(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116894080 A
(43)申请公布日 2023.10.17

(21)申请号 202310260681.4

(22)申请日 2023.03.13

(71)申请人 京东科技信息技术有限公司
地址 100176 北京市大兴区北京经济技术开发区科创十一街18号院2号楼6层601

(72)发明人 刘瑞雪 袁韶祖 戴爱君 祝天刚 陈蒙

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 王治东

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/35(2019.01)
G06F 40/177(2020.01)
G06F 40/35(2020.01)

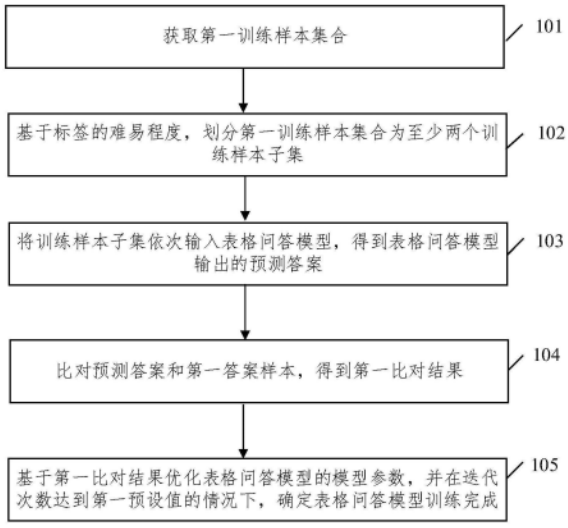
权利要求书2页 说明书11页 附图5页

(54)发明名称

模型训练方法、问题应答方法、装置、设备及介质

(57)摘要

本公开提供一种模型训练方法、问题应答方法、装置、设备及介质，模型训练方法包括：获取第一训练样本集合；基于标签的难易程度，划分第一训练样本集合为至少两个训练样本子集；将训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案；比对预测答案和第一答案样本，得到第一比对结果；基于第一比对结果优化表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定表格问答模型训练完成。本公开用以解决现有技术中表格问答模型预测精准率低的缺陷，实现提高表格问答模型的预测精准率。



CN 116894080 A

1.一种模型训练方法,其特征在于,包括:

获取第一训练样本集合,所述第一训练样本集合包括:表格样本、第一问题样本、所述第一问题样本对应的标签、以及第一答案样本,所述标签用于表征所述第一问题样本的难易程度;

基于所述标签的难易程度,划分所述第一训练样本集合为至少两个训练样本子集;

将所述训练样本子集依次输入表格问答模型,得到所述表格问答模型输出的预测答案;

比对所述预测答案和所述第一答案样本,得到第一比对结果;

基于所述第一比对结果优化所述表格问答模型的模型参数,并在迭代次数达到第一预设值的情况下,确定所述表格问答模型训练完成。

2.根据权利要求1所述的模型训练方法,其特征在于,所述将所述训练样本子集依次输入表格问答模型,得到所述表格问答模型输出的预测答案,包括:

基于所述标签的难易程度,对所述训练样本子集进行由易到难的排序;

将排序后的所述训练样本子集依次输入所述表格问答模型,得到所述表格问答模型输出的所述预测答案。

3.根据权利要求1所述的模型训练方法,其特征在于,所述获取第一训练样本集合之前,还包括:

获取所述表格样本、所述第一问题样本和所述第一答案样本;

将所述表格样本、所述第一问题样本和所述第一答案样本输入难易程度确定模型,得到所述难易程度确定模型输出的所述第一问题样本对应的难易程度;

将所述难易程度确定模型输出的所述难易程度标记至对应的所述第一问题样本上;

其中,所述难易程度确定模型基于表格训练样本、第一问题训练样本和第一答案训练样本训练得到,所述第一问题训练样本标记有表征所述难易程度的标签。

4.根据权利要求3所述的模型训练方法,其特征在于,所述将所述表格样本、所述第一问题样本和所述第一答案样本输入难易程度确定模型,得到所述难易程度确定模型输出的所述第一问题样本对应的难易程度,包括:

将所述表格样本、所述第一问题样本和所述第一答案样本输入难易程度确定模型,确定所述第一问题样本与所述表格样本的相似度,所述第一答案样本在所述表格样本的所占长度比例,所述第一问题样本的语法复杂度,所述第一答案样本的获得方式,以及所述第一答案样本的获得难度,基于所述相似度、所述所占长度比例、所述语法复杂度、所述获得方式以及所述获得难度,得到所述第一问题样本对应的难易程度。

5.根据权利要求1-4任一项所述的模型训练方法,其特征在于,所述获取第一训练样本集合之前,还包括:

获取待使用数据集合,所述待使用数据集合包括:表格、与所述表格对应的描述文本;

确定所述表格与所述描述文本的重合内容,确定所述重合内容为答案;

将所述描述文本和所述答案输入问题生成模型,得到所述问题生成模型输出的所述描述文本对应的问题;

基于预先确定的所述表格与所述描述文本的对应关系,得到所述表格、所述答案和所述问题的数据组;

将所述数据组添加至所述第一训练样本集合；

其中，所述问题生成模型基于第二训练样本集合训练得到，所述第二训练样本集合包括：描述文本样本、第二答案样本和第二问题样本。

6.根据权利要求5所述的模型训练方法，其特征在于，所述问题生成模型的训练过程包括：

获取第二训练样本集合；

将所述第二训练样本集合输入所述问题生成模型，得到所述问题生成模组输出的预测问题；

比对所述预测问题和所述第二问题样本，得到第二比对结果；

基于所述第二比对结果优化所述问题生成模型的模型参数，并在迭代次数达到第二预设值时，确定所述问题生成模型训练完成。

7.一种问题应答方法，其特征在于，包括：

获取目标表格与目标问题；

将所述目标表格和所述目标问题输入表格问答模型，得到所述表格问答模型输出的与所述目标问题对应的目标答案，所述表格问答模型基于权利要求1-6任一项的所述模型训练方法得到。

8.一种模型训练装置，其特征在于，包括：

获取模块，用于获取第一训练样本集合，所述第一训练样本集合包括：表格样本、第一问题样本、所述第一问题样本对应的标签、以及第一答案样本，所述标签用于表征所述第一问题样本的难易程度；

