(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113519001 A
(43)申请公布日 2021.10.19

(21)申请号 202080018455.X

(22)申请日 2020.02.24

(30)优先权数据
62/813,697 2019.03.04 US
16/393,801 2019.04.24 US

(85)PCT国际申请进入国家阶段日
2021.09.02

(86)PCT国际申请的申请数据
PCT/US2020/019453 2020.02.24

(87)PCT国际申请的公布数据
WO2020/180518 EN 2020.09.10

(71)申请人 易享信息技术有限公司
地址 美国加利福尼亚州

(72)发明人 N·拉贾尼 B·麦卡恩

(74)专利代理机构 北京市联德律师事务所 11361
代理人 黄大正 张来光

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/08(2006.01)

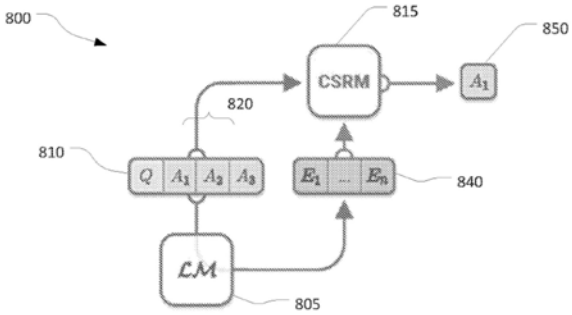
权利要求书2页 说明书10页 附图8页

(54)发明名称

利用语言模型生成常识解释

(57)摘要

根据一些实施方案,提供了用于开发或提供常识自动生成解释(CAGE)的系统和方法,用于由人工智能、神经网络或深度学习模型使用的推理进行预测。在一些实施方案中,系统和方法使用对语言模型(LM)的有监督微调而生成这样的解释。这些解释然后可以用于下游分类。



CN 113519001 A

1.一种方法,包括:

通过嵌入模块,针对问答集的结构化源文本进行编码和嵌入,所述问答集包括问题和多个回答选择;

通过多层变换器模块,基于生成的与来自先前迭代的结构化解释文本相关的令牌而迭代地解码所述嵌入模块的输出,以生成用于推断哪个回答选择对于所述问题正确的解释文本,其中所述来自先前迭代的结构化解释文本包括由人类注释器生成的解释文本;

将生成的解释文本提供给分类模块;和

使用所述生成的解释文本,在所述分类模块处生成哪一个回答选择对于所述问题正确的预测。

2.根据权利要求1所述的方法,其中所述问答集的结构化源文本包括自然语言形式的文本。

3.根据权利要求1或2所述的方法,包括从所述人类注释器收集所述结构化解释文本。

4.根据权利要求3所述的方法,其中收集所述结构化解释文本包括:

给所述人类注释器提供训练问答集;和

响应于所述训练问答集,从所述人类注释器接收所述结构化解释文本。

5.根据权利要求1-4中任一项所述的方法,包括将所述问答集提供给所述分类模块,其中当被提供给所述分类模块时,所述问题、所述多个回答选择和所述生成的解释文本由分隔符分开。

6.根据权利要求1-5中任一项所述的方法,其中所述嵌入模块和所述多层变换器模块包括自然语言模型的至少一部分。

7.根据权利要求1-6中任一项所述的方法,其中所述分类模块包括多层变换器编码器。

8.一种系统,包括:

嵌入模块,其用于针对问答集编码和嵌入结构化源文本,所述问答集包括问题和多个回答选择;

多层变换器模块,其用于基于生成的与来自先前迭代的结构化解释文本相关的令牌而迭代地解码所述嵌入模块的输出,以生成用于推断哪个回答选择对于问题正确的解释文本,其中所述来自先前迭代的结构化解释文本包括由人类注释器生成的解释文本;和

分类模块,其用于使用生成的解释文本而生成哪一个回答选择对于所述问题正确的预测。

9.根据权利要求8所述的系统,其中所述问答集的结构化源文本包括自然语言形式的文本。

10.根据权利要求8或9所述的系统,其中所述嵌入模块和所述多层变换器模块包括至少一部分神经网络。

11.根据权利要求8-10中任一项所述的系统,其中由分隔符分开的所述问题、所述多个回答选择和所述生成的解释文本被提供给所述分类模块。

12.根据权利要求8-11中任一项所述的系统,其中所述嵌入模块和所述多层变换器模块包括自然语言模型的至少一部分。

13.根据权利要求8-12中任一项所述的系统,其中所述分类模块包括多层变换器编码器。

14.一种非暂时性机器可读介质,包括可执行代码,当由与计算机相关联的一个或多个处理器执行时,所述可执行代码适于使所述一个或多个处理器执行包括如下的方法:

通过嵌入模块,针对问答集的结构化源文本进行编码和嵌入,所述问答集包括问题和多个回答选择;

通过多层变换器模块,基于生成的与来自先前迭代的结构化解释文本相关的令牌而迭代地解码嵌入模块的输出,以生成用于推断哪个回答选择对于问题正确的解释文本,其中所述来自先前迭代的结构化解释文本包括由人类注释器生成的解释文本;

将生成的解释文本提供给分类模块;和

使用所述生成的解释文本,在所述分类模块处生成哪一个回答选择对于所述问题正确的预测。

15.根据权利要求14所述的非暂时性机器可读介质,其中所述问答集的结构化源文本包括自然语言形式的文本。

16.根据权利要求14或15所述的非暂时性机器可读介质,包括当由所述一个或多个处理器执行时适于使所述一个或多个处理器从所述人类注释器收集所述结构化解释文本的可执行代码。

17.根据权利要求14-16中任一项所述的非暂时性机器可读介质,包括当由所述一个或多个处理器执行时适于使所述一个或多个处理器执行如下操作的可执行代码:

给所述人类注释器提供训练问答集;和

响应于所述训练问答集,从所述人类注释器接收所述结构化解释文本。

18.根据权利要求14-17中任一项所述的非暂时性机器可读介质,包括当由所述一个或多个处理器执行时适于使所述一个或多个处理器给所述分类模块提供所述问答集的可执行代码,其中所述问题、所述多个回答选择和所述生成的解释文本当被提供给所述分类模块时由分隔符分开。

19.根据权利要求14-18中任一项所述的非暂时性机器可读介质,其中所述嵌入模块和所述多层变换器模块包括自然语言模型的至少一部分。

20.根据权利要求14-19中任一项所述的非暂时性机器可读介质,其中所述分类模块包括多层变换器编码器。

利用语言模型生成常识解释

[0001] 发明人:N·拉贾尼和B·麦卡恩

[0002] 相关申请

[0003] 本申请要求2019年3月4日提交的美国临时专利申请号62/813,697和2019年4月24日提交的美国非临时专利申请号16/393,801的优先权,其全部内容通过引用并入本文。

[0004] 版权声明

技术领域

[0006] 公开内容大体上涉及自然语言处理,并且更具体地涉及利用(leveraging)语言模型生成推理(reasoning)或合理化(rationalization)的常识解释(commonsense explanation)。

背景技术

[0007] 使用神经网络和深度学习模型实现的人工智能已经展现出作为以人类似的(human-like)准确度自动分析真实世界信息的技术的巨大前景。然而,人工智能或深度学习模型通常不能解释它们的预测背后的推理或该预测的合理化,或不能解释该推理或合理化在何种程度上基于常识知识。这使得人类难以理解和信任这样的模型。

[0008] 因此,具有如下系统和方法将是有利的,该系统和方法提供、实现或改进了人工智能或深度学习模型中的常识推理或合理化,并且另外生成或提供了对该推理或合理化的解释。

附图说明

[0009] 图1是根据一些实施方案的计算设备的简化图。

[0010] 图2示出了根据一些实施方案的可以被包括在常识解释(CoS-E)数据集中的问题、回答和人类生成的解释(human generated explanation)的示例。

[0011] 图3示出了根据一些实施方案的在CoS-E数据集中收集的解释的示例分布。

[0012] 图4示出了根据一些实施方案的训练常识自动生成解释(Commonsense Auto-Generated Explanations,CAGE)语言模型以从CoS-E数据集生成解释的示例时间步骤。

