



(12) 发明专利

(10) 授权公告号 CN 109978142 B

(45) 授权公告日 2022. 11. 29

(21) 申请号 201910248461.3

(22) 申请日 2019.03.29

(65) 同一申请的已公布的文献号

申请公布号 CN 109978142 A

(43) 申请公布日 2019.07.05

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518000 广东省深圳市南山区高新区

科技中一路腾讯大厦35层

(72) 发明人 国雍 郑胤 赵沛霖 黄俊洲

(74) 专利代理机构 北京派特恩知识产权代理有

限公司 11270

专利代理师 王姗姗 张颖玲

(51) Int.Cl.

G06N 3/04 (2006.01)

(56) 对比文件

CN 109284820 A, 2019.01.29

CN 107909147 A, 2018.04.13

CN 108921291 A, 2018.11.30

CN 108229644 A, 2018.06.29

CN 108932548 A, 2018.12.04

CN 107316079 A, 2017.11.03

CN 108416187 A, 2018.08.17

CN 107944556 A, 2018.04.20

CN 108596988 A, 2018.09.28

US 2018046900 A1, 2018.02.15

WO 02073529 A1, 2002.09.19

WO 2008129021 A2, 2008.10.30

US 2017364799 A1, 2017.12.21

US 2018373975 A1, 2018.12.27

US 2018204110 A1, 2018.07.19

杨文柱等. “一种用于卷积神经网络压缩的混合剪枝方法”. 《小型微型计算机系统》. 2018, 第39卷(第12期),

黄聪等. “基于权值相似性的神经网络剪枝”. 《计算机科学与探索》. 2018, 第12卷(第8期),

Wei Wen等. “Learning Structured Sparsity in Deep Neural Networks”. 《arXiv》. 2016,

He Y H等. “Channel pruning for accelerating very deep neural networks”. 《Proceedings of the 2017 IEEE International Conference on Computer Vision》. 2017,

(续)

审查员 刘倩倩

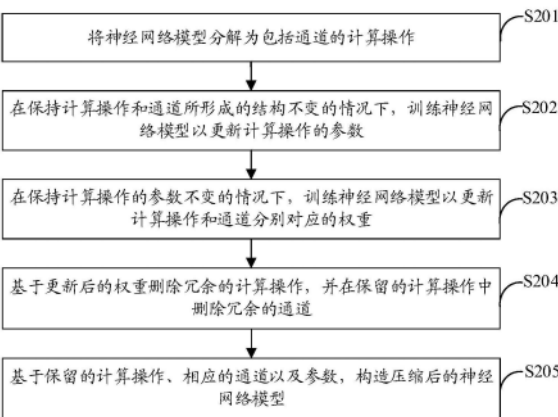
权利要求书2页 说明书15页 附图6页

(54) 发明名称

神经网络模型的压缩方法和装置

(57) 摘要

本发明实施例公开了一种神经网络模型压缩的方法和装置。方法包括：将神经网络模型分解为包括通道的计算操作；在保持计算操作和通道所形成的结构不变的情况下，训练神经网络模型以更新计算操作的参数；在保持计算操作的参数不变的情况下，训练神经网络模型以更新计算操作和通道分别对应的权重；基于更新后的权重删除冗余的计算操作，并在保留的计算操作中删除冗余的通道；基于保留的计算操作、相应的通道以及参数，构造压缩后的神经网络模型。通过本发明实施例能够有效简化神经网络模型。



CN 109978142 B

[接上页]

(56) 对比文件

张弛等.“神经网络模型压缩方法综述”.《计算机科学》.2018,第45卷(第10期),

林绍辉等.“深度神经网络压缩与加速综述”.《计算机研究与发展》.2018,第55卷(第9期),

Anwar S等.“Coarse pruning of convolutional neural networks with random

masks”.《Proceedings of the 2017 International Conference on Learning Representations》.2017,

Li D等.“Deeprebirth:Accelerating deep neural network execution on mobile devices”.《32nd AAAI Conf. on Artificial Intelligence》.2018,

1. 一种神经网络模型的压缩方法,其特征在于,所述方法包括:

将神经网络模型分解为包括通道的计算操作;

在保持所述计算操作和所述通道所形成的结构不变的情况下,训练所述神经网络模型以更新所述计算操作的参数;

在保持所述计算操作的参数不变的情况下,根据所述神经网络模型的目标设备的运行参数,确定对应的通道稀疏程度,其中,所述运行参数是体现神经网络模型运行时需求的硬件条件;

基于验证数据集训练所述神经网络模型以更新所述计算操作和所述通道分别对应的权重,直至所述计算操作和所述通道分别对应的权重满足所述通道稀疏程度;

基于更新后的权重删除冗余的计算操作,并在保留的计算操作中删除冗余的通道;

基于保留的计算操作、相应的通道以及参数,构造压缩后的神经网络模型。

2. 根据权利要求1所述的方法,其特征在于,所述将神经网络模型分解为通过通道而连接的计算操作,包括:

根据所述神经网络模型包括的网络层,将所述神经网络模型分解为对应所述网络层的计算操作;

将所述网络层中的不同信息流,分解为与所述信息流对应的通道。

3. 根据权利要求1所述的方法,其特征在于,所述训练所述神经网络模型以更新所述计算操作的参数,包括:

在保持所述计算操作和所述通道分别对应的权重不变的情况下,基于训练数据集训练所述神经网络模型以更新所述计算操作的参数。

4. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

交替更新所述计算操作的参数、以及所述神经网络模型的权重,直至所述神经网络模型满足收敛条件;其中,所述神经网络模型的权重包括:所述计算操作的权重和所述通道的权重。

5. 根据权利要求1所述的方法,其特征在于,所述方法还包括:

训练所述神经网络模型之前,在所述神经网络模型中添加候选计算操作,以使所述神经网络模型具有梯度下降的特性。

6. 根据权利要求1至5任一项所述的方法,其特征在于,所述方法还包括:

训练所述神经网络模型之前,初始化所述神经网络模型中的原始计算操作和候选计算操作。

7. 根据权利要求6所述的方法,其特征在于,

所述初始化所述神经网络模型中的原始计算操作和候选计算操作,包括:

当所述原始计算操作不存在参数或通道权重时,对所述原始计算操作对应的候选计算操作的参数进行随机初始化;

初始化所述原始计算操作对应的权重,且大于所述候选计算操作的初始化的权重。

8. 根据权利要求6所述的方法,其特征在于,

所述初始化所述神经网络模型中的原始计算操作和候选计算操作,包括:

当所述原始计算操作存在参数时,通过最小化所述原始计算操作与所对应的候选计算操作的重构误差的方式,初始化原始计算操作对应的所述候选计算操作的参数。

9. 一种神经网络模型的压缩方法,其特征在于,所述方法包括:

在模型压缩系统的客户端中接收神经网络模型、以及所述神经模型的目标设备的运行参数;

调用所述模型压缩系统,以根据所述目标设备的运行参数确定通道稀疏程度,以及

以所述通道稀疏程度为目标,对所述神经网络模型进行压缩,得到压缩后的神经网络模型,其中,所述对所述神经网络模型进行压缩是通过权利要求1所述的一种神经网络模型的压缩方法实现的。

10. 一种神经网络模型的压缩装置,其特征在于,包括:

分解单元,用于将神经网络模型分解为包括通道的计算操作;

第一训练单元,用于在保持所述计算操作和所述通道所形成的结构不变的情况下,训练所述神经网络模型以更新所述计算操作的参数;

第二训练单元,用于在保持所述计算操作的参数不变的情况下,根据所述神经网络模型的目标设备的运行参数,确定对应的通道稀疏程度,其中,所述运行参数是体现神经网络模型运行时需求的硬件条件;基于验证数据集训练所述神经网络模型以更新所述计算操作和所述通道分别对应的权重,直至所述计算操作和所述通道分别对应的权重满足所述通道稀疏程度;

删除单元,用于基于更新后的权重删除冗余的计算操作,并在保留的计算操作中删除冗余的通道;

构造单元,用于基于保留的计算操作、相应的通道以及参数,构造压缩后的神经网络模型。

11. 根据权利要求10所述的装置,其特征在于,

所述第一训练单元具体用于:

在保持所述计算操作和所述通道分别对应的权重不变的情况下,基于训练数据集训练所述神经网络模型以更新所述计算操作的参数;或

所述第二训练模块具体用于:

在保持所述计算操作的参数不变的情况下,基于验证数据集训练所述神经网络模型以更新所述计算操作和所述通道分别对应的权重。

12. 根据权利要求10所述的装置,其特征在于,所述装置还包括:

添加模块,用于训练所述神经网络模型之前,在所述神经网络模型中添加候选计算操作,以使所述神经网络模型具有梯度下降的特性。

13. 根据权利要求10所述的装置,其特征在于,所述装置还包括:

初始化模块,用于训练所述神经网络模型之前,初始化所述神经网络模型中的原始计算操作和候选计算操作。

14. 一种电子设备,其特征在于,所述电子设备包括:

存储器,用于存储计算机可执行指令;

处理器,用于执行所述存储器中存储的计算机可执行指令时,实现权利要求1至9任一项所述的神经网络模型的压缩方法。

15. 一种计算机可读存储介质,存储有计算机可执行指令,其特征在于,所述计算机可执行指令被处理器执行时实现权利要求1至9任一项所述的神经网络模型的压缩方法。

神经网络模型的压缩方法和装置

技术领域

[0001] 本发明涉及计算机技术,尤其涉及一种神经网络模型的压缩方法和装置。

背景技术

[0002] 以神经网络模型为基础的深度学习技术得到快速发展,在包括图片识别、目标检测、语义分割、语音识别和自然语言处理等多个应用领域都取得了突破性成果。然而,深度神经网络模型往往存在大量的冗余参数,导致模型冗余。模型冗余不仅带来高昂的计算代价,还大大增加了模型过拟合的风险。这些问题严重限制了神经网络模型在现场可编程门阵列(Field-Programmable GateArray,FPGA)、精简指令集计算机(Reduced Instruction Set Computer,RISC)微处理器(Advanced RISC Machine,ARM)等资源受限的计算平台中的应用。

