(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111651635 A
(43)申请公布日 2020.09.11

(21)申请号 202010467416.X

(22)申请日 2020.05.28

(71)申请人 拾音智能科技有限公司
地址 710000 陕西省西安市航天基地神州四路航创广场C座七楼708-119号

(72)发明人 王春辉 胡勇

(74)专利代理机构 北京中北知识产权代理有限公司 11253
代理人 卢业强

(51)Int.Cl.
G06F 16/78(2019.01)

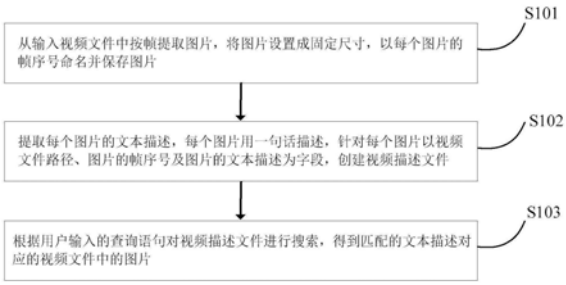
权利要求书1页 说明书5页 附图1页

(54)发明名称

一种基于自然语言描述的视频检索方法

(57)摘要

本发明公开一种基于自然语言描述的视频检索方法。所述方法包括:从输入视频文件中按帧提取图片,将图片设置成固定尺寸,以每个图片的帧序号命名并保存图片;提取每个图片的文本描述,针对每个图片,以视频文件路径、图片的帧序号及对应的文本描述为字段,创建视频描述文件;根据用户输入的查询语句对视频描述文件进行搜索,得到匹配的视频文件中的图片。由于每帧图像的文本描述是在查询之前生成的,所以用户在查询已生成文本描述的视频时能快速得到所要查询的视频和时间定位,提高了视频检索的速度。



CN 111651635 A

1.一种基于自然语言描述的视频检索方法,其特征在于,包括以下步骤:

步骤1,从输入视频文件中按帧提取图片,将图片设置成固定尺寸,以每个图片的帧序号命名并保存图片;

步骤2,提取每个图片的文本描述,每个图片用一句话描述,针对每个图片,以视频文件路径、图片的帧序号及图片的文本描述为字段,创建视频描述文件;

步骤3,根据用户输入的查询语句对视频描述文件进行搜索,得到匹配的文本描述对应的视频文件中的图片。

2.根据权利要求1所述的基于自然语言描述的视频检索方法,其特征在于,所述步骤2提取图片的文本描述的方法包括:

步骤2.1,利用DenseCap模型的卷积网络提取每个图片的特征图谱;

步骤2.2,确定候选区域并提取候选区域内的特征向量:首先,将所述特征图谱输入全卷积网络,以特征图谱中的每个像素点为锚点,将其反向映射到原始图像中,然后,基于所述锚点画出不同宽高比和不同大小的锚箱即初始边框,定位层通过回归模型预测初始边框的置信分数和位置信息;采用非极大抑制方式滤除与置信分数极高的区域重叠面积超过70%的初始边框后得到候选边框;最后,采用双线性插值法将每个候选边框内的区域提取为固定大小的特征向量,所有特征向量组成一个特征矩阵;

步骤2.3,利用一个全连接层将每个图片的特征矩阵展开成一个一维列向量;

步骤2.4,将所述一维列向量输入一个RNN网络,得到一个编码x_{-1},构建一个长度为T+2的单词向量序列$x_{-1},x_0,x_1,x_2,\cdots,x_T$,$x_0$为开始标志,$x_1,x_2,\cdots,x_T$,为图片文本描述的词系列编码;将所述向量序列输出到RNN中训练出一个预测模型;将x_{-1},x_0输入预测模型,得到单词向量y_0,根据y_0预测出第一个单词,然后将第一个单词再作为下一层RNN网络的输入,预测出第二个单词,直到输出的单词是END标志为止,得到图片的文本描述。

3.根据权利要求1所述的基于自然语言描述的视频检索方法,其特征在于,所述步骤3具体包括:

步骤3.1,读取视频描述文件,将视频描述文件中图片的文本描述输入分词组件,去掉标点符号和停用词、进行分词处理得到词元;将词元输入语言处理组件,将词元变成小写转换成词根形式,所述词根即为索引;

步骤3.2,对用户输入的查询语句进行词法分析,识别单词和关键词;进行语法分析,根据查询语句的语法规则构建语法树;进行语言处理,对查询语句进行加工;搜索索引得到符合语法树的文档即图片的文本描述;

