(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112884146 B
(45)授权公告日 2024.02.13

(21)申请号 202110211440.1

(22)申请日 2021.02.25

(65)同一申请的已公布的文献号
申请公布号 CN 112884146 A

(43)申请公布日 2021.06.01

(73)专利权人 香港理工大学深圳研究院
地址 518057 广东省深圳市南山区粤海街道高新技术产业园南区粤兴一道18号香港理工大学产学研大楼205室

(72)发明人 郭嵩 周祺华 谢鑫

(74)专利代理机构 深圳市君胜知识产权代理事务所(普通合伙) 44268
专利代理师 谢松 徐凯凯

(51)Int.Cl.
G06N 3/0495(2023.01)
G06N 3/0464(2023.01)
G06N 3/084(2023.01)
G06N 3/098(2023.01)

(56)对比文件
CN 110363281 A,2019.10.22
CN 111937010 A,2020.11.13
CN 110555508 A,2019.12.10
CN 111612147 A,2020.09.01
CN 112101097 A,2020.12.18
KR 20020040019 A,2002.05.30
US 2019138882 A1,2019.05.09
US 2020257960 A1,2020.08.13
井小浩.基于深层循环神经网络的陀螺仪降噪方法研究.空间控制技术与应用.2020,第46卷(第5期),全文.
孙浩然.基于参数量化的轻量级图像压缩神经网络研究.信息技术.2020,(第10期),全文.
Xu, YH .Deep Neural Network Compression with Single and Multiple Level Quantization.The 32 AAAI Conference on Artificial Intelligence.2018,第32卷(第1期),全文.

审查员 张天晶

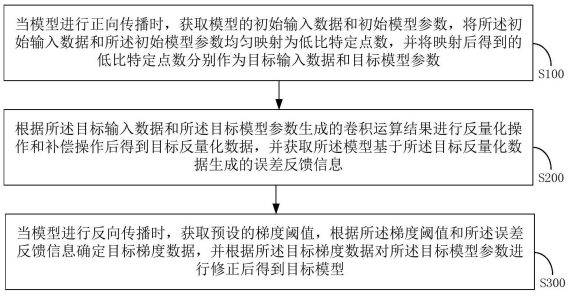
权利要求书3页 说明书9页 附图3页

(54)发明名称

一种训练基于数据量化与硬件加速的模型的方法及系统

(57)摘要

本发明公开了一种训练基于数据量化与硬件加速的模型的方法及边缘智能系统,通过在模型训练的前向传播阶段将边缘智能模型的处理数据转换为低比特定点数,从而使得边缘智能模型的计算成本有效降低,并采用误差补偿机制保障最终模型的质量和推理结果的准确性。在模型训练的后向传播阶段采用梯度截断机制,保障模型更新过程的平稳性。解决了现有技术中边缘智能设备端的模型在训练与推理过程中的计算与存储开销较大,且模型的预测准确性较低,难以胜任高动态的实时性任务的问题。



CN 112884146 B

1.一种训练基于数据量化与硬件加速的模型的方法,其特征在于,所述方法应用于图像分类的模型,模型的输入数据为图像数据,输出数据为图像数据对应的不同类别的概率;所述方法包括:

当模型进行正向传播时,获取模型的初始输入数据和初始模型参数,将所述初始输入数据和所述初始模型参数由多比特浮点数均匀映射为低比特定点数,并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数;

根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据,其中,所述反量化操作后得到初始反量化数据,所述补偿操作包括:将所述初始反量化数据输入补偿层,所述补偿层用于修正量化导致的误差;获取所述补偿层的数据分布期望数据、补偿缩放因子、补偿平移量数据;将所述补偿缩放因子与所述数据分布期望数据的乘积与所述补偿平移量数据相加后得到的数据作为所述补偿层对应的补偿项数据;根据所述补偿项数据对所述初始反量化数据进行补偿操作以后得到待调整反量化数据;对所述待调整反量化数据进行基于正态分布的调整操作后得到目标反量化数据;

获取所述模型基于所述目标反量化数据生成的误差反馈信息;

当模型进行反向传播时,获取预设的梯度阈值,根据所述梯度阈值和所述误差反馈信息确定目标梯度数据,并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

2.根据权利要求1所述的一种训练基于数据量化与硬件加速的模型的方法,其特征在于,所述当模型进行正向传播时,获取模型的初始输入数据和初始模型参数,将所述初始输入数据和所述初始模型参数由多比特浮点数均匀映射为低比特定点数,并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数包括:

当模型进行正向传播时,获取初始输入数据和初始模型参数;

对所述初始输入数据依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标输入数据;

对所述初始模型参数依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标模型参数。

3.根据权利要求2所述的一种训练基于数据量化与硬件加速的模型的方法,其特征在于,所述对所述初始输入数据依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标输入数据包括:

获取所述初始输入数据的数据分布信息,根据所述初始输入数据的数据分布信息确定第一缩放因子、第一平移量数据;

将所述初始输入数据与所述第一缩放因子相除后得到第一输入数据;

根据所述第一平移量数据对所述第一输入数据进行平移操作后得到第二输入数据;

获取预设的离散化取整区间,根据所述离散化取整区间对所述第二输入数据进行离散化取整操作后得到低比特定点数形式的目标输入数据。

4.根据权利要求3所述的一种训练基于数据量化与硬件加速的模型的方法,其特征在于,所述对所述初始模型参数依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标模型参数包括:

获取模型的初始模型参数的数据分布信息,根据所述初始模型参数的数据分布信息确定第二缩放因子、第二平移量数据;

将所述初始模型参数与所述第二缩放因子相除后得到第一模型参数;

根据所述第二平移量数据对所述第一模型参数进行平移操作后得到第二模型参数;

获取预设的离散化取整区间,根据所述离散化取整区间对所述第二模型参数进行离散化取整操作后得到低比特定点数形式的目标模型参数。

5.根据权利要求4所述的一种训练基于数据量化与硬件加速的模型的方法,其特征在于,所述根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据,并获取所述模型基于所述目标反量化数据生成的误差反馈信息包括:

根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作后得到初始反量化数据;

对所述初始反量化数据进行补偿操作后得到目标反量化数据;

获取所述模型基于所述目标反量化数据生成的最终层输出数据;

将所述最终层输出数据输入所述模型的损失函数中,并获取所述损失函数基于所述最终层输出数据生成的误差反馈信息。

6.根据权利要求5所述的一种训练基于数据量化与硬件加速的模型的方法,其特征在于,所述根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作后得到初始反量化数据包括:

根据所述目标输入数据和所述目标模型参数进行卷积操作后得到卷积运算数据;

根据所述第一平移量数据与所述第二平移量数据之和对所述卷积运算数据进行平移操作后得到平移数据;

将所述第一缩放因子与所述第二缩放因子之积与所述平移数据相乘后得到初始反量化数据。

7.根据权利要求1所述的一种训练基于数据量化与硬件加速的模型的方法,其特征在于,所述当模型进行反向传播时,获取预设的梯度阈值,根据所述梯度阈值和所述误差反馈信息确定目标梯度数据,并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型包括:

当模型进行反向传播时,根据所述误差反馈信息获取所述目标模型参数对应的梯度数据;

获取预设的梯度阈值,将所述目标模型参数对应的梯度数据与所述梯度阈值进行比较;

当所述目标模型参数对应的梯度数据大于所述梯度阈值时,将所述梯度阈值作为所述目标模型参数对应的目标梯度数据;

根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

8.一种训练基于数据量化与硬件加速的模型的系统,其特征在于,所述系统应用于图像分类的模型,模型的输入数据为图像数据,输出数据为图像数据对应的不同类别的概率;所述系统包括:

正向传播模块,用于当模型进行正向传播时,获取模型的初始输入数据和初始模型参

数,将所述初始输入数据和所述初始模型参数由多比特浮点数均匀映射为低比特定点数,并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数;

误差补偿模块,用于根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据,其中,所述反量化操作后得到初始反量化数据,所述补偿操作包括:将所述初始反量化数据输入补偿层,所述补偿层用于修正量化导致的误差;获取所述补偿层的数据分布期望数据、补偿缩放因子、补偿平移量数据;将所述补偿缩放因子与所述数据分布期望数据的乘积与所述补偿平移量数据相加后得到的数据作为所述补偿层对应的补偿项数据;根据所述补偿项数据对所述初始反量化数据进行补偿操作以后得到待调整反量化数据;对所述待调整反量化数据进行基于正态分布的调整操作后得到目标反量化数据;

获取所述模型基于所述目标反量化数据生成的误差反馈信息;

反向传播模块,用于模型当进行反向传播时,获取预设的梯度阈值,根据所述梯度阈值和所述误差反馈信息确定目标梯度数据,并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

一种训练基于数据量化与硬件加速的模型的方法及系统

技术领域

[0001] 本发明涉及机器学习领域,尤其涉及的是一种训练基于数据量化与硬件加速的模型的方法及系统。

背景技术

[0002] 目前的边缘智能往往针对特定的应用场景进行设计,同时需要额外专门硬件的支持,缺乏算法的可移植性、接口的易用性与模型的通用性。此外,目前的方法大多基于数据模拟层面的算法设计,难以真正发挥底层硬件的加速性能。同时,机器学习应用通常涉及到模型参数的训练,以往的方法大多用于预测与推理,不适用于网络的训练环境,因此无法满足现实场景下边缘智能的实时性与动态性需求。简言之,现有的边缘智能设备端的模型在训练与推理过程中的计算与存储开销较大,难以实现底层硬件的加速,且模型的预测准确性较低,难以胜任高动态的实时性任务。

[0003] 因此,现有技术还有待改进和发展。

发明内容

[0004] 本发明要解决的技术问题在于,针对现有技术的上述缺陷,提供一种训练基于数据量化与硬件加速的模型的方法及系统,旨在解决现有技术中边缘智能设备端的模型在训练与推理过程中的计算与存储开销较大,且模型的预测准确性较低,难以胜任高动态的实时性任务的问题。

[0005] 本发明解决问题所采用的技术方案如下:

[0006] 第一方面,本发明实施例提供一种训练基于数据量化与硬件加速的模型的方法,其中,所述方法包括:

[0007] 当模型进行正向传播时,获取模型的初始输入数据和初始模型参数,将所述初始输入数据和所述初始模型参数均匀映射为低比特定点数,并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数;

[0008] 根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据,并获取所述模型基于所述目标反量化数据生成的误差反馈信息;

