(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115455171 B
(45)授权公告日 2023.05.23

(21)申请号 202211389266.0
(22)申请日 2022.11.08
(65)同一申请的已公布的文献号
申请公布号 CN 115455171 A
(43)申请公布日 2022.12.09
(73)专利权人 苏州浪潮智能科技有限公司
地址 215100 江苏省苏州市吴中经济开发区郭巷街道官浦路1号9幢
(72)发明人 李仁刚 王立 范宝余 郭振华
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
专利代理师 张志梅
(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/335(2019.01)
G06F 16/783(2019.01)
G06N 3/0442(2023.01)
G06N 3/048(2023.01)
G06N 3/08(2023.01)

审查员 高沛沛

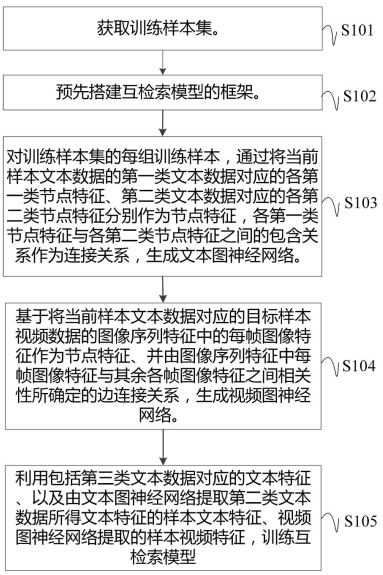
权利要求书5页 说明书21页 附图8页

(54)发明名称

文本视频的互检索以及模型训练方法、装置、设备及介质

(57)摘要

本申请公开了一种用于视频数据与文本数据之间互检索的模型训练方法及装置、视频数据与文本数据之间的互检索方法及装置、互检索设备、可读存储介质，应用于信息检索技术。其中，方法包括对训练样本集各组训练样本，通过将当前样本文本数据对应的节点特征作为节点特征，各节点特征间的包含关系作为连接关系，生成文本图神经网络；基于将目标样本视频数据的图像序列特征中的每帧图像特征作为节点特征、由各帧图像特征之间相关性所确定的边连接关系，生成视频图神经网络；利用融合第三类文本数据特征和文本图神经网络提取的第二类文本数据特征的样本文本特征和视频图神经网络提取的样本视频特征训练互检索模型，可有效提高视频文本的互检索精度。



CN 115455171 B

1.一种用于视频数据与文本数据之间互检索的模型训练方法,其特征在于,包括:

通过将当前样本文本数据的第一类文本数据对应的各第一类节点特征、第二类文本数据对应的各第二类节点特征分别作为节点特征,各第一类节点特征与各第二类节点特征之间的包含关系作为连接关系,生成文本图神经网络;所述第一类文本数据存在于所述第二类文本数据;训练样本集包括多组训练样本,每组训练样本均包括样本文本数据和对应的样本视频数据;

基于将所述当前样本文本数据对应的目标样本视频数据的图像序列特征中的每帧图像特征作为节点特征,以及由所述图像序列特征中每帧图像特征与其余各帧图像特征之间相关性所确定的边连接关系,生成视频图神经网络;

利用包括第三类文本数据对应的文本特征、以及由所述文本图神经网络提取所述第二类文本数据所得文本特征的样本文本特征,所述视频图神经网络提取的样本视频特征,训练互检索模型;所述互检索模型包括所述文本图神经网络和所述视频图神经网络;所述第三类文本数据用于概括所述第一类文本数据和所述第二类文本数据;

其中,所述由所述图像序列特征中每帧图像特征与其余各帧图像特征之间相关性所确定的边连接关系,包括:

对所述图像序列特征的每个图像特征,依次计算当前图像特征与其余各图像特征之间的相似度;

若当前节点的图像特征与目标节点的图像特征的相似度满足相似度条件,则所述当前节点与所述目标节点具有连接关系;调用边权重关系式,计算每两个节点之间的权重值,并基于各权重值生成邻接关系矩阵;所述边权重关系式:

$$A_{ij}=\begin{cases}1-\dfrac{rank(v_i,v_j)}{T}, v_j\in V\\ 0 \qquad\qquad\qquad ,\text{其它}\end{cases};$$

其中,A_{ij}为所述邻接关系矩阵A的元素,T为所述邻接关系矩阵的维度,v_i为第i个节点,v_j为第j个节点,V为图像序列特征集合,$rank(v_i,v_j)$为节点v_j在节点v_i与所有节点相似程度排序中的排序值;所述邻接关系矩阵用于表示每两个节点之间的关联关系。

2.根据权利要求1所述的用于视频数据与文本数据之间互检索的模型训练方法,其特征在于,所述利用包括第三类文本数据对应的文本特征、以及由所述文本图神经网络提取所述第二类文本数据所得文本特征的样本文本特征,所述视频图神经网络提取的样本视频特征,训练互检索模型,包括:

基于所述文本图神经网络提取的样本文本特征、所述视频图神经网络提取的样本视频特征,调用损失函数指导互检索模型的训练过程;所述损失函数为:

$$L^b_{TriHard}=\frac{1}{N}\sum_{a=1}^{N}[d(e_a^{video},e_p^{rec})-\min_{y_n\neq y_a}d(e_a^{video},e_n^{rec})+\nabla]_+ + \frac{1}{N}\sum_{a=1}^{N}[d(e_a^{rec},e_p^{video})-\min_{y_n\neq y_a}d(e_a^{rec},e_n^{video})+\nabla]_+ ;$$

式中,$L^b_{TriHard}$为所述损失函数,N为训练样本组数,e_a^{video}为所述训练样本集中所包含的所有样本视频数据中的第a个样本视频数据,e_p^{rec}为所述训练样本集中所包含的所有样本文

本数据中第p个样本文本数据、且其与第a个样本视频数据相对应，e_n^{rec}为在所有样本文本数据中的第n个样本文本数据、且其与第a个样本视频数据不对应，e_a^{rec}为所有样本文本数据中的第a个样本文本数据，e_p^{video}为所有样本视频数据中第p个样本视频数据、且其与第a个样本文本数据相对应，e_n^{video}为所有样本视频数据中的第n个样本视频数据、且其与第a个样本文本数据不对应，▽为超参数。

3.一种视频数据与文本数据之间的互检索方法，其特征在于，包括：

提取目标文本数据的待匹配文本特征；所述目标文本数据包括第一类文本数据、第二类文本数据和第三类文本数据，且所述第二类文本数据包含所述第一类文本数据，所述第三类文本数据用于概括所述第一类文本数据和所述第二类文本数据；所述待匹配文本特征包括第三类文本数据对应的文本特征、和利用互检索模型的文本图神经网络提取的所述第二类文本数据的文本特征；

提取目标视频数据的待匹配视频特征；

基于所述待匹配视频特征和所述待匹配文本特征，调用所述互检索模型生成所述目标文本数据和所述目标视频数据的检索结果；

其中，所述互检索模型利用如权利要求1或2所述用于视频数据与文本数据之间互检索的模型训练方法训练所得。

4.根据权利要求3所述的视频数据与文本数据之间的互检索方法，其特征在于，所述提取目标视频数据的待匹配视频特征，包括：

通过提取目标视频数据的多帧图像的图像特征，生成所述目标视频数据的图像序列特征；

基于将所述图像序列特征的每个图像特征作为节点特征、并由所述图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，生成视频图神经网络；

利用所述视频图神经网络，获取所述目标视频数据的待匹配视频特征。

5.根据权利要求4所述的视频数据与文本数据之间的互检索方法，其特征在于，所述基于将所述图像序列特征的每个图像特征作为节点特征、并由所述图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，生成视频图神经网络，包括：

所述视频图神经网络包括多层，每一层均包括当前层图结构网络、与所述当前层图结构网络相连的归一化层以及激活层；

所述视频图神经网络的各层图结构网络的神经输入特征图和神经输出特征图跳跃连接；经跳跃连接所得特征图与所述归一化层的归一输出特征图的特征相加和为所述激活层的输入；

其中，基于将所述图像序列特征的每个图像特征作为节点特征、并由所述图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，确定所述视频图神经网络每层的图结构网络。

6.根据权利要求4所述的视频数据与文本数据之间的互检索方法，其特征在于，所述通过提取目标视频数据的多帧图像的图像特征，生成所述目标视频数据的图像序列特征，包括：

预先训练图像特征提取模型;所述图像特征提取模型包含第一预设个数的卷积层和第二预设个数的残差模块,每个残差模块均包含多层卷积层、归一化层和 ReLU 激活函数层;

将目标视频数据的多帧图像输入至所述图像特征提取模型,得到每帧图像的图像特征;

根据各帧图像的图像特征,生成所述目标视频数据的图像序列特征。

7.根据权利要求6所述的视频数据与文本数据之间的互检索方法,其特征在于,所述将目标视频数据的多帧图像输入至所述图像特征提取模型,得到每帧图像的图像特征,包括:

接收图像提取指令,通过解析所述图像提取指令获取图像提取规则;

按照所述图像提取规则,从所述目标视频数据中提取相应帧图像。

8.根据权利要求3所述的视频数据与文本数据之间的互检索方法,其特征在于,所述对所述图像序列特征的每个图像特征,依次计算当前图像特征与其余各图像特征之间的相似度之后,还包括:

若当前节点的图像特征与目标节点的图像特征的相似度不满足相似度条件,则所述当前节点与所述目标节点无连接关系。

9.根据权利要求4至8任意一项所述的视频数据与文本数据之间的互检索方法,其特征在于,所述利用所述视频图神经网络,获取所述目标视频数据的待匹配视频特征,包括:

对所述视频图神经网络的每一层图结构网络,根据当前层图结构网络的图像特征、各节点之间的关联关系、当前层图结构网络的网络参数,更新当前层图神经网络的图像特征;

将更新后的所述视频图神经网络的每一层图结构网络的图像特征,作为所述目标视频数据的待匹配视频特征。

10.根据权利要求9所述的视频数据与文本数据之间的互检索方法,其特征在于,所述根据当前层图结构网络的图像特征、各节点之间的关联关系、当前层图结构网络的网络参数,更新当前层图神经网络的图像特征,包括:

调用视频特征更新关系式,更新所述视频图神经网络的各层图神经网络的图像特征;所述视频特征更新关系式为:

$$Z^{(l)_g}=\sigma(\tilde{D}_{qq}^{-\frac{1}{2}}\tilde{A}_{qm}\tilde{D}_{qq}^{-\frac{1}{2}}Z^{(l)}W^{(l)});$$

式中,$Z^{(l)g}$为所述视频图神经网络更新后的第l层图神经网络的图像特征,$Z^{(l)}$为所述视频图神经网络的第l层图神经网络的图像特征,σ为超参数,$W^{(l)}$为所述视频图神经网络的第l层图结构网络的网络参数,$\tilde{A}_{qm}$为邻接关系矩阵的变换矩阵,$\tilde{A}_{qm}=A+I$,A为邻接关系矩阵,I为单位矩阵,$\tilde{D}_{qq}$为对角矩阵,$\tilde{D}_{qq}=\sum_{m}\tilde{A}_{qm}$,$q$、$m$为矩阵维数。

11.一种用于视频数据与文本数据之间互检索的模型训练装置,其特征在于,包括:

文本图神经网络生成模块,用于通过将当前样本文本数据的第一类文本数据对应的各第一类节点特征、第二类文本数据对应的各第二类节点特征分别作为节点特征,各第一类节点特征与各第二类节点特征之间的包含关系作为连接关系,生成文本图神经网络;所述第一类文本数据存在于所述第二类文本数据;训练样本集包括多组训练样本,每组训练样本均包括样本文本数据和对应的样本视频数据;

视频图神经网络生成模块,用于基于将所述当前样本文本数据对应的目标样本视频数

据的图像序列特征中的每帧图像特征作为节点特征，以及由所述图像序列特征中每帧图像特征与其余各帧图像特征之间相关性所确定的边连接关系，生成视频图神经网络；

模型训练模块，用于利用包括第三类文本数据对应的文本特征、以及由所述文本图神经网络提取所述第二类文本数据所得文本特征的样本文本特征、所述视频图神经网络提取的样本视频特征，训练互检索模型；所述互检索模型包括所述文本图神经网络和所述视频图神经网络；所述第三类文本数据用于概括所述第一类文本数据和所述第二类文本数据；

其中，所述视频图神经网络生成模块进一步用于：

对所述图像序列特征的每个图像特征，依次计算当前图像特征与其余各图像特征之间的相似度；

若当前节点的图像特征与目标节点的图像特征的相似度满足相似度条件，则所述当前节点与所述目标节点具有连接关系；调用边权重关系式，计算每两个节点之间的权重值，并基于各权重值生成邻接关系矩阵；所述边权重关系式：

$$A_{ij}=\begin{cases}1-\dfrac{rank(v_i,v_j)}{T}, v_j\in V\\ 0 \quad ,其它\end{cases};$$

其中，A_{ij}为所述邻接关系矩阵A的元素，T为所述邻接关系矩阵的维度，v_i为第i个节点，v_j为第j个节点，V为图像序列特征集合，$rank(v_i,v_j)$为节点v_j在节点v_i与所有节点相似程度排序中的排序值；所述邻接关系矩阵用于表示每两个节点之间的关联关系。

12.一种视频数据与文本数据之间的互检索装置，其特征在于，包括：

文本特征提取模块，用于提取目标文本数据的待匹配文本特征；所述目标文本数据包括第一类文本数据、第二类文本数据和第三类文本数据，且所述第二类文本数据包含所述第一类文本数据，所述第三类文本数据用于概括所述第一类文本数据和所述第二类文本数据；所述待匹配文本特征包括第三类文本数据对应的文本特征、和利用互检索模型的文本图神经网络提取第二类文本数据的文本特征；

