Fudan at TRECVID 2015: Adaptive Feature Fusion for Multimedia Event Detection in Videos

Zuxuan Wu, Hao Ye, Yu-Gang Jiang*, Xiangyang Xue
School of Computer Science, Fudan University, Shanghai
{zxwu, haoye10, ygj, xyxue}@fudan.edu.cn

ABSTRACT

TRECVID 2015 [4] Multimedia Event Detection (MED) is an interesting and challenging task on the detection of high level complex events in Internet videos [1]. In this notebook paper, we present an overview of our system, focusing on combining multiple feature representations to improve the performance. Specifically, with the outputs of the multiple features, we adopt a simple yet effective fusion method to generate the final predictions, where the optimal fusion weights are learned adaptively for each class, and the learning process is regularized by automatically estimated class relationships. Our MED submissions include 5 system runs for the Pre-Specified (PS) sub-task and 2 runs for the AdHoc (AH) sub-task under 010Ex training condition. Very competitive results are obtained, which verify the effectiveness of both deep features and our fusion method.

Table 1: A summary of our submissions.

		Features	Fusion
AH	Run-1	VGG19-fc_{67}, VGG19-20K, FCVID-233, Conventional Fea.	Adaptive
	Run-2	VGG19-20K, FCVID-233, Conventional Fea.	Adaptive
PS	Run-1	VGG19-fc_{67}, VGG19-20K, FCVID-233, Conventional Fea.	Adaptive Param-1
	Run-2	VGG19-fc_{67}, VGG19-20K, FCVID-233, Conventional Fea.	Average
	Run-3	VGG19-fc_{67}, VGG19-20K, FCVID-233, Conventional Fea.	Adaptive Param-2
	Run-4	VGG19-20K, FCVID-233, Conventional Fea.	Adaptive
	Run-5	VGG19-fc_{67}, VGG19-20K, Conventional Fea.	Adaptive

1. SYSTEM DESCRIPTION

Figure 1 gives an overview of our system, consisting of three components, namely feature extraction, classification and adaptive fusion. We briefly introduce each of them in the following.

1.1 Features

In order to model videos from different perspectives, we compute both deep features extracted from Convolutional Neural Networks (CNN) and state-of-the-art hand-crafted features.

*Corresponding author.

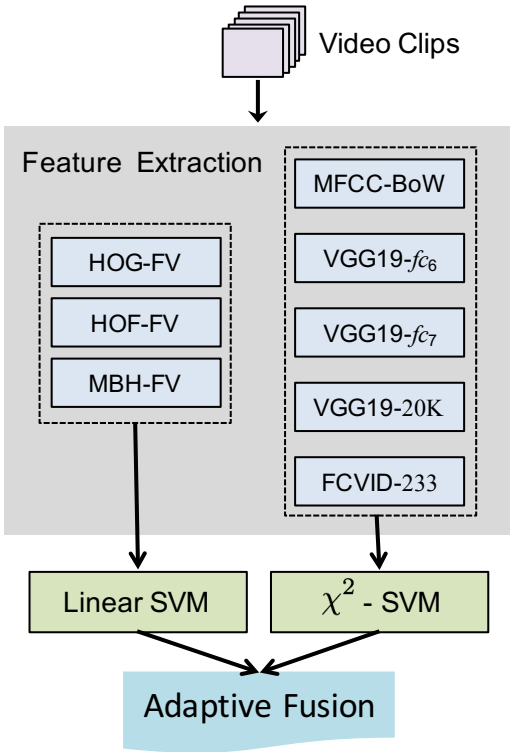

Figure 1: An overview of the key components in our system.

VGG19-fc_{67}: CNN models have demonstrated superior results on many visual recognition tasks, including image classification, object detection, etc. In our system, we adopt the state-of-the-art VGG19 model proposed by Simonyan [5], which consists of nineteen layers, including sixteen convolutional layers and three fully connected layers. In addition, the size of all the convolutional filters decreases to 3×3 and the stride reduces to only 1 pixel, which enables the network to explore finer-grained details from the feature maps. We finetune the original VGG19 model using all 14 million images annotated into 20,574 classes from ImageNet. Given a video clip, we extract the outputs from the first (fc_6) and second (fc_7) fully connected layer as features for each video frame respectively, and then average all frame-level descriptors into video-level features.

VGG19-20K: We also adopt the softmax outputs from our retrained VGG19 model on the full ImageNet dataset as

high-level semantic concepts, covering as many as 20K categories. With the high reliability of the VGG19 model, the 20,574-d response scores indicate the likelihood of presence of the corresponding classes in each frame. We then average the response vectors of all frames in order to compute the video-level representations.

FCVID-233: To explore helpful information from existing large video benchmarks, we select 233 video categories from the recently released Fudan-Columbia Video Dataset (FCVID) [2] to train video class detectors with the CNN. FCVID contains about 90K videos annotated into 239 videos, covering a wide range of topics like social events (e.g., "tailgate part"), procedural events (e.g., "making cak"), objects (e.g.,"panda"), scenes (e.g., "beach"), etc. We adopt spatial frames as inputs to fine-tune the VGG19 model for classifying the 233 video categories as in [9]. Similarly, the frame-level responses are averaged to obtain a 233-d video-level feature for all the MED videos.

