

一种神经网络构建方法以及系统

本申请要求于 2020 年 05 月 19 日提交中国专利局、申请号为“202010425173.3”、申请名称为“一种神经网络构建方法以及系统”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。

5 技术领域

本申请涉及人工智能领域，尤其涉及一种神经网络构建方法以及系统。

背景技术

在人工智能领域中，神经网络近年来在图像、视频和语音等多种媒体信号的处理与分析任务中取得了卓越的成就。一个性能良好的神经网络往往拥有精妙的网络架构，需要具有高超技能和丰富经验的人类专家花费大量精力进行设计。对于普通非专家用户而言，针对具体问题进行神经网络的设计是不现实的。骨干网络是神经网络中的重要网络架构，与神经网络的具体任务通常并无关联，如可以是针对大部分任务都需要的特征提取网络或残差网络等。例如，常用计算机视觉任务主要用于目标检测以及语义分割，通常可以先通过骨干网络对输入信息进行特征提取，然后将提取到的特征输入到预测模块，得到预测结果。而随着待处理场景的复杂度增加，对骨干网络特征提取能力的需求也增加，对骨干网络的设计难度也随之增加。

已经提出了一些方案用于模型的架构搜索，然而，现有方案中的架构搜索通常只针对单一任务，不同的场景则需要重新设计或者重新搜索网络架构。因此在面向新应用场景进行迁移的过程中需要大量的重构，调试的时间开销大。

发明内容

本申请提供了一种神经网络构建方法以及系统，用于通过将初始骨干网络中的部分基本单元替换为占位模块的方式来构建目标神经网络，从而可以根据不同的场景构建出不同的目标神经网络，泛化能力强，且用户友好。

第一方面，本申请提供一种神经网络构建方法，包括：获取初始骨干网络和候选集，初始骨干网络用于构建目标神经网络；将初始骨干网络中的至少一个基本单元替换为至少一个占位模块，得到待定网络，候选集中包括与至少一个占位模块对应的多种结构的参数；根据候选集进行采样，得到至少一个采样结构的信息；根据待定网络和至少一个采样结构的信息获取网络模型，其中，至少一个采样结构的信息用于确定至少一个占位模块的结构；若网络模型符合预设条件，则将网络模型作为目标神经网络。

因此，本申请实施方式中，通过在骨干网络中设置占位模块，改变占位模块的结构来改变骨干网络的结构。可以针对不同的场景来改变占位模块的结构或者位置等，可以适应不同的场景，泛化能力强。并且，即使在面向新的应用场景，也无需进行大量的迁移或者重构，降低了代码的调试时间，提高用户体验。此外，本申请实施方式中，用户可以仅提供初始骨干网络，或者可选地还提供候选集，本申请实施方式提供的神经网络构建方法，即可得到满足预设条件的目标神经网络，降低了用户的学习难度，易用性高，提供了一种

用户友好的神经网络构建方法。

在一种可能的实施方式中，若网络模型不符合预设条件，则重新根据候选集进行采样，并根据重新采样得到的至少一个采样结构的信息更新网络模型。

因此，本申请实施方式中，当网络模型不符合预设条件时，可以重新根据候选集进行采样，从而通过重新改变占位模块的结构或者位置等，得到更新后的网络模型，进一步提
5 高得到符合预设条件的网络模型的可能性。

在一种可能的实施方式中，在根据候选集进行采样，得到至少一个采样结构的信息之前，该方法还包括：根据候选集构建参数空间，参数空间中包括与候选集中的多种结构的参数对应的架构参数；上述的根据候选集进行采样，得到至少一个采样结构的信息，可以
10 包括：对参数空间进行采样，得到与至少一个采样结构对应的至少一组采样参数。

因此，本申请实施方式中，可以根据候选集构建参数空间，该参数空间中包括了候选集中的结构的参数对应的架构参数，后续可以采集参数空间中的架构参数，相对于直接从候选架构中采集架构，减少了采样的数据量，提高采样效率。

在一种可能的实施方式中，前述的根据待定网络和至少一个采样结构的信息获取网络模型，可以包括：根据至少一组采样参数对待定网络中的至少一个占位模块的结构进行转换，得到网络模型。
15

因此，本申请实施方式中，可以根据至少一组采样参数对待定网络中的至少一个占位模块的结构进行转换，从而将该至少一个占位模块的结构转换为与该至少一组采样参数对应的结构，从而得到网络模型，提供了一种得到网络模型的具体方式。

在一种可能的实施方式中，在根据待定网络和至少一个采样结构的信息获取网络模型之前，上述方法还可以包括：根据候选集和待定网络构建多种结构，多种结构组成结构搜索空间；根据待定网络和至少一个采样结构的信息获取网络模型，可以包括：根据至少一组采样参数从结构搜索空间搜索网络模型。
20

因此，本申请实施方式中，可以根据采样参数直接从结构搜索空间中搜索网络模型，可以快速搜索得到网络模型。
25

在一种可能的实施方式中，根据候选集进行采样的采样方式，包括：随机采样；或者，按照预设规则采样。本申请实施方式中，提供了更丰富的采样方式，可以根据实际应用场景选择合适的采样方式。

在一种可能的实施方式中，若采样方式为按照预设规则采样，则在确定网络模型不符合预设条件之后，前述方法还可以包括：根据网络模型的评估结果采用预设的优化算法更新预设规则。因此，本申请实施方式中，可以根据网络模型的评估结果采用优化算法更新采样方式，从而使得下一次采样时，可以采样到更优的采样参数。
30

在一种可能的实施方式中，上述的优化算法可以包括但不限于：进化算法、强化学习算法、贝叶斯优化算法或梯度优化算法。

在一种可能的实施方式中，前述的预设条件包括以下一项或者多项：获取网络模型的次数超过预设次数、获取网络模型的时长超过预设时长或者网络模型的输出结果符合预设要求。因此，本申请实施方式中，可以得到符合要求的网络模型。
35

在一种可能的实施方式中，候选集中包括以下一项或者多项：算子的种类、算子的属性信息或者算子之间的连接方式。因此，本申请实施方式中，候选集中包括了算子的多种信息，从而使得后续可以根据候选集中所包括的结构的信息，确定占位模块的具体结构，从而根据不同的需要选择不同的算子结构，实现不同场景下对网络模型的搜索。

5 在一种可能的实施方式中，目标神经网络用于进行图片识别、语义分割或者目标检测中的至少一种。因此，本申请实施方式提供的神经网络构建方法，可以应用于多种场景，如图片识别、语义分割或者目标检测等场景。

在一种可能的实施方式中，在将网络模型作为目标神经网络之后，前述的方法还可以包括：通过预设的数据集对目标神经网络进行训练，得到训练后的目标神经网络。因此，
10 本申请实施方式中，得到目标神经网络之后，还可以对目标神经网络进行训练，从而使目标神经网络的输出精确度更高。

在一种可能的实施方式中，获取初始骨干网络和候选集，包括：接收用户输入数据；从用户输入数据中获取初始骨干网络和候选集。因此，本申请实施方式中，可以由用户选择初始骨干网络和候选集，从而可以使用户可以在已有的初始骨干网络的基础上，使用候
15 选集对初始骨干网络中的部分基本单元进行替换，从而得到更优的目标神经网络。

第二方面，本申请提供一种神经网络构建系统，可以包括：输入模块、采样模块、架构构建器、架构评估器；

输入模块，用于获取初始骨干网络和候选集，初始骨干网络用于构建目标神经网络；

架构构建器，用于将初始骨干网络中的至少一个基本单元替换为至少一个占位模块，
20 得到待选网络，候选集中包括与至少一个占位模块对应的多种结构的参数；

采样模块，用于根据候选集进行采样，得到至少一个采样结构的信息；

架构构建器，还用于根据占位骨干网络待选网络和至少一个采样结构的信息获取网络模型，其中，至少一个采样结构的信息用于确定至少一个占位模块的结构；

架构评估器，用于评估网络模型是否符合预设条件，若网络模型符合预设条件，则将
25 网络模型作为目标神经网络；若网络模型不符合预设条件，则由采样模块重新根据候选集进行采样，架构构建器根据重新采样得到的至少一个采样结构的信息更新网络模型。

第二方面及第二方面任一种可能的实施方式产生的有益效果可参照第一方面及第一方面任一种可能实施方式的描述。

在一种可能的实施方式中，若网络模型不符合预设条件，则采样模块还用于重新根据
30 候选集进行采样，架构构建器还用于根据重新采样得到的至少一个采样结构的信息更新网络模型。

