

一种神经网络构建方法以及装置

本申请要求于 2020 年 05 月 25 日提交中国专利局、申请号为 202010448652.7、申请名称为“一种神经网络构建方法以及装置”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装置。

背景技术

10 在人工智能领域中，计算机视觉技术在自动驾驶，手机终端及安防监控等场景有着广泛的应用。主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，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。常用计算机视觉任务主要用于目标检测以及语义分割，通常可以先通过主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，对主干网络特征提取能力的需求也随之增加。

15 然而，在现有方案中，通常采用人工设计的或者搜索得到的分类网络的特征提取网络作为主干网络，或者，通过搜索得到的构建单元堆叠得到主干网络。现有技术中所获取的主干网络往往输出准确度低，获取主干网络的效率低下，并且，通过堆叠得到的主干网络，通常只能用于固定的场景中，在不同的场景中需要再次搜索构建单元并堆叠主干网络，泛化能力弱。

20

发明内容

本申请提供一种神经网络构建方法以及装置，用于准确高效地构建目标神经网络，构建出的目标神经网络输出的准确度高，还可以应用于不同的应用场景中，泛化能力强。

25 第一方面，本申请提供一种神经网络构建方法，包括：获取起点网络，该起点网络包括多个串行网络；基于预设的第一搜索空间对起点网络进行至少一次变形，得到串行网络，第一搜索空间包括对起点网络进行变形使用的参数的范围；若串行网络满足预设条件，则通过预设的数据集对串行网络进行训练，得到训练后的串行网络；若训练后的串行网络满足终止条件，则根据训练后的串行网络得到目标神经网络。

30 本申请实施方式中，可以对起点网络进行至少一次变形得到串行网络，当串行网络满足预设条件时，才通过预设的数据集对串行网络进行训练，而无需每次变形都对变形后的起点网络进行训练，减少了工作量，提高得到目标神经网络的效率。并且，变形后的起点网络可以继承变形前的起点网络的权重值，并通过轻量数据集对继承的权重值进行更新，从而实现对变形后的起点网络的快速训练，快速得到符合需求的串行网络，进而进一步提高得到目标神经网络的效率。并且，本申请可以通过对起点网络进行变形，从而使目标神经网络的输出的准确度更高。此外，本申请提供的神经网络构建的方法，可以适应不同的场景，可以根据不同的需求来搜索串行网络。可以面向更复杂的视觉任务，泛化能力强。35 从而可以节约构建不同的神经网络的成本，提高构建的神经网络的输出的准确性，提高用

户体验。

在一种可能的实施方式中，前述的基于第一搜索空间对起点网络进行至少一次变形中的其中一次变形，包括：在第一搜索空间中搜索变形策略，该变形策略包括对起点网络进行变形使用的一组参数；根据变形策略对起点网络进行变形，得到变形后的起点网络，变形后的起点网络的权重继承变形前的起点网络的权重；通过轻量数据集对变形后的起点网络的权重更新，得到权重更新后的起点网络，预设的数据集所包括的数据多于轻量数据集所包括的数据；对权重更新后的起点网络进行评估，得到第一评估结果；若第一评估结果满足预设的更新条件，则在第一搜索空间中重新搜索变形策略，并根据重新搜索得到的变形策略对变形前的起点网络进行重新变形；若第一评估结果不满足预设的更新条件，则将权重更新后起点网络作为串行网络。

因此，本申请实施方式中，在对起点网络进行变形时，变形后的起点网络可以继承变形前的起点网络的权重值，并根据轻量数据集对权重更新，可以减少权重值的训练过程，提高后续得到串行网络的效率，进而提高了得到目标神经网络的效率。并且，可以通过在第一搜索空间中搜索变形策略来对起点网络进行变形，第一搜索空间包括起点网络中各个特征层次相关的信息，如深度、宽度等，并且通常，在第一搜索空间中重新搜索到的变形策略，与上一次对起点网络进行变形所使用的变形策略不相同，从而避免无意义的变形，进而提高得到不符合更新条件的起点网络的效率。

在一种可能的实施方式中，上述方法还包括：若串行网络不满足预设条件，则将串行网络作为新的起点网络，或者若训练后的串行网络不满足终止条件，则将训练后的串行网络作为新的起点网络；更新第一搜索空间，并在更新后的第一搜索空间搜索变形策略，然后根据搜索到的变形策略对新的起点网络进行变形。

因此，本申请实施方式中，若串行网络不满足预设条件，则可以将串行网络作为新的起点网络，并重新对新的起点网络进行变形，直到选出满足预设条件或者终止条件的串行网络。因此，本申请实施例可以适用于多种不同的场景，针对目标任务选择出合适的串行网络，泛化能力强。

在一种可能的实施方式中，更新条件可以包括但不限于：对起点网络的变形次数不超过预设次数，或者，第一评估结果不满足预设的收敛条件，收敛条件包括但不限于以下一项或者多项：第一评估结果优于阈值，或者，第一评估结果与上一次对起点网络进行评估得到的评估结果之前的差值小于预设值。

因此，本申请实施方式中，对起点网络的变形次数不超过预设次数，或者第一评估结果不满足预设的收敛条件，即可继续对起点网络进行变形，直到得到满足不符合更新条件的起点网络。

在一种可能的实施方式中，预设条件可以包括但不限于：串行网络的第二评估结果满足第一预设评估条件，或者，串行网络的变形次数超过预设变形次数，或者，距离上一次对起点网络进行变形的时长超过预设值，第二评估结果为对串行网络进行评估得到的。

因此，本申请实施方式中，当串行网络的评估结果满足第一预设评估条件、串行网络的变形次数超过预设变形次数或者距离上一次对起点网络进行变形的时长超过预设值，则

可以对串行网络进行训练，从而得到训练后的串行网络，通常训练后的串行网络的输出结果更准确。

可选地，第一预设评估条件可以包括但不限于串行网络的损失值不大于第一损失阈值、输出结果的精度值大于第一精度阈值或者推理时长小于第一推理阈值等。

5 在一种可能的实施方式中，起点网络中的每个串行子网络由一个或者多个基本单元组成，对起点网络的变形，具体可以包括：交换多个串行子网络中相邻两个串行子网络所包括的基本单元；或者，在多个串行子网络中的一个或多个中插入处理后的基本单元，处理后的基本单元可以包括对基本单元的权重进行单位矩阵初始化和/或对基本单元进行分组归一化等得到。

10 本申请实施方式中，可以通过交换或转移串行子网络中相邻串行子网络包括的基本单元，或者在串行子网络中新增基本单元，来对起点网络进行变形，从而得到变形后的起点网络。

在一种可能的实施方式中，终止条件可以包括但不限于：串行网络的第二评估结果满足第二预设评估条件，或者，对串行网络的训练次数超过预设训练次数。

15 本申请实施方式中，当串行网络的评估结果满足第二预设评估条件，或者对串行网络的训练次数超过预设训练次数，即可终止对串行网络的变形，并根据该串行网络构建执行目标任务的目标神经网络。

可选地，第二预设评估条件可以包括串行网络的损失值不大于第二损失阈值、输出结果的精度值大于第二精度阈值或者推理时长小于第二推理阈值等。

20 在一种可能的实施方式中，前述的根据训练后的串行网络得到目标神经网络，可以包括：若训练后的串行网络满足输出条件，则将训练后的串行网络作为目标神经网络；若训练后的串行网络不满足输出条件，则获取与训练后的串行网络对应的并行子网络，并根据训练后的串行网络和与训练后的串行网络对应的并行子网络构建目标神经网络。

25 本申请实施方式中，若训练后的串行网络满足输出条件，则可以将训练后的串行网络作为目标神经网络，从而高效、准确地得到目标神经网络。若训练后的串行网络满足终止条件，而不满足输出条件，则可以获取与训练后的串行网络对应的并行子网络，并根据训练后的串行网络和与训练后的串行网络对应的并行子网络构建目标神经网络。可以理解为，可以获取与训练后的串行网络对应的并行子网络，并根据串行网络和并行子网络构建目标神经网络，从而提高得到的目标神经网络的输出结果的准确性，使根据串行网络和并行子网络构建得到的神经网络可以满足输出条件，从而可以得到输出结果更准确的目标神经网络。

30

在一种可能的实施方式中，串行网络可以包括多个串行的串行子网络，每个串行子网络包括一个或多个基本单元，上述的获取与训练后的串行网络对应的并行子网络，可以包括：在预设的数量范围内确定与多个串行子网络中每个串行子网络对应的并行子网络的数量；根据每个串行子网络对应的并行子网络的数量构建每个串行子网络对应的并行子网络。

35

因此，本申请实施方式中，可以在预设的数量范围内中搜索每个串行子网络对应的并行子网络的数量，然后该数量来构建每个串行子网络对应的并行子网络，从而可以构建输

出结果更准确的目标神经网络。

在一种可能的实施方式中，前述的根据串行网络和与每个串行子网络对应的并行子网络构建目标神经网络，可以包括：根据串行网络和每个串行子网络对应的并行子网络构建临时神经网络；对临时神经网络进行评估，得到第三评估结果；若第三评估结果满足输出条件，则将临时神经网络作为目标神经网络；若第三评估结果不满足输出条件，则重新获取与串行网络对应的并行子网络。

