(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 109886397 A
(43)申请公布日 2019.06.14

(21)申请号 201910218652.5

(22)申请日 2019.03.21

(71)申请人 西安交通大学
地址 710049 陕西省西安市咸宁西路28号

(72)发明人 梅魁志 张良 张增 薛建儒 鄢健宇 常藩 张向楠 王晓 陶纪安

(74)专利代理机构 西安通大专利代理有限责任公司 61200
代理人 徐文权

(51)Int.Cl.
G06N 3/04(2006.01)

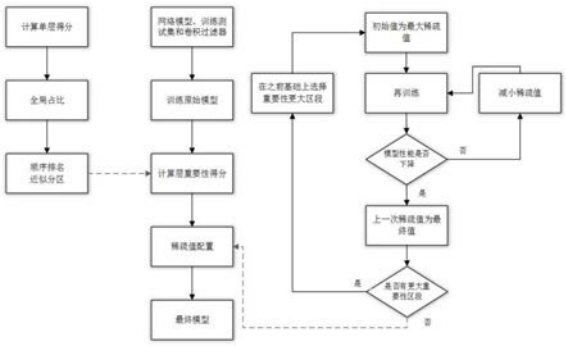
权利要求书2页 说明书5页 附图4页

(54)发明名称

一种针对卷积层的神经网络结构化剪枝压缩优化方法

(57)摘要

本发明公开了一种针对卷积层的神经网络结构化剪枝压缩优化方法,包括:(1)各卷积层稀疏值分配:(1.1)训练原始模型得到各可剪枝卷积层的权值参数,并计算得到各卷积层重要性分数;(1.2)按照重要性分数从小到大的顺序,并参照最大最小值进行平均刻度分段,依次对各区段卷积层进行稀疏值从小到大的配置,经过模型再训练调整,得到全部可剪枝卷积层的稀疏值配置;(2)结构化剪枝:根据步骤(1.2)确定的稀疏值选择卷积过滤器,进行结构化剪枝训练;其中,每层卷积层只使用一种卷积过滤器。本发明的优化方法,能够让深度神经网络在资源受限的平台上更便捷地运行,既能节省参数存储空间又能加速模型运算。



CN 109886397 A

1.一种针对卷积层的神经网络结构化剪枝压缩优化方法,其特征在于,包括:

(1)各卷积层稀疏值分配,包括:

(1.1)训练原始模型得到各可剪枝卷积层的权值参数,并计算得到各卷积层重要性分数;

(1.2)按照重要性分数从小到大的顺序,并参照最大最小值进行平均刻度分段,依次对各区段卷积层进行稀疏值从小到大的配置,经过模型再训练调整,得到全部可剪枝卷积层的稀疏值配置;

(2)结构化剪枝,包括:

根据步骤(1.2)确定的稀疏值选择卷积过滤器,进行结构化剪枝训练;

其中,每层卷积层只使用一种卷积过滤器。

2.根据权利要求1所述的一种针对卷积层的神经网络结构化剪枝压缩优化方法,其特征在于,步骤1中,训练原始模型得到各可剪枝卷积层的权值参数具体包括:

权值参数$k_{l,nchw}$,其中l为该层序号,n,c,h,w是卷积层权值参数的4-D张量指数,n为输入通道数,c为输出通道数,h、w分别为卷积核的高和宽;N为输入通道总数,C为输出通道总数,H、W分别为卷积核的总高和总宽,n,c,h,w为正整数且,n∈[1,N]、c∈[1,C]、h∈[1,H]、w∈[1,W]。

3.根据权利要求1所述的一种针对卷积层的神经网络结构化剪枝压缩优化方法,其特征在于,步骤1中,各卷积层重要性分数的计算表达式为:

$$M_l = \frac{\sum_{N,C}\left(\sum_{H,W}\left|k_{l,nchw}\right|\right)^2}{N \times C \times W \times H}$$

式中,对于指定的层l,使用M_l来表示该层的卷积核算子和值平方的平均值,n,c,h,w是卷积层权值参数的4-D张量指数,n为输入通道数,c为输出通道数,h、w分别为卷积核的高和宽;N为输入通道总数,C为输出通道总数,H、W分别为卷积核的总高和总宽,n,c,h,w为正整数且,n∈[1,N]、c∈[1,C]、h∈[1,H]、w∈[1,W]。

4.根据权利要求1所述的一种针对卷积层的神经网络结构化剪枝压缩优化方法,其特征在于,步骤2中,对各区段卷积层进行稀疏值从小到大配置的具体步骤包括:

每个可剪枝卷积层的稀疏值配置包括:改变其稀疏值,对模型进行再训练;如若模型性能保持良好,则继续增大其稀疏值,如若模型性能有较大的损失,则取上一次的稀疏值为其最终稀疏值;

重复卷积层稀疏值配置,直至完成最后一个区间的卷积层稀疏值配置,得到全可剪枝卷积层结构化剪枝的稀疏值初始配置;

