



(12) 发明专利申请

(10) 申请公布号 CN 116668615 A

(43) 申请公布日 2023. 08. 29

(21) 申请号 202310647053.1

(22) 申请日 2023.06.02

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市大兴区北京经济技术
开发区科创十一街18号院2号楼6层
601

(72) 发明人 袁韶祖 祝天刚 陈蒙

(74) 专利代理机构 中原信达知识产权代理有限
责任公司 11219

专利代理师 王安娜 李阳

(51) Int.Cl.

H04N 5/278 (2006.01)

H04N 21/488 (2011.01)

G06N 3/0499 (2023.01)

G06N 3/09 (2023.01)

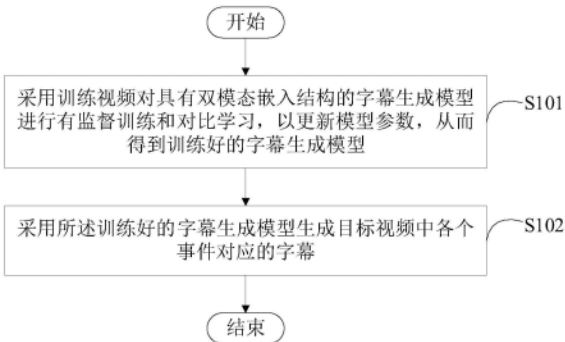
权利要求书2页 说明书12页 附图4页

(54) 发明名称

一种字幕生成方法和装置

(57) 摘要

本发明公开了一种字幕生成方法和装置,涉及人工智能技术领域。该方法的一具体实施方式包括:采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,从而得到训练好的字幕生成模型;采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。该实施方式能够解决不同模态特征没有很好地对齐以及对比学习不能直接应用于密集视频字幕的技术问题。



1. 一种字幕生成方法,其特征在于,包括:

采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型;

采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。

2. 根据权利要求1所述的方法,其特征在于,所述字幕生成模型包括依次串联的事件定位器、特征提取层、单流编码器、双流编码器和双模态解码器。

3. 根据权利要求2所述的方法,其特征在于,采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型,包括:

将训练视频输入到事件定位器,从而输出所述训练视频中的各个事件;

将所述各个事件输入到特征提取层,从而输出视觉流特征和光流特征;

将所述视觉流特征和所述光流特征输入到单流编码器,从而输出视觉流单编码特征和光流单编码特征;

将所述视觉流单编码特征和所述光流单编码特征输入到双流编码器,从而输出视觉流融合编码特征和光流融合编码特征;

对所述视觉流单编码特征和所述光流单编码特征进行对比学习;以及,采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,以更新模型参数,从而得到训练好的字幕生成模型。

4. 根据权利要求3所述的方法,其特征在于,所述字幕生成模型还包括注意力池化层;

对所述视觉流单编码特征和所述光流单编码特征进行对比学习,包括:

将所述视觉流单编码特征和视觉流检索编码特征输入到注意力池化层,从而输出视觉流池化特征;

将所述光流单编码特征和光流检索编码特征输入到注意力池化层,从而输出光流池化特征;

对所述视觉流池化特征和所述光流池化特征进行对比学习。

5. 根据权利要求4所述的方法,其特征在于,对所述视觉流池化特征和所述光流池化特征进行对比学习,包括:

从特征队列中获取历史视频的视觉流池化特征和光流池化特征;

计算所述训练视频的视觉流池化特征到所述历史视频的光流池化特征和所述训练视频的光流池化特征的交叉熵函数,从而得到第一交叉熵函数;

计算所述训练视频的光流池化特征到所述历史视频的视觉流池化特征和所述训练视频的视觉流池化特征的交叉熵函数,从而得到第二交叉熵函数;

将所述第一交叉熵函数和所述第二交叉熵函数相加,得到对比学习函数,从而进行对比学习。

6. 根据权利要求3所述的方法,其特征在于,所述字幕生成模型还包括字幕生成器;

采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,包括:

将所述视觉流融合编码特征和所述光流融合编码特征输入到双模态解码器中,从而输出事件解码特征;

将所述事件解码特征输入到字幕生成器,从而输出所述训练视频的字幕,同时以所述训练视频对应的训练字幕为目标,进行有监督训练。

7.根据权利要求2所述的方法,其特征在于,所述单流编码器包括串联的第一多头自注意力网络 and 第一前馈神经网络;所述双流编码器包括依次串联的第二多头自注意力网络、第一多头交叉注意力网络 and 第二前馈神经网络;所述双模态解码器包括依次串联的第三多头自注意力网络、第二多头交叉注意力网络、第三前馈神经网络 and 第四前馈神经网络。

8.根据权利要求4所述的方法,其特征在于,还包括:

在更新模型参数的过程中,所述注意力池化层的模型参数基于动量系数进行更新,所述动量系数小于0.5。

9.一种字幕生成装置,其特征在于,包括:

训练模块,用于采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练 and 对比学习,以更新模型参数,从而得到训练好的字幕生成模型;

生成模块,用于采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。

10.一种电子设备,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行时,所述一个或多个处理器实现如权利要求1-8中任一所述的方法。

11.一种计算机可读介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行时实现如权利要求1-8中任一所述的方法。

12.一种计算机程序产品,包括计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1-8中任一项所述的方法。

一种字幕生成方法和装置

技术领域

[0001] 本发明涉及人工智能技术领域,尤其涉及一种字幕生成方法和装置。

背景技术

[0002] 视频描述是为一个视频生成相应的视频描述。现实场景中视频通常分为不同的事件,每个事件所表达的信息都是不同的,因此只通过一句话并不能很好的刻画视频内容。密集视频字幕(dense video caption也叫DVC)是一项更具有挑战性的任务,它首先从一个视频中定位多个事件,然后分别为每个事件生成相应的字幕。

[0003] 在实现本发明过程中,发明人发现现有技术中至少存在如下问题:

[0004] 1)现有的多模态融合方法仅仅关注到使用不同模态特征的融合,但是却没有很好地对齐这些特征;2)对比学习不能直接应用于密集视频字幕:一方面,因为密集视频字幕存在多个事件,如何标注并区分这些事件是一个问题,如果标记错误可能导致错误对比的问题;另一方面,对比学习通常只在一个batch(批量大小)内进行,因此模型一次学习到的样本有限,不能全局感知到所有数据分布,这限制了模型的表达能力。

发明内容

[0005] 有鉴于此,本发明实施例提供一种字幕生成方法和装置,以解决不同模态特征没有很好地对齐以及对比学习不能直接应用于密集视频字幕的技术问题。

[0006] 为实现上述目的,根据本发明实施例的一个方面,提供了一种字幕生成方法,包括:

[0007] 采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型;

[0008] 采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。

[0009] 可选地,所述字幕生成模型包括依次串联的事件定位器、特征提取层、单流编码器、双流编码器和双模态解码器。

[0010] 可选地,采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型,包括:

[0011] 将训练视频输入到事件定位器,从而输出所述训练视频中的各个事件;

[0012] 将所述各个事件输入到特征提取层,从而输出视觉流特征和光流特征;

[0013] 将所述视觉流特征和所述光流特征输入到单流编码器,从而输出视觉流单编码特征和光流单编码特征;

[0014] 将所述视觉流单编码特征和所述光流单编码特征输入到双流编码器,从而输出视觉流融合编码特征和光流融合编码特征;

[0015] 对所述视觉流单编码特征和所述光流单编码特征进行对比学习;以及,采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,以更新模型参数,从而得到训练好的字幕生成模型。

- [0016] 可选地,所述字幕生成模型还包括注意力池化层;
- [0017] 对所述视觉流单编码特征和所述光流单编码特征进行对比学习,包括:
- [0018] 将所述视觉流单编码特征和视觉流检索编码特征输入到注意力池化层,从而输出视觉流池化特征;
- [0019] 将所述光流单编码特征和光流检索编码特征输入到注意力池化层,从而输出光流池化特征;
- [0020] 对所述视觉流池化特征和所述光流池化特征进行对比学习。
- [0021] 可选地,对所述视觉流池化特征和所述光流池化特征进行对比学习,包括:
- [0022] 从特征队列中获取历史视频的视觉流池化特征和光流池化特征;
- [0023] 计算所述训练视频的视觉流池化特征到所述历史视频的光流池化特征和所述训练视频的光流池化特征的交叉熵函数,从而得到第一交叉熵函数;
- [0024] 计算所述训练视频的光流池化特征到所述历史视频的视觉流池化特征和所述训练视频的视觉流池化特征的交叉熵函数,从而得到第二交叉熵函数;
- [0025] 将所述第一交叉熵函数和所述第二交叉熵函数相加,得到对比学习函数,从而进行对比学习。
- [0026] 可选地,所述字幕生成模型还包括字幕生成器;
- [0027] 采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,包括:
- [0028] 将所述视觉流融合编码特征和所述光流融合编码特征输入到双模态解码器中,从而输出事件解码特征;
- [0029] 将所述事件解码特征输入到字幕生成器,从而输出所述训练视频的字幕,同时以所述训练视频对应的训练字幕为目标,进行有监督训练。
- [0030] 可选地,所述单流编码器包括串联的第一多头自注意力网络和第一前馈神经网络;所述双流编码器包括依次串联的第二多头自注意力网络、第一多头交叉注意力网络和第二前馈神经网络;所述双模态解码器包括依次串联的第三多头自注意力网络、第二多头交叉注意力网络、第三前馈神经网络和第四前馈神经网络。
- [0031] 可选地,还包括:
- [0032] 在更新模型参数的过程中,所述注意力池化层的模型参数基于动量系数进行更新,所述动量系数小于0.5。
- [0033] 另外,根据本发明实施例的另一个方面,提供了一种字幕生成装置,包括:
- [0034] 训练模块,用于采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型;
- [0035] 生成模块,用于采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。
- [0036] 可选地,所述字幕生成模型包括依次串联的事件定位器、特征提取层、单流编码器、双流编码器和双模态解码器。
- [0037] 可选地,所述训练模块还用于:
- [0038] 将训练视频输入到事件定位器,从而输出所述训练视频中的各个事件;
- [0039] 将所述各个事件输入到特征提取层,从而输出视觉流特征和光流特征;

- [0040] 将所述视觉流特征和所述光流特征输入到单流编码器,从而输出视觉流单编码特征和光流单编码特征;
- [0041] 将所述视觉流单编码特征和所述光流单编码特征输入到双流编码器,从而输出视觉流融合编码特征和光流融合编码特征;
- [0042] 对所述视觉流单编码特征和所述光流单编码特征进行对比学习;以及,采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,以更新模型参数,从而得到训练好的字幕生成模型。
- [0043] 可选地,所述字幕生成模型还包括注意力池化层;
- [0044] 所述训练模块还用于:
- [0045] 将所述视觉流单编码特征和视觉流检索编码特征输入到注意力池化层,从而输出视觉流池化特征;
- [0046] 将所述光流单编码特征和光流检索编码特征输入到注意力池化层,从而输出光流池化特征;
- [0047] 对所述视觉流池化特征和所述光流池化特征进行对比学习。
- [0048] 可选地,所述训练模块还用于:
- [0049] 从特征队列中获取历史视频的视觉流池化特征和光流池化特征;
- [0050] 计算所述训练视频的视觉流池化特征到所述历史视频的光流池化特征和所述训练视频的光流池化特征的交叉熵函数,从而得到第一交叉熵函数;
- [0051] 计算所述训练视频的光流池化特征到所述历史视频的视觉流池化特征和所述训练视频的视觉流池化特征的交叉熵函数,从而得到第二交叉熵函数;
- [0052] 将所述第一交叉熵函数和所述第二交叉熵函数相加,得到对比学习函数,从而进行对比学习。
- [0053] 可选地,所述字幕生成模型还包括字幕生成器;
- [0054] 所述训练模块还用于:
- [0055] 将所述视觉流融合编码特征和所述光流融合编码特征输入到双模态解码器中,从而输出事件解码特征;
- [0056] 将所述事件解码特征输入到字幕生成器,从而输出所述训练视频的字幕,同时以所述训练视频对应的训练字幕为目标,进行有监督训练。
- [0057] 可选地,所述单流编码器包括串联的第一多头自注意力网络 and 第一前馈神经网络;所述双流编码器包括依次串联的第二多头自注意力网络、第一多头交叉注意力网络 and 第二前馈神经网络;所述双模态解码器包括依次串联的第三多头自注意力网络、第二多头交叉注意力网络、第三前馈神经网络 and 第四前馈神经网络。
- [0058] 可选地,所述训练模块还用于:
- [0059] 在更新模型参数的过程中,所述注意力池化层的模型参数基于动量系数进行更新,所述动量系数小于0.5。
- [0060] 根据本发明实施例的另一个方面,还提供了一种电子设备,包括:
- [0061] 一个或多个处理器;
- [0062] 存储装置,用于存储一个或多个程序,
- [0063] 当所述一个或多个程序被所述一个或多个处理器执行时,所述一个或多个处理器