发明内容

[0003] 本发明实施例提供一种神经网络模型的压缩方法、装置和存储介质,能够有效简化神经网络模型。

[0004] 本发明实施例的技术方案是这样实现的:

[0005] 本发明实施例提供一种神经网络模型的压缩方法,包括:

[0006] 将神经网络模型分解为包括通道的计算操作;

[0007] 在保持所述计算操作和所述通道所形成的结构不变的情况下,训练所述神经网络模型以更新所述计算操作的参数;

[0008] 在保持所述计算操作的参数不变的情况下,训练所述神经网络模型以更新所述计算操作和所述通道分别对应的权重;

[0009] 基于更新后的权重删除冗余的计算操作,并在保留的计算操作中删除冗余的通道;

[0010] 基于保留的计算操作、相应的通道以及参数,构造压缩后的神经网络模型。

[0011] 本发明实施例提供一种网络模型压缩方法,包括:

[0012] 在模型压缩系统的客户端中接收神经网络模型、以及所述神经模型的目标设备的运行参数;

[0013] 调用所述模型压缩系统,以根据所述目标设备的运行参数确定通道稀疏程度,以及

[0014] 以所述通道稀疏程度为目标,对所述神经网络模型中的计算操作、以及连接所述计算操作的通道分别进行压缩,得到压缩后的神经网络模型。

[0015] 本发明实施例提供一种网络模型压缩装置,包括:

[0016] 分解单元,用于将神经网络模型分解为包括通道的计算操作;

[0017] 第一训练单元,用于在保持所述计算操作和所述通道所形成的结构不变的情况下,训练所述神经网络模型以更新所述计算操作的参数;

[0018] 第二训练单元,用于在保持所述计算操作的参数不变的情况下,训练所述神经网络模型以更新所述计算操作和所述通道分别对应的权重;

[0019] 删除单元,用于基于更新后的权重删除冗余的计算操作,并在保留的计算操作中删除冗余的通道;

[0020] 构造单元,用于基于保留的计算操作、相应的通道以及参数,构造压缩后的神经网络模型。

[0021] 在一些实施例中,所述分解单元具体用于:

[0022] 根据所述神经网络模型包括的网络层,将所述神经网络模型分解为对应所述网络层的计算操作;

[0023] 将所述网络层中的不同信息流,分解为与所述信息流对应的通道。

[0024] 在一些实施例中,所述第二训练模块具体用于:

[0025] 根据用于神经网络模型的目标设备的运行参数,确定对应的通道稀疏程度;

[0026] 基于验证数据集训练所述神经网络模型以更新所述计算操作和所述通道分别对应的权重,直至所述计算操作和所述通道分别对应的权重满足所述通道稀疏程度。

[0027] 在一些实施例中,所述装置还包括:

[0028] 交替模块,用于交替更新所述计算操作的参数、以及所述神经网络模型的权重,直至所述神经网络模型满足收敛条件;

[0029] 其中,所述神经网络模型的权重包括:所述计算操作的权重和所述通道的权重。

[0030] 在一些实施例中,所述装置还包括:

[0031] 初始化模块,用于训练所述神经网络模型之前,初始化所述神经网络模型中的原始计算操作和候选计算操作。

[0032] 在一些实施例中,所述初始化模块,具体用于:

[0033] 当所述原始计算操作不存在参数或通道权重时,对所述原始计算操作对应的候选计算操作的参数进行随机初始化;

[0034] 初始化所述原始计算操作对应的权重,且大于所述候选计算操作的初始化的权重。

[0035] 在一些实施例中,所述初始化模块,具体用于:

[0036] 当所述原始计算操作存在参数时,根据所述原始计算操作的参数,通过最小化所述原始计算操作与所对应的候选计算操作的重构误差的方式,初始化原始计算操作对应的所述候选计算操作的参数。

[0037] 本发明实施例提供一种网络模型压缩装置,包括:

[0038] 确定单元,用于在模型压缩系统的客户端中接收神经网络模型、以及所述神经模型的目标设备的运行参数;

[0039] 压缩单元,用于调用所述模型压缩系统,以根据所述目标设备的运行参数确定通道稀疏程度,以及

[0040] 以所述通道稀疏程度为目标,对所述神经网络模型中的计算操作、以及连接所述计算操作的通道分别进行压缩,得到压缩后的神经网络模型。

[0041] 本发明实施例提供一种神经网络模型的压缩装置,包括:

[0042] 存储器,用于存储可执行指令;

[0043] 处理器,用于执行所述存储器中存储的可执行指令时,实现上述神经网络模型的压缩方法。

[0044] 本发明实施例提供一种存储介质,所述存储介质中存储有计算机程序,所述计算机程序被处理器执行时,实现上述神经网络模型的压缩方法。

[0045] 本发明实施例具有以下有益效果:

[0046] 通过从神经网络模型的计算操作和通道两个维度进行训练,基于计算操作和通道分别对应的权重,对计算操作进行删除以消除神经网络模型的结构冗余,对通道对应进行删除能够有效消除神经网络模型的参数冗余,从结构和参数两个维度对神经模型进行有效压缩。

附图说明

[0047] 图1A是本发明实施例模型压缩系统的示例性的架构示意图

[0048] 图1B是本发明实施例压缩前的神经网络模型的结构示意图;

[0049] 图1C是本发明实施例压缩后的神经网络模型的结构示意图;

[0050] 图2是本发明实施例神经网络模型的压缩方法的流程示意图一;

[0051] 图3是本发明实施例网络层和信息流的结构示意图;

[0052] 图4是本发明实施例神经网络模型的压缩方法的流程示意图二;

[0053] 图5是本发明实施例通道减枝示意图;

[0054] 图6是本发明实施例DARTS示意图;

[0055] 图7是本发明实施例搜索子空间示意图;

[0056] 图8是本发明实施例神经网络模型的压缩方法应用场景示意图;

[0057] 图9是本发明实施例神经网络模型的压缩装置的结构示意图一;

[0058] 图10是本发明实施例神经网络模型的压缩装置的结构示意图二。

具体实施方式

[0059] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合附图对本发明作进一步地详细描述,所描述的实施例不应视为对本发明的限制,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例,都属于本发明保护的范围。

[0060] 在以下的描述中,涉及到“一些实施例”,其描述了所有可能实施例的子集,但是可以理解,“一些实施例”可以是所有可能实施例的相同子集或不同子集,并且可以在不冲突的情况下相互结合。

[0061] 除非另有定义,本发明实施例所使用的所有的技术和科学术语与属于本发明实施例的技术领域的技术人员通常理解的含义相同。本发明实施例所使用的术语只是为了描述本发明实施例的目的,不是旨在限制本发明。

[0062] 对本发明实施例进行进一步详细说明之前,对本发明实施例中涉及的名词和术语进行说明,本发明实施例中涉及的名词和术语适用于如下的解释。

[0063] 1) 压缩,剔除神经网络模型中冗余的参数和通道的过程。例如,神经网络模型的组成单元为网络层。网络层基于功能不同可分为卷积层、激活层、池化层、映射层和批量标准化层等,对神经网络模型的压缩,包括删除网络层的参数进行压缩时,可从以下方面来进行

处理:卷积核的大小、卷积核的数量,其中,卷积核的大小包括:通道数、高、宽。

[0064] 2) 搜索子空间,是网络模型压缩的一种方式,将待压缩的神经网络模型作为搜索空间,可划分为多个连续的搜索子空间(也称为子网络或单元)而得到。搜索子空间是由N个有序节点x组成的有向无环图,比如包括:2个输入节点、中间节点和一个输出节点。其中输入节点为前一个搜索子空间的输出,中间节点为前继通过计算操作再求和得到,输出节点为每一个中间节点的合并。每个有向边表示节点 x_i 和 x_j 之间的某种计算操作,记为 $o^{(i,j)}$, $i < j < N$ 。

[0065] 3) 计算操作,与神经网络模型的网络层对应,计算操作的类型可包括:卷积层、激活层、池化层、映射层和批量标准化层等,比如:采用宽高为 3×3 的卷积核的卷积层、均值池化层和最大池化层等。

[0066] 4) 通道,神经网络模型的网络层中信息流传递的路径。

[0067] 5) 操作权重,即计算操作对应的权重,表征两个节点之间的该计算操作的重要性,一个计算操作越重要,则该计算操作对应的权重越大。

[0068] 6) 通道权重,即通道对应的权重,表示该通道的重要性。

[0069] 本发明实施例可提供为神经网络模型的压缩方法、装置和存储介质。实际应用中,神经网络模型的可由神经网络模型的装置实现神经网络模型的压缩装置中的各功能实体可以由设备(如终端设备、服务器或服务器集群)的硬件资源,如处理器等计算资源、通信资源(如用于支持实现光缆、蜂窝等各种方式通信)协同实现。

[0070] 首先说明实现本发明实施例的神经网络模型的模型压缩系统的示例性应用。

[0071] 参见图1A,图1A是本发明实施例提供的模型压缩系统的示例性的架构示意图,包括模型压缩系统13和终端12。终端12中运行的客户端11可以为对应模型压缩系统13的面向用户的浏览器页面或者专用客户端。用户可通过客户端11提交待压缩的神经网络模型14,以及终端12的运行参数,由客户端11调用模型压缩系统13的接口对神经网络模型14进行压缩,得到压缩后的神经网络模型15。

[0072] 作为示例,图1A中所示的神经网络模型14的结构可如图1B所示,神经网络模型14的节点包括:节点0、节点1、节点2、节点3、节点4和节点5;节点0和节点1之间的计算操作141为卷积 3×3 即Conv 3×3 ,节点1和节点2之间的计算操作142为Conv 3×3 ,节点3和节点4之间的计算操作143为Conv 3×3 ,节点4和节点5之间的计算操作144为Conv 3×3 。

