

# (12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织  
国际局

(43) 国际公布日  
2023 年 8 月 31 日 (31.08.2023)



(10) 国际公布号  
**WO 2023/159751 A1**

(51) 国际专利分类号:  
**G06N 3/04** (2006.01)

(21) 国际申请号: PCT/CN2022/090010

(22) 国际申请日: 2022 年 4 月 28 日 (28.04.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:  
202210162485.9 2022 年 2 月 22 日 (22.02.2022) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];  
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518048 (CN)。

(72) 发明人: 王晓锐(WANG, Xiaorui); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518048 (CN)。  
郑强(ZHENG, Qiang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518048 (CN)。

(74) 代理人: 北京辰权知识产权代理有限公司(BEIJING CHENQUAN INTELLECTUAL PROPERTY LAW FIRM); 中国北京市海淀区中关村东路 66 号世纪科贸大厦 C 座 2601 室, Beijing 100190 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,

(54) Title: MODEL PRUNING METHOD AND APPARATUS, COMPUTING DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 模型剪枝方法、装置、计算设备及存储介质

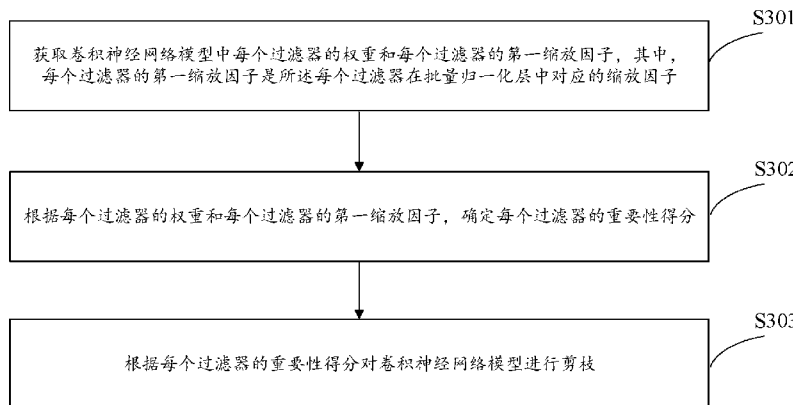


图 3

- S301 Obtain a weight of each filter in a convolutional neural network model and a first scaling factor of each filter, wherein the first scaling factor of each filter is a scaling factor corresponding to each filter in a batch normalization layer
- S302 Determine an importance score of each filter according to the weight of each filter and the first scaling factor of each filter
- S303 Prune the convolutional neural network model according to the importance score of each filter

(57) Abstract: The present application relates to the technical field of artificial intelligence. Disclosed are a model pruning method and apparatus, a computing device, and a storage medium. The method comprises: first, obtaining a weight of each filter in a convolutional neural network model and a first scaling factor of each filter, wherein the first scaling factor of each filter is a scaling factor corresponding to each filter in a batch normalization layer, and a convolutional layer where each filter is located and the batch normalization layer are adjacent network layers; then determining an importance score of each filter according to the weight of each filter and the first scaling factor of each filter; and finally, pruning the convolutional neural network model according to the importance score of each filter. According to the embodiments of the present application, relatively unimportant filters in a convolutional neural network model



GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

can be accurately selected, and the precision loss after model pruning is very small.

(57) 摘要: 本申请涉及人工智能技术领域, 公开了一种模型剪枝方法、装置、计算设备及存储介质。该方法包括: 首先, 获取卷积神经网络模型中每个过滤器的权重和每个过滤器的第一缩放因子, 其中, 每个过滤器的第一缩放因子是每个过滤器在批量归一化层中对应的缩放因子, 每个过滤器所在的卷积层和批量归一化层是相邻的网络层; 然后, 根据每个过滤器的权重和每个过滤器的第一缩放因子, 确定每个过滤器的重要性得分; 最后, 根据每个过滤器的重要性得分对卷积神经网络模型进行剪枝。本申请实施例能够准确地选出卷积神经网络模型中相对不重要的过滤器, 模型剪枝后的精度损失很小。

## 模型剪枝方法、装置、计算设备及存储介质

## 优先权申明

本申请要求于 2022 年 2 月 22 日提交中国专利局、申请号为 202210162485.9, 发明名称为“模型剪枝方法、装置、计算设备及存储介质”的中国专利申请的优先权, 其全部内容通过引用结合在本申请中。

## 技术领域

本申请涉及人工智能技术领域, 尤其涉及一种模型剪枝方法、装置、计算设备及存储介质。

## 背景技术

目前, 基于卷积神经网络 (convolutional neural network, CNN) 的模型在很多任务中都表现出了良好的性能。但是这些模型中常常含有大量的冗余信息, 需要耗费大量的计算开销, 因此, 模型压缩成为了必不可少的一步。

模型剪枝是一种常用的模型压缩方法, 过滤器 (filter) 剪枝又是用于卷积神经网络模型的一种结构化剪枝类型, 过滤器剪枝的操作过程主要可以分为两步: 首先, 选出模型中相对不重要的 filter 并剪除, 得到剪枝后的模型; 然后, 对剪枝后的模型进行微调 (fine-tuning), 以恢复模型精度。如何选出相对不重要的 filter, 成为了过滤器剪枝中的关键技术问题。

然而, 发明人意识到在一些现有的模型剪枝方法中, 所采用的 filter 重要性的衡量标准还不够好, 导致模型剪枝后的精度损失过大。

## 发明内容

为了解决现有技术中的上述问题, 本申请实施例中提供一种模型剪枝方法、装置、计算设备及存储介质, 能够更准确地选出模型中相对不重要的 filter, 使得模型剪枝后的精度损失很小。

第一方面, 本申请提供了一种模型剪枝方法, 该方法包括: 获取卷积神经网络模型中每个过滤器的权重和每个过滤器的第一缩放因子, 其中, 每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子, 所述每个过滤器所在的卷积层和所述批量归一化层是相邻的网络层; 根据每个过滤器的权重和每个过滤器的第一缩放因子, 确定每个过滤器的重要性得分; 根据每个过滤器的重要性得分对卷积神经网络模型进行剪枝。

第二方面, 本申请提供了一种模型剪枝装置, 该装置包括: 获取模块, 用于获取卷积神经网络模型中每个过滤器的权重和每个过滤器的第一缩放因子, 其中, 每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子, 所述每个过滤器所在的卷积层和所述批量归一化层是相邻的网络层; 确定模块, 用于根据每个过滤器的权重和每个过滤器的第一缩放因子, 确定每个过滤器的重要性得分; 处理模块, 用于根据每个过滤器的重要性得分对卷积神经网络模型进行剪枝。

第三方面, 本申请实施例提供一种计算机可读存储介质, 其中, 包括程序或指令, 当所述程序或指令在计算机上运行时, 使得计算机执行以下步骤:

获取卷积神经网络模型中每个过滤器的权重和所述每个过滤器的第一缩放因子, 其中, 所述每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子, 所述每个过滤器所在的卷积层和所述批量归一化层是相邻的网络层;

根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子, 确定所述每个过滤器的重要性得分;

根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝。

第四方面，本申请实施例提供了一种计算设备，其中，包括处理器和存储器；

所述存储器，用于存储计算机程序；

所述处理器，用于执行所述存储器中存储的计算机程序，以使得所述设备执行以下步骤的指令：

获取卷积神经网络模型中每个过滤器的权重和所述每个过滤器的第一缩放因子，其中，所述每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子，所述每个过滤器所在的卷积层和所述批量归一化层是相邻的网络层；

根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子，确定所述每个过滤器的重要性得分；

根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝。

综上所述，本申请实施例通过获取卷积神经网络模型中每个过滤器的权重和每个过滤器的缩放因子，能够确定该卷积神经网络中每个过滤器的重要性得分。其中，重要性得分用于衡量卷积神经网络模型中的各个过滤器的重要性，过滤器的重要性得分越高，则说明该过滤器在卷积神经网络模型中的重要性越高，过滤器的重要性得分越低，则说明该过滤器在卷积神经网络模型中的重要性越低。因此，根据重要性得分便可以准确地选出卷积神经网络模型中相对不太重要的过滤器，进而将重要性较低的这些过滤器给剪除，既实现了模型压缩的目的，减少模型中的冗余信息，又使得模型剪枝后的精度损失较小，可以很好地保留原始模型的精度和性能。

## 附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例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申请的一些实施例，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。

图1是本申请实施例提供的一种卷积神经网络模型中的两个相邻卷积层的示意图；

图2是本申请实施例提供的一种过滤器与卷积核的关系示意图；

图3是本申请实施例提供的一种模型剪枝方法的流程示意图；

图4是本申请实施例提供的一种剪除部分过滤器之后的示意图；

图5是本申请实施例提供的一种模型剪枝装置的结构示意图；

图6是本申请实施例提供的一种计算设备的结构示意图。

## 具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述。显然，下面所描述的实施例仅仅是本申请的一部分实施例，而不是全部的实施例。基于本申请中的实施例，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他实施例，都属于本申请保护的范围。

