(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110909768 B
(45)授权公告日 2023.03.07

(21)申请号 201911066305.1

(22)申请日 2019.11.04

(65)同一申请的已公布的文献号
申请公布号 CN 110909768 A

(43)申请公布日 2020.03.24

(73)专利权人 北京地平线机器人技术研发有限公司
地址 100086 北京市海淀区中关村大街1号3层318

(72)发明人 杜森垚

(74)专利代理机构 北京嘉科知识产权代理事务所(特殊普通合伙) 11687
专利代理师 杨波

(51)Int.Cl.
G06F 18/214(2023.01)
G06F 18/241(2023.01)
G06N 3/045(2023.01)
G06N 3/08(2023.01)

审查员 甘宇

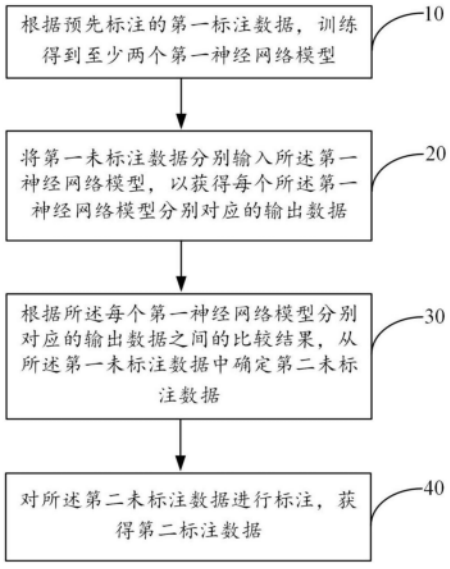
权利要求书2页 说明书10页 附图8页

(54)发明名称

一种标注数据获取方法及装置

(57)摘要

公开了一种标注数据获取方法、装置、计算机可读存储介质及电子设备，该方法包括：根据预先标注的第一标注数据，训练得到至少两个第一神经网络模型；将第一未标注数据分别输入所述第一神经网络模型，以获得每个所述第一神经网络模型分别对应的输出数据；根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果，从所述第一未标注数据中确定第二未标注数据；对所述第二未标注数据进行标注，获得第二标注数据。本申请有效提高了数据标注的效率，减少了进行数据标注的时间，有助于改善神经网络模型的性能。



CN 110909768 B

1.一种标注数据获取方法,所述方法包括:

根据预先标注的第一标注数据,训练得到至少两个第一神经网络模型;

将第一未标注数据分别输入所述第一神经网络模型,以获得每个所述第一神经网络模型分别对应的输出数据,所述第一未标注数据为点云地图,所述输出数据包括至少一个所述点云地图对应的预测的类别属性及所述类别属性对应的概率值;

根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果,从所述第一未标注数据中确定第二未标注数据;

对所述第二未标注数据进行标注,获得第二标注数据;

所述根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果,从所述第一未标注数据中确定第二未标注数据,包括:

确定所述每个第一神经网络模型分别对应的输出数据中概率值最大的类别属性为候选类别属性,所述候选类别属性对应的概率为候选概率;

根据所述每个第一神经网络模型分别对应的候选类别属性,获取所述每个第一神经网络模型分别对应的候选概率之间的比较结果;

根据所述比较结果,选取预设数量的所述第一未标注数据作为第二未标注数据;

所述根据所述每个第一神经网络模型分别对应的候选类别属性,获取所述每个第一神经网络模型分别对应的候选概率之间的比较结果,包括:

判断所述每个第一神经网络模型分别对应的候选类别属性是否相同;

若所述每个第一神经网络模型分别对应的候选类别属性相同,则将所述每个第一神经网络模型分别对应的候选概率之间的差值作为比较结果;

若所述每个第一神经网络模型分别对应的候选类别属性不相同,则将所述每个第一神经网络模型分别对应的候选概率中最大的候选概率作为比较结果。

2.根据权利要求1所述的方法,其中,所述根据所述比较结果,选取预设数量的所述第一未标注数据作为第二未标注数据,包括:

根据所述比较结果,按照预设方式对所述第一未标注数据进行排序;

按照所述第一未标注数据的排序,选取预设数量的所述第一未标注数据作为第二未标注数据。

3.根据权利要求1所述的方法,其中,所述根据所述比较结果,选取预设数量的所述第一未标注数据作为第二未标注数据,包括:

判断所述比较结果是否大于预设值;

若所述比较结果大于所述预设值,则将所述比较结果对应的所述第一未标注数据作为第二未标注数据。

4.根据权利要求1~3任一项所述的方法,其中,所述根据预先标注的第一标注数据,确定至少两个第一神经网络模型,包括:

获取第二神经网络模型;

采用所述预先标注的第一标注数据对所述第二神经网络模型依次进行至少两次训练,以获得对应的至少两个第一神经网络模型。

5.根据权利要求4所述的方法,其中,所述对所述第二未标注数据进行标注,获得第二标注数据步骤后,还包括:

根据所述第二标注数据对所述第二神经网络模型进行训练。

6.一种标注数据获取装置,所述装置包括:

第一获取模块,用于根据预先标注的第一标注数据,训练得到至少两个第一神经网络模型;

第一数据获取模块,用于将第一未标注数据分别输入所述第一神经网络模型,以获得每个所述第一神经网络模型分别对应的输出数据,所述第一未标注数据为点云地图,所述输出数据包括至少一个所述点云地图对应的预测的类别属性及所述类别属性对应的概率值;

第二数据获取模块,用于根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果,从所述第一未标注数据中确定第二未标注数据;

标注数据获取模块,用于对所述第二未标注数据进行标注,获得第二标注数据;

所述第一数据获取模块包括候选数据获取单元、比较结果获取单元以及第一数据获取单元;

所述候选数据获取单元用于:确定所述每个第一神经网络模型分别对应的输出数据中概率值最大的类别属性为候选类别属性,所述候选类别属性对应的概率为候选概率;

所述比较结果获取单元用于:根据所述每个第一神经网络模型分别对应的候选类别属性,获取所述每个第一神经网络模型分别对应的候选概率之间的比较结果;具体包括:判断所述每个第一神经网络模型分别对应的候选类别属性是否相同;若所述每个第一神经网络模型分别对应的候选类别属性相同,则将所述每个第一神经网络模型分别对应的候选概率之间的差值作为比较结果;若所述每个第一神经网络模型分别对应的候选类别属性不相同,则将所述每个第一神经网络模型分别对应的候选概率中最大的候选概率作为比较结果;

所述第一数据获取单元用于:根据所述比较结果,选取预设数量的所述第一未标注数据作为第二未标注数据。

7.一种计算机可读存储介质,所述存储介质存储有计算机程序,所述计算机程序用于执行上述权利要求1-5任一所述的标注数据获取方法。

8.一种电子设备,所述电子设备包括:

处理器;

用于存储所述处理器可执行指令的存储器;

所述处理器,用于从所述存储器中读取所述可执行指令,并执行所述指令以实现上述权利要求1-5任一所述的标注数据获取方法。

一种标注数据获取方法及装置

技术领域

[0001] 本申请涉及深度学习技术领域，且更具体地，涉及一种标注数据获取方法及装置。

背景技术

[0002] 深度学习(Deep Learning，简称DL)是学习样本数据的内在规律和表示层次，这些学习过程中获得的信息对诸如文字，图像和声音等数据的解释有很大的帮助，其目标是让机器能够具有分析学习能力，能够识别文字、图像和声音等数据。深度学习在搜索技术，数据挖掘，机器学习，机器翻译，自然语言处理，多媒体学习，语音，推荐和个性化技术，以及其他相关领域得到了广泛应用，取得了很多成果。

