(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 112100459 A
(43)申请公布日 2020.12.18

(21)申请号 202011026459.0

(22)申请日 2020.09.25

(71)申请人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 希滕 张刚 温圣召

(74)专利代理机构 北京市铸成律师事务所 11313
代理人 郭丽祥 杨瑾瑾

(51)Int.Cl.
G06F 16/903(2019.01)
G06F 16/909(2019.01)

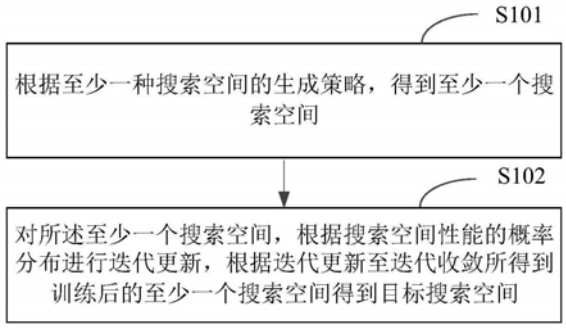
权利要求书3页 说明书12页 附图2页

(54)发明名称

搜索空间的生成方法及装置、电子设备及存储介质

(57)摘要

本申请公开了搜索空间的生成方法、基于概率分布的搜索方法及装置、电子设备及存储介质，涉及人工智能领域、计算机视觉、深度学习以及智能云技术等领域。其中，搜索空间的生成方法，其具体实现方案为：根据至少一种搜索空间的生成策略，得到至少一个搜索空间；对所述至少一个搜索空间，根据搜索空间性能的概率分布进行迭代更新，根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到目标搜索空间。采用本申请，至少可以提高诸如处理速度、处理精度等硬件性能。



CN 112100459 A

1.一种搜索空间的生成方法,该方法包括:

根据至少一种搜索空间的生成策略,得到至少一个搜索空间;

对所述至少一个搜索空间,根据搜索空间性能的概率分布进行迭代更新,根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到目标搜索空间。

2.根据权利要求1所述的方法,所述搜索空间的生成策略,包括如下任意一种或多种:

根据选定训练模型中层的类型及所述训练模型中层的个数,来生成所述搜索空间;

根据选定训练模型中层的类型,来生成所述搜索空间;

根据选定训练模型中层的卷积属性及所述训练模型中层的个数,来生成所述搜索空间;

根据选定训练模型中层的拓扑结构,来生成所述搜索空间;其中,所述拓扑结构包括单分支的拓扑结构、或多分支的拓扑结构。

3.一种基于概率分布的搜索方法,该方法包括:

根据至少一种搜索空间的生成策略,得到至少一个搜索空间;

对所述至少一个搜索空间,根据搜索空间性能的概率分布以迭代更新的方式进行训练,在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个搜索空间;

响应于第一搜索操作,在所述训练后的至少一个搜索空间中搜索得到目标搜索空间。

4.根据权利要求3所述的方法,其中,所述第一目标参数,包括:用于所述搜索空间性能评估的第一超参数。

5.根据权利要求3所述的方法,还包括:

对所述目标搜索空间中的至少一个模型结构,根据模型结构性能的概率分布以迭代更新的方式进行训练,在所述至少一个模型结构根据第二目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个模型结构;

响应于第二搜索操作,在所述训练后的至少一个模型结构中搜索得到目标模型结构。

6.根据权利要求5所述的方法,其中,所述第二目标参数,包括:用于所述模型结构性能评估的第二超参数。

7.根据权利要求3所述的方法,其中,所述对所述至少一个搜索空间,根据搜索空间性能的概率分布以迭代更新的方式进行训练,在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个搜索空间,包括:

对所述至少一个搜索空间,根据搜索空间性能的概率分布进行建模,得到第一概率模型;

将用于所述搜索空间性能评估的第一超参数作为所述第一目标参数;

根据所述第一超参数,迭代更新所述第一概率模型,以基于所述第一概率模型对所述至少一个搜索空间进行迭代更新至迭代收敛。

8.根据权利要求5所述的方法,其中,所述对所述目标搜索空间中的至少一个模型结构,根据模型结构性能的概率分布以迭代更新的方式进行训练,在所述至少一个模型结构根据第二目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个模型结构,包括:

对所述至少一个模型结构,根据模型结构性能的概率分布进行建模,得到第二概率模

型；

将用于所述模型结构性能评估的第二超参数作为所述第二目标参数；

根据所述第二超参数，迭代更新所述第二概率模型，以基于所述第二概率模型对所述至少一个模型结构进行迭代更新至迭代收敛。

9.根据权利要求3-8中任一项所述的方法，还包括：

获取待处理的图像；

将所述待处理的图像，输入根据所述目标搜索空间搜索得到的目标模型结构进行图像处理，得到目标图像；

其中，所述图像处理包括：图像分类、图像识别、图像检测中的至少一种处理。

10.一种搜索空间的生成装置，所述装置包括：

第一处理模块，用于根据至少一种搜索空间的生成策略，得到至少一个搜索空间；

第二处理模块，用于对所述至少一个搜索空间，根据搜索空间性能的概率分布进行迭代更新，根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到目标搜索空间。

11.根据权利要求10所述的装置，所述搜索空间的生成策略，包括如下任意一种或多种：

根据选定训练模型中层的类型及所述训练模型中层的个数，来生成所述搜索空间；

根据选定训练模型中层的类型，来生成所述搜索空间；

根据选定训练模型中层的卷积属性及所述训练模型中层的个数，来生成所述搜索空间；

根据选定训练模型中层的拓扑结构，来生成所述搜索空间；其中，所述拓扑结构包括单分支的拓扑结构、或多分支的拓扑结构。

12.一种基于概率分布的搜索装置，该装置包括：

第一搜索模块，用于根据至少一种搜索空间的生成策略，得到至少一个搜索空间；

第二搜索模块，用于对所述至少一个搜索空间，根据搜索空间性能的概率分布以迭代更新的方式进行训练，在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下，结束所述训练，得到训练后的至少一个搜索空间；

第三搜索模块，用于响应于第一搜索操作，在所述训练后的至少一个搜索空间中搜索得到目标搜索空间。

13.根据权利要求12所述的装置，其中，所述第一目标参数，包括：用于所述搜索空间性能评估的第一超参数。

14.根据权利要求12所述的装置，还包括第四搜索模块，用于：

对所述目标搜索空间中的至少一个模型结构，根据模型结构性能的概率分布以迭代更新的方式进行训练，在所述至少一个模型结构根据第二目标参数迭代更新至迭代收敛的情况下，结束所述训练，得到训练后的至少一个模型结构；