因此，本申请实施方式中，可以根据串行网络和对应的并行子网络构建符合输出条件的目标神经网络，从而可以准确、高效地得到目标神经网络，可以应用于不同的场景，泛化能力强。

在一种可能的实施方式中，输出条件包括以下一项或者多项：精度大于第一阈值，或者，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，每秒浮点运算次数不大于第五阈值，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。通常，输出条件可以是根据目标任务确定的，例如可以从用户输入数据来获取该输出条件，从而得到可以应用于不同场景的目标神经网络，提高用户体验。

在一种可能的实施方式中，第一搜索空间包括的参数的类型包括：宽度、深度、下采样的数量和下采样的位置，宽度为起点网络中的每一个串行子网络所包括的基本单元的输入通道和输出通道的数量，深度为起点网络中的每一个串行子网络所包括的基本单元的数量，下采样用于分割起点网络中两个相邻的串行子网络。因此，本申请实施方式中，可以在第一搜索空间中采样得到宽度、深度、下采样的数量或者下采样的位置等参数，从而根据搜索到的参数来对串行网络进行变形，提供了多种不同的变形方式，从而丰富了对串行网络进行变形的变形网络，从而可以准确地得到符合条件的串行网络。

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，预设的数据集包括 ImageNet 数据集。因此，本申请实施方式中，可以通过 ImageNet 数据集对串行网络进行训练，从而得到输出结果更准确的串行网络。

在一种可能的实施方式中，获取起点网络的具体方式可以包括：通过用户输入数据获取到目标任务，然后根据该目标任务来获取起点网络，而构建得到的目标神经网络可以用于执行该目标任务。

第二方面，本申请提供一种神经网络构建装置，包括：

获取模块，用于获取起点网络，起点网络包括多个串行子网络；

变形模块，用于基于预设的第一搜索空间对起点网络进行至少一次变形，得到串行网络，第一搜索空间包括对起点网络进行变形使用的参数的范围；

训练模块，用于若串行网络满足预设条件，则通过预设的数据集对串行网络进行训练，得到训练后的串行网络；

构建模块，用于若训练后的串行网络满足终止条件，则根据训练后的串行网络得到目标神经网络。

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，变形模块，具体可以用于：在预设的第一搜索空间中搜索变形策略，变形策略包括对起点网络进行变形使用的一组参数；根据变形策略对起点网络进行变形，得到变形后的起点网络，变形后的起点网络的权重为继承变形前的起点网络的权重得到；通过轻量数据集对变形后的起点网络的权重更新，得到权重更新后的起点网络，预设的数据集所包括的数据多于轻量数据集所包括的数据；对权重更新后的起点网络进行评估，得到第一评估结果；若第一评估结果满足预设的更新条件，则可以在第一搜索空间中重新搜索变形策略，并根据重新搜索到的变形策略对变形前起点网络进行重新变形；若第一评估结果不满足预设的更新条件，则将权重更新后起点网络作为串行网络。

在一种可能的实施方式中，神经网络构建装置还可以包括：

更新模块，用于若串行网络不满足预设条件，则将串行网络作为新的起点网络，或者，若训练后的串行网络不满足终止条件，则将训练后的串行网络作为新的起点网络，更新第一搜索空间，并根据更新后的第一搜索空间重新搜索变形策略，根据重新搜索到的变形策略对新的起点网络进行再次变形。

在一种可能的实施方式中，更新条件包括：对起点网络的变形次数不超过预设次数，或者，第一评估结果不满足预设的收敛条件，收敛条件包括以下一项或者多项：第一评估结果优于阈值，或者，第一评估结果与上一次对起点网络进行评估得到的评估结果之前的差值小于预设值。

在一种可能的实施方式中，预设条件可以包括但不限于：串行网络的第二评估结果满足第一预设评估条件，或者，串行网络的变形次数超过预设变形次数，或者，距离上一次对起点网络进行变形的时长超过预设值，第二评估结果为对串行网络进行评估得到的，第一预设评估条件包括以下一项或者多项：串行网络的损失值不大于第一损失阈值、串行网络的输出结果的精度值大于第一精度阈值，或者串行网络的推理时长小于第一推理阈值。

在一种可能的实施方式中，起点网络由多个串行子网络组成，其中，每个串行子网络由一个或者多个基本单元组成，变形模块，具体可以用于：交换多个串行子网络中相邻两个串行子网络所包括的基本单元；或者，在多个串行子网络中的一个或多个中插入处理后的基本单元，插入的处理后的基本单元为对基本单元的权重进行单位矩阵初始化的权重和/或对基本单元进行分组归一化等得到。

在一种可能的实施方式中，终止条件可以包括但不限于：串行网络的第二评估结果满足第二预设评估条件，或者，对串行网络的训练次数超过预设训练次数。

在一种可能的实施方式中，构建模块，具体可以用于：若训练后的串行网络满足输出条件，则将训练后的串行网络作为目标神经网络；若训练后的串行网络不满足输出条件，则获取与训练后的串行网络对应的并行子网络，并根据训练后的串行网络和与训练后的串行网络对应的并行子网络构建目标神经网络。

在一种可能的实施方式中，串行网络包括多个串行子网络，每个串行子网络包括一个或多个基本单元，构建模块，具体可以用于：在预设的数量范围内确定与多个串行子网络中每个串行子网络对应的并行子网络的数量；根据每个串行子网络对应的并行子网络的数

量构建每个串行子网络对应的并行子网络，每个串行子网络的结构和每个串行子网络对应的并行子网络的结构相同。

在一种可能的实施方式中，构建模块，具体可以用于：根据串行网络和每个串行子网络对应的并行子网络构建临时神经网络；对临时神经网络进行评估，得到第三评估结果；
5 若第三评估结果满足输出条件，则将临时神经网络作为目标神经网络；若第三评估结果不满足输出条件，则重新获取与串行网络对应的并行子网络。

在一种可能的实施方式中，输出条件包括但不限于以下一项或者多项：精度大于第一阈值，或者，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，每秒浮点运算次数不大于第五阈值，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。
10

在一种可能的实施方式中，第一搜索空间包括的参数的类型可以包括但不限于：宽度、深度、下采样的数量和下采样的位置，宽度为起点网络中的每一个串行子网络所包括的基本单元的输入通道和输出通道的数量，深度为起点网络中的每一个串行子网络所包括的基本单元的数量，下采样用于分割起点网络中两个相邻的串行子网络。

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，预设的数据集包括 ImageNet 数据集。
15

在一种可能的实施方式中，获取起点网络的具体方式可以包括：通过用户输入数据获取到目标任务，然后根据该目标任务来获取起点网络，而构建得到的目标神经网络可以用于执行该目标任务。

第三方面，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。
20

第四方面，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。
25

第五方面，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。

第六方面，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。
30

附图说明

- 图 1 为本申请应用的一种人工智能主体框架示意图；
35 图 2 为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；
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 为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网络变形方式示意图；
图 8 为本申请实施例提供的另一种网络变形方式示意图；
5 图 9 为本申请实施例提供的一种神经网络构建方式示意图；
图 10 为本申请实施例提供的另一种神经网络构建方法的流程示意图；
图 11 为本申请实施例提供的另一种网络变形方式示意图；
图 12 为本申请实施例提供的另一种神经网络构建方式示意图；
图 13 为本申请实施例提供的一种神经网络结构示意图；
10 图 14 为本申请实施例提供的一种神经网络构建装置的结构示意图；
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

15 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) 20 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。

25 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。

30 “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。

35 （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, 5 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、10 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。

15 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、25 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-图 13 中的相关描述。

30 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 发送目标任务, 请求执行设备构建神经网络, 并向执行设备 210 发送用于训练的数据库。

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。

计算模块 211 使用目标模型/规则 201 对输入的数据进行处理。具体地，计算模块 211 用于：根据目标任务获取起点网络基于预设的第一搜索空间对起点网络进行至少一次变形，得到串行网络，串行搜索空间包括对起点网络进行变形使用的参数；若串行网络满足预设条件，则通过预设的数据集对串行网络进行训练，得到训练后的串行网络；若训练后的串行网络满足终止条件，则根据训练后的串行网络得到目标神经网络，目标神经网络用于执行目标任务。可以理解为，计算模块 211 可以用于构建神经网络的主干网络。

关联功能模块 213 和关联功能模块 214 可以用于搜索与目标任务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神经网络。

更深层地，构建模块 202 可以针对不同的目标任务，基于不同的数据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与目标任务关联的数据存入数据库 230。

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。

在本申请所提及的训练或者更新过程可以由构建模块 202 来执行。可以理解的是，神经网络的训练过程即学习控制空间变换的方式，更具体即学习权重矩阵。训练神经网络的目的是使神经网络的输出尽可能接近期望值，因此可以通过比较当前网络的预测值和期望值，再根据两者之间的差异情况来更新神经网络中的每一层神经网络的权重向量（当然，在第一次更新之前通常可以先对权重向量进行初始化，即为深度神经网络中的各层预先配置参数）。例如，如果网络的预测值过高，则调整权重矩阵中的权重的值从而降低预测值，经过不断的调整，直到神经网络输出的值接近期望值或者等于期望值。具体地，可以通过损失函数(loss function)或目标函数(objective function)来衡量神经网络的预测值和期望值之间的差异。以损失函数举例，损失函数的输出值(loss)越高表示差异越大，神经网络的训练可以理解为尽可能缩小 loss 的过程。本申请以下实施方式中更新起点网络的权重以及对串行网络进行训练的过程可以参阅此过程，以下不再赘述。

