UTS-CMU at THUMOS 2015

Zhongwen Xu[†] Linchao Zhu[†] Yi Yang[†] Alexander G. Hauptmann[§]
[†]QCIS, University of Technology, Sydney [§]SCS, Carnegie Mellon University
{zhongwen.s.xu, zhulinchao7, yee.i.yang}@gmail.com alex@cs.cmu.edu

Abstract

This notebook paper describes our solution from UTS-CMU team in the THUMOS 2015 action recognition challenge. Our system contains two major components, video representation generated by VLAD encoding from ConvNet features and multi-skip improved Dense Trajectories. In addition, we explore optical flow ConvNet and acoustic features such as MFCC and ASR in our system. We demonstrate that our complete system can achieve state-of-the-art performance in large-scale action recognition tasks.

1. Introduction

This paper describes the solution from UTS-CMU team in the THUMOS action recognition challenge 2015. We investigate different state-of-the-art visual features to see how they perform in the action recognition task, especially for the untrimmed real-world videos. Action recognition has attracted much research attention in recent years. Along with the advances of visual features designed specifically for the action recognition task, great improvements on this task have been witnessed. The features designed for the action recognition have been shown very powerful for general video analysis as well, such as action localization task and multimedia event detection. However, most of the previous datasets utilized in the action recognition tasks are trimmed manually into clips with duration of several seconds, which is not realistic in the real-world application. In this notebook paper, we investigate this task in a different scenario, where the videos are temporally untrimmed and without any manually preprocessing. In THUMOS 2015 challenge, the whole dataset contains over 430 hours of video data and 45 million frames (70% larger than THUMOS 2014). Our investigation shows that we can achieve very promising performance in the large-scale real-world datasets. For the details for the THUMOS 2015 challenge, please refer to the challenge description [2].

2. Convolutional Neural Networks Video Representation

Our main component of the solution is the novel Convolutional Neural Networks (CNN) video representation proposed by Xu *et al*. [14], which is a general framework to adapt the CNN frame-level descriptors to generate the video representation. Average pooling on CNN based descriptors has been shown worse performance than the state-of-the-art hand-crafted feature improve Dense Trajectories (IDT) [13]. Instead of applying standard approaches such average pooling and max pooling on frame-level features, Xu *et al*. [14] utilize state-of-the-art encoding methods such as Fisher vectors (FV) [9, 10] and Vectors of Locally Aggregated Descriptors (VLAD) [5, 6] to generate the video representation. After extracting the CNN descriptors from fully-connected layers such as fc_6 and fc_7 for each frame, we aggregate all the frames into single video representation. For the utilization of the features from convolutional layers, which contain spatial information and may potentially improve the recognition accuracy, Xu *et al*. [14] propose a novel descriptor called latent concept descriptors (LCD). The latent concept descriptors (LCD) are generated by formulating the output of convolutional layers and pooling layers into multiple M-dimensional descriptors for each spatial location. The same encoding techniques as the features fc_6 and fc_7 are employed on LCD descriptors to generate the final representation. With these two contributions, the proposed video CNN representation achieves more than 30% relative improvement over the state-of-the-art video representation on the large scale TRECVID Multimedia Event Detection (MED) dataset. To accelerate the execution process in the video search, Xu *et al*. [14] conduct Product Quantization (PQ) [4] techniques to compress the representation, which saves the storage space by a factor of 32 and remains almost the same performance as the original representation.

In this notebook paper, we show that the schemes proposed in Xu *et al*. [14] are not specific for multimedia event detection tasks but also show great performance advantages over the hand-crafted features IDT in more general video

analysis tasks such as the THUMOS challenge [2].

3. Enhancement on Improved Dense Trajectories

In this submission, we employ recent enhancement [7] on improved Dense Trajectories [13] which extracts improved Dense Trajectories using a family of differential filters parameterized with multiple time skips and encodes shift-invariance into the frequency space. Lan *et al*. [7] is proposed by the observation that the same action may occur in different frequencies in the temporal scales, so we should consider multiple temporal scales when we conduct action recognition.

4. Experiment Results

For the Convolutional Neural Networks video representation part, we apply the state-of-the-art ConvNet model from Simonyan and Zisserman [12]. Specifically, we utilize the VGG-16 model kindly shared by Simonyan and Zisserman [12]. We experiment with the VGG-16 model and extract the features from layer fc_6, fc_7 and apply the VLAD encoding. In addition, we extract the LCD descriptors from the same model. We apply encoding on LCD descriptors extracted by a GoogLeNet with Batch Normalization [3] (denoted as Inception) reproduced by ourselves as well. To change into a better ConvNet, we can demonstrate the performance of video analysis can be directly enhanced by building the encoding framework upon a better underlying ConvNet model. For simplicity, we apply PCA reduction on all kinds of descriptors into 256 dimension, and the number of centers utilized in VLAD encoding (K) is mostly 256 except that we utilize $K = 512$ as well for LCD encoding from VGG-16 and Inception.

Beside the major two features, we utilize acoustic features MFCC and ASR as common practice in TRECVID MED task [15][1]. Furthermore, we reproduce the temporal stream ConvNet in two-stream ConvNets as described in [11] on UCF-101 datasets, and utilize the optical flow stream ConvNet from UCF-101 split-1 (which is from training part of THUMOS 2015) to extract features for the THUMOS data. We only sample 50 video segments with stacking length $L = 10$ due to our limited preparation time in the THUMOS challenge.

