(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 114241279 A
(43)申请公布日 2022.03.25

(21)申请号 202111651496.5
(22)申请日 2021.12.30
(71)申请人 中科讯飞互联(北京)信息科技有限公司
地址 100193 北京市海淀区西北旺东路10号院东区5号楼三层311-2
申请人 科大讯飞股份有限公司
(72)发明人 陈致鹏 崔一鸣 陈志刚
(74)专利代理机构 北京励诚知识产权代理有限公司 11647
代理人 周慧云
(51)Int.Cl.
G06V 10/80(2022.01)
G06F 16/56(2019.01)
G06F 16/36(2019.01)
G06F 16/33(2019.01)

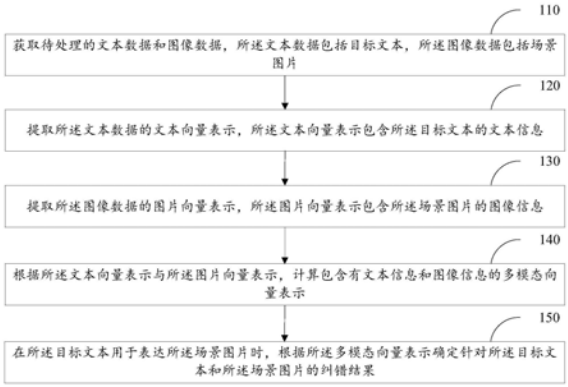
权利要求书3页 说明书15页 附图5页

(54)发明名称

图文联合纠错方法、装置、存储介质及计算机设备

(57)摘要

本申请公开一种图文联合纠错方法、装置、存储介质及计算机设备。该方法包括:获取待处理的文本数据和图像数据,文本数据包括目标文本,图像数据包括场景图片;提取文本数据的文本向量表示,文本向量表示包含目标文本的文本信息;提取图像数据的图片向量表示,图片向量表示包含场景图片的图像信息;根据文本向量表示与图片向量表示,计算包含有文本信息和图像信息的多模态向量表示;在目标文本用于表达场景图片时,根据多模态向量表示确定针对目标文本和场景图片的纠错结果,以根据纠错结果对图文进行纠错,实现了图文联合纠错,提升了纠错能力。



CN 114241279 A

1.一种图文联合纠错方法,其特征在于,所述方法包括:

获取待处理的文本数据和图像数据,所述文本数据包括目标文本,所述图像数据包括场景图片;

提取所述文本数据的文本向量表示,所述文本向量表示包含所述目标文本的文本信息;

提取所述图像数据的图片向量表示,所述图片向量表示包含所述场景图片的图像信息;

根据所述文本向量表示与所述图片向量表示,计算包含有文本信息和图像信息的多模态向量表示;

在所述目标文本用于表达所述场景图片时,根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果。

2.如权利要求1所述的图文联合纠错方法,其特征在于,所述提取所述文本数据的文本向量表示,包括:

通过词表将所述文本数据中的每个词转换成每个词在所述词表中对应的序号,并根据所述序号查找所述文本数据的文本向量表示。

3.如权利要求1所述的图文联合纠错方法,其特征在于,所述提取所述图像数据的图片向量表示,包括:

根据目标检测模型对所述场景图片进行目标检测以及特征提取,以得到所述图片向量表示,其中,所述图片向量表示包括所述场景图片中每个图像目标的图像信息向量表示和整个图片的图像信息向量表示。

4.如权利要求1所述的图文联合纠错方法,其特征在于,所述根据所述文本向量表示与所述图片向量表示,计算包含文本信息和图像信息的多模态向量表示,包括:

基于自注意力模型对所述文本向量表示与所述图片向量进行处理,获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息;

对所述全局交互信息进行归一化处理,得到第一归一化信息;

根据所述全局交互信息与所述第一归一化信息,确定包含文本信息和图像信息的多模态向量表示。

5.如权利要求4所述的图文联合纠错方法,其特征在于,所述将基于自注意力模型对所述文本向量表示与所述图片向量进行处理,获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息,包括:

将根据所述文本向量表示与所述图片向量表示确定的嵌入向量表示输入自注意力模型,根据所述嵌入向量表示与所述嵌入向量表示的转置矩阵之间的乘积,计算匹配矩阵;

根据所述匹配矩阵与所述嵌入向量表示的乘积,确定所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息。

6.如权利要求4所述的图文联合纠错方法,其特征在于,所述根据所述全局交互信息与所述第一归一化信息,确定包含文本信息和图像信息的多模态向量表示,包括:

将所述全局交互信息与所述第一归一化信息进行相加,以得到第一求和信息;

将所述第一求和信息输入全连接层进行处理后,对所述全连接层的输出结果进行归一化处理,得到第二归一化信息;

将所述第一求和信息与所述第二归一化信息进行相加，得到所述包含文本信息和图像信息的多模态向量表示。

7.如权利要求1所述的图文联合纠错方法，其特征在于，所述方法还包括：

获取位置向量表示和类型向量表示，所述位置向量表示用于标注所述文本数据中每个词的位置，所述类型向量表示用于区分文本类型和图像类型；

所述根据所述文本向量表示与所述图片向量表示，计算包含有文本信息和图像信息的多模态向量表示，包括：

根据所述文本向量表示、所述图片向量表示、所述位置向量表示和所述类型向量表示，计算包含有文本信息和图像信息的多模态向量表示。

8.如权利要求1-7任一项所述的图文联合纠错方法，其特征在于，所述根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果，包括：

将所述多模态向量表示连接一个全连接层，得到第一特征向量，所述第一特征向量包含有所述目标文本中每个词和所述场景图片中每个图像目标分别对应的向量表示；

根据所述第一特征向量，确定所述目标文本中每个词和所述场景图片中每个图像目标分别对应的纠错操作；

根据所述纠错操作和所述第一特征向量计算纠错结果，以根据所述纠错结果对所述目标文本和/或所述场景图片进行纠错处理。

9.如权利要求8所述的图文联合纠错方法，其特征在于，所述根据所述纠错操作和所述第一特征向量确定纠错结果，还包括：

还包括：

若所述纠错操作为无错，则确定所述纠错结果为输出与所述无错的纠错操作对应的词；或者

若所述纠错操作为删除操作，则确定所述纠错结果为将所述目标文本中与所述删除操作对应的词进行删除；或者

若所述纠错操作为修改操作，则确定所述纠错结果为将所述目标文本中与所述修改操作对应的词修改为预测词，或者将所述场景图片中与所述修改操作对应的图像目标修改为预测图像目标。

10.一种图文联合纠错装置，其特征在于，所述装置包括：

获取单元，用于获取待处理的文本数据和图像数据，所述文本数据包括目标文本，所述图像数据包括场景图片；

第一提取单元，用于提取所述文本数据的文本向量表示，所述文本向量表示包含所述目标文本的文本信息；

第二提取单元，用于提取所述图像数据的图片向量表示，所述图片向量表示包含所述场景图片的图像信息；

计算单元，用于根据所述文本向量表示与所述图片向量表示，计算包含有文本信息和图像信息的多模态向量表示；

确定单元，用于在所述目标文本用于表达所述场景图片时，根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果。

11.一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存储有计算机程

序,所述计算机程序适于处理器进行加载,以执行如权利要求1-9任一项所述的图文联合纠错方法中的步骤。

12.一种计算机设备,其特征在于,所述计算机设备包括处理器和存储器,所述存储器中存储有计算机程序,所述处理器通过调用所述存储器中存储的所述计算机程序,用于执行权利要求1-9任一项所述的图文联合纠错方法中的步骤。

图文联合纠错方法、装置、存储介质及计算机设备

技术领域

[0001] 本申请涉及计算机技术领域，具体涉及一种图文联合纠错方法、装置、存储介质及计算机设备。

背景技术

[0002] 目前针对文本进行纠错的系统，大多数是以语言模型为基础，再通过拼音以及其它一些限制条件对文本进行纠错。最常见如基于统计语言模型的算法(N-gram)对词句进行打分，然后判断出当前句子是否有明显的字词使用错误。然后再结合拼音以及常见易混淆词表来综合判断句子中是否存在字词句法错误。针对的问题基本都是单一模态的文本输入信息，目前暂时没有发现有关多模态的纠错相关系统。N-gram一般都是通过大规模语料统计出来的，也有相关工作使用屏蔽语言模型(MLM)，但是本质上也都是在纯文本上进行纠错，没有借助于类似图像的相关信息。对于同一句话，检错和纠错的结果一般是固定不变的，因为通过语言模型对当前句子计算得分是固定不变的，虽然有时候可以通过调整检错和纠错的阈值对系统进行调整，但是目前相关多模态的文本纠错系统在解决文本撰写的检错和纠错问题时，只能处理单一模态的文本输入，而无法对包含图文的网络数据进行纠错，且有的纠错系统只能纠正错别字，有的只能纠正语法错误，导致纠错能力不足。

