



# (12)发明专利申请

(10)申请公布号 CN 111553480 A

(43)申请公布日 2020.08.18

(21)申请号 202010661954.2

(22)申请日 2020.07.10

(71)申请人 腾讯科技(深圳)有限公司

地址 518057 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 柳露艳 刘松伟 马锴 郑冶枫

(74)专利代理机构 深圳市隆天联鼎知识产权代  
理有限公司 44232

代理人 叶虹

(51)Int.Cl.

G06N 3/08(2006.01)

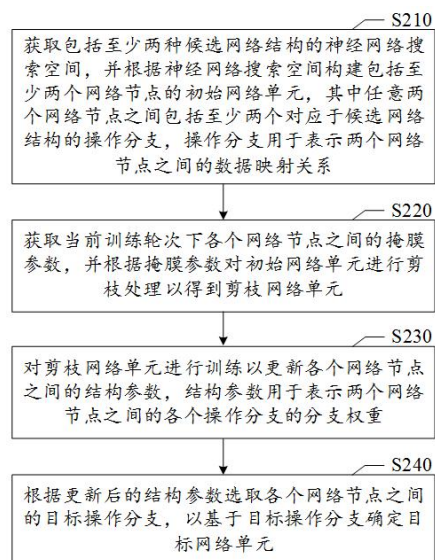
权利要求书3页 说明书18页 附图11页

## (54)发明名称

神经网络搜索方法、装置、计算机可读介质  
及电子设备

## (57)摘要

本申请属于人工智能技术领域,具体涉及一种神经网络搜索方法、神经网络搜索装置、计算机可读介质以及电子设备。该方法包括:获取包括至少两种候选网络结构的神经网络搜索空间,并根据神经网络搜索空间构建包括至少两个网络节点的初始网络单元;获取当前训练轮次下各个网络节点之间的掩膜参数,并根据掩膜参数对初始网络单元进行剪枝处理以得到剪枝网络单元;对剪枝网络单元进行训练以更新各个网络节点之间的结构参数;根据更新后的结构参数选取各个网络节点之间的目标操作分支,以基于目标操作分支确定目标网络单元。该方法可以降低计算成本,而且可以加速产生网络单元内各种操作之间的差异性,进而提高网络搜索的效率和准确性。



1. 一种神经网络搜索方法,其特征在于,包括:

获取包括至少两种候选网络结构的神经网络搜索空间,并根据所述神经网络搜索空间构建包括至少两个网络节点的初始网络单元,其中任意两个网络节点之间包括至少两个对应于所述候选网络结构的操作分支,所述操作分支用于表示两个网络节点之间的数据映射关系;

获取当前训练轮次下各个网络节点之间的掩膜参数,并根据所述掩膜参数对所述初始网络单元进行剪枝处理以得到剪枝网络单元;

对所述剪枝网络单元进行训练以更新所述各个网络节点之间的结构参数,所述结构参数用于表示两个网络节点之间的各个操作分支的分支权重;

根据更新后的结构参数选取各个网络节点之间的目标操作分支,以基于所述目标操作分支确定目标网络单元。

2. 根据权利要求1所述的神经网络搜索方法,其特征在于,所述获取当前训练轮次下各个网络节点之间的掩膜参数,包括:

获取当前训练轮次的轮次信息,并根据所述轮次信息确定当前训练轮次是否为掩膜参数更新轮次;

若当前训练轮次不是所述掩膜参数更新轮次,则将上一训练轮次的掩膜参数保持为当前训练轮次的掩膜参数;

若当前训练轮次是所述掩膜参数更新轮次,则根据当前训练轮次的剪枝率以及各个网络节点之间的结构参数确定当前训练轮次的掩膜参数,所述结构参数用于表示两个网络节点之间的各个操作分支的分支权重。

3. 根据权利要求2所述的神经网络搜索方法,其特征在于,所述根据当前训练轮次下的剪枝率以及各个网络节点之间的结构参数确定当前训练轮次的掩膜参数,包括:

获取当前训练轮次的轮次信息,并根据所述轮次信息确定当前训练轮次的剪枝率;

根据当前训练轮次的剪枝率以及各个网络节点之间的操作分支的分支数量确定所述操作分支的剪枝数量和保留数量;

根据各个网络节点之间的结构参数选取所述剪枝数量的操作分支作为剪枝分支并选取所述保留数量的操作分支作为保留分支;

为所述剪枝分支分配第一掩膜数值并为所述保留分支分配第二掩膜数值,以得到由所述第一掩膜数值和所述第二掩膜数值组成的当前训练轮次的掩膜参数。

4. 根据权利要求3所述的神经网络搜索方法,其特征在于,所述根据所述轮次信息确定当前训练轮次下的剪枝率,包括:

获取预设的迭代训练次数,并对所述迭代训练次数进行划分以得到多个剪枝率更新阶段,其中各个剪枝率更新阶段具有分段增大的剪枝率;

根据所述轮次信息确定当前训练轮次所在的剪枝率更新阶段,并将该剪枝率更新阶段的剪枝率确定为当前训练轮次的剪枝率。

5. 根据权利要求3所述的神经网络搜索方法,其特征在于,所述根据所述轮次信息确定当前训练轮次的剪枝率,包括:

获取用于在不同训练轮次下连续增大剪枝率的剪枝率更新模型;

通过所述剪枝率更新模型对所述轮次信息进行映射处理以得到当前训练轮次下的剪

枝率。

6. 根据权利要求3所述的神经网络搜索方法,其特征在于,所述根据各个网络节点之间的结构参数选取所述剪枝数量的操作分支作为剪枝分支并选取所述保留数量的操作分支作为保留分支,包括:

对各个网络节点之间的结构参数进行归一化映射以得到每个操作分支的选取概率;

按照所述选取概率将概率最低的所述剪枝数量的操作分支选取为剪枝分支并将概率最高的所述保留数量的操作分支选取为保留分支。

7. 根据权利要求1所述的神经网络搜索方法,其特征在于,所述根据所述掩膜参数对所述初始网络单元进行剪枝处理以得到剪枝网络单元,包括:

根据所述掩膜参数将所述初始网络单元中的部分操作分支确定为可更新分支,并将除所述可更新分支以外的其他操作分支确定为不可更新分支;

将所述可更新分支的网络参数和结构参数配置为可更新参数并将所述不可更新分支的网络参数和结构参数配置为不可更新参数;

将所述可更新分支以及所述不可更新分支组成的网络单元确定为剪枝网络单元。

8. 根据权利要求1所述的神经网络搜索方法,其特征在于,所述对所述剪枝网络单元进行训练以更新所述各个网络节点之间的结构参数,包括:

获取用于更新所述各个网络节点之间的结构参数的第一样本集合以及用于更新所述剪枝网络单元的网络参数的第二样本集合;

分别使用所述第一样本集合和所述第二样本集合对所述剪枝网络单元进行训练以交替更新所述各个网络节点之间的结构参数和所述剪枝网络单元的网络参数。

9. 根据权利要求8所述的神经网络搜索方法,其特征在于,所述分别使用所述第一样本集合和所述第二样本集合对所述剪枝网络单元进行训练以交替更新所述各个网络节点之间的结构参数和所述剪枝网络单元的网络参数,包括:

固定所述剪枝网络单元的网络参数,并使用所述第一样本集合对所述剪枝网络单元进行训练以更新所述各个网络节点之间的结构参数;

固定所述各个网络节点之间的结构参数,并使用所述第二样本集合对所述剪枝网络单元进行训练以更新所述剪枝网络单元的网络参数;

交替更新所述各个网络节点之间的结构参数以及所述剪枝网络单元的网络参数直至训练结束。

10. 根据权利要求9所述的神经网络搜索方法,其特征在于,所述使用所述第一样本集合对所述剪枝网络单元进行训练以更新所述各个网络节点之间的结构参数,包括:

通过所述剪枝网络单元对所述第一样本集合中的训练样本进行映射处理以得到与所述训练样本相关的损失误差;

根据所述损失误差确定所述各个网络节点之间的结构参数的结构参数误差梯度;

获取当前训练轮次的学习率,并根据所述学习率以及所述结构参数误差梯度更新所述各个网络节点之间的结构参数。

11. 根据权利要求9所述的神经网络搜索方法,其特征在于,所述使用所述第二样本集合对所述剪枝网络单元进行训练以更新所述剪枝网络单元的网络参数,包括:

通过所述剪枝网络单元对所述第二样本集合中的训练样本进行映射处理以得到与所

述训练样本相关的损失误差；

根据所述损失误差确定所述剪枝网络单元的网络参数的网络参数误差梯度；

获取当前训练轮次的学习率，并根据所述学习率以及所述网络参数误差梯度更新所述剪枝网络单元的网络参数。

12. 根据权利要求1所述的神经网络搜索方法，其特征在于，所述根据更新后的结构参数选取各个网络节点之间的目标操作分支，包括：

根据更新后的结构参数比较两个网络节点之间的操作分支的分支权重，并根据比较结果选取分支权重最大的一个操作分支作为候选操作分支；

分别确定与各个当前网络节点具有数据映射关系的前继网络节点，所述前继网络节点是通过操作分支向所述当前网络节点输入数据的网络节点；

比较所述当前网络节点与各个前继网络节点之间的候选操作分支的分支权重，并根据比较结果选取分支权重最大的一个或者多个候选操作分支作为目标操作分支。

13. 一种神经网络搜索装置，其特征在于，包括：

网络构建模块，被配置为获取包括至少两种候选网络结构的神经网络搜索空间，并根据所述神经网络搜索空间构建包括至少两个网络节点的初始网络单元，其中任意两个网络节点之间包括至少两个对应于所述候选网络结构的操作分支，所述操作分支用于表示两个网络节点之间的数据映射关系；

网络剪枝模块，被配置为获取当前训练轮次下各个网络节点之间的掩膜参数，并根据所述掩膜参数对所述网络单元进行剪枝处理以得到剪枝网络单元；

参数更新模块，被配置为对所述剪枝网络单元进行训练以更新所述各个网络节点之间的结构参数，所述结构参数用于表示两个网络节点之间的各个操作分支的分支权重；

网络确定模块，被配置为根据更新后的结构参数选取各个网络节点之间的目标操作分支，以基于所述目标操作分支确定目标网络单元。

14. 一种计算机可读介质，其上存储有计算机程序，该计算机程序被处理器执行时实现权利要求1至12中任意一项所述的神经网络搜索方法。

15. 一种电子设备，其特征在于，包括：

处理器；以及

存储器，用于存储所述处理器的可执行指令；

其中，所述处理器配置为经由执行所述可执行指令来执行权利要求1至12中任意一项所述的神经网络搜索方法。

## 神经网络搜索方法、装置、计算机可读介质及电子设备

### 技术领域

[0001] 本申请涉及人工智能技术领域,并涉及机器学习技术。具体而言,本申请涉及一种神经网络搜索方法、神经网络搜索装置、计算机可读介质以及电子设备。

### 背景技术

[0002] 人工神经网络是基于计算机技术模仿生物神经网络行为特征以进行分布式并行信息处理的算法数学模型。神经网络依靠系统的复杂程度,通过调整内部大量节点之间相互连接的关系,从而达到处理信息的目的。

[0003] 随着人工智能技术的发展,神经网络结构的设计,正在从手工设计转型为机器自动设计。神经网络结构搜索技术将网络结构和超参数的选择自动化,可以针对特定的数据集生成定制化的网络模型并达到超越专家模型的效果。然而,基于结构搜索的神经网络设计方案通常需要在数据量庞大的搜索空间中反复进行网络的筛选和训练,不仅需要耗费大量的计算资源,而且普遍存在搜索效率低、搜索结果不稳定等问题。

[0004] 需要说明的是,在上述背景技术部分公开的信息仅用于加强对本申请的背景的理解,因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

### 发明内容

[0005] 本申请的目的在于提供一种神经网络搜索方法、神经网络搜索装置、计算机可读介质以及电子设备,至少在一定程度上克服相关技术中存在的计算成本高、搜索效率低及稳定性差等技术问题。

[0006] 本申请的其他特性和优点将通过下面的详细描述变得显然,或部分地通过本申请的实践而习得。

[0007] 根据本申请实施例的一个方面,提供一种神经网络搜索方法,该方法包括:获取包括至少两种候选网络结构的神经网络搜索空间,并根据神经网络搜索空间构建包括至少两个网络节点的初始网络单元,其中任意两个网络节点之间包括至少两个对应于候选网络结构的分支,所述分支用于表示两个网络节点之间的数据映射关系;获取当前训练轮次下各个网络节点之间的掩膜参数,并根据所述掩膜参数对所述初始网络单元进行剪枝处理以得到剪枝网络单元;对所述剪枝网络单元进行训练以更新所述各个网络节点之间的结构参数,所述结构参数用于表示两个网络节点之间的各个分支的权重;根据更新后的结构参数选取各个网络节点之间的目标分支,以基于所述目标分支确定目标网络单元。

[0008] 根据本申请实施例的一个方面,提供一种神经网络搜索装置,该装置包括:网络构建模块,被配置为构建包括至少两个网络节点的初始网络单元,其中任意两个网络节点之间包括至少两个操作分支,所述分支用于表示两个网络节点之间的数据映射关系;网络剪枝模块,被配置为获取当前训练轮次下各个网络节点之间的掩膜参数,并根据所述掩膜参数对所述网络单元进行剪枝处理以得到剪枝网络单元;参数更新模块,被配置为对所

