Predicting Media Memorability with Audio, Video, and Text representations

Alison Reboud*, Ismail Harrando*, Jorma Laaksonen+ and Raphaël Troncy*

*EURECOM, Sophia Antipolis, France
+Aalto University, Espoo, Finland
{alison.reboud,ismail.harrando,raphael.troncy}@eurecom.fr
jorma.laaksonen@aalto.fi

ABSTRACT

This paper describes a multimodal approach proposed by the MeMAD team for the MediaEval 2020 "Predicting Media Memorability" task. Our best approach is a weighted average method combining predictions made separately from visual, audio, textual and visiolinguistic representations of videos. Our best model achieves Spearman scores of 0.101 and 0.078, respectively, for the short and long term predictions tasks.

1 INTRODUCTION

Considering video memorability as a useful tool for digital content retrieval as well as for sorting and recommending an ever growing number of videos, the Predicting Media Memorability task aims at fostering the research in the field by asking its participants to automatically predict both a short and a long term memorability score for a given set of annotated videos. The full description for this task is provided in [5]. Last year's best approaches for both the long term [10] and short term tasks [2] rely on multimodal features. Our method is inspired from last year's best approaches but also acknowledges the specifics of the 2020's edition dataset. More specifically, because in comparison to last year's set of videos, the TRECVid videos contain more actions, our model uses video features and image features for multiple frames. In addition, because this year sound was included in the videos, our model includes audio features. Finally, a key contribution of our approach is to test the relevance of visiolinguistic representation for the Media Memorability task. Our final model[1] is a multimodal weighted average with visual and audio deep features extracted from the videos, textual features from the provided captions and visiolinguistic features.

2 APPROACH

We trained separate models for the short and long term predictions using originally a 6-fold cross-validation of the training set, which means that we typically had 492 samples for training and 98 samples for testing each model.

2.1 Audio-Visual Approach

Our audio-visual memorability prediction scores are based on using a feed-forward neural network with a concatenation of video and audio features in the input, one hidden layer of units and one unit in the output layer. The best performance was obtained with 2575-dimensional features consisting of the concatenation of 2048-dimensional I3D [3] video features and 527-dimensional audio features. Our audio features encode the occurrence probabilities of the 527 classes of the Google AudioSet Ontology [6] in each video clip. The hidden layer uses ReLU activations and dropout during the training phase, while the output unit is sigmoidal. The training of the network used the Adam optimizer. The features, the number of training epochs and the number of units in the hidden layer were selected with the 6-fold cross-validation. For short term memorability prediction, the optimal number of epochs was 750 and the optimal hidden layer size 80 units, whereas for the long term prediction these figures were 260 and 160, respectively.

We also experimented with other types of features and their combinations. These include the ResNet [7] features extracted just from the middle frames of the clips as this approach worked very well last year. The contents of this year's videos are, however, such that genuine video features I3D and C3D [13] work better than still image features. When I3D and AudioSet features are used, C3D features do not bring any additional advantage.

2.2 Textual Approach

Our textual approach leverages the video descriptions provided by the organizers. First, all the provided descriptions are concatenated by video identifier to get one string per video. To generate the textual representation of the video content, we used the following methods:

- Computing TF-IDF, removing rare (less than 4 occurrences) and stopwords and accounting for frequent 2-grams.
- Averaging GloVe embeddings for all non-stopwords words using the pre-trained 300d version [9].
- Averaging BERT [4] token representations (keeping all the words in the descriptions up to 250 words per sentence).
- Using Sentence-BERT [11] sentence representations. We use the distilled version that is fine-tuned for the STS Textual Similarity Benchmark[2].

For each representation, we experimented with multiple regression models and finetuned the hyper-parameters for each model

[1] https://github.com/MeMAD-project/media-memorability

[2] https://huggingface.co/sentence-transformers/distilbert-base-nli-stsb-mean-tokens

MediaEval'20, 14-15 December 2020, Online

using the 6-fold cross-validation on the training set. For our submission, we used the *Averaging GloVe embeddings* with a Support Machine Regressor with an RBF kernel and a regulation parameter $C = 1e - 5$.

We also attempted enhancing the provided descriptions with additional captions automatically generated using the DeepCaption[3] software. We did not see an improvement in the results, which is probably due to the nature of the clips provided for this year's edition (as DeepCaption is trained on static stock images from MS COCO and TGIF datasets).

2.3 Visiolinguistic Approach

ViLBERT [8] is a task-agnostic extension of BERT that aims to learn the associations and links between visual and linguistic properties of a concept. It has a two-stream architecture, first modelling each modality (i.e. visual and textual) separately, and then fusing them through a set of attention-based interactions (co-attention). ViLBERT is pre-trained using the Conceptual Captions data set (3.3M image-caption pairs) [12] on masked multi modal learning and multi-modal alignment prediction. We used a frozen pre-trained model which was fine-tuned twice, first on the task of Video-Question Answering (VQA) [1] and then on the 2019 MediaEval Memorability task and dataset.

