Exploiting Image-trained CNN Architectures for Unconstrained Video Classification

Shengxin Zha[1]
szha@u.northwestern.edu

Florian Luisier[2]
fluisier@bbn.com

Walter Andrews[2]
wandrews@bbn.com

Nitish Srivastava[3]
nitish@cs.toronto.edu

Ruslan Salakhutdinov[3]
rsalakhu@cs.toronto.edu

[1] Northwestern University
Evanston IL USA

[2] Raytheon BBN Technologies
Cambridge, MA USA

[3] University of Toronto
Toronto, Ontario, Canada

In this paper, we propose an efficient approach to exploit off-the-shelf *image-trained* CNN architectures for video classification and evaluate on the challenging TRECVID MED'14 dataset and UCF-101 dataset. Our work is closely related to other research efforts towards the efficient use of CNN for video classification. While it is now clear that CNN-based approaches outperform most state-of-the-art handcrafted features for image classification, it is not yet obvious that this holds true for video classification. Moreover, there seems to be mixed conclusions regarding the benefit of training a spatiotemporal vs. applying an image-trained CNN architecture on videos. Although the specificity of the considered video datasets might play a role, the way the 2D CNN architecture is exploited for video classification is certainly the main reason behind these contradictory observations. The additional computational cost of training on videos is also an element that should be taken into account when comparing the two options. Prior to training a spatiotemporal CNN architecture, it thus seems legitimate to fully exploit the potential of image-trained CNN architectures. Obtained on a highly heterogeneous video dataset, we believe that our results can serve as a strong 2D CNN baseline against which to compare CNN architectures specifically trained on videos.

We conduct an in-depth exploration of different strategies for doing event detection in videos using convolutional neural networks (CNNs) trained for image classification (Figure 1, 2). We study different ways of performing spatial and temporal pooling, feature normalization, choice of CNN layers as well as choice of classifiers. Making judicious choices along these dimensions led to a very significant increase in performance over more naive approaches that have been used till now. The modality fusion of image-trained CNN features and motion-based Fisher vectors shows considerably improvement in classification performance. On TRECVID MED'14 dataset, our methods, based entirely on image-trained CNN features, can outperform several state-of-the-art non-CNN models. Our proposed late fusion of CNN- and motion-based features can further increase the mean average precision (mAP) on MED'14 from 34.95% to 38.74%, achieving the state-of-the-art (Table 1). On the challenging UCF-101 dataset, the image-based approach outperforms other image-trained CNN approaches; the fusion approach yields 89.6% classification accuracy (Table 2).

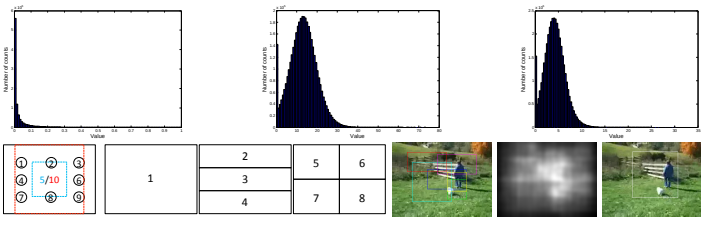

Figure 1: CNN features in output- and hidden-layer (top); spatial pyramid pooling and objectness-based pooling (bottom)

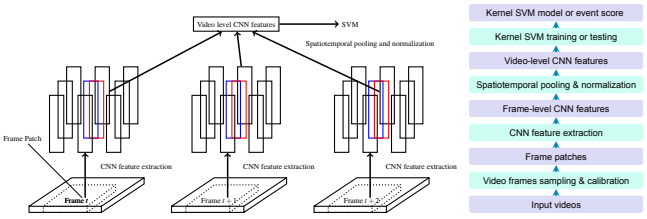

Figure 2: Overview of the proposed video classification pipeline.

Table 1: Result comparison on TRECVID MED'14 100Ex

Method	CNN	mAP
D-SIFT [3]+FV [5]	no	24.84%
IDT [7]+FV [5]	no	28.45%
D-SIFT+FV, IDT+FV (fusion)	no	33.09%
MIFS [2]	no	29.0 %
CNN-LCD$_{\text{VLAD}}$ with multi-layer fusion [8]	yes	36.8 %
proposed: CNN-hidden6	yes	33.54%
proposed: CNN-hidden7	yes	34.95%
proposed: CNN-hidden7, IDT+FV	yes	**38.74%**

Table 2: Result comparison on UCF-101

Method	Mean acc.
Spatial stream ConvNet [6]	73.0%
Temporal stream ConvNet [6]	83.7%
Two-stream ConvNet fusion by avg [6]	86.9%
Two-stream ConvNet fusion by SVM [6]	88.0%
Slow-fusion spatiotemporal ConvNet [1]	65.4%
Single-frame model [4]	73.3%
LSTM (image + optical flow) [4]	88.6%
MIFS [2]	89.1%
proposed: CNN-hidden6 only	79.34%
proposed: CNN-hidden6, IDT+FV (avg. fusion)	**89.62%**
proposed: CNN-hidden7, IDT+FV (avg. fusion)	89.30%

[1] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] Zhen-Zhong Lan, Ming Lin, Xuanchong Li, Alexander G. Hauptmann, and Bhiksha Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[4] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, pages 143–156, 2010.

[6] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[7] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[8] Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. A discriminative CNN video representation for event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.