需要说明的是，在本申请实施例中使用的术语仅仅出于描述特定实施例的目的，而非旨在限制本申请。在本申请实施例和权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指包含一个或多个相关联的列出项目的任意或所有可能组合。

为了便于理解本申请实施例中的技术方案，下面对本申请实施例中涉及的部分术语及概念进行简单介绍。

1、卷积层(convolutional layer)、过滤器(filter)及卷积核(kernel)：

卷积层是构成卷积神经网络的基本结构，用于特征提取。

过滤器是卷积层的重要组成部分，一个卷积层中可以包含一个或者多个过滤器，不同过滤器可以分别探测不同维度的特征，比如形状、颜色、对比度等等。

卷积核是过滤器的重要组成部分，一个过滤器可以由一个或者多个不同的卷积核来构成。

需要说明的是，卷积神经网络可以包括一个或多个卷积层，每个卷积层都有对应的输入通道(input channel)和输出通道(output channel)。为了便于描述，本申请实施例用  $Cin(m,k)$  表示第  $m$  个卷积层对应的第  $k$  个输入通道，用  $Cout(m,k)$  表示第  $m$  个卷积层对应的第  $k$  个输出通道，其中， $m$ 、 $k$  均为正整数。每个卷积层都包含一个或多个 filter，同一个卷积层中的不同 filter 分别产生本卷积层的一个输出通道的数据，换句话说，卷积层的输出通道数目由该卷积层中 filter 的个数所决定，filter 与输出通道一一对应。每个 filter 又包含一个或多个卷积核，同一个 filter 中的不同卷积核分别对应本卷积层的一个输入通道，卷积核只负责与对应的输入通道的数据进行卷积。

举例来说，图 1 是本申请实施例提供的一种卷积神经网络模型中的两个相邻卷积层的示意图，这两个卷积层分别是卷积层  $m$  和卷积层  $m+1$ ，依次代表的是该模型中的第  $m$  个卷积层和第  $m+1$  个卷积层， $m$  为正整数。应理解，卷积神经网络模型中包含多个卷积层，各卷积层之间具有先后顺序，卷积层  $m$  是卷积层  $m+1$  的上一个卷积层，卷积层  $m+1$  是卷积层  $m$  的下一个卷积层，也可以说卷积层  $m+1$  是卷积层  $m$  的下一个相邻卷积层。

如图 1 所示，卷积层  $m$  有对应的 3 个输入通道，分别用  $Cin(m,1)$ 、 $Cin(m,2)$  和  $Cin(m,3)$  来表示，卷积层  $m$  还有对应的 4 个输出通道，分别用  $Cout(m,1)$ 、 $Cout(m,2)$ 、 $Cout(m,3)$  和  $Cout(m,4)$  来表示。这里假设批尺寸(batch-size)等于 1，特征图 A (一个 3 通道特征图，比如是 RGB 三通道的图像) 作为卷积层  $m$  的输入，包含了上述 3 个输入通道的数据，其中，特征图 A 中的以左斜线填充的方块表示输入通道  $Cin(m,1)$  的数据，点填充的方块表示输入通道  $Cin(m,2)$  的数据，竖线填充的方块表示输入通道  $Cin(m,3)$  的数据。卷积层  $m$  中包含了 4 个过滤器，分别用 filter1、filter2、filter3 和 filter4 来表示，上述每一个过滤器都包含与卷积层  $m$  的输入通道同等数目的卷积核，同一个过滤器中的各卷积核分别对应卷积层  $m$  的一个输入通道，分别负责与对应输入通道的数据进行卷积。

以上述 filter1 为例，如图 2 所示，filter1 包括三个卷积核，分别用卷积核 1、卷积核 2、卷积核 3 来表示，它们依次对应卷积层  $m$  的输入通道  $Cin(m,1)$ 、 $Cin(m,2)$  和  $Cin(m,3)$ 。因此，卷积核 1 负责与输入通道  $Cin(m,1)$  的数据进行卷积，卷积核 2 负责与  $Cin(m,2)$  的数据进行卷积，卷积核 3 负责与输入通道  $Cin(m,1)$  的数据进行卷积。需要特别说明的是，本申请实施例中所述的过滤器与卷积核并非是一个概念，如图 2 所示，卷积核是由权重组成的二维数组，而过滤器是由多个卷积核堆叠形成的三维架构，二者处在不同的层次结果当中，具有包含关系。

如图 1 所示，特征图 A 输入卷积层  $m$  之后，分别与卷积层  $m$  中的各个 filter 进行卷积，每个 filter 都会产生一个输出通道的特征图(feature map)，即单通道特征图，或者说是单个通道的数据。其中，filter1 对特征图 A 进行卷积得到的输出结果为 Map1，这里用 Map1 的颜色(白色)表示它对应的是输出通道  $Cout(m,1)$ ；filter2 与特征图 A 卷积得到 Map2，用浅灰色表示 Map2 对应的是输出通道  $Cout(m,2)$ ；filter3 与特征图 A 卷积得到 Map3，用深灰色表示 Map3 对应的是输出通道  $Cout(m,3)$ ；filter4 与特征图 A 卷积得到 Map4，用黑色表示 Map4 对应的是输出通道  $Cout(m,4)$ 。

## 2、批量归一化(batch normalization, BN)层和激励层(activating layer):

BN 层通常用于对数据进行归一化处理以改善数据分布，具有减轻过拟合、改善梯度传播、容许较高的学习率、提高训练和收敛速度等优点。BN 层中有两种可学习参数，分别是缩放因子  $\gamma$  和偏差  $\beta$ 。

激励层通常用于增加神经网络模型的非线性，将特征映射到高维的非线性区间进行解释，以解决线性模型所不能解决的问题。激励层中常用的激励函数有 sigmoid、tanh、ReLU、Leaky Relu、ELU、Maxout 等等，本申请不具体限定。

需要说明的是，卷积层+BN 层+激励层的结构是卷积神经网络模型中的一种常用结构。针对卷积层的每一个输出通道，在该卷积层的相关 BN 层中都有一组对应的  $\gamma$  和  $\beta$ ， $\gamma$  和  $\beta$  分

别用于缩放和平移对应通道的数据，而卷积层中的每一个过滤器分别产生一个输出通道的数据，因此，卷积层中的过滤器与相关 BN 层中的缩放因子具有一一对应关系。卷积层的输出结果经过 BN 层的处理之后，激励层可以进一步把 BN 层输出的结果做非线性映射，以增加非线性。

承接上例，如图 1 所示，卷积层 m 有一个对应的 BN 层和一个对应的激励层，分别用 BN 层 m 和激励层 m 来表示。可以理解的是，卷积层 m、BN 层 m 和激励层 m 便组成了卷积神经网络中常用的卷积层+BN 层+激励层的结构，可以将这三个网络层看为一组，同一组内的三个网络层是相关的网络层。其中，BN 层 m 中的缩放因子  $\gamma 1$ 、 $\gamma 2$ 、 $\gamma 3$  和  $\gamma 4$  依次对应卷积层 m 的输出通道 Cout (m,1)、Cout (m,2)、Cout (m,3) 和 Cout (m,4)，负责缩放对应的输出通道的数据；BN 层 m 中的参数  $\beta 1$ 、 $\beta 2$ 、 $\beta 3$ 、 $\beta 4$  与卷积层 m 的各输出通道也是一一对应的，负责平移对应输出通道的数据。需要说明的是，图中未画出 BN 层 m 中的这些  $\beta$  参数。

如图 1 所示，卷积层 m 中的每个 filter 分别与特征图 A 进行卷积计算得到的输出结果都要经过 BN 层 m 的处理，即卷积层 m 的输出结果 Map1、Map2、Map3 和 Map4 会输入到 BN 层 m 进行处理。可以理解的是，因为 Map1 对应的是输出通道 Cout (m,1)，所以 BN 层 m 会采用  $\gamma 1$  对 Map1 的数据进行缩放并且用  $\beta 1$  进行平移，从而得到 Map1 ( $\gamma 1$ )。同理，BN 层 m 分别对 Map2、Map3、Map4 进行处理后得到 Map2 ( $\gamma 2$ )、Map3 ( $\gamma 3$ ) 和 Map4 ( $\gamma 4$ )。

然后，BN 层 m 将上述结果输入到激励层 m 中，激励层 m 分别对它们做非线性映射得到 Map1\* ( $\gamma 1$ )、Map2\* ( $\gamma 2$ )、Map3\* ( $\gamma 3$ ) 和 Map4\* ( $\gamma 4$ )，上述结果便组成了特征图 B (一个 4 通道特征图)。显然，特征图 B 包含了 4 个通道的数据，它将作为卷积层 m+1 的输入，所以卷积层 m+1 需要有对应的四个输入通道，即 Cin(m+1,1)、Cin (m+1,2)、Cin (m+1,3) 和 Cin (m+1, 4)。因此，卷积层 m+1 的输入通道的数目等于卷积层 m 中的过滤器的个数，即本卷积层的输入通道的数目由上一个相邻卷积层中的过滤器的个数所决定，上一卷积层的过滤器与本卷积层的输入通道具有一一对应关系。需要说明的是，卷积层 m+1 也有对应的 BN 层和激励层，但图 1 没有画出这部分。

