(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
2018 年 7 月 5 日 (05.07.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/121472 A1

(51) 国际专利分类号:
G06N 3/063 (2006.01)

(21) 国际申请号: PCT/CN2017/118264

(22) 国际申请日: 2017 年 12 月 25 日 (25.12.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201611234438.1 2016年12月28日 (28.12.2016) CN

(71) 申请人: 上海寒武纪信息科技有限公司 (SHANGHAI CAMBRICON INFORMATION TECHNOLOGY CO., LTD) [CN/CN]; 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (CN)。

(72) 发明人: 陈云霁 (CHEN, Yunji); 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (CN)。 宋新开 (SONG, Xinkai); 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (CN)。 刘少礼 (LIU, Shaoli); 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (CN)。 陈天石 (CHEN, Tianshi); 中国上海市浦东新区祖冲之路2290号1号楼1004, Shanghai 201203 (CN)。

(74) 代理人: 中科专利商标代理有限责任公司 (CHINA SCIENCE PATENT & TRADEMARK AGENT LTD.); 中国北京市海淀区西三环北路87号4-1105室, Beijing 100089 (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, 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,

(54) **Title:** COMPUTATION METHOD

(54) 发明名称: 一种运算方法

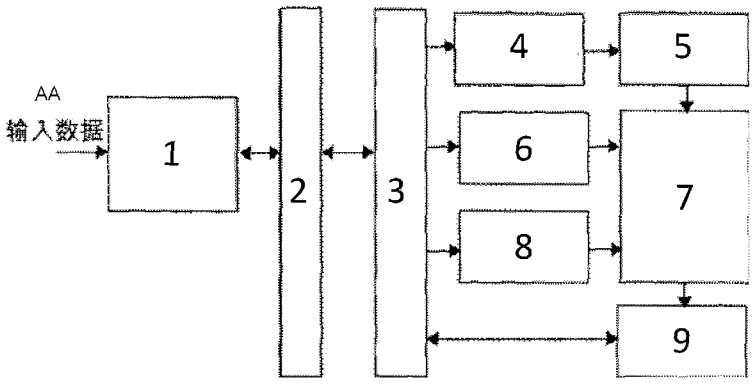

图 2

AA Input data

(57) **Abstract:** Provided is a computation method, using an artificial neural network calculation device to perform a computation on a multilayer artificial neural network.

(57) 摘要: 一种运算方法，使用人工神经网络计算装置对多层人工神经网络进行运算。

[见续页]

WO 2018/121472 A1

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))。

一种运算方法

技术领域

本公开涉及数据处理技术领域，更具体地涉及一种运算方法。

背景技术

人工神经网络（Artificial Neural Networks，ANNs）简称为神经网络（NNs），它是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间的相互连接关系，从而达到处理信息的目的。神经网络用到的算法就是向量乘法，并且广泛采用符号函数及其各种逼近。

就像大脑里的神经网络一样，神经网络由一些互相连接的节点组成，如图 1 所示，每个圆圈表示一个神经元，每个箭头表示两个神经元之间的连接，连接又被称为权值。

神经元的计算公式可以简单的描述成：$y=f(\sum_{i=0}^{n} w_i * x_i)$。其中，$x_i$ 表示所有和输出神经元相连接的输入神经元，w_i 表示 x_i 和输出神经元之间对应的权值。f(x)是一个非线性函数，通常称作激活函数，常用的函数如：$\frac{1}{1+e^{-x}}$，$\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$等。

神经网络被广泛应用于各种应用场景：计算视觉、语音识别和自然语言处理等。在近几年的时间里，神经网络的规模一直在增长。在 1998 年，Lecun 用于手写字符识别的神经网络的规模小于 1M 个权值；在 2012 年，krizhevsky 用于参加 ImageNet 竞赛的规模是 60M 个权值。

公开内容

根据本公开的一个方面，提供了一种运算方法，使用人工神经网络计算装置对多层人工神经网络进行训练，所述人工神经网络计算装置包括：映射单元、存储单元和运算单元，所述运算方法包括：对多层人工神经网络中的各层依次执行正向运算，对于所述多层人工神经网络的每一层，所

述正向运算包括：所述映射单元对输入数据进行处理，以得到处理后的输入数据； 所述存储单元存储所述处理后的输入数据；所述运算单元读取所述处理后的输入数据，进行人工神经网络运算，得到人工神经网络的输出数据；按照与所述正向运算相反的顺序，对所述多层人工神经网络中的各层依次执行反向运算；对所述多层人工神经网络中的各层进行权值更新；重复执行上述各个步骤多次，完成多层人工神经网络的训练。

在本公开的一些实施例中，所述人工神经网络计算装置还包括：第一输入缓存和第二输入缓存；所述对多层人工神经网络中的各层进行权值更新包括：对于所述人工神经网络的每一层，所述第一输入缓存和所述第二输入缓存分别存储该层的权值和权值梯度；所述运算单元从所述第二输入缓存接收所述权值梯度，然后使用所述权值梯度对所述权值进行更新，得到更新后的权值。

在本公开的一些实施例中，所述输入数据包括：输入神经元和权值；所述映射单元对输入数据进行处理，以得到处理后的输入数据包括：所述映射单元对权值进行处理，得到处理后的权值和权值的连接关系；所述映射单元基于所述权值的连接关系数据对输入神经元进行处理，得到处理后的输入神经元。

在本公开的一些实施例中，所述输入数据包括输入神经元和权值；所述映射单元对输入数据进行处理，以得到处理后的输入数据包括：所述映射单元对输入神经元进行处理，得到处理后的输入神经元和输入神经元的连接关系；所述映射单元基于所述输入神经元的连接关系数据对权值进行处理，得到处理后的权值。

在本公开的一些实施例中，所述映射单元对权值进行处理包括：所述映射单元将无连接的权值删除，保留有连接的权值作为处理后的权值。

在本公开的一些实施例中，所述映射单元对权值进行处理包括：所述映射单元将所述权值中数值或绝对值小于或者等于第一阈值的权值删除，得到处理后的权值。

在本公开的一些实施例中，所述映射单元对输入神经元进行处理包括：所述映射单元将所述输入神经元中数值或绝对值小于或者等于第二阈值的输入神经元删除，得到处理后的输入神经元。

在本公开的一些实施例中，所述输入数据包括处理后的权值、输入神经元和所述输入神经元的连接关系数据；所述映射单元对输入数据进行处理，以得到处理后的输入数据包括；所述映射单元根据所述输入神经元的连接关系数据对所述输入神经元进行处理，以得到处理后的输入神经元。

在本公开的一些实施例中，所述输入数据包括处理后的输入神经元、权值和所述权值的连接关系数据；所述映射单元对输入数据进行处理，以得到处理后的输入数据包括；所述映射单元根据所述权值的连接关系数据对所述权值进行处理，以得到处理后的权值。

在本公开的一些实施例中，所述映射单元对输入数据进行处理，以得到处理后的输入数据包括：所述映射单元对至少一个输入神经元进行处理，得到输入神经元的连接关系数据；对至少一个权值进行处理，得到权值的连接关系数据；所述映射单元基于输入神经元的连接关系数据和权值的的连接关系数据得到第三连接关系数据，并根据第三连接关系数据对输入神经元和权值进行处理，得到处理后的输入神经元和处理后的权值。

在本公开的一些实施例中，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以直接索引或者步长索引的形式表示。

在本公开的一些实施例中，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，所述连接关系数据为由 0 和 1 组成的字符串，0 表示所述输入神经元的值为零，1 表示所述输入神经元的值非零；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，所述连接关系数据为值为非零的输入神经元与上一个值为非零的输入神经元之间的距离值组成的字符串。

在本公开的一些实施例中，当所述权值的连接关系数据以直接索引的形式表示时，所述连接关系数据为由 0 和 1 组成的字符串，0 表示所述权值为无连接的权值，1 表示所述权值为有连接的权值；当所述权值的连接关系数据以步长索引的形式表示时，所述连接关系数据为与输出神经元有连接的输入神经元的与上一个与所述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。

在本公开的一些实施例中，对于多层人工神经网络的每一层，所述反向运算包括：权值梯度计算步骤和输入神经元梯度计算步骤；所述权值梯

度计算步骤包括：所述运算单元由人工神经网络每一层的输出神经元梯度和输入神经元得到该层的权值梯度；所述输入神经元梯度计算步骤包括：所述映射单元将输出神经元梯度和处理后的权值存储至存储单元；所述运算单元读取输出神经元梯度和处理后的权值，进行人工神经网络运算，得到输入神经元梯度。

在本公开的一些实施例中，将人工神经网络上一层的输出数据作为下一层的输入数据；或者，对人工神经网络上一层的输出数据进行层间操作，将操作结果作为下一层的输入数据；其中，对于人工神经网络的任意相邻两层：第一层与第二层，在正向运算中，所述上一层指所述第一层，所述下一层指所述第二层；在反向运算中，所述上一层指所述第二层，所述下一层指所述第一层。

本公开的运算方法具有以下有益效果：将输入神经元与权值、输出神经元与输入神经元、输出神经元与权值以一一对应的格式进行存储，进行前向运算和反向运算时，可以直接对一一对应的二者进行运算，无需通过索引数组寻找对应的输入、输出神经元和权值，减少了访问存储器的次数，简化了运算，提高了运算速度和性能。可以对稀疏人工神经网络的连接关系数据进行更新，提高了稀疏人工神经网络运算的精度；采用针对稀疏多层人工神经网络运算的专用指令和运算单元，解决了 CPU 和 GPU 运算性能不足，前端译码开销大的问题，有效提高了对多层人工神经网络运算算法的支持。

