(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105389772 B
(45)授权公告日 2018.09.07

(21)申请号 201510876116.6

(22)申请日 2015.12.02

(65)同一申请的已公布的文献号
申请公布号 CN 105389772 A

(43)申请公布日 2016.03.09

(73)专利权人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦三层

(72)发明人 胡娜 付晓寅 王桂彬

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 宋合成

(51)Int.Cl.
G06T 1/20(2006.01)
G06F 9/38(2006.01)

(56)对比文件
CN 104680235 A,2015.06.03,
CN 104732274 A,2015.06.24,
US 2014201126 A1,2014.07.17,
US 9015093 B1,2015.04.21,

审查员 张禹

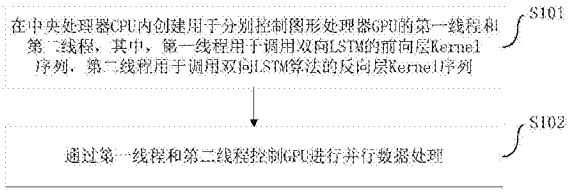
权利要求书2页 说明书8页 附图4页

(54)发明名称

基于图形处理器的数据处理方法和装置

(57)摘要

本发明提出一种基于图形处理器的数据处理方法和装置。其中,该数据处理方法包括:在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程,其中,所述第一线程用于调用双向LSTM算法的前向层Kernel序列,所述第二线程用于调用双向LSTM算法的反向层Kernel序列;通过所述第一线程和所述第二线程控制所述GPU进行并行数据处理。本发明实施例的数据处理方法,有效的提高了GPU的执行效率,缩短了LSTM的计算过程的执行时间。



CN 105389772 B

1.一种基于图形处理器的数据处理方法，其特征在于，包括以下步骤：

在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，所述第一线程用于调用双向LSTM的前向层Kernel序列，所述第二线程用于调用双向LSTM算法的反向层Kernel序列；

通过所述第一线程和所述第二线程控制所述GPU进行并行数据处理。

2.如权利要求1所述的数据处理方法，其特征在于，通过所述第一线程和所述第二线程控制所述图形处理器进行并行数据处理，包括：

将所述前向层Kernel序列和所述反向层Kernel序列分别派发至所述GPU的两条数据流中，以使所述GPU并行执行所述前向层Kernel序列和所述反向层Kernel序列。

3.如权利要求1或2所述的数据处理方法，其特征在于，所述前向层Kernel序列和所述反向层Kernel序列包括多个Kernel程序，所述处理方法还包括：

分别获取所述双向LSTM的前向层和反向层计算过程中的多个矩阵单元；

将至少两个无数据处理相关性的矩阵单元合并为一个，并应用一个所述Kernel程序处理合并后的矩阵单元。

4.如权利要求3所述的数据处理方法，其特征在于，每个Kernel程序中包括多个Kernel计算过程，所述处理方法还包括：

针对有数据处理相关性的矩阵单元，应用一个所述Kernel计算过程处理每个矩阵单元中至少两个无数据处理相关性的元素。

5.如权利要求3所述的数据处理方法，其特征在于，所述双向LSTM的前向层和反向层计算过程中的多个矩阵单元包括输入门矩阵、输出门矩阵、遗忘门矩阵和CELL矩阵。

6.一种基于图形处理器的数据处理装置，其特征在于，包括：

创建模块，用于在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，所述第一线程用于调用双向LSTM的前向层Kernel序列，所述第二线程用于调用双向LSTM算法的反向层Kernel序列；

第一处理模块，用于通过所述第一线程和所述第二线程控制所述GPU进行并行数据处理。

7.如权利要求6所述的数据处理装置，其特征在于，所述第一处理模块还用于：

将所述前向层Kernel序列和所述反向层Kernel序列分别派发至所述GPU的两条数据流中，以使所述GPU并行执行所述前向层Kernel序列和所述反向层Kernel序列。

