(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 112329748 B
(45)授权公告日 2021.04.30

(21)申请号 202110001077.0
(22)申请日 2021.01.04
(65)同一申请的已公布的文献号
申请公布号 CN 112329748 A
(43)申请公布日 2021.02.05
(73)专利权人 中国科学院自动化研究所
地址 100190 北京市海淀区中关村东路95号
(72)发明人 陶建华 连政 刘斌 孙立才
(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
代理人 苟冬梅
(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
(56)对比文件
CN 110969106 A,2020.04.07
CN 110222331 A,2019.09.10
CN 110427803 A,2019.11.08
CN 112017671 A,2020.12.01
CN 109793526 A,2019.05.24
CN 108520753 A,2018.09.11
Licai Sun等.Multi-modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism.《ACM》.2020,
Zheng Lian.Conversational Emotion Analysis via Attention Mechanisms.《arXiv》.2019,
审查员 王楠

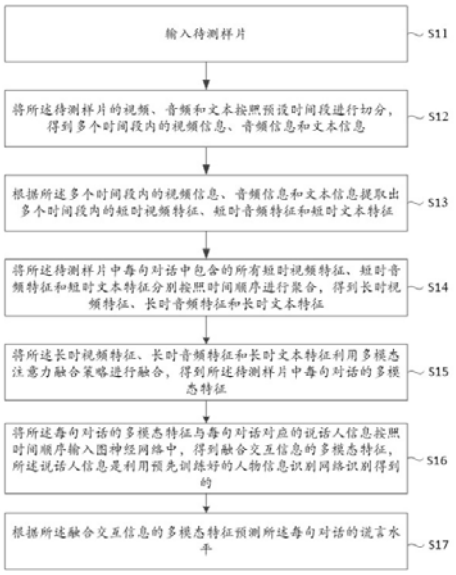
权利要求书4页 说明书15页 附图2页

(54)发明名称

面向交互场景的自动谎言检测方法、装置、设备及介质

(57)摘要

本申请实施例涉及一种面向交互场景的自动谎言检测方法、装置、设备及介质，旨在提高自动谎言检测的准确率。所述方法包括：将待测样片分别按照视频、音频和文本这三个模态进行切分，提取出这三个模态的短时特征，将待测样片中每句对话中三个模态的短时特征分别进行融合，得到每句对话对应的三个模态的长时特征，使用自注意力机制对三个模态的长时特征进行融合，得到每句对话的多模态特征，通过图神经网络为每句对话的多模态特征融合交互信息，得到融合交互信息的多模态特征，根据融合交互信息的多模态特征，预测每句对话的谎言水平。



CN 112329748 B

1.一种面向交互场景的自动谎言检测方法,其特征在于,所述方法包括:

输入待测样片;

将所述待测样片的视频、音频和文本按照预设时间段进行切分,得到多个时间段内的视频信息、音频信息和文本信息;

根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征;

将所述待测样片中每句对话中包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合,得到长时视频特征、长时音频特征和长时文本特征;

将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合,得到所述待测样片中每句对话的多模态特征;

将所述每句对话的多模态特征与每句对话对应的说话人信息按照时间顺序输入图神经网络中,得到融合交互信息的多模态特征,所述说话人信息是利用预先训练好的人物信息识别网络识别得到的;

根据所述融合交互信息的多模态特征预测所述每句对话的谎言水平;

其中,将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合,得到所述待测样片中每句对话的多模态特征的步骤包括:

将所述长时视频特征、长时音频特征和长时文本特征进行拼接,得到多模态拼接特征;

计算每个模态的注意力权重,依据所述注意力权重对所述多模态拼接特征进行融合,得到所述待测样片中每句对话的多模态特征,一个模态为视频、音频或文本;

其中,将所述每句对话的多模态特征与每句对话对应的说话人信息按照时间顺序输入图神经网络中,得到融合交互信息的多模态特征的步骤包括:

将所述每句对话的多模态特征作为所述图神经网络中每个句子节点的特征表示,将所述每句对话对应的说话人信息作为所述图神经网络中每个说话人节点的特征表示;

将所述图神经网络中每个句子节点与其邻域内的句子使用边进行连接,建立所述待测样片中每句对话之间的时序关系模型,得到所述待测样片中每句对话的时序关系;其中,连接所述句子节点的边分为两种类型,一种是前向边,一种是后向边,所述前向边的出发节点时序靠后,所述前向边指向的节点时序靠前,所述后向边的出发节点时序靠前,所述后向边指向的节点时序靠后;

将所述图神经网络中每个句子节点与该句子节点对应的说话人节点使用边进行连接,建立所述待测样片中每句对话的说话人信息模型,得到所述待测样片中每句对话与对应说话人之间的关系;

将所述每句对话的多模态特征与所述时序关系和所述每句对话与对应说话人之间的关系进行融合,得到所述融合交互信息的多模态特征。

2.根据权利要求1所述的方法,其特征在于,将所述待测样片的视频、音频和文本按照预设时间段进行切分,得到多个时间段内的视频信息、音频信息和文本信息,包括:

将所述待测样片的视频按照秒为单位进行切分,将所述待测样片中一秒内出现的所有的图像帧作为一个时间段内的视频信息,并将所述视频信息中的所有图像帧归一化,得到尺寸为第一预设尺寸的图像;

将所述待测样片的音频按照秒为单位进行切分,将所述待测样片中一秒内出现所有的

音频采样点作为一个时间内的音频信息；

将所述待测样片的文本按照秒为单位进行切分，将所述待测样片中一秒内说话人说出的文字作为一个时间段内的文本信息。

3.根据权利要求1所述的方法，其特征在于，根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征，包括：

从所述多个时间段内的短时视频信息中抽取出人脸区域图像；

对所述人脸区域图像进行矫正和归一化操作，得到尺寸为第二预设大小的标准人脸图像；

利用预先训练的表情识别网络，从所述标准人脸图像中抽取出隐含层特征表示，得到标准人脸特征；

对所述标准人脸特征进行特征统计，得到一个时间段内的所述标准人脸特征，作所为述短时视频特征；

通过计算得到所述多个时间段内的音频信息中的韵律特征、音质特征和普参数特征及其统计量，作为所述短时音频特征；

通过计算得到所述多个时间段内的文本信息中的词向量及其统计量，作为所述短时文本特征。

4. 根据权利要求1所述的方法，其特征在于，将所述待测样片中每句对话包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合，得到长时视频特征、长时音频特征和长时文本特征，包括：抽取所述待测样片中每句对话包含的所有短时视频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时视频特征进行计算，得到所述待测样片中每句对话的长时视频特征；

抽取所述待测样片中每句对话包含的所有短时音频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时音频特征进行计算，得到所述待测样片中每句对话的长时音频特征；

抽取所述待测样片中每句对话包含的所有短时文本特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时文本特征进行计算，得到所述待测样片中每句对话的长时文本特征。

5.根据权利要求4所述的方法，其特征在于，抽取所述待测样片中每句对话包含的所有短时视频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时视频特征进行计算，得到所述待测样片中每句对话的长时视频特征，包括：

