

一种基于神经网络的任务处理方法及相关设备

本申请要求在2018年10月10日提交中国专利局、申请号为201811180174.5的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机网络技术领域，例如一种基于神经网络的任务处理方法及相关设备。

背景技术

随着人工智能技术的快速发展，以深度神经网络为代表的机器学习方法在计算机视觉、语音识别等领域取得了实际应用，成为研究热点。

在实际部署基于神经网络的应用时，不仅要考虑网络本身运算开销，还需考虑应用整体的延时和吞吐量控制。目前，在实际应用，尤其是部署在移动端的实时类应用，通常使用多核处理器能力，来加速神经网络中的每一层的运算，即将神经网络中的每一网络层的运算任务分配到多个处理器核心进行处理，以通过多个处理器核心来完成同一层的运算。但是，将神经网络中的每一网络层的运算任务分配到多个处理器核心进行处理时，将运算任务分配至多核以及从多核收集计算结果的耗时可能会超过运算本身的耗时，如在单层耗时基本在0.5毫秒以下的情况下，多核之间调度所带来的额外开销（Overhead）可能会比运算本身的开销高，影响加速比。

发明内容

本申请实施例提供一种基于神经网络的任务处理方法及相关设备，以提高加速比，避免了相关技术中神经网络相关应用在多核处理器上的运算效率低的情况。

第一方面，本申请实施例提供了一种基于神经网络的任务处理方法，包括：获取输入数据，其中，所述输入数据用于触发线程任务，所述输入数据为源输入数据或缓存交换数据；根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，产生处理结果数据；其中，所述至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别

对应；将所述处理结果数据输出至缓存，以作为除所述至少两个模块线程之外的模块线程的缓存交换数据，或，输出所述处理结果数据，以作为源输入数据的处理结果。

第二方面，本申请实施例还提供了一种基于神经网络的任务处理装置，包括：输入数据获取模块，设置为获取输入数据，其中，所述输入数据用于触发线程任务，所述输入数据为源输入数据或缓存交换数据；模块线程调度模块，设置为根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，产生处理结果数据；其中，所述至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别对应；处理结果数据输出模块，设置为将所述处理结果数据输出至缓存，以作为除所述至少两个模块线程之外的模块线程的缓存交换数据，或，输出所述处理结果数据，以作为源输入数据的处理结果。

第三方面，本申请实施例还提供了一种设备，包括：处理器和存储器；所述存储器中存储有至少一条指令，所述指令由所述处理器执行，使得所述设备执行如第一方面所述的基于神经网络的任务处理方法。

第四方面，本申请实施例还提供了一种计算机可读存储介质，所述存储介质中的指令由设备的处理器执行时，使得设备能够执行如第一方面所述的基于神经网络的任务处理方法。

附图说明

图 1 是本申请的一种基于神经网络的任务处理方法实施例的步骤流程示意图；

图 2 是本申请一实施例中的两个模块线程并行执行两个网络模块的任务的示意图；

图 3 是本申请一实施例中的两个模块线程进行数据交换的示意图；

图 4 是本申请一实施例中的起始模块线程、中间模块线程以及末尾模块线程之间进行数据交换的示意图；

图 5 是本申请一实施例中的一种基于神经网络的任务处理装置实施例的结构方框示意图；

图 6 是本申请一实施例中的一种设备的结构方框示意图。

具体实施方式

申请人在实现本申请时发现，神经网络中的运算任务是顺序执行任务，即输入到神经网络中的数据依次通过该神经网络不同层的处理，得到最后的输出结果。相关技术中经常使用多核处理器能力来加速神经网络相关应用的任务处理。例如，对于不需要同步执行的应用，如云端的人脸识别等，通常是通过批量处理的方式，把许多类似的不相关任务分配到不同的处理器核上进行运算，以在每个任务的耗时接近时，获得理论上最高的并行度，但是这种方案不支持要求同步的应用，即需要实时回显的应用无法使用该方案。对于要求同步的应用，即在需要同步执行的场合，通常采用另一种方案——以神经网络中单层通道为单位，基于多核处理器来加速每一层的运算，如利用一个加速神经网络计算的加速包 nnPack 或一个开源的矩阵计算库 OpenBlas 中所包含的多处理器核优化的层算法来进行加速，唯一的并行度存在神经网络的单层内部，因此这种方案在单层耗时比较大时有效，而在单层耗时较小时无法有效利用多核进行加速，如对于移动端或实时级别的神经网络，结构上的网络层数较多而单层通道数量较小，单层耗时基本在 0.5 毫秒以下，此时多核调度带来的额外开销通常和单层网络运算本身的开销不相上下，甚至多核处理比单核处理慢。由此可知，相关技术给出一个两难问题：在需用同步执行的场合应用以单层通道为单位的方案仅在单层耗时比较大的情况下有效，在不需要同步执行的场合需要每个任务耗时接近且不适用于需要实时回显的应用。

为避免上述情况，本申请实施例提出了一种新的基于神经网络的任务处理方法。本申请实施例可以通过并行调度至少两个与神经网络的网络模块对应的模块线程，对输入数据进行处理，以极大地掩盖神经网络运算的耗时，增大吞吐量，即充分利用多处理器核性能来加速神经网络相关的应用，使得应用能够实时回显。其中，网络模块为预先依据神经网络中的网络层进行划分得到的，一个网络模块可以包括至少一个网络层，本申请实施例对此不作限制。