8.如权利要求6或7所述的数据处理装置，其特征在于，所述前向层Kernel序列和所述反向层Kernel序列包括多个Kernel程序，所述处理装置还包括：

获取模块，用于分别获取所述双向LSTM的前向层和反向层计算过程中的多个矩阵单元；

第二处理模块，用于将至少两个无数据处理相关性的矩阵单元合并为一个，并应用一个所述Kernel程序处理合并后的矩阵单元。

9.如权利要求8所述的数据处理装置，其特征在于，每个Kernel程序中包括多个Kernel计算过程，所述处理装置还包括：

第三处理模块，用于针对有数据处理相关性的矩阵单元，应用一个所述Kernel计算过程处理每个矩阵单元中至少两个无数据处理相关性的元素。

10.如权利要求8所述的数据处理装置，其特征在于，所述双向LSTM的前向层和反向层计算过程中的多个矩阵单元包括输入门矩阵、输出门矩阵、遗忘门矩阵和CELL矩阵。

基于图形处理器的数据处理方法和装置

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种基于图形处理器的数据处理方法和装置。

背景技术

[0002] 长短期记忆人工神经网络(Long-Short Term Memory，LSTM)是一种时间递归神经网络，适于处理和预测时间序列中间隔和延迟非常长的重要事件。双向LSTM从历史和未来两个方向学习输入特征，具有更高的识别精度，然而双向LSTM引入了更大的计算量，增大了模型训练的时间。

[0003] 当前，GPU(Graphics Processing Unit，图形处理器)已经成为深度学习平台广泛使用的加速部件，支持GPU加速计算典型的深度学习平台有MXNet、Kaldi、TensorFlow、Nervana等。其中，MXNet、Kaldi、TensorFlow都提供了双向LSTM的算法实现，其GPU线性代数库大多采用NVidia提供的cuBLAS库。而与前三者不同的是，Nervana的目的是构建一套跨平台的线性代数库。

[0004] 然而，目前存在的问题是，采用逐帧递推方式的双向LSTM的算法包含大量细粒度计算过程，而GPU在细粒度计算中难以充分发挥其海量计算资源优势，而且GPU的调用具有不可忽略的运行开销，因此存在GPU利用率低的问题。

发明内容

[0005] 本发明旨在至少在一定程度上解决相关技术中的技术问题之一。

[0006] 为此，本发明的第一个目的在于提出一种基于图形处理器的数据处理方法，该数据处理方法有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0007] 本发明的第二个目的在于提出一种基于图形处理器的数据处理装置。

[0008] 为达上述目的，本发明第一方面实施例提出了一种基于图形处理器的数据处理方法，包括：在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，所述第一线程用于调用双向LSTM的前向层Kernel序列，所述第二线程用于调用双向LSTM算法的反向层Kernel序列；通过所述第一线程和所述第二线程控制所述GPU进行并行数据处理。

[0009] 本发明实施例的基于图形处理器的数据处理方法，通过将双向LSTM的前向层和反向层的计算过程分派在GPU的两条数据流中，结合GPU体系的结构特点对双向LSTM的计算过程进行加速优化，从而有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0010] 为达上述目的，本发明第二方面实施例提出了一种基于图形处理器的数据处理装置，包括：创建模块，用于在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，所述第一线程用于调用双向LSTM的前向层Kernel序列，所述第二线程用于调用双向LSTM算法的反向层Kernel序列；第一处理模块，用于通过所述第一线程和所述第二线程控制所述GPU进行并行数据处理。

[0011] 本发明实施例的基于图形处理器的数据处理装置，通过将双向LSTM的前向层和反向层的计算过程分派在GPU的两条数据流中，结合GPU体系的结构特点对双向LSTM的计算过程进行加速优化，从而有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0012] 本发明附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0013] 本发明上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

[0014] 图1是本发明一个实施例的基于图形处理器的数据处理方法的流程图；

[0015] 图2是本发明一个具体实施例的基于图形处理器的数据处理方法的流程图；

[0016] 图3是本发明另一个具体实施例的基于图形处理器的数据处理方法的流程图；

[0017] 图4是本发明一个实施例的LSTM的计算过程的优化流程图；

[0018] 图5是本发明一个实施例的基于图形处理器的数据处理装置的结构示意图；

[0019] 图6是本发明一个具体实施例的基于图形处理器的数据处理装置的结构示意图；

[0020] 图7是本发明另一个具体实施例的基于图形处理器的数据处理装置的结构示意图。

具体实施方式

[0021] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，旨在用于解释本发明，而不能理解为对本发明的限制。

[0022] 此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。在本发明的描述中，“多个”的含义是两个或两个以上，除非另有明确具体的限定。

[0023] 流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本发明的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本发明的实施例所属技术领域的技术人员所理解。