[0003] 深度学习需要采用经过标注的数据来对其模型进行训练，数据量越大，越有利于深度学习模型的训练，从而有助于提升其性能。然而，由于深度学习需要海量的标注数据进行训练，而目前数据标注需要耗费巨大的时间、金钱和人力成本，导致数据标注成本高昂。

发明内容

[0004] 为了解决上述技术问题，提出了本申请。本申请的实施例提供了一种标注数据获取方法、装置、计算机可读存储介质及电子设备，有效提高了数据标注的效率，减少了进行数据标注的时间，有助于改善神经网络模型的性能。

[0005] 根据本申请的第一方面，提供了一种标注数据获取方法，包括：

[0006] 根据预先标注的第一标注数据，训练得到至少两个第一神经网络模型；

[0007] 将第一未标注数据分别输入所述第一神经网络模型，以获得每个所述第一神经网络模型分别对应的输出数据；

[0008] 根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果，从所述第一未标注数据中确定第二未标注数据；

[0009] 对所述第二未标注数据进行标注，获得第二标注数据。

[0010] 根据本申请的第二方面，提供了一种标注数据获取装置，包括：

[0011] 第一获取模块，用于根据预先标注的第一标注数据，训练得到至少两个第一神经网络模型；

[0012] 第一数据获取模块，用于将第一未标注数据分别输入所述第一神经网络模型，以获得每个所述第一神经网络模型分别对应的输出数据；

[0013] 第二数据获取模块，用于根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果，从所述第一未标注数据中确定第二未标注数据；

[0014] 标注数据获取模块，用于对所述第二未数据进行标注，获得第二标注数据。

[0015] 根据本申请的第三方面，提供了一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序用于执行上述的标注数据获取方法。

[0016] 根据本申请的第四方面，提供了一种电子设备，所述电子设备包括：

[0017] 处理器；

[0018] 用于存储所述处理器可执行指令的存储器；

[0019] 所述处理器，用于从所述存储器中读取所述可执行指令，并执行所述指令以实现上述的标注数据获取方法。

[0020] 与现有技术相比，本申请提供的标注数据获取方法、装置、计算机可读存储介质及电子设备，至少包括以下有益效果：

[0021] 本申请采用不同的第一神经网络模型对同一未标注数据进行处理，获取不同第一神经网络模型的输出数据之间的比较结果，并根据比较结果来获取未标注数据，从而挑选出神经网络模型的训练结果不稳定的样本数据进行标注，有效提高了数据标注的效率，减少了进行数据标注的时间，有助于改善神经网络模型的性能。

附图说明

[0022] 通过结合附图对本申请实施例进行更详细的描述，本申请的上述以及其他目的、特征和优势将变得更加明显。附图用来提供对本申请实施例的进一步理解，并且构成说明书的一部分，与本申请实施例一起用于解释本申请，并不构成对本申请的限制。在附图中，相同的参考标号通常代表相同部件或步骤。

[0023] 图1是本申请一示例性实施例提供的标注数据获取方法的流程示意图一。

[0024] 图2是本申请一示例性实施例提供的标注数据获取方法中获取第一神经网络模型的流程示意图。

[0025] 图3是本申请一示例性实施例提供的标注数据获取方法中获取输出数据的流程示意图。

[0026] 图4是本申请一示例性实施例提供的标注数据获取方法中获取比较结果的流程示意图。

[0027] 图5是本申请一示例性实施例提供的标注数据获取方法中获取第二未标注数据的一种流程示意图。

[0028] 图6是本申请一示例性实施例提供的标注数据获取方法中获取第二未标注数据的另一种流程示意图。

[0029] 图7是本申请一示例性实施例提供的标注数据获取方法的流程示意图二。

[0030] 图8是本申请一示例性实施例提供的标注数据获取装置的示意图一。

[0031] 图9是本申请一示例性实施例提供的标注数据获取装置中第一获取模块的示意图。

[0032] 图10是本申请一示例性实施例提供的标注数据获取装置中第一数据获取模块的示意图。

[0033] 图11是本申请一示例性实施例提供的标注数据获取装置的示意图二。

[0034] 图12是本申请一示例性实施例提供的电子设备的结构图。

具体实施方式

[0035] 下面，将参考附图详细地描述根据本申请的示例实施例。显然，所描述的实施例仅仅是本申请的一部分实施例，而不是本申请的全部实施例，应理解，本申请不受这里描述的示例实施例的限制。

[0036] 应注意到:除非另外具体说明,否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本公开的范围。

[0037] 本领域技术人员可以理解,本公开实施例中的“第一”、“第二”等术语仅用于区别不同步骤、设备或模块等,既不代表任何特定技术含义,也不表示它们之间的必然逻辑顺序。

[0038] 还应理解,在本公开实施例中,“多个”可以指两个或两个以上,“至少一个”可以指一个、两个或两个以上。

[0039] 还应理解,对于本公开实施例中提及的任一部件、数据或结构,在没有明确限定或者在前后文给出相反启示的情况下,一般可以理解为一个或多个。

[0040] 另外,本公开中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本公开中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0041] 还应理解,本公开对各个实施例的描述着重强调各个实施例之间的不同之处,其相同或相似之处可以相互参考,为了简洁,不再一一赘述。

[0042] 同时,应当明白,为了便于描述,附图中所示出的各个部分的尺寸并不是按照实际的比例关系绘制的。

[0043] 以下对至少一个示例性实施例的描述实际上仅仅是说明性的,决不作为对本公开及其应用或使用的任何限制。

[0044] 对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论,但在适当情况下,所述技术、方法和设备应当被视为说明书的一部分。

[0045] 应注意到:相似的标号和字母在下面的附图中表示类似项,因此,一旦某一项在一个附图中被定义,则在随后的附图中不需要对其进行进一步讨论。

[0046] 本公开实施例可以应用于终端设备、计算机系统、服务器等电子设备,其可与众多其它通用或专用计算系统环境或配置一起操作。适于与终端设备、计算机系统、服务器等电子设备一起使用的众所周知的终端设备、计算系统、环境和/或配置的例子包括但不限于:个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任何系统的分布式云计算技术环境,等等。

[0047] 终端设备、计算机系统、服务器等电子设备可以在由计算机系统执行的计算机系统可执行指令(诸如程序模块)的一般语境下描述。通常,程序模块可以包括例程、程序、目标程序、组件、逻辑、数据结构等等,它们执行特定的任务或者实现特定的抽象数据类型。计算机系统/服务器可以在分布式云计算环境中实施,分布式云计算环境中,任务是由通过通信网络链接的远程处理设备执行的。在分布式云计算环境中,程序模块可以位于包括存储设备的本地或远程计算系统存储介质上。

[0048] 申请概述

[0049] 深度学习在计算机视觉(ComputerVision,简写为CV)、自然语言处理(Natural Language Processing,简写为NLP)等诸多领域具有非常广泛的应用。深度学习需要采用经过标注的数据来对其模型进行训练,数据量越大,越有利于神经网络模型的训练,因而为了获得具有良好性能的神经网络模型,需要提供海量的标注数据进行训练。目前,用于深度学