述剪枝网络单元进行训练以更新所述各个网络节点之间的结构参数,所述结构参数用于表示两个网络节点之间的各个操作分支的分支权重;网络确定模块,被配置为根据更新后的结构参数选取各个网络节点之间的目标操作分支,以基于所述目标操作分支确定目标网络单元。

[0009] 在本申请的一些实施例中,基于以上技术方案,所述网络剪枝模块包括:轮次确定单元,被配置为获取当前训练轮次的轮次信息,并根据所述轮次信息确定当前训练轮次是否为掩膜参数更新轮次;掩膜参数确定单元,被配置为若当前训练轮次不是所述掩膜参数更新轮次,则将上一训练轮次的掩膜参数保持为当前训练轮次的掩膜参数;掩膜参数更新单元,被配置为若当前训练轮次是所述掩膜参数更新轮次,则根据当前训练轮次的剪枝率以及各个网络节点之间的结构参数确定当前训练轮次的掩膜参数,所述结构参数用于表示两个网络节点之间的各个操作分支的分支权重。

[0010] 在本申请的一些实施例中,基于以上技术方案,所述掩膜参数更新单元包括:剪枝率确定子单元,被配置为获取当前训练轮次的轮次信息,并根据所述轮次信息确定当前训练轮次的剪枝率;分支数量确定子单元,被配置为根据当前训练轮次的剪枝率以及各个网络节点之间的操作分支的分支数量确定所述操作分支的剪枝数量和保留数量;操作分支选取子单元,被配置为根据各个网络节点之间的结构参数选取所述剪枝数量的操作分支作为剪枝分支并选取所述保留数量的操作分支作为保留分支;掩膜参数确定子单元,被配置为为所述剪枝分支分配第一掩膜数值并为所述保留分支分配第二掩膜数值,以得到由所述第一掩膜数值和所述第二掩膜数值组成的当前训练轮次的掩膜参数。

[0011] 在本申请的一些实施例中,基于以上技术方案,所述剪枝率确定子单元包括:阶段划分子单元,被配置为获取预设的迭代训练次数,并对所述迭代训练次数进行划分以得到多个剪枝率更新阶段,其中各个剪枝率更新阶段具有分段增大的剪枝率;第一剪枝率确定子单元,被配置为根据所述轮次信息确定当前训练轮次所在的剪枝率更新阶段,并将该剪枝率更新阶段的剪枝率确定为当前训练轮次的剪枝率。

[0012] 在本申请的一些实施例中,基于以上技术方案,所述剪枝率确定子单元包括:模型获取子单元,被配置为获取用于在不同训练轮次下连续增大剪枝率的剪枝率更新模型;第二剪枝率确定子单元,被配置为通过所述剪枝率更新模型对所述轮次信息进行映射处理以得到当前训练轮次下的剪枝率。

[0013] 在本申请的一些实施例中,基于以上技术方案,所述操作分支选取子单元包括:参数映射子单元,被配置为对各个网络节点之间的结构参数进行归一化映射以得到每个操作分支的选取概率;概率选取子单元,被配置为按照所述选取概率将概率最低的所述剪枝数量的操作分支选取为剪枝分支并将概率最高的所述保留数量的操作分支选取为保留分支。

[0014] 在本申请的一些实施例中,基于以上技术方案,所述网络剪枝模块还包括:分支分类单元,被配置为根据所述掩膜参数将所述初始网络单元中的部分操作分支确定为可更新分支,并将除所述可更新分支以外的其他操作分支确定为不可更新分支;参数配置单元,被配置为将所述可更新分支的网络参数和结构参数配置为可更新参数并将所述不可更新分支的网络参数和结构参数配置为不可更新参数;分支组合单元,被配置为将所述可更新分支以及所述不可更新分支组成的网络单元确定为剪枝网络单元。

[0015] 在本申请的一些实施例中,基于以上技术方案,所述参数更新模块包括:样本集合

获取单元,被配置为获取用于更新所述各个网络节点之间的结构参数的第一样本集合以及用于更新所述剪枝网络单元的网络参数的第二样本集合;参数交替更新单元,被配置为分别使用所述第一样本集合和所述第二样本集合对所述剪枝网络单元进行训练以交替更新所述各个网络节点之间的结构参数和所述剪枝网络单元的网络参数。

[0016] 在本申请的一些实施例中,基于以上技术方案,所述参数交替更新单元包括:结构参数更新子单元,被配置为固定所述剪枝网络单元的网络参数,并使用所述第一样本集合对所述剪枝网络单元进行训练以更新所述各个网络节点之间的结构参数;网络参数更新子单元,被配置为固定所述各个网络节点之间的结构参数,并使用所述第二样本集合对所述剪枝网络单元进行训练以更新所述剪枝网络单元的网络参数;参数交替更新子单元,被配置为交替更新所述各个网络节点之间的结构参数以及所述剪枝网络单元的网络参数直至训练结束。

[0017] 在本申请的一些实施例中,基于以上技术方案,所述结构参数更新子单元包括:第一网络映射子单元,被配置为通过所述剪枝网络单元对所述第一样本集合中的训练样本进行映射处理以得到与所述训练样本相关的损失误差;第一梯度确定子单元,被配置为根据所述损失误差确定所述各个网络节点之间的结构参数的结构参数误差梯度;第一参数更新子单元,被配置为获取当前训练轮次的学习率,并根据所述学习率以及所述结构参数误差梯度更新所述各个网络节点之间的结构参数。

[0018] 在本申请的一些实施例中,基于以上技术方案,所述网络参数更新子单元包括:第二网络映射子单元,被配置为通过所述剪枝网络单元对所述第二样本集合中的训练样本进行映射处理以得到与所述训练样本相关的损失误差;第二梯度确定子单元,被配置为根据所述损失误差确定所述剪枝网络单元的网络参数的网络参数误差梯度;第二参数更新子单元,被配置为获取当前训练轮次的学习率,并根据所述学习率以及所述网络参数误差梯度更新所述剪枝网络单元的网络参数。

[0019] 在本申请的一些实施例中,基于以上技术方案,所述网络确定模块包括:分支候选单元,被配置为根据更新后的结构参数比较两个网络节点之间的操作分支的分支权重,并根据比较结果选取分支权重最大的一个操作分支作为候选操作分支;节点确定单元,被配置为分别确定与各个当前网络节点具有数据映射关系的前继网络节点,所述前继网络节点是通过操作分支向所述当前网络节点输入数据的网络节点;分支选取单元,被配置为比较所述当前网络节点与各个前继网络节点之间的候选操作分支的分支权重,并根据比较结果选取分支权重最大的一个或者多个候选操作分支作为目标操作分支。

[0020] 根据本申请实施例的一个方面,提供一种计算机可读介质,其上存储有计算机程序,该计算机程序被处理器执行时实现如以上技术方案中的神经网络搜索方法。

[0021] 根据本申请实施例的一个方面,提供一种电子设备,该电子设备包括:处理器;以及存储器,用于存储所述处理器的可执行指令;其中,所述处理器被配置为经由执行所述可执行指令来执行如以上技术方案中的神经网络搜索方法。

[0022] 根据本申请的一个方面,提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述各个技术方案中提供的神经网络搜索方法。

[0023] 在本申请实施例提供的技术方案中,首先构建可以重复使用的网络单元,再对其进行搜索训练以得到目标网络单元,然后将目标网络单元堆叠组成神经网络,在对其进行验证训练后得到可用的神经网络。在搜索训练目标网络单元的过程中,基于掩膜参数对网络单元进行剪枝处理可以降低网络训练的参数规模,即仅对网络单元中的部分参数进行更新,从而可以降低计算成本,而且可以加速产生网络单元内各种操作之间的差异性,进而提高网络搜索的效率和准确性。

[0024] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本申请。

## 附图说明

[0025] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本申请的实施例,并与说明书一起用于解释本申请的原理。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。

[0026] 图1示意性地示出了应用本申请技术方案的示例性系统架构框图。

[0027] 图2示意性地示出了本申请一些实施例中的神经网络搜索方法的步骤流程图。

[0028] 图3示意性地示出了一种网络单元的组成结构示意图。

[0029] 图4示意性地示出了本申请一些实施例中更新确定掩膜参数的方法步骤流程图。

[0030] 图5示意性地示出了本申请一些实施例中选取目标操作分支的方法步骤流程图。

[0031] 图6示意性地示出了在一应用场景下由目标网络单元堆叠形成神经网络的网络结构。

[0032] 图7A示意性地示出了在第k个训练轮次中的网络单元结构及其内部的软剪枝效果。

[0033] 图7B示意性地示出了在第k+1个训练轮次中的网络单元结构及其内部的软剪枝效果。

[0034] 图8A示出了本申请技术方案在一应用场景中搜索得到的网络单元Normal Cell(上)和网络单元Reduction Cell(下)的网络结构示意图。

[0035] 图8B示出了应用现有技术中的DARTS算法搜索得到的网络单元Normal Cell(上)和网络单元Reduction Cell(下)的网络结构示意图。

[0036] 图9示意性地示出了本申请实施例提供的神经网络搜索装置的结构框图。

[0037] 图10示意性地示出了适于用来实现本申请实施例的电子设备的计算机系统结构框图。

## 具体实施方式

[0038] 现在将参考附图更全面地描述示例实施方式。然而,示例实施方式能够以多种形式实施,且不应被理解为限于在此阐述的范例;相反,提供这些实施方式使得本申请将更加全面和完整,并将示例实施方式的构思全面地传达给本领域的技术人员。

[0039] 此外,所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施例中。在下面的描述中,提供许多具体细节从而给出对本申请的实施例的充分理解。然而,

本领域技术人员将意识到,可以实践本申请的技术方案而没有特定细节中的一个或更多,或者可以采用其它的方法、组元、装置、步骤等。在其它情况下,不详细示出或描述公知方法、装置、实现或者操作以避免模糊本申请的各方面。

[0040] 附图中所示的方框图仅仅是功能实体,不一定必须与物理上独立的实体相对应。即,可以采用软件形式来实现这些功能实体,或在一个或多个硬件模块或集成电路中实现这些功能实体,或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0041] 附图中所示的流程图仅是示例性说明,不是必须包括所有的内容和操作/步骤,也不是必须按所描述的顺序执行。例如,有的操作/步骤还可以分解,而有的操作/步骤可以合并或部分合并,因此实际执行的顺序有可能根据实际情况改变。

[0042] 在对本申请实施例提供的神经网络搜索方法、神经网络搜索装置等技术方案作出详细说明之前,先对本申请涉及的人工智能技术进行简单介绍。

[0043] 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。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0044] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。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、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。

[0045] 计算机视觉技术(Computer Vision, CV)计算机视觉是一门研究如何使机器“看”的科学,更进一步的说,就是指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等机器视觉,并进一步做图形处理,使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科,计算机视觉研究相关的理论和技术,试图建立能够从图像或者多维数据中获取信息的人工智能系统。计算机视觉技术通常包括图像处理、图像识别、图像语义理解、图像检索、OCR、视频处理、视频语义理解、视频内容/行为识别、三维物体重建、3D技术、虚拟现实、增强现实、同步定位与地图构建等技术,还包括常见的人脸识别、指纹识别等生物特征识别技术。

[0046] 机器学习(Machine Learning, ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0047] 随着人工智能技术研究和进步,人工智能技术在多个领域展开研究和应用,例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服等,相信随着技术的发展,人工智能技术将在更多的领域得到应用,并发挥越来越重要的价值。

[0048] 本申请实施例涉及了人工智能技术中的神经网络结构搜索技术(Neural

Architecture Search, NAS), NAS算法是用来设计神经网络的策略。具体是通过人为设定一定的搜索空间,设计搜索策略去在搜索空间中寻找在验证集上表现最好的网络结构。在NAS算法中,基于人为设定的搜索空间可以构造一个超网(super-net),超网是在神经网络结构搜索过程中包含所有可能的子网络的集合。通过设计一个大的搜索空间,这个搜索空间便组成一个超网,里面包含多个子网,每个子网经过训练后都可以被评测性能指标,NAS算法所做的便是如何从这些子网中找出性能指标最好的子网。一个子网要想被评测就需要先被训练,这个过程非常耗时,对搜索空间中的所有子网进行训练然后再评估是低效且无法实现的,各类NAS算法的主要区别便是如何快速高效评估子网,搜索子网。

[0049] 本申请针对现有的可微分NAS算法存在的收敛速度慢、搜索结果不稳定等问题,提出了一种软剪枝加速的可微分神经网络架构搜索算法。该算法在可微分架构搜索过程中应用软剪枝技术,在保证模型容量相对不变的情况下对搜索单元中的候选操作实行剪枝,每次仅激活和更新一定比例的候选操作和其权重参数,加快操作之间偏差的产生,使搜索进程加快,搜索结果不确定性更小,结果方差更小而且准确率也更高。

[0050] 图1示意性地示出了应用本申请技术方案的示例性系统架构框图。

[0051] 如图1所示,系统架构100可以包括终端设备110、网络120和服务器130。终端设备110可以包括智能手机、平板电脑、笔记本电脑、台式电脑等各种电子设备。服务端130可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云计算服务的云服务器。网络120可以是能够在终端设备110和服务器130之间提供通信链路的各种连接类型的通信介质,例如可以是有线通信链路或者无线通信链路。

