s Mechanical Turk (AMT) watched a continuous stream of three-second video clips and were asked to press the space bar when they saw a repeated video. Importantly, we varied the lag for repeated videos (the number of videos between the first and the second showing of a repeat), which allowed us to study the evolution of memorability over time. Our policies came from [26,30] and a summary of paradigm details can be found in the supplement.

Three seconds, the length of a Memento clip, is about the average duration of human working memory [2,4] and most human visual memory performances plateau at three seconds of exposure [9,10]. This makes three seconds a good “atomic” length for a single item held in memory. Previous work has shown that machine learning models can learn robust features even from short clips [37].

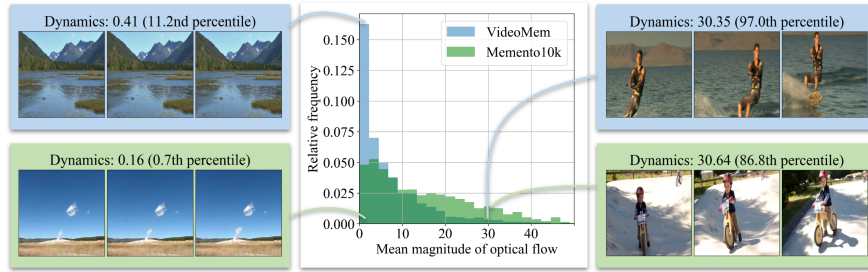


Fig.3. Video distribution by level of motion for VideoMem and Memento10k. The motion metric for each video is calculated by averaging optical flow magnitude over the entire video. A high percentage of videos in the VideoMem dataset are almost static, while Memento10k is much more balanced. We show select examples with low and high levels of motion.

Dynamic, In-The-Wild Videos. Memento10k is composed of natural videos scraped from the Internet². To limit our clips to non-artificial scenes with everyday context, we asked crowdworkers whether each clip was a “home video” and discarded videos that did not meet this criterion. After removing clips that contained undesirable properties (i.e. watermarks), we were left with 10,000 “clean” videos, which we break into train (7000), validation (1500), and test (1500) sets.

The Memento10k dataset is a significant step towards understanding memorability of real-world events. First, it is the most dynamic memorability dataset to date with videos containing a variety of motion patterns including camera motion and moving objects. The mean magnitude of optical flow in Memento10k is nearly double that of VideoMem [13] (approximately 15.476 for Memento vs. 7.296 for VideoMem, see Fig. 3), whose clips tend to be fairly static. Second, Memento10k’s diverse, natural content enables the study of memorability in an everyday context: Memento10k was compiled from in-the-wild amateur videos like those found on social media or real-life scenes, while VideoMem is composed of professional footage. Third, Memento10k’s greater number of annotations (90 versus 38 per video in VideoMem), spread over lags of 30 seconds to 10 minutes, leads to higher ground-truth human consistency and allows for a robust estimation of a video’s decay rate. Finally, we provide more semantic information such as action labels as well as 5 detailed captions per video. In this paper, we use both benchmarks to evaluate the generalization of our model.

Semantic Annotations. We augment our dataset with captions, providing a source of rich textual data that we can use to relate memorability to semantic concepts (examples in Fig. 4). We asked crowdworkers to describe the events in the video clip in full sentences and we manually vetted the captions for quality and corrected spelling mistakes. Each video has 5 unique captions from different crowdworkers. More details on caption collection are in the supplement.

² The Memento videos have partial overlap with the Moments in Time [37] dataset.

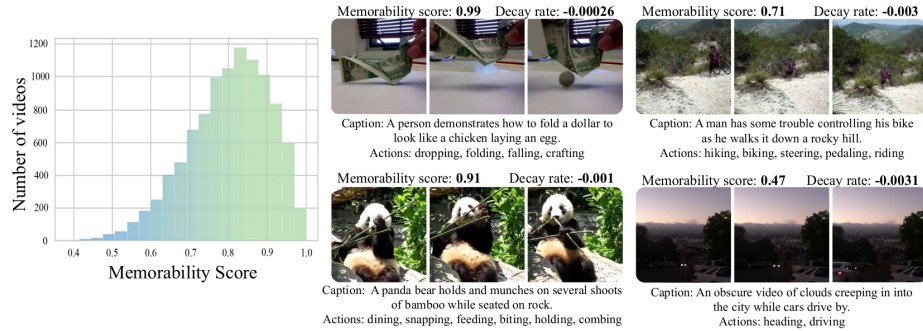


Fig. 4. The Memento10k dataset contains the memorability scores, alpha scores (decay rates), action labels, and five unique captions for 10,000 videos. **Left:** The distribution of memorability scores over the entire dataset. **Right:** Example clips from the Memento10k dataset along with their memorability score, decay rate, actions, and an example caption.