实现上述任一实施例所述的方法。

[0064] 根据本发明实施例的另一个方面,还提供了一种计算机可读介质,其上存储有计算机程序,所述程序被处理器执行时实现上述任一实施例所述的方法。

[0065] 根据本发明实施例的另一个方面,还提供了一种计算机程序产品,包括计算机程序,所述计算机程序被处理器执行时实现上述任一实施例所述的方法。

[0066] 上述发明中的一个实施例具有如下优点或有益效果:因为通过训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型的技术手段,所以克服了现有技术中不同模态特征没有很好地对齐以及对比学习不能直接应用于密集视频字幕的技术问题。本发明实施例将对比学习应用于密集视频字幕,并采用事件对齐来捕获更多的负样本,从而增强对比学习的效果,由此训练得到的字幕生成模型可以准确地生成密集视频字幕。

[0067] 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具体实施方式加以说明。

附图说明

[0068] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。其中:

[0069] 图1是根据本发明实施例的字幕生成方法的流程图;

[0070] 图2是根据本发明实施例的字幕生成模型的结构示意图;

[0071] 图3是根据本发明一个可参考实施例的字幕生成方法的流程图;

[0072] 图4是根据本发明一个可参考实施例的字幕生成模型的结构示意图;

[0073] 图5是根据本发明实施例的字幕生成装置的示意图;

[0074] 图6是本发明实施例可以应用于其中的示例性系统架构图;

[0075] 图7是适于用来实现本发明实施例的终端设备或服务器的计算机系统的结构示意图。

具体实施方式

[0076] 以下结合附图对本发明的示范性实施例做出说明,其中包括本发明实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本发明的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0077] 需要说明的是,本公开的技术方案中,所涉及的用户个人信息的采集、收集、更新、分析、处理、使用、传输、存储等方面,均符合相关法律法规的规定,被用于合法的用途,且不违背公序良俗。对用户个人信息采取必要措施,防止对用户个人信息数据的非法访问,维护用户个人信息安全、网络安全和国家安全。

[0078] 图1是根据本发明实施例的字幕生成方法的流程图。作为本发明的一个实施例,如图1所示,所述字幕生成方法可以包括:

[0079] 步骤101,采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型。

[0080] 在本发明的实施例中,字幕生成模型具有双模态嵌入结构,以适应不同的模态的对比学习。可选地,如图2所示,所述字幕生成模型包括依次串联的事件定位器(BMT)、特征提取层、单流编码器、双流编码器和双模态解码器。通过事件定位器可以定位出视频中的每个事件,对各个事件进行特征提取,得到每个事件的视觉流特征(RGB)和光流特征(flow,采用光流法提取),因此本发明实施例采用事件级别的对比学习拉近不同模态之间的语义空间。

[0081] 光流(optical flow)是空间运动物体在观察成像平面上的像素运动的瞬时速度。光流法是利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧跟当前帧之间存在的对应关系,从而计算出相邻帧之间物体的运动信息的一种方法。

[0082] 可选地,步骤101可以包括:将训练视频输入到事件定位器,从而输出所述训练视频中的各个事件;将所述各个事件输入到特征提取层,从而输出视觉流特征和光流特征;将所述视觉流特征和所述光流特征输入到单流编码器,从而输出视觉流单编码特征和光流单编码特征;将所述视觉流单编码特征和所述光流单编码特征输入到双流编码器,从而输出视觉流融合编码特征和光流融合编码特征;对所述视觉流单编码特征和所述光流单编码特征进行对比学习;以及,采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,以更新模型参数,从而得到训练好的字幕生成模型。如图2所示,将训练视频输入到事件定位器,事件定位器对视频中的事件进行定位,从而输出视频中的各个事件 $P = \{(start_j, end_j, score_j)\}_{j=1}^K$,接着将各个事件输入到特征提取层,特征提取层提取出每个事件的视觉流特征 $Z^V \in R^{L^j \times d^v}$ 和光流特征 $Z^F \in R^{L^j \times d^f}$ 。单流编码器利用自注意力网络表示学习能力,对事件对齐的视觉流特征和光流特征进行编码,得到视觉流单编码特征和光流单编码特征;双流编码器与单流编码器相似,双流编码器比单流编码器多了多头交叉注意力网络,双流编码器能够让不同的模态特征之间进行初步的交互,方便以后的语义学习,双流编码器输出视觉流融合编码特征和光流融合编码特征;双模态解码器采用了自注意力网络的Transformer架构,还增加了一个两层前馈神经网络来集成双模态输出解码结果,通过字幕生成器为每个事件生成相应的字幕,并以所述训练视频对应的真实字幕为目标,进行有监督训练;同时对视觉流单编码特征和光流单编码特征进行对比学习,以更新模型参数,从而得到训练好的字幕生成模型。

[0083] 对比学习是自监督学习的一种,也就是说,不依赖标注数据,要从无标注图像中自己学习知识。对比学习的指导原则是:通过自动构造相似实例和不相似实例,要求习得一个表示学习模型,通过这个模型,使得相似的实例在投影空间中比较接近,而不相似的实例在投影空间中距离比较远。

[0084] 为了避免同一视频中的不同事件表达的内容不同,本发明实施例只将同一事件作为对比学习的正样本,除此之外的其他事件均视为负样本。但是,同一视频的内容仍然有可能和当前事件的内容相似,从而导致错误的对比目标。为了解决这个问题,本发明实施例标记来自于同一视频的其他事件为负样本,在计算对比损失时,对那些与当前事件相同的负样本进行mask(屏蔽),避免影响到网络更新。

[0085] 步骤102,采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。

[0086] 训练好字幕生成模型后,将目标视频输入到训练好的字幕生成模型中,从而输出该目标视频中各个事件对应的字幕,即生成密集视频字幕。

[0087] 根据上面所述的各种实施例,可以看出本发明实施例通过采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型的技术手段,解决了现有技术中不同模态特征没有很好地对齐以及对比学习不能直接应用于密集视频字幕的技术问题。本发明实施例将对比学习应用于密集视频字幕,并采用事件对齐来捕获更多的负样本,从而增强对比学习的效果,由此训练得到的字幕生成模型可以准确地生成密集视频字幕。

[0088] 图3是根据本发明一个可参考实施例的字幕生成方法的流程图。作为本发明的又一个实施例,如图3所示,所述字幕生成方法可以包括:

[0089] 步骤301,将训练视频输入到事件定位器,从而输出所述训练视频中的各个事件。

[0090] 将训练视频输入到事件定位器,事件定位器对视频中的事件进行定位,从而输出视频中的各个事件 $P = \{(start_j, end_j, score_j)\}_{j=1}^K$ 。

