(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117034133 A
(43)申请公布日 2023.11.10

(21)申请号 202211263524.0

(22)申请日 2022.10.10

(71)申请人 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 王安然

(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 陈梅君

(51)Int.Cl.
G06F 18/2415(2023.01)
G06V 10/764(2022.01)
G06V 10/774(2022.01)
G06F 16/35(2019.01)
G06F 18/214(2023.01)
G06N 3/047(2023.01)
G06N 3/088(2023.01)
G06N 3/09(2023.01)
G06N 3/0499(2023.01)

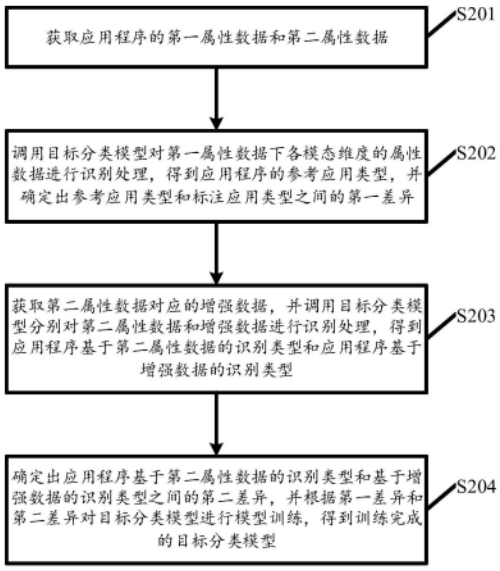
权利要求书4页 说明书24页 附图6页

(54)发明名称

一种数据处理方法、装置、设备和介质

(57)摘要

本申请实施例公开了一种数据处理方法、装置、设备和介质，该方法包括：获取应用程序的第一属性数据和第二属性数据；第一属性数据和第二属性数据分别是存在标注标签和不存在标注标签的属性数据；调用目标分类模型对第一属性数据下各模态维度的属性数据进行识别处理，得到应用程序的参考应用类型，并确定出参考应用类型和标注应用类型之间的第一差异；获取第二属性数据对应的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，可得到两个识别类型；确定出两个识别类型之间的第二差异，并根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型。这样可增强模型的鲁棒性，提升模型的类型识别能力。



CN 117034133 A

1.一种数据处理方法,其特征在于,所述方法包括:

获取应用程序的第一属性数据和第二属性数据;所述第一属性数据是存在标注标签及一个或多个模态维度的属性数据,所述标注标签用于指示所述应用程序的标注应用类型;所述第二属性数据是不存在标注标签的属性数据;

调用目标分类模型对所述第一属性数据下各模态维度的属性数据进行识别处理,得到所述应用程序的参考应用类型,并确定出所述参考应用类型和所述标注应用类型之间的第一差异;

获取所述第二属性数据对应的增强数据,并调用所述目标分类模型分别对所述第二属性数据和所述增强数据进行识别处理,得到所述应用程序基于所述第二属性数据的识别类型和所述应用程序基于所述增强数据的识别类型;

确定出所述应用程序基于所述第二属性数据的识别类型和基于所述增强数据的识别类型之间的第二差异,并根据所述第一差异和所述第二差异对所述目标分类模型进行模型训练,得到训练完成的目标分类模型;所述训练完成的目标分类模型用于进行应用类型的识别。

2.如权利要求1所述的方法,其特征在于,所述获取应用程序的第一属性数据,包括:

获取对应用程序进行应用属性描述的模态维度,所述模态维度包含以下一个或多个:文本维度、图像维度和语音维度;

基于模态维度获取所述应用程序在相应模态维度下的属性描述信息,并获取所述应用程序的标注标签;

将所述标注标签和相应的属性描述信息进行关联,并将关联所述标注标签的属性描述信息作为第一属性数据。

3.如权利要求1所述的方法,其特征在于,所述目标分类模型是已完成预训练的目标分类模型,其中,所述已完成预训练的目标分类模型包括已完成预训练的编码模块;所述调用目标分类模型对所述第一属性数据下各模态维度的属性数据进行识别处理,得到所述应用程序的参考应用类型,包括:

调用所述目标分类模型中已完成预训练的编码模块对所述第一属性数据进行特征编码处理,得到所述第一属性数据的编码特征;所述目标分类模型还包含未完成训练的识别模块;

采用所述未完成训练的识别模块,根据所述第一属性数据的编码特征对所述应用程序的应用类型进行识别处理,得到所述应用程序的参考应用类型。

4.如权利要求3所述的方法,其特征在于,若所述第一属性数据包含多个模态维度的属性描述信息;则所述调用所述目标分类模型中已完成预训练的编码模块对所述第一属性数据进行特征编码处理,得到所述第一属性数据的编码特征,包括:

调用所述目标分类模型中已完成预训练的编码模块,分别对所述第一属性数据中包含的不同模态维度的属性描述信息进行特征编码处理,得到所述第一属性数据中相应模态维度下的属性描述信息对应的描述特征;

将不同模态维度下的属性描述信息对应的描述特征进行拼接处理,并将拼接后的描述特征作为所述第一属性数据的编码特征。

5.如权利要求4所述的方法,其特征在于,当任一模态维度下的属性描述信息为多个,

且所述任一模态维度下的多个属性描述信息的信息类型不同时;所述调用所述目标分类模型中已完成预训练的编码模块,对所述第一属性数据中任一模态维度的属性描述信息进行特征编码处理,得到相应模态维度下的属性描述信息对应的描述特征的方式,包括:

调用所述目标分类模型中已完成预训练的编码模块,对所述第一属性数据中任一模态维度下对应不同信息类型的属性描述信息分别进行编码处理,得到所述任一模态维度下相应信息类型的属性描述信息对应的描述特征;

将得到的各信息类型的属性描述信息对应的描述特征均作为所述任一模态维度下的属性描述信息对应的描述特征;或者,将基于各信息类型的属性描述信息对应的描述特征得到的拼接描述特征,作为所述任一模态维度下的属性描述信息对应的描述特征。

6.如权利要求1-5中任一项所述的方法,其特征在于,所述第二属性数据为包含多个模态维度的属性描述信息的数据;所述获取所述第二属性数据对应的增强数据,包括:

确定出所述第二属性数据中的各属性描述信息分别对应的模态维度,并获取与相应模态维度匹配的数据增强算法;

采用匹配的数据增强算法,对相应模态维度下的属性描述信息进行数据增强处理,并将增强处理后的属性描述信息作为所述第二属性数据的增强数据。

7.如权利要求6所述的方法,其特征在于,在所述第二属性数据包括对应模态维度为文本维度的属性描述信息时,确定出的匹配的数据增强算法是与所述文本维度匹配的数据增强算法;所述采用匹配的数据增强算法,对相应模态维度下的属性描述信息进行数据增强处理,包括:

从所述第二属性数据中获取文本维度的属性描述信息,并对获取出的文本维度的属性描述信息进行翻译处理,得到文本维度的属性描述信息对应的翻译文本;

对所述翻译文本进行回译处理,得到所述翻译文本的回译文本;所述回译文本被作为对获取出的文本维度的属性描述信息进行数据增强处理后的属性描述信息。

8.如权利要求6所述的方法,其特征在于,在所述第二属性数据包括对应模态维度为文本维度的属性描述信息时,确定出的匹配的数据增强算法是与所述文本维度匹配的数据增强算法;所述采用匹配的数据增强算法,对相应模态维度下的属性描述信息进行数据增强处理,包括:

从所述第二属性数据中获取文本维度的属性描述信息,并对获取出的文本维度的属性描述信息进行随机的信息删除处理,得到文本维度的属性描述信息对应的删除文本;

其中,所述删除文本被作为对获取出的文本维度的属性描述信息进行数据增强处理后的属性描述信息。

9.如权利要求6所述的方法,其特征在于,在所述第二属性数据包括对应模态维度为图像维度的属性描述信息时,确定出的匹配的数据增强算法是与所述图像维度匹配的数据增强算法;所述采用匹配的数据增强算法,对相应模态维度下的属性描述信息进行数据增强处理,包括:

从所述第二属性数据中获取图像维度的属性描述信息,并获取目标噪声;

将所述目标噪声叠加到所述图像维度的属性描述信息中,得到所述图像维度的属性描述信息对应的噪声图像;所述噪声图像被作为对获取出的图像维度的属性描述信息进行数据增强处理后的属性描述信息。

10.如权利要求6所述的方法，其特征在于，在所述第二属性数据包括对应模态维度为图像维度的属性描述信息时，确定出的匹配的数据增强算法是与所述图像维度匹配的数据增强算法；所述采用匹配的数据增强算法，对相应模态维度下的属性描述信息进行数据增强处理，包括：

从所述第二属性数据中获取图像维度的属性描述信息，并从所述图像维度的属性描述信息中确定出目标图像区域；

对处于所述目标图像区域的图像像素的像素值进行随机替换处理，得到所述目标图像区域对应的遮挡图像区域；

其中，包含所述遮挡图像区域的图像维度的属性描述信息被作为对获取出的图像维度的属性描述信息进行数据增强处理后的属性描述信息。

11.如权利要求1所述的方法，其特征在于，所述根据所述第一差异和所述第二差异对所述目标分类模型进行模型训练，得到训练完成的目标分类模型，包括：

获取为所述第一差异设置的第一训练权重以及为所述第二差异设置的第二训练权重；

采用所述第一训练权重和所述第二训练权重分别对所述第一差异和所述第二差异进行加权求和处理，得到用于训练所述目标分类模型的目标差异，并采用所述目标差异训练所述目标分类模型，得到训练完成的目标分类模型。

12.如权利要求11所述的方法，其特征在于，在对所述目标分类模型进行模型训练时，包含至少两个训练阶段；还包括：

获取当前训练阶段及为所述目标分类模型设置的目标训练阶段；

在所述当前训练阶段为所述目标训练阶段之前的训练阶段时，调整所述第二训练权重的取值，以使调整后的第二训练权重小于所述第一训练权重；

在所述当前训练阶段为所述目标训练阶段时，调整所述第二训练权重的取值，以使调整后的第二训练权重大于所述第一训练权重。

13.如权利要求1所述的方法，其特征在于，若所述应用程序存在的标注标签的数量为多个，则调用所述目标分类模型识别所述第一属性数据得到的参考应用类型的数量为多个；所述确定出所述参考应用类型和所述标注应用类型之间的第一差异，包括：

从所述目标分类模型输出的多个参考应用类型中，确定出与所述应用程序的一个标注标签关联的一个参考应用类型；

基于关联的一个标注标签和一个参考应用类型构建一个子差异，并将得到的全部子差异作为所述参考应用类型和所述标注应用类型之间的第一差异。

14.一种数据处理装置，其特征在于，包括：

获取模块，用于获取应用程序的第一属性数据和第二属性数据；所述第一属性数据是存在标注标签及一个或多个模态维度的属性数据，所述标注标签用于指示所述应用程序的标注应用类型；所述第二属性数据是不存在标注标签的属性数据；

识别模块，用于调用目标分类模型对所述第一属性数据下各模态维度的属性数据进行识别处理，得到所述应用程序的参考应用类型；

确定模块，用于确定出所述参考应用类型和所述标注应用类型之间的第一差异；

所述获取模块，还用于获取所述第二属性数据对应的增强数据；

所述识别模块，还用于调用所述目标分类模型分别对所述第二属性数据和所述增强数

据进行识别处理，得到所述应用程序基于所述第二属性数据的识别类型和所述应用程序基于所述增强数据的识别类型；

所述确定模块，还用于确定出所述应用程序基于所述第二属性数据的识别类型和基于所述增强数据的识别类型之间的第二差异；

训练模块，用于根据所述第一差异和所述第二差异对所述目标分类模型进行模型训练，得到训练完成的目标分类模型；所述训练完成的目标分类模型用于进行对象类型应用类型的识别。

15.一种数据处理设备，其特征在于，包括：处理器、存储器以及网络接口；所述处理器与所述存储器、所述网络接口相连，其中，所述网络接口用于提供网络通信功能，所述存储器用于存储程序代码，所述处理器用于调用所述程序代码，以执行权利要求1-13中任一项所述的数据处理方法。