[0052] 根据实现需要,本申请实施例中的系统架构可以具有任意数目的客户端、网络和服务端。例如,服务器130可以由多个服务器设备组成的服务器群组。另外,本申请实施例提供的技术方案可以应用于终端设备110,也可以应用于服务器130,或者可以由终端设备110和服务器130共同实施,本申请对此不做特殊限定。

[0053] 举例而言,本申请实施例可以在终端设备110上进行神经网络搜索和训练,训练得到的神经网络模型可以直接布局在终端设备110上进行应用。本申请实施例也可以在服务器130上进行神经网络搜索和训练,训练得到的神经网络模型可以布局在终端设备110上或者布局在服务器130上。

[0054] 本申请实施例提供的神经网络搜索方法等技术方案可以应用于各种自动机器学习(AutoML)平台,为需要深度学习服务的用户提供模型与数据解耦的算法服务,特别是针对分类场景有比较好的适用性。例如,本申请实施例提供的技术方案可以广泛应用于诸如自动驾驶等终端场景。另外也可以进一步应用到分割领域,比如医学图像分割,未来在终端部署医学深度学习服务背后离不开高性能的网络压缩或者搜索算法来寻找更为轻量级的高效模型。

[0055] 下面结合具体实施方式对本申请实施例提供的神经网络搜索方法、神经网络搜索装置、计算机可读介质以及计算设备等技术方案做出详细说明。

[0056] 图2示意性地示出了本申请一些实施例中的神经网络搜索方法的步骤流程图,该神经网络方法可以由图1所示的终端设备110执行,也可以由服务器130执行,或者可以由终端设备110和服务器130共同执行。如图2所示,该神经网络搜索方法主要可以包括如下的步骤S210~步骤S240。

[0057] 步骤S210:获取包括至少两种候选网络结构的神经网络搜索空间,并根据神经网络搜索空间构建包括至少两个网络节点的初始网络单元,其中任意两个网络节点之间包括至少两个对应于候选网络结构的操作分支,操作分支用于表示两个网络节点之间的数据映射关系。

[0058] 步骤S220:获取当前训练轮次下各个网络节点之间的掩膜参数,并根据掩膜参数对初始网络单元进行剪枝处理以得到剪枝网络单元。

[0059] 步骤S230:对剪枝网络单元进行训练以更新各个网络节点之间的结构参数,结构参数用于表示两个网络节点之间的各个操作分支的分支权重。

[0060] 步骤S240:根据更新后的结构参数选取各个网络节点之间的目标操作分支,以基于目标操作分支确定目标网络单元。

[0061] 在本申请实施例提供的神经网络搜索方法中,通过构建可以重复使用的网络单元,再对其进行搜索训练可以得到目标网络单元,然后将目标网络单元堆叠组成神经网络,在对其进行验证训练后得到可用的神经网络。在搜索训练目标网络单元的过程中,基于掩膜参数对网络单元进行剪枝处理可以降低网络训练的参数规模,即仅对网络单元中的部分参数进行更新,从而可以降低计算成本,而且可以加速产生网络单元内各种操作之间的差异性,进而提高网络训练的效率和准确性。使用本申请实施例提供的神经网络搜索方法可以助力深度学习在边缘计算场景的落地,降低AI在终端落地的门槛。作为一种高效的神经网络压缩技术,通过设计轻量级的搜索空间,以及添加针对平台延时和计算量的损失约束,可以搜索出针对特定平台定制化的模型结构,其具有较小的参数量和较低的运算量,较快的推理速度,因而可以降低神经网络的训练和使用门槛,也避免了人为设计的局限性。

[0062] 下面分别对以上实施例中的各个步骤的实施方式做详细说明。

[0063] 在步骤S210中,获取包括至少两种候选网络结构的神经网络搜索空间,并根据神经网络搜索空间构建包括至少两个网络节点的网络单元,其中任意两个网络节点之间包括至少两个对应于候选网络结构的操作分支,操作分支用于表示两个网络节点之间的数据映射关系。

[0064] 网络单元是用于组成神经网络的重复单元,对其进行搜索训练的目标是确定网络单元中包含哪些操作以及各个网络节点如何连接。在搜索得到目标网络节点后,可以对其重复堆叠以形成深度神经网络。图3示意性地示出了一种网络单元的组成结构示意图。如图3所示,初始网络单元301包括自上向下依次分布的四个网络节点,即节点0、节点1、节点2和节点3。任意两个网络节点之间均通过多个操作分支进行连接,每个操作分支对应于神经网络搜索空间中的一种候选网络结构,表示一种从前端的网络节点向后端的网络节点进行的数据映射操作,网络单元中的网络节点通过操作分支进行连接形成有向无环图(DAG)。表1示出了由多种候选网络结构组成的神经网络搜索空间示例,每个候选网络结构用于实现一种操作类型。

[0065] 表1 搜索空间示例

| 类型编号 | 操作类型(Operation)          |
|------|--------------------------|
| 0    | 零操作(Zero)                |
| 1    | 残差连接(Skip-connect)       |
| 2    | 3x3最大池化(3x3 max pooling) |

|   |                                                |
|---|------------------------------------------------|
| 3 | 3x3平均池化(3x3 average pooling)                   |
| 4 | 3x3可分离卷积(3x3 separable convolutions)           |
| 5 | 5x5可分离卷积(5x5 separable convolutions)           |
| 6 | 3x3空洞可分离卷积(3x3 dilated separable convolutions) |
| 7 | 5x5空洞可分离卷积(5x5 dilated separable convolutions) |

在初始网络单元301中,每两个网络节点之间均通过神经网络搜索空间中的所有候选网络结构对应的操作分支建立连接,这些操作分支组合在一起可以形成一个混合操作mixed-operation。例如,节点0和节点1之间通过对应于8种操作类型的8个操作分支进行连接,那么可以将节点0对应的特征图进行8中计算操作后再进行加权求和得到节点1对应的特征图,其中每个操作分支的分支权重代表了该操作分支的重要程度。

[0066] 如图3中所示,通过步骤S210建立的初始网络单元301是由神经的网络搜索空间中全量的操作类型形成的超网络(super-net),任意两个网络节点之间都包含了所有的操作类型,而本申请实施例对其进行搜索训练的目标是不断更新任意两个网络节点之间的各个操作分支的重要程度,从而保留重要性程度高的部分操作分支以得到结构简化的目标网络单元302。

[0067] 在步骤S220中,获取当前训练轮次下各个网络节点之间的掩膜参数,并根据掩膜参数对初始网络单元进行剪枝处理以得到剪枝网络单元。

[0068] 掩膜参数是用于在搜索训练过程中对初始网络单元进行剪枝处理的参数,基于掩膜参数进行的剪枝处理是一种软剪枝机制,即各个操作分支在搜索训练过程中并不会被丢弃,而是仅仅在网络的反向传播过程中对其进行更新屏蔽。基于掩膜参数可以在初始网络单元中选取一部分操作分支进行参数更新,而未被选中的操作分支将保持其相关参数不被更新。

[0069] 本申请实施例采用的软剪枝机制本质上是一种对结构参数的稀疏正则化,通过设置一定的剪枝比例,每次仅选择每个mixed-operation中的一定的操作分支进行权重参数和结构参数的训练更新,选择的准则可以是对应的分支权重越大越容易被选中,产生的效果就是大的更大,小的更小,从而可以加速操作分支之间分支权重的差异性的产生,加速搜索过程的收敛,减小搜索训练阶段和验证训练阶段的性能差异,并减小搜索训练结果的方差。

[0070] 同时软剪枝机制在不同的迭代步骤会重新选择新的剪枝分支,之前被剪掉的操作分支可能会被重新激活更新,这在一定程度上保持了搜索过程中的网络容量,同时也会削弱某些初始化占优的操作分支的优势,进而使神经网络的收敛选择更加准确。

[0071] 掩膜参数一经确定也并非一成不变,本申请实施例可以预先配置一定的剪枝间隔,基于该剪枝间隔可以对掩膜参数进行更新。

[0072] 在本申请的一些可选的实施方式中,步骤S220中的获取当前训练轮次下各个网络节点之间的掩膜参数,可以进一步包括如下的步骤S221~步骤S223。

[0073] 步骤S221:获取当前训练轮次的轮次信息,并根据轮次信息确定当前训练轮次是否为掩膜参数更新轮次。

[0074] 步骤S222:若当前训练轮次不是掩膜参数更新轮次,则将上一训练轮次的掩膜参数保持为当前训练轮次的掩膜参数。

[0075] 步骤S223:若当前训练轮次是掩膜参数更新轮次,则根据当前训练轮次的剪枝率以及各个网络节点之间的结构参数确定当前训练轮次的掩膜参数,结构参数用于表示两个网络节点之间的各个操作分支的分支权重。

[0076] 通过设置一定的剪枝间隔,可以将部分训练轮次确定为掩膜参数更新轮次,以便在达到一定的训练次数后对掩膜参数做出更新。如果当前训练轮次不是掩膜参数更新轮次,那么需要延续使用上一训练轮次的掩膜参数。而如果当前训练轮次是掩膜参数更新轮次,那么需要根据经过多轮训练更新后的结构参数对当前的掩膜参数做出更新,以便调整操作分支的剪枝方式。通过更新掩膜参数,可以提高网络训练的灵活性,原本被更新屏蔽的操作分支可能新的剪枝步骤中被选中更新,一方面可以保证神经网络被充分地搜索训练,另一方面不会过多地影响神经网络的模型容量。

[0077] 掩膜参数可以是根据预设的剪枝率以及当前训练轮次下各个网络节点之间的结构参数来进行确定和更新的。图4示意性地示出了本申请一些实施例中更新确定掩膜参数的方法步骤流程图。如图4所示,在以上实施例的基础上,根据当前训练轮次的剪枝率以及各个网络节点之间的结构参数确定当前训练轮次的掩膜参数,可以包括如下的步骤S410~步骤S440。

[0078] 步骤S410:获取当前训练轮次的轮次信息,并根据轮次信息确定当前训练轮次的剪枝率。

[0079] 剪枝率用于表示对两个网络节点之间的操作分支进行剪枝处理的剪枝比例,不同的训练轮次下可以具有相同的剪枝率,也可以具有不同的剪枝率。

[0080] 步骤S420:根据当前训练轮次的剪枝率以及各个网络节点之间的操作分支的分支数量确定操作分支的剪枝数量和保留数量。

[0081] 例如,当前训练轮次的剪枝率为75%,两个网络节点之间的操作分支的分支数量为8,那么由此可以计算确定剪枝数量为 $75\% \times 8 = 6$ ,相应的保留数量为 $(1 - 75\%) \times 8 = 2$ 。

[0082] 步骤S430:根据各个网络节点之间的结构参数选取剪枝数量的操作分支作为剪枝分支并选取保留数量的操作分支作为保留分支。

[0083] 对各个网络节点之间的结构参数可以进行归一化映射以得到每个操作分支的选取概率。例如通过softmax函数对各个操作分支的分支权重进行映射处理可以得到数值在0~1之间的各个操作分支的选取概率。按照选取概率将概率最低的剪枝数量的操作分支选取为剪枝分支并将概率最高的保留数量的操作分支选取为保留分支。

[0084] 步骤S440:为剪枝分支分配第一掩膜数值并为保留分支分配第二掩膜数值,以得到由第一掩膜数值和第二掩膜数值组成的当前训练轮次的掩膜参数。

[0085] 掩膜参数可以由第一掩膜数值和第二掩膜数值组成的一维向量,例如第一掩膜数值为0,第二掩膜数值为1,那么掩膜参数可以由0和1组成的向量,向量形式的掩膜参数可以直接作为输入参数参与至搜索训练的参数更新过程。

[0086] 在本申请的一些实施例中,根据掩膜参数将网络单元中的部分操作分支确定为可更新分支,并将除可更新分支以外的其他操作分支确定为不可更新分支。在此基础上,将可更新分支的网络参数和结构参数配置为可更新参数并将不可更新分支的网络参数和结构参数配置为不可更新参数,进而将可更新分支以及不可更新分支组成的网络单元确定为剪枝网络单元。举例而言,掩膜参数为一个一维向量(0, 1, 0, 1),那么第一个和第三个操作

分支将在第一掩膜数值0的作用下而被确定为不可更新分支,与之对应的网络参数和结构参数也将被配置为不可更新参数;第二个和第四个操作分支将在第二掩膜数值1的作用下而被确定为可更新分支,与之对应的网络参数和结构参数将被配置为可更新参数。

