

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 4 月 4 日 (04.04.2024)



(10) 国际公布号
WO 2024/065865 A1

(51) 国际专利分类号:
G06F 9/50 (2006.01) **G06N 3/04** (2023.01)
G06F 12/02 (2006.01)

中国浙江省杭州市余杭区中泰街道科创大道之江实验室南湖总部, Zhejiang 311121 (CN)。

(21) 国际申请号: PCT/CN2022/124000

(22) 国际申请日: 2022 年 10 月 9 日 (09.10.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211177786.5 2022年9月27日 (27.09.2022) CN

(71) 申请人: 之江实验室(ZHEJIANG LAB) [CN/CN]; 中国浙江省杭州市余杭区中泰街道科创大道之江实验室南湖总部, Zhejiang 311121 (CN)。

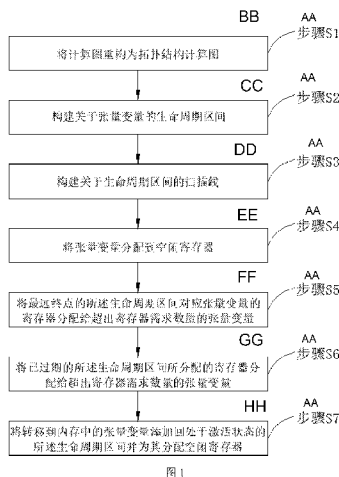
(72) 发明人: 王宏升(WANG, Hongsheng); 中国浙江省杭州市余杭区中泰街道科创大道之江实验室南湖总部, Zhejiang 311121 (CN)。 陈光(CHEN, Guang);

(74) 代理人: 北京志霖恒远知识产权代理有限公司(BEIJING ZHILIN HENGYUAN INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市东城区北三环东路36号环球贸易中心C座2001-2007室, Beijing 100013 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE,

(54) Title: MEMORY OPTIMIZATION METHOD AND APPARATUS FOR NEURAL NETWORK CALCULATION

(54) 发明名称: 一种面向神经网络计算的内存优化方法和装置



AA Step
BB Reconstruct a calculation graph into a topological structure calculation graph
CC Construct life cycle intervals regarding tensor variables
DD Construct a scanning line regarding the life cycle intervals
EE Allocate the tensor variables to idle registers
FF Allocate, to a tensor variable which exceeds a number required by a register, a register for a tensor variable corresponding to a life cycle interval at the farthest endpoint
GG Allocate, to the tensor variable which exceeds the number required by the register, a register which is allocated to an expired life cycle interval
HH Add, back to a life cycle interval in an active state, a tensor variable which is transferred into a memory, and allocate an idle register to the tensor variable which is transferred into the memory

(57) Abstract: Disclosed in the present invention are a memory optimization method and apparatus for neural network calculation. The method comprises the following steps: S1, reconstructing a calculation graph into a topological structure calculation graph; S2, constructing life cycle intervals regarding tensor variables; S3, constructing a scanning line regarding the life cycle intervals; S4, allocating the tensor variables to idle registers; S5, allocating, to a tensor variable which exceeds a number required by a register, a register for a tensor variable corresponding to a life cycle interval at the farthest endpoint; S6, allocating, to the tensor variable which exceeds the number required by the register, a register which is allocated to an expired life cycle interval; and S7, adding, back to a life cycle interval in an active state, a tensor variable which is transferred into a memory, and allocating an idle register to the tensor variable which is transferred into the memory. By means of the present invention, a memory for a data stream of a calculation graph for neural network calculation is optimized, thereby reducing memory overheads required by tensor variables in the data stream, and lowering requirements of a large model for memory resources of hardware.

SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA,
UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA,
RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本发明公开了一种面向神经网络计算的内存优化方法和装置, 包括以下步骤: 步骤S1: 将计算图重
构为拓扑结构计算图; 步骤S2: 构建关于张量变量的生命周期区间; 步骤S3: 构建关于生命周期区间的扫描
线; 步骤S4: 将张量变量分配到空闲寄存器; 步骤S5: 分配给超出寄存器需求数量的张量变量; 步骤S6: 将已
过期的所述生命周期区间所分配的寄存器分配给超出寄存器需求数量的张量变量; 步骤S7: 将转移到内存中的
张量变量添加回处于激活状态的所述生命周期区间并为其分配空闲寄存器。本发明优化用于神经网络计算的计
算图的数据流的内存, 减少数据流中张量变量所需的内存开销, 降低大模型对于硬件内存资源的要求。

一种面向神经网络计算的内存优化方法和装置

本申请要求于 2022 年 9 月 27 日向中国国家知识产权局提交的发明专利申请号为 202211177786.5，发明名称为“一种面向神经网络计算的内存优化方法和装置”的中国专利申请的优先权权益，其全部内容通过引用合并于本申请。

技术领域

[0001] 本发明涉及一种基于特定计算模型的计算机系统技术领域，尤其涉及一种面向神经网络计算的内存优化方法和装置。

背景技术

[0002] 随着工业界复杂场景对大规模神经网络应用的需求越来越紧迫，大模型对于内存空间的占用不断增大，人工智能硬件操作系统自身的内存资源不能满足大模型训练对于内存的需求，所以优化面向神经网络计算的内存技术变得极为重要。

[0003] 为此，我们提出一种面向神经网络计算的内存优化方法和装置。

发明内容

[0004] 本发明的目的在于提供一种面向神经网络计算的内存优化方法和装置，以解决如何优化减少张量变量对于深度学习操作系统内存资源的持久依赖和占用，减少数据流中张量变量所需的内存开销，并降低大模型对于硬件内存资源的要求的问题。

[0005] 本发明采用的技术方案如下：

一种面向神经网络计算的内存优化方法，包括以下步骤：

步骤 S1：将计算图重构为拓扑结构计算图；

步骤 S2：构建关于张量变量的生命周期区间；

步骤 S3：构建关于生命周期区间的扫描线；

步骤 S4：将张量变量分配到空闲寄存器；

步骤 S5：将最远终点的所述生命周期区间对应张量变量的寄存器分配给超出寄存器需求数量的张量变量；

步骤 S6：将已过期的所述生命周期区间所分配的寄存器分配给超出寄存器需求数量的张量变量；

步骤 S7：将转移到内存中的张量变量添加回处于激活状态的所述生命周期区间并为其分配空闲寄存器。

[0006] 进一步地，所述步骤 S1 具体包括以下子步骤：

步骤 S11: 后序顺序遍历计算图, 得到子图访问列表;

步骤 S12: 逆序后序所述子图访问列表, 得到计算图的拓扑结构顺序;

步骤 S13: 根据所述拓扑结构顺序重构计算图, 得到拓扑结构计算图。

[0007] 进一步地, 所述后序顺序为当访问所述计算图的某个节点时, 则优先递归地访问所述节点的后继节点。

[0008] 进一步地, 所述步骤 S2 具体为构建关于每个节点中包含张量变量的生命周期区间, 所述节点包含的张量变量对应的关于生命周期区间起始于所述张量变量处于存活状态的第一个节点的位置, 终止于所述张量变量处于存活状态的最后一个节点的位置。

[0009] 进一步地, 所述步骤 S3 具体为在所述拓扑结构计算图的起始节点处, 构建一条与所述生命周期区间平行的扫描线, 所述扫描线用于从所述生命周期区间的起始端向所述生命周期区间的终止端移动过程中, 观察是否存在空闲寄存器可以分配给数据流执行过程中的张量变量。

[0010] 进一步地, 所述步骤 S5 具体为执行流位于某个节点的位置时, 当所述节点既没有空闲寄存器, 也没有可以从处于激活状态的所述生命周期区间中移除的已经被扫描过期的所述生命周期区间, 则将最远终点的所述生命周期区间对应的张量变量所分配的寄存器中的张量变量转移到内存中, 然后将所释放的寄存器分配给超出寄存器需求数量的张量变量。

[0011] 进一步地, 所述步骤 S6 具体为执行流位于某个节点的位置时, 当所述扫描线已经穿过张量变量所分配寄存器对应的所述生命周期区间, 将张量变量从处于激活状态的所述生命周期区间中移除, 对应分配的寄存器回收到空闲寄存器列表中, 将所述空闲寄存器分配给超出寄存器需求数量的张量变量。

[0012] 进一步地, 所述步骤 S7 具体为执行流位于某个节点的位置时, 当存在空闲寄存器, 则将转移到内存中的张量变量添加回处于激活状态的所述生命周期区间, 并且将空闲寄存器分配给对应的所述生命周期区间。

[0013] 本发明还提供一种面向神经网络计算的内存优化装置, 包括存储器和一个或多个处理器, 所述存储器中存储有可执行代码, 所述一个或多个处理器执行所述可执行代码时, 用于实现上述实施例任一项所述的一种面向神经网络计算的内存优化方法。

