



(12) 发明专利申请

(10) 申请公布号 CN 117711441 A

(43) 申请公布日 2024. 03. 15

(21) 申请号 202311542716.X

G10L 15/06 (2013.01)

(22) 申请日 2023.11.17

(71) 申请人 京东城市(北京)数字科技有限公司

地址 100086 北京市海淀区知春路76号(写字楼)1号楼9层1-7-5号

申请人 京东科技信息技术有限公司

(72) 发明人 曾孙禄 王培英 何晓冬 陈俊卿
范璐 陈蒙

(74) 专利代理机构 北京律智知识产权代理有限公司 11438

专利代理师 陈媛婧

(51) Int. Cl.

G10L 25/63 (2013.01)

G10L 25/06 (2013.01)

G10L 25/30 (2013.01)

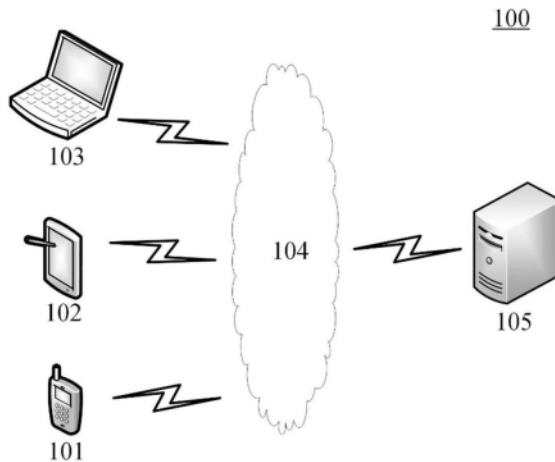
权利要求书2页 说明书10页 附图7页

(54) 发明名称

音频情绪识别模型的训练方法、装置、设备和可读介质

(57) 摘要

本公开提供一种音频情绪识别模型的训练方法、装置、设备和可读介质,其中,音频情绪识别模型的训练方法包括:提取音频样本中的关键帧的语音状态;根据预设的嵌入向量编码信息确定关键帧的语音状态对应的情绪标签嵌入向量;根据情绪标签嵌入向量和音频样本的独热向量确定损失函数的输入;基于损失函数对音频情绪识别模型进行训练。通过本公开实施例,使得音频情绪识别模型在训练过程中能够学习到更加泛化的情绪特征,更准确且可靠地进行情绪识别。



1. 一种音频情绪识别模型的训练方法,其特征在于,包括:
提取音频样本中的关键帧的语音状态;
根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量;
根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入;
基于所述损失函数对所述音频情绪识别模型进行训练。
2. 如权利要求1所述的音频情绪识别模型的训练方法,其特征在于,提取音频样本中的关键帧的语音状态包括:
将音频样本输入至音频情绪识别模型的预训练模型;
基于所述预训练模型的前端的CNN模块和量化模块获得帧级别的离散值,每一帧对应的离散值表征当前帧的语音状态;
采用TF-IDF关键词提取算法确定所述离散值对应的TF-IDF值;
根据所述TF-IDF值确定所述音频样本中的关键帧,并确定所述关键帧对应的语音状态。
3. 如权利要求1所述的音频情绪识别模型的训练方法,其特征在于,根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量包括:
获取音频情绪识别模型的预训练模型的嵌入向量编码信息;
根据所述嵌入向量编码信息确定所述音频样本中的关键帧的语音状态对应于的指定个数的嵌入向量;
对所述指定个数的嵌入向量进行平均化处理;
根据所述平均化处理的结果确定所述情绪标签嵌入向量。
4. 如权利要求1所述的音频情绪识别模型的训练方法,其特征在于,根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入包括:
将所述语音样本输入至音频情绪识别模型,并获取所述wav2vec2.0模型输出的语音特征图;
对所述语音特征图进行均值池化处理,并将所述均值池化处理的结果压缩为情绪嵌入向量;
将所述情绪嵌入向量输入至全连接层,以获得情绪分数向量;
根据所述情绪嵌入向量、所述情绪分数向量和所述独热向量确定所述损失函数的输入。
5. 如权利要求1-4中任一项所述的音频情绪识别模型的训练方法,其特征在于,还包括:
根据所述情绪分数向量与所述独热向量之间的交叉熵函数构建第一目标函数;
根据所述情绪标签嵌入向量与所述情绪嵌入向量之间的均方误差函数构建第二目标函数;
根据所述第一目标函数、所述第二目标函数和预设的权重值确定所述损失函数。
6. 如权利要求1-4中任一项所述的音频情绪识别模型的训练方法,其特征在于,还包括:
在对所述音频情绪识别模型进行训练的过程中,保持所述CNN模块的指定参数不更新。
7. 如权利要求1-4中任一项所述的音频情绪识别模型的训练方法,其特征在于,还包

括：

设置所述音频情绪识别模型的transformer网络层学习率不超过0.0001,和/或控制所述音频情绪识别模型的训练周期数小于或等于20次。

8.一种音频情绪识别模型的训练装置,其特征在于,包括：

提取模块,设置为提取音频样本中的关键帧的语音状态；

确定模块,设置为根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量；

所述确定模块,设置为根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入；

训练模块,设置为基于所述损失函数对所述音频情绪识别模型进行训练。

9.一种电子设备,其特征在于,包括：

存储器；以及

耦合到所述存储器的处理器,所述处理器被配置为基于存储在所述存储器中的指令,执行如权利要求1-7任一项所述的音频情绪识别模型的训练方法。

10.一种计算机可读存储介质,其上存储有程序,该程序被处理器执行时实现如权利要求1-7任一项所述的音频情绪识别模型的训练方法。

音频情绪识别模型的训练方法、装置、设备和可读介质

技术领域

[0001] 本公开涉及机器识别技术领域,具体而言,涉及一种音频情绪识别模型的训练方法、装置、设备和可读介质。

背景技术

[0002] 目前,现有的语音情绪识别技术,主要是利用神经网络对音频进行特征提取与情绪解析,得到给定音频中的说话人的情绪分类。

[0003] 在相关技术中,将独热向量作为训练标签对音频情绪识别模型进行训练。

[0004] 但是,由于独热向量是一种独立的、无语义信息的表征,忽略了情绪类别本身的先验知识,可能会导致音频情绪识别模型在进行情绪特征提取时会忽略一些有用的潜在类别信息,使得音频情绪识别模型在一些复杂应用场景上的泛化能力较差。