发明内容

[0003] 本申请实施例提供一种图文联合纠错方法、装置、存储介质及计算机设备，可以实现图文联合纠错，提升了纠错能力。

[0004] 一方面，提供一种图文联合纠错方法，所述方法包括：获取待处理的文本数据和图像数据，所述文本数据包括目标文本，所述图像数据包括场景图片；提取所述文本数据的文本向量表示，所述文本向量表示包含所述目标文本的文本信息；提取所述图像数据的图片向量表示，所述图片向量表示包含所述场景图片的图像信息；根据所述文本向量表示与所述图片向量表示，计算包含有文本信息和图像信息的多模态向量表示；在所述目标文本用于表达所述场景图片时，根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果。

[0005] 可选的，所述提取所述文本数据的文本向量表示，包括：通过词表将所述文本数据中的每个词转换成每个词在所述词表中对应的序号，并根据所述序号查找所述文本数据的文本向量表示。

[0006] 可选的，所述提取所述图像数据的图片向量表示，包括：根据目标检测模型对所述场景图片进行目标检测以及特征提取，以得到所述图片向量表示，其中，所述图片向量表示包括所述场景图片中每个图像目标的图像信息向量表示和整个图片的图像信息向量表示。

[0007] 可选的，所述根据所述文本向量表示与所述图片向量表示，计算包含文本信息和图像信息的多模态向量表示，包括：基于自注意力模型对所述文本向量表示与所述图片向量进行处理，获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信

息;对所述全局交互信息进行归一化处理,得到第一归一化信息;根据所述全局交互信息与所述第一归一化信息,确定包含文本信息和图像信息的多模态向量表示。

[0008] 可选的,所述将基于自注意力模型对所述文本向量表示与所述图片向量进行处理,获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息,包括:将根据所述文本向量表示与所述图片向量表示确定的嵌入向量表示输入自注意力模型,根据所述嵌入向量表示与所述嵌入向量表示的转置矩阵之间的乘积,计算匹配矩阵;根据所述匹配矩阵与所述嵌入向量表示的乘积,确定所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息。

[0009] 可选的,所述根据所述全局交互信息与所述第一归一化信息,确定包含文本信息和图像信息的多模态向量表示,包括:将所述全局交互信息与所述第一归一化信息进行相加,以得到第一求和信息;将所述第一求和信息输入全连接层进行处理后,对所述全连接层的输出结果进行归一化处理,得到第二归一化信息;将所述第一求和信息与所述第二归一化信息进行相加,得到所述包含文本信息和图像信息的多模态向量表示。

[0010] 可选的,所述方法还包括:获取位置向量表示和类型向量表示,所述位置向量表示用于标注所述文本数据中每个词的位置,所述类型向量表示用于区分文本类型和图像类型;

[0011] 所述根据所述文本向量表示与所述图片向量表示,计算包含有文本信息和图像信息的多模态向量表示,包括:根据所述文本向量表示、所述图片向量表示、所述位置向量表示和所述类型向量表示,计算包含有文本信息和图像信息的多模态向量表示。

[0012] 可选的,所述根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果,包括:将所述多模态向量表示连接一个全连接层,得到第一特征向量,所述第一特征向量包含有所述目标文本中每个词和所述场景图片中每个图像目标分别对应的向量表示;根据所述第一特征向量,确定所述目标文本中每个词和所述场景图片中每个图像目标分别对应的纠错操作;根据所述纠错操作和所述第一特征向量计算纠错结果,以根据所述纠错结果对所述目标文本和/或所述场景图片进行纠错处理。

[0013] 可选的,所述根据所述纠错操作和所述第一特征向量确定纠错结果,还包括:若所述纠错操作为无错,则确定所述纠错结果为输出与所述无错的纠错操作对应的词;或者若所述纠错操作为删除操作,则确定所述纠错结果为将所述目标文本中与所述删除操作对应的词进行删除;或者若所述纠错操作为修改操作,则确定所述纠错结果为将所述目标文本中与所述修改操作对应的词修改为预测词,或者将所述场景图片中与所述修改操作对应的图像目标修改为预测图像目标。

[0014] 另一方面,提供一种图文联合纠错装置,所述装置包括:

[0015] 获取单元,用于获取待处理的文本数据和图像数据,所述文本数据包括目标文本,所述图像数据包括场景图片;

[0016] 第一提取单元,用于提取所述文本数据的文本向量表示,所述文本向量表示包含所述目标文本的文本信息;

[0017] 第二提取单元,用于提取所述图像数据的图片向量表示,所述图片向量表示包含所述场景图片的图像信息;

[0018] 计算单元,用于根据所述文本向量表示与所述图片向量表示,计算包含有文本信

息和图像信息的多模态向量表示；

[0019] 确定单元，用于在所述目标文本用于表达所述场景图片时，根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果。

[0020] 另一方面，提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序适于处理器进行加载，以执行如上任一实施例所述的图文联合纠错方法中的步骤。

[0021] 另一方面，提供一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有计算机程序，所述处理器通过调用所述存储器中存储的所述计算机程序，用于执行如上任一实施例所述的图文联合纠错方法中的步骤。

[0022] 另一方面，提供一种计算机程序产品，包括计算机指令，所述计算机指令被处理器执行时实现如上任一实施例所述的图文联合纠错方法中的步骤。

[0023] 本申请实施例通过获取待处理的文本数据和图像数据，文本数据包括目标文本，图像数据包括场景图片；提取文本数据的文本向量表示，文本向量表示包含目标文本的文本信息；提取图像数据的图片向量表示，图片向量表示包含场景图片的图像信息；根据文本向量表示与图片向量表示，计算包含有文本信息和图像信息的多模态向量表示；在目标文本用于表达场景图片时，根据多模态向量表示确定针对目标文本和场景图片的纠错结果。本申请实施例通过Transformer模型实现图片和文本输入的多模态图文联合纠错，模型通过同时输入图像数据与包含目标文本的文本数据，通过Transformer模型内部的注意力(attention)机制，计算包含有文本信息和图像信息的多模态向量表示，以将有用的图片信息和文本信息过滤出来，然后根据多模态向量表示中的文本信息来检出图片中的错误信息，确定针对目标文本和场景图片的纠错结果，以根据纠错结果对图文进行纠错，实现了图文联合纠错，提升了纠错能力。

附图说明

[0024] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例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申请的一些实施例，对于本领域技术人员来讲，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。

[0025] 图1为本申请实施例提供的图文联合纠错系统的结构框架图。

[0026] 图2为本申请实施例提供的图文联合纠错方法的第一流程示意图。

[0027] 图3为本申请实施例提供的图文联合纠错方法的第一应用场景示意图。

[0028] 图4为本申请实施例提供的图文联合纠错方法的第二流程示意图。

[0029] 图5为本申请实施例提供的图文联合纠错方法的第二应用场景示意图。

[0030] 图6为本申请实施例提供的图文联合纠错方法的第三流程示意图。

[0031] 图7为本申请实施例提供的图文联合纠错方法的第三应用场景示意图。

[0032] 图8为本申请实施例提供的图文联合纠错装置的结构示意图。

[0033] 图9为本申请实施例提供的计算机设备的结构示意图。

具体实施方式

[0034] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0035] 本申请实施例提供一种图文联合纠错方法、装置、计算机设备和存储介质。具体地，本申请实施例的图文联合纠错方法可以由计算机设备执行，其中，该计算机设备可以为终端或者服务器等设备。该终端可以为智能手机、平板电脑、笔记本电脑、智能电视、智能音箱、穿戴式智能设备、个人计算机(Personal Computer,PC)等设备，终端还可以包括客户端，该客户端可以是视频客户端、浏览器客户端或即时通信客户端等。服务器可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network,CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0036] 本申请实施例可应用于人工智能、语音识别、智慧交通等各种场景。

[0037] 首先，在对本申请实施例进行描述的过程中出现的部分名词或者术语作如下解释：

[0038] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0039] 机器学习(Machine Learning,ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0040] 深度学习(Deep Learning,DL)是机器学习的分支，是一种试图使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法。深度学习是学习训练样本数据的内在规律和表示层次，这些学习过程中获得的信息对文字、图像和声音等数据的解释有很大的帮助。深度学习的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据。深度学习是一个复杂的机器学习算法，在语音和图像识别方面取得的效果，远远超过先前相关技术。

[0041] 神经网络(Neural Network,NN)是在机器学习和认知科学领域的一种模仿生物神经网络结构和功能的深度学习模型。

[0042] Transformer模型是一种NLP(自然语言处理)经典模型。Transformer模型完全基于注意力来编码输入和计算输出,而不依赖于序列对齐的循环神经网络或卷积神经网络,Transformer模型使用Self-Attention(自注意力)机制,而不采用RNN的顺序结构,使得模型可以并行化训练,而且能够拥有全局信息。

[0043] 目前相关的文本纠错系统在解决文本撰写的检错和纠错问题的时候主要存在以下几点不足的地方:

[0044] 1、系统只能够处理以文本为输入的数据。目前几乎所有的文本纠错系统都只能处理单一模态的文本输入,而无法对包含图文的网络数据进行纠错。这造成当前图文满天下的网络数据无法被纠错系统正确的输入处理。

[0045] 2、目前纠错系统都是单一模态,无法根据多模态的信息,来辅助提升纠错的效果。纠错本身就是一个多因素联合的处理系统,目前的文本纠错系统,虽然会根据不同的文本以及不同的应用场景来进行纠错,但是其本质上是单一模态的纯文本场景,无法利用出文本外的额外信息来辅助纠错。