[0073] 压缩后的神经网络模型15的结构可如图1C所示,节点0和节点1之间的计算操作151为Conv 3×3 ,节点1和节点2之间的计算操作152为Conv 5×5 ,节点3和节点4之间的计算操作153为最大池化,节点4和节点5之间的计算操作154为Conv 3×3 ,另外,节点0和节点3之间还包括计算操作155且为均值池化,节点2和节点4之间包括计算操作156且为身份映射。其中,Conv $kw \times kh$ 表征卷积层的卷积核的宽和高分别为kw和kh。

[0074] 可见,压缩后的神经网络模型15相对于神经网络模型14来说,节点1和节点2之间的计算操作由Conv 3×3 调整为Conv 5×5 ,将节点3和节点4之间的计算操作由Conv 3×3 调整为最大池化,并在节点0和节点3之间增加了均值池化,以及在节点2和节点4之间增加了身份映射,从而计算操作和通道都得到简化。

[0075] 压缩后的神经网络模型15可以被返回至终端12,终端12运行压缩后的神经网络模型15可以对接收到的样本进行分类操作并得到分类结果,例如基于图像样本的人脸识别,

基于音频样本的语音识别等。

[0076] 参见图2,图2是本发明实施例提供的神经网络模型压缩方法的流程示意图,将结合图2示出的步骤进行说明。

[0077] 步骤S201,将神经网络模型分解为包括通道的计算操作。

[0078] 在一些实施例中,对于用户预训练完成的神经网络模型,由于其结构和参数上存在冗余,需要进行压缩。根据神经网络模型所包括的网络层,将神经网络模型分解为一系列连接的计算操作,对于两个连接的计算操作来说,第一个计算操作的输出通道将作为第二计算操作的输入通道;计算操作中的信息流将对应被分解为通道。以网络层为卷积层为例,其通道的数量与卷积核的通道数量对应,如果卷积层中有3个卷积核,则该卷积层中包括3个通道。

[0079] 以图1B中的神经网络模型14为例,对分解过程进行说明,神经网络模型14包括4个卷积层 $\text{Conv}3 \times 3$,则分解得到的计算操作包括:计算操作141: $\text{Conv}3 \times 3$,计算操作142: $\text{Conv}3 \times 3$,计算操作143: $\text{Conv}3 \times 3$,计算操作144: $\text{Conv}3 \times 3$ 。

[0080] 假设计算操作1中的卷积核数量为3,则每个卷积核可以分别针对计算操作1的输入进行卷积操作,并独立输出,形成3个通道;相应地,计算操作1被分解为3个通道。

[0081] 这里,将第一个卷积层与第二个卷积层之间不同的信息流分解为多个通道,将第三个卷积层与第四个卷积层之间不同的信息流分解为多个通道。其中,各信息流的通道的数量与该信息流所输出的网络层的卷积核的数量有关。

[0082] 在实际应用中,可将神经网络模型作为整体分解为计算操作和通道,对神经网络模型整体上进行压缩。当然,也可在将神经网络模型分解为计算操作前,将神经网络模型划分为多个连续的搜索子空间(也即是子网络),以搜索子空间为单位,将每个搜索子空间分解为包括通道的计算操作,以在后续S202和S203中分别进行压缩,然后合成。

[0083] 步骤S202,在保持计算操作和通道所形成的结构不变的情况下,训练神经网络模型以更新计算操作的参数。

[0084] 在一些实施例中,保持计算操作和通道所形成的结构不变,即保持计算操作对应的权重和各通道对应的权重不变,通过如下的方式进行训练:构造神经网络模型的损失函数,将训练数据集代入损失函数,通过梯度下降的方式求解计算操作的参数,通过迭代计算来不断更新参数。

[0085] 步骤S203,在保持计算操作的参数不变的情况下,训练神经网络模型以更新计算操作和通道分别对应的权重。

[0086] 在一些实施例中,在保持计算操作的参数不变的情况下,通过如下的方式进行训练:将验证数据集(与训练数据集不同)代入神经网络模型的损失函数,求解计算操作和通道分别对应的权重,通过迭代计算的方式来不断更新权重。

[0087] 作为示例,在基于验证数据集训练神经网络模型以更新计算操作和通道分别对应的权重时,根据神经网络模型的目标设备(也就是用于运行神经网络模型的设备,例如运行FPGA、RISC和ARM等资源受限的计算设备)的运行参数,确定对应的通道稀疏程度;基于验证数据集训练神经网络模型,以更新计算操作和通道分别对应的权重,直到计算操作和通道分别对应的权重满足通道稀疏程度。

[0088] 其中,运行参数用于表示参数量、计算速度等体现神经网络模型的运行条件的参

数,运行参数与通道稀疏程度具有关联性,例如线性关联。举例来说,参考模型在设备运行时需求的硬件资源是运行参数的两倍,参考模型的通道稀疏程度为0.6,则运行参数对应的特定的通道稀疏程度为0.3。

[0089] 在一些实施例中,当执行步骤S202和步骤S203之后,还可以继续交替执行S202和S203,即交替更新计算操作的参数、以及神经网络模型的权重(包括计算操作的权重和通道的权重),直至神经网络模型的损失函数满足收敛条件,并继续执行S204。

[0090] 作为示例,将神经网络模型作为整体应用上述步骤S202和步骤S203,从而能够从整体上提取训练的效率;或者,当神经网络模型被划分为多个子网络时,针对每个子网络应用上述步骤S202和步骤S203,以便于更准确定位冗余的通道和计算操作,并通过训练来赋予与其重要性相应的权重。

[0091] 步骤S204,基于更新后的权重删除冗余的计算操作,并在保留的计算操作中删除冗余的通道。

[0092] 在一些实施例中,当神经网络模型收敛后,对神经网络模型中任意两个节点之间的所有计算操作的权重进行排序,排序在前的预定数量计算操作、排序在前的预定比例的计算操作、操作权重超出操作权重阈值的计算操作中的至少一种被保留,其余的计算操作作为冗余计算操作将被删除。在确定保留的计算操作后,根据保留的计算操作中的通道对应的权重,继续删除保留的计算操作中冗余的通道。

[0093] 作为冗余通道的示例,任意两个节点之间的计算操作所包括的通道,根据通道权重进行排序,排序在前的预定数量通道、排序在前的预定比例的通道、权重超出通道权重阈值的计算操作中的至少一种将被保留,其余的通道将被标记为冗余通道并删除。

[0094] 作为上述的两个节点的示例,当将神经网络中每个子网络分别应用上述步骤S202和步骤S203时,可以是每个子网络的输入节点和输出节点,相应地,每个子网络将应用上述步骤S204以删除冗余的计算操作和通道;当将神经网络模型作为整体应用上述步骤S202和步骤S203时,上述的两个节点可以为神经网络模型的输入节点和输出节点。

[0095] 作为删除冗余计算操作的例外,为了保证数据处理的完整性,当两个节点之间的只有一个计算操作时,由于不存在冗余,该计算操作将不被识别为冗余计算操作,从而得以保留。

[0096] 继续结合图3对冗余通道的删除进行说明。如图3所示的神经网络模型,A、B和C是特征图,代表神经网络模型中的输入节点、中间节点和输出节点。节点A、B中间的计算操作包括6个卷积核(图3中示例性示出了一个卷积核31),从而产生6个通道(记为W1)的输出。当部分通道被删除后(图3中以虚线表示被删除的通道),只有3个输出通道传递到节点B,从而节点B包括3个特征图,经过节点B、C之间的计算操作(使用一个卷积核32),产生1个输出通道到达节点C。

[0097] 步骤S205,基于保留的计算操作、相应的通道以及参数,构造压缩后的神经网络模型。

[0098] 在S204中将冗余的计算操作删除,并将保留的计算操作中的冗余通道删除后,基于保留的计算操作中保留的通道,以及保留的通道对应的参数,构建压缩后的神经网络模型。

[0099] 假设神经网络模型的计算操作包括:计算操作1、计算操作2、计算操作3、计算操作

4,其中,计算操作2和计算操作3的节点相同,计算操作1中的通道包括C11、C12和C13,计算操作2中的通道包括:C21、C22、C23、C24和C25,计算操作3中的通道包括:C31、C32、C33、C34和C35,计算操作4中的通道包括:C41和C42。

[0100] 当神经网络模型应用上述步骤S202和步骤203后,计算操作和通道分别被赋予权重,假设计算操作1、计算操作2、计算操作3、计算操作4对应的权重分别为: α_1 、 α_2 、 α_3 和 α_4 ,其中, $\alpha_2 > \alpha_3$ 。

[0101] 删除计算操作和通道的示例性处理如下:将计算操作3作为冗余计算操作删除。将计算操作1、计算操作2和计算操作4作为保留的计算操作。对于计算操作1,通道C11、C12和C13对应的权重分别为 β_{11} 、 β_{12} 和 β_{13} ,且 β_{11} 、 β_{12} 和 β_{13} 都大于0,则所有的通道都保留。对于计算操作2,通道C21、C22、C23、C24和C25对应的权重分别为 β_{21} 、 β_{22} 、 β_{23} 、 β_{24} 和 β_{25} ,且为 β_{21} 、 β_{23} 为0, β_{22} 、 β_{24} 和 β_{25} 都大于0,则将C21、C23删除,将C22、C24和C25保留。对于计算操作4,C41和C42对应的权重分别为 β_{41} 、 β_{42} , β_{41} 、 β_{42} 都大于0,则所有的通道都保留。基于计算操作1中的通道C11、C12和C13、计算操作2中的C22、C24和C25和计算操作4中的C41和C42以及各计算操作中相应的参数构成压缩后的神经网络模型

[0102] 需要说明的是,当将神经网络模型划分为多个搜索子空间时,通过步骤S202至步骤S205对各搜索子空间进行压缩,得到多个压缩后的搜索子空间,此时,将多个压缩后的搜索子空间进行堆叠,得到压缩后的神经网络模型。

