(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116868209 A
(43)申请公布日 2023.10.10

(21)申请号 202280015513.2

(22)申请日 2022.03.01

(30)优先权数据
2021-036254 2021.03.08 JP

(85)PCT国际申请进入国家阶段日
2023.08.17

(86)PCT国际申请的申请数据
PCT/JP2022/008658 2022.03.01

(87)PCT国际申请的公布数据
WO2022/190966 JA 2022.09.15

(71)申请人 欧姆龙株式会社
地址 日本京都府京都市下京区盐小路通堀川东入南不动堂町801番地

(72)发明人 长江成典

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
专利代理师 杨文娟 黄健

(51)Int.Cl.
G06N 20/00(2006.01)

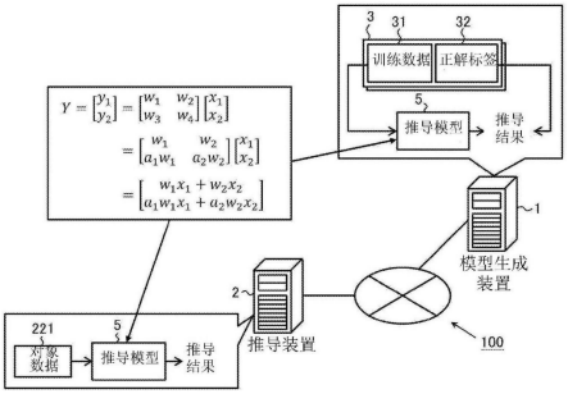
权利要求书4页 说明书36页 附图10页

(54)发明名称

推导装置、模型生成装置、推导方法以及推导程序

(57)摘要

本发明的一方面的推导装置获取对象数据，使用通过机器学习而训练完毕的推导模型，对所获取的所述对象数据执行推导任务，且输出与执行推导任务的结果相关的信息。推导模型的多个参数的至少一部分是以矩阵来表达。矩阵包含第一局部矩阵及第二局部矩阵。第一局部矩阵及第二局部矩阵的行及列各自的要素的数量相同，第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致。



CN 116868209 A

1.一种推导装置,包括:

数据获取部,获取对象数据;

推导部,使用通过机器学习而训练完毕的推导模型,对所获取的所述对象数据执行推导任务;以及

输出部,输出与执行所述推导任务的结果相关的信息,

所述推导模型的多个参数的至少一部分是以矩阵来表达,

所述矩阵包含第一局部矩阵及第二局部矩阵,

所述第一局部矩阵及所述第二局部矩阵的行及列各自的要素的数量相同,

所述第二局部矩阵的各要素的值被调节为与所述第一局部矩阵及对角矩阵的积一致。

2.根据权利要求1所述的推导装置,其中

在所述矩阵的至少一部分中,有如下所述的定标关系成立,即,

所述矩阵的至少一部分以行及列各自的要素的数量相同的局部矩阵在行上各排列M个且在列上各排列N个的方式,被分割为M×N个局部矩阵,

在各列中,配置在任一行上的局部矩阵构成相对于配置在除了所述任一行以外的其他行上的各局部矩阵的所述第一局部矩阵,且

配置在所述其他行上的各局部矩阵构成所述第二局部矩阵。

3.根据权利要求2所述的推导装置,其中

所述M及所述N为相同的素数S。

4.根据权利要求2或3所述的推导装置,其中

通过反复进行构成所述第一局部矩阵的局部矩阵内的所述定标关系的成立,从而在所述矩阵的至少一部分内,所述定标关系递归地成立。

5.根据权利要求4所述的推导装置,其中

所述M及所述N为相同的素数S,

所述矩阵的至少一部分包含要素的数量为素数S的幂的正方矩阵。

6.根据权利要求2至5中任一项所述的推导装置,其中

所述M及所述N分别为2。

7.根据权利要求1至6中任一项所述的推导装置,其中

所述推导模型包含神经网络,

所述矩阵的各要素构成为,与所述神经网络中的各神经元间的结合的权重对应。

8.根据权利要求1至7中任一项所述的推导装置,其中

所述对象数据包含映照有产品的图像数据,

所述推导任务是判定映照在所述图像数据中的产品是否存在缺陷。

9.一种模型生成装置,包括:

数据获取部,获取多个学习数据集,所述多个学习数据集分别包含训练数据及正解标签的组合,所述正解标签表示针对所述训练数据的推导任务的正解;以及

学习处理部,使用所述多个学习数据集来实施推导模型的机器学习,且所述学习处理部中,

所述推导模型的多个参数的至少一部分是以矩阵来表达,

所述矩阵包含第一局部矩阵及第二局部矩阵,

所述第一局部矩阵及所述第二局部矩阵的行及列各自的要素的数量相同，

所述机器学习是通过下述方式而构成，即，关于所述各学习数据集来训练所述推导模型，以使得使用所述推导模型来对所述训练数据执行所述推导任务的结果符合由所述正解标签所表示的正解，并且所述第二局部矩阵的各要素的值被调节为与所述第一局部矩阵及对角矩阵的积一致。

10.根据权利要求9所述的模型生成装置，其中

训练所述推导模型的处理包含：调节所述矩阵的各要素的值，以使得在所述矩阵的至少一部分中有如下所述的定标关系成立，即，所述矩阵的至少一部分以行及列各自的要素的数量相同的局部矩阵在行上各排列M个且在列上各排列N个的方式被分割为M×N个局部矩阵，在各列中，配置在任一行上的局部矩阵构成相对于配置在除了所述任一行以外的其他行上的各局部矩阵的所述第一局部矩阵，且配置在所述其他行上的各局部矩阵构成所述第二局部矩阵。

11.根据权利要求9所述的模型生成装置，其中

所述推导模型包含神经网络，

所述矩阵的各要素构成为，与所述神经网络中的各神经元间的结合的权重对应，

训练所述推导模型的处理包含：反复进行构成所述第一局部矩阵的局部矩阵内的所述定标关系的成立直至各局部矩阵成为1×1矩阵为止，由此来调节所述矩阵的至少一部分的各要素的值，以使得在所述矩阵的至少一部分内，所述定标关系递归地成立，

调节所述矩阵的至少一部分的各要素的值的处理包含：

以所述定标关系递归地成立的方式给予所述矩阵的至少一部分的各要素的初始值；

在正向传播的阶段中，导出对所述各学习数据集的训练数据执行推导任务的结果；以及

在反向传播的阶段中，对所导出的推导任务的执行结果以及由对应的正解标签所表示的正解之间的误差的梯度进行反向传播，由此来调节所述矩阵的至少一部分的各要素以及所述对角矩阵的各要素的值，

所述正向传播的阶段的运算包含：

第一步骤，计算构成初始的第一局部矩阵的所述1×1矩阵及输入向量的积；

第二步骤，计算所获得的所述初始的第一局部矩阵的积的结果以及所述对角矩阵的对应的要素的积，由此导出构成初始的第二局部矩阵的所述1×1矩阵及输入向量的积；

第三步骤，根据在对象层级中获得的所述第一局部矩阵的积的结果以及所述第二局部矩阵的积的结果，导出构成后续层级的第一局部矩阵的局部矩阵及输入向量的积；

第四步骤，计算所获得的后续层级的所述第一局部矩阵的积的结果以及所述对角矩阵的对应的要素的积，由此导出构成后续层级的所述第二局部矩阵的局部矩阵及输入向量的积；以及

第五步骤，作为在所述对象层级中获得的所述第一局部矩阵的积的结果以及所述第二局部矩阵的积的结果各自的初始值，分别代入在所述第一步骤以及所述第二步骤中分别获得的所述初始的第一局部矩阵的积的结果以及所述初始的第二局部矩阵的积的结果，且递归地反复进行所述第三步骤及所述第四步骤，由此导出所述矩阵的至少一部分及输入向量的积，

所述反向传播的阶段的运算包含：

第一步骤，获取相对于所述矩阵的至少一部分及输入向量的积的所述误差的梯度；

第二步骤，基于所获取的所述误差的梯度、与在所述正向传播的阶段的递归性反复的最终层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的值，导出关于在所述最终层级中使用的对角矩阵的各要素的误差的梯度；

第三步骤，基于所获取的所述误差的梯度以及在所述最终层级中使用的对角矩阵的各要素的值，导出关于在所述最终层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度；

第四步骤，基于所获取的所述误差的梯度、在所述正向传播的阶段的递归性反复的对象层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的值、以及关于在所述正向传播的阶段的递归性反复中在所述对象层级的后续层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度，导出关于在所述对象层级中使用的对角矩阵的各要素的误差的梯度；

第五步骤，基于所获取的所述误差的梯度、在所述对象层级中使用的对角矩阵的各要素的值、以及关于在所述后续层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度，导出关于在所述对象层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度；

第六步骤，作为关于在所述后续层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度的初始值，代入在所述反向传播的阶段的所述第三步骤中导出的、关于在所述最终层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度，且递归地反复进行所述反向传播的阶段的所述第四步骤及所述第五步骤，直至导出关于在所述正向传播的阶段的递归性反复的初始层级中使用的对角矩阵的各要素的误差的梯度为止，由此导出关于在各层级中使用的对角矩阵的各要素的误差的梯度；

第七步骤，基于所述输入向量以及通过所述第六步骤而导出的、关于在所述初始层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度，导出关于构成所述初始的第一局部矩阵的所述1×1矩阵的要素的误差的梯度；以及

第八步骤，基于关于在所述各层级中使用的对角矩阵的各要素以及构成所述初始的第一局部矩阵的所述1×1矩阵的要素分别导出的所述误差的梯度，调节在所述各层级中使用的对角矩阵的各要素以及构成所述初始的第一局部矩阵的所述1×1矩阵的要素各自的值。

12.一种推导方法，由计算机执行下述步骤：

获取对象数据；

使用通过机器学习而训练完毕的推导模型，对所获取的所述对象数据执行推导任务；以及

输出与执行所述推导任务的结果相关的信息，所述推导方法中，

所述推导模型的多个参数的至少一部分是以矩阵来表达，

所述矩阵包含第一局部矩阵及第二局部矩阵，

所述第一局部矩阵及所述第二局部矩阵的行及列各自的要素的数量相同，

所述第二局部矩阵的各要素的值被调节为与所述第一局部矩阵及对角矩阵的积一致。

13.一种推导程序，用于使计算机执行下述步骤：

获取对象数据；

使用通过机器学习而训练完毕的推导模型，对所获取的所述对象数据执行推导任务；以及

输出与执行所述推导任务的结果相关的信息，所述推导程序中，

所述推导模型的多个参数的至少一部分是以矩阵来表达，

所述矩阵包含第一局部矩阵及第二局部矩阵，

所述第一局部矩阵及所述第二局部矩阵的行及列各自的要素的数量相同，

所述第二局部矩阵的各要素的值被调节为与所述第一局部矩阵及对角矩阵的积一致。

推导装置、模型生成装置、推导方法以及推导程序

技术领域

[0001] 本发明涉及一种推导装置、模型生成装置、推导方法以及推导程序。

背景技术

[0002] 以往，在制造线等制造产品的场景中，正在推进下述技术的开发，即，通过拍摄装置来拍摄所制造的产品，并对所获得的图像数据进行分析，由此来检查产品的良否。例如，专利文献1中提出一种检查装置，其使用训练完毕的第一神经网络来判定映照在图像中的对象物是正常还是异常，若判定为异常，则使用训练完毕的第二神经网络来对所述异常的种类进行分类。

[0003] 根据利用训练完毕的神经网络的方法，即便未通过人工来详细规定图像处理的内容，也能够基于训练完毕的神经网络的运算结果来实施产品的外观检查。因此，能够简化外观检查的信息处理，降低制作检查程序的工时。此作用效果并不限于在利用神经网络的情形(case)下获得。在利用神经网络以外的、训练完毕的机器学习模型(例如，通过主成分分析所获得的主成分向量、支持向量机等)的方法中，也能够简化外观检查的信息处理，降低制作检查程序的工时。

[0004] 除此以外，可由训练完毕的机器学习模型来执行的推导任务并不限于外观检查以及对图像的推导。通过在机器学习中使用与所期望的信息处理对应的学习数据，便能够生成获得了对规定种类的数据执行所期望的推导任务的能力的、训练完毕的机器学习模型。因此，根据利用训练完毕的机器学习模型的方法，能够简化对规定种类的数据执行所期望的推导任务的信息处理，降低制作推导程序的工时。

[0005] 一般而言，机器学习模型具有在推导任务的运算中使用的参数。参数的值通过机器学习进行调节，以使得可相对于所期望的输入而获得所期望的输出(即，获得执行所期望的推导任务的能力)。在对机器学习模型采用神经网络的情况下，此参数例如为各神经元间的结合的权重等。存在下述倾向，即，越要提高推导任务的执行精度，则构成机器学习模型的参数的数量越增加。尤其，在对机器学习模型采用神经网络的情形中，为了提高推导任务的执行精度，存在加深神经网络的层级的倾向，神经网络的层级越深，则构成神经网络的参数的数量越增加。

[0006] 当然，构成机器学习模型的参数的数量越增加，则越能预见推导任务的执行精度的提高，但会导致机器学习模型的运算处理所耗费的计算量增加。因此，在机器学习的场景以及执行推导任务的场景这两种场景下，例如会产生计算时间延迟、压迫存储器、因计算耗费时间导致消耗电力变高、若非昂贵的计算机则无法完成运算等伴随计算量增加造成的各种问题(即，对计算资源造成负担)。

[0007] 因此，在专利文献2以及专利文献3中，提出了通过分散处理来使机器学习模型的运算处理高速化的方法。具体而言，在专利文献2中提出一种方法：通过在多个学习装置以及分散深层学习装置之间交换量化梯度，从而分散地进行深层学习。而且，在专利文献3中提出了一种方法：经由能够单向通信的环形通信网络来连接多个学习节点，通过各学习节

点间的协调处理来分散地进行深层学习。除此以外，在非专利文献1中提出了一种方法：将输入数据及参数的积的计算通过高速傅里叶变换而变换为其他表达的计算式，由此来降低卷积运算中的计算次数。

[0008] 现有技术文献

[0009] 专利文献

[0010] 专利文献1：日本专利特开2012-026982号公报

[0011] 专利文献2：日本专利特开2018-120441号公报

[0012] 专利文献3：日本专利特开2020-003848号公报

[0013] 非专利文献

[0014] 非专利文献1：Tyler Highlander、Andres Rodriguez，“使用快速傅立叶变换和重叠加法的卷积神经网络的高效训练(Very Efficient Training of Convolutional Neural Networks using Fast Fourier Transform and Overlap-and-Add)”，arXiv:1601.06815[cs.NE]，2016年1月25日

发明内容

[0015] 发明所要解决的问题

[0016] 本申请发明人发现在所述以往的方法中存在下述问题。即，在仅依据分散处理的方法中，通过连接多个计算机，虽能降低每一台计算机的负担且使运算处理高速化，但难以降低机器学习模型的运算处理所耗费的计算量。反而，因在各计算机间对计算过程的信息进行通信，而导致机器学习模型的运算处理所耗费的计算量相应地增加，在以整体来看的情况下，对计算资源造成的负担增加。而且，在像非专利文献1那样的、使用高速傅里叶变换的方法中，虽能使卷积运算高速化，但难以适用于其他运算。除此以外，因与机器学习模型的参数信息一同保持通过高速傅里叶变换而获得的其他表达的信息，对计算资源造成的负担相应地增加(尤其是存储器受到压迫)。因此，以往的方法中，难以降低机器学习模型的运算处理所耗费的计算量，抑制对计算资源造成的负担。

[0017] 本发明在一方面是有鉴于此种情况而完成，其目的在于提供一种技术，用于降低机器学习模型的运算处理所耗费的计算量，抑制对计算资源造成的负担。

[0018] 解决问题的技术手段

[0019] 为了解决所述问题，本发明采用以下的结构。

[0020] 即，本发明的一方面的推导装置包括：数据获取部，获取对象数据；推导部，使用通过机器学习而训练完毕的推导模型，对所获取的所述对象数据执行推导任务；以及输出部，输出与执行所述推导任务的结果相关的信息。所述推导模型的多个参数的至少一部分是以矩阵来表达。所述矩阵包含第一局部矩阵及第二局部矩阵。所述第一局部矩阵及所述第二局部矩阵的行及列各自的要素的数量相同，所述第二局部矩阵的各要素的值被调节为与所述第一局部矩阵及对角矩阵的积一致。

[0021] 所述结构中，推导模型的多个参数的至少一部分可通过包含第一局部矩阵及第二局部矩阵的矩阵来表达，第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致。因此，即便未保持第二局部矩阵的信息，也能够回头利用第一局部矩阵的运算结果来获得第二局部矩阵的运算结果。因此，根据所述结构，能够实质上降低构成推导模型的

参数的数量。具体而言，能够将参数的数量降低与第二局部矩阵的要素数与对角矩阵的0以外的要素数的差值相应的量。由此，能够降低推导模型的运算处理所耗费的计算量，抑制对计算资源造成的负担。

[0022] 所述一方面的推导装置中，也可为，在所述矩阵的至少一部分中，有如下所述的定标关系成立，即，所述矩阵的至少一部分以行及列各自的要素的数量相同的局部矩阵在行上各排列M个且在列上各排列N个的方式，被分割为M×N个局部矩阵，在各列中，配置在任一行上的局部矩阵构成相对于配置在除了所述任一行以外的其他行上的各局部矩阵的所述第一局部矩阵，且配置在所述其他行上的各局部矩阵构成所述第二局部矩阵。根据所述结构，通过定标关系的成立，能够有效地削减推导模型的参数数量。由此，能够有效地降低推导模型的运算处理所耗费的计算量，抑制对计算资源造成的负担。

[0023] 所述一方面的推导装置中，所述M及所述N可为相同的素数S。根据所述结构，能够容易地调节所述定标关系的成立范围。

[0024] 所述一方面的推导装置中，也可为，通过反复进行构成所述第一局部矩阵的局部矩阵内的所述定标关系的成立，从而在所述矩阵的至少一部分内，所述定标关系递归地成立。根据所述结构，通过定标关系递归地成立，从而即便在最终(即，在最外侧)成立的定标关系中的第一部分列内，也能够实现参数数量的降低。因而，能够更有效地降低推导模型的运算处理所耗费的计算量，进一步抑制对计算资源造成的负担。另外，定标关系的成立可反复进行至各局部矩阵成为任意尺寸的矩阵(例如1×1矩阵)为止。

[0025] 所述一方面的推导装置中，所述M及所述N可为相同的素数S，所述矩阵的至少一部分可包含要素的数量为素数S的幂的正方矩阵。根据所述结构，能够容易地调节所述定标关系的成立范围。

[0026] 所述一方面的推导装置中，所述M及所述N可分别为2。根据所述结构，能够容易地调节所述定标关系的成立范围。

[0027] 所述一方面的推导装置中，所述推导模型可包含神经网络，所述矩阵的各要素可构成为，与所述神经网络中的各神经元间的结合的权重对应。根据所述结构，在采用神经网络来作为推导模型(机器学习模型)的场景中，能够降低推导模型的运算处理所耗费的计算量，抑制对计算资源造成的负担。

[0028] 所述一方面的推导装置中，所述对象数据可包含映照有产品的图像数据，所述推导任务可为判定映照在所述图像数据中的产品是否存在缺陷。根据所述结构，在使用训练完毕的推导模型来进行产品的外观检查的场景中，能够降低推导模型的运算处理所耗费的计算量，抑制对计算资源造成的负担。

[0029] 而且，本发明的形态可不限于所述推导装置。本发明的一方面也可为模型生成装置，其生成在所述任一形态的推导装置中使用的训练完毕的推导模型。

[0030] 例如，本发明的一方面的模型生成装置包括：数据获取部，获取多个学习数据集，所述多个学习数据集分别包含训练数据及正解标签的组合，所述正解标签表示针对所述训练数据的推导任务的正解；以及学习处理部，使用所述多个学习数据集来实施推导模型的机器学习。所述推导模型的多个参数的至少一部分是以矩阵来表达。所述矩阵包含第一局部矩阵及第二局部矩阵。所述第一局部矩阵及所述第二局部矩阵的行及列各自的要素的数量相同。并且，所述机器学习是通过下述方式而构成，即，关于所述各学习数据集来训练所

述推导模型,以使得使用所述推导模型来对所述训练数据执行所述推导任务的结果符合由所述正解标签所表示的正解,并且所述第二局部矩阵的各要素的值被调节为与所述第一局部矩阵及对角矩阵的积一致。根据所述结构,能够降低推导模型的运算处理所耗费的计算量,抑制生成训练完毕的推导模型时的对计算资源造成的负担。

[0031] 所述一方面的模型生成装置中,训练所述推导模型的处理可包含:以在所述矩阵的至少一部分中有如下所述的定标关系成立的方式来调节所述矩阵的各要素的值,即,所述矩阵的至少一部分以行及列各自的要素的数量相同的局部矩阵在行上各排列M个且在列上各排列N个的方式被分割为M×N个局部矩阵,在各列中,配置在任一行上的局部矩阵构成相对于配置在除了所述任一行以外的其他行上的各局部矩阵的所述第一局部矩阵,且配置在所述其他行上的各局部矩阵构成所述第二局部矩阵。根据所述结构,能够与定标关系成立相应地降低推导模型的运算处理所耗费的计算量,抑制对计算资源造成的负担。而且,能够基于所述定标关系的成立来管理对推导模型的运算处理所耗费的计算量进行降低的量。另外,定标关系的成立可在构成第一局部矩阵的局部矩阵内反复进行至各局部矩阵成为任意尺寸的矩阵(例如1×1矩阵)为止。

