Multi-Modal Ensemble Models for Predicting Video Memorability

Tony Zhao[1], Irving Fang[1], Jeffrey Kim[1], Gerald Friedland[1]
[1]University of California, Berkeley
tonyzhao@berkeley.edu,irvingf7@berkeley.edu,jeffjohn3@berkeley.edu,fractor@eecs.berkeley.edu

ABSTRACT

Modeling media memorability has been a consistent challenge in the field of machine learning. The Predicting Media Memorability task in MediaEval2020 is the latest benchmark among similar challenges addressing this topic. Building upon techniques developed in previous iterations of the challenge, we developed ensemble methods with the use of extracted video, image, text, and audio features. Critically, in this work we introduce and demonstrate the efficacy and high generalizability of extracted audio embeddings as a feature for the task of predicting media memorability.

1 INTRODUCTION AND RELATED WORK

The MediaEval2020 Predicting Media Memorability Task aims to predict how memorable videos are. The dataset consists of videos with short-term and long-term memorability scores, pre-extracted images and video features, and other information described in detail in [5]. Participants predict the probability that each video will be remembered in both the short (after a few minutes) and long (after 24-72 hours) term.

Image-based features extracted from pre-trained convolutional neural networks (CNNs) have been demonstrated as useful in predicting the memorability of videos in the dataset. These features in combination with semantic-embedding and image captioning models have also proved effective for predicting memorability scores [4]. Video-based features, such as C3D and I3D, have also recently been considered in the study of video memorability [13]. Finally, the best performing models of the 2019 iteration of the Predicting Media Memorability challenge utilized ensemble models with the above-mentioned features [1].

2 APPROACH

The training set comprises of 590 short videos (1-8 seconds long) with 2-5 human-annotated captions each. Another development set of 410 additional videos was provided later, but was not used as we discovered the quality of annotated memorability scores to be worse than that of the original set. Each video has corresponding short-term and long-term memorability scores, which were adjusted using methods outlined in previous work [8] [4], namely the following update functions to calculate decay rate α and memorability m_T at duration T

$$\alpha \leftarrow \frac{\sum_{i=1}^{N} \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} \log(\frac{t_j^{(i)}}{T})[x_j^{(i)} - m_T^{(i)}]}{\sum_{i=1}^{N} \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} [\log(\frac{t_j^{(i)}}{T})]^2}$$

$$m_T^{(i)} \leftarrow \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} [x_j^{(i)} - \alpha \log(\frac{t_j^{(i)}}{T})]$$

where we have $n^{(i)}$ observations for video i given by $x^{(i)} \in \{0, 1\}$ and $t_j^{(i)}$ where $x_j = 1$ implies that the repeated video was correctly detected when shown after time t_j. The average t was 74.96, so we adjusted the memorability score for each video to be m_{75} after 10 iterations with α converging to -0.0264.

Models were then trained on a 80-20 training-validation split on video id. Since concatenating multiple multi-modal features resulted in extremely high-dimensional feature vectors which were difficult to train with, we trained various models (Support Vector Regressor, Bayesian Ridge Regressor, Linear Models) on each individual feature independently.

We also explored several prediction aggregation functions for when a video has multiple embeddings of a specific feature, such as several different model predictions due to multiple captions for a given video. While previous work generally defaulted to a simple average, we discovered that taking the median value performed consistently better than either mean, max, or min. In the opposite scenario, where a video has no feature such as a soundless video for audio-based models, we defaulted to using the average of all predicted memorability scores.

We noticed high variance in the performance of the models, which we attributed to the relatively small dataset. Therefore we ran the feature models over 5 random seeds and took the best performing features of each modality (VGGish for audio, ResNet152 for image, C3D for video, GloVe for text). The predictions of these models were then used to perform grid-search over permutations of buckets of 5% to calculate a weighted average as our final ensemble model.

2.1 Audio Features

Using VGGish [7], a pre-trained CNN model (trained on AudioSet), we extracted 128-dimensional embeddings for each second of video audio. Each video had on average had 5.6 embeddings extracted, with one soundless video having none. Bayesian Ridge Regressor provided the best performance on these features.

2.2 Image/Video Features

Local Binary Patterns (LBP), VGG [9], and Convolutional 3D (C3D) proved to have notable performance among the provided features. We discovered that Support Vector Regressor (SVR) [2] models produced the best results with these features.

We also extracted 8 equally spaced frames from each video, which were used for our own image-based feature extraction. The penultimate layer of a pre-trained ResNet-152 [6] (trained on ImageNet) was used to extract a 2048-dimensional feature vector. Similarly,

MediaEval'20, December 14-15 2020, Online

Table 1: Spearman's rank correlation coefficient (SRCC) of features for short-term (ST) and long-term (LT) memorability over 5 runs trained and evaluated on training set (80-20 train-validation split on 590 videos)

Modality	Feature	Model	Mean (ST)	Variance (ST)	Mean(LT)	Variance (LT)
video	**C3D**	SVR	0.152	0.081	0.076	0.059
image	**ResNet152**	SVR	0.233	0.096	0.133	0.066
image	VGG	SVR	0.167	0.058	**0.144**	0.092
image	LBP	SVR	0.139	0.072	0.024	0.067
audio	**VGGish**	Bayesian Ridge	**0.246**	**0.017**	0.059	**0.026**
text	**GloVe**	GRU	0.200	0.029	0.104	0.091

we found that the best performance came from SVR models trained on the extracted features.

