



(21) 申请号 202311467677.1

(22) 申请日 2023.11.06

(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 张振羽 宋阳 陈蒙

(74) 专利代理机构 北京世辉律师事务所 16093

专利代理师 罗利娜

(51) Int. Cl.

G06F 40/126 (2020.01)

G06F 40/35 (2020.01)

G06N 3/0499 (2023.01)

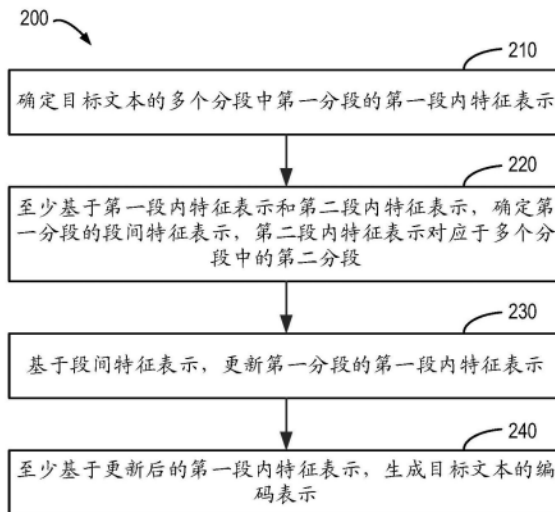
权利要求书2页 说明书12页 附图3页

(54) 发明名称

用于编码的方法、装置、设备和可读介质

(57) 摘要

本公开的实施例提供用于编码的方法、装置、设备和可读介质。该方法包括：确定目标文本的多个分段中第一分段的第一段内特征表示；至少基于第一段内特征表示和第二段内特征表示，确定第一分段的段间特征表示，第二段内特征表示对应于多个分段中的第二分段；基于段间特征表示，更新第一段的第一段内特征表示；以及至少基于更新后的第一段内特征表示，生成目标文本的编码表示。以此方式，可以通过结合文本的各个分段的段内特征表示和段间特征表示来充分考虑文本中的结构信息和语义信息，能够强化局部信息和全局信息交互，有助于提升编码效果。此外，还可以通过同时编码多个分段来提高编码的效率，能够有效编码长文本。



1. 一种用于编码的方法,包括:
确定目标文本的多个分段中第一分段的第一段内特征表示;
至少基于第一段内特征表示和第二段内特征表示,确定所述第一分段的段间特征表示,所述第二段内特征表示对应于所述多个分段中的第二分段;
基于所述段间特征表示,更新所述第一分段的所述第一段内特征表示;以及
至少基于更新后的所述第一段内特征表示,生成所述目标文本的编码表示。
2. 根据权利要求1所述的方法,还包括:
基于预设长度对所述目标文本进行划分,以确定所述目标文本的所述多个分段。
3. 根据权利要求1所述的方法,其中确定目标文本的多个分段中第一分段的第一段内特征表示包括:
确定所述第一分段中的一组词;
确定与所述一组词相关联的一组词特征表示;以及
基于所述一组词特征表示,确定所述第一分段的所述第一段内特征表示。
4. 根据权利要求3所述的方法,其中基于所述一组词特征表示确定所述第一分段的所述第一段内特征表示包括:
通过聚合所述一组词特征表示,确定第一聚合结果;以及
基于所述第一聚合结果和所述一组词特征表示,确定所述第一分段的所述第一段内特征表示。
5. 根据权利要求4所述的方法,其中基于所述第一聚合结果和所述一组词特征表示确定所述第一分段的所述第一段内特征表示包括:
基于所述第一聚合结果和所述一组词特征表示,确定与所述一组词特征表示对应的第一组权重,所述第一组权重指示所述第一聚合结果和所述一组词特征表示之间的注意力;以及
通过将所述第一组权重应用于所述一组词特征表示,确定所述第一分段的所述第一段内特征表示。
6. 根据权利要求3所述的方法,其中基于所述段间特征表示,更新所述第一分段的所述第一段内特征表示包括:
基于所述段间特征表示和所述一组词特征表示,更新所述第一分段的所述第一段内特征表示。
7. 根据权利要求6所述的方法,其中基于所述段间特征表示和所述一组词特征表示更新所述第一分段的所述第一段内特征表示包括:
基于所述段间特征表示和所述一组词特征表示,确定与所述段间特征表示对应的第二组权重,所述第二组权重指示所述段间特征表示和所述一组词特征表示之间的注意力;
通过将所述第二组权重应用于所述段间特征表示,确定所述段间特征表示的加权结果;
通过计算所述加权结果和所述一组词特征表示,更新所述第一分段的所述一组词特征表示;以及
基于更新后的一组词特征表示,确定更新后的第一段内特征表示。
8. 根据权利要求1所述的方法,其中至少基于更新后的所述第一段内特征表示,生成所

述目标文本的编码表示包括：

通过聚合至少所述第一分段的所述更新后的第一段内特征表示和所述第二分段的更新后的第二段内特征表示，确定所述目标文本对应的第二聚合结果；以及

至少基于所述第二聚合结果，确定所述目标文本的所述编码表示。

9. 根据权利要求1所述的方法，其中所述目标文本为对话文本。

10. 一种用于编码的装置，包括：

第一确定模块，被配置为确定目标文本的多个分段中第一分段的第一段内特征表示；

第二确定模块，被配置为至少基于第一段内特征表示和第二段内特征表示，确定所述第一分段的段间特征表示，所述第二段内特征表示对应于所述多个分段中的第二段；

更新模块，被配置为基于所述段间特征表示，更新所述第一分段的所述第一段内特征表示；以及

编码表示生成模块，被配置为至少基于更新后的所述第一段内特征表示，生成所述目标文本的编码表示。

11. 一种电子设备，包括：

至少一个处理单元；以及

至少一个存储器，所述至少一个存储器被耦合到所述至少一个处理单元并且存储用于由所述至少一个处理单元执行的指令，所述指令在由所述至少一个处理单元执行时使所述电子设备执行根据权利要求1至9任一项所述的方法。

12. 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序可由处理器执行以实现根据权利要求1至9任一项所述的方法。

用于编码的方法、装置、设备和可读介质

技术领域

[0001] 本公开的示例实施例总体涉及计算机技术领域,并且更具体地,涉及用于编码的方法、装置、设备和计算机可读存储介质。

背景技术

[0002] 随着信息技术的飞速发展,越来越多的应用或平台等均提供问答功能,给广大用户带来了诸多便利。具有问答功能的应用或平台可以基于智能对话机器人(Intelligent Robot)向用户提供问答服务。在进行对话中,智能对话机器人首先需要对当前对话进行理解,然后基于理解的对话状态采取对应的应答策略。由于对话理解的核心基础是对对话进行编码,并基于编码表示进行状态分类、策略预测等,因此,人们期望能够提高对对话进行编码的效果。

