Leveraging Audio Gestalt to Predict Media Memorability

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

ABSTRACT

Memorability determines what evanesces into emptiness, and what worms its way into the deepest furrows of our minds. It is the key to curating more meaningful media content as we wade through daily digital torrents. The Predicting Media Memorability task in MediaEval 2020 aims to address the question of media memorability by setting the task of automatically predicting video memorability. Our approach is a multimodal deep learning-based late fusion that combines visual, semantic, and auditory features. We used audio gestalt to estimate the influence of the audio modality on overall video memorability, and accordingly inform which combination of features would best predict a given video's memorability scores.

1 INTRODUCTION AND RELATED WORK

Our memories make us who we are—holding together the very fabric of our being. The vast majority of the population predominantly rely on visual information to remember and identify people, places, and things [16]. It is known that people generally tend to remember and forget the same images [3, 14], which implies that there are intrinsic qualities or characteristics that make visual content more or less memorable. While there is some evidence to suggest that sounds similarly have such intrinsic properties [21], it is generally accepted that auditory memory is inferior to visual memory, and decays more quickly [4, 6, 18]. However, it is important not to resultantly resort to dismissing the role of the audio modality in memorability—as multisensory experiences exhibit increased recall accuracy compared to unisensory ones [26, 27]—but to avail of the potential contextual priming information sounds provide [22]. In our work, we seek to explore the influence of the audio modality on overall video memorability, and employ [21]'s concept of audio gestalt in order to do so.

Image memorability is commonly defined as the probability of an observer detecting a repeated image in a stream of images a few minutes after exposition [3, 13–15]. This paper outlines our participation in the 2020 MediaEval Predicting Media Memorability Task [10] where memorability is sub-categorised into short-term (after minutes) and long-term (after 24-72 hours) memorability. The task requires participants to create systems that can predict the short-term and long-term memorability of a set of viral videos. The dataset, annotation protocol, pre-computed features, and ground-truth data are described in the overview paper [10]. Previous attempts at computing memorability have consistently demonstrated that off-the-shelf pre-computed features, such as C3D; HMP; LBP; etc., have mostly been unfruitful [5, 7, 11, 24], with captions being the exception [25, 29]. The high usefulness of captions is likely due to the semantically rich nature of text—the medium with the highest cued recall and free recall for narrative [9]—capturing the most information about the elements in the videos with the smallest semantic gap. Recent attempts highlighted the effectiveness of deep features in conjunction with other semantically rich features, such as emotions or actions [2, 28]. To our knowledge, the influence of audio on overall video memorability has yet to be explored.

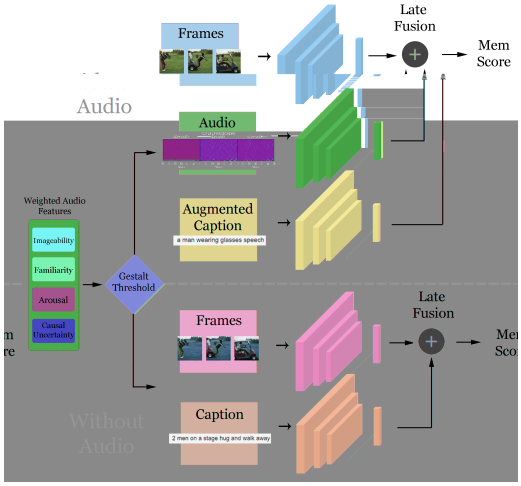

Figure 1: Our multimodal deep-learning based late fusion framework, using conditional audio gestalt based threshold

2 OUR APPROACH

The dataset is comprised of three sets—an initial development set, a supplemental development set, and a test set. Both development sets contain ground-truth (GT) scores and anonymised user annotation data for 590 and 410 videos respectively, while the test set contains 500 videos without GT scores. Upon inspection of the supplemental development set, we noticed that the short-term GT scores were true-negative rates (probabilities that unseen videos would not be falsely remembered), rather than true-positive rates, like the rest of the scores. We accordingly decided to generate our own short-term GT scores by employing collaborative filtering with the provided reaction time annotations. The resulting matrix of predicted reaction times, for each user-video combination, was used to calculate a short-term memorability score for each video. Predicted reaction times that were more than two standard deviations from the mean reaction time were counted as misses. We then divided this new development set into a training (800 videos) and testing set (200 videos). Our approach is predicated on the conditional exclusion/inclusion of audio related features depending on audio gestalt levels.