在实际处理中，为了实现在多核处理器上加速神经网络应用的任务处理，可以预先对神经网络中的网络层进行划分，得到至少两个网络模块。每一个网络模块可以包括至少一个网络层。在运行开始时，可以基于神经网络划分后得到的网络模型，创建包含至少两个线程的线程池，且线程池中的每个线程可以对应唯一网络模块的执行。本申请实施例中，可以将该线程池中的与网络模块对应的线程称为模块线程，以通过并行调度模块线程来分别执行神经网络中每

个网络模块所需要执行的任务，提高加速比，从而提升神经网络相关应用在多核处理器上的运行效率。

参照图 1，示出了本申请的一种基于神经网络的任务处理方法实施例的步骤流程图，包括步骤 110 至步骤 130。

在步骤 110 中，获取输入数据，其中，所述输入数据用于触发线程任务，所述输入数据为源输入数据或缓存交换数据。

本申请实施例在需要进行任务处理的情况下，可以针对当前所需要执行的任务获取输入数据，以根据当前获取到的输入数据触发相应的线程任务进行处理。其中，当前获取到的输入数据可以用于触发线程任务，包括与任务处理相关的数据，如可以包括：源输入数据、缓存交换数据等，本申请实施例对此不作限制。

需要说明的是，本申请实施例中的源输入数据可以是指任务处理所需要的源数据，如可以是图像识别任务所需要的图像帧数据等。缓存交换数据可以是指不同线程之间交换的缓存数据，且缓存数据可以存储在缓存空间中。

在步骤 120 中，根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，产生处理结果数据；其中，所述至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别对应。

本申请实施例在获取到的输入数据后，可以根据该输入数据所触发的至少两个线程任务，从线程池中并行调度至少两个模块线程，以通过调度到的至少两个模块，对所述输入数据进行处理，产生处理结果数据。其中，处理结果数据可以是指模块线程基于其所对应的网络模块进行任务处理后得到的结果数据。

例如，在并行调度了两个模块线程后，如在并行调度第一模块线程和第二模块线程后，通过调度的这两个模块线程对输入数据进行并行处理后，产生的处理结果数据可以包括这两个模块线程所产生的第一处理结果数据和第二处理结果数据；其中，第一处理结果数据可以是指：第一模块线程基于其所对应的第一网络模块进行任务处理后所产生的结果数据；第二处理结果数据可以是指：第二模块线程基于其所对应的第二网络模块进行任务处理后所产生的结果数据。需要说明的是，第一模块线程可以与第一网络模块进行绑定，以对应该第一网络模块的执行；同理，第二模块线程可以与第二网络模块进行绑定，以对应该第二网络模块的执行。

当然，在并行调度至少三个模块线程的情况下，产生的处理结果数据可以

包括通过调度的至少三个模块线程进行处理后得到的结果数据，本申请实施例对此不作限制。例如，在并行调度第一模块线程、第二模块线程以及第三模块线程后，通过第一模块线程、第二模块线程以及第三模块线程分别对获取到的源输入数据或缓存交换数据进行处理后，产生的处理结果数据可以包括有第一模块线程所产生的第一处理结果数据、第二模块线程所产生的第二处理结果数据以及第三模块线程所产生的第三处理结果数据等。

在步骤 130 中，将所述处理结果数据输出至缓存，以作为除所述至少两个模块线程之外的模块线程的缓存交换数据，或，输出所述处理结果数据，以作为源输入数据的处理结果。

本申请实施例在模块线程产生处理结果数据后，可以基于模块线程所对应的网络模块，对处理结果数据进行输出。示例性的，在模块线程所对应的网络模块为神经网络中的起始网络模块或中间网络模块的情况下，可以将该模块线程所产生的处理结果数据输出至缓存中，作为与除所述至少两个模块线程之外的模块线程进行交互的缓存交换数据，即将该模块线程所产生的处理结果数据作为下一个模块线程的输入数据，从而使得下一个模块线程可以基于该模块线程所产生的处理结果数据进行下一步的任务处理，如执行神经网络中的下一个网络模块的任务。在模块线程所对应的网络模块为神经网络中的末尾网络模块的情况下，可以确定该模块线程所产生的处理结果数据为源输入数据的处理结果并输出，从而使得神经网络相关应用可以利用该处理结果进行业务处理，满足业务需求。

需要说明的是，本申请实施例中的源输入数据的处理结果可以用于表征神经网络对源输入数据进行处理后输出的结果。

本申请实施例中的起始网络模块可以是指神经网络划分后得到的第一个网络模块，设置为接收传输给神经网络的源输入数据，还设置为执行该起始网络模块所包含的网络层所需要执行的任务，起始网络模块包括神经网络中用于接收源输入数据的输入网络层（即神经网络的输入层），或者，还可以包括神经网络中其他的至少一个网络层等，本申请实施例对此不作限制。

另外，本申请实施例中的末尾网络模块可以是指神经网络划分后得到的最后一个网络模块，设置为输出源输入数据的处理结果，还设置为执行该末尾网络模块所包含的网络层所需要执行的任务，末尾网络模块包括神经网络中用于输出处理结果的网络层（即神经网络的输出层），或者，还可以包括神经网络中