Human Results. The Memento10k dataset contains over 900,000 individual annotations, making it the biggest memorability dataset to date. We measured human consistency of these annotations following [13,26]: we randomly split our participant pool into two groups and calculate the Spearman’s rank correlation between the memorability rankings produced by each group, where the rankings are generated by sorting videos by raw hit rate. The average rank correlation (ρ) over 25 random splits is 0.73 (compared to 0.68 for images in [30], and 0.616 for videos in [13]). This high consistency between human observers confirms that videos have strong intrinsic visual, dynamic or semantic features that a model can learn from to predict memorability of new videos.

Fig. 5 illustrates some qualitative results of our experiment. We see some similar patterns as with image memorability: memorable videos tend to contain saturated colors, people and faces, manipulable objects, and man-made spaces, while less memorable videos are dark, cluttered, or inanimate. Additionally, videos with interesting motion patterns can be highly memorable whereas static videos often have low memorability.

4 Memory Decay: A Theoretical Formulation

Most memories decay over time. In psychology this is known as the forgetting curve, which estimates how the memory of an item naturally degrades. Because Memento10k’s memorability annotations occur at lags of anywhere from 9 videos (less than 30 seconds) to 200 videos (around 9 minutes), we have the opportunity to calculate the strength of a given video clip’s memory at different lags.

A naive method for calculating a memorability score is to simply take the video’s target hit rate, or the fraction of times that the repeated video was correctly detected. However, since we expect a video’s hit rate to go down with



Fig. 5. Examples of high- and low-memorability video clips. Videos involving people, faces, hands, man-made spaces, and moving objects are in general more memorable, while clips containing distant/outdoor landscapes or dark, cluttered, or static content tend to be less memorable.

time, annotations at different lags are not directly comparable. Instead, we derive an equation for how each video’s hit rate declines as a function of lag.

First, for the lags tested in our study, we observe that hit rate decays linearly as a function of lag. This is notable because previous work on image memorability has found that images follow a log-linear decay curve [22,25,43]. Fig. 6 (left) shows that a linear trend best fits our raw annotations; this holds for videos across the memorability distribution (see the supplement for more details).

Second, in contrast to prior work, we find that different videos decay in memory at different rates. Instead of assuming that all stimuli decay at one universal decay rate, α , as in [30], we assume that each video decays at its own rate, $\alpha^{(v)}$. Following the procedure laid out in [30], we find a memorability score and decay rate for each video that approximates our annotations. We define the memorability of video v as $m_T^{(v)} = \alpha^{(v)}T + c^{(v)}$, where T is the lag (the interval in videos between the first and second presentation) and $c^{(v)}$ is the base memorability of the video. If we know $m_T^{(v)}$ and $\alpha^{(v)}$, we can then calculate the video’s memorability at a different lag t with $m_t^{(v)} = m_T^{(v)} + \alpha^{(v)}(t - T)$ (1).

To obtain values for $m_T^{(v)}$ and $\alpha^{(v)}$, we minimize the L2 norm between the raw binary annotations from our experiment $x_j^{(v)}$, $j \in \{0, \dots, n^{(v)}\}$ and the predicted memorability score at the corresponding lag, $m_t^{(v)}$. The error equation is:

$$E(\alpha^{(v)}, m_T^{(v)}) = \sum_{j=1}^{n^{(v)}} \|x_j^{(v)} - m_t^{(v)}\|_2^2 = \sum_{j=1}^{n^{(v)}} \|x_j^{(v)} - [m_T^{(v)} + \alpha^{(v)}(t_j^{(v)} - T)]\|_2^2$$

