(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 112232497 A
(43)申请公布日 2021.01.15

(21)申请号 202011083320.X

(22)申请日 2020.10.12

(71)申请人 苏州浪潮智能科技有限公司
地址 215100 江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢

(72)发明人 沈付旺 景璐

(74)专利代理机构 北京连和连知识产权代理有限公司 11278
代理人 杨帆 张腾

(51)Int.Cl.
G06N 3/063(2006.01)
G06F 8/41(2018.01)

权利要求书2页 说明书9页 附图2页

(54)发明名称

一种编译AI芯片的方法、系统、设备及介质

(57)摘要

本发明公开了一种编译AI芯片的方法、系统、设备和存储介质，方法包括：对AI芯片的网络模型进行量化，并基于网络模型的权值对量化后的网络模型进行精度调整；在精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化网络模型的计算流程；以及获取硬件架构参数，并根据硬件架构参数和网络模型生成连续的计算流。本发明通过对AI芯片进行量化并进行精度调整，使得性能更加完善；最大程度的优化网络的计算，减少计算的步骤，最大化MAC计算单元的利用率；通过生成可使计算设备在网络推理过程中一直处于忙碌状态的计算流，提高了AI设备的利用率。

CN 112232497 A

1.一种编译AI芯片的方法,其特征在于,包括以下步骤:

对AI芯片的网络模型进行量化,并基于所述网络模型的权值对量化后的网络模型进行精度调整;

在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程;以及

获取硬件架构参数,并根据所述硬件架构参数和所述网络模型生成连续的计算流。

2.根据权利要求1所述的方法,其特征在于,所述基于所述网络模型的权值对量化后的网络模型进行精度调整包括:

对所述网络模型的权值进行线性量化以生成第一衡量数据;

生成所述网络模型的特征图并基于所述特征图对所述网络模型进行量化以生成第二衡量数据;以及

利用所述第一衡量数据和所述第二衡量数据进行量化推理,将推理结果与原始网络模型的推理结果进行对比,并根据对比结果对所述网络模型进行调整。

3.根据权利要求1所述的方法,其特征在于,所述在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程包括:

判断输入所述网络模型的图片的尺寸是否超过阈值;以及

响应于输入所述网络模型的图片的尺寸超过阈值,将所述图片进行分片,并将分片后的子图片发送到不同的AI芯片核心。

4.根据权利要求1所述的方法,其特征在于,所述在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程包括:

判断所述网络模型输入图片的通道数量是否超过第二阈值;以及

响应于所述网络模型输入图片的通道数量超过第二阈值,将所述网络模型的权值进行分片,并将分片后的子权值发送到不同的AI芯片核心。

5.根据权利要求1所述的方法,其特征在于,所述根据所述硬件架构参数和所述网络模型生成连续的计算流包括:

响应于生成新的计算流,获取当前每个时间段计算力的使用情况,并将所述新的计算流设置到当前计算力的使用最小的时间段。

6.根据权利要求1所述的方法,其特征在于,还包括:

对所述网络模型进行重训练以对所述网络模型进行裁剪和压缩。

7.根据权利要求1所述的方法,其特征在于,还包括:

对所述网络模型的计算图中的计算进行同类化以简化计算流程。

8.一种编译AI芯片的系统,其特征在于,包括:

量化模块,配置用于对AI芯片的网络模型进行量化,并基于所述网络模型的权值对量化后的网络模型进行精度调整;

计算模块,配置用于在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程;以及

流模块,配置用于获取硬件架构参数,并根据所述硬件架构参数和所述网络模型生成

连续的计算流。

9.一种计算机设备,其特征在于,包括:

至少一个处理器;以及

存储器,所述存储器存储有可在所述处理器上运行的计算机指令,所述指令由所述处理器执行时实现权利要求1-7任意一项所述方法的步骤。

10.一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1-7任意一项所述方法的步骤。

一种编译AI芯片的方法、系统、设备及介质

技术领域

[0001] 本发明涉及人工智能领域，更具体地，特别是指一种编译AI芯片的方法、系统、计算机设备及可读介质。

背景技术

[0002] 人工智能(AI)作为目前最热门、最具有潜力的技术，在强大的计算设备的加持下，近年来得到了长足的发展。相关应用渗透到了生活的方方面面，比如天气预测、自动驾驶、自然语言翻译等等。但AI技术的基础是庞大的数学计算，这些都要依托强大的计算设备。在AI技术应用的初期，因为没有AI计算专用的设备，往往使用通用的计算架构处理器如CPU和GPU来完成相关的计算和AI算例的部署，这些通用的计算设备虽然灵活性较强，通用性好，但是在进行AI相关计算的时候，往往计算效率较低，功耗较高，硬件成本也并不便宜，这就使得人们寻求一种AI计算专用的计算设备AI ASIC(AI专用芯片)。AI专用芯片的目的就是要使得AI计算效率更高，AI设备算力更强，能耗、硬件成本也更低，从而从整体上降低部署AI设备的成本，提高AI设备的处理速度。