划分模块，用于基于所述标签的难易程度，划分所述第一训练样本集合为至少两个训练样本子集；

预测模块，用于将所述训练样本子集依次输入表格问答模型，得到所述表格问答模型输出的预测答案；

比对模块，用于比对所述预测答案和所述第一答案样本，得到第一比对结果；

训练模块，用于基于所述第一比对结果优化所述表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定所述表格问答模型训练完成。

9.一种电子设备，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程序时实现如权利要求1至6任一项所述的模型训练方法，或，权利要求7所述的问题应答方法。

10.一种非暂态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6任一项所述的模型训练方法，或，权利要求7所述的问题应答方法。

模型训练方法、问题应答方法、装置、设备及介质

技术领域

[0001] 本公开涉及课程学习技术领域，尤其涉及一种模型训练方法、问题应答方法、装置、设备及介质。

背景技术

[0002] 表格问答是一种围绕表格进行自动回复的技术，即，根据用户问题和相关表格，自动进行内容查询从而生成回复答案。

[0003] 在现有技术中，在训练表格问答模型时，随机将所有的文本数据和用户问题放入模型中进行训练，忽略了表格问答和文本问答的差别，使得最终得到的表格问答模型的预测精准率较低。

发明内容

[0004] 本公开提供一种模型训练方法、问题应答方法、装置、设备及介质，用以解决现有技术中表格问答模型预测精准率低的缺陷，实现提高表格问答模型的预测精准率。

[0005] 本公开提供一种模型训练方法，包括：

[0006] 获取第一训练样本集合，所述第一训练样本集合包括：表格样本、第一问题样本、所述第一问题样本对应的标签、以及第一答案样本，所述标签用于表征所述第一问题样本的难易程度；

[0007] 基于所述标签的难易程度，划分所述第一训练样本集合为至少两个训练样本子集；

[0008] 将所述训练样本子集依次输入表格问答模型，得到所述表格问答模型输出的预测答案；

[0009] 比对所述预测答案和所述第一答案样本，得到第一比对结果；

[0010] 基于所述第一比对结果优化所述表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定所述表格问答模型训练完成。

[0011] 根据本公开提供的一种模型训练方法，所述将所述训练样本子集依次输入表格问答模型，得到所述表格问答模型输出的预测答案，包括：

[0012] 基于所述标签的难易程度，对所述训练样本子集进行由易到难的排序；

[0013] 将排序后的所述训练样本子集依次输入所述表格问答模型，得到所述表格问答模型输出的所述预测答案。

[0014] 根据本公开提供的一种模型训练方法，所述获取第一训练样本集合之前，还包括：

[0015] 获取所述表格样本、所述第一问题样本和所述第一答案样本；

[0016] 将所述表格样本、所述第一问题样本和所述第一答案样本输入难易程度确定模型，得到所述难易程度确定模型输出的所述第一问题样本对应的难易程度；

[0017] 将所述难易程度确定模型输出的所述难易程度标记至对应的所述第一问题样本上；

[0018] 其中，所述难易程度确定模型基于表格训练样本、第一问题训练样本和第一答案训练样本训练得到，所述第一问题训练样本标记有表征所述难易程度的标签。

[0019] 根据本公开提供的一种模型训练方法，所述将所述表格样本、所述第一问题样本和所述第一答案样本输入难易程度确定模型，得到所述难易程度确定模型输出的所述第一问题样本对应的难易程度，包括：

[0020] 将所述表格样本、所述第一问题样本和所述第一答案样本输入难易程度确定模型，确定所述第一问题样本与所述表格样本的相似度，所述第一答案样本在所述表格样本的所占长度比例，所述第一问题样本的语法复杂度，所述第一答案样本的获得方式，以及所述第一答案样本的获得难度，基于所述相似度、所述所占长度比例、所述语法复杂度、所述获得方式以及所述获得难度，得到所述第一问题样本对应的难易程度。

[0021] 根据本公开提供的一种模型训练方法，所述获取第一训练样本集合之前，还包括：

[0022] 获取待使用数据集合，所述待使用数据集合包括：表格、与所述表格对应的描述文本；

[0023] 确定所述表格与所述描述文本的重合内容，确定所述重合内容为答案；

[0024] 将所述描述文本和所述答案输入问题生成模型，得到所述问题生成模型输出的所述描述文本对应的问题；

[0025] 基于预先确定的所述表格与所述描述文本的对应关系，得到所述表格、所述答案和所述问题的数据组；

[0026] 将所述数据组添加至所述第一训练样本集合；

[0027] 其中，所述问题生成模型基于第二训练样本集合训练得到，所述第二训练样本集合包括：描述文本样本、第二答案样本和第二问题样本。

[0028] 根据本公开提供的一种模型训练方法，所述问题生成模型的训练过程包括：

[0029] 获取第二训练样本集合；

[0030] 将所述第二训练样本集合输入所述问题生成模型，得到所述问题生成模组输出的预测问题；

[0031] 比对所述预测问题和所述第二问题样本，得到第二比对结果；

[0032] 基于所述第二比对结果优化所述问题生成模型的模型参数，并在迭代次数达到第二预设值时，确定所述问题生成模型训练完成。

[0033] 本公开还提供一种问题应答方法，包括：

[0034] 获取目标表格与目标问题；

[0035] 将所述目标表格和所述目标问题输入表格问答模型，得到所述表格问答模型输出的与所述目标问题对应的目标答案，所述表格问答模型基于上述任一模型训练方法得到。

[0036] 本公开还提供一种模型训练装置，包括：

[0037] 获取模块，用于获取第一训练样本集合，所述第一训练样本集合包括：表格样本、第一问题样本、所述第一问题样本对应的标签、以及第一答案样本，所述标签用于表征所述第一问题样本的难易程度；

[0038] 划分模块，用于基于所述标签的难易程度，划分所述第一训练样本集合为至少两个训练样本子集；

[0039] 预测模块，用于将所述训练样本子集依次输入表格问答模型，得到所述表格问答

模型输出的预测答案；

[0040] 比对模块，用于比对所述预测答案和所述第一答案样本，得到第一比对结果；

[0041] 训练模块，用于基于所述第一比对结果优化所述表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定所述表格问答模型训练完成。

[0042] 本公开还提供一种电子设备，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程序时实现如上述任一种所述的模型训练方法，或，问题应答方法。

