中国地质大学
CHINA UNIVERSITY OF GEOSCIENCES

Learning Temporal Co-Attention Models for Unsupervised Video Action Localization

Guoqiang Gong Xinghan Wang Yadong Mu Qi Tian

Accepted to CVPR 2020

小组成员：周宁

Outline

01 背景介绍

02 研究方法与思路

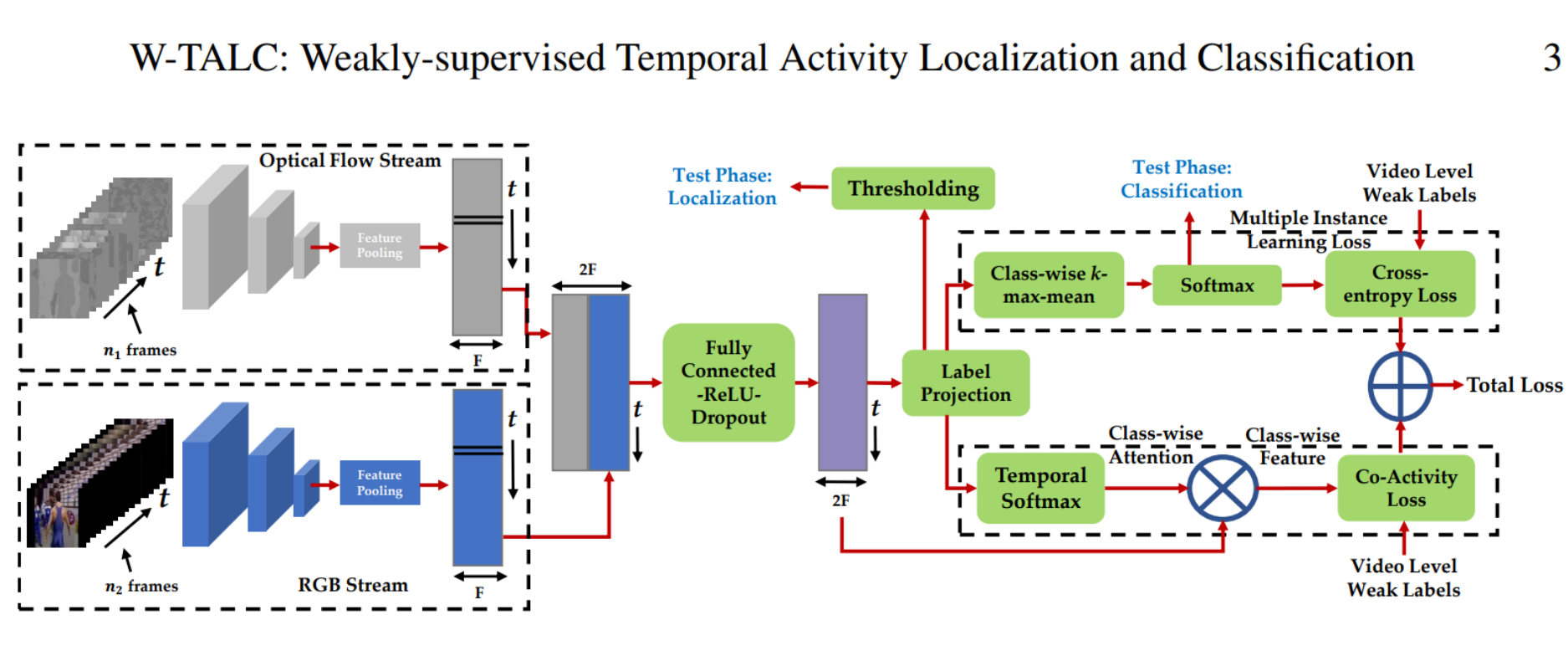
03 实验结果

04 总结

01 背景介绍

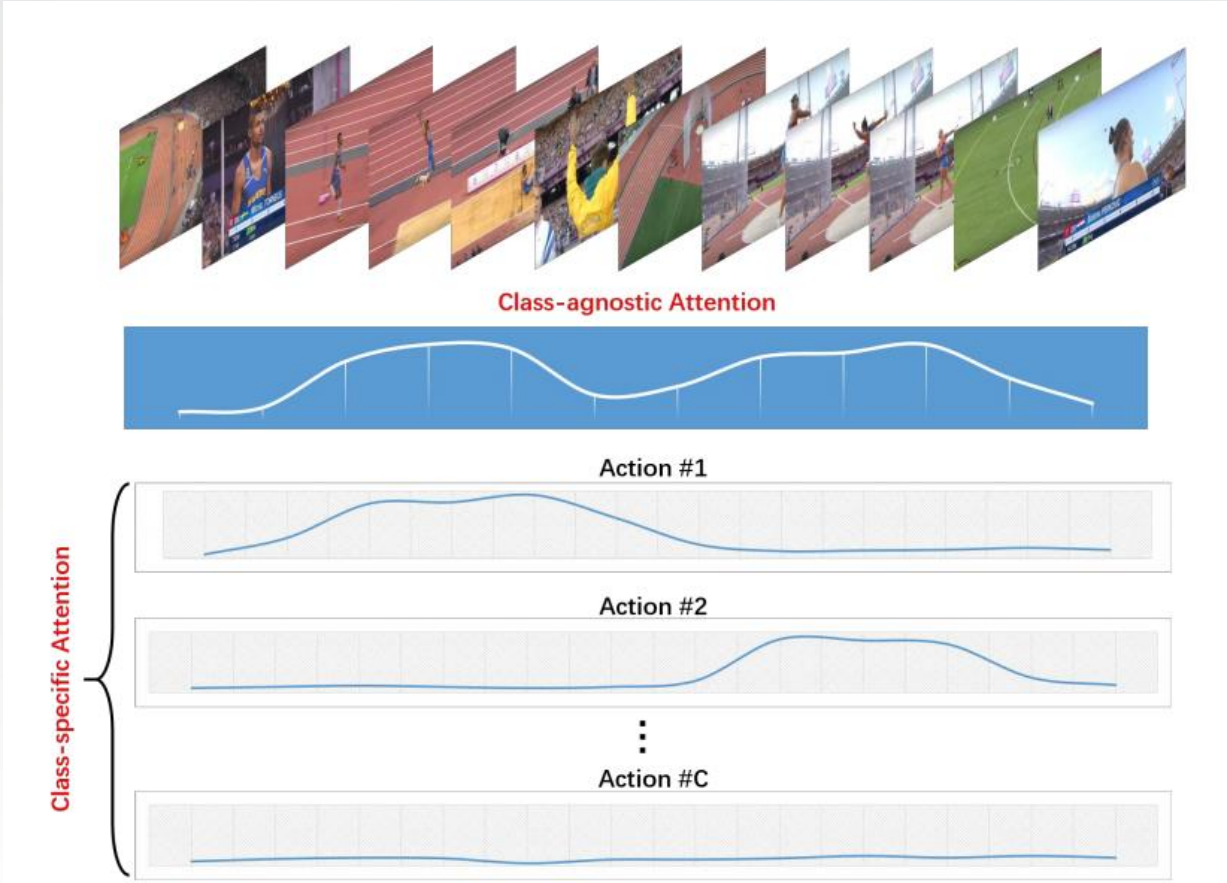
背景介绍
研究方法与思路
实验结果
总结

视频的时间动作定位（Temporal action localization，TAL），成为了一个重要的研究方向。目前有全监督、弱监督等时间动作定位的方法出现，但是无监督的方法还没有人提出。



