# RUC at MediaEval 2017: Predicting Media Interestingness Task

Shuai Wang, Shizhe Chen, Jinming Zhao, Wenxuan Wang, Qin Jin
Renmin University of China, China
{shuaiwang,cszhe1,qjin}@ruc.edu.cn
zhaojinming@bjfu.edu.cn,wangwenxuan@hust.edu.cn

## ABSTRACT

Predicting the interestingness of images or videos can greatly improve people's satisfaction in many applications, such as video retrieval and recommendations. In this paper, we present our methods in the 2017 Predicting Media Interestingness Task. We propose deep ranking model based on aural and visual modalities which simulates the human annotation procedures for more reliable interestingness prediction.

## 1 INTRODUCTION

The interestingness prediction task [1] aims to predict people's general preferences for images and videos, which has a wide range of applications such as video recommendation.

We propose an interestingness prediction model based on aural and visual modalities and deep ranking model to calculate interestingness score by the given images or video clips.

## 2 APPROACH

### 2.1 Aural-Visual Features

**Aural Features** We extract 39-dim Mel-Frequency Cepstral Coefficients (MFCCs) features from each video segment and create their bag of words features with 128 codewords denoting the responding segment. L1-norm is used to get the probability distributions on the codebook for each video.

**Visual Features** We utilize officially provided features including Alex_fc7, Alex_prob, ColorHist, DenseSIFT, GIST, HOG and LBP. Additionally, we consider 2048-dim frame-level features from the penultimate layer of InceptionV3 network, which is trained on 1.2 million images of ImageNet challenge dataset[3].

### 2.2 Deep Ranking Model

*2.2.1 Ranking Loss.* Suppose we have a set of video segments pairs $P$ sampled from the original video segment pool. In $P$, each pair contains a segment $p_i$ with higher interestingness and a segment $n_i$ with lower score. If function $f$ denotes the output of branches, we can get a score pair as follows:

$$(f(p_i), f(n_i)), \quad \forall (p_i, n_i) \in P \tag{1}$$

We set the margin in the loss as 1 by default according to [2]. By using this deviation namely the loss value, we can update the weights of the previous network and make it closer to the target of Equation 2. This kind of training will make the network more and more effective in recognition of video interestingness. Eventually, the network will give higher scores to attractive videos, and boring video will get a lower score.

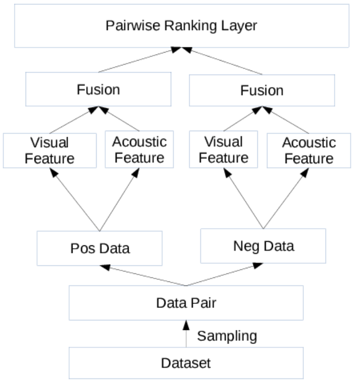

**Figure 1: The Network Structure of Deep Ranking Model**

$$minimization : \sum_{i=1}^{n} max(1 - f(p_i) + f(n_i), 0) \tag{2}$$

*2.2.2 Pairwise Generation.* Input data is an essential factor of the training stage. We try to use different strategies to sample data pairs as inputs. Different principles impact the training process and result in big difference. Let x and y denote float numbers in the range of 0 to 1. Our four strategies can be presented as follows:

$$f(p_i) - f(n_i) > x \tag{3}$$

$$f(p_i) - f(n_i) < y \tag{4}$$

$$x < f(p_i) - f(n_i) < y \tag{5}$$

$$f(p_i) - f(n_i) < x \quad or \quad f(p_i) - f(n_i) > y \tag{6}$$

Our basic empirical parameter of sampling is to set the distance of ground truth interestingness labels of the two videos in the same pair as 0.55. At first, big distances and small distances are both taken into account but does not show significant performance. We suppose that the network cannot learn much from two pretty similar videos and huge

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

gap between the two videos, which are the reasons that the network result in worse results.

# 3 RESULTS AND ANALYSIS

## 3.1 Experimental Setting

There are 7396 images or video clips in each subtask. We use video with id from 0 to 61 as the local training set, 62 to 69 as local validation set and 70 to 77 as local testing set.

We utilize the Support Vector Regression (SVR) and Random Forest Regression (RF) as our baseline models as the comparison with the deep ranking models. For SVM, RBF kernel is applied and the cost is searched from $2^2$ to $2^{10}$. And for Random Forest, the number of trees is searched from 100 to 1000 with step 100 and the depth of the tree is searched from 2 to 16.

