Predicting Media Memorability: Comparing Visual, Textual and Auditory Features

Lorin Sweeney, Graham Healy, Alan F. Smeaton
Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland
lorin.sweeney8@mail.dcu.ie,graham.healy@dcu.ie,alan.smeaton@dcu.ie

ABSTRACT

This paper describes our approach to the Predicting Media Memorability task in MediaEval 2021, which aims to address the question of media memorability by setting the task of automatically predicting video memorability. This year we tackle the task from a comparative standpoint, looking to gain deeper insights into each of three explored modalities, and using our results from last year's submission (2020) as a point of reference. Our best performing short-term memorability model (0.132) tested on the TRECVid2019 dataset—just like last year—was a frame based CNN that was not trained on any TRECVid data, and our best short-term memorability model (0.524) tested on the Memento10k dataset, was a Bayesian Ride Regressor fit with DenseNet121 visual features.

1 INTRODUCTION AND RELATED WORK

In the ever expanding storm of social media, the need for tools that help us wade through daily digital torrents will only grow. It can be argued that memorability is a measure whose shape uniquely fits the jagged edged problem of media content curation. Our lack of meta-cognitive insight into what we will ultimately remember or forget [6], casting clouds of obscuring cover over answers that will thread together our sense of self, is what motivates and brings meaning to the exploration of memorability—generally known as the likelihood of an observer remembering a repeated piece of media in a stream of media.

This paper outlines our participation in the 2021 MediaEval Predicting Media Memorability Task [8], which includes an extended subset of last year's TRECVid 2019 Video-to-Text dataset [2], and Memento10k [10]—a large and diverse short-term video memorability dataset. This year, short-term (after minutes) memorability is further sub-categorised into *raw* and *normalised*, and long-term (after 24-72 hours) memorability is kept the same. Additionally, two video memorability prediction sub-tasks were put forward, the first (sub-task 1) following the standard train with provided data to generate predictions, and the second (sub-task 2) taking the form of a constrained generalisation task—where the training and testing data must be from different sources. Further information about the datasets, annotation protocol, pre-computed features, and ground-truth data can be found in the task overview paper [8].

With last year's task [2] including audio as part of the video data for the first time, its impact in the context of multi-modal media was thrust into the limelight. While no conclusive findings were established, the best long-term memorability prediction came from an xResNet34 trained purely on audio spectrograms [14], suggesting that the audio modality provides a degree of useful information during video memorability recognition tasks. Additionally, follow-on work found evidence to suggest that "audio plays a contextualising role, with the potential to act as a signal or a trigger that aids recognition" depending on the extent of high-level human understandable information it contains, and the context in which it is presented [15].

Many previous works have firmly established the utility of combining features from more than one modality, and highlighted the effectiveness of combining deep visual features in conjunction with semantically rich features, such as captions; emotions; or actions in order to predict media memorability [1, 10, 12, 16]. However, given that this year's sub-task 1 could be viewed as a natural extension of the previous year's task, that this year's sub-task 2 is a generalisation task, and that the previous years official results were abnormally low across the board, we opted treat this year's task as one of insight rather than optimisation, keeping modalities separate, rather than following state of the art by combining features across modalities—which ultimately obscures the extent to which each modality contributes to the final memorability score prediction—and limiting each of our runs to one modality.

2 APPROACH

Both datasets are comprised of three subsets, a training set; a development set; and a test set, with the TRECVid training set comprising of 588 videos, and Memento10k 7,000 videos. The development sets contain 1,116 and 1,500 videos respectively, and the test sets contain 500 and 1,500 videos respectively. Our approach this year was to use the task as an opportunity to compare our results from last year, cutting down the complexity and focusing on one of three modalities, visual, textual, and auditory.

