Technicolor@MediaEval 2016 Predicting Media Interestingness Task

Yuesong Shen[1,2], Claire-Hélène Demarty[2], Ngoc Q. K. Duong[2]
[1]École polytechnique, France
[2]Technicolor, France
yuesong.shen@polytechnique.edu
claire-helene.demarty@technicolor.com
quang-khanh-ngoc.duong@technicolor.com

ABSTRACT

This paper presents the work done at Technicolor regarding the MediaEval 2016 Predicting Media Interestingness Task, which aims at predicting the interestingness of individual images and video segments extracted from Hollywood movies. We participated in both the image and video subtasks.

1. INTRODUCTION

The MediaEval 2016 Predicting Media Interestingness Task aims at predicting the level of interestingness of multimedia content, *i.e.,* frames and/or video excerpts from Hollywood-type movies. The task is divided in two subtasks depending on the type of content, *i.e.,* images or video segments. A complete description of the task can be found in [2].

For the image subtask, Technicolor's contribution is double: a Support Vector Machine (SVM)-based system is compared with several deep neural network (DNN) structures: Multi-layer perceptrons (MLP), Residual networks (ResNet), Highway networks. For the video subtask, we compare different systems built on DNN including both the existing Long Short Term Memory (LSTM) with ResNet blocks, and the proposed architecture named Circular State-Passing Recurrent Neural Network (CSP-RNN).

The paper is divided in two main parts corresponding to the two subtasks. Before this, section 2 gives insight on the features used. In each subtask's section, we present the systems built, then give some details on the derived runs. The results for the two subtasks are discussed in Section 5.

2. FEATURES

For both subtasks, input features for the visual modality are the CNN features extracted from the fc7 layer of the pre-trained CaffeNet model [5]. They were computed for each image of the dataset, after being resized to 227 for its smaller dimension and center-cropped so that the aspect ratio is preserved. The mean image that comes with the Caffe model was also subtracted for normalization purpose. The final feature dimension is 4096 per image, or per video frame for the video subtask.

For the audio modality, 60 Mel Frequency Cepstral Coefficients (MFCCs) concatenated with their first and second derivatives were extracted from windows of size 80 ms centered around each frame, resulting in an audio feature vector of size 180 per video frame. The mean value of the feature vector over the whole dataset is then subtracted for the normalization purpose.

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

3. IMAGE SUBTASK

For the image subtask, the philosophy was to experiment using several DNN structures and to compare with a SVM-based baseline. For all system types and both subtasks (image and video), the best parameter configurations were chosen, by splitting the development set (either MediaEval data or external data) into some training (80%) and validation sets (20%). As the MediaEval development dataset is not very large, we proceeded to a final training on the whole development set, when building the final models, except for the CSP-based runs, for which this final training was omitted.

SVM-based system We tested SVM with different kernels: linear, RBF and polynomial and with different parameter settings on the development dataset. We observed that the validation accuracy varies from one run to another (even for the same parameter setting) as the training samples change due to the random partition into the training and validation sets. This may suggest that the dataset is not large enough. Also, because of the class imbalance, the validation accuracy used for training makes it difficult to choose the best SVM configuration during grid search, that targets the optimization of the official MAP metric.

MLP-based system Several variations of network structures have been tested, with different number of layers, layer sizes, activation functions and topologies among which simple MLP, residual [3] and highway [7] networks. These structures were first trained on a balanced dataset of 200,000 images, extracted using the Flickr interestingness API [1]. As this API uses some social metadata associated to content, it may lead to a social definition of interestingness, instead of a more content-driven interestingness, which may bias the system performance on the official dataset. The best performance in terms of accuracy for the Flickr dataset was obtained by a simple structure: a first dense layer of size 1000 and rectified linear unit (ReLU) activation, with a dropout of 0.5, followed by a final softmax layer. This structure was then re-trained on the MediaEval image development dataset, but with the addition of some resampling or upsampling steps of the training data, to cope with the imbalance of the two classes in the official dataset. During resampling, a training sample is selected randomly from one class or another depending on a probability fixed beforehand. Upsampling consists of putting multiple occurrences of each interesting sample into the list of training data, resulting in potentially interesting samples being used multiple times during training. In both cases, different probabilities

(0.3 to 0.6 for the interesting class) or upsampling proportions (5 to 13 times more interesting samples) were tested.

Run #1: SVM-based SVM in Python Scikit-learn package[1] with RBF kernel, $gamma = 1/n\_features$, $c = 0.1$ and default parameter settings elsewhere, is used. Upsampling strategy to enlarge interesting samples by factor of 11 is used.

Run #2: MLP-based A simple structure with 2 layers of sizes (1000, 2) (cf. section 3) was selected for its performance on the Flickr dataset. Best performances with this structure were obtained with a learning rate of 0.1, decay rate of 0.1, ReLu activation function, Adadelta optimizing method and a maximum of 10 epochs. Resampling with probability of 0.6 for the interesting class gives the best MAP value on the MediaEval development set.

