



(12) 发明专利申请

(10) 申请公布号 CN 115100484 A

(43) 申请公布日 2022. 09. 23

(21) 申请号 202210551037.8

G06N 5/04 (2006.01)

(22) 申请日 2022.05.18

(71) 申请人 北京旷视科技有限公司

地址 100096 北京市海淀区西三旗建材城
内建中路12幢一层1268号

(72) 发明人 王塑 金宇 周昕宇 张朔 吕杰
厉莲莲 颜露 冀涛

(74) 专利代理机构 北京超凡宏宇专利代理事务
所(特殊普通合伙) 11463
专利代理师 董艳芳

(51) Int.Cl.

G06V 10/774 (2022.01)

G06V 10/764 (2022.01)

G06K 9/62 (2022.01)

G06N 3/08 (2006.01)

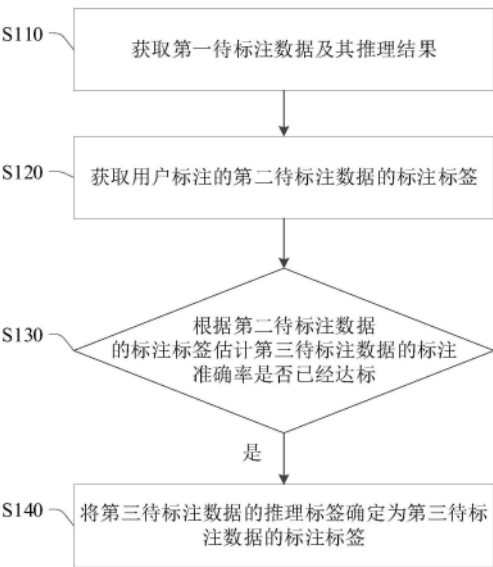
权利要求书5页 说明书25页 附图8页

(54) 发明名称

数据标注方法、计算机程序产品、存储介质
及电子设备

(57) 摘要

本申请实施例提供一种数据标注方法、计算机程序产品、存储介质及电子设备。其中,数据标注方法包括:获取第一待标注数据及其推理结果,第一待标注数据的推理结果为利用标注模型对其进行预测得到的信息;获取用户标注的第二待标注数据的标注标签,第二待标注数据为第一待标注数据中的部分待标注数据;根据第二待标注数据的标注标签估计第一待标注数据中尚未标注的第三待标注数据的标注准确率是否已经达标,第三待标注数据的标注准确率为根据其推理结果所确定的推理标签的准确率;若是,则将第三待标注数据的推理标签确定为第三待标注数据的标注标签。该方法显著提高了数据标注的效率,并且标注准确率也得到了有效保障。



1. 一种数据标注方法,其特征在于,包括:

获取第一待标注数据及其推理结果;其中,所述第一待标注数据的推理结果为利用标注模型对所述第一待标注数据进行预测得到的信息;

获取用户标注的第二待标注数据的标注标签;其中,所述第二待标注数据为所述第一待标注数据中的部分待标注数据;

根据所述第二待标注数据的标注标签估计所述第一待标注数据中尚未标注的第三待标注数据的标注准确率是否已经达标;其中,所述第三待标注数据的标注准确率为根据所述第三待标注数据的推理结果所确定的推理标签的准确率;

若所述第三待标注数据的标注准确率已经达标,则将所述第三待标注数据的推理标签确定为所述第三待标注数据的标注标签。

2. 根据权利要求1所述的数据标注方法,其特征在于,所述根据所述第二待标注数据的标注标签估计所述第一待标注数据中的第三待标注数据的标注准确率是否已经达标,包括:

根据所述第二待标注数据的标签确定所述第一待标注数据中的第四待标注数据;其中,所述第四待标注数据中包含标注正确的第二待标注数据和/或尚未标注过的第三待标注数据,所述第二待标注数据标注正确是指:所述第二待标注数据的标注标签和根据所述第二待标注数据的推理结果所确定的所述第二待标注数据的推理标签相同;

若所述第四待标注数据中包含尚未标注的第三待标注数据,则根据所述第四待标注数据中标注正确的第二待标注数据的数量,估计所述第三待标注数据的标注准确率是否达标。

3. 根据权利要求2所述的数据标注方法,其特征在于,从所述获取用户标注的第二待标注数据的标注标签,直至所述将所述第三待标注数据的推理标签确定为所述第三待标注数据的标注标签的步骤,通过迭代的方式执行,每轮迭代包括以下步骤:

从所述第四待标注数据中抽样出在本轮迭代中提供给用户标注的所述第二待标注数据;其中,对于首轮迭代,所述第四待标注数据为所述第一待标注数据;

获取用户标注的所述第二待标注数据的标注标签,并根据所述第二待标注数据的标注标签对所述第四待标注数据进行筛选;

经过筛选后,若所述第四待标注数据中包含尚未标注的第三待标注数据,则根据所述第四待标注数据中标注正确的第二待标注数据的数量,估计所述第三待标注数据的标注准确率是否达标;

若所述第三待标注数据的标注准确率已经达标,则将所述第三待标注数据的推理标签确定为所述第三待标注数据的标注标签。

4. 根据权利要求3所述的数据标注方法,其特征在于,所述标注模型为分类模型,所述第一待标注数据的推理结果包括所述第一待标注数据分属每个类别的概率;

在执行首轮迭代之前,所述方法还包括:根据所述第一待标注数据的推理结果创建与每个类别对应的标注候选池;其中,每个标注候选池中包含所述第一待标注数据及其推理结果中的、所述第一待标注数据属于该标注候选池对应的类别的概率;

每轮迭代包括以下步骤:

从所有标注候选池中的所述第四待标注数据中抽样出在本轮迭代中提供给用户标注

的所述第二待标注数据;其中,对于首轮迭代,每个标注候选池中的所述第四待标注数据为该标注候选池中的所述第一待标注数据;

获取用户标注的所述第二待标注数据的标注类别;其中,所述第二待标注数据的标注类别为所述第二待标注数据的标注标签;

对于每个标注候选池,执行以下步骤:

针对每条第二待标注数据,若该条第二待标注数据的标注类别和该标注候选池对应的类别不同,且该条第二待标注数据包含在该标注候选池中的所述第四待标注数据中,则将该标注候选池中的所述第四待标注数据缩减为仅包含满足以下条件的待标注数据:其推理结果中属于该标注候选池对应的类别的概率大于该条第二待标注数据的推理结果中属于该标注候选池对应的类别的概率;

在该标注候选池中的所述第四待标注数据经过缩减后,所述第四待标注数据中包含尚未标注的第三待标注数据,则根据所述第四待标注数据中标注正确的第二待标注数据的数量,估计所述第三待标注数据的标注准确率是否达标;

若所述第三待标注数据的标注准确率已经达标,则将该标注候选池对应的类别确定为所述第三待标注数据的标注类别;其中,该标注候选池对应的类别为所述第三待标注数据的推理标签,所述第三待标注数据的标注类别为所述第三待标注数据的标注标签。

5. 根据权利要求4所述的数据标注方法,其特征在于,在每个标注候选池中,所述第一待标注数据按照其推理结果中的、属于该标注候选池对应的类别的概率降序排序;

所述若该条第二待标注类别和该标注候选池对应的类别不同,且该条第二待标注数据包含在该标注候选池中的所述第四待标注数据中,则将该标注候选池中的所述第四待标注数据缩减为仅包含满足以下条件的待标注数据:其推理结果中属于该标注候选池对应的类别的概率大于该条第二待标注数据的推理结果中属于该标注候选池对应的类别的概率,包括:

若该条第二待标注数据的标注类别和该标注候选池对应的类别不同,且该条第二待标注数据的排序索引不大于该标注候选池中的所述第四待标注数据的最大排序索引,则将该标注候选池中的所述第四待标注数据缩减为仅包含满足以下条件的待标注数据:其排序索引小于该条第二待标注数据的排序索引。

6. 根据权利要求4或5所述的数据标注方法,其特征在于,每条第一待标注数据均包括标注准确率属性和标注类别属性,在所述获取用户标注的所述第二待标注数据的标注类别之后,所述方法还包括:

将每条第二待标注数据的标注准确率属性的值设置为第一数值,并将其标注类别属性的值设置为所述标注类别;

对于每个标注候选池,还执行以下步骤:

在该标注候选池中的所述第四待标注数据经过缩减后,若所述第四待标注数据中包含尚未标注的第三待标注数据,则获取所述第三待标注数据的标注准确率;其中,所述标注准确率根据所述第四待标注数据中标注正确的第二待标注数据的数量估计得到;

针对每条第三待标注数据,若该条第三待标注数据的标注准确率属性的值小于所述标注准确率,则将该条第三待标注数据的标注准确率属性的值设置为所述标注准确率,并将其标注类别属性的值设置为该标注候选池对应的类别。

7. 根据权利要求3所述的数据标注方法,其特征在于,所述标注模型为分类模型,所述第一待标注数据的推理结果包括所述第一待标注数据分属每个类别的概率;

每轮迭代包括以下步骤:

从所述第四待标注数据中抽样出在本轮迭代中提供给用户标注的所述第二待标注数据;其中,对于首轮迭代,所述第四待标注数据为所述第一待标注数据;

获取用户标注的所述第二待标注数据的标注类别;其中,所述第二待标注数据的标注类别为所述第二待标注数据的标注标签;

针对每条第二待标注数据,若该条第二待标注数据的标注类别和其推理结果中的最大概率对应的类别不同,且该条第二待标注数据包含在所述第四待标注数据中,则将所述第四待标注数据缩减为仅包含满足以下条件的待标注数据:其推理结果中的最大概率大于该条第二待标注数据的推理结果中的最大概率;

在所述第四待标注数据经过缩减后,若所述第四待标注数据中包含尚未标注的第三待标注数据,则根据所述第四待标注数据中标注正确的第二待标注数据的数量,估计所述第三待标注数据的标注准确率是否达标;

若所述第三待标注数据的标注准确率已经达标,则将所述第三待标注数据的推理结果中的最大概率对应的类别确定为所述第三待标注数据的标注类别;其中,所述第三待标注数据的推理结果中的最大概率对应的类别为所述第三待标注数据的推理标签,所述第三待标注数据的标注类别为所述第三待标注数据的标注标签。

8. 根据权利要求7所述的数据标注方法,其特征在于,所述第一待标注数据按照其推理结果中的、所述第一待标注数据分属每个类别的概率中的最大概率降序排序;

所述若该条第二待标注数据的标注类别和其推理结果中的最大概率对应的类别不同,且该条第二待标注数据包含在所述第四待标注数据中,则将所述第四待标注数据缩减为仅包含满足以下条件的待标注数据:其推理结果中的最大概率大于该条第二待标注数据的推理结果中的最大概率,包括:

若该条第二待标注数据的标注类别和其推理结果中的最大概率对应的类别不同,且该条第二待标注数据的排序索引不大于所述第四待标注数据中的最大排序索引,则将所述第四待标注数据缩减为仅包含满足以下条件的待标注数据:其排序索引小于该条第二待标注数据的排序索引。

9. 根据权利要求1-8中任一项所述的数据标注方法,其特征在于,在所述获取用户标注的第二待标注数据的标注标签之前,所述方法还包括:

根据所述第二待标注数据的推理结果,确定所述第二待标注数据的预测标签;

在用户标注界面上显示所述第二待标注数据及其预测标签。

10. 根据权利要求9所述的数据标注方法,其特征在于,所述标注模型为分类模型,所述第一待标注数据的推理结果包括所述第一待标注数据分属每个类别的概率,通过执行至少一轮迭代对所述第一待标注数据进行标注,在执行首轮迭代之前,根据所述第一待标注数据的推理结果创建与所述每个类别对应的标注候选池,且在每个标注候选池中,所述第一待标注数据按照其推理结果中的、属于该标注候选池对应的类别的概率降序排序;

执行每一轮迭代时,按照以下方式确定任意一条第二待标注数据的预测类别,所述第二待标注数据的预测类别为所述第二待标注数据的预测标签:

获取该条第二待标注数据在每个标注候选池中的排序索引；

将每个排序索引对应的标注候选池中的第四待标注数据的数量进行归一化，得到归一化后的排序索引；

将所有归一化后的排序索引中的最小排序索引对应的类别确定为该条第二待标注数据的预测类别。

11. 根据权利要求9所述的数据标注方法，其特征在于，所述标注模型为分类模型，所述第一待标注数据的推理结果包括所述第一待标注数据分属每个类别的概率，通过执行至少一轮迭代对所述第一待标注数据进行标注；

按照以下方式确定任意一条第二待标注数据的预测类别，所述第二待标注数据的预测类别为所述第二待标注数据的预测标签：

将该条第二待标注数据的推理结果中的最大概率对应的类别确定为该条第二待标注数据的预测类别。

12. 根据权利要求9-11中任一项所述的数据标注方法，其特征在于，所述标注模型为分类模型，所述第二待标注数据的预测标签为所述第二待标注数据的预测类别，所述在用户标注界面上显示所述第二待标注数据及其预测标签，包括：

按照所述第二待标注数据的预测类别，在所述用户标注界面上分类显示所述第二待标注数据。

13. 根据权利要求12所述的数据标注方法，其特征在于，所述第二待标注数据及其预测类别显示在所述用户标注界面的第一区域；

所述方法还包括：

响应在所述第一区域中触发的、针对所述第二待标注数据的数据选择操作，将被选中的第二待标注数据显示在所述用户标注界面的第二区域，并在所述第二区域中显示被选中的第二待标注数据的候选类别；

其中，若一条第二待标注数据被用户选中，则表明该条第二待标注数据的预测类别不被用户认可。

14. 根据权利要求9-13中任一项所述的数据标注方法，其特征在于，所述第二待标注数据为图像，所述第二待标注数据及其预测标签显示在所述用户标注界面的第一区域；

所述方法还包括：

在所述用户标注界面的第三区域中显示参考图像，所述参考图像为从所述第二待标注数据中选择出的一张图像；

记录在所述第三区域中触发的、针对所述参考图像的图像变换操作，将所述图像变换操作应用至所述第二待标注数据中的每张图像，并在所述第一区域中刷新显示执行所述图像变换操作后的第二待标注数据；

其中，所述图像变换操作包括图像平移、图像旋转、图像缩放、框选图像局部区域中的至少一种操作。

15. 根据权利要求1-14中任一项所述的数据标注方法，其特征在于，所述获取第一待标注数据及其推理结果，包括：

从训练数据池中抽样出至少零批次训练数据对所述标注模型进行训练，以及，利用所述标注模型对从推理数据候选池中抽样出的至少一批次的待标注数据进行推理，得到所述

至少一批次的待标注数据的推理结果；

将所述至少一批次的待标注数据及其推理结果写入推理结果池；

从所述推理结果池中读取所述第一待标注数据及其推理结果。

16. 一种计算机程序产品,其特征在于,包括计算机程序指令,所述计算机程序指令被处理器读取并运行时,执行如权利要求1-15中任一项所述的方法。

17. 一种计算机可读存储介质,其特征在于,所述计算机可读存储介质上存储有计算机程序指令,所述计算机程序指令被处理器读取并运行时,执行如权利要求1-15中任一项所述的方法。

18. 一种电子设备,其特征在于,包括:存储器以及处理器,所述存储器中存储有计算机程序指令,所述计算机程序指令被所述处理器读取并运行时,执行权利要求1-15中任一项所述的方法。

数据标注方法、计算机程序产品、存储介质及电子设备

技术领域

[0001] 本申请涉及人工智能技术领域,具体而言,涉及一种数据标注方法、计算机程序产品、存储介质及电子设备。

背景技术

[0002] 数据标注是人工智能领域的一项重要任务,所谓数据标注,即为数据打上可靠的标签信息。例如,在图像分类任务中,数据标注可以指为训练图像标出其真实类别,从而可以以此真实类别为监督信号,对图像分类模型进行有监督训练,使得模型能够学习到训练图像的内容和其真实类别之间的对应关系,进而可以对其他图像进行有效分类。