[0013] 图5是根据一些实施方案的语言模块或模型的简化图。

[0014] 图6示出了根据一些实施方案的生成预测的分类模型或模块的示例时间步骤。

[0015] 图7是根据一些实施方案的分类模型或模块的简化图。

[0016] 图8是示出根据一些实施方案的通过人工智能或深度学习模型生成推理的常识解释的系统的简化图。

[0017] 图9是根据一些实施方案的通过人工智能或深度学习模型生成推理的常识解释的

方法的简化图。

[0018] 图10是示出根据一些实施方案的通过人工智能或深度学习模型生成合理化的常识解释的系统的简化图。

[0019] 图11是根据一些实施方案的通过人工智能或深度学习模型生成合理化的常识解释的方法的简化图。

[0020] 图12示出了表格，该表格示出根据一些实施方案的推理和合理化的来自常识QA(CommonsenseQA)、CoS-E和CAGE样本的示例集合。

[0021] 图13示出了表格，该表格示出结果比较。

[0022] 在附图中，具有相同标记的元件具有相同或相似的功能。

具体实施方式

[0023] 示出各个方面、实施方案、实现或应用的描述和附图不应当被认为是限制性的——权利要求书限定了受保护的发明。在不偏离说明书和权利要求书的精神和范围的情况下，可以进行各种机械、组成、结构、电气和操作改变。在一些情况下，没有详细示出或描述公知的电路、结构或技术，因为这些是本领域技术人员已知的。两个或更多附图中的相同数字表示相同或相似的元件。

[0024] 在说明书中，阐述了描述根据公开内容的一些实施方案的具体细节。为了提供对实施方案的透彻理解，阐述了许多具体细节。然而，对于本领域技术人员显而易见的是，可以在没有这些具体细节中的一些或全部的情况下实施一些实施方案。本文公开的具体实施方案旨在是说明性而非限制性的。本领域的技术人员可以认识到尽管未在此具体描述，但其他元件在公开内容的范围和精神内。此外，为了避免不必要的重复，结合一个实施方案显示和描述的一个或多个特征可以结合入其他实施方案，除非另外具体描述或该一个或多个特征将使实施方案不起作用。

[0025] 概述

[0026] 使用神经网络和深度学习模型实现的人工智能已经展现出作为以人类似的准确度自动分析真实世界信息的技术的巨大前景。通常，这样的神经网络和深度学习模型接收输入信息并且基于该输入信息进行预测。然而，这些模型可能面临应用常识推理或合理化以发展或解释其预测的挑战。常识推理或合理化是现代机器学习方法的具有挑战性的任务。人工智能或深度学习模型通常不能解释其预测背后的推理或合理化(常识或其他)，这使得人类难以理解和信任这样的模型。

[0027] 应用常识推理或合理化并且进行解释将有助于使深层神经网络对人类更加透明并且建立信任。

[0028] 根据一些实施方案，公开内容提供了利用预先训练的语言模型生成对常识推理或合理化有用的解释的系统和方法。在一些实施方案中，常识自动生成解释(CAGE)被提供作为用于对生成常识问答(常识QA)的解释的框架。常识QA是为开发具有常识推理能力的自然语言处理(NLP)模型而提出的多项选择问答数据集，如Talmor et al.,“A Question Answering Challenge Targeting Commensense Knowledge,”arXiv:1811.00937v2(2018年11月2日)详细描述的，其通过引用并入本文。存在多个版本的常识QA(例如，v1.0、v1.1)，它们中的任何一个可以在一个或多个实施方案中使用。NLP是神经网络可能适用的一类问

题。NLP可以被用于给新的神经网络灌输(instill)对各个单词和短语的理解。

[0029] 在一些实施方案中,常识推理的人类解释被生成并且作为常识解释(CoS-E)在常识QA的集库(corpus)之外被建立或被添加至常识QA的集库。在一些实施方案中,CoS-E包含开放式(open-ended)自然语言解释以及突出显示的跨距注释(highlighted span annotation)二者的形式的人类解释,该突出显示的跨距注释将由人类选择的词语表示为对于预测正确的回答是重要的。

[0030] 根据一些实施方案,常识推理的任务被分解为两个阶段。在第一阶段,公开内容的系统和方法提供了常识QA示例以及针对语言模型的相应的CoS-E解释。语言模型决定来自示例的问题和回答选择,并且被训练以生成CoS-E解释。在第二阶段中,公开内容的系统和方法使用语言模型针对常识QA的训练和验证集中的每个示例生成解释。这些常识自动生成解释(CAGE)通过将第二常识推理模型连接(concatenate)至原始问题、回答选择和语言模型的输出的末尾而被提供给第二常识推理模型。两阶段CAGE框架获得了现有技术水平的结果,其超过了最佳报告基线10%,并且还产生了解释以证明其预测是常识自动产生的解释(CAGE)。

[0031] 总之而言,公开内容提出了一种新的常识解释(CoS-E)数据集以研究神经常识推理。公开内容提供了一种用于基于常识QA自动生成解释的新方法(CAGE),该解释取得了大约65%的现有技术准确性。

[0032] 计算设备

[0033] 图1是根据一些实施方案的计算设备的简化图。如图1所示,计算设备100包括耦合至存储器120的处理器110。计算设备100的操作由处理器110控制。而且虽然计算设备100被显示仅具有一个处理器110,但是应当理解,处理器110可以表示计算设备100中的一个或多个中央处理单元、多核处理器、微处理器、微控制器、数字信号处理器、现场可编程门阵列(FPGA)、专用集成电路(ASIC)、图形处理单元(GPU)等。计算设备100可以被实现为独立子系统、添加到计算设备的面板(board)和/或虚拟机。

[0034] 存储器120可以被用于存储由计算设备100执行的软件和/或在计算设备100的操作期间使用的一个或多个数据结构。存储器120可以包括一种或多种类型的机器可读介质。机器可读介质的一些常见形式可以包括软盘、软磁盘、硬盘、磁带、任何其它磁介质、CD-ROM、任何其它光学介质、打孔卡、纸带、具有孔图案的任何其它物理介质、RAM、PROM、EPROM、FLASH-EPROM、任何其它存储器芯片或盒式磁带和/或处理器或计算机适于读取的任何其它介质。

[0035] 处理器110和/或存储器120可以以任何合适的物理布置被布置。在一些实施方案中,处理器110和/或存储器120可以在相同的面板上、在相同的封装中(例如,系统级封装)、在相同的芯片上(例如,系统级芯片)等实现。在一些实施方案中,处理器110和/或存储器120可以包括分布式、虚拟化和/或容器化计算资源。与这样的实施方案一致,处理器110和/或存储器120可以位于一个或多个数据中心和/或云计算设施中。

[0036] 如图所示,存储器120包括可以被用于实现和/或仿真系统和模型,和/或实现本文进一步描述的任何方法的常识解释模块130。在一些示例中,常识解释模块130可以被用于开发、导出或生成预测,应用常识推理或合理化,并且生成或提供与本文进一步描述的相同的解释。在一些示例中,常识解释模块130还可以处理用于生成预测、应用常识推理或合理

化并且生成或提供解释的系统或模型的迭代训练和/或评估。在一些示例中，存储器120可以包括非暂时性、有形的机器可读介质，其包括可执行代码，当由一个或多个处理器(例如，处理器110)运行时，该可执行代码可以致使一个或多个处理器执行本文进一步详细描述的方法。在一些示例中，可以使用硬件、软件和/或硬件与软件的组合而实现常识解释模块130。

[0037] 如图所示，计算设备100接收被提供给常识解释模块130的输入数据140和自然语言解释文本145。输入数据140可能涉及期望应用人工智能、神经网络或深度学习模型而分析和进行预测的任何情况、场景、问题等，例如，用于问答(QA)或一些其他NLP任务。在一些实施方案中，自然语言解释文本145可以包括常识推理的人类解释，其可以是常识解释(CoS-E)。人类解释可以是开放式自然语言解释形式以及原始输入实例中突出显示的注释的形式。在一些实施方案中，自然语言解释文本145可以包括自动生成的解释。自然语言解释文本145可以被用于常识解释模块130的微调或训练。在一些实施方案中，该训练可以在由常识解释模块130执行或进行的一个或多个迭代之上发生。

[0038] 常识解释模块130对输入数据140进行操作以开发、导出或生成预测或结果，在进行上述操作时使用自然语言解释文本145以支持或应用常识推理。模块130还可以生成或提供其推理或合理化的解释。在一些实施方案中，常识解释模块130实现或结合可以生成解释的语言模型(LM)。在一些实施方案中，常识解释模块130实现或结合至少部分地基于来自语言模型(LM)的解释而开发或生成预测或结果的常识推理模型(CSRM)或分类模型。在一些实施方案中，常识解释模块130使用或结合生成性预先训练-变换器(Generative Pre-Trained Transformer，GPT)语言模型(如Radford et al.，“Improving language understanding by generative pre-training，”https://s3-us-west-2.amazonaws.com/openai-assets/research-overs/language-unsupervised/language understanding a paper.pdf.进一步描述的，其通过引用并入本文)并且通过对问题、回答选择和人类生成的解释的确定而在常识QA训练数据之上对其进行微调。结果和解释被提供为来自计算设备100的输出150。