[0032] 所述一方面的模型生成装置中,所述推导模型可包含神经网络。所述矩阵的各要素可构成为,与所述神经网络中的各神经元间的结合的权重对应。训练所述推导模型的处理可包含:反复进行构成所述第一局部矩阵的局部矩阵内的所述定标关系的成立直至各局部矩阵成为1×1矩阵为止,由此来调节所述矩阵的至少一部分的各要素的值,以使得在所述矩阵的至少一部分内,所述定标关系递归地成立。调节所述矩阵的至少一部分的各要素的值的处理可包含:以所述定标关系递归地成立的方式给予所述矩阵的至少一部分的各要素的初始值;在正向传播的阶段中,导出对所述各学习数据集的训练数据执行推导任务的结果;以及在反向传播的阶段中,对所导出的推导任务的执行结果以及由对应的正解标签所表示的正解之间的误差的梯度进行反向传播,由此来调节所述矩阵的至少一部分的各要素以及所述对角矩阵的各要素的值。所述正向传播的阶段的运算可包含:第一步骤,计算构成初始的第一局部矩阵的所述1×1矩阵及输入向量的积;第二步骤,计算所获得的所述初始的第一局部矩阵的积的结果以及所述对角矩阵的对应的要素的积,由此导出构成初始的第二局部矩阵的所述1×1矩阵及输入向量的积;第三步骤,根据在对象层级中获得的所述第一局部矩阵的积的结果以及所述第二局部矩阵的积的结果,导出构成后续层级的第一局部矩阵的局部矩阵及输入向量的积;第四步骤,计算所获得的后续层级的所述第一局部矩阵的积的结果以及所述对角矩阵的对应的要素的积,由此导出构成后续层级的所述第二局部矩阵的局部矩阵及输入向量的积;以及第五步骤,作为在所述对象层级中获得的所述第一局部矩阵的积的结果以及所述第二局部矩阵的积的结果各自的初始值,分别代入在所述第一步骤以及所述第二步骤中分别获得的所述初始的第一局部矩阵的积的结果以及所述初始的第二局部矩阵的积的结果,且递归地反复进行所述第三步骤及所述第四步骤,由此导出所述矩阵的至少一部分及输入向量的积。所述反向传播的阶段的运算可包含:第一步骤,获取相对于所述矩阵的至少一部分及输入向量的积的所述误差的梯度;第二步骤,基于所获取的所述误差的梯度、与在所述正向传播的阶段的递归性反复的最终层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的值,导出关于在所述最终层级中使用的对角矩阵的各要素的误差的梯度;第三步骤,基于所获取的所述误差的梯度以及在所述最终层级

中使用的对角矩阵的各要素的值，导出关于在所述最终层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度；第四步骤，基于所获取的所述误差的梯度、在所述正向传播的阶段的递归性反复的对象层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的值、以及关于在所述正向传播的阶段的递归性反复中在所述对象层级的后续层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度，导出关于在所述对象层级中使用的对角矩阵的各要素的误差的梯度；第五步骤，基于所获取的所述误差的梯度、在所述对象层级中使用的对角矩阵的各要素的值、以及关于在所述后续层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度，导出关于在所述对象层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度；第六步骤，作为关于在所述后续层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度的初始值，代入在所述反向传播的阶段的所述第三步骤中导出的、关于在所述最终层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度，且递归地反复进行所述反向传播的阶段的所述第四步骤及所述第五步骤，直至导出关于在所述正向传播的阶段的递归性反复的初始层级中使用的对角矩阵的各要素的误差的梯度为止，由此导出关于在各层级中使用的对角矩阵的各要素的误差的梯度；第七步骤，基于所述输入向量以及通过所述第六步骤而导出的、关于在所述初始层级中构成所述第一局部矩阵的局部矩阵及输入向量的积的误差的梯度，导出关于构成所述初始的第一局部矩阵的所述1×1矩阵的要素的误差的梯度；以及第八步骤，基于关于在所述各层级中使用的对角矩阵的各要素以及构成所述初始的第一局部矩阵的所述1×1矩阵的要素分别导出的所述误差的梯度，调节在所述各层级中使用的对角矩阵的各要素以及构成所述初始的第一局部矩阵的所述1×1矩阵的要素各自的值。根据所述结构，在采用神经网络来作为推导模型(机器学习模型)的场景中，能够在机器学习的处理过程中调节推导模型的各参数的值，以使定标关系递归地成立。其结果，即便在最终成立的定标关系中的第一局部矩阵内，也能够实现参数数量的降低。因而，根据所述结构，能够进一步降低推导模型的运算处理所耗费的计算量，进一步抑制对计算资源造成的负担。另外，此形态中，1×1矩阵可替换为任意尺寸的单位矩阵。

[0033] 而且，作为所述各形态的推导装置以及模型生成装置各自的其他形态，本发明的一方面也可为实现以上的各结构的全部或其一部分的信息处理方法，也可为程序，还可为存储有此种程序的、计算机或其他装置、机械等可读取的存储介质。此处，所谓计算机等可读取的存储介质，是指通过电、磁、光学、机械或化学作用来储存程序等信息的介质。而且，本发明的一方面也可为包含所述任一形态的推导装置以及模型生成装置的推导系统。

[0034] 例如，本发明的一方面的推导方法是一种信息处理方法，由计算机执行下述步骤：获取对象数据；使用通过机器学习而训练完毕的推导模型，对所获取的所述对象数据执行推导任务；以及输出与执行所述推导任务的结果相关的信息，所述信息处理方法中，所述推导模型的多个参数的至少一部分是以矩阵来表达，所述矩阵包含第一局部矩阵及第二局部矩阵，所述第一局部矩阵及所述第二局部矩阵的行及列各自的要素的数量相同，所述第二局部矩阵的各要素的值被调节为与所述第一局部矩阵及对角矩阵的积一致。

[0035] 例如，本发明的一方面的推导程序是一种程序，用于使计算机执行下述步骤：获取对象数据；使用通过机器学习而训练完毕的推导模型，对所获取的所述对象数据执行推导任务；以及输出与执行所述推导任务的结果相关的信息，所述程序中，所述推导模型的多个

参数的至少一部分是以矩阵来表达，所述矩阵包含第一局部矩阵及第二局部矩阵，所述第一局部矩阵及所述第二局部矩阵的行及列各自的要素的数量相同，所述第二局部矩阵的各要素的值被调节为与所述第一局部矩阵及对角矩阵的积一致。

[0036] 发明的效果

[0037] 根据本发明，能够降低机器学习模型的运算处理所耗费的计算量，抑制对计算资源造成的负担。

附图说明

[0038] [图1]图1示意性地例示适用本发明的场景的一例。

[0039] [图2]图2示意性地例示实施方式的模型生成装置的硬件结构的一例。

[0040] [图3]图3示意性地例示实施方式的推导装置的硬件结构的一例。

[0041] [图4]图4示意性地例示实施方式的模型生成装置的软件结构的一例。

[0042] [图5A]图5A表示第一局部矩阵及第二局部矩阵的设定方法的一例。

[0043] [图5B]图5B表示在4×4的参数矩阵内定标关系递归地成立的场景的一例。

[0044] [图5C]图5C表示定标关系递归地成立且在各层级第一行的局部矩阵构成第一局部矩阵时的要素分割的一般化的一例。

[0045] [图6]图6示意性地例示实施方式的推导装置的软件结构的一例。

[0046] [图7]图7是表示实施方式的模型生成装置的处理流程的一例的流程图。

[0047] [图8A]图8A表示正向传播的阶段中的参数矩阵(2×2矩阵)的运算过程的一例。

[0048] [图8B]图8B表示反向传播的阶段中的参数矩阵(2×2矩阵)的运算过程的一例。

[0049] [图9A]图9A表示正向传播的阶段中的参数矩阵(4×4矩阵)的运算过程的一例。

[0050] [图9B]图9B表示反向传播的阶段中的参数矩阵(4×4矩阵)的运算过程的一例。

[0051] [图10A]图10A表示正向传播的阶段中的参数矩阵(任意尺寸的矩阵)的运算过程的一例。

[0052] [图10B]图10B表示反向传播的阶段中的参数矩阵(任意尺寸的矩阵)的运算过程的一例。

[0053] [图11]图11是表示实施方式的推导装置的处理流程的一例的流程图。

[0054] [图12]图12示意性地例示适用本发明的其他场景的一例。

[0055] [图13]图13示意性地例示适用本发明的其他场景的一例。

[0056] [图14]图14示意性地例示适用本发明的其他场景的一例。

具体实施方式

[0057] 以下，基于附图来说明本发明的一方面的实施方式(以下也称作“本实施方式”)。但是，以下说明的本实施方式在所有方面不过是本发明的示例。当然可不脱离本发明的范围而进行各种改良或变形。即，在本发明的实施时，也可适当采用与实施方式相应的具体结构。另外，通过自然语言说明了本实施方式中出现的数据，但更具体而言，是以计算机可识别的伪语言、命令、参数、机器语言等来指定。

[0058] §1适用例

[0059] 图1示意性地例示适用了本发明的场景的一例。如图1所示，本实施方式的推导系

统100包括模型生成装置1及推导装置2。

[0060] 本实施方式的模型生成装置1是构成为生成通过机器学习而训练完毕的推导模型5的计算机。本实施方式中，模型生成装置1获取多个学习数据集3。各学习数据集3包含训练数据31及正解标签32的组合。训练数据31可根据推导任务等适当构成。正解标签32构成为表示针对训练数据31的推导任务的正解(真值)。

[0061] 可选择对数据中所含的特征进行推导的任意种类的任务来作为推导模型5的推导任务。作为一例，在生成用于进行基于图像的外观检查的训练完毕推导模型5的情况下，推导任务可为：判定映照在图像数据中的产品是否存在缺陷(例如可包含缺陷的有无、缺陷的种类、存在缺陷的范围的提取等任务)。此时，训练数据31可包含映照有产品的图像数据，正解标签32可构成为表示针对所关联的训练数据31的所述推导任务的正解(例如缺陷的有无、缺陷的种类、存在缺陷的范围等)。模型生成装置1使用多个学习数据集3来实施推导模型5的机器学习。

[0062] 推导模型5包含包括多个参数的机器学习模型。机器学习模型的种类可无特别限定，可根据实施方式来适当选择。对于推导模型5，例如可采用神经网络、通过主成分分析获得的主成分向量、支持向量机等。以下，为了便于说明，表示采用了神经网络作为构成推导模型5的机器学习模型的一例(后述的图4)。在推导模型5包含神经网络的情况下，各神经元(节点)间的结合的权重为参数的一例。

[0063] 本实施方式中，推导模型5的多个参数的至少一部分构成为可通过矩阵来表达。矩阵构成为包含第一局部矩阵及第二局部矩阵。第一局部矩阵及第二局部矩阵构成为，行及列各自的要素的数量相同。各局部矩阵的行及列的数量分别可根据实施方式来适当选择。行的数量与列的数量既可一致，或者也可互不相同。第一局部矩阵及第二局部矩阵也可为标量(1×1的矩阵)。

[0064] 机器学习是通过下述方式而构成，即，关于各学习数据集3来训练推导模型5，以使得使用推导模型5对训练数据31执行推导任务的结果符合由正解标签32所表示的正解，并且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致。由此，能够生成获得了执行推导任务的能力且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致的、训练完毕的推导模型5。

[0065] 图1的示例中，w_1-w_4为以矩阵来表达的参数的一例。这些中，w_1以及w_2为第一局部矩阵的一例。w_3以及w_4为第二局部矩阵的一例。(y_1,y_2)表示推导模型5中的、以此矩阵来表达的构成部分的运算结果(输出)。(x_1,x_2)表示对所述构成部分的输入。作为一例，在构成神经网络的各层通过所述矩阵来表达的情况下，对各层的输入对应于(x_1,x_2)，来自各层的输出对应于(y_1,y_2)。(a_1,a_2)为对角矩阵的对角成分的要素的一例。在对角矩阵的对角成分的各要素的值相同的情况下，对角矩阵可用一个值(标量)来表达。

[0066] 另一方面，本实施方式的推导装置2是构成为使用训练完毕的推导模型5来执行推导任务的计算机。本实施方式中，推导装置2获取对象数据221。接下来，推导装置2使用通过机器学习而训练完毕的推导模型5，对所获取的对象数据221执行推导任务。如上所述，推导模型5的多个参数的至少一部分是以矩阵来表达。所述矩阵包含第一局部矩阵及第二局部矩阵。第一局部矩阵及第二局部矩阵的行及列各自的要素的数量相同。第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致。推导装置2输出与执行推导任务

的结果相关的信息。

[0067] 如上所述，本实施方式中，推导模型5的多个参数的至少一部分可通过包含第一局部矩阵及第二局部矩阵的矩阵来表达。根据本实施方式的模型生成装置1，在训练完毕的推导模型5中，第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致。因此，即便未保持第二局部矩阵的信息，也能够回头利用第一局部矩阵的运算结果来获得第二局部矩阵的运算结果。图1的示例中，即便不执行w_3x_1以及w_4x_2的运算，也能够根据第一局部矩阵的运算结果(w_1x_1及w_2x_2)来获得第二局部矩阵的运算结果($a_1w_1x_1$及$a_2w_2x_2$)。因此，根据本实施方式，能够实质上降低构成推导模型5的参数的数量。具体而言，能够将参数的数量降低与第二局部矩阵的要素数与对角矩阵的0以外的要素数的差值相应的量。由此，能够降低推导模型5的运算处理所耗费的计算量，抑制对计算资源造成的负担。

[0068] 另外，图1的示例中，模型生成装置1及推导装置2经由网络相互连接。网络的种类例如可从国际互联网、无线通信网、移动通信网、电话网、专用网等中适当选择。但是，在模型生成装置1及推导装置2之间交换数据的方法也可不限定于此种示例，可根据实施方式来适当选择。例如，在模型生成装置1及推导装置2之间，可利用存储介质来交换数据。

[0069] 而且，图1的示例中，模型生成装置1及推导装置2分别包含独立的计算机。但是，本实施方式的推导系统100的结构也可不限定于此种示例，可根据实施方式来适当决定。例如，模型生成装置1及推导装置2也可为一体的计算机。而且，例如模型生成装置1及推导装置2中的至少一者也可包含多台计算机。

[0070] §2结构例

[0071] [硬件结构]

[0072] ＜模型生成装置＞

[0073] 图2示意性地例示本实施方式的模型生成装置1的硬件结构的一例。如图2所示，本实施方式的模型生成装置1是由控制部11、存储部12、通信接口13、外部接口14、输入装置15、输出装置16以及驱动器17电连接而成的计算机。另外，图2中，将通信接口以及外部接口记作“通信I/F”以及“外部I/F”。

[0074] 控制部11包含作为硬件处理器的中央处理器(Central Processing Unit，CPU)、随机存取存储器(Random Access Memory，RAM)、只读存储器(Read Only Memory，ROM)等，且构成为基于程序以及各种数据来执行信息处理。存储部12为存储器的一例，例如包含硬盘驱动器、固态硬盘等。本实施方式中，存储部12存储模型生成程序81、多个学习数据集3、学习结果数据125等各种信息。

[0075] 模型生成程序81是用于使模型生成装置1执行后述的机器学习的信息处理(图7)的程序，所述机器学习生成训练完毕的推导模型5。模型生成程序81包含所述信息处理的一连串命令。多个学习数据集3被用于训练完毕的推导模型5的生成。学习结果数据125表示与所生成的训练完毕的推导模型5相关的信息。本实施方式中，学习结果数据125是作为执行模型生成程序81的结果而生成。详情将后述。

[0076] 通信接口13例如为有线局域网(Local Area Network，LAN)模块、无线LAN模块等，是用于进行经由网络的有线或无线通信的接口。模型生成装置1可利用通信接口13来与其他信息处理装置之间执行经由网络的数据通信。外部接口14例如为通用串行总线(Universal Serial Bus，USB)端口、专用端口等，是用于与外部装置连接的接口。外部接口

14的种类以及数量可任意选择。模型生成装置1可经由通信接口13以及外部接口14的至少一者连接于用于获得训练数据31的设备(例如在训练数据31为图像数据的情况下,为摄像机)。

[0077] 输入装置15例如为鼠标、键盘等用于进行输入的装置。而且,输出装置16例如为显示器、扬声器等用于进行输出的装置。用户等作业员通过利用输入装置15以及输出装置16,从而可操作模型生成装置1。

[0078] 驱动器17例如为光盘(Compact Disc,CD)驱动器、数字多功能光盘(Digital Versatile Disc,DVD)驱动器等,是用于读取存储在存储介质91中的程序等各种信息的驱动装置。存储介质91是以计算机或其他装置、机械等能够读取所存储的程序等各种信息的方式而通过电、磁、光学、机械或化学作用来保存所述程序等信息的介质。所述模型生成程序81以及多个学习数据集3的至少任一者也可被存储于存储介质91中。模型生成装置1也可从所述存储介质91获取所述模型生成程序81以及多个学习数据集3的至少任一者。另外,图2中,作为存储介质91的一例,例示了CD、DVD等盘型的存储介质。但是,存储介质91的种类也可不限于盘型,也可为盘型以外。作为盘型以外的存储介质,例如可列举快闪存储器等半导体存储器。驱动器17的种类可根据存储介质91的种类来任意选择。

[0079] 另外,关于模型生成装置1的具体硬件结构,可根据实施方式来适当地进行构成元件的省略、替换以及追加。例如,控制部11也可包含多个硬件处理器。硬件处理器可包含微处理器、现场可编程门阵列(Field-Programmable Gate Array,FPGA)、数字信号处理器(Digital Signal Processor,DSP)等。存储部12也可包含控制部11中所含的RAM以及ROM。通信接口13、外部接口14、输入装置15、输出装置16以及驱动器17的至少任一者也可予以省略。模型生成装置1也可包含多台计算机。此时,各计算机的硬件结构既可一致,也可不一致。而且,模型生成装置1除了专为所提供的服务设计的信息处理装置以外,也可为通用的服务器装置、个人计算机(Personal Computer,PC)、工业个人计算机(Industrial Personal Computer,IPC)等。

[0080] ＜推导装置＞

[0081] 图3示意性地例示本实施方式的推导装置2的硬件结构的一例。如图3所示,本实施方式的推导装置2是由控制部21、存储部22、通信接口23、外部接口24、输入装置25、输出装置26以及驱动器27电连接而成的计算机。

[0082] 推导装置2的控制部21～驱动器27以及存储介质92可分别与所述模型生成装置1的控制部11～驱动器17以及存储介质91各自同样地构成。控制部21包含作为硬件处理器的CPU、RAM、ROM等,且构成为基于程序以及数据来执行各种信息处理。存储部22例如包含硬盘驱动器、固态硬盘等。本实施方式中,存储部22存储推导程序82、学习结果数据125等各种信息。

[0083] 推导程序82是用于使推导装置2执行使用训练完毕的推导模型5来执行推导任务的、后述的信息处理(图11)的程序。推导程序82包含所述信息处理的一连串命令。推导程序82以及学习结果数据125的至少任一者也可被存储于存储介质92中。而且,推导装置2也可从存储介质92获取推导程序82以及学习结果数据125的至少任一者。

[0084] 推导装置2可经由通信接口23及外部接口24的至少任一者连接于用于获取对象数据221的设备(传感器、以及计算机等)。推导装置2可通过输入装置25及输出装置26的利用

来受理来自用户等作业员的操作及输入。

[0085] 另外,关于推导装置2的具体硬件结构,可根据实施方式来适当地进行构成元件的省略、替换以及追加。例如,控制部21也可包含多个硬件处理器。硬件处理器可包含微处理器、FPGA、DSP等。存储部22也可包含控制部21中所含的RAM以及ROM。通信接口23、外部接口24、输入装置25、输出装置26以及驱动器27的至少任一者也可予以省略。推导装置2也可包含多台计算机。此时,各计算机的硬件结构既可一致,也可不一致。而且,推导装置2除了专为所提供的服务设计的信息处理装置以外,也可为通用的服务器装置、通用的PC、工业PC、可编程逻辑控制器(Programmable Logic Controller,PLC)等。

[0086] [软件结构]

[0087] ＜模型生成装置＞

[0088] 图4示意性地例示本实施方式的模型生成装置1的软件结构的一例。模型生成装置1的控制部11将存储于存储部12的模型生成程序81展开到RAM中。并且,控制部11通过CPU来解释及执行在RAM中展开的模型生成程序81中所含的命令,以控制各构成元件。由此,如图4所示,本实施方式的模型生成装置1作为包括数据获取部111、学习处理部112以及保存处理部113作为软件模块的计算机而运行。即,本实施方式中,模型生成装置1的各软件模块是通过控制部11(CPU)而实现。