[0003] 然而,目前的数据标注基本上是完全依赖人工完成的,其效率十分低下。

发明内容

[0004] 本申请实施例的目的在于提供一种数据标注方法、计算机程序产品、存储介质及电子设备,以改善上述技术问题。

[0005] 为实现上述目的,本申请提供如下技术方案:

[0006] 第一方面,本申请实施例提供一种数据标注方法,包括:获取第一待标注数据及其推理结果;其中,所述第一待标注数据的推理结果为利用标注模型对所述第一待标注数据进行预测得到的信息;获取用户标注的第二待标注数据的标注标签;其中,所述第二待标注数据为所述第一待标注数据中的部分待标注数据;根据所述第二待标注数据的标注标签估计所述第一待标注数据中尚未标注的第三待标注数据的标注准确率是否已经达标;其中,所述第三待标注数据的标注准确率为根据所述第三待标注数据的推理结果所确定的推理标签的准确率;若所述第三待标注数据的标注准确率已经达标,则将所述第三待标注数据的推理标签确定为所述第三待标注数据的标注标签。

[0007] 上述方法只需对第一待标注数据中的第二待标注数据进行人工标注,基于其标注结果,就可以对第一待标注数据中的尚未标注的第三待标注数据进行标注准确率评估,并在其准确率达标时对第三待标注数据进行自动标注,而无需对所有待标注数据都进行人工标注。从而,该方法显著提高了数据标注的效率,并且标注准确率也得到了有效保障。其中,第三待标注数据的标注准确率达到预期的标准(即达标),而第二待标注数据由于是人工标注的,在用户处于正常状态时,其标注准确率可以认为接近或等于100%。

[0008] 第二方面,本申请实施例提供一种计算机程序产品,包括计算机程序指令,所述计算机程序指令被处理器读取并运行时,执行第一方面或第一方面的任意一种可能的实现方式提供的方法。

[0009] 第三方面,本申请实施例提供一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机程序指令,所述计算机程序指令被处理器读取并运行时,执行第一方面或第一方面的任意一种可能的实现方式提供的方法。

[0010] 第四方面,本申请实施例提供一种电子设备,包括:存储器以及处理器,所述存储

器中存储有计算机程序指令,所述计算机程序指令被所述处理器读取并运行时,执行第一方面或第一方面的任意一种可能的实现方式提供的方法。

附图说明

[0011] 为了更清楚地说明本申请实施例的技术方案,下面将对本申请实施例中所需要使用的附图作简单地介绍,应当理解,以下附图仅示出了本申请的某些实施例,因此不应被看作是对范围的限定,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他相关的附图。

[0012] 图1示出了本申请实施例提供的标注方法的流程;

[0013] 图2(A)和图2(B)示出了用户标注界面的一种实现方式;

[0014] 图3(A)和图3(B)示出了用户标注界面的另一种实现方式;

[0015] 图4示出了图1中方法的一种迭代实现方式的流程;

[0016] 图5示出了第四待标注数据的缩减原理;

[0017] 图6示出了本申请实施例提供的标注系统的架构及其工作原理;

[0018] 图7示出了图6中的系统进行模型训练和推理的一种实现方式;

[0019] 图8示出了本申请实施例提供的标注装置包含的功能模块;

[0020] 图9示出了本申请实施例提供的电子设备的结构。

具体实施方式

[0021] 近年来,基于人工智能的计算机视觉、深度学习、机器学习、图像处理、图像识别等技术研究取得了重要进展。人工智能(Artificial Intelligence,简称AI)是研究、开发用于模拟、延伸人的智能的理论、方法、技术及应用系统的新兴科学技术。人工智能学科是一门综合性学科,涉及芯片、大数据、云计算、物联网、分布式存储、深度学习、机器学习、神经网络等诸多技术种类。计算机视觉作为人工智能的一个重要分支,具体是让机器识别世界,计算机视觉技术通常包括人脸识别、活体检测、指纹识别与防伪验证、生物特征识别、人脸检测、行人检测、目标检测、行人识别、图像处理、图像识别、图像语义理解、图像检索、文字识别、视频处理、视频内容识别、三维重建、虚拟现实、增强现实、同步定位与地图构建、计算摄影、机器人导航与定位等技术。随着人工智能技术的研究和进步,该项技术在众多领域展开了应用,例如安全防控、城市管理、交通管理、楼宇管理、园区管理、人脸通行、人脸考勤、物流管理、仓储管理、机器人、智能营销、计算摄影、手机影像、云服务、智能家居、穿戴设备、无人驾驶、自动驾驶、智能医疗、人脸支付、人脸解锁、指纹解锁、人证核验、智慧屏、智能电视、摄像机、移动互联网、网络直播、美颜、美妆、医疗美容、智能测温等领域。本申请实施例中的数据标注方法也属于人工智能技术的范畴。

[0022] 在介绍本申请实施例中的技术方案之前,先简单说明数据标注的含义以及进行数据标注的意义。

[0023] 所谓数据标注,即通过某种手段(例如,人工确认)为数据打上标签的过程,而标签可以理解为数据所具有的某项(或某几项)属性的值,如果标注效果是理想的,那么标签应为该项属性的真实值。

[0024] 在人工智能领域,数据标注往往针对某个模型(比如,神经网络模型)的训练样本

进行,被标注好的训练样本可用于模型的有监督训练,训练好的模型可用于执行特定的任务,为后文方便表述,不妨将其称为任务模型。这些任务可以概括为:将任务数据输入任务模型,经任务模型运算后,预测出任务数据的某项(或某几项)属性的值,预测出的值在具体问题中可能具有实际的物理含义。

[0025] 例如,一种可能的训练过程为:将训练样本输入任务模型,得到模型输出的、针对训练样本的某项属性的预测结果,根据预测结果和训练样本的标签,代入预设的损失函数计算预测损失,该预测损失表征预测结果和训练样本的标签之间的差异,迭代地优化任务模型的参数,直至该损失收敛。

[0026] 若训练样本的标签代表了训练样本的某项属性的真实值,那么训练过程也可以视为通过调整任务模型的参数,使其输出的、针对训练样本的该项属性预测结果和其真实值越来越接近的过程。从而,任务模型在训练好以后,再处理和训练样本同类型的任务数据,也能够对任务数据的该项属性预测出比较合理的结果,即能够较好地完成特定的任务。

[0027] 例如,任务模型可以是图像分类模型,图像分类模型用于预测图像中特定对象的类别,比如,预测图像中的手写数字是0~9中的哪个数字,此时的训练样本可以是预先采集的手写数字图像,训练样本的标签为预先标注的图像中的真实数字。

[0028] 又例如,任务模型可以是目标检测模型,目标检测模型用于检测图像中特定对象的位置,比如,检测图像中车辆的位置,此时的训练样本可以是预先采集的道路图像,训练样本的标签为预先标注的图像中车辆的真实位置。

[0029] 又例如,任务模型可以是词性分析模型,词性分析模型用于预测语句中每个单词的词性,比如,名词、动词、介词等,此时的训练样本可以是预先收集的语句,训练样本的标签为预先标注的语句中单词的真实词性。

[0030] 总之,本申请的方案中并不限定要标注的数据的具体形式,例如可以是图像、文字、语音等;也不限定标签的具体形式,这与任务模型要执行的具体任务有关,例如可以是图像中对象的真实类别、真实位置等。在后文中,为简单起见,主要以待标注数据是图像的情况为例介绍本申请提出的数据标注方法,但这不应视为对本申请保护范围的限制。

[0031] 进一步的,由于在人工智能领域,通常是将人的认知结果作为正确结果,任务模型只是对人在某个特定任务上的认知过程的模拟,因此目前的数据标注通常是通过人工方式进行的,将人工标注的标签作为数据中属性的真实值。然而,任务模型的训练很可能需要大量的带有标签的训练样本,导致标注人员的工作量极大,标注效率十分低下。

[0032] 本申请提出的数据标注方法旨在将人工标注和机器自动标注的方式结合起来,从而显著提升数据标注的效率,并且确保标注结果的准确率能够满足任务要求。

[0033] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行描述。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定义和解释。

[0034] 术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0035] 术语“第一”、“第二”等仅用于将一个实体或者操作与另一个实体或操作区分开来,而不能理解为指示或暗示相对重要性,也不能理解为要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。

[0036] 图1示出了本申请实施例中的数据标注方法的流程,该方法可以、但不限于在图9示出的电子设备上执行,该电子设备的具体结构可参考后文关于图9的阐述。参照图1,数据标注方法包括:

[0037] 步骤S110:获取第一待标注数据及其推理结果。

[0038] 第一待标注数据为一批待标注的数据,例如,第一待标注数据可以是尚未标注的训练样本,在标注好后则可以作为任务模型的正式训练样本。第一待标注数据的推理结果为利用标注模型对第一待标注数据进行预测得到的信息,或者说将第一待标注数据输入标注模型后,得到的模型输出就是第一待标注数据的推理结果。例如,若标注模型为分类模型,第一待标注数据的推理结果可以是模型预测出的第一待标注数据分属每个类别的概率,若标注模型为回归模型,第一待标注数据的推理结果可以是模型预测出的回归数值。

[0039] 标注模型是数据标注阶段使用的模型,标注模型可以和前文提到的任务模型具有相同形式的输入和输出。例如,任务模型是图像分类模型,其输入为图像,输出为该图像分属每个类别的概率(比如,对于手写数字识别,就是该图像分属数字0~9的概率,默认图像中只有一个手写数字),那么标注模型也可以是图像分类模型,其输入为图像,输出为该图像分属每个类别的概率。又例如,任务模型是目标检测模型,其输入为图像,输出为该图像中每个目标的位置(比如,对于车辆检测,就是该图像中每辆车的位置),那么标注模型也可以是目标检测模型,其输入为图像,输出为该图像中每个目标的位置,等等。

[0040] 但标注模型和任务模型可能、但不一定是相同的模型,即二者有可能结构相同但参数不同,或者干脆结构也不同。导致二者被设计为不同的模型的因素有很多:

[0041] 例如,在某场景中,希望尽可能地给待标注数据赋予准确的标签,即在进行数据标注时无需过多考虑资源消耗问题,此时,可以采用一个较复杂的模型作为标注模型;但对于任务模型,很可能是要部署到实际的生产环境中的(比如,手机、摄像头上),其可以使用的资源是严格受限的,只能采用一个较简单的模型。

[0042] 又例如,根据后文内容可知,在数据标注方法的某些实现方式中,标注模型只会使用人工标注的数据进行训练,而不会使用机器自动标注的数据进行训练,因此直接将其用于执行任务并不会取得太好的效果,因为它“见过”的训练数据十分有限,而任务模型在训练时则可能会使用人工标注的数据和机器自动标注的数据,因此其执行任务的效果更好,因为它“见过”的训练数据更多。关于标注模型如何训练,这里暂不展开阐述。

[0043] 为简单起见,在后文所举的例子中,可以认为标注模型和任务模型是同类的模型(比如,都是图像分类模型),且具有相同形式的输入和输出,但这不应当视为对本申请保护范围的限制。

[0044] 步骤S110中的“获取”泛指通过各种手段得到:例如,可以是在执行步骤S110之前已经得到了第一待标注数据及其推理结果并存储在某处,在步骤S110执行时直接读取第一待标注数据及其推理结果;又例如,也可能在执行步骤S110之前还没有得到第一待标注数据及其推理结果,在执行步骤S110时才从某处(比如,后文提到推理结果池)得到第一待标注数据,并利用标注模型计算出第一待标注数据的其推理结果,等等。

[0045] 步骤S120:获取用户标注的第二待标注数据的标注标签。

[0046] 第二待标注数据为第一待标注数据中的部分待标注数据。第二待标注数据可以从第一待标注数据中抽样得到,抽样方法不限,例如可以是随机抽样,也可以是按照固定模式抽样,等等。第二待标注数据中可以不存在重复的数据,避免进行重复标注。

[0047] 获得第二待标注数据后,可将其提供给用户进行人工标注,此处的用户就是指标注人员。将第二待标注数据提供给用户的方式有多种:例如,若用户进行数据标注所使用的设备(后文称用户终端)和执行步骤S120的设备(后文称当前设备)并非同一设备,则可以将第二待标注数据从当前设备发送给用户终端,并在用户终端的用户标注界面(比如,一个网页或者客户端界面)上进行显示,供用户标注;又例如,若用户终端和当前设备是同一设备,则可以直接将第二待标注数据在当前设备的用户标注界面上进行显示,供用户标注,等等。

[0048] 应当理解,虽然用户必须以某种方式获知第二待标注数据的内容才能对其进行标注,但获知方式并不限于将其在用户标注界面显示,例如,还可以用语音播放第二待标注数据的内容(如果其可以语音播放的话),或者将第二待标注数据打印出来给用户观看等等。

[0049] 用户对第二待标注数据进行标注,就是人工为其打上标签,将此标签称为标注标签,若默认人工标注是准确的,则标注标签也就是第二待标注数据最终的标签。例如,若第一待标注数据是用于训练分类模型的(比如,任务模型和标注模型都是分类模型),则用户可将自己认为的第一待标注数据的类别作为其标注标签;又例如,若第一待标注数据是用于训练回归模型的(比如,任务模型和标注模型都是回归模型),则用户可将自己认为的第一待标注数据的回归数值作为其标注标签。在介绍图2(A)、图2(B)、图3(A)和图3(B)时,还会对用户的标注过程作更多阐述,此处暂不展开。

[0050] 用户标注完成后,可以自行触发代表标注完成的确认操作(例如,在用户标注界面上点击“提交”),将标注结果,即第二待标注数据的标注标签返回给当前设备(若用户本来就在当前设备上标注则可以认为是返回给了执行数据标注流程的主程序),从而当前设备获得了第二待标注数据的标注标签。在执行完步骤S120后,第二待标注数据就标注完了。

[0051] 步骤S130:根据第二待标注数据的标注标签估计第三待标注数据的标注准确率是否已经达标。

[0052] 第三待标注数据为第一待标注数据中尚未标注的数据,由于第二待标注数据只是第一待标注数据中的部分待标注数据,所以在第二待标注数据被标注后,第一待标注数据中必然还有尚未标注的数据,从而可能会产生第三待标注数据。但需注意,第一待标注数据中所有尚未标注的数据可能、但不一定都属于第三待标注数据,这里存在不同的实现方式。例如,根据后文可知,在某些实现方式,一些被评估为标注准确率较低的未标注数据会被排除在第三待标注数据之外。

[0053] 第三待标注数据的标注准确率为根据第三待标注数据的推理结果所确定的推理标签的准确率。其中,由于第三待标注数据属于第一待标注数据,而步骤S110中已经获得了第一待标注数据的推理结果,所以第三待标注数据的推理结果在执行步骤S130时是已知的。

[0054] 第三待标注数据的推理标签是基于第三待标注数据的推理结果,并结合特定的规则得到的标签,推理标签在形式上和前面介绍的标注标签并无区别,只是推理标签是机器自动给出的标签,而并非是人工标注的标签,因此其准确率通常不是100%(默认人工标注

的标签是正确的),该标注准确率可以根据第二待标注数据的标注标签进行估计,其原理是:第二待标注数据也具有推理标签,且其得到推理标签的规则和第三待标注数据是相同的,从而第二待标注数据的推理标签和标注标签(正确的标注结果)的差异就可以反映出按照此种规则算出的推理标签的准确率到底如何。得到推理标签的规则有多种,后文会给出例子,此处暂不展开。

[0055] 需要指出,以上引出推理标签的概念仅仅是为了解释第三待标注数据的标注准确率的含义,并不代表在估计该标注准确率时一定要将第三待标注数据的推理标签计算出来,也不代表在估计该标注准确率时一定要给第三待标注数据打上推理标签,而可以这样理解该标注准确率的定义:若第三待标注数据按照其推理标签进行标注,则标注结果的准确率符合估计出的准确率。总之,第三待标注数据具有推理标签和其处于尚未标注的状态并不矛盾。

