NUS Submission to ActivityNet Challenge 2017: Evaluating frame-based and spatio-temporal CNN features for video classification

An Tran, Loong-Fah Cheong
Department of Electrical & Computer Engineering, National University of Singapore
an.tran@u.nus.edu, eleclf@nus.edu.sg

Abstract

This paper presents our findings when participating in ActivityNet challenges in untrimmed and trimmed video classification on ActivityNet and Kinetics dataset. Flow data are not as reliable as RGB data in challenging datasets such as ActivityNet. Furthermore, frame-based BN-Inception architecture performs better than spatio-temporal based C3D models.

1. Introduction

With the appearance of big video datasets (such as ActivityNet [1], Kinetics [4]), it would be interesting to evaluate different convolutional neural networks (CNNs) for action recognition. Two popular deep learning based representations for videos are frame-based CNNs features (*e.g.*, Inception [2]) and spatio-temporal CNNs features (*e.g.*, C3D [7]). In recent work, Wang *et al.* [9] shows that frame-based CNNs features achieves the state-of-the-art results on small video datasets such as UCF101 [6], HMDB51 [5]. In this submission to the challenge, we aims to evaluate the performance of frame-based and spatio-temporal CNNs on large video dataset. [1]

2. Our approach

Models. For frame-based features, we deploy BN-Inception [2] architecture with temporal-segment-networks (TSNs) proposed in [9]. For spatio-temporal features, we utilize the C3D architecture [7]. In C3D architecture, all convolution and pooling layers are made up of 3D operations [7]. C3D network has 8 convolution and 3 fully connected layers. C3D architecture is then given by: $C(3,64,1) - RL - P(2,2) - C(3,128,1) - RL - P(2,2) - C(3,256,1) - RL - C(3,256,1) - RL - P(2,2) - C(3,512,1) - RL - C(3,512,1) - RL - P(2,2) - C(3,512,1) - RL - C(3,512,1) - RL - P(2,2) - FC(4096) - D(0.9) - FC(4096) - D(0.8) - FC(101)$. Table 1 shows the number of parameters in BN-Inception and C3D architecture. As can be observed, C3D has more parameters than BN-Inception.

We also reports performances of some long-term temporal convolutions (LTC) models [8]. LTC models develop the ideas of C3D architectures into longer temporal dimension by reducing spatial dimension and increasing the temporal length.

Inception model is initialized from a pre-trained model on ImageNet and C3D weights are initialized from a pre-trained model on 1M-Sports dataset.

Regularization. For regularizing the capacity of a model, the TSNs [9] utilize batch normalization and high dropout ratios. We also set high dropout ratios for C3D networks with ratios of 0.9 and 0.8 for fully connected layers fc6 and fc7 respectively.

Input. The inputs for TSN models are 1 RGB frame for a segment of spatial-CNN stream and 5 stacked optical flow frames for a segment of temporal-CNN stream. The TSN networks average 3 segments of frames. In total, TSN networks operates on 3 RGB frames and 15 flows frames. The C3D networks operates on 16 continuous frames for both RGB and optical flows. The flows are extracted from OpenCV implementations of TV-L1 [10]. All the RGB and flow images are saved in resolution (128, 171). The flows have been compensated to remove camera motions. For the BN-Inception architecture in TSNs, the input images are resized into (256, 340) on the fly in our modifications of Caffe software [3]. On the other hand, C3D models operate on inputs of size (128, 171). Although this implementation is convenient for working with both BN-Inception and C3D architecture, it might slightly reduce performance of BN-Inception architecture because of lower quality resized image frames. Spatial and temporal models are trained individually.

Data augmentation. Data augmentation is shown to help deep convolutional models prevent severe over-fitting. For training, we adopt widely used data augmentation such as corner cropping, horizontal flipping, scale-jittering [9].

[1] Codes and models will be available at https://github.com/antran89/activitynet2017-nus.

