

(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
2022 年 9 月 1 日 (01.09.2022)



(10) 国际公布号
WO 2022/178908 A1

(51) 国际专利分类号:
G06N 3/08 (2006.01)

(21) 国际申请号: PCT/CN2021/078455

(22) 国际申请日: 2021 年 3 月 1 日 (01.03.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110216506.6 2021年2月26日 (26.02.2021) CN

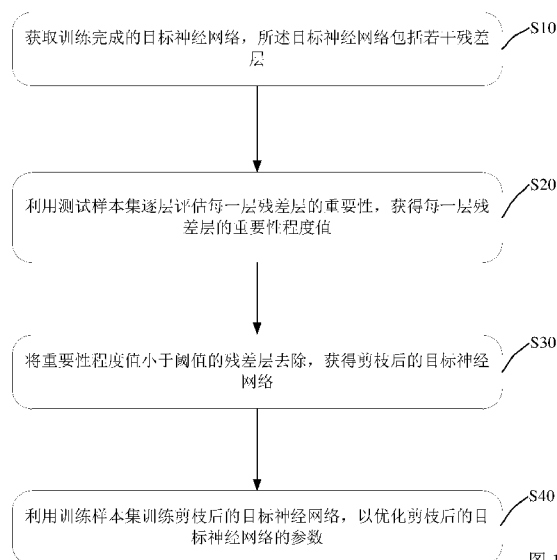
(71) 申请人: 中国科学院深圳先进技术研究院 (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES) [CN/CN]; 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(72) 发明人: 梁栋(LIANG, Dong); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。 王位(WANG, Wei); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。 朱燕杰(ZHU, Yanjie); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。 崔卓须(CUI, Zhuoxu); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。 刘新(LIU, Xin); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。 郑海荣(ZHENG, Hairong); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(74) 代理人: 深圳市铭粤知识产权代理有限公司(MING & YUE INTELLECTUAL PROPERTY LAW FIRM);

(54) Title: NEURAL NETWORK PRUNING METHOD AND APPARATUS, AND STORAGE MEDIUM

(54) 发明名称: 神经网络的剪枝方法及其剪枝装置、存储介质



S10 Acquire a completely trained target neural network, wherein the target neural network comprises several residual layers
S20 Evaluate the importance of all the residual layers layer by layer using a test sample set, in order to obtain an importance level value of each residual layer
S30 Remove a residual layer, the importance level value of which is less than a threshold, in order to obtain a pruned target neural network
S40 Train the pruned target neural network using a training sample set so as to optimize parameters of the pruned target neural network

(57) Abstract: A neural network pruning method and apparatus, and a storage medium. The pruning method comprises: acquiring a completely trained target neural network, wherein the target neural network comprises several residual layers (S10); evaluating the importance of all the residual layers layer by layer by using a test sample set, in order to obtain an importance level value of each residual layer (S20); removing a residual layer, the importance level value of which is less than a threshold, in order to obtain a pruned target neural network (S30); and training the pruned target neural network using a training sample set so as to optimize parameters of the pruned target neural network (S40). According to the method, the importance of parameters of network layers can be tested layer by layer, and the importance of the parameters of the layers is intuitively provided by means of presentation of results in a test set, thereby improving the universality and interpretability of the pruning method.

中国广东省深圳市南山区南山街道前海路泛海城市广场2栋604室, Guangdong 518066 (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, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, 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, 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))。

(57) 摘要: 一种神经网络的剪枝方法及其剪枝装置、存储介质, 其中剪枝方法包括: 获取训练完成的目标神经网络, 所述目标神经网络包括若干残差层(S10); 利用测试样本集逐层评估每一层残差层的重要性, 获得每一层残差层的重要性程度值(S20); 将重要性程度值小于阈值的残差层去除, 获得剪枝后的目标神经网络(S30); 利用训练样本集训练剪枝后的目标神经网络, 以优化剪枝后的目标神经网络的参数(S40)。该方法可以逐层测试网络各层参数的重要性, 通过在测试集上结果的表现直观地给出了各层各参数的重要性, 提高了剪枝方法的通用性和可解释性。

神经网络的剪枝方法及其剪枝装置、存储介质

技术领域

- 5 本发明属于机器学习技术领域，具体地讲，涉及一种神经网络的剪枝方法及其剪枝装置、计算机可读存储介质。

背景技术

- 深度学习目前广泛应用于各行，但是在许多现实应用中，深度学习网络的部署是一个很大的难题。一方面深度学习网络参数庞大，对内存的需求很高，另一方面深度学习需要大量计算。尤其是常用的卷积神经网络，其存在大量冗余参数，往往需要进行网络剪枝。网络剪枝的基本思想就是利用不同参数的重要性不同，裁剪掉那些冗余参数，使网络更精简，减少存储占用和计算量。针对参数重要性剪枝思路是建立一个评估标准，然后按照评估标准对网络的参数进行排序，设定阈值，然后剪除小于阈值的那些网络参数。

例如，文章(Li, Kadav, Durdanovic, Samet & Graf, 2016)中提出思路是：

- 1)计算卷积核中所有权值的绝对值之和；2)根据求和大小排列卷积核；3)删除数值较小的卷积核；4)对删除之后的卷积核重新组合，生成新的卷积核矩阵。

- 另一文章(Liu et al. 2017)指出，通过 BN 层的缩放因子进行网络剪枝，裁剪值小的缩放因子对应卷积核。缩放因子的选取取决于我们为整个网络所有层设置的一个全局阈值，它被定义为所有缩放因子值的一个比例，比如我们将剪掉整个网络中 70%的通道，那么我们先对缩放因子的绝对值排个序，然后取从小到大排序的缩放因子中 70%的缩放因子为阈值。

