Exploring Multimodality, Perplexity and Explainability for Memorability Prediction

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 several approaches proposed by the MeMAD Team for the MediaEval 2021 "Predicting Media Memorability" task. Our best approach is based on early fusion of multimodal (visual and textual) features. We also designed one of our run to be explainable in order to give new insights into the topic of audio visual content memorability. Finally, one of our runs is an experiment in analysing the potential role played by text perplexity in video content memorability.

1 APPROACH

The description of the task as well as the metrics used for its evaluation is detailed in [8]. We have experimented in the past with approaches combining textual and visual features [12] as well as using visio-linguistic models [13] for predicting short and long term media memorability. This year, we have explored other methods including: i) performing early fusion of multimodal features, ii) attempting to explain whether some phrases could trigger memorability or not and iii) estimating the perplexity of video descriptions. Our code to enable reproducibility of our approaches is available at https://github.com/MeMAD-project/media-memorability.

1.1 Early Fusion of Multimodal Features

Textual features. Our textual approach uses the video descriptions (or captions) provided by the task organizers. First, we concatenate the video descriptions to obtain one string for each video. Then, to get a textual representation of the video content, we experimented with 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 [11].
- Averaging BERT [4] token representations (keeping all the words in the descriptions up to 250 words per sentence).
- Using Sentence-BERT [14] sentence representations and in particular the distilled version that is fine-tuned for the STS Textual Similarity Benchmark[1]
- Using again Sentence-BERT with the model fine-tuned on the Yahoo answers topics dataset, comprising of questions and answers from Yahoo Answers, classified into 10 topics.

For each representation, we experimented with multiple regression models and fine-tuned the hyper-parameters using a fixed 6-fold cross-validation on the training set. For our submission, we used the *Sentence-BERT on Yahoo answers topic dataset* model.

Visual features. We extracted 2048-dimensional I3D [3] features to describe the visual content of the videos. The I3D features are extracted from the *Mixed_5c* layer of the readily-available model trained with the Kinetics-400 dataset [7]. These features performance are superior to those extracted from the 400-dimensional classification output and the C3D [15] features provided by the task organizers.

Audio features. We used 527-dimensional audio features that encode the occurrence probabilities of the 527 classes of the Google AudioSet Ontology [5] in each video clip. The model uses the readily-available VGGish feature extraction model [6].

Prediction model. In all our early fusion experiments, the respective features were concatenated to create multimodal input feature vectors. We used a feed-forward network with one hidden layer to predict the memorability score. We varied the number of units in the hidden layer and optimized it together with the number of training epochs. We used ReLU non-linearity and dropout between the layers and simple sigmoid output for the regression result. The experiments used the same 6-fold cross-validation on the training set. The best models typically consisted of 600 units in the hidden layer and needed 700 training epochs to produce the maximal Spearman correlation score. We have also experimented with a weighted average to combine modalities, but early fusion turned out to be more successful.

1.2 Exploring Explainability

We have experimented with different simple text-based models that offer the possibility to quantify the relation between the caption and the predicted memorability score in an explainable manner. We train the models given the specific sub-task and dataset, i.e. for the short-term memorability predictions, we train the models on the short-term memorability scores.

We compare feeding simple linear models (regressors) interpretable input features: bag of words, TF-iDF, and topic distributions produced by an LDA model [2] trained on the corpus made

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

MediaEval'21, December 13-15 2021, Online

of captions. Upon evaluating the performance of each model/input feature pair in a cross-fold validation protocol, we obtain the best results using TF-iDF features with a Linear Support Vector Regression (LinearSVR[2]). On one hand, this model allows us to somewhat understand the correspondence between some input words and the final score of classification. For example, the top words for raw and normalized short-term memorability on both `Memento10K` and `TRECVID` is *woman*. On the other hand, the empirical performance on both subtasks falls significantly behind other models, demonstrating both the non-linear and multimodal nature of memorability.

1.3 Exploring Perplexity

It has been suggested that memorable content can be found in sparse areas of an attribute space [1]. For example, images with convolutional neural networks features sparsely distributed have been found to be more memorable [9]. Additionally, we observe that the results obtained on the TRECVID dataset (made of short videos from Vine) are considerably worse than those obtained on the Memento10K dataset which may be due to the fact that the TRECVID dataset is smaller but also much more diverse. One hypothesis is that popular vines break with expectations. Backing this hypothesis, we have found in the TRECVID dataset that videos depicting a person eating a car, or a chicken coming out of an egg to have a high memorability score. Therefore, inspired by [10] who predicts the novelty of a caption, we wanted to test the hypothesis that the novelty of a caption influences its memorability.

We explore the (pseudo-)perplexity of each video description using a pretrained RoBERTa-large model. The score for each caption is computed by adding up the log probabilities of each masked token in the caption, and the aggregation between captions is done with a max function. We select the caption with the highest perplexity for each video. All runs have identical scores for each dataset as we do not use the training set at all in this method.

