



(12) 发明专利

(10) 授权公告号 CN 110162669 B

(45) 授权公告日 2021. 07. 02

(21) 申请号 201910273083.4

(22) 申请日 2019.04.04

(65) 同一申请的已公布的文献号
申请公布号 CN 110162669 A

(43) 申请公布日 2019.08.23

(73) 专利权人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市南山区高新区
科技中一路腾讯大厦35层

(72) 发明人 何奕江 郑茂

(74) 专利代理机构 广州华进联合专利商标代理
有限公司 44224
代理人 黄晶晶

(51) Int. Cl.
G06F 16/75 (2019.01)
G06K 9/62 (2006.01)

(56) 对比文件

CN 109359636 A, 2019.02.19

CN 109359636 A, 2019.02.19

CN 109508584 A, 2019.03.22

CN 109472232 A, 2019.03.15

CN 104504365 A, 2015.04.08

US 2018144194 A1, 2018.05.24

张玉珍等. 基于多模态融合的足球视频语义
分析.《计算机科学》.2010,第37卷(第7期),

审查员 曾婉琳

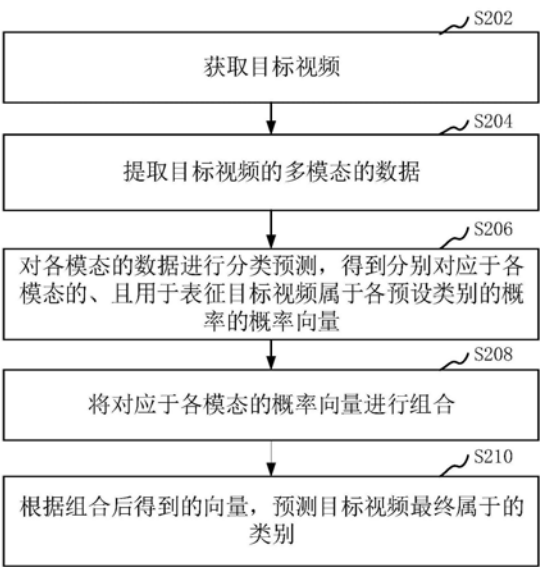
权利要求书4页 说明书18页 附图8页

(54) 发明名称

视频分类处理方法、装置、计算机设备及存
储介质

(57) 摘要

本申请涉及一种视频分类处理方法、装置、
计算机设备及存储介质,该方法包括:获取目标
视频;提取所述目标视频的多模态的数据;对各
模态的数据进行分类预测,得到分别对应于各模
态的、且用于表征目标视频属于各预设类别的概
率的概率向量;将对应于各模态的概率向量进行
组合;根据组合后得到的向量,预测所述目标视
频最终属于的类别。本申请的方案,能够提高视
频分类的准确性。



1. 一种视频分类处理方法,所述方法包括:

获取目标视频;

提取所述目标视频的多模态的数据;

对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量;

将对应于各模态的概率向量进行组合;

将组合后得到的向量分别输入预先训练的梯度提升分类回归树模型的分分类回归树中进行分类处理,得到各分类回归树所预测的用于表征目标视频属于各预设类别的概率的子概率向量;

获取每棵分类回归树相应的决策权重;

将各棵分类回归树所预测的子概率向量按照相应的决策权重进行加权平均,得到用于表征目标视频属于各预设类别的概率的最终概率向量;

选取所述最终概率向量中的最大概率值所对应的预设类别,得到所述目标视频最终属于的类别。

2. 根据权利要求1所述的方法,其特征在于,所述多模态的数据包括从目标视频中提取的属于视频帧模态的视频帧、属于视频文本模态的描述文本、以及属于音频模态的音频文件中的至少两种。

3. 根据权利要求1所述的方法,其特征在于,所述多模态的数据包括属于视频文本模态的描述文本;所述对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量包括:

编码对描述文本进行分词后的词片段,生成各词片段相应的隐含状态向量;

获取各隐含状态向量对应的注意力权重;

根据各隐含状态向量和相应注意力权重,得到最终隐含状态向量;

解码所述最终隐含状态向量,得到对应于视频文本模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

4. 根据权利要求3所述的方法,其特征在于,所述编码对描述文本进行分词后的词片段,生成各词片段相应的隐含状态向量包括:

将对描述文本进行分词后的词片段编码为词向量,得到词向量序列;

按照词向量序列中各词向量的先后顺序,对各词向量依次进行编码,在每次编码时,根据当前词向量和对前一词向量编码输出的隐含状态向量进行编码,生成所述当前词向量相应的隐含状态向量。

5. 根据权利要求4所述的方法,其特征在于,所述在每次编码时,根据当前词向量和对前一词向量编码输出的隐含状态向量进行编码,生成所述当前词向量相应的隐含状态向量包括:

在每次编码时,将当前词向量和前一编码单元输出的隐含状态向量输入文本分类模型中的当前编码单元,生成当前编码单元的临时单元状态向量、所述临时单元状态向量相应的输入权重、上一单元状态向量的遗忘权重和当前单元状态向量的输出权重;

将所述上一单元状态向量按照所述遗忘权重进行保留,得到第一向量;所述第一向量用于表征上一单元状态需保留信息;

将所述临时单元状态向量按照所述输入权重进行保留,得到第二向量;所述第二向量用于表征当前词向量需保留信息;

根据第一向量和第二向量,得到所述当前编码单元的当前单元状态向量;

将所述当前单元状态向量按照所述输出权重进行输出,得到当前词向量相应的隐含状态向量。

6. 根据权利要求1所述的方法,其特征在于,所述多模态的数据包括从目标视频中提取的属于视频帧模态的视频帧;

所述对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量包括:

分别卷积处理各视频帧,得到与各视频帧相应的视频帧特征向量;

对各视频帧特征向量进行平均池化处理;

对平均池化处理后的向量进行归一化处理,生成对应于所述视频帧模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

7. 根据权利要求6所述的方法,其特征在于,所述分别卷积处理各视频帧,得到与各视频帧相应的视频帧特征向量包括:

将每个视频帧输入卷积神经网络模型中进行多层卷积处理,在每层卷积处理时,将所述视频帧和所述视频帧的前一层卷积结果作为当层卷积处理的输入,进行联合卷积处理,得到当层卷积结果;

将最后一层卷积结果,作为所述视频帧的视频帧特征向量。

8. 根据权利要求1所述的方法,其特征在于,所述多模态的数据包括属于音频模态的音频文件;所述对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量包括:

将所述音频文件进行切割,得到音频片段;

提取各所述音频片段的频率特征;

分别对各所述频率特征进行卷积处理,得到相应的音频特征向量;

根据各所述音频特征向量进行分类预测,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

9. 根据权利要求8所述的方法,其特征在于,所述根据各所述音频特征向量进行分类预测,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量包括:

按照卷积后的音频特征向量的先后顺序,循环地将当前音频特征向量和前次编码后的编码向量作为当次编码的输入;

对最后一次编码后的编码向量进行归一化处理,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

10. 根据权利要求8所述的方法,其特征在于,所述将所述音频文件进行切割,得到音频片段包括:

将音频文件中的多通道转换为单通道;

按照预设采样率,对转换后的音频文件中的音频信号进行重采样;

将重采样后的音频文件切割为音频片段。

11. 根据权利要求1至10中任一项所述的方法,其特征在于,所述目标视频是内容推荐

平台中的视频；

所述方法还包括：

获取所述内容推荐平台中的用户对象所偏好的视频类别；

将目标视频最终属于的类别与各用户对象所偏好的视频类别进行匹配；

向匹配成功的用户对象相应的终端推荐所述目标视频。

12. 一种视频分类处理装置，其特征在于，所述装置包括：

提取模块，用于获取目标视频；提取所述目标视频的多模态的数据；

分类预测模块，用于对各模态的数据进行分类预测，得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量；

组合模块，用于将对应于各模态的概率向量进行组合；

所述分类预测模块还用于将组合后得到的向量分别输入预先训练的梯度提升分类回归树模型的各分类回归树中进行分类处理，得到各分类回归树所预测的用于表征目标视频属于各预设类别的概率的子概率向量；获取每棵分类回归树相应的决策权重；将各棵分类回归树所预测的子概率向量按照相应的决策权重进行加权平均，得到用于表征目标视频属于各预设类别的概率的最终概率向量；选取所述最终概率向量中的最大概率值所对应的预设类别，得到所述目标视频最终属于的类别。

13. 根据权利要求12所述的装置，其特征在于，所述多模态的数据包括从目标视频中提取的属于视频帧模态的视频帧、属于视频文本模态的描述文本、以及属于音频模态的音频文件中的至少两种。

14. 根据权利要求12所述的装置，其特征在于，所述多模态的数据包括属于视频文本模态的描述文本；所述分类预测模块还用于编码对描述文本进行分词后的词片段，生成各词片段相应的隐含状态向量；获取各隐含状态向量对应的注意力权重；根据各隐含状态向量和相应注意力权重，得到最终隐含状态向量；解码所述最终隐含状态向量，得到对应于视频文本模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

15. 根据权利要求14所述的装置，其特征在于，所述分类预测模块还用于将对描述文本进行分词后的词片段编码为词向量，得到词向量序列；按照词向量序列中各词向量的先后顺序，对各词向量依次进行编码，在每次编码时，根据当前词向量和对前一词向量编码输出的隐含状态向量进行编码，生成所述当前词向量相应的隐含状态向量。

16. 根据权利要求15所述的装置，其特征在于，所述分类预测模块还用于在每次编码时，将当前词向量和前一编码单元输出的隐含状态向量输入文本分类模型中的当前编码单元，生成当前编码单元的临时单元状态向量、所述临时单元状态向量相应的输入权重、上一单元状态向量的遗忘权重和当前单元状态向量的输出权重；将所述上一单元状态向量按照所述遗忘权重进行保留，得到第一向量；所述第一向量用于表征上一单元状态需保留信息；将所述临时单元状态向量按照所述输入权重进行保留，得到第二向量；所述第二向量用于表征当前词向量需保留信息；根据第一向量和第二向量，得到所述当前编码单元的当前单元状态向量；将所述当前单元状态向量按照所述输出权重进行输出，得到当前词向量相应的隐含状态向量。

17. 根据权利要求12所述的装置，其特征在于，所述多模态的数据包括从目标视频中提取的属于视频帧模态的视频帧；所述分类预测模块还用于分别卷积处理各视频帧，得到与

各视频帧相应的视频帧特征向量;对各视频帧特征向量进行平均池化处理;对平均池化处理后的向量进行归一化处理,生成对应于所述视频帧模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

18.根据权利要求17所述的装置,其特征在于,所述分类预测模块还用于将每个视频帧输入卷积神经网络模型中进行多层卷积处理,在每层卷积处理时,将所述视频帧和所述视频帧的前一层卷积结果作为当层卷积处理的输入,进行联合卷积处理,得到当层卷积结果;将最后一层卷积结果,作为所述视频帧的视频帧特征向量。

19.根据权利要求12所述的装置,其特征在于,所述多模态的数据包括属于音频模态的音频文件;所述分类预测模块还用于将所述音频文件进行切割,得到音频片段;提取各所述音频片段的频率特征;分别对各所述频率特征进行卷积处理,得到相应的音频特征向量;根据各所述音频特征向量进行分类预测,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

20.根据权利要求19所述的装置,其特征在于,所述分类预测模块还用于按照卷积后的音频特征向量的先后顺序,循环地将当前音频特征向量和前次编码后的编码向量作为当次编码的输入;对最后一次编码后的编码向量进行归一化处理,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