应理解，输入通道和输出通道中的“输入”和“输出”实际是相对于某个卷积层来说的，都可以直接称为通道，所以说，卷积层 m 的输出通道 Cout (m,1)、Cout (m,2)、Cout (m,3)、Cout (m,4) 与卷积层 m+1 的输入通道 Cin(m+1,1)、Cin (m+1,2)、Cin (m+1,3)、Cin (m+1, 4) 其实是一一对应的，对应的输入通道和输出通道其实指的是相同类别的通道。比如，卷积层 m 中的 Cout (m,1) 指是白色代表的这个通道类型，而卷积层 m+1 中的 Cin(m+1,1)也是指白色这个通道类型，所以说 Cout (m,1) 和 Cin(m+1,1)对应的其实是相同类型的通道。

基于上述内容，下面介绍本申请中的模型剪枝方法的实施例。

请参阅图 3，图 3 是本申请实施例提供的一种模型剪枝方法的流程示意图，该方法包括以下步骤：

S301、获取卷积神经网络模型中每个过滤器的权重和每个过滤器的第一缩放因子，其中，每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子。

其中，上述卷积神经网络模型可以是任意一个基于卷积神经网络的模型，本申请不具体限定。比如，卷积神经网络模型可以是用于图像识别的卷积神经网络模型，也可以是用于语音识别的卷积神经网络模型，还可以是用于语义理解的卷积神经网络模型。

由前述内容可知，每个过滤器所在的卷积层都可以有一个相邻的 BN 层（即相邻的两个网络层，该卷积层的输出是该 BN 层的输入），每个过滤器在其相邻的 BN 层中都会有一个对应的缩放因子，即某一个具体的  $\gamma$  参数，应理解， $\gamma$  参数是该卷积神经网络模型在之前的模型训练过程中学习得到的。某个过滤器的第一缩放因子，指的就是该过滤器在 BN 层中对应的那个缩放因子。

在可能的实施例中，步骤 S301 还可以是：获取卷积神经网络模型中多个过滤器的权重和

多个过滤器的第一缩放因子，多个过滤器中的每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子。需要说明的是，上述多个过滤器可以是卷积神经网络模型中的所有过滤器，也可以是卷积神经网络模型中的部分过滤器，也就是说，可以只对卷积神经网络模型中的部分过滤器的重要性进行评估，而不需要对卷积神经网络模型中的所有过滤器的重要性进行评估。

还需要说明的是，本申请实施例中所指的“权重”，指的是一个或多个权重值，多个权重值可以用权重向量或权重矩阵来表示。比如，某个  $3 \times 3$  的卷积核的权重指的是该卷积核中的所有权重值，即 9 个权重值，该卷积核中的所有权重值可以用矩阵（即二维数组）来表示，称为该卷积核的权重矩阵，也可以用向量来表示，称为该卷积核的权重向量。再如，某个过滤器的权重由构成该过滤器的所有卷积核的权重组成，即过滤器的权重是一组权重矩阵/权重向量，也可以用多维数组来表示，还可以直接理解为组成该过滤器的所有卷积核中的所有权重值。

S302、根据每个过滤器的权重和每个过滤器的第一缩放因子，确定每个过滤器的重要性得分。

其中，重要性得分用于衡量过滤器的重要性。若过滤器的重要性得分越高，则说明该过滤器在卷积神经网络模型中的重要性越高；若过滤器的重要性得分越低，则说明该过滤器在卷积神经网络模型中的重要性越低。因此，根据重要性得分便可以准确地选出卷积神经网络模型中相对不太重要的过滤器，进而可以将重要性较低的这些过滤器给剪除，避免将模型中较为重要的过滤器给剪除，能够降低对模型精度的影响。

在一种可能的实施例中，所述根据每个过滤器的权重和每个过滤器的第一缩放因子，确定每个过滤器的重要性得分，可以包括：确定每个过滤器的权重的范数值；根据每个过滤器的权重的范数值和每个过滤器的第一缩放因子，确定每个过滤器的重要性得分。

其中，上述范数值可以是 L1 范数值（或者 m1 范数值），也可以是 L2 范数值（或者 m2 范数值），本申请不做具体限定。应理解，L1 范数指的是向量/矩阵中各个元素的绝对值之和，L2 范数指的是向量/矩阵中各个元素的平方和，然后再求平方根。若无特殊说明，下文皆以 L2 范数为例进行介绍。

具体来说，过滤器的权重，即过滤器自身的权重，显然可以作为过滤器重要性的一个评价指标。本申请实施例通过计算过滤器权重的 L2 范数值，其值越大，在一定程度上表明该过滤器的重要性越高。这里用  $W_{i,j}^m$  表示第 m 个卷积层中的第 i 个 filter 中的第 j 个卷积核的权重向量，用  $L_i^m$  表示第 m 个卷积层中的第 i 个 filter 权重的 L2 范数值，因为过滤器的权重由构成该过滤器的所有卷积核的权重组成，所以  $L_i^m$  可以通过下式（1）进行计算：

$$L_i^m = \sqrt{\sum_j (W_{i,j}^m)^2} \quad (1)$$

其中，i、j、m 均为正整数。本申请实施例将  $L_i^m$  作为第 m 个卷积层中的第 i 个 filter 的第一得分。

由前文的介绍可知，每个卷积层都可以有一个相应的 BN 层，该卷积层的输出作为相应 BN 层的输入，而且该卷积层中的 filter 与相应 BN 层中的缩放因子具有一一对应的关系，BN 层是根据对应的缩放因子将 filter 的输出结果进行缩小或者放大的。因此，相应 BN 层中的缩放因子也可以作为过滤器重要性的一个评价指标，其值越大，在一定程度上表明该过滤器的重要性越高。所以，本申请实施例直接以某个过滤器在 BN 层中对应的缩放因子作为该过滤器的第二得分。

最后，根据某个过滤器的第一得分和第二得分，便可以确定该过滤器的重要性得分。

在可能的实施例中，可以将过滤器的第一得分和第二得分相乘得到该过滤器的重要性得分；也可以对过滤器的第一得分和第二得分进行加权求和得到该过滤器的重要性得分；还可以对过滤器的第一得分和第二得分求平均得到该过滤器的重要性得分，本申请不做具体限定。

若无特殊说明,下文皆以第一得分和第二得分相乘的方式进行介绍。

以 filter1 为例,请参见图 2,卷积层 m 中的 filter1 由卷积核 1、卷积核 2 和卷积核 3 构成,卷积核 1、卷积核 2、卷积核 3 的权重向量依次表示为  $W_{1,1}^m$ 、 $W_{1,2}^m$ 、 $W_{1,3}^m$ ,于是,filter1 权重的 L2 范数值  $L_1^m$  可以通过下式 (2) 进行计算,将计算出的  $L_1^m$  作为 filter1 的第一得分:

$$L_1^m = \sqrt{(W_{1,1}^m)^2 + (W_{1,2}^m)^2 + (W_{1,3}^m)^2} \quad (2)$$

再参见图 1,filter1 位于卷积层 m 中,并且卷积层 m 和 BN 层 m 是相邻的两个网络层,卷积层 m 的输出是 BN 层 m 的输入,卷积层 m 中的每个过滤器在 BN 层中都有一个对应的缩放因子。其中,filter1 在 BN 层 m 中对应的缩放因子为  $\gamma_1$ , $\gamma_1$  是该卷积神经网络模型在之前的训练过程中学习得到的,这里直接以  $\gamma_1$  作为 filter1 的第二得分。

最后,将 filter1 的第一得分和第二得分相乘,便可以得到 filter1 的重要性得分  $S_1^m$ , $S_1^m$  的计算公式 (3) 如下:

$$S_1^m = L_1^m * \gamma_1 \quad (3)$$

同理,可以计算出卷积神经网络模型中的每一个过滤器 (或者只计算部分过滤器) 的重要性得分。

在一种可能的实施例中,上一个步骤 S301 还可以包括:获取每个过滤器的权重信息;于是,步骤 S302 还可以是:根据每个过滤器的权重、每个过滤器对应的缩放因子和每个过滤器的权重信息,确定每个过滤器的重要性得分。

其中,所述每个过滤器所在的卷积层包括多个第一卷积核;所述每个过滤器所在的卷积层的下一个相邻卷积层包括多个第二卷积核;所述每个过滤器的权重信息包括所述每个过滤器对应的所述第二卷积核的权重。