习的数据需要经过人工进行标注，然后才能将标注数据用于神经网络模型的训练，需要耗费巨大的时间和人力成本，导致数据标注成本高昂。

[0050] 主动学习通过对未标注的数据进行筛选，可以利用少量的标注数据取得较高的学习准确度，因此，深度学习中的主动学习方法也成为了研究的热点。本实施例则提出了一种标注数据的获取方法，通过获取经过不同第一神经网络模型后输出数据差异较大的未标注数据，从而挑选出神经网络模型的训练结果不稳定的样本进行标注，可以有效减少进行数据标注的时间，提高数据标注的效率，改善神经网络模型的性能。

[0051] 示例性方法

[0052] 图1是本申请一示例性实施例提供的标注数据获取方法的流程示意图。本实施例可应用在服务器上，如图1所示，包括如下步骤：

[0053] 步骤10：根据预先标注的第一标注数据，训练得到至少两个第一神经网络模型。

[0054] 为了对神经网络模型进行训练和优化，需要采用数据集来对神经网络模型不断进行优化。因此，在对神经网络进行训练时，首先需要获取经过预先标注的标注数据构成一个数据集，然后采用该数据集对神经网络模型不断进行训练。在本实施例中，经过人工预先标注的数据记为第一标注数据，第一标注数据的数量可以根据需要进行选择。在进行训练时，对神经网路模型经过至少两个epoch(将包括第一标注数据的完整数据集通过神经网络模型一次并且返回一次，这个过程称为一个epoch)，每个epoch后均可获得一个对应的经过训练的神经网络模型，记为第一神经网络模型。

[0055] 步骤20：将第一未标注数据分别输入所述第一神经网络模型，以获得每个所述第一神经网络模型分别对应的输出数据。

[0056] 经过训练得到的第一神经网络模型可用于对第一未标注数据进行处理。可以理解的是，不同的第一神经网络，对于相同的输入，其输出数据也可能会出现不同。因此本实施中，在对第一未标注数据进行处理时，可以将同一第一未标注数据分别输入至少两个第一神经网络模型中，并获得对应的输出数据。

[0057] 步骤30：根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果，从所述第一未标注数据中确定第二未标注数据。

[0058] 根据步骤20中所述，不同的第一神经网络，对于相同的输入，其输出数据也可能会出现不同。因此，在获得了同一个第一未标注数据经过不同第一神经网络模型的输出数据后，可以获取输出数据之间的比较结果。可以理解的是，神经网络模型的的训练越充分，则训练获得的不同第一神经网络模型之间的差异就越小，意味着同样的数据经过不同的第一神经网络模型后的输出数据之间的差异越小。相反，神经网络模型的的训练越不充分，则训练获得的不同第一神经网络模型之间的差异就越大，意味着同样的数据经过不同的第一神经网络模型后的输出数据之间的差异越大。而差异越大的输出数据对应的第一未标注数据对于神经网络模型的训练来说具有更大的价值，此时需要对该输出数据对应的第一未标注数据进行处理。因此，本实施例基于不同输出数据之间的比较结果从第一未标注数据中选取符合要求的数据作为第二未标注数据。

[0059] 步骤40：对所述第二未标注数据进行标注，获得第二标注数据。

[0060] 通过步骤30所获得的第二未标注数据对于神经网络模型的训练而言具有更大的价值，因此在获得了第二未标注数据后，需要对第二未标注数据进行标注，以获得第二标注

数据。在本实施例中，可以采用人工的方式对第二未标注数据进行标注，也可以采用其他方式，此处不做限制。

[0061] 本实施例提供的标注数据获取方法的有益效果至少在于：

[0062] 本实施例采用不同的第一神经网络模型对同一未标注数据进行处理，获取不同第一神经网络模型的输出数据之间的比较结果，并根据比较结果来获取未标注数据，从而挑选出神经网络模型的训练结果不稳定的样本数据进行标注，有效提高了数据标注的效率，减少了进行数据标注的时间，有助于改善神经网络模型的性能。

[0063] 图2示出了如图1所示的实施例中根据预先标注的第一标注数据，训练得到至少两个第一神经网络模型步骤的流程示意图。

[0064] 如图2所示，在上述图1所示实施例的基础上，本申请一个示例性实施例中，步骤10所示获取至少两个第一神经网络模型步骤，具体可以包括：

[0065] 步骤101：获取第二神经网络模型。

[0066] 此处的第二神经网络模型可以是未经任何训练的神经网络模型，也可以是已经经过部分训练的神经网络模型。神经网络模型的类型可以根据需要进行选择，可以是各个领域的神经网络模型，本实施例不做任何限制。

[0067] 步骤102：采用所述预先标注的第一标注数据对所述第二神经网络模型依次进行至少两次训练，以获得对应的至少两个第一神经网络模型。

[0068] 可以理解的是，当数据集中第一标注数据的数量一定时，对神经网络模型训练的次数越多，所获得的神经网络模型的性能就越好。因此，通常都需要对神经网络模型进行多个epoch，越往后获得的神经网络模型的性能越佳，在选取第一神经网络模型时，通常选择最后的至少2个epoch所得到的第一神经网络模型。

[0069] 例如，当对神经网络模型进行2个epoch时，则选择该2个epoch所得到的第一神经网络模型。

[0070] 当对神经网络模型进行多个epoch，且所要选择的第一神经网络模型的数量为2个时，则选择最后2个epoch所得到的第一神经网络模型(例如，当进行N个epoch时，选择的epoch为第N-1个以及第N个epoch)。

[0071] 当对神经网络模型进行多个epoch，且所要选择的第一神经网络模型的数量为多个时，则选择最后的几个epoch所得到的第一神经网络模型。

[0072] 当然，在其他实施例中，第一神经网络模型也可以通过其他方式进行选取，并不仅限于上述的方式，此处不做限制。

[0073] 在本实施例中，通过选择最后的至少2个epoch所得到的第一神经网络模型，可以有效确保所获得的第一神经网络模型的性能，从而有利于后续对未标注数据进行处理，提高所获取的未标注数据的价值。

[0074] 在一个实施例中，步骤20中，将第一未标注数据输入不同的第一神经网络模型后，可以获得对应的输出数据，输出数据包括至少一个预测的类别属性及该类别属性对应的概率值。例如，当输入的第一未标注数据是一张点云地图时，其中可能包括树的点云、路面点云、车辆点云等等，每一个点云就是一个第一未标注数据，其包括了该点云是每一种类别的概率，例如该点云是树的概率是P_1，是路面的概率是P_2，是车辆的概率是P_3。当然，在其他实施例中，输出数据中还可以包括其他类型的数据，并不仅限于上述的情形。

[0075] 图3示出了如图1所示的实施例中根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果，从所述第一未标注数据中确定第二未标注数据步骤的流程示意图。

[0076] 如图3所示，在上述图1所示实施例的基础上，本申请一个示例性实施例中，步骤20所示获得每个所述第一神经网络模型分别对应的输出数据步骤，具体可以包括：

[0077] 步骤201：确定所述每个第一神经网络模型分别对应的输出数据中概率值最大的类别属性为候选类别属性，所述候选类别属性对应的概率为候选概率。

[0078] 当通过第一神经网络模型后获得的输出数据中包括多种类别属性和对应的概率值时，需要首先从中选择进行后续比较的一组。概率值越高，则该第一未标注数据属于该类别的可能性就越大。因此，选择概率最大的类别属性作为该第一未标注数据的候选类别属性，有助于提高后续确定第二未标注数据的准确性和有效性。