21.根据权利要求19所述的装置,其特征在于,所述分类预测模块还用于将音频文件中的多通道转换为单通道;按照预设采样率,对转换后的音频文件中的音频信号进行重采样;将重采样后的音频文件切割为音频片段。

22.根据权利要求12至21中任一项所述的装置,其特征在于,所述目标视频是内容推荐平台中的视频;

所述装置还包括:

推荐模块,用于获取所述内容推荐平台中的用户对象所偏好的视频类别;将目标视频最终属于的类别与各用户对象所偏好的视频类别进行匹配;向匹配成功的用户对象相应的终端推荐所述目标视频。

23.一种计算机设备,其特征在于,包括存储器和处理器,所述存储器中存储有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行权利要求1至11中任一项所述方法的步骤。

24.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质上存储有计算机程序,所述计算机程序被处理器执行时,使得所述处理器执行权利要求1至11中任一项所述方法的步骤。

视频分类处理方法、装置、计算机设备及存储介质

技术领域

[0001] 本发明涉及计算机技术领域,特别是涉及一种视频分类处理方法、装置、计算机设备及存储介质。

背景技术

[0002] 随着科学技术的飞速发展,大量先进技术不断涌现,视频处理技术即为其中一个非常重要的技术。在视频处理技术的广泛应用中,如何通过机器自动对视频进行准确分类成为一个研究的热点问题。

[0003] 传统方法中,是通过从视频中抽取视频帧,根据抽取的视频帧来进行分类处理,从而得到视频的类别。由于抽取的视频帧相对于原始的视频来说,信息量减少,因此,基于较少信息量的视频帧来进行分类,导致分类的准确性比较低。

发明内容

[0004] 基于此,有必要针对传统视频分类处理方法存在的准确性比较低的问题,提供一种视频分类处理方法、装置、计算机设备及存储介质。

[0005] 一种视频分类处理方法,所述方法包括:

[0006] 获取目标视频;

[0007] 提取所述目标视频的多模态的数据;

[0008] 对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量;

[0009] 将对应于各模态的概率向量进行组合;

[0010] 根据组合后得到的向量,预测所述目标视频最终属于的类别。

[0011] 一种视频分类处理装置,所述装置包括:

[0012] 提取模块,用于获取目标视频;提取所述目标视频的多模态的数据;

[0013] 分类预测模块,用于对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量;

[0014] 组合模块,用于将对应于各模态的概率向量进行组合;

[0015] 所述分类预测模块还用于根据组合后得到的向量,预测所述目标视频最终属于的类别。

[0016] 一种计算机设备,包括存储器和处理器,所述存储器中存储有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行如下步骤:

[0017] 获取目标视频;

[0018] 提取所述目标视频的多模态的数据;

[0019] 对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量;

[0020] 将对应于各模态的概率向量进行组合;

- [0021] 根据组合后得到的向量,预测所述目标视频最终属于的类别。
- [0022] 一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机程序,所述计算机程序被处理器执行时,使得所述处理器执行如下步骤:
- [0023] 获取目标视频;
- [0024] 提取所述目标视频的多模态的数据;
- [0025] 对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量;
- [0026] 将对应于各模态的概率向量进行组合;
- [0027] 根据组合后得到的向量,预测所述目标视频最终属于的类别。
- [0028] 上述视频分类处理方法、装置、计算机设备及存储介质,提取所述目标视频的多模态的数据,分别考虑目标视频不同模态的数据来对目标视频进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量。即,考虑了多模态的语义信息,来分别对目标视频进行分类预测。将分类得到的各模态的概率向量进行组合,根据组合后得到的向量,预测所述目标视频最终属于的类别,将多种分类概率情况进行了融合,相较于仅仅根据抽取的信息量较少的视频帧来进行分类而言,实现了基于多种模态下的信息对目标视频进行分类,提高了分类所参考的信息量,提高了视频分类的准确性。

附图说明

- [0029] 图1为一个实施例中视频分类处理方法的应用场景图;
- [0030] 图2为一个实施例中视频分类处理方法的流程示意图;
- [0031] 图3为一个实施例中视频分类处理方法的流程简化示意图;
- [0032] 图4为一个实施例中的文本分类模型的架构示意图;
- [0033] 图5为一个实施例中视频帧分类模型的架构示意图;
- [0034] 图6为一个实施例中音频分类模型的架构示意图;
- [0035] 图7为一个实施例中视频推荐的流程简化示意图;
- [0036] 图8为一个实施例中视频分类处理装置的框图;
- [0037] 图9为另一个实施例中视频分类处理装置的框图;
- [0038] 图10为一个实施例中计算机设备的框图。

具体实施方式

[0039] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0040] 图1为一个实施例中视频分类处理方法的应用场景图。参照图1,该应用场景中包括第一终端110、服务器120和第二终端130。第一终端110和第二终端130皆可以是智能电视机、智能音箱、台式计算机或移动终端,移动终端可以包括手机、平板电脑、笔记本电脑、个人数字助理和穿戴式设备等中的至少一种。服务器120可以是具备视频分类功能的视频处理平台,服务器120可以用独立的服务器或者是多个物理服务器组成的服务器集群来实现。

[0041] 可以理解,第一终端110上传源视频至服务器120。比如,多个拍摄或提供视频的用

户分别通过各自使用的第一终端110拍摄视频,并将视频上传至服务器120进行存储。服务器120可以执行本申请各实施例中的视频分类处理方法对存储的各视频进行分类处理。服务器120可以根据视频分类处理结果,将所存储的视频向使用第二终端120的用户进行推荐。

[0042] 需要说明的是,第一终端110和第二终端130的身份互换,即,第一终端110除了可以上传视频至服务器120以外,也可以接收服务器120所推荐的视频;第二终端120除了可以接收服务器120推荐的视频以外,也可以上传视频至服务器120。

[0043] 服务器120在对所存储的视频进行分类处理时,可以获取目标视频;提取所述目标视频的多模态的数据;对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量。服务器120可以将对应于各模态的概率向量进行组合;根据组合后得到的向量,预测所述目标视频最终属于的类别。进而,服务器120可以依照视频的分类,将该视频推荐给相应用户所使用的第二终端130。

[0044] 需要说明的是,图1仅是一个实施例中,并不限定于对视频分类后必须用于推荐这一个使用场景,也可以是按照分类类别对视频进行统一管理或其他方面的场景应用。

[0045] 图2为一个实施例中视频分类处理方法的流程示意图。本实施例主要以该视频分类处理方法应用于计算机设备为例进行举例说明,该计算机设备可以为图1中的服务器120。参照图2,该方法具体包括如下步骤:

[0046] S202,获取目标视频。

[0047] 其中,目标视频,是待分类的视频。目标视频可以是任意时长的视频内容。其中,视频内容中包括一帧一帧具有图像内容的视频帧。

[0048] 在一个实施例中,目标视频可以包括长视频和短视频中的至少一种。

[0049] 长视频,是指时长长于短视频的、且经过专业人员制作得到的视频。可以理解,经过专业制作的长视频,通常由于数据量过大,不便于在互联网平台上快速传播。在一个实施例中,长视频可以包括影视作品和规范化教学视频等至少一种。

[0050] 短视频,即短片视频,一般是时长便于在互联网平台上快速传播的视频。比如,新闻媒体平台中经常播放的时长较短的视频。比如,腾讯新闻中播放的新闻视频。

[0051] 在一个实施例中,计算机设备可以主动地或者根据预设的触发机制被触发地执行步骤S202~S210以对目标视频进行分类处理。在另一个实施例中,计算机设备也可以在接收到发送的针对目标视频的视频分类请求后,被触发执行步骤S202~S210。

[0052] S204,提取目标视频的多模态的数据。

[0053] 其中,目标视频的多模态的数据,是指能够反映目标视频的特征的不同形态的数据。

[0054] 在一个实施例中,目标视频的多模态数据包括从目标视频中提取的属于视频帧模态的视频帧、属于视频文本模态的描述文本、以及属于音频模态的音频文件中的至少两种。

[0055] 其中,视频帧模态,即为从视频帧形态反映目标视频的特征。所以,视频帧模态的数据,即为能够反映目标视频的特征的视频帧模态的视频帧。

[0056] 视频文本模态,即为从文本形态反映目标视频的特征。所以,视频文本模态的数据,即为能够反映目标视频的特征的关于目标视频的描述文本。其中,描述文本,是描述目标视频的相关文本内容。在一个实施例中,描述文本包括视频的标题、简介和用户提供的视

频标签等中的至少一种。其中,视频标签,是用户在上传视频时,针对视频添加的自定义标注。

[0057] 音频模态,即为从音频形态反映目标视频的特征,所以,音频模态的数据,是指能够反映目标视频的特征的音频文件。可以理解,由于视频不仅具有图像还具有声音,所以,音频文件是指与目标视频中图像相匹配的声音数据所形成的文件。

[0058] 在一个实施例中,计算机设备可以从目标视频中提取视频帧,得到视频帧模态的视频帧(即为视频帧模态的数据)。在一个实施例中,计算机设备可以按照预设时间间隔,从目标视频中提取视频帧。在另一个实施例中,计算机设备也可以将目标视频划分为多个视频片段,根据视频片段中各视频帧之间的距离,提取距离差异大于预设阈值的关键帧,将提取的关键帧作为视频帧模态的视频帧。

[0059] S206,对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0060] 具体地,计算机设备可以针对每个模态的数据,分别进行分类预测处理,分别得到对应于该模态的、且用于表示目标视频属于各预设类别的概率的概率向量。比如,当对2个模态的数据进行分类预测处理,则可以得到分别对应于各模态的概率向量,即可以得到2个概率向量,当对3个模态的数据进行分类预测处理,则可以得到分别对应于各模态的概率向量,即可以得到3个概率向量。

[0061] 其中,概率向量中包括的向量元素是用于表示目标视频属于各预设类别的概率。比如,预设类别有5个A~E,针对某个模态的数据进行分类预测,得到目标视频属于A~E这5个类别的概率分别为0.1、0.1、0.7、0.05、0.05,那么,能够得到的概率向量即可以为(0.1, 0.1, 0.7, 0.05, 0.05)。

[0062] 具体地,计算机设备可以直接将各模态的数据分别输入相应预先训练的机器学习分类模型中进行分类预测,输出得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0063] 可以理解,不同模态的数据所输入的机器学习分类模型不同。各个机器学习分类模型可以是一个完整独立的机器学习模型,也可以是多个模型组合后的机器学习模型。对此不作限定。

[0064] 可以理解,计算机设备可以预先根据样本视频的多模态的样本数据以及样本视频的视频标签作为训练数据,迭代地进行有监督地机器学习训练,来不断地调整对应于不同模态的机器学习分类模型的模型参数,直至得到训练完成的对应于不同模态的各机器学习分类模型。

[0065] S208,将对应于各模态的概率向量进行组合。

[0066] 在一个实施例中,计算机设备可以直接以拼接组合的方式,将各模态的概率向量进行拼接。可以理解,计算机设备可以不限定各模态的概率向量之间拼接的前后顺序,也可以按照预设顺序,对各模态的概率向量进行拼接。比如,有2个模态,4个预设类别,模态1对应的概率向量为(0.1, 0.35, 0.45, 0.1),模态2对应的概率向量为(0.05, 0.41, 0.45, 0.06),那么,将二者拼接组合,即可以得到向量(0.1, 0.35, 0.45, 0.1, 0.05, 0.41, 0.45, 0.06)。