[0039] 在一些示例中，常识解释模块130可以包括具有适当的预处理、编码、解码和输出层的单层或多层神经网络。神经网络已经展现作为以人类似的准确度自动分析真实世界信息的技术的巨大前景。通常，神经网络模型接收输入信息并且基于该输入信息进行预测。而分析真实世界信息的其它方法可能涉及硬编码过程、统计分析等，神经网络通过反复试验过程使用机器学习过程进行学习以逐步进行预测。可以使用大量训练示例训练给定的神经网络模型，迭代地进行直到神经网络模型开始从训练示例一致地做出人类可能做出的类似推断。尽管将常识解释模块130描绘为软件模块，但是其可以使用硬件、软件和/或硬件与软件的组合而实现。

[0040] 常识解释(CoS-E)

[0041] 根据一些实施方案，公开内容的语言模型系统和方法可以使用或利用常识推理的人类解释，其可以在常识解释(CoS-E)数据集中。在一些实施方案中，CoS-E数据集被添加到现有常识QA数据集或构建在现有常识QA数据集之上，以便在公开内容的语言模型系统和方法中使用。常识QA数据集由两个分割(split)组成，如Talmor et al.，“A Question Answering Challenge Targeting Commensense Knowledge，”arXiv:1811.00937v2(2018

年11月2日)描述的,其通过引用并入本文。在一些实施方案中,公开内容的CoS-E数据集和语言模型使用更困难的随机分割,即主要评估分割。常识QA中的每个示例由问题q,三个回答选择c0、c1、c2和标记回答a组成。CoS-E数据集添加了人类解释e_h,用于解释为什么a是最适当的选择。

[0042] 在一些实施方案中,可以例如使用亚马逊土耳其机器人(Amazon Mechanical Turk,MTurk)收集用于CoS-E数据集的常识推理的人类解释。如图2所示的示例中显示的,系统向人类参与者呈现或提出一个或多个问题210(例如,“与朋友一起吃汉堡时,人们试图做什么?”)和回答选择220(例如,“玩的开心、美味或消化不良”)以及真实(ground-truth)回答选择230(例如,“玩的开心”,例如粗体所示)。系统提示人类参与者出现下列问题:“为什么预测的输出是最适合的回答呢?”。系统指示人类参与者突出问题210中证明真实回答选择230正确的相关词语240(例如,“汉堡、与朋友一起”),并且基于突出的证明提供简短的开放式解释250(例如,“通常与朋友一起吃汉堡表示美好时光”),该突出的证明可以用作问题背后的常识推理。系统收集这些解释以添加到或建立在常识QA训练-随机-分割和去-随机-分割之上,这些解释可以分别具有7610个和950个示例的大小。所得的CoS-E数据集包括问题、回答选择以及真实回答选择的自由形式的解释和突出的文本二者。数据集中的突出显示的文本或词语240可以被称为“CoS-E-选择的”,而自由形式解释250可以被称为“CoS-E-开放式的”。

[0043] 关于收集常识推理的人类生成的解释,可能难以控制由与系统交互的参与者提供的开放式注释(例如,解释250)的质量。因此,在一些实施方案中,系统可以执行浏览器内检查以避免或拒绝明显错误的解释。在一些实施方案中,如果她/他没有突出问题210中的相关词语240或如果解释250的长度小于四个词语,则不允许人类注释器在系统中向前进行。系统还可以在没有任何其他额外词语的情况下检查解释250不是问题20或回答选择220的子串。在一些实施方案中,系统从每个示例的一个注释器收集这些解释250。系统还可以执行一个或多个后收集检查以捕捉未被其他过滤器捕捉或标识的示例。系统可以过滤出可以被分类为模板的解释250。例如,形式“<回答>是唯一的选项,即[正确明显]”的解释可以被系统删除,然后由相同或不同的人类参与者重新呈现以进行注释。

[0044] 图3示出了在一些实施方案中在CoS-E数据集中收集的解释(例如,图2的开放式解释250)的示例分布300。如图3所示,来自CoS-E数据集的58%的解释包含真实回答选择(例如,真实回答选择230)——情况“A”。7%的解释包括干扰物(或问题的错误选择)——情况“B”。12%的解释包括真实和干扰物(A和B),而23%的解释不包括真实或干扰物(A和B都不)。42%的解释具有与问题(例如,问题210)的二字组(bigram)重叠,而22%的解释具有与问题的三字组(trigram)重叠。

[0045] 在一些实施方案中,可以提供CoS-E数据集的人类生成的解释(例如,图2的解释250),例如,作为输入到计算设备100(图1)以供常识和解释模块130使用的自然语言解释文本145。根据一些实施方案,例如,如在模块130中实现或并入的那样,将CoS-E数据集添加到用于语言模型系统和方法的现有常识QA数据集中。语言模型(LM)使用CoS-E数据集的有效性并不局限于数据集的那些具体示例。在一些实施方案中,语言模型仅在训练期间通过使用CoS-E数据集来获得现有技术水平的结果。实证结果表明,即使仅使用那些不与任何回答选择有任何词语重叠的解释,性能也完全超过了不使用CoS-E数据集的基线的性能。还观察

到,在CoS-E数据集中也存在相当大比例的干扰物选择,并且在进一步的分析中,我们发现对于那些示例,注释器通过消除错误的选择而进行解释。这表明,即使对人类来说,也难以推导出常识QA的许多示例。CoS-E还向常识QA数据集添加了多样性的观点,尤其是对世界知识的多样性推理。即使许多解释在质量控制检查之后仍然是有噪声的,CoS-E数据集的解释也具有足够的质量以训练产生常识推理的语言模型。

[0046] 常识自动生成解释(CAGE)

[0047] 语言模型系统和方法可以开发、导出或生成NLP任务(例如问答)的预测或结果。根据一些实施方案,公开内容的语言模型系统和方法针对它们的预测或结果生成或输出它们的推理和原理(rationale)的解释——常识自动生成解释(CAGE)。在一些实施方案中,例如,语言模型或模块——如在常识解释模块130中实现或结合的——响应于或使用输入数据140和自然语言解释文本145生成这些解释。解释由语言模型生成并且被用作分类模型或模块的补充输入。

[0048] 在一些实施方案中,提供了CAGE并且将其应用于常识QA任务。如先前描述的,常识QA中的每个示例由问题q、三个回答选择c0、c1、c2和标记回答a组成;并且CoS-E数据集添加了为什么a是最适当的选择的人类解释e_h。CAGE的输出是语言模型生成的解释e,其被训练为接近e_h。

[0049] 根据一些实施方案,为了向分类模型提供CAGE,对语言模型(LM)进行微调或修改以从CoS-E数据集生成解释。在一些实施方案中,公开内容的语言模型可以被实现或结合预先训练的OpenAI生成预先训练-变换器(Generative Pre-Trained Transformer,GPT),如在Radford et al.,“Improving Language Understanding by Generative Pre-Training”,https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language understanding paper.pdf,2018中进一步详细描述的,其通过引用并入本文。GPT是多层变换器(参见Vaswani et al.,2017,通过引用并入本文)解码器。

[0050] 在一些实施方案中,语言模型(LM)(例如,GPT的LM)在常识QA数据集和CoS-E数据集的组合上被微调或训练。这在例如图4和5中示出。图4示出了训练CAGE语言模型(LM)或模块405以从CoS-E数据集生成解释的一个时间步骤。在一些实施方案中,语言模型可以在常识解释模块130(图1)中实现或成为其一部分。如所图示的,语言模型405是基于与回答选择令牌(token)A_1、A_2、$A_3$420和先前人类生成的解释令牌E_1、...E_{i-1} 430串联的问题令牌Q410而进行训练或调节。训练语言模型(LM)或模块405以生成解释令牌E_i 440。

[0051] 图5是根据一些实施方案的语言模块或模型505的简化图。在一些实施方案中,语言模型505可以与常识解释模块130和/或语言模型405一致。在一些示例中,语言模型505是多层神经网络。如图5所示,在一些实施方案中,该多层神经网络可以是包括嵌入模块510和变换器模块512的多层变换器编码器。在一些实施方案中,嵌入模块510可以包括嵌入层(E_1、E_2、...E_N),并且变换器模块512可以包括一层或多层变换器(Trm)。在一些实施方案中,每个变换器(Trm)可以使用长短期存储器(LSTM)实现。语言模型或模块505接收问题(Q)和回答选择形式的结构化源文本x,例如输入数据140。在一些实施方案中,结构化源文本x是自然语言形式。结构化源文本x被传递至嵌入层(E_1、E_2、...E_N),其将结构化源文本分解为令牌x_i,其中每个令牌x_i可以对应于单词、数字、标签等。在一些实施方案中,如图所示,语言模

