DevNet: A Deep Event Network for Multimedia Event Detection and Evidence Recounting

Chuang Gan[1]* Naiyan Wang[2]* Yi Yang[3] Dit-Yan Yeung[2] Alexander G. Hauptmann[4]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, China. [2] Hong Kong University of Science and Technology [3] Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, Australia [4] School of Computer Science, Carnegie Mellon University, USA

* denotes equal contribution

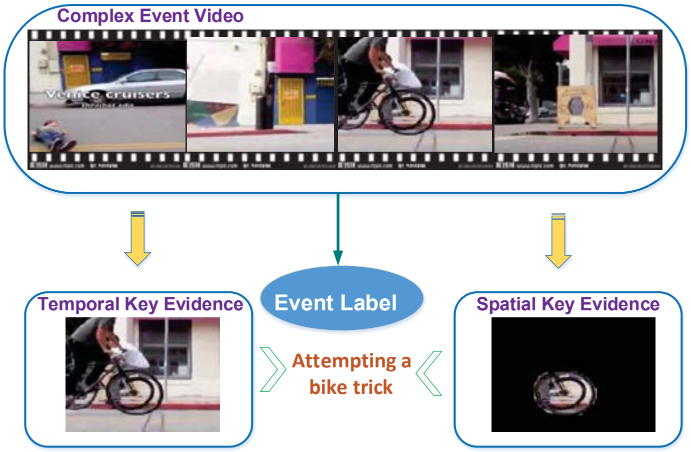

Figure 1: Given a video for testing, DevNet not only provides an event label but also spatial-temporal key evidences.

Learning features with Convolutional Neural Networks (CNNs) [1], has shown great potentials in various computer vision tasks giving state-of-the-art performance in image recognition and promising results in action recognition. The successes of CNNs also shed light on the multimedia event detection and recounting problems. However, whether and how the CNN architecture could be exploited for the video event detection and recounting problems has never been studied before, mainly due to the complexity and diversity of video events.

In this paper, we propose a flexible deep CNN infrastructure, namely Deep Event Network (DevNet), that simultaneously detects pre-defined events and provides key spatial-temporal evidences as shown in Figure 1. Taking key frames of videos as input, we first detect the event of interest at the video level by aggregating the CNN features of the key frames. The pieces of evidences which recount the detection results, are also automatically localized, both temporally and spatially. The challenge is that we only have video level labels, while the key evidences usually take place at the frame levels. Based on the intrinsic property of CNNs, we first generate a spatial-temporal saliency map by back passing through DevNet, which then can be used to find the key frames which are most indicative to the event, as well as to localize the specific spatial position, usually an object, in the frame of the highly indicative area.

Dataset	Evaluation Metric	IDTFV (SVM)	IDTFV (KR)	DevNet (SVM)	DevNet (KR)
MEDTest 14	AP	0.2696	0.2743	0.3288	**0.3329**
	MinNDC	0.4674	0.4493	**0.3687**	0.3699

Table 1: Event detection results comparing with improved dense trajectory Fisher vector (IDTFV). LOWER MinNDC / HIGHER AP indicates BETTER performance. The best results are highlighted in bold.

The framework of our DevNet is illustrated in Figure 2. To reduce the influence of limited training data, we first pre-train the DevNet using the largest image dataset to date, ImageNet, and then transfer the image-level features and train a new video-level event detector by fine-tuning the network. Next, we exploit the intrinsic property of CNNs [3] to generate a spatial-temporal saliency map without resorting to additional training steps. We only need to rank the saliency scores on the key frame level to localize the informative temporal evidences. For the top ranked key frames, we apply the graph-cut algorithm to the segmentation of discriminative regions as the spatial key evidences. Note that the localization process only utilizes the video-level event label without requiring the annotations of key frames and bounding boxes. Our work makes the following contributions:

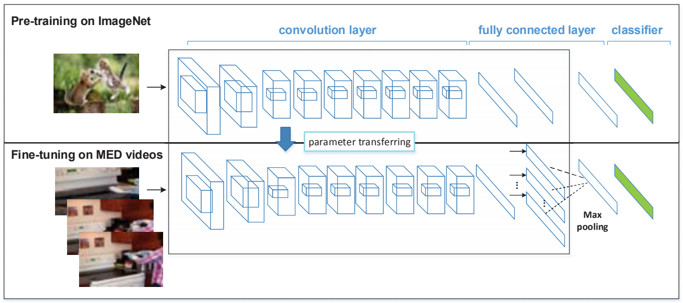

Figure 2: An illustration of the infrastructure of DevNet. We first pre-trained the DevNet using the ImageNet, and then fine-tuning on the MED video dataset.

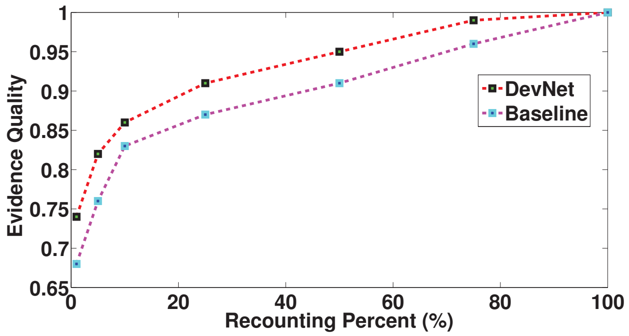

Figure 3: Comparison in terms of evidence quality against recounting percentage.

- To the best of our knowledge, we are the first to conduct high-level video event detection and spatial-temporal key evidence localization based on CNNs.
- This is the first paper that attempts to not only localize temporal key evidences (informative key frames and shots), but also provide discriminative spatial regions for evidence recounting.
- We show that our framework significantly outperforms state-of-the-art hand-crafted shallow features [2, 4] on event detection tasks as shown in Table 1 and achieves satisfactory results for localizing spatial-temporal key evidences as shown in Figure 3, which confirm the importance of representation learning for the event detection and evidence recounting tasks.

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[2] Dan Oneata, Jakob Verbeek, Cordelia Schmid, et al. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013.

[3] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[4] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.