附图说明

附图是用来提供对本公开的进一步理解，并且构成说明书的一部分，与下面的具体实施方式一起用于解释本公开，但并不构成对本公开的限制。在附图中：

图 1 是一种人工神经网络的示意图；

图 2 是本公开一实施例的人工神经网络计算装置的结构示意图；

图 3 是图 2 所示实施例的映射关系示意图；

图 4 是本公开一实施例的一种人工神经网络的示意图；

图 5 是图 4 所示的人工神经网络计算装置连接关系示意图；

图 6 是本公开另一实施例的人工神经网络计算装置的结构示意图；

图 7 是图 6 所示实施例的映射关系示意图；

图 8 是本公开又一实施例的人工神经网络计算装置的结构示意图；

图 9 是图 8 所示实施例的映射关系示意图；

图 10 是本公开再一实施例的人工神经网络计算装置的结构示意图；

图 11 是图 10 所示实施例的映射关系示意图；

图 12 是本公开实施例的一种运算方法的正向运算流程图；

图 13 是本公开实施例的一种运算方法的权值更新流程图；

图 14 是本公开实施例的一种运算方法的反向运算流程图；

图 15 是本公开实施例的一种运算方法的流程图。

图 16 是本公开又一实施例的一种运算方法的流程图。

图 17 是本公开又一实施例运算方法相邻两层的运算过程图。

符号说明

1-映射单元；2-存储单元；3-DMA；4-指令缓存；5-控制单元；6-第一输入缓存；7-运算单元；8-第二输入缓存；9-输出缓存。

具体实施方式

为使本公开的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本公开进一步详细说明。

本公开一个实施例提出了一种人工神经网络计算装置，参见图 2a，包括：存储单元 2、映射单元 1、运算单元 7、DMA3、指令缓存 4、第一输入缓存 6、第二输入缓存 8、输出缓存 9 和控制单元 5。

上述映射单元 1，用于接收输入数据之后，对所述输入数据进行处理，以得到处理后的输入数据，所述输入数据包括至少一个输入神经元和至少一个权值，所述处理后的输入数据包括处理后的输入神经元和处理后的权值。

进一步地，上述映射单元 1 将上述处理后的输入神经元中的每个处理后的输入神经元与其对应的处理后的权值作为一个输入神经元-权值对，所有的输入神经元-权值对组成一个数据集，将该数据集存储到上述存储单元 2 中。

具体来说，上述映射单元 1 将无连接的权值删除，保留有连接的权值

作为处理后的权值，映射单元 1 还得到权值的连接关系数据，根据权值的连接关系数据对输入神经元进行处理，得到处理后的输入神经元。

例如，当权值的形式为 w_{ij}，该 w_{ij} 表示第 i 个输入神经元与第 j 个输出神经元之间的权值；映射单元 1 根据权值确定上述权值的连接关系数据，将无连接的权值删除，保留有连接的权值作为处理后的权值。

或者，具体来说，上述映射单元 1 将数值或绝对值小于或等于第一阈值的权值删除，得到处理后的权值，映射单元 1 还得到权值的连接关系数据，根据权值的连接关系数据对输入神经元进行处理，得到处理后的输入神经元。

例如，当权值的形式为 w_{ij}，该 w_{ij} 表示第 i 个输入神经元与第 j 个输出神经元之间的权值；映射单元 1 根据权值确定上述权值的连接关系数据，并将权值或权值的绝对值小于或者等于第一阈值的权值删除，得到处理后的权值。

上述映射单元 1 还可以将数值或绝对值小于或等于第二阈值的输入神经元删除，以得到去零后的输入神经元，作为处理后的输入神经元，映射单元 1 还得到输入神经元的连接关系数据，根据输入神经元的连接关系数据对权值进行处理，得到处理后的权值。

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。

可选地，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。

需要指出的是，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，对于权值连接关系数据，0 表示该权值或该权值的绝对值小于或者等于上述第一阈值，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 表示该权值或该权值的绝对值大于上述第一阈值，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。

对于输入神经元的连接关系数据，0 表示该输入神经元或该输入神经元的绝对值小于或者等于上述第二阈值，1 表示该输入神经元或该输入神经元的绝对值大于上述第二阈值。

对于权值的连接关系数据，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。对于输入神经元，以步长索引表示的连接关系数据以当前非零的输入神经元与上一个非 0 的输入神经元之间的距离值组成的字符串表示。

举例说明，假设上述第一阈值为 0.01，输入神经元为 i1、i2、i3 和 i4，对于输出神经元 o1，权值为 w_{11}，w_{21}，w_{31} 和 w_{41}；对于输出神经元 o2，权值为 w_{12}，w_{22}，w_{32} 和 w_{42}，其中权值 w_{21}，w_{12} 和 w_{42} 的值为 0，其绝对值均小于上述第一阈值 0.01，映射单元 1 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”）。在本实施例中，连接关系数据可以用高维动态数组，可以用链表等等表示。

对于上述输出神经元 o1，上述映射单元 101 将上述 i1 与 w_{11}，i3 与 w_{31} 和 i4 与 w_{41} 分别作为一个输入神经元-权值对，所有的输入神经元-权值对组成一个数据集，并将该数据集存储到上述存储单元 2 中；对于输出神经元 o2，上述映射单元 101 将上述 i2 与 w_{22} 和 i3 与 w_{32} 分别作为一个输入神经元-权值对，所有的输入神经元-权值对组成一个数据集，并将该数据集存储到上述存储单元 2 中。

所述输入神经元-权值对不是一种真正的数据存储结构，仅仅是表示输入神经元和权值的对应关系。例如，输入神经元存储于向量 A 中，权值存储于向量 B 中，向量 A 和 B 的长度相同，向量 A 和 B 的同一位置的分量组合在一起被认为是一个输入神经元-权值对。

在另一个示例中，上述映射单元 101 对至少一个输入神经元进行处理，得到输入神经元的连接关系数据，对至少一个权值进行处理，得到权值的的连接关系数据，并基于输入神经元的连接关系数据和权值的的连接关系数据得到第三连接关系数据，根据第三连接关系数据对输入神经元和权值进行处理，得到处理后的输入神经元和处理后的权值。

输入神经元的连接关系数据为用于表示该至少一个输入神经元中值为非 0 的输入神经元的位置的字符串；权值的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，上述输入神经元的连接关系数据和权值的连接关系数据均可以步长索引或者直接索引的形式表示，具体可参见上一示例中相关的描述。

具体地，当上述输入神经元的连接关系数据和权值的连接关系数据均以直接索引的形式表示时，上述映射单元 1 对上述输入神经元的连接关系数据和权值的连接关系数据进行与操作。

当上述输入神经元的连接关系数据和权值的连接关系数据均以步长索引的形式表示时，上述映射单元 1 从左到右将上述输入神经元的连接关系数据字符串中的每一个元素与位于其之前的元素进行累加，得到的新的元素组成输入神经元的连接关系数据；同理，上述映射单元 1 对上述权值的连接关系数据的字符串进行同样的处理。然后上述映射单元 1 从上述输入神经元的连接关系数据和权值的连接关系数据的字符串中，选取相同的元素，按照其在上述输入神经元的连接关系数据的顺序或权值的连接关系数据的顺序，组成一个新的字符串。上述映射单元 1 将上述新的字符串中将每一个元素与其相邻的前一个元素进行相减，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输入神经元的连接关系数

据和权值的连接关系数据，上述权值的连接关系数据的字符串为“01111”，上述输入神经元的连接关系数据的字符串为“022”，将上述权值的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，对上述输入神经元的连接关系数据的字符串进行相同的处理后得到的第五连接关系数据为“024”。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”。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”。

上述映射单元 1 将上述处理后的输入神经元中的每个处理后的输入神经元与其对应的处理后的权值作为一个输入神经元-权值对，所有的输入神经元-权值对组成一个数据集，将该数据集存储出上述存储单元 2 中。

可以看出，在本公开实施例的方案中，上述映射单元对输入神经元和权值进行处理，剔除值为 0 的输入神经元和权值，减少了输入神经元和权值的数量，减少了额外的开销，运算单元根据处理后的输入神经元和权值进行人工神经网络运算，提高了运算的效率。

在本实施例中，当人工神经网络计算装置用于正向运算时，

映射单元 1 读取存储单元 2 存储的输入神经元和权值并进行处理，获取人工神经网络的连接关系数据。

该连接关系数据可以采用上述直接索引的形式表示。例如对于如图 4 所示的人工神经网络，有 4 个输入神经元：i1、i2、i3、i4，有 2 个输出神经元：o1、o2。对于该人工神经网络，如图 5 所示，o1 与 i1、i3、i4 有连接，o1 与 i2 无连接，则输出神经元 o1 与输入神经元 i1、i2、i3、i4 的连接状态为：1011。类似地，o2 与 i2、i3 有连接，o2 与 i1、i4 无连接，则输出神经元 o2 与输入神经元 i1、i2、i3、i4 的连接状态为：0110，将输出神经元 o1、o2 与所有输入神经元的连接状态依次排列，第一直接索引形式存储的人工神经网络的连接关系数据为 10110110。

该连接关系数据也可以采用第二直接索引形式存储。

例如对于如图 5 所示的人工神经网络，i1 与 o1 有连接、i1 与 o2 无连接，i2 与 o1 无连接、i2 与 o2 有连接，i3 与 o1、o2 有连接，i4 与 o1 有连接、i4 与 o2 无连接，则输入神经元 i1 与输出神经元 o1、o2、o3、o4 的连

接状态为：10，类似地，输入神经元 i2 与输出神经元 o1、o2、o3、o4 的连接状态为：01，输入神经元 i3 与输出神经元 o1、o2、o3、o4 的连接状态为 11，输入神经元 i4 与输出神经元 o1、o2、o3、o4 的连接状态为 10，将输入神经元 i1、i2、i3、i4 与所有输出神经元的连接状态依次排列，第二直接索引形式存储的人工神经网络的连接关系数据为 10011110。

该连接关系数据还可以采用步长索引形式存储。

例如对于如图 4 所示的人工神经网络，输出神经元 o1 与输入神经元 i1、i3、i4 有连接，输出神经元 o1 的连接状态第一个字符为 0，即和 o1 有连接的第一个输入神经元 i1，与人工神经网络的第一个输入神经元，也是 i1，的距离为 0，和 o1 有连接的第二个输入神经元 i3 与和 o1 有连接的第一个输入神经元 i1 的距离为 2，和 o1 有连接的第三个输入神经元 i4 与和 o1 有连接的第二个输入神经元 i3 的距离为 1，因此，输出神经元 o1 的连接状态为 021。输出神经元 o2 与输入神经元 i2、i3 有连接，类似地，输出神经元 o2 的连接状态为 11，则将输出神经元 o1、o2 与所有输入神经元的连接状态依次排列，步长索引形式存储的人工神经网络的连接关系数据为 02111。