现有技术的看似具有通用性，但是具有一个明显的缺陷，那就是没有考虑到网络结构的特异性，网络不同层的差异性和方法的可解释性。首先不同网络结构是不同的，而他们对参数的敏感性也不同，不可以所有网络同等对待。网

络从前到后是一个有层次的学习过程，而在此过程参数的敏感度也在变化，网络的前后参数不可同等对待。最后是方法的可解释性，目前的方法无法给出令人信服的解释。

5 发明内容

（一）本发明所要解决的技术问题

本发明解决的技术问题是：如何提供一种具有通用性和可解释性的神经网络的剪枝方法。

（二）本发明所采用的技术方案

10 一种神经网络的剪枝方法，所述剪枝方法包括：

获取训练完成的目标神经网络，所述目标神经网络包括若干残差层；

利用测试样本集逐层评估每一层残差层的重要性，获得每一层残差层的重要性程度值；

将重要性程度值小于阈值的残差层去除，获得剪枝后的目标神经网络；

15 利用训练样本集训练剪枝后的目标神经网络，以优化剪枝后的目标神经网络的参数。

可选择地，利用测试样本集逐层评估每一层残差层的重要性，获得每一层残差层的重要性程度值的方法包括：

逐层将残差层的卷积核替换为零；

20 利用测试样本集测试各个残差层的卷积核替换为零之前和之后的目标神经网络的性能指标值，并计算得到目标神经网络的性能指标变化值；

根据性能指标变化值确定残差层的重要性程度值。

可选择地，当性能指标变化值小于预设变化值时，残差层的重要性程度值为第一程度值；当性能指标变化值大于或等于预设变化值时，残差层的重要性程度值为第二程度值，其中所述第一程度值大于所述阈值，所述第二程度值小于所述阈值。

25

可选择地，所述性能指标值包括精确率、准确率、召回率、Dice 值、Iou

值中的任意一种或多种。

本申请还公开了一种神经网络的剪枝装置，所述剪枝装置包括：

获取单元，用于获取训练完成的目标神经网络，所述目标神经网络包括若干残差层；

- 5 评估单元，用于利用测试样本集评估每一层残差层的重要性，获得每一层残差层的重要性程度值；

剪枝单元，用于将重要性程度值小于阈值的残差层去除，获得剪枝后的目标神经网络；

- 10 训练单元，用于利用训练样本集训练剪枝后的目标神经网络，以优化剪枝后的目标神经网络的参数。

所述评估单元具体用于：

将残差层的卷积核替换为零；

利用测试样本集测试残差层的卷积核替换为零之前和之后的目标神经网络的性能指标值，并计算得到目标神经网络的性能指标变化值；

- 15 根据性能指标变化值确定残差层的重要性程度值。

可选择地，当性能指标变化值小于预设变化值时，所述评估单元还用于将残差层的重要性程度值设置为第一程度值；当性能指标变化值大于或等于预设变化值时，所述评估单元还用于将残差层的重要性程度值为第二程度值，其中所述第一程度值大于所述阈值，所述第二程度值小于所述阈值。

- 20 可选择地，所述性能指标值包括精确率、准确率、召回率、Dice 值、Iou 值中的任意一种或多种。

本申请还公开了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神经网络的剪枝程序，所述神经网络的剪枝程序被处理器执行时实现上述的神经网络的剪枝方法。

- 25 本申请还公开了一种计算机设备，所述计算机设备包括计算机可读存储介质、处理器和存储在所述计算机可读存储介质中的神经网络的剪枝程序，所述神经网络的剪枝程序被处理器执行时实现上述的神经网络的剪枝方法。

（三）有益效果

本发明公开了一种神经网络的剪枝方法，相对于传统的剪枝方法，具有如下技术效果：

利用测试样本集分别对训练完成的目标神经网络的每一层残差层进行评估，获得每一层残差层的重要性程度值，接着将重要性程度值小于阈值的残差层去除，保持其他参数不变，获得剪枝后的目标神经网络，这样可以逐层测试网络各层参数可重用性，通过在测试集上结果的表现直观地给出了各层各参数的重要性，提高了剪枝方法的通用性和可解释性。

附图说明

- 10 图 1 为本发明的实施例一的神经网络的剪枝方法的流程图；
图 2 为本发明的实施例一的神经网络的剪枝方法的详细步骤示意图；
图 3 为本发明的实施例一的目标神经网络的局部结构示意图；
图 4 为本发明的实施例二的神经网络的剪枝装置的示意图；
图 5 为本发明的实施例四的计算机设备示意图。

15

具体实施方式

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，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。

- 20 在详细描述本申请的各个实施例之前，首先简单描述本申请的发明构思：
现有的神经网络剪枝算法由于没有考虑网络结构的特异性以及不同网络层的差异性，导致算法不具有通用性和可解释性，本方案首先利用测试样本集分别对训练完成的目标神经网络的每一层残差层进行评估，获得每一层残差层的重要性程度值，接着将重要性程度值小于阈值的残差层去除，保持其他参数不变，
25 获得剪枝后的目标神经网络，这样可以逐层测试网络各层参数可重用性，通过在测试集上结果的表现直观地给出了各层各参数的重要性，提高了剪枝方法的通用性和可解释性。