[0043] 本公开还提供一种非暂态计算机可读存储介质，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上述任一种所述的模型训练方法，或，问题应答方法。

[0044] 本公开提供的模型训练方法、问题应答方法、装置、设备及介质，通过获取第一训练样本集合，第一训练样本集合包括：表格样本、第一问题样本、第一问题样本对应的标签、以及第一答案样本，标签用于表征第一问题样本的难易程度，可见，本公开的模型训练样本为符合表格问答的表格样本，为提高表格问答模型的精准率提供了有效的数据基础；进而，基于标签的难易程度，划分第一训练样本集合为至少两个训练样本子集；将训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案；比对预测答案和第一答案样本，得到第一比对结果；基于第一比对结果优化表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定表格问答模型训练完成，可见，本公开对表格问答中的用户问题进行难易程度的区分，对模型进行分类学习，增强了模型对用户问题涉及的不同类型问题和表格的语义学习表示，提高了模型的训练效率和模型的预测精准率。

附图说明

[0045] 为了更清楚地说明本公开或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图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。

[0046] 图1是本公开提供的模型训练方法的流程示意图；

[0047] 图2是本公开提供的问题获取方式的示意图；

[0048] 图3是本公开提供的问题生成模型的训练过程示意图；

[0049] 图4是本公开提供的确定问题难易程度的示意图；

[0050] 图5是本公开提供的难易程度确定模型的训练过程示意图；

[0051] 图6是本公开提供的模型训练装置的结构示意图；

[0052] 图7是本公开提供的问题应答装置的结构示意图；

[0053] 图8是本公开提供的电子设备的结构示意图。

具体实施方式

[0054] 为使本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例中的附图，对本公开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公开实施例一部分实施例，而不是全部的实施例。基于本公开实施例中的实施例，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，都属于本公开实施

例保护的范围。

[0055] 下面结合图1-图5描述本公开实施例的模型训练方法。

[0056] 近年来,随着深度学习技术的发展,大多使用深度学习方法解决表格问答(Table Question Answering,简称TableQA)的问题,通过借助模型训练技巧在有限的数据中提高模型的应答精准率。

[0057] 各类模型的训练通常依赖大规模的标注数据,进行监督式学习来实现。但是在实际应用场景中,通常可以很容易获取到大规模的表格和表格对应的文本描述,但是很难获取到表格-问句-答案的标注数据。因此训练数据的缺乏将限制各种深度学习模型在TableQA问题的训练。

[0058] 不仅如此,在针对TableQA问题的模型训练过程中,多数仅随机的将所有数据进行洗牌后进行模型训练。这种训练方法通常对TableQA的问题不做任何的处理,一视同仁不加群分的将所有问题和文本数据随机放入模型中训练,这忽略了表格问答的问题跟文本问答存在的差别:即针对表格问答的问题是存在难度差别的。

[0059] 例如,根据表格中存在的明显信息进行回答的问题通常很简答(例如,“明星A在电影B中的导演是谁”,可以通过电影B的表格介绍很容易获取答案),但是根据多行表格信息进行推理的则是非常难(例如,“明星A在2015年以后一共拍了多少部电影”,需要根据表格中每个电影的时间进行推断和相加获取答案)。这种不加区分的进行模型训练的方式不利于模型通过逐步学习的方式逐步增强表格问答的能力。

[0060] 可见,在现有技术中针对表格问答,存在众多问题,例如,缺乏大规模的表格-问题-答案的标注数据(训练样本数据),并且,构建该形式的标注数据需要耗费大量的人力成本;又例如,在模型训练过程中,将训练数据随机洗牌放入模型,忽略了表格问答所涉及的问题难度高低不一的情况;等。

[0061] 为了解决上述各个问题,本公开实施例提供了一种模型训练方法,该方法可以应用在智能终端,例如,手机、电脑、平板等,也可以应用在服务器中。下面,以该方法应用在服务器为例进行说明,但需要说明的是仅为举例说明,并不用于对本公开的保护范围进行限定。本公开实施例中的一些其他说明,也是举例说明,并不用于对本公开的保护范围进行限定,之后便不再一一说明。该方法的具体实现如图1所示:

[0062] 步骤101,获取第一训练样本集合。

[0063] 其中,第一训练样本集合包括:表格样本、第一问题样本、第一问题样本对应的标签、以及第一答案样本,标签用于表征第一问题样本的难易程度。

[0064] 具体的,第一训练样本集合为表格-问题-答案三元组形式的样本数据,由于该类数据很难获取到,因此,预先得到表格-问题-答案三元组形式的样本数据。

[0065] 一个具体实施例中,获得表格-问题-答案三元组形式的样本数据的具体实现为:

[0066] 获取待使用数据集合,待使用数据集合包括:表格、与表格对应的描述文本;确定表格与描述文本的重合内容,确定重合内容为答案;将描述文本和答案输入问题生成模型,得到问题生成模型输出的描述文本对应的问题;基于预先确定的表格与描述文本的对应关系,得到表格、答案和问题的数据组;将数据组添加至第一训练样本集合;其中,问题生成模型基于第二训练样本集合训练得到,第二训练样本集合包括:描述文本样本、第二答案样本和第二问题样本。

[0067] 具体的，初始化第一训练样本集合，第一训练样本集合可以为空集，也可以为预先存储有表格-问题-答案三元组形式的样本数据的集合。

[0068] 具体的，由于表格-描述文本的数据容易大量得到，因此，本公开预先通过网络爬虫的方式获取大量的表格以及与表格对应的描述文本；然后，通过语义匹配的方式，匹配表格与描述文本的重合内容，将重合的内容确定为答案，其中，由于表格和描述问题会存在多个重合内容，因此会得到多个答案；进而，将描述文本和答案输入到预先训练好的问题生成模型中，得到问题生成模型输出的问题，其中，由于存在多个答案，因此会得到多个问题；最后，基于表格与描述文本的对应关系，得到表格-问题-答案的三元组，并将其存储在第一训练样本集合中。

[0069] 下面，通过图2对得到表格的问题的获取方式进行示意表示。

[0070] 例如，图2中的表格见表1，描述文本为“2004年雅典运动会男子110米栏，以12秒91的成绩追平了由选手乙创造的实际记录，夺得冠军”，重合内容为“冠军”，生的为问题为“甲在2004年雅典运动会中取得了什么样的成绩”。

