Transfer learning for video memorability prediction

Romain Cohendet, Claire-Hélène Demarty and Ngoc Q. K. Duong
Technicolor
romain.cohendet@laposte.net,claire-helene.demarty@technicolor.com,quang-khanh-ngoc.duong@technicolor.com

ABSTRACT

This paper summarizes Technicolor's computational models to predict memorability of videos within the MediaEval 2018 Predicting Media Memorability Task. Our systems are based on deep learning features and architectures, and exploit the use of both semantic and multimodal features. Based on the obtained results, we discuss our findings and some scientific perspectives for the task.

1 INTRODUCTION

Understanding and predicting memorability of media such as images and videos has recently gained a significant attention from the research community. To facilitate the expansion of this research field, the Predicting Media Memorability Task is proposed at MediaEval 2018, which releases a large dataset of 10,000 videos, manually annotated with scores of memorability. A complete description of the task can be found in [4].

In order to automatically predict "short-term" and "long-term" memorability (as referred in the two proposed subtasks), we investigated different approaches, summarized in figure 1. Our first two approaches were intended to serve as a baseline for systems of video memorability prediction. We therefore re-used available high performance models for image memorability (IM) prediction and applied them directly to video memorability (VM) prediction (Section 2). Our second set of approaches (Section 3) investigated different features, including multi-modal ones. In a last approach (Section 4), instead of using an existing model as fixed feature extractor, we fine-tuned an entire state-of-the-art ResNet model to adapt it to the task of memorability prediction.

All the above models are frame-based. As input, we extracted seven frames (one per second) from each video, each frame being assigned the ground-truth score of its corresponding video. We then assess the VM score of a given video by simply averaging the seven predicted frame-based scores. When possible, we trained the models on short-term or long-term memorability ground-truth scores to build specific runs for the two subtasks. We also split the development set into 80% for training and 20% for validation. This random split was done at the video level, to enforce that frames from a single video were kept together in one part.

2 PRE-TRAINED IMAGE MEMORABILITY BASED APPROACHES

To construct a performance baseline of VM prediction, we tested two high-performance models available in the literature for IM prediction. Both were trained on the LaMem dataset [10], the largest dataset for IM to date (ca. 60,000 images from diverse sources).

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

MemNet-based system. The first network for large-scale IM prediction was presented in [10]. Based on the assumption that memorability depends on both scenes and objects, authors fine-tuned the training using a convolutional neural network (CNN) pre-trained both on the ImageNet and Places databases. They showed that fine-tuned deep features outperform other features by a large margin. We used this model as is to generate memorability scores for our video frames. We further averaged them to obtain the VM scores proposed in Run#1, identical for the two subtasks.

CNN and Image captioning based system. A more recent model that, to our knowledge, obtained the best performance up-to-now for IM prediction, was presented in [11]. It exploits both CNN-based and semantic Image captioning (IC)-based image features. The authors used the pre-trained VGG16 network for their CNN feature (extracted from the last layer), and a pre-trained IC model as an extractor for a more semantic image feature. The IC model builds an encoder consisting of a CNN and a long short-term memory recurrent network (LSTM) that enables to learn a joint image-text embedding by projecting the CNN image feature and the word2vec representation of the image caption on a 2D embedding space. Finally, the authors merged the two features using a Multilayer Perceptron (MLP). We also re-used this model as is as a second baseline which produces scores at frame level. Again, Run#2 is set to be identical for both subtasks.

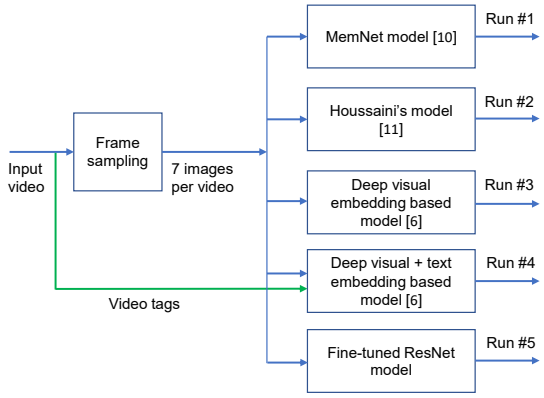

Figure 1: Summary of our approaches for VM prediction.

3 DEEP SEMANTICS EMBEDDING-BASED MULTIMODAL APPROACHES

We tested different features for VM prediction, including video-dedicated and frame-based features. Video-dedicated features included: *C3D* [13], *HMP* [1]. Frame-based features were extracted on three key-frames for each video and included: *Color histograms*, *InceptionV3 features* [12], *LBP* [8] and a set of *Aesthetic visual features* [7]. Please refer to [4] for more details on these features, as provided by the task's organizers.

Motivated by the finding that IC features perform well on both IM [11] and VM [5] prediction, we used the model proposed in [6] to extract some additional IC features from the frames. We also took advantage of this model to extract an additional text embedding feature from the titles provided with each video. As such a feature corresponds to a mapping of natural language words, i.e., a video description in our case, we expected an improvement of our system's capacity to capture semantics. We generated a new multimodal feature (image-text) by simply concatenating the two previous IC-based image and text features.

We then trained simple MLP (with one hidden layer of 100 neurons) on top of each single feature, and a concatenation of the 3 best non IC-based features. Again, these are frame-based models. Each time, two versions of the networks were trained on the short-term and long-term scores respectively. Table 1 shows the performance of each individual system on the validation data. From these results we decided to keep only the system with IC-image based features as input for Run#3 and the multimodal IC-(image+text) based features as input for Run#4, as the best performing features.