	BN-Inception [2]	**C3D [7]**
parameters	10,373,765	78,409,573

Table 1. Number of parameters of different convolution networks: CaffeNet, and C3D.

3. Results

Table 2 shows performance of different models on ActivityNet validation set. The results show that both C3D and TSN-BN-Inception obtain better performances on RGB data than on compensated optical flows. For examples, rgb-C3D-size112-len16 out-performs flow counterpart flow-C3D-size-112-len16 17.09% (65.36% vs. 48.27% top-1 accuracy). The similar phenomenon happens with TSN-BN-Inception model (72.73% vs. 53.40%). On contrary to performances of deep models [9] on UCF101 and HMDB51, in challenging datasets such as ActivityNet, the compensated flows are not as reliable as RGB data. It is because motion patterns in ActivityNet and Kinetics dataset are more complex than small datasets (*e.g*., UCF101, HMDB51).

The second observation is that Inception models generally outperform C3D models both in RGB modality, while LTC-C3D models have more advantages in flow modality with longer temporal length. Furthermore, two-stream BN-Inception model perform better two-stream C3D. It indicates it would be hard to fit C3D models with spatio-temporal video data.

The third observation is that more information we can feed into a C3D model, more successful the model is. The C3D model has better performance if we increase temporal length both RGB and flow data. As can be observed, spatial resolution is more important for rgb-C3D (spatial-stream), while temporal resolution is more important for flow-C3D (temporal-stream). It can be explained as spatial-stream exploits context in whole videos, while temporal-stream focus more long-term human motion informations (*e.g*., human silhouettes, or shapes).

Due to limited time and resources, we take a chance to submit our evaluations on the test set of the challenge server with only rgb-LTC-maxpool-size112-len32 model. We got about 34% top-1 on ActivityNet untrimmed video classification task, and 30% average error (top-1 and top-5 errors) on Kinetics trimmed video classification.

4. Conclusions

This paper reports some comparisons between current frame-based BN-Inception and spatio-temporal C3D models. Without regards of models, RGB data are more reliable than (compensated) optical flows in challenging datasets such as ActivityNet and Kinetics. With regards of models, frame-based BN-Inception currently perform better than spatio-temporal based C3D models.

Models	**Top-1 acc.** (%)	**Top-3 acc.** (%)
rgb-C3D-size112-len16	65.36	81.43
rgb-LTC-maxpool-size112-len32	67.29	82.22
rgb-LTC-maxpool-size58-len32	56.13	72.80
flow-C3D-size-112-len16	48.27	65.96
flow-LTC-maxpool-size112-len32	59.10	74.53
flow-LTC-maxpool-size58-len32	53.65	70.19
rgb-TSN-BN-Inception	72.73	**87.16**
flow-TSN-BN-Inception	53.40	70.60
Two-stream C3D	70.36	83.95
Two-stream TSN-BN-Inception	**73.25**	86.30

Table 2. Performance of TSNs and C3D models on ActivityNet untrimmed video validation set. We show the results in top-1 and top-3 accuracy metrics. The suffix size112 means that the C3D model has input image of size (112, 112) and len16 means that the C3D model has temporal length of 16.

References

[1] B. G. Fabian Caba Heilbron Victor Escorcia and J. C. Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[2] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. feb 2015.

[3] 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*, jun 2014.

[4] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. may 2017.

[5] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563, 2011.

[6] K. Soomro, A. R. Zamir, and M. Shah. UCF101 : A Dataset of 101 Human Actions Classes From Videos in The Wild. Technical Report November, 2012.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, dec 2015.

[8] G. Varol, I. Laptev, and C. Schmid. Long-term Temporal Convolutions for Action Recognition. *arXiv:1604.04494*, apr 2016.

[9] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV 2016 - European Conference on Computer Vision*, aug 2016.

[10] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *In Ann. Symp. German Association Patt. Recogn*, pages 214–223, 2007.