将所述待测样片中每句对话包含的短时视频特征赋值到三个变量之中，对所述三个变量进行m次线性映射，记录从第一次映射到第m次映射中每一次映射后的三个变量；

对所述每一次映射后的三个变量进行注意力计算，根据计算得到的注意力权重进行加权求和，得到每一次映射后的短时视频特征；

将所述每一次映射后的短时视频特征按顺序进行拼接，得到融合时序信息的短时视频特征；

对所述融合时序信息的短时视频特征进行注意力计算，根据计算得到的注意力权重对所述融合时序信息的短时视频特征进行加权求和，得到所述待测样片中每句对话的长时视

频特征;所述长时音频特征与所述长时文本特征的计算方法与所述长时视频特征的计算方法相同。

6.一种面向交互场景的自动谎言检测装置,其特征在于,所述装置包括:

样片输入模块,用于输入待测样片;

样片切分模块,用于将所述待测样片的视频、音频和文本按照预设时间段进行切分,得到多个时间段内的视频信息、音频信息和文本信息;

短时特征提取模块,用于根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征;

长时特征提取模块,用于将所述待测样片中每句对话中包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合,得到长时视频特征、长时音频特征和长时文本特征;

多模态融合模块,用于将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合,得到所述待测样片中每句对话的多模态特征;

交互建模模块,用于将所述每句对话的多模态特征与每句对话对应的说话人信息按照时间顺序输入图神经网络中,得到融合交互信息的多模态特征,所述说话人信息是利用预先训练好的人物信息识别网络识别得到的;

谎言水平预测模块,用于根据所述融合交互信息的多模态特征预测所述每句对话的谎言水平;

其中,所述多模态融合模块包括:

特征拼接子模块,用于将所述长时视频特征、长时音频特征和长时文本特征进行拼接,得到多模态拼接特征;

多模态特征计算子模块,用于计算每个模态的注意力权重,依据所述注意力权重对所述多模态拼接特征进行融合,得到所述待测样片中每句对话的多模态特征,一个模态为视频、音频或文本;

其中,所述交互建模模块包括:

时序关系建立子模块,用于将所述图神经网络中每个句子节点与其邻域内的句子使用边进行连接,建立所述待测样片中每句对话之间的时序关系模型,得到所述待测样片中每句对话的时序关系;其中,连接所述句子节点的边分为两种类型,一种是前向边,一种是后向边,所述前向边的出发节点时序靠后,所述前向边指向的节点时序靠前,所述后向边的出发节点时序靠前,所述后向边指向的节点时序靠后;

人物关系建立子模块,用于将所述图神经网络中每个句子节点与该句子节点对应的说话人节点使用边进行连接,建立所述待测样片中每句对话的说话人信息模型,得到所述待测样片中每句对话与对应说话人之间的关系;

融合交互子模块,用于将所述每句对话的多模态特征与所述时序关系和所述每句对话与对应说话人之间的关系进行融合,得到所述融合交互信息的多模态特征。

7.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该计算机程序被处理器执行时,实现如权利要求1至5任一所述的方法中的步骤。

8.一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述计算机程序时,实现如权利要求1至5任一所述的

方法的步骤。

面向交互场景的自动谎言检测方法、装置、设备及介质

技术领域

[0001] 本申请实施例涉及谎言分析技术领域，具体而言，涉及一种面向交互场景的自动谎言检测方法、装置、设备及介质。

背景技术

[0002] 自动谎言检测技术是总结人在说谎和非说谎条件下，面部表情、语音、文本的变化规律，提出相应的模型和方法来建立起机器在捕获谎言线索方面的能力。自动谎言检测在安全领域具有广泛的应用前景，能够帮助刑侦人员加快审讯速度，提升审讯效率。现有的自动谎言检测方法是分别对人说每句话时的面部表情、语音进行分析，判断此人这句话是否是谎言。

[0003] 现有技术中存在的问题是，当被测人员说话时，只分析被测人员当前一句话的面部表情、语音，并未将被测人员的面部表情和语音结合起来进行考虑，忽略了被测人员人之前时刻的谎言状态对当前时刻的谎言状态的影响，还忽略了被测人员和与其对话人员之间的交互信息，检测结果不准确。

发明内容

[0004] 本申请实施例提供一种面向交互场景的自动谎言检测方法、装置、设备及介质，旨在提高自动谎言检测的准确率。

[0005] 本申请实施例第一方面提供一种面向交互场景的自动谎言检测方法，所述方法包括：

[0006] 输入待测样片；

[0007] 将所述待测样片的视频、音频和文本按照预设时间段进行切分，得到多个时间段内的视频信息、音频信息和文本信息；

[0008] 根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征；

[0009] 将所述待测样片中每句对话中包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合，得到长时视频特征、长时音频特征和长时文本特征；

[0010] 将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合，得到所述待测样片中每句对话的多模态特征；

[0011] 将所述每句对话的多模态特征与每句对话对应的说话人信息按照时间顺序输入图神经网络中，得到融合交互信息的多模态特征，所述说话人信息是利用预先训练好的人物信息识别网络识别得到的；

[0012] 根据所述融合交互信息的多模态特征预测所述每句对话的谎言水平。

[0013] 可选地，将所述待测样片的视频、音频和文本按照预设时间段进行切分，得到多个时间段内的视频信息、音频信息和文本信息，包括：

[0014] 将所述待测样片的视频按照秒为单位进行切分，将所述待测样片中一秒内出现的

所有的图像帧作为一个时间段内的视频信息，并将所述视频信息中的所有图像帧归一化，得到尺寸为第一预设尺寸的图像；

[0015] 将所述待测样片的音频按照秒为单位进行切分，将所述待测样片中一秒内出现所有的音频采样点作为一个时间内的音频信息；

[0016] 将所述待测样片的文本按照秒为单位进行切分，将所述待测样片中一秒内说话人说出的文字作为一个时间段内的文本信息。

[0017] 可选地，根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征，包括：

[0018] 从所述多个时间段内的短时视频信息中抽取出人脸区域图像；

[0019] 对所述人脸区域图像进行矫正和归一化操作，得到尺寸为第二预设大小的标准人脸图像；

[0020] 利用预先训练的表情识别网络，从所述标准人脸图像中抽取出隐含层特征表示，得到标准人脸特征；

[0021] 对所述标准人脸特征进行特征统计，得到一个时间段内的所述标准人脸特征，作为述短时视频特征；

[0022] 通过计算得到所述多个时间段内的音频信息中的韵律特征、音质特征和普参数特征及其统计量，作为所述短时音频特征；

[0023] 通过计算得到所述多个时间段内的文本信息中的词向量及其统计量，作为所述短时文本特征。

[0024] 可选地，将所述待测样片中每句对话包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合，得到长时视频特征、长时音频特征和长时文本特征，包括：

[0025] 抽取所述待测样片中每句对话包含的所有短时视频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时视频特征进行计算，得到所述待测样片中每句对话的长时视频特征；

[0026] 抽取所述待测样片中每句对话包含的所有短时音频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时音频特征进行计算，得到所述待测样片中每句对话的长时音频特征；

[0027] 抽取所述待测样片中每句对话包含的所有短时文本特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时文本特征进行计算，得到所述待测样片中每句对话的长时文本特征。

[0028] 可选地，将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合，得到所述待测样片中每句对话的多模态特征，所述说话人信息是利用预先训练好的人物信息识别网络识别得到的，包括：