其他的至少一个网络层等，本申请实施例对此也不作限制。

相应的，本申请实施例中的中间网络模块可以包括神经网络中的至少一个中间网络层，且该中间网络层可以是指神经网络中除了第一个网络层和最后一个网络层之外的网络层。例如，在神经网络包括 5 个网络层的情况下，中间网络模块可以包括该神经网络中依次连接的第二个网络层、第三个网络层以及第四个网络层；或者，中间网络模块可以包括第三个网络层和第四个网络层；或者，中间网络模块可以包括第三个网络层。示例性的，在起始网络模块包括第一个网络层，末尾网络模块包含第五个网络层的情况下，中间网络模块可以包括第二个网络层、第三个网络层以及第四个网络层；在起始网络模块包括依次连接的第一个网络层和第二网络层，末尾网络模块包含第五个网络层的情况下，中间网络模块可以包括第三个网络层和第四个网络层；在起始网络模块包括依次连接的第一个网络层和第二网络层，末尾网络模块包含第五个网络层和第四个网络层的情况下，中间网络模块可以只包括第三个网络层，等等。

综上，本申请实施例在获取到用于触发线程任务的输入数据后，可以根据触发的至少两个线程任务，并行调度对应的至少两个模型线程，对输入数据进行处理，且并行调度的至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别对应，即将神经网络中的不同网络模块的任务分配到不同的模块线程上并行执行，极大掩盖神经网络运算的耗时，提升了神经网络相关应用在多核处理器上的运行效率，即避免了相关技术中神经网络相关应用的多核处理器上的运算效率低的情况，使得实时应用能够充分利用多核计算能力进行加速，应用范围广。

在一实施例中，本申请实施例提供的任务处理方法可以作为工程优化方法，可以以库或源码的形式作为任意神经网络相关应用的依赖组件，应用到基于神经网络部署的各种应用中，使得配置有多处理器核的设备可以充分利用多处理器核性能，对神经网络相关应用进行加速，达到实时要求。例如，可以根据本申请实施例提供的方法编写相关多线程逻辑，再以库或源码形式集成至应用中，使得诸如手机等设备在处理该应用的运算任务时，可以充分利用多核性能来进行加速，达到充分利用硬件资源的目的。

在一实施例中，本申请实施例可以预先对神经网络中进行划分，以将该神经网络中所包含的网络层划分到 N 个网络模块中，从而可以通过该 N 个网络模块所对应的 N 个模块线程，分别对神经网络中网络层所要执行的任务进行处理，

实现并行执行神经网络的每个网络模块的任务。其中，N可以为大于1的整数，可以用于表征神经网络划分后得到的网络模块的数量。因此，在本申请的一个实施例中，所述获取输入数据之前，还可以包括：对所述神经网络中的网络层进行划分，得到至少两个网络模块。需要说明的是，划分后得到的每个网络模块可以包括神经网络中的至少一个网络层，如可以包括卷积神经网络中的至少一个卷积层等，本申请实施例对此不作限制。

在一实施例中，为了简化不同网络模块之间的数据交互复杂度，本申请实施例可以选择神经网络中具有一个输入输出的网络层，作为网络模块之间连接点。例如，上述对所述神经网络中的网络层进行划分，得到至少两个网络模块，包括：分别确定所述神经网络中每两个相邻网络层之间的通道数量；在所述相邻网络层之间的通道数量为一个的情况下，将所述相邻网络层中的在前网络层划分为输入网络模块的输入层，将所述相邻网络层中的在后网络层划分为输出网络模块的输出层；基于所述输出网络模块的输出层和所述输入网络模块的输入层，生成至少两个网络模块。

在一实施例中，本申请实施例在进行神经网络划分时，可确定该神经网络的拓扑图结构，确定出该神经网络中所包含的网络层，如确定出卷积神经网络（Convolutional Neural Network, CNN）中所包含的输入层、卷积层、采样层等；并且可确定该神经网络中所包含的网络层之间的通道数量，如可以确定该神经网络中相邻的两个网络层之间的通道数量，随后可基于该神经网络中相邻的两个网络层之间的通道数量对该神经网络中的网络层进行划分，得到至少两个网络模块。

在一实施例中，可以通过判断神经网络中相邻的两个网络层之间的通道数量是否为一个，来确定是否将该相邻的两个网络层划分到不同的网络模块中。例如，在神经网络中相邻的两个网络层之间的通道数量为一个的情况下，即在相邻网络层之间的通道数量为一个的情况下，可以确定该相邻网络层中的在前网络层通过一个通道将数据输出给在后网络层，然后可将该相邻网络层中的在前网络层划分为输入网络模块的输入层，将该相邻网络层中的在后网络层划分为输出网络模块的输出层，随后可以基于划分后得到的输出网络模块的输出层和输入网络模块的输入层，确定出至少两个网络模块。其中，在前网络层所输出的数据可以作为在后网络层的输入数据，亦即，相邻网络层中的在后网络层通过一个通道获取在前网络层所输出的输入数据。

