Predicting Media Memorability using a Pre-Trained Convolutional Neural Network as a Feature Extractor

An approach for maximising short-term and long-term memorability without using descriptive titles (captions) attached to the videos

Colin Hehir
School of Computing
Dublin City University
Dublin, Ireland
colin.hehir3@mail.dcu.ie

ABSTRACT

The ability to be remembered is a significant problem in the media. People are now attempting to organise and allow online media to be accessible in our everyday lives. The CA684 Machine Learning module's Predicting Media Memorability task addresses this issue by posing a challenge to accurately anticipate memorability scores without the descriptive titles attached to the videos[1] [1]. This paper is influenced by the ensemble methods of the 2019 task winner [2]. It follows on from the conclusions observed where a competitive model is presented to predict memorability in short and long-term media without using captions with a pre-trained CNN, namely ResNet152.

CCS CONCEPTS

• Computing Methodologies; Machine Learning Techniques

KEYWORDS

Predicting Media Memorability, Spearman's Rank Correlation Coefficient, HMP, C3D, ResNet152, Convolutional Neural Network (CNN), Bayesian Ridge Regression, Support Vector Regression, Weighted Average Ensemble

1 Introduction and Related Work

The MediaEval Predicting Media Memorability Task's objective is to determine and develop models to assess the kind of media that is visually memorable for viewers. The data set consists of silent video clips, each of which has two memory ratings, which refer to the likelihood of memorability after two memory retention times of short-term (24 hours) and long-term (72 hours).

Recently, Computer Vision has attracted media memorability research interest [3, 4, 5]. When trained on substantial data sets such as ImageNet, Convolutional Neural Networks (CNNs) were more successful at predicting memory values than with multimedia captions and pre-computed features [6, 7]. The MediaEval 2019 entry of DCU has strongly influenced this paper, where at the 2019 competition, DCU performed best by ensembling various models [8].

2 Approach

The memorability dataset comprises 8,000 videos, a test set of 2,000 videos, and a development set of 6,000 videos. The development set consisted of training sets (5400 videos) and validation sets (600 videos). The validation set was used to select hyperparameter tuning, and the efficiency of each model was assessed.

The first frame was extracted from each source video and one frame after each video's seven seconds. Traditional Machine Learning and regularised linear models were developed [7], including Support Vector Regression [9] and Bayesian Ridge Regression [10].

The features used included off-the-shelf pre-computed video specialised features such as C3D (101 features per video) and HMP (6,075 features per video) as a histogram of motion patterns. ResNet152 (16,384 features per video) [11] was used as a feature extractor using transfer learning and a pre-trained model (on ImageNet) before an optimal weighted average ensembling model was selected for both short-term and long-term memorability.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

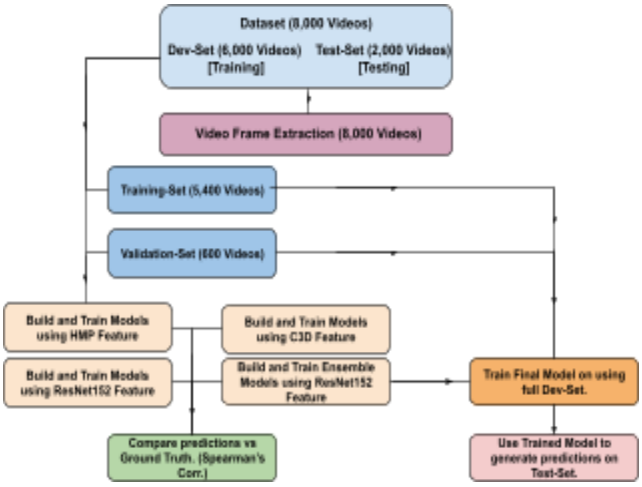

Figure 1: **Process flow diagram of the proposed solution.**

3 Experimental Results and Analysis

The following subsections each present a table of performance results for each feature explored using different models to predict short-term and long-term memorability on a validation set of 600 videos.

3.1 HMP Feature

Table 2 below shows the performance of HMP. This feature is the histogram of motion patterns for each video, where the models used include Bayesian Ridge Regression and Support Vector Regression.

Model	Validation (600 videos)	
	Spearman's Correlation Coefficient	
	Short-Term	Long-Term
Off the shelf pre-computed feature: HMP		
Bayesian Ridge Regression (Default Parameters)	0.274	0.160
Support Vector Regression (Default Parameters)	0.232	0.118
Support Vector Regression (kernel='linear', C=100.0, epsilon =0.001, gamma = 'scale ')	0.267	0.153

Table 1: **Performance of HMP feature for different models.**

3.2 C3D Feature

Table 2 below shows the performance of C3D. This feature is the final classification layer of the C3D model, where the models used include Bayesian Ridge Regression and Support Vector Machine Regression.