[0046] 3、系统迁移性较差。因为一套纠错系统最基本就是语言模型,而语言模型是通过大规模文本语料通过统计或者MLM训练得到的,一旦定向就无法无缝的迁移到别的领域。如在司法领域文本训练的语言模型,在日常对话的场景中使用效果就会不好。

[0047] 4、将纠错问题划分的很细,有的纠错系统只能纠错别字,有的就只能纠语法错误。很多纠错系统通过将纠错中的问题孤立化,然后去单独解决某类问题,这导致检错和纠错过程会出现不恰当的情况,或者同一错误被多个模块检测到错误,从而导致纠正结果无法统一。

[0048] 其中,传统的纠错都基于文本的纠错,而现实生活场景中,对于多个模态联合纠错的需求在不断提高。例如,一个句子“有个人正在吃苹果。”,其对应的场景是一个人在玩苹果手机,若采用传统的文本纠错系统来对这段文字进行纠错,则无法发现错误,或者识别为“有个人在低头吃苹果手机。”。或者,传统的文本纠错系统可能会纠正为“有个人在低头吃苹果。”,而不会纠正为“有个人在低头玩苹果手机。”,此时就需要多模态图文联合纠错来完成这个任务。随着互联网的快速发展,一条同时包含图片和文本的信息会越来越多,所以,图文联合纠错会变得越来越有应用价值,而传统的单一模态的纠错,对于一些场景的应用也会受到限制,并且无法根据特定的图片等信息,来辅助纠错。所以,单一的模态的纠错本身存在比较大的限制。

[0049] 本申请实施例是在单一文本的基础上进一步引入图像信息,来进行联合纠错,这样不但可以解决更加复杂场景的纠错,还能够通过图像和文本信息的互相增强,来提升发现错误和正确纠正错误的能力。

[0050] 本申请实施例提出的基于神经网络和深度学习模型的多模态的图文联合纠错系统可以克服以上几个问题。首先,本申请实施例可以将图文同时输入纠错系统,同时利用图文信息进行纠错。现在很多互联网数据都是图文混合的,单独只是对于文本进行纠错的系统,以后会逐渐跟不上时代潮流。其次,图文两个模态的信息可以互相利用,进而提高整体的纠错效果,避免一些单独只通过文本无法判断的错误无法被纠正。另外,本申请实施例会根据文本信息来检出图片中的错误信息,进而在纠错过程中,提升文本配图的质量,同时还可以对图文进行纠错。

[0051] 请参阅图1,图1为本申请实施例提供的图文联合纠错系统的结构框架图。该图文联合纠错系统包括一个多模态的Transformer模型和一个纠错模块。首先文本模态的文本数据的输入,该包括目标文本。然后是图像模态的图像数据的输入,图像数据包括场景图片。然后是一个多模态的Transformer模型。可以通过Transformer模型计算出一个融合了文本信息和图像信息的多模态向量表示。然后将该多模态向量表示再输入一个纠错模块,计算出文本和图像中可能出现错误的信息。

[0052] 通过同时输入目标文本和场景图片的信号,通过Transformer模型内部的attention机制,将有用的图像信息和文本信息过滤出来,然后根据这些过滤出来的信息,通过纠错模块计算出每个字(或每个词)和场景图片中每个图像目标是否存在错误。

[0053] 文本和图像的两个模态的特征表示,是在图片和文本同时输入的时候学习到的,模型能够准确的得到有用的表示信息。最后再准确的判断出这个场景图片是不是与当前的目标文本存在错误表达的问题,如果目标文本存在错误表达,则对错误地方进行纠正,如果场景图片中存在错误的地方,则对场景图片作出纠正或者输出图片哪个地方有问题。例如,可以将场景图片中的各个与文字信息对应的检测目标作为图像目标,通过图像目标中图像信息与文本信息的匹配度,来确定图像目标是否存在错误,即确定场景图片中存在错误的地方。

[0054] 由于整个图文联合纠错系统是直接通过一个模型来完成多模态的图文纠错,所以总体结构比较简洁,性能也比较好。

[0055] 以下分别进行详细说明。需说明的是,以下实施例的描述顺序不作为对实施例优先顺序的限定。

[0056] 本申请各实施例提供了一种图文联合纠错方法,该方法可以由终端或服务器执行,也可以由终端和服务器共同执行;本申请实施例以图文联合纠错方法由服务器执行为例来进行说明。

[0057] 请参阅图2至图7,图2、图4及图6均为本申请实施例提供的图文联合纠错方法的流程示意图,图3、图5及图7均为本申请实施例提供的图文联合纠错方法的应用场景示意图。该方法包括:

[0058] 步骤110,获取待处理的文本数据和图像数据,所述文本数据包括目标文本,所述图像数据包括场景图片。

[0059] 例如,该文本数据包括目标文本。比如,获取的待处理的目标文本为“图中漂亮女人戴着眼镜”。该目标文本也可以作为问题。

[0060] 例如,该图像数据包括场景图片,该场景图片是对应该目标文本描述的场景提供的图片,比如用户在作答过程中,可以通过观察该场景图片来回答该确定该目标文本(问题)是否为该场景图片的正确表述。

[0061] 步骤120,提取所述文本数据的文本向量表示,所述文本向量表示包含所述目标文本的文本信息。

[0062] 可选的,所述提取所述文本数据的文本向量表示,包括:

[0063] 通过词表将所述文本数据中的每个词转换成每个词在所述词表中对应的序号,并根据所述序号查找所述文本数据的文本向量表示。

[0064] 首先,可以对文本数据进行向量化的表示,以将目标文本中的每个词映射到一个

特定的空间,以得到文本数据的文本向量表示。

[0065] 例如,结合图3进行说明,该图文联合纠错系统可以包括多模态的数据输入处理模块、多模态的特征抽取模块和纠错模块,其中,该特征抽取模块可以采用Transformer模型。

[0066] 例如,将输入该数据输入处理模块的TXT格式的文本数据(即TXT格式的目标文本对应的原始文本),通过词表将原始文本转换成每个词对应词表中的序号(ID),然后将原始文本通过ID查找每个词对于词表的嵌入(embedding)向量表示。例如,目标文本:“图中漂亮女人戴着眼镜”,通过词表转化为ID[1,4,3,6,7,0,12,87,98],因为“图”在词表中的ID是1,所以就转换成1,然后再通过ID找到每个词对应的向量表示(w1,w2,w3,w4,w5,w6,w7,w8,w9,w10),得到词向量序列。然后将这一串词向量序列后续作为Transformer模型的输入参数,可以定义为文本向量表示,该文本向量表示为seq_len乘hid_size的矩阵,其中,seq_len表示文本长度,hid_size表示词向量的大小。

[0067] 步骤130,提取所述图像数据的图片向量表示,所述图片向量表示包含所述场景图片的图像信息。

[0068] 可选的,所述提取所述图像数据的图片向量表示,包括:

[0069] 根据目标检测模型对所述场景图片进行目标检测以及特征提取,以得到所述图片向量表示,其中,所述图片向量表示包括所述场景图片中每个图像目标的图像信息向量表示和整个图片的图像信息向量表示。

[0070] 例如,将输入该数据输入处理模块的图像数据,首先通过目标检测模型(Fast RCNN)抽取物理的数值信息,然后通过模型学习,得到一个能够正确抽取物理信息的Fast RCNN模型。Fast RCNN(Fast Regions with CNN features)一种快速的基于区域的卷积网络方法,用于目标检测。

[0071] 例如,结合图3进行说明,输入包含有场景图片的图像数据,首先通过Fast-RCNN对图像数据中的场景图片进行目标检测以及特征抽取,得到图片中每个图像目标的图像信息向量表示和整个图片的图像信息向量表示。其中,该图像目标为对应于文本数据中需要关注的对象,比如文本数据中提到“图中漂亮女人”,那么就需要关注场景图片中的人物,比如图中的男人和女人。其中,整个图片的图像信息向量表示会分别应用到每个文本字上面,而单个图像目标的图像信息向量表示会对应到专门表示图像信息的文本向量上,也就是图3示出的img文本向量。最终图像数据作为Transformer模型的输入参数也是一个seq_len乘hid_size的矩阵,可定义为图片向量表示。

[0072] 步骤140,根据所述文本向量表示与所述图片向量表示,计算包含有文本信息和图像信息的多模态向量表示。

[0073] 例如,将包含图像和文本的Embedding向量表示输入到Transformer模型中,计算出融合了文本信息和图像信息的多模态向量表示。多模态向量表示是一个既包含文本信息又包含图像信息的向量。

[0074] 可选的,所述方法还包括:获取位置向量表示和类型向量表示,所述位置向量表示用于标注所述文本数据中每个词的位置,所述类型向量表示用于区分文本类型和图像类型;

[0075] 所述根据所述文本向量表示与所述图片向量表示,计算包含有文本信息和图像信息的多模态向量表示,包括:

[0076] 根据所述文本向量表示、所述图片向量表示、所述位置向量表示和所述类型向量表示，计算包含有文本信息和图像信息的多模态向量表示。

[0077] 其中，位置向量表示用于标注文本数据中每个词的位置，位置向量表示的大小为seq_len乘hid_size的矩阵。类型向量表示用于区分文本类型和图像类型，类型向量表示的大小为seq_len乘hid_size的矩阵，比如分文本类型表示为0，图像类型表示为1。