在一实施例中，神经网络划分后得到的网络模块之间的处理耗时的差异值小于设定阈值。例如，本申请实施例可以基于神经网络中每个网络模型进行任务处理所需要的耗时，对神经网络进行划分，使得划分后得到的网络模块之间的处理耗时的差异值可以小于设定阈值，以在资源开销和并行性能之间达到较好的权衡。在神经网络本身有前后处理部分的情况下，视其耗时长短可以作为一个模块或某个模块的部分参与划分，即可以基于网络本身前、后部分的处理耗时，将网络本身前部分作为一个模块或某个模块的部分参与网络模块的划分，同时可以将网络本身后部分作为一个模块或某个模块的部分参与网络模块的划分。实际执行时，网络本身前后部分的耗时可以被掩盖起来，最终表现出来的耗时可以等同于划分后得到的众模块中耗时最长的那一个网络模块的处理耗时。

在一实施例中，可在离线状态或在线状态下，将CNN拆分为N个网络模块，且网络模块之间的耗时大致可以相等。每个网络模块的任务可以被分配到不同的模块线程上执行。对于CNN应用所需要处理的每一帧图像数据而言，可以依次由神经网络划分后的得到的N个网络模块所对应的N个模块线程按照顺序处理，总耗时理论上不变；但是从CNN应用整体来看，单位时间内所处理的图像数据的帧数增加了N倍。例如，在N为2的情况下，可以将前处理划分到第一个网络模块CNN1所对应的模型线程进行处理，而将后处理划分到第二个网络模块CNN2所对应的模型线程进行处理，如图2所示，从而可以将图像帧处理所需要的耗时 Δt_1 缩短为原来串行执行方法所需要的耗时 Δt_2 的一半，进而可以增加CNN应用在单位时间内所处理的图像帧数据对应的帧数，缩短图像帧的刷新时间，提高CNN应用显示的帧率，进而能够提升用户体验。

可见，基于本申请实施例提供的基于神经网络处理方法，可以将相邻图像帧数据纳入并行处理中，即可通过调度的至少两个网络模块对相邻图像帧数据进行并行处理，能够在保持图像帧的执行顺序的前提下充分利用多核进行加速。

在一实施例中，模块线程可以与处理器核一一对应绑定，从而可以通过多个不同处理器核来并行执行划分后的网络模块的任务，实现在多核处理器上加速CNN应用，即能够有效利用多核进行加速，达到充分利用硬件资源的目的。例如，在一些配置有图形处理器（Graphics Processing Unit，GPU）设备上，当CNN应用在该设备上执行诸如渲染，显示等操作，可以采用本申请实施例提供的任务处理方法，利用空闲的GPU资源对该CNN应用所需要执行的运算任务进行加速，以极大掩盖CNN运算耗时，达到实时要求。

在一实施例中，所述至少两个模块线程至少可以包括起始模块线程和末尾模块线程；起始模块线程可以是指与划分神经网络后得到的起始网络模块对应的线程，可以用于执行起始网络模块的任务；末尾模块线程可以是指与划分神经网络后得到的末尾网络模块对应的线程，可以用于执行末尾网络模块的任务。上述根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对上述输入数据进行处理，包括：根据触发的线程任务，调度起始模块线程，对输入数据进行处理；根据触发的线程任务，调度末尾模块线程，对输入数据进行处理。

在一实施例中，根据触发的线程任务，调度起始模块线程，对输入数据进行处理，包括：调用起始模块线程对源输入数据进行前处理，并基于与起始模块线程所对应的神经网络的起始网络模块进行任务处理，将处理结果数据输出至缓存，作为缓存交换数据。根据触发的线程任务，调度末尾模块线程，对输入数据进行处理，包括：调用末尾模块线程从缓存中获取对应的缓存交换数据作为输入数据，基于与末尾模块线程所对应的神经网络的末尾网络模块进行任务处理，将处理结果数据进行后处理并输出，将处理结果数据作为源输入数据的处理结果。

在一实施例中，在将卷积神经网络中的网络层划分到 2 个网络模块后，可以将划分到的第一个网络模块 CNN1 称为起始网络模块，将第二个网络模块 CNN2 称为末尾网络模块，并且可以将与起始网络模块进行绑定的线程称为起始模块线程，以及，可以将与末尾网络模块进行绑定的线程称为末尾模块线程。如图 3 所示，在获取到输入的图像帧数据后，可以根据该图像帧数据所触发的线程任务，调度与第一个网络模块 CNN1 对应的起始模块线程，对输入的图像帧数据进行前处理，并基于神经网络的起始网络模块（即第一个网络模块 CNN1）进行任务处理，产生第一处理结果数据，以及将所述第一处理结果数据输出至缓存，以在同步阶段将该第一处理结果作为缓存交换数据，传输给与第二个网络模块 CNN2 对应的末尾模块线程，即将该缓存交换数据作为末尾模块线程的输入数据，以触发末尾模块线程基于第二个网络模块 CNN2 进行任务处理。

需要说明的是，新的图像帧数据可以在同步阶段写入到起始模块线程对应的输入缓存中，并且也通过乒乓缓存交换，将起始模块线程输出的第一处理结果数据作为交换缓存数据，交换到末尾模块线程对应的输入缓存中，作为该末尾模块线程的输入。如图 3 所示，可以通过预先设置的同步屏障，将起始模块

线程输出至输出缓存的第一处理结果数据，交换到末尾模块线程对应的输入缓存中，作为末尾模块线程的输入，从而使得末尾模块线程可以从其输入缓存中获取到起始网络模块所产生的第一处理结果数据进行下一步的任务处理。

