(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110832507 B
(45)授权公告日 2024.06.28

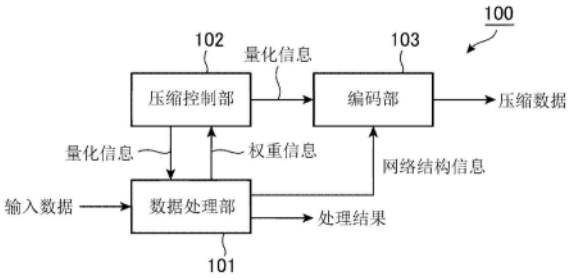
(21)申请号 201780092806.X
(22)申请日 2017.07.07
(65)同一申请的已公布的文献号
申请公布号 CN 110832507 A
(43)申请公布日 2020.02.21
(85)PCT国际申请进入国家阶段日
2019.12.31
(86)PCT国际申请的申请数据
PCT/JP2017/024992 2017.07.07
(87)PCT国际申请的公布数据
WO2019/008752 JA 2019.01.10
(73)专利权人 三菱电机株式会社
地址 日本东京都
(72)发明人 峯泽彰 守屋芳美 王梦雄
杉本和夫
(74)专利代理机构 北京三友知识产权代理有限公司 11127
专利代理师 马建军 邓毅
(51)Int.Cl.
G06N 3/02(2006.01)
H03M 7/30(2006.01)
(56)对比文件
CN 106066783 A,2016.11.02
CN 106485316 A,2017.03.08
Sajid Anwar.Fixed point optimization of deep convolutional neural networks for object recognition.2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).2015,1132-1-1133页.
审查员 李轲
权利要求书1页 说明书12页 附图10页

(54)发明名称

数据处理装置、数据处理方法以及存储介质

(57)摘要

数据处理部(101)使用神经网络来处理输入数据。压缩控制部(102)生成定义了量化步长的量化信息。编码部(103)对网络结构信息和量化信息进行编码而生成压缩数据,网络结构信息包含以由压缩控制部(102)决定的量化步长量化后的参数数据。



CN 110832507 B

1.一种数据处理装置,其特征在于,该数据处理装置具备:

数据处理部,其使用神经网络来处理输入数据;

压缩控制部,其决定对所述神经网络的参数数据进行量化时的量化步长,生成定义了量化步长的量化信息;以及

编码部,其对网络结构信息和所述量化信息进行编码而生成压缩数据,所述网络结构信息包含以由所述压缩控制部决定的量化步长量化后的参数数据;

其中,所述压缩控制部按照每个边、每个节点、每个核心或神经网络的每个层切换量化步长;

所述编码部对定义了每个所述边、每个所述节点、每个所述核心或神经网络的每个所述层的量化步长的所述量化信息进行编码。

2.一种数据处理装置,其特征在于,该数据处理装置具备:

数据处理部,其使用神经网络来处理输入数据;以及

解码部,其对压缩数据进行解码,该压缩数据是对定义了对所述神经网络的参数数据进行量化时的量化步长的量化信息和包含以所述量化信息中的量化步长量化后的参数数据的网络结构信息进行编码而成的,所述量化步长按照每个边、每个节点、每个核心或神经网络的每个层被切换,所述量化信息是定义了每个所述边、每个所述节点、每个所述核心或神经网络的每个所述层的量化步长的量化信息;

所述数据处理部使用由所述解码部从压缩数据解码出的所述量化信息和所述网络结构信息对参数数据进行逆量化,使用包含逆量化后的参数数据的所述网络结构信息构成所述神经网络。

3.根据权利要求1或2所述的数据处理装置,其特征在于,

所述神经网络的参数数据是对所述神经网络中的连接节点之间的边赋予的权重信息。

4.一种数据处理方法,其特征在于,该数据处理方法具备如下步骤:

解码部对压缩数据进行解码,该压缩数据是对定义了对神经网络的参数数据进行量化时的量化步长的量化信息和包含以所述量化信息中的量化步长量化后的参数数据的网络结构信息进行编码而成的;所述量化步长按照每个边、每个节点、每个核心或神经网络的每个层被切换,所述量化信息是定义了每个所述边、每个所述节点、每个所述核心或神经网络的每个所述层的量化步长的量化信息;以及

数据处理部使用由所述解码部从压缩数据解码出的所述量化信息和所述网络结构信息对参数数据进行逆量化,使用包含逆量化后的参数数据的所述网络结构信息构成所述神经网络,使用该神经网络来处理输入数据。

5.一种存储有压缩数据的存储介质,该压缩数据是对定义了对神经网络的参数数据进行量化时的量化步长的量化信息和包含以所述量化信息中的量化步长量化后的参数数据的网络结构信息进行编码而成的,所述量化步长按照每个边、每个节点、每个核心或神经网络的每个层被切换,所述量化信息是定义了每个所述边、每个所述节点、每个所述核心或神经网络的每个所述层的量化步长的量化信息;

其特征在于,

数据处理装置使用解码出的所述量化信息和所述网络结构信息对参数数据进行逆量化,使用包含逆量化后的参数数据的所述网络结构信息构成所述神经网络。

数据处理装置、数据处理方法以及存储介质

技术领域

[0001] 本发明涉及对与神经网络的结构相关的信息进行编码并压缩的数据处理装置、数据处理方法以及存储有压缩数据的存储介质。

背景技术

[0002] 作为解决输入数据的分类问题以及回归问题的方法,具有机器学习。

[0003] 机器学习具有对脑的神经回路(神经元)进行模拟的神经网络这样的手法。在神经网络中,利用通过神经元相互连接而成的网络来表现的概率模型(识别模型、生成模型)进行输入数据的分类(识别)或回归。

[0004] 进而,在不仅具有全连接层(Fully-connected Layer)而且具有卷积层(Convolution Layer)和池化层(Pooling Layer)的神经网络即卷积神经网络(Convolutional Neural Network)中,能够生成实现数据的滤波处理的网络等实现分类以及回归以外的数据处理的网络。例如,能够将图像或声音作为输入,通过卷积神经网络实现如下处理:实现输入信号的噪声去除或高品质化等的图像或声音的滤波处理、压缩声音等失去高频的声音的高频复原处理、一部分区域缺损的图像的复原处理(inpainting)、图像的超分辨率处理等。

[0005] 除此以外,近年来还发表了如下的对抗性生成网络(Generative Adversarial Network)这样的新神经网络:输入到判定通过生成模型生成的数据是否是真的数据(是否不是通过生成模型生成的数据)的识别模型来进行数据的真伪判定,将生成模型和识别模型组合起来构筑网络,以使生成模型不被识别模型识破生成数据是生成数据,且识别模型识破生成数据是生成数据的方式,对抗性地进行学习,从而高精度地生成生成模型。

[0006] 在这些神经网络中,通过使用大量数据的学习而优化网络参数,从而能够实现高性能化。

[0007] 但是,神经网络的数据大小存在大容量化的倾向,使用神经网络的计算机的计算负载也增加。

[0008] 相对于此,在非专利文献1中,记载有对作为神经网络的参数的边的权重进行标量量化并编码的技术。通过对边的权重进行标量量化并编码,对与边相关的数据的数据大小进行压缩。

[0009] 现有技术文献

[0010] 专利文献

[0011] 非专利文献1:Vincent Vanhoucke,Andrew Senior,Mark Z.Mao,“Improving the speed of neural networks on CPUs”,Proc.Deep Learning and Unsuper vised Feature Learning NIPS Workshop,2011.