Following the methods outlined in [1], pre-trained facial-emotion models were used to extract emotion-based features on the video frames. Ultimately, these features were not used in any of the final models, as only approximately a third of the videos had detectable faces as well as overall poor performance from all emotion-based models.

2.3 Text Features

Provided with several human-annotated captions for each video, we explored several different methods to extract semantic features. Past work suggested that simpler methods like bag-of-words could outperform more sophisticated methods in terms of both effectiveness and efficiency [12]. We vectorized the captions using bag-of-words before training with ordinary least squares, ridge, and lasso regression models. Bag-of-words vectorization and linear models performed the worst among our text-based approaches and were not used in the final model.

We tokenized the captions before extracting 300-dimensional GloVe [11] word embeddings (trained on Wikipedia 2014 and Gigaword 5) to vectorize the caption tokens. These vectors were used as input to train a recurrent neural network with gated recurrent units (GRU) [3]. Our final text-based model had 64 units in the initial GRU layer with a dropout of 0.8, followed by 4 dense layers with a dropout of 0.25 and ReLU activation, trained for 150 epochs with early stopping, a learning rate of 0.001, batch size of 64, and an Adam optimizer.

We also explored machine-generated captions [10] to augment our text-based approaches. After generating captions for each video based on the first, middle, and last frames and mixing the generated captions with the human-annotated captions, we discovered that the performance improvement was insignificant. These automatic captions were not included in our final models.

3 RESULTS AND ANALYSIS

The notable features are seen in Table 1, with the bolded features being selected for our final ensemble models, seen in Table 2 and Table 3. Given the small dataset, we observe relatively high variance in performance. VGGish had the best performance, smallest variance, and thus presumably the best generalization for short-term memorability. However, VGGish features tended to perform poorly for predicting long-term memorability despite maintaining low variance.

Table 2: Short-term memorability ensemble models, Validation SRCC on 20% of training set, Test SRCC on testing set

Model	C3D	ResNet152	VGGish	GloVe	Valid	Test
1	0.20	0.35	0.45	0.00	0.343	**0.136**
2	0.00	0.20	0.35	0.45	0.345	0.116
3	0.05	0.00	0.50	0.45	**0.370**	0.085
4	0.00	0.50	0.15	0.35	0.357	0.091
5	0.35	0.00	0.30	0.35	0.317	0.102

Table 3: Long-term memorability ensemble models, Validation SRCC on 20% of training set, Test SRCC on testing set

Model	C3D	ResNet152	VGGish	GloVe	Valid	Test
1	0.00	0.40	0.15	0.45	**0.289**	0.012
2	0.55	0.10	0.00	0.35	0.192	0.076
3	0.00	0.55	0.45	0.00	0.118	0.044
4	0.25	0.35	0.20	0.20	0.168	**0.077**
5	0.30	0.05	0.00	0.65	0.201	0.056

The difference in validation and test Spearman's rank correlation in our final ensemble models are likely due in part to overfitting after performing grid-search over a relatively small dataset. In addition, we noticed that the quality of annotations between datasets were different, which may cause distribution differences in memorability scores and consequently poorer performance at test time.

4 DISCUSSION AND OUTLOOK

Our main contribution is the demonstration that audio-based models perform well for predicting short-term memorability and generalize much more readily than other methods with the dataset. This may be in part due to the low dimensionality of the extracted audio embeddings. We reiterate that ensembling models of different modalities achieve the best performance as each model represents a different high-level abstraction of the data.

The size of the dataset was notably smaller than similar datasets for the task of predicting media memorability. Future work would ideally iterate on our findings for much larger datasets.

REFERENCES

[1] David Azcona, Enric Moreu, Feiyan Hu, Tomás E. Ward, and Alan F. Smeaton. 2019. Predicting Media Memorability Using Ensemble Models. In *Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019 (CEUR Workshop Proceedings)*, Vol. 2670. CEUR-WS.org. http://ceur-ws.org/Vol-2670/MediaEval_19_paper_15.pdf

[2] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[4] Romain Cohendet, Claire-Helene Demarty, Ngoc Q. K. Duong, and Martin Engilberge. 2019. VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[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] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[7] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://arxiv.org/abs/1609.09430

[8] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. 2015. Understanding and Predicting Image Memorability at a Large Scale. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 2390–2398. https://doi.org/10.1109/ICCV.2015.275

[9] S. Liu and W. Deng. 2015. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 730–734. https://doi.org/10.1109/ACPR.2015.7486599

[10] R. Luo, G. Shakhnarovich, S. Cohen, and B. Price. 2018. Discriminability Objective for Training Descriptive Captions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6964–6974. https://doi.org/10.1109/CVPR.2018.00728

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[12] Alison Reboud, Ismail Harrando, Jorma Laaksonen, Danny Francis, Raphaël Troncy, and Hector Laria Mantecon. 2019. Combining textual and visual modeling for predicting media memorability. In *MediaEval 2019, 10th MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop, 27-29 October 2019, Sophia Antipolis, France*. Sophia Antipolis, FRANCE. http://www.eurecom.fr/publication/6062

[13] Ricardo Manhães Savii, Samuel Felipe dos Santos, and Jurandy Almeida. 2018. GIBIS at MediaEval 2018: Predicting Media Memorability Task. In *Working Notes Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29-31 October 2018 (CEUR Workshop Proceedings)*, Vol. 2283. CEUR-WS.org. http://ceur-ws.org/Vol-2283/MediaEval_18_paper_40.pdf