The 1024-dimensional features extracted for the two modalities can be combined in different ways.In our experiment, multiplying textual and visual feature vectors performed the best for short term memorability prediction but using the sole visual feature vectors worked better for long term memorability prediction. Averaging the features extracted from 6 frames performed better than only using only the middle frame. We experimented with the same set of regression models as for the textual approach. In our submission, we used a Support Machine Regressor with a regulation parameter $C = 1e - 5$ and an RBF or Poly kernel respectively for short and long term scores prediction.

3 RESULTS AND ANALYSIS

We have prepared 5 different runs following the task description defined as follows:

- run1 = Audio-Visual Score
- run2 = Visiolinguistic Score
- run3 = Textual Score
- run4 = 0.5 * run1 + 0.2 * run2 + 0.3 * run3
- run5 = run4 with LT scores for LT task

For the Long Term task, all models except *run5* use exclusively short-term scores. For runs 4 and 5, we normalise the scores obtained from runs 1, 2 and 3 before combining them.

Table 1 provides the Spearman score obtained for each run when performing a 6-folds cross-validation on the training set. We observe that our models use only the training set, as the annotations on the later-provided development set did not yield better results. We hypothesize that this is due to the fewer number of annotations per video available as many videos had a score for 1, for instance, which we do not observe on the training set.

[3]https://github.com/aalto-cbir/DeepCaption

Table 1: Average Spearman score obtained on a 6-folds cross validation of the Training set

Method	Short Term	Long Term
run1	0.2899	0.179
run2	0.214	0.1309
run3	0.2506	0.1372
run4	**0.3104**	**0.2038**
run5	0.067	0.1700

Table 2: Results on the Test set for Short Term (ST) and Long Term (LT) memorability

Method	SpearmanST	PearsonST	SpearmanLT	PearsonLT
run1	0.099	0.09	0.077	0.0855
run2	0.098	0.085	-0.017	0.011
run3	0.073	0.091	0.019	0.049
run4	**0.101**	**0.09**	**0.078**	**0.085**
run5	0.101	0.09	0.067	0.066
AvgTeams	0.058	0.066	0.036	0.043

We present in Table 2 the final results obtained on the test set using models trained on the full training set composed of 590 videos. We observe that the weighted average method which uses short term scores works the best for both short and long term prediction, obtaining results which are approximately double the mean Spearman score obtained across the teams. Our best results (Spearman scores) on the test set are however significantly worse than the ones we obtained on average over the 6-folds of the training set suggesting that the test set is quite different from the training set. The results for Long Term prediction are always worse than the ones for Short Term prediction. Finally, both our scores and the mean score across team are below the ones obtained for the 2018 and 2019 videos.

4 DISCUSSION AND OUTLOOK

This paper describes a multimodal weighted average method proposed for the 2020 Predicting Media Memorability task of MediaEval. One of the key contribution of this paper is to have shown that based on our experiments during the model construction or testing phase, in comparison to image, audio and text, video features performed the best. Similarly to last year, short term scores predictions correlated better with long term scores than the predictions made when training directly on long term scores. Finally considering the difference of results obtained between the training and test set, it would be interesting to investigate further the differences between these datasets in terms of content (video, audio and text) and annotation. We conclude that generalizing this type of task to different video genres and characteristics remain a scientific challenge.

Acknowledgements

This work has been partially supported by the European Union's Horizon 2020 research and innovation programme via the project MeMAD (GA 780069).

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile.

[2] David Azcona, Enric Moreu, Feiyan Hu, Tomás E Ward, and Alan F Smeaton. 2019. Predicting media memorability using ensemble models. In *MediaEval 2019: Multimedia Benchmark Workshop*. Sophia Antipolis, France.

[3] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4724–4733.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. ACL, Minneapolis, Minnesota, USA, 4171—-4186.

[5] Alba García Seco de Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Claire-Hélène Demarty, Faiyaz Doctor, Bogdan Ionescu, and Alan F. Smeaton. 2020. Overview of MediaEval 2020 Predicting Media Memorability task: What Makes a Video Memorable?. In *Working Notes Proceedings of the MediaEval 2020 Workshop*.

[6] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, Louisiana, USA, 776–780.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, Nevada, USA, 770–778.

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *33^{rd} Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Melbourne, Australia, 1532—-1543.

[10] Alison Reboud, Ismail Harrando, Jorma Laaksonen, Danny Francis, Raphaël Troncy, and Héctor Laria Mantecón. 2019. Combining Textual and Visual Modeling for Predicting Media Memorability. In *MediaEval 2019: Multimedia Benchmark Workshop*. Sophia Antipolis, France.

[11] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Hong Kong, China, 3982—-3992.

[12] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, Melbourne, Australia, 2556–2565.

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