其中,模型的性能的评价标准为准确率或者目标识别中mAP值;若保持准确率或者mAP值不下降则为模型性能保持良好,若下降超过预设阈值则表示模型性能有较大的损失。

5.根据权利要求4所述的一种针对卷积层的神经网络结构化剪枝压缩优化方法,其特征在于,得到全可剪枝卷积层结构化剪枝的稀疏值初始配置后,对靠近重要性区段两端的卷积层进行微调;

微调包括:将数值小的一端稀疏值变大,将数值大的一端稀疏值变小,并遵循改变一次立即进行再训练操作,得到最终全可剪枝卷积层的稀疏值配置。

6.根据权利要求1所述的一种针对卷积层的神经网络结构化剪枝压缩优化方法,其特征在于,步骤(2)中,卷积过滤器为与卷积核算子尺寸一样的剪枝模板。

7.根据权利要求1所述的一种针对卷积层的神经网络结构化剪枝压缩优化方法,其特征在于,步骤(2)中,卷积过滤器使用三个参数进行描述,Kp_stride为剪枝或保留的步长,Kp_offset＝i为减去的第一个值的位置编号为i,Kp_keepset＝j为保留的第一个值的位置编号为j。

一种针对卷积层的神经网络结构化剪枝压缩优化方法

技术领域

[0001] 本发明属于计算机人工智能领域、深度神经网络优化技术领域以及图片识别技术领域,特别涉及一种针对卷积层的神经网络结构化剪枝压缩优化方法。

背景技术

[0002] 在人工智能领域,深度神经网络作为基石之一,其复杂性以及可移植性直接影响人工智能在生活中的应用。对深度网络的加速与压缩优化的研究,可使得人工智能更加方便的实现、更方便的服务于生活。

[0003] 目前,常见的深度网络的加速与压缩方法有如下几种:1.Low-Rank:低秩分解;2.Pruning:剪枝,剪枝方法又分为:结构化剪枝、核剪枝、梯度剪枝,使用范围较广;3.Quantization:量化,量化又分为:低比特量化、总体训练加速量化、分布式训练梯度量化;4.Knowledge Distillation:知识蒸馏;5.Compact Network Design:紧凑网络设计,这是从网络结构层面对模型进行优化。

[0004] 本发明主要针对第二种压缩方法剪枝进行进一步的改进,现有技术方案中也使用了结构化剪枝的思想,它的方法是每层卷积层使用多个卷积过滤器,而且这些卷积过滤器的类型是通过训练得到的;现有方法不仅训练周期很长耗费计算资源巨大(导致无法流畅的使用大型的训练数据集),而且这样的结构化剪枝并不能在模型前向计算过程节省更多的计算、存储资源。

[0005] 综上,亟需一种新型的神经网络结构化剪枝压缩优化方法。

发明内容

[0006] 本发明的目的在于提供一种针对卷积层的神经网络结构化剪枝压缩优化方法,以解决上述存在的一个或多个技术问题。本发明的优化方法,能够让深度神经网络在资源受限的平台上更便捷地运行,既能节省参数存储空间又能加速模型运算。

[0007] 为达到上述目的,本发明采用以下技术方案:

[0008] 一种针对卷积层的神经网络结构化剪枝压缩优化方法,包括:

[0009] (1)各卷积层稀疏值分配,包括:

[0010] (1.1)训练原始模型得到各可剪枝卷积层的权值参数,并计算得到各卷积层重要性分数;

[0011] (1.2)按照重要性分数从小到大的顺序,并参照最大最小值进行平均刻度分段,依次对各区段卷积层进行稀疏值从小到大的配置,经过模型再训练调整,得到全部可剪枝卷积层的稀疏值配置;

[0012] (2)结构化剪枝,包括:

[0013] 根据步骤(1.2)确定的稀疏值选择卷积过滤器,进行结构化剪枝训练;

[0014] 其中,每层卷积层只使用一种卷积过滤器。

[0015] 本发明的进一步改进在于,步骤1中,训练原始模型得到各可剪枝卷积层的权值参

数具体包括:权值参数$k_{l,nchw}$,其中l为该层序号,n,c,h,w是卷积层权值参数的4-D张量指数,n为输入通道数,c为输出通道数,h、w分别为卷积核的高和宽;N为输入通道总数,C为输出通道总数,H、W分别为卷积核的总高和总宽,n,c,h,w为正整数且,n∈[1,N]、c∈[1,C]、h∈[1,H]、w∈[1,W]。

[0016] 本发明的进一步改进在于,步骤1中,各卷积层重要性分数的计算表达式为:

[0017] $$M_l = \frac{\sum_{N,C}\left(\sum_{H,W}\left|k_{l,nchw}\right|\right)^2}{N \times C \times W \times H}$$