[0079] 步骤202：根据所述每个第一神经网络模型分别对应的候选类别属性，获取所述每个第一神经网络模型分别对应的候选概率之间的比较结果。

[0080] 由于第一神经网络模型的数量至少为两个，且每个第一神经网络模型的训练程度不同，当同一个第一未标注数据经过不同的第一神经网络模型后，输出数据可能会不同，因此需要对不同的输出数据进行比较，以获得比较结果。

[0081] 步骤203：根据所述比较结果，选取预设数量的所述第一未标注数据作为第二未标注数据。

[0082] 第二未标注数据的需求数量可以根据需要进行预先设定，例如可以根据比较结果是否满足预设需求来设定选取第二未标注数据的数量。

[0083] 本实施例通过将每个第一神经网络模型的输出数据中概率值最大的类别属性和概率确定为候选类别属性和候选概率，并将不同第一神经网络模型的输出数据进行比较以获得比较结果，最终根据比较结果来确定第二未标注数据，可以有效提高所获取的第二未标注数据的准确性和有效性。

[0084] 图4示出了如图3所示的实施例中根据所述每个第一神经网络模型分别对应的候选类别属性，获取所述每个第一神经网络模型分别对应的候选概率之间的比较结果步骤的流程示意图。

[0085] 如图4所示，在上述图3所示实施例的基础上，本申请一个示例性实施例中，步骤202所示获取比较结果步骤，具体可以包括：

[0086] 步骤2021：判断所述每个第一神经网络模型分别对应的候选类别属性是否相同。

[0087] 由于第一神经网络模型的数量至少为两个，且每个第一神经网络模型的训练程度不同，当同一个第一未标注数据经过不同的第一神经网络模型后，输出数据的类别属性可能会不相同，因此首先需要判断不同神经网络模型的输出数据的类别是否相同，从而选择不同的处理方式。

[0088] 以第一神经网络模型的数量为2个为例。当经过两个第一神经网络模型后输出数据的候选类别属性相同时，此时需要将两个候选概率进行比较，比较的方式可以是获取两者之间的差值，并将差值作为比较结果。当经过两个第一神经网络模型后输出数据的类型属性不相同时，则无法获取两个候选概率之间的差值，而可以将两个候选概率进行比较，并将比较的结果作为比较结果。比较结果可以反映两个候选概率之间的偏差。

[0089] 若所述每个第一神经网络模型分别对应的候选类别属性相同，则：

[0090] 步骤2022：将所述每个第一神经网络模型分别对应的候选概率之间的差值作为比较结果。

[0091] 以第一神经网络模型的数量为2个为例，当两个第一神经网络模型的候选类别属性相同时，此时可以将两个候选概率两者之间的差值作为比较结果：

[0092] $epoch\_score = |\max(P(y_1|x)) - \max(P(y_2|x))|$

[0093] 其中，epoch_score为比较结果；

[0094] $\max(P(y_1|x))$为第一个神经网络模型的候选概率最大值；

[0095] $\max(P(y_2|x))$为第二个神经网络模型的候选概率最大值。

[0096] 若所述每个第一神经网络模型分别对应的候选类别属性不相同，则：

[0097] 步骤2023：将所述每个第一神经网络模型分别对应的候选概率中最大的候选概率作为比较结果。

[0098] 以第一神经网络模型的数量为2个为例，当两个第一神经网络模型的候选类别属性不相同时，直接比较不同神经网络模型的输出数据中的概率，将值最大的候选概率作为比较结果：

[0099] $epoch\_score = \max(|\max(P(y_1|x))|, |\max(P(y_2|x))|)$

[0100] 当然，第一神经网络模型的数量为3个及以上时，采取的方式类似，只需要比较多个第一神经网络模型输出数据中的概率，将值最大的候选概率作为比较结果即可。

[0101] 本实施例根据不同第一神经网络模型输出数据中类别属性是否相同来选择不同的方式进行处理，从而可以有效获取比较结果，有助于后续根据比较结果来获取第二未标注数据。

[0102] 进一步地，在获取了比较结果后，可以采用不同的方式从第一未标注数据中确定第二未标注数据。

[0103] 如图5所示，在上述图1所示实施例的基础上，本申请一个示例性实施例中，步骤30所示确定第二未标注数据步骤，具体可以包括：

[0104] 步骤301：根据所述比较结果，按照预设方式对所述第一未标注数据进行排序。

[0105] 例如，当每个第一神经网络模型分别对应的候选类别属性相同时，所获得的比较结果为两个候选概率两者之间的差值，差值越大，则说明第一神经网络模型之间的差异越大，该第一未标注数据经过第一神经网络模型后的结果不收敛，因此需要挑选出该第一未标注数据进行标注。根据差值大小进行排序，可以是从差值大的输出数据对应的第一未标注数据依次排到差值小的输出数据对应的第一未标注数据，也可以是从差值小的输出数据对应的第一未标注数据依次排到差值大的输出数据对应的第一未标注数据。在本实施例中，优选从差值大的输出数据对应的第一未标注数据依次排到差值小的输出数据对应的第一未标注数据。

[0106] 再如，当每个第一神经网络模型分别对应的候选类别属性不相同时，所获得的比较结果为至少两个候选概率中概率最大的值。该概率值越大，则意味着第一神经网络模型的训练程度越高；相反，该概率值越小，则意味着第一神经网络模型的训练程度越缺乏，因此需要挑选出该第一未标注数据进行标注。根据概率值大小进行排序，可以是从概率值大的输出数据对应的第一未标注数据依次排到概率值小的输出数据对应的第一未标注数据，

也可以是从概率值小的输出数据对应的第一未标注数据依次排到概率值大的输出数据对应的第一未标注数据。在本实施例中，优选从概率值小的输出数据对应的第一未标注数据依次排到概率值大的输出数据对应的第一未标注数据。

[0107] 步骤302：按照所述第一未标注数据的排序，选取预设数量的所述第一未标注数据作为第二未标注数据。

[0108] 当根据预设方式将第一未标注数据进行排列后，可以根据预设数量依次选取对应的第一未标注数据即可，选取的数量可以根据需要进行设置。

[0109] 本实施例根据预设方式对第一未标注数据进行排序后再选取预设数量的第一未标注数据作为第二未标注数据，可以有效提高选取的未标注数据的质量，有助于提高数据标注的效率，进而有助于改善神经网络模型的性能。

[0110] 如图6所示，在上述图1所示实施例的基础上，本申请一个示例性实施例中，步骤30所示确定第二未标注数据步骤，具体可以包括：

[0111] 步骤303：判断所述比较结果是否大于预设值。

[0112] 若所述比较结果大于所述预设值，则：

[0113] 步骤304：将所述比较结果对应的所述第一未标注数据作为第二未标注数据。

[0114] 若所述比较结果大于所述预设值，则：

[0115] 步骤305：不选取该比较结果对应的所述第一未标注数据。

[0116] 例如，当每个第一神经网络模型分别对应的候选类别属性相同时，所获得的比较结果为两个候选概率两者之间的差值，通过判断该差值是否大于预设值，来对第一未标注数据进行筛选。当差值大于预设值时，说明不同第一神经网络模型之间的差异太大，还没有达到预设要求，因此需要挑选出该第一未标注数据进行标注。

[0117] 再如，当每个第一神经网络模型分别对应的候选类别属性不相同时，所获得的比较结果为至少两个候选概率中概率最大的值。当差值大于预设值时，说明第一神经网络模型的训练程度太低，还没有达到预设要求，因此需要挑选出该第一未标注数据进行标注。