视频特征提取模块，用于提取目标视频数据的待匹配视频特征；

互检索模块，用于基于所述待匹配视频特征和所述待匹配文本特征，调用所述互检索模型生成所述目标文本数据和所述目标视频数据的检索结果；其中，所述互检索模型利用如权利要求1或2所述用于视频数据与文本数据之间互检索的模型训练方法训练所得。

13.一种互检索设备，其特征在于，包括处理器、存储器、人机交互组件以及通信组件；

所述人机交互组件用于通过信息输入/信息输出接口，接收用户输入的训练样本集选择请求、模型训练请求、检索请求以及向用户显示目标文本数据和目标视频数据的检索结果；

所述通信组件用于传输互检索模型训练过程以及所述目标文本数据和所述目标视频数据的互检索任务执行过程中的数据及指令；

所述处理器用于执行所述存储器中存储的计算机程序时实现如权利要求1或2所述用于视频数据与文本数据之间互检索的模型训练方法和/或如权利要求3至10任一项所述视频数据与文本数据之间的互检索方法的步骤。

14.一种可读存储介质，其特征在于，所述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实现如权利要求1或2所述用于视频数据与文本数据之间互检索

的模型训练方法和/或如权利要求3至10任一项所述视频数据与文本数据之间的互检索方法的步骤。

文本视频的互检索以及模型训练方法、装置、设备及介质

技术领域

[0001] 本申请涉及信息检索技术领域，特别是涉及一种用于视频数据与文本数据之间互检索的模型训练方法及装置、视频数据与文本数据之间的互检索方法及装置、互检索设备、可读存储介质。

背景技术

[0002] 随着计算机技术以及网络技术的快速发展和广泛使用，日常工作生活均与之息息相关，不仅导致数据量呈现爆炸式增长，数据类型也越来越大，如图像数据、文本数据、音频数据、视频数据等。不同用户对同一应用场景或是同一目标物的描述往往会采用不同类型的数据进行描述，举例来说，对于同一款服务器来说，既可以采用文本数据描述该服务器的物理参数和性能信息，也可以直接以视频方式描述该服务器的物理参数和性能信息。

[0003] 不可避免的，用户可能会希望基于目标检索词如服务器型号检索到所有相关的、且不同多媒体格式的数据，也可能基于某一类多媒体数据检索到与之相同的其他类型的多媒体数据，举例来说，基于文本信息检索到视频数据。基于此，为了满足用户的检索需求，为用户呈现更加丰富的检索数据，不同媒体间的数据检索或者是称为跨媒体检索成为信息检索技术的趋势。

[0004] 其中，对于多媒体数据类型中的视频数据和文本数据之间的互检索，相关技术提出了一种神经多模态协同学习(Neural Multimodal Cooperative Learning，NMCL)模型，该方法通过学习图像、文本和语音的跨模态互补融合特征来帮助提升短视频分类任务的性能。在智能语音领域，语音识别和语音合成等任务说明了语音和文本之间密切的关联关系。这一系列的证据表面，在图像-文本匹配任务中添加语音信息，有助于提升图像-文本匹配任务的性能。但是，不同媒体数据所含信息量不对等，且其对视频与文本细粒度特征之间的关联关系挖掘并不充分，导致最终的视频数据与文本数据的互检索精度不高。

[0005] 鉴于此，如何提升视频数据和文本数据之间的互检索精度，是所属领域技术人员需要解决的技术问题。

发明内容

[0006] 本申请提供了一种用于视频数据与文本数据之间互检索的模型训练方法及装置、视频数据与文本数据之间的互检索方法及装置、互检索设备、可读存储介质，有效提升视频数据和文本数据之间的互检索精度。

[0007] 为解决上述技术问题，本发明实施例提供以下技术方案：

[0008] 本发明实施例第一方面提供了一种用于视频数据与文本数据之间互检索的模型训练方法，包括：

[0009] 通过将当前样本文本数据的第一类文本数据对应的各第一类节点特征、第二类文本数据对应的各第二类节点特征分别作为节点特征，各第一类节点特征与各第二类节点特征之间的包含关系作为连接关系，生成文本图神经网络；所述第一类文本数据存在于所述

第二类文本数据；训练样本集包括多组训练样本，每组训练样本均包括样本文本数据和对应的样本视频数据；

[0010] 基于将所述当前样本文本数据对应的目标样本视频数据的图像序列特征中的每帧图像特征作为节点特征，以及由所述图像序列特征中每帧图像特征与其余各帧图像特征之间相关性所确定的边连接关系，生成视频图神经网络；

[0011] 利用包括第三类文本数据对应的文本特征，以及由所述文本图神经网络提取所述第二类文本数据所得文本特征的样本文本特征、所述视频图神经网络提取的样本视频特征，训练互检索模型；所述互检索模型包括所述文本图神经网络和所述视频图神经网络；所述第三类文本数据用于概括所述第一类文本数据和所述第二类文本数据。

[0012] 可选的，所述利用包括第三类文本数据对应的文本特征，以及由所述文本图神经网络提取所述第二类文本数据所得文本特征的样本文本特征、所述视频图神经网络提取的样本视频特征，训练互检索模型，包括：

[0013] 基于所述文本图神经网络提取的样本文本特征、所述视频图神经网络提取的样本视频特征，调用损失函数指导互检索模型的训练过程；所述损失函数为：

[0014]
$$L_{TriHard}^{b}=\frac{1}{N}\sum_{a=1}^{N}[d(e_a^{video},e_p^{rec})-\min_{y_n\neq y_a}d(e_a^{video},e_n^{rec})+\nabla]_+ + \frac{1}{N}\sum_{a=1}^{N}[d(e_a^{rec},e_p^{video})-\min_{y_n\neq y_a}d(e_a^{rec},e_n^{video})+\nabla]_+ ;$$

[0015] 式中，$L_{TriHard}^{b}$为所述损失函数，N为训练样本组数，e_a^{video}为所述训练样本集中所包含的所有样本视频数据中的第a个样本视频数据，e_p^{rec}为所述训练样本集中所包含的所有样本文本数据中第p个样本文本数据、且其与第a个样本视频数据相对应，e_n^{rec}为在所有样本文本数据中的第n个样本文本数据、且其与第a个样本视频数据不对应，e_a^{rec}为所有样本文本数据中的第a个样本文本数据，e_p^{video}为所有样本视频数据中第p个样本视频数据、且其与第a个样本文本数据相对应，e_n^{video}为所有样本视频数据中的第n个样本视频数据、且其与第a个样本文本数据不对应，▽为超参数。

[0016] 本发明实施例第二方面提供了一种用于视频数据与文本数据之间互检索的模型训练装置，包括：

[0017] 提取目标文本数据的待匹配文本特征；所述目标文本数据包括第一类文本数据、第二类文本数据和第三类文本数据，且所述第二类文本数据包含所述第一类文本数据，所述第三类文本数据用于概括所述第一类文本数据和所述第二类文本数据；所述待匹配文本特征包括第三类文本数据对应的文本特征、和利用互检索模型的文本图神经网络提取第二类文本数据的文本特征；

[0018] 提取目标视频数据的待匹配视频特征；

[0019] 基于所述待匹配视频特征和所述待匹配文本特征，调用互检索模型生成所述目标文本数据和所述目标视频数据的检索结果；

[0020] 其中，所述互检索模型利用如前任意一项所述用于视频数据与文本数据之间互检索的模型训练方法训练所得。

[0021] 可选的，所述提取目标视频数据的待匹配视频特征，包括：

[0022] 通过提取目标视频数据的多帧图像的图像特征，生成所述目标视频数据的图像序列特征；

[0023] 基于将所述图像序列特征的每个图像特征作为节点特征、并由所述图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，生成视频图神经网络；

[0024] 利用所述视频图神经网络，获取所述目标视频数据的待匹配视频特征。

[0025] 可选的，所述基于将所述图像序列特征的每个图像特征作为节点特征、并由所述图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，生成视频图神经网络，包括：

[0026] 所述视频图神经网络包括多层，每一层均包括当前层图结构网络、与所述当前层图结构网络相连的归一化层以及激活层；

[0027] 所述视频图神经网络的各层图结构网络的神经输入特征图和神经输出特征图跳跃连接；经跳跃连接所得特征图与所述归一化层的归一输出特征图的特征相加和为所述激活层的输入；

[0028] 其中，基于将所述图像序列特征的每个图像特征作为节点特征、并由所述图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，确定所述视频图神经网络每层的图结构网络。

[0029] 可选的，所述通过提取目标视频数据的多帧图像的图像特征，生成所述目标视频数据的图像序列特征，包括：

[0030] 预先训练图像特征提取模型；所述图像特征提取模型包含第一预设个数的卷积层和第二预设个数的残差模块，每个残差模块均包含多层卷积层、归一化层和ReLU激活函数层；

[0031] 将目标视频数据的多帧图像输入至所述图像特征提取模型，得到每帧图像的图像特征；

[0032] 根据各帧图像的图像特征，生成所述目标视频数据的图像序列特征。

[0033] 可选的，所述将目标视频数据的多帧图像输入至所述图像特征提取模型，得到每帧图像的图像特征，包括：

[0034] 接收图像提取指令，通过解析所述图像提取指令获取图像提取规则；

[0035] 按照所述图像提取规则，从所述目标视频数据中提取相应帧图像。

[0036] 可选的，所述由所述图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，包括：

[0037] 对所述图像序列特征的每个图像特征，依次计算当前图像特征与其余各图像特征之间的相似度；

[0038] 若当前节点的图像特征与目标节点的图像特征的相似度满足相似度条件，则所述当前节点与所述目标节点具有连接关系；若当前节点的图像特征与目标节点的图像特征的相似度不满足相似度条件，则所述当前节点与所述目标节点无连接关系。

[0039] 可选的，所述若当前节点的图像特征与目标节点的图像特征的相似度满足相似度条件，则所述当前节点与所述目标节点具有连接关系之后，还包括：

[0040] 调用边权重关系式，计算每两个节点之间的权重值，并基于各权重值生成邻接关系矩阵；所述边权重关系式：

[0041] $$A_{ij}=\begin{cases}1-\dfrac{rank(v_i,v_j)}{T}, v_j\in V\\ 0 \quad ,其它\end{cases};$$

[0042] 其中，A_{ij}为所述邻接关系矩阵A的元素，T为所述邻接关系矩阵的维度，v_i为第i个节点，v_j为第j个节点，V为图像序列特征集合，$rank(v_i,v_j)$为节点v_j在v_i与所有节点相似程度排序中的排序值；所述邻接关系矩阵用于表示每两个节点之间的关联关系。

[0043] 可选的，所述利用所述视频图神经网络，获取所述目标视频数据的待匹配视频特征，包括：

[0044] 对所述视频图神经网络的每一层图结构网络，根据当前层图结构网络的图像特征、各节点之间的关联关系、当前层图结构网络的网络参数，更新所述当前层图神经网络的图像特征；

[0045] 将更新后的所述视频图神经网络的每一层图结构网络的图像特征，作为所述目标视频数据的待匹配视频特征。

[0046] 可选的，所述根据当前层图结构网络的图像特征、各节点之间的关联关系、当前层图结构网络的网络参数，更新所述当前层图神经网络的图像特征，包括：

[0047] 调用视频特征更新关系式，更新所述视频图神经网络的各层图神经网络的图像特征；所述视频特征更新关系式为：

[0048] $$Z^{(l)_g}=\sigma(\tilde{D}_{qq}^{-\frac{1}{2}}\tilde{A}_{qm}\tilde{D}_{qq}^{-\frac{1}{2}}Z^{(l)}W^{(l)});$$

[0049] 式中，$Z^{(1)g}$为所述视频图神经网络更新后的第1层图神经网络的图像特征，$Z^{(1)}$为所述视频图神经网络的第1层图神经网络的图像特征，σ为超参数，$W^{(1)}$为所述视频图神经网络的第1层图结构网络的网络参数，$\tilde{A}_{qm}$为邻接关系矩阵的变换矩阵，$\tilde{A}_{qm}=A+I$，A为邻接关系矩阵，I为单位矩阵，$\tilde{D}_{qq}$为对角矩阵，$\tilde{D}_{qq}=\sum_m \tilde{A}_{qm}$，q、m为矩阵维数。

[0050] 本发明实施例第三方面提供了一种视频数据与文本数据之间的互检索方法，包括：

[0051] 文本图神经网络生成模块，用于通过将当前样本文本数据的第一类文本数据对应的各第一类节点特征、第二类文本数据对应的各第二类节点特征分别作为节点特征，各第一类节点特征与各第二类节点特征之间的包含关系作为连接关系，生成文本图神经网络；所述第一类文本数据存在于所述第二类文本数据；训练样本集包括多组训练样本，每组训练样本均包括样本文本数据和对应的样本视频数据；

[0052] 视频图神经网络生成模块，用于基于将所述当前样本文本数据对应的目标样本视频数据的图像序列特征中的每帧图像特征作为节点特征，以及由所述图像序列特征中每帧

图像特征与其余各帧图像特征之间相关性所确定的边连接关系，生成视频图神经网络；

[0053] 模型训练模块，用于利用包括第三类文本数据对应的文本特征、以及由所述文本图神经网络提取所述第二类文本数据所得文本特征的样本文本特征、所述视频图神经网络提取的样本视频特征，训练互检索模型；所述互检索模型包括所述文本图神经网络和所述视频图神经网络；所述第三类文本数据用于概括所述第一类文本数据和所述第二类文本数据。