[0089] 数据获取部111构成为获取多个学习数据集3。各学习数据集3包含训练数据31及正解标签32的组合。训练数据31可根据推导任务等来适当构成。正解标签32构成为表示针对训练数据31的推导任务的正解(真值)。

[0090] 学习处理部112构成为,使用多个学习数据集3来实施推导模型5的机器学习。推导模型5包含包括多个参数的机器学习模型。只要推导模型5的多个参数的至少一部分可通过矩阵来表达,则推导模型5的种类可无特别限定,可根据实施方式来适当选择。

[0091] 本实施方式中,推导模型5的多个参数的至少一部分构成为可通过矩阵来表达。矩阵(以下也称作参数矩阵)构成为包含第一局部矩阵及第二局部矩阵。第一局部矩阵及第二局部矩阵构成为,行及列各自的要素的数量相同。机器学习是通过下述方式而构成,即,关于各学习数据集3来训练推导模型5,以使得使用推导模型5来对训练数据31执行推导任务的结果符合由正解标签32所表示的正解,并且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致。

[0092] 保存处理部113构成为,生成与通过机器学习而生成的训练完毕的推导模型5相关的信息来作为学习结果数据125,并将所生成的学习结果数据125保存至规定的存储区域。学习结果数据125可适当构成为,包含用于再现训练完毕的推导模型5的信息。

[0093] (推导模型)

[0094] 如图4所示,推导模型5的一例可包含神经网络。图4的示例中,构成推导模型5的神经网络包括输入层51、一个以上的中间(隐藏)层52以及输出层53。中间层52的数量可根据实施方式来适当决定。各层51～53包括一个或多个神经元(节点)。各层51～53中所含的神经元的数量可根据实施方式来适当决定。而且,各层51～53中所含的神经元间的连接关系也可根据实施方式来适当决定。一例中,各层51～53中所含的各神经元可与邻接的层的所有神经元相结合。由此,推导模型5可包含全结合型神经网络。

[0095] 对于各层51～53的各结合,设定有权重(结合负载)。对于各神经元设定有阈值,基

本上，根据各输入与各权重的积的和是否超过阈值来决定各神经元的输出。阈值可通过激活函数来表达。此时，通过将各输入与各权重的积的和输入至激活函数，执行激活函数的运算，从而决定各神经元的输出。激活函数的种类可任意选择。各层51～53中所含的各神经元间的结合的权重为推导模型5的参数的一例。即，在推导模型5包含神经网络的情况下，参数矩阵的各要素可构成为与神经网络中的各神经元间的结合的权重对应。

[0096] 另外，在推导模型5包含图4中例示的神经网络的情况下，各层51～53的参数可通过各自的矩阵来表达。此时，根据机器学习的结果，构成神经网络的各层51～53中的至少任一层的至少一部分可包含参数矩阵，所述参数矩阵包含第一局部矩阵及第二局部矩阵。一例中，各层51～53可包含互不相同的参数矩阵。此时，各参数矩阵中的所述(x_1,x_2)对应于针对各层51～53的输入数据，所述(y_1,y_2)对应于各层51～53的运算结果(输出)。

[0097] 在机器学习中，学习处理部112使用各学习数据集3中的训练数据31来作为输入数据，使用正解标签32来作为教学信号。即，学习处理部112将各学习数据集3的训练数据31输入至输入层51，执行推导模型5的正向传播的运算处理(正向传播的阶段)。作为所述正向传播的运算处理的结果，学习处理部112从输出层53获取与对训练数据31执行推导任务的结果(即，对推导任务的解进行推导的结果)对应的输出值。从输出层53获得的输出值的格式只要可确定推导结果，则可无特别限定，可根据实施方式来适当决定。一例中，输出层53的输出值也可构成为直接表示推导结果。另一例中，输出层53的输出值构成为，通过经由阈值处理等任意的运算处理来间接地表示推导结果。

[0098] 学习处理部112算出所获得的推导任务的执行结果与由跟所输入的训练数据31关联的正解标签32所表示的之间的误差，并进一步算出所算出的误差的梯度。并且，学习处理部112通过误差反向传播法，对所算出的误差的梯度进行反向传播，以调节各参数的值(反向传播的阶段)。在所述调节时，学习处理部112进行调节，以使第二局部矩阵的各要素的值与第一局部矩阵及对角矩阵的积一致。学习处理部112关于各学习数据集3反复进行调节各参数的值以使所算出的误差之和变小的、所述一连串处理。作为所述机器学习的结果，能够生成获得了执行推导任务的能力且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致的、训练完毕的推导模型5。

[0099] 保存处理部113生成用于再现通过所述机器学习而生成的训练完毕的推导模型5的学习结果数据125。只要能够再现训练完毕的推导模型5，则学习结果数据125的结构可无特别限定，可根据实施方式来适当决定。作为一例，学习结果数据125可包含表示通过所述机器学习的调节而获得的各参数的值的信息。此时，第二局部矩阵的各要素的值与第一局部矩阵及对角矩阵的积一致，因此表示构成第二局部矩阵的参数的值的信息可予以省略。即，学习结果数据125可构成为包含表示通过所述机器学习的调节而获得的构成第一局部矩阵的参数的值以及对角矩阵的对角成分的各值的信息。根据情况，学习结果数据125也可进而包含表示推导模型5的结构的信息。结构例如可根据从输入层直至输出层为止的层数、各层的种类、各层中所含的神经元的数量、邻接的层的神经元彼此的结合关系等来确定。保存处理部113将所生成的学习结果数据125保存至规定的存储区域。

[0100] (第一局部矩阵及第二局部矩阵之间的关系)

[0101] 此处，对第一局部矩阵及第二局部矩阵之间的关系的一例进行说明。本实施方式中，只要参数矩阵的一部分要素对应于第一局部矩阵，剩余中的至少一部分要素对应于第

二局部矩阵(即,只要在训练完毕的推导模型5的参数矩阵的至少一部分中有比例关系成立),便可获得削减参数数量的效果。因此,所述比例关系只要在参数矩阵的至少一部分中成立即可。但为了有效地削减参数的数量,优选的是,在参数矩阵内,相当于第一局部矩阵及第二局部矩阵(尤其是第二局部矩阵)的范围大。

[0102] 图5A表示用于有效地削减参数数量的、第一局部矩阵及第二局部矩阵的设定方法的一例。所述设定方法中,在参数矩阵的至少一部分中,以通过以下的条件(1)～(3)所定义的定标关系成立的方式设定第一局部矩阵及第二局部矩阵。

[0103] • 条件(1):参数矩阵的至少一部分以行及列各自的要素的数量相同的局部矩阵在行上各排列M个且在列上各排列N个的方式分割为M×N个局部矩阵。

[0104] • 条件(2):在各列中,配置在任一行上的局部矩阵构成与配置在除了所述任一行以外的其他行上的各局部矩阵对应的第一局部矩阵。

[0105] • 条件(3):配置在其他行上的各局部矩阵构成第二局部矩阵。

[0106] 图5A的示例中,W对应于参数矩阵的至少一部分,X对应于针对所述一部分的输入数据,Y对应于所述一部分的运算结果(输出)。w_1-w_{MN}相当于M×N个局部矩阵。各局部矩阵w_1-w_{MN}的行及列各自的要素的数量只要在各局部矩阵w_1-w_{MN}间相同,则可无特别限定,可在实施方式中适当决定。各局部矩阵w_1-w_{MN}也可为1×1矩阵(标量)。各输入(x_1-x_N)的要素数对应于各局部矩阵w_1-w_{MN}的列的要素数。

[0107] 图5A的示例中,各局部矩阵w_1-w_{MN}中的第一行的各局部矩阵w_1-w_N构成相对于各列中的其他行(第二行以后的各行)的局部矩阵的第一局部矩阵。并且,剩余的各局部矩阵w_{N+1}-w_{MN}构成相对于相应列中的第一行的局部矩阵(第一局部矩阵)的第二局部矩阵。即,第一行的局部矩阵w_t构成相对于第二行以后的各行的局部矩阵w_{rN+t}的第一局部矩阵,各局部矩阵w_{rN+t}构成相对于局部矩阵w_t的第二局部矩阵(t为1至N的自然数,r为1至M-1的自然数)。$a_{2,1}$-$a_{M,N}$为对角矩阵。$a_{2,1}$-$a_{M,N}$的要素数对应于局部矩阵w_1-w_N的行的要素数。

[0108] M及N的各值只要为2以上的自然数,则可无特别限定,可根据实施方式来适当决定。一例中,M及N可为相同的素数S。作为具体例,M及N可分别为2(图1的示例)。此时,参数矩阵的至少一部分的行及列分别被一分为二,从而能够容易地掌握与第一局部矩阵及第二局部矩阵分别对应的局部矩阵。另一例中,M及N可为互不相同的值。

[0109] 另外,图5A的示例中,配置在各列的第一行上的局部矩阵构成相对于配置在各列的其他行(第二行以后的各行)上的局部矩阵的第一局部矩阵,配置在各列的第二行以后的各行上的局部矩阵构成第二局部矩阵。但是,第一局部矩阵及第二局部矩阵的配置可不限定于此种示例。各列中,第一局部矩阵可配置在第二行以后的任一行上。而且,图5A的示例中,各列的同一行的局部矩阵构成第一局部矩阵,但第一局部矩阵的配置可不限定于此种示例。即,在各列之间,构成第一局部矩阵的局部矩阵所配置的行也可不同(例如也可为,在第一列中,第一行的局部矩阵构成第一局部矩阵,与此相对,在第二列中,第二行以后的任一行的局部矩阵构成第一局部矩阵)。

[0110] 为了更有效地削减参数的数量,也可为,在参数矩阵的至少一部分内,所述定标关系递归地成立(即,也可为,在各局部矩阵内进而有所述定标关系成立)。定标关系递归地成立的次数可无特别限定,可根据实施方式来适当决定。定标关系的成立可反复进行至各局部矩阵成为任意尺寸的矩阵为止。此时,任意尺寸的矩阵可称作“单位矩阵”。但是,为了更

有效地削减参数的数量，理想的是，以下述方式使定标关系递归地成立，即，构成第一局部矩阵的局部矩阵内的定标关系的成立反复进行至各局部矩阵成为1×1矩阵(标量)为止。由此，能够有效地降低推导模型5的运算处理所耗费的计算量，抑制对计算资源造成的负担。

[0111] 图5B表示在4×4的参数矩阵内，定标关系递归地成立至各局部矩阵成为1×1矩阵为止的场景的一例。首先，4×4的参数矩阵W被分割为2×2个局部矩阵(尺寸为2×2)。其中，各列的第一行的W_1及W_2构成第一局部矩阵，各列的第二行的局部矩阵构成第二局部矩阵。A1及A2(尺寸为2×2)为在此层级中使用的对角矩阵。由此，在4×4的参数矩阵内，定标关系成立。并且，各第一局部矩阵(W_1,W_2)进一步被分割为2×2个局部矩阵(尺寸为1×1)。其中，各局部矩阵内的各列的第一行的局部矩阵(w_1-w_4)构成第一局部矩阵，各列的第二行的局部矩阵构成第二局部矩阵。a_1-a_4(尺寸为1×1)为在此层级中使用的对角矩阵。在各第一局部矩阵(W_1,W_2)内，进而有定标关系成立。对角矩阵的信息只要仅保持对角成分即可。因此，图5B的示例中，能够将W中的16个参数削减至12个(w_1-w_4、a_1-a_4、A_1、A_2)为止。

[0112] 另外，图5B的示例中，定标关系在两层级中递归地成立，各层级中的分割数M×N分别相同(2×2)。但是，递归地成立的次数可不限定于此种示例，可根据实施方式来适当决定。而且，各层级中的分割数M×N也可在至少一部分中不同。例如也可为，在第一层级的分割数M_1×N_1与第二层级的分割数M_2×N_2之间，M_i及N_i的至少其中任一者不同。

[0113] 图5C表示定标关系递归地成立且在各层级第一行的局部矩阵构成第一局部矩阵时的要素分割的一般化的一例。在假定为W_j^i被分割为p_{i-1}×q_{i-1}个局部矩阵(也记载为小矩阵)，且在各层级中第一行的局部矩阵构成第一局部矩阵的情况下，可通过图5C的数式来表达各层级中的相应矩阵W_j^i内的小矩阵。I对应于递归的层级。在i中代入1至l+1为止的自然数。图5C的要素分割在i为2至l+1为止的范围内成立。L表示定标关系递归地成立的次数。J对应于各层级中的相应矩阵的配置列的编号。在j中代入1至J_i为止的自然数。J_i是通过以下的式1来表示。

[0114] [数1]

[0115] $$J_i=\begin{cases}1 & (i=l+1)\\ \prod_{e=i}^{l} q_e & (1\le i\le l)\end{cases} \quad \cdots(式1)$$

[0116] W_1^{l+1}对应于通过定标关系来分割要素之前的参数矩阵的至少一部分(即，参数矩阵中的定标关系递归地成立的范围的部分)。W_j^1-W_j^1对应于各层级中的第一局部矩阵。k(i,j)(=q_{i-1}×(j-1))是用于在与同一层级的其他矩阵(例如W_{j-1}^i)之间以连续编号来显示相应矩阵W_j^i内的小矩阵的偏置。A_g^{i-1}为尺寸M_{i-1}×M_{i-1}的对角矩阵。G为(2,k(i,j)+1)-(p_{i-1},k(i,j)+q_{i-1})。在与一层级前的要素数的关系上，W_j^i的行的要素数M_i可表达为M_{i+1}/p_i，W_j^i的列的要素数N_i可表达为N_{i+1}/q_i。最终分割中的W_j^1可为1×1矩阵(标量)。此时，W_1^{l+1}的行的要素数为p_1×…×p_1，W_1^{l+1}的列的要素数为q_1×…×q_1。在定标关系像这样递归地成立的情况下，只要将最终分割中的W_j^1以及各层级的对角矩阵A_g^1-A_g^1的信息保持作为学习结果数据125，便可再现训练完毕的推导模型5而执行参数矩阵的至少一部分的运算处理。即，可将参数矩阵的至少一部分的参数数量由p_1×…×p_1×q_1×…×q_1削减至W_j^1以及各层级的对角矩阵A_g^1-A_g^1的要素数为止。

[0117] 各层级中的p_{i-1}以及q_{i-1}对应于各层级中的分割数M及N。p_{i-1}以及q_i-1的值可无特

别限定，可根据实施方式来适当决定。各层级中的分割数既可在各层级间相同，或者也可不同。p_{i-1}以及q_{i-1}的值既可彼此相同，或者也可不同。作为一例，p_{i-1}以及q_i-1可为相同的素数S，在各层级间，分割数也可相同。例如，素数S可为2。此时，参数矩阵的至少一部分(W_1^{l+1})包含要素的数量为素数S的幂(S^{21})的正方矩阵。由此，能够容易地掌握与第一局部矩阵及第二局部矩阵分别对应的局部矩阵。

[0118] 另外，以下，为了便于说明，只要未特别明示，则假定构成各层级中的各列的第一局部矩阵的局部矩阵的配置为第一行，第二行以后的局部矩阵构成第二局部矩阵。但是，构成各层级中的各列的第一局部矩阵的局部矩阵的配置也可不限于第一行。各层级中的各列的第一局部矩阵可配置在第二行以后的任一行。各列的第一局部矩阵的配置既可在各层级间相同，或者也可不同。而且，最终分割中的W_j^1可不限于1×1矩阵。另一例中，最终分割中的W_j^1可为任意尺寸的矩阵。

[0119] ＜推导装置＞

[0120] 图6示意性地例示本实施方式的推导装置2的软件结构的一例。推导装置2的控制部21将存储在存储部22中的推导程序82展开到RAM中。并且，控制部21通过CPU来解释及执行在RAM中展开的推导程序82中所含的命令，从而控制各构成元件。由此，如图6所示，本实施方式的推导装置2作为包括数据获取部211、推导部212及输出部213作为软件模块的计算机而运行。即，本实施方式中，推导装置2的各软件模块也与模型生成装置1同样地，通过控制部21(CPU)来实现。

[0121] 数据获取部211构成为获取对象数据221。推导部212通过保持有学习结果数据125，从而包括通过机器学习而训练完毕的推导模型5。推导部212构成为，使用训练完毕的推导模型5来对所获取的对象数据221执行推导任务。输出部213构成为，输出与执行推导任务的结果相关的信息。

[0122] ＜其他＞

[0123] 关于模型生成装置1及推导装置2的各软件模块，将在后述的动作例中详细说明。另外，本实施方式中，对模型生成装置1及推导装置2的各软件模块均通过通用的CPU来实现的示例进行了说明。但是，也可为，所述软件模块的一部分或全部通过一个或多个专用的处理器(例如图形处理器)来实现。所述各模块也可作为硬件模块而实现。而且，关于模型生成装置1及推导装置2各自的软件结构，也可根据实施方式来适当地进行软件模块的省略、替换及追加。

[0124] §3动作例

[0125] [模型生成装置]

[0126] 图7是表示与本实施方式的模型生成装置1所进行的机器学习相关的处理流程的一例的流程图。以下说明的模型生成装置1的处理流程为模型生成方法的一例。但以下说明的模型生成装置1的处理流程不过是一例，各步骤可尽可能地变更。而且，对于以下的处理流程，可根据实施方式来适当地进行步骤的省略、替换及追加。

[0127] (步骤S101)

[0128] 步骤S101中，控制部11作为数据获取部111而运行，获取多个学习数据集3，所述多个学习数据集3分别包含训练数据31及正解标签32的组合，所述正解标签32表示针对所述训练数据31的推导任务的正解。

[0129] 各学习数据集3可适当生成。例如可在实际空间或虚拟空间中设想执行推导任务的环境，通过在所述环境中观测任意对象来获取训练数据31(训练样本)。观测对象可根据推导任务来适当选择。对于训练数据31的获取，可使用任意的传感器。作为一例，在推导任务为产品的外观检查的情况下，可通过摄像机来拍摄存在或不存在缺陷的产品，由此来获取训练数据31。接下来，获取表示对所获取的训练数据31执行推导任务的结果(正解/真值)的信息作为正解标签32。在此场景中，推导任务的执行可通过作业员等的人工来进行。并且，将所获得的正解标签32关联至所述训练数据31。由此，能够生成各学习数据集3。

[0130] 各学习数据集3既可通过计算机的动作而自动地生成，或者也可通过至少局部包含作业员的操作而手动地生成。而且，各学习数据集3的生成既可由模型生成装置1来进行，也可由模型生成装置1以外的其他计算机来进行。在由模型生成装置1生成各学习数据集3的情况下，控制部11自动地、或者通过作业员经由输入装置15的操作来手动地执行所述一连串的生成处理，由此来获取多个学习数据集3。另一方面，在由其他计算机生成各学习数据集3的情况下，控制部11例如经由网络、存储介质91等来获取由其他计算机所生成的多个学习数据集3。也可为，一部分学习数据集3由模型生成装置1生成，而其他的学习数据集3由一个或多个其他计算机生成。

[0131] 要获取的学习数据集3的件数可任意选择。当获取多个学习数据集3时，控制部11将处理推进至接下来的步骤S102。

[0132] (步骤S102)

[0133] 步骤S102中，控制部11作为学习处理部112而运行，使用多个学习数据集3来实施推导模型5的机器学习。本实施方式中，机器学习是通过下述方式而构成，即，关于各学习数据集3来训练推导模型5，以使得使用推导模型5对训练数据31执行推导任务的结果符合由正解标签32所表示的正解，并且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致。

[0134] 在生成所述定标关系成立的、训练完毕的推导模型5的情况下，训练推导模型5的处理包含：调节参数矩阵的至少一部分的各要素的值，以使定标关系成立。进而，在生成所述定标关系递归地成立的、训练完毕的推导模型5的情况下，训练推导模型5的处理包含：反复进行构成第一局部矩阵的局部矩阵内的定标关系的成立，由此来调节参数矩阵的至少一部分的各要素的值，以使得在参数矩阵的至少一部分内，定标关系递归地成立。所述定标关系的递归性的成立可反复进行至各局部矩阵成为1×1矩阵(标量)为止。即，通过基于定标关系的最终分割所获得的第一局部矩阵(W_j^1)可为1×1矩阵。机器学习的方法可根据构成推导模型5的机器学习模型的种类来适当选择。

[0135] 一例中，推导模型5的训练(调节参数矩阵的至少一部分的各要素的值)包含：

[0136] • 进行推导模型5的初始设定(对各参数给予初始值)；

[0137] • 在正向传播的阶段中，导出对各学习数据集3的训练数据31试行地执行推导任务的结果；以及

[0138] • 在反向传播的阶段中，对所导出的推导任务的执行结果以及由对应的正解标签32所表示的正解之间的误差的梯度进行反向传播，由此来调节推导模型5的参数(参数矩阵的至少一部分的各要素以及对角矩阵的各要素)的值。

[0139] (初始设定)

[0140] 首先，控制部11进行作为机器学习处理对象的推导模型5的初始设定。本实施方式中，控制部11对于神经网络的结构(例如层数、各层的种类、各层中所含的神经元的数量、邻接的层的神经元彼此的结合关系等)以及各神经元间的结合的权重的初始值，既可通过模板来给予，也可通过作业员的输入来给予。而且，在进行再学习的情况下，控制部11也可基于通过过去的机器学习所获得的学习结果数据来进行神经网络的初始设定。

