UntrimmedNets for Action Recognition and Detection

Hongzhi Liu

May 15, 2018

1 The Concept of UntrimmedNet

Action recognition has attracted extensive research attention in the past few years, and much progress has been made in computer vision community. However, current action recognition methods heavily rely on trimmed videos for model training which is expensive and time-consuming.

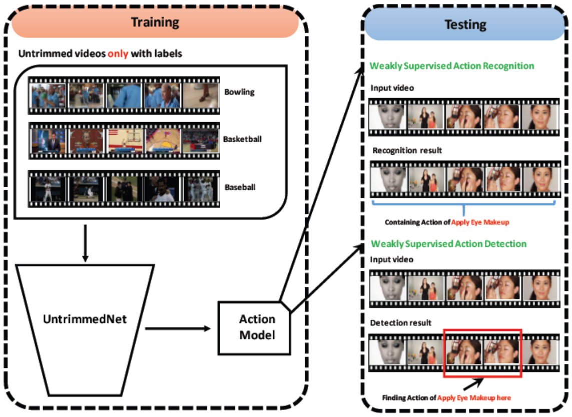
Training
Untrimmed videos only with labels
Bowling
Basketball
Baseball
UntrimmedNet
Action Model
Testing
Weakly Supervised Action Recognition
Input video
Recognition result
Containing Action of Apply Eye Makeup
Weakly Supervised Action Detection
Input video
Detection result
Finding Action of Apply Eye Makeup here

Figure 1: Weakly supervised action recognition and detection: during training phase, Professor Wang simply has untrimmed videos without temporal annotation and he trains action models from these untrimmed videos directly; during test phase, the learned action models could be applied to action recognition (WSR) and detection (WSD) in untrimmed videos.

In order to overcome the above limitations of using trimmed videos for training, Professor Wang introduces a more efficient setting of directly learning action recognition models from untrimmed videos as shown in Figure 1. The team call this new problem as *weakly supervised action recognition (WSR) and detection (WSD)* because of not having precise temporal annotations of action instances in training.

Professor Wang copes with the challenges of the WSR and WSD problems by proposing a new weakly supervised end-to-end architecture, called *UntrimmedNet* [1] to deal with the problems of WSR and WSD. The UntrimmedNet is mainly composed of three components, namely a *feature extraction module*, *classification module* and a *selection module*, which handle the problems of learning action models and detecting action instances respectively.

2 the Architecture of UntrimmedNet

In the paper of Professor Wang, I can learn the architecture of UntrimmedNet which is composed of a feature extraction module, a classification module, and a selection module as shown in Figure 2.

First, I learn about feature extraction module. After proposal generation, these shot clips are fed into deep networks for feature extraction. Formally, given a video V with a set of clip proposals $C = \{c_i\}_{i=1}^N$, they extract the representation as $\theta(V; c) \in \mathbb{R}^D$ for each clip proposal c. Their UntrimmedNet is a general framework for weakly supervised action recognition and detection, and does not depend on the choice of feature extraction network.

Then in the classification module, Professor Wang aims to classify each clip proposal c into the