如图 1 和图 2 所示，本实施例一的神经网络的剪枝方法包括如下步骤：

步骤 S10：获取训练完成的目标神经网络，所述目标神经网络包括若干残差层。

步骤 S20：利用测试样本集逐层评估每一层残差层的重要性，获得每一层残差层的重要性程度值。

5 步骤 S30：将重要性程度值小于阈值的残差层去除，获得剪枝后的目标神经网络。

步骤 S40：利用训练样本集训练剪枝后的目标神经网络，以优化剪枝后的目标神经网络的参数。

10 作为优选实施例，在步骤 S10 中，如图 3 所示，目标神经网络以 ResNet 网络为例。

ResNet 的网络结构包括短连接结构和若干残差层，其中短连接结构主要负责信息存储，残差层主要负责微调学习到的信息。因为短连接中存储了学习到的全部信息，所以相对信息存储结构，负责微调信息的残差层就相对不是很重要，因此残差层可以删除一部分。其中短连接结构和残差层中有两个特殊的层：
15 短连接特征分解层；参差特征分解层。短连接特征分解层负责对上面学习到的特征进行分解和组合，是一个层块（具有相同特征数的连续层称为一个层块）中最重要的部分。短连接特征分解层负责分解和组合上一个层块学到的信息，更重要的是存储学到的信息。步骤 S10 利用训练集训练目标神经网络，得到训练完成的目标神经网络。

20 在步骤 S20 中，利用测试样本集逐层评估每一层残差层的重要性，获得每一层残差层的重要性程度值包括如下步骤：

步骤 S21：逐层将残差层的卷积核替换为零；

步骤 S22：利用测试样本集测试各残差层的卷积核替换为零之后的目标神经网络的性能指标值，并计算得到目标神经网络的性能指标变化值；

25 步骤 S23：根据性能指标变化值确定残差层的重要性程度值。

需要说明的是逐层是指每层进行测试时，仅将该层的卷积核替换为零，其他层的卷积核保持不变。具体来说，步骤 S20 主要应用了零替换卷积核操作。ResNet 残差单元具有多种形式，可以表示为如下形式：

$$\mathbf{x}_{l+1} = \sigma(\mathbf{x}_l + \text{BN}(\sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l') * \mathbf{w}_l'')))$$

$$\mathbf{x}_{l+1} = \sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l^{1*1}) + \text{BN}(\sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l') * \mathbf{w}_l'')))$$

用零分别替换各卷积核后可以表示为:

$$\mathbf{x}_{l+1}' = \sigma(\mathbf{x}_l + \text{BN}(\sigma(\text{BN}(\mathbf{x}_l * \mathbf{0}) * \mathbf{w}_l'')) = \sigma(\mathbf{x}_l + \text{BN}(\sigma(\beta') * \mathbf{w}_l''))$$

$$\mathbf{x}_{l+1}'' = \sigma(\mathbf{x}_l + \text{BN}(\sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l') * \mathbf{0}))) = \sigma(\mathbf{x}_l + \beta'')$$

$$\mathbf{x}_{l+1}' = \sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l^{1*1}) + \text{BN}(\sigma(\text{BN}(\mathbf{x}_l * \mathbf{0}) * \mathbf{w}_l'')) = \sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l^{1*1}) + \text{BN}(\sigma(\beta') * \mathbf{w}_l''))$$

$$\mathbf{x}_{l+1}'' = \sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l^{1*1}) + \text{BN}(\sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l') * \mathbf{0}))) = \sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l^{1*1}) + \beta'')$$

$$\mathbf{x}_{l+1}''' = \sigma(\text{BN}(\mathbf{x}_l * \mathbf{0}) + \text{BN}(\sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l') * \mathbf{w}_l'')) = \sigma(\beta''') + \text{BN}(\sigma(\text{BN}(\mathbf{x}_l * \mathbf{w}_l') * \mathbf{w}_l''))$$

$\mathbf{x}_l, \mathbf{x}_{l+1}$ 分别表示第 $l, l+1$ 个残差单元的输入输出特征图;

10 $\mathbf{x}_{l+1}', \mathbf{x}_{l+1}'', \mathbf{x}_{l+1}'''$ 表示第 $l+1$ 个残差单元分别以不同方式进行零替换的输出特征图;

$\mathbf{w}_l', \mathbf{w}_l''$ 分别表示第 l 个残差单元的第二个卷积核和第一个卷积核;

*表示卷积操作; BN (Batch normalization) 表示批标准化; σ 表示激活函数;

15 $\beta', \beta'', \beta'''$ 分别表示第 l 个残差单元不同 BN 层的偏置。

进一步地, 当性能指标变化值小于预设变化值时, 残差层的重要性程度值为第一程度值; 当性能指标变化值大于或等于预设变化值时, 残差层的重要性程度值为第二程度值, 其中所述第一程度值大于所述阈值, 所述第二程度值小于所述阈值。其中性能指标值根据任务定义, 如针对分类任务可以是精确率、准确率、召回率, 针对分割任务可以是 Dice 值、Iou 值 (intersection over union, 交并比值)。示例性地, 性能指标值选为准确率, 当准确率变化值小于 5% 时, 20 则该残差层的重要性程度较低, 可以被删除, 当准确率变化值大于或等于 5% 时, 则该残差层的重要性程度较高, 不可以被删除。重复上述步骤 S20, 直至确定每一层残差层的重要性程度值。

25 在步骤 S30 中, 将重要性程度值为第二程度值的残差层去除, 获得剪枝后

的目标神经网络。剪枝完成后，需要利用训练样本集进一步训练目标神经网络，以优化目标神经网络的参数。

本实施例一提供的神经网络的剪枝方法，利用测试样本集分别对训练完成的目标神经网络的每一层残差层进行评估，获得每一层残差层的重要性程度值，接着将重要性程度值小于阈值的残差层去除，保持其他参数不变，获得剪枝后的目标神经网络，这样可以逐层测试网络各层参数可重用性，通过在测试集上结果的表现直观地给出了各层各参数的重要性，提高了剪枝方法的通用性和可解释性。