[0103] 可见,通过将待压缩的神经网络模型分解为计算操作和通道两个维度,并通过操作权重来删除冗余计算操作以消除结构冗余,通过通道权重来删除冗余通道以消除参数冗余,实现模型的有效精简。

[0104] 在实际应用中,可将文字、图片、音频、视频等待分类的数据输入经过本发明实施例提供的神经网络模型的压缩方法压缩的神经网络模型,得到输出的分类结果。基于压缩的神经网络模型的参数相对于压缩前,从网络层和通道两个层面,最大可能地减少冗余参数,从而减少神经网络模型在分类过程中的计算开销以及所占用的硬件资源,能够加快分类速度。

[0105] 在一些实施例中,步骤S201至步骤S205所应用的神经网络模型可以是针对生产环境中的特定应用(例如人脸识别、语音识别等)而预训练(简称为预训练模型或原始神经网络模型)。

[0106] 在另一些实施例中,鉴于预训练模型(尤其是深度神经网络)的不可微特性而导致在训练是难以使用梯度下降而求解参数的情况,步骤S201至步骤S205所应用的神经网络模型还可以是针对预训练神经网络模型中添加候选计算操作而得到的重构神经网络模型,使节点之间的连接松弛且连续,具有梯度下降的特性,能够通过梯度下降法对重构神经网络模型的参数或权重进行更新,加快神经网络模型的收敛速度。

[0107] 作为示例,在预训练神经网络所包括的计算操作即原始计算操作的基础上,在预训练神经网络的不同节点之间添加候选计算操作,在得到重构网络模型中,两个节点间包括有原始计算操作和的候选计算操作。在应用上述步骤201时,根据重构神经网络模型所包括的原始计算操作和候选计算操作将重构神经网络模型分解为通过通道连接的计算操作。

[0108] 以图1B的神经网络结构14为原始神经网络模型为例,在节点0和节点1之间不添加操作,在节点1和节点2之间已存在原始计算操作Conv 3×3 的情况下,添加候选计算操作

Conv 5×5 ,在节点0和节点3之间添加候选计算操作均值池化,在节点2和节点4之间添加候选计算操作映射,在节点3和节点4之间已存在原始计算操作Conv 3×3 的情况下,添加候选计算操作即最大池化层。

[0109] 这里,将与原始计算操作的两端节点相同的候选计算操作称为原始计算操作对应的候选计算操作。比如,在节点1和节点2中包括原始计算操作Conv 3×3 ,在节点1和节点2之间包括候选计算操作Conv 5×5 ,则Conv 5×5 为Conv 3×3 对应的候选计算操作。

[0110] 当重构神经网络模型在应用上述步骤203时,在保持重构神经网络模型中的计算操作和通道所形成的结构不变的情况下,训练重构神经网络模型以更新计算操作的参数;在保持计算操作的参数不变的情况下,训练重构神经网络模型以更新计算操作和通道分别对应的权重。此时的计算操作包括:原始神经网络模型中的原始计算操作和添加的候选计算操作。

[0111] 在一些实施例中,当应用上述步骤S201至步骤S205神经网络模型重构神经网络模型时,当对神经网络模型进行训练即应用上述步骤S202和步骤S203之前,还可以初始化神经网络模型中的原始计算操作和候选计算操作,以使神经网络模型在应用上述步骤S202和步骤S203后,原始计算操作的重要性显著于候选计算操作的重要性,从而使预训练神经网络模型的性能不会被候选计算操作干扰。

[0112] 下面结合具体的示例进行说明。基于原始计算操作的参数的不同,初始化可包括以下如下的初始化方式。

[0113] 在初始化方式一中,当原始计算操作不存在参数或通道权重时,对原始计算操作对应的候选计算操作的参数进行随机初始化;初始化原始计算操作对应的权重且大于所述候选计算操作的初始化的权重。

[0114] 当原始计算操作的类型为映射时,原始计算操作不存在参数。当原始计算操作的类型为最大池化、均值池化时,输入的信息流和输出的信息流的通道数一致,该计算操作对应的网络层的神经元与邻接网络层的神经元之间的信息流的数量不变,则不需要进行通道筛选,不存在通道权重。

[0115] 对于原始计算操作的参数,保留各参数,对于候选计算操作,对候选计算操作的参数进行随机初始化。对于计算操作对应的权重,初始化原始计算操作对应的权重大于候选计算操作的权重,使得初始化后的重构神经网络模型更偏向于选择原始计算操作,比如:将原始计算操作对应的权重初始化为1,将候选计算操作的对应的权重初始化为0,从而不会破坏神经网络模型的训练过程。其中,不同的候选计算操作对应的权重可相同,也可不同。

[0116] 在初始化方式二中,当原始计算操作存在参数时,根据原始计算操作的参数,初始化原始计算操作对应的候选计算操作的参数,使得原始计算操作与所对应的候选计算操作的重构误差最小化。

[0117] 当原始计算操作为卷积、批量标准化等有参数的计算操作,假设节点之间的计算操作共有5个,原始计算操作是第3个计算操作,候选计算操作为第1个、第2个、第4个、第5个计算操作,基于最小化第3个计算操作分别与第1个、第2个、第4个、第5个计算操作的重构误差,对第1个、第2个、第4个、第5个计算操作的参数进行初始化。

[0118] 以图1B所示的神经网络模型14为例,节点1和2中原始计算操作是Conv 3×3 ,对于引入的候选计算操作Conv 5×5 而言,通过最小化重构误差使得Conv 5×5 实现和Conv 3×3 得

到几乎完全一样的计算结果。在对Conv 5x5的参数进行初始化时,记录的神经网络模型11输出的结果1,将神经网络模型14中Conv 3x3替换为Conv 5x5,得到将Conv 5x5替换Conv 3x3后新的神经网络模型的输出的结果2,通过对Conv 5x5的参数的调整,最小化输出的结果1和结果2的重构误差,实现Conv 5x5和Conv 3x3得到几乎完全一样的计算结果。

[0119] 本发明实施例还提供一种神经网络模型的压缩方法,将结合图4示出的步骤进行说明。

[0120] 步骤S401,在模型压缩系统的客户端中接收神经网络模型、以及神经模型的目标设备的运行参数。

[0121] 模型压缩系统的客户端接收用户输入的或目标设备发送的神经网络模型以及目标设备的运行参数。运行参数可为用户在模型压缩系统的客户端输入的,也可为模型压缩系统的客户端从运行目标设备所接收的。其中,运行神经网络模型的目标设备可通过对自身运行环境的检测获得运行参数,并发送至模型压缩系统的客户端。

[0122] 步骤S402,调用模型压缩系统,以根据目标设备的运行参数确定通道稀疏程度,以及以通道稀疏程度为目标,对神经网络模型中的计算操作、以及连接计算操作的通道分别进行压缩,得到压缩后的神经网络模型。

[0123] 模型压缩系统的客户端调用模型压缩系统,由所调用的模型压缩系统根据运行参数确定通道系数程度,并基于通道稀疏程度,应用上述步骤S201至步骤S205以构造压缩后的神经网络模型。

[0124] 其中,模型压缩系统根据参考运行参数和参考通道稀疏程度,确定获得的运行参数对应的通道稀疏程度,参考运行参数为参考神经网络模型的参数量、内存开销、计算速度等运行成本,参考通道稀疏程度为参考运行参数对应的通道稀疏程度。

[0125] 模型压缩系统评估参考运行参数,并与输入的运行参数进行比较,按比例对通道数量进行缩放,把输入的运行参数转化为特定的通道稀疏程度。比如:参考神经网络模型的计算成本是输入的运行参数的两倍,参考神经网络模型的计算成本对应的通道稀疏程度为1,则接收的运行参数对应的特定的通道稀疏程度为0.5。

[0126] 继续说明本发明的示例性实施。在实施本发明实施例的过程中,发现相关技术提供一种降低网络层的宽度的网络模型压缩算法:通道减枝。

[0127] 通道剪枝是通过对网络层中输入的信息流的冗余通道进行识别并剔除,以达到压缩模型的目的。其中,在网络层中输入的信息流的不同通道上引入一个通道权重 β ,通过通道权重 β 表示各通道的重要性。通过最小化一通道引入通道权重 β 前后节点的差值,同时使得操作权重 β 稀疏化的方式,对模型的参数以及通道权重进行训练,最终把通道权重小的通道去除以达到模型压缩的目的。

[0128] 如图5所示,网络层1-1的神经元和网络层1的神经元之间不同信息流对应的通道对应的权重构成权重向量,权重向量的元素为各通道对应的通道权重 β ,初始化各通道的通道权重 β 为1。通过对网络层1-1和网络层1所在的神经网络的训练,最小化一通道引入通道权重 β 前后信息流的差值,同时使得通道权重 β 稀疏化的方式,对网络层的参数以及通道权重进行更新。训练后,权重向量中的通道权重根据对应通道的重要性进行更新,比如:一通道的操作权重 β 的值增大为1.1,部分通道的操作权重 β 的值减小至0.8、0等,将网络层1-1中操作权重 β 为0的通道(如虚线所示)剔除,且基于剔除以后的通道来得到网络层1。

[0129] 可见,通道减枝能够减小网络层的神经元的数量,减小网络层的参数,从而减小神经网络模型的参数量,但是通道减枝会固定神经网络模型的结构,因此压缩后的神经网络模型的结构可能已经不是最优的,导致压缩后的神经网络模型存在严重的结构冗余,限制模型的训练性能。

[0130] 在实施本发明实施例的过程中,发现相关技术还提供一种网络结构调整算法:可微结构搜索(Differentiable Architecture Search,DARTS)。其中,以图6所示的模型结构对DARTS进行说明,在图6中,在节点0、节点1、节点2和节点3的每两个节点之间包括有通过不同的线条代表的不同计算操作,实线表示通过权重所选择的计算操作,即保留的计算操作,虚线表示没有选择到的冗余的计算操作。

