



(12) 发明专利申请

(10) 申请公布号 CN 115545185 A

(43) 申请公布日 2022. 12. 30

(21) 申请号 202211245725.8

(22) 申请日 2022.10.12

(71) 申请人 京东科技信息技术有限公司

地址 100176 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 祝天刚 陈蒙 刘瑞雪 袁韶祖
戴爱君

(74) 专利代理机构 中原信达知识产权代理有限
责任公司 11219

专利代理师 王安娜 李阳

(51) Int.Cl.

G06N 3/08 (2006.01)

G06N 3/04 (2006.01)

G06F 40/30 (2020.01)

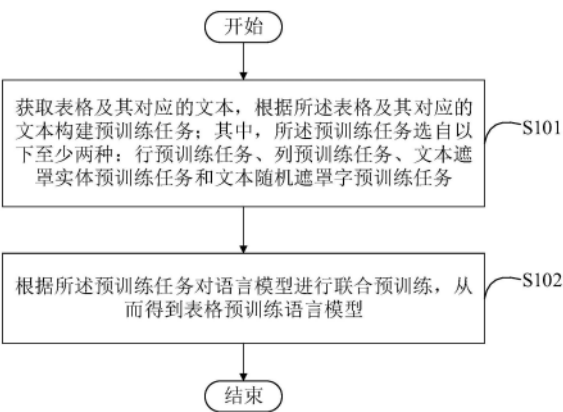
权利要求书2页 说明书10页 附图3页

(54) 发明名称

一种表格预训练方法和装置

(57) 摘要

本发明公开了一种表格预训练方法和装置，涉及人工智能技术领域。该方法的一具体实施方式包括：获取表格及其对应的文本，根据所述表格及其对应的文本构建预训练任务；其中，所述预训练任务选自以下至少两种：行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务；根据所述预训练任务对语言模型进行联合预训练，从而得到表格预训练语言模型。该实施方式能够解决缺乏对文本与表格进行深度结构化语义交互的建模的技术问题。



1. 一种表格预训练方法,其特征在于,包括:

获取表格及其对应的文本,根据所述表格及其对应的文本构建预训练任务;其中,所述预训练任务选自以下至少两种:行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务;

根据所述预训练任务对语言模型进行联合预训练,从而得到表格预训练语言模型。

2. 根据权利要求1所述的方法,其特征在于,若所述预训练任务为行预训练任务,则根据所述表格及其对应的文本构建预训练任务,包括:

根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在行的行号,从而构建得到有监督的行预训练任务。

3. 根据权利要求1所述的方法,其特征在于,若所述预训练任务为列预训练任务,则根据所述表格及其对应的文本构建预训练任务,包括:

根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在列的行号,从而构建得到有监督的列预训练任务。

4. 根据权利要求1所述的方法,其特征在于,若所述预训练任务为文本遮罩实体预训练任务,

则根据所述表格及其对应的文本构建预训练任务,包括:

将所述表格中的列名和所述列名对应的单元格内容作为实体;

按照第一预设遮罩比例对出现在所述文本中的实体进行随机遮罩,从而构建得到无监督的文本遮罩实体预训练任务。

5. 根据权利要求1所述的方法,其特征在于,若所述预训练任务为文本随机遮罩字预训练任务,

则根据所述表格及其对应的文本构建预训练任务,包括:

按照第二预设遮罩比例对所述文本中的字进行随机遮罩,从而构建得到无监督的文本随机遮罩字预训练任务。

6. 根据权利要求4或5所述的方法,其特征在于,所述第一预设遮罩比例为10-25%;

和/或,

所述第二预设遮罩比例为10-25%。

7. 根据权利要求1所述的方法,其特征在于,所述语言模型包括依次串联的输入层、嵌入层、编码层和输出层;其中,所述输出层包括多个并联的子输出层,所述子输出层的数量与所述预训练任务的数量相同。

8. 一种表格预训练装置,其特征在于,包括:

构建模块,用于获取表格及其对应的文本,根据所述表格及其对应的文本构建预训练任务;其中,所述预训练任务选自以下至少两种:行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务;

训练模块,用于根据所述预训练任务对语言模型进行联合预训练,从而得到表格预训练语言模型。

9. 一种电子设备,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行时,所述一个或多个处理器实现如权利要求1-7中任一所述的方法。

10.一种计算机可读介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行时实现如权利要求1-7中任一所述的方法。

11.一种计算机程序产品,包括计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1-7中任一项所述的方法。

一种表格预训练方法和装置

技术领域

[0001] 本发明涉及人工智能技术领域,尤其涉及一种表格预训练方法和装置。

背景技术

[0002] 表格预训练是通过对结构化的表格数据和与之相关的自然语言句子进行的预训练,增强“文本”与“表格”两种多模态数据的对齐。在一系列以表格为基础的下流任务(如自然语言转sql语句,表格事实验证,表格选择,表格分类)中,都需要借助表格预训练来使用表格的结构化信息,以达到更好效果。编码文本和表格信息,避免编码文本与表格时出现不一致的问题,是表格预训练的关键。

[0003] 关于表格预训练模型,业界的主流方法是在大规模文本-表格对的训练数据上,对文本和表格分别建立预训练任务,进行联合预训练。即只在联合预训练目标上进行交互,来体现文本与表格之间的关系。但是这样的交互是十分浅层次的,并没有对表格的结构化信息进行建模,会导致编码文本与表格时语义不一致问题的出现。