[0141] 本实施方式中，参数矩阵的各要素对应于各神经元间的结合的权重。因此，给予各神经元间的结合的权重的初始值对应于给予参数矩阵的各要素的初始值。第二局部矩阵是以第一局部矩阵及对角矩阵的积来表达，因此给予与第二局部矩阵对应的要素的初始值的处理可予以省略。即，给予各神经元间的结合的权重的初始值的处理可通过下述处理来构成，即，给予与第一局部矩阵对应的要素以及对角矩阵的对角成分的各要素的初始值。

[0142] 在生成定标关系递归地成立的、训练完毕的推导模型5的情况下，控制部11以定标关系递归地成立的方式来给予参数矩阵的至少一部分的各要素的初始值。作为具体例，控制部11给予在最终分割(即，最终成立的定标关系)中构成第一局部矩阵的W_j^1的各要素以及在各层级中使用的对角矩阵的各要素的初始值。

[0143] (正向传播的阶段)

[0144] 接下来，在正向传播的阶段中，控制部11将各学习数据集3的训练数据31输入至输入层51，执行推导模型5的正向传播的运算处理。在所述正向传播的运算处理时，首先，控制部11计算构成第一局部矩阵的局部矩阵及输入向量的积。在参数矩阵对应于输入层51的情况下，输入向量为各学习数据集3的训练数据31。在参数矩阵对应于中间层52或输出层53的情况下，输入向量为对应的层之前的层的计算结果(输出)。接下来，控制部11计算第一局部矩阵的积的结果以及对角矩阵的对应的要素的积，由此导出构成第二局部矩阵的局部矩阵及输入向量的积。即，控制部11将第一局部矩阵的积的结果以及对角矩阵相乘，获取通过相乘所得的计算结果来作为第二局部矩阵及输入向量的积的结果。

[0145] 在生成定标关系递归地成立的、训练完毕的推导模型5的情况下，正向传播的阶段的运算可包含以下的第一步骤～第五步骤的运算。

[0146] • 第一步骤：在使定标关系递归地成立的参数矩阵的至少一部分内，计算构成初始的第一局部矩阵的局部矩阵及输入向量的积

[0147] • 第二步骤：计算所获得的初始的第一局部矩阵的积的结果以及对角矩阵的对应的要素的积，由此导出构成初始的第二局部矩阵的局部矩阵及输入向量的积

[0148] • 第三步骤：根据在递归性反复的对象层级中获得的第一局部矩阵的积的结果以及第二局部矩阵的积的结果，导出构成后续层级的第一局部矩阵的局部矩阵及输入向量的积

[0149] • 第四步骤：计算所获得的后续层级的第一局部矩阵的积的结果以及对角矩阵的对应的要素的积，由此导出构成后续层级的第二局部矩阵的局部矩阵及输入向量的积

[0150] • 第五步骤：作为在对象层级中获得的第一局部矩阵的积的结果以及第二局部矩阵的积的结果各自的初始值，分别代入在第一步骤及第二步骤中分别获得的初始的第一局部矩阵的积的结果以及初始的第二局部矩阵的积的结果，且递归地反复进行(即，将所获得的后续层级的各局部矩阵的积的结果作为新的对象层级的各局部矩阵的积的结果而代入，反复进行第三步骤及第四步骤的处理)第三步骤及第四步骤，由此导出参数矩阵的至少一

部分及输入向量的积

[0151] 另外,初始的第一局部矩阵及初始的第二局部矩阵可包含1×1矩阵。

[0152] (反向传播的阶段)

[0153] 在反向传播的阶段中,控制部11算出通过正向传播的阶段所获得的推导任务的执行结果与由跟所输入的训练数据31关联的正解标签32所表示的正解之间的误差,进而算出所算出的误差的梯度。并且,控制部11通过误差反向传播法对所算出的误差的梯度进行反向传播,以调节各参数的值。在此调节时,控制部11调节各参数的值,以使第二局部矩阵的各要素的值与第一局部矩阵及对角矩阵的积一致。即,控制部11以与第二局部矩阵的各要素对应的方式调节第一局部矩阵及对角矩阵的各要素的值。

[0154] 在生成定标关系递归地成立的、训练完毕的推导模型5的情况下,反向传播的阶段的运算可包含以下的第一步骤～第八步骤的运算。

[0155] • 第一步骤:获取相对于参数矩阵的至少一部分及输入向量的积的误差的梯度

[0156] • 第二步骤:基于所获取的误差的梯度的对应的各要素、与在正向传播的阶段的第五步骤的递归性反复的最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的值,导出关于在所述最终层级中使用(即,乘以最终层级的第一局部矩阵)的对角矩阵的各要素的误差的梯度

[0157] • 第三步骤:基于所获取的误差的梯度以及在最终层级中使用的对角矩阵的各要素的值,导出关于在最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度

[0158] • 第四步骤:基于所获取的误差的梯度、在正向传播的阶段的递归性反复的对象层级中构成第一局部矩阵的局部矩阵及输入向量的积的值、以及关于在正向传播的阶段的递归性反复中在所述对象层级的后续层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度,导出关于在所述对象层级中使用的对角矩阵的各要素的误差的梯度

[0159] • 第五步骤:基于所获取的误差的梯度、在对象层级中使用的对角矩阵的各要素的值、以及关于在后续层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度,导出关于在对象层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度

[0160] • 第六步骤:作为关于在后续层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度的初始值,代入在反向传播的阶段的第三步骤中导出的、关于在最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度,且递归地反复进行反向传播的阶段的第四步骤及第五步骤,直至导出关于在正向传播的阶段的递归性反复的初始层级中使用的对角矩阵的各要素的误差的梯度为止,由此导出关于在各层级中使用的对角矩阵的各要素的误差的梯度

[0161] • 第七步骤:基于输入向量以及通过第六步骤而导出的、关于在初始层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度,导出关于构成初始的第一局部矩阵的局部矩阵的要素的误差的梯度

[0162] • 第八步骤:基于关于在各层级中使用的对角矩阵的各要素以及构成初始的第一局部矩阵的局部矩阵的要素分别导出的误差的梯度,调节在各层级中使用的对角矩阵的各要素以及构成初始的第一局部矩阵的局部矩阵的要素各自的值

[0163] 在参数矩阵对应于输出层53的情况下，第一步骤的误差的梯度是根据通过正向传播的阶段而获得的推导任务的执行结果与由跟所输入的训练数据31关联的正解标签32所表示的正解之间的误差所算出。在参数矩阵对应于中间层52或输入层51的情况下，第一步骤的误差的梯度是从对应的层之后的层反向传播的梯度。另外，如上所述，初始的第一局部矩阵可包含1×1矩阵。

[0164] (A)第一具体例(2×2矩阵)

[0165] 首先，使用图8A及图8B来说明在简单的情形(2×2矩阵)中定标关系成立时的正向传播及反向传播的运算过程的一例。图8A及图8B表示定标关系成立的参数矩阵(2×2矩阵)的正向传播及反向传播的阶段中的运算过程的一例。本情形相当于图1的各局部矩阵w_1-w_4为1×1矩阵的情形。

[0166] 在正向传播的阶段中，控制部11计算构成第一局部矩阵的局部矩阵(w_1,w_2)以及输入向量(x_1,x_2)的积。若像以下的式2及式3那样表达各个要素，则通过此计算，可获得Y_1^1以及Y_2^1。

[0167] [数2]

[0168] $$\begin{bmatrix} Z_1^1 \\ Z_2^1 \end{bmatrix} = \begin{bmatrix} a_1 Y_1^1 \\ a_2 Y_2^1 \end{bmatrix} \left(Y_1^1 = w_1 x_1 、\ Y_2^1 = w_2 x_2 \right) \quad \cdots (式2)$$

[0169] [数3]

[0170] $$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} Y_1^1 + Y_2^1 \\ Z_1^1 + Z_2^1 \end{bmatrix} \quad \cdots (式3)$$

[0171] 继而，控制部11计算第一局部矩阵的积的运算结果(Y_1^1,Y_2^1)以及对角矩阵的对应的要素(a_1,a_2)的积，由此导出构成第二局部矩阵的局部矩阵及输入向量的积(Z_1^1,Z_2^1)。图8A例示至此为止的运算过程。此运算处理的结果可导出参数矩阵及输入向量的积的结果(y_1,y_2)。在参数矩阵对应于输入层51或中间层52的情况下，控制部11将所获得的积的结果正向传播至后续的层。另一方面，在参数矩阵对应于输出层53的情况下，控制部11获取积的结果作为推导任务的执行结果。

[0172] 在反向传播的阶段中，控制部11获取相对于参数矩阵及输入向量的积的误差L的梯度($\partial L/\partial y_1$, $\partial L/\partial y_2$)。如上所述，在参数矩阵对应于输出层53的情况下，控制部11对通过正向传播的阶段而获得的推导任务的执行结果与由对应的正解标签32所表示的正解之间的误差进行偏微分，由此可获得误差的梯度($\partial L/\partial y_1$, $\partial L/\partial y_2$)。另一方面，在参数矩阵为中间层52或输入层51的情况下，控制部11可获得从对应的层之后的层反向传播的梯度来作为误差的梯度($\partial L/\partial y_1$, $\partial L/\partial y_2$)。

[0173] [数4]

[0174] $$\frac{\partial L}{\partial a_1} = \frac{\partial L}{\partial Z_1^1} \frac{\partial Z_1^1}{\partial a_1} \quad \cdots (式4)$$

[0175] [数5]

[0176] $$\frac{\partial L}{\partial Z_1^1} = \frac{\partial L}{\partial y_2} \frac{\partial y_2}{\partial Z_1^1} \quad \cdots (式5)$$

[0177] 对角矩阵的要素($\partial L/\partial a_1$)可根据链式法则(chain rule)而通过所述式4及式5来表达。根据所述式3,($\partial y_2/\partial Z_1^1$)为1。因此,($\partial L/\partial Z_1^1$)成为($\partial L/\partial y_2$)。而且,根据所述式2,($\partial Z_1^1/\partial a_1$)为($Y_1^1$)。因而,($\partial L/\partial a_1$)成为($\partial L/\partial y_2$)×($Y_1^1$)。当对角矩阵的另一个要素($\partial L/\partial a_2$)也同样运算时,($\partial L/\partial a_2$)成为($\partial L/\partial y_2$)×($Y_2^1$)。因此,控制部11可基于所获取的误差的梯度的对应的各要素($\partial L/\partial y_2$)、以及构成第一局部矩阵的局部矩阵及输入向量的积的值(Y_1^1,Y_2^1),来导出关于对角矩阵的各要素的误差的梯度($\partial L/\partial a_1$,$\partial L/\partial a_2$)。

[0178] [数6]

[0179] $$\frac{\partial L}{\partial w_1}=\frac{\partial L}{\partial Y_1^1}\frac{\partial Y_1^1}{\partial w_1} \quad \cdots(式6)$$

[0180] [数7]

[0181] $$\frac{\partial L}{\partial Y_1^1}=\frac{\partial L}{\partial y_1}\frac{\partial y_1}{\partial Y_1^1}+\frac{\partial L}{\partial Z_1^1}\frac{\partial Z_1^1}{\partial Y_1^1} \quad \cdots(式7)$$

[0182] 继而,控制部11算出关于构成第一局部矩阵的局部矩阵的各要素的误差的梯度($\partial L/\partial w_1$,$\partial L/\partial w_2$)。第一局部矩阵的误差的梯度中的($\partial L/\partial w_1$)可根据链式法则而通过所述式6及式7来表达。如图8A所示,Y_1^1的变化传递至y_1及Z_1^1。因此,($\partial L/\partial Y_1^1$)通过式7来表达。

[0183] 根据所述式3,($\partial y_1/\partial Y_1^1$)为1。根据所述式2,($\partial Z_1^1/\partial Y_1^1$)为$a_1$。而且,($\partial Y_1^1/\partial w_1$)为$x_1$。因此,($\partial L/\partial Y_1^1$)成为"($\partial L/\partial y_1$)+($\partial L/\partial Z_1^1$)×$a_1$",($\partial L/\partial w_1$)成为"($\partial L/\partial Y_1^1$)×$x_1$"。同样地,($\partial L/\partial Y_2^1$)成为"($\partial L/\partial y_1$)+($\partial L/\partial Z_2^1$)×$a_2$",($\partial L/\partial w_2$)成为"($\partial L/\partial Y_2^1$)×$x_2$"。($\partial L/\partial Z_1^1$)及($\partial L/\partial Z_2^1$)通过先前的运算(例如式5)已计算出。因此,控制部11基于所获取的误差的梯度($\partial L/\partial y_1$,$\partial L/\partial y_2$)以及对角矩阵的各要素的值(a_1,a_2),导出关于构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_1^1$,$\partial L/\partial Y_2^1$)。并且,控制部11基于输入向量($x_1$,$x_2$)以及关于第一局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_1^1$,$\partial L/\partial Y_2^1$),导出关于构成第一局部矩阵的局部矩阵的各要素的误差的梯度($\partial L/\partial w_1$,$\partial L/\partial w_2$)。

[0184] [数8]

[0185] $$\frac{\partial L}{\partial x_1}=\frac{\partial L}{\partial Y_1^1}\frac{\partial Y_1^1}{\partial x_1} \quad \cdots(式8)$$

[0186] 进而,为了将误差反向传播至之前的层,控制部11算出关于输入向量的误差的梯度。关于输入向量的误差的梯度中的($\partial L/\partial x_1$)可根据链式法则而通过所述式8来表达。($\partial L/\partial Y_1^1$)通过先前的运算已计算出。根据所述式2,($\partial Y_1^1/\partial x_1$)为$w_1$。因此,($\partial L/\partial x_1$)成为"($\partial L/\partial Y_1^1$)×$w_1$"。同样地,($\partial L/\partial x_2$)成为"($\partial L/\partial Y_2^1$)×$w_2$"。因此,控制部11基于构成第一局部矩阵的局部矩阵的各要素的值(w_1,w_2)、以及关于第一局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_1^1$,$\partial L/\partial Y_2^1$),算出关于输入向量的误差的梯度($\partial L/\partial x_1$,$\partial L/\partial x_2$)。在参数矩阵对应于中间层52或输出层53的情况下,控制部11将所算出的误差的梯度反向传播至之前的层。图8B例示至此为止的运算过程。此运算处理的结果可获得用于对第一局部矩阵的各要素以及对角矩阵的各要素进行调节的各个误差的梯度。控制部11根据所获得的误差的梯度

来调节各参数(第一局部矩阵的各要素以及对角矩阵的各要素)的值。

[0187] (B)第二具体例(4×4矩阵)

[0188] 接下来,使用图9A及图9B来说明在简单的情形(4×4矩阵)中定标关系递归地成立时的正向传播及反向传播的运算过程的一例。图9A及图9B表示定标关系递归地成立的参数矩阵(4×4矩阵)的正向传播及反向传播的阶段中的运算过程的一例。本情形相当于图5B的最终分割中的各局部矩阵(w_1-w_4等)为1×1矩阵的情形。

[0189] 在正向传播的阶段中,控制部11计算构成初始的第一局部矩阵的局部矩阵(w_1-w_4)以及输入向量(x_1-x_4)的积。此计算相当于所述正向传播阶段的第一步骤的运算。若像以下的式9及式10那样表达各个要素,则通过所述计算,可获得初始的第一局部矩阵的积的结果(Y_1^1,Y_2^1,Y_3^1,Y_4^1)。

[0190] [数9]

[0191] $$\begin{bmatrix} Z_1^1 \\ Z_2^1 \\ Z_3^1 \\ Z_4^1 \end{bmatrix} = \begin{bmatrix} a_1 Y_1^1 \\ a_2 Y_2^1 \\ a_3 Y_3^1 \\ a_4 Y_4^1 \end{bmatrix} \left(Y_1^1 = w_1 x_1 、\ Y_2^1 = w_2 x_2 、\ Y_3^1 = w_3 x_3 、\ Y_4^1 = w_4 x_4 \right) \quad \cdots (式9)$$

[0192] [数10]

[0193] $$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} Y_1^1 + Y_2^1 + Y_3^1 + Y_4^1 \\ Z_1^1 + Z_2^1 + Z_3^1 + Z_4^1 \end{bmatrix} \quad \cdots (式10)$$

[0194] 继而,控制部11计算第一局部矩阵的积的运算结果(Y_1^1-Y_4^1)以及对角矩阵的对应的各要素(a_1-a_4)的积,由此导出构成初始的第二局部矩阵的局部矩阵及输入向量的积(Z_1^1-Z_4^1)。此计算相当于所述正向传播阶段的第二步骤的运算。(w_1,w_2,a_1w_1,a_2w_2)以及(w_3,w_4,$_3w_3$,a_4w_4)分别相当于后续层级的第一局部矩阵(W_1,W_2)。后续层级的第一局部矩阵的各要素可通过以下的式11及式12来表达。

[0195] [数11]

[0196] $$W_1 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} Y_1^1 + Y_2^1 \\ Z_1^1 + Z_2^1 \end{bmatrix} = \begin{bmatrix} Y_1^2 \\ Y_2^2 \end{bmatrix} \quad \cdots (式11)$$

[0197] [数12]

[0198] $$W_2 \begin{bmatrix} x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} Y_3^1 + Y_4^1 \\ Z_3^1 + Z_4^1 \end{bmatrix} = \begin{bmatrix} Y_3^2 \\ Y_4^2 \end{bmatrix} \quad \cdots (式12)$$

[0199] 因此,控制部11根据初始的第一局部矩阵的积的结果(Y_1^1-Y_4^1)以及第二局部矩阵的积的结果(Z_1^1-Z_4^1),导出构成后续层级的第一局部矩阵的局部矩阵及输入向量的积(Y_1^2-Y_4^2)。此计算相当于所述正向传播阶段的第三步骤的运算,所述正向传播阶段的第三步骤是分别代入在第一步骤及第二步骤中分别获得的初始的第一局部矩阵的积的结果以及初始的第二局部矩阵的积的结果来作为在对象层级中获得的第一局部矩阵的积的结果以及第二局部矩阵的积的结果各自的初始值而执行。

[0200] 并且,控制部11计算所获得的后续层级的第一局部矩阵的积的结果(Y_1^2-Y_4^2)以及

对角矩阵的对应的各要素(A_1,A_2)的积,由此导出构成后续层级的第二局部矩阵的局部矩阵及输入向量的积(Z_1^2-Z_4^2)。此计算相当于所述正向传播阶段的第四步骤的运算。第二局部矩阵的积可通过以下的式13来表达。

[0201] [数13]

[0202] $$A_1W_1\begin{bmatrix}x_1\\x_2\end{bmatrix}+A_2W_2\begin{bmatrix}x_3\\x_4\end{bmatrix}=\begin{bmatrix}Z_1^2\\Z_2^2\end{bmatrix}+\begin{bmatrix}Z_3^2\\Z_4^2\end{bmatrix}=\begin{bmatrix}y_3\\y_4\end{bmatrix}\quad\cdots(式13)$$

[0203] 本情形中,定标关系递归地成立的次数为一次,因此通过所述运算处理,达成所述正向传播阶段的第五步骤的运算。即,控制部11可导出参数矩阵及输入向量的积的结果(y_1-y_4)。图9A例示至此为止的运算过程。在参数矩阵对应于输入层51或中间层52的情况下,控制部11将所获得的积的结果正向传播至后续的层。另一方面,在参数矩阵对应于输出层53的情况下,控制部11获取积的结果作为推导任务的执行结果。

[0204] 在反向传播的阶段中,控制部11获取相对于参数矩阵及输入向量的积的误差L的梯度($\partial L/\partial y_1$-$\partial L/\partial y_4$)。获取方法与所述第一具体例同样。此处理相当于所述反向传播阶段的第一步骤的处理。

[0205] 继而,控制部11导出关于在最终层级中使用的对角矩阵(A_1,A_2)的各要素的误差的梯度。若将对角矩阵A_1的对角成分表达为diag(A_{11},A_{12}),将对角矩阵A_2的对角成分表达为diag(A_{21},A_{22}),则与所述第一具体例同样地,根据链式法则,经由最终层级中的第二局部矩阵的误差的梯度($\partial L/\partial Z_1^2$-$\partial L/\partial Z_4^2$),各要素的梯度($\partial L/\partial A_{11}$, $\partial L/\partial A_{12}$, $\partial L/\partial A_{21}$, $\partial L/\partial A_{22}$)满足以下的式14至式17。

[0206] [数14]

[0207] $$\frac{\partial L}{\partial A_{11}}=Y_1^2\frac{\partial L}{\partial y_3}\quad\cdots(式14)$$

[0208] [数15]

[0209] $$\frac{\partial L}{\partial A_{12}}=Y_2^2\frac{\partial L}{\partial y_4}\quad\cdots(式15)$$

[0210] [数16]

[0211] $$\frac{\partial L}{\partial A_{21}}=Y_3^2\frac{\partial L}{\partial y_3}\quad\cdots(式16)$$

[0212] [数17]

[0213] $$\frac{\partial L}{\partial A_{22}}=Y_4^2\frac{\partial L}{\partial y_4}\quad\cdots(式17)$$