Visual: For our visual approach, we implemented two methods, the first of which was a Bayesian Ridge Regressor (BRR) that we fit with default sklearn [11] parameters using the provided DenseNet121 [5] features (which were extracted from the first, middle, and last video frames), and the second method was an ImageNet-pretrained xResNet50 that was either fine-tuned (for 50 epochs, with a maximum learning rate of 1e-3, and weight decay of 1e-2) on the Memento10k training data and then further fine-tuned (for 10 epochs, with a maximum learning rate of 3e-2, and weight decay of 1e-1) on the TRECVid development set videos, fine-tuned on the Memento10k training data, or fine-tuned on the LaMem [7] dataset depending on the run and its restrictions.

Textual: For our textual approach, we implemented a caption model, the AWD-LSTM (ASGD Weight-Dropped LSTM) architecture [9], a highly regularised and competitive language model. Transfer learning was used in order to fully avail of the high-level representations that a language model offers. The specific transfer

MediaEval'21, 13-15 December 2021, Online

learning method employed was UMLFiT [4], which uses discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing. The language model was pre-trained on the Wiki-103 dataset, and fine-tuned (for 10 epochs, with a maximum learning rate of 2e-3, a weight decay of 1e-2, and a dropout multiplier of 0.5) on the first 300,000 captions from Google's Conceptual Captions dataset [13]. The encoder from that fine-tuned language model was then used in each of our caption models, which were either trained (for 15 epochs, with a maximum learning rate of 1e-3, a weight decay of 1e-2, and a dropout multiplier of 0.8) on a paragraph of all five Memento10k training captions, or additionally fine-tuned on the first TRECVid development set captions to predict memorability scores rather than the next word in a sentence.

Auditory: Initially, we extracted Mel-frequency cepstral coefficients from the videos that had audio, stacked them together with their delta coefficients in order to create a three channel spectrogram images, and used them to train an ImageNet-pretrained xResNet34. However, after experimenting with VGGish [3] features—extracting 128-dimensional embeddings for each second of the first three seconds of audio, resulting in a 384-dimensional feature set per video—and using them to fit a BRR, we noticed marginally, but consistently better results, and opted to use them in favour of spectrogram images in our final run submissions.

3 DISCUSSION AND OUTLOOK

Tables 1 and 2 show the Spearman scores (r_s) and Pearson scores (r) for our runs from sub-task 1, with Table 1 showing the scores for our runs tested on the official TRECVid test set, and Table 2 showing the scores which came from the official Memento10k test set. Table 3 shows the r_s and r scores for our runs from sub-task 2—the generalisation task—which were trained without any TRECVid data, and tested on the official TRECVid test set.

Although more TRECVid videos were provided this year compared to last year (1,704 vs 1,000), and even though our short-term TRECVid test scores (Table 1) are roughly double what they were last year, the scores are still quite low compared to expected results from training validation, and the long-term scores are shockingly poor. While it is not possible to pinpoint the exact cause of this, it is quite likely that either there is lack of distributional overlap between videos used to train and test our models, or that there still are not enough videos to be able to properly generalise. Both of these possibilities are supported by the fact that our best performing TRECVid run—just like last year—came from a model not trained on any TRECVid data, but purely on memento10k data (Table 3. *xResNet50 Frames Memento*), which is a much larger, varied, and "in-the-wild" video memorability dataset than TRECVid.

Results from Table 2 show that the best performing model on the Memento10k dataset was a BRR fit on DenseNet121 features, indicating that visual features contribute quite a lot to the overall memorability of a video. The next best model was a BRR trained on VGGish audio features, which is very interesting as the Memento10k ground-truth scores were gathered with the videos being played without sound. The stark order of magnitude difference in performance between a BRR trained on Memento10k data (0.524) and one trained on TRECVid data (0.053), raises some interesting questions concerning the nature of the differences in visual content between these two datasets, which unfortunately cannot be answered in this paper.