[0078] 例如，如图3所示，最终输入Transformer模型的输入参数可以包括由文本向量表示加上图片向量表示加上位置向量表示和类型向量表示构成的嵌入向量表示Embedding，记为E。例如，将目标文本和场景图片的Embedding向量表示输入到Transformer模型中，计算出文本与图像之间的多模态向量表示。多模态向量表示是一个既包含文本信息又包含图像信息的向量，即目标文本中的每个词的向量表示与场景图片中每个图像目标的向量表示通过Transformer模型内部的计算，通过各组共现的方式抽取的最优的特征向量，最后将这个多模态的表示输出给纠错模块，在该图文联合纠错系统中就是通过纠错模块，计算出目标文本中每个词是否需要修改、怎么修改，以及场景图片中的每个图像目标是否有错误，是否需要修改等。

[0079] 例如，结合图3进行说明，可以通过根据多模态的特征抽取模块对文本向量表示与图片向量表示进行处理，计算包含有文本信息和图像信息的多模态向量表示，其中，该特征抽取模块可以采用Transformer模型。该特征抽取模块的主要功能是计算融合了文本信息和图像信息的多模态向量表示。如图3所示，通过Transformer模型处理后最终会得到一个融合了文本信息和图像信息的多模态向量表示，而这个多模态向量表示可以用于纠错模块计算目标文本和场景图片中是否存在错误，以及是否需进行“增”、“删”、“改”等处理。如图3所示，计算的就是“女人”这个表述是否和图片匹配，以及图片中是不是女人戴着眼镜。该模块设计在这里主要是为了计算图像文本之间的关联关系，并且输出匹配的特征矩阵。

[0080] 可选的，如图4所示，步骤140可通过步骤141至步骤143来实现，具体为：

[0081] 步骤141，基于自注意力模型对所述文本向量表示与所述图片向量进行处理，获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息。

[0082] 可选的，所述将基于自注意力模型对所述文本向量表示与所述图片向量进行处理，获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息，包括：将根据所述文本向量表示与所述图片向量表示确定的嵌入向量表示输入自注意力模型，根据所述嵌入向量表示与所述嵌入向量表示的转置矩阵之间的乘积，计算匹配矩阵；根据所述匹配矩阵与所述嵌入向量表示的乘积，确定所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息。

[0083] 其中，全局交互信息的长度维度与嵌入向量表示的长度维度相同。

[0084] 例如，请参阅图5，特征抽取模块的输入为：图像和文本的向量表示，其大小为seq_len乘hid_size的矩阵。例如，该图像和文本的向量表示可以为文本向量表示加上图片向量表示构成的Embedding向量表示，记为E。该图像和文本的向量表示也可以为文本向量表示加上图片向量表示加上位置向量表示和类型向量表示，构成的Embedding向量表示，记为E。

[0085] 特征抽取模块的输出为：融合了图像信息和所有文本信息的多模态向量表示，其大小为seq_len乘hid_size的矩阵。

[0086] 例如，请参阅图5，通过特征抽取模块内部的自注意力(self_attention)模型计算

匹配矩阵，输入为seq_len乘hid_size的embedding向量表示E，该embedding向量表示即为图像和文本的向量表示；输出为全局交互信息H_s，H_s的大小为seq_len乘hid_size。具体计算过程中，self_attention即自己和自己计算注意力(attention)表示，E矩阵乘E^T，得到匹配矩阵M，其中，匹配矩阵M的大小为seq_len乘seq_len，然后M矩阵乘E得到H_s，H_s的大小为seq_len乘hid_size。其中，E^T是E矩阵的转置矩阵。

[0087] 步骤142，对所述全局交互信息进行归一化处理，得到第一归一化信息。

[0088] 例如，请参阅图5，对self_attention模型输出的全局交互信息H_s进行归一化，即norm，得到第一归一化信息H_n，第一归一化信息H_n的大小为seq_len乘hid_size，归一化不会影响矩阵大小。其中，全局交互信息的长度维度与嵌入向量表示的长度维度相同，全局交互信息的词向量大小与嵌入向量表示的词向量大小相同。

[0089] 步骤143，根据所述全局交互信息与所述第一归一化信息，确定包含文本信息和图像信息的多模态向量表示。

[0090] 可选的，所述根据所述全局交互信息与所述第一归一化信息，确定包含文本信息和图像信息的多模态向量表示，包括：将所述全局交互信息与所述第一归一化信息进行相加，以得到第一求和信息；将所述第一求和信息输入全连接层进行处理后，对所述全连接层的输出结果进行归一化处理，得到第二归一化信息；将所述第一求和信息与所述第二归一化信息进行相加，得到所述包含文本信息和图像信息的多模态向量表示。

[0091] 例如，请参阅图5，全连接层的输入为H_s+H_n，即将全局交互信息H_s与第一归一化信息H_n进行相加得到第一求和信息，然后将将第一求和信息输入全连接层进行处理，其中全连接层的输出结果的大小为seq_len乘hid_size。然后再对全连接层的输出结果进行归一化处理(norm)，得到第二归一化信息，并再次将得到第二归一化信息与第一求和信息相加，得到多模态向量表示H_{nn}。由于H_{nn}与输入的E矩阵大小一样，self_attention模型这里可以叠加多层，一般可以设置为12层或者24层。

[0092] 例如，多模态向量表示H_{nn}最后通过直接输出output，output的大小为seq_len乘hid_size的矩阵，该多模态向量表示为融合了图像信息和所有文本信息的多模态向量表示。其中，多模态向量表示的长度维度与嵌入向量表示的长度维度相同。

[0093] 通过该特征抽取模块，计算得到融合了文本和图像的多模态向量表示，该多模态向量表示包含文本和图像之间的高度抽象的语义匹配关系，为后续模块根据匹配信息语义检错纠错提供了丰富的信息。同时也将以往只是通过简单的字符级别的文本匹配转换到了向量空间之间的匹配。该特征抽取模块使得文本匹配上升到了语义空间级别。

[0094] 步骤150，在所述目标文本用于表达所述场景图片时，根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果。

[0095] 例如，可以将特征抽取模块的输出的output输出结果(融合了文本信息和图像信息的多模态向量表示)，通过一个纠错模块进行处理，计算出目标文本中每个词是否需要修改、怎么修改，以及场景图片中的每个图像目标是否有错误，是否需要修改等。

[0096] 其中，该纠错模块主要是通过对特征抽取模块的output输出结果进行处理，以对目标文本中的每个词和场景图片中的每个图像目标部分进行纠错，以图3示出的目标文本“图中漂亮女人戴着眼镜”为例，该目标文本本身是没有任何语法表达上的错误，同时场景图片上也没有任何有问题的地方，但是将该目标文本用于表达这个场景图片时就是有错误

的，即该场景图片与目标文本存在不匹配的错误信息。该纠错模块的主要任务就是计算并发现错误，且纠正错误。

[0097] 可选的，如图6所示，步骤150可通过步骤151至步骤153来实现，具体为：

[0098] 步骤151，将所述多模态向量表示连接一个全连接层，得到第一特征向量，所述第一特征向量包含有所述目标文本中每个词和所述场景图片中每个图像目标分别对应的向量表示。

[0099] 例如，请参阅图7，融合了文本信息和图像信息的多模态向量表示，接一个全连接，得到第一特征向量$H=[h_1,h_2,h_3,\cdots,h_{15}]$，比如seq_len＝15，H包含的目标文本中所有词和场景图片中所有图像目标分别对应的向量总共为15个。

[0100] 步骤152，根据所述第一特征向量，确定所述目标文本中每个词和所述场景图片中每个图像目标分别对应的纠错操作。

[0101] 例如，请参阅图7，将第一特征向量H输入到识别纠错方式模块中，通过这个识别纠错方式模块，可以得到长度维度为seq_len上的每个词和每个图像目标对应的纠错操作$O=[O_1,O_2,O_3,\cdots,O_{15}]$，该纠错操作主要包括三种：无错，这种无错的操作说明不需要做纠错；删除操作，这种删除操作说明当前词冗余了需要删除；修改操作，这种修改操作说明当前词是一个错误的词，需要做纠正；或者说明当前图像目标是一个错误的图像目标，需要做纠正。

[0102] 步骤153，根据所述纠错操作和所述第一特征向量计算纠错结果，以根据所述纠错结果对所述目标文本和/或所述场景图片进行纠错处理。

[0103] 可选的，所述根据所述纠错操作和所述第一特征向量确定纠错结果，还包括：

[0104] 若所述纠错操作为无错，则确定所述纠错结果为输出与所述无错的纠错操作对应的词；或者

[0105] 若所述纠错操作为删除操作，则确定所述纠错结果为将所述目标文本中与所述删除操作对应的词进行删除；或者

[0106] 若所述纠错操作为修改操作，则确定所述纠错结果为将所述目标文本中与所述修改操作对应的词修改为预测词，或者将所述场景图片中与所述修改操作对应的图像目标修改为预测图像目标。