发明内容

[0004] 有鉴于此,本发明实施例提供一种表格预训练方法和装置,以解决缺乏对文本与表格进行深度结构化语义交互的建模的技术问题。

[0005] 为实现上述目的,根据本发明实施例的一个方面,提供了一种表格预训练方法,包括:

[0006] 获取表格及其对应的文本,根据所述表格及其对应的文本构建预训练任务;其中,所述预训练任务选自以下至少两种:行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务;

[0007] 根据所述预训练任务对语言模型进行联合预训练,从而得到表格预训练语言模型。

[0008] 可选地,若所述预训练任务为行预训练任务,

[0009] 则根据所述表格及其对应的文本构建预训练任务,包括:

[0010] 根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在行的行号,从而构建得到有监督的行预训练任务。

[0011] 可选地,若所述预训练任务为列预训练任务,

[0012] 则根据所述表格及其对应的文本构建预训练任务,包括:

[0013] 根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在列的行号,从而构建得到有监督的列预训练任务。

[0014] 可选地,若所述预训练任务为文本遮罩实体预训练任务,

[0015] 则根据所述表格及其对应的文本构建预训练任务,包括:

[0016] 将所述表格中的列名和所述列名对应的单元格内容作为实体;

[0017] 按照第一预设遮罩比例对出现在所述文本中的实体进行随机遮罩,从而构建得到

无监督的文本遮罩实体预训练任务。

[0018] 可选地,若所述预训练任务为文本随机遮罩字预训练任务,

[0019] 则根据所述表格及其对应的文本构建预训练任务,包括:

[0020] 按照第二预设遮罩比例对所述文本中的字进行随机遮罩,从而构建得到无监督的文本随机遮罩字预训练任务。

[0021] 可选地,所述第一预设遮罩比例为10-25%;

[0022] 和/或,

[0023] 所述第二预设遮罩比例为10-25%。

[0024] 可选地,所述语言模型包括依次串联的输入层、嵌入层、编码层和输出层;其中,所述输出层包括多个并联的子输出层,所述子输出层的数量与所述预训练任务的数量相同。

[0025] 另外,根据本发明实施例的另一个方面,提供了一种表格预训练装置,包括:

[0026] 构建模块,用于获取表格及其对应的文本,根据所述表格及其对应的文本构建预训练任务;其中,所述预训练任务选自以下至少两种:行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务;

[0027] 训练模块,用于根据所述预训练任务对语言模型进行联合预训练,从而得到表格预训练语言模型。

[0028] 可选地,若所述预训练任务为行预训练任务,

[0029] 则所述构建模块还用于:

[0030] 根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在行的行号,从而构建得到有监督的行预训练任务。

[0031] 可选地,若所述预训练任务为列预训练任务,

[0032] 则所述构建模块还用于:

[0033] 根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在列的行号,从而构建得到有监督的列预训练任务。

[0034] 可选地,若所述预训练任务为文本遮罩实体预训练任务,

[0035] 则所述构建模块还用于:

[0036] 将所述表格中的列名和所述列名对应的单元格内容作为实体;

[0037] 按照第一预设遮罩比例对出现在所述文本中的实体进行随机遮罩,从而构建得到无监督的文本遮罩实体预训练任务。

[0038] 可选地,若所述预训练任务为文本随机遮罩字预训练任务,

[0039] 则所述构建模块还用于:

[0040] 按照第二预设遮罩比例对所述文本中的字进行随机遮罩,从而构建得到无监督的文本随机遮罩字预训练任务。

[0041] 可选地,所述第一预设遮罩比例为10-25%;

[0042] 和/或,

[0043] 所述第二预设遮罩比例为10-25%。

[0044] 可选地,所述语言模型包括依次串联的输入层、嵌入层、编码层和输出层;其中,所述输出层包括多个并联的子输出层,所述子输出层的数量与所述预训练任务的数量相同。

[0045] 根据本发明实施例的另一个方面,还提供了一种电子设备,包括:

[0046] 一个或多个处理器；

[0047] 存储装置，用于存储一个或多个程序，

[0048] 当所述一个或多个程序被所述一个或多个处理器执行时，所述一个或多个处理器实现上述任一实施例所述的方法。

[0049] 根据本发明实施例的另一个方面，还提供了一种计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现上述任一实施例所述的方法。

[0050] 根据本发明实施例的另一个方面，还提供了一种计算机程序产品，包括计算机程序，所述计算机程序被处理器执行时实现上述任一实施例所述的方法。

[0051] 上述发明中的一个实施例具有如下优点或有益效果：因为采用构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的至少两种预训练任务，并根据预训练任务对语言模型进行联合预训练的技术手段，所以克服了现有技术中缺乏对文本与表格进行深度结构化语义交互的建模的技术问题。本发明实施例通过构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的至少两种预训练任务，强化文本与表格的深度结构化语义交互，从而得到高质量的表格预训练语言模型。

[0052] 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具体实施方式加以说明。

附图说明

[0053] 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。其中：

[0054] 图1是根据本发明实施例的表格预训练方法的主要流程的示意图；

[0055] 图2是根据本发明实施例的语言模型的结构示意图；

[0056] 图3是根据本发明一个可参考实施例的表格预训练方法的主要流程的示意图；

[0057] 图4是根据本发明实施例的表格预训练装置的主要模块的示意图；

[0058] 图5是本发明实施例可以应用于其中的示例性系统架构图；

[0059] 图6是适于用来实现本发明实施例的终端设备或服务器的计算机系统的结构示意图。

具体实施方式

[0060] 以下结合附图对本发明的示范性实施例做出说明，其中包括本发明实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本发明的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0061] 本发明实施例主要解决现有表格预训练方法中，缺乏对文本与表格进行深度结构化语义交互的建模，出现编码文本与表格时语义不一致问题，从而导致表格预训练质量较低，影响下游任务的使用。

