(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119539078 A
(43)申请公布日 2025.02.28

(21)申请号 202411599133.5
(22)申请日 2024.11.11
(71)申请人 香港中文大学(深圳)
地址 518172 广东省深圳市龙岗区龙城街道龙翔大道2001号
(72)发明人 王方鑫 吴攀龙
(74)专利代理机构 成都巾帼知识产权代理有限公司 51260
专利代理师 邢伟
(51)Int.Cl.
G06N 5/04(2023.01)
G06N 5/022(2023.01)
G06N 3/006(2023.01)
G06N 3/0455(2023.01)
G06N 3/098(2023.01)
G06F 16/3329(2025.01)

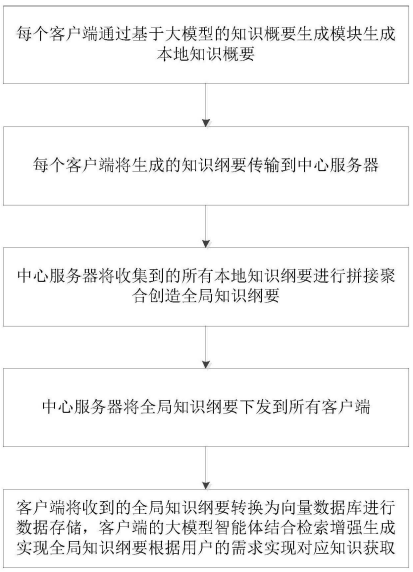
权利要求书2页 说明书4页 附图2页

(54)发明名称

一种基于联邦上下文学习和大模型智能体的知识查询方法

(57)摘要

本发明公开了一种基于联邦上下文学习和大模型智能体的知识查询方法,包括以下步骤:S1.每个客户端通过基于大模型的知识概要生成模块生成本地知识概要;S2.每个客户端将生成的知识纲要传输到中心服务器;S3.中心服务器将收集到的所有本地知识纲要进行拼接聚合创造全局知识纲要;S4.中心服务器将全局知识纲要下发到所有客户端;S5.客户端将收到的全局知识纲要转换为向量数据库进行数据存储,客户端的大模型智能体结合检索增强生成实现全局知识纲要根据用户的需求实现对应知识获取。本发明基于检索增强生成的知识学习使用模型设计,有效解决了大模型智能体联邦学习的带宽、计算需求,数据异构问题,能够提供高效且性能优良的联邦学习过程。



CN 119539078 A

1.一种基于联邦上下文学习和大模型智能体的知识查询方法,其特征在于:包括以下步骤:

S1.每个客户端通过基于大模型的知识概要生成模块生成本地知识概要;

S2.每个客户端将生成的知识纲要传输到中心服务器;

S3.中心服务器将收集到的所有本地知识纲要进行拼接聚合创造全局知识纲要;

S4.中心服务器将全局知识纲要下发到所有客户端;

S5.客户端将收到的全局知识纲要转换为向量数据库进行数据存储,客户端的大模型智能体结合检索增强生成实现全局知识纲要根据用户的需求实现对应知识获取。

2.根据权利要求1所述的一种基于联邦上下文学习和大模型智能体的知识查询方法,其特征在于:所述步骤S1包括:

S101.本地工具知识纲要指令生成,包含:

系统指令,含有用户问题和相应的答案工具名称生成指令;

工具详细描述生成指令,用于帮助用户清楚地了解工具的主要功能和用途;

工具使用场景生成指令,用于帮助大模型智能体了解工具的有效性,并确保选择合适的工具来解决问题;

工具注意事项生成指令,用于指导大模型智能体如何有效使用工具,避免错误请求,从而提高应用程序的性能;

工具交互协同生成指令,用于指导大模型智能体对工具的连锁调用来解决问题;

S102.将本地工具知识纲要指令输入大模型,大模型根据输入指令进行知识纲要生成:

采用的大模型为DeepSeek-v2,当接收到本地工具知识纲要指令后,大模型基于预训练得到的丰富的知识生成相应领域的知识纲要。

3.根据权利要求2所述的一种基于联邦上下文学习和大模型智能体的知识查询方法,其特征在于:所述知识纲要的内容包括:工具名称、工具描述、工具使用场景注意事项、工具输入数据的数据格式、工具调用的版本信息以及调用工具需要的权限。