[0071]

日期	比赛	成绩	名次
2004年	世界室内田径锦标赛	7.45秒	冠军
2004年	国际田径大奖赛	13.06秒	冠军
2004年	国际田径大奖赛	13.11秒	冠军
2004年	运动会	12.91秒	冠军
2005年	第十届世界田径锦标赛	13.08秒	冠军
2005年	黄金田径联赛	13.2秒	冠军

[0072] 表1表格

[0073] 一个具体实施例中，问题生成模型的训练过程包括：

[0074] 获取第二训练样本集合；将第二训练样本集合输入问题生成模型，得到问题生成模组输出的预测问题；比对预测问题和第二问题样本，得到第二比对结果；基于第二比对结果优化问题生成模型，并在迭代次数达到第二预设值时，确定问题生成模型训练完成。

[0075] 下面，通过图3对问题生成模型的训练过程进行示意表示。

[0076] 例如，在图3中的描述文本为“在2002年NBA选秀中，丙以状元秀身份被NBA的休斯顿火箭队选中，2003-09年连续6个赛季(生涯共8次)入选NBA全明星”，答案为“火箭队”，生成的问题为“请问丙在2002年是被NBA哪支球队以状元秀身份选中的呢”。

[0077] 本公开通过问题生成模型，解决了基于深度学习的表格问答模型需要大量的表格-问题-答案三元组形式的训练样本数据数量较少的问题，并且，得到三元组的过程无需人工标注，节省了人工成本。

[0078] 一个具体实施例中，在得到表格-问题-答案三元组之后，确定各个问题对应的难易程度，具体实现如下所示：

[0079] 获取表格样本、第一问题样本和第一答案样本；将表格样本、第一问题样本和第一答案样本输入难易程度确定模型，得到难易程度确定模型输出的第一问题样本对应的难易程度；将难易程度确定模型输出的难易程度标记至对应的第一问题样本上；其中，难易程度确定模型基于表格训练样本、第一问题训练样本、第一答案训练样本和难易程度标签样本训练得到。

[0080] 其中,难易程度包括:容易、中等和困难等几个等级。

[0081] 具体的,将各个表格-问题-答案三元组输入难易程度确定模型,得到难易程度确定模型输出的问题对应的难易程度,得到难易程度对应的具体值之后,将其标记在问题上,并将表格-问题(携带有难易程度标签)-答案三元组存储在第一训练样本集合中。

[0082] 例如,容易对应标签1,中等对应标签2,困难对应标签3。

[0083] 一个具体实施例中,由于表格问答的问题具有难度高低之分,为了区分表格问答中问题的难易程度,通过融合了多种信息的难易程度确定模型来确定第一问题样本的难易程度,具体确定方式为:

[0084] 将表格样本、第一问题样本和第一答案样本输入难易程度确定模型,确定第一问题样本与表格样本的相似度,第一答案样本在表格样本的所占长度比例,第一问题样本的语法复杂度,第一答案样本的获得方式,以及第一答案样本的获得难度,基于相似度、所占长度比例、语法复杂度、获得方式以及获得难度,得到第一问题样本对应的难易程度。

[0085] 具体的,通过词汇特征确定第一问题样本与表格样本的相似度,比如,确定第一问题样本与表格样本相同的词汇的个数,基于预先确定的个数与相似度的匹配关系,得到第一问题样本与表格样本的相似度;统计第一答案样本的第一文本长度以及表格样本中所有文本的第二文本长度,计算第一文本长度与第二文本长度的比值,得到第一答案样本在表格样本的所占长度比例;分析第一问题样本的语法结构,基于语法结构,确定第一问题样本的语法复杂度;确定第一答案样本的获取方式是能够直接从表格中得到的方式,还是通过推理从表格中得到的方式;判断第一答案样本是通过直接从表格中得到的方式得到,还是通过推理从表格中得到的方式得到,若是,通过直接从表格中得到的方式得到,得到与之对应的获得难度,若是,通过推理从表格中得到的方式得到,基于得到第一答案样本所执行的运算得到与之对应的获得难度。

[0086] 具体的,相似度、所占长度比例、语法复杂度、获得方式以及获得难度都会存在一个对应的分值,即,通过难易度确定模型得到与相似度、所占长度比例、语法复杂度、获得方式以及获得难度对应的序列,基于得到序列确定第一问题样本对应的难易程度。

[0087] 下面,通过图4对难易程度确定模型确定第一问题样本的难易程度进行示意说明。

[0088] 例如,在图4中的表格为表1;问题包括6个,分别为“Q1:甲在2004年雅典运动会中取得了什么样的成绩;Q2:甲在2004年一共获得多少冠军;Q3:甲在2005年最快成绩是多少;Q4:甲在第十届世界田径锦标赛比运动会慢多少;Q4:甲在那一年的运动会得到了冠军;Q5:甲在2004年和2005年获得了多少冠军”;得到的难易程度结果为“Q1和Q5为容易,Q2和Q6为中等,Q3和Q4为困难”。

[0089] 一个具体实施例中,难易程度确定模型的训练过程包括:

[0090] 获取表格训练样本、第一问题训练样本和第一答案训练样本,其中,第一问题训练样本标记有表征难易程度的标签;将表格训练样本、第一问题训练样本和第一答案训练样本输入难易程度确定模型,得到难易程度模型输出的第一问题训练样本的预测难易值;比对预测难易值和表征难易程度的标签对应的值;得到第三比对结果;基于第二比对结果优化难易程度确定模型的模型参数,并在迭代次数达到第三预设值时,确定难易程度确定模型训练完成。

[0091] 具体的,将表格训练样本、第一问题训练样本和第一答案训练样本输入难易程度

确定模型，确定第一问题训练样本与表格训练样本的相似度样本，第一答案训练样本在表格训练样本的所占长度比例样本，第一问题训练样本的语法复杂度样本，第一答案训练样本的获得方式样本，以及第一答案训练样本的获得难度样本，基于相似度样本、所占长度比例样本、语法复杂度样本、获得方式样本以及获得难度样本，得到第一问题训练样本对应的难易程度。