响应于第二搜索操作，在所述训练后的至少一个模型结构中搜索得到目标模型结构。

15.根据权利要求14所述的装置，其中，所述第二目标参数，包括：用于所述模型结构性能评估的第二超参数。

16.根据权利要求12所述的装置，其中，所述第二搜索模块，用于：

对所述至少一个搜索空间，根据搜索空间性能的概率分布进行建模，得到第一概率模

型；

将用于所述搜索空间性能评估的第一超参数作为所述第一目标参数；

根据所述第一超参数，迭代更新所述第一概率模型，以基于所述第一概率模型对所述至少一个搜索空间进行迭代更新至迭代收敛。

17.根据权利要求14所述的装置，其中，所述第四搜索模块，用于：

对所述至少一个模型结构，根据模型结构性能的概率分布进行建模，得到第二概率模型；

将用于所述模型结构性能评估的第二超参数作为所述第二目标参数；

根据所述第二超参数，迭代更新所述第二概率模型，以基于所述第二概率模型对所述至少一个模型结构进行迭代更新至迭代收敛。

18.根据权利要求12-17中任一项所述的装置，还包括图像处理模块，用于：

获取待处理的图像；

将所述待处理的图像，输入根据所述目标搜索空间搜索得到的目标模型结构进行图像处理，得到目标图像；

其中，所述图像处理包括：图像分类、图像识别、图像检测中的至少一种处理。

19.一种电子设备，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-9中任一项所述的方法。

20.一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令用于使所述计算机执行权利要求1-9中任一项所述的方法。

搜索空间的生成方法及装置、电子设备及存储介质

技术领域

[0001] 本申请涉及人工智能领域。本申请尤其涉及计算机视觉、深度学习以及智能云技术等领域等领域。

背景技术

[0002] 在信息处理领域中，无论是对文本信息、对包括音频或视频在内的多媒体信息、对图像信息、对视频处理中帧提取得到的图像信息等各种信息而言，都需要使用具备更好硬件性能的硬件(如终端或服务器及其芯片)、或通过多个硬件组合来架设具备更好硬件性能的硬件系统，才可以得到最优的处理效果。

[0003] 然而，相关技术中，对如何提高硬件性能，如处理速度、处理精度等，并未提供有效的解决方案。

发明内容

[0004] 本申请提供了一种搜索空间的生成方法、基于概率分布的搜索方法及装置、电子设备及存储介质。

[0005] 根据本申请的一方面，提供了一种搜索空间的生成方法，包括：

[0006] 根据至少一种搜索空间的生成策略，得到至少一个搜索空间；

[0007] 对所述至少一个搜索空间，根据搜索空间性能的概率分布进行迭代更新，根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到目标搜索空间。

[0008] 根据本申请的另一方面，提供了一种基于概率分布的搜索方法，包括：

[0009] 根据至少一种搜索空间的生成策略，得到至少一个搜索空间；

[0010] 对所述至少一个搜索空间，根据搜索空间性能的概率分布以迭代更新的方式进行训练，在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下，结束所述训练，得到训练后的至少一个搜索空间；

[0011] 响应于第一搜索操作，在所述训练后的至少一个搜索空间中搜索得到目标搜索空间。

[0012] 根据本申请的另一方面，提供了一种搜索空间的生成装置，包括：

[0013] 第一处理模块，用于根据至少一种搜索空间的生成策略，得到至少一个搜索空间；

[0014] 第二处理模块，用于对所述至少一个搜索空间，根据搜索空间性能的概率分布进行迭代更新，根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到目标搜索空间。

[0015] 根据本申请的另一方面，提供了一种基于概率分布的搜索装置，包括：

[0016] 第一搜索模块，用于根据至少一种搜索空间的生成策略，得到至少一个搜索空间；

[0017] 第二搜索模块，用于对所述至少一个搜索空间，根据搜索空间性能的概率分布以迭代更新的方式进行训练，在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下，结束所述训练，得到训练后的至少一个搜索空间；

[0018] 第三搜索模块，用于响应于第一搜索操作，在所述训练后的至少一个搜索空间中搜索得到目标搜索空间。

[0019] 根据本申请的另一方面，提供了一种电子设备，包括：

[0020] 至少一个处理器；以及

[0021] 与该至少一个处理器通信连接的存储器；其中，

[0022] 该存储器存储有可被该至少一个处理器执行的指令，该指令被该至少一个处理器执行，以使该至少一个处理器能够执行本申请任意一实施例所提供的方法。

[0023] 根据本申请的另一方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，该计算机指令用于使该计算机执行本申请任意一项实施例所提供的方法。

[0024] 采用本申请，可以根据至少一种搜索空间的生成策略，得到至少一个搜索空间，对所述至少一个搜索空间，根据搜索空间性能的概率分布进行迭代更新，根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到性能最优的搜索空间，将该性能最优的搜索空间作为目标搜索空间。由于经基于概率分布的迭代更新训练，该目标搜索空间是在众多搜索空间(即至少一个搜索空间)可能性中性能最优的搜索空间，那么，后续该目标搜索空间中搜索得到的模型结构，该模型结构的性能也是最优的，因此，将该目标搜索空间及搜索得到的模型结构，应用到图像处理(如图像分类、图像识别、图像检测)等场景中，可以提高图像处理等场景中的硬件性能，比如硬件的处理速度、处理精度等，而且，随着该硬件性能的提高，还可以降低硬件的使用数量，比如用少于以往数量的硬件同样可以达到以往同等的硬件性能，从而降低了硬件成本。

[0025] 应当理解，本部分所描述的内容并非旨在标识本申请的实施例的关键或重要特征，也不用于限制本申请的范围。本申请的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0026] 附图用于更好地理解本方案，不构成对本申请的限定。其中：

[0027] 图1是根据本申请实施例的搜索空间的生成方法的流程示意图；

[0028] 图2是根据本申请实施例的基于概率分布的搜索方法的流程示意图；

[0029] 图3是根据本申请实施例的搜索空间的生成装置的组成结构示意图；

[0030] 图4是根据本申请实施例的基于概率分布的搜索装置的组成结构示意图；

[0031] 图5是用来实现本申请实施例的搜索空间的生成方法或基于概率分布的搜索方法的电子设备的框图。

具体实施方式

[0032] 以下结合附图对本申请的示范性实施例做出说明，其中包括本申请实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本申请的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0033] 本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。本文中术语“至少一种”表示多种中的任意一种或多种中的至少两种的任意组合，例如，包括A、B、C中