在一实施例中，本申请实施例中的神经网络可以为卷积神经网络；获取到的源输入数据可以为图像帧数据。相应的，上述前处理可以为图像前处理，后处理可以为图像后处理。在图像后处理后，还可以进行其他处理，如图像渲染处理等等，本申请实施例对此不作限制。

在一实施例中，上述模块线程还可以包括至少一个中间模块线程。例如，划分神经网络后得到的网络模块还可以包括至少一个中间网络模块。因此，并行调度的至少两个模块线程还可以包括至少一个与中间网络模块对应的中间模块线程，且该中间模块线程可以用于执行中间网络模块的任务。上述根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，包括：根据触发的线程任务，调度至少一个中间模块线程，对输入数据进行处理。

在一实施例中，根据触发的线程任务，调度至少一个中间模块线程，对输入数据进行处理，包括：调用至少一个中间模块线程从缓存中获取对应的缓存交换数据作为输入数据，基于神经网络的中间网络模块进行任务处理，将处理结果数据输出至缓存，作为缓存交换数据。示例性的，本申请实施例可以在任务处理过程中，可以根据起始模块线程所输出的缓存交换数据，触发中间线程任务，并可根据触发的中间线程任务，调用与神经网络的中间网络模块对应的中间模块线程，对该起始模块线程所输出的缓存交换数据进行任务处理，产生中间处理结果数据，将所述中间处理结果数据输出至缓存，作为与下一个中间模块线程或末尾模块线程进行数据交换的缓存交换数据。

由于神经网络划分后得到的网络模块可以与模块线程一一对应，如图 4 所示，在将卷积神经网络中的网络层划分到 3 个网络模块后，可以将划分到的第一个网络模块 CNN1 称为起始网络模块，将第二个网络模块 CNN2 称为中间网络模块，将第三个网络模块 CNN3 称为末尾网络模块，以及可以将与第一个网络模块 CNN1 对应的模块线程称为起始模块线程，将与第二个网络模块 CNN2 对应的模块线程称为中间模块线程，将与第三个网络模块 CNN3 对应的模块线程称为末尾模块线程。另外，网络模块与模块线程之间的连接关系可以一一对应的，因此每个模块线程间的数据交换可以通过监控与之相关的唯一消费线程

即可实行。

例如，如图 4 所示，起始模块线程对输入的图像帧数据进行图像前处理，并基于神经网络的第一个网络模块 CNN1 进行任务处理，产生第一处理结果数据，并可该第一处理结果数据输出至缓存 A 中。中间模块线程可以获取写入到其所需要执行的线程任务对应的目标缓存空间中的缓存交换数据，并基于神经网络的第二个网络模块 CNN2 进行任务处理，产生中间处理结果数据，并可该中间处理结果数据输出至缓存 B 中。末尾模块线程可以获取写入到其所需要执行的线程任务对应的目标缓存空间(即图 4 中的缓存 C)中的缓存交换数据，并基于神经网络的第三个网络模块 CNN3 进行任务处理，随后可对任务处理后得到的处理结果数据进行图像后处理，以及进行渲染处理，得到源输入数据的处理结果，并依据该处理结果进行输出，如在卷积神经网络相关应用的应用界面进行显示等等。

在数据交换的过程中，数据交换的顺序可以是后往前依次交换，如图 4 所示，可首先将缓存 B 和缓存 C 进行数据交换，然后将缓存 A 和被交换到第二个线程的缓存 C 进行交换，即数据交换的顺序可以与模块线程的执行顺序相反。在一实施例中，第二网络模块 CNN2 的耗时可以约等于第一网络模块 CNN1 加图像前处理的耗时，或者，可以约等于第三网络模块 CNN3 加后处理的耗时，以使得触发的三个模块线程达到最高效率。

在一实施例中，获取输入数据包括：在监听到线程任务所对应读取的目标缓存空间中写入数据的情况下，确定产生线程任务的触发条件，将目标缓存空间中写入的数据作为输入数据，其中，模块线程与线程任务一一对应绑定，线程任务与目标缓存空间绑定。示例性的，本申请实施例可以通过线程任务所对应读取的目标缓存空间是否写入数据，确定是否产生线程任务的触发条件；在监听到线程任务所对应读取的目标缓存空间中写入数据的情况下，可以确定产生线程任务的触发条件，并且可以将目标缓存空间中写入的数据作为输入数据，以依据该输入数据触发相应的线程任务。在从目标缓存空间获取到输入数据后，可以根据触发的线程任务并行调度对应的模块线程，以完成对输入数据的任务处理。

综上，本申请实施例将完整的设计网络划分为多个耗时相近的网络模块，每一个网络模块可以与模块线程一一对应，从而可以将不同网络模块的任务分配至不同的模块线上执行，且每个模块线程可以分配至不同处理器核心，进而

可以通过不同的处理器核心来执行不同网络模块的任务，充分利用硬件资源，如充分利用设备的图形处理器资源、充分利用诸如手机等移动设备的多核性能，使得实时应用也能充分利用多核计算能力进行加速，增大吞吐量，极大地掩盖网络运算的耗时，如提升了卷积神经网络相关应用在多核处理器上的运行效率，对 CNN 相关应用进行加速以达到实时要求。

