(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
2019 年 7 月 4 日 (04.07.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/129302 A1

(51) 国际专利分类号:
G06N 3/063 (2006.01)

(21) 国际申请号: PCT/CN2018/125801

(22) 国际申请日: 2018 年 12 月 29 日 (29.12.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:

201711499267.X	2017年12月30日 (30.12.2017)	CN
201711499268.4	2017年12月30日 (30.12.2017)	CN
201711499265.0	2017年12月30日 (30.12.2017)	CN
201711499266.5	2017年12月30日 (30.12.2017)	CN

(71) 申请人: 北京中科寒武纪科技有限公司 (CAMBRICON TECHNOLOGIES CORPORATION LIMITED) [CN/CN]; 中国北京市海淀区科学院南路6号科研综合楼644室, Beijing 100000 (CN)。

(72) 发明人: 陈天石 (CHEN, Tianshi); 中国北京市海淀区科学院南路6号科研综合楼644室, Beijing 100000 (CN)。 刘少礼 (LIU, Shaoli); 中国北京市海淀区科学院南路6号科研综合楼644室, Beijing 100000 (CN)。 王秉睿 (WANG, Bingrui); 中国北京市海淀区科学院南路6号科研综合楼644室, Beijing 100000 (CN)。 张尧 (ZHANG, Yao); 中国北京市海淀区科学院南路6号科研综合楼644室, Beijing 100000 (CN)。 胡帅 (HU, Shuai); 中国北京市海淀区科学院南路6号科研综合楼644室, Beijing 100000 (CN)。 宋新开 (SONG, Xinkai); 中国北京市海淀区科学院南路6号科研综合楼644室, Beijing 100000 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路80号汇华商贸大厦1508室, Guangdong 510070 (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,

(54) **Title:** INTEGRATED CIRCUIT CHIP DEVICE AND RELATED PRODUCT

(54) **发明名称**: 集成电路芯片装置及相关产品

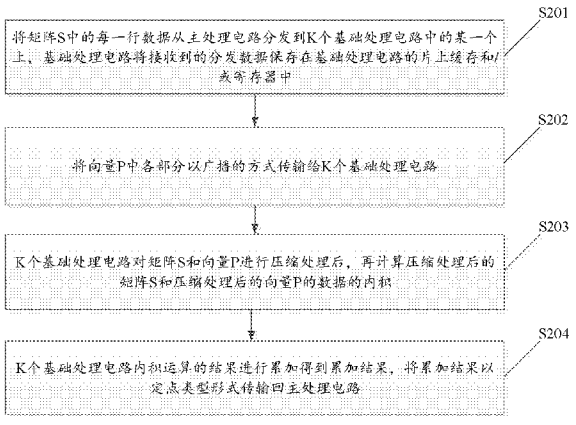

图 2

S201 Distribute each row of data in a matrix S from a main processing circuit to a certain basic processing circuit from among K basic processing circuits, the basic processing circuit storing the received distributed data in a cache and/or a register on a chip of the basic processing circuit

S202 Transmit each part of a vector P to the K basic processing circuits by means of broadcasting

S203 Once the K basic processing circuits have performed compression processing on the matrix S and the vector P, further calculate the inner product of data of the compressed matrix S and the compressed vector P

S204 Accumulate the results of the inner product operations of the K basic processing circuits to obtain an accumulation result, and transmit the accumulation result back to the main processing circuit in a fixed point-type form

(57) **Abstract:** An integrated circuit chip device and a related product, the integrated circuit chip device comprising: a main processing circuit and a plurality of basic processing circuits; the main processing circuit or at least one basic processing circuits from among the plurality of basic processing circuits comprises: a compression mapping circuit (101), the compression mapping circuit (101) being used to execute the compression processing of each piece of data in a neural network computation. The integrated circuit chip device and the related product have the advantages of calculation volume being small and power consumption being low.

(57) **摘要**: 一种集成电路芯片装置及相关产品，所述集成电路芯片装置包括：主处理电路以及多个基础处理电路；所述主处理电路或多个基础处理电路中至少一个基础处理电路包括：压缩映射电路（101），所述压缩映射电路（101）用于执行神经网络运算中的各个数据的压缩处理。所述集成电路芯片装置及相关产品具有计算量小，功耗低的优点。

WO 2019/129302 A1

[见续页]

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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))。

集成电路芯片装置及相关产品

相关申请：

本申请要求 2017 年 12 月 30 日提交，申请号为 201711499267.X，发明名称为“集成电路芯片装置及相关产品”的优先权；

本申请要求 2017 年 12 月 30 日提交，申请号为 201711499268.4，发明名称为“集成电路芯片装置及相关产品”的优先权；

本申请要求 2017 年 12 月 30 日提交，申请号为 201711499265.0，发明名称为“集成电路芯片装置及相关产品”的优先权；

本申请要求 2017 年 12 月 30 日提交，申请号为 201711499266.5，发明名称为“集成电路芯片装置及相关产品”的优先权。

技术领域

本披露涉及神经网络领域，尤其涉及一种集成电路芯片装置及相关产品。

背景技术

人工神经网络（Artificial Neural Network，ANN），是 20 世纪 80 年代以来人工智能领域兴起的研究热点。它从信息处理角度对人脑神经元网络进行抽象， 建立某种简单模型，按不同的连接方式组成不同的网络。在工程与学术界也常直接简称为神经网络或类神经网络。神经网络是一种运算模型，由大量的节点（或称神经元）之间相互联接构成。现有的神经网络的运算基于中央处理器（Central Processing Unit，CPU）或图形处理器（Graphics Processing Unit，GPU）来实现神经网络的运算，此种运算的计算量大，功耗高。

发明内容

本披露实施例提供了一种集成电路芯片装置及相关产品，可提升计算装置的处理速度，提高效率。

第一方面，提供一种集成电路芯片装置，所述集成电路芯片装置包括：主处理电路、k 个分支电路以及 k 组基础处理电路，所述主处理电路与所述 k 个分支电路分别连接，k 个分支电路中的每个分支电路和 k 组基础处理电路中的一组基础处理电路一一对应，所述一组基础处理电路包括至少一个基础处理电路；

所述分支电路包括：压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及和与其相连的所述 k 个分支电路传输数据；

所述 k 个分支电路，用于在主处理电路与 k 组基础电路之间转发所述传输数据，依据所述传输数据的运算控制是否启动所述压缩映射电路对所述传输数据进行压缩处理；

所述 k 个基础处理电路，用于依据所述传输数据或压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

第二方面，提供一种集成电路芯片装置，所述集成电路芯片装置包括：主处理电路以及多个基础处理电路；

所述基础处理电路包括：压缩映射电路；所述压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及向所述多个基础处理电路传输数据；

所述多个基础处理电路，用于依据所述传输数据的运算控制是否启动所述压缩映射电路对所述传输数据进行压缩处理；依据所述传输数据或压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

第三方面，提供一种集成电路芯片装置，所述集成电路芯片装置包括：主处理电路以及多个基础处理电路；

所述多个基础处理电路呈阵列分布；每个基础处理电路与相邻的其他基础处理电路连接，所述主处理电路连接所述多个基础处理电路中的 k 个基础处理电路，所述 k 个基础电路为：第 1 行的 n 个基础处理电路、第 m 行的 n 个基础处理电路以及第 1 列的 m 个基础处理电路；

所述多个基础处理电路包括：压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及和与所述 k 个基础处理电路传输数据；

所述 k 个基础处理电路，用于在所述主处理电路以及多个基础处理电路之间的数据转发；

所述多个基础处理电路，用于依据传输数据的运算控制确定是否启动所述压缩映射电路对所述传输数据进行压缩处理，依据压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

第四方面，提供一种集成电路芯片装置，所述集成电路芯片装置包括：主处理电路以及多个基础处理电路；

所述多个基础处理电路呈阵列分布；每个基础处理电路与相邻的其他基础处理电路连接，所述主处理电路连接所述多个基础处理电路中的 k 个基础处理电路，所述 k 个基础电路为：第 1 行的 n 个基础处理电路以及第 1 列的 m 个基础处理电路；

所述 k 个基础处理电路包括：压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及和与其相连的所述基础处理电路传输数据；

所述 k 个基础处理电路，用于依据传输数据的运算控制确定是否启动所述压缩映射电路对所述传输数据进行压缩处理，并将压缩处理后的传输数据发送给与所述 k 个基础处理电路连接的基础处理电路；

所述多个基础处理电路，用于依据压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

第五方面，提供一种神经网络运算装置，所述神经网络运算装置包括上述第一方面至第四方面中的任一方面所提供的集成电路芯片装置。

第六方面，提供一种组合处理装置，所述组合处理装置包括：第五方面提供的神经网络运算装置、通用互联接口和通用处理装置；

所述神经网络运算装置通过所述通用互联接口与所述通用处理装置连接。

第七方面，提供一种芯片，所述芯片集成上述第一方面至第六方面中的任一方面所提供的装置。

第八方面，提供一种电子设备，所述电子设备包括第七方面的芯片。

第九方面，提供一种神经网络的运算方法，所述方法应用在集成电路芯片装置内，所述集成电路芯片装置包括：第一方面至第四方面中的任一方面所述的集成电路芯片装置，所述集成电路芯片装置用于执行神经网络的运算。

可以看出，通过本披露实施例，提供压缩映射电路将数据块压缩处理后再进行运算，节省了传输资源以及计算资源，所以其具有功耗低，计算量小的优点。

附图说明

图 1a 是一种集成电路芯片装置结构示意图。

图 1b 是另一种集成电路芯片装置结构示意图。

图 1c 是一种基础处理电路的结构示意图。

图 1d 为本申请实施例提供的一种压缩映射电路的局部结构示意图。

图 1e 为本申请实施例提供的一种神经网络结构示意图。

图 1f 为本申请实施例提供的另一种压缩映射电路的局部结构示意图。

图 1g 为本申请实施例提供的另一种压缩映射电路的局部结构示意图。

图 1h 为本申请实施例提供的另一种压缩映射电路的局部结构示意图。

图 1i 为本申请实施例提供的另一种压缩映射电路的局部结构示意图。

图 1j 为本申请实施例提供的另一种压缩映射电路的局部结构示意图。

图 1k 为本申请实施例提供的另一种压缩映射电路的局部结构示意图。

图 2 为一种矩阵乘以向量流程示意图。

图 2a 是矩阵乘以向量的示意图。

图 2b 为一种矩阵乘以矩阵流程示意图。

图 2c 是矩阵 Ai 乘以向量 B 的示意图。

图 2d 是矩阵 A 乘以矩阵 B 的示意图。

图 2e 是矩阵 Ai 乘以矩阵 B 的示意图。

图 3a 为神经网络训练示意图。

图 3b 为卷积运算示意图。

图 4a 是另一种集成电路芯片装置结构示意图。

图 4b 是另一种集成电路芯片装置结构示意图。

图 4c 是一种基础处理电路的结构示意图。

图 5a 是一种基础处理电路的使用方法示意图。

图 5b 是一种主处理电路传输数据示意图。

图 5c 是矩阵乘以向量的示意图。

图 5d 是一种集成电路芯片装置结构示意图。

图 5e 是另一种集成电路芯片装置结构示意图。

图 5f 是矩阵乘以矩阵的示意图。

图 6a 为卷积输入数据示意图。

图 6b 为卷积核示意图。

图 6c 为输入数据的一个三维数据块的运算窗口示意图。

图 6d 为输入数据的一个三维数据块的另一运算窗口示意图。

图 6e 为输入数据的一个三维数据块的又一运算窗口示意图。

图 7 为本申请实施例提供的一种神经网络芯片的结构示意图。

具体实施方式

为了使本技术领域的人员更好地理解本披露方案，下面将结合本披露实施例中的附图，对本披露实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本披露一部分实施例，而不是全部的实施例。基于本披露中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本披露保护的范围。

本申请装置中，所述主处理电路，用于执行神经网络运算中的各个连续的运算以及向所述多个基础处理电路传输数据；所述 k 组基础处理电路，用于依据所述传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

在可选实施例中，装置还包括 k 个分支电路，所述主处理电路与所述 k 个分支电路分别连接，k 个分支电路中每个分支电路对应 k 组基础处理电路中的一组基础处理电路，用于在所述主处理电路与所述 k 组基础处理电路之间转发传输数据。

在可选实施例中，所述基础处理电路包括压缩映射电路；所述压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；所述 k 组基础处理电路，具体用于依据所述传输数据的运算控制是否启动所述压缩映射电路对所述传输数据进行压缩处理；依据所述传输数据或压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

在可选实施例中，所述主处理电路，用于获取待计算的数据块以及运算指令，依据该运算指令对所述待计算的数据块划分成分发数据块以及广播数据块；对所述分发数据块进行拆分处理得到多个基本数据块，将所述多个基本数据块分发至与其连接的电路，将所述广播数据块广播至与其连接的电路；所述基础处理电路，用于依据所述运算控制启动所述压缩映射电路对所述基本数据块与所述广播数据块进行压缩处理后再执行内积运算得到运算结果，将所述运算结果发送至主处理电路；所述主处理电路，用于对所述运算结果处理得到所述待计算的数据块以及运算指令的指令结果；其中，所述待计算的数据块为待计算的至少一个输入神经元，和/或，至少一个权值。

在可选实施例中，所述分支电路包括：压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；所述主处理电路，用于执行神经网络运算中的各个连续的运算以及和与其相连的所述 k 个分支电路传输数据；所述 k 个分支电路，用于在主处理电路与 k 组基

础电路之间转发所述传输数据，依据所述传输数据的运算控制是否启动所述压缩映射电路对所述传输数据进行压缩处理；所述 k 个基础处理电路，用于依据所述传输数据或压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

在可选实施例中，所述主处理电路，用于获取待计算的数据块以及运算指令，依据该运算指令对所述待计算的数据块划分成分发数据块以及广播数据块；对所述分发数据块进行拆分处理得到多个基本数据块，将所述多个基本数据块分发至与其连接的所述 k 个分支电路，将所述广播数据块广播至与其连接的所述 k 个分支电路；所述 k 个分支电路，用于接收基本数据块以及广播数据块，启动压缩映射电路将该基本数据块以及广播数据块进行压缩处理；将压缩处理后的基本数据块以及压缩处理后的广播数据块转发至 k 组基础处理电路；所述基础处理电路，用于对所述压缩处理后的基本数据块与所述压缩处理后的广播数据块执行内积运算得到运算结果，将所述运算结果发送至所述主处理电路；所述主处理电路，用于对所述运算结果处理得到所述待计算的数据块以及运算指令的指令结果；其中，所述分发数据块以及所述广播数据块为至少一个输入神经元或者，至少一个权值。

在可选实施例中，所述主处理电路，具体用于将所述广播数据块通过一次广播至所述 k 个分支电路。

在可选实施例中，所述主处理电路，具体用于将所述广播数据块分成多个部分广播数据块，将所述多个部分广播数据块通过多次广播至所述 K 个分支电路。

在可选实施例中，所述基础处理电路，具体用于将所述部分广播数据块与所述基本数据块执行一次内积处理后得到内积处理结果，将所述内积处理结果累加得到部分运算结果，将所述部分运算结果发送至所述主处理电路。

在可选实施例中，所述基础处理电路，具体用于复用 n 次该部分广播数据块执行该部分广播数据块与该 n 个基本数据块内积运算得到 n 个部分处理结果，将 n 个部分处理结果分别累加后得到 n 个部分运算结果，将所述 n 个部分运算结果发送至主处理电路，所述 n 为大于等于 2 的整数。

在可选实施例中，所述主处理电路包括：主寄存器或主片上缓存电路；

或所述分支电路包括：基本寄存器或基本片上缓存电路；

或所述基础处理电路包括：基本寄存器或基本片上缓存电路。

在可选实施例中，所述主处理电路包括：向量运算器电路、算数逻辑单元电路、累加器电路、矩阵转置电路、直接内存存取电路、压缩映射电路或数据重排电路中的一种或任意组合。

在可选实施例中，所述数据为：向量、矩阵、三维数据块、四维数据块以及 n 维数据块中一种或任意组合。

在可选实施例中，如所述运算指令为乘法指令，所述主处理电路确定乘数数据块为广播数据块，被乘数数据块为分发数据块；

如所述运算指令为卷积指令，所述主处理电路确定输入数据块为广播数据块，卷积核为分发数据块。

在可选实施例中，本申请涉及的神经网络的运算包括：卷积运算、矩阵乘矩阵运算、矩阵乘向量运算、偏执运算、全连接运算、GEMM 运算、GEMV 运算、激活运算中的一种

或任意组合。

参阅图 1a，图 1a 为一种集成电路芯片装置的结构示意图，如图 1a 所示，该芯片装置包括：主处理电路、基本处理电路和分支处理电路（可选的）。其中，集成电路芯片装置包括：主处理电路、k 个分支电路(如图 1a 所示，k=4，当然在实际应用中也可以为其他数值，例如 8、16 等等数值)以及 k 组基础处理电路，所述主处理电路与所述 k 个分支电路分别连接，k 个分支电路中每个分支电路对应 k 组基础处理电路中的一组基础处理电路，所述一组基础处理电路包括至少一个基础处理电路。在实际应用中，压缩映射电路可设置在基础处理电路或者分支电路中，如图以虚线框所示。该压缩映射电路用于对数据进行压缩处理，具体在本申请下文所述。

主处理电路（如图 1d 所示）可以包括寄存器和/或片上缓存电路，该主处理电路还可以包括:控制电路、向量运算器电路、算数逻辑单元（arithmetic and logic unit，ALU）电路、累加器电路、直接内存存取 (Direct Memory Access，DMA)电路等电路，当然在实际应用中，上述主处理电路还可以添加，转换电路（例如矩阵转置电路）、数据重排电路或激活电路等等其他的电路；

可选的，主处理电路可以包括：压缩映射电路，压缩映射电路可以用于对接收或发送的数据进行压缩处理，在实际应用中例如将为 0 或者小于预设阈值（如 0.1）的数据进行剔除。所述预设阈值为用户侧或终端设备侧自定义设置的，例如 0.1、0.05 等等。本申请并不限制上述压缩映射电路的具体形式。关于所述压缩处理将在下文进行具体阐述。

主处理电路还包括数据发送电路、数据接收电路或接口，该数据发送电路可以集成数据分发电路以及数据广播电路，当然在实际应用中，数据分发电路以及数据广播电路也可以分别设置；在实际应用中上述数据发送电路以及数据接收电路也可以集成在一起形成数据收发电路。对于广播数据，主处理电路即需要将该广播数据发送给每个基础处理电路的数据。对于分发数据，主处理电路即需要有选择的将分发数据发送给部分基础处理电路的数据，具体的选择方式可以由主处理电路依据负载以及计算方式进行具体的确定。对于广播发送方式，即将广播数据以广播形式发送至每个基础处理电路。(在实际应用中，通过一次广播的方式将广播数据发送至每个基础处理电路，也可以通过多次广播的方式将广播数据发送至每个基础处理电路，本申请具体实施方式并不限制上述广播的次数)，对于分发发送方式，即将分发数据有选择的发送给部分基础处理电路。

在实现分发数据时，主处理电路的控制电路向部分或者全部基础处理电路传输数据，该数据可以相同，也可以不同。具体的，如果采用分发的方式发送数据，各个接收数据的基础处理电路收到的数据可以不同，当然也可以有部分基础处理电路收到的数据相同；

具体地，广播数据时，主处理电路的控制电路向部分或者全部基础处理电路传输数据，各个接收数据的基础处理电路可以收到相同的数据。

可选的，上述主处理电路的向量运算器电路可以执行向量运算,包括但不限于：两个向量加减乘除，向量与常数加、减、乘、除运算，或者对向量中的每个元素执行任意运算。其中，连续的运算具体可以为，向量与常数加、减、乘、除运算、激活运算、累加运算等等。

每个基础处理电路可以包括基础寄存器和/或基础片上缓存电路；每个基础处理电路还

可以包括：内积运算器电路、向量运算器电路、累加器电路等中一个或任意组合。上述内积运算器电路、向量运算器电路、累加器电路都可以是集成电路，上述内积运算器电路、向量运算器电路、累加器电路也可以为单独设置的电路。

该芯片装置可选的还可以包括一个或多个分支处理电路，如该芯片装置具有分支处理电路时，其中主处理电路与分支处理电路连接，该分支处理电路与基本处理电路连接，该基本处理电路的内积运算器电路用于执行数据块之间的内积运算，该主处理电路的控制电路控制数据接收电路或数据发送电路收发外部数据，以及通过控制电路 控制 数据发送电路将外部数据分发至分支处理电路，该分支处理电路用于收发主处理电路或基本处理电路的数据。如图 1a 所示的结构适合复杂数据的计算，因为对于主处理电路来说，其连接的单元的数量有限，所以需要在主处理电路与基本处理电路之间添加分支处理电路以实现更多的基本处理电路的接入，从而实现对复杂数据块的计算。分支处理电路和基础处理电路的连接结构可以是任意的，不局限在图 1a 的 H 型结构。可选的，主处理电路到基础处理电路是广播或分发的结构，基础处理电路到主处理电路是收集（Gather）的结构。广播结构，分发结构和收集结构的定义如下：对于分发或广播结构，此时的基础处理电路的数量大于主处理电路，即 1 个主处理电路对应多个基础处理电路，即从主处理电路到多个基础处理电路为广播或分发的结构，反之，从多个基础处理电路到主处理电路可以为收集结构。

基础处理电路，接收主处理电路分发或者广播的数据保存到基础处理电路的片上缓存中，可以进行运算产生结果，可以向主处理电路发送数据。

基础处理电路中所涉及到的数据可以是经过压缩处理后的数据，其中压缩处理涉及的具体实施方式将在后续进行阐述。

可选的，每个基础处理电路均可以包括压缩映射电路，也可以在部分基础处理电路配置压缩映射电路；压缩映射电路可以用于对接收或发送的数据进行压缩处理。本申请并不限制上述压缩映射电路的具体形式。

可选的，该基础处理电路的向量运算器电路可以对压缩处理后的两个向量执行向量运算，当然在实际应用中，基础处理电路的内积运算器电路可以对压缩处理后的两个向量执行内积运算，累加器电路也可以对内积运算的结果进行累加。

在一种可选方案中，两个向量可以存放在片上缓存和/或寄存器中，基础处理电路可以根据实际计算的需要提取两个向量执行运算。该运算包括但不限于：内积运算、乘法运算、加法运算或其他的运算。

在一种可选方案中，内积运算的结果可以累加到片上缓存和/或寄存器上；其可选方案的优点是，减少了基础处理电路和主处理电路之间的数据传输量，提高了运算效率，降低了数据传输功耗。

在一种可选方案中，内积运算的结果不进行累加，直接作为结果传输；此技术方案的优点是，减少了基础处理电路内部的运算量，提高基础处理电路的运算效率。

在一种可选方案中，每个基础处理电路可以执行多组两个向量的内积运算，也可以对多组内积运算的结果分别进行累加；

在一种可选方案中，多组的两个向量数据可以存放在片上缓存和/或寄存器中；

在一种可选方案中，多组内积运算的结果可以分别累加到片上缓存和/或寄存器中；

在一种可选方案中，各组内积运算的结果可以不进行累加，直接作为结果传输；

在一种可选方案中，每个基础处理电路可以执行同一个向量与多个向量分别进行内积运算的操作（“一对多”内积，即多组内积里每组的两个向量中有一个向量是共享的），并将每个向量对应的内积结果分别进行累加。此技术方案可以实现同一套权值对不同的输入数据进行多次计算，增加了数据复用，减少基础处理电路内部数据的数据传输量，提高计算效率，降低功耗。

具体地，计算内积使用的数据中，各组共享的向量和每组的另一个向量（即每组之间不同的那个向量）的数据来源可以不同：

在一种可选方案中，在计算内积时，各组共享的向量来自主处理电路或者分支处理电路的广播或者分发；

在一种可选方案中，在计算内积时，各组共享的向量来自片上缓存；

在一种可选方案中，在计算内积时，各组共享的向量来自寄存器；

在一种可选方案中，在计算内积时，每组的另一个非共享向量来自主处理电路或者分支处理电路的广播或者分发；

在一种可选方案中，在计算内积时，每组的另一个非共享向量来自从片上缓存；

在一种可选方案中，在计算内积时，每组的另一个非共享向量来自寄存器；

在一种可选方案中，在进行多组的内积运算时，每组共享的向量在基础处理电路的片上缓存和/寄存器中保留任意份数；

在一种可选方案中，共享向量可以对应每组内积各保留一份；

在一种可选方案中，共享向量可以只保留一份；

具体地，多组内积运算的结果可以分别累加到片上缓存和/或寄存器中；

具体地，各组内积运算的结果可以不进行累加，直接作为结果传输；

参阅图 1a 所示的结构，其包含一主处理电路（可以执行向量操作），多基础处理电路（可以执行内积操作）。这样组合的好处是：装置不仅能使用基础处理电路执行矩阵和向量乘法运算，也能使用主处理电路执行其他任意的向量运算，使装置在有限的硬件电路的配置下，能够更快的完成更多的运算，减少了与装置外部进行数据传输的次数，提高了计算效率，降低了功耗。另外，本芯片在基础处理电路和/或主处理电路均可以设置压缩映射电路，这样在进行神经网络计算时能够减少计算的数据量，并且本芯片可以依据各个电路（主要是主处理电路和基础处理电路）的运算量（负载量）动态地分配由哪个电路来进行数据的压缩处理，这样能够减少数据计算的复杂程序，降低功耗，并且动态的分配数据的压缩处理能够实现不影响芯片的计算效率。该分配的方式包括但不限于：负载均衡、负载最小值分配等等方式。

参阅图 1b 所示的装置，图 1b 所示的装置包括主处理电路和基础处理电路，可选地还可包括分支处理电路。如图 1b 所示的装置包括：主处理电路以及 N 个基础处理电路，其中，主处理电路（具体的结构如图 1c 所示）与 N 个基础处理电路可以直接或间接连接，如为间接连接的方式时 ，一种可选的方案如图 1a 所示可以包括 N/4 个分支处理电路，每个分支处理电路分别连接 4 个基础处理电路，对于主处理电路以及 N 个基础处理电路分别包含的电路可以参见上述如图 1a 所示的描述，这里不再赘述，这里需要说明的是，上述基础

处理电路还可以设置在分支处理电路内，另外，每个分支处理电路连接基础处理电路的数量也可以不局限于 4 个，厂家可以根据实际需要进行配置。该上述主处理电路和/或 N 个基础处理电路均可以包括压缩映射电路，具体的，可以是主处理电路包括压缩映射电路，也可以是 N 个基础处理电路或其中的一部分包括压缩映射电路，也可以是主处理电路和 N 个基础处理电路或其中的一部分均包括。上述主处理电路可以根据神经网络计算指令动态的分配数据压缩处理步骤的操作实体，具体的，主处理电路可以根据自身的负载确定是否对接收到的数据执行数据压缩处理，具体的，可以将负载的值设置多个区间，每个区间对应分配数据压缩处理步骤的执行主体，例如，以 3 个区间为例，区间 1 的负载值较低，可以由主处理电路单独执行数据压缩处理步骤，区间 2 负载值位于区间 1 以及区间 3 之间，可以由主处理电路或 N 个基础处理电路共同执行数据压缩处理步骤，区间 3 负载值较高，可以由 N 个基础处理电路执行数据压缩处理步骤。对此，可以以明示的方式来执行，例如主处理电路可以配置一个特殊指示或指令，当基础处理电路接收到该特殊指示或指令时，确定执行数据压缩处理步骤，如基础处理电路未接收到特殊指示或指令时，确定不执行数据压缩处理步骤。又如，可以以暗示的方式来执行，例如，基础处理电路接收到稀疏数据（即含 0，或包括小于预设阈值的数据大于预设数量）且确定需要执行内积运算时，将该稀疏数据进行压缩处理。

下面阐述本申请涉及的数据压缩处理的相关实施例。需要说明的是，本申请中的数据可以是神经网络中的输入神经元或权值，其具体可为矩阵数据或向量数据等，本申请不做限定。也即是本申请下文阐述的数据或数据块可为神经网络中的输入神经元或权值，它们可以矩阵或向量等形式体现。