型或模块505在变换器(Trm)层使用受约束的自我注意,其中每个令牌只能注意其左侧的语境(context)。这些仅左侧语境变换器(Trm)层共同充当文本生成的变换器解码器。生成的文本(T_1、T_2、...T_N)用于常识解释E_i。在一些实施方案中,可以使用这样的解释而推断哪个回答选择对于问题是正确的。

[0052] 鉴于来自CoS-E的人类解释或来自语言模型或模块(例如,405或505)的推理/解释,公开内容的系统和方法可以学习以执行对常识QA任务的预测。在一些实施方案中,例如图6和图7所示,分类模型或模块生成或导出针对输入问答集作出的预测。图6示出了生成预测的分类模型(CRSM)615的一个时间步骤。在一些实施方案中,分类模型可以在常识解释模块130(图1)中实现或成为其一部分。如图所示,分类模型或模块615接收与回答选择令牌A_1、A_2、A_3 620串联的问题令牌Q 610,并且生成或导出预测令牌A_1 650。

[0053] 在一些实施方案中,分类模型或模块615可以被实现或采用语言表示模型,比如来自变换器的双向编码器表示(Bidirectional Encoder Representations from Transformers,BERT)模型,如在Devlin et al.,"BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding,"arXiv preprint arXiv:1810.04805(2018年10月11日)中描述的,其通过引用并入本文。在一些实施方案中,分类模型615可以被实现或采用$BERT_{LARGE}$模型,其可以通过添加简单的二元分类器而被微调以用于多选择问答。分类器将对应于在BERT模型的所有输入开始时放置的特殊[CLS]令牌的最终状态作为输入。对于数据集中的每个示例,分类模型615构造用于微调$BERT_{LARGE}$模型的三个输入序列。说明的输入表示与问题的输入表示相同。

[0054] 图7是根据一些实施方案的分类模型或模块715的简化图。在一些实施方案中,分类模型715可以与常识解释模块130和/或分类模型615一致。在一些示例中,分类模型715是多层神经网络。如图7所示,在一些实施方案中,多层神经网络可以是包括嵌入模块710和变换器模块712的多层变换器编码器。在一些实施方案中,嵌入模块710可以包括嵌入层(E_1、E_2、...E_N),并且变换器模块712可以包括一层或多层变换器(Trm)。在一些实施方案中,可以使用长期短期存储器(LSTM)层而不是变换器层。分类模型或模块715接收问题(Q)和回答选择形式的结构化源文本x,例如输入数据140。在一些实施方案中,结构化文本还可以包括例如由训练的语言模型(例如,405或505)生成的解释。问题、回答选择和解释由输入数据中的分隔符[SEP]隔开。在一些实施方案中,每个序列是问题、分隔符令牌[SEP]以及回答选择中的一种的串联。如果方法需要来自CoS-E或如在CAGE中自动生成的解释,则分类模型或模块715将问题[SEP]、解释[SEP]和回答选择串联起来。结构化源文本x被传递至嵌入层(E_1、E_2、...E_N),其将结构化源文本分解为令牌x_i,其中每个令牌x_i可以对应于单词、数字、标签等。在一些实施方案中,如图所示,分类模型715在变换器(Trm)层使用双向自我注意,其中每个令牌可以注意其左侧和右侧的语境。这些变换器(Trm)层共同用作变换器编码器。分类模型或模块715生成或导出输入问题的回答选择的预测。

[0055] 用于生成解释和预测的两种设置或可能性可以是:(1)解释-然后-预测("推理");和(2)预测-然后-解释("合理化")。

[0056] 推理:根据图8和9说明推理。图8是示出根据一些实施方案的通过人工智能或深度学习模型生成用于推理的常识解释的系统的简化图。图9是用于系统800的对应方法900的简化图。方法900的过程910-940中的一个或多个可以至少部分地以存储在非暂时性、有形、

机器可读介质上的可执行代码的形式实现，当由一个或多个处理器运行时，可以使一个或多个处理器执行过程910-940中的一个或多个。在一些实施方案中，系统800可以在图1的计算设备100(例如，常识解释模块130)中实现，并且方法900可以由图1的计算设备100(例如，常识解释模块130)执行。

[0057] 利用推理，如图8和9所示，使用训练的CAGE语言模型805(其可以与语言模型或模块405和505一致)生成下游分类或常识推理模型(CSRM)815的解释840。

[0058] 对于训练，在过程910，语言模型805接收自然语言解释文本。在一些示例中，自然语言解释文本(例如，文本145)可以包括从人类收集或由人类开发的问题q和回答选择c0、c1、c2，以及解释e_h。

[0059] 在一些实施方案中，收集或开发来自人类的解释的任务由两部分组成。在第一部分中，人类注释器被指示在证明输出正确的问题中突出相关词语。在第二部分中，注释器被要求针对为什么预测输出正确而非其他选择提供简要的开放解释。这些指令促使注释器提供解释，这些解释实际上提供了问题背后的常见推理。在一些实施方案中，自然语言解释文本被用于训练、测试和运行语言模型805。

[0060] 通过推理，根据问题q，回答选择c0、c1、c2和人类生成的解释e_h而不是实际预测的标签或回答a来微调语言模型(LM)805。因此，训练期间的输入语境C_{RE}定义如下：

[0061] C_{RE}＝“q、c0、c1或c2?常识说”

[0062] 根据条件语言建模目标训练语言模型805以生成解释e。

[0063] 在训练了系统800(例如，语言模型805)之后，在过程920，语言模型805和分类模型或模块815接收输入数据(例如，输入数据140)。输入数据可以涉及期望应用人工智能、神经网络或深度学习模型进行分析和预测的任何情况、场景、问题等。在一些实施方案中，如图所示，输入数据可包括问题Q 810和回答选择A_1、A_2、A_3 820。

[0064] 在过程930，语言模型805生成或开发用于输入数据的潜在预测或结果的常识推理的解释E 840。这可以例如关于图4和5的语言模型405和505所描述的那样被实现。将机器生成的常识解释840提供给分类模型815。

[0065] 在过程940，分类模型或模块815(其可以与分类模型或模块615和715一致)对输入数据(例如，问题集810和回答选择820)进行操作以开发、导出或生成预测或结果850。在一些示例中，分类模型815使用机器生成的解释840支持或在其分析中应用常识推理。这可以例如关于图6和7的分类模型615和715描述的那样被实现。

[0066] 在一些实施方案中，目标是最大化：

[0067] $$\sum_i \log P(e_i|e_{i-k},\ldots,e_{i-1},C_{RE};\Theta)$$

[0068] 其中k是语境窗口的大小(在此情况下，k总是大于e的长度，使得整个解释在语境内)。条件概率P通过具有以C_{RE}和先前解释令牌为条件的参数Θ的神经网络来建模。这种解释可以称为“推理”，因为它可以在推理期间自动生成，以便为常识问答提供额外的语境。下面显示，该方法比常识QA的所报告的现有技术水平胜过10％。

[0069] 常识推理的结果和解释被提供为输出(例如，来自常识解释模块130的输出150)。

[0070] 合理化：推理的逆向方法是合理化的。关于图10和11示出合理化。图10是示出根据一些实施方案的通过人工智能或深度学习模型生成用于合理化的常识解释的系统1000的

简化图。图11是用于系统1000的对应方法1100的简化图。方法1100的过程1110-1140中的一个或多个可以至少部分地以存储在非暂时性、有形、机器可读介质上的可执行代码的形式来实现，当由一个或多个处理器运行时，可使一个或多个处理器执行过程1110-1140中的一个或多个。在一些实施方案中，系统1000可以在图1的计算设备100(例如，常识解释模块130)中实现，并且方法1100可以由图1的计算设备100(例如，常识解释模块130)来执行。

[0071] 通过合理化，如图10和11所示，分类模型或模块1015(其可以与分类模型或模块615和715一致)首先做出预测a，然后语言模型或模块1005(其可以与语言模型或模块405和505一致)基于那些标签生成解释。

[0072] 对于训练，在过程1110，分类模型1015对输入数据(例如，问题集合1010和回答选择1020)进行操作以开发、导出或生成预测或结果1050。语言模型或模块1005接收自然语言解释文本。在一些示例中，自然语言解释文本(例如，文本145)可以包括如前所述的从人收集或由人开发的问题q和回答选择c0、c1、c2以及解释e_h。

[0073] 在过程1120，语言模型1005和分类模型1015接收输入数据(例如，输入数据140)。输入数据可以涉及期望应用人工智能、神经网络或深度学习模型进行分析和预测的任何情况、场景、问题等。在一些实施方案中，如图所示，输入数据可包括问题Q 1010和回答选择A_1、A_2、A_3 1020。