[0118] 本实施例根据预设方式对第一未标注数据进行筛选，将筛选后的第一未标注数据作为第二未标注数据，可以有效提高选取的未标注数据的质量，有助于提高数据标注的效率，进而有助于改善神经网络模型的性能。

[0119] 进一步地，如图7所示，步骤40对所述第二未标注数据进行标注，获得第二标注数据后，还可以包括如下步骤：

[0120] 步骤50：根据所述第二标注数据对所述第二神经网络模型进行训练。

[0121] 即在获得了第二标注数据后，可以将第二标注数据加入到数据集中，进一步用于对神经网络模型进行训练，从而可以提高神经网络模型的性能。上述过程可以不断进行重复，重复的次数可以根据需要进行设定，例如可以根据神经网络模型训练的预算需求、性能需求等进行设定。

[0122] 示例性装置

[0123] 图8是本申请一示例性实施例提供的标注数据获取装置的示意图。标注数据获取装置包括第一获取模块61、第一数据获取模块62、第二数据获取模块63以及标注数据获取模块64。其中，第一获取模块61用于根据预先标注的第一标注数据，训练得到至少两个第一神经网络模型；第一数据获取模块62用于将第一未标注数据分别输入所述第一神经网络模

型，以获得每个所述第一神经网络模型分别对应的输出数据；第二数据获取模块63用于根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果，从所述第一未标注数据中确定第二未标注数据；标注数据获取模块64用于对所述第二未数据进行标注，获得第二标注数据。

[0124] 进一步地，请参阅图9，第一获取模块61包括模型获取单元611和第一获取单元612。其中，模型获取单元611用于获取第二神经网络模型，第一获取单元612用于采用所述预先标注的第一标注数据对所述第二神经网络模型依次进行至少两次训练，以获得对应的至少两个第一神经网络模型。

[0125] 进一步地，请参阅图10，第一数据获取模块62包括候选数据获取单元621、比较结果获取单元622以及第一数据获取单元623。其中，候选数据获取单元621用于确定所述每个第一神经网络模型分别对应的输出数据中概率值最大的类别属性为候选类别属性，所述候选类别属性对应的概率为候选概率；比较结果获取单元622用于根据所述每个第一神经网络模型分别对应的候选类别属性，获取所述每个第一神经网络模型分别对应的候选概率之间的比较结果；第一数据获取单元623用于根据所述比较结果，选取预设数量的所述第一未标注数据作为第二未标注数据。

[0126] 进一步地，请参阅图11，标注数据获取装置还包括训练模块65，训练模块65用于根据所述第二标注数据对所述第二神经网络模型进行训练。

[0127] 示例性电子设备

[0128] 图12图示了根据本申请实施例的电子设备的框图。如图12所示，电子设备70包括一个或多个处理器71和存储器72。

[0129] 处理器71可以是中央处理单元(CPU)或者具有数据处理能力和/或指令执行能力的其他形式的处理单元，并且可以控制电子设备70中的其他组件以执行期望的功能。

[0130] 存储器72可以包括一个或多个计算机程序产品，所述计算机程序产品可以包括各种形式的计算机可读存储介质，例如易失性存储器和/或非易失性存储器。所述易失性存储器例如可以包括随机存取存储器(RAM)和/或高速缓冲存储器(cache)等。所述非易失性存储器例如可以包括只读存储器(ROM)、硬盘、闪存等。在所述计算机可读存储介质上可以存储一个或多个计算机程序指令，处理器71可以运行所述程序指令，以实现上文所述的本申请的各个实施例的标注数据获取方法以及/或者其他期望的功能。

[0131] 在一个示例中，电子设备70还可以包括：输入装置73和输出装置74，这些组件通过总线系统和/或其他形式的连接机构(未示出)互连。该输入装置73可以用于捕捉声源的输入信号。此外，该输入装置73还可以包括例如键盘、鼠标等等。该输出装置74可以向外部输出各种信息，输出装置74可以包括例如显示器、扬声器、打印机、以及通信网络及其所连接的远程输出设备等。

[0132] 当然，为了简化，图12中仅示出了该电子设备70中与本申请有关的组件中的一些，省略了诸如总线、输入/输出接口等等的组件。除此之外，根据具体应用情况，电子设备70还可以包括任何其他适当的组件。

[0133] 示例性计算机程序产品和计算机可读存储介质

[0134] 除了上述方法和设备以外，本申请的实施例还可以是计算机程序产品，其包括计算机程序指令，所述计算机程序指令在被处理器运行时使得所述处理器执行本说明书上述

“示例性方法”部分中描述的根据本申请各种实施例的标注数据获取方法中的步骤。

[0135] 所述计算机程序产品可以以一种或多种程序设计语言的任意组合来编写用于执行本申请实施例操作的程序代码，所述程序设计语言包括面向对象的程序设计语言，诸如Java、C++等，还包括常规的过程式程序设计语言，诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。

[0136] 此外，本申请的实施例还可以是计算机可读存储介质，其上存储有计算机程序指令，所述计算机程序指令在被处理器运行时使得所述处理器执行本说明书上述“示例性方法”部分中描述的根据本申请各种实施例的标注数据获取方法中的步骤。

[0137] 所述计算机可读存储介质可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以包括但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0138] 以上结合具体实施例描述了本申请的基本原理，但是，需要指出的是，在本申请中提及的优点、优势、效果等仅是示例而非限制，不能认为这些优点、优势、效果等是本申请的各个实施例必须具备的。另外，上述公开的具体细节仅是为了示例的作用和便于理解的作用，而非限制，上述细节并不限制本申请为必须采用上述具体的细节来实现。

[0139] 本申请中涉及的器件、装置、设备、系统的方框图仅作为例示性的例子并且不意图要求或暗示必须按照方框图示出的方式进行连接、布置、配置。如本领域技术人员将认识到的，可以按任意方式连接、布置、配置这些器件、装置、设备、系统。诸如“包括”、“包含”、“具有”等等的词语是开放性词汇，指“包括但不限于”，且可与其互换使用。这里所使用的词汇“或”和“和”指词汇“和/或”，且可与其互换使用，除非上下文明确指示不是如此。这里所使用的词汇“诸如”指词组“诸如但不限于”，且可与其互换使用。

[0140] 还需要指出的是，在本申请的装置、设备和方法中，各部件或各步骤是可以分解和/或重新组合的。这些分解和/或重新组合应视为本申请的等效方案。

[0141] 提供所公开的方面的以上描述以使本领域的任何技术人员能够做出或者使用本申请。对这些方面的各种修改对于本领域技术人员而言是非常显而易见的，并且在此定义的一般原理可以应用于其他方面而不脱离本申请的范围。因此，本申请不意图被限制到在此示出的方面，而是按照与在此公开的原理和新颖的特征一致的最宽范围。

[0142] 为了例示和描述的目的已经给出了以上描述。此外，此描述不意图将本申请的实施例限制到在此公开的形式。尽管以上已经讨论了多个示例方面和实施例，但是本领域技术人员将认识到其某些变型、修改、改变、添加和子组合。