进一步地，如图 4 所示，本实施例二提供了一种神经网络的剪枝装置，所述剪枝装置包括获取单元 100、评估单元 200、剪枝单元 300 和训练单元 400。其中，获取单元 100 用于获取训练完成的目标神经网络，所述目标神经网络包括若干残差层；评估单元 200 用于利用测试样本集评估每一层残差层的重要性，获得每一层残差层的重要性程度值；剪枝单元 300 用于将重要性程度值小于阈值的残差层去除，获得剪枝后的目标神经网络；训练单元 400 用于利用训练样本集训练剪枝后的目标神经网络，以优化剪枝后的目标神经网络的参数。

进一步地，所述评估单元 200 具体用于：将残差层的卷积核替换为零；利用测试样本集测试残差层的卷积核替换为零之前和之后的目标神经网络的性能指标值，并计算得到目标神经网络的性能指标变化值；根据性能指标变化值确定残差层的重要性程度值。

当性能指标变化值小于预设变化值时，所述评估单元 200 还用于将残差层的重要性程度值设置为第一程度值；当性能指标变化值大于或等于预设变化值时，所述评估单元 200 还用于将残差层的重要性程度值为第二程度值，其中所述第一程度值大于所述阈值，所述第二程度值小于所述阈值。评估单元 200 的具体评估过程参考实施例一中关于步骤 S20 的描述，在此不进行赘述。

本申请的实施例三还公开了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神经网络的剪枝程序，所述神经网络的剪枝程序被处理器执行时实现上述的神经网络的剪枝方法。

本申请的另一实施例还公开了一种计算机设备，在硬件层面，如图 5 所示，该终端包括处理器 12、内部总线 13、网络接口 14、计算机可读存储介质 11。处理器 12 从计算机可读存储介质中读取对应的计算机程序然后运行，在逻辑

层面上形成请求处理装置。当然，除了软件实现方式之外，本说明书一个或多个实施例并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限于各个逻辑单元，也可以是硬件或逻辑器件。所述计算机可读存储介质 11 上存储有神经网络的剪枝程序，所述神经网络

5 神经网络的剪枝程序被处理器执行时实现上述的神经网络的剪枝方法。

计算机可读存储介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机可读存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、

10 其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘只读存储器（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带、磁盘存储、量子存储器、基于石墨烯的存储介质或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。

15 上面对本发明的具体实施方式进行了详细描述，虽然已表示和描述了一些实施例，但本领域技术人员应该理解，在不脱离由权利要求及其等同物限定其范围的本发明的原理和精神的情况下，可以对这些实施例进行修改和完善，这些修改和完善也应在本发明的保护范围内。

权利要求书

1、一种神经网络的剪枝方法，其中，所述剪枝方法包括：

获取训练完成的目标神经网络，所述目标神经网络包括若干残差层；

5 利用测试样本集逐层评估每一层残差层的重要性，获得每一层残差层的重要性程度值；

将重要性程度值小于阈值的残差层去除，获得剪枝后的目标神经网络；

利用训练样本集训练剪枝后的目标神经网络，以优化剪枝后的目标神经网络的参数。

10 2、根据权利要求 1 所述的神经网络的剪枝方法，其中，利用测试样本集逐层评估每一层残差层的重要性，获得每一层残差层的重要性程度值的方法包括：

逐层将残差层的卷积核替换为零；

利用测试样本集测试各个残差层的卷积核替换为零之前和之后的目标神经网络的性能指标值，并计算得到目标神经网络的性能指标变化值；

15 根据性能指标变化值确定残差层的重要性程度值。

3、根据权利要求 2 所述的神经网络的剪枝方法，其中，当性能指标变化值小于预设变化值时，残差层的重要性程度值为第一程度值；当性能指标变化值大于或等于预设变化值时，残差层的重要性程度值为第二程度值，其中所述第一程度值大于所述阈值，所述第二程度值小于所述阈值。

20 4、根据权利要求 2 所述的神经网络的剪枝方法，其中，所述性能指标值包括精确率、准确率、召回率、Dice 值、Iou 值中的任意一种或多种。

5、一种神经网络的剪枝装置，其中，所述剪枝装置包括：

获取单元，用于获取训练完成的目标神经网络，所述目标神经网络包括若干残差层；

25 评估单元，用于利用测试样本集评估每一层残差层的重要性，获得每一层残差层的重要性程度值；

剪枝单元，用于将重要性程度值小于阈值的残差层去除，获得剪枝后的目

标神经网络；

训练单元，用于利用训练样本集训练剪枝后的目标神经网络，以优化剪枝后的目标神经网络的参数。

5 6、根据权利要求 5 所述的神经网络的剪枝装置，其中，所述评估单元具体用于：

逐层将残差层的卷积核替换为零；

利用测试样本集测试各个残差层的卷积核替换为零之前和之后的目标神经网络的性能指标值，并计算得到目标神经网络的性能指标变化值；

根据性能指标变化值确定残差层的重要性程度值。

10 7、根据权利要求 6 所述的神经网络的剪枝装置，其中，当性能指标变化值小于预设变化值时，所述评估单元还用于将残差层的重要性程度值设置为第一程度值；当性能指标变化值大于或等于预设变化值时，所述评估单元还用于将残差层的重要性程度值为第二程度值，其中所述第一程度值大于所述阈值，所述第二程度值小于所述阈值。

15 8、根据权利要求 6 所述的神经网络的剪枝装置，其中，所述性能指标值包括精确率、准确率、召回率、Dice、Iou 值中的任意一种或多种。

9、一种计算机可读存储介质，其中，所述计算机可读存储介质存储有神经网络的剪枝程序，所述神经网络的剪枝程序被处理器执行时实现权利要求 1 所述的神经网络的剪枝方法。