[0092] 下面，通过图5对难易程度确定模型的训练过程进行示意表示。

[0093] 例如，在图5中以难易程度确定模型为基于Seq2Seq为例进行说明，当然此处仅为举例说明，并不对难易程度确定模型的结构进行限定。

[0094] 将表格和问题输入难易程度确定模型，通过编码器和解码器的对数据的处理，得到问题的难易程度，另外还得到该问题的答案。

[0095] 本公开通过难易程度确定模型得到第一问题样本的难易程度，为提高表格问答模型的预测精准度提供了有效的数据基础。

[0096] 步骤102，基于标签的难易程度，划分第一训练样本集合为至少两个训练样本子集。

[0097] 例如，难易程度分为容易和困难时，将第一训练样本集合划分为2个训练样本子集，一个训练样本子集中存储有难易程度被标记为容易的数据，一个训练样本子集存储有难易程度被标记为困难的数据。

[0098] 又例如，难易程度分为容易、中等和困难时，将第一训练样本集合划分为3个训练样本子集，一个训练样本子集中存储有难易程度被标记为容易的数据，一个训练样本子集中存储有难易程度被标记为中等的数据，一个训练样本子集存储有难易程度被标记为困难的数据。

[0099] 具体的划分基础依赖于预先确定的难易程度的等级划分。

[0100] 本公开通过对第一训练样本集合进行难易程度的划分，为后续表格问答模型的训练提供了有效的数据基础。

[0101] 步骤103，将训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案。

[0102] 一个具体实施例中，基于标签的难易程度，对训练样本子集进行由易到难的排序；将排序后的训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案。

[0103] 其中，表格问答模型可以为Seq2Seq生成式模型，包括编码器和解码器的结构，也可以为卷积神经网络的结构，等，对于表格问答模型的结构本公开并不做限制。

[0104] 由于在模型训练过充不区分问题的难度而直接进行训练，极大会影响模型的学习效率，从上文的示例可以看出，针对难度低的问题“甲在2004年雅典运动会中获得什么样的成绩”，模型只要理解2004/运动会/成绩等关键词，便可以快速定位到“冠军”的答案，然而，针对难度高的问题“甲在2005年最快成绩是多少”，模型不仅需要理解最快，还需要通过自动对比不同行的答案，选择出“最快”的成绩所对应的表格。而此类高难度问题将考验模型的语义理解能力和推理运算能力。

[0105] 另外，基于Seq2Seq学习框架的传统方法很难从嘈杂和难度参差不齐的训练数据中生成自然的答案。课程学习能够从中学习低质量和简单问题答案对。然后，逐渐从培训实例中生成正确自然的答案。

[0106] 因此，在训练过程中，本公开将标记过的标签按照先易后难的顺序，依次放入模型中进行训练，从而模仿人类学习过程，将提高模型针对不同问题的回答准确率。

[0107] 步骤104，比对预测答案和第一答案样本，得到第一比对结果。

[0108] 具体的，比对预测答案和第一答案样本的一致性，将该一致性作为第一比对结果。

[0109] 步骤105，基于第一比对结果优化表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定表格问答模型训练完成。

[0110] 一个具体实施例中，表格问答模型训练完成之后，并可以进行模型的应用(本公开实施例公开的问题应答方法)，具体实现为：获取目标表格与目标问题；将目标表格和目标问题输入表格问答模型，得到表格问答模型输出的与目标问题对应的目标答案。

[0111] 基于本公开训练的表格问答模型能够准确、快速的输出与目标问题对应的目标答案，另外，还可以在输出目标答案的同时还可以输出该目标问题的难易程度，以在课程学习领域，使用户知晓该目标问题是否符合自己的学习规划。

[0112] 本公开提供的模型训练方法，通过获取第一训练样本集合，第一训练样本集合包括：表格样本、第一问题样本、第一问题样本对应的标签、以及第一答案样本，标签用于表征第一问题样本的难易程度，可见，本公开的模型训练样本为符合表格问答的表格样本，为提高表格问答模型的精准率提供了有效的数据基础；进而，基于标签的难易程度，划分第一训练样本集合为至少两个训练样本子集；将训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案；比对预测答案和第一答案样本，得到第一比对结果；基于第一比对结果优化表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定表格问答模型训练完成，可见，本公开对表格问答中的用户问题进行难易程度的区分，对模型进行分类学习，增强了模型对用户问题涉及的不同类型问题和表格的语义学习表示，提高了模型的训练效率和模型的预测精准率。

[0113] 下面对本公开实施例提供的模型训练装置进行描述，下文描述的模型训练装置与上文描述的模型训练方法可相互对应参照，重复之处不再赘述，如图6所示，该模型训练装置包括：

[0114] 获取模块601，用于获取第一训练样本集合，第一训练样本集合包括：表格样本、第一问题样本、第一问题样本对应的标签、以及第一答案样本，标签用于表征第一问题样本的难易程度；

[0115] 划分模块602，用于基于标签的难易程度，划分第一训练样本集合为至少两个训练样本子集；

[0116] 预测模块603，用于将训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案；

[0117] 比对模块604，用于比对预测答案和第一答案样本，得到第一比对结果；

[0118] 训练模块605，用于基于第一比对结果优化表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定表格问答模型训练完成。

[0119] 一个具体实施例中，预测模块603，具体用于基于标签的难易程度，对训练样本子集进行由易到难的排序；将排序后的训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案。

[0120] 一个具体实施例中，获取模块601，还用于获取表格样本、第一问题样本和第一答

案样本;将表格样本、第一问题样本和第一答案样本输入难易程度确定模型,得到难易程度确定模型输出的第一问题样本对应的难易程度;将难易程度确定模型输出的难易程度标记至对应的第一问题样本上;其中,难易程度确定模型基于表格训练样本、第一问题训练样本和第一答案训练样本训练得到,第一问题训练样本标记有表征难易程度的标签。

[0121] 一个具体实施例中,获取模块601,还用于将表格样本、第一问题样本和第一答案样本输入难易程度确定模型,确定第一问题样本与表格样本的相似度,第一答案样本在表格样本的所占长度比例,第一问题样本的语法复杂度,第一答案样本的获得方式,以及第一答案样本的获得难度,基于相似度、所占长度比例、语法复杂度、获得方式以及获得难度,得到第一问题样本对应的难易程度。