[0024] 图1是本发明一个实施例的基于图形处理器的数据处理方法的流程图。

[0025] 如图1所示，基于图形处理器的数据处理方法包括：

[0026] S101，在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，第一线程用于调用双向LSTM的前向层Kernel序列，第二线程用于调用双向LSTM算法的反向层Kernel序列。

[0027] 在本发明的一个实施例中将前向层Kernel序列和反向层Kernel序列分别派发至GPU的两条数据流中，以使GPU并行执行前向层Kernel序列和反向层Kernel序列。

[0028] 具体地，双向LSTM的算法中，前向层和反向层的计算过程是相互独立的，因此可以

利用GPU加速部件的硬件支持，同一个GPU可以并发执行前向层和反向层的计算过程。具体而言，在主机的CPU上派生两个线程，即第一线程和第二线程，使用同一个GPU上两条不同的数据流，将前向层和反向层的Kernel序列分别派发在两条数据流中，以使GPU的硬件完成Kernel序列的调度过程。换言之，在对双向LSTM计算过程的优化中，首先以较大的优化粒度对LSTM的计算过程进行优化，判断LSTM的计算过程中是否存在可以并发执行的Kernel序列，例如LSTM的前向层Kernel序列和反向层Kernel序列，基于CUDA(Compute Unified Device Architecture，一种由NVIDIA推出的通用并行计算架构)提供的流机制，将并发的前向层Kernel序列和反向层Kernel序列分派之GPU的不同的数据流中，使之并发执行。

[0029] S102，通过第一线程和第二线程控制GPU进行并行数据处理。

[0030] 本发明实施例的基于图形处理器的数据处理方法，通过将双向LSTM的前向层和反向层的计算过程分派在GPU的两条数据流中，结合GPU体系的结构特点对双向LSTM的计算过程进行加速优化，从而有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0031] 图2是本发明一个具体实施例的基于图形处理器的数据处理方法的流程图。

[0032] 如图2所示，基于图形处理器的数据处理方法包括：

[0033] S201，在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，第一线程用于调用双向LSTM的前向层Kernel序列，第二线程用于调用双向LSTM算法的反向层Kernel序列。

[0034] 在本发明的一个实施例中将前向层Kernel序列和反向层Kernel序列分别派发至GPU的两条数据流中，以使GPU并行执行前向层Kernel序列和反向层Kernel序列。

[0035] 具体地，双向LSTM的算法中，前向层和反向层的计算过程是相互独立的，因此可以利用GPU加速部件的硬件支持，同一个GPU可以并发执行前向层和反向层的计算过程。具体而言，在主机的CPU上派生两个线程，即第一线程和第二线程，使用同一个GPU上两条不同的数据流，将前向层和反向层的Kernel序列分别派发在两条数据流中，以使GPU的硬件完成Kernel序列的调度过程。换言之，在对双向LSTM计算过程的优化中，首先以较大的优化粒度对LSTM的计算过程进行优化，判断LSTM的计算过程中是否存在可以并发执行的Kernel序列，例如LSTM的前向层Kernel序列和反向层Kernel序列，基于CUDA(Compute Unified Device Architecture，一种由NVIDIA推出的通用并行计算架构)提供的流机制，将并发的前向层Kernel序列和反向层Kernel序列分派之GPU的不同的数据流中，使之并发执行。

[0036] S202，通过第一线程和第二线程控制GPU进行并行数据处理。

[0037] S203，分别获取双向LSTM的前向层和反向层计算过程中的多个矩阵单元。

[0038] 其中，Kernel序列中包括多个Kernel程序，每个Kernel程序用于计算双向LSTM的前向层中的多个矩阵单元，或者用于计算双向LSTM的后向层中的多个矩阵单元。具体而言，双向LSTM算法的正向层和反向层计算过程中，均包含对输入门矩阵、输出门矩阵、遗忘门矩阵和CELL矩阵的计算过程，这些矩阵的计算过程之间有些有数据处理相关性，有些没有数据处理相关性。其中，无数据处理相关性是指矩阵的计算过程不依赖于其它矩阵的计算结果，例如，前向层中包含Ka、Kb和Kc三个矩阵的计算过程，如果Kb的计算过程依赖于Ka的计算结果，则表示Ka和Kb的计算过程存在数据处理相关性，而如果Kc的计算过程不依赖于Kb的计算结果，则表示Kb和Kc的计算过程不存在数据处理相关性。因此，对于无数据处理相关性的矩阵，可以并行执行两个或者两个以上矩阵的计算过程。