[0003] 对于AI芯片或者AI设备要进行实例计算的流程是：1)从现有的深度学习框架如TensorFlow、PyTorch、Caffe、ONNX中获取要进行实例运算的算法或者深度网络模型文件；2)利用AI编译器生成AI设备上可直接执行的指令集或者计算配置文件；3)利用顶层应用或者运行时将2)中的指令集或者文件直接分发给驱动；4)驱动将指令集或者相关计算配置再转发给AI设备；5)AI设备完成计算，输出计算结果或者精度数据。

[0004] 但是，现有技术中对于编译AI芯片具有以下缺点：

[0005] (1)量化功能不够完善，量化精度也有待提高；

[0006] (2)计算优化仅能支持算子融合类的优化，并不能很好地对计算做更加全面深度地优化；

[0007] (3)在灵活性上不能直接支持PyTorch等一些主流的框架，为用户带来不少的限制；

[0008] (4)只针对某些特定的硬件平台，无法进行扩展；

[0009] (5)使用过程太过复杂，不利于用户部署实际的算例；

[0010] (6)计算设备在进行网络推理计算的时候存在空闲状态，设备的利用率有待提高；

[0011] (7)计算优化仅能支持非常初级的算子融合类的优化，无法最大程度的优化网络的计算，减少计算的步骤，最大化MAC计算单元的利用率。

发明内容

[0012] 有鉴于此，本发明实施例的目的在于提出一种编译AI芯片的方法、系统、计算机设备及计算机可读存储介质，通过对AI芯片进行量化并进行精度调整，使得性能更加完善；最大程度的优化网络的计算，减少计算的步骤，最大化MAC计算单元的利用率；通过生成可使计算设备在网络推理过程中一直处于忙碌状态的计算流，提高了AI设备的利用率。

[0013] 基于上述目的，本发明实施例的一方面提供了一种编译AI芯片的方法，包括如下步骤：对AI芯片的网络模型进行量化，并基于所述网络模型的权值对量化后的网络模型进行精度调整；在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程；以及获取硬件架构参数，并根据所述硬件架构参数和所述网络模型生成连续的计算流。

[0014] 在一些实施方式中，所述基于所述网络模型的权值对量化后的网络模型进行精度调整包括：对所述网络模型的权值进行线性量化以生成第一衡量数据；生成所述网络模型的特征图并基于所述特征图对所述网络模型进行量化以生成第二衡量数据；以及利用所述第一衡量数据和所述第二衡量数据进行量化推理，将推理结果与原始网络模型的推理结果进行对比，并根据对比结果对所述网络模型进行调整。

[0015] 在一些实施方式中，所述在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程包括：判断输入所述网络模型的图片的尺寸是否超过阈值；以及响应于输入所述网络模型的图片的尺寸超过阈值，将所述图片进行分片，并将分片后的子图片发送到不同的AI芯片核心。

[0016] 在一些实施方式中，所述在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程包括：判断所述网络模型输入图片的通道数量是否超过第二阈值；以及响应于所述网络模型输入图片的通道数量超过第二阈值，将所述网络模型的权值进行分片，并将分片后的子权值发送到不同的AI芯片核心。

[0017] 在一些实施方式中，所述根据所述硬件架构参数和所述网络模型生成连续的计算流包括：响应于生成新的计算流，获取当前每个时间段计算力的使用情况，并将所述新的计算流设置到当前计算力的使用最小的时间段。

[0018] 在一些实施方式中，方法还包括：对所述网络模型进行重训练以对所述网络模型进行裁剪和压缩。

[0019] 在一些实施方式中，方法还包括：对所述网络模型的计算图中的计算进行同类化以简化计算流程。

[0020] 本发明实施例的另一方面，还提供了一种编译AI芯片系统，包括：量化模块，配置用于对AI芯片的网络模型进行量化，并基于所述网络模型的权值对量化后的网络模型进行精度调整；计算模块，配置用于在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程；以及流模块，配置用于获取硬件架构参数，并根据所述硬件架构参数和所述网络模型生成连续的计算流。

[0021] 本发明实施例的又一方面，还提供了一种计算机设备，包括：至少一个处理器；以及存储器，所述存储器存储有可在所述处理器上运行的计算机指令，所述指令由所述处理器执行时实现如上方法的步骤。