由于神经网络是一个高计算量和高访存的算法，权值越多，计算量和访存量都会增大。特别是，针对权值较小（如为 0，或小于设定数值的权值）的情况下，为提高计算速率、减小开销需对这些权值较小的数据进行压缩处理。在实际应用中，数据压缩处理在稀疏神经网络中应用，效果最为明显，如减小数据计算的工作量、减小数据额外开销，提高数据计算速率等。

以输入数据为例，具体阐述所述压缩映射电路的数据压缩处理实施例。所述输入数据包括但不限于至少一个输入神经元和/或至少一个权值。

第一种实施例中：压缩映射电路对输入神经元和权值均进行压缩处理

压缩映射电路 101 接收到输入数据（具体可为主压缩处理电路发送的所述待计算的数据块）之后，可对所述输入数据进行压缩处理，以得到压缩处理后的输入数据，所述输入数据包括至少一个输入神经元和至少一个权值，所述压缩处理后的输入数据包括压缩处理后的输入神经元和压缩处理后的权值。

上述输入数据包括至少一个输入神经元和至少一个权值。上述压缩映射电路 101 确定所述至少一个输入神经元中每个输入神经元的绝对值是否大于第一阈值。当上述输入神经元的绝对值小于或者等于该第一阈值时，上述压缩映射电路 101 将该输入神经元删除；当上述输入神经元的绝对值大于上述第一阈值时，上述压缩映射电路 101 保留该输入神经元，该压缩映射电路 101 将删除后的输出神经元输出，作为压缩处理后的输入神经元。上述压缩映射电路 101 获取输入神经元的连接关系数据，该输入神经元的连接关系数据表示上述

至少一个输入神经元中绝对值大于上述第一阈值 的输入神经元的位置信息。上述压缩映射电路 101 确定上述至少一个权值中每个权值的绝对值是否大于第二阈值。当权值的绝对值小于或者等于上述第二阈值时，上述压缩映射电路 101 将该权值删除，并根据上述输入神经元的连接关系数据将从上述删除后的权值中选择相关的权值输出，作为压缩处理后的权值。

在一种可行的实施例中，上述输入数据包括至少一个输入神经元和至少一个权值。上述压缩映射电路 101 确定所述至少一个权值中每个权值的绝对值是否大于第二阈值。当上述权值的绝对值小于或者等于该第二阈值时，上述压缩映射电路 101 将该权值删除；当上述权值的绝对值大于上述第二阈值时，上述压缩映射电路 101 保留该权值，该压缩映射电路 101 将删除后的权值输出，作为压缩处理后的权值。上述压缩映射电路 101 获取权值的连接关系数据，该权值的连接关系数据表示上述至少一个输入神经元与输出神经元之间的连接关系的数据。上述压缩映射电路 101 确定上述至少一个输入神经元中每个输入神经元的绝对值是否大于第一阈值。当输入神经元的绝对值小于或者等于上述第一阈值时，上述压缩映射电路 101 将该输入神经元删除，并根据上述权值的连接关系数据将从上述删除后的输入神经元中选择相关的输入神经元输出，作为压缩处理后的输入神经元。

进一步地，上述压缩映射电路 101 将上述压缩处理后的输入神经元和压缩处理后的权值按照一一对应的格式存储到存储电路中。

具体地，上述压缩映射电路 101 对上述压缩处理后的输入神经元和上述压缩处理后的权值按照一一对应的格式进行存储的具体方式是将上述压缩处理后的输入神经元中的每个压缩处理后的输入神经元和与其对应的压缩处理后的权值作为一个数据集，并将该数据集存储到存储电路中。

具体地，如图 1d 所示，上述压缩映射电路 101 包括：

第一稀疏处理单元 1011，用于对第二输入数据进行压缩处理，以得到第三输出数据和第二输出数据，并将所述第三输出数据传输至第一数据处理单元 1012。

第一数据处理单元 1012，用于接收第一输入数据和接收所述第三输出数据，并根据上述第三输出数据和第一输入数据输出第一输出数据。

其中，当所述第一输入数据包括至少一个输入神经元，所述第二输入数据包括至少一个权值时，所述第一输出数据为压缩处理后的输入神经元，所述第二输出数据为压缩处理后的权值，所述第三输出数据为权值的连接关系数据；当所述第一输入数据包括至少一个权值，所述第二输入数据包括至少一个输入神经元时，所述第一输出数据为压缩处理后的权值，所述第二输出数据为压缩处理后的输入神经元，所述第三输出数据为输入神经元的连接关系数据。

具体地，当上述第二输入数据为权值时，且权值的形式为 wij，该 wij 表示第 i 个输入神经元与第 j 个输出神经元之间的权值；上述第一稀疏处理单元 1011 根据权值确定上述连接关系数据（即上述第三输出数据），并将上述权值中绝对值小于或者等于第二阈值的权值删除，得到压缩处理后的权值（即上述第二输出数据）；当上述第二输入数据为输入神经元时，上述第一稀疏处理单元 1011 根据输入神经元得到连接关系数据，并将该输入神经元中的绝对值小于或等于上述第一阈值的输入神经元删除，以得到压缩处理后的输入神经元。

可选地，上述第一阈值可为 0.1、0.08、0.05、0.02、0.01、0 或者其他值。上述第二阈值可为 0.1、0.08、0.06、0.05、0.02、0.01、0 或者其他值。上述第一阈值和上述第二阈值可以一致，也可以不一致。

其中，上述连接关系数据可以步长索引或者直接索引的形式表示。

具体地，以直接索引形式表示的连接关系数据为由 0 和 1 组成的字符串，当上述第二输入数据为权值时，0 表示该权值的绝对值小于或者等于上述第二阈值，即该权值对应的输入神经元与输出神经元之间没有连接，1 表示该权值的绝对值大于上述第二阈值，即该权值对应的输入神经元与输出神经元之间有连接。以直接索引形式表示的连接关系数据有两种表示顺序：以每个输出神经元与所有输入神经元的连接状态组成一个 0 和 1 的字符串来表示权值的连接关系；或者每个输入神经元与所有输出神经元的连接状态组成一个 0 和 1 的字符串来表示权值的连接关系。当上述第二输入数据为输入神经元时，0 表示该输入神经元的绝对值小于或者等于上述第一阈值，1 表示该输入神经元的绝对值大于上述第一阈值。

应理解的，所述连接关系数据也可用向量/矩阵等形式体现，其中，0 表示该位置对应的输入神经元/权值的数据为 0 或者小于第一阈值；相应地，1 表示该位置对应的输入神经元/权值的数据不为 0 或者大于第一阈值等，本申请不做限定。可选的，所述数据的连接关系数据也可称为标记 mask 矩阵/mask 向量。

当上述第二输入数据为权值时，以步长索引形式表示的连接关系数据为与输出神经元有连接的输入神经元与上一个与该输出神经元有连接的输入神经元之间的距离值组成的字符串；当上述第二输入数据为输入神经元时，以步长索引表示的数据以当前绝对值大于上述第一阈值的输入神经元与上一个绝对值大于上述第一阈值的输入神经元之间的距离值组成的字符串表示。

举例说明，假设上述第一阈值和上述第二阈值均为 0.01，参见图 1e，图 1e 为本申请实施例提供的一种神经网络的示意图。如图 1e 中的 a 图所示，上述第一输入数据为输入神经元，包括输入神经元 i1、i2、i3 和 i4，上述第二输入数据为权值。对于输出神经元 o1，权值为 w11，w21，w31 和 w41；对于输出神经元 o2，权值 w12，w22，w32 和 w42，其中权值 w21，w12 和 w42 的值为 0，其绝对值均小于上述第一阈值 0.01，上述第一稀疏处理单元 1011 确定上述输入神经元 i2 和输出神经元 o1 没有连接，上述输入神经元 i1 和 i4 与输出神经元 o2 没有连接，上述输入神经元 i1、i3 和 i4 与上述输出神经元 o1 有连接，上述输入神经元 i2 和 i3 与输出神经元 o2 有连接。以每个输出神经元与所有输入神经元的连接状态表示上述连接关系数据，则上述输出神经元 o1 的连接关系数据为“1011”，输出神经元 o2 的连接关系数据为“0110”（即上述连接关系数据为“10110110”）；以每个输入神经元与所有输出神经元的连接关系，则输入神经元 i1 的连接关系数据为“10”，输入神经元 i2 的连接关系数据为“01”，输入神经元 i3 的连接关系数据为“11”，输入神经元 i4 的连接关系数据为“10”（即上述连接关系数据为“10011110”）。

对于上述输出神经元 o1，上述压缩映射电路 101 将上述 i1 与 w11，i3 与 w31 和 i4 与 w41 分别作为一个数据集，并将该数据集存储到存储电路中；对于输出神经元 o2，上述压缩映射电路 101 将上述 i2 与 w22 和 i3 与 w32 分别作为一个数据集，并将该数据集存储到

存储电路中。

针对上述输出神经元 o1，上述第二输出数据为 w11，w31 和 w41；针对上述输出神经元 o2，上述第二输出数据为 w22 和 w32。

当上述第二输入数据为输入神经元 i1、i2、i3 和 i4，且该输入神经元的值分别为 1，0，3，5 则上述连接关系数据（即上述第三输出数据）为“1011”，上述第二输出数据为 1，3，5。

如图 1e 中的 b 图所示，上述第一输入数据包括输入神经元 i1、i2、i3 和 i4，上述第二输入数据为权值。对于输出神经元 o1，权值为 w11，w21，w31 和 w41；对于输出神经元 o2，权值 w12，w22，w32 和 w42，其中权值 w21，w12 和 w42 的值为 0，上述稀疏处理单元 1011 确定上述输入神经元 i1、i3 和 i4 与上述输出神经元 o1 有连接，上述输入神经元 i2 和 i3 与输出神经元 o2 有连接。上述输出神经元 o1 与输入神经元之间的连接关系数据为“021”。其中，该连接关系数据中第一个数字“0”表示第一个与输出神经元 o1 有连接的输入神经元与第一个输入神经元之间的距离为 0，即第一个与输出神经元 o1 有连接的输入神经元为输入神经元 i1；上述连接关系数据中第二个数字“2”表示第二个与输出神经元 o1 有连接的输入神经元与第一个与输出神经元 o1 有连接的输入神经元（即输入神经元 i1）之间的距离为 2，即第二个与输出神经元 o1 有连接的输入神经元为输入神经元 i3；上述连接关系数据中第三个数字“1”表示第三个与输出神经元 o1 有连接的输入神经元与第二个与该输出神经元 o1 有连接的输入神经元之间的距离为 1，即第三个与输出神经元 o1 有连接的输入神经元为输入神经元 i4。

上述输出神经元 o2 与输入神经元之间的连接关系数据为“11”。其中，该连接关系数据中的第一数字“1”表示第一个与输出神经元 o2 有连接的输入神经元与第一个输入神经元（即输入神经元 i1）之间的距离为，即该第一个与输出神经元 o2 有连接关系的输入神经元为输出神经元 i2；上述连接关系数据中的第二数字“1”表示第二个与输出神经元 o2 有连接的输入神经元与第一个与输出神经元 o2 有连接的输入神经元的距离为 1，即第二个与输出神经元 o2 有连接的输入神经元为输入神经元 i3。

对于上述输出神经元 o1，上述压缩映射电路 101 将上述 i1 与 w11，i3 与 w31 和 i4 与 w41 分别作为一个数据集，并将该数据集存储到存储电路中；对于输出神经元 o2，上述压缩映射电路 101 将上述 i2 与 w22 和 i3 与 w32 分别作为一个数据集，并将该数据集存储到存储电路中。

针对上述输出神经元 o1，上述第二输出数据为 w11，w31 和 w41；针对上述输出神经元 o2，上述第二输出数据为 w22 和 w32。

当上述第二输入数据为输入神经元 i1、i2、i3 和 i4，且该输入神经元的值分别为 1，0，3，5 则上述连接关系数据即上述第三输出数据为“021”，上述第二输出数据为 1，3，5。

当上述第一输入数据为输入神经元时，则上述第二输入数据为权值，上述第三输出数据为输出神经元与上述输入神经元之间的连接关系数据。上述第一数据处理单元 1012 接收到上述输入神经元后，将该输入神经元中绝对值小于或等于上述第二阈值的输入神经元剔除，并根据上述连接关系数据，从剔除后的输入神经元中选择与上述权值相关的输入神经元，作为第一输出数据输出。

举例说明，假设上述第一阈值为 0，上述输入神经元 i1、i2、i3 和 i4，其值分别为 1，0，3 和 5，对于输出神经元 o1，上述第三输出数据（即连接关系数据）为“021”，上述第二输出数据为 w11，w31 和 w41。上述第一数据处理单元 1012 将上述输入神经元 i1、i2、i3 和 i4 中值为 0 的输入神经元剔除，得到输入神经元 i1、i3 和 i4。该第一数据处理单元 1012 根据上述第三输出数据“021”确定上述输入神经元 i1、i3 和 i4 均与上述输出神经元均有连接，故上述数据处理单元 1012 将上述输入神经元 i1、i3 和 i4 作为第一输出数据输出，即输出 1，3，5。

当上述第一输入数据为权值，上述第二输入数据为输入神经元时，上述第三输出数据为上述输入神经元的连接关系数据。上述第一数据处理单元 1012 接收到上述权值 w11，w21，w31 和 w41 后，将该权值中绝对值小于上述第一阈值的权值剔除，并根据上述连接关系数据，从上述剔除后的权值中选择与该上述输入神经元相关的权值，作为第一输出数据并输出。

举例说明，假设上述第二阈值为 0，上述权值 w11，w21，w31 和 w41，其值分别为 1，0，3 和 4，对于输出神经元 o1，上述第三输出数据（即连接关系数据）为“1011”，上述第二输出数据为 i1，i3 和 i5。上述第一数据处理单元 1012 将上述权值 w11，w21，w31 和 w41 中值为 0 的输入神经元剔除，得到权值 w11，w21，w31 和 w41。该第一数据处理单元 1012 根据上述第三输出数据“1011”确定上述输入神经元 i1、i2，i3 和 i4 中的输入神经元 i2 的值为 0，故上述第一数据处理单元 1012 将上述输入神经元 1，3 和 4 作为第一输出数据输出。

在一种可行的实施例中，第三输入数据和第四输入数据分别为至少一个权值和至少一个输入神经元，上述压缩映射电路 101 确定上述至少一个输入神经元中绝对值大于上述第一阈值的输入神经元的位置，并获取输入神经元的连接关系数据；上述压缩映射电路 101 确定上述至少一个权值中绝对值大于上述第二阈值的权值的位置，并获取权值的连接关系数据。上述压缩映射电路 101 根据上述权值的连接关系数据和输入神经元的连接关系数据得到一个新的连接关系数据，该连接关系数据表示上述至少一个输入神经元中绝对值大于上述第一阈值的输入神经元与输出神经元之间的关系和对应的权值的值。压缩映射电路 101 根据该新的连接关系数据、上述至少一个输入神经元和上述至少一个权值获取压缩处理后的输入神经元和压缩处理后的权值。

进一步地，上述压缩映射电路 101 将上述压缩处理后的输入神经元和压缩处理后的权值按照一一对应的格式存储到存储电路中。

具体地，上述压缩映射电路 101 对上述压缩处理后的输入神经元和上述压缩处理后的权值按照一一对应的格式进行存储的具体方式是将上述压缩处理后的输入神经元中的每个压缩处理后的输入神经元和与其对应的压缩处理后的权值作为一个数据集，并将该数据集存储到存储电路中。

对于压缩映射电路 101 包括第一稀疏处理单元 1011 和第一数据处理单元 1012 的情况，压缩映射电路 101 中的稀疏处理单元 1011 对输入神经元或者权值进行稀疏化压缩处理，减小了权值或者输入神经元的数量，进而减小了运算单元进行运算的次数，提高了运算效率。

具体地，如图 1f 所示，上述压缩映射电路 101 包括：

第二稀疏处理单元 1013，用于接收到第三输入数据后，根据所述第三输入数据得到第

一连接关系数据，并将该第一连接关系数据传输至连接关系处理单元 1015；

第三稀疏处理单元 1014，用于接收到第四输入数据后，根据所述第四输入数据得到第二连接关系数据，并将该第二连接关系数据传输至所述连接关系处理单元 1015；

所述连接关系处理单元 1015，用于根据所述第一连接关系数据和所述第二连接关系数据，以得到第三连接关系数据，并将该第三连接关系数据传输至第二数据处理单元 1016；

所述第二数据处理单元 1016，用于在接收到所述第三输入数据，所述第四输入数据和所述第三连接关系数据后，根据所述第三连接关系数据对所述第三输入数据和所述第四输入数据进行压缩处理，以得到第四输出数据和第五输出数据；

其中，当所述第三输入数据包括至少一个输入神经元，第四输入数据包括至少一个权值时，所述第一连接关系数据为输入神经元的连接关系数据，所述第二连接关系数据为权值的连接关系数据，所述第四输出数据为压缩处理后的输入神经元，所述第五输出数据为压缩处理后的权值；当所述第三输入数据包括至少一个权值，所述第四输入数据包括至少一个输入神经元时，所述第一连接关系数据为权值的连接关系数据，所述第二连接关系数据为输入神经元的连接关系数据，所述第四输出数据为压缩处理后的权值，所述第五输出数据为压缩处理后的输入神经元。

当上述第三输入数据包括至少一个输入神经元时，上述第一连接关系数据为用于表示该至少一个输入神经元中绝对值大于上述第一阈值的输入神经元的位置的字符串；当上述第三输入数据包括至少一个权值时，上述第一连接关系数据为用于表示输入神经元与输出神经元之间是否有连接的字符串。

当上述第四输入数据包括至少一个输入神经元时，上述第二连接关系数据为用于表示该至少一个输入神经元中绝对值大于上述第一阈值的输入神经元的位置的字符串；当上述第四输入数据包括至少一个权值时，上述第二连接关系数据为用于表示输入神经元与输出神经元之间是否有连接的字符串。

需要说明的是，上述第一连接关系数据、第二连接关系数据和第三连接关系数据均可以步长索引或者直接索引的形式表示，具体可参见上述相关描述。

换句话说，上述连接关系处理单元 1015 对上述第一连接关系数据和上述第二连接关系数据进行压缩处理，以得到第三连接关系数据。该第三连接关系数据可以直接索引或者步长索引的形式表示。

具体地，当上述第一连接关系数据和上述第二连接关系数据均以直接索引的形式表示时，上述连接关系处理单元 1015 对上述第一连接关系数据和上述第二连接关系数据进行与操作，以得到第三连接关系数据，该第三连接关系数据是以直接索引的形式表示的。

需要说明的是，表示上述第一连接关系数据和第二连接关系数据的字符串在内存中是按照物理地址高低的顺序存储的，可以是由高到低的顺序存储的，也可以是由低到高的顺序存储的。

当上述第一连接关系数据和上述第二连接关系数据均以步长索引的形式表示，且表示上述第一连接关系数据和第二连接关系数据的字符串是按照物理地址由低到高的顺序存储时，上述连接关系处理单元 1015 将上述第一连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第四连接关系

数据；同理，上述连接关系处理单元 1015 对上述第二连接关系数据的字符串进行同样的压缩处理，得到第五连接关系数据。然后上述连接关系处理单元 1015 从上述第四连接关系数据的字符串和上述第五连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成一个新的字符串。上述连接关系处理单元 1015 将上述新的字符串中将每一个元素与其相邻且值小于该元素值的元素进行相减，以得到一个新的元素。按照该方法，对上述新的字串中的每个元素进行相应的操作，以得到上述第三连接关系数据。

举例说明，假设以步长索引的形式表示上述第一连接关系数据和上述第二连接关系数据，上述第一连接关系数据的字符串为“01111”，上述第二连接关系数据的字符串为“022”，上述连接关系处理单元 1015 将上述第一连接关系数据的字符串中的每个元素与其相邻的前一个元素相加，得到第四连接关系数据“01234”；同理，上述连接关系处理单元 1015 对上述第二连接关系数据的字符串进行相同的压缩处理后得到的第五连接关系数据为“024”。上述连接关系处理单元 1015 从上述第四连接关系数据“01234”和上述第五连接关系数据“024”选组相同的元素，以得到新的字符串“024”。上述连接关系处理单元 1015 将该新的字符串中的每个元素与其相邻的前一个元素进行相减，即 0，(2-0)，(4-2)，以得到上述第三连接数据“022”。

当上述第一连接关系数据和上述第二连接关系数据中的任意一个以步长索引的形式表示，另一个以直接索引的形式表示时，上述连接关系处理单元 1015 将上述以步长索引表示的连接关系数据转换成以直接索引的表示形式或者将以直接索引表示的连接关系数据转换成以步长索引表示的形式。然后上述连接关系处理单元 1015 按照上述方法进行压缩处理，以得到上述第三连接关系数据（即上述第五输出数据）。

可选地，当上述第一连接关系数据和上述第二连接关系数据均以直接索引的形式表示时，上述连接关系处理单元 1015 将上述第一连接关系数据和上述第二连接关系数据均转换成以步长索引的形式表示的连接关系数据，然后按照上述方法对上述第一连接关系数据和上述第二连接关系数据进行压缩处理，以得到上述第三连接关系数据。

具体地，上述第三输入数据可为输入神经元或者权值、第四输入数据可为输入神经元或者权值，且上述第三输入数据和第四输入数据不一致。上述第二数据处理单元 1016 根据上述第三连接关系数据从上述第三输入数据（即输入神经元或者权值）中选取与该第三连接关系数据相关的数据，作为第四输出数据；上述第二数据处理单元 1016 根据上述第三连接关系数据从上述第四输入数据中选取与该第三连接关系数据相关的数据，作为第五输出数据。

进一步地，上述第二数据处理单元 1016 将上述压缩处理后的输入神经元与其对应的压缩处理后的权值作为一个数据集，将该数据集存储出存储电路中。

举例说明，假设上述第三输入数据包括输入神经元 i1，i2，i3 和 i4，上述第四输入数据包括权值 w11，w21，w31 和 w41，上述第三连接关系数据以直接索引方式表示，为“1010”，则上述第二数据处理单元 1016 输出的第四输出数据为输入神经元 i1 和 i3，输出的第五输出数据为权值 w11 和 w31。上述第二数据处理单元 1016 将输入神经元 i1 与权值 w11 和输入神经元 i3 与权值 w31 分别作为一个数据集，并将这两个数据集存储到存储电路中。

对于压缩映射电路 101 包括第二稀疏处理单元 1013，第三稀疏处理单元 1014、连接关

系处理单元 1015 和第二数据处理单元 1016 的情况，压缩映射电路 101 中的稀疏处理单元对输入神经元和权值均进行稀疏化压缩处理，使得输入神经元和权值的数量进一步减小，进而减小了运算单元的运算量，提高了运算效率。

可选地，所述压缩映射电路 101 对所述输入数据进行压缩处理之前，所述压缩映射电路 101 还用于：

对所述至少一个输入神经元进行分组，以得到 M 组输入神经元，所述 M 为大于或者等于 1 的整数；

判断所述 M 组输入神经元的每一组输入神经元是否满足第一预设条件，所述第一预设条件包括一组输入神经元中绝对值小于或者等于第三阈值的输入神经元的个数小于或者等于第四阈值；

当所述 M 组输入神经元任意一组输入神经元不满足所述第一预设条件时，将该组输入神经元删除；

对所述至少一个权值进行分组，以得到 N 组权值，所述 N 为大于或者等于 1 的整数；

判断所述 N 组权值的每一组权值是否满足第二预设条件，所述第二预设条件包括一组权值中绝对值小于或者等于第五阈值的权值的个数小于或者等于第六阈值；

当所述 N 组权值任意一组权值不满足所述第二预设条件时，将该组权值删除。

可选地，上述第三阈值可为 0.5，0.2，0.1，0.05，0.025，0.0，0 或者其他值。上述第四阈值与上述一组输入神经元中输入神经元的个数相关。可选地，该第四阈值=一组输入神经元中的输入神经元个数-1 或者该第四阈值为其他值。可选地，上述第五阈值可为 0.5，0.2，0.1，0.05，0.025，0.01，0 或者其他值。其中，上述第六阈值与上述一组权值中的权值个数相关。可选地，该第六阈值=一组权值中的权值个数-1 或者该第六阈值为其他值。

需要说明的是，上述第三阈值和上述第五阈值可相同或者不同，上述第四阈值和上述第六阈值可相同或者不同。可选的，存储电路可用于存储上述压缩处理后的输入神经元、压缩处理后的权值和相关的运算指令。

在可选实施例中，如图 1g 所示的压缩映射电路在已知输入数据的连接关系数据的情况下，可利用该输入数据的连接关系数据对所述输入数据进行压缩处理。所述输入数据包括至少一个输入神经元或者至少一个权值。具体如图 1g 所示，上述压缩映射电路 601 包括：

输入数据缓存单元 6011，用于缓存第一输入数据，该第一输入数据包括至少一个输入神经元或者至少一个权值。

连接关系缓存单元 6012，用于缓存第一输入数据的连接关系数据，即上述输入神经元的连接关系数据或者上述权值的连接关系数据。

其中，上述输入神经元的连接关系数据为用于表示该输入神经元中绝对值是否小于或者等于第一阈值的字符串，上述权值的连接关系数据为表示该权值绝对值是否小于或者等于上述第一阈值的字符串，或者为表示该权值对应的输入神经元和输出神经元之间是否有连接的字符串。该输入神经元的连接关系数据和权值的连接关系数据可以直接索引或者步长索引的形式表示。

需要说明的是，上述直接索引和步长索引的描述可参见上述图 1b 所示实施例的相关描述。

第四稀疏处理单元 6013，用于根据所述第一输入数据的连接关系数据对所述第一输入数据进行压缩处理，以得到压缩处理后的第一输入数据，并将该压缩处理后的第一输入数据存储到上述第一输入缓存单元中 605。

其中，当上述第一输入数据为至少一个输入神经元时，上述第四稀疏处理单元 6013 在一个时钟周期压缩处理一个输入神经元和一个连接关系，即在一个时钟周期从 S1 个输入神经元中选择一个有效的输入神经元，S1 为大于 1 的整数。

在一种可行的实施例中，上述第四稀疏处理单元 6013 在一个时钟周期压缩处理多个输入神经元和多个连接关系数据，即一个时钟周期从 S1 个输入神经元中选出有效的 S2 个输入数据，上述 S2 为大于 0 且小于或者等于该 S1 的整数。

举例说明，如图 1h 所示，上述输入神经元为 i1，i2，i3 和 i4，以直接索引的形式表示的连接关系数据为“1011”，并且上述第四稀疏处理单元 6013 在一个时钟周期可从 4 个输入神经元选择 1 个有连接（即有效）的输入神经元。上述第四稀疏处理单元 6013 从上述输入数据缓存单元 6011 和上述连接关系缓存单元 6012 中分别获取上述输入神经元 i1，i2，i3 和 i4 和上述连接关系数据“1011”后，上述第四稀疏处理单元 6013 根据该连接关系数据“1011”从上述输入神经元 i1，i2，i3 和 i4 选取有连接的输入神经元 i1，i3 和 i4。由于上述第四稀疏处理单元 6013 在一个时钟周期可从 4 个输入神经元选择 1 个有连接（即有效）的输入神经元，该第四稀疏处理单元 6013 在三个时钟周期内依次输出输入神经元 i1，i3 和 i4，如图 1h 所示。上述第四稀疏处理单元 6013 将上述输入神经元 i1，i3 和 i4 存储到第一输入缓存单元中。

再举例说明，如图 1i 所示，输入神经元为 i1，i2，i3 和 i4，以直接索引的形式表示的连接关系数据有两组，分别为“1011”和“0101”，上述第四稀疏处理单元 6013 在一个时钟周期可从 4 个输入神经元中选择 2 个有连接（即有效）的输入神经元。上述第四稀疏处理单元 6013 根据上述连接关系数据“1011”从上述输入神经元 i1，i2，i3 和 i4 中选择有连接的输入神经元 i1，i3 和 i4；根据上述连接关系数据“0101”从上述输入神经元 i1，i2，i3 和 i4 中选择有连接的输入神经元 i2 和 i4。由于上述第四稀疏处理单元 6013 在一个时钟周期可从 4 个输入神经元选择 2 个有连接（即有效）的输入神经元，对于连接关系数据“1011”，该第四稀疏处理单元 6013 在第一个时钟周期从上述输入神经元 i1，i2 和 i4 中选择输入神经元 i1 和 i3，并将该输入神经元 i1 和 i3 存储到上述第一输入缓存单元 606 中，在第二个时钟周期从上述输入神经元 i1，i2 和 i4 中选择输入神经元 i4，并将该输入神经元 i4 存储到上述第一输入缓存单元 606 中；对于连接关系数据“0101”，该第四稀疏处理单元 6013 在一个时钟周期从上述输入神经元 i2 和 i4 中选择输入神经元 i2 和 i4，如图 1i 所示。上述第四稀疏处理单元 6013 将上述输出神经元 i2 和 i4 和存储到第一输入缓存单元中。

