(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
2021 年 5 月 27 日 (27.05.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/098821 A1

(51) 国际专利分类号:
G06N 3/04 (2006.01)

(21) 国际申请号: PCT/CN2020/130393

(22) 国际申请日: 2020 年 11 月 20 日 (20.11.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201911144635.8 2019年11月20日 (20.11.2019) CN

(71) 申请人: 华为技术有限公司 (**HUAWEI TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 清华大学(**TSINGHUA UNIVERSITY**) [CN/CN]; 中国北京市海淀区清华大学, Beijing 100084 (CN)。

(72) 发明人: 高滨 (**GAO, Bin**); 中国北京市海淀区清华大学, Beijing 100084 (CN)。 姚鹏(**YAO, Peng**); 中国北京市海淀区清华大学, Beijing 100084 (CN)。 王侃文(**WANG, Kanwen**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 廖健行 (**LIAO, Jianxing**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 王铁英 (**WANG, Tieying**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 吴华强 (**WU, Huaqiang**); 中国北京市海淀区清华大学, Beijing 100084 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT,

(54) **Title:** METHOD FOR DATA PROCESSING IN NEURAL NETWORK SYSTEM, AND NEURAL NETWORK SYSTEM

(54) 发明名称: 神经网络系统中数据处理的方法、神经网络系统

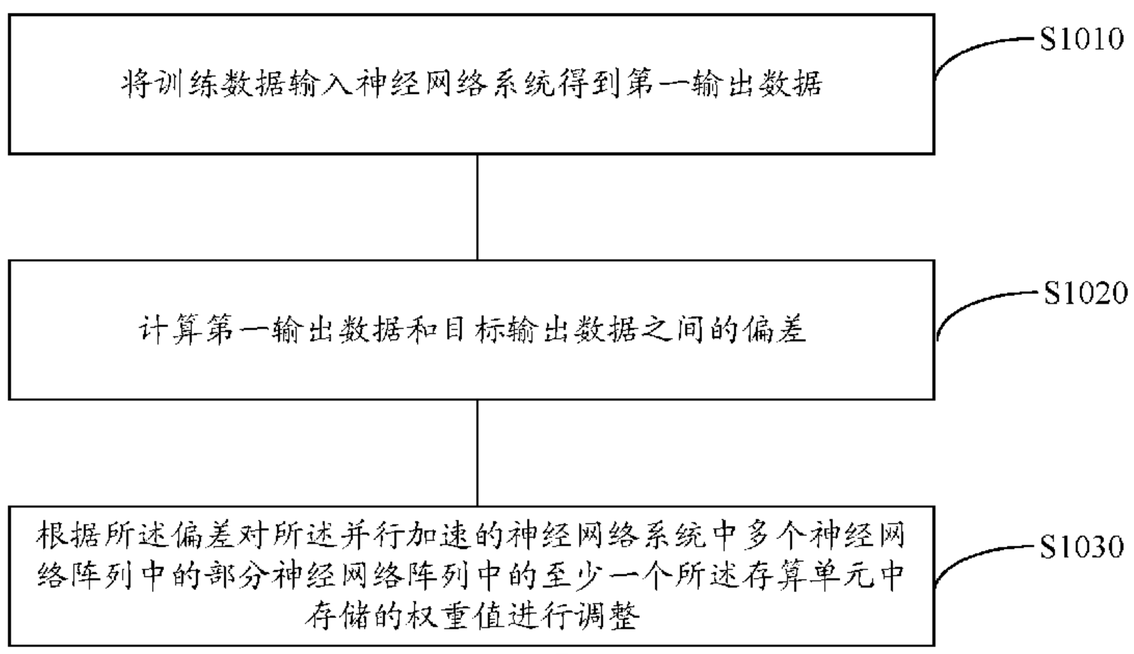


图 10

S1010 INPUT TRAINING DATA INTO A NEURAL NETWORK SYSTEM TO OBTAIN FIRST OUTPUT DATA

S1020 CALCULATE THE DEVIATION BETWEEN THE FIRST OUTPUT DATA AND A TARGET OUTPUT DATA

S1030 ADJUST, ACCORDING TO THE DEVIATION, WEIGHT VALUES STORED IN AT LEAST ONE STORAGE AND CALCULATION UNIT OF PART OF A PLURALITY OF NEURAL NETWORK ARRAYS IN A PARALLEL ACCELERATION NEURAL NETWORK SYSTEM

(57) **Abstract:** A method for data processing in a neural network system, and a neural network system. Said method comprises: inputting training data into a neural network system to obtain first output data, and according to the deviation between the first output data and a target output data, adjusting weight values stored in at least one storage and calculation unit of part of a plurality of neural network arrays in a parallel acceleration neural network system. Part of the neural network arrays are used to implement the calculation of part of neural network layers in the neural network system. Said method improves the performance of a neural network system and the accuracy of identification.

[见续页]

WO 2021/098821 A1

JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, 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, 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, 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, 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：一种神经网络系统中数据处理的方法、神经网络系统，该方法包括：将训练数据输入神经网络系统得到第一输出数据，并根据第一输出数据和目标输出数据之间的偏差，对并行加速的神经网络系统中多个神经网络阵列中的部分神经网络阵列中的至少一个存算单元中存储的权重值进行调整。其中，部分神经网络阵列用于实现神经网络系统中部分神经网络层的计算。该方法可以提高神经网络系统的性能以及识别的准确率。

神经网络系统中数据处理的方法、神经网络系统

技术领域

本申请涉及神经网络领域，并且更具体地，涉及神经网络系统中数据处理的方法、神经网络系统。

背景技术

人工智能（artificial intelligence，AI）是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能领域的研究包括机器人，自然语言处理，计算机视觉，决策与推理，人机交互，推荐与搜索，AI 基础理论等。

在 AI 领域中，深度学习是一种基于深层次的人工神经网络（artificial neural network，ANN）算法的学习技术。神经网络的训练过程是一种数据为中心的任务，需要计算硬件具有高性能、低功耗的处理能力。

基于多个神经网络阵列的神经网络系统可以实现存算一体，可以处理深度学习的任务。例如，神经网络阵列中的至少一个存算单元可以存储对应的神经网络层的权重值。由于网络结构或系统架构设计的原因，各神经网络阵列的处理速度会不一致，此时可以使用多个神经网络阵列进行并行处理，共同计算来加速处于速度瓶颈的神经网络阵列。然而参与并行加速的各神经网络间由于存算单元自身存在的一些非理想特性，例如，器件波动、电导漂移、阵列良率等，将降低神经网络系统的整体性能，该神经网络系统的准确率较低。

发明内容

本申请提供一种在并行加速的神经网络系统中数据处理的方法、神经网络系统，可以解决使用并行加速技术时的由于器件非理想特性带来的影响，提高神经网络系统的性能以及识别的准确率。

第一方面，提供了一种神经网络系统中数据处理的方法，包括：在采用并行加速的神经网络系统中，将训练数据输入神经网络系统得到第一输出数据，其中，该神经网络系统中包括多个神经网络阵列，该多个神经网络阵列中的每个神经网络阵列中包括多个存算单元，每个存算单元用于存储对应的神经网络中神经元的权重值；计算该第一输出数据和目标输出数据之间的偏差；根据该偏差对该多个神经网络阵列中的部分神经网络阵列中的至少一个该存算单元中存储的权重值进行调整，其中，该部分神经网络阵列用于实现该神经网络系统中部分神经网络层的计算。

上述技术方案中，可以根据神经网络阵列的实际输出数据和目标输出数据之间的偏差，对多个神经网络阵列中的部分神经网络阵列中存算单元存储的权重值进行调整和更新，可以实现对存算单元非理想特性的兼容，提高系统的识别率和系统的性能，从而避免由于存算单元的非理想特性所造成的系统性能下降。

在第一方面的一种可能的实现方式中，多个神经网络阵列包括第一神经网络阵列和第

二神经网络阵列，该第一神经网络阵列的输入数据包括该第二神经网络阵列的输出数据。

在第一方面的另一种可能的实现方式中，该第一神经网络阵列包括用于实现神经网络中的全连接层计算的神经网络阵列。

上述技术方案中，可以仅对实现全连接层计算的神经网络阵列的存算单元存储的权重值进行调整和更新，便可以实现对存算单元非理想特性的兼容，提高系统的识别率和系统的性能，其成本较低，且有效，易于实现。

在第一方面的另一种可能的实现方式中，根据该第一神经网络阵列的输入值和该偏差，对该第一神经网络阵列中至少一个该存算单元中存储的权重值进行调整。

在第一方面的另一种可能的实现方式中，该多个神经网络阵列还包括第三神经网络阵列，该第三神经网络阵列和该第二神经网络阵列用于并行实现神经网络中的卷积层的计算。

在第一方面的另一种可能的实现方式中，根据该第二神经网络阵列的输入数据和该偏差，对该第二神经网络阵列中的至少一个该存算单元中存储的权重值进行调整；根据该第三神经网络阵列的输入数据和该偏差，对该第三神经网络阵列中的至少一个该存算单元中存储的权重值进行调整。

上述技术方案中，还可以对并行实现神经网络中卷积层计算的多个神经网络阵列中存算单元存储的权重值进行调整和更新，提高调整精度，从而提高神经网络系统输出的准确率。

在第一方面的另一种可能的实现方式中，将该偏差划分为至少两个子偏差，其中，该至少两个子偏差中的第一子偏差与该第二神经网络阵列的输出数据对应；根据所述第一子偏差和该第二神经网络阵的输入数据，对该第二神经网络阵列中的至少一个该存算单元中存储的权重值进行调整；根据所述第二子偏差和该第三神经网络阵的输入数据，对该第三神经网络阵列中的至少一个该存算单元中存储的权重值进行调整。

在第一方面的另一种可能的实现方式中，根据存算单元中更新后的权重值确定脉冲个数，根据该脉冲个数，对该神经网络阵列中至少一个该存算单元中存储的权重值进行改写。

第二方面，提供了一种神经网络系统，包括：

处理模块，用于将训练数据输入神经网络系统得到第一输出数据，其中，该神经网络系统中包括多个神经网络阵列，该多个神经网络阵列中的每个神经网络阵列中包括多个存算单元，每个存算单元用于存储对应的神经网络中神经元的权重值；

计算模块，用于计算该第一输出数据和目标输出数据之间的偏差；

调整模块，用于根据该偏差对该多个神经网络阵列中的部分神经网络阵列中的至少一个该存算单元中存储的权重值进行调整，其中，该部分神经网络阵列用于实现该神经网络系统中部分神经网络层的计算。

在第二方面的一种可能的实现方式中，该多个神经网络阵列包括第一神经网络阵列和第二神经网络阵列，该第一神经网络阵列的输入数据包括该第二神经网络阵列的输出数据。

在第二方面的另一种可能的实现方式中，该第一神经网络阵列包括用于实现神经网络中的全连接层计算的神经网络阵列。

在第二方面的另一种可能的实现方式中，该调整模块具体用于：

根据该第一神经网络阵列的输入值和该偏差，对该第一神经网络阵列中至少一个该存算单元中存储的权重值进行调整。

在第二方面的另一种可能的实现方式中，该多个神经网络阵列还包括第三神经网络阵列，该第三神经网络阵列和该第二神经网络阵列用于并行实现神经网络中的卷积层的计算。

在第二方面的另一种可能的实现方式中，该调整模块具体用于：

根据该第二神经网络阵列的输入数据和该偏差，对该第二神经网络阵列中的至少一个该存算单元中存储的权重值进行调整；根据该第三神经网络阵列的输入数据和该偏差，对该第三神经网络阵列中的至少一个该存算单元中存储的权重值进行调整。

在第二方面的另一种可能的实现方式中，该调整模块具体用于：

将该偏差划分为至少两个子偏差，其中，该至少两个子偏差中的第一子偏差与该第二神经网络阵列的输出数据对应，该至少两个子偏差中的第二子偏差与该第三神经网络阵列的输出数据对应；

根据所述第一子偏差和该第二神经网络阵的输入数据，对该第二神经网络阵列中的至少一个该存算单元中存储的权重值进行调整；

根据所述第二子偏差和该第三神经网络阵的输入数据，对该第三神经网络阵列中的至少一个该存算单元中存储的权重值进行调整。

在第二方面的另一种可能的实现方式中，调整模块具体用于：根据存算单元中更新后的权重值确定脉冲个数，根据该脉冲个数，对该神经网络阵列中至少一个该存算单元中存储的权重值进行改写。

第二方面和第二方面的任意一个可能的实现方式的有益效果和第一方面以及第一方面的任意一个可能的实现方式的有益效果是对应的，对此，不再赘述。

第三方面，提供了一种神经网络系统，包括处理器和存储器，其中该存储器用于存储计算机程序，该处理器用于从存储器中调用并运行该计算机程序，使得神经网络系统执行第一方面或第一方面任意一种可能的实现方式提供的方法。

可选地，在具体实现中，该处理器的个数不做限制。该处理器可以是通用处理器，可以通过硬件来实现也可以通过软件来实现。当通过硬件实现时，该处理器可以是逻辑电路、集成电路等；当通过软件来实现时，该处理器可以是一个通用处理器，通过读取存储器中存储的软件代码来实现，该存储器可以集成在处理器中，可以位于该处理器之外，独立存在。

第四方面，提供了一种芯片，该芯片上设置有如第二方面或第二方面中任一种可能的实现方式中该的神经网络系统。

该芯片包括处理器与数据接口，其中，处理器通过该数据接口读取存储器上存储的指令，以执行第一方面或第一方面任意一种可能的实现方式中的方法。在具体实现过程中，该芯片可以以中央处理器（central processing unit，CPU）、微控制器（micro controller unit，MCU）、微处理器（micro processing unit，MPU）、数字信号处理器（digital signal processing，DSP）、片上系统（system on chip，SoC）、专用集成电路（application-specific integrated circuit，ASIC）、现场可编程门阵列（field programmable gate array，FPGA）或可编辑逻辑器件（programmable logic device，PLD）的形式实现。

第五方面，提供了一种计算机程序产品，该计算机程序产品包括：计算机程序代码，

当该计算机程序代码在计算机上运行时，使得计算机执行上述第一方面或第一方面任意一种可能的实现方式中的方法。

第六方面，提供了一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序代码，当该计算机程序代码在计算机上运行时，使得计算机执行上述执行第一方面或第一方面任意一种可能的实现方式中的方法。这些计算机可读存储包括但不限于如下的一个或者多个：只读存储器（read-only memory，ROM）、可编程 ROM（programmable ROM，PROM）、可擦除的 PROM（erasable PROM，EPROM）、Flash 存储器、电 EPROM（electrically EPROM，EEPROM）以及硬盘驱动器（hard drive）。

附图说明

图 1 为本申请提供的一种神经网络系统 100 的结构示意图。

图 2 为本申请提供的另一种神经网络系统 200 的结构示意图。

图 3 是神经网络与神经网络阵列之间的映射关系的示意图。

图 4 是本申请提供的一种可能的权重矩阵示意图。

图 5 是一种可能的神经网络模型的示意图。

图 6 是本申请提供的一种神经网络系统的示意图。

图 7 是本申请提供的并行计算的多组忆阻器阵列的输入数据和输出数据的结构示意图。

图 8 的（a）是本申请提供的一种对输入数据进行加速的并行计算的多组忆阻器阵列。

图 8 的（b）是本申请提供的一种具体的数据拆分的示意图。

图 9 是本申请提供的另一种对输入数据进行加速的并行计算的多组忆阻器阵列。

图 10 是本申请提供的一种神经网络系统中数据处理的方法的示意性流程图。

图 11 是本申请提供的一种前向操作和反向操作过程的示意图。

图 12 是本申请提供的一种对多个忆阻器阵列中实现全连接层计算的第一忆阻器阵列中存储的权重值进行更新的示意图。

图 13 是本申请提供的另一种对多个忆阻器阵列中实现全连接层计算的第一忆阻器阵列中存储的权重值进行更新的示意图。

图 14 是一种对实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的示意图。

图 15 是一种根据残差值对实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的示意图。

图 16 是另一种对实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的示意图。

图 17 是另一种根据残差值对实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的示意图。

图 18 是本申请提供的一种增加神经网络阵列中的至少一个存算单元存储的权重值的示意图。

图 19 是本申请提供的一种减小神经网络阵列中的至少一个存算单元存储的权重值的示意图。

图 20 是本申请提供的一种通过边读边写的方式增加神经网络阵列中的至少一个存算

单元存储的权重值的示意图。

图 21 是本申请提供的一种通过边读边写的方式减小神经网络阵列中的至少一个存算单元存储的权重值的示意图。

图 22 是本申请实施例提供的一种神经网络的训练过程的示意性流程图。

图 23 是本申请实施例提供的一种神经网络系统 2300 的示意性结构图。

具体实施方式

下面将结合附图，对本申请中的技术方案进行描述。

人工智能（artificial intelligence，AI）是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能领域的研究包括机器人，自然语言处理，计算机视觉，决策与推理，人机交互，推荐与搜索，AI 基础理论等。

在 AI 领域中，深度学习是一种基于深层次的人工神经网络（artificial neural network，ANN）算法的学习技术。人工神经网络（artificial neural network，ANN），简称为神经网络（neural network，NN）或类神经网络。在机器学习和认知科学领域，是一种模仿生物神经网络（动物的中枢神经系统，特别是大脑）的结构和功能的数学模型或计算模型，用于对函数进行估计或近似。人工神经网络可以包括卷积神经网络(convolutional neural network，CNN)、多层感知器（multilayer perceptron，MLP）、循环神经网络（recurrent neural network，RNN）等神经网络。

神经网络的训练过程也就是学习参数矩阵的过程，其最终目的是得到训练好的神经网络的每一层神经元的参数矩阵（每一层神经元的参数矩阵包括该层神经元包括的每个神经元对应的权重）。通过训练得到的权重形成的各个参数矩阵可以从用户输入的待推理图像中提取像素信息，从而帮助神经网络对该待推理图像进行正确的推理，使得训练好的神经网络输出的预测值尽可能的接近训练数据的先验知识（prior knowledge）。

应理解，先验知识也被称为真实值（ground truth），一般包括由用户提供的训练数据对应的真实结果。

上述神经网络的训练过程是一种以数据为中心的任务，需要计算硬件具有高性能、低功耗的处理能力。基于传统冯诺依曼架构的计算由于存储单元和计算单元相分离，需要大量数据搬移，无法实现高能效的处理。

下面结合图 1-图 2 描述本申请的系统架构图。

图 1 为本申请实施例提供的一种神经网络系统 100 的结构示意图。如图 1 所示，神经网络系统 100 可以包括主机 105 以及神经网络电路 110。

神经网络电路 110 通过主机接口与主机 105 连接。主机接口可以包括标准的主机接口以及网络接口（network interface）。例如，主机接口可以包括快捷外设互联标准（peripheral component interconnect express，PCIE）接口。

作为示例，如图 1 所示，神经网络电路 110 可以通过 PCIE 总线 106 与主机 105 连接。因此，数据通过 PCIE 总线 106 输入至神经网络电路 110 中，并通过 PCIE 总线 106 接收

神经网络电路 110 处理完成后的数据。并且，主机 105 也可以通过主机接口监测神经网络电路 110 的工作状态。

主机 105 可以包括处理器 1052 以及内存 1054。需要说明的是，除了图 1 所示的器件外，主机 105 还可以包括通信接口以及作为外存的磁盘等其他器件，在此不做限制。