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)、公路网络或残差网络其他神经网络等。

示例性地,下面以卷积神经网络(convolutional neural network, CNN)为例。

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。

5 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。

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。

20 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。

25 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。

30 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。

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；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-图 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。

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-图 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。

自动机器学习（AutoML）：是指设计一系列高级的控制系统去操作机器学习模型，使得模型可以自动化地学习到合适的参数和配置而无需人工干预。在基于深度神经网络的学习模型中，自动计算学习主要包括网络架构搜索与全局参数设定。其中，网络架构搜索用于根据数据让计算机生成最适应问题的神经网络架构，具有训练复杂度高，性能提升大的特

点。

残差网络 (Residual Networks): 一种卷积神经网络。在图像识别中, 相比于传统的卷积神经网络识别准确率更高。在残差网络设计中, 会很多结构相同的子模块, 通常会用 ResNet 连接一个数字表示子模块重复的次数, 比如 ResNet50 表示其中有 50 个子模块。

5 主干网络 (Backbone): 在检测器、分割器或分类器等中用来对输入信息做特征提取的网络结构。通常, 在神经网络中, 除了主干网络之外, 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, 用于对主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。

10 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。

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、获取起点网络。

15 其中, 该起点网络可以是包括多个串行子网络, 该多个串行子网络通过串行的方式连接, 每个串行子网络由一个或者多个基本单元组成。

具体地, 该起点网络可以是分类、分割或者检测任务中常用的主干网络, 如未携带预测头的分类网络、分割网络或者检测网络等, 具体可以根据实际应用场景进行调整。

20 可选地, 起点网络可以根据目标任务选择的, 该目标任务即需要构建的神经网络所执行的任务。当然, 起点网络也可以是随机选择的网络, 具体可以根据实际应用进行调整, 本申请通过目标任务选择起点网络, 使选择的起点网络与目标任务的需求接近, 可以减少训练过程, 提高得到目标神经网络的效率。

25 在一种可能的实施方式中, 在步骤 601 之前, 还可以获取目标任务, 该目标任务可以是需要构建的目标神经网络执行的任务。若本申请可以由前述的执行设备 210 来执行, 该目标任务可以是接收到的其他设备 (如前述的客户设备 240 或者本地设备 301 或本地设备 302 等) 发送的任务, 也可以是根据执行设备 210 自身需求生成的目标任务。

30 此外, 还可以获取与目标任务关联的信息, 例如, 与目标任务关联的数据集, 执行目标任务的目标神经网络满足的精度、计算复杂度、推理速度、推理时长等信息, 具体可以根据该与目标任务关联的信息来选择起点网络。例如, 可以根据目标神经网络满足的精度、计算复杂度、推理速度、推理时长等信息来选择起点网络, 使起点网络满足或者接近目标神经网络需要满足的精度、计算复杂度、推理速度、推理时长等。具体例如, 若目标任务为构建平均精度为 80% 的图片识别神经网络, 则在选择起点网络时, 可以选择平均精度大于 80% 或者接近 80% 的分类网络作为起点网络。因此, 本申请实施方式中, 可以根据目标任务对应的需求来选择起点网络, 在给定的计算资源的限制下, 挑选出能够获得最大效益的起点网络。如根据精度、计算复杂度、显存占用量、每秒浮点运算次数 (floating-point operations per second, FLOPS)、CPU 占用率、推理速度、推理时长等来选择起点网络, 从而获得条件限制下的最优的起点网络, 进而快速得到串行网络, 可以减少得到目标神经网络的训练过程, 提高后续得到目标神经网络的效率。

在一些可能的场景中，目标神经网络可以用于进行图片识别、语义分割或者目标检测等中的一种或者多种。例如，客户设备 240 可以向执行设备 210 发送目标任务，该目标任务可以为准确度为 80% 的图片识别任务，或者，准确度为 95% 的语义分割任务或者精度为推理时长为 100ms 的目标检测任务等。

5 602、基于第一搜索空间对起点网络进行至少一次变形，得到串行网络。

在选择了起点网络，或者串行网络不满足预设条件，或者训练后的串行网络不满足终止条件之后，可以基于第一搜索空间对起点网络进行至少一次变形，将变形后的起点网络作为串行网络。该第一搜索空间中包括对起点网络进行变形使用的参数的范围，如深度的取值范围或者宽度的深度的取值范围等。

10 其中，本步骤中的提及的变形，可以是在选择了起点网络之后，基于第一搜索空间对起点网络进行至少一次变形；或者，也可以是在确定串行网络不满足预设条件（参阅以下步骤 603）或者训练后的串行网络不满足终止条件（参阅以下步骤 605）之后，可以将不满足预设条件的串行网络或者不满足终止条件的训练后的串行网络作为新的起点网络，并在对第一搜索空间进行更新之后，再对新的起点网络进行变形。可以理解为，若在确定串行网络不满足预设条件或者训练后的串行网络不满足终止条件之后，则可以将串行网络或者训练后的串行网络作为新的起点网络进行变形，从而再次得到串行网络。

具体地，对起点网络进行至少一次变形中的其中一个变形的具体流程可以包括：在第一搜索空间中搜索变形策略，即对起点网络进行变形所使用的一组参数，然后根据该变形策略对起点网络进行变形，得到变形后的起点网络，变形后的起点网络的权重继承了变形前的起点网络的权重。然后通过轻量数据集对变形后的起点网络的权重值进行更新，得到权重更新后的起点网络。然后对权重更新后的起点网络进行评估，得到第一评估结果。判断该第一评估结果是否满足预设的更新条件，若该第一评估结果满足更新条件，则在第一搜索空间中重新搜索变形策略，并根据重新搜索到的变形策略对起点网络进行重新变形，其中，在对起点网络进行重新变形时，可以在第一搜索空间中重新搜索变形策略进行变形，且重新搜索到的变形策略与上一次变形所使用的变形策略通常是不相同的。

25 可以理解的是，本步骤中可能对起点网络进行了多次变形，并进行权重更新，直到得到不符合更新条件的起点网络。在此过程中，得到了多个权重更新后的起点网络，若对起点网络的更新次数已达到限定值，则可以从该多个权重更新后的起点网络中确定出一个最优的起点网络作为最终的串行网络。例如，若对起点网络进行了多次变形，得到了多个权重更新后的起点网络，该多个权重更新后的起点网络的输出结果的精度分别为 50%、60% 以及 80%，则将精度为 80% 的起点网络作为最终的串行网络。

下面分别对起点网络的变形的流程进行更详细的解释。

30 其中，该轻量数据集包括的数据用于更新变形后的起点网络的权重值。可以理解为，可以根据轻量数据集所包括的数据对变形后的起点网络进行快速训练，从而高效地更新变形后的起点网络的权重值。具体地，该轻量数据集可以是数据采集设备 260 采集到的数据，也可以是由客户设备发送至执行设备 210 的数据。例如，该轻量数据集中所包括的数据可以是本地 ImageNet 数据集中的部分数据，或者，该轻量数据集中所包括的数据可以是数据

库 230 中包括的部分数据。

具体地，该第一评估结果可以包括但不限于起点网络的损失值、输出结果的精度值或者推理时长等等。相应地，对起点网络的评估的具体过程可以包括：计算起点网络的损失函数的值。计算起点网络的输出结果的准确度、或者计算起点网络从输入到输出的推理时长等，得到第一评估结果。

可选地，该更新条件可以包括但不限于以下一项或多项：对起点网络的变形次数不超过预设次数，或者，第一评估结果不满足预设的收敛条件，该预设的收敛条件可以根据第一评估结果所包括的数据确定，具体地，可以是该第一评估结果所包括的数据优于阈值，或者，第一评估结果所包括的数据与上一次对起点网络进行评估得到的评估结果之前的差值小于预设值等。例如，若第一评估结果中包括起点网络的损失值，则该收敛条件可以是损失值小于 20，或者第一评估结果与上一次对起点网络进行评估得到的损失值之间的差值小于 10。又例如，若第一评估结果中包括起点网络的进行分类的精度值，则该收敛条件可以是该精度值大于 80%，或者，该收敛条件可以是本次评估得到的精度值与上一次对起点网络进行评估得到的精度值之间的差值小于 5%等。

该第一搜索空间中包括对起点网络进行变形使用的参数的范围。具体可以包括起点网络中各个特征层次 (stage) 相关的信息，如一个 stage 的宽度、深度或者下采样的位置等取值范围。其中，起点网络中的一个特征层次可以理解为一个串行子网络。

通常，起点网络或者串行网络可以由多个串行的串行子网络组成，每个串行子网络中可以包括一个或者多个基本单元 (block)。具体地，起点网络中所包括的基本单元，可以分为基础单元 (basic block, BB)、瓶颈单元 (bottle neck block, BNB) 以及瓶颈分组 (ResNeXt) 单元等。

