Two-Stream Convolutional Networks for Action Recognition in Videos

Liangjie Cao

July 28, 2018

Abstract

The authors investigate architectures of discriminatively trained deep Convolutional Networks (ConvNets) for action recognition in video. The challenge is to capture the complementary information on appearance from still frames and motion between frames. They also aim to generalise the best performing hand-crafted features within a data-driven learning framework. Their contribution is three-fold. First, they propose a two-stream ConvNet architecture which incorporates spatial and temporal networks. Second, they demonstrate that a ConvNet trained on multi-frame dense optical flow is able to achieve very good performance in spite of limited training data. Finally, they show that multitask learning, applied to two different action classification datasets, can be used to increase the amount of training data and improve the performance on both. Their architecture is trained and evaluated on the standard video actions bench- marks of UCF-101 and HMDB-51, where it is competitive with the state of the art. It also exceeds by a large margin previous attempts to use deep nets for video classification.

1. Introduction

Recognition of human actions in videos is a challenging task which has received a significant amount of attention in the research community. Compared to still image classification, the temporal component of videos provides an additional (and important) clue for recognition, as a number of actions can be reliably recognised based on the motion information. Additionally, video provides natural data augmentation (jittering) for single image (video frame) classification. In this work, they aim at extending deep Convolutional Networks (ConvNets), a state-of-the-art still image representation, to action recognition in video data. This task has recently been addressed in [2] by using stacked video frames as input to the network, but the results were significantly worse than those of the best hand-crafted shallow representations. They investigate a different architecture based on two separate recognition streams (spatial and temporal), which are then combined by late fusion. The spatial stream performs action recognition from still video frames, whilst the temporal stream is trained to recognise action from motion in the form of dense optical flow. Both streams are implemented as ConvNets. Decoupling the spatial and temporal nets also allows us to exploit the availability of large amounts of annotated image data by pre-training the spatial net on the ImageNet challenge dataset. Our proposed architecture is related to the two-streams hypothesis, according to which the human visual cortex contains two pathways: the ventral stream (which performs object recognition) and the dorsal the ventral stream (which performs object recognition) and the dorsal stream (which recognises motion); though they do not investigate this connection any further here.

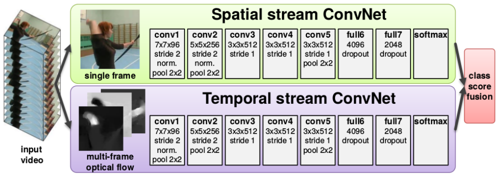

Figure 1. Two-stream architecture for video classification

2. Related work

Video recognition research has been largely driven by the advances in image recognition methods, which were often adapted and extended to deal with video data. A large family of video action recognition methods is based on shallow high-dimensional encodings of local spatio-temporal features. For instance, the algorithm of [3] consists in detecting sparse spatio-temporal interest points, which are then described using local spatio-temporal features: Histogram of Oriented Gradients (HOG) [1] and Histogram of Optical Flow (HOF). The features are then encoded into the Bag Of Features (BoF) representation, which is pooled over several spatio-temporal grids (similarly to spatial pyramid pooling) and combined with an SVM classifier. In a later work [4], it was shown that dense sampling of local features outperforms sparse interest points.

3. Two-stream architecture for video recognition

Video can naturally be decomposed into spatial and temporal components. The spatial part, in the form of individual frame appearance, carries information about scenes and objects depicted in the video. The temporal part, in the form of motion across the frames, conveys the movement of the observer (the camera) and the objects. They devise our video recognition architecture accordingly, dividing it into two streams, as shown in Fig. 1. Each stream is implemented using a deep ConvNet, softmax scores of which are combined by late fusion. We consider two fusion methods: averaging and training a multi-class linear SVM on stacked L 2 -normalised softmax scores as features.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1

[4] H. Wang, M. M. Ullah, A. Klser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 1