论文题目

Point-Based Multi-View Stereo Network

论文作者

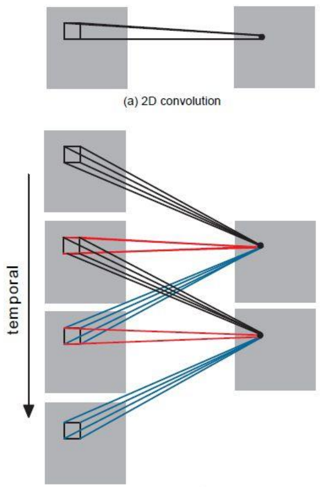
Rui Chen, Songfang Han, Jing Xu, Hao Su
Tsinghua University
The Hong Kong University of Science and Technology
University of California, San Diego

代码地址

http://hansf.me/projects/PMVSNet/
https://github.com/callmeray/PointMVSNet

论文的前提

一般的 2D 的 CNN 不能很好的捕捉时序信息,3D 的 CNN 会有一个 depth 的概念来捕捉相邻帧的时序信息,下面的图中 depth=3



(a) 2D convolution

(b) 3D convolution

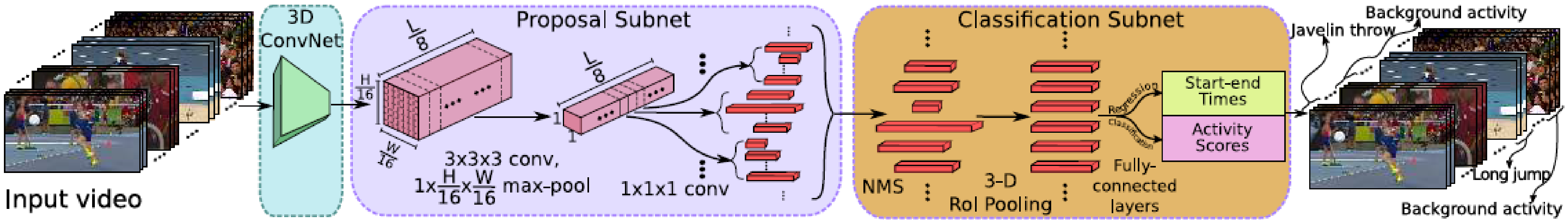
论文观点

以 C3D 网络为基础,借鉴了 Faster RCNN 的思路,对于任意的输入视频 L,先进行 proposal,然后 3D-pooling,最后后进行分类和回归操作