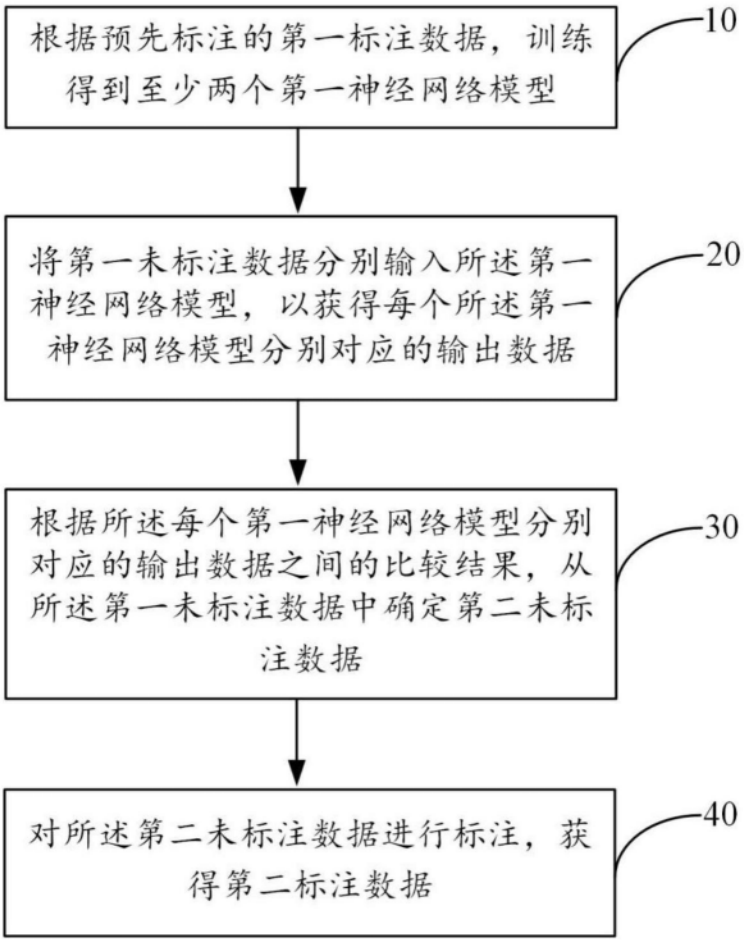

根据预先标注的第一标注数据，训练得到至少两个第一神经网络模型
10
将第一未标注数据分别输入所述第一神经网络模型，以获得每个所述第一神经网络模型分别对应的输出数据
20
根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果，从所述第一未标注数据中确定第二未标注数据
30
对所述第二未标注数据进行标注，获得第二标注数据
40

图1

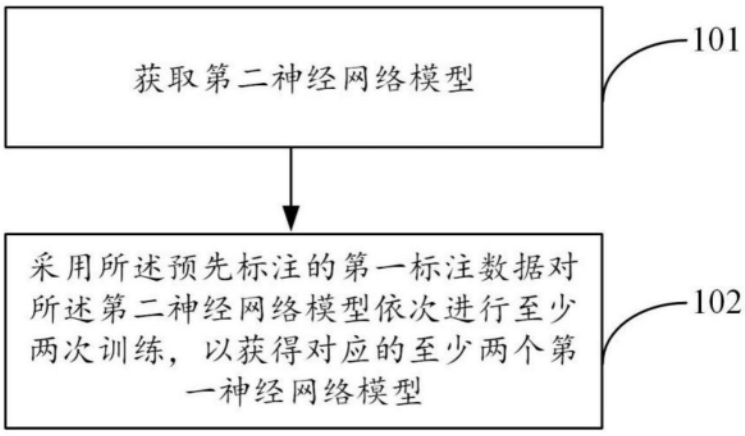

获取第二神经网络模型
101
采用所述预先标注的第一标注数据对所述第二神经网络模型依次进行至少两次训练，以获得对应的至少两个第一神经网络模型
102

图2

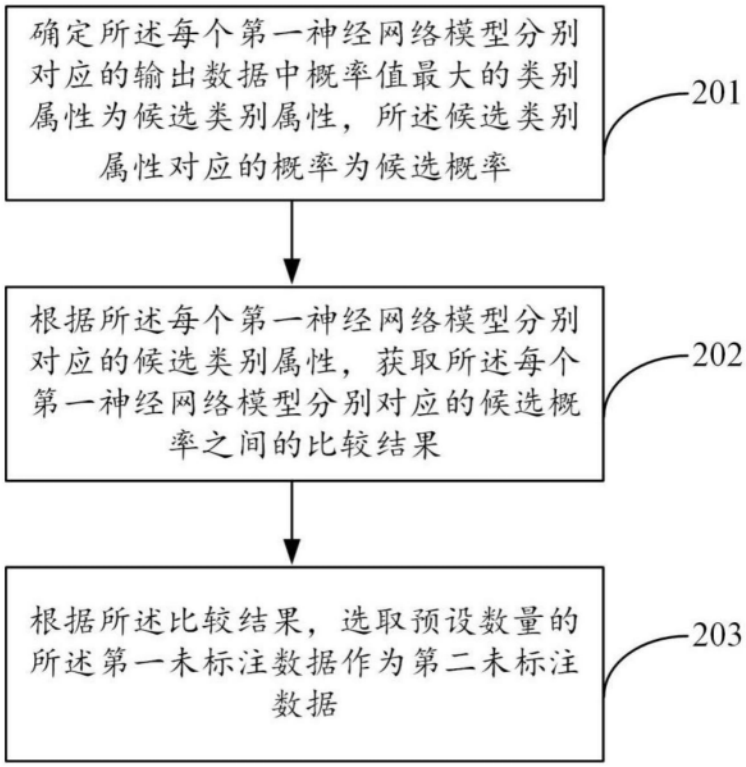

确定所述每个第一神经网络模型分别对应的输出数据中概率值最大的类别属性为候选类别属性，所述候选类别属性对应的概率为候选概率
201
根据所述每个第一神经网络模型分别对应的候选类别属性，获取所述每个第一神经网络模型分别对应的候选概率之间的比较结果
202
根据所述比较结果，选取预设数量的所述第一未标注数据作为第二未标注数据
203

图3

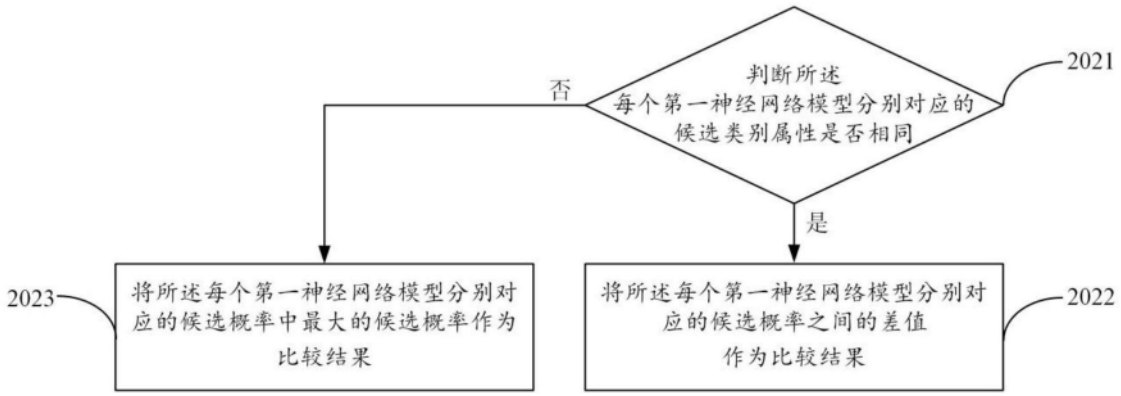

判断所述每个第一神经网络模型分别对应的候选类别属性是否相同
2021
否
是
将所述每个第一神经网络模型分别对应的候选概率中最大的候选概率作为比较结果
2023
将所述每个第一神经网络模型分别对应的候选概率之间的差值作为比较结果
2022