[0062] 本发明实施例提供的表格预训练方法基于深度结构化语义交互进行建模,在通过文本-表格对进行表格预训练时,构建了多个文本到表格不同角度的结构化语义交互的预训练任务,用以训练一个高质量的表格预训练语言模型,用来更好的编码文本和表格,避免二者语义不一致的出现,从而更好供下游任务使用。

[0063] 图1是根据本发明实施例的表格预训练方法的主要流程的示意图。作为本发明的一个实施例,如图1所示,所述表格预训练方法可以包括:

[0064] 步骤101,获取表格及其对应的文本,根据所述表格及其对应的文本构建预训练任务;其中,所述预训练任务选自以下至少两种:行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务。

[0065] 首先,获取表格及其对应的文本,也就是文本-表格对,然后根据文本-表格对构建多种预训练任务,其中所述预训练任务选自行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的任意两种、任意三种或者四种。

[0066] 例如,根据文本-表格对构建行预训练任务和列预训练任务;根据文本-表格对构建文本遮罩实体预训练任务和文本随机遮罩字预训练任务;根据文本-表格对构建行预训练任务和文本遮罩实体预训练任务;根据文本-表格对构建列预训练任务和文本遮罩实体预训练任务;根据文本-表格对构建行预训练任务、列预训练任务和文本遮罩实体预训练任务;根据文本-表格对构建行预训练任务、列预训练任务和文本随机遮罩字预训练任务;根据文本-表格对构建列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务;根据文本-表格对构建行预训练任务、列预训练任务和文本随机遮罩字预训练任务;根据文本-表格对构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务。可以根据实际需要构建多种预训练任务,本发明实施例对此不作限制。

[0067] 可选地,若所述预训练任务为行预训练任务,则根据所述表格及其对应的文本构建预训练任务,包括:根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在行的行号,从而构建得到有监督的行预训练任务。行预训练任务为有监督的预训练任务,因此需要标记出标签,也就是目标词在表格中的行号。例如,表格和文本中都出现了目标词【天气】,则在表格中查找目标词【天气】所在行的行号,并将其标记出来,将该行号作为预训练任务的标签。需要指出的是,如果在多个行中都出现了目标词,那么将这些行的行号都标记出来,都作为该预训练任务的标签。

[0068] 可选地,若所述预训练任务为列预训练任务,则根据所述表格及其对应的文本构建预训练任务,包括:根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在列的行号,从而构建得到有监督的列预训练任务。列预训练任务为有监督的预训练任务,因此需要标记出标签,也就是目标词在表格中的列号。例如,表格和文本中都出现了目标词【天气】,则在表格中查找目标词【天气】所在列的列号,并将其标记出来,将该列号作为预训练任务的标签。需要指出的是,如果在多个列中都出现了目标词,那么将这些列的列号都标记出来,都作为该预训练任务的标签。

[0069] 可选地,若所述预训练任务为文本遮罩实体预训练任务,则根据所述表格及其对应的文本构建预训练任务,包括:将所述表格中的列名和所述列名对应的单元格内容作为实体;按照第一预设遮罩比例对出现在所述文本中的实体进行随机遮罩,从而构建得到无监督的文本遮罩实体预训练任务。文本遮罩实体预训练任务为无监督的预训练任务,首先

将表格中的列名和各个列名对应的单元格内容作为实体,比如,列名为【姓名】,该列名对应的单元格内容包括【张三】、【李四】、【王五】,那么【姓名】、【张三】、【李四】、【王五】均为实体;从表格中识别出各个实体后,按照预设的遮罩比例对出现在文本中的实体进行随机遮罩,比如遮罩【张三】,从而构建得到无监督的文本遮罩实体预训练任务。

[0070] 可选地,所述第一预设遮罩比例为10-25%,按照10-25%进行遮罩可以提高模型的训练效果。具体地,可以按照10%、12%、15%、18%或者20%进行随机遮罩。

[0071] 可选地,若所述预训练任务为文本随机遮罩字预训练任务,则根据所述表格及其对应的文本构建预训练任务,包括:按照第二预设遮罩比例对所述文本中的字进行随机遮罩,从而构建得到无监督的文本随机遮罩字预训练任务。文本随机遮罩字预训练任务为无监督的预训练任务,按照预设的遮罩比例对文本中的字进行随机遮罩,可以遮罩一个字也可以遮罩多个字,而且多个字之间可以有间隔(非连续遮罩)也可以没有间隔(连续遮罩),因此遮罩的结果可能是单个字,也可能是词语,还有可能包括单个字和词语。

[0072] 可选地,所述第二预设遮罩比例为10-25%,按照10-25%进行遮罩可以提高模型的训练效果。具体地,可以按照10%、11%、13%、15%或者20%进行随机遮罩。

[0073] 步骤102,根据所述预训练任务对语言模型进行联合预训练,从而得到表格预训练语言模型。

[0074] 通过步骤101构建好多种预训练任务之后,根据这些预训练任务对语言模型进行联合预训练,从而训练得到表格预训练语言模型。通过步骤101构建的多种预训练任务进行联合预训练,本发明实施例深度建模了文本与表格之间的结构化关系,可以训练得到一个更高质量的表格预训练语言模型,避免文本、表格编码不一致问题。