处理器（processor）1052 是主机 105 的运算核心和控制核心（control unit）。处理器 1052 中可以包括多个处理器核（core）。处理器 1052 可以是一块超大规模的集成电路。在处理器 1052 中安装有操作系统和其他软件程序，从而使得处理器 1052 能够实现对内存 1054、缓存、磁盘及外设设备（如图 1 中的神经网络电路）的访问。可以理解的是，在本申请实施例中，处理器 1052 中的 core 例如可以是中央处理器（central processing unit，CPU），还可以是其他特定集成电路（application specific integrated circuit，ASIC）。

应理解，本申请实施例中的处理器 1052 还可以为其他通用处理器、数字信号处理器（digital signal processor，DSP）、专用集成电路（application specific integrated circuit，ASIC）、现成可编程门阵列（field programmable gate array，FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

内存 1054 是主机 105 的主存。内存 1054 通过双倍速率（double data rate，DDR）总线和处理器 1052 相连。内存 1054 通常用来存放操作系统中各种正在运行的软件、输入和输出数据以及与外存交换的信息等。为了提高处理器 1052 的访问速度，内存 1054 需要具备访问速度快的优点。在传统的计算机系统架构中，通常采用动态随机存取存储器（dynamic random access memory，DRAM）作为内存 1054。处理器 1052 能够通过内存控制器（图 1 中未示出）高速访问内存 1054，对内存 1054 中的任意一个存储单元进行读操作和写操作。

还应理解，本申请实施例中的内存 1054 可以是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（read-only memory，ROM）、可编程只读存储器（programmable ROM，PROM）、可擦除可编程只读存储器（erasable PROM，EPROM）、电可擦除可编程只读存储器（electrically EPROM，EEPROM）或闪存。易失性存储器可以是随机存取存储器（random access memory，RAM），其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的随机存取存储器（random access memory，RAM）可用，例如静态随机存取存储器（static RAM，SRAM）、动态随机存取存储器（DRAM）、同步动态随机存取存储器（synchronous DRAM，SDRAM）、双倍数据速率同步动态随机存取存储器（double data rate SDRAM，DDR SDRAM）、增强型同步动态随机存取存储器（enhanced SDRAM，ESDRAM）、同步连接动态随机存取存储器（synchlink DRAM，SLDRAM）和直接内存总线随机存取存储器（direct rambus RAM，DR RAM）。

图 1 所示的神经网络电路 110 可以是由多个神经网络芯片（chip）和多个路由器 120 组成的芯片阵列。为了描述方便，本申请实施例将神经网络芯片 115 简称为芯片 115。多个芯片 115 通过路由器 120 相互连接。例如，一个芯片 115 可以与一个或多个路由器 120 连接。多个路由器 120 可以组成一种或多种网络拓扑。芯片 115 之间可以通过所述多种网络拓扑进行数据传输和信息交互。

图 2 为本申请实施例提供的另一种神经网络系统 200 的结构示意图。如图 2 所示，神

经网络系统200可以包括主机105以及神经网络电路210。

神经网络电路210通过主机接口与主机105连接，如图2所示，神经网络电路210可以通过PCIE总线106与主机105。连接主机105中可以包括处理器1052以及内存1054。具体有关主机105的描述，请参考图2中的说明，此处不再赘述。

图2所示的神经网络电路210可以是由多个芯片115组成的芯片阵列，其中，该多个芯片115挂靠在PCIE总线106上。芯片115之间通过PCIE总线106进行数据传输和信息交互。

可选地，图1-图2中的神经网络系统的架构仅作为示例，本领域技术人员能够理解，在实践中，上述神经网络系统可以包括比图1或图2中更多或更少的单元。或者神经网络系统中的模块、单元或电路也可以被其它具有类似功能的模块、单元或电路替代，本申请实施例对此不作限定。例如，在另一些示例中，上述神经网络系统也可以由基于数字计算的图形处理器(graphics processing unit，GPU)或现场可编程门阵列(field programmable gate array，FPGA ）实现。

在一些示例中，上述神经网络电路可以由存算一体的多个神经网络矩阵来实现。多个神经网络矩阵中的每一个神经网络矩阵可以包括多个存算单元，每个存算单元中用于存储对应的神经网络中每一层神经元的权重值，用于实现神经网络层的计算。

本申请实施例对存算单元不作具体限定，可以包括但不限于：忆阻器（memristor）、静态RAM（static RAM，SRAM）、NOR Flash、磁性RAM（magnetism RAM，MRAM）、铁电栅场效应晶体管（ferroelectric gate field-effect transistors，FeFET）、电化学RAM（electrochemistry RAM，ECRAM）。其中，忆阻器可以包括但不限于：阻变式随机访问存储器（resistive random-access memory，ReRAM）、导电桥接RAM（conductive-bridging RAM，CBRAM）、相变存储器（phase-change memory，PCM）。

例如，神经网络矩阵是由ReRAM组成的ReRAM交叉矩阵（ReRAM crossbar）。上述神经网络系统中可以包括多个ReRAM crossbar。

在本申请实施例中，ReRAM crossbar也可以称为忆阻器交叉阵列、ReRAM器件或者ReRAM。包括一个或多个ReRAM crossbar的芯片可以称为ReRAM芯片。

ReRAM crossbar 是一种全新的非冯诺依曼计算架构。该架构集存储和计算功能于一身，具有灵活的可配置特性，并利用了模拟计算方式，有望实现比传统的计算构架更快速度、更低能耗的矩阵向量乘，在神经网络计算中具有广泛应用前景。

下面结合图3，以神经网络阵列为ReRAM crossbar作为示例，对通过ReRAM crossbar实现神经网络层计算的具体实现过程进行详细描述。

图3是神经网络与神经网络阵列之间的映射关系的示意图。神经网络110包括多个神经网络层。

在本申请实施例中，神经网络层为逻辑的层概念，一个神经网络层是指要执行一次神经网络操作。每一层神经网络计算均是由计算节点（也可以称为神经元）来实现。实际应用中，神经网络层可以包括卷积层、池化层、全连接层等。

本领域技术人员知道，在执行神经网络计算（例如，卷积计算）时，神经网络系统中的计算节点可以将输入数据与对应神经网络层的权重进行计算。在神经网络系统中，权重通常用一个实数矩阵表示，权重矩阵中的每一个元素代表一个权重值。权重通常用于表示输入数据对于输出数据的重要程度。如图4所示，图4所示的m行n列的权重矩阵可以

是一个神经网络层的一个权重，该权重矩阵中的每一个元素代表一个权重值。

由于每一层神经网络的计算可以由 ReRAM crossbar 来实现，又由于 ReRAM 具有存算一体的优势。因此，权重可以在计算之前被配置到 ReRAM crossbar 的多个 ReRAM 单元（cell）上。从而，可以通过 ReRAM crossbar 实现输入数据与配置的权重的矩阵乘加操作。

应理解，本申请实施例中 ReRAM 单元也可以称为忆阻器单元。忆阻器单元在计算之前被配置权重可以理解为在忆阻器单元中存储对应的神经网络中神经元的权重值，具体的，可以是通过忆阻器单元的阻值或电导值来指示神经网络中神经元的权重值。

还应理解，实际应用中，ReRAM crossbar 和神经网络层之间可以是一对一的映射关系，或者还可以是一对多的映射关系。下面会结合附图进行详细描述，此处不再赘述。

为了描述清楚，下面对 ReRAM crossbar 如何实现矩阵乘加操作的过程进行简单的描述。

需要说明的是，图 3 中以神经网络 110 中的第一神经网络层为例介绍数据处理过程。在神经网络系统中，第一神经网络层可以是神经网络系统中的任意一层，为描述方便，第一神经网络层也可以简称为“第一层”。

图 3 所示的 ReRAM crossbar 120 为 $m \times n$ 的交叉阵列，其中，ReRAM crossbar 120 中可以包括多个忆阻器单元（例如，$G_{1,1}$，$G_{1,2}$ 等），每一列的忆阻器单元的位数据线（bit line，BL）连接在一起，每一行的忆阻器单元的源数据线（source line，SL）连接在一起。

在本申请实施例中，可以通过忆阻器的电导值来表示神经网络中神经元的权重。具体的，作为示例，可以将图 4 所示的权重矩阵中的每个元素通过位于 BL 与 SL 交点的忆阻器的电导值来表示。例如，图 3 的 $G_{1,1}$ 表示图 4 中的权重元素 $W_{0,0}$，图 3 的 $G_{1,2}$ 表示图 4 中的权重元素 $W_{0,1}$ 等。

忆阻器单元的电导值不同可以表示忆阻器单元存储的神经网络中神经元的权重不同。

在执行神经网络计算的过程中，可以通过加载到忆阻器的 BL 的电压值来表示 n 个输入数据 Vi，例如图 3 中的：V1，V2，V3，...Vn。输入数据可以通过电压表示，从而使得加载到忆阻器的输入数据与忆阻器存储的权重值实现点乘运算，得到如图 3 中所示的 m 个输出数据。所述 m 个输出数据可以通过 SL 的电流表示，例如图 3 中的：I1，I2，...Im。

应理解，加载到忆阻器的电压值的实现方式有多种，本申请实施例对此不做具体限定。例如，可以通过电压的脉冲幅值来表示，又如，还可以通过电压的脉冲宽度来表示，又如，还可以通过电压的脉冲个数来表示，又如，还可以通过电压的脉冲个数与电压的脉冲幅值的结合来表示。

需要说明的是，上文中以一个神经网络阵列为例，对神经网络阵列完成神经网络中相应的乘累加计算的过程进行了详细的描述。在实际应用中，是由多个神经网络阵列共同完成完整的神经网络所需要的乘累加计算的。

多个神经网络阵列中的一个神经网络阵列可以和一个神经网络层对应，该神经网络阵列用于实现一个神经网络层的计算。或者，多个神经网络阵列可以和一个神经网络层对应，用于实现该一个神经网络层的计算。或者，多个神经网络阵列中的一个神经网络阵列可以和多个神经网络层对应，用于实现该多个神经网络层的计算。

下面结合图 5-图 6，对神经网络阵列和神经网络层之间的对应关系进行详细描述。

为了便于描述，下文中以忆阻器阵列为神经网络阵列作为示例进行说明。

图 5 是一种可能的神经网络模型的示意图。该神经网络模型可以包括多个神经网络层。

在本申请实施例中，神经网络层为逻辑的层概念，一个神经网络层是指要执行一次神经网络操作。每一层神经网络计算均是由计算节点来实现。神经网络层可以包括卷积层、池化层、全连接层等。

如图 5 所示，神经网络模型中可以包括 n 个神经网络层（又可以被称为 n 层神经网络），其中，n 为大于或等于 2 的整数。图 5 示出了神经网络模型中的部分神经网络层，如图 5 所示，神经网络模型可以包括第一层 302、第二层 304、第三层 306、第四层 308、第五层 310 至第 n 层 312。其中，第一层 302 可以执行卷积操作，第二层 304 可以是对第一层 302 的输出数据执行池化操作或激活运算的操作，第三层 306 可以是对第二层 304 的输出数据执行卷积操作，第四层 308 可以对第三层 306 的输出结果执行卷积操作，第五层 310 可以对第二层 304 的输出数据以及第四层 308 的输出数据执行求和操作。第 n 层 312 可以执行全连接层的操作。

应理解，池化操作或激活运算的操作可以是由外部的数字电路模块来实现。具体的，外部的数字电路模块可以是通过 PCIE 总线 106 与神经网络电路 110 连接（图 1 或图 2 中未示出）。

可以理解的是，图 5 只是对神经网络模型中的神经网络层的一个简单示例和说明，并不对每一层神经网络的具体操作进行限制，例如，第四层 308 也可以是池化运算，第五层 310 也可以是做卷积操作或池化操作等其他的神经网络操作。

图 6 是本申请实施例提供的一种神经网络系统的示意图。如图 6 所示，该神经网络系统可以包括多个忆阻器阵列。例如，第一忆阻器阵列，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列等。

第一忆阻器阵列可以实现神经网络中全连接层的计算。具体的，可以将神经网络中全连接层的权重存储在第一忆阻器阵列中，该忆阻器阵列中的每个忆阻器单元的电导值可以用于在指示全连接层的权重，并实现神经网络中全连接层的乘累加计算过程。

需要说明的是，神经网络中全连接层也可以对应多个忆阻器阵列，并由该多个忆阻器阵列共同完成全连接层的计算。本申请对此不做具体限定。

图 6 所示的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）可以实现神经网络中卷积层的计算。对于卷积层的操作而言，卷积核每次滑窗之后都会有新的输入，导致一次完整的卷积层的计算过程需要处理不同的输入。因此，可以增加神经网络网络系统层次的并行度，通过多组忆阻器阵列实现网络中相同位置的权重，从而实现对不同输入进行并行加速。即将关键位置的卷积权重用多组忆阻器阵列分别实现。在计算时，各组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）并行处理不同的输入数据，彼此之间并行工作，从而提升卷积计算效率和系统性能。

应理解，一个卷积核（kernel）代表神经网络计算过程中的一种特征提取方式。例如，在神经网络系统中进行图像处理时，给定输入图像，在输出图像中每一个像素是输入图像中一个小区域中像素的加权平均，其中权值由一个函数定义，这个函数即称为卷积核。在计算过程中，卷积核按照一定的步长（stride）依次划过输入特征图（feature map）产生特征提取后的输出数据（也被称为输出特征图）。因此，卷积核大小（kernel size）也用于指示神经网络系统中的计算节点执行一次计算的数据量的大小。本领域技术人员可以知

道，卷积核可以用一个实数矩阵来表示，例如图 8 的（a）示出了一个 3 行 3 列的卷积核，该卷积核中的每一个元素代表一个权重值。实际应用中，一个神经网络层可以包括多个卷积核。在神经网络计算过程中，可以将输入数据与卷积核进行乘加计算。

并行计算的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输入数据可以包括其他忆阻器阵列的输出数据或者外部的输入数据，该多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出数据可以作为共享的第一忆阻器阵列的输入数据。也就是说，第一忆阻器阵列的输入数据可以包括多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出数据。

并行计算的多组忆阻器阵列的输入数据和输出数据的结构可以有多种，本申请对此不做具体限定。

下面结合图 7-图 9，对并行计算的多组忆阻器阵列的输入数据和输出数据的结构进行详细描述。

图 7 是本申请实施例提供的并行计算的多组忆阻器阵列的输入数据和输出数据的结构示意图。

一种可能的实现方式中，如图 7 所示的方式一，并行计算的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输入数据组合在一起成为一个完整的输入数据，该并行计算的多组忆阻器阵列的输出数据组合在一起成为一个完整的输出数据。

例如，第二忆阻器阵列的输入数据为数据 1，第三忆阻器阵列的输入数据为数据 2，第四忆阻器阵列的输入数据为数据 3，对于卷积层而言，一个完整的输入数据包括数据 1、数据 2 以及数据 3 的组合。同样的，第二忆阻器阵列的输出数据为结果 1，第三忆阻器阵列的输出数据为结果 2，第四忆阻器阵列的输出数据为结果 3，对于卷积层而言，一个完整的输出数据包括结果 1、结果 2、结果 3 的组合。

具体的，参见图 8 的（a），可以将一张输入图片拆分为不同的部分，并分别输入至多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）进行并行计算。该多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出结果的组合可以作为该输入图片对应的完整的输出数据。

图 8 的（b）示出了一种可能的图片拆分的示意图。如图 8 的（b），将一张图像拆分成三部分分别送入三组并行加速的阵列中进行计算。第一部分送入图 8 的（a）所示的第二忆阻器阵列，得到图 7 方式一对应的“结果 1”，对应于完整输出中第二忆阻器阵列的输出结果。第二部分和第三部分可以完成类似的处理。根据卷积核的大小和滑窗步长确定每部分之间的交叠部分（如本实例中每部分之间有 2 行交叠），实现三组阵列处理后的输出结果能共同构成完整的输出。在训练过程中，当得到该层的完整残差时，按照前向计算过程的对应关系，利用第二忆阻器阵列计算对应的神经元的残差值和第一部分输入，对第二忆阻器阵列进行原位更新。第二组和第三组的阵列更新类似。具体的更新过程请参见下文中的描述，此处不再赘述。

另一种可能的实现方式中，如图 7 所示的方式二，并行计算的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输入数据分别是一个完整的输入数据，该并行计算的多组忆阻器阵列的输出数据分别是一个完整的输出数据。

例如，第二忆阻器阵列的输入数据为数据 1，对于卷积层而言，该数据 1 为一个完整

的输入数据，其输出数据为结果 1，该结果 1 为一个完整的输出数据。同样的，第三忆阻器阵列的输入数据为数据 2，对于卷积层而言，该数据 2 为一个完整的输入数据，其输出数据为结果 2，该结果 2 为一个完整的输出数据。第四忆阻器阵列的输入数据为数据 3，对于卷积层而言，该数据 3 为一个完整的输入数据，其输出数据为结果 3，该结果 3 为一个完整的输出数据。

具体的，参见图 9，可以将多个不同的完整输入数据分别输入至多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）进行并行计算。该多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出结果分别对应一个完整的输出数据。

如果神经网络阵列中的存算单元存在一些非理想特性的影响，例如，器件波动、电导漂移、阵列良率等，则该存算单元不能无损的实现权重，从而导致神经网络系统的系统整体性能下降，其识别率降低。

本申请实施例提供的技术方案，可以提高神经网络系统的性能以及识别的准确率。

下面结合图 10-图 22，对本申请实施例提供的方法侧实施例进行详细描述。

需要说明的是，本申请实施例的技术方案可以应用于各种神经网络，例如：卷积神经网络（convolutional neural network，CNN），在处理自然语言和语音中有广泛应用的循环神经网络，以及卷积神经网络和循环神经网络结合的深度神经网络等等。卷积神经网络的处理过程类似于动物的视觉系统，使得其非常适用于图片识别领域。卷积神经网络可以应用于各个图片识别领域如安防领域、计算机视觉、平安城市等，还可以应用于语音识别、搜索引擎、机器翻译等等。实际应用中，巨大的参数量和计算量对神经网络在实时性高、功耗低的场景的应用带来极大挑战。

图 10 是本申请实施例提供的一种神经网络系统中数据处理的方法的示意性流程图。如图 10 所示，该方法可以包括步骤 1010-1030，下面分别对步骤 1010-1030 进行详细描述。

步骤 1010：将训练数据输入神经网络系统得到第一输出数据。

本申请实施例中并行加速的神经网络系统中可以包括多个神经网络阵列，所述多个神经网络阵列中的每个神经网络阵列中可以包括多个存算单元，每个存算单元用于存储对应的神经网络中神经元的权重值。

步骤 1020：计算第一输出数据和目标输出数据之间的偏差。

目标输出数据可以是实际输出的第一输出数据的理想值。

本申请实施例中的偏差可以是计算第一输出数据和目标输出数据之间的差值，或者还可以是计算第一输出数据和目标输出数据之间的残差，或者还可以是计算第一输出数据和目标输出数据之间其它形式的损失函数。

步骤 1030：根据所述偏差对所述并行加速的神经网络系统中多个神经网络阵列中的部分神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整。

本申请实施例中部分神经网络阵列可以用于实现所述神经网络系统中部分神经网络层的计算。也就是说，神经网络阵列和神经网络层之间的对应关系可以是一对一的关系，或者还可以是一对多的关系，或者还可以是多对一的关系。

例如，图 6 所示的第一忆阻器阵列与神经网络中的全连接层对应，用于实现全连接层的计算。又如，图 6 所示的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）与神经网络中的卷积层对应，用于实现卷积层的计算。

