论文题目

You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization

论文作者

Okan Köpüklü*, Xiangyu Wei*, Gerhard Rigoll

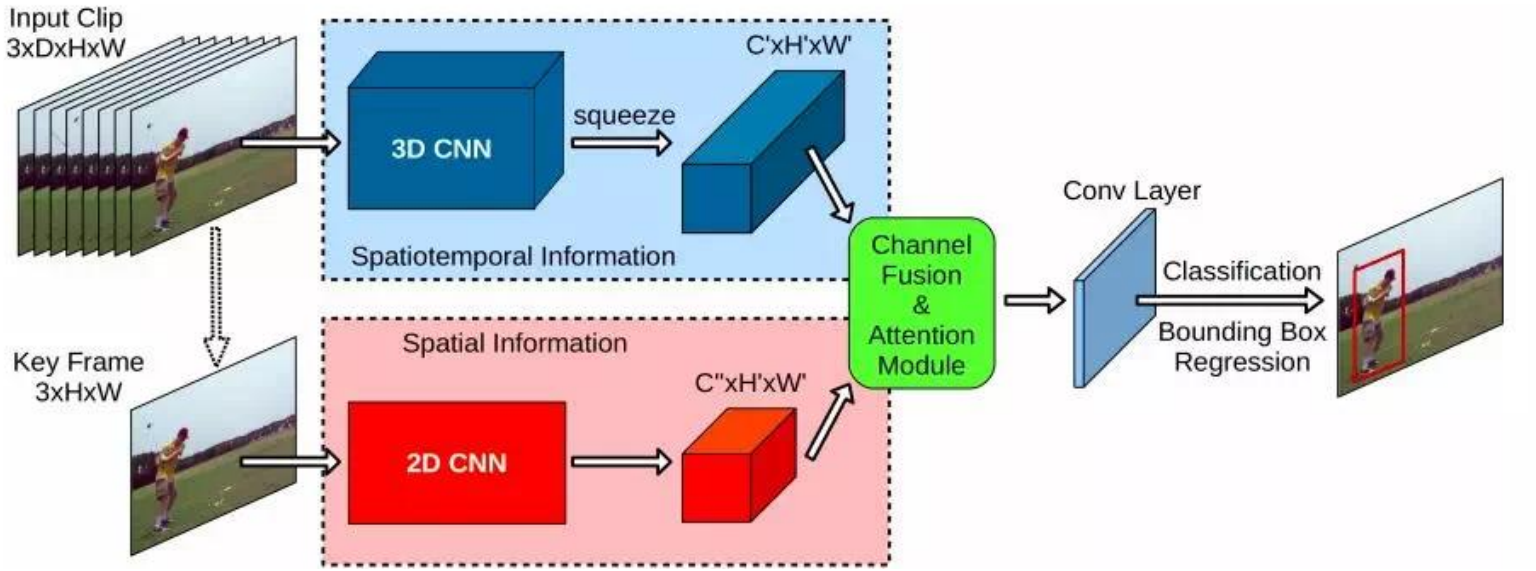
Technical University of Munich, Germany

论文代码地址

https://arxiv.org/pdf/1911.06644.pdf
https://github.com/wei-tim/YOWO

论文做法

1.整体网络图



2.提取特征部分

使用 3DCNN 获得前面帧的时间信息(3D 对时间信息 get 更有效)

使用 2DCNN 对关键帧提取空间信息(2D 对空间信息 get 更有效)