在一种可能的实施方式中，架构构建器，还用于根据候选集构建参数空间，参数空间中包括与多种结构的参数对应的架构参数；

采样模块，具体用于对参数空间进行采样，得到与至少一个采样结构对应的至少一组
35 采样参数。

在一种可能的实施方式中，架构构建器，具体用于根据至少一个采样结构对待选网络中的至少一个占位模块的结构进行转换，得到网络模型。

在一种可能的实施方式中，架构构建器，还用于在根据待定网络和至少一个采样结构的信息获取网络模型之前，根据候选集和待定网络构建多种结构，多种结构组成结构搜索空间；

架构构建器，具体用于根据至少一组采样参数从结构搜索空间搜索网络模型。

5 在一种可能的实施方式中，根据候选集进行采样的采样方式，包括：随机采样；或者，按照预设规则采样。

在一种可能的实施方式中，若采样方式为按照预设规则采样，则采样模块，还用于在确定网络模型不符合预设条件之后，根据网络模型的评估结果采用预设的优化算法更新预设规则。

10 在一种可能的实施方式中，优化算法包括：进化算法、强化学习算法、贝叶斯优化算法或梯度优化算法。

在一种可能的实施方式中，预设条件包括以下一项或者多项：获取网络模型的次数超过预设次数、获取网络模型的时长超过预设时长或者网络模型的输出结果符合预设要求。

15 在一种可能的实施方式中，候选集中包括以下一项或者多项：算子的种类、算子的属性信息或者算子之间的连接方式。

在一种可能的实施方式中，目标神经网络用于进行图片识别、语义分割或者目标检测中的至少一种。

在一种可能的实施方式中，神经网络构建系统还包括：

20 训练模块，用于在将网络模型作为目标神经网络之后，通过预设的数据集对目标神经网络进行训练，得到训练后的目标神经网络。

在一种可能的实施方式中，输入模块，具体用于：接收用户输入数据；从用户输入数据中获取初始骨干网络和候选集。

25 第三方面，本申请实施例提供一种神经网络构建装置，该神经网络构建装置具有实现上述第一方面神经网络构建方法的功能。该功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。该硬件或软件包括一个或多个与上述功能相对应的模块。

第四方面，本申请实施例提供一种神经网络构建装置，包括：处理器和存储器，其中，处理器和存储器通过线路互联，处理器调用存储器中的程序代码用于执行上述第一方面任一项所示的神经网络构建方法中与处理相关的功能。可选地，该神经网络构建装置可以是芯片。

30 第五方面，本申请实施例提供了一种神经网络构建装置，该神经网络构建装置也可以称为数字处理芯片或者芯片，芯片包括处理单元和通信接口，处理单元通过通信接口获取程序指令，程序指令被处理单元执行，处理单元用于执行如上述第一方面或第一方面任一可选实施方式中与处理相关的功能。

35 第六方面，本申请实施例提供了一种计算机可读存储介质，包括指令，当其在计算机上运行时，使得计算机执行上述第一方面或第一方面任一可选实施方式中的方法。

第七方面，本申请实施例提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述第一方面或第一方面任一可选实施方式中的方法。

附图说明

- 图 1 本申请应用的一种人工智能主体框架示意图；
图 2 本申请提供的一种系统架构示意图；
图 3 为本申请实施例提供的一种卷积神经网络结构示意图；
5 图 4 为本申请实施例提供的另一种卷积神经网络结构示意图；
图 5 本申请提供的另一种系统架构示意图；
图 6 为本申请实施例提供的一种神经网络构建方法的流程示意图；
图 7 为本申请实施例提供的一种神经网络变形方式示意图；
图 8 为本申请实施例提供的一种占位模块的替换结构示意图；
10 图 9 为本申请实施例提供的一种神经网络构建系统的结构示意图；
图 10 为本申请实施例提供的另一种神经网络构建系统的结构示意图；
图 11 为本申请实施例提供的另一种神经网络构建系统的结构示意图；
图 12 为本申请实施例提供的另一种占位模块的替换结构示意图；
图 13 为本申请实施例提供的另一种占位模块的替换结构示意图；
15 图 14 为本申请实施例提供的另一种占位模块的替换结构示意图；
图 15 为本申请实施例提供的另一种神经网络构建装置的结构示意图；
图 16 为本申请实施例提供的一种芯片的结构示意图。

具体实施方式

- 20 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描述，显然，所描述的实施例仅仅是本申请一部分实施例，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。

- 本申请提供的神经网络构建方法可以应用于人工智能 (artificial intelligence, AI) 场景中。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分支，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作出反应的智能机器。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。人工智能领域的研究包括机器人，自然语言处理，计算机视觉，决策与推理，人机交互，推荐与搜索，AI 基础理论等。

- 30 图 1 示出一种人工智能主体框架示意图，该主体框架描述了人工智能系统总体工作流程，适用于通用的人工智能领域需求。

- 下面从“智能信息链”（水平轴）和“IT 价值链”（垂直轴）两个维度对上述人工智能主题框架进行阐述。

- 35 “智能信息链”反映从数据的获取到处理的一系列过程。举例来说，可以是智能信息感知、智能信息表示与形成、智能推理、智能决策、智能执行与输出的一般过程。在这个过程中，数据经历了“数据—信息—知识—智慧”的凝练过程。

“IT 价值链”从人智能的底层基础设施、信息（提供和处理技术实现）到系统的产业生态过程，反映人工智能为信息技术产业带来的价值。

（1）基础设施：

基础设施为人工智能系统提供计算能力支持，实现与外部世界的沟通，并通过基础平台实现支撑。通过传感器与外部沟通；计算能力由智能芯片，如中央处理器（central processing unit, CPU）、网络处理器（neural-network processing unit, NPU）、图形处理器（英语：graphics processing unit, GPU）、专用集成电路（application specific integrated circuit, ASIC）或现场可编程逻辑门阵列（field programmable gate array, FPGA）等硬件加速芯片）提供；基础平台包括分布式计算框架及网络等相关的平台保障和支持，可以包括云存储和计算、互联互通网络等。举例来说，传感器和外部沟通获取数据，这些数据提供给基础平台提供的分布式计算系统中的智能芯片进行计算。

（2）数据

基础设施的上一层的数据用于表示人工智能领域的数据来源。数据涉及到图形、图像、语音、视频、文本，还涉及到传统设备的物联网数据，包括已有系统的业务数据以及力、位移、液位、温度、湿度等感知数据。

（3）数据处理

数据处理通常包括数据训练，机器学习，深度学习，搜索，推理，决策等方式。

其中，机器学习和深度学习可以对数据进行符号化和形式化的智能信息建模、抽取、预处理、训练等。

推理是指在计算机或智能系统中，模拟人类的智能推理方式，依据推理控制策略，利用形式化的信息进行机器思维和求解问题的过程，典型的功能是搜索与匹配。

决策是指智能信息经过推理后进行决策的过程，通常提供分类、排序、预测等功能。

（4）通用能力

对数据经过上面提到的数据处理后，进一步基于数据处理的结果可以形成一些通用的能力，比如可以是算法或者一个通用系统，例如，翻译，文本的分析，计算机视觉的处理（如图像识别、目标检测等），语音识别等等。

（5）智能产品及行业应用

智能产品及行业应用指人工智能系统在各领域的产品和应用，是对人工智能整体解决方案的封装，将智能信息决策产品化、实现落地应用，其应用领域主要包括：智能制造、智能交通、智能家居、智能医疗、智能安防、自动驾驶，平安城市，智能终端等。

参见附图 2，本申请实施例提供了一种系统架构 200。该系统架构中包括数据库 230、客户设备 240。数据采集设备 260 用于采集数据并存入数据库 230，构建模块 202 基于数据库 230 中维护的数据生成目标模型/规则 201。下面将更详细地描述构建模块 202 如何基于数据得到目标模型/规则 201，目标模型/规则 201 即本申请以下实施方式中构建得到的神经网络，具体参阅以下图 6-13 中的相关描述。

计算模块可以包括构建模块 202，构建模块 202 得到的目标模型/规则可以应用不同的系统或设备中。在附图 2 中，执行设备 210 配置收发器 212，该收发器 212 可以是无线收

发器、光收发器或有线接口（如 I/O 接口）等，与外部设备进行数据交互，“用户”可以通过客户设备 240 向收发器 212 输入数据，例如，本申请以下实施方式，客户设备 240 可以向执行设备 210 发送初始骨干网络，请求执行设备基于初始骨干网络构建目标神经网络。可选地，客户设备 240 还可以向执行设备 210 发送用于构建目标神经网络的数据库，即本

5 申请以下提及的候选集，此处不再赘述。

执行设备 210 可以调用数据存储系统 250 中的数据、代码等，也可以将数据、指令等存入数据存储系统 250 中。

10 计算模块 211 对输入的数据进行处理。具体地，计算模块 211 用于：将初始骨干网络中的至少一个基本单元替换为至少一个占位模块，得到待定网络；然后根据候选集得到至少一个采样结构的信息；然后根据待定网络和至少一个采样结构的信息获取网络模型，若网络模型不符合预设条件，则可以重新根据候选集进行采样，并根据重新采样得到的至少一个采样结构更新网络模型；若网络模型符合预设条件，则将该网络模型作为目标神经网络，即图 2 中所示的目标模型/规则 201。