[0014] 本发明还提供一种计算机可读存储介质, 其上存储有程序, 该程序被处理器执行时, 实现上述实施例任一项所述的一种面向神经网络计算的内存优化方法。

[0015] 本发明的有益效果是: 本发明提供了计算图执行过程中产生的张量变量与物理寄存器和内存的映射关系, 并且提供了基于所述映射关系的优化方法。寄存器可以存储计算图执行

过程中产生的张量变量在内存中的存储位置。传统的张量变量存储方法是直接将张量变量的值存储到内存中。由于张量变量的值既可以存储在内存中也可以存储在寄存器中，考虑到寄存器允许中央处理器单元直接访问和访问速度快的特性，所以，本发明提出的借助寄存器优化内存的方法优化了用于神经网络计算的计算图的数据流的内存，减少数据流中张量变量所需的内存开销，并降低了大模型对于硬件内存资源的要求。所述的用于神经网络计算的内存优化方法提高整个计算图的计算效率，节约硬件和时间成本。

附图说明

[0016] 图 1 为本发明一种面向神经网络计算的内存优化方法的流程示意图；

图 2 为实施例 1 将计算图重构为拓扑结构的过程示意图；

图 3 为实施例 1 拓扑结构计算图；

图 4 为实施例 1 构建关于拓扑结构计算图节点包含张量变量生命周期的区间；

图 5 为实施例 1 将拓扑结构计算图节点包含的前两个张量变量分配到两个寄存器；

图 6 为实施例 1 将寄存器中张量变量转移到内存并将新张量变量分配到已空闲寄存器；

图 7 为实施例 2 用于神经网络计算的计算图；

图 8 为实施例 2 构建关于数据流中张量变量生命周期区间；

图 9 为实施例 2 构建关于张量变量生命周期区间的扫描线；

图 10 为实施例 2 将寄存器 r_3 分配给节点 V_1 处的变量 x ；

图 11 为实施例 2 将寄存器 r_1 分配给节点 V_2 处的变量 y ；

图 12 为实施例 2 将寄存器 r_2 分配给节点 V_3 处的变量 z ；

图 13 为实施例 2 将最远终点区间 l_x 对应张量变量 x 的寄存器 r_3 分配给超出寄存器需求数量的张量变量 b ；

图 14 为实施例 2 将已过期生命周期区间 l_y 所分配的寄存器 r_1 分配给超出寄存器需求数量的张量变量 w ；

图 15 为实施例 2 将已过期生命周期区间所对应的张量变量从处于激活状态的生命周期区间列表中移除并回收寄存器；

图 16 为实施例 2 将已过期生命周期区间对应的张量变量从处于激活状态的生命周期区间列表中移除并回收寄存器；

图 17 为实施例 2 将空闲寄存器 r_3 分配给 l_{r3} 对应生命周期区间；

图 18 为实施例 3 中的一种面向神经网络计算的内存优化装置的示意图。

具体实施方式

[0017] 以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本发明及其应用或使用的任何限制。基于本发明中的实施例，本领域普通技术人员在没有付出创造性劳动的前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0018] 参见图 1，一种面向神经网络计算的内存优化方法，包括以下步骤：

步骤 S1：将计算图重构为拓扑结构计算图；

步骤 S11：后序顺序遍历计算图，得到子图访问列表；

所述后序顺序为当访问所述计算图的某个节点时，则优先递归地访问所述节点的后继节点。

[0019] 步骤 S12：逆序后序所述子图访问列表，得到计算图的拓扑结构顺序；

步骤 S13：根据所述拓扑结构顺序重构计算图，得到拓扑结构计算图。

[0020] 步骤 S2：构建关于张量变量的生命周期区间；

具体为构建关于每个节点中包含张量变量的生命周期区间，所述节点包含的张量变量对应的关于生命周期区间起始于所述张量变量处于存活状态的第一个节点的位置，终止于所述张量变量处于存活状态的最后一个节点的位置。

[0021] 步骤 S3：构建关于生命周期区间的扫描线；

在所述拓扑结构计算图的起始节点处，构建一条与所述生命周期区间平行的扫描线，所述扫描线用于从所述生命周期区间的起始端向所述生命周期区间的终止端移动过程中，观察是否存在空闲寄存器可以分配给数据流执行过程中的张量变量。

[0022] 步骤 S4：将张量变量分配到空闲寄存器；

步骤 S5：将最远终点的所述生命周期区间对应张量变量的寄存器分配给超出寄存器需求数量的张量变量；

为执行流位于某个节点的位置时，当所述节点既没有空闲寄存器，也没有可以从处于激活状态的所述生命周期区间中移除的已经被扫描过期的所述生命周期区间，则将最远终点的所述生命周期区间对应的张量变量所分配的寄存器中的张量变量转移到内存中，然后将所释放的寄存器分配给超出寄存器需求数量的张量变量。

[0023] 步骤 S6：将已过期的所述生命周期区间所分配的寄存器分配给超出寄存器需求数量的张量变量；

为执行流位于某个节点的位置时，当所述扫描线已经穿过张量变量所分配寄存器对应的所述生命周期区间，将张量变量从处于激活状态的所述生命周期区间中移除，对应分配的寄存器回收到空闲寄存器列表中，将所述空闲寄存器分配给超出寄存器需求数量的张量变量。

[0024] 步骤 S7：将转移到内存中的张量变量添加回处于激活状态的所述生命周期区间并为其

分配空闲寄存器。

[0025] 为执行流位于某个节点的位置时，当存在空闲寄存器，则将转移到内存中的张量变量添加回处于激活状态的所述生命周期区间，并且将空闲寄存器分配给对应的所述生命周期区间。

[0026] 对下述实施例中对应附图的函数定义如下：

$\text{tf.random_uniform}([5,3])$ ：表示随机生成形状为 5 行 3 列的张量。

[0027] $\text{goto } V_i$ ：表示进入执行 V_i 节点的计算流。

[0028] if 表达式 $\text{goto } V_i$ ：表示判断表达式的值是否为真，如果为真，则执行 V_i 节点的计算流；否则执行其他分支节点的计算流。

[0029] $\text{tf.add}(x, y)$ ：表示张量 x 与张量 y 进行相加操作。

[0030] $\text{tf.ones}(a_i.\text{shape})$ ：表示创建一个与张量 a_i 形状相同且所有元素都为 1 的张量。

[0031] $\emptyset(a_i, a_j)$ ：表示张量变量 a_i 和张量变量 a_j 关于张量变量 a 的正确定义的路由选择器。

[0032] $\text{tf.relu}(x)$ ：表示将张量 x 输入整流线性单元。

[0033] $\text{tf.matmul}(x, y)$ ：表示张量 x 与张量 y 进行矩阵乘法操作。

[0034] $\text{return } b_i$ ：表示返回执行包含 b_i 张量变量的分支。

[0035] I_x ：表示张量变量 x 的生命周期区间。

[0036] $\text{tf.subtract}(x, y)$ ：表示张量 x 与张量 y 进行相减操作。

[0037] r_i ：表示将空闲寄存器 r_i 分配给所对应生命周期区间的张量变量。

[0038] s_{r_i} ：表示存储操作，代表将寄存器 r_i 中的张量变量 a_0 存储到内存中。

[0039] l_{r_i} ：表示存储操作，代表将内存中的张量变量 a_0 加载到寄存器 r_i 中。

[0040] 实施例 1：

参见图 2，步骤 S1：将计算图重构为拓扑结构计算图；

步骤 S11：后序顺序遍历计算图，得到子图访问列表；

按照后序顺序遍历计算图，得到子图访问列表为：D，B，E，C，F，A；

所述后序顺序为当访问所述计算图的某个节点时，则优先递归地访问所述节点的后继节点。

[0041] 每当计算图中的某个节点 C 被按照后序顺序被访问完成时, 那么与所述节点 V_C 的所有连边都已经被访问过了。所述后序顺序的遍历可以确保计算图遍历中关于从节点 V_A 指向节点 V_B 的路径中节点 V_B 一定优先于节点 V_A 被访问。

[0042] 步骤 S12: 逆序后序所述子图访问列表, 得到计算图的拓扑结构顺序;
逆序后序所述子图访问列表, 得到计算图的拓扑结构顺序为: A, F, C, E, B, D;
所述逆序后序节点列表是指将第一步后序顺序访问所得节点的列表进行逆序。所述逆序后序节点列表确保如果图中存在从节点 V_A 指向节点 V_B 的路径, 那么所得拓扑顺序的列表中节点 V_A 出现在节点 V_B 之前。所述逆序后序的过程保证了拓扑结构的计算图在访问由某个节点 V_C 连接的任何其他节点之前, 需要优先访问所述的 V_C 节点。

[0043] 步骤 S13: 根据所述拓扑结构顺序重构计算图, 得到拓扑结构计算图, 参见图 3。

[0044] 参见图 4, 步骤 S2: 构建关于张量变量的生命周期区间;