[0022] 本发明实施例的再一方面，还提供了一种计算机可读存储介质，计算机可读存储介质存储有被处理器执行时实现如上方法步骤的计算机程序。

[0023] 本发明具有以下有益技术效果：

[0024] (1)解决了不同的深度学习框架PyTorch、TensorFlow、Caffe、ONNX的接口问题，并能够提供全面且可靠的量化功能；

[0025] (2)提供了全面的计算优化功能，包括算子融合、裁剪压缩、稀疏部署以及基于AutoML和强化学习的针对硬件的自适应优化功能，以及最优计算图的自动选择、实现最大化的MAC计算单元的利用率，使得计算速度最大化，延迟最小化；

[0026] (3)根据硬件架构和网络计算流程，自动生成最大化利用AI设备的计算流，使得计算效率逼近理想的100％；

[0027] (4)给用户提供了多种灵活性开发方式，比如自定义算子和自定义网络的开发，用户可选择自己擅长的开发语言实现自定义和现编译器还不支持的算子和网络；

[0028] (5)通过LLVM(Low Level Virtual Machine，底层虚拟机)后端的实现方法，可满足用户针对自己的硬件的AI编译器和硬件的适配，不用因为硬件架构的变化而进行大量的AI编译器的二次开发等，可为用户节省大量的研发时间，让用户聚焦于硬件架构的设计之上。

附图说明

[0029] 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实施例。

[0030] 图1为本发明提供的编译AI芯片的方法的实施例的示意图；

[0031] 图2为本发明提供的自定义算子和网络的流程图；

[0032] 图3为本发明提供的定制化底层编译器的流程图；

[0033] 图4为本发明提供的编译AI芯片的计算机设备的实施例的硬件结构示意图。

具体实施方式

[0034] 为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明实施例进一步详细说明。

[0035] 需要说明的是，本发明实施例中所有使用“第一”和“第二”的表述均是为了区分两个相同名称非相同的实体或者非相同的参量，可见“第一”“第二”仅为了表述的方便，不应理解为对本发明实施例的限定，后续实施例对此不再一一说明。

[0036] 基于上述目的，本发明实施例的第一个方面，提出了一种编译AI芯片的方法的实施例。图1示出的是本发明提供的编译AI芯片的方法的实施例的示意图。如图1所示，本发明实施例包括如下步骤：

[0037] S1、对AI芯片的网络模型进行量化，并基于网络模型的权值对量化后的网络模型进行精度调整；

[0038] S2、在精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化网络模型的计算流程；以及

[0039] S3、获取硬件架构参数，并根据硬件架构参数和网络模型生成连续的计算流。

[0040] AI技术的基础是深度学习，专用的AI计算芯片，其实根本上就是一款专门针对深度学习计算的芯片。而设计一款计算力强大且高效的芯片，要从算法顶层、AI编译器、体系结构以及硬件设计等多个方面去综合考虑。算法主要是深度神经网络，比如卷积类网络、循

环类神经网络等等。这些算法都是已有的，并经过相关数据集进行精度验证过的开源算法，未来用户需要结合自己的实际应用场景选择对应的算法。设计的芯片就是要更加灵活、全面、高效地去支持这些算法。

[0041] 对AI芯片的网络模型进行量化，并基于网络模型的权值对量化后的网络模型进行精度调整。本发明实施例具有对网络模型进行量化的功能，可将原始基于TensorFlow、PyTorch、ONNX、Caffe等的FP32的高比特网络模型量化到INT16/Fp16、INT8甚至是INT4的低比特类型上。且网络的量化精度满足关于精度的指标。

[0042] 目前深度学习的算法较多，实现这些算法的深度学习框架也较多，AI编译器要能够编译出可以在硬件上直接运行的指令集或者配置文件，首要的就是能够解析这些原始的基于不同框架实现的算法，这就需要较为全面的模型接口与解析的功能，本发明实施例提供的接口能够全面地支持不同的深度学习框架，且能够全面的解析已有的深度学习算法。

[0043] 在一些实施方式中，所述基于所述网络模型的权值对量化后的网络模型进行精度调整包括：对所述网络模型的权值进行线性量化以生成第一衡量数据；生成所述网络模型的特征图并基于所述特征图对所述网络模型进行量化以生成第二衡量数据；以及利用所述第一衡量数据和所述第二衡量数据进行量化推理，将推理结果与原始网络模型的推理结果进行对比，并根据对比结果对所述网络模型进行调整。首先，对网络模型的权值进行线性量化，并生成权值的Scale数据(第一衡量数据)；然后，部署网络模型在相应数据集上的推理，生成FP32的推理数据结果，也即是特征图文件；接着，对网络模型进行特征图或者激活值量化，生成量化系数或者层间的Scale数据(第二衡量数据)；最后，利用量化的权值以及权值和特征图的Scale数据，进行量化推理，将最终的推理结果与原始FP32推理结果进行对比，并根据某层误差的大小或者数据集上的精度进行微调，最终实现精度不损失或者精度损失在1％的量化模型。