[0087] 为了获得更加稳定的剪枝搜索过程并提高剪枝搜索效率,本申请实施例可以根据预先设置的剪枝率更新策略对不同训练轮次下的剪枝率进行更新。

[0088] 在一些可选的实施方式中,可以对剪枝率进行分段增大更新。具体可以先获取获取预设的迭代训练次数,并对迭代训练次数进行划分以得到多个剪枝率更新阶段,其中各个剪枝率更新阶段具有分段增大的剪枝率;根据轮次信息确定当前训练轮次所在的剪枝率更新阶段,并将该剪枝率更新阶段的剪枝率确定为当前训练轮次的剪枝率。举例而言,一种可选的剪枝率更新策略可以是:根据预设的迭代训练次数 $K$ 划分出四个剪枝率更新阶段,即 $0 \sim K/4$ 、 $K/4 \sim K/2$ 、 $K/2 \sim 3K/4$ 以及 $3K/4 \sim K$ ,在 $0$ 、 $K/4$ 、 $K/2$ 和 $3K/4$ 四个训练轮次下分别更新剪枝率为 $0$ 、 $25\%$ 、 $50\%$ 和 $75\%$ ,从而在不同的剪枝率更新阶段内使用不同的剪枝率进行剪枝操作。

[0089] 在另一些可选的实施方式中,可以对剪枝率进行连续增大更新。具体可以先获取用于在不同训练轮次下连续增大剪枝率的剪枝率更新模型,然后再通过剪枝率更新模型对轮次信息进行映射处理以得到当前训练轮次下的剪枝率。

[0090] 基于分段增大或者连续增大的剪枝率更新方案,在搜索训练的初期可以设置较低的剪枝率以使得各个操作分支均能得到一定程度的训练,避免参数初始化带来的性能偏差对搜索过程造成干扰。

[0091] 在步骤S230中,对剪枝网络单元进行训练以更新各个网络节点之间的结构参数,结构参数用于表示两个网络节点之间的各个操作分支的分支权重。

[0092] 在本申请的一些可选的实施方式中,可以采用交替更新网络参数和结构参数的方式实现对剪枝网络单元的搜索训练。具体而言,可以先获取用于更新各个网络节点之间的结构参数的第一样本集合以及用于更新剪枝网络单元的网络参数的第二样本集合;然后分别使用第一样本集合和第二样本集合对剪枝网络单元进行训练以交替更新各个网络节点之间的结构参数和剪枝网络单元的网络参数。

[0093] 在结构参数的更新轮次下,固定剪枝网络单元的网络参数,并使用第一样本集合对剪枝网络单元进行训练以更新各个网络节点之间的结构参数;在网络参数的更新轮次下,固定各个网络节点之间的结构参数,并使用第二样本集合对剪枝网络单元进行训练以更新剪枝网络单元的网络参数;重复执行对结构参数和网络参数的更新步骤,交替更新各个网络节点之间的结构参数以及剪枝网络单元的网络参数直至训练结束。

[0094] 本申请实施例可以采用梯度下降法分别对结构参数和网络参数进行更新。

[0095] 在结构参数的更新轮次下,通过剪枝网络单元对第一样本集合中的训练样本进行映射处理以得到与训练样本相关的损失误差;根据损失误差确定各个网络节点之间的结构参数的结构参数误差梯度;获取当前训练轮次的学习率,并根据学习率以及结构参数误差梯度更新各个网络节点之间的结构参数。

[0096] 在网络参数的更新轮次下,通过剪枝网络单元对第二样本集合中的训练样本进行映射处理以得到与训练样本相关的损失误差;根据损失误差确定剪枝网络单元的网络参数的网络参数误差梯度;获取当前训练轮次的学习率,并根据学习率以及网络参数误差梯度

更新剪枝网络单元的网络参数。

[0097] 在更新结构参数和网络参数的过程中,随着更新轮次的增加可以逐步减小学习率。逐步减小的学习率与逐步增大的剪枝率相配合可以获得更加稳定的剪枝搜索过程。

[0098] 在步骤S240中,根据更新后的结构参数选取各个网络节点之间的目标操作分支,以基于目标操作分支确定目标网络单元。

[0099] 在达到预设的训练终止条件时,可以根据最终更新得到的结构参数选取目标操作分支并基于目标操作分支确定目标网络单元。训练终止条件例如可以是迭代训练次数达到预设次数或者损失函数输出损失误差小于误差阈值。

[0100] 图5示意性地示出了本申请一些实施例中选取目标操作分支的方法步骤流程图。如图5所示,在以上实施例的基础上,根据更新后的结构参数选取各个网络节点之间的目标操作分支,可以包括如下的步骤S510~步骤S530。

[0101] 步骤S510:根据更新后的结构参数比较两个网络节点之间的各个操作分支的分支权重,并根据比较结果选取分支权重最大的一个操作分支作为候选操作分支。

[0102] 根据结构参数的更新结果,首先可以对任意两个网络节点之间的操作分支进行比较和筛选,从中选取分支权重最大的一个操作分支作为候选操作分支。经过本步骤进行分支选取后,每两个网络节点之间可以仅保留一个候选操作分支,而丢弃除候选操作分支以外的其他操作分支。

[0103] 步骤S520:分别确定与各个当前网络节点具有数据映射关系的前继网络节点,前继网络节点是通过操作分支向当前网络节点输入数据的网络节点。

[0104] 以图3中的网络单元为例,节点3的前继网络节点包括节点0、节点1和节点2,节点2的前继网络节点包括节点0和节点1,节点1的前继网络节点包括节点0。

[0105] 步骤S530:比较当前网络节点与各个前继网络节点之间的候选操作分支的分支权重,并根据比较结果选取分支权重最大的一个或者多个候选操作分支作为目标操作分支。

[0106] 一个当前网络节点可以通过候选操作分支与一个或者多个前继网络节点相连,本步骤可以对各个前继网络节点对应的候选操作分支的分支权重进行比较,从而根据比较结果选取分支权重最大的若干个候选操作分支作为目标操作分支。一般而言,针对一个当前网络节点,可以保留分支权重最大的一个或者两个候选操作分支作为目标操作分支。

[0107] 通过执行步骤S510~步骤S530可以基于结构参数的更新对操作分支进行筛选,从而得到结构简化且表征能力强的目标网络单元。

[0108] 在通过搜索训练得到目标网络单元后,可以将目标网络单元堆叠组成神经网络,并对神经网络进行训练以更新神经网络的网络参数。

[0109] 在得到目标网络单元后,可以根据实际需要的网络深度对目标网络单元进行堆叠形成神经网络,组成该神经网络的各个目标网络单元具有完全相同的结构参数和网络参数。再利用训练样本对该神经网络进行训练,可以对其内部的网络参数进行更新,从而得到满足特定应用需求的神经网络。

[0110] 图6示意性地示出了在一应用场景下由目标网络单元堆叠形成神经网络的网络结构。如图6所示,在该神经网络中包括堆叠设置的Normal Cell和Reduction Cell两种网络单元。具体地,在连续堆叠n个Normal Cell后可以插入一个Reduction Cell。其中,Normal Cell内部的所有操作的步长stride均为1,使得其输入特征图和输出特征图具有相同的分

分辨率;Reduction Cell内部的所有操作的步长stride为2,使其具有对特征图进行下采样的功能。每个网络单元均包括两个输入端口,例如第k个网络单元的第一输入端口对应于第k-1个网络单元的输出端口,第k个网络单元的第二输入端口对应于第k-2个网络单元的输出端口。

[0111] 需要说明的是,图6所示的神经网络的网络结构仅为示例,本申请实施例可以根据不同应用场景下的实际需求选用不同形式的网络堆叠结构,本申请对此不多特殊限定。

[0112] 下面结合图7A和图7B对本申请实施例在一应用场景中的网络训练方法做出详细说明。图7A示意性地示出了在第k个训练轮次中的网络单元结构及其内部的软剪枝效果,图7B示意性地示出了在第k+1个训练轮次中的网络单元结构及其内部的软剪枝效果。

[0113] 如图7A所示,网络单元Cell被建模为一个有向无环图(DAG),假设每个Cell包含4个网络节点(0,1,2,3),每个网络节点都和它之前的所有网络节点存在基于多种操作分支的混合操作连接(mixed-operation),以网络节点对(0,1)为例,两个网络节点之间存在连接(mixed-operation),每个mixed-operation是候选操作集合中的一组操作的加权和。

[0114] 网络节点对(0,1)之间的mixed-operation由四个候选操作组成,对应的结构参数 $\alpha$ 经过softmax函数进行归一化映射之后可以得到连续松弛的选取概率信息。如图7A所示,在第k个训练轮次下,某一网络单元Normal Cell中的选取概率信息为 $\text{softmax}(\alpha)=[0.18, 0.46, 0.12, 0.24]$ 。连续松弛是一种将离散空间连续化的方法,本质上就是利用softmax函数将结构参数 $\alpha$ 的序列从离散空间变换到连续空间。

[0115] 本申请引入软剪枝机制,基于剪枝率引入一个掩膜参数 $M$ ,设定剪枝率 $\text{rate}=50\%$ ,那么只选中 $\text{softmax}(\alpha)$ 中概率最大的两个分量对应的操作保留,如图7A所示,第k个训练轮次下的掩膜参数 $M=[0, 1, 0, 1]$ ,在该训练轮次下只有被选中的操作及其对应的结构参数 $\alpha$ 的分量会被更新。例如,可以采用如下公式进行参数更新。

$$\delta^{i,j}(x) = \sum_{o \in O} \frac{\exp(\alpha_o^{i,j})}{\sum_{o' \in O} \exp(\alpha_{o'}^{i,j})} M_o^{i,j} o(x)$$

$$x^j = \sum_{i < j} o^{i,j} x^i$$

[0116] 其中, $O$ 代表有向图中连线上操作的集合, $o$ 和 $o'$ 表示操作集合 $O$ 中的任意操作, $\alpha_o^{i,j}$ 和 $\alpha_{o'}^{i,j}$ 分别代表网络节点 $i$ 和网络节点 $j$ 的连线上各个操作 $o$ 和 $o'$ 的权值, $x^i$ 和 $x^j$ 分别表示网络节点 $i$ 和网络节点 $j$ 的特征图, $M_o^{i,j}$ 表示网络节点 $i$ 和网络节点 $j$ 之间的掩膜参数。该公式的含义是使用连线上每个操作对相应的网络结点的特征图进行操作,然后按照一定的权值将结果组合起来作为连线的输出 $\delta^{i,j}(x)$ 。

[0117] 掩膜参数 $M$ 一旦确定下来并不是一直不变的,本申请可以具体设定一定的剪枝间隔,在新的剪枝步骤中掩膜参数 $M$ 会被重新更新,本次没有被选中的操作在下次剪枝步骤中也有可能被重新选中。如图7B所示,在第k+1个训练轮次,选取概率信息被更新为 $\text{softmax}(\alpha)=[0.18, 0.45, 0.19, 0.18]$ ,剪枝率仍为 $\text{rate}=50\%$ ,此时的掩膜参数被更新为 $M=[0, 1, 1, 0]$ 。

[0118] 在该应用场景下,神经网络的训练过程整体可以分为搜索训练和验证训练两个部

分。

[0119] 在搜索训练部分,可微分神经网络结构搜索将网络训练和搜索建模为如以下公式的两级优化问题:

$$\begin{aligned} \min_{\alpha} \quad & \mathcal{L}_{val}(w^*(\alpha), \alpha) \\ \text{s.t.} \quad & w^*(\alpha) = \operatorname{argmin}_w \mathcal{L}_{train}(w, \alpha) \end{aligned}$$

其中, $\alpha$ 为结构参数, $w$ 为网络参数, $val$ 和 $train$ 代表两个不同的训练样本集合。

[0120] 整个搜索过程分为两部分:网络权重参数(包含的所有操作的网络参数 $w$ )的训练和结构权重参数(每个mixed-operation对应的结构参数 $\alpha$ )优化。将原始的训练集按照1:1分为 $train$ 和 $val$ 两部分,交替固定训练两部分参数。

[0121] 具体而言,可以分为以下步骤。

[0122] 步骤S1:划分数据集得到 $train$ 和 $val$ ,构建超网 $F(.;w;\alpha)$ ,设定 $w$ 和 $\alpha$ 的学习率,设定总的训练轮次数量 $epoch$ ,剪枝间隔(多少个 $epoch$ 更新一次各个mixed-operation对应的掩膜参数 $M$ ),以及剪枝率变化策略 $R(.)$ 。

[0123] 步骤S2:初始化各个可训练参数。

[0124] 步骤S3:根据当前的 $epoch$ 和剪枝率变化策略更新当前的剪枝率。

[0125] 步骤S4:根据步骤S3更新的剪枝率以及当前各个mixed-operation的结构参数 $\alpha$ 更新掩膜参数 $M$ (如果处于剪枝 $epoch$ 便更新,否则保持上一次更新得到的 $M$ )。

[0126] 步骤S5:利用 $train$ 数据集,基于当前结构参数 $\alpha$ 及对应的掩膜参数 $M$ 更新网络参数 $w$ :

$$w_{k+1} \leftarrow w_k - \eta_w \cdot \nabla_{w_k} \mathcal{L}_{train}(w_k; \alpha_k \cdot M_k)$$

步骤S6:利用 $val$ 数据集,基于当前的网络参数 $w$ 和掩膜参数 $M$ 更新结构参数 $\alpha$ :

$$\alpha_{k+1} \leftarrow \alpha_k - \eta_{\alpha} \cdot \nabla_{\alpha_k} \mathcal{L}_{val}(w_k + 1; \alpha_k \cdot M_k)$$

