(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114846508 B
(45)授权公告日 2025.06.27

(21)申请号 202080087306.9
(22)申请日 2020.12.16
(65)同一申请的已公布的文献号
申请公布号 CN 114846508 A
(43)申请公布日 2022.08.02
(30)优先权数据
2019-226239 2019.12.16 JP
(85)PCT国际申请进入国家阶段日
2022.06.15
(86)PCT国际申请的申请数据
PCT/JP2020/046887 2020.12.16
(87)PCT国际申请的公布数据
WO2021/125206 JA 2021.06.24
(73)专利权人 富士胶片株式会社
地址 日本国东京都
(72)发明人 滨口侑也
(74)专利代理机构 中科专利商标代理有限责任公司 11021
专利代理师 高颖
(51)Int.Cl.
*G06T 7/00*(2017.01)
*G06V 30/19*(2022.01)
*G06V 30/42*(2022.01)
*G06V 10/82*(2022.01)
*G06N 3/0464*(2023.01)
*G06N 3/084*(2023.01)
*G06N 3/09*(2023.01)
(56)对比文件
Joshua Staker等.Molecular Structure Extraction From Documents Using Deep Learning.Journal of Chemical Information and Modeling.2019,第1-15页.
审查员 王姣

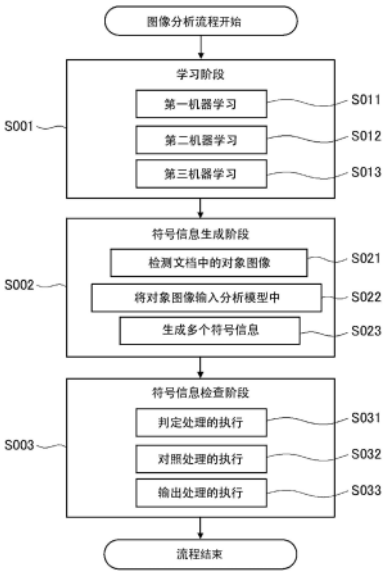
权利要求书3页 说明书12页 附图8页

(54)发明名称

图像分析装置、图像分析方法及计算机程序产品

(57)摘要

本发明提供一种在根据表示化合物的结构式的图像生成结构式的字符信息时,能够对应结构式的写法变化的图像分析装置、图像分析方法、及用于实现图像分析方法的程序。本发明的一实施方式所涉及的图像分析装置具备处理器,处理器通过分析模型,基于表示对象化合物的结构式的对象图像的特征量,生成用线性标记法表示对象化合物的结构式的符号信息。分析模型是通过使用了学习用图像和用线性标记法表示学习用图像所表示的化合物的结构式的符号信息的机器学习构建的模型。




CN 114846508 B

1.一种图像分析装置,其具备处理器,对表示化合物的结构式的图像进行分析,其中,

所述处理器通过分析模型,基于表示对象化合物的结构式的对象图像的特征量,生成用线性标记法表示所述对象化合物的结构式的符号信息,

所述分析模型是通过使用了学习用图像和用线性标记法表示所述学习用图像所表示的化合物的结构式的符号信息的机器学习构建的,

所述分析模型是通过机器学习构建的,所述机器学习使用了所述学习用图像、用所述线性标记法表示所述学习用图像所表示的化合物的结构式的所述符号信息、以及用不同于所述线性标记法的描述方法描述所述学习用图像所表示的化合物的结构式的描述信息,

所述分析模型包括:

特征量输出模型,其通过被输入所述对象图像而输出所述特征量;

描述信息输出模型,其通过被输入所述对象图像而输出所述对象化合物的结构式的所述描述信息;以及

符号信息输出模型,其通过被输入将所输出的所述特征量及所述描述信息合成所得的合成信息,而输出与所述合成信息对应的所述符号信息。

2.根据权利要求1所述的图像分析装置,其中,

所述特征量输出模型输出矢量化的所述特征量,

所述描述信息输出模型输出由矢量化的分子指纹构成的所述描述信息。

3.一种图像分析装置,其具备处理器,对表示化合物的结构式的图像进行分析,其中,

所述处理器通过分析模型,基于表示对象化合物的结构式的对象图像的特征量,生成用线性标记法表示所述对象化合物的结构式的符号信息,

所述分析模型是通过使用了学习用图像和用线性标记法表示所述学习用图像所表示的化合物的结构式的符号信息的机器学习构建的,

所述分析模型包括:

特征量输出模型,通过被输入所述对象图像而输出所述特征量;以及

符号信息输出模型,通过被输入所述特征量而输出与所述特征量对应的所述符号信息,

所述处理器

针对所述符号信息输出模型所输出的所述符号信息中的每一个,执行判定有无标记上的异常的判定处理,

输出所述符号信息输出模型所输出的所述符号信息中无所述异常的正常符号信息,来作为所述对象化合物的结构式的所述符号信息,

所述处理器

通过对照模型,根据所述对象图像生成用不同于所述线性标记法的描述方法描述所述对象化合物的结构式的第一描述信息,

生成用所述描述方法描述所述正常符号信息所表示的结构式的第二描述信息,

将所述第一描述信息和所述第二描述信息进行对照,

根据所述第一描述信息与所述第二描述信息的重合度,输出所述正常符号信息来作为所述对象化合物的结构式的所述符号信息。

4.根据权利要求3所述的图像分析装置,其中,

所述特征量输出模型包括卷积神经网络，

所述符号信息输出模型包括递归神经网络。

5.根据权利要求3所述的图像分析装置，其中，

所述对象化合物的结构式的所述符号信息由多个符号构成，

所述符号信息输出模型从所述符号信息的开头依次确定构成与所述特征量对应的所述符号信息的所述符号，输出所述符号按所确定的顺序排列的所述符号信息。

6.根据权利要求5所述的图像分析装置，其中，

所述处理器通过所述分析模型，基于所述对象图像的所述特征量，对于所述对象化合物的结构式生成多个所述符号信息，

所述符号信息输出模型

对于各个所述符号信息，计算构成所述符号信息的多个所述符号各自的输出概率，且基于计算出的多个所述符号各自的所述输出概率计算所述符号信息的输出分数，

根据计算出的所述输出分数，输出预先确定的个数的所述符号信息。

7.根据权利要求3所述的图像分析装置，其中，

所述对照模型是通过机器学习构建的，所述机器学习使用了第二学习用图像和用所述描述方法描述所述第二学习用图像所表示的化合物的结构式的描述信息。

8.根据权利要求3所述的图像分析装置，其中，

所述对照模型包括：

特征量输出模型，其通过被输入所述对象图像而输出所述特征量；以及

描述信息输出模型，其通过被输入从所述特征量输出模型输出的所述特征量而输出与所述特征量对应的所述第一描述信息。

9.根据权利要求1~8中任一项所述的图像分析装置，其中，

所述处理器

从包含所述对象图像的文档中检测所述对象图像，

通过将检测到的所述对象图像输入所述分析模型中，来生成所述对象化合物的结构式的所述符号信息。

10.根据权利要求9所述的图像分析装置，其中，

所述处理器利用对象检测算法从所述文档中检测所述对象图像。

11.根据权利要求9所述的图像分析装置，其中，

所述处理器

从包含多个所述对象图像的所述文档中检测多个所述对象图像，

通过将检测到的多个所述对象图像按每个所述对象图像输入所述分析模型中，来生成多个所述对象图像各自所表示的所述对象化合物的结构式的所述符号信息。

12.根据权利要求1至8中任一项所述的图像分析装置，其中，

所述线性标记法为简化分子线性输入规范记法或规范简化分子线性输入规范记法。

13.一种图像分析方法，其对表示化合物的结构式的图像进行分析，其中，

处理器实施通过分析模型基于表示对象化合物的结构式的对象图像的特征量来生成用线性标记法表示所述对象化合物的结构式的符号信息的步骤，

所述分析模型是通过机器学习构建的，所述机器学习使用了学习用图像和用线性标记

法表示所述学习用图像所表示的化合物的结构式的符号信息，

所述分析模型是通过机器学习构建的，所述机器学习使用了所述学习用图像、用所述线性标记法表示所述学习用图像所表示的化合物的结构式的所述符号信息、以及用不同于所述线性标记法的描述方法描述所述学习用图像所表示的化合物的结构式的描述信息，