[0054] 本发明实施例第四方面提供了一种视频数据与文本数据之间的互检索装置，包括：

[0055] 文本特征提取模块，用于提取目标文本数据的待匹配文本特征；所述目标文本数据包括第一类文本数据、第二类文本数据和第三类文本数据，且所述第二类文本数据包含所述第一类文本数据，所述第三类文本数据用于概括所述第一类文本数据和所述第二类文本数据；所述待匹配文本特征包括第三类文本数据对应的文本特征、和利用互检索模型的文本图神经网络提取第二类文本数据的文本特征；

[0056] 视频特征提取模块，用于提取目标视频数据的待匹配视频特征；

[0057] 互检索模块，用于基于所述待匹配视频特征和所述待匹配文本特征，调用所述互检索模型生成所述目标文本数据和所述目标视频数据的检索结果；其中，所述互检索模型利用前任意一项所述用于视频数据与文本数据之间互检索的模型训练方法训练所得。

[0058] 本发明实施例还提供了一种互检索设备，包括处理器、存储器、人机交互组件以及通信组件；

[0059] 所述人机交互组件用于通过信息输入/信息输出接口，接收用户输入的训练样本集选择请求、模型训练请求、检索请求以及向用户显示目标文本数据和目标视频数据的检索结果；

[0060] 所述通信组件用于传输互检索模型训练过程以及所述目标文本数据和所述目标视频数据的互检索任务执行过程中的数据及指令；

[0061] 所述处理器用于执行所述存储器中存储的计算机程序时实现如前任意一项所述用于视频数据与文本数据之间互检索的模型训练方法和/或如前任一项所述视频数据与文本数据之间的互检索方法的步骤。

[0062] 本发明实施例最后还提供了一种可读存储介质，所述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实现如前任意一项所述用于视频数据与文本数据之间互检索的模型训练方法和/或如前任一项所述视频数据与文本数据之间的互检索方法的步骤。

[0063] 本申请提供的技术方案的优点在于，分别基于文本和视频所包含的数据及其内部关系，构建用于提取相应特征的图神经网络，从而有利于提取可反映现实世界中的文本及其内在关联关系的文本特征，反映现实世界中视频及其内在关联关系的视频特征，将概括文本数据的第三类文本数据和第二类文本数据的融合特征作为执行匹配任务的文本特征，可进一步挖掘文本数据之间的内在关系，最后基于提取的文本特征及视频特征进行模型训练，有利于充分挖掘视频与文本细粒度特征之间的关联关系，从而得到高精度的视频文本互检索模型，有效提高视频数据与文本数据的互检索精度。

[0064] 此外，本发明实施例还针对用于视频数据与文本数据之间互检索的模型训练方法

提供了视频数据与文本数据之间的互检索方法及各自相应的装置、互检索设备、可读存储介质，进一步使得该方法更具有实用性，所述用于视频数据与文本数据之间互检索的模型训练装置、视频数据与文本数据之间的互检索方法及装置、互检索设备、可读存储介质具有相应的优点。

[0065] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性的，并不能限制本公开。

附图说明

[0066] 为了更清楚的说明本发明实施例或相关技术的技术方案，下面将对实施例或相关技术描述中所需要使用的附图作简单的介绍，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。

[0067] 图1为本发明实施例提供的一种用于视频数据与文本数据之间互检索的模型训练方法的流程示意图；

[0068] 图2为本发明实施例提供的文本图神经网络在一种可选实施方式下的结构框架示意图；

[0069] 图3为本发明实施例提供的一种视频数据与文本数据之间的互检索方法的流程示意图；

[0070] 图4为本发明实施例提供的图像特征提取模型在一种可选实施方式下的模型结构示意图；

[0071] 图5为本发明实施例提供的图像特征提取模型的一种可选的网络参数示意图；

[0072] 图6为本发明实施例提供的视频图神经网络在一种可选实施方式下的结构框架示意图；

[0073] 图7为本发明实施例提供的一个示例性应用场景的框架示意图；

[0074] 图8为本发明实施例提供的互检索模型结构的示意图；

[0075] 图9为本发明实施例提供的文本特征提取模型在一种可选实施方式下的模型结构示意图；

[0076] 图10为本发明实施例提供的用于视频数据与文本数据之间互检索的模型训练装置的一种具体实施方式结构图；

[0077] 图11为本发明实施例提供的视频数据与文本数据之间的互检索装置的一种具体实施方式结构图；

[0078] 图12为本发明实施例提供的互检索设备的一种具体实施方式结构图。

具体实施方式

[0079] 为了使本技术领域的人员更好地理解本发明方案，下面结合附图和具体实施方式对本发明作进一步的详细说明。显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0080] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第

四”等是用于区别不同的对象，而不是用于描述特定的顺序。此外术语“包括”和“具有”以及他们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可包括没有列出的步骤或单元。

[0081] 在介绍了本发明实施例的技术方案后，下面详细的说明本申请的各种非限制性实施方式。

[0082] 首先参见图1，图1为本发明实施例提供的一种用于视频数据与文本数据之间互检索的模型训练方法的流程示意图，本发明实施例可包括以下内容：

[0083] S101：获取训练样本集。

[0084] 本步骤的训练样本集包括多组训练样本，每组训练样本均包括相对应的一个样本文本和一个样本视频，也就是样本文本和样本视频为相匹配的一组样本数据，训练样本集所包含的训练样本组数可根据实际训练需求以及实际应用场景来确定，本申请对此不作任何限定。训练样本集中的样本文本可从任何一种已有数据库中获取，该样本文本对应的视频样本可从相应的数据库中获取。当然，为了扩充训练样本集的数量。样本文本或视频文本也可为对原始样本文本或视频文本样本进行裁剪、拼接、拉伸等处理后的数据。本实施例的样本文本或者是待检索文本包括至少三种性质完全不同的数据，且其中两类数据之间具有包含关系。为了便于描述，可称为第一类文本数据和第二类文本数据，所谓的包含关系是指第一类文本数据所包含的特征均存在于第二类文本数据的特征中，另一类数据为概括第一类文本数据和第二类文本数据的文本数据，以菜谱文本举例来说，第一类文本数据可为菜谱成分，第二类文本数据可为做菜步骤，第三类文本数据可为菜名；以服务器工作原理说明文档举例来说，第一类文本数据可为服务器结构组成，第二类文本数据可为工作原理，第三类文本数据为服务器工作原理；以电子设备说明书为例，第一类文本数据可为电子设备的产品结构，第二类文本数据为使用说明书，第三类文本数据为电子设备说明书。

[0085] S102：预先搭建互检索模型的框架。

[0086] 本实施例的互检索模型用于执行文本数据与视频数据之间的互检索任务，所谓的互检索任务是指互检索模型可以基于待检索文本数据从已知视频数据库中确定与之相匹配的视频数据，也可基于待检索视频数据从已知文本数据库中确定与之相匹配的文本数据。本实施例的互检索模型包括文本图神经网络和视频图神经网络。文本图神经网络用于对输入文本数据如样本文本或待检索文本的第二类文本数据进行处理并最终输出该文本数据对应的文本特征，视频图神经网络用于对输入视频数据如样本视频或待检索视频进行处理，并输出该视频数据的最终视频特征。文本图神经网络和视频图神经网络可基于任何技术中的任何一种图结构进行搭建，这均不影响本申请的实现。

[0087] S103：对训练样本集的每组训练样本，通过将当前样本文本数据的第一类文本数据对应的各第一类节点特征、第二类文本数据对应的各第二类节点特征分别作为节点特征，各第一类节点特征与各第二类节点特征之间的包含关系作为连接关系，生成文本图神经网络。

[0088] 在本实施例中，文本图神经网络是基于图结构搭建的神经网络模型，文本图神经网络除了包括图结构，还包括文本特征提取功能对应结构和文本特征输出对应的结构，对于文本特征提取功能对应结构和文本特征输出对应的结构可采用任何一种现有的机器学习模型如Bert(Bidirectional Encoder Representation from Transformers，预训练的

语言表征模型)、word2vec(word to vector,词向量模型)、双向长短期记忆神经网络,长短期记忆神经网络实现相应功能的模型结构,其中,图结构包括节点和连接边,通过文本特征提取功能提出输入文本的每类文本数据的文本特征,文本图神经网络的节点即为文本样本的文本特征,样本文本由于所包含的数据类型至少包括两种性质不同的数据,故这两种数据类型对应的文本特征可作为文本图神经网络的异构节点,每类数据均包含多个文本特征,一个文本特征对应一个节点。文本图神经网络的连接边由各异构节点对应文本特征之间是否具有包含关系来决定,如果某两个异构节点对应的文本特征之间具有包含关系,也即如果第一类文本数据的第一个特征也即第一个第一类节点特征在第二类文本数据的第一个第一类节点特征中出现,则第一类文本数据的第一个第一类节点特征对应的节点与第二类文本数据的第一个第一类节点特征对应的节点具有连接边。举例来说,如图2所示,样本文本包括两类文本数据,第一类文本数据的各第一类节点特征包括电源v_1^{ins}、开关键v_2^{ins}、指示灯v_3^{ins}、指示器v_4^{ins},第二类文本数据的各第二类节点特征包括连接电源v_1^{ing}、开启开关键,指示灯闪烁v_2^{ing}、若指示灯停止闪烁时,则进入工作状态v_3^{ing},则文本图神经网络的节点包括v_1^{ins}、v_2^{ins}、v_3^{ins}、v_4^{ins}、v_1^{ing}、v_2^{ing}、v_3^{ing},由于v_2^{ing}、v_3^{ing}中包含的v_3^{ing}特征,也即v_3^{ins}和v_2^{ing}、v_3^{ing}均有关联关系,则v_3^{ins}和v_2^{ing}、v_3^{ing}具有连接边e_{32}、e_{33};由于v_1^{ins}和v_1^{ing}有包含关系,所以v_1^{ins}和v_1^{ing}之间也具有连接边e_{11}。从图结构数据中可以提取样本文本的空间特征,基于提取的空间特征结合特征输出功能生成最终的文本特征。

[0089] S104:基于将当前样本文本数据对应的目标样本视频数据的图像序列特征中的每帧图像特征作为节点特征、并由图像序列特征中每帧图像特征与其余各帧图像特征之间相关性所确定的边连接关系,生成视频图神经网络。

[0090] 在上个步骤确定文本特征之后,由于训练样本为一对,本步骤可针对该文本特征对应的视频样本的视频特征进行提取处理,也即确定用于生成视频特征的视频图神经网络。同样的,本申请用于处理视频数据的网络模型基于图结构,其除了包括图结构,还包括图像特征提取功能对应结构和视频特征输出对应的结构,对于图像特征提取功能对应结构和视频特征输出对应的结构可采用任何一种现有的机器学习模型如人工卷积神经网络、VGG16(Visual Geometry Group Network,目视图像生成器)、Resnet(Deep residual network,深度残差网络)等实现相应功能的模型结构。其中,对于视频图神经网络的图结构,基于视频图神经网络的图像特性提取功能对输入视频的关键帧的图像特征,得到一组图像特征也即本步骤所称为的图像序列特征,对于该组图像序列特征,本实施例将每个图像特征对应作为图结构的一个节点,该图像序列特征中的每个图像特征与该图像序列特征中的其余图像特征之间的相关性来判断这两个节点之间是否具有连接边,两个图像特征的相关性可通过相似度来衡量,进一步的,两个图像特征的相关性可利用欧式距离、余弦距离、马氏距离等等来确定特征间相似度。对于相似度值大于等于预设相似度阈值的两个图像特征,其二者对应的节点之间具有连接边,对于相似度值小于预设相似度阈值的两个图像特征,其二者对应的节点之间没有连接边。从图结构数据中可以提取样本视频的空间特征,基于提取的空间特征确定最终的视频特征。

[0091] S105:利用包括第三类文本数据对应的文本特征、以及由文本图神经网络提取第二类文本数据所得文本特征的样本文本特征、视频图神经网络提取的样本视频特征,训练互检索模型。

[0092] 在本实施例中,一个样本文本的文本特征对应一个样本视频的视频特征,本实施例的每个样本文本的文本特征均为融合特征,融合的是该样本文本的第三类文本数据对应的文本特征以及其第二类文本数据由文本图神经网络提取所得到的特征。对于第三类文本数据对应的文本特征可采用任何一种文本特征提取模型提取得到,本实施例对此不做任何限定。模型训练包括前向传播阶段和反向传播阶段,前向传播阶段是数据由低层次向高层次传播的阶段,反向传播阶段是当前向传播得出的结果与预期不相符时,将误差从高层次向底层次进行传播训练的阶段。训练过程可采用损失函数来指导,然后通过诸如梯度反传等模型参数更新方式实现对文本图神经网络以及视频图神经网络的各网络参数的更新,直至达到迭代次数或者取得满意的收敛为止。举例来说,先随机初始化互检索模型中的文本图神经网络及视频图神经网络的所有网络层的权重值,然后输入样本视频和文本视频经过文本图神经网络及视频图神经网络各层的前向传播得到输出值;计算互检索模型的模型输出值,并基于损失函数计算该输出值的损失值。将误差反向传回互检索模型中,依次求得文本图神经网络及视频图神经网络的各层的反向传播误差,根据各层的反向传播误差对文本图神经网络及视频图神经网络的所有权重系数进行调整,实现权重的更新。再次随机从训练样本集中选取一对新的视频样本和文本样本,然后重复上述过程,无限往复迭代,直至计算得到的模型输出值与标签之间的误差小于预设阈值,结束模型训练,并将此刻模型所有层参数作为训练好的互检索模型网络参数。

[0093] 在进行互检索模型训练过程中,可采用如L1范数损失函数、均方误差损失函数、交叉熵损失等任何一种损失函数,为了进一步提高互检索模型的精准度,本申请还给出了一种损失函数的可选实施方式,也即可基于文本图神经网络提取的样本文本特征、视频图神经网络提取的样本视频特征,调用损失函数指导互检索模型的训练过程;该损失函数可表述为:

[0094]
$$L_{TriHard}^{b}=\frac{1}{N}\sum_{a=1}^{N}[d(e_a^{video},e_p^{rec})-\min_{y_n\neq y_a}d(e_a^{video},e_n^{rec})+\nabla]_+ +$$
$$\frac{1}{N}\sum_{a=1}^{N}[d(e_a^{rec},e_p^{video})-\min_{y_n\neq y_a}d(e_a^{rec},e_n^{video})+\nabla]_+ ;$$

[0095] 式中,$L_{TriHard}^{b}$为损失函数,N为训练样本组数,e_a^{video}为训练样本集中所包含的所有样本视频数据中的第a个样本视频数据,e_p^{rec}为训练样本集中所包含的所有样本文本数据中第p个样本文本数据、且其与第a个样本视频数据相对应,e_n^{rec}为在所有样本文本数据中的第n个样本文本数据、且其与第a个样本视频数据不对应,e_a^{rec}为所有样本文本数据中的第a个样本文本数据,e_p^{video}为所有样本视频数据中第p个样本视频数据、且其与第a个样本文本数据相对应,e_n^{video}为所有样本视频数据中的第n个样本视频数据、且其与第a个样本文本数据不对应,▽为超参数。

[0096] 在本发明实施例提供的技术方案中，分别基于文本和视频所包含的数据及其内部关系，构建用于提取相应特征的图神经网络，从而有利于提取可反映现实世界中的文本及其内在关联关系的文本特征，反映现实世界中视频及其内在关联关系的视频特征，将概括文本数据的第三类文本数据和第二类文本数据的融合特征作为执行匹配任务的文本特征，可进一步挖掘文本数据之间的内在关系，最后基于提取的文本特征及视频特征进行模型训练，有利于充分挖掘视频与文本细粒度特征之间的关联关系，从而得到高精度的视频文本互检索模型，有效提高视频数据与文本数据的互检索精度。

[0097] 此外，基于上述实施例，本申请还提供了视频数据与文本数据之间的互检索方法，请参阅图3，图3为本发明实施例提供的另一种视频数据与文本数据之间的互检索方法的流程示意图，可包括以下内容：

[0098] S301：预先训练用于执行视频数据与文本数据之间互检索任务的互检索模型。

[0099] 利用上述实施例中用于视频数据与文本数据之间互检索的模型训练方法训练得到互检索模型。

[0100] S302：提取目标文本数据的待匹配文本特征。

[0101] 本实施例的目标文本数据至少包括三类文本数据，为了便于描述，可称为第一类文本数据、第二类文本数据和第三类文本数据，第二类文本数据且包含第一类文本数据，且第三类文本数据可概括第一类文本数据和第二类文本数据。所谓的包含关系是指第一类文本数据的特征会存在于第二类文本数据中。进一步的，本申请的第二类文本数据所包含的各类文本特征也即各第二类文本特征之间具有计算依赖关系或者是先后执行顺序关系的，如第二类文本数据可为使用步骤数据，如第二类文本数据可为物理参数计算数据等。待匹配文本特征为融合第三类文本数据对应的文本特征、和利用互检索模型的文本图神经网络提取第二类文本数据的文本特征所得。在一个文本数据中，不同类型的文本数据之间具有关联性，可先通过文本图神经网络中实现文本特征提取功能的结构抽取文本特征，抽取得到的文本特征表达作为文本图神经网络中的图结构的输入，通过学习不同文本特征之间的互补特征，确定各类文本数据之间的潜在联系。至于待匹配文本特征的提取过程可参阅上述实施例中样本文本的文本特征提取方式，此处，便不再做任何赘述。

[0102] S303：提取目标视频数据的待匹配视频特征。

[0103] 本步骤中，可先获取目标视频的某些帧或者是全部帧的图像特征，然后基于这些图像特征整合生成视频特征，为了描述不引起歧义，可称为待匹配视频特征，同理，目标文本数据的文本特征称为待匹配文本特征。在一个视频中，不同的图像帧具有语义相似性，帧与帧之间可能依赖，对于不同的关键帧，可通过视频图神经网络中实现图像特征提取功能的结构抽取关键帧特征，抽取得到的特征表达作为视频图神经网络中的图结构的输入，学习不同帧特征之间的互补特征，建立不同视频帧间的潜在联系。至于待匹配视频特征的提取过程可参阅上述实施例中样本视频的视频特征提取方式，此处，便不再做任何赘述。

[0104] S304：基于待匹配视频特征和待匹配文本特征，调用互检索模型生成目标文本数据和目标视频数据的检索结果。

[0105] 若用户待检索请求是从目标数据库中检索指定文本也即目标文本数据对应的视频，则待检索请求携带目标文本数据，目标视频数据为目标数据库中随机选择的一个视频数据，通过依次比对目标文本数据与目标数据库的各视频之间的相似性，最终确定与该目

标文本数据最匹配的视频数据，并输出最终确定的最匹配的视频。若用户待检索请求是从目标数据库中检索指定视频也即目标视频数据对应的文本，则待检索请求携带目标视频数据，目标文本数据为目标数据库中随机选择的一个文本数据，通过依次比对目标文本数据与目标数据库的各文本数据之间的相似性，最终确定与该目标视频数据最匹配的文本数据，并输出最终确定的最匹配的文本。

[0106] 由上可知，本发明实施例可有效提升视频数据和文本数据之间的互检索精度。

[0107] 需要说明的是，本申请中各步骤之间没有严格的先后执行顺序，只要符合逻辑上的顺序，则这些步骤可以同时执行，也可按照某种预设顺序执行，图1和图3只是一种示意方式，并不代表只能是这样的执行顺序。

[0108] 在上述实施例中，对于如何执行步骤S103并不做限定，本实施例中给出目标视频数据的待匹配视频特征的一种可选的提取方式，可包括下述内容：

[0109] 通过提取目标视频数据的多帧图像的图像特征，生成目标视频数据的图像序列特征；

[0110] 基于将图像序列特征的每个图像特征作为节点特征、并由图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，生成视频图神经网络；

[0111] 利用视频图神经网络，获取目标视频数据的待匹配视频特征。

[0112] 其中，多帧图像可为目标视频数据的每一帧图像，也可为每1s从目标视频中提取对应帧图像，还可为将目标视频数据分为多段视频，并提取每段视频的第一帧图像构成多帧图像。可选的，用户可下发图像提取指令，图像提取指令携带图像提取规则，也即如何提取目标视频数据的图像帧的方法，系统接收图像提取指令，通过解析图像提取指令获取图像提取规则；按照图像提取规则，从目标视频数据中提取相应帧图像。在从目标视频数据中确定提取图像特征的图像帧之后，可利用S301步骤训练好的视频图神经网络的图像特征提取功能，提取这些图像帧的图像特征，以作为图像序列特征。可选的，实现视频图神经网络的图像特征提取功能的结构可称为图像特征提取模型，将目标视频数据的多帧图像输入至该图像特征提取模型，得到每帧图像的图像特征；根据各帧图像的图像特征，生成目标视频数据的图像序列特征。本实施例还给出了图像特征提取模型的一种可选结构方式，图像特征提取模型可包含第一预设个数的卷积层和第二预设个数的残差模块，每个残差模块均包含多层卷积层、归一化层和ReLU激活函数。举例来说，图像特征提取模型可采用ResNet50网络提取目标视频数据中每帧图像的特征，如图4及图5所示，ResNet50可包含1个卷积层和4个残差层，每个残差模块包含多层卷积、归一化层和ReLU激活函数层。图5中，[ ]内代表残差块的基本组成，×n代表堆叠次数，输出尺寸代表经过不同网络层后输处的特征图的尺度。残差模块由1×1的卷积核和一个3×3的卷积核组成。为了保证残差模块的输入维度和输出维度保持一致，可先用一个1×1的卷积核对输入特征的通道数进行降维，然后用3×3的卷积的进行特征变换，最后加一个1×1的卷积核提升特征维度至原始输入的特征维度，以减少模型参数，提升计算效率。若目标视频数据的输入是256×128×3，随机采样该段视频中的连续图像序列$R=[l_1,l_2,\cdots,l_T]$，T是从目标视频数据中采样的帧的数量。对于采样得到的关键帧图像，通过ResNet50网络提取图像特征，每张图像输入尺寸为256×128×3，从而可得到其通过网络后的输出特征维度为16×8×2048。通过全局平均池化层，对该帧图像进行池化操作，即计算16×8×2048的特征向量前两维图像矩阵的平均值，得到1×2048

维的特征向量。遍历图像序列$R=[1_1,1_2,\cdots,1_T]$，得到图像序列特征$F=[f_1,f_2,\cdots,f_T]$。为了实现视频图神经神经网络对节点信息传播与聚合，基于图像序列特征构建图结构。基本的图结构定义为$G=(V,E,A)$，其中，V代表视频图神经神经网络的节点集合$V=[v_1,v_2,\cdots,v_T]$，其中$v_1,v_2,\cdots,v_T$代表节点特征，也即分别对应$f_1,f_2,\cdots,f_T$；E代表视频图神经神经网络的节点的连接边$E=[e_{ij},\cdots]$，$e_{ij}=(v_i,v_j)$。$A\in R^{T\times T}$代表邻接矩阵，其中每个元素代表A_{ij}表示节点(v_i,v_j)之间的关系。

[0113] 在获取到目标视频数据的多帧图像的图像特征之后，将这些图像特征作为视频图神经网络的图结构的输入，图神经网络的本质就是提取图结构数据的空间特征，聚合邻居节点的信息生成新的节点特征表示。为了提取更丰富的图像特征，可采用累加多层的图神经网络实现节点信息的传播和汇聚，也即本申请的视频图神经网络包括多层，如三层，每一层图神经网络的结构均相同，如图6所示，前一层图神经网络的输出是后一层图神经网络的输入，最后一层图神经网络的输出是整个视频图神经网络的输出。对于每一层图神经网络，均包括当前层图结构网络、与当前层图结构网络相连的归一化层(也即BN层)以及激活层；视频图神经网络的各层图结构网络的神经输入特征图和神经输出特征图跳跃连接；经跳跃连接所得特征图与归一化层的归一输出特征图的特征相加和为激活层的输入。通过将对应特征图矩阵直接相加后经过非线性激活层LeakyReLU传入下一层，可以提取更准确的特征表达。每层图结构网络为基于将图像序列特征的每个图像特征作为节点特征、并由图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系，确定视频图神经网络每层的图结构网络。

[0114] 利用本实施例所提供的图像特征提取模型进行图像特征提取，有利于提升互检索效率；视频图神经网络为叠加多层结构的网络模型，可以挖掘同一视频的不同帧之间的互补信息，有利于进一步提升模型精度，提升互检索精准度。

[0115] 上述实施例对如何确定视频图神经网络中图结构的各节点之间是否具有边连接关系，并不做任何限定，基于此，本实施例还提供了边连接关系的一种可选的确定方式，可包括下述内容：

[0116] 对图像序列特征的每个图像特征，依次计算当前图像特征与其余各图像特征之间的相似度；

[0117] 若当前节点的图像特征与目标节点的图像特征的相似度满足相似度条件，则当前节点与目标节点具有连接关系；若当前节点的图像特征与目标节点的图像特征的相似度不满足相似度条件，则当前节点与目标节点无连接关系。

[0118] 其中，可通过调用下述关系式计算每两个图像特征之间的特征相似度：

[0119] $$d_{i,j}=\frac{\sum_{k=1}^{K}v_{ik}v_{jk}}{\sqrt{\sum_{k=1}^{K}v_{ik}^{2}}\sqrt{\sum_{k=1}^{K}v_{jk}^{2}}}$$

；

[0120] 其中，v_{ik}代表节点特征向量v_i的第k个元素，该特征向量共有K个元素组成，如$K=128$。同理，v_{jk}代表节点特征向量v_j的第k个元素。

[0121] 本实施例的相似度条件基于相似度的计算方式和实际互检索精度需求来确定，例

如相似度计算方式为通过余弦相似度方式计算，而余弦相似度的值越接近1表示其越相似，相似度条件可为两节点的相似度值大于0.98。此外，还可对当前节点如节点v_i与其他每个节点之间的特征相似度先升序排序，形成当前节点如v_i节点的前k个最相似的节点集合S，并与其最近的邻居相连形成边。

[0122] 进一步的，为了确定节点之间的关联关系，还可考虑不同节点的重要性，对各个相连的边赋予权重信息，基于此，在确定当前节点与目标节点具有连接关系之后，还可包括：

[0123] 调用边权重关系式，计算每两个节点之间的权重值，并基于各权重值生成邻接关系矩阵；边权重关系式：

[0124] $$A_{ij}=\begin{cases}1-\dfrac{rank(v_i,v_j)}{T}, v_j\in V\\ 0 \quad ,\text{其它}\end{cases};$$

[0125] 其中，A_{ij}为所述邻接关系矩阵A的元素，T为所述邻接关系矩阵的维度，v_i为第i个节点，v_j为第j个节点，V为图像序列特征集合，$rank(v_i,v_j)$为节点v_j在节点v_i与所有节点相似程度排序中的排序值，也即用于表示节点v_j在节点v_i的第几个相似。邻接关系矩阵用于表示每两个节点之间的关联关系，构造的视频图神经网络的图结构可以用一个邻接矩阵A反映任意两个节点之间的关系。若$A_{ij}=0$则表示节点v_j和节点v_i之间没有连接。

[0126] 上述实施例对如何使用视频图神经网络获取目标频帧间的互补信息，得到更鲁棒性的视频特征表示，并没有进行任何限定，基于此，本申请还给出利用视频图神经网络获取目标视频数据的待匹配视频特征的一种可选的实施方式，包括：

