VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability
Supplementary Material

Romain Cohendet
Technicolor, Rennes, France
romain.cohendet@laposte.net

Claire-Hélène Demarty
InterDigital, Rennes, France
claire-helene.demarty@interdigital.com

Ngoc Q. K. Duong
InterDigital, Rennes, France
quang-khanh-ngoc.duong@interdigital.com

Martin Engilberge
InterDigital, Rennes, France
martin.engilberge@interdigital.com

1. Additional results

We present in this section the prediction results obtained by additional models on the VideoMem dataset. The purpose is to add some additional comparison with the models proposed in the main submission. We present all results (additional and main) in Table. 1.

1.1. Image captioning-based model

The first additional model leverages the use of some frame-based Image Captioning (IC) system as *feature extractor*. The goal was here to get a baseline of how a state-of-this-art IC model would perform on VideoMem, compared to our advanced fine-tuned IC model. We used the IC system in [1]. As this model is solely based on image information, similarly to what was done for the advanced IC model, for each video we extracted 7 features, one for each of the 7 frames in the advanced IC model. Each feature (of dimension 1024) corresponds to the projection in a joint image-text 2D embedding space of each frame. Instead of using these features individually as input to a model, and then averaging the 7 resulting scores as for the advanced model, we tested a concatenation of the 7 features as input to some MLP in an attempt to add some modelling of the temporal evolution of each video. The MLP was then trained to predict the VM score, with the mean square error (MSE) measure as regression loss. The final result is obtained with the MLP parameters (after some grid search): one hidden layer with 1500 neurons, optimizer=IBLGS, activation=tanh, learning rate (lr)=1e-3. From the results in table 1, it is clear that the advanced IC model did bring some substantial increase of performance compared to this baseline. Compared to the dedicated fine-tuning and new ranking loss used, even the attempt of temporal modeling by concatenation of the 7 features did not succeed in equalling the performance of the advanced system.

1.2. Fine-tuned ResNet model

The rational behind this second additional model was also to have a baseline to compare with our fine-tuned ResNet3D model as proposed in the main submission. For this, we also fine-tuned the original frame-based ResNet model [2] similarly to what we did for ResNet3D. As with ResNet3D, we replaced the last fully connected layer of ResNet by a new one dedicated to our considered regression task. This last layer was first trained alone for 5 epochs (Adam optimizer, batchsize=32, lr=1e-3), then the whole network was re-trained for more epochs (same parameters, but lr=1e-5).

Because this new model is frame-based, input data was all 7 frames pre-extracted from each video (one per second, each frame being assigned the same ground-truth score as the video) from VideoMem, mixed with images from LaMem [3], to enlarge the size of the overall training dataset. For the latter images, we normalized the ground-truth scores to be in the same range as those of VideoMem. Some data augmentation was conducted: random center cropping of 224x224 after resizing of the original images and horizontal flip, followed by a mean normalization computed on ImageNet. These two last settings i.e., training on additional LaMem data and pursuing some data augmentation differ to what was done with the fine-tuned ResNet3D. Note that we also tried another variant of ResNet: ResNet101. These changes were chosen in an attempt to challenge the results of the ResNet3D model, by trying to improve the performance of the frame-based ResNet version. In the end, to get a memorability score per video, we simply average the 7 resulting frame-based scores for all 7 frames of the video. Note that the ResNet101

Models	short-term memorability			long-term memorability		
	validation	test	test (500)	validation	test	test (500)
MemNet [3]	0.397	0.385	0.426	0.195	0.168	0.213
Squalli *et al.* [4]	0.401	0.398	0.424	0.201	0.182	0.232
C3D	0.319	0.322	0.331	0.175	0.154	0.158
HMP	0.469	0.314	0.398	0.222	0.129	0.134
ResNet (Sec. 1.2)	0.498	0.46	0.527	0.222	0.218	0.219
ResNet3D	0.508	0.462	0.535	0.23	0.191	0.202
IC-based model (Sec. 1.1)	0.492	0.442	0.514	0.22	0.201	0.188
Semantic embedding model	**0.503**	**0.494**	**0.565**	**0.26**	**0.256**	**0.275**

Table 1: Results in terms of Spearman's rank correlation between predicted and ground truth memorability scores, on the validation and test sets, and on the 500 most annotated videos of the dataset (test (500)) that were placed in the test set.

(a) Category #1. Images with one quite large face as main object.

(b) Category #1. Images with rare information in the background.

Figure 1: Visualization of the attention mechanism's output for frames in category #1. The model focuses either on close enough faces (a) or main objects when background texture is uniform or blurry.

was fine-tuned only to predict short-term memorability as LaMem dataset contains only short-term scores. Therefore, the comparison with the long-term performance is biased as we reused the model trained on short-term to predict long-term scores.