发明内容

[0012] 发明要解决的课题

[0013] 然而,对神经网络中的多个边分别赋予的权重的最优值根据网络的学习结果而不同,并非恒定。

[0014] 因此,边的权重的压缩大小产生偏差,在非专利文献1记载的技术中,存在无法实现与神经网络的边相关的参数数据的高压缩这样的课题。

[0015] 本发明用于解决上述课题,其目的在于,得到能够对神经网络的参数数据进行高压缩的数据处理装置、数据处理方法以及存储有压缩数据的存储介质。

[0016] 用于解决课题的手段

[0017] 本发明的数据处理装置具备数据处理部、压缩控制部以及编码部。数据处理部使用神经网络对输入数据进行处理。压缩控制部决定对神经网络的参数数据进行量化时的量化步长,生成定义了量化步长的量化信息。编码部对网络结构信息和量化信息进行编码而生成压缩数据,网络结构信息包含以由压缩控制部决定的量化步长量化后的参数数据。

[0018] 发明效果

[0019] 根据本发明,对定义了对神经网络的参数数据进行量化时的量化步长的量化信息和包含以量化信息中的量化步长量化后的参数数据的网络结构信息进行编码而生成压缩数据。由此,能够对神经网络的参数数据进行高压缩。

[0020] 通过使用从压缩数据解码出的量化信息和网络结构信息,能够在解码侧构成在编码侧优化后的神经网络。

附图说明

[0021] 图1是示出本发明的实施方式1的数据处理装置(编码器)的结构的框图。

[0022] 图2是示出实施方式1的数据处理装置(解码器)的结构的框图。

[0023] 图3A是示出实现实施方式1的数据处理装置的功能的硬件结构的框图。图3B是示出执行实现实施方式1的数据处理装置的功能的软件的硬件结构的框图。

[0024] 图4是示出实施方式1的数据处理装置(编码器)的动作的流程图。

[0025] 图5是示出实施方式1的数据处理装置(解码器)的动作的流程图。

[0026] 图6是示出实施方式1中的神经网络的结构例的图。

[0027] 图7是示出实施方式1中的一维数据的卷积处理的例子的图。

[0028] 图8是示出实施方式1中的二维数据的卷积处理的例子的图。

[0029] 图9是示出神经网络的第1层的层中的每个节点的边的权重信息的矩阵的图。

[0030] 图10是示出神经网络的第1层的层中的每个节点的边的权重信息的量化步长的矩阵的图。

[0031] 图11是示出卷积层中的边的权重信息的矩阵的图。

[0032] 图12是示出卷积层中的边的权重信息的量化步长的矩阵的图。

[0033] 图13是示出实施方式1中的构成量化信息的信息的语法的图。

[0034] 图14是示出实施方式1中的构成量化信息的信息的矩阵单位的语法的图。

[0035] 图15是示出实施方式1中的构成量化信息的信息的层单位的语法的图。

具体实施方式

[0036] 以下,为了更详细地说明本发明,参照附图对用于实施本发明的方式进行说明。

[0037] 实施方式1

[0038] 图1是示出本发明的实施方式1的数据处理装置100的结构的框图。在图1中,数据处理装置100使用已学习的神经网络来处理输入数据,输出处理结果。

[0039] 并且,数据处理装置100作为对量化信息和网络结构信息进行编码的编码器发挥功能,具备数据处理部101、压缩控制部102以及编码部103。

[0040] 数据处理部101使用上述神经网络来处理输入数据。

[0041] 并且,数据处理部101输入由压缩控制部102生成的量化信息,以量化信息中定义的量化步长对神经网络的参数数据进行量化。然后,数据处理部101将包含量化后的上述参数数据的网络结构信息输出到编码部103。

[0042] 在数据处理部101中使用的上述神经网络既可以使用预先确定的网络,也可以通过学习进行参数数据的优化。

[0043] 在对神经网络的参数数据进行学习的情况下,针对预先确定的初始状态(参数数据的初始值)的神经网络,使用作为学习对象的输入数据进行神经网络的学习,然后输入由压缩控制部102生成的量化信息,以量化信息中定义的量化步长对神经网络的参数数据进行量化。

[0044] 然后,将该量化后的神经网络作为接下来的学习的初始状态实施上述的学习和量化。将作为反复进行L次(L为1以上的整数)该学习和量化处理的结果而得到的神经网络作为网络结构信息的一部分而输出到编码部103。

[0045] 另外,在L=1的情况下,不再学习量化后的神经网络,因此,可以说是与使用在数据处理部101中未学习而在外部已学习的神经网络相同的处理。换言之,区别仅在于在数据处理部101中进行学习还是在外部进行学习。

[0046] 网络结构信息是表示神经网络的结构的信息,例如包含网络的层数、每个层的节点数、连接节点之间的边、按照每个边赋予的权重信息、表示节点的输出的活性化函数以及每个层的类别信息(例如卷积层、池化层、全连接层)等。

[0047] 神经网络的参数数据例如有对神经网络中的连接节点之间的边赋予的权重信息。

[0048] 压缩控制部102决定对神经网络的参数数据进行量化时的量化步长,生成定义了量化步长的量化信息。

[0049] 例如,压缩控制部102决定神经网络中的按照每个边、每个节点、每个核心或每个层切换的量化步长。

[0050] 量化信息是定义了对神经网络的参数数据进行量化时的量化步长的信息。量化步长是指对参数数据进行量化时的幅度(量化幅度),量化步长越大,则参数数据被分解得越粗略,因此压缩率越高。量化步长越小,则参数数据被分解得越细致,因此压缩率越低。

[0051] 具体而言,量化值k由下述式(1)表示。

[0052] $k=\mathrm{floor}((x/Q)+d_0)+d_1$ (1)

[0053] 在上述式(1)中,x是量化对象参数的值,Q是量化步长,d_0 ($0\leq d_0<1$)是与各个量化值对应的量化对象值的范围的调整偏移,d_1 ($0\leq d_1<1$)是对量化值进行调整的偏移,floor()表示小数点舍去处理函数。进而,还有设置设针对上述式(1)设定的范围的量化对象值x的量化值k为0的死区的方法。

[0054] 并且,已量化参数的值y如下述式(2)所示。

[0055] $y=kQ$ (2)

[0056] 在设上述说明的量化步长的最小切换单位为边单位到层单位的情况下,基于量化的参数数据的压缩率变高,因此,能够削减编码前的参数数据。

[0057] 编码部103对包含由数据处理部101量化后的参数数据的网络结构信息以及由压缩控制部102生成的量化信息进行编码而生成压缩数据。

[0058] 另外,从数据处理部101输入到编码部103的网络结构信息是包含由数据处理部101以由压缩控制部102决定的量化步长量化后的参数数据的网络结构信息。

[0059] 图2是示出实施方式1的数据处理装置200的结构的框图。在图2中,数据处理装置200使用对压缩数据进行解码而得到的神经网络来处理输入数据,输出处理结果。在处理结果中,与数据处理装置100同样地具有输入数据的分类结果或回归分析结果。

[0060] 数据处理装置200作为从压缩数据解码出量化信息和网络结构信息的解码器发挥功能,具备解码部201以及数据处理部202。

[0061] 解码部201从如上所述由编码部103编码后的压缩数据,解码出量化信息和网络结构信息。

[0062] 由解码部201解码出的网络结构信息包含有对由作为编码器的数据处理装置100侧的学习结果而优化的边的权重信息等的参数数据进行量化的结果(量化值k)。

[0063] 在由解码部201解码出的量化信息中,定义了对参数数据进行量化时的量化步长Q。从对上述参数数据进行量化的结果k和上述量化步长Q,按照上述式(2)解码出已量化参数y。这些解码结果从解码部201输出到数据处理部202。

[0064] 数据处理部202使用神经网络来处理输入数据。

[0065] 并且,数据处理部202使用由解码部201从压缩数据解码出的量化信息和网络结构信息,对作为参数数据的边的权重信息进行逆量化。进而,数据处理部202使用包含逆量化后的参数数据的网络结构信息构成神经网络。

[0066] 这样,数据处理部202使用从压缩数据解码出的信息,构成包含有通过数据处理装置100侧的学习结果而优化的边的权重信息等的参数数据的神经网络,使用该神经网络而处理输入数据。由此,能够通过数据处理装置100和数据处理装置200使优化的参数数据的压缩大小恒定,能够实现参数数据的高压缩。