[0056] 第三待标注数据的准确率达标可以指其标注准确率估值大于目标准确率。其中,目标准确率可以是开始数据标注之前,根据任务模型的训练需求设定好的一个准确率,例如可以是99%、98%等。若估计出第三待标注数据的准确率达标,则执行步骤S140,若估计出第三待标注数据的准确率不达标,则可能采取不同的处理方式:例如,抽样出更多的第二待标注数据,并执行步骤S120~S130,重新估计第三待标注数据的标注准确率,当然此时的第三待标注数据可能会发生变化;又例如,重新训练标注模型,并执行步骤S110~S130,重新估计第三待标注数据的标注准确率;又例如,直接对第三待标注数据进行人工标注,等等。

[0057] 步骤S140:将第三待标注数据的推理标签确定为第三待标注数据的标注标签。

[0058] 既然第三待标注数据的标注准确率能够满足需求,那么直接按照推理标签对第三待标注数据进行标注就可以了,第三待标注数据的标注标签也就是其最终的标签。如果在之前未计算过第三待标注数据的推理标签,则在步骤S140中需要先计算出第三待标注数据的推理标签。由于第三待标注数据是自动标注的,所以其标注效率远高于人工标注,并且其标注准确率也是有保障的,例如,足以满足任务模型的训练需求。

[0059] 在执行完步骤S140后,第一待标注数据中的第二待标注数据和第三待标注数据都已经标注完成,但不排除第一待标注数据中还剩余有数据尚未标注,对于这些数据,可以采取不同的处理方式:例如,直接由人工进行标注;又例如,重新训练标注模型后再进行标注(此时推理结果可能发生变化,从而改变标注结果),等等。

[0060] 简单总结图1中的方法,该方法只需对第一待标注数据中的第二待标注数据进行人工标注,基于其标注结果,就可以对第一待标注数据中的尚未标注的第三待标注数据进行标注准确率评估,并在其准确率达标时对第三待标注数据进行自动标注,而无需对所有待标注数据都进行人工标注。从而,该方法显著提高了数据标注的效率,并且标注准确率也得到了有效保障,比如,可以满足任务模型的训练需求。其中,第三待标注数据的标注准确率达到预期的标准,而第二待标注数据由于是人工标注的,在用户处于正常状态时,其标注准确率可以认为接近或等于100%。

[0061] 注意,发明人研究发现,待标注数据往往数量庞大,第一待标注数据很可能只是其中的一批数据,即使并非每一批第一待标注数据都能触发步骤S140的执行,还是有很多批的第一待标注数据能够触发步骤S140的执行,即自动标注几乎总是会执行的,上述有益效

果并非只在少数特殊情况下才能实现。

[0062] 另外,需要指出,虽然上面在阐述时,总是将数据标注和任务模型的训练联系在一起,但实际上二者并不具有绑定关系,即标注好的数据(第二待标注数据和第三待标注数据)并不一定用于训练任务模型,例如也有可能仅仅是用于展示、存储目的,即其用途并不限定。

[0063] 前文在阐述步骤S120时提到,为了获得第二待标注数据的标注标签,可以先将第二待标注数据在用户标注界面上进行显示,由用户进行标注。由于第二待标注数据此前并未经过标注,所以在一种实现方式中,用户标注界面上可以显示标签未知的第二待标注数据,完全由用户给出其标注标签,此种实现方式比较简单,但用户的工作量比较大。

[0064] 因此,在另一种实现方式中,可以先根据第二待标注数据的推理结果,确定第二待标注数据的预测标签,在用户标注界面上则不仅要显示第二待标注数据,还要显示其预测标签。

[0065] 其中,预测标签是基于第二待标注数据的推理结果,并结合特定的规则得到的标签,预测标签在形式上和前面介绍的标注标签并无区别,只是预测标签并非是人工标注的标签。得到预测标签的特定规则有多种,后文会给出例子,此处暂不展开,但需要注意,计算预测标签的特定规则和计算推理标签的特定规则可能相同也可能不同,从而同一条第二待标注数据预测标签和推理标签可能相同也可能不同。

[0066] 预测标签可以表示根据标注模型的推理结果所预测出的标注结果,若标注模型的推理结果足够精确,则预测标签可能和标注标签十分接近,从而预测标签可以作为用户进行人工标注时的重要参考信息,使得用户可以无需修改用户标注界面上显示的预测标签,直接将其确认为标注标签,或者,只需少量修改用户标注界面上显示的预测标签(修改为用户认为合适的标注标签),将大部分预测标签都确认为标注标签,从而显著提高了用户的标注效率。

[0067] 可选的,若标注模型是分类模型,则第二待标注数据的预测标签为第二待标注数据的预测类别,此时可以按照第二待标注数据的预测类别,在用户标注界面上分类显示第二待标注数据。

[0068] 参照图2(A),标注模型是情绪分类模型,第二待标注数据为人脸图像,第二待标注数据的推理结果可以是图像中的人脸分属每种情绪类别(例如,高兴、难过、惊讶等)的概率,第二待标注数据的预测类别可以是图像所属的情绪类别,例如高兴、难过、惊讶等。

[0069] 在图2(A)中,预测类别为高兴类别的人脸图像显示在一行,预测类别为难过类别的人脸图像显示在另一行,以此类推。对人脸图像分类显示后,用户很容易确认每张人脸图像所属的情绪类别,并且也更容易从中挑选出预测类别错误的图像并更改其类别,因为在大量具有共性的同类数据中,少量的异类数据会更加明显。例如,对于高兴类别中的第二张人脸图像,其嘴角弯曲方向和其他人脸图像明显不同,用户可以很快确认出其预测类别存在错误。

[0070] 进一步的,第二待标注数据及其预测类别可以显示在用户标注界面的第一区域,用户可以在第一区域中执行数据选择操作,若一条第二待标注数据被用户选中,则表明该条第二待标注数据的预测类别不被用户认可。

[0071] 例如,参照图2(B),用户标注界面的上半部分为第一区域,若用户使用PC机进行标

注,则数据选择操作可以是鼠标左键的单击操作,假设用户认为高兴类别中的第二张人脸图像的预测类别存在错误,则可以将屏幕上的光标移动至此人脸图像上然后点击鼠标左键,此时该人脸图像下方会出现未知类别的字样,表示该图像的类别待定。

[0072] 响应于在第一区域中触发的数据选择操作,用户标注界面可以将被选中的第二待标注数据显示界面中的第二区域内,并在第二区域中显示被选中的第二待标注数据的候选类别,供用户进行人工标注时选择。

[0073] 例如,参照图2(B),用户标注界面的下半部分为第二区域,高兴类别中的第二张人脸图像被显示在此处,并且其右侧显示出了所有可能的情绪类别供用户选择为标注标签。特别地,未知类别也可以作为一个候选类别,若用户选择了未知类别,表示该人脸图像的标注标签是未知的,或者说也可以视为其未进行过标注。

[0074] 应当理解,第一区域和第二区域在用户标注界面中的布局方式不限于图2(A)和图2(B)示出的方式,甚至第一区域和第二区域也可以存在重叠。

[0075] 按照上述方式进行标注的有益效果在于:在第一区域中,用户只需关注某条第二待标注数据的预测类别是否正确,即只需进行判断(判断“是”或“否”),而无需进行选择(选择正确的标注标签),在第二区域中,用户则只需进行选择无需进行判断。从而,用户可以先在第一区域中选择出预测类别错误的所有第二待标注数据,然后在第二区域中统一进行类别修改(改为标注标签中的类别),即在每个区域内,用户只需按照单一的标注逻辑进行操作,无需在两种思维逻辑之间来回切换,实验表明,此种方式能显著提升用户的标注速度。

[0076] 应当理解,在标注界面上划分第一区域和第二区域不是必须的,例如,也可以只设置第一区域,用户在第一区域中每选中一条第二待标注数据后,就直接在其显示位置处进行类别的更改,但此时用户需要在判断和选择两种思维逻辑之间来回切换,标注效率可能会降低。

[0077] 另外,即使标注模型是分类模型,在用户标注界面上也不一定要按照预测类别分类显示第二待标注数据,例如也可以将所有类别的第二待标注数据混合显示,在每条第二待标注数据的显示位置处也显示其预测类别即可。

[0078] 进一步的,发明人研究发现,用户在进行标注时,很可能会从第二待标注数据的局部来判断其预测标签是否正确,或者来决定其标注标签到底是什么。例如,在图2(B)中,用户可能主要是从人脸图像的嘴角弯曲方向来判断情绪类别的(上弯为高兴,下弯为难过),对人脸图像的其余部分则并不太关注。因此,若用户标注界面在显示第二待标注数据时,能够应用户的需求,将显示内容聚焦到用户关注的重点部分(即用户注意力集中的位置),则更有利于用户快速、准确地给出标注标签。

[0079] 对于第二待标注数据是图像的情况,基于用户注意力进行数据显示的可能做法如下:

[0080] 在用户标注界面的第一区域中显示第二待标注数据及其预测标签,在用户标注界面的第三区域中显示参考图像。其中,参考图像为从当前要标注的所有第二待标注数据中选择出的一张图像,可以由用户选择,也可以由程序自动选择(比如,自动选择排在第一张图像)。

[0081] 用户可以在第三区域中执行针对参考图像的图像变换操作,其目的可以是将参考图像中与标注结果密切相关的用户注意力集中区域以较显著的方式显示出来。其中,图像

变换操作包括图像平移、图像旋转、图像缩放、框选图像局部区域中的至少一种操作。

[0082] 响应于在第三区域中触发的图像变换操作,用户标注界面(或者后台)可以将图像变换操作记录下来。由于每种操作都对应特定的操作参数,因此记录操作实际上就是记录这些操作参数。例如,对于图像平移可以记录平移量,对于图像旋转可以记录旋转角度,对于图像缩放可以记录缩放倍数,对于框选图像局部区域可以记录局部区域的位置,等等。

[0083] 最后,可以将记录的图像变换操作应用到第二待标注数据中的每张图像上,并在第一区域中刷新显示执行图像变换操作后的第二待标注数据。从而,第二待标注数据中的每张图像也和参考图像一样,图像中与用户注意力集中区域以较显著的方式被显示出来,便于用户进行观察并标注。

[0084] 例如,参照图3(A),用户标注界面的左侧部分为第三区域,右侧部分的上半为第一区域,右侧部分的下半则为第二区域。第三区域中有一个固定尺寸的窗口(灰色),高兴类别中的第一张人脸图像被选为参考图像后,显示在该窗口中。

[0085] 之后,用户可以在该固定窗口中进行图像的平移、旋转、缩放、框选局部区域中的至少一种操作,其操作参数将被记录,变换过程中超出固定窗口尺寸的部分不显示。参照图3(B),用户进行了平移操作,将图像中的人脸移动到窗口正中,此时可以记录下图像中心相对于窗口中心的偏移量作为平移参数;用户还进行了放大操作,使图像中的人脸充满整个窗口,此时可以记录下图像的放大倍数作为缩放参数;用户还框选了图像在标注时需要重点关注的嘴部区域,如粗虚线方框所示,此时可以记录下嘴部区域的位置作为框选局部区域的参数。

[0086] 最后,根据记录的平移参数和缩放参数可以生成一个变换矩阵,对第二待标注数据中的每张人脸图像应用该变换矩阵,就可以将人脸移动到画面正中并放大。然后,再根据记录的嘴部区域的位置从各人脸中框选出同位置的区域,并进一步放大框选部分直至充满第二待标注图像的原始显示区域(也可以是一个固定尺寸的窗口),最终的刷新显示效果如图3(B)所示。可见,在第一区域中,人脸图像已经被其嘴部的局部放大图像所替代,根据这些放大图像,用户可以更清楚地看到其中的嘴角弯曲方向,从而可以快速选择出错误的预测标签,或者给出正确的标注标签。

[0087] 注意,第一,图像变换操作在应用到第二待标注数据中的每张图像上时,并不一定和其应用到参考图像上的效果完全相同。例如,在图3(B)中,对于框选图像局部区域这种操作,在应用到参考图像上时,只是给出嘴部区域的位置,而在应用到第二待标注数据中的每张图像上时,则不仅要确定出嘴部区域的位置,还要对嘴部区域进行放大显示。

[0088] 第二,第一区域和第三区域在用户标注界面中的布局方式不限于图3(A)和图3(B)示出的方式,甚至第一区域和第三区域也可以存在重叠。

[0089] 第三,虽然图3(A)和图3(B)中示出了第一区域、第二区域和第三区域,但在上面基于用户注意力进行数据显示的方案中,是否要设置第二区域是可选的。

[0090] 第四,上面基于用户注意力进行数据显示的方案与标注模型是否为分类模型无关,例如,若标注模型是目标检测模型,也可以应用该方案,方便用户更精确地标注目标的位置。

[0091] 下面在以上实施例的基础上,继续介绍步骤S130中关于第三待标注数据的标注准确率是否达标的估计方法。在一种实现方式中,步骤S130可以按照以下流程执行:

[0092] 根据第二待标注数据的标签,结合特定的规则从第一待标注数据中确定出第四待标注数据;若第四待标注数据中包含尚未标注的第三待标注数据,则根据第四待标注数据中标注正确的第二待标注数据的数量,估计第三待标注数据的标注准确率是否达标。

[0093] 其中,第四待标注数据中可能包含标注正确的第二待标注数据也可能尚未标注过的第三待标注数据,或者同时包含这两种待标注数据,具体包含哪种数据与上述特定规则的设计相关。第二待标注数据标注正确是指:第二待标注数据的标注标签和根据第二待标注数据的推理结果所确定的第二待标注数据的推理标签相同,关于推理标签和标注标签的含义在前文已经阐述,不再重复。

[0094] 不妨先考虑第四待标注数据中包含标注正确的第二待标注数据和尚未标注过的第三待标注数据的情况,此时的第四待标注数据相当于从第一待标注数据中划定了一个范围,该范围内的部分待标注数据的推理标签已经被人工确认为是正确的,由于每条待标注数据的推理标签的确定规则是统一的,所以根据该范围内标注正确的数据的数量,可以估计剩余待标注数据的推理标签的准确率。

[0095] 若第四待标注数据中仅包含尚未标注过的第三待标注数据,则等价于第四待标注数据中标注正确的第二待标注数据的数量为0的情况,可以和标注正确的第二待标注数据的数量不为0的情况进行统一处理。若第四待标注数据中仅包含标注正确的第二待标注数据,则说明此时这批第四待标注数据已经标注完成,没有需要自动标注的数据。应当理解,确定第四待标注数据的规则不应当使得第四待标注数据中总是仅包含尚未标注过的第三待标注数据,因为这样标注准确率必然难以达标,也不应当总是仅包含标注正确的第二待标注数据,因为这样相当于总是不进行自动标注。

[0096] 进一步的,在步骤S130采用上面的实现方式时,图1中的方法还可以实现为图4中的迭代流程:

[0097] 步骤S210:获取第一待标注数据及其推理结果。

[0098] 步骤S210类似步骤S110,不再重复阐述。注意,虽然图4从整体来看是一种迭代流程,但步骤S210并不参与迭代。步骤S220~S260为步骤S120~S140的一种迭代实现方式,或者说每一轮迭代在逻辑上都包括步骤S220~S260(当然,并不一定每一轮迭代步骤S220~S260都会全部执行)。

[0099] 步骤S220:从第四待标注数据中抽样出在本轮迭代中提供给用户标注的第二待标注数据。

[0100] 其中,步骤S220中的第四待标注数据既可以认为是上一轮迭代结束时的第四待标注数据,也可以认为是本轮迭代开始时的第四待标注数据,在每一轮迭代中,第四待标注数据可能、但不一定会更新(见步骤S230)。对于首轮迭代,由于不存在上一轮,因此首轮迭代开始时的第四待标注数据为第一待标注数据。