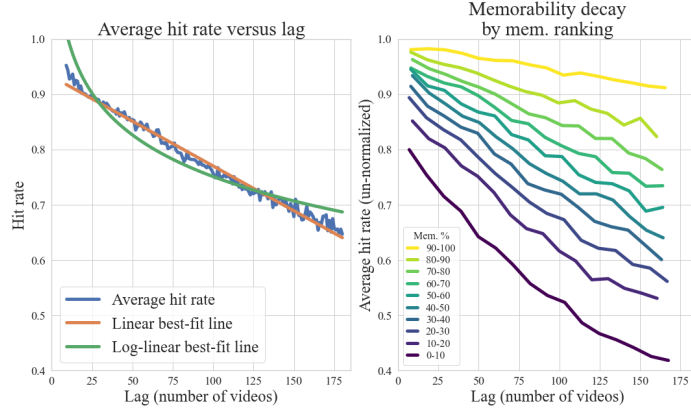


Fig. 6. Our data suggests a **memory model where each video decays linearly** in memory according to an individual decay rate $\alpha^{(v)}$. **Left:** A linear trend is a better approximation for our raw data ($r = -0.991$) than a log-linear trend ($r = -0.953$). **Right:** We confirm our assumption that α varies by video by grouping videos into deciles based on their normalized memorability score and plotting group average hit rate as a function of lag. Videos with lower memorability show a faster rate of decay.

We find update equations for $m_T^{(v)}$ and $\alpha^{(v)}$ by taking the derivative with respect to each and setting it to zero:

$$\alpha^{(v)} \leftarrow \frac{\frac{1}{n^{(v)}} \sum_{j=1}^{n^{(v)}} (t_j^{(v)} - T) [x_j^{(v)} - m_T^{(v)}]}{\frac{1}{n^{(v)}} \sum_{j=1}^{n^{(v)}} [t_j^{(v)} - T]^2} \quad m_T^{(v)} \leftarrow \frac{1}{n^{(v)}} \sum_{j=1}^{n^{(v)}} [x_j^{(v)} - \alpha^{(v)}(t_j^{(v)} - T)] \quad (2)$$

We initialize $\alpha^{(v)}$ to $-5e-4$ and $m_T^{(v)}$ to each video’s mean hit rate. We set our base lag T to 80 and optimize for 10 iterations to produce $\alpha^{(v)}$ and $m_{80}^{(v)}$ for each video. We thus define a video’s “memorability score”, for the purposes of memorability ranking, as its hit rate at a lag of 80; however, we can use equation 1 to calculate its hit rate at an arbitrary lag within the range that we studied.

Next, we validate our hypothesis that videos decay in memory at different rates. We bucket the Memento10k videos into 10 groups based on their normalized memorability scores and plot the raw data (average hit rate as a function of lag) for each group. Fig. 6 (right) confirms that different videos decay at different rates.

5 Modeling Experiments

In this section, we explore different architecture choices for modeling video memorability that take into consideration both visual and semantic features. We also move away from a conception of memorability as a single value, and instead predict the *decay curve* of a video, resulting in the first model that predicts the raw probability that a video will be remembered at a particular lag.

	Approach	RC - Memento10k (test set)	RC - VideoMem (validation set)
	Human consistency	0.730	0.616
(Sec. 5.1)	Flow stream only	0.579	0.425
	Frames stream only	0.595	0.527
	Video stream only	0.596	0.492
	Flow + Frames + Video	0.659	0.555
(Sec. 5.2)	Video stream + captions	0.602	0.512
	Video stream + triplet loss	0.599	-
	SemanticMemNet (ours)	0.663	0.556

Table 1. SemanticMemNet ablation study. We experiment with different ways of incorporating visual and semantic features into memorability prediction. We measure performance by calculating the Spearman’s rank correlation (RC) of the predicted memorability rankings with ground truth rankings on both Memento10k and VideoMem.

Metrics. A commonly-used metric for evaluating memorability is the Spearman rank correlation (RC) between the memorability ranking produced by the ground-truth memorability scores versus the predicted scores. This is a popular metric [13,26,30] because memorability rankings are generally robust across experimental designs and choice of lag; however, it does not measure the accuracy of predicted memorability values. As such, when evaluating the quality of the decay curve produced by our model, we will also consider R^2 for our predictions.

5.1 Modeling Visual Features

Baseline: Static Frames. We evaluate the extent to which static visual features contribute to video memorability by training a network to predict a video’s memorability from a single frame. We first train an ImageNet-pretrained DenseNet-121 to predict image memorability by training on the LaMem dataset [30], then finetune on the video memorability datasets. At test time, a video’s memorability score is calculated by averaging predictions over every 4th frame (about 22 frames total for Memento videos).