[0067] 图3A是示出实现数据处理装置100的功能的硬件结构的框图。在图3A中,处理电路300是作为数据处理装置100发挥功能的专用电路。图3B是示出执行实现数据处理装置100的功能的软件的硬件结构的框图。在图3B中,处理器301以及存储器302通过信号总线而相互连接。

[0068] 数据处理装置100中的数据处理部101、压缩控制部102以及编码部103各自的功能通过处理电路来实现。

[0069] 即,数据处理装置100具备用于执行使用图4后述的步骤ST1～步骤ST3的处理的处理电路。

[0070] 处理电路可以是专用硬件,也可以是执行存储于存储器的程序的CPU(Central Processing Unit:中央处理单元)。

[0071] 在上述处理电路为图3A所示的专用硬件的情况下,处理电路300例如为单一电路、复合电路、程序化的处理器、并行程序化的处理器、ASIC(Application Specific

Integrated Circuit:面向特定用途的集成电路)、FPGA(Field-Programmable Gate Array:现场可编程门阵列)或它们的组合。

[0072] 另外,数据处理部101、压缩控制部102以及编码部103各自的功能既可以通过不同的处理电路来实现,也可以将它们的功能汇总而通过1个处理电路来实现。

[0073] 在上述处理电路为图3B所示的处理器的情况下,数据处理部101、压缩控制部102以及编码部103各自的功能通过软件、固件或软件和固件的组合来实现。

[0074] 将软件或固件记作程序而存储到存储器302。

[0075] 处理器301读出并执行存储于存储器302的程序而实现数据处理部101、压缩控制部102以及编码部103各自的功能。即,数据处理装置100具备用于存储在由处理器301执行时作为结果而执行图4所示的步骤ST1～步骤ST3的处理的程序的存储器302。

[0076] 这些程序使计算机执行数据处理部101、压缩控制部102以及编码部103的过程或方法。

[0077] 存储器302也可以是存储有用于使计算机作为数据处理部101、压缩控制部102以及编码部103发挥功能的程序的计算机可读存储介质。

[0078] 存储器302例如为RAM(Random Access Memory:随机存取存储器)、ROM(Read Only Memory:只读存储器)、闪存、EPROM(Erasable Programmable Read Only Memory:可擦除可编程只读存储器)、EEPROM(Electrically-EPROM:电可擦除可编程只读存储器)等非易失性或易失性的半导体存储器、磁盘、软盘、光盘、高密度盘、迷你盘、DVD等。

[0079] 另外,关于数据处理部101、压缩控制部102以及编码部103各自的功能,也可以由专用硬件实现一部分并由软件或固件实现一部分。

[0080] 例如,也可以是,关于数据处理部101,由作为专用硬件的处理电路来实现其功能,关于压缩控制部102以及编码部103,通过处理器301读出并执行存储于存储器302的程序来实现其功能。

[0081] 这样,处理电路能够通过硬件、软件、固件或它们的组合来实现上述功能的各个功能。

[0082] 另外,对数据处理装置100进行了说明,但数据处理装置200也是同样的。例如,数据处理装置200具备用于执行使用图5后述的步骤ST1a～步骤ST4a的处理的处理电路。该处理电路可以是专用硬件,也可以是执行存储于存储器的程序的CPU。

[0083] 如果上述处理电路为图3A所示的专用硬件,则处理电路300例如为单一电路、复合电路、程序化的处理器、并行程序化的处理器、ASIC、FPGA或它们的组合。

[0084] 另外,解码部201以及数据处理部202各自的功能既可以通过不同的处理电路来实现,也可以将它们的功能汇总而通过1个处理电路来实现。

[0085] 在上述处理电路为图3B所示的处理器时,解码部201以及数据处理部202各自的功能通过软件、固件或软件和固件的组合来实现。

[0086] 将软件或固件记作程序而存储到存储器302。

[0087] 处理器301通过读出并执行存储于存储器302的程序而实现解码部201以及数据处理部202各自的功能。

[0088] 即,数据处理装置200具备用于存储在由处理器301执行时作为结果而执行图5所示的步骤ST1a～步骤ST4a的处理的程序的存储器302。

[0089] 这些程序使计算机执行解码部201以及数据处理部202的过程或方法。

[0090] 存储器302也可以是存储有用于使计算机作为解码部201以及数据处理部202发挥功能的程序的计算机可读存储介质。

[0091] 另外,关于解码部201以及数据处理部202各自的功能,也可以通过专用硬件实现一部分并通过软件或固件实现一部分。

[0092] 例如,也可以是,关于解码部201,通过作为专用硬件的处理电路来实现其功能,关于数据处理部202,通过处理器301读出并执行存储于存储器302的程序来实现其功能。

[0093] 接下来,对动作进行说明。

[0094] 图4是示出数据处理装置100的动作的流程图。

[0095] 以下,对神经网络的参数数据为边的权重信息的情况进行说明。

[0096] 压缩控制部102决定对构成已学习的神经网络的多个边各自的权重信息进行量化时的量化步长,生成定义了量化步长的量化信息(步骤ST1)。量化信息从压缩控制部102输出到数据处理部101以及编码部103。

[0097] 当从压缩控制部102输入量化信息时,数据处理部101以量化信息中的量化步长,对上述神经网络的边的权重信息进行量化(步骤ST2)。数据处理部101生成包含量化后的边的权重信息的网络结构信息而输出到编码部103。

[0098] 编码部103对从数据处理部101输入的上述网络结构信息和从压缩控制部102输入的上述量化信息进行编码(步骤ST3)。

[0099] 将由编码部103编码后的上述网络结构信息以及上述量化信息的压缩数据输出到数据处理装置200。

[0100] 图5是示出数据处理装置200的动作的流程图。

[0101] 解码部201从由编码部103编码后的上述压缩数据解码出量化信息和网络结构信息(步骤ST1a)。将量化信息和网络结构信息从解码部201输出到数据处理部202。

[0102] 接下来,数据处理部202使用由解码部201从压缩数据解码出的量化信息和网络结构信息,计算逆量化后的边的权重信息(步骤ST2a)。

[0103] 接下来,数据处理部202使用包含有逆量化后的边的权重信息的网络结构信息构成神经网络(步骤ST3a)。

[0104] 由此,数据处理装置200能够构成在数据处理装置100中已学习的神经网络。

[0105] 数据处理部202使用在步骤ST3a中构成的神经网络来处理输入数据(步骤ST4a)。

[0106] 图6是示出实施方式1中的神经网络的结构例的图。

[0107] 在图6所示的神经网络中,在各个层中处理输入数据($x_1, x_2, \cdots, x_{N1}$)而输出处理结果($y_1, \cdots, y_{NL}$)。

[0108] 在图6中,N_l($l=1,2,\cdots,L$)表示第l层的层的节点数,L表示神经网络的层数。

[0109] 如图6所示,神经网络具有输入层、隐藏层以及输出层,在这些层的各个层中,成为多个节点用边连接的结构。

[0110] 能够根据用边连接之前的层的节点的输出值、边的权重信息以及按照每个层设定的活性化函数,计算多个节点各自的输出值。

[0111] 作为神经网络的例子有CNN(Convolutional Neural Network:卷积神经网络)。在CNN的隐藏层中交替地连接有卷积层(Convolutional layer)和池化层(Pooling layer),

按照最终的输出设置有全连接的神经网络层(全连接层:Fully-connected layer)。卷积层的活性化函数例如使用ReLU函数。

[0112] 另外,被称作DNN(Deep Neural Network:深度神经网络)的网络(也被称作深度规划、DCNN(Deep CNN)等)是对CNN的层数进行多层化而成的。