发明内容

[0003] 在本公开的第一方面,提供一种用于编码的方法。该方法包括:确定目标文本的多个分段中第一分段的第一段内特征表示;至少基于第一段内特征表示和第二段内特征表示,确定第一分段的段间特征表示,第二段内特征表示对应于多个分段中的第二分段;基于段间特征表示,更新第一分段的第一段内特征表示;以及至少基于更新后的第一段内特征表示,生成目标文本的编码表示。

[0004] 在本公开的第二方面,提供一种用于编码的装置。该装置包括:第一确定模块,被配置为确定目标文本的多个分段中第一分段的第一段内特征表示;第二确定模块,被配置为至少基于第一段内特征表示和第二段内特征表示,确定第一分段的段间特征表示,第二段内特征表示对应于多个分段中的第二分段;更新模块,被配置为基于段间特征表示,更新第一分段的第一段内特征表示;以及编码表示生成模块,被配置为至少基于更新后的第一段内特征表示,生成目标文本的编码表示。

[0005] 在本公开的第三方面,提供了一种电子设备。该电子设备包括至少一个处理单元;以及至少一个存储器,至少一个存储器被耦合到至少一个处理单元并且存储用于由至少一个处理单元执行的指令,指令在由至少一个处理单元执行时使电子设备执行本公开第一方面的方法。

[0006] 在本公开的第四方面,提供了一种计算机可读存储介质。该计算机可读存储介质上存储有计算机程序,其可由处理器执行以执行根据本公开的第一方面的方法。

[0007] 应当理解,发明内容部分中所描述的内容并非旨在限定本公开的实施例的关键特征或重要特征,也不用于限制本公开的范围。本公开的其它特征将通过以下的描述而变得容易理解。

附图说明

[0008] 在下文中,结合附图并参考以下详细说明,本公开各实现方式的上述和其他特征、

优点及方面将变得更加明显。在附图中,相同或相似的附图标记表示相同或相似的元素,其中:

- [0009] 图1示出了本公开的实施例能够在其中实现的示例环境的示意图;
- [0010] 图2示出了根据本公开的一些实施例的用于编码的过程的流程图;
- [0011] 图3示出了根据本公开的一些实施例的编码器的示例架构的示意图;
- [0012] 图4示出了根据本公开的一些实施例的用于编码的装置的示意性结构框图;以及
- [0013] 图5示出了其中可以实施本公开的一个或多个实施例的计算设备的框图。

具体实施方式

[0014] 下面将参照附图更详细地描述本公开的实施例。虽然附图中示出了本公开的一些实施例,然而应当理解的是,本公开可以通过各种形式来实现,而且不应该被解释为限于这里阐述的实施例,相反,提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是,本公开的附图及实施例仅用于示例性作用,并非用于限制本公开的保护范围。

[0015] 在本公开的实施例的描述中,术语“包括”及其类似用语应当理解为开放性包含,即“包括但不限于”。术语“基于”应当理解为“至少部分地基于”。术语“一个实施例”或“该实施例”应当理解为“至少一个实施例”。术语“一些实施例”应当理解为“至少一些实施例”。下文还可能包括其它明确的和隐含的定义。

[0016] 如本文中所使用的,术语“模型”可以从训练数据中学习到相应的输入与输出之间的关联,从而在训练完成后可以针对给定的输入,生成对应的输出。模型的生成可以基于机器学习技术。深度学习是一种机器学习算法,通过使用多层处理单元来处理输入和提供相应输出。在本文中,“模型”也可以被称为“机器学习模型”、“机器学习网络”、“神经网络”或“网络”,这些术语在本文中可互换地使用。

[0017] “神经网络”是一种基于深度学习的机器学习网络。神经网络能够处理输入并且提供相应输出,其通常包括输入层和输出层以及在输入层与输出层之间的一个或多个隐藏层。在深度学习应用中使用的神经网络通常包括许多隐藏层,从而增加网络的深度。神经网络的各个层按顺序相连,从而前一层的输出被提供作为后一层的输入,其中输入层接收神经网络的输入,而输出层的输出作为神经网络的最终输出。神经网络的每个层包括一个或多个节点(也称为处理节点或神经元),每个节点处理来自上一层的输入。

[0018] 通常,机器学习大致可以包括三个阶段,即训练阶段、测试阶段和应用阶段(也称为推理阶段)。在训练阶段,给定的模型可以使用大量的训练数据进行训练,不断迭代更新参数值,直到模型能够从训练数据中获得一致的满足预期目标的推理。通过训练,模型可以被认为能够从训练数据中学习从输入到输出之间的关联(也称为输入到输出的映射)。训练后的模型参数值被确定。在测试阶段,将测试输入应用到训练后的模型,测试模型是否能够提供正确的输出,从而确定模型的性能。在应用阶段,模型可以被用于基于训练得到的参数值,对实际的输入进行处理,确定对应的输出。

[0019] 需要说明的是,本公开的技术方案中,所涉及的用户个人信息的获取,存储和应用等,均符合相关法律法规的规定,且不违背公序良俗。

[0020] 可以理解的是,在使用本公开各实施例公开的技术方案之前,均应当根据相关法律法规通过适当的方式对本公开所涉及个人信息的类型、使用范围、使用场景等告知用户

并获得用户的授权。

[0021] 例如,在响应于接收到用户的主动请求时,向用户发送提示信息,以明确地提示用户,其请求执行的操作将需要获得和使用到用户的个人信息,从而使得用户可以根据提示信息来自主地选择是否向执行本公开技术方案的操作的电子设备、应用程序、服务器或存储介质等软件或硬件提供个人信息。

[0022] 作为一种可选的但非限制性的实现方式,响应于接收到用户的主动请求,向用户发送提示信息的方式,例如可以是弹出窗口的方式,弹出窗口中可以以文字的方式呈现提示信息。此外,弹出窗口中还可以承载供用户选择“同意”或“不同意”向电子设备提供个人信息的选择控件。

[0023] 如上文所讨论的,具有问答功能的应用或平台可以基于智能对话机器人向用户提供问答服务。在进行对话中,智能对话机器人首先需要对当前对话进行理解,然后基于理解的对话状态采取对应的应答策略。对话理解的核心基础是对对话进行编码,并基于编码表示进行状态分类、策略预测等。传统的对话编码方法主要分为四类:(1)拼接对话文本,然后基于语言模型编码器获取对话状态;(2)基于堆叠LSTM和transformer分别编码对话语句和对话整体状态;(3)基于对话编码状态缓存的方法,将对话分段编码缓存,利用循环模式递增编码;(4)基于稀疏注意力编码对话。