得到人工神经网络的连接关系数据后，可以据此得到处理后的输入神经元和权值。映射单元 1 基于连接关系数据，将处理后的权值与输入神经元组成的数据集存储至存储单元 2，处理后的权值是指有连接的权值，或者，是指数值或绝对值大于第一阈值的权值。例如，第一直接索引形式存储的人工神经网络的连接关系数据为 10110110，处理后的权值 w11、w31、w41、w22、w32 对应的输入神经元分别为 i1、i3、i4、i2、i3，则将输入神经元与处理后的权值组成数据集：i1、w11，i3、w31，i4、w41，i2、w22，i3、w32，并存储至存储单元 2。

第二直接索引形式存储的人工神经网络的连接关系数据为 10011110，将输入神经元与处理后的权值组成数据集：i1、w11，i2、w22，i3、w31，i3、w32，i4、w41，并存储至存储单元 2。步长索引形式存储的人工神经网络的连接关系数据为 02111，将输入神经元与处理后的权值组成数据集：i1、w11，i3、w31，i4、w41，i2、w22，i3、w32，并存储至存储单元 2。

在其他示例中，也可以如图 3 所示，映射单元 1 读取存储单元 2 的输

入神经元、权值和连接关系数据并进行处理，映射单元 1 将处理后的权值与输入神经元组成的数据集格式分别存储至第一输入缓存 6 和第二输入缓存 8。

控制单元 5 控制指令缓存 4 读取存储单元 2 中的指令，并基于指令控制运算单元 7 运算。运算单元 7 在控制单元 5 的控制下，将存储单元 2 中存储的数据集的输入神经元和权值分别存入第一输入缓存 6 和第二输入缓存 8，运算单元 7 分别从第一输入缓存 6 和第二输入缓存 8 中读取输入神经元和权值，并进行人工神经网络前向运算，得到输出神经元并存入输出缓存 9，输出缓存 9 中的输出神经元经 DMA3 存入存储单元 2。

运算单元 7 将输入神经元和权值数据一一对应相乘得到加权输出神经元，将加权输出神经元通过加法树逐级相加得到求和输出神经元，或者，将加权输出神经元通过和偏置相加得到加偏置输出神经元，将求和输出神经元或加偏置输出神经元进行激活函数运算，得到人工神经网络的输出神经元，完成人工神经网络的前向运算。激活函数可以是 sigmoid 函数、tanh 函数或 ReLU 函数。

本实施例中，运算单元 7、指令缓存 4、第一输入缓存 6、第二输入缓存 8、输出缓存 9 和控制单元 5 由专用处理器实现，第一输入缓存 6、第二输入缓存 8、输出缓存 9 均为片上缓存。

当人工神经网络计算装置用于权值更新时，

映射单元 1 基于处理后的权值将其对应的有连接的输出神经元梯度与输入神经元存储至存储单元 2。

运算单元 7 在控制单元 5 的控制下，将存储单元 2 中存储的输出神经元梯度和输入神经元分别存入第一输入缓存 6 和第二输入缓存 8，运算单元 7 分别从第一输入缓存 6 和第二输入缓存 8 中读取输出神经元梯度和输入神经元，将输出神经元梯度和输入神经元一一对应相乘，并将相乘结果乘以学习率，得到处理后的权值的更新量；将处理后的权值减去对应的更新量，得到更新的处理后的权值并存入输出缓存 9，输出缓存 9 中的更新的处理后的权值经 DMA3 存入存储单元 2。

例如，第一直接索引形式存储的人工神经网络的连接关系数据为 10110110，处理后的权值 w11、w31、w41、w22、w32 对应的输入神经元

分别为 i1、i3、i4、i2、i3，则将处理后的权值对应的输出神经元梯度与输入神经元组成数据集：To1、i1，To1、i3，To1、i4，To2、i2，To2、i3，并存储至存储单元 2，To1 和 To2 分别表示输出神经元 o1 和 o2 的梯度。To1 与 i1，To1 与 i3，To1 与 i4，To2 与 i2，To2 与 i3 对应相乘，相乘结果再乘以学习率，得到处理后的权值的更新量 Δw11、Δw31、Δw41、Δw22、Δw32，处理后的权值 w11、w31、w41、w22、w32 分别减去更新量，得到更新的处理后的权值 w11’、w31’、w41’、w22’、w32’。

类似地，第二直接索引形式存储的人工神经网络的连接关系数据为 10011110，处理后的权值为 w11，w22，w31，w32，w41 对应的输入神经元分别为 i1、i2、i3、i3、i4，则将输出神经元梯度与输入神经元组成数据集：To1、i1，To2、i2，To1、i3，To2、i3，To1、i4，并存储至存储单元 2。To1 与 i1，To2 与 i2，To1 与 i3，To2 与 i3，To1 与 i4 对应相乘，相乘结果再乘以学习率，得到处理后的权值的更新量 Δw11，Δw22，Δw31，Δw32，Δw41，处理后的权值 w11，w22，w31，w32，w41 分别减去更新量，得到更新的处理后的权值 w11’，w22’，w31’，w32’，w41’。

步长索引形式存储的人工神经网络的连接关系数据为 02111，处理后的权值 w11、w31、w41、w22、w32 对应的输入神经元分别为 i1、i3、i4、i2、i3，则将输出神经元梯度与输入神经元组成数据集：To1、i1，To1、i3，To1、i4，To2、i2，To2、i3，并存储至存储单元 2。To1 与 i1，To1 与 i3，To1 与 i4，To2 与 i2，To2 与 i3 对应相乘，相乘结果再乘以学习率，得到处理后的权值的更新量 Δw11、Δw31、Δw41、Δw22、Δw32，处理后的权值 w11、w31、w41、w22、w32 分别减去更新量，得到更新的处理后的权值 w11’、w31’、w41’、w22’、w32’。

当人工神经网络计算装置用于反向运算时，

映射单元 1 将输出神经元梯度和处理后的权值存储至存储单元 2。

映射单元 1 选取处理后的权值所对应的输出神经元，并将输出神经元梯度与处理后的权值存储至存储单元 2，其中处理后的权值是指在更新前有连接的权值，或者，数值或绝对值大于第一阈值的权值。

例如，第一直接索引形式存储的人工神经网络的连接关系数据为 10110110，将输出神经元梯度与处理后的权值组成数据集：To1，w11，To1，

w31，To1，w41，To2，w22，To2，w32，并存储至存储单元 2。第二直接索引形式存储的人工神经网络的连接关系数据为 10011110，将输出神经元梯度与处理后的权值组成数据集：To1、w11，To2、w22，To1、w31，To2、w32，To1、w41，并存储至存储单元 2。步长索引形式存储的人工神经网络的连接关系数据为 02111，将输出神经元梯度与处理后的权值组成数据集：To1，w11，To1，w31，To1，w41，To2，w22，To2，w32，并存储至存储单元 2。

运算单元 7 从存储单元 2 中读取输出神经元梯度和处理后的权值，并分别存入第一输入缓存 6 和第二输入缓存 8，运算单元 7 读取第一输入缓存 6 和第二输入缓存 8 中的输出神经元梯度和处理后的权值，将输出神经元梯度和处理后的权值一一对应相乘，并将相乘结果通过加法树逐级相加，对相加结果进行激活函数反运算，得到输入神经元梯度并存入输出缓存 9，输出缓存 9 中的输入神经元梯度经 DMA3 存入存储单元 2，完成人工神经网络的反向运算。

本公开的人工神经网络计算装置，将处理后的权值与输入神经元、输出神经元与输入神经元、输出神经元与处理后的权值进行存储，进行前向运算和反向运算时，可以直接对二者进行运算，无需通过索引数组寻找对应的输入、输出神经元和权值，减少了访问存储器的次数，简化了运算，提高了运算速度和性能；并且通过采用针对多层人工神经网络运算的专用指令和运算单元 7，解决了 CPU 和 GPU 运算性能不足，前端译码开销大的问题，有效提高了对多层人工神经网络运算算法的支持；通过采用针对多层人工神经网络运算的专用片上缓存，充分挖掘了输入神经元和权值数据的重用性，避免了反复向内存读取这些数据，降低了内存访问带宽，避免了内存带宽成为多层人工神经网络运算及其训练算法性能瓶颈的问题。

进一步地，上述运算单元 7 包括但不限定于三个部分，分别为乘法器、一个或多个加法器（可选的，多个加法器组成加法树）和激活函数单元。上述乘法器将第一输入数据（in1）和第二输入数据（in2）相乘得到第一输出数据（out1），过程为：out1=in1*in2；上述加法树将第三输入数据（in3）通过加法树逐级相加得到第二输出数据（out2），其中 in3 是一个长度为 N 的向量，N 大于 1，过称为：out2＝in3[1]+in3[2]+...+in3[N]，和/或将第三

输入数据（in3）通过加法树累加之后得到的结果和第四输入数据（in4）相加得到第二输出数据（out2），过程为：out2=in3[1]+in3[2]+...+in3[N]+in4，或者将第三输入数据（in3）和第四输入数据（in4）相加得到第二输出数据（out2），过称为：out2=in3+in4；上述激活函数单元将第五输入数据（in5）通过激活函数（active）运算得到第三输出数据（out3），过程为：out3=active(in5)，激活函数 active 可以是 sigmoid、tanh、relu、softmax 等函数，除了做激活操作，激活函数单元可以实现其他的非线性函数运算，可将输入数据（in）通过函数（f）运算得到输出数据（out），过程为：out=f(in)。

本公开另一个实施例的人工神经网络计算装置，与前述实施例相同的特征不再重复描述。

在反向运算中，映射单元 1 将输出神经元梯度与全部输入神经元存储至存储单元 2。

例如，当采用第一直接索引形式和步长索引形式存储连接关系数据时，输出神经元梯度与输入神经元的数据集为：To1、i1，To1、i2，To1、i3，To1、i4，To2、i1，To2、i2，To2、i3，To2、i4，并存储至存储单元 2，当采用第二直接索引形式存储连接关系数据时，输出神经元梯度与输入神经元的数据集为：To1、i1，To2、i1，To1、i2，To2、i2，To1、i3，To2、i3，To1、i4，To2、i4。

运算单元 7 从存储单元 2 中读取输出神经元梯度和输入神经元，并分别存入第一输入缓存 6 和第二输入缓存 8，运算单元 7 读取第一输入缓存 6 和第二输入缓存 8 的输出神经元梯度和输入神经元，将输出神经元梯度和输入神经元一一对应相乘，并将相乘结果乘以学习率，得到所有权值的更新量；将所有权值减去对应的更新量，得到更新后的所有权值。

