



(12) 发明专利申请

(10) 申请公布号 CN 115840744 A

(43) 申请公布日 2023. 03. 24

(21) 申请号 202211064426.4

(22) 申请日 2022.09.01

(71) 申请人 京东科技信息技术有限公司

地址 100000 北京市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 祝天刚 赵天云 袁韶祖
赵宇明 戴爱君 陈蒙

(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 赵春华

(51) Int. Cl.

G06F 16/22 (2019.01)

G06F 16/28 (2019.01)

G06F 40/30 (2020.01)

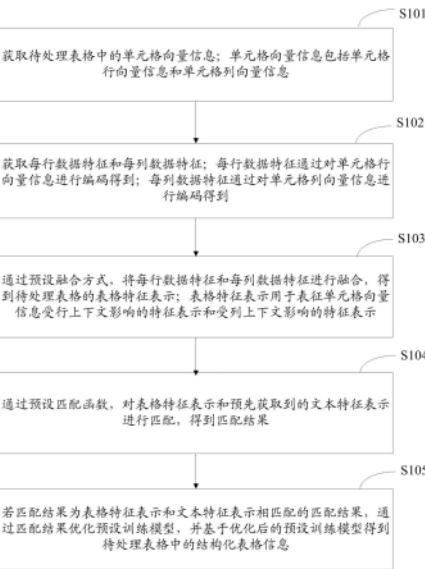
权利要求书2页 说明书10页 附图4页

(54) 发明名称

一种表格数据的处理方法、系统、存储介质及电子设备

(57) 摘要

本申请公开了一种表格数据的处理方法、系统、存储介质及电子设备,获取待处理表格中的单元格向量信息、文本向量信息、获取每行数据特征和每列数据特征,将每行数据特征和每列数据特征进行融合得到待处理表格的表格特征表示,对表格特征表示和文本特征表示进行匹配,得到表格特征表示和文本特征表示相匹配的匹配结果,通过匹配结果优化预设训练模型,基于优化后的预设训练模型得到待处理表格中的结构化表格信息。通过上述,联合学习自由文本和结构化表格的语义表征,从不同空间和层次将不同的信息源数据,如表格的单元格向量信息、文本向量信息、结构化表格的语义表征等数据进行融合得到多粒度信息,通过该多粒度信息有效获取表格的结构化信息。



1. 一种表格数据的处理方法,其特征在于,所述方法包括:

获取待处理表格中的单元格向量信息;所述单元格向量信息包括单元格行向量信息和单元格列向量信息;

获取每行数据特征和每列数据特征;所述每行数据特征通过对所述单元格行向量信息进行编码得到;所述每列数据特征通过对所述单元格列向量信息进行编码得到;

通过预设融合方式,将所述每行数据特征和所述每列数据特征进行融合,得到所述待处理表格的表格特征表示;所述表格特征表示用于表征所述单元格向量信息受行上下文影响的特征表示和受列上下文影响的特征表示;

通过预设匹配函数,对所述表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果;

若所述匹配结果为所述表格特征表示和所述文本特征表示相匹配的匹配结果,通过所述匹配结果优化预设训练模型,并基于优化后的预设训练模型得到所述待处理表格中的结构化表格信息。

2. 根据权利要求1所述的方法,其特征在于,所述获取待处理表格中的单元格向量信息,包括:

通过预设训练模型获取待处理表格中的单元格向量信息。

3. 根据权利要求1所述的方法,其特征在于,获取每行数据特征,包括:

按照预设行顺序对所述单元格行向量信息进行编码,得到行特征表示;所述行特征表示为所述单元格向量信息受行上下文影响的特征表示;

对所述行特征表示添加第一预设字符,得到每行数据特征;所述第一预设字符用于表征每行表格特征的字符。

4. 根据权利要求1所述的方法,其特征在于,获取每列数据特征,包括:

按照预设列顺序对所述单元格列向量信息进行编码,得到列特征表示;所述列特征表示为所述单元格向量信息受列上下文影响的特征表示;

对所述列特征表示添加第二预设字符,得到每列数据特征;所述第二预设字符用于表征每列表格特征的字符。

5. 根据权利要求1所述的方法,其特征在于,所述通过预设匹配函数,对所述表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果,包括:

获取待处理表格中的文本向量信息,并通过所述文本向量信息确定文本特征表示;

在通过预设匹配函数对所述表格特征表示与所述文本特征表示进行匹配时,获取所述表格特征表示对应的表格和所述文本特征表示对应的文本之间的匹配相似度分数;

当所述匹配相似度分数大于预设相似度分数时,确定所述表格特征表示和所述文本特征表示匹配,并生成表征所述表格特征表示和所述文本特征表示相匹配的匹配结果;

当所述匹配相似度分数小于预设相似度分数时,确定所述表格特征表示和所述文本特征表示不匹配,并生成表征所述表格特征表示和所述文本特征表示不匹配的结果。

6. 一种表格数据的处理系统,其特征在于,所述系统包括:

第一获取单元,用于获取待处理表格中的单元格向量信息;所述单元格向量信息包括单元格行向量信息和单元格列向量信息;

第二获取单元,用于获取每行数据特征和每列数据特征;所述每行数据特征通过对所

述单元格行向量信息进行编码得到;所述每列数据特征通过对所述单元格列向量信息进行编码得到;

融合单元,用于通过预设融合方式,将所述每行数据特征和所述每列数据特征进行融合,得到所述待处理表格的表格特征表示;所述表格特征表示用于表征所述单元格向量信息受行上下文影响的特征表示和受列上下文影响的特征表示;

匹配单元,用于通过预设匹配函数,对所述表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果;

第三获取单元,用于若所述匹配结果为所述表格特征表示和所述文本特征表示相匹配的匹配结果,通过所述匹配结果优化预设训练模型,并基于优化后的预设训练模型得到所述待处理表格中的结构化表格信息。

7. 根据权利要求6所述的系统,其特征在于,所述第一获取单元,具体用于:

通过通过预设训练模型获取待处理表格中的单元格向量信息。

8. 根据权利要求6所述的系统,其特征在于,获取每行数据特征的第二获取单元,包括:

第一编码模块,用于按照预设行顺序对所述单元格行向量信息进行编码,得到行特征表示;所述行特征表示为所述单元格向量信息受行上下文影响的特征表示;

第一添加模块,用于对所述行特征表示添加第一预设字符,得到每行数据特征;所述第一预设字符用于表征每行表格特征的字符。

9. 一种存储介质,其特征在于,所述存储介质包括存储的指令,其中,在所述指令运行时控制所述存储介质所在的设备执行如权利要求1至5任意一项所述的表格数据的处理方法。

10. 一种电子设备,其特征在于,包括存储器,以及一个或者一个以上的指令,其中一个或者一个以上指令存储于存储器中,且经配置以由一个或者一个以上处理器执行如权利要求1至5任意一项所述的表格数据的处理方法。

一种表格数据的处理方法、系统、存储介质及电子设备

技术领域

[0001] 本发明涉及数据处理技术领域,更具体地说,涉及一种表格数据的处理方法、系统、存储介质及电子设备。

背景技术

[0002] 随着文本的自然语言(NL)理解任务的预训练语言模型(LMs)蓬勃发展,预训练的语言表征模型(Bidirectional Encoder Representations from Transformer, BERT)等预训练模型借助大量的长文本自然语言作为预料,通过设置自监督的预训练任务对模型进行训练,通过训练后的模型来进行文本分类、实体识别、文本匹配等操作。

[0003] 虽然BERT,预训练的模型(TaBERT),生成式的预训练模型(Generative Pre-Training, GPT)等预训练模型在文本分类、文本匹配、文本生成等下游任务中具有很显著的效果,但是通常此类模型主要针对自然语言长文本的数据进行训练,其设置的预训练任务也只针对自然文本,从而无法有效获取到表格中结构化的表格信息。

[0004] 因此,如何有效获取到表格中结构化的表格信息,成为本领域技术人员亟待解决的问题。

发明内容

[0005] 有鉴于此,本申请公开了一种表格数据的处理方法、系统、存储介质及电子设备,旨在通过该多粒度信息有效获取表格的结构化信息。

[0006] 为了实现上述目的,其公开的技术方案如下:

[0007] 本申请第一方面公开了一种表格数据的处理方法,所述方法包括:

[0008] 获取待处理表格中的单元格向量信息;所述单元格向量信息包括单元格行向量信息和单元格列向量信息;

[0009] 获取每行数据特征和每列数据特征;所述每行数据特征通过对所述单元格行向量信息进行编码得到;所述每列数据特征通过对所述单元格列向量信息进行编码得到;

[0010] 通过预设融合方式,将所述每行数据特征和所述每列数据特征进行融合,得到所述待处理表格的表格特征表示;所述表格特征表示用于表征所述单元格向量信息受行上下文影响的特征表示和受列上下文影响的特征表示;

[0011] 通过预设匹配函数,对所述表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果;

[0012] 若所述匹配结果为所述表格特征表示和所述文本特征表示相匹配的匹配结果,通过所述匹配结果优化预设训练模型,并基于优化后的预设训练模型得到所述待处理表格中的结构化表格信息。

[0013] 优选的,所述获取待处理表格中的单元格向量信息,包括:

[0014] 通过预设训练模型获取待处理表格中的单元格向量信息。

[0015] 优选的,获取每行数据特征,包括:

[0016] 按照预设行顺序对所述单元格行向量信息进行编码,得到行特征表示;所述行特征表示为所述单元格向量信息受行上下文影响的特征表示;

[0017] 对所述行特征表示添加第一预设字符,得到每行数据特征;所述第一预设字符用于表征每行表格特征的字符。

[0018] 优选的,获取每列数据特征,包括:

[0019] 按照预设列顺序对所述单元格列向量信息进行编码,得到列特征表示;所述列特征表示为所述单元格向量信息受列上下文影响的特征表示;

[0020] 对所述列特征表示添加第二预设字符,得到每列数据特征;所述第二预设字符用于表征每列表格特征的字符。

[0021] 优选的,所述通过预设匹配函数,对所述表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果,包括:

[0022] 获取待处理表格中的文本向量信息,并通过所述文本向量信息确定文本特征表示;

[0023] 在通过预设匹配函数对所述表格特征表示与所述文本特征表示进行匹配时,获取所述表格特征表示对应的表格和所述文本特征表示对应的文本之间的匹配相似度分数;

[0024] 当所述匹配相似度分数大于预设相似度分数时,确定所述表格特征表示和所述文本特征表示匹配,并生成表征所述表格特征表示和所述文本特征表示相匹配的匹配结果;

[0025] 当所述匹配相似度分数小于预设相似度分数时,确定所述表格特征表示和所述文本特征表示不匹配,并生成表征所述表格特征表示和所述文本特征表示不匹配的结果。

[0026] 本申请第二方面公开了一种表格数据的处理系统,所述系统包括:

[0027] 第一获取单元,用于获取待处理表格中的单元格向量信息;所述单元格向量信息包括单元格行向量信息和单元格列向量信息;

[0028] 第二获取单元,用于获取每行数据特征和每列数据特征;所述每行数据特征通过对所述单元格行向量信息进行编码得到;所述每列数据特征通过对所述单元格列向量信息进行编码得到;

[0029] 融合单元,用于通过预设融合方式,将所述每行数据特征和所述每列数据特征进行融合,得到所述待处理表格的表格特征表示;所述表格特征表示用于表征所述单元格向量信息受行上下文影响的特征表示和受列上下文影响的特征表示;

[0030] 匹配单元,用于通过预设匹配函数,对所述表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果;

[0031] 第三获取单元,用于通过所述匹配结果优化预设训练模型,并基于优化后的预设训练模型得到所述待处理表格中的结构化表格信息。

[0032] 优选的,所述第一获取单元,具体用于:

[0033] 通过通过预设训练模型获取待处理表格中的单元格向量信息。

[0034] 优选的,获取每行数据特征的第二获取单元,包括:

[0035] 第一编码模块,用于按照预设行顺序对所述单元格行向量信息进行编码,得到行特征表示;所述行特征表示为所述单元格向量信息受行上下文影响的特征表示;

[0036] 第一添加模块,用于对所述行特征表示添加第一预设字符,得到每行数据特征;所述第一预设字符用于表征每行表格特征的字符。

[0037] 本申请第三方面公开了一种存储介质,所述存储介质包括存储的指令,其中,在所述指令运行时控制所述存储介质所在的设备执行如第一方面任意一项所述的表格数据的处理方法。

[0038] 本申请第四方面公开了一种电子设备,包括存储器,以及一个或者一个以上的指令,其中一个或者一个以上指令存储于存储器中,且经配置以由一个或者一个以上处理器执行如第一方面任意一项所述的表格数据的处理方法。

[0039] 经由上述技术方案可知,本申请公开了一种表格数据的处理方法、系统、存储介质及电子设备,获取待处理表格中的单元格向量信息,单元格向量信息包括单元格行向量信息和单元格列向量信息,获取每行数据特征和每列数据特征,每行数据特征通过对所述单元格行向量信息进行编码得到,每列数据特征通过对单元格列向量信息进行编码得到,通过预设融合方式,将每行数据特征和每列数据特征进行融合,得到待处理表格的表格特征表示;表格特征表示用于表征单元格向量信息受行上下文影响的特征表示和受列上下文影响的特征表示,通过预设匹配函数,对表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果,若匹配结果为表格特征表示和文本特征表示相匹配的匹配结果,通过匹配结果优化预设训练模型,并基于优化后的预设训练模型得到待处理表格中的结构化表格信息。通过上述方案,联合学习自由文本和结构化表格的语义表征,从不同空间和层次将不同的信息源数据,如表格的单元格向量信息、文本向量信息、结构化表格的语义表征等数据进行融合得到多粒度信息,通过该多粒度信息有效获取表格的结构化信息。

附图说明

[0040] 为了更清楚地说明本发明实施例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简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的实施例,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,还可以根据提供的附图获得其他的附图。

[0041] 图1为本申请实施例公开的一种表格数据的处理方法的流程示意图;

[0042] 图2为本申请实施例公开的预训练的BERT模型的示例图;

[0043] 图3为本申请实施例公开的预训练表格和文本数据的示例图;

[0044] 图4为本申请实施例公开的列名预测示例图;

[0045] 图5为本申请实施例公开的列值预测示例图;

[0046] 图6为本申请实施例公开的文本及表格相关度预测的示例图;

[0047] 图7为本申请实施例公开的一种表格数据的处理系统的结构示意图;

[0048] 图8为本申请实施例公开的一种电子设备的结构示意图。

具体实施方式

[0049] 下面将结合本发明实施例中的附图,对本发明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。

[0050] 在本申请中,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、物品或者设备中还存在另外的相同要素。

[0051] 由背景技术可知,虽然BERT, TaBERT, GPT等预训练模型在文本分类、文本匹配、文本生成等下游任务中具有很显著的效果,但是通常此类模型主要针对自然语言长文本的数据进行训练,其设置的预训练任务也只针对自然文本,从而无法有效获取到表格中结构化的表格信息。因此,如何有效获取到表格中结构化的表格信息,成为本领域技术人员亟待解决的问题。

[0052] 为了解决上述问题,本申请实施例公开了一种表格数据的处理方法、系统、存储介质及电子设备,联合学习自由文本和结构化表格的语义表征,从不同空间和层次将不同的信息源数据,如表格的单元格向量信息、文本向量信息、结构化表格的语义表征等数据进行融合得到多粒度信息,通过该多粒度信息有效获取表格的结构化信息。具体实现方式通过下述实施例具体进行说明。

[0053] 参考图1所示,为本申请实施例公开的一种表格数据的处理方法的流程示意图,该表格数据的处理方法主要包括如下步骤:

[0054] S101:获取待处理表格中的单元格向量信息;单元格向量信息包括单元格行向量信息和单元格列向量信息。

[0055] 在S101中,通过预设训练模型获取待处理表格中的单元格向量信息。

[0056] 其中,待处理表格即为未通过预设训练模型进行处理的表格。

[0057] 预设训练模型可以是预训练的(Bidirectional Encoder Representation from Transformers, BERT)模型,也可以是其他类型的模型,具体预设训练模型的确定由技术人员根据实际情况进行设置,本申请不做具体限定。本申请的预设训练模型优选预训练的BERT模型。

[0058] 预训练的BERT模型包括RowTransformer模型、ColumnTransformer模型和TextTransformer模型。为了方便理解预训练的BERT模型,这里结合图2进行说明,图2示出了预训练的BERT模型的示例图。

[0059] 按照传统的BERT预训练方式中,表格中的列和行信息是组合成一个线性的长的文本序列进行训练的(例如将上面表格内容线性化成“套餐名称长寿大王卡4G[SEP]价格39[SEP]合作商京东[SEP]套餐名称淘宝大王卡…”)。但是这种数据输入方式在进行模型训练时,使得不同列、行之间的单元格值被独立,无法获取期间的上下文依存关系(比如图2中“长寿卡4G套餐”和“淘宝大王卡5G套餐”同属于一个列,具有一定相似度)。

[0060] 因此,需要通过RowTransformer模型对单元格行向量信息进行编码,得到每行数据特征,通过ColumnTransformer模型对单元格列向量信息进行编码,得到每列数据特征,通过TextTransformer模型对文本向量信息进行编码,得到文本向量信息,通过每行数据特征、每列数据特征和文本向量信息得到上下文依存关系。

[0061] 文本向量信息是通过预训练的BERT模型获取得到。

[0062] 通过TextTransformer模型对文本向量信息进行编码,得到文本特征表示。文本特征表示用于与下述得到的表格特征表示进行匹配来得到待处理表格中的结构化表格信息。