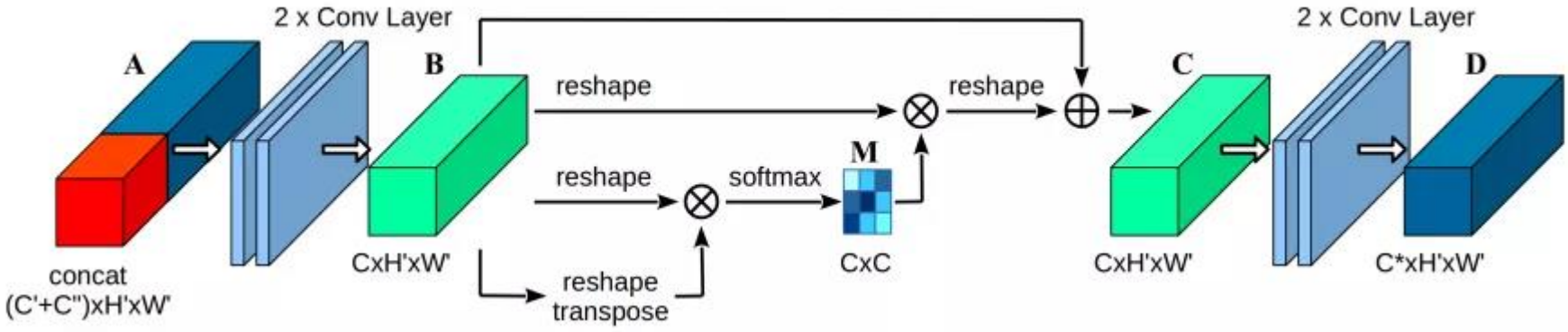
images ➔ 3DCNN(3DResNet-101):提取时空信息

输入(3*D*H*W)➔C_1*D*(H/32)*(W/32)➔C_2(C_1*D)*(H/32)*(W/32)

images ➔ 2DCNN(darknet-19):对关键帧提取空间信息

输入(3*H*W)➔C_3*(H/32)*(W/32)

3.融合部分(对两部分特征=>通道融合(concat) + 注意力机制➔特征聚合)



输入特征 concat

输入(C_2)*(H/32)*(W/32)+(C_3)*(H/32)*(W/32)➔ C_4(C_2+C_3)*(H/32)*(W/32)

经过两次卷积对不同 C 和不同 D(最后的 C_4 肯定不同)做处理

输入 C_4*(H/32)*(W/32)➔C_5*(H/32)*(W/32)[B]

对 B reshape 成 F(C_5*N((H/32)*(W/32))),F*F_T=M(C_5*N ** N*C_5)➔M(C_5*C_5),其中 M_ij 是第 j 通道对第 i 通道的影响

M_T*F=(C_5*N)➔reshape➔alpha*(C_5*(H/32)*(W/32))+[B]➔ C_5*(H/32)*(W/32)[C]

对[C]做两次卷积➔C_6*(H/32)*(W/32)[D]

4.卷积分类+框回归

一共有(H/32)*(W/32)个 cell,每个 cell 有 C_6 维度的信息,通过每个 cell 的 C_6 维度信息我们为这个 cell 预测 5 个 anchor

一个 anchor 是有框(4 个点)+这个框的 conf 以及这个框的类别的 softmax 组成(假设类别是 80 类,则是一个 80 维度的加和为 1 的向量)

最终要预测的是:(H/32)*(W/32)*5*[80+4+1]

最终效果

Method	Frame-mAP	Video-mAP		
		0.2	0.5	0.75
Peng w/o MR [24]	56.9	71.1	70.6	48.2
Peng w/ MR [24]	58.5	74.3	73.1	-
ROAD [32]	-	73.8	72.0	44.5
T-CNN [13]	61.3	78.4	76.9	-
ACT [17]	65.7	74.2	73.7	52.1
P3D-CTN [38]	71.1	84.0	80.5	-
TPnet [31]	-	74.8	74.1	**61.3**
YOWO (16-frame)	**74.4**	**87.8**	**85.7**	58.1

Table 5: Performance on dataset J-HMDB-21 and comparison with SOTA results by frame-mAP (%) under IOU threshold 0.5 and video-mAP (%) under different IOU thresholds.

Method	Frame-mAP	Video-mAP		
		0.1	0.2	0.5
Peng w/o MR [24]	64.8	49.5	41.2	-
Peng w/ MR [24]	65.7	50.4	42.3	-
ROAD [32]	-	-	73.5	46.3
T-CNN [13]	41.4	51.3	47.1	-
ACT [17]	69.5	-	77.2	**51.4**
MPS [1]	-	82.4	72.9	41.1
STEP [41]	75.0	**83.1**	**76.6**	-
YOWO (16-frame)	**87.2**	82.5	75.8	48.8