[0009] 当模型进行反向传播时,获取预设的梯度阈值,根据所述梯度阈值和所述误差反馈信息确定目标梯度数据,并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

[0010] 在一种实施方式中,所述当模型进行正向传播时,获取模型的初始输入数据和初始模型参数,将所述初始输入数据和所述初始模型参数均匀映射为低比特定点数,并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数包括:

[0011] 当模型进行正向传播时,获取初始输入数据和初始模型参数;

[0012] 对所述初始输入数据依次进行缩放操作、平移操作以及离散化取整操作后得到低

比特定点数形式的目标输入数据；

[0013] 对所述初始模型参数依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标模型参数。

[0014] 在一种实施方式中，所述对所述初始输入数据依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标输入数据包括：

[0015] 获取所述初始输入数据的数据分布信息，根据所述初始输入数据的数据分布信息确定第一缩放因子、第一平移量数据；

[0016] 将所述初始输入数据与所述第一缩放因子相除后得到第一输入数据；

[0017] 根据所述第一平移量数据对所述第一输入数据进行平移操作后得到第二输入数据；

[0018] 获取预设的离散化取整区间，根据所述离散化取整区间对所述第二输入数据进行离散化取整操作后得到低比特定点数形式的目标输入数据。

[0019] 在一种实施方式中，所述对所述初始模型参数依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标模型参数包括：

[0020] 获取模型的初始模型参数的数据分布信息，根据所述初始模型参数的数据分布信息确定第二缩放因子、第二平移量数据；

[0021] 将所述初始模型参数与所述第二缩放因子相除后得到第一模型参数；

[0022] 根据所述第二平移量数据对所述第一模型参数进行平移操作后得到第二模型参数；

[0023] 获取预设的离散化取整区间，根据所述离散化取整区间对所述第二模型参数进行离散化取整操作后得到低比特定点数形式的目标模型参数。

[0024] 在一种实施方式中，所述根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据，并获取所述模型基于所述目标反量化数据生成的误差反馈信息包括：

[0025] 根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作后得到初始反量化数据；

[0026] 对所述初始反量化数据进行补偿操作后得到目标反量化数据；

[0027] 获取所述模型基于所述目标反量化数据生成的最终层输出数据；

[0028] 将所述最终层输出数据输入所述模型的损失函数中，并获取所述损失函数基于所述最终层输出数据生成的误差反馈信息。

[0029] 在一种实施方式中，所述根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作后得到初始反量化数据包括：

[0030] 根据所述目标输入数据和所述目标模型参数进行卷积操作后得到卷积运算数据；

[0031] 根据所述第一平移量数据与所述第二平移量数据之和对所述卷积运算数据进行平移操作后得到平移数据；

[0032] 将所述第一缩放因子与所述第二缩放因子之积与所述平移数据相乘后得到初始反量化数据。

[0033] 在一种实施方式中，所述对所述初始反量化数据进行补偿操作后得到目标反量化数据包括：

[0034] 将所述初始反量化数据输入补偿层；

[0035] 获取所述补偿层对应的补偿项数据；

[0036] 根据所述补偿项数据对所述初始反量化数据进行补偿操作以后得到待调整反量化数据；

[0037] 对所述待调整反量化数据进行基于正态分布的调整操作后得到目标反量化数据。

[0038] 在一种实施方式中，所述获取所述补偿层对应的补偿项数据包括：

[0039] 获取所述补偿层的数据分布期望数据、补偿缩放因子、补偿平移量数据；

[0040] 将所述补偿缩放因子与所述数据分布期望数据的乘积与所述补偿平移量数据相加后得到的数据作为所述补偿层对应的补偿项数据。

[0041] 在一种实施方式中，所述当模型进行反向传播时，获取预设的梯度阈值，根据所述梯度阈值和所述误差反馈信息确定目标梯度数据，并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型包括：

[0042] 当模型进行反向传播时，根据所述误差反馈信息获取所述目标模型参数对应的梯度数据；

[0043] 获取预设的梯度阈值，将所述目标模型参数对应的梯度数据与所述梯度阈值进行比较；

[0044] 当所述目标模型参数对应的梯度数据大于所述梯度阈值时，将所述梯度阈值作为所述目标模型参数对应的目标梯度数据；

[0045] 根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

[0046] 第二方面，本发明实施例还提供一种训练基于数据量化与硬件加速的模型的系统，其中，所述系统包括：

[0047] 正向传播模块，用于当模型进行正向传播时，获取模型的初始输入数据和初始模型参数，将所述初始输入数据和所述初始模型参数均匀映射为低比特定点数，并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数；

[0048] 误差补偿模块，用于根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据，并获取所述模型基于所述目标反量化数据生成的误差反馈信息；

[0049] 反向传播模块，用于当模型进行反向传播时，获取预设的梯度阈值，根据所述梯度阈值和所述误差反馈信息确定目标梯度数据，并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

[0050] 本发明的有益效果：本发明实施例通过在模型训练的前向传播阶段将边缘智能模型的处理数据转换为低比特定点数，从而使得边缘智能模型的计算成本有效降低，并采用误差补偿机制保障最终模型的质量和推理结果的准确性。在模型训练的后向传播阶段采用梯度截断机制，保障模型更新过程的平稳性。解决了现有技术中边缘智能设备端的模型在训练与推理过程中的计算与存储开销较大，且模型的预测准确性较低，难以胜任高动态的实时性任务的问题。