上述范数值可以是 L1 范数值,也可以是 L2 范数值。可以理解的是,如果采用两类范数值,对于过滤器重要性的衡量会更加全面,比如,可以将某个过滤器权重的 L1 范数值和该过滤器权重的 L2 范数值相乘或者加权求和,以作为该过滤器的权重的范数值。若无特殊说明,下文皆以 L2 范数为例进行介绍。

在可能的实施例中,每个过滤器对应的所述第二卷积核所对应的通道和所述每个过滤器对应的通道相同。也就是说,每个过滤器与其在下一个卷积层中对应的第二个卷积核所具有的“对应关系”,指的是它们对应的通道相同。

例如,如图 1 所示,卷积层 m 中包括 4 个过滤器,它们中的每个过滤器都包含多个卷积核,卷积层 m+1 中包括 5 个过滤器,它们中的每个过滤器也都包含多个卷积核。为了便于描述以及区分两个相邻卷积层中的卷积核,在卷积层 m 和卷积层 m+1 这两个相邻的卷积层中,将卷积层 m 包括的卷积核都称为第一卷积核,即卷积层 m 中的每一个过滤器所包含的卷积核都称为第一卷积核,并且将卷积层 m+1 包括的卷积核都称为第二卷积核,即卷积层 m+1 中的每一个过滤器所包含的卷积核都称为第二卷积核。可以理解的是,卷积层 m 中的每个过滤器在卷积层 m+1 中都有对应的第二卷积核,即卷积层 m 中的每个过滤器分别与卷积层 m+1 中的部分第二卷积核对应。比如,卷积层 m 中的 filter1 与卷积层 m+1 的输入通道  $C_{in}(m+1,1)$  具有对应关系,卷积层 m+1 中的每个过滤器中的以白色小方块所表示的卷积核对应的也是卷积层 m+1 的输入通道  $C_{in}(m+1,1)$ 。因为 filter1 和卷积层 m+1 中的每个过滤器中的以白色小方块表示的卷积核对应的都是  $C_{in}(m+1,1)$ ,它们对应相同的通道,所以 filter1 和卷积层 m+1 中的每个过滤器中的以白色小方块表示的卷积核也具有对应关系。于是,卷积层 m 中的 filter1 对应的第二卷积核,即卷积层 m 中的 filter1 在卷积层 m+1 中对应的第二卷积核,指的就是卷积层 m+1 中的以白色小方块表示的这些卷积核。

具体来说,由前文的介绍可知,某个过滤器与其下一个相邻卷积层的一个输入通道具有对应关系,并且卷积层中的卷积核与该卷积层的输入通道也具有对应关系,于是,某个过滤器与其下一个相邻卷积层中的部分卷积核 (与该过滤器对应相同输入通道的卷积核) 也具有对应关系。因此,某个过滤器在下一个卷积层中对应的那些卷积核的权重,也可以作为该过

滤波器重要性的一个评价指标。

本申请实施例将某个过滤器在下一个卷积层中对应的卷积核(即上文所述的第二卷积核)的权重称为该过滤器的相关权重,通过计算过滤器的相关权重的 L2 范数值,其值越大,在一定程度上表明该过滤器的重要性越高。这里用  $L_i^{m,m+1}$  表示第  $m$  个卷积层中的第  $i$  个 filter 的相关权重的 L2 范数值,并且第  $m$  个卷积层中的第  $i$  个 filter 过滤器的相关权重由该过滤器在第  $m+1$  个卷积层中对应的(第二)卷积核的权重组成,于是  $L_i^{m,m+1}$  可以通过下式(4)进行计算:

$$L_i^{m,m+1} = \sqrt{\sum_j (W_{j,i}^{m+1})^2} \quad (4)$$

其中,  $W_{j,i}^{m+1}$  表示卷积层  $m+1$  中的第  $j$  个过滤器中的第  $i$  个卷积核的权重向量。可以看出,第  $m$  个卷积层中的第  $i$  个过滤器的相关权重,是由卷积层  $m+1$  中的所有过滤器中的第  $i$  个卷积核的权重向量组成的。将上面计算出的  $L_i^{m,m+1}$  作为第  $m$  个卷积层中的第  $i$  个过滤器的第三得分。

关于过滤器的第一得分和第二得分的计算,与前文方法一致,这里不重复介绍。

最后,根据某个过滤器的第一得分、第二得分和第三得分,便可以确定该过滤器的重要性得分。

在可能的实施例中,可以根据某个过滤器的第一得分、第二得分和第三得分的乘积,确定该过滤器的重要性得分;也可以对某个过滤器的第一得分、第二得分和第三得分进行加权求和得到该过滤器的重要性得分;还可以对某个过滤器的第一得分、第二得分和第三得分求平均得到该过滤器的重要性得分,本申请不做具体限定。若无特殊说明,下文皆以第一得分、第二得分和第三得分相乘得到重要性得分的方式进行介绍。

举例来说,如图 1 所示,卷积层  $m+1$  中包含了 5 个 filter,分别用 filter5、filter6、filter7、filter8 和 filter9 来表示,它们依次代表卷积层  $m+1$  中的第 1 至第 5 个过滤器。上述 5 个过滤器中的每一个过滤器都包含 4 个卷积核,同一过滤器中的不同卷积核分别对应卷积层  $m+1$  的一个输入通道。并且,卷积层  $m$  中的各个过滤器与卷积层  $m+1$  的输入通道存在一一对应关系。因此,卷积层  $m$  中的过滤器与卷积层  $m+1$  中的卷积核也具有对应关系。

比如,卷积层  $m$  中的 filter1 对应卷积层  $m+1$  的输入通道  $Cin(m+1,1)$ ,卷积层  $m+1$  中的各个过滤器(filter5 至 filter9)中的白色小方块所代表的卷积核,对应的也是  $Cin(m+1,1)$ ,于是,filter1 与卷积层  $m+1$  中的上述卷积核之间也是对应的,上述这些卷积核的权重即为 filter1 的相关权重。假设 filter5 至 filter9 中的白色小方块所代表的卷积核,它们的权重向量依次为  $W_{1,1}^{m+1}$ 、 $W_{2,1}^{m+1}$ 、 $W_{3,1}^{m+1}$ 、 $W_{4,1}^{m+1}$ 、 $W_{5,1}^{m+1}$ ,于是 filter1 的相关权重  $L_1^{m,m+1}$  可以通过下式

(5) 进行计算,将计算出的  $L_i^{m,m+1}$  作为 filter1 的第三得分:

$$L_1^{m,m+1} = \sqrt{(W_{1,1}^{m+1})^2 + (W_{2,1}^{m+1})^2 + (W_{3,1}^{m+1})^2 + (W_{4,1}^{m+1})^2 + (W_{5,1}^{m+1})^2} \quad (5)$$

关于 filter1 的第一得分和第二得分的计算方式,与前文一致,这里不重复介绍。

最后,将 filter1 的第一得分、第二得分和第三得分相乘,便可以确定 filter1 的重要性得分  $S_1^m$ ,  $S_1^m$  的计算公式(6)如下:

$$S_1^m = L_1^m * \gamma_1 * L_1^{m,m+1} \quad (6)$$

同理,可以计算出卷积神经网络模型中的每一个过滤器的重要性得分。

需要说明的是,对于卷积神经网络模型中的最后一个卷积层中的卷积核,由于其之后没有卷积层了(没有下一个相邻的卷积层),所以无法用前面的方法计算这些卷积核的第三得分。在一种方案中,可以直接忽略/不考虑最后一个卷积层中的过滤器,不计算这些过滤器的重要性得分,即这些过滤器不参与模型剪枝,不进行过滤器之间重要性的比较。在另一种方案中,由于模型中的最后一个卷积层之后通常还会接着线性层(全连接层),最后一个卷积层中的这些过滤器与后面这个线性层的输入通道也会具有对应关系,因此,可以将最后一个卷积层中的每个过滤器在其之后的线性层中对应的输入通道的权重,作为该过滤器的相关权重,然后

计算过滤器的相关权重的范数值，得到该过滤器的第三得分，进而可以确定最后一个卷积层中的这些过滤器各自的重要性得分。

S303、根据每个过滤器的重要性得分对卷积神经网络模型进行剪枝。

可以理解的是，对于模型的剪枝率，可以由技术人员根据模型的复杂度、应用场景等进行合理选择，本申请不具体限定。

在可能的实施例中，根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝，可以包括：对每个过滤器的重要性得分进行排序，获得排序结果；根据排序结果和剪枝率确定卷积神经网络模型中的目标过滤器；将卷积神经网络模型中的目标过滤器剪除，得到剪枝后的模型。

具体来说，通过前面的步骤 S301 和 S302 可以计算出模型中的每个过滤器各自的重要性得分。然后，对每个过滤器的重要性得分进行排序，可以得到相应的排序结果，比如，可以将按照重要性得分从高到低对卷积神经网络模型中的所有过滤器进行排序。接着，将卷积神经网络模型中重要性得分较低的过滤器确定为该模型中的目标过滤器，将目标过滤器进行剪除便得到剪枝后的模型。