运算单元 7 判断更新后的权值是否仍是有连接的权值，若更新后的权值仍是有连接的权值，则将该权值设置为有连接的权值，否则，则将该权值设置为无连接的权值，即该权值对应的输入神经元与输出神经元之间无连接，得到更新的处理后的权值。

或者，运算单元 7 将更新后的权值与第一阈值进行比较，若更新后的权值或权值的绝对值大于第一阈值，则将该权值保留，即该权值对应的输入神经元与输出神经元之间有连接；否则，则将该权值删除，即该权值对

应的输入神经元与输出神经元之间无连接，得到更新的处理后的权值。

映射单元 1 还基于更新的处理后的权值选取其对应的输出神经元，并将处理后的权值与输出神经元梯度存储至存储单元 2，其中处理后的权值是指该处理后的权值在更新之前的值。

运算单元 7 从存储单元 2 中读取输出神经元梯度和处理后的权值，并分别存入第一输入缓存 6 和第二输入缓存 8，运算单元 7 读取第一输入缓存 6 和第二输入缓存 8 的输出神经元梯度和处理后的权值，并进行人工神经网络反向运算，得到输入神经元梯度。

由此可见，本公开还可以对人工神经网络的连接关系数据进行更新，提高了人工神经网络运算的精度。

本公开又一个实施例的人工神经网络计算装置，与前述实施例相同的特征不再重复描述。如图 6 所示，本实施例中，映射单元 1 也位于专用处理器片内。

当人工神经网络计算装置用于正向运算时，

映射单元 1 读取存储单元 2 存储的输入神经元和权值并进行处理，得到输入神经元-权值对，所有的输入神经元-权值对组成一个数据集，数据集分别存储至第一输入缓存 6 和第二输入缓存 8。在其他示例中，也可以如图 7 所示，映射单元 1 读取存储单元 2 的输入神经元、权值和连接关系数据并进行处理，得到输入神经元-权值对，所有的输入神经元-权值对组成一个数据集，数据集分别存储至第一输入缓存 6 和第二输入缓存 8。

当人工神经网络计算装置用于权值更新时，

映射单元 1 基于处理后的权值将其对应的有连接的输出神经元梯度与输入神经元分别存储至第一输入缓存 6 和第二输入缓存 8。

当人工神经网络计算装置用于反向运算时，

映射单元 1 将输出神经元梯度和处理后的权值分别存储至第一输入缓存 6 和第二输入缓存 8。

本公开再一个实施例的人工神经网络计算装置，与前述实施例相同的特征不再重复描述。如图 8 所示，本实施例中，映射单元 1 也位于专用处理器片内且连接第一输入缓存 6，不连接第二输入缓存 8。

当人工神经网络计算装置用于正向运算时，

映射单元 1 读取存储单元 2 存储的输入神经元和输入神经元的连接关系数据并进行处理，并将处理后的输入神经元存储至第一输入缓存 6；存储单元 2 将处理后的权值存储至第二输入缓存 8，如图 9 所示。

当人工神经网络计算装置用于权值更新时，

映射单元 1 基于处理后的权值将其对应的输出神经元梯度存储至第一输入缓存 6，存储单元 2 直接将与输出神经元梯度有连接的输入神经元存储至第二输入缓存 8。

当人工神经网络计算装置用于反向运算时，

映射单元 1 将输出神经元梯度存储至第一输入缓存 6，存储单元 2 直接将处理后的权值存储至第二输入缓存 8。

本公开再一个实施例的人工神经网络计算装置，与前述实施例相同的特征不再重复描述。如图 10 所示，本实施例中，映射单元 1 也位于专用处理器片内且连接第二输入缓存 8，不连接第一输入缓存 6。

当人工神经网络计算装置用于正向运算时，

映射单元 1 读取存储单元 2 存储的权值和权值的连接关系数据并进行处理，并将处理后的权值存储至第二输入缓存 8；存储单元 2 将处理后的输入神经元存储至第一输入缓存 6。映射单元 1 读取存储单元 2 存储的权值和权值的连接关系数据，对权值和权值的连接关系数据进行处理，并将处理后的权值存储至第二输入缓存 8；存储单元 2 将处理后的输入神经元存储至第一输入缓存 6，如图 11 所示。

当人工神经网络计算装置用于权值更新时，

映射单元 1 基于处理后的权值将其对应的输入神经元存储至第一输入缓存 6，存储单元 2 直接将与输入神经元有连接的输出神经元梯度存储至第二输入缓存 8。

当人工神经网络计算装置用于反向运算时，

映射单元 1 将处理后的权值存储至第一输入缓存 6，存储单元 2 直接将输出神经元梯度存储至第二输入缓存 8。

需要说明的是，上述人工神经网络计算装置不仅可以进行稀疏神经网络运算，还可以进行稠密神经网络运算。上述人工神经网络计算装置特别适用于稀疏神经网络的运算，是因为稀疏神经网络里 0 值数据或者绝对值

很小的数据非常多。通过映射单元可以提出这些数据，在保证运算精度的情况下，可提高运算的效率。

稀疏神经网络的定义为：在神经网络中，值为 0 的权值的数目很多，并且值为非 0 的权值分布没有规律，则称该神经网络为稀疏神经网络。定义权值为0的元素数量与神经网络所有权值元素数量的比值为神经网络的稀疏度。

本公开再一个实施例还提出了一种运算方法，其利用上述人工神经网络计算装置进行人工神经网络计算，运算方法包括人工神经网络的正向运算，如图 12，包括：

映射单元 1 对输入神经元和权值进行处理，获取人工神经网络的连接关系数据；

映射单元 1 基于连接关系数据得到输入神经元-权值对，所有的输入神经元-权值对组成一个数据集，数据集存储至存储单元 2，处理后的权值是指有连接的权值，或者，是指数值或绝对值大于第一阈值的权值。

运算单元 7 读取输入神经元和处理后的权值，进行人工神经网络前向运算，得到输出神经元。

运算方法包括人工神经网络的权值更新，如图 13，包括：

映射单元 1 基于处理后的权值将其对应的有连接的输出神经元梯度与输入神经元存储至存储单元 2。

运算单元 7 读取输出神经元梯度和输入神经元，对处理后的权值进行更新。

具体来说，将输出神经元梯度与输入神经元存储至存储单元；将存储单元中存储的输出神经元梯度和输入神经元分别存入第一输入缓存和第二输入缓存。

从第一输入缓存和第二输入缓存中读取输出神经元梯度和输入神经元；将输出神经元梯度和输入神经元一一对应相乘，并将相乘结果乘以学习率，得到处理后的权值的更新量；将处理后的权值减去对应的更新量，得到更新的处理后的权值。

运算方法包括人工神经网络的反向运算，如图 14，包括：

映射单元 1 将输出神经元梯度和处理后的权值存储至存储单元 2。

具体来说，选取处理后的权值所对应的输出神经元，并将输出神经元梯度与处理后的权值存储至存储单元 2；从存储单元 2 中读取输出神经元梯度和处理后的权值，并分别存入第一输入缓存 6 和第二输入缓存 8。

运算单元 7 读取输出神经元梯度和处理后的权值，进行人工神经网络反向运算，得到输入神经元梯度。

具体来说，运算单元 7 读取第一输入缓存 6 和第二输入缓存 8 中的输出神经元梯度和处理后的权值；将输出神经元梯度和有连接的权值一一对应相乘，并将相乘结果通过加法树逐级相加；对相加结果进行激活函数反运算，得到输入神经元梯度。激活函数包括 sigmoid 函数、tanh 函数或 ReLU 函数。

其中，可以仅存储处理后的权值，也可以将所有权值均进行存储，当存储所有权值时，

映射单元 1 将输出神经元梯度与输入神经元存储至存储单元 2。

运算单元 7 读取输出神经元梯度和输入神经元，对处理后的权值进行更新。

权值的更新量包括所有权值的更新量，将所有权值减去对应的更新量，得到更新后的所有权值。对更新后的所有权值进行判断，若更新后的权值为有连接的权值，则将该权值设置为有连接的权值，否则，将该权值设置为无连接的权值。或者，将更新后的所有权值与门限值进行比较，若更新后的权值或权值的绝对值大于第一阈值，则将保留该权值，否则，将该权值删除。

映射单元 1 还基于更新的处理后的权值选取其对应的输出神经元，并将处理后的权值与输出神经元梯度存储至存储单元 2，其中处理后的权值是指该处理后的权值在更新之前的值。

运算单元 7 读取输出神经元梯度和处理后的权值，进行人工神经网络反向运算，得到输入神经元梯度。

运算方法可以包括人工神经网络的正向运算、权值更新和反向运算，如图 15 所示。

本公开又一个实施例还提出了一种运算方法，具体来说是一种利用上述人工神经网络计算装置进行人工神经网络训练的方法，参见图 16，运算

方法包括：

正向运算步骤：首先，对多层人工神经网络中的各层依次执行正向运算，得到各层的输出神经元。

对于人工神经网络的每一层，可以按照上一实施例的正向运算步骤进行该层的正向运算。参见图 17，在该步骤中，当上一层人工神经网络的正向运算执行完毕后，在运算指令的控制下，运算单元会将上一层人工神经网络的输出数据作为下一层人工神经网络的输入数据，即将上一层人工神经网络的输出神经元作为下一层人工神经网络的输入神经元，或者是对上一层人工神经网络的输出神经元进行层间操作，将操作结果作为下一层人工神经网络的输入神经元。得到输入神经元后，下一层人工神经网络进行正向运算，同时，将权值也替换为下一层人工神经网络的权值。其中该层间操作可以是向量运算、非线性运算。

反向运算步骤：然后，按照与正向运算相反的顺序，对多层人工神经网络中的各层依次执行反向运算，得到各层的权值梯度和输入神经元梯度。

该步骤包括第一运算部分和第二运算部分。第一运算部分用于计算权值梯度。对于人工神经网络的每一层，由该层的输出神经元梯度和输入神经元通过矩阵乘法或卷积得到该层的权值梯度。

