(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113837216 B
(45)授权公告日 2024.05.10

(21)申请号 202110610877.2
(22)申请日 2021.06.01
(65)同一申请的已公布的文献号
申请公布号 CN 113837216 A
(43)申请公布日 2021.12.24
(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层
(72)发明人 谭维 李松南
(74)专利代理机构 深圳市联鼎知识产权代理有限公司 44232
专利代理师 王鹏健
(51)Int.Cl.
G06F 18/24(2023.01)
G06F 18/214(2023.01)

(56)对比文件
CN 111177371 A,2020.05.19
CN 112353402 A,2021.02.12
CN 112232524 A,2021.01.15
US 2020356851 A1,2020.11.12
CN 111177569 A,2020.05.19
CN 111626063 A,2020.09.04
CN 111737521 A,2020.10.02
CN 111783861 A,2020.10.16
CN 112182229 A,2021.01.05
CN 112417150 A,2021.02.26
WO 2021087985 A1,2021.05.14

审查员 丁彬

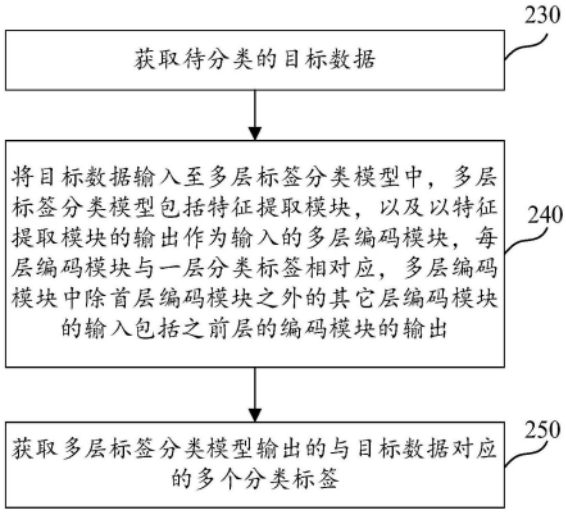
权利要求书4页 说明书17页 附图8页

(54)发明名称

数据分类方法、训练方法、装置、介质及电子设备

(57)摘要

本申请的实施例提供了一种数据分类方法、训练方法、装置、介质及电子设备。该数据分类方法包括:获取待分类的目标数据;将目标数据输入至多层标签分类模型中,多层标签分类模型包括特征提取模块,以及以特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;获取所述多层标签分类模型输出的与所述目标数据对应的多个分类标签。本申请实施例的技术方案可以利用了标签的层次结构信息,通过训练一个模型就能完成层次结构标签分类任务,可以保证预测结果中不会出现层次结构错误,并可以减少使用模型进行训练及预测时所消耗的资源。



CN 113837216 B

1.一种数据分类方法,其特征在于,包括:

获取待分类的目标数据;

将所述目标数据输入至多层标签分类模型中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;所述多层编码模块中除首层编码模块之外的其它层编码模块包括第一编码单元和第二编码单元;其中,所述第一编码单元的输入包括所述特征提取模块的输出,所述第二编码单元的输入包括所述第一编码单元的输出、首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出;所述多层标签分类模型的损失函数是通过如下方式生成的:根据所述首层编码模块的输出与样本数据的首层标签之间的差值,生成所述首层编码模块对应的损失函数;根据所述多层编码模块中除首层编码模块之外的其它层编码模块的输出与样本数据对应层级的标签之间的差值,以及所述其它层编码模块的输出与前一层编码模块的输出之间的归属关系,生成所述其它层编码模块对应的损失函数;根据所述首层编码模块对应的损失函数和所述其它层编码模块对应的损失函数,生成所述多层标签分类模型的损失函数;

获取所述多层标签分类模型输出的与所述目标数据对应的多个分类标签。

2.根据权利要求1所述的数据分类方法,其特征在于,在将所述目标数据输入至多层标签分类模型中之前,所述方法还包括:

获取样本数据集,所述样本数据集中的样本数据包括样本和与所述样本对应的多层标签;

基于所述样本数据集对所述多层标签分类模型进行训练。

3.根据权利要求2所述的数据分类方法,其特征在于,所述获取样本数据集,包括:

获取样本和与所述样本对应的指定层级标签;

基于所述指定层级标签查询标签层次表,以获取到与所述指定层级标签相关联的其它层级标签;

根据所述指定层级标签、所述其它层级标签和所述样本生成样本数据;

根据所述样本数据建立样本数据集。

4.根据权利要求1所述的数据分类方法,其特征在于,所述数据分类方法还包括:

对所述其它层编码模块所包含的第一编码单元的输出、所述首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出进行融合处理,得到所述其它层编码模块的输出结果。

5.根据权利要求4所述的数据分类方法,其特征在于,所述多层编码模块包括第一层编码模块、第二层编码模块和第三层编码模块;

所述第一层编码模块、所述第二层编码模块和所述第三层编码模块根据如下公式获得输出结果:

$$\begin{cases} pre1 = S_1(f_1) \\ pre2 = S_2\left(FC_2\,\hat{}\,(A\times f_1 + B\times f_2)\right) \\ pre3 = S_3\left(FC_3\,\hat{}\,(A\times f_1 + B\times f_2 + C\times f_3)\right) \end{cases}$$

其中,pre1、pre2和pre3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块的输出结果;S_1、S_2和S_3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块中的激活函数;f_1表示经过所述第一层编码模块处理后得到的特征;f_2表示经过所述第二层编码模块所包含的第一编码单元处理后得到的特征;f_3表示经过所述第三层编码模块所包含的第一编码单元处理后得到的特征;$FC_2\,\hat{}$表示将特征f_1和特征f_2进行融合;$FC_3\,\hat{}$表示将特征f_1、特征f_2和特征f_3进行融合;A、B和C为参数。

6.根据权利要求1所述的数据分类方法,其特征在于,所述多层编码模块包括第一层编码模块、第二层编码模块和第三层编码模块;所述多层标签分类模型的损失函数如下:

$$\begin{cases} L1 = pre1 - gt1 \\ L2 = (pre2 - gt2)\times F21 \\ L3 = (pre3 - gt3)\times F32 \end{cases}$$

其中,pre1、pre2和pre3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块的输出结果;gt1、gt2和gt3分别表示样本数据的第一层标签、第二层标签和第三层标签;L1、L2和L3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块对应的损失函数;F21和F32表示权重,若pre2归属于pre1,则F21=1,若pre2不归属于pre1,则F21>1;若pre3归属于pre2,则F32=1,若pre3不归属于pre2,则F32>1。

7.根据权利要求1至6中任一项所述的数据分类方法,其特征在于,所述多层编码模块中的各层编码模块均包括全连接层,所述全连接层的输入包括所述特征提取模块的输出。

8.根据权利要求7所述的数据分类方法,其特征在于,所述全连接层中的卷积核基于共享的权重进行特征处理。

9.根据权利要求1至6中任一项所述的数据分类方法,其特征在于,所述数据分类方法还包括:

在获取所述多层标签分类模型输出的与所述目标数据对应的多个分类标签之后,基于标签层次表,对所述多个分类标签的层级关系进行校验;

若对所述多个分类标签的层级关系校验通过,则输出所述多个分类标签。

10.一种多层标签分类模型的训练方法,其特征在于,包括:

获取样本数据集,所述样本数据集中的样本数据包括样本和与所述样本对应的多层标签;

将所述样本数据集中的样本数据输入至多层标签分类模型中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;所述多层编码模块中除首层编码模块之外的其它层编码模块包括第一编码单元和第二编码单元;其中,所述第一编码单元的输入包括所述特

征提取模块的输出,所述第二编码单元的输入包括所述第一编码单元的输出、首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出;所述多层标签分类模型的损失函数是通过如下方式生成的:根据所述首层编码模块的输出与样本数据的首层标签之间的差值,生成所述首层编码模块对应的损失函数;根据所述多层编码模块中除首层编码模块之外的其它层编码模块的输出与样本数据对应层级的标签之间的差值,以及所述其它层编码模块的输出与前一层编码模块的输出之间的归属关系,生成所述其它层编码模块对应的损失函数;根据所述首层编码模块对应的损失函数和所述其它层编码模块对应的损失函数,生成所述多层标签分类模型的损失函数;

根据所述多层标签分类模型的输出结果与所述样本对应的多层标签之间的损失值,调整所述多层标签分类模型的参数,以对所述多层标签分类模型进行训练。

11.一种数据分类装置,其特征在于,包括:

第一获取单元,用于获取待分类的目标数据;

输入单元,用于将所述目标数据输入至多层标签分类模型中,其中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应;当对所述多层标签分类模型进行训练时,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;所述多层编码模块中除首层编码模块之外的其它层编码模块包括第一编码单元和第二编码单元;其中,所述第一编码单元的输入包括所述特征提取模块的输出,所述第二编码单元的输入包括所述第一编码单元的输出、首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出;

所述输入单元还用于:根据所述首层编码模块的输出与样本数据的首层标签之间的差值,生成所述首层编码模块对应的损失函数;根据所述多层编码模块中除首层编码模块之外的其它层编码模块的输出与样本数据对应层级的标签之间的差值,以及所述其它层编码模块的输出与前一层编码模块的输出之间的归属关系,生成所述其它层编码模块对应的损失函数;根据所述首层编码模块对应的损失函数和所述其它层编码模块对应的损失函数,生成所述多层标签分类模型的损失函数;

第二获取单元,用于获取所述多层标签分类模型输出的与所述目标数据对应的多个分类标签。

12.一种多层标签分类模型的训练装置,其特征在于,包括:

样本数据集获取单元,用于获取样本数据集,所述样本数据集中的样本数据包括样本和与所述样本对应的多层标签;

样本数据输入单元,用于将所述样本数据集中的样本数据输入至多层标签分类模型中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;所述多层编码模块中除首层编码模块之外的其它层编码模块包括第一编码单元和第二编码单元;其中,所述第一编码单元的输入包括所述特征提取模块的输出,所述第二编码单元的输入包括所述第一编码单元的输出、首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块

之间层级的编码模块所包含的第一编码单元的输出;所述多层标签分类模型的损失函数是通过如下方式生成的:根据所述首层编码模块的输出与样本数据的首层标签之间的差值,生成所述首层编码模块对应的损失函数;根据所述多层编码模块中除首层编码模块之外的其它层编码模块的输出与样本数据对应层级的标签之间的差值,以及所述其它层编码模块的输出与前一层编码模块的输出之间的归属关系,生成所述其它层编码模块对应的损失函数;根据所述首层编码模块对应的损失函数和所述其它层编码模块对应的损失函数,生成所述多层标签分类模型的损失函数;

训练单元,用于根据所述多层标签分类模型的输出结果与所述样本对应的多层标签之间的损失值,调整所述多层标签分类模型的参数,以对所述多层标签分类模型进行训练。

13.一种计算机可读介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至9中任一项所述的数据分类方法。

14.一种电子设备,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序,当所述一个或多个程序被所述一个或多个处理器执行时,使得所述一个或多个处理器实现如权利要求1至9中任一项所述的数据分类方法。