[0122] 一个具体实施例中,获取模块601,还用于获取待使用数据集合,待使用数据集合包括:表格、与表格对应的描述文本;确定表格与描述文本的重合内容,确定重合内容为答案;将描述文本和答案输入问题生成模型,得到问题生成模型输出的描述文本对应的问题;基于预先确定的表格与描述文本的对应关系,得到表格、答案和问题的数据组;将数据组添加至第一训练样本集合;其中,问题生成模型基于第二训练样本集合训练得到,第二训练样本集合包括:描述文本样本、第二答案样本和第二问题样本。

[0123] 一个具体实施例中,训练模块605,还用于获取第二训练样本集合;将第二训练样本集合输入问题生成模型,得到问题生成模组输出的预测问题;比对预测问题和第二问题样本,得到第二比对结果;基于第二比对结果优化问题生成模型的模型参数,并在迭代次数达到第二预设值时,确定问题生成模型训练完成。

[0124] 下面对本公开实施例提供的问题应答装置进行描述,下文描述的问题应答装置与上文描述的问题应答方法可相互对应参照,重复之处不再赘述,如图7所示,该问题应答装置包括:

[0125] 目标获取模块701,用于获取目标表格与目标问题;

[0126] 目标预测模块702,用于将目标表格和目标问题输入表格问答模型,得到表格问答模型输出的与目标问题对应的目标答案。

[0127] 图8示例了一种电子设备的实体结构示意图,如图8所示,该电子设备可以包括:处理器(processor)801、通信接口(Communications Interface)802、存储器(memory)803和通信总线804,其中,处理器801,通信接口802,存储器803通过通信总线804完成相互间的通信。处理器801可以调用存储器803中的逻辑指令,以执行模型训练方法,该方法包括:获取第一训练样本集合,第一训练样本集合包括:表格样本、第一问题样本、第一问题样本对应的标签、以及第一答案样本,标签用于表征第一问题样本的难易程度;基于标签的难易程度,划分第一训练样本集合为至少两个训练样本子集;将训练样本子集依次输入表格问答模型,得到表格问答模型输出的预测答案;比对预测答案和第一答案样本,得到第一比对结果;基于第一比对结果优化表格问答模型的模型参数,并在迭代次数达到第一预设值的情况下,确定表格问答模型训练完成,或,执行问题应答方法,该方法包括:获取目标表格与目标问题;将目标表格和目标问题输入表格问答模型,得到表格问答模型输出的与目标问题对应的目标答案。

[0128] 此外,上述的存储器803中的逻辑指令可以通过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

公开实施例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，或者网络设备等)执行本公开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、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0129] 另一方面，本公开还提供一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，计算机能够执行上述各方法所提供的模型训练方法，该方法包括：获取第一训练样本集合，第一训练样本集合包括：表格样本、第一问题样本、第一问题样本对应的标签、以及第一答案样本，标签用于表征第一问题样本的难易程度；基于标签的难易程度，划分第一训练样本集合为至少两个训练样本子集；将训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案；比对预测答案和第一答案样本，得到第一比对结果；基于第一比对结果优化表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定表格问答模型训练完成，或，执行问题应答方法，该方法包括：获取目标表格与目标问题；将目标表格和目标问题输入表格问答模型，得到表格问答模型输出的与目标问题对应的目标答案。

[0130] 又一方面，本公开还提供一种非暂态计算机可读存储介质，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以执行上述各提供的模型训练方法，该方法包括：获取第一训练样本集合，第一训练样本集合包括：表格样本、第一问题样本、第一问题样本对应的标签、以及第一答案样本，标签用于表征第一问题样本的难易程度；基于标签的难易程度，划分第一训练样本集合为至少两个训练样本子集；将训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案；比对预测答案和第一答案样本，得到第一比对结果；基于第一比对结果优化表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定表格问答模型训练完成，或，执行问题应答方法，该方法包括：获取目标表格与目标问题；将目标表格和目标问题输入表格问答模型，得到表格问答模型输出的与目标问题对应的目标答案。

[0131] 以上所描述的装置实施例仅仅是示意性的，其中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模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0132] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，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，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0133] 最后应说明的是：以上实施例仅用以说明本公开的技术方案，而非对其限制；尽管

参照前述实施例对本公开进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本公开各实施例技术方案的精神和范围。