[0091] 步骤302,将所述各个事件输入到特征提取层,从而输出视觉流特征和光流特征。

[0092] 将各个事件输入到特征提取层,特征提取层提取出每个事件的视觉流特征 $Z^V \in R^{L^j \times d^v}$ 和光流特征 $Z^F \in R^{L^j \times d^f}$ 。

[0093] 步骤303,将所述视觉流特征和所述光流特征输入到单流编码器,从而输出视觉流单编码特征和光流单编码特征。

[0094] 如图4所示,所述单流编码器包含N个编码层,每个编码层包括串联的第一多头自注意力网络和第一前馈神经网络(FFN,图中未示出)。FFN=Feedward+Relu,单流编码器利用多头自注意力网络表示学习能力,对事件对齐的视觉流特征和光流特征进行编码,得到视觉流单编码特征和光流单编码特征:

$$[0095] \quad \begin{bmatrix} v^V = FFN(SelfAttention(Z^V, Z^V)) \\ v^F = FFN(SelfAttention(Z^F, Z^F)) \end{bmatrix}^N$$

[0096] 其中, v^V 表示视觉流单编码特征, v^F 表示光流单编码特征。

[0097] 步骤304,将所述视觉流单编码特征和所述光流单编码特征输入到双流编码器,从而输出视觉流融合编码特征和光流融合编码特征。

[0098] 双流编码器在结构上与单流编码器相似,除了在每个编码层增加额外的多头交叉注意来融合每个模态的输出。如图4所示,所述双流编码器包括依次串联的第二多头自注意力网络、第一多头交叉注意力网络和第二前馈神经网络(FFN,图中未示出)。双流编码器的定义如下:

$$[0099] \quad \begin{bmatrix} \chi^V = SelfAttention(v^V, v^V) \\ \chi^F = SelfAttention(v^F, v^F) \\ \chi^{VF} = FFN(CrossAttention(\chi^V, v^F)) \\ \chi^{FV} = FFN(CrossAttention(\chi^F, v^V)) \end{bmatrix}^N$$

[0100] 其中, x^V 表示第二多头自注意力网络输出的视觉流单编码特征, x^F 表示第二多头自注意力网络输出的光流单编码特征; $\chi^{VF} \in R^{L^j \times d^v}$ 表示光流参与的视觉流融合编码特征,

$\chi^{FV} \in R^{L^j \times d^f}$ 表示视觉流参与的光流融合编码特征。

[0101] 多头交叉注意力网络能够让不同的模态特征之间进行初步的交互,方便以后的语义学习。

[0102] 步骤305,对所述视觉流单编码特征和所述光流单编码特征进行对比学习;以及,采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,以更新模型参数,从而得到训练好的字幕生成模型。

[0103] 可选地,所述字幕生成模型还包括注意力池化层;对所述视觉流单编码特征和所述光流单编码特征进行对比学习,包括:将所述视觉流单编码特征和视觉流检索编码特征输入到注意力池化层,从而输出视觉流池化特征;将所述光流单编码特征和光流检索编码特征输入到注意力池化层,从而输出光流池化特征;对所述视觉流池化特征和所述光流池化特征进行对比学习。需要说明的是,视觉流检索编码特征和光流检索编码特征是字幕生成模型中的参数,代表关键的特征,字幕生成模型在训练过程中,学习得到视觉流检索编码特征和光流检索编码特征。注意力池化层对特征进行降维,输出一维的视觉流池化特征和光流池化特征。

[0104] 可选地,对所述视觉流池化特征和所述光流池化特征进行对比学习,包括:从特征队列中获取历史视频的视觉流池化特征和光流池化特征;计算所述训练视频的视觉流池化特征到所述历史视频的光流池化特征和所述训练视频的光流池化特征的交叉熵函数,从而得到第一交叉熵函数;计算所述训练视频的光流池化特征到所述历史视频的视觉流池化特征和所述训练视频的视觉流池化特征的交叉熵函数,从而得到第二交叉熵函数;将所述第一交叉熵函数和所述第二交叉熵函数相加,得到对比学习函数,从而进行对比学习。

[0105] 可选地,所述字幕生成模型还包括字幕生成器;采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,包括:将所述视觉流融合编码特征和所述光流融合编码特征输入到双模态解码器中,从而输出事件解码特征;将所述事件解码特征输入到字幕生成器,从而输出所述训练视频的字幕,同时以所述训练视频对应的训练字幕为目标,进行有监督训练。

[0106] 可选地,如图4所示,所述双模态解码器包括依次串联的第三多头自注意力网络、第二多头交叉注意力网络、第三前馈神经网络 (FFN) 和第四前馈神经网络 (FFN)。双模态解码器采用了基于N层自注意力网络的Transformer架构,还增加了一个两层前馈神经网络来集成双模态输出解码结果。

[0107] 假设时间步长t的输入为 $W_{t-1} = (w_0, w_1, \dots, w_{t-1})$, $w_i \in R^{d \times 1}$, w_i 表示字幕,生成的字幕输入到嵌入层和第三多头自注意力网络,输出的字幕特征 $\hat{K}_t$ 分别与视觉流融合编码特征 x^{VF} 、光流融合编码特征 x^{FV} 进行融合,得到字幕视觉流融合编码特征 κ_t^{VF} 和字幕光流融合编码特征 κ_t^{FV} ,最终采用第三前馈神经网络和第四前馈神经网络进行学习,输出事件解码特征 κ_t 。

[0108] 双模态解码器定义为:

$$[0109] \quad \begin{bmatrix} \hat{\kappa}_t = SelfAttention(w_{t-1}, W_{t-1}) \\ \kappa_t^{VF} = CrossAttention(\hat{\kappa}_t, \chi^{VF}) \\ \kappa_t^{FV} = CrossAttention(\hat{\kappa}_t, \chi^{FV}) \\ \kappa_t = FFN(FFN([\kappa_t^{VF}, \kappa_t^{FV}])) \end{bmatrix}^N$$

[0110] 其中, κ_t 表示时间步长 t 的输出。

[0111] 字幕生成器包括前馈网络 (Feedward) 和归一化网络 (Softmax), 字幕生成器生成当前词 $p(w_t | W_{t-1})$ 的概率, 这样依次就可以为每个事件生成相应的字幕。

[0112] 可以看出, 训练字幕生成模型的目标包括字幕目标 and 对比度目标, 其中字幕目标采用有监督学习方式进行学习, 对比度目标采用对比学习方式进行学习:

[0113] 1) 字幕目标

[0114] 对于字幕生成, 本发明实施例采用直接优化KL散度作为损失函数, 定义为:

$$[0115] \quad \mathcal{L}_{cap} = \sum_{i=1}^T KL(\log(p(w_t | W_{t-1})), y_t^*)$$

[0116] 其中, $y_t^* \in |\Sigma|$ 表示当前的真实字幕。

[0117] 2) 对比度目标

[0118] 计算所述训练视频的视觉流池化特征到所述历史视频的光流池化特征和所述训练视频的光流池化特征的交叉熵函数, 从而得到第一交叉熵函数:

$$[0119] \quad C(v \rightarrow f) = \frac{\exp((\hat{q}^v \cdot \hat{q}_m^f)/\tau)}{\sum_{m=1}^M \exp((\hat{q}^v \cdot \hat{q}_m^f)/\tau)}$$

[0120] 计算所述训练视频的光流池化特征到所述历史视频的视觉流池化特征和所述训练视频的视觉流池化特征的交叉熵函数, 从而得到第二交叉熵函数:

$$[0121] \quad C(f \rightarrow v) = \frac{\exp((\hat{q}^f \cdot \hat{q}_m^v)/\tau)}{\sum_{m=1}^M \exp((\hat{q}^f \cdot \hat{q}_m^v)/\tau)}$$

[0122] 其中, $\hat{q}^v$ 表示训练视频的视觉流池化特征, $\hat{q}^f$ 表示训练视频的光流池化特征, $\hat{q}_m^f$ 表示历史视频的光流池化特征, $\hat{q}_m^v$ 表示历史视频的视觉流池化特征, C 表示交叉熵函数, M 表示一个batch的大小, 也就是队列中的历史视频的各个事件和当前的训练视频的各个事件。

[0123] 将所述第一交叉熵函数和所述第二交叉熵函数相加, 得到对比学习函数:

$$[0124] \quad \mathcal{L}_{con} = C(v \rightarrow f) + C(f \rightarrow v)$$

[0125] 其中, $\mathcal{L}_{con}$ 表示对比学习函数。

[0126] 本发明实施例通过鼓励匹配的事件对获得更高的相似性分数来促进模态间的融合, 从而得到更好地进行特征表示。

[0127] 本发明实施例提出了两个改进来适应密集的视频字幕任务:

[0128] (1) 为了避免同一视频中的不同事件表达的内容不同, 本发明实施例只将同一事件作为对比学习的正样本, 除此之外的其他事件均视为负样本。但是, 同一视频的内容仍然有可能和当前事件的内容相似, 从而导致错误的对比目标。为了解决这个问题, 本发明实施

例标记来自于同一视频的其他事件为负样本,在计算对比损失时,对那些与当前事件相同的负样本进行mask(屏蔽),避免影响到网络更新。

[0129] (2)为了扩大负采样的空间,本发明实施例维护一个特征队列,队列中里面存储了大量负样本用于对比学习。虽然队列扩大了负采样空间,但在训练迭代过程中,队列中的那些单模态的视觉流池化特征、光流池化特征随着梯度更新,也必然会被修改。这种频繁更新会导致训练不稳定,使模型难以收敛。为了解决这一问题,本发明实施例采用慢更新方式对模型参数进行优化,以稳定地产生视觉流池化特征、光流池化特征。

[0130] $\theta \leftarrow \beta \theta_m + (1-\beta) \theta_u$

[0131] 其中, θ_m 表示注意力池化层的参数, θ_u 表示整个模型的参数, $\beta \in [0,1)$ 为动量系数, θ 表示整个模型的参数。

[0132] 因此,注意力池化层的参数有两份,一份为 θ_m ,用于对比学习,另一份为 θ_u ,用于反向传播。

[0133] 可选地,在更新模型参数的过程中,所述注意力池化层的模型参数基于动量系数进行更新,所述动量系数小于0.5,比如0.2、0.1、0.05或者0.01等。

[0134] 因此,本发明实施例设计了一个慢更新的特征队列来存储来自其他视频的负样本,从而扩大了负采样空间,以促进对比学习,增强对比学习的效果。

[0135] 需要说明的是,对该训练视频完成训练后,可以将该训练视频的视觉流池化特征和光流池化特征放入特征队列,作为下一个训练视频的历史视频。

[0136] 步骤306,采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。

[0137] 另外,在本发明一个可参考实施例中字幕生成方法的具体实施内容,在上面所述字幕生成方法中已经详细说明了,故在此重复内容不再说明。

[0138] 图5是根据本发明实施例的字幕生成装置的示意图。如图5所示,所述字幕生成装置500包括训练模块501和生成模块502;其中,训练模块501用于采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型;生成模块502用于采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。

[0139] 可选地,所述字幕生成模型包括依次串联的事件定位器、特征提取层、单流编码器、双流编码器和双模态解码器。

[0140] 可选地,所述训练模块501还用于:

[0141] 将训练视频输入到事件定位器,从而输出所述训练视频中的各个事件;

[0142] 将所述各个事件输入到特征提取层,从而输出视觉流特征和光流特征;

[0143] 将所述视觉流特征和所述光流特征输入到单流编码器,从而输出视觉流单编码特征和光流单编码特征;

[0144] 将所述视觉流单编码特征和所述光流单编码特征输入到双流编码器,从而输出视觉流融合编码特征和光流融合编码特征;

[0145] 对所述视觉流单编码特征和所述光流单编码特征进行对比学习;以及,采用双模态解码器对所述视觉流融合编码特征和所述光流融合编码特征进行解码,并根据解码结果进行有监督学习,以更新模型参数,从而得到训练好的字幕生成模型。

[0146] 可选地,所述字幕生成模型还包括注意力池化层;

[0147] 所述训练模块501还用于：

[0148] 将所述视觉流单编码特征和视觉流检索编码特征输入到注意力池化层，从而输出视觉流池化特征；

[0149] 将所述光流单编码特征和光流检索编码特征输入到注意力池化层，从而输出光流池化特征；

[0150] 对所述视觉流池化特征和所述光流池化特征进行对比学习。

[0151] 可选地，所述训练模块501还用于：

[0152] 从特征队列中获取历史视频的视觉流池化特征和光流池化特征；

[0153] 计算所述训练视频的视觉流池化特征到所述历史视频的光流池化特征和所述训练视频的光流池化特征的交叉熵函数，从而得到第一交叉熵函数；

[0154] 计算所述训练视频的光流池化特征到所述历史视频的视觉流池化特征和所述训练视频的视觉流池化特征的交叉熵函数，从而得到第二交叉熵函数；

[0155] 将所述第一交叉熵函数和所述第二交叉熵函数相加，得到对比学习函数，从而进行对比学习。

[0156] 可选地，所述字幕生成模型还包括字幕生成器；

[0157] 所述训练模块501还用于：

[0158] 将所述视觉流融合编码特征和所述光流融合编码特征输入到双模态解码器中，从而输出事件解码特征；

[0159] 将所述事件解码特征输入到字幕生成器，从而输出所述训练视频的字幕，同时以所述训练视频对应的训练字幕为目标，进行有监督训练。

[0160] 可选地，所述单流编码器包括串联的第一多头自注意力网络和第一前馈神经网络；所述双流编码器包括依次串联的第二多头自注意力网络、第一多头交叉注意力网络和第二前馈神经网络；所述双模态解码器包括依次串联的第三多头自注意力网络、第二多头交叉注意力网络、第三前馈神经网络和第四前馈神经网络。