15 关联功能模块 213 和关联功能模块 214 为可选模块，可以用于搜索与目标神经网络关联的除主干网络之外的其他网络，如区域生成网络(region proposal network, RPN)、特征金字塔网络(feature Pyramid network, FPN)等。

最后，收发器 212 将构建得到的目标神经网络返回给客户设备 240，以在客户设备 240 或者其他设备中部署该目标神经网络。

20 更深层地，构建模块 202 可以针对不同的目标任务，基于不同的候选集得到相应的目标模型/规则 201，以给用户提供更佳的结果。

在附图 2 中所示情况下，可以根据用户的输入数据确定输入执行设备 210 中的数据，例如，用户可以在收发器 212 提供的界面中操作。另一种情况下，客户设备 240 可以自动地向收发器 212 输入数据并获得结果，若客户设备 240 自动输入数据需要获得用户的授权，用户可以在客户设备 240 中设置相应权限。用户可以在客户设备 240 查看执行设备 210 输出的结果，具体的呈现形式可以是显示、声音、动作等具体方式。客户设备 240 也可以作为数据采集端将采集到的数据存入数据库 230。

30 需要说明的是，附图 2 仅是本申请实施例提供的一种系统架构的示例性的示意图，图中所示设备、器件、模块等之间的位置关系不构成任何限制。例如，在附图 2 中，数据存储系统 250 相对执行设备 210 是外部存储器，在其它场景中，也可以将数据存储系统 250 置于执行设备 210 中。

本申请提及的神经网络可以包括多种类型，如深度神经网络(deep neural network, DNN)、卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural networks, RNN)或残差网络其他神经网络等。

示例性地，下面以 CNN 为例。

35 CNN 是一种带有卷积结构的深度神经网络。CNN 是一种深度学习(deep learning)架构，深度学习架构是指通过机器学习的算法，在不同的抽象层级上进行多个层次的学习。作为一种深度学习架构，CNN 是一种前馈(feed-forward)人工神经网络，该前馈人工神

神经网络中的各个神经元对输入其中的图像中的重叠区域作出响应。卷积神经网络包含了一个由卷积层和子采样层构成的特征抽取器。该特征抽取器可以看作是滤波器，卷积过程可以看作是使用一个可训练的滤波器与一个输入的图像或者卷积特征平面(feature map)做卷积。卷积层是指卷积神经网络中对输入信号进行卷积处理的神经元层。在卷积神经网络的卷积层中，一个神经元可以只与部分邻层神经元连接。一个卷积层中，通常包含若干个特征平面，每个特征平面可以由一些矩形排列的神经单元组成。同一特征平面的神经单元共享权重，这里共享的权重就是卷积核。共享权重可以理解为提取图像信息的方式与位置无关。这其中隐含的原理是：图像的某一部分的统计信息与其他部分是一样的。即意味着在某一部分学习的图像信息也能用在另一部分上。所以对于图像上的所有位置，我们都能使用同样的学习得到的图像信息。在同一卷积层中，可以使用多个卷积核来提取不同的图像信息，一般地，卷积核数量越多，卷积操作反映的图像信息越丰富。一个或者多个卷积核可以组成一个基本单元。

卷积核可以以随机大小的矩阵的形式初始化，在卷积神经网络的训练过程中卷积核可以通过学习得到合理的权重。另外，共享权重带来的直接好处是减少卷积神经网络各层之间的连接，同时又降低了过拟合的风险。

卷积神经网络可以采用误差反向传播(back propagation, BP)算法在训练过程中修正初始的超分辨率模型中参数的大小，使得超分辨率模型的重建误差损失越来越小。具体地，前向传递输入信号直至输出会产生误差损失，通过反向传播误差损失信息来更新初始的超分辨率模型中参数，从而使误差损失收敛。反向传播算法是以误差损失为主导的反向传播运动，旨在得到最优的超分辨率模型的参数，例如权重矩阵。

如图3所示，卷积神经网络(CNN) 100可以包括输入层110，卷积层/池化层120，其中池化层为可选的，以及神经网络层130。

如图3所示卷积层/池化层120可以包括如示例121-126层，在一种实现中，121层为卷积层，122层为池化层，123层为卷积层，124层为池化层，125为卷积层，126为池化层；在另一种实现方式中，121、122为卷积层，123为池化层，124、125为卷积层，126为池化层。即卷积层的输出可以作为随后的池化层的输入，也可以作为另一个卷积层的输入以继续进行卷积操作。

以卷积层121为例，卷积层121可以包括很多个卷积算子，卷积算子也称为核，其在图像处理中的作用相当于一个从输入图像矩阵中提取特定信息的过滤器，卷积算子本质上可以是一个权重矩阵，这个权重矩阵通常被预先定义。在对图像进行卷积操作的过程中，权重矩阵通常在输入图像上沿着水平方向一个像素接着一个像素（或两个像素接着两个像素……这取决于步长stride的取值）的进行处理，从而完成从图像中提取特定特征的工作。该权重矩阵的大小应该与图像的大小相关。需要注意的是，权重矩阵的纵深维度(depth dimension)和输入图像的纵深维度是相同的，在进行卷积运算的过程中，权重矩阵会延伸到输入图像的整个深度。因此，和一个单一的权重矩阵进行卷积会产生一个单一纵深维度的卷积化输出，但是大多数情况下不使用单一权重矩阵，而是应用维度相同的多个权重矩阵。每个权重矩阵的输出被堆叠起来形成卷积图像的纵深维度。不同的权重矩阵可以用来

提取图像中不同的特征，例如一个权重矩阵用来提取图像边缘信息，另一个权重矩阵用来提取图像的特定颜色，又一个权重矩阵用来对图像中不需要的噪点进行模糊化等。该多个权重矩阵维度相同，经过该多个维度相同的权重矩阵提取后的特征图维度也相同，再将提取到的多个维度相同的特征图合并形成卷积运算的输出。

5 通常，权重矩阵中的权重值在实际应用中需要经过大量的训练得到，通过训练得到的权重值形成的各个权重矩阵可以从输入图像中提取信息，从而帮助卷积神经网络 100 进行正确的预测。

当卷积神经网络 100 有多个卷积层时，初始的卷积层（例如 121）往往提取较多的一般特征，该一般特征也可以称之为低级别的特征；随着卷积神经网络 100 深度的加深，越往后的卷积层（例如 126）提取到的特征越来越复杂，比如高级别的语义之类的特征，语义越高的特征越适用于待解决的问题。

池化层：

由于常常需要减少训练参数的数量，因此卷积层之后常常需要周期性的引入池化层，即如图 3 中 120 所示例的 121-126 各层，可以是一层卷积层后面跟一层池化层，也可以是多层卷积层后面接一层或多层池化层。在图像处理过程中，池化层的唯一目的就是减少图像的空间大小。池化层可以包括平均池化算子和/或最大池化算子，以用于对输入图像进行采样得到较小尺寸的图像。平均池化算子可以在特定范围内对图像中的像素值进行计算产生平均值。最大池化算子可以在特定范围内取该范围内值最大的像素作为最大池化的结果。另外，就像卷积层中用权重矩阵的大小应该与图像大小相关一样，池化层中的运算符也应该与图像的大小相关。通过池化层处理后输出的图像尺寸可以小于输入池化层的图像的尺寸，池化层输出的图像中每个像素点表示输入池化层的图像的对应子区域的平均值或最大值。

神经网络层 130：

在经过卷积层/池化层 120 的处理后，卷积神经网络 100 还不足以输出所需要的输出信息。因为如前所述，卷积层/池化层 120 只会提取特征，并减少输入图像带来的参数。然而为了生成最终的输出信息（所需要的类信息或别的相关信息），卷积神经网络 100 需要利用神经网络层 130 来生成一个或者一组所需要的类的数量的输出。因此，在神经网络层 130 中可以包括多层隐含层（如图 3 所示的 131、132 至 13n）以及输出层 140。在本申请中，该卷积神经网络为：对选取的起点网络进行至少一次变形得到串行网络，然后根据训练后的串行网络得到。该卷积神经网络可以用于图像识别，图像分类，图像超分辨率重建等等。