应理解，神经网络层为逻辑的层概念，一个神经网络层是指要执行一次神经网络操作。具体的请参见图 5 中的说明，此处不再赘述。

存算单元中的电阻值或者电导值可以用于指示神经网络层中的权重值，本申请实施例中可以根据计算的偏差，对多个神经网络阵列中的部分神经网络阵列中的至少一个存算单元对电阻值或者电导值进行调整或者改写。

对存算单元对电阻值或者电导值进行调整或者改写的实现方式有多种。一种可能的实现方式中，可以根据偏差确定存算单元的电阻值或者电导值的更新值，并根据该更新值对存算单元施加固定个数的编程脉冲。另一种可能的实现方式中，根据偏差确定存算单元的电阻值或者电导值的更新值，通过边读边写的方式对存算单元施加编程脉冲。另一种可能的实现方式中，还可以根据不同的存算单元的特性施加不同数目的编程脉冲，以实现其电阻值或者电导值的调整或改写。下面会结合具体的实施例进行描述，此处不再赘述。

需要说明的是，本申请实施例可以通过偏差对多个神经网络阵列中用于实现全连接层的神经神经网络阵列的电阻值或者电导值进行调整，或者还可以通过偏差对多个神经网络阵列中用于实现卷积层的神经神经网络阵列的电阻值或者电导值进行调整，或者还可以通过偏差同时对用于实现全连接层的神经神经网络阵列以及用于实现卷积层的神经神经网络阵列的电阻值或者电导值进行调整。下面会结合图 11-图 17 进行详细描述，此处不再赘述。

为了便于描述，下面以计算实际输出值和目标输出值之间的残差作为示例，先对残差的计算过程进行详细描述。

在前向传播（forward propagation，FP）计算过程中，获取训练数据集，例如输入图像的像素信息，将训练数据集的数据输入到神经网络中。经过从第一层神经网络到最后一层神经网络的传递，从最后一层的神经网络的输出得到实际的输出值。

在反向传播（back propagation，BP）计算过程中，希望神经网络实际的输出值尽可能的接近训练数据的先验知识（prior knowledge），先验知识也被称为真实值（ground truth）或理想的输出值，一般包括由人提供的训练数据对应的预测结果。所以可以通过比较当前实际的输出值和理想的输出值，再根据两者之间的偏差计算残差值。具体的可以是计算目标损失函数的偏导数。根据残差值计算需要更新的权重大小，从而可以根据需要更新的权重大小对神经网络阵列中的至少一个存算单元中存储的权重值进行更新。

作为一个示例，可以计算神经网络实际的输出值和理想的输出值之间的差值的平方，并采用该平方数对权重矩阵中的权重求导数，得到残差值。

根据上述中确定的残差值以及权重值对应的输入数据，通过公式（1）确定需要更新的权重大小。

$$\Delta W = \eta \sum_{i}^{N} V_i \delta_i \qquad (1)$$

其中，ΔW 表示需要更新的权重大小，η 表示学习率，N 表示有 N 组输入数据，V 表示本层的输入数据值，δ 表示本层的残差值。

具体的，参见图 11，如图 11 所示的 N x M 阵列，SL 表示源数据线（source line），BL 表示位数据线（bit line）。

前向操作中在 BL 输入电压，在 SL 输出电流，完成 Y=XW 的矩阵向量乘计算（X 对应输入电压 V，Y 对应输出电流 I），X 是输入计算数据，可以用于前向推理。

在反向操作中在 SL 输入电压，在 BL 输出电流，即 $Y=XW^T$ 计算（X 对应输入电压 V，Y 对应输出电流 I），X 是残差值，即完成残差值的反向传播计算。忆阻器阵列更新操作（也称为原位更新）可以完成权重按照梯度方向变化的过程。

可选的，在一些实施例中，对于该层第 m 行第 n 列得到的累积更新权重，还可以根据如下的公式（2）确定是否对该层第 m 行第 n 列的权重值进行更新。

$$\Delta W_{m,n}=\begin{cases}\Delta W_{m,n} & \left|\Delta W_{m,n}\right|\geq Threshold\\ 0 & \left|\Delta W_{m,n}\right|<Threshold\end{cases} \tag{2}$$

其中，阈值（threshold）表示预设阈值。

对于该层第m行第n列得到的累积更新权重 $\Delta W_{m,n}$，采用如公式（2）所示的阈值更新法则，即没有达到阈值（threshold）要求的权重不进行更新。具体的，如果 $\Delta W_{m,n}$ 大于或等于预设阈值，则可以对该层第m行第n列的权重值进行更新。如果 $\Delta W_{m,n}$ 小于预设阈值，则不对该层第m行第n列的权重值进行更新。

下面结合图12-图13，以不同的数据组织结构为例，对实现全连接层计算的第一忆阻器阵列中存储的权重值进行更新的具体实现过程进行详细描述。

图 12 是一种对多个忆阻器阵列中实现全连接层计算的第一忆阻器阵列中存储的权重值进行更新的示意图。

如图 12 所示，可以将提前训练好的神经网络层的权重写到多个忆阻器阵列中。也就是说，在多个忆阻器阵列中存储对应神经网络层的权重。例如，第一忆阻器阵列可以实现神经网络中全连接层的计算，可以将神经网络中全连接层的权重存储在第一忆阻器阵列中，该忆阻器阵列中的每个忆阻器单元的电导值可以用于指示全连接层的权重，并实现神经网络中全连接层的乘累加计算过程。又如，多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）可以实现神经网络中卷积层的计算。通过多组忆阻器阵列实现卷积层中相同位置的权重，从而实现对不同输入进行并行加速。

如图 8 的（a）所示，将一张输入图片拆分为不同的部分，并分别输入至多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）进行并行计算。该多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出结果可以作为输入数据，输入到第一忆阻器阵列，并通过该第一忆阻器阵列得到第一输出数据。

本申请实施例中可以根据上文中计算残差值的方法，根据第一输出数据和理想输出数据计算残差值。并根据公式（1）对实现全连接层计算的第一忆阻器阵列中的各个忆阻器存储的权重值进行原位更新。

图 13 是另一种对多个忆阻器阵列中实现全连接层计算的第一忆阻器阵列中存储的权重值进行更新的示意图。

如图 9 所示，将多个不同的输入数据分别输入至多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）进行并行计算。该多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出结果可以作为输入数据，输入第一忆阻器阵列，并通过该第一忆阻器阵列得到第一输出数据。

本申请实施例中可以根据上文中计算残差值的方法，根据第一输出数据和理想输出数据计算残差值。并根据公式（1）对实现全连接层计算的第一忆阻器阵列中的各个忆阻器存储的权重值进行原位更新。

下面结合图14-图17，以不同的数据组织结构为例，对并行实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的具体实现过程进行详细描述。

图 14 是一种对实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的示意图。

如图 8 的（a）所示，将一张输入图片拆分为不同的部分，并分别输入至多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）进行并行计算。该多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出结果的组合可以作为该输入图片对应的完整的输出数据。

本申请实施例中可以根据上文中计算残差值的方法，根据输出数据和理想输出数据计算残差值。并根据公式（1）对并行实现卷积层计算的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）中的各个忆阻器存储的权重值进行原位更新。具体的实现方式有多种。

一种可能的实现方式中，可以根据多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出值和对应的理想输出值计算残差值，并根据该残差值对并行实现卷积层计算的多组忆阻器阵列中的各个忆阻器存储的权重值进行原位更新。

另一种可能的实现方式中，还可以根据第一忆阻器阵列的第一输出值和对应的理想输出值计算残差值，并根据该残差值对并行实现卷积层计算的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）中的各个忆阻器存储的权重值进行原位更新。下面结合图 15 进行详细描述。

图 15 是一种根据残差值对实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的示意图。

如图 15 所示，完整残差可以是共享的第一忆阻器阵列的残差值，该残差值是根据第一忆阻器阵列的输出值和对应的理想输出值确定的。本申请实施例中可以将完整残差划分为多个子残差，例如，残差 1，残差 2，残差 3。

每一个子残差分别和并行计算的多组忆阻器阵列中每个忆阻器阵列的输出数据对应。例如，残差 1 和第二忆阻器阵列的输出数据对应，残差 2 和第三忆阻器阵列的输出数据对应，残差 3 和第四忆阻器阵列的输出数据对应。

本申请实施例中，根据多组忆阻器阵列中每个忆阻器阵列的输入数据和子残差，结合公式（2）对忆阻器阵列中各个忆阻器存储的权重值进行原位更新。

图 16 是另一种对实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的示意图。

根据如图 9 所示的输入数据的结构，将多个不同的输入数据分别输入至多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）进行并行计算。该多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）的输出结果分别对应一个完整的输出数据。

本申请实施例中可以根据上文中计算残差值的方法，根据输出数据和理想输出数据计算残差值。并根据公式（1）对并行实现卷积层计算的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）中的各个忆阻器存储的权重值进行改写。具体的实现方式有多种。

一种可能的实现方式中，可以根据多组忆阻器阵列（例如，第二忆阻器阵列，第三忆

阻器阵列，第四忆阻器阵列）的输出值和对应的理想输出值计算残差值，并根据该残差值对并行实现卷积层计算的多组忆阻器阵列中的各个忆阻器存储的权重值进行原位更新。

另一种可能的实现方式中，还可以根据第一忆阻器阵列的第一输出值和对应的理想输出值计算残差值，并根据该残差值对并行实现卷积层计算的多组忆阻器阵列（例如，第二忆阻器阵列，第三忆阻器阵列，第四忆阻器阵列）中的各个忆阻器存储的权重值进行原位更新。下面结合图 17 进行详细描述。

图 17 是另一种根据残差值对实现卷积层计算的多组忆阻器阵列中存储的权重值进行更新的示意图。

如图 17 所示，完整残差可以是共享的第一忆阻器阵列的残差值，该残差值是根据第一忆阻器阵列的输出值和对应的理想输出值确定的。由于参与并行加速的每组忆阻器阵列都会处理得到的完整的输出结果，每组忆阻器阵列都可以依据相关的完整残差数据进行更新。假设上述完整残差是根据第二忆阻器阵列的输出结果 1 得到的，因此，可以通过上述公式（1），根据完整残差和第二忆阻器阵列的输入数据 1，对第二忆阻器阵列中各个忆阻器存储的权重值进行原位更新。

可选的，本申请实施例中还可以对并行实现卷积层计算的多个忆阻器阵列的前级阵列存储的权重值进行调整，可以通过反向传播的方式计算各层神经元的残差值。具体的请参见上文中描述的方法，此处不再赘述。应理解，并行实现卷积层计算的多个忆阻器阵列的前级神经网络阵列而言，这些阵列的输入数据可以是更前级忆阻器阵列的输出数据，也可以是外部输入的原始数据，如图像、文本、语音等；这些阵列的输出数据作为并行实现卷积层计算的多个忆阻器阵列的输入数据。

上文通过图 11-图 17 描述了对实现全连接层计算的神经网络矩阵存储的权重值进行调整，以及对并行实现卷积层计算的多个神经网络矩阵中存储的权重值进行调整。应理解，本申请实施例还可以同时对实现全连接层计算的神经网络矩阵存储的权重值以及并行实现卷积层计算的多个神经网络矩阵中存储的权重值进行调整。方法类似，此处不再赘述。

下面结合图 18-图 21，以向位于 BL 与 SL 交点的目标忆阻器单元写目标数据为例，介绍置位（set）操作和复位（reset）操作。

其中，set 操作用于将忆阻器单元的电导从低电导向高电导调整，reset 操作用于将忆阻器单元的电导从高电导向低电导调整。

如图 18，假设目标忆阻器单元的目标电导范围可以表示目标权重 W_{11}。在向目标忆阻器单元写 W_{11} 的过程中，若目标忆阻器单元的当前电导低于目标电导范围的下限，则可以通过 set 操作，增大目标忆阻器单元的电导。此时，可以通过 SL_{11} 向需要调整的目标忆阻器单元中晶体管的栅极加载电压开启晶体管，使得目标忆阻器单元处于选通状态。同时将目标忆阻器单元连接的 SL 以及交叉阵列中的其他 BL 也接地，然后，再向目标忆阻器单元所在的 BL 施加 set 脉冲，以调整目标忆阻器单元的电导。

如图 19，若目标忆阻器单元的当前电导高于目标电导范围上限，则可以通过 reset 操作，减小目标忆阻器单元的电导。此时，可以通过 SL 向需要调整的目标忆阻器单元中晶体管的栅极加载电压，使得目标忆阻器单元处于选通状态。同时将目标忆阻器单元连接的 BL 以及交叉阵列中的其他 SL 接地。然后，再向目标忆阻器单元所在的 SL 施加 reset 脉冲，以调整目标忆阻器单元的电导。

调整目标忆阻器单元的电导的具体实现方式有多种。例如，可以对该目标忆阻器单元

施加固定个数的编程脉冲。又如，或者还可以通过边读边写的方式对目标忆阻器单元施加编程脉冲。又如，或者还可以对不同的忆阻器单元施加不同的数目的编程脉冲，以实现其电导值的调整。

下面基于 set 操作、reset 操作，结合图 20-图 21，对通过边读边写的方式对目标忆阻器单元施加编程脉冲的具体实现方式进行详细描述。

本申请实施例可以基于增量步进编程脉冲（incremental step program pulse，ISPP）策略向目标忆阻器单元写目标数据。具体的，ISPP 策略通常采用“读验证-矫正”的方式调整目标忆阻器单元的电导，使得目标忆阻器单元的电导最终调整至目标数据对应的目标电导。

参见图 20，器件 1、器件 2…为选中的忆阻器阵列中的目标忆阻器单元。首先，可以向目标忆阻器单元施加一个读脉冲（V_{read}），读出目标忆阻器单元的当前电导。与目标电导进行比较，如果当前电导小于目标电导，则可以向目标忆阻器单元加载 set 脉冲（V_{set}）以增大目标忆阻器单元的电导。然后再通过读脉冲（V_{read}）读取调整后的电导，如果当前还电导小于目标电导，则再向目标忆阻器单元加载 set 脉冲（V_{set}），使得目标忆阻器单元的电导向目标电导调整。

参见图 21，器件 1、器件 2…为选中的忆阻器阵列中的目标忆阻器单元。首先，可以向目标忆阻器单元施加一个读脉冲（V_{read}），读出目标忆阻器单元的当前电导。与目标电导进行比较，如果当前电导大于目标电导，则可以向目标忆阻器单元加载 reset 脉冲（V_{reset}）以减小目标忆阻器单元的电导。然后再通过读脉冲（V_{read}）读取调整后的电导，如果当前还电导大于目标电导，则再向目标忆阻器单元加载 set 脉冲（V_{reset}），使得目标忆阻器单元的电导向目标电导调整。

应理解，V_{read} 可以是一个小于阈值电压的读电压脉冲，V_{set} 或 V_{reset} 可以是一个大于阈值电压的读电压脉冲。

本申请实施例可以分别通过上述边读边写的方式使得目标忆阻器单元的电导最终调整至目标数据对应的目标电导。可选地，终止条件可以是该行所有选中的器件的电导增加量都达到要求。

图 22 是本申请实施例提供的一种神经网络的训练过程的示意性流程图。如图 22 所示，该方法可以包括步骤 2210-2255，下面分别对步骤 2210-2255 进行详细描述。

步骤 2210：根据神经网络信息确定需要加速的网络层。

本申请实施例中可以根据以下中的一种或多种确定需要加速的网络层：神经网络的层数，参数信息，训练数据集的大小等。

步骤 2215：在外部个人电脑（personal computer，PC）进行离线训练确定初始训练权重。

可以在外部 PC 上通过前向计算和反向计算等步骤对神经网络的神经元上的权重参数进行训练，确定初始训练权重。

步骤 2220：将初始训练权重分别映射到存算一体架构中实现并行加速的网络层计算的神经网络阵列以及实现非并行加速的网络层计算的神经网络阵列。

本申请实施例中可以根据图 3 所示的方法，将初始训练权重分别映射到存算一体架构多个神经网络阵列的至少一个存算单元中，从而，可以通过神经网络阵列实现输入数据与配置的权重的矩阵乘加操作。

多个神经网络阵列中可以包括实现非并行加速的网络层计算的神经网络阵列以及实现并行加速的网络层计算的神经网络阵列。

步骤 2225：将一组训练数据输入到存算一体架构的多个神经网络阵列，得到基于存算一体架构的实际硬件的前向计算的输出结果。

步骤 2230：判断神经网络系统的准确率是否达到要求或是否达到预设训练次数。

如果神经网络系统的准确率达到要求或达到预设训练次数，可以执行步骤 2235。

如果神经网络系统的准确率未达到要求或未达到预设训练次数，可以执行步骤 2240。

步骤 2235：训练结束。

步骤 2240：判断训练数据是否是最后一组训练数据。

如果训练数据是最后一组训练数据，可以执行步骤 2245，步骤 2255。

如果训练数据不是最后一组训练数据，可以执行步骤 2250-2255。

步骤 2245：重新加载训练数据。

步骤 2250：基于提出的并行训练存算一体系统的训练方法，经过反向传播等计算，对并行加速阵列或其他阵列的电导权重进行片上原位训练和更新。

具体的更新方法请参见上文中的说明，此处不再赘述。

步骤 2255：加载下一组训练数据。

加载下一组训练数据后，继续执行步骤 2225 中的操作。也就是说，将加载的训练数据输入到存算一体架构的多个神经网络阵列，得到基于存算一体架构的实际硬件的前向计算的输出结果。

应理解，在本申请的各种实施例中，上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

上文结合图 1 至图 22，详细描述了本申请实施例提供的一种神经网络系统中数据处理的方法，下面结合图 23 详细描述本申请的装置的实施例。应理解，方法实施例的描述与装置实施例的描述相互对应，因此，未详细描述的部分可以参见前面方法实施例。

需要说明的是，装置实施例是从产品和设备角度考虑本申请方案的实现，本申请的装置实施例和前面描述的本申请方法实施例的部分内容是对应的或者互为补充的。在涉及到方案的实现，以及对权利要求范围的支持上是可以通用的。

下面结合图 23 描述本申请的装置侧实施例。

图 23 是本申请实施例提供的一种神经网络系统 2300 的示意性结构图。应理解，图 23 示出的神经网络系统 2300 仅是示例，本申请实施例的装置还可包括其他模块或单元。应理解，神经网络系统 2300 能够执行图 10 -图 22 的方法中的各个步骤，为了避免重复，此处不再详述。

如图 23 所示，神经网络系统 2300 可以包括：

处理模块 2310，用于将训练数据输入神经网络系统得到第一输出数据，其中，所述神经网络系统中包括多个神经网络阵列，所述多个神经网络阵列中的每个神经网络阵列中包括多个存算单元，每个存算单元用于存储对应的神经网络中神经元的权重值；

计算模块 2320，用于计算所述第一输出数据和目标输出数据之间的偏差；

调整模块 2330，用于根据所述偏差对所述多个神经网络阵列中的部分神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整，其中，所述部分神经网络阵列用于

实现所述神经网络系统中部分神经网络层的计算。

可选地，一种可能的实现方式中，所述多个神经网络阵列包括第一神经网络阵列和第二神经网络阵列，所述第一神经网络阵列的输入数据包括所述第二神经网络阵列的输出数据。

在第二方面的另一种可能的实现方式中，所述第一神经网络阵列包括用于实现神经网络中的全连接层计算的神经网络阵列。

可选地，另一种可能的实现方式中，所述调整模块 2330 具体用于:

根据所述第一神经网络阵列的输入值和所述偏差，对所述第一神经网络阵列中至少一个所述存算单元中存储的权重值进行调整。

可选地，另一种可能的实现方式中，所述多个神经网络阵列还包括第三神经网络阵列，所述第三神经网络阵列和所述第二神经网络阵列用于并行实现神经网络中的卷积层的计算。

可选地，另一种可能的实现方式中，

所述调整模块 2330 具体用于:

根据所述第二神经网络阵列的输入数据和所述偏差，对所述第二神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整;

根据所述第三神经网络阵列的输入数据和所述偏差，对所述第三神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整。

在第二方面的另一种可能的实现方式中，

所述调整模块 2330 具体用于:

将所述偏差划分为至少两个子偏差，其中，所述至少两个子偏差中的第一子偏差与所述第二神经网络阵列的输出数据对应，所述至少两个子偏差中的第二子偏差与所述第三神经网络阵列的输出数据对应;

根据第一子偏差和所述第二神经网络阵的输入数据，对所述第二神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整;

根据第二子偏差和所述第三神经网络阵的输入数据，对所述第三神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整。

可选地，另一种可能的实现方式中，调整模块 2330 具体用于：根据存算单元中更新后的权重值确定脉冲个数，根据所述脉冲个数，对所述神经网络阵列中至少一个所述存算单元中存储的权重值进行改写。

应理解，这里的神经网络系统 2300 以功能模块的形式体现。这里的术语“模块”可以通过软件和/或硬件形式实现，对此不作具体限定。例如，“模块”可以是实现上述功能的软件程序、硬件电路或二者结合。当以上任一模块以软件实现的时候，所述软件以计算机程序指令的方式存在，并被存储在存储器中，处理器可以用于执行所述程序指令以实现以上方法流程。所述处理器可以包括但不限于以下至少一种：中央处理单元（central processing unit，CPU）、微处理器、数字信号处理器（digital signal processing，DSP）、微控制器（microcontroller unit，MCU）、或人工智能处理器等各类运行软件的计算设备，每种计算设备可包括一个或多个用于执行软件指令以进行运算或处理的核。该处理器可以是个单独的半导体芯片，也可以跟其他电路一起集成为一个半导体芯片，例如，可以跟其他电路(如编解码电路、硬件加速电路或各种总线和接口电路)构成一个片上系统(system

on chip，SoC），或者也可以作为一个专用集成电路（application-specific integrated circuit，ASIC）的内置处理器集成在所述 ASIC 当中，该集成了处理器的 ASIC 可以单独封装或者也可以跟其他电路封装在一起。该处理器除了包括用于执行软件指令以进行运算或处理的核外，还可进一步包括必要的硬件加速器，如现场可编程门阵列（field programmable gate array，FPGA）、可编程逻辑器件（programmable logic device，PLD）、或者实现专用逻辑运算的逻辑电路。

当以上模块以硬件电路实现的时候，所述硬件电路可能以通用中央处理器（central processing unit，CPU）、微控制器（micro controller unit，MCU）、微处理器（micro processing unit，MPU）、数字信号处理器（digital signal processing，DSP）、片上系统（system on chip，SoC）来实现，当然也可以采用专用集成电路（application-specific integrated circuit，ASIC）实现，或可编程逻辑器件（programmable logic device，PLD）实现，上述 PLD 可以是复杂程序逻辑器件（complex programmable logical device，CPLD），现场可编程门阵列（field-programmable gate array，FPGA），通用阵列逻辑（generic array logic，GAL）或其任意组合，其可以运行必要的软件或不依赖于软件以执行以上方法流程。

上述实施例，可以全部或部分地通过软件、硬件、固件或其他任意组合来实现。当使用软件实现时，上述实施例可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令或计算机程序。在计算机上加载或执行所述计算机指令或计算机程序时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以为通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线（例如红外、无线、微波等）方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集合的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质（例如，软盘、硬盘、磁带）、光介质（例如，DVD）、或者半导体介质。半导体介质可以是固态硬盘。

应理解，本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况，其中 A，B 可以是单数或者复数。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系，但也可能表示的是一种“和/或”的关系，具体可参考前后文进行理解。

本申请中，“至少一个”是指一个或者多个，“多个”是指两个或两个以上。“以下至少一项(个)”或其类似表达，是指的这些项中的任意组合，包括单项（个）或复数项（个）的任意组合。例如，a，b，或 c 中的至少一项（个），可以表示：a，b，c，a-b，a-c，b-c，或 a-b-c，其中 a，b，c 可以是单个，也可以是多个。

应理解，在本申请的各种实施例中，上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以

硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

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、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请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。例如，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，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。

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，在本申请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。

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，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分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）执行本申请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：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。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权 利 要 求 书

1、一种神经网络系统中数据处理的方法，其特征在于，包括:

将训练数据输入神经网络系统得到第一输出数据，其中，所述神经网络系统中包括多个神经网络阵列，所述多个神经网络阵列中的每个神经网络阵列中包括多个存算单元，每个存算单元用于存储对应的神经网络中神经元的权重值;

计算所述第一输出数据和目标输出数据之间的偏差;

根据所述偏差对所述多个神经网络阵列中的部分神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整，其中，所述部分神经网络阵列用于实现所述神经网络系统中部分神经网络层的计算。

2、根据权利要求1所述的方法，其特征在于，所述多个神经网络阵列包括第一神经网络阵列和第二神经网络阵列，所述第一神经网络阵列的输入数据包括所述第二神经网络阵列的输出数据。

3、根据权利要求2所述的方法，其特征在于，所述第一神经网络阵列包括用于实现神经网络中的全连接层计算的神经网络阵列。

4、根据权利要求3所述的方法，其特征在于，所述根据所述偏差对所述多个神经网络阵列中的部分神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整，包括:

根据所述第一神经网络阵列的输入值和所述偏差，对所述第一神经网络阵列中至少一个所述存算单元中存储的权重值进行调整。

5、根据权利要求2至4中任一项所述的方法，其特征在于，所述多个神经网络阵列还包括第三神经网络阵列，所述第三神经网络阵列和所述第二神经网络阵列用于并行实现神经网络中的卷积层的计算。

6、根据权利要求5所述的方法，其特征在于，所述根据所述偏差对所述多个神经网络阵列中的部分神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整，包括:

根据所述第二神经网络阵列的输入数据和所述偏差，对所述第二神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整;

根据所述第三神经网络阵列的输入数据和所述偏差，对所述第三神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整。

7、根据权利要求5所述的方法，其特征在于，所述根据所述偏差对所述多个神经网络阵列中的部分神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整，包括:

将所述偏差划分为至少两个子偏差，其中，所述至少两个子偏差中的第一子偏差与所述第二神经网络阵列的输出数据对应，所述至少两个子偏差中的第二子偏差与所述第三神经网络阵列的输出数据对应;

根据所述第一子偏差和所述第二神经网络阵的输入数据，对所述第二神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整;

根据所述第二子偏差和所述第三神经网络阵的输入数据，对所述第三神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整。

8、一种神经网络系统，其特征在于，包括:

处理模块，用于将训练数据输入神经网络系统得到第一输出数据，其中，所述神经网络系统中包括多个神经网络阵列，所述多个神经网络阵列中的每个神经网络阵列中包括多个存算单元，每个存算单元用于存储对应的神经网络中神经元的权重值；

计算模块，用于计算所述第一输出数据和目标输出数据之间的偏差；

调整模块，用于根据所述偏差对所述多个神经网络阵列中的部分神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整，其中，所述部分神经网络阵列用于实现所述神经网络系统中部分神经网络层的计算。

9、根据权利要求 8 所述的神经网络系统，其特征在于，所述多个神经网络阵列包括第一神经网络阵列和第二神经网络阵列，所述第一神经网络阵列的输入数据包括所述第二神经网络阵列的输出数据。

10、根据权利要求 9 所述的神经网络系统，其特征在于，所述第一神经网络阵列包括用于实现神经网络中的全连接层计算的神经网络阵列。

11、根据权利要求 10 所述的神经网络系统，其特征在于，所述调整模块具体用于：

根据所述第一神经网络阵列的输入值和所述偏差，对所述第一神经网络阵列中至少一个所述存算单元中存储的权重值进行调整。

12、根据权利要求 9 至 11 中任一项所述的神经网络系统，其特征在于，所述多个神经网络阵列还包括第三神经网络阵列，所述第三神经网络阵列和所述第二神经网络阵列用于并行实现神经网络中的卷积层的计算。

13、根据权利要求 12 所述的神经网络系统，其特征在于，所述调整模块具体用于：

根据所述第二神经网络阵列的输入数据和所述偏差，对所述第二神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整；

根据所述第三神经网络阵列的输入数据和所述偏差，对所述第三神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整。

14、根据权利要求 12 所述的神经网络系统，其特征在于，所述调整模块具体用于：

将所述偏差划分为至少两个子偏差，其中，所述至少两个子偏差中的第一子偏差与所述第二神经网络阵列的输出数据对应，所述至少两个子偏差中的第二子偏差与所述第三神经网络阵列的输出数据对应；

根据所述第一子偏差和所述第二神经网络阵的输入数据，对所述第二神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整；

根据所述第二子偏差和所述第三神经网络阵的输入数据，对所述第三神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整。

15、一种神经网络系统，其特征在于，包括：

存储器，用于存储程序；

处理器，用于从所述存储器中调用并运行所述程序以执行权利要求 1 至 7 中任一项所述的方法。

16、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储用于设备执行的程序代码，该程序代码包括用于执行如权利要求 1 至 7 中任一项所述的方法。

17、一种芯片，其特征在于，所述芯片包括处理器与数据接口，所述处理器通过所述数据接口读取存储器上存储的指令，以执行如权利要求 1 至 7 中任一项所述的方法。

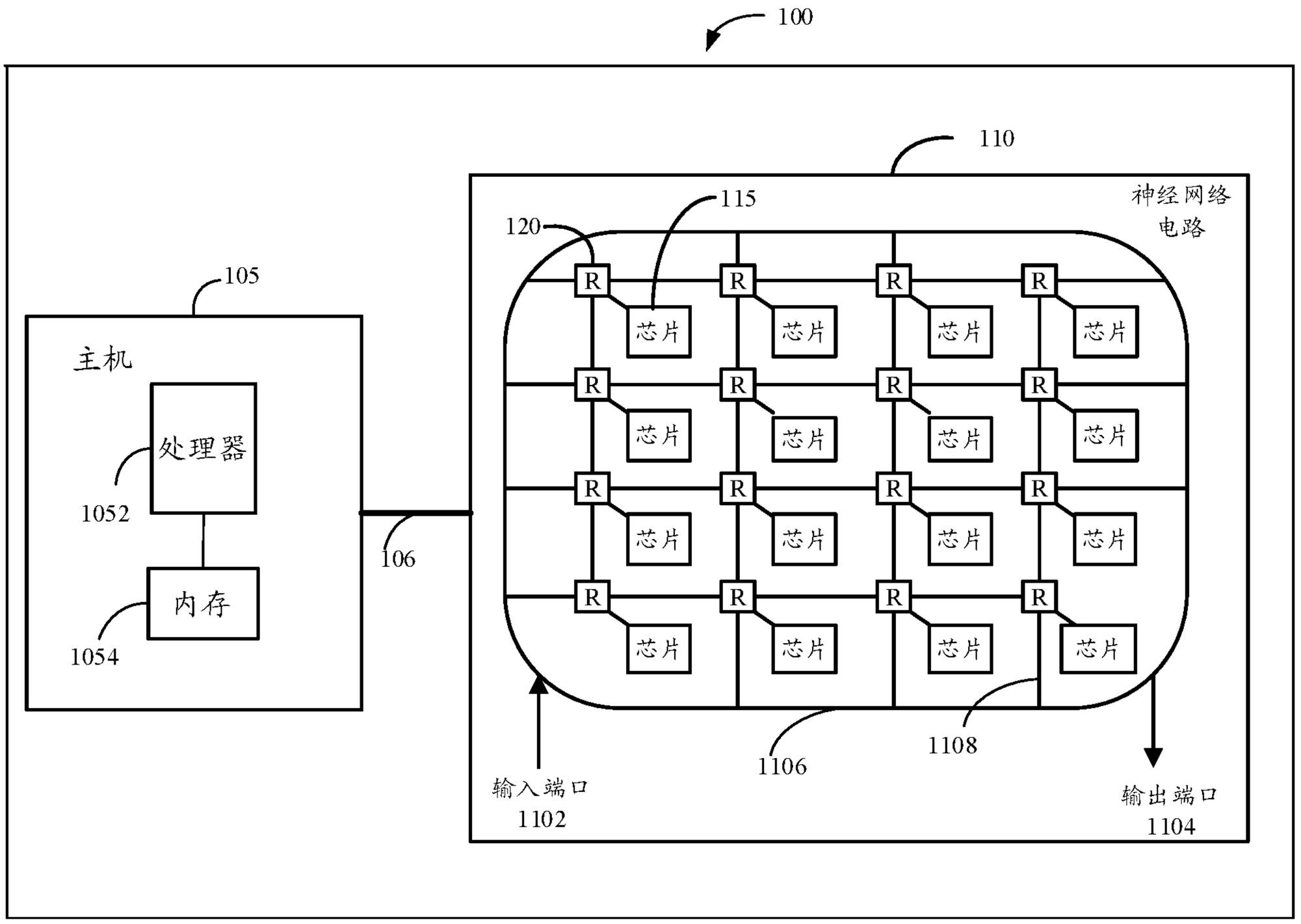

100
110
神经网络
电路
115
120
R
芯片
105
主机
处理器
1052
内存
1054
106
输入端口
1102
1106
1108
输出端口
1104

图 1

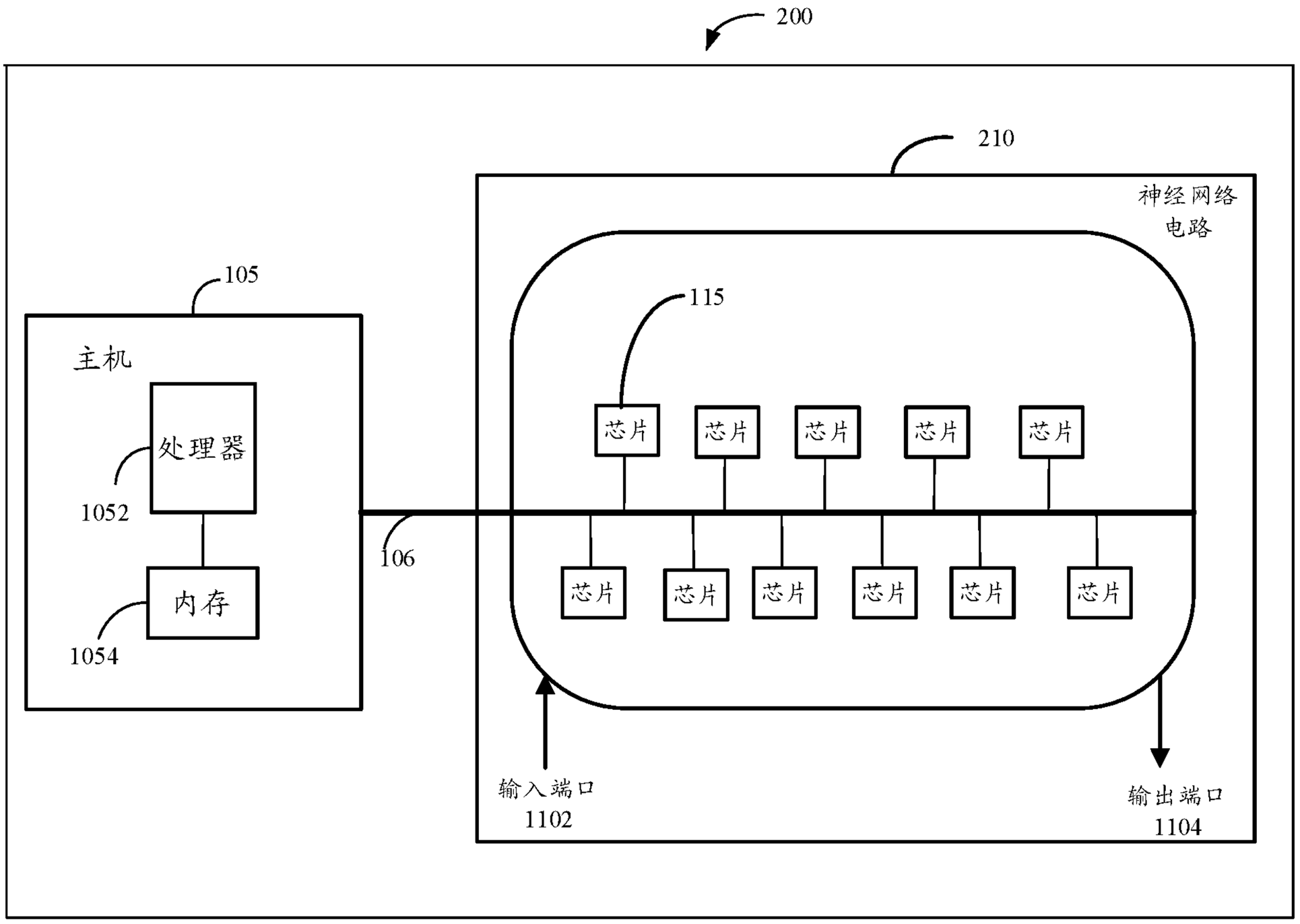

200
210
神经网络
电路
115
芯片
105
主机
处理器
1052
内存
1054
106
输入端口
1102
输出端口
1104

图 2

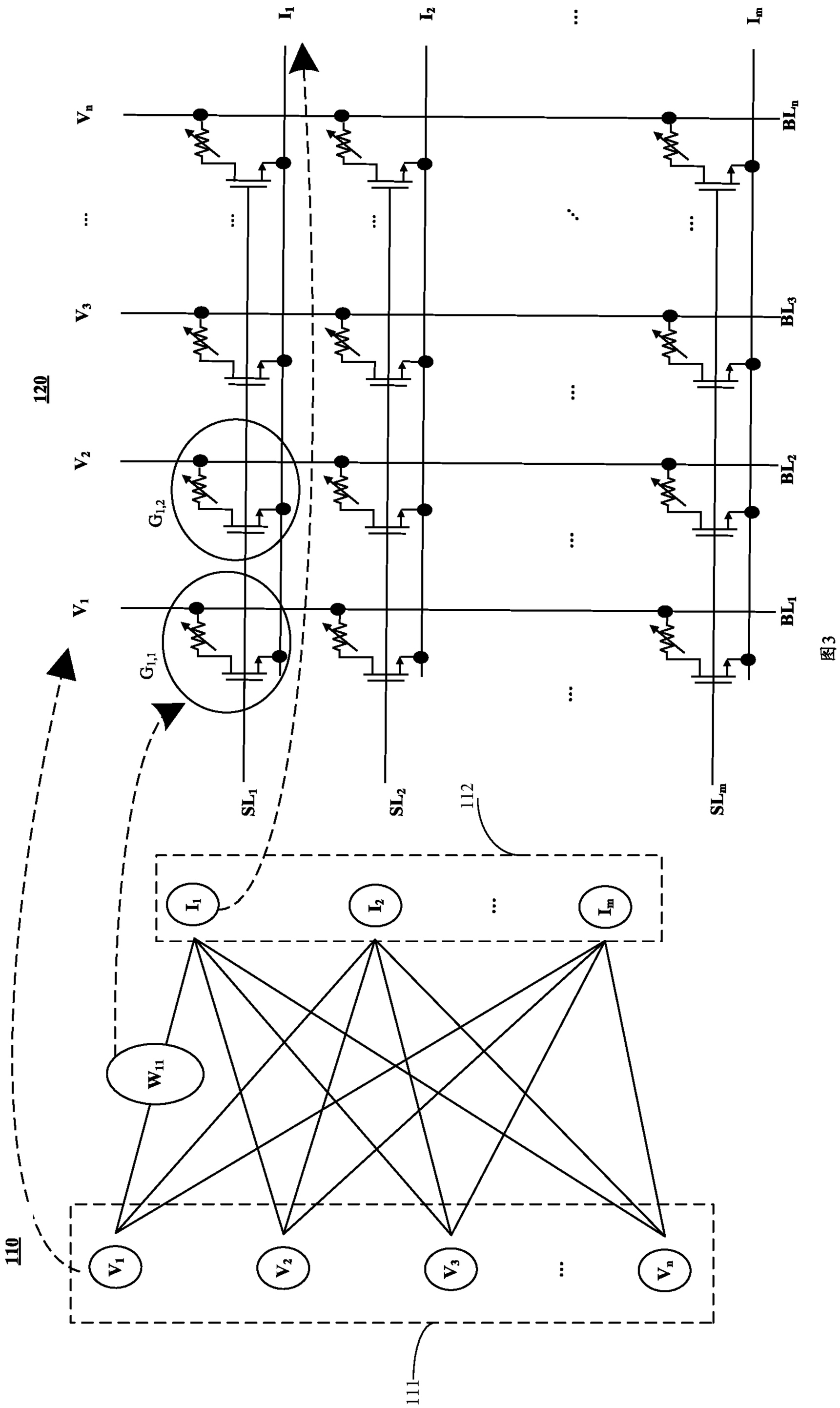

I1
I2
...
Im
Vn
BLn
V3
V2
V1
BL3
BL2
BL1
120
G1,2
G1,1
SL1
SL2
SLm
112
I1
I2
...
Im
W11
110
V1
V2
V3
...
Vn
111

图3

$$\begin{matrix} W_{0,0} & W_{0,1} & * & * & * & W_{0,n-1} \\ W_{1,0} & W_{1,1} & * & * & * & W_{1,n-1} \\ W_{2,0} & W_{2,1} & * & * & * & W_{2,n-1} \\ W_{3,0} & W_{3,1} & * & * & * & W_{3,n-1} \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ W_{m-1,0} & W_{m-1,1} & * & * & * & W_{m-1,n-1} \end{matrix}$$

