NTT-Fudan Team @ TRECVID 2015: Multimedia Event Detection

Yongqing Sun[1], Zuxuan Wu[2], Xi Wang[2], Kyoko Sudo[1], Yukinobu Taniguchi[1,3], Tetsuya Kinebuchi[1], and Yu-Gang Jiang[2]

[1] NTT Media Intelligence Laboratories, Japan
[2] Fudan University, China
[3] Faculty of Engineering, Tokyo University of Science, Japan

1 Summary

In this notebook paper, we present an overview and results analysis of our system designed for TREVID 2015 [1] Multimedia Event Detection (MED) task. Motivated by the great success of deep learning, we focus on exploiting various deep features to capture visual appearance and temporal dynamics in video clips. In order to fully utilize knowledge from existing large-scale image and video benchmarks, our system also incorporates high-level semantic features generated by pre-trained Convolutional Neural Networks. Then we performed classification with SVMs using different features and average the results carefully to obtain the final prediction scores. We submmited results for full evaluation of both Pre-Specified (PS) and Ad-Hoc (AH) sub-tasks in the 010Ex training condition. Our runs are submitted below.

Table 1. Summary of submitted runs for TRECVID 2015 MED

AH	baseline-1	IDT + MFCC + VGG19-fc_6 + VGG19-fc_7 + C-20K + C-233
	contrast-1	VGG19-fc_6 + VGG19-fc_7 + C-20K + C-233
PS	baseline-1	IDT + MFCC + VGG19-fc_6 + VGG19-fc_7 + C-20K + C-233
	contrast-1	IDT + MFCC + VGG19-fc_6 + VGG19-fc_7 + LSTM + C-20K + C-233
	contrast-2	IDT + MFCC + VGG19-fc_6 + VGG19-fc_7
	contrast-3	VGG19-fc_6 + VGG19-fc_7 + C-20K + C-233
	contrast-4	IDT + MFCC

2 System Overview

In this section, we elaborate the technical components of our system. First, we describe the adopted features as well as their corresponding encoding strategies. Then we introduce the classifiers for model training and different fusion approaches.

2.1 Feature Representation

In TRECVID 2015, we adopt three sets of features in our system to capture the rich multimodal information in videos, including traditional features, deep features and concept representations. All the features used in our MED system are summarized in Table 2 and the detailed descriptions are given below.

Table 2. Features adopted in our MED system

	Features
Traditional Features	IDT (MBH, HOG, HOF) MFCC
Deep Features	VGG19-fc_6, VGG19-fc_7, LSTM
Concept Feature	C-20K, C-233

Traditional Features

– Improved Dense Trajectory (IDT): We extract the state-of-the-art improved dense trajectory features [2], which exhibit top-notch performance on action recognition tasks. Along with the densely extracted trajectories, three features are computed: HOG, HOF, and MBH. We first reduce the dimension of HOG, HOF and MBH descriptors by a factor of two using Principal Component Analysis (PCA). Then these features are further quantized respectively using the FV representation with the vocabulary size being 256.

– MFCC: In addition to the above visual features, audio features can provide complementary clues. For this, we adopt the well-known Mel-Frequency Cepstral Coefficients (MFCC). It is first computed over each 32ms time-window (with 16ms overlap) of the soundtrack and then all the descriptors are quantized into a single BoW feature representation.

Deep Features

– VGG19-fc_6 and VGG19-fc_7: Inspired by the great success of CNN, we adopt VGG19 model proposed by Simonyan [3] in our system. Compared to AlexNet, VGG19 not only further reduces the size of convolutional filters and the stride, but more importantly, it also extends the depth of the network. With this much deeper architecture, VGG19 possess strong capabilities of learning more discriminative features and the high-level final predictions. It can produce a 7.1% top-5 error rate on the ILSVRC-2012 validation set. In order to increase the generalization ability of the VGG19 model, we finetune the model using the full ImageNet dataset, which consists of 14 million images annotated into 20K classes. Given a video clip, we extract the outputs from the two fully-connected layers (i.e., VGG19-fc_6 and VGG19-fc_7) of each frame and then average them frame-level features into video-level representations.

– LSTM Feature: In order to further model the long-term dynamic information that is mostly discarded in the spatial CNNs, we utilize our recently developed LSTM model, as shown in Figure 1. Different from a traditional Recurrent Neural Network (RNN) unit, the LSTM unit has a built-in memory cell.