3D Architectures: Video and Optical Flow. Training on the RGB videos allows the network to access information on both motion and visual features, while training on optical flow lets us isolate the effects of motion. We train I3D architectures [12] on raw video and optical flow (computed using OpenCV’s TV-L1 implementation). Our models were pretrained on the ImageNet and Kinetics datasets. We test the different visual feature architectures on both Memento10k and VideoMem videos. Our results are in the top section of Table 1. Out of the three input representations (frames, flow, and video), optical flow achieves the poorest performance, probably because of the lack of access to explicit visual features like color and shape. Static frames perform remarkably well on their own, outperforming a 3D-video representation on VideoMem (as VideoMem is a fairly static dataset, this result is reasonable). Even with the relatively high level of motion in the Memento10k dataset (see Fig. 3), the video and frames streams

perform comparably. For both datasets, combining the three streams maximizes performance, which is consistent with previous work [12] and reinforces that both visual appearance and motion are relevant to predicting video memorability. In fact, our three-stream approach leverages motion information from the optical flow stream to refine its predictions, as shown in Fig. 7.

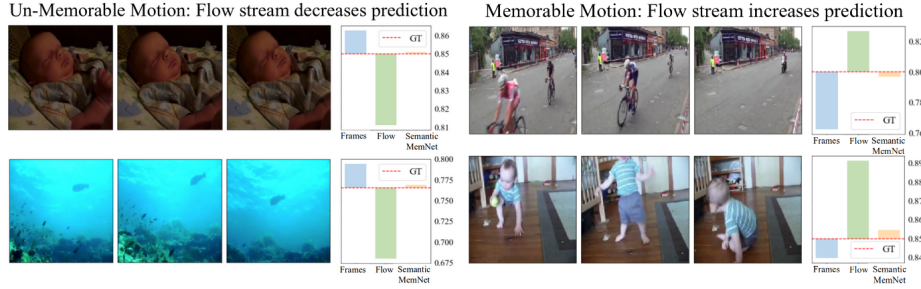


Fig. 7. Our model leverages visual and motion information to produce accurate memorability scores. Here, we compare the contributions of the frames stream (static information only) and the optical flow stream (motion information only) to separate the contributions of visual features and temporal dynamics. **Left:** Flow decreases the frames prediction. The frames stream detects memorable features like a human face or saturated colors, but the flow stream detects a static video and predicts lower memorability. **Right:** Flow increases the frames prediction. The flow stream picks up on dynamic patterns like fast bikers or a baby falling and increases the memorability prediction.

5.2 Modeling Semantic Features

It is well-known that semantics are an important contributor to memorability [13,15,33,39]. To increase our model’s ability to extract semantic information, we jointly train it on memorability prediction and a captioning task, which ensures that the underlying representation learned for both problems contains relevant event semantics. To test this approach, we enhance the video stream from the previous section with an additional module that aims to solve one of two tasks: generating captions or learning a joint text-video embedding. We augment the video stream as opposed to the flow or frames streams because it contains both visual and motion information required to reconstruct a video caption.

Caption Generation. Our first approach is to predict captions directly. This has the benefit of forcing the model to encode a rich semantic representation of the image, taking into account multiple objects, actions, and colors. However, it also involves learning an English language model, which is tangential to the task of predicting memorability. We feed the output features of the video I3D base into an LSTM that learns to predict the ground-truth captions. For Memento10k, we tokenize our 5 ground-truth captions and create a vocabulary of 3,870 words

that each appear at least 5 times in our training set. We use pre-trained FastText word embeddings [6] to map our tokens to 300-dimensional feature vectors that we feed to the recurrent module. The VideoMem dataset provides a single brief description with each video, which we process the same way. We train with teacher forcing and at test-time feed the output of the LSTM back into itself.

Mapping Videos into a Semantic Embedding. Our second approach is to learn to map videos into a sentence-level semantic embedding space using a triplet loss. We pre-compute sentence embeddings for the captions using the popular transformer-based network BERT [16].³ At training time, we stack a fully-connected layer on top of the visual encoder’s output features and use a triplet loss with squared distance to ensure that the embedded representation is closer to the matching caption than a randomly selected one from our dataset. This approach has the benefit that our network does not have to learn a language model, but it may not pick up on fine-grained semantic actors in the video.