[0113] 图7是示出实施方式1中的一维数据的卷积处理的例子的图,示出进行一维数据的卷积处理的卷积层。一维数据例如有声音数据、时间序列数据。

[0114] 图7所示的卷积层在上一层具备9个节点10-1～10-9,在下一层具备3个节点11-1～11-3。

[0115] 对边12-1、12-6、12-11赋予相同的权重,对边12-2、12-7、12-12赋予相同的权重,对边12-3、12-8、12-13赋予相同的权重,对边12-4、12-9、12-14赋予相同的权重,对边12-5、12-10、12-15赋予相同的权重。并且,边12-1～12-5的权重有时全部是不同的值,有时多个权重是相同的值。

[0116] 上一层的9个节点10-1～10-9中的5个节点以上述权重连接于下一层的1个节点。核心大小K是5,核心由这些权重的组合规定。

[0117] 例如,如图7所示,节点10-1经由边12-1连接到节点11-1,节点10-2经由边12-2连接到节点11-1,节点10-3经由边12-3连接到节点11-1,节点10-4经由边12-4连接到节点11-1,节点10-5经由边12-5连接到节点11-1。核心由边12-1～12-5的权重的组合规定。

[0118] 节点10-3经由边12-6连接到节点11-2,节点10-4经由边12-7连接到节点11-2,节点10-5经由边12-8连接到节点11-2,节点10-6经由边12-9连接到节点11-2,节点10-7经由边12-10连接到节点11-2。核心由边12-6～12-10的权重的组合规定。

[0119] 节点10-5经由边12-11连接到节点11-3,节点10-6经由边12-12连接到节点11-3,节点10-7经由边12-13连接到节点11-3,节点10-8经由边12-14连接到节点11-3,节点10-9经由边12-15连接到节点11-3。核心由边12-11～12-15的权重的组合规定。

[0120] 在使用CNN的输入数据的处理中,数据处理部101或数据处理部202使用卷积层的边的权重的组合,按照每个核心以步长数S(在图7中S＝2)的间隔实施卷积运算。按照每个核心通过学习来决定边的权重的组合。

[0121] 另外,在图像辨识用途的CNN中,往往由具有多个核心的卷积层构成网络。

[0122] 图8是示出实施方式1中的二维数据的卷积处理的例子的图,示出图像数据这样的二维数据的卷积处理。

[0123] 图8所示的二维数据中的核心20是x方向的大小为K_x、y方向的大小为K_y的块区域。核心大小K是$K＝K_x×K_y$。

[0124] 数据处理部101或数据处理部202在二维数据中,以x方向步长数S_x的间隔以及y方向步长数S_y的间隔实施每个核心20的数据的卷积运算。在此,步长S_x、S_y是1以上的整数。

[0125] 图9是示出作为神经网络的全连接层的第l(l＝1,2,…,L)层的层中的每个节点的边的权重信息的矩阵的图。

[0126] 图10是示出作为神经网络的全连接层的第l(l＝1,2,…,L)层的层中的每个节点的边的权重信息的量化步长的矩阵的图。

[0127] 在神经网络中,图9所示的每个层的权重w_{ij}的组合成为构成网络的数据。因此,在DNN这样的多层的神经网络中,一般而言成为几百Mbyte以上的数据量,还需要大的存储器

大小。另外,i是节点索引,i＝1,2,…,N_l。j是边索引,j＝1,2,…,N_{l-1}。

[0128] 因此,在实施方式1的数据处理装置100中,为了削减边的权重信息的数据量,对权重信息进行量化。如图10所示,按照边的每个权重w_{ij}设定量化步长q_{ij}。

[0129] 进而,也可以在多个节点索引或多个边索引或多个节点索引和边索引中共用量化步长。由此,能够削减应编码的量化信息。

[0130] 图11是示出卷积层中的边的权重信息的矩阵的图。

[0131] 图12是示出卷积层中的边的权重信息的量化步长的矩阵的图。在卷积层中,在全部节点中共用针对1个核心的边的权重,能够减小按照每1个节点连接的边数即核心大小K而将核心设为小区域。

[0132] 图11是按照每个核心设定有边的权重$w_{i',j}$的数据,图12是按照每个核心设定有量化步长$q_{i',j}$的数据。

[0133] 另外,i'是核心索引,i'＝1,2,…,M_l(l＝1,2,…,L)。j'是边索引,j'＝1,2,…,K_l。

[0134] 进而,也可以在多个核心索引或多个边索引或多个核心索引和边索引中共用量化步长。由此,能够削减应编码的量化信息。

[0135] 压缩控制部102在图4的步骤ST1中决定在数据处理部101的权重量化处理中使用的量化步长,作为量化信息输出到数据处理部101。量化步长是图10所示的量化步长q_{ij}以及图12所示的量化步长$q_{i',j'}$。

[0136] 在图4的步骤ST2中,数据处理部101以图10所示的量化步长q_{ij}对图9所示的边的权重w_{ij}进行量化,将包含量化后的权重w_{ij}的网络结构信息输出到编码部103。

[0137] 同样地,在图4的步骤ST2中,数据处理部101以图12所示的量化步长$q_{i',j'}$对图11所示的边的权重$w_{i',j'}$进行量化,将包含量化后的权重$w_{i',j'}$的网络结构信息输出到编码部103。

[0138] 另外,在网络结构信息中,除了量化后的权重以外,还包含网络的层数、每个层的节点数、连接节点之间的边、按照每个边赋予的权重信息、表示节点的输出的活性化函数、每个层的类别信息(卷积层、池化层、全连接层)等。但是,在数据处理装置100与数据处理装置200之间预先固定(定义)的信息不包含在要编码的网络结构信息中。

[0139] 图13是示出实施方式1中的构成量化信息的信息的语法的图。

[0140] 图14是示出实施方式1中的构成量化信息的信息的矩阵单位的语法的图。

[0141] 图15是示出实施方式1中的构成量化信息的信息的层单位的语法的图。

[0142] 在图13中,标志quant_enable_flag、标志layer_adaptive_quant_flag、标志matrix_adaptive_quant_flag以及量化步长fixed_quant_step是由编码部103编码的量化信息的编码参数。

[0143] 并且,L是层数。

[0144] 在图14中,量化步长base_quant_step[j]、标志prev_quant_copy_flag[i-1]以及差分值diff_quant_value[i-1][j]是由编码部103编码的量化信息的编码参数。

[0145] 并且,C是节点数N_{layer_id}或核心数M_{layer_id}。进而,E是边数N_{layer_id-1}或核心大小K_{layer_id}。

[0146] 在图15中,量化步长base_layer_quant_step、标志layer_quant_copy_flag[i-2]以及量化步长layer_quant_step[i-2]是由编码部103编码的量化信息的编码参数。并且,L是层数。

[0147] 图13所示的信息包含有对网络中有无边的权重信息的量化进行设定的标志quant_enable_flag。

[0148] 在标志quant_enable_flag为0(伪)的情况下,网络中的全部边的权重信息未被量化。即,量化步长未被设定于量化信息。

[0149] 另一方面,在标志quant_enable_flag为1(真)的情况下,压缩控制部102参考标志layer_adaptive_quant_flag。

[0150] 压缩控制部102在标志layer_adaptive_quant_flag为0(伪)的情况下,将对网络中的全部边共用的量化步长fixed_quant_step设定于量化信息。

[0151] 在标志layer_adaptive_quant_flag为1(真)的情况下,压缩控制部102参考标志matrix_adaptive_quant_flag。

[0152] 在标志matrix_adaptive_quant_flag为0(伪)时,压缩控制部102决定按照层单位共用的量化步长,作为网络中的多个边各自的权重信息的量化步长。

[0153] 但是,由于输入层(第1层)不具有边,因此不设定量化步长。

[0154] 另外,图15示出与按照层单位共用的量化步长相关的语法。

[0155] 在标志matrix_adaptive_quant_flag为1(真)时,压缩控制部102决定图10所示的量化步长或图12所示的量化步长,作为网络中的多个边各自的权重信息的量化步长。图14示出图10或图12所示的量化步长的语法。

