(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120086355 A
(43)申请公布日 2025.06.03

(21)申请号 202510525474.6

(22)申请日 2025.04.25

(71)申请人 深圳平安通信科技有限公司
地址 518000 广东省深圳市前海深港合作区南山街道兴海大道3048号前海自贸大厦2701(04单元)

(72)发明人 瞿晓阳 王健宗 陶伟 卢昊骋

(74)专利代理机构 广东良马律师事务所 44395
专利代理师 刘毋凡

(51)Int.Cl.
G06F 16/334(2025.01)
G06F 16/35(2025.01)
G06F 40/30(2020.01)
G06N 5/04(2023.01)

权利要求书3页 说明书28页 附图4页

(54)发明名称

模型量化推理加速方法、装置、设备及介质

(57)摘要

本发明涉及人工智能技术领域,可应用于医疗健康及金融科技等业务场景中,公开了一种模型量化推理加速方法、装置、设备及介质,包括:将输入文本划分为多个处理块,对非首个处理块进行重要性评分,按评分结果分配计算精度格式,确定每个处理块的统一量化配置;将网络模块划分为配置共享组,组内共享对应处理块的量化配置;根据统一量化配置执行块级量化推断,生成模型推理结果。本发明通过基于token重要性分数统一确定每个处理块的量化配置,并在网络模块组内复用该配置,实现了块级别的精度分配与并行量化推理,在保障推理精度的同时大幅降低显存开销和配置时间开销,有效提升长文本推理任务中的执行效率与显存利用率。

CN 120086355 A

1.一种模型量化推理加速方法,其特征在于,包括以下步骤:

将输入文本划分为多个处理块,将首个处理块的处理精度格式固定为高精度格式,并禁用对所述首个处理块的量化处理;

对所述多个处理块中除首个处理块以外的其他处理块,通过语言模型生成每个其他处理块的自注意力矩阵,并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和,并将所述全体元素数值之和作为每个token位置的重要性分数;

将重要性分数大于第一阈值的token位置分配为高精度格式,将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式,将重要性分数小于第二阈值的token位置分配为低精度格式;

统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量,选择数量最多的精度格式作为对应处理块的统一量化配置;

将所述语言模型的网络模块划分为多个配置共享组,每个配置共享组至少包含两个网络模块;

在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块;

根据每个处理块对应的统一量化配置,对所有处理块执行块级批量量化并完成模型推理,生成模型推理结果。

2.如权利要求1所述的模型量化推理加速方法,其特征在于,将输入文本划分为多个处理块,将首个处理块的处理精度格式固定为高精度格式,并禁用对所述首个处理块的量化处理,包括:

根据预设块长度将输入文本划分为多个等长的处理块;

对首个处理块的所有token位置禁用量化参数调整;

将所述首个处理块在语言模型中的嵌入层、自注意力层及前馈网络层的处理精度格式固定为高精度格式;

当输入文本的末尾存在不足预设块长度的剩余token时,将所述剩余token组成的文本片段作为独立处理块;

对所述独立处理块填充无效token至所述预设块长度,并将填充后的独立处理块的处理精度格式固定为高精度格式,且禁用对所述独立处理块的量化操作;

记录所有处理块的起始位置索引和结束位置索引。

3.如权利要求1所述的模型量化推理加速方法,其特征在于,对所述多个处理块中除首个处理块以外的其他处理块,通过语言模型生成每个其他处理块的自注意力矩阵,并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和,并将所述全体元素数值之和作为每个token位置的重要性分数,包括:

将每个其他处理块输入至语言模型的多头自注意力层,得到每个其他处理块对应的多个注意力头的局部自注意力矩阵;

对多个注意力头的局部自注意力矩阵进行加权平均或算术平均处理,生成对应每个其他处理块的最终自注意力矩阵;

从所述最终自注意力矩阵中提取每个token位置对应的列向量;

确定每个列向量中全体元素的数值之和,并将所述数值之和作为每个列向量对应的

token位置的重要性分数;

若其他处理块中存在填充的无效token位置,则在确定所述填充的无效token位置的重要性分数时,将无效token位置对应列向量的所有元素数值置零。

4.如权利要求1所述的模型量化推理加速方法,其特征在于,将重要性分数大于第一阈值的token位置分配为高精度格式,将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式,将重要性分数小于第二阈值的token位置分配为低精度格式,包括:

根据处理块的长度确定第一阈值和第二阈值;

若存在无效token位置,则将无效token位置的处理精度格式分配为低精度格式,并在统计每个处理块的精度格式数量时排除所述无效token位置;

将每个有效token位置的重要性分数与所述第一阈值和第二阈值比较;

将重要性分数大于第一阈值的有效token位置的处理精度格式分配为高精度格式;

将重要性分数处于第二阈值之上且处于第一阈值之下的有效token位置的处理精度格式分配为中等精度格式;

将重要性分数小于第二阈值的有效token位置的处理精度格式分配为低精度格式。

5.如权利要求1所述的模型量化推理加速方法,其特征在于,统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量,选择数量最多的精度格式作为对应处理块的统一量化配置,包括:

遍历当前处理块的所有token位置,识别被分配为高精度格式、中等精度格式及低精度格式的token位置;

在统计过程中,若存在无效token位置,则排除所有无效token位置,仅统计有效token位置的精度格式分配结果;

分别统计有效token位置中被分配为高精度格式、中等精度格式及低精度格式的计数,生成高精度计数、中等精度计数及低精度计数;

比较所述高精度计数、中等精度计数及低精度计数的数值大小;

将数值最大的计数对应的精度格式作为所述当前处理块的统一量化配置;

若存在多个精度格式的计数相同且为最大值,则选择多个精度格式中的最高精度格式作为所述当前处理块的统一量化配置。

6.如权利要求1所述的模型量化推理加速方法,其特征在于,在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块,包括:

为每个处理块分配唯一块标识符,并为每个配置共享组分配唯一组标识符;

在配置映射表中建立唯一组标识符与唯一块标识符的绑定关系;

基于所述唯一组标识符与唯一块标识符的绑定关系,将每个配置共享组的第一个网络模块与对应处理块的统一量化配置参数相关联;

将所述统一量化配置参数写入每个配置共享组的共享内存区域,并为每个配置共享组的所有网络模块分配相同的内存地址映射;

同一配置共享组内的其他网络模块在执行量化处理时,通过所述内存地址映射从所述共享内存区域读取所述统一量化配置参数。

7.如权利要求1所述的模型量化推理加速方法,其特征在于,根据每个处理块对应的统一量化配置,对所有处理块执行块级批量量化并完成模型推理,生成模型推理结果,包括:

为每个处理块加载对应的统一量化配置参数,所述统一量化配置参数包括精度格式标识;

根据所述精度格式标识,为不同处理块配置独立的处理核函数;

在图形处理器或张量处理器的并行处理单元中,根据处理块的索引顺序分配处理资源,并发执行所有处理块的量化处理;

对每个处理块的处理结果进行块内token位置校验,以剔除无效token位置对应的中间结果;

将所有处理块的有效中间结果按起始位置索引排序,拼接为完整的模型输出序列;

对拼接后的输出序列执行后处理操作,生成最终的模型推理结果。

8.一种模型量化推理加速装置,其特征在于,所述模型量化推理加速装置包括:

输入文本预处理模块,用于将输入文本划分为多个处理块,将首个处理块的处理精度格式固定为高精度格式,并禁用对所述首个处理块的量化处理;

自注意力分析模块,用于对所述多个处理块中除首个处理块以外的其他处理块,通过语言模型生成每个其他处理块的自注意力矩阵,并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和,并将所述全体元素数值之和作为每个token位置的重要性分数;

精度分配模块,用于将重要性分数大于第一阈值的token位置分配为高精度格式,将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式,将重要性分数小于第二阈值的token位置分配为低精度格式;

量化配置决策模块,用于统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量,选择数量最多的精度格式作为对应处理块的统一量化配置;

网络模块分组控制模块,用于将所述语言模型的网络模块划分为多个配置共享组,每个配置共享组至少包含两个网络模块;

配置共享管理模块,用于在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块;

推理执行模块,用于根据每个处理块对应的统一量化配置,对所有处理块执行块级批量量化并完成模型推理,生成模型推理结果。

9.一种计算机设备,其特征在于,所述计算机设备包括存储器、处理器以及存储至所述存储器上并可以在所述处理器上运行的模型量化推理加速程序,所述模型量化推理加速程序被所述处理器执行时实现如权利要求1-7中任一项所述的模型量化推理加速方法的步骤。

10.一种计算机可读存储介质,其特征在于,所述存储介质上存储有模型量化推理加速程序,所述模型量化推理加速程序被处理器执行时实现如权利要求1-7中任一项所述的模型量化推理加速方法的步骤。

模型量化推理加速方法、装置、设备及介质

技术领域

[0001] 本发明涉及人工智能技术领域,尤其涉及一种模型量化推理加速方法、装置、设备及存储介质。

背景技术

[0002] 近年来,大语言模型(LLM)在自然语言处理任务中表现出色,广泛应用于对话系统、机器翻译、问答系统等场景。然而,随着模型规模的持续扩展,其在执行长文本推理任务时面临显著的显存占用问题。大多数主流LLM拥有数十亿甚至数千亿参数,推理过程中需将模型及相关中间状态完全加载至GPU显存中,对硬件资源提出了极高要求。

[0003] 当应用于长文本推断任务时,例如长篇文档生成、复杂文献摘要提取或多轮对话处理,输入序列往往包含数千至上万个token,这种超长序列显著增加了注意力机制的计算复杂度与显存需求。具体而言,注意力矩阵的维度随输入序列长度的平方增长,大量的中间激活值也需临时存储于显存中,极易导致内存溢出或推断失败。现有的显存优化手段,如梯度检查点、激活值卸载等,主要针对训练阶段,难以直接适用于推理过程中。此外,分布式推理虽然可缓解单设备的资源瓶颈,但在实际部署中对基础设施要求高、配置复杂,且难以在边缘端或中低资源环境下稳定运行。

[0004] 在医疗健康业务领域,LLM被用于病历摘要生成、医疗问答系统和诊疗记录分析等场景。这些任务往往涉及大量医学术语与上下文关联信息,对模型在长文本下的推理能力依赖性更强。当前采用的通用量化压缩方法虽然可降低显存占用,但往往忽略了医疗文本中某些关键信息的保留,导致模型精度下降,甚至出现医疗信息理解偏差,影响结果可信度。

[0005] 在金融科技业务领域,大语言模型被广泛用于合同解析、财务摘要生成、客户风险评估等文本密集型任务。这些任务通常涉及结构复杂、文本超长的合规报告或历史交易记录,对推理过程中的显存效率提出极高要求。然而,当前模型在处理金融文档等长文本输入时容易因显存不足而失败,严重影响了模型的稳定性与可扩展性。

[0006] 为缓解显存压力,量化技术已成为主流的压缩手段之一,尤其在推理场景中被广泛采用。现有研究尝试使用混合精度量化方法,即将高重要性的token分配更高位宽,而低重要性的token使用低位宽,从而在保持模型精度的同时减少总体显存使用。然而,混合精度量化在推理过程中面临配置选择效率低的问题。由于当前方法需在每轮推理过程中动态分析每个token的重要性并确定位宽策略,这一过程本身耗时严重,削弱了量化所带来的速度优势,特别是在长文本任务中影响更为显著。

[0007] 综上所述,现有技术在应对大语言模型长文本推理过程中的显存开销问题方面仍存在关键不足,特别是在量化配置策略的效率、跨任务领域的重要性识别机制及块级推理资源调度等方面,仍需进一步优化以满足金融科技与医疗健康等领域对高效、精准推理的实际需求。

发明内容

[0008] 本发明的主要目的在于提供一种模型量化推理加速方法、装置、设备及存储介质，旨在解决现有技术在推理过程中需要为每个token确定量化位宽，导致量化配置效率低下，尤其在长文本推理任务中严重影响推理速度与显存优化效果的技术问题。

[0009] 为实现上述目的，本发明提供一种模型量化推理加速方法，包括：

将输入文本划分为多个处理块，将首个处理块的处理精度格式固定为高精度格式，并禁用对所述首个处理块的量化处理；

对所述多个处理块中除首个处理块以外的其他处理块，通过语言模型生成每个其他处理块的自注意力矩阵，并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和，并将所述全体元素数值之和作为每个token位置的重要性分数；

将重要性分数大于第一阈值的token位置分配为高精度格式，将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式，将重要性分数小于第二阈值的token位置分配为低精度格式；

统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量，选择数量最多的精度格式作为对应处理块的统一量化配置；

将所述语言模型的网络模块划分为多个配置共享组，每个配置共享组至少包含两个网络模块；

在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块；

根据每个处理块对应的统一量化配置，对所有处理块执行块级批量量化并完成模型推理，生成模型推理结果。

[0010] 进一步地，为实现上述目的，本发明提供一种模型量化推理加速装置，包括：

输入文本预处理模块，用于将输入文本划分为多个处理块，将首个处理块的处理精度格式固定为高精度格式，并禁用对所述首个处理块的量化处理；

自注意力分析模块，用于对所述多个处理块中除首个处理块以外的其他处理块，通过语言模型生成每个其他处理块的自注意力矩阵，并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和，并将所述全体元素数值之和作为每个token位置的重要性分数；