步骤3.3,将得到的每个文档和查询语句均看作是一个词序列,按下式计算每个词的权重:

$$w=TF\times\log_e(n/d) \qquad (1)$$

式中,w为权重,TF为词在文档中出现的次数,d为包含所述词的文档的数量,n为文档总数;

将每个文档和查询语句中的每个词用其权重替换,得到每个文档和查询语句的向量表示,计算每个文档的向量与查询语句向量的余弦相似度,余弦相似度最大的文档对应的图片即为所要查询的图片。

一种基于自然语言描述的视频检索方法

技术领域

[0001] 本发明属于自然语言理解技术领域,具体涉及一种基于自然语言描述的视频检索方法。

背景技术

[0002] 视频检索定位是一个复杂并具有挑战性的问题,在视频中响应文本查询的特定时刻的定位与许多视觉任务有关,这些任务包括视频检索、时间动作定位以及视频描述和问题解答。

[0003] 视频检索,是给定一组视频候选者和一种语言查询的任务,利用视频检索算法检索与查询匹配的视频。有一种检索模型,将视频中的视觉概念与通过解析句子描述而生成的语义图进行匹配;通过为给定视频的每个句子和一组具有时间顺序的句子分配时间间隔,来解决视频文本对齐问题。最近,Hendricks等人提出了一种联合视频语言模型,用于基于纹理查询来检索视频中的时刻。但是,这些模型只能验证包含相应力矩的线段,返回结果中存在许多背景噪声。尽管可以密集采样不同比例的视频时刻,并利用这些模型来检索相应的视频时刻,但不仅计算量大,而且随着搜索空间的增加,匹配任务也更具挑战性。

[0004] 关于时间定位,Gaidon等人提出了在未修剪的视频中临时定位动作的问题,重点关注有限的动作。模型3DConvNets提出了一种基于端到端段的3D卷积神经网络(CNN)框架,该框架通过同时捕获时空信息而优于其它基于递归神经网络(RNN)的方法。还有一种新颖的时间单位回归网络模型,该模型可以联合预测动作建议并通过时间坐标回归细化时间边界。由于这些方法仅限于预先定义的动作列表,因此,有学者建议使用自然语言查询来本地化活动。他们利用了当前输入周围的所有上下文时刻,而没有明确考虑输入查询的语义信息。

[0005] 关于视频问答任务,注意力机制在神经机器翻译、视频字幕和视频问答中取得令人印象深刻的结果。用于视频字幕的视觉注意模型在每个时间步都利用视频帧,而无需明确考虑预测词的语义属性。这是不必要的,甚至是误导的。为了解决这个问题,有人利用分层的长短期记忆(LSTM)网络,该网络具有用于视频字幕的调整后的时间注意模型。后来,对注意力模型进行扩展,使其不仅选择性地参与特定时间或空间区域,而且选择性地参与输入的特定形式,例如图像特征、运动特征和音频特征。最近,一种多模态注意力LSTM网络发展较快,该网络充分利用了多模态流和时间注意力以在句子生成过程中选择性地关注特定元素。

[0006] 现有视频检索定位的方法,都在一定程度上合并了上文提到的在其它任务中的方法,以提高模型效果。但它们是端到端模式,对于一个新的查询或者新的视频都需要从头开始运行模型,运行时间长,不能快速定位,降低了用户的使用兴趣。

发明内容

[0007] 为了解决现有技术中存在的上述问题,本发明提出一种基于自然语言描述的视频

检索方法。

[0008] 为实现上述目的,本发明采用如下技术方案:

[0009] 一种基于自然语言描述的视频检索方法,包括以下步骤:

[0010] 步骤1,从输入视频文件中按帧提取图片,将图片设置成固定尺寸,以每个图片的帧序号命名并保存图片;

[0011] 步骤2,提取每个图片的文本描述,每个图片用一句话描述,针对每个图片,以视频文件路径、图片的帧序号及对应的文本描述为字段,创建视频描述文件;

[0012] 步骤3,根据用户输入的查询语句对视频描述文件进行搜索,得到匹配的文本描述对应的视频文件中的图片。

[0013] 与现有技术相比,本发明具有以下有益效果:

[0014] 本发明通过从输入视频文件中按帧提取图片,提取每个图片的文本描述,针对每个图片以视频文件路径、图片的帧序号及对应的文本描述为字段,创建视频描述文件,根据用户输入的查询语句对视频描述文件进行搜索,得到匹配的文本描述对应的视频文件中的图片,实现了基于自然语言描述的视频检索。由于每帧图像的描述是在查询之前生成的,所以用户在查询已生成文本描述的视频时能较快地得到所要查询的视频和时间定位,提高了视频检索的速度。