Features	short-term	long-term
C3D	.28	.126
HMP	.275	.114
ColorHist	.134	.05
InceptionV3	.16	.058
LBP	.267	.128
Aesthetics	.283	.127
C3D+LBP+Aesthetics	.347	.128
IC-image (Run#3)	.492	.22
IC-(image+text) (Run#4)	.436	.222

Table 1: Results in terms of Spearman's correlation obtained by a simple MLP for different video-dedicated and frame-based features, on the validation dataset.

4 FINE-TUNED RESNET101

As in [3, 10], where fine-tuned DNN outperformed classical approaches, we tried a transfer learning approach by fine-tuning a state-of-the-art ResNet model to the problem of IM prediction.

For this, we classically replaced the last fully connected layer of ResNet to a new one dedicated to our regression task of memorability prediction. This last layer was first trained alone for a few epochs (5), before re-training the complete network for more epochs. The following parameters were used: optimizer, Adam; batch size, 32. We used the *Mean Square Error* as loss function to stick to our regression task. 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. We trained on an augmented dataset composed of the 80% of the development set and LaMem (because of the latter, we processed to a normalization of the scores from the two datasets).

We fine-tuned two variants of ResNet: ResNet18 and ResNet101. We kept ResNet101 to generate scores for Run#5, as it gave the best performance on the validation set. We did not trained separate models for the short-term and long-term subtasks, due to time constraints. Note that, as LaMem images are provided with short-term memorability scores only, we would still have biased the network for long-term memorability prediction in doing so, but at least we could have improved the performance by using the long-term memorability scores of our dataset. So, Run#5 is identical for both subtasks.

5 RESULTS AND DISCUSSION

Results are summarized in Table 2. The results of the first two runs for the short-term subtask show that it is possible to achieve quite good results in VM prediction using models designed for IM prediction This means that the memorability of a video is correlated to some extent with the memorability of its constituent frames. We may also note the poor performance of all models for the long-term subtask, compared to the short-term subtask. For runs #1, #2 and #5, this may be explained by the fact that the training was done with the use of LaMem for which only short-term scores are available. This may also come from the significantly lower number of annotations for the long-term scores in the task's dataset [4]. It may also highlight that there is a significant difference between short-term and long-term memorability and that it might be more difficult to predict the latter. However, these results also prove that long-term memorability is correlated – though not perfectly – with short-term memorability. In accordance with the literature, the model of [11] performed a little better than the model of [10] for memorability prediction.

Runs	short-term mem.				long-term mem.			
	Spearman		Pearson		Spearman		Pearson	
	val.	test	val.	test	val.	test	val.	test
1-MemNet	.397	.385	.414	.406	.195	.168	.188	.184
2-CNN&IC	.401	.398	.405	.402	.201	.182	.199	.191
3-IC	.492	.442	.501	**.493**	.22	.201	.233	.216
4-Multi	.452	.418	.48	.451	.212	.208	.23	**.228**
5-ResNet	.498	**.46**	.512	.491	.198	**.219**	.217	.217

Table 2: Official results on the test set, and results on the validation set. (Official metric: Spearman's corr.)

Runs #3 and #4 perform better than runs #1 and #2. As in [11] and [5], IC features performed well for memorability prediction tasks, especially when fine-tuned on the new dataset (Run#3 can be seen as a fine-tuned version of Run#2). Indeed, IC features convey high semantics: high-level visual attributes and scene semantics (actions, movements, appearance of objects, emotions, etc.) have been founded to be linked to memorability [9, 10]. It also shows that training of long-term scores helps improving the performance for long-term memorability. The multimodal approach gave slightly worse results than IC features alone. However, due to time constraints, we did not proceed to any optimizing of the set of parameters, to deal with the possible redundancy between IC image and text embedding features.

The most accurate memorability prediction were obtained by the fine-tuned ResNet101, which confirms that transfer learning from an image classification problem to yet another task such as memorability prediction works well. This validates also the quality of the dataset at least for the short-term annotations. As perspectives, it will be interesting to test systems incorporating temporal evolution of the videos such as motion information or latest architectures such as TCN [2] to see how it improves the performances.

REFERENCES

[1] Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. 2011. Comparison of video sequences with histograms of motion patterns. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*. 3673–3676.

[2] S. Bai, J.Z. Kolter, and Koltun V. 2018. *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling.* Technical Report. arXiv preprint arXiv:1803.01271.

[3] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. 2016. Deep Learning for Image Memorability Prediction: the Emotional Bias. In *Proc. ACM International Conference on Multimedia (ACMM)*. 491–495.

[4] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. MediaEval 2018: Predicting Media Memorability Task. In *Proc. of the MediaEval Workshop*.

[5] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicting long-term video memorability. In *Proc. of the ICMR 2018 Workshop, Yokohama, Japan, June 11-14*.

[6] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. 2018. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. 3984–3993.

[7] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, and others. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *PeerJ* 3 (2015), e1390.

[8] Dong-Chen He and Li Wang. 1990. Texture unit, texture spectrum, and texture analysis. *IEEE Transactions on Geoscience and Remote Sensing* 28, 4 (1990), 509–512.

[9] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1469–1482.

[10] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 2390–2398.

[11] Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. 2018. Deep learning for predicting image memorability. In *Proc. IEEE International Conference on Audio, Speech and Language Processing (ICASSP)*.

[12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 4489–4497.