附图说明

[0051] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现

有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明中记载的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0052] 图1是本发明实施例提供的一种训练基于数据量化与硬件加速的模型的方法的流程示意图。

[0053] 图2是本发明实施例提供的数据量化与硬件加速的模型的内部结构示意图。

[0054] 图3是本发明实施例提供的一种训练基于数据量化与硬件加速的系统的模块示意图。

具体实施方式

[0055] 为使本发明的目的、技术方案及优点更加清楚、明确,以下参照附图并举实施例对本发明进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0056] 随着万物互联时代的到来,网络边缘设备产生的数据量快速增加,带来了更高的数据传输带宽需求,传统云计算模型要求应用数据传送到云计算中心,再请求数据处理结果,不仅增大了系统延迟还造成了网络带宽很大的数据传输压力需,所以传统云计算模型已经无法有效应对当前新型应用对数据处理的实时性要求,因此,边缘计算应运而生。边缘智能是边缘节点在边缘侧提供的高级数据分析、场景感知、实时决策、自组织与协同等服务,属于机器学习的范畴,其目标是在各类移动设备、物联网传感器、手持终端等场景下部署高性能的人工智能应用。

[0057] 边缘智能要求从根本上实现基于端到端的自动学习范式,即从输入端到输出端会得到一个预测结果,与真实结果相比较会得到一个误差,这个误差会在模型中的每一层反向传播,每一层的表示都会根据这个误差来做调整,直到模型收敛或者达到预期的效果才结束,一个典型的端到端网络就是神经网络。

[0058] 然而,目前的边缘智能往往针对特定的应用场景进行设计,同时需要额外专门硬件的支持,缺乏算法的可移植性、接口的易用性与模型的通用性。此外,目前的方法大多基于数据模拟层面的算法设计,难以真正发挥底层硬件的加速性能。同时,机器学习应用通常涉及到模型参数的训练,以往的方法大多用于预测与推理,不适用于网络的训练环境,因此无法满足现实场景下边缘智能的实时性与动态性需求。

[0059] 简言之,现有的边缘智能设备端的模型在训练与推理过程中的计算与存储开销较大,难以实现底层硬件的加速,且模型的预测准确性较低,难以胜任高动态的实时性任务。

[0060] 针对现有技术的上述缺陷,本发明提供了一种训练基于数据量化与硬件加速的模型的方法及系统,通过在模型训练的前向传播阶段将边缘智能模型的处理数据转换为低比特定点数,从而使得边缘智能模型的计算成本有效降低,并采用误差补偿机制保障最终模型的质量和推理结果的准确性。在模型训练的后向传播阶段采用梯度截断机制,保障模型更新过程的平稳性。解决了现有技术中边缘智能设备端的模型在训练与推理过程中的计算与存储开销较大,且模型的预测准确性较低,难以胜任高动态的实时性任务的问题。

[0061] 如图1所示,所述方法包括如下步骤:

[0062] 步骤S100、当模型进行正向传播时,获取模型的初始输入数据和初始模型参数,将

所述初始输入数据和所述初始模型参数均匀映射为低比特定点数,并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数。

[0063] 具体地,模型的正向传播是指对神经网络沿着输入层到输出层的顺序,依次计算并存储模型的中间变量以及输出,例如如图2所示,其中正向传播按从左往右,即从第一层计算到最后一层的顺序执行。为了降低边缘智能设备端的计算成本,因此本实施例中需要首先将边缘智能设备所处理的数据转换为低比特定点数的表达形式。具体地,首先需要获取边缘智能设备端的初始输入数据和初始模型参数,这两种数据的数据形式为多比特浮点数,因此需要先将这两种数据进行非对称量化,将它们从原本的多比特浮点数形式均匀映射为低比特定点数,即相当于将原本更多字节的数据压缩为更少字节的数据。举例说明,当初始输入数据和初始模型参数为32比特浮点数,可以将32比特浮点数映射为8比特定点数,即相当于将4字节的数据压缩为1字节,大多数的处理器对8比特定点数的运算更快,因此映射为8比特顶点数后再进行计算拥有更好的实现效率,且计算所消耗的能量更少。映射完毕以后,再将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数,就可以达到降低模型的计算开销的目的。

[0064] 在一种实现方式中,所述步骤S100具体包括如下步骤:

[0065] 步骤S110、当模型进行正向传播时,获取初始输入数据和初始模型参数;

[0066] 步骤S120、对所述初始输入数据依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标输入数据;