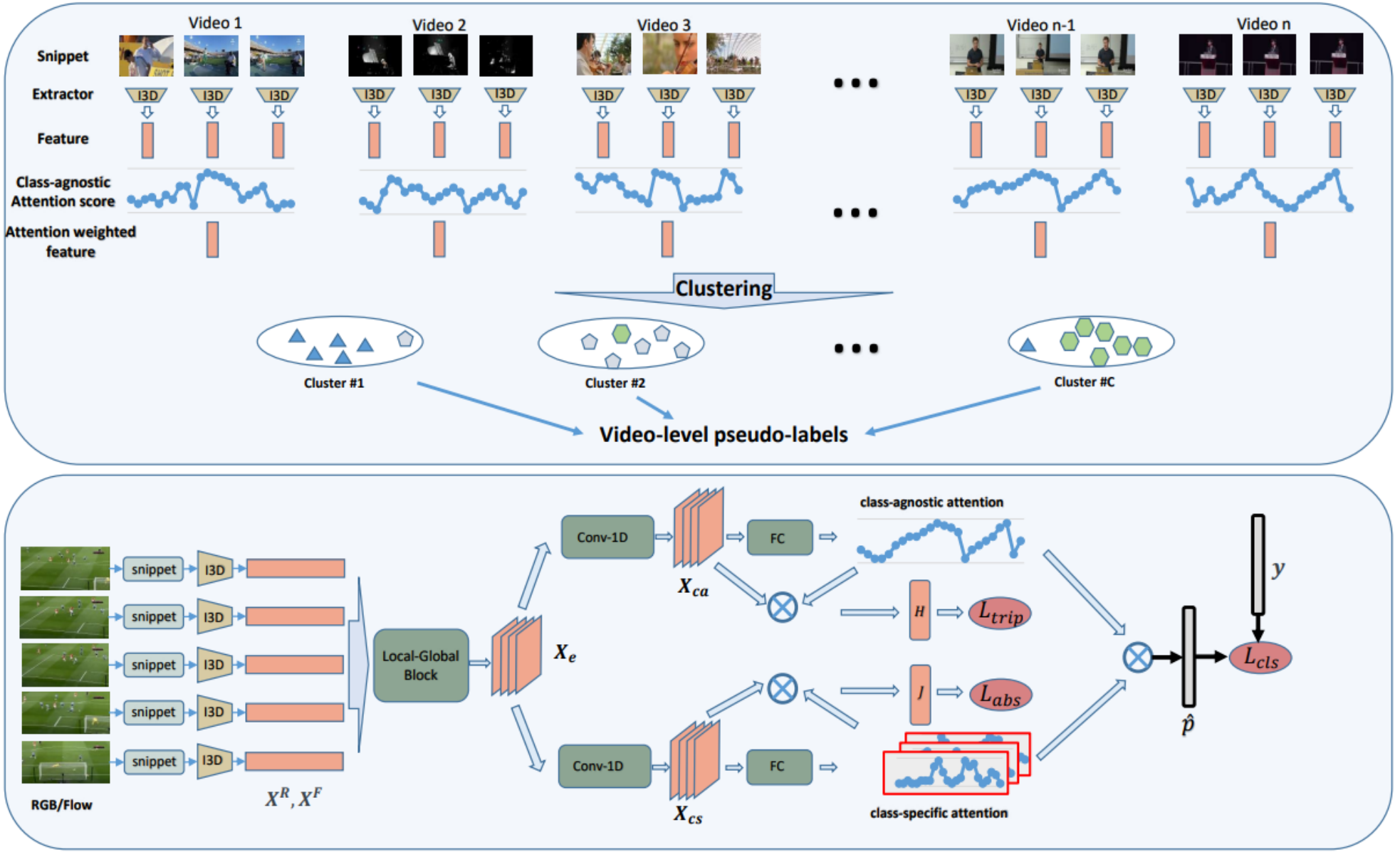
背景介绍
研究方法与思路
实验结果
总结

在这样的背景下，提出了一个“聚类+定位”的方法，解决无监督视频时间动作定位。
聚类步骤为定位步骤提供了noisy的伪标记，而定位步骤提供了时间共关注模型，从而提高了聚类性能，这两个过程相辅相成。