[0063] 具体BERT模型的预训练过程包括对BERT模型的表格&文本数据的预训练、基于掩码的语言模型 (Masked Language Modeling, MLM) 建模、掩码实体词预测Masked Entity Prediction、掩码实体词预测Masked Entity Prediction、列名预测Column Name Prediction、列值预测RowValue Prediction和文本表格相关度预测NL&TableMatch。

[0064] 对BERT模型的表格&文本数据的预训练:

[0065] 在表格和文本数据的预训练任务中,通过电商领域和公开网站等收集大量表格和表格相关的文本数据,组成Table&NL (natural language) 预训练对进行预训练,参考图3所示。图3示出了预训练表格和文本数据的示例图。

[0066] 图3中,表格内容包括节假日、日期和假期天数(不含调休)。其中,节假日包括新年(元旦)、春节、清明、劳动节、端午、中秋和国庆节;日期包括1月1日、农历正月初一、初二、初三、清明节当日、5月1日、农历端午当日、农历中秋当日、10月1日、2日、3日;假期天数(不含调休)包括1天、3天等。

[0067] 文本数据的内容为“中国重大的传统节日有春节、元宵节、清明节等。此外少数民族也都保留自己传统节日,诸如傣族的泼水节,蒙古族的那达慕大会…”。

[0068] 基于掩码的语言模型 (Masked Language Modeling, MLM) 建模:

[0069] 该任务参考BERT模型中的预训练任务,针对所有的文本的自然语言 (NL), 随机对某个单词使用 [Masked] 词汇进行掩码,通过上下文NL来预测被掩码的词汇。例如,针对图3中的“此外少数民族也都保留自己传统节日,诸如傣族的泼水节,蒙古族的那达慕大会…”随机替换成“此外少数民 [MASK] 也都保 [MASK] 自己传统节日,诸如傣 [MASK] 的泼水节,蒙古 [MASK] 的那达 [MASK] 大…”,通过语言模型预测被掩盖的单词。

[0070] 掩码实体词预测Masked Entity Prediction:

[0071] 传统的MLM仅仅针对整个自然语言的上下文特征对某个掩码词汇进行预测,无法获取表格信息特征。为了增强表格预训练任务NL和Table的信息交互,捕获其中的知识共享,申请提出新的预训练任务。该任务首先使用n-gram字符匹配的方式,将在NL中提到的Table某个单元格实体词(列名、列值均可)识别出来,随后将该实体词按照一定比例进行掩码,随后通过NL和Table的上下文整体特征来预测被掩盖的实体词。例如在NL“中国重大的传统节日有春节,元宵节,清明节等”,其中,“春节,元宵,清明”等均是表格列值中提到的实体词,将其中词汇随机掩盖,使用NL和Table特征进行预测:“中国重大的传统节日有 [MASK] [MASK],元宵节,清明节等”。

[0072] 列名预测Column Name Prediction:

[0073] 为了加强表格内部的结构化信息学习,同时提出两个针对表格信息学习的预训练任务,其中Column Name Prediction主要是将表格中的列名进行随机的掩码,随后根据该列相关的空间信息和层次信息对该列进行预测。列名预测的具体过程结合图4进行说明。

[0074] 图4中,将“合作商”进行掩码,得到空间信息(横向根据列名获取的信息)“套餐名称 [SEP] 价格 [SEP] 合作商”->“套餐名称 [SEP] 价格 [SEP] [MASK] [MASK] [MASK]”和层次信息“[COL] | 合作商 | 京东 | 阿里 | …”->“[COL] | [MASK] [MASK] [MASK] | 京东 | 阿里 | …”,将利用横向信息和纵向信息预测该列名。

[0075] 列值预测RowValue Prediction:

[0076] 列值预测的具体过程结合图5进行说明。图5中,列值预测也是随机将表格中的某

个列值进行掩盖,随后利用该列的空间信息和该行的层次信息作为上下文预测被掩盖的列值。

[0077] 文本表格相关度预测NL&TableMatch:

[0078] 为了进一步增强模型NL和Table信息表征的交互知识融合,额外增加了文本表格相关度预测的任务。在数据收集过程中,将收集到与Table相关的NL文本描述作为正例样本,将NL与其他随机Table组合后构建成本负例样本。具体文本表格相关度预测NL&TableMatch结合图6进行说明,图6示出了文本及表格相关度预测的示例图。

[0079] 图6中,simScore为文本和某列/某行进行匹配的时候的相似度分数(similarity score),相似度分数越高说明两者匹配,分数低则说明两者不匹配。

[0080] 在训练过程中,我们将Row transformer获取到的每行数据特征[ROW]和ColumnTransformer获取到的每列数据特征[COL]通过自注意力学习获取到表格table表征(图6中[ROW]和[COL])。随后通过BERT transformer获取到[CLS]表征作为文本特征表示;最后通过分类函数获取两个表征是否匹配。

[0081] S102:获取每行数据特征和每列数据特征;每行数据特征通过对单元格行向量信息进行编码得到;每列数据特征通过对单元格列向量信息进行编码得到。

[0082] 在S102中,通过RowTransformer模型对单元格行向量信息进行编码,得到每行数据特征;通过ColumnTransformer模型对单元格列向量信息进行编码,得到每列数据特征。

[0083] 获取每行数据特征的过程如A1-A2所示,并结合图2进行说明。

[0084] A1:按照预设行顺序对单元格行向量信息进行编码,得到行特征表示;行特征表示为单元格向量信息受行上下文影响的特征表示。

[0085] 其中,按照图2中每行由左至右的顺序对单元格信息进行编码,假设在Row_i包含c_1,c_2,c_n等单元格的向量信息,将该序列融合到RowTransformer模型中,随后通过BERT相同的自注意力机制等获取到每个单元格信息受行上下文影响的行特征表示r_1,r_2,r_3等。