[0161] 可选地，所述训练模块501还用于：

[0162] 在更新模型参数的过程中，所述注意力池化层的模型参数基于动量系数进行更新，所述动量系数小于0.5。

[0163] 需要说明的是，在本发明所述字幕生成装置的具体实施内容，在上面所述字幕生成方法中已经详细说明了，故在此重复内容不再说明。

[0164] 图6示出了可以应用本发明实施例的字幕生成方法或字幕生成装置的示例性系统架构600。

[0165] 如图6所示，系统架构600可以包括终端设备601、602、603，网络604和服务器605。网络604用以在终端设备601、602、603和服务器605之间提供通信链路的介质。网络604可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0166] 用户可以使用终端设备601、602、603通过网络604与服务器605交互，以接收或发送消息等。终端设备601、602、603上可以安装有各种通讯客户端应用，例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等(仅为示例)。

[0167] 终端设备601、602、603可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0168] 服务器605可以是提供各种服务的服务器,例如对用户利用终端设备601、602、603所浏览的购物类网站提供支持的后台管理服务器(仅为示例)。后台管理服务器可以对接收到的物品信息查询请求等数据进行分析等处理,并将处理结果反馈给终端设备。

[0169] 需要说明的是,本发明实施例所提供的字幕生成方法一般由服务器605执行,相应地,所述字幕生成装置一般设置在服务器605中。本发明实施例所提供的字幕生成方法也可以由终端设备601、602、603执行,相应地,所述字幕生成装置可以设置在终端设备601、602、603中。

[0170] 应该理解,图6中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器的。

[0171] 下面参考图7,其示出了适于用来实现本发明实施例的终端设备的计算机系统700的结构示意图。图7示出的终端设备仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0172] 如图7所示,计算机系统700包括中央处理单元(CPU)701,其可以根据存储在只读存储器(ROM)702中的程序或者从存储部分708加载到随机访问存储器(RAM)703中的程序而执行各种适当的动作和处理。在RAM 703中,还存储有系统700操作所需的各种程序和数据。CPU 701、ROM 702以及RAM703通过总线704彼此相连。输入/输出(I/O)接口705也连接至总线704。

[0173] 以下部件连接至I/O接口705:包括键盘、鼠标等的输入部分706;包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分707;包括硬盘等的存储部分708;以及包括诸如LAN卡、调制解调器等网络接口卡的通信部分709。通信部分709经由诸如因特网的网络执行通信处理。驱动器710也根据需要连接至I/O接口705。可拆卸介质711,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器710上,以便于从其上读出的计算机程序根据需要被安装入存储部分708。

[0174] 特别地,根据本发明公开的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本发明公开的实施例包括一种计算机程序,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分709从网络上被下载和安装,和/或从可拆卸介质711被安装。在该计算机程序被中央处理单元(CPU)701执行时,执行本发明的系统中限定的上述功能。

[0175] 需要说明的是,本发明所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本发明中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本发明中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限

于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、电线、光缆、RF等等,或者上述的任意合适的组合。

[0176] 附图中的流程图和框图,图示了按照本发明各种实施例的系统、方法和计算机程序的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0177] 描述于本发明实施例中涉及到的模块可以通过软件的方式实现,也可以通过硬件的方式来实现。所描述的模块也可以设置在处理器中,例如,可以描述为:一种处理器包括训练模块和生成模块,其中,这些模块的名称在某种情况下并不构成对该模块本身的限定。

[0178] 作为另一方面,本发明还提供了一种计算机可读介质,该计算机可读介质可以是上述实施例中描述的设备中所包含的;也可以是单独存在,而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被一个该设备执行时,该设备实现如下方法:采用训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型;采用所述训练好的字幕生成模型生成目标视频中各个事件对应的字幕。

[0179] 作为另一方面,本发明实施例还提供了一种计算机程序产品,包括计算机程序,所述计算机程序被处理器执行时实现上述任一实施例所述的方法。

[0180] 根据本发明实施例的技术方案,因为通过训练视频对具有双模态嵌入结构的字幕生成模型进行有监督训练和对比学习,以更新模型参数,从而得到训练好的字幕生成模型的技术手段,所以克服了现有技术中不同模态特征没有很好地对齐以及对比学习不能直接应用于密集视频字幕的技术问题。本发明实施例将对对比学习应用于密集视频字幕,并采用事件对齐来捕获更多的负样本,从而增强对比学习的效果,由此训练得到的字幕生成模型可以准确地生成密集视频字幕。

[0181] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明白的是,取决于设计要求和因素,可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明保护范围之内。