MediaEval'20, 14-15 December 2020, Online

Audio Gestalt: Defined in [21] as high-level conceptual audio features, imageability; human causal uncertainty (Hcu); arousal; and familiarity, were found to be strongly correlated with audio memorability. We predict audio gestalt by doing a weighted sum of these four features. Imageability is based on whether the audio is music or not using the PANNs [17] network. Hcu and arousal scores are predicted with an xResNet34 pre-trained on ImageNet [8], and fine-tuned on HCU400 [1]. For familiarity we chose to use the top audio-tag confidence score of the PANNs [17] network, as we had observed a correlation between the two scores. These scores were then weighted and summed to produce an audio gestalt score for a video. Depending on a threshold, one of two pathways—without audio, using captions and frames, and with audio, using audio augmented captions, frames, and audio spectrograms—was used to predict memorability scores.

Without Audio: Deep Neural Network (DNN) frame-based and caption-based models were chosen, as they had proven to be quite effective in previous video memorability prediction attempts [25, 29]. For our caption model, given that overfitting was a primary concern, we decided to use the AWD-LSTM (ASGD Weight-Dropped LSTM) architecture [19], as it is highly regularised, and is used in state-of-the-art language modelling. In order to fully take advantage of the high level representations that a language model offers, we used transfer learning. The specific transfer learning method employed was UMLFiT [12], a method that uses discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing to avoid catastrophic forgetting. A language model was pre-trained on the Wiki-103 dataset, and fine-tuned on the first 300,000 captions from Google's Conceptual Captions dataset [23]. The encoder from that fine-tuned language model was then re-used in another model of the same architecture, but trained on 800 of the development set captions to predict short-term and long-term memorability scores rather than the next word in a sentence. For our frame based model, we transfer trained an xResNet50 model pre-trained on ImageNet [8]. It was first fine-tuned on Memento10k [20], then fine-tuned on 800 of the development set videos.

With Audio: We incorporated audio features by augmenting captions with audio tags, and training a CNN on audio spectrograms. We kept the same frame based model as a control. We opted to augment captions with audio tags, rather than relying on audio tags alone, so that the context provided by captions was not lost. We used the PANNs model [17] to extract audio tags, append them to their corresponding development set captions, and re-trained the aforementioned caption model. For the audio spectrogram model, we extracted Mel-frequency cepstral coefficients from the audio, and stacked them with their delta coefficients in order to create three channel spectrogram images. We then used the same transfer training procedure as our frame based model.

For each stream, final predictions were the result of a weighted sum of their constituent model predictions.

3 DISCUSSION AND OUTLOOK

Table 1 shows the scores for our runs when tested on 200 dev set videos kept for validation. Runs 1-3 were intended as controls for our 4th run in which we tested our proposed framework. Run 1 was the "audio only" control, using augmented captions and audio spectrograms; run 2 was the "no audio" control, using captions and video frames; and run 3 was the "everything" control, using all features irrespective of audio gestalt scores. These preliminary result suggest that the audio gestalt-based conditional inclusion of audio features does indeed improve memorability prediction (run 3 vs run 4).

Table 1: Results on 200 dev-set videos kept for validation for each of our runs.

	Short-term	Long-term
Run	**Spearman**	**Spearman**
run1-required	0.345	0.365
run2-required	0.338	0.437
run3-required	0.319	0.425
run4-required	**0.364**	**0.470**
memento10k	0.314	-

Table 2: Official results on test-set for each of our submitted runs.

	Short-term		Long-term	
Run	**Spearman**	Pearson	**Spearman**	Pearson
run1-required	0.054	0.044	**0.113**	0.121
run2-required	0.05	0.072	0.059	0.071
run3-required	-	-	0.109	0.119
run4-required	0.076	0.092	0.041	0.058
memento10k	**0.137**	0.13	-	-

Table 2 shows the scores achieved by each of our submitted runs. Unfortunately, due to a mistake in the submission process, we do not have official short-term results for our 3rd run, and are therefore unable to properly compare it with run 4 and validate our preliminary results. The stark difference between the official results and our preliminary results highlight the inherent difficulty of achieving generalisability when learning from limited data. Unfortunately, there is no way yet of knowing if these differences are simply due the lack of distributional overlap between training and testing videos, or if it is a case of overfitting. Surprisingly, our highest short-term score was achieved by a model not trained on any data from the task—an xResNet50 model pre-trained on ImageNet [8], and fine-tuned on Memento10k [20]. We believe that this result is most likely due to the better generalisability of models trained on larger and more diverse datasets. Our highest long-term score was achieved by our "audio only" run, which indicates that the audio modality does indeed provide some contextual information during video memorability recognition tasks.