图 4

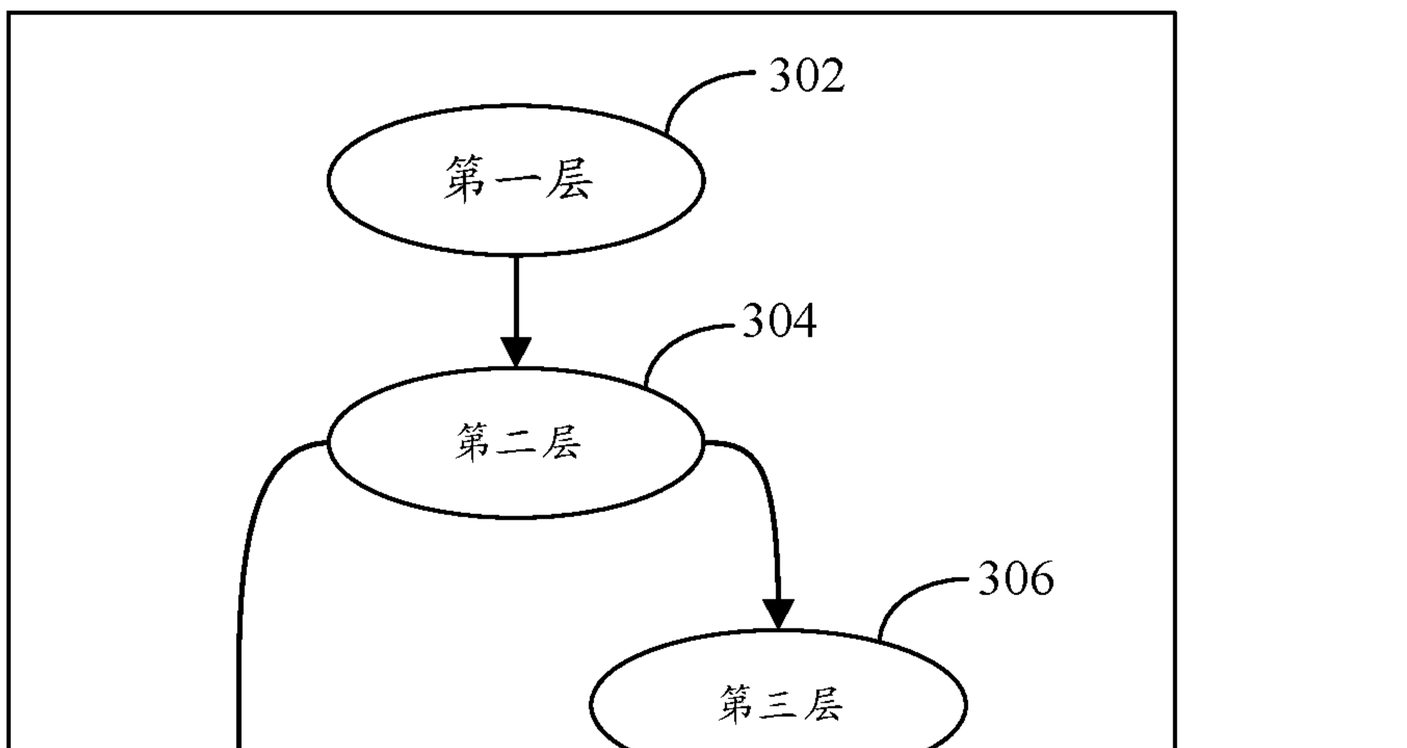


图 5

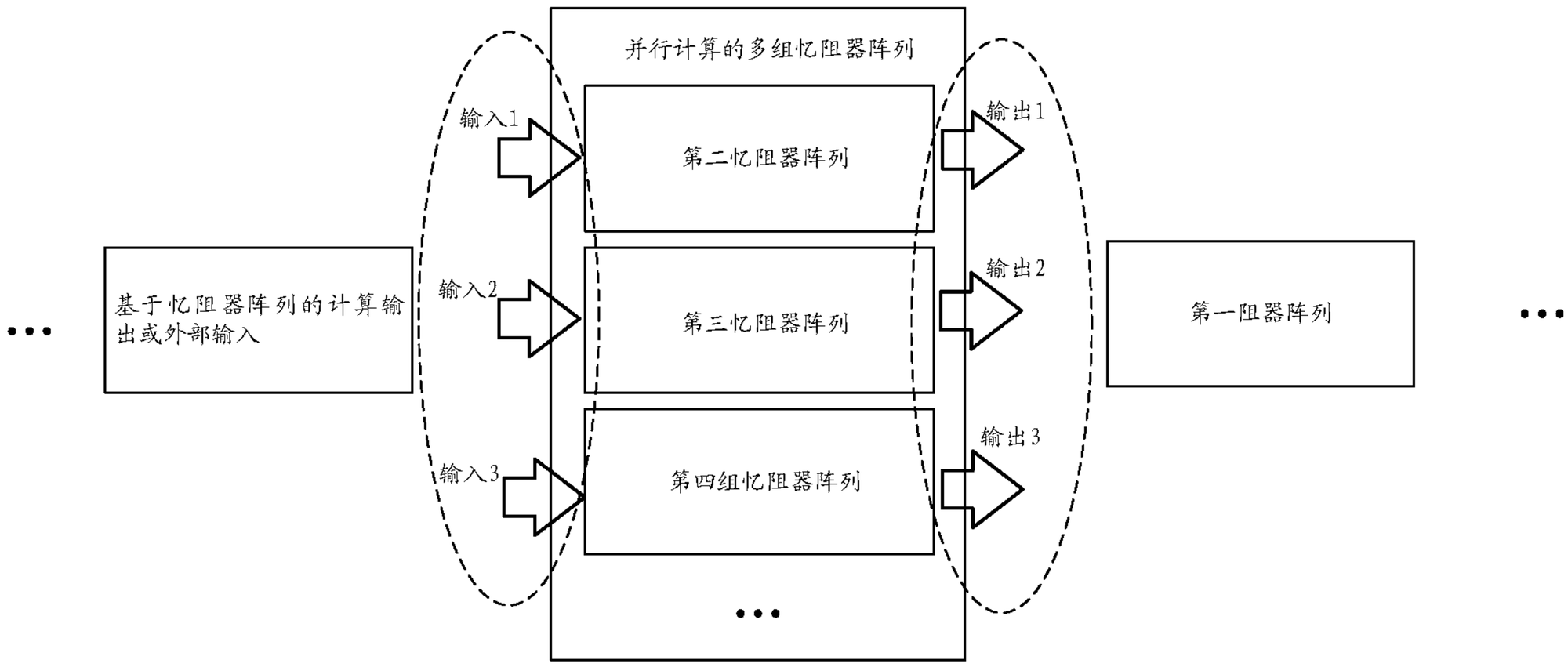

并行计算的多组忆阻器阵列
输入1
第二忆阻器阵列
输出1
基于忆阻器阵列的计算输出或外部输入
输入2
第三忆阻器阵列
输出2
第一阻器阵列
输入3
第四组忆阻器阵列
输出3

图 6

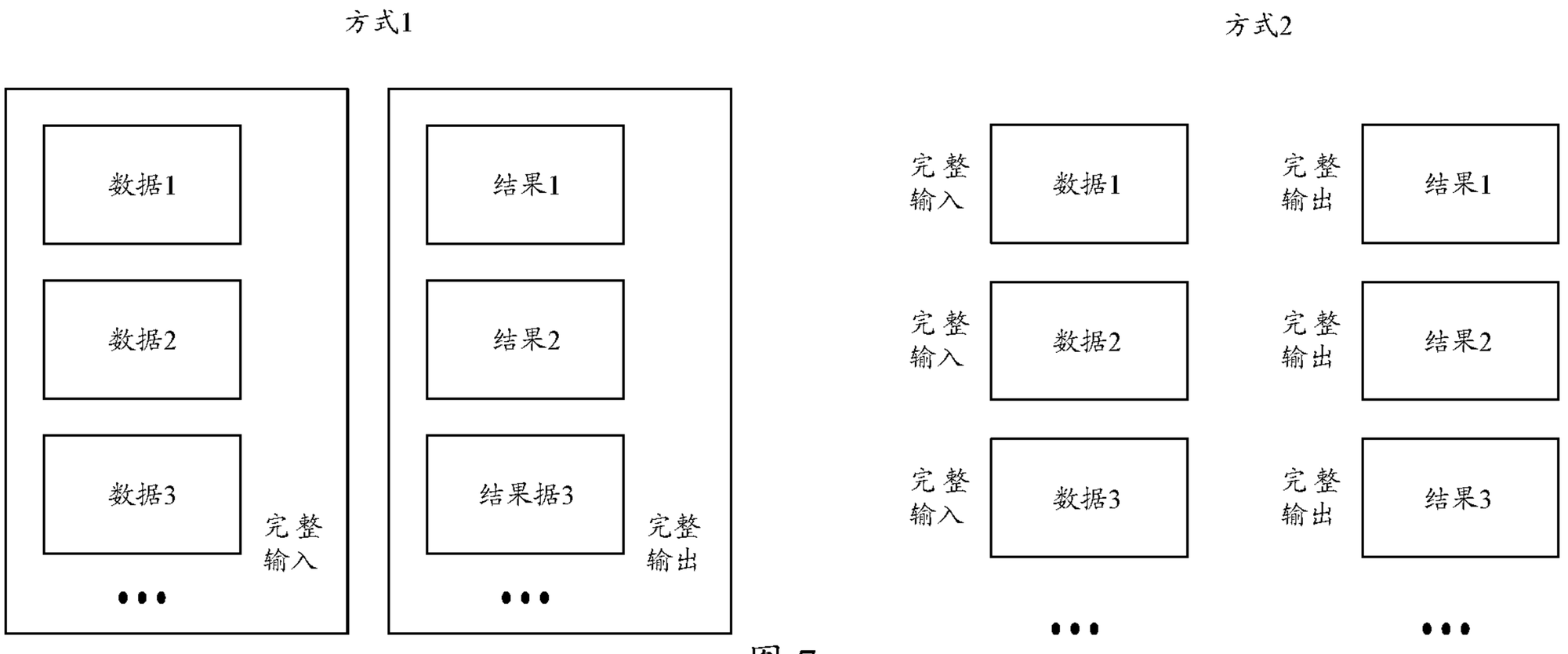
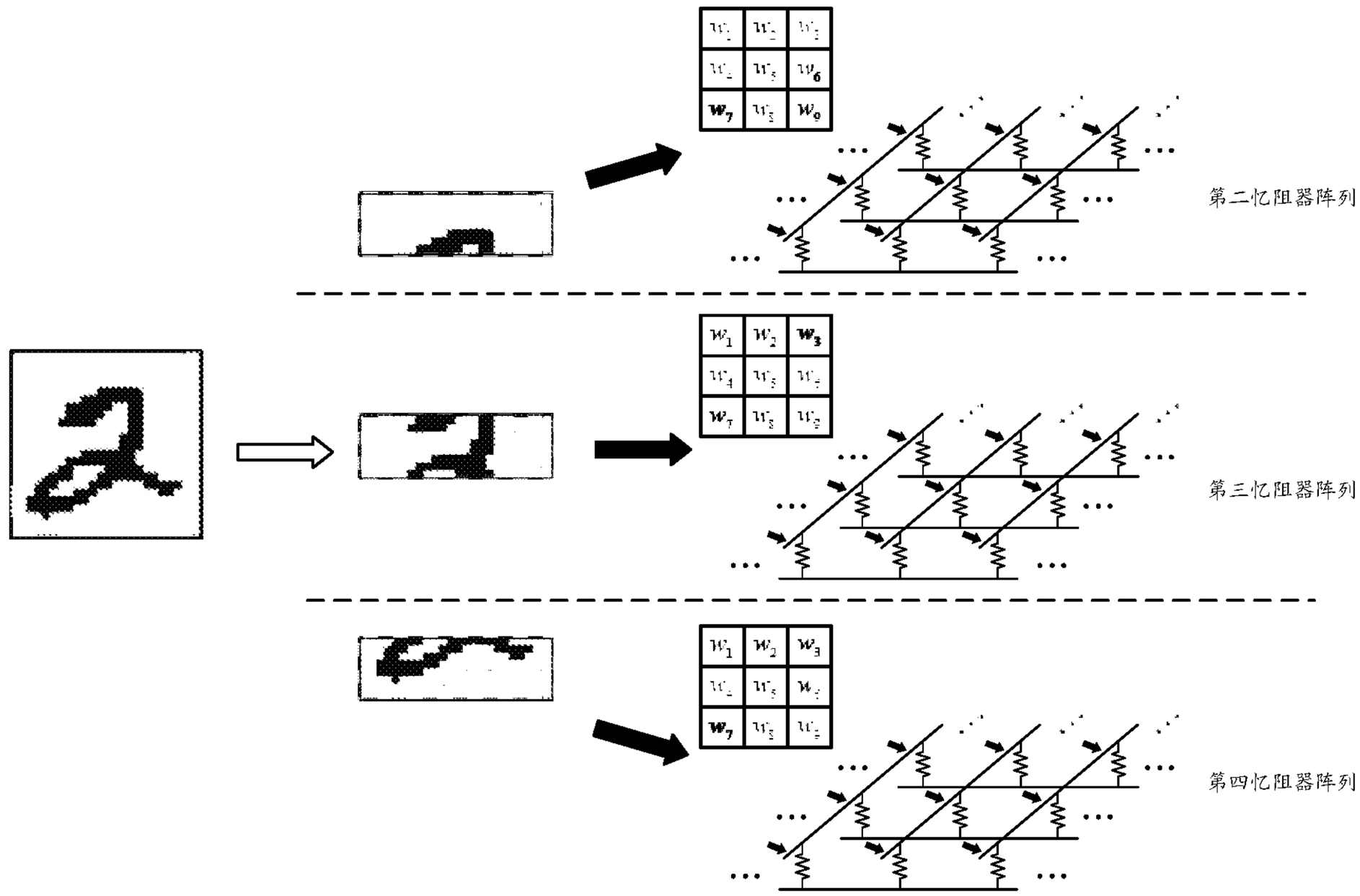
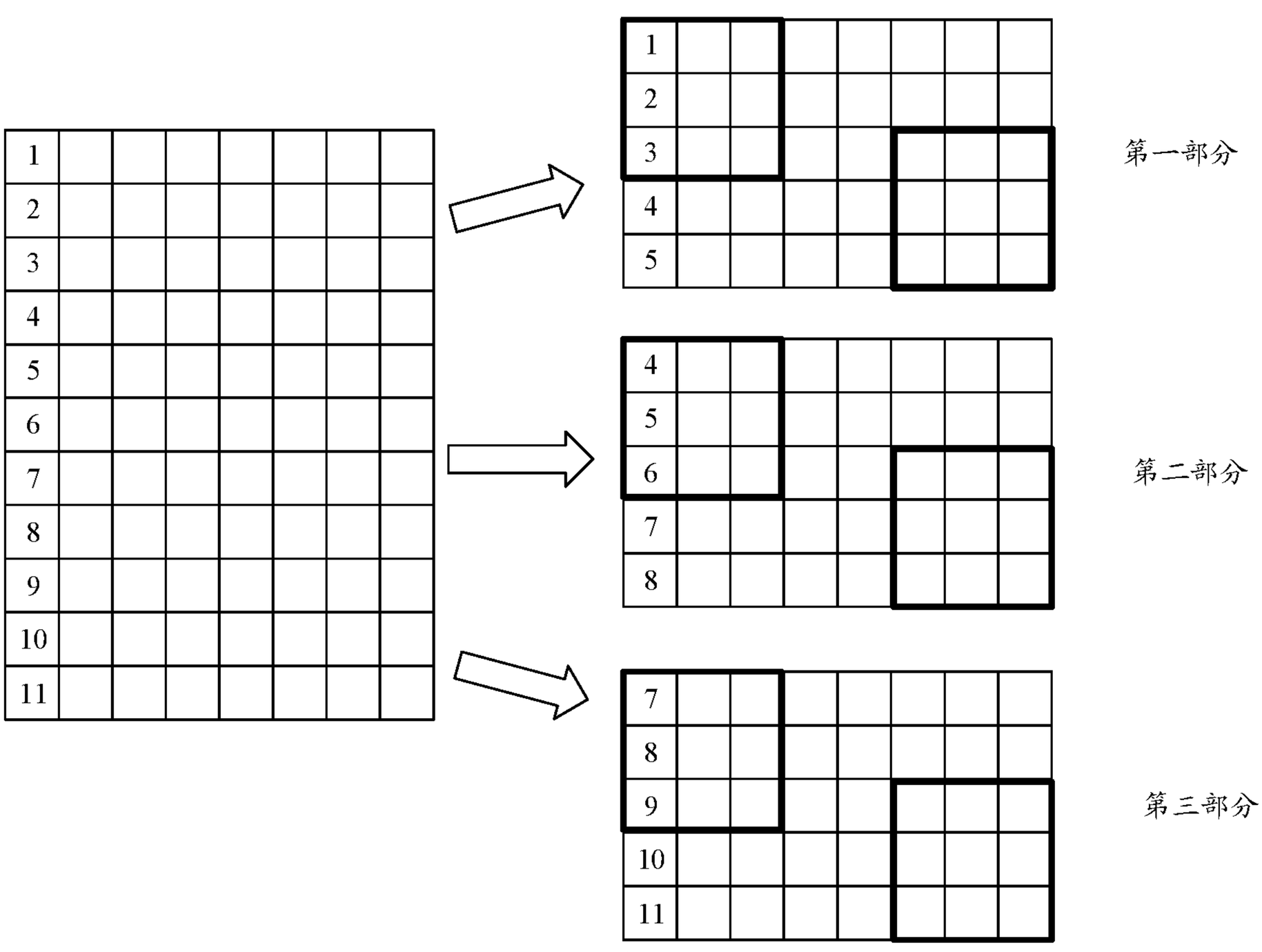