[0044] 在完成上述网络模型量化之后，进一步对整个网络模型的计算流程进行优化、合并，最大程度地减少数据的搬移以及计算步骤，最大化MAC计算单元的利用率。

[0045] 在一些实施方式中，方法还包括：对所述网络模型进行重训练以对所述网络模型进行裁剪和压缩。例如，编译器对网络模型进行评估，并对网络模型进行必要的裁剪，具体来说，网络模型的裁剪就是通过对网络模型中不重要的层或者通道，进行重训练的方式在将其去掉，直到裁剪之后精度不能恢复为止，达到最大的裁剪比例且不损失精度的结果，实现网络模型的小型化，实现计算量大比例的减少。在经过网络的裁剪和压缩之后，每层的计算结构就变得稀疏，这样稀疏的计算要与硬件架构进行匹配，完成稀疏计算的部署。

[0046] 在一些实施方式中，方法还包括：对所述网络模型的计算图中的计算进行同类化以简化计算流程。例如，对于一个算子中或者一个计算步骤中，涉及到的不同种类的计算进行同类化，消除公共子，简化算子的计算步骤，进一步实现计算的简化。通过将计算图中可以合并的算子或者操作融合，实现计算步骤的缩减，减少数据在每次计算前后的搬移，节省大量计算时间。

[0047] 可以利用最新的强化学习以及自适应学习AutoML，针对用户当前的AI设备进行自适应学习的计算优化，使得计算优化达到计算与硬件的最佳匹配。

[0048] 在精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化网络模型的计算流程。一个卷积网络往往由很多个卷积层构成，每个卷积层的

计算的基本过程都是:特征图(Feature)数据输入、权值(Weight)加载、乘加计算、结果输出或者暂存。当有多张图片输入,每张图片可分别分配给各核心,因为权值相同,所以各核心的权值是共享复用的。

[0049] 在一些实施方式中,所述在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程包括:判断输入所述网络模型的图片的尺寸是否超过阈值;以及响应于输入所述网络模型的图片的尺寸超过阈值,将所述图片进行分片,并将分片后的子图片发送到不同的AI芯片核心。有的实用场景下,图片的尺寸可能非常大,当超过预先设置的阈值时,可通过对图片进行分片的方式送入到不同的AI芯片核心上,同样此时的权值是共享复用的。

[0050] 在一些实施方式中,所述在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程包括:判断所述网络模型输入图片的通道数量是否超过第二阈值;以及响应于所述网络模型输入图片的通道数量超过第二阈值,将所述网络模型的权值进行分片,并将分片后的子权值发送到不同的AI芯片核心。网络计算过程中,图片的尺寸可能越来越小,但是图片的通道数量可能越来越多,当通道数量超过预先设置的第二阈值时,可通过将权值进行分片的方式分别送入不同地计算核心上,此时输入(特征图)是共享复用的。

[0051] 获取硬件架构参数,并根据硬件架构参数和网络模型生成连续的计算流。本发明的AI编译器可根据硬件配置以及网络模型的计算流程进行自适配,保证硬件在部署计算任务之后,不存在因为数据等待、搬移以及硬件子任务配置等所导致的硬件空闲的状态,可使得硬件的利用率接近理想的100%。具体的:首先,导入硬件体系结构文件,如果没有该参数或者不导入文件,编译器默认为自带的硬件架构;导入网络模型文件;然后根据硬件架构、网络模型和相关计算配置文件数据,生成高效、无缝衔接的计算流。

[0052] 在一些实施方式中,所述根据所述硬件架构参数和所述网络模型生成连续的计算流包括:响应于生成新的计算流,获取当前每个时间段计算力的使用情况,并将所述新的计算流设置到当前计算力的使用最小的时间段。

[0053] 对于新型的网络或者用户自定义的网络,本发明的AI编译器提供自定义网络以及自定义算子的实现接口。用户可以选择使用自己擅长的语言进行自定义编程,比如使用Python,C++实现自己算子和网络的定义。

[0054] 图2示出的是本发明提供的自定义算子和网络的流程图。如图2所示,分析算子的具体功能和数学表达,明确输入和输出,确定该算子的开发方式和使用的计算接口,确定算子实现的文件名、算子名、算子类型;分别进行算子信息和原型定义,并通过代码实现对应的算子功能。算子编译包括:对实现的算子进行Makefile编写,并make编译。算子测试包括:通过特定输入数据,测试算子的输出与实际是否相符,如果符合则进行下一步,不符合则检查算子定义和实现代码,重新开发直至最后测试通过。对AI编译器进行重编译,AI编译器会自动将新开发的算子添加到AI编译器的算子库中。根据重新生成的AI编译器进一步对自定义的网络模型进行模型转换,生成硬件可直接执行的指令集或者配置文件等,验证网络推理结果。同样根据最后的推理结果进行下一步任务还是检查算子开发过程中的问题。