[0039] S204，将至少两个无数据处理相关性的矩阵单元合并为一个，并应用一个Kernel程序处理合并后的矩阵单元。

[0040] 具体地，将多个矩阵单元中的两个或者两个以上的无数据处理相关性的矩阵单元合并为一个，例如，原有的两个矩阵分别是100*100的矩阵，若判断这两个矩阵无数据处理相关性，则将这两个矩阵合并为一个100*200的矩阵，应用Kernel程序处理合并后的矩阵单元。换言之，将无数据处理相关性的矩阵单元对应的Kernel程序合并成一个Kernel程序，应用合并后的Kernel程序，利用GPU的硬件多线程机制完成对无数据处理相关性的矩阵的计算过程。

[0041] 应当理解的是，本实施例中进一步以较小的优化粒度对同一数据流中的Kernel序列进行优化，针对GPU的同一数据流中的Kernel序列中，判断是否存在无数据处理相关性的多个Kernel程序，若存在则将无数据处理相关性的多个Kernel程序进行合并。

[0042] 本发明实施例的基于图形处理器的数据处理方法，将无数据处理相关性的多个矩阵单元对应的Kernel程序合并为一个Kernel程序，通过GPU完成合并后的Kernel程序的计算过程，从而增大了GPU的计算粒度，减少了GPU的调用次数，有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0043] 图3是本发明另一个具体实施例的基于图形处理器的数据处理方法的流程图。

[0044] 如图3所示，基于图形处理器的数据处理方法包括：

[0045] S301，在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，第一线程用于调用双向LSTM的前向层Kernel序列，第二线程用于调用双向LSTM算法的反向层Kernel序列。

[0046] 在本发明的一个实施例中将前向层Kernel序列和反向层Kernel序列分别派发至GPU的两条数据流中，以使GPU并行执行前向层Kernel序列和反向层Kernel序列。

[0047] 具体地，双向LSTM的算法中，前向层和反向层的计算过程是相互独立的，因此可以利用GPU加速部件的硬件支持，同一个GPU可以并发执行前向层和反向层的计算过程。具体而言，在主机的CPU上派生两个线程，即第一线程和第二线程，使用同一个GPU上两条不同的数据流，将前向层和反向层的Kernel序列分别派发在两条数据流中，以使GPU的硬件完成Kernel序列的调度过程。换言之，在对双向LSTM计算过程的优化中，首先以较大的优化粒度对LSTM的计算过程进行优化，判断LSTM的计算过程中是否存在可以并发执行的Kernel序列，例如LSTM的前向层Kernel序列和反向层Kernel序列，基于CUDA(Compute Unified Device Architecture，一种由NVIDIA推出的通用并行计算架构)提供的流机制，将并发的前向层Kernel序列和反向层Kernel序列分派之GPU的不同的数据流中，使之并发执行。

[0048] S302，通过第一线程和第二线程控制GPU进行并行数据处理。

[0049] S303，分别获取双向LSTM的前向层和反向层计算过程中的多个矩阵单元。

[0050] 其中，Kernel序列中包括多个Kernel程序，每个Kernel程序用于计算双向LSTM的前向层中的多个矩阵单元，或者用于计算双向LSTM的后向层中的多个矩阵单元。具体而言，双向LSTM算法的正向层和反向层计算过程中，均包含对输入门矩阵、输出门矩阵、遗忘门矩阵和CELL矩阵的计算过程，这些矩阵的计算过程之间有些有数据处理相关性，有些没有数据处理相关性。其中，无数据处理相关性是指矩阵的计算过程不依赖于其它矩阵的计算结果，例如，前向层中包含Ka、Kb和Kc三个矩阵的计算过程，如果Kb的计算过程依赖于Ka的计

算结果，则表示Ka和Kb的计算过程存在数据处理相关性，而如果Kc的计算过程不依赖于Kb的计算结果，则表示Kb和Kc的计算过程不存在数据处理相关性。因此，对于无数据处理相关性的矩阵，可以并行执行两个或者两个以上矩阵的计算过程。