[0029] 将所述长时视频特征、长时音频特征和长时文本特征进行拼接，得到多模态拼接特征；

[0030] 计算每个模态的注意力权重，依据所述注意力权重对所述多模态拼接特征进行融合，得到所述待测样片中每句对话的多模态特征，一个模态为视频、音频或文本。

[0031] 可选地，将所述每句对话的多模态特征与每句对话对应的说话人信息按照时间顺

序输入图神经网络中，得到融合交互信息的多模态特征，包括：

[0032] 将所述每句对话的多模态特征作为所述图神经网络中每个句子节点的特征表示，将所述每句对话对应的说话人信息作为所述图神经网络中每个说话人节点的特征表示；

[0033] 将所述图神经网络中每个句子节点与其邻域内的句子使用边进行连接，建立所述待测样片中每句对话之间的时序关系模型，得到所述待测样片中每句对话的时序关系；其中，连接所述句子节点的边分为两种类型，一种是前向边，一种是后向边，所述前向边的出发节点时序靠后，所述前向边指向的节点时序靠前，所述后向边的出发节点时序靠前，所述后向边指向的节点时序靠后；

[0034] 将所述图神经网络中每个句子节点与该句子节点对应的说话人节点使用边进行连接，建立所述待测样片中每句对话的说话人信息模型，得到所述待测样片中每句对话与对应说话人之间的关系；

[0035] 将所述每句对话的多模态特征与所述时序关系和所述每句对话与对应说话人之间的关系进行融合，得到所述融合交互信息的多模态特征。

[0036] 可选地，抽取所述待测样片中每句对话包含的所有短时视频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时视频特征进行计算，得到所述待测样片中每句对话的长时视频特征，包括：

[0037] 将所述待测样片中每句对话包含的短时视频特征赋值到三个变量之中，对所述三个变量进行m次线性映射，记录从第一次映射到第m次映射中每一次映射后的三个变量；

[0038] 对所述每一次映射后的三个变量进行注意力计算，根据计算得到的注意力权重进行加权求和，得到每一次映射后的短时视频特征；

[0039] 将所述每一次映射后的短时视频特征按顺序进行拼接，得到融合时序信息的短时视频特征；

[0040] 对所述融合时序信息的短时视频特征进行注意力计算，根据计算等得到的注意力权重对所述融合时序信息的短时视频特征进行加权求和，得到所述待测样片中每句对话的长时视频特征；所述长时音频特征与所述长时文本特征的计算方法与所述长时视频特征的计算方法相同。

[0041] 本申请实施例第二方面提供一种面向交互场景的自动谎言检测装置，所述装置包括：

[0042] 样片输入模块，用于输入待测样片；

[0043] 样片切分模块，用于将所述待测样片的视频、音频和文本按照预设时间段进行切分，得到多个时间段内的视频信息、音频信息和文本信息；

[0044] 短时特征提取模块，用于根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征；

[0045] 长时特征提取模块，用于将所述待测样片中每句对话中包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合，得到长时视频特征、长时音频特征和长时文本特征；

[0046] 多模态融合模块，用于将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合，得到所述待测样片中每句对话的多模态特征；

[0047] 交互建模模块，用于将所述每句对话的多模态特征与每句对话对应的说话人信息

按照时间顺序输入图神经网络中，得到融合交互信息的多模态特征，所述说话人信息是利用预先训练好的人物信息识别网络识别得到的；

[0048] 谎言水平预测模块，用于根据所述融合交互信息的多模态特征预测所述每句对话的谎言水平。

[0049] 可选地，所述样片切分模块包括：

[0050] 视频切分子模块，用于将所述待测样片的视频按照秒为单位进行切分，将所述待测样片中一秒内出现的所有的图像帧作为一个时间段内的视频信息，并将所述视频信息中的所有图像帧归一化，得到尺寸为第一预设尺寸的图像；

[0051] 音频切分子模块，用于将所述待测样片的音频按照秒为单位进行切分，将所述待测样片中一秒内出现所有的音频采样点作为一个时间内的音频信息；

[0052] 文本切分子模块，用于将所述待测样片的文本按照秒为单位进行切分，将所述待测样片中一秒内说话人说出的文字作为一个时间段内的文本信息。

[0053] 可选地，所述短时特征提取模块包括：

[0054] 第一图像抽取子模块，用于从所述多个时间段内的短时视频信息中抽取出人脸区域图像；

[0055] 第一图像处理子模块，用于对所述人脸区域图像进行矫正和归一化操作，得到尺寸为第二预设大小的标准人脸图像；

[0056] 表情识别子模块，用于利用预先训练的表情识别网络，从所述标准人脸图像中抽取出隐含层特征表示，得到标准人脸特征；

[0057] 短时视频特征提取子模块，用于对所述标准人脸特征进行特征统计，得到一个时间段内的所述标准人脸特征，作为述短时视频特征；

[0058] 短时音频特征提取子模块，用于通过计算得到所述多个时间段内的音频信息中的韵律特征、音质特征和普参数特征及其统计量，作为所述短时音频特征；

[0059] 短时文本特征提取子模块，用于通过计算得到所述多个时间段内的文本信息中的词向量及其统计量，作为所述短时文本特征。

[0060] 可选地，所述长时特征提取模块包括：

[0061] 长时视频特征提取子模块，用于抽取所述待测样片中每句对话包含的所有短时视频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时视频特征进行计算，得到所述待测样片中每句对话的长时视频特征；

[0062] 长时音频特征提取子模块，用于抽取所述待测样片中每句对话包含的所有短时音频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时音频特征进行计算，得到所述待测样片中每句对话的长时音频特征；

[0063] 长时文本特征提取子模块，用于抽取所述待测样片中每句对话包含的所有短时文本特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时文本特征进行计算，得到所述待测样片中每句对话的长时文本特征。

[0064] 可选地，所述多模态特征融合模块包括：

[0065] 特征拼接子模块，用于将所述长时视频特征、长时音频特征和长时文本特征进行拼接，得到多模态拼接特征；

[0066] 多模态特征计算子模块，用于计算每个模态的注意力权重，依据所述注意力权重

对所述多模态拼接特征进行融合，得到所述待测样片中每句对话的多模态特征，一个模态为视频、音频或文本。

[0067] 可选地，所述交互建模模块包括：

[0068] 节点表示子模块，用于将所述每句对话的多模态特征作为所述图神经网络中每个句子节点的特征表示，将所述每句对话对应的说话人信息作为所述图神经网络中每个说话人节点的特征表示；

[0069] 时序关系建立子模块，用于将所述图神经网络中每个句子节点与其邻域内的句子使用边进行连接，建立所述待测样片中每句对话之间的时序关系模型，得到所述待测样片中每句对话的时序关系；其中，连接所述句子节点的边分为两种类型，一种是前向边，一种是后向边，所述前向边的出发节点时序靠后，所述前向边指向的节点时序靠前，所述后向边的出发节点时序靠前，所述后向边指向的节点时序靠后；

[0070] 人物关系建立子模块，用于将所述图神经网络中每个句子节点与该句子节点对应的说话人节点使用边进行连接，建立所述待测样片中每句对话的说话人信息模型，得到所述待测样片中每句对话与对应说话人之间的关系；

[0071] 融合交互子模块，用于将所述每句对话的多模态特征与所述时序关系和所述每句对话与对应说话人之间的关系进行融合，得到所述融合交互信息的多模态特征。

[0072] 可选的，所述长时视频特征提取子模块包括：

[0073] 特征映射子模块，用于将所述待测样片中每句对话包含的短时视频特征赋值到三个变量之中，对所述三个变量进行m次线性映射，记录从第一次映射到第m次映射中每一次映射后的三个变量；