[0127] 对视频图神经网络的每一层图结构网络，根据当前层图结构网络的图像特征、各节点之间的关联关系、当前层图结构网络的网络参数，更新当前层图神经网络的图像特征；

[0128] 将更新后的视频图神经网络的每一层图结构网络的图像特征，作为目标视频数据的待匹配视频特征。

[0129] 视频图神经网络为多层结构，为了便于描述，不引起歧义，每一层可称为图神经网络，每一层图神经网络包括图结构网络、与图结构网络相连的归一化层以及激活层。通过视频图神经网络获取目标视频数据的待匹配视频特征，其实就是计算图结构数据，图结构数据的计算是对某一个顶点和其邻居顶点加权求和的过程，本领域技术人员可根据实际情况选择任何一种图结构计算方法来提取图结构特征，这均不影响本申请的实现。可选的，本实施例还可通过调用视频特征更新关系式，更新视频图神经网络的各层图神经网络的图像特征；视频特征更新关系式可表示为：

[0130] $$Z^{(l)_g}=\sigma(\tilde{D}_{qq}^{-\frac{1}{2}}\tilde{A}_{qm}\tilde{D}_{qq}^{-\frac{1}{2}}Z^{(l)}W^{(l)});$$

[0131] 式中，$Z^{(1)g}$为所述视频图神经网络更新后的第1层图神经网络的图像特征，$Z^{(1)}$为所述视频图神经网络的第1层图神经网络的图像特征，σ为超参数，$W^{(1)}$为所述视频图神经网络的第1层图结构网络的网络参数，$\tilde{A}_{qm}$为邻接关系矩阵的变换矩阵，$\tilde{A}_{qm}=A+I$，A为邻接关系矩阵，I为单位矩阵，$\tilde{D}_{qq}$为对角矩阵，$\tilde{D}_{qq}=\sum_m \tilde{A}_{qm}$，q、m为矩阵维数。

[0132] 在确定节点特征之后，目标视频数据的视频特征可通过计算所有节点特征的均值

获取，也即可调用下述关系式确定最终视频特征e_{video}：

[0133] $$e_{video}=\frac{1}{T}\sum_{i\in[1,T]}v_i \;;V=[v_1,v_2,\cdots,v_T]。$$

[0134] 最后，为了使所属领域技术人员更加清楚明白本申请的实施方式，本实施例还提供了一个实现视频文本互检索的示意性的例子，其所依赖的硬件系统如图7所示，可包括通过网络相连的第一电子设备71和第二电子设备72，该示意性例子用于实现菜谱文本与菜谱视频的互检索任务，相应的，第一电子设备71可为菜谱检索终端设备，第二电子设备72可为菜谱服务器，用户可以在菜谱检索终端设备上执行人机交互操作，菜谱检索终端设备通过网络实现与菜谱服务器的交互，菜谱服务器可以部署如图8所示的互检索模型，基于该硬件系统，菜谱视频和菜谱文本的互检索任务执行过程可包括下述内容：

[0135] 为了实现菜谱文本与菜谱视频互检索的功能，菜谱服务器需要首先对互检索模型进行训练。在训练过程中，可以由菜谱检索终端设备向菜谱服务器传输训练样本集，训练样本集可包含有多组训练样本，每组训练样本包括相对应的一个菜谱文本样本和一个菜谱视频样本，每个菜谱文本样本包括操作步骤(instruction list)、成分信息(ingredients)和菜名(Title)。Instructions为做菜的步骤，在下文中统一用步骤表示。Ingredients为菜的成分，在下文统一用成分表示。

[0136] 服务器在获取到训练样本集后，分别对菜谱文本和菜谱视频进行特征编码。本实施例可采用文本图神经网络对文本信息进行编码。本实施例将文本特征构建成一种图结构，图结构包括节点及节点特征和连接关系，如图2所示。成分和步骤从构造到性质都是不同的，所以称为异质节点。本实施例中每一个步骤称为1个节点，同理每1个成分称为1个节点。节点是由1句话或者1个词组组成，本实施例可使用如图9所示Bert模型提取每句话或每个单词的特征，实现方式如下：所有菜谱文本从最下方的文本信息输入，同时还会输入与菜谱文本信息相伴随的位置信息和文本类型。位置信息是指若一句话中有5个单词“peel and slice the mango”，则其位置信息分别为“1,2,3,4,5”。文本类型是指：若输入文本是步骤，其文本类型为1；若输入文本是成分，其文本类型为2。通过该Bert模型，可以获得每句话和每个单词的编码特征，该特征用于代表节点特征，即成分节点特征和步骤节点特征，成分节点特征和步骤节点特征都是一个高维向量，其维度均为$\mathbb{R}^d$维度(d维实向量)。在确定节点特征之后，如果该主成分存在该操作步骤中，则该成分节点和步骤节点需要有一条边连接，也即两个节点之间具有连接关系。可选的，可通过文本比对的方法，遍历步骤信息，提取每个步骤文本，然后依次查找主成分，如果该主成分中的单词在该步骤中出现，则该步骤和该主成分之间连接一条边即有连接关系。通过遍历所有步骤文本，可以构建步骤节点预成分节点的连接关系，即异质图的连接关系。在异质图建立之后，异质图信息更新可采用图注意力网络实现特征聚合与更新，更新方法是依次遍历每个异质节点进行更新。通过图运算来实现文本特征的聚合与提取，计算方法可如下所示：

[0137] 首先对步骤节点进行更新，h_q^{ins}是步骤节点的第q个节点的节点特征，h_p^{ing}代表成分节点的第p个节点的特征。若步骤节点的第q个节点与成分节点的第p个节点有连接(边)，则用成分节点的第p个节点的特征去更新步骤节点的第q个节点特征。在更新过程中，需要考虑各节点之间的相关性，本实施例可通过赋予权重来表示节点间的关联性，可选的，可调用

下述关系式(1)计算步骤节点的第q个节点与成分节点的第p个节点特征的相关权重z_{qp}。对于每个步骤节点，例如h_q^{ins}，遍历所有与其有相连的边的成分节点，假设有N_p个，都会得到与其对应的相关权重z_{qp}。

[0138] $$z_{qp} = LeakyReLU(W_c[W_a h_q^{ins}; W_b h_p^{ing}]) \quad (1)$$

[0139] 其中，W_a、W_b、W_c为已知的$\mathbb{R}^{d\times d}$维矩阵，$W_a h_q^{ins}$代表矩阵乘法，也即向量映射。

[0140] 在更新完各步骤节点之后，可对所有与步骤节点相连的边的成分节点进行相关权重的归一化，也即可调用下述关系式(2)得到归一化的相关权重α_{qp}：

[0141] $$\alpha_{qp} = \frac{\exp(z_{qp})}{\sum_{l\in N_P}\exp(z_{ql})} \quad (2)$$

[0142] 式中，exp代表求指数函数，$\sum_{l\in N_P}\exp(z_{ql})$代表求取所有与步骤节点相连的边的成分节点的相关权重的总和。最后通过归一化的相关权重对步骤节点的节点特征进行更新，也即调用下述关系式(3)进行计算：

[0143] $$h_q^{ins} = \sigma\left(\sum_{p\in N_P}\alpha_{qp} W_v h_p^{ing}\right) \quad (3)$$

[0144] 其中，σ代表超参数，在[0,1]区间。W_v是$\mathbb{R}^{d\times d}$维矩阵，h_q^{ins}是被与其相连的成分节点更新后的新的特征向量。

[0145] 进一步，基于残差网络的思想，调用下述关系式(4)可将更新后的h_q^{ins}与未更前的初始特征h_q^{ins}相加：

[0146] $$h_q^{ins} = \sigma\left(\sum_{p\in N_P}\alpha_{qp}^k W_v^k h_p^{ing}\right) + h_q^{ins} \quad (4)$$

[0147] 同理，可调用关系式(5)对成分节点也做相同的计算与更新：

[0148] $$h_p^{ing} = \sigma\left(\sum_{q\in N_Q}\alpha_{qp}^k W_v^k h_q^{ins}\right) + h_p^{ing} \quad (5)$$

[0149] 遍历完所有的成分节点和步骤节点，即完成图注意力网络一层的网络更新。通常，可叠加T层图注意力网络，用t代表第t层的图注意力网络，每一层的节点特征的更新方式都如上所述。通常会在每层图注意力网络后面加入集成全连接层，实现对节点特征(包括成分节点和步骤节点)特征的再编码，如下述关系式(6)所示：

[0150] $$\langle h_q^{ins}\rangle^{t+1} = FFN(\langle h_q^{ins}\rangle^{t})$$

[0151] $$\langle h_p^{ing}\rangle^{t+1} = FFN(\langle h_p^{ing}\rangle^{t}) \quad (6)$$

[0152] FFN代表全连接层，$\langle h_p^{ing}\rangle^{t+1}$、$\langle h_q^{ins}\rangle^{t+1}$代表t+1层的图注意力网络的初始化节点特征。

[0153] 如上完成了对本节点特征的更新，为了实现与菜谱视频的检索，还需要将所有文字节点的特征如操作步骤、成分信息和菜名进行归纳和综合。在本实施例中，由于步骤节点融合了成分节点信息，成分节点通过文本图神经网络更新，以关键词的形式对相关步骤节点特征进行了强调。同时，由于菜名信息中包含重要的主材信息和烹饪手段，同时，菜名文本在基于菜谱的图文互检任务中通常是一个广泛的存在。基于此，本实施例还可通过Bert模型提取菜名的特征。在获取各文本特征之后，可采用BiLSTM(双向长短期记忆神经网络)方法进一步挖掘步骤节点的时序信息，实现对文字节点特征的归纳综合，并将其打包成一个向量。

[0154] 本实施例可调用下述关系式(7)和(8)提取所有步骤节点的时序信息特征：

[0155] $$\overrightarrow{\hbar_q} = \overrightarrow{LSTM}(\langle h_q^{ins}\rangle^T, \overrightarrow{\hbar_{q-1}}), q \in [1, Q] \tag{7}$$

[0156] $$\overleftarrow{\hbar_q} = \overleftarrow{LSTM}(\langle h_q^{ins}\rangle^T, \overleftarrow{\hbar_{q+1}}), q \in [Q, 1] \tag{8}$$

[0157] 其中，向左和向右的箭头代表LSTM编码的方向，即步骤节点特征正序编码和倒序编码。$\hbar_q$代表BiLSTM中第q个单元的输出，箭头方向不同代表按照步骤节点输入顺序不同得到的BiLSTM编码输出。同理，$\hbar_{q-1}$则代表BiLSTM中第q-1个单元的输出，也即上一个状态的输出。假设菜谱步骤共有Q步，$\hbar_0$为0，$\langle h_q^{ins}\rangle^T$代表第T层的图神经网络的第q个步骤节点的特征。按照步骤的顺序和逆序，依次输入到其对应的BiLSTM网络中，最后得到所有步骤节点的BiLSTM编码，如下述关系式(9)所示：

[0158] $$e_{rec} = \left(\sum_{q=1}^{Q} \overrightarrow{\hbar_q} + \sum_{q=1}^{Q} \overleftarrow{\hbar_q}\right) / 2Q \tag{9}$$

[0159] 在获取所有BiLSTM单元的输出之后，可通过求和后取平均值得到整个文本特征的输出。其中，e_{rec}代表文本特征的输出，用来进行下一步的检索。将e_{rec}特征与菜名title特征进行融合$e_{rec}=[e_{rec}, e_{ttl}]$，[ ]代表特征拼接，即特征首尾相连。$e_{rec}$特征最后会经过一个全连接层进行特征映射，也即$e_{rec}=fc(e_{rec})$，得到新维度的向量，也即菜谱文本的文本特征信息，其用于作为与菜谱视频的编码特征进行匹配。

[0160] 对于菜谱视频的编码过程，可将样本视频作为菜谱视频，提取菜谱视频的全部图像帧输入至图像特征提取模型接到菜谱图像序列特征，基于菜谱图像序列特征作为视频图神经网络中的图结构的输入，学习不同帧特征之间的互补特征，建立不同视频帧间的潜在联系，最终得到菜谱视频特征。可采用上述任意一个实施例实现基于视频图神经网络生成菜谱视频特征，此处，便不再赘述。再得到训练样本集的每组训练样本的菜谱视频特征和菜谱文本特征信息之后，可采用上述实施例的损失函数指导视频文本互检模型的训练，使其收敛。

[0161] 菜谱检索终端设备可以包括显示屏、输入接口、输入键盘、无线传输模块。当显示屏为触摸屏时，输入键盘可以是在显示屏上呈现的软键盘。输入接口可以用于实现与外部设备如U盘的连接。输入接口可以有多个。在实际应用中，用户可以通过输入键盘向菜谱检索终端设备输入待检索菜谱文本或待检索视频，也可以将待检索菜谱文本或待检索视频写入U盘，将U盘插入菜谱检索终端设备的输入接口。用户向菜谱检索终端设备输入检索请求，检索请求携带待检索的菜谱文本或待检索的菜谱视频，菜谱检索终端可以通过无线传输模块向菜谱服务器发送该检索请求，菜谱服务器基于训练好的互检索模型检索相应的数据库，以将最终确定的目标菜谱视频或目标菜谱文本反馈至菜谱检索终端设备，菜谱检索终端设备可以通过显示屏向用户展示所检索到的目标菜谱视频或目标菜谱文本。

[0162] 本发明实施例还针对用于视频数据与文本数据之间互检索的模型训练方法以及视频数据与文本数据之间的互检索方法提供了相应的装置，进一步使得方法更具有实用性。其中，装置可从功能模块的角度和硬件的角度分别说明。下面对本发明实施例提供的用于视频数据与文本数据之间互检索的模型训练装置以及视频数据与文本数据之间的互检索装置进行介绍，下文描述的视频数据与文本数据之间的互检索装置与上文描述用于视频数据与文本数据之间互检索的模型训练方法以及视频数据与文本数据之间的互检索方法可相互对应参照。

