BigVid at MediaEval 2016: Predicting Interestingness in Images and Videos

Baohan Xu[13], Yanwei Fu[23], Yu-Gang Jiang[13]
[1]School of Computer Science, Fudan University, China
[2]School of Data Science, Fudan University, China
[3]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China
{bhxu14, yanweifu, ygj}@fudan.edu.cn

ABSTRACT

Despite growing research interest, the tasks of predicting the interestingness of images and videos remain as an open challenge. The main obstacles come from both the diversity and complexity of video content and highly subjective and varying judgements of interestingness of different persons. In the MediaEval 2016 Predicting Media Interestingness Task, our team of BigVid@Fudan had submitted five runs exploring various methods of extraction, and modeling the low-level features (from visual and audio modalities) and hundreds of high-level semantic attributes; and fusing these features for classification. We not only investigated the use of the SVM (Support Vector Machine) model; but the recent deep learning methods were explored as well. We had submitted 5 runs using SVM/Ranking-SVM (Run1, Run3 and Run4) and Deep Neural Networks (Run2 and Run5) respectively. We achieved a mean average precision of 0.23 for the image subtask and 0.15 for the video subtask. Furthermore, our experiments revealed some insights of this task which are interesting and potential useful. For example, our results show that the visual features and high-level attributes are complementary to each other.

1. INTRODUCTION

The problem of automatically predicting the interestingness of images and videos has started to receive increasing attention. Interestingness prediction has a number of real-world applications, such as interestingness-based video recommendation system for social media platform.

MediaEval introduced the “2016 Predicting Media Interestingness Task”. This task requires participants to automatically select images and/or video segments which are considered to be the most interesting for a common viewer. Interestingness of the media is to be judged based on visual appearance, audio information and text accompanying the data. To solve the task, participants are strongly encouraged to deploy multimodal approaches. For the definitions, dataset and evaluation of the task, please refer to the official document [2].

This paper describes the first participation of MediaEval 2016 from the team of BigVid@Fudan. For this task we developed an approach to investigate how features and classifiers affect the interestingness in images and videos. Both visual features and high-level attributes were explored in our framework. We also compared SVM with deep neural networks to further study the relations between different features.

2. SYSTEM DESCRIPTION

Figure 1 gives an overview of our system. The whole system is composed of two key components: feature extraction and classifiers.

2.1 Feature Extraction

There are several pre-computed features provided by the organizers, such as denseSIFT [3], pre-trained CNN *fc7* layer features using ImageNet model and face features. To enlarge the useful information in data, we also consider two other types of high-level features. These features have been shown very useful in the tasks of aesthetics and interestingness prediction in [5] and [1]. The average pooling of all the descriptors from all sampled frames is used to form the video-level representation for each feature modality.

Style Attributes: We have considered the photographic style attributes [5] as high-level descriptors. These attributes have been shown highly related to aesthetics and interestingness in [5]. To compute these high-level features, the descriptor is formed by concatenating the classification outputs of 14 photographic styles (e.g., *Complementary Colors*, *Duotones*, *Rule of Thirds*, *Vanishing Point*, etc).

SentiBank: There are 1,200 concepts in SentiBank, and each is defined as an adjective-noun pair, e.g., ”*crazy cat*” and ”*lovely girl*”, where the adjective is strongly related to emotions and the noun corresponds to objects and scenes that are expected to be automatically detectable. Models for detecting the concepts were trained on Flickr images [1]. This set of attributes is intuitively effective on the emotion-related objects and scenes. Since interesting images/videos often related with strong emotions, the attribute is expected to be a very helpful clue for predicting interestingness.

2.2 Classifiers

Several classifiers are investigated here in order to be robustness to the diversity and complexity of similar visual content. Particularly, we discussed the SVM, Ranking-SVM and Deep Neural Networks (DNN) for feature fusion and classification. We explain them as follows,

SVM: χ^2 kernel was adopted for the bag-of-words features (denseSIFT), and Gaussian RBF kernel was used for the oth-

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

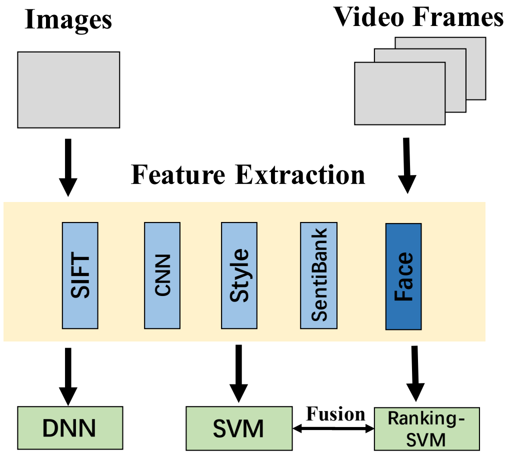

Figure 1: An overview of the key components in our proposed methods. We use DNN and SVM for both subtask, while Ranking-SVM is only used in video subtask. The face feature is computed according to the movement of face in the video shot, which is also only used in video subtask.

ers. For feature fusion, kernel-level average fusion was used for the features, which linearly combines kernels computed on different features.

