GIBIS at MediaEval 2019: Predicting Media Memorability Task

Samuel Felipe dos Santos and Jurandy Almeida

GIBIS Lab, Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo – UNIFESP
12247-014, São José dos Campos, SP – Brazil
{felipe.samuel,jurandy.almeida}@unifesp.br

ABSTRACT

This paper presents the GIBIS team experience in the *Predicting Media Memorability Task* at MediaEval 2019. In this task, the teams were requested to develop an approach to predict a score reflecting whether videos are memorable or not, considering short-term memorability and long-term memorability. Our proposal relies on late fusion of multiple regression models learned with both hand-crafted and data-driven features and by different regression algorithms.

1 INTRODUCTION

People's experience in watching a video is essential to making it remembered or forgotten after a while. Due to this subjectiveness, the challenging task of automatically predicting whether a video is memorable or not has attracted a lot of attention. Since 2018, the *Predicting Media Memorability Task* [4] at MediaEval has been challenging participants to assign a memorability score for a video reflecting its probability to be remembered. For this, it is provided a dataset composed of 10,000 short, soundless videos, which are splitted into 8,000 videos for the development set and 2,000 videos for the test set. For more details about this task, please, refer to [4].

In this paper, we describe the work developed by the GIBIS team in the context of the MediaEval 2019 Predicting Media Memorability Task. Our starting point was the approach we proposed last year [8]. Roughly speaking, it relies on regression models learned with hand-crafted and data-driven features and by different regression algorithms. This year we focused on improving our previous approach by exploiting new features, regressors, and late fusion.

2 APPROACH

Both short-term and long-term memorability subtasks were approached with the same strategies. The starting point for our proposal is the work of Savii et al. [8], where visual features were extracted from videos and then used to train regression models.

Different visual features were evaluated by our approach: (1) hand-crafted motion features extracted with HMP[1] (*Histogram of Motion Patterns*) [1] and (2) data-driven features learned with I3D[2] (*Inflated 3D ConvNet*) [3]. One limitation of I3D is its capacity to capture subtle but long-term motion dynamics, as it requires to break a video into small clips. Unlike I3D, HMP captures motion dynamics of a video as a whole, and not just parts.

HMP [1] considers the video movement by the transitions between frames. For each frame, motion features are extracted from the video stream. After that, each feature is encoded as a unique pattern, representing its spatio-temporal configuration. Finally, those patterns are accumulated to form a normalized histogram.

I3D [3] generalizes a 2D ConvNet into a 3D ConvNet. For that, 2D convolutional filters of the Inception-V1 [5] architecture are *inflated* into 3D convolutions, thus adding a temporal dimension. The I3D model was first initialized by repeating and rescaling the weights of the Inception-V1 model pre-trained on ImageNet and then trained on the Kinetics Human Action Video Dataset[3] [3]. To extract the I3D features, the classification layers of this pre-trained model were replaced by a global average pooling layer. Next, each video was resized to 256×256 resolution and then splitted into 64-frame clips with an overlap of 32 frames between two consecutive clips. After that, a single center crop with size 224×224 was extracted from each of those clips and passed through the network, producing multiple I3D features for each video. Finally, different strategies were used to combine clip-based features into a single video representation: (1) *average*, where the multiple I3D features are averaged; and (2) *concatenation*, where they are concatenated together.

Each of the above features was used as input to train different regression algorithms: (1) KNR (*k-Nearest Neighbor Regressor*) and (2) SVR (*Support Vector Regression*) [7]. The KNR and SVR implementations from the scikit-learn python package[4] [7] were used for easy reproducibility. For training such regressors, we first divided the development set into training and validation sets, with an 80%-20% split. Then, we randomly splitted the training set into n equal-size subsets and trained one regression model for each subset, thus obtaining n different regression models. Next, they were combined as an ensemble model to predict memorability scores for the videos in both validation and test sets. For that, the final score was computed by averaging their individual scores and we used the 95% confidence interval as the output confidence. In our experiments, the values tested for n were 1, 5, and 10. For KNR, the values tested for the parameter k were 1, 3, and 5. For SVR, we used RBF kernel with the parameter ϵ set to 0.1 and values ranging from 0.5 to 16 with step of 0.5 were tested for the C parameter.