在神经网络层 130 中的多层隐含层之后，也就是整个卷积神经网络 100 的最后层为输出层 140，该输出层 140 具有类似分类交叉熵的损失函数，具体用于计算预测误差，一旦整个卷积神经网络 100 的前向传播（如图 3 由 110 至 140 的传播为前向传播）完成，反向传播（如图 3 由 140 至 110 的传播为反向传播）就会开始更新前面提到的各层的权重值以及偏差，以减少卷积神经网络 100 的损失及卷积神经网络 100 通过输出层输出的结果和理想结果之间的误差。

需要说明的是，如图 3 所示的卷积神经网络 100 仅作为一种卷积神经网络的示例，在

具体的应用中，卷积神经网络还可以以其他网络模型的形式存在，例如，如图 4 所示的多个卷积层/池化层并行，将分别提取的特征均输入给全神经网络层 130 进行处理。

参见附图 5，本申请实施例还提供了一种系统架构 300。执行设备 210 由一个或多个服务器实现，可选的，与其它计算设备配合，例如：数据存储、路由器、负载均衡器等设备；
5 执行设备 210 可以布置在一个物理站点上，或者分布在多个物理站点上。执行设备 210 可以使用数据存储系统 250 中的数据，或者调用数据存储系统 250 中的程序代码实现本申请以下图 6-13 对应的神经网络构建方法的步骤。

用户可以操作各自的本地设备（例如本地设备 301 和本地设备 302）与执行设备 210 进行交互。每个本地设备可以表示任何计算设备，例如个人计算机、计算机工作站、智能手机、平板电脑、智能摄像头、智能汽车或其他类型蜂窝电话、媒体消费设备、可穿戴设备、机顶盒、游戏机等。

每个用户的本地设备可以通过任何通信机制/通信标准的通信网络与执行设备 210 进行交互，通信网络可以是广域网、局域网、点对点连接等方式，或它们的任意组合。具体地，该通信网络可以包括无线网络、有线网络或者无线网络与有线网络的组合等。该无线网络包括但不限于：第五代移动通信技术（5th-Generation, 5G）系统，长期演进（long term evolution, LTE）系统、全球移动通信系统（global system for mobile communication, GSM）或码分多址（code division multiple access, CDMA）网络、宽带码分多址（wideband code division multiple access, WCDMA）网络、无线保真（wireless fidelity, WiFi）、蓝牙（bluetooth）、紫蜂协议（Zigbee）、射频识别技术（radio frequency identification, RFID）、远程（Long Range, Lora）无线通信、近距离无线通信（near field communication, NFC）中的任意一种或多种的组合。该有线网络可以包括光纤通信网络或同轴电缆组成的网络等。

在另一种实现中，执行设备 210 的一个方面或多个方面可以由每个本地设备实现，例如，本地设备 301 可以为执行设备 210 提供本地数据或反馈计算结果。

25 需要注意的，执行设备 210 的所有功能也可以由本地设备实现。例如，本地设备 301 实现执行设备 210 的功能并为用户提供服务，或者为本地设备 302 的用户提供服务。

基于前述图 1-5 提供的系统架构或者神经网络，下面对本申请提供的神经网络构建方法进行详细介绍。

首先，为便于理解，对本申请涉及的一些名词进行解释。

30 基本单元（block）：通常基本单元由卷积层组成，或者，一个基本单元可以理解为一个卷积模组。

神经网络算子（Neural Network Operator）：定义了对输入数据进行某种计算得到输出的方式，通常可以作为神经网络的基本单元。算子的属性通常有类型、宽度或深度等。计算机视觉任务网络常用的算子的类型可以包括卷积（Convolution）、池化（Pooling）、
35 激活函数（Activation）等。多个算子之间相互连接形成的有向计算图即构成神经网络。

神经网络架构（Neural Architecture）：包括神经网络的内部各算子的属性定义和算子间的连接方式等。神经网络架构通常由重复子结构组成，如重复单元（Cell），残差块

(Residual Block) 等。一个网络架构对应一个从输入数据到输出数据的完整计算图，如从图像到图像类别，图像到物体目标，从文本到语义信息等。

骨干网络架构 (Backbone Architecture): 或者也可以称为骨干网络或者初始骨干网络，指用户提供的原始网络架构，作为架构搜索方案要提升性能的对象，通常为经典的堆叠得到的网络或者人工设计架构及其变种。某些任务中划分的多个子网络架构统称一个骨架网络。如图像分类任务中的分类器网络为一个骨架网络，而目标检测模型中的特征提取网络和检测网络等合在一起称为一个骨架网络架构。通常，在神经网络中，除了骨干网络之外，还可以包括其他的功能性网络，如区域生成网络 (region proposal network, RPN)、特征金字塔网络 (feature Pyramid network, FPN) 等网络，用于对骨干网络提取到的特征进行进一步处理，如识别特征的分类、对特征进行语义分割等。

参阅图 6，本申请提供的一种神经网络构建方法的流程示意图，如下所述。

601、获取初始骨干网络和候选集。

其中，该初始骨干网络用于构建目标神经网络，该初始骨干网络可以是人工设计的网络或者其变种，或者也可以是分类、分割或者检测任务中常用的骨干网络等。

具体地，该初始骨干网络可以通过用户输入数据得到的网络。如前述图 2 中，本申请提供的神经网络构建方法可以由执行设备 210 来执行，客户设备 240 可以将初始骨干网络发送至执行设备 210，然后由执行设备执行以下步骤 602-606。

可选地，候选集可以是接收到用户输入数据之后，从用户输入数据中获取到的，或者，候选集也可以是从本地数据中获取到的。例如，在一些常用的场景中，如分类、分割或者检测任务所需的骨干网络的架构类似，则在分类、分割或者检测任务等场景中，可以使用相同的候选集。

具体地，候选集中可以包括多种结构的参数，具体可以包括算子 (operator) 的结构、算子的属性、算子之间的连接方式等。例如，该候选集中可以包括卷积算子 (也可以称为卷积核或者卷积模组)、梯度算子或者微分算子等算子的结构、算子的宽度、算子之间的连接方式等。

602、将初始骨干网络中的至少一个基本单元替换为至少一个占位模块，得到待定网络。

在获取到初始骨干网络之后，将初始骨干网络中的至少一个基本单元替换为至少一个占位模块，得到待定网络。

其中，该占位模块的结构可以是初始结构或者为空。初始骨干网络中被替换为占位模块的基本单元的位置可以是预先设定的，也可以是随机选择的，还可以是对初始骨干网络进行评估后确定的。例如，可以预先设定目标神经网络在不同的场景中可被替换的基本单元的位置，在构建执行目标任务的目标神经网络时，将根据目标任务确定可被替换的基本单元的位置，然后根据该位置将用户提供的初始骨干网络中相应位置的基本单元替换为占位模块。又例如，可以对初始骨干网络进行评估，确定初始骨干网络输出结果的准确度，然后根据输出结果的准确性来确定被替换的基本单元的数量，如输出结果的准确度越低，被替换的基本单元的数量也就越多，然后根据该数量随机或者按照预设规则来选择初始骨干网络中被替换的基本单元的位置。

示例性地，如图 7 所示，首先接收用户发送的初始骨干网络，其中包括四个模块 (module)，即 module 1、module 2、module 3 和 module 4。当然，此处的四个模块仅仅是示例性说明，也可以是更多或者更少的模块。然后可以将其中的 module 1、module 2 和 module 3 替换为占位模块 (slot) 1、slot 2 和 slot 3，得到待定网络。

5 603、根据候选集进行采样，得到至少一个采样结构的信息。

在得到待定网络之后，根据候选集进行采样，得到至少一个采样结构的信息。该至少一个结果为步骤 602 中至少一个占位模块的结构。

其中，得到至少一个采样结构的信息的方式有多种，下面示例性地对几种常用的方式进行介绍。

10 方式一、直接从候选集中采样

其中，可以直接从候选集中进行采样，得到至少一个采样结构的参数，如算子结构、算子的属性或者算子之间的连接方式等。

例如，若候选集中包括 10 个算子的结构、5 种连接方式和宽度的取值范围（如 1-6），则可以从候选集中采样 2 个算子结构、3 种连接方式和宽度范围（如 1、5、6），得到至少一个采样结构的信息。

15 因此，本申请实施方式中，可以直接从候选集中进行采样，流程简单，采集效率高。

方式二、在参数空间中采集至少一组采样参数

在本方式中，可选地，在步骤 603 之前，即根据所述候选集进行采样，得到至少一个采样结构的信息之前，本申请提供的方法还可以包括：根据候选集构建参数空间，参数空间中包括与候选集中的多种结构的参数对应的架构参数。该参数空间可以理解为定义了各类架构参数的集合，为后续的结构采样提供了接口 (Interface)。通常，在一些场景中，参数空间记录所有已定义的架构参数，并支持通过统一的接口进行查找，遍历，赋值，导出等操作。此外，参数空间提供了自定义的参数回调接口，可在参数更新的时候调用用户自定义的功能。

