Supplemental - 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

1 Introduction

This supplementary material provides more details about our memorability dataset collection, the memorability trends we observe, and the baselines we use. We also provide additional model predictions.

2 Memorability experiments

2.1 Background: Measures of Human Memory

Whereas memorability can be measured using different paradigms, in this work we used the classical old new recognition paradigm. In this paradigm, people are shown a long stream of stimuli and asked to press a key whenever they recognize a particular stimuli (as in the large scale behavioral experiments of [1, 11]. This task has several advantages: 1) it allows to collect **objective measurements** of human memory, which are needed to quantify the results of neural networks and compare models (i.e. correlation rank); 2) it can **scale up** using crowdsourcing experiments; 3) it allows to quantify memory performances at **different time scales** (measured as the number of intervening stimuli between the first and second repetition of a stimulus). Importantly, normalizing the responses for time (as [10] did) is mandatory for **evaluating decay rate**, a novel feature of our work.

Other works on video memorability have used different memory paradigms. For instance, [8] collected neuroimaging data (using fMRI on a few human subjects) to augment model's learning; [12] used a semantic recall task where participants are asked to indicate which descriptions match a video they have seen before; [4] used a small dataset (660 videos, 11 annotations per video) of clips from common movies that participants have seen before, thus contaminating memorability measurements with previous memories. Importantly, none of these other tasks can scale to the near-million data points needed to train deep learning models for both memorability estimation and forgetting, nor allow to measure precisely the decay rate of each individual video as we pioneer here (more than two data points over time are needed to reliably estimate the slope of memory decay), making these results unsuitable to compare with the standard old-new recognition paradigm used in most computational memorability works.

2.2 Memento: The Video Memory Game

Our memorability experiment was based on the old-new recognition paradigms used in previous large scale experiments [9, 10]. In *Memento: The Video Memory Game*, 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. Unlike [9, 10], where each image is separated by a blank screen, our videos are shown back-to-back like a "movie trailer" to keep the pace engaging and game-like. When participants press the spacebar, they receive feedback in the form of a red flash for an incorrect response or a green flash for a correct response. The flow diagram of our task can be seen in Figure 2 of the main paper. When a worker correctly identifies a repeat, that is known as a "hit" and the stream skips ahead to the next video; there is no feedback for missed repeats.

Each level of the memory game contains on average 204 videos (with repeats) and lasts around 9 minutes. The number of intervening videos between the first and second occurrence of a repeated video is known as the "lag". The game consists of "vigilance" repeats that occur at short lags of 2-3 videos and are used to filter out inattentive workers and "target" repeats at lags of 9-200 videos that provide memorability data. In order to ensure high-quality annotations, we invalidate a level of the game if a participant's vigilance accuracy is below 80%, if their false positive rate is above 50%, or if the participant fails a quality check early in the level. These checks discarded around 15% of levels started.

Target repeats, which comprise around 20% of the video presentations in each level, form the core of our dataset. A target video's "hit rate" is the fraction of people who correctly identified it as a repeat. This wide range in target repeat lags allows us to measure how a video's hit rate changes as a function of lag. After an initial burn-in period, target repeats occur with an approximately uniform probability (25%) at each position. This uniformity is important so that participants cannot infer when the next repeat will come. All target repeats occur within one level of the game and videos are not reused across levels.

Each of our 10,000 videos has on average over 90 valid annotations per video. Responses were collected from 22,226 valid levels played by 4,967 players. The final quality-filtered dataset contains over 900,000 individual annotations, making it the biggest memorability dataset to date.

Caption collection. We used a two-stage process to collect captions for our videos. First, we collected a round of captions, hand-selected participants who produced good captions (detailed, grammatical, and accurate), and collected the remaining data from them. No participant captioned the same video multiple times. For all captions collected, we ensure that participants input captions with at least ten words to obtain detailed descriptions of the actors and actions in the video. We used an automatic spell-checker to correct common spelling and grammar errors. Then three researchers read through all the captions and eliminated those that did not describe the video, were vague, or contained serious grammar errors.