16.一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存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时，执行权利要求1-13中任一项所述的数据处理方法。

一种数据处理方法、装置、设备和介质

技术领域

[0001] 本申请涉及计算机技术领域，尤其涉及一种数据处理方法、装置、设备和介质。

背景技术

[0002] 随着人工智能技术的高速发展，各种各样基于神经网络构建的模型可用于实现不同功能的自动化应用。在分类领域中，使用分类模型对各种对象(例如植物、动物、应用程序等等)进行的自动化分类处理也日趋成熟。对分类模型的训练是模型应用于实际分类前十分重要的一个环节，模型训练效果能够直接影响分类准确度。通常为了使得分类模型处理更加准确，可以使用大量的标注数据来对分类模型进行训练，但是在一些应用场景下，存在标注的样本数据的累积花费时间较长，且标注成本往往是比较高的，虽然可以使用少量标注数据来训练出分类模型，但是目前训练得到的分类模型还存在稳定性不足的问题，模型的鲁棒性还有一定的提升空间。

发明内容

[0003] 本申请实施例提供一种数据处理方法，可以有效地增强分类模型的鲁棒性，提升分类模型的应用类型识别能力。

[0004] 一方面，本申请实施例提供了一种数据处理方法，包括：

[0005] 获取应用程序的第一属性数据和第二属性数据；第一属性数据是存在标注标签及一个或多个模态维度的属性数据，标注标签用于指示应用程序的标注应用类型；第二属性数据是不存在标注标签的属性数据；

[0006] 调用目标分类模型对第一属性数据下各模态维度的属性数据进行识别处理，得到应用程序的参考应用类型，并确定出参考应用类型和标注应用类型之间的第一差异；

[0007] 获取第二属性数据对应的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到应用程序基于第二属性数据的识别类型和应用程序基于增强数据的识别类型；

[0008] 确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异，并根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型；训练完成的目标分类模型用于进行应用类型的识别。

[0009] 一方面，本申请实施例提供了一种数据处理装置，包括：

[0010] 获取模块，用于获取应用程序的第一属性数据和第二属性数据；第一属性数据是存在标注标签及一个或多个模态维度的属性数据，标注标签用于指示应用程序的标注应用类型；第二属性数据是不存在标注标签的属性数据；

[0011] 识别模块，用于调用目标分类模型对第一属性数据下各模态维度的属性数据进行识别处理，得到目标对象的参考对象类型；

[0012] 确定模块，用于确定出参考应用类型和标注应用类型之间的第一差异；

[0013] 获取模块，还用于获取第二属性数据对应的增强数据；

[0014] 识别模块，还用于调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到应用程序基于第二属性数据的识别类型和应用程序基于增强数据的识别类型；

[0015] 确定模块，还用于确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异；

[0016] 训练模块，用于根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型；训练完成的目标分类模型用于进行应用类型的识别。

[0017] 相应地，本申请实施例提供了一种数据处理设备，包括：处理器、存储器以及网络接口；处理器与存储器、网络接口相连，其中，网络接口用于提供网络通信功能，存储器用于存储程序代码，处理器用于调用程序代码，以执行本申请实施例中数据处理方法。

[0018] 相应地，本申请实施例提供了一种计算机可读存储介质，计算机可读存储介质存储有计算机程序，计算机程序包括程序指令，程序指令当被处理器执行时，执行本申请实施例的数据处理方法。