[0067] 步骤S130、对所述初始模型参数依次进行缩放操作、平移操作以及离散化取整操作后得到低比特定点数形式的目标模型参数。

[0068] 为了实现将所述初始输入数据和初始模型参数转换为低比特定点数形式的目标输入数据和目标模型参数,本实施例需要对所述初始输入数据进行缩放操作、平移操作以及离散化取整操作。具体地,本实施例首先需要获取所述初始输入数据的数据分布信息,模型的输入数据通常是一系列高维向量(矩阵)的形式,因此对于每个向量,数据的分布信息通过数值的频次统计即可得到。然后根据所述初始输入数据的数据分布信息来确定第一缩放因子和第一平移量数据,可以理解的是缩放因子可以用于指示数据的缩放程度,而平移量则可以用于指示数据的平移程度。然后将所述初始输入数据与所述第一缩放因子相除后得到第一输入数据,并根据所述第一平移量数据对所述第一输入数据进行平移操作后得到第二输入数据,为了使所述第二输入数据为定点数形式,本实施例还需要对所述第二输入数据进行进行离散化取整操作。具体地,系统中预先设定好了离散化取整的区间,获取该离散化取整区间对所述第二输入数据进行离散化取整操作后就可以得到8比特定点数形式的目标输入数据。举例说明,假设初始输入数据为32比特的浮点数,本实施例需要将32比特的浮点数转换为8比特的定点数。首先根据该初始输入数据的数据分布情况确定其对应的缩放因子和平移量,然后将原本的32比特的浮点数除以该缩放因子后再根据该平移量进行平移,然后获取系统内预设的离散化取整区间[-127,+127]或[0,255],将平移后的数据的值域约束在[-127,+127]或[0,255]区间内,再进行取整数的操作,使得数据都以整数形式表达,由于所有的数据只有256种不同的取值可能性,即2的8次方,因此可以被8比特所容纳。

[0069] 此外,本实施例还需要将初始模型参数也转换为低比特定点数形式的目标模型参数,转换过程与上述初始输入数据的转换过程相似。简单来说,同样需要首先获取模型的初

始模型参数的数据分布信息,然后根据所述初始模型参数的数据分布信息确定第二缩放因子、第二平移量数据,再将所述初始模型参数与所述第二缩放因子相除后得到第一模型参数,然后根据所述第二平移量数据对所述第一模型参数进行平移操作后得到第二模型参数。最后获取预设的离散化取整区间,根据所述离散化取整区间对所述第二模型参数进行离散化取整操作后得到低比特定点数形式的目标模型参数。需要说明的是,在量化过程中,模型会自动调整平移操作以及离散化取整操作,使得量化后的数据分布更加贴近真实值。

[0070] 如图1所示,所述方法还包括:

[0071] 步骤S200、根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据,并获取所述模型基于所述目标反量化数据生成的误差反馈信息。

[0072] 具体地,当得到目标输入数据和目标模型参数以后,通过模型中的卷积层对目标输入数据和目标模型参数进行计算,就可以完成一次模型的前向计算过程,然后需要对卷积层得到的卷积运算结果进行反量化操作,将定点数回算到浮点数域中,再依次传递给下一层进行相应的仿射运算。举例说明,如图2所示,神经网络是层叠堆积的结构,主要有若干个卷积层和全连接层组成,图中展示的1个卷积层加2个全连接层的神经网络。卷积层主要是用于计算数据特征,而全连接层主要用于对卷积层得到的数据特征进行矩阵内积,将数据从高维形式转换为低维形式,并输出一维向量,从而以用户可以理解的形式来表达推理的结果。其中,每个卷积层和全连接层(除了最后一个全连接层)内部的末尾都有一个激活函数。激活函数主要采用修正线性单元(Rectified Linear Unit,ReLU),其将小于0的输入数据全部变为0,大于0的输入数据则原值保留,从而使得神经网络能够根据不同的输入值具有更高的区分能力。而最后一个全连接层输出的结果会进入该模型的损失函数中,与预先设置的真实标签进行误差对比,从而衡量模型的训练效果,并根据衡量的结果来修正模型中各层的参数。

[0073] 在一种实现方式中,所述步骤S200具体包括如下步骤:

[0074] 步骤S210、根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作后得到初始反量化数据;

[0075] 步骤S220、对所述初始反量化数据进行补偿操作后得到目标反量化数据;

[0076] 步骤S230、获取所述模型基于所述目标反量化数据生成的最终层输出数据;

[0077] 步骤S240、将所述最终层输出数据输入所述模型的损失函数中,并获取所述损失函数基于所述最终层的输出数据生成的误差反馈信息。

[0078] 为了获得误差反馈信息,本实施例首先需要根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作后得到初始反量化数据。可以理解的是反量化操作相当于量化操作的逆过程.具体地,首先将所述目标输入数据和所述目标模型参数输入卷积层中进行卷积计算得到卷积运算结果,然后开始对所述卷积运算结果进行反量化操作:首先根据前述量化操作中确定的所述第一平移量数据与所述第二平移量数据之和对所述卷积运算数据进行平移操作,并得到平移数据。然后,再获取前述量化操作中确定的所述第一缩放因子与所述第二缩放因子之积,将所述第一缩放因子与所述第二缩放因子之积与所述平移数据相乘以后得到初始反量化数据。

[0079] 由于数据在卷积过程中会丢失原有的计算精度,因此本实施例引入了误差补偿机

制来修正量化导致的误差。本实施例将补偿操作封装成了一个专门的层结构，名为补偿层。然后将所述初始反量化数据输入所述补偿层中，对所述初始反量化数据进行补偿操作后得到目标反量化数据。具体地，首先本实施例需要确定补偿层中的补偿项数据，所述补偿项数据相当于对所述初始反量化数据进行补偿操作时需要用到的参数。在一种实现方式中，所述补偿项数据主要由三种类型的参数组成，即补偿层的数据分布期望数据、补偿缩放因子和补偿平移量数据，然后通过将所述补偿缩放因子与所述数据分布期望数据的乘积与所述补偿平移量数据相加，得到本实施例所需的补偿项数据。需要说明的是，本实施例中补偿层的数据分布期望数据、补偿缩放因子和补偿平移量数据的具体数值均与前述量化操作或者反量化操作中使用的缩放因子、平移量数据等无关，本实施例正是通过补偿层对所述反量化操作后得到的数据进行调整，实现了逐层且动态的补偿，从而提高了模型的预测准确性。