[0131] DARTS在选择计算操作时,在任意两个节点中同时考虑多种不同的计算操作(包括卷积层、池化层等)。对于任意一个中间节点,可以通过一种计算操作组合的方式得到最终的计算结果如公式(1):

$$x^{(i)} = \sum_{j < i} o^{(i,j)}(x^{(j)}) \quad \text{公式(1)},$$

[0133] 其中, $o^{(i,j)}$ 表示节点 x_i 和节点 x_j 之间的计算操作。为了实现不同计算操作的选择,可以对不同操作赋予权重 α , α 表示两个节点之间该计算操作的重要性,最终选择权重最高的计算操作作为最终的计算方式。

[0134] 在神经网络模型的训练过程中,通过使用基于梯度的优化方法对不同操作的权重进行训练,会使得能够提高模型准确率的计算操作的对应权重上升,其他降低模型准确率的计算操作对应的权重下降,从而达到自动搜搜神经网络模型的结构的目的。

[0135] 但是,基于DARTS的网络结构搜索方法会固定每一种计算操作的通道数,最终搜索到的模型的结构往往结构复杂,包含大量的参数冗余,不能达到设计精简模型的目的。

[0136] 针对相关技术提供的模型精简方案的上述问题,本发明实施例提供如图2所示的神经网络模型压缩方法,在模型压缩的过程中引入网络结构自动调整机制,以寻找减枝后的模型最优的网络结构,以得到高效精简的深度神经网络,从而能够实现动态的网络结构压缩,得到准确率高且参数量少的精简模型。

[0137] 在对神经网络模型进行压缩时,可先将神经网络模型划分为多个搜索子空间(记为cell),以图7所示神经网络模型的结构为例,对cell的划分进行说明。其中,在图7中,输入701作为神经网络模型的输入,之后接着3个重复的模块702、703和704,得到输出结构705。每个模块包括N个标准搜索子空间(Normal cell)和一个下降搜索子空间,其中,N个Normal cell用来增加神经网络模型的非线性表达能力,Reduction cell类似于pooling操作,用来降低神经网络模型所提取的特征的空间大小。两种cell对应于两种不同的网络结构,但Normal cell中的操作保证特征空间大小不变,而Reduction cell会将空间大小减半。两种cell都可能含有大量模型冗余,二者的结构都需要被精简。

[0138] 将神经网络模型划分为多个连续的cell,且对各cell进行压缩,得到压缩后的cell,将各压缩的cell进行堆叠,构成压缩后的神经网络模型。

[0139] 下面以一个cell例对本发明实施例提供的神经网络的压缩方法进行说明。

[0140] 这里,引入了 α 和 β 这两个权重,分别表示不同计算操作和不同通道的重要性,从而同时对通道数量和网络结构即网络的网络层之间的连接关系进行优化。当引入 α 和 β 后,网

络层的计算结果 y 可表示为公式(2)：

$$[0141] \quad y = \sum_{i=1}^{\|O\|} \alpha_i \sum_{c=1}^C \beta_c \alpha(x_c, \theta_{i,c}) \quad \text{公式(2)},$$

[0142] 其中, α_i 表示第 i 个计算操作的重要性, β_c 表示第 c 个通道的重要性, α_i 表示第 i 个计算操作中第 c 个通道输入的信息流, $\theta_{i,c}$ 表示第 i 个计算操作中第 c 个通道所包含的参数。

[0143] 上述的 α_i 和 β_c 作为变量可以通过基于梯度的优化方法进行训练。在神经网络模型的训练中不断的更新 β_c ,使得对神经网络模型的最终分类结果准确率贡献大的通道权重上升,贡献小的通道的权重下降,最后按照 β_c 的大小进行排序,判别每个通道的重要性,把权重小的通道也就是不重要的通道作为冗余通道直接剔除,以实现通道压缩的目的。在神经网络模型的训练中不断的更新 α_i ,使得任意两节点之间能够提升模型准确率的计算操作对应的 α_i 上升,降低会影响模型准确率的计算操作的相应 α_i ,最终选择 α_i 最高的计算操作以达到自动调整神经网络模型的结构的目的。

[0144] 在 α_i 和 β_c 的更新过程中,两个权重的调整是相互影响的。因为在优化过程中两个权重是一起通过梯度下降方法进行更新,两个权重的取值会共同影响最终的模型计算结果。

[0145] 以一个cell的结构同图1B所示的神经网络模型14的结构为例,继续对本发明实施例提供的神经网络模型的压缩方法进行说明。

[0146] 如图1B所示,神经网络模型14为原始的卷积神经网络,是具有固定卷积核大小的卷积层堆叠在一起构成的模型,且每一层的通道数量都很宽,存在大量的参数冗余;神经网络模型15为经过结构调整后的精简的神经网络,每一层的卷积核数量都被大幅度压缩,同时根据压缩后的参数相应的结构也会进行调整,以找到最优的网络结构。可见,本发明实施例提供的神经网络模型的压缩方法,能够针对图片识别、视频动作识别等应用场景,对已有的神经网络模型的参数和结构进行压缩,最终得到参数量少、计算开销低的精简神经网络,以满足不同应用场景对计算资源和计算速度的要求。

[0147] 下面,以图1B所示的结构一个cell为例,对神经网络模型的压缩流程进行说明。

[0148] 第一步:在已有的网络模型中加入其他可选的计算操作,构造适用于网络结构搜索的模型结构。

[0149] 如图1B所示的神经网络模型14中,节点1和节点2之间的原始计算操作是Conv 3x3,为了进行结构自动调整,在神经网络模型14的节点1和节点2间增加其他可选的计算操作即候选计算操作,比如:Conv 5x5、最大池化(MaxPool)等计算操作。在本发明实施例中,对所增加的候选计算操作的类型不进行任何的限定。

[0150] 第二步:利用基于重构的方式对神经网络模型中全部已有的计算操作即原始计算操作和新增的计算操作即候选计算操作的参数进行初始化。

[0151] 这里,通过重构的方式对原始计算操作和候选计算操作的参数进行初始化,能够避免以随机方式初始化的候选计算操作被选择概率过小的问题,保证不同的计算操作能够被公平地选择。

[0152] 比如:在图1B所示的神经网络模型14中,节点1和节点2之间原始计算操作为Conv 3x3,在节点1和节点2之间增加候选计算操作Conv 5x5后,基于重构的方式对Conv 5x5的参数进行初始化。

[0153] 通过重构的方式对候选计算操作的参数进行初始化,能够避免候选计算操作的参

数没有经过训练,而严重影响神经网络模型最终的性能的问题。

[0154] 第三步:在训练数据集上,固定模型的结构和通道数量,通过梯度下降算法对模型中所有操作的参数进行优化。

[0155] 节点之间不同计算操作对应的权重和不同通道对应的权重都会固定,通过优化模型最终的准确率来更新神经网络模型中所有计算操作的参数。其中,这里,所更新的参数所属的计算操作包括神经网络模型的原始计算操作和在第二部引入的候选计算操作。

[0156] 第四步:在验证数据集上,固定当前神经网络模型的参数,根据网络层通道稀疏程度的要求,同时对通道数量和网络结构进行调整。

[0157] 验证数据集与训练数据集的数据集不重叠,验证数据集中的数据为神经网络模型在训练过程中没见过的数据。

[0158] 以图1B为例,神经网络模型14的节点1和2之间的原始计算操作为Conv 3x3,在增加Conv 5x5后,通过对神经网络模型的训练,确定Conv 5x5的效果要优于Conv 3x3,则增加Conv 5x5对应的权重而降低Conv 3x3对应的权重,最终在所有备选的计算操作(比如:Conv 3x3和Conv 5x5)中选择权重最高的权重对应的计算操作(比如:Conv 5x5)替换候选计算操作Conv 3x3,以实现结构自动调整。

[0159] 对于通道数量而言,在模型训练过程中会自动判别不同通道的重要性,最终把重要性低的通道剔除以实现通道数量压缩。

[0160] 第五步:通过交替训练的方式重复第三步和第四步,直到神经网络模型收敛。

[0161] 本发明实施例提供的神经网络模型的压缩方法,将搜索到的神经网络模型在训练数据集上微调参数,并在验证数据集上验证神经网络模型的性能,完成网络结构自动压缩过程。

[0162] 下面,对第二步中描述的基于重构的神经网络模型的参数的初始化进行说明。

[0163] 对于一个预训练的网络结构即原始神经网络模型,节点之间的计算操作只有一种的参数是已有的,如果对于其他的计算操作没有一个好的初始化,将会严重影响模型的训练。而以下基于重构的神经网络模型的参数的初始化能够有效解决该问题。参数的初始化过程如下:

[0164] 第一步:将预训练模型中已有的计算操作的参数赋值到待优化的模型中。

[0165] 预训练模型指的是已经训练完成的原始神经网络(对应图1B中的14),待优化模型为为了实现模型自动压缩,在原始计算操作的基础上添加候选计算操作所构造的模型,也就是重构神经网络。

[0166] 第二步:对于预训练模型的无参数操作或无通道权重的操作,将原始计算操作对应的权重设置为1,其余计算操作的参数随机初始化并设置对应的权重为0。

[0167] 无参数操作包括同一映射等不存在参数的计算操作,无通道权重的操作包括最大池化、平均池化等输入的信息流和输出的信息流的通道数相同的计算操作,对于无通道权重操作,每个通道上,没有需要训练的权重。

[0168] 当原始计算操作为无参数操作或无通道权重操作,无法直接通过优化重构误差对引入的候选计算操作的参数进行初始化,为了避免破坏模型的性能,将原始计算操作对应的权重设置为1,其余计算操作的参数随机初始化并设置对应的权重为0,这样模型会更偏向于选择原始计算操作,不会破坏模型的训练过程。