Captioning Results. The results of our captioning experiments are in the second section of Table 1. Caption generation outperforms the semantic embedding approach on Memento10k. Learning to generate captions provides a boost over only the video stream for both datasets. Fig. 9 contains examples of captions generated by our model.

5.3 Modeling Memorability Decay

Up until this point, we have evaluated our memorability predictions by converting them to rankings and comparing them to the ground truth. However, the Memento10k data and our parameterization of the memorability decay curve unlocks a richer representation of memorability, where $m_t^{(v)}$ is the true probability that an arbitrary person remembers video v at lag t . Thus, we also investigate techniques for predicting the ground-truth values of the memorability decay curve. Again, we consider two alternative architectures.

Mem- α Model. “Mem- α ” models produce two outputs by regressing to a video’s memory score and decay coefficient. To train these models, we define a loss that consists of uniformly sampling 100 values along the true and predicted memorability curves and calculating the Mean Absolute Error on the resulting pairs. Equation 1 can then be used to predict the raw hit rate at a different lag.

Recurrent Decay Model. This model directly outputs multiple probability values corresponding to different points on the decay curve. It works by injecting the feature vector produced by the video encoder into the hidden state of an 8-cell LSTM, where the cells represent evenly spaced lags from $t = 40$ through 180. At each time step, the LSTM modifies the encoded video representation, which is then fed into a multi-layer perceptron to generate the hit rate at that lag. The ground truth values used during training are calculated from $\alpha^{(v)}$ and $m_{80}^{(v)}$ using Equation 1.

³ Computed using [46]

Table 2. Multi-lag memorability prediction: Rank correlation (RC) and raw predictions (R^2) at 3 different lags (t , representing the number of intervening videos).

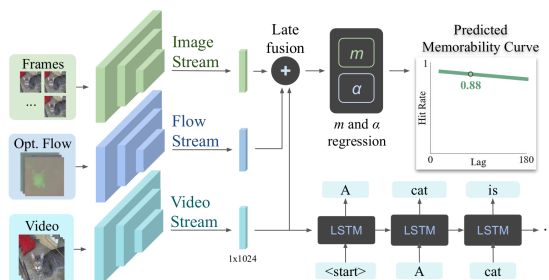
Approach	RC	R^2		
		$t=40$	$t=80$	$t=160$
Mem- α	0.604	0.146	0.227	0.121
Recur. head	0.599	0.298	0.364	0.219

Decay Results. We evaluate the models in two ways. First, we calculate rank correlation with ground truth, based on the memorability scores (defined as $m_{80}^{(v)}$). We also compare their raw predictions for different values of t , for which we report R^2 . The results are in Table 2.

We find that the Mem- α has better performance than the recurrent decay model in terms of rank correlation. However, the recurrent decay model outperforms the Mem- α model at predicting the raw memorability values, and exhibits good results for low and high lags as well. It makes sense that the performance of the Mem- α model falls off at lags further away from $T = 80$, since any error in the prediction of alpha (the slope of the decay curve) is amplified as we extrapolate away from the reference lag. These two models present a trade-off between simplicity and ranking accuracy (mem- α) and numerical accuracy along the entire decay curve (recurrent decay). Because of its relative simplicity and strong RC score, we use an Mem- α architecture for our final predictions.

6 Model results

Fig. 8. The architecture of SemanticMemNet. An I3D is jointly trained to predict memorability and semantic captions for an input video. Its memorability predictions are combined with a frames-based and optical flow stream to produce m_{80} and α , the parameters of the memorability decay curve.



SemanticMemNet (Fig. 8) combines our findings from the three previous sections. We use a three-stream encoder that operates on three different input representations: 1) the raw frames, 2) the entire video as a 3D unit, and 3) the 3D optical flow. We jointly train the video stream to output memorability scores and captions for the video. Each of our streams predicts both the memorability and the decay rate, which allows us to predict the probability that an observer will recall the video at an arbitrary lag within the range we studied.

To evaluate the effectiveness of our model, we compare against prior work in memorability prediction. MemNet [30] is a strong image memorability baseline;

we apply it to our videos by averaging its predictions over 7 frames uniformly sampled from the video. “Feature extraction and regression” is based on the approach from Shekhar et al. [39], where semantic, spatio-temporal, saliency, and color features are first extracted from the video and then a regression is learned to predict the memorability score. The final two baselines are the best-performing models from Cohendet et al. [13]. (Further details about baseline implementation can be found in the supplement.) The results of our evaluations are in Table 3. Our model achieves state-of-the-art performance on both Memento10k and VideoMem. Example predictions generated by our model are in Fig. 9.