[0163] 基于功能模块的角度，首先参见图10，图10为本发明实施例提供的用于视频数据与文本数据之间互检索的模型训练装置在一种具体实施方式下的结构图，该装置可包括：

[0164] 文本图神经网络生成模块101，用于通过将当前样本文本数据的第一类文本数据对应的各第一类节点特征、第二类文本数据对应的各第二类节点特征分别作为节点特征，各第一类节点特征与各第二类节点特征之间的包含关系作为连接关系，生成文本图神经网络；第二类文本数据包括第一类文本数据；训练样本集包括多组训练样本，每组训练样本均包括样本文本数据和对应的样本视频数据。

[0165] 视频图神经网络生成模块102，用于基于将当前样本文本数据对应的目标样本视频数据的图像序列特征中的每帧图像特征作为节点特征、并由图像序列特征中每帧图像特征与其余各帧图像特征之间相关性所确定的边连接关系，生成视频图神经网络；

[0166] 模型训练模块103，用于利用包括第三类文本数据对应的文本特征、以及由文本图神经网络提取第二类文本数据所得文本特征的样本文本特征、视频图神经网络提取的样本视频特征，训练互检索模型；互检索模型包括文本图神经网络和视频图神经网络。第三类文本数据用于概括第一类文本数据和第二类文本数据。

[0167] 其次，请参见图11，图11为本发明实施例提供的视频数据与文本数据之间的互检索装置在一种具体实施方式下的结构图，该装置可包括：

[0168] 文本特征提取模块，用于提取目标文本数据的待匹配文本特征；目标文本数据包括第一类文本数据、第二类文本数据和第三类文本数据，且所述第二类文本数据包含所述第一类文本数据，第三类文本数据用于概括第一类文本数据和第二类文本数据；待匹配文本特征包括第三类文本数据对应的文本特征、和利用互检索模型的文本图神经网络提取第二类文本数据的文本特征；

[0169] 视频特征提取模块，用于提取目标视频数据的待匹配视频特征；

[0170] 互检索模块，用于基于待匹配视频特征和待匹配文本特征，调用所述互检索模型

生成目标文本数据和目标视频数据的检索结果;其中,互检索模型利用前任意一个实施例中用于视频数据与文本数据之间互检索的模型训练方法训练所得。

[0171] 可选的,作为本实施例的一种可选的实施方式,上述视频特征提取模块还可用于:通过提取目标视频数据的多帧图像的图像特征,生成目标视频数据的图像序列特征;基于将图像序列特征的每个图像特征作为节点特征、并由图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系,生成视频图神经网络;利用视频图神经网络,获取目标视频数据的待匹配视频特征。

[0172] 作为上述实施例的一种可选的实施方式,视频图神经网络包括多层,每一层均包括当前层图结构网络、与当前层图结构网络相连的归一化层以及激活层;视频图神经网络的各层图结构网络的神经输入特征图和神经输出特征图跳跃连接;经跳跃连接所得特征图与归一化层的归一输出特征图的特征相加和为激活层的输入;其中,基于将图像序列特征的每个图像特征作为节点特征、并由图像序列特征中每个图像特征与其余各图像特征之间相关性所确定的边连接关系,确定视频图神经网络每层的图结构网络。

[0173] 作为上述实施例的另一种可选的实施方式,上述视频特征提取模块可包括特征提取单元,该单元用于:预先训练图像特征提取模型;图像特征提取模型包含第一预设个数的卷积层和第二预设个数的残差模块,每个残差模块均包含多层卷积层、归一化层和ReLU激活函数层;将目标视频数据的多帧图像输入至图像特征提取模型,得到每帧图像的图像特征;根据各帧图像的图像特征,生成目标视频数据的图像序列特征。

[0174] 作为上述实施例的再一种可选的实施方式,上述视频特征提取模块还可包括图像提取单元,该单元用于:接收图像提取指令,通过解析图像提取指令获取图像提取规则;按照图像提取规则,从目标视频数据中提取相应帧图像。

[0175] 可选的,作为本实施例的另一种可选的实施方式,上述视频特征提取模块可进一步用于:对图像序列特征的每个图像特征,依次计算当前图像特征与其余各图像特征之间的相似度;若当前节点的图像特征与目标节点的图像特征的相似度满足相似度条件,则当前节点与目标节点具有连接关系;若当前节点的图像特征与目标节点的图像特征的相似度不满足相似度条件,则当前节点与目标节点无连接关系。

[0176] 作为上述实施例的一种可选的实施方式,上述视频特征提取模块还可进一步用于:调用边权重关系式,计算每两个节点之间的权重值,并基于各权重值生成邻接关系矩阵;边权重关系式:

[0177] $$A_{ij}=\begin{cases}1-\dfrac{rank(v_i,v_j)}{T}, v_j\in V\\ 0 \quad \text{,其它}\end{cases};$$

[0178] 其中,A_{ij}为所述邻接关系矩阵A的元素,T为所述邻接关系矩阵的维度,v_i为第i个节点,v_j为第j个节点,V为图像序列特征集合,$rank(v_i,v_j)$为节点v_j在v_i与所有节点相似程度排序中的排序值;所述邻接关系矩阵用于表示每两个节点之间的关联关系。

[0179] 可选的,作为本实施例的其他一种可选的实施方式,上述视频特征提取模块还可进一步包括特征更新单元,该单元用于对视频图神经网络的每一层图结构网络,根据当前层图结构网络的图像特征、各节点之间的关联关系、当前层图结构网络的网络参数,更新当

前层图神经网络的图像特征;将更新后的视频图神经网络的每一层图结构网络的图像特征,作为目标视频数据的待匹配视频特征。

[0180] 作为上述实施例的一种可选的实施方式,上述特征更新单元还可进一步用于:调用视频特征更新关系式,更新视频图神经网络的各层图神经网络的图像特征;视频特征更新关系式为:

[0181] $$Z^{(l)_g}=\sigma(\tilde{D}_{qq}^{-\frac{1}{2}}\tilde{A}_{qm}\tilde{D}_{qq}^{-\frac{1}{2}}Z^{(l)}W^{(l)})$$;

[0182] 式中,$Z^{(1)g}$为所述视频图神经网络更新后的第1层图神经网络的图像特征,$Z^{(1)}$为所述视频图神经网络的第1层图神经网络的图像特征,σ为超参数,$W^{(1)}$为所述视频图神经网络的第1层图结构网络的网络参数,$\tilde{A}_{qm}$为邻接关系矩阵的变换矩阵,$\tilde{A}_{qm}=A+I$,A为邻接关系矩阵,I为单位矩阵,$\tilde{D}_{qq}$为对角矩阵,$\tilde{D}_{qq}=\sum_{m}\tilde{A}_{qm}$,q、m为矩阵维数。

[0183] 本发明实施例用于视频数据与文本数据之间互检索的模型训练装置以及视频数据与文本数据之间的互检索装置的各功能模块的功能可根据上述方法实施例中的方法具体实现,其具体实现过程可以参照上述方法实施例的相关描述,此处不再赘述。

[0184] 由上可知,本发明实施例可有效提升视频数据和文本数据之间的互检索精度。

[0185] 上文中提到的用于视频数据与文本数据之间互检索的模型训练装置以及视频数据与文本数据之间的互检索装置是从功能模块的角度描述,进一步的,本申请还提供一种互检索设备,是从硬件角度描述。图12为本申请实施例提供的互检索设备在一种实施方式下的结构示意图。如图12所示,该互检索设备包括存储器120,用于存储计算机程序;处理器121,用于执行存储器中存储的计算机程序时实现如前任意一个实施例所述的用于视频数据与文本数据之间互检索的模型训练方法和/或如前任意一个实施例所述的视频数据与文本数据之间的互检索方法的步骤;人机交互组件122用于通过信息输入/信息输出接口,接收用户输入的训练样本集选择请求、模型训练请求、检索请求以及向用户显示目标文本数据和目标视频数据的检索结果;通信组件123用于传输互检索模型训练过程以及目标文本数据和目标视频数据的互检索任务执行过程中的数据及指令。

[0186] 其中,处理器121可以包括一个或多个处理核心,比如4核心处理器、8核心处理器,处理器121还可为控制器、微控制器、微处理器或其他数据处理芯片等。处理器121可以采用DSP(Digital Signal Processing,数字信号处理)、FPGA(Field-Programmable Gate Array,现场可编程门阵列)、PLA(Programmable Logic Array,可编程逻辑阵列)中的至少一种硬件形式来实现。处理器121也可以包括主处理器和协处理器,主处理器是用于对在唤醒状态下的数据进行处理的处理器,也称CPU(Central Processing Unit,中央处理器);协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中,处理器121可以集成有GPU(Graphics Processing Unit,图像处理器),GPU用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中,处理器121还可以包括AI(Artificial Intelligence,人工智能)处理器,该AI处理器用于处理有关机器学习的计算操作。

[0187] 存储器120可以包括一个或多个计算机可读存储介质,该计算机可读存储介质可以是非暂态的。存储器120还可包括高速随机存取存储器以及非易失性存储器,比如一个或

多个磁盘存储设备、闪存存储设备。存储器120在一些实施例中可以是互检索设备的内部存储单元，例如服务器的硬盘。存储器120在另一些实施例中也可以是互检索设备的外部存储设备，例如服务器上配备的插接式硬盘，智能存储卡(Smart Media Card，SMC)，安全数字(Secure Digital，SD)卡，闪存卡(Flash Card)等。进一步地，存储器120还可以既包括互检索设备的内部存储单元也包括外部存储设备。存储器120不仅可以用于存储安装于互检索设备的应用软件及各类数据，例如：执行用于视频数据与文本数据之间互检索的模型训练过程中以及视频数据与文本数据之间的互检索过程中的程序的代码等，还可以用于暂时地存储已经输出好的互检索模型以及互检索结果或者将要输出的数据。本实施例中，存储器120至少用于存储以下计算机程序1201，其中，该计算机程序被处理器121加载并执行之后，能够实现前述任一实施例公开的用于视频数据与文本数据之间互检索的模型训练方法以及视频数据与文本数据之间的互检索方法的相关步骤。另外，存储器120所存储的资源还可以包括操作系统1202和数据1203等，存储方式可以是短暂存储或者永久存储。其中，操作系统1202可以包括Windows、Unix、Linux等。数据1203可以包括但不限于用于视频数据与文本数据之间互检索的模型训练过程中以及视频数据与文本数据之间的互检索过程中所生成的数据以及检索结果、模型训练结果数据等。

[0188] 人机交互组件122可包括有显示屏、信息输入/信息输出接口如键盘或鼠标，显示屏、信息输入/信息输出接口属于用户接口，可选的用户接口还可以包括标准的有线接口、无线接口等。可选地，在一些实施例中，显示器可以是LED显示器、液晶显示器、触控式液晶显示器以及OLED(Organic Light-Emitting Diode，有机发光二极管)触摸器等。显示器也可以适当的称为显示屏或显示单元，用于显示在互检索设备中处理的信息以及用于显示可视化的用户界面。通信组件123可包括通信接口或者称为网络接口、通信总线等，通信接口可选的可以包括有线接口和/或无线接口，如WI-FI接口、蓝牙接口等，通常用于在互检索设备与其他互检索设备之间建立通信连接。通信总线可以是外设部件互连标准(peripheral component interconnect，简称PCI)总线或扩展工业标准结构(extended industry standard architecture，简称EISA)总线等。该总线可以分为地址总线、数据总线、控制总线等。为便于表示，图12中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。在一些实施例中，上述互检索设备还可包括电源124以及实现各类功能的传感器125。本领域技术人员可以理解，图12中示出的结构并不构成对该互检索设备的限定，可以包括比图示更多或更少的组件。

[0189] 本发明实施例互检索设备的各功能模块的功能可根据上述方法实施例中的方法具体实现，其具体实现过程可以参照上述方法实施例的相关描述，此处不再赘述。

[0190] 由上可知，本发明实施例可有效提升视频数据和文本数据之间的互检索精度。

[0191] 可以理解的是，如果上述实施例中的用于视频数据与文本数据之间互检索的模型训练方法以及视频数据与文本数据之间的互检索方法以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，执行本申请各个实施例方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、电可擦除可编

程ROM、寄存器、硬盘、多媒体卡、卡型存储器(例如SD或DX存储器等)、磁性存储器、可移动磁盘、CD-ROM、磁碟或者光盘等各种可以存储程序代码的介质。

[0192] 基于此,本发明实施例还提供了一种可读存储介质,存储有计算机程序,计算机程序被处理器执行时如上任意一实施例用于视频数据与文本数据之间互检索的模型训练方法以及视频数据与文本数据之间的互检索方法的步骤。

[0193] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其它实施例的不同之处,各个实施例之间相同或相似部分互相参见即可。对于实施例公开的硬件包括装置及电子设备而言,由于其与实施例公开的方法相对应,所以描述的比较简单,相关之处参见方法部分说明即可。

[0194] 专业人员还可以进一步意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本发明的范围。

[0195] 以上对本申请所提供的一种用于视频数据与文本数据之间互检索的模型训练方法及装置、视频数据与文本数据之间的互检索方法及装置、互检索设备、可读存储介质进行了详细介绍。本文中应用了具体个例对本发明的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本发明的方法及其核心思想。应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以对本申请进行若干改进和修饰,这些改进和修饰也落入本申请权利要求的保护范围内。