[0051] S304，将至少两个无数据处理相关性的矩阵单元合并为一个，并应用一个Kernel程序处理合并后的矩阵单元。

[0052] 具体地，将多个矩阵单元中的两个或者两个以上的无数据处理相关性的矩阵单元合并为一个，例如，原有的两个矩阵分别是100*100的矩阵，若判断这两个矩阵无数据处理相关性，则将这两个矩阵合并为一个100*200的矩阵，应用Kernel程序处理合并后的矩阵单元。换言之，将无数据处理相关性的矩阵单元对应的Kernel程序合并成一个Kernel程序，应用合并后的Kernel程序，利用GPU的硬件多线程机制完成对无数据处理相关性的矩阵的计算过程。

[0053] S305，针对有数据处理相关性的矩阵单元，应用一个Kernel计算过程处理每个矩阵单元中至少两个无数据处理相关性的元素。

[0054] 其中，每个Kernel程序中包括多个Kernel计算过程，每个Kernel计算过程用于一个矩阵单元中一个元素的计算过程。具体而言，在双向LSTM的矩阵计算过程中存在大量的元素级操作，本实施例中进一步以最小的优化粒度对Kernel序列的计算过程进行优化，对于矩阵单元中的多个元素，将两个或者两个以上的无数据处理相关性的元素的对应的Kernel计算过程合并为一个计算过程。

[0055] 进而，将多个针对元素的Kernel计算过程合并为一个Kernel计算过程，再将多个Kernel计算过程合并为一个Kernel程序，在CPU创建的线程内顺序地完成有数据处理相关性的矩阵单元的计算过程。

[0056] 本发明实施例的基于图形处理器的数据处理方法，将无数据处理相关性的矩阵单元的Kernel程序中的Kernel计算过程合并为一个Kernel计算过程，进而将多个Kernel计算过程合并为一个Kernel程序，通过GPU完成合并后的Kernel程序的计算过程，从而减少了对GPU外部存储器的访问次数，减少了GPU的调用次数，有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0057] 应当理解的是，如图4所示，图4中示出了应用本发明优化方法的具体应用流程，按照自上而下的步骤展开，其优化粒度逐渐变小。首先，判断双向LTSM的计算过程中是否存在可并发执行的Kernel序列，通过CPU创建的不同线程将Kernel序列分派至不同的流中，使之并发执行。其次，在同一数据流内的Kernel序列中，判断是否存在无数据处理相关性的多个Kernel程序，如果存在无数据处理相关性的多个Kernel程序，则应用图2实施例中的方法将无数据处理相关性的多个Kernel程序合并为一个Kernel程序。最后，针对同一数据流内并且存在数据处理相关性的Kernel程序，判断每个Kernel程序中的元素级的计算过程是否满足合并条件，如果满足合并条件则应用图3实施例中的方法将多个元素的Kernel计算过程合并为一个Kernel计算过程，进而合并多个Kernel计算过程为一个Kernel程序。针对双向LTSM的计算过程，基于深度学习的声学训练过程计算量和数据量都很大，模型训练时间极大地制约了新技术新方法的验证周期，基于本发明的双向LTSM的计算过程的优化方法，可以有效缩短模型的训练时间，降低研发成本。此外，语音识别的速度直接影响用户的体验，基于本发明的双向LTSM的计算过程的优化方法，可以有效地缩短语音识别的延迟，提高线

上语音识别的速度。另外，本发明的双向LTSM的计算过程的优化方法还可以作为普适方法用于其他深度学习框架中。

[0058] 为了实现上述实施例，本发明还提出一种基于图形处理器的数据处理装置。

[0059] 图5是本发明一个实施例的基于图形处理器的数据处理装置的结构示意图。

[0060] 如图5所示，基于图形处理器的数据处理装置包括：创建模块100和第一处理模块200。

[0061] 其中，创建模块100用于在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，第一线程用于调用双向LSTM的前向层Kernel序列，第二线程用于调用双向LSTM算法的反向层Kernel序列。第一处理模块200用于通过第一线程和第二线程控制GPU进行并行数据处理。