[0080] 根据所述补偿项数据对所述初始反量化数据进行补偿操作以后，还需要对该数据进行调整，才能生成目标反量化数据。具体地，本实施例将进行补偿操作后得到的数据作为待调整反量化数据，然后需要将所述待调整反量化数据输入模型中的归一化层，使得所述待调整反量化数据在所述归一化层中进行基于正态分布的调整操作。概括地讲，基于正态分布的调整操作是将数据调整为满足期望为0，方差为1的正态分布。调整前的数据分布是杂乱不确定的，调整后是一个比较光滑、近似正态分布、中轴在0附近的、两侧很稀疏的钟型分布。从而使得模型参数能够适应不同初始化条件下的训练，并且加快模型的收敛速率。调整完毕以后，即得到目标反量化数据。

[0081] 然后将所述目标反量化数据卷积层的下一层。例如如图2所示，将所述目标反量化数据输入激活函数层，并传递给后续的卷积和全连接层，并对全连接层的输入数据进行相应的仿射运算，具体为针对一个向量(矩阵)，进行一次线性变换，然后接上一个平移，使该向量变换到另一个向量空间。最后模型会基于所述目标反量化数据生成最终层输出数据，该输出结果代表了模型对于某一具体任务的推断结果。举例说明，假设当前神经网络的目标是对图像数据进行分类，则该模型的最终层输出数据就是输入图像数据对应的不同类别的概率，且所有概率相加的和为1。

[0082] 得到最终层输出数据以后，为了衡量模型的训练效果，还需要将所述最终层输出数据输入该模型的损失函数中，然后获取所述损失函数基于所述最终层输出数据生成的误差反馈信息。具体地，在机器学习的过程中如果参数过多，模型过于复杂，则容易产生过拟合的问题，即模型在训练样本数据上表现的很好，但在实际测试样本上表现的较差，不具备良好的泛化能力，因此在一种实现方式中，本实施例在所述损失函数中引入基于所述补偿层的L2正则项。所述L2正则项的目的是限制参数过多或者过大，避免模型更加复杂。可以理解的是，当提高该L2正则项的权重时，将扩大补偿层对模型的影响。反之，当降低该L2正则项的权重时，将减小补偿层对模型的影响。

[0083] 如图1所示，所述方法还包括如下步骤：

[0084] 步骤S300、当模型进行反向传播时，获取预设的梯度阈值，根据所述梯度阈值和所述误差反馈信息确定目标梯度数据，并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

[0085] 完成模型的正向传播以后，还需要进行模型的反向传播。反向传播指的是计算神经网络参数梯度的方法，主要是针对神经网络优化的过程中进行。概括地讲，反向传播阶段

会根据获取到的误差反馈信息,利用动态梯度下降的优化方法对模型参数进行更新,使得模型能够迭代式地逼近最优值。如图2所示,与正向传播的顺序相反,反向传播从神经网络的尾部开始,向前推进。每一层的参数在反向传播过程中都会得到对应的梯度,算法将利用该梯度对模型进行修正。因此在反向传播阶段,正向传播涉及到的量化操作、补偿操作和归一化层中的调整操作的参数都将得到更新。由于梯度值过大将会对模型更新的过程产生较大的波动,因此本实施例还预先设定一个梯度阈值,通过所述梯度阈值对每层的梯度数据进行约束,从而使得模型更新过程更加平滑。

[0086] 在一种实现方式中,所述步骤S300具体包括如下步骤:

[0087] 步骤S310、当模型进行反向传播时,根据所述误差反馈信息获取所述目标模型参数对应的梯度数据;

[0088] 步骤S320、获取预设的梯度阈值,将所述目标模型参数对应的梯度数据与所述梯度阈值进行比较;

[0089] 步骤S330、当所述目标模型参数对应的梯度数据大于所述梯度阈值时,将所述梯度阈值作为所述目标模型参数对应的目标梯度数据;

[0090] 步骤S340、根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

[0091] 具体地,为了对边缘智能模型中各层的参数进行优化,本实施例需要在反向传播阶段中,根据在前向传播阶段获取到的误差反馈信息计算出神经网络各层的参数(即步骤S100中获取到的目标模型参数)对应的梯度数据,然后获取预先设定好的梯度阈值,判断计算出的梯度数据的数值是否超出所述梯度阈值,若未超出所述梯度阈值,则表示所述梯度数据的数值不大,并不会对模型的更新阶段产生较大的波动,因此可以保留计算的梯度数据;若超出所述梯度阈值,则表示所述梯度数据的数值较大,有可能会对模型的更新阶段产生较大的波动,因此需要对所述梯度数据进行截取,将其约束为所述梯度阈值的大小,从而保证所述梯度数据的值域被控制在梯度阈值范围内,使得模型的更新过程更加平滑。