[0101] 从第四待标注数据中抽样第二待标注数据的方法不限,每一轮迭代时第二待标注数据都要重新抽样,并且可选的,对于那些在之前轮次中已经被抽样过的第二待标注数据,由于其在之前的迭代轮次中已经被用户标注过,可以不再重复抽样,即可以限定每次抽样的第二待标注数据,都是从上一轮迭代后的第四待标注数据中的、尚未被标注的第三待标注数据中抽样,从而可以避免重复对数据进行标注。

[0102] 步骤S230:获取用户标注的第二待标注数据的标注标签,并根据第二待标注数据

的标注标签对第四待标注数据进行筛选。

[0103] 抽样出的第二待标注数据可以送给用户进行标注,并获得其标注标签,具体实现方式可以参考前文,不再重复阐述。

[0104] 若第四待标注数据仅包括标注正确的第二待标注数据和/或尚未标注过的第三待标注数据,则在本轮中的第二待标注数据被标注完后,目前的第四待标注数据可能已经无法满足该要求,所以需要按照特定的规则对第四待标注数据进行筛选,得到新的第四待标注数据,新的第四待标注数据仅包括标注正确的第二待标注数据和/或尚未标注过的第三待标注数据,从而可以对其中第三待标注数据的标注准确率继续进行有效评估。

[0105] 至于被筛选掉的待标注数据,其中包括标注错误的第二待标注数据,还可能包括某些尚未标注的待标注数据,其处理方式不限,例如,标注错误的第二待标注数据仅仅是说其推理标签存在错误,但其标注标签并无错误(默认人工标注是准确的),因此可以认为其已经标注完成,而对于尚未标注的待标注数据,则可以在标注模型更新后重新进行推理、重新标注,等等。

[0106] 注意,在步骤S230之后的步骤中,凡是提到第四待标注数据,均指筛选后的得到的新的第四待标注数据,为简单起见,仍然简称为第四待标注数据。

[0107] 步骤S240:判断第四待标注数据中是否包含尚未标注的第三待标注数据。

[0108] 若第四待标注数据中包含尚未标注的第三待标注数据,则继续执行步骤S250。若第四待标注数据中不包含尚未标注的第三待标注数据,即第四待标注数据中仅包含标注正确的第二待标注数据,此时也无法进行自动标注,从而可以结束对目前这批第四待标注数据的标注。又由于第一待标注数据中除第四待标注数据以外的部分都是因为步骤S230中的筛选操作产生的,若对这部分数据暂时不再进行标注,此时也可以结束对目前这批第一待标注数据的标注。

[0109] 步骤S250:根据第四待标注数据中标注正确的第二待标注数据的数量,估计第三待标注数据的标注准确率是否达标。

[0110] 若第三待标注数据的标注准确率已经达标,则继续执行步骤S260,若第三待标注数据的标注准确率尚未达标,则可以开始下一轮迭代,即跳转到步骤S220继续执行。注意,步骤S250中标注正确的第二待标注数据,并不限于本轮迭代中标注的第二待标注数据(即不限于步骤S220中抽样出的第二待标注数据),历次迭代中标注的第二待标注数据均应纳入考虑。

[0111] 步骤S260:将第三待标注数据的推理标签确定为第三待标注数据的标注标签。

[0112] 步骤S260类似步骤S140,不再重复阐述。步骤S260执行完后,这批第四待标注数据的标注也就结束了,并且,若暂时不再对第一待标注数据中除第四待标注数据以外的部分进行标注,此时也可以结束对目前这批第一待标注数据的标注。

[0113] 注意,虽然步骤S260名义上属于迭代步骤(例如,该步骤可以包含在迭代部分的代码中),但实际上在整个迭代过程中,该步骤至多只在满足条件时被执行一次,执行完后就会跳出迭代。

[0114] 上述迭代流程的关键在于不断对第四待标注数据进行筛选,使得第四待标注数据中的第三待标注的标注准确率估值越来越高,直至达标(当然,也有可能最终无法达标)。由于图4中的流程仅仅是图1中方法的一种实现方式,因此在阐述图4中的步骤时,未提及之

处,均可以参考图1中相关步骤的阐述。

[0115] 下面,以标注模型是分类模型的情况为例,举例说明迭代步骤S220~S260的两种实现方式。此时,步骤S210中第一待标注数据的推理结果包括第一待标注数据分属每个类别的概率。

[0116] 方式1

[0117] 在执行首轮迭代之前,先根据第一待标注数据的推理结果创建与每个类别对应的标注候选池,一个标注候选池可以理解为一个数据集,或者也可以理解为一个存储空间。例如,若标注模型的分类结果共有 nc 个类别,则需要创建 nc 个标注候选池,标注候选池 i 对应类别 i , i 为类别序号,比如 i 可以取 $0 \sim nc-1$ 之中的任意整数。

[0118] 其中,每个标注候选池中包含第一待标注数据及其推理结果中的、第一待标注数据属于该标注候选池对应的类别的概率。例如,标注候选池 i 中包含第一待标注数据及其推理结果中的、第一待标注数据属于类别 i 的概率。

[0119] 注意,所谓标注候选池中包含第一待标注数据,并不是说一定要为每个标注候选池复制一份第一待标注数据,也可以只是将第一待标注数据的标识(例如,每条第一待标注数据的编号)在标注候选池中保存一份,由于根据第一待标注数据的标识可以找到对应的第一待标注数据,因此等效于在标注候选池中保存了第一待标注数据的副本。对于后文中提到的其他数据池,若存在不同的数据池保存相同数据的情况,可以类似理解。

[0120] 创建好标注候选池后,每轮迭代可以包括以下步骤:

[0121] 步骤A1:从所有标注候选池中的第四待标注数据中抽样出在本轮迭代中提供给用户标注的第二待标注数据。

[0122] 在步骤A1中,每个标注候选池中的第四待标注数据既可以认为是上一轮迭代结束时该标注候选池中的第四待标注数据,也可以认为是本轮迭代开始时该标注候选池中的第四待标注数据,在每一轮迭代中,每个标注候选池中第四待标注数据可能、但不一定会更新(见步骤C1)。对于首轮迭代,由于不存在上一轮,因此首轮迭代开始时,每个标注候选池中的第四待标注数据均为该标注候选池中的第一待标注数据。

[0123] 从所有标注候选池中的第四待标注数据中抽样第二待标注数据的方法不限,每一轮迭代时第二待标注数据都要重新抽样。例如,可以对所有标注候选池中的第四待标注数据取并集,以去除重复的待标注数据,然后从该并集中抽样产生本轮要标注的第二待标注数据。进一步的,还可以对所有标注候选池中的第四待标注数据中尚未标注的第三待标注数据取并集,然后从该并集中抽样产生本轮要标注的第二待标注数据,从而避免已经标注过的数据被重复抽样。

[0124] 步骤B1:获取用户标注的第二待标注数据的标注类别。

[0125] 由于此时针对的是分类问题,所以第二待标注数据的标注类别就是步骤S230中提到的第二待标注数据的标注标签。

[0126] 对于每个标注候选池,执行以下步骤,以任意的标注候选池 i 为例:

[0127] 步骤C1:针对步骤A1得到的每条第二待标注数据,以任意的第二待标注数据 j 为例,若第二待标注数据 j 的标注类别和标注候选池 i 对应的类别(即类别 i)不同,且第二待标注数据 j 包含在标注候选池 i 中的第四待标注数据中,则将标注候选池 i 中的第四待标注数据缩减为仅包含满足以下条件的待标注数据:其推理结果中属于类别 i 的概率大于第二待

标注数据j的推理结果中属于类别i的概率。

[0128] 步骤C1中的缩减就是步骤230中的筛选的一种实现方式。类别i就是标注候选池i中所有的第一待标注数据的推理标签,从而若第二待标注数据j的标注类别和类别i不同,则表明第二待标注数据j的推理标签是错误的,而根据前文对第四待标注数据的定义,其中可以仅包含标注正确的第二待标注数据和/或尚未标注过的第三待标注数据,所以第二待标注数据j应当从当前的第四待标注数据中移除。

[0129] 另外,对于标注候选池i中的、除标注错误的第二待标注数据j之外的每条第四待标注数据(第二待标注数据本身也是第四待标注数据),若其推理结果中属于类别i的概率,不大于第二待标注数据j的推理结果中属于类别i的概率,则表明该条第四待标注数据的推理标签只能持平,甚至还不如第二待标注数据j的推理标签(已经是错误标签)可靠,如果将其保留在第四待标注数据中,将使得第三待标注数据的标注准确率估计难以进一步提高,从而这样的第四待标注数据也可以从当前的第四待标注数据中移除。

[0130] 不难看出,除开被移除的待标注数据,第四待标注数据中的剩余数据满足步骤C1中提到的条件。

[0131] 图5示出了第四待标注数据的缩减原理。图5中的细线箭头表示标注正确的第二待标注数据(不限于本轮标注的第二待标注数据),粗线箭头表示标注错误的第二待标注数据,尚未标注的第三待标注数据则未示出。

[0132] 参照图5,假设第二待标注数据j是左起第一个标注错误的第二待标注数据,第四待标注数据位于第二待标注数据j左侧的部分,包括若干标注正确的第二待标注数据以及未示出的第三待标注数据,由于其推理结果中属于类别i的概率大于第二待标注数据j的推理结果中属于类别i的概率(图5最下面为属于类别i的概率的变化趋势),所以在缩减后被保留,第二待标注数据j本身,以及第四待标注数据位于第二待标注数据j右侧的部分,包括一条标注错误的第二待标注数据、若干标注正确的第二待标注数据以及未示出的若干尚未标注的待标注数据,由于其推理结果中属于类别i的概率不大于第二待标注数据j的推理结果中属于类别i的概率,所以在缩减后被移除,从而代表第四待标注数据的矩形方框后半部分变为虚线。

[0133] 注意,按照上面介绍的缩减方式,在移除标注错误的第二待标注数据j时,也可能“顺便”将第二待标注数据k也移除掉,所以在步骤C1中,还要求在进行缩减之前,先确认第二待标注数据j是否仍包含在标注候选池i中的第四待标注数据中,避免重复移除。

[0134] 参照图5,假设目前要处理的第二待标注数据k是左起第二个标注错误的第二待标注数据,由于之前在处理第二待标注数据j时,已经将第二待标注数据k移除了,所以第二待标注数据k目前已经不在第四待标注数据中,也不会导致第四待标注数据被缩减。

[0135] 此外,在缩减过程中被移除掉的第四待标注数据,并不一定要从标注数据池i中删除,其也可以继续保留在标注数据池i中。

[0136] 在步骤C1之后的本轮迭代步骤中,凡是提到第四待标注数据,均指缩减后的得到的第四待标注数据,为简单起见,仍然简称为第四待标注数据。

[0137] 步骤D1:判断标注数据池i中的第四待标注数据中是否包含尚未标注的第三待标注数据。

[0138] 若第四待标注数据中包含尚未标注的第三待标注数据,则继续执行步骤E1。若第

四待标注数据中不包含尚未标注的第三待标注数据,即第四待标注数据中仅包含标注正确的第二待标注数据,此时也无法进行自动标注,从而可以跳出迭代过程,结束对标注数据池*i*的标注。

[0139] 步骤E1:根据标注数据池*i*中的第四待标注数据中标注正确的第二待标注数据的数量,估计第三待标注数据的标注准确率是否达标。

[0140] 这里给出一个估计标注准确率是否达标的判断公式: $1/(1+w^2/C_i) > acc$ 。

[0141] 其中,大于号左侧为标注准确率的估算公式,右侧的 acc 为阐述步骤S130时提到的目标准确率,若左侧大于右侧,则表明标注准确率达标。在左侧的表达式中, C_i 为标注候选池*i*的第四待标注数据中标注正确的第二待标注数据的数量(注意,这 C_i 条第二待标注数据不一定是本轮中的,之前各迭代轮次中标准正确的第二待标注数据若仍属于本轮中的第四待标注数据,则也对 C_i 产生贡献), w 为与置信度相关的预设值, w 的值越大,估计出的准确率的置信度越高,例如, $w=2$ 时,置信度为95%。

[0142] 应当理解,估计标注准确率也可能采用其他公式,对此本申请并不限定。另外,判断标注准确率是否达标,并不意味着一定要显式计算出标注准确率估值,例如,把上面的公式变形一下,得到 $C_i > w^2/(1/acc-1)$,即用 C_i 和某个阈值比较也可以进行判断。

[0143] 若第三待标注数据的标注准确率已经达标,则继续执行步骤F1,若第三待标注数据的标注准确率尚未达标,则可以开始下一轮迭代,即跳转到步骤A1继续执行。

[0144] 步骤F1:将类别*i*确定为第三待标注数据的标注类别。

[0145] 其中,根据步骤C1中的阐述,类别*i*为第三待标注数据的推理标签,而第三待标注数据的标注类别就是步骤S260中提到的第三待标注数据的标注标签。步骤F1执行完后,可以跳出迭代过程,结束对标注数据池*i*的标注。

[0146] 可选的,方式1在创建标注候选池时,还可以将每个标注候选池中的第一待标注数据,按照其推理结果中的、属于该标注候选池对应的类别的概率降序排序。此时步骤C1中可以按照如下方式实现:

[0147] 针对步骤A1得到的每条第二待标注数据,若第二待标注数据*j*的标注类别和类别*i*不同,且第二待标注数据*j*的排序索引不大于标注候选池*i*中的第四待标注数据的最大排序索引,则将标注候选池*i*中的第四待标注数据缩减为仅包含满足以下条件的待标注数据:其排序索引小于第二待标注数据*j*的排序索引。

[0148] 其中,待标注数据的排序索引可以指其在标注候选池*i*中的排序序号,由于标注候选池*i*中的第一待标注数据是按照其推理结果中对应于类别*i*的概率降序排序的,所以若某条第四待标注数据的排序索引小于第二待标注数据*j*的排序索引,表明其推理结果中属于类别*i*的概率大于第二待标注数据*j*的推理结果中属于类别*i*的概率,即与步骤C1中的条件是一致的。

[0149] 不难看出,按照以上方式进行缩减,第四待标注数据的排序索引的分布范围是单方向缩小的。例如,若排序索引从0开始递增计数,则最开始时,第四待标注数据的排序索引的分布范围是 $[0, m-1]$, m 为第一待标注数据的数量,一轮迭代之后,第四待标注数据的排序索引的分布范围是 $[0, m_1-1]$, $m_1 \leq m$,两轮迭代之后,第四待标注数据的排序索引的分布范围是 $[0, m_2-1]$, $m_2 \leq m_1$,以此类推。

[0150] 进一步的,为避免重复移除数据,步骤C1中还要求在缩减之前应满足条件“第二待

标注数据j包含在标注候选池i中的第四待标注数据中”，对于标注候选池按概率进行了降序排序的情况，此条件可以等价为第二待标注数据j的排序索引不大于标注候选池i中的第四待标注数据的最大排序索引。

[0151] 可见，若先对标注候选池中的数据进行排序，在执行第四待标注数据的缩减时，仅需进行简单的排序索引的比较，而不用进行概率值的比较，因此提高了处理效率。

[0152] 假设待标注数据的排序索引从0开始递增计数，用 L_i 表示标注候选池i中第四待标注数据的数量（此时第四待标注数据的最大排序索引为 L_i-1 ），则上述对步骤C1的实现还可以进一步简化描述为：

[0153] 若第二待标注数据j的标注类别和类别i不同，则将标注候选池i中的第四待标注数据的数量 L_i 缩减为 $\min(\text{index}(j, i), L_i)$ 。