数据分类方法、训练方法、装置、介质及电子设备

技术领域

[0001] 本申请涉人工智能技术领域,具体而言,涉及一种数据分类方法、多层标签分类模型的训练方法、装置、计算机可读介质及电子设备。

背景技术

[0002] 目前,对于多层级标签的分类任务,需要分别训练与多个层级标签对应的多个模型。在该方式下,一方面,会出现预测结果中的下级标签不属于上级标签的逻辑错误;另一方面,由于每一层级标签都需要一个独立的模型,因此在使用模型进行训练及预测时所消耗的资源较大。

发明内容

[0003] 本申请的实施例提供了一种数据分类方法、多层标签分类模型的训练方法、装置、计算机可读介质及电子设备,进而至少在一定程度上可以在减少标签从属关系中出现逻辑错误的缺陷,并可以减少使用模型进行训练及预测时所消耗的资源。

[0004] 本申请的其他特性和优点将通过下面的详细描述变得显然,或部分地通过本申请的实践而习得。

[0005] 根据本申请实施例的一个方面,提供了一种数据分类方法,包括:获取待分类的目标数据;将所述目标数据输入至多层标签分类模型中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;获取所述多层标签分类模型输出的与所述目标数据对应的多个分类标签。

[0006] 根据本申请实施例的一个方面,提供了一种多层标签分类模型的训练方法,包括:获取样本数据集,所述样本数据集中的样本数据包括样本和与所述样本对应的多层标签;将所述样本数据集中的样本数据输入至多层标签分类模型中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;根据所述多层标签分类模型的输出结果与所述样本对应的多层标签之间的损失值,调整所述多层标签分类模型的参数,以对所述多层标签分类模型进行训练。

[0007] 根据本申请实施例的一个方面,提供了一种数据分类装置,包括:第一获取单元,用于获取待分类的目标数据;输入单元,用于将所述目标数据输入至多层标签分类模型中,其中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应;当对所述多层标签分类模型进行训练时,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;第二获取单元,用于获取所述多层标签分类模型输出的与所述目标数

据对应的多个分类标签。

[0008] 在本申请的一些实施例中,基于前述方案,在将所述目标数据输入至多层标签分类模型中之前,所述第一获取单元还用于:获取样本数据集,所述样本数据集中的样本数据包括样本和与所述样本对应的多层标签;基于所述样本数据集对所述多层标签分类模型进行训练。

[0009] 在本申请的一些实施例中,基于前述方案,所述第一获取单元配置为:获取样本和与所述样本对应的指定层级标签;基于所述指定层级标签查询标签层次表,以获取到与所述指定层级标签相关联的其它层级标签;根据所述指定层级标签、所述其它层级标签和所述样本生成样本数据;根据所述样本数据建立样本数据集。

[0010] 在本申请的一些实施例中,基于前述方案,所述多层编码模块中除首层编码模块之外的其它层编码模块包括第一编码单元和第二编码单元;其中,所述第一编码单元的输入包括所述特征提取模块的输出,所述第二编码单元的输入包括所述第一编码单元的输出、首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出。

[0011] 在本申请的一些实施例中,基于前述方案,所述输入单元还用于:对所述其它层编码模块所包含的第一编码单元的输出、所述首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出进行融合处理,得到所述其它层编码模块的输出结果。

[0012] 在本申请的一些实施例中,基于前述方案,所述多层编码模块包括第一层编码模块、第二层编码模块和第三层编码模块;所述第一层编码模块、所述第二层编码模块和所述第三层编码模块根据如下公式获得输出结果:

[0013]
$$\begin{cases} pre1 = S_1(f_1) \\ pre2 = S_2\left(FC_2\hat{}\left(A\times f_1 + B\times f_2\right)\right) \\ pre3 = S_3\left(FC_3\hat{}\left(A\times f_1 + B\times f_2 + C\times f_3\right)\right) \end{cases}$$

[0014] 其中,pre1、pre2和pre3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块的输出结果;S_1、S_2和S_3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块中的激活函数;f_1表示经过所述第一层编码模块处理后得到的特征;f_2表示经过所述第二层编码模块所包含的第一编码单元处理后得到的特征;f_3表示经过所述第三层编码模块所包含的第一编码单元处理后得到的特征;$FC_2\hat{}$表示将特征f_1和特征f_2进行融合;$FC_3\hat{}$表示将特征f_1、特征f_2和特征f_3进行融合;A、B和C为参数。

[0015] 在本申请的一些实施例中,基于前述方案,所述输入单元还用于:根据所述首层编码模块的输出与样本数据的首层标签之间的差值,生成所述首层编码模块对应的损失函数;根据所述多层编码模块中除首层编码模块之外的其它层编码模块的输出与样本数据对应层级的标签之间的差值,以及所述其它层编码模块的输出与前一层编码模块的输出之间的归属关系,生成所述其它层编码模块对应的损失函数;根据所述首层编码模块对应的损失函数和所述其它层编码模块对应的损失函数,生成所述多层标签分类模型的损失函数。

[0016] 在本申请的一些实施例中,基于前述方案,所述多层编码模块包括第一层编码模

块、第二层编码模块和第三层编码模块;所述多层标签分类模型的损失函数如下:

[0017] $$\begin{cases} L1 = pre1 - gt1 \\ L2 = (pre2 - gt2) \times F21 \\ L3 = (pre3 - gt3) \times F32 \end{cases}$$

[0018] 其中,pre1、pre2和pre3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块的输出结果;gt1、gt2和gt3分别表示样本数据的第一层标签、第二层标签和第三层标签;L1、L2和L3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块对应的损失函数;F21和F32表示权重,若pre2归属于pre1,则F21=1,若pre2不归属于pre1,则F21>1;若pre3归属于pre2,则F32=1,若pre3不归属于pre2,则F32>1。

[0019] 在本申请的一些实施例中,基于前述方案,所述多层编码模块中的各层编码模块均包括全连接层,所述全连接层的输入包括所述特征提取模块的输出。

[0020] 在本申请的一些实施例中,基于前述方案,所述全连接层中的卷积核基于共享的权重进行特征处理。

[0021] 在本申请的一些实施例中,基于前述方案,所述输入单元还用于:在获取所述多层标签分类模型输出的与所述目标数据对应的多个分类标签之后,基于标签层次表,对所述多个分类标签的层级关系进行校验;若对所述多个分类标签的层级关系校验通过,则输出所述多个分类标签。

[0022] 根据本申请实施例的一个方面,提供了一种多层标签分类模型的训练装置,包括:样本数据集获取单元,用于获取样本数据集,所述样本数据集中的样本数据包括样本和与所述样本对应的多层标签;样本数据输入单元,用于将所述样本数据集中的样本数据输入至多层标签分类模型中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;训练单元,用于根据所述多层标签分类模型的输出结果与所述样本对应的多层标签之间的损失值,调整所述多层标签分类模型的参数,以对所述多层标签分类模型进行训练。

[0023] 根据本申请实施例的一个方面,提供了一种计算机可读介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现如上述实施例中所述的数据分类方法。

[0024] 根据本申请实施例的一个方面,提供了一种电子设备,包括:一个或多个处理器;存储装置,用于存储一个或多个程序,当所述一个或多个程序被所述一个或多个处理器执行时,使得所述一个或多个处理器实现如上述实施例中所述的数据分类方法。

[0025] 在本申请的一些实施例所提供的技术方案中,利用了构建的多层标签分类模型实现了数据分类,并且多层标签分类模型中包括特征提取模块和多层编码模块,其中多层编码模块以特征提取模块的输出作为输入,每层编码模块与一层分类标签相对应,并且除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出,因此整个多层标签分类模型利用了标签的层次结构信息,具有多个层级标签的数据可用于同一模型的训练,整个多层标签分类模型也能够根据一个数据的输入而一次性输出对应的多层级标签,通过训练一个模型就能完成层次结构标签分类任务,可以保证预测结果中不会出现层次结构错误,并可以减少使用模型进行训练及预测时所消耗的资源。

[0026] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本申请。

附图说明

[0027] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本申请的实施例,并与说明书一起用于解释本申请的原理。显而易见地,下面描述中的附图仅仅是本申请的一些实施例,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的附图。在附图中:

[0028] 图1示出了可以应用本申请实施例的技术方案的示例性系统架构的示意图;

[0029] 图2示出了根据本申请的一个实施例的数据分类方法的流程图;

[0030] 图3示出了根据本申请的一个实施例的图2中步骤230之前步骤的流程图;

[0031] 图4示出了根据本申请的一个实施例的用于获取样本数据的页面的示意图;

[0032] 图5示出了根据本申请的一个实施例的样本数据集的获取过程的流程图;

[0033] 图6示出了根据本申请的一个实施例的基于一种方式训练多层标签分类模型的模型架构示意图;

[0034] 图7示出了根据本申请的一个实施例的基于另一种方式训练多层标签分类模型的模型架构示意图;

[0035] 图8示出了根据本申请的一个实施例的对多层标签分类模型的输出结果进行校验的流程图;

[0036] 图9示出了根据本申请的一个实施例的多层标签分类模型的训练方法的流程图;

[0037] 图10示出了根据本申请的一个实施例的数据分类装置的框图;

[0038] 图11示出了根据本申请的一个实施例的多层标签分类模型的训练装置的框图;

[0039] 图12示出了适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

具体实施方式

[0040] 现在将参考附图更全面地描述示例实施方式。然而,示例实施方式能够以多种形式实施,且不应被理解为限于在此阐述的范例;相反,提供这些实施方式使得本申请将更加全面和完整,并将示例实施方式的构思全面地传达给本领域的技术人员。

[0041] 此外,所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施例中。在下面的描述中,提供许多具体细节从而给出对本申请的实施例的充分理解。然而,本领域技术人员将意识到,可以实践本申请的技术方案而没有特定细节中的一个或更多,或者可以采用其它的方法、组元、装置、步骤等。在其它情况下,不详细示出或描述公知方法、装置、实现或者操作以避免模糊本申请的各方面。

[0042] 附图中所示的方框图仅仅是功能实体,不一定必须与物理上独立的实体相对应。即,可以采用软件形式来实现这些功能实体,或在一个或多个硬件模块或集成电路中实现这些功能实体,或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0043] 附图中所示的流程图仅是示例性说明,不是必须包括所有的内容和操作/步骤,也不是必须按所描述的顺序执行。例如,有的操作/步骤还可以分解,而有的操作/步骤可以合并或部分合并,因此实际执行的顺序有可能根据实际情况改变。

[0044] 分类是机器学习和人工智能领域的重要任务之一,分类模型在很多场景下都存在着广泛的应用。

[0045] 通常利用包含标签的样本数据进行分类模型的训练,由于样本数据中包含标签,因此分类模型实际上是通过监督学习进行训练的。然而,就像同一事物可以属于不同的类别一样,同一对象可能同时与多层级标签相对应。多层级标签是树状结构的标签体系,其具有层次结构,后一级标签属于前一级,例如,狗-中型犬-哈士奇便为具有层次结构的多层级标签,其中,哈士奇这个标签属于中型犬这一标签,中型犬这个标签属于狗这一标签。