精度分配模块，用于将重要性分数大于第一阈值的token位置分配为高精度格式，将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式，将重要性分数小于第二阈值的token位置分配为低精度格式；

量化配置决策模块，用于统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量，选择数量最多的精度格式作为对应处理块的统一量化配置；

网络模块分组控制模块，用于将所述语言模型的网络模块划分为多个配置共享组，每个配置共享组至少包含两个网络模块；

配置共享管理模块，用于在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块；

推理执行模块，用于根据每个处理块对应的统一量化配置，对所有处理块执行块

级批量量化并完成模型推理,生成模型推理结果。

[0011] 进一步地,为实现上述目的,本发明还提供一种计算机设备,所述计算机设备包括存储器、处理器以及存储至所述存储器上并可在所述处理器上运行的模型量化推理加速程序,所述模型量化推理加速程序被所述处理器执行时实现如上述所述的模型量化推理加速方法的步骤。

[0012] 进一步地,为实现上述目的,本发明还提供一种计算机可读存储介质,所述存储介质上存储有模型量化推理加速程序,所述模型量化推理加速程序被处理器执行时实现如上所述的模型量化推理加速方法的步骤。

[0013] 有益效果:本发明涉及人工智能技术领域,可应用于医疗健康及金融科技等业务场景中,公开了一种模型量化推理加速方法,包括:将输入文本划分为多个处理块,固定首个处理块的计算精度格式为高精度格式并禁用量化操作;针对除首个处理块外的其他处理块,通过语言模型生成自注意力矩阵,并依据每个token位置在自注意力矩阵中对应列的数值总和计算重要性分数;基于两个预设阈值将各token位置分配为高精度、中等精度或低精度格式;统计每个处理块内各精度格式token数量,选取数量最多的精度格式作为统一量化配置;将网络模块划分为多个配置共享组,并在组内共享处理块的统一量化配置;依据处理块的统一量化配置执行块级批量量化并完成模型推理,生成推理结果。本发明通过基于token重要性分数统一确定每个处理块的量化配置,并在网络模块组内复用该配置,实现了块级别的精度分配与并行量化推理,在保障推理精度的同时大幅降低显存开销和配置时间开销,有效提升长文本推理任务中的执行效率与显存利用率。

附图说明

[0014] 下面将结合附图及实施例对本发明作进一步说明,附图中:

图1为本发明一实施例中模型量化推理加速方法的一应用环境示意图;

图2为本发明模型量化推理加速方法一实施例的流程示意图;

图3为本发明模型量化推理加速装置较佳实施例的功能模块示意图;

图4为本发明一实施例中计算机设备的一结构示意图;

图5为本发明一实施例中计算机设备的另一结构示意图。

具体实施方式

[0015] 应当理解,此处所描述的具体实施例仅用以解释本发明,并不用于限定本发明。

[0016] 本发明实施例提供的模型量化推理加速方法,可应用在如图1的应用环境中,其中,用户端通过网络与服务端进行通信。服务端可以通过用户端将输入文本划分为多个处理块,固定首个处理块的计算精度格式为高精度格式并禁用量化操作;针对除首个处理块外的其他处理块,通过语言模型生成自注意力矩阵,并依据每个token位置在自注意力矩阵中对应列的数值总和计算重要性分数;基于两个预设阈值将各token位置分配为高精度、中等精度或低精度格式;统计每个处理块内各精度格式token数量,选取数量最多的精度格式作为统一量化配置;将网络模块划分为多个配置共享组,并在组内共享处理块的统一量化配置;依据处理块的统一量化配置执行块级批量量化并完成模型推理,生成推理结果。本发明通过基于token重要性分数统一确定每个处理块的量化配置,并在网络模块组内复用该

配置,实现了块级别的精度分配与并行量化推理,在保障推理精度的同时大幅降低显存开销和配置时间开销,有效提升长文本推理任务中的执行效率与显存利用率。其中,用户端可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备。服务端可以用独立的服务器或者是多个服务器组成的服务器集群来实现。下面通过具体的实施例对本发明进行详细的描述。

[0017] 请参阅图2,图2为本发明提供的模型量化推理加速方法一实施例的流程示意图。需要说明的是,虽然在流程图中示出了逻辑顺序,但是在某些情况下,可以以不同于此处的顺序执行所示出或描述的步骤。

[0018] 如图2所示,本发明提出的模型量化推理加速方法包括如下步骤:

S10,将输入文本划分为多个处理块,将首个处理块的处理精度格式固定为高精度格式,并禁用对所述首个处理块的量化处理;

在本实施例中,将输入文本划分为多个处理块是为了控制长文本输入在模型推理中的计算资源分布,并对模型的激活值、注意力矩阵和显存使用进行局部管理。在实际应用中,文本处理块可以通过固定长度的滑动窗口进行生成,例如每512个token划为一块,该长度依据模型的训练窗口或推理窗口的最大处理长度确定。该划分方式来源于序列建模中广泛使用的窗口分片机制,其目的是将长文本转化为局部上下文块,既保证局部语义连贯,又避免一次性载入全部token带来的计算瓶颈。

[0019] 将首个处理块的处理精度格式固定为高精度格式,是为了在推理初始阶段为模型提供稳定的计算起点。高精度格式一般指的是浮点格式,如FP32或FP16,其相较于低位宽定点格式(如INT8或INT4)在数值表达范围与表达精度上具备更强的鲁棒性与容错性。

[0020] 禁用首个处理块的量化处理,是为了进一步确保该段输入在全精度表达下运行,避免量化误差在模型传播初期对后续块传播路径产生累积性干扰。禁用量化不仅包括不对该处理块的权重和激活值应用量化操作,还包括跳过该处理块的量化配置生成过程。其核心目的是提供量化决策中的参考基准,使得首个处理块成为后续块重要性评分与精度配置判断的相对锚点。

[0021] 实际应用中,量化处理通常涉及对模型的权重参数进行离散化处理,并将激活值压缩到有限位宽表达域中。若对首个处理块仍执行此操作,会导致初始attention结果失真,从而降低后续token位置的精度分配准确度。因此,禁用首个处理块的量化处理,不仅是一个数值稳定性考量,更是整个推理路径中动态精度调整策略的基础前提。

[0022] 由于模型通常采用残差连接和归一化机制,高精度首块有助于稳定前几层的统计行为,进而提高后续处理块在量化策略决策中的判别准确率。该特征还具备良好的泛化性,不依赖于具体模型结构,因此可适用于GPT、T5、BERT等不同架构的大语言模型中。

[0023] 在具体实现中,可以将输入文本加载为token序列后,按预设的处理块长度进行分段,每段构成一个处理块。首个处理块的token范围可以为token序列的前512个token,划分完成后,为该处理块配置计算精度格式为FP16或FP32。在模型推理时,跳过该块的量化配置生成过程,即不参与重要性评分、不执行位宽决策。为了禁用量化处理,在调用模型执行时,针对首个处理块,强制绕过模型中部署的量化算子。这可以通过配置推理引擎的量化mask参数实现,也可以在模型图转换过程中,将该块的数据路径设定为非量化路径。进一步地,可以在该处理块经过模型的嵌入层、多头注意力层及前馈层时,全部保持浮点精度执行,不

调用低精度计算核函数。在实际的模型运行环境中,如果采用TensorRT等推理框架,可在构建engine时标记首块为常驻高精度区域,或通过编译时添加首块精度锁定指令实现。在多块执行流程中,首个处理块的输出还可作为参考输入,用于后续处理块的重要性分析与精度分配策略生成。

[0024] 示例说明:在医疗健康业务领域,当处理包含长篇医学诊断记录的文本时,往往存在关键病症描述出现在文首的情况。如果首个处理块由于量化而丢失信息,可能导致模型未能正确提取疾病症状与对应分析。在采用固定首块高精度处理后,模型可以更精准捕获首段高价值内容,有利于后续生成准确的诊断总结或病程预测。

[0025] 在金融科技业务领域,当处理一段完整的风险报告时,报告开头通常包含全局风险分类与总结信息。若将其置于低精度计算路径中,容易导致模型误判风险等级,从而影响后续判断过程。通过对首个处理块保持高精度计算并禁用量化,可以有效捕捉报告核心要点,确保风控模型推理准确度与稳定性,提升金融服务决策的可靠性与安全性。

[0026] 通过在推理过程中固定首个处理块为高精度格式并禁用其量化处理,可以显著提升后续处理块在精度策略判别中的稳定性与准确性。高精度首块提供了完整的上下文表示能力,使得其输出可作为权重分配与量化级别选择的基准,同时避免了初始信息传播因量化误差而造成的模型偏移。在长文本场景中,这一策略可有效降低由于首段误判而导致的全局精度退化问题,从而提升整体推理性能与稳定性。

[0027] S20,对所述多个处理块中除首个处理块以外的其他处理块,通过语言模型生成每个其他处理块的自注意力矩阵,并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和,并将所述全体元素数值之和作为每个token位置的重要性分数;

在本实施例中,对于多个处理块中除首个处理块以外的其他处理块,引入自注意力机制生成对应的自注意力矩阵,主要用于度量同一处理块内不同token位置之间的信息依赖关系。在自然语言处理任务中,大语言模型(Large Language Model)通常采用基于Transformer结构的自注意力机制来构建token之间的上下文表示。通过对每个处理块单独构造自注意力矩阵,可以局部化计算范围,从而有效控制内存占用并提升计算效率。

[0028] 每个token位置对应列的全体元素数值之和被用作重要性分数,该操作本质上是一种列向量求和,其技术含义是衡量某个token在该处理块中的整体注意力集中程度。该数值越大,表示该token与其他token之间的联系越强,其语义信息在块内传播范围更广,从而被认为更“重要”。

[0029] 在自注意力计算中,每个token位置对应列向量表示该token作为目标token时对所有源token分配的注意力权重。在此基础上,对列向量求和可得其全局关注程度。通过这种方式获得的重要性分数不会依赖于具体任务标签,从而具有良好的通用性,适用于机器翻译、问答系统、信息抽取等多种任务中对token选择性精度控制的需求。

[0030] 为避免引入不必要的噪声和冗余计算,自注意力矩阵在生成时应当跳过填充token,即对paddingtoken位置可采取mask策略,将其对应的注意力值强制为零,防止其干扰其他token的重要性评估。这一设计在多头注意力机制(Multi-head Attention,多头注意力)中同样适用,各个注意力头计算的局部矩阵可先进行平均融合,统一为最终的自注意力矩阵。

[0031] 在一种具体实施方式中,每个处理块的token序列被送入多头自注意力模块进行

前向传播,得到多个注意力头的注意力矩阵。通过将这些局部矩阵进行加权平均或简单算术平均,生成当前处理块的最终自注意力矩阵。随后,遍历该矩阵的每一列,对列向量中的所有数值进行求和操作,作为该列所对应的token位置的重要性分数。在此过程中,填充token所对应的列向量将被替换为全零向量,从而其重要性分数恒为零,确保其不参与后续精度配置的选择。

[0032] 对于计算性能要求较高的场景,可将自注意力矩阵的生成与列向量求和操作融合为一个GPU内核函数(CUDA Kernel),在GPU并行框架下批量处理多个处理块,提升整体处理效率。此外,对于attention矩阵中数值精度较低的情况(如FP16),可以引入归一化或数值平滑机制,避免token分数受数值漂移影响而产生误判。

[0033] 示例说明:在医疗健康领域,医生的问诊记录、病人自述、检测报告等文本中,往往包含大量token,而真正对后续决策产生影响的词语是其中极少数高度专业化且上下文高度相关的医学术语。以一段电子病历为例,其中可能包含病人年龄、基础疾病、主诉、检查结果等信息片段。在生成该处理块的自注意力矩阵时,例如"肝功能异常"这一词组,其对应的token位置在矩阵中与"ALT升高""黄疸""乙肝表面抗原阳性"等token位置之间呈现出较强的注意力连接。通过计算该token对应列的元素和,可以量化其在整个片段中的语义重要性。进而,这些高分token将被标记为重要token,为后续的诊断分类或自动生成病历摘要任务提供更高的计算精度支持。而类似"病人性别""此次复诊"等通用性高但上下文语义权重低的token,其对应列的元素和较小,自然被赋予较低的重要性分数,从而在后续推理中以较低精度计算,节省算力资源而不影响核心语义的表达。

[0034] 在金融领域的智能客服对话、用户投诉处理或交易日志解析中,重要的上下文信息通常分布在包含"风险""冻结""诈骗""延迟到账"等词汇的token中,而这类词汇的上下文依赖性通常也较高。在自注意力矩阵中,"诈骗"这个token与"转账失败""资金冻结""陌生联系人"等token形成强连接,其对应列向量的元素和显著高于其他背景性词汇如"你好""请问""谢谢"等token,从而在该步骤中被标注为高重要性位置。在后续的风险判断模型推理中,重点token被分配为高精度计算,确保对敏感表达的理解不因低位宽量化而失真,尤其在语义边界模糊、表达不规范的用户输入中尤为关键。通过这种基于注意力的token级精度分配策略,可以在不牺牲风控敏感性和精准度的前提下,有效压缩推断时的显存占用和计算负载。

