A Dataset for Movie Description

**影片描述的数据集**



计算机视觉阅读笔记 14S103030 徐竟祎

1. 问题背景：

音频描述（Audio Description，简称AD）是一种影片的语言描述，能够让很多视力受损的人理解电影。这样的描述通常是视觉来设计的，从而自然而然的为计算机视觉和计算机语言学提供了数据源；

音频描述（AD）能够让千千万万视觉受损的人看到电影；但现在大部分的音频描述都是由人工完成的，据统计，一部2小时的电影要超过60人力来转录，所以，视力受损的人现在只能看到少部分电影和电视节目。所以我们要提出一种能够自动转换的方法；

一般来说，剧本还没有用做是视频的描述，其主要原因是因为由于电影和剧本存在不符的情况，所以不能用做自动对齐。即使是能完美对齐到电影，剧本同样也不如音频描述（AD）那么准确，因为剧本一样都是在电影拍摄前就已经创造完成了的；

1. 主要研究内容：

在这篇文章中，我们主要是提出了一种新颖的数据集，其中包括转录的音频描述（AD），并在时间上和完整的高清电影对齐。除此之外，我们还收集在以前的工作中用到的对齐的电影剧本，并比较两个不同的源的描述。在总的MPII电影描述的数据集（MPII-MD）包含从94部高清电影中得到的超过68K的句子和片段的视频并行语料库。我们以基准不同的方法来生成视频描述数据集的特征。比较AD的剧本，我们发现，音频描述（AD）能更为直观和准确描述电影显示了什么而不是什么应该按照之前的电影生产创建的剧本发生。

我们提出一个新的数据集，包含已经对齐完整高清电影的转录的音频描述；为此，我们从一个高清蓝光光盘中获取音频流，分割出音频描述的音频，再通过一个转录服务来进行转录；当音频描述没有完全对齐到视频中的活动时，我们就人为的将句子对齐到电影；所以和其他的数据集相比，我们的数据集是能够对准视频中的各项活动，而不是仅仅对齐音频中的描述；

1. MPII电影描述数据集：

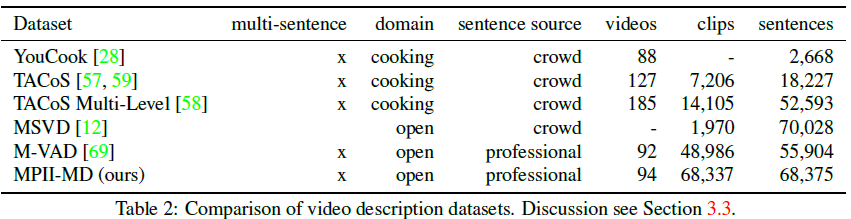
第一个工作就是设定几个电影描述的基准；首先是用最先进的可视化特点，采用邻近检索从训练数据集中获取句子；其次，自动从语义分析中获取语义代表；

这篇文章最大的贡献是我们的MPH 电影表述数据集（MPII-MD），其中提供了转录对齐的音频描述和剧本数据；下面我们来介绍我们的获取音频描述和剧本，并且对齐到电影得到半自动化方法；

1. 获取音频描述（AD）：我们在英国的亚马逊上在音频描述这一块中搜索蓝光电影，并且选出了55部不同的电影类型；由于ADs只能是音频格式，所以先从蓝光影碟中获取它的音频流，然后再用我们接下来要介绍的半自动化分割AD音频的方法（AD音频中包含原始的音频流，要分离出来）；分离出来的音频片段通过转录服务，会得到一个对每一句对话都有对应的时间戳。但可能在对话和视频会有微小的不一致，所以我们要人为的进行对齐；

半自动化分割Ads：我们首先要估计Ads和原始音频（这也是AD的部分）之间的时间对准差异，因为他们可能会缺失掉几个时间帧；这种精确的对准是非常重要的，依次来计算两个流之间的相似性。相似性和对齐性这两个步骤都是用快速傅里叶变化来计算他们的光谱图；如果两者之间的音频流大于某个假设的音频阈值，则我们假定AD就包含在该时间点上的音频描述；阈值的确定是通过不同的电影样本，但不同的电影有不同的解说员和混合音频流，所以阈值也不同；

1. 获取剧本数据：我们选择了50部电影，首先分析剧本，处理一些默认格式的偏离的剧本；其次，从蓝光影碟中提取字母，再用动态规划的方法将剧本对齐到字母，并且推断出描述性句子的时间戳；我们选取的句子都是可靠对齐得分在0.5以上的；
2. 比较其他数据集：



有三点不同：A. 我们的语料库包含剧本和音频描述，然而其他的可能只包含音频描述；B. 我们会手动将每句话对齐到视频中的各项活动，其他是自动化的方法，所以其精确度下降；C. 我们用蓝光电影，其他用的是DVD；