[0024] 对话具有不同于普通文本的特点,其一般具有更长的长度以及不同于普通文本的结构信息,而且普遍存在大量的全局信息和局部信息的交互。但是目前的方法在编码中没有很好综合考虑这些特点,不利于高效捕获对话信息。具体地,拼接对话文本的简单编码方法,完全没有考虑到对话的结构特点,因此其很难有效表征对话中轮次迭代过程、角色信息等。由于对话往往是一句句分开的,这种方式相当于把对话当作一整段文本进行编码,会丢失很多信息。此外,由于目前文本编码的主流模型都是类似BERT模型的编码器(encoder),其基于自注意力(SA)计算表征,其计算复杂度是输入对话长度 N 的 $O(N^2)$ 的复杂度,对于较长的对话很不友好。

[0025] 对于基于层次注意力的模型,往往首先使用长短期记忆(Long Short Term Memory,LSTM)网络或者循环神经(Gate Recurrent Unit,GRU)网络对每一句对话进行编码,然后将编码后的所有对话句子向量送入transformer网络或者另一个LSTM网络进行编码。虽然这种模型考虑到了对话中句子的结构特点,但是由于模型架构上没有直接让句子的单词表征进行充分的交互,会在处理一些指代消解、长距离细粒度上下文交互上出现信息丢失的问题。

[0026] 利用循环模式递增编码的方法首先会将输入的句子的编码的结果缓存起来,然后在输入下一句话时,会基于缓存的内容和当前句子的编码计算注意力,然后缓存当前句子的编码结果。通过不断重复这样的过程,不断循环迭代地编码所有的对话内容。因为这个编码过程需要进行多次循环迭代,会导致效率低下。此外,由于前面的缓存内容无法和后面的句子内容进行注意力计算获得编码状态的更新,导致缓存内容的上下文编码缺失对话的下文信息。这个方法虽然依靠缓存增加了编码对话的长度,但是却降低了对话编码的计算效率和上下文理解能力。

[0027] 基于稀疏注意力的方法主要用来优化目前最先进的基础网络结构transformer的平方复杂度的注意力计算过程,从而可大幅度提升编码的对话文本长度。但是,由于稀疏

注意力的效果不如标准的稠密注意力,所以这种编码方案在提升对话编码长度的过程中,会导致显著的效果下降,使得相关对话任务上准确率降低。综上所述,传统的对话编码方法存在准确性低、效率低的问题。

[0028] 有鉴于此,本公开的实施例提供了一种用于编码的方法。该方法包括:确定目标文本的多个分段中第一分段的第一段内特征表示。至少基于第一段内特征表示和第二段内特征表示,确定第一分段的段间特征表示,第二段内特征表示对应于多个分段中的第二段。基于段间特征表示,更新第一分段的第一段内特征表示。至少基于更新后的第一段内特征表示,生成目标文本的编码表示。以此方式,本公开的实施例可以通过结合文本的各个分段的段内特征表示和段间特征表示来充分考虑文本中的结构信息和语义信息,能够强化局部信息和全局信息交互,有助于提升编码效果。此外,还可以通过同时编码多个分段来提高编码的效率,能够有效编码长文本。

[0029] 图1示出了本公开的实施例能够在其中实现的示例环境100的示意图。如图1所示,环境100可以包括电子设备110。

[0030] 电子设备110可以确定目标文本105对应的编码表示115。目标文本105可以是存储在电子设备110本地的文本,也可以是电子设备110实时获取到的文本。目标文本105可以包括多个分段101(包括分段101-1、101-2、……、101-N,其中N是正整数)。这多个分段101可以是基于句子的数目被划分得到的。例如,每个分段均包括5个句子。这多个分段101还可以是基于单词的数目被划分得到的,例如,每个分段均包括50个单词。可以理解,可以通过任意适当的方式对目标文本105进行划分以得到多个分段101,本公开对此不做限定。

[0031] 电子设备110中安装有编码器112。这里的编码器112例如可以是机器学习模型中的编码器,例如transformer网络中的编码器encoder。电子设备110可以将目标文本105输入编码器112,并获取编码器112输出的编码表示115。在一些实施例中,电子设备110可以将目标文本105的多个分段101输入编码器112。这多个分段101可以是依次输入的,也可以是同时输入的。

[0032] 电子设备110可以包括具有计算能力的任何计算系统,例如各种计算设备/系统、终端设备、服务端设备等。终端设备可以是任意类型的移动终端、固定终端或便携式终端,包括移动手机、台式计算机、膝上型计算机、笔记本计算机、上网本计算机、平板计算机、媒体计算机、多媒体平板、掌上电脑、便携式游戏终端、VR/AR设备、个人通信系统(Personal Communication System,PCS)设备、个人导航设备、个人数字助理(Personal Digital Assistant,PDA)、音频/视频播放器、数码相机/摄像机、定位设备、电视接收器、无线电广播接收器、电子书设备、游戏设备或者前述各项的任意组合,包括这些设备的配件和外设或者其任意组合。服务端设备可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络、以及大数据和人工智能平台等基础云计算服务的云服务器。服务端设备例如可以包括计算系统/服务器,诸如大型机、边缘计算节点、云环境中的计算设备,等等。

[0033] 应当理解,仅出于示例性的目的描述环境100中各个元素的结构和功能,而不暗示对于本公开的范围的任何限制。

[0034] 以下将继续参考附图描述本公开的一些示例实施例。

[0035] 图2示出了根据本公开的一些实施例的用于编码的过程200的流程图。过程200可以被实现在电子设备110处。为便于讨论,将参考图1的环境100来描述过程200。

[0036] 在框210,电子设备110确定目标文本的多个分段中第一分段的第一段内特征表示。目标文本可以是任意格式的文本,包括但不限于doc、docx、txt等等。在一些实施例中,目标文本可以为对话文本。这样的对话文本可以是存储在电子设备110本地的文本,也可以是电子设备110实时获取到的文本。这里的第一分段可以是目标文本中的任意分段。

[0037] 在一些实施例中,在确定目标文本的多个分段中第一分段的第一段内特征表示之前,电子设备110需要先确定目标文本的多个分段。电子设备110可以基于预设长度对目标文本进行划分以确定目标文本的多个分段。这里的预设长度例如可以与单词数目相关联。例如,可以将预设长度设置为300,表示目标文本的每个分段均包含300个单词。在一些实施例中,目标文本的多个分段的长度相同。在这种情况下,在目标文本的长度(也即包含的单词数目)可并非预设长度的倍数的情况下,电子设备110还可以对长度不足预设长度的分段添加掩码以使得其达到预设长度。示例性的,如预设长度为300,目标文本为1000,则电子设备110对目标文本进行划分可以得到4个分段,并且其中第四个分段只包含100个单词。在这种情况下,电子设备110可以对第四个分段添加掩码,添加掩码后的第四个分段包含100个单词和200个掩码,此时第四个分段的长度同样为300。预设长度例如还可以与目标文本中的每一行文本相关联(例如每个分段包含5行文本)、与目标文本所包括的句子相关联(例如每个分段包含10个句子)等等,本公开对此不做限定。