Results from the generalisation task (Table 3) further highlight the aforementioned potential distributional problems with the TRECVid dataset. Given that the performance of both the frame based and caption based models is worse on the TRECVid test set when further fine-tuned on the TRECVid training and development data, a detailed exploration and investigation into the nature and distributions of the TRECVid subsets could be very fruitful.

While insights into possible causes of last year's uncharacteristically low task-wide scores across participant submissions were gained, few tangible insights into the influence of each of the explored modalities—visual, textual, and auditory—were obtained. In order to fully reveal the influence of each of the modalities, independent ground-truth memorability scores are required to elucidate the role they each play when coinciding with one another in a multi-modal medium such as video, and should be a focus of future memorability tasks and research.

Table 1: Official results on test-set for sub-task 1 for the TRECVid dataset.

	Short-norm		Long	
Run	$\mathbf{r_s}$	r	$\mathbf{r_s}$	r
BayesianRidge Dense121	0.053	0.071	-	-
xResNet50 Transfer Frames	0.105	0.13	-.021	-.036
AWD-LSTM Transfer Caption	0.105	0.083	0.002	0.013

Table 2: Official results on test-set for sub-task 1 for the Memento10k dataset.

	Short-raw		Short-norm	
Run	$\mathbf{r_s}$	r	$\mathbf{r_s}$	r
BayesianRidge Dense121	**0.523**	0.522	**0.524**	0.522
BayesianRidge Vggish	0.29	0.289	0.272	0.274
AWD-LSTM Caption	-	-	0.174	0.172
AWD-LSTM Transfer Caption	-	-	0.181	0.163
xResNet50 Frames	-	-	0.129	0.114

Table 3: Official results on test-set for sub-task 2 for the TRECVid dataset.

	Short-raw		Short-norm	
Run	$\mathbf{r_s}$	r	$\mathbf{r_s}$	r
BayesianRidge Vggish Memento	0.018	0.008	0.021	0.012
xResNet50 Frames LaMem	0.093	0.073	0.088	0.076
xResNet50 Frames Memento	**0.116**	0.131	**0.132**	0.145
AWD-LSTM Caption Memento	0.114	0.12	0.106	0.11

ACKNOWLEDGEMENTS

This work was funded by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

REFERENCES

[1] David Azcona, Enric Moreu, Feiyan Hu, Tomás Ward, and Alan F Smeaton. 2019. Predicting media memorability using ensemble models. In *Proceedings of MediaEval 2019, Sophia Antipolis, France*. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2670/

[2] 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 *Proceedings of the MediaEval 2020 Workshop*.

[3] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[4] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 328–339.

[5] Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens Van Der Maaten, and Kilian Weinberger. 2019. Convolutional Networks with Dense Connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[6] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2013. What makes a photograph memorable. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2013), 1469–1482.

[7] 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*. 2390–2398.

[8] Rukiye Savran Kiziltepe, Mihai Gabriel Constantin, Claire-Helène 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 *Proceedings of the MediaEval 2021 Workshop*.

[9] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and Optimizing LSTM Language Models. In *International Conference on Learning Representations*.

[10] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. 2020. Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 223–240.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[12] Alison Reboud, Ismail Harrando, Jorma Laaksonen, Raphaël Troncy, and others. 2020. Predicting Media Memorability with Audio, Video, and Text representations. In *Proceedings of the MediaEval 2020 Workshop*.

[13] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[14] Lorin Sweeney, Graham Healy, and Alan F Smeaton. 2020. Leveraging Audio Gestalt to Predict Media Memorability. In *Proceedings of the MediaEval 2020 Workshop*.

[15] Lorin Sweeney, Graham Healy, and Alan F. Smeaton. 2021. The Influence of Audio on Video Memorability with an Audio Gestalt Regulated Video Memorability System. In *Proceedings of the 2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–6.

[16] Tony Zhao, Irving Fang, Jeffrey Kim, and Gerald Friedland. 2020. Multimodal Ensemble Models for Predicting Video Memorability. In *Proceedings of the MediaEval 2020 Workshop*.