[0075] 对于行预训练任务:预测在对应表格中,哪些行出现在了文本中。通过此任务可以强化表格预训练语言模型建立文本与表格的行之间的联系,从而加强文本与表格的深度结构化语义交互。

[0076] 对于列预训练任务:预测在对应表格中,哪些列出现在了文本中。通过此任务可以强化表格预训练语言模型建立文本与表格的列之间的联系,从而加强文本与表格的深度结构化语义交互。

[0077] 对于文本遮罩实体预训练任务:用文本中未被遮罩的字来预测遮罩的实体。通过此任务,既可以很好地训练文本本身的语言模型,同时也可以通过对遮罩实体的预测,使文本与表格之间的联系更紧密。

[0078] 对于文本随机遮罩字预训练任务:用文本中未被遮罩的字来预测遮罩的字。通过此任务,可以很好的训练文本本身的语言模型,使文本的表达能力更强。

[0079] 可选地,所述语言模型包括依次串联的输入层、嵌入层、编码层和输出层;其中,所述输出层包括多个并联的子输出层,所述子输出层的数量与所述预训练任务的数量相同。如图2所示,通过输入层将文本-表格对输入到语言模型中,通过嵌入层进行嵌入表示后,通过编码层进行编码,最终通过输出层输出预训练任务;以行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务为例,第一输出层输出行预训练任务,第二输出层输出列预训练任务,第三输出层输出文本遮罩实体预训练任务,第四输出层输出文本随机遮罩字预训练任务。

[0080] 根据上面所述的各种实施例,可以看出本发明实施例通过构建行预训练任务、列

预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的至少两种预训练任务,并根据预训练任务对语言模型进行联合预训练的技术手段,解决了现有技术中缺乏对文本与表格进行深度结构化语义交互的建模的技术问题。本发明实施例通过构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的至少两种预训练任务,强化文本与表格的深度结构化语义交互,从而得到高质量的表格预训练语言模型。

[0081] 图3是根据本发明一个可参考实施例的表格预训练方法的主要流程的示意图。作为本发明的又一个实施例,如图3所示,所述表格预训练方法可以包括:

[0082] 步骤301,获取表格及其对应的文本。

[0083] 可以从数据集中获取文本-表格对。例如,文本:张三是个男生,来自北京市海淀区,是一名公务员。对应的表格如下:

[0084] 表1文本对应的表格

[0085]

列序号	0	1	2	3	4
行序号	姓名	性别	年龄	身高	体重
0	张三	男	24	182	80
1	李四	女	19	164	51
2	王五	男	28	187	85

[0086] 步骤302,根据所述表格及其对应的文本构建四种预训练任务;其中,所述预训练任务包括行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务。

[0087] 在该实施例中,构建四种预训练任务,分别为行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务。以步骤301中的文本-表格对为例,分别构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务。

[0088] 行预训练任务:是有监督的行预训练任务,需要确定同时出现在表格和文本中的目标词,然后在表格中标记出目标词所在行的行号,也就是表1中的行序号0。该预训练任务就是通过预测当前文本涉及表格的第几行,来训练表格预训练语言模型。

[0089] 列预训练任务:是有监督的列预训练任务,需要确定同时出现在表格和文本中的目标词,然后在表格中标记出目标词所在列的列号,也就是表1中的列序号0和1。该预训练任务就是通过预测当前文本涉及表格的第几列,来训练表格预训练语言模型。

[0090] 文本遮罩实体预训练任务:是无监督的列预训练任务,首先将表格中的列名和各个列名对应的单元格内容作为实体,也就是表1中的【姓名】、【张三】、【性别】、【男】,以预设的遮罩比例对这些实体进行随机遮罩,比如【张三】被遮罩。该预训练任务用文本中未被遮罩的字来预测遮罩的实体,即预测被遮罩的【张三】的过程中,训练了表格预训练语言模型。

[0091] 文本随机遮罩字预训练任务:是无监督的列预训练任务,对文本中的字进行随机遮罩,用文本中未被遮罩的字来预测遮罩的字。遮罩示例如:张三是个男生,来自<北京市>海淀区,是一名<公务员>,这里【北京市】和【公务员】是被随机遮罩的,预测他们的过程中,训练了表格预训练语言模型。

[0092] 步骤303,根据所述预训练任务对语言模型进行联合预训练,从而得到表格预训练

语言模型。

[0093] 通过步骤302构建好四种预训练任务之后,根据这四种预训练任务对语言模型进行联合预训练,语言模型的结构如图2所示,从而训练得到表格预训练语言模型。

[0094] 对于行预训练任务:预测在对应表格中,哪些行出现在了文本中。通过此任务可以强化表格预训练语言模型建立文本与表格的行之间的联系,从而加强文本与表格的深度结构化语义交互。

[0095] 对于列预训练任务:预测在对应表格中,哪些列出现在了文本中。通过此任务可以强化表格预训练语言模型建立文本与表格的列之间的联系,从而加强文本与表格的深度结构化语义交互。

[0096] 对于文本遮罩实体预训练任务:用文本中未被遮罩的字来预测遮罩的实体。通过此任务,既可以很好地训练文本本身的语言模型,同时也可以通过对遮罩实体的预测,使文本与表格之间的联系更紧密。