此外，本申请实施例可以将神经网络应用所需的前后处理部分作为第三方模块，加入到网络层划分中，减少数据处理及显示、渲染等开销；并且，本申请实施例中的模块线程与线程任务进行绑定，从而可以通过简单的乒乓缓存即可完成不同模块线程之间的数据同步，减少多线程间数据同步开销，避免了因同步开销过大而导致多线程效率低的问题，并且在不改变图像帧数据对应的任务执行顺序的前提下，支持要求同步的应用。

需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请实施例并不受所描述的动作顺序的限制，因为依据本申请实施例，某些步骤可以采用其他顺序或者同时进行。

参照图 5，示出了本申请实施例中的一种基于神经网络的任务处理装置实施例的结构框图，该基于神经网络的任务处理装置包括输入数据获取模块 510、模块线程调度模块 520 以及处理结果数据输出模块 530。

输入数据获取模块 510，设置为获取输入数据，其中，所述输入数据用于触发线程任务，所述输入数据为源输入数据或缓存交换数据。

模块线程调度模块 520，设置为根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，产生处理结果数据；其中，所述至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别对应。

处理结果数据输出模块 530，设置为将所述处理结果数据输出至缓存，以作为除所述至少两个模块线程之外的模块线程的缓存交换数据，或，输出所述处理结果数据，以作为源输入数据的处理结果。

在一实施例中，上述基于神经网络的任务处理装置还包括：网络层划分模块。该网络层划分模块设置为对所述神经网络中的网络层进行划分，得到至少两个网络模块。

在一实施例中，所述网络层划分模块可以包括通道数量确定子模块、网络

层划分子模块以及网络模块生成模块。

通道数量确定子模块，设置为分别确定所述神经网络中每两个相邻网络层之间的通道数量。

网络层划分子模块，设置为在所述相邻网络层之间的通道数量为一个的情况下，将所述相邻网络层中的在前网络层划分为输入网络模块的输入层，将所述相邻网络层中的在后网络层划分为输出网络模块的输出层。

网络模块生成模块，设置为基于所述输出网络模块的输出层和所述输入网络模块的输入层，生成至少两个网络模块。

在一实施例中，网络模块之间的处理耗时的差异值小于设定阈值。

在一实施例中，上述至少两个模块线程至少可以包括起始模块线程和末尾模块线程。上述模块线程调度模块 520，设置为根据触发的线程任务，调度起始模块线程，对输入数据进行处理；并且可以根据触发的线程任务，调度末尾模块线程，对输入数据进行处理。其中，根据触发的线程任务，调度起始模块线程，对输入数据进行处理，可以包括：调用起始模块线程对源输入数据进行前处理，并基于与起始模块线程所对应的神经网络的起始网络模块进行任务处理，将处理结果数据输出至缓存，作为缓存交换数据。根据触发的线程任务，调度末尾模块线程，对输入数据进行处理包括：调用末尾模块线程从缓存中获取对应的缓存交换数据作为输入数据，基于与末尾模块线程所对应的神经网络的末尾网络模块进行任务处理，将处理结果数据进行后处理并输出，将处理结果数据作为源输入数据的处理结果。

在一实施例中，模块线程调度模块 520 包括起始模块线程调度子模块以及末尾模块线程调度子模块。

起始模块线程调度子模块，设置为调用起始模块线程对源输入数据进行前处理，并基于与起始模块线程所对应的神经网络的起始网络模块进行任务处理，将处理结果数据输出至缓存，作为缓存交换数据；

末尾模块线程调度子模块，设置为调用末尾模块线程从缓存中获取对应的缓存交换数据作为输入数据，基于与末尾模块线程所对应的神经网络的末尾网络模块进行任务处理，将处理结果数据进行后处理并输出，将处理结果数据作为源输入数据的处理结果。

当然，本申请实施例中的模块线程调度模块 520 还可以包括其他子模块，例如还可以包括至少一个中间模块线程，以根据触发的线程任务，调度中间模

块线程，对输入数据进行处理，产生中间模块线程对应的处理结果数据，等等。在以实施例1中，可以将中间模块线程对应的处理结果数据输出至缓存，作为缓存交换数据，以作为与该中间模块线程关联的下一个模块线程的输入数据，使得下一个模块线程可以获取该缓存交换数据进行任务处理。

在一实施例中，模块线程还包括至少一个中间模块线程。模块线程调度模块520，设置为根据触发的线程任务，调度至少一个中间模块线程，对输入数据进行处理。其中，根据触发的线程任务，调度至少一个中间模块线程，对输入数据进行处理包括：调用至少一个中间模块线程从缓存中获取对应的缓存交换数据作为输入数据，基于神经网络的中间网络模块进行任务处理，将处理结果数据输出至缓存，作为缓存交换数据。

模块线程调度模块520，还可以包括中间模块线程调度子模块。该中间模块线程调度子模块，设置为调用中间模块线程从缓存中获取对应的缓存交换数据作为输入数据，基于神经网络的中间网络模块进行任务处理，将处理结果数据输出至缓存，作为缓存交换数据。