所述分析模型包括：

特征量输出模型，其通过被输入所述对象图像而输出所述特征量；

描述信息输出模型，其通过被输入所述对象图像而输出所述对象化合物的结构式的所述描述信息；以及

符号信息输出模型，其通过被输入将所输出的所述特征量及所述描述信息合成所得的合成信息，而输出与所述合成信息对应的所述符号信息。

14.一种计算机程序产品，包括程序，所述程序用于使处理器实施权利要求13所述的图像分析方法的步骤。

# 图像分析装置、图像分析方法及计算机程序产品

## 技术领域

[0001] 本发明涉及图像分析装置、图像分析方法及程序，特别是涉及对表示化合物的结构式的图像进行分析的图像分析装置、图像分析方法及程序。

## 背景技术

[0002] 化合物的结构式被作为图像数据进行处理的情况很多，例如在因特网上公开或并入文档数据中。但是，作为图像数据来处理的化合物的结构式很难通过普通的检索方法进行检索。

[0003] 另一方面，为了能够检索图像所表示的化合物的结构式，开发了一种利用基于计算机的自动识别技术，从化合物的结构式的图像中识别结构式的技术。作为具体例，可举出专利文献1及2所记载的技术。

[0004] 专利文献1所记载的技术对化学结构图中的字符信息(例如，构成化学物质的原子)进行图案识别，通过规定的算法识别化学结构图的线图信息(例如，原子间的键)。

[0005] 在专利文献2所记载的技术中，读入化合物的结构式的图像，对图像中表示原子符号的区域(像素)分配表示原子符号的属性的值，对表示键符号的区域(像素)分配表示键符号的属性的值。

[0006] 以往技术文献

[0007] 专利文献

[0008] 专利文献1：日本特开2013-61886号公报

[0009] 专利文献2：日本特开2014-182663号公报

## 发明内容

[0010] 发明要解决的技术课题

[0011] 在专利文献1及2所记载的技术中，将表示化合物的结构式的图像中表示结构式中的部分结构(构成要素)的部分与该部分结构的对应关系规则化，按照该规则识别图像中的结构式。

[0012] 但是，在结构式的记载样式中有很多等效的样式，另外，结构式中的键合线的粗细及方向等也可以因写法而改变。在该情况下，为了对应结构式的写法的不同，需要准备许多识别用各种写法记载的部分结构的规则。

[0013] 另外，在专利文献1及2所记载的技术中，例如，由于对于用新的写法记载的结构式的图像没有准备识别规则，因此有可能无法进行识别。

[0014] 本发明是鉴于上述情况而创建的，是解决上述的以往技术的问题点的发明。具体而言，本发明的目的在于，提供一种在根据表示化合物的结构式的图像生成结构式的字符信息时，能够对应结构式的写法的变化的图像分析装置、图像分析方法、及用于实现图像分析方法的程序。

[0015] 用于解决技术课题的手段

[0016] 为了实现上述目的,本发明的图像分析装置具备处理器,对表示化合物的结构式的图像进行分析,其特征在于,处理器通过分析模型,基于表示对象化合物的结构式的对象图像的特征量,生成用线性标记法表示对象化合物的结构式的符号信息,分析模型是通过使用了学习用图像和用线性标记法表示学习用图像所表示的化合物的结构式的符号信息的机器学习构建的。

[0017] 另外,优选的是,处理器从包含对象图像的文档中检测对象图像,通过将检测到的对象图像输入分析模型中,来生成对象化合物的结构式的符号信息。

[0018] 此外,更优选的是,处理器利用对象检测算法从文档中检测对象图像。

[0019] 此外,进一步优选的是,处理器从包含多个对象图像的文档中检测多个对象图像,通过将检测到的多个对象图像按每个对象图像输入分析模型中,来生成多个对象图像各自所表示的对象化合物的结构式的符号信息。

[0020] 另外,也可以是,分析模型包括:特征量输出模型,其通过被输入对象图像而输出特征量;以及符号信息输出模型,其通过被输入特征量而输出与特征量对应的符号信息。

[0021] 此外,也可以是,特征量输出模型包括卷积神经网络,符号信息输出模型包括递归神经网络。

[0022] 另外,优选的是,对象化合物的结构式的符号信息由多个符号构成,符号信息输出模型从符号信息的开头依次确定构成与特征量对应的符号信息的符号,输出符号按所确定的顺序排列的符号信息。

[0023] 此外,也可以是,处理器通过分析模型,基于对象图像的特征量,对于对象化合物的结构式生成多个符号信息。在该情况下,更优选的是,符号信息输出模型对于各个符号信息,计算构成符号信息的多个符号各自的输出概率,且基于计算出的多个符号各自的输出概率计算符号信息的输出分数,根据计算出的输出分数,输出预先确定的个数的符号信息。

[0024] 此外,进一步优选的是,处理器针对符号信息输出模型所输出的符号信息中的每一个,执行判定有无标记上的异常的判定处理,输出符号信息输出模型所输出的符号信息中无异常的正常符号信息,来作为对象化合物的结构式的符号信息。

[0025] 另外,进一步优选的是,处理器通过对照模型,根据对象图像生成用不同于线性标记法的描述方法描述对象化合物的结构式的第一描述信息,生成用描述方法描述正常符号信息所表示的结构式的第二描述信息,将第一描述信息和第二描述信息进行对照,根据第一描述信息与第二描述信息的重合度,输出正常符号信息来作为对象化合物的结构式的符号信息。

[0026] 另外,进一步优选的是,对照模型是通过机器学习构建的,所述机器学习使用了第二学习用图像和用上述描述方法描述第二学习用图像所表示的化合物的结构式的描述信息。

[0027] 此外,进一步优选的是,对照模型包括:特征量输出模型,其通过被输入对象图像而输出特征量;以及描述信息输出模型,其通过被输入从特征量输出模型输出的特征量而输出与特征量对应的第一描述信息。

[0028] 另外,也可以是,分析模型是通过机器学习构建的,所述机器学习使用了学习用图像、用线性标记法表示学习用图像所表示的化合物的结构式的符号信息、以及用不同于线性标记法的描述方法描述学习用图像所表示的化合物的结构式的描述信息。在该情况下,

也可以是,分析模型包括:特征量输出模型,其通过被输入对象图像而输出特征量;描述信息输出模型,其通过被输入对象图像而输出对象化合物的结构式的描述信息;以及符号信息输出模型,其通过被输入将所输出的特征量及描述信息合成所得的合成信息,而输出与合成信息对应的符号信息。

[0029] 此外,优选的是,特征量输出模型输出矢量化的特征量,描述信息输出模型输出由矢量化的分子指纹构成的描述信息。

[0030] 另外,也可以是,线性标记法为简化分子线性输入规范(Simplified Molecular Input Line Entry System)记法或规范简化分子线性输入规范(canonical Simplified Molecular Input Line Entry System)记法。

[0031] 另外,所述目的可以通过图像分析方法来实现,所述图像分析方法对表示化合物的结构式的图像进行分析,其中,处理器实施通过分析模型基于表示对象化合物的结构式的对象图像的特征量来生成用线性标记法表示对象化合物的结构式的符号信息的步骤,分析模型是通过机器学习构建的,所述机器学习使用了学习用图像和用线性标记法表示学习用图像所表示的化合物的结构式的符号信息。

[0032] 另外,也可以实现一种程序,所述程序用于使处理器实施上述图像分析方法的步骤。

[0033] 发明效果

[0034] 根据本发明,能够对应结构式的写法的变化,能够根据表示化合物的结构式的图像适当地生成结构式的字符信息。

## 附图说明

[0035] 图1是关于结构式的线性标记的说明图。

[0036] 图2是分析模型的概念图。

[0037] 图3是表示本发明的一实施方式所涉及的图像分析装置的硬件结构的图。

[0038] 图4是表示图像分析流程的流程的图。

[0039] 图5是关于分子指纹的说明图。

[0040] 图6是从一个文档中检测到多个对象图像的状态的图。

[0041] 图7是对照模型的概念图。

[0042] 图8是变形例所涉及的分析模型的概念图。

## 具体实施方式

[0043] 下面,参照附图对本发明的一实施方式(以下,称为“本实施方式”。)所涉及的图像分析装置、图像分析方法及程序进行说明。

[0044] 此外,以下实施方式只是为了将本发明说明得容易理解而举出的一个例子,不是限定本发明。即,本发明不限于下述的实施方式,在不脱离本发明的主旨的范围内可进行各种改良或变更。另外,当然,其等同物也包含在本发明中。

[0045] 另外,在以后的说明中,除非特别说明,设为“文档”及“图像”是电子化(数据化)的文档及图像,是能够通过计算机进行处理的信息(数据)。

[0046] <本实施方式的图像分析装置的功能>

[0047] 本实施方式的图像分析装置具备处理器,对表示化合物的结构式的图像进行分析。本实施方式的图像分析装置的主要功能是对表示对象化合物的结构式的图像(对象图像)进行分析,并生成对象图像所表示的结构式的符号信息。在此,“对象化合物”是成为生成结构式的符号信息的对象的化合物,例如,在文档中包含的图像中示出有结构式的有机化合物等符合。