[0169] 第三步:对于预训练模型中有参数的操作,则通过最小化与已有操作的重构误差,对于其他有参数的操作的参数进行初始化,并设置所有计算操作对应的权重为1。

[0170] 有参数的操作为卷积、批量标准化等计算操作。以图1B所示的神经网络模型14为例,节点1和2中原本的操作是Conv 3x3,对于引入的候选计算操作Conv 5x5而言,通过最小化重构误差使得Conv 5x5实现和Conv 3x3得到几乎完全一样的计算结果。在对Conv 5x5的参数进行初始化时,记录原有cell的输出结果,cell中Conv 3x3替换为Conv 5x5,得到将Conv 5x5替换Conv 3x3后新的cell的输出结果,通过对Conv 5x5的参数调整,最小化这两种cell输出结果的重构误差,实现Conv 5x5和Conv 3x3得到几乎完全一样的计算结果。

[0171] 通过本发明实施例提供的候选计算操作的参数的初始化,能够使得重构神经网络模型中各个操作得到比较好的初始化。

[0172] 下面,结合图8对用户输入运行参数场景下的神经网络模型压缩方法进行进一步描述。其中,图8中的模型压缩系统800为一运行于服务器802上的平台,可自动化机器学习(AutoML)平台,终端801能够接收用户输入的运行参数,并基于运行参数对已有的神经网络模型14进行压缩,得到精简模型即压缩后的神经网络模型15。可见,AutoML能够根据用户对模型参数量、内存开销、计算速度等运行参数,自动对已有的神经网络模型进行结构压缩,以得到参数少且准确率高的精简模型。其中,终端801上运行有模型压缩系统800的客户端。

[0173] 以图片识别任务为例,在实际应用中,深度神经网络模型往往运行在低功率的硬件平台上,比如FPGA和移动端设备,因此,对于深度神经网络模型的参数量和计算速度有较严格的要求。为了将已有的深度神经网络模型应用在这些资源受限的硬件平台中,同时在保证模型性能不损失的情况下,必须对模型参数和模型结构进行压缩。压缩的流程包括:

[0174] 步骤S8001,接收用户输入的运行参数。

[0175] 用户可根据具体应用场景的硬件条件,将参数量、计算速度等运行参数输入到AutoML中。

[0176] 步骤S8002,根据参考运行参数和参考通道稀疏程度,确定运行参数对应的通道稀疏程度。

[0177] 参考运行参数为运行参考神经网络模型的参数量、内存开销、计算速度等计算成本,参考通道稀疏程度为参考运行参数对应的通道稀疏程度。

[0178] 评估参考运行参数,并与输入的运行参数进行比较,按比例对通道数量进行缩放,把输入的运行参数转化为特定的通道稀疏程度。比如:参考神经网络模型的计算成本是运行参数的两倍,参考神经网络模型的计算成本对应的通道稀疏程度为1,则接收的运行参数对应的通道稀疏程度为0.5。

[0179] 步骤S8003,以通道稀疏程度为目标,对预训练的深度神经网络模型进行网络结构压缩。

[0180] 利用图2所示的神经网络模型压缩算法对预训练的深度神经网络14进行自动网络结构压缩,最终得到满足要求的精简神经网络15。

[0181] 针对用户只需要提供参数量、内存开销等运行参数作为运行条件,不需要用户具备相关的机器学习和神经网络的专业知识的情况,AutoML平台能够根据用户的要求生成合适的精简模型给用户使用,即AutoML平台的自动网络结构压缩模块会大大降低用户在实际

资源受限平台上部署深度模型的门槛。本发明实施例中提供的神经网络模型的压缩方法借助AutoML平台更好地为用户提供满足部署要求的精简深度神经网络。这里,可在AutoML平台中设置一自动网络结构压缩模块,步骤S8001、S8002、S8003由自动网络结构压缩模块实现。

[0182] 图9所示的神经网络的压缩装置91和图10所示的神经网络装置92为本发明实施例提供的神经网络的压缩方法的示例性的功能结构。神经网络的压缩装置91的硬件层可以部署图8所示的运行模型压缩系统的服务器802的方式实现,神经网络的压缩装置92的硬件层可以采用图8所示的终端801的方式实现。

[0183] 参见图9或图10,图9或图10是本发明实施例提供的服务器802或终端801的示例性的结构示意图,包括至少一个处理器1001、存储器1002、至少一个网络接口1003和用户接口1004。神经网络的压缩装置100或110的各个组件通过总线系统1005耦合在一起。可理解,总线系统1005用于实现这些组件之间的连接通信。总线系统1005除包括数据总线之外,还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见,在图9或图10中将各种总线都标为总线系统1005。用户接口1004可以包括显示器、键盘、鼠标、轨迹球、点击轮、按键、按钮、触感板或者触摸屏等。存储器1002可以是易失性存储器或非易失性存储器,也可包括易失性和非易失性存储器两者。

[0184] 本发明实施例中的存储器1002能够存储数据以支持服务器802或终端801的操作。这些数据的示例包括:用于在服务器802或终端801上操作的任何数据,如:训练数据集、验证数据集,又如:参考运行参数,还包括任何计算机程序,如操作系统和应用程序。其中,操作系统包含各种系统程序,用于实现各种基础业务以及处理基于硬件的任务。

[0185] 作为本发明实施例提供的方法采用软硬件结合实施的示例,本发明实施例所提供的神经网络的压缩方法可以直接体现为由处理器1001执行的软件模块组合,软件模块可以位于计算机可读存储介质中,计算机可读存储介质位于存储器1002,处理器1001读取存储器1002中软件模块包括的可执行指令,结合必要的硬件(例如,包括处理器1001以及连接到总线1005的其他组件)完成本发明实施例提供的神经网络的压缩方法。

[0186] 下面结合图9说明存储器1002中的软件模块组合的示例,包括:分解单元911,用于将神经网络模型分解为通过包括通道的计算操作;第一训练单元912,用于在保持计算操作和通道所形成的结构不变的情况下,训练神经网络模型以更新计算操作的参数;第二训练单元913,用于在保持计算操作的参数不变的情况下,训练神经网络模型以更新计算操作和通道分别对应的权重;删除单元914,用于基于更新后的权重删除冗余的计算操作,并在保留的计算操作中删除冗余的通道;构造单元915,用于基于保留的计算操作、相应的通道以及参数,构造压缩后的神经网络模型。

[0187] 在一些实施例中,分解单元911,包括:第一分解模块,用于根据神经网络模型包括的网络层,将神经网络模型分解为对应网络层的计算操作;第二分解模块,用于将网络层的神经元与邻接网络层的神经元之间的不同信息流,分解为与信息流对应的通道。

[0188] 在一些实施例中,第一训练单元912,具体用于:在保持计算操作和通道分别对应的权重不变的情况下,基于训练数据集训练神经网络模型以更新计算操作的参数。

[0189] 在一些实施例中,第二训练模块913,具体用于:在保持计算操作的参数不变的情况下,基于验证数据集训练神经网络模型以更新计算操作和通道分别对应的权重。

[0190] 在一些实施例中,第二训练模块914,具体用于:根据用于神经网络模型的目标设备的运行参数,确定对应的通道稀疏程度;基于验证数据集训练神经网络模型以更新计算操作和通道分别对应的权重,直到计算操作和通道分别对应的权重满足通道稀疏程度。

[0191] 在一些实施例中,神经网络的压缩装置91,还包括:交替模块,用于交替更新计算操作的参数、以及神经网络模型的权重,直至神经网络模型满足收敛条件;其中,神经网络模型的权重包括:计算操作的权重和通道的权重。

[0192] 在一些实施例中,神经网络的压缩装置91还包括:添加模块,用于训练神经网络模型之前,在神经网络模型中添加候选计算操作,以使神经网络模型具有梯度下降的特性。

[0193] 在一些实施例中,神经网络的压缩装置91还包括:初始化模块,具体用于训练神经网络模型之前,初始化神经网络模型中的原始计算操作和候选计算操作。

[0194] 在一些实施例中,初始化模块,具体用于:当原始计算操作不存在参数或通道权重时,对原始计算操作对应的候选计算操作的参数进行随机初始化;初始化原始计算操作对应的权重,且大于所述候选计算操作的初始化的权重。

[0195] 在一些实施例中,初始化模块,用于:当原始计算操作存在参数时,根据原始计算操作的参数,通过最小化原始计算操作与所对应的候选计算操作的重构误差的方式,初始化原始计算操作对应的候选计算操作的参数。

[0196] 本发明实施例中的图10所示的存储器1002中的软件模块组合的示例,包括:确定单元921,用于在模型压缩系统的前端页面或客户端中接收神经网络模型、以及神经模型的目标设备的运行参数;压缩单元922,用于调用模型压缩系统,以根据目标设备的运行参数确定通道稀疏程度,以及以通道稀疏程度为目标,对神经网络模型中的计算操作、以及连接计算操作的通道分别进行压缩,得到压缩后的神经网络模型。其中,模型压缩系统可由图9中的分解单元911、第一训练单元912、第二训练单元913、删除单元914、构造单元915实现。

[0197] 本发明实施例还提供一种存储介质,即计算机可读存储介质,计算机可读存储介质中存储有计算机程序,计算机程序被处理器执行时,实现本发明实施例提供的神经网络模型的压缩方法。

[0198] 综上,在通过通道减枝能够减少模型参数的基础上,解决了减枝后模型结构不能调整的问题,可以得到准确率更高、更精简的神经网络模型;由于在进行自动网络结构压缩前,不同计算操作的参数能够较好的初始化,可以加快算法的收敛速度,并得到更好的神经网络模型。

[0199] 以上所述,仅为本发明的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应以权利要求的保护范围为准。