在一种可能的实施方式中，第一搜索空间中包括的参数类型可以包括：宽度、深度、下采样的数量和下采样的位置等。其中，宽度为起点网络中每个串行子网络所包括的基本单元的输入通道和输出通道的数量，深度为起点网络中每个串行子网络所包括的基本单元的数量，下采样用于分割起点网络中两个相邻的串行子网络，下采样的数量用于确定起点网络中串行子网络的数量，下采样的位置用于分割两个相邻的串行子网络。下采样可以理解为起点网络的各个串行子网络在对图片处理的过程中，可以对图片进行下采样之后再处理，以减少需要处理的图片的数据量，从而降低工作量。第一搜索空间所包括的内容例如：网络深度：8-60 个基本单元；特征层次层数：5-7 个特征层次，即串行子网络的数量；网络宽度：表示串行子网络中的基础单元的输入通道和输出通道的数量，更进一步地，还可以表示为串行网络中通道大小加倍的位置。

例如，如图 7 所示，起点网络中可以包括 4 个串行子网络，即图 7 中所示的串行子网络 1、串行子网络 2、串行子网络 3 和串行子网络 4。“1”表示对应的基本单元的输出通道是输入通道的 1 倍，当然，此处的“1”也可以替换为“2”或者“3”等，具体可以根据实际应用场景进行调整。“,”表示下采样，例如，block2 和 block3 之间存在“,”，表示 block2 和 block 处于不同的串行子网络中，具体地，在对图片进行处理的过程中，可以对图片进行下采样，从而减少需要处理的图片的数据量，进而降低工作量。当前，串行网络中的每

个串行子网络也可以不对图片进行下采样，即每个串行子网络处理的图片的数据量都相同，从而提高检测结果的准确性，具体可以根据实际应用场景确定是否进行下采样，本申请对此不作限定。

更进一步地，对起点网络进行变形的具体过程可以包括：交换多个串行子网络中相邻两个串行子网络所包括的基本单元，或者也可以理解为，交换下采样和基本单元的位置；或者，在多个串行子网络中的一个或多个中插入基本单元，插入的基本单元为根据单位矩阵初始化的权重和分组归一化得到。为便于理解，可以将多个串行子网络中的一个或多个中插入基本单元的过程称为扩展。即对起点网络的变形操作可以包括交换或者扩展。示例性地，如图 8 所示，可以将串行子网络 1 中的 block2 调整至串行子网络 2 中，还可以将串行子网络 3 中的 block5 调整至串行子网络 2 中，即调整下采样“，”的位置，使串行子网络 1 中的 block2 调整至串行子网络 2 以及将串行子网络 3 中的 block5 调整至串行子网络 2 中。此外，还可以在串行子网络 4 中扩展 block8。

相应地，对起点网络进行变形的具体方式可以包括：在第一搜索空间中搜索对起点网络进行变形的变形策略，该变形策略中包括对起点网络进行变形所述使用的参数，如起点网络中各个串行子网络的宽度、深度、起点网络中下采样的位置或者下采样的数量等。然后根据该变形策略，调整起点网络中各个串行子网络的宽度、深度以及下采样的位置或者下采样的数量等，得到变形后的起点网络。例如，当某一个串行子网络的深度小于变形策略中搜索到的深度时，可以在该串行子网络中插入新的基本单元，从而增加该串行子网络的深度。又例如，当某一个串行子网络的宽度小于变形策略中的宽度时，可以调整该串行子网络中的基本单元的输入通道和输出通道，使变形后的串行子网络符合变形策略所包括的宽度。

并且，在进行交换和扩展时，为使变形后的起点网络可以继承变形前的起点网络的权重值，还需要对交换或者扩展的基本单元进行处理。具体地，在进行交换的过程中，可以改变相邻基本单元的卷积步长，使得交换后的起点网络，可以复载变形前的起点网络的网络参数，如权重值，降低调整起点网络的网络参数的工作量。在扩展的过程中，新插入的单元为采用单位矩阵初始化的权重和分组归一化进行处理得到，使得变形后的起点网络可以继承变形前的起点网络的参数。可以理解为，在交换或者扩展起点网络所包括的基本单元之后，对于交换或者扩展的基本单元，可以通过改变其结构，使得交换或者扩展的基本单元与起点网络中对应的串行子网络中的基本单元结构类似，例如，新插入的单元的宽度与插入的串行子网络的宽度相同，或者，变形后的基本单元与变形前的基本单元的宽度相等，从而可以继承变形前的起点网络的权重值，从而减少权重训练的过程。因此，本申请实施方式中，可以在短时间内，快速得到变形后的起点网络，并得到变形后的起点网络的评估结果，提高得到串行网络的效率。

603、判断串行网络是否满足预设条件，若是。则执行步骤 604，若否，则执行步骤 602。

其中，在对起点网络进行至少一次变形得到串行网络之后，判断串行网络是否满足预设条件。若串行网络满足预设条件，则可以通过预设的数据集对串行网络进行训练，若串行网络不满足预设条件，则可以将该串行网络作为新的起点网络，继续在第一搜索空间中

搜索变形策略，从而继续对新的起点网络进行变形。

在一种可能的实施方式中，预设条件可以包括但不限于：串行网络的第二评估结果满足第一预设评估条件，或者，串行网络的变形次数超过预设变形次数，或者，对串行网络的训练次数不超过预设训练次数，第二评估结果为对串行网络进行评估得到的。其中，该第二评估结果可以是在前述步骤 602 中的进行至少一次变形之后的起点网络的第一评估结果，也可以是得到串行网络之后对串行网络进行再次评估得到的评估结果。该第二评估结果可以包括但不限于起点网络的损失值、输出结果的精度值或者推理时长等等。相应地，该第一预设评估条件可以包括起点网络或者串行网络的损失值不大于第一损失阈值、输出结果的精度值大于第一精度阈值或者推理时长小于第一推理阈值等。串行网络的变形次数即前述步骤 602 中对起点网络进行变形的次数。对串行网络的训练次数不超过预设训练次数即以下步骤 604 中通过预设的数据集对串行网络进行训练的次数。

604、通过预设的数据集对串行网络进行训练，得到训练后的串行网络。

在确定起点网络满足预设条件之后，通过预设的数据集对串行网络进行训练，得到训练后的串行网络。

其中，该预设的数据集可以理解为全量数据集，其中包括大量的训练数据。例如，若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，则该全量数据集可以是 ImageNet 数据集，其中包括大量的图片数据。因此，本申请实施方式中，并非在每次对起点网络进行变形即通过 ImageNet 数据集进行训练，而是在对起点网络进行至少一次变形，并满足预设条件之后，才对串行网络进行训练，可以降低对串行网络进行训练的工作量，提高得到目标神经网络的效率。

可以理解的是，通常，为了提高最终得到的串行网络的输出结果的准确性，需要通过大量的数据集进行训练，如 ImageNet 数据集或者其他包括大量数据的数据集，而通过大量的数据集进行训练需要耗费更多的资源，而本申请中，在串行网络满足预设条件之后才通过大量数据对串行网络进行训练，而无需在每次变形之后都进行训练，从而可以减少对串行网络进行训练的工作量，提高得到最终的串行网络的效率。

605、判断训练后的串行网络是否满足终止条件，若是，则执行步骤 606，若否，则执行步骤 602。

在通过预设的数据集对串行数据进行训练之后，判断训练后的串行网络是否满足终止条件。若训练后的串行网络满足终止条件，则可以根据训练后的串行网络来得到目标网络，即执行步骤 606。若训练后的串行网络不满足终止条件，则可以在更新第一搜索空间之后，将该训练后的串行网络作为新的起点网络，并对新的起点网络进行至少一次变形，然后再将变形后的起点网络作为串行网络。

在一种可能的实施方式中，终止条件可以包括但不限于：串行网络的第二评估结果满足第二预设评估条件，或者，对串行网络的训练次数超过预设训练次数。

其中，该第二评估结果可以是在前述步骤 602 中的进行至少一次变形之后的起点网络的第一评估结果，也可以是得到串行网络之后对串行网络进行再次评估得到的评估结果。该第二评估结果可以包括但不限于串行网络的损失值、输出结果的精度值或者推理时长等

等。相应地，该第二预设评估条件可以包括串行网络的损失值不大于第二损失阈值、输出结果的精度值大于第二精度阈值或者推理时长小于第二推理阈值等。对串行网络的训练次数不超过预设训练次数即步骤 604 中通过预设的数据集对串行网络进行训练的次数。

应理解，此处的第二预设评估条件与前述的第一预设评估条件可以相同也可以不相同，
5 当第二预设评估条件与前述的第一预设评估条件不同时，通常第二预设评估条件对应的网络优于第一预设评估条件对应的网络，可以是第二损失阈值低于第一损失阈值，第二精度阈值高于第一精度阈值，或者第二推理阈值低于第一推理阈值等。

606、根据训练后的串行网络得到目标神经网络。

其中，在确定训练后的串行网络满足终止条件之后，即可根据该训练后的串行网络得
10 到目标神经网络。

在一种可能的实施方式中，还可以获取目标任务，该目标神经网络用于执行该目标任务。例如，在步骤 601 之前，可以获取到目标任务，在步骤 601 中，可以根据该目标任务来选择起点网络，后续得到的目标神经网络即可用于执行该目标任务。

具体地，根据训练后的串行网络得到目标神经网络，可以包括：若训练后的串行网络
15 满足输出条件，则将训练后的串行网络作为目标神经网络；若训练后的串行网络不满足输出条件，则获取与训练后的串行网络对应的并行子网络，并根据训练后的串行网络和与训练后的串行网络对应的并行子网络构建目标神经网络。