Ranking-SVM: As the interestingness level also affects the classification result, we consider training a model to compare the interestingness of different images/videos. We therefore adopt Joaquims' Ranking SVM [4] to enhance the final results. To fully use the training data, we organized them in form of pairs, with ground-truth labels indicating which one is more interesting for each pair. Score-level average late fusion was adopted to combine the results of SVM and Ranking-SVM.

DNN: We also adopted a DNN-based classifier proposed in our recent work [6]. The fusion methods for the SVM classifiers may take advantage of different features; however, they often neglect the hidden relations shared among features. We proposed a regularized DNN to explore the relationship of distinct features, which is found useful for image/video classification. Specifically, for each input feature, a layer of neurons was first used to perform feature abstraction. Then, feature fusion is performed by another layer with carefully designed structural-norm regularization on network weights. The feature relationships is also considered in the regularized DNN. And the fused representation was finally used to construct a classification model in the last layer. With this special network, we are able to fuse features by considering both feature correlation and feature diversity, as well as perform classification simultaneously. Please see [6] for more details.

3. SUBMITTED RUNS AND RESULTS

There are two subtasks in this year's evaluation, namely *predicting video interestingness* and *predicting image interestingness*. We submitted 5 runs for official evaluation, among which 2 runs for the image subtask and 3 runs for the video subtask. Run1 and Run4 used SVM for video and image subtasks respectively, Run 2 and Run 5 used DNN for video and image subtasks respectively. Run 3 used SVM fusion with Ranking-SVM for video subtask.

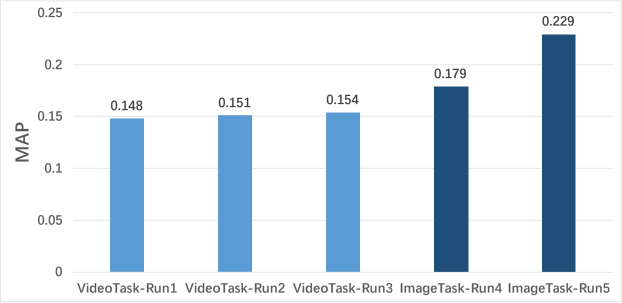

Figure 2: Performance of our 5 submitted runs on both video and image subtasks. AP is computed on a per trailer basis over the top N best ranked images/video shots. And MAP averaged over all trailers.

Figure 2 summarized the results of all the submissions. The official performance measure is MAP for both video and image subtasks. For image subtask, the DNN (Run5) significantly outperforms the SVM classifier (Run4) since feature correlation plays an important role in feature fusion for the interestingness task. This also clearly confirms the effectiveness of our proposed deep networks. Our experiments also verify that the high-level attributes are complementary to visual features and CNN features.

For the video subtask, besides the visual and high-level features, we combined the face features. The experiments show these features are complementary with each other, which means visual and high-level attribute both make contribution to determine whether a video clip is interesting or not. We found that adding the audio feature such as MFCC (Mel-Frequency Cepstrum Coefficient) may cause worse results. This is possibly due to the fact that the video shots are very short and cannot provide continuous and useful audio information. We also considered adding ranking information for video tasks (Run3); it shows slightly improvement over SVM (Run1) and DNN (Run2). The result also indicates that interestingness level may further improve the result.

It's also worth mentioning that the results of the image subtask are better than for the video subtask. It may be caused by the fact that the average of frame features weaken the weights of interesting information. How to fully use the effective information in video clips is a future direction.

4. CONCLUSIONS

We have explored both SVM model and DNN to achieve better classification on image and video interestingness. Our experiments have shown that DNN-based method outperforms the SVM model by considering feature correlation. Additionally, the high-level attributes are complementary to visual and CNN features on predicting interestingness. Nevertheless, our experimental results indicate that the visual and audio features may lack of discrimination about interestingness. Thus, as the future work of predicting the interestingness, we will consider extracting from the image and videos the text information which may contain the textual descriptions of interestingness (from linguistic perspective).

5. REFERENCES

[1] D. Borth, T. Chen, R. Ji, and S. F. Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM MM*, 2013.

[2] 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.

[3] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang. Super fast event recognition in internet videos. *IEEE TMM*, 17(8):1–13, 2015.

[4] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, 2002.

[5] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

[6] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *ACM MM*, 2014.