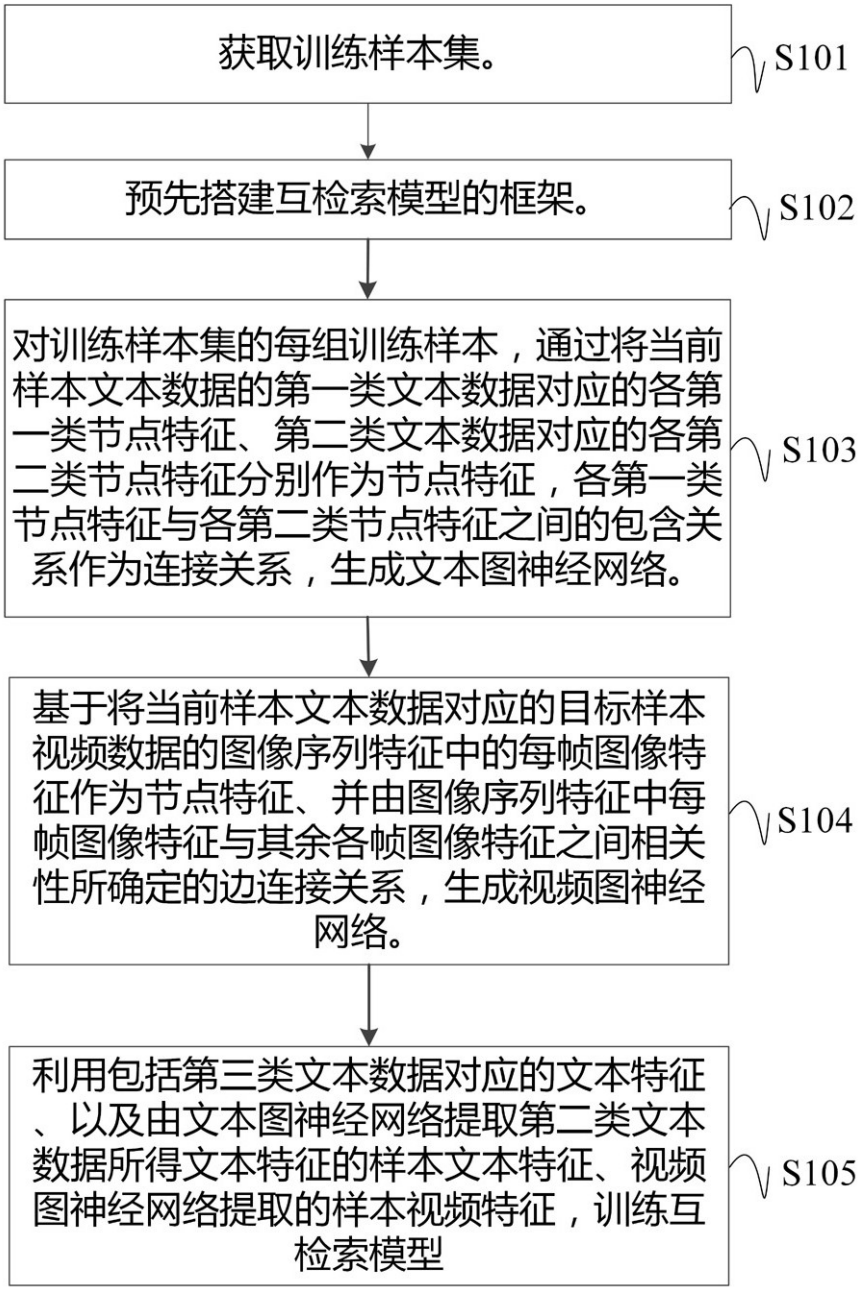

获取训练样本集。
S101
预先搭建互检索模型的框架。
S102
对训练样本集的每组训练样本，通过将当前样本文本数据的第一类文本数据对应的各第一类节点特征、第二类文本数据对应的各第二类节点特征分别作为节点特征，各第一类节点特征与各第二类节点特征之间的包含关系作为连接关系，生成文本图神经网络。
S103
基于将当前样本文本数据对应的目标样本视频数据的图像序列特征中的每帧图像特征作为节点特征、并由图像序列特征中每帧图像特征与其余各帧图像特征之间相关性所确定的边连接关系，生成视频图神经网络。
S104
利用包括第三类文本数据对应的文本特征、以及由文本图神经网络提取第二类文本数据所得文本特征的样本文本特征、视频图神经网络提取的样本视频特征，训练互检索模型
S105

图1

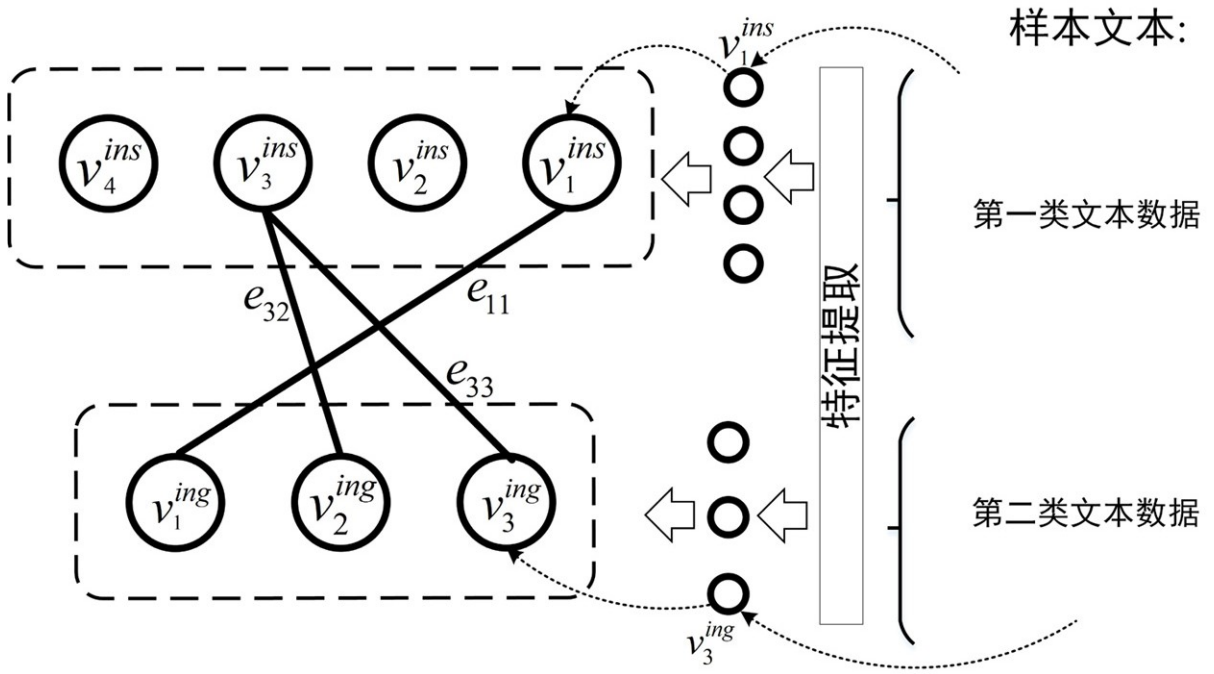

样本文本:
v_1^{ins}
v_4^{ins}
v_3^{ins}
v_2^{ins}
v_1^{ins}
e_{32}
e_{11}
e_{33}
v_1^{ing}
v_2^{ing}
v_3^{ing}
特征提取
第一类文本数据
第二类文本数据
v_3^{ing}

图2

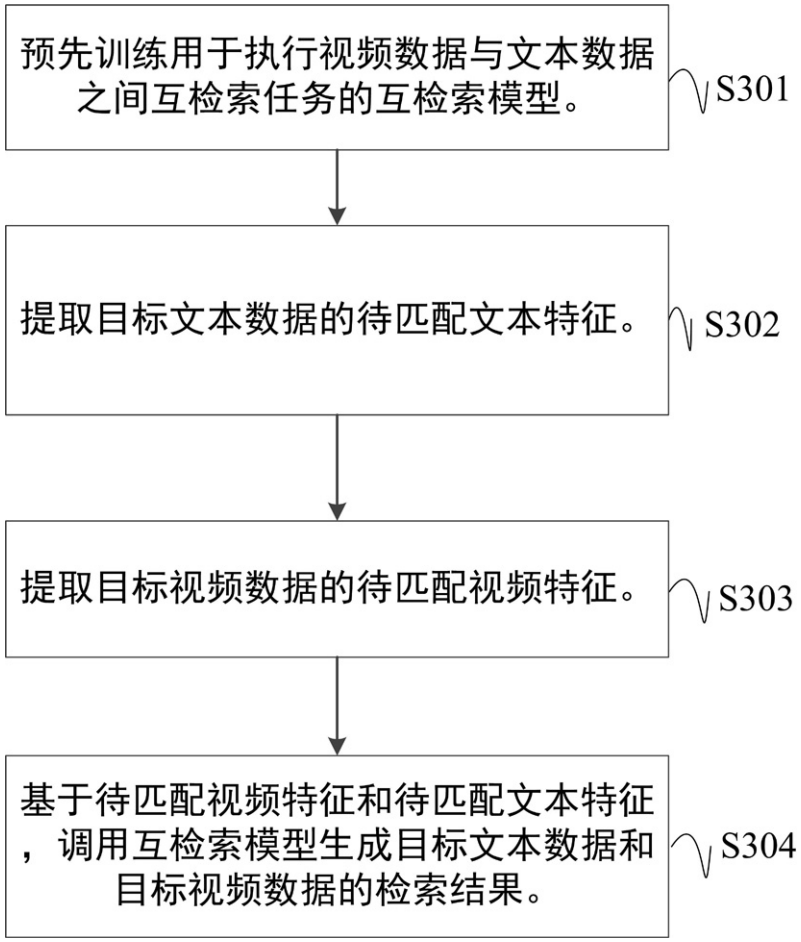

预先训练用于执行视频数据与文本数据之间互检索任务的互检索模型。
S301
提取目标文本数据的待匹配文本特征。
S302
提取目标视频数据的待匹配视频特征。
S303
基于待匹配视频特征和待匹配文本特征，调用互检索模型生成目标文本数据和目标视频数据的检索结果。
S304

图3

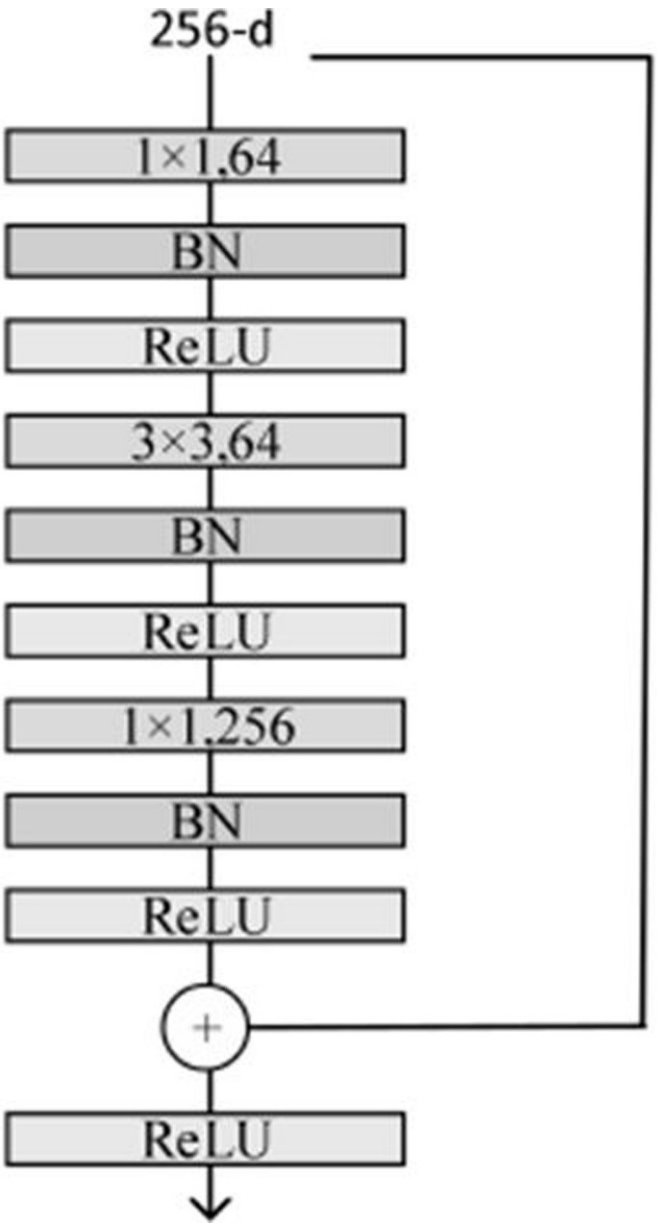


图4

网络层	卷积核参数	输出尺寸
Conv_1	7×7,64， stride 2	256×128×3
	3×3 Max pooling， stride 2	
Conv2_x	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	128×64×256
Conv3_x	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	64×32×512
Conv4_x	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	32×16×1024
Conv5_x	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	16×8×2048