例如，假设某个卷积神经网络模型中总共有 100 个过滤器，模型的剪枝率设定为 5%，即需要剪除该模型中 5% 的过滤器。通过步骤 S301 和 S302 得到该模型中的所有过滤器的重要性得分之后，按照重要性得分从高到低对该模型中的所有过滤器进行排序，得到排序结果，然后再根据剪枝率，将排序最靠后的五个过滤器确定为该模型中的目标过滤器，将该模型中的目标过滤器剪除便可以得到剪枝后的模型，剪枝完成。

需要说明的是，由于本卷积层中的过滤器与本卷积层的输出通道是一一对应，将本卷积层中的部分过滤器剪除之后，会导致本卷积层的输出通道数目减少。而且，由于本卷积层中的过滤器和下一个相邻卷积层中的卷积核也具有对应关系，当剪除本卷积层中的某个过滤器时，会导致该过滤器在下一个相邻卷积层中对应的输入通道也就没有了，下一个卷积层中的每个过滤器中的卷积核数目要相应减少，这些过滤器都需要移除对应该输入通道的卷积核。

例如，如图 4 所示，假设卷积层 m 中的 filter2 被确定为目标过滤器，在 filter2 被剪除之后，卷积层 m 所对应的输出通道相应减少了，即 filter2 在卷积层 m 中对应的输出通道 Cout(m,2) 没有了，卷积层 m 的输出通道变为了三个。因此，当特征图 A 再次输入卷积层 m 时，得到的输出结果中便没有 Cout(m,2) 通道的数据，不会产生 Map2，于是 BN 层 m 中的 Map ( $\gamma_2$ ) 以及激励层 m 中的 Map\* ( $\gamma_2$ ) 也不会有了。所以，特征图 B 从原来的 4 个通道减少为了三个通道，只有三个通道的数据，将这个特征图 B 输入卷积层 m+1 时，卷积层 m+1 所需要的输入通道数目也相应减少了，没有了 Cin (m+1,2)，因此，卷积层 m+1 中的各个 filter 中的以浅灰色所表示的卷积核（这些卷积核对应的是 Cin (m+1,2)）也可以移除，移除的这些过滤器对应的其实就是卷积层 m 中被剪枝的 filter2。

在可能的实施例中，在上述步骤 S303 之后，该方法还可以包括：获取训练数据，通过训练数据对上述剪枝后的模型进行微调得到目标模型。其中，训练数据包括图像数据、语音数据、文本数据中的一种或多种类型，由卷积神经网络模型模型的具体类型/功能来决定，本申请不做具体限定。

可以理解的是，模型剪枝后会有一定的精度损失，这里的微调就是用训练数据对剪枝之后的模型重新进行训练，使其尽量恢复到原始模型（即剪枝之前的模型）的精度水平。关于模型微调，本申请不做过多介绍。

综上所述，本申请实施例通过获取卷积神经网络模型中每个过滤器的权重和每个过滤器的缩放因子，能够确定该卷积神经网络中每个过滤器的重要性得分。其中，重要性得分用于衡量过滤器的重要性，若过滤器的重要性得分越高，则说明该过滤器在卷积神经网络模型中的重要性越高，若过滤器的重要性得分越低，则说明该过滤器在卷积神经网络模型中的重要性越低。本申请实施例还可以获取每个过滤器的权重信息，在每个过滤器的权重和每个过滤器的第一缩放因子的基础上，再结合每个过滤器的权重信息来确定每个过滤器的重要性得分。

可以看出,本申请实施例可以以 filter 自身的权重、filter 在 BN 层中对应的缩放因子(即第一缩放因子)以及 filter 的相关权重作为 filter 重要性的评价指标,根据这三种评价指标对卷积神经网络模型中的过滤器的重要性进行评估,从而确定出相对不重要的 filter 并剪除。可以理解的是,通过融合以上三种评价指标,对 filter 重要性的衡量更加全面,能够更准确地选出模型中相对不重要的 filter 进行剪枝,既实现了模型压缩的目的,减少模型中的冗余信息,又使得模型剪枝后的精度损失较小,可以很好地保留原始模型的精度和性能。

请参见图 5,图 5 是本申请实施例提供的一种模型剪枝装置 500 的结构示意图,该模型剪枝装置 500 包括获取模块 501、确定模块 502 和处理模块 503。

获取模块 501 用于,获取卷积神经网络模型中每个过滤器的权重和每个过滤器的第一缩放因子,其中,每个过滤器的第一缩放因子是每个过滤器在批量归一化层中对应的缩放因子,每个过滤器所在的卷积层和批量归一化层是相邻的网络层

确定模块 502 用于,根据每个过滤器的权重和每个过滤器的第一缩放因子,确定每个过滤器的重要性得分

处理模块 503 用于,根据每个过滤器的重要性得分对卷积神经网络模型进行剪枝。

在可能的实施例中,获取模块 501 还用于,获取每个过滤器的权重信息,其中,所述每个过滤器所在的卷积层包括多个第一卷积核,所述每个过滤器所在的卷积层的下一个相邻卷积层包括多个第二卷积核,所述每个过滤器的权重信息包括所述每个过滤器对应的所述第二卷积核的权重;确定模块 502 还用于:根据每个过滤器的权重、每个过滤器的第一缩放因子和每个过滤器的权重信息,确定每个过滤器的重要性得分。

在可能的实施例中,所述每个过滤器对应的所述第二卷积核所对应的通道和所述每个过滤器对应的通道相同。

在可能的实施例中,确定模块 502 还用于:确定每个过滤器的权重的范数值和每个过滤器对应的第二卷积核的权重的范数值;根据每个过滤器的权重的范数值、每个过滤器的第一缩放因子和每个过滤器对应的第二卷积核的权重的范数值,确定每个过滤器的重要性得分。

在可能的实施例中,上述范数值是 L1 范数值和 L2 范数值中的一种或多种。

在可能的实施例中,确定模块 502 还用于:根据每个过滤器的权重的范数值、每个过滤器的第一缩放因子和每个过滤器对应的第二卷积核的权重的范数值的乘积,确定每个过滤器的重要性得分。

在可能的实施例中,处理模块 503 还用于:对每个过滤器的重要性得分进行排序,获得排序结果;根据排序结果和剪枝率确定卷积神经网络模型中的目标过滤器;将卷积神经网络模型中的目标过滤器剪除,得到剪枝后的模型。

在可能的实施例中,处理模块 503 还用于:获取训练数据,其中,训练数据包括图像数据、语音数据、文本数据中的一种或多种;通过训练数据对剪枝后的模型进行微调,得到目标模型。

上述模型剪枝装置 500 中的各个模块具体用于实现图 3 中的任一实施例的方法,请参见上文相关介绍,这里不重复。

图 6 是本申请实施例提供的一种计算设备 600 的结构示意图。计算设备 600 可以是笔记本电脑、平板电脑以及服务器等计算设备,本申请不做具体限定。

计算设备 600 包括处理器 601、存储器 602 以及通信接口 603,所述计算设备 600 具体用于实现图 3 的模型剪枝方法中的任一实施例。其中,处理器 601、存储器 602 以及通信接口 603 可以通过内部总线 604 相互连接,也可通过无线传输等其他手段实现通信。本申请实施例以通过总线 604 连接为例,总线 604 可以是外设部件互连标准(peripheral component interconnect, PCI)总线或扩展工业标准结构(extended industry standard architecture, EISA)总线等。总线 604 可以分为地址总线、数据总线、控制总线等。为便于表示,图 6 中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

处理器 601 可以由至少一个通用处理器构成, 例如中央处理器(central processing unit, CPU), 或者 CPU 和硬件芯片的组合。上述硬件芯片可以是专用集成电路(application-specific integrated circuit, ASIC)、可编程逻辑器件(programmable logic device, PLD)或其组合。上述 PLD 可以是复杂可编程逻辑器件(complex programmable logic device, CPLD)、现场可编程逻辑门阵列(field-programmable gate array, FPGA)、通用阵列逻辑(generic array logic, GAL)或其任意组合。处理器 601 执行各种类型的数字存储指令, 例如存储在存储器 602 中的软件或者固件程序, 它能使计算设备 600 提供多种服务。

存储器 602 用于存储程序代码, 并由处理器 601 来控制执行。存储器 602 可以包括易失性存储器(volatile memory), 例如随机存取存储器(random access memory, RAM); 存储器 602 也可以包括非易失性存储器(non-volatile memory), 例如只读存储器(read-only memory, ROM)、快闪存储器(flash memory)、硬盘(hard disk drive, HDD)或固态硬盘(solid-state drive, SSD); 存储器 602 还可以包括上述种类的组合。存储器 602 可以存储有程序代码, 具体可以包括用于执行图 3 的模型剪枝方法的任一实施例的程序代码, 这里不再进行赘述。

通信接口 402 可以为有线接口、无线接口或者内部接口, 用于与其他模块或设备进行通信。有线接口可以是以太接口、局域互连网络(local interconnect network, LIN)等, 无线接口可以是蜂窝网络接口或使用无线局域网接口等, 内部接口可以是 PCIe 接口等等。

