

神经网络运行方法、装置、电子设备及存储介质

技术领域

5 本公开涉及深度学习技术领域，具体而言，涉及一种神经网络运行方法、装置、电子设备及存储介质。

背景技术

随着技术的发展，大型神经网络已被应用于各种场景下，比如，自动驾驶场景、图像识别场景等。在大型神经网络构建之后，可以通过计算设备运行该大型神经网络。

发明内容

10 有鉴于此，本公开至少提供一种神经网络运行方法、装置、电子设备及存储介质。

第一方面，本公开提供了一种神经网络运行方法，包括：确定目标神经网络中的待处理网络层；从确定的多个算子和多种分块策略中，确定所述目标神经网络中所述待处理网络层对应的目标算子和目标分块策略；所述多个算子中的每个算子均用于实现所述待处理网络层对应的功能，所述多种分块策略中的每个分块策略均匹配用于运行所述目标神经网络的计算设备的运行要求；基于所述待
15 处理网络层对应的所述目标分块策略，运行包含所述目标算子的所述目标神经网络。

上述方法中，在确定了目标神经网络中的待处理网络层之后，可以从确定的多个算子和多种分块策略中，确定待处理网络层对应的目标算子和目标分块策略，由于分块策略满足计算设备的运行要求，使得基于待处理网络层对应的目标分块策略，运行包含目标算子的目标神经网络时，能够满足计算设备的运行要求。同时，由于目标分块策略可以对匹配的待处理网络层对应的目标算子的参数数据进行分块，使得基于分块后的参数数据运行待处理网络层的资源消耗最小，比如该资源消耗可以用总计算开销表征，即在满足计算设备的运行要求的同时，使得基于至少一个待处理网络层分别对应的目标分块策略，运行包含目标算子的目标神经网络的效率较高。

一种可能的实施方式中，所述分块策略用于对所述待处理网络层对应的目标算子的参数数据进行分块；在所述多种分块策略中，基于采用所述目标分块策略对所述目标算子的参数数据进行分块
25 得到的参数数据，运行所述待处理网络层的资源消耗最小。

一种可能的实施方式中，在所述待处理网络层为多个的情况下，所述从确定的多个算子和多种分块策略中，确定所述目标神经网络中待处理网络层对应的目标算子和目标分块策略，包括：针对所述目标神经网络中的每个待处理网络层，从所述多个算子中确定所述待处理网络层对应目标候选算子、并从所述多种分块策略中确定与所述目标候选算子匹配的目标候选分块策略；在存在任一待
30 处理网络层对应的所述目标候选算子为多个和/或所述目标候选分块策略为多个的情况下，基于各个待处理网络层分别对应的目标候选算子和目标候选分块策略，确定每个待处理网络层对应的所述目标算子和所述目标分块策略。

上述实施方式中，可以先分别确定每个待处理网络层对应的目标候选算子、和与目标候选算子匹配的目标候选分块策略，实现了每个待处理网络层的目标候选算子和目标候选分块策略的局部优选。进一步的，在存在任一待处理网络层对应的目标候选算子为多个和/或目标候选分块策略为多个
35 的情况下，基于各个待处理网络层分别对应的目标候选算子和目标候选分块策略，确定每个待处理网络层对应的目标算子和目标分块策略，实现了各个待处理网络层的目标候选算子和目标候选分块策略的全局优选。

一种可能的实施方式中，所述基于各个待处理网络层分别对应的目标候选算子和目标候选分块

策略，确定每个待处理网络层对应的所述目标算子和所述目标分块策略，包括：基于各个待处理网络层分别对应的目标候选算子、和与所述目标候选算子对应的目标候选分块策略，确定所述目标神经网络对应的多个测试网络；其中，每个测试网络中包括各个所述待处理网络层对应的一个所述目标候选算子、和与该目标候选算子匹配的一个目标候选分块策略；分别运行所述多个测试网络，得到多个测试结果，其中，每个测试网络对应一个测试结果；基于所述多个测试结果，从所述多个测试网络中选取目标测试网络；将所述目标测试网络中所述待处理网络层的目标候选算子和目标候选分块策略，确定为所述目标神经网络中所述待处理网络层对应的所述目标算子和目标分块策略。

上述实施方式中，通过基于各个待处理网络层对应的目标候选算子、和与目标候选算子对应的目标候选分块策略，确定目标神经网络对应的多个测试网络；再利用计算设备运行多个测试网络，确定每个测试网络的测试结果；基于测试结果，确定目标测试网络，比如，在测试结果为计算开销时，可以选择计算开销最小的测试网络作为目标测试网络，将目标测试网络中各个待处理网络层的目标候选算子和目标候选分块策略，确定为目标神经网络中各个待处理网络层分别对应的目标算子和目标分块策略，实现了目标算子和目标分块策略的全局优选。

一种可能的实施方式中，针对所述目标神经网络中的每个待处理网络层，从所述多个算子中确定所述待处理网络层对应目标候选算子、并从所述多种分块策略中确定与所述目标候选算子匹配的目标候选分块策略，包括：针对所述待处理网络层，从所述多个算子中，确定一个或多个第一候选算子；基于所述第一候选算子在所述多种分块策略中的每种分块策略下的资源消耗情况，从所述第一候选算子、和所述多种分块策略中，选择所述待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的目标候选分块策略。

这里，可以针对每个待处理网络层，在确定了待处理网络层对应的一个或多个第一候选算子之后，可以基于第一候选算子在多种分块策略中的每种分块策略下的资源消耗情况，从第一候选算子、和多种分块策略中，选择待处理网络层对应的一个或多个目标候选算子、以及与目标候选算子对应的目标候选分块策略，比如，可以选择资源消耗最小的第一候选算子和分块策略，作为目标候选算子和目标候选分块策略，实现了每个待处理网络层对应的目标候选算子和目标候选分块策略的局部优选。

一种可能的实施方式中，所述资源消耗情况用计算开销值表示，根据以下步骤确定所述第一候选算子在所述每种分块策略下的计算开销值：确定所述第一候选算子在预设尺寸下对应的受限场景，其中，所述受限场景为基于在预设尺寸下所述第一候选算子对应的数据容量的计算耗时和传输耗时确定的；在所述受限场景属于带宽受限场景的情况下，基于所述分块策略进行分块的分块结果，确定所述第一候选算子在所述分块策略下对应的直接内存操作 DMA 数据传输总量、DMA 任务个数、和数据转换开销；基于所述 DMA 数据传输总量、所述 DMA 任务个数、所述数据转换开销、和所述计算设备对应的 DMA 速率、DMA 任务开销，确定所述第一候选算子在所述分块策略下的计算开销值；其中，所述数据转换开销为按照所述第一候选算子对应的目标数据排布方式，对所述第一候选算子对应的输入数据进行数据排布方式转换所消耗的时间；在所述受限场景属于计算受限场景的情况下，基于所述分块策略进行分块的分块结果，确定所述第一候选算子在所述分块策略下对应的参数数据的计算耗时、所述第一候选算子的算子调用次数、初始数据传输总量、DMA 任务个数、和数据转换开销；基于所述计算耗时、所述算子调用次数、初始数据传输总量、数据转换开销、DMA 任务开销、DMA 任务个数、和所述计算设备对应的 DMA 速率，确定所述第一候选算子在所述分块策略下的计算开销值。

上述实施方式中，可以确定第一候选算子在预设尺寸下对应的受限场景，不同的受限场景对应不同计算开销值确定方式。比如，在带宽受限场景下，可以基于 DMA 数据传输总量、DMA 任务个

数、数据转换开销、DMA 速率、DMA 任务开销，确定计算开销值；在计算受限场景下，可以基于计算耗时、算子调用次数、初始数据传输总量、数据转换开销、DMA 任务开销、DMA 任务个数、和 DMA 速率，确定计算开销值。

一种可能的实施方式中，所述基于所述第一候选算子在所述多种分块策略中的每种分块策略下的资源消耗情况，从所述第一候选算子、和所述多种分块策略中，选择所述待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的一个或多个目标候选分块策略，包括：从所述第一候选算子对应的多个所述资源消耗情况中，选取满足预设条件的目标资源消耗情况；其中，一个第一候选算子在一种分块策略下对应一个所述资源消耗情况；将所述目标资源消耗情况对应的分块策略，确定为候选分块策略，并基于所述候选分块策略，运行包含所述目标资源消耗情况对应的第二候选算子的待处理网络层，确定与所述候选分块策略及所述第二候选算子对应的测试结果；基于所述测试结果，确定所述待处理网络层对应的一个或多个目标候选算子以及与所述目标候选算子对应的目标候选分块策略。

采用上述方法，可以先利用资源消耗情况，从第一候选算子和多种分块策略中，选择第二候选算子、和与第二候选算子匹配的候选分块策略；并对第二候选算子和候选分块策略进行测试，再利用测试结果，确定待处理网络层对应的至少一个目标候选算子和目标候选分块策略，使得确定的待处理网络层对应的至少一个目标候选算子和目标候选分块策略为较优选择。

一种可能的实施方式中，从所述第一候选算子、和所述多种分块策略中，选择所述待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的目标候选分块策略之前，还包括：基于确定的所述目标神经网络对应的最小颗粒度信息，对所述第一候选算子对应的参数数据进行对齐操作，得到所述第一候选算子对应的对齐后的参数数据；其中，所述最小颗粒度信息中包括所述参数数据在不同维度下对应的最小颗粒度；所述对齐后的参数数据在不同维度下的尺寸，是所述最小颗粒度信息指示的对应维度下的最小颗粒度的整数倍。

这里，可以基于目标神经网络对应的最小颗粒度信息，对每个第一候选算子对应的参数数据进行对齐操作，得到第一候选算子对应的对齐后的参数数据，对齐后的参数数据在不同维度下的尺寸，是最小颗粒度信息指示的对应维度下的最小颗粒度的整数倍，降低后续基于目标分块策略，运行目标神经网络时，造成参数数据丢失的情况的发生概率。

一种可能的实施方式中，在所述参数数据包括输入数据和常数数据的情况下，所述多种分块策略包括以下至少一种：将全部输入数据作为初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果；所述初始数据为计算设备运行所述目标神经网络时，写入直接内存操作 DMA 任务所分配的初始数据区域内的数据；将全部常数数据作为所述初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果；将部分输入数据作为所述初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果；其中，部分输入数据的目标尺寸为根据所述输入数据的第一维度的最小颗粒度确定；将部分常数数据作为所述初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果；其中，部分常数数据的目标尺寸为根据所述常数数据的第一维度的最小颗粒度确定的。

一种可能的实施方式中，所述将部分输入数据作为初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果，包括：基于所述输入数据的第一维度的最小颗粒度的 i 倍，确定所述部分输入数据的目标尺寸；分别将所述目标尺寸的所述部分输入数据作为初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果；其中， i 为在确定了部分输入数据的目标尺寸后，使得所述部分输入数据的数据容量，

以及基于所述常数数据的维度参数的最小颗粒度，确定的常数数据块的数据容量，满足计算设备的内存要求的正整数。

一种可能的实施方式中，所述将部分常数数据作为初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果，包括：基于所述常数数据的第一维度的最小颗粒度的 j 倍，确定所述部分常数数据的目标尺寸；分别将所述目标尺寸的所述部分常数数据作为初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果；其中， j 为在确定了部分常数数据的目标尺寸后，使得所述部分常数数据的数据容量，以及基于所述输入数据的维度参数的最小颗粒度，确定的输入数据块的数据容量，满足计算设备的内存要求的正整数。

10 这里，设置多种分块策略，可以使得每个待处理网络层能够选择较优的目标算子和与目标算子匹配的目标分块策略。

一种可能的实施方式中，在所述指定维度为一维，所述维度参数包括第一维度的情况下，分别将所述常数数据和所述输入数据作为目标数据，基于确定的所述目标数据的第一维度，对所述目标数据进行一维分块，得到分块结果，包括：将所述目标数据的第一维度对应的最小颗粒度的 k 倍，
15 确定为目标分块尺寸，基于所述目标分块尺寸，将所述目标数据按照所述第一维度进行一维分块，得到所述目标数据对应的多个目标数据块；其中， k 为正整数；在确定所述多个目标数据块和所述初始数据满足设置的分块条件的情况下，将所述目标数据的第一维度对应的最小颗粒度的 $k+1$ 倍作为更新后的目标分块尺寸，返回至基于所述目标分块尺寸，将所述目标数据按照所述第一维度进行一维分块的步骤，直至确定所述多个目标数据块和所述初始数据不满足设置的分块条件，将所述目标数据的第一维度对应的最小颗粒度的 k 倍确定为所述分块结果；在所述初始数据、和 k 等于 1 时
20 生成的所述多个目标数据块不满足设置的分块条件的情况下，确定所述分块结果为一维分块失败。