[0035] 这类示例场景表明,在长文本任务中,通过自注意力矩阵捕捉token之间的语义耦合程度,并基于列向量求和确定重要性,不仅实现了token级精度控制的技术目标,同时具备跨领域的通用适应性和高推理效率保障。

[0036] 通过对每个处理块的自注意力矩阵进行逐列求和并计算重要性分数,可以基于token的上下文依赖强度精确识别出对语义表达贡献最大的token位置。在此基础上进一步进行精度分配时,不再依赖外部特征提取或复杂规则匹配,大幅降低推断阶段的前处理计算开销,实现重要token的识别与精度控制的统一。

[0037] S30,将重要性分数大于第一阈值的token位置分配为高精度格式,将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式,将重要性分数小于第二阈值的token位置分配为低精度格式;

在本实施例中,将每个token位置的重要性分数用于驱动后续计算精度的配置,是

实现差异化量化策略的关键路径。这里的“重要性分数”来源于前一阶段基于自注意力矩阵生成的、用于反映每个token在其上下文中的信息影响力的数值表示。该分数越高,说明该token在当前处理块中被其他token关注的程度越高,即它在语义传播中的作用越强,语义遮蔽的容忍度越低。因此,该值直接决定token在推理中所需的计算精度。

[0038] 高精度格式通常对应浮点计算格式,例如半精度浮点(FP16)或单精度浮点(FP32),在计算复杂度与表达精度之间取得平衡。中等精度格式可以是中位宽定点量化,例如INT4,能够在保证主要语义表达准确性的同时节省一部分显存资源与运算开销。低精度格式则通常采用更小位宽的定点量化格式,例如INT2或更进一步的INT1(二值化)表示,其适用于冗余性较高、语义容错性强的token位置,显著压缩计算负载。

[0039] 通过设置第一阈值与第二阈值实现分级控制,其中第一阈值大于第二阈值。第二阈值用于识别最低重要性的token,而第一阈值识别最重要的一批token,二者之间的token则落入中等精度范围。第一阈值与第二阈值的设置原则可以根据处理块长度、模型规模和当前硬件资源动态设定,也可通过经验性模型训练确定。这种双阈值机制使得精度配置既具有可控的策略灵活性,也具备对计算资源的细粒度调度能力。

[0040] 将分数与两个阈值进行比较的过程,通常可在预处理阶段的权重计算线程中并行执行,每个token位置通过一次浮点比较操作即可确定其精度等级,之后将其标注为对应的精度标签,以供后续配置生成与推理分配使用。整个分配机制不依赖于模型结构的修改,只作用于量化策略层,具备良好的系统兼容性与部署适应性。

[0041] 可以通过静态阈值设定方式实现分级精度配置。例如,第二阈值取0.2,第一阈值取0.8,token重要性分数在[0,1]范围内均匀分布。则大于0.8的token分配为FP16,中间段(0.2-0.8之间,包含0.8及0.2)采用INT4,低于0.2则为INT2。还可以基于每个处理块的分数分布进行动态阈值计算,例如采用分位数(quantile)法,将前10%作为高精度,后30%作为低精度,中间为中等精度。此外,也可以融合任务感知因子,对特定语义标注类任务中已知重要token优先标为高精度,即进行注意力引导下的精度增强。

[0042] 在具体实现上,每个token位置记录一个精度标签(例如2位标识:00低精度、01中等、10高精度),以位图形式存储于显存中,量化配置模块读取该标识后调用相应位宽的量化核函数完成低层推理绑定。多个token可批量标记,借助SIMD(Single Instruction Multiple Data)指令集并发分配,提高吞吐效率。对于分布式场景,可在每个GPU上本地完成精度分配标记并上传配置bitmap至共享控制模块,实现跨块一致性调度。

[0043] 通过在token级别按重要性分数分配精度格式,可以实现计算资源的重点倾斜,即将更高的算力用于关键token,降低对低权重信息的精度资源投入。在模型整体保持高推理准确率的同时,有效控制了显存开销与推理延迟。

[0044] S40,统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量,选择数量最多的精度格式作为对应处理块的统一量化配置;

在本实施例中,在多级精度分配完成后,需要对每个处理块内部不同精度格式的分布情况进行统计,以便为后续的统一量化配置选择提供决策基础。每个token位置在经过前序重要性分数比较后已经被分配为高精度格式、中等精度格式或低精度格式,该步骤的目标是通过遍历整个处理块的有效token位置,分别统计三类精度格式对应的token位置数量。

[0045] 每个处理块的量化配置必须统一,以匹配硬件计算单元中块级并行执行的调度需求。例如,在GPU的张量核心(Tensor Core)执行INT8或FP16等计算时,要求每次操作针对统一的矩阵精度,因此需要预先确定处理块的整体计算格式。该步骤通过量化后的精度统计结果,确定当前处理块的主导精度分布。

[0046] 遍历过程中无效token位置不参与统计,以避免填充token对真实分布造成偏差。实际实现中通常利用在填充阶段记录的mask掩码,快速排除掉padding区域,仅对有效位置的精度标签进行计数。在计数完成后,比较三类精度格式的数量大小,选择其中数量最多的精度格式作为该处理块的统一量化配置。

[0047] 若在某些特殊情况下出现两个或三个精度格式的token数量相同且为最大值,为了确保模型推理精度的稳定性与保守性,优先选择精度等级更高的格式。例如,高精度与中等精度数量相等时,选择高精度格式作为该处理块的量化配置。这种保守偏向策略确保模型在不增加计算负担的前提下,尽可能保留关键语义路径的信息表达能力。

[0048] 该操作不仅是一种基于统计主导趋势的配置策略,也是一种确保处理块内部量化策略一致性的技术机制。通过这一机制,每个处理块被赋予一个唯一的统一量化配置,从而在后续推理中调用对应核函数并实现硬件级块并行加速。

[0049] 可以使用位图结构表示每个token位置的精度标签,每一位代表当前token的精度类别,例如使用两位二进制标识分别表示高、中、低精度。在统计阶段,使用逻辑与运算提取每个处理块的精度标志位,根据掩码跳过无效token位置,分别对三种精度标记进行位计数。

[0050] 在多核系统中,统计操作可分块并行执行,每个线程处理一个处理块的统计任务,并将结果写入共享内存或量化配置寄存表。在共享内存中保存三类计数值后,执行一次比较操作判断最大值,并依据优先级策略生成精度配置标识。若采用动态精度调整策略,可以在统计完成后同步调整当前处理块的动态量化位宽设置。

[0051] 为了进一步提高处理块统一量化配置的智能性与适应性,在基于精度格式统计结果选择精度配置的基础上,可以引入多种语义与历史上下文特征,对统计计数结果进行加权微调,以反映语言模型对局部语义变化的敏感性,并动态适应推理任务中不同语义区域对计算资源的实际需求。

[0052] 首先,在处理自然语言中的长文本输入时,存在token语义密度随上下文语义场变化而呈现出不均匀分布的现象。以法律合同、医学检验报告等高结构化文档为例,其文本片段可能出现大量重复、高频且信息冗余的token(如术语、连接符、单位名等),这类token虽然在字面上频繁出现,但在实际语义上往往并不构成信息重点。如果仅依据attention计算得出的原始重要性分数进行精度格式划分,可能造成统计过程中中等或低精度token数量偏少,从而误将该类处理块配置为高精度,增加计算冗余。

[0053] 为此,可以引入基于滑动窗口的高频token重要性抑制机制。该机制在统计过程中对出现在多个相邻处理块中的token进行频率计数,如果某个token在滑动窗口内的出现频率超过预设阈值(如70%块中出现),则对该token的重要性分数适当降低一个比例因子(如乘以0.8),从而在统计时更倾向于将其划分为中等或低精度格式,避免高频冗余token主导整体精度配置选择。该滑动窗口可以为长度3~5的处理块序列,窗口长度可根据输入总长度动态调整,以兼顾响应速度与局部特征捕捉能力。

[0054] 其次,可以引入前序语义块影响因子。由于处理块之间往往具有强语义连贯性,如在金融舆情分析中,“信用卡逾期”处理块之后紧跟“催收、违约、征信”类语义块,在统计精度时应参考前一处理块中的精度倾向,从而增强当前块与上文语义一致性的精度配置策略。实现方式上,可以记录前一处理块中各类精度格式token数量所占比例,并在当前处理块精度计数基础上引入加权因子进行调整,例如当前中等精度token数量为200,若前一块中高精度比例占比高,则可将当前高精度计数乘以1.1的前向调整因子以提升其配置选择可能性。

[0055] 再者,针对某些任务的特定需求,还可以引入上下文权重感知模型,通过轻量级前向网络模块(如一层FFN+Softmax)动态生成当前处理块的精度配置权重向量。输入特征包括该块原始三类精度token计数、滑动窗口高频token比例、前序块精度分布等,通过前向模型学习到的权重调整每一类精度的相对权重,从而更细致地控制最终精度配置的选择逻辑。例如,在医疗健康文本生成场景下,该感知模型能够识别“病理状态转折”或“治疗反应变化”时点,倾向于选择更高精度配置。

[0056] 最后,当高、中、低三类精度格式计数数值接近时,为避免量化抖动带来的上下文精度不连续问题,可以引入动态门控机制。门控模块基于精度分布熵(例如Shannon熵或Top-k占比)判断当前处理块精度配置分布是否均衡,当判定为“无明显主导精度”时,强制选择中等精度配置,以此实现对系统总体计算资源的软性调控。该机制对大规模模型部署场景尤为适用,能有效控制功耗峰值,提升模型整体推理吞吐量。

[0057] 示例说明:在医疗健康领域中,面对病历摘要任务时,输入中可能包含诊断结论、症状描述、治疗方案等内容。例如,处理块中包含“患者出现间断性头痛与视物模糊,怀疑颅内压升高”。在前序步骤中,“头痛”“模糊”“颅内压”被标为高精度,“患者”“出现”“怀疑”等被标为中等精度,其余连接词被标为低精度。统计过程中发现中等精度token数量略多于高精度,但高精度与中等精度数量接近。系统根据保守策略,选择高精度作为处理块的统一配置,保证后续诊断模型的推理精度。

[0058] 在金融业务场景下,分析用户交易历史行为生成风险预测时,某一处理块包括“用户今日转账三次,每笔金额超过限额”。在重要性评分阶段,“转账”“限额”“金额”被分配为高精度格式,其余为中等或低精度。经过精度统计后,高精度token数量占据主导,系统将该处理块配置为FP16精度,并统一调用浮点核函数进行处理。该策略确保系统能够对异常交易模式进行精确识别与早期预警。

[0059] 通过统计各精度格式的token数量并采用最大占比原则进行配置选择,有效实现处理块内部精度策略的标准化,使得后续推理阶段可基于块级统一配置进行并行量化计算,提升了系统执行效率并简化了核函数调度路径。同时引入精度优先级规则,使得在配置冲突时优先保证模型语义表达能力不受损,从而在节省资源的同时保持模型精度稳定性。

[0060] S50,将所述语言模型的网络模块划分为多个配置共享组,每个配置共享组至少包含两个网络模块;

在本实施例中,语言模型通常由多个连续堆叠的网络模块(network modules)构成,每个模块包含嵌入层、前馈网络层(Feed-Forward Network)、多头自注意力层(Multi-Head Self Attention)等组成部分。为了在长文本输入情境下控制显存开销并提升推断效率,可将这些网络模块按处理顺序划分为若干“配置共享组”(configuration sharing

group),每组内模块共享相同的精度配置与量化参数,以减少不必要的重复计算。

[0061] 按处理顺序是指按照网络模块在模型执行路径中的自然前后排列进行分组,而非随机打乱或基于任务分配动态组织,这种顺序划分方式能够保证精度配置传播逻辑的稳定性和实现简便性。每个配置共享组内部包含若干相邻的网络模块,其数量可以由一个预设参数决定,该参数可以根据模型的深度、每层计算量、可用计算资源等维度进行设定。常见的预设数量包括4、8、12等,具体值可以通过实验优化确定。每个配置共享组内至少会包含两个网络模块。

[0062] 配置共享组的引入核心是为了形成“块级推理-组级复用”的执行结构。通过统一配置共享组中所有模块的量化方案,可以显著降低在执行过程中需要独立管理的量化配置数量,减少模型推断中不同模块频繁加载配置的显存与调度开销。这种按顺序划分并设定组内固定数量的策略,在结构上可以与模块数量整除,从而在硬件编排中实现统一调度和分配。

[0063] 值得注意的是,配置共享组与前面输入文本分块生成的“处理块”是两个维度的结构:前者作用于模型结构内部,用于网络模块之间的量化配置优化;后者作用于输入数据的组织结构,用于控制输入长度和计算划分,两者之间通过绑定机制(如组内首模块绑定对应的处理块)建立联系。

[0064] 例如,可以将一个包含96个网络模块的大语言模型划分为12个配置共享组,每组包含8个连续模块。划分时按照模块编号顺序进行,例如Group_1包括Module_1至Module_8,Group_2包括Module_9至Module_16,以此类推。在配置阶段,为每个共享组分配一个唯一组标识符,并建立组标识与处理块绑定的映射关系。

