The MLPBOON Predicting Media Interestingness System for MediaEval 2016

Jayneel Parekh
Indian Institute of Technology, Bombay, India
jayneelparekh@gmail.com

Sanjeel Parekh
Technicolor, Cesson Sévigné, France
sanjeelparekh@gmail.com

ABSTRACT

This paper describes the system developed by team MLPBOON for MediaEval 2016 Predicting Media Interestingness Image Subtask. After experimenting with various features and classifiers on the development dataset, our final system involves use of CNN features (fc7 layer of AlexNet) for the input representation and logistic regression as the classifier. For the proposed method, the MAP for the best run reaches a value of 0.229.

1. INTRODUCTION

The MediaEval 2016 Predicting Media Interestingness Task [1] requires to automatically select images and/or video segments which are considered to be the most interesting for a common viewer. We will be focusing on solving the image interestingness subtask which involves automatically identifying images from a given set of key-frames extracted from a certain movie that the viewers report to be interesting. We will only use the visual content and no additional metadata.

The solution should essentially involve encoding into features many generic factors that are taken into account by humans while judging interestingness of an image [3]. However, there is an intrinsic difficulty this task presents which makes it extremely challenging to have reliable datasets and features - subjectivity [2]. One can observe the high level of subjectivity by realizing that a given image could be labeled as highly interesting or non-interesting depending upon the parts of the world in which it is surveyed. Even though current methods of annotating datasets tend to reduce [2] this factor but none can eliminate it.

Therefore, while taking into account subjectivity, we wish to determine features good for satisfactorily solving the task. In this context, several efforts have been made to understand factors that affect, or cues that contribute to interestingness of an image, even at an individual level. Katti *et al.* [5] attempt to understand the effect of human cognition and perception in interestingness. Work by Gygli *et al.* [3] shows how interestingness is related to features capturing unusualness, aesthetics and general preferences such as GIST, SIFT, Color Histograms *etc.* Further, [8] tries to learn attributes that can be used to predict interestingness at an individual level. Moreover, recent advances in application of neural networks to tasks in image processing and computer vision makes use of convolutional neural network [4] based features very promising [6].

Our approach was inspired by the following line of thought: if the right set of features are identified then any simple classifier should produce good results. Thus, we decided upon the proposed system after experimenting with different feature sets.

2. SYSTEM DESCRIPTION

We have opted for a more traditional machine learning pipeline involving - feature selection & preprocessing, training of classification model and then the predictions.

Given the training data feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$ consisting of N examples, each described by a F-dimensional vector, we first standardize it and apply principal component analysis (PCA) to reduce its dimensionality. The transformed feature matrix $\mathbf{Z} = (z_i)_i \in \mathbb{R}^{N \times M}$ is used to experiment with various classifiers. Here M depends on the number of top eigenvalues we wish to consider.

After preliminary testing (discussed in section 3), we decided to move ahead with logistic regression as our classifier. Logistic regression minimizes the following cost function [9].

$$\text{Cost } (w) = C \sum_{i=1}^{N} \log(1 + e^{-y_i w^T z_i}) + \frac{1}{2} w^T w, \qquad (1)$$

where w denotes the weight vector, $C > 0$ denotes penalty parameter, z_i denotes feature vector for the i^{th} instance of training data, while y_i denotes its label (0 if non-interesting and 1 if interesting). Note that a column of ones is appended to $\mathbf{Z}$ to include the hyperplane intercept as a coefficient of w. Now given a test data instance t, its label y is assigned according to equation (2).

$$y = \begin{cases} 1, & \text{if } w^T t \geq 0 \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

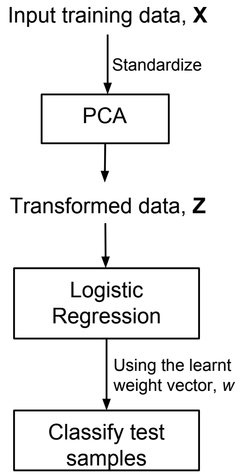
After experimenting with various descriptors (as discussed later in Section 3.1), we use CNN features extracted from fc7 layer of the AlexNet as our input feature representation for building $\mathbf{X}$. In the following section we discuss our experimental results obtained by varying different parameters of the above stated system.