[0018] 式中,对于指定的层l,使用M_l来表示该层的卷积核算子和值平方的平均值,n,c,h,w是卷积层权值参数的4-D张量指数,n为输入通道数,c为输出通道数,h、w分别为卷积核的高和宽;N为输入通道总数,C为输出通道总数,H、W分别为卷积核的总高和总宽,n,c,h,w为正整数且,n∈[1,N]、c∈[1,C]、h∈[1,H]、w∈[1,W]。

[0019] 本发明的进一步改进在于,步骤2中,对各区段卷积层进行稀疏值从小到大配置的具体步骤包括:每个可剪枝卷积层的稀疏值配置包括:改变其稀疏值,对模型进行再训练;如若模型性能保持良好,则继续增大其稀疏值,如若模型性能有较大的损失,则取上一次的稀疏值为其最终稀疏值;重复卷积层稀疏值配置,直至完成最后一个区间的卷积层稀疏值配置,得到全可剪枝卷积层结构化剪枝的稀疏值初始配置;

[0020] 其中,模型的性能的评价标准为准确率或者目标识别中的mAP值;若保持准确率或者mAP值不下降则为模型性能保持良好,若下降超过预设阈值则表示模型性能有较大的损失。

[0021] 本发明的进一步改进在于,得到全可剪枝卷积层结构化剪枝的稀疏值初始配置后,对靠近重要性区段两端的卷积层进行微调;

[0022] 微调包括:将数值小的一端稀疏值变大,将数值大的一端稀疏值变小,并遵循改变一次立即进行再训练操作,得到最终全可剪枝卷积层的稀疏值配置。

[0023] 本发明的进一步改进在于,步骤(2)中,卷积过滤器为与卷积核算子尺寸一样的剪枝模板。

[0024] 本发明的进一步改进在于,步骤(2)中,卷积过滤器使用三个参数进行描述,Kp_stride为剪枝或保留的步长,Kp_offset＝i为减去的第一个值的位置编号为i,Kp_keepset＝j为保留的第一个值的位置编号为j。

[0025] 与现有技术相比,本发明具有以下有益效果:

[0026] 本发明的优化方法,根据各卷积层的重要性得分合理分配每层的稀疏值,进行一层一种卷积过滤器的卷积算子级别的结构化剪枝,经过调参、再训练、调参的训练模式得到最终模型;在性能无明显降低的前提下,可使得整个卷积神经网络得到合理的结构化剪枝压缩优化,不仅能够大大降低参数存储空间,还具备了巨大的运算优化的潜力。另外,结构化剪枝后,一张数据流只需做一次部分规律的数据读取工作,读取的数据就能被反复利用,这将节省巨大的硬件平台的存储资源,并且节省大量的运算操作,具备很大的运算加速潜力,能够让深度神经网络在资源受限的平台上更便捷地运行,既能节省参数存储空间又能加速模型运算。

[0027] 进一步地,得到全可剪枝卷积层结构化剪枝的稀疏值初始配置后,由于之前是按照重要性分区段进行一个或者几个卷积层同时改变为相同稀疏值的操作,这个步骤会导致

靠近重要性区段两端的卷积层的稀疏值不是那么准确，后续需要对这些靠近重要性区段两端的一些卷积层进行数值小的一端稀疏值变大、数值大的一端稀疏值变小的微调，遵循改变一次立即进行再训练操作，比较模型性能前后的变化，得到最终全可剪枝卷积层的稀疏值配置，模型性能的评判依据准确率或者目标识别中的mAP值。

[0028] 进一步地，本发明选择人工调整稀疏率并且每一个卷积层只使用一种卷积过滤器，使得不需要进行漫长的训练来确定各层的稀疏值，并且由于每层只使用一种卷积过滤器，使得模型前向计算过程能节省大量的存储、计算资源。

附图说明

[0029] 图1为本发明实施例的优化方法中结构化剪枝原理示意图；

[0030] 图2为本发明实施例的优化方法中结构化剪枝运算原理示意图；

[0031] 图3为本发明实施例的优化方法中稀疏值配置原理示意图；

[0032] 图4为本发明实施例的优化方法中卷积算子尺寸3*3的卷积过滤器结构示意图。

具体实施方式

[0033] 下面结合附图和具体实施例对本发明作进一步详细说明。

[0034] 请参阅图1，图1所示为整个结构化剪枝压缩优化原理示意图。本发明实施例的一种针对深度神经网络卷积层的结构化剪枝压缩优化方法，具体步骤包括：各卷积层稀疏值分配和结构化剪枝两部分。