[0074] 第一短时视频特征计算子模块，用于对所述每一次映射后的三个变量进行注意力计算，根据计算得到的注意力权重进行加权求和，得到每一次映射后的短时视频特征；

[0075] 第二短时视频特征计算子模块，用于将所述每一次映射后的短时视频特征按顺序进行拼接，得到融合时序信息的短时视频特征；

[0076] 长时视频特征计算子模块，用于对所述融合时序信息的短时视频特征进行注意力计算，根据计算等得到的注意力权重对所述融合时序信息的短时视频特征进行加权求和，得到所述待测样片中每句对话的长时视频特征；所述长时音频特征与所述长时文本特征的计算方法与所述长时视频特征的计算方法相同。

[0077] 本申请实施例第三方面提供一种可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时，实现如本申请第一方面所述的方法中的步骤。

[0078] 本申请实施例第四方面提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述计算机程序时，实现本申请第一方面所述的方法的步骤。

[0079] 采用本申请提供的自动谎言检测方法，对待测样片中的视频、音频和文本分别以秒为单位进行切分，得到短时视频信息、短时音频信息和短时文本信息。从多个短时视频信息、短时音频信息和短时文本信息中提取出多个短时视频特征、短时音频特征和短时文本特征。将多个短时视频信息聚合成句子级别的长时视频特征，将多个短时音频特征聚合成句子级别的长时音频特征，将多个短时文本特征聚合成句子级别的长时文本特征。利用自注意力机制将长时视频特征、长时音频特征和长时文本特征融合成多模态特征，将多模态

特征和人物信息特征输入图神经网络中，得到融合了交互信息的多模态特征，人物信息特征是由预先训练好的人物信息识别网络提取出来的。根据融合了交互信息的多模态特征对对话的谎言水平进行预测。这种方法通过自注意力机制将视频、音频和文本特征融合成了多模态特征，将三种模态的特征结合起来进行考虑，并且通过图神经网络建模了对话的时序信息与每个句子与对应的说话人之间的信息，将多模态特征与交互信息进行融合考虑，提高了谎言检测的准确性。

附图说明

[0080] 为了更清楚地说明本申请实施例的技术方案，下面将对本申请实施例的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，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0081] 图1是本申请一实施例提出的面向交互场景的自动谎言检测方法的流程图；

[0082] 图2是本申请一实施例提出的面向交互场景的自动谎言检测装置的示意图。

具体实施方式

[0083] 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，本领域普通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。

[0084] 参考图1，图1是本申请一实施例提出的面向交互场景的自动谎言检测方法的流程图。如图1所示，该方法包括以下步骤：

[0085] S11：输入待测样片。

[0086] 本实施例中，待测样片可以是一段对话视频，对话视频中需要有人的面部特征，对话的语音即可，在此不做限制。

[0087] S12：将所述待测样片的视频、音频和文本按照预设时间段进行切分，得到多个时间段内的视频信息、音频信息和文本信息。

[0088] 本实施例中，待测样片中的画面是视频，待测样片中的语音是音频，待测样片中每句对话的内容是文本。将视频、音频和文本切分成较短的片段，可以更加有效的提取每句对话的特征，有助于谎言检测准确率的提升。

[0089] 本实施例中，将待测样片的视频、音频和文本按照预设时间段进行切分，得到多个时间段内的视频信息、音频信息和文本信息的具体步骤为：

[0090] S12-1：将所述待测样片的视频按照秒为单位进行切分，将所述待测样片中一秒内出现的所有的图像帧作为一个时间段内的视频信息，并将所述视频信息中的所有图像帧归一化，得到尺寸为第一预设尺寸的图像。

[0091] 本实施例中，从视频中可以看出人面部表情的特征变化，将视频以秒为单位进行切分，将一秒内的所有的图像帧作为一个时间段内的视频信息，所有的图像帧不能出现重叠，即切分时间不能出现重叠。例如，待测样片时长为60秒，待测样片的帧率为25帧/秒，则将这25帧的画面作为一个时间段内的视频信息，即取0-1秒内的25帧画面作为第一个时间

段内的视频信息，1-2秒内的25帧画面作为第二个时间段内的视频信息，直至59-60秒内的25帧画面作为第60个时间段内的视频信息。

[0092] 本实施例中，将视频中的所有图像帧进行归一化以便于对图像的进一步处理，例如，可以将所有图像帧归一化到300×300大小。

[0093] S12-2：将所述待测样片的音频按照秒为单位进行切分，将所述待测样片中一秒内出现所有的音频采样点作为一个时间内的音频信息。

[0094] 本实施例中，通过待测样片中的音频信息可以提取到说话人的声音，说话语气，说话速度等信息。将一秒内出现的所有的音频采样点作为一个时间段内的音频信息，音频采样点不能出现重叠，即切分的时间段不能出现重叠。例如，待测样片时长为60秒，待测样片的音频采样率为16000采样点/秒，那么，将0-1秒内的16000个采样点作为第一个时间段内的音频信息，将1-2秒内的16000个采样点作为第二个时间段内的音频信息，直至59-60秒内的16000个采样点作为第60个时间段内的音频信息。

[0095] S12-3：将所述待测样片的文本按照秒为单位进行切分，将所述待测样片中一秒内说话人说出的文字作为一个时间段内的文本信息。

[0096] 本实施例中，待测样片的文本为样片中说话人说话内容，结合说话人的用词，前后文的信息也有助于谎言的检测，将一秒内出现的所有文字作为一个时间段内的文本信息，文字不能出现重叠，即切分的时间段不能出现重叠。

[0097] 例如，待测样片时长为60秒，那么将0-1秒内出现的文字作为第一个时间段内的文本信息，将1-2秒内出现的文字作为第二个时间段内的文本信息，直至第59-60秒内出现的文字作为第60个时间段内的文本信息。

[0098] S13：根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征。

[0099] 本实施例中，短时视频特征、短时音频特征，短时文本特征实质上是包含了短时视频特征、短时音频特征和短时文本特征的向量。

[0100] 本实施例中，根据多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征的具体步骤为：

[0101] S13-1：从所述多个时间段内的短时视频信息中抽取出人脸区域图像。

[0102] 本实施例中，神经网络对接收到的短时视频信息进行处理，可以抽取出人脸区域的图像，抽取出人脸区域图像，用于对说话人的面部特征进行提取。

[0103] S13-2：对所述人脸区域图像进行矫正和归一化操作，得到尺寸为第二预设大小的标准人脸图像。

[0104] 本实施例中，先对人脸区域图像进行矫正，矫正的目的是将人脸区域图像中的人脸矫正到一个正常角度，这样有利于人脸信息的提取。将人脸矫正到正常角度后将图像再次进行归一化，得到尺寸为第二预设大小的标准人脸图像，例如100×100大小的标准人脸图像。

[0105] S13-3：利用预先训练的表情识别网络，从所述标准人脸图像中抽取出隐含层特征表示，得到标准人脸特征。

[0106] 本实施例中，利用收集的好的标注了表情的图片可以训练出表情识别网络，从标准人脸图像中抽取出隐含层特征表示，就是人脸图像进入神经网络之后，隐藏层对人脸图