[0067] 在另一个实施例中,计算机设备也可以采取向量元素重组的方式,对各模态的概率向量进行组合,得到组合后的向量。同样以模态1对应概率向量(0.1, 0.35, 0.45, 0.1),模

态2对应概率向量(0.05,0.41,0.45,0.06)为例进行解释说明,那么,以向量元素重组的方式进行组合,就可以组合得到向量(0.1,0.35,0.05,0.41,0.45,0.1,0.45,0.06)。

[0068] S210,根据组合后得到的向量,预测目标视频最终属于的类别。

[0069] 具体地,计算机设备可以将组合后得到的向量输入预先训练的最终的机器学习分类模型中,通过该最终的机器学习分类模型输出目标视频最终属于的类别。可以理解,目标视频最终属于的类别,是预设类别中的其中一个类别。

[0070] 其中,最终的机器学习分类模型,是指用于预测输出最终的类别的机器学习分类模型。

[0071] 在一个实施例中,该最终的机器学习分类模型可以为梯度提升分类回归树模型(GBDT,Gradient Boosting Decision Tree)。可以理解,梯度提升分类回归树模型中包括多棵分类回归树。计算机设备可以将组合后得到的向量输入梯度提升分类回归树模型中的各棵分类回归树中进行分类预测,得到该目标视频最终属于的类别。

[0072] 需要说明的是,在其他实施例中,该最终的机器学习分类模型也可以是其他机器学习分类模型,比如,随机森林模型(Random Forest)或自适应提升分类模型(Adaboost)等由多个弱分类器组成的强分类模型。此外,最终的机器学习分类模型可以不限定于由多个弱分类器组成的强分类模型,也可以是单独的机器学习模型。

[0073] 图3为一个实施例中视频分类处理方法的流程简化示意图。图3是以多模态的数据为属于视频帧模态的视频帧、属于视频文本模态的描述文本、以及属于音频模态的音频文件为例进行说明。参照图3,与视频文本模态对应的机器学习分类模型为文本分类模型,与视频帧模态对应的机器学习分类模型为视频帧分类模型,与音频模态对应的机器学习分类模型为音频分类模型。计算机设备可以在接收到视频分类请求后,对视频分类请求所针对的目标视频进行多模态文件抽取,以抽取属于视频帧模态的视频帧、属于视频文本模态的描述文本以及属于音频模态的音频文件,计算机设备可以分别将抽取的视频帧输入视频帧分类模型进行分类预测、将抽取的描述文本输入文本分类模型进行分类预测、以及将抽取的音频文件输入音频分类模型中进行分类预测。可以理解,对上述三个模态的数据进行分类预测的先后顺序不做限定。接着,计算机设备可以将三个分类预测结果进行融合,将基于融合得到的最终的结果返回至视频分类请求方。

[0074] 上述视频分类处理方法,提取所述目标视频的多模态的数据,分别考虑目标视频不同模态的数据来对目标视频进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量。将分类得到的各模态的概率向量进行组合,根据组合后得到的向量,预测所述目标视频最终属于的类别,相当于,将多种分类概率情况进行了融合,相较于仅仅根据抽取的信息量较少的视频帧来进行分类而言,实现了基于多种模态下的信息对目标视频进行分类,提高了分类所参考的信息量,提高了视频分类的准确性。

[0075] 此外,考虑到基于多模态的信息来进行视频分类,这样一来,在某个模态表现不佳时其他模态能够进行信息的补充,保证了分类所参考的信息量的稳定性,从而提升了视频分类的鲁棒性。

[0076] 在一个实施例中,当多模态的数据包括属于视频文本模态的描述文本时,本实施例中,步骤S206包括:编码对描述文本进行分词后的词片段,生成各词片段相应的隐含状态向量;获取各隐含状态向量对应的注意力权重;根据各隐含状态向量和相应注意力权重,得

到最终隐含状态向量;解码最终隐含状态向量,得到对应于视频文本模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0077] 其中,描述文本,是描述目标视频的相关文本内容。在一个实施例中,描述文本包括视频的标题、简介和用户提供的视频标签等中的至少一种。其中,视频标签,是用户在上传视频时,针对视频添加的自定义标注。

[0078] 具体地,计算机设备可以直接获取对描述文本进行分词处理后得到的词片段,也可以对描述文本进行分词处理,进而得到词片段。计算机设备可以对词片段进行编码,生成与各词片段相应的隐含状态向量。

[0079] 可以理解,隐含状态向量,是由神经网络中的隐藏层编码输出的向量。隐藏层的意义就是把输入数据的特征,抽象到另一个维度空间,来展现其更抽象化的特征。所以,隐含状态向量是能够让输入数据的特征能更好的进行线性划分的一种数据。可以理解,隐含状态向量不直接对用户进行输出显示,而是神经网络处理过程中的一个中间态数据。

[0080] 计算机设备可以根据注意力模型,获取各隐含状态向量对应的注意力权重,计算机设备可以根据各隐含状态向量和相应注意力权重,得到最终隐含状态向量,得到最终隐含状态向量。计算机设备可以解码最终隐含状态向量,得到对应于视频文本模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0081] 其中,注意力模型(Attention Model),是用于从隐含状态向量中,提取出对分类预测更关键的隐含状态向量的机器学习模型。可以理解,注意力模型可以根据各隐含状态向量对分类预测的关键影响程度为各隐含状态向量分配相应的注意力权重。

[0082] 在一个实施例中,计算机设备可以按照以下公式计算最终隐含状态向量:

$$[0083] \quad h_{att} = \sum_t a_t \tanh(W_w h_t + b_w); \quad (\text{公式1})$$

[0084] 其中, h_{att} 为最终隐含状态向量; a_t 为第t个编码单元输出的隐含状态向量的注意力权重; h_t 为第t个编码单元输出的隐含状态向量; $\tanh()$ 为双曲正切函数, W_w 和 b_w 分别为双曲正切函数 $\tanh()$ 中的参数。

[0085] 在一个实施例中,计算机设备可以对最终隐含状态向量进行归一化处理,得到对应于视频文本模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0086] 在一个实施例中,计算机设备可以按照以下公式来计算对应于文本状态的概率向量:

$$[0087] \quad \text{pre} = \text{softmax}(W_{att} h_{att} + b_{att}); \quad (\text{公式2})$$

[0088] 其中,pre表示预测得到的概率向量;softmax()表示归一化指数函数; h_{att} 为最终隐含状态向量; W_{att} 和 b_{att} 分别为归一化指数函数中的参数。

[0089] 上述实施例中,将对描述文本进行分词后的词片段编码为隐含状态向量,更有利于进行线性划分。然后结合注意力模型,将各隐含状态向量和相应注意力权重,得到最终隐含状态向量。可以理解,结合注意力模型,可以在利用全局信息的同时又可以专注于局部比较有用的信息,最终隐含状态向量中相当于综合了单个的隐含状态向量中更加有用的特征信息,因此,对最终隐含状态向量解码生成的概率向量更加的准确。

[0090] 在一个实施例中,编码对描述文本进行分词后的词片段,生成各词片段相应的隐含状态向量包括:将对描述文本进行分词后的词片段编码为词向量,得到词向量序列;按照

词向量序列中各词向量的先后顺序,对各词向量依次进行编码,在每次编码时,根据当前词向量和对前一词向量编码输出的隐含状态向量进行编码,生成当前词向量相应的隐含状态向量。

[0091] 具体地,计算机设备可以将各词片段分别编码为相应的词向量,根据各词向量得到词向量序列。计算机设备可以按照词向量序列中各词向量的先后顺序,对各词向量依次进行编码。比如,有3个词片段,分别编码得到的词向量为 $x_1 \sim x_3$,则可以得到词向量序列为 (x_1, x_2, x_3) ,计算机设备则可以依次对 x_1 编码,在对 x_1 编码完毕后,会对 x_2 进行编码,在 x_2 编码完毕后接着对 x_3 进行编码。

[0092] 可以理解,在每次编码时,计算机设备可以对当前词向量编码输出相应的隐含状态向量。具体地,在每次编码时,计算机设备可以根据当前词向量和对前一词向量编码输出的隐含状态向量进行编码,生成当前词向量相应的隐含状态向量。需要说明的是,在对首位词向量进行编码时,不存在对前一词向量编码输出的隐含状态向量,所以可以使用默认值来充当对前一词向量编码输出的隐含状态向量。

[0093] 同样以词向量序列 (x_1, x_2, x_3) 为例,计算机设备可以对 x_1 进行编码输出相应隐含状态向量 h_1 ,那么,在对 x_2 进行编码时,计算机设备可以根据对前一词向量 x_1 编码输出的隐含状态向量 h_1 和当前词向量 x_2 进行编码,输出 x_2 相应的隐含状态向量 h_2 。在对 x_3 进行编码时,计算机设备可以根据对前一词向量 x_2 编码输出的隐含状态向量 h_2 和当前词向量 x_3 进行编码,输出 x_3 相应的隐含状态向量 h_3 。

[0094] 在一个实施例中,计算机设备可以将对描述文本进行分词后的词片段编码为词向量,得到词向量序列;将词向量序列输入循环神经网络或者长短期记忆神经网络中,按照词向量序列中各词向量的先后顺序,将各词向量依次输入各编码单元中进行编码,在每次编码时,通过当前编码单元根据当前词向量和对前一词向量编码输出的隐含状态向量进行编码,生成当前词向量相应的隐含状态向量。

[0095] 图4为一个实施例中的文本分类模型的架构示意图。参照图4可知,文本分类模型可以为包括注意力模型和长短期记忆神经网络模型的组合模型。图4中, $x_1 \sim x_n$ 为对 n 个词片段进行编码输出的词向量,计算机设备可以将词向量 $x_1 \sim x_n$ 输入LSMT单元(即LSTM模型中的编码单元)中,各LSMT单元可以输出相应的隐含状态向量,分别得到隐含状态向量 $h_1 \sim h_n$ 。除了第一个LSMT单元以外,其余的LSMT单元在编码时,都会将当前词向量和前一编码单元输出的隐含状态向量输入至当前编码单元中进行编码。比如,计算机设备可以将第一个LSMT单元输出的隐含状态向量 h_1 和当前词向量 x_2 输入至第2个LSMT单元中进行编码,由第2个LSMT单元输出隐含状态向量 h_2 。然后,将在词向量序列中排序第3的词向量 x_3 作为当前词向量,将当前词向量 x_3 和由第2个LSMT单元输出隐含状态向量 h_2 输入至第3个LSMT单元中编码输出隐含状态向量 h_3 。依次类推,将当前词向量 x_n 和由第 $n-1$ 个LSMT单元输出隐含状态向量 h_{n-1} 输入至第 n 个LSMT单元中编码输出隐含状态向量 h_n 。图4中的 $a_1 \sim a_n$ 分别为隐含状态向量 $h_1 \sim h_n$ 的注意力权重,计算机设备可以将隐含状态向量 $h_1 \sim h_n$ 按照相应的注意力权重进行加权求和,得到最终隐含状态向量 h_{att} 。计算机设备可以解码最终隐含状态向量 h_{att} 进行预测,预测得到对应于视频文本模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0096] 上述实施例中,在对当前词向量进行编码时,考虑了对前一词向量编码输出的隐

含状态向量,相当于在编码当前词向量时考虑了历史信息,从而提高了编码准确性。此外,结合了注意力机制,进一步地提高了编码准确性。从而,能够解码得到更为准确地视频分类结果。