通常，串行网络包括多个串行的串行子网络，每个串行子网络包括一个或者多个基本单元，上述的获取与训练后的串行网络对应的并行子网络，具体可以包括：在预设的数量
20 范围内确定与该串行网络中的每个串行子网络对应的并行子网络的数量，然后根据每个串行子网络对应的并行子网络的数量构建与每个串行子网络对应的并行子网络。并且，每个串行子网络对应的并行子网络的结构通常与该串行子网络的结构相同。例如，若串行子网络 1 对应的并行子网络的数量为 1，则可以构建该串行子网络 1 对应的 1 个并行子网络，该并行子网络的结构与串行子网络 1 相同，或者是对串行子网络 1 进行变形之后得到。

25 在一种具体的实施方式中，上述的根据串行网络和与每个串行子网络对应的并行子网络构建目标神经网络，可以包括：根据串行网络和每个串行子网络对应的并行子网络构建临时神经网络；对临时神经网络进行评估，得到第三评估结果；若第三评估结果满足输出条件，则将临时神经网络作为目标神经网络；若第三评估结果不满足输出条件，则重新获取与串行网络对应的并行子网络。其中，第三评估结果与前述的第一评估结果或者第二评估结果类似，此处不再赘述。更进一步地，通过构建与串行网络中的每个串行子网络对应的并行子网络来得到目标神经网络的具体流程可以参阅以下步骤 1012-1017 的相关描述，
30 此处不再赘述。

可以理解的是，当串行网络满足输出条件时，可以直接将串行网络作为目标神经网络，从而高效准确地得到目标神经网络。当串行网络满足终止条件，而不满足输出条件时，

35 示例性地，如图 9 所示，串行网络包括串行子网络 1、串行子网络 2、串行子网络 3 和串行子网络 4，串行子网络 1 对应的并行子网络的数量为 2，串行子网络 2 对应的并行子网络的数量为 3，串行子网络 3 对应的并行子网络的数量为 0，串行子网络 4 对应的并行子网

络的数量为 2。在图 9 所示的神经网络中，将串行子网络 1 至串行子网络 4 的传输方向称为行，将串行子网络 1 和与之并列的并行子网络称为列。通常，在构建每个串行子网络对应的并行子网络的结构之后，通常处于同一列的相邻两行之间连接，例如，如图 9 中所示，block30 的输入为 block20 的输出。当某一个并行子网络的同一列中不存在相邻的并行网络时，可以与相邻的串行网络中的串行子网络连接。例如，可以将串行子网络 3 中的 block6 的输出作为 block70 和 block71 的输入，block70 和 block71 为串行子网络 4 对应的并行子网络中的基本单元。

在一种可选的实施方式中，输出条件可以包括但不限于：精度大于第一阈值，或者，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，每秒浮点运算次数不大于第五阈值等等。其中，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。

因此，在本申请实施方式中，可以对起点网络进行至少一次变形得到串行网络，当串行网络满足预设条件时，才通过预设的数据集对串行网络进行训练，而无需每次变形都对变形后的起点网络进行训练，减少了工作量，提高得到目标神经网络的效率。并且，变形后的起点网络可以继承变形前的起点网络的权重值，并通过轻量数据集对继承的权重值进行更新，从而实现对变形后的起点网络的快速训练，快速得到符合需求的串行网络，进而进一步提高得到目标神经网络的效率。此外，本申请提供的神经网络构建的方法，可以适应不同的场景，可以根据不同的需求来搜索串行网络。可以面向更复杂的视觉任务，泛化能力强。从而可以节约构建不同的神经网络的成本，提高构建的神经网络的输出的准确性，提高用户体验。

前述对本申请提供的神经网络构建方法的流程进行了详细介绍，为便于理解，下面示例性地，对本申请提供的神经网络构建方法的完整流程进行更详细的介绍。参阅图 10，本申请提供的另一种神经网络构建方法的流程示意图，如下所述。

需要说明的是，为便于理解，本申请将构建目标神经网络分为多个阶段，包括第一阶段和第二阶段。其中，第一阶段为串行搜索阶段，为得到串行网络的过程，第二阶段为并行搜索阶段，为得到并行子网络并构建目标神经网络的过程。第二阶段为可选阶段，当第一阶段得到的串行网络符合输出条件时，即可将该串行网络作为目标神经网络，具体可以根据实际应场景进行调整，下面分别进行介绍。

第一阶段：串行搜索阶段

1001、获取起点网络。

1002、在第一搜索空间中搜索变形策略。

1003、对起点网络进行变形，得到变形后的起点网络。

1004、通过轻量数据集对变形后的起点网络的权重更新，得到权重更新后的起点网络。

1005、对权重更新后的起点网络进行评估，得到第一评估结果，若符合更新条件，则执行步骤 1003，若不符合更新条件，则执行步骤 1006。

1006、将权重更新后起点网络作为串行网络。

1007、判断串行网络是否满足预设条件，若否，则执行步骤 1002，若是，则执行步骤

1008。

1008、通过预设的数据集对串行网络进行训练，得到训练后的串行网络

1009、判断训练后的串行网络是否满足终止条件，若是，则执行步骤 1010，若否，则执行步骤 1002。

5 1010、判断训练后的串行网络是否满足输出条件，若是，则执行步骤 1011，若否，则执行步骤 1012。

1011、将训练后的串行网络作为所述目标神经网络的主干网络。

其中，步骤 1001-1011 可以参阅前述步骤 601-606 中的相关描述，此处不再赘述。

10 示例性地，以一个场景为例，如图 11 所示，起点网络中包括串行子网络 1、串行子网络 2、串行子网络 3 和串行子网络 4，在第一搜索空间中搜索到变形策略之后，对起点网络所包括的串行子网络进行变形，具体可以包括交换或者扩展。其中，可以改变交换后的基本单元的卷积步长，从而使改变了卷积步长之后的基本单元可以继承变形前的起点网络中的串行子网络的权重值，并且，扩展的基本单元为根据单元矩阵进行分组归一化处理后得到，因此，扩展的基本单元也可以继承对应的串行子网络的权重值。当串行网络，即变形后的起点网络满足预设条件时，即可对串行网络进行进一步训练，即启动重点火过程，进而得到目标神经网络。该预设条件即可理解为对串行网络进行重点火的触发条件。若变形后的起点网络不满足预设条件，则将训练后的串行网络作为新的起点网络，并对新的起点网络进行重新变形，从而使后续可以得到满足点后条件的串行网络。

第二阶段：并行搜索阶段

20 1012、在预设的数量范围内确定与多个串行子网络中每个串行子网络对应的并行子网络的数量。

其中，在得到满足终止条件，但不满足输出条件的串行网络之后，在预设的数量范围内确定与每个串行子网络对应的并行子网络的数量，并根据该数量构建并行子网络。

25 例如，串行网络可以包括 L 个特层层级，并行子网络与其处于同一特征层级的串行子网络的结构相同，若每个特征层级允许存在的最大并行子网络的数量为 K ，并行子网络的结构可以编码为“ $k_1, k_2, \dots, k_L$ ”，可以构建第二搜索空间，第二搜索空间中包括每个串行子网络对应的并行子网络的数量范围，用于搜索确定每个串行子网络对应的并行子网络的数量，第二搜索空间的大小为 K^L 。例如，若每个特征层级允许添加的并行子网络的数量为 3，即 $K=3$ ，特征层级的层数为 10，则第二搜索空间的大小可以是 3^{10} 。并且，也可以使用“,”来分割不同分辨率的特征层级，即表示下采样的位置。通常，相邻的两个特征层级之间，后一特征层级对应的分辨率相较于前一特征层级对应的分辨率减半，当然，后一特征层级对应的分辨率也可以与前一特征层级对应的分辨率相同。

30 此外，在预设的数量范围内搜索每个特征层级对应的并行子网络的数量时，可以采用随机搜索的方式，也可以将预设的资源约束条件作为约束来进行采样。例如，可以以 FLOPS、推理时长、显存占用率或者 CPU 占用率等作为资源约束条件来进行采样。例如，通常每个特征层级添加并行子网络的概率与待添加的串行子网络的 FLOPS 成反比，又或者，每个特征层级添加的并行子网络的数量与推理时长、显存占用率或者 CPU 占用率等成反比等。因

此, 本申请实施方式中, 可以采用随机采样、演化算法或者资源约束条件等多种方式进行采样, 从而可以得到与目标任务的需求对应的并行子网络的数量, 加快了并行子网络的数量搜索进程, 提高得到目标神经网络的效率。

1013、根据每个串行子网络对应的并行子网络的数量构建每个串行子网络对应的并行子网络。

在得到串行网络中每个串行子网络对应的并行子网络的数量之后, 构建每个串行子网络对应的并行子网络。通常, 每个串行子网络对应的并行子网络的结构与对应的串行网络中的串行子网络的结构相同。例如, 若串行网络中某一个串行子网络中包括 block2 和 block3, 对应的并行子网络的数量为 3, 则可以拷贝 3 次 block2 和 block3, 得到 3 个并行子网络。