20 10、根据权利要求 6 所述的计算机可读存储介质，其中，利用测试样本集逐层评估每一层残差层的重要性，获得每一层残差层的重要性程度值的方法包括：

逐层将残差层的卷积核替换为零；

25 利用测试样本集测试各个残差层的卷积核替换为零之前和之后的目标神经网络的性能指标值，并计算得到目标神经网络的性能指标变化值；

根据性能指标变化值确定残差层的重要性程度值。

11、根据权利要求 10 所述的计算机可读存储介质，其中，当性能指标变化值小于预设变化值时，残差层的重要性程度值为第一程度值；当性能指标变

化值大于或等于预设变化值时，残差层的重要性程度值为第二程度值，其中所述第一程度值大于所述阈值，所述第二程度值小于所述阈值。

12、根据权利要求 2 所述的计算机可读存储介质，其中，所述性能指标值包括精确率、准确率、召回率、Dice 值、Iou 值中的任意一种或多种。

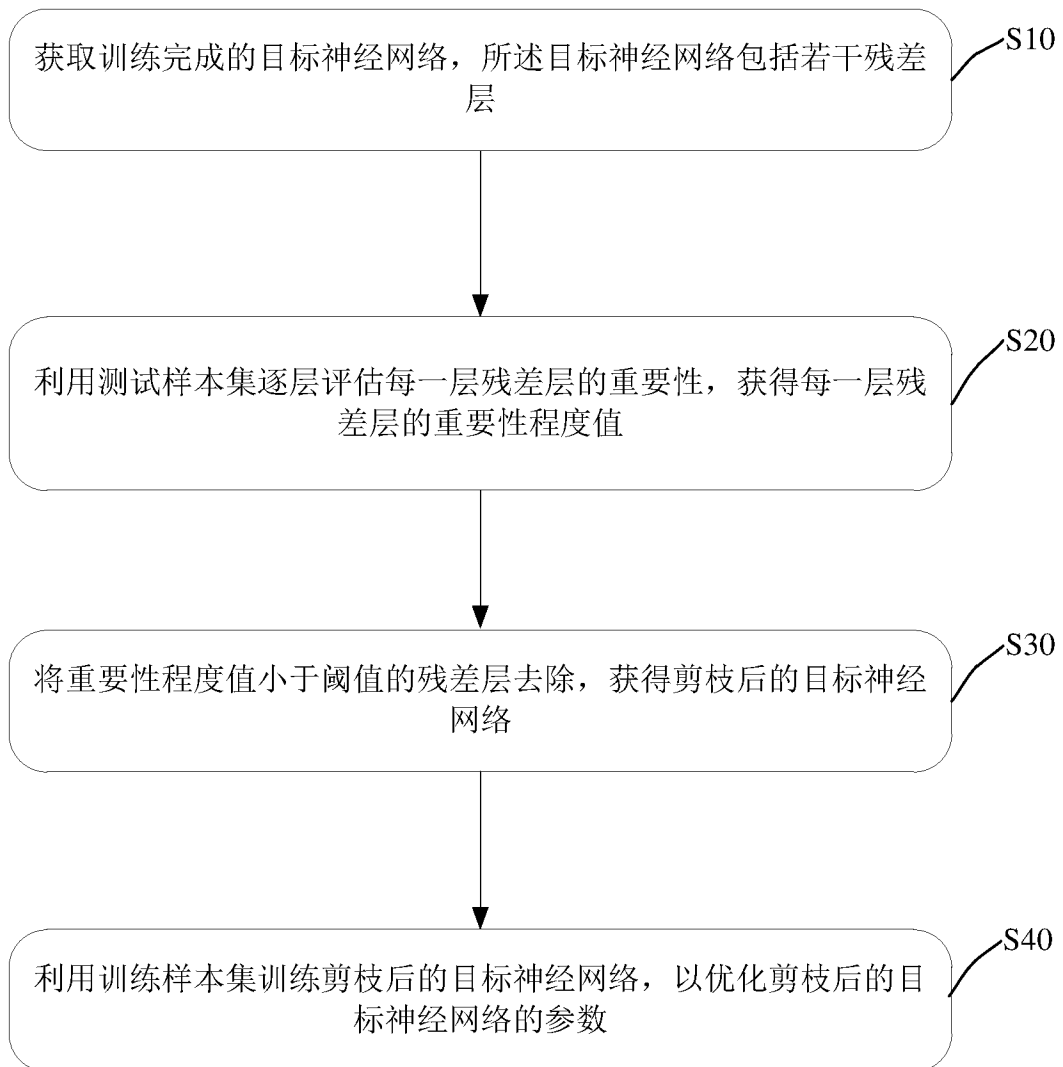


图 1

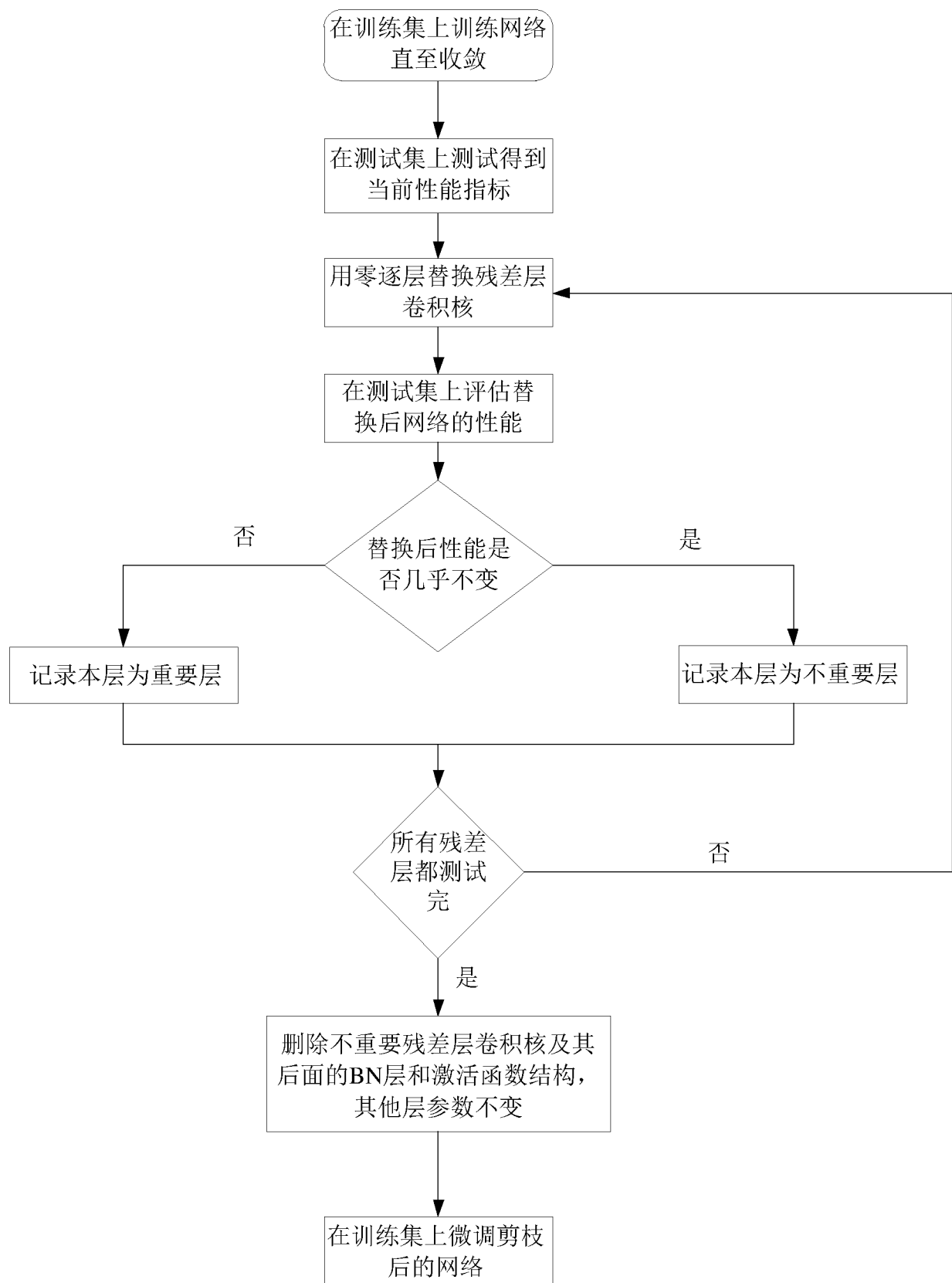


图 2

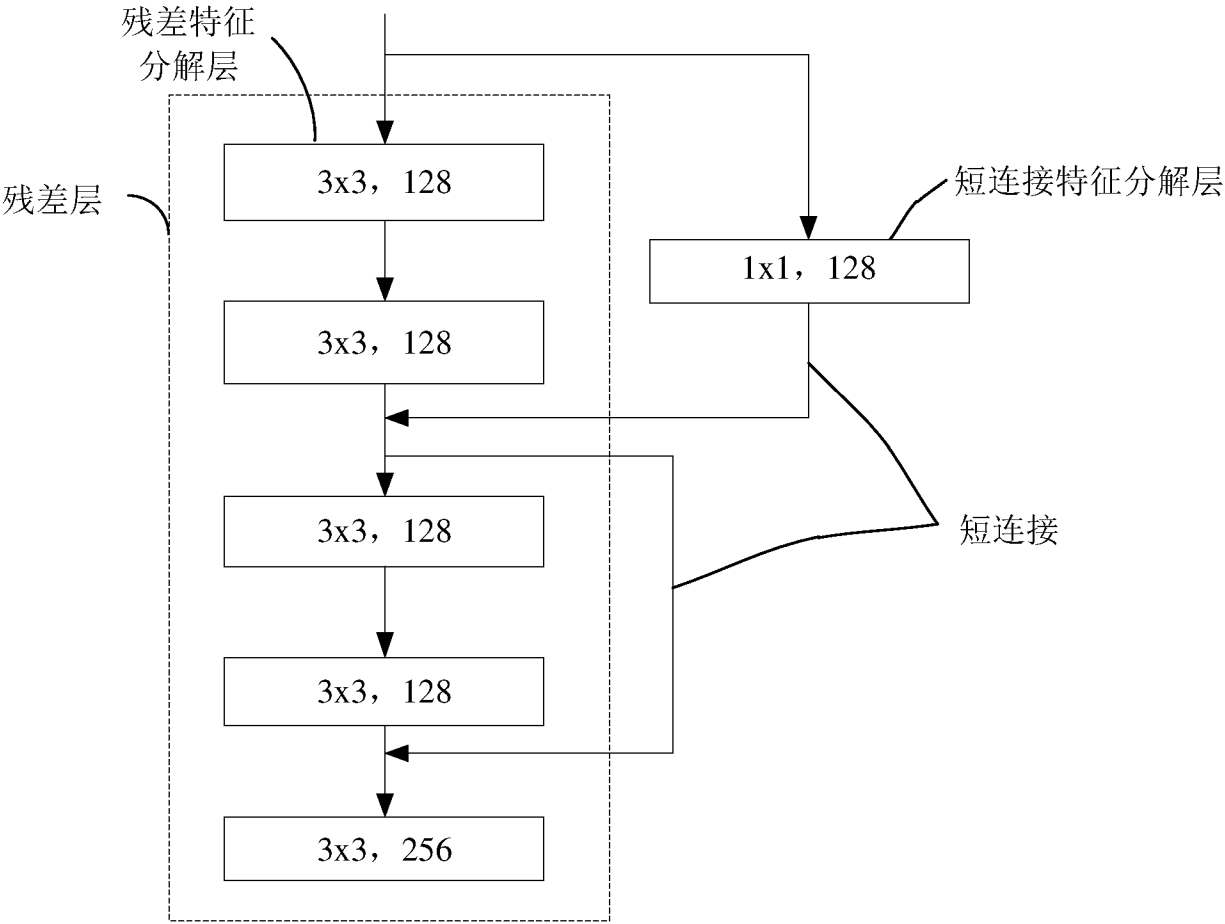