需要说明的是, 本实施例可以是通用的物理服务器实现的, 例如, ARM 服务器或者 X86 服务器, 服务器可以是独立的服务器, 也可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network, CDN)以及大数据和人工智能平台等基础云计算服务的云服务器。本实施例也可以是基于通用的物理服务器结合 NFV 技术实现的虚拟机实现的, 虚拟机指通过软件模拟的具有完整硬件系统功能的、运行在一个完全隔离环境中的完整计算机系统, 本申请不作具体限定。应理解, 图 6 所示的计算设备 600 还可以是至少一个服务器构成的服务器集群, 本申请不作具体限定。

还需要说明的, 图 6 仅仅是本申请实施例的一种可能的实现方式, 实际应用中, 计算设备 600 还可以包括更多或更少的部件, 这里不作限制。关于本申请实施例中未出示或未描述的内容, 可参见前述图 3 的模型剪枝方法的实施例中的相关阐述, 这里不再赘述。

本申请实施例中还提供一种计算机可读存储介质, 计算机可读存储介质中存储有指令, 当其在处理器上运行时, 图 3 的任一实施例的方法得以实现。所述计算机可读存储介质可以是非易失性, 也可以是易失性。

本申请实施例中还提供一种计算机程序产品, 当计算机程序产品在处理器上运行时, 图 3 的任一实施例的方法得以实现。

本领域普通技术人员可以理解实现上述方法实施例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, 所述的程序可存储于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, 所述的存储介质可为磁碟、光盘、只读存储记忆体(read-only memory, ROM)或随机存储记忆体(random access memory, RAM)等等。

本申请实施例可以基于人工智能技术对相关的数据进行获取和处理。其中, 人工智能(Artificial Intelligence, AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能, 感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、机器人技术、生物识别技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

以上所揭露的仅为本申请一种较佳实施例而已, 当然不能以此来限定本申请之权利范围,

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流程，并依本申请权利要求所作的等同变化，仍属于发明所涵盖的范围。

## 权 利 要 求 书

1. 一种模型剪枝方法, 其中, 所述方法包括:

获取卷积神经网络模型中每个过滤器的权重和所述每个过滤器的第一缩放因子, 其中, 所述每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子, 所述每个过滤器所在的卷积层和所述批量归一化层是相邻的网络层;

根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子, 确定所述每个过滤器的重要性得分;

根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝。

2. 根据权利要求1所述的方法, 其中, 所述方法还包括: 获取所述每个过滤器的权重信息, 其中, 所述每个过滤器所在的卷积层包括多个第一卷积核, 所述每个过滤器所在的卷积层的下一个相邻卷积层包括多个第二卷积核, 所述每个过滤器的权重信息包括所述每个过滤器对应的所述第二卷积核的权重;

所述根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子, 确定所述每个过滤器的重要性得分, 包括:

根据所述每个过滤器的权重、所述每个过滤器的第一缩放因子和所述每个过滤器的权重信息, 确定所述每个过滤器的重要性得分。

3. 根据权利要求2所述的方法, 其中, 所述每个过滤器对应的所述第二卷积核所对应的通道和所述每个过滤器对应的通道相同。

4. 根据权利要求3所述的方法, 其中, 所述根据所述每个过滤器的权重、所述每个过滤器的第一缩放因子和所述每个过滤器的权重信息, 确定所述每个过滤器的重要性得分, 包括:

确定所述每个过滤器的权重的范数值和所述每个过滤器对应的所述第二卷积核的权重的范数值;

根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值, 确定所述每个过滤器的重要性得分。

5. 根据权利要求4所述的方法, 其中, 所述根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值, 确定所述每个过滤器的重要性得分, 包括:

根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值的乘积, 确定所述每个过滤器的重要性得分。

6. 根据权利要求5所述的方法, 其中, 所述根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝, 包括:

对所述每个过滤器的重要性得分进行排序, 获得排序结果;

根据所述排序结果和剪枝率确定所述卷积神经网络模型中的目标过滤器;

将所述卷积神经网络模型中的所述目标过滤器剪除, 得到剪枝后的模型。

7. 根据权利要求6所述的方法, 其中, 所述方法还包括:

获取训练数据, 所述训练数据包括图像数据、语音数据、文本数据中的一种或多种;

通过所述训练数据对所述剪枝后的模型进行微调, 得到目标模型。

8. 一种模型剪枝装置, 其中, 所述装置包括:

获取模块, 用于获取卷积神经网络模型中每个过滤器的权重和所述每个过滤器的第一缩放因子, 其中, 所述每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子, 所述每个过滤器所在的卷积层和所述批量归一化层是相邻的网络层;

确定模块, 用于根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子, 确定所述每个过滤器的重要性得分;

处理模块, 用于根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝。

9. 一种计算设备, 其中, 包括处理器和存储器;

所述存储器, 用于存储计算机程序;

所述处理器，用于执行所述存储器中存储的计算机程序，以使得所述设备执行以下步骤的指令：

获取卷积神经网络模型中每个过滤器的权重和所述每个过滤器的第一缩放因子，其中，所述每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子，所述每个过滤器所在的卷积层和所述批量归一化层是相邻的网络层；

根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子，确定所述每个过滤器的重要性得分；

根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝。

10. 根据权利要求 9 所述的计算设备，其中，所述步骤还包括：获取所述每个过滤器的权重信息，其中，所述每个过滤器所在的卷积层包括多个第一卷积核，所述每个过滤器所在的卷积层的下一个相邻卷积层包括多个第二卷积核，所述每个过滤器的权重信息包括所述每个过滤器对应的所述第二卷积核的权重；

所述根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子，确定所述每个过滤器的重要性得分，包括：

根据所述每个过滤器的权重、所述每个过滤器的第一缩放因子和所述每个过滤器的权重信息，确定所述每个过滤器的重要性得分。

11. 根据权利要求 10 所述的计算设备，其中，所述每个过滤器对应的所述第二卷积核所对应的通道和所述每个过滤器对应的通道相同。

12. 根据权利要求 11 所述的计算设备，其中，所述根据所述每个过滤器的权重、所述每个过滤器的第一缩放因子和所述每个过滤器的权重信息，确定所述每个过滤器的重要性得分，包括：

确定所述每个过滤器的权重的范数值和所述每个过滤器对应的所述第二卷积核的权重的范数值；

根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值，确定所述每个过滤器的重要性得分。

13. 根据权利要求 12 所述的计算设备，其中，所述根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值，确定所述每个过滤器的重要性得分，包括：

根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值的乘积，确定所述每个过滤器的重要性得分。

14. 根据权利要求 13 所述的计算设备，其中，所述根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝，包括：

对所述每个过滤器的重要性得分进行排序，获得排序结果；

根据所述排序结果和剪枝率确定所述卷积神经网络模型中的目标过滤器；

将所述卷积神经网络模型中的所述目标过滤器剪除，得到剪枝后的模型。

15. 一种计算机可读存储介质，其中，包括程序或指令，当所述程序或指令在计算机上运行时，使得计算机执行以下步骤：

获取卷积神经网络模型中每个过滤器的权重和所述每个过滤器的第一缩放因子，其中，所述每个过滤器的第一缩放因子是所述每个过滤器在批量归一化层中对应的缩放因子，所述每个过滤器所在的卷积层和所述批量归一化层是相邻的网络层；

根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子，确定所述每个过滤器的重要性得分；

根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝。

16. 根据权利要求 15 所述的计算机可读存储介质，其中，所述步骤还包括：获取所述每个过滤器的权重信息，其中，所述每个过滤器所在的卷积层包括多个第一卷积核，所述每个过滤器所在的卷积层的下一个相邻卷积层包括多个第二卷积核，所述每个过滤器的权重信息

包括所述每个过滤器对应的所述第二卷积核的权重；

所述根据所述每个过滤器的权重和所述每个过滤器的第一缩放因子，确定所述每个过滤器的重要性得分，包括：

根据所述每个过滤器的权重、所述每个过滤器的第一缩放因子和所述每个过滤器的权重信息，确定所述每个过滤器的重要性得分。

17. 根据权利要求 16 所述的计算机可读存储介质，其中，所述每个过滤器对应的所述第二卷积核所对应的通道和所述每个过滤器对应的通道相同。

18. 根据权利要求 17 所述的计算机可读存储介质，其中，所述根据所述每个过滤器的权重、所述每个过滤器的第一缩放因子和所述每个过滤器的权重信息，确定所述每个过滤器的重要性得分，包括：

确定所述每个过滤器的权重的范数值和所述每个过滤器对应的所述第二卷积核的权重的范数值；

根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值，确定所述每个过滤器的重要性得分。

19. 根据权利要求 18 所述的计算机可读存储介质，其中，所述根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值，确定所述每个过滤器的重要性得分，包括：