[0086] A2:对行特征表示添加第一预设字符,得到每行数据特征;第一预设字符用于表征每行表格特征的字符。

[0087] 其中,在每行数据的开头添加如图2中的特殊字符[ROW],即第一预设字符,得到每行的表格特征。

[0088] 具体获取每列数据特征的过程如B1-B2所示,并结合图2进行说明。

[0089] B1:按照预设列顺序对单元格列向量信息进行编码,得到列特征表示;列特征表示为单元格向量信息受列上下文影响的特征表示。

[0090] 其中,按照图2中每列由上至下的顺序对单元格信息进行编码,假设在Column_i包含c_1,c_2,c_n等单元格的向量信息,将该序列融合到ColumnTransformer单元中,随后通过BERT相同的自注意力机制等获取到每个单元格信息受列上下文影响的列特征表示c_1,c_2,c_3等。

[0091] B2:对列特征表示添加第二预设字符,得到每列数据特征;第二预设字符用于表征每列表格特征的字符。

[0092] 在每列数据的开头添加如图2中的特殊字符[COL],即第二预设字符,得到每列的表格特征。

[0093] S103:通过预设融合方式,将每行数据特征和每列数据特征进行融合,得到待处理表格的表格特征表示;表格特征表示用于表征单元格向量信息受行上下文影响的特征表示和受列上下文影响的特征表示。

[0094] 其中,通过预设融合方式可以是均值池化 (Mean_pooling) 的融合方式,也可以是其他类型的融合方式,具体预设融合方式的确定由技术人员根据实际情况进行设置,本申请不做具体限定。本申请的预设融合方式优选均值池化的融合方式。

[0095] 每个单元格在经过RowTransformer模型和ColumnTransformer模型编码后,分别有两个不同的特征表示(在RowTransformer模型中,图2的“京东”表示的是第二行的某个表格值向量特征,在ColumnTransformer模型中,图2的“京东”表示的是第三列的某个表格值的向量特征),通过均值池化 (Mean_pooling) 的方式将两个粒度的信息(上述两个不同的特征表示)进行融合,从而获取到该单元格的最终特征表示。

[0096] 为了获取到整体的表格特征,将每列数据特征[COL]、每行数据特征[ROW]和单元格的最终特征表示进行自注意力学习后进行融合成单一向量,作为整体特征表示,即待处理表格的表格特征表示。

[0097] S104:通过预设匹配函数,对表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果。

[0098] 在S104中,预设匹配函数可以是n-gram字符匹配函数,也可以是其他匹配函数,具体预设匹配函数的确定由技术人员根据实际情况进行设置,本申请不做限定。本申请的预设匹配函数优选n-gram字符匹配函数。

[0099] 通过预训练的BERT模型获取文本向量信息,并通过TextTransformer模型对文本向量信息进行编码,得到文本特征表示。

[0100] 假设文本中包含 $w_1, w_2, w_3 \cdots w_n$ 多个字符(其中,n的取值为大于等于1的整数)经过编码后,获取到 $h_1, h_2, h_3 \cdots h_n$ 等文本特征表示。

[0101] 具体通过预设匹配函数,对表格特征表示和文本特征表示进行匹配,得到匹配结果的过程如C1-C4所示。

[0102] C1:获取待处理表格中的文本向量信息,并通过文本向量信息确定文本特征表示。

[0103] 例如,参考图2所示,“这款长寿卡4G套餐是京东合作推出的”即为文本向量信息。

[0104] C2:在通过预设匹配函数对表格特征表示与文本特征表示进行匹配时,获取表格特征表示对应的表格和文本特征表示对应的文本之间的匹配相似度分数。

[0105] C3:当匹配相似度分数大于预设相似度分数时,确定表格特征表示和文本特征表示相匹配,并生成表征表格特征表示和文本特征表示相匹配的匹配结果。

[0106] C4:当匹配相似度分数小于预设相似度分数时,确定表格特征表示和文本特征表示不匹配,并生成表征表格特征表示和文本特征表示不匹配的结果。

[0107] S105:若匹配结果为表格特征表示和文本特征表示相匹配的匹配结果,通过匹配结果优化预设训练模型,并基于优化后的预设训练模型得到待处理表格中的结构化表格信息。

[0108] 在S105中,若表格相关度和文本相关度一致,得到表征表格特征表示和文本特征表示相匹配的匹配结果,并通过匹配结果优化预设训练模型,基于优化后的预设训练模型得到待处理表格中的结构化表格信息。

[0109] 在预设训练模型训练过程中,通过不断优化匹配结果,来优化模型对文本信息和表格中的结构化信息的学习。

[0110] 待处理表格中的结构化信息是指待处理表格中的信息经过分析后可分解成多个互相关联的组成部分,各组成部分间有明确的层次结构。

[0111] 本申请实施例中,联合学习自由文本和结构化表格的语义表征,从不同空间和层次将不同的信息源数据,如表格的单元格向量信息、文本向量信息、结构化表格的语义表征等数据进行融合得到多粒度信息,通过该多粒度信息有效获取表格的结构化信息。

[0112] 基于上述实施例图1公开了一种表格数据的处理方法,本申请实施例还对应公开了一种表格数据的处理系统,参考图7所示,该表格数据的处理系统包括第一获取单元701、第二获取单元702、融合单元703、匹配单元704和第三获取单元705。

[0113] 第一获取单元701,用于获取待处理表格中的单元格向量信息;单元格向量信息包括单元格行向量信息和单元格列向量信息。

[0114] 第二获取单元702,用于获取每行数据特征和每列数据特征;每行数据特征通过对单元格行向量信息进行编码得到;每列数据特征通过对单元格列向量信息进行编码得到。

[0115] 融合单元703,用于通过预设融合方式,将每行数据特征和每列数据特征进行融合,得到待处理表格的表格特征表示;表格特征表示用于表征单元格向量信息受行上下文影响的特征表示和受列上下文影响的特征表示。