方式1
数据1
数据2
数据3
完整输入
结果1
结果2
结果据3
完整输出
方式2
完整输入
数据1
数据2
数据3
完整输入
完整输入
完整输出
结果1
结果2
结果3
完整输出
完整输出

图 7



(a)

以 3x3 的卷积核为例，slide = 1



(b)

图 8

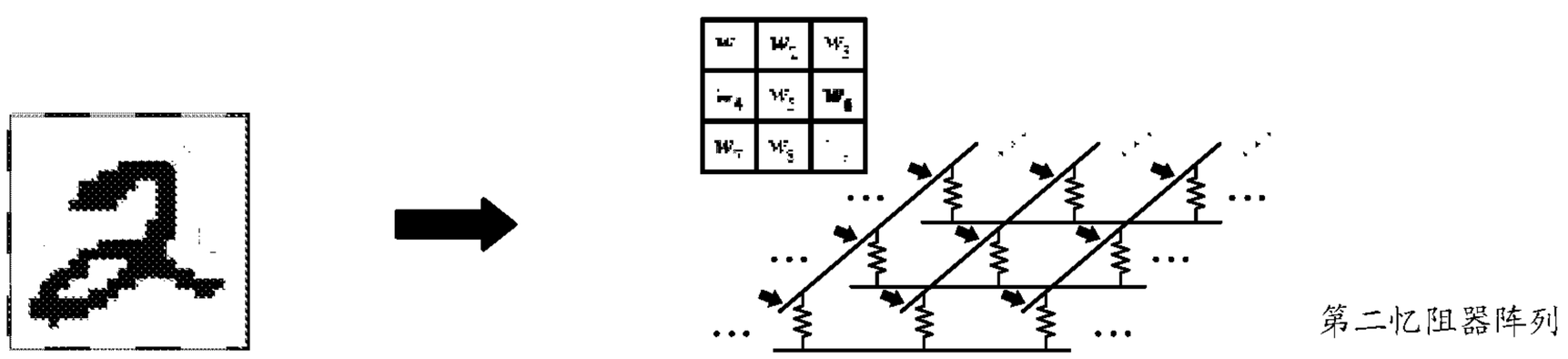

第二忆阻器阵列
第三忆阻器阵列
第四忆阻器阵列

图 9

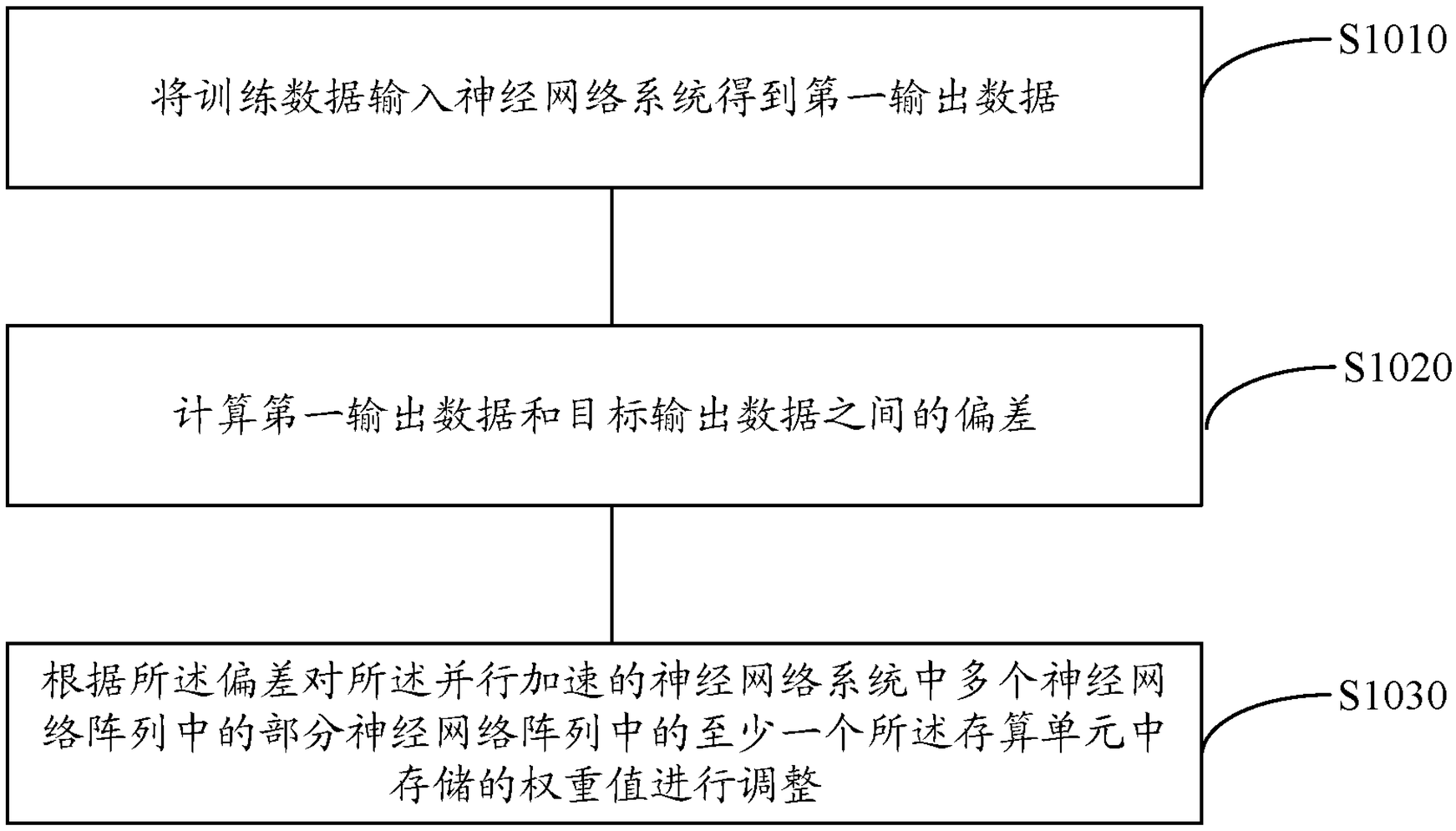
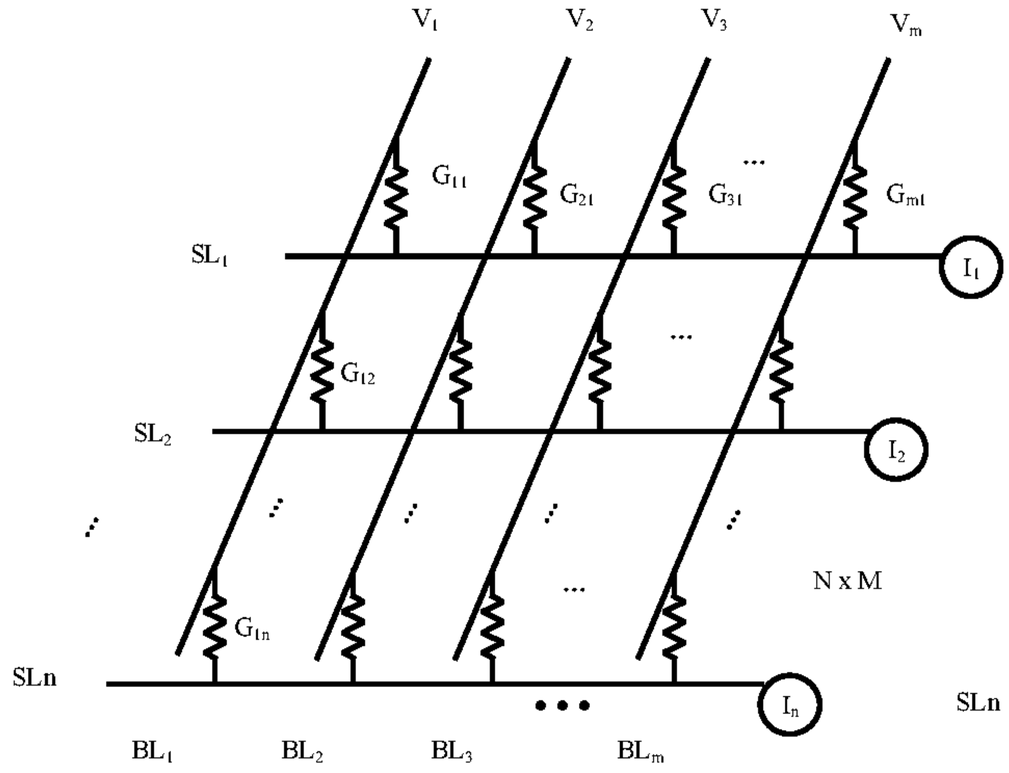
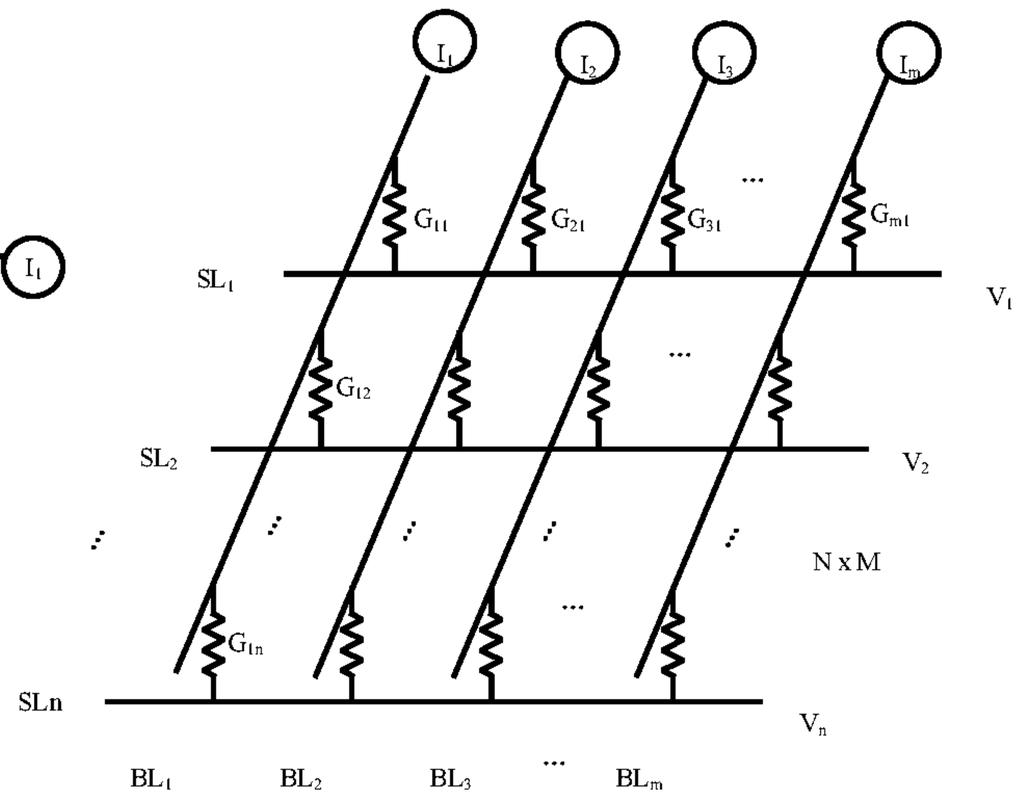

将训练数据输入神经网络系统得到第一输出数据
S1010
计算第一输出数据和目标输出数据之间的偏差
S1020
根据所述偏差对所述并行加速的神经网络系统中多个神经网络阵列中的部分神经网络阵列中的至少一个所述存算单元中存储的权重值进行调整
S1030