2 RESULTS AND DISCUSSION

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

- run1 = Explainable (Section 1.2)
- run2 = Early Fusion of Textual+Visual+Audio features
- run3 = Early Fusion of Textual+Visual features
- run4 = Perplexity-based (Section 1.3)
- run5 = Early fusion of Textual+Visual features trained on the combined (TRECVID + Memento10k) datasets

All models except the *run1* use exclusively short-term scores for predicting the long-term score.

We present in Tables 1 and 2 the final results obtained on the test set of respectively the TRECVID and the Memento10k datasets. We comment on the Spearman Rank scores as this is the official evaluation metrics. We observe that the early fusion method which uses short term scores works the best for both short and long term predictions. Adding the audio modality features did not improve the results. We can also observe that the results for Long Term prediction are always worse than the ones for Short Term prediction and the results for Memento10K are always better. Combining the datasets did not yield better results. This is not very surprising for the Memento10K results since it is a bigger dataset. However, the fact that augmenting the TRECVID dataset did not lead to significant improvement suggests that beyond a size difference, there is a difference in nature between the datasets that leads to a bad generalisation in terms of prediction. This fact is confirmed by the generalisation subtask which yields significantly worse results for both Memento10K and TRECVID. Finally the scores obtained with the perplexity run were by far the lowest, only reaching 0.073 for Memento10K when our best run obtained 0.658. With this run, rather than obtaining the best results, we wanted to evaluate the potential for adding a caption perplexity measure. At this stage, these results do not suggest a strong relationship between perplexity and memorability.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVR.html

Table 1: Results on the TRECVID Test set for Short Term Raw (STr), Short Term Normalized (STn) and Long Term (LT) memorability (Sp = Spearman, Pe= Pearson)

Method	SpSTr	PeSTr	SpSTn	PeSTn	SpLT	PeLT
run1	0.127	0.153	0.158	0.168	0.016	0.014
run2	0.216	0.212	0.221	0.209	0.060	0.090
run3	**0.220**	0.214	**0.226**	0.218	**0.063**	**0.098**
run4	−0.050	0.013	−0.052	0.018	−0.043	0.024
run5	0.196	**0.215**	0.211	**0.222**	0.062	0.059

Table 2: Results on the Memento10K Test set for Short Term Raw (STr) and Short Term Normalized (STn) memorability

Method	SpSTr	PeSTr	SpSTn	PeSTn
run1	0.464	0.460	0.463	0.458
run2	**0.658**	**0.674**	0.657	0.674
run3	0.655	0.672	**0.658**	**0.675**
run4	0.073	0.064	0.077	0.069
run5	0.654	0.672	0.651	0.671

Table 3: Generalisation subtask: results on the TRECVID Test set for Short Term Raw (STr), Short Term Normalized (STn) and Long Term (LT) memorability

Method	SpSTr	PeSTr	SpSTn	PeSTn	SpLT	PeLT
run1	0.076	0.099	0.068	0.091	-0.013	0.021
run2	0.140	0.165	0.146	0.170	0.045	0.042

Table 4: Generalisation subtask: results on the Memento10K Test set for Short Term Raw (STr) and Short Term Normalized (STn) memorability

Method	SpSTr	PeSTr	SpSTn	PeSTn
run1	0.196	0.196	0.181	0.184
run2	0.310	0.313	0.320	0.316

REFERENCES

[1] Wilma A Bainbridge. 2021. Shared memories driven by the intrinsic memorability of items. *Human Perception of Visual Information: Psychological and Computational Perspectives* (2021).

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993—-1022.

[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] 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.

[6] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. (2017). arXiv:cs.SD/1609.09430

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. (2017). arXiv:cs.CV/1705.06950

[8] Rukiye Savran Kiziltepe, Mihai Gabriel Constantin, Claire-Hélène Demarty, Graham Healy, Camilo Fosco, Alba García Seco de Herrera, Sebastian Halder, Bogdan Ionescu, Ana Matran-Fernandez, Alan F. Smeaton, and Lorin Sweeney. 2021. Overview of The MediaEval 2021 Predicting Media Memorability Task. In *Multimedia Benchmark Workshop (MediaEval)*.

[9] Jiří Lukavskỳ and Filip Děchtěrenko. 2017. Visual properties and memorising scenes: Effects of image-space sparseness and uniformity. *Attention, Perception, & Psychophysics* 79, 7 (2017), 2044–2054.

[10] Nianzu Ma, Alexander Politowicz, Sahisnu Mazumder, Jiahua Chen, Bing Liu, Eric Robertson, and Scott Grigsby. 2021. Semantic Novelty Detection in Natural Language Descriptions. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 866–882.

[11] 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.

[12] Alison Reboud, Ismail Harrando, Jorma Laaksonen, Danny Francis, Raphael Troncy, and Hector Laria Mantecon. 2019. Combining Textual and Visual Modeling for Predicting Media Memorability. In *Multimedia Benchmark Workshop (MediaEval) (CEUR Workshop Proceedings)*, Vol. 2670.

[13] Alison Reboud, Ismail Harrando, Jorma Laaksonen, and Raphael Troncy. 2020. Predicting Media Memorability with Audio, Video, and Text representations. In *Multimedia Benchmark Workshop (MediaEval) (CEUR Workshop Proceedings)*, Vol. 2882.

[14] 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—-10.

[15] 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.