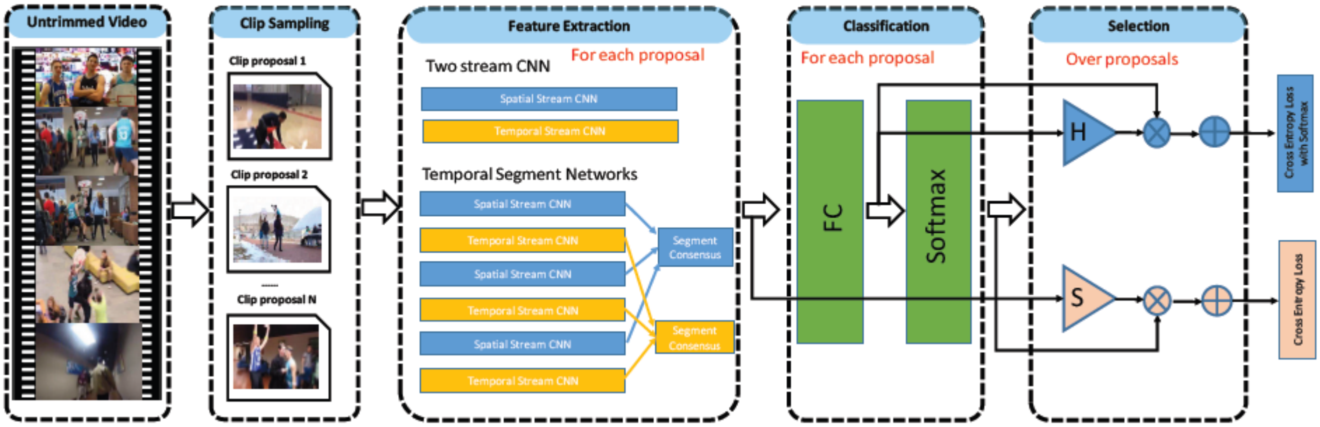

Figure 2: Pipeline of learning from untrimmed videos: the UntrimmedNets start with clip proposal generation, where we sample a set of short clips from the continuous untrimmed videos. Then, these clip proposals are separately fed into pre-trained networks for feature extraction. After this, a classification module is designed to perform action recognition for each clip proposal independently, and a selection module is proposed to detect or rank important clip proposals. Finally, the outputs of classification module and selection module are combined to yield the video-level prediction.

predefined action categories based on the extracted features $\theta(c)$. Suppose having C action classes, they learn a linear mapping $W^c \in \mathbb{R}^{C \times D}$ to transform the feature representation (c) into a C-dimensional score vector $X^c(c)$, i.e., $X^c(c) = W^c\theta(c)$, where C is the number of action categories and W^c are the model parameters. This score vector can be also passed through a softmax layer as Equation (1):

$$\bar{x}_i^c(c) = \frac{exp(x_i^c(c))}{\sum_{k=1}^{c} exp(x_k^c(c))} \tag{1}$$

where $x_i^c(c)$ denotes the i^{th} dimension of $x^c(c)$. For clarity, they use the notation $x^c(c)$ to denote the original classification score of clip proposal c and $\bar{x}^c(c)$ to represent the softmax classification score.

Furthermore, the selection module aims to select those clip proposals of most probably containing action instances. In the hard selection method, they try to identify a subset of k clip proposals for each action class. Besides, in the soft they want to combine the classification scores of all clip proposals and learn an importance weight to rank different clip proposals.

3 Evaluation of UntrimmedNet

Professor Wang evaluates their UntrimmedNet on two large datasets, namely THUMOS14 and ActivityNet [2]. These two datasets are suitable to evaluate the method because they provide the original untrimmed videos. In this experiment, they use the two stream CNNs for feature extraction in the UntrimmedNet and seven clips are randomly sampled from each video. The numerical results are summarized in Figure 3.

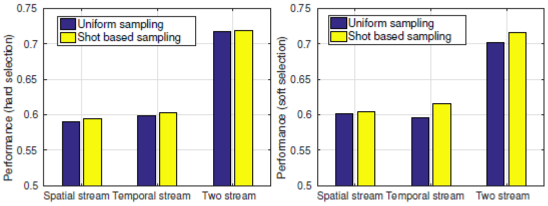

Figure 3: Comparison of different clip proposal sampling methods on the THUMOS14 dataset.

The team choose two baseline methods to compare: the standard temporal segment network with the average aggregation function (TSN), which is the state-of-the-art action recognition method, and TSN with more segments, which uses more segments during training. The numerical results are summarized in Table 1. From these results, we first see that our UntrimmedNet equipped with a hard or soft selection module outperforms the original TSN frameworks on both datasets.

Table 1: Effectiveness of selection module on the problem of weakly supervised action recognition (WSR).

Method	THUMOS14	ActivityNet (a)	ActivityNet (b)
TSN (3 seg) [3]	67.7%	85.0%	88.5%
TSN (21 seg)	68.5%	86.3%	90.5%
UntrimmedNet (hard)	73.6%	87.7%	91.3%
UntrimmedNet (soft)	74.2%	86.9%	90.9%

References

[1] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Computer Vision and Pattern Recognition*, pages 6402–6411, 2017.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. pages 961–970, 2015.

[3] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. 22(1):20–36, 2016.