3 Linear vs. Log-Linear Trend

We performed an additional analysis to test how well a linear versus a log-linear trend fits high-memorability versus low-memorability videos. We sorted our videos by memorability score, broke them into ten groups, and fit both a linear and a log-linear curve to each group (in other words, we performed the same analysis as in Fig. 6 (left) for each of the curves plotted in Fig. 6 (right)). The results are in Table 3. For the lowest-memorability decile (lowest 10% of videos, dark purple curve in Fig.6 left), the R^2 values for both regressions were similar, with the value for the log-linear regression being a little higher (0.940 for the linear fit, 0.942 for the log-linear fig). For all other deciles, the linear regression was a better fit (see values below).[1] This shows that a linear trend is a good approximation for the memorability decay of videos across the memorability spectrum.

Decile	**Linear R^2**	**Log-Linear R^2**
10%	0.940	0.942
20%	0.943	0.916
30%	0.932	0.916
40%	0.922	0.876
50%	0.940	0.872
60%	0.932	0.855
70%	0.922	0.825
80%	0.939	0.842
90%	0.879	0.770
100%	0.781	0.653

4 Video Memorability Baselines

We compare Semantic MemNet against four models from prior work: the image memorability model MemNet [10], a ResNet3D, a feature-extraction pipeline followed by a regression inspired by [12], and the semantic embedding model from [3]. Here we provide more detail as to how we computed the latter two of these baselines.

ResNet3D. As in [3], we train a ResNet3D-34. We train it to map Memento videos to memorability scores.

Semantic embedding model. We retrain this model from scratch following the authors' procedure as outlined in [3, 7]. First, we train a SoDeep sorting module [7] to take in batches of scores in the range 0 to 1 and output the rank of each element in the batch (we use batch size 100). We use this module to define a ranking-based loss function, the Spearman's Rank Correlation between

[1] Regressions are calculated on values up to lag 180; for later lags, we do not have enough data to robustly estimate hit rate at each decile and each lag.

the ground-truth data and the predicted scores output by the model. For the backbone architecture, we take the visual stream of the semantic embedding model proposed in [6], which was pretrained on MS-COCO. We fine-tune this image-based network on the LaMem dataset [10] and finally on Memento10k by taking 7 evenly-sampled frames from each video and associating each with the memorability score of the overall video. This model's final score on the VideoMem dataset was taken from the paper [3].

Feature extraction and regression. [12] extracts several features from the input videos and builds a regression on top of those to generate a score for a video. The authors find that a combination of semantic, saliency, spatio-temporal, and color features produce the best results. We construct a similar baseline by extracting static features from the Memento10k and VideoMem videos and training an SVM on top of each feature set individually. We average the memorability scores predicted by each feature to get the final prediction. Here is a brief description of the features we computed

- *Semantic features:* We calculated the BERT [5] sentence embeddings for the ground-truth verbal descriptions of each video[2]. For Memento10k, we randomly selected one human-written caption per video and took its embedding. For VideoMem, we took the single embedding for the brief description provided of each video. Note that this feature is a function of ground-truth human annotations and as such is not automatic from pixels; however, it serves to capture the contribution of purely semantic content on memorability.
- *Saliency features:* We produced a saliency heatmap for 10 uniformly-sampled frames from each video. As in [12], we average the heatmaps for a given video and resize them to be 50 by 50 pixels.
- *Spatio-temporal/deep features:* We extract length-5120 deep features from a kinetics-pretrained I3D architecture [2].
- *Color features:* We follow the procedure described in [12].

We trained an SVM for each feature separately and then averaged the results across the four features to produce the result included in the main paper.

5 Additional Predictions

Figure 1 shows representative predictions from SemanticMemNet across the memorability spectrum.