[0048] “表示结构式的图像”是表示结构式的线图的图像。在结构式的记载方法中有多种等效的记载方法,例如,可举出省略氢原子(H)的单键的标记省略、骨架的碳原子(C)的标记省略、及官能团的缩略标记等。另外,线图可以根据绘制方法(例如,原子间的键线的粗细、长度、及线延伸的方向等)而改变。另外,在本实施方式中,在结构式的写法中包括表示结构式的图像的分辨率。

[0049] “符号信息”是用线性标记法表示化合物的结构式的信息,将多个符号(例如,ASCII符号)排列而构成。作为线性标记法,可举出SMILES(Simplified Molecular Input Entry System)记法、标准化(canonical)SMILES、SMARTS(Smiles Arbitrary Target Specification)记法、SLN(Sybyl Line Notation)记法、WLN(Wiswesser Line-Formula Notation)记法、ROSDAL(Representation of structure diagram arranged linearly)记法、InChI(International Chemical Identifier)及InChI Key(哈希InChI)等。

[0050] 也可以使用上述线性标记法中的任意一种标记法,但在比较简便且广泛使用这一点上,优选SMILES记法。另外,在考虑到分子内的原子的序列及顺序而将标记唯一确定这一点上,也优选标准化SMILES。此外,在本实施方式中,采用生成按照SMILES记法表示结构式的符号信息的标记法。另外,以下,也将用SMILES记法进行标记称为SMILES标记。

[0051] SMILES记法是将化合物的结构式转换为由多个符号构成的一行符号信息(字符信息)的标记法。SMILES记法所用的符号表示原子的种类(元素)、原子间的键、分支结构、及将环结构切断而形成链状结构时的切断部位等,按照规定的规则而定。

[0052] 此外,作为用SMILES记法标记的化合物的结构式、即符号信息的一个例子,在图1中示出了(S)-溴氯氟甲烷的例子。图1中,左侧表示结构式,右侧表示符号信息(SMILES标记的结构式)。

[0053] 本实施方式的图像分析装置使用表示化合物的结构式的学习用图像和学习用图像所表示的结构式的符号信息(正确的标志信息)作为学习数据集来实施机器学习。通过该机器学习,构建分析模型,上述分析模型基于表示化合物的结构式的图像的特征量,生成该图像所表示的结构式的符号信息。关于分析模型,将在后面的部分详细描述。

[0054] 另外,本实施方式的图像分析装置具有从包含表示化合物的结构式的图像的文档中检测图像(对象图像)的功能。然后,通过将检测到的对象图像输入上述分析模型中,来生成对象图像所表示的结构式的符号信息。

[0055] 通过如上那样的功能,当在论文或专利说明书等文档中包含表示化合物的结构式的图像时,能够检测该图像,并将该图像所表示的化合物的结构式转换为符号信息。

[0056] 另外,转换为符号信息的结构式自此以后可用作检索关键词,因此能够容易地检索包含表示作为目的(目标)的化合物的结构式的图像的文档。

[0057] 此外,本实施方式的图像分析装置具有检查由分析模型生成的符号信息的正误的功能。更详细地进行说明,在本实施方式中,从一个对象图像的特征量中可得到多个符号信

息,对于各个符号信息,判定有无标记上的异常(例如,SMILES记法中的错误标记)。

[0058] 此外,对于未发现异常的符号信息(正常符号信息)中的每一个,实施后述的对照处理。然后,根据对照处理的结果,输出规定数量的正常符号信息作为对象化合物的结构式的符号信息。

[0059] 如上所述,通过检查由分析模型生成的符号信息,作为对象化合物的结构式的符号信息,能够获得准确的信息。

[0060] <关于分析模型>

[0061] 对本实施方式中使用的分析模型(以下,为分析模型M1)进行说明。如图2所示,分析模型M1由特征量输出模型Ma和符号信息输出模型Mb构成。分析模型M1是通过机器学习构建的,所述机器学习将表示化合物的结构式的学习用图像和学习用图像中所示的结构式的符号信息(正确数据)作为学习数据集,且使用多个学习数据集。

[0062] 此外,关于用于机器学习的学习数据集的数量,从提高学习精度的观点来看,多为宜,优选设为5万个以上为佳。

[0063] 在本实施方式中,机器学习是监督学习,其方法是深度学习(即,多层神经网络),但不限于此。关于机器学习的种类(算法),也可以是无监督学习、半监督学习、巩固学习或转导。

[0064] 另外,关于机器学习的技术,也可以是遗传编程、归纳逻辑编程、支持向量机、聚类、贝叶斯网络、极限学习机(ELM)或决策树学习。

[0065] 另外,作为在神经网络的机器学习中将目标函数(损耗函数)最小化的方法,可以使用梯度下降法,或者也可以使用误差逆传播法。

[0066] 特征量输出模型Ma是通过被输入表示对象化合物的结构式的图像(对象图像)而输出对象图像的特征量的模型,例如,由在中间层具有卷积层及池化层的卷积神经网络(CNN)构成。在此,图像的特征量是指卷积神经网络CNN中的学习的特征量,是指在一般的图像识别(图案识别)的过程中确定的特征量。在本实施方式中,特征量输出模型Ma输出矢量化的特征量。

[0067] 此外,在本实施方式中,特征量输出模型Ma也可以使用用于图像分类的网络模型,作为这种模型,可举出例如Oxford visual geometry group的16层CNN(VGG16)、Google公司的Inception模型(GoogLeNet)、Kaiming He氏的152层CNN(Resnet)及Chollet氏的改良Iception模型(Xception)。

[0068] 输入到特征量输出模型Ma的图像的尺寸没有特别限定,但对于化合物的图像,也可以为例如高宽为75×75的尺寸。或者,出于提高模型的输出精度的理由,也可以将化合物的图像的尺寸设为更大的尺寸(例如300×300)。另外,在彩色图像的情况下,出于减轻计算处理的理由,也可以转换成黑白的单色图像,将该单色图像输入特征量输出模型Ma中。

[0069] 另外,在中间层内重复配置卷积层和池化层后,配置全连接层,从该全连接层输出多维矢量化的特征量。此外,从全连接层输出的特征量(多维矢量)在通过了线性层后,被输入符号信息输出模型Mb中。

[0070] 符号信息输出模型Mb是通过被输入从特征量输出模型Ma输出的特征量而输出对象化合物的结构式的符号信息(将结构式进行了SMILES标记的字符信息)的模型。符号信息输出模型Mb例如由作为递归神经网络(RNN)的一种的LSTM(Long Short Term Memory)网络

构成。LSTM将RNN的隐藏层置换为LSTM层。

[0071] 此外,在本实施方式中,如图2所示,可在各LSTM层的前段设置嵌入层(Embedding layer:在图2中标记为Wemb),并针对向各LSTM层的输入赋予固有矢量。另外,对来自各LSTM层的输出,应用softmax函数(在图2中标记为softmax),将来自各LSTM层的输出转换为概率。应用了softmax函数的n个(n为自然数)输出概率的总和为1.0。在本实施方式中,通过softmax函数将来自各LSTM层的输出转换为概率,使用交叉熵误差作为损耗函数求出损耗(学习结果与正确数据之间的差异)。

[0072] 在本实施方式中,符号信息输出模型Mb由LSTM网络构成,但不限于此,符号信息输出模型Mb也可以由GRU(Gated Recurrent Unit)构成。

[0073] 当将对象图像输入分析模型M1中时,如上所述构成的分析模型M1基于对象图像的特征量,对于对象化合物的结构式生成多个符号信息。

[0074] 对符号信息的生成步骤进行说明,当对象图像被输入特征量输出模型Ma中时,特征量输出模型Ma输出对象图像的特征量,该特征量被输入符号信息输出模型Mb中。符号信息输出模型Mb从该符号信息的开头依次确定构成与所输入的特征量对应的符号信息的符号,输出符号按所确定的顺序排列的符号信息。

[0075] 更详细地进行说明,当符号信息输出模型Mb输出由m个(m为2以上的自然数)符号构成的符号信息时,对于第一～第m个符号中的每一个,从对应的LSTM层输出多个候补。根据针对第一～第m个符号中的每一个确定的候补的组合,确定符号信息。例如,在m=3的实例中,在第一个符号的候补为3个、第二个符号的候补为4个、第三个符号的候补为5个的情况下,确定出60种(=3×4×5)符号信息。

[0076] 此外,符号的组合数量(即,符号信息的个数)不限定于将针对第一～第m个符号中的每一个确定的多个候补全部组合时的数量。例如,出于减轻计算处理的负荷的目的,也可以对于针对第一～第m个符号中的每一个确定的多个候补应用束搜索等搜索算法,采用多个候补中前K个(K为自然数)符号。