的至少一种，可以表示包括从A、B和C构成的集合中选择的任意一个或多个元素。本文中术语“第一”、“第二”表示指代多个类似的技术用语并对其进行区分，并不是限定顺序的意思，或者限定只有两个的意思，例如，第一特征和第二特征，是指代有两类/两个特征，第一特征可以为一个或多个，第二特征也可以为一个或多个。

[0034] 另外，为了更好的说明本申请，在下文的具体实施方式中给出了众多的具体细节。本领域技术人员应当理解，没有某些具体细节，本申请同样可以实施。在一些实例中，对于本领域技术人员熟知的方法、手段、元件和电路未作详细描述，以便于凸显本申请的主旨。

[0035] 随着人工智能、深度学习技术的发展，为了改善硬件性能，可以根据硬件性能的指标来训练神经网络，以达到将训练后神经网络应用于硬件可以达到符合所预期的该指标的目的。其中，神经网络结构的好坏是至关重要的，神经网络结构的好坏，对最终在硬件上加载基于该神经网络结构的模型所能执行的硬件性能好坏有非常重要的影响。人工设计网络拓扑结构需要非常丰富的经验和众多尝试，并且众多参数会产生爆炸性的组合，采用随机搜索(Random Search)几乎不可行，因此，近期兴起的神经网络架构搜索技术(Neural Architecture Search,NAS)成为研究热点。

[0036] 在NAS中，搜索空间非常重要，以往的NAS工作搜索空间是人工设计好的确定的搜索空间，在确定好的该搜索空间中搜索模型结构。这种搜索空间的人工设计存在局限性，搜索空间确定好之后就固定下来使用，而不能根据用户需求来调整，且搜索空间确定好，人工能设计出来的搜索空间其数量是有限的，即搜索存在上限而不是具备更多的可能性，即便通过该确定好的搜索空间中搜索得到当前所能搜索得到的最优模型结构，该最优模型结构的性能可能也会差强人意，若该确定好的搜索空间本身由于人工设计时不同人理解经验值的差异，导致搜索空间本身就设计的不合适，那么通过该确定好的搜索空间中搜索得到当前所能搜索得到的模型结构，就连是否为最优模型结构也有待评估。

[0037] 采用本申请，可以根据至少一种搜索空间的生成策略，得到至少一个搜索空间，对该至少一个搜索空间进行迭代更新的训练，由于搜索空间非人工设定好，而是可以对其进行迭代更新的自主训练，从而可以得到具备多种可能性的搜索空间，搜索空间呈现多样化，非确定性。而且，搜索模型结构的过程不再是搜索模型结构，而是先从迭代更新训练后得到的至少一个搜索空间中搜索得到一个性能最优的搜索空间，并作为目标搜索空间，然后在该目标搜索空间中搜索模型结构，从而，不仅可以从目标搜索空间中找到最优的模型结构，还可以指定用户据此来更好的设计最优的模型结构，当部署该最优的模型结构到相应的硬件上时，就可以达到预期的最优硬件性能，如最优的处理速度、最优的处理精度等，同时，随着该硬件性能的提高，还可以降低硬件的使用数量，比如用少于以往数量的硬件同样可以达到以往同等的硬件性能，从而降低了硬件成本。

[0038] 本申请的适用范围，除了可以适用于上述提及的人工智能、深度学习、云计算、图像处理等领域，还可以适用于模型压缩的PaddleSlim、用于云计算的Paddlecloud、用于图像识别的EasyDL及人工智能(Artificial Intelligenc,AI)小程序等等。其中，所述Paddle为一个深度学习框架的称谓，所谓Paddle，指并行分布式深度学习(Parallel Distributed Deep Learning)的缩写，可以基于该深度学习框架部署适应多种应用场景的模型结构训练。PaddleSlim除了可以在模型压缩中实现量化功能，还可以在模型压缩中集成剪裁、蒸馏、模型结构搜索、模型硬件搜索等。Paddlecloud是适应于“云计算+大数据+人工智能”三

位一体的定位趋势，可以在云端部署所需的模型并去分担大量的计算处理逻辑。EasyDL是定制化图像识别平台，可以基于生成的模型结构得到SDK或API接口服务)。AI可以实现各种针对不同应用场景的小程序等，通过AI可以开发用于模拟、延伸和扩展人工智能的技术，可以通过所需的模型来呈现对应不同应用场景且利用到人类智能的技术，侧重于用户交互。

[0039] 根据本申请的实施例，提供了一种搜索空间的生成方法，图1是根据本申请实施例的搜索空间的生成方法的流程示意图，该方法可以应用于搜索空间的生成装置，例如，该装置可以部署于终端或服务器或其它处理设备执行的情况下，可以执行诸如图像分类、图像识别、图像检测等图像处理场景中，以及对视频提取出视频帧后的分类、识别、检测等视频处理场景等等。其中，终端可以为用户设备(UE,User Equipment)、移动设备、蜂窝电话、无绳电话、个人数字处理(PDA,Personal Digital Assistant)、手持设备、计算设备、车载设备、可穿戴设备等。在一些可能的实现方式中，该方法还可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。如图1所示，包括：

[0040] S101、根据至少一种搜索空间的生成策略，得到至少一个搜索空间。

[0041] 一示例中，搜索空间的生成策略，包括如下任意一种或多种的组合：

[0042] 1)根据选定训练模型中层的类型及所述训练模型中层的个数，来生成所述搜索空间；

[0043] 2)根据选定训练模型中层的类型，来生成所述搜索空间；

[0044] 3)根据选定训练模型中层的卷积属性及所述训练模型中层的个数，来生成所述搜索空间；

[0045] 4)根据选定训练模型中层的拓扑结构，来生成所述搜索空间；其中，所述拓扑结构包括单分支的拓扑结构、或多分支的拓扑结构。

[0046] 一示例中，可以通过搜索空间生成器及至少一种搜索空间的生成策略，得到至少一个搜索空间。

[0047] S102、对所述至少一个搜索空间，根据搜索空间性能的概率分布进行迭代更新，根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到目标搜索空间。