[0107] 例如，请参阅图7，根据纠错操作O和第一特征向量H，最终计算纠错结果。例如，“无错”直接输出与无错的纠错操作对应的词，或者若目标文本中所有词对应的纠错操作均为无错，则可以直接输出该目标文本。“删除”，则与该删除操作对应的词输出为空。“修改”，则将当前h_i(如图7的示例为h_1)接上全连接预测一个新的词(或字)输出，如果当前h_i是图像目标，对于场景图片中这个图像目标所在的地方将通过全连接输出一个新的图像目标。例如，对于输入的目标文“图中漂亮女人戴着眼镜”，纠错结果为“漂、亮、女”三个字对应的纠错操作是修改，预测的新词分别为“帅、气、男”，根据纠错结果对目标文本进行纠错处理时的输出为“图中帅气男人戴着眼镜”。

[0108] 本申请实施例提供的图文联合纠错系统在纠错的过程中不但可以对文本进行纠错，还可以同时识别图片中有问题的地方，如果是有错误的图片，还会生成纠错后的图片作为修改的建议。

[0109] 其中，在使用该图文联合纠错系统之前，还可以提供给足够多的跨模态数据，对该图文联合纠错系统进行模型学习。通过学习过后，整个图文联合纠错系统就可以自动对包

含图片和文本的输入进行纠错，当然也可以单独处理文本或者图片。

[0110] 本申请实施例提供的图文联合纠错方法，通过图像信息辅助文本纠错，通过文本信息发现图片中的一些常识性错误，开创性的设计了一套联合图片和文本信息的纠错系统，将纠错由单一模态任务，提升到一个多模态夸模态的层面，极具创新性的将多模态技术应用在了纠错任务上，而且是多模态的联合纠错，本申请实施例虽然只详细描述了图像文本的联合纠错，实际上可以加入更多其它模态的信息，如语音、视频等来辅助纠错。图文纠错方式的创新，将纠错定义为两个阶段，第一阶段识别出纠错需要采取的方法和手段，第二阶段具体纠错方法下使用对应的纠结结果进行纠错。将复杂的纠错系统，设计成了end-to-end的模型方法，这样非常有利于后期维护和部署。多个模态互相增强的纠错方法，多模态的信号处理是以后人工智能发展的必然趋势，本系统前瞻的多模态图文联合纠错方法，将智能纠错提升到一个新的高度，能够解决常规纯文本纠错无法解决的问题，并且可以帮助机器学习到多个模态的知识以及尝试，这对单一模态的纠错也是非常有帮助的。本申请实施例可以根据图像可以发现文本中存在的表述不当。同时，还可以借助文本信息，对一些包含错误的图片进行纠正。

[0111] 本申请实施例提供了一种多模态的图文联合纠错方法以及图文联合纠错系统，相比于以往的只能对文本进行纠错的方法，本申请实施例可以同时对文本和图片进行纠错，在对文本纠错的同时，还可以在图片上找出对应的论据，增加纠错的可理解性。

[0112] 上述所有的技术方案，可以采用任意结合形成本申请的可选实施例，在此不再一一赘述。

[0113] 本申请实施例通过获取待处理的文本数据和图像数据，文本数据包括目标文本，图像数据包括场景图片；提取文本数据的文本向量表示，文本向量表示包含目标文本的文本信息；提取图像数据的图片向量表示，图片向量表示包含场景图片的图像信息；根据文本向量表示与图片向量表示，计算包含有文本信息和图像信息的多模态向量表示；在目标文本用于表达场景图片时，根据多模态向量表示确定针对目标文本和场景图片的纠错结果。本申请实施例通过Transformer模型实现图片和文本输入的多模态图文联合纠错，模型通过同时输入图像数据与包含目标文本的文本数据，通过Transformer模型内部的注意力(attention)机制，计算包含有文本信息和图像信息的多模态向量表示，以将有用的图片信息和文本信息过滤出来，然后根据多模态向量表示中的文本信息来检出图片中的错误信息，确定针对目标文本和场景图片的纠错结果，以根据纠错结果对图文进行纠错，实现了图文联合纠错，提升了纠错能力。

[0114] 为便于更好的实施本申请实施例的图文联合纠错方法，本申请实施例还提供一种图文联合纠错装置。请参阅图8，图8为本申请实施例提供的图文联合纠错装置的结构示意图。其中，该图文联合纠错装置200可以包括：

[0115] 获取单元201，用于获取待处理的文本数据和图像数据，所述文本数据包括目标文本，所述图像数据包括场景图片；

[0116] 第一提取单元202，用于提取所述文本数据的文本向量表示，所述文本向量表示包含所述目标文本的文本信息；

[0117] 第二提取单元203，用于提取所述图像数据的图片向量表示，所述图片向量表示包含所述场景图片的图像信息；

[0118] 计算单元204，用于根据所述文本向量表示与所述图片向量表示，计算包含有文本信息和图像信息的多模态向量表示；

[0119] 确定单元205，用于在所述目标文本用于表达所述场景图片时，根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果。

[0120] 可选的，所述第一提取单元202，可以用于通过词表将所述文本数据中的每个词转换成每个词在所述词表中对应的序号，并根据所述序号查找所述文本数据的文本向量表示。

[0121] 可选的，所述第二提取单元203，可以用于根据目标检测模型对所述场景图片进行目标检测以及特征提取，以得到所述图片向量表示，其中，所述图片向量表示包括所述场景图片中每个图像目标的图像信息向量表示和整个图片的图像信息向量表示。

[0122] 可选的，所述计算单元204，可以具体用于：基于自注意力模型对所述文本向量表示与所述图片向量进行处理，获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息；对所述全局交互信息进行归一化处理，得到第一归一化信息；根据所述全局交互信息与所述第一归一化信息，确定包含文本信息和图像信息的多模态向量表示。

[0123] 可选的，所述计算单元204，在将基于自注意力模型对所述文本向量表示与所述图片向量进行处理，获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息时，可以用于：将根据所述文本向量表示与所述图片向量表示确定的嵌入向量表示输入自注意力模型，根据所述嵌入向量表示与所述嵌入向量表示的转置矩阵之间的乘积，计算匹配矩阵；根据所述匹配矩阵与所述嵌入向量表示的乘积，确定所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息。

[0124] 可选的，所述计算单元204，在根据所述全局交互信息与所述第一归一化信息，确定包含文本信息和图像信息的多模态向量表示时，可以用于：将所述全局交互信息与所述第一归一化信息进行相加，以得到第一求和信息；将所述第一求和信息输入全连接层进行处理后，对所述全连接层的输出结果进行归一化处理，得到第二归一化信息；将所述第一求和信息与所述第二归一化信息进行相加，得到所述包含文本信息和图像信息的多模态向量表示。

[0125] 可选的，所述获取单元201，还可以用于获取位置向量表示和类型向量表示，所述位置向量表示用于标注所述文本数据中每个词的位置，所述类型向量表示用于区分文本类型和图像类型；

[0126] 所述计算单元204，可以用于根据所述文本向量表示、所述图片向量表示、所述位置向量表示和所述类型向量表示，计算包含有文本信息和图像信息的多模态向量表示。

[0127] 可选的，所述确定单元205，可以具体用于：将所述多模态向量表示连接一个全连接层，得到第一特征向量，所述第一特征向量包含有所述目标文本中每个词和所述场景图片中每个图像目标分别对应的向量表示；根据所述第一特征向量，确定所述目标文本中每个词和所述场景图片中每个图像目标分别对应的纠错操作；根据所述纠错操作和所述第一特征向量计算纠错结果，以根据所述纠错结果对所述目标文本和/或所述场景图片进行纠错处理。

[0128] 可选的，所述确定单元205在根据所述纠错操作和所述第一特征向量确定纠错结果时，还可以用于：若所述纠错操作为无错，则确定所述纠错结果为输出与所述无错的纠错

操作对应的词;或者若所述纠错操作为删除操作,则确定所述纠错结果为将所述目标文本中与所述删除操作对应的词进行删除;或者若所述纠错操作为修改操作,则确定所述纠错结果为将所述目标文本中与所述修改操作对应的词修改为预测词,或者将所述场景图片中与所述修改操作对应的图像目标修改为预测图像目标。

[0129] 需要说明的是,本申请实施例中的图文联合纠错装置200中各模块的功能可对应参考上述各方法实施例中任意实施例的具体实现方式,这里不再赘述。

[0130] 上述图文联合纠错装置中的各个单元可全部或部分通过软件、硬件及其组合来实现。上述各个单元可以以硬件形式内嵌于或独立于计算机设备中的处理器中,也可以以软件形式存储于计算机设备中的存储器中,以便于处理器调用执行上述各个单元对应的操作。

[0131] 图文联合纠错装置200例如可以集成在具备储存器并安装有处理器而具有运算能力的终端或服务器中,或者该图文联合纠错装置200为该终端或服务器。该终端可以为智能手机、平板电脑、笔记本电脑、智能电视、智能音箱、穿戴式智能设备、个人计算机(Personal Computer,PC)等设备,终端还可以包括客户端,该客户端可以是视频客户端、浏览器客户端或即时通信客户端等。服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network,CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0132] 图9为本申请实施例提供的计算机设备的结构示意图,如图9所示,计算机设备300可以包括:通信接口301,存储器302,处理器303和通信总线304。通信接口301,存储器302,处理器303通过通信总线304实现相互间的通信。通信接口301用于装置300与外部设备进行数据通信。存储器302可用于存储软件程序以及模块,处理器303通过运行存储在存储器302的软件程序以及模块,例如前述方法实施例中的相应操作的软件程序。