[0035] (1)各卷积层稀疏值分配步骤如下：首先，训练原始模型得到各可剪枝卷积层的参数数据，并计算单层重要性分数。对每层的重要性得分M_l求和得M，计算每层重要性全局占比 $D_l = \frac{M_l}{\mathrm{M}} * 100\%$。按照从小到大对$D_l$进行顺序排名，根据$D_l$最大、最小值进行等间距区间分段，具体区间数需要通过观察D_l数据规律结合经验给出，遵循总段数不超过层总数的一半，尽量区分开各卷积层。依次从重要性得分小到大的区段卷积层进行稀疏值从小到到的配置；每改变一次稀疏值进行一次再训练操作，如若模型性能保持良好，继续减小稀疏值，直至模型性能有较大的损失，用此次试验上一次的稀疏值为最终值。然后，在前面基础上对下一个重要性得分区间的卷积层重复上述工作，直至完成最后一个区间的卷积层稀疏值配置工作，得到全可剪枝卷积层结构化剪枝的稀疏值初始配置。最后，可以修改少数卷积层的稀疏值进行再训练微调，得到最终全可剪枝卷积层的稀疏值配置。其中，模型的性能的评价标准为准确率或者目标识别中的mAP值；若保持准确率或者mAP值不下降则为模型性能保持良好，若下降超过预设阈值则表示模型性能有较大的损失。

[0036] 本发明实施例选择的神经网络原始模型为目标检测网络YOLOv3，它的卷积算子的尺寸为3*3和1*1，选择尺寸为3*3的卷积层进行结构化剪枝。

[0037] 选择Pascal VOC的2012年的数据集，它的测试集包括11540＝5717+5823张图片，测试集包括4952张图片。然后使用官方配置文件yolov3-voc.cfg，训练得到原始模型，并测试得到此时模型的mAP值。

[0038] 根据原始模型参数得到卷积算子尺寸为3*3的卷积层的重要性得分：

[0039] $$M_l = \frac{\sum_{N,C}\left(\sum_{H,W}\left|k_{l,nchw}\right|\right)^2}{N \times C \times W \times H}$$

[0040] 式中，对于指定的层l，使用M_l来表示该层的卷积核算子和值平方的平均值，N,C,H,W是卷积层权值参数的4-D张量指数；其中，N为输入通道数，C为输出通道数，H、W分别为卷积核的高和宽，n,c,h,w为正整数且，n∈[1,N]、c∈[1,C]、h∈[1,H]、w∈[1,W]。

[0041] 请参阅图3，训练原始模型得到各可剪枝卷积层的参数数据，并计算单层重要性分数。对每层的重要性得分M_l求和得M，计算每层重要性全局占比$D_l = \frac{M_l}{\mathrm{M}} * 100\%$。按照从小到大对$D_l$进行顺序排名，根据$D_l$最大、最小值进行等间距区间分段，具体区间数需要通过观察D_l数据规律结合经验给出，遵循总段数不超过层总数的一半，尽量区分开各卷积层。每改变一次稀疏值进行一次再训练操作，如若模型性能保持良好，继续增大稀疏值，直至模型性能有较大的损失，用此次试验上一次的稀疏值为最终值；模型的性能的评价标准为准确率或者目标识别中的mAP值；若保持准确率或者mAP值不下降则为模型性能保持良好，若下降超过预设阈值则表示模型性能有较大的损失。然后，在前面基础上对下一个重要性得分区间的卷积层重复上述工作，直至完成最后一个区间的卷积层稀疏值配置工作，得到全可剪枝卷积层结构化剪枝的稀疏值初始配置。最后，可以修改少数卷积层的稀疏值进行再训练微调，得到最终全可剪枝卷积层的稀疏值配置；由于之前是按照重要性分区段进行一个或者几个卷积层同时改变为相同稀疏值的操作，这个步骤会导致靠近重要性区段两端的卷积层的稀疏值不是那么准确，后续需要对这些靠近重要性区段两端的一些卷积层进行数值小的一端稀疏值变大、数值大的一端稀疏值变小的微调，遵循改变一次立即进行再训练操作，比较模型性能前后的变化，得到最终全可剪枝卷积层的稀疏值配置，模型性能的评判依据准确率或者目标识别中的mAP值。

[0042] (2)结构化剪枝步骤如下：根据稀疏值随机选择该稀疏值对应的卷积过滤器中的一种，但必须遵循一个卷积层只能选择一种卷积过滤器，卷积过滤器就是和卷积核算子尺寸一样的剪枝模板。遵循每层卷积层使用同一种卷积过滤器，进行结构化剪枝训练。

[0043] 请参阅图2和图4，根据配置的稀疏值，从如图4所示中选择的3*3的卷积过滤器。按照如图2所示的结构化剪枝运算原理进行再训练操作。卷积过滤器使用三个参数进行描述，Kp_stride为剪枝(或保留)的步长，Kp_offset＝i为减去的第一个值的位置编号为i，Kp_keepset＝j为保留的第一个值的位置编号为j。