Model	Validation (600 videos)	
	Spearman's Correlation Coefficient	
	Short-Term	Long-Term
Off the shelf pre-computed feature: C3D		
Support Vector Regression (Default Parameters)	0.297	0.089
Bayesian Ridge Regression (Default Parameters)	0.325	0.095
Bayesian Ridge Regression (lambda_2=1e-06, alpha_1=0.1, alpha_2=1e-06, normalize=False, lambda_1=0.1')	0.325	0.095

Table 2: **Performance of C3D feature for different models.**

3.3 ResNet152 Feature

Table 3 below shows the performance of ResNet152. This feature is a pre-trained CNN as a feature extractor, also known as transfer learning, where the models used include Bayesian Ridge Regression and Support Vector Regression.

Model	Validation (600 videos)	
	Spearman's Correlation Coefficient	
	Short-Term	Long-Term
Pre-trained CNN as feature extractor: ResNet152		
Support Vector Regression (Default Parameters)	0.477	0.255
Bayesian Ridge Regression (Default Parameters)	0.505	0.308
Support Vector Regression (kernel='rbf', C=0.1, epsilon=0.001, gamma='scale')	0.502	0.267

Table 3: **Performance of ResNet152 feature for different models.**

3.4 Weighted Average Ensemble Model

Table 4 below shows the performance of a weighted average ensemble of the ResNet152 feature using Bayesian Ridge Regression and Support Vector Regression, where the most optimal weights were selected via a trial and error process.

Model	Validation (600 videos)	
	Spearman's Correlation Coefficient	
	Short-Term	Long-Term
Pre-trained CNN as feature extractor: ResNet152		
0.4* Bayesian Ridge Regression (Default Parameters) + 0.6* Support Vector Regression (kernel='rbf', C=0.1, epsilon=0.001, gamma='scale')	0.509	0.291

Table 4: **Performance of a weighted average ensemble of the ResNet152 feature using different models.**

3.5 Final Model Selected

The best performing model for both short-term and long-term memorability (Table 5) was selected to make predictions on the 2,000 videos kept as part of the test-set.

Final Model	Short-Term	Long-Term
	Pre-trained CNN as feature extractor: ResNet152	
	0.4* Bayesian Ridge Regression (Default Parameters) + 0.6* Support Vector Regression (kernel='rbf', C=0.1, epsilon=0.001, gamma='scale')	

Table 5: **The model selected to predict short-term and long-term memorability on the test-set.**

4 Concluding Discussion and Outlook

This paper describes a multimodal weighted average method that may outperform the Predicting Media Memorability Task's best results without using descriptive titles such as captions. Off-the-shelf pre-computed features such as HMP and C3D were explored, and modest predictions for short-term and long-term memorability were made using Bayesian Ridge Regression and Support Vector Regression models. One of this paper's critical contributions is to have demonstrated that using a pre-trained CNN as a feature extractor such as ResNet152 can achieve very high results using Bayesian Ridge Regression. Simultaneously, peak short-term and long-term memorability predictions were found in an ensemble model using Bayesian Ridge Regression and Support Vector Regression. An increase in the number of annotations to the long-term memorability dataset would likely decrease the obvious noise present, resulting in more optimal predictions. Further work should aim to increase this model's accuracy using other models and pre-trained CNNs with the final objective of ensembling it. Also, it can be safely argued that the hyperparameter tuning of the Bayesian Ridge Regression models of the ResNet152 feature will improve these results even further. Other areas to explore in the future include TimeSformer, a completely new video comprehension architecture introduced in March 2021 by Facebook AI. It is the first multimedia architecture focused solely on Transformers. In recent times, this video architecture has become a preferred paradigm to a wide range of applications in natural language processing (NLP) [12].

ACKNOWLEDGMENTS

The writer would like to thank Professor Tomas Ward and Eoin Brophy of Dublin City University for the excellent resources and guidance provided in the preparation of this assignment.

REFERENCES

[1] Constantin, M.G., Ionescu, B., Demarty, C.H., Duong, N.Q., Alameda-Pineda, X. and Sjöberg, M., 2019, October. Predicting Media Memorability Task at MediaEval 2019. In Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France.

[2] Azcona, D., Moreu, E., Hu, F., Ward, T.E. and Smeaton, A.F., 2020, September. Predicting media memorability using ensemble models. CEUR Workshop Proceedings.

[3] Romain Cohendet, Karthik Yadati, Ngoc QK Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicting long-term video memorability. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. ACM, 178–186.

[4] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and recall: Learning what makes videos memorable. In Proceedings of the IEEE International Conference on Computer Vision. 2730–2739.

[5] Hammad Squalli-Houssaini, Ngoc QK Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. 2018. Deep learning for predicting image memorability. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2371–2375.

[6] Duy-Tue Tran-Van, Le-Vu Tran, and Minh-Triet Tran. 2018. Predicting Media Memorability Using Deep Features and Recurrent Network.. In MediaEval

[7] Rohit Gupta and Kush Motwani. 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features.. In MediaEval.

[8] Azcona, D. 2019. Insight@DCU in the Memorability Challenge at MediaEval2019 https://github.com/dazcona/memorability.

[9] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2, 3 (2011), 27.

[10] Christopher M Bishop. 2006. Pattern recognition and machine learning. springer.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[12] Bertasius, G., Wang, H. and Torresani, L., 2021. Is Space-Time Attention All You Need for Video Understanding?. arXiv preprint arXiv:2102.05095