[0065] 在执行阶段,当模型推断任务到达某个配置共享组时,系统首先从绑定的处理块中读取该组对应的统一量化配置参数,并将该参数映射到组内所有模块使用。所有组内模块不再进行单独的量化配置初始化,从而节省了时间和存储开销。

[0066] 在特殊场景下的大型深度模型中,由于不同深度的网络模块在语义建模强度、信息流动特性和量化敏感性方面存在非均匀性,因此采用统一数量划分配置共享组可能无法实现精度和效率的最佳平衡。此时可引入非均匀划分方式,通过灵活调整每个配置共享组包含的网络模块数量,使划分策略更加契合模型层级结构与具体任务需求。

[0067] 例如,靠近输入侧的模块通常承担底层语义编码任务(如位置嵌入、词汇建模等),其计算结构相对简单,且对精度损失不敏感,因而可将多个相邻模块合并为较大的配置共享组,提升参数复用效率并减少配置切换频率。反之,靠近输出侧的模块主要完成高层语义抽象与决策信息整合,这些模块对量化精度敏感度更高,若采用统一配置共享策略,易造成精度下降。因此可将这些模块划分为更小的配置共享组,甚至为每两个模块单独配置,从而在输出层保留更多量化灵活度,适应高维输出特征的表示需求。

[0068] 为了进一步提升自适应性,该非均匀划分策略可以与模型结构搜索(Neural Architecture Search,NAS)机制结合。具体做法是将不同划分策略(如共享组大小、共享组位置分布、绑定块编号等)作为结构搜索空间的一部分,在预训练或蒸馏阶段引入控制变量进行性能-效率联合优化。结构搜索结果可以反馈哪些深度区域对精度更敏感,从而引导共享组布局动态调整,最终形成兼顾算力分布和精度稳定性的自适应配置共享机制。

[0069] 此外,在部署面向医疗健康对话系统或金融交易行为建模等任务的深层

Transformer结构时,也可以引入语义密度驱动的动态划分机制:例如根据输入token在模型中引起的梯度激活分布情况,实时动态调整各段网络的组划分方式。语义密度高的区域(如疾病诊断结论或交易行为断点)可配置更小粒度的组划分,以提高配置保真度;而信息冗余区域(如重复问询或无效行为)可使用大组共享,从而减少不必要的计算资源浪费。

[0070] 综上所述,非均匀划分不仅提供了更具弹性的配置复用路径,也为高性能量化推理提供了结构感知优化通路,是复杂模型在资源受限环境中实现高效部署的重要增强策略。

[0071] 通过将语言模型的网络模块划分为多个包含预设数量模块的配置共享组,能够显著降低在推断阶段针对每层模块重复计算量化配置的时间与显存消耗,简化配置调度逻辑。该策略有效引入模块结构层级的配置复用机制,在保障量化灵活性的同时实现计算资源的批量控制。通过组内共享精度配置,避免了大量冗余的配置加载与切换操作,使得长文本推理任务的执行效率获得整体提升。

[0072] S60,在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块;

在本实施例中,为了提升块级量化推理的执行效率与精度一致性,将网络模块划分为多个配置共享组,并在每个组内复用第一个网络模块关联的处理块的统一量化配置。这一操作的本质是构建一种跨模块的参数共享机制,避免在推理过程中为每一个网络模块单独生成或加载量化配置,从而减少冗余配置开销,提升整体运行效率和内存使用率。

[0073] 在具体实现中,首先需要将语言模型的网络模块按照处理顺序划分为多个配置共享组,每个共享组中包含预设数量的网络模块,例如每组包含4个连续的Transformer子层,或者按照模型深度等距划分。随后,从每个共享组中选定第一个网络模块,并将其所关联的处理块的统一量化配置作为该组的基准配置。这一处理块可能来自模型输入阶段实际的推理数据,也可能由模拟数据或任务先验数据产生,确保所选配置具有代表性。

[0074] 为了实现高效共享,需要将统一量化配置写入一个共享内存区域,并为同一配置共享组内的所有网络模块映射相同的内存地址,从而避免重复配置加载与转换操作。共享的量化配置通常包括量化位宽(如INT8或INT4)、缩放因子(scale)、零点(zero point)以及计算精度格式标识(如是否为混合精度浮点计算等)。对于采用加权平均、历史累计或语义重要性聚合策略生成的配置参数,也可以直接写入共享区域,使得共享机制不影响原始配置生成策略的灵活性。

[0075] 这种配置共享机制不仅提升了推理阶段的运行效率,同时也增强了精度的局部一致性,尤其是在模型深层模块之间存在语义递归增强的结构设计中,使用一致的量化配置可以减少误差积累与量化抖动效应。为了保证在动态任务场景下的有效性,可进一步在共享组中引入轻量级的同步机制,例如若检测到首个网络模块绑定的处理块发生配置更新,系统可通过配置映射表自动触发共享内存刷新操作,保证后续模块读取到的是最新配置。

[0076] 在Transformer架构中,假设存在48个网络模块,将其划分为12个配置共享组,每组4个模块。每组中第一个模块绑定一个具有代表性的重要性分数分布的处理块(例如在高频词片段或对话重点区域提取的块),根据该处理块的token精度分配情况生成统一量化配置。生成的配置被写入共享内存区域,每个共享组的其他模块在推理过程中通过相同的内存指针访问此配置,完成权重与激活值的统一量化操作。在多卡部署环境中,还可以在每张

卡的共享内存中同步存储该组的量化配置,通过指针或IPC映射减少通信延迟。若检测到处理块的结构发生剧烈变化或跨场景输入切换,系统可动态更新绑定处理块与配置,并重映射共享内存地址,确保配置共享的适应性。

[0077] 通过构建基于处理块的配置共享组,并在组内复用第一个网络模块关联的量化配置,能够显著减少模型推理过程中的量化参数生成和切换频率,在保证语义保真度的同时降低了重复计算开销。强化了模型内部的计算一致性,并提升了大模型推理在多场景部署下的执行稳定性和可扩展性。

[0078] S70,根据每个处理块对应的统一量化配置,对所有处理块执行块级批量量化并完成模型推理,生成模型推理结果。

[0079] 在本实施例中,为了提升大语言模型在长文本推理任务中的计算效率与显存利用率,需基于每个处理块对应的统一量化配置,执行块级别的批量量化计算,并完成对应的模型推理流程。该操作在技术路径上整合了配置驱动量化计算与处理块并行调度的能力,确保每个处理块按照其重要性精度需求完成精细化推理处理,并最终拼接生成完整的推理结果。

[0080] 每个处理块的统一量化配置,来源于前期重要性分数分析与精度统计操作,通常包括精度格式标识(如FP16、INT8、INT4)、权重参数的量化信息(如缩放因子scale_w与零点zero_point_w)以及激活值的量化信息(scale_a与zero_point_a)。量化计算的执行通常依赖于高效计算核函数的选择,在处理块精度配置生效后,需根据精度格式分配相应的执行路径,例如浮点计算使用基于CUDA或ROCm实现的FP16核函数,定点量化路径则调用INT核函数,同时加载相关量化参数。

[0081] 为提升运行效率,所有处理块被映射到图形处理器(GPU,Graphics Processing Unit)或张量处理器(TPU,Tensor Processing Unit)上的并行处理单元(SM,Streaming Multiprocessor),依照处理块的起始索引或处理顺序进行任务调度。在调度过程中,每个处理块加载自身绑定的统一量化配置,执行对应的矩阵乘、归一化、激活函数及残差连接等操作。过程中,激活值量化参数用于动态调整输入范围,权重量化参数用于将模型权重从定点表示恢复为可计算的量化表达。

[0082] 推理计算完成后,需要对每个处理块的中间结果进行精度校验与无效位置清除操作。这一步通常依据记录的无效token位置索引表完成,将无效token对应位置上的中间张量结果置零或删除,避免其对最终输出产生干扰。所有有效计算结果再按处理块的起始位置索引拼接重构为完整的模型输出序列。

[0083] 最后,对拼接完成的输出序列执行标准的后处理流程,包括维度对齐(如将不同块的张量补齐到统一形状)、归一化(如LayerNorm标准化激活输出),以及解码(如通过Greedy Decoding或Beam Search生成自然语言文本),最终输出用户可直接使用的推理结果文本或结构化向量。

[0084] 例如,在文本生成任务中,将输入文本按512个token一块分割成处理块序列,每个处理块经过步骤4、5分析后获得对应的统一量化配置,如处理块A为高精度格式(FP16)、块B为中等精度(INT8)、块C为低精度(INT4)。在推理执行阶段,GPU加载各处理块的统一量化配置后,将处理块A分配至高精度执行通道,处理块B与C分配至定点执行通道,并通过不同的核函数进行矩阵乘与非线性变换计算。执行完每个处理块的计算后,系统根据预处理时记

录的token位置索引,清除填充token对应的计算结果。例如块C包含12个填充token,则其末尾12个位置的输出将被置零。之后,将块A、B、C的有效输出按起始索引排序拼接,并进行后处理(如归一化与解码),得到完整的文本生成结果。在多路输入的批量推理场景中,还可以结合微批调度机制,在同一GPU中调度多个输入样本的处理块,并在加载共享配置时共享部分高频结构的权重缓存,进一步提升吞吐率。

[0085] 示例说明:在医疗健康业务领域,一家智能辅助诊断平台希望部署大语言模型对超长电子病历文本进行结构化分析与推理,目标是从上万字的住院记录中提取主要诊断、关键症状演化过程及病因推断结论。在该任务中,原始输入文本长度超过模型的标准窗口限制,因此需采用处理块方式进行分段推理。首先,将整个病历文本按预设长度(例如每512个token)划分为多个处理块。系统将第一个处理块标记为高精度处理区域,该区域通常包含病人的基本信息、主诉、现病史等高语义密度部分,直接决定后续诊断推理的可信度,因此禁用该处理块的量化操作,仅使用高精度算子进行浮点计算。对剩余的处理块,平台在模型执行前自动生成每个处理块对应的自注意力矩阵,并对矩阵中每一列进行加和,得到每个token在整个语境中的被关注程度,即重要性分数。在此基础上,系统按照分数与两个动态阈值进行比较,将重要性分数较高的token配置为高精度格式,一般对应于核心病程节点或病因信息,将中间分数的token配置为中等精度格式,用于处理描述性段落或病情观察记录,较低分数的token如格式信息、表格数据则配置为低精度格式。每个处理块完成token级精度标注后,系统统计不同精度格式token的数量,并选择数量最多的精度等级作为该处理块的统一量化配置。例如某处理块中,中等精度格式token占比最高,则该块被整体配置为中等精度计算路径。此策略避免频繁切换精度,提升执行效率。之后,系统将模型的推理模块按顺序划分为若干配置共享组,每组内包含若干连续的网络模块。每组第一个模块所处理的块被绑定并配置精度参数,其余模块共享该配置,避免每层重复执行精度选择逻辑。最终,所有处理块依据各自的统一量化配置,分别调用对应精度的量化计算核函数并行执行推理。高精度块使用FP计算核函数,中等与低精度块调用定点核函数,在GPU或TPU并行单元中并发运行。推理中会自动跳过padding token对应的位置,确保输出结果有效。推理完成后,系统将所有处理块的有效结果按照原始token位置索引拼接为一整段推理结果文本,并进行统一的维度对齐与归一化,最终解码生成诊断结论、关键症状时间线及疑似病因列表。整个推理过程在控制显存占用的同时确保了诊断核心信息的精度与稳定性,实现高性能长文本医疗理解任务的部署能力。

[0086] 在金融科技业务领域,一家面向中小企业的信用评估平台需要基于企业提交的完整经营资料、历史信贷合同、公开年报、客户反馈记录等长文本数据,对企业的信用风险进行大语言模型推理判断。由于这些文档往往结构松散、内容繁杂,整体token数量动辄上万,直接输入模型将面临显存爆炸与响应时延严重问题。系统首先将全部文本信息按预设长度(如每1024 token)划分为多个处理块,保证每块长度可控且计算资源可接受。第一个处理块通常包含企业主体信息、注册资本、核心业务和近三年营业收入等高权重字段,是影响信用评分的核心,因此平台策略上固定该块采用高精度计算路径,禁用量化处理,以最大限度保留关键财务特征和法律声明信息的表达细节。对于其余处理块,平台引入多头自注意力机制分析每个token在上下文中的被关注程度,构建注意力矩阵后,提取每列表示每个token接收的全局关注权重之和,作为token的重要性指标。随后,系统依据两个动态阈值对