[0077] 接着,符号信息输出模型Mb对于各个符号信息计算构成符号信息的m个符号各自的输出概率。例如,在对于对象化合物的结构式的符号信息中的第i个(i=1～m)符号输出j个(j为自然数)候补的情况下,通过所述softmax函数计算j个符号各自的输出概率$P_{i1}$、$P_{i2}$、$P_{i3}$…$P_{ij}$。

[0078] 之后,符号信息输出模型Mb基于计算出的各符号的输出概率计算各符号信息的输出分数。在此,输出分数是将构成各符号信息的m个符号各自的输出概率全部相加时的总和。但是,并不限于此,也可以将构成各符号信息的m个符号各自的输出概率相乘时的积作为输出分数。

[0079] 然后,符号信息输出模型Mb根据计算出的输出分数,输出预先确定的个数的符号信息。在本实施方式中,从计算出的输出分数高的符号信息开始依次输出Q个符号信息。在此,关于所输出的符号信息的数量Q,也可以任意确定,但优选为2～20个左右。但是,并不限于此,关于对象化合物的结构式,也可以仅输出一个输出分数最高的符号信息。或者,也可以输出数量相当于将各符号的候补全部组合后的组合数量的符号信息。

[0080] <本实施方式的图像分析装置的结构>

[0081] 接下来,参照图3对本实施方式的图像分析装置(以下为图像分析装置10)的结构

例进行说明。此外,在图3中,将外部接口记载为“外部I/F”。

[0082] 如图3所示,图像分析装置10是处理器11、存储器12、外部接口13、输入装置14、输出装置15及储存器16相互电连接的计算机。此外,在图3所示的结构中,图像分析装置10由一台计算机构成,但图像分析装置10也可以由多台计算机构成。

[0083] 处理器11构成为执行后述的程序21并实施与图像分析相关的一系列处理。此外,处理器11由一个或多个CPU(Central Processing Unit)及后述的程序21构成。

[0084] 构成处理器11的硬件处理器不限于CPU,也可以是FPGA(Field Programmable Gate Array)、DSP(Digital Signal Processor)、ASIC(Application Specific Integrated Circuit)、GPU(Graphics Processing Unit)、MPU(Micro-Processing Unit)或其他IC(Integrated Circuit),或者也可以是将它们组合而成的硬件处理器。另外,处理器11也可以是以SoC(System on Chip)等为代表,发挥图像分析装置10整体的功能的一个IC(Integrated Circuit)芯片。

[0085] 此外,上述的硬件处理器也可以是将半导体元件等电路元件组合成的电路(Circuitry)。

[0086] 存储器12由ROM(Read Only Memory)及RAM(Random Access Memory)等半导体存储器构成,通过临时存储程序及数据而向处理器11提供作业区域,并且也临时存储通过处理器11所执行的处理而生成的各种数据。

[0087] 存储于存储器12的程序中包括图像分析用的程序21。该程序21包括用于实施机器学习并构建分析模型M1的程序、用于从文档中检测对象图像的程序、及用于通过分析模型M1根据对象图像的特征量生成对象化合物的结构式的符号信息的程序。另外,在本实施方式中,程序21还包括用于对于所生成的符号信息执行判定处理及对照处理的程序。

[0088] 此外,程序21可以通过从计算机可读取的记录介质(介质)中读入来获取,或者也可以通过经由因特网或内联网等网络进行接收(下载)来获取。

[0089] 外部接口13是用于与外部装置连接的接口。图像分析装置10经由外部接口13与外部装置、例如扫描仪或因特网上的其他计算机通信。通过这种通信,图像分析装置10能够获取机器学习用的数据的一部分或全部,另外,能够获取登载有对象图像的文档。

[0090] 输入装置14例如由鼠标及键盘等构成,接收用户的输入操作。图像分析装置10例如能够通过用户经由输入装置14输入相当于符号信息的字符信息来获取机器学习用的数据的一部分。

[0091] 输出装置15例如由显示器及扬声器等构成,是用于显示由分析模型M1生成的符号信息或者进行音频播放的装置。

[0092] 储存器16例如由闪存、HDD(Hard Disc Drive)、SSD(Solid State Drive)、FD(Flexible Disc)、MO盘(Magneto-Optical disc)、CD(Compact Disc)、DVD(Digital Versatile Disc)、SD卡(Secure Digital card)及USB存储器(Universal Serial Bus memory)等构成。在储存器16中存储有包括机器学习用的数据的各种数据。此外,储存器16中还存储有以分析模型M1为首的、通过机器学习构建的各种模型的数据。此外,能够将由分析模型M1生成的对象化合物的结构式的符号信息存储于储存器16,并预先登记为数据库。

[0093] 此外,在本实施方式中,储存器16是内置于图像分析装置10的设备,但不限于此,储存器16也可以是与图像分析装置10连接的外置型的设备,或者,也可以是以可经由网络

进行通信的方式连接的外部的计算机(例如,云服务用的服务器计算机)。

[0094] 关于图像分析装置10的硬件结构,不限定于上述结构,可以根据具体实施方式适当地追加、省略及置换构成设备。

[0095] 〈关于图像分析流程〉

[0096] 接下来,对使用了图像分析装置10的图像分析流程进行说明。

[0097] 此外,在以下说明的图像分析流程中,采用本发明的图像分析方法。即,在以下说明中,包括对于本发明的图像分析方法的说明。另外,图像分析流程中的各步骤构成本发明的图像分析方法。

[0098] 如图4所示,本实施方式的图像分析流程按照学习阶段S001、符号信息生成阶段S002及符号信息检查阶段S003的顺序进行。下面,对各阶段进行说明。

[0099] [学习阶段]

[0100] 学习阶段S001是为了在构建以后的阶段中所需的模型而实施机器学习的阶段。在学习阶段S001中,如图4所示,实施第一机器学习S011、第二机器学习S012及第三机器学习S013。

[0101] 第一机器学习S011是用于构建分析模型M1的机器学习,如上所述,使用学习用图像和学习用图像所表示的化合物的结构式的符号信息作为学习用数据集来进行。

[0102] 第二机器学习S012是用于构建在符号信息检查阶段S003中所用的对照模型的机器学习。对照模型是根据对象图像生成用不同于所述线性标记法的描述方法描述对象化合物的结构式的描述信息的模型。

[0103] 作为不同于线性标记法的描述方法,例如,可举出基于分子指纹的描述方法。分子指纹用于辨识具有某种特征的分子,如图5所示将结构式转换为表示结构式中的各种类的部分结构(断片)的有无的二进制多维矢量。在此,部分结构是指表示结构式中的一部分的要素,包括多个原子和原子间的键。

[0104] 构成分子指纹的矢量的维数可任意确定,例如设定为几十～几千维数。在本实施方式中,模仿作为代表性指纹的MACCS Keys,使用用167维矢量表示的分子指纹。

[0105] 此外,不同于线性标记法的描述方法不限于分子指纹,也可以是除此以外的描述方法,例如基于KEGG(Kyoto Encyclopedia of Genes and Genomes) Chemical Function格式(KCF格式)、Molecular Design Limited公司运营的化学结构数据库(MACCS)的输入格式即MOL记法、及MOL的变法即SDF法等的描述方法。

[0106] 第二机器学习S012是使用表示化合物的结构式的学习用图像(第二学习用图像)和第二学习用图像所表示的结构式的描述信息(详细而言,由分子指纹构成的描述信息)作为学习用数据集来进行的。在此,用于第二机器学习S012的第二学习用图像也可以是与第一机器学习S011中使用的学习用图像相同的图像,或者也可以是与第一机器学习S011中使用的学习用图像分开准备的图像。

[0107] 然后,通过使用上述学习用数据进行第二机器学习S012,构建对照模型。关于对照模型,将在后面详细描述。

[0108] 第三机器学习S013是用于构建从登载有表示化合物的结构式的图像的文档中检测该图像的模型(以下,称为图像检测模型)的机器学习。图像检测模型是利用对象检测算法从文档中检测结构式的图像的模型。作为目标检测算法,例如,可以利用R-CNN(Region-

based CNN)、Fast R-CNN、YOLO(You only Look Once)及SDD(Single Shot Multibox Detector)。在本实施方式中,从检测速度的观点来看,构建利用了YOLO的图像检测模型。

[0109] 用于第三机器学习S013的学习用数据(教师数据)通过对表示化合物的结构式的学习用图像应用注释工具来制作。注释工具是对成为对象的数据赋予正确标签(标识)及对象物的坐标等相关信息作为注释的工具。启动注释工具,显示包括学习用图像的文档,由边界框包围表示化合物的结构式的区域,对该区域进行注释,由此制作学习用数据。