[0097] 在一个实施例中,在每次编码时,根据当前词向量和对前一词向量编码输出的隐含状态向量进行编码,生成当前词向量相应的隐含状态向量包括:在每次编码时,将当前词向量和前一编码单元输出的隐含状态向量输入文本分类模型中的当前编码单元,生成当前编码单元的临时单元状态向量、临时单元状态向量相应的输入权重、上一单元状态向量的遗忘权重和当前单元状态向量的输出权重;将上一单元状态向量按照遗忘权重进行保留,得到第一向量;第一向量用于表征上一单元状态需保留信息;将临时单元状态向量按照输入权重进行保留,得到第二向量;第二向量用于表征当前词向量需保留信息;根据第一向量和第二向量,得到当前编码单元的当前单元状态向量;将当前单元状态向量按照输出权重进行输出,得到当前词向量相应的隐含状态向量。

[0098] 其中,文本分类模型可以是对文本进行分类预测的模型。文本分类模型可以是单独的分类模型,也可以是由多个模型组合而成的分类模型。编码单元,是用于实现编码处理的神经元。可以理解,文本分类模型中可以包括多个编码单元。

[0099] 在一个实施例中,文本分类模型可以是基于注意力机制的长短时记忆神经网络模型(Long Short Term Memory Network,LSTM),即注意力模型和长短时记忆神经网络模型组合得到的模型。

[0100] 可以理解,计算机设备可以按照词向量序列中各词向量的先后顺序,将各词向量依次输入文本分类模型中的编码单元进行编码。在每次编码时,计算机设备可以通过编码单元对词向量进行编码,输出相应的隐含状态向量。

[0101] 具体地,在每次编码时,计算机设备可以将当前词向量和前一编码单元输出的隐含状态向量输入文本分类模型中的当前编码单元,通过用于表征该当前编码单元中的遗忘门的激活函数,对当前词向量和前一编码单元输出的隐含状态向量进行编码,映射生成上一单元状态向量的遗忘权重。

[0102] 其中,遗忘门,用于决定从上一单元状态向量中丢弃多少信息。所以,遗忘权重,用于表示上一单元状态向量不被遗忘的(即能够保留的)权重。遗忘权重实质上可以为权重矩阵。

[0103] 可以理解,计算机设备可以通过用于表征遗忘门的激活函数,对当前词向量和前一编码单元输出的隐含状态向量进行编码,映射到0到1之间的数值,得到上一单元状态向量的遗忘权重。其中,0表示完全舍弃,1表示完全接受(即完全保留)。

[0104] 在一个实施例中,计算机设备可以按照以下公式计算得到遗忘权重:

[0105] $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$; (公式3)

[0106] 其中, f_t 即为第t个编码单元生成的针对上一单元状态向量的遗忘权重; h_{t-1} 表示的是第t-1个编码单元输出的隐含状态向量, x_t 表示的是当前词向量, σ 表示激活函数sigmoid函数, W_f 和 b_f 表示sigmoid函数的参数。 $[h_{t-1}, x_t]$ 表示将 h_{t-1} 和 x_t 进行组合。

[0107] 因此,基于公式(3)可以得知,计算机设备可以通过当前编码单元将当前词向量和前一编码单元输出的隐含状态向量进行组合,然后将组合后的向量输入激活函数中映射生成针对上一单元状态向量的遗忘权重。

[0108] 在每次编码时,计算机设备还可以通过用于表征该当前编码单元中的输入门的激活函数,对当前词向量和前一编码单元输出的隐含状态向量进行编码,生成当前编码单元的临时单元状态向量以及该临时单元状态向量相应的输入权重。

[0109] 其中,临时单元状态向量相应的输入权重,用于决定让多少新的信息加入到当前单元状态中来。

[0110] 计算机设备将所述临时单元状态向量按照所述输入权重进行保留,得到第二向量;所述第二向量用于表征当前词向量需保留信息;

[0111] 在一个实施例中,计算机设备可以按照以下公式来计算临时单元状态向量相应的输入权重:

[0112] $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$; (公式4)

[0113] 其中, i_t 表示第t个编码单元的临时单元状态向量的输入权重; h_{t-1} 表示的是第t-1个编码单元输出的隐含状态向量, x_t 表示的是当前词向量; σ 表示激活函数sigmoid函数, W_i 和 b_i 分别表示sigmoid函数的参数。

[0114] 在一个实施例中,计算机设备可以按照以下公式来计算当前编码单元的临时单元状态向量:

[0115] $\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$; (公式4)

[0116] 其中, $\tilde{C}_t$ 即为当前编码单元的临时单元状态向量; h_{t-1} 表示的是第t-1个编码单元输出的隐含状态向量, x_t 表示的是当前词向量; $\tanh$ 表示正曲双切函数, W_c 和 b_c 分别表示tanh函数的参数。

[0117] 此外,在每次编码时,计算机设备还可以通过用于表征该当前编码单元中的输出门的激活函数,对当前词向量和前一编码单元输出的隐含状态向量进行编码,生成当前单元状态向量的输出权重。

[0118] 其中,当前单元状态向量的输出权重,用于决定将当前编码单元的当前单元状态向量的哪些信息进行输出。

[0119] 在一个实施例中,计算机设备可以按照以下公式计算当前单元状态向量的输出权重:

[0120] $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$; (公式6)

[0121] 其中, o_t 表示第t个编码单元的当前单元状态向量的输出权重; W_o 和 b_o 分别表示(公式6)中sigmoid函数的参数; h_{t-1} 表示的是第t-1个编码单元输出的隐含状态向量, x_t 表示的是当前词向量; σ 表示激活函数sigmoid函数。

[0122] 进一步地,计算机设备可以将上一单元状态向量按照遗忘权重进行保留,得到第一向量。其中,第一向量用于表征上一单元状态需保留信息。计算机设备可以将所述临时单元状态向量按照所述输入权重进行保留,得到第二向量;所述第二向量用于表征当前词向量需保留信息。计算机设备可以将第一向量和第二向量相加,得到所述当前编码单元的当前单元状态向量。

[0123] 在一个实施例中,计算机设备可以按照以下公式计算当前单元状态向量:

[0124] $C_t = f_t * C_{t-1} + i_t * \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$; (公式7)

[0125] 其中, C_t 表示当前第t个编码单元的单元状态向量; C_{t-1} 表示第t-1个编码单元的单

元状态向量; f_t 即为第 t 个编码单元生成的针对上一单元状态向量 C_{t-1} 的遗忘权重; i_t 表示第 t 个编码单元的临时单元状态向量的输入权重; $\tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$ 即为当前编码单元的临时单元状态向量; h_{t-1} 表示的是第 $t-1$ 个编码单元输出的隐含状态向量, x_t 表示的是当前词向量; $\tanh$ 表示正曲双切函数; W_c 和 b_c 分别表示 $\tanh$ 函数的参数。

[0126] 计算机设备可以将当前单元状态向量按照输出权重进行输出,得到当前词向量相应的隐含状态向量。

[0127] 在一个实施例中,计算机设备可以按照以下公式来计算当前词向量相应的隐含状态向量:

[0128] $h_t = o_t * \tanh(C_t)$; (公式8)

[0129] 其中, h_t 为第 t 个编码单元输出的隐含状态向量; $\tanh()$ 为双曲正切函数, C_t 表示当前第 t 个编码单元的单元状态向量。

[0130] 上述实施例中,在每次编码时,所确定的遗忘权重、输入权重以及输出权重,就相当于一些阀门,能够控制记忆和存储一些长期信息,因而,能够使得编码依赖更多的长期历史信息,从而提高了编码准确性。进而,能够解码得到更为准确地视频分类结果。

[0131] 在一个实施例中,当多模态的数据包括从目标视频中提取的属于视频帧模态的视频帧时,步骤S206包括:分别卷积处理各视频帧,得到与各视频帧相应的视频帧特征向量;对各视频帧特征向量进行平均池化处理;对平均池化处理后的向量进行归一化处理,生成对应于视频帧模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0132] 其中,视频帧特征向量,是用于表征视频帧的特征的向量。

[0133] 具体地,计算机设备可以将所提取的各视频帧分别输入至卷积神经网络模型中进行卷积处理,得到与各视频帧相应的视频帧特征向量。

[0134] 在一个实施例中,计算机设备可以对提取的各视频帧进行预处理,以将各视频帧的大小调整一致。然后将调整一致后的各视频帧分别输入至卷积神经网络模型中进行卷积处理,得到与各视频帧相应的视频帧特征向量。可以理解,针对大小调整一致的视频帧进行卷积处理,能够提高卷积处理的准确性。

[0135] 在一个实施例中,卷积神经网络模型可以是以Inception-ResNet-v2(Google公司研发的一种深度卷积神经网络)模型为基础构架的卷积神经网络模型。本实施例中,分别卷积处理各视频帧,得到与各视频帧相应的视频帧特征向量包括:将每个视频帧输入卷积神经网络模型中进行多层卷积处理,在每层卷积处理时,将所述视频帧和所述视频帧的前一层卷积结果作为当层卷积处理的输入,进行联合卷积处理,得到当层卷积结果;将最后一层卷积结果,作为所述视频帧的视频帧特征向量。

[0136] 需要说明的是,当进行首层卷积处理时,不存在该视频帧的前一层卷积结果,所以,首层卷积处理时的输入仅为该视频帧本身,从第二层卷积处理起,将所述视频帧和所述视频帧的前一层卷积结果作为当层卷积处理的输入。

[0137] 可以理解,Inception-ResNet-v2比先前的Inception V3(Google公司在Inception-ResNet-v2之前研发的一种深度卷积神经网络)还要深一些,在模型中使用了重复的残差区块。另外简化了inception结构,比先前的Inception V3种要包含更少的并行塔,Inception-ResNet-v2架构的精确度比之前的最优模型更高。

[0138] 为了便于理解,现举例说明。假如有5个视频帧 $frame_1 \sim frame_5$,以 $frame_1$ 为例,计

计算机设备可以将 frame_1 输入至卷积神经网络模型中进行3层卷积处理。具体地,计算机设备可以将 frame_1 输入至第一层卷积层进行卷积,得到第一层卷积结果 F_1 ,然后将 frame_1 和第一层卷积结果 F_1 一起输入至第2层卷积层进行卷积,得到第2层卷积结果 F_2 ,然后将 frame_1 和第2层卷积结果 F_2 一起输入至第3层卷积层进行卷积,得到第3层卷积结果 F_3 。计算机设备则将第3层卷积结果 F_3 进行输出,作为 frame_1 的视频帧特征向量。

[0139] 计算机设备可以对各视频帧特征向量中同一位置的向量元素求平均,即相当于进行平均池化处理,根据求平均后得到的各向量元素构成平均池化处理后的向量。计算机设备可以对平均池化处理后的向量进行归一化处理,即,将平均池化处理后的向量中各向量元素映射为取值在0到1之间的数值。可以理解,归一化处理后的各向量元素即用于表征目标视频属于各预设类别的概率。进而,计算机设备可以根据归一化后的向量元素,生成对应于视频帧模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0140] 在一个实施例中,计算机设备可以通过归一化指数函数(softmax函数)对平均池化处理后的向量进行归一化处理。

[0141] 在一个实施例中,使用的卷积神经网络结构如表1中所示,其中,以conv命名的为卷积层,以pool命名的为池化层,以fc命名的为全连接层:

[0142] 表1

神经网络 层名	神经元个数或 者特征图数	卷积核大小或 者池化大小	步 长
conv1	64	3*3	1
pool1		2*2	2
conv2	128	3*3	1
pool2		2*2	2
conv3_1	256	3*3	1
conv3_2	256	3*3	1
pool3		2*2	2
conv4_1	512	3*3	1
conv4_2	512	3*3	1
pool4		2*2	2
fc1_1	4096		
fc1_2	4096		
fc_2	128		