[0214] 因此,控制部11基于所获取的误差的梯度的对应的各要素($\partial L/\partial y_3$, $\partial L/\partial y_4$)、以及在正向传播的阶段的第五步骤的递归性反复的最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的值(Y_1^2-Y_4^2),导出关于在所述最终层级中使用的对角矩阵的各要素的误差的梯度($\partial L/\partial A_{11}$, $\partial L/\partial A_{12}$, $\partial L/\partial A_{21}$, $\partial L/\partial A_{22}$)。此运算相当于所述反向传播阶段的第二步骤。第一局部矩阵的积的值(Y_1^2-Y_4^2)已通过正向传播的阶段而计算出。

[0215] 同样地,根据链式法则,经由关于最终层级中的第二局部矩阵的积的误差的梯度

($\partial L/\partial Z_1^2$- $\partial L/\partial Z_4^2$),关于在最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_1^2$-$\partial L/\partial Y_4^2$)满足以下的式18至式21。

[0216] [数18]

[0217] $$\frac{\partial L}{\partial Y_1^2} = A_{11}\frac{\partial L}{\partial y_3} \quad \cdots(式18)$$

[0218] [数19]

[0219] $$\frac{\partial L}{\partial Y_2^2} = A_{12}\frac{\partial L}{\partial y_4} \quad \cdots(式19)$$

[0220] [数20]

[0221] $$\frac{\partial L}{\partial Y_3^2} = A_{21}\frac{\partial L}{\partial y_3} \quad \cdots(式20)$$

[0222] [数21]

[0223] $$\frac{\partial L}{\partial Y_4^2} = A_{22}\frac{\partial L}{\partial y_4} \quad \cdots(式21)$$

[0224] 因此,控制部11基于所获取的误差的梯度的对应的各要素($\partial L/\partial y_3$, $\partial L/\partial y_4$)以及在最终层级中使用的对角矩阵的各要素的值(A_{11},A_{12},A_{21},A_{22}),导出关于在最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_1^2$-$\partial L/\partial Y_4^2$)。此运算相当于所述反向传播阶段的第三步骤。

[0225] 接下来,控制部11导出关于在前一个层级(对象层级)中使用的对角矩阵的各要素的误差的梯度($\partial L/\partial a_1$-$\partial L/\partial a_4$)。根据链式法则,经由关于所述层级中的第二局部矩阵的积的误差的梯度($\partial L/\partial Z_1^1$-$\partial L/\partial Z_4^1$),各要素的误差的梯度($\partial L/\partial a_1$-$\partial L/\partial a_4$)满足以下的式22至式25。

[0226] [数22]

[0227] $$\frac{\partial L}{\partial a_1} = Y_1^1\left(\frac{\partial L}{\partial y_2} + \frac{\partial L}{\partial Y_2^2}\right) \quad \cdots(式22)$$

[0228] [数23]

[0229] $$\frac{\partial L}{\partial a_2} = Y_2^1\left(\frac{\partial L}{\partial y_2} + \frac{\partial L}{\partial Y_2^2}\right) \quad \cdots(式23)$$

[0230] [数24]

[0231] $$\frac{\partial L}{\partial a_3} = Y_3^1\left(\frac{\partial L}{\partial y_2} + \frac{\partial L}{\partial Y_4^2}\right) \quad \cdots(式24)$$

[0232] [数25]

[0233] $$\frac{\partial L}{\partial a_4} = Y_4^1\left(\frac{\partial L}{\partial y_2} + \frac{\partial L}{\partial Y_4^2}\right) \quad \cdots(式25)$$

[0234] 因此,控制部11基于所获取的误差的梯度($\partial L/\partial y_2$)、在对象层级中构成第一局部矩阵的局部矩阵及输入向量的积的值(Y_1^1-Y_4^1)、以及关于在对象层级的后续层级(在此情况

下为最终层级)中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_2^2$, $\partial L/\partial Y_4^2$),导出关于在对象层级中使用的对角矩阵的各要素的误差的梯度($\partial L/\partial a_1$-$\partial L/\partial a_4$)。此运算相当于所述反向传播阶段的第四步骤的运算,所述反向传播阶段的第四步骤是代入在所述第三步骤中导出的、关于在最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度来作为关于在后续层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度的初始值而执行。第一局部矩阵的积的值(Y_1^1-Y_4^1)已通过正向传播的阶段而计算出。

[0235] 接下来,控制部11导出关于在对象层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_1^1$-$\partial L/\partial Y_4^1$)。根据链式法则,经由关于所述层级中的第二局部矩阵的积的误差的梯度($\partial L/\partial Z_1^1$-$\partial L/\partial Z_4^1$),关于对象层级中的第一局部矩阵的积的误差的梯度($\partial L/\partial Y_1^1$-$\partial L/\partial Y_4^1$)满足以下的式26至式29。

[0236] [数26]

[0237] $$\frac{\partial L}{\partial Y_1^1}=\frac{\partial L}{\partial y_1}+a_1\left(\frac{\partial L}{\partial y_2}+\frac{\partial L}{\partial Y_2^2}\right)+\frac{\partial L}{\partial Y_1^2} \quad \cdots(式26)$$

[0238] [数27]

[0239] $$\frac{\partial L}{\partial Y_2^1}=\frac{\partial L}{\partial y_1}+a_2\left(\frac{\partial L}{\partial y_2}+\frac{\partial L}{\partial Y_2^2}\right)+\frac{\partial L}{\partial Y_1^2} \quad \cdots(式27)$$

[0240] [数28]

[0241] $$\frac{\partial L}{\partial Y_3^1}=\frac{\partial L}{\partial y_1}+a_3\left(\frac{\partial L}{\partial y_2}+\frac{\partial L}{\partial Y_4^2}\right)+\frac{\partial L}{\partial Y_3^2} \quad \cdots(式28)$$

[0242] [数29]

[0243] $$\frac{\partial L}{\partial Y_4^1}=\frac{\partial L}{\partial y_1}+a_4\left(\frac{\partial L}{\partial y_2}+\frac{\partial L}{\partial Y_4^2}\right)+\frac{\partial L}{\partial Y_3^2} \quad \cdots(式29)$$

[0244] 因此,控制部11基于所获取的误差的梯度($\partial L/\partial y_1$, $\partial L/\partial y_2$)、在对象层级中使用的对角矩阵的各要素的值(a_1-a_4)、以及关于在后续层级(在此情况下为最终层级)中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_1^2$-$\partial L/\partial Y_4^2$),导出关于在对象层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_1^1$-$\partial L/\partial Y_4^1$)。此运算相当于所述反向传播阶段的第五步骤的运算。本情形中,定标关系递归地成立的次数为一次,因此通过所述运算处理,达成所述反向传播阶段的第六步骤的运算。即,通过至此为止的处理,导出关于在各层级中使用的对角矩阵的各要素的误差的梯度($\partial L/\partial A_{11}$, $\partial L/\partial A_{12}$, $\partial L/\partial A_{21}$, $\partial L/\partial A_{22}$, $\partial L/\partial a_1$-$\partial L/\partial a_4$)、以及关于在初始层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_1^1$-$\partial L/\partial Y_4^1$)。

[0245] 接下来,控制部11导出关于构成初始的第一局部矩阵的局部矩阵的要素的误差的梯度($\partial L/\partial w_1$-$\partial L/\partial w_4$)。根据链式法则,初始的第一局部矩阵的误差的梯度($\partial L/\partial w_1$-$\partial L/\partial w_4$)满足以下的式30至式33。

[0246] [数30]

[0247] $$\frac{\partial L}{\partial w_1} = x_1 \frac{\partial L}{\partial Y_1^1} \quad \cdots(式30)$$

[0248] [数31]

[0249] $$\frac{\partial L}{\partial w_2} = x_2 \frac{\partial L}{\partial Y_2^1} \quad \cdots(式31)$$

[0250] [数32]

[0251] $$\frac{\partial L}{\partial w_3} = x_3 \frac{\partial L}{\partial Y_3^1} \quad \cdots(式32)$$

[0252] [数33]

[0253] $$\frac{\partial L}{\partial w_4} = x_4 \frac{\partial L}{\partial Y_4^1} \quad \cdots(式33)$$

[0254] 因此，控制部11基于输入向量(x_1-x_4)以及通过第六步骤而导出的、关于在初始层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_1^1$-$\partial L/\partial Y_4^1$)，导出关于构成初始的第一局部矩阵的局部矩阵的要素的误差的梯度($\partial L/\partial w_1$-$\partial L/\partial w_4$)。此运算相当于所述反向传播阶段的第七步骤。通过至此为止的运算，可获得在参数矩阵的调节(更新)中使用的误差的梯度，即，关于构成初始的第一局部矩阵的局部矩阵的要素以及在各层级中使用的对角矩阵的各要素的误差的梯度(本情形中，为$\partial L/\partial w_1$-$\partial L/\partial w_4$、$\partial L/\partial A_{11}$、$\partial L/\partial A_{12}$、$\partial L/\partial A_{21}$、$\partial L/\partial A_{22}$、$\partial L/\partial a_1$-$\partial L/\partial a_4$)的信息。控制部11基于关于在各层级中使用的对角矩阵的各要素以及构成初始的第一局部矩阵的局部矩阵的要素分别导出的误差的梯度，朝与推导结果的误差变小的方向调节在各层级中使用的对角矩阵的各要素以及构成初始的第一局部矩阵的局部矩阵的要素各自的值。调节量可根据学习率来适当调整。此运算相当于所述反向传播阶段的第八步骤。

[0255] 另外，在参数矩阵对应于中间层52或输出层53的情况下，控制部11为了将误差反向传播至之前的层，而导出关于输入向量的误差的梯度($\partial L/\partial x_1$-$\partial L/\partial x_4$)。根据链式法则，关于输入向量的误差的梯度($\partial L/\partial x_1$-$\partial L/\partial x_4$)满足以下的式34至式37。

[0256] [数34]

[0257] $$\frac{\partial L}{\partial x_1} = w_1 \frac{\partial L}{\partial Y_1^1} \quad \cdots(式34)$$

[0258] [数35]

[0259] $$\frac{\partial L}{\partial x_2} = w_2 \frac{\partial L}{\partial Y_2^1} \quad \cdots(式35)$$

[0260] [数36]

[0261] $$\frac{\partial L}{\partial x_3} = w_3 \frac{\partial L}{\partial Y_3^1} \quad \cdots(式36)$$

[0262] [数37]

[0263] $$\frac{\partial L}{\partial x_4} = w_4 \frac{\partial L}{\partial Y_4^1} \quad \cdots(式37)$$

[0264] 因此，控制部11基于在初始层级中构成第一局部矩阵的局部矩阵的各要素的值(w_1-w_4)、以及关于在初始层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_1^1$- $\partial L/\partial Y_4^1$)，导出关于输入向量的误差的梯度($\partial L/\partial x_1$-$\partial L/\partial x_4$)。图9B例示至此为止的运算过程。控制部11将所导出的误差的梯度反向传播至之前的层。

[0265] (C)第三具体例(任意尺寸的矩阵)

[0266] 接下来，使用图10A及图10B来说明在经一般化的情形(任意尺寸的矩阵)中定标关系递归地成立时的正向传播及反向传播的运算过程的一例。图10A及图10B表示定标关系递归地成立的参数矩阵的正向传播及反向传播的阶段中的运算过程的一例。本情形相当于图5C中所例示的参数矩阵的情形。本情形的运算过程除了要素数经一般化的方面以外，与所述第二具体例同样。

[0267] 在正向传播的阶段中，作为第一步骤的处理，控制部11计算构成初始的第一局部矩阵的局部矩阵(W_j^1)以及输入向量(x_j)的积(式38)。通过此计算，控制部11获得初始的第一局部矩阵的积的结果(Y_j^1)。

[0268] [数38]

[0269] $$Y_j^1 = x_j W_j^1 \left(1 \le j \le J_1\right) \quad \cdots(式38)$$

[0270] 继而，作为第二步骤的处理，控制部11计算第一局部矩阵的积的运算结果(Y_j^1)以及对角矩阵($A_{k,j}^1$)的积，由此导出构成初始的第二局部矩阵的局部矩阵及输入向量的积($Z_{k,j}^1$)(式39的i＝1的情形)。

[0271] [数39]

[0272] $$Z_{k,j}^i = A_{k,j}^i Y_j^i \begin{pmatrix} 1 \le i \le l \\ 2 \le k \le p_i \\ 1 \le j \le J_i \end{pmatrix} \quad \cdots(式39)$$

[0273] 另外，J_i是通过式1而算出。

[0274] 接下来，作为第三步骤的处理，控制部11根据在递归性反复的对象层级中获得的第一局部矩阵的积的结果(Y_j^i)以及第二局部矩阵的积的结果($Z_{k,j}^i$)，导出构成后续层级的第一局部矩阵的局部矩阵及输入向量的积(Y_j^{i+1})。此处，将Y_j^i及$Z_{k,j}^i$的各要素分别设为$Y_{j,m}^i$、$Z_{k,j,m}^i$(式40)。而且，作为第四步骤的处理，控制部11计算所获得的后续层级的第一局部矩阵的积的结果(Y_j^{i+1})以及对角矩阵($A_{k,j}^{i+1}$)的积，由此导出构成后续层级的第二局部矩阵的局部矩阵及输入向量的积($Z_{k,j}^{i+1}$)(式39)。

[0275] [数40]

[0276] $$Y_{j,kN_i+m}^{i+1} = \begin{cases} \sum_{s=1}^{q_i} Y_{q_i(j-1)+s,m}^i & (k=0) \\ \sum_{s=1}^{q_i} Z_{k+1,q_i(j-1)+s,m}^i & (k \ge 1) \end{cases} \begin{pmatrix} 1 \le i \le l \\ 0 \le k \le p_i - 1 \\ 1 \le j \le J_{i+1} \\ 1 \le m \le N_i \\ 1 \le s \le q_i \end{pmatrix} \quad \cdots(式40)$$

[0277] 控制部11分别代入在第一步骤及第二步骤中分别获得的初始的第一局部矩阵的积的结果(Y_j^1)以及初始的第二局部矩阵的积的结果($Z_{k,j}^1$)，以作为在对象层级中获得的第

一局部矩阵的积的结果(Y_j^i)以及第二局部矩阵的积的结果($Z_{k,j}^i$)各自的初始值。并且，控制部11递归地反复进行第三步骤及第四步骤。这一连串处理相当于所述正向传播阶段的第五步骤的处理。本情形中，控制部11将第三步骤及第四步骤的处理重复一次。通过此运算结果，控制部11可导出参数矩阵及输入向量的积的结果(式40的i＝1的情形)。图10A例示至此为止的运算过程。在参数矩阵对应于输入层51或中间层52的情况下，控制部11将所获得的积的结果正向传播至后续的层。另一方面，在参数矩阵对应于输出层53的情况下，控制部11获取积的结果作为推导任务的执行结果。

[0278] 在反向传播的阶段中，作为第一步骤的处理，控制部11获取相对于参数矩阵及输入向量的积的误差L的梯度($\partial L/\partial Y_j^{l+1}$)。获取方法与所述第一具体例同样。继而，作为第二步骤的处理，

[0279] 控制部11基于所获取的误差的梯度($\partial L/\partial Y_j^{l+1}$)的对应的各要素、以及在正向传播的阶段的第五步骤的递归性反复的最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的值(Y_j^l)，导出关于在最终层级中使用的对角矩阵($A_{k,j}^l$)的各要素($A_{k,j,m}^l$)的误差的梯度(式41及式42的i＝1的情形)。

[0280] [数41]

[0281] $$\frac{\partial L}{\partial A_{k,j,m}^i}=Y_{j,m}^i\frac{\partial L}{\partial Z_{k,j,m}^i}\begin{pmatrix}1\le i\le l\\2\le k\le p_i\\1\le j\le J_i\\1\le m\le N_i\end{pmatrix}\quad\cdots(式41)$$

[0282] [数42]

[0283] $$\frac{\partial L}{\partial Z_{k+1,q_i(j-1)+s,m}^i}=\frac{\partial L}{\partial Y_{j,kN_i+m}^{i+1}}\begin{pmatrix}1\le i\le l\\1\le k\le p_i-1\\1\le j\le J_{i+1}\\1\le m\le N_i\\1\le s\le q_i\end{pmatrix}\quad\cdots(式42)$$

[0284] 接下来，作为第三步骤的处理，控制部11基于所获取的误差的梯度($\partial L/\partial Y_j^{l+1}$)以及在最终层级中使用的对角矩阵($A_{k,j}^l$)的各要素($A_{k,j,m}^l$)的值，导出关于在最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_j^l$)的各要素($\partial L/\partial Y_{j,m}^l$)(式42、式43的i＝1的情形)。

[0285] [数43]

[0286] $$\frac{\partial L}{\partial Y_{q_i(j-1)+s,m}^i}=\frac{\partial L}{\partial Y_{j,m}^{i+1}}+\sum_{k=2}^{p_i}A_{k,q_i(j-1)+s,m}^i\frac{\partial L}{\partial Z_{k,q_i(j-1)+s,m}^i}\begin{pmatrix}1\le i\le l\\1\le j\le J_{i+1}\\1\le m\le N_i\\1\le s\le q_i\end{pmatrix}\quad\cdots(式43)$$

[0287] 继而，作为第四步骤的处理，控制部11基于所获取的误差的梯度($\partial L/\partial Y_j^{l+1}$)、在正向传播的阶段的递归性反复的对象层级中构成第一局部矩阵的局部矩阵及输入向量的积

的值(Y_j^i)、以及关于在正向传播的阶段的递归性反复中在所述对象层级的后续层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_j^{i+1}$),导出关于在所述对象层级中使用的对角矩阵的各要素的误差的梯度($\partial L/\partial A_{k,j}^i$)(式41、式42)。

[0288] 作为第五步骤的处理,控制部11基于所获取的误差的梯度($\partial L/\partial Y_j^{i+1}$)、在对象层级中使用的对角矩阵($A_{k,j}^i$)的各要素的值、以及关于在后续层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_j^{i+1}$),导出关于在对象层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_j^i$)(式42、式43)。

[0289] 控制部11代入在反向传播的阶段的第三步骤中导出的、关于在最终层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_j^1$),以作为关于在后续层级中构成第一局部矩阵的局部矩阵及输入向量的积的各要素的误差的梯度($\partial L/\partial Y_j^{i+1}$)的初始值。并且,控制部11递归地反复进行所述反向传播阶段的第四步骤及第五步骤,直至导出关于在正向传播的阶段的递归性反复的初始层级中使用的对角矩阵的各要素的误差的梯度($\partial L/\partial A_{k,j}^1$)为止,由此导出关于在各层级中使用的对角矩阵的各要素的误差的梯度($\partial L/\partial A_{k,j}^i$,i为1至1的自然数)。这一连串处理相当于所述反向传播阶段的第六步骤的处理。通过至此为止的处理,导出关于在各层级中使用的对角矩阵的各要素的误差的梯度($\partial L/\partial A_{k,j}^i$)、以及关于在初始层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_j^1$)。

[0290] 接下来,作为第七步骤的处理,控制部11基于输入向量(x_j)以及通过第六步骤而导出的、关于在初始层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_j^1$),导出关于构成初始的第一局部矩阵的局部矩阵的要素的误差的梯度($\partial L/\partial W_j^1$)。

[0291] [数44]

[0292] $$\frac{\partial L}{\partial W_j^1}=x_j\frac{\partial L}{\partial Y_j^1}\left(1\le j\le J_1\right)\qquad\cdots(式44)$$

[0293] 作为第八步骤的处理,控制部11基于关于在各层级中使用的对角矩阵的各要素以及构成初始的第一局部矩阵的局部矩阵的要素分别导出的误差的梯度($\partial L/\partial A_{k,j}^i$, $\partial L/\partial W_j^1$),对在各层级中使用的对角矩阵的各要素以及构成初始的第一局部矩阵的局部矩阵的要素各自的值($A_{k,j}^i$,W_j^1)进行调节。

[0294] 而且,在参数矩阵对应于中间层52或输出层53的情况下,控制部11为了将误差反向传播至之前的层,基于在初始层级中构成第一局部矩阵的局部矩阵的各要素的值(W_j^1)、以及关于在初始层级中构成第一局部矩阵的局部矩阵及输入向量的积的误差的梯度($\partial L/\partial Y_j^1$),导出关于输入向量的误差的梯度($\partial L/\partial x_j$)(式45)。图10B例示至此为止的运算过程。控制部11将所导出的误差的梯度反向传播至之前的层。

[0295] [数45]

[0296] $$\frac{\partial L}{\partial x_j}=W_j^1\frac{\partial L}{\partial Y_j^1}\left(1\le j\le J_1\right)\qquad\cdots(式45)$$

[0297] (小结)

[0298] 通过以上的各运算过程,在各情形中,控制部11可执行正向传播的阶段以及反向传播的阶段的处理,并调节推导模型5的参数的值。控制部11也可反复执行正向传播的阶段以及反向传播的阶段的各处理,直至满足规定的条件(例如执行规定次数、所算出的误差之和成为阈值以下)为止。