附图说明

[0015] 图1为本发明实施例一种基于自然语言描述的视频检索方法的流程图。

具体实施方式

[0016] 下面结合附图对本发明作进一步详细说明。

[0017] 本发明实施例一种基于自然语言描述的视频检索方法,包括以下步骤:

[0018] S101、从输入视频文件中按帧提取图片,将图片设置成固定尺寸,以每个图片的帧序号命名并保存图片;

[0019] S102、提取每个图片的文本描述,每个图片用一句话描述,针对每个图片,以视频文件路径、图片的帧序号及对应的文本描述为字段,创建视频描述文件;

[0020] S103、根据用户输入的查询语句对视频描述文件进行搜索,得到匹配的文本描述对应的视频文件中的图片。

[0021] 在本实施例中,步骤S101主要用于按帧提取视频图像,即以视频文件作为输入得到一系列的图片。可利用python模型的FFmpeg模块按视频帧数取帧,将提取出来的图片处理成相同尺寸,如720×480(像素点×像素点)。以每个图片图片的帧序号命名并保存图片。图片的帧序号即按帧提取图片时的先后顺序号,例如,第一个提取的图片的帧序号为1。

[0022] 在本实施例中,步骤S102主要用于获取图片的文本描述。可利用DenseCap模型从图片中提取文本描述。文本描述包括全局描述和局部描述,为了提高查询速度,只取模型生成的全局描述,即一张图片对应一句文本描述。DenseCap模型由三部分组成:卷积网络(Convolutional Network)、全卷积定位层(Fully Convolutional Localization Layer)和RNN语言模型。DenseCap模型能将图片中的局部细节用自然语言描述出来。该模型可以说是目标检测和普通图片描述的一种结合,即当描述的对象是一个单词的时候,就可以看作

是目标检测;当描述的对象是整幅图片的时候,就成了图片描述生成。本实施例的描述对象都是整幅图片。文本描述生成后,将视频路径、图片的帧序号和图片描述文字组合起来,最后生成以视频帧为单位的视频描述文件。

[0023] 在本实施例中,步骤S103主要用于根据查询语句对视频描述文件进行搜索获取相应的视频文件及图片。可利用python中的whoosh库搭建搜索框架,该框架基于全文检索方法进行搜索。全文检索包含索引创建和索引搜索两个过程,先建立索引,然后对索引进行搜索,得到与查询语句匹配的文本描述。所述文本描述对应的图片就是所要查询的图片。

[0024] 作为一种可选实施例,所述步骤S102提取图片的文本描述的方法包括:

[0025] S1021、利用DenseCap模型的卷积网络提取每个图片的特征图谱;

[0026] S1022、确定候选区域并提取候选区域内的特征向量:首先,将所述特征图谱输入全卷积网络,以特征图谱中的每个像素点为锚点,将其反向映射到原始图像中,然后,基于所述锚点画出不同宽高比和不同大小的锚箱即初始边框,定位层通过回归模型预测初始边框的置信分数和位置信息;采用非极大抑制方式滤除与置信分数极高的区域重叠面积超过70%的初始边框后得到候选边框;最后,采用双线性插值法将每个候选边框内的区域提取为固定大小的特征向量,所有特征向量组成一个特征矩阵;

[0027] S1023、利用全连接层将所述特征矩阵展开成一个一维列向量;

[0028] S1024、将所述一维列向量输入一个RNN网络,得到一个编码x_{-1},构建一个长度为T+2的单词向量序列$x_{-1},x_0,x_1,x_2,\cdots,x_T$,$x_0$为开始标志,$x_1,x_2,\cdots,x_T$,为图片文本描述的词系列编码;将所述向量序列输出到RNN中训练出一个预测模型;将x_{-1},x_0输入预测模型,得到单词向量y_0,根据y_0预测出第一个单词,然后将第一个单词再作为下一层RNN网络的输入,预测出第二个单词,直到输出的单词是END标志为止,得到图片的文本描述。

[0029] 本实施例给出了提取图片文本描述的一种技术方案。共包含4个步骤S1021～S1024。

[0030] 步骤S1021主要用于利用卷积网络提取图片的特征图谱。特征图包含多种类型特征,比如图片的纹理、光线强度、形状等,每一处的数值代表了某个特征强弱值。由于卷积神经网络的特点,随着层数的加深,获取的特征会越抽象并包含更多的语义信息。DenseCap模型的卷积网络采用了基于VGG-16的网络结构,包含13层卷积核为3×3的卷积层和4层池化核为2×2的最大池化层。对于大小为3×720×480(三维矩阵,三维指红、绿、蓝三个色通道)的图片,经过卷积网络后,输出结果是512×45×30的特征图谱。所述特征图谱是下一层全卷积定位层FCL的输入。

