《CROSS-ATTENTIONAL AUDIO-VISUAL FUSION FOR WEAKLY-SUPERVISED ACTION LOCALIZATION》

弱监督动作定位的交叉注意视听融合

其他只依赖视觉特征来进行动作定位，我们提出依靠视听特征进行弱监督动作定位

1. 提出多阶段注意力交叉机制来协助视听融合
2. 提出Open-max分类器将背景作为Open-set,来建模背景与前景帧
3. 提出一致性损失，来增强动作类预测的时间连续性与精准动作定位

Introduction

(Tian et al., 2018; Lin et al., 2019; Xuan et al., 2020)提出视听融合，事件具有强音频提示动作--->我们提出的是一个基于具有弱音频信息，或者缺乏音频信息的事件模型

以一种既能利用互补性质，又能保持特定形态信息的方式融合音频和视频数据是一项关键挑战

提出多阶段注意交叉机制：

在多个阶段逐步学习每个模态的特征，在每个阶段只有通过交叉注意才允许模态间相互作用，在最后阶段视听融合（视觉感知的听觉特征，听觉感知的视觉特征）视频每个片段获得一个audio-visual 特征。

从动作/事件中分离背景是时间定位中的常见问题：

前景可靠性估计与分类，通过open-max分类器与时间连续性损失

对于每个视频片段，open-max预测背景和动作分数（open-max有两并行分支组成：动作分类、前景可靠性）

相关工作

弱监督动作定位：

《Audio-visual event localization in unconstrained videos.》

引入了一个共同活动相似性损失，它在包含一个共同动作类的一对视频中寻找相似的时间区域。

《Category count and center loss for weakly-supervised action localization》

Center loss for global-level,counting loss for local-level

为缓解背景给动作定位带来的影响：

《Weakly-supervised action localiza-tion with background modeling》

提出了自上而下的类引导注意力模型

《Temporal structure mining for weakly supervised action detection.》

利用视频片段的时间关系

《Learning from trimmed videos to localize actions.》

将一段视频分成可解释的片段，称为动作字节，并有效的用于动作proposal

为了区分动作和上下文片段：

《Weakly-supervised action localization by generative attention modeling》

使用条件变分自动编码器设计了以注意力为条件的类无关帧方式概率。

《Weakly-supervised action localization with expectation-maximization multi-instance learning. 》

提出了一个期望最大化的多实例学习框架，其中关键实例被建模为一个隐藏变量

视听事件定位：

将每一个时间步长分类到一个事件类或背景中

《Audio-visual event localization in unconstrained videos. 》

提出了一种具有听觉引导注意力的网络，该网络展示了视听事件定位和跨模态同步事件定位的典型结果。

为了在事件定位中利用全局和局部信息：

《Dual-modality seq2seq network for audio-visual event localization.》

使用多个LSTMs在视频级和片段级进行视听融合

《Dual attention matching for audio-visual event localization.》

假设单个事件视频，使用一个模态中的视频级特征来确定事件类别，并且通过将视频级特征与来自另一个模态的片段级特征相匹配来检测事件相关片段。

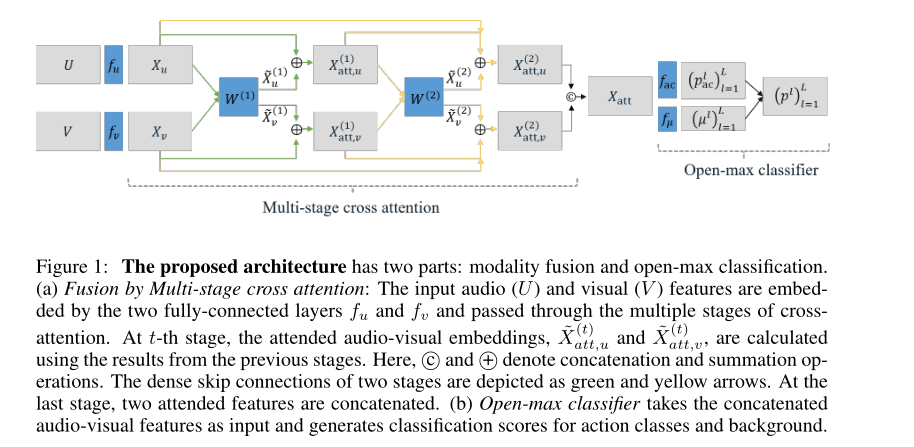
为了解决视听模式之间的时间不一致：

《Cross-modal attention network for temporal inconsistent audio-visual event localization》