[0299] 作为所述机器学习的处理结果,控制部11能够生成获得了执行推导任务的能力且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致的、训练完毕的推导模型5。尤其,根据所述第二具体例及第三具体例,能够生成各参数的值被调节为在参数矩阵的至少一部分内定标关系递归地成立的、训练完毕的推导模型5。当机器学习的处理完成时,控制部11将处理推进至下个步骤S103。

[0300] (步骤S103)

[0301] 返回图7,步骤S103中,控制部11作为保存处理部113而运行,生成与通过机器学习而生成的训练完毕的推导模型5相关的信息来作为学习结果数据125。在学习结果数据125中保持用于再现训练完毕的推导模型5的信息。作为一例,学习结果数据125可构成为包含表示通过所述机器学习的调节而获得的构成第一局部矩阵的参数的值以及对角矩阵的对角成分的各值的信息,与构成第二局部矩阵的参数相关的信息可予以省略。作为另一例,在定标关系递归地成立的情况下,学习结果数据125可构成为包含表示通过所述机器学习的调节而获得的构成初始的第一局部矩阵的参数的值(W_j^l)以及在各层级中使用的对角矩阵的各要素的值($A_{k,j}^i$)的信息,与这些以外的参数相关的信息可予以省略。控制部11将所生成的学习结果数据125保存在规定的存储区域。

[0302] 规定的存储区域例如可为控制部11内的RAM、存储部12、外部存储装置、存储介质或它们的组合。存储介质例如可为CD、DVD等,控制部11也可经由驱动器17来将学习结果数据125保存至存储介质。外部存储装置例如可为网络附属存储器(Network Attached Storage,NAS)等数据服务器。此时,控制部11也可利用通信接口13,经由网络将学习结果数据125保存至数据服务器中。而且,外部存储装置例如也可为经由外部接口14连接于模型生成装置1的外置存储装置。

[0303] 当学习结果数据125的保存完成时,控制部11结束与本动作例相关的模型生成装置1的处理流程。

[0304] 另外,所生成的学习结果数据125可在任意时机被提供给推导装置2。例如,控制部11也可作为步骤S103的处理或者独立于步骤S103的处理,而将学习结果数据125传输至推导装置2。推导装置2也可通过接收所述传输来获取学习结果数据125。而且,例如推导装置2也可利用通信接口23而经由网络来对模型生成装置1或数据服务器进行存取,由此来获取学习结果数据125。而且,例如推导装置2也可经由存储介质92来获取学习结果数据125。而且,例如学习结果数据125也可被预先装入推导装置2中。

[0305] 进而,控制部11也可通过定期或不定期地反复进行所述步骤S101～步骤S103的处理来更新或新生成学习结果数据125。在所述反复时,可适当执行用于机器学习的学习数据集3的至少一部分的变更、修正、追加、删除等。并且,控制部11也可通过将更新或新生成的学习结果数据125以任意方法提供给推导装置2来更新推导装置2所保持的学习结果数据125。

[0306] [推导装置]

[0307] 图11是表示与本实施方式的推导装置2对推导任务的执行相关的处理流程的一例的流程图。以下说明的推导装置2的处理流程为推导方法的一例。但是,以下说明的推导装置2的处理流程不过是一例,各步骤可尽可能地变更。而且,对于以下的处理流程,可根据实施方式来适当地进行步骤的省略、替换以及追加。

[0308] (步骤S201)

[0309] 步骤S201中,控制部21作为数据获取部211而运行,获取对象数据221。对象数据221是成为推测任务执行对象的规定种类的数据的样本。获取对象数据221的方法可根据数据的种类来适当决定。作为一例,可通过利用传感器(例如摄像机)来观测推导对象而生成对象数据221。而且,获取对象数据221的途径可无特别限定,可根据实施方式来适当选择。作为一例,控制部21也可从传感器等装置直接获取对象数据221。另一例中,控制部21也可经由其他计算机、存储介质92等来间接地获取对象数据221。当获取对象数据221时,控制部21将处理推进至接下来的步骤S202。

[0310] (步骤S202)

[0311] 步骤S202中,控制部21作为推导部212而运行,参照学习结果数据125来进行训练完毕的推导模型5的设定。并且,控制部21使用训练完毕的推导模型5来对所获取的对象数据221执行推导任务。

[0312] 步骤S202中的推导处理可与所述机器学习的训练处理中的正向传播(即,推导试行)的阶段的运算处理同样。即,控制部21将对象数据221输入至输入层51,执行推导模型5的正向传播的运算处理。在所述正向传播的运算处理时,控制部21计算构成第一局部矩阵的局部矩阵及输入向量的积。在参数矩阵对应于输入层51的情况下,输入向量为对象数据221。在参数矩阵对应于中间层52或输出层53的情况下,输入向量为对应的层之前的层的计算结果(输出)。接下来,控制部21计算第一局部矩阵的积的结果以及对角矩阵的对应的要素的积,由此导出构成第二局部矩阵的局部矩阵及输入向量的积。即,控制部21将第一局部矩阵的积的结果以及对角矩阵相乘,获取通过相乘所得的计算结果来作为第二局部矩阵及输入向量的积的结果。在定标关系递归地成立的情况下,控制部21执行所述正向传播阶段的第一步骤至第五步骤的运算。这些运算的结果为,控制部21从输出层53获取与对于对象数据221执行推导任务的结果对应的输出值。当推导处理完成时,控制部21将处理推进至接下来的步骤S203。

[0313] (步骤S203)

[0314] 步骤S203中,控制部21作为输出部213而运行,输出与执行推导任务的结果(推导结果)相关的信息。

[0315] 输出目标以及要输出的信息的内容可分别根据实施方式来适当决定。例如,控制部21也可将通过步骤S202而获得的推导结果直接输出至输出装置26。而且,控制部21也可基于所获得的推导结果来执行任意的信息处理。并且,控制部21也可将执行所述信息处理的结果作为与推导结果相关的信息予以输出。执行所述信息处理的结果的输出可包含根据推导结果来对控制对象装置的动作进行控制等。输出目标例如可为输出装置26、以及计算机的输出装置、控制对象装置等。

[0316] 当与执行推导任务的结果相关的信息的输出完成时,控制部21结束与本动作例相

关的推导装置2的处理流程。另外，控制部21也可持续地反复执行步骤S201～步骤S203的一连串信息处理。反复的时机可根据实施方式来适当决定。由此，推导装置2可构成为持续地反复执行所述推导任务。

[0317] [特征]

[0318] 如上所述，根据本实施方式的模型生成装置1，通过步骤S101～步骤S103的处理，能够生成可通过包含第一局部矩阵及第二局部矩阵的矩阵来表达且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致的、训练完毕的推导模型5。由此，在推导装置2的步骤S202的处理中，即便未保持第二局部矩阵的信息，也能够回头利用第一局部矩阵的运算结果来获得第二局部矩阵的运算结果。同样地，通过以满足此种关系的方式对各参数给予初始值(即，设定第一局部矩阵及对角矩阵的各要素的值)，从而在步骤S102的处理中，也无须保持第二局部矩阵的信息，而可回头利用第一局部矩阵的运算结果来获得第二局部矩阵的运算结果。因此，根据本实施方式，能够在学习结果数据125中省略与第二局部矩阵相关的信息，从而实质上降低构成推导模型5的参数的数量。具体而言，能够将参数的数量降低与第二局部矩阵的要素数与对角矩阵的0以外的要素数的差值相应的量。在定标关系递归地成立的情况下，对于参数矩阵的至少一部分的要素数，能够将参数的数量降低至构成初始的第一局部矩阵的局部矩阵的要素以及在各层级中使用的对角矩阵的要素的数量为止。因此，根据本实施方式，能够降低推导模型5的运算处理所耗费的计算量，抑制对计算资源造成的负担。另外，本实施方式中，也可并用在专利文献2及专利文献3中例示的分散处理、在非专利文献1中例示的卷积运算的高速处理等其他高速化方法。

[0319] §4变形例

[0320] 以上，详细说明了本发明的实施方式，但直至前述为止的说明在所有方面不过是本发明的示例。当然可不脱离本发明的范围而进行各种改良或变形。例如，可进行如下所述的变更。另外，以下，关于与所述实施方式同样的构成元件，使用同样的符号，对于与所述实施方式同样的点，则适当省略了说明。以下的变形例可适当组合。

[0321] ＜4.1＞

[0322] 所述实施方式的推导系统100可适用于对规定种类的数据执行任意推导任务的所有场景。推导任务例如可为判定映照在图像数据中的产品是否存在缺陷、识别映照在图像数据中的对象物的种类、以及推导由传感数据所表示的对象物的状态等。训练数据31及对象数据221例如可为图像数据、声音数据、数值数据、文本数据、以及通过各种传感器获得的测量数据等。训练数据31及对象数据221可为通过借助传感器来观测任意对象而生成的传感数据。传感器例如可为摄像机、麦克风、编码器、环境传感器、生命传感器、医疗检查装置、车载传感器、家庭安全传感器等。环境传感器例如可为气压计、温度计、湿度计、声压计、声音传感器、紫外线传感器、照度计、雨量计、气体传感器等。生命传感器例如可为血压计、脉搏计、心率计、心电图记录仪、肌电图描记器、体温计、皮肤电反应计、微波传感器、脑波仪、脑磁仪、活动测量仪、血糖值测量仪、眼电位传感器、眼球运动测量仪等。医疗检查装置例如可为计算机断层扫描(Computed Tomography，CT)装置、磁共振成像(Magnetic Resonance Imaging，MRI)装置等。车载传感器例如可为图像传感器、光探测和测距(Light detection and ranging，Lidar)传感器、毫米波雷达、超声波传感器、加速度传感器等。家庭安全传感器例如可为图像传感器、红外线传感器、活动(声音)传感器、气体(CO_2等)传感器、电流传感

器、智能仪表(smart meter)(测量家电、照明等的电力使用量的传感器)等。以下例示限定了适用场景的变形例。

[0323] (A)外观检查的场景

[0324] 图12示意性地例示第一变形例的检查系统100A的适用场景的一例。本变形例是将所述实施方式适用于利用映照有产品的图像数据来实施产品的外观检查的场景的示例。本变形例的检查系统100A为所述推导系统100的一例,包括模型生成装置1及检查装置2A。与所述实施方式同样地,模型生成装置1及检查装置2A可经由网络相互连接。

[0325] 本变形例中所处理的规定种类的数据(训练数据31A及对象数据221A)包含映照有产品RA的图像数据(图像样本)。图像数据可通过利用摄像机SA来拍摄产品RA而获得。推导任务为:判定映照在图像数据中的产品RA是否存在缺陷。判定是否存在缺陷包含:识别缺陷的有无、推测产品RA包含缺陷的概率、识别产品RA中所包含的缺陷的种类(也可包含表示“无缺陷”的种类)、提取产品RA中所包含的缺陷的范围或者它们的组合。除了这些限定以外,本变形例的检查系统100A可与所述实施方式的推导系统100同样地构成。

[0326] 另外,产品RA例如可为电子机器、电子零件、汽车零件、药品、食品等在制造线上受到搬送的产品。电子零件例如可为底座、贴片电容器、液晶、继电器的绕组等。汽车零件例如可为连杆(con rod)、轴、发动机缸体(engine block)、电动车窗开关、仪表板等。药品例如可为包装好的片剂、未包装的片剂等。产品RA既可为在制造过程完成后所生成的最终品,也可为在制造过程的中途生成的中间品,还可为在经过制造过程之前所准备的初始品。缺陷例如可为划痕、脏污、裂纹、磕痕、毛刺、色差、异物混入等。

[0327] (模型生成装置)

[0328] 本变形例中,模型生成装置1在所述步骤S101中获取多个学习数据集3A。各学习数据集3A包含训练数据31A及正解标签32A的组合。训练数据31A包含映照有产品RA的图像数据。对于训练数据31A的获取,可使用任意种类的摄像机。正解标签32A例如构成为表示缺陷的有无、缺陷的种类、存在缺陷的范围等与映照在训练数据31A中的产品缺陷相关的推导任务的正解。

[0329] 模型生成装置1通过所述步骤S102的处理,使用所获取的多个学习数据集3A来实施推导模型5的机器学习。由此,能够生成获得了判定映照在图像数据中的产品RA是否存在缺陷的能力且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致的、训练完毕的推导模型5。模型生成装置1通过所述步骤S103的处理,将与训练完毕的推导模型5相关的信息作为学习结果数据125而适当保存至规定的存储区域。所生成的学习结果数据125可在任意时机被提供给检查装置2A。

[0330] (检查装置)

[0331] 检查装置2A为所述推导装置2的一例。检查装置2A的硬件结构及软件结构可与所述实施方式的推导装置2同样。本变形例中,检查装置2A可经由通信接口或外部接口而连接于摄像机SA。或者,也可构成为,摄像机SA连接于其他计算机,检查装置2A连接于所述其他计算机,由此,能够从摄像机SA获取图像数据。摄像机SA例如可为一般的RGB摄像机、深度摄像机、红外线摄像机等。摄像机SA可适当配置在能够拍摄产品RA的场所。例如,摄像机SA可配置在搬送产品RA的输送器装置的附近。

[0332] 检查装置2A通过与所述推导装置2同样的处理流程,执行与外观检查相关的一连

串信息处理。即，在步骤S201中，检查装置2A从摄像机SA获取对象数据221A。对象数据221A包含映照有作为检查对象的产品RA的图像数据。在步骤S202中，检查装置2A使用训练完毕的推导模型5来判定映照在对象数据221A(图像数据)中的产品RA是否存在缺陷。接下来，在步骤S203中，检查装置2A输出与判定产品RA是否存在缺陷的结果相关的信息。一例中，检查装置2A可将判定产品RA是否存在缺陷的结果直接输出至输出装置。另一例中，检查装置2A也可在判定为产品RA包含缺陷的情况下，将用于告知此情况的警告输出至输出装置。这些信息的输出目标并不限于检查装置2A的输出装置，也可为其他计算机的输出装置。又一例中，在将检查装置2A连接于搬送产品RA的输送器装置的情况下，检查装置2A也可基于判定结果来控制输送器装置，以在不同的线上搬送有缺陷的产品与无缺陷的产品。

[0333] (特征)

[0334] 根据本变形例，通过实质上降低构成推导模型5的参数的数量，能够降低在外观检查中使用的推导模型5的运算处理所耗费的计算量，抑制对计算资源造成的负担。由此，能够期待外观检查的高速化、利用廉价的计算机实施外观检查等效果。

[0335] (B)图像识别的场景

[0336] 图13示意性地例示第二变形例的识别系统100B的适用场景的一例。本变形例是将所述实施方式适用于对映照在图像数据中的对象物进行识别的场景的示例。本变形例的识别系统100B为所述推导系统100的一例，包括模型生成装置1及识别装置2B。与所述实施方式同样地，模型生成装置1及识别装置2B可经由网络相互连接。

[0337] 本变形例中所处理的规定种类的数据(训练数据31B及对象数据221B)包含映照有对象物RB的图像数据(图像样本)。图像数据可通过利用摄像机SB来拍摄对象物RB而获得。推导任务为：对映照在图像数据中的对象物RB的种类进行识别。识别对象物RB的种类的处理可包含分割(提取映照有对象的范围)。对象物RB既可为人物，也可为任意物体。映照在图像数据中的范围既可为对象物RB的特定部位(例如脸部等)，或者也可为对象物RB整体。在对象物RB为人物的情况下，识别对象例如也可为脸部等身体的一部分。识别人物的种类例如既可为推测个人，或者也可为推测身体部位(脸部、臂、脚、关节等)。对于任意物体也同样。除了这些限定以外，本变形例的识别系统100B可与所述实施方式的推导系统100同样地构成。

[0338] (模型生成装置)

[0339] 本变形例中，模型生成装置1在所述步骤S101中获取多个学习数据集3B。各学习数据集3B包含训练数据31B及正解标签32B的组合。训练数据31B包含映照有对象物RB的图像数据。对于训练数据31B的获取，可使用任意种类的摄像机。正解标签32B构成为表示映照在训练数据31B中的对象物RB的种类(正解)。

[0340] 模型生成装置1通过所述步骤S102的处理，使用所获取的多个学习数据集3B来实施推导模型5的机器学习。由此，能够生成获得了对映照在图像数据中的对象物RB的种类进行识别的能力且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致的、训练完毕的推导模型5。模型生成装置1通过所述步骤S103的处理，将与训练完毕的推导模型5相关的信息作为学习结果数据125而适当保存至规定的存储区域。所生成的学习结果数据125可在任意时机被提供给识别装置2B。

[0341] (识别装置)

[0342] 识别装置2B为所述推导装置2的一例。识别装置2B的硬件结构及软件结构可与所述实施方式的推导装置2同样。本变形例中，识别装置2B可经由通信接口或外部接口而连接于摄像机SB。或者，也可构成为，摄像机SB连接于其他计算机，识别装置2B连接于所述其他计算机，由此可从摄像机SB获取图像数据。摄像机SB例如可为一般的RGB摄像机、深度摄像机、红外线摄像机等。摄像机SB可适当配置在能够拍摄对象物RB的场所。

[0343] 识别装置2B通过与所述推导装置2同样的处理流程，执行与对象物RB的识别相关的一连串信息处理。即，在步骤S201中，识别装置2B从摄像机SB获取对象数据221B。对象数据221B包含映照有作为识别对象的对象物RB的图像数据。在步骤S202中，识别装置2B使用训练完毕的推导模型5，对映照在对象数据221B(图像数据)中的对象物RB的种类进行识别。接下来，在步骤S203中，识别装置2B输出与识别对象物RB的种类的结果相关的信息。一例中，识别装置2B可将识别对象物RB的结果直接输出至输出装置。另一例中，识别装置2B可根据识别对象物RB的结果来执行任意的信息处理。例如，识别装置2B也可在所识别的对象人物(对象物RB)为特定的个人的情况下，执行解除锁定等信息处理。

[0344] (特征)

[0345] 根据本变形例，通过实质上降低构成推导模型5的参数的数量，能够降低在对象物RB的识别中使用的推导模型5的运算处理所耗费的计算量，抑制对计算资源造成的负担。其结果，能够期待对象物RB的识别处理的高速化、利用廉价的计算机实施识别处理等效果。

[0346] (C)状态推导的场景

[0347] 图14示意性地例示第三变形例的推导系统100C的适用场景的一例。本变形例是将所述实施方式适用于对由传感数据所表示的对象物RC的状态进行推导(识别/回归)的场景的示例。本变形例的推导系统100C为所述推导系统100的一例，包括模型生成装置1及推导装置2C。与所述实施方式同样地，模型生成装置1及推导装置2C可经由网络相互连接。

[0348] 本变形例中所处理的规定种类的数据(训练数据31C及对象数据221C)包含通过利用传感器SC来观测对象物RC而生成的传感数据。传感器SC例如可为摄像机、麦克风、编码器、环境传感器、生命传感器、医疗检查装置、车载传感器、家庭安全传感器等。推导任务为：对由传感数据所表示的对象物RC的状态进行推导(识别/回归)。推导可包含预测(即，推导从获得传感数据的时间点至未来的状态)。除了这些限定以外，本变形例的推导系统100C可与所述实施方式的推导系统100同样地构成。

[0349] 另外，传感器SC的种类可根据推导任务来适当选择。作为一例，对象物RC为人物，推导对象物RC的状态可为推导对象人物的状态。此时，传感器SC例如可包含摄像机、麦克风、生命传感器以及医疗检查装置的至少任一者。推导对象人物的状态例如可为推导患发规定疾病的概率、产生身体健康变化的概率等健康状态。或者，对象人物例如可为车辆的驾驶员，推导对象人物的状态可为推导驾驶员的状态(例如瞌睡程度、疲劳度、从容度等)。

[0350] 作为另一例，对象物RC为工业用机械，推导对象物RC的状态可为推导(探测或预测)工业用机械是否存在异常。此时，传感器SC例如可包含麦克风、编码器以及环境传感器的至少任一者。传感数据可包含马达的编码器值、温度、运行声等。

[0351] 作为另一例，对象物RC为存在于车辆外部的物体，推导对象物RC的状态可为推导车辆外部的状况。此时，传感器SC例如可包含摄像机以及车载传感器的至少任一者。推导车辆外部的状况例如可为推导存在于车辆外部的物体的属性、推导拥堵状况、推导事故的风

险等。存在于车辆外部的物体例如可为道路、信号器、障碍物(人、物)等。推导存在于车辆外部的物体的属性例如可包含推导人或车辆的突然出现、急加速、紧急停车、车道变更等事件的发生。

[0352] 作为另一例,对象物RC例如为存在于室外、规定的室内(例如乙烯大棚(vinyl house)内等)的特定场所的物体,推导对象物RC的状态可为推导特定场所的状况。此时,传感器SC例如可包含摄像机、麦克风以及环境传感器的至少任一者。作为具体例,对象物RC可为植物,推导特定场所的状况可为推导植物的种植状况。

[0353] 作为另一例,对象物RC例如为存在于住宅内的物体,推导对象物RC的状态可为推导住宅内的状况。此时,传感器SC例如可包含摄像机、麦克风、环境传感器以及家庭安全传感器的至少任一者。

[0354] (模型生成装置)