步骤S7:重复执行步骤S3至步骤S6直至训练结束,根据结构参数 $\alpha$ 得到最佳的子网。

[0127] 在验证训练部分,利用搜索得到的Normal Cell和Reduction Cell堆叠成目标网络,然后在完整的训练集上面执行从头开始的训练,在本该应用场景中基于CIFAR10数据集进行搜索训练,同样可以基于CIFAR10数据集进行验证训练,训练超参数设置为: $epochs=600$ , $batch\ size=96$ , $learning\ rate=0.025$ ,优化器选择SGD,并采用余弦衰减的学习率变化策略。CIFAR10数据集是分类任务中应用非常广泛的样本数据集,包含10个类别6万张图片,其中训练集5万张,测试集1万张,单张图片为32x32x3的RGB图片。

[0128] 图8A示出了本申请技术方案在一应用场景中搜索得到的网络单元Normal Cell(上)和网络单元Reduction Cell(下)的网络结构示意图,图8B示出了应用现有技术中的DARTS算法搜索得到的网络单元Normal Cell(上)和网络单元Reduction Cell(下)的网络结构示意图。由图8A和图8B的对比可以发现,利用本申请提供的软剪枝技术加速的可微分神经网络架构搜索方法可搜索得到的Normal cell和DARTS方法搜索得到的Normal Cell相比具有更深的网络结构,理论上具有更高的特征学习能力,而Reduce Cell同样也具有更多的含参数操作,在下采样的同时具有更强的学习能力。

[0129] 表2示出了本申请技术方案与现有技术中部分NAS算法在CIFAR10数据集上进行实

验的数据对比。

[0130] 表2 实验数据对比

| 方案                     | 实验误差 (%) | 数据量 (M) | 计算成本 (GPU-天) | 训练方法           |
|------------------------|----------|---------|--------------|----------------|
| DenseNet-BC            | 3.46     | 25.6    | -            | RL             |
| NASNet-A               | 2.65     | 3.3     | 1800         | evolution      |
| AmoebaNet-A            | 3.34     | 3.2     | 3150         | evolution      |
| AmoebaNet-B            | 2.55     | 2.8     | 3150         | evolution      |
| Hierarchical Evolution | 3.75     | 15.7    | 300          | evolution      |
| PNAS                   | 3.41     | 3.2     | 225          | SMB0           |
| ENAS                   | 2.89     | 4.6     | 0.5          | RL             |
| DARTS                  | 3.00     | 3.3     | 4            | Gradient-based |
| 本申请                    | 2.59     | 3.5     | 0.3          | Gradient-based |

其中,RL表示强化学习,evolution表示基于基因算法的NAS。在CIFAR10上面的实验结果证实,本申请提供的技术方案可以缩短基于梯度的可微分神经网络搜索算法的耗时,同时取得更高的准确率。

[0131] 应当注意,尽管在附图中以特定顺序描述了本申请中方法的各个步骤,但是,这并非要求或者暗示必须按照该特定顺序来执行这些步骤,或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的,可以省略某些步骤,将多个步骤合并为一个步骤执行,以及/或者将一个步骤分解为多个步骤执行等。

[0132] 以下介绍本申请的装置实施例,可以用于执行本申请上述实施例中的神经网络搜索方法。图9示意性地示出了本申请实施例提供的神经网络搜索装置的结构框图。

[0133] 如图9所示,神经网络搜索装置900包括:网络构建模块910,被配置为获取包括至少两种候选网络结构的神经网络搜索空间,并根据所述神经网络搜索空间构建包括至少两个网络节点的初始网络单元,其中任意两个网络节点之间包括至少两个对应于所述候选网络结构的操作分支,操作分支用于表示两个网络节点之间的数据映射关系;网络剪枝模块920,被配置为获取当前训练轮次下各个网络节点之间的掩膜参数,并根据掩膜参数对网络单元进行剪枝处理以得到剪枝网络单元;参数更新模块930,被配置为对剪枝网络单元进行训练以更新各个网络节点之间的结构参数,结构参数用于表示两个网络节点之间的各个操作分支的分支权重;网络确定模块940,被配置为根据更新后的结构参数选取各个网络节点之间的目标操作分支,以基于目标操作分支确定目标网络单元。

[0134] 在本申请的一些实施例中,基于以上各实施例,网络剪枝模块包括:轮次确定单元,被配置为获取当前训练轮次的轮次信息,并根据轮次信息确定当前训练轮次是否为掩膜参数更新轮次;掩膜参数确定单元,被配置为若当前训练轮次不是掩膜参数更新轮次,则将上一训练轮次的掩膜参数保持为当前训练轮次的掩膜参数;掩膜参数更新单元,被配置为若当前训练轮次是掩膜参数更新轮次,则根据当前训练轮次的剪枝率以及各个网络节点之间的结构参数确定当前训练轮次的掩膜参数,结构参数用于表示两个网络节点之间的各个操作分支的分支权重。

[0135] 在本申请的一些实施例中,基于以上各实施例,掩膜参数更新单元包括:剪枝率确定子单元,被配置为获取当前训练轮次的轮次信息,并根据轮次信息确定当前训练轮次的剪枝率;分支数量确定子单元,被配置为根据当前训练轮次的剪枝率以及各个网络节点之

间的操作分支的分支数量确定操作分支的剪枝数量和保留数量;操作分支选取子单元,被配置为根据各个网络节点之间的结构参数选取剪枝数量的操作分支作为剪枝分支并选取保留数量的操作分支作为保留分支;掩膜参数确定子单元,被配置为为剪枝分支分配第一掩膜数值并为保留分支分配第二掩膜数值,以得到由第一掩膜数值和第二掩膜数值组成的当前训练轮次的掩膜参数。

[0136] 在本申请的一些实施例中,基于以上各实施例,剪枝率确定子单元包括:阶段划分子单元,被配置为获取预设的迭代训练次数,并对迭代训练次数进行划分以得到多个剪枝率更新阶段,其中各个剪枝率更新阶段具有分段增大的剪枝率;第一剪枝率确定子单元,被配置为根据轮次信息确定当前训练轮次所在的剪枝率更新阶段,并将该剪枝率更新阶段的剪枝率确定为当前训练轮次的剪枝率。

[0137] 在本申请的一些实施例中,基于以上各实施例,剪枝率确定子单元包括:模型获取子单元,被配置为获取用于在不同训练轮次下连续增大剪枝率的剪枝率更新模型;第二剪枝率确定子单元,被配置为通过剪枝率更新模型对轮次信息进行映射处理以得到当前训练轮次下的剪枝率。

[0138] 在本申请的一些实施例中,基于以上各实施例,操作分支选取子单元包括:参数映射子单元,被配置为对各个网络节点之间的结构参数进行归一化映射以得到每个操作分支的选取概率;概率选取子单元,被配置为按照选取概率将概率最低的剪枝数量的操作分支选取为剪枝分支并将概率最高的保留数量的操作分支选取为保留分支。

[0139] 在本申请的一些实施例中,基于以上各实施例,网络剪枝模块还包括:分支分类单元,被配置为根据掩膜参数将初始网络单元中的部分操作分支确定为可更新分支,并将除可更新分支以外的其他操作分支确定为不可更新分支;参数配置单元,被配置为将可更新分支的网络参数和结构参数配置为可更新参数并将不可更新分支的网络参数和结构参数配置为不可更新参数;分支组合单元,被配置为将可更新分支以及不可更新分支组成的网络单元确定为剪枝网络单元。

[0140] 在本申请的一些实施例中,基于以上各实施例,参数更新模块包括:样本集合获取单元,被配置为获取用于更新各个网络节点之间的结构参数的第一样本集合以及用于更新剪枝网络单元的网络参数的第二样本集合;参数交替更新单元,被配置为分别使用第一样本集合和第二样本集合对剪枝网络单元进行训练以交替更新各个网络节点之间的结构参数和剪枝网络单元的网络参数。

[0141] 在本申请的一些实施例中,基于以上各实施例,参数交替更新单元包括:结构参数更新子单元,被配置为固定剪枝网络单元的网络参数,并使用第一样本集合对剪枝网络单元进行训练以更新各个网络节点之间的结构参数;网络参数更新子单元,被配置为固定各个网络节点之间的结构参数,并使用第二样本集合对剪枝网络单元进行训练以更新剪枝网络单元的网络参数;参数交替更新子单元,被配置为交替更新各个网络节点之间的结构参数以及剪枝网络单元的网络参数直至训练结束。

[0142] 在本申请的一些实施例中,基于以上各实施例,结构参数更新子单元包括:第一网络映射子单元,被配置为通过剪枝网络单元对第一样本集合中的训练样本进行映射处理以得到与训练样本相关的损失误差;第一梯度确定子单元,被配置为根据损失误差确定各个网络节点之间的结构参数的结构参数误差梯度;第一参数更新子单元,被配置为获取当前

训练轮次的学习率,并根据学习率以及结构参数误差梯度更新各个网络节点之间的结构参数。

[0143] 在本申请的一些实施例中,基于以上各实施例,网络参数更新子单元包括:第二网络映射子单元,被配置为通过剪枝网络单元对第二样本集合中的训练样本进行映射处理以得到与训练样本相关的损失误差;第二梯度确定子单元,被配置为根据损失误差确定剪枝网络单元的网络参数的网络参数误差梯度;第二参数更新子单元,被配置为获取当前训练轮次的学习率,并根据学习率以及网络参数误差梯度更新剪枝网络单元的网络参数。

[0144] 在本申请的一些实施例中,基于以上各实施例,网络确定模块包括:分支候选单元,被配置为根据更新后的结构参数比较两个网络节点之间的操作分支的分支权重,并根据比较结果选取分支权重最大的一个操作分支作为候选操作分支;节点确定单元,被配置为分别确定与各个当前网络节点具有数据映射关系的前继网络节点,前继网络节点是通过操作分支向当前网络节点输入数据的网络节点;分支选取单元,被配置为比较当前网络节点与各个前继网络节点之间的候选操作分支的分支权重,并根据比较结果选取分支权重最大的一个或者多个候选操作分支作为目标操作分支。

[0145] 本申请各实施例中提供的神经网络搜索装置的具体细节已经在对应的方法实施例中进行了详细的描述,此处不再赘述。

[0146] 图10示意性地示出了用于实现本申请实施例的电子设备的计算机系统结构框图。

[0147] 需要说明的是,图10示出的电子设备的计算机系统1000仅是一个示例,不应对本申请实施例的功能和使用范围带来任何限制。

[0148] 如图10所示,计算机系统1000包括中央处理器1001(Central Processing Unit, CPU),其可以根据存储在只读存储器1002(Read-Only Memory, ROM)中的程序或者从存储部分1008加载到随机访问存储器1003(Random Access Memory, RAM)中的程序而执行各种适当的动作和处理。在随机访问存储器1003中,还存储有系统操作所需的各种程序和数据。中央处理器1001、在只读存储器1002以及随机访问存储器1003通过总线1004彼此相连。输入/输出接口1005(Input /Output接口,即I/O接口)也连接至总线1004。

[0149] 以下部件连接至输入/输出接口1005:包括键盘、鼠标等的输入部分1006;包括诸如阴极射线管(Cathode Ray Tube, CRT)、液晶显示器(Liquid Crystal Display, LCD)等以及扬声器等的输出部分1007;包括硬盘等的存储部分1008;以及包括诸如局域网卡、调制解调器等的网络接口卡的通信部分1009。通信部分1009经由诸如因特网的网络执行通信处理。驱动器1010也根据需要连接至输入/输出接口1005。可拆卸介质1011,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器1010上,以便于从其上读出的计算机程序根据需要被安装入存储部分1008。

[0150] 特别地,根据本申请的实施例,各个方法流程图中所描述的过程可以被实现为计算机软件程序。例如,本申请的实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分1009从网络上被下载和安装,和/或从可拆卸介质1011被安装。在该计算机程序被中央处理器1001执行时,执行本申请的系统中限定的各种功能。

[0151] 需要说明的是,本申请实施例所示的计算机可读介质可以是计算机可读信号介质

或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是一——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(Erasable Programmable Read Only Memory, EPROM)、闪存、光纤、便携式紧凑磁盘只读存储器(Compact Disc Read-Only Memory, CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中,计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、有线等等,或者上述的任意合适的组合。

[0152] 附图中的流程图和框图,图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0153] 应当注意,尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元,但是这种划分并非强制性的。实际上,根据本申请的实施方式,上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之,上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0154] 通过以上的实施方式的描述,本领域的技术人员易于理解,这里描述的示例实施方式可以通过软件实现,也可以通过软件结合必要的硬件的方式来实现。因此,根据本申请实施方式的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM, U盘, 移动硬盘等)中或网络上,包括若干指令以使得一台计算设备(可以是个人计算机、服务器、触控终端、或者网络设备等)执行根据本申请实施方式的方法。