[0055] 编程语言最终都会编译成相应的机器指令用于硬件的执行。然而这些语言的编译器,本身生成的机器代码的目标平台主要是CPU平台,要使得这些代码编译之后不仅能够运

行在通用的CPU等架构平台上,还能够在设计的AI芯片上高效地执行,这就需要定制针对特定硬件平台的编译器。为此本发明实施例提拱了基于LLVM的针对用户自己硬件架构的定制编译器开发功能。LLVM编译器架构主要分为前、中、后三部分。前端开发主要是增加对一门语言的支持,透过LLVM可以解析语言的语法,并生成中间代码LLVM IR(intermediate representation,中间代码),中间部分为LLVM优化器,是语言和目标平台独立的部分,前端和中间部分不需要用户去关心,可使用已有的实现。后端则主要为生成目标平台的可执行代码,本发明AI编译器便实现了支持特定的AI芯片ASIC的定制化LLVM后端。

[0056] 图3示出的是本发明提供的定制化底层编译器的流程图。如图3所示,定制化底层编译器包括创建目标机器、目标机器注册、创建寄存器集、指令集选择、指令调度和JIT支持。创建目标机器,即通过创建TargetMachine子类,用于描述自己的目标机器的特性。目标注册是通过TargetRegistry接口来注册目标机器。创建目标机器硬件的寄存器集是通过TableGen来生成有关寄存器定义、寄存器别名和寄存器类的代码,通过实现继承TargetResigerInfo类的子类来表示有助于寄存器分配和寄存器间交互的信息。指令集选择是将设备无关的IR指令转换成设备相关DAG(Directed Acycle Graph)节点。指令调度是通过表调度(一种贪婪启发式算法)方式实现指令的的高效调度和执行。JIT(Just-In-Time,及时编译)支持是通过编写继承自TargetJITInfo类的子类,实现运行时调用函数或者程序段的编译执行。

[0057] 本发明实施例解决了不同的深度学习框架PyTorch,TensorFlow,Caffe,ONNX的接口问题,并能够提供全面且可靠的量化功能。本发明提供了全面的计算优化功能,包括算子融合,裁剪压缩,稀疏部署以及基于AutoML和强化学习的针对硬件的自适应优化功能,以及最优计算图的自动选择,最大化MAC计算单元的利用率,使得计算速度最大化,延迟最小化。本发明可根据硬件架构和网络计算流程,自动生成最大化利用AI设备的计算流,使得计算效率逼近理想的100%。本发明给用户提供了多种灵活性开发方式,比如自定义算子和自定义网络的开发,用户可选择自己擅长的开发语言实现自定义和先编译器还不支持的算子和网络。本发明的可扩展性,通过LLVM后端的实现方法,可满足用户针对自己的硬件的AI编译器和硬件的适配,不用因为硬件架构的变化而进行大量的AI编译器的二次开发等,可为用户节省大量的研发时间,让用户聚焦于硬件架构的设计之上。

[0058] 需要特别指出的是,上述编译AI芯片的方法的各个实施例中的各个步骤均可以相互交叉、替换、增加、删减,因此,这些合理的排列组合变换之于编译AI芯片的方法也应当属于本发明的保护范围,并且不应将本发明的保护范围局限在实施例之上。

[0059] 基于上述目的,本发明实施例的第二个方面,提出了一种编译AI芯片的系统,包括:量化模块,配置用于对AI芯片的网络模型进行量化,并基于所述网络模型的权值对量化后的网络模型进行精度调整;计算模块,配置用于在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程;以及流模块,配置用于获取硬件架构参数,并根据所述硬件架构参数和所述网络模型生成连续的计算流。

[0060] 在一些实施方式中,所述量化模块配置用于:对所述网络模型的权值进行线性量化以生成第一衡量数据;生成所述网络模型的特征图并基于所述特征图对所述网络模型进行量化以生成第二衡量数据;以及利用所述第一衡量数据和所述第二衡量数据进行量化推

理，将推理结果与原始网络模型的推理结果进行对比，并根据对比结果对所述网络模型进行调整。

[0061] 在一些实施方式中，所述计算模块配置用于：判断输入所述网络模型的图片的尺寸是否超过阈值；以及响应于输入所述网络模型的图片的尺寸超过阈值，将所述图片进行分片，并将分片后的子图片发送到不同的AI芯片核心。

[0062] 在一些实施方式中，所述计算模块配置用于：判断所述网络模型输入图片的通道数量是否超过第二阈值；以及响应于所述网络模型输入图片的通道数量超过第二阈值，将所述网络模型的权值进行分片，并将分片后的子权值发送到不同的AI芯片核心。

