DUT-MMSR at MediaEval 2017

Predicting Media Interestingness Task

Reza Aditya Permadi, Reza; Septian Gilang Permana Putra, Septian; Helmiriawan, Helmi; Liem, Cynthia

Publication date
2017

Document Version
Final published version

Published in
Working Notes Proceedings of the MediaEval 2017 Workshop

Citation (APA)
Reza Aditya Permadi, R., Septian Gilang Permana Putra, S., Helmiriawan, H., & Liem, C. (2017). DUT-MMSR at MediaEval 2017: Predicting Media Interestingness Task. In G. Gravier, B. Bischke , C-H. Demarty, M. Zaharieva, M. Riegler, E. Dellandrea, D. Bogdanov, R. Sutcliffe, G. J. F. Jones, & M. Larson (Eds.), *Working Notes Proceedings of the MediaEval 2017 Workshop* (pp. 1-3). (CEUR Workshop Proceedings; Vol. 1984).

DUT-MMSR at MediaEval 2017: Predicting Media Interestingness Task

Reza Aditya Permadi, Septian Gilang Permana Putra, Helmiriawan, Cynthia C. S. Liem
Delft University of Technology, The Netherlands
{r.a.permadi,septiangilangpermanaputra,helmiriawan}@student.tudelft.nl,c.c.s.liem@tudelft.nl

ABSTRACT

This paper describes our approach for the submission to the Mediaeval 2017 Predicting Media Interestingness Task, which was particularly developed for the Image subtask. An approach using a late fusion strategy is employed, combining classifiers from different features by stacking them using logistic regression (LR). As the task ground truth was based on pairwise evaluation of shots or keyframe images within the same movie, next to using precomputed features as-is, we also include a more contextual feature, considering averaged feature values over each movie. Furthermore, we also consider evaluation outcomes for the heuristic algorithm that yielded the highest MAP score on the 2016 Image subtask. Considering results obtained for the development and test sets, our late fusion method shows consistent performance on the Image subtask, but not on the Video subtask. Furthermore, clear differences can be observed between MAP@10 and MAP scores.

1 INTRODUCTION

The main challenge of the Media Interestingness task is to rank sets of images and video shots from a movie, based on their interestingness level. The evaluation metric of interest for this task is the Mean Average Precision considering the first 10 documents (MAP@10). A complete overview of the task, along with the description of the dataset, is given in [4].

Due to the similarity of this year's task to the 2016 Predicting Media Interestingness task, we considered the strategies used in submissions to last year's task to inform the strategy of our submission to this year's task. [3] and [2] both use an early fusion strategy by combining features that perform relatively well individually. A late fusion strategy with average weighting is used in [6], combining classifiers from different modalities. A Support Vector Machine (SVM) is used as the final combining classifier. [8] finds that logistic regression gives good results, using CNN features which have been transformed by PCA.

[7] proposed a heuristic approach, based on observing clear presence of people in images. This approach performed surprisingly well, even yielding the highest MAP score in the 2016 Image subtask. While we will mainly focus on a fusion-based approach this year, we will also include results of the best-performing configuration from [7] in unaltered form, so a reference point to state-of-the-art of last year is retained.

2 APPROACH

We would like to devise a strategy which is computationally efficient and yet gives a good interpretability of the results. Our approach therefore consists of a fairly straightforward machine learning pipeline, evaluating performance of individual features first, and subsequently applying classifier stacking to find the best combinations of the best-performing classifiers on the best-performing features. For our implementation, we use sklearn [9]. As base classifier, we use logistic regression and aim to find optimal parameters. Further details of our approach are described below.

2.1 Features of interest

For the Image subtask, we initially consider all the pre-computed visual features from [5]: Color Histogram (HSV), LBP, HOG, GIST, denseSIFT, and the Alexnet-based features (fully connected (fc7) layer and probability output). In all cases, we consider pre-computed features and their dimensionalities as-is.