举例说明，如图 1j 所示，输入数据为输入神经元 i1，i2，i3 和 i4，以步长索引的形式表示的连接关系数据为“021”，并且上述第四稀疏处理单元 6013 在一个时钟周期可从 4 个输入神经元选择 1 个有连接（即有效）的输入神经元。上述第四稀疏处理单元 6013 从上述输入数据缓存单元 6011 和上述连接关系缓存单元 6012 中分别获取上述输入神经元 i1，i2，i3 和 i4 和上述连接关系数据“021”后，上述第四稀疏处理单元 6013 根据该连接关系数据“1011”从上述输入神经元 i1，i2，i3 和 i4 选取有连接的输入神经元 i1，i3 和 i4。由于上述

第四稀疏处理单元 6013 在一个时钟周期可从 4 个输入神经元选择 1 个有连接（即有效）的输入神经元，该第四稀疏处理单元 6013 在三个时钟周期内依次输出输入神经元 i1，i3 和 i4，如图 1j 所示。上述第四稀疏处理单元 6013 将上述输入神经元 i1，i3 和 i4 存储到第一输入缓存单元中。

再举例说明，如图 1k 所示，输入数据为输入神经元 i1，i2，i3 和 i4，以步长索引的形式表示的连接关系数据有两组，分别为“021”和“22”，上述第四稀疏处理单元 6013 在一个时钟周期可从 4 个输入神经元中选择 2 个有连接（即有效）的输入神经元。上述第四稀疏处理单元 6013 根据上述连接关系数据“021”从上述输入神经元 i1，i2，i3 和 i4 中选择有连接的输入神经元 i1，i3 和 i4；根据上述连接关系数据“22”从上述输入神经元 i1，i2，i3 和 i4 中选择有连接的输入神经元 i2 和 i4。由于上述第四稀疏处理单元 6013 在一个时钟周期可从 4 个输入神经元选择 2 个有连接（即有效）的输入神经元，对于连接关系数据“021”，该第四稀疏处理单元 6013 在第一个时钟周期从选择输入神经元 i1 和 i3，并将该输入神经元 i1 和 i3 存储到上述第一输入缓存单元 606 中。在第二个时钟周期从选择输入神经元 i4 并将该输入神经元 i4 存储到上述第一输入缓存单元 606 中；对于连接关系数据“22”，该第四稀疏处理单元 6013 在一个时钟周期从选择输入神经元 i2 和 i4 并输出，如图 1k 所示。上述第四稀疏处理单元 6013 将上述输入神经元 i2 和 i4 存储到第一输入缓存单元中。

在一种可行的实施例中，上述输入数据缓存单元 6011 用于缓存的第一输入数据包括至少一个权值，上述连接关系缓存单元 6012 缓存的数据为上述权值的连接关系数据，且上述至少一个权值的绝对值均大于第一阈值时，上述第四稀疏处理单元 6013 根据上述权值的连接关系数据，将没有连接关系的输入神经元和输出神经元之间的权值的值置为 0，并将该值为 0 的权值和上述至少一个权值存储到第二输入缓存单元中。

举例说明，权值的形式为 wij，表示第 i 个输入神经元与第 j 个输出神经元之间的权值。假设输入神经元包括 i1，i2，i3 和 i4，输出神经元包括 o1，上述第一输入数据（权值）为 w11，w31，w41，上述第一输入数据的连接关系数据（即上述权值的连接关系数据）以直接索引的形式表示，为 1011，上述第四稀疏处理单元 6013 根据上述第二输入数据确定上述输入神经元 i2 与上述输出神经元 o1 之间没有连接，上述第四稀疏处理单元 6013 将该上述输入神经元 i2 与上述输出神经元 o1 之间的权值 w21 的值置为 0，并将 w11，w21（0），w31，w41 存储到第二输入缓存单元中。

上述第一输入缓存单元，用于缓存上述压缩处理后的输入神经元。上述第二输入缓存单元，用于缓存从存储电路中读取的压缩处理的权值。

在一种可行的实施例中，当上述第一输入数据为至少一个权值时，上述第四稀疏处理单元 6013 在一个时钟周期压缩处理一个权值和一个连接关系，即在一个时钟周期从 S3 个权值中选择一个有效的权值，该 S3 为大于 1 的整数。

可选地，上述第四稀疏处理单元 6013 在一个时钟周期压缩处理多个权值和多个连接关系数据，即一个时钟周期从 S3 个权值中选出有效的 S4 个权值，上述 S4 为大于 0 且小于或者等于该 S3 的整数。

上述第一输入缓存单元，用于缓存上述压缩处理后的权值。上述第二输入缓存单元，用于缓存从存储电路中读取的压缩处理的输入神经元。

需要说明的是，上述相关描述可参见前述实施例中的相关描述，在此不再叙述。

可选地，所述压缩映射电路 601 对所述第一输入数据进行压缩处理之前，所述压缩映射电路 601 还用于：对所述至少一个输入神经元进行分组，以得到 M 组输入神经元，所述 M 为大于或者等于 1 的整数；判断所述 M 组输入神经元的每一组输入神经元是否满足第一预设条件，所述第一预设条件包括一组输入神经元中绝对值小于或者等于第三阈值的输入神经元的个数小于或者等于第四阈值；当所述 M 组输入神经元任意一组输入神经元不满足所述第一预设条件时，将该组输入神经元删除；对所述至少一个权值进行分组，以得到 N 组权值，所述 N 为大于或者等于 1 的整数；判断所述 N 组权值的每一组权值是否满足第二预设条件，所述第二预设条件包括一组权值中绝对值小于或者等于第五阈值的权值的个数小于或者等于第六阈值；当所述 N 组权值任意一组权值不满足所述第二预设条件时，将该组权值删除。

需要说明的是，上述相关描述可参见前述实施例中的相关描述，在此不再叙述。上述第一阈值、第二阈值、第三阈值、第四阈值、第五阈值和第六阈值可均存储在存储电路或者第一输出缓存单元中；上述第一阈值、第二阈值、第三阈值、第四阈值、第五阈值和第六阈值中部分阈值存储在存储电路、部分阈值存储在第一输出缓存单元中。上述第一输入缓存单元、上述第二输入缓存单元和上述输出缓存单元均可为所述压缩映射电路或所述主处理电路中的功能单元，也可为其他处理电路共享的功能单元，本申请不做限定。

在一种可选实施例中，所述输入神经元的连接关系数据和所述权值的连接关系数据是由 0 或 1 表示的字符串/矩阵组成，其中 0 表示所述输入神经元/所述权值的绝对值小于或等于第一阈值，1 表示所述输入神经元/所述权值的绝对值大于第一阈值，与输出神经元无关。本实施例中，连接关系数据（即所述神经元/权值的连接关系数据）也可称为 mask 矩阵。

本申请中权值的连接关系数据和/或神经元的连接关系数据的表示方式除了直接索引和步长索引之外，还可为以下几种情况：列表的列表（List of Lists，LIL）、坐标列表（Coordinate list，COO）、压缩稀疏行（Compressed Sparse Row，CSR）、压缩稀疏列（Compressed Sparse Column，CSC）、（ELL Pack，ELL）以及混合（Hybird，HYB）等等，本申请不做详述。

此外，本申请实施例中提到的输入神经元和输出神经元并非是指整个神经网络的输入层中的神经元和输出层中的神经元，而是对于神经网络中任意相邻的两层神经元，处于网络前馈运算下层中的神经元即为输入神经元，处于网络前馈运算上层中的神经元即为输出神经元。以卷积神经网络为例，假设一个卷积神经网络有 L 层，K=1,2,3...L-1，对于第 K 层和第 K+1 层来说，第 K 层被称为输入层，该层中的神经元为上述输入神经元，第 K+1 层被称为输入层，该层中的神经元为上述输出神经元，即除了顶层之外，每一层都可以作为输入层，其下一层为对应的输出层。

下面提供一种采用如图 1a 所示的装置实现计算的方法，该计算的方法具体可以为神经网络的计算方式，例如神经网络的正向运算，神经网络的训练，在实际应用中，正向运算依据不同的输入数据可以执行矩阵乘矩阵、卷积运算、激活运算、变换运算等等运算，上述运算均可以采用如图 1a 所示的装置实现。

主处理电路的控制电路将数据通过分支处理电路传输给基础处理电路；其中，分支处

理电路可通过压缩映射电路先对数据进行压缩处理然后再转发给基础处理电路运算。例如，分支处理电路的压缩处理电路对数据进行压缩处理后再将压缩处理后的数据传输给基础处理电路，其优点是可以减少传输数据的数据量，减少传输的总比特数量，基础处理电路执行数据运算的效率也更高，功耗更低。

如分支处理电路接收到的数据为稀疏数据，那么分支处理电路可以收到数据后由压缩映射电路对数据进行压缩处理然后再进行计算，例如，分支处理电路收到主处理电路传输过来的稀疏数据，压缩映射电路将其进行压缩处理，然后发送给基础处理电路的内积运算器电路、向量运算器电路或累加器电路对压缩处理后的数据进行运算，提高运算效率，降低功耗。

主处理电路将待计算的数据传输到全部或者一部分基础处理电路上；以矩阵乘以向量计算为例，主处理电路的控制电路可以将矩阵数据拆分每列作为一个基础数据，例如 m*n 矩阵，可以拆分成 n 个 m 行的向量，主处理电路的控制电路将拆分后的 n 个 m 行的向量分发给多个基础处理电路。对于向量，主处理电路的控制电路可以将向量整体广播给每个基础处理电路。如果 m 的值比较大，那么控制电路可以先将 m*n 矩阵拆分成 x*n 个向量，以 x=2 为例，具体的可以拆分成，2n 个向量，每个向量包含 m/2 行，即将 n 个 m 行的向量中每个向量均分成 2 个向量，以第一行为例，如 n 个 m 行的向量的第一个向量为 1000 行，那么均分成 2 个向量可以为，将前 500 行组成第一向量，将后 500 行组成第二向量，控制电路通过 2 个广播将 2 个向量广播给多个基础处理电路。

所述数据传输的方式可以是广播或者分发，或者其他任何可能的传输方式；

基础处理电路接收到数据后，执行运算，得到运算结果；

基础处理电路将运算结果传输回主处理电路；

所述运算结果可以是中间运算结果，也可以是最终运算结果。

使用如图 1a 所示装置完成矩阵乘向量的运算；

（矩阵乘向量可以是矩阵中的每一行分别与向量进行内积运算，并将这些结果按对应行的顺序摆放成一个向量。）

下面描述计算尺寸是 M 行 L 列的矩阵 S 和长度是 L 的向量 P 的乘法的运算，如图 2a 所示，（矩阵 S 中的每一行与向量 P 长度相同，他们中的数据按位置一一对应）所述神经网络计算装置拥有 K 个基础处理电路：

参阅图 2，图 2 提供了了一种矩阵乘向量的实现方法，具体可以包括：

步骤 S201、主处理电路的控制电路将矩阵 S 中的每一行数据分发到 K 个基础处理电路中的某一个上，基础处理电路将接收到的分发数据保存在基础处理电路的片上缓存和/或寄存器中；

在一种可选方案中，当装置包括分支电路时，分支电路中包括压缩映射电路。主处理电路的控制电路将输入矩阵 S（M 行 L 列）中的每一行数据通过分支处理电路进行压缩处理后再分发到 K 个基础处理电路中的某一个上，基础处理电路将接收到的分发数据保存在基础处理电路的片上缓存和/或寄存器中。

具体的，分支处理电路可接收到主处理电路分发的输入矩阵 S1（M1 行 L1 列）其中，M1 小于等于 M，L1 小于等于 L。即 S1 属于 S 的一部分，即前文所述的分发数据块。进一

步地，分支处理电路的压缩映射电路将输入矩阵 S1（M1 行 L1 列）中的每一行数据进行压缩处理得到压缩处理后的矩阵 S2（M2 行 L2 列）。然后再将压缩处理后的矩阵 S2 转发给基础处理电路。其中， M 大于等于 M1，且大于等于 M2。L 大于等于 L1，且大于等于 L2。

例如，压缩映射电路将输入矩阵 S2 和矩阵 P2 中数据为指定数值（如 0）和/或数据小于预设阈值（如 0.1）所对应的数据剔除，具体实现时可根据矩阵 S2 和矩阵 P2 各自对应的 mask 矩阵来剔除，例如剔除 mask 矩阵中数据为 0 时对应的相同位置上矩阵 S2/P2 中的数据，具体可参见前述关于数据压缩处理实施例中的相关阐述，这里不再赘述。应理解的，这里的矩阵 S 和矩阵 P 也可对应理解为前述实施例中的输入神经元（也可称为输入神经原矩阵）和权值（也可称为权值矩阵）等。

在一种可选方案中，如果矩阵 S 的行数 M<=K 则，主处理电路的控制电路给 K 个基础处理电路分别分发 S 矩阵的一行数据；

在一种可选方案中，如果矩阵 S 的行数 M>K，则主处理电路的控制电路给每个基础处理电路分别分发 S 矩阵中一行或多行的数据。

分发到第 i 个基础处理电路的 S 中的行的集合为 Ai，共有 Mi 个行，如图 2c 表示第 i 个基础处理电路上将要执行的计算。

在一种可选方案中，在每个基础处理电路中，例如第 i 个基础处理电路中，可以将接收到的分发数据例如矩阵 Ai 保存在第 i 个基础处理电路的寄存器和/或片上缓存中；优点是减少了之后的分发数据的数据传输量，提高了计算效率，降低了功耗。

步骤 S202、 主处理电路的控制电路将向量 P 中各部分以广播的方式传输给 K 个基础处理电路；

在一种可选方案中，当装置包括分支电路时，分支电路中包括压缩映射电路。主处理电路的控制电路将输入向量 P（长度为 L）中各个部分以广播的方式通过对应的分支处理电路进行压缩处理后再传输给 K 个基础处理电路；

具体的，分支处理电路可接收到主处理电路分发的输入向量 P1（长度 L1）其中，L1 小于等于 L。P1 属于 P 的一部分，即前文所述的广播数据块。进一步地，分支处理电路的压缩映射电路将输入向量 P1（长度 L1）中的数据进行压缩处理得到压缩处理后的向量 P2（L2 列）。然后在将压缩处理后的向量 P2 转发给基础处理电路。其中， L2 小于等于 L1，且小于等于 L。

在一种可选方案中，主处理电路的控制电路可以将向量 P 中各部分只广播一次到各个基础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对这一次得到的向量 P 的数据进行充分地复用，完成对应与矩阵 Ai 中每一行的内积运算。优点是，减少从主处理电路到基础处理电路的向量 P 的重复传输的数据传输量，提高执行效率，降低传输功耗。

在一种可选方案中，主处理电路的控制电路可以将向量 P 中各部分多次广播到各个基础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对每次得到的向量 P 的数据不进行复用，分次完成对应于矩阵 Ai 中的每一行的内积运算；优点是，减少基础处理电路内部的单次传输的向量 P 的数据传输量，并可以降低基础处理电路缓存和/或寄存器的容量，提高执行效率，降低传输功耗，降低成本。

在一种可选方案中，主处理电路的控制电路可以将向量 P 中各部分多次广播到各个基

础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对每次得到的向量 P 的数据进行部分复用，完成对应于矩阵 Ai 中的每一行的内积运算；优点是，减少从主处理电路到基础处理电路的数据传输量，也减少基础处理电路内部的数据传输量，提高执行效率，降低传输功耗。

步骤 S203、 K 个基础处理电路各自的内积运算器电路计算矩阵 S 和向量 P 的数据的内积，例如第 i 个基础处理电路，计算矩阵 Ai 的数据和向量 P 的数据的内积；

在一种可选方案中，当装置中的压缩映射电路设置在基础处理电路时，

基础处理电路接收到主处理电路发送的矩阵 S 和向量 P 后，可利用基础处理电路中的压缩映射电路先对矩阵 S 和向量 P 进行压缩处理，然后再利用内积运算器电路计算压缩处理后的矩阵 S 和向量 P 的数据的内积。

具体的，压缩映射电路对输入矩阵 S（M1 行 L1 列）进行压缩处理得到压缩处理后的矩阵 S（M 行 L 列）。例如，将输入矩阵 S 和向量 P 中数据为指定数值（如 0）和/或数据小于预设阈值（如 0.1）所对应的数据剔除，具体实现时可根据矩阵 S 和向量 P 各自对应的 mask 矩阵来剔除，例如剔除 mask 矩阵中数据为 0 时对应的相同位置上矩阵 S/P 中的数据，具体可参见前述关于数据压缩处理实施例中的相关阐述，这里不再赘述。应理解的，这里的矩阵 S 和矩阵 P 也可对应理解为前述实施例中的输入神经元（也可称为输入神经元矩阵）和权值（也可称为权值矩阵）等。

步骤 S204、 K 个基础处理电路的累加器电路将内积运算的结果进行累加得到累加结果，将累加结果以定点类型形式传输回主处理电路。

在一种可选方案中，可以将每次基础处理电路执行内积运算得到的部分和（部分和即累加结果的一部分，例如累加结果为：F1*G1+ F2*G2+ F3*G3+ F4*G4+ F5*G5,那么部分和可以为：F1*G1+ F2*G2+ F3*G3 的值）传输回主处理电路进行累加；优点是，减少了基础处理电路内部的运算量，提高基础处理电路的运算效率。

在一种可选方案中，也可以将每次基础处理电路执行的内积运算得到的部分和保存在基础处理电路的寄存器和/或片上缓存中，累加结束之后传输回主处理电路；优点是，减少了基础处理电路和主处理电路之间的数据传输量，提高了运算效率，降低了数据传输功耗。

在一种可选方案中，也可以将每次基础处理电路执行的内积运算得到的部分和在部分情况下保存在基础处理电路的寄存器和/或片上缓存中进行累加，部分情况下传输到主处理电路进行累加，累加结束之后传输回主处理电路；优点是，减少了基础处理电路和主处理电路之间的数据传输量，提高了运算效率，降低了数据传输功耗，减少了基础处理电路内部的运算量，提高基础处理电路的运算效率。

参阅图 2b，使用如图 1a 所示的装置完成矩阵乘矩阵的运算；

下面描述计算尺寸是 M 行 L 列的矩阵 S 和尺寸是 L 行 N 列的矩阵 P 的乘法的运算,（矩阵 S 中的每一行与矩阵 P 的每一列长度相同，如图 2d 所示）所述神经网络计算装置拥有 K 个基础处理电路：

步骤 S201b、主处理电路的控制电路将矩阵 S 中的每一行数据分发到 K 个基础处理电路中的某一个上，基础处理电路将接收到的数据保存在片上缓存和/或寄存器中；

在一种可选方案中，分支处理电路中设置有压缩映射电路，主处理电路的控制电路将

矩阵 S 中的每一行数据通过分支处理电路进行压缩处理后再分发到 K 个基础处理电路中的某一个上，基础处理电路将接收到的数据保存在片上缓存和/或寄存器中；

具体的，主处理电路的控制电路将输入矩阵 S（M 行 L 列）中的每一行数据通过分支处理电路进行压缩处理后再分发到 K 个基础处理电路中的某一个上。相应地，分支处理电路可接收到主处理电路分发的输入矩阵 S1（M1 行 L1 列）其中， M1 小于等于 M，L1 小于等于 L。进一步地，分支处理电路的压缩映射电路将输入矩阵 S1（M1 行 L1 列）中的每一行数据进行压缩处理得到压缩处理后的矩阵 S2（M2 行 L2 列）。然后在将压缩处理后的矩阵 S2 转发给对应的基础处理电路。其中， M 大于等于 M1，且大于等于 M2。L 大于等于 L1，且大于等于 L2。

例如，压缩映射电路将输入矩阵 S2 和矩阵 P2 中数据为指定数值（如 0）和/或数据小于预设阈值（如 0.1）所对应的数据剔除，具体实现时可根据矩阵 S2 和矩阵 P2 各自对应的 mask 矩阵来剔除，例如剔除 mask 矩阵中数据为 0 时对应的相同位置上矩阵 S2/P2 中的数据，具体可参见前述关于数据压缩处理实施例中的相关阐述，这里不再赘述。应理解的，这里的矩阵 S 和矩阵 P 也可对应理解为前述实施例中的输入神经元（也可称为输入神经原矩阵）和权值（也可称为权值矩阵）等。

在一种可选方案中，如果 S 的行数 M<=K 则，主处理电路的控制电路给 M 个基础处理电路分别分发 S 矩阵的一行；

在一种可选方案中，如果 S 的行数 M>K，主处理电路的控制电路给每个基础处理电路分别分发 S 矩阵中一行或多行的数据。

S 中有 Mi 行分发到第 i 个基础处理电路，这 Mi 行的集合称为 Ai，如图 2e 表示第 i 个基础处理电路上将要执行的计算。

在一种可选方案中，在每个基础处理电路中，例如第 i 个基础处理电路中：

接收的由主处理电路分发的矩阵 Ai，将矩阵 Ai 保存在第 i 个基础处理电路寄存器和/或片上缓存中；优点是减少了之后的数据传输量，提高了计算效率，降低了功耗。

步骤 S202b、 主处理电路的控制电路将矩阵 P 中各部分以广播的方式传输给各个基础处理电路；

在一种可选方案中，分支处理电路中设置有压缩映射电路，主处理电路的控制电路将矩阵 P 中各部分以广播的方式通过分支处理电路压缩处理后再传输给各个基础处理电路；

具体的，分支处理电路可接收到主处理电路分发的输入向量 P1（长度 L1）其中，L1 小于等于 L。P1 属于 P 的一部分，即前文所述的广播数据块。进一步地，分支处理电路的压缩映射电路将输入向量 P1（长度 L1）中的数据进行压缩处理得到压缩处理后的向量 P2（L2 列）。然后再将压缩处理后的向量 P2 转发给基础处理电路。其中， L2 小于等于 L1，且小于等于 L。

在一种可选方案中，可以将矩阵 P 中各部分只广播一次到各个基础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对这一次得到的矩阵 P 的数据进行充分地复用，完成对应与矩阵 Ai 中每一行的内积运算；本实施例中的复用具体可以为基础处理电路在计算中重复使用，例如矩阵 P 的数据的复用，可以是对矩阵 P 的数据在多次使用。

在一种可选方案中，主处理电路的控制电路可以将矩阵 P 中各部分多次广播到各个基

础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对每次得到的矩阵 P 的数据不进行复用，分次完成对应于矩阵 Ai 中的每一行的内积运算；

在一种可选方案中，主处理电路的控制电路可以将矩阵 P 中各部分多次广播到各个基础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对每次得到的矩阵 P 的数据进行部分复用，完成对应于矩阵 Ai 中的每一行的内积运算；

在一种可选方案中，每个基础处理电路，例如第 i 个基础处理电路，计算矩阵 Ai 的数据和矩阵 P 的数据的内积；

步骤 S203b、每个基础处理电路的累加器电路将内积运算的结果进行累加并传输回主处理电路。

在一种可选方案中，基础处理电路中设置有压缩映射电路，则内积运算的结果可为基础处理电路对矩阵 S 和矩阵 P 进行压缩处理后，再利用内积运算器电路计算压缩处理后的矩阵 S 和向量 P 的数据的内积的结果。

具体的，压缩映射电路对输入矩阵 S（M1 行 L1 列）和输入矩阵 P（L1 行 N1 列）进行压缩处理得到压缩处理后的矩阵 S（M 行 L 列）和矩阵 P（L 行 N 列）；进一步地基础处理单元的运算器可对压缩处理后的矩阵 S 和矩阵 P 进行内积运算，以得到内积运算的结构。例如，将输入矩阵 S 和矩阵 P 中数据为指定数值（如 0）和/或数据小于预设阈值（如 0.1）所对应的数据剔除，具体实现时可根据矩阵 S 和矩阵 P 各自对应的 mask 矩阵来剔除，例如剔除 mask 矩阵中数据为 0 时对应的相同位置上矩阵 S/P 中的数据，具体可参见前述关于数据压缩处理实施例中的相关阐述，这里不再赘述。应理解的，这里的矩阵 S 和矩阵 P 也可对应理解为前述实施例中的输入神经元（也可称为输入神经原矩阵）和权值（也可称为权值矩阵）等。

在一种可选方案中，基础处理电路可以将每次执行内积运算得到的部分和传输回主处理电路进行累加；

在一种可选方案中，也可以将每次基础处理电路执行的内积运算得到的部分和保存在基础处理电路的寄存器和/或片上缓存中，累加结束之后传输回主处理电路；

在一种可选方案中，也可以将每次基础处理电路执行的内积运算得到的部分和在部分情况下保存在基础处理电路的寄存器和/或片上缓存中进行累加，部分情况下传输到主处理电路进行累加，累加结束之后传输回主处理电路；

参阅图 3a，使用如图 1a 所示的装置完成全连接运算：

如果全连接层的输入数据是一个向量（即神经网络的输入是单个样本的情况），则以全连接层的权值矩阵作为矩阵 S，输入向量作为向量 P，按照所述装置的使用方法一执行如图 2 所示的矩阵乘向量的运算；

如果全连接层的输入数据是一个矩阵（即神经网络的输入是多个样本作为 batch 的情况），则以全连接层的权值矩阵作为矩阵 S，输入向量作为矩阵 P，或者以全连接层的权值矩阵作为矩阵 P，输入向量作为矩阵 S，按照所述装置的使用如图 2c 所示的矩阵乘矩阵的执行运算；

参阅图 3b，使用如图 1a 所示的装置完成卷积运算：

对于一个卷积层，记其卷积核的数量为 M；

步骤 S301、主处理电路的控制电路将卷积层权值中的每一个卷积核的权值分发到 K 个基础处理电路中的某一个上，保存在基础处理电路的片上缓存和/或寄存器中；

在一种可选方案中，分支处理电路中包括压缩映射电路，则主处理电路的控制电路将卷积层权值中的每一个卷积核的权值通过分支处理电路压缩处理后再分发到 K 个基础处理电路中的某一个上，保存在基础处理电路的片上缓存和/或寄存器中；

具体的，分支处理电路接收到主处理电路发送的所述卷积层权值中的每一个卷积核的权值后，可利用分支处理电路的压缩映射电路对卷积层权值中的每一个卷积核的权值进行压缩处理，以对应得到压缩处理后的所述卷积层权值中的每一个卷积核的权值，然后再转发给基础处理电路进行运算。关于数据的压缩处理可参见前述实施例中的相关阐述，这里不再赘述。

在一种可选方案中，如果卷积核的个数 M<=K 则，主处理电路的控制电路给 M 个基础处理电路分别分发一个卷积核的权值；

在一种可选方案中，如果卷积核的个数 M>K，主处理电路的控制电路给每个基础处理电路分别分发一个或多个卷积核的权值。

共有 Mi 个卷积核分发到第 i 个基础处理电路，这些卷积核权值的集合称为 Ai。

在一种可选方案中，在每个基础处理电路中，例如第 i 个基础处理电路中：

将收到的由主处理电路分发的卷积核权值 Ai 保存在其寄存器和/或片上缓存中；

步骤 S302、主处理电路的控制电路将输入数据 P 中各部分以广播的方式传输给各个基础处理电路；

在一种可选方案中，分支处理电路包括压缩映射电路，则主处理电路的控制电路将输入数据 P 中各部分以广播的方式通过相应地分支处理电路压缩处理后再转发给各个基础处理电路，这里不再赘述。

在一种可选方案中，主处理电路的控制电路可以将输入数据 P 中各部分只广播一次到各个基础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对这一次得到的输入数据 P 的数据进行充分地复用，完成对应与 Ai 中每一个卷积核的内积运算；