[0038] 关于确定目标文本的多个分段中第一分段的第一段内特征表示,在一些实施例中,电子设备110可以利用编码器的分段内部自注意力(inner-segment self-attention)层来确定目标文本的多个分段中第一分段的第一段内特征表示。在一些实施例中,电子设备110可以将目标文本输入编码器的分段内部自注意力层,以便分段内部自注意力层确定第一分段中的一组词,并确定与这一组词相关联的一组词特征表示。这里的一组词即第一分段中所包括的所有单词,也可以被称之为分段所包括的隐含状态。示例性的,第一分段包含B个单词,也可以称之为包含B个隐含状态。相应的,一组词特征表示也可以被称之为隐含状态表征。备选地或者附加地,在一些实施例中,电子设备110还可以直接确定第一分段中的一组词以及与这一组词相关联的一组词特征表示。在这种情况下,电子设备110可以直接将第一分段的一组词特征表示提供给分段内部自注意力层。

[0039] 分段内部自注意力层进而可以基于这一组词特征表示确定第一分段的第一段内特征表示。具体地,分段内部自注意力层可以通过聚合这一组词特征表示来确定第一聚合结果,进而基于第一聚合结果和一组词特征表示,确定第一分段的第一段内特征表示。

[0040] 在一些实施例中,分段内部自注意力层可以基于第一聚合结果和一组词特征表示,确定与一组词特征表示对应的第一组权重。这里的第一组权重可以指示第一聚合结果和一组词特征表示之间的注意力。分段内部自注意力层例如可以基于如下公式来确定第一组权重:

$$[0041] \quad Agg(H_{inn_i}) = \frac{1}{\sum e^{M_j}} \sum_{j=1}^B H_{inn_{i,j}} \times e^{M_j} \quad (1)$$

$$[0042] \quad \alpha_{ij} = \text{softmax} \left(\frac{\text{Agg}(H_{inn_i}) H_{inn_{i,j}}^T}{\sqrt{d}} \right), j \in [1, B] \quad (2)$$

[0043] 其中, B表示分段内包含B个单词。 $H_{inn_{i,j}}$ 为B个单词对应的一组词特征表示, $H_{inn_{i,j}}^T$ 为 $H_{inn_{i,j}}$ 的转置, 其中i指示在第i分段, j指示在该分段内的第j个单词。 M_j 为掩码, $M_j \in \{0, -\text{inf}\}$, 在第j个单词为掩码的情况下, M_j 为负无穷, 此时 e^{M_j} 近似于0, 在第j个单词为单词的情况下, M_j 为0, 此时 e^{M_j} 为1。 $\text{Agg}(H_{inn_i})$ 即为第i分段的一组词特征表示的聚合结果, 也可以被称之为代表表征。 d 为模型维度, 也即编码器的维度。 α_{ij} 表示第一组权重, 其可以指示代表表征 $\text{Agg}(H_{inn_i})$ 和B个 $H_{inn_{i,j}}^T$ (也即B个隐含状态) 之间的注意力。

[0044] 分段内部自注意力层进而通过将第一组权重应用于一组词特征表示, 确定第一分段的段内特征表示:

$$[0045] \quad \widetilde{H_{inn_i}} = W_P \left(\sum_{j=1}^B H_{inn_{i,j}} \times \alpha_{ij} \right)^T + b_P, W_P \in R^{d \times d}, b_P \in R^d \quad (3)$$

[0046] 其中, W_P 和 b_P 均为编码器的可学习参数, $\widetilde{H_{inn_i}}$ 表示第i分段的段内特征表示。由此, 分段内部自注意力层可以基于上述获取到的第一组权重 α_{ij} , 对每个词特征表示 $H_{inn_{i,j}}$ 进行进一步基于注意力的加权, 以获得最终的段内特征表示。以此方式, 可以避免采用简单平均加权导致的信息丢失, 可以保留更多段内局部的信息。

[0047] 在框220, 电子设备110至少基于第一段内特征表示和第二段内特征表示, 确定第一分段的段间特征表示, 第二段内特征表示对应于多个分段中的第二段。第二段是不同与第一分段的任意分段。在一些实施例中, 电子设备110可以基于目标文本所包括的所有分段来确定任意分段对应的段间特征表示。

[0048] 在一些实施例中, 电子设备110同样利用编码器来确定第一分段的段间特征表示。编码器的分段内部自注意力层可以输出目标文本的多个分段对应的多个段内特征表示, 这多个段内特征表示被提供给编码器的分段间自注意力(inter-segment self-attention)层。分段间自注意力层可以基于如下公式确定第一分段的段间特征表示:

$$[0049] \quad Q = W_Q \widetilde{H_{inn}}, Q \in R^{n \times d}, W_Q \in R^{n \times d} \quad (4)$$

$$[0050] \quad K = W_K \widetilde{H_{inn}}, K \in R^{n \times d}, W_K \in R^{n \times d} \quad (5)$$

$$[0051] \quad V = W_V \widetilde{H_{inn}}, V \in R^{n \times d}, W_V \in R^{n \times d} \quad (6)$$

$$[0052] \quad H_{int} = SA(\widetilde{H_{inn}}) = \frac{QK^T}{\sqrt{d}} V, \widetilde{H_{inn}} \in R^{n \times d}, H_{int} \in R^{\frac{n}{B} \times d} \quad (7)$$

[0053] 其中, W_Q 、 W_K 、 W_V 为编码器的可学习参数, d 是编码器的维度, Q 表示查询特征, K 表示

键特征, V 表示值特征, H_{int} 表示段间特征表示。上述处理可以被理解为, 利用 Q 查询特征和键特征 K 计算注意力权重矩阵, 并用该注意力权重矩阵来对值特征 V 进行加权求和。以此方式, 分段间自注意力层可以确定捕获了对话整体的上下信息的段间特征表示。段间特征表示也可以被称之为全局表征。

[0054] 在框230, 电子设备110基于段间特征表示, 更新第一分段的第一段内特征表示。在一些实施例中, 编码器还包括分段状态更新器(segment updater)。分段状态更新器可以获取分段内部自注意力层输出的多个段内特征表示以及分段间自注意力层输出的多个段间特征表示。分段状态更新器进而可以基于段间特征表示来更新第一分段的第一段内特征表示。

[0055] 在一些实施例中, 分段状态更新器还可以获取多个分段各自的一组词特征表示。示例性的, 针对第一分段, 分段状态更新器可以获取第一分段对应的第一段内特征表示、一组词特征表示以及段间特征表示。分段状态更新器可以基于段间特征表示和一组词特征表示, 更新第一分段的第一段内特征表示。具体地, 分段状态更新器可以基于段间特征表示和一组词特征表示, 确定与段间特征表示对应的第二组权重。这里的第二组权重可以指示段间特征表示和一组词特征表示之间的注意力。分段状态更新器可以基于如下公式确定第二组权重:

$$[0056] \quad \beta_{ij} = \text{softmax} \left(\frac{H_{inn_{i,j}} H_{int_i}^T}{\sqrt{d}} \right), j \in [1, B] \quad (8)$$