像的每个像素点进行分析，从而的到人的面部表情特征。

[0107] S13-4：对所述标准人脸特征进行特征统计，得到一个时间段内的所述标准人脸特征，作为述短时视频特征。

[0108] 本实施例中，一个时间段内的每一帧中都有一张人脸图片，对一个时间段内的所有人脸特征进行统计，将统计到的人脸特征作为短时视频特征。短时视频特征中可以包括人嘴角上扬的角度，人的瞳孔大小，人的眉头是否紧皱等。

[0109] S13-5：通过计算得到所述多个时间段内的音频信息中的韵律特征、音质特征和普参数特征及其统计量，作为所述短时音频特征。

[0110] 本实施例中，韵律特征包括了音高、音长和音强等特征，音质特征可以随着不同人改变，普参数特征是一个人声纹形成的频谱，每个人因为声道的形状不同，普参数都不同。通过计算每个时间段内的所有采样点，可以得出音频信息中的韵律特征、音质特征和普参数特征。统计量是对这些数据进行分析、检验的变量，有效反映了音频特征。

[0111] S13-6：通过计算得到所述多个时间段内的文本信息中的词向量及其统计量，作为所述短时文本特征。

[0112] 本实施例中，对每个时间段内的短时文本信息中的词向量进行计算，根据这些词向量以及词向量的统计量，可以得到短时文本特征。

[0113] S14：将所述待测样片中每句对话中包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合，得到长时视频特征、长时音频特征和长时文本特征。

[0114] 本实施例中，长时视频特征、长时音频特征和长时文本特征是由短时视频特征、短时音频特征和短时文本特征分别聚合而成的，聚合出的长时特征是待测样片中每个句子的特征。

[0115] 本实施例中，将待测样片中每句对话中包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合，得到长时视频特征、长音频特征和长时文本特征的具体步骤是：

[0116] S14-1：抽取所述待测样片中每句对话包含的所有短时视频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时视频特征进行计算，得到所述待测样片中每句对话的长时视频特征。

[0117] 本实施例中，自注意力机制是对注意力机制的改进，相较于注意力机制，自注意力机制减少了对外部信息的依赖，更擅长捕捉数据或特征的内部相关性。使用自注意力机制对待测样片中每句对话包含的所有短时视频特征计算，得到待测样片中每句对话的长时视频特征的具体步骤包括：

[0118] S14-1-1：将所述待测样片中每句对话包含的短时视频特征赋值到三个变量之中，对所述三个变量进行m次线性映射，记录从第一次映射到第m次映射中每一次映射后的三个变量。

[0119] 本实施例中，假设待测样片中有对话假设对话$U=[u_1,\cdots,u_N]$，其中$u_j(j=1,\cdots,N)$是对话中的第j句话。从u_j中抽取音频模态或文本模型的短时特征，标记为S_j。将S_j赋值到三个变量Q,K,V中，每个变量经过m次线性映射，从而得到：

[0120] $Q_i = QW_i^Q (i=1, \cdots, m)$ (1)

[0121] $K_i = KW_i^K (i=1, \cdots, m)$ (2)

[0122] $V_i = VW_i^V (i=1, \cdots, m)$ (3)

[0123] 其中W_i^Q, W_i^K, W_i^V分别是Q,K,V的映射参数，Q_i，K_i，V_i变量经过第m线性映射后的值。

[0124] S14-1-2：对所述每一次映射后的三个变量进行注意力计算，根据计算得到的注意力权重进行加权求和，得到每一次映射后的短时视频特征。

[0125] 本实施例中，将Q_i，K_i，V_i输入到注意力模块中，注意力模块的输出标记为$head_i$：

[0126] $head_i = softmax(Q_i K_i^T) V_i$ (4)

[0127] 其中$softmax(Q_i K_i^T)$计算的是注意力权重，用于对V_i进行加权求和，$head_i$为第m次映射后的短时视频特征。

[0128] S14-1-3：将所述每一次映射后的短时视频特征按顺序进行拼接，得到融合时序信息的短时视频特征。

[0129] 本实施例中，将$head_i (i=1, \cdots, m)$拼接起来，作为融合时序信息的短时特征Z：

[0130] $Z = Concat(head_1, \cdots, head_m)$ (5)

[0131] 其中Z是融合了视讯信息的短时特征，contact（）为将向量进行拼接。

[0132] S14-1-4：对所述融合时序信息的短时视频特征进行注意力计算，根据计算等得到的注意力权重对所述融合时序信息的短时视频特征进行加权求和，得到所述待测样片中每句对话的长时视频特征；所述长时音频特征与所述长时文本特征的计算方法与所述长时视频特征的计算方法相同。

[0133] 将短时特征Z输入到段级别注意力模块中，从而获取长时特征G。段级别注意力模块能够关注于重点段信息，提升谎言特征的区分性：

[0134] $A = softmax(ZW_z)$ (6)

[0135] $G = AZ$ (7)

[0136] 其中softmax（）为用于计算的函数，W_z为可训练参数，A为每段特征所对应的注意力权重，G为基于注意力A对短时特征Z加权融合后的结果，即每段对话的长时视频特征。

[0137] S14-2抽取所述待测样片中每句对话包含的所有短时音频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时音频特征进行计算，得到所述待测样片中每句对话的长时音频特征。

[0138] 本实施例中，长时音频同样可用S14-1-1至S14-1-4计算得到，在此不再叙述。

[0139] S14-3：抽取所述待测样片中每句对话包含的所有短时文本特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时文本特征进行计算，得到所述待测样片中每

句对话的长时文本特征。

[0140] 本实施例中，长时文本特征同样可用S14-1-1至S14-1-4计算得到，在此不再叙述。

[0141] S15：将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合，得到所述待测样片中每句对话的多模态特征。

[0142] 本实施例中，利用多模态注意力融合策略可从视频特征、音频特征和文本特征三个模态进行谎言检测。将长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合，得到待测样片中每句话的多模态特征的具体步骤为：

[0143] S15-1：将所述长时视频特征、长时音频特征和长时文本特征进行拼接，得到多模态拼接特征。

[0144] 本实施例中，将三个模态的长时特征拼接成为多模态特征，首先要将每句对话中三个模态的特征向量直接进行拼接，得到多模态拼接特征。假设G_A为长时音频特征，G_L为长时文本特征，G_V为长时面部特征，将多模态特征拼接：

[0145] $$G_{cat}=Concat(G_A,G_L,G_V) \tag{8}$$

[0146] 其中G_{cat}为拼接后的向量，contact( )表示将三个向量进行拼接。

[0147] S15-2：计算每个模态的注意力权重，依据所述注意力权重对所述多模态拼接特征进行融合，得到所述待测样片中每句对话的多模态特征，一个模态为视频、音频或文本。

[0148] 本实施例中，计算每个模态的注意力权重，依据注意力权重融合多模态特征：

[0149] $$P_F=tanh(W_F G_{cat}) \tag{9}$$

[0150] $$A_F=softmax(w_F^T P_F) \tag{10}$$

[0151] $$F=G_{cat}A_F \tag{11}$$

[0152] 其中，W_F和w_F为可训练参数，A_F为每个模态的注意力权重。P_F为中间值，F为多模态特征。

[0153] 本实施例中，依据注意力机制将三个模态的长时特征融合成为多模态特征，神经网络在接收到每句话的多模态特征时会重点分析权重较高的部分，例如待测样片中说话人说一个句子时表情和说的话正常，但说话声音突然加重或减小，这时神经网络就会重点分析多模态特征中表征此人说话声音的部分。