[0044] 图2中常规运算是由上层得的输入数据(5*5)，按照卷积核(3*3)，经过重排图像块为矩阵列(im2col)得到输入数据矩阵(9*9)，再和卷积核(9*1)相乘得到结果(9*1)．而结构化剪枝则是将输入数据(5*5)，按照卷积过滤器(3*3)，不读取剪枝部分数据的im2col操作，得到输入数据矩阵(9*4)，再和经过卷积过滤器剪枝后的卷积核(4*1)相乘得到结果(9*1)．由于每层只选择一种卷积过滤器，于是上层的输入数据只需要做一次im2col得到的输入数据矩阵(9*4)能被该层其他经过卷积过滤器剪枝后的卷积核使用，从而不需要因为一层中有几种不同类型的卷积过滤器而对输入数据做多次im2col操作。不仅减少运算量而且节省了大量的存储资源。

[0045] 图4中本发明选择卷积核大小3*3进行卷积过滤器介绍，图中“0”值表示对卷积核相应位置进行剪枝，而“1”值表示保留卷积核对应位置的权值参数。而对于每个稀疏值下的

卷积过滤器形状通过枚举方法得到。对Kp_stride、Kp_offset和Kp_stride、Kp_keepset两种组合进行Kp_stride、Kp_offset、Kp_keepset∈[0,$ksize^2$-1],其中三个卷积过滤器参数为整数,ksize为卷积核边长。然后剔除掉“1”和“0”值不对称的卷积过滤器。

[0046] 现有的非结构化剪枝压缩方法虽然能到达很高的压缩率,但是压缩后的模型难以进行运算优化,不利于卷积神经网络在资源受限的硬件平台上实现。针对这个问题,本发明对每个卷积层进行结构化剪枝操作。首先,根据卷积层卷积算子的尺寸,选择出数据读取更流畅的卷积过滤器,并将其根据稀疏值进行分类;然后对每个可剪枝卷积层相对整个卷积神经网络重要性进行评估,对每个可剪枝的卷积层分配合适的稀疏值及合适的卷积过滤器;使用再训练、调参、再训练的训练模式,在性能无明显降低的前提下,使得整个卷积神经网络得到合理的结构化剪枝压缩优化,不仅大大降低参数存储空间还具备了巨大的运算优化的潜力。

[0047] 综上,本发明的针对卷积层的神经网络结构化剪枝压缩优化方法,属于剪枝方法中的结构化剪枝方法,剪枝的对象是卷积核算子;通过对每层卷积层依据其重要性得分配置合适的稀疏值,然后每一层使用同一种卷积过滤器,因此深度网络模型不仅在参数存储空间上大大减小,而且这种一层一种卷积过滤器的设置方式,能够带来很大的运算加速效果。通过本发明方法结构化剪枝后,一张数据流只需做一次部分规律的数据读取工作,读取的数据就能被反复利用,这将将节省巨大的硬件平台的存储资源,并且节省大量的运算操作,具备很大的运算加速潜力。

[0048] 以上实施例仅用以说明本发明的技术方案而非对其限制,尽管参照上述实施例对本发明进行了详细的说明,所属领域的普通技术人员依然可以对本发明的具体实施方式进行修改或者等同替换,这些未脱离本发明精神和范围的任何修改或者等同替换,均在申请待批的本发明的权利要求保护范围之内。