4.根据权利要求1所述的一种基于联邦上下文学习和大模型智能体的知识查询方法,其特征在于:所述步骤S3包括:

中心服务器从各个客户端收集所有用户的本地知识纲要$\zeta_1,\zeta_2,\cdots,\zeta_n$,接着将收集到的本地知识纲要进行拼接,得到全局知识纲要ζ_{global}:

$\zeta_{global}=\{\zeta_1,\zeta_2,\ldots,\zeta_n\}$。

5.根据权利要求1所述的一种基于联邦上下文学习和大模型智能体的知识查询方法,其特征在于:所述步骤S5包括:

S501:使用基于深度神经网络的嵌入模型bge-large-en-v1_5,将全局知识纲要储存为向量数据库,处理后的知识纲要被转换成多维向量表示;

此嵌入模型基于神经网络架构来学习语义和结构信息,将语义相似度和上下文关系嵌入到向量表示中;

将生成的向量数据存储到向量数据库中,以便于后续的存储和检索操作;

向量数据库同时支持实时更新、批量插入、以及复杂的查询优化机制,

S502:当接收到用户请求时,首先使用嵌入模型bge-large-en-v1_5将用户请求转换成为多维向量;接着使用转换后的多维向量和上一步储存在向量数据库中的知识纲要向量进

行相似度计算：

依据预设阈值或排序机制选出最匹配的知识条目作为上下文信息,上下文信息含有完成请求所需的具体工具和技术的相关知识,包含了所需要工具的知识纲要信息；

接着将选定的上下文信息与原始用户请求相结合,生成一个包含相关知识纲要信息和具体要求的输入序列；

大模型智能体接收到相关的输入序列后,通过上下文学习能够获取到相关使用工具需要的知识,从而达到工具学习的目的；

大模型输出工具调用所需要的工具名称和输入参数,进行对应工具的调用来完成用户的请求。

6.根据权利要求5所述的一种基于联邦上下文学习和大模型智能体的知识查询方法,其特征在于:所述相似度计算使用余弦相似度或Jaccard相似度的度量标准来评估请求向量与存储向量之间的相似程度。

一种基于联邦上下文学习和大模型智能体的知识查询方法

技术领域

[0001] 本发明涉及大模型智能体的联邦学习领域,特别是涉及一种基于联邦上下文学习和大模型智能体的知识查询方法。

背景技术

[0002] 大模型往往拥有数十亿以上的参数量,并且在海量的数据上经过训练,这使得大模型具有很多传统小模型不具有的能力,其中就包括了作为智能体使用工具的能力。通过工具的使用,大模型能够在日常生活中为人类提供很多帮助,包括天气查询,旅行计划安排等等。然而大模型智能体在下游任务重的表现常受限于高质量数据的稀缺性。大多数数据都是本地存储且私有的,这限制了大型语言模型智能体通过更多数据的训练来提升性能。联邦学习作为一种有前景的方法应运而生,它能够在不直接交换私有数据的情况下,实现多个客户端之间模型的协同改进。在联邦学习中,通过聚合不同客户端的模型权重来促进知识共享,这种方式有效保护了数据隐私。然而,大模型参数量极大的特性导致联邦学习中大模型参数的聚合所需要的带宽远远超出了主流设备的网络传输能力,客户端的计算也给算力设备带来了严峻挑战,并且不同用户的数据异构问题使得联邦学习的效果容易受到显著的负面影响。因此,优化联邦学习中的通信,计算开销和缓解数据异构问题带来的影响成为当务之急。

[0003] 当前的研究已经提出了几种策略来优化,包括基于LoRA等高效参数微调方法的联邦学习算法,基于正则化的联邦学习算法等。然而目前流行的算法在大模型智能体联邦学习的场景下都不能高效并且有效地完成训练。为了解决这些问题,我们提出了首个具有隐私保护性的基于上下文学习的大模型智能体联邦学习算法。FICAL包括一下创新点:设计基于大模型的本地知识纲要生成模块,进行具有隐私保护性的知识纲要生成。提出知识纲要的聚合与分发代替传统联邦学习中的参数聚合,实现通信开销大幅降低。基于检索增强生成的工具学习与使用模块设计,指导大模型智能体通过上下文学习进行知识的摄取和工具的使用。