[0154] 其中， $\text{index}(j, i)$ 表示第二待标注数据j在标注候选池i中的排序索引，下面简单解释该公式的含义：

[0155] 根据前文可知，由于在缩减过程中，第四待标注数据的排序索引的分布范围是单方向缩小的，所以第四待标注数据的数量 L_i 缩减为 $\min(\text{index}(j, i), L_i)$ ，等价于第四待标注数据的排序索引的分布范围缩减为 $[0, \min(\text{index}(j, i), L_i) - 1]$ 。显然，区间 $[0, \min(\text{index}(j, i), L_i) - 1]$ 中的任意一个数都小于 $\text{index}(j, i)$ ，即缩减后的第四待标注数据的排序索引小于第二待标注数据j的排序索引。

[0156] 并且，只有在 $\text{index}(j, i) < L_i$ 或者说 $\text{index}(j, i) \leq L_i - 1$ 时， L_i 才会发生变化，否则 L_i 仍然维持原取值，即仅当第二待标注数据j的排序索引不大于标注候选池i中的第四待标注数据的最大排序索引 $L_i - 1$ 时，缩减才有可能发生。

[0157] 可选的，方式1在执行步骤B1之前，还可以根据第二待标注数据的推理结果，确定第二待标注数据的预测类别（即预测标签），并在用户标注界面上显示第二待标注数据及其预测类别，以提高用户标注的效率。

[0158] 若标注候选池中的第一待标注数据，按照其推理结果中的属于该标注候选池对应的类别的概率进行了降序排序，则在每一轮迭代时，第二待标注数据的预测类别可以按照以下方式来确定，以任意一条第二待标注数据j为例：

[0159] 首先，获取第二待标注数据j在每个标注候选池中的排序索引，例如 $\text{index}(j, i)$ ，i取遍 $0 \sim n_c - 1$ ，即共获得 n_c 个排序索引。

[0160] 然后，将每个排序索引用对应的标注候选池中的第四待标注数据的数量进行归一化，得到归一化后的排序索引。例如， $\text{index}(j, i)$ 归一化后可以是 $\text{index}(j, i) / L_i$ 。注意，由于给出预测标签的步骤是在步骤B1之前执行的，此时本轮迭代中关于第四待标注数据的缩减还未执行，所以 L_i 应为本轮迭代开始时第四待标注数据的数量。

[0161] 最后，将所有归一化后的排序索引中的最小排序索引对应的类别确定为第二待标注数据j的预测类别。例如，将 $\text{argmin}_i(\text{index}(j, i) / L_i)$ 作为第二待标注数据j的预测类别。其中， argmin_i 表示计算使得后面括号内的数值取最小值时的i值。

[0162] 通过此方式确定的预测类别，由于用标注候选池中的第四待标注数据的数量对排序索引进行了归一化，相当于削弱了不同类别的数据在分布上的差异，从而有利于改善预测类别的精度，进而提高人工标注的效率。

[0163] 方式2

[0164] 方式2无需创建标注候选池,或者也可以认为方式2中的标注候选池只有一个,全体第一待标注数据都在该标注候选池中。每轮迭代可以包括以下步骤:

[0165] 步骤A2:从第四待标注数据中抽样出在本轮迭代中提供给用户标注的第二待标注数据。

[0166] 其中,首轮迭代开始时的第四待标注数据为第一待标注数据。步骤A2和步骤S220、步骤A1类似,不再重复阐述。

[0167] 步骤B2:获取用户标注的第二待标注数据的标注类别。

[0168] 由于此时针对的是分类问题,所以第二待标注数据的标注类别就是步骤S230中提到的第二待标注数据的标注标签。

[0169] 步骤C2:针对步骤A2得到的每条第二待标注数据,以任意的第二待标注数据j为例,若第二待标注数据j的标注类别和其推理结果中的最大概率对应的类别不同,且第二待标注数据j包含在第四待标注数据中,则将第四待标注数据缩减为仅包含满足以下条件的待标注数据:其推理结果中的最大概率大于第二待标注数据j的推理结果中的最大概率。

[0170] 步骤C2中的缩减就是步骤230中的筛选的一种实现方式。所有第一待标注数据的推理标签就是其推理结果中的、第一待标注数据分属每个类别的概率中的最大值(简称推理结果中的最大概率)所对应的类别。从而若第二待标注数据j的标注类别和其推理结果中的最大概率对应的类别不同,则表明第二待标注数据j的推理标签是错误的,而根据前文对第四待标注数据的定义,其中可以仅包含标注正确的第二待标注数据和/或尚未标注过的第三待标注数据,所以第二待标注数据j应当从当前的第四待标注数据中移除。

[0171] 另外,对于除标注错误的第二待标注数据j之外的每条第四待标注数据(第二待标注数据本身也是第四待标注数据),若其推理结果中的最大概率不大于第二待标注数据j的推理结果中的最大概率,则表明该条第四待标注数据的推理标签只能持平,甚至还不如第二待标注数据j的推理标签(已经是错误标签)可靠,如果将其保留在第四待标注数据中,将使得第三待标注数据的标注准确率估计难以进一步提高,从而这样的第四待标注数据也可以从当前的第四待标注数据中移除。

[0172] 不难看出,除开被移除的待标注数据,第四待标注数据中的剩余数据满足步骤C2中提到的条件。

[0173] 在步骤C2之后的本轮迭代步骤中,凡是提到第四待标注数据,均指缩减后的得到的第四待标注数据,为简单起见,仍然简称为第四待标注数据。

[0174] 步骤D2:判断第四待标注数据中是否包含尚未标注的第三待标注数据。

[0175] 若第四待标注数据中包含尚未标注的第三待标注数据,则继续执行步骤E2。若第四待标注数据中不包含尚未标注的第三待标注数据,即第四待标注数据中仅包含标注正确的第二待标注数据,此时也无法进行自动标注,从而可以跳出迭代过程,结束对第一待标注数据的标注。

[0176] 步骤E2:根据第四待标注数据中标注正确的第二待标注数据的数量,估计第三待标注数据的标注准确率是否达标。

[0177] 估计标注准确率是否达标的判断公式可以为: $1/(1+w^2/C) > acc$ 。其中,C为第四待标注数据中标注正确的第二待标注数据的数量(注意,这C条第二待标注数据不一定是本轮中的,之前各迭代轮次中标准正确的第二待标注数据若仍属于本轮中的第四待标注数据,

则也对C产生贡献), w 为与置信度相关的预设值, acc 为阐述步骤S130时提到的目标准确率。不难看出,该公式的和前面的公式 $1/(1+w^2/C_i) > acc$ 是类似的,只是 C_i 换成了 C ,因为方式2本来就可以视为只有一个标注候选池。

[0178] 应当理解,判断标注准确率是否达标也可能采用其他公式,对此本申请并不限定。

[0179] 若第三待标注数据的标注准确率已经达标,则继续执行步骤F2,若第三待标注数据的标注准确率尚未达标,则可以开始下一轮迭代,即跳转到步骤A2继续执行。

[0180] 步骤F2:将第三待标注数据的推理结果中的最大概率对应的类别确定为第三待标注数据的标注类别。

[0181] 其中,根据步骤C2中的阐述,第三待标注数据的推理结果中的最大概率对应的类别为第三待标注数据的推理标签,而第三待标注数据的标注类别就是步骤S260中提到的第三待标注数据的标注标签。步骤F2执行完后,可以跳出迭代过程,结束对第一待标注数据的标注。

[0182] 可选的,方式2在开始迭代之前,还可以将第一待标注数据按照其推理结果中的、第一待标注数据分属每个类别的概率中的最大概率降序排序。此时步骤C2中可以按照如下方式实现:

[0183] 针对步骤A2中得到的每条第二待标注数据,若第二待标注数据 j 的标注类别和其推理结果中的最大概率对应的类别不同,且第二待标注数据 j 的排序索引不大于第四待标注数据中的最大排序索引,则将第四待标注数据缩减为仅包含满足以下条件的待标注数据:其排序索引小于第二待标注数据 j 的排序索引。

[0184] 其中,待标注数据的排序索引可以指其在第一待标注数据中的排序序号,由于第一待标注数据是按照其推理结果中的最大概率降序排序的,所以若某条第四待标注数据的排序索引小于第二待标注数据 j 的排序索引,表明其推理结果中的最大概率大于第二待标注数据 j 的推理结果中的最大概率,即与步骤C2中的条件是一致的。不难看出,按照以上方式进行缩减,第四待标注数据的排序索引的分布范围是单方向缩小的。

[0185] 进一步的,为避免重复移除数据,步骤C2中还要求在缩减之前应满足条件“第二待标注数据 j 包含在第四待标注数据中”,对于第一待标注数据按概率进行了降序排序的情况,此条件可以等价为第二待标注数据 j 的排序索引不大于第四待标注数据的最大排序索引。

[0186] 可见,若先对第一待标注数据进行排序,在执行第四待标注数据的缩减时,仅需进行简单的排序索引的比较,而不用进行概率值的比较,因此提高了处理效率。

[0187] 假设待标注数据的排序索引从0开始递增计数,用 L 表示第四待标注数据的数量(此时第四待标注数据的最大排序索引为 $L-1$),则上述对步骤C2的实现还可以进一步简化描述为:

[0188] 若第二待标注数据 j 的标注类别和其推理结果中的最大概率对应的类别不同,则将第四待标注数据的数量 L 缩减为 $\min(\text{index}(j), L)$ 。其中, $\text{index}(j)$ 表示第二待标注数据 j 在第一待标注数据中的排序索引,该公式的以可以参照方式1中的公式 $\min(\text{index}(j, i), Li)$ 进行理解,不再重复阐述。

[0189] 可选的,方式2在执行步骤B2之前,还可以根据第二待标注数据的推理结果,确定第二待标注数据的预测类别(即预测标签),并在用户标注界面上显示第二待标注数据及其

预测类别,以提高用户标注的效率。

[0190] 若第一待标注数据,按照其推理结果中的最大概率进行了降序排序,第二待标注数据的预测类别可以按照以下方式来确定(与迭代轮次无关),以任意一条第二待标注数据j为例:

[0191] 将第二待标注数据j的推理结果中的最大概率对应的类别确定为第二待标注数据j的预测类别。通过此方式确定预测类别十分简单快速。

[0192] 简单对比方式1和方式2,方式2在实现上比较简单,无需创建多个标注候选池。当方式1的优势在于,针对每个类别的数据都有一个标注候选池,每个标注候选池中的第四待标注数据的数量是不同的,相当于自适应地为每个类别都设置了一个判别阈值,来决定哪些数据可以被标注为该类别,而并非每个类别都设置相同的判别阈值(如,概率值 $1/nc$),从而可以在一定程度上克服待标注数据中存在的偏置(即每个类别的数据分布不同),改善标注的准确率。

[0193] 应当理解,即使对于标注模型是分类模型的情况,迭代步骤S220~S260的也不限于方式1和方式2两种实现方式,比如,在方式2的某种实现方式中,按照推理结果中的最大概率对第一待标注数据进行降序排序,该最大概率也可以替换成其他指标。

[0194] 在阐述方式1和方式2的步骤时,未提及之处,均可以参考图1及图4中相关步骤的阐述,或者两种方式相互参照。

[0195] 根据前文的阐述可知,图4中的数据标注方法只对当前的第四待标注数据中尚未标注的第三待标注数据给出标注准确率估计,对于其他待标注数据,其标注准确率则是未知,并且只在第三待标注数据的标注准确率达标时,才给出其标签,标注准确率未达标时,第三待标注数据的标签则是未知,并且其他尚未标注的待标注数据(步骤S230中被筛选掉的),其标签也是未知。

[0196] 即,并不能保证每条待标注数据在任何时刻都有一个当前合理的标注准确率和一个当前合理的标签,这给某些应用场景带来了困难:例如,用户标注了一段时间之后,因故不得不停止标注,并希望根据目前的标注进度,获取尽可能多的、标注准确率大于96%(目标准确率为99%)的数据及其标签用于任务模型的训练,但按照之前介绍的方法,至多只能给出目前的第三待标注数据的标注准确率估值,难以满足用户需求。

[0197] 针对上述问题,在一种改进方案中,为每条待标注数据都增设两项属性:标注准确率属性和标注标签属性,标注准确率属性的中存储该条待标注数据当前的标注准确率估值,标注标签属性中存储该条待标注数据当前的标签估值。这样,用户随时可以结束标注,并根据每条待标注数据的标注准确率属性的值来筛选其是否可作为任务模型的训练样本,并根据其标注标签属性的值为训练样本打上标签。

[0198] 下面,结合前面方式1中的步骤,来举例说明如何计算每条待标注数据的标注准确率属性和标注标签属性,由于方式1针对的是分类问题,所以标注标签属性也可以改称标注类别属性。

[0199] 首先,执行初始化步骤,可以在执行迭代步骤之前为第一待标注数据的这两项属性设置初始值,例如,对于标注准确率属性,可以将其初始值设置为0、 $1/nc$ 等,对于标注标签属性,可以将其初始值设置为标签未知(可视为一个特殊标签),或者也可以设置为其推理标签等。

[0200] 注意,如果某条第一待标注数据的这两项数据本来就有值,且并非初始值,则无需对其进行初始化,原因在于,这些待标注数据在之前可能标注过(比如之前的某批第一待标注数据),之前标注的时候就计算过这两项属性的值,但在之前标注的时候被缩减掉了,并未被打上最终的标注标签,所以本次又被加入到当前的一批第一待标注数据中,重新标注(由于此时标注模型可能已经优化,所以其对同一条待标注数据的推理结果也很可能和之前的不同,重新标注可能会得到新的结果)。

[0201] 在执行步骤B1之后,可以将每条第二待标注数据的标注准确率属性的值设置为第一数值,并将其标注类别属性的值设置为用户给出的标注类别。其中,第一数值可以是一个预设的值,例如,若默认用户给出的标注类别是正确的,则第一数值可以是1,当然也不排除第一数值取其他值。

[0202] 在执行步骤C1时,可能会有部分待标注数据被移除第四待标注数据,被移除的待标注数据包括三种成分:其一,是第二待标注数据,其标注准确率属性的值为第一数值,标注标签属性的值为用户给出的标签;其二,是首轮迭代即被缩减掉的、尚未标注的数据,由于进行了初始化,其标注准确率属性和标注标签属性均为各自的初始值;其三,是非首轮迭代中被缩减掉的、尚未标注的数据,根据下面的内容可知,其标注准确率属性和标注标签属性可能会被赋予特定的值。

[0203] 在执行步骤D1之后,若第四待标注数据中包含尚未标注的第三待标注数据,则获取第三待标注数据的标注准确率。其中,若步骤E1中进行判断时计算了第三待标注数据的标注准确率,则此处可以直接获取E1中的计算结果,若步骤E1中进行判断时并未计算第三待标注数据的标注准确率(如前所述,估计标注准确率是否达标,并不一定要显式计算标注准确率估值),则此时需要计算第三待标注数据的标注准确率。

[0204] 针对每条第三待标注数据,以任意的第三待标注数据s为例,若第三待标注数据s的标注准确率属性的值小于获取到的标注准确率,则将第三待标注数据s的标注准确率属性的值设置为获取到的标注准确率,并将其标注类别属性的值设置为类别i(以标注候选池i中的情况为例)。

[0205] 例如,根据公式 $1/(1+w^2/C_i)$ 可以估计标注候选池i中的第三待标注数据s的标注准确率,将第三待标注数据s的标注准确率属性的值记为 $p(s)$ 、标注类别属性的值记为 $c(s)$,若 $p(s) < 1/(1+w^2/C_i)$,则将 $p(s)$ 设置为 $1/(1+w^2/C_i)$,将 $c(s)$ 设置为i,否则 $p(s)$ 和 $c(s)$ 的值保持不变。特别地,若 $p(s) > acc$,则 $c(s)$ 的值也就是第三待标注数据最终的标注标签。

[0206] 注意,由于同一条待标注数据s在每个标注候选池中都会保留副本,所以可能在多个标注候选池中都会估计标注准确率,但按照上面的公式,会自动选择计算结果中的最大值 $p(s)$ 。