采用上述方法，不断增大目标分块尺寸，通过不断尝试的方式确定使得计算设备的内存使用率较高的分块结果，有利于减少计算设备的内存资源浪费。

一种可能的实施方式中，在所述指定维度为二维，所述维度参数包括第二维度的情况下，分别
25 将所述常数数据和所述输入数据作为目标数据，基于确定的所述目标数据的第一维度、第二维度，对所述目标数据进行二维分块，得到分块结果，包括：将所述目标数据的第一维度对应的最小颗粒度的 y 倍，确定为第一目标分块尺寸，基于所述第一目标分块尺寸，将所述目标数据按照第一维度进行一维分块，得到所述目标数据对应的多个中间数据块；其中， y 为正整数；将所述目标数据的第二维度对应的最小颗粒度的 x 倍，确定为第二目标分块尺寸；基于所述第二目标分块尺寸，将每个
30 中间数据块按照所述第二维度进行二维分块，得到各个中间数据块分别对应的多个目标数据块；其中， x 为正整数；在确定所述多个目标数据块和所述初始数据满足设置的分块条件的情况下，则将所述目标数据的第二维度对应的最小颗粒度的 $x+1$ 倍作为更新后的第二目标分块尺寸，返回至基于所述第二目标分块尺寸，将每个中间数据块按照第二维度进行二维分块的步骤，直至确定所述多个目标数据块和所述初始数据不满足设置的分块条件为止，将所述目标数据的第二维度对应的最小
35 颗粒度的 x 倍确定为所述分块结果。

一种可能的实施方式中，在所述待处理网络层对应的参数数据还包括输出数据的情况下，确定所述多个目标数据块和所述初始数据满足设置的分块条件，包括：在确定所述初始数据、所述输出数据、和每个目标数据块分别满足所述计算设备的内存要求，以及所述初始数据、所述输出数据、和每个目标数据块分别满足所述计算设备中 DMA 传输要求的情况下，确定所述多个目标数据块和
40 所述初始数据满足设置的分块条件。

采用上述方法，在初始数据、输出数据和每个目标数据块，满足计算设备的内存要求和计算设

备中 DMA 传输要求时，确定多个目标数据块和所述初始数据满足设置的分块条件，保证了分块策略与计算设备的运行要求相匹配。

以下装置、电子设备等的效果描述参见上述方法的说明，这里不再赘述。

5 第二方面，本公开提供了一种神经网络运行装置，包括：第一确定模块，用于确定目标神经网络中的待处理网络层；第二确定模块，用于从确定的多个算子和多种分块策略中，确定所述目标神经网络中待处理网络层对应的目标算子和目标分块策略；所述多个算子中的每个算子均用于实现所述待处理网络层对应的功能，所述多种分块策略中的每个分块策略均匹配用于运行所述目标神经网络的计算设备的运行要求；运行模块，用于基于所述待处理网络层分别对应的所述目标分块策略，运行包含所述目标算子的所述目标神经网络。

10 第三方面，本公开提供一种电子设备，包括：处理器、存储器和总线，所述存储器存储有所述处理器可执行的机器可读指令，当电子设备运行时，所述处理器与所述存储器之间通过总线通信，所述机器可读指令被所述处理器执行时执行如上述第一方面或任一实施方式所述的神经网络运行方法的步骤。

15 第四方面，本公开提供一种计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行如上述第一方面或任一实施方式所述的神经网络运行方法的步骤。

第五方面，本公开提供了一种计算机程序，包括计算机可读代码，当所述计算机可读代码在电子设备中运行时，所述电子设备中的处理器执行上述方法。

为使本公开的上述目的、特征和优点能更明显易懂，下文特举实施例，并配合所附附图，作详细说明如下。

20

附图说明

25 为了更清楚地说明本公开实施例的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，此处的附图被并入说明书中并构成本说明书中的一部分，这些附图示出了符合本公开的实施例，并与说明书一起用于说明本公开的技术方案。应当理解，以下附图仅示出了本公开的某些实施例，因此不应被看作是对范围的限定，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他相关的附图。

图 1 示出了本公开实施例所提供的一种神经网络运行方法的流程示意图；

图 2 示出了本公开实施例所提供的一种神经网络运行方法中，确定目标神经网络中待处理网络层对应的目标算子和目标分块策略的流程示意图；

30 图 3 示出了本公开实施例所提供的一种神经网络运行方法中，从多个算子中确定待处理网络层对应目标候选算子、并从多种分块策略中确定与目标候选算子匹配的目标候选分块策略的流程示意图；

图 4 示出了本公开实施例所提供的一种神经网络运行方法中，计算设备的软件硬件调度的示意图；

35 图 5 示出了本公开实施例所提供的一种神经网络运行装置的架构示意图；

图 6 示出了本公开实施例所提供的一种电子设备的结构示意图。

具体实施方式

为使本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本公开一部分实施例，而不是全部的实施例。通常在此处附图中描述和示出的本公开实施例的组件可以以各种不同的配置来布置和设计。因此，以下对在附图中提供的本公开的实施例的详细描述并非旨在限制要求保护的本公开的范围，而是仅仅表示本公开的选定实施例。基于本公开的实施例，本领域技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本公开保护的范围。

一般的，对于依赖直接内存操作（Direct Memory Access, DMA）进行数据传输的计算设备，该计算设备的数据缓冲 Data cache 效率不高或者无 Data cache，故在使用该计算设备推理大型神经网络时，由于计算设备的内存有限，将可能遇到对大型神经网络的单层任务进行分块(tile)、调度等问题。

具体实现时，在计算设备运行该大型神经网络时，可以使用计算设备的生成厂商设置的官方推理库，在计算设备上运行大型神经网络，但是该官方推理库是针对特定的基础神经网络的，而在用户对基础神经网络进行优化后，可能会造成官方推理库无法使用，或者，可能会造成计算设备利用该官方推理库运行优化后的神经网络的效率较低。其中，官方推理库是一种可用的推理部署方案，比如，官方推理库可以为 ceva dsp 的 cdnn 库。

进一步的，针对优化后的神经网络，可以对官方推理库进行二次开发，以使得开发后的推理库能够适用于优化后的神经网络，但是开发过程成本较高、效率较低，且开发后的推理库仅适用于该优化后的神经网络，不适用于其他神经网络，使得开发后的推理库的复用率较低。

因此，为了解决上述问题，本公开实施例提供了一种神经网络运行方法、装置、电子设备及存储介质。

针对以上方案所存在的缺陷，均是发明人在经过实践并仔细研究后得出的结果，因此，上述问题的发现过程以及下文中本公开针对上述问题所提出的解决方案，都应该是发明人在本公开过程中对本公开做出的贡献。

应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步定义和解释。

为便于对本公开实施例进行理解，首先对本公开实施例所公开的一种神经网络运行方法进行详细介绍。本公开实施例所提供的神经网络运行方法的执行主体一般为具有一定计算能力的计算机设备，该计算机设备可以为运行神经网络的计算设备，也可以为其他计算设备，该计算机设备例如包括：终端设备或服务器或其它处理设备，终端设备可以为用户设备（User Equipment, UE）、移动设备、用户终端、终端、蜂窝电话、无绳电话、个人数字助理（Personal Digital Assistant, PDA）、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，该神经网络运行方法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。

参见图 1 所示，为本公开实施例所提供的神经网络运行方法的流程示意图，该方法包括 S101-S103，其中：

S101，确定目标神经网络中的待处理网络层。

S102，从确定的多个算子和多种分块策略中，确定目标神经网络中待处理网络层对应的目标算子和目标分块策略。

其中，多个算子中的每个算子均用于实现待处理网络层对应的功能，多种分块策略中的每个分块策略均匹配用于运行目标神经网络的计算设备的运行要求。

S103，基于待处理网络层对应的目标分块策略，运行包含目标算子的目标神经网络。

上述方法中，在确定了目标神经网络中的待处理网络层之后，可以从确定的多个算子和多种分

块策略中，确定待处理网络层对应的目标算子和目标分块策略，由于分块策略满足计算设备的运行要求，使得基于待处理网络层对应的目标分块策略，运行包含目标算子的目标神经网络时，能够满足计算设备的运行要求。同时，由于目标分块策略可以对匹配的待处理网络层对应的目标算子的参数数据进行分块，使得基于分块后的参数数据运行待处理网络层的资源消耗最小，比如该资源消耗可以用总计算开销表征，即在满足计算设备的运行要求的同时，使得基于至少一个待处理网络层分别对应的目标分块策略，运行包含目标算子的目标神经网络的效率较高。

在一种可能的实施方式中，目标神经网络用于实现图像处理任务，所述图像处理任务包括：图像识别、图像分类、图像分割、关键点检测中的至少一种。

下述对 S101-S103 进行具体说明。

针对 S101：这里，目标神经网络可以为进行图层次的优化处理（即图优化处理）后的任一神经网络，图优化处理后的神经网络为计算图已确定的神经网络，即图优化处理后的神经网络为每一网络层的任务和参数尺寸已确定的神经网络，每一网络层的参数尺寸可以为该网络层中包括的参数数据的尺寸。比如，目标神经网络的第一网络层的任务可以为特征提取，在第一网络层包括输入数据时，输入数据的参数尺寸可以为 $256 \times 256 \times 128$ 。其中，每一网络层的任务和参数尺寸可以根据实际情况进行设置，此处仅为示例性说明。

待处理网络层可以为目标神经网络上的待处理的网络操作（operator，OP）层。比如，待处理网络层可以为目标神经网络中尺寸大于设定阈值的网络 OP 层；或者，也可以为用户根据需要选择的网络 OP 层。确定的待处理网络层的数量可以为一个或多个。

示例性的，每个网络 OP 层可以近似为卷积层，比如，全连接层可以近似为特征图与卷积核尺寸一致的卷积层，不存在权值的常规层可以使为权值为 0 的卷积层等。

针对 S102：这里，在待处理网络层的数量为多个时，可以为每个待处理网络层，确定该待处理网络层对应的目标算子和目标分块策略。其中，多种分块策略中的每个分块策略均匹配用于运行目标神经网络的计算设备的运行要求；多个算子中的每个算子均用于实现待处理网络层对应的功能，每个算子可以对应于一种操作或者一个基本网络结构单元。比如，预设的算子例如包括：卷积算子、池化算子、全连接算子等。计算设备为直接处理目标神经网络推理计算的设备，比如，计算设备可以为数字信号处理（Digital Signal Processing，DSP）等。

一种可能的实施方式中，分块策略用于对待处理网络层对应的目标算子的参数数据进行分块；在多种分块策略中，基于采用目标分块策略对目标算子的参数数据进行分块得到的参数数据，运行待处理网络层的资源消耗最小。

这里，资源消耗最小可以指运行待处理网络层的运行耗时最小。具体实施时，每个待处理网络层的目标分块策略用于对该待处理网络层对应的目标算子的参数数据进行分块，以使得计算设备基于分块后的参数数据，运行各个待处理网络层的资源消耗最小，比如该资源消耗可以用总计算开销表征，即运行各个待处理网络层的总计算开销最小。其中，算子的参数数据可以包括输入输出数据和常数数据，输入输出数据可以包括输入数据和输出数据，常数数据可以包括权值数据和/或偏差数据。

示例性的，输入数据可以为三维数据，比如可以包括宽度维度、高度维度、输入通道维度；输出数据可以为三维数据，比如可以包括输出宽度维度、输出高度维度、和输出通道维度；权值数据可以为四维数据，比如可以包括宽度维度、高度维度、输入通道维度、输出通道维度；偏差数据可以为一维数据，比如可以包括输出通道维度。其中，上述输入数据、输出数据、权值数据和偏差数据的维度信息可以根据实际情况进行设置，此处仅为示例性说明。

具体实施时，在待处理网络层为多个的时，可以针对目标神经网络中的各个待处理网络层，按

照目标神经网络中待处理网络层的顺序，逐层的确定每个待处理网络层的目标算子和目标候选分块策略；或者，也可以随机确定各个待处理网络层中每个待处理网络层的目标算子和目标候选分块策略。比如，在需要确定当前的待处理网络层的输入数据的数据排布是否与设置的目标数据排布方式一致时，则需要用到当前的待处理网络层之前的待处理网络层的输出数据，此时，需要逐层确定每个待处理网络层对应的目标候选算子和目标候选分块策略。

一种可选实施方式中，参见图 2 所示，在所述待处理网络层为多个的情况下，从确定的多个算子和多种分块策略中，确定目标神经网络中待处理网络层对应的目标算子和目标分块策略，包括：

S201，针对目标神经网络中的每个待处理网络层，从多个算子中确定待处理网络层对应目标候选算子、并从多种分块策略中确定与目标候选算子匹配的目标候选分块策略。

S202，在存在任一待处理网络层对应的目标候选算子为多个和/或目标候选分块策略为多个的情况下，基于各个待处理网络层分别对应的目标候选算子和目标候选分块策略，确定每个待处理网络层对应的目标算子和目标分块策略。