[0057] 其中, β_{ij} 表示第二组权重, 其可以指示全局表征和一组隐含状态表征之间的注意力。

[0058] 分段状态更新器进而可以通过将第二组权重应用于段间特征表示, 确定段间特征表示的加权结果。进一步地, 分段状态更新器可以通过计算加权结果和一组词特征表示, 更新第一分段的一组词特征表示:

$$[0059] \quad \widetilde{H_{seg_{i,j}}} = \beta_{ij} \times H_{int_i} + H_{inn_{i,j}} \quad (9)$$

[0060] 其中, $\widetilde{H_{seg_{i,j}}}$ 即为第一分段的更新后的一组词特征表示。

[0061] 分段状态更新器可以基于更新后的一组词特征表示, 确定更新后的第一段内特征表示。在一些实施例中, 分段状态更新器可以采用上述公式(1)至公式(3)所示方法来确定更新后的第一段内特征表示 $\widetilde{H_{seg_i}}$ 。

[0062] 以此方式, 可以避免获得了段间特征表示后, 段内特征表示尚未被更新所导致的信息流成为单项的情况。可以避免局部信息丢失, 在保持对话全局信息高效建模的同时, 局部信息也得到同步更新, 从而有效地将对话结构信息和语义进行信息编码。

[0063] 在框240, 电子设备110至少基于更新后的第一段内特征表示, 生成目标文本的编码表示。

[0064] 在一些实施例中, 多个更新后的段内特征表示将被拼接起来得到 $\tilde{H} = \left\{ \widetilde{H_{seg_{i,j}}} \mid 1 \leq i \leq \frac{n}{B}, 1 \leq j \leq B \right\}$ 。这个拼接结果将被输入编码器的前馈神经网络

络(feed-forward)层,前馈神经网络层可以通过如下公式确定输出结果:

$$H_l = FF(\tilde{H}) = W_2 \text{Relu}(W_1 \tilde{H} + b_1) + b_2 \quad (10)$$

[0065]

$$W_1 \in R^{d \times 4d}, b_1 \in R^{4d}, W_2 \in R^{4d \times d}, b_2 \in R^d$$

[0066] 其中, W_1 、 b_1 、 W_2 、 b_2 均为可学习参数。 H_l 表示为当前层(第1层)的输出结果(假定编码模型一共包含M层网络层,每层网络层均包含分段内部自注意力层、分段间自注意力层、分段状态更新器以及前馈神经网络层。当前层为第1层, $0 < 1 < M$)。

[0067] 图3示出了根据本公开的一些实施例的编码器的示例架构的示意图。如图3所示,编码器300包括分段内部自注意力层310、分段间自注意力层320、分段状态更新器330以及前馈神经网络层340。电子设备110获取到目标文本后,确定目标文本包括的多个分段301(包括分段301-1、301-2、……、301-N,N是正整数)。为提高编码效率,在一些实施例中,电子设备110可以将这多个分段301同步提供给编码器300。编码器300的分段内部自注意力层310可以确定这多个分段301各自对应的多个段内特征表示311(包括段内特征表示311-1、311-2、……、311-N,N是正整数)。示例性的,如分段301-2为第一分段,则段内特征表示311-2为第一段内特征表示。在一些实施例中,分段内部自注意力层310可以同步输出多个分段301对应的多个段内特征表示311。若段内特征表示为 $\widetilde{H}_{inn_i}$,则分段内部自注意力层310可以输出 $\widetilde{H}_{inn} = [\widetilde{H}_{inn_1}, \widetilde{H}_{inn_2}, \dots, \widetilde{H}_{inn_{n/B}}]$,其中n表示目标文本共包含的单词数目,

B每个分段包含的单词数目,n/B表示分段数目, $\widetilde{H}_{inn}$ 表示目标文本包含的多个分段301对应的多个段内特征表示311的集合。

[0068] 多个段内特征表示311被提供给分段间自注意力层320。分段间自注意力层320可以基于多个段内特征表示311输出与多个分段301对应的多个段间特征表示321(包括段间特征表示321-1、321-2、……、321-N,N是正整数)。这多个段间特征表示321与多个段内特征表示311一同被提供给分段状态更新器330。分段状态更新器330可以基于多个段间特征表示321,更新各个分段的段内特征表示311以得到更新后的多个段内特征表示331(包括更新后的段内特征表示331-1、331-2、……、331-N,N是正整数)。

[0069] 最后,更新后的多个段内特征表示331将被拼接,拼接结果被提供给前馈神经网络层340。前馈神经网络层340可以输出目标文本对应的编码表示345。

[0070] 图3示出了编码器仅包括单层次化分段自注意力神经网络的示例,可以理解,编码器还可以包含多层次化分段自注意力神经网络,本公开对此不作限定。

[0071] 以此方式,可以通过堆叠多层(例如12层,24层等)如上层次化分段自注意力神经网络(也即上述分段内部自注意力层、分段间自注意力层、分段状态更新器),形成一个比较深的对话编码器。此外,整个网络相比于传统的transformer网络(仅包含自注意力层和前馈神经网络层),用层次化分段自注意力神经网络(也即上述分段内部自注意力层、分段间自注意力层、分段状态更新器)代替了传统的自注意力层却没有新增学习参数,因此,可以直接加载BERT、Roberta等训练模型的权重,避免从头训练编码模型。

[0072] 以此方式,每个层内部都可以模拟人类理解对话过程:首先阅读局部信息(对应确定段内特征表示),然后进行对话内容全局的理解(对应确定段间特征表示),反馈局部信息

的进一步理解(对应更新段内特征表示)。这可以保证全局信息和局部信息在交替流动,有利于很多下游对话任务,例如指代消解、上下文理解、状态预测等等。

[0073] 此外,由于这样的网络结构采用了分段的层次化注意力计算方法,原本的 $O(n^2)$ 复杂度大幅度降低了到了 $O\left(nB + \left(\frac{n}{B}\right)^2 + n\right)$,这可以提高编码推理速度(例如如果 $n=1024, B=64$,则计算复杂度只有原先的6.4%左右),这可以在实际应用时降低对电子设备110的计算量的需求。

[0074] 综上所述,本公开的实施例可以通过结合文本的各个分段的段内特征表示和段间特征表示来充分考虑文本中的结构信息和语义信息,能够强化局部信息和全局信息交互,有助于提升编码效果。此外,还可以通过同时编码多个分段来提高编码的效率,能够有效编码长文本。

[0075] 本公开的实施例还提供了用于实现上述方法或过程的相应装置。图4示出了根据本公开的某些实施例的用于编码的装置400的示意性结构框图。装置400可以被实现为或者被包括在电子设备110中。装置400中的各个模块/组件可以由硬件、软件、固件或者它们的任意组合来实现。