[0048] 一示例中，针对迭代更新而言，迭代次数为0时，对应的是该至少一个搜索空间，即初始的搜索空间，随着不断迭代更新，比如迭代次数达到迭代规则中预设的迭代次数100时，对应的是迭代更新训练后得到的训练后至少一个搜索空间，即该迭代收敛所得到的搜索空间，以便从所述训练后至少一个搜索空间中搜索得到该目标搜索空间。

[0049] 一示例中，在通过搜索空间生成器及至少一种搜索空间的生成策略得到至少一个搜索空间的情况下，若所述搜索空间生成器的更新次数未达预设的迭代规则，则继续迭代更新训练所述至少一个搜索空间，直至满足该迭代规则则迭代收敛，结束训练，搜索得到该目标搜索空间。其中，针对预设的迭代规则而言，预设的迭代规则可以是达到预设迭代次数，比如预设迭代次数为100或200次等等，已经达到该100或200次等，则迭代收敛，结束训练；预设的迭代规则也可以是基于目标参数进行迭代训练，相应的性能已经连续达到预设次数，比如预设次数为50次或100次等等，如果该性能连续50次或100次达到目标且没有继续优化，则迭代收敛，结束训练。

[0050] 相关技术中，是采用一个人工设计好的搜索空间(搜索空间是确定的，不具备多种可能性)中存在多个模型结构，从搜索空间中，搜索得到硬件性能“如处理速度、处理精度”

最优的模型结构，只能在限定数量的该搜索空间内搜索最优的模型结构，这限制了搜索空间的更多可能性。比如，以Mnasnet为例，模型结构只能限定在搜索通道数，膨胀系数等，模型结构实际上还是mobilenet_v2-like结构，也就是说，在确定好的搜索空间中搜索模型结构，一方面，限制了可以搜索得到的模型结构的上界；另一方面，如果在不合适的搜索空间中，即使可以找到最优的模型结构，模型结构的性能也会很差。

[0051] 采用本申请，可以根据至少一种搜索空间的生成策略，得到至少一个搜索空间，根据搜索空间性能的概率分布进行迭代更新，在迭代更新至迭代收敛的情况下可以得到性能最优的搜索空间，将该性能最优的搜索空间作为目标搜索空间。本申请中，由于搜索空间是根据各自可能性的定义自动生成的，也就是说，搜索空间具备多种可能性，是不确定的，搜索空间的生成过程中，可以不断迭代训练，且基于概率分布进行迭代训练，比如，可以根据搜索空间性能的概率分布对迭代训练得到的所有搜索空间进行有条件的搜索空间采样，得到在该不断迭代训练中采样得到的至少一个搜索空间，以便从该采样得到的至少一个搜索空间中搜索最优的搜索空间，也就是说，经迭代更新的训练，该目标搜索空间是在众多搜索空间(即至少一个搜索空间)可能性中性能最优的搜索空间，基于概率分布进行迭代训练还可以更快速及准确的得到该性能最优的搜索空间。那么，后续从该目标搜索空间中搜索得到的模型结构，该模型结构的性能也是最优的，则得到该性能最优的搜索空间之后，从该最优的搜索空间中可以搜索得到硬件性能“如处理速度、处理精度”最好的模型结构，因此，将该目标搜索空间及搜索得到的模型结构，应用到图像处理(如图像分类、图像识别、图像检测)等场景中，可以提高图像处理等场景中的硬件性能，比如硬件的处理速度、处理精度等，而且，随着该硬件性能的提高，还可以降低硬件的使用数量，比如用少于以往数量的硬件同样可以达到以往同等的硬件性能，从而降低了硬件成本。

[0052] 就上述搜索空间的生成策略而言，可以是针对该搜索空间的规则定义，一些示例中，该搜索空间中的空间/集合包含了所有搜索空间的可能性可以包括如下内容：

[0053] 1)以训练模型中层(block)为例，如果搜索空间选定层的类型为残差网络中的残差层(residual block)，层的个数选定为resnet50(resnet50为卷积网络结构的设计)，则据此生成的搜索空间可以搜索任意resnet50-like结构的模型结构。其中，block在训练模型中也可以称为块，不限定具体的名称，只要是构成训练模型中的模块或层，都在本申请的保护范围之内。

[0054] 2)基于上述1)，如果搜索空间如果不限定层数，只选定层的类型为residual block，则据此生成的搜索空间可以搜索任意resnet-like结构的模型结构。

[0055] 3)以训练模型中层(block)为例，如果搜索空间选定层的卷积属性为depth-wise block，depth-wise block为深度可分离卷积属性的层类型，层的个数选定为mobilenet_v2(mobilenet_v2为深度可分离卷积网络结构中的一种设计)，则据此生成的搜索空间可以搜索任意mobilenet_v2-like结构的模型结构。

[0056] 4)如果选定训练模型中层的拓扑结构为darts，darts为单分支的拓扑结构，则据此生成的搜索空间可以搜索任意darts结构的模型结构。

[0057] 需要指出的是：搜索空间除了可以为上述根据该搜索空间的生成策略得到，还可以通过人工方式得到已确定的搜索空间，也就是说，本申请中的搜索空间不限定于人工设定的已确定的搜索空间，还包括上述根据该搜索空间的生成策略得到的非确定的各种搜索

空间的可能性(不限于上述示例中的可能性),并且,已确定的搜索空间、及非确定的各种搜索空间的可能性还可以进一步根据用户设计并应用该方法的硬件(如芯片上某功能模块的算法逻辑)需求来自由的任意组合。

[0058] 需要说明的是:上述“空间/集合”实际上也是搜索空间,只是由于本申请中的搜索空间不确定,而是具备多种可能性且能不断迭代更新,则存在不止一个搜索空间,为了方便描述,相比于更宽泛含义的“搜索空间”自身,也可以将“搜索空间”中的“空间/集合”称之为“子搜索空间”。

[0059] 根据本申请的实施例,提供了一种基于概率分布的搜索方法,图2是根据本申请实施例的基于概率分布的搜索方法的流程示意图,如图2所示,包括:

[0060] S201、根据至少一种搜索空间的生成策略,得到至少一个搜索空间。

[0061] 一示例中,搜索空间的生成策略,包括如下任意一种或多种的组合:

[0062] 1)根据选定训练模型中层的类型及所述训练模型中层的个数,来生成所述搜索空间;