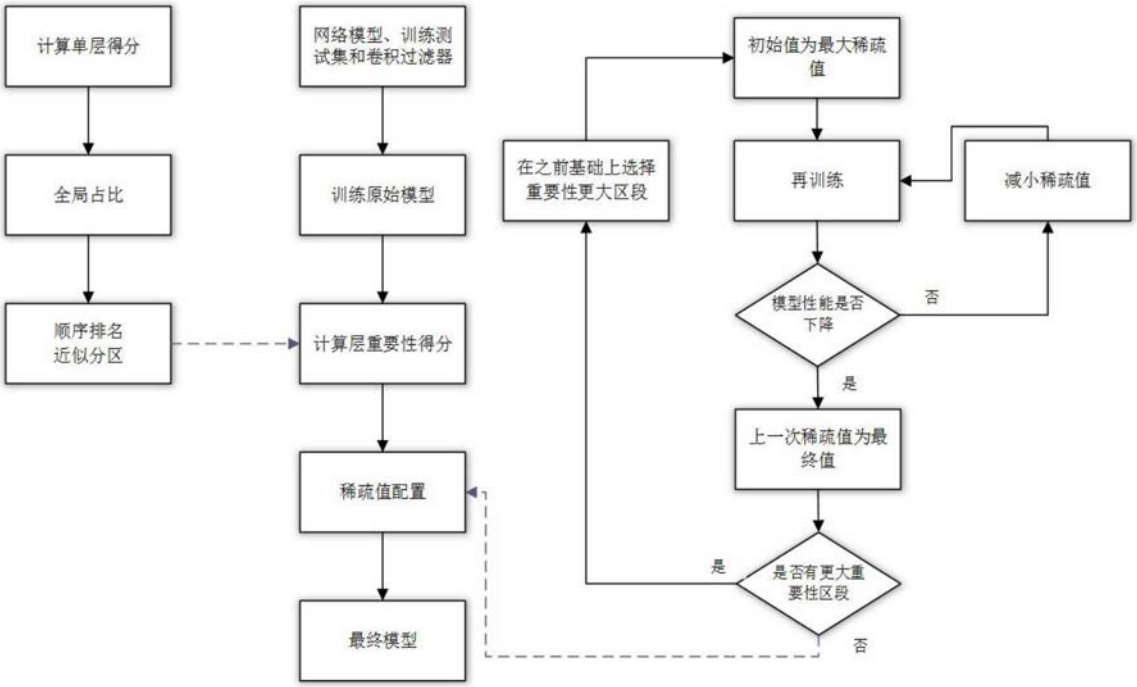

计算单层得分
全局占比
顺序排名
近似分区
网络模型、训练测试集和卷积过滤器
训练原始模型
计算层重要性得分
稀疏值配置
最终模型
初始值为最大稀疏值
再训练
减小稀疏值
模型性能是否下降
否
是
上一次稀疏值为最终值
是否有更大重要性区段
是
否
在之前基础上选择重要性更大区段