图 3

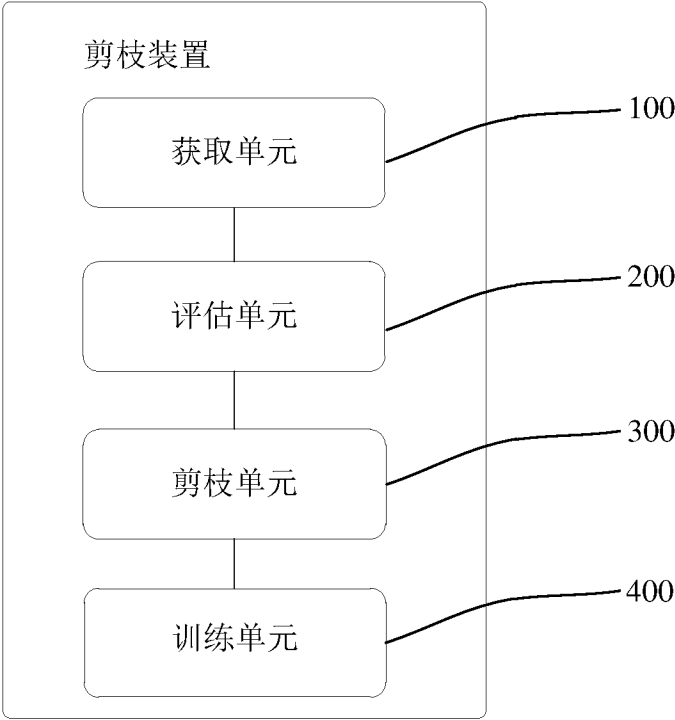


图 4

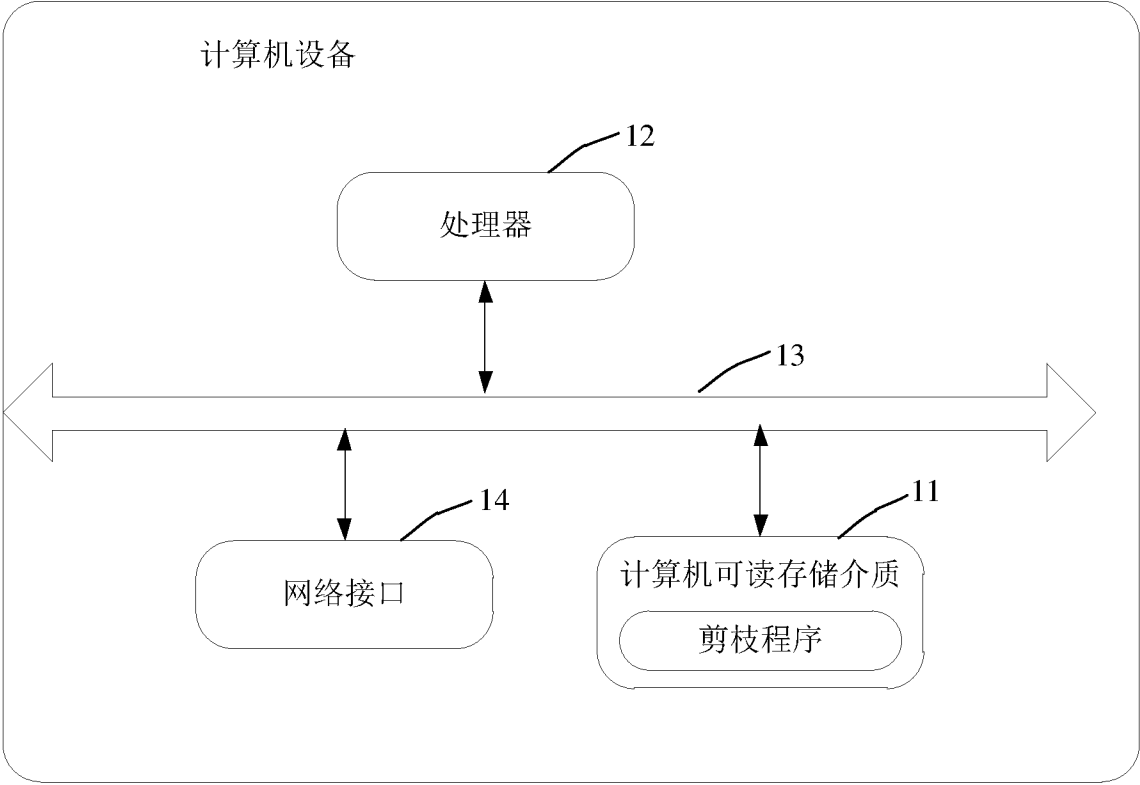


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/078455

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(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)