[0154] S16：将所述每句对话的多模态特征与每句对话对应的说话人信息按照时间顺序输入图神经网络中，得到融合交互信息的多模态特征，所述说话人信息是利用预先训练好的人物信息识别网络识别得到的。

[0155] 本实施例中，利用预先标注好的人物图片和对应的人物信息作为训练集，对人物信息识别网络进行训练，得到训练好的人物信息识别网络，人物信息识别网络可以在接收到人物图片时识别出人物的特征信息，例如身份信息等。

[0156] 本实施例中，将每句对话的多模态特征与每句对话对应的说话人信息按照时间顺序输入图神经网络中，得到融合交互信息的多模态特征的具体步骤为：

[0157] S16-1:将所述每句对话的多模态特征作为所述图神经网络中每个句子节点的特征表示,将所述每句对话对应的说话人信息作为所述图神经网络中每个说话人节点的特征表示。

[0158] 本实施例中,图神经网络也是图卷积网络,由多个节点和边构成。假设对话$U=[u_1,\cdots,u_N]$,其中$u_j(j=1,\cdots,N)$是对话中的第j句话,每个句子经过多模态特征融合模块后,获取每句话的多模态特征表示$f_j(j=1,\cdots,N)$。将整个对话U表示为图神经网络,其中包括两种类型节点:句子节点和说话人节点。采用f_j作为每个句子节点的特征表示,采用预先训练好的人物信息识别网络识别出的人物特征作为说话人节点,例如,可以是x-vector,x-vector是一种神经网络,在人物信息识别上效果较好。

[0159] S16-2:将所述图神经网络中每个句子节点与其邻域内的句子使用边进行连接,建立所述待测样片中每句对话之间的时序关系模型,得到所述待测样片中每句对话的时序关系;其中,连接所述句子节点的边分为两种类型,一种是前向边,一种是后向边,所述前向边的出发节点时序靠后,所述前向边指向的节点时序靠前,所述后向边的出发节点时序靠前,所述后向边指向的节点时序靠后。

[0160] S16-3:将所述图神经网络中每个句子节点与该句子节点对应的说话人节点使用边进行连接,建立所述待测样片中每句对话的说话人信息模型,得到所述待测样片中每句对话与对应说话人之间的关系。

[0161] S16-4:将所述每句对话的多模态特征与所述时序关系和所述每句对话与对应说话人之间的关系进行融合,得到所述融合交互信息的多模态特征。

[0162] 本实施例中,多模态特征经过图神经网络后,获取了融合交互信息的特征表示。对于单层图神经网络而言:

[0163] $$h_i^{(1)}=ReLU\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{w_{ij}}{|N_i^r|}W_r^{(1)}h_j^{(0)}\right) \quad (12)$$

[0164] 其中,$Re\mathrm{LU}$ ()为线性整流函数,R为边类型集合,N_i^r为在关系r 下与节点i相连的节点集合,$|N_i^r|$为集合N_i^r包含的元素个数。w_{ij}为节点i和节点j之间边的权重,$W_r^{(1)}$为关系r所对应的权重参数,$h_j^{(0)}$为节点j的初始特征,对于句子节点$h_j^{(0)}$为f_j,对于说话人节点,$h_j^{(0)}$为人物特征向量,h_i^1是单层神经网络中节点i输出的融合交互信息的多模态特征。

[0165] 对于多层图神经网络而言:

[0166] $$h_i^{(l)}=ReLU\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{w_{ij}}{|N_i^r|}W_r^{(l)}h_j^{(l-1)}\right) \quad (13)$$

[0167] 假设有L层网络,那么最终节点i的输出为$h_i^{(L)}$,h_i^l是L层网络中节点i输出的融

合交互信息的多模态特征。

[0168] 本实施例中，在多模态特征中融合了时序信息和交互信息，通过时序信息可以考虑到前一时刻的谎言状态对当前时刻谎言状态的影响，例如前一时刻说的话是谎言的几率较大，那么这一时刻说的话是谎言的几率也会随之提高。交互信息同样会影响人的谎言状态，例如在审讯过程中，被审讯人员可能会受到审讯人员的影响而改变谎言状态。

[0169] S17：根据所述融合交互信息的多模态特征预测所述每句对话的谎言水平。

[0170] 本实施例中，图神经网络输出融合交互信息的多模态特征，根据每句对话的多模态特征对对话的谎言水平进行预测。例如可以使用单层感知机来预测每句对话的谎言水平，计算方法为：

[0171] $$P_i = softmax(h_i W_l) \qquad (14)$$

[0172] 其中$soft\ max\ (\ )$为用于计算的函数，W_l是单层感知机的参数。P_i 是谎言水平预测结果，预测结果显示出来是这句话为真话的概率和这句话为假话的概率，h_i 是节 i 点输出的融合交互信息的多模态特征。

[0173] 采用本实施例的S11-S17，首先将待测样片进行切分，得到多段短时视频信息、短时音频信息和短时文本信息，从多段短时视频信息、短时音频信息和短时文本信息中提取出短时视频特征、短时音频特征和短时文本特征，利用自注意力机制将每句对话的所有短时视频特征、短时音频特征和短时文本特征进行融合，得到每句对话的长时视频特征、长时音频特征和长时文本特征，利用利用多模态注意力融合策略将每句对话的长时视频特征、长时音频特征和长时文本特征进行融合，得到每句对话的多模态特征，将每句对话的多模态特征与人物特征输入图神经网络中，得到融合交互信息的多模态特征，根据融合交互信息的多模态特征来预测每句对话的谎言水平。这样结合了多模态信息和交互信息进行谎言检测，大大提高了谎言检测的准确性。

[0174] 基于同一发明构思，本申请一实施例提供一种面向交互场景的自动谎言检测装置。参考图2，图2是本申请一实施例提出的面向交互场景的自动谎言检测装置200的示意图。如图2所示，该装置包括：

[0175] 样片输入模块201，用于输入待测样片；

[0176] 样片切分模块202，用于将所述待测样片的视频、音频和文本按照预设时间段进行切分，得到多个时间段内的视频信息、音频信息和文本信息；

[0177] 短时特征提取模块203，用于根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征；

[0178] 长时特征提取模块204，用于将所述待测样片中每句对话中包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合，得到长时视频特征、长时音频特征和长时文本特征；

[0179] 多模态融合模块205，用于将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合，得到所述待测样片中每句对话的多模态特征；

[0180] 交互建模模块206，用于将所述每句对话的多模态特征与每句对话对应的说话人信息按照时间顺序输入图神经网络中，得到融合交互信息的多模态特征，所述说话人信息

是利用预先训练好的人物信息识别网络识别得到的；

[0181] 谎言水平预测模块207，用于根据所述融合交互信息的多模态特征预测所述每句对话的谎言水平。

[0182] 可选地，所述样片切分模块包括：

[0183] 视频切分子模块，用于将所述待测样片的视频按照秒为单位进行切分，将所述待测样片中一秒内出现的所有的图像帧作为一个时间段内的视频信息，并将所述视频信息中的所有图像帧归一化，得到尺寸为第一预设尺寸的图像；