具体为构建关于每个节点中包含张量变量的生命周期区间, 所述节点包含的张量变量对应的关于生命周期区间起始于所述张量变量处于存活状态的第一个节点的位置, 终止于所述张量变量处于存活状态的最后一个节点的位置。

[0045] 对于节点包含的张量变量 v , 所述张量变量对应的关于生命周期区间 l_v 起始于张量变量 v 处于存活状态的第一个节点的位置, 终止于张量变量 v 处于存活状态的最后一个节点的位置。

[0046] 步骤 1: 构建关于张量变量 a_0 的生命周期区间 l_{a_0} , 所述关于张量变量 a_0 的生命周期区间 l_{a_0} 起始于节点 V_1 , 终止于节点 V_3 。

[0047] 步骤 2: 构建关于张量变量 a_1 的生命周期区间 l_{a_1} , 所述关于张量变量 a_1 的生命周期区间 l_{a_1} 起始于节点 V_4 , 由于子图 E 与子图 D 之间存在由子图 E 指向子图 D 的连边, 所以张量变量 a_1 会穿过节点 V_8 到达子图 D, 因此关于张量变量 a_1 的生命周期区间 l_{a_1} 终止于节点 V_8 。

[0048] 步骤 3: 构建关于张量变量 a_2 的生命周期区间 l_{a_2} 。所述关于张量变量 a_2 的生命周期区间 l_{a_2} 起始于节点 V_5 , 由于子图 E 与子图 D 之间存在由子图 E 指向子图 D 的连边, 所

以张量变量 a_2 会穿过节点 V_8 到达子图 D, 因此关于张量变量 a_2 的生命周期区间 I_{a_2} 终止于节点 V_8 。

[0049] 步骤 S3: 构建关于生命周期区间的扫描线;

在所述拓扑结构计算图的起始节点处, 构建一条与所述生命周期区间平行的扫描线, 所述扫描线用于从所述生命周期区间的起始端向所述生命周期区间的终止端移动过程中, 观察是否存在空闲寄存器可以分配给数据流执行过程中的张量变量。

[0050] 参见图 5, 步骤 S4: 将张量变量分配到空闲寄存器;

所述将拓扑结构计算图节点包含的张量变量分配到两个寄存器 r_0 和 r_1 中包含以下过程:

步骤 1: 将张量变量 a_0 分配到寄存器 r_0 中。

[0051] 步骤 2: 将张量变量 a_1 分配到寄存器 r_1 中。

[0052] 步骤 S5: 将最远终点的所述生命周期区间对应张量变量的寄存器分配给超出寄存器需求数量的张量变量;

为执行流位于某个节点 V_i 的位置时, 当所述节点既没有空闲寄存器, 也没有可以从处于激活状态的所述生命周期区间中移除的已经被扫描过期的所述生命周期区间, 则将最远终点的所述生命周期区间对应的张量变量 i 所分配的寄存器 r_i 中的张量变量 i 转移到内存中, 然后将所释放的寄存器 r_i 分配给超出寄存器需求数量的张量变量 j 。

[0053] 步骤 S6: 将已过期的所述生命周期区间 I_i 所分配的寄存器分配给超出寄存器需求数量的张量变量 j ;

为执行流位于某个节点 V_i 的位置时, 当所述扫描线已经穿过张量变量 i 所分配寄存器 r_i 对应的所述生命周期区间 I_i , 将张量变量 i 从处于激活状态的所述生命周期区间中移除, 对应分配的寄存器 r_i 回收到空闲寄存器列表中, 将所述空闲寄存器 r_i 分配给超出寄存器需求数量的张量变量 j 。

[0054] 参见图 6, 步骤 S7: 将转移到内存中的张量变量添加回处于激活状态的所述生命周期区间并为其分配空闲寄存器。

[0055] 为执行流位于某个节点 V_i 的位置时, 当存在空闲寄存器 r_i , 则将转移到内存中的张量变量 i 添加回处于激活状态的所述生命周期区间, 并且将空闲寄存器 r_i 分配给对应的所述生命周期区间 I_i 。

[0056] 每当数据流流经包含张量变量 i 的重定义节点时, 需要将寄存器 r_i 中张量变量 i 存储到内存中; 每当数据流流经包含张量变量 i 的使用节点时, 需要将张量变量 i 从内存中加载到

寄存器 r_1 中。所述将转移到内存中的张量变量添加回处于激活状态的区间列表的过程 I_{r_0} 标记所示位置。

[0057] 第一步、由于节点 V_1 和 V_9 处均包含张量变量 a_0 的定义，所以需要将节点 V_1 和 V_9 处寄存器 r_0 中的张量变量 a_0 存储到内存中。如图 6 中 S_{r_0} 标记所示位置。

[0058] 第二步、由于节点 V_2 , V_4 , V_5 , V_9 和 V_3 处均包含张量变量 a_0 的使用，所以需要在所述节点处将张量变量 a_0 从内存中加载到寄存器 r_0 中。

[0059] 参见图 7，实施例 2：一种面向神经网络计算的内存优化方法，在内存优化过程中用于神经网络计算的计算图执行流中的张量变量分配 3 个寄存器，具体如下：

步骤 S1：将计算图重构为拓扑结构计算图；如图 8 左边所示的计算图。

[0060] 步骤 S2：构建关于张量变量的生命周期区间；如图 8 右边边所示的计算图。

[0061] 步骤 S3：构建关于生命周期区间的扫描线；

在拓扑结构计算图的起始节点 V_1 处，构建一条与所述生命周期区间起始线平行的扫描线。所述扫描线用来辅助观察空闲寄存器和张量变量的状态。扫描线的工作方式是在扫描线从生命周期区间的起始端向生命周期区间的终止端移动的过程中，观察是否存在空闲寄存器可以分配给数据流执行过程中的张量变量，参见图 9，顶部横线代表扫描线。

[0062] 步骤 S4：将张量变量分配到空闲寄存器；

参见图 10，将空闲寄存器 r_3 分配给张量变量 x ，扫描线的起始位置，也就是节点 V_1 处，发现存在空闲寄存器 r_3 可以分配给张量变量 x 。

[0063] 参见图 11，将寄存器 r_1 分配给节点 V_2 处的张量变量 y 。扫描线扫描至节点 V_2 的位置时，发现扫描线已经穿过寄存器 r_1 的所述生命周期区间，所以可以将寄存器 r_1 的所述生命周期区间从处于激活状态的所述生命周期区间列表中移除，将寄存器 r_1 回收至空闲寄存器列表中。最后将所述空闲寄存器 r_1 可以分配给张量变量 y 。

[0064] 参见图 12，将寄存器 r_2 分配给节点 V_3 处的张量变量 z 。扫描线扫描至节点 V_3 的位置时，发现扫描线已经穿过寄存器 r_2 的所述生命周期区间，所以可以将寄存器 r_2 的所述生命周期区间从处于激活状态的所述生命周期区间列表中移除，将寄存器 r_2 回收至空闲寄存器列表中。最后将所述空闲寄存器 r_2 可以分配给张量变量 z 。

[0065] 步骤 S5：将最远终点的所述生命周期区间对应张量变量的寄存器分配给超出寄存器需求数量的张量变量；

参见图 13，扫描线扫描至节点 V_4 的位置时，发现既没有空闲寄存器，也没有可以从处于激

活状态的所述生命周期区间列表中移除的已经被扫描过期的所述生命周期区间。所以需要将最远终点的所述生命周期区间对应张量变量 x 所分配的寄存器 r_3 中的张量变量转移到内存中，然后将所释放的寄存器 r_3 分配给超出寄存器需求数量的张量变量 b 。由于张量变量 x 被存储到内存中，所以张量变量 x 对应的所述生命周期区间更新为虚线。

[0066] 参见图 14，将已过期的生命周期区间 l_y 所分配的寄存器分配给超出寄存器需求数量的张量变量 w 。扫描线扫描至节点 V_5 的位置时，发现扫描线已经穿过张量变量 y 所分配寄存器 r_1 对应的生命周期区间 l_y ，所以可以将张量变量 y 从处于激活状态的生命周期区间列表中移除，将寄存器 r_1 回收到空闲寄存器列表中。最后将所述空闲寄存器 r_1 可以分配给超出寄存器需求数量的张量变量 w 。

[0067] 步骤 S6：将已过期的所述生命周期区间所分配的寄存器分配给超出寄存器需求数量的张量变量；

参见图 15，将已过期生命周期区间所分配的寄存器回收到空闲寄存器列表中。扫描线扫描至节点 V_8 的结束位置时，发现扫描线已经穿过张量变量 z 所分配寄存器 r_2 对应的生命周期区间 l_z 和张量变量 w 所分配寄存器 r_1 对应的生命周期区间 l_w 。所以将已过期生命周期区间 l_z 和 l_w 所对应的张量变量 z 和 w 从处于激活状态的生命周期区间列表中移除，将寄存器 r_2 和 r_1 回收到空闲寄存器列表中。