这里，可以先确定每个待处理网络层对应的目标候选算子和与该目标候选算子匹配的目标候选分块策略，实现对每个待处理网络层的目标候选算子和目标候选分块策略的局部优选；进一步的，在存在任一待处理网络层对应的目标候选算子为多个和/或目标候选分块策略为多个的情况下，在综合考虑各个待处理网络层，确定目标神经网络中各个待处理网络层分别对应的目标算子和目标分块策略，实现各个待处理网络层的目标算子和目标分块策略的全局优选。

这里，可以预先设置有算子集合和分块策略集合，算子集合中包括设置的全部算子、以及分块策略中包括设置的全部分块策略。为了提高待处理网络层的目标算子和目标分块策略的确定效率，针对每个待处理网络层，可以从算子集合和分块策略集合中，确定该待处理网络层对应的多种算子和多种分块策略。不同的待处理网络层对应的多个算子可以相同、也可以不同；以及不同的待处理网络层对应的多种分块策略可以相同、也可以不同。其中，可以根据实际情况确定每个待处理网络层对应的多个算子和多种分块策略。

示例性的，可以基于历史经验数据，确定每个待处理网络层对应的多个算子和/或多种分块策略。比如，基于历史经验数据，可以确定待处理网络层一对应的多个算子包括算子一、算子二、算子三，对应的多种分块策略包括分块策略一、分块策略二、分块策略四；待处理网络层二对应的多个算子包括算子一、算子三、算子四、算子五，对应的多种分块策略包括分块策略二、分块策略五。

在 S201 中，一种可选实施方式中，参见图 3 所示，针对目标神经网络中的每个待处理网络层，从多个算子中确定待处理网络层对应目标候选算子、并从多种分块策略中确定与目标候选算子匹配的目标候选分块策略，包括：

S301，针对待处理网络层，从多个算子中，确定一个或多个第一候选算子。

S302，基于第一候选算子在多种分块策略中的每种分块策略下的资源消耗情况，从第一候选算子、和多种分块策略中，选择待处理网络层对应的一个或多个目标候选算子、以及与目标候选算子对应的目标候选分块策略。

这里，可以针对每个待处理网络层，在确定了待处理网络层对应的一个或多个第一候选算子之后，可以基于第一候选算子在每种分块策略下的资源消耗情况，从至少一个第一候选算子、和多种分块策略中，选择待处理网络层对应的一个或多个目标候选算子、以及与目标候选算子对应的目标候选分块策略，比如，可以选择资源消耗最小的第一候选算子和分块策略，作为目标候选算子和目标候选分块策略，实现了每个待处理网络层对应的目标候选算子和目标候选分块策略的局部优选。

针对 S301，针对目标神经网络中的每个待处理网络层，可以从多个算子中，确定该待处理网络层对应的一个或多个第一候选算子。比如，可以根据每个待处理网络层的任务，从多个算子中选择

能够完成该任务的算子，作为该待处理网络层对应的第一候选算子；或者，也可以为用户基于目标神经网络的需求，确定待处理网络层对应的第一候选算子。

针对 S302，可以先确定每个第一候选算子在每种分块策略下的资源消耗情况，再基于第一候选算子在多种分块策略中的每种分块策略下的资源消耗情况，从至少一个候选算子和多个分块策略中，
5 确定该待处理网络层对应的一个或多个目标候选算子、以及与目标候选算子对应的目标候选分块策略。其中，资源消耗情况为计算设备基于分块策略，运行第一候选算子时资源消耗数据，比如，该资源消耗情况可以用计算开销值表征，该计算开销值表征计算设备运行包含目标算子的待处理网络层时所消耗的时间。

示例一，若待处理网络层一对应的第一候选算子包括第一候选算子一、第一候选算子二，待处理网络层对应的分块策略包括分块策略一、分块策略二、分块策略三，则针对第一候选算子一，可以计算得到分块策略一对应的计算开销值、分块策略二对应的计算开销值、分块策略一对应的计算开销值，针对第一候选算子二，可以计算得到分块策略一对应的计算开销值、分块策略二对应的计算开销值、分块策略一对应的计算开销值，则可以基于计算得到的 6 个计算开销值，确定待处理网络层一对应的目标候选算子和目标候选分块策略。
10

一种可选实施方式中，在得到每个第一候选算子对应的多个计算开销值之后，可以直接利用计算开销值，确定待处理网络层对应的至少一个目标候选算子和目标候选分块策略。
15

比如，可以通过下述两种方式，利用计算开销值，确定待处理网络层对应的至少一个目标候选算子和目标候选分块策略：

方式一中，从计算得到的各个计算开销值中，选择开销值最小的第一候选算子和分块策略，作为待处理网络层对应的目标候选算子和目标候选分块策略。
20

承接上述示例一继续说明，在得到 6 个计算开销值之后，选择最小的计算开销值，比如，若第一候选算子一在分块策略一下的计算开销值最小，则将第一候选算子一确定为待处理网络层一对应的目标候选算子、将分块策略一确定为待处理网络层一对应的目标候选分块策略。

方式二中，可以设置开销阈值，从计算得到的待处理网络层对应的多个计算开销值中，选择计算开销值小于开销阈值的候选开销值，将候选开销值对应的第一候选算子确定为待处理网络层对应的目标候选算子、以及将候选开销值对应的分块策略确定为与目标候选算子匹配的目标分块策略。
25

承接上述示例一继续说明，在得到 6 个计算开销值之后，若第一候选算子一在分块策略一下的计算开销值小于设置的开销阈值、以及第一候选算子二在分块策略三下的计算开销值小于设置的开销阈值，则将第一候选算子一确定为待处理网络层一对应的一个目标候选算子、将分块策略一确定为与第一候选算子一匹配的目标候选分块策略；以及将第一候选算子二确定待处理网络层一对应的一个目标候选算子、将分块策略三确定为与第一候选算子三匹配的目标候选分块策略，即确定了待处理网络层一对应的目标候选算子和目标候选分块策略。
30

另一种实施方式中，S302 中，基于第一候选算子在多种分块策略中的每种分块策略下的资源消耗情况，从第一候选算子、和多种分块策略中，选择待处理网络层对应的一个或多个目标候选算子、
35 以及与目标候选算子对应的一个或多个目标候选分块策略，包括：

步骤一、从第一候选算子对应的多个所述资源消耗情况中，选取满足预设条件的目标资源消耗情况；其中，一个第一候选算子在一种分块策略下对应一个资源消耗情况。

步骤二、将目标资源消耗情况对应的分块策略，确定为候选分块策略，并基于候选分块策略，运行包含目标资源消耗情况对应的第二候选算子的待处理网络层，确定与候选分块策略及第二候选算子对应的测试结果。
40

步骤三、基于测试结果，确定待处理网络层对应的一个或多个目标候选算子以及与目标候选算

子对应的目标候选分块策略。

采用上述方法，可以先利用资源消耗情况，从第一候选算子和多种分块策略中，选择第二候选算子、和与第二候选算子匹配的候选分块策略；并对第二候选算子和候选分块策略进行测试，再利用测试结果，确定待处理网络层对应的至少一个目标候选算子和目标候选分块策略，使得确定的待

5 处理网络层对应的至少一个目标候选算子和目标候选分块策略为较优选择。

在步骤一中，一个第一候选算子在一种分块策略下对应一个资源消耗情况。比如，第一候选算子一对应的分块策略包括 4 种时，则第一候选算子对应了 4 个资源消耗情况。

下述以资源消耗情况用计算开销值表征为例进行说明，预设条件可以根据实际需要进行设置。比如，预设条件可以为开销值最小；和/或，预设条件可以为小于设置的开销阈值；和/或，预设条件还可以为选取最小开销值、和与最小开销值之间的差值小于设置的差值阈值的次小开销值。

比如，若计算得到的第一候选算子在设置的多种分块策略下的计算开销值包括：计算开销值一为 10、计算开销值二为 12、计算开销值三为 18、计算开销值四为 20，则可以从多个计算开销值中选择最小开销值，即将计算开销值一确定为目标开销值；或者，可以设置开销阈值 15，将计算开销值一和计算开销值二确定为目标开销值；或者，还可以设置差值阈值 5，（可知计算开销值二与计算

15 开销值一之间的差值小于设置的差值阈值），将计算开销值一和计算开销值二确定为目标开销值，即该目标开销值对应第二候选算子、以及与第二候选算子匹配的候选分块策略。

在步骤二中，可以对目标开销值（即目标资源消耗情况）对应的第二候选算子和候选分块策略进行实测，得到各个目标开销值对应的测试结果。即针对每个目标开销值，可以将目标资源消耗情况对应的分块策略，确定为候选分块策略，并基于候选分块策略，运行包含该目标开销值对应的第二候选算子的待处理网络层，确定该目标开销值对应的测试结果，即确定与候选分块策略和第二候选算子对应的测试结果。

20

在步骤三中，可以基于测试结果，确定待处理网络层对应的一个或多个目标候选算子、以及与目标候选算子对应的目标候选分块策略。比如，在测试结果为运行时间时，可以选择运行时间最短对应的第二候选算子，确定为该待处理网络层的目标候选算子，将运行时间最短的第二候选算子对应的候选分块策略，确定为目标候选算子。其中，第一候选算子、第二候选算子可以为能够实现待处理网络层功能的算子。

25

或者，还可以设置运行时间阈值，将小于运行时间阈值的测试结果确定为目标测试结果，将目标测试结果对应的第二候选算子确定为目标候选算子，将目标测试结果中目标候选算子对应的候选分块策略，确定为目标候选分块策略。

30 在一种可选实施方式中，资源消耗情况用计算开销值表示，可以根据以下步骤确定第一候选算子在每种分块策略下的计算开销值：

步骤一、确定第一候选算子在预设尺寸下对应的受限场景，其中，受限场景为基于在预设尺寸下所述第一候选算子对应的数据容量的计算耗时和传输耗时确定的；

步骤二、在受限场景属于带宽受限场景的情况下，基于分块策略进行分块的分块结果，确定第一候选算子在分块策略下对应的直接内存操作 DMA 数据传输总量、DMA 任务个数、和数据转换开销；基于 DMA 数据传输总量、DMA 任务个数、数据转换开销、和计算设备对应的 DMA 速率、DMA 任务开销，确定第一候选算子在分块策略下的计算开销值；其中，数据转换开销为按照第一候选算子对应的目标数据排布方式，对第一候选算子对应的输入数据进行数据排布方式转换所消耗的时间；

35

步骤三、在受限场景属于计算受限场景的情况下，基于分块策略进行分块的分块结果，确定第一候选算子在分块策略下对应的参数数据的计算耗时、第一候选算子的算子调用次数、初始数据传输总量、DMA 任务个数、和数据转换开销；基于计算耗时、算子调用次数、初始数据传输总量、数

40

据转换开销、DMA 任务开销、DMA 任务个数、和计算设备对应的 DMA 速率，确定第一候选算子在分块策略下的计算开销值。

在步骤一中，可以确定每个第一候选算子在预设尺寸下对应的受限场景。其中，预设尺寸可以根据需要设置的、较大的尺寸。

- 5 在具体实施时，每个待处理网络层对应的目标算子的参数数据可以存储在除计算设备的内存之外的其他外部存储器中，比如，可以存储在双倍速率同步动态随机存储器（Double Data Rate, DDR）中，再运行每个待处理网络层时，DMA 可以从 DDR 中获取该待处理网络层对应的目标算子的参数数据（比如输入数据、常数数据等），并将获取到的参数数据传输到计算设备的内存中，计算设备在计算完成后，DMA 再将数据结果（即输出数据）传输到 DDR 中，以便目标神经网络中与该待处理网络层相邻的下一网络层（该网络层可以为待处理网络层）进行使用。其中，DMA 可以使用乒乓调度策略，对获取到的参数数据进行传输。

- 10 由此可知，DMA 在传输预设尺寸下的第一候选算子的参数数据时存在一个传输耗时，以及计算设备在对预设尺寸下的第一候选算子的参数数据进行处理时存在一个计算耗时。进而，若计算耗时大于传输耗时，表征在 DMA 将当前参数数据传输至计算设备的内存上时，计算设备对上一参数数据的处理过程没有结束，此时计算设备需要进行等待，在计算设备对上一参数数据的处理过程结束后，将当前参数数据传输至计算设备的内存上，这种情况对应的场景为计算受限场景；若计算耗时小于传输耗时，表征在计算设备将上一参数数据处理完成后，DMA 还未将当前参数数据传输至计算设备的内存上，DMA 需要进行等待，直至接收到 DMA 传输的当前参数数据，这种情况对应的场景可以为带宽受限场景。

- 20 进而，在受限场景为带宽受限场景，可以使用带宽受限场景对应的第一开销函数计算开销值；在受限场景为计算受限场景，可以使用计算受限场景对应的第二开销函数计算开销值。