图4

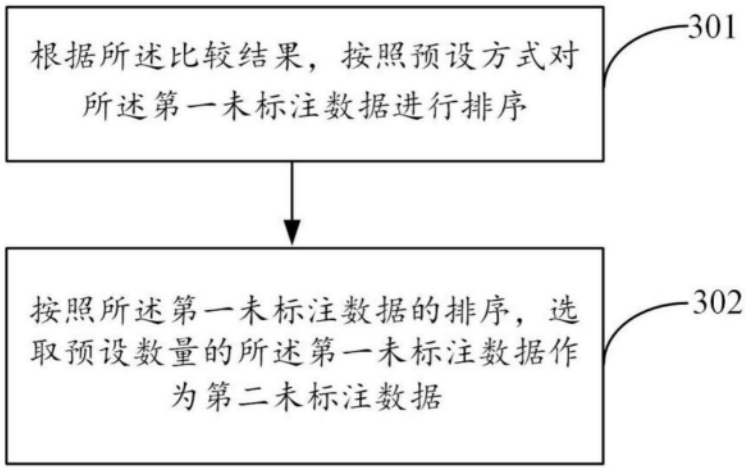

根据所述比较结果，按照预设方式对所述第一未标注数据进行排序
301
按照所述第一未标注数据的排序，选取预设数量的所述第一未标注数据作为第二未标注数据
302

图5

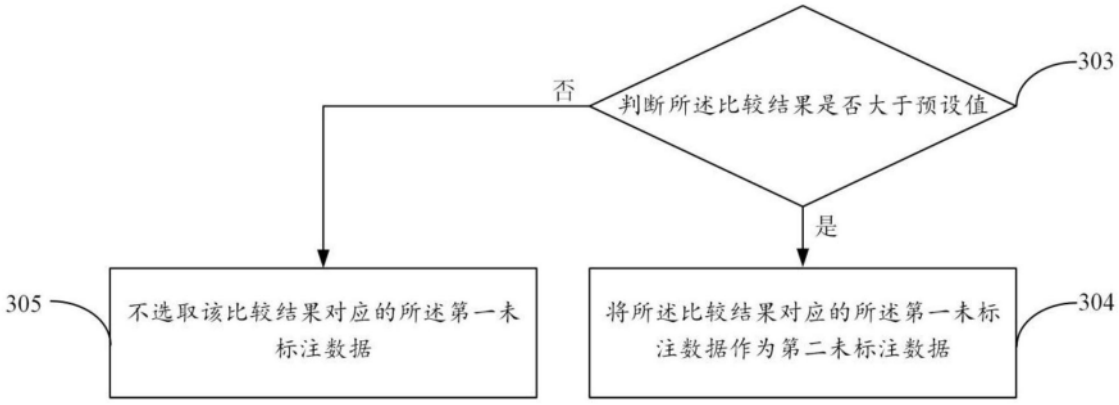

判断所述比较结果是否大于预设值
303
否
是
不选取该比较结果对应的所述第一未标注数据
305
将所述比较结果对应的所述第一未标注数据作为第二未标注数据
304

图6

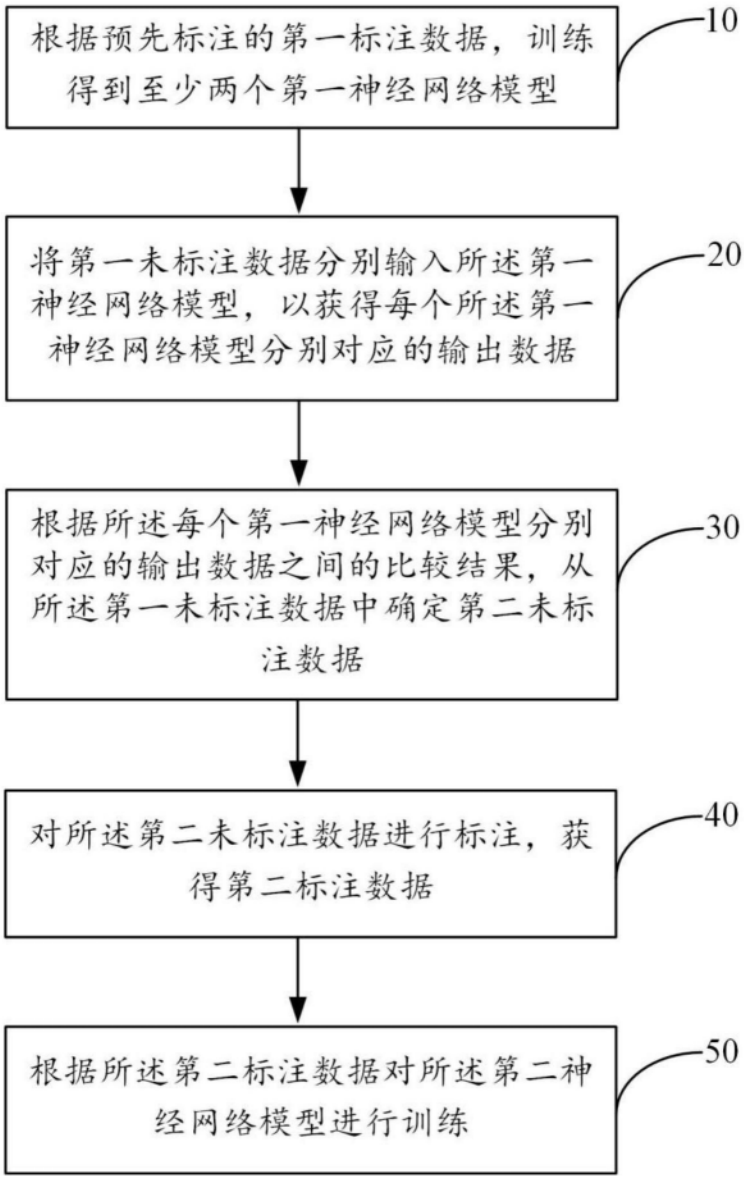

根据预先标注的第一标注数据，训练得到至少两个第一神经网络模型
10
将第一未标注数据分别输入所述第一神经网络模型，以获得每个所述第一神经网络模型分别对应的输出数据
20
根据所述每个第一神经网络模型分别对应的输出数据之间的比较结果，从所述第一未标注数据中确定第二未标注数据
30
对所述第二未标注数据进行标注，获得第二标注数据
40
根据所述第二标注数据对所述第二神经网络模型进行训练
50