[0063] 在一些实施方式中，所述流模块配置用于：响应于生成新的计算流，获取当前每个时间段计算力的使用情况，并将所述新的计算流设置到当前计算力的使用最小的时间段。

[0064] 在一些实施方式中，系统还包括：重训练模块，配置用于对所述网络模型进行重训练以对所述网络模型进行裁剪和压缩。

[0065] 在一些实施方式中，系统还包括：简化模块，配置用于对所述网络模型的计算图中的计算进行同类化以简化计算流程。

[0066] 基于上述目的，本发明实施例的第三个方面，提出了一种计算机设备，包括：至少一个处理器；以及存储器，存储器存储有可在处理器上运行的计算机指令，指令由处理器执行以实现如下步骤：S1、对AI芯片的网络模型进行量化，并基于网络模型的权值对量化后的网络模型进行精度调整；S2、在精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化网络模型的计算流程；以及S3、获取硬件架构参数，并根据硬件架构参数和网络模型生成连续的计算流。

[0067] 在一些实施方式中，所述基于所述网络模型的权值对量化后的网络模型进行精度调整包括：对所述网络模型的权值进行线性量化以生成第一衡量数据；生成所述网络模型的特征图并基于所述特征图对所述网络模型进行量化以生成第二衡量数据；以及利用所述第一衡量数据和所述第二衡量数据进行量化推理，将推理结果与原始网络模型的推理结果进行对比，并根据对比结果对所述网络模型进行调整。

[0068] 在一些实施方式中，所述在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程包括：判断输入所述网络模型的图片的尺寸是否超过阈值；以及响应于输入所述网络模型的图片的尺寸超过阈值，将所述图片进行分片，并将分片后的子图片发送到不同的AI芯片核心。

[0069] 在一些实施方式中，所述在所述精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化所述网络模型的计算流程包括：判断所述网络模型输入图片的通道数量是否超过第二阈值；以及响应于所述网络模型输入图片的通道数量超过第二阈值，将所述网络模型的权值进行分片，并将分片后的子权值发送到不同的AI芯片核心。

[0070] 在一些实施方式中，所述根据所述硬件架构参数和所述网络模型生成连续的计算流包括：响应于生成新的计算流，获取当前每个时间段计算力的使用情况，并将所述新的计算流设置到当前计算力的使用最小的时间段。

[0071] 在一些实施方式中，步骤还包括：对所述网络模型进行重训练以对所述网络模型进行裁剪和压缩。

[0072] 在一些实施方式中，步骤还包括：对所述网络模型的计算图中的计算进行同类化以简化计算流程。

[0073] 如图4所示，为本发明提供的上述编译AI芯片的计算机设备的一个实施例的硬件结构示意图。

[0074] 以如图4所示的装置为例，在该装置中包括一个处理器301以及一个存储器302，并还可以包括：输入装置303和输出装置304。

[0075] 处理器301、存储器302、输入装置303和输出装置304可以通过总线或者其他方式连接，图4中以通过总线连接为例。

[0076] 存储器302作为一种非易失性计算机可读存储介质，可用于存储非易失性软件程序、非易失性计算机可执行程序以及模块，如本申请实施例中的编译AI芯片的方法对应的程序指令/模块。处理器301通过运行存储在存储器302中的非易失性软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例的编译AI芯片的方法。

[0077] 存储器302可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据编译AI芯片的方法的使用所创建的数据等。此外，存储器302可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实施例中，存储器302可选包括相对于处理器301远程设置的存储器，这些远程存储器可以通过网络连接至本地模块。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0078] 输入装置303可接收输入的用户名和密码等信息。输出装置304可包括显示屏等显示设备。

[0079] 一个或者多个编译AI芯片的方法对应的程序指令/模块存储在存储器302中，当被处理器301执行时，执行上述任意方法实施例中的编译AI芯片的方法。

[0080] 执行上述编译AI芯片的方法的计算机设备的任何一个实施例，可以达到与之对应的前述任意方法实施例相同或者相类似的效果。

[0081] 本发明还提供了一种计算机可读存储介质，计算机可读存储介质存储有被处理器执行时执行如上方法的计算机程序。

[0082] 最后需要说明的是，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，可以通过计算机程序来指令相关硬件来完成，编译AI芯片的方法的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，程序的存储介质可为磁碟、光盘、只读存储记忆体(ROM)或随机存储记忆体(RAM)等。上述计算机程序的实施例，可以达到与之对应的前述任意方法实施例相同或者相类似的效果。

