Essex-NLIP at MediaEval Predicting Media Memorability 2020 Task

Janadhip Jacutprakart, Rukiye Savran Kiziltepe, John Q. Gan,
Giorgos Papanastasiou, Alba G. Seco de Herrera
University of Essex, UK
{j.jacutprakart,rs16419,jqgan,g.papanastasiou,alba.garcia}@essex.ac.uk

ABSTRACT

In this paper, we present the methods of approach and the main results from the Essex NLIP Team's participation in the MediEval 2020 Predicting Media Memorability task. The task requires participants to build systems that can predict short-term and long-term memorability scores on real-world video samples provided. The focus of our approach is on the use of colour-based visual features as well as the use of the video annotation meta-data. In addition, hyper-parameter tuning was explored. Besides the simplicity of the methodology, our approach achieves competitive results. We investigated the use of different visual features. We assessed the performance of memorability scores through various regression models where Random Forest regression is our final model, to predict the memorability of videos.

1 INTRODUCTION

The number of published videos has been increasing, and the need for improved content analysis has been of great interest for different areas of research in media memorability. The MediaEval Predicting Media Memorability task focuses on how video contents are memorable to viewers. Participants of the task are expected to develop systems that predict automatically short-term and long-term memorability scores for given video samples [8]. The task was introduced in 2018 with a soundless dataset including 10,000 short videos [3]. In 2019 the task continued to explore the short-term and long-term video memorability, during which people were watching it for an extended period with no sound videos [5]. However, this year, García et al. [8] have released a new dataset based on real-world data with motion and audio information within videos. The dataset, annotation collection procedure, pre-computed features, and ground truth data are described in the task overview paper [8].

This article describes the participation of the Essex-NLIP[1] research group in the MediaEval Predicting Media Memorability 2020 task. The Essex-NLIP group participated in this task in 2019 for the first time [10]. In 2020, the team focuses on the use of visual features based on colour and based on the video annotation metadata. Hyper-parameter tuning was also explored.

[1]Essex-NLIP is the Natural Language and Information Processing research group at the University of Essex, UK (see https://essexnlip.uk/).

MediaEval'20, December 14-15 2020, Online

2 RELATED WORK

In the 2019 MediaEval task, various types of regression models were used to explore the use of image and video features for predicting media memorability. Dos Santos and Almeida [7] used K-Nearest Neighbour Regression (KNR) and Leyva et al. [10] proposed the method of Support Vector Regression (SVR), using a regularised method and LassoLarsCV. SVR and Bayesian Ridge Regression using an ensemble method to detect the capability of the predictions on each model was employed by Azcona et al. [1]. In contrast to the above complex regressions, Wang et al. [18] used Random Forest regression and SVR. In this work, we use Random Forest after exploring several regression models (see Section 3.3).

In the last 20 years, many visual video descriptors have been explored [4, 11, 12]. A set of pre-computed visual descriptors were provided in this task for the videos in the collection (see Section 3.1).

In order to combine features, both Rattani et al. [14] and Ross and Govindarajan [15] used a simple concatenated method to fuse different features.In this work, we also used the concatenate method to fuse the descriptor we have on each feature together (see Section 3.2)

Bergstra and Bengio [2] implemented the hyper-parameter optimisation method using both RandomizedsearchCV and GridsearchCV similar to our approach in this work (see Section 3.3).

3 APPROACH

This section describes the basic techniques that we used on the 5 runs, submitted for this work. The selected features in Section 3.2 were explored using Random Forest regression described in Section 3.3 in order to obtain the memorability score. Figure 3 shows an overview of the process.

Figure 1: Overview of the approach used for predicting the media memorability score.

3.1 Dataset

In 2020, the collection is composed of 1500 short videos. A set of pre-extracted features were also distributed, including seven visual features. More details about the task and the collection can be found at García et al. [8].

3.2 Features

All the pre-computed visual features provided by the task were explored (AlexNetFC7, HOG, HSVHist, RGBHist, LBP, VGGFC7, C3D)

Table 1: Short-term and Long-term Spearman's correlation scores achieved by the proposed runs on the development set and the test set. The results are compared with the Mean and Variance of the results from the other participants in the MedieEval 2020 Predicting Media Memorability task.