[0046] 在相关技术中,只能利用同一层级标签的数据训练得到对应的分类模型,该分类模型也只能输出该层级的标签,因此,对于多层级标签的数据,只能训练多个分类模型,每个分类模型对应一个层级标签。以三层级标签分类任务为例,在基于相关技术训练完成该任务的模型时,通常独立训练三个模型,这三个模型分别对应三个层级标签。

[0047] 基于上述相关技术中的技术方案虽然也能获得数据对应的多层级标签,但至少存在以下缺陷:

[0048] 第一,每一层级的标签对应的模型是独立进行训练的,浪费了层次结构的信息;

[0049] 第二,在使用模型进行预测时,由于缺失层次结构信息,会出现预测结果里下级标签不属于上级标签的逻辑错误;

[0050] 第三,每一层级的标签都需要一个独立的模型进行预测,训练及预测的资源消耗较大。

[0051] 为此,本申请首先提供了一种数据分类方法。本申请实施例提供的数据分类方法可以克服上述缺陷。本申请实施例提供的数据分类方法所能够分类的数据包括但不限于文本数据、图片数据、音频数据、视频数据等,因而,本申请实施例提供的数据分类方法能够应用于文本分类、图片分类、音频分类、视频分类等分类任务中。

[0052] 图1示出了可以应用本申请实施例的技术方案的示例性系统架构的示意图。

[0053] 如图1所示,系统架构可以包括终端设备(如图1中所示智能手机101、平板电脑102和便携式计算机103中的一种或多种,当然也可以是台式计算机等等)、网络104和服务器105。网络104用以在终端设备和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线通信链路、无线通信链路等等。

[0054] 应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器。比如服务器105可以是多个服务器组成的服务器集群等。

[0055] 在本申请的一个实施例中,各终端设备中至少一部分终端设备的用户为数据标注员,各终端设备上部署有应用程序,数据标注员通过应用程序可以获得来自服务器105的待标注样本数据,然后,数据标注员通过操作应用程序可以为待标注样本数据标注上对应的多层标签,最后,待标注样本数据和对应的多层标签可以通过网络104发送至服务器105上,由此,服务器105可以获得包含多层标签的样本数据。

[0056] 在本申请的一个实施例中,服务器105上部署有多层标签分类模型,服务器105可以利用已获得的包含多层标签的样本数据对该多层标签分类模型进行训练,经过训练的多层标签分类模型能够针对一个数据输出对应的多层标签。

[0057] 在本申请的一个实施例中,各终端设备中至少一部分终端设备能够向服务器105

发送需要进行分类的数据,服务器105在获得了这些数据之后,可以利用经过训练的多层标签分类模型对这些数据进行分类,并输出与每一数据对应的一个或多个分类标签,这些分类标签具有层次或层级关系;服务器105上的多层标签分类模型在输出分类标签之后,可以通过网络104将分类标签发送至对应的终端设备。

[0058] 在本申请的一个实施例中,待标注样本数据为图片数据,待标注样本数据标注对应的多层标签是图片数据中记录的物体的多个类别;经过训练的多层标签分类模型能够根据图片数据的输入而输出该图片数据中记录的物体的各个类别。

[0059] 需要说明的是,虽然本申请实施例中样本数据和待分类的数据均来自实施终端之外的终端,但在本申请的其他实施例中,样本数据和待分类的数据可以均存储于本地;虽然本申请实施例中,待标注样本数据为图片数据,但在本申请的其他实施例或者具体应用中,待标注样本数据可以为文本数据、视频数据等其他类型的数据。本申请实施例对此不作任何限定,本申请的保护范围也不应因此而受到任何限制。

[0060] 并且,易于理解,本申请实施例所提供的数据分类方法一般由服务器105执行,相应地,数据分类装置一般设置于服务器105中。但是,在本申请的其它实施例中,终端设备也可以与服务器具有相似的功能,从而执行本申请实施例所提供的数据分类的方案。

[0061] 本申请实施例可以由服务器对来自终端的数据进行分类。服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN(Content Delivery Network,内容分发网络)、以及大数据和人工智能平台等基础云计算服务的云服务器。终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等,但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接,本申请在此不做限制。

[0062] 本申请实施例可以应用于云计算技术中。云计算(cloud computing)是一种计算模式,它将计算任务分布在大量计算机构成的资源池上,使各种应用系统能够根据需要获取计算力、存储空间和信息服务。提供资源的网络被称为“云”。“云”中的资源在使用者看来是可以无限扩展的,并且可以随时获取,按需使用,随时扩展,按使用付费。

[0063] 作为云计算的基础能力提供商,会建立云计算资源池(简称云平台,一般称为IaaS(Infrastructure as a Service,基础设施即服务))平台,在资源池中部署多种类型的虚拟资源,供外部客户选择使用。云计算资源池中主要包括:计算设备(为虚拟化机器,包含操作系统)、存储设备、网络设备。

[0064] 按照逻辑功能划分,在IaaS(Infrastructure as a Service,基础设施即服务)层上可以部署PaaS(Platform as a Service,平台即服务)层,PaaS层之上再部署SaaS(Software as a Service,软件即服务)层,也可以直接将SaaS部署在IaaS上。PaaS为软件运行的平台,如数据库、web容器等。SaaS为各式各样的业务软件,如web门户网站、短信群发器等。一般来说,SaaS和PaaS相对于IaaS是上层。

[0065] 以下对本申请实施例的技术方案的实现细节进行详细阐述:

[0066] 图2示出了根据本申请的一个实施例的数据分类方法的流程图,该数据分类方法可以由具有计算功能的设备来执行,比如可以是图1中所示的服务器105。参照图2所示,该数据分类方法至少包括以下步骤:

[0067] 在步骤230中,获取待分类的目标数据。

[0068] 目标数据可以是文本数据、图片数据、音频数据、视频数据等各种类型的数据。当待分类的目标数据为图片数据时,图片数据中记录了物体,对图片数据进行分类就是确定图片数据中记录的物体所属的类别。

[0069] 例如,图片数据可以为哈士奇的照片,那么对该图片数据进行分类时,可以将该图片数据分类为狗,因此,对图片数据进行分类也相当于对图片数据进行识别,即识别图片数据中物体的类别。

[0070] 本申请实施例能够为目标数据输出多层标签,并且与同一数据对应的多层标签之间具有归属关系。

[0071] 比如,与一个包含哈士奇的图片数据对应的多层标签可以分别为狗-中型犬-哈士奇,其中,每一层的标签归属于与该标签相邻且位于该标签之前的一层标签,比如,哈士奇归属于中型犬,中型犬又归于狗。基于此,多层标签中一层标签越靠前,那么该层标签的范围就越大,并包含该层标签之后的所有层标签,因此,每一层的标签实际上是归属于位于该标签之前的标签的,比如,哈士奇既归属于中型犬,中型犬又归于狗。

[0072] 多层标签是基于树状结构的标签体系建立起来的,比如,狗、猫等可以分为一层,大型犬、中型犬等可以分为一层,京巴、哈士奇等又可以分为一层,这样便形成了树状结构。

[0073] 在下文中,若非特别指明,在用层、级、层级、层次描述标签时都指代同样的意思,即表示能够覆盖不同范围的标签。

[0074] 在步骤240中,将目标数据输入至多层标签分类模型中,多层标签分类模型包括特征提取模块,以及以特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出。

[0075] 多层标签分类模型中无论是特征提取模块还是多层编码模块,均是利用神经网络建立起来的。

[0076] 在本申请的一个实施例中,特征提取模块为预训练模型。

[0077] 预训练模型具体可以为Bert等模型,预训练模型预先在大规模数据集上经过了训练,因此,在基于预训练模型构建多层标签分类模型时,不仅能够为所构建的多层标签分类模型引入更多信息,还能够加速多层标签分类模型的训练,从而能够节约训练模型消耗的资源以及训练成本。

[0078] 多层编码模块与特征提取模块相连,因此,多层编码模块能够以特征提取模块的输出作为输入。

[0079] 由于每层编码模块与一层分类标签相对应,因此,本申请实施例的多层标签分类模型能够为同一数据输出多层分类标签。

[0080] 比如,多层编码模块可以分别为1层编码模块、2层编码模块和3层编码模块,其中,1层编码模块为位于最前的首层编码模块,紧随首层编码模块之后的是2层编码模块,3层编码模块是位于最后的编码模块。此时,多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出是指:2层编码模块的输入包括1层编码模块的输出,3层编码模块的输入包括2层编码模块的输出。

[0081] 由于多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层

的编码模块的输出,并且每层编码模块与一层分类标签相对应,因此,在多层标签分类模型的训练和使用过程中都利用了标签的层次结构信息,可以保证预测结果中不会出现层次结构错误。

[0082] 多层标签分类模型需要经过训练才能用于对目标数据进行分类。下面,将介绍多层标签分类模型的训练过程。

[0083] 图3示出了根据本申请的一个实施例的图2中步骤230之前步骤的流程图。请参见图3,在步骤230之前可以包括以下步骤:

[0084] 在步骤210中,获取样本数据集,样本数据集中的样本数据包括样本和与样本对应的多层标签。

[0085] 样本数据集中包括多个样本数据。

[0086] 各样本对应的多层标签中标签的层数是相同的,并且,样本对应的多层标签中标签的层数等于多层标签分类模型的多层编码模块中编码模块的层数,同时,每层标签仅与多层编码模块中的一层编码模块相对应,因此,利用这样的样本数据便可以对能够输出多层标签的多层标签分类模型进行训练。

[0087] 样本数据中的样本以及与样本对应的多层标签可以是通过多种方式获得的。比如,可以通过基于数据挖掘的方式获得,可以通过利用爬虫从互联网中爬取的方式获得,还可以通过人工标注的方式获得。

[0088] 图4示出了根据本申请的一个实施例的用于获取样本数据的页面的示意图。请参见图4,该页面可以是提供给数据标注员,用以标注数据的页面。具体地,该页面在左侧显示了包含狗的照片,该照片即为样本,该页面的右侧列出了三个文本录入框,通过这些文本录入框,数据标注员可以录入与该照片对应的多层标签,实现对样本进行标签的标注,可以看到,已经录入的一层标签为狗。这样当多层标签录入完毕,便可以得到一个样本数据。该页面中还显示了“上一个”和“下一个”按钮,当数据标注员点击了“上一个”按钮后,便可以对已为其他样本标注的各个标签进行编辑,当数据标注员点击了“下一个”按钮后,便可以对尚未标注的其他样本进行标签的标注。图4所示的页面可以显示在数据标注员所使用的终端上,包含最后一个样本的页面中可以显示有“提交”按钮,当数据标注员对所有样本标注完毕,通过点击该“提交”按钮,可以将所有样本发送至本申请实施例的实施终端,从而获得多个样本数据。

[0089] 图5示出了根据本申请的一个实施例的样本数据集的获取过程的流程图。请参阅图5,样本数据集的获取过程可以包括如下步骤:

[0090] 在步骤510中,获取样本和与样本对应的指定层级标签。

[0091] 在本申请的一个实施例中,获取样本和与样本对应的指定层级标签,包括:获取样本和与样本对应的最后一层标签。

[0092] 与样本对应的最后一层标签归属于与样本对应的其他层标签。

[0093] 最后一层标签是覆盖范围最小的标签,也是与样本最密切相关的标签,比如,一个样本对应的多层标签可以分别为狗-中型犬-哈士奇,其中,哈士奇便是最后一层标签。由于本步骤中只能够获得与样本对应的最后一层标签,因此,还需要获取与样本对应的其他层标签以满足训练多层标签分类模型的需要。

[0094] 在步骤520中,基于指定层级标签查询标签层次表,以获取到与指定层级标签相关

联的其它层级标签。

[0095] 标签层次表可以根据人工经验建立。

[0096]

一层标签	二层标签	三层标签
狗	中型犬	哈士奇
狗	大型犬	苏格兰牧羊犬
狗	大型犬	拉布拉多猎犬
狗	小型犬	柴犬

[0097] 表1

[0098] 表1示意性地示出了标签层次表。在表1中,位于同一行的各层标签相对应,三层标签是最后一层标签,其归属于对应的二层标签,二层标签又归属于对应的一层标签。因而,只需要获得了三层标签,就可以通过查找该标签层次表而确定出对应的一层标签和二层标签,即能够根据指定层级标签获取到与指定层级标签相关联的其它层级标签。

[0099] 在步骤530中,根据指定层级标签、其它层级标签和样本生成样本数据。

[0100] 指定层级标签和其它层级标签可以组成与样本对应的多层标签,将多层标签与样本相组合,可以生成样本数据。

[0101] 在步骤540中,根据样本数据建立样本数据集。

[0102] 在获得了多个样本数据之后,利用多个样本数据构建样本数据集。

[0103] 在本申请实施例中,用户在获得了样本之后,可以仅为该样本标注指定层级标签,如最后一层标签,而该样本对应的其他层级标签可以查表而自动获得。因此,本申请实施例可以大大提高样本对应的多层标签的获取效率,从而提高样本数据和样本数据集的生成效率。

[0104] 在本申请的一个实施例中,多层编码模块中除首层编码模块之外的其它层编码模块包括第一编码单元和第二编码单元;其中,第一编码单元的输入包括特征提取模块的输出,第二编码单元的输入包括第一编码单元的输出、首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出。

[0105] 图6示出了根据本申请的一个实施例的基于一种方式训练多层标签分类模型的模型架构示意图。请参阅图6,该模型架构包括特征提取模块610以及多层编码模块,特征提取模块610是单一backbone,单一backbone为用于提取特征的主干网络,比如可以是预训练模型。多层编码模块分别为第一层编码模块620、第二层编码模块630以及第三层编码模块640,其中,第一层编码模块620为首层编码模块。第二层编码模块630和第三层编码模块640中均包括两个编码单元,其中,靠近特征提取模块的编码单元是第一编码单元,另一个编码单元则是第二编码单元。通过图6可以看到,特征提取模块的输出直接进入第一编码单元,第二层编码模块630中第二编码单元的输入包括第二层编码模块630中第一编码单元的输出和第一层编码模块620的输出;由于第一层编码模块620和第三层编码模块640之间还包括第二层编码模块630,因此,第三层编码模块640中第二编码单元的输入除了包括第三层编码模块640中第一编码单元的输出和第一层编码模块620的输出之外,还包括第二层编码模块630中第一编码单元的输出。

[0106] 在本申请的一个实施例中,数据分类方法还包括:对其它层编码模块所包含的第一编码单元的输出、首层编码模块的输出,以及处于首层编码模块与其它层编码模块之间

层级的编码模块所包含的第一编码单元的输出进行融合处理,得到其它层编码模块的输出结果。

[0107] 请继续参见图6,假如其它层编码模块为第三层编码模块640,因此,第三层编码模块640是通过对第三层编码模块640中第一编码单元的输出、第一层编码模块620的输出以及处于第一层编码模块620和第三层编码模块640之间的第二层编码模块630中第一编码单元的输出进行融合处理而得到输出结果的。

[0108] 在本申请的一个实施例中,多层编码模块包括第一层编码模块、第二层编码模块和第三层编码模块;第一层编码模块、第二层编码模块和第三层编码模块根据如下公式获得输出结果:

[0109] $$\begin{cases} pre1 = S_1(f_1) \\ pre2 = S_2\left(FC_2\hat{}\left(A\times f_1 + B\times f_2\right)\right) \\ pre3 = S_3\left(FC_3\hat{}\left(A\times f_1 + B\times f_2 + C\times f_3\right)\right) \end{cases}$$

[0110] 其中,pre1、pre2和pre3分别表示第一层编码模块、第二层编码模块和第三层编码模块的输出结果;S_1、S_2和S_3分别表示第一层编码模块、第二层编码模块和第三层编码模块中的激活函数;f_1表示经过第一层编码模块处理后得到的特征;f_2表示经过第二层编码模块所包含的第一编码单元处理后得到的特征;f_3表示经过第三层编码模块所包含的第一编码单元处理后得到的特征;$FC_2\hat{}$表示将特征f_1和特征f_2进行融合;$FC_3\hat{}$表示将特征f_1、特征f_2和特征f_3进行融合;A、B和C为参数。

[0111] 可以对多层编码模块中的参数施加一定的约束,比如,可以固定设置A+B=1。

[0112] 请继续参见图6,从上到下依次排列的是第一层编码模块620、第二层编码模块630以及第三层编码模块640,它们的输出分别是pre1、pre2和pre3。

[0113] 在本申请的一个实施例中,多层编码模块中的各层编码模块均包括全连接层,全连接层的输入包括特征提取模块的输出。

[0114] 首层编码模块包括全连接层,其它层编码模块中的第一编码单元和第二编码单元中可以均包括全连接层。

[0115] 具体地,在上述公式中,$f_1=FC_1(X)$,$f_2=FC_2(X)$,$f_3=FC_3(X)$,其中,X为特征提取模块的输出,f_1为第一层编码模块中的全连接层,f_2为第二层编码模块的第一编码单元中的全连接层,f_3为第三层编码模块的第一编码单元中的全连接层,并且,$FC_2\hat{}$可以为第二层编码模块的第二编码单元中的全连接层,$FC_3\hat{}$可以为第三层编码模块的第二编码单元中的全连接层,S_2和S_3可以为位于第二编码单元中的激活函数层,且该激活函数层可以位于所在的第二编码单元中的全连接层之后。

[0116] 虽然本申请实施例中,第二编码单元包括全连接层和激活函数层,但易于理解的是,第二编码单元中还可以包括池化层和Softmax层等其他层,这样可以使模型更完整,并可以减少网络复杂度。具体而言,第二编码单元中从靠近第一编码单元到远离第一编码单元的各层结构可以依次为:全连接层、池化层、激活函数层以及Softmax层,其中,激活函数层可以ReLU激活函数。

[0117] 请继续参见图3,在步骤220中,基于样本数据集对多层标签分类模型进行训练。

[0118] 在本申请的一个实施例中,基于样本数据集对多层标签分类模型进行训练的步骤,包括:

[0119] 按照预定比例将样本数据集分为训练数据集和测试数据集;基于训练数据集对多层标签分类模型进行训练。

[0120] 在对多层标签分类模型训练完毕之后,可以利用测试数据集对多层标签分类模型进行测试。

[0121] 在本申请的一个实施例中,全连接层中的卷积核基于共享的权重进行特征处理。

[0122] 具体来说,对一个特征矩阵的各个部分利用具有相同权重的卷积核进行特征过滤。请继续参阅图6,第一层编码模块620、第二层编码模块630以及第三层编码模块640所在的部分便可以是共享权重的。

[0123] 在本申请的一个实施例中,多层编码模块中除首层编码模块之外的其他层编码模块依据与该层编码模块相邻且位于该层编码模块之前的一层编码模块的输出结果进行训练。

[0124] 在本申请的一个实施例中,数据分类方法还包括:根据首层编码模块的输出与样本数据的首层标签之间的差值,生成首层编码模块对应的损失函数;根据多层编码模块中除首层编码模块之外的其它层编码模块的输出与样本数据对应层级的标签之间的差值,以及其它层编码模块的输出与前一层编码模块的输出之间的归属关系,生成其它层编码模块对应的损失函数;根据首层编码模块对应的损失函数和其它层编码模块对应的损失函数,生成多层标签分类模型的损失函数。

[0125] 在本申请的一个实施例中,多层编码模块包括第一层编码模块、第二层编码模块和第三层编码模块;数据分类方法还包括:根据第一层编码模块的输出与样本数据的第一层标签之间的差值,生成第一层编码模块对应的损失函数;根据第二层编码模块的输出与样本数据的第二层标签之间的差值,以及第二层编码模块的输出与第一层编码模块的输出之间的归属关系,生成第二层编码模块对应的损失函数;根据第三层编码模块的输出与样本数据的第三层标签之间的差值,以及第三层编码模块的输出与第二层编码模块的输出之间的归属关系,生成第三层编码模块对应的损失函数;根据第一层编码模块、第二层编码模块和第三层编码模块分别对应的损失函数,生成多层标签分类模型的损失函数。

[0126] 在本申请的一个实施例中,多层编码模块包括第一层编码模块、第二层编码模块和第三层编码模块;

[0127] 多层标签分类模型的损失函数如下:

[0128]
$$\begin{cases} L1 = pre1 - gt1 \\ L2 = \left(pre2 - gt2\right) \times F21 \\ L3 = \left(pre3 - gt3\right) \times F32 \end{cases}$$

[0129] 其中,pre1、pre2和pre3分别表示第一层编码模块、第二层编码模块和第三层编码模块的输出结果;gt1、gt2和gt3分别表示样本数据的第一层标签、第二层标签和第三层标签;L1、L2和L3分别表示第一层编码模块、第二层编码模块和第三层编码模块对应的损失函数;F21和F32表示权重,若pre2归属于pre1,则F21=1,若pre2不归属于pre1,则F21>1;若pre3归属于pre2,则F32=1,若pre3不归属于pre2,则F32>1。

[0130] 第一层标签是覆盖范围最大的标签,第二层标签是与第一层标签的层级相邻且归属于第一层标签的标签,第三层标签是与第二层标签的层级相邻且归属于第二层标签的标签。比如,第一层标签可以是狗,第二层标签可以是中型犬,第三层标签可以是哈士奇。

[0131] 请继续参见图6,可以看到,在层次loss部分,第一层编码模块620、第二层编码模块630以及第三层编码模块640分别对应一个损失函数,在图6中,F21被表示为F(pre2/pre1),F32则被表示为F(pre3/pre2)。

[0132] 通过不断调整模型的参数,使损失函数最小化,进而完成多层标签分类模型的训练,其中,A、B和C等参数也可以在模型的训练过程中进行调整。

[0133] 对通过上述公式可以看到,多层标签分类模型的多层编码模块中每层编码模块的训练是依赖于其它层编码模块的输出而进行的,并且每层编码模块还根据对应层级的标签进行训练,因此,本申请实施例的多层标签分类模型的训练有效地利用了多层标签的层次结构信息,可以保证预测结果中不会出现层次结构错误。