[0083] 以上是本发明公开的示例性实施例，但是应当注意，在不背离权利要求限定的本发明实施例公开的范围的前提下，可以进行多种改变和修改。根据这里描述的公开实施例的方法权利要求的功能、步骤和/或动作不需以任何特定顺序执行。此外，尽管本发明实施例公开的元素可以以个体形式描述或要求，但除非明确限制为单数，也可以理解为多个。

[0084] 应当理解的是，在本文中使用的，除非上下文清楚地支持例外情况，单数形式“一

个”旨在也包括复数形式。还应当理解的是，在本文中使用的“和/或”是指包括一个或者一个以上相关联地列出的项目的任意和所有可能组合。

[0085] 上述本发明实施例公开实施例序号仅仅为了描述，不代表实施例的优劣。

[0086] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0087] 所属领域的普通技术人员应当理解：以上任何实施例的讨论仅为示例性的，并非旨在暗示本发明实施例公开的范围(包括权利要求)被限于这些例子；在本发明实施例的思路下，以上实施例或者不同实施例中的技术特征之间也可以进行组合，并存在如上的本发明实施例的不同方面的许多其它变化，为了简明它们没有在细节中提供。因此，凡在本发明实施例的精神和原则之内，所做的任何省略、修改、等同替换、改进等，均应包含在本发明实施例的保护范围之内。

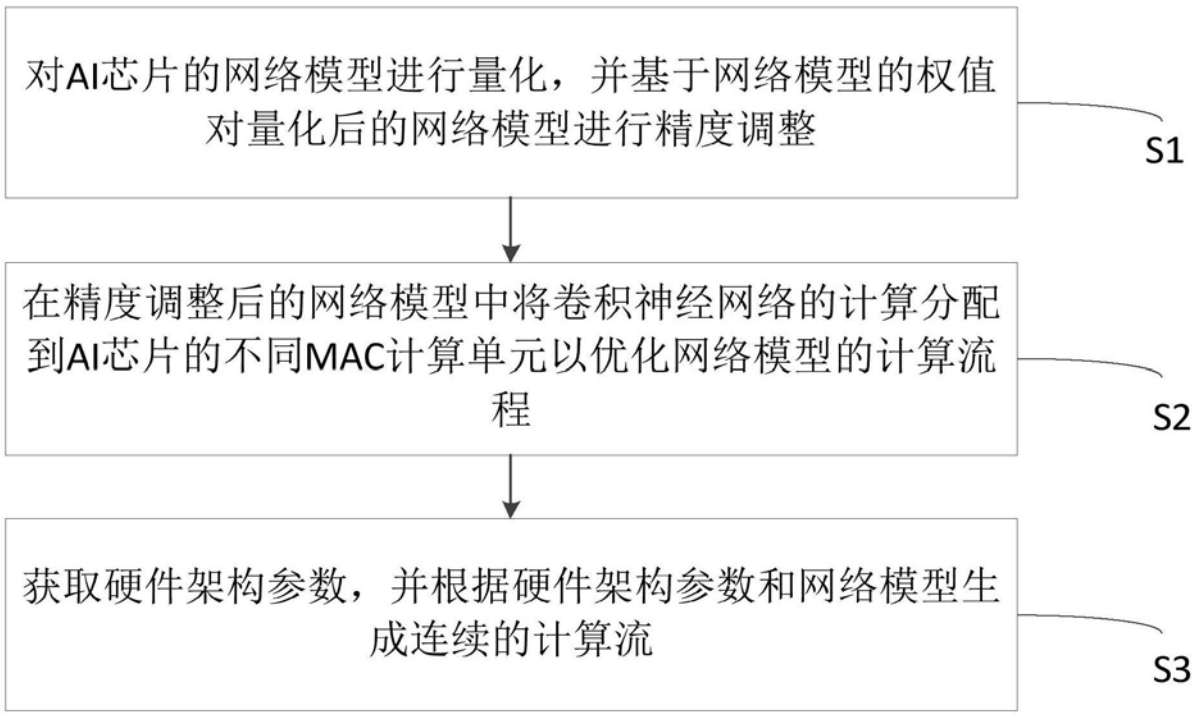
对AI芯片的网络模型进行量化，并基于网络模型的权值对量化后的网络模型进行精度调整
S1
在精度调整后的网络模型中将卷积神经网络的计算分配到AI芯片的不同MAC计算单元以优化网络模型的计算流程
S2
获取硬件架构参数，并根据硬件架构参数和网络模型生成连续的计算流
S3