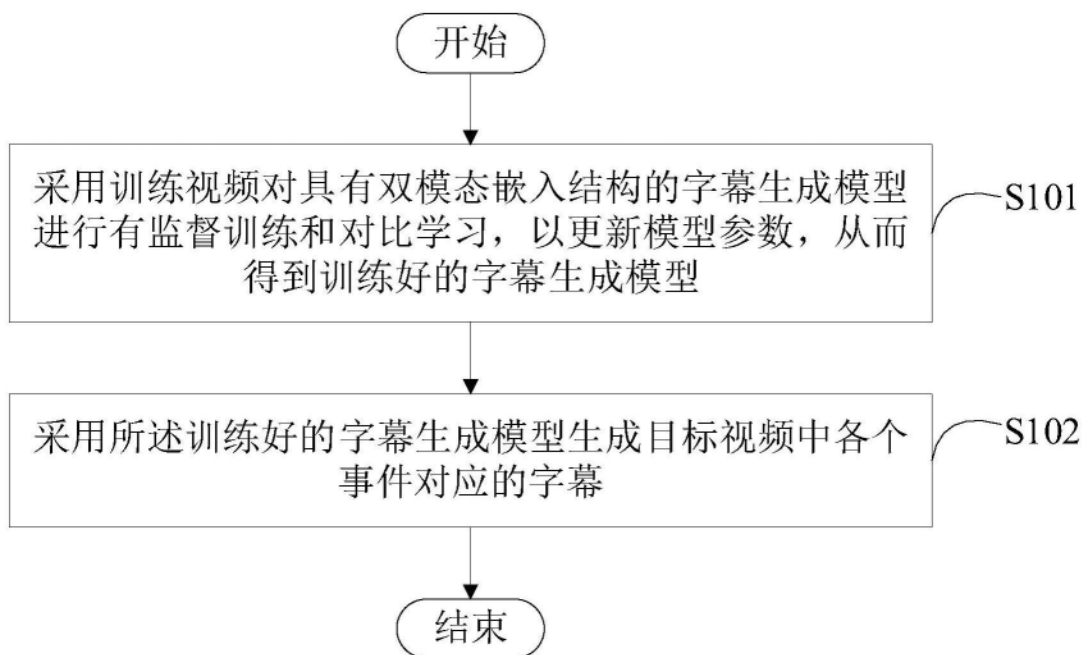


图1

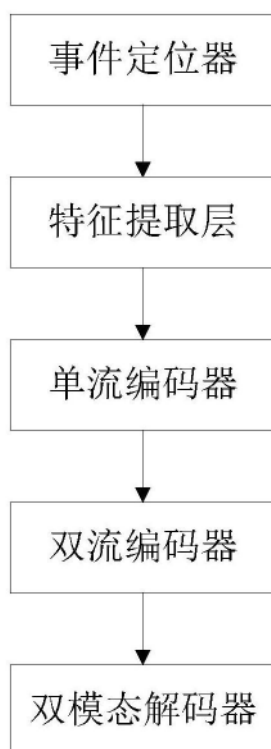


图2

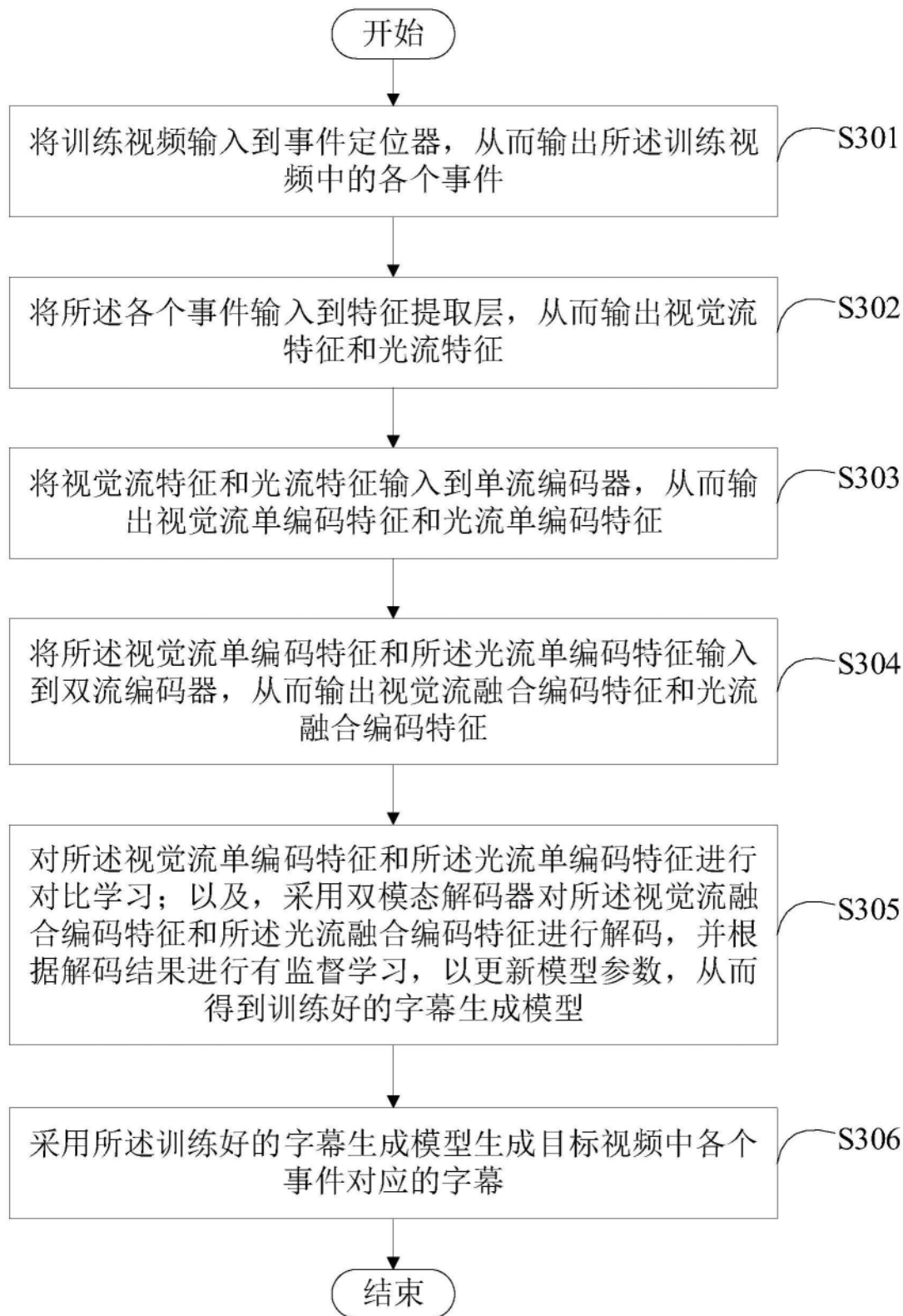
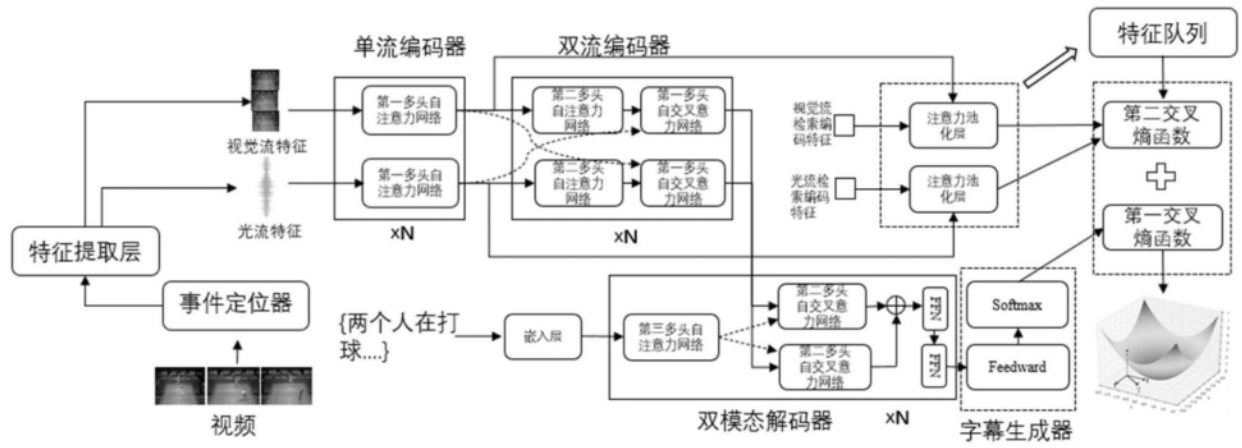
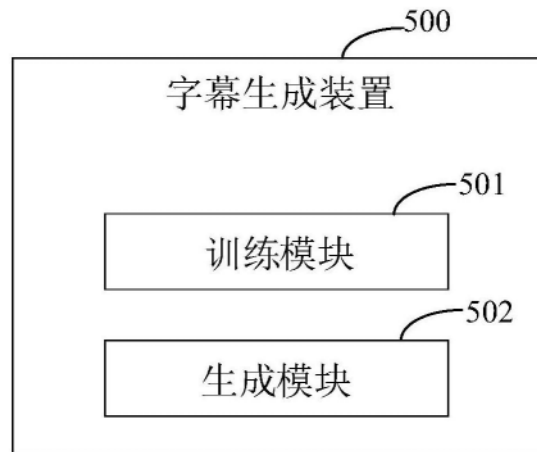