- 25 示例性，可以根据下述过程确定第一候选算子在预设尺寸下对应的受限场景：针对在预设尺寸下的第一候选算子，确定传输预设尺寸下的第一候选算子对应的参数数据需要的传输耗时，以及确定计算设备对预设尺寸下的第一候选算子的参数数据进行处理的处理耗时，根据传输耗时和计算耗时的大小，确定第一候选算子对应的受限场景。

- 30 示例性的，还可以根据下述过程确定第一候选算子在预设尺寸下的受限场景：第一，基于预设尺寸信息，确定计算设备基于第一候选算子对应的参数数据运行对应的待处理网络层所需的目标耗时、以及确定第一候选算子对应的参数数据的目标数据容量。第二，基于计算设备对应的 DMA 速率、和目标耗时，确定目标耗时内 DMA 可传输的数据容量。第三，基于目标耗时内 DMA 可传输的数据容量、与目标数据容量的比值，确定受限场景，即在比值小于或等于 1 时，确定受限场景为带宽受限场景；在比值大于 1 时，确定受限场景为计算受限场景。

- 35 这里，目标耗时内 DMA 可传输的数据容量与传输速度相关，目标数据容量与计算速度相关，在比值大于 1 时，表征传输速度大于计算速度（即传输耗时小于计算耗时），即为计算受限场景；在比值小于或等于 1 时，表征传输速度小于或等于计算速度（即传输耗时大于或等于计算耗时），即为带宽受限场景，进而后续针对不同的受限场景，可以选择不同的方式确定计算开销值。

- 40 可以基于第一候选算子的参数数据的预设尺寸信息，确定第一候选算子的参数数据在计算设备上对应的目标耗时，即确定计算设备在基于第一候选算子对应的参数数据运行对应的待处理网络时，所需的目标耗时。进而可以将计算设备对应的 DMA 速率和目标耗时相乘，得到目标耗时内 DMA 可传输的数据容量。

- 40 同时，可以基于第一候选算子的参数数据的预设尺寸信息，确定第一候选算子对应的参数数据的目标数据容量。比如，在第一候选算子为卷积算子时，该目标数据容量可以为常数数据（包括权

值数据和偏差数据)、输出数据和输入数据的总和。再可以基于计算的目标耗耗时 DMA 可传输的数据容量、与目标数据容量的比值, 确定受限场景。

具体实施时, 在确定了计算设备之后, 可以确定该计算设备对应的 DMA 任务开销, 单位为秒(s), 比如, 可以将每创建一个 DMA 任务需要消耗的周期 cycle, 转换为时间, 即得到 DMA 任务开销; 以及可以确定 DMA 速率, 即 DMA 的传输速率, 单位为 Byte/s。

在步骤二中, 可以使用第一开销函数确定第一候选算子在分块策略下的计算开销值。第一开销函数可以为: $\text{计算开销} = \text{DMA 数据传输总量} / \text{DMA 速率} + \text{DMA 任务个数} \times \text{DMA 任务开销} + \text{数据转换开销}$ 。

即在确定属于带宽受限场景时, 可以基于分块结果, 确定第一候选算子在分块策略下对应的 DMA 数据传输总量 (单位为 Byte), DMA 任务个数、和数据转换开销 (单位为秒 s)。其中, DMA 数据传输总量可以根据生成的 DMA 任务进行确定; DMA 任务个数可以在基于分块策略对参数数据进行分块后, 基于得到的参数数据的数据块的数量进行确定; 比如, 在一个数据块对应一个 DMA 任务、生成的数据块的数量为 10 个时, 则确定 DMA 的任务有 10 个。这里, DMA 数据传输总量、和 DMA 任务个数可以根据实际情况进行确定, 此处仅为示例性说明。比如, 在第一候选算子为卷积算子时, 可以根据卷积算子对应的卷积核尺寸、卷积步长等卷积参数, 确定分块结果后得到的 DMA 任务个数。

数据转换开销为按照第一候选算子对应的目标数据排布方式, 对第一候选算子对应的输入数据进行数据排布方式转换所消耗的时间。这里, 第一算子的输入数据的数据排布、与第一算子对应的目标数据排布方式一致时, 数据转换开销为 0; 在第一算子的输入数据的数据排布、与第一算子对应的目标数据排布方式不一致时, 可以根据下述公式计算数据转换开销: $\text{数据转换开销} = \text{输入数据的总数据容量} \times 2 / \text{DMA 速率}$ 。其中, 输入数据的总数据容量为分块之前、输入至待处理网络层的全部输入数据。

在步骤三中, 在确定属于计算受限场景时, 可以根据第二开销函数计算第一候选算子在分块策略下的计算开销值。第二开销函数为: $\text{计算开销} = \text{算子开销折算带宽} \times \text{算子调用次数} / \text{DMA 速率} + \text{初始数据传输总量} / \text{DMA 速率} + \text{DMA 任务个数} \times \text{DMA 任务开销} + \text{数据转换开销}$ 。

其中, 算子开销折算宽度为基于第一候选算子在预设尺寸下的计算耗时、和第一候选算子在分块策略下对应的参数数据的尺寸, 确定的算子传输数据量。比如, 在预设尺寸为 $1024 \times 1024 \times 128$ 、第一候选算子在预设尺寸下的计算耗时为 10 毫秒, 分块后的参数数据的尺寸为 $512 \times 512 \times 64$, 则第一候选算子在所述分块策略下对应的参数数据的计算耗时为 1.25 毫秒。再可以基于确定的计算速度、和第一候选算子在所述分块策略下对应的参数数据的计算耗时 (比如 1.25 毫秒), 确定分块后的第一候选算子对应的算子开销折算带宽。

具体的, 可以基于分块结果, 确定第一候选算子在分块策略下对应的参数数据的计算耗时、第一候选算子的算子调用次数、初始数据传输总量、DMA 任务个数、和数据转换开销。其中, 算子调用次数可以在基于分块策略对参数数据进行分块后, 基于得到的参数数据的数据块的数量进行确定; 比如, 若得到的参数数据的数据块的数量为 10 个, 则确定算子调用次数为 10 次; 初始数据传输总量为基于分块策略确定的初始数据的数据容量; 其中, 目标数据容量、算子调用次数、初始数据传输总量, 可以根据实际情况进行确定。

其中, 步骤二和步骤三中, 基于分块结果, 可以得到第一候选算子的对齐后的参数数据对应的目标数据容量、算子调用次数、初始数据传输总量、DMA 任务个数、和数据转换开销。

步骤三中的数据转换开销的确定过程、与步骤二中的数据转换开销的确定过程相同, 此处不再进行具体说明。本公开实施例中主要可以应用于带宽受限场景中, 即在满足带宽受限场景时, 使用

步骤二确定计算开销值；在不满足带宽受限场景时（即满足计算开销时），可以利用步骤三确定计算开销值。

上述实施方式中，可以确定第一候选算子在预设尺寸下对应的受限场景，不同的受限场景对应不同计算开销值确定方式。比如，在带宽受限场景下，可以基于 DMA 数据传输总量、DMA 任务个数、数据转换开销、DMA 速率、DMA 任务开销，确定计算开销值；在计算受限场景下，可以基于计算耗时、算子调用次数、初始数据传输总量、数据转换开销、和 DMA 速率，确定计算开销值。

在一种可选实施方式中，从第一候选算子、和多种分块策略中，选择待处理网络层对应的一个或多个目标候选算子、以及与目标候选算子对应的目标候选分块策略之前，该方法还包括：

基于确定的目标神经网络对应的最小颗粒度信息，对第一候选算子对应的参数数据进行对齐操作，得到第一候选算子对应的对齐后的参数数据；其中，最小颗粒度信息中包括参数数据在不同维度下对应的最小颗粒度；对齐后的参数数据在不同维度下的尺寸，是最小颗粒度信息指示的对应维度下的最小颗粒度的整数倍。

这里，最小颗粒度信息中包括参数数据在不同维度下对应的最小颗粒度，比如，参数数据包括权值数据时，权值数据对应的最小颗粒度信息包括宽度维度上的最小颗粒度、长度维度上的最小颗粒度、输入通道维度上的最小颗粒度、和输出通道维度上的最小颗粒度。其中，最小颗粒度信息可以根据计算设备的运行需求、和/或用户需求进行确定，此处仅为示例性说明。

可以利用确定的目标神经网络对应的最小颗粒度信息，对每个第一候选算子对应的参数数据进行对齐操作，得到第一候选算子对应的对齐后的参数数据，使得对齐后的参数数据在不同维度下的尺寸，是最小颗粒度信息指示的对应维度下的最小颗粒度的整数倍。比如，若最小颗粒度信息指示的宽度维度的尺寸为 32，在参数数据指示的宽度维度的尺寸为 33 时，则生成的对齐后的参数数据的宽度维度的尺寸为 64；在参数数据指示的宽度维度的尺寸为 31 时，则生成的对齐后的参数数据的宽度维度的尺寸为 32。

其中，对齐操作的具体过程可以根据实际需要进行选择。比如，可以使用常规的数据对齐方式（比如 padding 方式等），对参数数据进行对齐操作，生成对齐后的参数数据。

在另一种实施方式中，还可以计算设备还可以从 DDR 中获取对齐前的参数数据，使用 garbage 数据计算，然后在计算设备输出的数据中选取有效数据，将有效数据作为输出数据输入到 DDR 中。

这里，可以基于目标神经网络对应的最小颗粒度信息，对每个第一候选算子对应的参数数据进行对齐操作，得到第一候选算子对应的对齐后的参数数据，对齐后的参数数据在不同维度下的尺寸，是最小颗粒度信息指示的对应维度下的最小颗粒度的整数倍，降低后续基于目标分块策略，运行目标神经网络时，造成参数数据丢失的情况的发生概率。

针对 S202：一种可选实施方式中，S202 中，基于各个待处理网络层分别对应的目标候选算子和目标候选分块策略，确定每个待处理网络层对应的目标算子和目标分块策略，包括：

S2021，基于各个待处理网络层分别对应的目标候选算子、和与目标候选算子对应的目标候选分块策略，确定目标神经网络对应的多个测试网络；其中，每个测试网络中包括各个待处理网络层对应的一个目标候选算子、和与该目标候选算子匹配的一个目标候选分块策略。

S2022，分别运行多个测试网络，得到多个测试结果，其中，每个测试网络对应一个测试结果。

S2023，基于多个测试结果，从多个测试网络中选取目标测试网络。

S2024，将目标测试网络中待处理网络层的目标候选算子和目标候选分块策略，确定为目标神经网络中待处理网络层分别对应的目标算子和目标分块策略。

在 S2021 中，示例性的，若目标神经网络包括第一待处理网络层、第二待处理网络层和第三待处理网络层，第一待处理网络层中包括目标候选算子一、目标候选算子一对应的分块策略一，以及

目标候选算子二、目标候选算子二对应的分块策略二；第二待处理网络层中包括目标候选算子三、目标候选算子对应的分块策略一，以及目标候选算子四、目标候选算子四对应的分块策略一；第三待处理网络层中包括目标候选算子五、目标候选算子五对应的分块策略三。

进而可以得到目标神经网络对应的四个测试网络，第一测试网络中包括：目标候选算子一、目标候选算子一对应的分块策略一，目标候选算子三、目标候选算子对应的分块策略一，目标候选算子五、目标候选算子五对应的分块策略三。第二测试网络中包括：目标候选算子一、目标候选算子一对应的分块策略一，目标候选算子四、目标候选算子四对应的分块策略一，目标候选算子五、目标候选算子五对应的分块策略三。第三测试网络中包括：目标候选算子二、目标候选算子二对应的分块策略二，目标候选算子三、目标候选算子对应的分块策略一，目标候选算子五、目标候选算子五对应的分块策略三。第四测试网络中包括：目标候选算子二、目标候选算子二对应的分块策略二，目标候选算子四、目标候选算子四对应的分块策略一，目标候选算子五、目标候选算子五对应的分块策略三。

在 S2022 和 S2023 中，可以控制计算设备分别运行多个测试网络，确定每个测试网络的测试结果。比如，测试结果可以为每个测试网络对应的运行时间。再可以基于多个测试网络对应的测试结果，从多个测试网络中选取目标测试网络。比如，可以选取运行时间最短的测试网络，作为目标测试网络。

在 S2024 中，可以将目标测试网络中包括的各个待处理网络层的目标候选算子和目标候选分块策略，确定为目标神经网络中各个待处理网络层分别对应的目标算子和目标分块策略。

比如，若确定第二测试网络为目标测试网络，则确定目标候选算子一为第一待处理网络层的目标算子，分块策略一为第一待处理网络层对应的目标分块策略；目标候选算子四为第二待处理网络层的目标算子，分块策略一为第二待处理网络层的目标候选分块策略；目标候选算子五为第三代理网络层的目标算子，分块策略三为第三待处理网络层的目标候选分块策略。