在一实施例中，上述源输入数据可以为图像帧数据，上述前处理可以为图像前处理，上述后处理可以为图像渲染处理，上述神经网络可以为卷积神经网络。

在一个实施例中，输入数据获取模块510包括监听子模块以及确定子模块。

监听子模块，设置为监听线程任务所对应读取的目标缓存空间，例如监听线程任务所对应读取的目标缓存空间是否写入数据等；

确定子模块，设置为在所述监听子模块监听到线程任务所对应读取的目标缓存空间中写入数据的情况下，确定产生线程任务的触发条件，将目标缓存空间中写入的数据作为输入数据。其中，模块线程与线程任务一一对应绑定，线程任务与目标缓存空间绑定。

在一实施例中，线程任务可以与处理器核一一对应绑定。

需要说明的是，上述提供的基于神经网络的任务处理装置可执行本申请任意实施例所提供的视录制方法。

在一实施例中，上述基于神经网络的任务处理装置可以集成在设备中。该设备可以是至少两个物理实体构成，也可以是一个物理实体构成，如设备可以是个人计算机（Personal Computer，PC）、电脑、手机、平板设备、个人数字助理、服务器、消息收发设备、游戏控制台等。

本申请实施例还提供一种设备，包括：处理器和存储器。存储器中存储有至少一条指令，且指令由所述处理器执行，使得所述设备执行如上述方法实施例中所述的基于神经网络的任务处理方法。

参照图6，示出了本申请一个示例中的一种设备的结构示意图。如图6所示，该设备包括：处理器60、存储器61、具有触摸功能的显示屏62、输入装置63、输出装置64以及通信装置65。该设备中处理器60的数量可以是至少一个，图6中以一个处理器60为例。该设备中存储器61的数量可以是至少一个，图6中以一个存储器61为例。该设备的处理器60、存储器61、显示屏62、输入装置63、输出装置64以及通信装置65可以通过总线或者其他方式连接，图6中以通过总线连接为例。

存储器61作为一种计算机可读存储介质，设置为存储软件程序、计算机可执行程序以及模块，如本申请任意实施例所述的任务处理方法对应的程序指令/模块（例如，上述任务处理装置中的输入数据获取模块510、模块线程调度模块520以及处理结果数据输出模块530等）。存储器61可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作装置、至少一个功能所需的应用程序；存储数据区可存储根据设备的使用所创建的数据等。此外，存储器61可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中，存储器61还包括相对于处理器60远程设置的存储器，这些远程存储器可以通过网络连接至设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

显示屏62为具有触摸功能的显示屏62，其可以是电容屏、电磁屏或者红外屏。一般而言，显示屏62设置为根据处理器60的指示显示数据，还设置为接收作用于显示屏62的触摸操作，并将相应的信号发送至处理器60或其他装置。示例性的，在显示屏62为红外屏的情况下，其还包括红外触摸框，该红外触摸框设置在显示屏62的四周，该红外触摸框还设置为接收红外信号，并将该红外信号发送至处理器60或者其他设备。

通信装置65，设置为与其他设备建立通信连接，其可以是有线通信装置和无线通信装置中至少一种。

输入装置63设置为接收输入的数字或者字符信息，以及产生与设备的用户设置以及功能控制有关的键信号输入，还设置为获取图像的摄像头以及获取音

频数据的拾音设备。输出装置 64 可以包括扬声器等音频设备。需要说明的是，输入装置 63 和输出装置 64 的组成可以根据实际情况设定。

处理器 60 通过运行存储在存储器 61 中的软件程序、指令以及模块，从而执行设备的各种功能应用以及数据处理，即实现上述基于神经网络的任务处理方法。

在一实施例中，处理器 60 执行存储器 61 中存储的至少一个程序时，实现如下操作：获取输入数据，其中，所述输入数据用于触发线程任务，所述输入数据为源输入数据或缓存交换数据；根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，产生处理结果数据；其中，所述至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别对应；将所述处理结果数据输出至缓存，以作为除所述至少两个模块线程之外的模块线程的缓存交换数据，或，输出所述处理结果数据，以作为源输入数据的处理结果。

本申请实施例还提供一种计算机可读存储介质，所述存储介质中的指令由设备的处理器执行时，使得设备能够执行如上述方法实施例所述的基于神经网络的任务处理方法。示例性的，该基于神经网络的任务处理方法包括：获取输入数据，其中，所述输入数据用于触发线程任务，所述输入数据为源输入数据或缓存交换数据；根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，产生处理结果数据；其中，所述至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别对应；将所述处理结果数据输出至缓存，以作为除所述至少两个模块线程之外的模块线程的缓存交换数据，或，输出所述处理结果数据，以作为源输入数据的处理结果。

需要说明的是，对于装置、设备、存储介质实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

通过以上关于实施方式的描述，所属领域的技术人员可以清楚地了解到，本申请可借助软件及必需的通用硬件来实现，当然也可以通过硬件实现。基于这样的理解，本申请的技术方案本质上或者说对相关技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如计算机的软盘、只读存储器（Read-Only Memory, ROM）、随机存取存

储器 (Random Access Memory, RAM)、闪存 (FLASH)、硬盘或光盘等, 包括若干指令用以使得一台计算机设备 (可以是机器人, 个人计算机, 服务器, 或者网络设备等) 执行本申请任意实施例所述的基于神经网络的任务处理方法。