[0005] 需要说明的是,在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解,因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

[0006] 本公开的目的在于提供一种音频情绪识别模型的训练方法、装置、设备和可读介质,用于至少在一定程度上克服由于相关技术的限制和缺陷而导致的音频情绪识别模型的泛化性差等问题。

[0007] 根据本公开实施例的第一方面,提供一种音频情绪识别模型的训练方法,包括:提取音频样本中的关键帧的语音状态;根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量;根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入;基于所述损失函数对所述音频情绪识别模型进行训练。

[0008] 在本公开的一种示例性实施例中,提取音频样本中的关键帧的语音状态包括:

[0009] 将音频样本输入至音频情绪识别模型的预训练模型。

[0010] 基于所述预训练模型的前端的CNN模块和量化模块获得帧级别的离散值,每一帧对应的离散值表征当前帧的语音状态。

[0011] 采用TF-IDF关键词提取算法确定所述离散值对应的TF-IDF值。

[0012] 根据所述TF-IDF值确定所述音频样本中的关键帧,并确定所述关键帧对应的语音状态。

[0013] 在本公开的一种示例性实施例中,根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量包括:

[0014] 获取音频情绪识别模型的预训练模型的嵌入向量编码信息。

[0015] 根据所述嵌入向量编码信息确定所述音频样本中的关键帧的语音状态对应于的指定个数的嵌入向量。

[0016] 对所述指定个数的嵌入向量进行平均化处理。

[0017] 根据所述平均化处理的结果确定所述情绪标签嵌入向量。

[0018] 在本公开的一种示例性实施例中,根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入包括:

[0019] 将所述语音样本输入至音频情绪识别模型,并获取所述wav2vec2.0模型输出的语音特征图。

[0020] 对所述语音特征图进行均值池化处理,并将所述均值池化处理的结果压缩为情绪嵌入向量。

[0021] 将所述情绪嵌入向量输入至全连接层,以获得情绪分数向量。

[0022] 根据所述情绪嵌入向量、所述情绪分数向量和所述独热向量确定所述损失函数的输入。

[0023] 在本公开的一种示例性实施例中,还包括:

[0024] 根据所述情绪分数向量与所述独热向量之间的交叉熵函数构建第一目标函数。

[0025] 根据所述情绪标签嵌入向量与所述情绪嵌入向量之间的均方误差函数构建第二目标函数。

[0026] 根据所述第一目标函数、所述第二目标函数和预设的权重值确定所述损失函数。

[0027] 在本公开的一种示例性实施例中,还包括:

[0028] 在对所述音频情绪识别模型进行训练的过程中,保持所述CNN模块的指定参数不更新。

[0029] 在本公开的一种示例性实施例中,还包括:

[0030] 设置所述音频情绪识别模型的transformer网络层学习率不超过0.0001,和/或控制所述音频情绪识别模型的训练周期数小于或等于20次。

[0031] 根据本公开实施例的第二方面,提供一种音频情绪识别模型的训练装置,包括:

[0032] 提取模块,设置为提取音频样本中的关键帧的语音状态。

[0033] 确定模块,设置为根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量。

[0034] 所述确定模块,设置为根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入。

[0035] 训练模块,设置为基于所述损失函数对所述音频情绪识别模型进行训练。

[0036] 根据本公开的第三方面,提供一种电子设备,包括:存储器。以及耦合到所述存储器的处理器,所述处理器被配置为基于存储在所述存储器中的指令,执行如上述任意一项所述的方法。

[0037] 根据本公开的第四方面,提供一种计算机可读存储介质,其上存储有程序,该程序被处理器执行时实现如上述任意一项所述的音频情绪识别模型的训练方法。

[0038] 本公开实施例,通过提取音频样本中的关键帧的语音状态,并根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量,进而根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入,最后基于所述损失函数对所述音频情绪识别模型进行训练,使得音频情绪识别模型在训练过程中能够学习到更加泛化的情绪特征,更准确且可靠地进行情绪识别。

[0039] 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和解释性的,并不能限制本公开。

附图说明

[0040] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本公开的实施例,并与说明书一起用于解释本公开的原理。显而易见地,下面描述中的附图仅仅是本公开的一些实施例,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。

[0041] 图1示出了可以应用本发明实施例的音频情绪识别模型的训练方案的示例性系统架构的示意图;

[0042] 图2是本公开示例性实施例中一种音频情绪识别模型的训练方法的流程图;

[0043] 图3是本公开示例性实施例中另一种音频情绪识别模型的训练方法的流程图;

[0044] 图4是本公开示例性实施例中另一种音频情绪识别模型的训练方法的流程图;

[0045] 图5是本公开示例性实施例中另一种音频情绪识别模型的训练方法的流程图;

[0046] 图6是本公开示例性实施例中另一种音频情绪识别模型的训练方法的流程图;

[0047] 图7是本公开示例性实施例中另一种音频情绪识别模型的训练方法的流程图;

[0048] 图8是本公开示例性实施例中另一种音频情绪识别模型的训练方法的流程图;

[0049] 图9是本公开示例性实施例中一种音频情绪识别模型的训练方案中获取情绪标签嵌入向量的示意图;

[0050] 图10是相关技术中一种音频情绪识别模型的训练方案的示意框图;

[0051] 图11是本公开示例性实施例中一种音频情绪识别模型的训练方案的示意框图;

[0052] 图12是本公开示例性实施例中一种音频情绪识别模型的训练装置的方框图;

[0053] 图13是本公开示例性实施例中一种电子设备的方框图。

具体实施方式

[0054] 现在将参考附图更全面地描述示例实施方式。然而,示例实施方式能够以多种形式实施,且不应被理解为限于在此阐述的范例。相反,提供这些实施方式使得本公开将更加全面和完整,并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中,提供许多具体细节从而给出对本公开的实施方式的充分理解。然而,本领域技术人员将意识到,可以实践本公开的技术方案而省略所述特定细节中的一个或更多,或者可以采用其它的方法、组元、装置、步骤等。在其它情况下,不详细示出或描述公知技术方案以避免喧宾夺主而使得本公开的各方面变得模糊。