[0031] 步骤S1022主要用于确定候选区域并提取候选区域内的特征向量。该步骤主要由全卷积网络完成。全卷积定位层是整个模型的核心部分,与Faster R-Cnn类似,用于生成识别图片内物体边框。它的输入是来自卷积网络的特征图谱,输出是定长的多个(如300个)候选区域的特征向量,每个特征向量包含候选区域坐标、置信分数和候选区域特征三个数据。置信分数越大说明越接近真实区域。全卷积定位层的处理过程包括四个步骤:第一步是卷积锚点。首先以来自卷积网络尺寸为C×W'×H'的特征图谱中的每一个像素点作为一个锚点,将该点反向映射到原始图像中,然后基于该锚点画出不同宽高比和大小不同的锚箱,组合出来的锚箱数目为k(如k=12),对于每个锚箱,FCL中的定位层会通过回归模型来预测相应的置信分数和位置信息。具体的计算过程是将特征图片作为输入,经过一个卷积核为3×

3的卷积层，然后再经过一个卷积核为1×1的卷积层，卷积核数目为5k，所以这一层的最终输出的是5k×W'×H'的三维数组，包含了所有锚点对应的置信分数和位置信息。第二步进行边框回归。这是对初始边框的一次精修。由于上一步得到的边框与真实区域可能并不特别匹配，在真实区域的监督下利用线性回归得到边框的四个位移值，这四个位移值主要是用来更新候选区域中点坐标的横纵坐标值和候选边框的长与宽。第三步是采样，由于通过前两步得到的候选边框过多，为了降低运行成本，需要对候选边框进行采样，通过非极大抑制的方式选取300个候选边框，其中非极大抑制方法是去掉与具有极高置信分数区域重叠面积超过70％的候选边框，从而减少了重叠区域的输出，更加精细化的地定位目标位置。第四步是进行双线性插值。经采样后得到的各个候选区域是具有不同大小和宽高比的矩形框即候选区域，为了与后续的全连接层即识别网络和RNN语言模型建立连接，模型利用双线性插值法将候选区域提取为固定大小的特征向量，将所有候选区域的特征向量组合成特征矩阵。

[0032] 步骤S1023主要用于将上一步得到的特征矩阵展开成一个一维列向量，然后将所有正样本一维列向量组合成一个矩阵。该步骤主要由一个全连接的神经网络完成。它将每个候选区域的特征拉展平成一个一维列向量，令其经过两层全连接层，每次都使用ReLU激活函数和Dropout优化原则。最后，每一个候选区域都生成一个长度为D＝4096的一维向量。将所有一维向量存储起来，形成一个300×4096的矩阵，将该矩阵传送到下一步的RNN语言模型中。另外，还可以对候选区域的置信分数和位置信息进行二次精修，从而生成每个候选区域最终的置信分数和位置信息。这一次的精修与之前的边界回归基本是一样的，只不过是针对这个长度的向量又进行了一次边界回归而已。

[0033] 步骤S1024主要用于输出图片的文本描述。该步骤主要由利RNN网络(又称RNN语言模型)完成，它以上一步得到的一维特征向量作为输入，输出基于描述图片内容的自然语言序列。

[0034] DenseCap模型的关键点是FCLN结构并利用双线性插值使得定位层可导，从而可以支持从图片区域到自然语言描述之间的端到端训练。实验结果表明，本实施例的网络结构相较于以前的网络结构，不论是在生成的图片描述的质量上，还是在生成速度上，都有一定的提升。鉴于Densecap的优点，本实施例使用了Densecap预训练好了的模型，以图片的文本描述为输出，并构建以视频路径、图片的帧序号、图片的文本描述为字段的文件。由于Densecap是介于物体识别和普通描述之间，所以最后产生的图片描述相比普通的描述生成模型有更多局部区域的信息，提高了视频检索定位的准确性。

[0035] 作为一种可选实施例，所述步骤S103具体包括：

[0036] S1031、读取视频描述文件，将视频描述文件中图片的文本描述输入分词组件，去掉标点符号和停用词、进行分词处理得到词元；将词元输入语言处理组件，将词元变成小写转换成词根形式，所述词根即为索引；

[0037] S1032、对用户输入的查询语句进行词法分析，识别单词和关键词；进行语法分析，根据查询语句的语法规则构建语法树；进行语言处理，对查询语句进行加工；搜索索引得到符合语法树的文档即图片的文本描述；

[0038] S1033、将得到的每个文档和查询语句均看作是一个词序列，按下式计算每个词的权重：

[0039] $w = TF \times \log_e(n/d)$ (1)