为了减少运行测试网络所消耗的成本和运算资源，提高目标算子和目标分块策略的确定效率，在具体实施时，可以设置每个待处理网络层对应的匹配有目标分块策略的目标算子的最大数量。比如，设置的最大数量为 2 时，则每个待处理网络层可以包括一个匹配有目标分块策略的目标算子，比如，匹配有目标分块策略一的目标算子一；或者，每个待处理网络层可以包括两个匹配有目标分块策略的目标算子，比如，两个匹配有目标分块策略的目标算子可以为：匹配有目标分块策略一的目标算子一、匹配有目标分块策略二的目标算子一；或者，两个匹配有目标分块策略的目标算子可以为：匹配有目标分块策略一的目标算子一、匹配有目标分块策略二的目标算子二；再或者，两个匹配有目标分块策略的目标算子可以为：匹配有目标分块策略一的目标算子一、匹配有目标分块策略一的目标算子二等。

和/或，在具体实施时，可以设置目标神经网络对应的测试网络的数量阈值。比如，设置的数量阈值为 100，待处理网络层包括 10 层，若第一层待处理网络层至第六层待处理网络层对应的匹配有目标分块策略的目标算子的数量均为 2，则在第一层待处理网络层至第六待处理网络层中，基于每个待处理网络层对应的目标算子和目标分块策略，构成的局部测试网络的数量可以为 $2^6=64$ 个。进而，在确定第七层待处理网络层的目标算子和目标分块策略时，若第七层待处理网络层对应的匹配有目标分块策略的目标算子的数量为 2 个时，则在第一层待处理网络层至第七待处理网络层中，基于每个待处理网络层对应的目标算子和目标分块策略，构成的局部测试网络的数量可以为 $2^7=128$ 个，大于设置的数量阈值；在这种情况下，第七层待处理网络层、第八层待处理网络层、第九层待处理网络层、以及第十层待处理网络层中，每层待处理网络层对应的匹配有目标分块策略的目标算子的数量仅可以为 1 个。

上述实施方式中，通过基于各个待处理网络层对应的至少一个目标候选算子、和与目标候选算子对应的目标候选分块策略，确定目标神经网络对应的多个测试网络；再利用计算设备运行多个测试网络，确定每个测试网络的测试结果；基于测试结果，确定目标测试网络，比如，在测试结果为计算开销时，可以选择计算开销最小的测试网络作为目标测试网络，将目标测试网络中各个待处理网络层的目标候选算子和目标候选分块策略，确定为目标神经网络中各个待处理网络层分别对应的目标算子和目标分块策略，实现了目标算子和目标分块策略的全局优选。

一种可选实施方式中，在指定维度为一维时，维度参数为第一维度；在指定维度为N维时，维度参数包括第一维度至第N维度，N为大于2且小于常数数据或输入数据的维度。在参数数据包括输入数据和常数数据的情况下，多种分块策略包括以下至少一种：

10 方案一、将全部输入数据作为初始数据，基于确定的常数数据的第一维度，对常数数据进行一维分块，得到分块结果；初始数据为计算设备运行目标神经网络时，写入直接内存操作 DMA 任务所分配的初始数据区域内的数据。

方案二、将全部输入数据作为初始数据，基于确定的常数数据的第一维度、和第二维度，对常数数据进行二维分块，得到分块结果。

15 方案三、将全部常数数据作为初始数据，基于确定的输入数据的第一维度，对输入数据进行一维分块，得到分块结果。

方案四、将全部常数数据作为初始数据，基于确定的输入数据的第一维度和第二维度，对输入数据进行二维分块，得到分块结果。

20 方案五、将部分输入数据作为初始数据，基于确定的常数数据的第一维度，对常数数据进行一维分块，得到分块结果；其中，部分输入数据的目标尺寸为根据输入数据的第一维度的最小颗粒度确定。

方案六、将部分输入数据作为初始数据，基于确定的常数数据的第一维度和第二维度，对常数数据进行二维分块，得到分块结果；其中，部分输入数据的目标尺寸为根据输入数据的第一维度的最小颗粒度确定。

25 方案七、将部分常数数据作为初始数据，基于确定的输入数据的第一维度，对输入数据进行一维分块，得到分块结果；其中，部分常数数据的目标尺寸为根据常数数据的第一维度的最小颗粒度确定的。

30 方案八、将部分常数数据作为初始数据，基于确定的输入数据的第一维度和第二维度，对输入数据进行二维分块，得到分块结果；其中，部分常数数据的目标尺寸为根据常数数据的第一维度的最小颗粒度确定的。

这里，可以将全部的输入数据作为初始数据，将初始数据申请空间 `allocate` 在初始数据区域内。再基于确定的常数数据的第一维度，对常数数据进行一维分块，得到分块结果。或者，基于确定的常数数据的第一维度和第二维度，对常数数据进行二维分块，得到分块结果。

35 可以将全部的常数数据作为初始数据，基于确定的输入数据的第一维度，对输入数据进行一维分块，得到分块结果。或者，基于确定的输入数据的第一维度和第二维度，对输入数据进行二维分块，得到分块结果。

还可以将部分的输入数据作为初始数据，基于确定的输入数据的第一维度，对输入数据进行一维分块，得到分块结果；或者，基于确定的输入数据的第一维度和第二维度，对输入数据进行二维分块，得到分块结果。

40 一种可选实施方式中，方案五或方案六中，将部分输入数据作为初始数据，基于确定的常数数据的维度参数，对常数数据进行指定维度的分块，得到分块结果，包括：

一、基于输入数据的第一维度的最小颗粒度的 i 倍，确定部分输入数据的目标尺寸。

二、分别将目标尺寸的部分输入数据作为初始数据，基于确定的常数数据的维度参数，对常数数据进行指定维度的分块，得到分块结果。

其中， i 为在确定了部分输入数据的目标尺寸后，使得部分输入数据的数据容量，以及基于常数数据的维度参数的最小颗粒度，确定的常数数据块的数据容量，满足计算设备的内存要求的正整数。

这里，可以通过递增的方式，确定 i 的最大值。下述以方案五为例进行说明（即一维分块为例进行说明）， i 从 1 开始进行递增，即在 $i=1$ 时，部分输入数据的目标尺寸为输入数据的第一维度的最小颗粒度的 1 倍，将目标尺寸的部分输入数据作为初始数据，基于确定的常数数据的第一维度对常数数据进行一维分块，得到一维分块结果。

在 $i=1$ 对应的一维分块结果指示常数数据 `allocate` 失败时，则该方案五为不可用的方案；在 $i=1$ 对应的一维分块结果指示常数数据 `allocate` 成功时，则 i 的值加 1（得到 $i=2$ ），并返回至确定部分输入数据的目标尺寸的步骤，即目标尺寸为输入数据的第一维度的最小颗粒度的 2 倍，将目标尺寸的部分输入数据作为初始数据，基于确定的常数数据的第一维度对常数数据进行一维分块，得到一维分块结果。在 $i=2$ 对应的一维分块结果指示常数数据 `allocate` 失败时，则确定 i 的最大值为 1，递增过程截止；在一维分块结果指示常数数据 `allocate` 成功时，则 i 的值加 1（此时 $i=3$ ），再次返回至确定部分输入数据的目标尺寸的步骤，直至一维分块结果指示常数数据 `allocate` 失败为止。比如，若在 $i=6$ 时，确定一维分块结果指示常数数据 `allocate` 失败，则确定 i 的最大值为 5。在 i 的最大值为 5 时，则此方案可以得到 5 个分块结果。

其中，分块结果指示常数数据 `allocate` 失败可以为常数数据按照第一维度的最小颗粒度进行划分后，得到的常数数据块和初始数据不满足计算设备的内存要求。若调度策略为乒乓调度，在按照第一维度的最小颗粒度划分后的常数数据块的数据容量的 2 倍大于计算设备的调度区域的内存时，则输入数据 `allocate` 失败。

比如在 i 的最大值为 5 时，方案五中，可以包括以下 5 种分块策略：

方式一、将输入数据的第一维度的最小颗粒度的 1 倍确定为部分输入数据的目标尺寸，将部分输入数据作为初始数据，基于确定的常数数据的第一维度，对常数数据进行一维分块，得到一维分块结果；

方式二、将输入数据的第一维度的最小颗粒度的 2 倍确定为部分输入数据的目标尺寸，将部分输入数据作为初始数据，基于确定的常数数据的第一维度，对常数数据进行一维分块，得到一维分块结果；

.....

方式五、将输入数据的第一维度的最小颗粒度的 5 倍确定为部分输入数据的目标尺寸，将部分输入数据作为初始数据，基于确定的常数数据的第一维度，对常数数据进行一维分块，得到一维分块结果。

还可以将部分常数数据作为初始数据，基于确定的输入数据的第一维度，对输入数据进行一维分块。

一种可选实施方式中，方案七或方案八中，将部分常数数据作为初始数据，基于确定的输入数据的维度参数，对输入数据进行指定维度分块，得到分块结果，包括：

一、基于常数数据的第一维度的最小颗粒度的 j 倍，确定部分常数数据的目标尺寸；

二、分别将目标尺寸的部分常数数据作为初始数据，基于确定的输入数据的维度参数，对输入数据进行指定维度的分块，得到分块结果。

这里，可以通过递增的方式，确定 j 的最大值。下述以对方案七为例进行说明， j 从 1 开始进行

递增,即在 $j=1$ 时,部分输入数据的目标尺寸为输入数据的第一维度的最小颗粒度的 1 倍,将目标尺寸的部分输入数据作为初始数据,基于确定的常数数据的第一维度对常数数据进行一维分块,得到一维分块结果。

- 在 $j=1$ 对应的一维分块结果指示输入数据 `allocate` 失败时,则该方案七为不可用的方案;在 $j=1$ 对应的一维分块结果指示输入数据 `allocate` 成功时,则 j 的值加 1 (得到 $j=2$),并返回至确定部分输入数据的目标尺寸的步骤,直至得到的一维分块结果指示输入数据 `allocate` 失败为止。比如,若在 $j=6$ 时,确定得到的一维分块结果指示输入数据 `allocate` 失败,则确定 j 的最大值为 5。在 j 的最大值为 5 时,则此方案可以得到 5 个分块结果。

- 其中,分块结果指示输入数据 `allocate` 失败可以为输入数据按照第一维度的最小颗粒度进行划分后,得到的输入数据块和初始数据不满足计算设备的内存要求。若调度策略为乒乓调度,在按照第一维度的最小颗粒度划分后的输入数据块的数据容量的 2 倍大于计算设备的调度区域的内存时,则输入数据 `allocate` 失败。比如,若初始数据、调度数据乒乓 (按照第一维度的最小颗粒度划分后的输入数据块)、和调度数据乒乓 (按照第一维度的最小颗粒度划分后的输入数据块) 不满足计算设备的内存要求时,则确定输入数据 `allocate` 失败。

- 比如在 j 的最大值为 6 时,方案七中,可以包括以下 6 种分块策略:

方式一、将常数数据的第一维度的最小颗粒度的 1 倍确定为部分常数数据的目标尺寸,将部分常数数据作为初始数据,基于确定的输入数据的第一维度,对输入数据进行一维分块,得到一维分块结果;

- 方式二、将常数数据的第一维度的最小颗粒度的 2 倍确定为部分常数数据的目标尺寸,将部分常数数据作为初始数据,基于确定的输入数据的第一维度,对输入数据进行一维分块,得到一维分块结果;

.....