[0063] 2)根据选定训练模型中层的类型,来生成所述搜索空间;

[0064] 3)根据选定训练模型中层的卷积属性及所述训练模型中层的个数,来生成所述搜索空间;

[0065] 4)根据选定训练模型中层的拓扑结构,来生成所述搜索空间;其中,所述拓扑结构包括单分支的拓扑结构、或多分支的拓扑结构。

[0066] 一示例中,可以通过搜索空间生成器及至少一种搜索空间的生成策略,得到至少一个搜索空间。

[0067] S202、对所述至少一个搜索空间,根据搜索空间性能的概率分布以迭代更新的方式进行训练,在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个搜索空间。

[0068] 一示例中,该第一目标参数,可以包括:用于所述搜索空间性能评估的第一超参数,可以通过该第一超参数来衡量搜索空间性能的优劣,如衡量如平均性能、性能中位数、性能方差等搜索空间性能。

[0069] S203、响应于第一搜索操作,在所述训练后的至少一个搜索空间中搜索得到目标搜索空间。

[0070] 一示例中,该目标搜索空间可以为该训练后的至少一个搜索空间中的最优搜索空间,即从具备更多可能性的该训练后的至少一个搜索空间中,选取其中硬件性能最好的一个最优搜索空间。

[0071] 还可以包括:

[0072] S204、对所述目标搜索空间中的至少一个模型结构,根据模型结构性能的概率分布以迭代更新的方式进行训练,在所述至少一个模型结构根据第二目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个模型结构。

[0073] 一示例中,该第二目标参数,包括:用于所述模型结构性能评估的第二超参数,可以通过该第二超参数来衡量搜索空间性能的优劣,如衡量如平均性能、性能中位数、性能方差等搜索空间性能。

[0074] S205、响应于第二搜索操作,在所述训练后的至少一个模型结构中搜索得到目标

模型结构。

[0075] 一示例中，除了基于概率分布的迭代训练可以得到上述S203中的目标搜索空间，还可以执行本S204-S205，即从具备更多可能性的该训练后的至少一个搜索空间中，选取其中硬件性能最好的一个最优搜索空间之后，还可以在该最优搜索空间中去搜索模型结构的搜索过程中，即根据该最优搜索空间对所述目标搜索空间中的至少一个模型结构，根据模型结构性能的概率分布以迭代更新的方式进行训练，在所述至少一个模型结构根据第二目标参数迭代更新至迭代收敛的情况下得到最优的模型结构，该最优模型结构即为目标模型结构。由于通过上述S203可以得到最优搜索空间，进而通过S204-S205还可以根据该最优搜索空间，也基于概率分布的迭代训练以得到最优模型结构，因此，从该最优搜索空间中去搜索模型结构，就可以优中选优，得到该最优模型结构。

[0076] 采用本申请，可以根据至少一种搜索空间的生成策略，得到至少一个搜索空间，对所述至少一个搜索空间基于概率分布进行迭代更新，在迭代更新至迭代收敛的情况下可以得到性能最优的搜索空间，将该性能最优的搜索空间作为目标搜索空间。响应于第一搜索操作，在所述训练后的至少一个搜索空间中搜索得到目标搜索空间。进一步，对于该目标搜索空间中的至少一个模型结构，也可以基于概率分布进行迭代更新，在迭代更新至迭代收敛的情况下得到训练后的至少一个模型结构，响应于第二搜索操作，在训练后的至少一个模型结构中搜索得到最优的模型结构，该最优模型结构即为目标模型结构。本申请中，由于搜索空间是根据各自可能性的定义自动生成的，也就是说，搜索空间具备多种可能性，是不确定的，搜索空间的生成过程中，可以不断迭代训练，且基于概率分布(通过概率分布进行约束的有条件形式)进行搜索空间的迭代更新，以便从不断迭代训练的搜索空间中搜索最优的搜索空间，且相比无条件的迭代更新，更快速及准确，也就是说，经迭代更新的训练，该目标搜索空间是在众多搜索空间(即至少一个搜索空间)可能性中性能最优的搜索空间。之后，可以从该目标搜索空间中搜索得到模型结构，则该模型结构的性能也是最优的。也就是说，可以得到该性能最优的搜索空间之后，从该最优的搜索空间中可以搜索得到硬件性能“如处理速度、处理精度”最好的模型结构，且还可以基于概率分布(通过概率分布进行约束的有条件形式)进行模型结构的迭代更新，相比无条件的迭代更新，更快速及准确，因此，将该目标搜索空间及搜索得到的模型结构，应用到图像处理(如图像分类、图像识别、图像检测)等场景中，可以提高图像处理等场景中的硬件性能，比如硬件的处理速度、处理精度等，而且，随着该硬件性能的提高，还可以降低硬件的使用数量，比如用少于以往数量的硬件同样可以达到以往同等的硬件性能，从而降低了硬件成本。

[0077] 一实施方式中，所述根据至少一种搜索空间的生成策略，得到至少一个搜索空间，包括：初始化搜索空间生成器，比如，根据搜索空间中的子空间/子集合来初始化搜索空间生成器。根据该搜索空间生成器及该至少一种搜索空间的生成策略，得到该至少一个搜索空间。

[0078] 一实施方式中，对所述至少一个搜索空间，根据搜索空间性能的概率分布以迭代更新的方式进行训练，在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下，结束所述训练，得到训练后的至少一个搜索空间，包括：对所述至少一个搜索空间，根据搜索空间性能的概率分布进行建模，得到第一概率模型(如基于概率分布对搜索空间建模得到的概率模型)；将用于所述搜索空间性能评估的第一超参数作为所述第一目标参

数;根据所述第一超参数,迭代更新所述第一概率模型,以基于所述第一概率模型对所述至少一个搜索空间进行迭代更新至迭代收敛。

[0079] 一实施方式中,对所述目标搜索空间中的至少一个模型结构,根据模型结构性能的概率分布以迭代更新的方式进行训练,在所述至少一个模型结构根据第二目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个模型结构,包括:对所述至少一个模型结构,根据模型结构性能的概率分布进行建模,得到第二概率模型(如基于概率分布对模型结构建模得到的概率模型);将用于所述模型结构性能评估的第二超参数作为所述第二目标参数;根据所述第二超参数,迭代更新所述第二概率模型,以基于所述第二概率模型对所述至少一个模型结构进行迭代更新至迭代收敛。