图 10



前向过程的操作：Y=XW



反向过程的操作：$Y=XW^T$

图 11

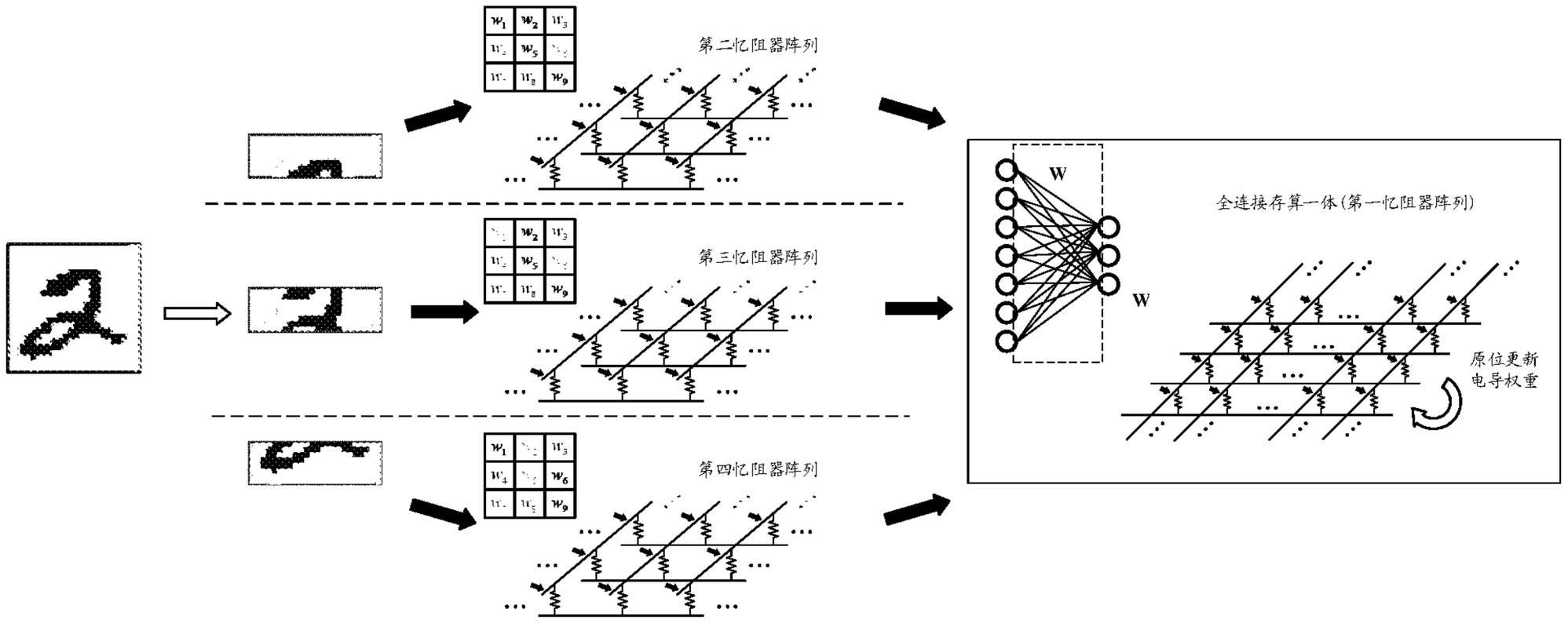


图 12

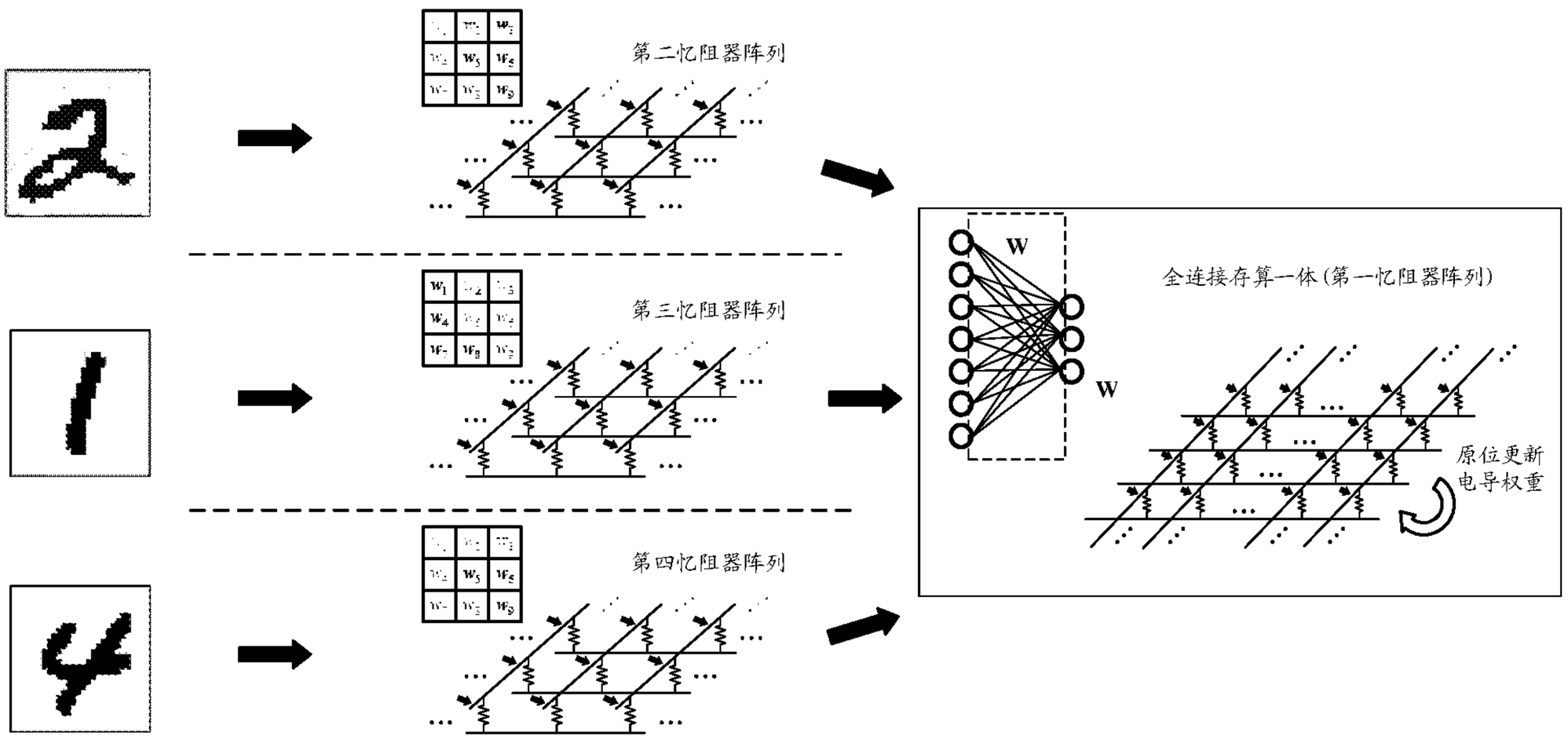

第二忆阻器阵列
第三忆阻器阵列
第四忆阻器阵列
W
W
全连接存算一体(第一忆阻器阵列)
原位更新
电导权重

图 13

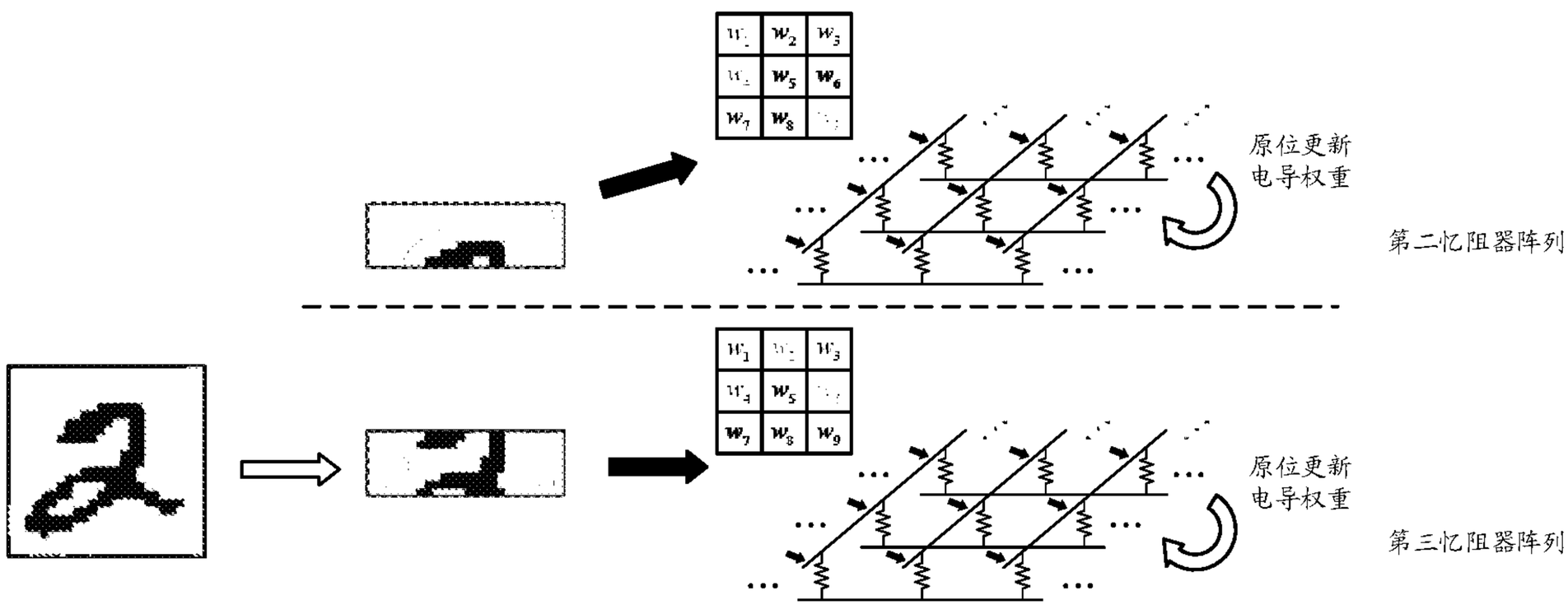

原位更新
电导权重
第二忆阻器阵列
原位更新
电导权重
第三忆阻器阵列
原位更新
电导权重
第四忆阻器阵列

图 14

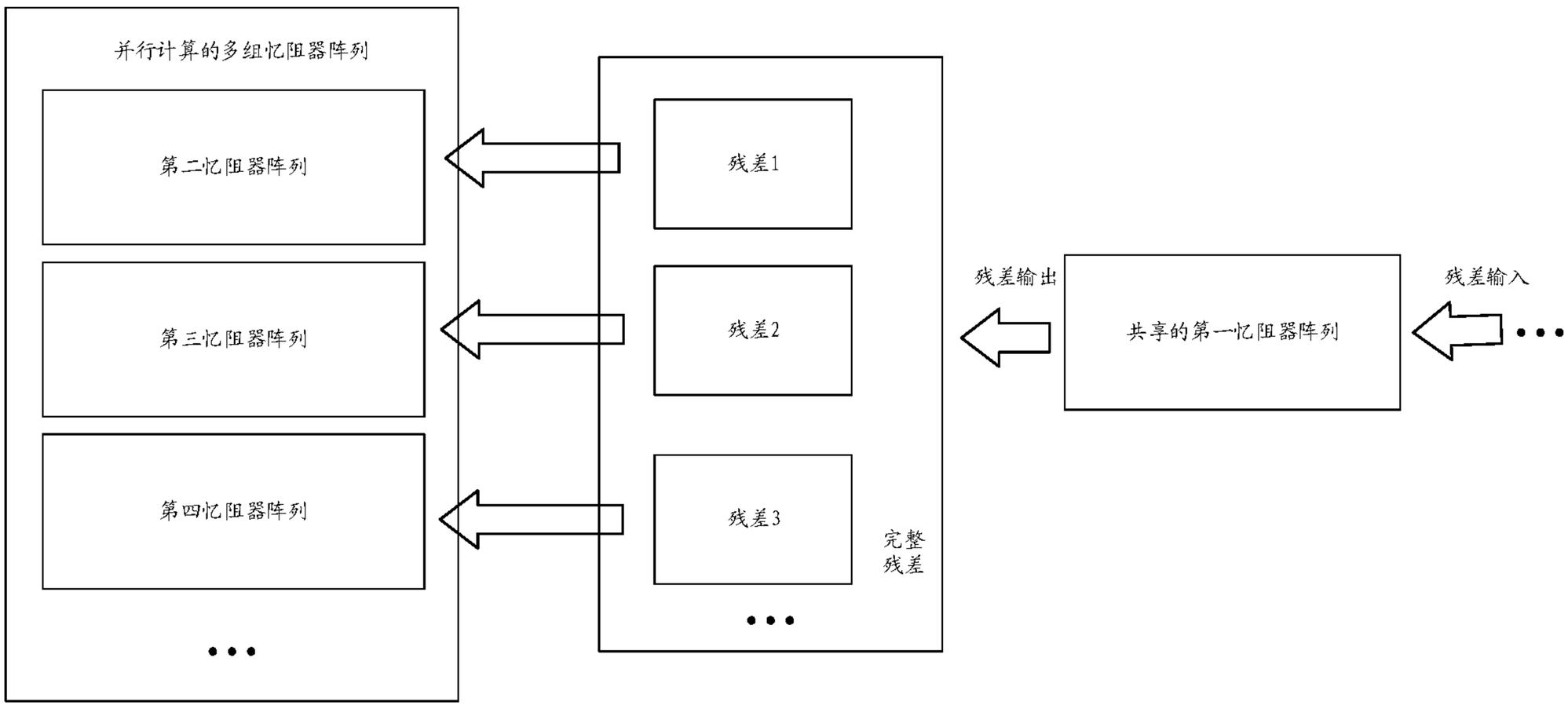

并行计算的多组忆阻器阵列
第二忆阻器阵列
第三忆阻器阵列
第四忆阻器阵列
残差1
残差2
残差3
完整残差
残差输出
共享的第一忆阻器阵列
残差输入

图 15

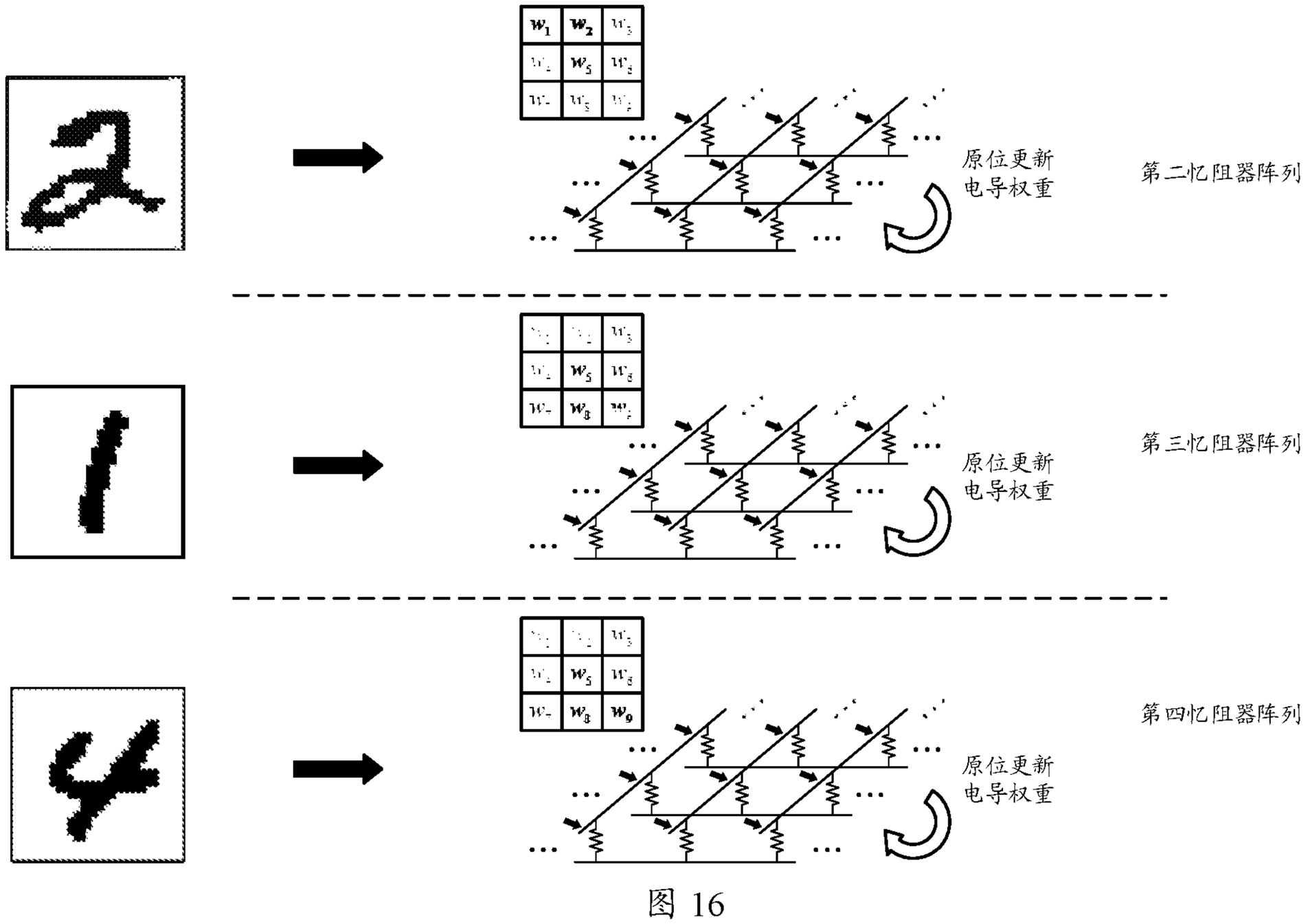

原位更新
电导权重
第二忆阻器阵列
第三忆阻器阵列
第四忆阻器阵列

图 16

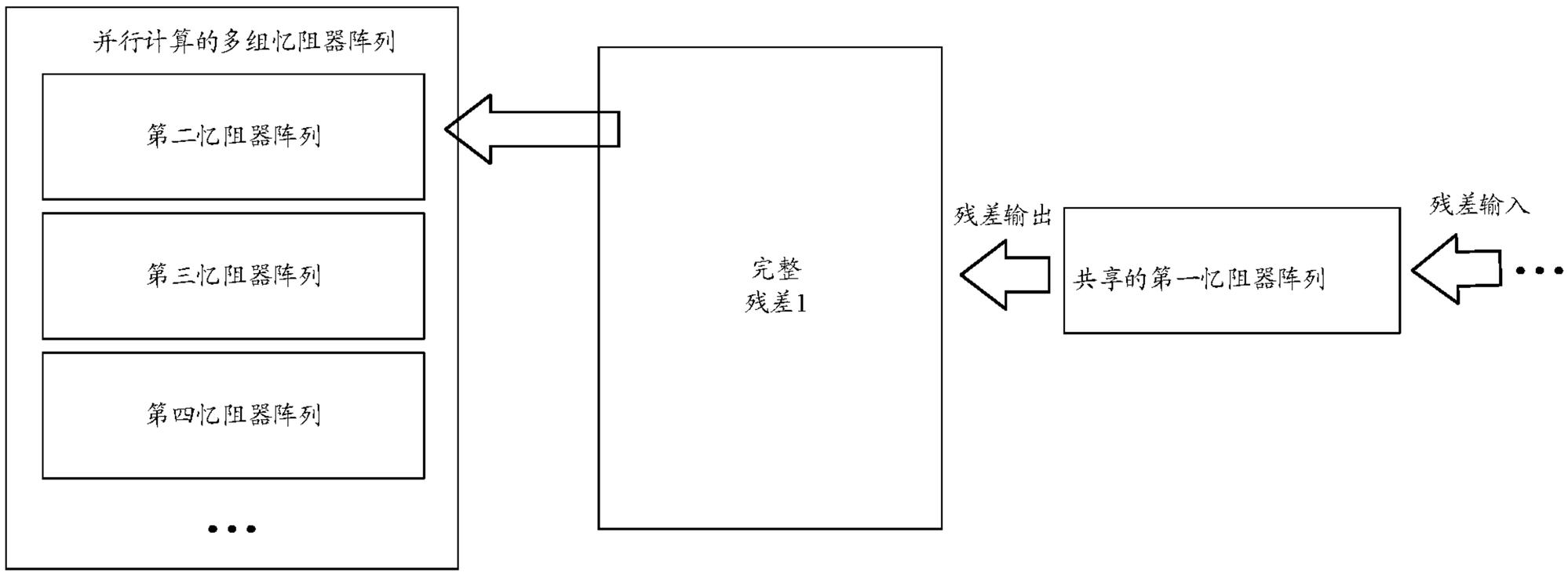
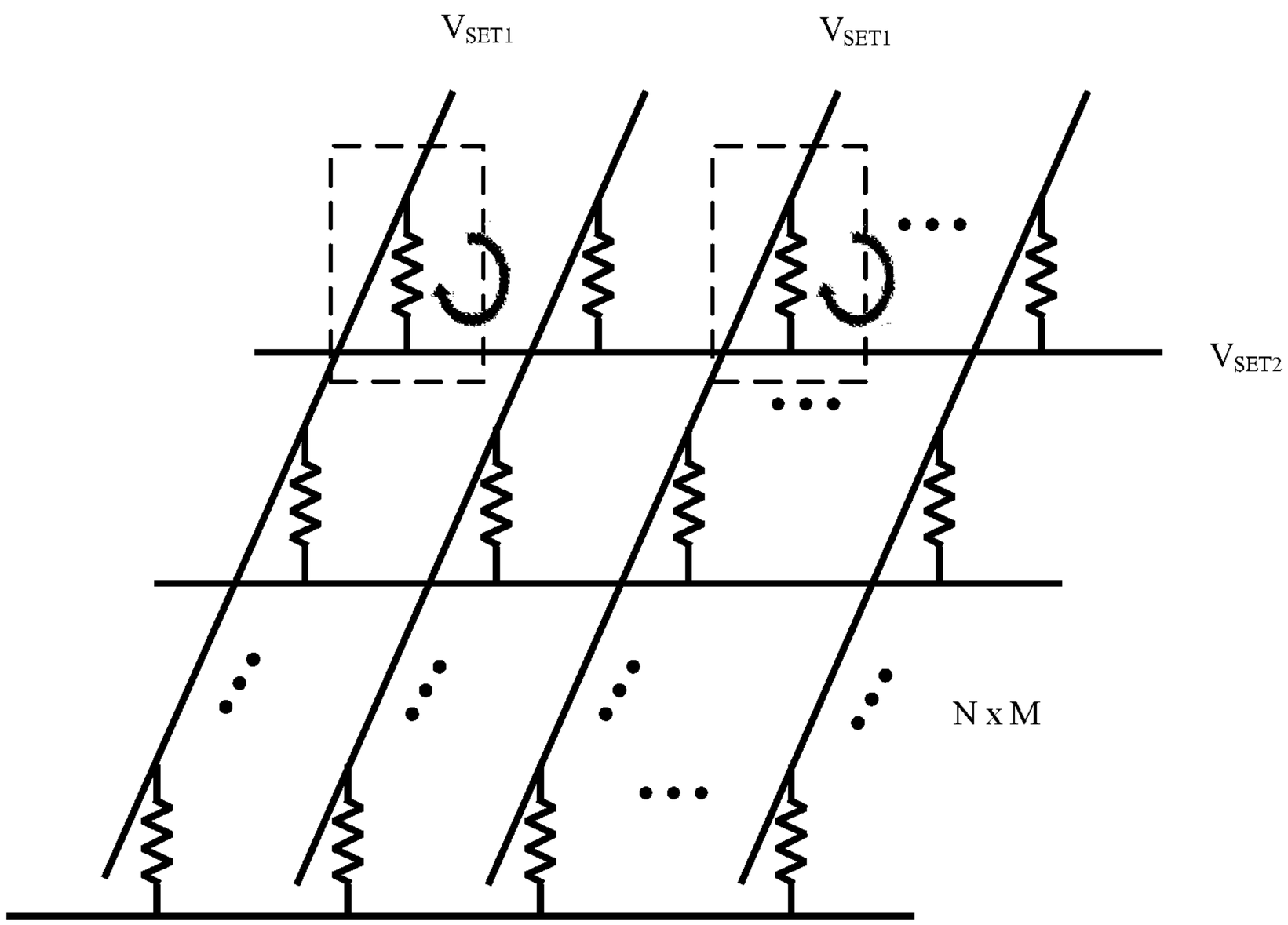
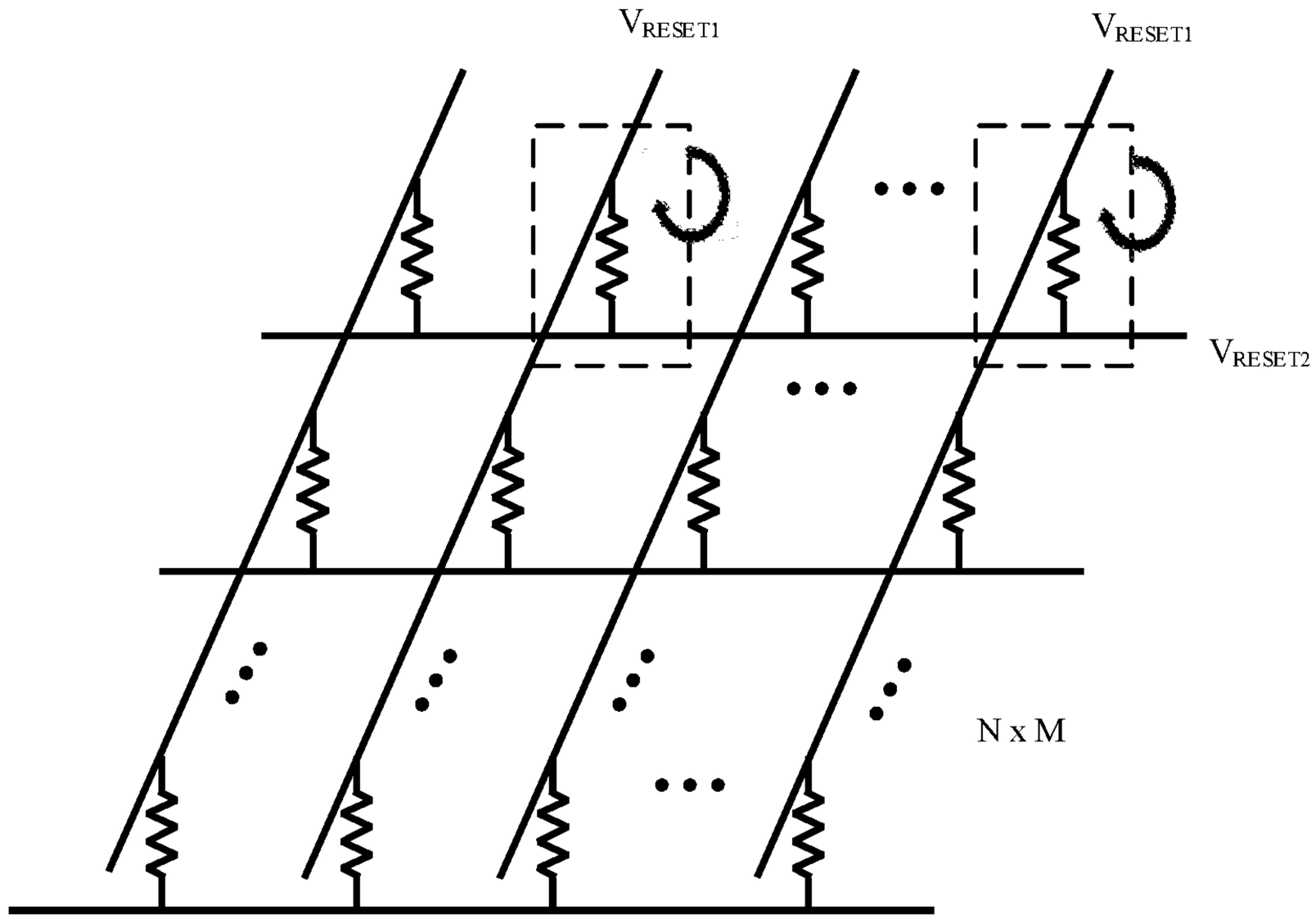