在一种可选方案中，主处理电路的控制电路可以将输入数据 P 中各部分多次广播到各个基础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对每次得到的输入数据 P 的数据不进行复用，分次完成对应于 Ai 中的每一个卷积核的内积运算；

在一种可选方案中，主处理电路的控制电路可以将输入数据 P 中各部分多次广播到各个基础处理电路的寄存器或者片上缓存中，第 i 个基础处理电路对每次得到的输入数据 P 的数据进行部分复用，完成对应于 Ai 中的每一个卷积核的内积运算；

步骤 S303、每个基础处理电路计算卷积核和输入数据 P 的数据内积，例如第 i 个基础处理电路，计算 Ai 的每一个卷积核和输入数据 P 的数据的内积；

在一种可选方案中，基础处理电路包括压缩映射电路时，则基础处理电路接收到主处理电路发送的卷积核和输入数据 P 后，可利用基础处理电路中的压缩映射电路先对卷积核和输入数据 P 进行压缩处理，然后再利用内积运算器电路计算压缩处理后的卷积核和输入数据 P 的数据的内积。例如，第 i 个基础处理电路，计算压缩处理后的 Ai 的每一个卷积核和压缩处理后的输入数据 P 的数据的内积。

步骤 S304、每个基础处理电路的累加器电路将内积运算的结果进行累加并传输回主处理电路：

在一种可选方案中，可基础处理电路以将每次执行内积运算得到的部分和传输回主处理电路进行累加；

在一种可选方案中，基础处理电路也可以将每次执行的内积运算得到的部分和保存在基础处理电路的寄存器和/或片上缓存中，累加结束之后传输回主处理电路；

在一种可选方案中，基础处理电路也可以将每次执行的内积运算得到的部分和在部分情况下保存在基础处理电路的寄存器和/或片上缓存中进行累加，部分情况下传输到主处理电路进行累加，累加结束之后传输回主处理电路；

使用如图 1a 所示的装置更新权值的方法：

利用主处理电路的向量运算器电路实现神经网络训练过程中的权值更新功能，具体地，权值更新是指使用权值的梯度来更新权值的方法。

在一种可选方案中，使用主处理电路的向量运算器电路对权值和权值梯度这两个向量进行加减运算得到运算结果，该运算结果即为更新权值。

在一种可选方案中，使用主处理电路的向量运算器电路在权值以及权值梯度乘以或除以一个数得到中间权值和中间权值梯度值，向量运算器电路对中间权值和中间权值梯度值进行加减运算得到运算结果，该运算结果即为更新权值。

在一种可选方案中，可以先使用权值的梯度计算出一组动量，然后再使用动量与权值进行加减计算得到更新后的权值。

本申请还提供一种芯片，该芯片包含计算装置，该计算装置包括：

一个主处理电路，主处理电路中所涉及到的数据可以是压缩处理后的数据，在一种可选实施例中，所述压缩处理后的数据包括至少一个输入神经元或至少一个权值，所述至少一个神经元中的每个神经元大于第一阈值或者，所述至少一个权值中的每个权值大于第二阈值。所述第一阈值和所述第二阈值为用户侧自定义设置的，它们可以相同，也可不同。

在一种可选方案中，主处理电路包括压缩映射电路；在一种可选方案中，主处理电路包括执行数据压缩处理的运算单元，例如向量运算单元；具体地，包含接收输入数据的数据输入接口；

在一种可选方案中，该计算装置还包括：一个分支处理电路，分支处理电路中所涉及到的数据可以是压缩处理后的数据，在一种可选实施例中，所述压缩处理后的数据包括至少一个输入神经元或至少一个权值，所述至少一个神经元中的每个神经元大于第一阈值或者，所述至少一个权值中的每个权值大于第二阈值。所述第一阈值和所述第二阈值为用户侧自定义设置的，它们可以相同，也可不同。

在一种可选方案中，分支处理电路包括压缩映射电路；

在一种可选方案中，分支处理电路包括执行数据压缩处理的运算单元，如向量运算单元等；具体地，包含接收输入数据的数据输入接口；

在一种可选方案中，所述接收的数据来源可以是：所述神经网络运算电路装置的外部或所述神经网络运算电路装置的部分或全部基础处理电路；

在一种可选方案中，所述数据输入接口可以有多个；具体地，可以包含输出数据的数

据输出接口；

在一种可选方案中，所述输出的数据的去向可以是：所述神经网络运算装置的外部或所述神经网络运算电路装置的部分或全部基础处理电路；

在一种可选方案中，所述数据输出接口可以有多个；

在一种可选方案中，所述分支处理电路包括片上缓存和/或寄存器；

在一种可选方案中，所述分支处理电路中包含运算单元，可以执行数据运算；

在一种可选方案中，所述分支处理电路中包含算术运算单元；

在一种可选方案中，所述分支处理电路中包含向量运算单元，可以同时对一组数据执行运算；具体地，所述算术运算和/或向量运算可以是任意类型的运算，包括但不限于：两个数相加减乘除，一个数与常数加减乘除，对一个数执行指数运算，幂次运算，对数运算，以及各种非线性运算，对两个数执行比较运算，逻辑运算等。两个向量相加减乘除，一个向量中的每一个元素与常数加减乘除，对向量中的每一个元素执行指数运算，幂次运算，对数运算，以及各种非线性运算等，对一个向量中的每两个对应的元素执行比较运算，逻辑运算等。

在一种可选方案中，所述主处理电路包括数据重排列单元，用于按照一定的顺序向基础处理电路传输数据，或者按照一定的顺序原地重新排列数据；

在一种可选方案中，所述数据排列的顺序包括：对一个多维数据块进行维度顺序的变换；所述数据排列的顺序还可以包括：对一个数据块进行分块以发送到不同的基础处理电路。

该计算装置还包括多个基础处理电路：每一个基础处理电路用于计算两个向量的内积，计算的方法是，基础处理电路收到的两组数，将这两组数中的元素对应相乘，并且将相乘的结果累加起来；内积的结果传输出去，这里传输出去根据基础处理电路的位置，有可能传输给其他基础处理电路，也可以直接传输给主处理电路。

基础处理电路中所涉及到的数据可以是压缩处理后的数据，在一种可选实施例中，所述压缩处理后的数据包括至少一个输入神经元或至少一个权值，所述至少一个神经元中的每个神经元大于第一阈值或者，所述至少一个权值中的每个权值大于第二阈值。所述第一阈值和所述第二阈值为用户侧自定义设置的，它们可以相同，也可不同。

在一种可选方案中，基础处理电路包括压缩映射电路；

在一种可选方案中，基础处理电路包括执行数据压缩处理的向量运算单元；

具体地，包括由片上缓存和/或寄存器构成的存储单元；

具体地，包括一个或多个接收数据的数据输入接口；

在一种可选方案中，包括两个数据输入接口，每次从两个数据输入接口处可以分别获得一个或多个数据；

在一种可选方案中，基础处理电路可以将从数据输入接口接收到输入数据后保存在寄存器和/或片上缓存中；

上述数据输入接口接收数据的来源可以是：其他基础处理电路和/或主处理电路。

所述神经网络运算电路装置的主处理电路；

所述神经网络运算电路装置的其他基础处理电路（所述神经网络运算电路装置拥有多

个基础处理电路）；

具体地，包括一个或多个传输输出数据的数据输出接口；

在一种可选方案中，可以将一个或多个数据从数据输出接口传输出去；

具体地，通过数据输出接口传输出去的数据可以是：从数据输入接口接收到的数据、保存在片上缓存和/或寄存器中的数据、乘法器运算结果、累加器运算结果或内积运算器运算结果中的一种或任意组合。

在一种可选方案中，包含三个数据输出接口，其中的两个分别对应于两个数据输入接口，每一层输出上一层从数据输入接口接收到的数据，第三个数据输出接口负责输出运算结果；

具体地，所述数据输出接口传输数据的去向可以是：上文数据来源和此处的数据去向决定了基础处理电路在装置中的连接关系。

所述神经网络运算电路装置的主处理电路；

所述神经网络运算电路装置的其他基础处理电路，所述神经网络运算电路装置拥有多个基础处理电路；

具体地，包括算术运算电路：该算术运算电路具体可以为：一个或多个乘法器电路、一个或多个累加器电路、一个或多个执行两组数内积运算的电路中的一个或任意组合。

在一种可选方案中，可以执行两个数的乘法运算，其结果可以保存在片上缓存和/或寄存器上，也可以直接累加到寄存器和/或片上缓存中；

在一种可选方案中，可以执行两组数据的内积运算，其结果可以保存在片上缓存和/或寄存器中，也可以直接累加到寄存器和/或片上缓存中；

在一种可选方案中，可以执行数据的累加运算，将数据累加到片上缓存和或寄存器中；

具体地，累加器电路被累加的数据，可以是：从数据输入接口接收到的数据、保存在片上缓存和/或寄存器中的数据、乘法器运算结果、累加器运算结果、内积运算器运算结果中的一个或任意组合。

需要说明的是，上述对基础处理电路的描述中所用到的“数据输入接口”和“数据输出接口”是指每一个基础处理电路的数据输入与输出接口，而不是整个装置的数据输入与输出接口。

在本申请另一方面提供的集成电路芯片装置中，包括：主处理电路以及多个基础处理电路；

所述多个基础处理电路呈阵列分布；每个基础处理电路与相邻的其他基础处理电路连接，所述主处理电路连接所述多个基础处理电路中的 k 个基础处理电路，所述 k 个基础电路为：第 1 行的 n 个基础处理电路以及第 1 列的 m 个基础处理电路；

所述多个基础处理电路中的部分或所有基础处理电路包括：压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及和与所述 k 个基础处理电路传输数据；

所述 k 个基础处理电路，用于在所述主处理电路以及多个基础处理电路之间的数据转

发；

所述部分或所有基础处理电路，用于依据传输数据的运算控制确定是否启动所述压缩映射电路对所述传输数据进行压缩处理，依据压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果通过与所述主处理电路连接的基础处理电路传输给所述主处理电路。

在一种可选方案中，在所述多个基础处理电路均包括压缩映射电路时，所述多个基础处理电路，用于依据传输数据的运算控制确定是否启动所述压缩映射电路对所述传输数据进行压缩处理，依据压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果通过与所述 k 个基础处理电路传输给所述主处理电路。

在一种可选方案中，所述主处理电路，用于获取待计算的数据块以及运算指令，依据该运算指令对所述待计算的数据块划分成分发数据块以及广播数据块；对所述分发数据块进行拆分处理得到多个基本数据块，将所述多个基本数据块分发至与所述 K 个基础处理电路，将所述广播数据块广播至与所述 k 个基础处理电路；所述多个基础处理电路，用于依据接收到的基础数据块、广播数据块以及运算指令启动所述压缩映射电路将基础数据块和广播数据块进行压缩处理，对压缩处理后的所述基本数据块与压缩处理后的所述广播数据块执行内积运算得到运算结果，将运算结果通过所述 k 个基础处理电路传输给所述主处理电路；所述主处理电路，用于对所述运算结果处理得到所述待计算的数据块以及运算指令的指令结果；其中，所述分发数据块以及所述广播数据块为至少一个输入神经元或者，至少一个权值。

在一种可选方案中，在所述多个基础处理电路中的所述 k 个基础处理电路均包括压缩映射电路时，所述 k 个基础处理电路，用于依据传输数据的运算控制确定是否启动所述压缩映射电路对所述传输数据行压缩处理，并将压缩处理后的传输数据发送给与所述 k 个基础处理电路连接的基础处理电路；所述多个基础处理电路，用于依据压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果通过与所述主处理电路连接的基础处理电路传输给所述主处理电路。

在一种可选方案中，所述主处理电路，用于获取待计算的数据块以及运算指令，依据该运算指令对所述待计算的数据块划分成分发数据块以及广播数据块；对所述分发数据块进行拆分处理得到多个基本数据块，将所述多个基本数据块分发至与所述 K 个基础处理电路，将所述广播数据块广播至与所述 k 个基础处理电路；所述 k 个基础处理电路，用于依据接收到的基础数据块、广播数据块以及运算指令启动所述压缩映射电路将基础数据块和广播数据块进行压缩处理，然后传输给与所述 k 个基础处理电路连接的基础处理电路；所述多个基础处理电路，用于对压缩处理后的所述基本数据块与所述广播数据块执行内积运算得到运算结果，并将所述运算结果发送至所述主处理电路；所述主处理电路，用于对所述运算结果处理得到所述待计算的数据块以及运算指令的指令结果；其中，所述分发数据块以及所述广播数据块为至少一个输入神经元或者，至少一个权值。

参阅图 4a，图 4a 为本披露提供的一种集成电路芯片装置，该集成电路芯片装置包括：主处理电路和多个基础处理电路，所述多个基础处理电路呈阵列排布（m*n 阵列），其中，m、n 的取值范围为大于等于 1 的整数且 m、n 中至少有一个值大于等于 2。对于 m*n 阵列

分布的多个基础处理电路，每个基础处理电路与相邻的基础处理电路连接，所述主处理电路连接多个基础处理电路中的 k 个基础处理电路，所述 k 个基础处理电路可以为：第 1 行的 n 个基础处理电路、第 m 行的 n 个基础处理电路以及第 1 列的 m 个基础处理电路。如图 1a 所示的集成电路芯片装置，主处理电路和/或多个基础处理电路可以包括压缩映射电路，具体的多个基础处理电路中可以有部分基础处理电路包括压缩映射电路，例如，在一个可选的技术方案中，可以将 k 个基础处理电路配置压缩映射电路，这样 n 个基础处理电路可以分别负责对本列的 m 个基础处理电路的数据进行数据压缩处理步骤。此设置能够提高运算效率，降低功耗，因为对于第 1 行的 n 个基础处理电路来说，由于其最先接收到主处理电路发送的数据，那么将该接收到的数据进行压缩处理可以减少后续基础处理电路的计算量以及与后续基础处理电路的数据传输的量，同理，对于第一列的 m 个基础处理电路配置压缩映射电路也具有计算量小和功耗低的优点。另外，依据该结构，主处理电路可以采用动态的数据发送策略，例如，主处理电路向第 1 列的 m 个基础处理电路广播数据，主处理电路向第 1 行的 n 个基础处理电路发送分发数据，此优点是，通过不同的数据输入口传递不同的数据到基础处理电路内，这样基础处理电路可以不区分该接收到的数据是何种数据，只需要确定该数据从哪个接收端口接收即可以获知其属于何种数据。

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及和与其相连的所述基础处理电路传输数据；上述连续的运算但不限于：累加运算、ALU 运算、激活运算等等运算。

所述多个基础处理电路，用于依据传输的数据以并行方式执行神经网络中的运算，并将运算结果通过与所述主处理电路连接的基础处理电路传输给所述主处理电路。上述并行方式执行神经网络中的运算包括但不限于：内积运算、矩阵或向量乘法运算等等。具体的，所述多个基础处理电路可先对传输的数据进行压缩处理，然后再依据压缩处理后的数据以并行方式执行神经网络中的运算。

主处理电路可以包括：数据发送电路、数据接收电路或接口，该数据发送电路可以集成数据分发电路以及数据广播电路，当然在实际应用中，数据分发电路以及数据广播电路也可以分别设置。对于广播数据，即需要发送给每个基础处理电路的数据。对于分发数据，即需要有选择的发送给部分基础处理电路的数据，具体的，如卷积运算，卷积运算的卷积输入数据需要发送给所有的基础处理电路，所以该卷积输入数据为广播数据，卷积核需要有选择的发送给部分基础数据块，所以卷积核为分发数据。分发数据具体的选择发送给哪个基础处理电路的方式可以由主处理电路依据负载以及其他分配方式进行具体的确定。对于广播发送方式，即将广播数据以广播形式发送至每个基础处理电路。(在实际应用中，通过一次广播的方式将广播数据发送至每个基础处理电路，也可以通过多次广播的方式将广播数据发送至每个基础处理电路，本披露具体实施方式并不限制上述广播的次数)，对于分发发送方式，即将分发数据有选择的发送给部分基础处理电路。

可选的，对于第 m 行 n 个基础处理电路的累加器电路可以执行内积运算的累加运算，因为对于第 m 行基础处理电路来说，其能够接收到本列所有的基础处理电路的乘积结果，而通过第 m 行的 n 个基础处理电路执行内积运算的累加运算，这样能够对计算资源进行有效的分配，具有节省功耗的优点。此技术方案尤其对于 m 数量较大时更为适用。

对于数据的压缩处理可以由主处理电路来分配执行的电路，具体的，可以通过显示或隐式的方式来分配执行的电路，对于显示方式，主处理电路可以配置一个特殊指示或指令，当基础处理电路接收到该特殊指示或指令时，确定执行数据压缩处理，如基础处理电路未接收到特殊指示或指令时，确定不执行数据压缩处理。又如，可以以暗示的方式来执行，例如，基础处理电路接收到数据类型为稀疏数据（即含 0，或包括小于预设阈值的数据大于预设数量）且确定需要执行内积运算时，将对稀疏数据进行压缩处理。对于显示配置的方式，特殊指令或指示可以配置一个递减序列，该递减序列每经过一个基础处理电路，数值减 1，基础处理电路读取该递减序列的值，如该值大于零，则执行数据压缩处理，如该值等于或小于零，则不执行数据压缩处理。此设置是依据阵列分配的基础处理电路所配置的，例如对于第 i 列的 m 个基础处理电路来说，主处理电路需要前面 5 个基础处理电路执行属于压缩处理，则主处理电路下发一个特殊指令，该特殊指令包含有递减序列，该递减序列的初始值可以为 5，则每经过一个基础处理电路，递减序列的值即减 1，到第 5 个基础处理电路时，该递减序列的值为 1，到第 6 个基础处理电路时，该递减序列为 0，此时第 6 个基础处理电路将不再执行该数据压缩处理，此种方式可以使得主处理电路可以动态的配置数据压缩处理的执行主体以及执行次数。

本披露一个实施例提供一种集成电路芯片装置，包括一个主处理电路（也可以称为主单元）和多个基础处理电路(也可以称为基础单元)；实施例的结构如图 4b 所示；其中，虚线框中是所述神经网络运算装置的内部结构；灰色填充的箭头表示主处理电路和基础处理电路阵列之间的数据传输通路，空心箭头表示基础处理电路阵列中各个基础处理电路（相邻基础处理电路）之间的数据传输通路。其中，基础处理电路阵列的长度宽度可以不同，即 m、n 的取值可以不同，当然也可以相同，本披露并不限制上述取值的具体值。

基础处理电路的电路结构如图 4c 所示；图中虚线框表示基础处理电路的边界，与虚线框交叉的粗箭头表示数据输入输出通道（指向虚线框内是输入通道，指出虚线框是输出通道）；虚线框中的矩形框表示存储单元电路（寄存器和/或片上缓存），包括输入数据 1，输入数据 2，乘法或内积结果，累加数据；菱形框表示运算器电路，包括乘法或内积运算器，加法器。

本实施例中，所述神经网络运算装置包括一个主处理电路和 16 个基础处理电路（16 个基础处理电路仅仅为了举例说明，在实际应用中，可以采用其他的数值）；

本实施例中，基础处理电路有两个数据输入接口，两个数据输出接口；在本例的后续描述中，将横向的输入接口（图 4b 中指向本单元的横向箭头）称作输入 0，竖向的输入接口（图 4b 中指向本单元的竖向箭头）称作输入 1；将每一个横向的数据输出接口（图 4b 中从本单元指出的横向箭头）称作输出 0，竖向的数据输出接口（图 4b 中从本单元指出的竖向箭头）称作输出 1。

每一个基础处理电路的数据输入接口和数据输出接口可以分别连接不同的单元，包括主处理电路与其他基础处理电路；

本例中，基础处理电路 0,4,8,12（编号见图 4b）这四个基础处理电路的输入 0 与主处理电路的数据输出接口连接；

本例中，基础处理电路 0,1,2,3 这四个基础处理电路的输入 1 与主处理电路的数据输出

接口连接；

本例中，基础处理电路 12,13,14,15 这四个基础处理电路的输出 1 与主处理电路的数据输入接口相连；

本例中，基础处理电路输出接口与其他基础处理电路输入接口相连接的情况见图 1b 所示，不再一一列举；

具体地，S 单元的输出接口 S1 与 P 单元的输入接口 P1 相连接，表示 P 单元将可以从其 P1 接口接收到 S 单元发送到其 S1 接口的数据。

本实施例包含一个主处理电路，主处理电路与外部装置相连接（即有输入接口也有输出接口），主处理电路的一部分数据输出接口与一部分基础处理电路的数据输入接口相连接；主处理电路的一部分数据输入接口与一部分基础处理电路的数据输出接口相连。

集成电路芯片装置的使用方法

本披露提供的使用方法中所涉及到的数据可以是经过压缩处理后的数据。关于如何实现数据的压缩处理具体可参见前述实施例中的相关阐述，例如图 1e~图 1k，这里不再赘述。

需要在基础处理电路中完成的运算，可以使用下述方法进行：

主处理电路的控制电路可将数据分发给基础处理电路运算。相应地，基础处理电路的压缩映射电路先对数据进行压缩处理后再运算，其优点是可以减少数据计算量，基础处理电路执行数据运算的效率也更高，功耗更低）

如基础处理电路接收到的数据为稀疏数据，那么基础处理电路可以收到数据后由压缩映射电路对数据进行压缩处理然后再进行计算，例如，基础处理电路收到主处理电路传输过来的稀疏数据，压缩映射电路将其进行压缩处理，然后基础处理电路的内积运算器电路、向量运算器电路或累加器电路对压缩处理后的数据进行运算，提高运算效率，降低功耗。

基础处理电路的使用方法（如图 5a）；

主处理电路从装置外部接收待计算的输入数据；

可选地，主处理电路利用本单元的各种运算电路，向量运算电路，内积运算器电路、累加器电路等对数据进行运算处理；

主处理电路通过数据输出接口向基础处理电路阵列（把所有基础处理电路的集合称作基础处理电路阵列）发送数据(如图 5b 所示)；

此处的发送数据的方式可以是向一部分基础处理电路直接发送数据，即多次广播方式；

此处发送数据的方式可以向不同的基础处理电路分别发送不同的数据，即分发方式；

基础处理电路阵列对数据进行计算；

基础处理电路接收到输入数据后进行运算；可选的，基础处理电路可根据该数据的运算指令确定是否启动所述基础处理电路中的压缩映射单元对数据进行压缩处理，然后对压缩处理后的数据进行运算。

可选地，基础处理电路接收到数据后将该数据从本单元的数据输出接口传输出去；（传输给其他没有直接从主处理电路接收到数据的基础处理电路，可选的，该数据也可为压缩处理后的数据。）

可选地，基础处理电路将运算结果从数据输出接口传输出去；（中间计算结果或者最终计算结果）

主处理电路接收到从基础处理电路阵列返回的输出数据；

可选地，主处理电路对从基础处理电路阵列接收到的数据继续进行处理（例如累加或激活操作）；

主处理电路处理完毕，将处理结果从数据输出接口传输给装置外部。

使用所述电路装置完成矩阵乘向量运算；

（矩阵乘向量可以是矩阵中的每一行分别与向量进行内积运算，并将这些结果按对应行的顺序摆放成一个向量。）

下面描述计算尺寸是 M 行 L 列的矩阵 S 和长度是 L 的向量 P 的乘法的运算，如下图 5c 所示。

此方法用到所述神经网络计算装置的全部或者一部分基础处理电路，假设用到了 K 个基础处理电路；

主处理电路将矩阵 S 的部分或全部行中的数据发送到 k 个基础处理电路中的每个基础处理电路；

在一种可选的方案中，主处理电路的控制电路将矩阵 S 中某行的数据每次发送一个数或者一部分数给某个基础处理电路；（例如，对于每次发送一个数，可以为对于某一个基础处理电路，第 1 次发送第 3 行第 1 个数，第 2 次发送第 3 行数据中的第 2 个数，第 3 次发送第 3 行的第 3 个数……，或者对于每次发送一部分数，第 1 次发送第 3 行前两个数（即第 1、2 个数），第二次发送第 3 行第 3 和第 4 个数，第三次发送第 3 行第 5 和第 6 个数……；）

在一种可选的方案中，主处理电路的控制电路将矩阵 S 中某几行的数据每次各发送一个数者一部分数给某个基础处理电路；（例如，对于某一个基础处理电路，第 1 次发送第 3,4,5 行每行的第 1 个数，第 2 次发送第 3,4,5 行每行的第 2 个数，第 3 次发送第 3,4,5 行每行的第 3 个数……，或者第 1 次发送第 3,4,5 行每行前两个数，第二次发送第 3,4,5 行每行第 3 和第 4 个数，第三次发送第 3,4,5 行每行第 5 和第 6 个数……。）

主处理电路的控制电路将向量 P 中的数据逐次发送到第 0 个基础处理电路；

第 0 个基础处理电路接收到向量 P 的数据之后，将该数据发送给与其相连接的下一个基础处理电路，即基础处理电路 1；

具体的，有些基础处理电路不能直接从主处理电路处获得计算所需的所有的数据，例如，图 2d 中的基础处理电路 1，只有一个数据输入接口与主处理电路相连，所以只能直接从主处理电路获得矩阵 S 的数据，而向量 P 的数据就需要依靠基础处理电路 0 输出给基础处理电路 1，同理，基础处理电路 1 也要收到数据后继续把向量 P 的数据输出给基础处理电路 2。

可选地，在所述 k 个基础处理电路中的每个基础处理电路接收到数据后，可先根据该数据的运算指令（即运算控制）确定是否启动对应的压缩映射电路来对该数据进行压缩处理，然后再利用压缩处理后的数据进行运算；可选的，还可将压缩处理后的数据传输给其他基础处理单元。

例如，基础处理电路在接收到输入矩阵 S 或矩阵 P 后，启用压缩映射电路将输入矩阵 S 和矩阵 P 中数据为指定数值（如 0）和/或数据小于预设阈值（如 0.1）所对应的数据剔除，具体实现时可根据矩阵 S 和矩阵 P 各自对应的 mask 矩阵来剔除，例如剔除 mask 矩阵为 0

所对应的相同位置在矩阵 S/P 中的数据，具体可参见前述关于数据压缩处理实施例中的相关阐述，这里不再赘述。应理解的，这里的矩阵 S 和矩阵 P 也可对应理解为前述实施例中的输入神经元（也可称为输入神经元矩阵）和权值（也可称为权值矩阵）等。

每一个基础处理电路对接收到的数据进行运算，该运算包括但不限于：内积运算、乘法运算、加法运算等等；

在一种可选方案中，基础处理电路每次计算一组或多组两个数据的乘法，然后将结果累加到寄存器和或片上缓存上；

在一种可选方案中，基础处理电路每次计算一组或多组两个向量的内积，然后将结果累加到寄存器和或片上缓存上；

基础处理电路计算出结果后，将结果从数据输出接口传输出去（即传输给与其连接的其他基础处理电路）；

在一种可选方案中，该计算结果可以是内积运算的最终结果或中间结果；

基础处理电路接收到来自其他基础处理电路的计算结果之后，将该数据传输给与其相连接的其他基础处理电路或者主处理电路；

主处理电路接收到各个基础处理电路内积运算的结果，将该结果处理得到最终结果（该处理可以为累加运算或激活运算等等）。

采用上述计算装置实现矩阵乘向量方法的实施例：

在一种可选方案中，方法所用到的多个基础处理电路按照如下图 5d 或者图 5e 所示的方式排列；

如图 4c 所示，主处理单元的控制电路将矩阵 S 的 M 行数据分成 K 组，分别由第 i 个基础处理电路负责第 i 组（该组数据中行的集合记为 Ai）的运算；具体的，第 i 个基础处理电路在负责第 i 组（该组数据中行的集合记为 Ai）的运算之前，可根据数据的运算指令确定是否需要先利用压缩映射电路对 Ai 进行压缩处理，然后再对压缩处理后的 Ai 执行运算。或者，针对装置中第 1 列或第 1 行中的各个基础处理单元在负责第 i 组（该组数据中行的集合记为 Ai）的运算之前，可根据数据的运算指令确定是否需要先利用压缩映射电路对 Ai 进行压缩处理，然后再对压缩处理后的 Ai 执行运算，本申请不做限定。关于数据压缩处理具体可参见前述实施例中的相关阐述，这里不再赘述。