[0074] 在过程1130，分类模型或模块1015对输入数据进行操作以开发、导出或生成预测或结果1050。这可以例如与图6和图7的分类模型或模块615和715的描述一致地实现。结果1050被提供给语言模型1015。

[0075] 在合理化中，在过程1140处，语言模型1015在预测标签a上的条件与输入一起生成因果合理化，或者换言之，生成用于进行预测的推理的解释。在语言模型1015的微调步骤期间，输入语境C_{RA}包含输出标签a并且如下构造：

[0076] C_{RA}＝“q、c0、c1或c2?a因为”

[0077] 语言模型1015在合理化方面的训练目标类似于推理中的训练目标，除了在这种情况下，模型1015在训练期间可以获得输入问题的真实标签。

[0078] 因为语言模型或模块1015以预测标签为条件，所以解释不被认为是常识推理。相反，它们提供了使模型更可访问和可解释的“合理化”。已经发现，这种合理化方法比现有技术模型好6％，如下所述。

[0079] 关于图8-11的系统和方法，诸如计算设备100等计算设备的一些示例可包括非暂时性、有形、机器可读介质，这些介质包括在由一个或多个处理器(例如，处理器110)运行时可使该一个或多个处理器执行方法900和1100的过程的可执行代码。可以包括方法900和1100的过程的机器可读介质的一些常见形式例如是软盘、软磁盘、硬盘、磁带、任何其它磁介质、CD-ROM、任何其它光学介质、打孔卡、纸带、具有孔图案的任何其它物理介质、RAM、PROM、EPROM、FLASH-EPROM、任何其它存储器芯片或盒式磁带，和/或处理器或计算机适于读取的任何其它介质。

[0080] 结果

[0081] 呈现了使用所提出的常识自动生成解释(CAGE)的变型的关于常识QA数据集的结果。$BERT_{LARGE}$模型用作基线，没有任何CoS-E或CAGE。

[0082] 图12示出了表格1200，该表格示出了来自常识QA、CoS-E，以及CAGE样品的示例集

合(出于理由和基本原理)。可以观察到,在一些实施方案中,CAGE-推理通常采用比CoS-E-开放更简单的构造。尽管如此,这种简单的声明性模式有时可以比CoS-E-开放更有信息性。实现CAGE的公开内容的系统和方法通过提供更明确的指导(如在表1200的最后示例1202中)或通过添加有意义的语境(如在第三示例1204中通过引入词语‘朋友’)来实现这一点。从表1200中观察到,在一些实施方案中,CAGE-推理包含该时间的43%的这些回答选择中的至少一个,在这些回答选择之中它包含该时间的21%的该模型的实际预测回答选择。这表明CAGE-推理比直接指向回答更有效。

[0083] 从表1200中可以看出,CAGE-合理化和CAGE-推理通常是相同的,或者仅仅在词序上是不同的,或者通过用另一个回答选择替换一个回答选择。仅基于CAGE-合理化42%的时间,人类可以预测回答,与CAGE-推理一样。虽然CAGE-理性化似乎比CAGE推理更好,但是我们发现在没有实际问题的情况下试图猜测正确回答时,它没有显着地改进模型的语言生成行为,即人类判断的行为。

[0084] 额外的实验设置仅使用开放的解释,其不包含来自任何回答选择的任何词语。这些解释可称为“CoS-E-限制开放”解释,因为它们在允许的词语选择上受到限制。我们观察到,即使使用这些有限的解释,也会改善BERT基线,这表明这些解释提供了有用的信息,而不仅仅是提到正确或不正确的回答。

[0085] 图13示出了示出使用仅使用常识QA输入的BERT基线实现的结果与根据公开内容的实施方案的使用包含来自CoS-E的说明的输入训练的系统和方法的比较的表1300。如表1300中所见,BERT基线模型达到64%准确度。在训练期间在问题旁边添加开放的人类解释(CoS-E-开放)导致问答模型的准确度提高2%。当在训练和验证期间进一步向模型提供通过CAGE-推理(不以真实为条件)产生的解释时,模型的准确度增加到72%。

[0086] 示出发明性方面、实施方案、实现方式或应用的此描述和附图不应被视为限制。在不偏离本说明书和权利要求书的精神和范围的情况下,可以进行各种机械的、组成的、结构的、电气的和操作的改变。在一些情况下,未详细展示或描述众所周知的电路、结构或技术以免混淆本发明的实施方案。两个或多个图中的相同数字表示相同或相似的元件。

[0087] 在本说明书中,阐述了描述根据公开内容的一些实施方案的具体细节。为了提供对实施方案的透彻理解,阐述了许多具体细节。然而,对于本领域技术人员显而易见的是,可以在没有这些具体细节中的一些或全部的情况下实施一些实施方案。本文所公开的具体实施方案旨在说明而非限制。本领域的技术人员可以认识到尽管未在此具体描述,但在公开内容的范围和精神内的其他元件。此外,为了避免不必要的重复,结合一个实施方案显示和描述的一个或多个特征可以结合到其他实施方案中,除非另外具体描述或者如果一个或多个特征将使实施方案不起作用。

[0088] 尽管已经示出并描述了示例性实施方案,但是在前述公开中以及在一些情况下,可以考虑更宽范围的修改、改变和替换,可以采用实施方案的一些特征而不对应地使用其他特征。本领域普通技术人员将认识到许多变化、替代和修改。因此,本发明的范围应当仅由所附权利要求来限定,并且适当的是,以与在此公开的实施方案的范围一致的方式宽泛地解释权利要求。