[0145] 需要说明的是,表1中的卷积神经网络结构是经过实验得到的对音频分类效果起到显著作用的结构。

[0146] 图5为一个实施例中视频帧分类模型的架构示意图。参照图5,从目标视频video中提取多个视频帧 $\text{frame}_1 \sim \text{frame}_n$,然后分别将 $\text{frame}_1 \sim \text{frame}_n$ 输入至卷积神经网络模型中进行卷积处理,得到各个视频帧对应的视频帧特征向量,然后将所有的视频帧特征向量输入平均池化层中进行平均池化处理,再通过softmax函数对平均池化后的向量进行归一化处理,从而得到对应于视频帧模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0147] 上述实施例中,分别对各视频帧卷积提取相应的视频帧特征向量,然后对各视频帧特征向量进行平均池化处理,相当于将各视频帧的特征进行了组合。因此,基于组合后的特征生成概率向量更加的准确。此外,在特征层面即实现了组合,而并不用一一对各视频帧进行复杂的分类处理,提高了视频分类预测的效率。

[0148] 在一个实施例中,当多模态的数据包括属于音频模态的音频文件时,本实施例中步骤S206包括:将音频文件进行切割,得到音频片段;提取各音频片段的频率特征;分别对各频率特征进行卷积处理,得到相应的音频特征向量;根据各音频特征向量进行分类预测,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0149] 在一个实施例中,计算机设备可以对音频文件进行预处理,以使得音频文件中的音频信号的采样率相同,然后对预处理后的音频文件进行切割,得到音频片段。

[0150] 在一个实施例中,将音频文件进行切割,得到音频片段包括:将音频文件中的多通道转换为单通道;按照预设采样率,对转换后的音频文件中的音频信号进行重采样;将重采样后的音频文件切割为音频片段。

[0151] 其中,通道,是音频中播放声音的通道。可以理解,一个音频文件可能有很多种声音通道,比如,背景音通道和说话主声通道等。多通道是有多维度的,如果针对多通道进行特征提取,就需要分别针对每个通道皆进行处理,非常的繁琐。因此,计算机设备将音频文件中的多通道转换为单通道,使其变为单一维度,这样就可以大大降低特征提取的复杂性。

[0152] 采样率,是指每秒从连续信号中提取并组成离散信号的采样个数。重采样,是指将原始的采样频率变换为新的采样频率以适应不同采样率的要求。

[0153] 可以理解,通过对转换为单通道的音频文件中的音频信号进行重采样,使得音频文件中的音频信号更加的统一,进而将重采样后的音频文件切割为音频片段,便于后续统一化的特征提取处理。

[0154] 在一个实施例中,计算机设备可以将音频文件切割为等长的音频片段。可以理解,都能够等长分割时最后一个音频片段的时间不够的话可以向前补充。比如,有58秒的音频文件,按照10秒的时间间隔,切分为6等分,最后一个音频片段为第51~58秒,不够10秒,则可以向前补充2秒,即取第49秒~58秒的片段,作为最后一个音频片段。可以理解,等长的音频片段能够保证处理数据的格式一致性,提高了后续处理的准确性。

[0155] 计算机设备可以分别提取各音频片段的频率特征。在一个实施例中,计算机设备可以针对各音频片段计算梅尔频率倒谱系数(MFCC, Mel Frequency Cepstrum Coefficient),将计算得到的梅尔频率倒谱系数作为相应音频片段的频率特征。

[0156] 在一个实施例中,计算机设备可以将音频片段中线形的自然频谱转换为体现人类

听觉特性的梅尔频谱;在梅尔频谱上面进行倒谱分析,得到各音频片段的梅尔频率倒谱系数。

[0157] 在一个实施例中,计算机设备可以分别将各音频片段的频率特征分别输入至卷积神经网络模型中进行卷积处理,得到相应的音频特征向量。计算机设备可以根据各音频特征向量进行分类预测,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0158] 在一个实施例中,根据各音频特征向量进行分类预测,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量包括:按照卷积后的音频特征向量的先后顺序,循环地将当前音频特征向量和前次编码后的编码向量作为当次编码的输入;对最后一次编码后的编码向量进行归一化处理,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0159] 具体地,计算机设备可以按照卷积后的音频特征向量的先后顺序,循环地将当前音频特征向量和前次编码后的向量输入当前编码器中,生成当次编码后的编码向量。以此,计算机设备可以循环迭代地针对每个音频特征向量进行编码,生成相应的编码向量。计算机设备可以对最后一次编码后的编码向量进行归一化处理,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。可以理解,这样一来,可以在编码时可以参考历史信息,提高了编码准确性。

[0160] 图6为一个实施例中音频分类模型的架构示意图。参照图6可知,音频分类模型可以为包括卷积神经网络模型CNN(Convolutional Neural Networks)和长短期记忆神经网络模型LSTM的组合模型。图6中,计算机设备可以对音频文件audio进行切割,生成均匀长度的音频片段 $\text{audio}_1 \sim \text{audio}_n$ 。分别提取 $\text{audio}_1 \sim \text{audio}_n$ 的频率特征1~频率特征n,将频率特征1~频率特征n输入至卷积神经网络模型CNN中进行卷积处理,得到频率特征1~频率特征n分别对应的音频特征向量 $\text{feature}_1 \sim \text{feature}_n$ 。计算机设备可以将第一个音频特征向量 feature_1 输入至第一个LSMT单元(即编码单元)中,生成相应的编码向量 E_1 ,然后将该编码向量 E_1 和第二个音频特征向量 feature_2 输入至第二个LSMT单元中,生成第二个编码向量 E_2 ,并将第二个编码向量 E_2 和第三个音频特征向量 feature_3 输入至第三个LSMT单元中进行编码,以此类推,直至将第n-1个编码向量 E_{n-1} 和第n个音频特征向量 feature_n 输入至第n个LSMT单元中进行编码,生成第n个编码向量 E_n 。计算机设备可以针对编码向量 E_n 进行归一化处理,生成对应于音频模态的概率向量。

[0161] 上述实施例中,通过将音频文件分割为音频片段进行特征提取,降低了数据处理复杂度,而且提高了特征提取的准确性。而且,频率特征能够很准确地反映音频片的特征,所以对各音频片的音频特征向量进行卷积处理,根据卷积处理得到的音频特征向量进行分类预测,提高了分类预测的准确性。此外,在分类预测时,充分考虑能够充分表征音频文件的特征的各音频片的音频特征向量,进一步地提高了分类预测的准确性。

[0162] 在一个实施例中,步骤S210根据组合后得到的向量,预测目标视频最终属于的类别包括:将组合后得到的向量分别输入预先训练的梯度提升回归树模型的各分类回归树中进行分类处理,得到各分类回归树所预测的用于表征目标视频属于各预设类别的概率的子概率向量;获取每棵回归树相应的决策权重;将各棵回归树所预测的子概率向量按照相应的决策权重进行加权平均,得到用于表征目标视频属于各预设类别的概率的最终概率向

量;选取最终概率向量中的最大概率值所对应的预设类别,得到目标视频最终属于的类别。

[0163] 具体地,计算机设备可以预先根据样本数据进行机器学习训练,得到梯度提升回归树模型。计算机设备可以将组合后得到的向量分别输入预先训练的梯度提升回归树模型的各分类回归树中进行分类处理,得到各分类回归树所预测的用于表征目标视频属于各预设类别的概率的子概率向量。计算机设备可以获取每棵回归树相应的决策权重;将各棵回归树所预测的子概率向量按照相应的决策权重进行加权平均,得到用于表征目标视频属于各预设类别的概率的最终概率向量。

[0164] 在一个实施例中,计算机设备可以按照以下公式计算最终概率向量:

$$[0165] \quad F(x; w) = \sum_{t=0}^T \alpha_t h_t(x; w_t) \quad ; \quad (\text{公式 } 9)$$

[0166] 其中, $F(x; w)$ 为最终概率向量; x 为组合后得到的向量, w 用于统一表征分类回归树对应的参数; h_t 为第 t 棵分类回归树预测的子概率向量, w_t 为第 t 棵分类回归树对应的参数, α_t 为第 t 棵分类回归树对应的决策权重。

[0167] 计算机设备可以从最终概率向量中选取最大概率值,获取该最大概率值所对应的预设类别,作为目标视频最终属于的类别。

[0168] 在一个实施例中,计算机设备可以将梯度提升回归树模型的训练过程和多模态对应的分类模型的训练过程结合在一起,使用一批样本数据同时训练出视多模态对应的分类模型以及梯度提升回归树模型,样本数据中包括样本视频的多模态的样本数据,以及样本视频的类别标签。那么,在迭代训练过程中的输入为样本数据中样本是多模态的样本数据,输出为样本视频的类别标签,在迭代的机器学习训练过程中,不断地调整多模态对应的分类模型以及梯度提升回归树模型中的模型参数,直至达到迭代停止条件(比如,迭代次数达到预设次数或模型参数稳定),从而训练出端到端的对视频进行分类的组合模型。

[0169] 在其他实施例中,计算机设备也可以将梯度提升回归树模型的训练过程和多模态对应的分类模型的训练过程分开。这样一来,计算机设备则可以根据第一批样本数据训练多模态对应的分类模型。然后,使用第二批样本数据作为训练好的多模态对应的分类模型的输入,通过各模态对应的分类模型分别对第二批样本数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量。然后,计算机设备可以将对应于各模态的概率向量进行组合,将组合后得到的向量以及第二批样本数据中携带的样本视频的视频标签作为梯度提升回归树模型的训练样本,迭代地对梯度提升回归树模型进行机器学习训练,直至得到训练好的梯度提升回归树模型。需要说明的是,第一批样本数据和第二批样本数据中皆包括样本视频的多模态的样本数据,以及样本视频的类别标签。可以理解,这种将梯度提升回归树模型的训练过程和多模态对应的分类模型的训练过程分开的方式,能够降低对机器硬件的要求,同时也能够保证分类准确性。

[0170] 上述实施例中,使用梯度提升回归树模型这一强分类器,将组合后得到的向量输入这一强分类器中的各个弱分类器,即各分类回归树,进行分类处理,综合考虑各分类回归树预测得到的子概率向量,得到最终概率向量。该最终概率向量相较于单一的分类器而言,提高了分类概率预测的准确性。进而,选取最终概率向量中的最大概率值所对应的预设类别,得到目标视频最终属于的类别,提高了视频分类的准确性。

[0171] 在一个实施例中,目标视频是内容推荐平台中的视频。该方法还包括:获取内容推

荐平台中的用户对象所偏好的视频类别;将目标视频最终属于的类别与各用户对象所偏好的视频类别进行匹配;向匹配成功的用户对象相应的终端推荐目标视频。

[0172] 其中,内容推荐平台,是对内容进行的推荐的互联网平台。

[0173] 在一个实施例中,内容推荐平台可以包括新闻媒体平台、内容搜索平台和社交网络平台等中的至少一种。社交网络平台包括即时通信平台和内容分享平台(比如,微博)。

[0174] 其中,新闻媒体平台是以新闻的形式推荐内容的平台。其中,新闻媒体平台可以包括独立的新闻媒体平台。比如,独立的腾讯(Tencent)新闻平台。

[0175] 新闻媒体平台还可以包括集成于即时通信平台中的新闻媒体平台。比如,集成于微信平台中的腾讯新闻平台。其中,微信(WeChat),是腾讯公司于2011年1月21日推出的一个为智能终端提供即时通讯服务的应用程序。