Conventional features: In addition to the deep features, we also extract the improved dense trajectories (IDT) features according to [6]. Briefly, densely sampled local frame patches are first tracked over time and three descriptors are then computed for each trajectory: a 96-d histogram of oriented gradients (HOG) descriptor, a 108-d histogram of optical flow (HOF) descriptor, and a 108-d motion boundary histogram (MBH) descriptor. We first reduce the dimension of these descriptors by a factor of two using Principle Component Analysis (PCA) then we encode these features into Fisher Vectors with a codebook of 256 words.

It is well-known that the audio soundtracks contain useful clues for identifying some video semantics. We utilize the popular MFCCs (Mel-Frequency Cepstral Coefficients), which are computed for every 32ms timewindow with 50% overlap and then quantized into a 4000-d bag-of-words representation [3].

1.2 Classification

We adopt SVM as the classifier. Linear kernel is applied for the IDT features, since it was found working well with the high-dimensional Fisher vector based representations. We adopt χ^2-kernel for all the other features.

Notice that direct classification with the CNN is difficult for the MED task, since the number of positive examples is too few to tune a good neural network based classifier.

1.3 Fusion

Given the prediction scores of multiple features, we are able to capture the video characteristics from different aspects. It is critical to effectively fuse the multiple scores to generate the final predictions. Different semantic classes usually associate with the multiple streams with different strength. For example, some classes are strongly related with particular objects that could be effectively recognized by the CNN features, while others may contain dramatic movements so motion features such as IDT can contribute more significantly.

Traditional fusion methods are performed uniformly without considering the class-specific preferences. Different from the uniform methods, we adopt our recently proposed adaptive multi-stream fusion to learn the optimal fusion weights adaptively for each class [7]. In addition, by regularizing the weight learning process using class relationships estimated without using additional labels, class relationship context is also injected into the final predictions to promote the overall results. See [7] for more details.

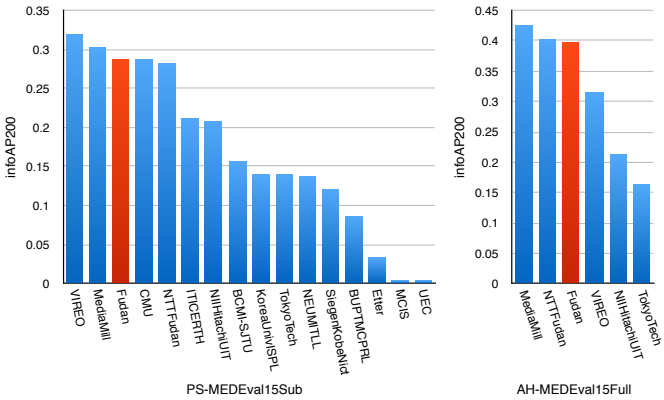

Figure 2: Performance of official submissions for PS and AH sub-tasks. The vertical axis shows the performance measured by infoAP200.

2. RESULT AND DISCUSSION

In TRECVID 2015, we submitted two runs for AH sub-task and five runs for the PS sub-task as listed in Table 1, all of which are measured by infoAP200 as defined by the task organizers. Figure 2 shows the performance of the official submissions for PS (MEDEval15Sub) and AH sub-tasks (MEDEval15Full). We can see that our system achieves competitive performance for both sub-tasks.

More specifically, for the AH sub-task, we achieved a 39.6% infoAP200 using all the features with adaptive fusion for each class. When we drop the VGG19-fc_{67} features, the performance significantly decreases to 33.7%, which verifies that CNN features are very crucial for video classification [8].

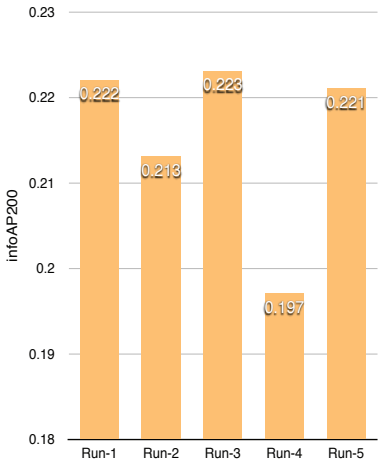

Figure 3: Performance of our 5 submitted runs for the PS sub-task.

Figure 3 shows the results of our five submissions for PS sub-task. Run-1, Run-2 and Run-3 are all based on the same set of features, but Run-2 only adopts average fusion to com-

bine the scores of multiple features. Run-1 and Run-3 learn the optimal fusion weights for each class, both of which offer much better results than average fusion. Since Run-3 emphasizes more on the use of class relationships (with a larger parameter value in the learning process), slightly better performance is achieved. Comparing Run-4 with Run-5, it is clear that the VGG19 features have a significant impact on the overall performance.

3. REFERENCES

[1] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

[2] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *arXiv preprint arXiv:1502.07209*, 2015.

[3] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[4] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quéenot, and R. Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *NIST TRECVID Workshop*, 2015.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[6] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[7] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, X. Xue, and J. Wang. Fusing multi-stream deep networks for video classification. *arXiv preprint arXiv:1509.06086*, 2015.

[8] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM Multimedia*, 2015.

[9] H. Ye, Z. Wu, R.-W. Zhao, X. Wang, Y.-G. Jiang, and X. Xue. Evaluating two-stream cnn for video classification. In *ACM ICMR*, 2015.