第二运算部分用于计算输入神经元梯度。对于人工神经网络的每一层，可以按照上一实施例的反向运算步骤进行该层的反向运算。参见图 17，在该步骤中，当上一层人工神经网络的反向运算执行完毕后，在运算指令的控制下，运算单元会将上一层人工神经网络的输出数据作为下一层人工神经网络的输入数据，但在反向运算中，上一层人工神经网络的输出数据为输入神经元梯度，下一层人工神经网络的输入数据为输出神经元梯度，即上一层人工神经网络的输入神经元梯度作为下一层人工神经网络的输出神经元梯度，或者是对上一层人工神经网络的输入神经元梯度进行层间操作，将操作结果作为下一层人工神经网络的输出神经元梯度。得到输出神经元梯度后，下一层人工神经网络进行反向运算，同时，将权值也替换为下一层人工神经网络的权值。

需要说明的是，在反向运算中，所述的“上一层”和“下一层”与正向运算相反，即对于人工神经网络的相邻两层：第一层与第二层，如果在

正向运算中，第一层是“上一层”，第二层是“下一层”，则在反向运算中，第一层是“下一层”，第二层是“上一层”。

权值更新步骤：接着，对多层人工神经网络中的各层进行权值更新，得到更新后的权值。

在执行完反向运算之后，就计算出了各层的权值梯度。在这个步骤中，对于人工神经网络的每一层，利用第一输入缓存和第二输入缓存分别存储该层的权值和权值梯度，然后在运算单元中利用权值梯度对权值进行更新，得到更新后的权值。

重复执行正向运算步骤、反向运算步骤和权值更新步骤多次，完成多层人工神经网络的训练。

整个训练方法需要多次重复执行上述过程，直至人工神经网络的参数达到要求，训练过程完毕。

举例来说，假设神经网络包括三层：输入层、中间层和输出层，其中：输入层各神经元负责接收来自外界的输入数据，并传递给中间层各神经元；中间层负责信息变换，根据信息变化能力的需求，中间层可以设计为单隐层或者多隐层结构，最后一个隐层传递到输出层各神经元的信息，经进一步处理后，完成一次学习的正向传播处理过程；输出层向外界输出信息处理结果。误差的反向传播阶段：误差通过输出层，按误差梯度下降的方式修正各层权值，向中间层、输入层逐层反传。

本发明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神经元和输出层中神经元，而是对于网络中任意相邻的两层，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，设一个卷积神经网络有L层，$K=1,2,...,L-1$，对于第K层和第$K+1$层来说，我们将第K层称为输入层，其中的神经元为所述输入神经元，第$K+1$层称为输出层，其中的神经元为所述输出神经元。即除最顶层外，每一层都可以作为输入层，其下一层为对应的输出层。

上文中提到的运算都是神经网络中的一层的运算，对于多层神经网络，其实现过程是，在正向运算中，当上一层人工神经网络执行完成之后，下一层的运算指令会将运算单元中计算出的输出神经元作为下一层的输入

神经元进行运算（或者是对该输出神经元进行某些操作再作为下一层的输入神经元），同时，将权值也替换为下一层的权值；在反向运算中，当上一层人工神经网络的反向运算执行完成后，下一层运算指令会将运算单元中计算出的输入神经元梯度作为下一层的输出神经元梯度进行运算（或者是对该输入神经元梯度进行某些操作再作为下一层的输出神经元梯度），同时将权值替换为下一层的权值。如图 17 所示，图中虚线的箭头表示反向运算，实线的箭头表示正向运算。

需要说明的是，在训练过程中，反向运算所述的“上一层”和“下一层”与正向运算正好相反，即对于人工神经网络的相邻两层：第一层与第二层，如果在正向运算中，第一层是“上一层”，第二层是“下一层”，则在反向运算中，第一层是“下一层”，第二层是“上一层”。

需要说明的是，在附图或说明书正文中，未绘示或描述的实现方式，均为所属技术领域中普通技术人员所知的形式，并未进行详细说明。此外，上述对各元件的定义并不仅限于实施例中提到的各种具体结构、形状，本领域普通技术人员可对其进行简单地更改或替换；本文可提供包含特定值的参数的示范，但这些参数无需确切等于相应的值，而是可在可接受的误差容限或设计约束内近似于相应值；实施例中提到的方向用语，例如“上”、“下”、“前”、“后”、“左”、“右”等，仅是参考附图的方向，并非用来限制本公开的保护范围；上述实施例可基于设计及可靠度的考虑，彼此混合搭配使用或与其他实施例混合搭配使用，即不同实施例中的技术特征可以自由组合形成更多的实施例。

各功能单元/模块都可以是硬件，比如该硬件可以是电路，包括数字电路，模拟电路等等。硬件结构的物理实现包括但不局限于物理器件，物理器件包括但不局限于晶体管，忆阻器等等。所述计算装置中的计算模块可以是任何适当的硬件处理器，比如 CPU、GPU、FPGA、DSP 和 ASIC 等等。所述存储单元可以是任何适当的磁存储介质或者磁光存储介质，比如 RRAM、DRAM、SRAM、EDRAM、HBM、HMC 等等。

本公开实施例还提供一种计算机存储介质，其中，该计算机存储介质可存储有程序，该程序执行时包括上述方法实施例中记载的任何一种神经网络处理方法的部分或全部步骤。

需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本公开并不受所描述的动作顺序的限制，因为依据本公开，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本公开所必须的。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

在本申请所提供的几个实施例中，应该理解到，所揭露的装置，可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，在本公开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。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元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器中。基于这样的理解，本公开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全部或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器中，包括若干指令用以使得一台计算机设备（可为个人计算机、服务器或者网络设备等）执行本公开各个实施例所述方法的全部或部分步骤。而前述的存储器包括：U 盘、只读存储器（ROM，Read-Only

Memory）、随机存取存储器（RAM，Random Access Memory）、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读存储器中，存储器可以包括：闪存盘、只读存储器（英文：Read-Only Memory ，简称：ROM）、随机存取器（英文：Random Access Memory，简称：RAM）、磁盘或光盘等。

以上对本公开实施例进行了详细介绍，本文中应用了具体个例对本公开的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本公开的方法及其核心思想；同时，对于本领域的一般技术人员，依据本公开的思想，在具体实施方式及应用范围上均会有改变之处，综上上述，本说明书内容不应理解为对本公开的限制。

权利要求

1、一种运算方法，其特征在于，使用人工神经网络计算装置对多层人工神经网络进行训练，所述人工神经网络计算装置包括：映射单元、存储单元和运算单元，所述运算方法包括：

对多层人工神经网络中的各层依次执行正向运算，对于所述多层人工神经网络的每一层，所述正向运算包括：

所述映射单元对输入数据进行处理，以得到处理后的输入数据；

所述存储单元存储所述处理后的输入数据；

所述运算单元读取所述处理后的输入数据，进行人工神经网络运算，得到人工神经网络的输出数据；

按照与所述正向运算相反的顺序，对所述多层人工神经网络中的各层依次执行反向运算；

对所述多层人工神经网络中的各层进行权值更新；

重复执行上述各个步骤多次，完成多层人工神经网络的训练。

2、如权利要求 1 所述的运算方法，其特征在于，所述人工神经网络计算装置还包括：第一输入缓存和第二输入缓存；所述对多层人工神经网络中的各层进行权值更新包括：

对于所述人工神经网络的每一层，所述第一输入缓存和所述第二输入缓存分别存储该层的权值和权值梯度；

所述运算单元从所述第二输入缓存接收所述权值梯度，然后使用所述权值梯度对所述权值进行更新，得到更新后的权值。

3、根据权利要求 1 或 2 所述的运算方法，其特征在于，

所述输入数据包括：输入神经元和权值；

所述映射单元对输入数据进行处理，以得到处理后的输入数据包括：

所述映射单元对权值进行处理，得到处理后的权值和权值的连接关系；

所述映射单元基于所述权值的连接关系数据对输入神经元进行处理，得到处理后的输入神经元。

4、根据权利要求 1 或 2 所述的运算方法，其特征在于，所述输入数据包括输入神经元和权值；

所述映射单元对输入数据进行处理，以得到处理后的输入数据包括：

所述映射单元对输入神经元进行处理，得到处理后的输入神经元和输入神经元的连接关系；

所述映射单元基于所述输入神经元的连接关系数据对权值进行处理，得到处理后的权值。

5、根据权利要求 3 所述的运算方法，其特征在于，所述映射单元对权值进行处理包括：

所述映射单元将无连接的权值删除，保留有连接的权值作为处理后的权值。

6、根据权利要求 3 所述的运算方法，其特征在于，所述映射单元对权值进行处理包括：

所述映射单元将所述权值中数值或绝对值小于或者等于第一阈值的权值删除，得到处理后的权值。

7、根据权利要求 4 所述的运算方法，其特征在于，所述映射单元对输入神经元进行处理包括：

所述映射单元将所述输入神经元中数值或绝对值小于或者等于第二阈值的输入神经元删除，得到处理后的输入神经元。

8、根据权利要求 1 或 2 所述的运算方法，其特征在于，所述输入数据包括处理后的权值、输入神经元和所述输入神经元的连接关系数据；

所述映射单元对输入数据进行处理，以得到处理后的输入数据包括；

所述映射单元根据所述输入神经元的连接关系数据对所述输入神经元进行处理，以得到处理后的输入神经元。

9、根据权利要求 1 或 2 所述的运算方法，其特征在于，所述输入数据包括处理后的输入神经元、权值和所述权值的连接关系数据；

所述映射单元对输入数据进行处理，以得到处理后的输入数据包括；

所述映射单元根据所述权值的连接关系数据对所述权值进行处理，以得到处理后的权值。

10、根据权利要求 1 或 2 所述的运算方法，其特征在于，所述映射单元对输入数据进行处理，以得到处理后的输入数据包括：

所述映射单元对至少一个输入神经元进行处理，得到输入神经元的连

接关系数据；对至少一个权值进行处理，得到权值的连接关系数据；

所述映射单元基于输入神经元的连接关系数据和权值的的连接关系数据得到第三连接关系数据，并根据第三连接关系数据对输入神经元和权值进行处理，得到处理后的输入神经元和处理后的权值。

11、根据权利要求 3-10 任一项所述的运算方法，其特征在于，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以直接索引或者步长索引的形式表示。

12、根据权利要求 11 所述的运算方法，其特征在于，

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，所述连接关系数据为由 0 和 1 组成的字符串，0 表示所述输入神经元的值为零，1 表示所述输入神经元的值非零；