此处对 M 行数据进行分组的方法是任意不会重复分配的分组方式；

在一种可选方案中，采用如下分配方式：将第 j 行分给第 j%K（%为取余数运算）个基础处理电路；

在一种可选方案中，对于不能平均分组的情况也可以先对一部分行平均分配，对于剩下的行以任意方式分配。

主处理电路的控制电路每次将矩阵 S 中部分或全部行中的数据依次发送给对应的基础处理电路；

在一种可选方案中，主处理电路的控制电路每次向第 i 个基础处理电路发送其负责的第 i 组数据 Mi 中的一行数据中的一个或多个数据；

在一种可选方案中，主处理电路的控制电路每次向第 i 个基础处理电路发送其负责的第 i 组数据 Mi 中的部分或全部行中的每行的一个或多个数据；

主处理电路的控制电路将向量 P 中的数据依次向第 1 个基础处理电路发送；

在一种可选方案中，主处理电路的的控制电路每次可以发送向量 P 中的一个或多个数据；

第 i 个基础处理电路接收到向量 P 的数据之后发送给与其相连的第 i+1 个基础处理电路；可选的，发送的向量 P 的数据可为压缩处理后的数据。

每个基础处理电路接收到来自矩阵 S 中某一行或者某几行中的一个或多个数据以及来自向量 P 的一个或多个数据后，进行运算（包括但不限于乘法或加法）；

在一种可选方案中，基础处理电路每次计算一组或多组两个数据的乘法，然后将结果累加到寄存器和或片上缓存上；

在一种可选方案中，基础处理电路每次计算一组或多组两个向量的内积，然后将结果累加到寄存器和或片上缓存上；

在一种可选方案中，基础处理电路接收到的数据也可以是中间结果，保存在寄存器和或片上缓存上；

基础处理电路将本地的计算结果传输给与其相连接的下一个基础处理电路或者主处理电路；

在一种可选方案中，对应于图 5d 的结构，只有每列的最后一个基础处理电路的输出接口与主处理电路相连接的，这种情况下，只有最后一个基础处理电路可以直接将本地的计算结果传输给主处理电路，其他基础处理电路的计算结果都要传递给自己的下一个基础处理电路，下一个基础处理电路传递给下下个基础处理电路直至全部传输给最后一个基础处理电路，最后一个基础处理电路将本地的计算结果以及接收到的本列的其他基础处理电路的结果执行累加计算得到中间结果，将中间结果发送至主处理电路；当然还可以是：最后一个基础处理电路将本列的其他基础电路的结果以及本地的处理结果直接发送给主处理电路。

在一种可选方案中，对应于图 5e 的结构，每一个基础处理电路都有与主处理电路相连接的输出接口，这种情况下，每一个基础处理电路都直接将本地的计算结果传输给主处理电路；

基础处理电路接收到其他基础处理电路传递过来的计算结果之后，传输给与其相连接的下一个基础处理电路或者主处理电路。

主处理电路接收到 M 个内积运算的结果，作为矩阵乘向量的运算结果。

使用所述电路装置完成矩阵乘矩阵运算；

下面描述计算尺寸是 M 行 L 列的矩阵 S 和尺寸是 L 行 N 列的矩阵 P 的乘法的运算，（矩阵 S 中的每一行与矩阵 P 的每一列长度相同，如图 5f 所示）

本方法使用所述装置如图 4b 所示的实施例进行说明；

主处理电路的控制电路将矩阵 S 的部分或全部行中的数据发送到通过横向数据输入接口直接与主处理电路相连的那些基础处理电路（例如，图 4b 中最上方的灰色填充的竖向数据通路）；

在一种可选方案中，主处理电路的控制电路将矩阵 S 中某行的数据每次发送一个数或者一部分数给某个基础处理电路；（例如，对于某一个基础处理电路，第 1 次发送第 3 行第

1 个数，第 2 次发送第 3 行数据中的第 2 个数，第 3 次发送第 3 行的第 3 个数……，或者第 1 次发送第 3 行前两个数，第二次发送第 3 行第 3 和第 4 个数，第三次发送第 3 行第 5 和第 6 个数……；）

在一种可选方案中，主处理电路的控制电路将矩阵 S 中某几行的数据每次各发送一个数或者一部分数给某个基础处理电路；（例如，对于某一个基础处理电路，第 1 次发送第 3,4,5 行每行的第 1 个数，第 2 次发送第 3,4,5 行每行的第 2 个数，第 3 次发送第 3,4,5 行每行的第 3 个数……，或者第 1 次发送第 3,4,5 行每行前两个数，第二次发送第 3,4,5 行每行第 3 和第 4 个数，第三次发送第 3,4,5 行每行第 5 和第 6 个数……；）

主处理电路的控制电路将矩阵 P 中的部分或全部列中的数据发送到通过竖向数据输入接口直接与主处理电路相连的那些基础处理电路（例如，图 4b 中基础处理电路阵列左侧的灰色填充的横向数据通路）；

在一种可选方案中，主处理电路的控制电路将矩阵 P 中某列的数据每次发送一个数或者一部分数给某个基础处理电路；（例如，对于某一个基础处理电路，第 1 次发送第 3 列第 1 个数，第 2 次发送第 3 列数据中的第 2 个数，第 3 次发送第 3 列的第 3 个数……，或者第 1 次发送第 3 列前两个数，第二次发送第 3 列第 3 和第 4 个数，第三次发送第 3 列第 5 和第 6 个数……；）

在一种可选方案中，主处理电路的控制电路将矩阵 P 中某几列的数据每次各发送一个数或者一部分数给某个基础处理电路；（例如，对于某一个基础处理电路，第 1 次发送第 3,4,5 列每列的第 1 个数，第 2 次发送第 3,4,5 列每列的第 2 个数，第 3 次发送第 3,4,5 列每列的第 3 个数……，或者第 1 次发送第 3,4,5 列每列前两个数，第二次发送第 3,4,5 列每列第 3 和第 4 个数，第三次发送第 3,4,5 列每列第 5 和第 6 个数……；）

基础处理电路接收到矩阵 S 的数据之后，将该数据通过其横向的数据输出接口传输给其相连接下一个基础处理电路（例如，图 4b 中基础处理电路阵列中间的白色填充的横向的数据通路）；基础处理电路接收到矩阵 P 的数据后，将该数据通过其竖向的数据输出接口传输给与其相连接的下一个基础处理电路（例如，图 4b 中基础处理电路阵列中间的白色填充的竖向的数据通路）；

可选地，在每个基础处理电路均包括压缩映射电路时，基础处理电路在接收到数据（具体可为矩阵 S 或矩阵 P 的数据）后，可根据数据的运算控制确定启动压缩映射电路对数据进行压缩处理；进一步地，可将压缩处理后的数据通过其横向或竖向的数据输出接口传输给其相连接下一个基础处理电路；

例如，基础处理电路在接收到输入矩阵 S 或矩阵 P 后，启用压缩映射电路将输入矩阵 S 和矩阵 P 中数据为指定数值（如 0）和/或数据小于预设阈值（如 0.1）所对应的数据剔除，具体实现时可根据矩阵 S 和矩阵 P 各自对应的 mask 矩阵来剔除，例如剔除 mask 矩阵为 0 所对应的相同位置在矩阵 S/P 中的数据，具体可参见前述关于数据压缩处理实施例中的相关阐述，这里不再赘述。应理解的，这里的矩阵 S 和矩阵 P 也可对应理解为前述实施例中的输入神经元（也可称为输入神经元矩阵）和权值（也可称为权值矩阵）等。

可选的，在第 1 列和第 1 行中的各个基础处理电路中均包括压缩映射电路时，针对装置中第 1 列或第 1 行中的各个基础处理电路在接收到数据（具体可为矩阵 S 或矩阵 P 的数

据）后，可根据该数据对应的运算控制确定是否需要启用各自基础处理电路中的压缩映射电路对数据进行压缩处理；进一步地，可将压缩处理后的数据通过其横向或竖向的数据输出接口传输给其相连接下一个基础处理电路；可选的，针对装置中第 1 列或第 1 行中的各个基础处理电路在接收到数据后可直接启动其内的压缩映射电路对数据进行压缩处理，接着进行后续操作，例如发送给其他基础处理电路或对其进行运算等。

每一个基础处理电路对接收到的数据进行运算，可选的，该接收到的数据可为压缩处理后的数据。

在一种可选方案中，基础处理电路每次计算一组或多组两个数据的乘法，然后将结果累加到寄存器和或片上缓存上；

在一种可选方案中，基础处理电路每次计算一组或多组两个向量的内积，然后将结果累加到寄存器和或片上缓存上；

基础处理电路计算出结果后，可以将结果从数据输出接口传输出去；

在一种可选方案中，该计算结果可以是内积运算的最终结果或中间结果；

具体地，如果该基础处理电路有直接与主处理电路相连接的输出接口则从该接口传输结果，如果没有，则向着能够直接向主处理电路输出的基础处理电路的方向输出结果（例如，图 4b 中，最下面一行基础处理电路将其输出结果直接输出给主处理电路，其他基础处理电路从竖向的输出接口向下传输运算结果）。

基础处理电路接收到来自其他基础处理电路的计算结果之后，将该数据传输给与其相连接的其他基础处理电路或者主处理电路；

向着能够直接向主处理电路输出的方向输出结果（例如，图 4b 中，最下面一行基础处理电路将其输出结果直接输出给主处理电路，其他基础处理电路从竖向的输出接口向下传输运算结果）；

主处理电路接收到各个基础处理电路内积运算的结果，即可得到输出结果。

“矩阵乘矩阵”方法的实施例：

方法用到按照如图 4b 所示方式排列的基础处理电路阵列，假设有 h 行，w 列；

主处理电路的控制电路将矩阵 S 的 h 行数据分成 h 组，分别由第 i 个基础处理电路负责第 i 组（该组数据中行的集合记为 Hi）的运算；

此处对 h 行数据进行分组的方法是任意不会重复分配的分组方式；

在一种可选方案中，采用如下分配方式：主处理电路的控制电路将第 j 行分给第 j%h 个基础处理电路；

在一种可选方案中，对于不能平均分组的情况也可以先对一部分行平均分配，对于剩下的行以任意方式分配。

主处理电路的控制电路将矩阵 P 的 W 列数据分成 w 组，分别由第 i 个基础处理电路负责第 i 组（该组数据中行的集合记为 Wi）的运算；

此处对 W 列数据进行分组的方法是任意不会重复分配的分组方式；

在一种可选方案中，采用如下分配方式：主处理电路的控制电路将第 j 行分给第 j%w 个基础处理电路；

在一种可选方案中，对于不能平均分组的情况也可以先对一部分列平均分配，对于剩

下的列以任意方式分配。

主处理电路的控制电路将矩阵 S 的部分或全部行中的数据发送到基础处理电路阵列中每行的第一个基础处理电路；

在一种可选方案中，主处理电路的控制电路每次向基础处理电路阵列中第 i 行的第一个基础处理电路发送其负责的第 i 组数据 Hi 中的一行数据中的一个或多个数据；

在一种可选方案中，主处理电路的控制电路每次向基础处理电路阵列中第 i 行的第一个基础处理电路发送其负责的第 i 组数据 Hi 中的部分或全部行中的每行的一个或多个数据；

主处理电路的控制电路将矩阵 P 的部分或全部列中的数据发送到基础处理电路阵列中每列的第一个基础处理电路；

在一种可选方案中，主处理电路的控制电路每次向基础处理电路阵列中第 i 列的第一个基础处理电路发送其负责的第 i 组数据 Wi 中的一列数据中的一个或多个数据；

在一种可选方案中，主处理电路的控制电路每次向基础处理电路阵列中第 i 列的第一个基础处理电路发送其负责的第 i 组数据 Ni 中的部分或全部列中的每列的一个或多个数据；

基础处理电路接收到矩阵 S 的数据之后，将该数据通过其横向的数据输出接口传输给其相连接下一个基础处理电路（例如，图 4b 中基础处理电路阵列中间的白色填充的横向的数据通路）；基础处理电路接收到矩阵 P 的数据后，将该数据通过其竖向的数据输出接口传输给与其相连接的下一个基础处理电路（例如，图 4b 中基础处理电路阵列中间的白色填充的竖向的数据通路）；

可选地，在每个基础处理电路均包括压缩映射电路时，基础处理电路在接收到数据（具体可为矩阵 S 或矩阵 P 的数据）后，可根据数据的运算控制确定启动压缩映射电路对数据进行压缩处理；进一步地，可将压缩处理后的数据通过其横向或竖向的数据输出接口传输给其相连接下一个基础处理电路；关于数据的压缩处理可参见前述实施例中的相关阐述，这里不再赘述。

可选的，在第 1 列和第 1 行中的各个基础处理电路中均包括压缩映射电路时，针对装置中第 1 列或第 1 行中的各个基础处理单元在接收到数据（具体可为矩阵 S 或矩阵 P 的数据）后，可对数据进行压缩处理；进一步地，可将压缩处理后的数据通过其横向或竖向的数据输出接口传输给其相连接下一个基础处理电路。具体可参见前述实施例中的相关阐述。

每一个基础处理电路对接收到的数据进行运算，可选的，该接收到的数据可为压缩处理后的数据；

在一种可选方案中，基础处理电路每次计算一组或多组两个数据的乘法，然后将结果累加到寄存器和或片上缓存上；

在一种可选方案中，基础处理电路每次计算一组或多组两个向量的内积，然后将结果累加到寄存器和或片上缓存上；

基础处理电路计算出结果后，可以将结果从数据输出接口传输出去；

在一种可选方案中，该计算结果可以是内积运算的最终结果或中间结果；

具体地，如果该基础处理电路有直接与主处理电路相连接的输出接口则从该接口传输

结果，如果没有，则向着能够直接向主处理电路输出的基础处理电路的方向输出结果（例如，最下面一行基础处理电路将其输出结果直接输出给主处理电路，其他基础处理电路从竖向的输出接口向下传输运算结果）。

基础处理电路接收到来自其他基础处理电路的计算结果之后，将该数据传输给与其相连接的其他基础处理电路或者主处理电路；

向着能够直接向主处理电路输出的方向输出结果（例如，最下面一行基础处理电路将其输出结果直接输出给主处理电路，其他基础处理电路从竖向的输出接口向下传输运算结果）；

主处理电路接收到各个基础处理电路内积运算的结果，即可得到输出结果。

以上描述中使用的“横向”，“竖向”等词语只是为了表述图 4b 所示的例子，实际使用只需要区分出每个单元的“横向”“竖向”接口代表两个不同的接口即可。

使用所述电路装置完成全连接运算：

如果全连接层的输入数据是一个向量（即神经网络的输入是单个样本的情况），则以全连接层的权值矩阵作为矩阵 S，输入向量作为向量 P，按照所述装置采用矩阵乘以向量方法执行运算；

如果全连接层的输入数据是一个矩阵（即神经网络的输入是多个样本的情况），则以全连接层的权值矩阵作为矩阵 S，输入向量作为矩阵 P，或者以全连接层的权值矩阵作为矩阵 P，输入向量作为矩阵 S，按照所述装置的矩阵乘以矩阵执行运算；

使用所述电路装置完成卷积运算：

下面描述卷积运算，下面的图中一个方块表示一个数据，输入数据用图 6a 表示（N 个样本，每个样本有 C 个通道，每个通道的特征图的高为 H，宽为 W），权值也即卷积核用图 6b 表示（有 M 个卷积核，每个卷积核有 C 个通道，高和宽分别为 KH 和 KW）。对于输入数据的 N 个样本，卷积运算的规则都是一样的，下面解释在一个样本上进行卷积运算的过程，在一个样本上，M 个卷积核中的每一个都要进行同样的运算，每个卷积核运算得到一张平面特征图，M 个卷积核最终计算得到 M 个平面特征图，（对一个样本，卷积的输出是 M 个特征图），对于一个卷积核，要在一个样本的每一个平面位置进行内积运算，然后沿着 H 和 W 方向进行滑动，例如，图 6c 表示一个卷积核在输入数据的一个样本中右下角的位置进行内积运算的对应图；图 6d 表示卷积的位置向左滑动一格和图 6e 表示卷积的位置向上滑动一格。

本方法使用所述装置如图 4b 所示的实施例进行说明；

主处理电路的控制电路将权值的部分或全部卷积核中的数据发送到通过横向数据输入接口直接与主处理电路相连的那些基础处理电路（例如，图 4b 中最上方的灰色填充的竖向数据通路）；

在一种可选方案中，主处理电路的控制电路将权值中某个卷积核的数据每次发送一个数或者一部分数给某个基础处理电路；（例如，对于某一个基础处理电路，第 1 次发送第 3 行第 1 个数，第 2 次发送第 3 行数据中的第 2 个数，第 3 次发送第 3 行的第 3 个数……，或者第 1 次发送第 3 行前两个数，第二次发送第 3 行第 3 和第 4 个数，第三次发送第 3 行第 5 和第 6 个数……；）

在一种可选方案中另一种情况是，主处理电路的控制电路将权值中某几个卷积核的数据每次各发送一个数或者一部分数给某个基础处理电路；(例如，对于某一个基础处理电路，第 1 次发送第 3,4,5 行每行的第 1 个数，第 2 次发送第 3,4,5 行每行的第 2 个数，第 3 次发送第 3,4,5 行每行的第 3 个数……，或者第 1 次发送第 3,4,5 行每行前两个数，第二次发送第 3,4,5 行每行第 3 和第 4 个数，第三次发送第 3,4,5 行每行第 5 和第 6 个数……；)

主处理电路的控制电路把输入数据按照卷积的位置进行划分，主处理电路的控制电路将输入数据中的部分或全部卷积位置中的数据发送到通过竖向数据输入接口直接与主处理电路相连的那些基础处理电路（例如，图 4b 中基础处理电路阵列左侧的灰色填充的横向数据通路）；

在一种可选方案中，主处理电路的控制电路将输入数据中某个卷积位置的数据每次发送一个数或者一部分数给某个基础处理电路；(例如，对于某一个基础处理电路，第 1 次发送第 3 列第 1 个数，第 2 次发送第 3 列数据中的第 2 个数，第 3 次发送第 3 列的第 3 个数……，或者第 1 次发送第 3 列前两个数，第二次发送第 3 列第 3 和第 4 个数，第三次发送第 3 列第 5 和第 6 个数……；)

在一种可选方案中另一种情况是，主处理电路的控制电路将输入数据中某几个卷积位置的数据每次各发送一个数或者一部分数给某个基础处理电路；(例如，对于某一个基础处理电路，第 1 次发送第 3,4,5 列每列的第 1 个数，第 2 次发送第 3,4,5 列每列的第 2 个数，第 3 次发送第 3,4,5 列每列的第 3 个数……，或者第 1 次发送第 3,4,5 列每列前两个数，第二次发送第 3,4,5 列每列第 3 和第 4 个数，第三次发送第 3,4,5 列每列第 5 和第 6 个数……；)

基础处理电路接收到权值的数据之后，将该数据通过其横向的数据输出接口传输给其相连接下一个基础处理电路（例如，图 4b 中基础处理电路阵列中间的白色填充的横向的数据通路）；基础处理电路接收到输入数据的数据后，将该数据通过其竖向的数据输出接口传输给与其相连接的下一个基础处理电路（例如，图 4b 中基础处理电路阵列中间的白色填充的竖向的数据通路）；可选的，基础处理电路在接收到数据（具体可为权值的部分或者全部卷积核中的数据）后，可根据数据的运算控制确定启动压缩映射电路对数据进行压缩处理；进一步地，可将压缩处理后的数据通过其横向或竖向的数据输出接口传输给其相连接下一个基础处理电路；具体可参见前述实施例中的相关阐述。

或者，针对装置中第 1 列或第 1 行中的各个基础处理单元在接收到数据（具体可为权值的部分或者全部卷积核中的数据）后，可对数据进行压缩处理；进一步地，可将压缩处理后的数据通过其横向或竖向的数据输出接口传输给其相连接下一个基础处理电路；具体可参见前述实施例中的相关阐述。

每一个基础处理电路对接收到的数据进行运算，该接收到的数据可为压缩处理后的数据；

在一种可选方案中，基础处理电路每次计算一组或多组两个数据的乘法，然后将结果累加到寄存器和/或片上缓存上；

在一种可选方案中，基础处理电路每次计算一组或多组两个向量的内积，然后将结果累加到寄存器和/或片上缓存上；

基础处理电路计算出结果后，可以将结果从数据输出接口传输出去；

在一种可选方案中，该计算结果可以是内积运算的最终结果或中间结果；具体地，如果该基础处理电路有直接与主处理电路相连接的输出接口则从该接口传输结果，如果没有，则向着能够直接向主处理电路输出的基础处理电路的方向输出结果（例如，图 4b 中，最下面一行基础处理电路将其输出结果直接输出给主处理电路，其他基础处理电路从竖向的输出接口向下传输运算结果）。

基础处理电路接收到来自其他基础处理电路的计算结果之后，将该数据传输给与其相连接的其他基础处理电路或者主处理电路；

向着能够直接向主处理电路输出的方向输出结果（例如，最下面一行基础处理电路将其输出结果直接输出给主处理电路，其他基础处理电路从竖向的输出接口向下传输运算结果）；

主处理电路接收到各个基础处理电路内积运算的结果，即可得到输出结果。

在一个实施例中，本申请公开了一种神经网络运算装置，其包括用于执行如上所述方法实施例中提供的所有或部分实施方式所对应的功能单元。

在一个实施例里，本申请公开了一种芯片（如图 7），用于执行如上所述方法实施例中提供的所有或部分实施方式。

在一个实施例里，本申请公开了一种电子装置，其包括用于执行如上所述方法实施例中的所有或部分实施方式的功能单元。

电子装置包括数据处理装置、机器人、电脑、打印机、扫描仪、平板电脑、智能终端、手机、行车记录仪、导航仪、传感器、摄像头、服务器、相机、摄像机、投影仪、手表、耳机、移动存储、可穿戴设备、交通工具、家用电器、和/或医疗设备。

所述交通工具包括飞机、轮船和/或车辆；所述家用电器包括电视、空调、微波炉、冰箱、电饭煲、加湿器、洗衣机、电灯、燃气灶、油烟机；所述医疗设备包括核磁共振仪、B 超仪和/或心电图仪。

权 利 要 求

1、一种集成电路芯片装置，其特征在于，所述集成电路芯片装置包括：主处理电路以及 k 组基础处理电路，所述主处理电路与所述 k 组基础处理电路连接，所述一组基础处理电路包括至少一个基础处理电路；

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及向所述多个基础处理电路传输数据；

所述 k 组基础处理电路，用于依据所述传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

2、根据权利要求 1 所述的集成电路芯片装置，其特征在于，

所述集成电路芯片装置还包括：k 个分支电路，所述主处理电路与所述 k 个分支电路分别连接，所述 k 个分支电路中每个分支电路对应 k 组基础处理电路中的一组基础处理电路，用于在所述主处理电路与所述 k 组基础处理电路之间转发传输数据。

3、根据权利要求 1 所述的集成电路芯片装置，其特征在于，

所述基础处理电路包括压缩映射电路；所述压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；

所述 k 组基础处理电路，具体用于依据所述传输数据的运算控制是否启动所述压缩映射电路对所述传输数据进行压缩处理；依据所述传输数据或压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果传输给所述主处理电路。

4、根据权利要求 3 所述的集成电路芯片装置，其特征在于，

所述主处理电路，用于获取待计算的数据块以及运算指令，依据该运算指令对所述待计算的数据块划分成分发数据块以及广播数据块；对所述分发数据块进行拆分处理得到多个基本数据块，将所述多个基本数据块分发至与其连接的电路，将所述广播数据块广播至与其连接的电路；

所述基础处理电路，用于依据所述运算控制启动所述压缩映射电路对所述基本数据块与所述广播数据块进行压缩处理后再执行内积运算得到运算结果，将所述运算结果发送至主处理电路；

所述主处理电路，用于对所述运算结果处理得到所述待计算的数据块以及运算指令的指令结果；

其中，所述待计算的数据块为待计算的至少一个输入神经元，和/或，至少一个权值。

5、根据权利要求 2 所述的集成电路芯片装置，其特征在于，

所述分支电路包括：压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及和与其相连的所述 k 个分支电路传输数据；

所述 k 个分支电路，用于在主处理电路与 k 组基础电路之间转发所述传输数据，依据所述传输数据的运算控制是否启动所述压缩映射电路对所述传输数据进行压缩处理；

所述 k 个基础处理电路，用于依据所述传输数据或压缩处理后的传输数据以并行方式

执行神经网络中的运算，并将运算结果传输给所述主处理电路。

6、根据权利要求 5 所述的集成电路芯片装置，其特征在于，

所述主处理电路，用于获取待计算的数据块以及运算指令，依据该运算指令对所述待计算的数据块划分成分发数据块以及广播数据块；对所述分发数据块进行拆分处理得到多个基本数据块，将所述多个基本数据块分发至与其连接的所述 k 个分支电路，将所述广播数据块广播至与其连接的所述 k 个分支电路；

所述 k 个分支电路，用于接收基本数据块以及广播数据块，启动压缩映射电路将该基本数据块以及广播数据块进行压缩处理；将压缩处理后的基本数据块以及压缩处理后的广播数据块转发至 k 组基础处理电路；

所述基础处理电路，用于对所述压缩处理后的基本数据块与所述压缩处理后的广播数据块执行内积运算得到运算结果，将所述运算结果发送至所述主处理电路；

所述主处理电路，用于对所述运算结果处理得到所述待计算的数据块以及运算指令的指令结果；

其中，所述分发数据块以及所述广播数据块为至少一个输入神经元或者，至少一个权值。

7、根据权利要求 4 或 6 所述的集成电路芯片装置，其特征在于，所述压缩映射电路包括第二稀疏处理单元、第三稀疏处理单元以及连接关系处理单元；

所述第二稀疏处理单元，用于接收到第三输入数据后，根据所述第三输入数据得到第一连接关系数据，并将该第一关系数据传输至连接关系处理单元；

所述第三稀疏处理单元，用于接收到第四输入数据后，根据所述第四输入数据得到第二连接关系数据，并将该第二关系数据传输至所述连接关系处理单元；

所述连接关系处理单元，用于根据所述第一连接关系数据和所述第二连接关系数据，以得到第三连接关系数据，并将该第三连接关系数据传输至第二数据处理单元；

所述第二数据处理单元，用于在接收到所述第三输入数据，所述第四输入数据和所述第三连接关系数据后，根据所述第三连接关系数据对所述第三输入数据和所述第四输入数据进行压缩处理，以得到第四输出数据和第五输出数据；

其中，当所述第三输入数据包括至少一个输入神经元，第四输入数据包括至少一个权值时，所述第一连接关系数据为输入神经元的连接关系数据，所述第二连接关系数据为权值的连接关系数据，所述第四输出数据为处理后的输入神经元，所述第五输出数据为处理后的权值；当所述第三输入数据包括至少一个权值，所述第四输入数据包括至少一个输入神经元时，所述第一连接关系数据为权值的连接关系数据，所述第二连接关系数据为输入神经元的连接关系数据，所述第四输出数据为处理后的权值，所述第五输出数据为处理后的输入神经元。

8、根据权利要求 7 所述的集成电路芯片装置，其特征在于，所述神经元的连接关系数据以及所述权值的连接关系数据均为由 0 和 1 组成的字符串或矩阵组成，且与输出神经元无关；或者，

所述输入神经元的连接关系数据和所述权值的连接关系数据均以直接索引或者步长索引的形式表示；

其中，当所述输入神经元的连接关系数据以直接索引的形式表示时，该连接关系数据为由 0 和 1 组成的字符串，0 表示所述输入神经元的绝对值小于或者等于第一阈值，1 表示所述输入神经元的绝对值大于所述第一阈值；

当所述输入神经元的连接关系数据以步长索引形式表示时，该连接关系数据为绝对值大于所述第一阈值的输入神经元与上一个绝对值大于所述第一阈值的输入神经元之间的距离值组成的字符串；