值得注意的是, 上述基于神经网络的任务处理装置中, 所包括的每个单元和模块只是按照功能逻辑进行划分的, 但并不局限于上述的划分, 只要能够实现相应的功能即可; 另外, 每个功能单元的名称也只是为了便于相互区分, 并不用于限制本申请的保护范围。

应当理解, 本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中, 多个步骤或方法可以用存储在存储器中且由合适的指令执行装置执行的软件或固件来实现。例如, 如果用硬件来实现, 和在另一实施方式中一样, 可用本领域公知的下列技术中的任一项或他们的组合来实现: 具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路, 具有合适的组合逻辑门电路的专用集成电路, 可编程门阵列 (Programmable Gate Array, PGA), 现场可编程门阵列 (Field-Programmable Gate Array, FPGA) 等。

在本说明书的描述中, 参考术语 “一个实施例”、“一些实施例”、“示例”、“具体示例”、或 “一些示例” 等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请的至少一个实施例或示例中。在本说明书中, 对上述术语的示意性表述不一定指的是相同的实施例或示例。而且, 描述的具体特征、结构、材料或者特点可以在至少一个实施例或示例中以合适的方式结合。

权利要求书

1、一种基于神经网络的任务处理方法，包括：

获取输入数据，其中，所述输入数据用于触发线程任务，所述输入数据为源输入数据或缓存交换数据；

根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，产生处理结果数据；其中，所述至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别对应；

将所述处理结果数据输出至缓存，以作为除所述至少两个模块线程之外的模块线程的缓存交换数据，或，输出所述处理结果数据，以作为源输入数据的处理结果。

2、根据权利要求1所述的方法，所述获取输入数据之前，还包括：

对所述神经网络中的网络层进行划分，得到至少两个网络模块。

3、根据权利要求2所述的方法，其中，对所述神经网络中的网络层进行划分，得到至少两个网络模块，包括：

分别确定所述神经网络中每两个相邻网络层之间的通道数量；

在所述相邻网络层之间的通道数量为一个的情况下，将所述相邻网络层中的在前网络层划分为输入网络模块的输入层，将所述相邻网络层中的在后网络层划分为输出网络模块的输出层；

基于所述输出网络模块的输出层和所述输入网络模块的输入层，生成至少两个网络模块。

4、根据权利要求1所述的方法，其中，网络模块之间的处理耗时的差异值小于设定阈值。

5、根据权利要求1所述的方法，其中，所述至少两个模块线程至少包括起始模块线程和末尾模块线程；

根据触发的线程任务，调度所述起始模块线程，对输入数据进行处理，包括：调用所述起始模块线程对源输入数据进行前处理，并基于与所述起始模块线程所对应的神经网络的起始网络模块进行任务处理，将处理结果数据输出至缓存，作为缓存交换数据；

根据触发的线程任务，调度所述末尾模块线程，对输入数据进行处理，包括：调用所述末尾模块线程从缓存中获取对应的缓存交换数据作为输入数据，基于与所述末尾模块线程所对应的神经网络的末尾网络模块进行任务处理，将处理结果数据进行后处理并输出，将处理结果数据作为所述源输入数据的处理

结果。

6、根据权利要求 5 所述的方法，所述模块线程还包括至少一个中间模块线程，根据触发的线程任务，调度所述至少一个中间模块线程，对输入数据进行处理，包括：

调用所述至少一个中间模块线程从缓存中获取对应的缓存交换数据作为输入数据，基于与所述中间模块线程所对应的神经网络的中间网络模块进行任务处理，将处理结果数据输出至缓存，作为缓存交换数据。

7、根据权利要求 5 或 6 所述的方法，其中，所述源输入数据为图像帧数据，所述前处理为图像前处理，所述后处理为图像后处理，所述神经网络为卷积神经网络。

8、根据权利要求 1 所述的方法，其中，所述获取输入数据包括：

在监听到所述线程任务所对应读取的目标缓存空间中写入数据的情况下，确定产生所述线程任务的触发条件，将目标缓存空间中写入的数据作为输入数据，其中，所述模块线程与所述线程任务一一对应绑定，所述线程任务与所述目标缓存空间绑定。

9、根据权利要求 1 所述的方法，其中，所述模块线程与处理器核一一对应绑定。

10、一种基于神经网络的任务处理装置，包括：

输入数据获取模块，设置为获取输入数据，其中，所述输入数据用于触发线程任务，所述输入数据为源输入数据或缓存交换数据；

模块线程调度模块，设置为根据触发的至少两个线程任务，并行调度对应的至少两个模块线程，对所述输入数据进行处理，产生处理结果数据；其中，所述至少两个模块线程与依据神经网络中的网络层进行划分的至少两个网络模块分别对应；

处理结果数据输出模块，设置为将所述处理结果数据输出至缓存，以作为除所述至少两个模块线程之外的模块线程的缓存交换数据，或，输出所述处理结果数据，以作为源输入数据的处理结果。

11、一种设备，包括：处理器和存储器；

所述存储器中存储有至少一条指令，所述指令由所述处理器执行，使得所述设备执行如权利要求 1 至 9 任一项所述的基于神经网络的任务处理方法。

12、一种计算机可读存储介质，所述存储介质中的指令由设备的处理器执

行时，使得设备能够执行如权利要求 1 至 9 任一项所述的基于神经网络的任务处理方法。