Approach	RC - Memento10k (test set)	RC - VideoMem (validation set ⁴)
Human consistency	0.730	0.616
MemNet Baseline [30]	0.485	0.425
Feature extraction + regression (as in [39])*	0.615	0.427
Cohendet et al. (ResNet3D) [13]	0.574	0.508
Cohendet et al. (Semantic)[13]	0.552	0.503
SemanticMemNet	0.663	0.556

Table 3. Comparison to state-of-the-art on Memento10k and VideoMem. Our approach, SemanticMemNet, approaches human consistency and outperforms previous approaches. *Uses ground-truth captions at test-time.

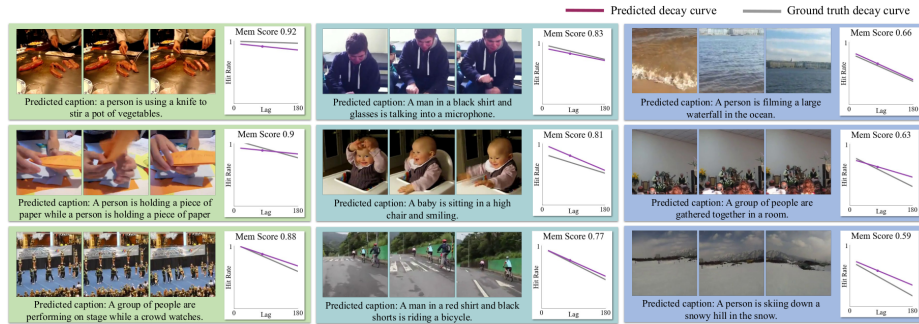


Fig. 9. Memorability and captions predictions from SemanticMemNet. For each example, we plot the predicted memorability decay curve based on SemanticMemNet’s values in purple, as well as the ground truth in gray.

⁴ We use the VideoMem validation set as the test set has not been made public.

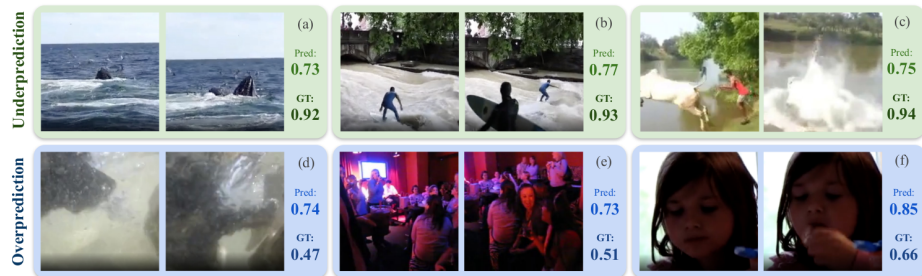


Fig. 10. Under and overpredictions of SemanticMemNet. Our network underestimates the memorability of visually bland scenes with a single distinctive element, like a whale sighting (a). It can fail on out-of-context scenes, like someone surfing on a flooded concrete river (b), or surprising events, like a man getting dragged into a lake by a cow (c). By contrast, it overestimates the memorability of choppy, dynamic scenes without clear semantic content (d) and of scenes that contain memorable elements, such as humans and faces, but that are overly cluttered (e), dark (f), or shaky.

7 Conclusion

Our contributions. We introduce a novel task (memorability decay estimation) and a new dynamic dataset with memorability scores at different delays. We propose a mathematical formulation for memorability decay and a model that takes advantage of it to estimate both memorability and decay.

Limitations and Future Work. Memorability is not a solved problem. Fig. 10 analyzes instances where our model fails because of competing visual attributes or complex semantics. Furthermore, there is still room for improvement in modeling memorability decay (Table 2) and extending our understanding of memorability to longer sequences. Our approach makes progress towards continuous memorability prediction for long videos (i.e. first-person live streams, YouTube videos) where memorability models should handle past events *and* their decay rates, to assess memorability of events at t different points in the past. To encourage exploration in this direction, we have released a live demo⁵ of SemanticMemNet that extracts memorable segments from longer video clips.