[0110] 此外,作为注释工具,例如,可以利用tzutalin公司的labeImg及microsoft公司的VoTT等。

[0111] 然后,通过使用上述的学习用数据进行第三机器学习S013,构建作为YOLO形式的物体检测模型的图像检测模型。

[0112] [符号信息生成阶段]

[0113] 符号信息生成阶段S002是对文档中包含的对象化合物的结构式的图像(对象图像)进行分析,生成对象化合物的结构式的符号信息的阶段。

[0114] 在符号信息生成阶段S002中,首先,图像分析装置10的处理器11对包括对象图像的文档应用所述图像检测模型,检测文档中的对象图像(S021)。即,在本步骤S021中,处理器11利用对象检测算法(具体而言,为YOLO),从文档中检测对象图像。

[0115] 另外,在一个文档中包含多个对象图像的情况下,如图6所示,处理器11从上述文档中检测多个图像(在图6中,用虚线包围的部分的图像)。

[0116] 接下来,处理器11将检测到的对象图像输入分析模型M1中(S022)。在分析模型M1中,在前段的特征量输出模型Ma中输出对象图像的特征量,在后段的符号信息输出模型Mb中,基于输入的对象图像的特征量,输出对象化合物的结构式的符号信息。此时,如上所述,从输出分数高的符号信息开始依次输出预先确定的个数的符号信息。像以上那样,处理器11通过分析模型M1,基于对象图像的特征量,对于对象化合物的结构式生成多个符号信息(S023)。

[0117] 另外,当在步骤S021中检测到多个对象图像时,处理器11针对每个对象图像将检测到的多个对象图像输入分析模型M1中。在该情况下,对于多个对象图像各自所表示的对象化合物的结构式,对每个对象图像生成多个符号信息。

[0118] [符号信息检查阶段]

[0119] 符号信息检查阶段S003是对在符号信息生成阶段S002中对于对象化合物的结构式生成的多个符号信息中的每一个执行判定处理及对照处理的阶段。

[0120] 在符号信息检查阶段S003中,首先,处理器11执行判定处理(S031)。判定处理是对从分析模型M1的符号信息输出模型Mb输出的规定个数的符号信息中的每一个,判定有无SMILES标记上的异常的处理。

[0121] 详细地进行说明,处理器11为了针对符号信息输出模型Mb所输出的各符号信息判定形成各符号信息的字符串是否是正确的SMILES标记的词序,尝试从该字符串向结构式的转换。在此,如果向结构式的转换成功,则判定为该符号信息没有标记上的异常(换言之,该符号信息正常)。在此,以下将无异常的符号信息称为“正常符号信息”。

[0122] 此外,作为从字符串转换为结构式的算法,可以利用与搭载在ChemDraw(注册商标)及RDKit那样的公知的结构式绘图软件中的转换功能同样的算法。

[0123] 执行判定处理后,处理器11对正常符号信息执行对照处理(S032)。对照处理是对由对照模型生成的对象化合物的结构式的第一描述信息和从正常符号信息生成的第二描述信息进行对照的处理。第一描述信息是用分子指纹的描述方式描述对象化合物的结构式的信息。在本实施方式中,第一描述信息是通过将对象图像输入图7所图示的对照模型M2中而生成的。

[0124] 对照模型M2是通过所述的第二机器学习S012构建的,如图7所示,包括特征量输出模型Mc和描述信息输出模型Md。

[0125] 特征量输出模型Mc与分析模型M1的特征量输出模型Ma同样,是通过被输入表示对象化合物的结构式的图像(对象图像)而输出对象图像的特征量的模型,在本实施方式中由CNN构成。另外,在本实施方式中,特征量输出模型Mc与特征量输出模型Ma同样,输出矢量化的特征量。

[0126] 描述信息输出模型Md是通过被输入从特征量输出模型Mc输出的特征量而输出与特征量对应的描述信息(详细而言,由分子指纹构成的描述信息)的模型。在本实施方式中,描述信息输出模型Md例如由神经网络(NN)构成。描述信息输出模型Md输出由矢量化的分子指纹构成的描述信息作为第一描述信息。从描述信息输出模型Md输出的描述信息是对象化合物的结构式的描述信息。

[0127] 此外,作为对照模型M2的特征量输出模型Mc,也可以兼用分析模型M1的特征量输出模型Ma。即,也可以在特征量输出模型Ma、Mc之间将CNN的中间层的权重设定为共同值。在该情况下,第二机器学习S012原样固定由第一机器学习S011确定的CNN的中间层的权重,并确定作为描述信息输出模型Md的NN的中间层的权重,能够减轻模型构建的负荷(计算负荷)。但是,对照模型M2也可以不兼用分析模型M1的CNN(特征量输出模型Ma)而由另外的CNN构成。

[0128] 第二描述信息是用分子指纹的描述方式描述正常符号信息所表示的结构式的描述信息。在本实施方式中,第二描述信息是通过按照转换规则将SMILES标记的符号信息转换成分子指纹而生成的。此时使用的转换规则是通过对于许多化合物确定SMILES标记的结构式与分子指纹的对应关系并将其规则化而规定的。

[0129] 在对照处理中,将如上所述生成的第一描述信息和第二描述信息进行对照,计算两描述信息之间的重合度。在存在多个正常符号信息的情况下,根据每个正常符号信息生成第二描述信息,对于各第二描述信息计算与第一描述信息的重合度。此外,作为重合度的计算方法,可使用计算分子指纹间的相似度的公知的方法,例如,可以利用Tanimoto系数的计算方法。

[0130] 执行对照处理后,处理器11执行输出处理(S033)。输出处理是根据对照处理中计算出的重合度,最终输出(例如,显示)正常符号信息作为对象化合物的结构式的符号信息的处理。在此,根据重合度输出正常符号信息,例如,可以是仅输出重合度超过基准值的正常符号信息,或者也可以是从重合度高的正常符号信息开始依次输出。

[0131] <关于本实施方式的有效性>

[0132] 本实施方式的图像分析装置10可利用通过第一机器学习构建的分析模型M1,基于表示对象化合物的结构式的对象图像的特征量,生成将该结构式进行了SMILES标记的符号信息。其结果,能够适当地对应对象图像中的结构式的写法的变化。

[0133] 对上述的效果进行详细描述,在以往的技术中,将表示化合物的结构式的图像的一部分与该部分中出现的结构式中的部分结构之间的对应关系规则化,并按照该识别规则识别结构式。但是,当结构式的写法发生了改变时,如果未准备可适应该写法的识别规则,则无法识别该结构式。其结果,在上述情况下,难以生成结构式的符号信息。

[0134] 与此相对,在本实施方式中,利用作为机器学习的成果的分析模型M1,根据对象图像的特征量生成符号信息。即,在本实施方式中,即使结构式的写法发生了改变,也能够确定表示该结构式的图像的特征量,如果能够确定特征量,则能够根据该特征量生成符号信息。

[0135] 如上所述,根据本实施方式,即使在对象化合物的结构式的写法发生了改变的情况下,也能够适当地获取符号信息。

[0136] <其他实施方式>

[0137] 综上所述,举出具体例对本发明的图像分析装置、图像分析方法及程序进行了说明,但上述的实施方式只是一个例子,也可以考虑其他实施方式。

[0138] 例如,作为构成图像分析装置的计算机,也可以是用于ASP(Application Service Provider)、SaaS(Software as a Service)、PaaS(Platform as a Service)或IaaS(Infrastructure as a Service)等的服务器。在该情况下,利用上述ASP等服务的用户操作未图示的终端,将包含对象图像的文档发送到服务器。服务器接收到从用户发来的文档时,从文档中检测对象图像,基于对象图像的特征量,生成对象图像所表示的对象化合物的结构式的符号信息。然后,服务器向用户的终端输出(发送)所生成的符号信息。在用户侧,显示从服务器发送来的符号信息或者进行音频播放。

[0139] 另外,在上述实施方式中,设为了对于由分析模型M1生成的符号信息,执行判定有无标记上的异常的判定处理。此外,在上述实施方式中,设为了执行对基于对象图像的特征量生成的分子指纹(第一描述信息)和从正常符号信息转换所得的分子指纹(第二描述信息)进行对照的对照处理。

[0140] 但是,并不限于此,也可以仅执行判定处理及对照处理中的任意一个处理,或者也可以不执行任何处理。

[0141] 另外,在上述实施方式中,设为了用于构建各种模型的机器学习(第一~第三机器学习)通过图像分析装置10来进行,但不限于此。一部分或全部的机器学习也可以通过不同于图像分析装置10的其他装置(计算机)来进行。在该情况下,图像分析装置10获取通过其他装置进行的机器学习构建的模型。