[0004] 与传统的联邦联邦学习相比,大模型智能体的联邦学习面临几个关键的挑战:1)高带宽消耗与现代通信系统之间的不匹配。例如,流行的大模型如LLaMA3.1-405B在典型的100Mbps通信网络速率下,需要在两个分布式节点之间传输超过十个小时。在联邦学习中,客户端与中央服务器之间每轮共享大模型参数所需要的数据量对现代通信系统构成了巨大负担。2)高计算消耗与现代计算硬件之间存在不匹配。流行的现代大模型通常拥有数十亿个以上的参数,这比传统模型大数千倍之多,而硬件的发展速度无法跟上这一急剧增长的步伐。训练这些大模型计算密集,不仅导致训练时间显著延长,而且还因为需要获取具有高处理能力和大内存容量的GPU而产生了显著的成本。3)不同客户端上的异质数据分布。由于用户地理位置或行业的差异,不同客户端可能拥有非独立同分布(non-IID)的数据分布。这可能损害模型参数的聚合,并进一步加剧联邦学习过程的复杂性。

发明内容

[0005] 本发明的目的在于克服现有技术的不足,提供一种基于联邦上下文学习和大模型智能体的知识查询方法,基于检索增强生成的知识学习使用模型设计,有效减小了大模型智能体联邦学习的带宽、计算需求,并考虑了不同客户端的异质数据分布问题,能够提供高效且性能优良的联邦学习过程。

[0006] 本发明的目的是通过以下技术方案来实现的:一种基于联邦上下文学习和大模型智能体的知识查询方法,包括以下步骤:

[0007] S1.每个客户端通过基于大模型的知识概要生成模块生成本地知识概要;

[0008] S2.每个客户端将生成的知识纲要传输到中心服务器;

[0009] S3.中心服务器将收集到的所有本地知识纲要进行拼接聚合创造全局知识纲要;

[0010] S4.中心服务器将全局知识纲要下发到所有客户端;

[0011] S5.客户端将收到的全局知识纲要转换为向量数据库进行数据存储,客户端的大模型智能体结合检索增强生成实现全局知识纲要根据用户的需求实现对应知识获取。

[0012] 本发明的有益效果是:本发明基于检索增强生成的知识学习使用模型设计,有效降低了大模型智能体联邦学习的带宽、计算需求,并解决数据异构问题(在这里是指不同客户端的数据分布不同,比如有的用户对于某一个工具的数据比较多,有的用户对另一个工具的数据比较多,数据分布的不同往往会造成联邦学习效果的下降),能够提供高效且性能优良的联邦学习过程:通过知识纲要的传递,用户能够在互相保护隐私的条件下学习到不同工具的使用方法,从而获取知识,并且在大模型的场景下大大减少了通信开销,提高了大模型智能体联邦学习的通信效率和计算效率和算法有效性。

附图说明

[0013] 图1为本发明的方法流程图;

[0014] 图2为工具知识纲要生成指令架构图。

具体实施方式

[0015] 下面结合附图进一步详细描述本发明的技术方案,但本发明的保护范围不局限于以下所述。

[0016] 本发明的目的在于有效处理大模型智能体联邦学习的高带宽,高计算量,数据异构等挑战,以提高联邦学习的算法效果和降低资源消耗。本发明从架构角度包括本地知识纲要生成模块,全局知识纲要生成模块,工具学习与利用模块。技术层面上,本发明采用了联邦学习技术,上下文学习技术,检索增强生成技术与大模型技术,充分利用了大模型的上下文学习能力,通过联邦学习和上下文学习的结合,实现了用户请求应答的性能提升,并且极大地减少了联邦学习中的通信消耗。联邦学习利用但是不会泄露不同用户的数据。在本申请中体现为利用了不同用户发出的知识纲要,但是不会泄露用户本身的数据(即每个用户私有的“问题-回答”数据),具体地:

[0017] 如图1所示,一种基于联邦上下文学习和大模型智能体的知识查询方法,包括以下步骤:

[0018] S1.每个客户端通过基于大模型的知识概要生成模块生成本地知识概要;

[0019] S101:本地工具知识纲要指令生成。包含系统指令,含有用户问题和相应的答案(包括要使用的工具的名称和正确使用工具的输入参数)的示例,工具名称生成指令,工具详细描述生成指令,工具使用场景生成指令,工具注意事项生成指令,工具交互协同生成指令。其中工具详细描述生成指令用于帮助用户清楚地了解工具的主要功能和用途。工具使用场景用于帮助大模型智能体了解工具在特定情况下的有效性,并确保选择合适的工具来解决特定的问题。工具注意事项指导大模型智能体如何有效使用工具,避免错误请求,从而提高应用程序的性能。工具交互协同信息指导大模型智能体对工具的连锁调用来解决问题。工具知识纲要生成指令架构如图2所示。

[0020] S102:本地工具知识纲要指令输入大模型,大模型根据输入指令进行知识纲要生成。本发明利用了先进的人工智能技术中的大模型架构,以实现高效且精确的知识纲要生成。大模型是指具有大量参数的深度学习模型,通常经过大规模数据集预训练,以捕捉复杂的数据模式。本发明所采用的大模型DeepSeek-v2是一种基于Transformer架构的混合专家模型,在自然语言处理领域表现出色具备长距离依赖捕捉能力(支持128K tokens的上下文长度)同时具有经济的训练成本和高效的推理能力。当接收到用户请求时,混合专家模型根据用户请求所属不同领域智能稀疏激活不同的专家进行回答,达到推理的加速与成本的降低。该大模型预先在海量文本数据上进行了训练,以学习语言规则、上下文关系以及其他复杂的语义特征。当接收到本地知识纲要时,文本首先会被送入模型的输入处理系统,其中文本将被分词并转换为token ID,接着转换为嵌入向量。接着模型使用Rotary Position Embedding(ROPE)进行位置编码。模型通过Multi-head Latent Attention(MLA)模块有效地捕捉序列中不同位置的单词之间的依赖关系,灵活地处理序列中不固定长度的本地工具知识纲要指令。当接收到本地工具知识纲要指令后,大模型能够迅速理解指令意图,并基于其预训练得到的丰富的知识生成相应领域的知识纲要,显著提升生成内容的质量和准确性。知识纲要的内容包括以下几个部分:工具名称;工具描述,这部分内容包含了对于工具的详细描述,比如工具支持的平台,工具设计被用于解决什么问题,工具可以在哪些操作系统或者硬件上运行,工具如何实施数据加密,隐私保护等安全措施,工具的使用流程等等;工具使用场景,这部分包含了有哪些可能的情况下这个工具应该被使用来提升大模型智能体回复质量;注意事项,这部分内容包含了工具使用需要注意的细节,如工具请求的频率限制,工具输入数据的数据格式,工具调用的版本信息,调用工具需要的权限等等。

[0021] S2.每个客户端将生成的知识纲要传输到中心服务器;

[0022] S3.中心服务器将收集到的所有本地知识纲要进行拼接聚合创造全局知识纲要;

[0023] 中心服务器从联邦学习系统(系统可以认为是中心服务器和客户端)的用户中收集所有用户的本地知识纲要$\zeta_1,\zeta_2,\cdots,\zeta_n$。接着将收集到的本地知识纲要进行拼接得到全局知识纲要ζ_{global}:

[0024] $\zeta_{global}=\{\zeta_1,\zeta_2,\cdots,\zeta_n\}$

[0025] S4.中心服务器将全局知识纲要下发到所有客户端;

[0026] S5.客户端将收到的全局知识纲要转换为向量数据库进行数据存储,客户端的大模型智能体结合检索增强生成实现全局知识纲要根据用户的需求实现对应知识获取。

[0027] S501:使用基于深度神经网络的嵌入模型bge-large-en-v1_5,将全局知识纲要储存为向量数据库,处理后的知识纲要被转换成多维向量表示。此嵌入模型基于神经网络架