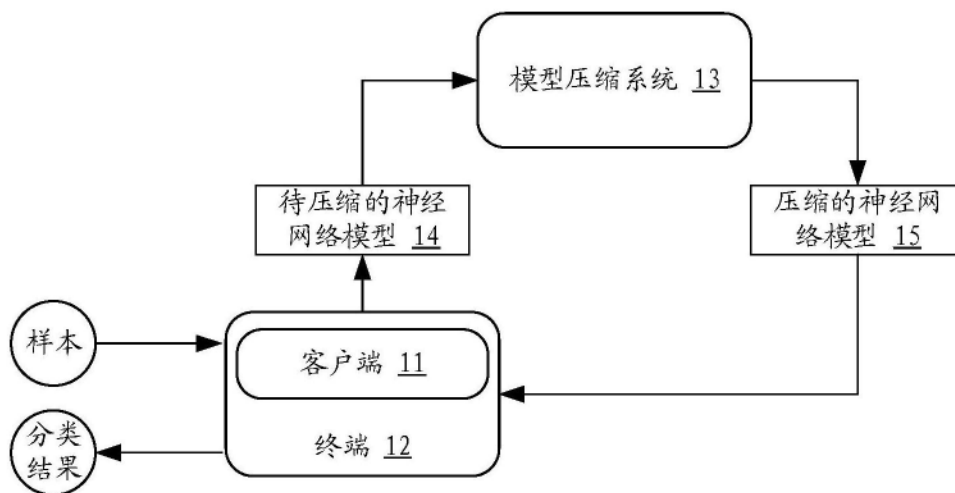


图1A

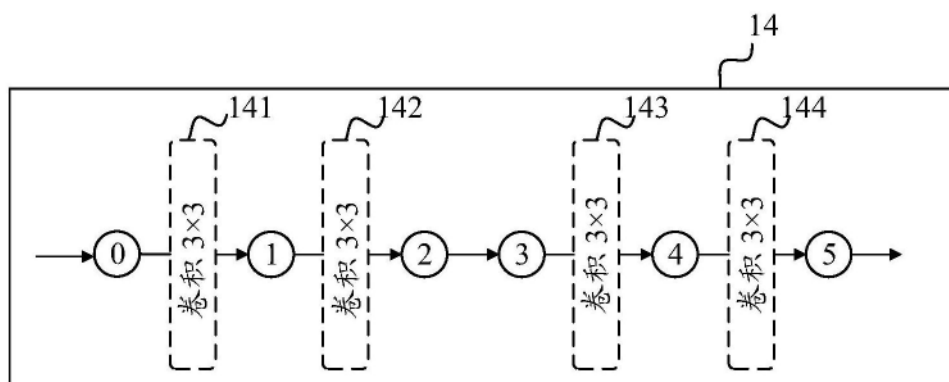


图1B

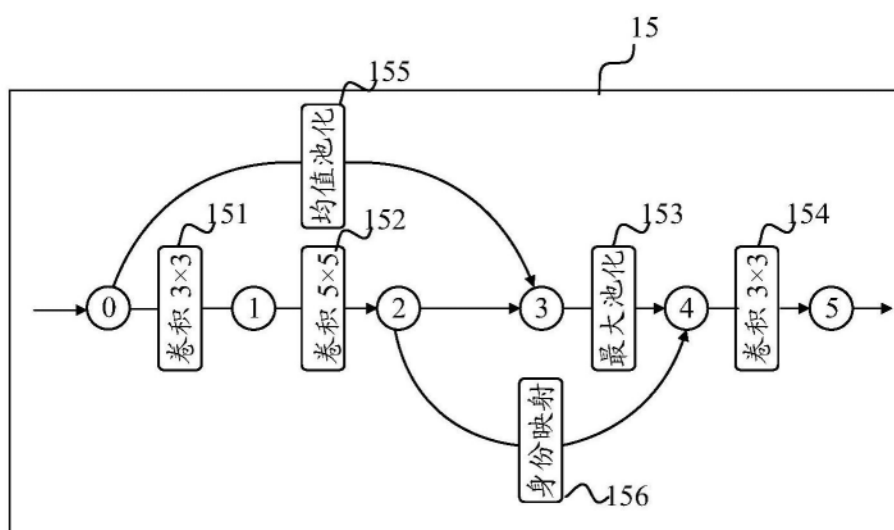


图1C

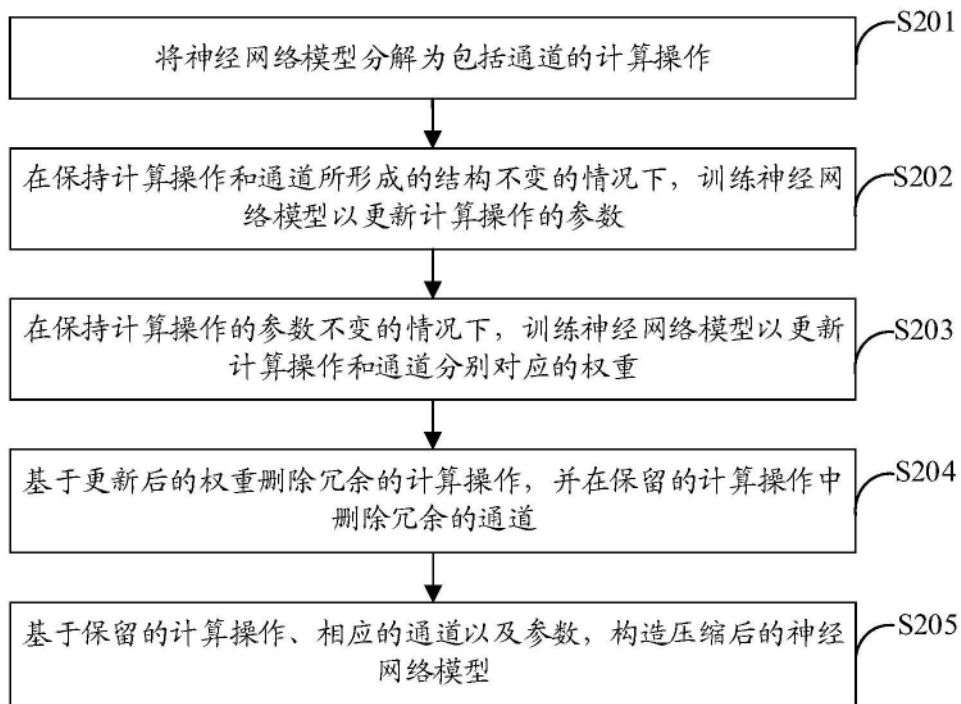


图2

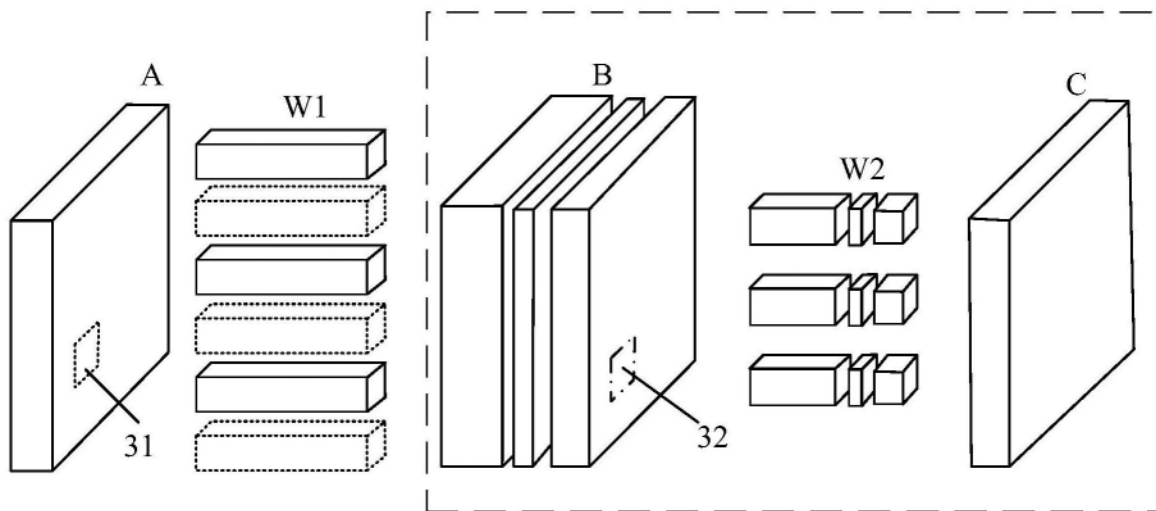


图3

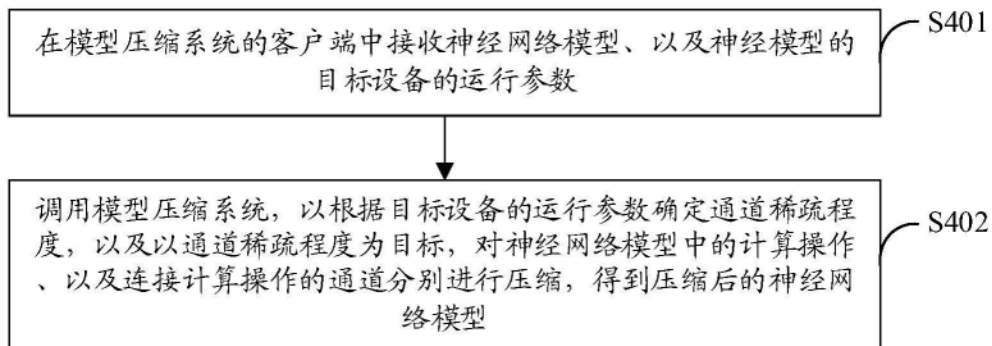