[0142] 例如,在第一机器学习通过其他装置来进行的情况下,图像分析装置10从其他装置获取通过第一机器学习构建的分析模型M1。然后,图像分析装置10通过获取到的分析模型M1,对对象图像进行分析,对于该图像所表示的对象化合物的结构式生成符号信息。

[0143] 另外,在上述实施方式中,上述分析模型M1是通过机器学习构建的,所述机器学习使用了学习用图像和用线性标记表示学习用图像所表示的化合物的结构式的符号信息。而且,分析模型M1基于对象图像的特征量,生成对象图像所表示的对象化合物的结构式的符号信息。

[0144] 但是,并不限于此,作为生成对象化合物的结构式的符号信息的分析模型,可考虑其他模型,例如可举出图8所示的分析模型(以下,为变形例所涉及的分析模型M3)。

[0145] 如图8所示,变形例所涉及的分析模型M3具有特征量输出模型Me、描述信息输出模型Mf、以及符号信息输出模型Mg。变形例的分析模型M3是通过机器学习(以下,为变形例所涉及的机器学习)构建的。变形例所涉及的机器学习使用表示化合物的结构式的学习用图像、学习用图像所表示的化合物的结构式的符号信息(例如,SMILES标记的符号信息)、以及学习用图像所表示的化合物的结构式的描述信息(例如,由分子指纹构成的描述信息)作为学习数据集来进行。

[0146] 特征量输出模型Me与分析模型M1的特征量输出模型Ma同样,通过被输入表示对象化合物的结构式的图像(对象图像)而输出对象图像的特征量,例如由CNN构成。特征量输出模型Me输出矢量化的特征量(例如,2048维矢量)。

[0147] 描述信息输出模型Mf是通过被输入对象图像而输出对象化合物的结构式的描述信息(详细而言,由分子指纹构成的描述信息)的模型。描述信息输出模型Mf是依据所述对照模型M2的模型,例如由CNN构成,输出由矢量化的分子指纹构成的描述信息(例如,167维矢量)。

[0148] 在变形例所涉及的分析模型M3中,如图8所示,对从特征量输出模型Me输出的特征量和从描述信息输出模型Mf输出的描述信息进行合成,生成矢量化的合成信息。合成信息的矢量维数是将特征量的矢量维数和描述信息的矢量维数相加所得的值(即,2215维)。

[0149] 符号信息输出模型Mg是通过被输入上述合成信息而输出与合成信息对应的符号信息(详细而言,SMILES标记的符号信息)的模型。符号信息输出模型Mg与分析模型M1的符号信息输出模型Mb几乎相同,例如由RNN构成,作为其一个例子,可以利用LSTM网络。

[0150] 即使在使用了如上所述构成的变形例所涉及的分析模型M3的情况下,也能够根据对象图像的特征量生成用线性标记表示对象化合物的结构式的符号信息。

[0151] 符号说明

| | | |
|---|---|---|
| [0152] | 10 | 图像分析装置 |
| [0153] | 11 | 处理器 |
| [0154] | 12 | 存储器 |
| [0155] | 13 | 外部接口 |
| [0156] | 14 | 输入装置 |
| [0157] | 15 | 输出装置 |
| [0158] | 16 | 储存器 |
| [0159] | 21 | 程序 |
| [0160] | M1 | 分析模型 |
| [0161] | M2 | 对照模型 |
| [0162] | M3 | 变形例所涉及的分析模型 |
| [0163] | Ma、Mc、Me | 特征量输出模型 |
| [0164] | Mb、Mg | 符号信息输出模型 |
| [0165] | Md、Mf | 描述信息输出模型。 |