Considering the way in which ground truth for this task was established, human raters were asked to perform pairwise annotations on shots or keyframe images from the same movie. Hence, it likely that overall properties of the movie may have affected the ratings: for example, if a movie consistently is shot in a dark background, a dark shot may not stand out as much as in another movie. In other words, we assume that the same feature vector may be associated to different interestingness levels, depending on the context of the movie it occurs in. Therefore, apart from the pre-computed features by the organizers, we also consider a contextual feature, based on the average image feature values per movie.

Let X_i be an $m \times n$ feature matrix for a movie, where m is the number of images offered for the movie, and n the length of the feature vector describing each image. For our contextual feature, we then take the average value of X_i across its columns, yielding a new vector μ_i of size $1 \times n$. In our subsequent discussion, we will denote the contextual feature belonging to a feature type F by 'mean$F$' (e.g. HSV $\rightarrow$ meanHSV). This feature is then concatenated to the original feature vector.

For the Video subtask, in comparison to the Image subtask, we now also have information from the audio modality in the form of Mel-Frequency Cepstral Coefficients (MFCC). We further use the pre-computed fully-connected layer (fc6) of a C3D deep neural network [10]. As pre-computed features are given at the keyframe resolution, we simply average over all the keyframes to obtain the values for a particular feature representation. Again, we consider pre-computed features and their dimensionalities as-is.

2.2 Individual feature evaluation

For each feature type, we would like to individually find the best-performing classifier that will optimize the MAP@10 value. Before feeding the feature vector into the classifier, the values are scaled to have zero mean and unit variance, considering the overall statistics

MediaEval'17, 13-15 September 2017, Dublin, Ireland

Table 1: Evaluation on the test and development set

				Development Set		Test Set	
SubTask	**Run**	**Features**	**Late Fusion Classifier**	**MAP@10**	**MAP**	**MAP@10**	**MAP**
Image	1	[(lbp, meanlbp), (hog), (fc7)]	Logistic Regression, C=1	0.1387	0.3021	0.1310	0.3002
	2		Logistic Regression, C=100	**0.1390**	**0.3050**	**0.1385**	**0.3075**
	3		SVM polynomial kernel, degree=2, gamma=0.01	0.1373	0.3028	0.1349	0.3052
	4		SVM RBF kernel, gamma=0.01, C=0.01	0.1322	0.2972	0.1213	0.2887
	5	2016 unmodified heuristic feature [7]		0.0508	0.1864	0.0649	0.2105
Video	1	[(gist), (c3d)]	Logistic Regression, C=1000	0.1040	0.2295	0.0443	0.1734
	2		Logistic Regression, C=10	**0.1045**	**0.2299**	0.0465	0.1748
	3	[(hog), (c3d)]	Logistic Regression, C=1	0.1025	0.2249	0.0478	0.1770
	4		Logistic Regression, C=10	0.1025	0.2253	0.0482	0.1783
	5	2016 unmodified heuristic feature [7]		0.0530	0.1494	**0.0516**	**0.1791**

Table 2: Best individual feature classifier using LR

Task	**Feature**	**Size**	**C**	**MAP@10**
Image	lbp, meanlbp	118	10^{-3}	0.1145
	hog	300	10^{-8}	0.1092
	fc7	4096	10^{-5}	0.1081
Video	gist	512	10^{-4}	0.0767
	c3d	4096	10^{-5}	0.1017
	hog	300	10^{-5}	0.0744

of the training set. For logistic regression, the optimal penalty parameter C is searched on a logaritmic scale from 10^{-9} until 10^{0}.

To evaluate our model, 5-fold cross validation is used. Each fold is created based on the number of movies in the dataset, rather than on the number of individual instances of images or videos within a movie. This way, we make sure that training or prediction always considers all instances offered for a particular movie. For each evaluation, the cross validation procedure is run 10 times to avoid dependence on specific fold compositions, and the average MAP@10 value is considered.

2.3 Classifier stacking

After identifying the best classifier configuration per feature type, we stack the output of those classifiers and try different combinations of them, which are then trained again with several classifiers (logistic regression, SVM, AdaBoost, Linear Discriminant Analysis, and Random Forest). Finding that logistic regression and SVM perform quite well, we apply a more intensive grid search on these classifier types to optimize parameters.