[0097] 对于文本随机遮罩字预训练任务:用文本中未被遮罩的字来预测遮罩的字。通过此任务,可以很好的训练文本本身的语言模型,使文本的表达能力更强。

[0098] 另外,在本发明一个可参考实施例中表格预训练方法的具体实施内容,在上面所述表格预训练方法中已经详细说明了,故在此重复内容不再说明。

[0099] 图4是根据本发明实施例的表格预训练装置的主要模块的示意图。如图4所示,所述表格预训练装置400包括构建模块401和训练模块402;其中,构建模块401用于获取表格及其对应的文本,根据所述表格及其对应的文本构建预训练任务;其中,所述预训练任务选自以下至少两种:行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务;训练模块402用于根据所述预训练任务对语言模型进行联合预训练,从而得到表格预训练语言模型。

[0100] 可选地,若所述预训练任务为行预训练任务,

[0101] 则所述构建模块401还用于:

[0102] 根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在行的行号,从而构建得到有监督的行预训练任务。

[0103] 可选地,若所述预训练任务为列预训练任务,

[0104] 则所述构建模块401还用于:

[0105] 根据所述表格及其对应的文本,确定同时出现在所述表格和所述文本中的目标词,在所述表格中标记出所述目标词所在列的行号,从而构建得到有监督的列预训练任务。

[0106] 可选地,若所述预训练任务为文本遮罩实体预训练任务,

[0107] 则所述构建模块401还用于:

[0108] 将所述表格中的列名和所述列名对应的单元格内容作为实体;

[0109] 按照第一预设遮罩比例对出现在所述文本中的实体进行随机遮罩,从而构建得到无监督的文本遮罩实体预训练任务。

[0110] 可选地,若所述预训练任务为文本随机遮罩字预训练任务,

[0111] 则所述构建模块401还用于:

[0112] 按照第二预设遮罩比例对所述文本中的字进行随机遮罩,从而构建得到无监督的文本随机遮罩字预训练任务。

[0113] 可选地,所述第一预设遮罩比例为10-25%;

[0114] 和/或,

[0115] 所述第二预设遮罩比例为10-25%。

[0116] 可选地,所述语言模型包括依次串联的输入层、嵌入层、编码层和输出层;其中,所述输出层包括多个并联的子输出层,所述子输出层的数量与所述预训练任务的数量相同。

[0117] 根据上面所述的各种实施例,可以看出本发明实施例通过构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的至少两种预训练任务,并根据预训练任务对语言模型进行联合预训练的技术手段,解决了现有技术中缺乏对文本与表格进行深度结构化语义交互的建模的技术问题。本发明实施例通过构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的至少两种预训练任务,强化文本与表格的深度结构化语义交互,从而得到高质量的表格预训练语言模型。

[0118] 需要说明的是,在本发明所述表格预训练装置的具体实施内容,在上面所述表格预训练方法中已经详细说明了,故在此重复内容不再说明。

[0119] 图5示出了可以应用本发明实施例的表格预训练方法或表格预训练装置的示例性系统架构500。

[0120] 如图5所示,系统架构500可以包括终端设备501、502、503,网络504和服务器505。网络504用以在终端设备501、502、503和服务器505之间提供通信链路的介质。网络504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。

[0121] 用户可以使用终端设备501、502、503通过网络504与服务器505交互,以接收或发送消息等。终端设备501、502、503上可以安装有各种通讯客户端应用,例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等(仅为示例)。

[0122] 终端设备501、502、503可以是具有显示屏并且支持网页浏览的各种电子设备,包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0123] 服务器505可以是提供各种服务的服务器,例如对用户利用终端设备501、502、503所浏览的购物类网站提供支持的后台管理服务器(仅为示例)。后台管理服务器可以对接收到的物品信息查询请求等数据进行分析等处理,并将处理结果反馈给终端设备。

[0124] 需要说明的是,本发明实施例所提供的表格预训练方法一般由服务器505执行,相应地,所述表格预训练装置一般设置在服务器505中。本发明实施例所提供的表格预训练方法也可以由终端设备501、502、503执行,相应地,所述表格预训练装置可以设置在终端设备501、502、503中。

[0125] 应该理解,图5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。

[0126] 下面参考图6,其示出了适于用来实现本发明实施例的终端设备的计算机系统600的结构示意图。图6示出的终端设备仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0127] 如图6所示,计算机系统600包括中央处理单元(CPU)601,其可以根据存储在只读存储器(ROM)602中的程序或者从存储部分608加载到随机访问存储器(RAM)603中的程序而执行各种适当的动作和处理。在RAM 603中,还存储有系统600操作所需的各种程序和数据。

CPU 601、ROM 602以及RAM603通过总线604彼此相连。输入/输出 (I/O) 接口605也连接至总线604。

[0128] 以下部件连接至I/O接口605:包括键盘、鼠标等的输入部分606;包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分607;包括硬盘等的存储部分608;以及包括诸如LAN卡、调制解调器等网络接口卡的通信部分609。通信部分609经由诸如因特网的网络执行通信处理。驱动器610也根据需要连接至I/O接口605。可拆卸介质611,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器610上,以便于从其上读出的计算机程序根据需要被安装入存储部分608。

[0129] 特别地,根据本发明公开的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本发明公开的实施例包括一种计算机程序,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分609从网络上被下载和安装,和/或从可拆卸介质611被安装。在该计算机程序被中央处理单元(CPU) 601执行时,执行本发明的系统中限定的上述功能。

[0130] 需要说明的是,本发明所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是一——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。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本发明中,计算机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。而在本发明中,计算机可读的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,其中承载了计算机可读的程序代码。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。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、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。计算机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、电线、光缆、RF等等,或者上述的任意合适的组合。

[0131] 附图中的流程图和框图,图示了按照本发明各种实施例的系统、方法和计算机程序的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0132] 描述于本发明实施例中所涉及到的模块可以通过软件的方式实现,也可以通过硬