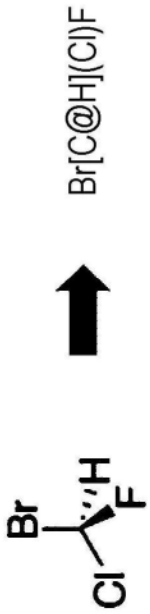



图1

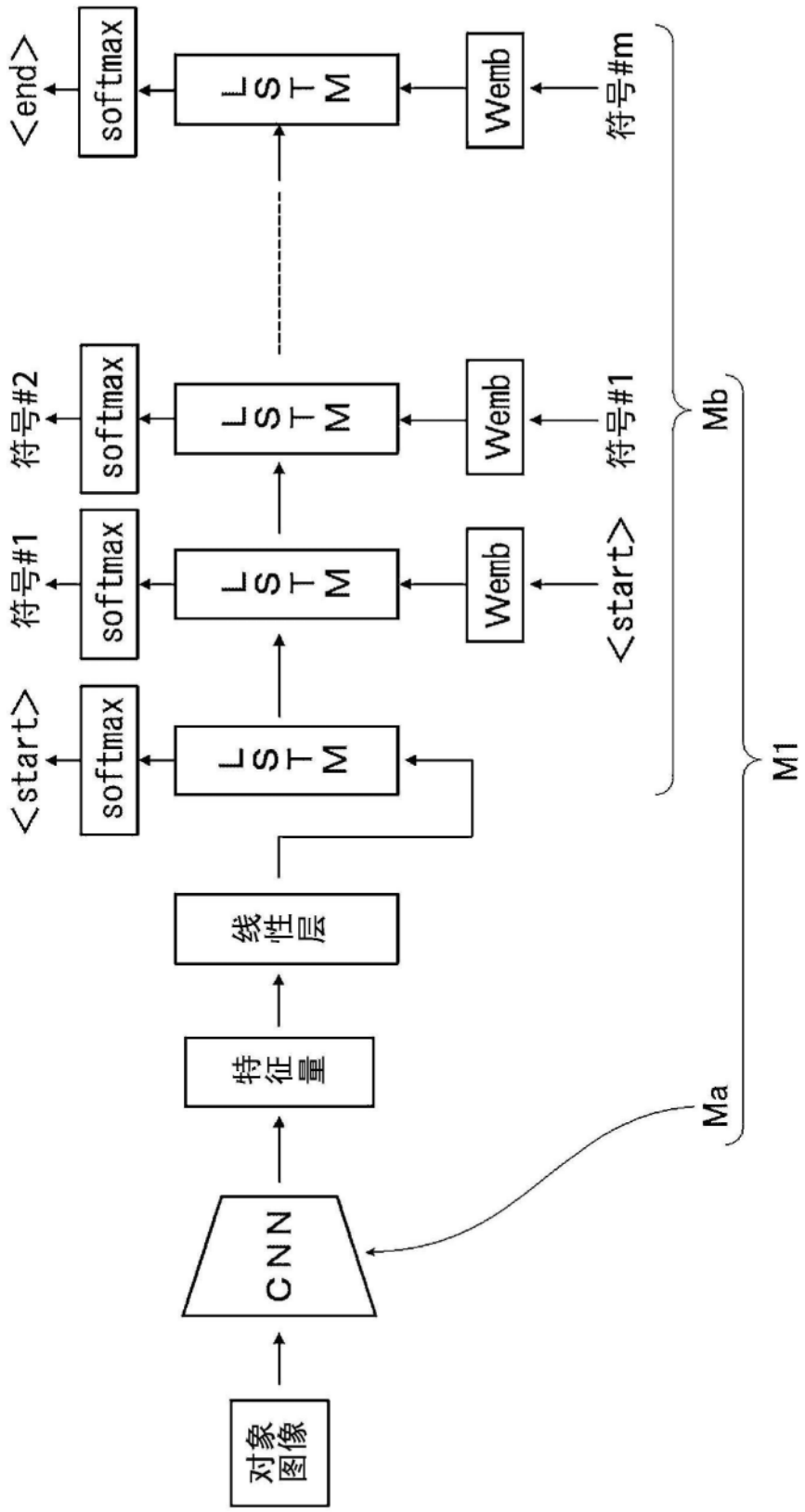

对象图像
CNN
特征量
线性层
LSTM
softmax
Wemb
<start>
符号#1
符号#2
符号#m
<end>
Ma
Mb
M1

图2

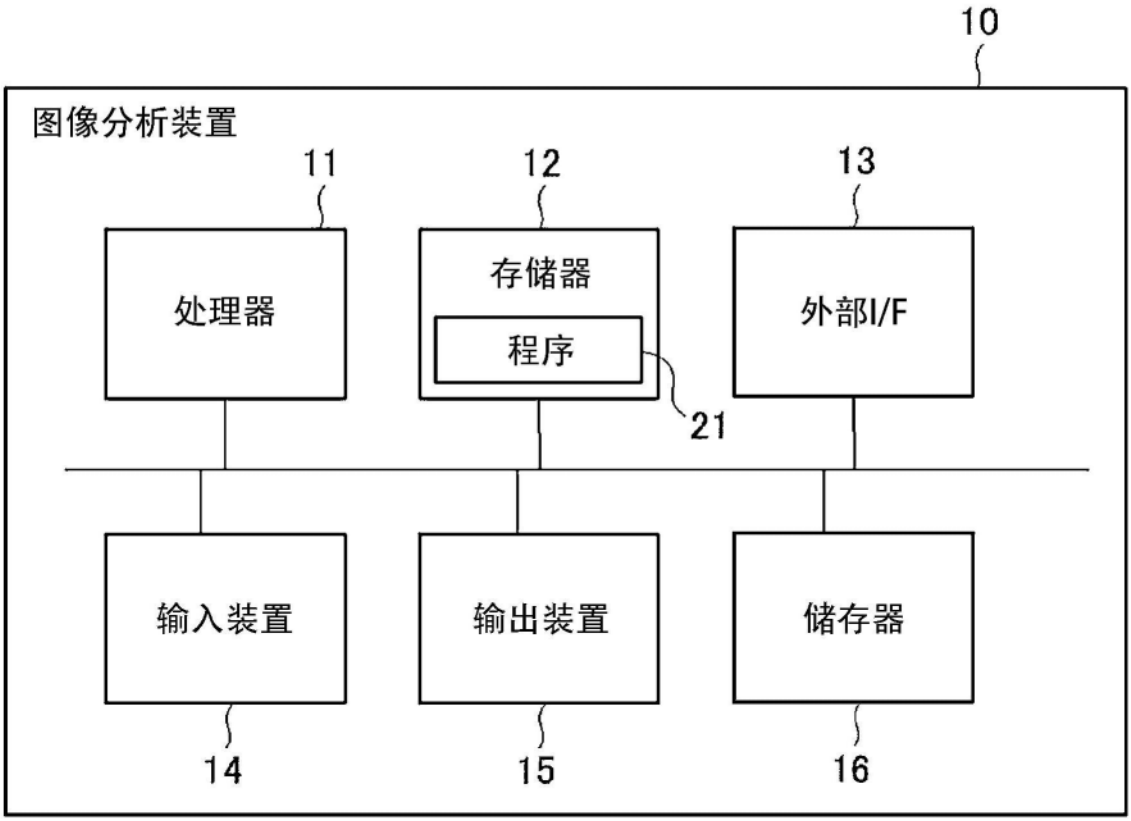



图3

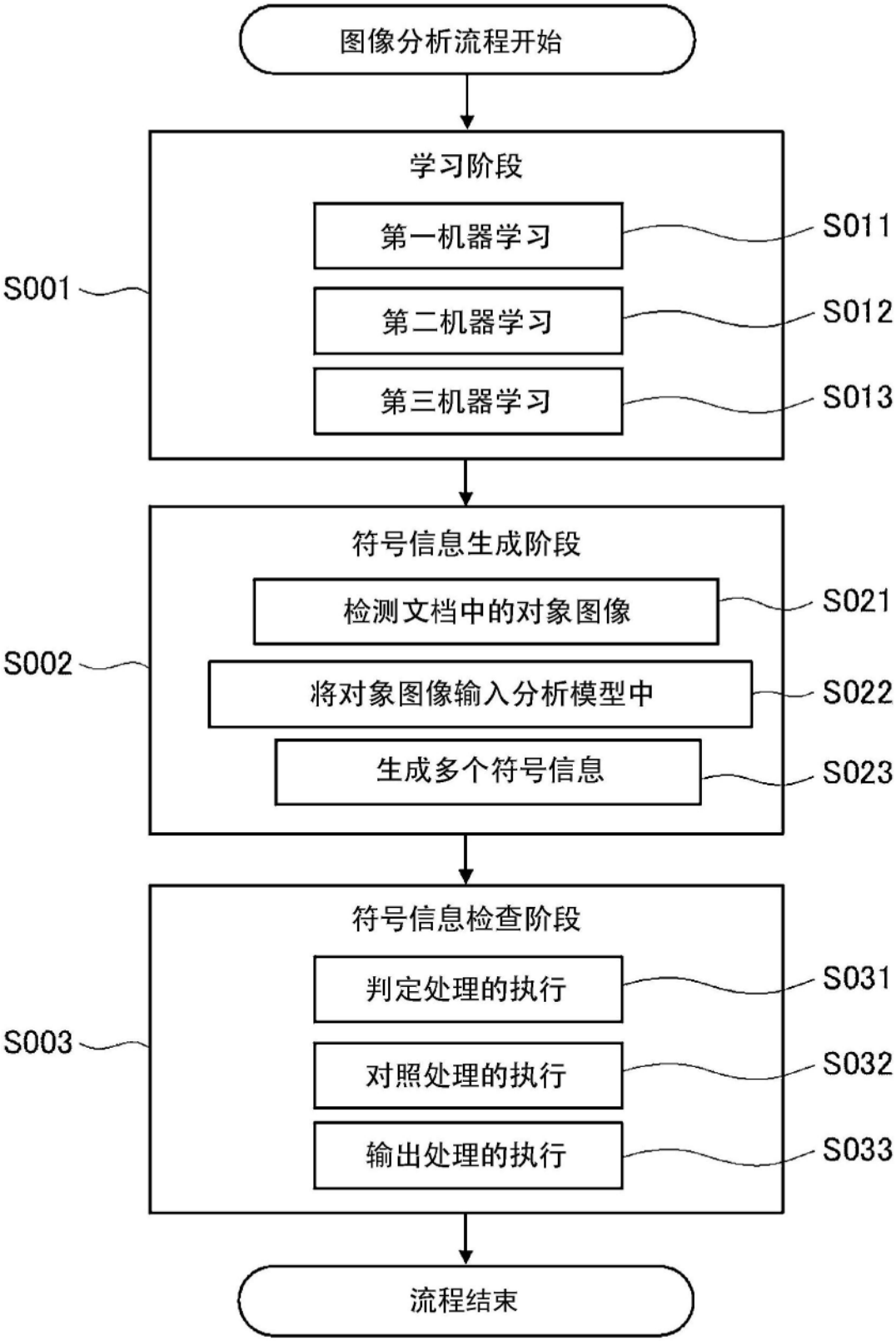

图像分析流程开始
学习阶段
第一机器学习
第二机器学习
第三机器学习
S001
S011
S012
S013
符号信息生成阶段
检测文档中的对象图像
将对象图像输入分析模型中
生成多个符号信息
S002
S021
S022
S023
符号信息检查阶段
判定处理的执行
对照处理的执行
输出处理的执行
S003
S031
S032
S033
流程结束

图4

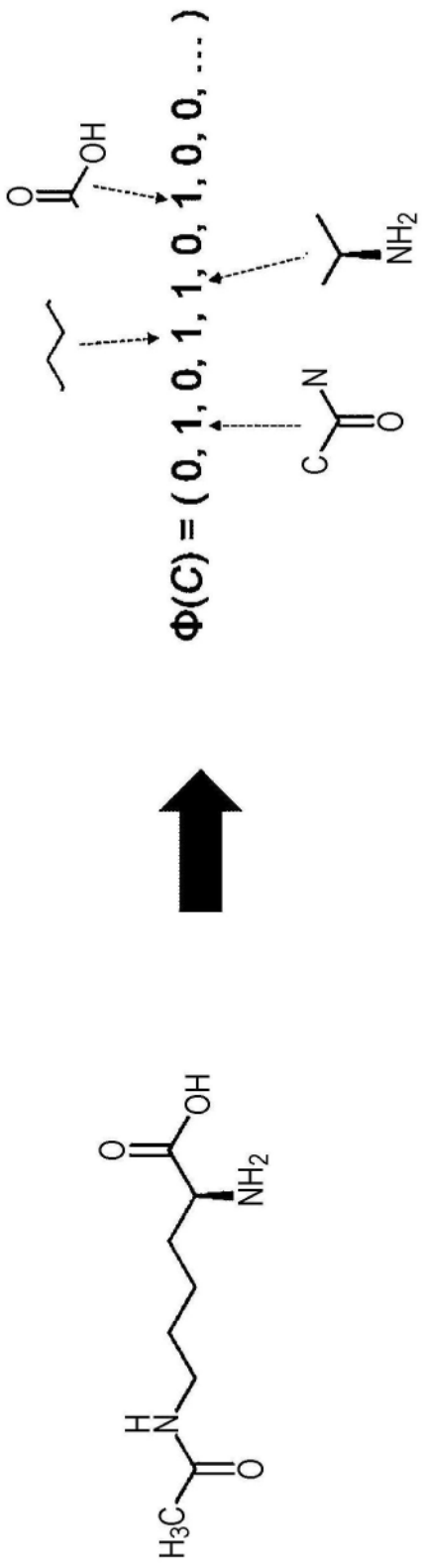


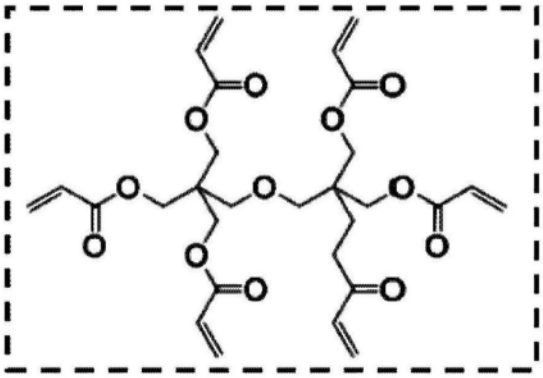




图5

(34) JP 2019-191507 A 2019.10.31

[化学式33]