[0134] 编码模块的层次关系也可以采用其他的方式进行定义。图7示出了根据本申请的一个实施例的基于另一种方式训练多层标签分类模型的模型架构示意图。请参见图7,第一层编码模块740、第二层编码模块730以及第三层编码模块720均与特征提取模块710相连。比如,pre1、pre2和pre3可以分别表示第一层编码模块740、第二层编码模块730和第三层编码模块720的输出结果;gt1、gt2和gt3分别表示样本数据的第一层标签、第二层标签和第三层标签。此时,F(pre2/pre1)和F(pre3/pre2)可以表示权重,F(pre2/pre1)的含义可以是:若pre1包括pre2(pre1包括pre2表示pre2是pre1的下级标签),则F(pre2/pre1)=1,若pre1不包括pre2,则F(pre2/pre1)>1;F(pre3/pre2)的含义可以是:若pre2包括pre3(pre2包括pre3表示pre3是pre2的下级标签),则F(pre3/pre2)=1,若pre2不包括pre3,则F(pre3/pre2)>1。

[0135] 在本申请的一个实施例中,数据分类方法还包括:从前向后依次对多层标签分类模型的多层编码模块进行训练。

[0136] 在前述实施例中,由于每层编码模块是根据该层编码模块的输出与之前层编码模块的输出之间的归属关系进行训练的,因此,若之前层编码模块的输出的准确性较低,会直接导致该层编码模块的训练受到该误差的影响,从而使训练效果差,也降低了训练进度。在本申请实施例中,通过从前向后依次对编码模块进行训练,能够提高训练效果和训练速度。

[0137] 在本申请的一个实施例中,多层编码模块中每层编码模块的训练次数相同。

[0138] 请继续参照图2,在步骤250中,获取多层标签分类模型输出的与目标数据对应的多个分类标签。

[0139] 与目标数据对应的多个分类标签即多层标签,各层标签之间具有归属关系,类似于狗-中型犬-哈士奇的形式。

[0140] 图8示出了根据本申请的一个实施例的对多层标签分类模型的输出结果进行校验的流程图。请参阅图8,包括以下步骤:

[0141] 在步骤810中,在获取多层标签分类模型输出的与目标数据对应的多个分类标签之后,基于标签层次表,对多个分类标签的层级关系进行校验。

[0142] 标签层次表保存了分类标签的对应关系,因此,通过标签层次表可以确定多个分类标签是否相对应。

[0143] 在步骤820中,若对多个分类标签的层级关系校验通过,则输出多个分类标签。

[0144] 当通过查询标签层次表,可以确认多层标签分类模型输出的多个分类标签是相互对应时,就校验通过。

[0145] 在本申请实施例中,通过利用标签层次表对多层标签分类模型的输出结果进行校验,并在校验通过后,再输出分类结果,从而进一步保证了分类结果的准确性。

[0146] 本申请实施例还提供了一种多层标签分类模型的训练方法。

[0147] 图9示出了根据本申请的一个实施例的多层标签分类模型的训练方法的流程图。请参阅图9,可以包括以下步骤:

[0148] 在步骤910中,获取样本数据集,样本数据集中的样本数据包括样本和与样本对应的多层标签。

[0149] 样本数据集包括多个样本数据。

[0150] 在步骤920中,将样本数据集中的样本数据输入至多层标签分类模型中,多层标签分类模型包括特征提取模块,以及以特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出。

[0151] 多层标签分类模型可以采用图6所示的模型架构,有关多层标签分类模型的具体细节,请参见上述实施例的方案。

[0152] 在步骤930中,根据多层标签分类模型的输出结果与样本对应的多层标签之间的损失值,调整多层标签分类模型的参数,以对多层标签分类模型进行训练。

[0153] 通过调整多层标签分类模型的参数,使损失值最小化,从而使经过训练的多层标签分类模型能够根据数据的输入而准确输出对应的多层标签。

[0154] 根据本申请实施例的技术方案可以端到端地训练多层标签分类模型,可以获得更高的预测准确率,同时保证预测结果的层次结构正确性。

[0155] 以下介绍本申请的装置实施例,可以用于执行本申请上述实施例中的数据分类方法。对于本申请装置实施例中未披露的细节,请参照本申请上述的数据分类方法的实施例。

[0156] 图10示出了根据本申请的一个实施例的数据分类装置的框图。

[0157] 参照图10所示,根据本申请的一个实施例的数据分类装置1000,包括:第一获取单元1010、输入单元1020和第二获取单元1030。

[0158] 其中,第一获取单元1010用于获取待分类的目标数据;输入单元1020用于将所述目标数据输入至多层标签分类模型中,其中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应;当对所述多层标签分类模型进行训练时,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;第二获取单元1030用于获取所述多层标签分类模型输出的与所述目标数据对应的多个分类标签。

[0159] 在本申请的一些实施例中,基于前述方案,在将所述目标数据输入至多层标签分类模型中之前,第一获取单元1010还用于:获取样本数据集,所述样本数据集中的样本数据包括样本和与所述样本对应的多层标签;基于所述样本数据集对所述多层标签分类模型进行训练。

[0160] 在本申请的一些实施例中,基于前述方案,第一获取单元1010配置为:获取样本和

与所述样本对应的指定层级标签;基于所述指定层级标签查询标签层次表,以获取到与所述指定层级标签相关联的其它层级标签;根据所述指定层级标签、所述其它层级标签和所述样本生成样本数据;根据所述样本数据建立样本数据集。

[0161] 在本申请的一些实施例中,基于前述方案,所述多层编码模块中除首层编码模块之外的其它层编码模块包括第一编码单元和第二编码单元;其中,所述第一编码单元的输入包括所述特征提取模块的输出,所述第二编码单元的输入包括所述第一编码单元的输出、首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出。

[0162] 在本申请的一些实施例中,基于前述方案,输入单元1020还用于:对所述其它层编码模块所包含的第一编码单元的输出、所述首层编码模块的输出,以及处于所述首层编码模块与所述其它层编码模块之间层级的编码模块所包含的第一编码单元的输出进行融合处理,得到所述其它层编码模块的输出结果。

[0163] 在本申请的一些实施例中,基于前述方案,所述多层编码模块包括第一层编码模块、第二层编码模块和第三层编码模块;所述第一层编码模块、所述第二层编码模块和所述第三层编码模块根据如下公式获得输出结果:

[0164]
$$\begin{cases} pre1 = S_1\left(f_1\right) \\ pre2 = S_2\left(FC_2\,\hat{}\left(A\times f_1 + B\times f_2\right)\right) \\ pre3 = S_3\left(FC_3\,\hat{}\left(A\times f_1 + B\times f_2 + C\times f_3\right)\right) \end{cases}$$

[0165] 其中,pre1、pre2和pre3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块的输出结果;S_1、S_2和S_3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块中的激活函数;f_1表示经过所述第一层编码模块处理后得到的特征;f_2表示经过所述第二层编码模块所包含的第一编码单元处理后得到的特征;f_3表示经过所述第三层编码模块所包含的第一编码单元处理后得到的特征;$FC_2\hat{}$表示将特征f_1和特征f_2进行融合;$FC_3\hat{}$表示将特征f_1、特征f_2和特征f_3进行融合;A、B和C为参数。

[0166] 在本申请的一些实施例中,基于前述方案,输入单元1020还用于:根据所述首层编码模块的输出与样本数据的首层标签之间的差值,生成所述首层编码模块对应的损失函数;根据所述多层编码模块中除首层编码模块之外的其它层编码模块的输出与样本数据对应层级的标签之间的差值,以及所述其它层编码模块的输出与前一层编码模块的输出之间的归属关系,生成所述其它层编码模块对应的损失函数;根据所述首层编码模块对应的损失函数和所述其它层编码模块对应的损失函数,生成所述多层标签分类模型的损失函数。

[0167] 在本申请的一些实施例中,基于前述方案,所述多层编码模块包括第一层编码模块、第二层编码模块和第三层编码模块;所述多层标签分类模型的损失函数如下:

[0168]
$$\begin{cases} L1 = pre1 - gt1 \\ L2 = \left(pre2 - gt2\right)\times F21 \\ L3 = \left(pre3 - gt3\right)\times F32 \end{cases}$$

[0169] 其中,pre1、pre2和pre3分别表示所述第一层编码模块、所述第二层编码模块和所

述第三层编码模块的输出结果;gt1、gt2和gt3分别表示样本数据的第一层标签、第二层标签和第三层标签;L1、L2和L3分别表示所述第一层编码模块、所述第二层编码模块和所述第三层编码模块对应的损失函数;F21和F32表示权重,若pre2归属于pre1,则F21＝1,若pre2不归属于pre1,则F21>1;若pre3归属于pre2,则F32＝1,若pre3不归属于pre2,则F32>1。

[0170] 在本申请的一些实施例中,基于前述方案,所述多层编码模块中的各层编码模块均包括全连接层,所述全连接层的输入包括所述特征提取模块的输出。

[0171] 在本申请的一些实施例中,基于前述方案,所述全连接层中的卷积核基于共享的权重进行特征处理。

[0172] 在本申请的一些实施例中,基于前述方案,输入单元1020还用于:在获取所述多层标签分类模型输出的与所述目标数据对应的多个分类标签之后,基于标签层次表,对所述多个分类标签的层级关系进行校验;若对所述多个分类标签的层级关系校验通过,则输出所述多个分类标签。

[0173] 图11示出了根据本申请的一个实施例的多层标签分类模型的训练装置的框图。

[0174] 参照图11所示,根据本申请的一个实施例的多层标签分类模型的训练装置1100,包括:样本数据集获取单元1110、样本数据输入单元1120和训练单元1130。

[0175] 其中,样本数据集获取单元1110用于获取样本数据集,所述样本数据集中的样本数据包括样本和与所述样本对应的多层标签;样本数据输入单元1120用于将所述样本数据集中的样本数据输入至多层标签分类模型中,所述多层标签分类模型包括特征提取模块,以及以所述特征提取模块的输出作为输入的多层编码模块,每层编码模块与一层分类标签相对应,所述多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出;训练单元1130用于根据所述多层标签分类模型的输出结果与所述样本对应的多层标签之间的损失值,调整所述多层标签分类模型的参数,以对所述多层标签分类模型进行训练。

[0176] 图12示出了适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

[0177] 需要说明的是,图12示出的电子设备的计算机系统1200仅是一个示例,不应对本申请实施例的功能和使用范围带来任何限制。

[0178] 如图12所示,计算机系统1200包括中央处理单元(Central Processing Unit,CPU)1201,其可以根据存储在只读存储器(Read-Only Memory,ROM)1202中的程序或者从存储部分1208加载到随机访问存储器(Random Access Memory,RAM)1203中的程序而执行各种适当的动作和处理,例如执行上述实施例中所述的方法。在RAM 1203中,还存储有系统操作所需的各种程序和数据。CPU 1201、ROM 1202以及RAM 1203通过总线1204彼此相连。输入/输出(Input/Output,I/O)接口1205也连接至总线1204。