图1

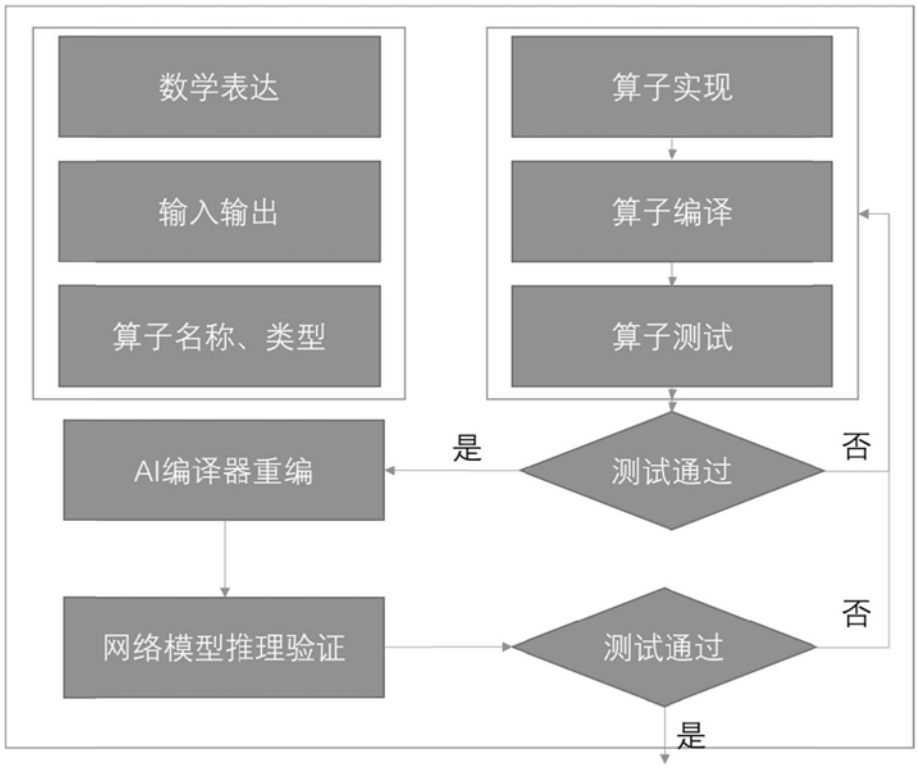
数学表达
输入输出
算子名称、类型
算子实现
算子编译
算子测试
测试通过
是
否
AI编译器重编
网络模型推理验证
测试通过
否
是

图2

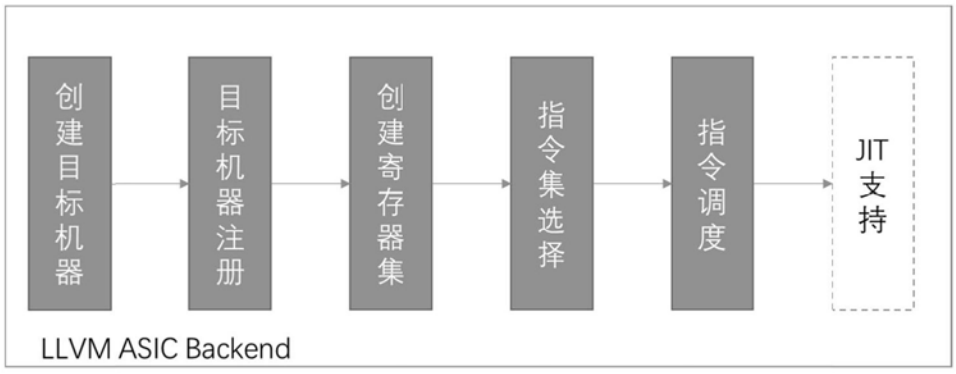
创建目标机器
目标机器注册
创建寄存器集
指令集选择
指令调度
JIT支持
LLVM ASIC Backend

图3

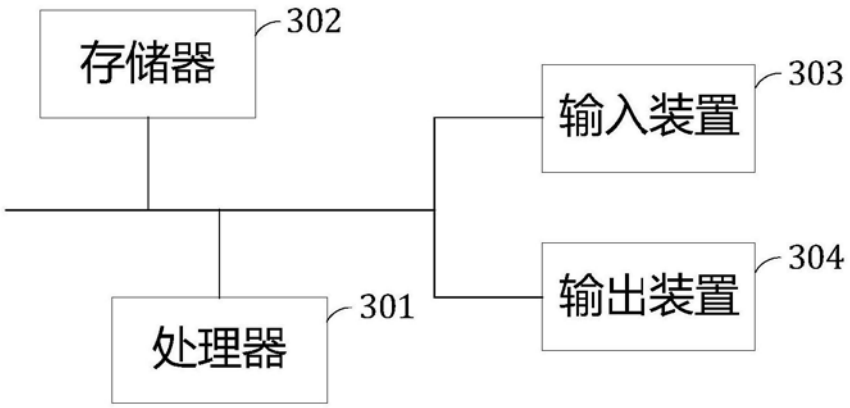
存储器
302
输入装置
303
输出装置
304
处理器
301

图4