25 其中，架构参数定义了可能的占位模块的结构中的参数的名称、种类以及取值范围，一方面，可以通过访问架构参数所定义的参数的信息来确定搜索参数的范围，从而在改范围内搜索占位模块的可能的结构的参数。另一方面，可以通过构建占位模块可能的结构和对应的参数，实现了结构到参数的映射。例如，参数结构的种类和取值范围可以包括：分类类型 (Categorical)：取值范围可以包括候选集中的多种类型，参数空间中所包括的分类类型可以通过预设的方式进行定义，如将分类类型定义为甲、乙、丙等；张量类型 (Tensor)：取值范围为给定形状以及数据类型的张量，如 n 维可求导数组；数值类型 (Numerical)：取值为单一数值，数据类型分为整型 (Integer) 或浮点型 (Real) 等。可以理解为，架构参数定义了候选集中的结构参数的索引或者取值范围，从而降低后续的采样数据量。

30 相应地，步骤 603 可以包括：对参数空间进行采样，得到与至少一个采样结构对应的至少一组采样参数。即步骤 603 中的至少一个采样结构的信息可以包括从参数空间采集到的至少一组采样参数。通常，一组采样参数可以对应一个采样结构，例如，一组采样参数可以包括分类类型、张量类型、数值类型等取值。当然，也可以是一个采样结构对应多组

采样参数，具体可以根据实际应用场景进行调整。可以理解为，参数空间中可以包括候选集中所都包括的算子类型的索引、算子之间的连接方式的索引或者参数的取值范围等。因此，可以从参数控制直接采集到算子类型的索引、算子之间的连接方式的索引或者参数的取值范围等，而无需从候选集中直接采样，提高了采样效率。

5 可选地，在本申请实施方式中，对候选集、参数空间或者结构搜索进行采样的方式可以是随机采样也可以是按照预设规则进行采样。该预设规则可以是预先设置的概率分布或者根据优化算法计算出的概率分布或者根据优化算法计算出的采样方式等。

可选地，更新预设规则的算法可以包括但不限于：进化算法、强化学习算法、贝叶斯算法或者梯度优化算法等。因此，本申请实施方式中，可以通过优化算法更新采样方式，
10 从而使后续采集到的参数对应的结构代入待定网络得到的评估结果更优，提高了得到最终的目标神经网络的效率。

604、根据待定网络和至少一个采样结构的信息获取网络模型。

其中，在获取到至少一个采样结构的信息之后，待定网络中的至少一个占位模块的结构可以是该至少一个采样结构的结构，从而得到网络模型。

15 具体地，根据待定网络和至少一个采样结构的信息获取网络模型的方式可以有多种下面分别进行说明。

方式一、直接从候选集中采样，并根据采样得到的结构的参数构建网络模型

其中，可以直接从候选集中搜索得到至少一个采样结构或者至少一个采样结构的属性、连接方式等信息，然后可以根据采集到的至少一个采样结构或者至少一个采样结构的属性、
20 连接方式等信息确定待定网络中的至少一个占位模块的结构，从而构建得到完整的网络模型。例如，若从候选集中采集到卷子算子 A，则可以将占位模块替换为卷积算子 A，或者说将占位模块的结构转换为卷积算子 A 的结构。

因此，在本方式中，可以直接从候选集中进行采样，可以直接采集到可以替换占位模块的结构，提高了采集效率。

25 方式二、根据采样参数从候选集中采集，并根据采样得到的结构的参数构建网络模型

其中，若根据候选集构建了参数空间，则在参数空间采集到至少一组采样参数之后，根据该至少一组采样参数从候选集中采集相应的结构的参数，并根据采集到的结构的参数确定待定网络中的至少一个占位模块的结构，从而得到网络模型。例如，若某一组采样参数中包括了卷积算子 B 的索引，则可以根据该索引，从候选集中采集到卷积算子 B 的结构，
30 然后将待定网络中的其中一个占位模块的结构替换为卷积算子 B 的结构，得到网络模型。

具体地，待定网络中的至少一个占位模块的结构可以是空白的，或者也可以是初始化的结构；在采集到至少一组采样参数之后，可以对待定网络中的至少一个占位模块的结构进行转换，得到网络模型。例如，可以将占位模块的结构转换为对应的采样参数的结构，或者，根据对应的采样参数改变占位模块的结构，如改变占位模块的通道数量、深度或者其他结构参数等。又例如，若占位模块由多个算子组成，采样参数中可以包括采集到的算子之间的连接方式，可以根据该采样参数调整或者改变占位模块包括的算子之间的连接方式，如串联、并联或者串并联结合等连接方式。例如，多个算子也可以组成多分支的结构，
35

如树形、有向无环图等结构，即存在多个分支，该多个分支的数据分配方式可以是将输入复制给每一个分支，然后对每一个分支的输出求和，或者也可以是将输入分割给每一个分支，然后对每一个分支的输出进行合并然后输出等。

因此，本申请实施方式中，可以采集参数空间中所包括的参数，然后根据从参数空间中采样到的参数从候选集中采集对应的结构的参数。因参数空间中包括的参数可以理解为候选集中所包括的参数的索引，因此，本方式中从参数空间中采样，采集到的数据也是候选集中所包括的参数的索引，采集效率高。

方式三、根据采样参数从结构搜索空间中采集网络模型

在本方式中，在步骤 605 之前，还可以根据候选集与待定网络构建结构搜索空间，该结构搜索空间中包括多种结构，该多种结构中包括候选集中所包括的结构或者参数与待定网络的所有可能的组合方式。例如，若候选集中包括卷积算子 A 和卷积算子 B，则可以将卷积算子 A 和卷积算子 B 与待定网络进行组合，得到卷积算子 A 和卷积算子 B 与待定网络的所有可能的组合结构，组成结构搜索空间。

若根据候选集构建了参数空间以及结构搜索空间，则在采集到至少一组采样参数之后，可以根据该至少一组采样参数，从结构搜索空间中所述对应的采样结构所在的结构，从而可以直接采集到网络模型。例如，若采集到的参数中包括卷积算子 C 的索引，则可以根据该索引，从结构搜索空间中所述该索引与待定网络组成的结构，从而直接得到网络模型。减少了在获得网络模型中的构建过程，提高了得到网络模型的效率。

示例性地，占位模块 slot1 的可能的结构可以如图 8 所示，其中，获取到用户输入的初始骨干网络，其中包括 module1 和 module2。可以将初始骨干网络中的 module1 替换为占位模块 slot1。slot1 可能的结构可以包括如图 8 中所示的算子 X、弹性结构、module1（即将 slot1 的结构转换为初始骨干网络中的 module1 的结构）、混合算子或者多个 slot 组成的结构等。下面分别对 slot1 的可能的结构分别进行说明，其中，混合算子可以包括候选集中的多个算子的结构组成的结构，即 slot1 的结构可以是多个算子组合起来的结构；多个 slot 即将 slot1 替换为多个可替换的 slot，如图 8 中所示的将 slot1 替换为 slot11、slot22 等占位模块；弹性结构即 slot1 的结构可以是可变的结构，可以在训练的过程中确定该 slot 的具体结构，如 slot1 的宽度、深度等可变；算子 X 可以是包括于候选集中的其中一个算子或者不包括于候选集的一个算子等。

可以理解为，方式三中，在步骤 604 之前，即可对网络模型进行了构建，构建方式可以参阅前述方式二中构建网络模型的方式，从而使得在进行采样时，可以直接采集到网络模型，提高了采样效率。

因此，本申请实施方式中，可以通过改变占位模块的结构来确定待定网络的结构，即使应用在不同的场景中，也仅需要改变选择的占位模块的结构即可得到不同的网络模型，泛化能力强。

605、判断网络模型不符合预设条件，若是，则执行步骤 606；可选地，若否，则执行步骤 603。

其中，在得到网络模型之后，可以对网络模型进行评估，然后判断网络模型的评估结

果是否符合预设条件。若网络模型的评估结果符合预设条件，则可以将该网络模型作为目标神经网络，即执行步骤 606；若网络模型的评估结果不符合预设条件，在可以重新根据候选集进行采样，并根据重新采样得到的至少一个采样结构的参数构建的结构作为新的网络模型。

5 在一种可能的实施方式中，若前述的采样方式为根据预设规则进行采样，且确定网络模型的评估结果不符合预设条件，则可以通过网络模型的评估结果以及优化算法来更新预设规则，该优化算法可以包括但不限于进化算法、强化学习算法、贝叶斯优化算法或梯度优化算法等。