构(如Transformer)来学习语义和结构信息,将语义相似度和上下文关系嵌入到向量表示中。将生成的向量数据存储到向量数据库中,以便于后续的存储和检索操作。向量数据库同时支持实时更新、批量插入、以及复杂的查询优化机制,确保大规模知识条目的存储和检索具备高效性、可扩展性和准确性。通过该过程,实现了全局知识条目的向量化管理,使得在向量空间中能够快速、高效地完成相关知识的匹配、检索和推理操作,极大提高了知识库的可用性和智能化水平。

[0028] S502:当接收到用户请求时,首先使用嵌入模型bge-large-en-v1_5将用户请求转换成为多维向量。接着使用转换后的多维向量和上一步储存在向量数据库中的知识纲要向量进行相似度计算。这里使用余弦相似度或Jaccard相似度等度量标准来评估请求向量与存储向量之间的相似程度,以确定最相关的知识条目。依据预设阈值或排序机制选出最匹配的知识条目作为上下文信息,这些信息含有完成特定请求所需的具体工具和技术的相关知识,包含了特定所需要工具的知识纲要信息,比如工具名称,工具描述,工具的使用场景,工具的使用注意事项等等。接着将选定的上下文信息与原始用户请求相结合,生成一个包含相关知识纲要信息和具体要求的输入序列。大模型智能体接收到相关的输入序列后,通过上下文学习能够获取到相关使用工具需要的知识,从而达到工具学习的目的。大模型输出工具调用所需要的工具名称和输入参数,进行对应工具的调用来完成用户的请求。

[0029] 在本专利申请的实施例中,相比于传统方案,本发明的性能显著超越现有的解决方案。在大量实验中,本发明在多个关键指标上均实现了显著提升,将本发明的方法记为FICAL。具体而言,FICAL在工具调用准确率上比现有的流行的联邦学习算法FedACG,FedDecorr,FedLoRA,FedAdam,FedYogi,FedProx在默认的设定下达到了26.09%～43.48%的工具调用准确率提升。同时相比于基线算法中随着模型参数量线性增加的O(N)的通信开销,FICAL将线性开销减小为O(1),并且在LLaMA-3 8b模型的使用设定下达到了3.33×10^5倍的通信开销降低。

[0030] 最后应说明的是:以上所述仅为本发明的优选实施例而已,并不用于限制本发明,尽管参照前述实施例对本发明进行了详细的说明,对于本领域的技术人员来说,其依然可以对前述各实施例所记载的方法进行修改,例如所述方法名称的变化等。凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