[0133] 可选的,该处理器303可以调用存储在存储器302的软件程序以及模块执行如下操作:获取待处理的文本数据和图像数据,所述文本数据包括目标文本,所述图像数据包括场景图片;提取所述文本数据的文本向量表示,所述文本向量表示包含所述目标文本的文本信息;提取所述图像数据的图片向量表示,所述图片向量表示包含所述场景图片的图像信息;根据所述文本向量表示与所述图片向量表示,计算包含有文本信息和图像信息的多模态向量表示;在所述目标文本用于表达所述场景图片时,根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果。

[0134] 可选的,该计算机设备300为该终端或服务器。该终端可以为智能手机、平板电脑、笔记本电脑、智能电视、智能音箱、穿戴式智能设备、个人计算机等设备。该服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0135] 可选的,本申请还提供了一种计算机设备,包括存储器和处理器,存储器中存储有计算机程序,该处理器执行计算机程序时实现上述各方法实施例中的步骤。

[0136] 本申请还提供了一种计算机可读存储介质,用于存储计算机程序。该计算机可读存储介质可应用于计算机设备,并且该计算机程序使得计算机设备执行本申请实施例中的

图文联合纠错方法中的相应流程,为了简洁,在此不再赘述。

[0137] 本申请还提供了一种计算机程序产品,该计算机程序产品包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得计算机设备执行本申请实施例中的图文联合纠错方法中的相应流程,为了简洁,在此不再赘述。

[0138] 本申请还提供了一种计算机程序,该计算机程序包括计算机指令,计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得计算机设备执行本申请实施例中的图文联合纠错方法中的相应流程,为了简洁,在此不再赘述。

[0139] 应理解,本申请实施例的处理器可能是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法实施例的各步骤可以通过处理器中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器可以是通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成,或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器,闪存、只读存储器,可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器,处理器读取存储器中的信息,结合其硬件完成上述方法的步骤。

[0140] 可以理解,本申请实施例中的存储器可以是易失性存储器或非易失性存储器,或可包括易失性和非易失性存储器两者。其中,非易失性存储器可以是只读存储器(Read-Only Memory,ROM)、可编程只读存储器(Programmable ROM,PROM)、可擦除可编程只读存储器(Erasable PROM,EPROM)、电可擦除可编程只读存储器(Electrically EPROM,EEPROM)或闪存。易失性存储器可以是随机存取存储器(Random Access Memory,RAM),其用作外部高速缓存。通过示例性但不是限制性说明,许多形式的RAM可用,例如静态随机存取存储器(Static RAM,SRAM)、动态随机存取存储器(Dynamic RAM,DRAM)、同步动态随机存取存储器(Synchronous DRAM,SDRAM)、双倍数据速率同步动态随机存取存储器(Double Data Rate SDRAM,DDR SDRAM)、增强型同步动态随机存取存储器(Enhanced SDRAM,ESDRAM)、同步连接动态随机存取存储器(Synchlink DRAM,SLDRAM)和直接内存总线随机存取存储器(Direct Rambus RAM,DR RAM)。应注意,本文描述的系统和方法的存储器旨在包括但不限于这些和任意其它适合类型的存储器。

[0141] 应理解,上述存储器为示例性但不是限制性说明,例如,本申请实施例中的存储器还可以是静态随机存取存储器(static RAM,SRAM)、动态随机存取存储器(dynamic RAM,DRAM)、同步动态随机存取存储器(synchronous DRAM,SDRAM)、双倍数据速率同步动态随机存取存储器(double data rate SDRAM,DDR SDRAM)、增强型同步动态随机存取存储器(enhanced SDRAM,ESDRAM)、同步连接动态随机存取存储器(synch link DRAM,SLDRAM)以及直接内存总线随机存取存储器(Direct Rambus RAM,DR RAM)等等。也就是说,本申请实

施例中的存储器旨在包括但不限于这些和任意其它适合类型的存储器。

[0142] 本领域普通技术人员可以意识到,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、或者计算机软件和电子硬件的结合来实现。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申请的范围。

[0143] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0144] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0145] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0146] 另外,在本申请实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。

[0147] 所述功能如果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,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器)执行本申请各个实施例所述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、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0148] 以上所述,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应所述以权利要求的保护范围为准。

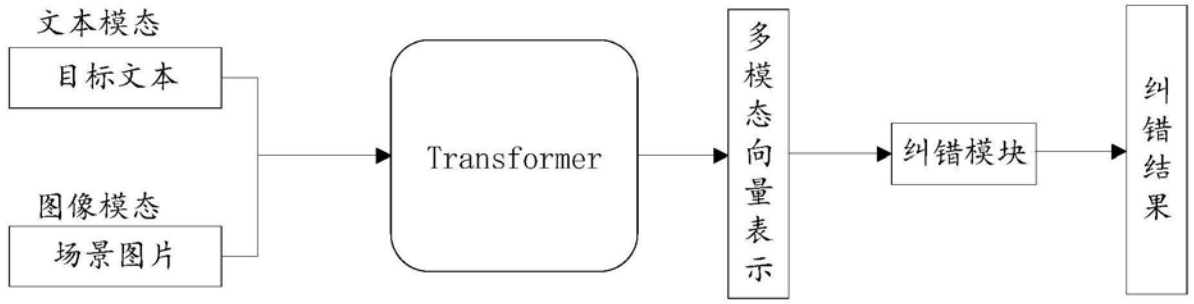

文本模态
目标文本
图像模态
场景图片
Transformer
多模态向量表示
纠错模块
纠错结果

图1

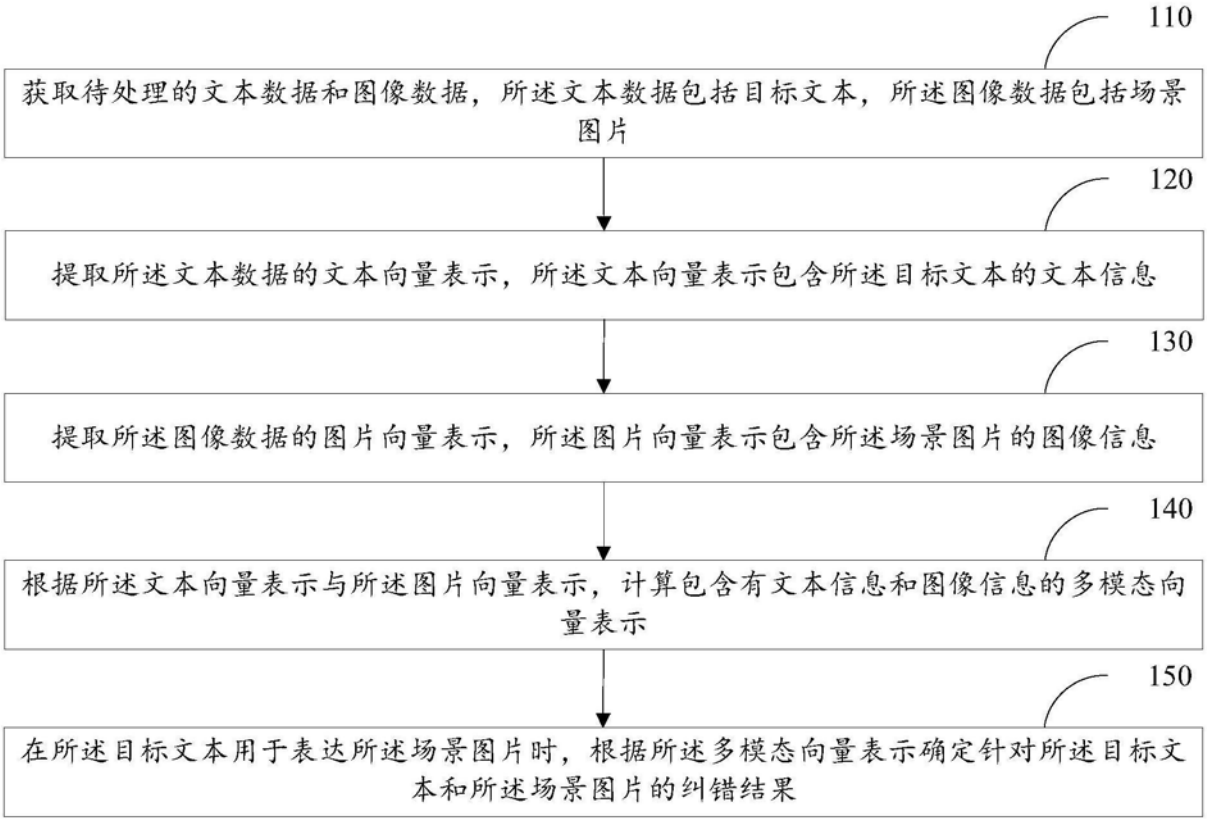

110
获取待处理的文本数据和图像数据，所述文本数据包括目标文本，所述图像数据包括场景图片
120
提取所述文本数据的文本向量表示，所述文本向量表示包含所述目标文本的文本信息
130
提取所述图像数据的图片向量表示，所述图片向量表示包含所述场景图片的图像信息
140
根据所述文本向量表示与所述图片向量表示，计算包含有文本信息和图像信息的多模态向量表示
150
在所述目标文本用于表达所述场景图片时，根据所述多模态向量表示确定针对所述目标文本和所述场景图片的纠错结果