当所述权值的连接关系数据以直接索引的形式表示时，该连接关系数据为由 0 和 1 组成的字符串，0 表示所述权值的绝对值小于或者等于第二阈值，即该权值对应的输入神经元与输出神经元之间没有连接，1 表示所述权值的绝对值大于所述第二阈值，即该权值对应的输入神经元与输出神经元之间有连接；以直接索引形式表示权值的连接关系数据有两种表示顺序：以每个输出神经元与所有输入神经元的连接状态组成一个 0 和 1 的字符串来表示所述权值的连接关系数据；或者每个输入神经元与所有输出神经元的连接状态组成一个 0 和 1 的字符串来表示所述权值的连接关系数据；

当所述权值的连接关系数据以步长索引的形式表示时，该连接关系数据为与输出神经元有连接的输入神经元与上一个与该输出神经元有连接的输入神经元之间的距离值组成的字符串。

9、根据权利要求 8 所述的集成电路芯片装置，其特征在于，当所述第一连接关系数据和所述第二连接关系数据均以步长索引的形式表示，且表示所述第一连接关系数据和所述第二连接关系数据的字符串是按照物理地址由低到高的顺序存储时，所述连接关系处理单元具体用于：

将所述第一连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第四连接关系数据；同理，对所述第二连接关系数据的字符串进行同样的处理，得到第五连接关系数据；

从所述第四连接关系数据的字符串和所述第五连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成新的字符串；

将所述新的字符串中每一个元素与其相邻的且值小于该元素值的元素进行相减，得到的元素组成所述第三连接关系数据。

10、根据权利要求 8 所述的集成电路芯片装置，其特征在于，当所述第一连接关系数据和所述第二连接关系数据均以直接索引的形式表示时，所述连接关系处理单元具体用于：

对所述第一连接关系数据和所述第二连接关系数据进行与操作，以得到第三连接关系数据。

11、根据权利要求 8 所述的集成电路芯片装置，其特征在于，当所述第一连接关系数据与所述第二连接关系数据中任意一个以步长索引的形式表示，另一个以直接索引的形式表示时，所述连接关系处理单元具体用于：

若所述第一关系数据是以步长索引的形式表示，将所述第一连接关系数据转换成以直接索引的形式表示的连接关系数据；

若所述第二关系数据是以步长索引的形式表示，将所述第二连接关系数据转换成以直接索引的形式表示的连接关系数据；

对所述第一连接关系数据和所述第二连接关系数据进行与操作，以得到第三连接关系数据。

12、根据权利要求 8 所述的集成电路芯片装置，其特征在于，当所述第一连接关系数据与所述第二连接关系数据中任意一个以步长索引的形式表示，另一个以直接索引的形式表示，且表示所述第一连接关系数据和所述第二连接关系数据的字符串是按照物理地址由低到高的顺序存储时，所述连接关系处理单元还具体用于：

若所述第一关系数据是以步长索引的形式表示，将所述第二连接关系数据转换成以步长索引的形式表示的连接关系数据；

若所述第二关系数据是以步长索引的形式表示，将所述第一连接关系数据转换成以步长索引的形式表示的连接关系数据；

将所述第一连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第四连接关系数据；同理，对所述第二连接关系数据的字符串进行同样的处理，得到第五连接关系数据；

从所述第四连接关系数据的字符串和所述第五连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成新的字符串；

将所述新的字符串中每一个元素与其相邻的且值小于该元素值的元素进行相减，得到的元素组成所述第三连接关系数据。

13、根据权利要求 6 所述的集成电路芯片装置，其特征在于，所述启动压缩映射电路将该基本数据块以及广播数据块进行压缩处理之前，还包括：

所述 K 个分支电路，还用于通过所述压缩映射电路对所述至少一个输入神经元进行分组，以得到 M 组输入神经元，所述 M 为大于或者等于 1 的整数；判断所述 M 组输入神经元的每一组输入神经元是否满足第一预设条件，所述第一预设条件包括一组输入神经元中绝对值小于或者等于第三阈值的输入神经元的个数小于或者等于第四阈值；当所述 M 组输入神经元任意一组输入神经元不满足所述第一预设条件时，将该组输入神经元删除；对所述至少一个权值进行分组，以得到 N 组权值，所述 N 为大于或者等于 1 的整数；判断所述 N 组权值的每一组权值是否满足第二预设条件，所述第二预设条件包括一组权值中绝对值小于或者等于第五阈值的权值的个数小于或者等于第六阈值；当所述 N 组权值任意一组权值不满足所述第二预设条件时，将该组权值删除。

14、根据权利要求 6 所述的集成电路芯片装置，其特征在于，

所述主处理电路，具体用于将所述广播数据块通过一次广播至所述 k 个分支电路；或者，

所述主处理电路，具体用于将所述广播数据块分成多个部分广播数据块，将所述多个部分广播数据块通过多次广播至所述 K 个分支电路。

15、根据权利要求 14 所述的集成电路芯片装置，其特征在于，

所述基础处理电路，具体用于将压缩处理后的所述部分广播数据块与压缩处理后的所述基本数据块执行一次内积处理后得到内积处理结果，将所述内积处理结果累加得到部分运算结果，将所述部分运算结果发送至所述主处理电路。

16、根据权利要求 15 所述的集成电路芯片装置，其特征在于，

所述基础处理电路，具体用于复用 n 次该部分广播数据块执行该部分广播数据块与该 n 个基本数据块内积运算得到 n 个部分处理结果，将 n 个部分处理结果分别累加后得到 n 个部分运算结果，将所述 n 个部分运算结果发送至主处理电路，所述 n 为大于等于 2 的整数。

17、一种集成电路芯片装置，其特征在于，所述集成电路芯片装置包括：主处理电路以及多个基础处理电路；

所述多个基础处理电路呈阵列分布；每个基础处理电路与相邻的其他基础处理电路连接，所述主处理电路连接所述多个基础处理电路中的 k 个基础处理电路，所述 k 个基础电路为：第 1 行的 n 个基础处理电路以及第 1 列的 m 个基础处理电路；

所述多个基础处理电路中的部分或所有基础处理电路包括：压缩映射电路，用于执行神经网络运算中的各个数据的压缩处理；

所述主处理电路，用于执行神经网络运算中的各个连续的运算以及和与所述 k 个基础处理电路传输数据；

所述 k 个基础处理电路，用于在所述主处理电路以及多个基础处理电路之间的数据转发；

所述部分或所有基础处理电路，用于依据传输数据的运算控制确定是否启动所述压缩映射电路对所述传输数据进行压缩处理，依据压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果通过与所述主处理电路连接的基础处理电路传输给所述主处理电路。

18、根据权利要求 1 所述的集成电路芯片装置，其特征在于，在所述多个基础处理电路均包括压缩映射电路时，

所述多个基础处理电路，用于依据传输数据的运算控制确定是否启动所述压缩映射电路对所述传输数据进行压缩处理，依据压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果通过与所述 k 个基础处理电路传输给所述主处理电路。

19、根据权利要求 18 所述的集成电路芯片装置，其特征在于，

所述主处理电路，用于获取待计算的数据块以及运算指令，依据该运算指令对所述待计算的数据块划分成分发数据块以及广播数据块；对所述分发数据块进行拆分处理得到多个基本数据块，将所述多个基本数据块分发至所述 k 个基础处理电路，将所述广播数据块广播至所述 k 个基础处理电路；

所述多个基础处理电路，用于依据接收到的基础数据块、广播数据块以及运算指令启动所述压缩映射电路将基础数据块和广播数据块进行压缩处理，对压缩处理后的所述基本数据块与压缩处理后的所述广播数据块执行内积运算得到运算结果，将运算结果通过所述 k 个基础处理电路传输给所述主处理电路；

所述主处理电路，用于对所述运算结果处理得到所述待计算的数据块以及运算指令的指令结果；

其中，所述分发数据块以及所述广播数据块为至少一个输入神经元或者，至少一个权值。

20、根据权利要求 1 所述的集成电路芯片装置，其特征在于，

在所述多个基础处理电路中的所述 k 个基础处理电路均包括压缩映射电路时，

所述 k 个基础处理电路，用于依据传输数据的运算控制确定是否启动所述压缩映射电路对所述传输数据进行压缩处理，并将压缩处理后的传输数据发送给与所述 k 个基础处理电路连接的基础处理电路；

所述多个基础处理电路，用于依据压缩处理后的传输数据以并行方式执行神经网络中的运算，并将运算结果通过与所述主处理电路连接的基础处理电路传输给所述主处理电路。

21、根据权利要求 20 所述的集成电路芯片装置，其特征在于，

所述主处理电路，用于获取待计算的数据块以及运算指令，依据该运算指令对所述待计算的数据块划分成分发数据块以及广播数据块；对所述分发数据块进行拆分处理得到多个基本数据块，将所述多个基本数据块分发至所述 k 个基础处理电路，将所述广播数据块广播至所述 k 个基础处理电路；

所述 k 个基础处理电路，用于依据接收到的基础数据块、广播数据块以及运算指令启动所述压缩映射电路将基础数据块和广播数据块进行压缩处理，然后传输给与所述 k 个基础处理电路连接的基础处理电路；

所述多个基础处理电路，用于对压缩处理后的所述基本数据块与所述广播数据块执行内积运算得到运算结果，并将所述运算结果发送至所述主处理电路；

所述主处理电路，用于对所述运算结果处理得到所述待计算的数据块以及运算指令的指令结果；

其中，所述分发数据块以及所述广播数据块为至少一个输入神经元或者，至少一个权值。

22、一种芯片，其特征在于，所述芯片集成如权利要求 1-16 任意一项所述的装置，或者所述芯片集成如权利要求 17-21 任意一项所述的装置。

23、一种智能设备，其特征在于，所述智能设备包括如权利要求 22 所述的芯片。

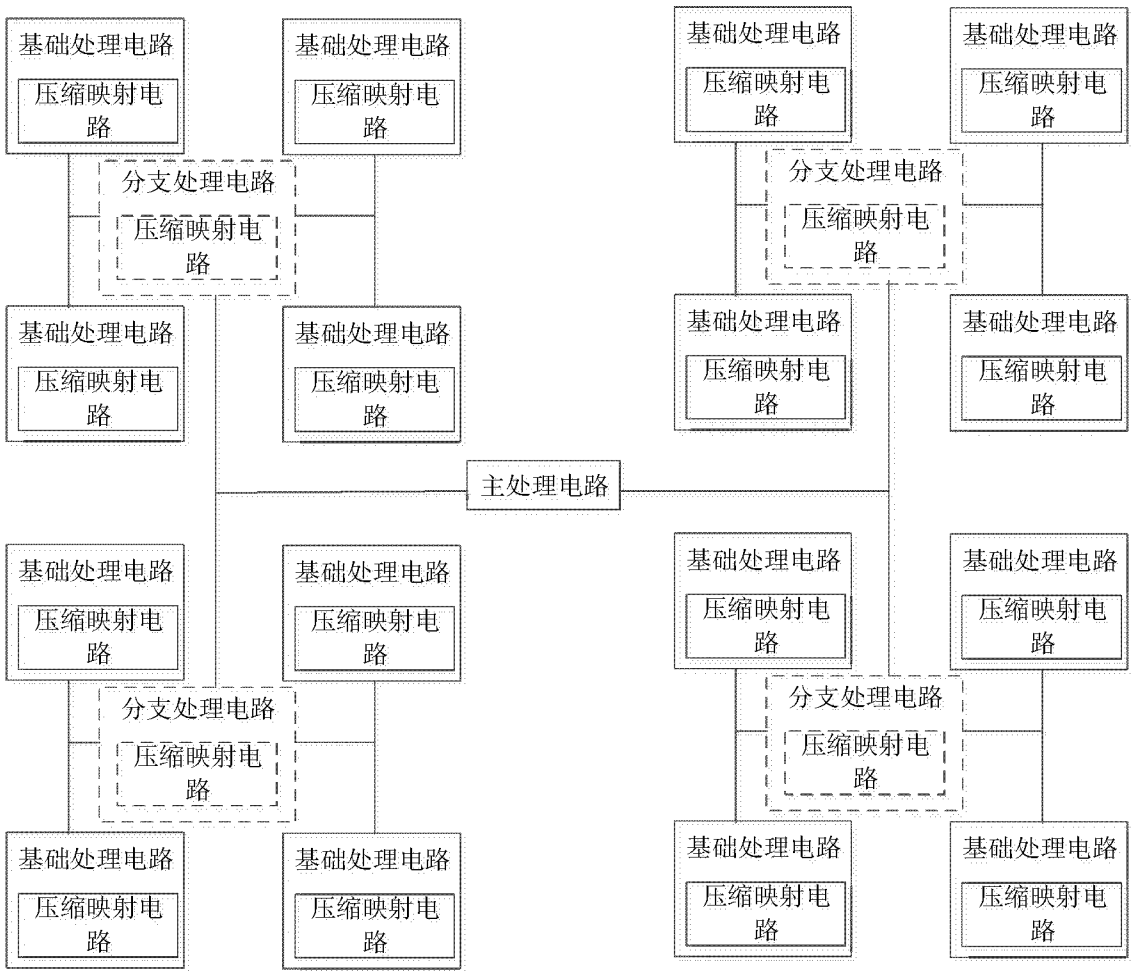

图 1a

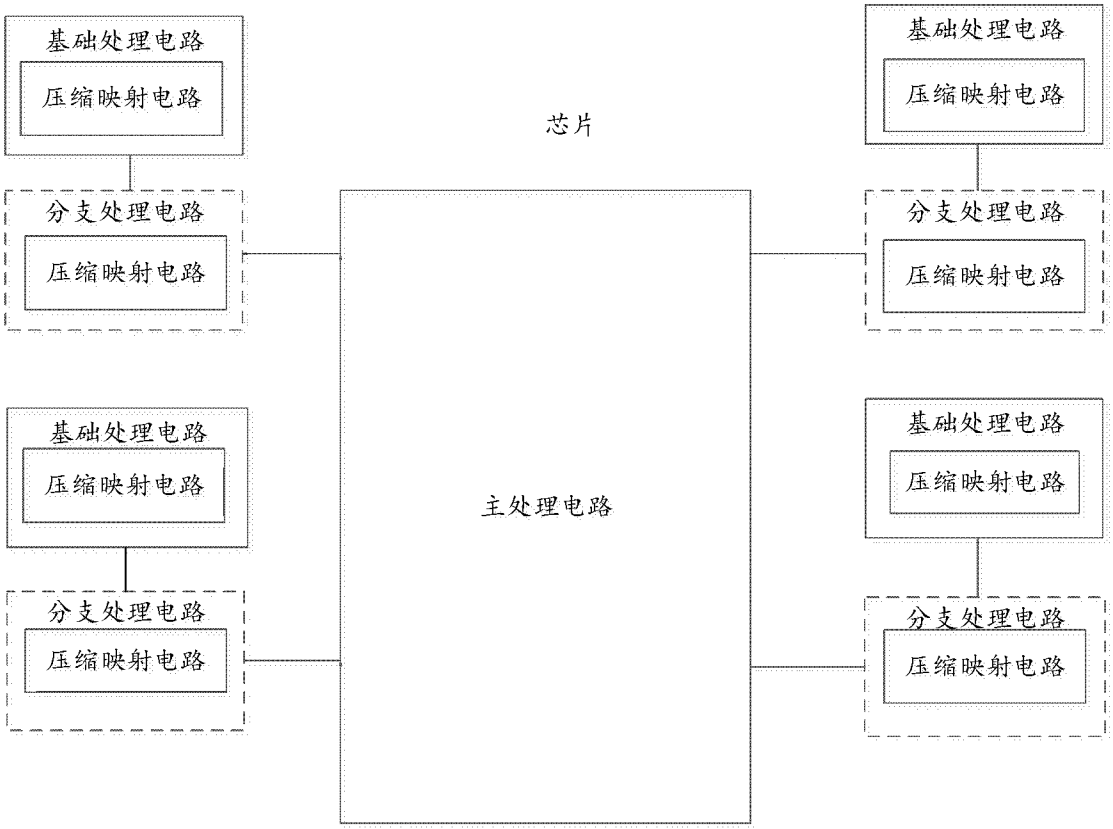
芯片
基础处理电路
压缩映射电路
分支处理电路
压缩映射电路
基础处理电路
压缩映射电路
分支处理电路
压缩映射电路
主处理电路
基础处理电路
压缩映射电路
分支处理电路
压缩映射电路
基础处理电路
压缩映射电路
分支处理电路
压缩映射电路

图 1b


主处理电路
矩阵转置电路
累加器电路
DMA电路
ALU电路
控制电路
向量运算器电路
数据重排电路
激活电路

图 1c

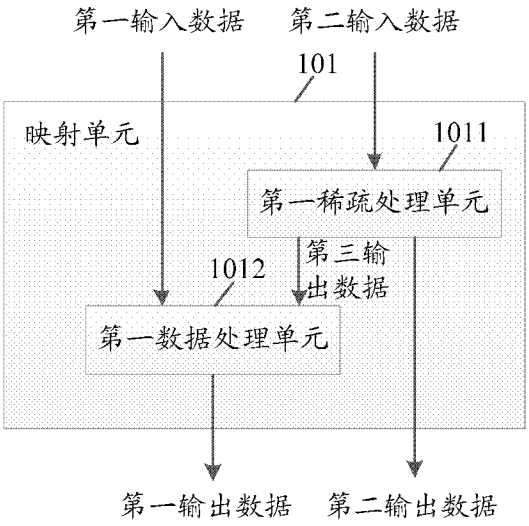
第一输入数据
第二输入数据
101
映射单元
1011
第一稀疏处理单元
第三输出数据
1012
第一数据处理单元
第一输出数据
第二输出数据

图 1d

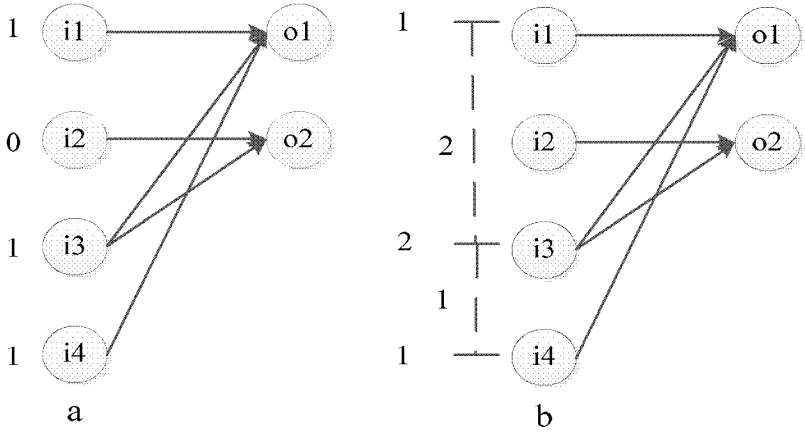
1
i1
o1
0
i2
o2
1
i3
1
i4
a
1
i1
o1
2
i2
o2
2
i3
1
1
i4
b

图 1e

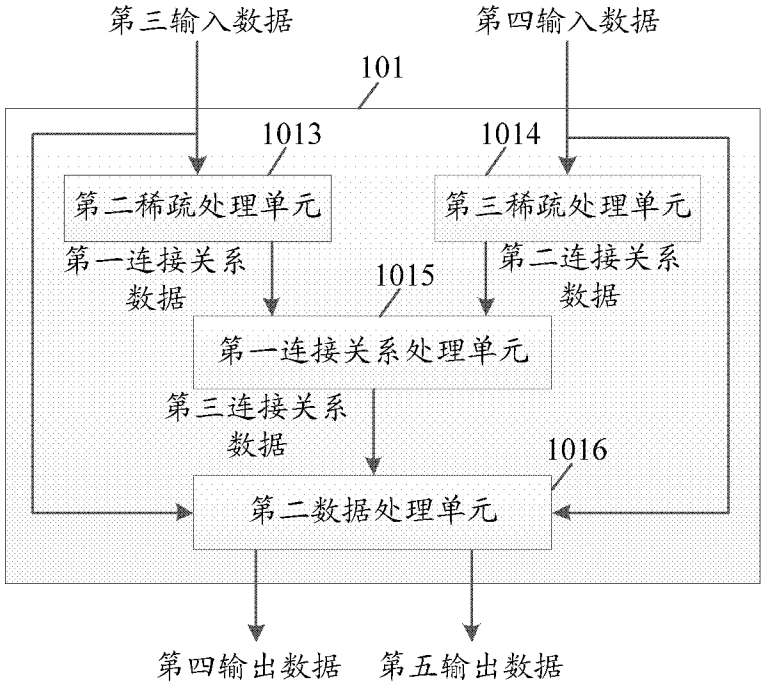
第三输入数据
第四输入数据
101
1013
1014
第二稀疏处理单元
第三稀疏处理单元
第一连接关系数据
第二连接关系数据
1015
第一连接关系处理单元
第三连接关系数据
1016
第二数据处理单元
第四输出数据
第五输出数据

图 1f

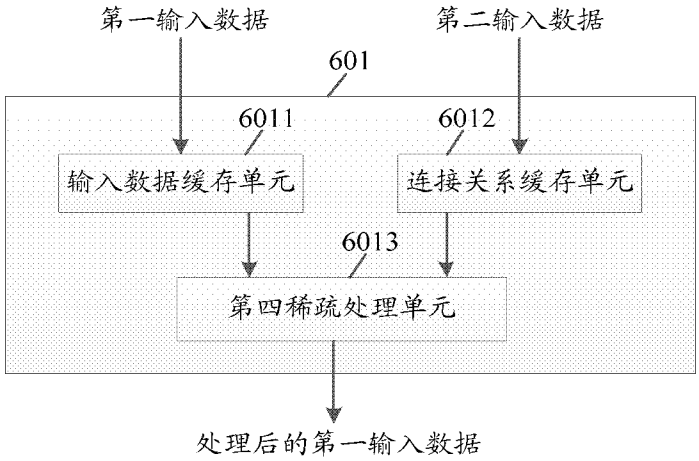
第一输入数据
第二输入数据
601
6011
6012
输入数据缓存单元
连接关系缓存单元
6013
第四稀疏处理单元
处理后的第一输入数据

图 1g

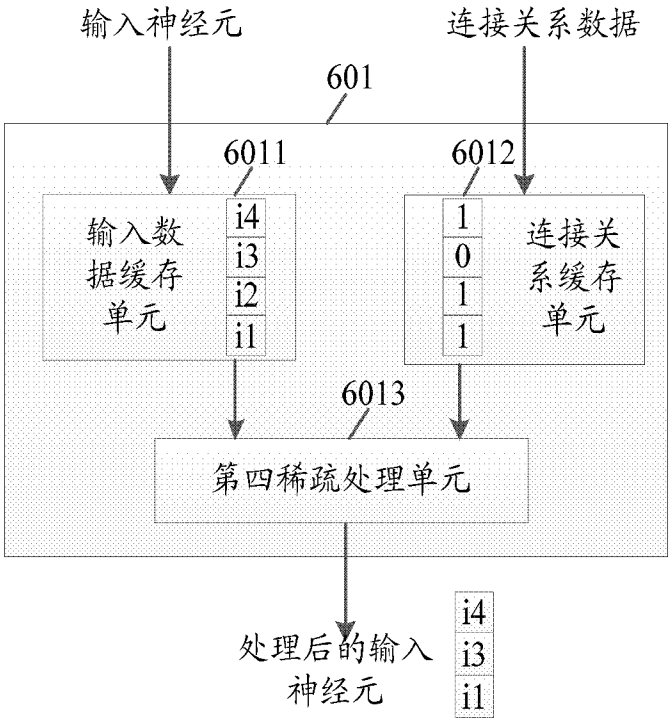
输入神经元
连接关系数据
601
6011
6012
输入数据缓存单元
i4
i3
i2
i1
1
0
1
1
连接关系缓存单元
6013
第四稀疏处理单元
处理后的输入神经元
i4
i3
i1

图 1h

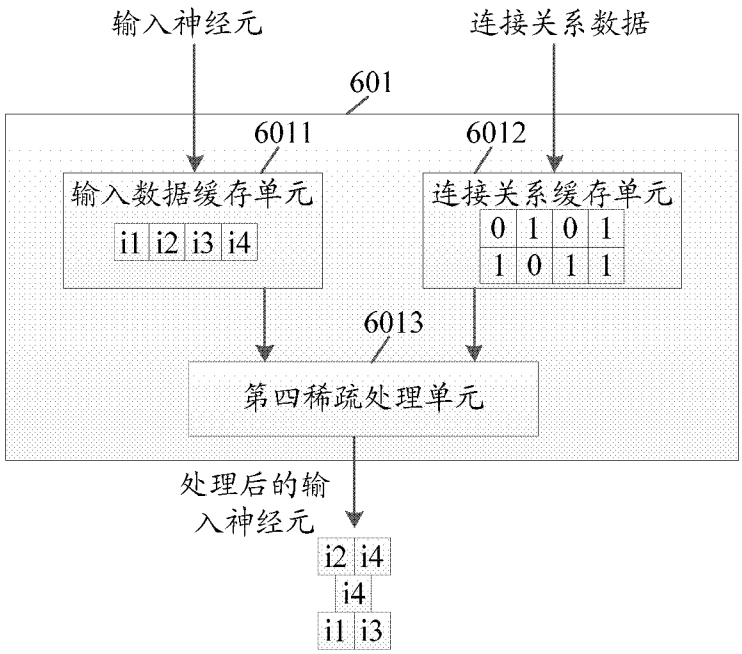
输入神经元
连接关系数据
601
6011
6012
输入数据缓存单元
i1 i2 i3 i4
连接关系缓存单元
0 1 0 1
1 0 1 1
6013
第四稀疏处理单元
处理后的输入神经元
i2 i4
i4
i1 i3

图 1i

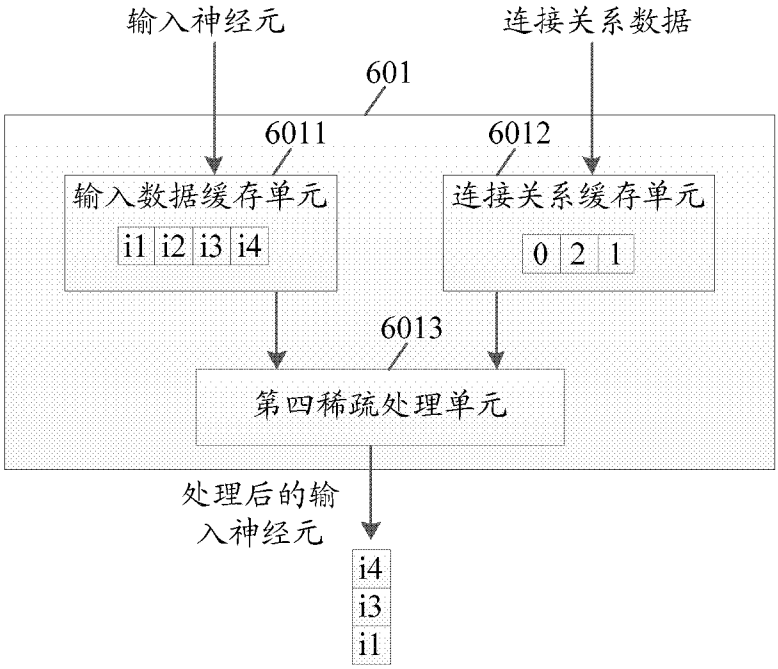
输入神经元
连接关系数据
601
6011
6012
输入数据缓存单元
i1 i2 i3 i4
连接关系缓存单元
0 2 1
6013
第四稀疏处理单元
处理后的输入神经元
i4
i3
i1

图 1j

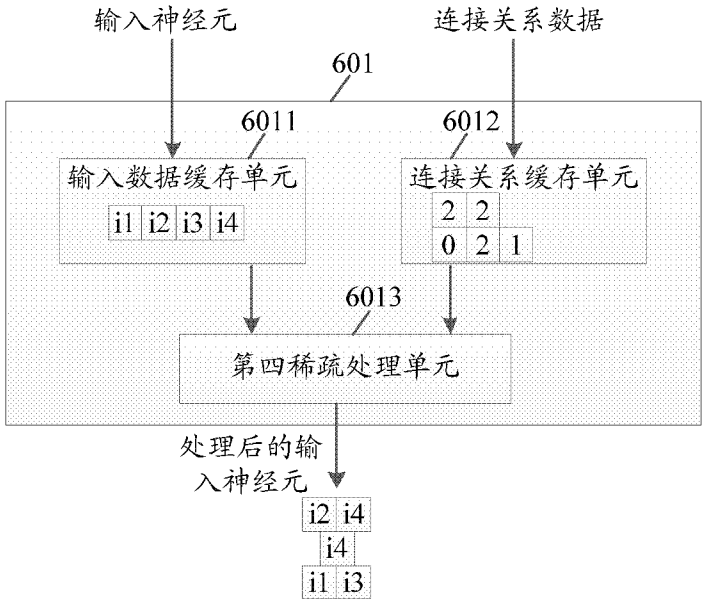
输入神经元
连接关系数据
601
6011
6012
输入数据缓存单元
i1 i2 i3 i4
连接关系缓存单元
2 2
0 2 1
6013
第四稀疏处理单元
处理后的输入神经元
i2 i4
i4
i1 i3