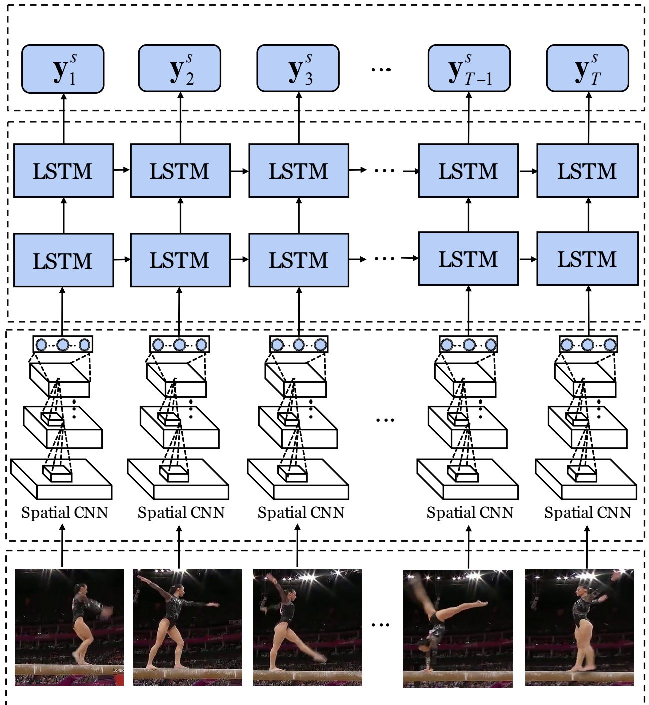

Fig. 1. The structure of the LSTM network.

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 by taking the feature representations from CNN at each time step. Due to time constraint of the evaluation, we directly adopt LSTM model trained with another video dataset (the UCF-101 dataset) and use the average output from all the time-steps of the last LSTM layers as the feature (512-d).

Concept Features

– C-20K: Since the softmax output of our finetuned VGG19 model demonstrates the probability of the 20K objects in a frame, we adopt this as our high-level semantic concept detector in our system. For each key frame in a given video, we obtain a 20,574-d concept score with the trained model and frame-level scores are then averaged to generate a video-level concept feature vector for further classification.

– C-233: We trained 233 concept detectors on the newly released Fudan-Columbia Video Dataset [4] using a VGG19 CNN model. Given a video clip, we obtain the 233 concept detector responses using the softmax layer of the CNN model. Then, a video level concept representation is obtained by average pooling the scores of all frames.

2.2 Classification and Fusion

To train event detection models, we employ two different types of classifiers in our system:

- Linear SVMs: To enhance classification performance, we first perform early fusion with the appearance feature and motion feature by concatenating them into a long vector. Since the concatenated vector is discriminative enough in the high-dimensional space, we adopt linear SVMs with C fixed to 100 to train the model.
- χ^2 SVMs: For MFCC audio feature, deep features and concept scores, we first map them into χ^2-kernel separately. Then, we train independent classifiers for each of these features.

With multiple classifiers, each video clip is accordingly associated with multiple output scores, which are then fused to compute the final prediction.

3 Results and Analysis

Our MED system is designed to combine multiple feature representations to fully model multiple clues in videos. We submitted 2 runs for AH task and 5 runs for the PS task in order to investigate the effectiveness of different features.

Figure 2 shows the results of all the submissions. The official performance measure is infoAP200 for both AH and PS tasks. For AH task, we can see that traditional features are highly complementary with features extracted from deep models (i.e., deep features and concept features). For PS task, as a first trial, we incorporate LSTM features trained on UCF-101 in order to capture the long-term temporal dynamics, which promote the performance by 0.6% (PS baseline-1 vs PS-contrast-1). We claim that if the LSTM models are trained on large video corpus, the features can be more discriminative and will offer better performance. Comparing PS baseline-1 and PS contrast-2, we found that concept features can slightly improve the performance. In addition, we can see that the deep learning based features (PS contrast-3) are significantly better than the conventional features (PS contrast-4), which corroborates the fact that deep learning features trained on ImageNet usually posses high generalization ability [5].

References

1. Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A.F., QuHenot, G., Ordelman, R.: Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: TRECVID 2015. (2015)
2. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. (2013)
3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR (2014)

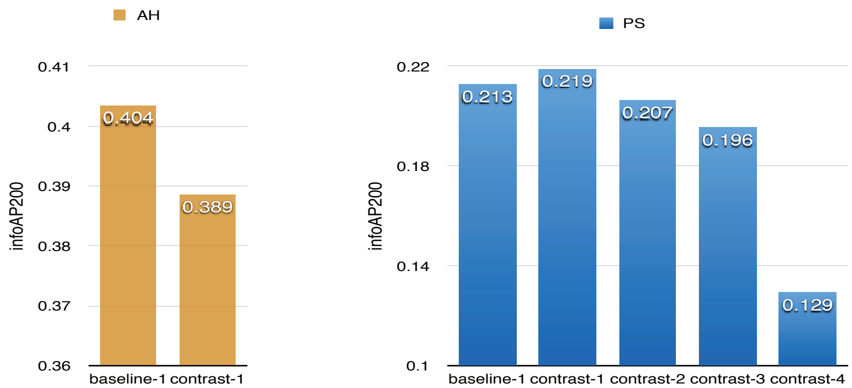

Fig. 2. The results of our submissions.

4. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. arXiv preprint arXiv:1502.07209 (2015)
5. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. CoRR (2014)