图 17



对选中的器件并行执行SET，增加器件电导

图 18



对选中的器件并行执行RESET，减小器件电导

图 19

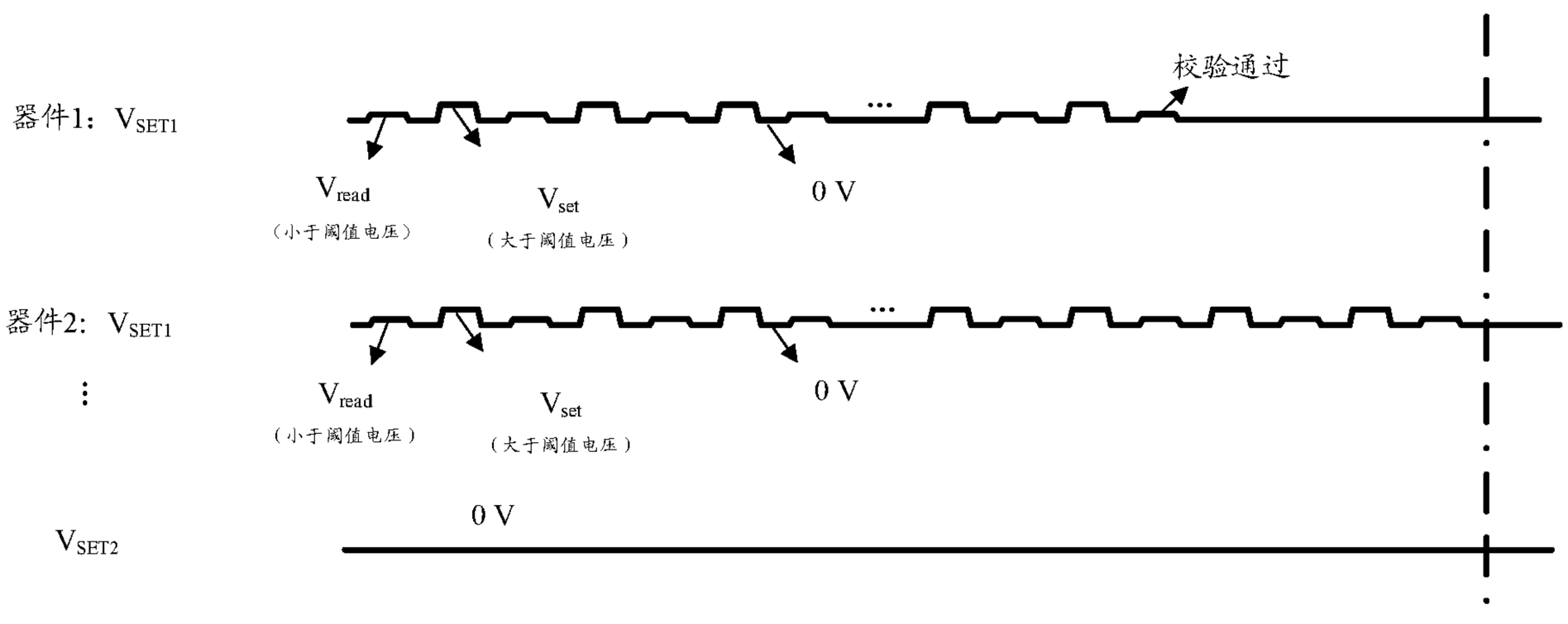


图 20

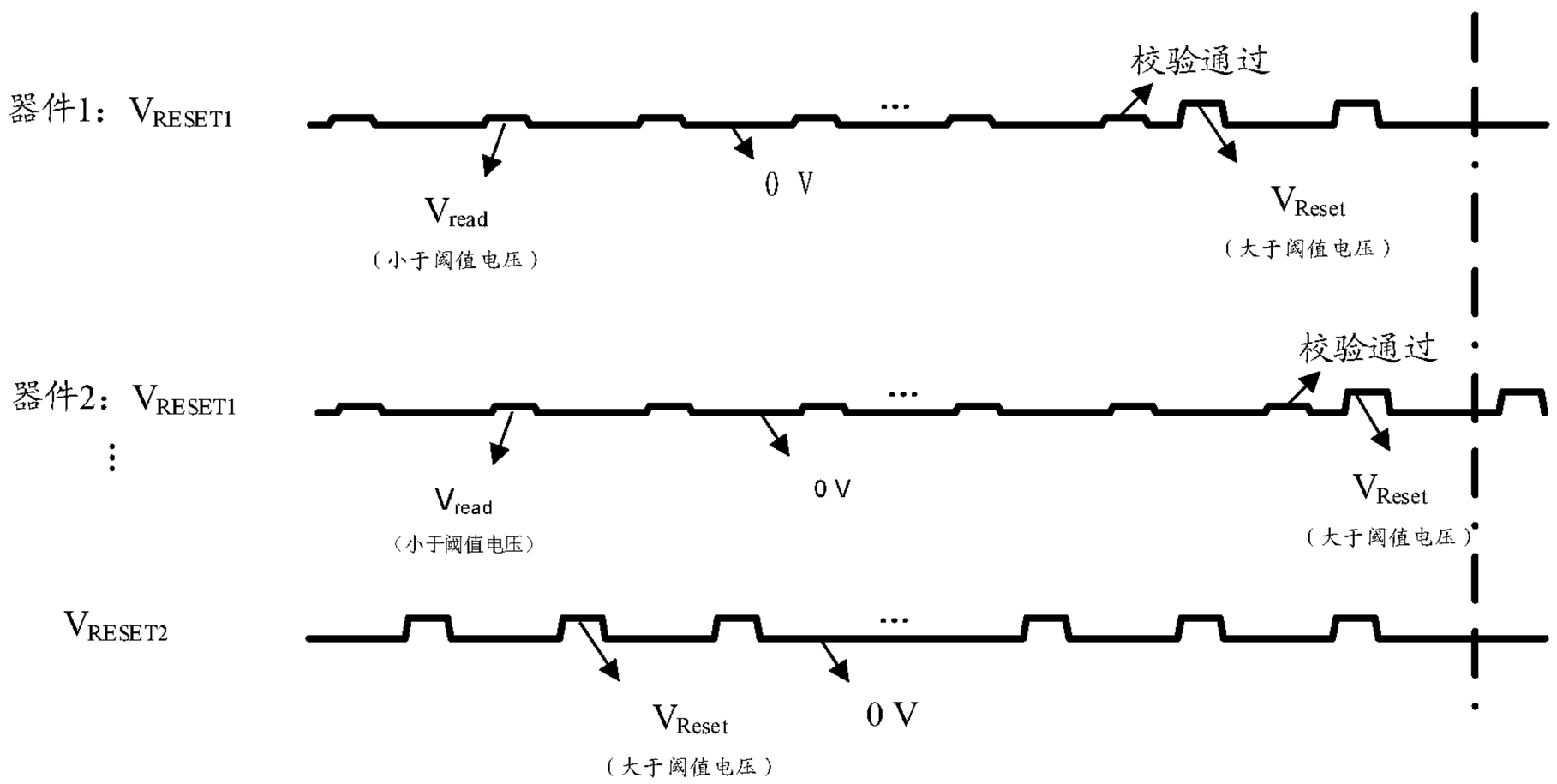

器件1：V_{RESET1}
校验通过
V_{read}
（小于阈值电压）
0 V
V_{Reset}
（大于阈值电压）
器件2：V_{RESET1}
校验通过
V_{read}
（小于阈值电压）
0 V
V_{Reset}
（大于阈值电压）
V_{RESET2}
V_{Reset}
（大于阈值电压）
0 V

图 21

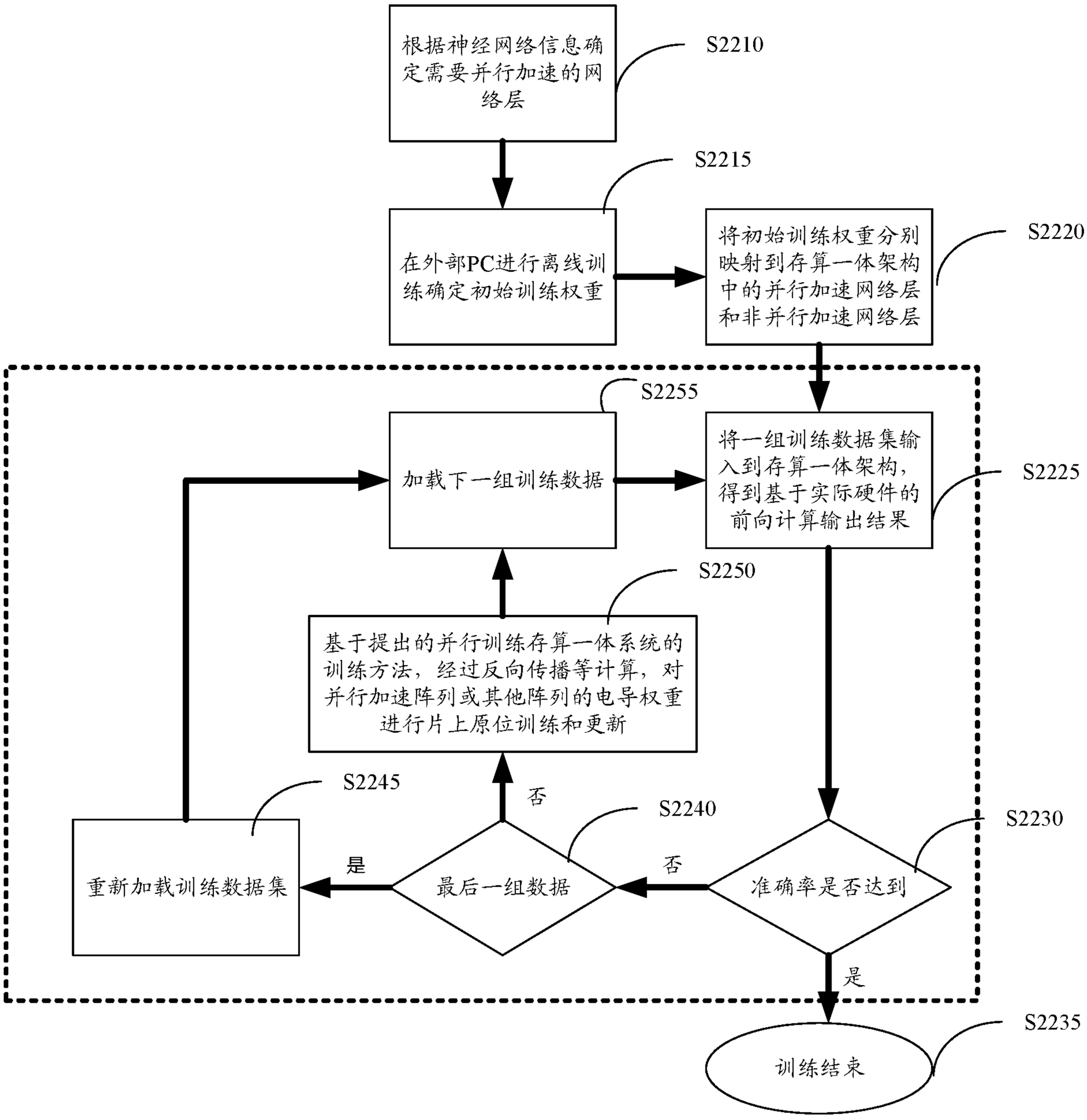

根据神经网络信息确定需要并行加速的网络层
S2210
在外部PC进行离线训练确定初始训练权重
将初始训练权重分别映射到存算一体架构中的并行加速网络层和非并行加速网络层
S2220
S2215
S2255
加载下一组训练数据
将一组训练数据集输入到存算一体架构，得到基于实际硬件的前向计算输出结果
S2225
S2250
基于提出的并行训练存算一体系统的训练方法，经过反向传播等计算，对并行加速阵列或其他阵列的电导权重进行片上原位训练和更新
S2245
否
S2240
S2230
是
否
重新加载训练数据集
最后一组数据
准确率是否达到
是
S2235
训练结束

图 22

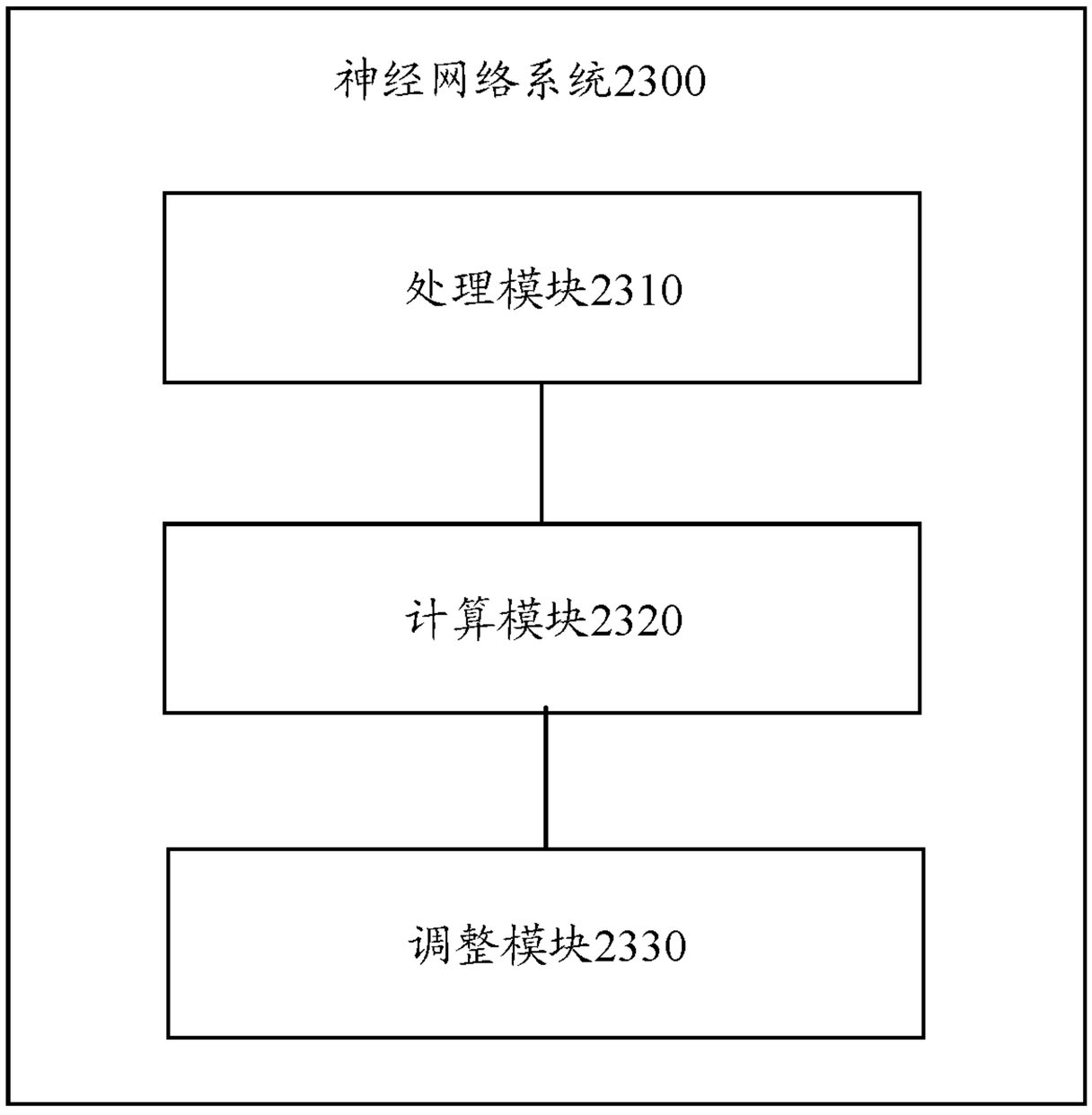


图 23

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/130393

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/04(2006.01)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)

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)

CNPAT, WPI, EPODOC, CNKI: 神经网络, 忆阻器, 存算, 输出, 误差, 偏差, 残差, 权重, neutral network, memristor, output, error, weight

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 109460817 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 12 March 2019 (2019-03-12) description, paragraphs [0003]-[0049]	1-17
X	CN 110443168 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 12 November 2019 (2019-11-12) claims 1-10	1-17
A	CN 109886393 A (HANGZHOU FLASH BILLION SEMICONDUCTOR CO., LTD.) 14 June 2019 (2019-06-14) entire document	1-17
A	CN 108009640 A (TSINGHUA UNIVERSITY) 08 May 2018 (2018-05-08) entire document	1-17

☐ 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
"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	Date of mailing of the international search report
08 February 2021	**22 February 2021**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/130393

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109460817 A	12 March 2019	WO 2020052342 A1	19 March 2020
		US 2020342301 A1	29 October 2020
CN 110443168 A	12 November 2019	None	
CN 109886393 A	14 June 2019	None	
CN 108009640 A	08 May 2018	None	

国际检索报告

国际申请号
PCT/CN2020/130393

A. 主题的分类

G06N 3/04(2006.01)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)

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））

CNPAT, WPI, EPODOC, CNKI：神经网络，忆阻器，存算，输出，误差，偏差，残差，权重，neutral network，memristor，output，error，weight

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
X	CN 109460817 A (华中科技大学) 2019年 3月 12日 (2019-03-12) 说明书第[0003]-[0049]段	1-17
X	CN 110443168 A (华中科技大学) 2019年 11月 12日 (2019-11-12) 权利要求1-10	1-17
A	CN 109886393 A (杭州闪亿半导体有限公司) 2019年 6月 14日 (2019-06-14) 全文	1-17
A	CN 108009640 A (清华大学) 2018年 5月 8日 (2018-05-08) 全文	1-17

☐ 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
“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	国际检索报告邮寄日期
2021年 2月 8日	2021年 2月 22日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 林婉娟 电话号码 86-(10)-53961343

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/130393

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	109460817	A	2019年 3月 12日	WO	2020052342	A1	2020年 3月 19日
				US	2020342301	A1	2020年 10月 29日
CN	110443168	A	2019年 11月 12日		无		
CN	109886393	A	2019年 6月 14日		无		
CN	108009640	A	2018年 5月 8日		无		