[0080] 基于上述任意本申请实施例、实施方式及其组合,还包括:可以获取待处理的图像,将所述待处理的图像,输入根据所述目标搜索空间搜索得到的目标模型结构进行图像处理,得到目标图像。其中,所述图像处理包括:图像分类、图像识别、图像检测中的至少一种处理。

[0081] 应用示例:

[0082] 应用本申请实施例一搜索模型结构的处理流程包括如下内容:

[0083] 一、收集一批目标场景的数据(如用于图像分类、图像识别、图像检测等场景的数据),数据可以是从数据库中获取的已标注数据,也可以是通过标注人员标注得到的所需标注数据,其中,每个标注人员可以根据自己的主观判断给数据进行打分。多个标注人员同时打分,数据标注结果根据最终标注得分得到,该最终标注得分可以为多个标注人员的平均值,最终得到该所需的标注数据,该标注数据可以用于以下的搜索空间及模型结构搜索的训练过程中。

[0084] 二、生成搜索空间及基于概率分布在搜索空间中搜索模型结构。

[0085] 1、根据第一概率模型建模搜索空间的性能,旨在通过第一概率模型预测任意搜索空间的平均性能,其中,第一概率模型中的第一超参数可以随机初始化,并随着基于该第一概率模型采样搜索空间而逐步更新。

[0086] 2、根据第一概率模型对所述至少一个搜索空间进行迭代更新至迭代收敛,得到训练后的至少一个搜索空间,以预测足够多的搜索空间的性能。

[0087] 3、从上述2中预测的该训练后的至少一个搜索空间中选取最优的搜索空间,并作为目标搜索空间。

[0088] 4、根据上述3中的目标搜索空间,建模该搜索空间中模型结构性能的第二概率模型,其中,第二概率模型中的第二超参数可以随机初始化,并随着基于该第二概率模型采样模型结构而逐步更新。

[0089] 5、根据第二概率模型对所述至少一个模型结构进行迭代更新至迭代收敛,得到训练后的至少一个模型结构,以预测足够多的模型结构的性能。

[0090] 6、从上述3中以采样方式选取top k的模型结构进行训练,并记录这些模型结构的性能,其中,k为对模型结构进行采样的采样个数。

[0091] 7、根据上述6中采样得到的top k模型结构的性能,更新第二概率模型中的第二超参数,其中,k为对模型结构进行采样的采样个数。

[0092] 8、若第二超参数更新次数未达预设的迭代规则,则返回上述5。

[0093] 9、根据采样的搜索空间中的模型结构的综合性能(如平均性能,中位性能)更新第一概率模型中的第一超参数。

[0094] 10、若第一超参数更新次数未达预设的迭代规则,则返回上述2。

[0095] 11、输出最优的搜索空间。

[0096] 12、输出最优搜索空间中的最优模型结构。

[0097] 采用本应用示例,搜索过程不再是在人工设定所确定好的搜索空间中搜索模型结构,而是对具备多种可能性的搜索空间进行搜索,以搜索得到该最优的搜索空间,基于该最优的搜索空间来搜索最优的模型结构,从而不仅可以从中找到最优的模型结构,还可以指导用户更好的设计模型结构。提升了模型在特定硬件上的处理速度、处理精度等硬件性能,还可以同时降低产品的硬件成本。

[0098] 根据本申请的实施例,提供了一种搜索空间的生成装置,图3是根据本申请实施例的搜索空间的生成装置的组成结构示意图,如图3所示,该装置包括第一处理模块31,用于根据至少一种搜索空间的生成策略,得到至少一个搜索空间;第二处理模块32,用于对所述至少一个搜索空间,根据搜索空间性能的概率分布进行迭代更新,根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到目标搜索空间。

[0099] 一实施方式中,所述搜索空间的生成策略,包括如下任意一种或多种:

[0100] 根据选定训练模型中层的类型及所述训练模型中层的个数,来生成所述搜索空间;

[0101] 根据选定训练模型中层的类型,来生成所述搜索空间;

[0102] 根据选定训练模型中层的卷积属性及所述训练模型中层的个数,来生成所述搜索空间;

[0103] 根据选定训练模型中层的拓扑结构,来生成所述搜索空间;其中,所述拓扑结构包括单分支的拓扑结构、或多分支的拓扑结构。

[0104] 根据本申请的实施例,提供了一种基于概率分布的搜索装置,图4是根据本申请实施例的基于概率分布的搜索装置的组成结构示意图,如图4所示,该装置包括第一搜索模块41,用于根据至少一种搜索空间的生成策略,得到至少一个搜索空间;第二搜索模块42,用于对所述至少一个搜索空间,根据搜索空间性能的概率分布以迭代更新的方式进行训练,在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个搜索空间;第三搜索模块43,用于响应于第一搜索操作,在所述训练后的至少一个搜索空间中搜索得到目标搜索空间。

[0105] 一实施方式中,所述第一目标参数,包括:用于所述搜索空间性能评估的第一超参数。

[0106] 一实施方式中,还包括第四搜索模块,用于对所述目标搜索空间中的至少一个模型结构,根据模型结构性能的概率分布以迭代更新的方式进行训练,在所述至少一个模型结构根据第二目标参数迭代更新至迭代收敛的情况下,结束所述训练,得到训练后的至少一个模型结构;响应于第二搜索操作,在所述训练后的至少一个模型结构中搜索得到目标模型结构。

[0107] 一实施方式中,所述第二目标参数,包括:用于所述模型结构性能评估的第二超参数。

[0108] 一实施方式中，所述第二搜索模块，用于对所述至少一个搜索空间，根据搜索空间性能的概率分布进行建模，得到第一概率模型；将用于所述搜索空间性能评估的第一超参数作为所述第一目标参数；根据所述第一超参数，迭代更新所述第一概率模型，以基于所述第一概率模型对所述至少一个搜索空间进行迭代更新至迭代收敛。

[0109] 一实施方式中，所述第四搜索模块，用于对所述至少一个模型结构，根据模型结构性能的概率分布进行建模，得到第二概率模型；将用于所述模型结构性能评估的第二超参数作为所述第二目标参数；根据所述第二超参数，迭代更新所述第二概率模型，以基于所述第二概率模型对所述至少一个模型结构进行迭代更新至迭代收敛。