[0062] 其中，第一处理模块200还用于将前向层Kernel序列和反向层Kernel序列分别派发至GPU的两条数据流中，以使GPU并行执行前向层Kernel序列和反向层Kernel序列。具体地，双向LSTM的算法中，前向层和反向层的计算过程是相互独立的，因此可以利用GPU加速部件的硬件支持，同一个GPU可以并发执行前向层和反向层的计算过程。具体而言，创建模块100在主机的CPU上派生两个线程，即第一线程和第二线程，第一处理模块200使用同一个GPU上两条不同的数据流，将前向层和反向层的Kernel序列分别派发在两条数据流中，以使GPU的硬件完成Kernel序列的调度过程。换言之，在对双向LSTM计算过程的优化中，首先以较大的优化粒度对LSTM的计算过程进行优化，判断LSTM的计算过程中是否存在可以并发执行的Kernel序列，例如LSTM的前向层Kernel序列和反向层Kernel序列，基于CUDA(Compute Unified Device Architecture，一种由NVIDIA推出的通用并行计算架构)提供的流机制，将并发的前向层Kernel序列和反向层Kernel序列分派之GPU的不同的数据流中，使之并发执行。

[0063] 本发明实施例的基于图形处理器的数据处理装置，通过将双向LSTM的前向层和反向层的计算过程分派在GPU的两条数据流中，结合GPU体系的结构特点对双向LSTM的计算过程进行加速优化，从而有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0064] 图6是本发明一个具体实施例的基于图形处理器的数据处理装置的结构示意图。

[0065] 如图6所示，基于图形处理器的数据处理装置包括：创建模块100、第一处理模块200、获取模块300和第二处理模块400。

[0066] 其中，获取模块300用于分别获取双向LSTM的前向层和反向层计算过程中的多个矩阵单元。第二处理模块400用于将至少两个无数据处理相关性的矩阵单元合并为一个，并应用一个Kernel程序处理合并后的矩阵单元。其中，Kernel序列中包括多个Kernel程序，每个Kernel程序用于计算双向LSTM的前向层中的多个矩阵单元，或者用于计算双向LSTM的后向层中的多个矩阵单元。具体而言，双向LSTM算法的正向层和反向层计算过程中，均包含对输入门矩阵、输出门矩阵、遗忘门矩阵和CELL矩阵的计算过程，这些矩阵的计算过程之间有些有数据处理相关性，有些没有数据处理相关性。其中，无数据处理相关性是指矩阵的计算过程不依赖于其它矩阵的计算结果，例如，前向层中包含Ka、Kb和Kc三个矩阵的计算过程，如果Kb的计算过程依赖于Ka的计算结果，则表示Ka和Kb的计算过程存在数据处理相关性，而如果Kc的计算过程不依赖于Kb的计算结果，则表示Kb和Kc的计算过程不存在数据处理相关性。因此，对于无数据处理相关性的矩阵，第二处理模块400可以并行执行两个或者两个

以上矩阵的计算过程。

[0067] 具体地，第二处理模块400将多个矩阵单元中的两个或者两个以上的无数据处理相关性的矩阵单元合并为一个，例如，原有的两个矩阵分别是100*100的矩阵，若判断这两个矩阵无数据处理相关性，则将这两个矩阵合并为一个100*200的矩阵，应用Kernel程序处理合并后的矩阵单元。换言之，第二处理模块400将无数据处理相关性的矩阵单元对应的Kernel程序合并成一个Kernel程序，应用合并后的Kernel程序，利用GPU的硬件多线程机制完成对无数据处理相关性的矩阵的计算过程。

[0068] 本发明实施例的基于图形处理器的数据处理装置，将无数据处理相关性的多个矩阵单元对应的Kernel程序合并为一个Kernel程序，通过GPU完成合并后的Kernel程序的计算过程，从而增大了GPU的计算粒度，减少了GPU的调用次数，有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0069] 图7是本发明另一个具体实施例的基于图形处理器的数据处理装置的结构示意图。

[0070] 如图7所示，基于图形处理器的数据处理装置包括：创建模块100、第一处理模块200、获取模块300、第二处理模块400、第三处理模块500。

[0071] 其中，第三处理模块500用于针对有数据处理相关性的矩阵单元，应用一个Kernel计算过程处理每个矩阵单元中至少两个无数据处理相关性的元素。其中，每个Kernel程序中包括多个Kernel计算过程，每个Kernel计算过程用于一个矩阵单元中一个元素的计算过程。具体而言，在双向LSTM的矩阵计算过程中存在大量的元素级操作，本实施例中进一步以最小的优化粒度对Kernel序列的计算过程进行优化，对于矩阵单元中的多个元素，第三处理模块500将两个或者两个以上的无数据处理相关性的元素的对应的Kernel计算过程合并为一个计算过程。