[0176] 内容搜索平台,是对内容进行搜索的平台。比如,浏览器平台即为内容搜索平台。其中,内容搜索平台可以包括独立的浏览器平台和集成于即时通信平台中的浏览器平台。

[0177] 具体地,计算机设备在预测出目标视频最终属于的类别后,可以获取内容推荐平台中的用户对象的用户信息,从用户信息中提取用户对象所偏好的视频类别。计算机设备可以将目标视频最终属于的类别与各用户对象所偏好的视频类别进行匹配。计算机设备可以向匹配成功的用户对象相应的终端推荐该目标视频。

[0178] 图7为一个实施例中视频推荐的流程简化示意图。图7中是以短视频为例进行举例说明。参照图7,计算机设备从短视频中分别提取属于视频帧模态的视频帧、属于视频文本模态的短视频文本(即描述文本)和音频文件,通过执行本申请各实施例中的视频分类处理方法的短视频分类系统,得到短视频的类别属性。并获取用户的用户属性,将短视频的类别属性和用户属性一并输入至推荐系统,由推荐系统从该用户属性中提取用户所偏好的视频类别,将短视频的类别属性和用户所偏好的视频类别进行匹配,得到推荐短视频结果。需要说明的是,推荐系统可以是运行于服务器的具备内容推荐功能的系统。

[0179] 上述实施例中,基于本申请实施例中的视频分类处理方法所得到的目标视频的类别,更加的准确,因而基于更为准确的类别对视频进行推荐,提高了推荐的准确性。

[0180] 需要说明的是,本申请各实施例中的视频分类处理方法并不局限于视频推荐这么一个使用场景,在其他实施例中,还可以按照分类类别对视频进行统一管理。比如,计算机设备可以对服务器中存储的各视频均执行本申请各实施例中的视频分类处理方法,进而得到各视频的分类类别,计算机设备可以按照各视频的分类类别进行分组。进一步地,计算机设备可以根据分组情况进行大数据分析,计算机设备可以根据相应的大数据分析结果,触发相应的处理。比如,分析哪个组别中的视频数量最多,或者分析哪个组别的视频被访问频率最高,从而得到哪些类别的视频为热门视频,然后可以向各用户推荐热门视频。

[0181] 通过将通过本申请各实施例中的视频分类处理方法所预测的视频的分类结果进行随机抽取进而进行人工评测发现,本申请各实施例中的视频分类处理方法相比现有方法在视频分类的准确率、召回率上都有明显的提升。此外,在一些内容推荐平台的视频推荐策略上,用户的点击率也明显得到了提高,相当于提高了推荐准确性。

[0182] 如图8所示,在一个实施例中,提供了一种视频分类处理装置800,该装置800包括:提取模块802、分类预测模块804以及组合模块806,其中:

[0183] 提取模块802,用于获取目标视频;提取所述目标视频的多模态的数据。

[0184] 分类预测模块804,用于对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0185] 组合模块806,用于将对应于各模态的概率向量进行组合;

[0186] 所述分类预测模块804还用于根据组合后得到的向量,预测所述目标视频最终属于的类别。

[0187] 在一个实施例中,所述多模态的数据包括从目标视频中提取的属于视频帧模态的视频帧、属于视频文本模态的描述文本、以及属于音频模态的音频文件中的至少两种。

[0188] 在一个实施例中,所述多模态的数据包括属于视频文本模态的描述文本;所述分类预测模块804还用于编码对描述文本进行分词后的词片段,生成各词片段相应的隐含状态向量;获取各隐含状态向量对应的注意力权重;根据各隐含状态向量和相应注意力权重,得到最终隐含状态向量;解码所述最终隐含状态向量,得到对应于视频文本模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0189] 在一个实施例中,所述分类预测模块804还用于将对描述文本进行分词后的词片段编码为词向量,得到词向量序列;按照词向量序列中各词向量的先后顺序,对各词向量依次进行编码,在每次编码时,根据当前词向量和对前一词向量编码输出的隐含状态向量进行编码,生成所述当前词向量相应的隐含状态向量。

[0190] 在一个实施例中,所述分类预测模块804还用于在每次编码时,将当前词向量和前一编码单元输出的隐含状态向量输入文本分类模型中的当前编码单元,生成当前编码单元的临时单元状态向量、所述临时单元状态向量相应的输入权重、上一单元状态向量的遗忘权重和当前单元状态向量的输出权重;将所述上一单元状态向量按照所述遗忘权重进行保留,得到第一向量;所述第一向量用于表征上一单元状态需保留信息;将所述临时单元状态向量按照所述输入权重进行保留,得到第二向量;所述第二向量用于表征当前词向量需保留信息;根据第一向量和第二向量,得到所述当前编码单元的当前单元状态向量;将所述当前单元状态向量按照所述输出权重进行输出,得到当前词向量相应的隐含状态向量。

[0191] 在一个实施例中,所述多模态的数据包括从目标视频中提取的属于视频帧模态的视频帧;所述分类预测模块804还用于分别卷积处理各视频帧,得到与各视频帧相应的视频帧特征向量;对各视频帧特征向量进行平均池化处理;对平均池化处理后的向量进行归一化处理,生成对应于所述视频帧模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0192] 在一个实施例中,所述分类预测模块804还用于将每个视频帧输入卷积神经网络模型中进行多层卷积处理,在每层卷积处理时,将所述视频帧和所述视频帧的前一层卷积结果作为当层卷积处理的输入,进行联合卷积处理,得到当层卷积结果;将最后一层卷积结果,作为所述视频帧的视频帧特征向量。

[0193] 在一个实施例中,所述多模态的数据包括属于音频模态的音频文件;所述分类预测模块804还用于将所述音频文件进行切割,得到音频片段;提取各所述音频片段的频率特征;分别对各所述频率特征进行卷积处理,得到相应的音频特征向量;根据各所述音频特征向量进行分类预测,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0194] 在一个实施例中,所述分类预测模块804还用于按照卷积后的音频特征向量的先

后顺序,循环地将当前音频特征向量和前次编码后的编码向量作为当次编码的输入;对最后一次编码后的编码向量进行归一化处理,得到对应于音频模态的、且用于表征目标视频属于各预设类别的概率的概率向量。

[0195] 在一个实施例中,所述分类预测模块804还用于将音频文件中的多通道转换为单通道;按照预设采样率,对转换后的音频文件中的音频信号进行重采样;将重采样后的音频文件切割为音频片段。

[0196] 在一个实施例中,所述分类预测模块804还用于将组合后得到的向量分别输入预先训练的梯度提升回归树模型的各分类回归树中进行分类处理,得到各分类回归树所预测的用于表征目标视频属于各预设类别的概率的子概率向量;获取每棵回归树相应的决策权重;将各棵回归树所预测的子概率向量按照相应的决策权重进行加权平均,得到用于表征目标视频属于各预设类别的概率的最终概率向量;选取所述最终概率向量中的最大概率值所对应的预设类别,得到所述目标视频最终属于的类别。

[0197] 在一个实施例中,所述目标视频是内容推荐平台中的视频。

[0198] 本实施例中,如图9所示,该装置800还包括:

[0199] 推荐模块808,用于获取所述内容推荐平台中的用户对象所偏好的视频类别;将目标视频最终属于的类别与各用户对象所偏好的视频类别进行匹配;向匹配成功的用户对象相应的终端推荐所述目标视频。

[0200] 图10为一个实施例中计算机设备的内部结构示意图。参照图10,该计算机设备可以是图1中终端110或服务器120。该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。其中,存储器包括非易失性存储介质和内存存储器。该计算机设备的非易失性存储介质可存储操作系统和计算机程序。该计算机程序被执行时,可使得处理器执行一种视频分类处理方法。该计算机设备的处理器用于提供计算和控制能力,支撑整个计算机设备的运行。该内存存储器中可储存有计算机程序,该计算机程序被处理器执行时,可使得处理器执行一种视频分类处理方法。计算机设备的网络接口用于进行网络通信。

[0201] 本领域技术人员可以理解,图10中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0202] 在一个实施例中,本申请提供的异常检测装置可以实现为一种计算机程序的形式,计算机程序可在如图10所示的计算机设备上运行,计算机设备的非易失性存储介质可存储组成该异常检测装置的各个程序模块,比如,图8所示的提取模块802、分类预测模块804以及组合模块806。各个程序模块所组成的计算机程序用于使该计算机设备执行本说明书中描述的本申请各个实施例的视频分类处理方法中的步骤,例如,计算机设备可以通过如图8所示的视频分类处理装置800中的提取模块802获取目标视频;提取所述目标视频的多模态的数据。计算机设备可以通过分类预测模块804对各模态的数据进行分类预测,得到分别对应于各模态的、且用于表征目标视频属于各预设类别的概率的概率向量。计算机设备可以通过组合模块806将对应于各模态的概率向量进行组合。计算机设备可以通过分类预测模块804根据组合后得到的向量,预测所述目标视频最终属于的类别。

[0203] 在一个实施例中,提供了一种计算机设备,包括存储器和处理器,存储器存储有计算机程序,计算机程序被处理器执行时,使得处理器执行上述视频分类处理方法的步骤。此

处视频分类处理方法的步骤可以是上述各个实施例的视频分类处理方法中的步骤。

[0204] 在一个实施例中,提供了一种计算机可读存储介质,存储有计算机程序,计算机程序被处理器执行时,使得处理器执行上述视频分类处理方法的步骤。此处视频分类处理方法的步骤可以是上述各个实施例的视频分类处理方法中的步骤。

[0205] 需要说明的是,本申请各实施例中的“第一”、“第二”和“第三”等仅用作区分,而并不用于大小、先后、从属等方面的限定。

[0206] 应该理解的是,虽然本申请各实施例中的各个步骤并不是必然按照步骤标号指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,各实施例中至少一部分步骤可以包括多个子步骤或者多个阶段,这些子步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,这些子步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0207] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的程序可存储于一非易失性计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink) DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0208] 以上所述实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0209] 以上所述实施例仅表达了本发明的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本发明构思的前提下,还可以做出若干变形和改进,这些都属于本发明的保护范围。因此,本发明的保护范围应以所附权利要求为准。