设计了模态哨兵，它过滤掉了与事件无关的模态

深度多模态表征学习

方法：



多阶段注意交叉机制：

在另一个模态的约束下为每个模态学习特征这样，学习的特征即表示模态间的信息也保持了模态内信息

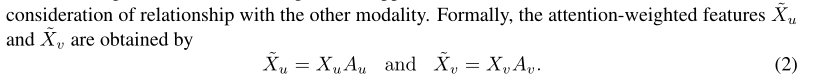
Ux,Vx分别通过FC学习-->l2标准化-->计算cross-correlation. （W可学习参数，）



Λ的l列表示X(v,l)相对于l音频片段的相关性，高相关系数意味着对应的音频和视频片段特征高度相关。

通过softmax Λ，与Λ的转置 得到视频特征权重A(u),音频特征权重A(v)（交叉注意权重）

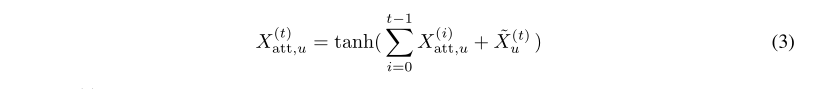
然后，对于每个模态，注意力权重被用于重新加权片段特征，以使它们在考虑与其他模态的关系时更具区别性

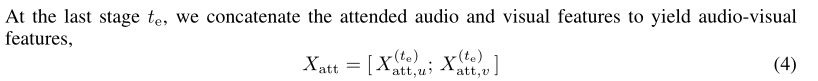


为防止原始模态的特征被过度抑制采用 dense skip-connection （Densely connected

convolutional networks cvpr2017）

使用下面公式获得阶段t的音频特征：





（Λ：相当于w(1)）

Open-max 分类器：

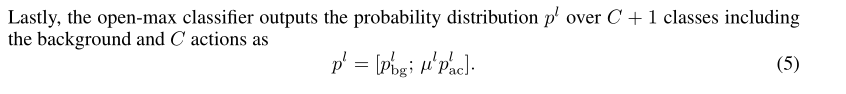
包括两个平行的FC,一个用来动作分类，另一个用来作前景可靠性估计

第一个FC:每个视频片段依次放入FC中

第一层输出h经过soft-max得出分类概率分数

第二次：FC输出经过singmod输出分类是任何动作的概率u(l)，如果概率低则是背景

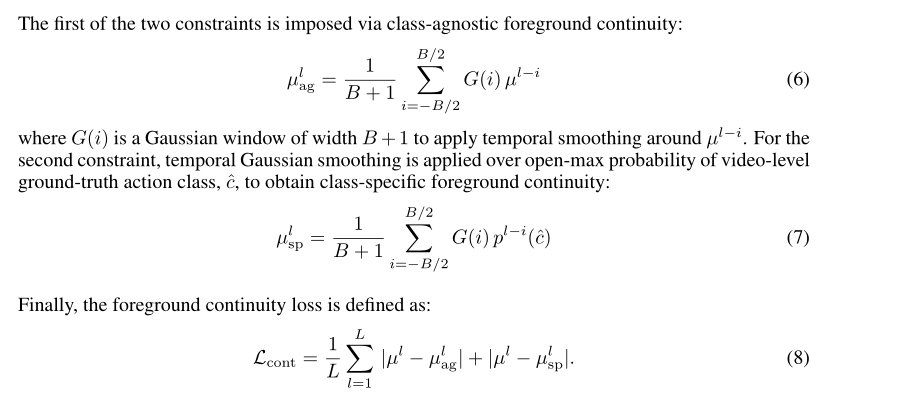
所以背景的概率p(bg,l)=1-u(l)

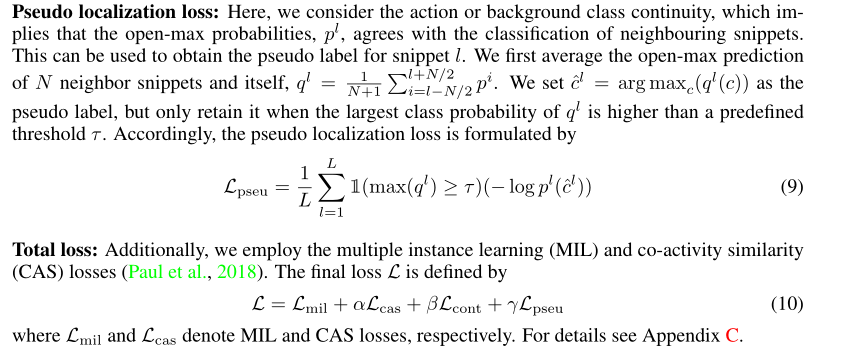


训练损失：

动作或前景不会随时间突然改变。为了施加这种约束，我们设计了两种类型的时间连续性损失：

前景连续损失：前景连续有两个重要属性，与类无关的类似前景可靠性，与目标前景一致的open-max概率





将每个视频一致采样l个片段，然后在放入预训练的网络提取音频，视频特征

使用I3D net和resNet提取视频特征，使用VGG-like提取音频特征