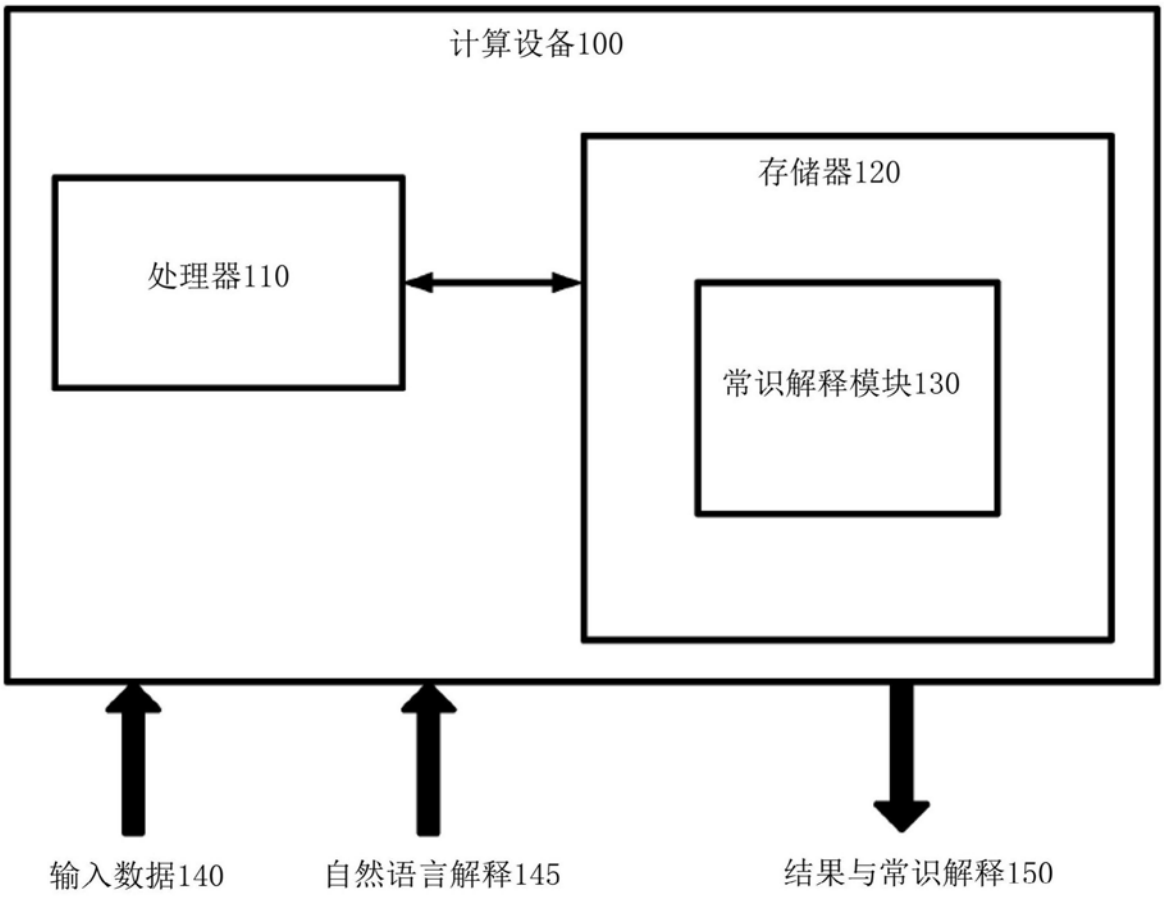

计算设备100
存储器120
处理器110
常识解释模块130
输入数据140
自然语言解释145
结果与常识解释150

图1

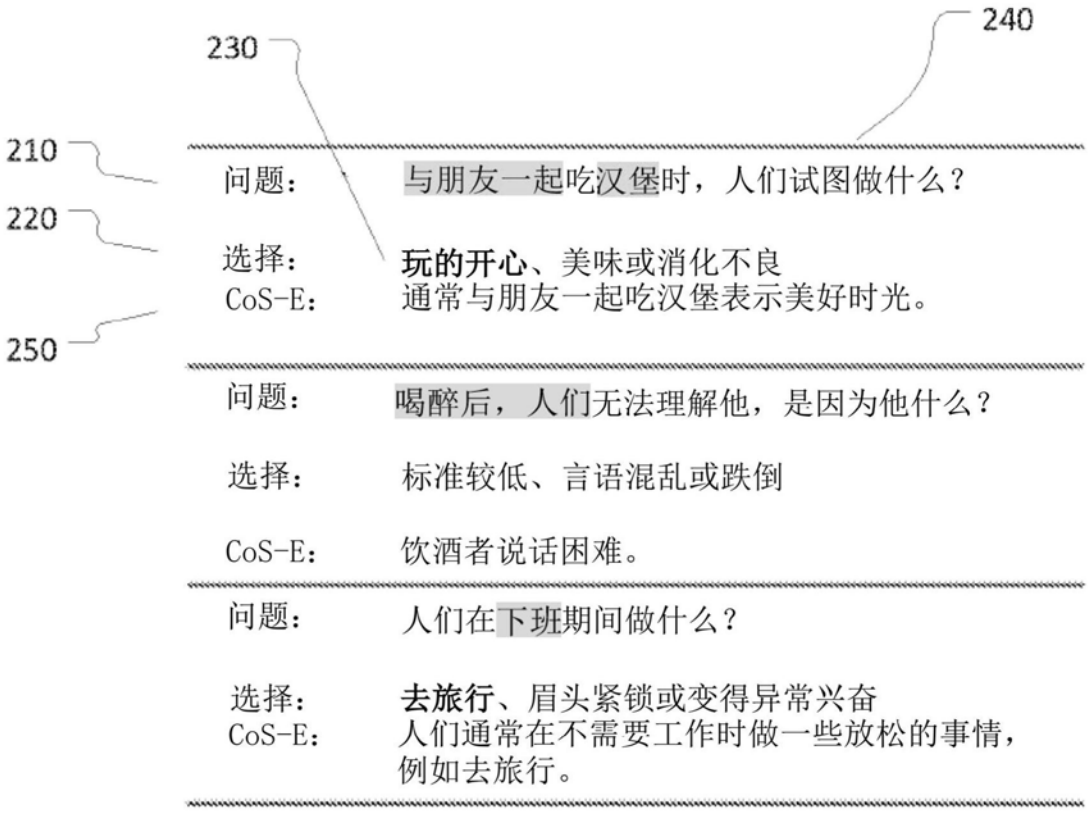


表1：来自我们的CoS-E数据集的示例。

图2

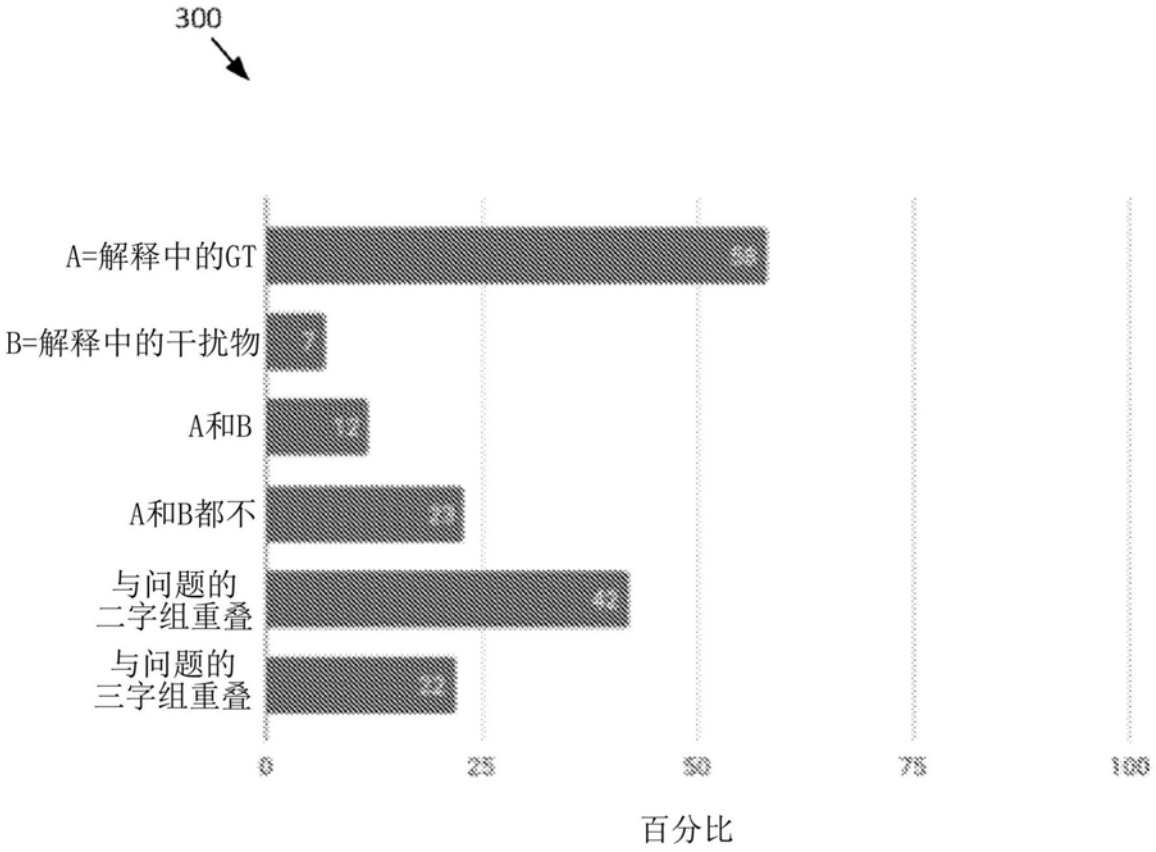

300
A=解释中的GT
B=解释中的干扰物
A和B
A和B都不
与问题的
二字组重叠
与问题的
三字组重叠
0
25
50
75
100
百分比

图3

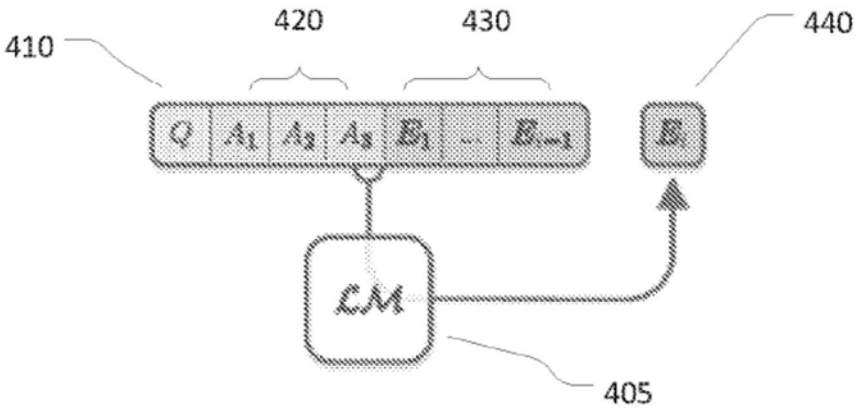

410
420
430
440
Q
A_1
A_2
A_3
E_1
...
E_{t-1}
E_t
CM
405

图4

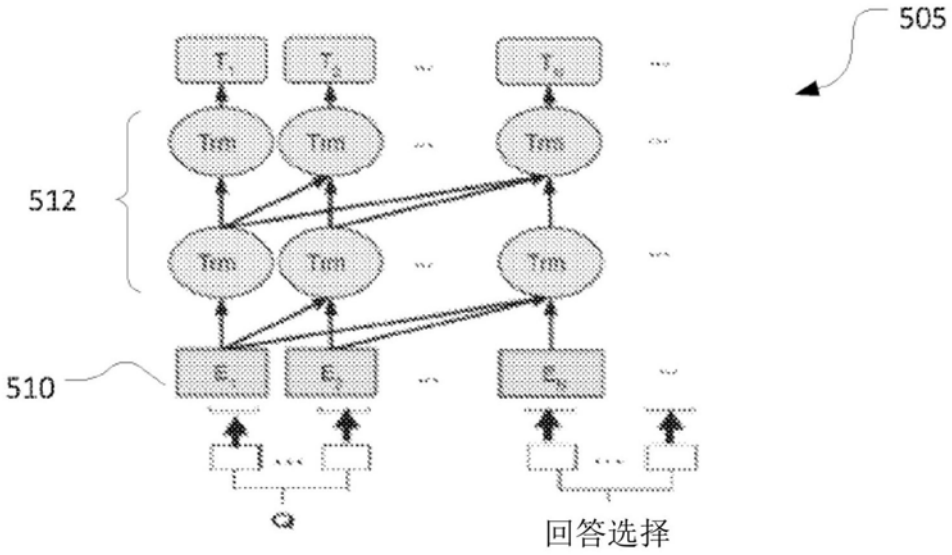


图5

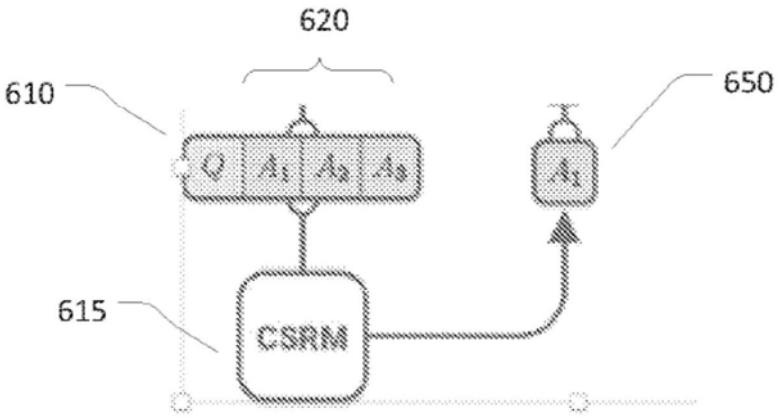


图6

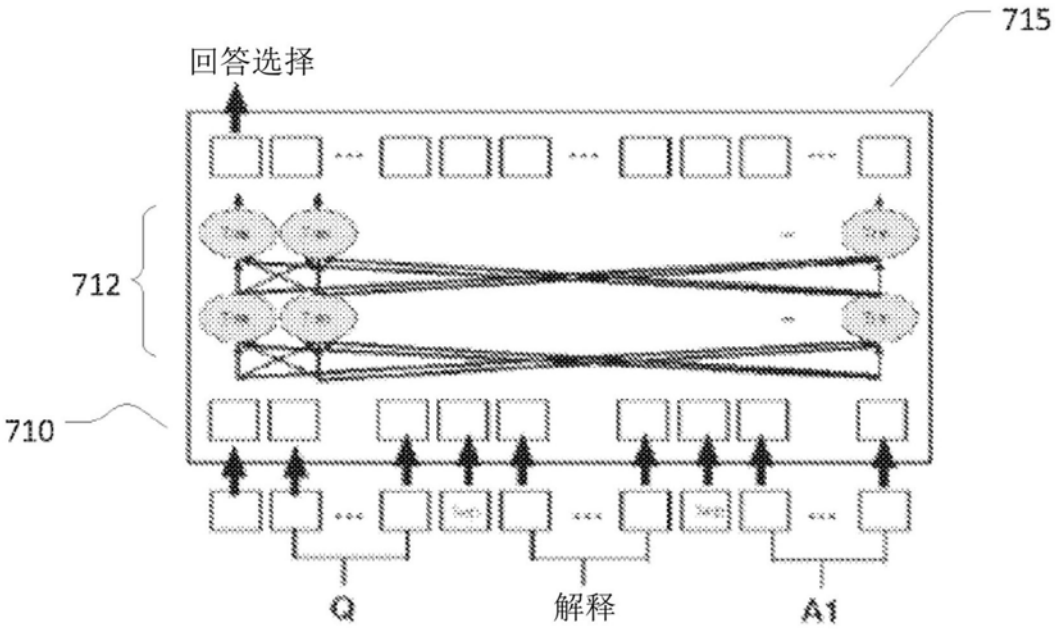

715
回答选择
712
710
Q
解释
A1

图7