[0155] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。

[0156] 应当理解的是,本申请并不局限于上面已经描述并在附图中示出的精确结构,并

且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

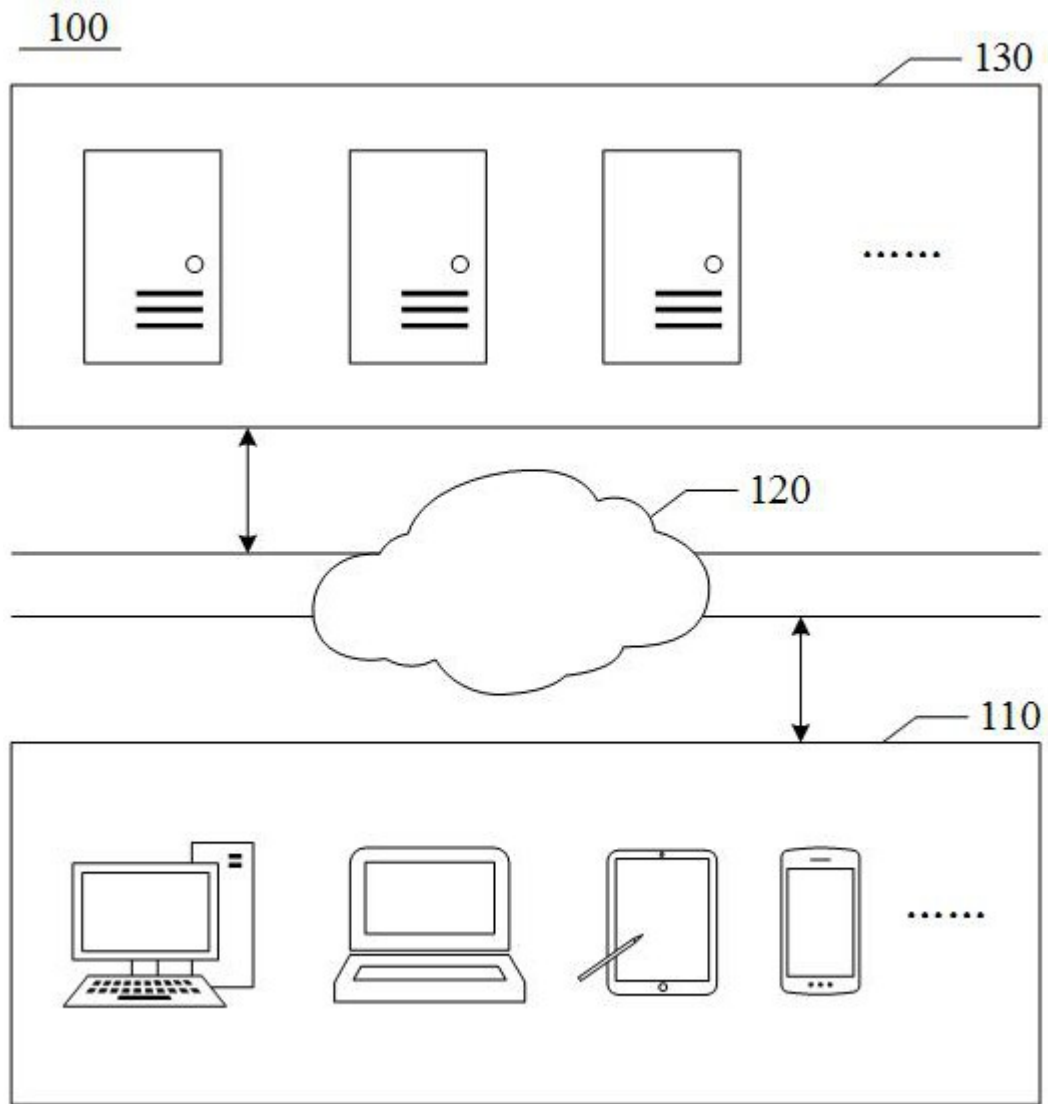


图1

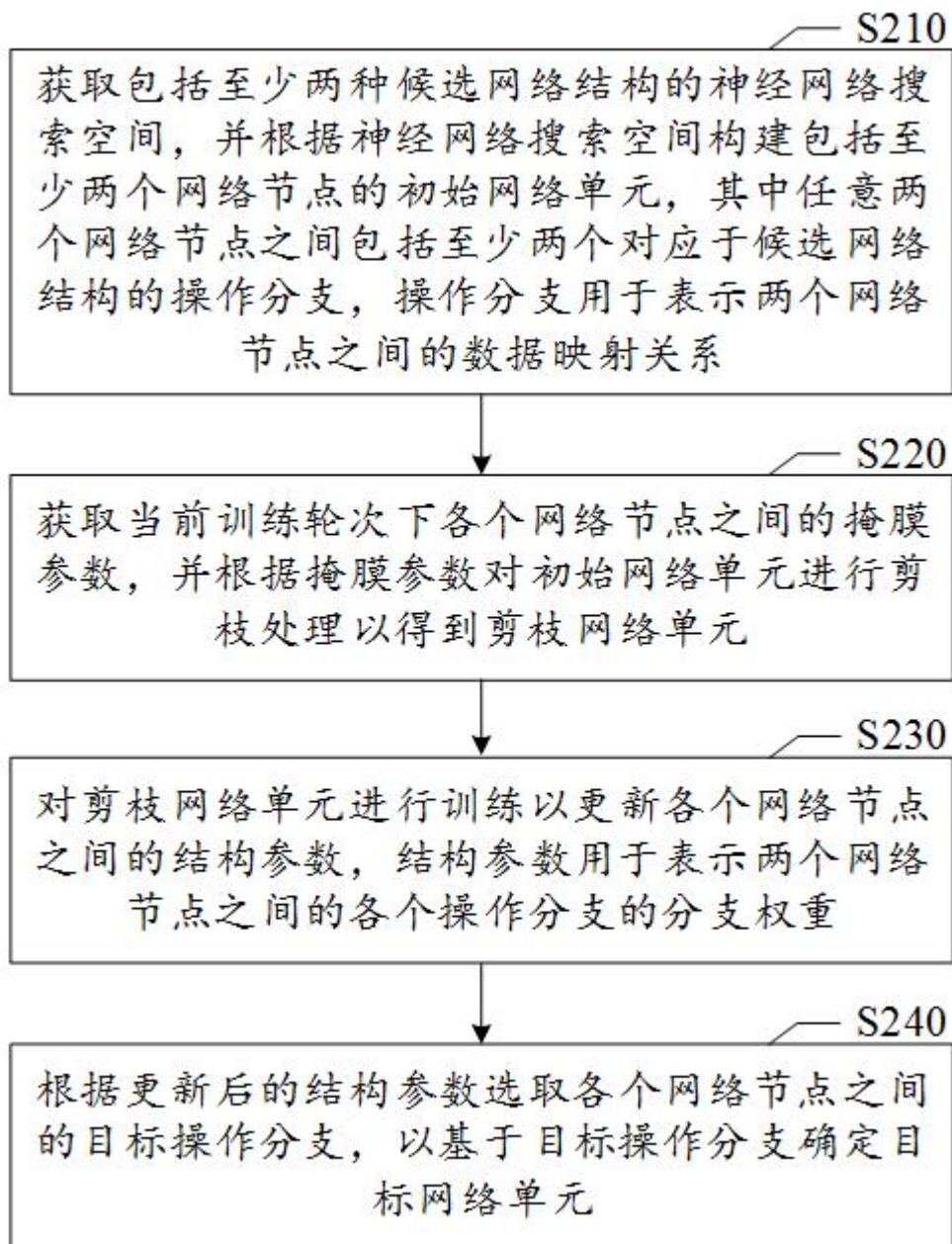


