Signal Processing: Image Communication 70 (2019) 14–20

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Temporal stochastic linear encoding networks

Zhigang Zhu, Hongbing Ji *, Wenbo Zhang, Guopeng Huang

School of Electronic Engineering, Xidian University, Xi'an, 710071, China

ARTICLE INFO

Keywords:
Convolutional neural network
Action recognition
Long-range temporal dynamics
Motion boundary

ABSTRACT

Convolutional Neural Networks (CNNs) have achieved great success for action recognition. Technically, extracting effective long-range temporal dynamics is critical for such temporal tasks. This paper proposes a temporal stochastic linear encoding (TSLE) to construct the global video representations for action recognition, which can be embedded inside of CNN as a layer. The advantages of temporal stochastic linear encoding networks (TSLEN) are: (a) Compared with algorithms focusing on the short-term motions it can implement an easy yet robust manipulation of long range temporal clues. (b) We propose an arbitrarily directional motion boundary (ADMB) unit, which can save the training time and hard disk space. (c) The proposed TSLE unit maps the highly-dimensional videos to the compact spatio-temporal representations. On the efficiency and recognition accuracy experimental results demonstrate that the proposed TSLENs achieve competitive performance among the effective algorithms.

1. Introduction

Dramatic progress has been achieved by convolutional neural networks (CNN) on video-based recognition tasks [1–4] owing to its applications in many areas like security and behavior analysis. The main problems remain to be solved, in action recognition, are how to extract long-range temporal dynamics [5–7]. Recent works such as [1,2,8–10] have pointed out that long-range temporal dynamics are very important cues for action classification. Indeed, efficient video architectures should allow for temporal evolution of the long-range video sequences. However, mainstream proposals generally enable ConvNet frameworks [5–7,11] to extract short-range motions, and some attempts can primarily be categorized into two practices [5,6]. The first type is that 2D convolutions are extended to 3D spatio-temporal filters for the sake of temporal evolution of multiple clips [12]. Limited to the large-scale parameters and computational costs caused by the extra temporal dimension, 3D-CNN seems relatively uncompetitive. Specifically, 3D CNNs abandon the benefits of transfer learning, in which the CNN architectures are initialized by the models matured in 2D image domain. By decoupling spatial and temporal domains the second type utilizes video frames and optical flow fields to train ConvNets [5] respectively. The former focuses on the single frame appearance while the latter, by stacking multiple optical flow images, exacts the short-range motions. However, the advantages of CNNs over traditional methods are not so evident under the constraints of modeling capacities of shallow networks and the extremely-starved video datasets. Therefore, very deep two-stream ConvNets (e.g. VGGNet [13], GoogLeNet [14]) are trained to boost the recognition performances [15,16]. Based on the decoupled idea, the trajectory-pooled deep convolutional descriptors [7] are proposed to accumulate the feature maps along motion trajectories, and achieve competitive performances of recognition.

Despite the competitive performances of short-range dynamics algorithms, long-range temporal dependency remains to be crucial to further improve the video classification. Following the pipeline that motivates long-range temporal models, there are a few attempts [1,2,8,17–21] to construct the long-range dependency. *First*, restricted in the postulate that a function is capable of ordering the video frames temporally, the parameters of the rank-pooling machines well capture long range temporal information for action recognition [1,2,8] and elaborately summarize the video contents into compact representations, namely dynamic image and dynamic feature map respectively [22]. More importantly, either dynamic image or dynamic map network can directly fine-tune the existing CNN models advanced in 2D image. Nevertheless, learning the parameters of the rank pooling machines in an end-to-end manner suffers the optimization difficulty. It is necessary to solve the precise but not approximate gradients in the stochastic gradient descent process, and solving the specific gradients with respect to the parameters is a non-trivial challenge. *Second*, complex long-range temporal dynamics can be modeled by Long-term Recurrent Convolutional Networks (LRCNs) [19], which are directly connected to modern visual CNN models. It also can be jointly trained to simultaneously learn temporal

* Corresponding author.
E-mail address: hbji@xidian.edu.cn (H. Ji).

https://doi.org/10.1016/j.image.2018.09.001
Received 16 March 2018; Received in revised form 31 July 2018; Accepted 4 September 2018
Available online 6 September 2018

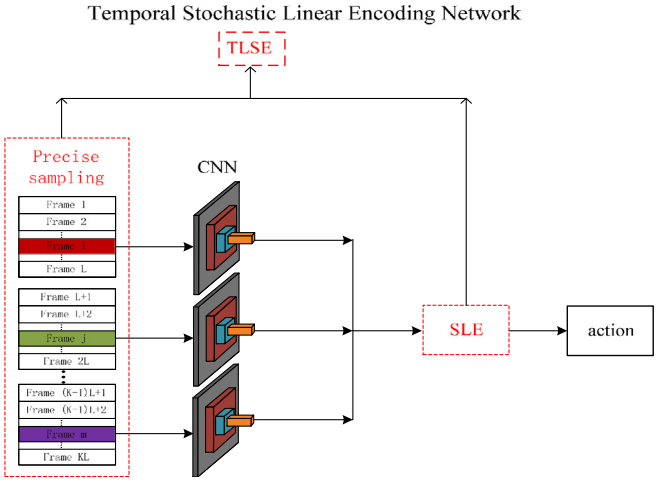

Fig. 1. Temporal Stochastic Linear Encoding network (TSLEN) for action recognition. It consists of two independent parts, precise sampling and stochastic linear encoding (SLE), which jointly encodes the long-range temporal dynamics.

dynamics and spatial representations. Incorporating nonlinearities into the network state update is appealing in that RNNs [23–25] can map variable-length inputs to variable-length outputs and model the complex dynamics of videos. Meanwhile the RNN [26–28] can be optimized with the back-propagation, which means that the CNN + LSTM [19] are end-to-end trainable.