[0355] 本变形例中,模型生成装置1在所述步骤S101中,获取多个学习数据集3C。各学习数据集3C包含训练数据31C及正解标签32C的组合。训练数据31C包含通过利用传感器SC来观测对象物RC而生成的传感数据。正解标签32C构成为表示由训练数据31C所表示的对象物RC的状态(正解)。

[0356] 模型生成装置1通过所述步骤S102的处理,使用所获取的多个学习数据集3C来实施推导模型5的机器学习。由此,能够生成获得了对由传感数据所表示的对象物RC的状态进行推导的能力且第二局部矩阵的各要素的值被调节为与第一局部矩阵及对角矩阵的积一致的、训练完毕的推导模型5。模型生成装置1通过所述步骤S103的处理,将与训练完毕的推导模型5相关的信息作为学习结果数据125适当保存至规定的存储区域。所生成的学习结果数据125可在任意时机被提供给推导装置2C。

[0357] (推导装置)

[0358] 推导装置2C为所述推导装置2的一例。推导装置2C的硬件结构及软件结构可与所述实施方式的推导装置2同样。本变形例中,推导装置2C可经由通信接口或外部接口而连接于传感器SC。或者可构成为,传感器SC连接于其他计算机,推导装置2C连接于所述其他计算机,由此可从传感器SC获取传感数据。

[0359] 推导装置2C通过与所述推导装置2同样的处理流程,执行与对象物RC的状态推导相关的一连串信息处理。即,在步骤S201中,推导装置2C从传感器SC获取对象数据221C。对象数据221C包含通过利用传感器SC来观测作为推导对象的对象物RC而生成的传感数据。在步骤S202中,推导装置2C使用训练完毕的推导模型5来推导由对象数据221C(传感数据)所表示的对象物RC的状态。

[0360] 接下来,在步骤S203中,推导装置2C输出与推导对象物RC的状态的结果相关的信息。一例中,推导装置2C可将推导对象物RC的状态的结果直接输出至输出装置。另一例中,推导装置2C可根据推导对象物RC的状态的结果来执行任意的信息处理。作为具体例,在推导对象物RC的状态为推导对象人物的健康状态的情况下,推导装置2C也可在判定为对象人物的健康状态存在异常(例如,规定疾病的患发概率超过阈值)时,输出用于告知此情况的警告。作为另一具体例,在推导对象物RC的状态为推导驾驶员的状态的情况下,推导装置2C也可在驾驶员的瞌睡程度或疲劳度超过阈值时,执行通知敦促驾驶休息的消息、禁止从自动驾驶切换到手动驾驶等输出。作为另一具体例,在推导对象物RC的状态为推导工业用机

械是否存在异常的情况下，也可在判定为工业用机械存在异常或者存在其预兆时，推导装置2C输出用于告知此情况的警告。作为另一具体例，在推导对象物RC的状态为推导车辆外部的状况的情况下，推导装置2C也可根据所推导的车辆外部的状况来决定对车辆的动作指令，通过所决定的动作指令来控制车辆(例如，在探测到人的突然出现的情况下，执行车辆的暂时停止)。

[0361] (特征)

[0362] 根据本变形例，通过实质上降低构成推导模型5的参数的数量，能够降低在对象物RC的状态推导中使用的推导模型5的运算处理所耗费的计算量，抑制对计算资源造成的负担。其结果，能够期待推导对象物RC的状态的处理的高速化、利用廉价的计算机实施推导处理等效果。

[0363] ＜4.2＞

[0364] 所述实施方式中，推导模型5包含全结合型神经网络。但是，构成推导模型5的神经网络的种类也可不限定于此种示例。另一例中，推导模型5可包含卷积神经网络、循环神经网络等。构成推导模型5的神经网络例如也可包含卷积层、池化层、归一化层、丢弃层等其他种类的层。

[0365] 而且，所述实施方式中，构成推导模型5的机器学习模型的种类也可不限于神经网络。只要能够通过矩阵来表达参数，则构成推导模型5的机器学习模型的种类可无特别限定，可根据实施方式来适当选择。作为另一例，推导模型5例如可包含通过主成分分析获得的主成分向量、支持向量机等。机器学习的方法可根据所采用的机器学习模型的种类来适当决定。

[0366] [符号的说明]

[0367] 1:模型生成装置

[0368] 11:控制部

[0369] 12:存储部

[0370] 13:通信接口

[0371] 14:外部接口

[0372] 15:输入装置

[0373] 16:输出装置

[0374] 17:驱动器

[0375] 81:模型生成程序

[0376] 91:存储介质

[0377] 111:数据获取部

[0378] 112:学习处理部

[0379] 113:保存处理部

[0380] 125:学习结果数据

[0381] 2:推导装置

[0382] 21:控制部

[0383] 22:存储部

[0384] 23:通信接口

[0385] 24:外部接口

[0386] 25:输入装置

[0387] 26:输出装置

[0388] 27:驱动器

[0389] 82:推导程序

[0390] 92:存储介质

[0391] 211:数据获取部

[0392] 212:推导部

[0393] 213:输出部

[0394] 221:对象数据

[0395] 3:学习数据集

[0396] 31:训练数据

[0397] 32:正解标签

[0398] 5:推导模型

[0399] 51:输入层

[0400] 52:中间(隐藏)层

[0401] 53:输出层

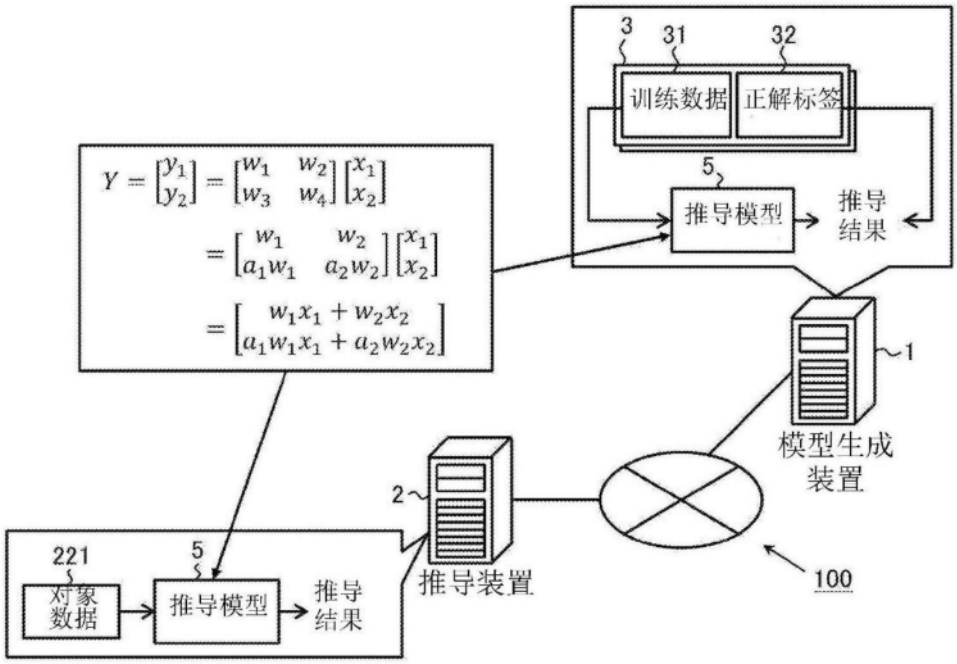

3
31
32
训练数据
正解标签
5
推导模型
推导结果
$Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} w_1 & w_2 \\ w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
$= \begin{bmatrix} w_1 & w_2 \\ a_1 w_1 & a_2 w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
$= \begin{bmatrix} w_1 x_1 + w_2 x_2 \\ a_1 w_1 x_1 + a_2 w_2 x_2 \end{bmatrix}$
1
模型生成装置
2
推导装置
100
221
对象数据
5
推导模型
推导结果

图1

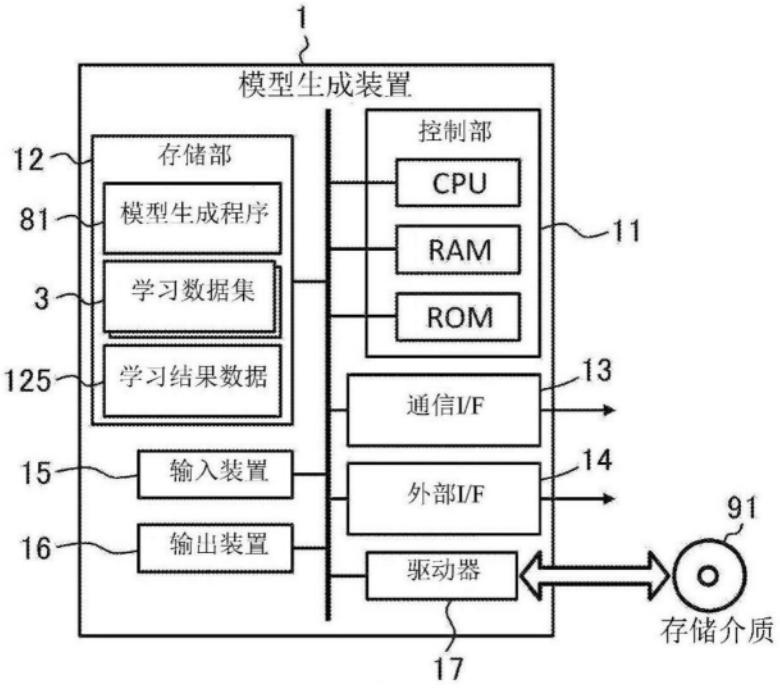

1
模型生成装置
控制部
CPU
RAM
ROM
11
12
存储部
81
模型生成程序
3
学习数据集
125
学习结果数据
13
通信I/F
14
外部I/F
15
输入装置
16
输出装置
驱动器
91
存储介质
17

图2

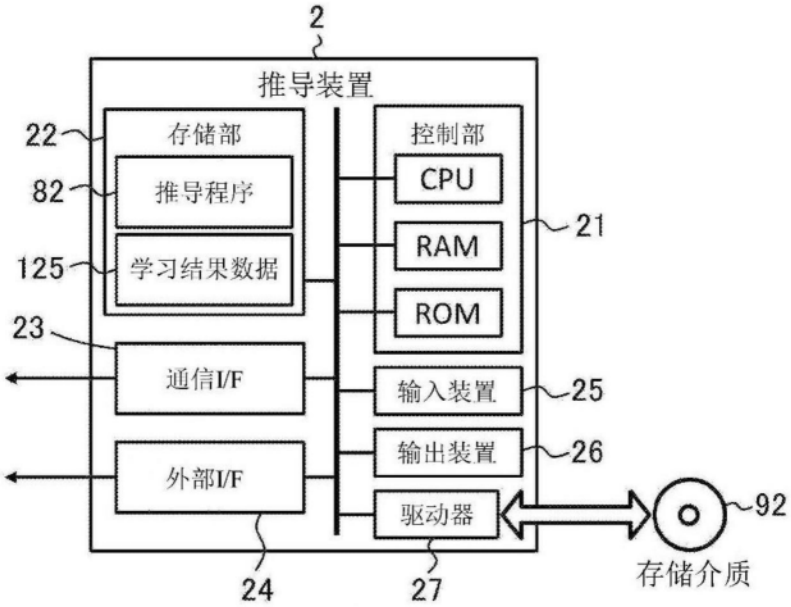

2
推导装置
22
存储部
82
推导程序
125
学习结果数据
23
通信I/F
外部I/F
24
控制部
CPU
RAM
ROM
21
输入装置
25
输出装置
26
驱动器
27
92
存储介质

图3

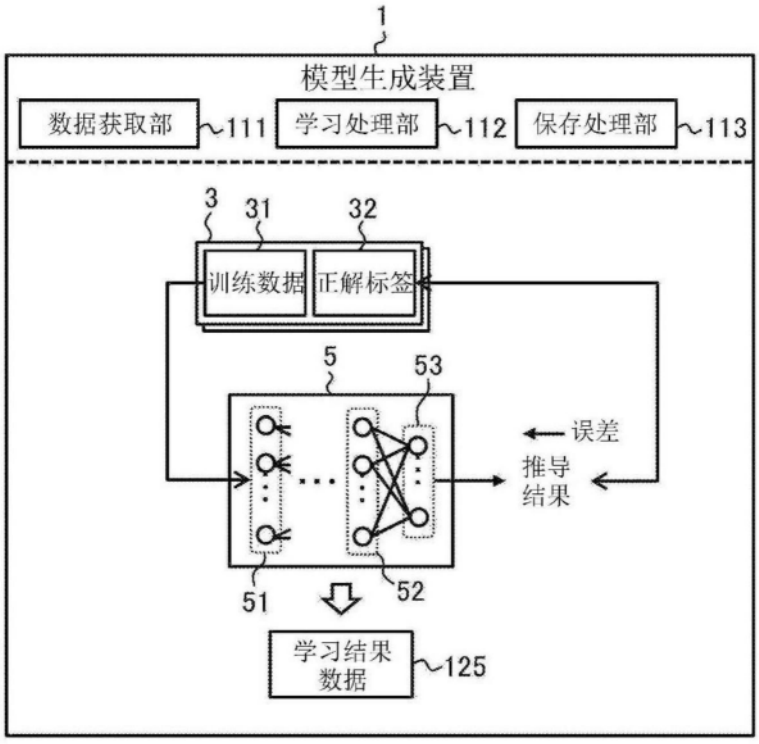

1
模型生成装置
数据获取部
111
学习处理部
112
保存处理部
113
3
31
32
训练数据
正解标签
5
53
误差
推导结果
51
52
学习结果数据
125

图4

$$Y = WX = \begin{bmatrix} w_1 & \cdots & w_N \\ w_{N+1} & \cdots & w_{2N} \\ \vdots & \ddots & \vdots \\ w_{(M-1)N+1} & \cdots & w_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

$$= \begin{bmatrix} w_1 & \cdots & w_N \\ a_{2,1}w_1 & \cdots & a_{2,N}w_N \\ \vdots & \ddots & \vdots \\ a_{M,1}w_1 & \cdots & a_{M,N}w_N \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

图5A

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = WX = \begin{bmatrix} \overbrace{\begin{matrix} w_1 & w_2 \\ a_1w_1 & a_2w_2 \end{matrix}}^{W_1} & \overbrace{\begin{matrix} w_3 & w_4 \\ a_3w_3 & a_4w_4 \end{matrix}}^{W_2} \\ A_1W_1 & A_2W_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

※ A1、A2为2×2的对角矩阵

图5B

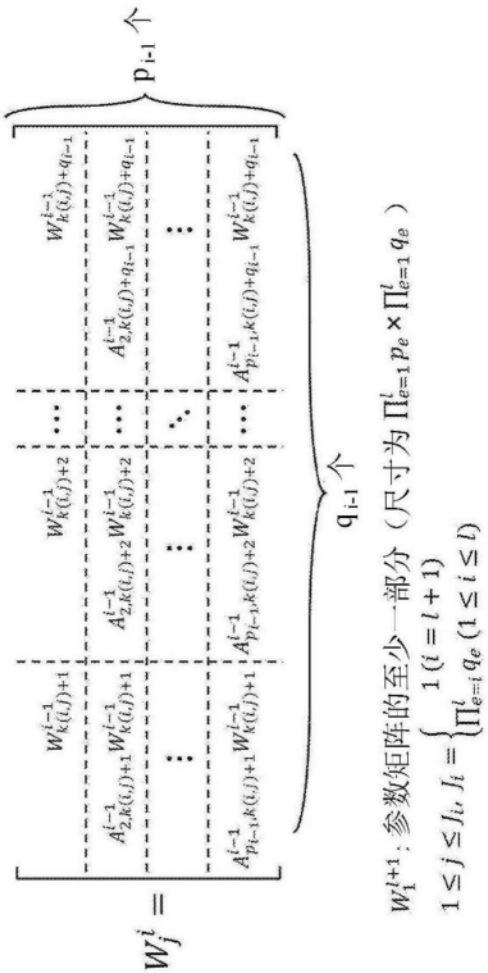

$W_j^i =$
p_{i-1} 个
q_{i-1} 个
W_1^{l+1}：参数矩阵的至少一部分（尺寸为 $\prod_{e=1}^{l} p_e \times \prod_{e=1}^{l} q_e$ ）
$1 \le j \le J_l,\ J_l = \begin{cases} 1\ (l = L+1) \\ \prod_{e=l}^{L} q_e\ (1 \le l \le L) \end{cases}$

图5C

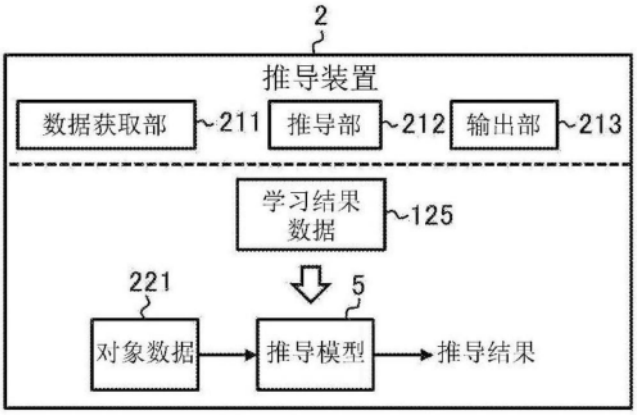

2
推导装置
数据获取部
211
推导部
212
输出部
213
学习结果数据
125
221
5
对象数据
推导模型
推导结果

图6

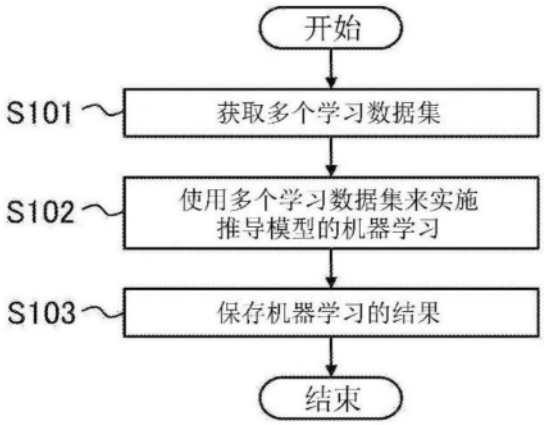

开始
S101
获取多个学习数据集
S102
使用多个学习数据集来实施
推导模型的机器学习
S103
保存机器学习的结果
结束

图7

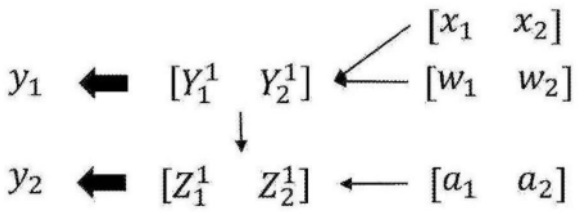

$[x_1 \quad x_2]$
y_1 $[Y_1^1 \quad Y_2^1]$ $[w_1 \quad w_2]$
y_2 $[Z_1^1 \quad Z_2^1]$ $[a_1 \quad a_2]$

图8A

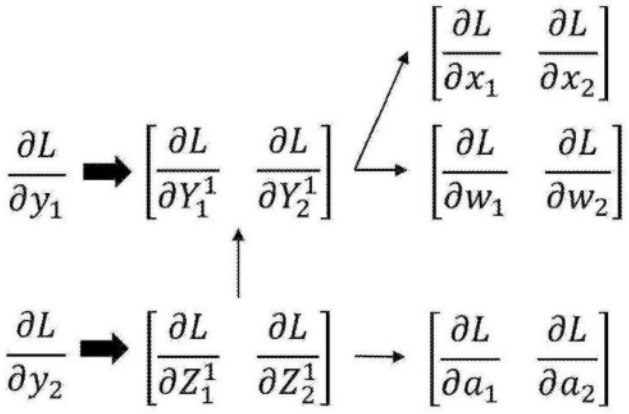

$\left[\frac{\partial L}{\partial x_1} \quad \frac{\partial L}{\partial x_2}\right]$
$\frac{\partial L}{\partial y_1}$ $\left[\frac{\partial L}{\partial Y_1^1} \quad \frac{\partial L}{\partial Y_2^1}\right]$ $\left[\frac{\partial L}{\partial w_1} \quad \frac{\partial L}{\partial w_2}\right]$
$\frac{\partial L}{\partial y_2}$ $\left[\frac{\partial L}{\partial Z_1^1} \quad \frac{\partial L}{\partial Z_2^1}\right]$ $\left[\frac{\partial L}{\partial a_1} \quad \frac{\partial L}{\partial a_2}\right]$