[0076] 如图4所示,装置400包括第一确定模块410,被配置为确定目标文本的多个分段中第一分段的第一段内特征表示。装置400还包括第二确定模块420,被配置为至少基于第一段内特征表示和第二段内特征表示,确定第一分段的段间特征表示,第二段内特征表示对应于多个分段中的第二段。装置400还包括更新模块430,被配置为基于段间特征表示,更新第一分段的第一段内特征表示。装置400还包括编码表示生成模块440,被配置为至少基于更新后的第一段内特征表示,生成目标文本的编码表示。

[0077] 在一些实施例中,装置400还包括分段模块,被配置为基于预设长度对目标文本进行划分,以确定目标文本的多个分段。

[0078] 在一些实施例中,第一确定模块410包括:词确定模块,被配置为确定第一分段中的一组词;词特征确定模块,被配置为确定与一组词相关联的一组词特征表示;以及段内特征确定模块,被配置为基于一组词特征表示,确定第一分段的第一段内特征表示。

[0079] 在一些实施例中,段内特征确定模块包括:第一聚合结果确定模块,被配置为通过聚合一组词特征表示,确定第一聚合结果;以及第三确定模块,被配置为基于第一聚合结果和一组词特征表示,确定第一分段的第一段内特征表示。

[0080] 在一些实施例中,第三确定模块包括:第一权重确定模块,被配置为基于第一聚合结果和一组词特征表示,确定与一组词特征表示对应的第一组权重,第一组权重指示第一聚合结果和一组词特征表示之间的注意力;以及第四确定模块,被配置为通过将第一组权重应用于一组词特征表示,确定第一分段的第一段内特征表示。

[0081] 在一些实施例中,更新模块430包括:第一更新模块,被配置为基于段间特征表示和一组词特征表示,更新第一分段的第一段内特征表示。

[0082] 在一些实施例中,第一更新模块包括:第二权重确定模块,被配置为基于段间特征表示和一组词特征表示,确定与段间特征表示对应的第二组权重,第二组权重指示段间特征表示和一组词特征表示之间的注意力;加权结果确定模块,被配置为通过将第二组权重应用于段间特征表示,确定段间特征表示的加权结果;第二更新模块,被配置为通过计算加

权结果和一组词特征表示,更新第一分段的一组词特征表示;以及第五确定模块,被配置为基于更新后的一组词特征表示,确定更新后的第一段内特征表示。

[0083] 在一些实施例中,编码表示生成模块440包括:第二聚合结果确定模块,被配置为通过聚合至少第一分段的更新后的第一段内特征表示和第二分段的更新后的第二段内特征表示,确定目标文本对应的第二聚合结果;以及编码表示确定模块,被配置为至少基于第二聚合结果,确定目标文本的编码表示。

[0084] 在一些实施例中,目标文本为对话文本。

[0085] 装置400中所包括的单元和/或模块可以利用各种方式来实现,包括软件、硬件、固件或其任意组合。在一些实施例中,一个或多个单元和/或模块可以使用软件和/或固件来实现,例如存储在存储介质上的机器可执行指令。除了机器可执行指令之外或者作为替代,装置400中的部分或者全部单元和/或模块可以至少部分地由一个或多个硬件逻辑组件来实现。作为示例而非限制,可以使用的示范类型的硬件逻辑组件包括现场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准品(ASSP)、片上系统(SOC)、复杂可编程逻辑器件(CPLD),等等。

[0086] 应当理解,以上方法中的一个或多个步骤可以由适当的电子设备或电子设备的组合来执行。这样的电子设备或电子设备的组合例如可以包括图1中的电子设备110。

[0087] 图5示出了其中可以实施本公开的一个或多个实施例的电子设备500的框图。应当理解,图5所示出的电子设备500仅仅是示例性的,而不应当构成对本文所描述的实施例的功能和范围的任何限制。图5所示出的电子设备500可以用于实现图1的电子设备110。

[0088] 如图5所示,电子设备500是通用电子设备的形式。电子设备500的组件可以包括但不限于一个或多个处理器或处理单元510、存储器520、存储设备530、一个或多个通信单元540、一个或多个输入设备550以及一个或多个输出设备560。处理单元510可以是实际或虚拟处理器并且能够根据存储器520中存储的程序来执行各种处理。在多个处理器系统中,多个处理单元并行执行计算机可执行指令,以提高电子设备500的并行处理能力。

[0089] 电子设备500通常包括多个计算机存储介质。这样的介质可以是电子设备500可访问的任何可以获得的介质,包括但不限于易失性和非易失性介质、可拆卸和不可拆卸介质。存储器520可以是易失性存储器(例如寄存器、高速缓存、随机访问存储器(RAM))、非易失性存储器(例如,只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、闪存)或它们的某种组合。存储设备530可以是可拆卸或不可拆卸的介质,并且可以包括机器可读介质,诸如闪存驱动、磁盘或者任何其他介质,其可以能够用于存储信息和/或数据并且可以在电子设备500内被访问。

[0090] 电子设备500可以进一步包括另外的可拆卸/不可拆卸、易失性/非易失性存储介质。尽管未在图5中示出,可以提供用于从可拆卸、非易失性磁盘(例如“软盘”)进行读取或写入的磁盘驱动和用于从可拆卸、非易失性光盘进行读取或写入的光盘驱动。在这些情况中,每个驱动可以由一个或多个数据介质接口被连接至总线(未示出)。存储器520可以包括计算机程序产品525,其具有一个或多个程序模块,这些程序模块被配置为执行本公开的各种实施例的各种方法或动作。

[0091] 通信单元540实现通过通信介质与其他电子设备进行通信。附加地,电子设备500的组件的功能可以以单个计算集群或多个计算机器来实现,这些计算机器能够通过通信连

接进行通信。因此,电子设备500可以使用与一个或多个其他服务器、网络个人计算机(PC)或者另一个网络节点的逻辑连接来在联网环境中进行操作。

[0092] 输入设备550可以是一个或多个输入设备,例如鼠标、键盘、追踪球等。输出设备560可以是一个或多个输出设备,例如显示器、扬声器、打印机等。电子设备500还可以根据需要通过通信单元540与一个或多个外部设备(未示出)进行通信,外部设备诸如存储设备、显示设备等,与一个或多个使得用户与电子设备500交互的设备进行通信,或者与使得电子设备500与一个或多个其他电子设备通信的任何设备(例如,网卡、调制解调器等)进行通信。这样的通信可以经由输入/输出(I/O)接口(未示出)来执行。

[0093] 根据本公开的示例性实现方式,提供了一种计算机可读存储介质,其上存储有计算机可执行指令,其中计算机可执行指令被处理器执行以实现上文描述的方法。根据本公开的示例性实现方式,还提供了一种计算机程序产品,计算机程序产品被有形地存储在非瞬态计算机可读介质上并且包括计算机可执行指令,而计算机可执行指令被处理器执行以实现上文描述的方法。