1014、根据串行网络和每个串行子网络对应的并行子网络构建临时神经网络。

在构建了串行网络中每个串行子网络对应的并行子网络之后, 在串行网络的基础上, 添加并行子网络, 得到临时神经网络。

示例性地, 如图 12 所示, 串行网络可以包括 L 个特征层次, 即 L 个 Stage, 串行网络可以表示为: $SerialNet = s_L(s_{L-1}(...s_1(x)))$, $l = 1, 2, ..., L$, 其中, x 为串行网络的输入数据; L 为串行网络的特征层次, 即串行网络所包括的串行子网络的数量; s_l 表示第 l 特征层次的串行子网络的结构。并行子网络与处于同一特征层次的串行网络中的串行子网络的结构相同, 可以理解为串行网络中的 Stage1 与处于同一层的并行子网络 Stage1 的结构相同, 如与串行子网络 Stage1 对应的并行子网络的结构可以与 Stage1 的结构相同的。构建了每个串行子网络对应的并行子网络之后, 构建每个并行子网络的连接方式。通常, 并行子网络的输入可以是上一个特征层次中处于同一列的并行子网络的输出。示例性地, 可以以 p_l^k 表示第 l 特征层次的第 k 个并行子网络, 通常, p_l^k 的输入为 p_{l-1}^k 的输出, 以此类推。若 p_{l-1}^k 不存在, 则可以将 s_{l-1} 的输出作为 p_l^k 的输入。因此, 本申请实施方式中, 在得到串行网络和每个串行子网络对应的并行子网络之后, 即可构建得到临时神经网络。

1015、对临时神经网络进行评估, 得到第三评估结果。

其中, 在得到临时神经网络之后, 可以对该临时神经网络进行评估, 得到第三评估结果。

具体地, 对临时神经网络进行评估的过程与对起点网络或者串行网络进行评估的过程类似, 此处不再赘述。该第三评估结果也可以包括临时神经网络的输出结果对应的损失值、精度、推理时长、FLOPS、平均精度等等。

1016、判断第三评估结果是否满足输出条件, 若是, 则执行步骤 1017, 若否, 则执行步骤 1012。

在得到临时神经网络的第三评估结果之后, 判断该第三评估结果是否满足输出条件。若该第三评估结果满足输出条件, 则可以将临时神经网络作为目标神经网络或者目标神经网络的主干网络; 若该第三评估结果不满足输出条件, 则可以继续执行在预设的数量范围中搜索每个串行子网络对应的并行子网络的数量, 即重复步骤 1012-1016, 直到得到的临时神经网络的评估结果满足输出条件。

具体地，该输出条件可以参阅前述步骤 606 中的相关介绍，此处不再赘述。

1017、将临时神经网络作为目标神经网络的主干网络。

其中，在确定临时神经网络的输出结果满足输出条件之后，即可将临时神经网络作为目标神经网络的主干网络。

5 通常，主干网络可以完成对图片的目标检测，而目标神经网络可以用于进行图片识别、语义分割或者目标检测中的一种或者多种，因此，在主干网络的基础上，还可以添加其他网络，用于使目标神经网络分别实现不同的功能。例如，可以添加具有不同功能的头(HEAD)，从而使目标神经网络具有不同的功能，具体可以根据实际应用场景进行调整。

10 示例性地，如图 13 所示，在通过串行搜索阶段和并行搜索阶段搜索得到主干网络之后，还在该主干网络的基础上添加 FPN、RPN 和 Head，添加的 FPN、RPN 和 Head 可以根据目标任务的具体需求确定，例如，若需要进行图片识别、语义分割或者目标检测等，则可以添加与进行图片识别、语义分割或者目标检测等中的一个或者多个分别对应的 Head。因此，相对于使用已有的分类网络作为主干网络，本申请实施方式中，可以在主干网络的基础上添加不同的功能性网络来适应不同的场景，从而使得本申请构建的主干网络泛化能力强，
15 在不同的场景中也可以具有较高的精度，提高了构建目标神经网络的效率。

因此，本申请实施方式中，构建目标神经网络的过程可以分为两个阶段，在串行搜索阶段中，可以对起点网络进行至少一次变形得到串行网络，当串行网络满足预设条件时，才通过预设的数据集对串行网络进行训练，而无需每次变形都对变形后的起点网络进行训练，减少了工作量，提高得到目标神经网络的效率。并且，变形后的起点网络可以继承变形前的起点网络的权重值，并通过轻量数据集对继承的权重值进行更新，从而实现对变形后的起点网络的快速训练，快速得到符合需求的串行网络，进而进一步提高得到目标神经网络的效率。在并行搜索阶段中，可以以目标任务对应的资源约束条件为约束搜索串行网络中每个串行子网络对应的并行子网络的数量，使搜索得到的并行子网络所构建得到的目标串行子网络，在满足资源约束的同时，还满足输出条件，得到输出结果更准确的神经网络。
25

为便于理解，下面以一些具体的应用场景对本申请提供的神经网络构建方法所实现的效果进行介绍。

在使用不同的数据集训练的主干网络的场景中，主干网络中的不同特征层次的计算量的占比可能不相同。下面如表 1 所示，示例性地，以不同的场景，如使用 ECP、BDD(Berkeley deep drive)、COCO、VOC(PASCAL VOC)等数据集进行训练的场景，使用不同的基本单元，如瓶颈单元(bottle neck block, BNB)、基础单元(Basic Block, BB)或者瓶颈分组单元(ResNeXt Block, XB)，为例进行说明。
30

数据集 (dataset)	基本单元 类型	目标神经网络结构	
		串行网络	并行子网络
ECP	BNB	111d1111, d1111, 11d, 1111, 1111	0, 2, 1, 1, 0
BDD		111111111d, 1111111111111111, d 1111111111111111, d111111, 111, 1	0, 1, 2, 3, 0, 0
COCO		1111d1, 11111111d, 11, 1111111111 11111d, 11111111111, 1	0, 0, 0, 0, 0, 2
VOC	BB	11d, 1111, d111111111, d111, 1, 111 1	0, 0, 0, 0, 0, 0,
COCO	XB	1111, d1111, d111111111111111111 111111, d11, 1	1, 0, 2, 1, 2

表 1

其中，第一列是不同的数据集的场景，第二列是网络基本单元选择；第三列是串行阶段搜索的串行网络结构，“,”表示不同的分辨率分隔每个特征层次阶段，后一阶段相较于前一阶段分辨率减半；“1”表示不改变输入输出通道的常规块，即输出通道是输入通道的1倍，“d”表示该块中基础通道数量增加一倍，即“2”；最后一列表示搜索到的并行网络结构，“,”表示不同的特征层次阶段，数字表示该特征层次添加的并行子网络个数。

下面对不同的数据集进行简要介绍。

COCO: COCO 2017 数据集包含 80 个目标物体种类，对应约 11.8 万张训练图片与 5 千张测试图片。

BDD: 是一个包含了 10 类目标物体的自动驾驶数据集。其包括了 7 万张训练图片与 1 万张测试图片。

VOC: 包含了 20 类目标物体。其包括了 1 万张训练图片与 4900 张测试图片。

ECP: 是目前最大的行人检测数据集，分为行人和骑行者两类。其包括 2.4 万训练图片与 4300 测试图片。

评价标准可以包括：平均精确度 (average precision, AP)，考虑了不同阈值下的精确度，与不同大小的物体的平均精确度和平均召回率。运行速度 (Inf time)，在同一显卡上对所有方法进行比较，得到在相同推理时间下均精确度更高的网络架构。对数平均漏检率 (log-average miss rate, LAMR)，表示测试集正例判别为负例的数目/测试集检索到想要的正例数加上未检测到不想要的 log 指数，该标准越低越好。

从表 1 中得到的目标神经网络的串行网络和并行子网络的结构可知，在面对目标复杂的 COCO 数据集时，需要更复杂的语义信息以及更大的视野区域来对图片中的目标进行检测。而在行人检测 (Euro City Persons, ECP) 数据集中，存在大量目标较小且类别单一的情况，因此，需要在保留更多的空间信息的特征层次中投入更多的计算量。

此外，在不同数据集中，通过本申请提供的神经网络构建方法搜索得到的主干网络（以下称为 SPNet），在 FLOPS 与检测精度上均大幅度优于目前已有的手工设计（如 ResNext、

HRNet、DetNet)及自动搜索算法(如 DetNAS、NAS-FPN 等)等得到的模型。

并且,在 COCO 数据集上,本申请搜索得到的架构 SPNet (XB)与 SPNet (BNB)对比经典的手工设计的分类主干网络 ResNext101,在提升 7% mAP 情况下,FLOPS 减少将近一半(8.1G vs 15.2G);对比于另一为检测任务手工设计的主干网络 HRNet-w48,本申请搜索得到的架构 E5 在未使用额外目标分割标注的情况下,以相似的运行速度,将 mAP 提升了将近 6%。与目标检测方法(如 AlignDet, NAS-FPN 等)相比,在运行速度与检测性能均有显著优势。在公开数据集 VOC 以及 BDD 上基于 FPN 检测器,本申请搜索到的主干网络比手工设计的 ResNet 网络或者其他自动搜索算法在精度与速度上有显著的性能提升。具体例如,在 VOC 场景中 SPNet 比 DetNAS、Auto-FPN AP50 等方案分别提升 3.2%,2.8%;在 BDD 场景中 SPNet 比 Auto-FPN 运行速度提升,mAP 提升 14%。并且,在公开数据集 VOC 的场景中,与 DetNAS 和 NAS-FPN 相比,本申请可以用更少的时间搜索得到性能更优的架构。并且,本申请搜索得到的网络架构在其他数据集上也有较好的泛化性能(均优于手工设计的主干网络 ResNet101),且在迁移场景越相似(自动驾驶场景 BDD、ECP),迁移效果越好。