The utility of memorability. Video memorability models open the door to many exciting applications in computer vision. They can be used to provide guidance to designers, educators, and models to generate clips that will be durable in memory. They can improve summarization by selecting segments likely to be retained. They can act as a measure of the utility of different segments in space-constrained systems; for instance, a camera in a self-driving car or a pair of virtual assistant glasses could discard data once it has fallen below a certain memorability threshold. Predicting visual memory will lead to systems that make intelligent decisions about what information to delete, enhance, and preserve.

Acknowledgments. We thank Zoya Bylinskii and Phillip Isola for their useful discussions and Alex Lascelles and Mathew Monfort for helping with the dataset.

⁵ <http://demo.memento.csail.mit.edu/>

References

1. Akagunduz, E., Bors, A.G., Evans, K.K.: Defining Image Memorability using the Visual Memory Schema. *Pattern Analysis and Machine Intelligence*, IEEE Transactions on (2019)
2. Baddeley, A.: Working memory. *Science* **255**(5044), 556–559 (1992). <https://doi.org/10.1126/science.1736359>, <https://science.sciencemag.org/content/255/5044/556>
3. Bainbridge, W., Isola, P., Aude, O.: The Intrinsic Memorability of Face Photographs. *Journal of Experimental Psychology: General* **142**, 1323–1334 (2013)
4. Barrouillet, P., Bernardin, S., Camos, V.: Time constraints and resource sharing in adults’ working memory spans. *Journal of experimental psychology. General* **133**, 83–100 (04 2004). <https://doi.org/10.1037/0096-3445.133.1.83>
5. Baveye, Y., Cohendet, R., Perreira Da Silva, M., Le Callet, P.: Deep Learning for Image Memorability Prediction: The Emotional Bias. In: *Proceedings of the 24th ACM International Conference on Multimedia*. pp. 491–495. ACM (2016)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* (2016)
7. Borkin, A., Bylinskii, Z., Kim, N., Bainbridge, C., Yeh, C., Borkin, D., Pfister, H., Oliva, A.: Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics* **22**(1), 519–528 (2016)
8. Borkin, M., Vo, A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., Pfister, H.: What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* **19**(12), 2306–2315 (2013)
9. Brady, T.F., Konkle, T., Alvarez, G.A., Oliva, A.: Visual Long-Term Memory Has a Massive Storage Capacity for Object Details. *Proceedings of the National Academy of Sciences* **105**(38), 14325–14329 (2008)
10. Brady, T.F., Konkle, T., Gill, J., Oliva, A., Alvarez, G.A.: Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science* **24**(6), 981–990 (2013). <https://doi.org/10.1177/0956797612465439>, <https://doi.org/10.1177/0956797612465439>, PMID: 23630219
11. Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., Oliva, A.: Intrinsic and Extrinsic Effects on Image Memorability. *Vision Research* **116**, 165–178 (2015)
12. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. pp. 4724–4733 (2017). <https://doi.org/10.1109/CVPR.2017.502>, <https://doi.org/10.1109/CVPR.2017.502>
13. Cohendet, R., Demarty, C., Duong, N.Q.K., Martin, E.: VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2531–2540 (2019)
14. Cohendet, R., Demarty, C., Duong, N.Q.K., Sjöberg, M., Ionescu, B., Do, T., Rennes, F.: MediaEval 2018: Predicting Media Memorability Task. *CoRR abs/1807.01052* (2018), <http://arxiv.org/abs/1807.01052>
15. Cohendet, R., Yadati, K., Duong, N.Q., Demarty, C.H.: Annotating, Understanding, and Predicting Long-Term Video Memorability. In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. pp. 178–186. ACM (2018)

16. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), <http://arxiv.org/abs/1810.04805>
17. Dubey, R., Peterson, J., Khosla, A., Yang, M.H., Ghanem, B.: What Makes an Object Memorable? In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1089–1097 (2015)
18. Engilberge, M., Chevallier, L., Pérez, P., Cord, M.: Finding beans in burgers: Deep semantic-visual embedding with localization. CoRR **abs/1804.01720** (2018), <http://arxiv.org/abs/1804.01720>
19. Fajtl, J., Argyriou, V., Monekosso, D., Remagnino, P.: Amnet: Memorability Estimation with Attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6363–6372 (2018)
20. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 2121–2129. Curran Associates, Inc. (2013), <http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf>
21. Goetschalckx, L., Andonian, A., Oliva, A., Isola, P.: GANalyze: Towards Visual Definition of Cognitive Image Properties. In: IEEE International Conference on Computer Vision, ICCV 2019, Seoul, Korea. pp. 5744–5753 (2019)
22. Goetschalckx, L., Moors, P., Wagemans, J.: Image memorability across longer time intervals. *Memory* **26**, 581–588 (5 2017). <https://doi.org/10.1080/09658211.2017.1383435>
23. Han, J., Chen, C., Shao, L., Xintao, H., Jungong, H., Tianming, L.: Learning computational models of video memorability from fMRI brain imaging. *Cybernetics, IEEE Transactions on* **45**(8), 1692–1703 (2015)
24. Isola, P., Parikh, D., Torralba, A., Oliva, A.: Understanding the intrinsic memorability of images. In: Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. pp. 2429–2437 (2011)
25. Isola, P., Xiao, J., Parikh, D., Torralba, A., Oliva, A.: What Makes a Photograph Memorable? Pattern Analysis and Machine Intelligence, *IEEE Transactions on* **36**(7), 1469–1482 (2014)
26. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What Makes an Image Memorable? In: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011. pp. 145–152 (2011). <https://doi.org/10.1109/CVPR.2011.5995721>, <https://doi.org/10.1109/CVPR.2011.5995721>
27. Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., Rust, N.: Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife* **8**, e47596 (2019)
28. Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions. CoRR **abs/1412.2306** (2014), <http://arxiv.org/abs/1412.2306>
29. Khosla, A., Bainbridge, W., Torralba, A., Oliva, A.: Modifying the Memorability of Face Photographs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3200–3207 (2013)
30. Khosla, A., Raju, A.S., Torralba, A., Oliva, A.: Understanding and Predicting Image Memorability at a Large Scale. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2390–2398 (2015)

31. Khosla, A., Xiao, J., Torralba, A., Oliva, A.: Memorability of Image Regions. In: *Advances in Neural Information Processing Systems*. pp. 305–313 (2012)
32. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. *CoRR* **abs/1411.2539** (2014), <http://arxiv.org/abs/1411.2539>
33. Konkle, T., Brady, T., Alvarez, G., Oliva, A.: Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General* **139**, 558–578 (3 2010). <https://doi.org/10.1037/a0019359>
34. Konkle, T., Brady, T., Alvarez, G., Oliva, A.: Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychological Science* **21**, 1551–6 (10 2010). <https://doi.org/10.1177/0956797610385359>
35. Koutstaal, W., Reddy, C., Jackson, E., Prince, S., Cendan, D., Schacter, D.: False recognition of abstract versus common objects in older and younger adults: Testing the semantic categorization account. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **29**, 499–510 (2003)
36. Mohsenzadeh, Y., Mullin, C., Oliva, A., Pantazis, D.: The perceptual neural trace of memorable unseen scenes. *Scientific Reports* **8**, 6033 (2019)
37. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, Y., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence* (2019)
38. Perera, S., Tal, A., Zelnik-Manor, L.: Is Image Memorability Prediction Solved? In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019)
39. Shekhar, S., Singal, D., Singh, H., Kedia, M., Shetty, A.: Show and Recall: Learning What Makes Videos Memorable. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2730–2739 (2017)
40. Sidorov, O.: Changing the Image Memorability: From Basic Photo Editing to GANs. In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019)
41. Squalli-Houssaini, H., Duong, N., Gwenaëlle, M., Demarty, C.H.: Deep learning for predicting image memorability. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2018)
42. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence – video to text. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015)
43. Vö, M.L.H., Bylinskii, Z., Oliva, A.: Image Memorability In The Eye Of The Beholder: Tracking The Decay Of Visual Scene Representations. *bioRxiv* p. 141044 (2017)
44. Vogt, S., Magnussen, S.: Long-term memory for 400 pictures on a common theme. *Experimental Psychology* **54**, 298–303 (2007)
45. Wiseman, S., Neisser, U.: Perceptual organization as a determinant of visual recognition memory. *American Journal of Psychology* **87**, 675–681 (1974)
46. Xiao, H.: bert-as-service. <https://github.com/hanxiao/bert-as-service> (2018)
47. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. *CoRR* **abs/1502.03044** (2015), <http://arxiv.org/abs/1502.03044>
48. Zarezadeh, S., Rezaeian, M., Sadeghi, M.T.: Image Memorability Prediction Using Deep Features. In: *2017 Iranian Conference on Electrical Engineering (ICEE)*. pp. 2176–2181. IEEE (2017)