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，所述连接关系数据为值为非零的输入神经元与上一个值为非零的输入神经元之间的距离值组成的字符串。

13、根据权利要求 11 所述的运算方法，其特征在于，

当所述权值的连接关系数据以直接索引的形式表示时，所述连接关系数据为由 0 和 1 组成的字符串，0 表示所述权值为无连接的权值，1 表示所述权值为有连接的权值；

当所述权值的连接关系数据以步长索引的形式表示时，所述连接关系数据为与输出神经元有连接的输入神经元的与上一个与所述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。

14、如权利要求 1 至 13 任一项所述的运算方法，其特征在于，对于多层人工神经网络的每一层，所述反向运算包括：权值梯度计算步骤和输入神经元梯度计算步骤；

所述权值梯度计算步骤包括：

所述运算单元由人工神经网络每一层的输出神经元梯度和输入神经元得到该层的权值梯度；

所述输入神经元梯度计算步骤包括：

所述映射单元将输出神经元梯度和处理后的权值存储至存储单元；

所述运算单元读取输出神经元梯度和处理后的权值，进行人工神经网

络运算，得到输入神经元梯度。

15、如权利要求 1 至 14 任一项所述的运算方法，其特征在于，将人工神经网络上一层的输出数据作为下一层的输入数据；或者，对人工神经网络上一层的输出数据进行层间操作，将操作结果作为下一层的输入数据；其中，

对于人工神经网络的任意相邻两层：第一层与第二层，在正向运算中，所述上一层指所述第一层，所述下一层指所述第二层；在反向运算中，所述上一层指所述第二层，所述下一层指所述第一层。