[0068] 参见图 16，将已过期生命周期区间所分配的寄存器回收到空闲寄存器池中并将空闲寄存器分配给处于激活状态的生命周期区间。扫描线扫描至节点 V_9 的位置时，发现扫描线已经穿过张量变量 b 所分配寄存器 r_3 对应的生命周期区间 l_b 。所以将已过期生命周期区间 l_b 所对应的张量变量 b 从处于激活状态的生命周期区间列表中移除，将寄存器 r_3 回收到空闲寄存器列表中。扫描线扫描至节点 V_9 的位置时，发现存在空闲寄存器 r_1 ，将空闲寄存器 r_1 分配给 l_x 对应生命周期区间。扫描线扫描至节点 V_{10} 的位置时，发现存在空闲寄存器 r_3 ，将空闲寄存器 r_3 分配给 l_z 对应生命周期区间。

[0069] 步骤 S7：将转移到内存中的张量变量添加回处于激活状态的所述生命周期区间并为其分配空闲寄存器。

[0070] 参见图 17，扫描线扫描至节点 V_{10} 的位置时，发现存在空闲寄存器 r_2 ，将转移到内存中的变量 x 添加回处于激活状态的生命周期区间列表，并且将空闲寄存器 r_2 分配给 l_x 对应的生命周期区间。

[0071] 与前述一种面向神经网络计算的内存优化方法的实施例相对应，本发明还提供了一种

面向神经网络计算的内存优化装置的实施例 3。

[0072] 参见图 18，本发明实施例 3 提供一种面向神经网络计算的内存优化装置，包括存储器和一个或多个处理器，所述存储器中存储有可执行代码，所述一个或多个处理器执行所述可执行代码时，用于实现上述实施例中的一种面向神经网络计算的内存优化方法。

[0073] 本发明一种面向神经网络计算的内存优化装置的实施例 3 可以应用在任意具备数据处理能力的设备上，该任意具备数据处理能力的设备可以为诸如计算机等设备或装置。装置实施例 3 可以通过软件实现，也可以通过硬件或者软硬件结合的方式实现。以软件实现为例，作为一个逻辑意义上的装置，是通过其所在任意具备数据处理能力的设备的处理器将非易失性存储器中对应的计算机程序指令读取到内存中运行形成的。从硬件层面而言，如图 18 所示，为本发明一种面向神经网络计算的内存优化装置所在任意具备数据处理能力的设备的一种硬件结构图，除了图 18 所示的处理器、内存、网络接口、以及非易失性存储器之外，实施例 3 中装置所在的任意具备数据处理能力的设备通常根据该任意具备数据处理能力的设备的实际功能，还可以包括其他硬件，对此不再赘述。

[0074] 上述装置中各个单元的功能和作用的实现过程具体详见上述方法中对应步骤的实现过程，在此不再赘述。

[0075] 对于装置实施例 3 而言，由于其基本对应于方法实施例，所以相关之处参见方法实施例的部分说明即可。以上所描述的装置实施例 3 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发明方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

[0076] 本发明实施例还提供一种计算机可读存储介质，其上存储有程序，该程序被处理器执行时，实现上述实施例中的一种面向神经网络计算的内存优化方法。

[0077] 所述计算机可读存储介质可以是前述任一实施例所述的任意具备数据处理能力的设备的内部存储单元，例如硬盘或内存。所述计算机可读存储介质也可以是任意具备数据处理能力的设备的外部存储设备，例如所述设备上配备的插接式硬盘、智能存储卡（Smart Media Card, SMC）、SD 卡、闪存卡（Flash Card）等。进一步的，所述计算机可读存储介质还可以既包括任意具备数据处理能力的设备的内部存储单元也包括外部存储设备。所述计算机可读存储介质用于存储所述计算机程序以及所述任意具备数据处理能力的设备所需的其他程序和数据，还可以用于暂时地存储已经输出或者将要输出的数据。

[0078] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1. 一种面向神经网络计算的内存优化方法，其特征在于，包括以下步骤：

步骤 S1：将计算图重构为拓扑结构计算图；

步骤 S2：构建关于张量变量的生命周期区间；

步骤 S3：构建关于生命周期区间的扫描线；

步骤 S4：将张量变量分配到空闲寄存器；

步骤 S5：将最远终点的所述生命周期区间对应张量变量的寄存器分配给超出寄存器需求数量的张量变量；

步骤 S6：将已过期的所述生命周期区间所分配的寄存器分配给超出寄存器需求数量的张量变量；

步骤 S7：将转移到内存中的张量变量添加回处于激活状态的所述生命周期区间并为其分配空闲寄存器。

2. 如权利要求 1 所述的一种面向神经网络计算的内存优化方法，其特征在于，所述步骤 S1 具体包括以下子步骤：

步骤 S11：后序顺序遍历计算图，得到子图访问列表；

步骤 S12：逆序后序所述子图访问列表，得到计算图的拓扑结构顺序；

步骤 S13：根据所述拓扑结构顺序重构计算图，得到拓扑结构计算图。

3. 如权利要求 2 所述的一种面向神经网络计算的内存优化方法，其特征在于，所述后序顺序为当访问所述计算图的某个节点时，则优先递归地访问所述节点的后继节点。

4. 如权利要求 1 所述的一种面向神经网络计算的内存优化方法，其特征在于，所述步骤 S2 具体为构建关于每个节点中包含张量变量的生命周期区间，所述节点包含的张量变量对应的关于生命周期区间起始于所述张量变量处于存活状态的第一个节点的位置，终止于所述张量变量处于存活状态的最后一个节点的位置。

5. 如权利要求 1 所述的一种面向神经网络计算的内存优化方法，其特征在于，所述步骤 S3 具体为在所述拓扑结构计算图的起始节点处，构建一条与所述生命周期区间平行的扫描线，所述扫描线用于从所述生命周期区间的起始端向所述生命周期区间的终止端移动过程中，观察是否存在空闲寄存器可以分配给数据流执行过程中的张量变量。

6. 如权利要求 1 所述的一种面向神经网络计算的内存优化方法，其特征在于，所述步骤 S5 具体为执行流位于某个节点的位置时，当所述节点既没有空闲寄存器，也没有可以从处于激活状态的所述生命周期区间中移除的已经被扫描过期的所述生命周期区间，则将最远终点的所述生命周期区间对应的张量变量所分配的寄存器中的张量变量转移到内存中，然后将所释

放的寄存器分配给超出寄存器需求数量的张量变量。

7. 如权利要求 1 所述的一种面向神经网络计算的内存优化方法，其特征在于，所述步骤 S6 具体为执行流位于某个节点的位置时，当所述扫描线已经穿过张量变量所分配寄存器对应的所述生命周期区间，将张量变量从处于激活状态的所述生命周期区间中移除，对应分配的寄存器回收到空闲寄存器列表中，将所述空闲寄存器分配给超出寄存器需求数量的张量变量。
8. 如权利要求 1 所述的一种面向神经网络计算的内存优化方法，其特征在于，所述步骤 S7 具体为执行流位于某个节点的位置时，当存在空闲寄存器，则将转移到内存中的张量变量添加回处于激活状态的所述生命周期区间，并且将空闲寄存器分配给对应的所述生命周期区间。
9. 一种面向神经网络计算的内存优化装置，其特征在于，包括存储器和一个或多个处理器，所述存储器中存储有可执行代码，所述一个或多个处理器执行所述可执行代码时，用于实现权利要求 1-8 中任一项所述的一种面向神经网络计算的内存优化方法。
10. 一种计算机可读存储介质，其特征在于，其上存储有程序，该程序被处理器执行时，实现权利要求 1-8 中任一项所述的一种面向神经网络计算的内存优化方法。