根据所述每个过滤器的权重的范数值、所述每个过滤器的第一缩放因子和所述每个过滤器对应的所述第二卷积核的权重的范数值的乘积，确定所述每个过滤器的重要性得分。

20. 根据权利要求 19 所述的计算机可读存储介质，其中，所述根据所述每个过滤器的重要性得分对所述卷积神经网络模型进行剪枝，包括：

对所述每个过滤器的重要性得分进行排序，获得排序结果；

根据所述排序结果和剪枝率确定所述卷积神经网络模型中的目标过滤器；

将所述卷积神经网络模型中的所述目标过滤器剪除，得到剪枝后的模型。

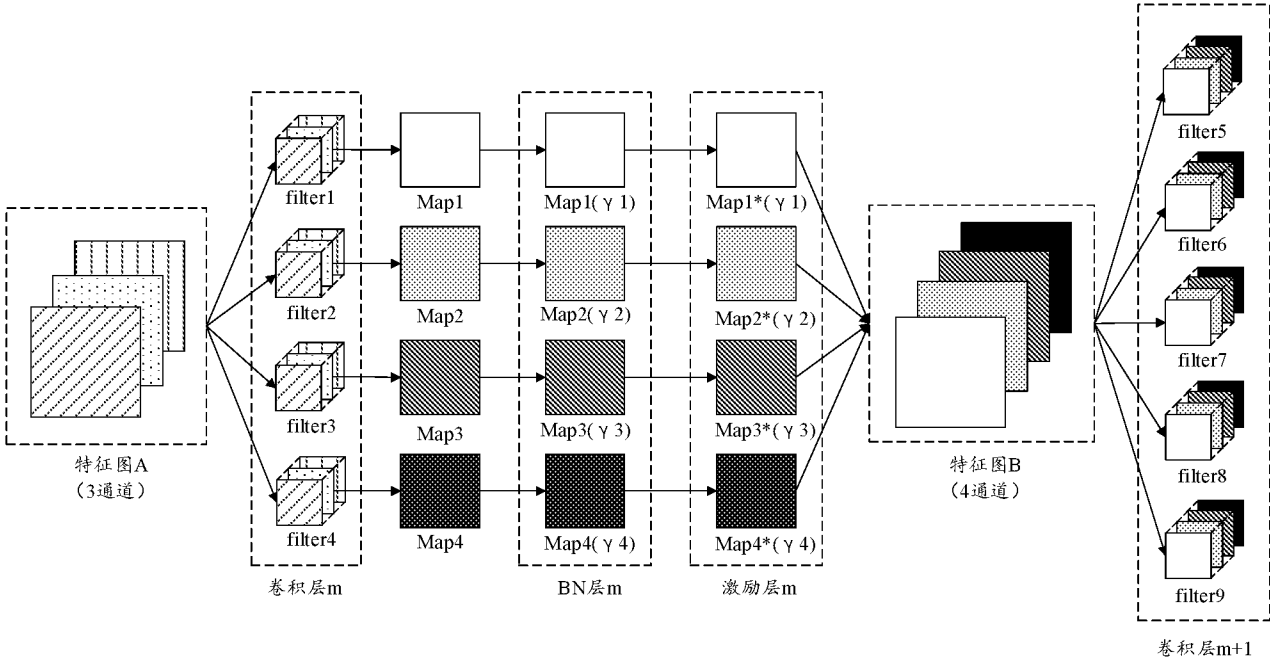


图 1

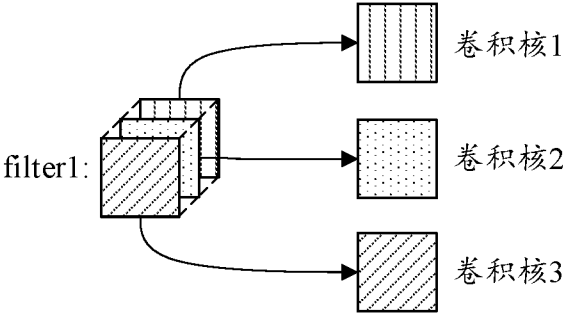


图 2

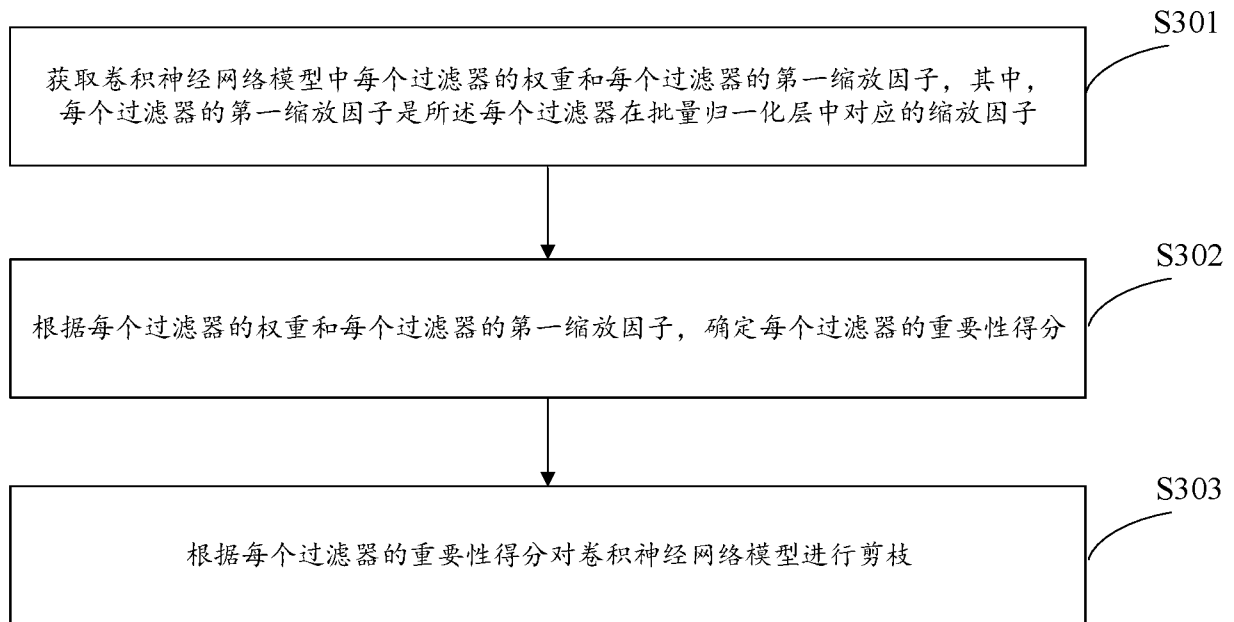


图 3

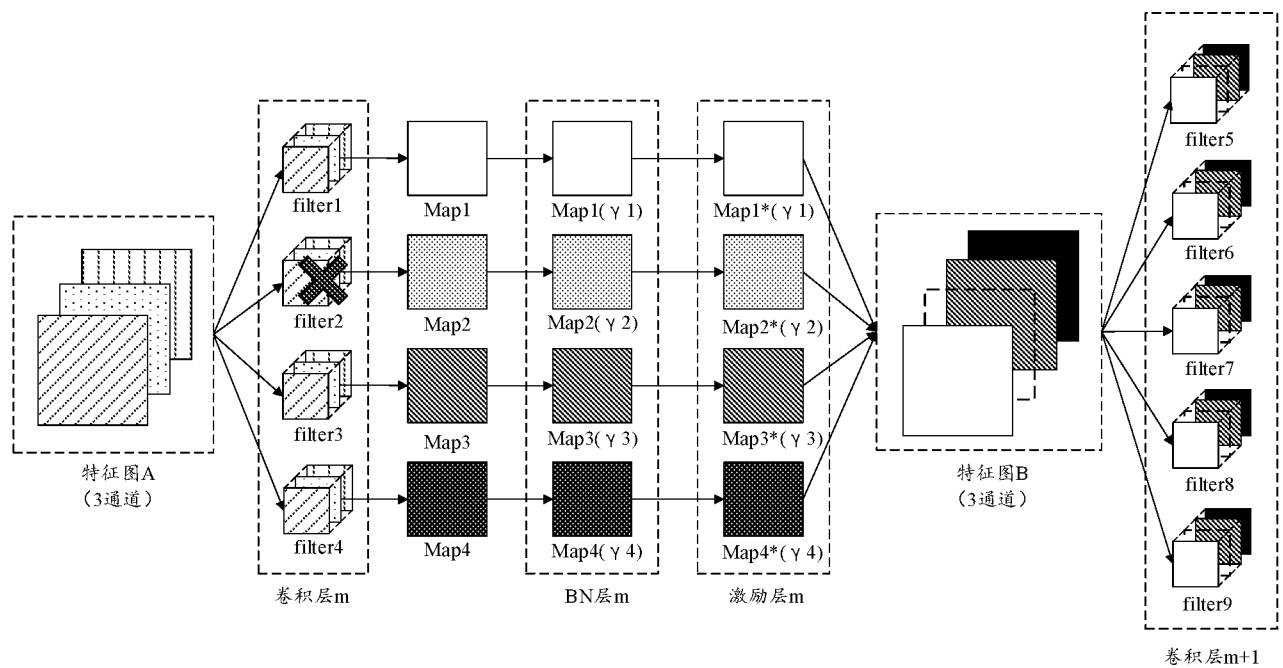


图 4

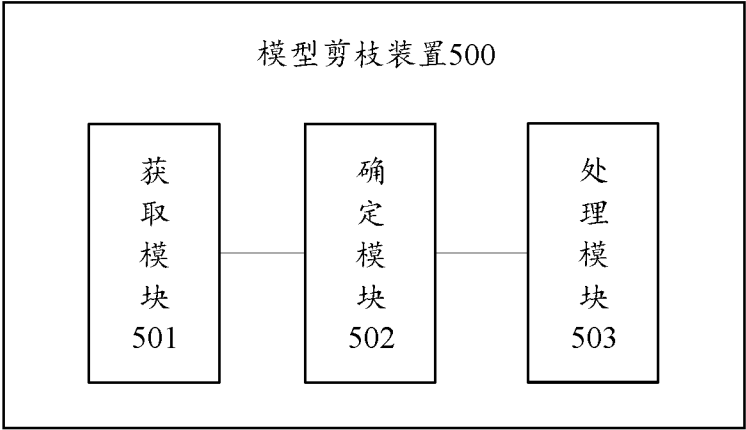


图 5

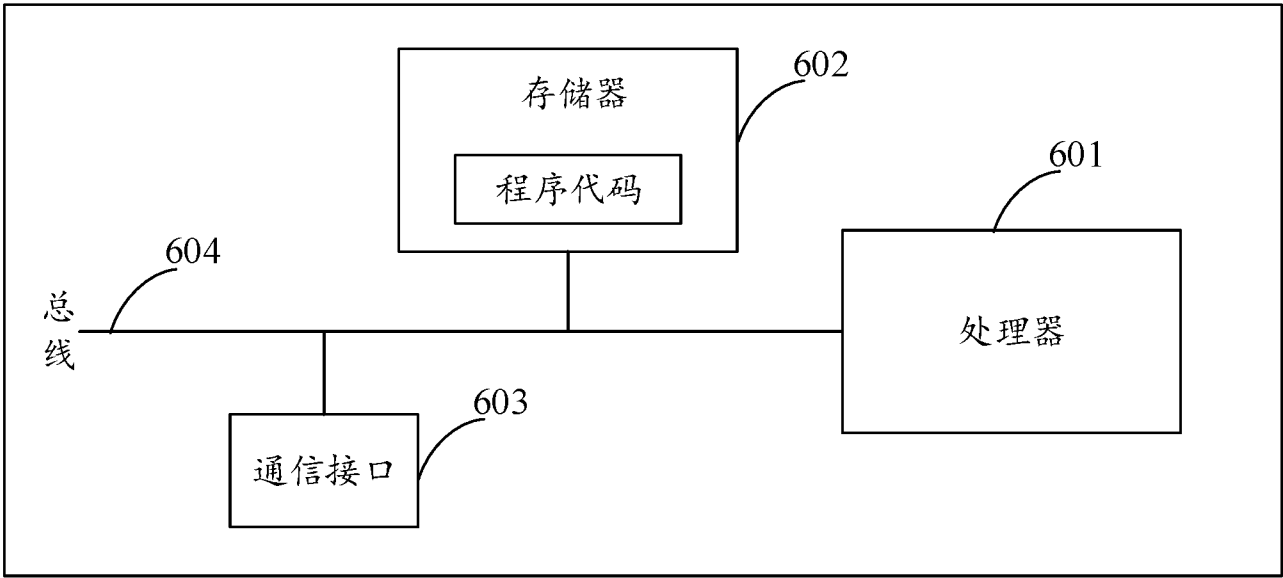


图 6

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/090010

**A. CLASSIFICATION OF SUBJECT MATTER**

G06N 3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; DWPI; VEN; USTXT; WOTXT; EPTXT; CNKI: 模型, 神经网络, 剪枝, 压缩, 卷积, 过滤器, 滤波器, 权重, 缩放, 拉伸, 批量归一化层, 重要性, 评分, 得分, model, neural net, pruning, compression, convolution, filter, weight, zoom, stretch, BN层, importance, significance, grade, score

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                           | Relevant to claim No. |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| PX        | CN 114330713 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 12 April 2022 (2022-04-12)<br>description, paragraphs 58-167                        | 1-20                  |
| X         | CN 112183748 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 05 January 2021 (2021-01-05)<br>description, paragraphs 49-85          | 1-20                  |
| X         | CN 110263841 A (NANJING UNIVERSITY OF INFORMATION SCIENCE & TECHNOLOGY) 20 September 2019 (2019-09-20)<br>description, paragraphs 5-41       | 1-20                  |
| X         | CN 111461322 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 28 July 2020 (2020-07-28)<br>description, paragraphs 119-228 | 1-20                  |
| A         | US 2019294929 A1 (THE REGENTS OF THE UNIVERSITY OF MICHIGAN et al.) 26 September 2019 (2019-09-26)<br>entire document                        | 1-20                  |



Further documents are listed in the continuation of Box C.



See patent family annex.

\* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance  
 “E” earlier application or patent but published on or after the international filing date  
 “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)  
 “O” document referring to an oral disclosure, use, exhibition or other means  
 “P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention  
 “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone  
 “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art  
 “&” document member of the same patent family

Date of the actual completion of the international search

01 November 2022

Date of mailing of the international search report

16 November 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/  
CN)  
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing  
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

**INTERNATIONAL SEARCH REPORT**  
**Information on patent family members**

International application No.

**PCT/CN2022/090010**

| Patent document<br>cited in search report |            |    | Publication date<br>(day/month/year) | Patent family member(s) | Publication date<br>(day/month/year) |