[0207] 应当理解,即使图4中的步骤不是通过方式1实现的,也可以按照类似的方式计算每条第一待标注数据的标注准确率属性和标注类别属性的值,不再详细阐述。

[0208] 本申请实施例还提供一种数据标注系统,该数据标注系统可以、但不限于部署在图9示出的电子设备,用于执行上面介绍的数据标注方法及其任意一种实现方式的步骤,图6示出了数据标注系统300的架构及其工作原理,下面将结合数据标注系统300继续介绍本申请实施例提供的标注方法。

[0209] 参照图6,数据标注系统300包括:

[0210] 两个功能模块:模型训练与推理模块310和标注策略模块330;其中,功能模块可以理解为数据标注程序的程序模块;

[0211] 四个数据池:标注结果池322、推理数据候选池324、训练数据池326和推理结果池328;其中,数据池可以理解为一个数据集,或者也可以理解为一个存储空间。

[0212] 以及,一个界面:用户标注界面340。

[0213] 结合图6,该系统的工作工程概述如下:

[0214] 步骤1:模型训练与推理模块310从训练数据池326中抽样出至少零批次训练数据对标注模型进行训练,以及,利用训练后得到的标注模型对从推理数据候选池324中抽样出的至少一批次的待标注数据进行推理,得到至少一批次的待标注数据的推理结果。

[0215] 步骤1中的抽样方式不限。训练数据池326为训练数据的数据源,训练数据池326中的训练数据包括训练样本及其标注标签(例如,图像和其标注类别),如何得到训练数据稍后说明。一批次(batch)的数据包括至少一条数据,若抽样出的训练数据为零批次,其实表明此时不进行数据抽样,从而也不进行标注模型的训练,继续使用原有的标注模型,只是为了表述上的简便性,将抽取零批次训练数据的情况与抽取至少一批次训练数据的情况合并在一起了而已。抽取零批次训练数据的情况是可能发生的,例如,认为当前标注模型的性能已经足够好,无需调整其参数,避免浪费计算资源。被抽样出的训练数据无需从训练数据池326中移除。关于标注模型的训练过程,可以参考现有技术,此处不具体解释。

[0216] 推理数据候选池326为待标注数据的数据源,待标注数据是无标签的数据。被抽样出的待标注数据暂时无需从推理数据候选池326中移除。关于标注模型的推理过程,可以参考现有技术,此处不具体解释。

[0217] 可选的,数据标注系统300初始化时,模型训练与推理模块310会将标注模型初始化,包括构建模型的结构、设置模型的初始参数、设置模型使用的损失函数等。

[0218] 然后,模型训练与推理模块310可将所有待标注数据都写入推理数据候选池326,这样推理数据候选池326中最初的数据就有了,从而可以支持步骤1中提到的数据抽样。然后,可以由标注策略模块330从推理数据候选池326中读取预设数量(例如,30条)的待标注数据,并将这些待标注数据直接交由用户在用户标注界面340上进行标注并获得用户标注界面340返回的标注标签,注意,这部分待标注数据不含预测标签,因为此时标注模型还没训练,无法给出预测标签。之后,标注策略模块330将这部分待标注数据及其标注标签写入训练数据池326,这样训练数据池326中最初的数据就有了,从而可以支持步骤1中提到的数据抽样。

[0219] 图7示步骤1中进行模型训练和推理的一种实现方式。参照图7,步骤1可实现为至少一次模型训练/推理循环,每次模型训练/推理循环中包括至少零次模型训练(从训练数据池326中每抽样一批次训练数据对应一次模型训练),以及,利用训练后得到的标注模型进行的至少零次模型推理(从推理数据候选池326中每抽样一批次待标注数据对应一次模型推理)。

[0220] 但需注意,在每一次模型训练/推理循环中训练和推理的次数不能都是零,并且,在全部的模型训练/推理循环中至少应包括一次模型推理,否则无法得到待标注数据的推理结果。例如,进行两次模型训练/推理循环,可以实现为顺序执行的1次模型训练、1次模型推理、1次模型训练、1次模型推理,也可以实现为0次模型训练、1次模型推理、1次模型训练、

0次模型推理,也可以实现为0次模型训练、1次模型推理、0次模型训练、2次模型推理,等等。

[0221] 步骤2:模型训练与推理模块310将步骤1中的至少一批次的待标注数据及其推理结果写入推理结果池328。

[0222] 可选的,数据标注系统300初始化时,模型训练与推理模块310可以将推理结果池328初始化为空。

[0223] 步骤3:标注策略模块330从推理结果池328中读取第一待标注数据及其推理结果。

[0224] 其中,第一待标注数据的数量不超过推理结果池328中目前的待标注数据的总数量,被读取出的第一待标注数据及其推理结果可以认为是从推理结果池328中移除了。

[0225] 步骤3有多种可能的实现方式:从读取时机上说,标注策略模块330可以每隔固定的时间从推理结果池328中读取第一待标注数据及其推理结果,也可以等待上一批的第一待标注数据的标注结束后(前文已经介绍了结束条件)再从推理结果池328中读取新一批的第一待标注数据及其推理结果,等等。从读取的数据量上说,标注策略模块330可以将推理结果池328中当前的全部待标注数据及其推理结果都读取出来作为第一待标注数据其推理结果,也可以总是从推理结果池328中读取固定条数的第一待标注数据其推理结果(除非推理结果池328中数据量不足),等等。

[0226] 步骤1~3也可以认为是步骤S110的一种实现方式。

[0227] 步骤4:标注策略模块330执行第一待标注数据的标注流程。

[0228] 步骤4对应步骤S120~S140,在标注过程中,还需要将第二待标注数据及其预测标签(可选)在用户标注界面340上进行显示,并通过用户标注界面340获得第二待标注数据的标注标签。

[0229] 对于第二待标注数据及其标注标签,标注策略模块330可以将其写入训练数据池326,用于增加其中训练数据的数量,从而在后续再次执行步骤1时提升标注模型的性能。对于自动标注出的第三待标注数据及其标注标签(步骤S140得到),也可以写入训练数据池326,但这不是必须的,因为毕竟自动标注结果未经过人工确认,其标注准确率仅仅是算法估计得到的,可能会对标注模型的性能产生负面影响。

[0230] 对于已经确定了标注标签的第二待标注数据和第三待标注数据,标注策略模块330可以将其从推理数据候选池324中移除,避免其重复推理、重复标注。不过,对于那些未确定标注标签的待标注数据,则可以继续将其保留在推理数据候选池324中,允许其再次推理、再次标注,随着标注模型的优化,针对这些数据有可能会推理出不同的结果,从而被成功标注。

[0231] 另外,数据标注系统300初始化时,模型训练与推理模块310可以将所有待标注数据及一个特殊标签(代表未知标签)写入标注结果池322中,在执行步骤4时,对于已经确定了标注标签的第二待标注数据和第三待标注数据,标注策略模块330可以利用其标注标签将标注结果池322中对应的特殊标签修改掉。可选的,对于前文提到的待标注数据的标注准确率属性和标注标签属性,也可以保存在标注结果池322中。

[0232] 在替代方案中,标注结果池322也可以初始化为空,标注策略模块330负责将已经确定了标注标签的第二待标注数据和第三待标注数据、以及二者的标注标签写入标注结果池322。

[0233] 步骤1~4可以重复执行,在一种实现方式中(简称同步模式),执行步骤3~4时,不

允许执行步骤1~2。在另一种实现方式中(简称异步模式),执行步骤3~4时,允许同时执行步骤1~2,即标注策略模块330从推理结果池328中读取第一待标注数据后,不等第一待标注数据标注结束,模型训练与推理模块310又可能将新的待标注数据及其推理结果写入推理结果池328。

[0234] 在异步模式下,由于模型训练与推理模块310和标注策略模块330都可以按照各自的逻辑独立工作,所以其标注效率高于同步模式,但对于异步模式,可能导致一个问题:虽然根据前文介绍,对于已经确定了标注标签的第二待标注数据和第三待标注数据,标注策略模块330会将其从推理数据候选池324中移除,但在移除之前,这些数据可能又被模型训练与推理模块310从推理数据候选池324中抽样出来,并进行推理,从而导致其再次进入推理结果池328,进而引起重复标注。

[0235] 为改善此问题,对于已经确定了标注标签的第二待标注数据和第三待标注数据,标注策略模块330还可以将其从推理数据候选池324中移除,如果其存在于推理数据候选池324中的话。

[0236] 进一步的,在异步模式下,由于标注策略模块330随时可能向训练数据池326写入新的训练数据,所以若步骤1采用训练/推理循环的方式实现,在不同轮次的训练/推理循环中,训练数据池326中的训练数据可能是不同的。

[0237] 应当理解,通过模型训练与推理模块310和标注策略模块330来执行步骤1~4只是一种可选方式,程序模块的划分是比较自由的,例如,也可以由同一个模块来执行步骤1~4。

[0238] 下面,在以上实施例的基础上,继续介绍本申请实施例提供的两种标注模式:

[0239] 可选的,数据标注过程可以在两种标注模式下进行,一种是半自动标注模式,一种是人工标注模式。其中,半自动标注模式是指用户标注结合自动标注的模式,例如,前面介绍的数据标注方法就是半自动标注模式,其中第二待标注数据通过用户标注,标注准确率达标的第三待标注数据通过机器自动标注,人工标注模式是指完全由用户标注的模式,即所有待标注数据都交由用户标注,后文会给出人工标注模式的例子。

[0240] 在满足标注模式切换条件时,这两种标注模式之间可以相互切换,具体为:满足人工标注模式切换条件时,从半自动标注模式切换为人工标注模式,在满足半自动标注模式切换条件时,从人工标注模式切换为半自动标注模式。

[0241] 例如,若将一个标注候选池中的第三待标注数据被自动标注称为执行了一次跳过策略,若连续M个(M大于某一个数阈值)标注候选池都未执行跳过策略,而是由于标注候选池中的第四待标注数据仅包含标注正确的第二待标注数据而结束标注,则表明此时的数据对于机器来说标注起来比困难(或者说说明此时的标注模型的性能达不到自动标注的要求),从而可以从半自动标注模式切换为人工标注模式。

[0242] 不过需要注意,即使继续采用半自动标注模式进行标注,也是可以标注完的,但通常可将人工标注模式下的标注方法设计得比半自动标注模式更简单一些(比如,不用创建若干标注候选池),使得在进行模式切换后标注效率能够提高。

[0243] 又例如,若标注策略模块330每次都从推理结果池328中读取全部的待标注数据及其推理结果作为第一待标注数据及其推理结果,则当标注策略模块330检测到推理结果池中待标注数据的数量小于某个数量阈值时,可以从半自动标注模式切换为人工标注模式。

[0244] 因为若模型训练与推理模块310大致以恒定的速度进行推理,推理结果池328中待标注数据的数量较少,表明标注进入了收尾阶段,而收尾阶段剩下的待标注数据,很可能是之前历次标注过程中未能成功标注的数据(比如,被缩减掉的),即难数据,从而可以合理地推断这些数据并不能触发、或者只能少量触发跳过策略,还不如直接由用户进行标注。

[0245] 其中,对于标注模型是分类模型,且分类结果为nc个类别,且采用方式1进行半自动标注的情况,该数量阈值可以、但不限于设置为 $(w^2/(1/\text{acc}-1)) \times nc \times h$,其中h为一个预设的倍数,例如可以取3、5等。

[0246] 下面再以标注模型是分类模型的情况为例,介绍一种人工标注模式的实现方法:

[0247] 对于一批待标注数据,假设叫第五待标注数据,第五待标注数据的推理结果包括第五待标注数据分属每个类别的概率,第五待标注数据在人工标注模式的标注方式为:

[0248] 首先,获取第五待标注数据及其推理结果;其中,第五待标注数据的推理结果利用标注模型对第五待标注数据进行推理得到。例如,可以从推理结果池328中读取第五待标注数据及其推理结果。

[0249] 然后,将第五待标注数据按照其推理结果中的最大概率降序排序,按照排序索引递增的顺序逐批(比如,每批100条数据)将第五待标注数据及其预测类别提供给用户进行标注;其中,每条第五待标注数据的预测类别为其推理结果中的最大概率对应的类别。

[0250] 获取用户标注的每批第五待标注数据的标注类别,直至第五待标注数据全部标注完成,或者,也可以在每标注完一批第五待标注数据后,就重新检查一下目前是否满足半自动标注模式切换条件,若满足切换条件,则从人工标注模式切换为半自动标注模式,并将剩余尚未标注的第五待标注数据改为半自动标注。

[0251] 上述标注方法的关键在于按照排序索引分批标注,各批第五待标注数据的标注难度从低到高(推理结果中的概率越低,表明标注模型对数据所属类别越不肯定,从而标注难度也越高),用户先以较高的效率标注低难度数据,有可能在低难度数据标注完后,因满足了半自动标注模式切换条件,剩余的数据就不用全部靠用户来标注了,因此可以避免用户完全自己去标注难度较高的数据。

[0252] 应当理解,采用两种标注模式不是必须的,例如也可以一直采用半自动标注模式进行标注。

[0253] 在一些应用场景中,希望对待标注数据中的简单数据集和难数据集进行区分,从而可以在这两个数据集上分别评估任务模型的性能,进而确定其在训练时的优化方向。

[0254] 借助本申请实施例提供的标注方法可以实现简单数据集和难数据集的划分:

[0255] 例如,在半自动标注模式下标注的数据都认为属于简单数据集,在人工标注模式下标注的数据都认为属于难数据集。又例如,在所有的待标注数据标注完成后,根据其标注准确率属性的值进行判断,若其大于某个阈值,则认为待标注数据属于简单数据集,否则认为其属于难数据集,等等。

[0256] 图8示出了本申请实施例提供的标注装置400的一种可能的结构。参照图8,标注装置400包括:

[0257] 数据获取单元410,用于获取第一待标注数据及其推理结果;其中,所述第一待标注数据的推理结果为利用标注模型对所述第一待标注数据进行预测得到的信息;

[0258] 人工标注单元420,用于获取用户标注的第二待标注数据的标注标签;其中,所述

第二待标注数据为所述第一待标注数据中的部分待标注数据；

[0259] 准确率评估单元430,用于根据所述第二待标注数据的标注标签估计所述第一待标注数据中尚未标注的第三待标注数据的标注准确率是否已经达标;其中,所述第三待标注数据的标注准确率为根据所述第三待标注数据的推理结果所确定的推理标签的准确率;

[0260] 自动标注单元440,用于在所述第三待标注数据的标注准确率已经达标时,将所述第三待标注数据的推理标签确定为所述第三待标注数据的标注标签。

[0261] 本申请实施例提供的标注装置400,其实现原理及产生的技术效果在前述方法实施例中已经介绍,其可实现前述方法实施例中的全部内容,为简要描述,该装置所包括各个单元所对应功能的具体实现过程可参考前述方法实施例中相应内容,此处不再赘述。

[0262] 图9示出了本申请实施例提供的电子设备500的结构。参照图9,电子设备500包括:处理器510、存储器520以及通信接口530,这些组件通过通信总线540和/或其他形式的连接机构(未示出)互连并相互通讯。

[0263] 其中,处理器510包括一个或多个(图中仅示出一个),其可以是一种集成电路芯片,具有信号的处理能力。上述的处理器510可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、微控制单元(Micro Controller Unit,简称MCU)、网络处理器(Network Processor,简称NP)或者其他常规处理器;还可以是专用处理器,包括图形处理器(Graphics Processing Unit,GPU)、神经网络处理器(Neural-network Processing Unit,简称NPU)、数字信号处理器(Digital Signal Processor,简称DSP)、专用集成电路(Application Specific Integrated Circuits,简称ASIC)、现场可编程门阵列(Field Programmable Gate Array,简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。并且,在处理器510为多个时,其中的一部分可以是通用处理器,另一部分可以是专用处理器。