1. 可视化特征：

DT：我们从相机的运动中获取改进密集轨迹；对于每个特征（轨迹，HOG, HOF，MBH），我们都创建了一个有4000集群的密码本，并计算相应的柱状图；再采用L1归一化方法获取柱状图并用做特征；

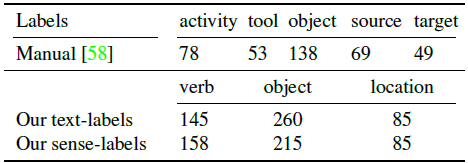
LSDA：我们用大型物体监测CNN用来区分7604中ImageNet类；我们用检测器在每秒提取帧上进行监测；在每一帧我们maxpool所有类的网络响应，然后做meanpooling在帧的视频片断中，并使用结果作为特征。

Places and Hybrid：我们将分类器用在我们数据集中获取的帧，即是maxpool每一个视频片段的帧，将结果作为特征；

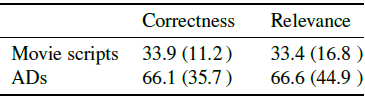
1. 视频描述方法
2. 最近邻居：从训练语料库中，采用上述说的可视化特征获取最近的句子；
3. SMT：将视频输入映射到语义代表（SR），之后将语义代表（SR）又映射到自然语句；
4. SMT 可视化词汇：我们将可视化的分类和词汇直接转化成一个句子；对于物体和场景来说，我们采用LSDA和Places；对于活动，采用最新的活动识别特征DT；我们用k-means方法将DT的柱状图聚集到300个词汇；每一类的索引即是DTi的中心，并被作为每一类的标签；所以，我们获得了元组：
5. 语义分析：从自然语言中提取中间的语义代表（SR）

语义分析方法：对于名词的识别，我们消除前面的词汇；对于动词，则将其前面的词汇和WordNet中联系起来进行识别；我们采用两层匹配来获得标签的角色；首先是语法匹配；每个在VerbNet中的动词都有固定的语法框架；所以我们首先将句子中的动词和VerbNet中的进行匹配，得到一些候选词汇进入下一步；之后再进行语义分析，VerbNet同样也给这些词提供了角色的限制，所以根据上一步的候选词汇进行角色的判定；

元组的表示方法是<Verb, Object, Location>,作为我们的语义代表（SR），之后用CRF方法，三个节点分别表示Verb，Object和Location；最后用CRF产生的SR恢复成自然语言；通过下图的对比发现，我们的自动分析方法可以取代之前的人工对齐方法；



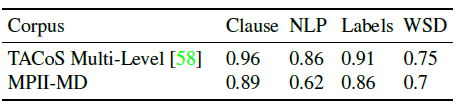
1. 评价：
2. AD和剧本数据的对比



可以看出来，AD在2/3的情况下都更正确

1. 语义分析评价

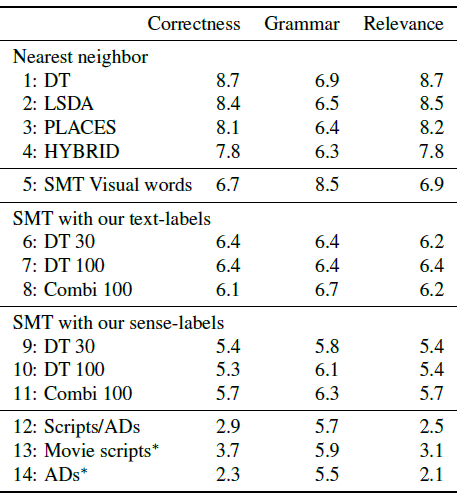
下表中呈现了不同的组件的语义分析情况，其中有从句分割（Clause），词性标注，分块（NLP），语义角色标注（Labels）和词义消岐（WSD）；总体来说，我们可以看出来MPII-MD数据集比TACoS Multi-Level表现更好



1. 视频描述

根据下表，我们可以得出以下结论

1. 我们采用SMT的分析方法比近邻基线和SMT可视化文字要好；
2. 实际电影的剧本/AD的平均等级要低，也就是说比其他的自动化方法都要好；
3. 当值计算AD的等级时，准确性和相关性都要比电影剧本低；



1. 结论：我们提出了一种剧本和AD对齐的音频描述数据集；我们首先和电影描述方法结合起来，将这个数据集利用在最近的视觉特征上；然后调整这种方法用在不住事的数据集上，只是依靠标签进行语义分析；结果表明，我们的结果表现非常好；我们的音频分析AD比之前所用的剧本数据更和电影准确相关；我们的结果为以后再更大规模的情况下创造了可能；