CNPAT, EPODOC, WPI, CNKI, ISI: 剪枝, 神经网络, 训练, 残差, 测试, 样本, 重要性, 阈值, 参数, 优化, 卷积核, pruning, neural, network, training, residual, test, sample, important, threshold, parameter, optimize, covolution, kernel

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 112052951 A (BEIJING SMARTER EYE TECHNOLOGY CO., LTD.) 08 December 2020 (2020-12-08) description, paragraphs [0044]-[0104] and figures 1-3	1-12
Y	CN 111008640 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 14 April 2020 (2020-04-14) description, paragraphs [0066]-[0134] and figure 1	1-12
A	CN 110689113 A (ZHEJIANG UNIVERSITY) 14 January 2020 (2020-01-14) entire document	1-12
A	CN 112149797 A (OPPO (CHONGQING) INTELLIGENT TECHNOLOGY CO., LTD.) 29 December 2020 (2020-12-29) entire document	1-12
A	CN 111291637 A (SHANGHAI INSTITUTE OF MICROSYSTEM AND INFORMATION TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 16 June 2020 (2020-06-16) entire document	1-12
A	US 2019244103 A1 (ROYAL BANK OF CANADA) 08 August 2019 (2019-08-08) entire document	1-12



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

05 November 2021

Date of mailing of the international search report

26 November 2021

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/CN2021/078455

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112052951	A	08 December 2020	None			
CN	111008640	A	14 April 2020	None			
CN	110689113	A	14 January 2020	None			
CN	112149797	A	29 December 2020	None			
CN	111291637	A	16 June 2020	None			
US	2019244103	A1	08 August 2019	CA	3033014	A1	07 August 2019

国际检索报告

国际申请号

PCT/CN2021/078455

A. 主题的分类 G06N 3/08 (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)) CNPAT, EPDOC, WPI, CNKI, ISI: 剪枝, 神经网络, 训练, 残差, 测试, 样本, 重要性, 阈值, 参数, 优化, 卷积核, pruning, neural, network, training, residual, test, sample, important, threshold, parameter, optimize, covolution, kernel																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>Y</td> <td>CN 112052951 A (北京中科慧眼科技有限公司) 2020年 12月 8日 (2020 - 12 - 08) 说明书第[0044]-[0104]段及图1-3</td> <td>1-12</td> </tr> <tr> <td>Y</td> <td>CN 111008640 A (平安科技深圳有限公司) 2020年 4月 14日 (2020 - 04 - 14) 说明书第[0066]-[0134]段以及图1</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>CN 110689113 A (浙江大学) 2020年 1月 14日 (2020 - 01 - 14) 全文</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>CN 112149797 A (OPPO重庆智能科技有限公司) 2020年 12月 29日 (2020 - 12 - 29) 全文</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>CN 111291637 A (中国科学院上海微系统与信息技术研究所) 2020年 6月 16日 (2020 - 06 - 16) 全文</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>US 2019244103 A1 (ROYAL BANK OF CANADA) 2019年 8月 8日 (2019 - 08 - 08) 全文</td> <td>1-12</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	Y	CN 112052951 A (北京中科慧眼科技有限公司) 2020年 12月 8日 (2020 - 12 - 08) 说明书第[0044]-[0104]段及图1-3	1-12	Y	CN 111008640 A (平安科技深圳有限公司) 2020年 4月 14日 (2020 - 04 - 14) 说明书第[0066]-[0134]段以及图1	1-12	A	CN 110689113 A (浙江大学) 2020年 1月 14日 (2020 - 01 - 14) 全文	1-12	A	CN 112149797 A (OPPO重庆智能科技有限公司) 2020年 12月 29日 (2020 - 12 - 29) 全文	1-12	A	CN 111291637 A (中国科学院上海微系统与信息技术研究所) 2020年 6月 16日 (2020 - 06 - 16) 全文	1-12	A	US 2019244103 A1 (ROYAL BANK OF CANADA) 2019年 8月 8日 (2019 - 08 - 08) 全文	1-12
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
Y	CN 112052951 A (北京中科慧眼科技有限公司) 2020年 12月 8日 (2020 - 12 - 08) 说明书第[0044]-[0104]段及图1-3	1-12																					
Y	CN 111008640 A (平安科技深圳有限公司) 2020年 4月 14日 (2020 - 04 - 14) 说明书第[0066]-[0134]段以及图1	1-12																					
A	CN 110689113 A (浙江大学) 2020年 1月 14日 (2020 - 01 - 14) 全文	1-12																					
A	CN 112149797 A (OPPO重庆智能科技有限公司) 2020年 12月 29日 (2020 - 12 - 29) 全文	1-12																					
A	CN 111291637 A (中国科学院上海微系统与信息技术研究所) 2020年 6月 16日 (2020 - 06 - 16) 全文	1-12																					
A	US 2019244103 A1 (ROYAL BANK OF CANADA) 2019年 8月 8日 (2019 - 08 - 08) 全文	1-12																					
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td> * 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 </td> <td> “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 “&” 同族专利的文件 </td> </tr> </table>			* 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 “&” 同族专利的文件																			
* 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 “&” 同族专利的文件																						
国际检索实际完成的日期 2021年 11月 5日		国际检索报告邮寄日期 2021年 11月 26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 李婉怡 电话号码 86-(10)-53961408																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/078455

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	112052951	A	2020年 12月 8日	无	
CN	111008640	A	2020年 4月 14日	无	
CN	110689113	A	2020年 1月 14日	无	
CN	112149797	A	2020年 12月 29日	无	
CN	111291637	A	2020年 6月 16日	无	
US	2019244103	A1	2019年 8月 8日	CA 3033014 A1	2019年 8月 7日