3 RESULTS AND CONCLUSIONS

Table 2 illustrates the top performance on the development set for individual classifiers, which will then also be considered in the test set. In several cases, addition of our contextual averaged feature yields slight improvements over the individual feature alone. This improvement is the biggest for LBP where there is an increase of 0.68 compared to the original features.

The final evaluation results for 5 different runs per subtask are shown in Table 1. As before, late fusion makes use of the output probability of the best classifiers trained on each feature (as in Table 2), rather than the feature values themselves. Our best-performing result on the Image development set is a MAP@10 value of 0.139 (Logistic Regression, $C = 100$). This is an improvement over the performance of the best individual classifier in Table 2. On the test set, our best result on the Image subtask is a MAP@10 value of 0.1385 for the same classifier configuration.

For the Video subtask, evaluation results on the test set show considerable differences in comparison to the development set. While somewhat surprising, we did notice considerable variation in results during cross-validation, and our reported development set results are an average of several cross-validation runs. As one particularly bad cross-validation result, using late fusion of GIST and c3d features with logistic regression ($C = 10$), with videos 0, 3, 6, 9, 12, 22, 23, 27, 28, 30, 46, 50, 60, 69, 71 as the evaluation fold, we only obtained a MAP@10 value of 0.0496.

Considering the results for the best-performing configuration (*histface*) of the heuristic approach from [7], we notice clear differences between MAP@10 and MAP as a metric. Generally spoken, the heuristic approach is especially outperformed on MAP@10, implying that clear presence of people is not the only criterion for the top-ranked items. Comparing results for this approach to our proposed late fusion approach, the late fusion approach consistently outperforms a heuristic approach on the Image subtask, but in the Video subtask, the heuristic approach still has reasonable scores, and outperforms the late fusion approach on the test set.

In conclusion, we employed the offered pre-computed features and included a contextual averaged feature, and then proposed a late fusion strategy based on the best-performing classifier settings for the best-performing features. Using fusion shows improvements over results obtained on individual features. In future work, alternative late fusion strategies as explained in [1] may be investigated.

For the Image subtask, we notice consistent results between the development and test set. However, on the Video subtask, we notice inconsistent results on the development set in comparison to the test set. Predicting interestingness in video likely needs a more elaborated approach that we have not yet covered thoroughly in our method. It also might be the case that the feature distribution of the test set turned out different from that of the training set, or that generally, the distribution of features across a video should be taken into account in more sophisticated ways, for example by taking into account temporal development aspects.

REFERENCES

[1] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16, 6 (01 Nov 2010), 345–379. https://doi.org/10.1007/s00530-010-0182-0

[2] Shizhe Chen, Yujie Dian, and Qin Jin. 2016. RUC at MediaEval 2016: Predicting Media Interestingness Task. In *MediaEval 2016 Working Notes Proceedings*.

[3] Mihai Gabriel Constantin, Bogdan Boteanu, and Bogdan Ionescu. 2016. LAPI at MediaEval 2016 Predicting Media Interestingness Task. In *MediaEval 2016 Working Notes Proceedings*.

[4] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Michael Gygli, and Ngoc QK Duong. Mediaeval 2017 Predicting Media Interestingness Task. In *MediaEval 2017 Working Notes Proceedings*.

[5] Yu-Gang Jiang, Qi Dai, Tao Mei, Yong Rui, and Shih-Fu Chang. 2015. Super Fast Event Recognition in Internet Videos. *IEEE Transactions on Multimedia* 17, 8 (Aug 2015), 1174–1186.

[6] Vu Lam, Tien Do, Sang Phan, Duy-Dinh Le, and Duc Anh Duong. 2016. NII-UIT at MediaEval 2016 Predicting Media Interestingness Task.. In *MediaEval 2016 Working Notes Proceedings*.

[7] Cynthia C. S. Liem. 2016. TUD-MMC at MediaEval 2016: Predicting Media Interestingness Task.. In *MediaEval 2016 Working Notes Proceedings*.

[8] Jayneel Parekh and Sanjeel Parekh. 2016. The MLPBOON Predicting Media Interestingness System for MediaEval 2016.. In *MediaEval 2016 Working Notes Proceedings*.

[9] 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, Oct (2011), 2825–2830.

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.