图2

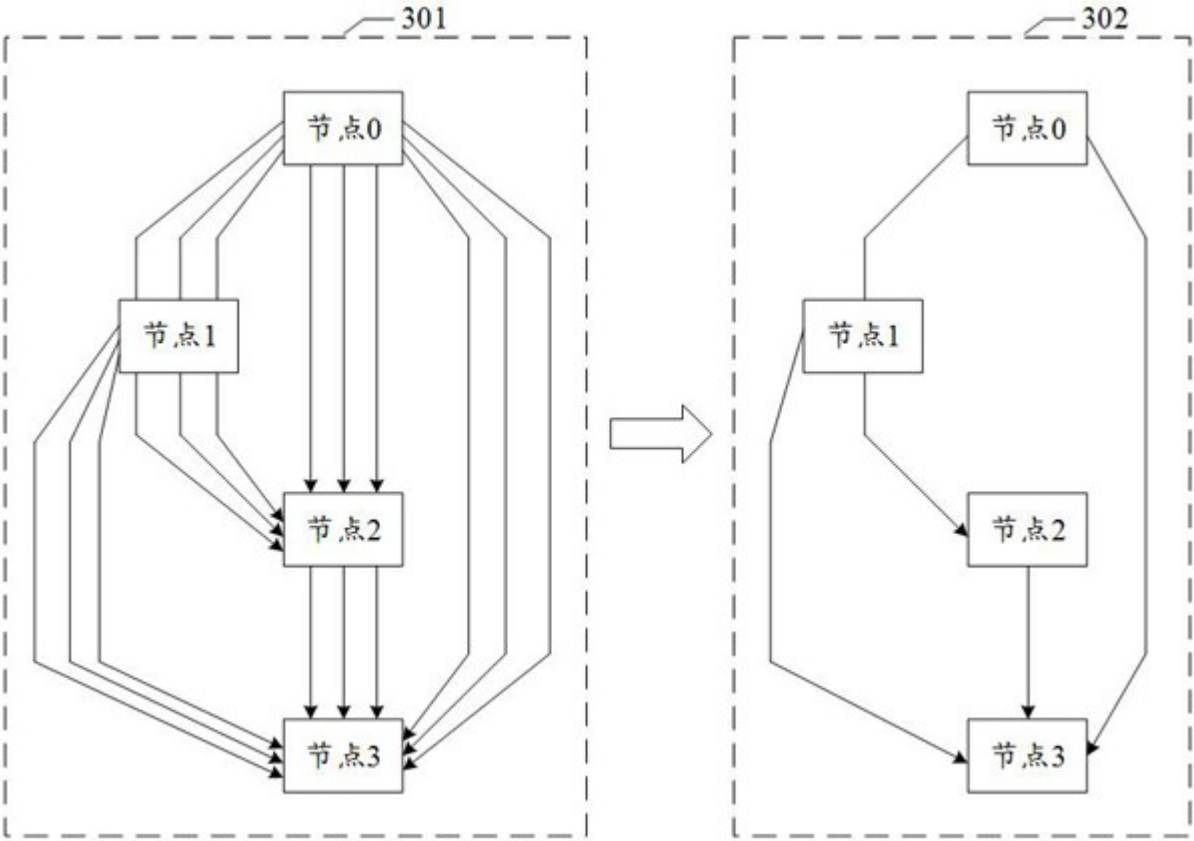


图3

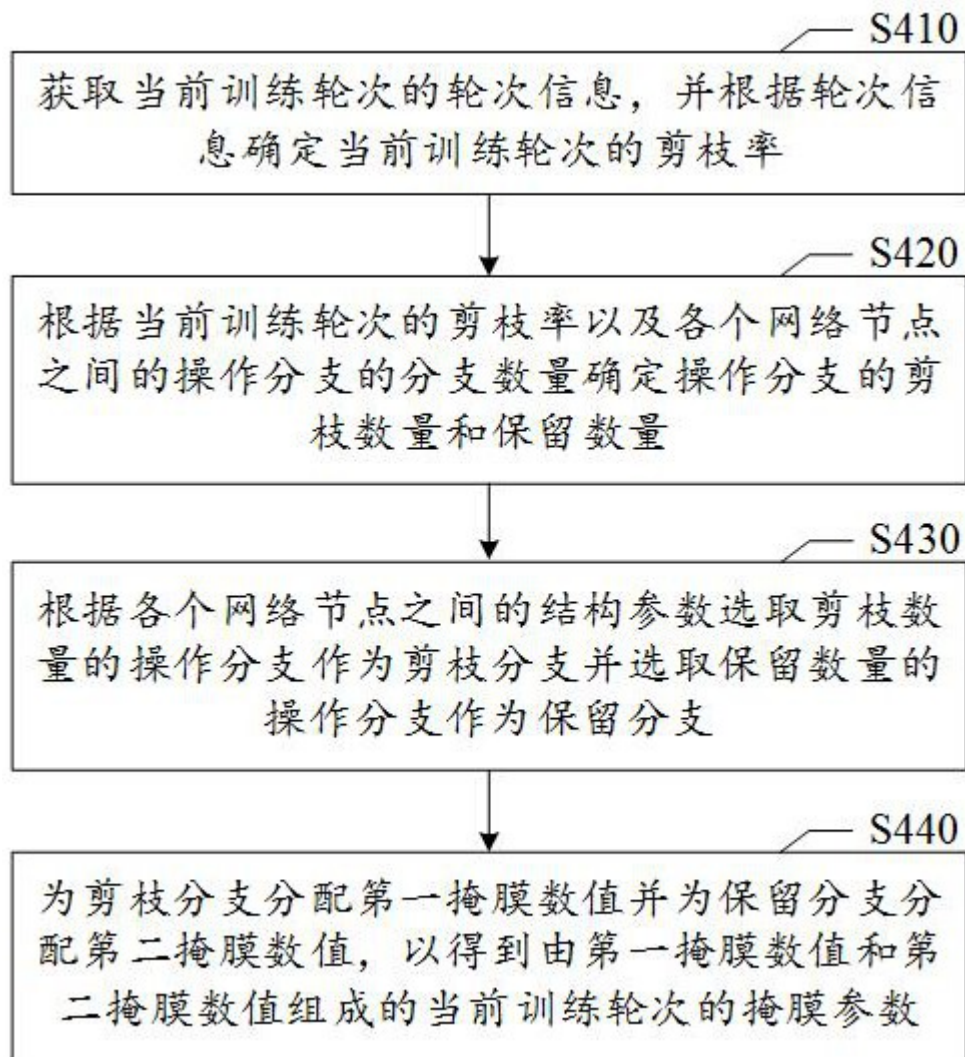


图4

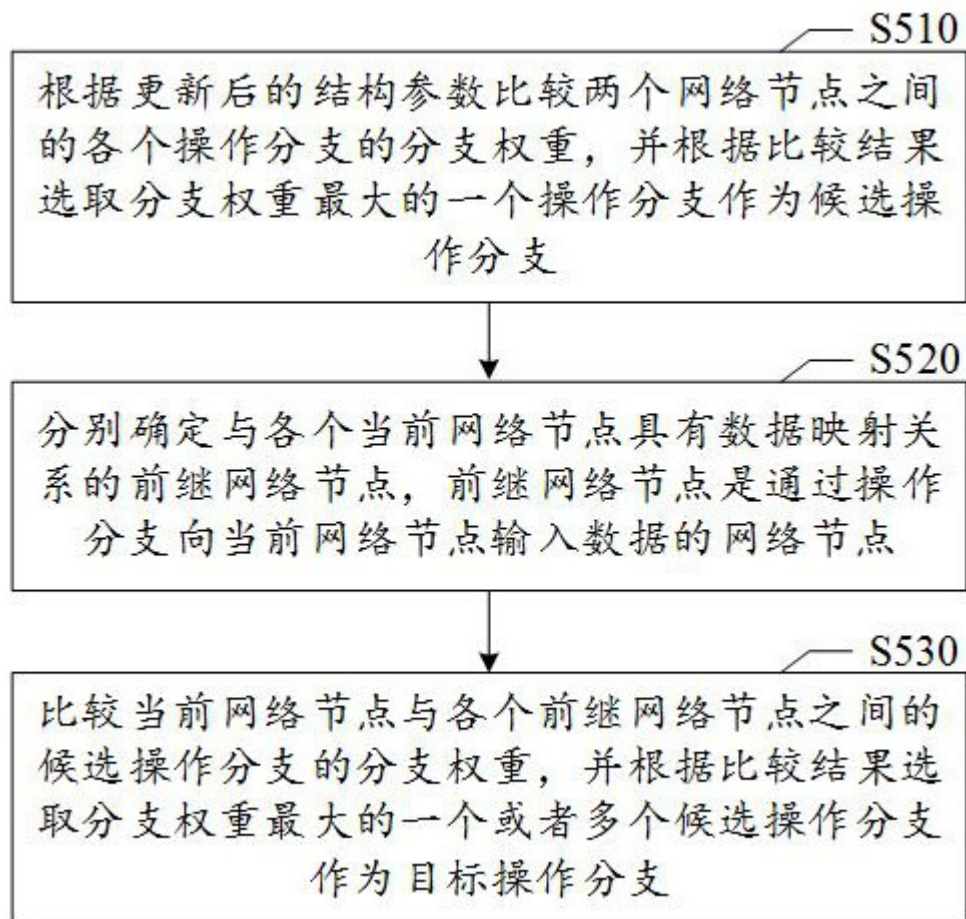


图5

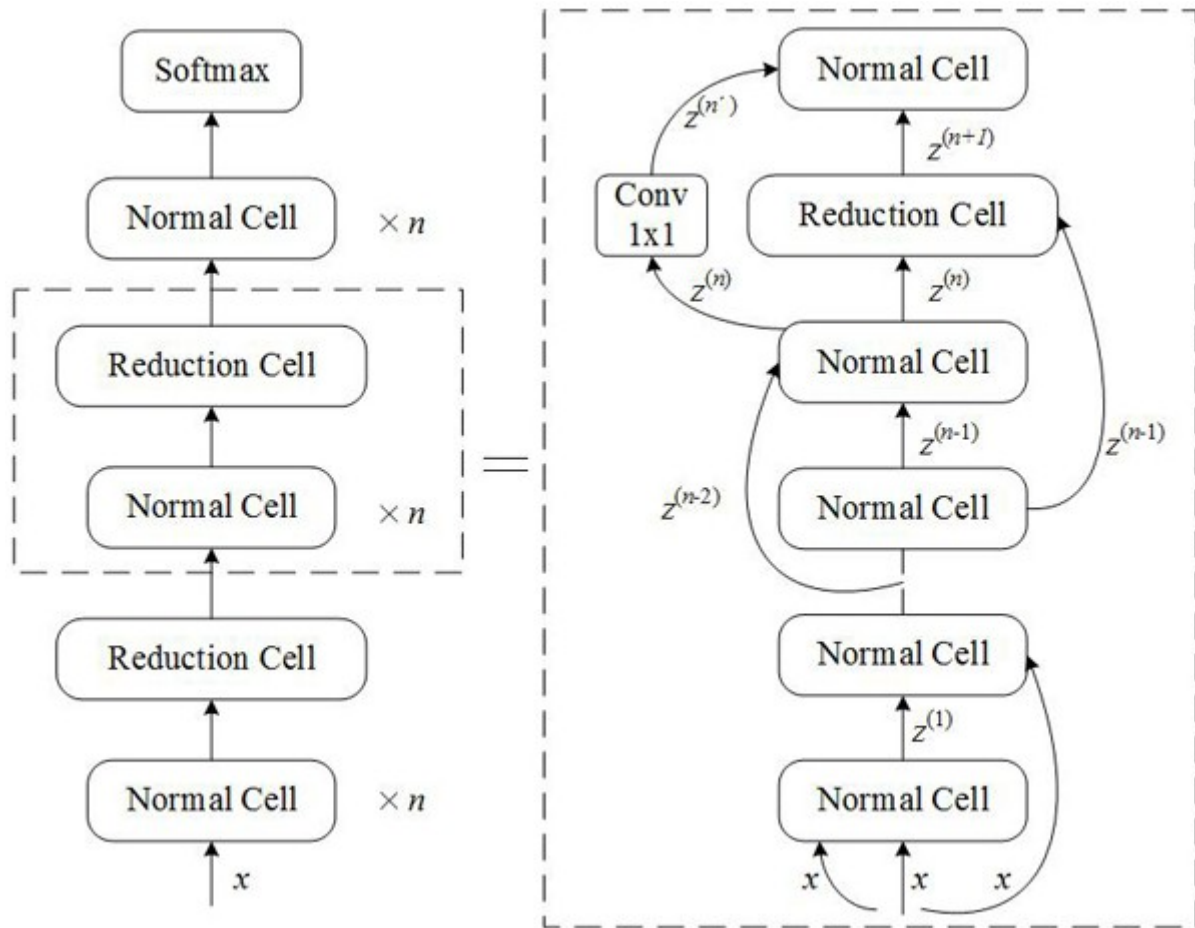


图6

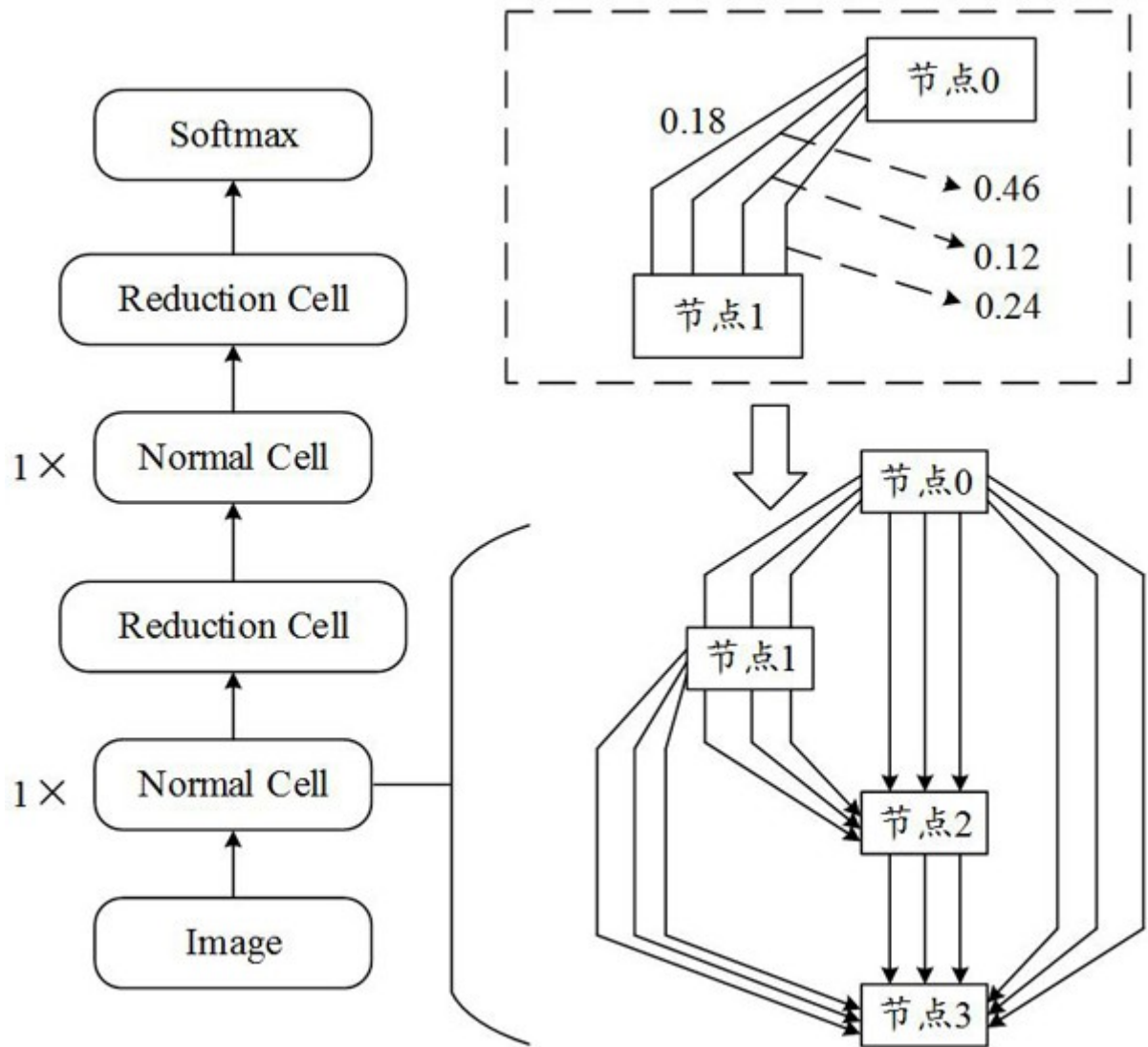