[0116] 匹配单元704,用于通过预设匹配函数,对表格特征表示和预先获取到的文本特征表示进行匹配,得到匹配结果。

[0117] 第三获取单元705,用于若匹配结果为表格特征表示和文本特征表示相匹配的匹配结果,通过匹配结果优化预设训练模型,并基于优化后的预设训练模型得到待处理表格中的结构化表格信息。

[0118] 进一步的,第一获取单元701具体用于通过预设训练模型获取待处理表格中的单元格向量信息。

[0119] 进一步的,获取每行数据特征的第二获取单元702包括第一编码模块和第一添加模块。

[0120] 第一编码模块,用于按照预设行顺序对单元格行向量信息进行编码,得到行特征表示;行特征表示为单元格向量信息受行上下文影响的特征表示。

[0121] 第一添加模块,用于对行特征表示添加第一预设字符,得到每行数据特征;第一预设字符用于表征每行表格特征的字符。

[0122] 进一步的,获取每列数据特征的第二获取单元702包括第二编码模块和第二添加模块。

[0123] 第二编码模块,用于按照预设列顺序对所述单元格列向量信息进行编码,得到列特征表示;列特征表示为单元格向量信息受列上下文影响的特征表示。

[0124] 第二添加模块,用于对列特征表示添加第二预设字符,得到每列数据特征;第二预设字符用于表征每列表格特征的字符。

[0125] 进一步的,匹配单元704包括第一获取模块、第二获取模块、第一确定模块和第二确定模块。

[0126] 第一获取模块,用于获取待处理表格中的文本向量信息,并通过文本向量信息确

定文本特征表示。

[0127] 第二获取模块,用于在通过预设匹配函数对所述表格特征表示与文本特征表示进行匹配时,获取表格特征表示对应的表格和文本特征表示对应的文本之间的匹配相似度分数。

[0128] 第一确定模块,用于当匹配相似度分数大于预设相似度分数时,确定表格特征表示和文本特征表示相匹配,并生成表征表格特征表示和文本特征表示相匹配的匹配结果。

[0129] 第二确定模块,用于当匹配相似度分数小于预设相似度分数时,确定表格特征表示和文本特征表示不匹配,并生成表征表格特征表示和文本特征表示不匹配的结果。

[0130] 本申请实施例中,联合学习自由文本和结构化表格的语义表征,从不同空间和层次将不同的信息源数据,如表格的单元格向量信息、文本向量信息、结构化表格的语义表征等数据进行融合得到多粒度信息,通过该多粒度信息有效获取表格的结构化信息。

[0131] 本申请实施例还提供了一种存储介质,存储介质包括存储的指令,其中,在指令运行时控制存储介质所在的设备执行上述表格数据的处理方法。

[0132] 本发明实施例还提供了一种电子设备,其结构示意图如图8所示,具体包括存储器801,以及一个或者一个以上的指令802,其中一个或者一个以上指令802存储于存储器801中,且经配置以由一个或者一个以上处理器803执行所述一个或者一个以上指令802执行上述表格数据的处理方法。

[0133] 上述各个实施例的具体实施过程及其衍生方式,均在本发明的保护范围之内。

[0134] 本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于系统或系统实施例而言,由于其基本相似于方法实施例,所以描述得比较简单,相关之处参见方法实施例的部分说明即可。以上所描述的系统及系统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劳动的情况下,即可以理解并实施。

[0135] 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。

[0136] 对所公开的实施例的上述说明,使本领域技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

[0137] 以上所述仅是本发明的优选实施方式,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,还可以做出若干改进和润饰,这些改进和润饰也应