[0072] 进而，第三处理模块500将多个针对元素的Kernel计算过程合并为一个Kernel计算过程，再将多个Kernel计算过程合并为一个Kernel程序，在CPU创建的线程内顺序地完成有数据处理相关性的矩阵单元的计算过程。

[0073] 本发明实施例的基于图形处理器的数据处理装置，将无数据处理相关性的矩阵单元的Kernel程序中的Kernel计算过程合并为一个Kernel计算过程，进而将多个Kernel计算过程合并为一个Kernel程序，通过GPU完成合并后的Kernel程序的计算过程，从而减少了对GPU外部存储器的访问次数，减少了GPU的调用次数，有效的提高了GPU的执行效率，缩短了LSTM的计算过程的执行时间。

[0074] 应当理解，本发明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系统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(PGA)，现场可编程门阵列(FPGA)等。

[0075] 在本发明中，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”、等术语应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或成一体；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连

通或两个元件的相互作用关系，除非另有明确的限定。对于本领域的普通技术人员而言，可以根据具体情况理解上述术语在本发明中的具体含义。

[0076] 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发明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必须针对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任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0077] 尽管上面已经示出和描述了本发明的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本发明的限制，本领域的普通技术人员在本发明的范围内可以对上述实施例进行变化、修改、替换和变型。

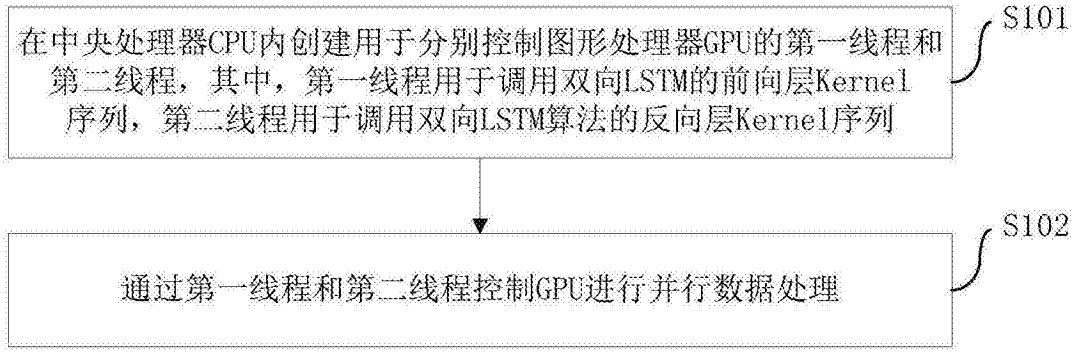

在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，第一线程用于调用双向LSTM的前向层Kernel序列，第二线程用于调用双向LSTM算法的反向层Kernel序列
S101
通过第一线程和第二线程控制GPU进行并行数据处理
S102