10 在一种可能的实施方式中，该预设条件具体可以包括但不限于以下一项或者多项：网络模型的输出的精度大于第一阈值，或者，网络模型的输出的平均精度大于第二阈值，或者，损失值不大于第三阈值，或者，推理时长不大于第四阈值，每秒浮点运算次数（FLOPS）不大于第五阈值等，平均精度为对神经网络进行多次评估得到的多个精度的平均值，推理时长为从神经网络根据输入得到输出结果的时长。

606、将网络模型作为目标神经网络。

15 其中，若网络模型的评估结果符合预设条件，则可以将该网络模型作为目标神经网络。

在一种可能的实施方式中，在将网络模型作为目标神经网络之后，还可以通过数据集对目标神经网络进行训练，从而得到训练后的目标神经网络，且通常训练后的目标神经网络的输出更准确。

20 在一种可能的实施方式中，在将网络模型作为目标神经网络之后，还可以根据目标神经网络需要执行的任务为目标神经网络添加其他与任务关联的模块，如区域生成网络（region proposal network, RPN）、特征金字塔网络（feature Pyramid network, FPN）等。

25 因此，本申请实施方式中，通过在骨干网络中设置占位模块，改变占位模块的结构来改变骨干网络的结构。可以针对不同的场景来改变占位模块的结构或者位置等，可以适应不同的场景，泛化能力强。并且，即使在面向新的应用场景，也无需进行大量的迁移或者重构，降低了代码的调试时间，提高用户体验。

此外，本申请实施方式中，无需用户手动绑定候选集合占位模块之间的关系，也无需用户提供占位模块的结构转换方式，用户仅需提供初始骨干网络，或者可选地还提供候选集，降低了用户学习门槛，对于用户更友好。

30 前述对本申请提供的神经网络构建方法进行了详细说明，下面基于前述的神经网络构建方法，对本申请提供的神经网络构建系统进行详细介绍，该神经网络构建系统用于执行前述图 6 所示的神经网络构建方法。

参阅图 9，本申请提供的一种神经网络构建系统示意图，如下所述。

35 该神经网络构建系统可以包括但不限于：输入模块 901、采样模块 902、架构构建器 903 和架构评估器 904。

输入模块 901，用于获取初始骨干网络和候选集，初始骨干网络用于构建目标神经网络；

架构构建器 903, 用于将初始骨干网络中的至少一个基本单元替换为至少一个占位模块, 得到待定网络, 候选集中包括与至少一个占位模块对应的多种结构的参数;

采样模块 902, 用于根据候选集进行采样, 得到至少一个采样结构的信息;

架构构建器 903, 还用于根据占位骨干网络待定网络和至少一个采样结构的信息获取
5 网络模型, 其中, 至少一个采样结构的信息用于确定至少一个占位模块的结构;

架构评估器 904, 用于评估网络模型是否符合预设条件, 若网络模型符合预设条件, 则将网络模型作为目标神经网络。

在一种可能的实施方式中, 若网络模型不符合预设条件, 则由采样模块 902 重新根据候选集进行采样, 架构构建器 903 根据重新采样得到的至少一个采样结构的信息更新网络
10 模型。

在一种可能的实施方式中, 架构构建器 903, 还用于根据候选集构建参数空间, 参数空间中包括与多种结构的参数对应的架构参数;

采样模块 902, 具体用于对参数空间进行采样, 得到与至少一个采样结构对应的至少一组采样参数。

在一种可能的实施方式中, 架构构建器 903, 具体用于根据至少一个采样结构对待定网络中的至少一个占位模块的结构进行转换, 得到网络模型。

在一种可能的实施方式中, 架构构建器 903, 还用于在根据待定网络和至少一个采样结构的信息获取网络模型之前, 根据候选集和待定网络构建多种结构, 多种结构组成结构搜索空间;

20 架构构建器 903, 具体用于根据至少一组采样参数从结构搜索空间搜索网络模型。

在一种可能的实施方式中, 根据候选集进行采样的采样方式, 包括: 随机采样; 或者, 按照预设规则采样。

在一种可能的实施方式中, 若采样方式为按照预设规则采样, 则采样模块, 还用于在确定网络模型不符合预设条件之后, 根据网络模型的评估结果采用预设的优化算法更新预设规则。
25

在一种可能的实施方式中, 优化算法包括: 进化算法、强化学习算法、贝叶斯优化算法或梯度优化算法。

在一种可能的实施方式中, 预设条件包括以下一项或者多项: 获取网络模型的次数超过预设次数、获取网络模型的时长超过预设时长或者网络模型的输出结果符合预设要求。

30 在一种可能的实施方式中, 候选集中包括但不限于以下一项或者多项: 算子的种类、算子的属性信息或者算子之间的连接方式。

在一种可能的实施方式中, 目标神经网络用于进行图片识别、语义分割或者目标检测中的至少一种。

在一种可能的实施方式中, 神经网络构建系统还包括: 训练模块 905, 用于在将网络模型作为目标神经网络之后, 通过预设的数据集对目标神经网络进行训练, 得到训练后的
35 目标神经网络。

在一种可能的实施方式中, 输入模块 901, 具体用于: 接收用户输入数据; 从用户输

入中获取初始骨干网络和候选集。

为便于理解，下面示例性地，对本申请提供的神经网络构建系统进行更详细地介绍。

如图 10 所示，下面首先对图 10 中的一些模块进行说明。

初始骨干网络 1001：可以是由用户提供的一些常用的任务中的神经网络的骨干网络，
5 或者根据不同的任务生成的骨干网络。

候选集 1003：也可以称为 candidate set，包括了多种结构的参数，如多种卷积算子、卷积算子的宽度、卷积算子之间的连接方式等。通常，每个 slot 可以与候选集 1003 中的全部或者部分参数具有绑定关系，用于后续确定 slot 的结构。例如，可以为一个 slot 绑定候选集 1003 中的其中一个子集，每个 slot 绑定的子集可以相同也可以不相同。

10 骨干网络（backbone）1002：即将初始骨干网络中的至少一个基本单元替换为占位模块，从而得到骨干网络 1002。在构建骨干网络 1002 的过程中，占位模块与候选集进行绑定，得到新的完整架构。该模块实现了骨架与候选集合的解耦和复用。在构建的过程中，可以根据不同情况进行重新绑定，即可得到新的网络架构或者结构搜索空间，并支持多层级的占位模块（如将一个基本单元替换为多个占位模块等）等功能。

15 参数空间 1004：定义了候选集 1003 中的多种结构的参数的索引或者取值范围，为参数优化器 902 提供了同一的访问结构（interface）。参数空间中记录了候选集中的所有的结构的参数对应的架构参数，并支持通过统一的接口进行查找、遍历、赋值或者导出等操作。例如，若候选集中包括了卷积算子 A 和卷积算子 B，则可以为卷积算子 A 和卷积算子 B 构建对应的索引，如 11 代表卷积算子 A，10 代表卷积算子 B 等，因此后续在参数空间 1004
20 中采样时，可以直接采集卷积算子 A 或者卷积算子 B 的索引，而无需采集卷积算子 A 或者卷积算子 B 的具体结构，提高了采样效率以及采样的数据量。此外，参数空间还提供了自定义的参数回调接口，可以在参数更新的时候调用用户自定义的功能。更具体地，该参数空间 1004 可以参阅前述步骤 603 中的描述。

25 结构搜索空间 1005：包括了候选集所包括的结构的参数与骨干网络 1002 的能的组合方式得到的结构。

参数优化器 902：定义了优化算法，在参数空间 1004 中寻找可能的最优解（即采样参数）。参数优化器 902 每次给出一组采样参数作为可能的最优解，然后架构评估器 904 通过预设的数据集对该可能的最优解对应的架构进行评估。评估后得到的指标用来更新参数优化器 902 的采样方式，以给出下一组更优的采样参数。

30 架构构建器 903：可以理解为具有两个主要的功能：构建和转化。其中，构建即通过候选集构建参数空间或者通过候选集合骨干网络 1002 构建结构搜索空间 1005。具体地，架构构建器 903 可以遍历整个骨干网络 1002 的模型，并将骨干网络 1002 中的每个占位模块（Slot）与候选集或者候选集的子集进行绑定，得到一个或多个可运行的完整架构，组成结构搜索空间 1005。转化即当不存在结构搜索空间 1005 时，根据参数优化器 902 采集
35 到的采样参数，将骨干网络 1002 按照采样参数进行转化，如将骨干网络 1002 中的 slot 的通道由 3 转化为 6，从而得到完整的网络模型。架构构建器 903 根据当前采样参数的值或对应的编码对骨干网络 1002 进行变换操作。常见变换操作包括重构整个架构，改变某个模