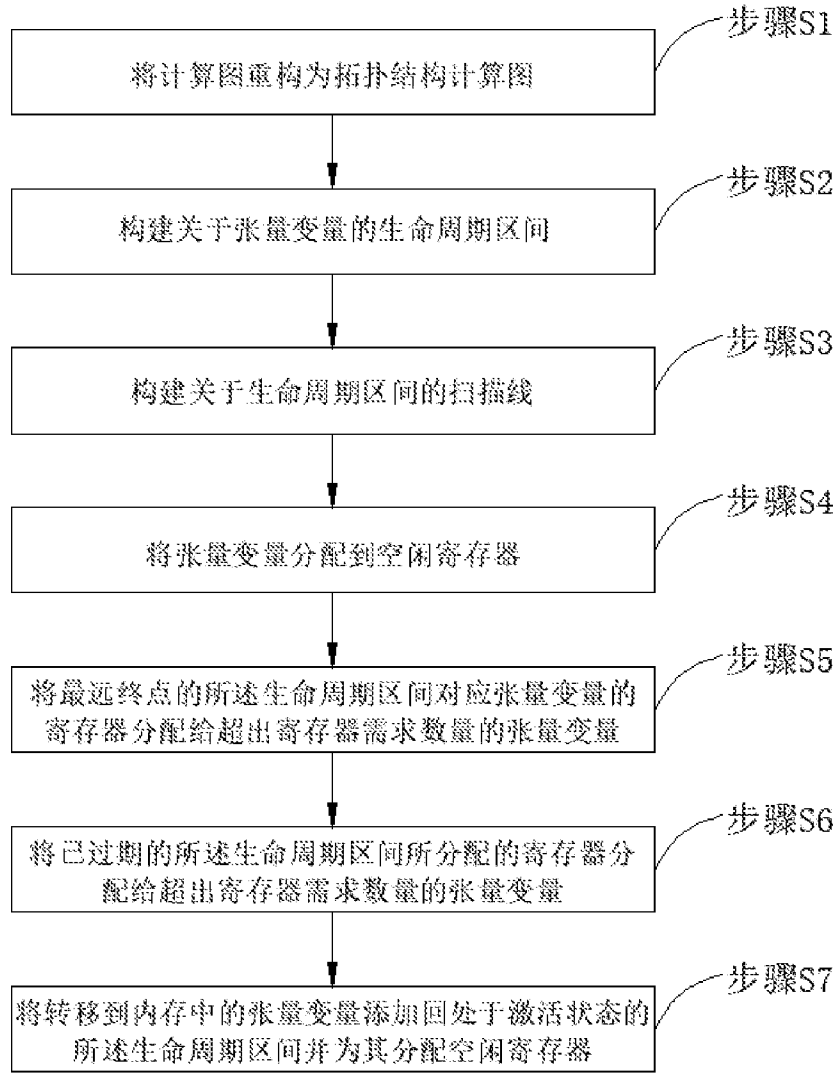


图 1

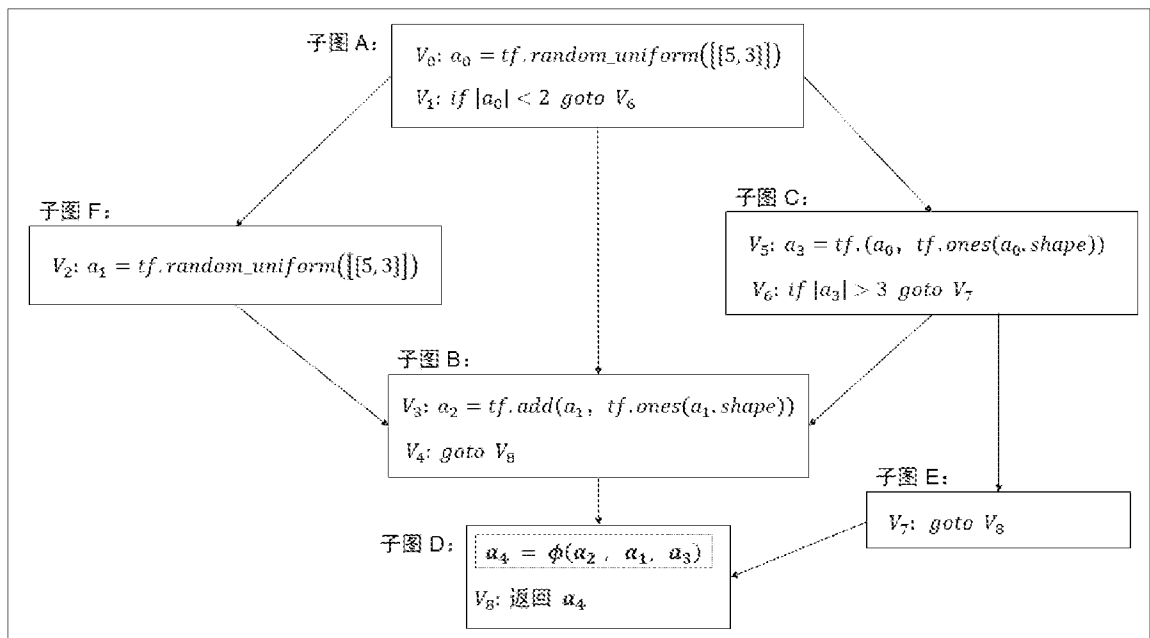


图 2

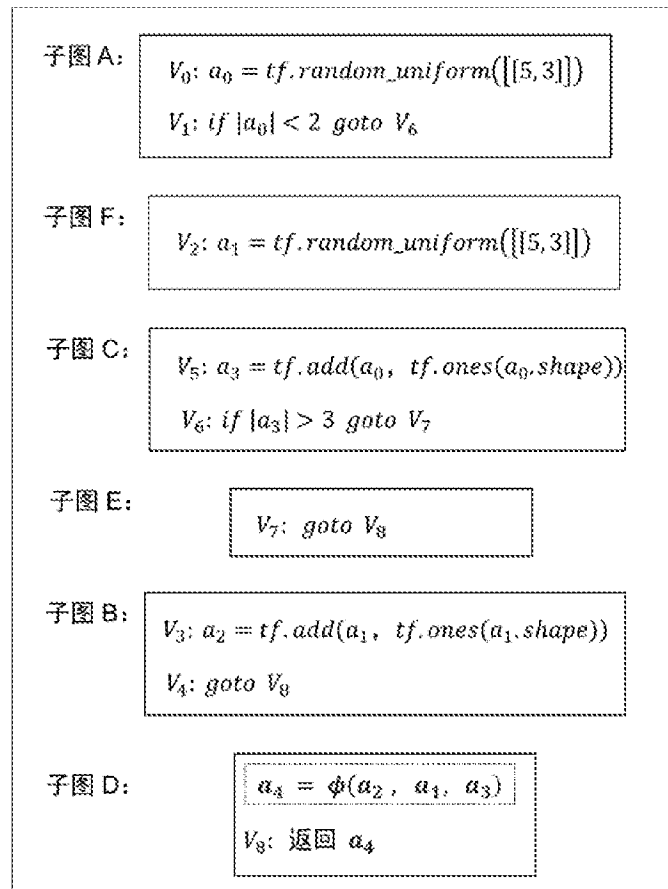


图 3

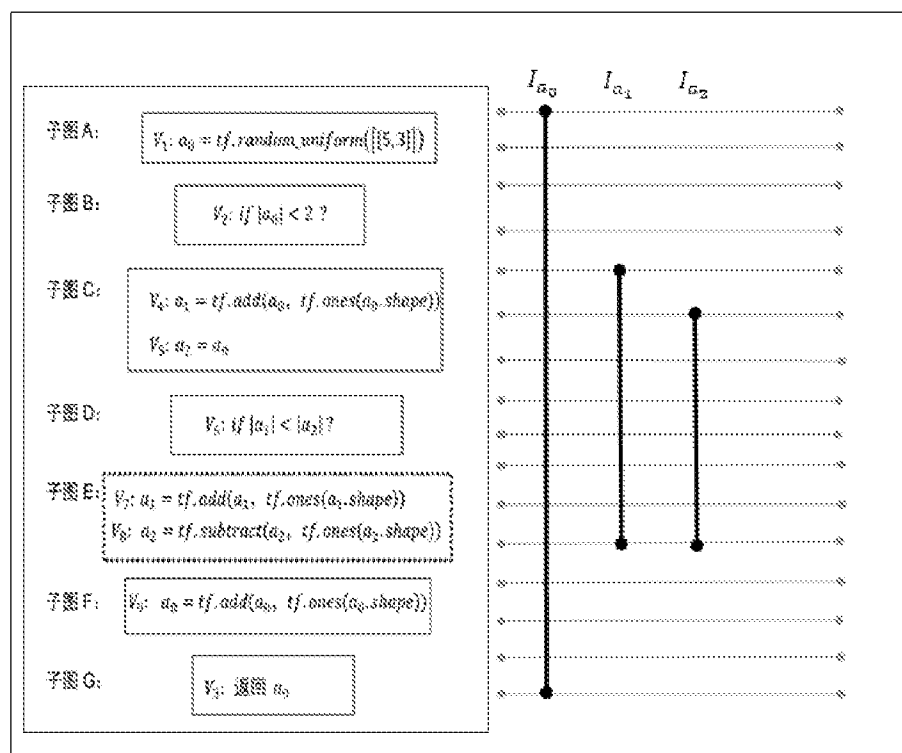


图 4

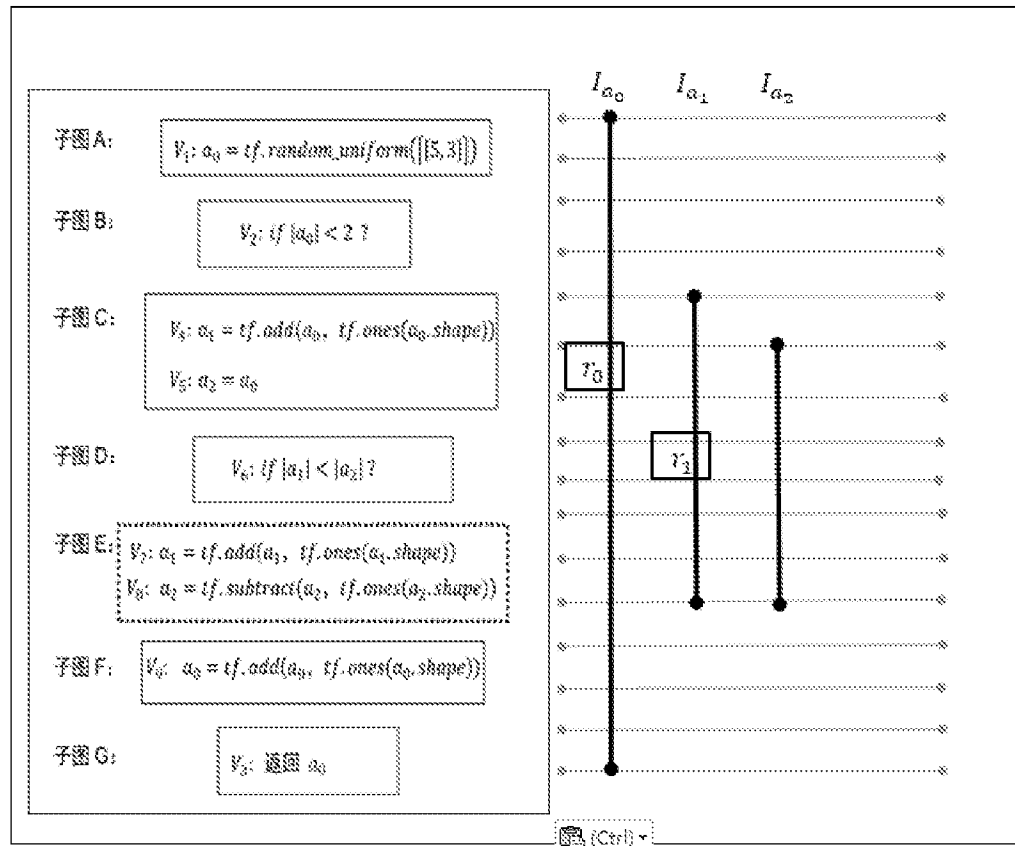


图 5

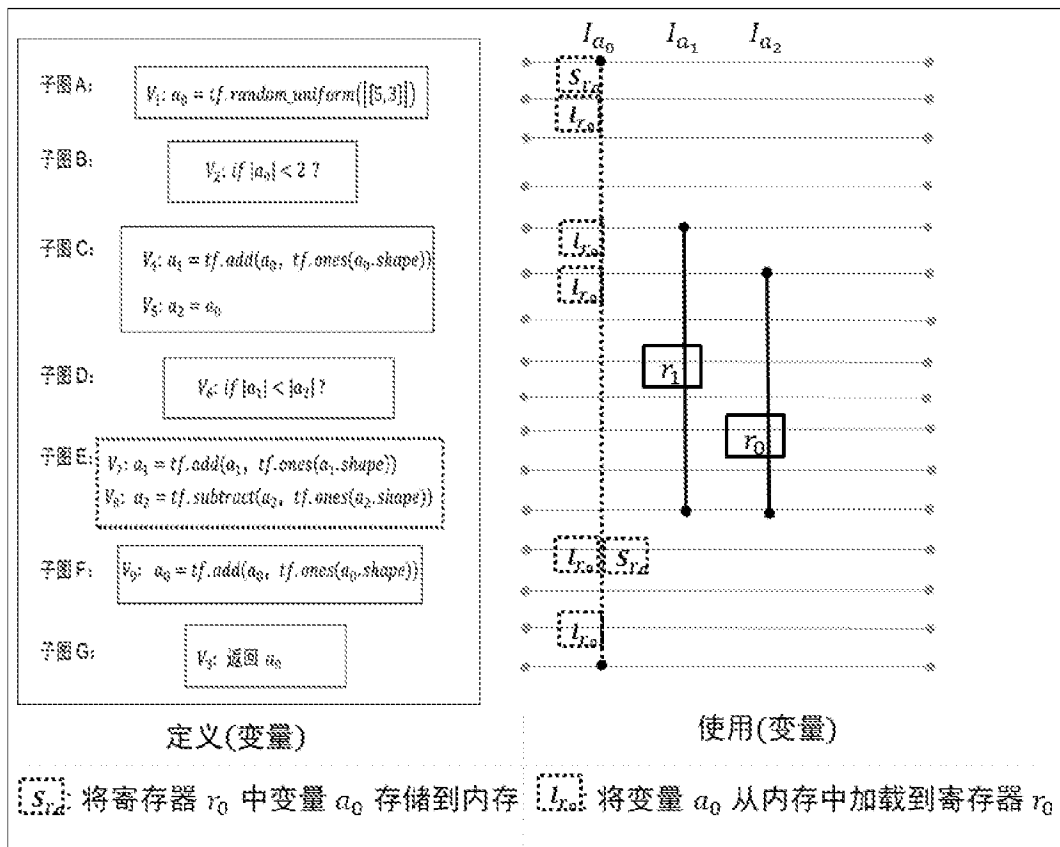


图 6

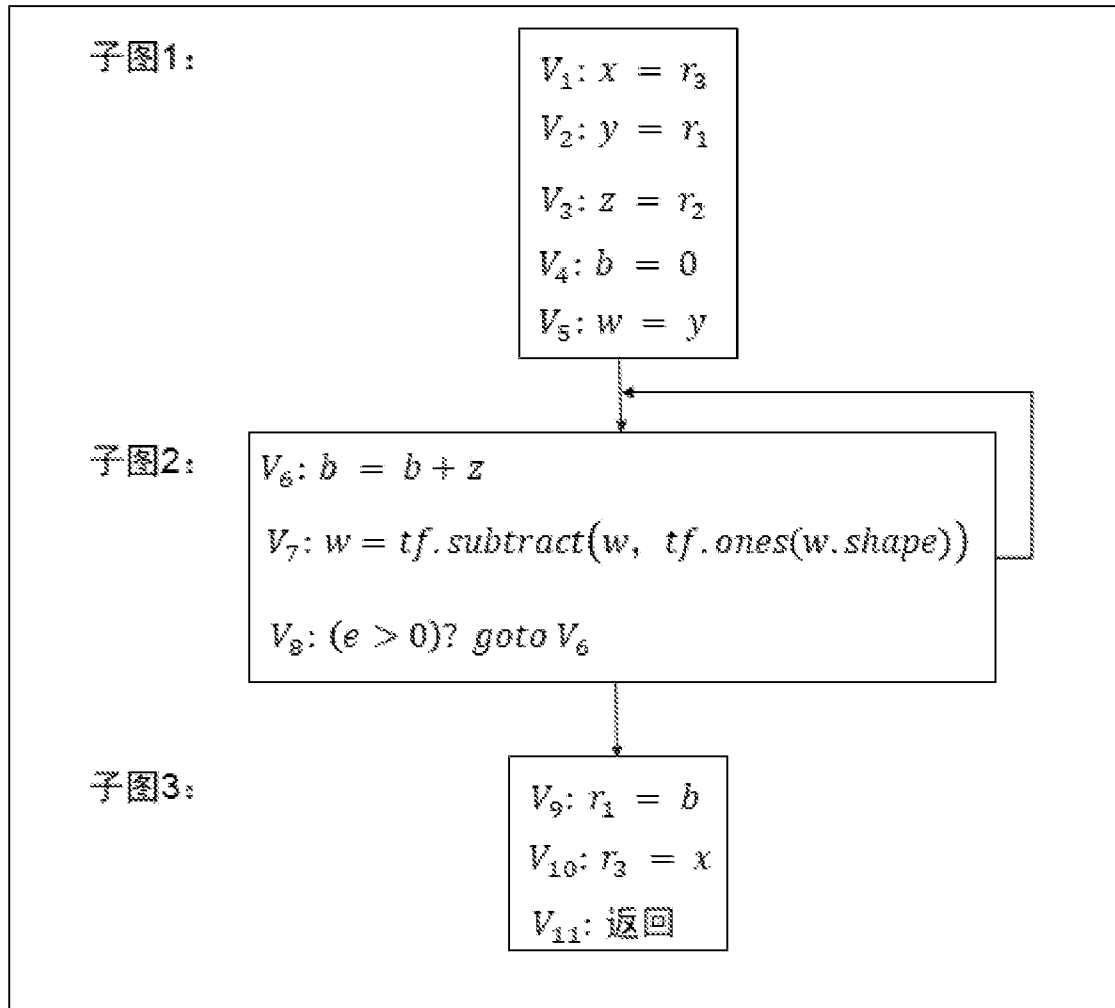


图 7

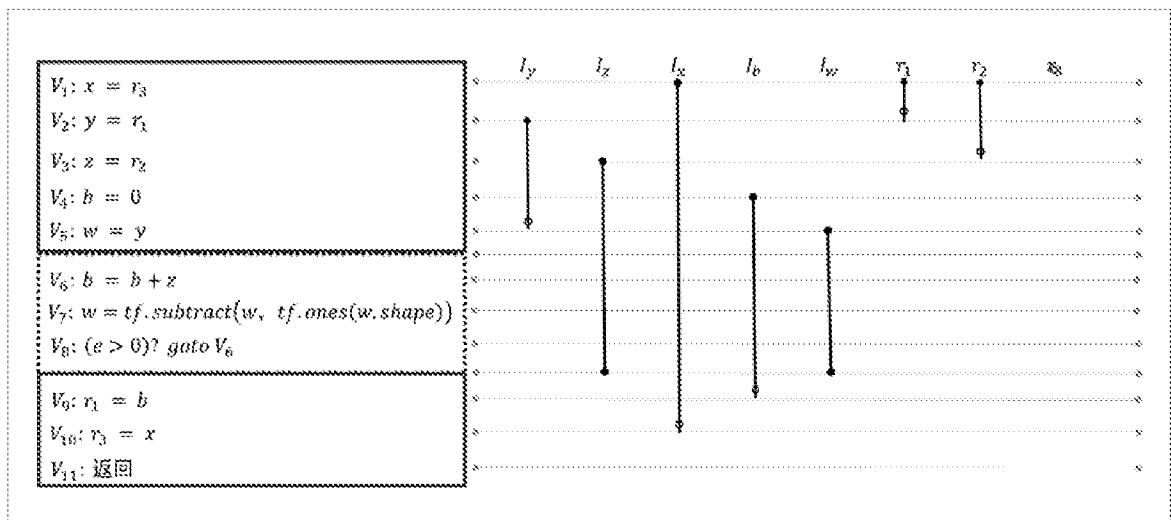


图 8

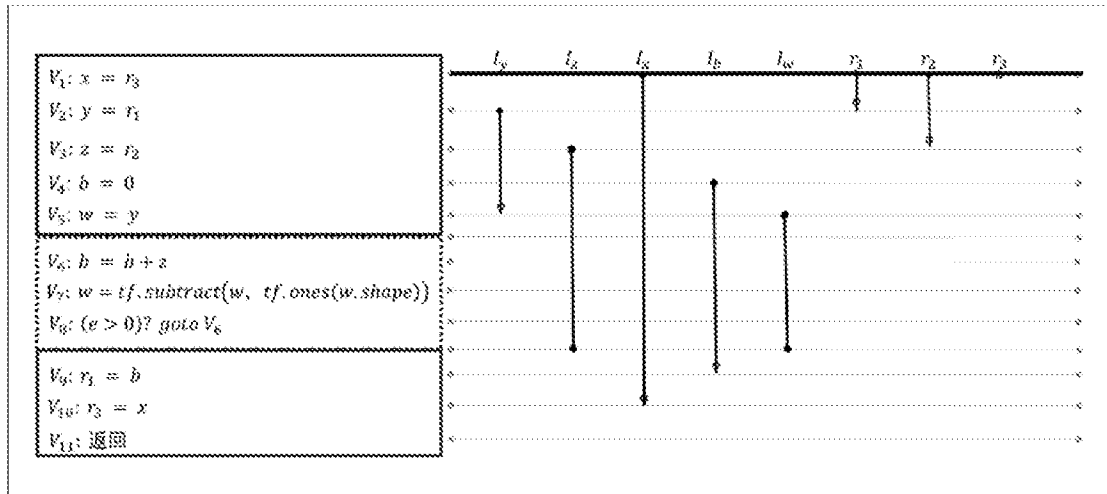


图 9

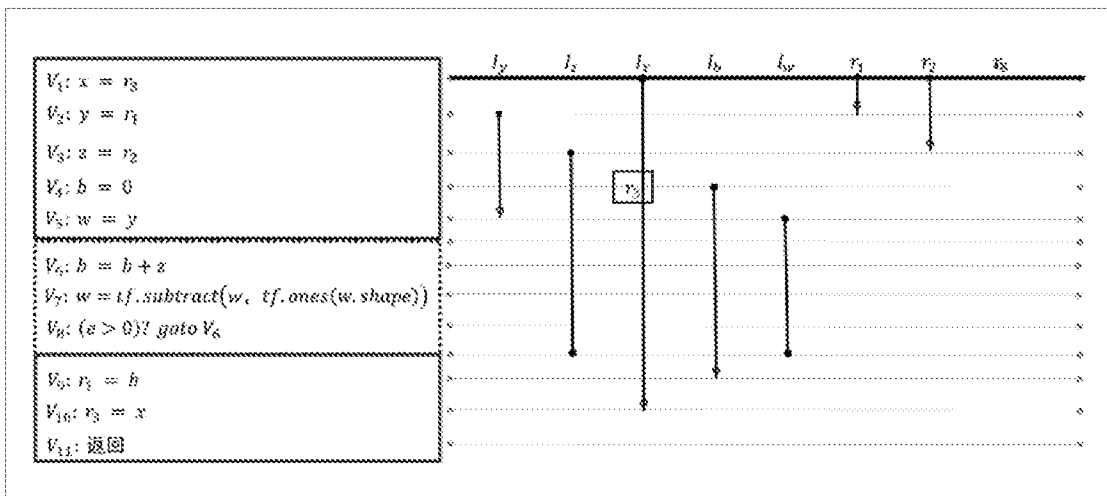


图 10

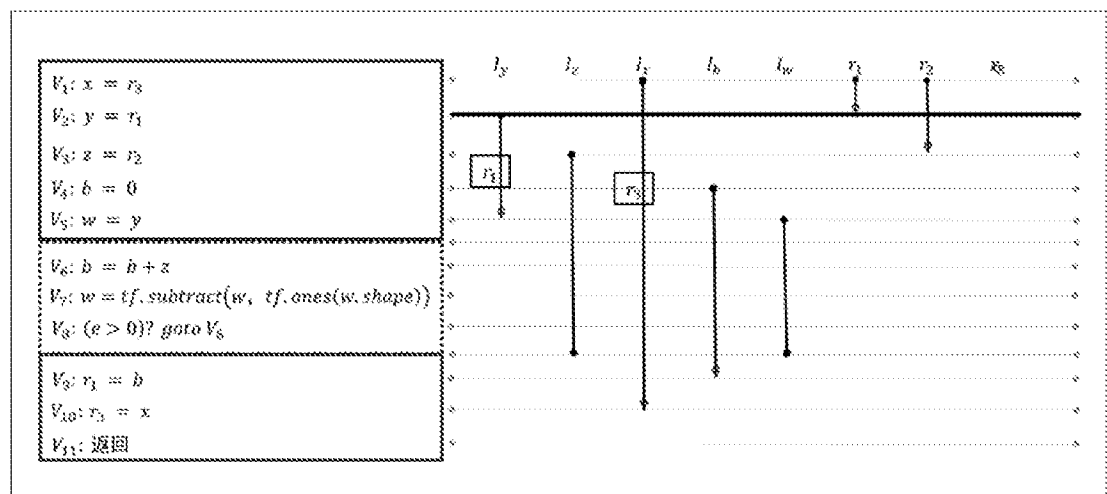


图 11

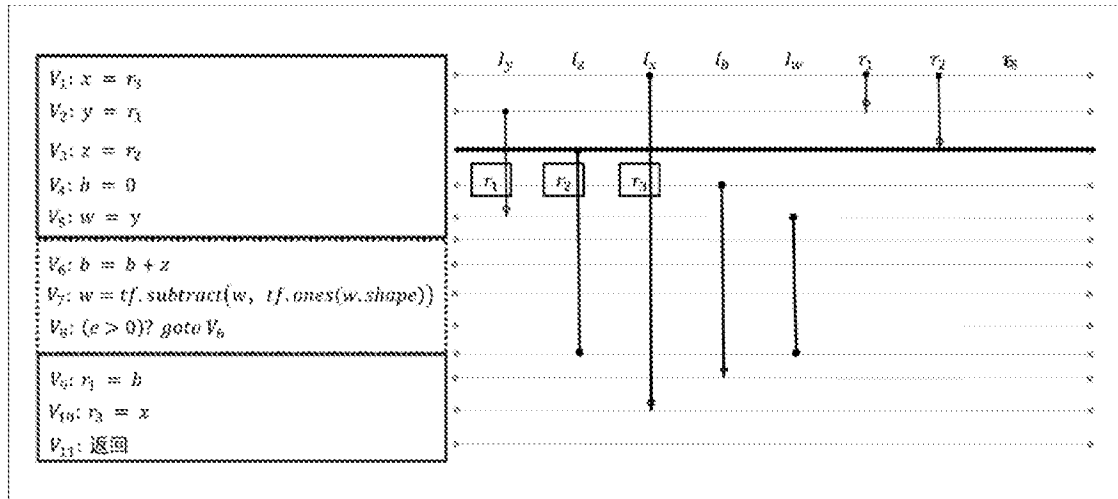


图 12

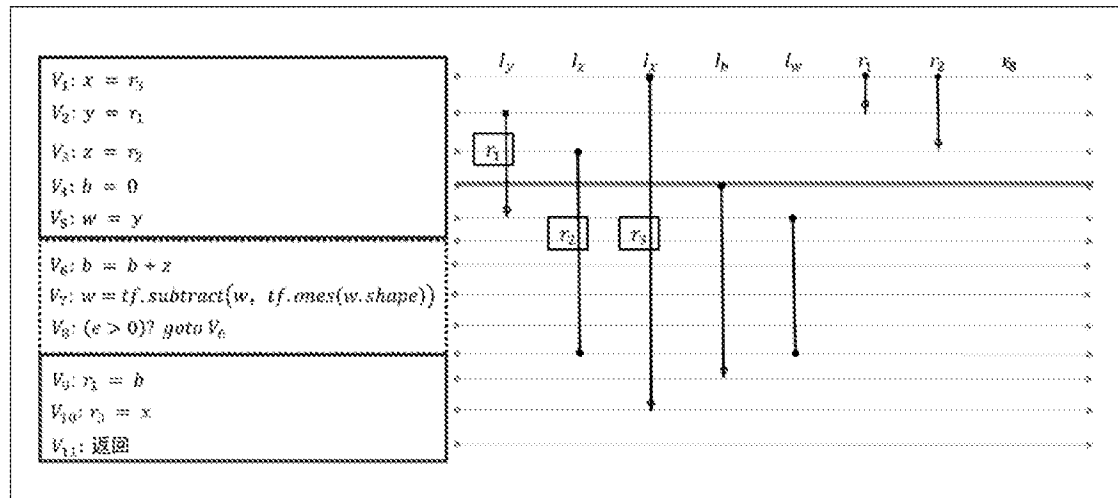


图 13

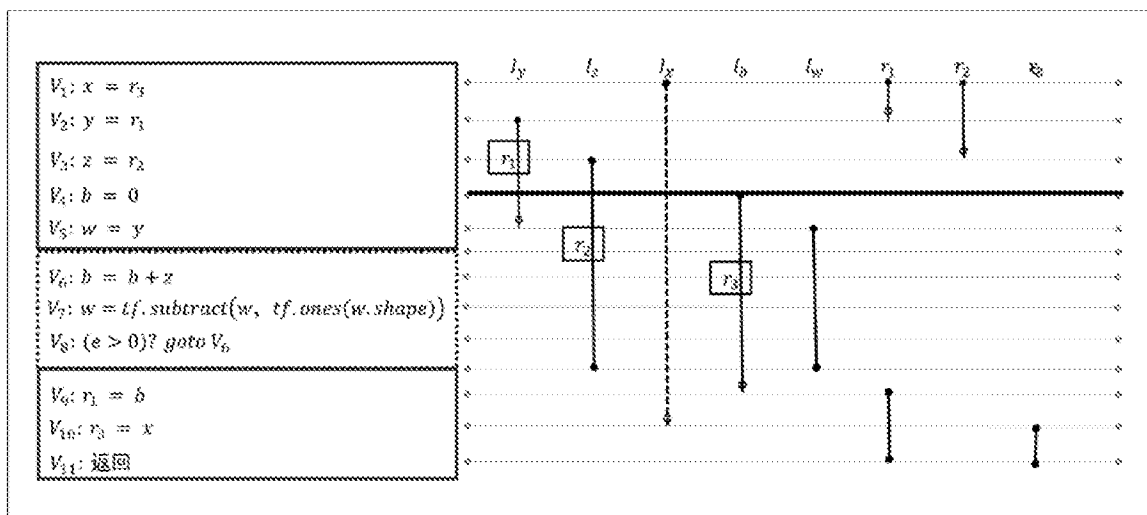


图 14

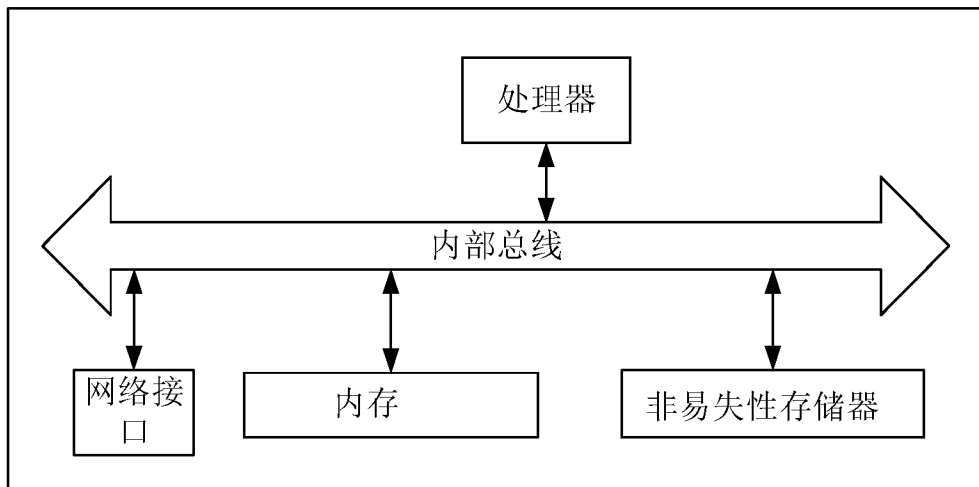


图 18

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/124000