Besides individual predictions provided by different combinations of features and regressors, we also explored late fusion for combining the top performing regression models learned with different features, by different regression algorithms, and using different hyperparameter settings. For that, we adopted the strategy proposed by Almeida et al. [2]. First, individual regression models obtained by all the different configurations (i.e., combination of features, regressors, and hyperparameter settings) were sorted in an decreasing order of their performance on the validation set according to the official metric for the task. Then, each of those individual

[1] https://github.com/jurandy-almeida/hmp (As of September, 2019)
[2] https://github.com/deepmind/kinetics-i3d (As of September, 2019)

[3] In this work, we used the I3D model pre-trained on Kinetics with RGB data only.
[4] https://scikit-learn.org/ (As of September, 2019)

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

regression models was selected according to its rank, i.e., the best was the first, the second best was the second, and so on. At each step, the next model was combined with all the previous ones by averaging their individual scores. This process was repeated until the performance degrades. At the end, the best set of regression models for the validation set was selected by this procedure and then used to predict memorability scores for videos in the test set.

Finally, we evaluted the use of the I3D model as a quantile regressor instead of a feature extractor. For that, we changed its output layer to have only 3 neurons representing the quantiles τ of 0.1, 0.5 and 0.9. The 0.5 quantile corresponds to the median and was taken as the memorability score whereas the other two were used to calculate the output confidence. The resulting model was initilialized with weights pre-trained on the Kinetics dataset and fine-tuned on the training set for 10 epochs with stochastic gradient descent using learning rate of 0.1, batch size of 20, and quantile loss function [6].

3 RESULTS AND ANALYSIS

Five different runs were submitted for each subtask. They were configured as shown in Table 1. The first three runs refer to the best parameter setting for each combination of feature & regressor in isolation, the fourth run refers to late fusion of the top performing feature & regressor combinations, and the last run refers to the deep quantile regression with the I3D model. All the evaluated approaches were calibrated on the development set using a holdout method (80% train/20% test). The evaluation metrics are: Spearman's rank correlation, Pearson correlation coefficient, and MSE (Mean Squared Error). The former is the official metric for the task.

Table 1: Configuration of the submitted runs.

Subtask	Run	Configuration
Long-term memorability	1	HMP & KNR($k = 1$) with $n = 1$
	2	$\text{I3D}_{\text{feature}}^{\text{average}}$ & KNR($k = 5$) with $n = 10$
	3	$\text{I3D}_{\text{feature}}^{\text{concatenation}}$ & KNR($k = 5$) with $n = 10$
	4	Late Fusion[5] (same as run 2)
	5	$\text{I3D}_{\text{regressor}}$
Short-term memorability	1	HMP & KNR($k = 3$) with $n = 10$
	2	$\text{I3D}_{\text{feature}}^{\text{average}}$ & KNR($k = 5$) with $n = 5$
	3	$\text{I3D}_{\text{feature}}^{\text{concatenation}}$ & SVR($C = 16$) with $n = 1$
	4	Late Fusion (of the six best combinations): $\text{I3D}_{\text{feature}}^{\text{average}}$ & KNR($k = 5$) with $n = 5$ $\text{I3D}_{\text{feature}}^{\text{average}}$ & KNR($k = 3$) with $n = 5$ $\text{I3D}_{\text{feature}}^{\text{average}}$ & KNR($k = 3$) with $n = 10$ $\text{I3D}_{\text{feature}}^{\text{average}}$ & SVR($C = 1$) with $n = 1$ $\text{I3D}_{\text{feature}}^{\text{average}}$ & SVR($C = 0.5$) with $n = 1$ $\text{I3D}_{\text{feature}}^{\text{average}}$ & SVR($C = 10$) with $n = 10$
	5	$\text{I3D}_{\text{regressor}}$

[5] The run 4 from the long-term memorability subtask was not submitted, since no performance gain was obtained on combining the best model in isolation with the other ones, being therefore identical to the run 2.