块的输入至或者输出值的权重，改变各模块的计算顺序等。通常，转化步骤可通过自定义转化函数或者通过参数更新的回调接口来实现。

架构评估器 904：用于对架构构建器 903 构建的网络模型进行评估，得到评估结果，并将评估结果返回参数优化器 902，从而使参数优化器 902 根据评估结果优化对参数空间的采样方式。架构评估器 904 具体可以通过预先设定的数据集 (dataset) (图 10 中未示出) 对网络模型进行评估，该数据集可以是用户输入的数据集，也可以是本地的数据集。该数据集中可以包括用于对网络模型进行评估的数据，例如可以包括图片以及图片包括的目标或者图片以及图片对应的分类结果等。

其中，本申请提供的神经网络构建系统具体执行的步骤可以包括如下所述。

接收用户输入的初始骨干网络 1001，然后将初始骨干网络 1001 中的部分基本单元替换为占位模块 (slot)，得到骨干网络 (backbone) 1002。初始骨干网络 1001 通常是一个完整的神经网络架构，通常是手工设计的神经网络或者根据任务生成的神经网络。其中初始骨干网络 1001 的某些关键部位可以用占位模块 (Slot) 进行替换，得到一个可变的待定架构，即包括了 slot 的骨干网络 1002。同时，对于每一个占位模块，可以由根据不同的任务或者由用户选择一个候选子集，包含了该 slot 的架构属性的可变范围。候选集通常包括了基础的网络算子或基本构建单元，或者抽象的网络属性 (如宽度，深度等) 等。对于一些常见的架构搜索场景，通常神经网络构建系统中可以内置多种候选集，此时则可以无需用户额外指定候选集。

候选集 1003 也可以是用户输入的数据，也可以是本地数据库中采集到的数据等。

可选地，可以根据候选集 1003 构建参数空间 (parameter space) 1004，该参数空间中包括了候选集中对应的多种结构的参数对应的架构参数，也可以理解为候选集中所包括的多种结构的参数的索引或者取值范围。

可选地，架构构建器 903 还可以根据候选集和骨干网络 1002 构建结构搜索空间 (Arch space) 1005，该结构搜索空间 1005 中包括了所有可能的组合方式构成的网络架构。

参数优化器 902 可以理解为前述的采样模块 902，该参数优化器可以从参数空间 1004 中进行采样，得到至少一组采样参数，然后将该至少一组采样参数反馈给架构构建器 903。

架构构建器可以根据参数优化器 902 采集到的至少一组采样参数搜索对应的结构，从而得到网络模型。具体地，架构构建器 903 可以从结构搜索空间 1005 中直接搜索完整的架构，得到网络模型，也可以是从候选集中搜索对应的占位模块的结构，并根据搜索到的结构对骨干网络 1002 中的 slot 的结构进行转换，例如，搜索到的通道为 6，而 slot 自身的通道为 3，则可以将 slot 的通道转换为 6，从而得到网络模型。

架构构建器 903 在得到网络模型之后，将该网络模型输入至架构评估器 904，该架构评估器 904 可以通过预设的数据集对网络模型进行评估，并将评估结果反馈至参数优化器 902。其中，若网络模型的评估结果符合预设条件，则可以直接将网络模型作为目标神经网络，并输出对应的采样参数或者直接输出目标神经网络等。若网络模型的评估结果不满足预设条件，则参数优化器 902 可以根据架构评估器 904 反馈的评估结果来优化采样方式，使下一次采样到的采样参数对应的网络模型的评估结果更接近或者符合预设条件，从而可

以提高得到最终的目标神经网络的效率。

此外，也可以无需构建参数空间 1004 或者结构搜索空间 1005，如图 11 所示，与前述图 10 的不同之处在于，无需构建参数空间 1004 或者结构搜索空间 1005，采样模块 902 可以直接从候选集 1003 中采集对应的结构的参数，如直接采集卷积算子、算子属性等。然后由架构构造器 903 直接对骨干网络 1002 中的 slot 的结构进行转换，从而得到网络模型。对于图 11 中与图 10 的相同之处此处不再赘述。

因此，本申请实施方式提供的神经网络构建系统，可以将架构搜索过程分解为参数空间，参数优化器，架构估计器和架构构建器几个独立模块间的统一交互流程，并支持基于离散，连续优化等多种搜索方案按此方式进行解耦及运行，通过改变占位模块的结构，来得到不同结构的网络模型，从而适应不同的场景，泛化能力强。并且将骨干网络与候选集解耦并模块化，以实现其代码复用及自由组合，减少了场景迁移所需的代码量，实现跨场景架构搜索应用的高效开发部署。在此基础上，提供了用户友好的调用接口，支持用户将自定义骨干网络架构通过极少代码自动转化为结构搜索空间，而无需手动定义搜索空间，实现用户友好的结构搜索方式，提高用户体验。

例如，搜索方案 A 中包含了参数优化算法 O1 和架构评估方法 E1，而为了设计新的搜索方案 B，采用优化算法 O2，保持评估方法 E1 不变，则只能全部重写 O2 和 E1，造成了评估方法实现的冗余。而本申请提供的神经网络构建方法中，可以通过较少的代码，实现跨场景架构搜索应用的高效开发部署。例如，针对不同算力开销的使用场景，本申请提供的神经网络构建系统可在最少 1 个 GPU 天，最多任意长的搜索时间内完成结构搜索流程，构建得到目标神经网络。本申请提供的神经网络构建系统不限制网络架构的具体任务，因此适用于图像分类、推荐搜索、自然语言处理等多种任务的架构搜索任务。通过本申请提供的神经网络构建系统，用户可以将自己的搜索算法在多种骨架网络上进行评估测试，并与其他算法进行公平比较。用户还可以使用本申请提供的神经网络构建系统已支持的算法，优化自身的网络模型，以提高网络模型的表现或降低模型计算开销，最终将模型部署到车载物体检测，人脸识别，应用搜索推荐等真实场景中。

示例性地，为便于理解，下面以一个更具体的场景对本申请提供的神经网络构建方法和神经网络构建系统进行更详细的介绍。MobileNetV2 骨干网络由 n 个重复的基本卷积单元 (Mobile Inverted Conv) 串联构成，每个单元的宽度是网络基础宽度的 6 倍，n 为正整数。因此，用户仅需通过一句代码即可将所有的基本单元转化为占位模块 (Slot)。例如，骨干网络中的代码可以是：`self.conv=MobileInvertedConv(chn_in,chn_out,stride,C=C,activation)`，更换为占位模块的代码可以是：`self.conv=Slot(chn_in,chn_out,stride,kwarg={ 'C' ;C, ' activation' :activation})`，实现的结果效果可以如图 12 所示，其中，初始骨干网络中的基本卷积单元，如 MBConv1、MBConv2 和 MBConv3 即可替换为 slot1、slot2 和 slot3。因此，可以使用较少的代码实现将基本单元替换为 slot 的结构，代码量较少，可以提高用户体验。

用户可以将修改后的网络定义（即修改后的代码）输入到本申请提供的神经网络构建系统中，同时指定目标神经网络的训练的数据集为 ImageNet，即可开始 slot 的结构搜索

流程。

首先框架内置的架构构建器 (Constructor) 将进行构建 (Construct) 步骤。如图 13 所示, 对于每一个占位模块, 架构构建器可以将其转换为一个与原始骨架定义相同, 但宽度可变的基本卷积单元, 即 MBConv(E), 并声明一个架构参数 p (如图 13 中所示的 $p1$ 、 $p2$ 和 $p3$), 如 p 的取值范围为基础宽度的 1, 3, 6, 9 倍, 即为 slot 绑定一个候选子集。同时为每个 MBConv(E) 绑定一个转化回调函数 (Handler), 使得每当 p 更新时, 在转化步骤根据 p 的具体取值, 动态地调整该单元的宽度为对应倍数。当遍历所有的占位模块后, 架构构建器的构建步骤结束。此时的参数空间包含 n 个代表宽度的架构参数。

下面进入搜索流程。如图 14 所示, 首先是参数更新步骤。参数优化器执行优化算法, 如进化算法, 并给出第一组用于评估的架构参数值, 假设分别为 3, 6, 9, 代表第一个到第三个单元的宽度。如图 14 所示, 将参数 $p1$ 的宽度从 3 更新为 6。

当架构参数值更新后, 进入架构转化 (Transform) 步骤。如图 14 所示, 对于每一个更新的参数, 如 p 的值从 3 变为 6, 与 p 绑定的转化函数将其对应的 MBConv(E) 的宽度从 3 倍动态调整为 6 倍。此时得到的网络架构即为转换后的架构。

转化后的架构进入架构评估步骤。架构评估器 (Estimator) 904 将该架构在 ImageNet 数据集上进行训练, 并在划分出的验证数据集上评估其性能, 假设为输出精确度为 80%。该性能指标被反馈给参数优化器, 并被用于更新进化算法的内部状态。此时进入下一轮循环, 重复上述流程, 直到得到输出精确度达到预定指标的模型。