[0184] 音频切分子模块，用于将所述待测样片的音频按照秒为单位进行切分，将所述待测样片中一秒内出现所有的音频采样点作为一个时间内的音频信息；

[0185] 文本切分子模块，用于将所述待测样片的文本按照秒为单位进行切分，将所述待测样片中一秒内说话人说出的文字作为一个时间段内的文本信息。

[0186] 可选地，所述短时特征提取模块包括：

[0187] 第一图像抽取子模块，用于从所述多个时间段内的短时视频信息中抽取出人脸区域图像；

[0188] 第一图像处理子模块，用于对所述人脸区域图像进行矫正和归一化操作，得到尺寸为第二预设大小的标准人脸图像；

[0189] 表情识别子模块，用于利用预先训练的表情识别网络，从所述标准人脸图像中抽取出隐含层特征表示，得到标准人脸特征；

[0190] 短时视频特征提取子模块，用于对所述标准人脸特征进行特征统计，得到一个时间段内的所述标准人脸特征，作为述短时视频特征；

[0191] 短时音频特征提取子模块，用于通过计算得到所述多个时间段内的音频信息中的韵律特征、音质特征和普参数特征及其统计量，作为所述短时音频特征；

[0192] 短时文本特征提取子模块，用于通过计算得到所述多个时间段内的文本信息中的词向量及其统计量，作为所述短时文本特征。

[0193] 可选地，所述长时特征提取模块包括：

[0194] 长时视频特征提取子模块，用于抽取所述待测样片中每句对话包含的所有短时视频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时视频特征进行计算，得到所述待测样片中每句对话的长时视频特征；

[0195] 长时音频特征提取子模块，用于抽取所述待测样片中每句对话包含的所有短时音频特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时音频特征进行计算，得到所述待测样片中每句对话的长时音频特征；

[0196] 长时文本特征提取子模块，用于抽取所述待测样片中每句对话包含的所有短时文本特征，利用自注意力机制对所述待测样片中每句对话包含的所有短时文本特征进行计算，得到所述待测样片中每句对话的长时文本特征。

[0197] 可选地，所述多模态特征融合模块包括：

[0198] 特征拼接子模块，用于将所述长时视频特征、长时音频特征和长时文本特征进行拼接，得到多模态拼接特征；

[0199] 多模态特征计算子模块，用于计算每个模态的注意力权重，依据所述注意力权重对所述多模态拼接特征进行融合，得到所述待测样片中每句对话的多模态特征，一个模态

为视频、音频或文本。

[0200] 可选地，所述交互建模模块包括：

[0201] 节点表示子模块，用于将所述每句对话的多模态特征作为所述图神经网络中每个句子节点的特征表示，将所述每句对话对应的说话人信息作为所述图神经网络中每个说话人节点的特征表示；

[0202] 时序关系建立子模块，用于将所述图神经网络中每个句子节点与其邻域内的句子使用边进行连接，建立所述待测样片中每句对话之间的时序关系模型，得到所述待测样片中每句对话的时序关系；其中，连接所述句子节点的边分为两种类型，一种是前向边，一种是后向边，所述前向边的出发节点时序靠后，所述前向边指向的节点时序靠前，所述后向边的出发节点时序靠前，所述后向边指向的节点时序靠后；

[0203] 人物关系建立子模块，用于将所述图神经网络中每个句子节点与该句子节点对应的说话人节点使用边进行连接，建立所述待测样片中每句对话的说话人信息模型，得到所述待测样片中每句对话与对应说话人之间的关系；

[0204] 融合交互子模块，用于将所述每句对话的多模态特征与所述时序关系和所述每句对话与对应说话人之间的关系进行融合，得到所述融合交互信息的多模态特征。

[0205] 可选的，所述长时视频特征提取子模块包括：

[0206] 特征映射子模块，用于将所述待测样片中每句对话包含的短时视频特征赋值到三个变量之中，对所述三个变量进行m次线性映射，记录从第一次映射到第m次映射中每一次映射后的三个变量；

[0207] 第一短时视频特征计算子模块，用于对所述每一次映射后的三个变量进行注意力计算，根据计算得到的注意力权重进行加权求和，得到每一次映射后的短时视频特征；

[0208] 第二短时视频特征计算子模块，用于将所述每一次映射后的短时视频特征按顺序进行拼接，得到融合时序信息的短时视频特征；

[0209] 长时视频特征计算子模块，用于对所述融合时序信息的短时视频特征进行注意力计算，根据计算等得到的注意力权重对所述融合时序信息的短时视频特征进行加权求和，得到所述待测样片中每句对话的长时视频特征；所述长时音频特征与所述长时文本特征的计算方法与所述长时视频特征的计算方法相同。

[0210] 基于同一发明构思，本申请另一实施例提供一种可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如本申请上述任一实施例所述的面向交互场景的自动谎言检测方法中的步骤。

[0211] 基于同一发明构思，本申请另一实施例提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行时实现本申请上述任一实施例所述的面向交互场景的自动谎言检测方法中的步骤。

[0212] 对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0213] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0214] 本领域内的技术人员应明白，本申请实施例的实施例可提供为方法、装置、或计算机程序产品。因此，本申请实施例可采用完全硬件实施例、完全软件实施例、或结合软件和

硬件方面的实施例的形式。而且，本申请实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0215] 本申请实施例是参照根据本申请实施例的方法、终端设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0216] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0217] 这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上，使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0218] 尽管已描述了本申请实施例的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请实施例范围的所有变更和修改。

[0219] 最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者终端设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

[0220] 以上对本申请所提供的一种面向交互场景的自动谎言检测方法、装置、设备及介质，进行了详细介绍，本文中应用了具体个例对本申请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申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