图2

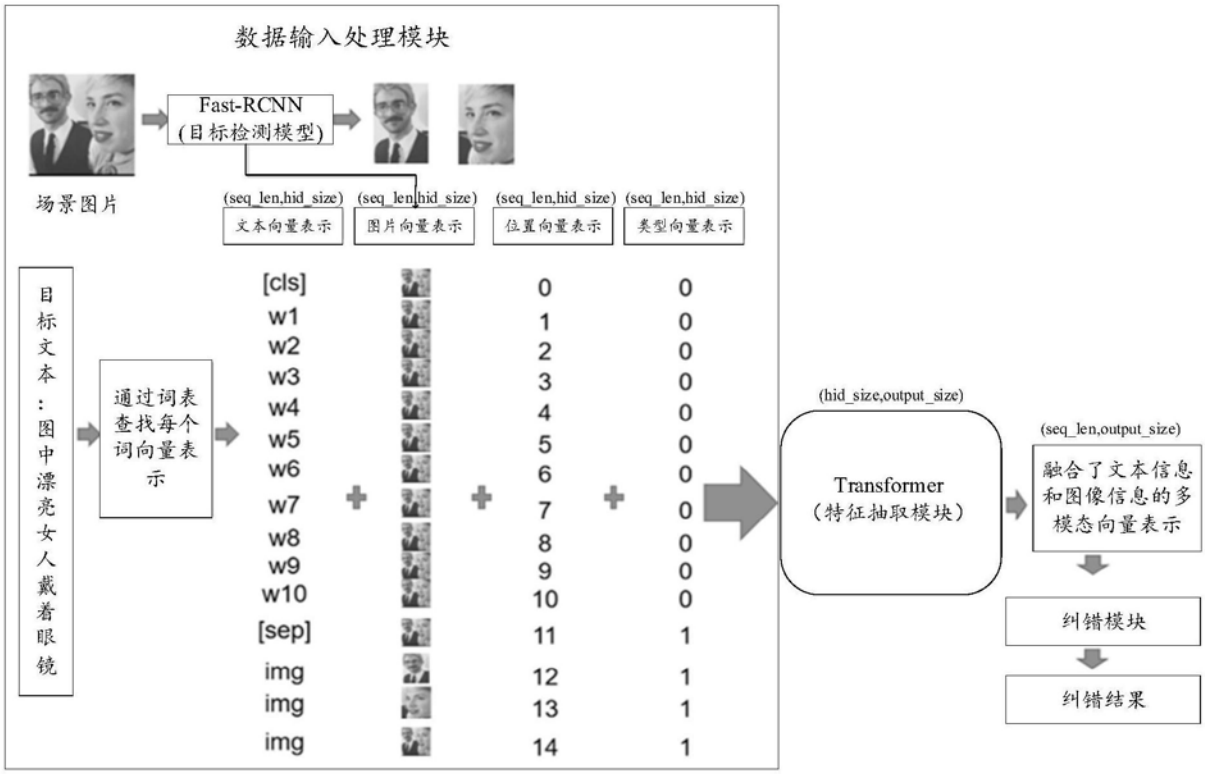

数据输入处理模块
Fast-RCNN
(目标检测模型)
场景图片
(seq_len,hid_size)
文本向量表示
(seq_len,hid_size)
图片向量表示
(seq_len,hid_size)
位置向量表示
(seq_len,hid_size)
类型向量表示
目标文本：图中漂亮女人戴着眼镜
通过词表查找每个词向量表示
[cls]
w1
w2
w3
w4
w5
w6
w7
w8
w9
w10
[sep]
img
img
img
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
0
0
0
0
0
0
0
0
0
0
0
1
1
1
1
(hid_size,output_size)
Transformer
(特征抽取模块)
(seq_len,output_size)
融合了文本信息和图像信息的多模态向量表示
纠错模块
纠错结果

图3

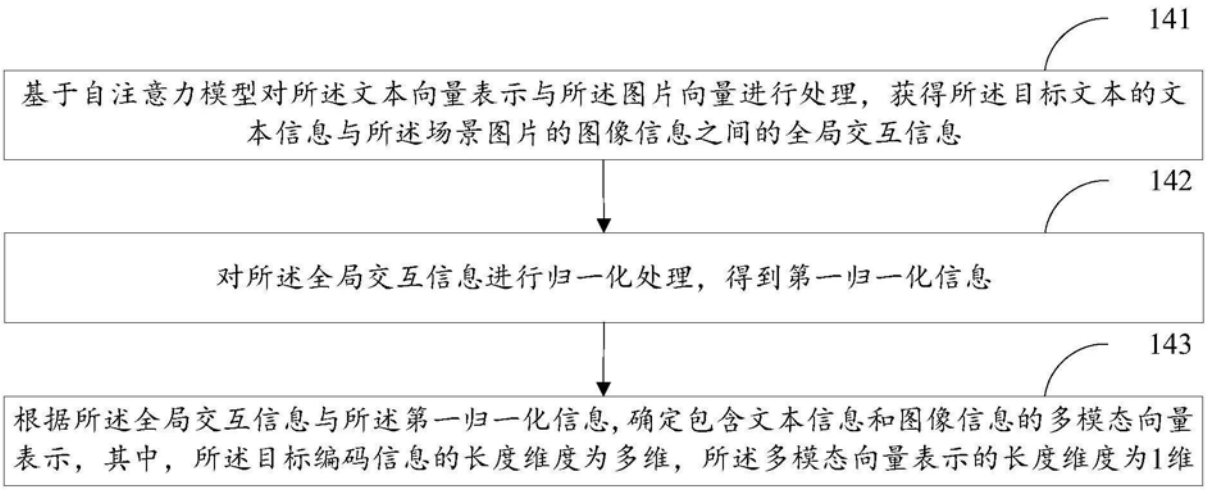

141
基于自注意力模型对所述文本向量表示与所述图片向量进行处理，获得所述目标文本的文本信息与所述场景图片的图像信息之间的全局交互信息
142
对所述全局交互信息进行归一化处理，得到第一归一化信息
143
根据所述全局交互信息与所述第一归一化信息，确定包含文本信息和图像信息的多模态向量表示，其中，所述目标编码信息的长度维度为多维，所述多模态向量表示的长度维度为1维

图4

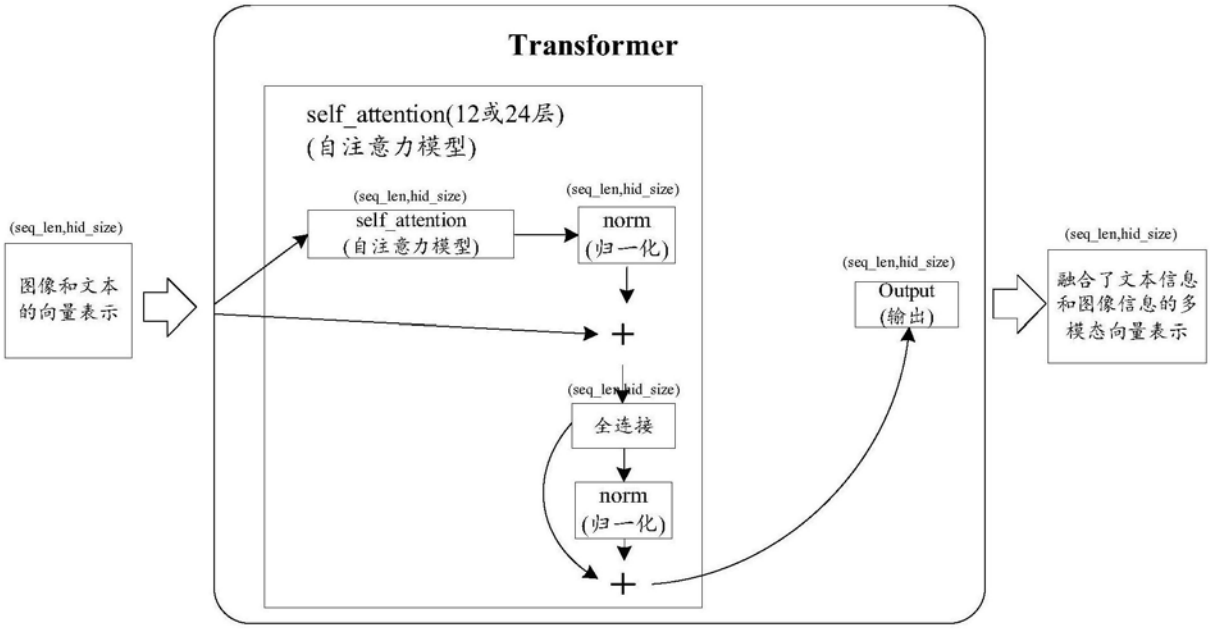

Transformer
self_attention(12或24层)
(自注意力模型)
(seq_len,hid_size)
图像和文本的向量表示
(seq_len,hid_size)
self_attention
(自注意力模型)
(seq_len,hid_size)
norm
(归一化)
(seq_len,hid_size)
全连接
norm
(归一化)
(seq_len,hid_size)
Output
(输出)
(seq_len,hid_size)
融合了文本信息和图像信息的多模态向量表示