[0156] 对图14所示的语法进行说明。

[0157] 如上所述,输入层(第1层)不具有边。

[0158] 因此,ID信息layer_id为1~L-1的第layer_id+1层的层成为量化步长的设定对象。

[0159] 首先,压缩控制部102在第layer_id+1层的层中设定表示图10所示的第1个节点(在图12中为第1个核心)的量化步长的base_quant_step[j](j=0,1,…,E-1)。

[0160] 另外,E是边数N_{layer_id-1}或核心大小K_{layer_id}。

[0161] 接下来,压缩控制部102关于从第2个起(i≥1)的节点(或核心),以节点(或核心)为单位,参考表示量化步长与前1个索引的节点(或核心)是否相同的标志prev_quant_copy_flag[i-1]。

[0162] 在标志prev_quant_copy_flag[i-1]为1(真)的情况下,第i+1个节点(或核心)的量化步长与第i个节点(或核心)相同。

[0163] 另一方面,在标志prev_quant_copy_flag[i-1]为0(伪)的情况下,压缩控制部102设定差分值diff_quant_value[i-1][j](i=1,2,…,C-1,j=0,1,…,E-1),作为生成第i+1个节点(或核心)的量化步长的信息。

[0164] 能够对差分值diff_quant_value[i-1][j]加上设定于前1个节点(或核心)的量化步长来生成量化步长。

[0165] 即,在第2个(i=1)节点(或核心)中,base_quant_step[j]+diff_quant_value[0][j]成为量化步长。在从第3个起(i≥2)的节点(或核心)中,diff_quant_value[i-2][j]+diff_quant_value[i-1][j]成为量化步长。

[0166] 另外,作为编码参数示出节点(或核心)之间的量化步长的差分值diff_quant_value[i-1][j],但也可以以节点(或核心)为单位设定独立的量化步长。

[0167] 在存在节点(或核心)之间的量化步长的相关性低的倾向的情况下,这样构成时的编码部103中的编码效率更高。

[0168] 对图15所示的语法进行说明。

[0169] 如上所述,输入层(第1层)不具有边。

[0170] 因此,压缩控制部102设定base_layer_quant_step作为第2层的层中的全部边的权重信息中共用的量化步长。

[0171] 接下来,压缩控制部102关于从第3层起(i≥2)的层,参考表示第i+1层的层中的全部边的权重信息中共用的量化步长是否与第i层的层中的全部边的权重信息共用的量化步长相同的标志layer_quant_copy_flag[i-2](i=2,3,…,L-1)。

[0172] 在标志layer_quant_copy_flag[i-2]为1(真)的情况下,压缩控制部102设第i+1层中的全部边的权重信息中共用的量化步长与第i层的层中的全部边的权重信息中共用的量化步长相同。另一方面,在标志layer_quant_copy_flag[i-2]为0(伪)的情况下,压缩控制部102设定layer_quant_step[i-2]作为第i+1层中的全部边的权重信息中共用的量化步长。

[0173] 另外,在此示出压缩控制部102以层为单位定义layer_quant_step[i-2]作为独立的量化步长的情况,但也可以定义layer_quant_step[i-2]作为与前1个层(第i层的层)的量化步长之间的差分值。通过设为差分值,产生大量0附近的差分值,因此,能够提高编码部103中的编码效率。

[0174] 编码部103将图13～图15中的编码参数作为量化信息进行编码来生成压缩数据。

[0175] 另外,在此将量化步长的最小切换单位作为边单位,但也可以如图10所示将节点单位(在图12中以核心为单位)作为量化步长的最小切换单位。这与在图14中设E=1是相同的意思。在该情况下,也可以以节点为单位(在图12中以核心为单位)独立地对量化步长进行编码。

[0176] 并且,量化步长的最小切换单位也可以是层单位。

[0177] 这与在图13中是标志layer_adaptive_quant_flag=1(真)的情况下无标志matrix_adaptive_quant_flag而始终仅执行layer_quant_coding()是相同的意思。通过这样使量化步长的最小切换单位大于边单位,能够削减编码前的量化信息的数据大小。

[0178] 如上所述,在实施方式1的数据处理装置100中,数据处理部101使用神经网络来处理输入数据。压缩控制部102决定量化步长并生成定义了量化步长的量化信息。编码部103对包含以由压缩控制部102决定的量化步长量化后的参数数据的网络结构信息和量化信息进行编码而生成压缩数据。

[0179] 特别是,以上对作为神经网络的参数数据而处理对神经网络中的连接节点之间的边赋予的权重信息的例子进行了说明。通过具有这些结构,将定义了量化步长的量化信息和包含以量化信息中的量化步长量化后的参数数据的网络结构信息编码成压缩数据。由此,能够对神经网络的参数数据进行高压缩。

[0180] 并且,通过使用从压缩数据解码出的量化信息和网络结构信息,能够在解码侧构成在编码侧优化的神经网络。

[0181] 在实施方式1的数据处理装置200中,数据处理部202使用神经网络来处理输入数据。解码部201对压缩数据进行解码。

[0182] 在该结构中,数据处理部202使用由解码部201从压缩数据解码出的量化信息和网络结构信息对参数数据进行逆量化,使用包含逆量化后的参数数据的网络结构信息构成神经网络。

[0183] 由此,能够使用从压缩数据解码出的量化信息和网络结构信息构成在编码侧优化的神经网络。

[0184] 在实施方式1的数据处理装置100中,压缩控制部102按照每个边切换量化步长。编码部103对定义了每个边的量化步长的量化信息进行编码。通过这样构成,能够高精度地量化参数数据。

[0185] 在实施方式1的数据处理装置100中,压缩控制部102按照每个节点或每个核心切换量化步长。编码部103对定义了每个节点或每个核心的量化步长的量化信息进行编码。

[0186] 这样构成也能够高精度地量化参数数据。

[0187] 在实施方式1的数据处理装置100中,压缩控制部102按照神经网络的每个层切换量化步长。编码部103对定义了神经网络的每个层的量化步长的量化信息进行编码。

[0188] 通过这样构成,基于量化的参数数据的压缩率升高,因此,能够削减编码前的权重信息的数据量。

[0189] 实施方式2

[0190] 在实施方式1中,对将神经网络的输出结果直接作为数据处理结果的例子进行了说明,但还有如下的应用例:将神经网络的中间层的输出用作以下述参考文献的图像检索(retrieval)或匹配(matching)为一例的针对图像数据和声音数据的数据处理的特征量,将其如下述参考文献所述通过另外的数据处理手法得到最终的数据处理结果。

[0191] 例如,在作为图像检索、匹配、物体追踪等图像处理的图像特征量使用神经网络的中间层的输出的情况下,通过进行针对在以往的上述图像处理中使用的图像特征量即HOG(Histogram of Oriented Gradients:定向梯度直方图)、SIFT(Scale Invariant Feature Transform:尺度不变特征变换)、SURF(Speeded Up Robust Features:加速稳健特征)等的图像特征量的置换或追加,能够以与使用上述以往的图像特征量的图像处理相同的处理流程来实现图像处理。

[0192] 在该情况下,在数据处理装置100中,作为网络结构信息、量化信息进行编码的是直至能得到作为数据处理特征量的输出的中间层为止的神经网络。

[0193] 进而,数据处理装置100使用上述数据处理特征量来进行图像检索等数据处理。数据处理装置200从压缩数据解码出直至上述中间层为止的神经网络,将输入输入数据而得到的输出作为数据处理特征量来实施图像检索等数据处理。

[0194] (参考文献)ISO/IEC JTC1/SC29/WG11/m39219,“Improved retrieval and matching with CNN feature for CDVA”,Chengdu,China,Oct.2016.