[0055] 此外,附图仅为本公开的示意性图解,图中相同的附图标记表示相同或类似的部分,因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体,不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体,或在一个或多个硬件模块或集成电路中实现这些功能实体,或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0056] 图1示出了可以应用本发明实施例的音频情绪识别模型的训练方案的示例性系统架构的示意图。

[0057] 如图1所示,系统架构100可以包括终端设备101、102、103中的一种或多种,网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的

介质。网络104可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。

[0058] 应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器的。比如服务器105可以是多个服务器组成的服务器集群等。

[0059] 用户可以使用终端设备101、102、103通过网络104与服务器105交互,以接收或发送消息等。终端设备101、102、103可以是具有显示屏的各种电子设备,包括但不限于智能手机、平板电脑、便携式计算机和台式计算机等等。

[0060] 在一些实施例中,本发明实施例所提供的音频情绪识别模型的训练方法一般由服务器105执行,相应地,音频情绪识别模型的训练装置一般设置于终端设备103(也可以是终端设备101或102)中。在另一些实施例中,某些终端可以具有与服务器设备相似的功能从而执行本方法。

[0061] 下面结合附图对本公开示例实施方式进行详细说明。

[0062] 图2是本公开示例性实施例中音频情绪识别模型的训练方法的流程图。

[0063] 参考图2,音频情绪识别模型的训练方法可以包括:

[0064] 步骤S202,提取音频样本中的关键帧的语音状态。

[0065] 在本公开的实施例中,对音频样本进行切割后得到帧级别的波段,提取上述波段中能够反映语音状态的帧即为关键帧,语音状态可以包括语音中高兴、愤怒、悲伤等情绪状态,但不限于此。

[0066] 步骤S204,根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量。

[0067] 在本公开的实施例中,预设的嵌入向量编码信息为音频情绪识别模型对应的码本或向量字典,基于嵌入向量编码信息将音频样本对应的编码转化为向量,即情绪标签嵌入向量,并基于情绪标签嵌入向量对音频情绪识别模型进行训练,以提高音频情绪识别模型的泛化性和可靠性。

[0068] 步骤S206,根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入。

[0069] 在本公开的实施例中,独热向量是一种独立的、无语义信息的表征,忽略了情绪类别本身的先验知识,若将其作为监督训练标签,则会导致模型在进行情绪特征提取时会忽略一些有用的潜在类别信息,使得模型在不同应用场景上的泛化能力较差,而通过将情绪标签嵌入向量和独热向量确定损失函数的输入,增强了音频情绪识别模型学习语言状态的泛化性和可靠性。

[0070] 步骤S208,基于所述损失函数对所述音频情绪识别模型进行训练。

[0071] 在本公开的实施例中,通过提取音频样本中的关键帧的语音状态,并根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量,进而根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入,最后基于所述损失函数对所述音频情绪识别模型进行训练,使得音频情绪识别模型在训练过程中能够学习到更加泛化的情绪特征,更准确且可靠地进行情绪识别。

[0072] 下面,对音频情绪识别模型的训练方法的各步骤进行详细说明。

[0073] 在本公开的一种示例性实施例中,如图3所示,提取音频样本中的关键帧的语音状

态包括:

[0074] 步骤S302,将音频样本输入至音频情绪识别模型的预训练模型。

[0075] 步骤S304,基于所述预训练模型的前端的CNN模块和量化模块获得帧级别的离散值,每一帧对应的离散值表征当前帧的语音状态。

[0076] 在本公开的实施例中,预训练模型用于对

[0077] 在本公开的实施例中,音频情绪识别模型可以采用开源的wav2vec2.0模型,对于训练集中每一个音频样本,通过音频情绪识别模型获得帧级别的离散值(40ms为一帧),每一帧离散值能够代表当前音频帧的语音状态。

[0078] 步骤S306,采用TF-IDF关键词提取算法确定所述离散值对应的TF-IDF值。

[0079] 在本公开的实施例中,通过TF-IDF关键词提取算法确定所述离散值对应的TF-IDF值,TF-IDF值越高,关键程度越高,重要程度越高,代表这个语音状态情感相关度越高。

[0080] 步骤S308,根据所述TF-IDF值确定所述音频样本中的关键帧,并确定所述关键帧对应的语音状态。

[0081] 在本公开的实施例中,可以将TF-IDF值最大的指定个数的帧级别语音状态确定为关键帧的语音状态。

[0082] 在本公开的一种示例性实施例中,如图4所示,根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量包括:

[0083] 步骤S402,获取音频情绪识别模型的预训练模型的嵌入向量编码信息。

[0084] 步骤S404,根据所述嵌入向量编码信息确定所述音频样本中的关键帧的语音状态对应于的指定个数的嵌入向量。

[0085] 在本公开的实施例中,指定个数可以设置为50、100、150、200等,但不限于此。

[0086] 步骤S406,对所述指定个数的嵌入向量进行平均化处理。

[0087] 在本公开的实施例中,对每一类情绪的指定个数的嵌入向量进行平均化处理,在不损失语言状态的前提下,减少了嵌入向量的数据量。

[0088] 步骤S408,根据所述平均化处理的结果确定所述情绪标签嵌入向量。

[0089] 在本公开的一种示例性实施例中,如图5所示,根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入包括:

[0090] 步骤S502,将所述语音样本输入至音频情绪识别模型,并获取所述wav2vec2.0模型输出的语音特征图。

[0091] 步骤S504,对所述语音特征图进行均值池化处理,并将所述均值池化处理的结果压缩为情绪嵌入向量。

[0092] 在本公开的实施例中,对所述语音特征图进行均值池化处理,并将所述均值池化处理的结果压缩为情绪嵌入向量,有效地降低了情绪嵌入向量的数据量,减小了音频情绪识别模型的训练量。

[0093] 步骤S506,将所述情绪嵌入向量输入至全连接层,以获得情绪分数向量。

[0094] 步骤S508,根据所述情绪嵌入向量、所述情绪分数向量和所述独热向量确定所述损失函数的输入。

[0095] 在本公开的一种示例性实施例中,如图6所示,音频情绪识别模型的训练方法还包括:

[0096] 步骤S602,根据所述情绪分数向量与所述独热向量之间的交叉熵函数构建第一目标函数。

[0097] 步骤S606,根据所述情绪标签嵌入向量与所述情绪嵌入向量之间的均方误差函数构建第二目标函数。

[0098] 步骤S606,根据所述第一目标函数、所述第二目标函数和预设的权重值确定所述损失函数。

[0099] 在本公开的实施例中,所述第一目标函数的权重值和所述第二目标函数的权重值的加和为1,通常将所述第一目标函数的权重值和所述第二目标函数的权重值设置为0.58。

[0100] 在本公开的一种示例性实施例中,如图7所示,音频情绪识别模型的训练方法还包括:

[0101] 步骤S702,在对所述音频情绪识别模型进行训练的过程中,保持所述CNN模块的指定参数不更新。

[0102] 在本公开的实施例中,CNN模块的指定参数是指根据应用场景选定的模型参数,在基于损失函数对音频情绪识别模型进行训练的过程中,通过保持指定参数不更新,来避免音频情绪识别模型过度拟合,进而进一步地提高了音频情绪识别模型的泛化性和可靠性。

[0103] 在本公开的一种示例性实施例中,如图8所示,音频情绪识别模型的训练方法还包括:

[0104] 步骤S802,设置所述音频情绪识别模型的transformer网络层学习率不超过0.0001。

[0105] 步骤S804,控制所述音频情绪识别模型的训练周期数小于或等于20次。

[0106] 在本公开的实施例中,通过设置所述音频情绪识别模型的transformer网络层学习率不超过0.0001,和/或控制所述音频情绪识别模型的训练周期数小于或等于20次,不仅能提高训练速度,而且能够有效地降低音频情绪识别模型的过拟合的概率。

[0107] 如图9所示,本公开示例性实施例中一种音频情绪识别模型的训练方案中获取情绪标签嵌入向量的处理过程900包括:

[0108] 基于开源的wav2vec2.0预训练模型,利用其前端的CNN模块和量化模块,对于训练集中每一个音频,获得模型输出的帧级别的离散值(40ms为一帧),每一帧离散值代表了当前音频帧的语音状态。

[0109] 将每个帧级别语音状态视作一个文本单元,将每一类情绪的所有语音状态序列作为一篇文章,利用TF-IDF关键词提取算法计算TF-IDF值,TF-IDF值越高,关键程度越高,重要程度越高,代表这个语音状态情感相关度越高,统计出每一类情绪音频中TF-IDF值最大的100个帧级别语音状态。

[0110] 并根据此状态从wav2vec2.0的codebook(即嵌入向量字典或者码本)中获取对应的100个嵌入向量。对每一类情绪的100个情绪最相关的嵌入向量进行平均,最终得到该类情绪的标签嵌入向量 $L \in \mathbb{R}^d$,d为特征维度。

[0111] 如图10所示,相关技术中一种音频情绪识别模型的训练方案1000包括:

[0112] 将独热向量作为训练标签,与模型的识别输出交叉熵作为损失函数进行监督训练。独热向量是一种独立的、无语义信息的表征,忽略了情绪类别本身的先验知识,将其作

为监督训练标签,会导致模型在进行情绪特征提取时会忽略一些有用的潜在类别信息,使得模型在不同应用场景上的泛化能力较差。

[0113] 为了解决上述图10所示的方案的技术问题,本公开提出了一种新的训练方案,即如图11所示的音频情绪识别模型的训练方案1100,包括以下关键步骤:

[0114] (1) 输入波形后,wav2vec2.0模型会输出语音特征图 $H \in \mathbb{R}^{T \times d}$,T为语音帧的数量,继而利用均值池化,在时间维度上将语音特征图压缩为情绪嵌入向量 $h \in \mathbb{R}^d$,情绪嵌入向量的表达式如下公式(1)所示:

[0115] $h = \text{meanpooling}(H)$, 公式(1)

[0116] (2) 通过全连接层以及激活函数softmax,得到最终的情绪分数向量 $\hat{y} \in \mathbb{R}^c$,其中c为情绪类别数量,激活函数softmax的表达式如下公式(2)所示:

[0117] $\hat{y} = \text{softmax}(\text{fullyconnect}(h))$, 公式(2)

[0118] (3) 情绪标签嵌入向量作为训练标签的方法所涉及到的训练目标函数有两个,第一个目标函数与现有的监督学习方法一致,模型输出分类结果向量 $\hat{y}$ 与情绪类别的独热向量 $y \in \mathbb{R}^c$ 之间交叉熵的表达式如下公式(3)所示:

[0119] $\text{loss}_{ce} = \text{Cross_Entropy}(y, \hat{y})$, 公式(3)

[0120] 其中,第二个目标函数为wav2vec2.0模型输出的表征h与对应标签嵌入向量L的均方误差函数,其表达式如下公式(4)所示:

[0121] $\text{loss}_{mse} = \text{MSE}(L, h)$, 公式(4)

[0122] (4) 利用情绪标签向量直接引导wav2vec2.0学习并有效提取情绪相关表征,则整体的损失函数的表达式如下公式(5)所示:

[0123] $\text{loss} = \lambda \text{loss}_{ce} + (1 - \lambda) \text{loss}_{mse}$, 公式(5)

[0124] 其中, λ 为可调整超参数,一般取值为0.5。

[0125] 另外,为了保持wav2vec2.0模型在领域外数据集的泛化能力,防止过拟合,在模型训练阶段,会保持前端CNN模型的部分参数不更新,剩余的transformer网络层学习率设置为0.0001以下,控制训练周期数为20个周期。

[0126] 对应于上述方法实施例,本公开还提供一种音频情绪识别模型的训练装置,可以用于执行上述方法实施例。

[0127] 图12是本公开示例性实施例中一种音频情绪识别模型的训练装置的方框图。

[0128] 参考图12,音频情绪识别模型的训练装置1200可以包括:

[0129] 提取模块1202,设置为提取音频样本中的关键帧的语音状态。

[0130] 确定模块1204,设置为根据预设的嵌入向量编码信息确定所述关键帧的语音状态对应的情绪标签嵌入向量。

[0131] 所述确定模块1204,设置为根据所述情绪标签嵌入向量和所述音频样本的独热向量确定损失函数的输入。

[0132] 训练模块1206,设置为基于所述损失函数对所述音频情绪识别模型进行训练。

[0133] 在本公开的一种示例性实施例中,提取模块1202还设置为:

[0134] 将音频样本输入至音频情绪识别模型的预训练模型;

- [0135] 基于所述预训练模型的前端的CNN模块和量化模块获得帧级别的离散值,每一帧对应的离散值表征当前帧的语音状态;
- [0136] 采用TF-IDF关键词提取算法确定所述离散值对应的TF-IDF值;
- [0137] 根据所述TF-IDF值确定所述音频样本中的关键帧,并确定所述关键帧对应的语音状态。
- [0138] 在本公开的一种示例性实施例中,确定模块1204还设置为:
- [0139] 获取音频情绪识别模型的预训练模型的嵌入向量编码信息;
- [0140] 根据所述嵌入向量编码信息确定所述音频样本中的关键帧的语音状态对应于的指定个数的嵌入向量;
- [0141] 对所述指定个数的嵌入向量进行平均化处理;
- [0142] 根据所述平均化处理的结果确定所述情绪标签嵌入向量。
- [0143] 在本公开的一种示例性实施例中,确定模块1204还设置为:
- [0144] 将所述语音样本输入至音频情绪识别模型,并获取所述wav2vec2.0模型输出的语音特征图;
- [0145] 对所述语音特征图进行均值池化处理,并将所述均值池化处理的结果压缩为情绪嵌入向量;
- [0146] 将所述情绪嵌入向量输入至全连接层,以获得情绪分数向量;
- [0147] 根据所述情绪嵌入向量、所述情绪分数向量和所述独热向量确定所述损失函数的输入。
- [0148] 在本公开的一种示例性实施例中,音频情绪识别模型的训练装置1200还设置为:
- [0149] 根据所述情绪分数向量与所述独热向量之间的交叉熵函数构建第一目标函数;
- [0150] 根据所述情绪标签嵌入向量与所述情绪嵌入向量之间的均方误差函数构建第二目标函数;
- [0151] 根据所述第一目标函数、所述第二目标函数和预设的权重值确定所述损失函数。
- [0152] 在本公开的一种示例性实施例中,音频情绪识别模型的训练装置1200还设置为:
- [0153] 在对所述音频情绪识别模型进行训练的过程中,保持所述CNN模块的指定参数不更新。
- [0154] 在本公开的一种示例性实施例中,音频情绪识别模型的训练装置1200还设置为:
- [0155] 设置所述音频情绪识别模型的transformer网络层学习率不超过0.0001,和/或控制所述音频情绪识别模型的训练周期数小于或等于20次。
- [0156] 由于装置1200的各功能已在其对应的方法实施例中予以详细说明,本公开于此不再赘述。
- [0157] 应当注意,尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元,但是这种划分并非强制性的。实际上,根据本公开的实施方式,上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之,上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。
- [0158] 在本公开的示例性实施例中,还提供了一种能够实现上述方法的电子设备。
- [0159] 所属技术领域的技术人员能够理解,本发明的各个方面可以实现为系统、方法或程序产品。因此,本发明的各个方面可以具体实现为以下形式,即:完全的硬件实施方式、完

全的软件实施方式(包括固件、微代码等),或硬件和软件方面结合的实施方式,这里可以统称为“电路”、“模块”或“系统”。

[0160] 下面参照图13来描述根据本发明的这种实施方式的电子设备1300。图13显示的电子设备1300仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0161] 如图13所示,电子设备1300以通用计算设备的形式表现。电子设备1300的组件可以包括但不限于:上述至少一个处理单元1310、上述至少一个存储单元1320、连接不同系统组件(包括存储单元1320和处理单元1310)的总线1330。

[0162] 其中,所述存储单元存储有程序代码,所述程序代码可以被所述处理单元1310执行,使得所述处理单元1310执行本说明书上述“示例性方法”部分中描述的根据本发明各种示例性实施方式的步骤。例如,所述处理单元1310可以执行如本公开实施例所示的方法。

[0163] 存储单元1320可以包括易失性存储单元形式的可读介质,例如随机存取存储单元(RAM) 13201和/或高速缓存存储单元13202,还可以进一步包括只读存储单元(ROM) 13203。

[0164] 存储单元1320还可以包括具有一组(至少一个)程序模块13205的程序/实用工具13204,这样的程序模块13205包括但不限于:操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0165] 总线1330可以为表示几类总线结构中的一种或多种,包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

[0166] 电子设备1300也可以与一个或多个外部设备1340(例如键盘、指向设备、蓝牙设备等)通信,还可与一个或者多个使得用户能与该电子设备1300交互的设备通信,和/或与使得该电子设备1300能与一个或多个其它计算设备进行通信的任何设备(例如路由器、调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口1350进行。并且,电子设备1300还可以通过网络适配器1360与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器1360通过总线1330与电子设备1300的其它模块通信。应当明白,尽管图中未示出,可以结合电子设备1300使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0167] 通过以上的实施方式的描述,本领域的技术人员易于理解,这里描述的示例实施方式可以通过软件实现,也可以通过软件结合必要的硬件的方式来实现。因此,根据本公开实施方式的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM, U盘,移动硬盘等)中或网络上,包括若干指令以使得一台计算设备(可以是个人计算机、服务器、终端装置、或者网络设备等)执行根据本公开实施方式的方法。

[0168] 在本公开的示例性实施例中,还提供了一种计算机可读存储介质,其上存储有能够实现本说明书上述方法的程序产品。在一些可能的实施方式中,本发明的各个方面还可以实现为一种程序产品的形式,其包括程序代码,当所述程序产品在终端设备上运行时,所述程序代码用于使所述终端设备执行本说明书上述“示例性方法”部分中描述的根据本发明各种示例性实施方式的步骤。

[0169] 根据本发明的实施方式的用于实现上述方法的程序产品可以采用便携式紧凑盘

只读存储器 (CD-ROM) 并包括程序代码,并可以在终端设备,例如个人电脑上运行。然而,本发明的程序产品不限于此,在本文件中,可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0170] 所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括:具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器 (RAM)、只读存储器 (ROM)、可擦式可编程只读存储器 (EPROM或闪存)、光纤、便携式紧凑盘只读存储器 (CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0171] 计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了可读程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质,该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0172] 可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于无线、有线、光缆、RF等等,或者上述的任意合适的组合。