图5

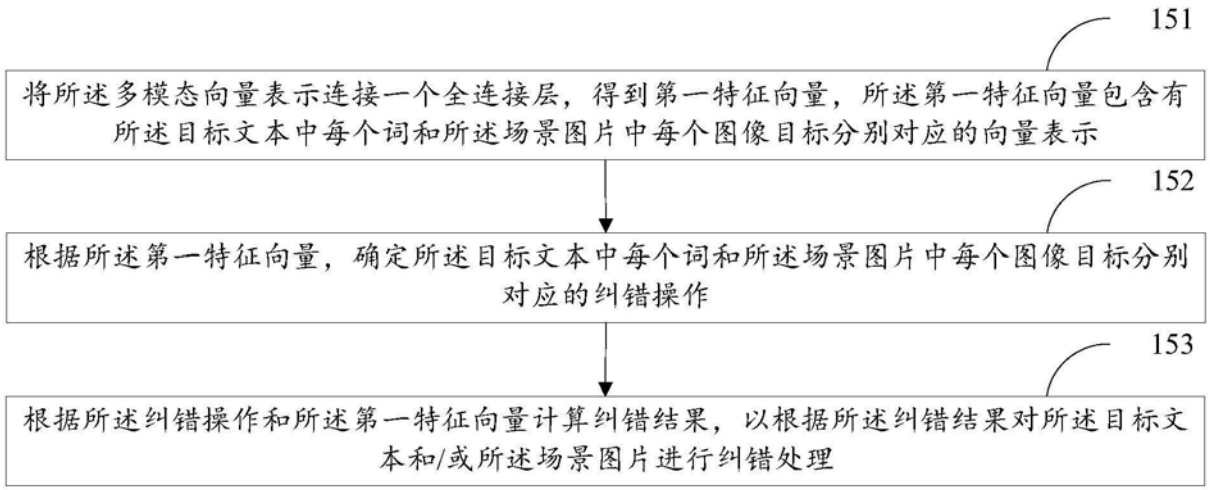

151
将所述多模态向量表示连接一个全连接层，得到第一特征向量，所述第一特征向量包含有所述目标文本中每个词和所述场景图片中每个图像目标分别对应的向量表示
152
根据所述第一特征向量，确定所述目标文本中每个词和所述场景图片中每个图像目标分别对应的纠错操作
153
根据所述纠错操作和所述第一特征向量计算纠错结果，以根据所述纠错结果对所述目标文本和/或所述场景图片进行纠错处理

图6

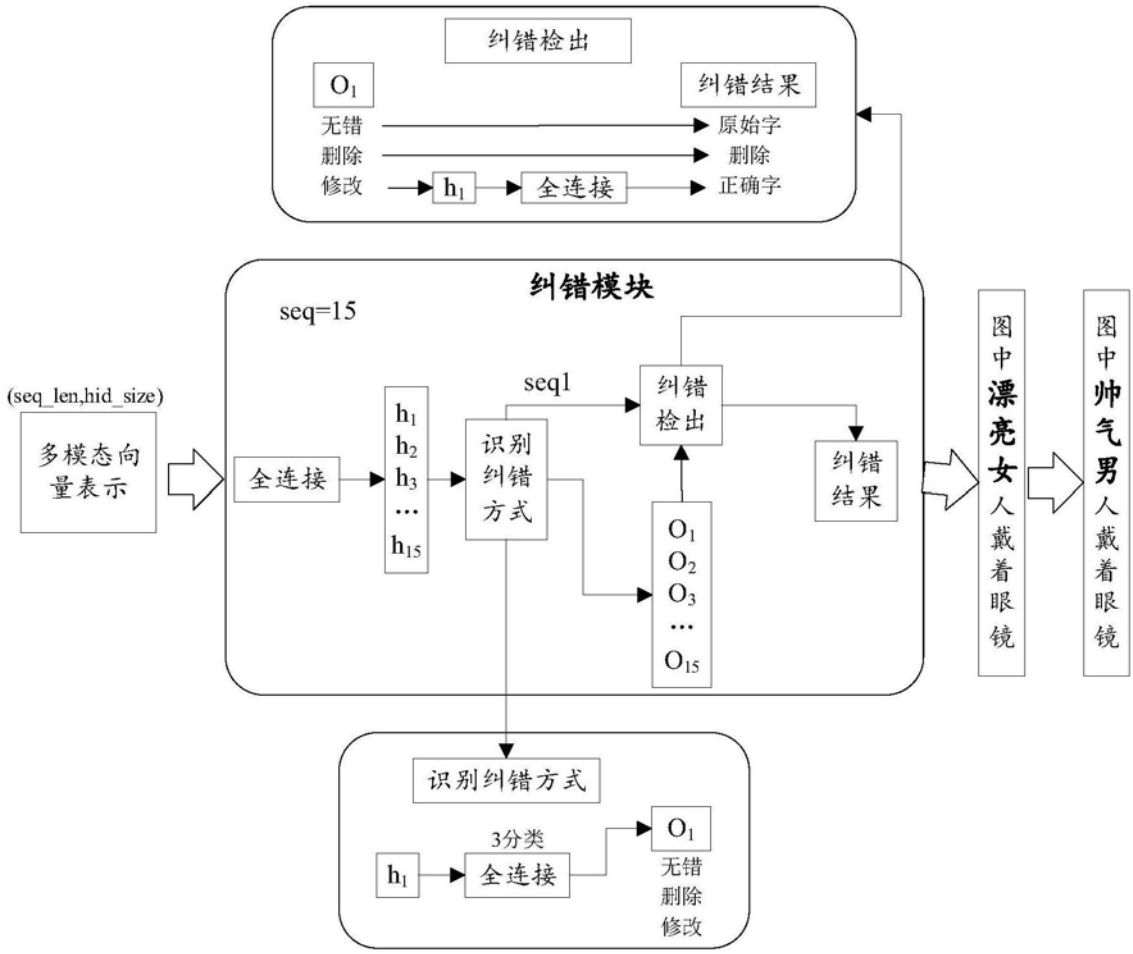

纠错检出
O1
纠错结果
无错
删除
修改
h1
全连接
原始字
删除
正确字
纠错模块
seq=15
(seq_len,hid_size)
多模态向量表示
全连接
h1
h2
h3
…
h15
识别纠错方式
seq1
纠错检出
纠错结果
O1
O2
O3
…
O15
图中漂亮女人戴着眼镜
图中帅气男人戴着眼镜
识别纠错方式
3分类
h1
全连接
O1
无错
删除
修改

图7

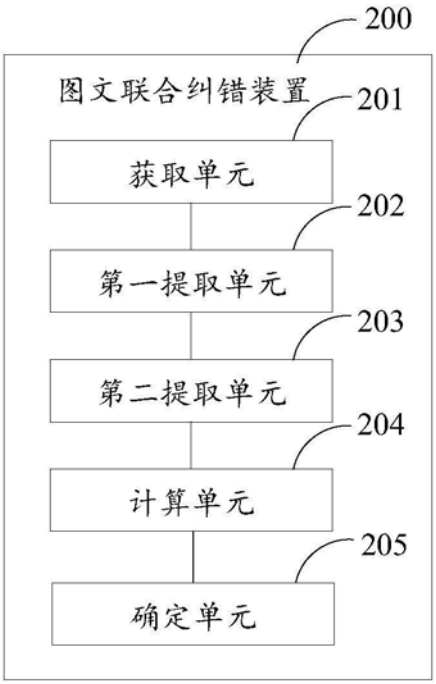

200
图文联合纠错装置
201
获取单元
202
第一提取单元
203
第二提取单元
204
计算单元
205
确定单元

图8

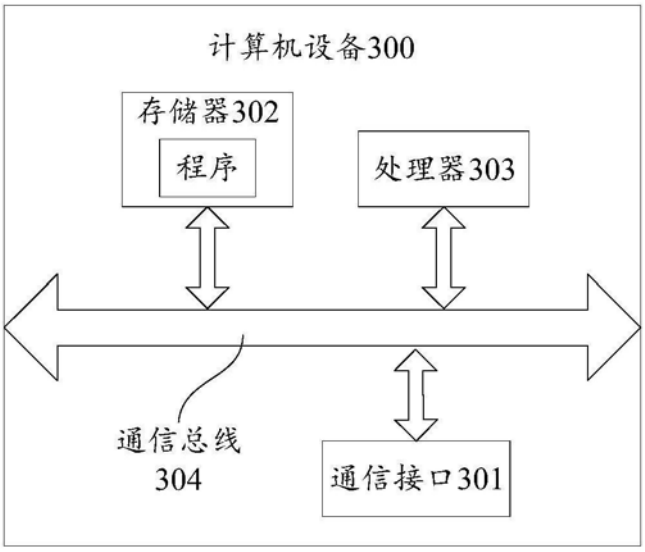

计算机设备300
存储器302
程序
处理器303
通信总线
304
通信接口301

图9