[0110] 基于上述任意本申请实施例、实施方式及其组合，还包括图像处理模块，用于获取待处理的图像；将所述待处理的图像，输入根据所述目标搜索空间搜索得到的目标模型结构进行图像处理，得到目标图像。其中，所述图像处理包括：图像分类、图像识别、图像检测中的至少一种处理。

[0111] 本申请实施例各装置中的各模块的功能可以参见上述方法中的对应描述，在此不再赘述。

[0112] 根据本申请的实施例，本申请还提供了一种电子设备和一种可读存储介质。

[0113] 如图5所示，是用来实现本申请实施例的搜索空间的生成方法或基于概率分布的搜索方法的电子设备的框图。该电子设备可以为前述部署设备或代理设备。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0114] 如图5所示，该电子设备包括：一个或多个处理器801、存储器802，以及用于连接各部件的接口，包括高速接口和低速接口。各个部件利用不同的总线互相连接，并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理，包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如，耦合至接口的显示设备)上显示GUI的图形信息的指令。在其它实施方式中，若需要，可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样，可以连接多个电子设备，各个设备提供部分必要的操作(例如，作为服务器阵列、一组刀片式服务器、或者多处理器系统)。图5中以一个处理器801为例。

[0115] 存储器802即为本申请所提供的非瞬时计算机可读存储介质。其中，所述存储器存储有可由至少一个处理器执行的指令，以使所述至少一个处理器执行本申请所提供的搜索空间的生成方法或基于概率分布的搜索方法。本申请的非瞬时计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行本申请所提供的搜索空间的生成方法或基于概率分布的搜索方法。

[0116] 存储器802作为一种非瞬时计算机可读存储介质，可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块，如本申请实施例中的搜索空间的生成方法或基于概率分布的搜索方法对应的程序指令/模块(例如，附图3所示的搜索空间的生成装置中的第一处理模块、第二处理模块等模块；附图4所示的搜索装置中的第一搜索模块、第二搜索模块、第

三搜索模块等模块)。处理器801通过运行存储在存储器802中的非瞬时软件程序、指令以及模块,从而执行服务器的各种功能应用以及数据处理,即实现上述方法实施例中的搜索空间的生成方法或基于概率分布的搜索方法。

[0117] 存储器802可以包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需要的应用程序;存储数据区可存储根据电子设备的使用所创建的数据等。此外,存储器802可以包括高速随机存取存储器,还可以包括非瞬时存储器,例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中,存储器802可选包括相对于处理器801远程设置的存储器,这些远程存储器可以通过网络连接至电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0118] 搜索空间的生成方法或基于概率分布的搜索方法的电子设备,还可以包括:输入装置803和输出装置804。处理器801、存储器802、输入装置803和输出装置804可以通过总线或者其他方式连接,图5中以通过总线连接为例。

[0119] 输入装置803可接收输入的数字或字符信息,以及产生与电子设备的用户设置以及功能控制有关的键信号输入,例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置804可以包括显示设备、辅助照明装置(例如,LED)和触觉反馈装置(例如,振动电机)等。该显示设备可以包括但不限于,液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中,显示设备可以是触摸屏。

[0120] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0121] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令,并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的,术语“机器可读介质”和“计算机可读介质”指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如,磁盘、光盘、存储器、可编程逻辑装置(PLD)),包括,接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0122] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0123] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据

服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0124] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

[0125] 采用本申请,可以根据至少一种搜索空间的生成策略,得到至少一个搜索空间,对所述至少一个搜索空间,根据搜索空间性能的概率分布进行迭代更新,根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到性能最优的搜索空间,将该性能最优的搜索空间作为目标搜索空间。由于经基于概率分布的迭代更新训练,该目标搜索空间是在众多搜索空间(即至少一个搜索空间)可能性中性能最优的搜索空间,那么,后续该目标搜索空间中搜索得到的模型结构,该模型结构的性能也是最优的,因此,将该目标搜索空间及搜索得到的模型结构,应用到图像处理(如图像分类、图像识别、图像检测)等场景中,可以提高图像处理等场景中的硬件性能,比如硬件的处理速度、处理精度等,而且,随着该硬件性能的提高,还可以降低硬件的使用数量,比如用少于以往数量的硬件同样可以达到以往同等的硬件性能,从而降低了硬件成本。

[0126] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本申请公开的技术方案所期望的结果,本文在此不进行限制。

[0127] 上述具体实施方式,并不构成对本申请保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等,均应包含在本申请保护范围之内。

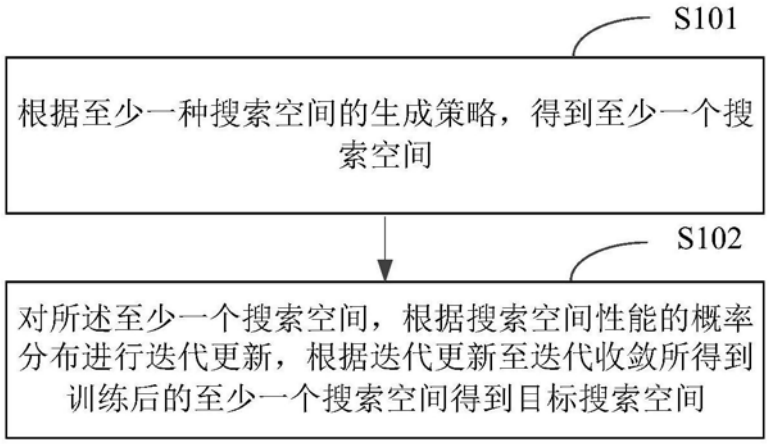

S101
根据至少一种搜索空间的生成策略，得到至少一个搜索空间
S102
对所述至少一个搜索空间，根据搜索空间性能的概率分布进行迭代更新，根据迭代更新至迭代收敛所得到训练后的至少一个搜索空间得到目标搜索空间