图1

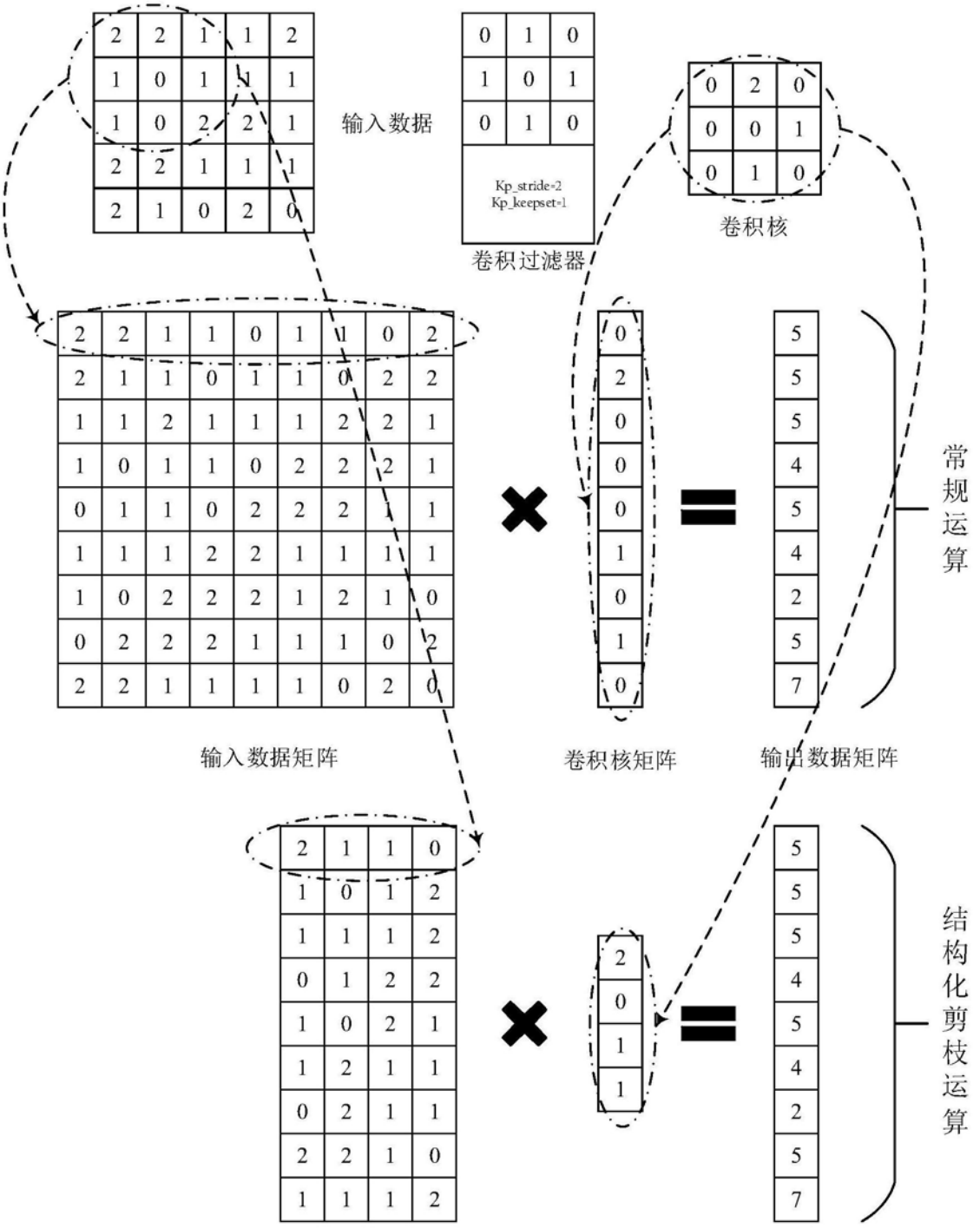

输入数据
Kp_stride=2
Kp_keepset=1
卷积过滤器
卷积核
输入数据矩阵
卷积核矩阵
输出数据矩阵
常规运算
结构化剪枝运算

图2

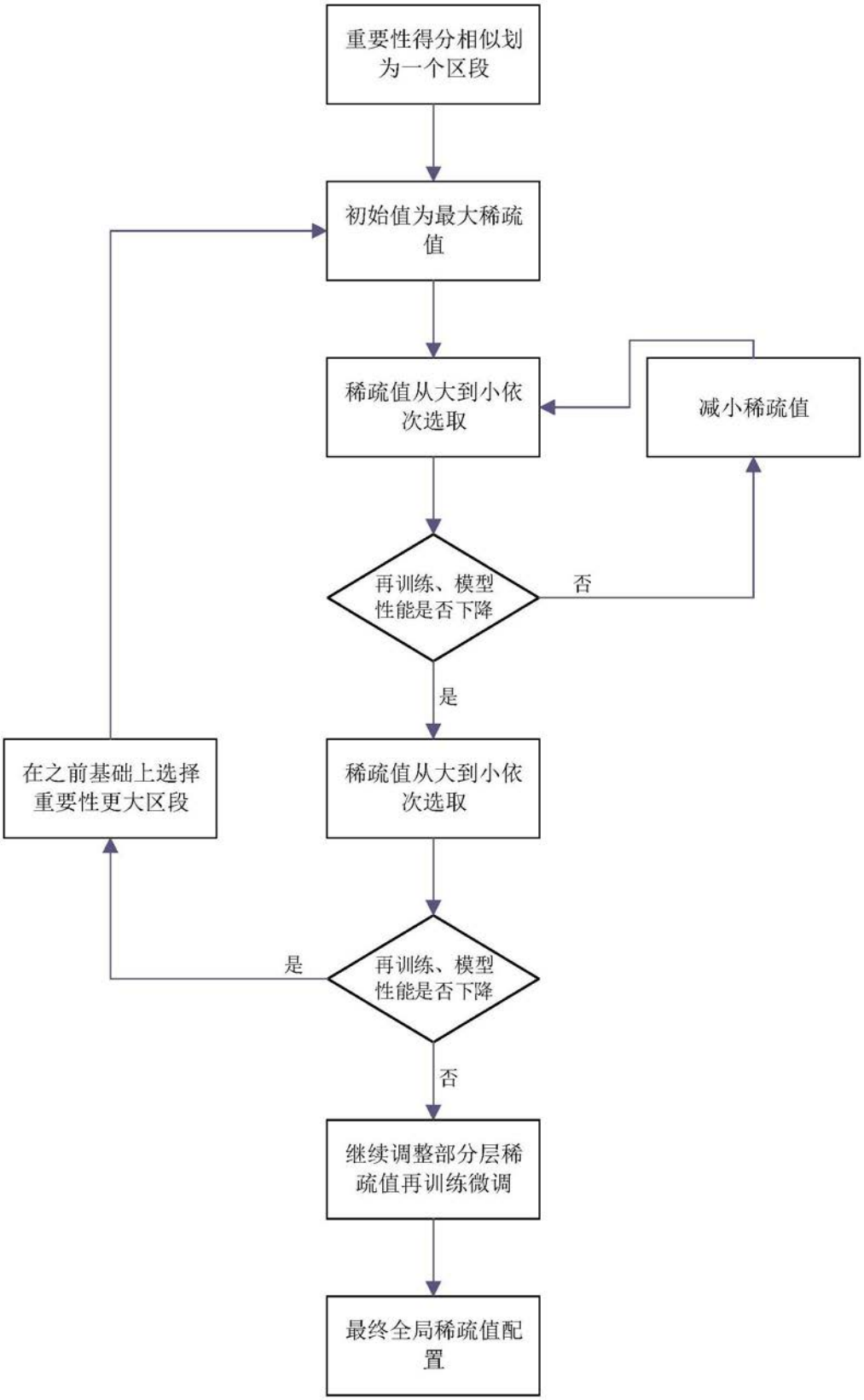

重要性得分相似划为一个区段
初始值为最大稀疏值
稀疏值从大到小依次选取
减小稀疏值
再训练、模型性能是否下降
否
是
在之前基础上选择重要性更大区段
稀疏值从大到小依次选取
是
再训练、模型性能是否下降
否
继续调整部分层稀疏值再训练微调
最终全局稀疏值配置