[0195] 因此,在实施方式2的数据处理装置100中基于量化的参数数据的压缩率升高,因此,能够削减编码前的权重信息的数据量。在实施方式2的数据处理装置200中,能够通过对从上述数据处理装置100输出的压缩数据进行解码来生成神经网络,从而实施数据处理。

[0196] 另外,本发明不限于上述实施方式,能够在本发明的范围内进行实施方式的任意结构要素的变形或实施方式的任意结构要素的省略。

[0197] 产业上的可利用性

[0198] 本发明的数据处理装置能够对神经网络的参数数据进行高压缩,因此,例如能够用于图像识别技术。

[0199] 符号说明

[0200] 10-1～10-9、11-1～11-3节点;12-1～12-15边;20核心;100、200数据处理装置;101、202数据处理部;102压缩控制部;103编码部;201解码部;300处理电路;301处理器;302存储器。

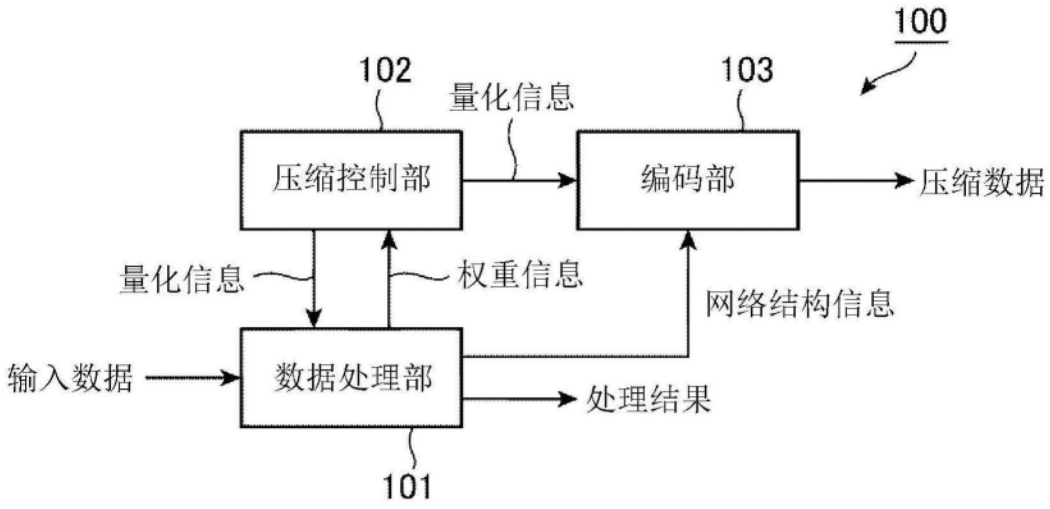

100
102
量化信息
103
压缩控制部
编码部
压缩数据
量化信息
权重信息
网络结构信息
输入数据
数据处理部
处理结果
101

图1

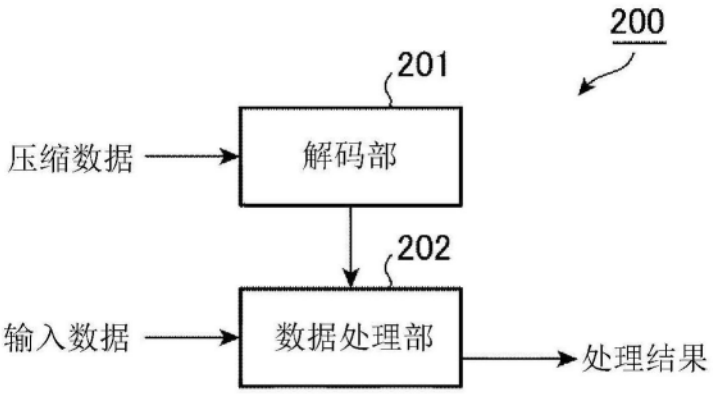

200
201
压缩数据
解码部
202
输入数据
数据处理部
处理结果

图2

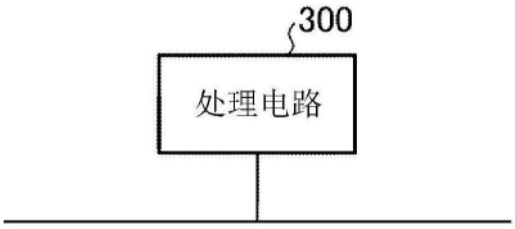

300
处理电路

图3A

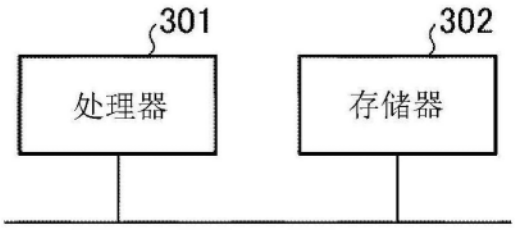

301
302
处理器
存储器

图3B

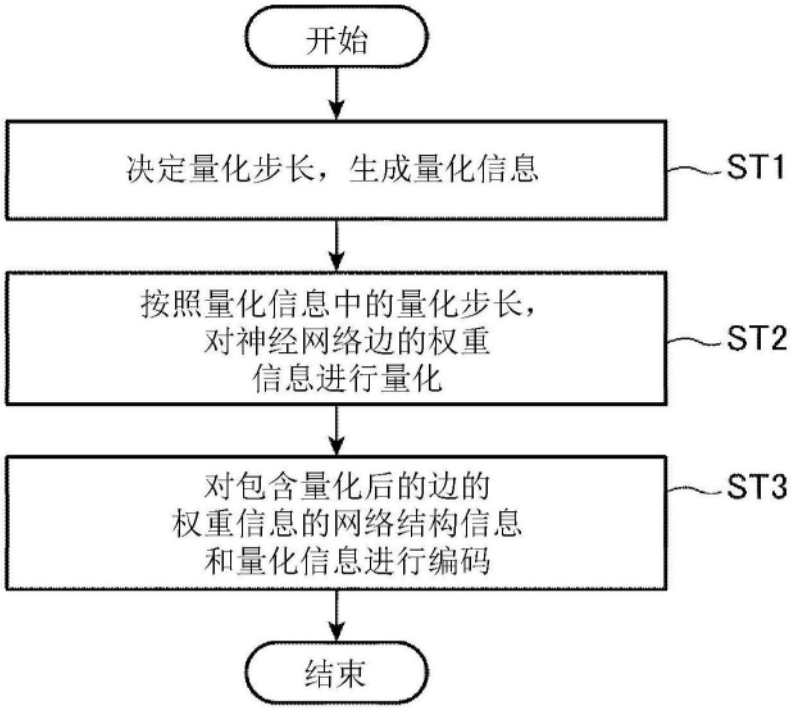

开始
决定量化步长，生成量化信息
ST1
按照量化信息中的量化步长，
对神经网络边的权重
信息进行量化
ST2
对包含量化后的边的
权重信息的网络结构信息
和量化信息进行编码
ST3
结束

图4

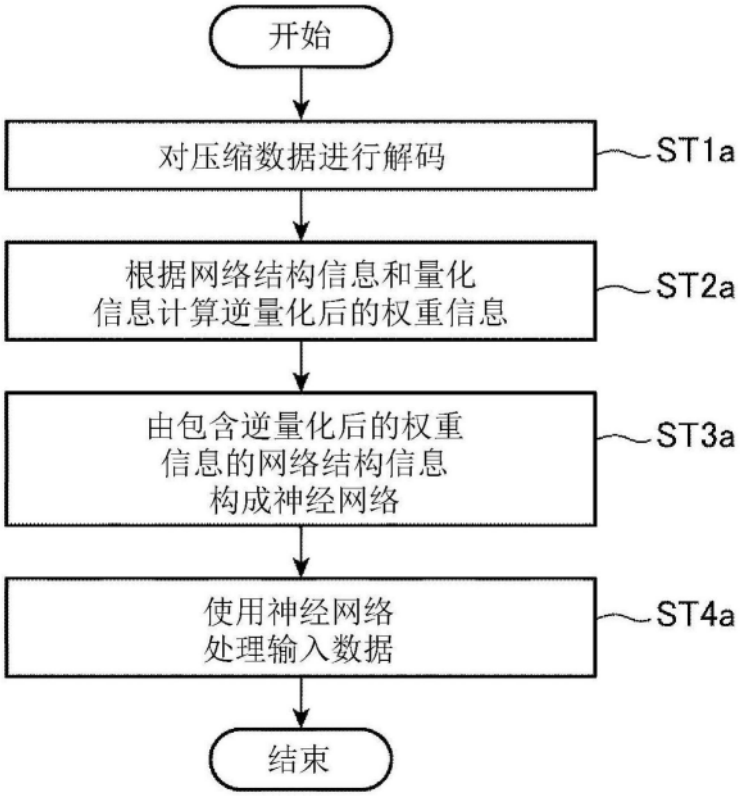

开始
对压缩数据进行解码
ST1a
根据网络结构信息和量化
信息计算逆量化后的权重信息
ST2a
由包含逆量化后的权重
信息的网络结构信息
构成神经网络
ST3a
使用神经网络
处理输入数据
ST4a
结束