前述对本申请提供的方法进行了详细介绍,下面对本申请提供的装置进行介绍。

参阅图 14,本申请提供的一种神经网络构建装置的结构示意图。该神经网络构建装置可以用于执行前述图 6-图 13 中所示的神经网络构建方法的流程。该神经网络构建装置包括:

获取模块 1401,用于获取起点网络,起点网络包括多个串行子网络;

变形模块 1402,用于基于预设的第一搜索空间对起点网络进行至少一次变形,得到串行网络,第一搜索空间包括对起点网络进行变形使用的参数的范围;

训练模块 1403,用于若串行网络满足预设条件,则通过预设的数据集对串行网络进行训练,得到训练后的串行网络;

构建模块 1404,用于若训练后的串行网络满足终止条件,则根据训练后的串行网络得到目标神经网络。

在一种可能的实施方式中,变形模块 1402,具体可以用于:在预设的第一搜索空间中搜索变形策略,变形策略包括对起点网络进行变形使用的一组参数;根据变形策略对起点网络进行变形,得到变形后的起点网络,变形后的起点网络的权重为继承变形前的起点网络的权重得到;通过轻量数据集对变形后的起点网络的权重更新,得到权重更新后的起点网络,预设的数据集所包括的数据多于轻量数据集所包括的数据;对权重更新后的起点网络进行评估,得到第一评估结果;若第一评估结果满足预设的更新条件,则在第一搜索空间中重新搜索变形策略,并根据重新搜索到的变形策略对起点网络进行重新变形;若第一评估结果不满足预设的更新条件,则将权重更新后起点网络作为串行网络。

在一种可能的实施方式中,神经网络构建装置还可以包括:

更新模块 1405,用于若串行网络不满足预设条件,则将串行网络作为新的起点网络,或者,若训练后的串行网络不满足终止条件,则将训练后的串行网络作为新的起点网络,更新第一搜索空间,并根据更新后的第一搜索空间重新搜索变形策略,根据重新搜索到的变形策略对新的起点网络进行再次变形。

在一种可能的实施方式中，更新条件包括：对起点网络的变形次数不超过预设次数，或者，第一评估结果不满足预设的收敛条件，收敛条件包括以下一项或者多项：第一评估结果优于阈值，或者，第一评估结果与上一次对起点网络进行评估得到的评估结果之前的差值小于预设值。

- 5 在一种可能的实施方式中，预设条件可以包括但不限于：串行网络的第二评估结果满足第一预设评估条件，或者，串行网络的变形次数超过预设变形次数，或者，距离上一次对起点网络进行变形的时长超过预设值，第二评估结果为对串行网络进行评估得到的；

 第一预设评估条件包括以下一项或者多项：串行网络的损失值不大于第一损失阈值、串行网络的输出结果的精度值大于第一精度阈值，或者串行网络的推理时长小于第一推理
10 阈值。

在一种可能的实施方式中，起点网络的每个串行子网络由一个或者多个基本单元组成，变形模块 1402，具体可以用于：交换多个串行子网络中相邻两个串行子网络所包括的基本单元；或者，在多个串行子网络中的一个或多个中插入基本单元，插入的基本单元为根据单位矩阵初始化的权重和分组归一化得到。

- 15 在一种可能的实施方式中，终止条件可以包括但不限于：串行网络的第二评估结果满足第二预设评估条件，或者，对串行网络的训练次数超过预设训练次数。

 在一种可能的实施方式中，构建模块 1404，具体可以用于：若训练后的串行网络满足输出条件，则将训练后的串行网络作为目标神经网络；若训练后的串行网络不满足输出条件，则获取与训练后的串行网络对应的并行子网络，并根据训练后的串行网络和与训练后的
20 串行网络对应的并行子网络构建目标神经网络。

在一种可能的实施方式中，串行网络包括多个串行子网络，每个串行子网络包括一个或多个基本单元，构建模块 1404，具体可以用于：在预设的数量范围内确定与多个串行子网络中每个串行子网络对应的并行子网络的数量；根据每个串行子网络对应的并行子网络的数量构建每个串行子网络对应的并行子网络，每个串行子网络的结构和每个串行子网络
25 对应的并行子网络的结构相同。

在一种可能的实施方式中，构建模块 1404，具体可以用于：根据串行网络和每个串行子网络对应的并行子网络构建临时神经网络；对临时神经网络进行评估，得到第三评估结果；若第三评估结果满足输出条件，则将临时神经网络作为目标神经网络；若第三评估结果不满足输出条件，则重新获取与串行网络对应的并行子网络。

- 30 在一种可能的实施方式中，输出条件包括但不限于以下一项或者多项：精度大于第一阈值，或者，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，每秒浮点运算次数不大于第五阈值，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。

 在一种可能的实施方式中，第一搜索空间包括的参数类型可以包括但不限于：宽度、
35 深度、下采样的数量和下采样的位置，宽度为起点网络中的每一个串行子网络所包括的基本单元的输入通道和输出通道的数量，深度为起点网络中的每一个串行子网络所包括的基本单元的数量，下采样用于分割起点网络中两个相邻的串行子网络。

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，预设的数据集可以包括 ImageNet 数据集。

在一种可能的实施方式中，获取起点网络的具体方式可以包括：通过用户输入数据获取到目标任务，然后根据该目标任务来获取起点网络，而构建得到的目标神经网络可以用于执行该目标任务。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 4 或图 8 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 4 或图 8 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 4 至图 8 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 或图 10 中任一实施例所示的神经网络构建装置执行的方法步骤。

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。

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 至图 13 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。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。

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的任何常规的处理器等。

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。

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。

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。

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。

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。在一些实现中, 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;

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。

5 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-图 13 的方法的程序执行的集成电路。

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。

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。

25 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。

权 利 要 求

1. 一种神经网络构建方法，其特征在于，包括：
获取起点网络，所述起点网络包括多个串行子网络；
基于第一搜索空间对所述起点网络进行至少一次变形，得到串行网络，所述第一搜索
5 空间包括对所述起点网络进行变形使用的参数的范围；
若所述串行网络满足预设条件，则通过预设的数据集对所述串行网络进行训练，得到
训练后的串行网络；
若所述训练后的串行网络满足终止条件，则根据所述训练后的串行网络得到目标神经网络。
- 10 2. 根据权利要求1所述的方法，其特征在于，所述基于第一搜索空间对所述起点网络
进行至少一次变形中的其中一次变形，包括：
在所述第一搜索空间中搜索变形策略，所述变形策略包括对所述起点网络进行变形使用
的一组参数；
根据所述变形策略对所述起点网络进行变形，得到变形后的起点网络，所述变形后的
15 起点网络的权重继承变形前的起点网络的权重；
通过轻量数据集对所述变形后的起点网络的权重更新，得到权重更新后的起点网络，
所述预设的数据集所包括的数据多于所述轻量数据集所包括的数据；
对所述权重更新后的起点网络进行评估，得到第一评估结果；
若所述第一评估结果满足预设的更新条件，则在所述第一搜索空间中重新搜索变形策
20 略，并通过重新搜索得到的变形策略对所述起点网络重新进行变形；
若所述第一评估结果不满足所述预设的更新条件，则将所述权重更新后的起点网络作
为所述串行网络。
3. 根据权利要求2所述的方法，其特征在于，所述方法还包括：
若所述串行网络不满足所述预设条件，则将所述串行网络作为新的起点网络，或者，
25 若所述训练后的串行网络不满足所述终止条件，则将所述训练后的串行网络作为新的起点
网络；
更新所述第一搜索空间，并根据更新后的所述第一搜索空间搜索变形策略，根据搜索
到的所述变形策略对所述新的起点网络进行变形。
- 30 4. 根据权利要求2或3所述的方法，其特征在于，所述更新条件包括：对所述起点网
络的变形次数不超过预设次数，或者，所述第一评估结果不满足预设的收敛条件，所述收
敛条件包括以下一项或者多项：所述第一评估结果优于阈值，或者，所述第一评估结果与
上一次对所述起点网络进行评估得到的评估结果之前的差值小于预设值。
5. 根据权利要求2-4中任一项所述的方法，其特征在于，所述预设条件包括：所述串
行网络的第二评估结果满足第一预设评估条件，或者，所述串行网络的变形次数超过预设
35 变形次数，或者，距离上一次对所述起点网络进行变形的时长超过预设值，所述第二评估
结果为对所述串行网络进行评估得到的；
所述第一预设评估条件包括以下一项或者多项：所述串行网络的损失值不大于第一损

失阈值、所述串行网络的输出结果的精度值大于第一精度阈值，或者所述串行网络的推理时长小于第一推理阈值。

6. 根据权利要求 1-5 中任一项所述的方法，其特征在于，所述起点网络中的每个串行子网络由一个或者多个基本单元组成，所述对所述起点网络的变形，包括：

5 交换所述多个串行子网络中相邻两个串行子网络所包括的基本单元；或者，

在所述多个串行子网络中的至少一个串行子网络中插入处理后的基本单元，所述处理后的基本单元为对基本单元的权重进行单位矩阵初始化和/或对基本单元进行分组归一化得到。

7. 根据权利要求 1-6 中任一项所述的方法，其特征在于，所述终止条件包括：

10 所述串行网络的第二评估结果满足第二预设评估条件，或者，对所述串行网络的训练次数超过预设训练次数。

8. 根据权利要求 1-7 中任一项所述的方法，其特征在于，所述根据所述训练后的串行网络得到目标神经网络，包括：

15 若所述训练后的串行网络满足输出条件，则将所述训练后的串行网络作为所述目标神经网络；

若所述训练后的串行网络不满足所述输出条件，则获取与所述训练后的串行网络对应的并行子网络，并根据所述训练后的串行网络和与所述训练后的串行网络对应的并行子网络构建所述目标神经网络。