图 1k

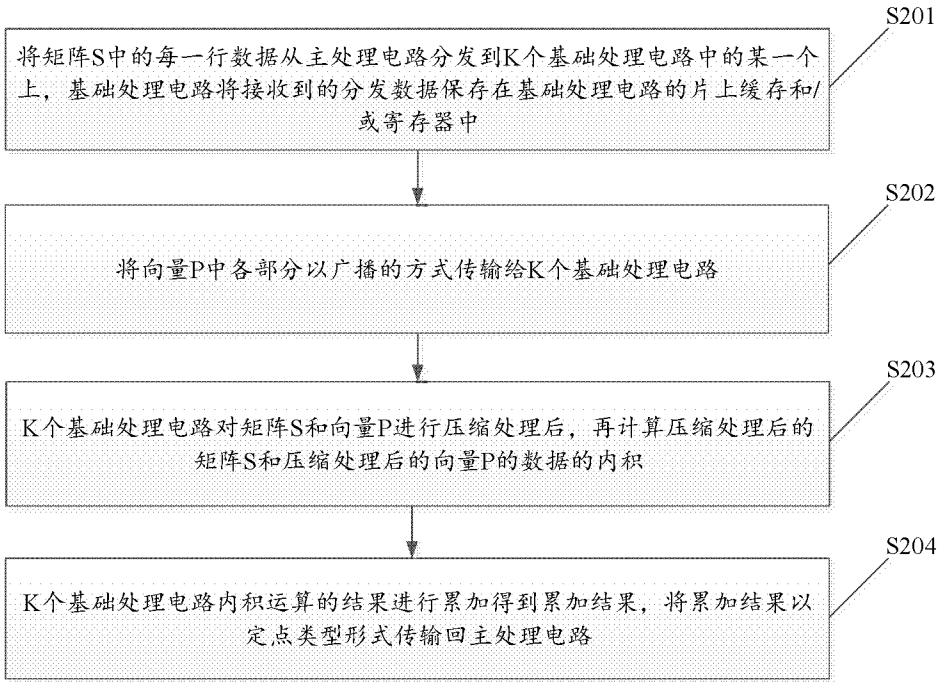
S201
将矩阵S中的每一行数据从主处理电路分发到K个基础处理电路中的某一个上，基础处理电路将接收到的分发数据保存在基础处理电路的片上缓存和/或寄存器中
S202
将向量P中各部分以广播的方式传输给K个基础处理电路
S203
K个基础处理电路对矩阵S和向量P进行压缩处理后，再计算压缩处理后的矩阵S和压缩处理后的向量P的数据的内积
S204
K个基础处理电路内积运算的结果进行累加得到累加结果，将累加结果以定点类型形式传输回主处理电路

图 2

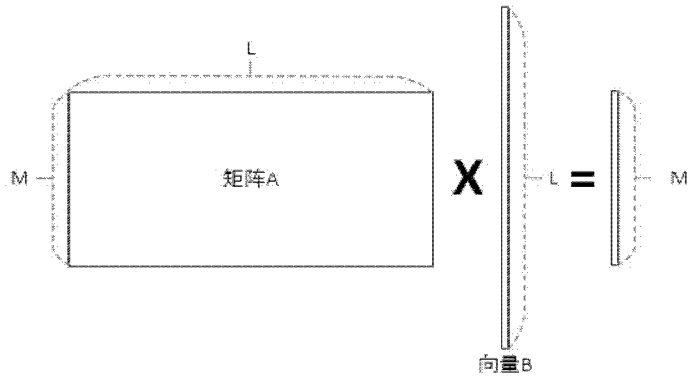
L
M
矩阵A
X
L
=
M
向量B

图 2a

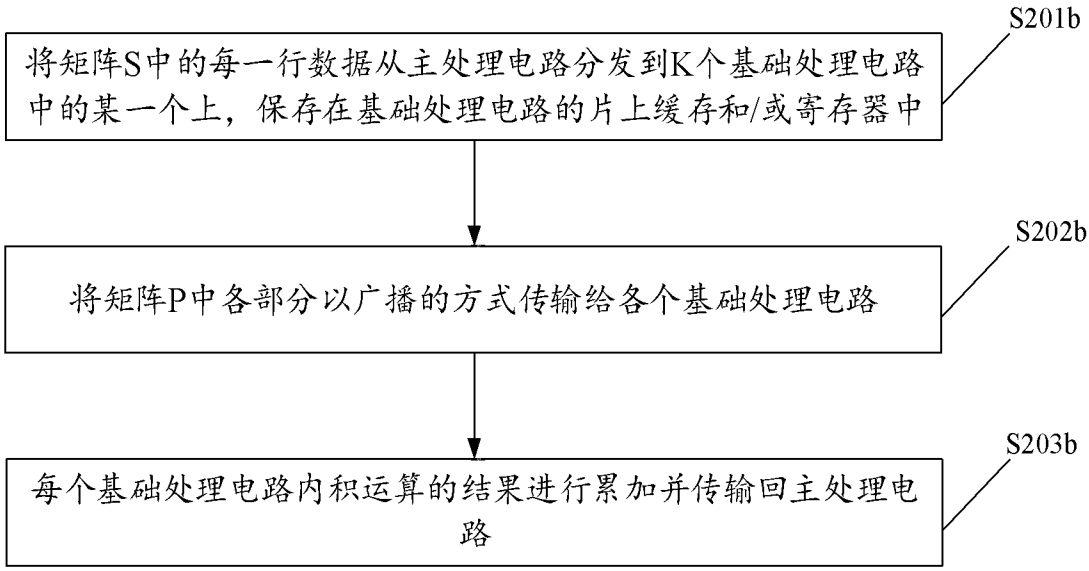
S201b
将矩阵S中的每一行数据从主处理电路分发到K个基础处理电路中的某一个上，保存在基础处理电路的片上缓存和/或寄存器中
S202b
将矩阵P中各部分以广播的方式传输给各个基础处理电路
S203b
每个基础处理电路内积运算的结果进行累加并传输回主处理电路

图 2b

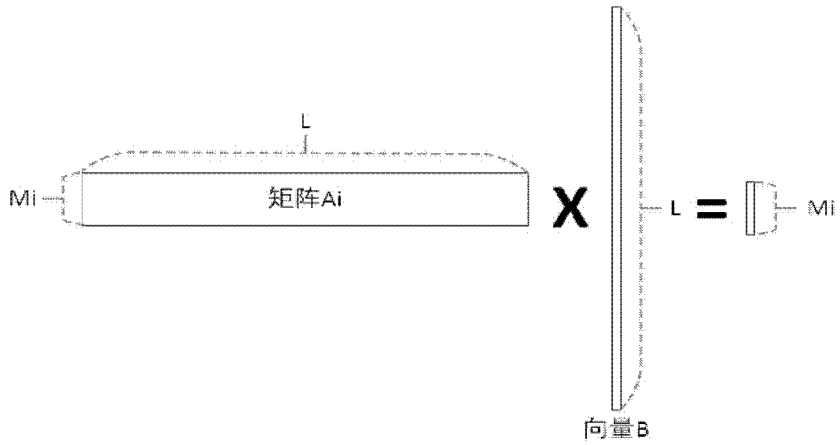
L
Mi
矩阵Ai
X
L
=
Mi
向量B

图 2c

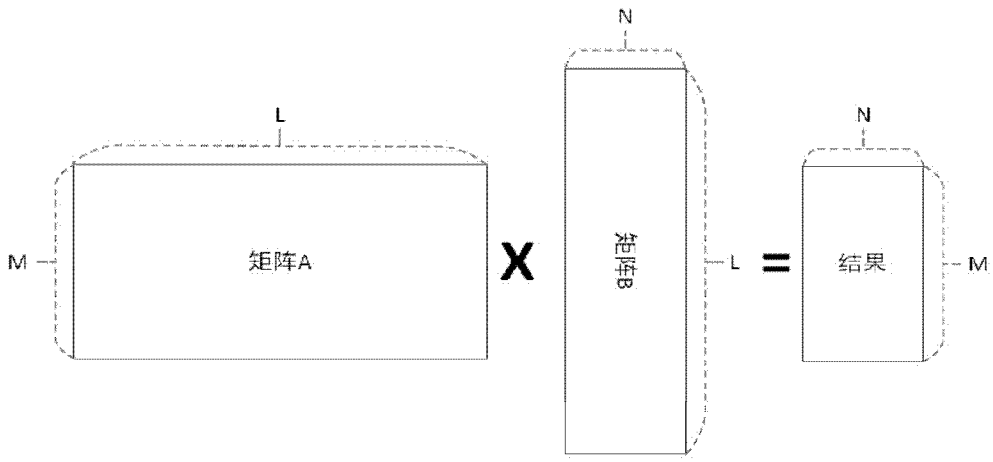
N
L
N
M
矩阵A
X
矩阵B
L
=
结果
M

图 2d

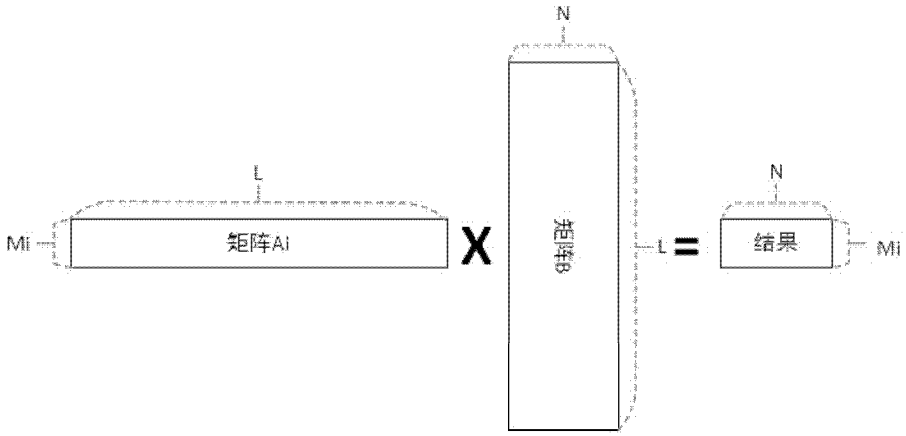
N
L
N
Mi
矩阵Ai
X
矩阵B
L
=
结果
Mi

图 2e

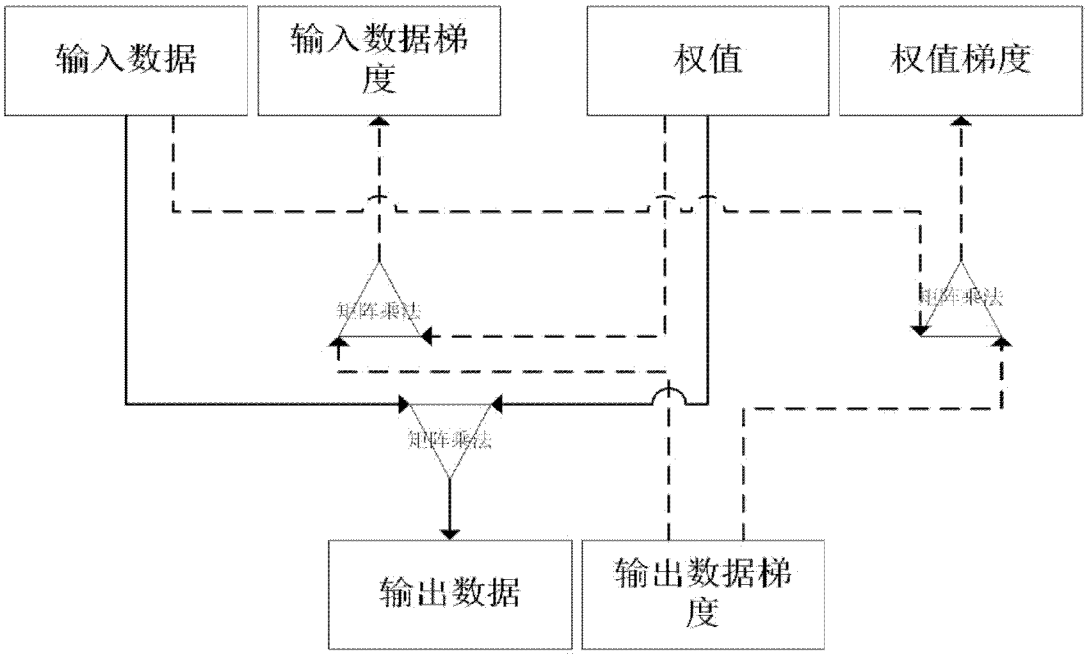
输入数据
输入数据梯度
权值
权值梯度
矩阵乘法
矩阵乘法
矩阵乘法
输出数据
输出数据梯度

图 3a

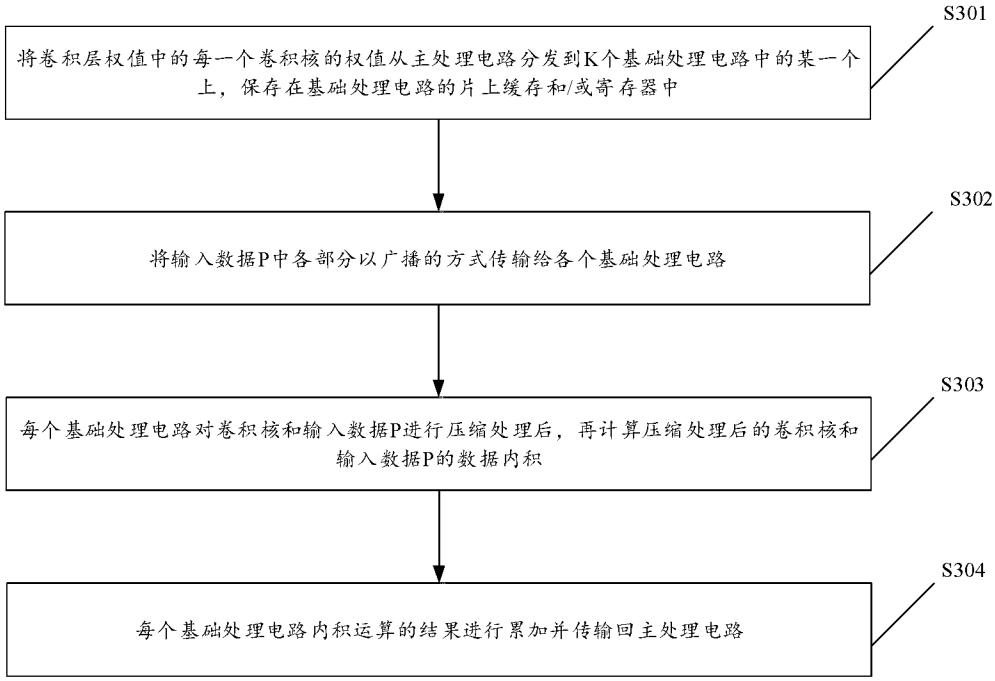
S301
将卷积层权值中的每一个卷积核的权值从主处理电路分发到K个基础处理电路中的某一个上，保存在基础处理电路的片上缓存和/或寄存器中
S302
将输入数据P中各部分以广播的方式传输给各个基础处理电路
S303
每个基础处理电路对卷积核和输入数据P进行压缩处理后，再计算压缩处理后的卷积核和输入数据P的数据内积
S304
每个基础处理电路内积运算的结果进行累加并传输回主处理电路

图 3b

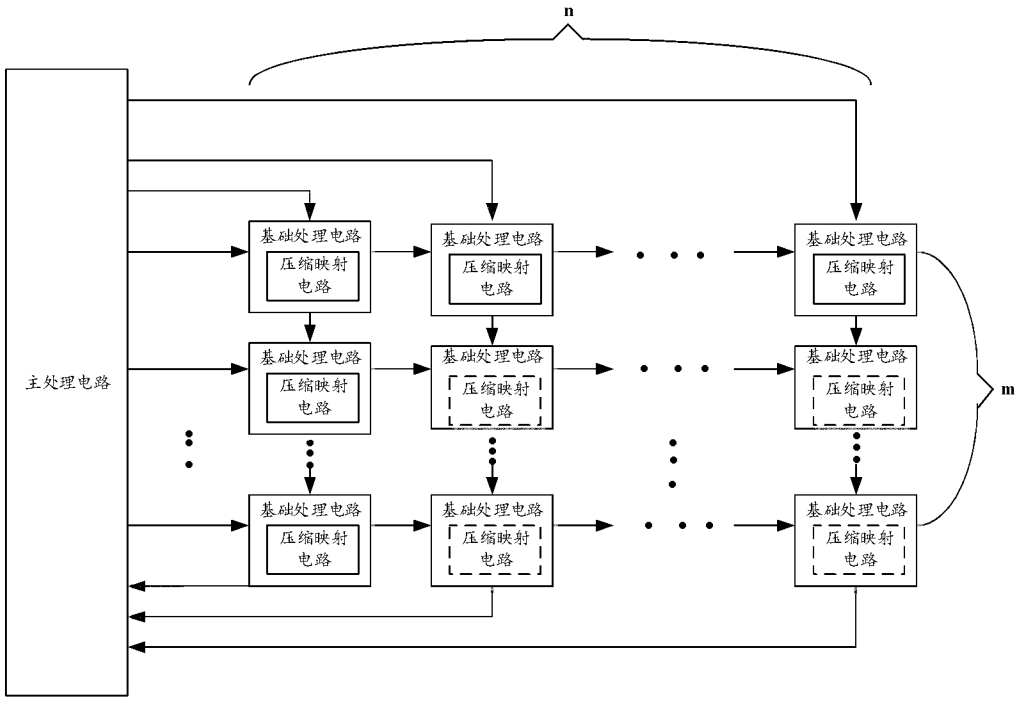
n
主处理电路
基础处理电路
压缩映射电路
m

图 4a

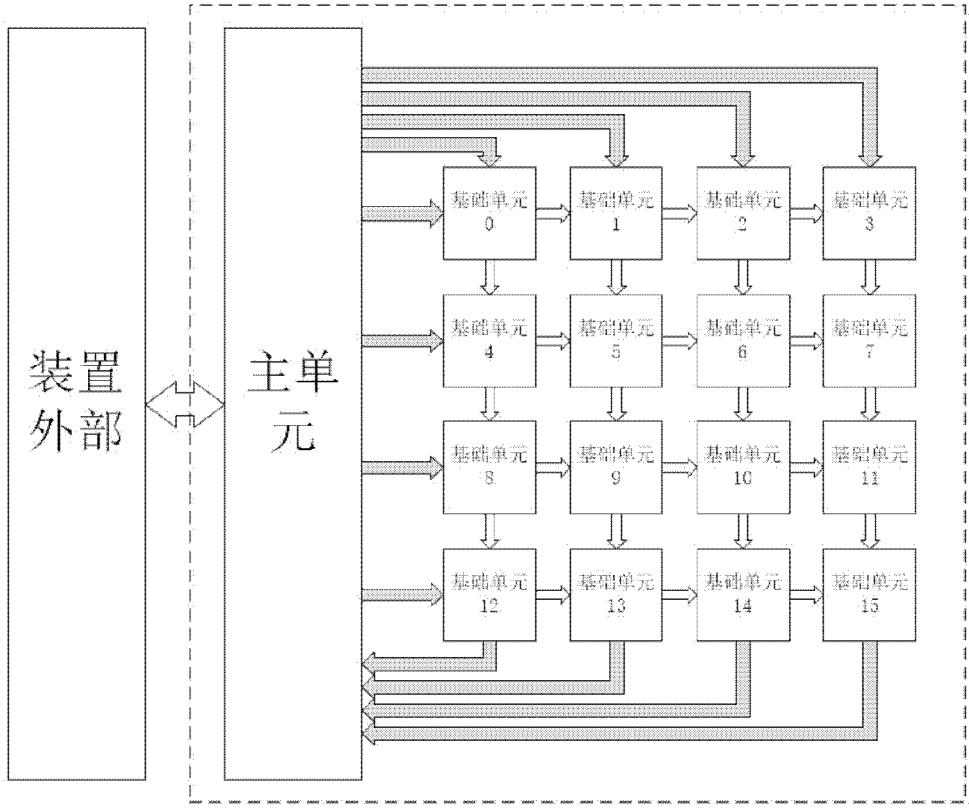
装置外部
主单元
基础单元0
基础单元1
基础单元2
基础单元3
基础单元4
基础单元5
基础单元6
基础单元7
基础单元8
基础单元9
基础单元10
基础单元11
基础单元12
基础单元13
基础单元14
基础单元15

图 4b

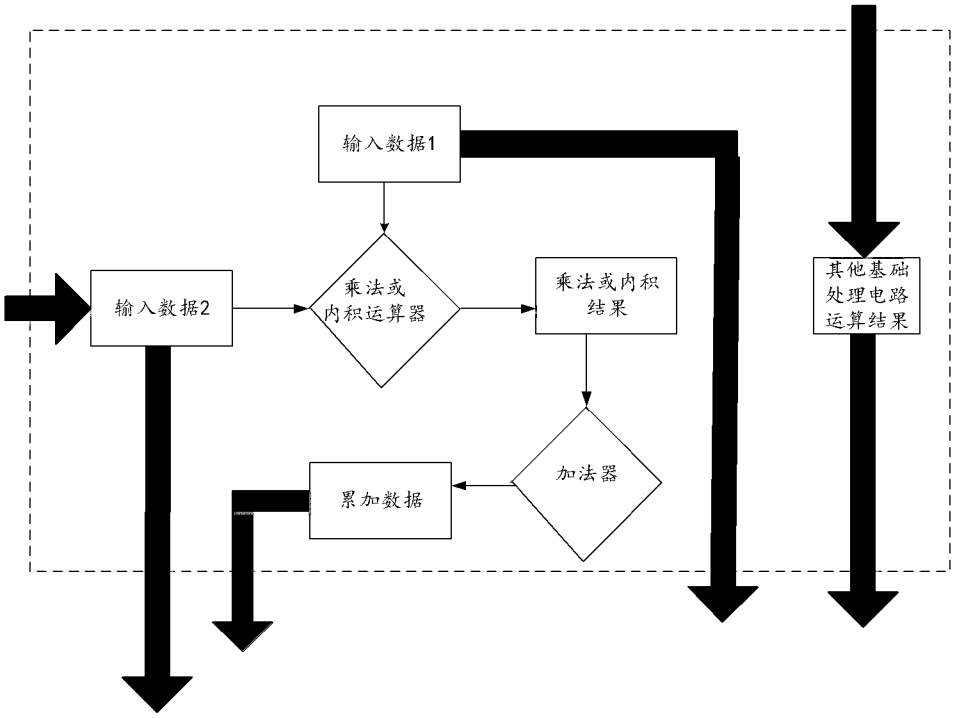
输入数据1
输入数据2
乘法或内积运算器
乘法或内积结果
加法器
累加数据
其他基础处理电路运算结果

图 4c

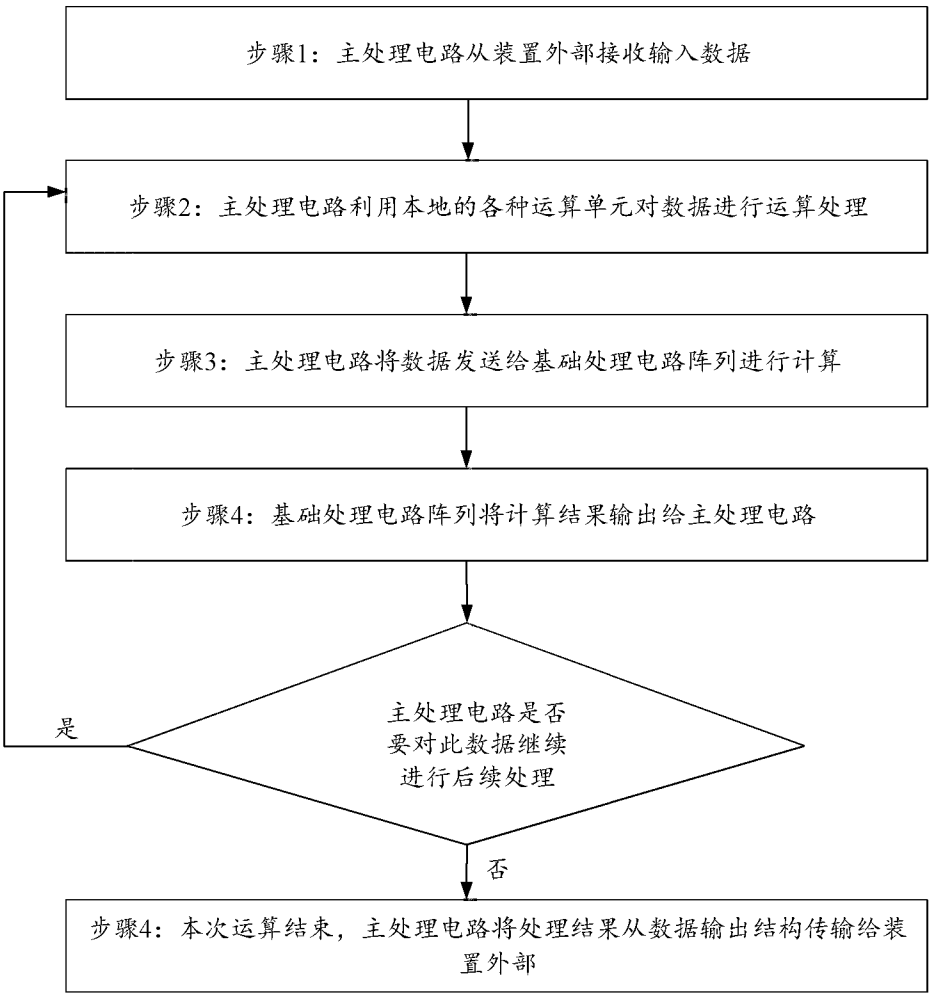
步骤1：主处理电路从装置外部接收输入数据
步骤2：主处理电路利用本地的各种运算单元对数据进行运算处理
步骤3：主处理电路将数据发送给基础处理电路阵列进行计算
步骤4：基础处理电路阵列将计算结果输出给主处理电路
主处理电路是否
要对此数据继续
进行后续处理
是
否
步骤4：本次运算结束，主处理电路将处理结果从数据输出结构传输给装置外部