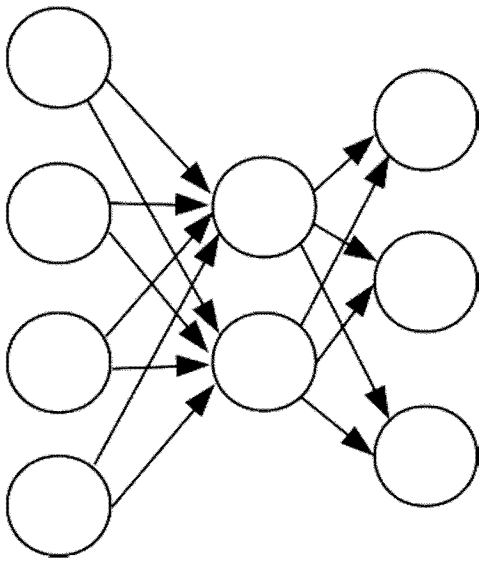

图 1

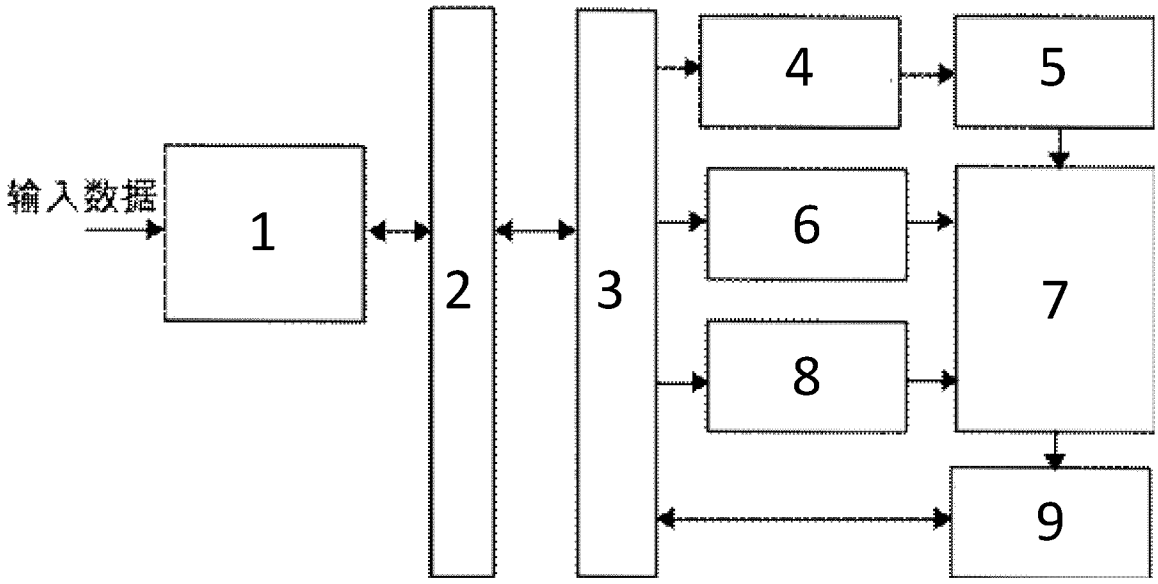

图 2

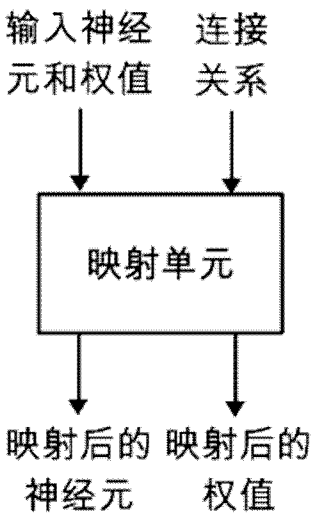

图 3

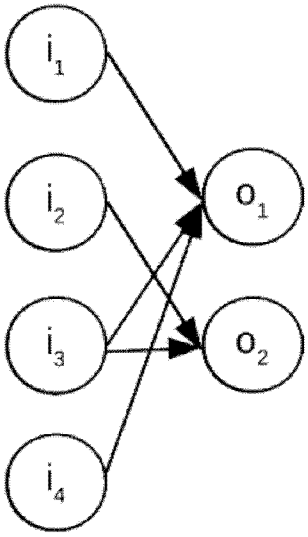

图 4

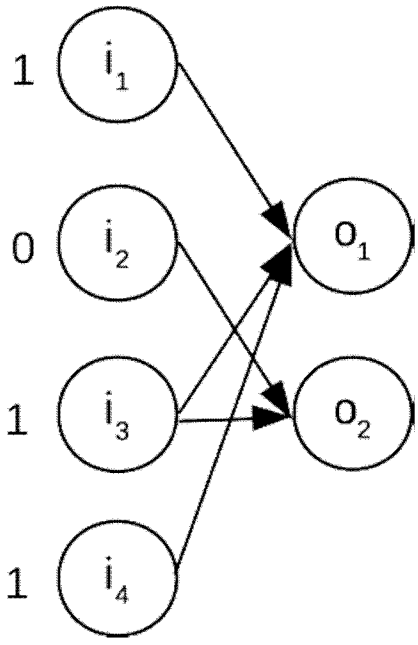

图 5

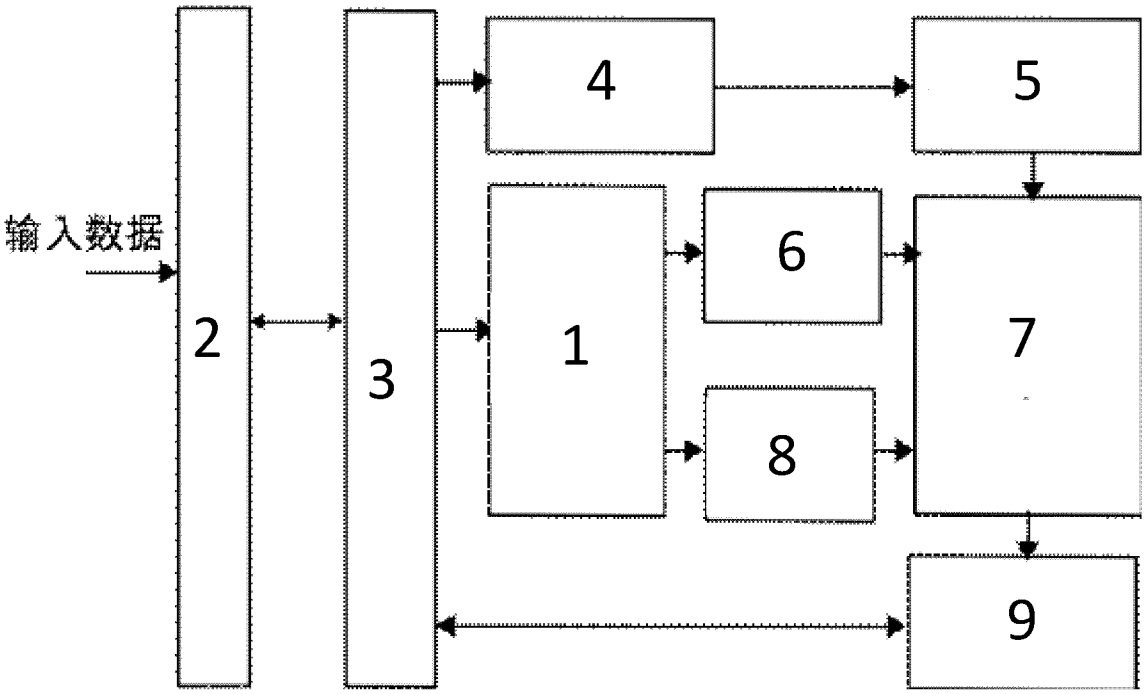

图 6

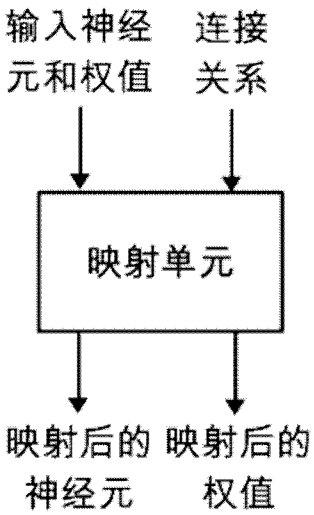

图 7

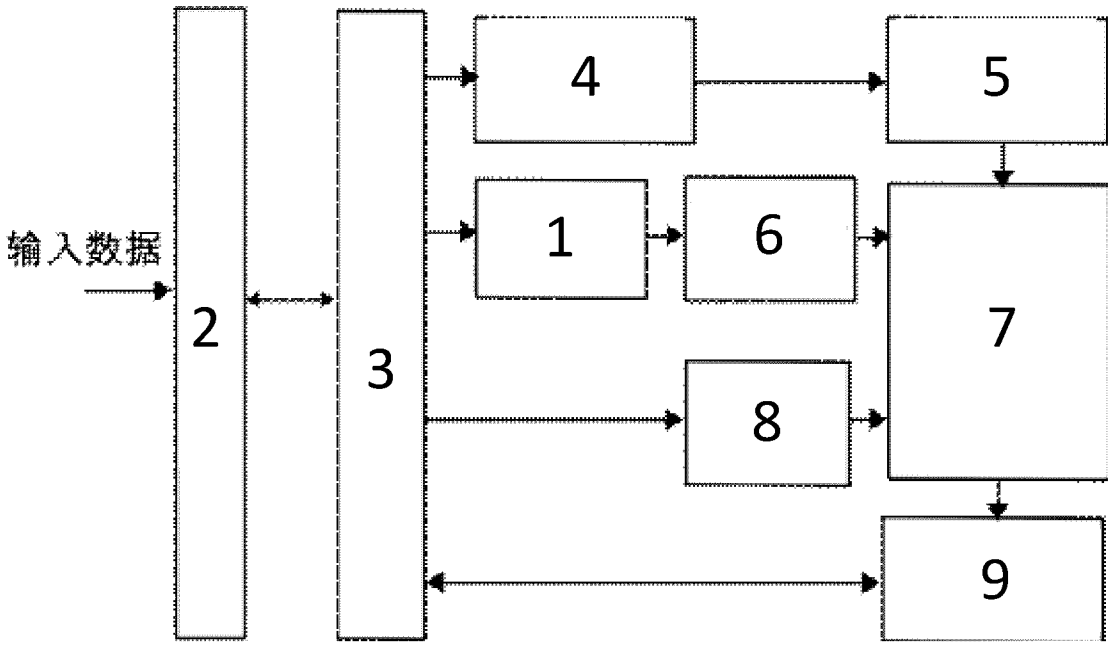

图 8

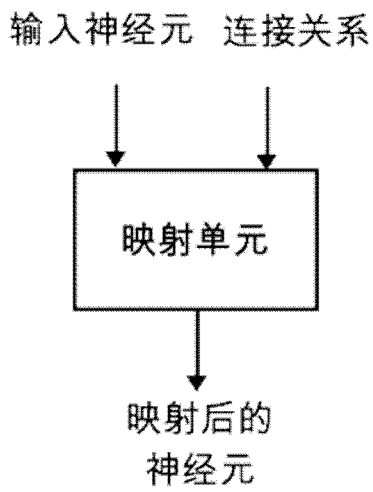

图 9

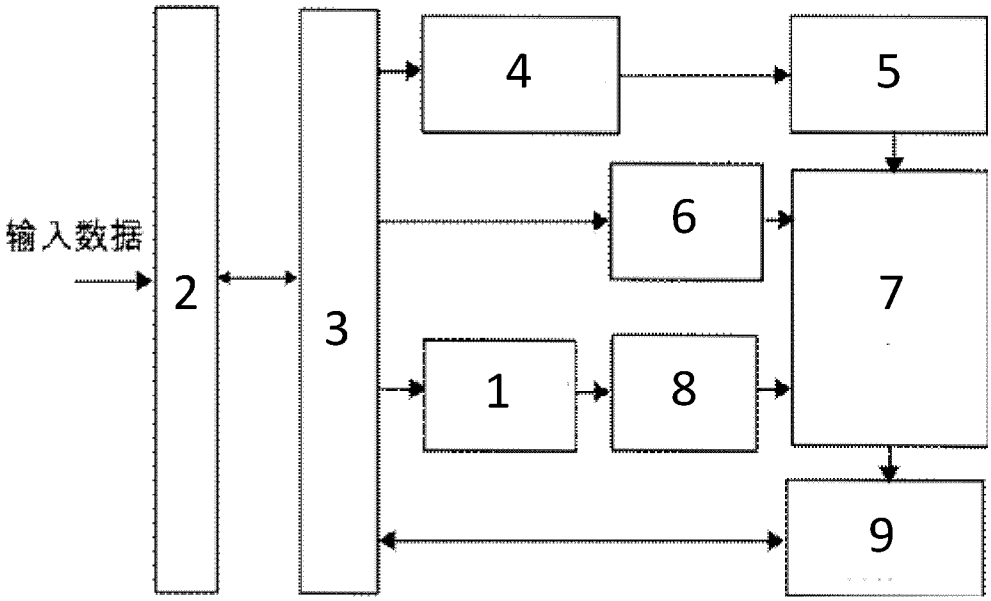

图 10

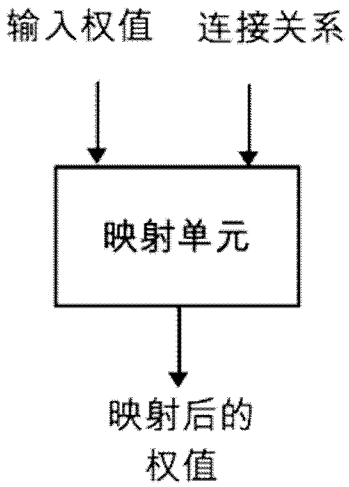

图 11

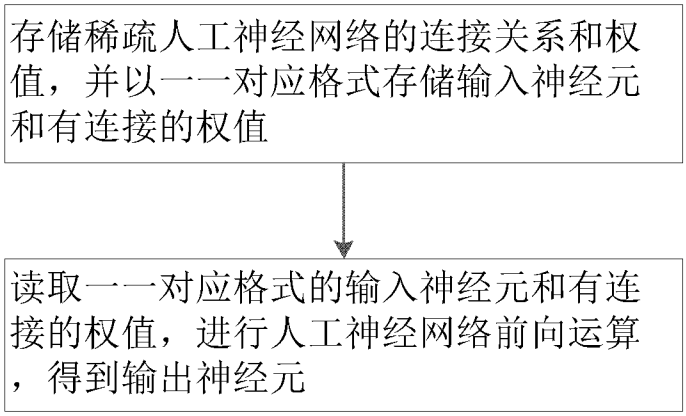
存储稀疏人工神经网络的连接关系和权值，并以一一对应格式存储输入神经元和有连接的权值
读取一一对应格式的输入神经元和有连接的权值，进行人工神经网络前向运算，得到输出神经元

图 12

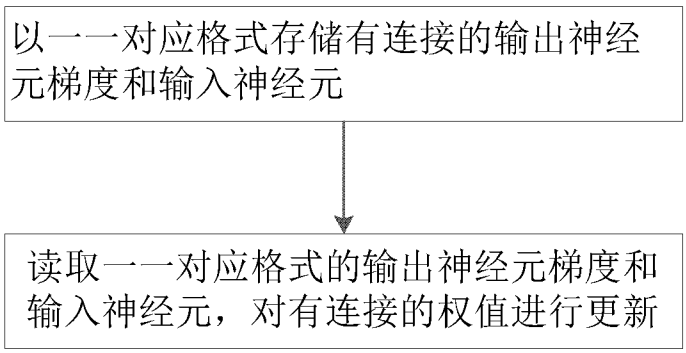
以一一对应格式存储有连接的输出神经元梯度和输入神经元
读取一一对应格式的输出神经元梯度和输入神经元，对有连接的权值进行更新

图 13

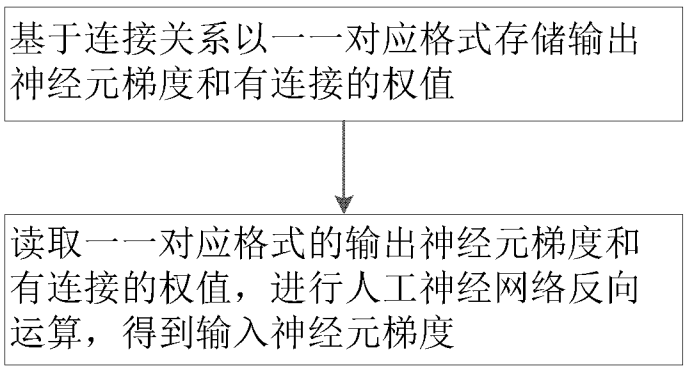
基于连接关系以一一对应格式存储输出神经元梯度和有连接的权值
读取一一对应格式的输出神经元梯度和有连接的权值，进行人工神经网络反向运算，得到输入神经元梯度

图 14

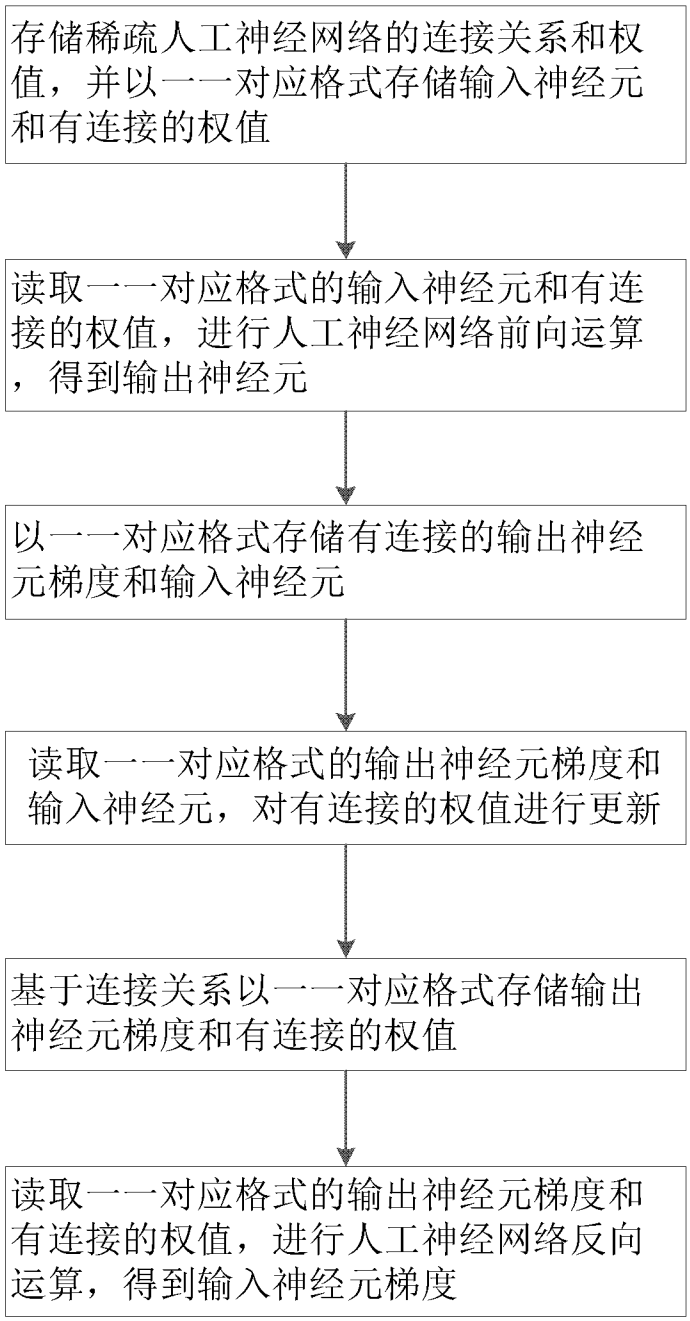
存储稀疏人工神经网络的连接关系和权值，并以一一对应格式存储输入神经元和有连接的权值
读取一一对应格式的输入神经元和有连接的权值，进行人工神经网络前向运算，得到输出神经元
以一一对应格式存储有连接的输出神经元梯度和输入神经元
读取一一对应格式的输出神经元梯度和输入神经元，对有连接的权值进行更新
基于连接关系以一一对应格式存储输出神经元梯度和有连接的权值
读取一一对应格式的输出神经元梯度和有连接的权值，进行人工神经网络反向运算，得到输入神经元梯度