			Short-term		Long-term	
Run	Feature	Parameter Tuning	Devset	Testset	Devset	Testset
Run1	HSV	Yes	0.415	0.042	0.419	0.032
Run2	RGB	No	0.455	-0.003	0.387	0.043
Run3	RGB	Yes	0.428	-0.015	0.391	0.032
Run4	RGB&HSV	Yes	0.463	-0.022	0.422	-0.017
Run5	Descriptive	Yes	0.508	0.02	0.001	-0.054
Mean	-	-	-	0.058	-	0.036
Variance	-	-	-	0.002	-	0.002

in combination with several regression models (see Section 3.3). For this paper, *RGBHist* and *HSVHist* were selected based on the results obtained on the experiments on the development set. Both RGBHist and HSVHist are colour histogram-based which extract vectorised pixels within a certain neighbourhood across each pixel, through either RGB colour channels or HSV (known as Hue / Saturation / Value colour). The descriptors were concatenated based on colour channels in order to preserve the data format of the features. For one of the experiments, we fused both, RGBHist and HSVHist, by concatenating both descriptors.

In addition, the metadata provided with the annotation was used as a video descriptor, which we call it "Descriptive". It contains the average value of video position and the number of annotations occurs per video.

3.3 Regression Model

After exploring several regression models (Random Forest [18], Decision Tree [17], Gradient Boosting [6], Extra Tree [16] and Sequential regression models [9]), in this work, we used Random Forest based on the results we obtained from the development set.

As for Random Forest regression, we explore performance based on both the default parameters and the hyper-parameter tuning method. RandomizedsearchCV is a method that chooses random numbers of hyper-parametric pairs from a given domain. GridsearchCV is a method that executed a complete search on the predefined parameter values for an estimator and returns the best result obtained from hyper-parametric combinations [13]. We have optimised the hyper-parameters for our GridsearchCV, based on the results of RandomizedsearchCV we acquired.

4 RESULTS AND ANALYSIS

This year, Essex-NLIP submitted 5 runs for both short-term and long-term, using the techniques described in Section 3:

- *Run 1* - This run uses *HSVHist* descriptor and hyper-parameter tuning.
- *Run 2* - This run uses *RGBHist* descriptor and no hyper-parameter tuning.
- *Run 3* - This run uses *RGBHist* descriptor and hyper-parameter tuning.
- *Run 4* - This run uses both *RGBHist* and *HSVHist* descriptors and hyper-parameter tuning.
- *Run 5* - This run uses the *Descriptive* descriptor and hyper-parameter tuning.

Table 1 presents the results from the development and test sets for both short-term and long-term memorability using Random Forest regression and the following features: RGBHist, HSVHist and Descriptive. Table 1 also indicates that the results achieved in this year challenge were very low, considering the mean and variance of participants' results. Besides using a simple approach, two of the submitted runs achieved competitive results over the test set. For short-term memorability best result was achieved with *Run 1* when using HSVHist and hyper-parameter tuning. In the case of long-term memorability, the best result was obtained on *Run 2* when using RGBHist without hyper-parameter tuning.

The results obtained on the development set were considerably higher compared to the ones achieved on the test set. The highest Spearman's correlation score (0.508) on the development set was achieved with *Run 5*. The best result for long-term memorability over the development set was obtained by *Run 4* when using the fusion features of RGBHist & HSVHist with hyper-parameter tuning. It obtained a 0.422 Spearman's correlation score. Further investigation is needed to increase the prediction performance on the test set.

Run 5 uses Descriptive feature presented in Section 3.2. Results indicated that taking into account only the number of annotations based on the video positions influences in how memorable a video is, even without considering any further video descriptor.

5 DISCUSSION AND OUTLOOK

This article describes the methods and results of the Essex-NLIP team for the MediaEval 2020 Predicting Media Memorability task. Five runs were submitted for both short-term and long-term memorability using Random Forest regression. After exploring all the features provided by the task organisation, we worked on colour-based features and metadata on the video position annotation which achieved the highest score on our development set. he results on the development set were higher compared to the test set, due to differences in the data set size (development set was larger). Besides the simplicity of the proposed approach, it achieved competitive results whilst explored how the video position and the number of annotations can affect the memorability score.