## 3.2 Results and Discussion

In both subtasks, we consider different prediction models and features. The results are shown in Figure 2 and Figure 3 respectively.

In image subtask, we can find the pairwise ranking model shows greater performance generally and the deep neural network features are distinctive for interestingness prediction. It is not surprising that deep neural network displays its state-of-art capability.

In video subtask, we test the MFCC BoAW feature on local testing set and get a MAP of 0.151. Early fusion is applied over various visual features and MFCC BoAW. The results are generally consistent with the conclusions of image subtask. In our ranking model, the greatest MAP on local testing set is 0.210, which overpasses the other results. While the fusion boosts the performance not very much for each visual feature. We suppose it is due to its low dimensionality.

Given the experiments results on local testing set, we pick the winners of various models and features, namely pairwise ranking model and InceptionV3 feature, as our final choice for submissions. The official results for both subtasks are shown in Table 1. We utilize two types of input in the experiments of InceptionV3 feature, which are original images and normalized images. Normalized images are scaled into 0 to 1 for each pixel. As the results shown, the InceptionV3 feature from original image performs a little better than the normalized one on official testing set.

As the results show, image interestingness prediction is generally accurate than video subtask. We think that it is easier to fetch distinctive features from static images than from videos. Firstly, audio displays completely different cues with images and the fusion of the two modalities may present brand new interestingness. Secondly, dynamic properties like changes of scene make videos more informative, so that we cannot capture the interestingness precisely only by the static images inside a video.

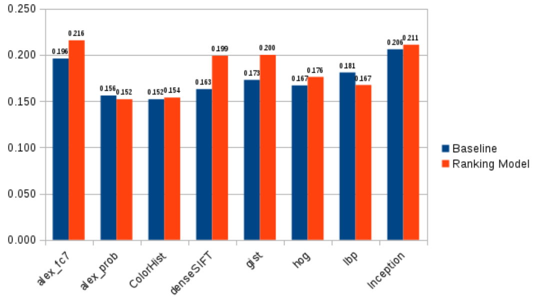

**Figure 2: MAP of Single Feature for Image Subtask on Local Testing Set**

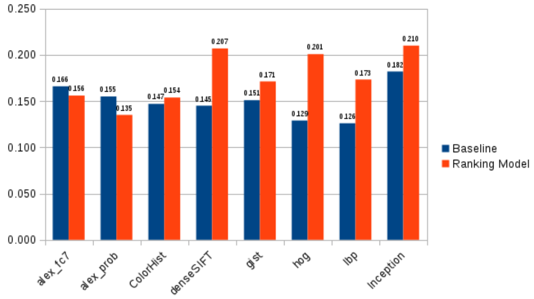

**Figure 3: MAP of Single Feature for Video Subtask on Local Testing Set**

**Table 1: Results of the official submitted runs**

| Runs | Subtask | Input | MAP | MAP@10 |
|---|---|---|---|---|
| 1 | Image | img_norm | 0.2655 | 0.0940 |
| 2 | Video | img_norm | 0.1830 | 0.0589 |
| 3 | Video | img_origin | 0.1897 | 0.0637 |

After investigating the testing set, we find out some interesting phenomena. For image subtask, images containing varied scenes can be ranked precisely, but a series of images with dark spectacles gains a low MAP. For video subtask, videos with changeless audio content obtain relative low MAP.

# 4 CONCLUSIONS

We develop an interestingness prediction system based on pairwise ranking. Comparing with basic regression models, we notice the effectiveness of ranking model and the InceptionV3 feature is distinctive for interestingness prediction task. In training process, optimizing the data pair sampling strategy is always considered a fundamental and essential point. In the future, we will also use more temporal cues to guarantee that the information within the internal frames of the same video is not wasted.

# ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Plan under Grant No.2016YFB1001202.

# REFERENCES

[1] Claire Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh Toan Do, Michael Gygli, and Ngoc Q K Duong. Sept. 13-15, 2017.. MediaEval 2017 Predicting Media Interestingness Task. In *Proc. of the MediaEval 2017 Workshop, Dublin, Ireland*.

[2] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 982–990.

[3] Hangjun Ye and Guangyou Xu. 2003. Hierarchical indexing scheme for fast search in a large-scale image database. 5286, 3-4 (2003), 974–979.