件的方式来实现。所描述的模块也可以设置在处理器中,例如,可以描述为:一种处理器包括构建模块和训练模块,其中,这些模块的名称在某种情况下并不构成对该模块本身的限定。

[0133] 作为另一方面,本发明还提供了一种计算机可读介质,该计算机可读介质可以是上述实施例中描述的设备中所包含的;也可以是单独存在,而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被一个该设备执行时,该设备实现如下方法:获取表格及其对应的文本,根据所述表格及其对应的文本构建预训练任务;其中,所述预训练任务选自以下至少两种:行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务;根据所述预训练任务对语言模型进行联合预训练,从而得到表格预训练语言模型。

[0134] 作为另一方面,本发明实施例还提供了一种计算机程序产品,包括计算机程序,所述计算机程序被处理器执行时实现上述任一实施例所述的方法。

[0135] 根据本发明实施例的技术方案,因为采用构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的至少两种预训练任务,并根据预训练任务对语言模型进行联合预训练的技术手段,所以克服了现有技术中缺乏对文本与表格进行深度结构化语义交互的建模的技术问题。本发明实施例通过构建行预训练任务、列预训练任务、文本遮罩实体预训练任务和文本随机遮罩字预训练任务中的至少两种预训练任务,强化文本与表格的深度结构化语义交互,从而得到高质量的表格预训练语言模型。

[0136] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明白的是,取决于设计要求和因素,可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明保护范围之内。