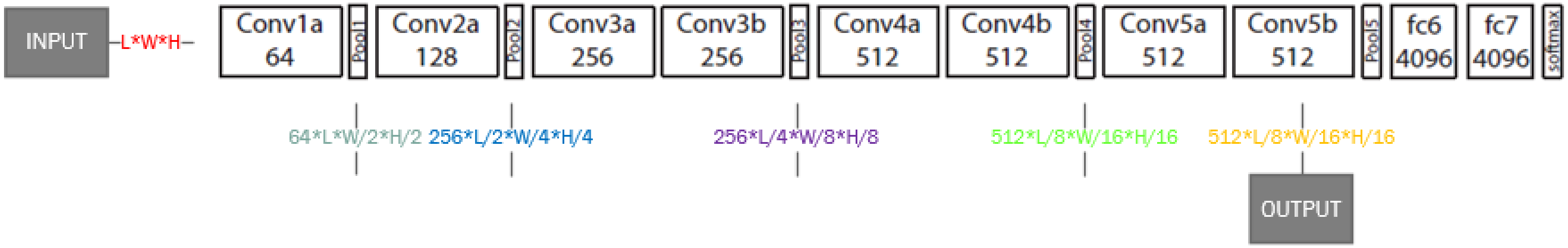
可以针对任意长度视频,任意分辨率,任意长度行为进行端到端的检测

共享 Proposal generation 和 Classification 网络的 C3D 参数使得速度很快

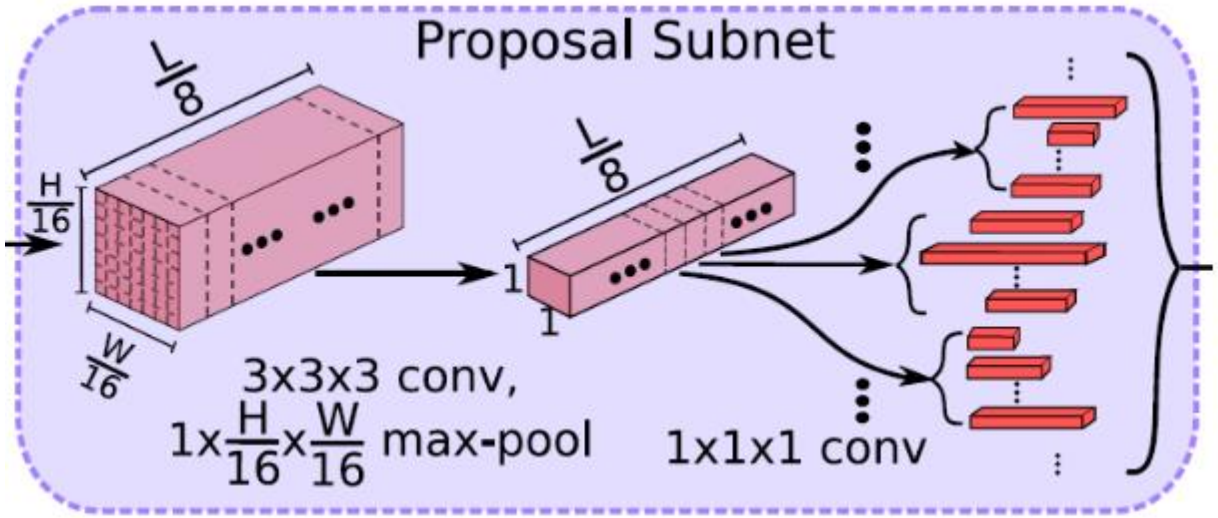
论文做法



1. C3D 特征提取网络(基本的 C3D 的 L=16,此处的更新在于 L 可以等于任意值)



2. Temporal Proposal(提取一系列可能存在行为的候选时序)



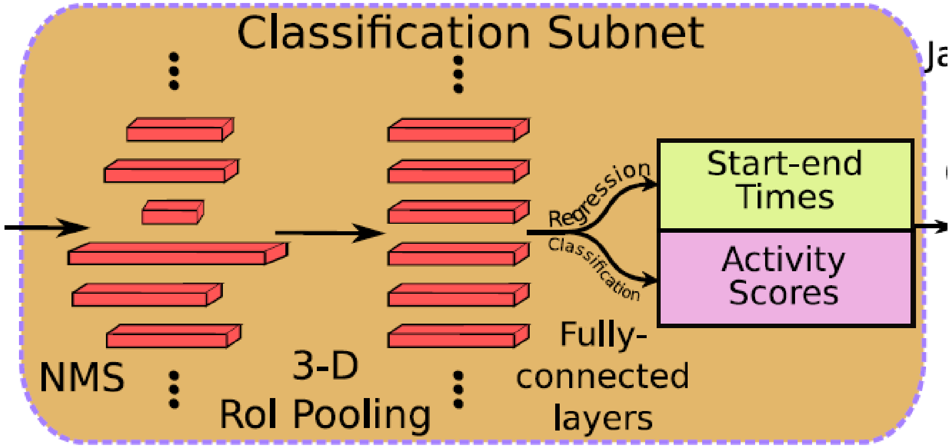
input: 512 * L/8 * H/16 * W/16

假设anchor均匀分布在L/8的时间域上,也就是有L/8个anchors,每个anchors生成K个不同scale的候选时序(时序上的操作)

为了获得每个时序点(anchor)上每段候选时序的中心位置偏移和时序的长度,将空间上H/16*W/16的特征图经过一个3*3*3的卷积核和一个3D pooling层下采样到1*1➔512*L/8*1*1

每个时序位置上的512维的特征向量用来预测中心位置和长度的相对偏移,同时它也预测了此Proposal是动作还是背景,最后两个1*1*1卷积预测提议偏移和提议分数

3. Activity Classification



非极大值抑制生成优质的 proposal➔3D ROI pooling 之后得到相同大小的 feature map
边框回归+类别分类

最终效果

Table 2. Per-class AP at IoU threshold $\alpha = 0.5$ on THUMOS'14 (in percentage).

	[20]	[39]	[24]	R-C3D (ours)
Baseball Pitch	8.6	14.6	14.9	**26.1**
Basketball Dunk	1.0	6.3	20.1	**54.0**
Billiards	2.6	**9.4**	7.6	8.3
Clean and Jerk	13.3	**42.8**	24.8	27.9
Cliff Diving	17.7	15.6	27.5	**49.2**
Cricket Bowling	9.5	10.8	15.7	**30.6**
Cricket Shot	2.6	3.5	**13.8**	10.9
Diving	4.6	10.8	17.6	**26.2**
Frisbee Catch	1.2	10.4	15.3	**20.1**
Golf Swing	**22.6**	13.8	18.2	16.1
Hammer Throw	34.7	28.9	19.1	**43.2**
High Jump	17.6	**33.3**	20.0	30.9
Javelin Throw	22.0	20.4	18.2	**47.0**
Long Jump	47.6	39.0	34.8	**57.4**
Pole Vault	19.6	16.3	32.1	**42.7**
Shotput	11.9	16.6	12.1	**19.4**
Soccer Penalty	8.7	8.3	**19.2**	15.8
Tennis Swing	3.0	5.6	**19.3**	16.6
Throw Discus	**36.2**	29.5	24.4	29.2
Volleyball Spiking	1.4	5.2	4.6	**5.6**
mAP@0.5	14.4	17.1	19.0	**28.9**

Table 3. Detection results on ActivityNet in terms of mAP@0.5 (in percentage). The top half of the table shows performance from methods using additional handcrafted features while the bottom half shows approaches using deep features only (including ours). Results for [29] are taken from [1]

	train data	validation	test
G. Singh *et. al.* [30]	train	34.5	36.4
B. Singh *et. al.* [29]	train+val	-	28.8
UPC [18]	train	22.5	22.3
R-C3D (ours)	train	**26.8**	**26.8**
R-C3D (ours)	train+val	-	**28.4**

Table 4. Activity detection results on Charades (in percentage). We report the results using the same evaluation metric as in [25].

	mAP	
	standard	post-process
Random [25]	4.2	4.2
RGB [25]	7.7	8.8
Two-Stream [25]	7.7	10.0
Two-Stream+LSTM [25]	8.3	8.8
Sigurdsson et al. [25]	9.6	12.1
R-C3D (ours)	**12.4**	**12.7**

[redacted]

Table 5. Activity detection speed during inference.

	FPS
S-CNN [24]	60
DAP [4]	134.1
R-C3D (ours on Titan X Maxwell)	**569**
R-C3D (ours on Titan X Pascal)	**1030**