A. CLASSIFICATION OF SUBJECT MATTER

G06F9/50(2006.01)i; G06F12/02(2006.01)i; G06N3/04(2023.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F, G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, USTXT, VEN, CNKI, IEEE: 神经网络, 图, 内存, 寄存器, 生命, 生存, 周期, 区间, 需求, 超出, 分配, 扫描, 回收,
neural network, graph, memory, register, life, cycle, requir+, interval, exceed+, distribute+, scan+, recycl+**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 114936099 A (ZHEJIANG LAB) 23 August 2022 (2022-08-23) description, paragraphs [0023]-[0042]	1-10
A	CN 111814971 A (HANGZHOU NATIONALCHIP SCIENCE & TECHNOLOGY CO., LTD.) 23 October 2020 (2020-10-23) entire document	1-10
A	CN 112948001 A (ANHUI CAMBRICON INFORMATION TECHNOLOGY CO., LTD.) 11 June 2021 (2021-06-11) entire document	1-10
A	US 2022035544 A1 (SIGMASTAR TECHNOLOGY LTD.) 03 February 2022 (2022-02-03) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

18 May 2023

Date of mailing of the international search report

24 May 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/124000

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	114936099	A	23 August 2022	None			
CN	111814971	A	23 October 2020	None			
CN	112948001	A	11 June 2021	None			
US	2022035544	A1	03 February 2022	CN	112199190	A	08 January 2021

A. 主题的分类

G06F9/50(2006.01)i; G06F12/02(2006.01)i; G06N3/04(2023.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: G06F, G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNTEXT, USTXT, VEN, CNKI, IEEE: 神经网络, 图, 内存, 寄存器, 生命, 生存, 周期, 区间, 需求, 超出, 分配, 扫描, 回收, neural network, graph, memory, register, life, cycle, requir+, interval, exceed+, distribute+, scan+, recycl+

C. 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 114936099 A (之江实验室) 2022年8月23日 (2022 - 08 - 23) 说明书第[0023]-[0042]段	1-10
A	CN 111814971 A (杭州国芯科技股份有限公司) 2020年10月23日 (2020 - 10 - 23) 全文	1-10
A	CN 112948001 A (安徽寒武纪信息科技有限公司) 2021年6月11日 (2021 - 06 - 11) 全文	1-10
A	US 2022035544 A1 (SIGMASTAR TECHNOLOGY LTD.) 2022年2月3日 (2022 - 02 - 03) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2023年5月18日

国际检索报告邮寄日期

2023年5月24日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

王琦瑶

电话号码 (+86) 010-53961363

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/124000

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	114936099	A	2022年8月23日	无	
CN	111814971	A	2020年10月23日	无	
CN	112948001	A	2021年6月11日	无	
US	2022035544	A1	2022年2月3日	CN 112199190 A	2021年1月8日