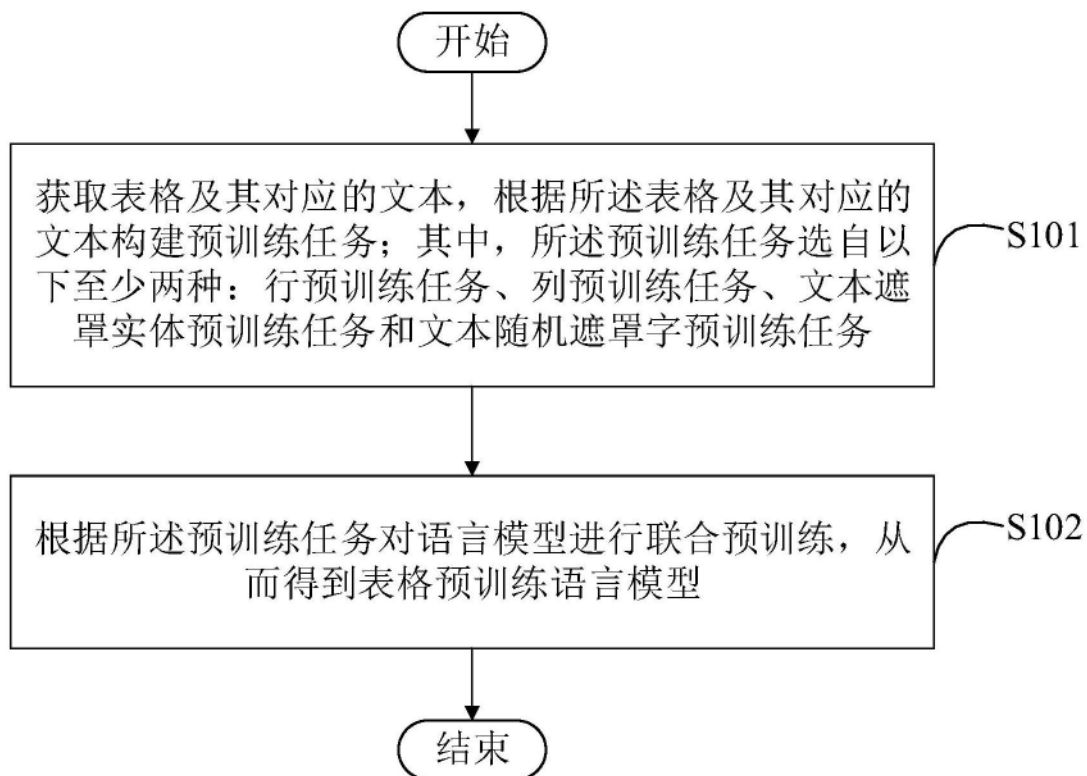


图1

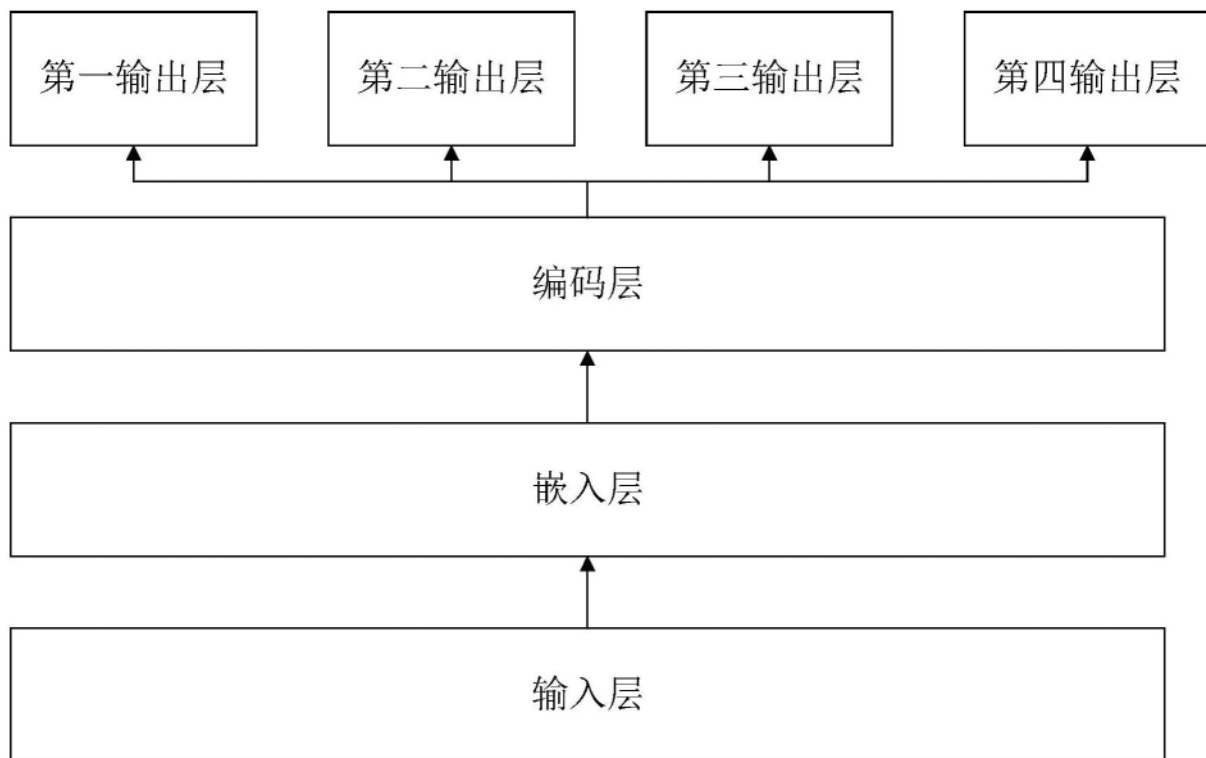


图2

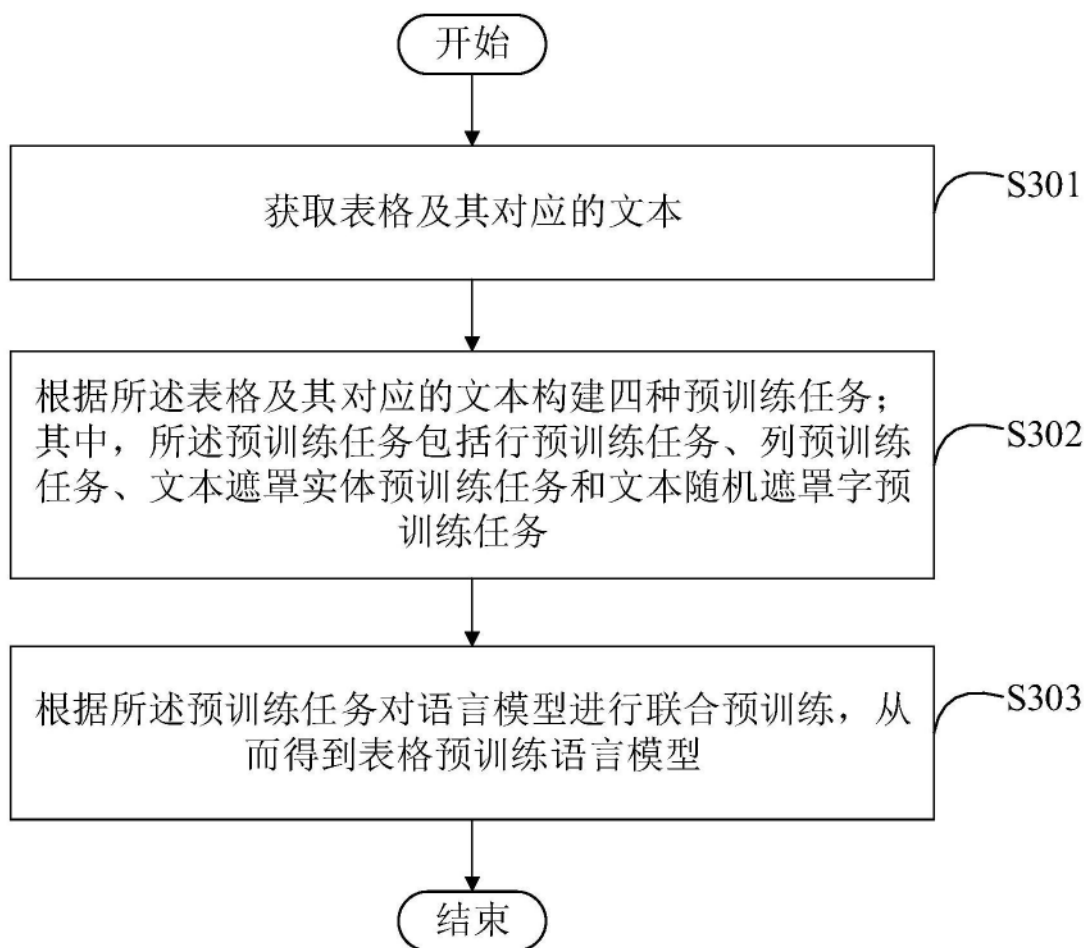


图3