[2] We calculated the BERT sentence embeddings using the code from [13]

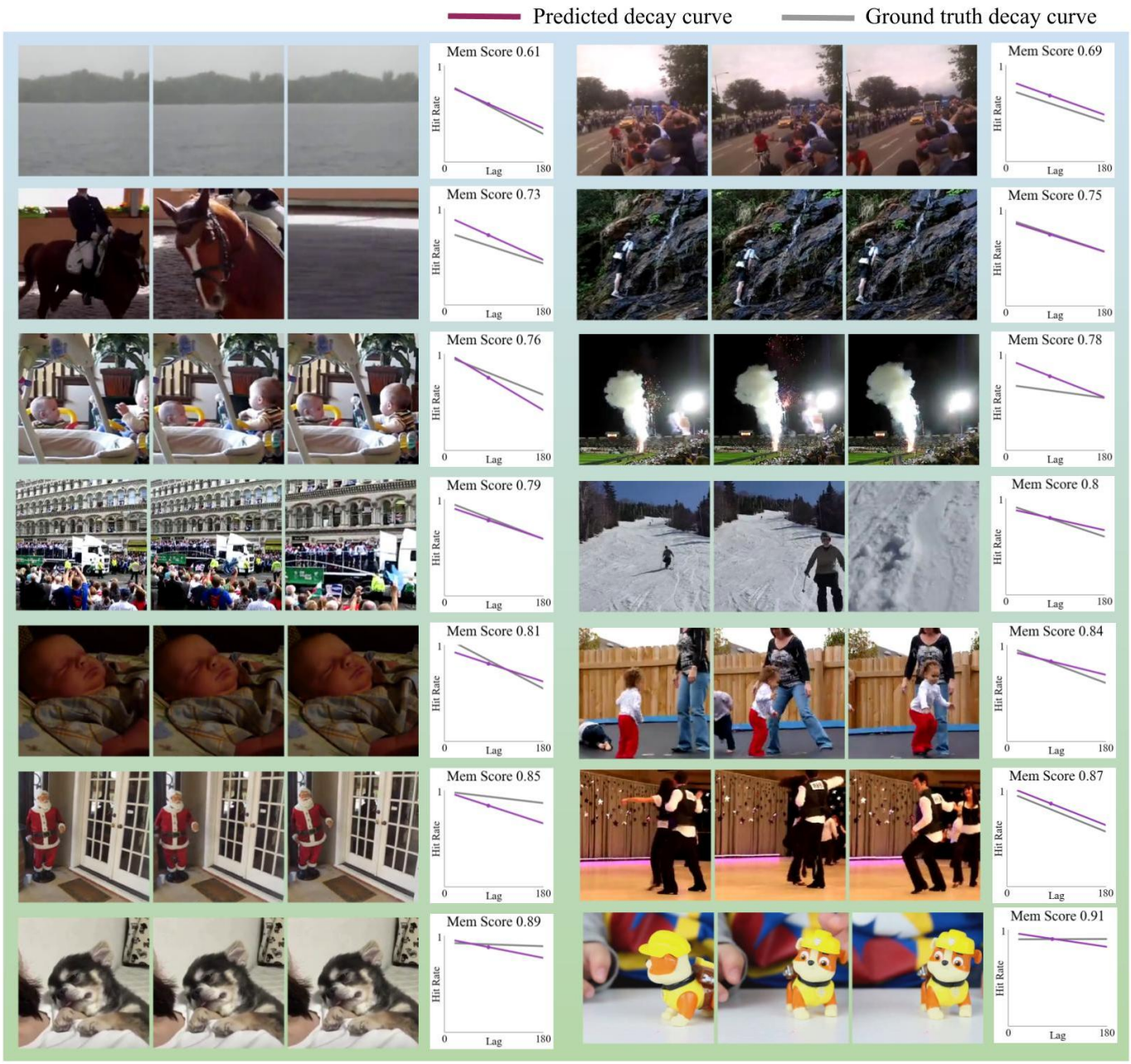

Fig. 1. Representative memorability predictions from SemanticMemNet on the Memento10k test set.

References

1. 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)
2. 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
3. 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)
4. 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)
5. 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
6. 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
7. Engilberge, M., Chevallier, L., Perez, P., Cord, M.: Sodeep: A sorting deep net to learn ranking loss surrogates. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
8. 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)
9. 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
10. 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)
11. 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/11.1177/0956797610385359
12. 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)
13. Xiao, H.: bert-as-service. https://github.com/hanxiao/bert-as-service (2018)