视为本发明的保护范围。

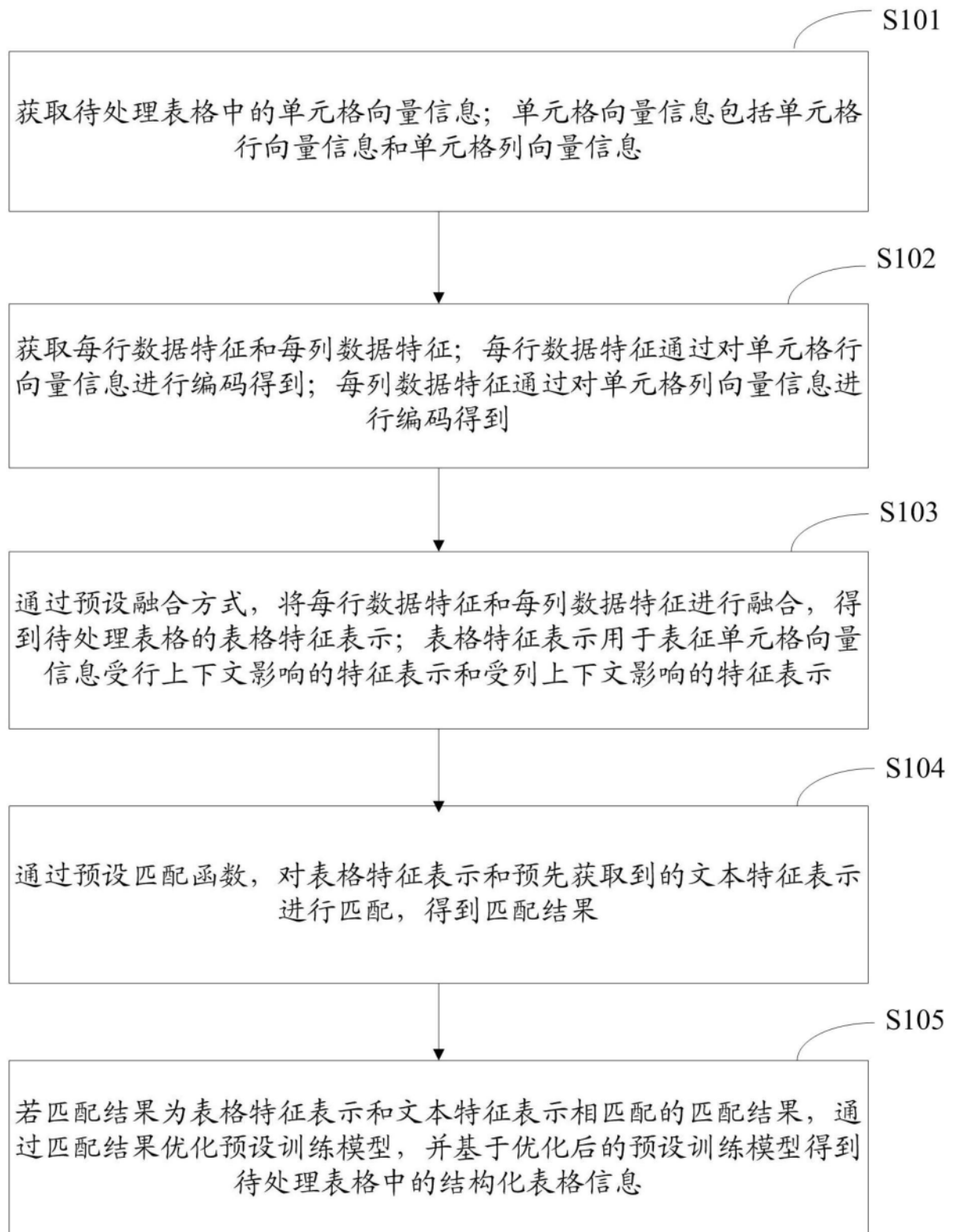


图1

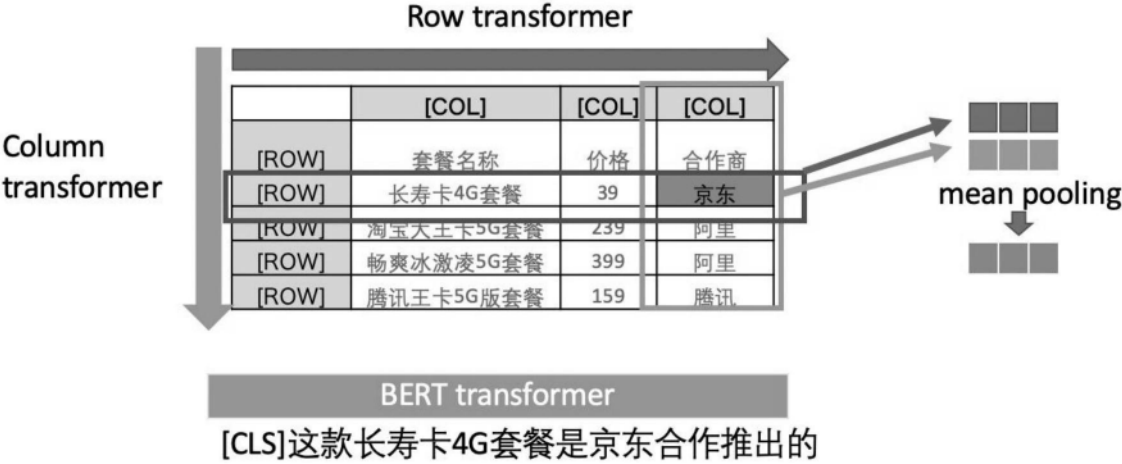


图2

节假日	日期	假期天数（不含调休）
新年（元旦）	1月1日	1天
春节	农历正月初一、初二、初三	3天
清明	清明节当日	1天
劳动节	5月1日	1天
端午	农历端午当日	1天
中秋	农历中秋当日	1天
国庆节	10月1日、2日、3日	3天

中国重大的传统节日有春节，元宵节，清明节等。此外少数民族也都保留自己传统节日，诸如傣族的泼水节，蒙古族的那达慕大会....