图1

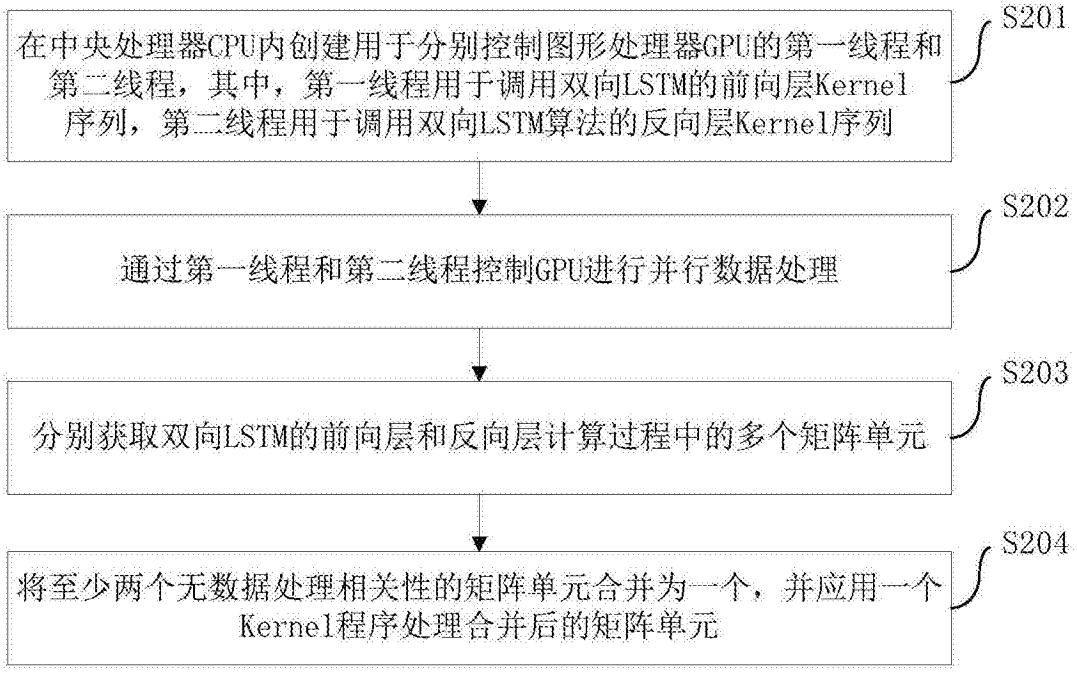

在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，第一线程用于调用双向LSTM的前向层Kernel序列，第二线程用于调用双向LSTM算法的反向层Kernel序列
S201
通过第一线程和第二线程控制GPU进行并行数据处理
S202
分别获取双向LSTM的前向层和反向层计算过程中的多个矩阵单元
S203
将至少两个无数据处理相关性的矩阵单元合并为一个，并应用一个Kernel程序处理合并后的矩阵单元
S204

图2


在中央处理器CPU内创建用于分别控制图形处理器GPU的第一线程和第二线程，其中，第一线程用于调用双向LSTM的前向层Kernel序列，第二线程用于调用双向LSTM算法的反向层Kernel序列
S301
通过第一线程和第二线程控制GPU进行并行数据处理
S302
分别获取双向LSTM的前向层和反向层计算过程中的多个矩阵单元
S303
将至少两个无数据处理相关性的矩阵单元合并为一个，并应用一个Kernel程序处理合并后的矩阵单元
S304
针对有数据处理相关性的矩阵单元，应用一个Kernel计算过程处理每个矩阵单元中至少两个无数据处理相关性的元素
S305

图3

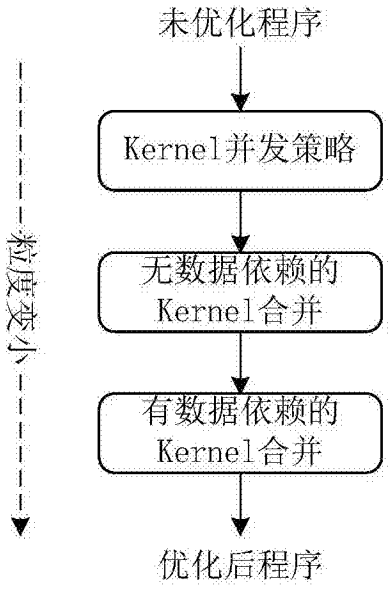

未优化程序
Kernel并发策略
无数据依赖的Kernel合并
有数据依赖的Kernel合并
优化后程序
粒度变小

图4

图5

图6

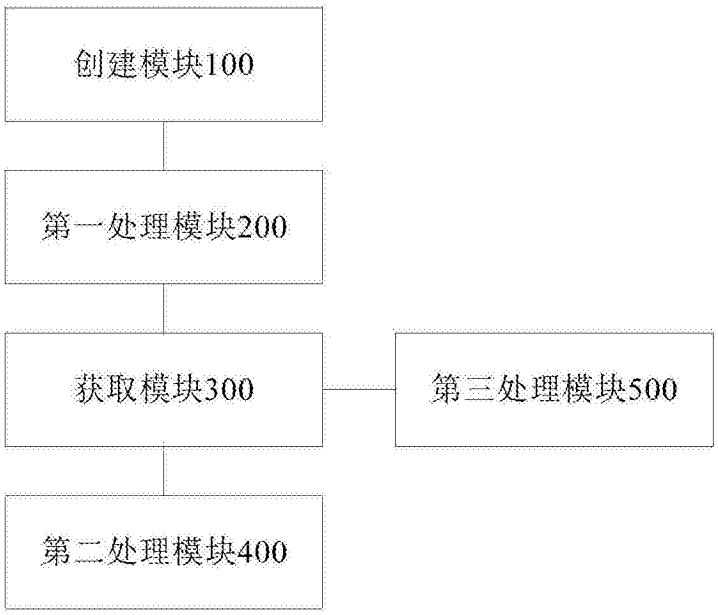


图7