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 (CEUR Workshop Proceedings)*, Vol. 2670.

[2] James Bergstra and Yoshua Bengio. 2012. Random Search for Hyper-Parameter Optimization. *The Journal of Machine Learning Research* 13, 1 (2012), 281–305.

[3] Romain Cohendet, Claire Hélène Demarty, Ngoc Q.K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh Toan Do. 2018. MediaEval 2018: Predicting Media Memorability. In *Working Notes Proceedings of the MediaEval 2018 Workshop (CEUR Workshop Proceedings)*, Vol. 2283.

[4] Miguel T Coimbra and JP Silva Cunha. 2006. MPEG-7 Visual Descriptors—Contributions for Automated Feature Extraction in Capsule Endoscopy. *IEEE Transactions on Circuits and Systems for Video Technology* 16, 5 (2006), 628–637.

[5] Mihai Gabriel Constantin, Bogdan Ionescu, Claire Hélène Demarty, Ngoc Q.K. Duong, Xavier Alameda-Pineda, and Mats Sjöberg. 2019. The Predicting Media Memorability Task at MediaEval 2019. In *Working Notes Proceedings of the MediaEval 2019 Workshop (CEUR Workshop Proceedings)*, Vol. 2670.

[6] Xu Dazhan, Wu Xiaoyu, and Sun Guoquan. 2020. Image Memorability Prediction Based on Machine Learning. In *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)*. IEEE, 91–94.

[7] Samuel Felipe Dos Santos and Jurandy Almeida. 2019. GIBIS at MediaEval 2019: Predicting Media Memorability Task. In *Working Notes Proceedings of the MediaEval 2019 Workshop (CEUR Workshop Proceedings)*, Vol. 2670.

[8] 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 (CEUR Workshop Proceedings)*.

[9] Nikhil Ketkar. 2017. *Deep Learning with Python*. Springer. 97–111 pages.

[10] Roberto Leyva, Faiyaz Doctor, Alba García Seco de Herrera, and Sohail Sahab. 2019. Multimodal Deep Features Fusion For Video Memorability Prediction. In *Working Notes Proceedings of the MediaEval 2019 Workshop (CEUR Workshop Proceedings)*, Vol. 2670.

[11] Ionuţ Mironică, Ionuţ Cosmin Duţă, Bogdan Ionescu, and Nicu Sebe. 2016. A Modified Vector of Locally Aggregated Descriptors Approach for Fast Video Classification. *Multimedia Tools and Applications* 75, 15 (2016), 9045–9072.

[12] Jens-Rainer Ohm, F Bunjamin, Wolfram Liebsch, Bela Makai, Karsten Müller, Aljoscha Smolic, and D Zier. 2000. A Set of Visual Feature Descriptors and their Combination in a Low-Level Description Scheme. *Signal Processing: Image Communication* 16, 1-2 (2000), 157–179.

[13] 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 (2011), 2825–2830.

[14] Ajita Rattani, Dakshina Ranjan Kisku, Manuele Bicego, and Massimo Tistarelli. 2007. Feature Level Fusion of Face and Fingerprint Biometrics. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*. IEEE, 1–6.

[15] Arun A Ross and Rohin Govindarajan. 2005. Feature Level Fusion Using Hand and Face Biometrics. In *Biometric Technology for Human Identification II*, Vol. 5779. International Society for Optics and Photonics, 196–204.

[16] Jaak Simm, Ildefons Magrans De Abril, and Masashi Sugiyama. 2014. Tree-Based Ensemble Multi-Task Learning Method for Classification and Regression. *IEICE Transactions on Information and Systems* 97, 6 (2014), 1677–1681.

[17] 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.

[18] Shuai Wang, Linli Yao, Jieting Chen, and Qin Jin. 2019. RUC at MediaEval 2019: Video Memorability Prediction Based on Visual Textual and Concept Related Features. In *Working Notes Proceedings of the MediaEval 2019 Workshop (CEUR Workshop Proceedings)*, Vol. 2670.