To construct the long-range temporal dynamics, one of the common practices is to sample multiple clips and classic approaches such as bag-of-visual-words can be considered to encode multiple features. We know that a complete action is composed of multiple sub-actions. However, owing to the completely random sampling process in each segment [8,17], the sampled clips in the adjacent segment may be too closer to represent the various sub-actions. What comes with it is information redundancy of the same sub-action, that is, the sampled ones between the consecutive segments have risk of missing the next sub-action, which can lead to the misclassification of ConvNets.

Due to the fact that almost all the actions in UCF101 and HMDB51 constitute of multiple sub-actions uniformly spanning the temporal dimension and complex actions tend to be composed of multiple stages, how to keep the uniformity of video sampling is the first question worth exploring. With the counter clips generated in the randomly sampling process, the learned parameters will be biased towards the direction that is not conducive to correct classification. In addition, various encoding methods are devised to aggregate the snippet-level features to a video-level extraction. TLE constructs the outer product of convolutional features accumulating the features into a super vector [21], Evenly averaging and maximum implement the point-wise sum and max operators [17], 3D pooling conducts the clip-level consensus in temporal axis, and VLAD introduces (i) partition of data, (ii) partition of feature, and (iii) local PCA for codebook enhancement [29], which can be integrated together to boost the performances. We make finding that the forms of encoding methods remain to be an open question.

Motivated by the above observations, we propose a novel spatio-temporal encoding method called temporal stochastic linear encoding (TSLE) shown in Fig. 1, which aims to be at the spot aggregating highly-dimensional video features into compact representations with lower dimensions. To this end, the TSLE can be implemented in two steps. First, we propose a novel precise sampling to ensure that the key sequences are sampled as much as possible in order to make the sampling process miss the counter samples. Second, the final feature representations over the whole video can be constructed through a novel stochastic linear encoding (SLE). Notably, both parts are indispensable to the goal, and combination of the precise sampling and the stochastic

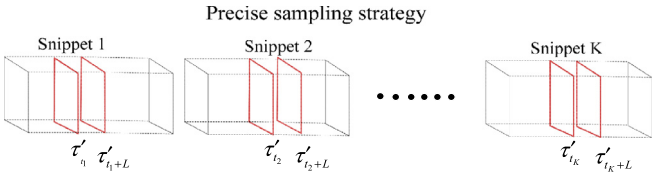

Fig. 2. Precise sampling strategy: One input video is divided into K segments and a short snippet is randomly selected from each segment. The selection range $\tau'_{t_i} \in [N/3K, 2N/3K - L]$ determines the uniform sampling process.

linear encoding can comprehensively construct the long-range temporal clues. The goal of the paper is not to achieve high performance, but to show that TSLEs are computationally efficient, robust, and compact. To confront with high risk of over-fitting, very deep ConvNets [13,30] are introduced to train the CNNs, and the Kinetics dataset is leveraged to unleash full potentials.

The rest of the paper is organized as follows: Section 2 describes our temporal stochastic linear encoding networks. This is followed by the datasets and implementation details in Section 3, and the experiments are clarified in Section 4. We conclude this paper in Section 5.

2. Temporal stochastic linear encoding (TSLE)

In this section, we give a detailed description of temporal stochastic linear encoding. We first verify whether the counter samples exist and then introduce the principles of the stochastic linear encoding (SLE) that conducts robust and compact representations. Finally, based on the hand-crafted motion boundary [31,32] we introduce how to construct arbitrarily directional motion boundary (ADMB) to save disk spaces and reduce the training time, which is embodied inside of CNNs and trained in an end-to-end manner.

2.1. Precise sampling strategy

In the common algorithms, the usual routine is to separate one video into multiple segments, and each segment generates one corresponding snippet randomly. That is, K segments spanning the same video generate K different stochastic clips, under the circumstances that the clips sampled in any position of segments do not amount to uniform segments. There is a situation that the adjacent snippets are randomly chosen at very close interval, which may result in the redundant characteristics and miss the key elements in the components of action. Therefore, the completely random sampling strategy may cause a degree of uneven sampling on the whole video. In this case certain snippets in the video sequences may be not relevant to the action itself, and we can call such samples counter clips. Once selected in the sampled process, they will be marked as the corresponding action label as well as the representative ones. This tends to move the trained CNN parameters towards an erroneous inductive bias path and lead to inferior performances in the test.

Motivated by the analysis above, we proposed a precise snippet-sampling strategy taking into account randomness and representativeness of clips (see Fig. 2). Given the video with N sequences, the stacked length of snippets L, and the segment set $S_1, S_2, \ldots, S_K$, we change the sampling range from $[0, N/K - L]$ to $[N/3K, 2N/3K - L]$. By the root this setting is the greatest degree of avoiding counter samples. In this way the precise sampling can select each clips located in the middle range of each segment. This implementation can keep the distance between the sub-actions from being too close, ensuring that the sampled clips evenly distribute throughout the whole video as much as possible.