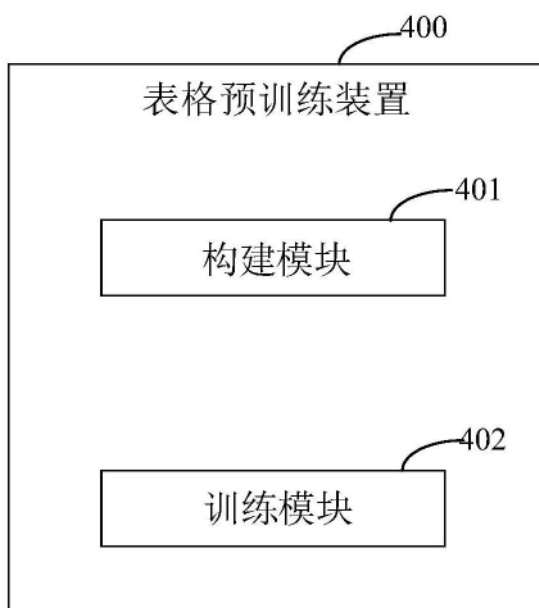


图4

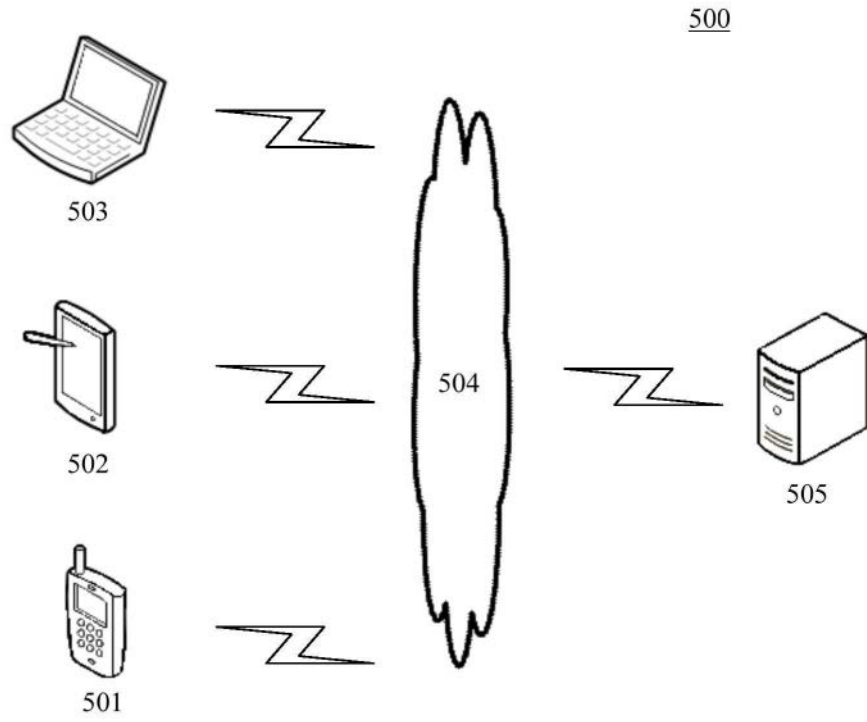


图5

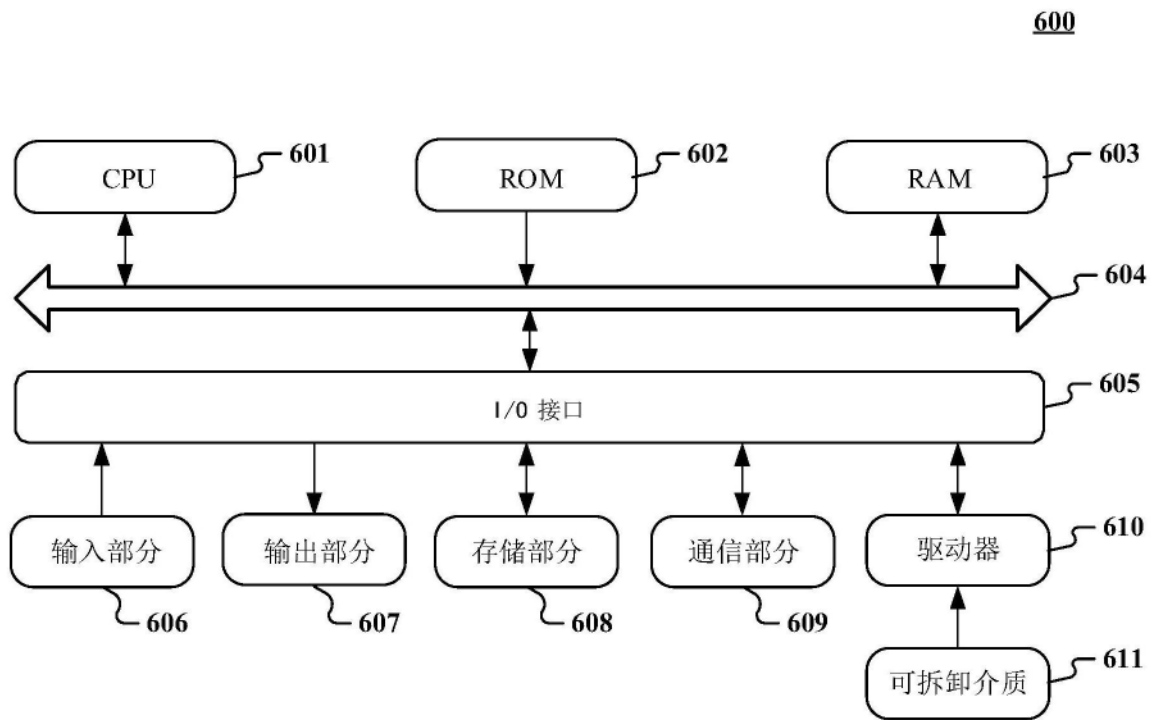


图6