Further investigation into the influence of the audio modality on overall video memorability is clearly required. Testing our proposed framework on a much larger video memorability dataset would be an interesting next step towards that goal.

Acknowledgements

This publication has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

REFERENCES

[1] Ishwarya Ananthabhotla, David B Ramsay, and Joseph A Paradiso. 2019. HCU400: An Annotated Dataset for Exploring Aural Phenomenology Through Causal Uncertainty. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 920–924.

[2] 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/

[3] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. 2013. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* 142, 4 (2013), 1323.

[4] James Bigelow and Amy Poremba. 2014. Achilles' ear? Inferior human short-term and recognition memory in the auditory modality. *PloS One* 9, 2 (2014), e89914.

[5] Ritwick Chaudhry, Manoj Kilaru, and Sumit Shekhar. 2018. Show and Recall@ MediaEval 2018 ViMemNet: Predicting Video Memorability. In *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2283/

[6] Michael A Cohen, Todd S Horowitz, and Jeremy M Wolfe. 2009. Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences* 106, 14 (2009), 6008–6010.

[7] Romain Cohendet, Claire-Hélène Demarty, and Ngoc QK Duong. 2018. Transfer Learning for Video Memorability Prediction.. In *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2283/

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[9] Adrian Furnham, Barrie Gunter, and Andrew Green. 1990. Remembering science: The recall of factual information as a function of the presentation mode. *Applied Cognitive Psychology* 4, 3 (1990), 203–212.

[10] 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*.

[11] Rohit Gupta and Kush Motwani. 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features. In *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2283/

[12] 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.

[13] 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.

[14] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 145–152.

[15] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*. 2390–2398.

[16] Edwin A Kirkpatrick. 1894. An experimental study of memory. *Psychological Review* 1, 6 (1894), 602.

[17] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.

[18] Maria Larsson and Lars Bäckman. 1998. Modality memory across the adult life span: evidence for selective age-related olfactory deficits. *Experimental Aging Research* 24, 1 (1998), 63–82.

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

[20] 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.

[21] David Ramsay, Ishwarya Ananthabhotla, and Joseph Paradiso. 2019. The Intrinsic Memorability of Everyday Sounds. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*.

[22] Annett Schirmer, Yong Hao Soh, Trevor B Penney, and Lonce Wyse. 2011. Perceptual and conceptual priming of environmental sounds. *Journal of Cognitive Neuroscience* 23, 11 (2011), 3241–3253.

[23] 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.

[24] Alan F Smeaton, Owen Corrigan, Paul Dockree, Cathal Gurrin, Graham Healy, Feiyan Hu, Kevin McGuinness, Eva Mohedano, and Tomás E Ward. 2018. Dublin's participation in the predicting media memorability task at MediaEval 2018. In *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2283/

[25] Wensheng Sun and Xu Zhang. 2018. Video Memorability Prediction with Recurrent Neural Networks and Video Titles at the 2018 MediaEval Predicting Media Memorability Task.. In *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2283/

[26] Antonia Thelen and Micah M Murray. 2013. The efficacy of single-trial multisensory memories. *Multisensory Research* 26, 5 (2013), 483–502.

[27] Antonia Thelen, Durk Talsma, and Micah M Murray. 2015. Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition* 138 (2015), 148–160.

[28] Duy-Tue Tran-Van, Le-Vu Tran, and Minh-Triet Tran. 2018. Predicting Media Memorability Using Deep Features and Recurrent Network.. In *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2283/

[29] Shuai Wang, Weiying Wang, Shizhe Chen, and Qin Jin. 2018. RUC at MediaEval 2018: Visual and Textual Features Exploration for Predicting Media Memorability.. In *Proceedings of MediaEval 2018, CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2283/