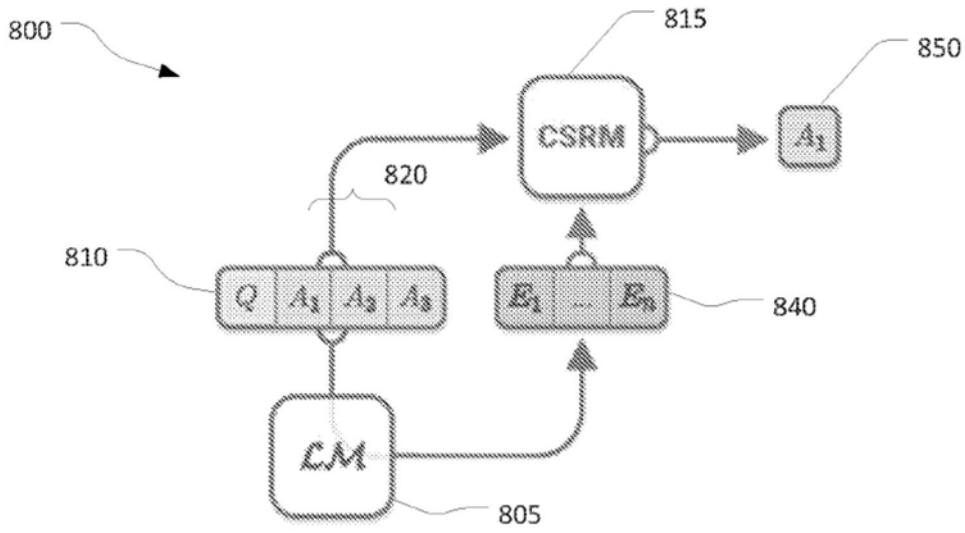

800
815
850
CSRM
820
810
Q
A1
A2
A3
E1
...
En
840
LM
805

图8

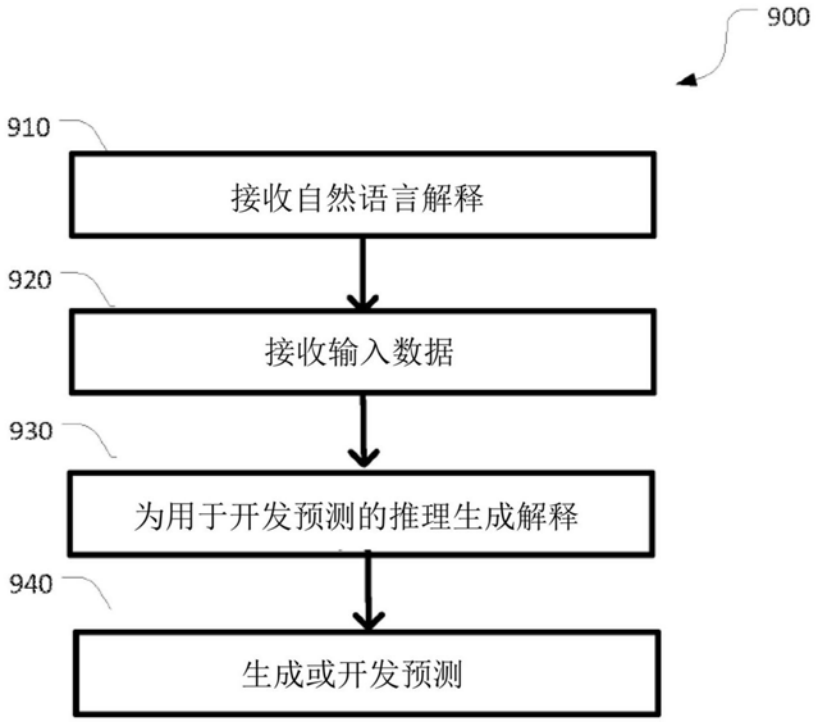

900
910
接收自然语言解释
920
接收输入数据
930
为用于开发预测的推理生成解释
940
生成或开发预测

图9

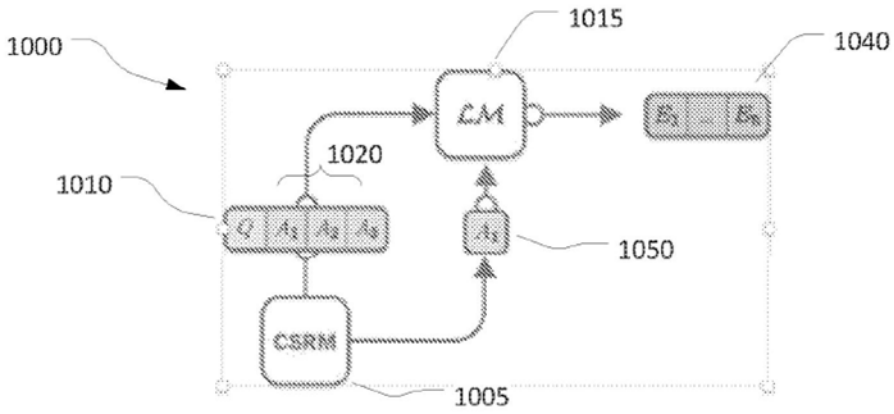

1000
1015
1040
CM
1020
1010
1050
CSRM
1005

图10

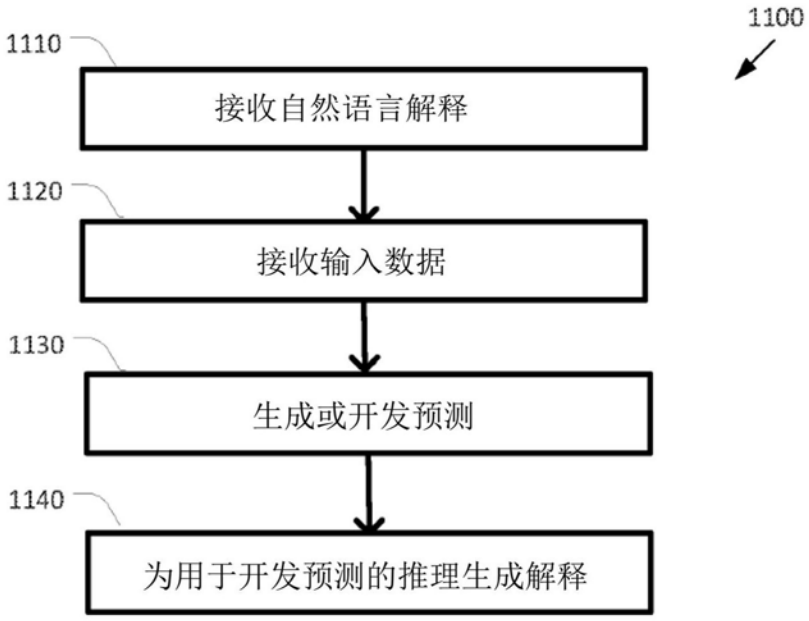

1100
1110
接收自然语言解释
1120
接收输入数据
1130
生成或开发预测
1140
为用于开发预测的推理生成解释

图11

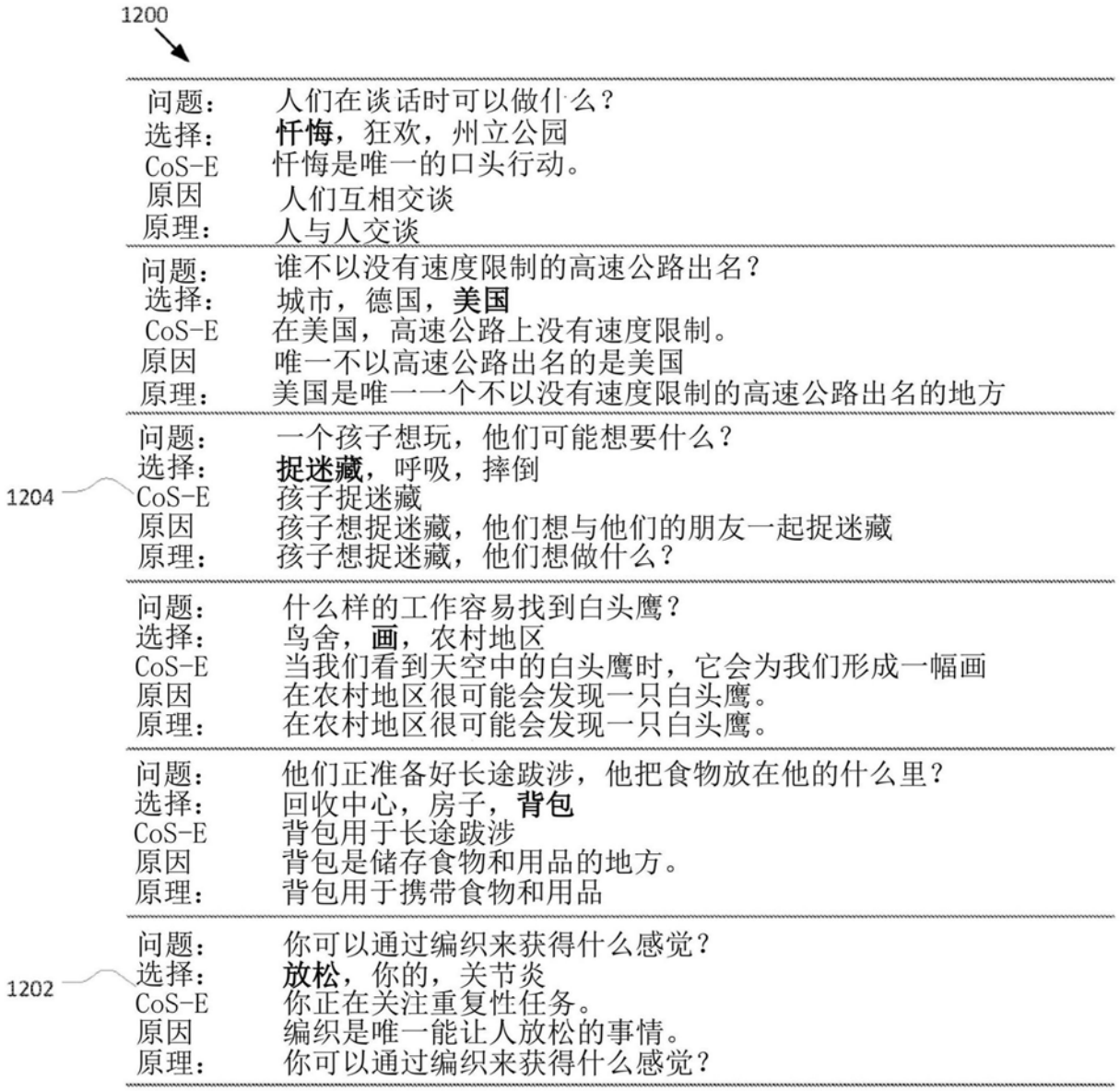

1200
问题： 人们在谈话时可以做什么？
选择： **忏悔**，狂欢，州立公园
CoS-E 忏悔是唯一的口头行动。
原因 人们互相交谈
原理： 人与人交谈
问题： 谁不以没有速度限制的高速公路出名？
选择： 城市，德国，**美国**