每个token分配不同精度等级,确保对如税务异常、合同违约、银行授信记录等高风险因子使用高精度格式处理,而对财务报表注释、客户描述等相对次要信息采用中等或低精度。接着,系统统计每个处理块中各类精度token的数量,并选择其中出现频次最高的精度等级作为该块的统一量化配置,以此简化执行路径并降低调度开销。如果出现高、中、低精度数量相近的边界情况,则偏向选择高精度,以避免信用判断中关键语义被削弱。平台将整个大语言模型按层级划分为多个配置共享组,每组包含固定数量的网络模块,并以组内第一个模块处理的块所采用的量化配置作为共享配置,记录于组标识符与块标识符绑定的配置映射表中,确保所有模块在处理本组任务时调用相同量化路径。在执行阶段,系统为每个处理块加载对应的统一量化配置参数,包括权重的缩放比例、激活值的校准范围以及当前块精度格式标识。不同块调度至支持混合精度的GPU并行单元中,其中高精度块调用浮点处理单元,中低精度块使用INT4/INT8指令集执行推理,同时剔除无效token位置对应中间结果,避免引入冗余干扰。最终,平台将所有有效的推理结果按原始token索引排序拼接成完整输出序列,并通过后处理模块进行结构对齐、风险标签归一化和多标签评分解码,生成企业在当前金融场景下的信用等级、风险敞口范围与预警维度建议清单,协助信审人员或自动化信贷引擎完成决策。

[0087] 通过基于统一量化配置执行块级批量量化计算,不仅实现了精度感知的异构执行策略,同时也充分释放了多核处理器的并行计算能力。在确保不同token区段按照重要性执行差异化计算的同时,有效减少了重复配置加载、动态量化生成等开销。降低了平均推理显存占用并提升处理吞吐率,适用于长文本任务中对资源敏感的大规模部署场景。

[0088] 本发明涉及人工智能技术领域,可应用于医疗健康及金融科技等业务场景中,公开了一种模型量化推理加速方法,包括:将输入文本划分为多个处理块,固定首个处理块的计算精度格式为高精度格式并禁用量化操作;针对除首个处理块外的其他处理块,通过语言模型生成自注意力矩阵,并依据每个token位置在自注意力矩阵中对应列的数值总和计算重要性分数;基于两个预设阈值将各token位置分配为高精度、中等精度或低精度格式;统计每个处理块内各精度格式token数量,选取数量最多的精度格式作为统一量化配置;将网络模块划分为多个配置共享组,并在组内共享处理块的统一量化配置;依据处理块的统一量化配置执行块级批量量化并完成模型推理,生成推理结果。本发明通过基于token重要性分数统一确定每个处理块的量化配置,并在网络模块组内复用该配置,实现了块级别的精度分配与并行量化推理,在保障推理精度的同时大幅降低显存开销和配置时间开销,有效提升长文本推理任务中的执行效率与显存利用率。

[0089] 在一个实施例中,上述步骤S10包括:

S101,根据预设块长度将输入文本划分为多个等长的处理块;

S102,对首个处理块的所有token位置禁用量化参数调整;

S103,将所述首个处理块在语言模型中的嵌入层、自注意力层及前馈网络层的处理精度格式固定为高精度格式;

S104,当输入文本的末尾存在不足预设块长度的剩余token时,将所述剩余token组成的文本片段作为独立处理块;

S105,对所述独立处理块填充无效token至所述预设块长度,并将填充后的独立处理块的处理精度格式固定为高精度格式,且禁用对所述独立处理块的量化操作;

S106,记录所有处理块的起始位置索引和结束位置索引。

[0090] 在本实施例中,为了减轻大语言模型在长文本推断中的显存压力,采用处理块(Processing Block)划分策略将输入文本按预设块长度切分成多个子区段。每个处理块是模型推理中的最小量化调度单元,具有明确的起始与结束位置索引。该划分不仅用于控制显存使用的局部性,还为后续量化策略生成提供边界支持。

[0091] 首个处理块的处理精度格式被固定为高精度格式,通常为FP16或FP32浮点格式,意味着模型在处理该块时采用全精度计算路径,并显式禁止该块参与任何形式的量化参数调整,包括但不限于权重参数与激活值的位宽压缩、动态缩放因子生成等,从而跳过该块在量化配置生成中的流程。

[0092] 嵌入层(Embedding Layer)、自注意力层(Self-Attention Layer)与前馈网络层(Feed-Forward Network Layer)共同构成语言模型中token表示变换的核心模块,对token的语义表征起着关键作用。对首个处理块的这些核心结构采用高精度格式,有助于稳定模型在推断初期的上下文感知能力,并提升后续token的嵌套推理表现,尤其在长文本输入初期包含关键主题提示的场景下更为显著。

[0093] 若文本长度无法被预设块长度整除,尾部剩余的token将被组装成独立的处理块。为保持块级推断结构的一致性与便于并行调度的对齐需求,该独立处理块将填充无效token(Paddingtoken)以补齐至完整块长度,并同样使用高精度格式处理。填充token在实际推断中不参与模型输出的有效计算,仅用于保持输入维度一致。

[0094] 系统在完成文本分块后,需记录每个处理块的起始位置与结束位置索引信息,作为后续计算调度、输出拼接、精度分配等步骤的关键辅助标识。这些索引信息可构建为双向映射表(如哈希表),以便在运行时快速定位token对应的处理块及其精度控制策略。

[0095] 可以通过设定块长度为模型支持的最大序列长度的子集(如512或1024个token),对输入文本进行顺序划分。划分操作可在预处理阶段离线完成,也可在运行时基于流式输入动态进行。在GPU推理框架中,可使用position encoding掩码区分实际token与填充token,配合高精度kernel函数处理指定块。

[0096] 在当前主流的量化推断硬件体系中,FPGA(Field Programmable Gate Array,现场可编程门阵列)、ASIC(Application-Specific Integrated Circuit,专用集成电路)以及TPU(Tensor Processing Unit,张量处理单元)广泛应用于高性能大语言模型的部署。针对这些平台,在实际部署时需要提供灵活的精度调度能力,以兼容首个处理块的高精度处理需求。

[0097] 在可编程芯片架构(如FPGA或AI加速芯片)中,一般会将模型执行路径中的各个操作映射为一系列可重构逻辑单元或算子链(operator chain)。为实现首个处理块的高精度执行,可以在中间计算路径上通过位宽控制寄存器(bit-width control register)或算子调度指令表设置当前块的计算精度模式。当识别到处理的token位置属于首个处理块时,系统控制器可通过软硬协同的方式,将对应算子链的精度控制字段设置为高精度标志位,例如启用FP16或FP32通道,并绕过量化缩放模块(quantization scaling unit),跳过零点调整、离散映射等操作,使数据以原始浮点格式直接传输至算子输入口,实现所谓的“高精度直通通道”(precision passthrough path)。

[0098] 这一机制的关键在于将处理块的位置信息与寄存器配置策略做绑定。可以通过在

调度单元中维护一个处理块-精度策略映射表,实现按块级粒度动态切换精度路径的目的。在处理流式token输入时,该机制可通过边界检测模块实现运行时配置更新。

[0099] 在TPU这类高度集成的异构计算平台中,内部通常包含多个支持不同计算精度的处理核心(Core),其中部分核心为高算力、支持浮点计算的主核心(High-Precision Core),而其他核心则优化为定点计算专用路径。TPU架构中常集成调度引擎(Scheduler)与资源管理器(Resource Allocator),用于在不同Core间动态调度工作负载。

[0100] 在处理首个处理块时,调度引擎可以识别其为高优先级计算单元,并将其任务直接分配至主核心,利用浮点计算核心执行Embedding、自注意力与前馈网络模块的计算,从而避免量化过程的引入;而其余处理块的推理任务则可以划分至其他低位宽Core中,以低精度高吞吐量模式执行量化推断。为了实现资源的动态复用,TPU的调度策略还可以支持时间片轮转(time slicing)与任务预取(prefetching),保证多个处理块之间调度的低延迟和高并行。

[0101] 若部署于通用GPU平台,也可以通过CUDA kernel launch参数传递高精度执行标识,实现对特定kernel路径中量化流程的禁用。例如为首个处理块启动一组不包含量化核函数的kernel,或在通用模型框架中通过精度标签注入(precision tag injection)机制强制调度float路径而非int路径。

[0102] 此外,在更高阶的系统中,也可以引入动态微调逻辑(precision adaptation controller),根据首个处理块的重要性分数分布或语义权重自动判断是否启用高精度路径,从而在性能与精度之间实现灵活平衡。

[0103] 本实施例通过将输入文本划分为多个等长处理块并对首个处理块禁用量化处理,可以在模型推断初始阶段保留更多原始语义与上下文信息,从而提高后续token重要性评估与精度分配的稳定性。有助于在不显著增加总体显存开销的前提下提升长文本输入的起始推断精度,为后续分块量化配置提供更加可靠的基线表示。

[0104] 在一个实施例中,上述步骤S20包括:

S201,将每个其他处理块输入至语言模型的多头自注意力层,得到每个其他处理块对应的多个注意力头的局部自注意力矩阵;

S202,对多个注意力头的局部自注意力矩阵进行加权平均或算术平均处理,生成对应每个其他处理块的最终自注意力矩阵;

S203,从所述最终自注意力矩阵中提取每个token位置对应的列向量;

S204,确定每个列向量中全体元素的数值之和,并将所述数值之和作为每个列向量对应的token位置的重要性分数;

S205,若其他处理块中存在填充的无效token位置,则在确定所述填充的无效token位置的重要性分数时,将无效token位置对应列向量的所有元素数值置零。

[0105] 在本实施例中,在对处理块进行精度分配之前,需要首先获取每个token在当前语义上下文中的重要性。为了实现这一目标,处理流程从多头自注意力层出发,为每个处理块构建注意力矩阵。每个处理块在输入语言模型后会经过多个注意力头,每个注意力头独立计算一份局部自注意力矩阵,每个矩阵都以处理块中所有token为单位,描述它们之间的相互关注关系。这些矩阵反映了语言模型从不同关注角度对句子内部结构的建模能力。

[0106] 为了获得统一且具备可比较性的关注度表示,需要将这些局部矩阵进行融合处

理。融合方式可以根据实际任务需求灵活设定,常见方式包括对所有注意力头的矩阵进行算术平均,即简单求均值;也可以通过为每个注意力头配置不同的权重系数,实现加权平均。这一融合过程生成了每个处理块的最终自注意力矩阵,该矩阵作为处理块内部token相互依赖关系的综合表示。

[0107] 在该最终矩阵中,每一列向量代表了一个特定token接收来自其他token的注意力聚焦程度。具体来说,固定某个token的位置,提取该位置对应的列向量,可以看作是“该token被所有其他token所关注的程度集合”。这是评估一个token是否在当前处理块中占据重要语义地位的重要依据。即:

当前token:正在计算其注意力输出的那个token,记为第i个位置的token。

[0108] 其他token:同一个处理块中除了第i个位置以外的所有有效token,即在块内的位置编号为0到N-1(其中N是处理块长度);不包括第i个位置本身;排除无效token(如paddingtoken、填充token)的位置。

[0109] 为了提取这种重要性信号,将每个列向量中的全部元素进行数值求和,得到一个用于度量该token被全局关注的聚合值。该聚合值即为该token的重要性分数,分数越高,说明该token对整体上下文语义结构的影响越大。在后续的精度分配过程中,这一分数将直接用于判断该token是否应使用更高精度的计算资源进行处理。

[0110] 考虑到处理块可能包含为填充长度而补入的无效token,这些token并不携带语义内容,如果参与注意力矩阵的计算,可能会对重要性分数造成干扰。为防止此类干扰,需在生成最终自注意力矩阵后,识别所有无效token对应的位置,并对其在矩阵中的列向量统一置零。这样可以有效避免无效token获得虚假的高关注度值,确保重要性计算的准确性和鲁棒性。该操作可以通过结构化掩码或显式索引机制实现,具体方法依据模型架构与部署平台不同而有所调整。

[0111] 在实际部署中,当处理块被送入语言模型进行推理时,模型的每一层自注意力机制都会输出若干注意力头所对应的注意力矩阵。这些矩阵通常以三维结构组织,其第一个维度为注意力头的数量,第二和第三个维度分别表示处理块中token的数量,即形成一个结构为“注意力头数×token数×token数”的三维张量。

[0112] 模型在前向推理过程中,可通过中间层拦截机制,在每一层的自注意力模块输出处缓存这些注意力张量。为了减少推理时的中断开销,缓存操作应当以非阻塞方式插入模型的计算图中,并由调度器统一管理其生命周期。在缓存完成后,通过在第一个维度(即注意力头维度)上进行操作,可以实现不同融合策略:一种方式是对所有注意力头的矩阵进行简单平均,得到一个二维矩阵;另一种方式是对每个头赋予不同的加权系数,进行加权平均处理,这种方式适用于事先评估出不同注意力头对语义感知能力的差异,在权重设置上做策略性调整,以增强融合矩阵的表达能力。

[0113] 完成注意力头融合后,将得到一个token数×token数的二维矩阵,表示处理块内部token间的综合关注关系。每个token的全局关注程度由该矩阵中对应列的元素总和决定。具体操作中,可以通过固定某一列索引,对整张矩阵进行列向量提取,这一操作可通过标准的矩阵切片函数在张量操作框架中实现(如PyTorch、TensorFlow、JAX等均支持此类切片访问)。随后,对该列向量中的所有数值进行累加,得到该位置token的注意力聚合值,即其重要性分数。由于向量加法在现代张量计算库中均有高度优化实现,重要性分数的计算