02 研究方法与思路

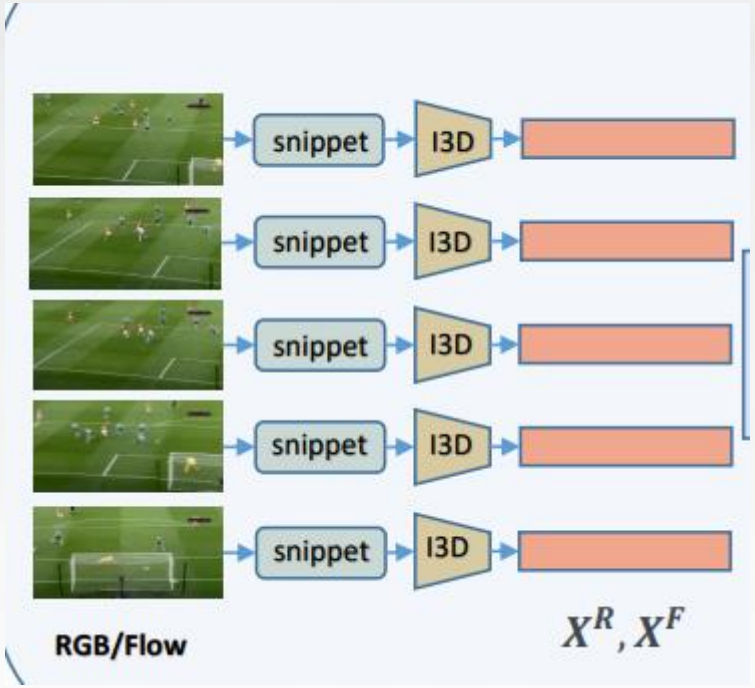
背景介绍
研究方法与思路
实验结果
总结



背景介绍
研究方法与思路
实验结果
总结

Video Feature Extraction

给定一个未修剪的视频，我们首先将其分成一组片段，每个片段由几个连续的帧组成。按照先前工作中的常规做法，我们提取每个片段的RGB和flow视频功能。$X^R, X^F \in \mathbb{R}^{T \times D}$，其中T代表片段的数量，D代表特征尺寸



背景介绍
研究方法与思路
实验结果
总结

Clustering

对于每个视频v,我们同样得到视频的RGB和flow特征X^R, X^F，令$S^R_{v,i}, S^F_{v,i}$为第i次迭代中的class-agnostic attention weights权重。因为这个是训练的时候才能得到所以最开始可以都设为$1/T$

$$S^R_{v,1}[j,1] = S^F_{v,1}[j,1] = \frac{1}{T_v}(1 \le j \le T_v)$$

对于视频v在迭代i产生的RGB特征和flow特征就能得到

$$f^R_v = L_2Norm((X^R_v)^\intercal S^R_{v,i}),\ f^F_v = L_2Norm((X^F_v)^\intercal S^F_{v,i}),$$

将f^R_v, f^F_v连接在一起就能得到总特征$\mathbf{f_i}$，利用f_i构建图结构

背景介绍
研究方法与思路
实验结果
总结

Clustering

对于图G = {V, E}，其中V表示顶点的集合，即训练集视频，E表示边的集合。其中v_i, v_j的权重$w_{i,j}$由

$$w_{ij} = \exp\left(-\frac{\|f_i - f_j\|_2^2}{2\sigma^2}\right)$$

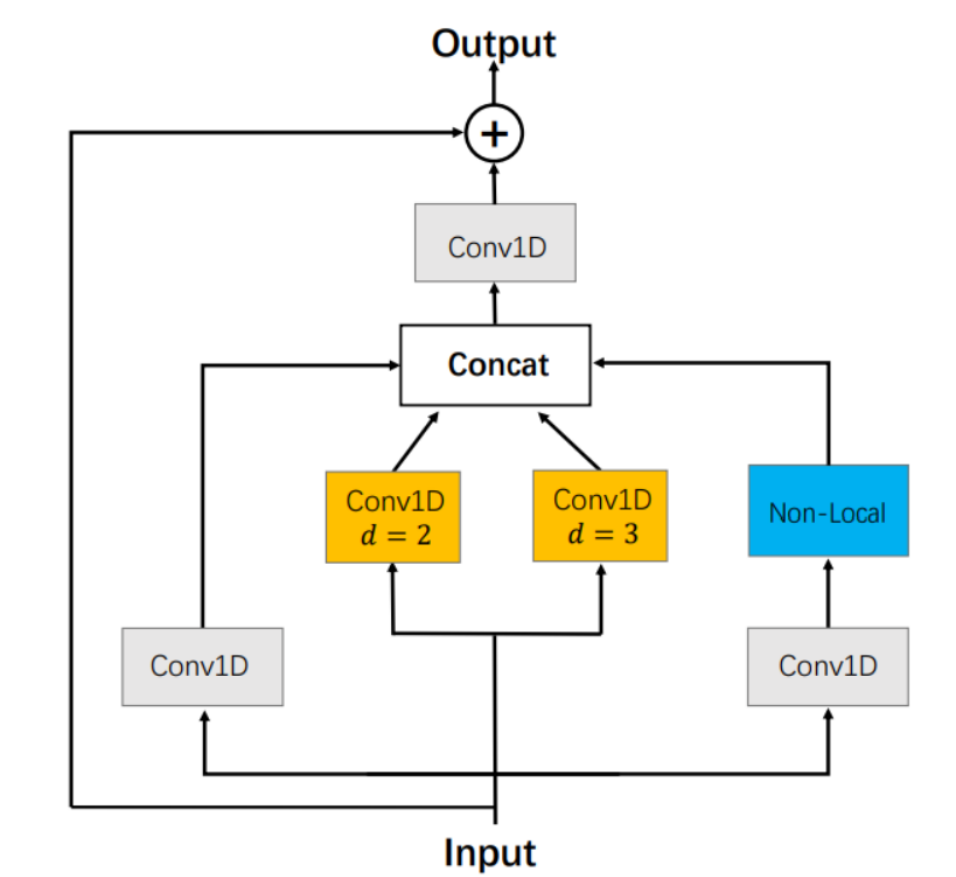
$$\sigma = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|f_i - f_j\|_2$$