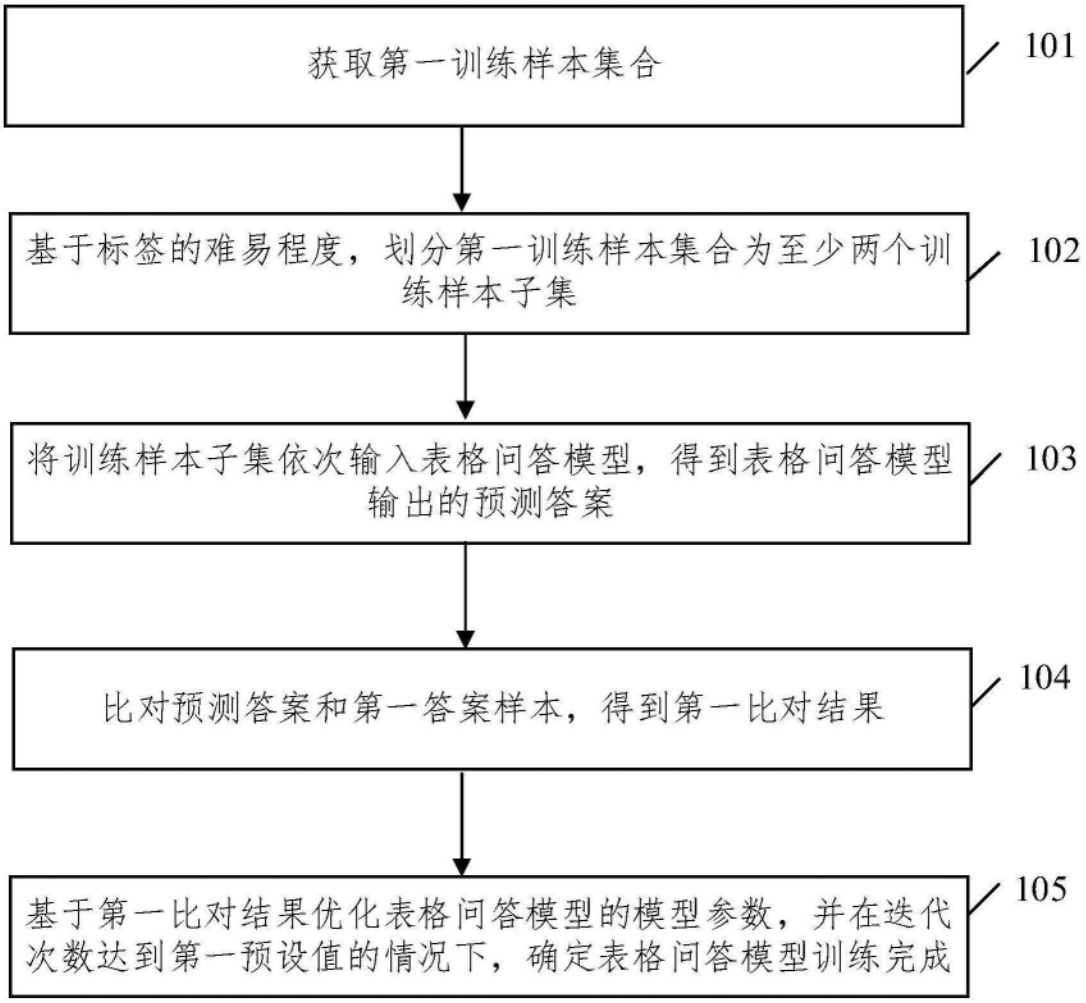

获取第一训练样本集合
101
基于标签的难易程度，划分第一训练样本集合为至少两个训练样本子集
102
将训练样本子集依次输入表格问答模型，得到表格问答模型输出的预测答案
103
比对预测答案和第一答案样本，得到第一比对结果
104
基于第一比对结果优化表格问答模型的模型参数，并在迭代次数达到第一预设值的情况下，确定表格问答模型训练完成
105