图5

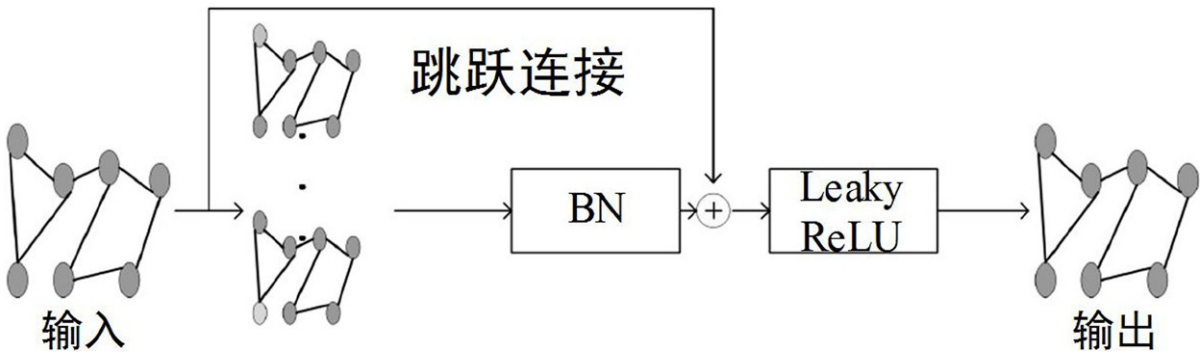


图6

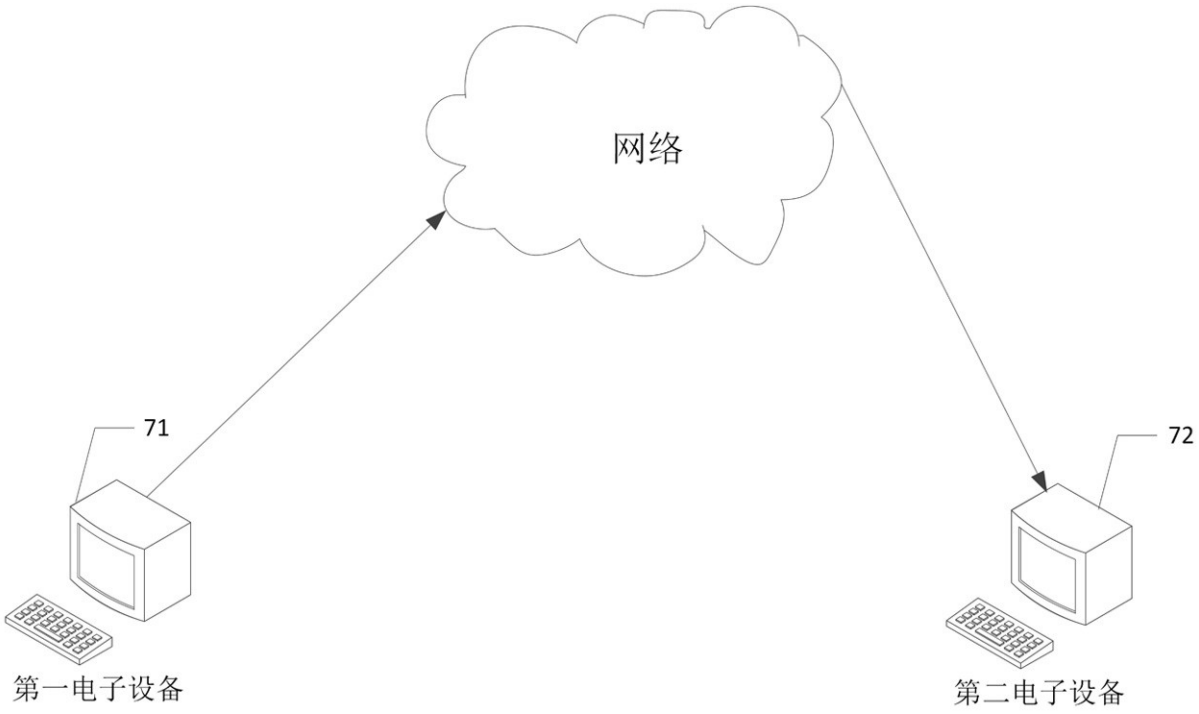

网络
71
72
第一电子设备
第二电子设备

图7

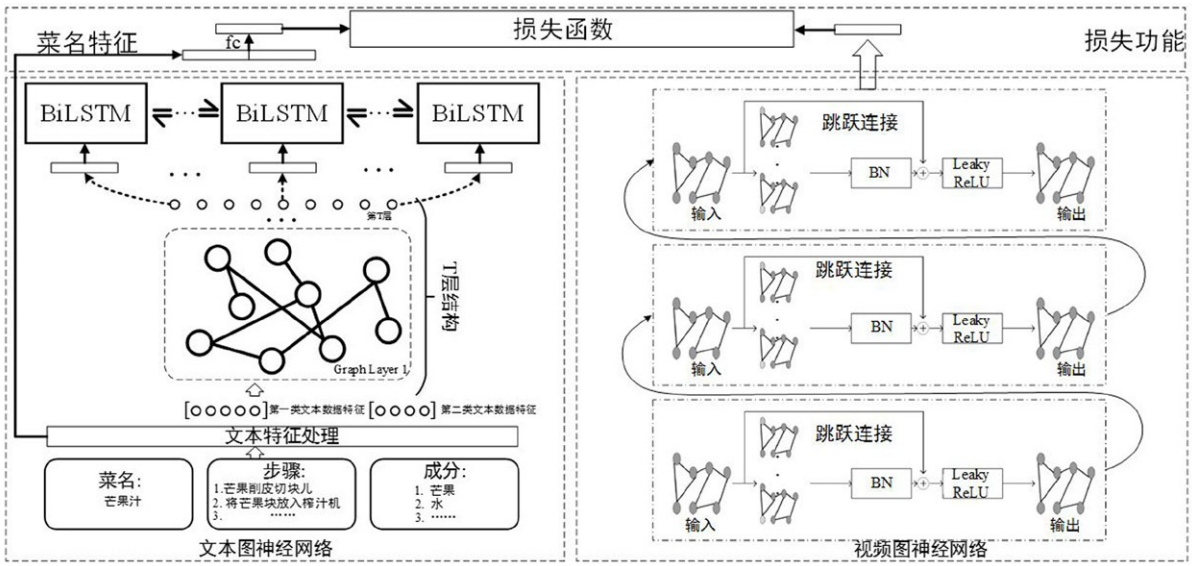

菜名特征
fc
损失函数
损失功能
BiLSTM
BiLSTM
BiLSTM
第T层
T层结构
Graph Layer 1
[○○○○○]第一类文本数据特征
[○○○○]第二类文本数据特征
文本特征处理
菜名:
芒果汁
步骤:
1.芒果削皮切块儿
2.将芒果块放入榨汁机
3.
……
成分:
1. 芒果
2. 水
3. ……
文本图神经网络
跳跃连接
BN
Leaky ReLU
输入
输出
跳跃连接
BN
Leaky ReLU
输入
输出
跳跃连接
BN
Leaky ReLU
输入
输出
视频图神经网络

图8

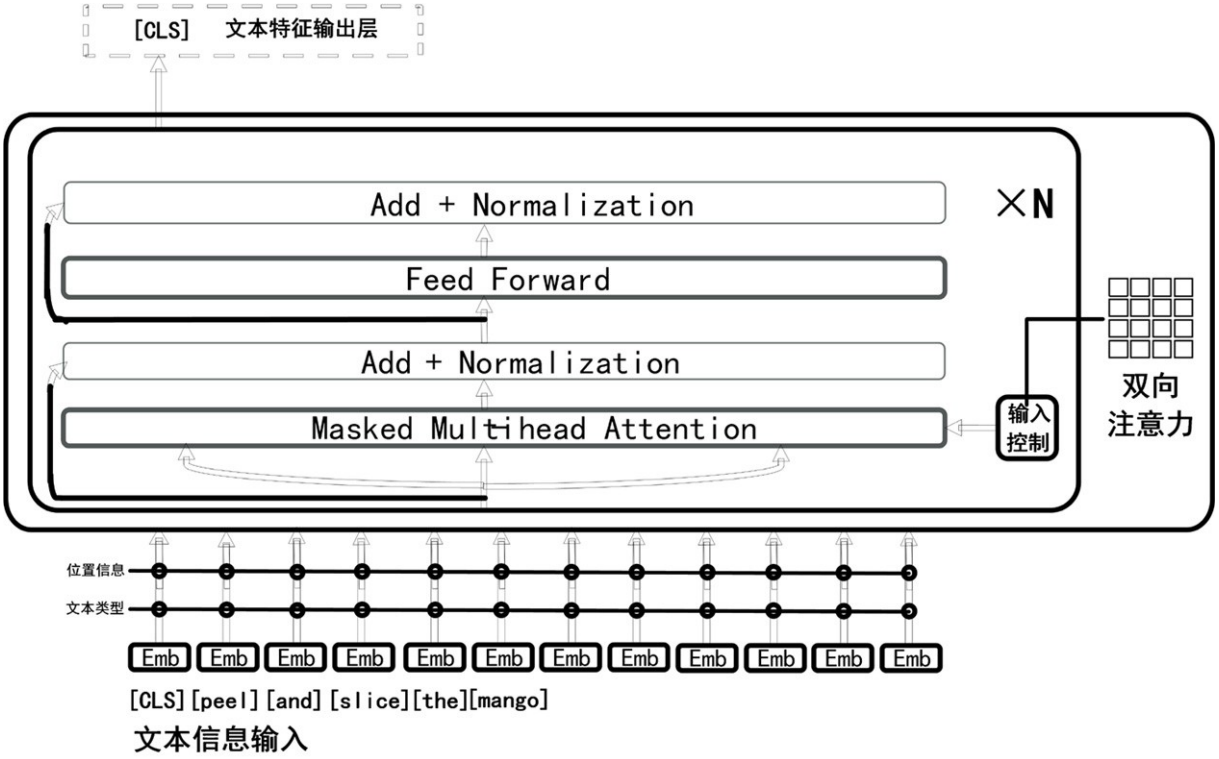

[CLS] 文本特征输出层
Add + Normalization
×N
Feed Forward
Add + Normalization
Masked Multihead Attention
输入控制
双向注意力
位置信息
文本类型
Emb
[CLS][peel][and][slice][the][mango]
文本信息输入

图9

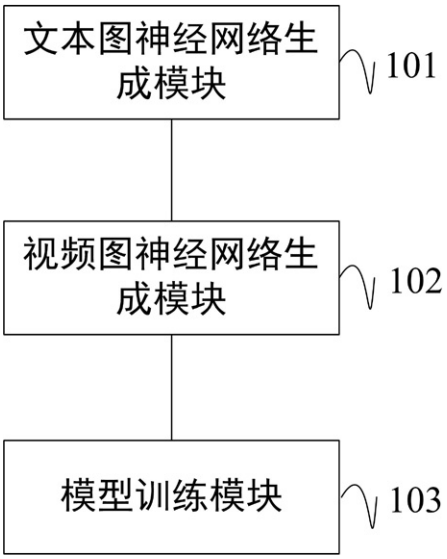

文本图神经网络生成模块
101
视频图神经网络生成模块
102
模型训练模块
103

图10

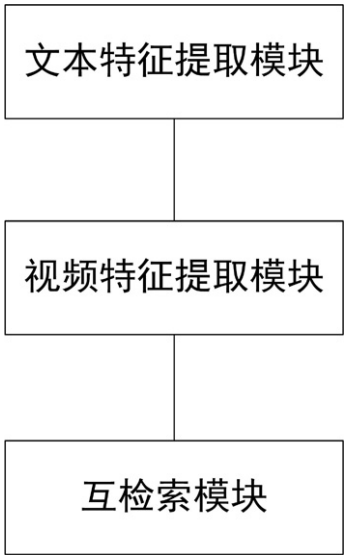

文本特征提取模块
视频特征提取模块
互检索模块

图11

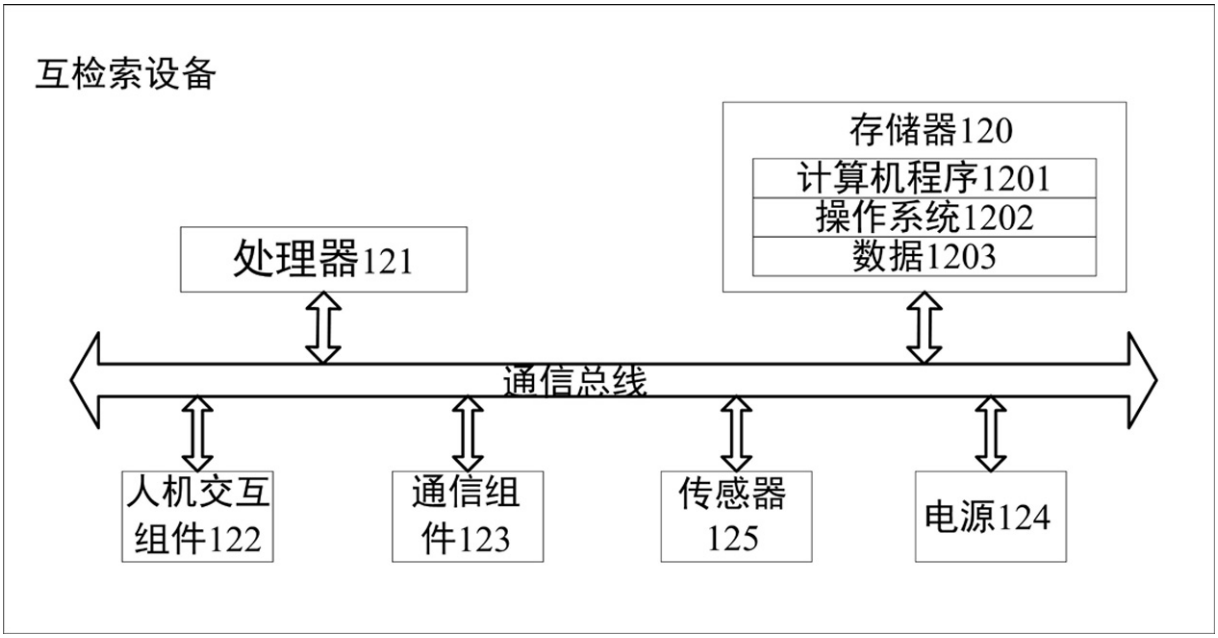

互检索设备
存储器120
计算机程序1201
操作系统1202
数据1203
处理器121
通信总线
人机交互组件122
通信组件123
传感器125
电源124

图12