图8B

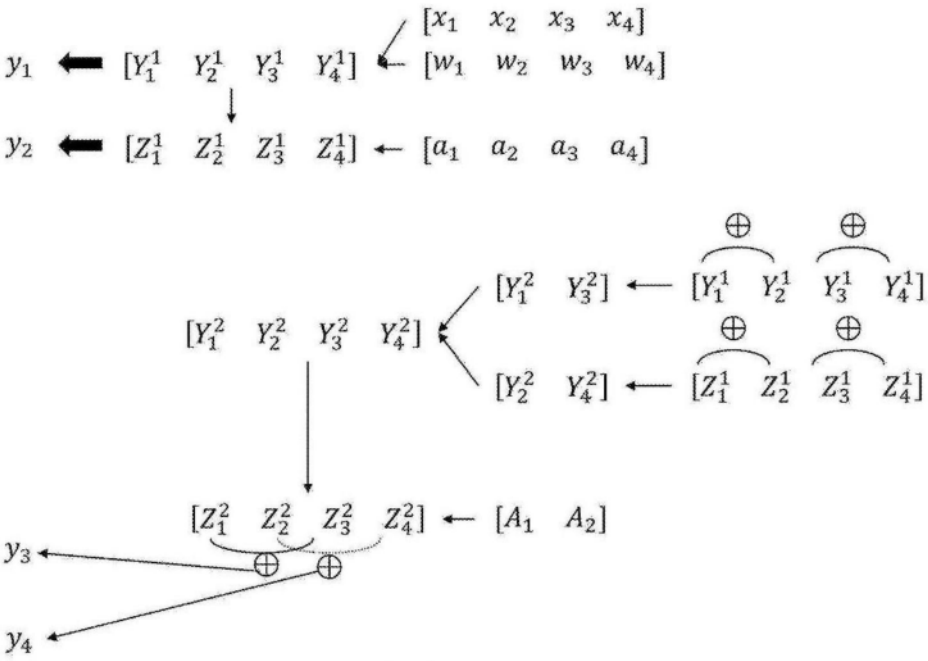

$[x_1 \quad x_2 \quad x_3 \quad x_4]$
y_1 $[Y_1^1 \quad Y_2^1 \quad Y_3^1 \quad Y_4^1]$ $[w_1 \quad w_2 \quad w_3 \quad w_4]$
y_2 $[Z_1^1 \quad Z_2^1 \quad Z_3^1 \quad Z_4^1]$ $[a_1 \quad a_2 \quad a_3 \quad a_4]$
$[Y_1^2 \quad Y_3^2]$ $[Y_1^1 \quad Y_2^1 \quad Y_3^1 \quad Y_4^1]$
$[Y_1^2 \quad Y_2^2 \quad Y_3^2 \quad Y_4^2]$
$[Y_2^2 \quad Y_4^2]$ $[Z_1^1 \quad Z_2^1 \quad Z_3^1 \quad Z_4^1]$
$[Z_1^2 \quad Z_2^2 \quad Z_3^2 \quad Z_4^2]$ $[A_1 \quad A_2]$
y_3
y_4

图9A

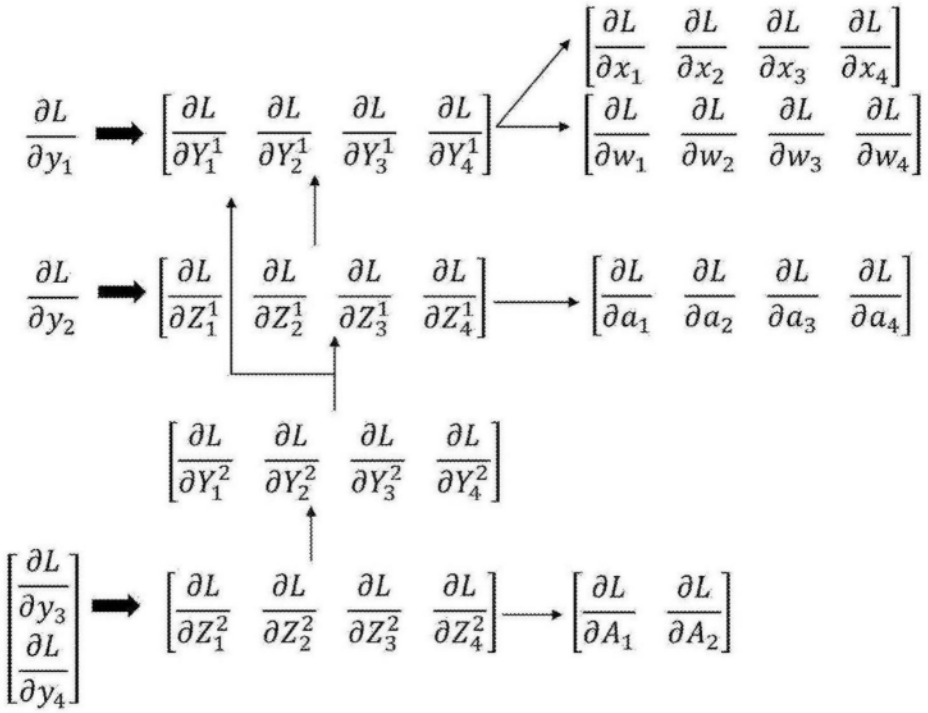

$\frac{\partial L}{\partial y_1}$ $\left[\frac{\partial L}{\partial Y_1^1} \quad \frac{\partial L}{\partial Y_2^1} \quad \frac{\partial L}{\partial Y_3^1} \quad \frac{\partial L}{\partial Y_4^1}\right]$
$\left[\frac{\partial L}{\partial x_1} \quad \frac{\partial L}{\partial x_2} \quad \frac{\partial L}{\partial x_3} \quad \frac{\partial L}{\partial x_4}\right]$
$\left[\frac{\partial L}{\partial w_1} \quad \frac{\partial L}{\partial w_2} \quad \frac{\partial L}{\partial w_3} \quad \frac{\partial L}{\partial w_4}\right]$
$\frac{\partial L}{\partial y_2}$ $\left[\frac{\partial L}{\partial Z_1^1} \quad \frac{\partial L}{\partial Z_2^1} \quad \frac{\partial L}{\partial Z_3^1} \quad \frac{\partial L}{\partial Z_4^1}\right]$ $\left[\frac{\partial L}{\partial a_1} \quad \frac{\partial L}{\partial a_2} \quad \frac{\partial L}{\partial a_3} \quad \frac{\partial L}{\partial a_4}\right]$
$\left[\frac{\partial L}{\partial Y_1^2} \quad \frac{\partial L}{\partial Y_2^2} \quad \frac{\partial L}{\partial Y_3^2} \quad \frac{\partial L}{\partial Y_4^2}\right]$
$\begin{bmatrix}\frac{\partial L}{\partial y_3} \\ \frac{\partial L}{\partial y_4}\end{bmatrix}$ $\left[\frac{\partial L}{\partial Z_1^2} \quad \frac{\partial L}{\partial Z_2^2} \quad \frac{\partial L}{\partial Z_3^2} \quad \frac{\partial L}{\partial Z_4^2}\right]$ $\left[\frac{\partial L}{\partial A_1} \quad \frac{\partial L}{\partial A_2}\right]$

图9B

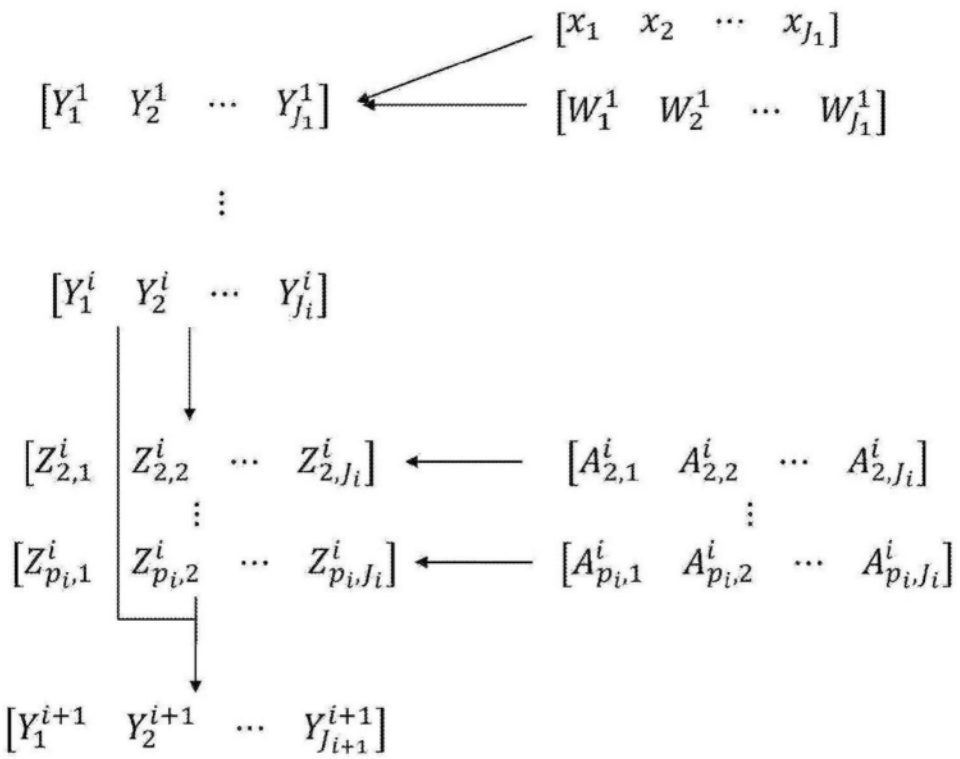

$[x_1 \quad x_2 \quad \cdots \quad x_{J_1}]$
$[Y_1^1 \quad Y_2^1 \quad \cdots \quad Y_{J_1}^1]$
$[W_1^1 \quad W_2^1 \quad \cdots \quad W_{J_1}^1]$
$[Y_1^i \quad Y_2^i \quad \cdots \quad Y_{J_i}^i]$
$[Z_{2,1}^i \quad Z_{2,2}^i \quad \cdots \quad Z_{2,J_i}^i]$
$[A_{2,1}^i \quad A_{2,2}^i \quad \cdots \quad A_{2,J_i}^i]$
$[Z_{p_i,1}^i \quad Z_{p_i,2}^i \quad \cdots \quad Z_{p_i,J_i}^i]$
$[A_{p_i,1}^i \quad A_{p_i,2}^i \quad \cdots \quad A_{p_i,J_i}^i]$
$[Y_1^{i+1} \quad Y_2^{i+1} \quad \cdots \quad Y_{J_{i+1}}^{i+1}]$

图10A

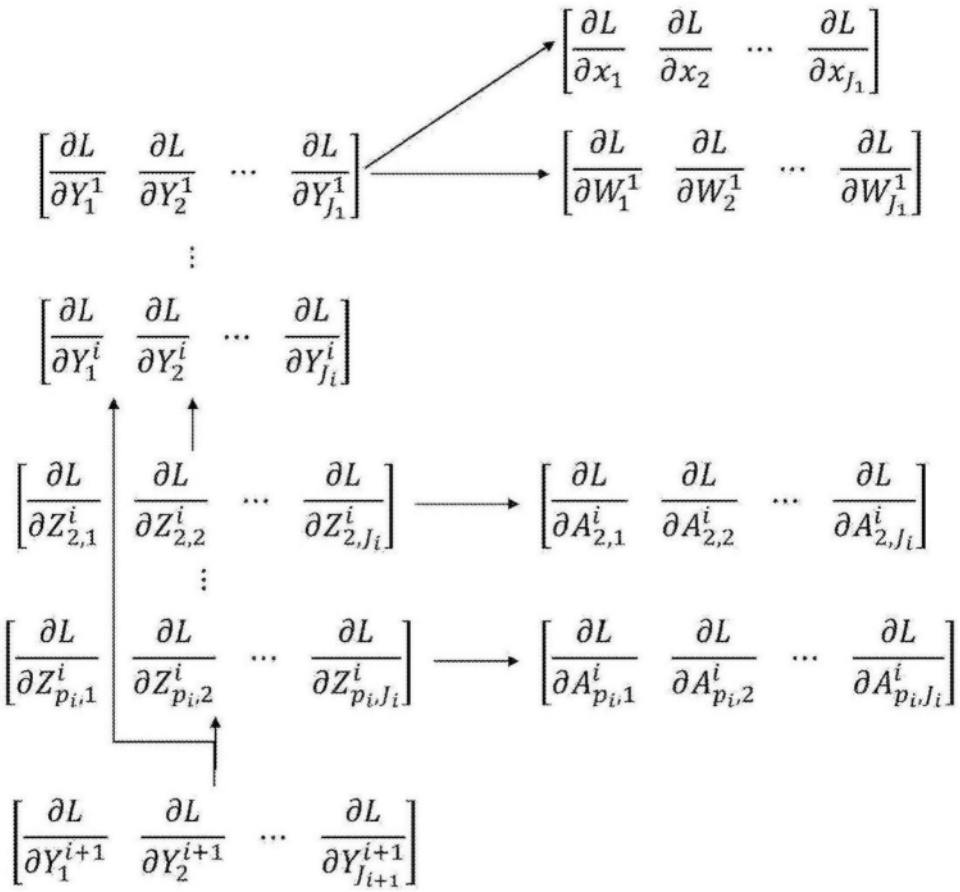

$\left[\frac{\partial L}{\partial Y_1^1} \quad \frac{\partial L}{\partial Y_2^1} \quad \cdots \quad \frac{\partial L}{\partial Y_{J_1}^1}\right]$
$\left[\frac{\partial L}{\partial x_1} \quad \frac{\partial L}{\partial x_2} \quad \cdots \quad \frac{\partial L}{\partial x_{J_1}}\right]$
$\left[\frac{\partial L}{\partial W_1^1} \quad \frac{\partial L}{\partial W_2^1} \quad \cdots \quad \frac{\partial L}{\partial W_{J_1}^1}\right]$
$\left[\frac{\partial L}{\partial Y_1^i} \quad \frac{\partial L}{\partial Y_2^i} \quad \cdots \quad \frac{\partial L}{\partial Y_{J_i}^i}\right]$
$\left[\frac{\partial L}{\partial Z_{2,1}^i} \quad \frac{\partial L}{\partial Z_{2,2}^i} \quad \cdots \quad \frac{\partial L}{\partial Z_{2,J_i}^i}\right]$
$\left[\frac{\partial L}{\partial A_{2,1}^i} \quad \frac{\partial L}{\partial A_{2,2}^i} \quad \cdots \quad \frac{\partial L}{\partial A_{2,J_i}^i}\right]$
$\left[\frac{\partial L}{\partial Z_{p_i,1}^i} \quad \frac{\partial L}{\partial Z_{p_i,2}^i} \quad \cdots \quad \frac{\partial L}{\partial Z_{p_i,J_i}^i}\right]$
$\left[\frac{\partial L}{\partial A_{p_i,1}^i} \quad \frac{\partial L}{\partial A_{p_i,2}^i} \quad \cdots \quad \frac{\partial L}{\partial A_{p_i,J_i}^i}\right]$
$\left[\frac{\partial L}{\partial Y_1^{i+1}} \quad \frac{\partial L}{\partial Y_2^{i+1}} \quad \cdots \quad \frac{\partial L}{\partial Y_{J_{i+1}}^{i+1}}\right]$

图10B

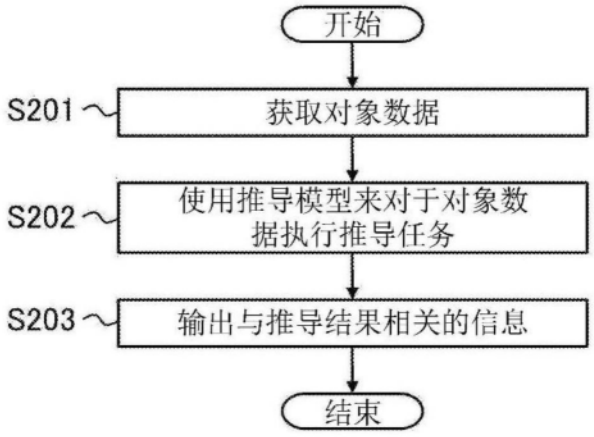

开始
S201
获取对象数据
S202
使用推导模型来对于对象数据执行推导任务
S203
输出与推导结果相关的信息
结束

图11

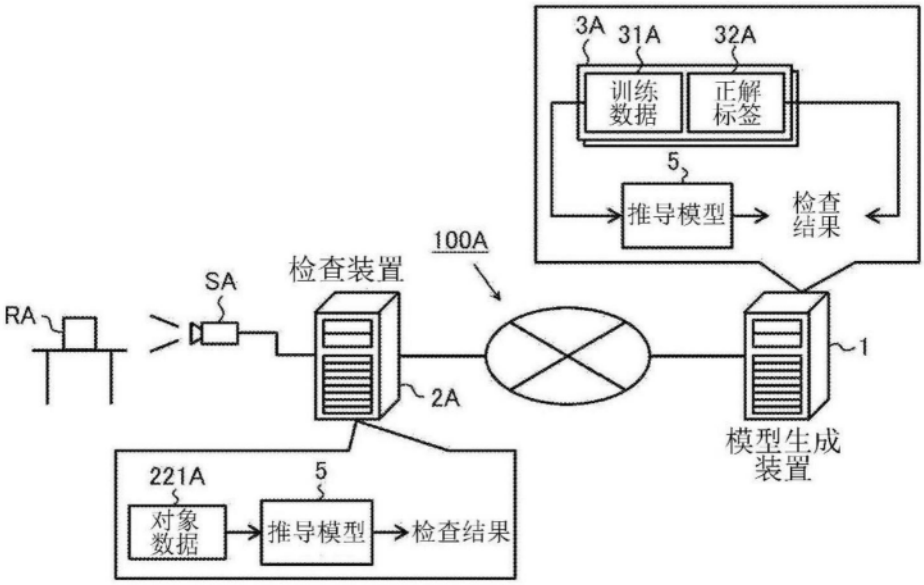

3A
31A
32A
训练数据
正解标签
5
推导模型
检查结果
100A
检查装置
SA
RA
2A
1
模型生成装置
221A
5
对象数据
推导模型
检查结果

图12

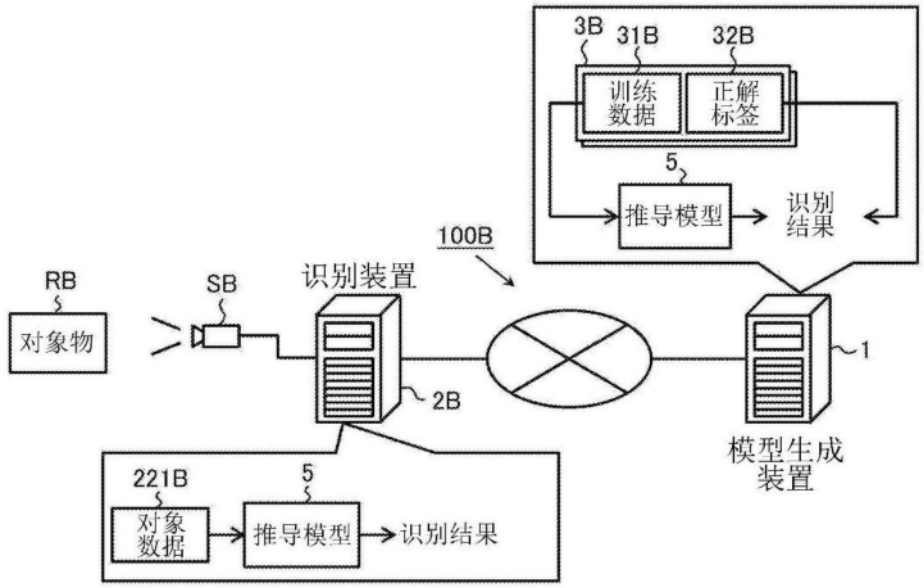

3B
31B
32B
训练数据
正解标签
5
推导模型
识别结果
100B
识别装置
RB
对象物
SB
2B
1
模型生成装置
221B
5
对象数据
推导模型
识别结果

图13

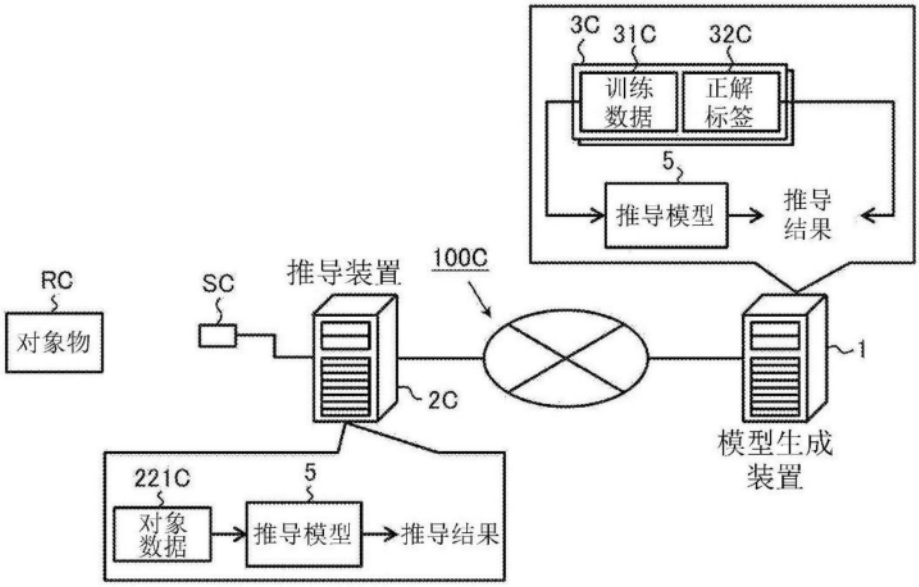

3C
31C
32C
训练数据
正解标签
5
推导模型
推导结果
RC
对象物
SC
推导装置
100C
2C
1
模型生成装置
221C
5
对象数据
推导模型
推导结果

图14