[0173] 可以以一种或多种程序设计语言的任意组合来编写用于执行本发明操作的程序代码,所述程序设计语言包括面向对象的程序设计语言—诸如Java、C++等,还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中,远程计算设备可以通过任意种类的网络,包括局域网 (LAN) 或广域网 (WAN),连接到用户计算设备,或者,可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0174] 此外,上述附图仅是根据本发明示例性实施例的方法所包括的处理的示意性说明,而不是限制目的。易于理解,上述附图所示的处理并不表明或限制这些处理的时间顺序。另外,也易于理解,这些处理可以是例如在多个模块中同步或异步执行的。

[0175] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本公开的其它实施方案。本申请旨在涵盖本公开的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本公开的真正范围和构思由权利要求指出。

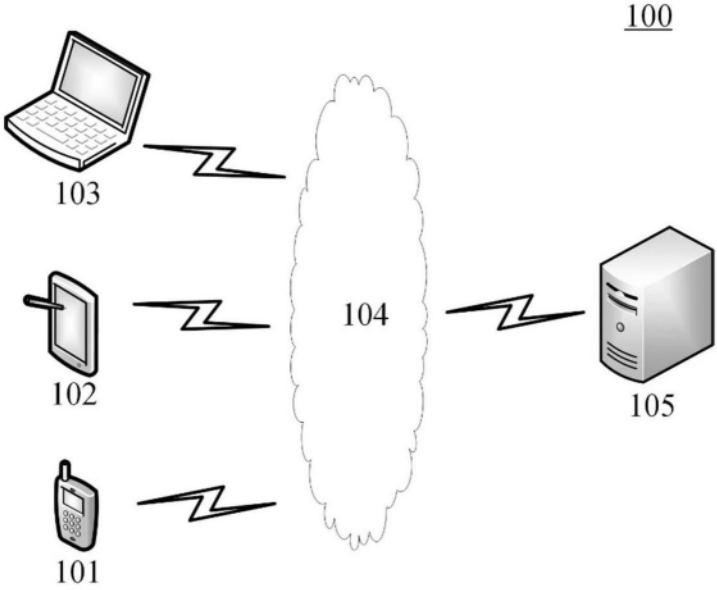


图1

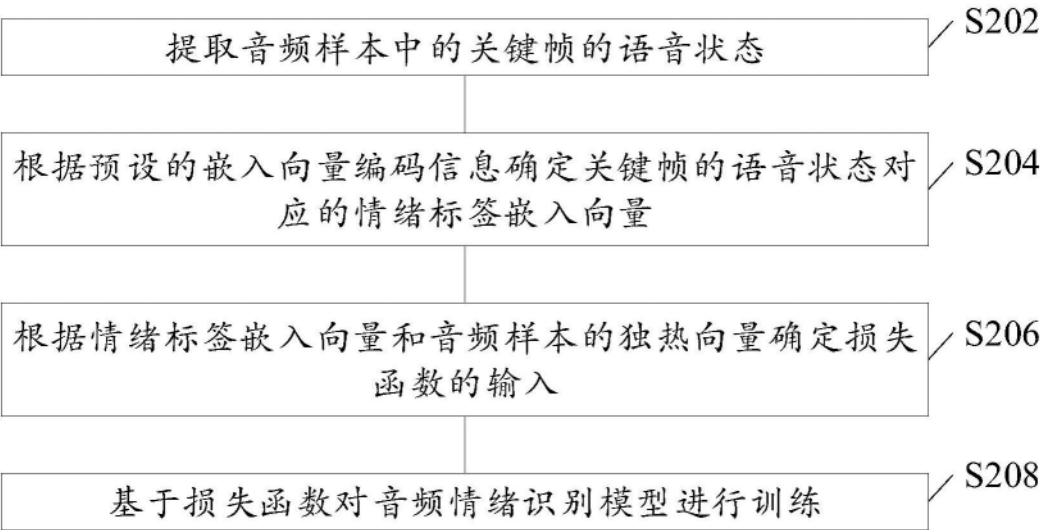


图2



图3

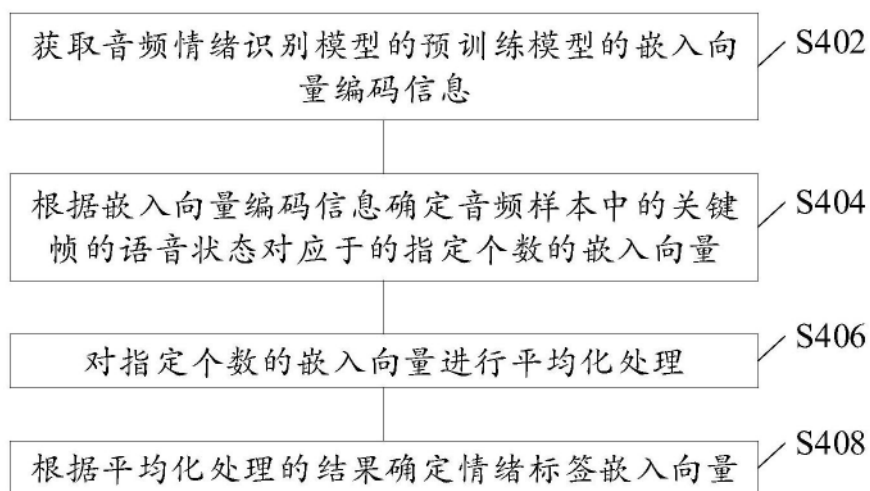


图4

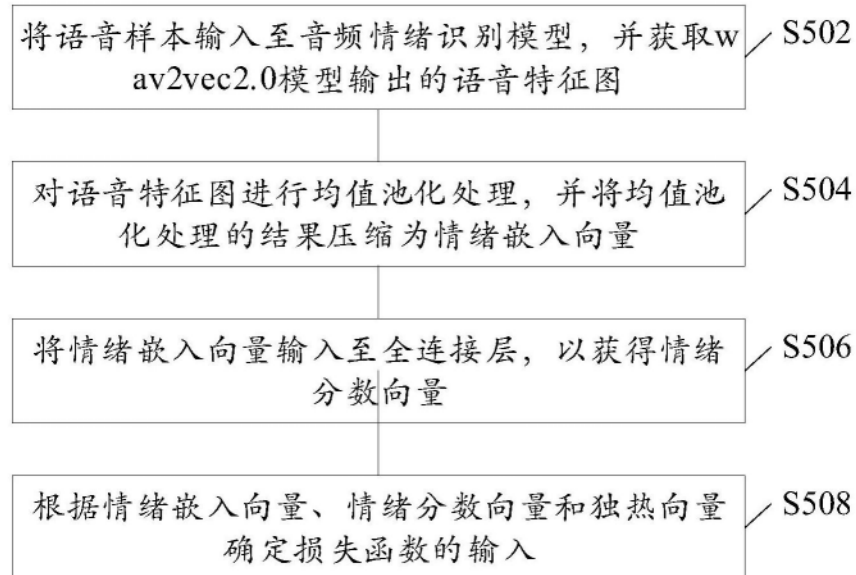


图5

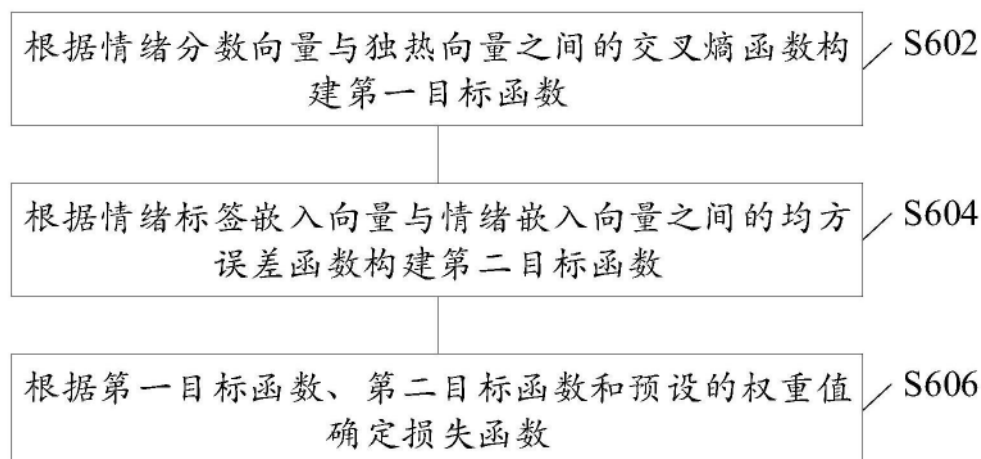