[0264] 存储器520包括一个或多个(图中仅示出一个),其可以是,但不限于,随机存取存储器(Random Access Memory,简称RAM)、只读存储器(Read Only Memory,简称ROM)、可编程只读存储器(Programmable Read-Only Memory,简称PROM)、可擦除可编程只读存储器(Erasable Programmable Read-Only Memory,简称EPROM)、电可擦除可编程只读存储器(Electric Erasable Programmable Read-Only Memory,简称EEPROM)等。

[0265] 处理器510以及其他可能的组件可对存储器520进行访问,读和/或写其中的数据。特别地,在存储器520中可以存储一个或多个计算机程序指令,处理器510可以读取并运行这些计算机程序指令,以实现本申请实施例提供的标注方法。

[0266] 通信接口530包括一个或多个(图中仅示出一个),可以用于和其他设备进行直接或间接地通信,以便进行数据的交互。通信接口530可以包括进行有线和/或无线通信的接口。

[0267] 可以理解,图9所示的结构仅为示意,电子设备500还可以包括比图9中所示更多或者更少的组件,或者具有与图9所示不同的配置。例如,电子设备500在不需要与其他设备进行通信时,也可以不包括通信接口530。

[0268] 图9中所示的各组件可以采用硬件、软件或其组合实现。电子设备500可能是实体设备,例如服务器、PC机、笔记本电脑、平板电脑、手机等,也可能是虚拟设备,例如虚拟机、容器等。并且,电子设备500也不限于单台设备,也可以是多台设备的组合或者大量设备构

成的集群。

[0269] 本申请实施例还提供一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序指令,这些计算机程序指令被处理器读取并运行时,执行本申请实施例提供的标注方法。例如,计算机可读存储介质可以实现为图9中电子设备500中的存储器520。

[0270] 本申请实施例还提供一种计算机程序产品,该计算机程序产品包括计算机程序指令,这些计算机程序指令被处理器读取并运行时,执行本申请实施例提供的标注方法。

[0271] 以上所述仅为本申请的实施例而已,并不用于限制本申请的保护范围,对于本领域的技术人员来说,本申请可以有各种更改和变化。凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的保护范围之内。

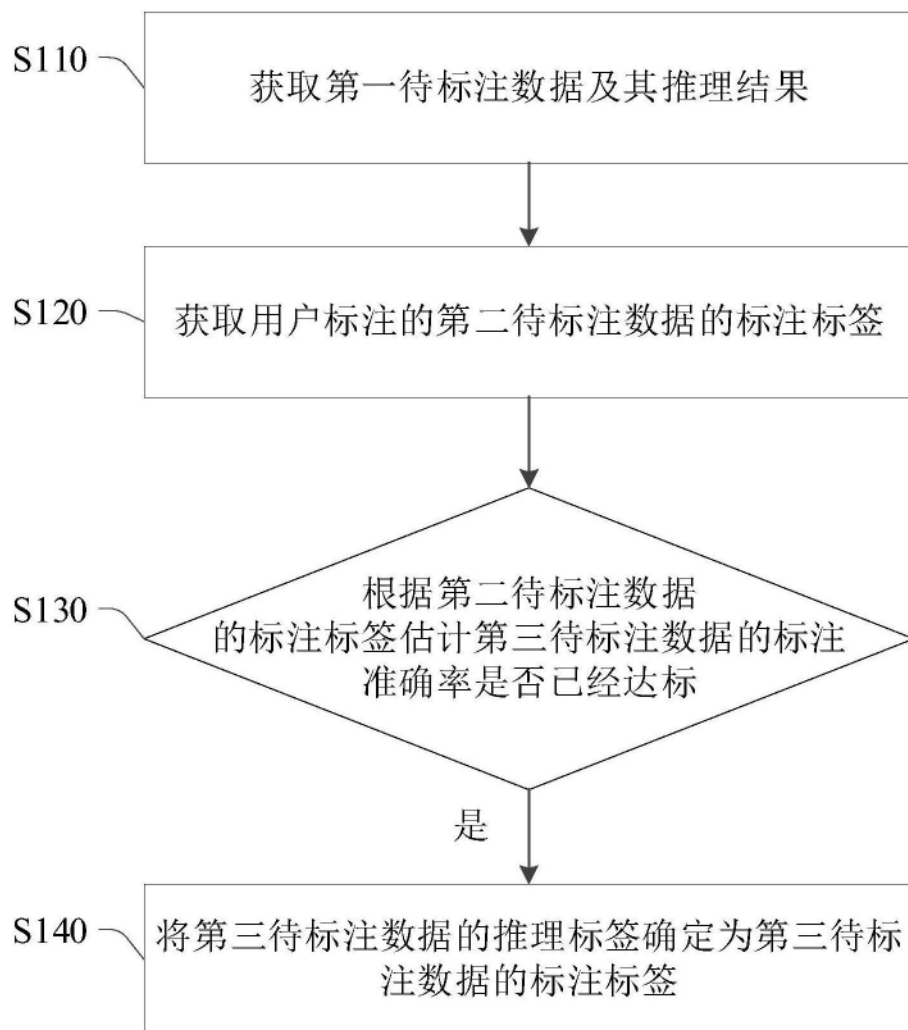


图1

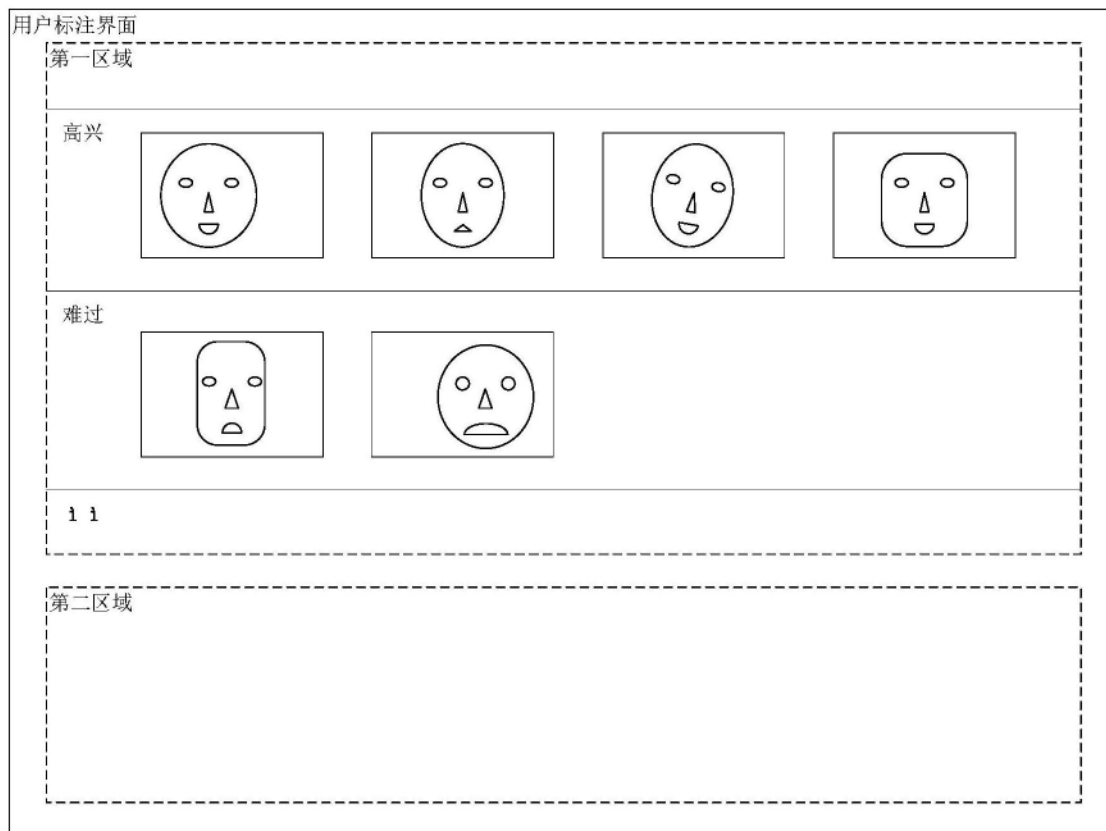


图2 (A)

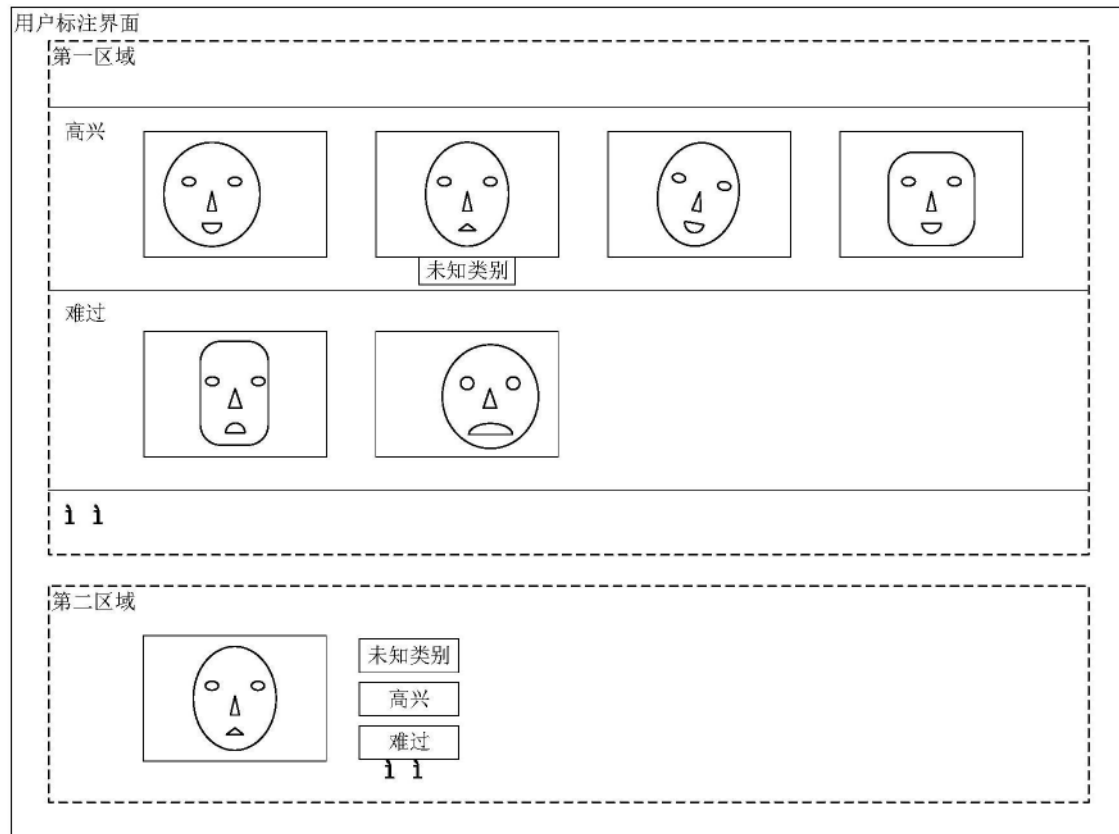


图2 (B)

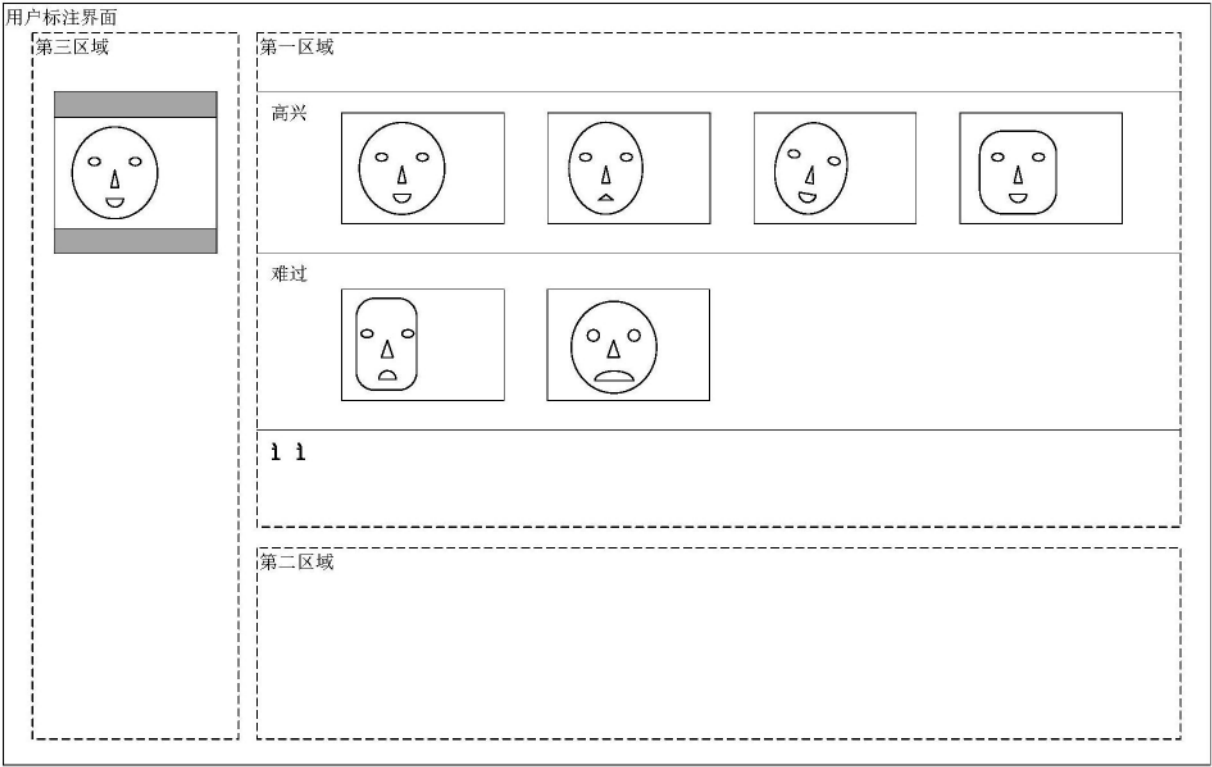


图3 (A)

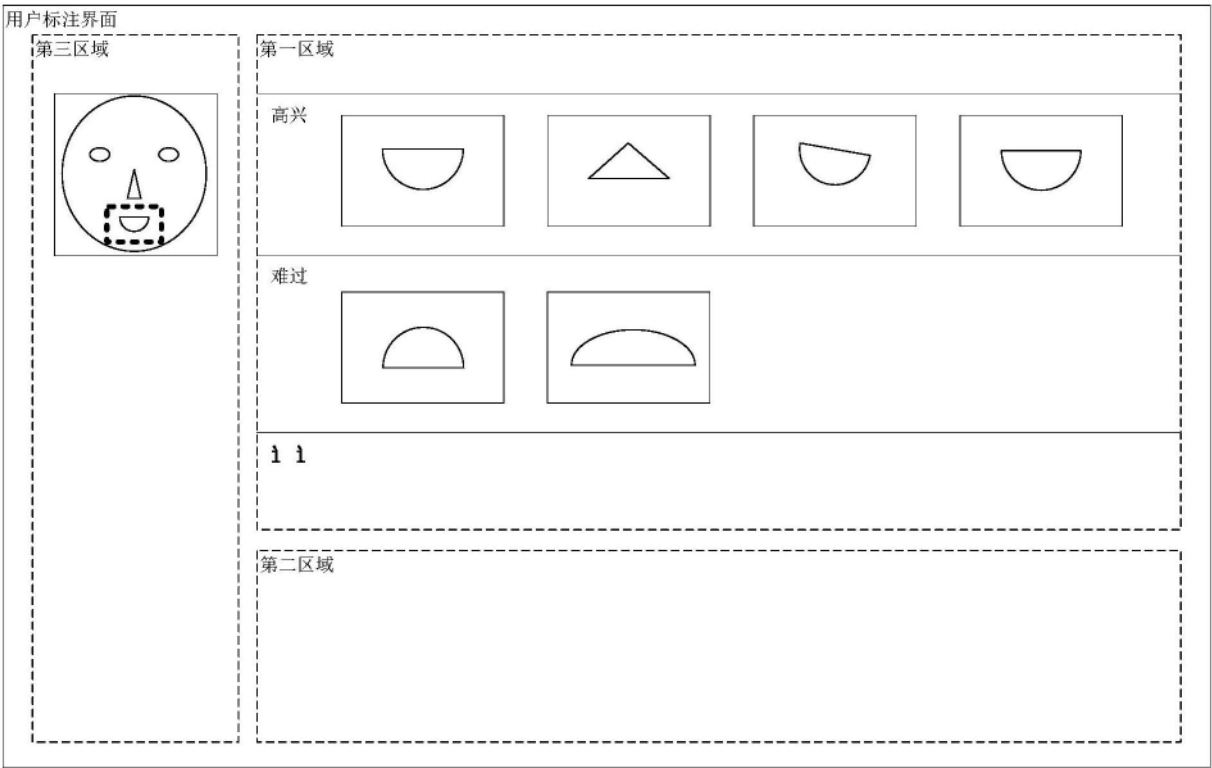


图3 (B)