图 5a

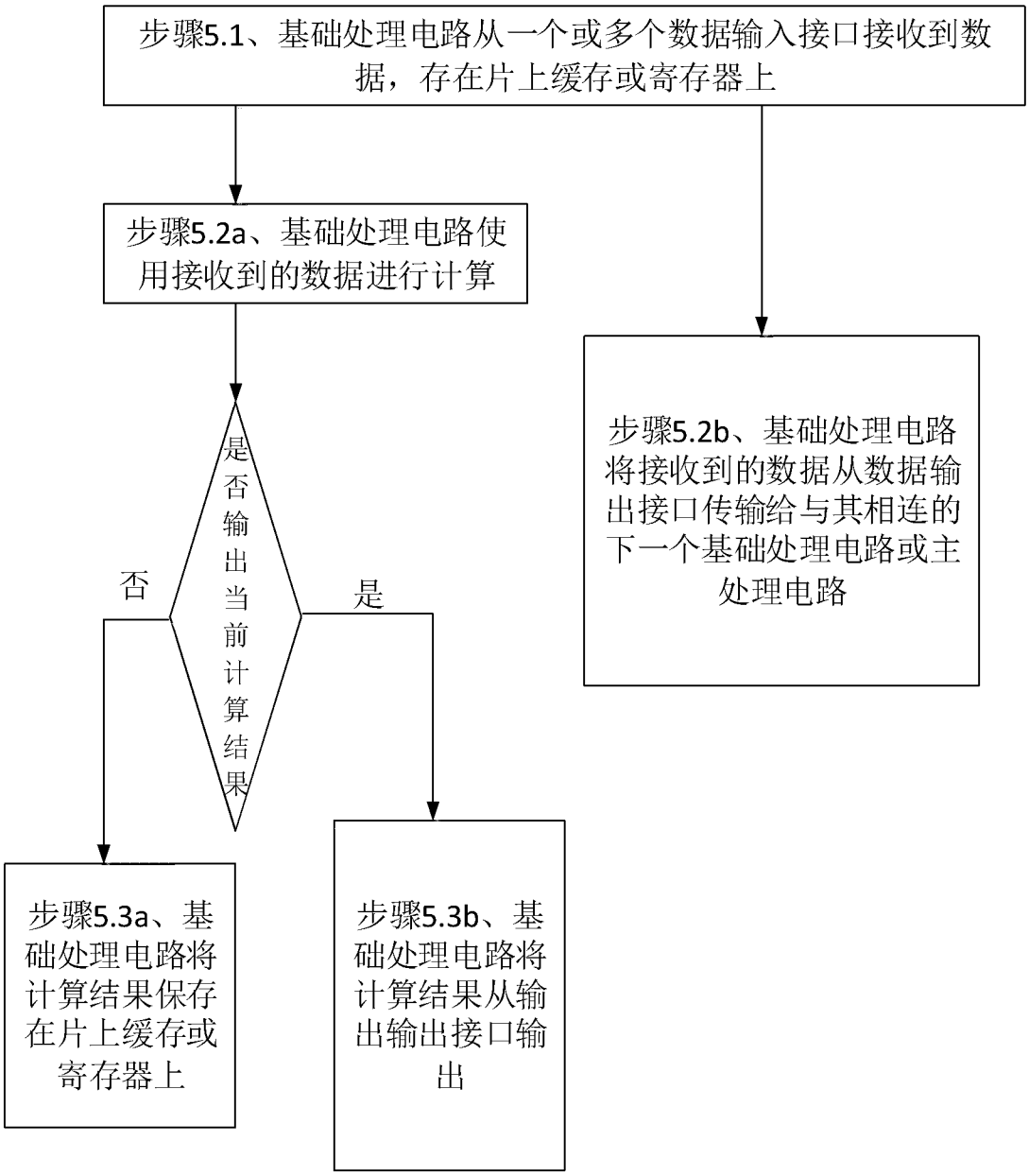
步骤5.1、基础处理电路从一个或多个数据输入接口接收到数据，存在片上缓存或寄存器上
步骤5.2a、基础处理电路使用接收到的数据进行计算
步骤5.2b、基础处理电路将接收到的数据从数据输出接口传输给与其相连的下一个基础处理电路或主处理电路
是否输出当前计算结果
否
是
步骤5.3a、基础处理电路将计算结果保存在片上缓存或寄存器上
步骤5.3b、基础处理电路将计算结果从输出输出接口输出

图 5b

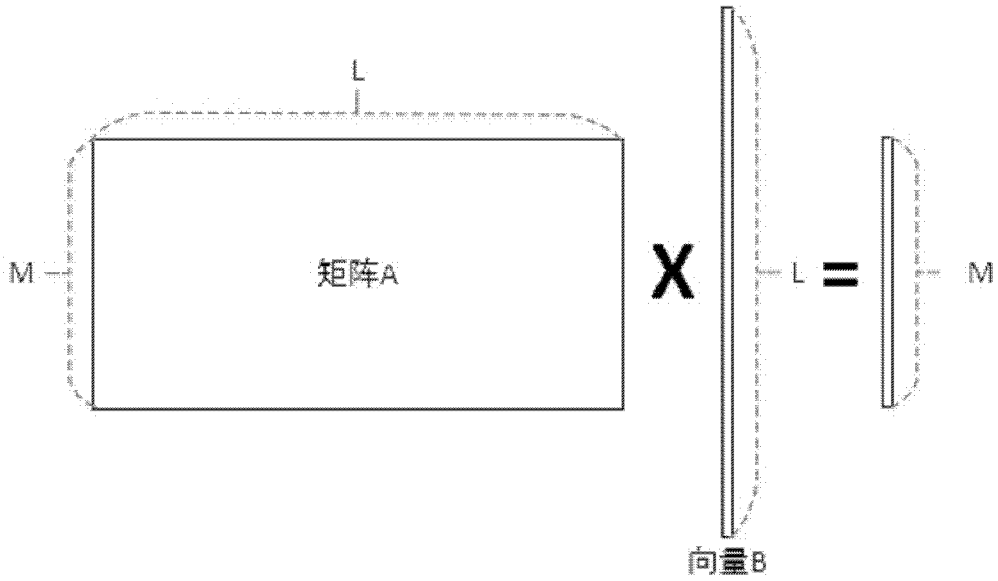

图 5c

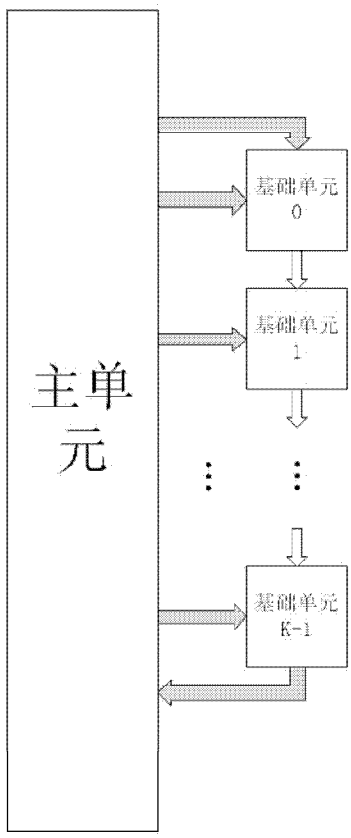

图 5d

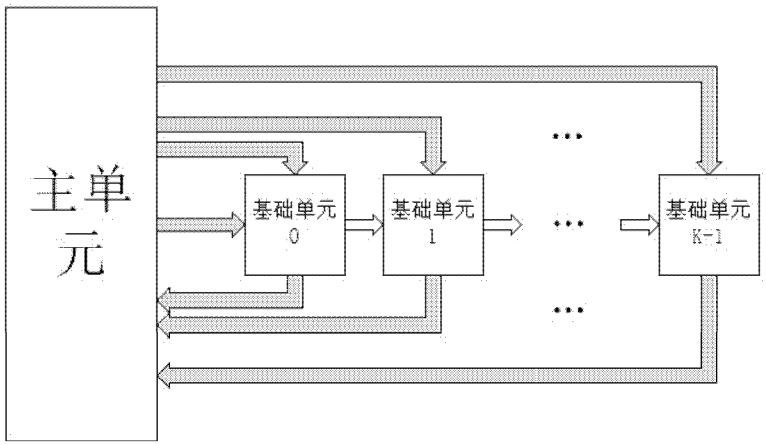

图 5e

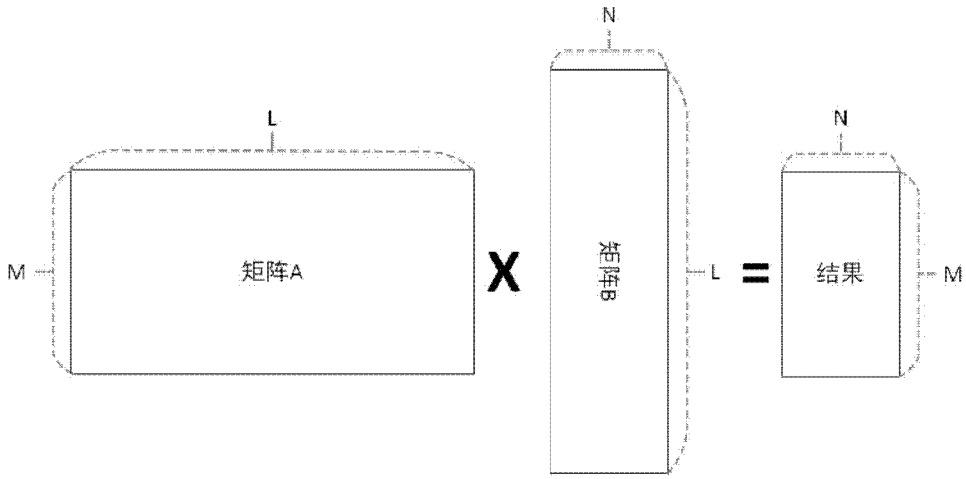

图 5f

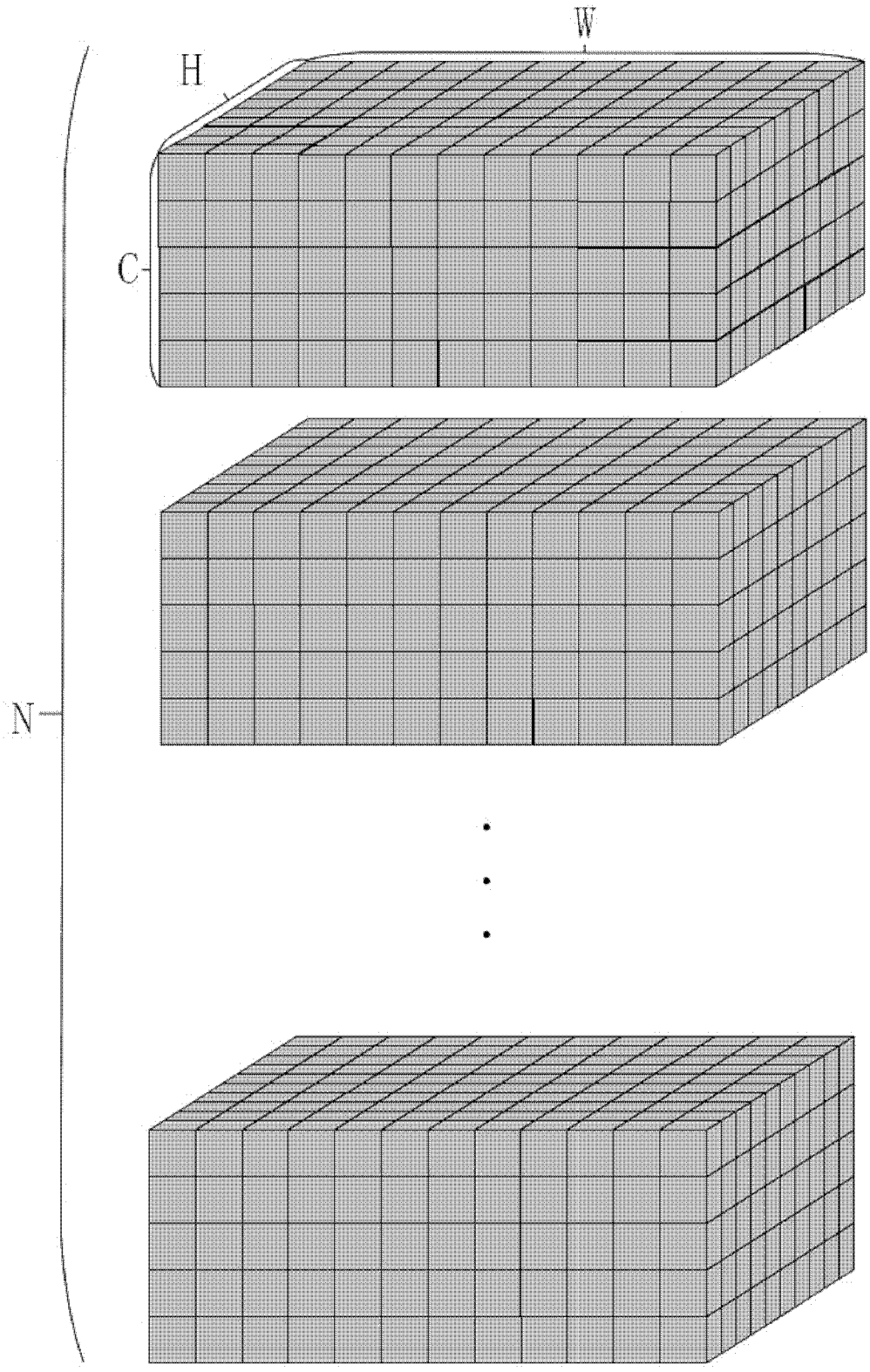

图 6a

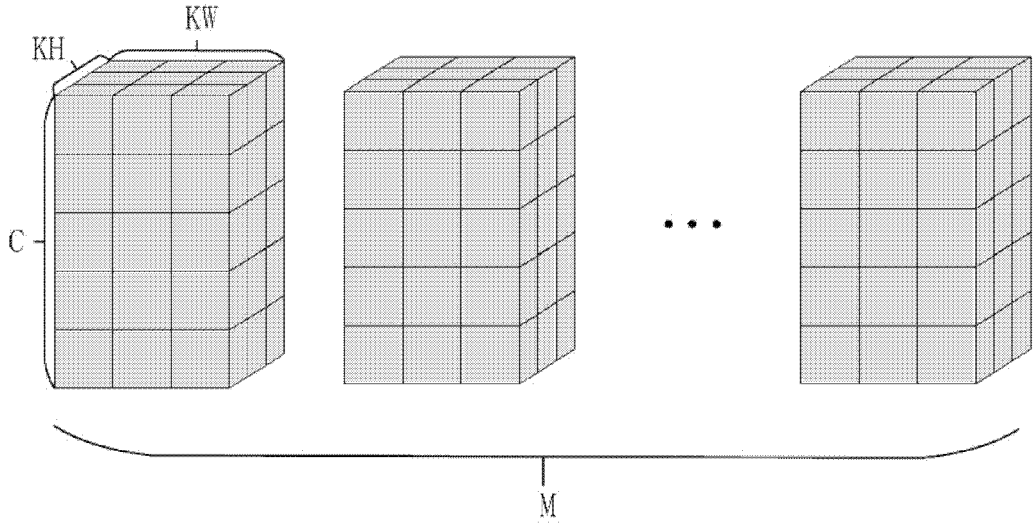

图 6b

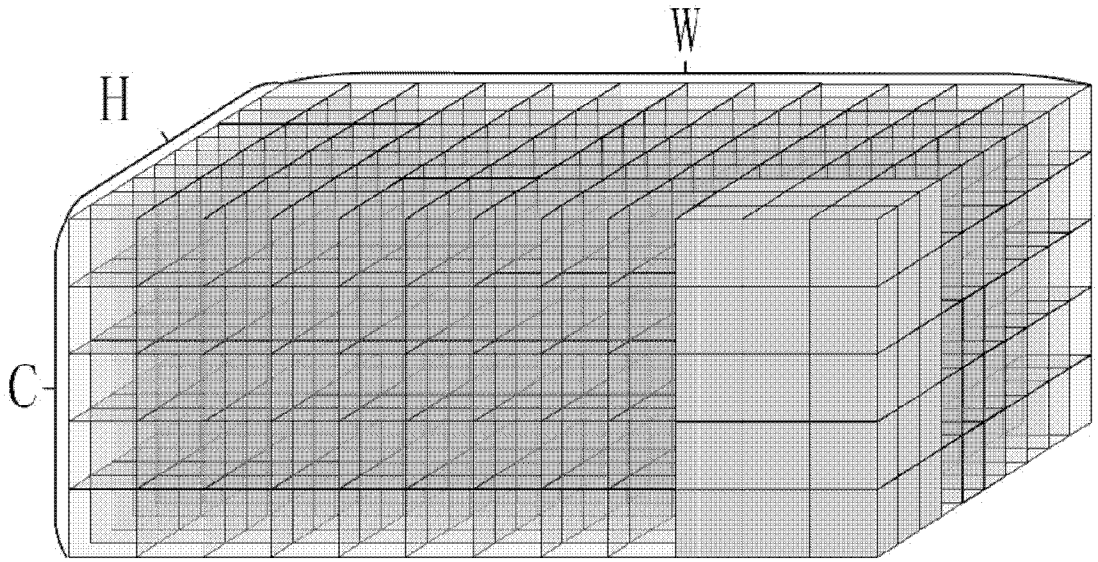

图 6c

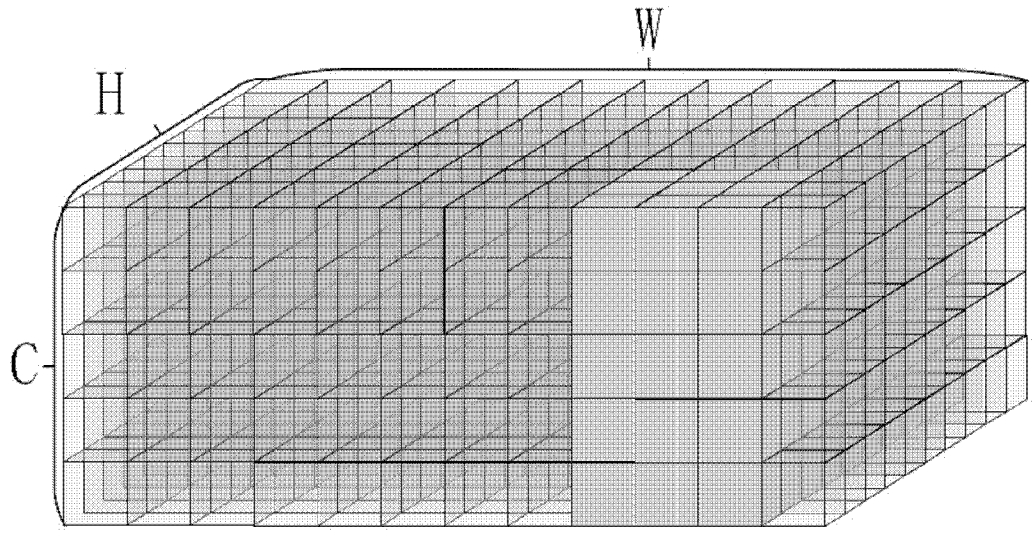

图 6d

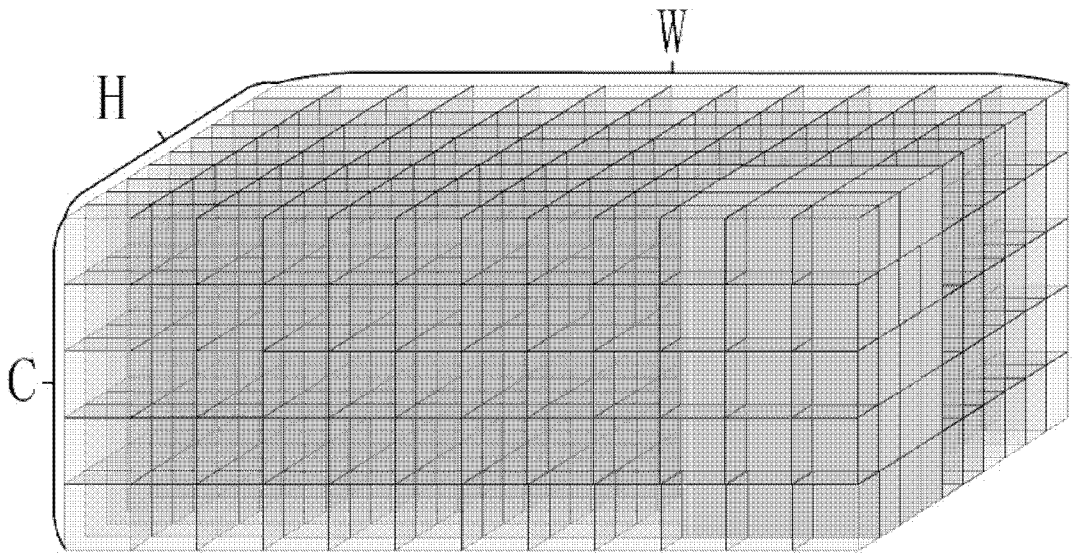

图 6e

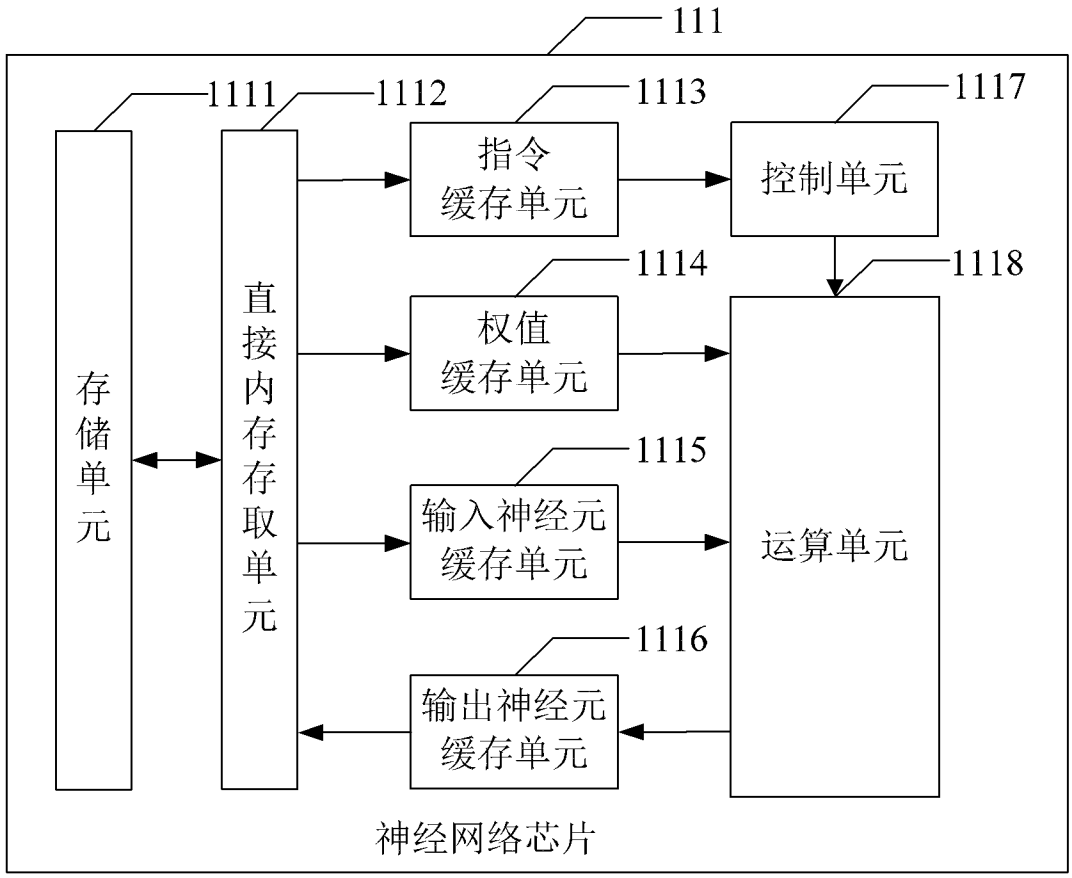
111
1111
1112
1113
1117
指令
缓存单元
控制单元
1114
1118
存储单元
直接内存存取单元
权值
缓存单元
1115
输入神经元
缓存单元
运算单元
1116
输出神经元
缓存单元
神经网络芯片

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/125801

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/063(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)

WPI, EPODOC, CNPAT, CNKI, IEEE: 神经, 网络, 电路, 集成, 芯片, 运算, 计算, 计示, 压缩, 映射, 传输, 传送, 输入, 输出, 数据, 神经元, 权值, 权重, 分发, 分配, 广播, neural network, "NN", neuro, integrated, circuit, chip, die, calculate, compress, data, distribute

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106126481 A (HUAWEI TECHNOLOGIES CO., LTD.) 16 November 2016 (2016-11-16) description, paragraphs [0046]-[0084], and [0123]-[0134], and figures 4 and 9	1, 2, 22, 23
Y	CN 106126481 A (HUAWEI TECHNOLOGIES CO., LTD.) 16 November 2016 (2016-11-16) description, paragraphs [0046]-[0084], and [0123]-[0134], and figures 4 and 9	3-6, 14-23
Y	CN 106447034 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 22 February 2017 (2017-02-22) description, paragraphs [0036]-[0063], and figures 1-5	3-6, 14-23
Y	CN 107229967 A (BEIJING DEEPHI TECHNOLOGY CO., LTD.) 03 October 2017 (2017-10-03) description, paragraphs [0159]-[0169]	4, 6, 14-16, 19, 21-23
A	CN 107220702 A (BEIJING TUSIMPLE FUTURE TECHNOLOGY CO., LTD.) 29 September 2017 (2017-09-29) entire document	1-23
A	EP 0631254 A2 (MOTOROLA, INC.) 28 December 1994 (1994-12-28) entire document	1-23

☐ 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
15 March 2019	**29 March 2019**

Name and mailing address of the ISA/CN	Authorized officer
National Intellectual Property Administration, PRC (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/125801

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106126481	A	16 November 2016	None			
CN	106447034	A	22 February 2017	None			
CN	107229967	A	03 October 2017	US	2018046913	A1	15 February 2018
				US	2018046894	A1	15 February 2018
				US	2018046903	A1	15 February 2018
				CN	107657263	A	02 February 2018
				CN	107729999	A	23 February 2018
				CN	107704916	A	16 February 2018
				CN	107239829	A	10 October 2017
				US	2018046897	A1	15 February 2018
CN	107220702	A	29 September 2017	None			
EP	0631254	A2	28 December 1994	EP	0631254	A3	15 February 1995
				MX	9404348	A	31 January 1995
				US	5720002	A	17 February 1998
				US	5517667	A	14 May 1996
				US	5781701	A	14 July 1998
				CN	1100541	A	22 March 1995
				JP	H0713949	A	17 January 1995
				CA	2125255	A1	15 December 1994
				US	5574827	A	12 November 1996

国际检索报告

国际申请号
PCT/CN2018/125801

A. 主题的分类

G06N 3/063(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））

WPI, EPODOC, CNPAT, CNKI, IEEE：神经，网络，电路，集成，芯片，运算，计算，计标，压缩，映射，传输，传送，输入，输出，数据，神经元，权值，权重，分发，分配，广播，neural network，“NN”,neuro，integrated, circuit，chip，die，calculate，compress，data，distribute

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 106126481 A (华为技术有限公司) 2016年 11月 16日 (2016-11-16) 说明书第0046-0084段、第0123-0134段及图4、9	1-2、22-23
Y	CN 106126481 A (华为技术有限公司) 2016年 11月 16日 (2016-11-16) 说明书第0046-0084段、第0123-0134段及图4、9	3-6、14-23
Y	CN 106447034 A (中国科学院计算技术研究所) 2017年 2月 22日 (2017-02-22) 说明书第0036-0063段及图1-5	3-6、14-23
Y	CN 107229967 A (北京深鉴智能科技有限公司) 2017年 10月 3日 (2017-10-03) 说明书第0159-0169段	4、6、14-16、19、21-23
A	CN 107220702 A (北京图森未来科技有限公司) 2017年 9月 29日 (2017-09-29) 全文	1-23
A	EP 0631254 A2 (MOTOROLA, INC.) 1994年 12月 28日 (1994-12-28) 全文	1-23

☐ 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
2019年 3月 15日	2019年 3月 29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 胡晓英 电话号码 010-53961456

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/125801

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	106126481	A	2016年 11月 16日	无			
CN	106447034	A	2017年 2月 22日	无			
CN	107229967	A	2017年 10月 3日	US	2018046913	A1	2018年 2月 15日
				US	2018046894	A1	2018年 2月 15日
				US	2018046903	A1	2018年 2月 15日
				CN	107657263	A	2018年 2月 2日
				CN	107729999	A	2018年 2月 23日
				CN	107704916	A	2018年 2月 16日
				CN	107239829	A	2017年 10月 10日
				US	2018046897	A1	2018年 2月 15日
CN	107220702	A	2017年 9月 29日	无			
EP	0631254	A2	1994年 12月 28日	EP	0631254	A3	1995年 2月 15日
				MX	9404348	A	1995年 1月 31日
				US	5720002	A	1998年 2月 17日
				US	5517667	A	1996年 5月 14日
				US	5781701	A	1998年 7月 14日
				CN	1100541	A	1995年 3月 22日
				JP	H0713949	A	1995年 1月 17日
				CA	2125255	A1	1994年 12月 15日
				US	5574827	A	1996年 11月 12日