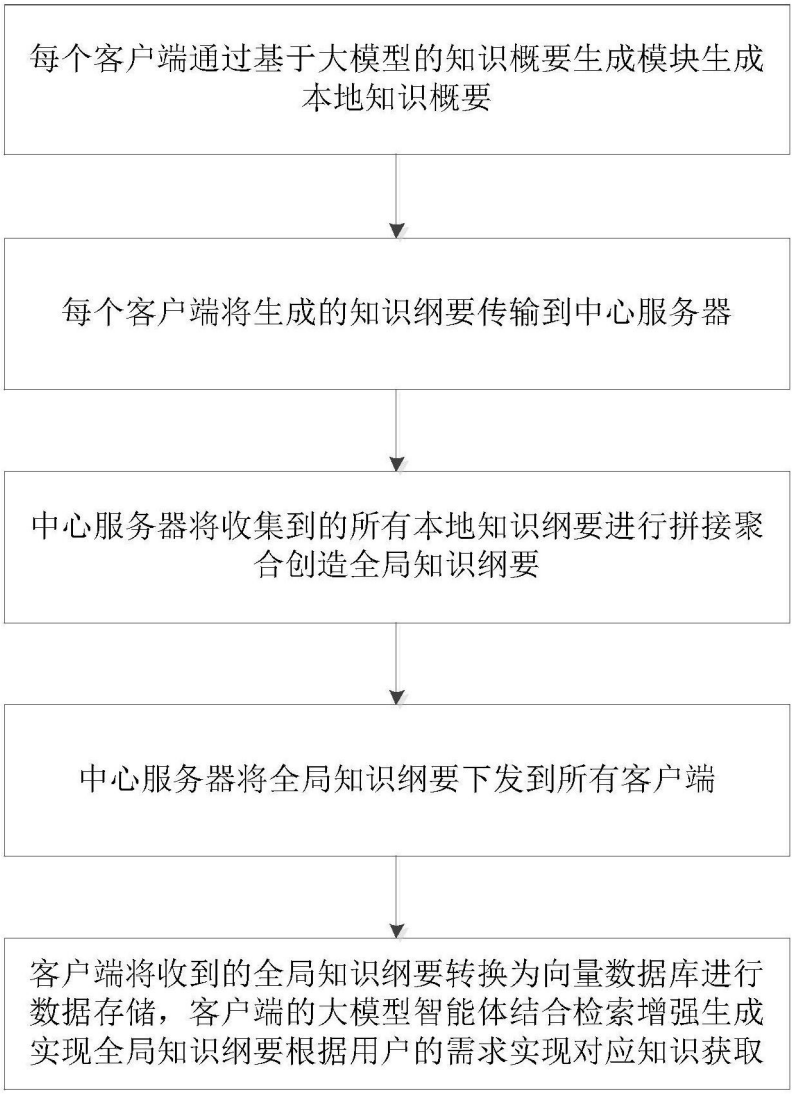

每个客户端通过基于大模型的知识概要生成模块生成本地知识概要
每个客户端将生成的知识纲要传输到中心服务器
中心服务器将收集到的所有本地知识纲要进行拼接聚合创造全局知识纲要
中心服务器将全局知识纲要下发到所有客户端
客户端将收到的全局知识纲要转换为向量数据库进行数据存储，客户端的大模型智能体结合检索增强生成实现全局知识纲要根据用户的需求实现对应知识获取

图1

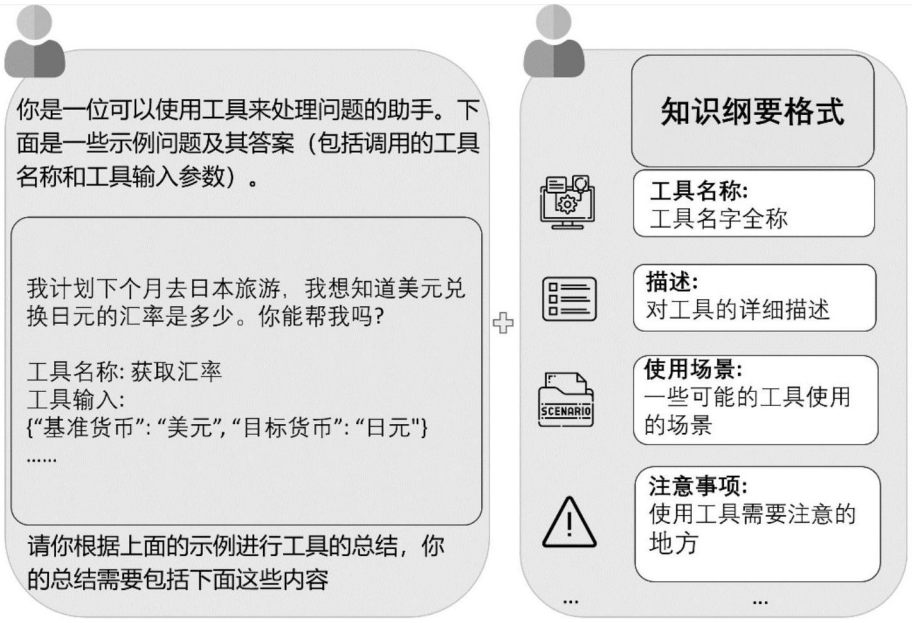

你是一位可以使用工具来处理问题的助手。下面是一些示例问题及其答案（包括调用的工具名称和工具输入参数）。
我计划下个月去日本旅游，我想知道美元兑换日元的汇率是多少。你能帮我吗?
工具名称: 获取汇率
工具输入:
{“基准货币”: “美元”, “目标货币”: “日元"}
.....
请你根据上面的示例进行工具的总结，你的总结需要包括下面这些内容
知识纲要格式
工具名称:
工具名字全称
描述:
对工具的详细描述
使用场景:
一些可能的工具使用的场景
注意事项:
使用工具需要注意的地方
...
...

图2