- 方式六、将常数数据的第一维度的最小颗粒度的 6 倍确定为部分常数数据的目标尺寸,将部分常数数据作为初始数据,基于确定的输入数据的第一维度,对输入数据进行一维分块,得到一维分块结果。

这里,对输入数据进行切块的第一维度、第二维度,可以根据运行需求和/或算子类型等信息进行设置;以及对常数数据进行切块的第一维度、第二维度,可以根据运行需求和/或算子类型等信息进行设置。比如,若算子为卷积算子时,常数数据的第一维度可以为输出通道 (output channel, 后续简称 OC) 维度,第二维度可以为输入通道 (input channel, 后续简称 IC) 维度。

- 这里,设置多种分块策略,可以使得每个待处理网络层能够选择较优的目标算子和与目标算子匹配的目标分块策略。

在一种可选实施方式中,在指定维度为一维,维度参数包括第一维度的情况下,分别将常数数据和输入数据作为目标数据,基于确定的目标数据的第一维度,对目标数据进行一维分块,得到一维分块结果,包括:

- A1,将目标数据的第一维度对应的最小颗粒度的 k 倍,确定为目标分块尺寸,基于目标分块尺寸,将目标数据按照第一维度进行一维分块,得到目标数据对应的多个目标数据块;其中, k 为正整数;

- A2,在确定多个目标数据块和初始数据满足设置的分块条件的情况下,将目标数据的第一维度对应的最小颗粒度的 $k+1$ 倍作为更新后的目标分块尺寸,返回至基于目标分块尺寸,将目标数据按照第一维度进行一维分块的步骤,直至确定多个目标数据块和初始数据不满足设置的分块条件,将目标数据的第一维度对应的最小颗粒度的 k 倍确定为分块结果。

A3, 在初始数据、和 k 等于 1 时生成的多个目标数据块不满足设置的分块条件的情况下, 确定分块结果为一维分块失败。

采用上述方法, 不断增大目标分块尺寸, 通过不断尝试的方式确定使得计算设备的内存使用率较高的分块结果, 有利于降低计算设备的内存资源浪费。

- 5 在步骤 A1 中, k 为正整数。从 $k=1$ 开始, 将目标数据的第一维度对应的最小颗粒度, 确定为目标分块尺寸, 按照目标分块尺寸, 将目标数据按照第一维度进行一维分块, 得到目标数据对应的多个目标数据块。得到的每个目标数据块的第一维度的尺寸与目标分块尺寸一致, 每个目标数据块的除第一维度之外的其他维度的尺寸与目标数据的对应维度的尺寸一致。

- 10 比如, 若第一维度的最小颗粒度为 32, 目标数据的尺寸信息为 $64 \times 64 \times 128$, 则目标分块尺寸为 32, 按照目标分块尺寸, 将目标数据按照第一维度进行一维分块, 得到多个目标数据块, 每个目标数据块的尺寸可以为 $32 \times 64 \times 128$ 。其中, 目标数据块的数量可以根据实际情况进行确定。

其中, 第一维度可以根据需要进行设置, 比如, 输入数据的第一维度可以为宽度 W 维度, 第二维度可以为输入通道 IC 维度; 常数数据的第一维度可以为输出通道 OC 维度, 第二维度可以为输入通道 IC 维度。

- 15 进而, 可以判断多个目标数据块和初始数据是否满足设置的分块条件, 若满足, 则将目标数据的第一维度对应的最小颗粒度的 2 倍作为更新后的目标分块尺寸, 返回至按照目标分块尺寸, 将目标数据按照第一维度进行一维分块的步骤, 直至确定多个目标数据块和初始数据不满足设置的分块条件, 并将目标数据的第一维度对应的最小颗粒度的 k 倍确定为分块结果。比如, 在 $k=5$ 时, 确定 $k=5$ 时生成的多个目标数据块和初始数据不满足设置的分块条件, 将目标数据的第一维度对应的最小颗粒度的 4 倍确定为分块结果。即在运行待处理网络层时, 可以将第一维度的最小颗粒度的 4 倍,
- 20 作为目标分块尺寸, 按照目标分块尺寸对待处理网络层的目标算子对应的目标数据进行一维分块。

若不满足 (即在 $k=1$ 时生成的多个目标数据块和初始数据不满足设置的分块条件), 则确定分块结果为一维分块失败。

- 25 一种可选实施方式中, 在指定维度为二维, 维度参数包括第二维度的情况下, 分别将常数数据和输入数据作为目标数据, 基于确定的目标数据的第一维度、第二维度, 对目标数据进行二维分块, 得到分块结果, 包括:

B1, 将目标数据的第一维度对应的最小颗粒度的 y 倍, 确定为第一目标分块尺寸, 基于第一目标分块尺寸, 将所述目标数据按照第一维度进行一维分块, 得到目标数据对应的多个中间数据块; 其中, y 为正整数;

- 30 B2, 将目标数据的第二维度对应的最小颗粒度的 x 倍, 确定为第二目标分块尺寸; 基于第二目标分块尺寸, 将每个中间数据块按照第二维度进行二维分块, 得到各个中间数据块分别对应的多个目标数据块; 其中, x 为正整数;

- 35 B3, 在确定多个目标数据块和初始数据满足设置的分块条件的情况下, 则将目标数据的第二维度对应的最小颗粒度的 $x+1$ 倍作为更新后的第二目标分块尺寸, 返回至基于第二目标分块尺寸, 将每个中间数据块按照第二维度进行二维分块的步骤, 直至确定多个目标数据块和初始数据不满足设置的分块条件为止, 将目标数据的第二维度对应的最小颗粒度的 x 倍确定为分块结果。

在 B1 中, y 为初始值为 1 的正整数, 比如在设置的 y 的最大值为 3 时, 则可以将 y 确定为 1, 执行步骤 B1~B3, 得到一个二维分块结果; 再将 y 确定为 2, 执行步骤 B1~B3, 得到一个二维分块结果; y 确定为 3, 执行步骤 B1~B3, 得到一个二维分块结果, 即可以得到 3 个二维分块结果。

- 40 以 $y=1$ 为例对二维分块过程进行说明, 若第一维度对应的最小颗粒度为 32, 目标数据的尺寸为 $128 \times 128 \times 256$, 则可以基于第一目标分块尺寸, 将目标数据按照第一维度进行一维分块, 得到目标

数据对应的多个目标中间数据块，每个目标中间数据块的尺寸可以为 $32 \times 128 \times 256$ 。其中，目标中间数据块的数量可以根据实际情况进行确定。

在 B2 中，承接 B1 中的示例继续说明， x 为正整数，从 $x=1$ 开始，将目标数据的第二维度对应的最小颗粒度的 1 倍，确定为第二目标分块尺寸，比如，若第二维度的最小颗粒度为 32，则第二目标分块尺寸为 32，基于第二目标分块尺寸，将每个中间数据块按照第二维度进行二维分块，得到各个中间数据块对应的多个目标数据块，即得到多个目标数据块，每个目标数据块的尺寸可以为 $32 \times 32 \times 256$ 。

在 B3 中，可以判断多个目标数据块和初始数据是否满足设置的分块条件，若满足，则将目标数据的第二维度对应的最小颗粒度的 2（即 $x+1$ ）倍作为更新后的第二目标分块尺寸，返回至基于第二目标分块尺寸，将每个中间数据块按照第二维度进行二维分块的步骤，直至确定多个目标数据块和初始数据不满足设置的分块条件为止，将目标数据的第二维度对应的最小颗粒度的 x 倍确定为分块结果。

比如，在 $x=3$ 时，确定 $x=3$ 时生成的多个目标数据块和初始数据不满足设置的分块条件，将目标数据的第二维度对应的最小颗粒度的 2 倍确定为分块结果。即在运行待处理网络层时，可以将第一维度的最小颗粒度，作为第一目标分块尺寸，将第二维度的最小颗粒度的 2 倍，作为第二目标分块尺寸，基于第一目标分块尺寸和第二目标分块尺寸，对待处理网络层的目标算子对应的目标数据进行二维分块。

一种可选实施方式中，在待处理网络层对应的参数数据还包括输出数据的情况下，确定多个目标数据块和初始数据满足设置的分块条件，包括：在确定初始数据、输出数据、和每个目标数据块分别满足计算设备的内存要求，以及初始数据、输出数据、和每个目标数据块分别满足计算设备中 DMA 传输要求的情况下，确定多个目标数据块和初始数据满足设置的分块条件。

这里，计算设备的内存要求可以根据用户需求和/或计算设备需求进行设置。比如，可以确定初始数据、输出数据、和每个目标数据块的数据容量总和，是否小于或等于设置的计算设备的内存容量，若是，则确定满足计算设备的内存要求。

或者，还可以确定初始数据的数据容量是否小于或等于在计算设备的内存上为初始数据分配的第一局部内存容量，确定输出数据的数据容量是否小于或等于在计算设备的内存上为输出数据分配的第二局部内存容量，以及确定每个目标数据块的数据容量是否小于或等于在计算设备的内存上为目标数据分配的三局部内存容量，若初始数据、输出数据和每个目标数据块均满足要求，则确定满足计算设备的内存要求。

在具体实施时，还可以设置专用内存和公共内存，若设置常数数据存储于公共内存中，输入数据和输出数据存储于专用内存上，则可以判断初始数据、输出数据、和每个目标数据块是否均满足对应的专用内存和公共内存的内存要求，若是，则确定满足计算设备的内存要求。即在初始数据为输入数据，目标数据块为常数数据对应的目标数据块，则判断初始数据和输出数据的数据容量是否小于或等于设置的专用内存的内存容量，以及判断每个目标数据块是否小于或等于设置的公共内存的内存容量，若均满足，则确定满足计算设备的内存要求。

示例性的，可以确定了每个目标数据块之后，对目标数据块、初始数据、输出数据进行 `allocate` 尝试，若 `allocate` 尝试成功，则确定初始数据、输出数据和每个目标数据块满足计算设备的内存要求。

其中，DMA 传输要求可以根据实际需要进行确定。比如，若确定初始数据、输出数据和每个目标数据块的数据容量总和，小于或等于 DMA 可传输的数据容量，即在确定 DMA 任务成功建立时，则确定满足计算设备中 DMA 传输要求。

在确定初始数据、输出数据和每个目标数据块满足计算设备的内存要求，以及满足计算设备中

DMA 传输要求时, 确定多个目标数据块和初始数据满足设置的分块条件。

采用上述方法, 在初始数据、输出数据和每个目标数据块, 满足计算设备的内存要求和计算设备中 DMA 传输要求时, 确定多个目标数据块和所述初始数据满足设置的分块条件, 保证了分块策略与计算设备的运行要求相匹配。

- 5 针对 S103: 在确定了目标神经网络每个待处理网络层对应的目标算子和目标分块策略之后, 可以基于至少一个待处理网络层分别对应的目标分块策略, 运行包含目标算子的目标神经网络。

10 比如, 可以将待处理图像输入至目标神经网络中, 计算设备利用每个待处理网络层分别对应的目标分块策略和目标算子, 对待处理图像进行特征提取, 确定待处理图像对应的检测结果, 比如, 该检测结果可以为待处理图像中包括的目标对象的类别、目标对象的位置信息、目标对象的轮廓信息等。

示例性的, 参见图 4 所示的一种神经网络运行方法中, 计算设备的软件硬件调度的示意图, 结合图 4 对使用乒乓调度对待处理网络层的参数数据的处理过程进行说明, 其中, 可以将计算设备的内存划分为初始数据区、调度数据区兵、调度数据区兵、输出数据区兵、输出数据区兵。在初始数据为输入数据时, 调度数据为常数数据; 在初始数据为常数数据时, 调度数据为输入数据。

- 15 由图 4 可知, 计算设备与 DMA 在并行运行, DMA 先将初始数据和调度兵 (即调度数据兵) 传输至计算设备的对应内存区域内 (即将初始数据传输至计算设备的初始数据区对应的内存区域、将调度数据兵输出至计算设备的调度数据区兵对应的内存区域中); 计算设备对初始数据和调度兵进行处理, 同时, DMA 还可以将调度兵 (即调度数据兵) 传输至计算设备的调度数据兵对应的内存区域。

20 在计算设备对初始数据和调度兵处理结束后, 生成输出兵 (即输出数据兵), 将输出兵放置在计算设备的输出数据区兵对应的内存区域, 并通过 DMA 从计算设备的输出数据区兵对应的内存区域中获取输出兵, 再将输出兵传输至对应外部内存 (比如 DDR)。进而计算设备对接收到的调度兵进行处理, 同时 DMA 再将下一调度兵传输至计算设备的调度兵对应的内存区域, 重复上述过程, 直至待处理层的参数数据处理完成为止。

- 25 本领域技术人员可以理解, 在具体实施方式的上述方法中, 各步骤的撰写顺序并不意味着严格的执行顺序而对实施过程构成任何限定, 各步骤的具体执行顺序应当以其功能和可能的内在逻辑确定。

基于相同的构思, 本公开实施例还提供了一种神经网络运行装置, 参见图 5 所示, 为本公开实施例提供的神经网络运行装置的架构示意图, 包括第一确定模块 501、第二确定模块 502、运行模块 503, 具体的:

- 30 第一确定模块 501, 用于确定目标神经网络中的待处理网络层;

第二确定模块 502, 用于从确定的多个算子和多种分块策略中, 确定所述目标神经网络中待处理网络层对应的目标算子和目标分块策略; 其中, 所述多个算子中的每个算子均用于实现所述待处理网络层对应的功能, 所述多种分块策略中的每个分块策略均匹配用于运行所述目标神经网络的计算设备的运行要求;

- 35 运行模块 503, 用于基于所述待处理网络层对应的所述目标分块策略, 运行包含所述目标算子的所述目标神经网络。

一种可能的实施方式中, 所述分块策略用于对所述待处理网络层对应的目标算子的参数数据进行分块;

- 40 在所述多种分块策略中, 基于采用所述目标分块策略对所述目标算子的参数数据进行分块得到的参数数据, 运行所述待处理网络层的资源消耗最小。

一种可能的实施方式中, 在所述待处理网络层为多个的情况下, 所述第二确定模块 502, 在从

确定的多个算子和多种分块策略中，确定所述目标神经网络中待处理网络层对应的目标算子和目标分块策略时，用于：

针对所述目标神经网络中的每个待处理网络层，从所述多个算子中确定所述待处理网络层对应目标候选算子、并从所述多种分块策略中确定与所述目标候选算子匹配的目标候选分块策略；