CoS-E 在美国，高速公路上没有速度限制。
原因 唯一不以高速公路出名的是美国
原理： 美国是唯一一个不以没有速度限制的高速公路出名的地方
问题： 一个孩子想玩，他们可能想要什么？
选择： **捉迷藏**，呼吸，摔倒
1204
CoS-E 孩子捉迷藏
原因 孩子想捉迷藏，他们想与他们的朋友一起捉迷藏
原理： 孩子想捉迷藏，他们想做什么？
问题： 什么样的工作容易找到白头鹰？
选择： 鸟舍，**画**，农村地区
CoS-E 当我们看到天空中的白头鹰时，它会为我们形成一幅画
原因 在农村地区很可能会发现一只白头鹰。
原理： 在农村地区很可能会发现一只白头鹰。
问题： 他们正准备好长途跋涉，他把食物放在他的什么里？
选择： 回收中心，房子，**背包**
CoS-E 背包用于长途跋涉
原因 背包是储存食物和用品的地方。
原理： 背包用于携带食物和用品
问题： 你可以通过编织来获得什么感觉？
选择： **放松**，你的，关节炎
1202
CoS-E 你正在关注重复性任务。
原因 编织是唯一能让人放松的事情。
原理： 你可以通过编织来获得什么感觉？

图12

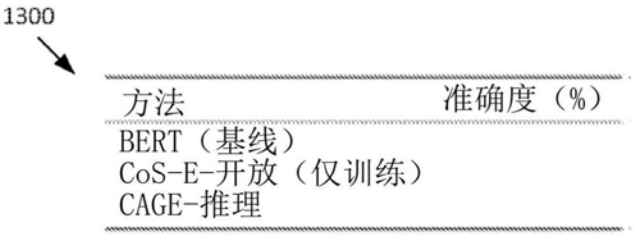

1300
方法 准确度（%）
BERT（基线）
CoS-E-开放（仅训练）
CAGE-推理

图13