图5

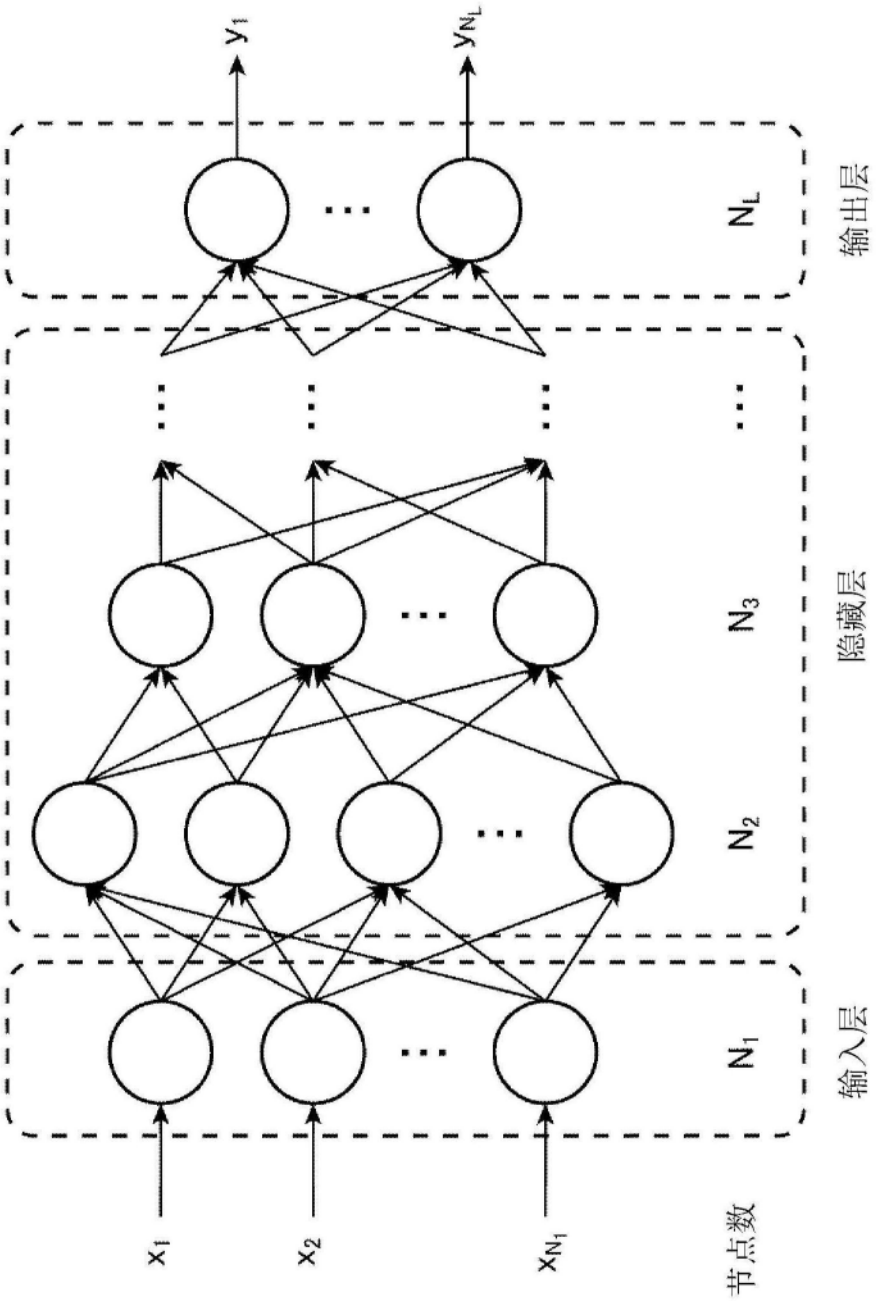


图6

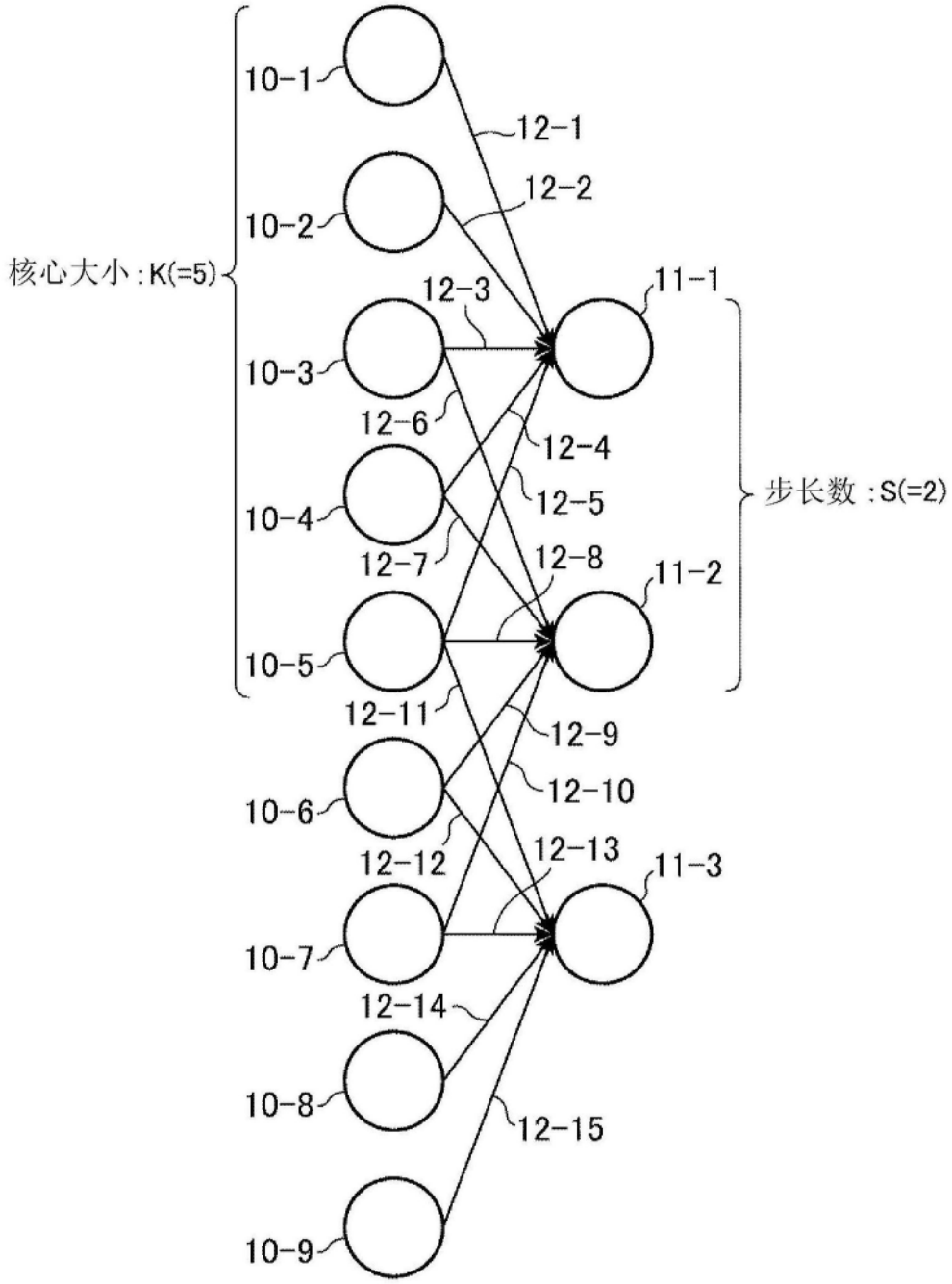

核心大小 :K(=5)
步长数 :S(=2)
10-1
10-2
10-3
10-4
10-5
10-6
10-7
10-8
10-9
11-1
11-2
11-3
12-1
12-2
12-3
12-4
12-5
12-6
12-7
12-8
12-9
12-10
12-11
12-12
12-13
12-14
12-15