[0094] 这里参照根据本公开实现的方法、装置、设备和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0095] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理单元,从而生产出一种机器,使得这些指令在通过计算机或其他可编程数据处理装置的处理单元执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0096] 可以把计算机可读程序指令加载到计算机、其他可编程数据处理装置、或其他设备上,使得在计算机、其他可编程数据处理装置或其他设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机、其他可编程数据处理装置、或其他设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0097] 附图中的流程图和框图显示了根据本公开的多个实现的系统、方法和计算机程序产品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、程序段或指令的一部分,模块、程序段或指令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。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,两个连续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。也要注意的,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动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。

[0098] 以上已经描述了本公开的各实现,上述说明是示例性的,并非穷尽性的,并且也不限于所公开的各实现。在不偏离所说明的各实现的范围和精神的情况下,对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择,旨在最好地解释各实现的原理、实际应用或对市场中的技术的改进,或者使本技术领域的其他普通

技术人员能理解本文公开的各个实现方式。

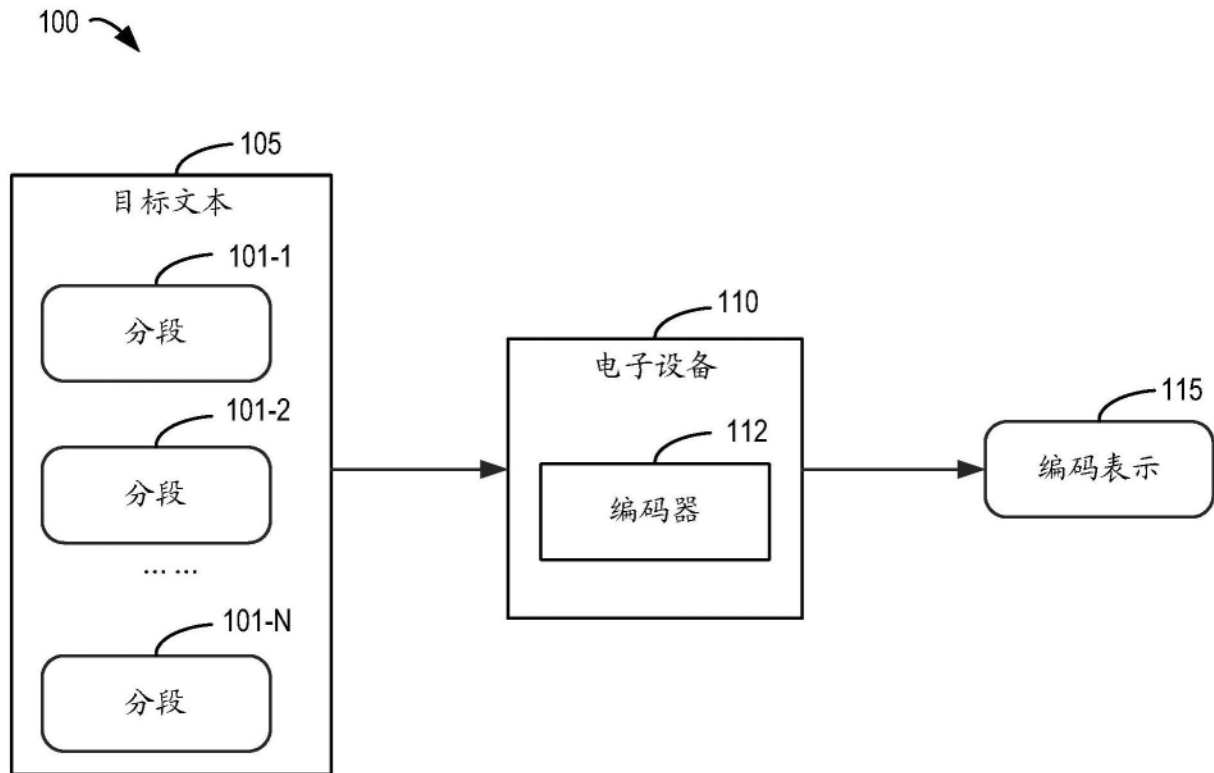


图1

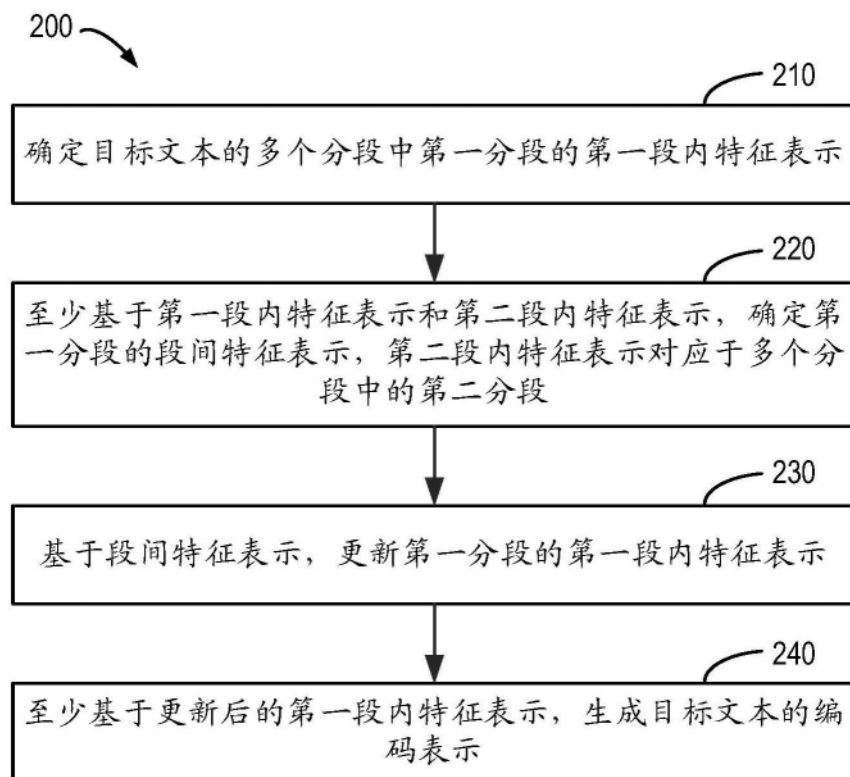


图2

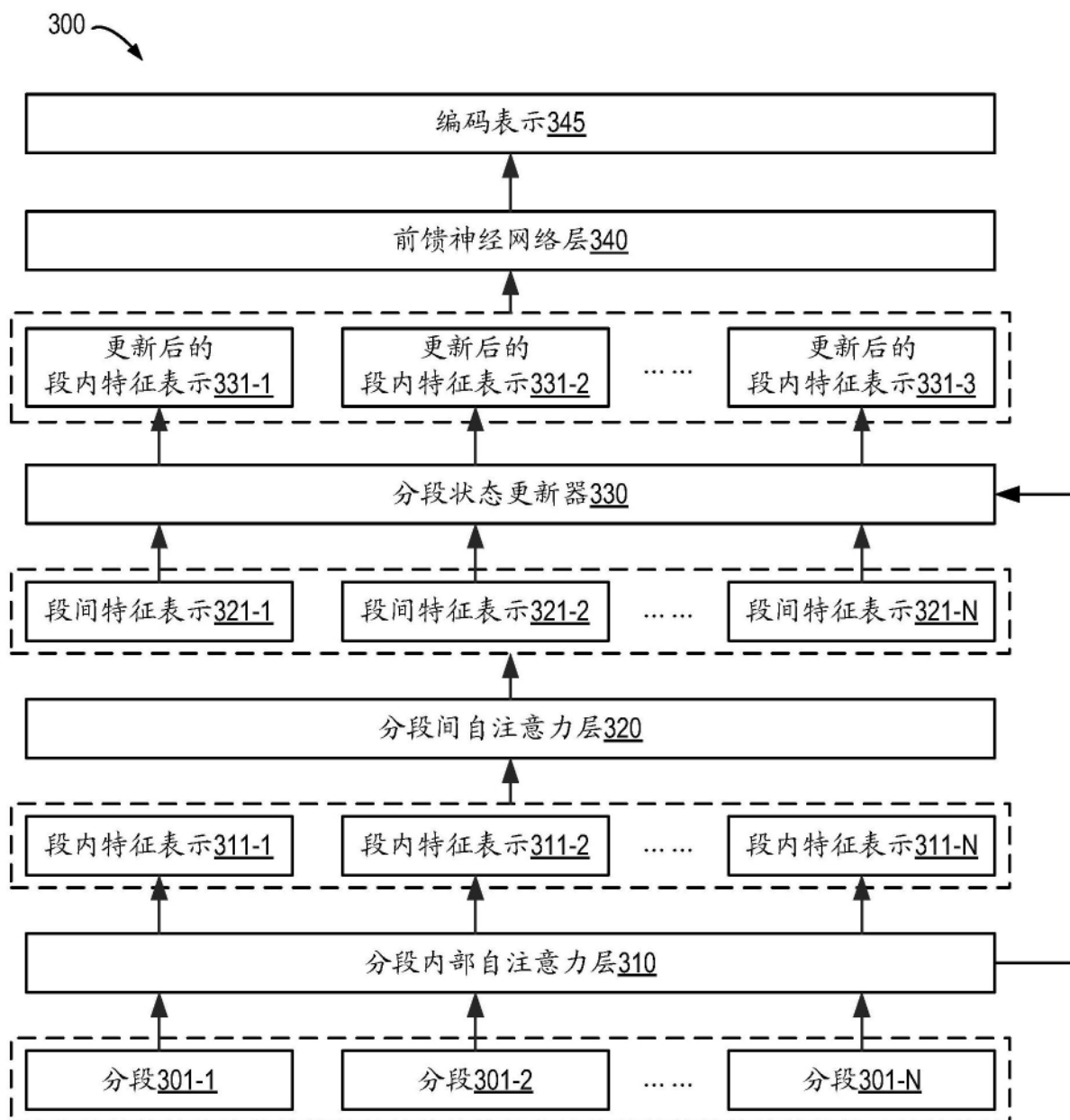


图3

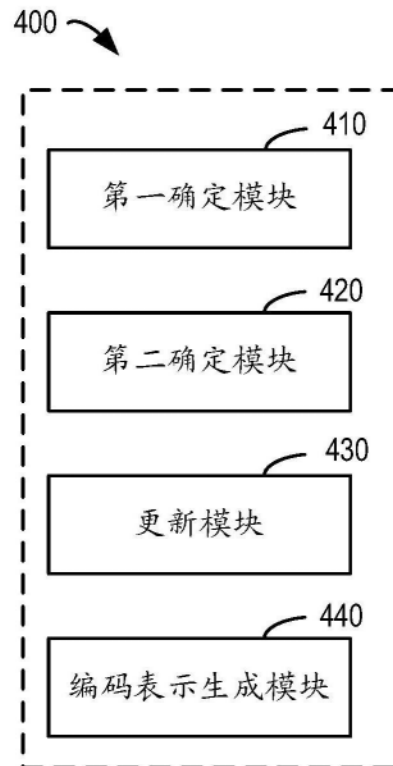


图4

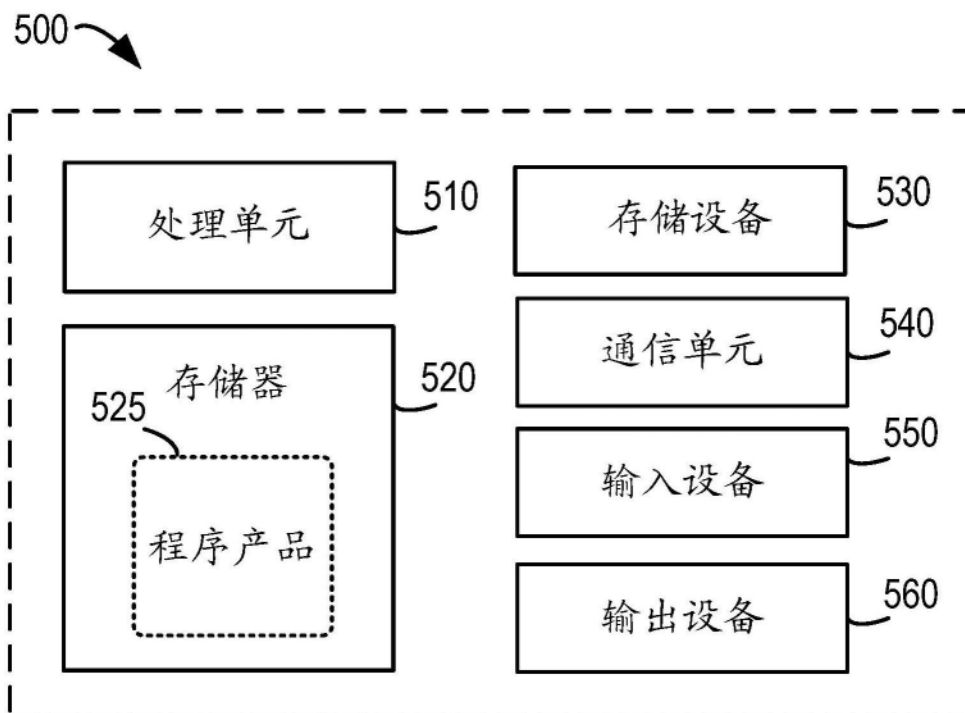


图5