For all of the experiments, we utilize Support Vector Machines (SVM) [1] to classify the actions. When classifying one specific action, we regard all of the remaining actions (100 actions) as the negative samples to train the model[2]. The decision values from the classifier (linearly scaled into $[0, 1]$) are served as the ranking scores for each action. We fix $C = 100$ in SVM as the parameter shows consistently good performance across different features over $C = 1$ or $C = 0.01$.

4.1. Results for Validation 15 Dataset

We firstly show the results for each feature we utilize on the standard dataset validation 15 provided by the organizers [2]. Tables 1 shows the great advantages of VLAD encoding on ConvNet features [14] over the standard approach average pooling to generate the video representation from frame-level features. Table 2 shows the performance of our main component, the features generated from the methods proposed in Xu *et al*. [14], while Table 3 shows the results from other features for comparison. We can see from the tables clearly that the features generated by VLAD encoding on frame-level CNN descriptors outperforms other features in a significant level. And for the most powerful feature VLAD encoding on LCD descriptors, a better ConvNet model provides better performance on the final video classification performance (our trained Inception ConvNet model has better performance than VGG-16 model on ImageNet validation set), which verifies the general utilization of the framework proposed in [14].

	Average pooling	VLAD encoding
fc_6	0.521	**0.589**
fc_7	0.493	**0.566**

Table 1. Performance comparisons between average pooling and VLAD encoding [14, 6] for fc features on THUMOS val15 dataset.

	LCD	LCD from Inception	fc_6	fc_7
mAP	**0.619**	**0.628**	0.589	0.566

Table 2. Performance comparisons of features generated from [14] on (VLAD encoding) THUMOS val15 dataset.

	FlowNet	multi-skip IDT
mAP	0.416	0.547

Table 3. Performance of other features on THUMOS val15 dataset.

Note that a large amount of training data (UCF-101) videos are silent so we did not report the relatively poor performance on MFCC feature and ASR feature in this setting. After late fusion of video representation from [14], multi-skip IDT [7] [13], and FlowNet [11], we can achieve mAP **0.689** for THUMOS val15 dataset.

4.2. Results on Cross-validation Sets Generated from Validation 15

We follow the experiments from LEAR team last year [8] and conduct cross-validation on the training and validation set. We select 10 samples for each action from the val15 set

[1] LEAR team [8] utilized acoustic features as well in THUMOS challenge 2014

[2] We do not utilize background videos.

into the training set and all the remaining sample videos in the val15 set serve as the test data. The experiments are repeated for 5 times and the average performance is reported. Similar to the official validation setting, we compare the performance of video representation generated from average pooling over the frames and from VLAD encoding over the frame-level features, as shown in Table 4. We show the results for VLAD encoding on LCD descriptors from VGG-16 and Inception as well in Table 5. In Table 6, we list the performance from other features on cross-validation sets for reference.

	Average pooling	VLAD encoding
fc_6	0.653	**0.702**
fc_7	0.625	**0.686**

Table 4. Performance comparisons between average pooling and VLAD encoding [14, 6] for fc features on cross-validation sets.

	LCD	LCD from Inception	fc_6	fc_7
mAP	**0.738**	**0.746**	0.702	0.686

Table 5. Performance comparisons of features generated from Xu *et al*. [14] (VLAD encoding) on cross-validation sets.

	multi-skip IDT	MFCC	ASR	FlowNet
mAP	0.691	0.185	0.180	0.551

Table 6. Performance of other features on cross-validation sets

With increasing K in VLAD encoding from 256 to 512, LCD from VGG-16 improves from 0.738 to 0.746, and LCD from Inception improves from 0.746 to 0.752.

When fusing all the features with logistic regression, we obtain mAP **0.842**.

4.3. Results on THUMOS 15 Test Data

In the submissions, we utilize the following schemes (numbered with the Run ID): (1) Fuse two different kinds of features each time with grid search on fusion weights; (2) Average late fusion; (3) Logistic regression fusion on visual features; (4) Logistic regression fusion on all features; (5) Average late fusion on VLAD encoded CNN features [14] only. We choose the fusion parameters in (1) (3) and (4) on the cross-validation sets, since the cross-validation sets are more similar to the testing condition than the official validation set. The performance for each submitted run is shown in Table 7 respectively, which is evaluated by the organizers.

	mAP on Test set
Run 1	**0.738**
Run 2	0.716
Run 3	0.701
Run 4	0.691
Run 5	0.647

Table 7. Results on THUMOS 15 Test data

References

[1] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 2

[2] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://www.thumos.info/, 2015. 1, 2

[3] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 2

[4] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1):117–128, 2011. 1

[5] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010. 1

[6] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 34(9):1704–1716, 2012. 1, 2, 3

[7] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition. In *CVPR*, 2015. 2

[8] D. Oneata, J. Verbeek, and C. Schmid. The LEAR submission at THUMOS 2014. 2014. 2

[9] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*. 2010. 1

[10] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher Vector: Theory and practice. *IJCV*, 105(3):222–245, 2013. 1

[11] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2

[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

[13] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 1, 2

[14] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. In *CVPR*, 2015. 1, 2, 3

[15] S.-I. Yu, L. Jiang, Z. Xu, Z. Lan, S. Xu, X. Chang, X. Li, Z. Mao, C. Gan, Y. Miao, et al. Informedia@TRECVID 2014 MED and MER. *NIST TRECVID Workshop*, 2014. 2