|-------------------------------------------|------------|----|--------------------------------------|-------------------------|--------------------------------------|
| CN                                        | 114330713  | A  | 12 April 2022                        | None                    |                                      |
| CN                                        | 112183748  | A  | 05 January 2021                      | None                    |                                      |
| CN                                        | 110263841  | A  | 20 September 2019                    | None                    |                                      |
| CN                                        | 111461322  | A  | 28 July 2020                         | None                    |                                      |
| US                                        | 2019294929 | A1 | 26 September 2019                    | US 10936913 B2          | 02 March 2021                        |

## 国际检索报告

国际申请号

PCT/CN2022/090010

## A. 主题的分类

G06N 3/04 (2006.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

## B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

CNABS; CNTXT; DWPI; VEN; USTXT; WOTXT; EPTXT; CNKI: 模型, 神经网络, 剪枝, 压缩, 卷积, 过滤器, 滤波器, 权重, 缩放, 拉伸, 批量归一化层, 重要性, 评分, 得分, model, neural net, pruning, compression, convolution, filter, weight, zoom, stretch, BN层, importance, significance, grade, score

## C. 相关文件

| 类 型* | 引用文件, 必要时, 指明相关段落                                                                                | 相关的权利要求 |
|------|--------------------------------------------------------------------------------------------------|---------|
| PX   | CN 114330713 A (平安科技深圳有限公司) 2022年4月12日 (2022 - 04 - 12)<br>说明书第58-167段                           | 1-20    |
| X    | CN 112183748 A (中国科学院自动化研究所) 2021年1月5日 (2021 - 01 - 05)<br>说明书第49-85段                            | 1-20    |
| X    | CN 110263841 A (南京信息工程大学) 2019年9月20日 (2019 - 09 - 20)<br>说明书第5-41段                               | 1-20    |
| X    | CN 111461322 A (中国科学院计算技术研究所) 2020年7月28日 (2020 - 07 - 28)<br>说明书第119-228段                        | 1-20    |
| A    | US 2019294929 A1 (THE REGENTS OF THE UNIVERSITY OF MICHIGAN 等) 2019年9月26日 (2019 - 09 - 26)<br>全文 | 1-20    |

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

\* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&amp;” 同族专利的文件

国际检索实际完成的日期

2022年11月1日

国际检索报告邮寄日期

2022年11月16日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

邢丽超

电话号码 86-(20)-28950382

国际检索报告  
关于同族专利的信息

国际申请号

PCT/CN2022/090010

| 检索报告引用的专利文件 |            |    | 公布日<br>(年/月/日) | 同族专利           | 公布日<br>(年/月/日) |
|-------------|------------|----|----------------|----------------|----------------|
| CN          | 114330713  | A  | 2022年4月12日     | 无              |                |
| CN          | 112183748  | A  | 2021年1月5日      | 无              |                |
| CN          | 110263841  | A  | 2019年9月20日     | 无              |                |
| CN          | 111461322  | A  | 2020年7月28日     | 无              |                |
| US          | 2019294929 | A1 | 2019年9月26日     | US 10936913 B2 | 2021年3月2日      |