图3

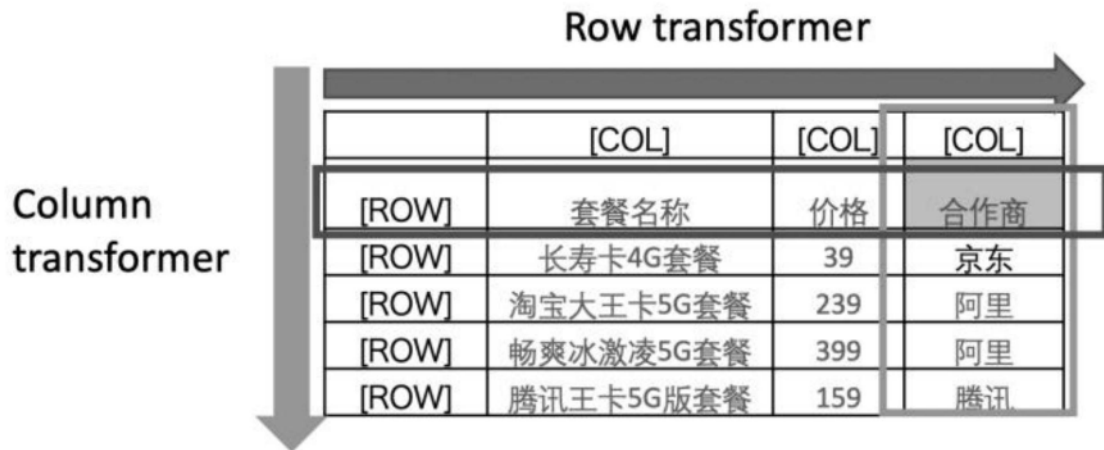


图4



图5

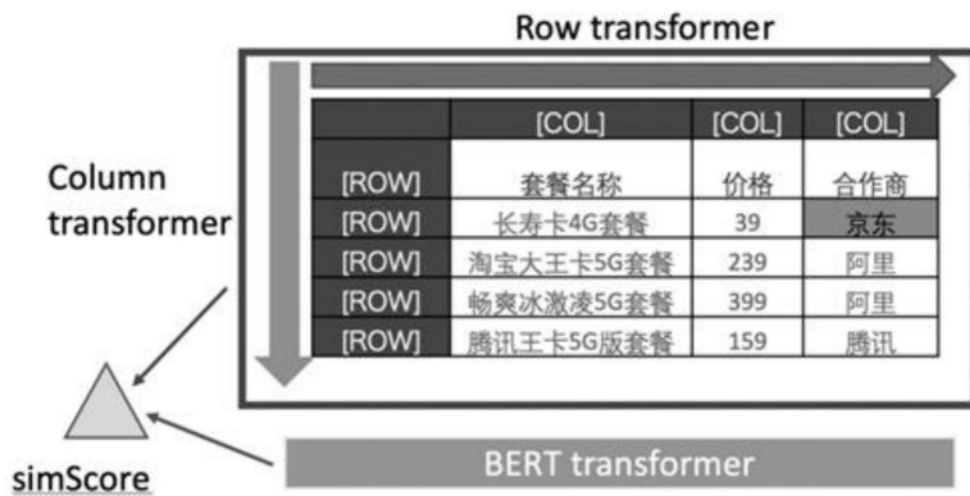


图6

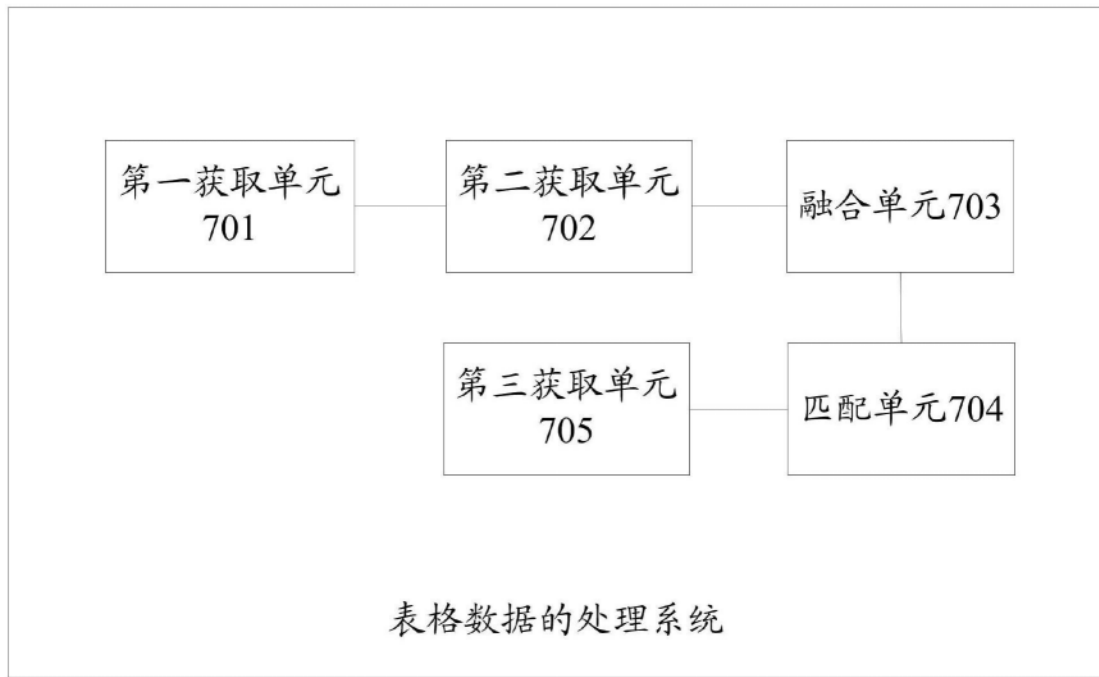


图7

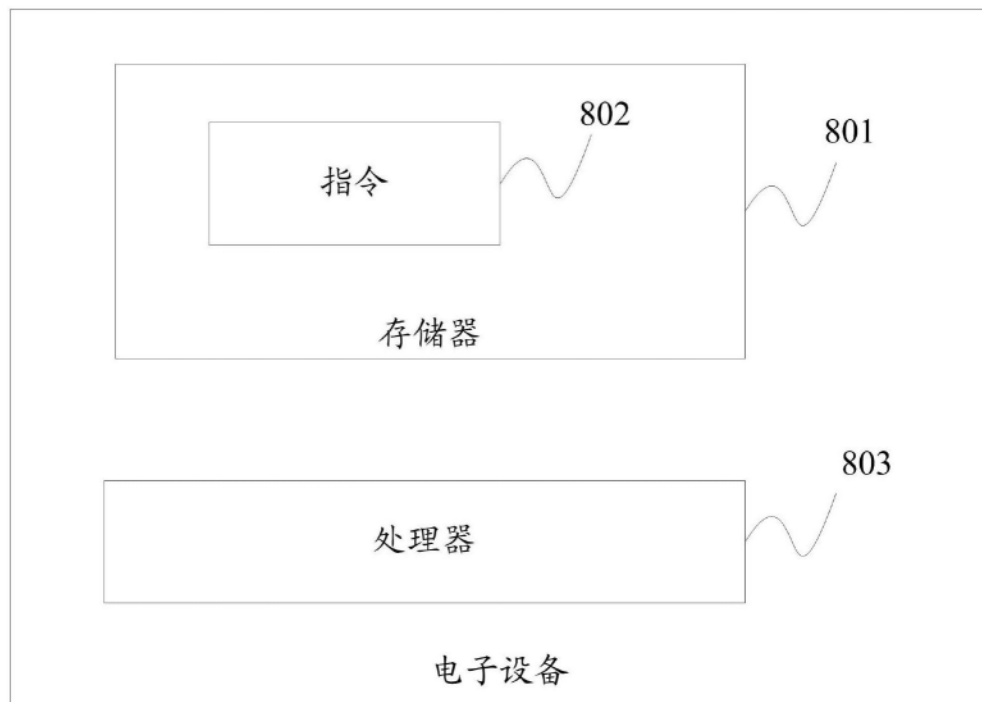


图8