计算得来。基于构造的图，使用频谱聚类算法将未修剪的视频分组为C个簇，每个簇都定义了一个伪动作。

背景介绍
研究方法与思路
实验结果
总结

Local-Global Feature Aggregation Block

由于每个段的特征仅包含当前片段的信息，因此缺少时间上下文信息。为了提高每个代码段特征的可分辨性，提出了局部全局特征聚合块（Local-Global Feature Aggregation Block，FAB）以提取局部和全局上下文信息



1. a 1D temporal convolution branch

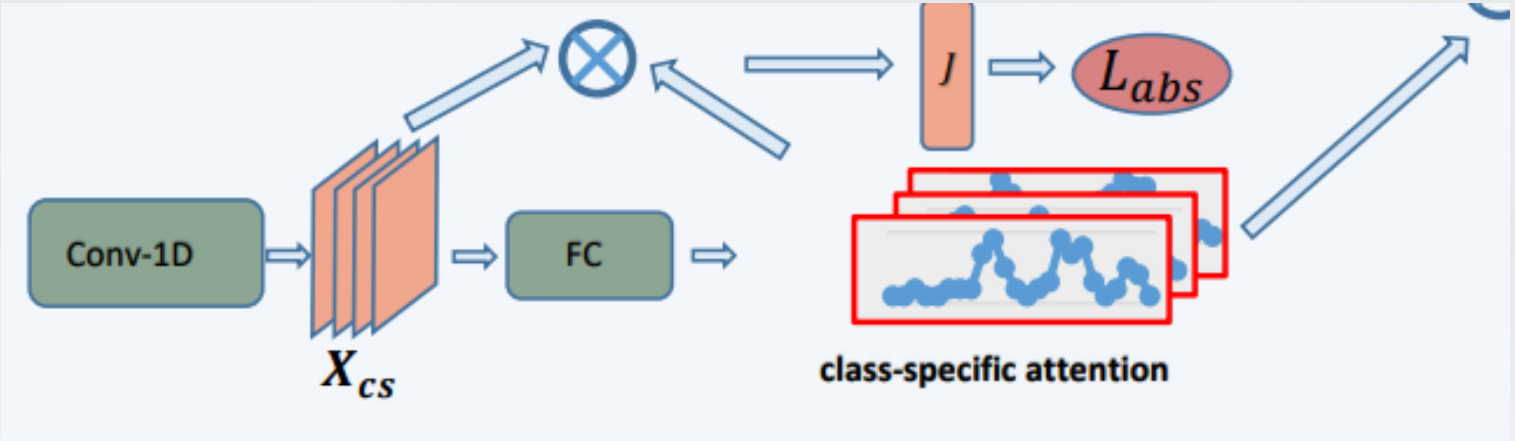
2. a dilated temporal pyramid branch

3. a global context branch

背景介绍
研究方法与思路
实验结果
总结

Class-Specific Temporal Attention Module

这个模块的功能主要是获得在不同时间出现的不同动作类别的概率。



以X_{cs}为中间层输入，输出类特定分数$A \in R^{T \times C}$,可以理解为A表示某一类动作的概率。最后再进行softmax来归一化

$$\hat{A} = softmax(A).$$

背景介绍
研究方法与思路
实验结果
总结

Class-Specific Temporal Attention Module

计算动作背景分离损失(action-background separation loss)。

对于一批训练视频，我们从随机训练集的C簇中，抽取出Z簇，再从Z簇中各自抽取出K个视频。定义，定义Vz为属于某以簇的K个视频的集合

$$V_z = \{v_k\}_{k=1}^{K}$$

计算，对于每个视频$\mathbf{v_k}$,我们计算动作特征和背景特征

$$J_k = X_{cs,k}{}^{\intercal} \hat{A}_k [:, z]$$

$$B_k = \frac{1}{T_k - 1} X_{cs,k}{}^{\intercal} \left(\mathbf{1} - \hat{A}_k [:, z] \right)$$