[0040] 式中，w为权重，TF为词在文档中出现的次数，d为包含所述词的文档的数量，n为文档总数；

[0041] 将每个文档和查询语句中的每个词用其权重替换，得到每个文档和查询语句的向量表示，计算每个文档的向量与查询语句向量的余弦相似度，余弦相似度最大的文档对应的图片即为所要查询的图片。

[0042] 本实施例给出了从视频描述文件中搜索与查询语句匹配的图片的一种技术方案。共包含3个步骤S1031～S1033。

[0043] 步骤S1031主要用于创建索引。创建索引是将视频描述文件中的文本描述进行语言处理从而用词元创建索引的过程。主要由分词组件和语言处理组件实现。分词组件去掉文本描述中的标点符号、停用词(无实际意义的词，如a、an等)，进行分词处理得到词元。例如，文本“I am driving a car on the road”经分词组件后得到词元“I”、“driving”、“car”、“road”。语言处理组件进一步将词元变成小写转换成词根形式，所述词根即为创建的索引。上面例子得到的索引为“i”、“driving”、“car”、“road”。

[0044] 步骤S1032主要用于搜索索引。首先对查询语句进行词法分析，即识别单词和关键词；然后对查询语句进行语法分析，即根据查询语句的语法规则构建语法树；还要进行语言处理，即对原始查询语句的进一步加工。最后搜索上一步建立的索引，得到符合语法树的文档，也就是图片的文本描述。

[0045] 步骤S1033主要用于从上一步得到的文档中筛选出与查询语句最匹配的文档，即图片的文本描述，从而得到所要查询的视频文件及图片。首先按照公式(1)计算每个文档及查询语句中每个词的权重；然后用每个词的权重替换每个词，得到每个文档及查询语句用权重表示的向量；计算每个文档向量与查询语句向量的余弦相似度，余弦相似度最大的文档即为所要查询图片的文本描述。有了图片的文本描述也就有了图片所在的视频文件名和图片的号码。

[0046] 上述仅对本发明中的几种具体实施例加以说明，但并不能作为本发明的保护范围，凡是依据本发明中的设计精神所做出的等效变化或修饰或等比例放大或缩小等，均应认为落入本发明的保护范围。

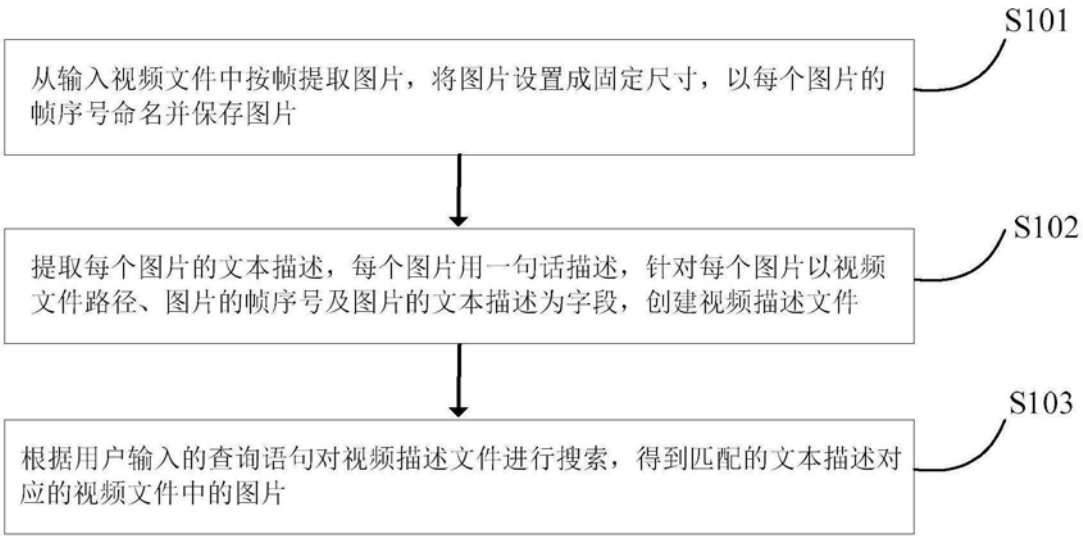

S101
从输入视频文件中按帧提取图片，将图片设置成固定尺寸，以每个图片的帧序号命名并保存图片
S102
提取每个图片的文本描述，每个图片用一句话描述，针对每个图片以视频文件路径、图片的帧序号及图片的文本描述为字段，创建视频描述文件
S103
根据用户输入的查询语句对视频描述文件进行搜索，得到匹配的文本描述对应的视频文件中的图片

图1