20 9. 根据权利要求 8 所述的方法，其特征在于，所述串行网络包括多个串行子网络，每个串行子网络包括一个或多个基本单元，

所述获取与所述训练后的串行网络对应的并行子网络，包括：

在预设的数量范围内确定与所述多个串行子网络中每个串行子网络对应的并行子网络的数量；

25 根据所述每个串行子网络对应的并行子网络的数量构建所述每个串行子网络对应的并行子网络，所述每个串行子网络的结构和所述每个串行子网络对应的并行子网络的结构相同。

10. 根据权利要求 9 所述的方法，其特征在于，所述根据所述训练后的串行网络和与所述训练后的串行网络对应的并行子网络构建所述目标神经网络，包括：

30 根据所述训练后的串行网络和所述每个串行子网络对应的并行子网络构建临时神经网络；

对所述临时神经网络进行评估，得到第三评估结果；

若所述第三评估结果满足所述输出条件，则将所述临时神经网络作为所述目标神经网络；

35 若所述第三评估结果不满足所述输出条件，则重新获取与所述每个串行子网络对应的并行子网络。

11. 根据权利要求 8-10 中任一项所述的方法，其特征在于，所述输出条件包括以下一项或者多项：精度大于第一阈值，或者，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，每秒浮点运算次数不大于第五阈值，所述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，所述推理时长为从神经网络根据输入得到输出结果的时长。

12. 根据权利要求 1-11 中任一项所述的方法，其特征在于，所述第一搜索空间包括的参数的类型包括：宽度、深度、下采样的数量和下采样的位置，所述宽度为所述起点网络中的每一个串行子网络所包括的基本单元的输入通道和输出通道的数量，所述深度为所述起点网络中的每一个串行子网络所包括的基本单元的数量，所述下采样用于分割起点网络中两个相邻的串行子网络。

13. 根据权利要求 1-12 中任一项所述的方法，其特征在于，所述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，所述预设的数据集包括 ImageNet 数据集。

14. 一种神经网络构建装置，其特征在于，包括：

获取模块，用于获取起点网络，所述起点网络包括多个串行子网络；

变形模块，用于基于第一搜索空间对所述起点网络进行至少一次变形，得到串行网络，所述第一搜索空间包括对所述起点网络进行变形使用的参数的范围；

训练模块，用于若所述串行网络满足预设条件，则通过预设的数据集对所述串行网络进行训练，得到训练后的串行网络；

构建模块，用于若所述训练后的串行网络满足终止条件，则根据所述训练后的串行网络得到目标神经网络。

15. 根据权利要求 14 所述的神经网络构建装置，其特征在于，所述变形模块，具体用于：

在所述第一搜索空间中搜索变形策略，所述变形策略包括对所述起点网络进行变形使用的一组参数；

根据所述变形策略对所述起点网络进行变形，得到变形后的起点网络，所述变形后的起点网络的权重继承变形前的起点网络的权重；

通过轻量数据集对所述变形后的起点网络的权重更新，得到权重更新后的起点网络，所述预设的数据集所包括的数据多于所述轻量数据集所包括的数据；

对所述权重更新后的起点网络进行评估，得到第一评估结果；

若所述第一评估结果满足预设的更新条件，则在所述第一搜索空间中重新搜索变形策略，并通过重新搜索得到的变形策略对所述起点网络进行重新变形；

若所述第一评估结果不满足所述预设的更新条件，则将所述权重更新后起点网络作为所述串行网络。

16. 根据权利要求 15 所述的神经网络构建装置，其特征在于，所述神经网络构建装置还包括：

更新模块，用于若所述串行网络不满足所述预设条件，则将所述串行网络作为新的起点网络，或者，若所述训练后的串行网络不满足所述终止条件，则将所述训练后的串行网络作为新的起点网络，更新所述第一搜索空间；

所述变形模块，还用于根据更新后的所述第一搜索空间搜索变形策略，根据搜索到的所述变形策略对所述新的起点网络进行变形。

17. 根据权利要求 15 或 16 所述的神经网络构建装置，其特征在于，所述更新条件包括：对所述起点网络的变形次数不超过预设次数，或者，所述第一评估结果不满足预设的收敛条件，所述收敛条件包括以下一项或者多项：所述第一评估结果优于阈值，或者，所述
5 所述第一评估结果与上一次对所述起点网络进行评估得到的评估结果之前的差值小于预设值。

18. 根据权利要求 15-17 中任一项所述的神经网络构建装置，其特征在于，所述预设条件包括：所述串行网络的第二评估结果满足第一预设评估条件，或者，所述串行网络的变形次数超过预设变形次数，或者，距离上一次对所述起点网络进行变形的时长超过预设
10 值，所述第二评估结果为对所述串行网络进行评估得到的；

所述第一预设评估条件包括以下一项或者多项：所述串行网络的损失值不大于第一损失阈值、所述串行网络的输出结果的精度值大于第一精度阈值，或者所述串行网络的推理时长小于第一推理阈值。

19. 根据权利要求 14-18 中任一项所述的神经网络构建装置，其特征在于，所述起点
15 网络中的每个串行子网络由一个或者多个基本单元组成，所述变形模块，具体用于：

交换所述多个串行子网络中相邻两个串行子网络所包括的基本单元；或者，

在所述多个串行子网络中的一个或多个中插入处理后的基本单元，所述处理后的基本单元为对基本单元的权重进行单位矩阵初始化和/或对基本单元进行分组归一化得到。

20. 根据权利要求 14-19 中任一项所述的神经网络构建装置，其特征在于，所述终止
20 条件包括：

所述串行网络的第二评估结果满足第二预设评估条件，或者，对所述串行网络的训练次数超过预设训练次数。

21. 根据权利要求 14-20 中任一项所述的神经网络构建装置，其特征在于，所述构建
模块，具体用于：

25 若所述训练后的串行网络满足输出条件，则将所述训练后的串行网络作为所述目标神经网络；

若所述训练后的串行网络不满足所述输出条件，则获取与所述训练后的串行网络对应的并行子网络，并根据所述训练后的串行网络和与所述训练后的串行网络对应的并行子网络构建所述目标神经网络。

30 22. 根据权利要求 21 所述的神经网络构建装置，其特征在于，所述串行网络包括多个串行子网络，每个串行子网络包括一个或多个基本单元，

所述构建模块，具体用于：

在预设的数量范围内确定与所述多个串行子网络中每个串行子网络对应的并行子网络的数量；

35 根据所述每个串行子网络对应的并行子网络的数量构建所述每个串行子网络对应的并行子网络，所述每个串行子网络的结构和所述每个串行子网络对应的并行子网络的结构相同。

23. 根据权利要求 22 所述的神经网络构建装置, 其特征在于, 所述构建模块, 具体用于:

根据所述训练后的串行网络和所述每个串行子网络对应的并行子网络构建临时神经网络;

5 对所述临时神经网络进行评估, 得到第三评估结果;

若所述第三评估结果满足所述输出条件, 则将所述临时神经网络作为所述目标神经网络;

若所述第三评估结果不满足所述输出条件, 则重新获取与所述串行网络对应的并行子网络。

10 24. 根据权利要求 21-23 中任一项所述的神经网络构建装置, 其特征在于, 所述输出条件包括以下一项或者多项: 精度大于第一阈值, 或者, 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, 每秒浮点运算次数不大于第五阈值, 所述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, 所述推理时长为从神经网络根据输入得到输出结果的时长。

15 25. 根据权利要求 14-24 中任一项所述的神经网络构建装置, 其特征在于, 所述第一搜索空间包括的参数的类型包括: 宽度、深度、下采样的数量和下采样的位置, 所述宽度为所述起点网络中的每一个串行子网络所包括的基本单元的输入通道和输出通道的数量, 所述深度为所述起点网络中的每一个串行子网络所包括的基本单元的数量, 所述下采样用于分割起点网络中两个相邻的串行子网络。

20 26. 根据权利要求 14-25 中任一项所述的神经网络构建装置, 其特征在于, 所述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, 所述预设的数据集包括 ImageNet 数据集。

27. 一种神经网络构建装置, 其特征在于, 包括处理器, 所述处理器和存储器耦合, 所述存储器存储有程序, 当所述存储器存储的程序指令被所述处理器执行时实现权利要求 1 至 13 中任一项所述的方法。

28. 一种计算机可读存储介质, 包括程序, 当其被处理单元所执行时, 执行如权利要求 1 至 13 中任一项所述的方法。

29. 一种神经网络构建装置, 其特征在于, 包括处理单元和通信接口, 所述处理单元通过所述通信接口获取程序指令, 当所述程序指令被所述处理单元执行时实现权利要求 1 至 13 中任一项所述的方法。

30