图7

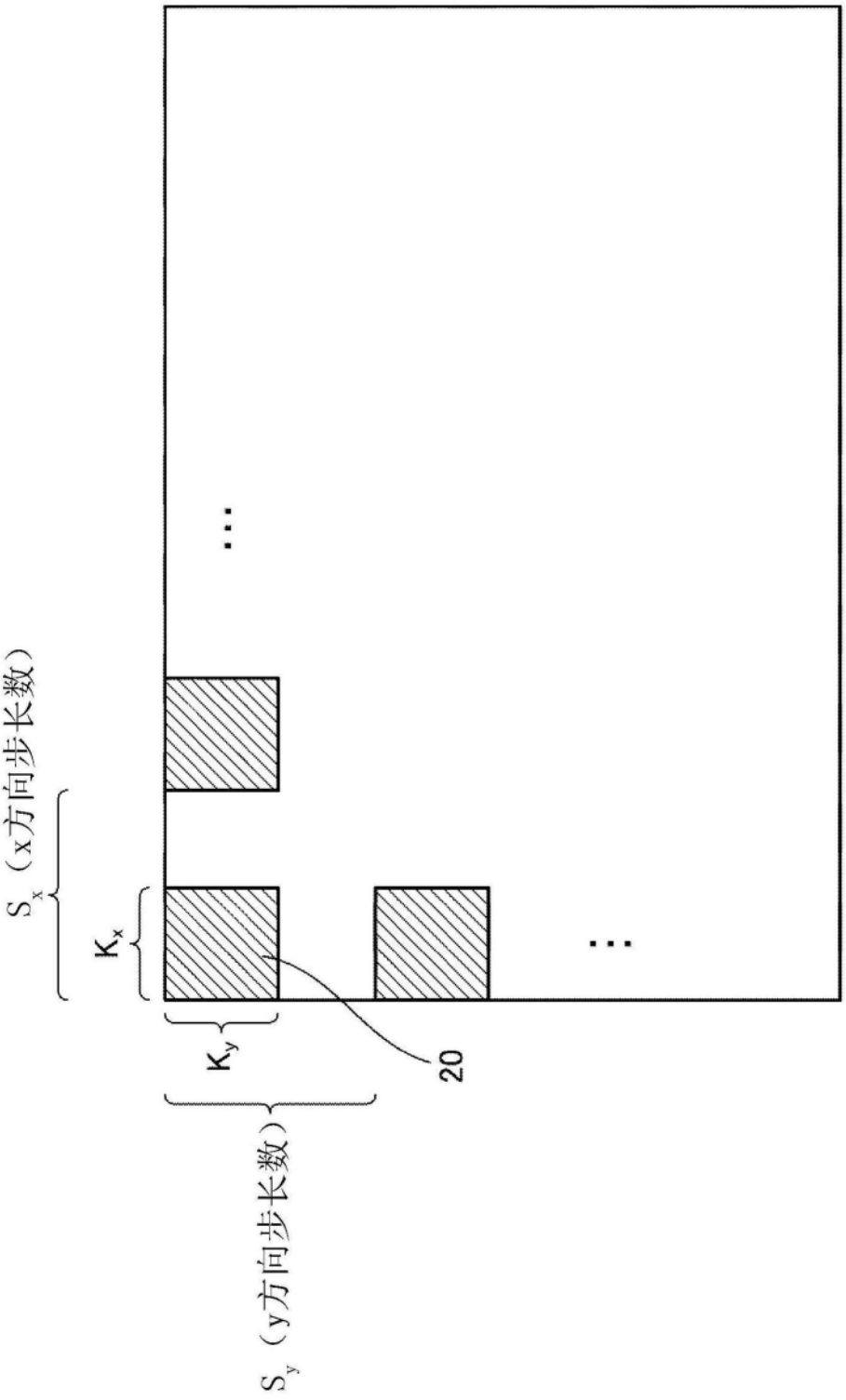


图8

		节点索引				
		1	2	3	…	N_l
边索引	1	w_{11}	w_{21}	w_{31}	…	w_{N_l1}
	2	w_{12}	w_{22}	w_{32}	…	w_{N_l2}
	3	w_{13}	w_{23}	w_{33}	…	w_{N_l3}
	⋮	⋮	⋮	⋮	…	⋮
	N_{l-1}	$w_{1N_{l-1}}$	$w_{2N_{l-1}}$	$w_{3N_{l-1}}$	…	$w_{N_lN_{l-1}}$

图9

		节点索引				
		1	2	3	…	N_l
边索引	1	q_{11}	q_{21}	q_{31}	…	q_{N_l1}
	2	q_{12}	q_{22}	q_{32}	…	q_{N_l2}
	3	q_{13}	q_{23}	q_{33}	…	q_{N_l3}
	⋮	⋮	⋮	⋮	…	⋮
	N_{l-1}	$q_{1N_{l-1}}$	$q_{2N_{l-1}}$	$q_{3N_{l-1}}$	…	$q_{N_lN_{l-1}}$

图10

		核心索引				
		1	2	3	…	M_l
边索引	1	w_{11}	w_{21}	w_{31}	…	w_{M_l1}
	2	w_{12}	w_{22}	w_{32}	…	w_{M_l2}
	3	w_{13}	w_{23}	w_{33}	…	w_{M_l3}
	⋮	⋮	⋮	⋮	…	⋮
	K_l	w_{1K_l}	w_{2K_l}	w_{3K_l}	…	$w_{M_lK_l}$

图11

		核心索引				
		1	2	3	…	M_l
边索引	1	q_{11}	q_{21}	q_{31}	…	q_{M_l1}
	2	q_{12}	q_{22}	q_{32}	…	q_{M_l2}
	3	q_{13}	q_{23}	q_{33}	…	q_{M_l3}
	⋮	⋮	⋮	⋮	…	⋮
	K_l	q_{1K_l}	q_{2K_l}	q_{3K_l}	…	$q_{M_lK_l}$

图12

```
quant_coding() {
  quant_enable_flag
  if (quant_enable_flag) {
    layer_adaptive_quant_flag
    if (layer_adaptive_quant_flag) {
      matrix_adaptive_quant_flag
      if (matrix_adaptive_quant_flag) {
        for (i =1; i < L; i++)
          matrix_quant_coding(i)
      } else {
        layer_quant_coding()
      }
    } else {
      fixed_quant_step
    }
  }
}
```

图13

```
matrix_quant_coding(layer_id) {
  for (j =0; j < E; j++)
    base_quant_step[j]
  for (i =1; i < C; i++) {
    prev_quant_copy_flag[i-1]
    if (!prev_quant_copy_flag[i-1]) {
      for (j =0; j < E; j++)
        diff_quant_value[i-1][j]
    }
  }
}
```

图14

```
layer_quant_coding() {
  base_layer_quant_step
  for (i =2; i < L; i++) {
    layer_quant_copy_flag[i-2]
    if (!layer_quant_copy_flag[i-2])
      layer_quant_step[i-2]
  }
}
```

图15