图3

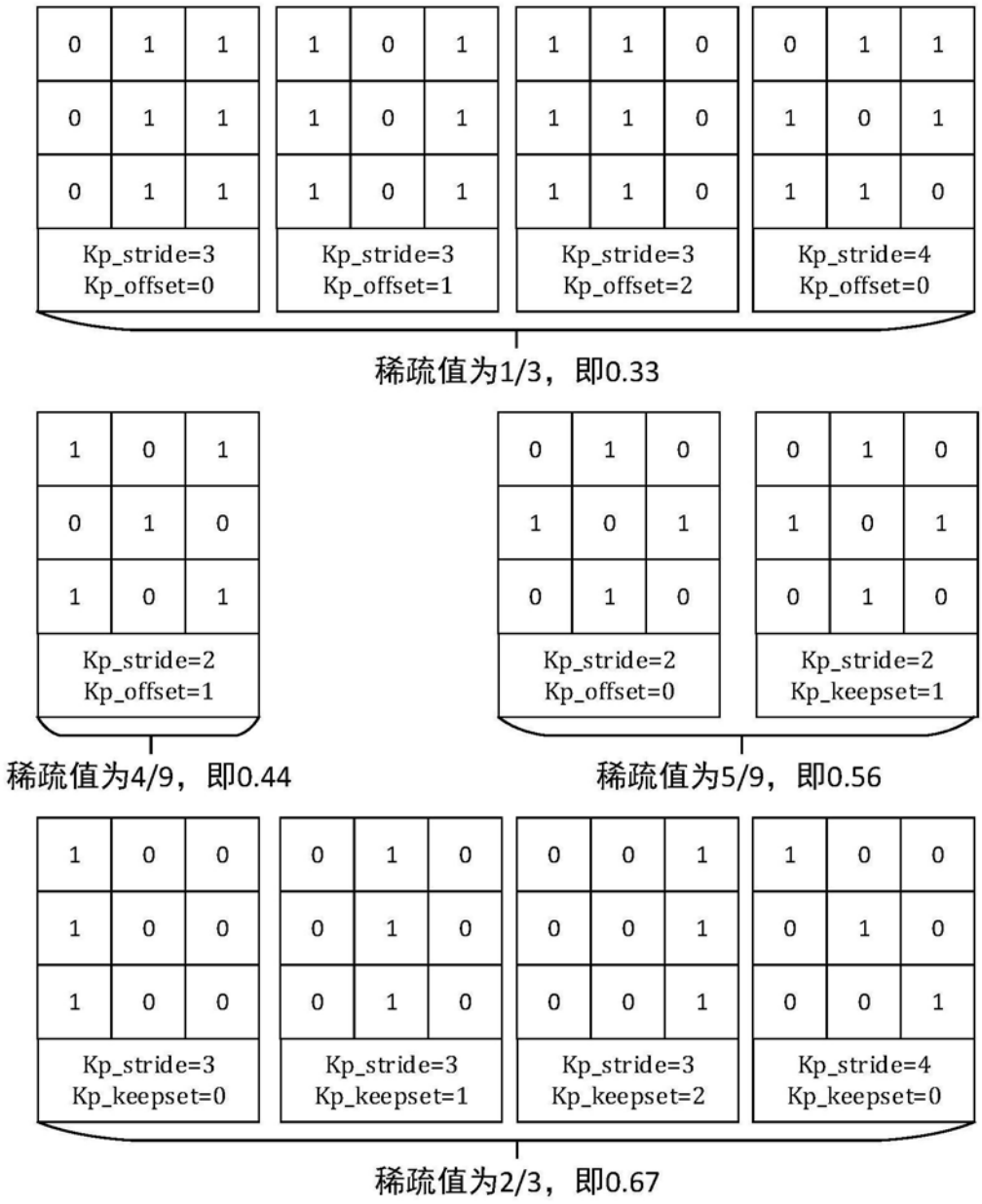

0 1 1
0 1 1
0 1 1
Kp_stride=3
Kp_offset=0
1 0 1
1 0 1
1 0 1
Kp_stride=3
Kp_offset=1
1 1 0
1 1 0
1 1 0
Kp_stride=3
Kp_offset=2
0 1 1
1 0 1
1 1 0
Kp_stride=4
Kp_offset=0
稀疏值为1/3，即0.33
1 0 1
0 1 0
1 0 1
Kp_stride=2
Kp_offset=1
稀疏值为4/9，即0.44
0 1 0
1 0 1
0 1 0
Kp_stride=2
Kp_offset=0
0 1 0
1 0 1
0 1 0
Kp_stride=2
Kp_keepset=1
稀疏值为5/9，即0.56
1 0 0
1 0 0
1 0 0
Kp_stride=3
Kp_keepset=0
0 1 0
0 1 0
0 1 0
Kp_stride=3
Kp_keepset=1
0 0 1
0 0 1
0 0 1
Kp_stride=3
Kp_keepset=2
1 0 0
0 1 0
0 0 1
Kp_stride=4
Kp_keepset=0
稀疏值为2/3，即0.67

图4