图 15

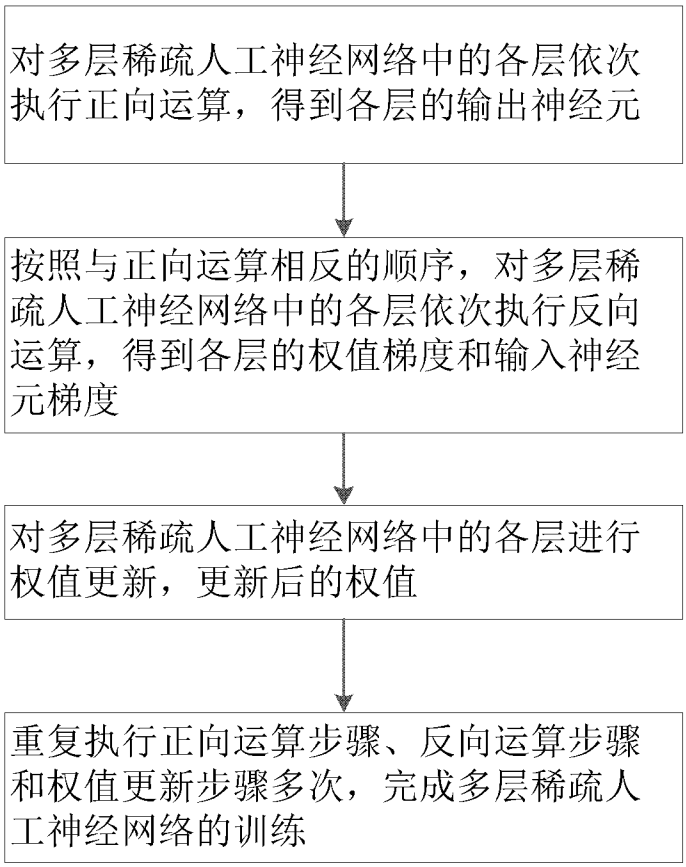
对多层稀疏人工神经网络中的各层依次执行正向运算，得到各层的输出神经元
按照与正向运算相反的顺序，对多层稀疏人工神经网络中的各层依次执行反向运算，得到各层的权值梯度和输入神经元梯度
对多层稀疏人工神经网络中的各层进行权值更新，更新后的权值
重复执行正向运算步骤、反向运算步骤和权值更新步骤多次，完成多层稀疏人工神经网络的训练

图 16

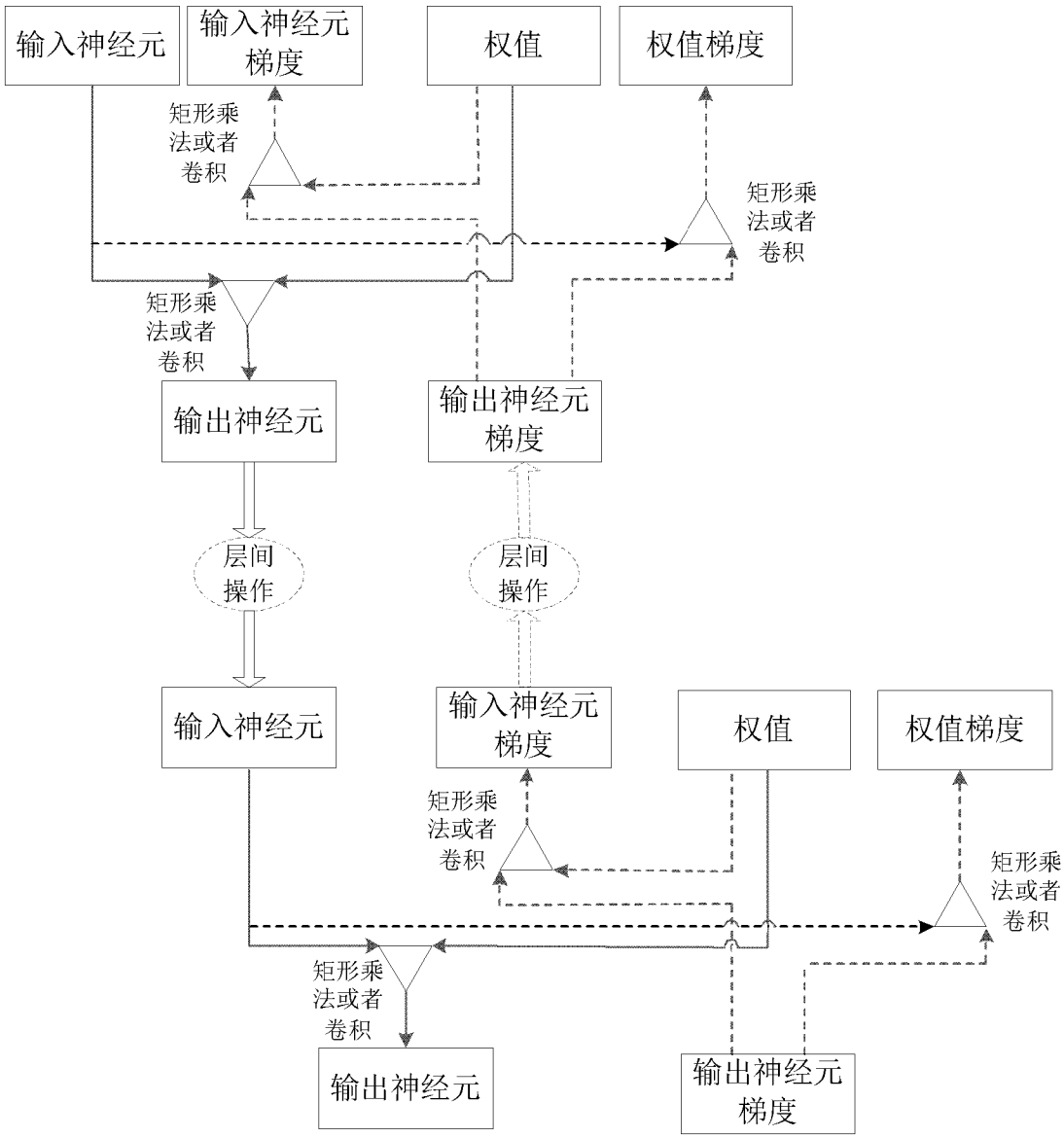
输入神经元
输入神经元梯度
权值
权值梯度
矩形乘法或者卷积
矩形乘法或者卷积
矩形乘法或者卷积
输出神经元
输出神经元梯度
层间操作
层间操作
输入神经元
输入神经元梯度
权值
权值梯度
矩形乘法或者卷积
矩形乘法或者卷积
矩形乘法或者卷积
输出神经元
输出神经元梯度

图 17

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/118264

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)

CNPAT; CNKI; WPI; EPODOC: 神经网络, 权重, 权值, 输入, 输出, 处理, 阈值, CNN, input, output, process

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104809426 A (NEC CORPORATION), 29 July 2015 (29.07.2015), description, paragraphs [0013]-[0023]	1-15
A	CN 104978601 A (TENCENT INC.), 14 October 2015 (14.10.2015), entire document	1-15
A	CN 104572504 A (INSPUR (BEIJING) ELECTRONIC INFORMATION INDUSTRY CO., LTD.), 29 April 2015 (29.04.2015), entire document	1-15
A	CN 103455843 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY), 18 December 2013 (18.12.2013), entire document	1-15

☐ 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
19 March 2018	28 March 2018

Name and mailing address of the ISA
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No. (86-10) 62019451

Authorized officer
CONG, Lei

Telephone No. (86-10) 53961305

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/118264

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104809426 A	29 July 2015	None	
CN 104978601 A	14 October 2015	None	
CN 104572504 A	29 April 2015	None	
CN 103455843 A	18 December 2013	None	

国际检索报告

国际申请号
PCT/CN2017/118264

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））

CNPAT；CNKI；WPI；EPODOC:神经网络,权重，权值，输入，输出，处理，阈值，CNN,input,output,process

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 104809426 A (日本电气株式会社) 2015年 7月 29日 (2015-07-29) 说明书第[0013]-[0023]段	1-15
A	CN 104978601 A (深圳市腾讯计算机系统有限公司) 2015年 10月 14日 (2015-10-14) 全文	1-15
A	CN 104572504 A (浪潮北京电子信息产业有限公司) 2015年 4月 29日 (2015-04-29) 全文	1-15
A	CN 103455843 A (华中科技大学) 2013年 12月 18日 (2013-12-18) 全文	1-15

☐ 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
2018年 3月 19日	2018年 3月 28日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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 丛磊 电话号码 (86-10)53961305

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/118264

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 104809426 A	2015年 7月 29日	无	
CN 104978601 A	2015年 10月 14日	无	
CN 104572504 A	2015年 4月 29日	无	
CN 103455843 A	2013年 12月 18日	无	