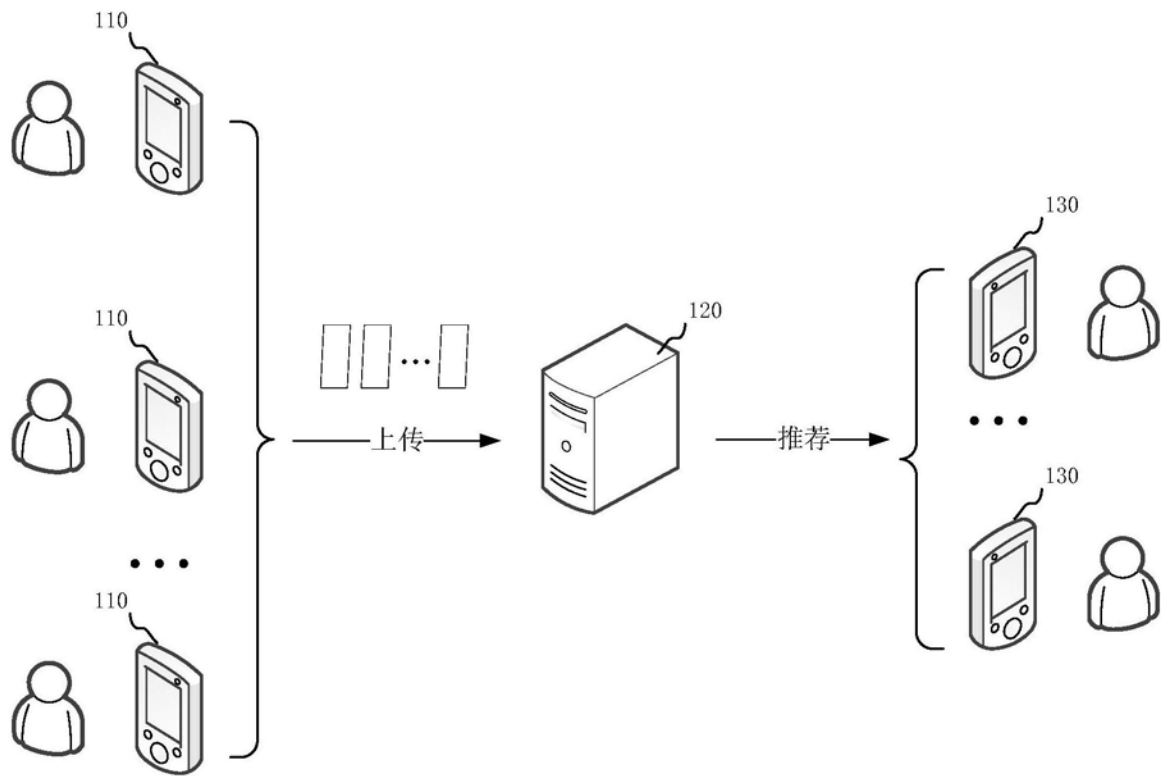


图1

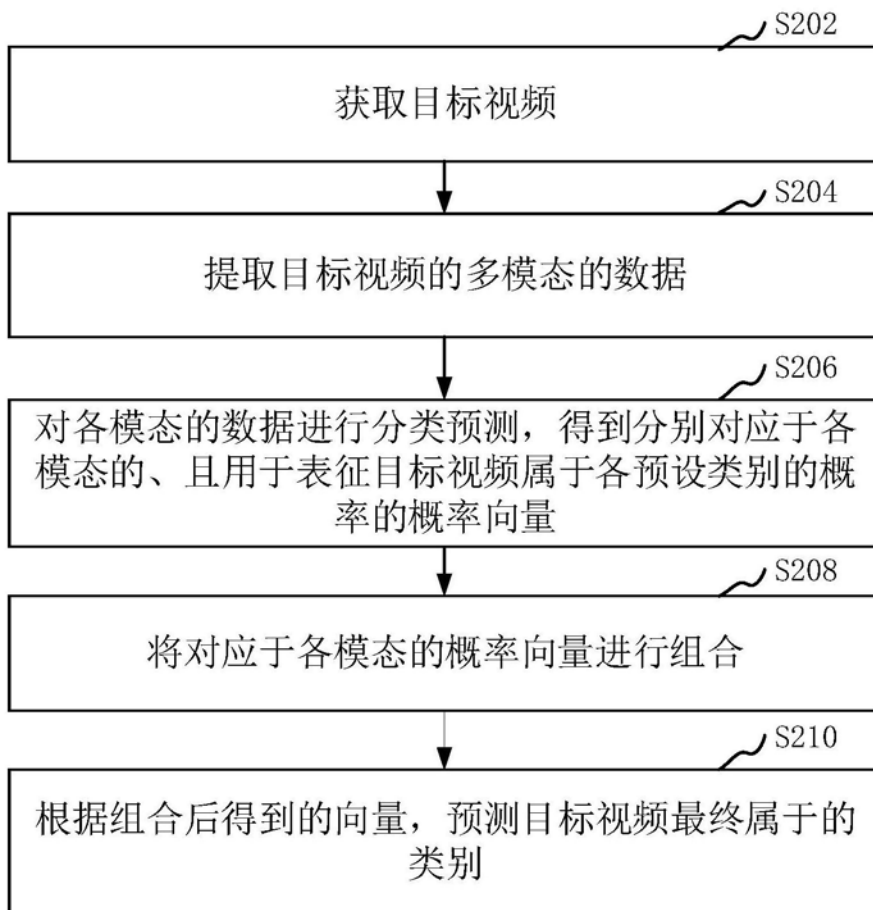


图2

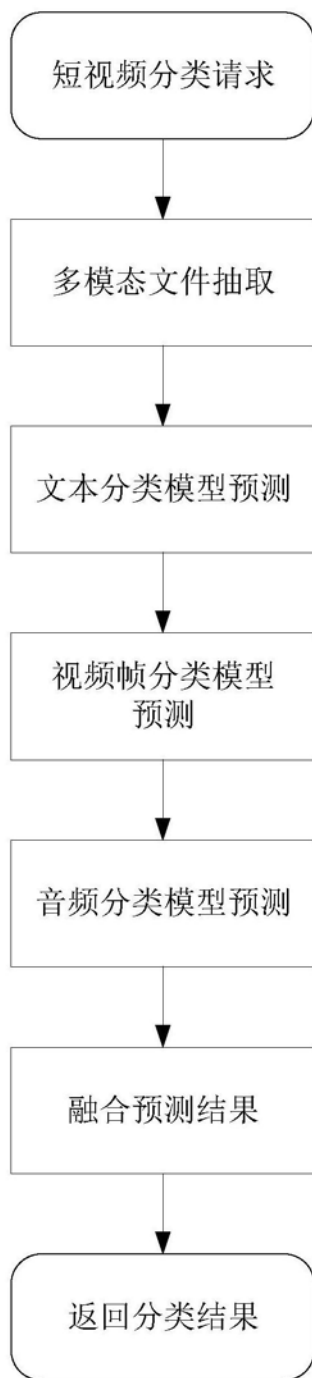


图3

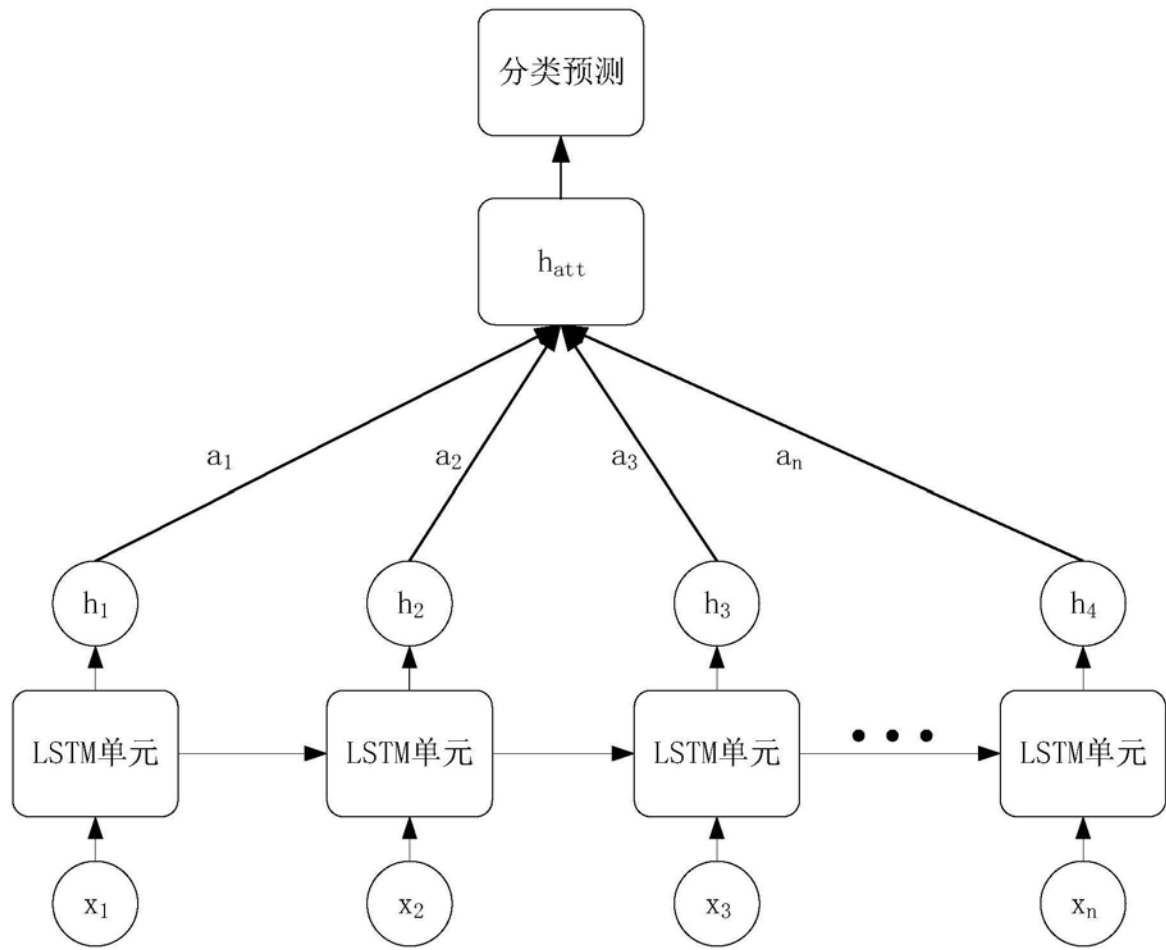


图4

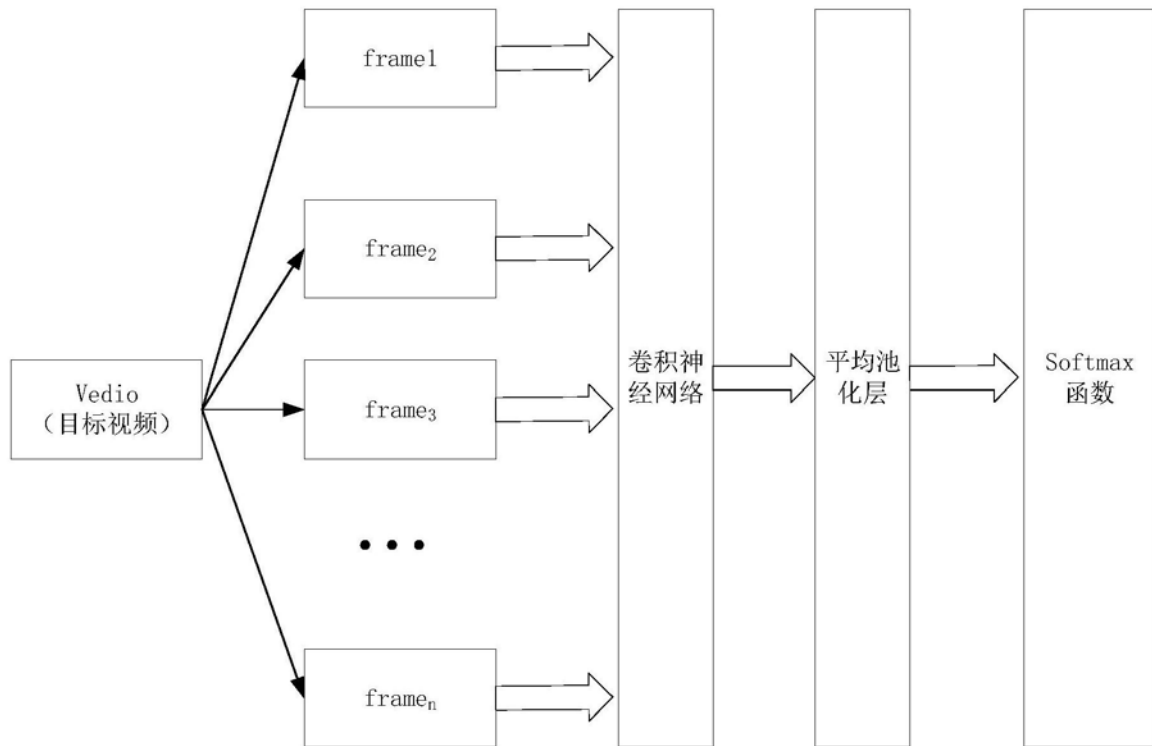


图5