As shown in table 1, for short-term prediction, even with

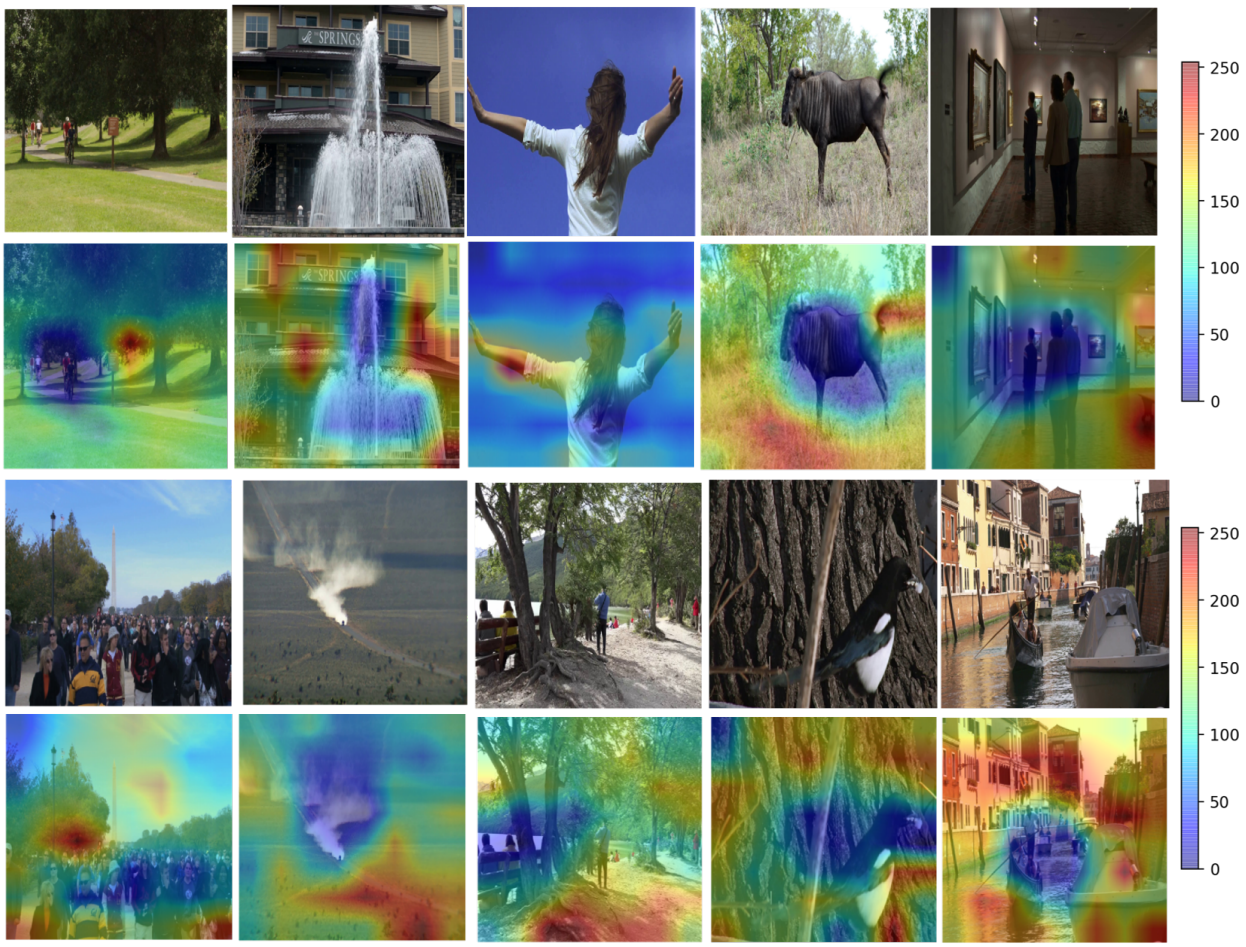

Figure 2: Visualization of the attention mechanism's output from frames in category #2. The model focuses on details in the background and not on the main objects.

the increased input data, the frame-based ResNet version provides slightly lower results than its temporal ResNet3D version. This tends to show that a real temporal modelling benefits to the task as we were able to reach slightly better results without requiring to additional data and with a less complicated model. Although not directly comparable, the results for long-term prediction are in contrary slightly higher with the frame-based version of ResNet, compared to ResNet3D. This confirms the finding of the main submission that, to some extend, both short-term and long-term memorabilities are correlated, as we were able to predict long-term scores with a model trained on short-term only.

2. Intra-memorability visualization

We present in Fig. 1 and Fig. 2 additional results for our model with attention mechanism, together with their original frames to provide a better and detailed visualization of the model's behavior.

As discussed in the main submission, we could empirically distinguish two main categories of frames and their associated results.

The first category was characterized by all attention maps focusing quite classically on the main object in the image, as it would have been expected intuitively. This matter of fact tends to happen in two cases, subdividing this first category in two groups: 1/ images with one rather large visible face and 2/ images with one main object and rare or no information in the background. For the former images, the model focuses on specific face features as a human would do when trying to remember a person, as it can be seen in Fig. 1, (a). Some additional quite important objects such as the music toy in image #2 are even *forgotten* in favor of faces. For the latter, as seen in Fig. 1, (b), the model focuses on the only source of information, because the background is dark or uniform (e.g., in image #2, a fireplace in front of a dark background).

In the second category, that groups all other frames, with several main and secondary objects, cluttered background,

etc., it seems on the contrary that the model focuses on some details which are out of the main objects/subjects of the images as it can be seen in Fig. 2. The model seems to behave as if it was trying to remember little details that will help it differentiate the image from another similar one. It might also be interpreted as a second memorization process, once the first one – focusing on the main object – is already achieved. For example, in the second row, in image #2, it focuses on building details in the background and not on the main fountain object, or in the fourth row, image #4, it concentrates on details of the tree bark and *forgets* the magpie.

References

[1] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and Claire-Hélène Demarty. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 11–14, 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[3] 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 (ICCV)*, pages 2390–2398, 2015.

[4] Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Gwenaëlle Marquant, and Claire-Hélène Demarty. Deep learning for predicting image memorability. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2371–2375, 2018.