Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning

Qi Dai[1], Rui-Wei Zhao[1], Zuxuan Wu[1], Xi Wang[1],
Zichen Gu[2], Wenhai Wu[2], Yu-Gang Jiang[1*]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Media Lab, Huawei Technologies Co. Ltd., China

ABSTRACT

Techniques for violent scene detection and affective impact prediction in videos can be deployed in many applications. In MediaEval 2015, we explore deep learning methods to tackle this challenging problem. Our system consists of several deep learning features. First, we train a Convolutional Neural Network (CNN) model with a subset of ImageNet classes selected particularly for violence detection. Second, we adopt a specially designed two-stream CNN framework [1] to extract features on both static frames and motion optical flows. Third, Long Short Term Memory (LSTM) models are applied on top of the two-stream CNN features, which can capture the longer-term temporal dynamics. In addition, several conventional motion and audio features are also extracted as complementary information to the deep learning features. By fusing all the advanced features, we achieve a mean average precision of 0.296 in the violence detection subtask, and an accuracy of 0.418 and 0.488 for arousal and valence respectively in the induced affect detection subtask.

1. SYSTEM DESCRIPTION

Figure 1 gives an overview of our system. In this short paper, we briefly describe each of the key components. For more information about the task definitions, interested readers may refer to [2].

1.1 Features

We extract several features, including both neural network based features and the conventional hand-crafted ones, as described in the following.

CNN-Violence: The effectiveness of CNN models has been verified on many visual recognition tasks like object recognition. We train an AlexNet [3] based model on video frames, which takes individual frames as network inputs followed by several convolutional layers, pooling layers and fully connected (FC) layers. Specially, a subset of ImageNet is used to tune the network. We manually pick 2614 classes which are relatively more related to violence (or its related semantic ingredients). These classes are mostly among the categories of scenes, people, weapons and actions. The outputs of FC6 (i.e., the sixth FC layer; 4096-d), FC7 (4096-d) and FC8 (2614-d) are used as the features.

*Corresponding author. Email: ygj@fudan.edu.cn.

MediaEval 2015 Workshop, Sep. 14-15, 2015, Wurzen, Germany

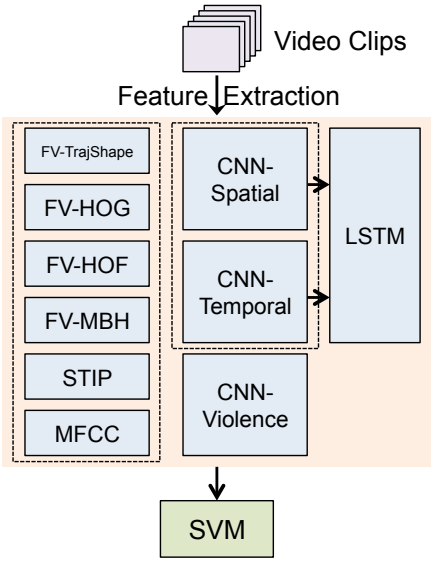

Figure 1: The key components in our system.

Two-stream CNN: Recent works [1, 4] have also revealed the effectiveness of the CNN models for video classification. Video data could be naturally decomposed into two components, namely spatial and temporal respectively. Thus we adopt a two-stream (spatial stream and temporal stream) CNN model to extract features. Specially, for the spatial stream, a CNN model which was pre-trained on the ImageNet Challenge dataset (different from the 2614 classes used in CNN-Violence) is used. The outputs of the last three FC layers are used as the features. For the temporal stream, which aims to capture the motion information, a CNN model is trained to take stacked optical flows as input. The output of the last FC layer is used as features. For more details of our two-stream CNN model used in this evaluation, please refer to [4]. Note that the models are not fine-tuned using MediaEval data.

LSTM: In order to further model the long-term dynamic information that is mostly discarded in the spatial and temporal stream CNNs, we utilize our recently developed LSTM model [5]. Different from a traditional Recurrent Neural Network (RNN) unit, the LSTM unit has a built-in memory cell. Several non-linear gates are used to govern the information flow into and out of the cell, which enables the model to explore long-range dynamics. Figure 2 shows the structure of the LSTM model. With a two-stream CNN model, video

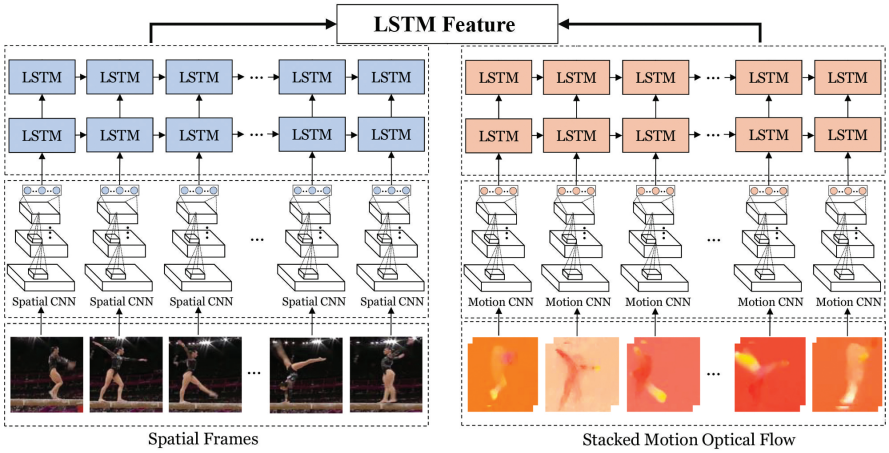

Figure 2: The structure of the LSTM network.

frames or stacked optical flows could be transformed to a series of fixed-length vector representations. The LSTM model is used to model these temporal information. Due to time constraint of the evaluation, we directly adopt LSTM model trained with another video dataset (the UCF-101 dataset [6]) and use the average output from all the time-steps of the last LSTM layers as the feature (512-d).