Class-Specific Temporal Attention Module

背景介绍
研究方法与思路
实验结果
总结

除此之外，还有以下限制：
假设我们有一对属于V_z的同族视频v_m, v_n，令d表示余弦距离函数，τ1和τ2分别表示两个cos余弦距离

$$d(J_m, J_n) \leq \tau_1.$$

$$\begin{aligned} d(J_m, B_m) - d(J_m, J_n) &\geq \tau_2, \\ d(J_n, B_n) - d(J_m, J_n) &\geq \tau_2. \end{aligned}$$

Class-Specific Temporal Attention Module

背景介绍
研究方法与思路
实验结果
总结

背景分离损失(action-background separation loss)

$$\mathcal{L}_{inter,z} = \sum_{m=1}^{K} \sum_{n=1,n\neq m}^{K} max\{d(J_m, J_n) - \tau_1, 0\}$$

$$\mathcal{L}_{intra,z} = \sum_{m=1}^{K} \sum_{n=1,n\neq m}^{K} max\{d(J_m, J_n) - d(J_m, B_m) + \tau_2, 0\}$$

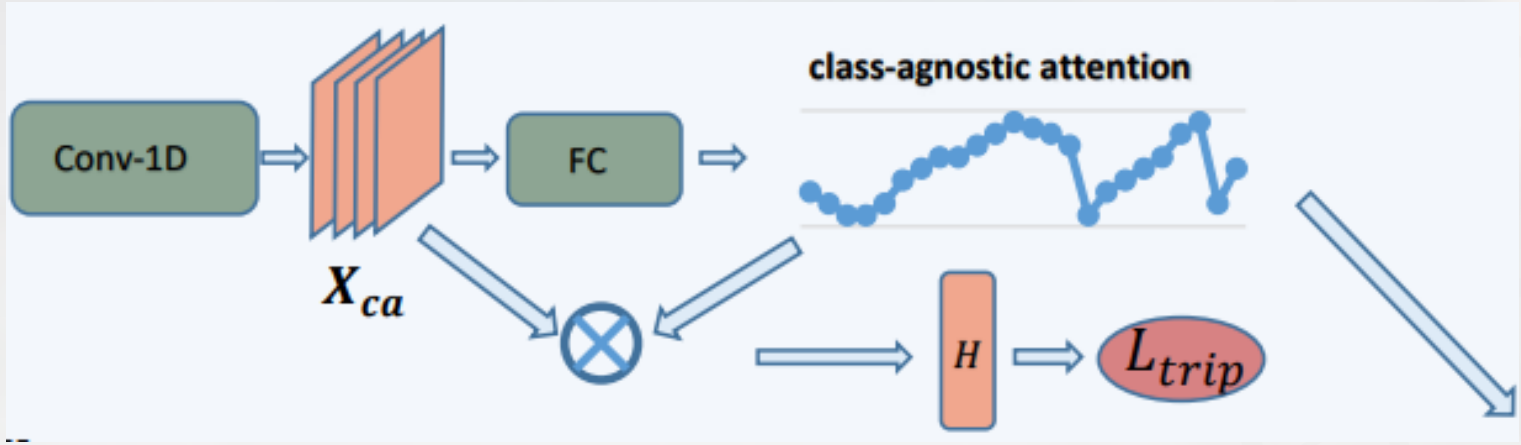
$$\mathcal{L}_{abs} = \sum^{Z} (L_{inter,z} + \theta \cdot L_{intra,z})$$