3. EXPERIMENTAL VALIDATION

The training dataset consised of 5054 images extracted from 52 movie trailers, while the test data consisted of 2342 images extracted from 26 movie trailers. [1] gives complete

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

Figure 1: Block diagram for the proposed system



information about the preparation of the dataset. WEKA and scikit-learn [7] were used to implement and test various configurations.

3.1 Results and Discussion

The run submission results are given in Table 1. The table gives the mean average precision (MAP) - the official metric, precision and recall on the interesting images of different runs corresponding to the respective penalty parameter and number of transformed features retained after PCA. The general strategy for the run submissions was to first decide and fix the number of PCA features and subsequently tune C for best MAP on development data.

As observed, C decreases with increasing PCA features. This trend can be possibly explained as a way to avoid overfitting. The 5^{th} run gives the best MAP, however, the MAP for all the runs seems comparable. This points towards the utility of dimensionality reduction which significantly reduces the redundancy without affecting the results much. It was observed that 400 and 780 transformed features capture about 95% and 98% variance of the data, respectively. The difference between MAP on development and test data for all the runs was very small and lied between 0.01-0.03. The maximum MAP on development data was 0.24 with 1^{st} run's system configuration.

System Design Decisions

We experimented with the following features provided by the task [1]: CNN (fc7 and prob layers of AlexNet), GIST and Color Histogram (HSV space) features [4], and trained their different combinations on various machine learning classifiers like SVM, Decision Trees, Logistic Regression with 4-fold or 5-fold cross-validation when experimenting on the development data. In this section we give a rationale for selected features and classifier in the proposed system.

Features: The results on the development data using the GIST (512 dimensional feature vector) and ColorHistogram (128 dimensional feature vector) features were not very positive over any classifier. The use of CNN features (4096-dimensional fc7 & 1000-dimensional prob layers) did show significant improvements with fc7 features in particular performing better over the prob features. We also observed that combination of CNN features with GIST and ColorHistogram features gave similar performance to the case when we use just CNN features. Hence we went forward with using just CNN features, in particular from fc7 layer.

Run	No. of features	C	MAP	Precision	Recall
1	780	0.001	0.2205	0.140	0.581
2	700	0.008	0.2023	0.128	0.381
3	700	0.05	0.1941	0.131	0.348
4	400	0.1	0.2170	0.137	0.427
5	2016	0.0001	0.2296	0.141	0.726

Table 1: Run Submission Results: MAP was the official metric

Classifier: After selecting CNN features we experimented with various classifiers with different parameters. Specifically, we tried (1) SVM with linear, polynomial and rbf kernels (2) ridge regression classifier (3) stochastic gradient descent classifier with hinge, log, modified-huber and squared-hinge loss functions (4) logistic regression [7] and (5) random trees (WEKA). In general, it was found that logistic regression performed better than the other classifiers with its MAP being greater than 0.2 on training data. The performance of SVM was reasonable with the prob features but it did not show any significant improvements with the fc7 features. It particularly did not perform well with the rbf kernel. Hence we went ahead with logistic regression.

4. CONCLUSIONS

In summary, we have presented a system for interestingness prediction in images. Despite its simplicity, we obtain reasonable mean average precision values with the maximum being 0.229. From an analysis of the system's development history we think that selection of features was more important than the selection of the classifier. We believe it would be useful to identify and incorporate high level features describing image composition and object expressivity such as facial expressions. Moreover, to analyze the issue of subjectivity, it would be interesting to check inter-annotator agreement over test images.

5. REFERENCES

[1] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. Duong, and F. Lefebvre. Mediaeval 2016 predicting media interestingness task. In *Proc. of the MediaEval 2016 Workshop, Hilversum, Netherlands, Oct. 20-21*, 2016.

[2] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao. Interestingness prediction by robust learning to rank. In *European Conference on Computer Vision*, pages 488–503. Springer, 2014.

[3] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1633–1640, 2013.

[4] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang. Super fast event recognition in internet videos. *IEEE Transactions on Multimedia*, vol. 177(8):1–13, 2015.

[5] H. Katti, K. Y. Bin, C. T. Seng, and M. Kankanhalli. Interestingness discrimination in images.

[6] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876. ACM, 2014.

[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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] M. Soleymani. The quest for visual interest. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 919–922. ACM, 2015.

[9] H.-F. Yu, F.-L. Huang, and C.-J. Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, 2011.