Conventional features: Same as last year [7], we also extract the improved dense trajectories (IDT) features according to [8]. Four trajectories based features, including histograms of oriented gradients (HOG), histograms of optical flow (HOF), motion boundary histograms (MBH) and trajectory shape (TrajShape) descriptors are computed. The features are encoded using the Fisher vectors (FV) with a codebook of 256 codewords. The other two kinds of conventional features include Space-Time Interest Points (STIP) [9] and Mel-Frequency Cepstral Coefficients (MFCC). The STIP describes the texture and motion features around local interest points, which are encoded using the bag-of-words framework with 4000 codewords. The MFCC is a very popular audio feature. It is extracted from every 32ms time-window with 50% overlap. The bag-of-words is also adopted to quantize the MFCC descriptors, using 4000 codewords.

1.2 Classification

We use SVM as the classifier. Linear kernel is used for the four IDT features, and χ^2 kernel is used for all the others. For feature fusion, kernel level fusion is adopted, which linearly combines kernels computed on different features.

Notice that direct classification with the CNN is feasible, which may lead to better results. However, tuning the models using MediaEval data requires additional computation.

2. SUBMITTED RUNS AND RESULTS

There are two subtasks in this year's evaluation, namely violence detection and induced affect detection. Induced affect detection requires participants to predict two emotional impacts, arousal and valence, of a video clip.

We submit five runs for each subtask. For both subtasks, Run 1 uses the conventional features, Run 2 uses all the deep learning features, Run 3 combines Run 1 and the CNN-Violence feature, Run 4 further includes the two-stream CNN features, and, finally, Run 5 fuses all the features.

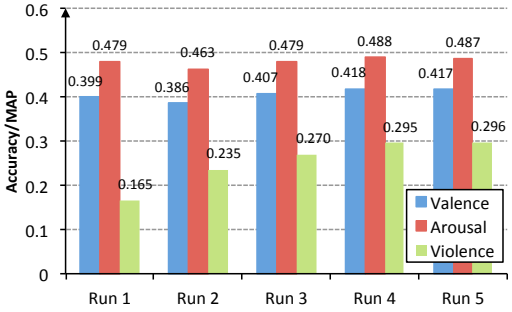

Figure 3: Performance of our 5 submitted runs on both affect and violence subtasks. For the affect subtask, accuracy is used as the official performance measure. For the violence subtask, MAP is used.

Figure 3 shows the results of all the submissions. The official performance measure is accuracy and MAP for the affect and violence subtasks respectively. We can see that the deep learning based features (Run 2) are significantly better than the conventional features (Run 1) for the violence subtask, and both are comparable for the affect subtask. This is possibly because the CNN-Violence feature is specially optimized for detecting violence. Comparing Run 3 with Run 1, it is obvious that the CNN-Violence feature could improve the result with a large margin for the violence subtask (from 0.165 to 0.27), but the gain is much less significant for the other subtask. In addition, the two-stream CNN also brings considerable improvement on both subtasks (Run 4). The LSTM models seem to be ineffective (Run 5 vs. Run 4). The reason is that the LSTM models were trained on the UCF-101 dataset, which is very different from the data used in MediaEval. We expect clear improvements from LSTM if the models can be re-trained. Also, the contributions from the CNN-based models could probably be even more significant if re-trained on MediaEval data. Overall, we conclude that deep learning features are very effective for this task and the room for improvements is huge with model re-training.

Acknowledgements

This work was supported in part by a Key Technologies Research and Development Program of China (#2013BAH09F01), a National 863 Program of China (#2014AA015101), a grant from NSFC (#61201387), and Huawei Technologies.

3. REFERENCES

[1] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[2] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, L. Chen. The MediaEval 2015 Affective Impact of Movies Task. In *MediaEval 2015 Workshop*, 2015.

[3] A. Krizhevsky, I. Sutskever, G. E. Hinton. Image-Net classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] H. Ye, Z. Wu, R.-W. Zhao, X. Wang, Y.-G. Jiang, X. Xue. Evaluating Two-Stream CNN for Video Classification points. In *ICMR*, 2015.

[5] Z. Wu, X. Wang, Y.-G. Jiang et al. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification. In *ACM Multimedia*, 2015.

[6] K. Soomro, A. R. Zamir and M. Shah. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. In *CRCV-TR-12-01*, 2012.

[7] Q. Dai, Z. Wu, Y.-G. Jiang et al. Fudan-NJUST at MediaEval 2014: Violent Scenes Detection Using Deep Neural Networks. In *MediaEval 2014 Workshop*, 2014.

[8] H. Wang, C. Schmid. Action Recognition With Improved Trajectories. In *ICCV*, 2013.

[9] I. Laptev. On space-time interest points. *IJCV*, 64:107–123, 2005.