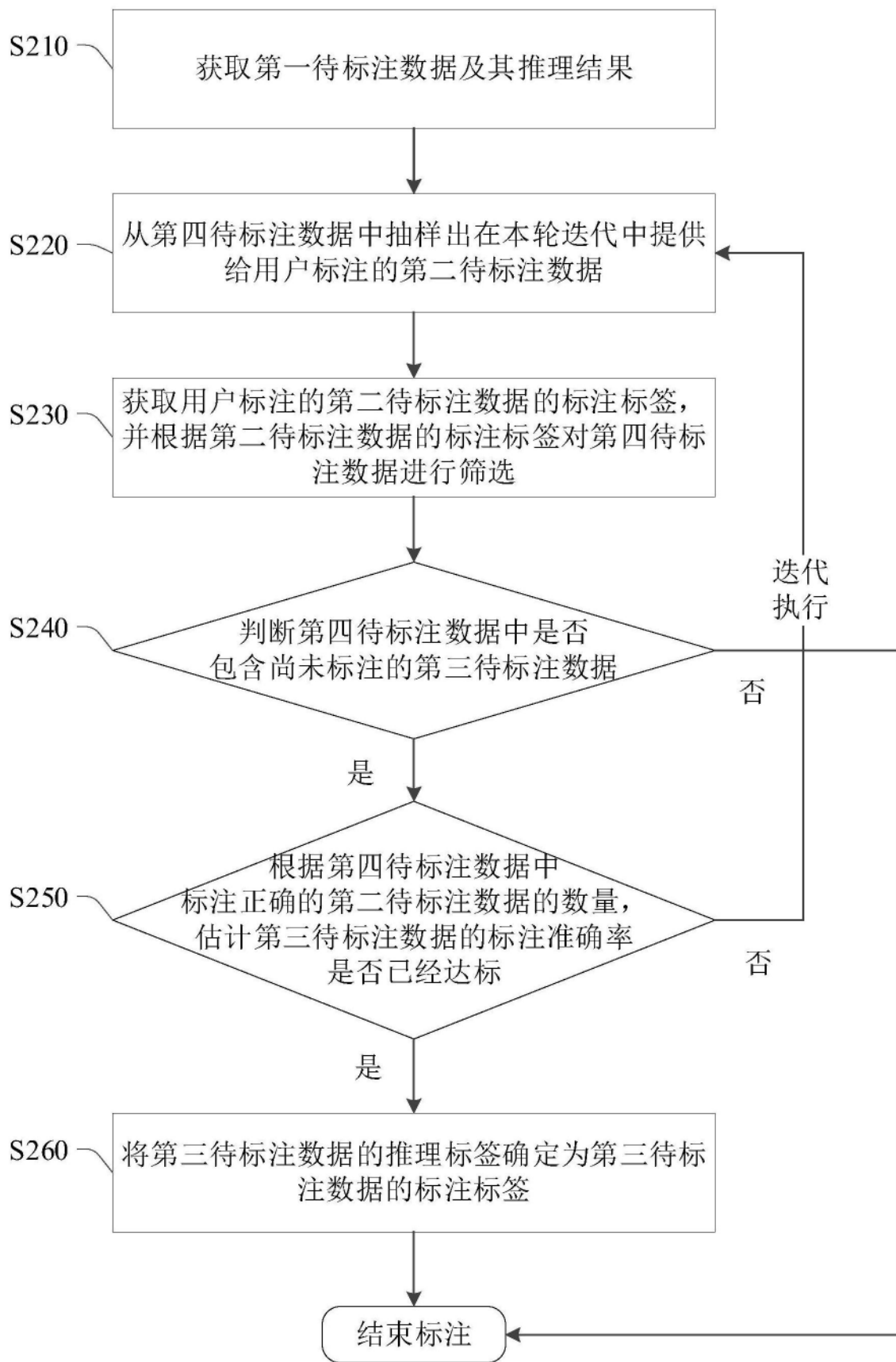


图4

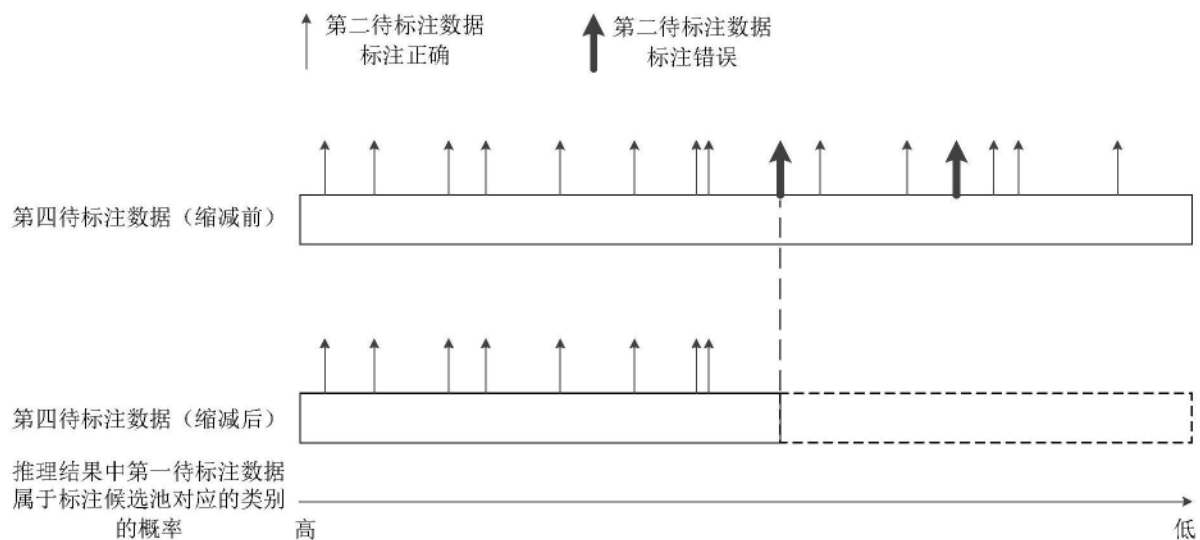


图5

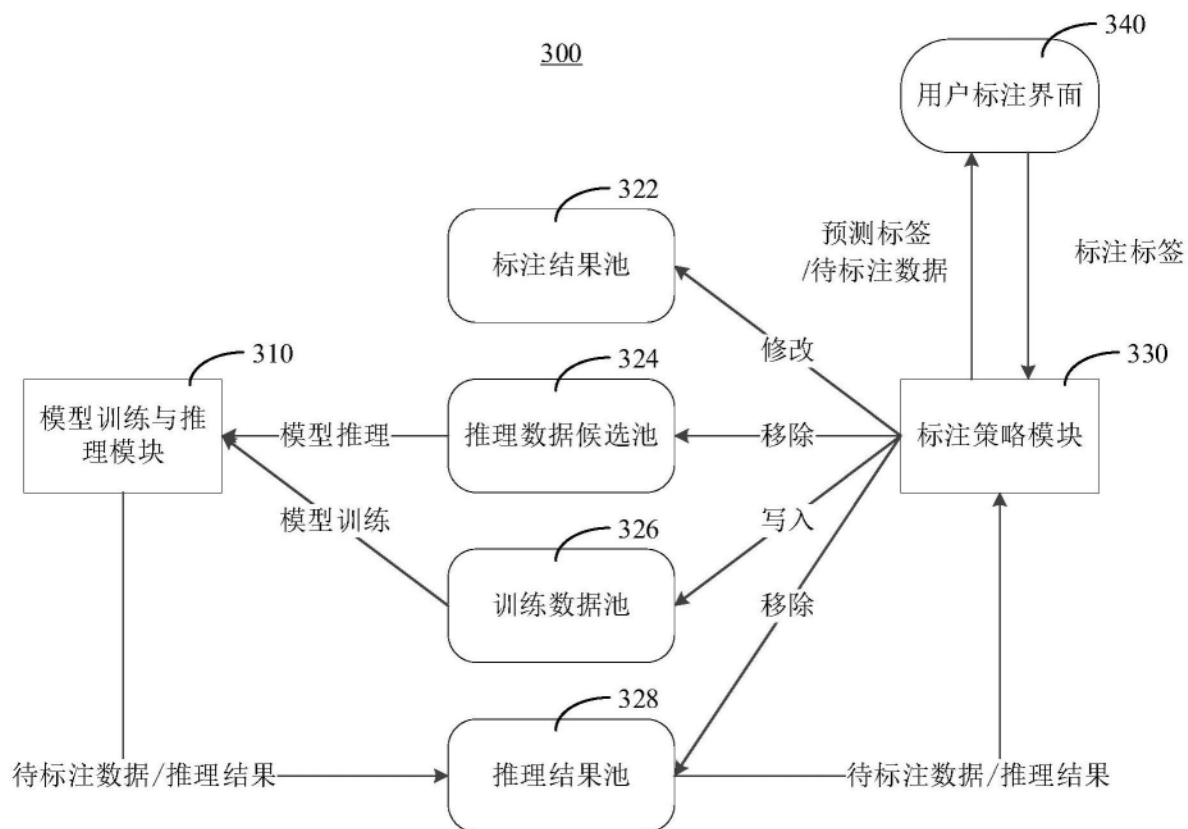


图6

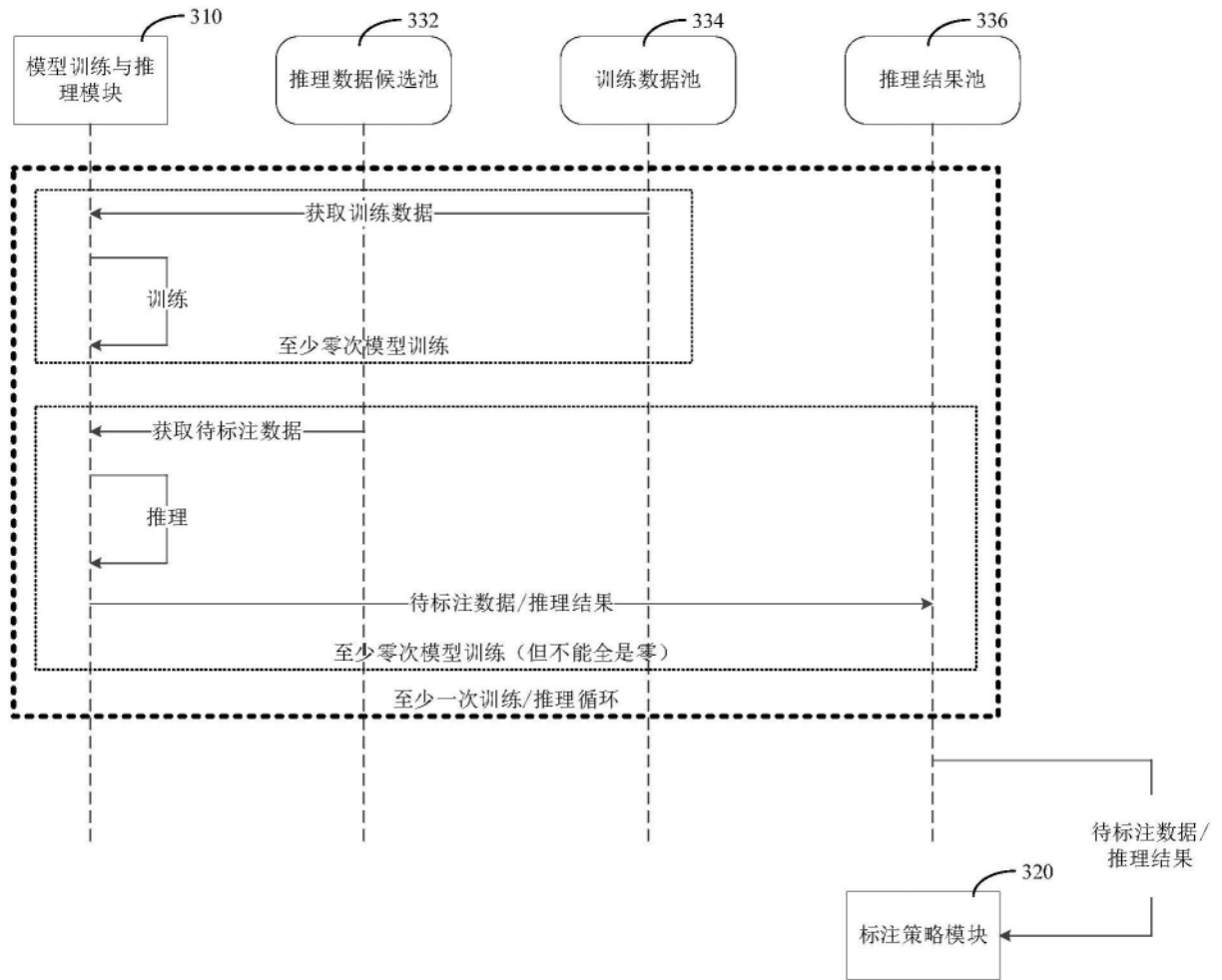


图7

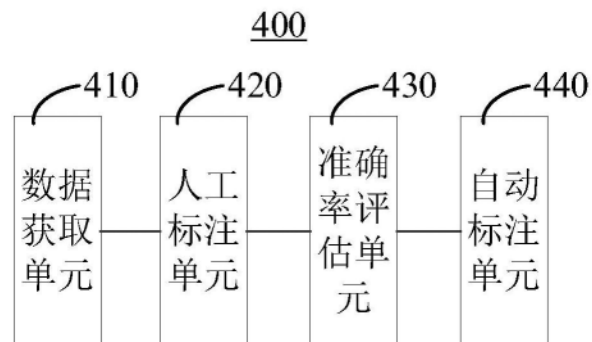


图8

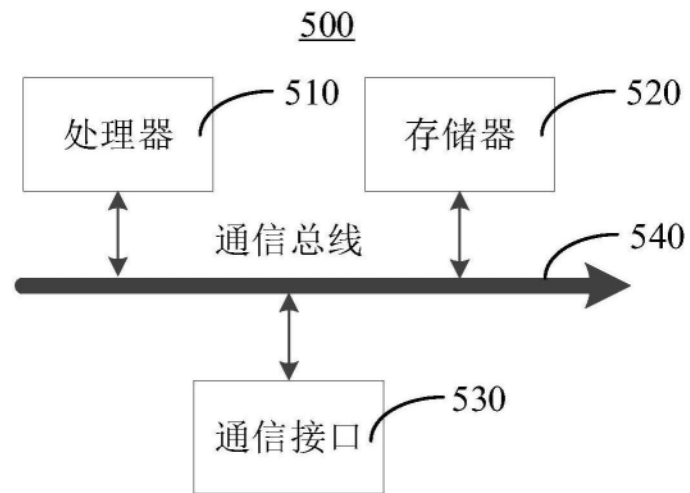


图9