图1

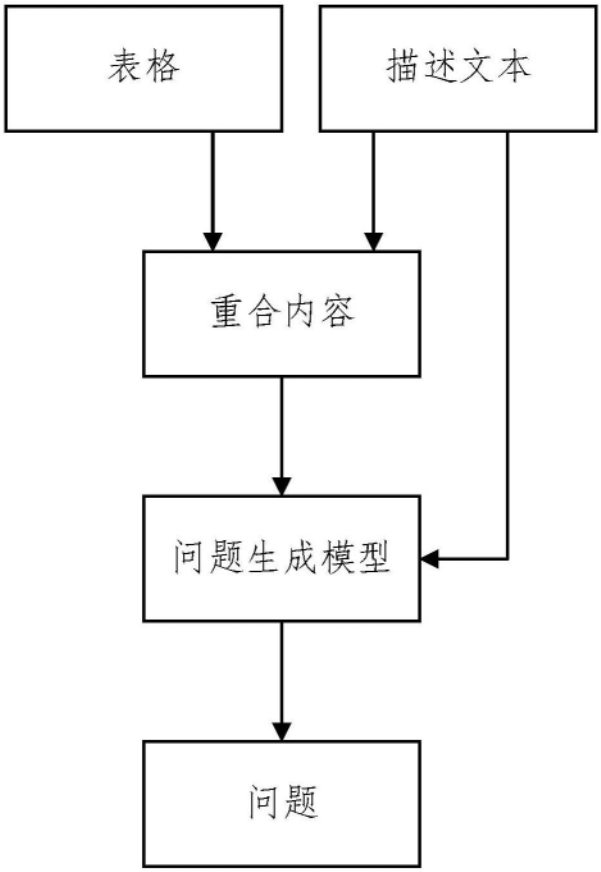


图2

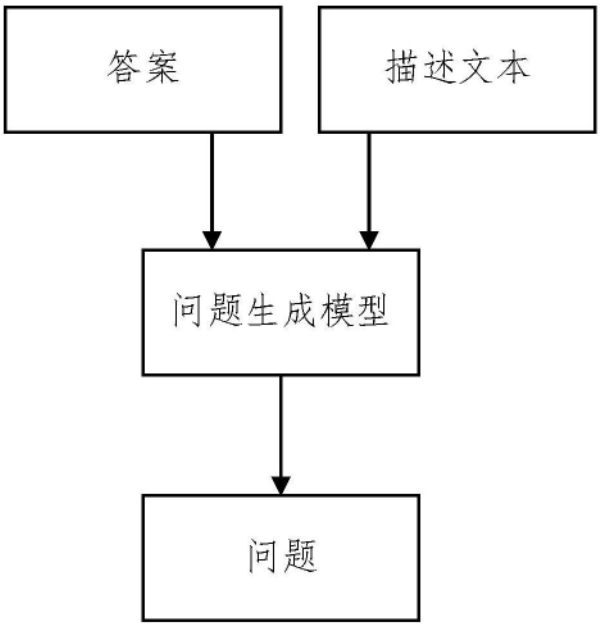


图3

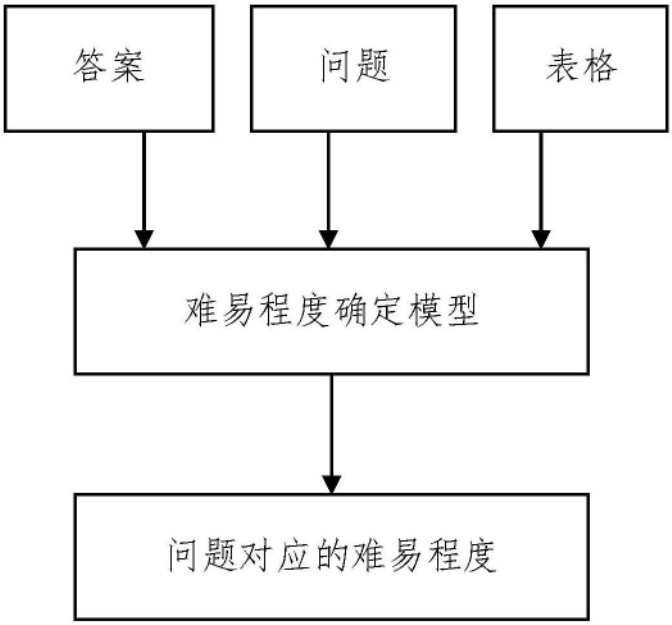


图4

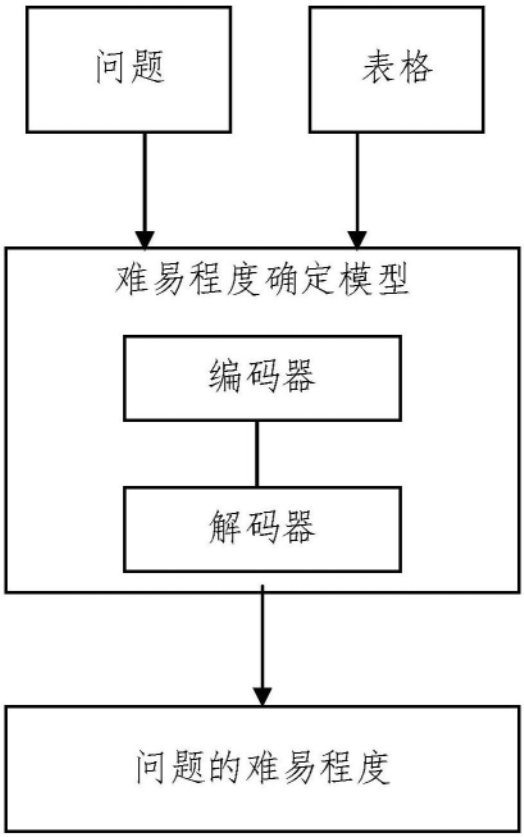


图5

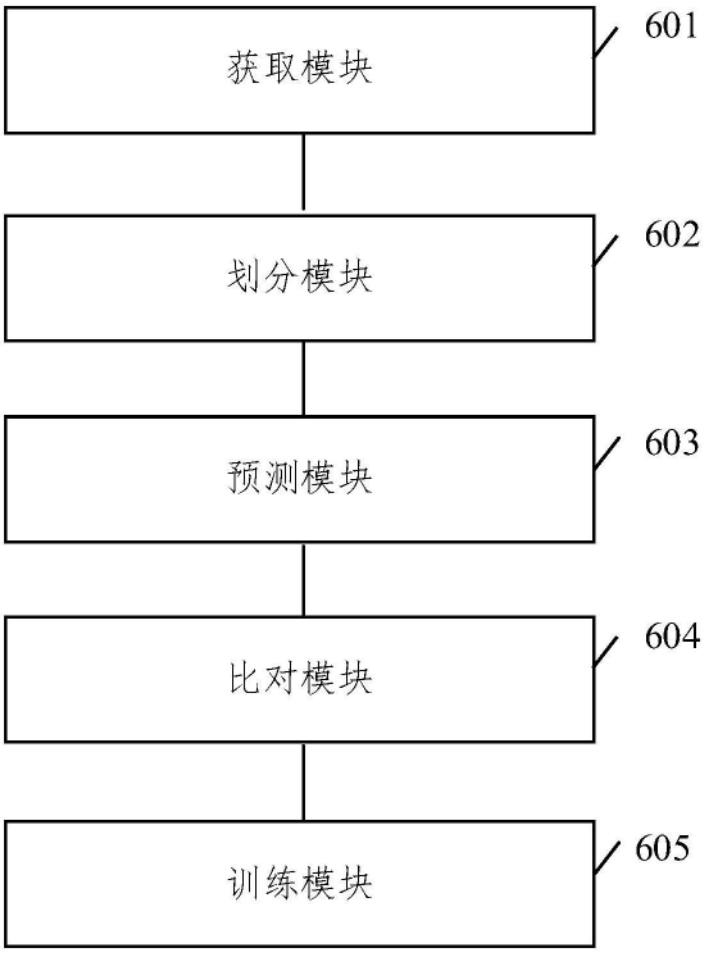

获取模块
601
划分模块
602
预测模块
603
比对模块
604
训练模块
605

图6

目标获取模块
701
目标预测模块
702

图7

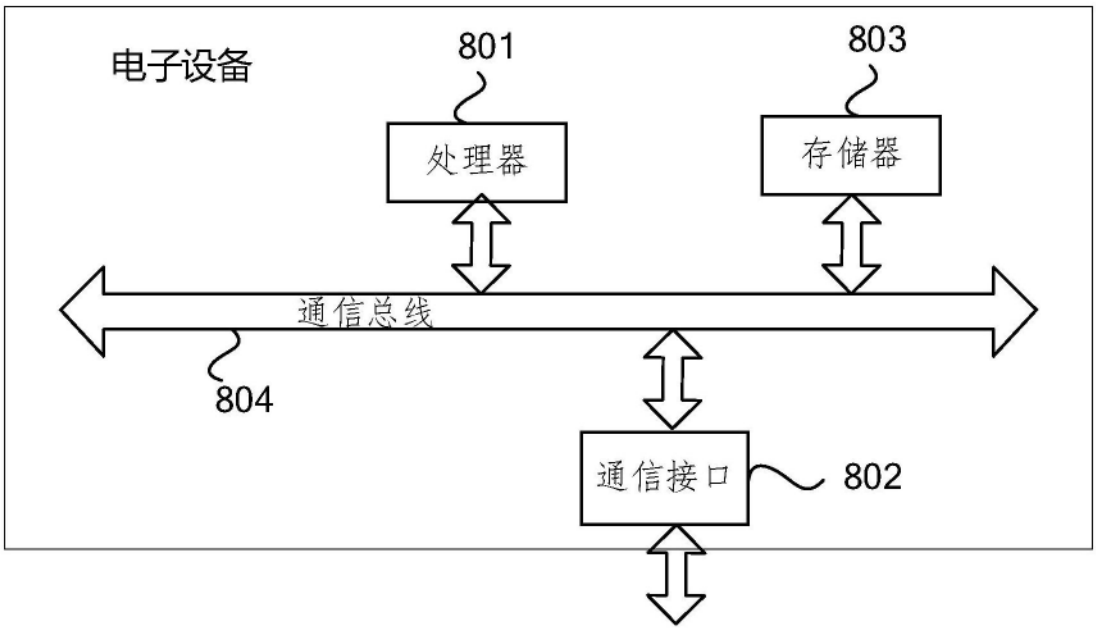


图8