- 5 在存在任一待处理网络层对应的所述目标候选算子为多个和/或所述目标候选分块策略为多个的情况下，基于各个待处理网络层分别对应的目标候选算子和目标候选分块策略，确定每个待处理网络层对应的所述目标算子和所述目标分块策略。

- 一种可能的实施方式中，所述第二确定模块 502，在基于各个待处理网络层分别对应的目标候选算子和目标候选分块策略，确定每个待处理网络层对应的所述目标算子和所述目标分块策略时，
10 用于：

基于各个待处理网络层分别对应的目标候选算子、和与所述目标候选算子对应的目标候选分块策略，确定所述目标神经网络对应的多个测试网络；其中，每个测试网络中包括各个所述待处理网络层对应的一个所述目标候选算子、和与该目标候选算子匹配的一个目标候选分块策略；

分别运行所述多个测试网络，得到多个测试结果，其中，每个测试网络对应一个测试结果；

- 15 基于所述多个测试结果，从所述多个测试网络中选取目标测试网络；

将所述目标测试网络中待处理网络层的目标候选算子和目标候选分块策略，确定为所述目标神经网络中待处理网络层分别对应的所述目标算子和目标分块策略。

- 一种可能的实施方式中，第二确定模块 502，在针对所述目标神经网络中的每个待处理网络层，从所述多个算子中确定所述待处理网络层对应目标候选算子、并从所述多种分块策略中确定与所述目标候选算子匹配的目标候选分块策略时，用于：
20

针对所述待处理网络层，从所述多个算子中，确定一个或多个第一候选算子；

基于所述第一候选算子在所述多种分块策略中的每种分块策略下的资源消耗情况，从所述第一候选算子、和所述多种分块策略中，选择所述待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的目标候选分块策略。

- 25 一种可能的实施方式中，所述资源消耗情况用计算开销值表示，所述第二确定模块 502，用于根据以下步骤确定所述第一候选算子在每种分块策略下的计算开销值：

确定所述第一候选算子在预设尺寸下对应的受限场景，其中，所述受限场景为基于在预设尺寸下所述第一候选算子对应的数据容量的计算耗时和传输耗时确定的；

- 30 在所述受限场景属于带宽受限场景的情况下，基于所述分块策略进行分块的分块结果，确定所述第一候选算子在所述分块策略下对应的直接内存操作 DMA 数据传输总量、DMA 任务个数、和数据转换开销；基于所述 DMA 数据传输总量、所述 DMA 任务个数、所述数据转换开销、和所述计算设备对应的 DMA 速率、DMA 任务开销，确定所述第一候选算子在所述分块策略下的计算开销值；其中，所述数据转换开销为按照所述第一候选算子对应的目标数据排布方式，对所述第一候选算子对应的输入数据进行数据排布方式转换所消耗的时间；

- 35 在所述受限场景属于计算受限场景的情况下，基于所述分块策略进行分块的分块结果，确定所述第一候选算子在所述分块策略下对应的参数数据的计算耗时、所述第一候选算子的算子调用次数、初始数据传输总量、DMA 任务个数、和数据转换开销；基于所述计算耗时、所述算子调用次数、初始数据传输总量、数据转换开销、DMA 任务开销、DMA 任务个数、和所述计算设备对应的 DMA 速率，确定所述第一候选算子在所述分块策略下的计算开销值。

- 40 一种可能的实施方式中，所述第二确定模块 502，在基于所述第一候选算子在所述多种分块策略中的每种分块策略下的资源消耗情况，从所述第一候选算子、和所述多种分块策略中，选择所述

待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的一个或多个目标候选分块策略时，用于：

从所述第一候选算子对应的多个所述资源消耗情况中，选取满足预设条件的目标资源消耗情况；其中，一个第一候选算子在一种分块策略下对应一个所述资源消耗情况；

- 5 将所述目标资源消耗情况对应的分块策略，确定为候选分块策略，并基于所述候选分块策略，运行包含所述目标资源消耗情况对应的第二候选算子的待处理网络层，确定与所述候选分块策略及所述第二候选算子对应的测试结果；

基于所述测试结果，确定所述待处理网络层对应的一个或多个目标候选算子以及与所述目标候选算子对应的目标候选分块策略。

- 10 一种可能的实施方式中，从所述第一候选算子、和所述多种分块策略中，选择所述待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的目标候选分块策略之前，还包括：

对齐模块 504，用于基于确定的所述目标神经网络对应的最小颗粒度信息，对所述第一候选算子对应的参数数据进行对齐操作，得到所述第一候选算子对应的对齐后的参数数据；

- 15 其中，所述最小颗粒度信息中包括所述参数数据在不同维度下对应的最小颗粒度；对齐后的参数数据在不同维度下的尺寸，是所述最小颗粒度信息指示的对应维度下的最小颗粒度的整数倍。

一种可能的实施方式中，在所述参数数据包括输入数据和常数数据的情况下，所述多种分块策略包括以下至少一种：

- 20 将全部输入数据作为初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果；所述初始数据为计算设备运行所述目标神经网络时，写入直接内存操作 DMA 任务所分配的初始数据区域内的数据；

将全部常数数据作为所述初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果；

- 25 将部分输入数据作为所述初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果；其中，部分输入数据的目标尺寸为根据所述输入数据的第一维度的最小颗粒度确定；

将部分常数数据作为所述初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果；其中，部分常数数据的目标尺寸为根据所述常数数据的第一维度的最小颗粒度确定的。

- 30 一种可能的实施方式中，所述将部分输入数据作为初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果，包括：

基于所述输入数据的第一维度的最小颗粒度的 i 倍，确定所述部分输入数据的目标尺寸；

分别将所述目标尺寸的所述部分输入数据作为初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果；

- 35 其中， i 为在确定了部分输入数据的目标尺寸后，使得所述部分输入数据的数据容量，以及基于所述常数数据的维度参数的最小颗粒度，确定的常数数据块的数据容量，满足计算设备的内存要求的正整数。

一种可能的实施方式中，所述将部分常数数据作为初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果，包括：

- 40 基于所述常数数据的第一维度的最小颗粒度的 j 倍，确定所述部分常数数据的目标尺寸；

分别将所述目标尺寸的所述部分常数数据作为初始数据，基于确定的所述输入数据的维度参数，

对所述输入数据进行指定维度的分块，得到分块结果；

其中， j 为在确定了部分常数数据的目标尺寸后，使得所述部分常数数据的数据容量，以及基于所述输入数据的维度参数的最小颗粒度，确定的输入数据块的数据容量，满足计算设备的内存要求的正整数。

- 5 一种可能的实施方式中，在所述指定维度为一维，所述维度参数包括第一维度的情况下，分别将所述常数数据和所述输入数据作为目标数据，基于确定的所述目标数据的第一维度，对所述目标数据进行一维分块，得到一维分块结果，包括：

将所述目标数据的第一维度对应的最小颗粒度的 k 倍，确定为目标分块尺寸，基于所述目标分块尺寸，将所述目标数据按照所述第一维度进行一维分块，得到所述目标数据对应的多个目标数据块；其中， k 为正整数；

10 在确定所述多个目标数据块和所述初始数据满足设置的分块条件的情况下，将所述目标数据的第一维度对应的最小颗粒度的 $k+1$ 倍作为更新后的目标分块尺寸，返回至基于所述目标分块尺寸，将所述目标数据按照所述第一维度进行一维分块的步骤，直至确定所述多个目标数据块和所述初始数据不满足设置的分块条件，将所述目标数据的第一维度对应的最小颗粒度的 k 倍确定为所述分块结果；

15 在所述初始数据、和 k 等于 1 时生成的所述多个目标数据块不满足设置的分块条件的情况下，确定所述分块结果为一维分块失败。

- 一种可能的实施方式中，在所述指定维度为二维，所述维度参数包括第二维度的情况下，分别将所述常数数据和所述输入数据作为目标数据，基于确定的所述目标数据的第一维度、第二维度，
- 20 对所述目标数据进行二维分块，得到二维分块结果，包括：

将所述目标数据的第一维度对应的最小颗粒度的 y 倍，确定为第一目标分块尺寸，基于所述第一目标分块尺寸，将所述目标数据按照第一维度进行一维分块，得到所述目标数据对应的多个中间数据块；其中， y 为正整数；

25 将所述目标数据的第二维度对应的最小颗粒度的 x 倍，确定为第二目标分块尺寸；基于所述第二目标分块尺寸，将每个中间数据块按照所述第二维度进行二维分块，得到各个中间数据块分别对应的多个目标数据块；其中， x 为正整数；

30 在确定所述多个目标数据块和所述初始数据满足设置的分块条件的情况下，则将所述目标数据的第二维度对应的最小颗粒度的 $x+1$ 倍作为更新后的第二目标分块尺寸，返回至基于所述第二目标分块尺寸，将每个中间数据块按照第二维度进行二维分块的步骤，直至确定所述多个目标数据块和所述初始数据不满足设置的分块条件为止，将所述目标数据的第二维度对应的最小颗粒度的 x 倍确定为所述分块结果。

一种可能的实施方式中，在所述待处理网络层对应的参数数据还包括输出数据的情况下，确定所述多个目标数据块和所述初始数据满足设置的分块条件，包括：

- 35 在确定所述初始数据、所述输出数据、和每个目标数据块分别满足所述计算设备的内存要求，以及所述初始数据、所述输出数据、和每个目标数据块分别满足所述计算设备中 DMA 传输要求的情况下，确定所述多个目标数据块和所述初始数据满足设置的分块条件。

在一些实施例中，本公开实施例提供的装置具有的功能或包含的模板可以用于执行上文方法实施例描述的方法，其具体实现可以参照上文方法实施例的描述，为了简洁，这里不再赘述。

- 40 基于同一技术构思，本公开实施例还提供了一种电子设备。参照图 6 所示，为本公开实施例提供的电子设备的结构示意图，包括处理器 601、存储器 602、和总线 603。其中，存储器 602 用于存储执行指令，包括内存 6021 和外部存储器 6022；这里的内存 6021 也称内存储器，用于暂时存放处

理器 601 中的运算数据, 以及与硬盘等外部存储器 6022 交换的数据, 处理器 601 通过内存 6021 与外部存储器 6022 进行数据交换, 当电子设备 600 运行时, 处理器 601 与存储器 602 之间通过总线 603 通信, 使得处理器 601 在执行以下指令:

确定目标神经网络中的待处理网络层;

- 5 从确定的多个算子和多种分块策略中, 确定所述目标神经网络中待处理网络层对应的目标算子和目标分块策略; 所述多个算子中的每个算子均用于实现所述待处理网络层对应的功能, 所述多种分块策略中的每个分块策略均匹配用于运行所述目标神经网络的计算设备的运行要求;

基于所述待处理网络层对应的所述目标分块策略, 运行包含所述目标算子的所述目标神经网络。

- 10 此外, 本公开实施例还提供一种计算机可读存储介质, 该计算机可读存储介质上存储有计算机程序, 该计算机程序被处理器运行时执行上述方法实施例中所述的神经网络运行方法的步骤。其中, 该存储介质可以是易失性或非易失的计算机可读取存储介质。

本公开实施例还提供一种计算机程序产品, 该计算机程序产品承载有程序代码, 所述程序代码包括的指令可用于执行上述方法实施例中所述的神经网络运行方法的步骤, 具体可参见上述方法实施例, 在此不再赘述。

- 15 其中, 上述计算机程序产品可以具体通过硬件、软件或其结合的方式实现。在一个可选实施例中, 所述计算机程序产品具体体现为计算机存储介质, 在另一个可选实施例中, 计算机程序产品具体体现为软件产品, 例如软件开发包 (Software Development Kit, SDK) 等等。

- 20 所属领域的技术人员可以清楚地了解到, 为描述的方便和简洁, 上述描述的系统 and 装置的具体工作过程, 可以参考前述方法实施例中的对应过程, 在此不再赘述。在本公开所提供的几个实施例中, 应该理解到, 所揭露的系统、装置和方法, 可以通过其它的方式实现。以上所描述的装置实施例仅仅是示意性的, 例如, 所述单元的划分, 仅仅为一种逻辑功能划分, 实际实现时可以有另外的划分方式, 又例如, 多个单元或组件可以结合或者可以集成到另一个系统, 或一些特征可以忽略, 或不执行。另一点, 所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些通信接口, 装置或单元的间接耦合或通信连接, 可以是电性, 机械或其它的形式。

- 25 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单元来实现本实施例方案的目的。

另外, 在本公开各个实施例中的各功能单元可以集成在一个处理单元中, 也可以是各个单元单独物理存在, 也可以两个或两个以上单元集成在一个单元中。

- 30 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时, 可以存储在一个处理器可执行的非易失的计算机可读取存储介质中。基于这样的理解, 本公开的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案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一个存储介质中, 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) 执行本公开各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:
- 35 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)、磁碟或者光盘等各种可以存储程序代码的介质。