[0092] 本实施例将最终确定好的梯度数据作为目标梯度数据,获取到所述目标梯度数据以后,还需要根据所述目标梯度数据对所述目标模型参数进行修正后才可以得到目标模型。具体地,在获取到所述目标梯度数据以后,还需要获取预设的学习率数据,所述学习率数据属于神经网络中的超参数,用于控制模型的更新幅度,例如可以将所述学习率设置为0.01。然后将所述学习率数据与所述目标梯度数据的乘积作为修改值,根据所述修改值对目标模型参数进行修正,具体地,可以将旧的目标模型参数与所述修改值的差值作为新的目标模型参数。修正完毕以后即得到目标模型。

[0093] 综合上述实施例来看,本发明能够有效保障模型在低比特定点数值域下的训练精度和收敛效率。具体地,由于本发明将应用执行过程中涉及的数据转换为低比特定点数表达形式,使得卷积操作在定点数格式下进行,因此能够充分发挥底层硬件的加速性能,从而有效降低系统的计算负载、内存用量、总线带宽、电量能耗等方面的开销,提升模型响应速度和推理吞吐量。此外,由于本发明还引入了补偿层以及梯度阈值的相关技术特征,因此能保障模型的预测准确性以及更新过程的平稳性。从而整体降低系统在计算负载、内存用量、总线带宽、电量能耗等方面的开销。

[0094] 鉴于本发明的上述优点,可以解决实际应用中的以下问题:

[0095] 1.使得边缘智能应用能够在实际场景中得到部署,推动相关产业的落地。

[0096] 2.节省了边缘设备上的各类资源开销,计算负载、内存用量、总线带宽、电量能耗等。

[0097] 3.保障了在资源受限的边缘设备端的模型质量与预测准确性,使得边缘智能能够胜任高动态的实时性任务。

[0098] 4.保护了设备端的数据和用户隐私,避免了传统云端智能范式的网络瓶颈与昂贵的资源消耗。

[0099] 基于上述实施例,本发明还提供了一种训练基于数据量化与硬件加速的模型的系统,如图3所示,所述系统包括:

[0100] 正向传播模块01,用于当模型进行正向传播时,获取模型的初始输入数据和初始模型参数,将所述初始输入数据和所述初始模型参数均匀映射为低比特定点数,并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数;

[0101] 误差补偿模块02,用于根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据,并获取所述模型基于所述目标反量化数据生成的误差反馈信息;

[0102] 反向传播模块03,用于当模型进行反向传播时,获取预设的梯度阈值,根据所述梯度阈值和所述误差反馈信息确定目标梯度数据,并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型。

[0103] 综上所述,本发明公开了一种训练基于数据量化与硬件加速的模型的方法,通过在模型训练的前向传播阶段将边缘智能模型的处理数据转换为低比特定点数,从而使得边缘智能模型的计算成本有效降低,并采用误差补偿机制保障最终模型的质量和推理结果的准确性。在模型训练的后向传播阶段采用梯度截断机制,保障模型更新过程的平稳性。解决了现有技术中边缘智能设备端的模型在训练与推理过程中的计算与存储开销较大,且模型的预测准确性较低,难以胜任高动态的实时性任务的问题。

[0104] 应当理解的是,本发明的应用不限于上述的举例,对本领域普通技术人员来说,可以根据上述说明加以改进或变换,所有这些改进和变换都应属于本发明所附权利要求的保护范围。