图7

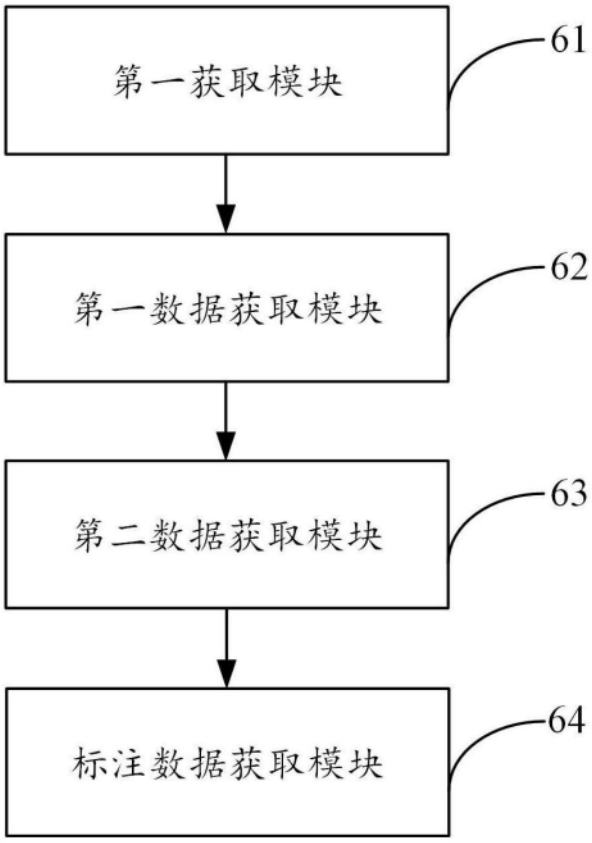

第一获取模块
61
第一数据获取模块
62
第二数据获取模块
63
标注数据获取模块
64

图8

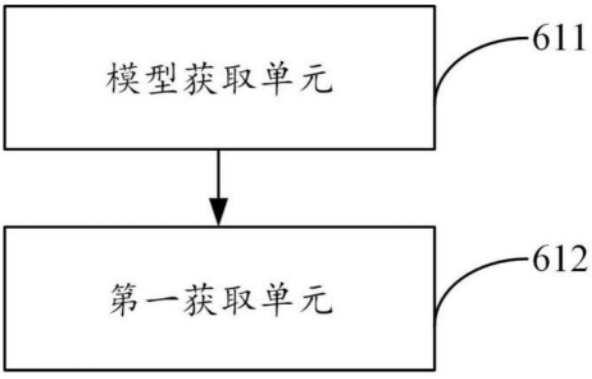

模型获取单元
611
第一获取单元
612

图9

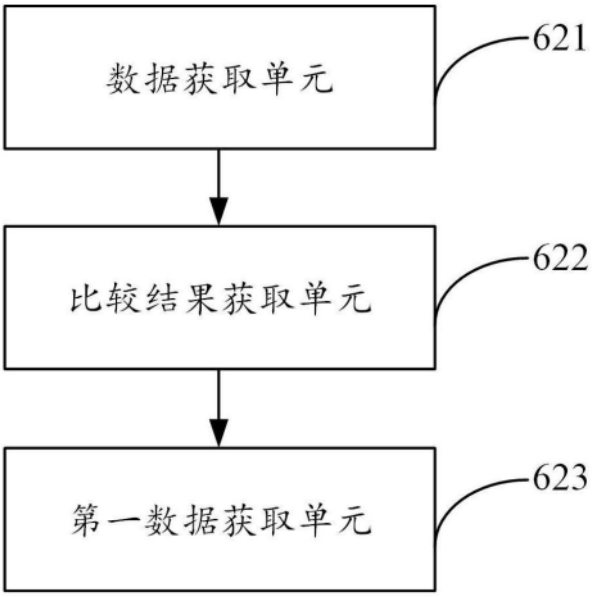


图10

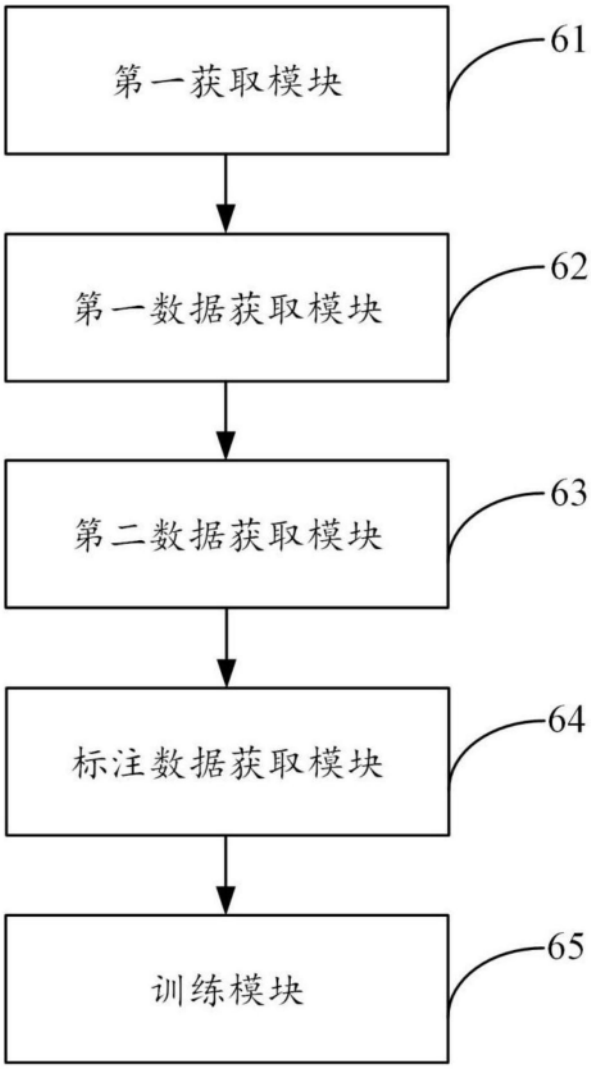


图11

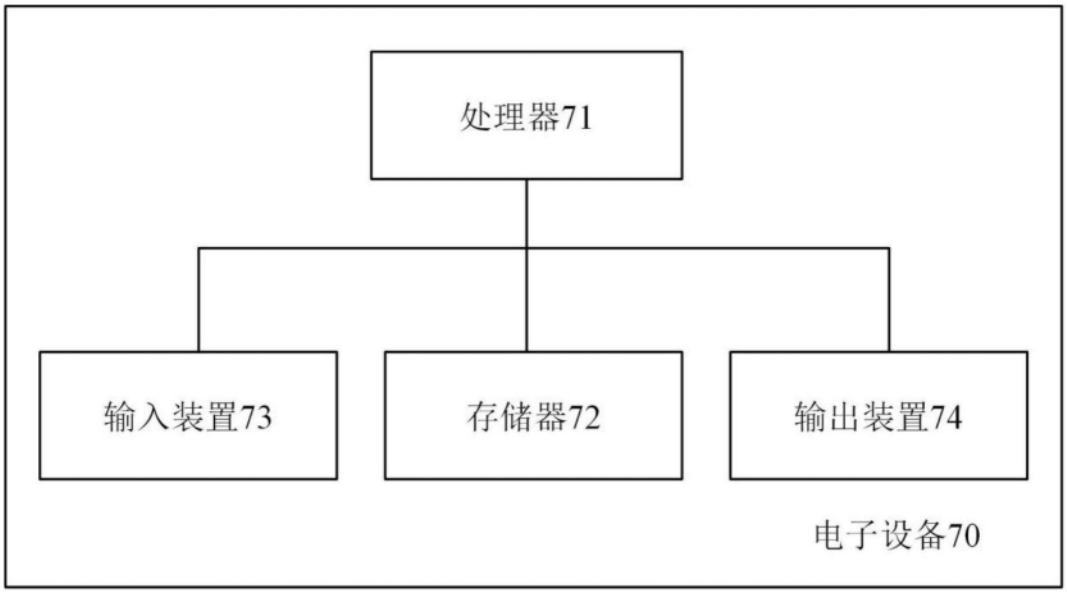


图12