假设下一个循环中的架构参数为 6, 6, 9, 评估后的精度为 85%。参数优化器更新之后则将倾向于选择宽度倍数更大的架构, 最终得到的架构将在性能上优于原始骨架网络。若在反馈中加入计算能效等指标, 则能够搜索到性能和计算能效之间平衡更优的架构。

因此, 本申请提供的神经网络构建方法和系统, 可以在跨场景的应用中基本无需代码重构, 而是通过更换框架各模块来实现, 减少开发调试成本, 针对同一使用场景可执行多种架构搜索方案及其组合, 提高架构搜索效率和最终表现, 且用户仅用少量的代码就能实现骨干网络的自动转换, 易用性较高, 实现了用户友好的神经网络构建方式, 提高用户体验。

请参阅图 15, 本申请提供的另一种神经网络构建装置的结构示意图, 如下所述。

该神经网络构建装置可以包括处理器 1501 和存储器 1502。该处理器 1501 和存储器 1502 通过线路互联。其中, 存储器 1502 中存储有程序指令和数据。

存储器 1502 中存储了前述图 6-14 中的步骤对应的程序指令以及数据。

处理器 1501 用于执行前述图 6-14 中任一实施例所示的神经网络构建装置执行的方法步骤。

可选地, 该神经网络构建装置还可以包括收发器 1503, 用于接收或者发送数据。

本申请实施例中还提供一种计算机可读存储介质, 该计算机可读存储介质中存储有用于生成车辆行驶速度的程序, 当其在计算机上行驶时, 使得计算机执行如前述图 6-14 所示实施例描述的方法中的步骤。

可选地, 前述的图 15 中所示的神经网络构建装置为芯片。

本申请实施例还提供了一种神经网络构建装置，该神经网络构建装置也可以称为数字处理芯片或者芯片，芯片包括处理单元和通信接口，处理单元通过通信接口获取程序指令，程序指令被处理单元执行，处理单元用于执行前述图 6-14 中任一实施例所示的神经网络构建装置执行的方法步骤。

5 本申请实施例还提供一种数字处理芯片。该数字处理芯片中集成了用于实现上述处理器 1501，或者处理器 1501 的功能的电路和一个或者多个接口。当该数字处理芯片中集成了存储器时，该数字处理芯片可以完成前述实施例中的任一个或多个实施例的方法步骤。当该数字处理芯片中未集成存储器时，可以通过通信接口与外置的存储器连接。该数字处理芯片根据外置的存储器中存储的程序代码来实现上述实施例中神经网络构建装置执行的动作。

10 本申请实施例中还提供一种包括计算机程序产品，当其在计算机上行驶时，使得计算机执行如前述图 6-14 所示实施例描述的方法中神经网络构建装置所执行的步骤。

本申请实施例提供的神经网络构建装置可以为芯片，芯片包括：处理单元和通信单元，所述处理单元例如可以是处理器，所述通信单元例如可以是输入/输出接口、管脚或电路等。
15 该处理单元可执行存储单元存储的计算机执行指令，以使服务器内的芯片执行上述图 4 至图 8 所示实施例描述的神经网络构建方法。可选地，所述存储单元为所述芯片内的存储单元，如寄存器、缓存等，所述存储单元还可以是所述无线接入设备端内的位于所述芯片外部的存储单元，如只读存储器（read-only memory, ROM）或可存储静态信息和指令的其他类型的静态存储设备，随机存取存储器（random access memory, RAM）等。

20 具体地，前述的处理单元或者处理器可以是中央处理器（central processing unit, CPU）、网络处理器（neural-network processing unit, NPU）、图形处理器（graphics processing unit, GPU）、数字信号处理器（digital signal processor, DSP）、专用集成电路（application specific integrated circuit, ASIC）或现场可编程逻辑门阵列（field programmable gate array, FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者也可以是任何常规的处理器等。

30 示例性地，请参阅图 16，图 16 为本申请实施例提供的芯片的一种结构示意图，所述芯片可以表现为神经网络处理器 NPU 160，NPU 160 作为协处理器挂载到主 CPU（Host CPU）上，由 Host CPU 分配任务。NPU 的核心部分为运算电路 1603，通过控制器 1604 控制运算电路 1603 提取存储器中的矩阵数据并进行乘法运算。

在一些实现中，运算电路 1603 内部包括多个处理单元（process engine, PE）。在一些实现中，运算电路 1603 是二维脉动阵列。运算电路 1603 还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中，运算电路 1603 是通用的矩阵处理器。

35 举例来说，假设有输入矩阵 A，权重矩阵 B，输出矩阵 C。运算电路从权重存储器 1602 中取矩阵 B 相应的数据，并缓存在运算电路中每一个 PE 上。运算电路从输入存储器 1601 中取矩阵 A 数据与矩阵 B 进行矩阵运算，得到的矩阵的部分结果或最终结果，保存在累加

器 (accumulator) 1608 中。

统一存储器 1606 用于存放输入数据以及输出数据。权重数据直接通过存储单元访问控制器 (direct memory access controller, DMAC) 1605, DMAC 被搬运到权重存储器 1602 中。输入数据也通过 DMAC 被搬运到统一存储器 1606 中。

5 总线接口单元 (bus interface unit, BIU) 1610, 用于 AXI 总线与 DMAC 和取指存储器 (instruction fetch buffer, IFB) 1609 的交互。

总线接口单元 1610 (bus interface unit, BIU), 用于取指存储器 1609 从外部存储器获取指令, 还用于存储单元访问控制器 1605 从外部存储器获取输入矩阵 A 或者权重矩阵 B 的原数据。

10 DMAC 主要用于将外部存储器 DDR 中的输入数据搬运到统一存储器 1606 或将权重数据搬运到权重存储器 1602 中或将输入数据数据搬运到输入存储器 1601 中。

向量计算单元 1607 包括多个运算处理单元, 在需要的情况下, 对运算电路的输出做进一步处理, 如向量乘, 向量加, 指数运算, 对数运算, 大小比较等等。主要用于神经网络中非卷积/全连接层网络计算, 如批归一化(batch normalization), 像素级求和, 对特征平面进行上采样等。

15 在一些实现中, 向量计算单元 1607 能将经处理的输出的向量存储到统一存储器 1606。例如, 向量计算单元 1607 可以将线性函数和/或非线性函数应用到运算电路 1603 的输出, 例如对卷积层提取的特征平面进行线性插值, 再例如累加值的向量, 用以生成激活值。在一些实现中, 向量计算单元 1607 生成归一化的值、像素级求和的值, 或二者均有。在实现中, 处理过的输出的向量能够用作到运算电路 1603 的激活输入, 例如用于在神经网络中的后续层中的使用。

控制器 1604 连接的取指存储器 (instruction fetch buffer) 1609, 用于存储控制器 1604 使用的指令;

25 统一存储器 1606, 输入存储器 1601, 权重存储器 1602 以及取指存储器 1609 均为 On-Chip 存储器。外部存储器私有于该 NPU 硬件架构。

其中, 循环神经网络中各层的运算可以由运算电路 1603 或向量计算单元 1607 执行。

其中, 上述任一处提到的处理器, 可以是一个通用中央处理器, 微处理器, ASIC, 或一个或多个用于控制上述图 6-14 的方法的程序执行的集成电路。

30 另外需说明的是, 以上所描述的装置实施例仅仅是示意性的, 其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的, 作为单元显示的部件可以是或者也可以不是物理单元, 即可以位于一个地方, 或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。另外, 本申请提供的装置实施例附图中, 模块之间的连接关系表示它们之间具有通信连接, 具体可以实现为一条或多条通信总线或信号线。

35 通过以上的实施方式的描述, 所属领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件的方式来实现, 当然也可以通过专用硬件包括专用集成电路、专用 CPU、专用存储器、专用元器件等来实现。一般情况下, 凡由计算机程序完成的功能都可以很容

易地用相应的硬件来实现,而且,用来实现同一功能的具体硬件结构也可以是多种多样的,例如模拟电路、数字电路或专用电路等。但是,对本申请而言更多情况下软件程序实现是更佳的实施方式。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在可读取的存储介质中,如计算机的软盘、U 盘、移动硬盘、只读存储器(read only memory, ROM)、随机存取存储器(random access memory, RAM)、磁碟或者光盘等,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本申请各个实施例所述的方法。

在上述实施例中,可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时,可以全部或部分地以计算机程序产品的形式实现。

所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时,全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中,或者从一个计算机可读存储介质向另一计算机可读存储介质传输,例如,所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存储的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质,(例如,软盘、硬盘、磁带)、光介质(例如,DVD)、或者半导体介质(例如固态硬盘(solid state disk, SSD))等。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等(如果存在)是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

最后应说明的是:以上,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以权利要求的保护范围为准。