图4

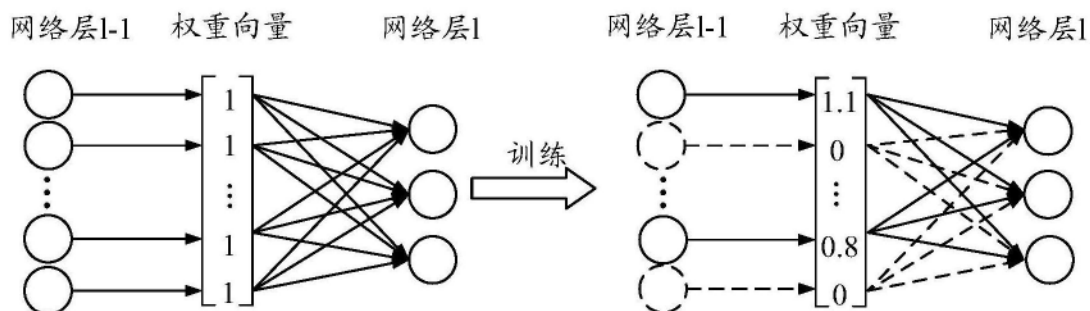


图5

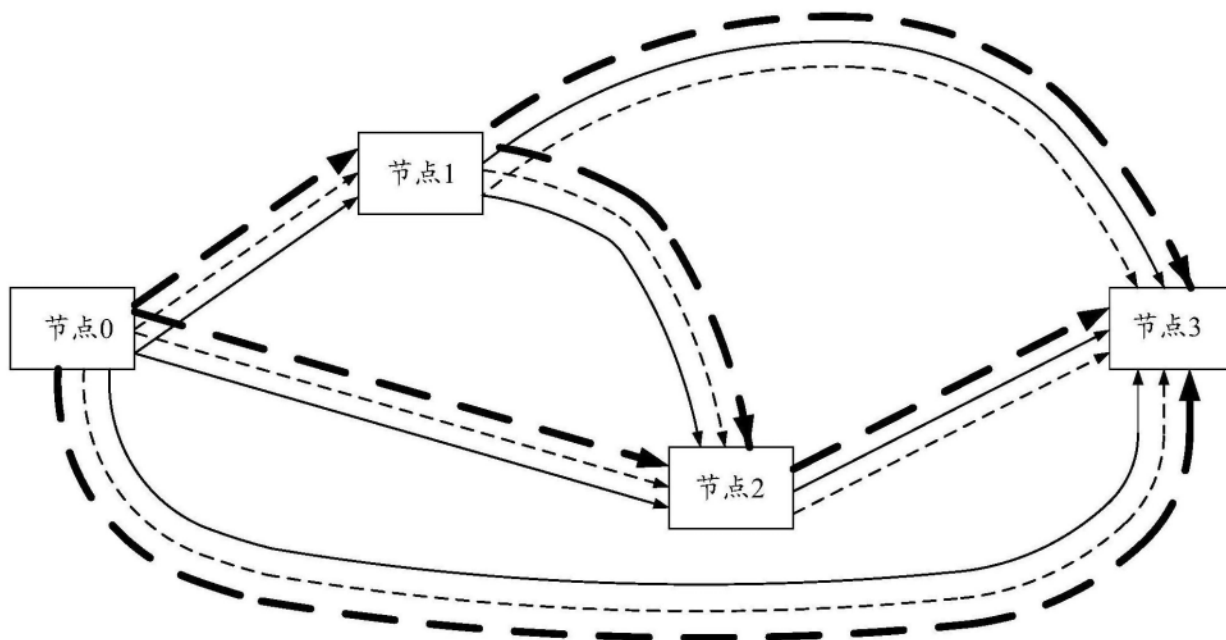


图6

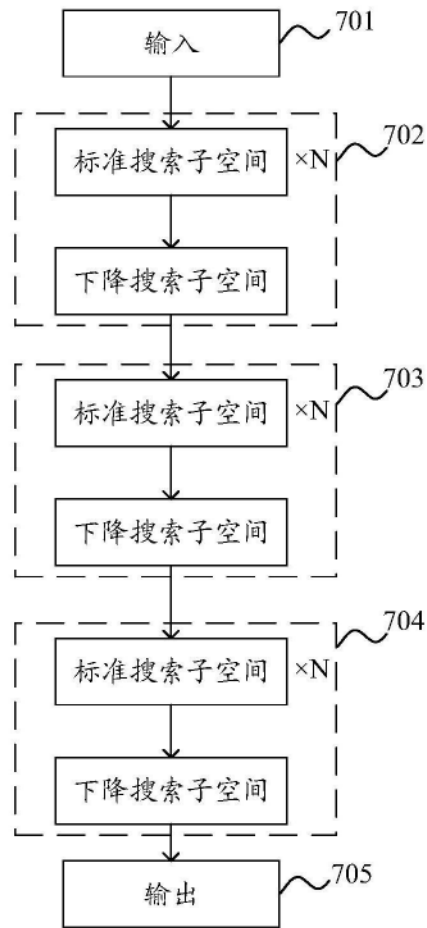


图7

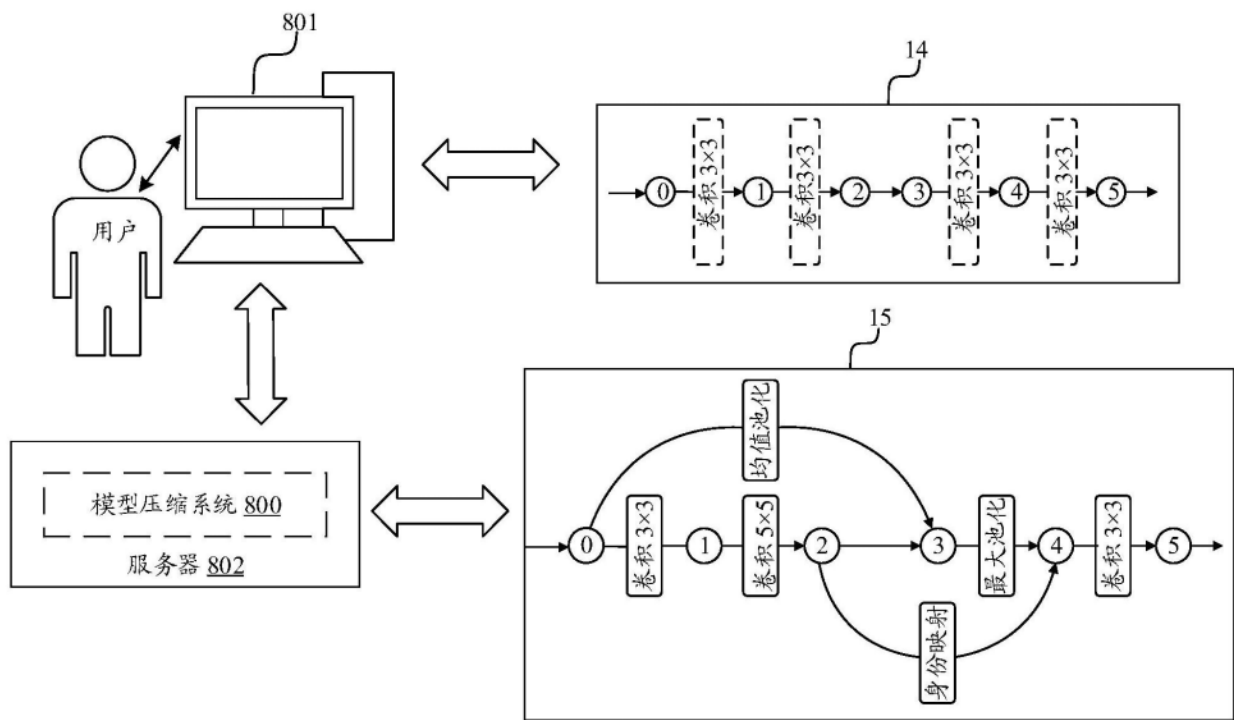


图8

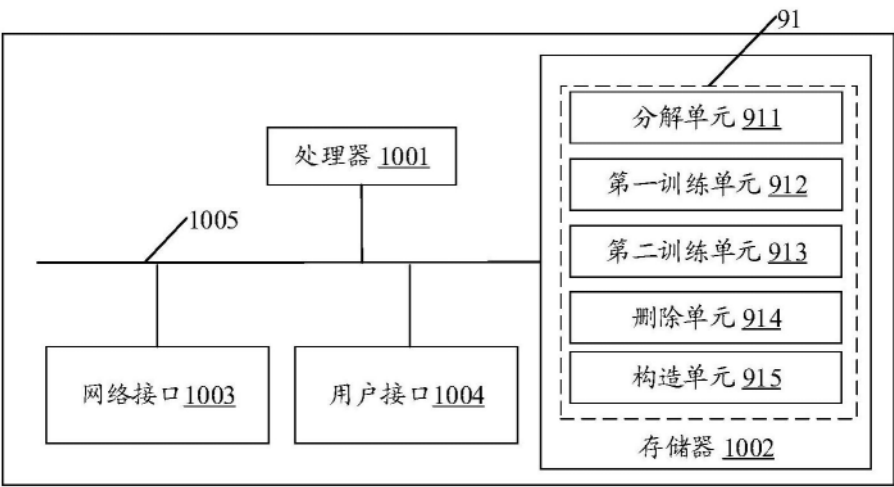


图9

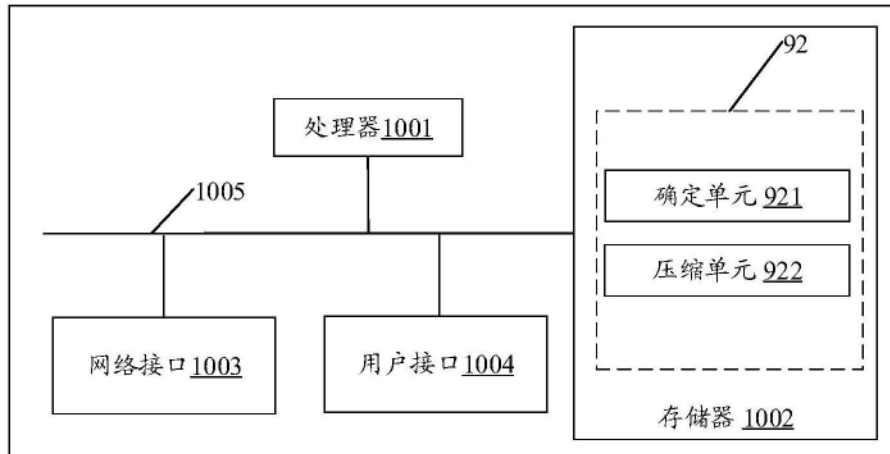


图10