图3



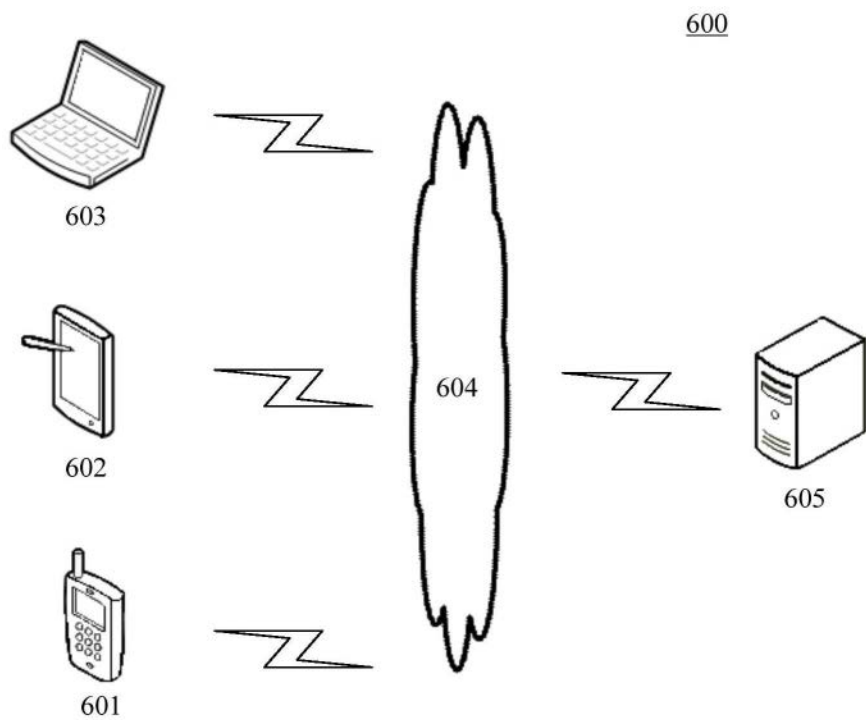


图6

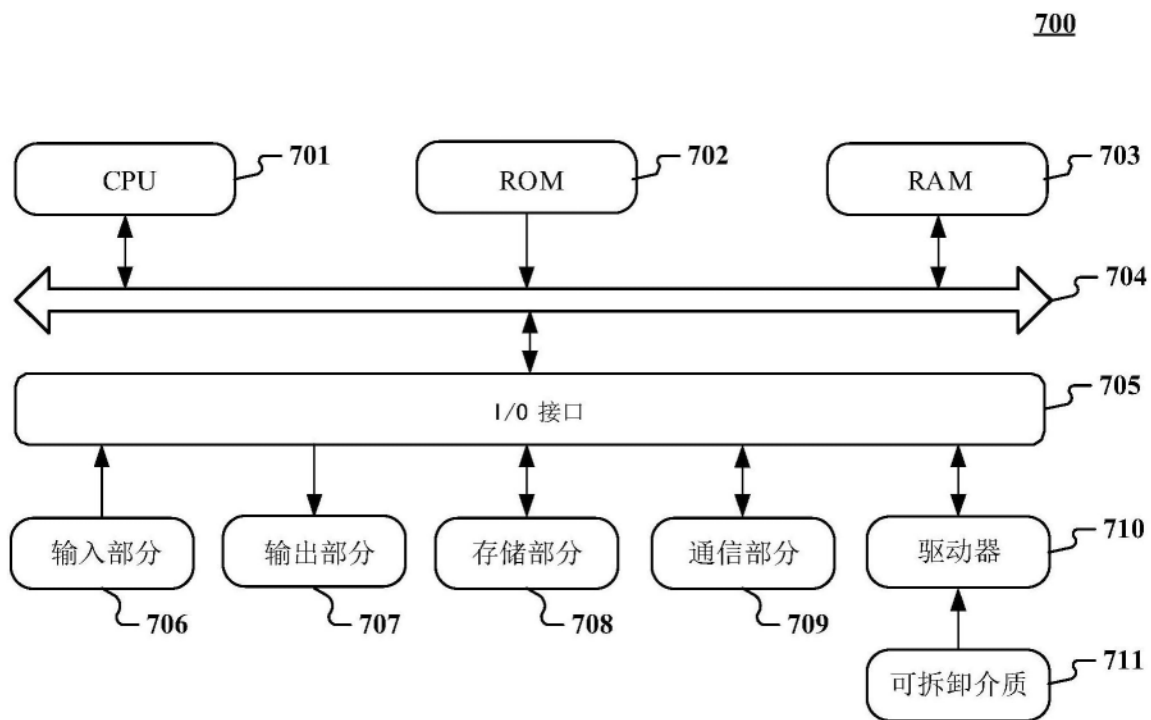


图7