[0179] 以下部件连接至I/O接口1205:包括键盘、鼠标等的输入部分1206;包括诸如阴极射线管(Cathode Ray Tube,CRT)、液晶显示器(Liquid Crystal Display,LCD)等以及扬声器等的输出部分1207;包括硬盘等的存储部分1208;以及包括诸如LAN(Local Area Network,局域网)卡、调制解调器等的网络接口卡的通信部分1209。通信部分1209经由诸如因特网的网络执行通信处理。驱动器1210也根据需要连接至I/O接口1205。可拆卸介质1211,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器1210上,以便于从其上读出的计算机程序根据需要被安装入存储部分1208。

[0180] 特别地,根据本申请的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本申请的实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分1209从网络上被下载和安装,和/或从可拆卸介质1211被安装。在该计算机程序被中央处理单元(CPU)1201执行时,执行本申请的系统中限定的各种功能。

[0181] 需要说明的是,本申请实施例所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——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。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(Erasable Programmable Read Only Memory,EPROM)、闪存、光纤、便携式紧凑磁盘只读存储器(Compact Disc Read-Only Memory,CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、有线等等,或者上述的任意合适的组合。

[0182] 附图中的流程图和框图,图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。其中,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0183] 描述于本申请实施例中所涉及到的单元可以通过软件的方式实现,也可以通过硬件的方式来实现,所描述的单元也可以设置在处理器中。其中,这些单元的名称在某种情况下并不构成对该单元本身的限定。

[0184] 作为一方面,本申请还提供了一种计算机可读介质,该计算机可读介质可以是上述实施例中描述的电子设备中所包含的;也可以是单独存在,而未装配入该电子设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被一个该电子设备执行时,使得该电子设备实现上述实施例中所述的方法。

[0185] 应当注意,尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单

元,但是这种划分并非强制性的。实际上,根据本申请的实施方式,上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之,上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0186] 通过以上的实施方式的描述,本领域的技术人员易于理解,这里描述的示例实施方式可以通过软件实现,也可以通过软件结合必要的硬件的方式来实现。因此,根据本申请实施方式的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中或网络上,包括若干指令以使得一台计算设备(可以是个人计算机、服务器、触控终端、或者网络设备等)执行根据本申请实施方式的方法。

[0187] 本领域技术人员在考虑说明书及实践这里公开的实施方式后,将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。

[0188] 应当理解的是,本申请并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

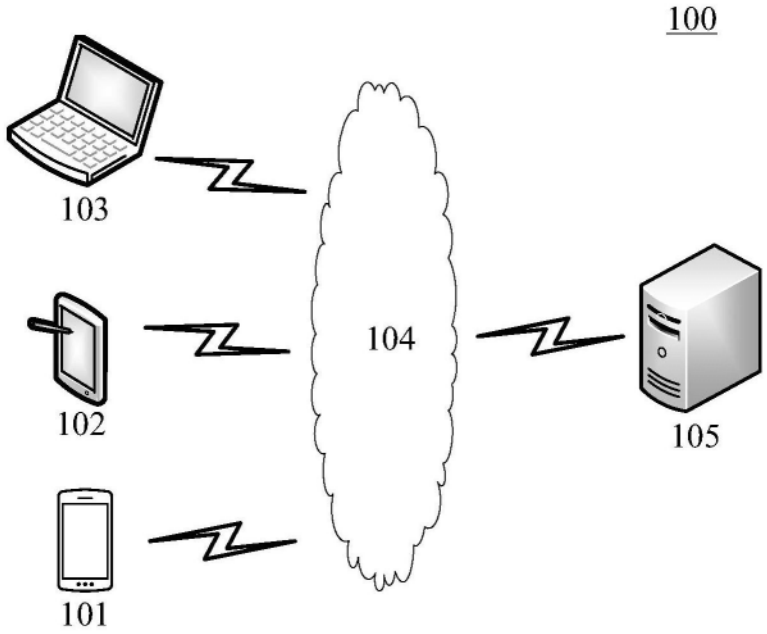

100
103
102
101
104
105

图1

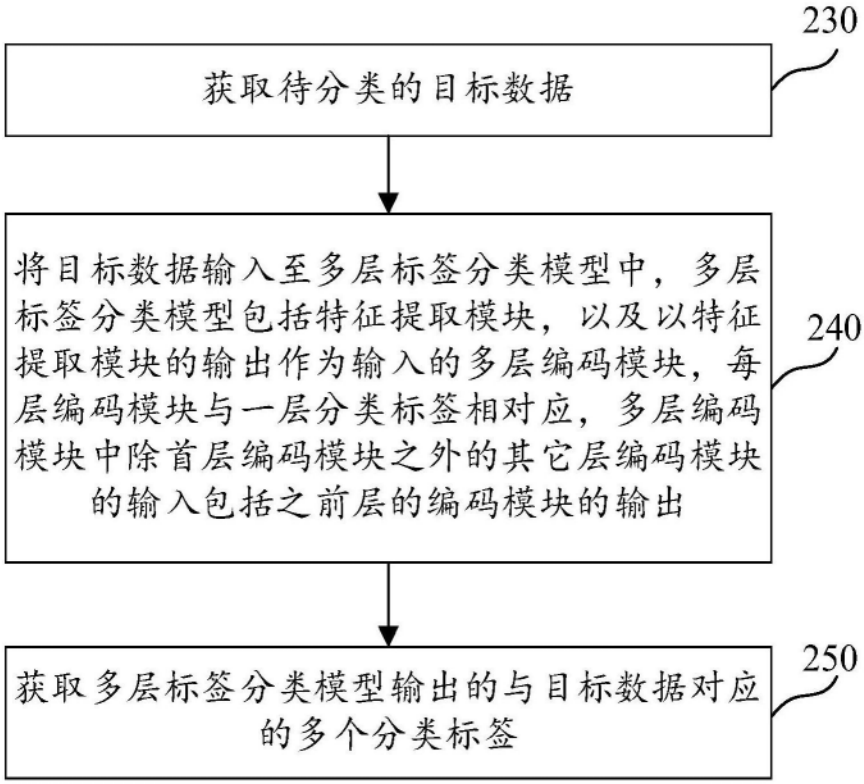

230
获取待分类的目标数据
将目标数据输入至多层标签分类模型中，多层标签分类模型包括特征提取模块，以及以特征提取模块的输出作为输入的多层编码模块，每层编码模块与一层分类标签相对应，多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出
240
获取多层标签分类模型输出的与目标数据对应的多个分类标签
250

图2

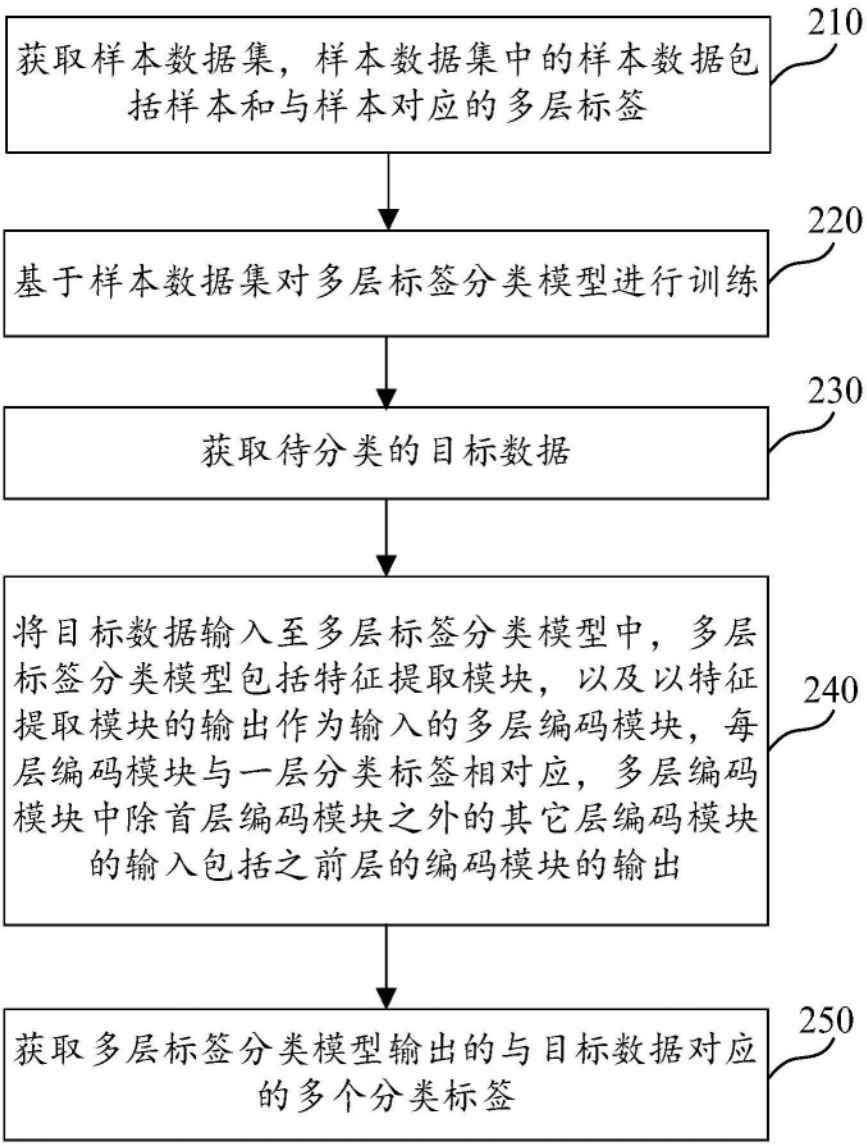

获取样本数据集，样本数据集中的样本数据包括样本和与样本对应的多层标签
210
基于样本数据集对多层标签分类模型进行训练
220
获取待分类的目标数据
230
将目标数据输入至多层标签分类模型中，多层标签分类模型包括特征提取模块，以及以特征提取模块的输出作为输入的多层编码模块，每层编码模块与一层分类标签相对应，多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出
240
获取多层标签分类模型输出的与目标数据对应的多个分类标签
250