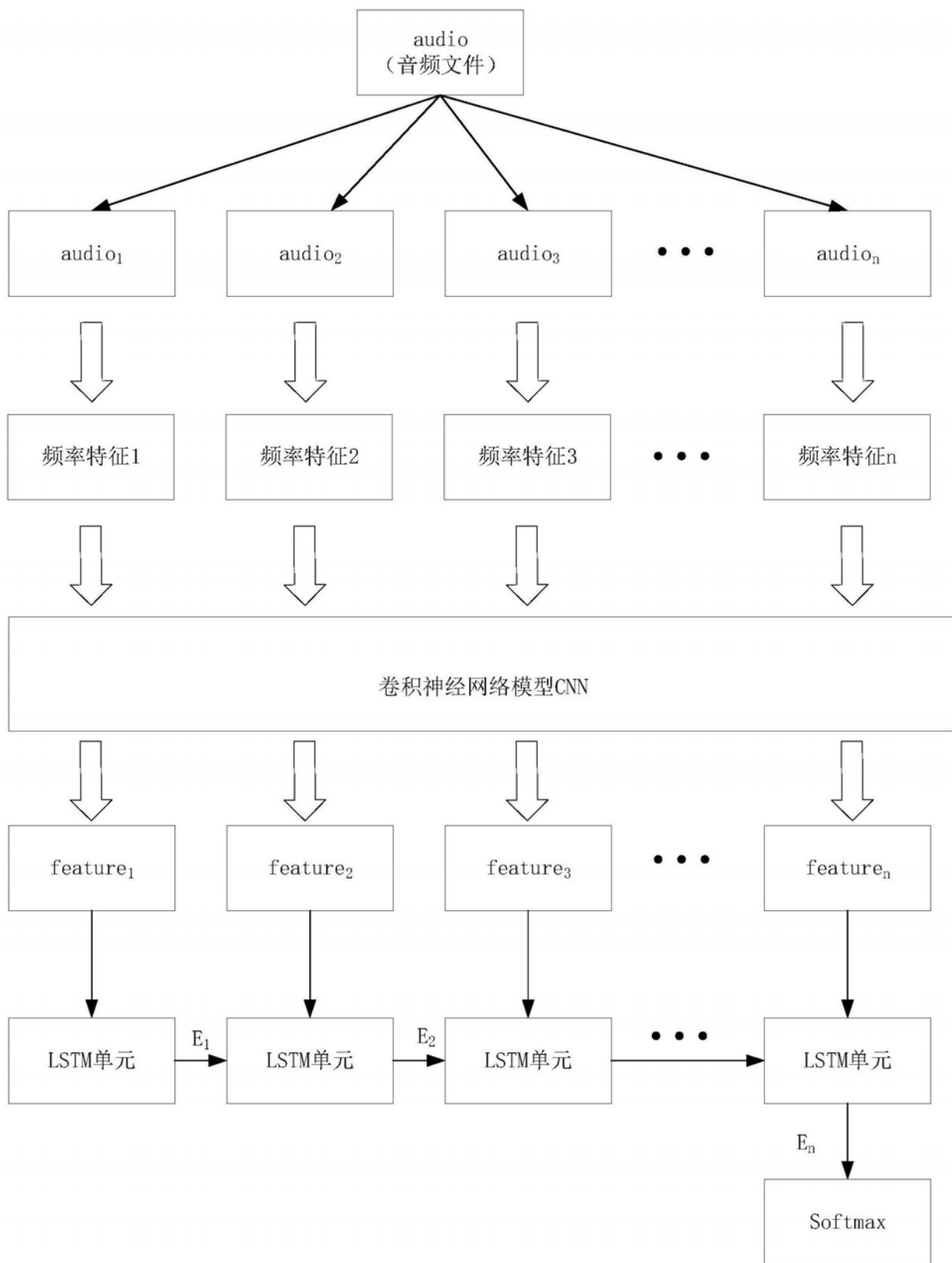


图6

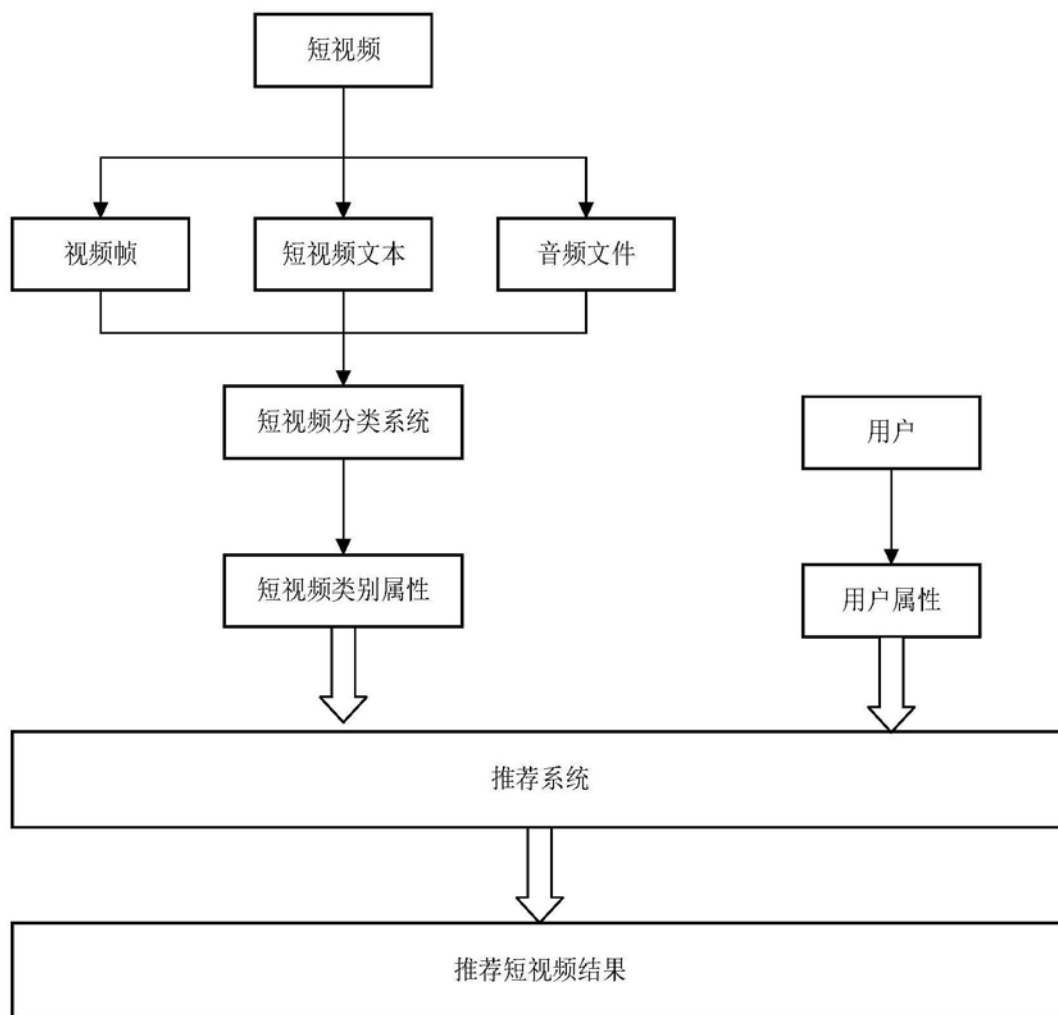


图7

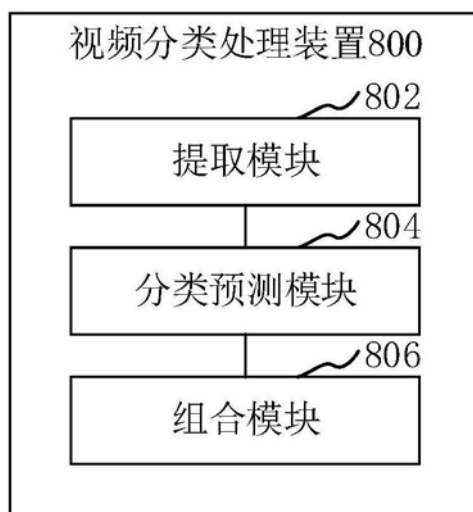


图8

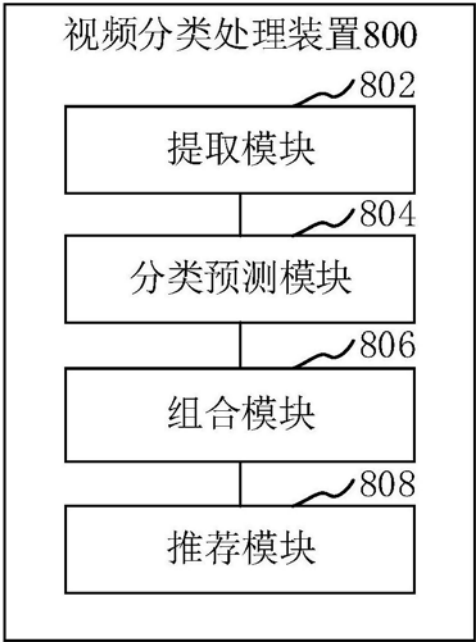


图9

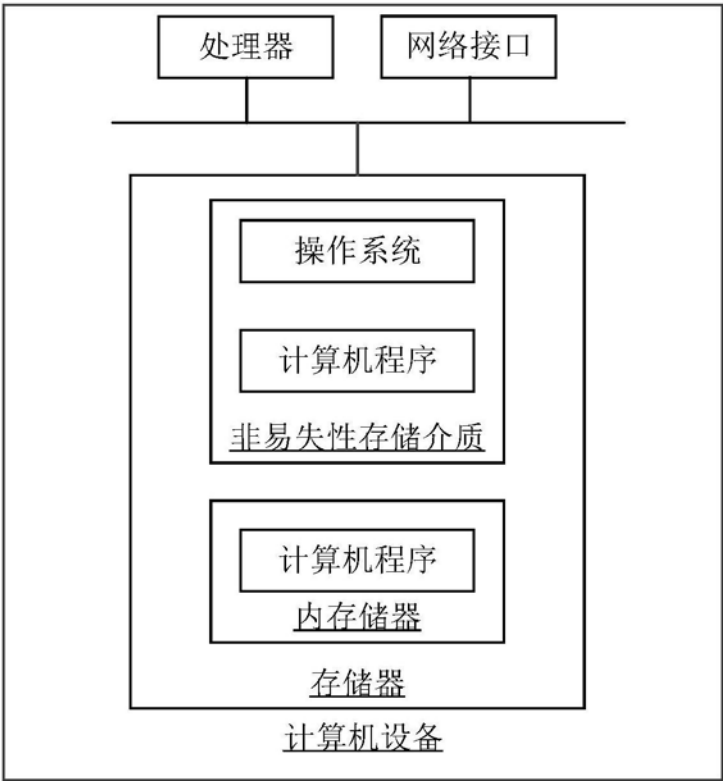


图10