loss的作用主要是加强同簇中视频的动作相似性和动作背景的分离性

背景介绍
研究方法与思路
实验结果
总结

Class-Agnostic Temporal Attention Module

这个模块的功能是为了学习和动作类别无关的部分即背景部分出现的概率



以X_{ca}为中间层输入，输出类特定分数$S \in R^{T\times 1}$，这个分数S和上面的A有相同的作用，可以理解为出现的概率

背景介绍
研究方法与思路
实验结果
总结

Class-Agnostic Temporal Attention Module

计算cluster-based triplet loss

提取class-agnostic video feature representation H

$$H = X_{ca}{}^{\intercal} S$$

抽取出某一簇内的一个视频v_a，假设v_n是不在群集z中并且与v_a的距离最小的视频，v_p是群集z中的视频并且与v_a的距离最大，有这样的限制：

$$d(H_a, H_n) - d(H_a, H_p) \geq m$$

背景介绍
研究方法与思路
实验结果
总结

Class-Agnostic Temporal Attention Module

计算cluster-based triplet loss

提取class-agnostic video feature representation H

$$H = X_{ca}{}^{\intercal} S$$

抽取出某一簇内的一个视频v_a，假设v_n是不在群集z中并且与v_a的距离最小的视频，v_p是群集z中的视频并且与v_a的距离最大，有这样的限制：

$$d(H_a, H_n) - d(H_a, H_p) \geq m$$

背景介绍
研究方法与思路
实验结果
总结

Class-Agnostic Temporal Attention Module

计算cluster-based triplet loss

$$\mathcal{L}_{trip} = \sum_{z=1}^{Z} \sum_{a=1}^{K} max\{d(H_a, H_p) - d(H_a, H_n) + m, 0\}$$

这个LOSS的意义很明确，为了将同一聚类的视频特征表示拉近，并将不同聚类的视频特征表示在特征空间中推得更远

背景介绍
研究方法与思路
实验结果
总结

Loss

最终loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{trip} + \beta\mathcal{L}_{abs}$$

其中L_{cls}是经典的交叉熵损失

$$\mathcal{L}_{cls} = -\sum_{n=1}^{ZK}\sum_{i=1}^{C} y_{n,i}\log\hat{p}_{n,i}$$

其中$y_{n,i}$为分配的标签，$\hat{p}_{n,i} = softmax(p), p = A\hat{S}$

03 实验结果

中国地质大学
CHINA UNIVERSITY OF GEOSCIENCES

实验结果

背景介绍

研究方法与思路

实验结果

总结

Purity↑	ARI↑	NMI↑
0.645	0.445	0.726
0.740	0.569	0.788
0.780	0.612	0.811

	Methods	mAP@IoU (%)				
		0.3	0.4	0.5	0.6	0.7
FS	SLM-mgram [33]	30.0	23.2	15.2	-	-
	Glimpse [50]	36.0	26.4	17.1	-	-
	PSDF [54]	33.6	26.1	18.8	-	-
	S-CNN [39]	36.3	28.7	19.0	10.3	5.3
	SSAD [17]	43.0	35.0	24.6	-	-
	CDC [37]	40.1	29.4	23.3	13.1	7.9
	R-C3D [48]	44.8	35.6	28.9	-	-
	SSN [57]	51.9	41.0	29.8	-	-
	TAL-Net [4]	53.2	48.5	42.8	33.8	20.8
WS	Hide-and-seek [41]	19.5	12.7	6.8	-	-
	UntrimmedNet [46]	28.2	21.1	13.7	-	-
	STPN [27]	35.5	25.8	16.9	9.9	4.3
	Autoloc [38]	35.8	29.0	21.2	13.4	5.8
	W-TALC [29]	40.1	31.1	22.8	-	7.6
	MAAN [55]	41.1	30.6	20.3	12.0	6.9
	CMCS [20]	41.2	32.1	23.1	15.0	7.0
	3C-Net [26]	44.2	34.1	26.6	-	8.1
	BM [28]	46.6	37.5	26.8	17.6	9.0
	TSM [53]	39.5	-	24.5	-	7.1
	CleanNet [14]	37.0	30.9	23.9	13.9	7.1
	Ours	**46.9**	**38.9**	**30.1**	**19.8**	**10.4**
US	Ours	39.6	32.9	25.0	16.7	8.9

04 总 结

背景介绍
研究方法与思路
实验分析
总结

- 优点

首次使用无监督进行定位，效果不错，并且提供的一个很好的思路

- 缺点

网络结构比较复杂，分帧分段的方式似乎只有局部特征性，全局性比较弱

参考：
https://blog.csdn.net/qq_43310834/article/details/108502214

中国地质大学
CHINA UNIVERSITY OF GEOSCIENCES

谢谢老师和同学的观看