图1

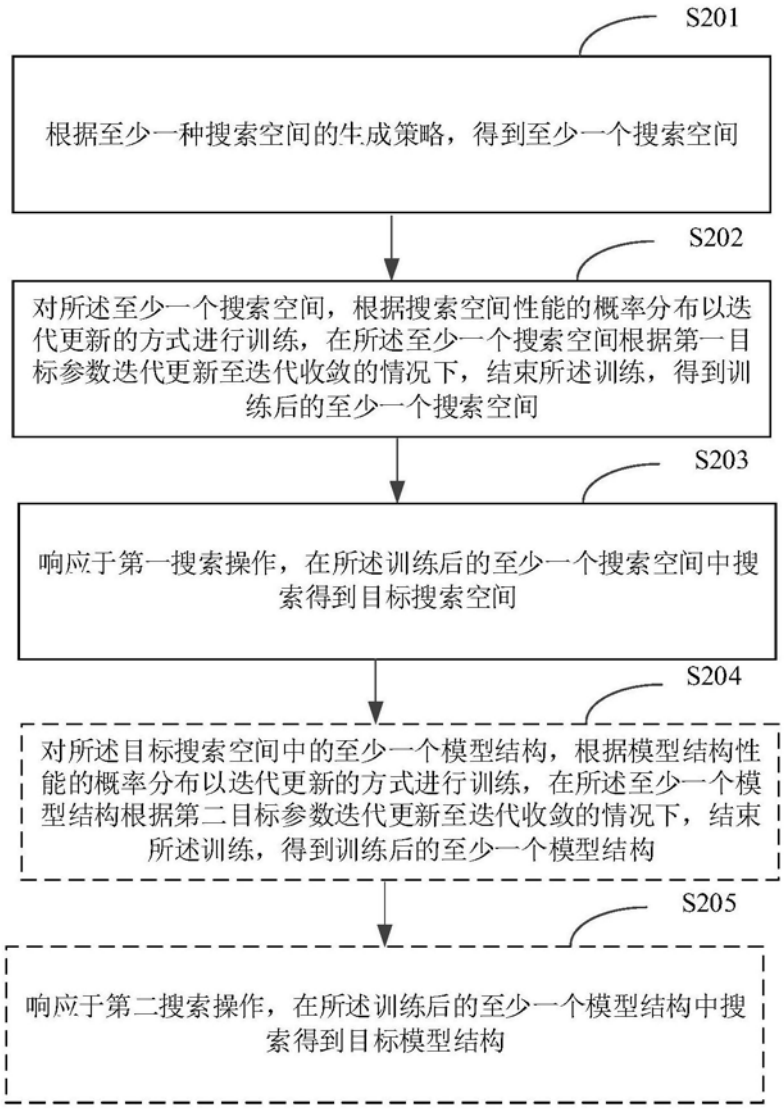

S201
根据至少一种搜索空间的生成策略，得到至少一个搜索空间
S202
对所述至少一个搜索空间，根据搜索空间性能的概率分布以迭代更新的方式进行训练，在所述至少一个搜索空间根据第一目标参数迭代更新至迭代收敛的情况下，结束所述训练，得到训练后的至少一个搜索空间
S203
响应于第一搜索操作，在所述训练后的至少一个搜索空间中搜索得到目标搜索空间
S204
对所述目标搜索空间中的至少一个模型结构，根据模型结构性能的概率分布以迭代更新的方式进行训练，在所述至少一个模型结构根据第二目标参数迭代更新至迭代收敛的情况下，结束所述训练，得到训练后的至少一个模型结构
S205
响应于第二搜索操作，在所述训练后的至少一个模型结构中搜索得到目标模型结构

图2

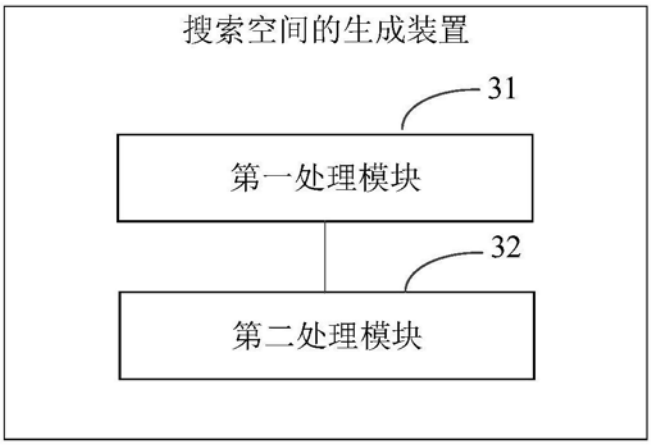

搜索空间的生成装置
31
第一处理模块
32
第二处理模块

图3

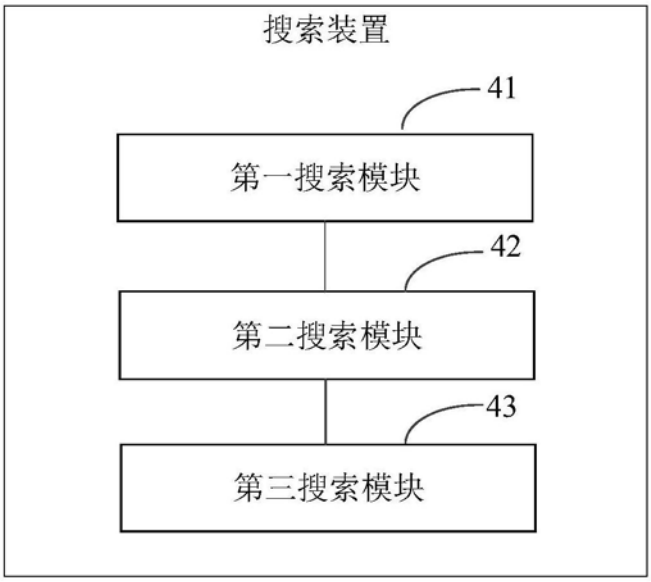

搜索装置
41
第一搜索模块
42
第二搜索模块
43
第三搜索模块

图4

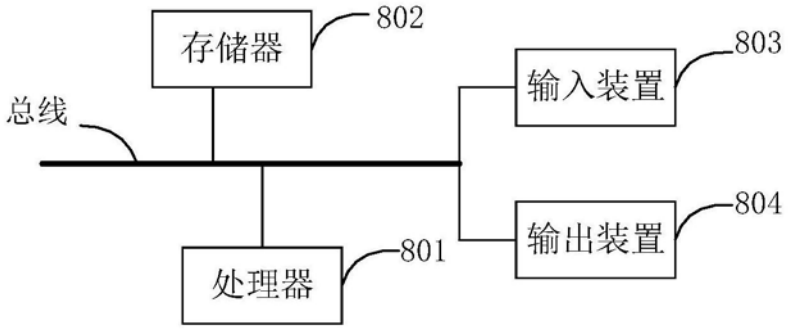

存储器
802
总线
输入装置
803
输出装置
804
处理器
801

图5