能够在GPU/TPU上以极低的时延完成,具备良好的工程可行性。

[0114] 针对填充token的处理,在前向推理时模型通常会生成一个padding mask(填充掩码),该掩码用于标记哪些token是由填充产生的无效位置。掩码一般为与处理块等长的布尔向量或0/1张量,其中1表示有效位置,0表示填充位置。可将该掩码扩展为矩阵形式,与最终自注意力矩阵的列结构对齐。在实施上,该掩码可广播为二维矩阵,应用于注意力矩阵的每一列,在矩阵级别将填充位置对应列向量的所有元素统一置零。

[0115] 为了实现效率最优,该掩码应用过程可在模型图构建阶段嵌入量化前处理模块,作为一个后置操作挂接在注意力矩阵计算节点之后。针对运行平台的不同,在GPU上可以通过CUDA kernel实现列级清零逻辑;在TPU上则可依托其内建的高通量张量映射操作,在单个时钟周期内完成批量列屏蔽。

[0116] 此外,为确保上下文场景中出现的特殊结构不会引发错误判断,如在部分任务中填充token可能处于序列中间而非尾部(如多段拼接输入),需要提前记录原始输入结构和padding mask生成逻辑,确保掩码准确性。可选方式是引入位置标签辅助机制,将每个token的位置语义编码与掩码联合使用,以强化无效token的识别稳定性。

[0117] 本实施例通过从融合自注意力矩阵中提取列向量并进行总权重求和,得到token的重要性分数,构建了从语言模型结构出发的量化精度感知机制。在不引入外部打分模块的前提下,充分利用模型自身已学习到的注意力信息,实现精度分配依据的内部化。避免了显式标签依赖,同时具有良好的计算图兼容性,有助于后续在加速芯片上的高效部署。

[0118] 在一个实施例中,上述步骤S30包括:

S301,根据处理块的长度确定第一阈值和第二阈值;

S302,若存在无效token位置,则将无效token位置的处理精度格式分配为低精度格式,并在统计每个处理块的精度格式数量时排除所述无效token位置;

S303,将每个有效token位置的重要性分数与所述第一阈值和第二阈值比较;

S304,将重要性分数大于第一阈值的有效token位置的处理精度格式分配为高精度格式;

S305,将重要性分数处于第二阈值之上且处于第一阈值之下的有效token位置的处理精度格式分配为中等精度格式;

S306,将重要性分数小于第二阈值的有效token位置的处理精度格式分配为低精度格式。

[0119] 在本实施例中,将重要性分数大于第一阈值的token位置分配为高精度格式,是在对每个处理块中所有有效token的注意力信息进行评估后,对其处理精度进行差异化配置的重要操作。高精度格式指代的是在推理过程中采用精度更高的表示与计算方式,例如使用更高位宽的数值表示或禁用量化操作的浮点运算。其目的是为了保留关键语义token的表达能力,避免精度下降对模型预测结果造成负面影响。中等精度格式与低精度格式则对应于资源受限场景下的中低位宽量化配置,用于处理重要性较低的token,节省显存与计算资源。

[0120] 处理块的长度对阈值的设定具有动态影响。因为长处理块中token数量增加,其整体信息密度分布趋于平均,为了保证筛选机制在不同长度处理块中具有相对一致的区分能力,需将第一阈值与第二阈值的取值相应下调,扩大高精度token的判定范围。可以通过设

定初始标准阈值,并结合当前块长度与标准长度的比例,对两个阈值进行线性或非线性缩放,从而实现动态自适应。

[0121] 在处理块中存在无效token位置的场景下,处理精度分配需要加入特殊判断。由于无效token不参与语义建模,因此在其注意力信息中本身就无语义贡献,其对应的重要性分数可以认为是结构性空值或强制置零。将其统一分配为低精度格式可以节省运算资源,并在后续统计精度比例时主动排除,以避免对当前处理块的统一量化配置决策产生干扰。

[0122] 每个有效token的重要性分数将依次与两个阈值进行比较。当其数值高于第一阈值,即代表其在全局注意力聚合中承担显著语义连接作用,需分配高精度格式;当其分数处于第一与第二阈值之间,则被认为为信息次重要区间,可使用中等精度格式处理;当其低于第二阈值,则表示该token对上下文构建作用较弱,可采用低精度格式执行推理计算。该多等级精度配置方案保证了语义表达、性能控制与计算效率三者的平衡。

[0123] 若处理块中不存在无效token位置,则所有的token位置均为有效token位置。

[0124] 在具体实现中,可以在初始化模型运行时,通过读取处理块的长度,调用预设的动态阈值计算函数生成当前块的第一阈值与第二阈值。该函数可通过查表方式或公式计算方式进行实现,具体实现可支持区间插值、指数递减、最小阈值保护等机制。随后,将当前处理块的所有token位置按顺序遍历。在遍历过程中,系统调用padding mask或token有效性标签判断当前token是否为无效token。若是无效token,则直接将其处理精度格式标记为低精度,并跳过重要性分数比较逻辑,进入下一token判断流程。对于有效token,则从缓存或当前attention模块中读取该token的重要性分数。该分数随后与当前处理块对应的两个阈值进行比较。若该分数大于第一阈值,则将其在处理块精度分配向量中标记为高精度;若该分数大于第二阈值且小于等于第一阈值,则标记为中等精度;若该分数小于等于第二阈值,则标记为低精度。处理完成后,将所有有效token的精度标签保存在处理块内部的精度控制向量中,并用于后续的统一量化配置决策与核函数选择。若需要加速处理,可将重要性分数比较逻辑向量化实现,在GPU上并行完成分配策略的判断,提高执行效率。

[0125] 本实施例通过上述步骤,可在不显著增加模型复杂度的前提下,根据token的实际语义重要性对其计算精度进行差异化配置,从而在保证模型推理精度的基础上大幅压缩整体显存消耗与计算负载。动态阈值机制进一步提升了算法在处理不同长度输入序列时的适配能力,避免固定阈值导致的精度分配不均问题。无效token的专门处理确保资源分配的精准性,并简化后续统计逻辑。

[0126] 在一个实施例中,上述步骤S40包括:

S401,遍历当前处理块的所有token位置,识别被分配为高精度格式、中等精度格式及低精度格式的token位置;

S402,在统计过程中,若存在无效token位置,则排除所有无效token位置,仅统计有效token位置的精度格式分配结果;

S403,分别统计有效token位置中被分配为高精度格式、中等精度格式及低精度格式的计数,生成高精度计数、中等精度计数及低精度计数;

S404,比较所述高精度计数、中等精度计数及低精度计数的数值大小;

S405,将数值最大的计数对应的精度格式作为所述当前处理块的统一量化配置;

S406,若存在多个精度格式的计数相同且为最大值,则选择多个精度格式中的最

高精度格式作为所述当前处理块的统一量化配置。

[0127] 在本实施例中,统计每个处理块内被分配为不同精度格式的token位置数量,是在完成token精度分配后,为确定当前处理块的整体量化策略而执行的关键分析操作。该过程首先需要遍历当前处理块中所有token位置,识别每个位置对应的精度标签,该标签可能为高精度、中等精度或低精度格式。

[0128] 遍历过程中如果处理块中包含无效token,例如末尾填充的padding token,需要将其在统计中排除。这是因为无效token不参与实际推理计算,不应对整体精度分布产生影响。仅统计有效token的精度标签,能够确保后续统一量化配置决策的合理性与准确性。

[0129] 对每种精度标签进行计数是该步骤的核心,系统将分别生成三个计数结果:高精度计数、中等精度计数、低精度计数。每个计数结果对应处理块中有效token被分配为该精度格式的位置数量,用于刻画该块内精度分布趋势。

[0130] 随后,系统比较三种计数结果的大小,寻找数量最多的精度格式,并将其作为当前处理块的统一量化配置,即当前处理块后续将以该格式进行统一量化与计算核函数调用。统一配置机制的设计目的是避免在同一处理块内部频繁切换精度计算路径,从而降低上下文切换代价,提高推理吞吐效率。

[0131] 如果多个精度计数结果相同且为最大值,如高精度与中等精度计数完全一致,则系统应使用精度优先原则,在并列最大值中选择精度等级更高者作为最终配置。该策略体现了对语义质量的保护倾向,即在资源允许的范围内优先保留更高计算精度,从而降低token误识或推理偏差的风险。

[0132] 例如,模型执行过程中,在完成每个处理块的token精度分配后,系统会启动精度统计模块,首先通过循环或向量化逻辑遍历该处理块中的所有token索引位置。每个token对应的精度标签存储于前序步骤的精度标记向量中,向量每个元素为该token的精度标识,如0表示低精度,1表示中等精度,2表示高精度。

[0133] 在遍历过程中,系统检查每个token是否为有效token。判断依据可以是padding mask或token有效性向量。若为无效token,其对应位置将在精度统计过程中跳过,不参与任何精度计数。

[0134] 对有效token,则根据其精度标识将其加入对应的精度计数器。系统同时维护三个独立的整数变量,分别累计高、中、低精度的token数量。完成遍历后,系统对三个计数器的值进行比较,找出最大值对应的精度格式。

[0135] 若某两个或三种精度计数均为最大值,例如高精度和中等精度均为100,则系统执行优先级判断,可通过精度等级数值进行比对,优先选择等级更高的精度格式。系统在此逻辑中还可扩展引入额外因素,如历史块配置惯性、上下文语义密度趋势等,以增强精度配置的稳定性与适应性。

[0136] 最终确定的精度格式将作为当前处理块的统一量化配置写入处理块控制结构中,并传递至后续的量化核函数选择逻辑中。该配置可在GPU上以batch形式批量应用,提高量化准备阶段的效率。

[0137] 本实施例通过对每个处理块内部的精度标签进行全面统计与优先级分析,能够以数据驱动的方式为每个块分配最符合其语义密度与精度要求的统一量化配置,显著降低计算路径复杂度与精度切换开销。在保留关键token高精度处理能力的同时,也通过块级合并

策略实现了推理效率提升与显存压缩。无效token的排除处理确保了统计结果不被冗余token干扰,进一步提升了系统的配置准确性与执行稳定性。精度优先策略的引入,使得系统在面对精度分布模糊场景时,依然能够保障高语义贡献token的处理质量,有助于模型输出的稳定性和可靠性。

[0138] 在一个实施例中,上述步骤S60包括:

S601,为每个处理块分配唯一块标识符,并为每个配置共享组分配唯一组标识符;

S602,在配置映射表中建立唯一组标识符与唯一块标识符的绑定关系;

S603,基于所述唯一组标识符与唯一块标识符的绑定关系,将每个配置共享组的第一个网络模块与对应处理块的统一量化配置参数相关联;

S604,将所述统一量化配置参数写入每个配置共享组的共享内存区域,并为每个配置共享组的所有网络模块分配相同的内存地址映射;

S605,同一配置共享组内的其他网络模块在执行量化处理时,通过所述内存地址映射从所述共享内存区域读取所述统一量化配置参数。

[0139] 在本实施例中,为了在大语言模型(LLM)推断过程中实现显存资源和计算路径的压缩优化,引入配置共享机制以在同一组内的多个网络模块之间复用统一量化配置。该机制的关键在于对处理块和网络模块的映射关系进行结构化组织,并利用共享内存实现参数访问路径的复用。

[0140] 首先,系统会在处理块划分阶段为每个处理块生成全局唯一的块标识符,用于标识该块所对应的一组token序列及其量化配置。在不同实现平台中,该标识符可以是整数索引、散列结果或者结合处理块位置与长度的组合编码,具备稳定性与可追溯性,便于在后续步骤中快速定位量化配置所属。

[0141] 与此同时,系统还会根据网络模块在模型推理图中的处理顺序,将其划分为若干配置共享组。配置共享组是一个逻辑结构,内部包含多个相邻或语义相关的网络模块,通常设置为固定数量(如每组包含4个、8个或16个模块)。每个配置共享组也会被分配一个唯一的组标识符,其生成机制应保持与处理块标识符一致的唯一性和可映射性。

[0142] 在建立这两类标识的基础上,系统通过构建配置映射表,形成组标识符与块标识符之间的绑定关系。这一映射表可以采用键值对形式,组标识符为键,对应的块标识符为值。该结构允许模型在运行时迅速查找当前配置共享组所绑定的处理块及其量化配置,从而确定共享访问路径。

[0143] 进一步地,为了实现量化配置在组内多模块间的复用,系统在绑定完成后会将该处理块的统一量化配置参数写入到一块共享内存区域中。这一共享区域在实现层面可以是GPU显存中的共享缓存(shared memory),也可以是支持并发访问的张量加速器寄存器区,或在支持异构架构的系统中由中间件分配的高带宽缓存块。

[0144] 统一量化配置参数通常包括权重量化参数、激活值量化参数、处理精度格式信息、位宽编码策略、量化范围信息及辅助偏置等。这些参数以结构化格式存储,例如键值字典、压缩张量或配置结构体,并通过共享内存地址暴露给配置共享组内部的所有网络模块。

[0145] 在地址映射方面,系统会将配置共享组内部所有模块的量化配置访问路径指向统一的共享内存地址。该映射过程既可以在模型加载时静态完成,也可以在推理过程中动态维护,具体取决于处理块生成与组内模块调用的时间耦合关系。需要特别注意的是,为了保