DPHA

[0143]

〔实施例13～14〕

在透明树脂层的形成中，将透明树脂层形成用组合物A变更为下述所示的透明树脂层形成用组合物B、C，除此以外，用与实施例12相同的方法，获得实施例13、14的层叠体。

[0144]

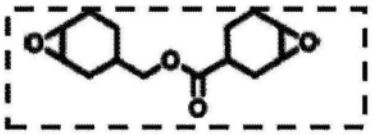
透明树脂层形成用组合物B

| | |
|---|---|
| • 下述CEL2021P（大赛璐公司制） | 144质量份 |
| • 下述IRGACURE127（BASF公司制） | 3质量份 |
| • 下述CPI-100P（碳酸丙烯酯溶液） | 6质量份 |
| • MEGAFACE RS-90（DIC制） | 0.3质量份 |
| • 甲乙酮（MEK） | 347质量份 |

[0145]

CEL2021P

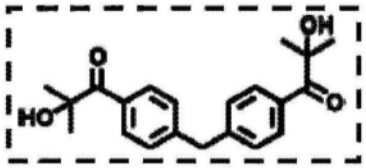
[化学式34]



[0146]

IRGACURE127

[化学式35]



[0147]

光阳离子聚合引发剂（CPI-100P）

图6

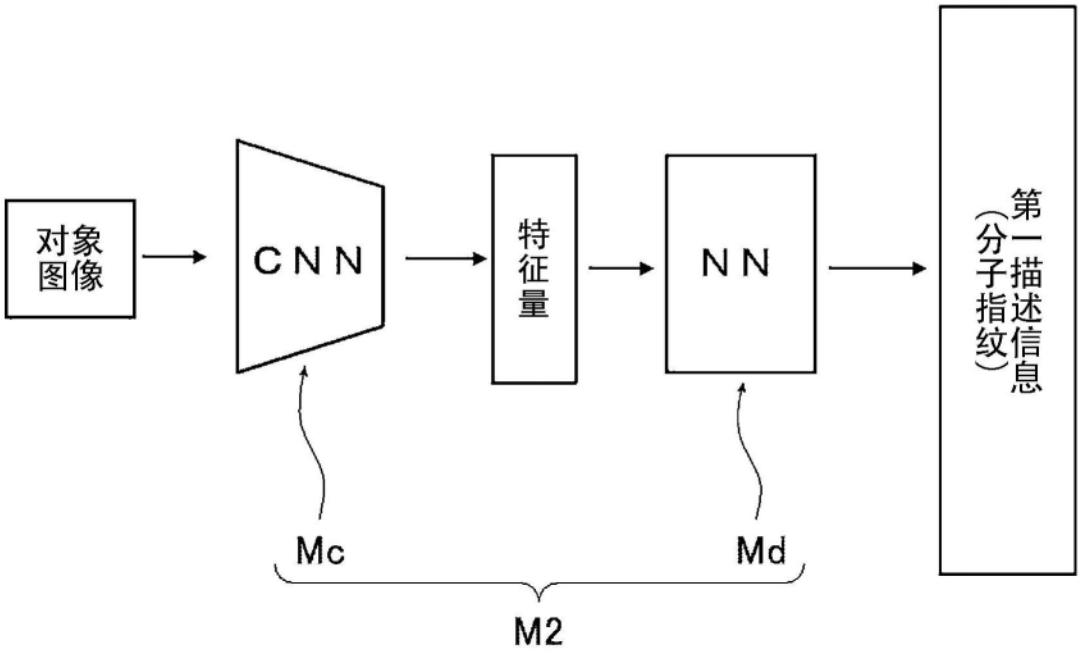



图7

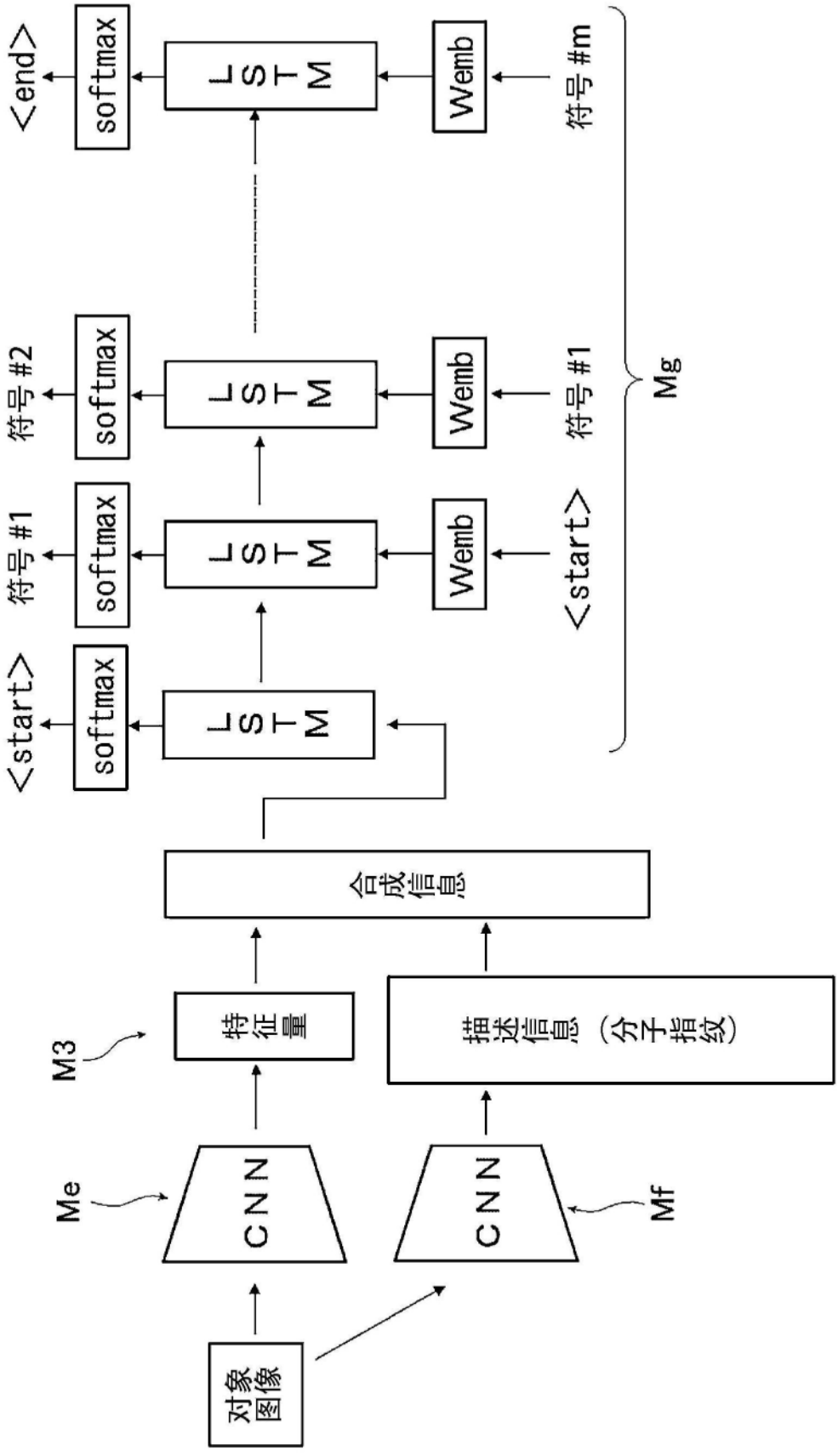

对象图像
CNN
Me
CNN
Mf
特征量
M3
描述信息（分子指纹）
合成信息
LSTM
softmax
Wemb
<start>
符号 #1
符号 #2
符号 #m
<end>
Mg

图8