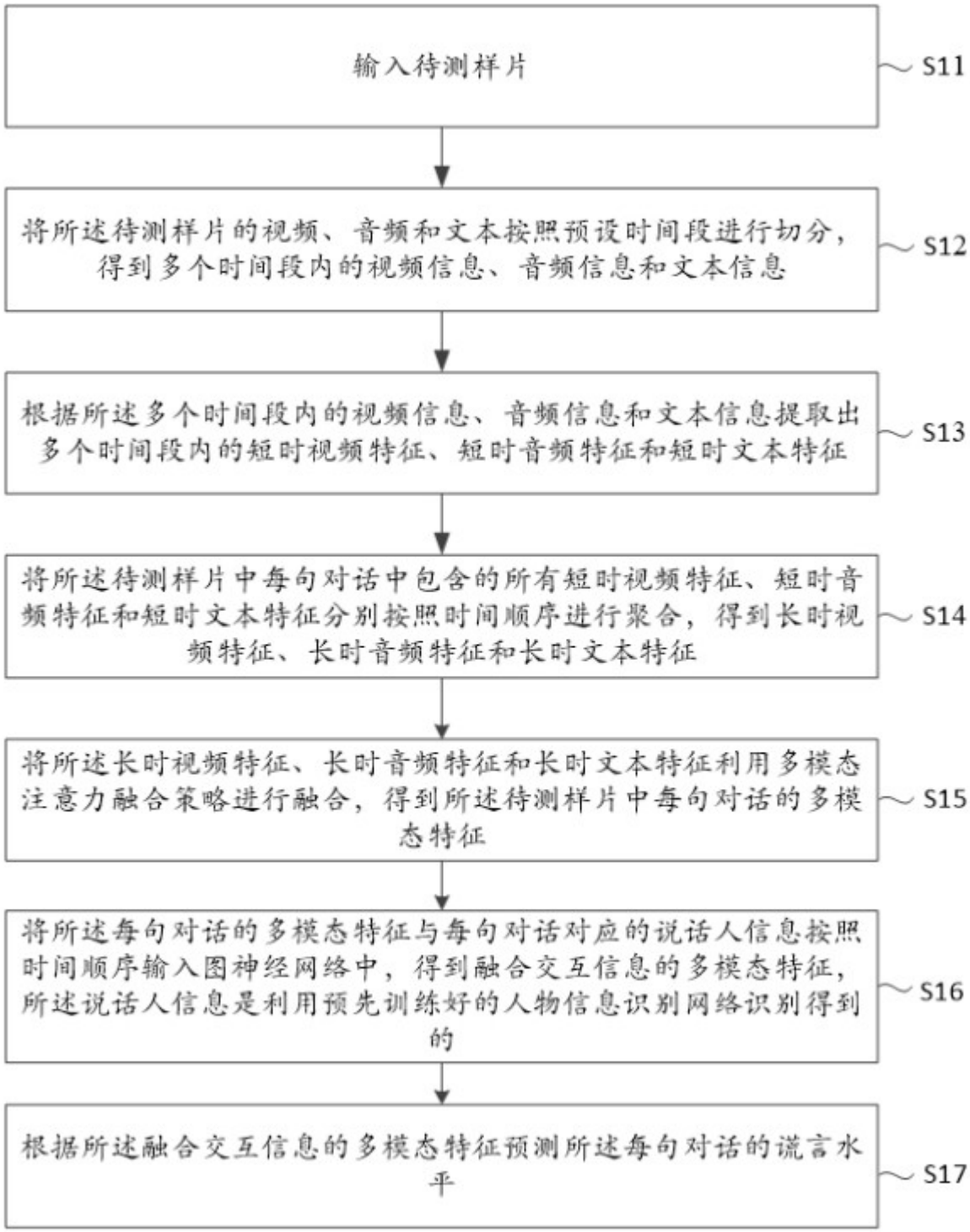

输入待测样片
S11
将所述待测样片的视频、音频和文本按照预设时间段进行切分，得到多个时间段内的视频信息、音频信息和文本信息
S12
根据所述多个时间段内的视频信息、音频信息和文本信息提取出多个时间段内的短时视频特征、短时音频特征和短时文本特征
S13
将所述待测样片中每句对话中包含的所有短时视频特征、短时音频特征和短时文本特征分别按照时间顺序进行聚合，得到长时视频特征、长时音频特征和长时文本特征
S14
将所述长时视频特征、长时音频特征和长时文本特征利用多模态注意力融合策略进行融合，得到所述待测样片中每句对话的多模态特征
S15
将所述每句对话的多模态特征与每句对话对应的说话人信息按照时间顺序输入图神经网络中，得到融合交互信息的多模态特征，所述说话人信息是利用预先训练好的人物信息识别网络识别得到的
S16
根据所述融合交互信息的多模态特征预测所述每句对话的谎言水平
S17

图1

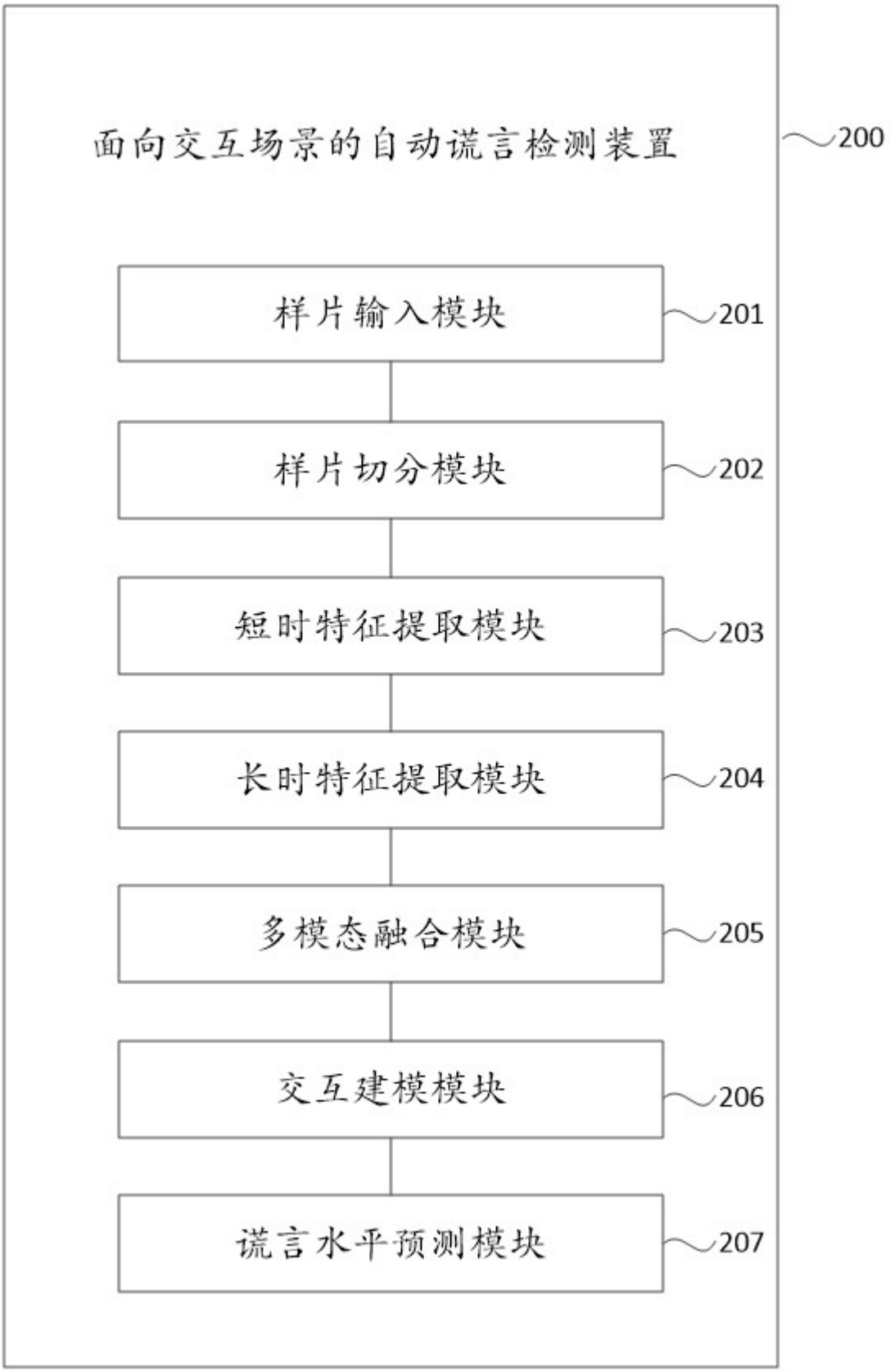

面向交互场景的自动谎言检测装置
200
样片输入模块
201
样片切分模块
202
短时特征提取模块
203
长时特征提取模块
204
多模态融合模块
205
交互建模模块
206
谎言水平预测模块
207

图2