图3

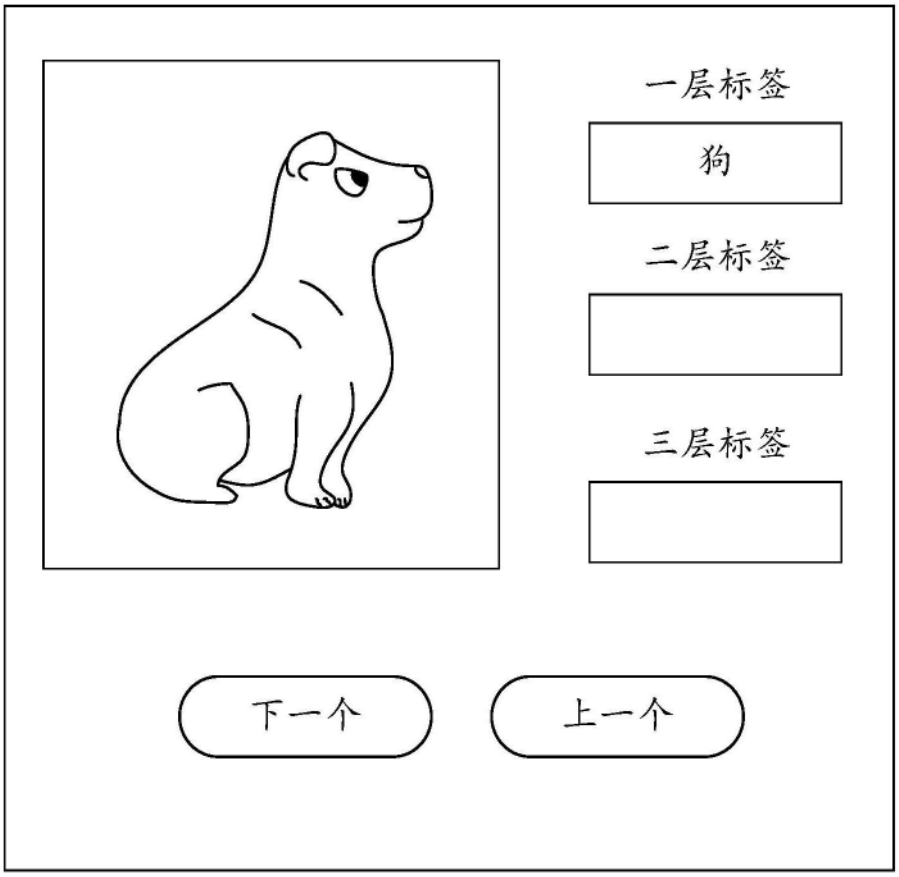


图4

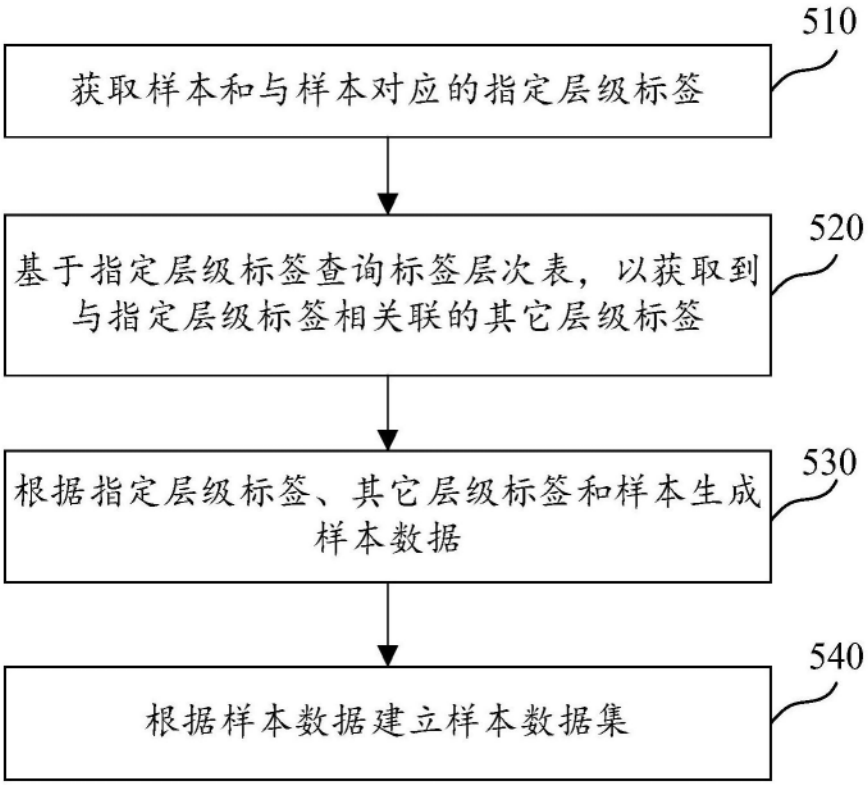

510
获取样本和与样本对应的指定层级标签
520
基于指定层级标签查询标签层次表，以获取到与指定层级标签相关联的其它层级标签
530
根据指定层级标签、其它层级标签和样本生成样本数据
540
根据样本数据建立样本数据集

图5

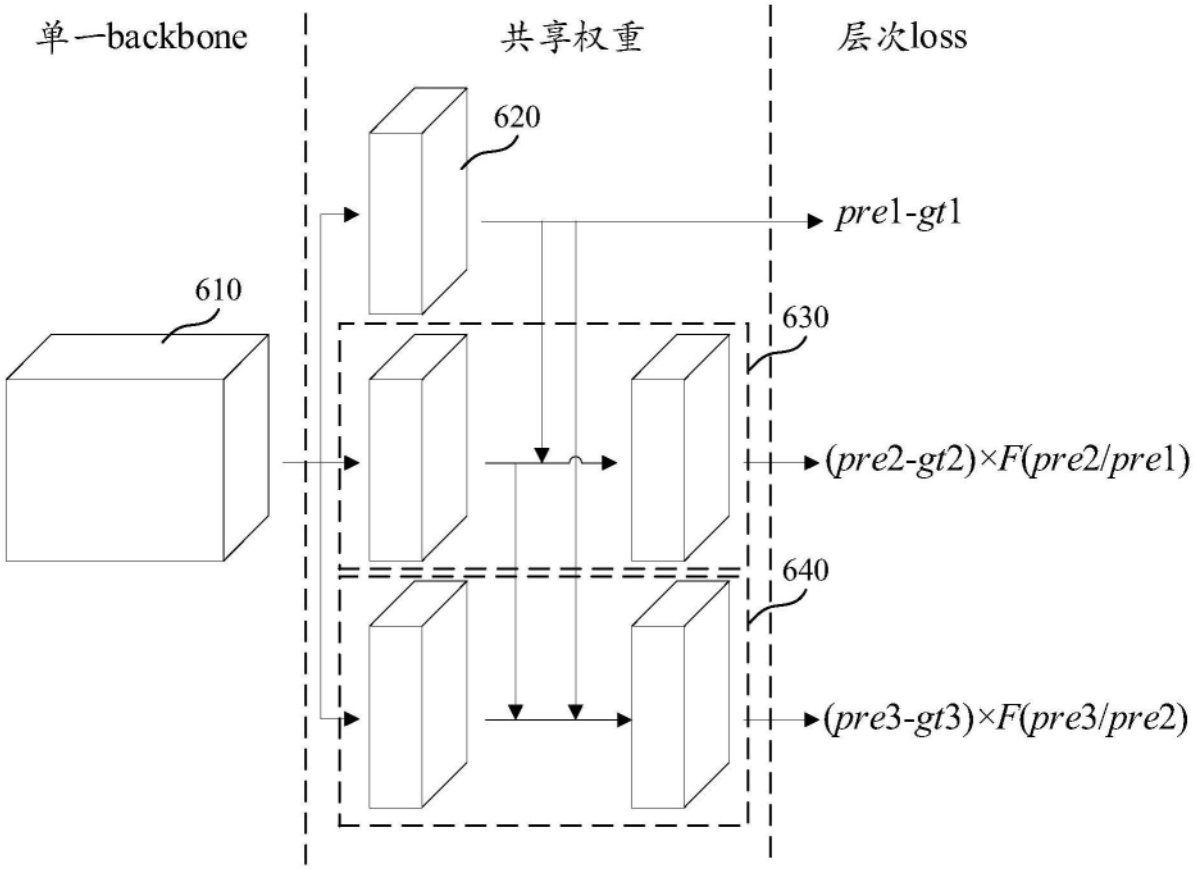

单一backbone
共享权重
层次loss
620
610
630
640
pre1-gt1
(pre2-gt2)×F(pre2/pre1)
(pre3-gt3)×F(pre3/pre2)

图6

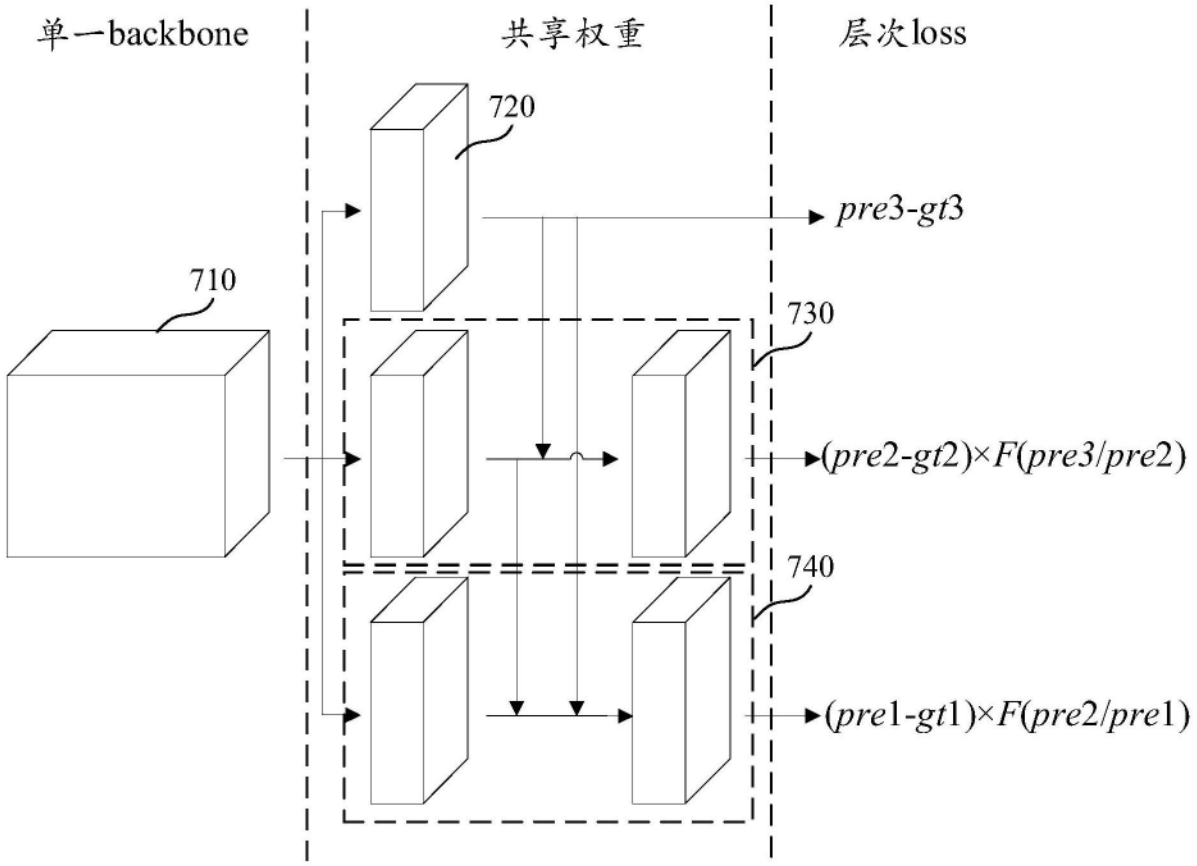

单一backbone
共享权重
层次loss
720
pre3-gt3
710
730
(pre2-gt2)×F(pre3/pre2)
740
(pre1-gt1)×F(pre2/pre1)