图6

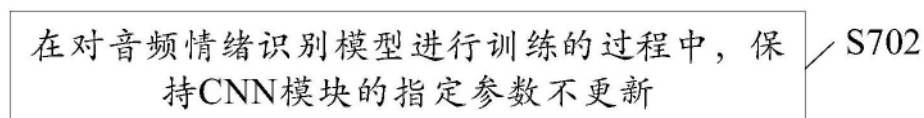


图7

设置音频情绪识别模型的transformer网络层学习率不超过0.0001 S802

控制音频情绪识别模型的训练周期数小于或等于20次 S804

图8

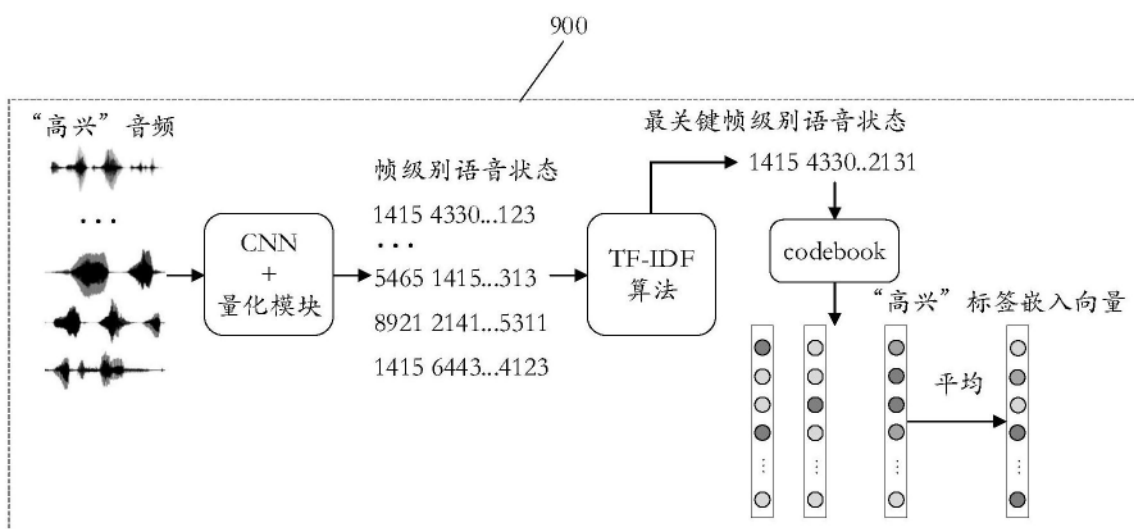


图9

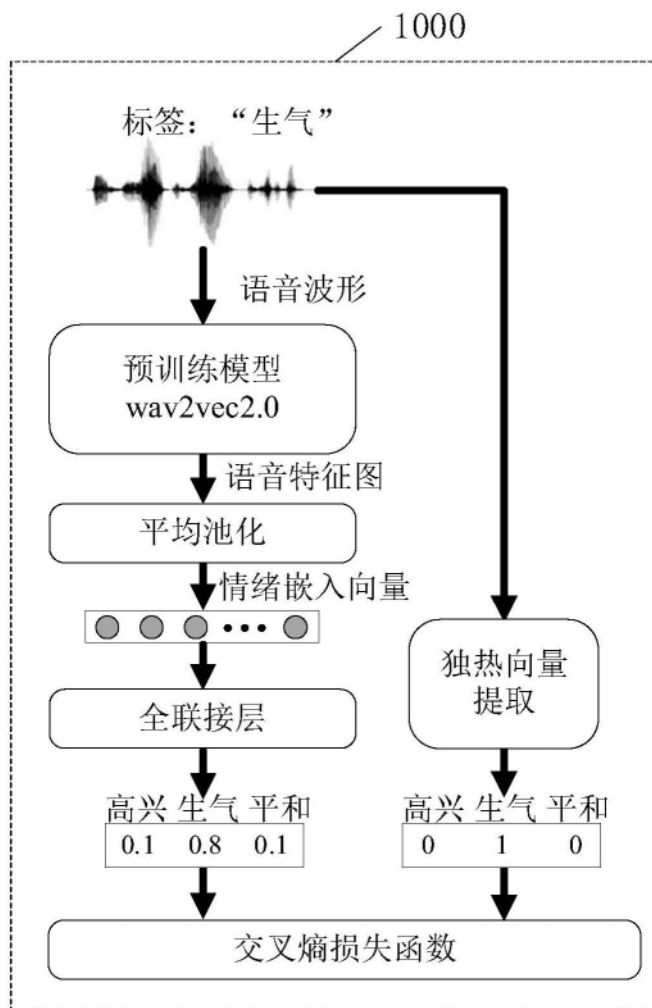


图10

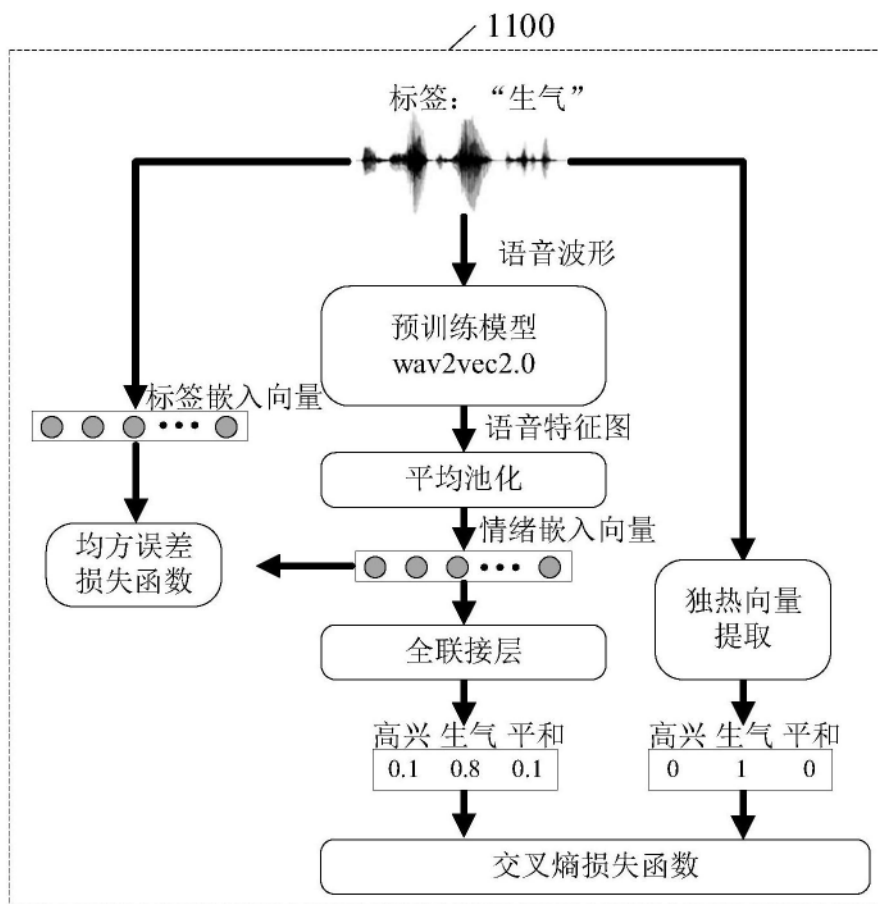


图11

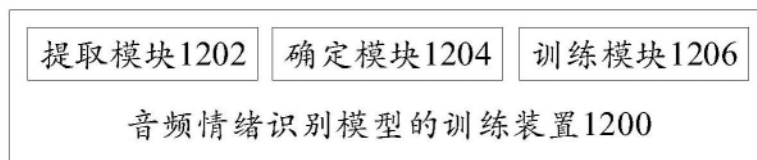


图12

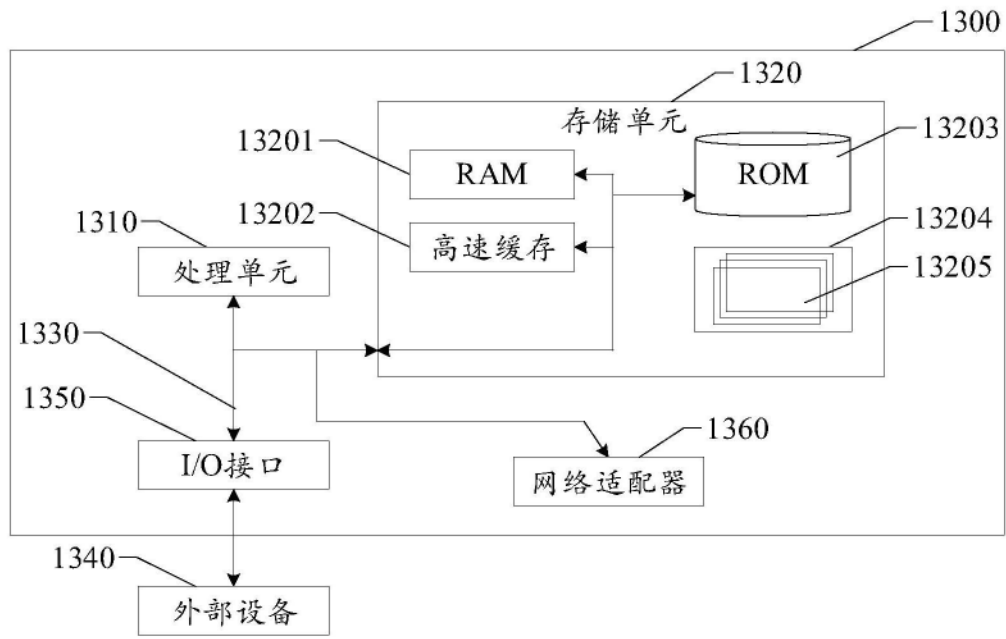


图13