以上仅为本公开的具体实施方式, 但本公开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公开揭露的技术范围内, 可轻易想到变化或替换, 都应涵盖在本公开的保护范围之内。因此, 本公开的保护范围应以权利要求的保护范围为准。

40

权 利 要 求 书

1、一种神经网络运行方法，其特征在于，包括：

确定目标神经网络中的待处理网络层；

5 从确定的多个算子和多种分块策略中，确定所述目标神经网络中所述待处理网络层对应的目标算子和目标分块策略；所述多个算子用于实现所述待处理网络层对应的功能，所述多种分块策略匹配用于运行所述目标神经网络的计算设备的运行要求；

基于所述待处理网络层对应的所述目标分块策略，运行包含所述目标算子的所述目标神经网络。

2、根据权利要求1所述的方法，其特征在于，所述分块策略用于对所述待处理网络层对应的目标算子的参数数据进行分块；

10 在所述多种分块策略中，基于采用所述目标分块策略对所述目标算子的参数数据进行分块得到的参数数据，运行所述待处理网络层的资源消耗最小。

3、根据权利要求1或2所述的方法，其特征在于，在所述待处理网络层为多个的情况下，所述从确定的多个算子和多种分块策略中，确定所述目标神经网络中待处理网络层对应的目标算子和目标分块策略，包括：

15 针对所述目标神经网络中的任一待处理网络层，从所述多个算子中确定所述待处理网络层对应目标候选算子、并从所述多种分块策略中确定与所述目标候选算子匹配的目标候选分块策略；

在存在任一待处理网络层对应的所述目标候选算子为多个和/或所述目标候选分块策略为多个的情况下，基于所述待处理网络层对应的目标候选算子和目标候选分块策略，确定所述待处理网络层对应的所述目标算子和所述目标分块策略。

20 4、根据权利要求3所述的方法，其特征在于，所述基于所述待处理网络层对应的目标候选算子和目标候选分块策略，确定所述待处理网络层对应的所述目标算子和所述目标分块策略，包括：

基于所述待处理网络层对应的目标候选算子、和与所述目标候选算子对应的目标候选分块策略，确定所述目标神经网络对应的多个测试网络；其中，所述多个测试网络中的任一测试网络包括所述待处理网络层对应的一个所述目标候选算子、和与该目标候选算子匹配的一个目标候选分块策略；

25 分别运行所述多个测试网络，得到多个测试结果，其中，所述多个测试网络中的任一测试网络对应一个测试结果；

基于所述多个测试结果，从所述多个测试网络中选取目标测试网络；

将所述目标测试网络中所述待处理网络层的目标候选算子和目标候选分块策略，确定为所述目标神经网络中所述待处理网络层对应的所述目标算子和目标分块策略。

30 5、根据权利要求3或4所述的方法，其特征在于，所述针对所述目标神经网络中的任一待处理网络层，从所述多个算子中确定所述待处理网络层对应目标候选算子、并从所述多种分块策略中确定与所述目标候选算子匹配的目标候选分块策略，包括：

针对所述待处理网络层，从所述多个算子中，确定一个或多个第一候选算子；

35 基于所述第一候选算子在所述多种分块策略中的每种分块策略下的资源消耗情况，从所述第一候选算子、和所述多种分块策略中，选择所述待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的目标候选分块策略。

6、根据权利要求5所述的方法，其特征在于，所述资源消耗情况用计算开销值表示，根据以下步骤确定所述第一候选算子在所述每种分块策略下的计算开销值：

40 确定所述第一候选算子在预设尺寸下对应的受限场景，其中，所述受限场景为基于在预设尺寸下所述第一候选算子对应的数据容量的计算耗时和传输耗时确定的；

在所述受限场景属于带宽受限场景的情况下，基于所述分块策略进行分块的分块结果，确定所

述第一候选算子在所述分块策略下对应的直接内存操作 DMA 数据传输总量、DMA 任务个数、和数据转换开销；基于所述 DMA 数据传输总量、所述 DMA 任务个数、所述数据转换开销、和所述计算设备对应的 DMA 速率、DMA 任务开销，确定所述第一候选算子在所述分块策略下的计算开销值；其中，所述数据转换开销为按照所述第一候选算子对应的目标数据排布方式，对所述第一候选算子对应的输入数据进行数据排布方式转换所消耗的时间；

5 在所述受限场景属于计算受限场景的情况下，基于所述分块策略进行分块的分块结果，确定所述第一候选算子在所述分块策略下对应的参数数据的计算耗时、所述第一候选算子的算子调用次数、初始数据传输总量、DMA 任务个数、和数据转换开销；基于所述计算耗时、所述算子调用次数、初始数据传输总量、数据转换开销、DMA 任务开销、DMA 任务个数、和所述计算设备对应的 DMA 速率，确定所述第一候选算子在所述分块策略下的计算开销值。

10 7、根据权利要求 5 或 6 所述的方法，其特征在于，所述基于所述第一候选算子在所述多种分块策略中的每种分块策略下的资源消耗情况，从所述第一候选算子、和所述多种分块策略中，选择所述待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的一个或多个目标候选分块策略，包括：

15 从所述第一候选算子对应的多个所述资源消耗情况中，选取满足预设条件的目标资源消耗情况；其中，一个第一候选算子在一种分块策略下对应一个所述资源消耗情况；

将所述目标资源消耗情况对应的分块策略，确定为候选分块策略，并基于所述候选分块策略，运行包含所述目标资源消耗情况对应的第二候选算子的待处理网络层，确定与所述候选分块策略及所述第二候选算子对应的测试结果；

20 基于所述测试结果，确定所述待处理网络层对应的一个或多个目标候选算子以及与所述目标候选算子对应的目标候选分块策略。

8、根据权利要求 5~7 任一所述的方法，其特征在于，从所述第一候选算子、和所述多种分块策略中，选择所述待处理网络层对应的一个或多个目标候选算子、以及与所述目标候选算子对应的目标候选分块策略之前，还包括：

25 基于确定的所述目标神经网络对应的最小颗粒度信息，对所述第一候选算子对应的参数数据进行对齐操作，得到所述第一候选算子对应的对齐后的参数数据；

其中，所述最小颗粒度信息中包括所述参数数据在不同维度下对应的最小颗粒度；所述对齐后的参数数据在不同维度下的尺寸，是所述最小颗粒度信息指示的对应维度下的最小颗粒度的整数倍。

30 9、根据权利要求 1~8 任一所述的方法，其特征在于，在所述参数数据包括输入数据和常数数据的情况下，所述多种分块策略包括以下至少一种：

将全部输入数据作为初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果；所述初始数据为计算设备运行所述目标神经网络时，写入直接内存操作 DMA 任务所分配的初始数据区域内的数据；

35 将全部常数数据作为所述初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果；

将部分输入数据作为所述初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果；其中，部分输入数据的目标尺寸为根据所述输入数据的第一维度的最小颗粒度确定；

40 将部分常数数据作为所述初始数据，基于确定的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果；其中，部分常数数据的目标尺寸为根据所述常数数据的第一维度的最小颗粒度确定的。

10、根据权利要求 9 所述的方法，其特征在于，所述将部分输入数据作为初始数据，基于确定的所述常数数据的维度参数，对所述常数数据进行指定维度的分块，得到分块结果，包括：

基于所述输入数据的第一维度的最小颗粒度的 i 倍，确定所述部分输入数据的目标尺寸；

分别将所述目标尺寸的所述部分输入数据作为初始数据，基于确定的所述常数数据的维度参数，
5 对所述常数数据进行指定维度的分块，得到分块结果；

其中， i 为在确定了部分输入数据的目标尺寸后，使得所述部分输入数据的数据容量，以及基于所述常数数据的维度参数的最小颗粒度，确定的常数数据块的数据容量，满足计算设备的内存要求的正整数。

11、根据权利要求 9 所述的方法，其特征在于，所述将部分常数数据作为初始数据，基于确定的
10 的所述输入数据的维度参数，对所述输入数据进行指定维度的分块，得到分块结果，包括：

基于所述常数数据的第一维度的最小颗粒度的 j 倍，确定所述部分常数数据的目标尺寸；

分别将所述目标尺寸的所述部分常数数据作为初始数据，基于确定的所述输入数据的维度参数，
对所述输入数据进行指定维度的分块，得到分块结果；

其中， j 为在确定了部分常数数据的目标尺寸后，使得所述部分常数数据的数据容量，以及基于
15 所述输入数据的维度参数的最小颗粒度，确定的输入数据块的数据容量，满足计算设备的内存要求的正整数。

12、根据权利要求 9~11 任一所述的方法，其特征在于，在所述指定维度为一维，所述维度参数包括第一维度的情况下，分别将所述常数数据和所述输入数据作为目标数据，基于确定的所述目标数据的第一维度，对所述目标数据进行一维分块，得到分块结果，包括：

20 将所述目标数据的第一维度对应的最小颗粒度的 k 倍，确定为目标分块尺寸，基于所述目标分块尺寸，将所述目标数据按照所述第一维度进行一维分块，得到所述目标数据对应的多个目标数据块；其中， k 为正整数；

在确定所述多个目标数据块和所述初始数据满足设置的分块条件的情况下，将所述目标数据的第一维度对应的最小颗粒度的 $k+1$ 倍作为更新后的目标分块尺寸，返回至基于所述目标分块尺寸，
25 将所述目标数据按照所述第一维度进行一维分块的步骤，直至确定所述多个目标数据块和所述初始数据不满足设置的分块条件，将所述目标数据的第一维度对应的最小颗粒度的 k 倍确定为所述分块结果；

在所述初始数据、和 k 等于 1 时生成的所述多个目标数据块不满足设置的分块条件的情况下，确定所述分块结果为一维分块失败。

30 13、根据权利要求 9~12 任一所述的方法，其特征在于，在所述指定维度为二维，所述维度参数包括第二维度的情况下，分别将所述常数数据和所述输入数据作为目标数据，基于确定的所述目标数据的第一维度、第二维度，对所述目标数据进行二维分块，得到分块结果，包括：

将所述目标数据的第一维度对应的最小颗粒度的 y 倍，确定为第一目标分块尺寸，基于所述第一目标分块尺寸，将所述目标数据按照第一维度进行一维分块，得到所述目标数据对应的多个中间
35 数据块；其中， y 为正整数；

将所述目标数据的第二维度对应的最小颗粒度的 x 倍，确定为第二目标分块尺寸；基于所述第二目标分块尺寸，将至少一个中间数据块按照所述第二维度进行二维分块，得到各个中间数据块分别对应的多个目标数据块；其中， x 为正整数；

在确定所述多个目标数据块和所述初始数据满足设置的分块条件的情况下，则将所述目标数据的第二维度对应的最小颗粒度的 $x+1$ 倍作为更新后的第二目标分块尺寸，返回至基于所述第二目标
40 分块尺寸，将至少一个中间数据块按照第二维度进行二维分块的步骤，直至确定所述多个目标数据

块和所述初始数据不满足设置的分块条件为止，将所述目标数据的第二维度对应的最小颗粒度的 x 倍确定为所述分块结果。

14、根据权利要求 12 或 13 所述的方法，其特征在于，在所述待处理网络层对应的参数数据包括输出数据的情况下，所述确定所述多个目标数据块和所述初始数据满足设置的分块条件，包括：

5 在确定所述初始数据、所述输出数据、和至少一个目标数据块分别满足所述计算设备的内存要求，以及所述初始数据、所述输出数据、和至少一个目标数据块分别满足所述计算设备中 DMA 传输要求的情况下，确定所述多个目标数据块和所述初始数据满足设置的分块条件。

15、一种神经网络运行装置，其特征在于，包括：

第一确定模块，用于确定目标神经网络中的待处理网络层；

10 第二确定模块，用于从确定的多个算子和多种分块策略中，确定所述目标神经网络中待处理网络层对应的目标算子和目标分块策略；所述多个算子中用于实现所述待处理网络层对应的功能，所述多种分块策略匹配用于运行所述目标神经网络的计算设备的运行要求；

运行模块，用于基于所述待处理网络层对应的所述目标分块策略，运行包含所述目标算子的所述目标神经网络。

15 16、一种电子设备，其特征在于，包括：处理器、存储器和总线，所述存储器存储有所述处理器可执行的机器可读指令，当电子设备运行时，所述处理器与所述存储器之间通过总线通信，所述机器可读指令被所述处理器执行时执行如权利要求 1 至 14 任一所述的神经网络运行方法的步骤。

17、一种计算机可读存储介质，其特征在于，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行如权利要求 1 至 14 任一所述的神经网络运行方法的步骤。

20 18、一种计算机程序，包括计算机可读代码，当所述计算机可读代码在电子设备中运行时，所述电子设备中的处理器执行用于实现权利要求 1-14 中的任一权利要求所述的方法。

25

30

35