证多模块并发读取的性能和一致性,应当确保共享内存区域具备线程安全性和非阻塞特性,这可通过锁机制、读写缓存或通信通道实现。

[0146] 一旦映射完成,配置共享组内的其余网络模块在执行量化处理时,将不再独立生成或加载量化配置,而是通过查找映射地址,从共享区域直接读取所需的参数数据。这种方式避免了重复计算、内存复制和配置切换,极大地压缩了推理路径的执行图宽度,提升了整体计算密度和吞吐率。

[0147] 更进一步地,如果该机制与量化感知训练(QAT)或后训练量化(PTQ)策略结合,配置共享不仅降低了硬件层的显存消耗,也提高了语义一致性,使得同一组内的模块能在一致量化配置下协同表达局部语义,有效避免因量化策略差异导致的表示抖动或语义漂移。

[0148] 此外,配置共享机制具备良好的可扩展性与硬件兼容性。在支持图执行框架(如TensorRT、ONNX Runtime)中可通过统一层参数注入策略实现,在低级硬件层(如TPU、NPU)则可通过内存地址绑定表与中断同步机制配合完成动态共享。

[0149] 本实施例通过配置共享机制,实现了处理块量化配置在配置共享组内的高效复用,显著降低了重复配置加载造成的显存消耗与计算负担。相较于传统每个模块独立加载配置的方式,使用统一内存地址映射可将配置访问时延降至常量级,并降低系统整体峰值内存占用。组内统一量化策略还提升了多模块协同执行的一致性,有效减少推理路径中的量化抖动,增强模型输出稳定性。在多块并发执行场景下,此机制可有效规避因量化策略不一致导致的语义漂移或输出断层问题。

[0150] 在一个实施例中,上述步骤S70包括:

S701,为每个处理块加载对应的统一量化配置参数,所述统一量化配置参数包括精度格式标识;

S702,根据所述精度格式标识,为不同处理块配置独立的处理核函数;

S703,在图形处理器或张量处理器的并行处理单元中,根据处理块的索引顺序分配处理资源,并发执行所有处理块的量化处理;

S704,对每个处理块的处理结果进行块内token位置校验,以剔除无效token位置对应的中间结果;

S705,将所有处理块的有效中间结果按起始位置索引排序,拼接为完整的模型输出序列;

S706,对拼接后的输出序列执行后处理操作,生成最终的模型推理结果。

[0151] 在本实施例中,为每个处理块加载统一量化配置参数是模型执行前的准备阶段。统一量化配置参数包括三类关键内容,其一为权重量化参数,用于将模型权重从浮点表示映射到量化表示的参数集合,通常包括缩放因子和零点,这些参数在静态量化或动态量化过程中均有广泛应用;其二为激活值量化参数,控制输入激活值的压缩精度,用于在推理时保持输入特征分布的稳定性;其三为精度格式标识,是指导处理块选择对应计算路径的标记,通常为枚举类型或控制寄存器标志,用于后续动态核函数调度。加载这些参数的具体实现可以通过配置表查找、内存预读或配置缓存实现,以减少推理阶段的访存延迟。

[0152] 依据精度格式标识配置独立的处理核函数是实现块级异构计算的关键环节。处理核函数是针对不同精度格式所定制的执行路径,高精度格式通常对应于FP16或FP32的浮点核函数,保留更高的数值精度以处理语义密集或位置关键的token;中等和低精度格式对应

于INT8、INT4等定点量化核函数,利用张量量化、权重共享等机制加速矩阵计算与内存传输。计算核函数的调度策略可通过函数指针映射表或硬件支持的kernel选择机制实现,使得处理块执行路径按需分配,提升整体吞吐量。

[0153] 在图形处理器(GPU)或张量处理器(TPU)中,为所有处理块分配并行处理资源是实现块级批量量化的物理保障。基于处理块的索引顺序进行资源分配可确保处理结果的输出顺序与输入顺序一致,同时避免线程间资源竞争。在GPU平台中,可通过CUDA block与stream机制控制处理核函数分布,而在TPU中则可通过矩阵乘单元(MXU)映射各处理块执行。此过程中权重量化参数与激活值量化参数被同步加载到对应的执行上下文中,分别作用于模型中权重张量的量化映射与输入张量的动态校准操作。

[0154] 处理完成后的中间结果需进行块内token位置校验。由于某些处理块包含padding填充的无效token,这些位置的计算结果应被过滤,以防影响后续模型输出。在该步骤中,可利用前序步骤中记录的token mask对中间结果按位置执行筛选操作,剔除无效token对应结果并保持其在拼接过程中的位置一致性。

[0155] 拼接所有处理块的有效中间结果时,需按照原始输入的起始位置索引进行排序与连接,确保语义信息按正确的上下文顺序被复原。该操作不仅需要基于token索引进行数据重组,还要求不同处理块输出格式保持一致性,包括张量维度、序列对齐方式等,避免拼接过程产生维度错位或信息重叠。

[0156] 最终,将拼接后的序列输入至后处理模块,执行维度对齐、归一化与解码操作。维度对齐用于标准化各段输出的张量形状,归一化用于压缩数值动态范围,提升解码准确性,解码则依据模型任务类型生成文本、标签或特征值等最终推理结果。整个后处理流程可以嵌入至推理图后段作为统一的输出层,确保高效执行。

[0157] 在某些实现中,权重量化参数与激活值量化参数可通过外部量化训练得到,保存在离线配置文件中,并在推理开始前预载入。另一种方式是运行时动态量化,根据当前输入激活值的最大最小值区间实时计算缩放比例。对于精度格式标识的赋值,可以在量化配置生成步骤中将其表示为整型编码,例如0表示低精度、1表示中等精度、2表示高精度,供处理核函数选择器解析。

[0158] 处理核函数的分配可以采用硬件支持的自适应kernel调用策略。例如在NVIDIA TensorRT框架中,根据精度标识自动匹配不同的GEMM实现模块,而在TPU平台上可通过XLA编译器将精度配置编译为相应的硬件指令序列。处理资源的分配也可以根据硬件架构进行优化,部分平台支持优先将高精度块调度至计算性能更强的核心资源区,以保障模型推理准确性。

[0159] 在实际部署中,为降低拼接过程的带宽开销,可以在每个处理块输出后立即将其有效结果写入预分配的全局输出缓存区,按token索引位置排布,避免后续排序操作。在维度对齐过程中,对于因精度格式不同导致的输出张量差异,可引入统一映射策略或padding补齐机制,确保数据结构一致。

[0160] 本实施例通过为每个处理块加载其对应的统一量化配置,并据此动态选择处理核函数,实现了计算精度与资源分配的协同优化。通过引入精度格式标识控制核函数路径,使得在保证高重要性区域处理精度的同时,最大化地压缩了低重要性区域的计算资源。并行处理单元按块执行进一步提升了执行效率,而无效token位置的剔除确保了拼接结果的精

度与一致性。最终的输出经过统一后处理操作后,既能保持上下文语义连贯性,又有效降低了整体推理延迟与显存消耗。

[0161] 在一实施例中,提供一种模型量化推理加速装置,该模型量化推理加速装置与上述实施例中模型量化推理加速方法一一对应。参照图3,图3为本发明模型量化推理加速装置一较佳实施例的功能模块示意图。输入文本预处理模块10、自注意力分析模块20、精度分配模块30、量化配置决策模块40、网络模块分组控制模块50、配置共享管理模块60和推理执行模块70。各功能模块详细说明如下:

输入文本预处理模块10,用于将输入文本划分为多个处理块,将首个处理块的处理精度格式固定为高精度格式,并禁用对所述首个处理块的量化处理;

自注意力分析模块20,用于对所述多个处理块中除首个处理块以外的其他处理块,通过语言模型生成每个其他处理块的自注意力矩阵,并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和,并将所述全体元素数值之和作为每个token位置的重要性分数;

精度分配模块30,用于将重要性分数大于第一阈值的token位置分配为高精度格式,将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式,将重要性分数小于第二阈值的token位置分配为低精度格式;

量化配置决策模块40,用于统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量,选择数量最多的精度格式作为对应处理块的统一量化配置;

网络模块分组控制模块50,用于将所述语言模型的网络模块划分为多个配置共享组,每个配置共享组至少包含两个网络模块;

配置共享管理模块60,用于在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块;

推理执行模块70,用于根据每个处理块对应的统一量化配置,对所有处理块执行块级批量量化并完成模型推理,生成模型推理结果。

[0162] 在一实施例中,输入文本预处理模块10,具体用于:

根据预设块长度将输入文本划分为多个等长的处理块;

对首个处理块的所有token位置禁用量化参数调整;

将所述首个处理块在语言模型中的嵌入层、自注意力层及前馈网络层的处理精度格式固定为高精度格式;

当输入文本的末尾存在不足预设块长度的剩余token时,将所述剩余token组成的文本片段作为独立处理块;

对所述独立处理块填充无效token至所述预设块长度,并将填充后的独立处理块的处理精度格式固定为高精度格式,且禁用对所述独立处理块的量化操作;

记录所有处理块的起始位置索引和结束位置索引。

[0163] 在一实施例中,自注意力分析模块20,具体用于:

将每个其他处理块输入至语言模型的多头自注意力层,得到每个其他处理块对应的多个注意力头的局部自注意力矩阵;

对多个注意力头的局部自注意力矩阵进行加权平均或算术平均处理,生成对应每

个其他处理块的最终自注意力矩阵;

从所述最终自注意力矩阵中提取每个token位置对应的列向量;

确定每个列向量中全体元素的数值之和,并将所述数值之和作为每个列向量对应的token位置的重要性分数;

若其他处理块中存在填充的无效token位置,则在确定所述填充的无效token位置的重要性分数时,将无效token位置对应列向量的所有元素数值置零。

[0164] 在一实施例中,精度分配模块30,具体用于:

根据处理块的长度确定第一阈值和第二阈值;

若存在无效token位置,则将无效token位置的处理精度格式分配为低精度格式,并在统计每个处理块的精度格式数量时排除所述无效token位置;

将每个有效token位置的重要性分数与所述第一阈值和第二阈值比较;

将重要性分数大于第一阈值的有效token位置的处理精度格式分配为高精度格式;

将重要性分数处于第二阈值之上且处于第一阈值之下的有效token位置的处理精度格式分配为中等精度格式;

将重要性分数小于第二阈值的有效token位置的处理精度格式分配为低精度格式。

[0165] 在一实施例中,量化配置决策模块40,具体用于:

遍历当前处理块的所有token位置,识别被分配为高精度格式、中等精度格式及低精度格式的token位置;

在统计过程中,若存在无效token位置,则排除所有无效token位置,仅统计有效token位置的精度格式分配结果;

分别统计有效token位置中被分配为高精度格式、中等精度格式及低精度格式的计数,生成高精度计数、中等精度计数及低精度计数;

比较所述高精度计数、中等精度计数及低精度计数的数值大小;

将数值最大的计数对应的精度格式作为所述当前处理块的统一量化配置;

若存在多个精度格式的计数相同且为最大值,则选择多个精度格式中的最高精度格式作为所述当前处理块的统一量化配置。

[0166] 在一实施例中,配置共享管理模块60,具体用于:

为每个处理块分配唯一块标识符,并为每个配置共享组分配唯一组标识符;

在配置映射表中建立唯一组标识符与唯一块标识符的绑定关系;

基于所述唯一组标识符与唯一块标识符的绑定关系,将每个配置共享组的第一个网络模块与对应处理块的统一量化配置参数相关联;

将所述统一量化配置参数写入每个配置共享组的共享内存区域,并为每个配置共享组的所有网络模块分配相同的内存地址映射;

同一配置共享组内的其他网络模块在执行量化处理时,通过所述内存地址映射从所述共享内存区域读取所述统一量化配置参数。

[0167] 在一实施例中,推理执行模块70,具体用于:

为每个处理块加载对应的统一量化配置参数,所述统一量化配置参数包括精度格

式标识;

根据所述精度格式标识,为不同处理块配置独立的处理核函数;

在图形处理器或张量处理器的并行处理单元中,根据处理块的索引顺序分配处理资源,并发执行所有处理块的量化处理;

对每个处理块的处理结果进行块内token位置校验,以剔除无效token位置对应的中间结果;

将所有处理块的有效中间结果按起始位置索引排序,拼接为完整的模型输出序列;

对拼接后的输出序列执行后处理操作,生成最终的模型推理结果。

[0168] 在一个实施例中,提供了一种计算机设备,该计算机设备可以是服务端,其内部结构图可以如图4所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性和/或易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的网络接口用于与外部的用户端通过网络连接通信。该计算机程序被处理器执行时以实现一种模型量化推理加速方法服务端侧的功能或步骤。

[0169] 在一个实施例中,提供了一种计算机设备,该计算机设备可以是用户端,其内部结构图可以如图5所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口、显示屏和输入装置。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统和计算机程序。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的网络接口用于与外部服务器通过网络连接通信。该计算机程序被处理器执行时以实现一种模型量化推理加速方法用户端侧的功能或步骤

在一个实施例中,提供了一种计算机设备,包括存储器、处理器及存储至存储器上并可在处理器上运行的计算机程序,处理器执行计算机程序时实现以下步骤:

将输入文本划分为多个处理块,将首个处理块的处理精度格式固定为高精度格式,并禁用对所述首个处理块的量化处理;

对所述多个处理块中除首个处理块以外的其他处理块,通过语言模型生成每个其他处理块的自注意力矩阵,并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和,并将所述全体元素数值之和作为每个token位置的重要性分数;

将重要性分数大于第一阈值的token位置分配为高精度格式,将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式,将重要性分数小于第二阈值的token位置分配为低精度格式;

统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量,选择数量最多的精度格式作为对应处理块的统一量化配置;

将所述语言模型的网络模块划分为多个配置共享组,每个配置共享组至少包含两个网络模块;

在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块;

根据每个处理块对应的统一量化配置,对所有处理块执行块级批量量化并完成模型推理,生成模型推理结果。

[0170] 在一个实施例中,提供了一种计算机可读存储介质,其上存储有计算机程序,计算机程序被处理器执行时实现以下步骤:

将输入文本划分为多个处理块,将首个处理块的处理精度格式固定为高精度格式,并禁用对所述首个处理块的量化处理;

对所述多个处理块中除首个处理块以外的其他处理块,通过语言模型生成每个其他处理块的自注意力矩阵,并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和,并将所述全体元素数值之和作为每个token位置的重要性分数;

将重要性分数大于第一阈值的token位置分配为高精度格式,将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式,将重要性分数小于第二阈值的token位置分配为低精度格式;

统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量,选择数量最多的精度格式作为对应处理块的统一量化配置;

将所述语言模型的网络模块划分为多个配置共享组,每个配置共享组至少包含两个网络模块;

在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块;

根据每个处理块对应的统一量化配置,对所有处理块执行块级批量量化并完成模型推理,生成模型推理结果。

[0171] 需要说明的是,上述关于计算机可读存储介质或计算机设备所能实现的功能或步骤,可对应参阅前述方法实施例中,服务端侧以及用户端侧的相关描述,为避免重复,这里不再一一描述。

[0172] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一非易失性计算机可读取存储介质中,该计算机程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0173] 所属领域的技术人员可以清楚地了解到,为了描述的方便和简洁,仅以上述各功能单元、模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能单元、模块完成,即将所述装置的内部结构划分成不同的功能单元或模块,以完成以上描述的全部或者部分功能。

[0174] 应当说明的是,本申请实施例中若出现了非本公司的软件工具或组件,仅仅是用于举例介绍,并不代表实际使用。以上所述实施例仅用以说明本发明的技术方案,而非对其

限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围,均应包含在本发明的保护范围之内。

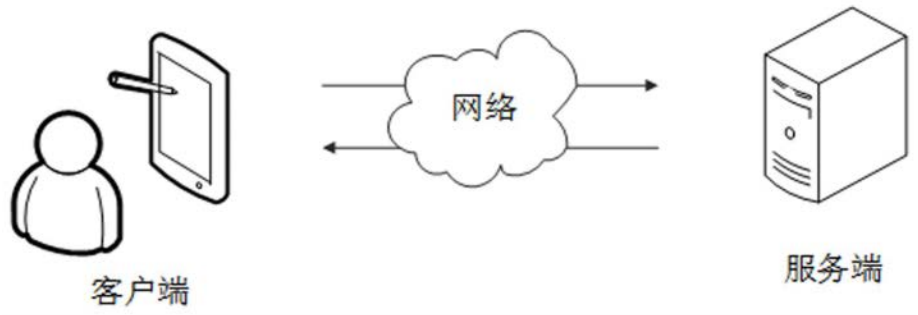
网络
客户端
服务端

图 1


将输入文本划分为多个处理块，将首个处理块的处理精度格式固定为高精度格式，并禁用对所述首个处理块的量化处理
S10
对所述多个处理块中除首个处理块以外的其他处理块，通过语言模型生成每个其他处理块的自注意力矩阵，并确定所述自注意力矩阵中每个token位置对应列的全体元素数值之和，并将所述全体元素数值之和作为每个token位置的重要性分数
S20
将重要性分数大于第一阈值的token位置分配为高精度格式，将重要性分数处于第二阈值之上且处于第一阈值之下的token位置分配为中等精度格式，将重要性分数小于第二阈值的token位置分配为低精度格式
S30
统计每个处理块内被分配为高精度格式、中等精度格式及低精度格式的token位置的数量，选择数量最多的精度格式作为对应处理块的统一量化配置
S40
将所述语言模型的网络模块划分为多个配置共享组，每个配置共享组至少包含两个网络模块
S50
在每个配置共享组内将第一个网络模块对应的处理块的统一量化配置共享给同一配置共享组内的其他网络模块
S60
根据每个处理块对应的统一量化配置，对所有处理块执行块级批量量化并完成模型推理，生成模型推理结果
S70

图 2

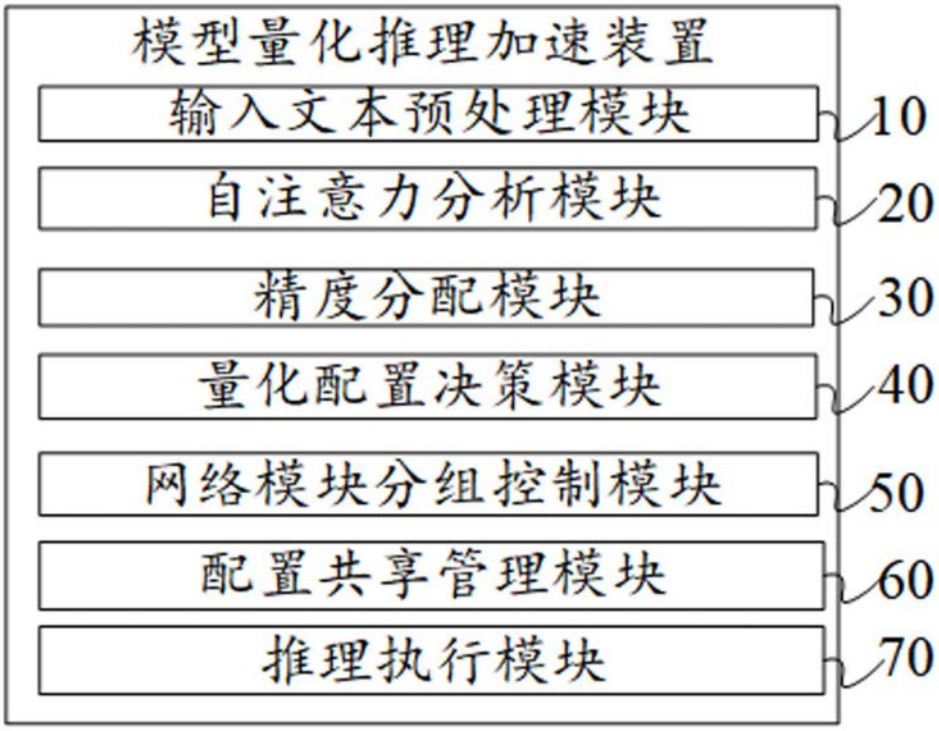
模型量化推理加速装置
输入文本预处理模块
10
自注意力分析模块
20
精度分配模块
30
量化配置决策模块
40
网络模块分组控制模块
50
配置共享管理模块
60
推理执行模块
70

图 3

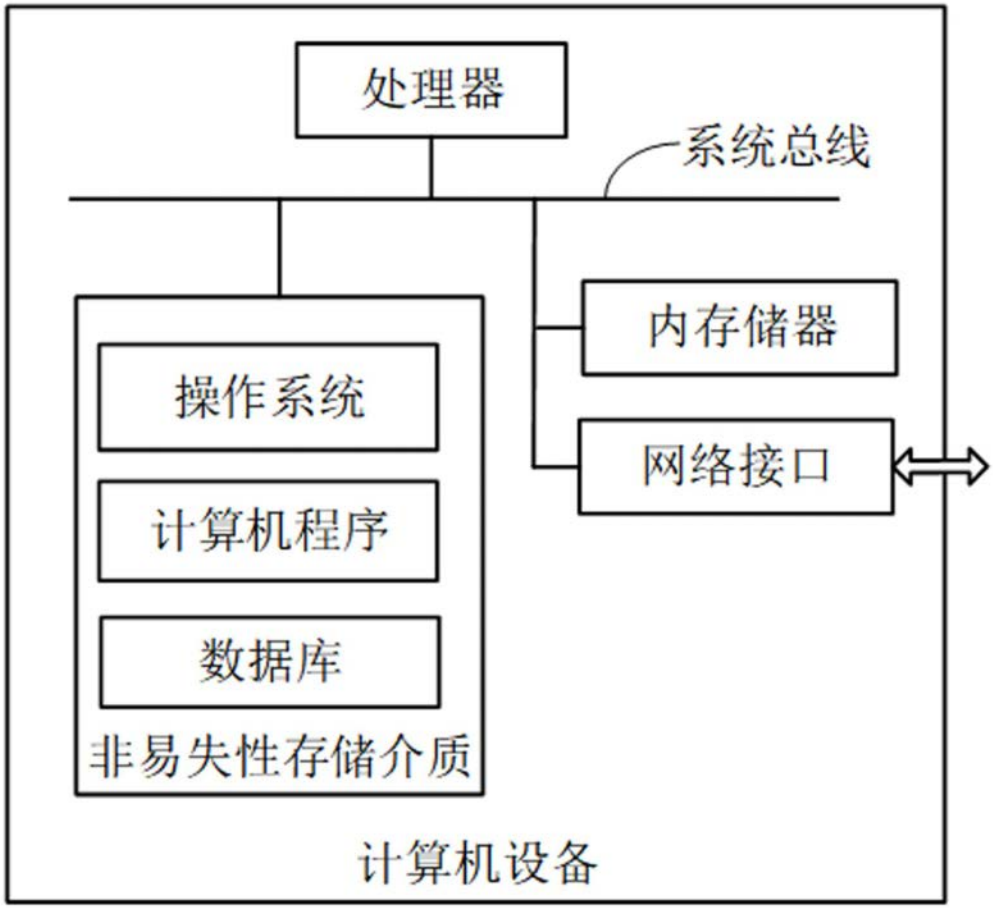

图 4

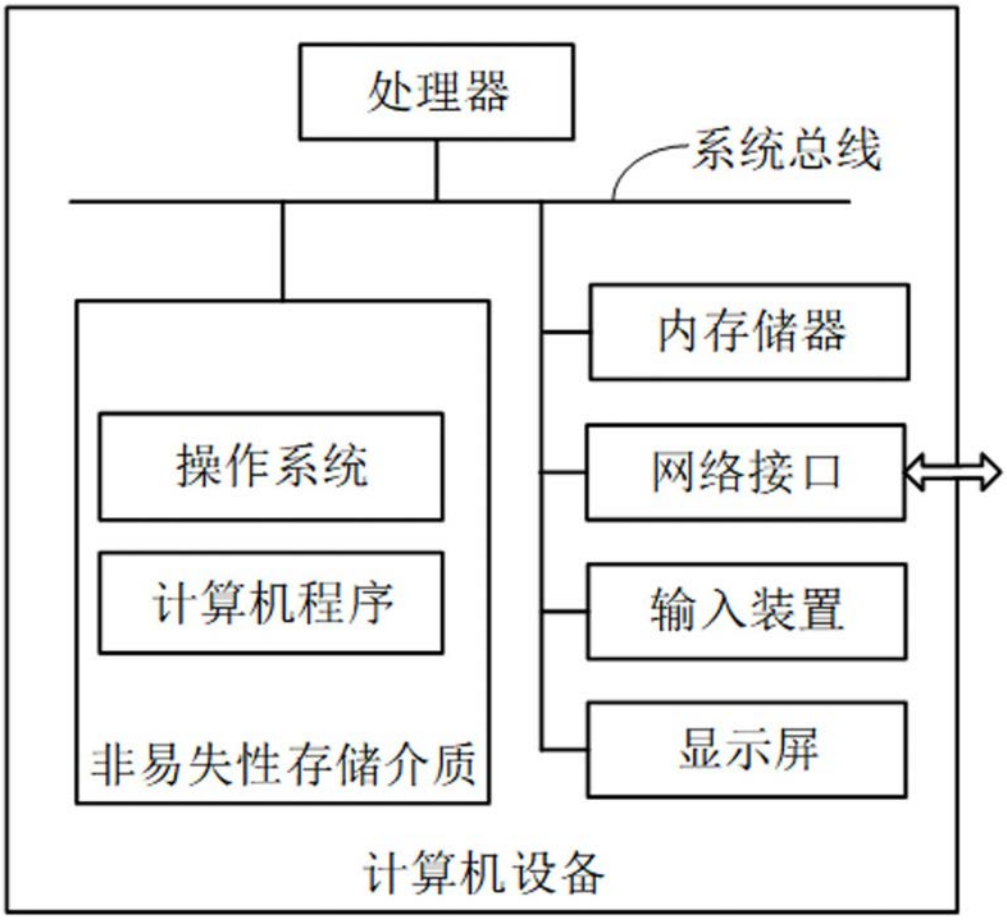

图 5