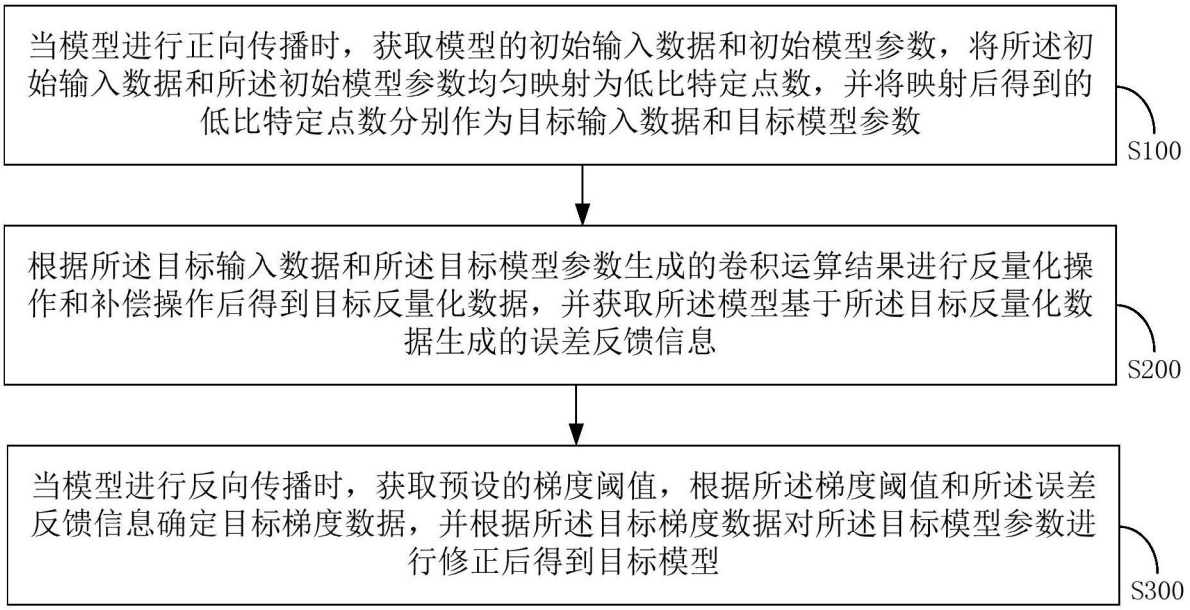

当模型进行正向传播时，获取模型的初始输入数据和初始模型参数，将所述初始输入数据和所述初始模型参数均匀映射为低比特定点数，并将映射后得到的低比特定点数分别作为目标输入数据和目标模型参数
S100
根据所述目标输入数据和所述目标模型参数生成的卷积运算结果进行反量化操作和补偿操作后得到目标反量化数据，并获取所述模型基于所述目标反量化数据生成的误差反馈信息
S200
当模型进行反向传播时，获取预设的梯度阈值，根据所述梯度阈值和所述误差反馈信息确定目标梯度数据，并根据所述目标梯度数据对所述目标模型参数进行修正后得到目标模型
S300

图1

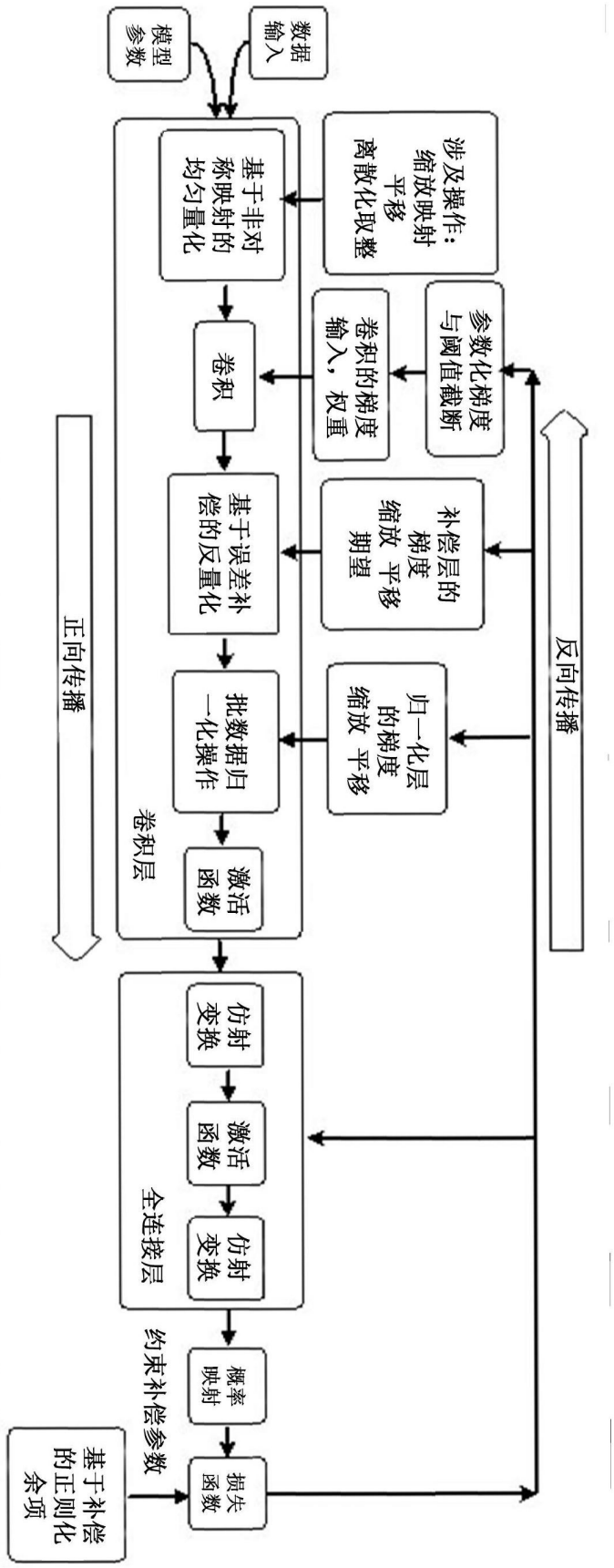

模型参数
数据输入
涉及操作：缩放映射平移离散化取整
基于非对称映射的均匀量化
参数化梯度与阈值截断
卷积的梯度输入，权重
卷积
补偿层的梯度缩放 平移期望
基于误差补偿的反量化
归一化层的梯度缩放 平移
批数据归一化操作
激活函数
卷积层
正向传播
反向传播
仿射变换
激活函数
仿射变换
全连接层
概率映射
约束补偿参数
损失函数
基于补偿的正则化余项

图2

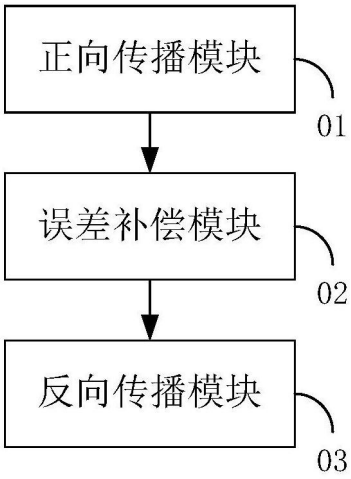


图3