图7A

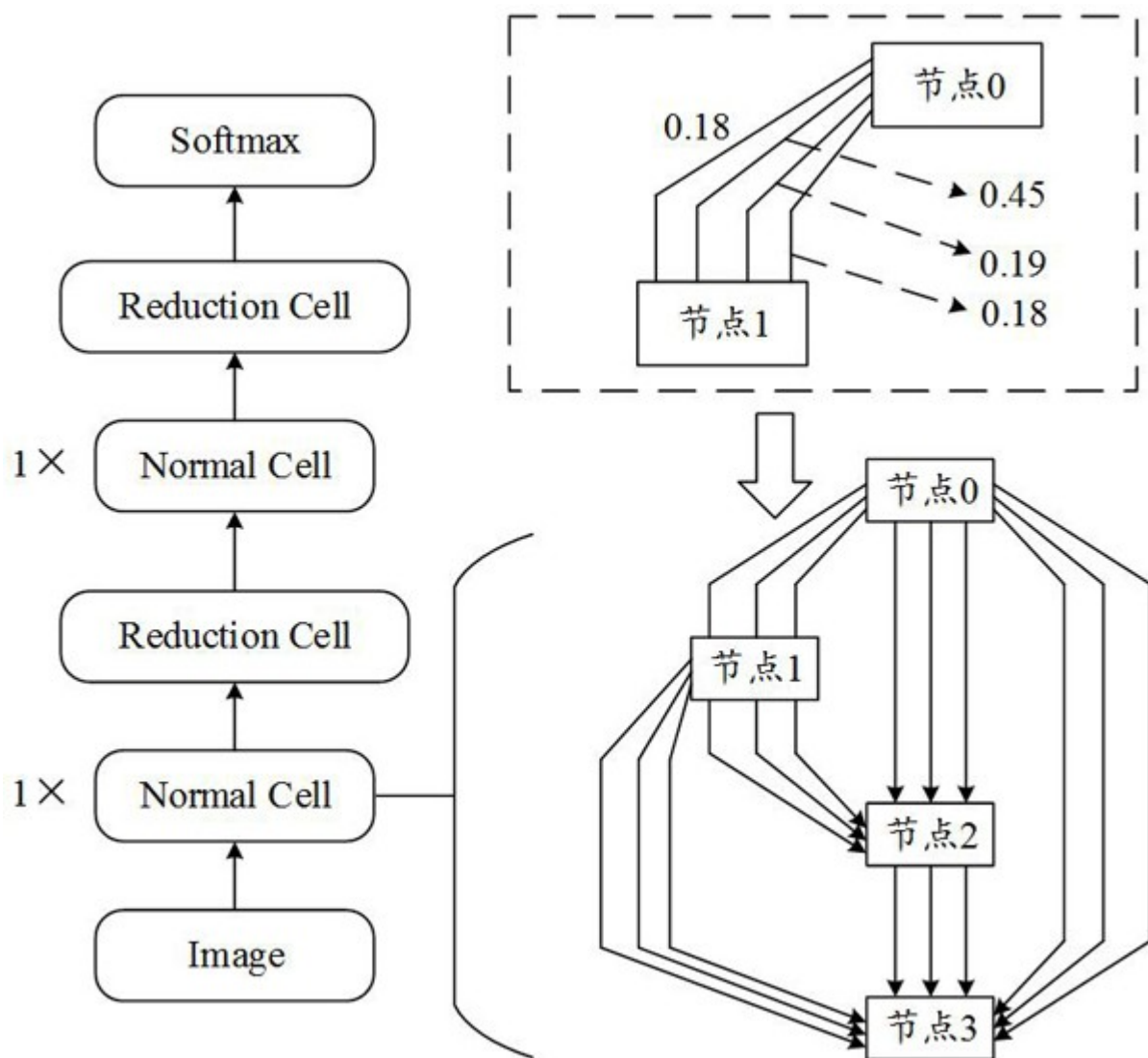


图7B

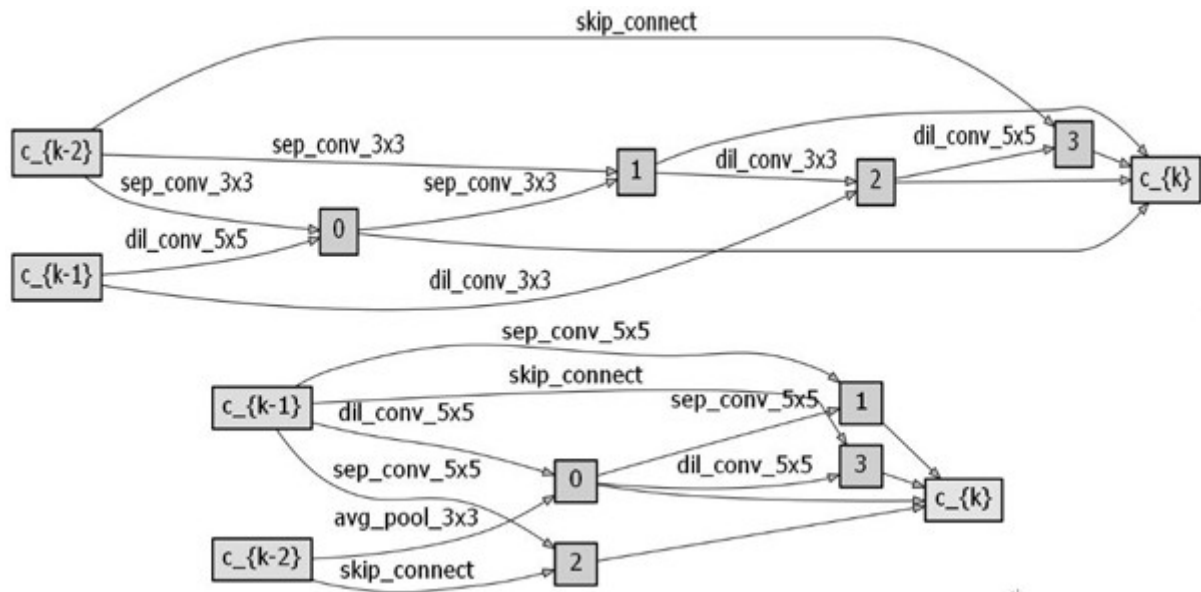


图8A

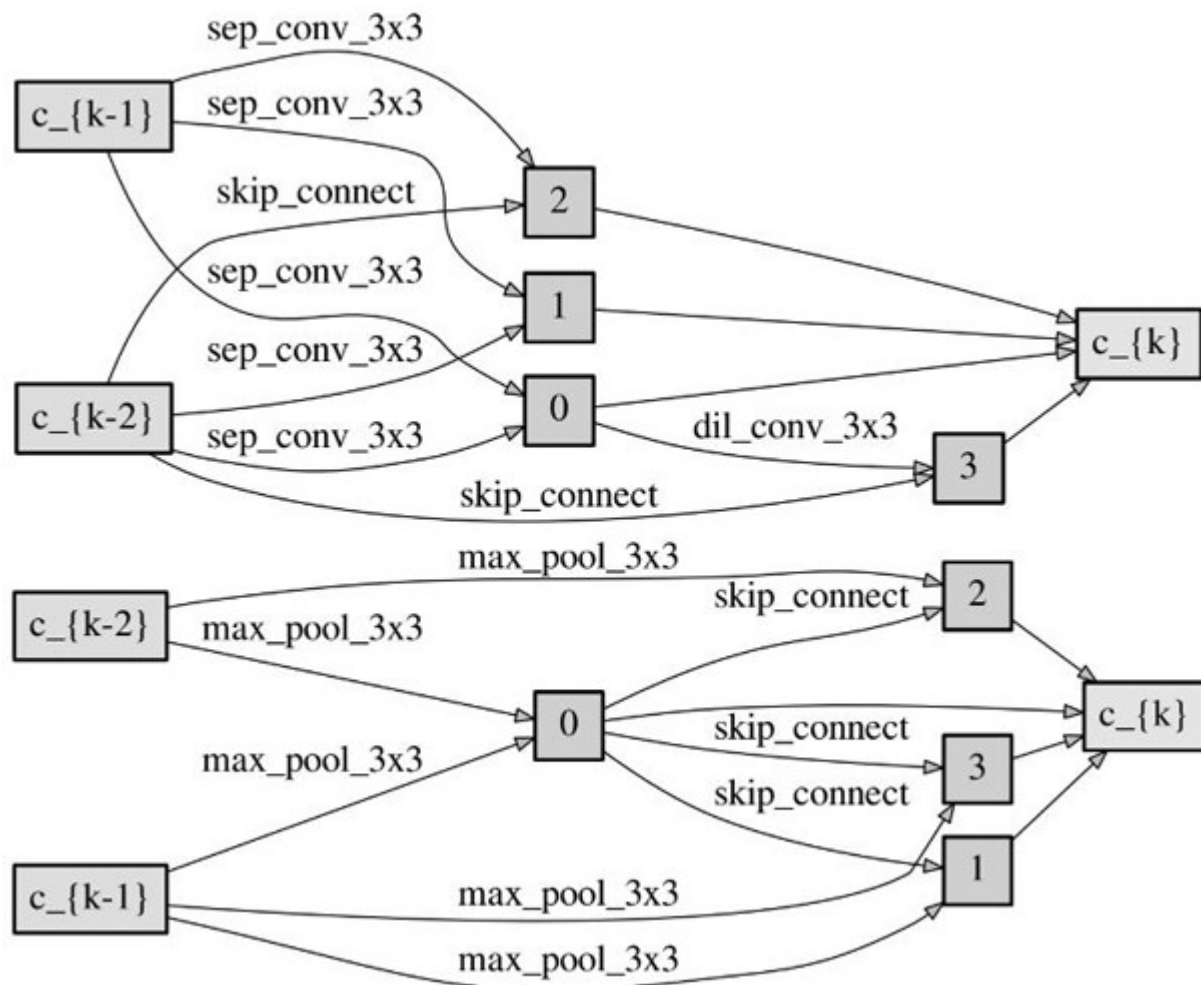


图8B

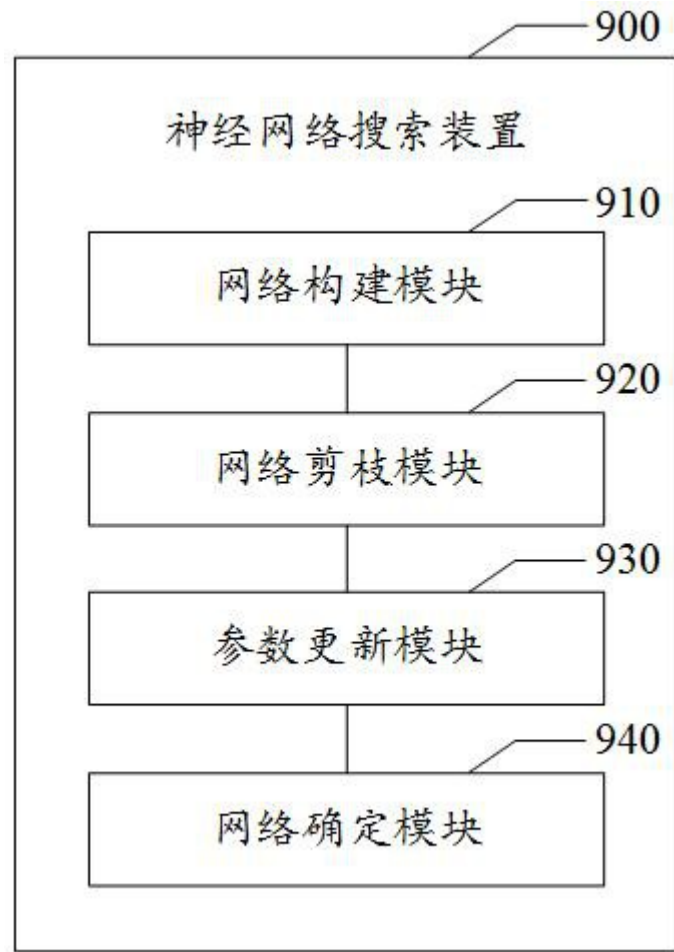


图9

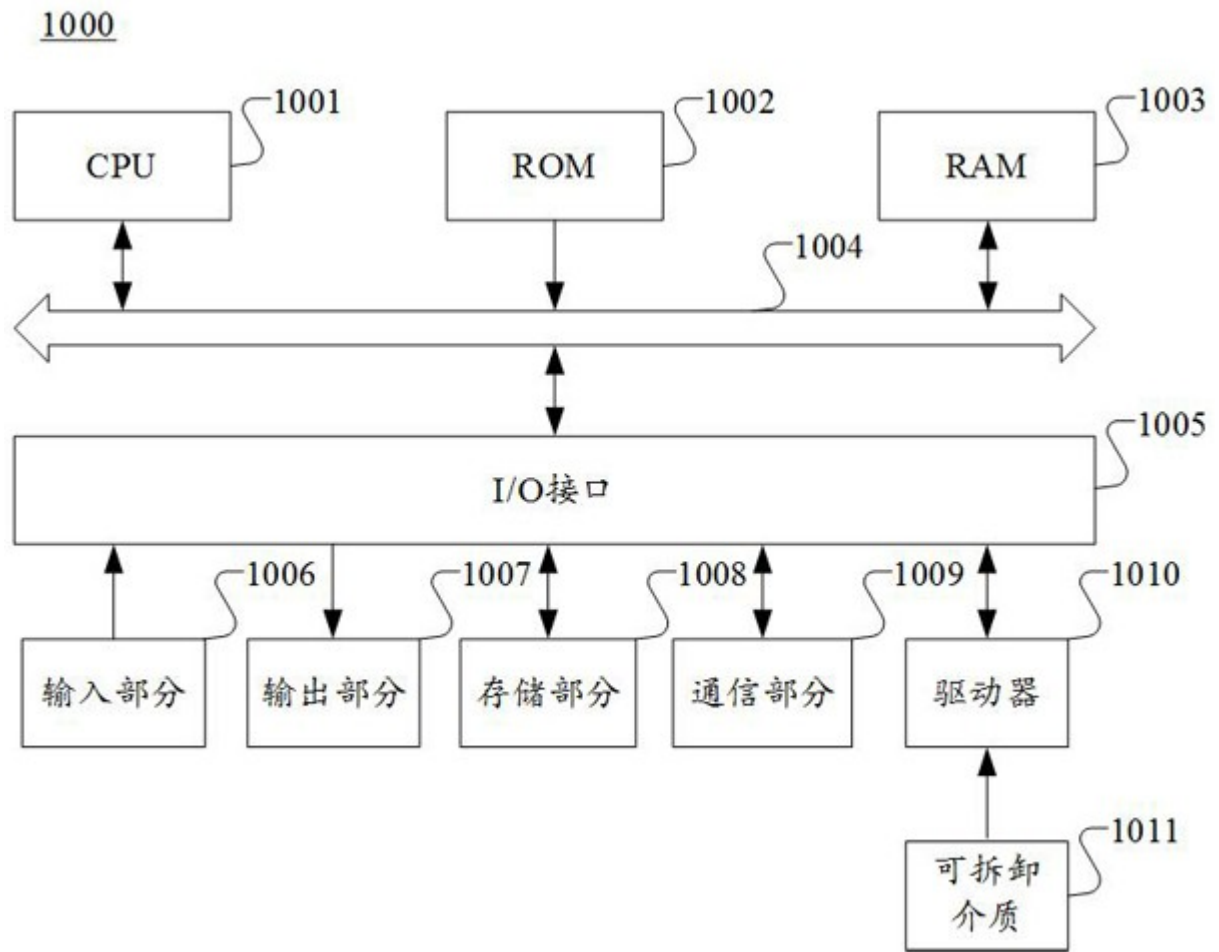


图10