图7

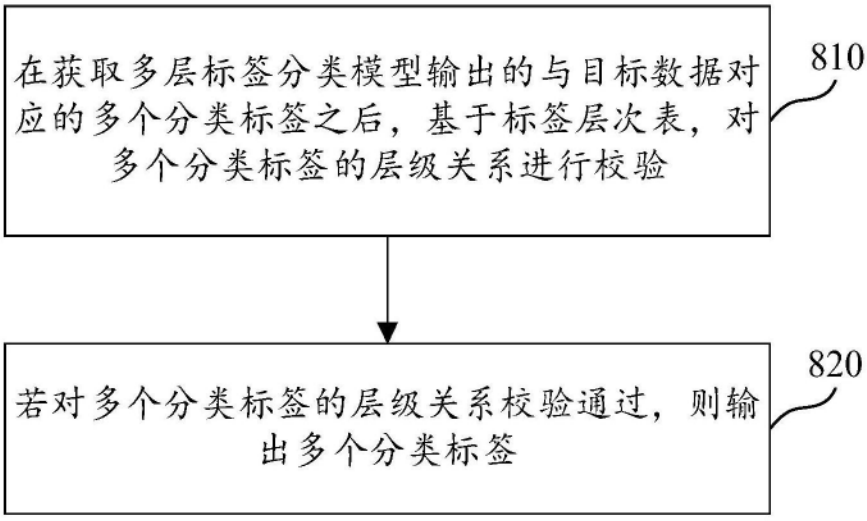

在获取多层标签分类模型输出的与目标数据对应的多个分类标签之后，基于标签层次表，对多个分类标签的层级关系进行校验
810
若对多个分类标签的层级关系校验通过，则输出多个分类标签
820

图8

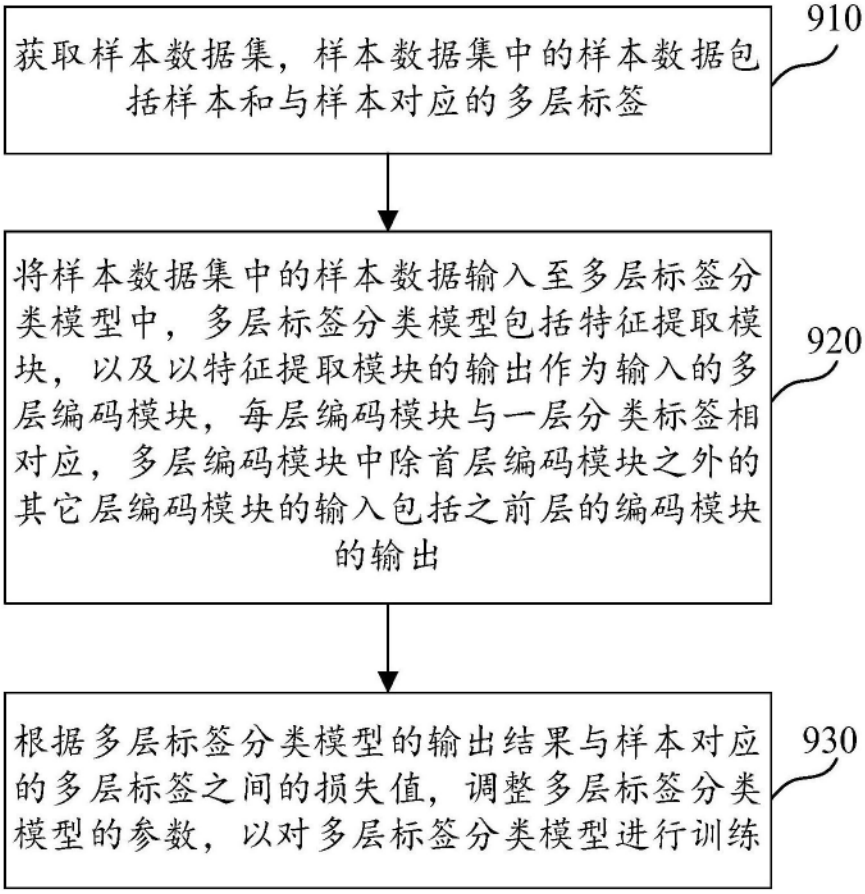

910
获取样本数据集，样本数据集中的样本数据包括样本和与样本对应的多层标签
920
将样本数据集中的样本数据输入至多层标签分类模型中，多层标签分类模型包括特征提取模块，以及以特征提取模块的输出作为输入的多层编码模块，每层编码模块与一层分类标签相对应，多层编码模块中除首层编码模块之外的其它层编码模块的输入包括之前层的编码模块的输出
930
根据多层标签分类模型的输出结果与样本对应的多层标签之间的损失值，调整多层标签分类模型的参数，以对多层标签分类模型进行训练

图9

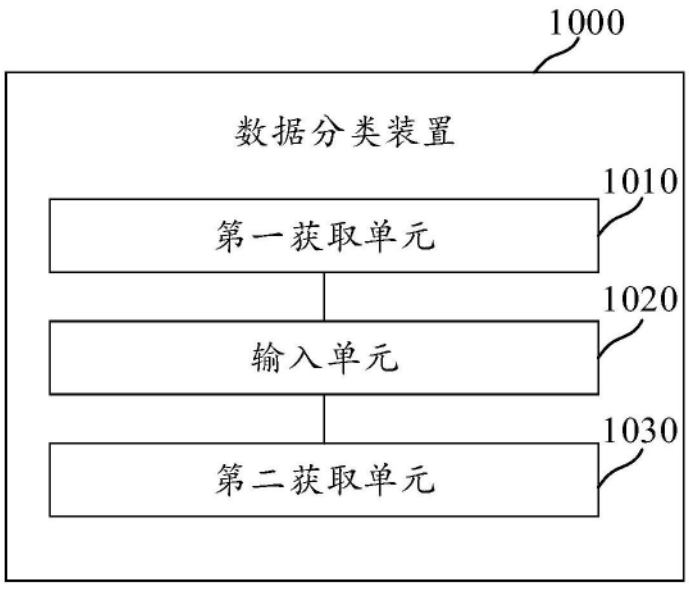

1000
数据分类装置
1010
第一获取单元
1020
输入单元
1030
第二获取单元

图10

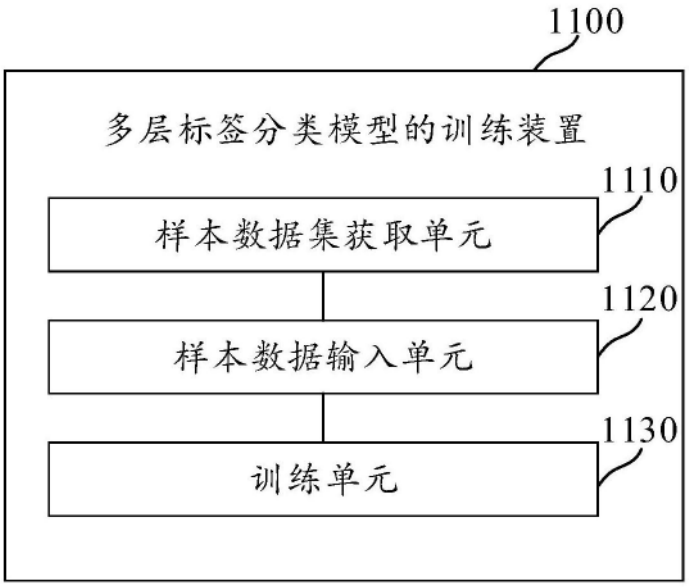

1100
多层标签分类模型的训练装置
1110
样本数据集获取单元
1120
样本数据输入单元
1130
训练单元

图11

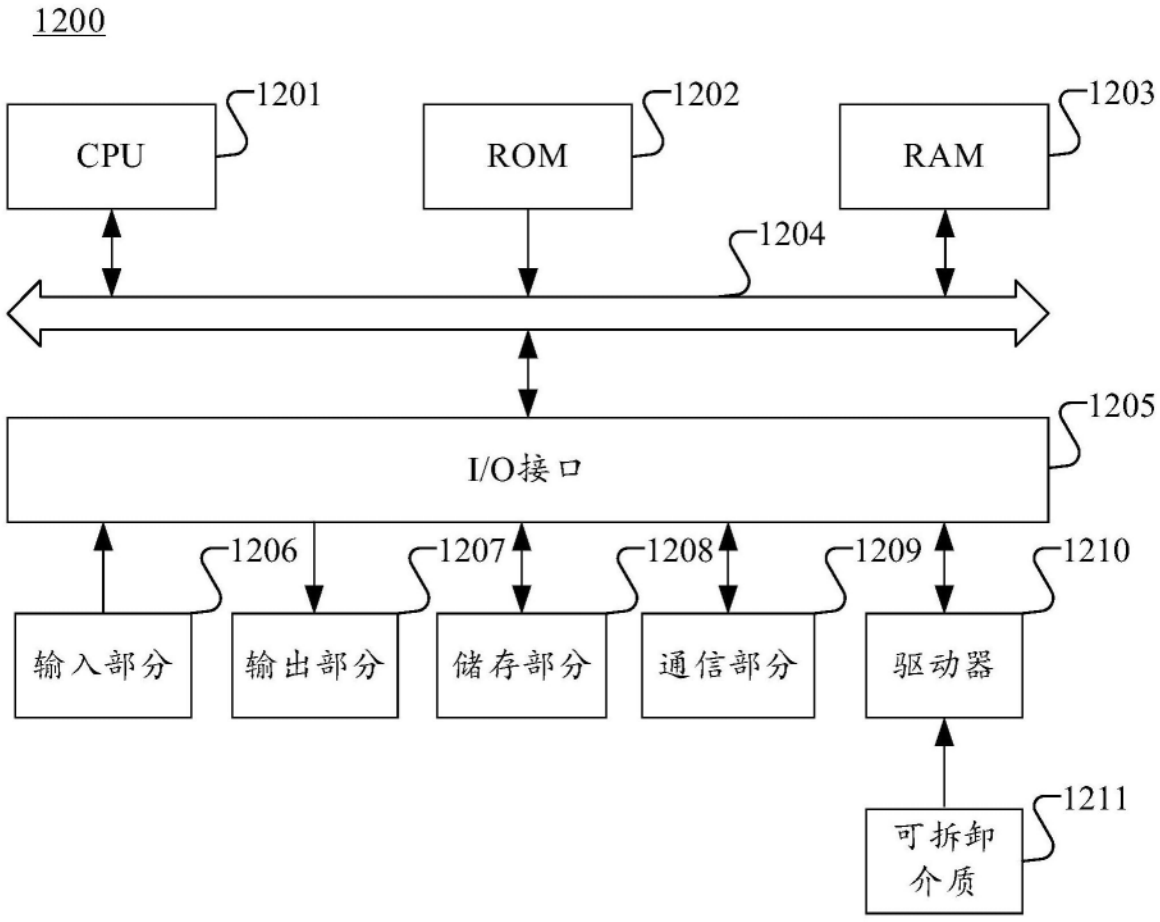

1200
CPU
1201
ROM
1202
RAM
1203
1204
I/O接口
1205
输入部分
1206
输出部分
1207
储存部分
1208
通信部分
1209
驱动器
1210
可拆卸介质
1211

图12