Table 2 presents the results for the development and test sets in the long-term memorability subtask. Our best result on the development set was obtained by $\text{I3D}_{\text{feature}}^{\text{average}}$ using an ensemble of $n = 10$ KNR($k = 5$), achieving a Spearman value of 0.213. In contrast, $\text{I3D}_{\text{feature}}^{\text{concatenation}}$ with an ensemble of $n = 10$ KNR($k = 5$) achieved the best result on the test set, yielding a Spearman value of 0.199.

Table 2: Long-term memorability results.

Set	Run	Spearman	Pearson	MSE
Dev. Set	1	0.091	0.101	0.04
	2	0.213	0.219	0.02
	3	0.189	0.203	0.02
	4	0.213	0.219	0.02
	5	0.071	0.077	0.02
Test Set	1	0.015	0.019	0.04
	2	0.197	0.214	0.02
	3	0.199	0.214	0.02
	4	0.197	0.214	0.02
	5	0.111	0.137	0.02

Table 3 presents the results for the development and test sets in the short-term memorability subtask. Our best result on both sets was obtained by the late fusion of the six best models among all the combinations of features & regressors, achieving a Spearman value of 0.453 for the development set and 0.438 for the test set.

Table 3: Short-term memorability results.

Set	Run	Spearman	Pearson	MSE
Dev. Set	1	0.215	0.256	0.01
	2	0.434	0.474	0.01
	3	0.416	0.454	0.01
	4	0.453	0.491	0.01
	5	0.262	0.281	0.01
Test Set	1	0.249	0.259	0.01
	2	0.417	0.46	0.01
	3	0.398	0.443	0.01
	4	0.438	0.477	0.01
	5	0.247	0.25	0.01

4 DISCUSSION AND OUTLOOK

In general, I3D performed better than HMP as feature extractor. The results for $\text{I3D}_{\text{feature}}^{\text{average}}$ and $\text{I3D}_{\text{feature}}^{\text{concatenation}}$ were similar with a small advantage to the first. An intent for future work is to analyze the use of smarter strategies for combining the clip-based features extracted with the I3D model, like RNNs (Recurrent Neural Networks).

Late fusion of the top performing models achieved our best results in the short-term subtask, but for the long-therm subtask it did not lead to performance gain. As future work, we also plan to evaluate different fusion strategies, for instance, the use of SVM to learn how to combine features and regressors effectively.

The performance of using the I3D model as a regressor was lower than expected. One of the reasons might be the small volume of data available for training and/or our choices for hyperparameters, since they were chosen arbitrarily. We want to conduct a deeper investigation of strategies to overcome those issues in future.

ACKNOWLEDGMENTS

This research was supported by the São Paulo Research Foundation - FAPESP (grant #2018/21837-0), the FAPESP-Microsoft Research Virtual Institute (grant #2017/25908-6), and the Brazilian National Council for Scientific and Technological Development - CNPq (grants #423228/2016-1 and #313122/2017-2).

REFERENCES

[1] J. Almeida, N. J. Leite, and R. S. Torres. 2011. Comparison of Video Sequences with Histograms of Motion Patterns. In *IEEE International Conference on Image Processing (ICIP'11)*. Brussels, Belgium, 3673–3676.

[2] J. Almeida, D. C. G. Pedronette, B. C. Alberton, L. P. C. Morellato, and R. S. Torres. 2016. Unsupervised Distance Learning for Plant Species Identification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9, 12 (2016), 5325–5338.

[3] J. Carreira and A. Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. Honolulu, HI, USA, 4724–4733.

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

[5] S. Ioffe and C. Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML'15)*. Lille, France, 448–456.

[6] R. Koenker. 2005. *Quantile Regression*. Cambridge University Press.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[8] R. M. Savii, S. F. dos Santos, and J. Almeida. 2018. GIBIS at MediaEval 2018: Predicting Media Memorability Task. In *Proc. of the MediaEval 2018 Workshop*. Sophia Antipolis, France.