4. VIDEO SUBTASK

Different DNN structures capable of handling the temporal evolution of the data were tested with variation of size and depth. We also investigated the performances of different modalities separately vs. in a multimodal approach.

Systems based on existing structures

Different simple RNN and LSTM [4] structures were tested by varying their number and size of layers, as they are well-known to be able to handle the temporal aspect of the data. We also experimented the idea of ResNet (recently proposed for CNN) in our implementation with RNN and LSTM. Monomodal systems (audio-only, visual-only) were also compared to multimodal (audio+visual modality) ones. For the latter, a mid-level multimodal fusion was implemented, *i.e.,* each modality was first processed independently through one or more layers before merging and processing by some additional layers, as illustrated in figure 1. The best structures and set of parameters were chosen while training on the Flickr part of the Jiang dataset [6]. Contrary to the MediaEval dataset, this dataset contains 1200 longer videos, equally balanced between the two classes. Once the structures and parameters were chosen, some upsampling/resampling process was applied while re-training on the MediaEval dataset.

Run #3: LSTM-ResNet-based The best structure obtained while validating on the Jiang dataset corresponds to figure 1, with a multimodal processing part composed of a residual block built upon 2 LSTM layers of size 436. After re-training on the MediaEval dataset, upsampling with a factor of 9 was applied to the input samples.

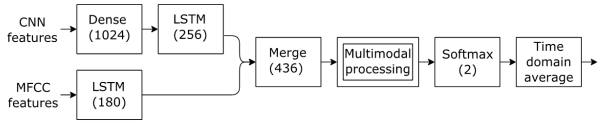

Figure 1: Proposed multimodal architecture.

Systems based on the proposed CSP

Figure 2 illustrates the philosophy of this new structure, which can be seen as a generalization of the traditional RNN, in which at a time instance t, N samples of the input sequence go through N recurrent nodes arranged in a circular structure (allowing to take into account both the past and the future over a temporal window of size N) to produce N internal outputs. These outputs are then combined to form a final output at t. This architecture targets a better modeling of temporal events in videos by considering several successive temporal samples and states to produce an output at each time instance (while RNN takes one input sample and one state to produce an output at each time instance). The CSP-RNN was trained directly on the MediaEval dataset, and for three fixed configurations only: audio, video and multimodal.

[1]http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC

Runs - image	MAP	Runs - video	MAP
Run #1	0.2336	**Run #3**	0.1465
Run #2	0.2315	**Run #4**	0.1365
		Run #5	0.1618
Random baseline	0.1656	**Random baseline**	0.1496

Table 1: Results in terms of MAP values.

Run #4: video and CSP-based For the audio and video configurations, the features were used directly as input to the CSP network, with some dimension reduction from 4096 to 256 for the video, thanks to a simple dense layer. No upsampling/resampling was applied. During validation on the MediaEval dataset, the audio-only CSP configuration has given lower performances than the video and multimodal configurations. We therefore kept the video system for run#4.

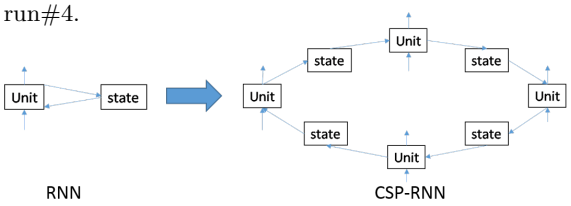

Figure 2: Proposed CSP-RNN structure.

Run #5: multimodal and CSP-based For the multimodal configuration the same framework proposed in figure 1 was kept: the multimodal processing part being replaced by a CSP of size 5. No upsampling/resampling was tested.

5. RESULTS AND DISCUSSION

The obtained results are reported in terms of MAP in Table 1, with some baseline values computed from a random ranking of the test data. At least on the image subtask our systems perform significantly higher than the baseline. For the video subtask, MAP values are lower and we may wonder whether these performances come in part from the difficulty of the task itself or the dataset which contains significant number of very short shots that were certainly difficult to annotate.

For the image subtask, we observed that simple SVM systems perform similarly (development set) and even slightly better (test set) than more sophisticated DNNs, leading to the conclusion that the size of the dataset was probably not large enough for DNN training. This is also supported by our test on the external Flickr dataset containing 200,000 images for which DNN reached more than 80% accuracy.

For the video subtask, several conclusions may be drawn. First, multimodality seems to bring benefit to the task (this was confirmed by some additional but not submitted runs). Second, the new CSP structure seems be able to capture the temporal evolution of the videos better than classic RNN and more sophisticated LSTM-ResNet structures, and this independently of the monomodal branches which were the same in both cases. This very first results support the need for further testing of this new structure in the future.

6. REFERENCES

[1] Flickr interestingness api. `https://www.flickr.com/services/api/flickr.interestingness.getList.html`.

[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] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *arXiv prepring arXiv:1506.01497*, 2015.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[6] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pages 1113–1119. AAAI Press, 2013.

[7] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.