[0019] 在本申请实施例中，可以获取应用程序的第一属性数据(存在标注标签)和第二属性数据(不存在标注标签)，标注标签用于指示应用程序的标注应用类型，该标注应用类型是应用程序在某个分类体系下的真实类别。通过调用目标分类模型对第一属性数据下各模态维度的属性数据进行识别处理，可得到应用程序的参考应用类型，进而可确定参考应用类型和标注应用类型之间的第一差异，通过第一差异可对目标分类模型进行有监督的模型训练，多模态的属性数据可增强模型的学习能力，综合可使得模型训练具备准确的类型识别能力。此外，还可引入第二属性数据进行数据增强处理后得到的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到两个识别类型并确定出两个识别类型之间的第二差异，通过第二差异可衡量应用程序的第二属性数据在数据增强前后的分类结果的差异，实现对目标分类模型进行无监督的模型训练，以及目标分类模型对于第二属性数据的细微变化的抖动情况的评估，从而强化模型对应用程序的应用类型进行识别的稳定性，进而在结合第一差异和第二差异对目标分类模型进行模型训练时，便可对目标分类模型进行有监督训练以及无监督训练，通过有监督训练可以保证训练完成的目标分类模型的进行应用类型识别的能力，通过无监督训练则可以有效提升训练完成的目标分类模型的抗干扰能力，进而增强模型的鲁棒性，使得目标分类模型对应用程序进行的类型识别更稳定。由此，在模型训练结束之后，得到的训练完成的目标分类模型可用于进行稳定、准确地应用类型的识别。

附图说明

[0020] 图1是本申请实施例提供的一种数据处理系统的架构图；

[0021] 图2是本申请实施例提供的一种数据处理方法的流程示意图；

[0022] 图3是本申请实施例提供的另一种数据处理方法的流程示意图；

[0023] 图4是本申请实施例提供的一种包含基于预训练模型进行特征表示构造的原理示意图；

[0024] 图5是本申请实施例提供的一种应用程序的分类方法整体流程图；

[0025] 图6a是本申请实施例提供的一种目标分类模型的训练流程示意图；

[0026] 图6b是本申请实施例提供的另一种目标分类模型的训练流程示意图；

[0027] 图7是本申请实施例提供的一种应用程序分类的处理示意图；

[0028] 图8是本申请实施例提供的一种数据处理装置的结构示意图；

[0029] 图9是本申请实施例提供的一种数据处理设备的结构示意图。

具体实施方式

[0030] 本申请实施例提出了一种数据处理方案，数据处理设备可获取应用程序的第一属性数据和第二属性数据，第一属性数据是携带标注标签及一个或多个模态维度的属性数据，第二属性数据是不携带标注标签的属性数据，其中，标注标签用于指示应用程序的标注应用类型，该标注应用类型是应用程序的真实类别；然后，数据处理设备可调用目标分类模型来识别第一属性数据下各个模态维度的属性数据，得到应用程序的参考应用类型，并可确定出参考应用类型和标注应用类型之间的第一差异。通过多模态的属性数据训练模型，可增强模型的分类能力，通过第一差异可衡量目标分类模型对应用程序的应用类型的识别准确度，该第一差异可用于对目标分类模型进行有监督的模型训练，从而能够训练模型学习能力和识别准确性。此外，数据处理设备可以对获取到的第二属性数据进行数据增强处理，得到第二属性数据对应的增强数据，之后，数据处理设备可调用目标分类模型对增强数据进行识别处理，得到应用程序基于该增强数据的识别类型，并可调用目标分类模型对第二属性数据进行识别处理，得到应用程序基于该第二属性数据的识别类型，接着，数据处理设备可确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异，通过第二差异可衡量应用程序的第二属性数据在数据增强前后的分类结果的差异，基于该第二差异可对目标分类模型进行无监督的模型训练，可获知目标分类模型对于输入的第二属性数据的细微变化的抖动情况并对模型进行调整，从而强化模型对应用程序的应用类型进行识别的稳定性。在得到第一差异和第二差异之后，数据处理设备可基于第一差异和第二差异对目标分类模型进行模型训练，通过两种差异可以对目标分类模型进行有监督训练以及无监督训练，基于存在标注标签的属性数据进行有监督训练可以保证训练完成的目标分类模型的识别准确率，基于第二属性数据在数据增强前后的类型识别结果的差异进行无监督训练，可以使得目标分类模型在无标注标签的第二属性数据下的类型识别结果更稳定，有效提升训练完成的目标分类模型的抗干扰能力，进而增强模型的鲁棒性。在模型训练结束之后，可以得到训练完成的目标分类模型，并可将其用于进行应用类型的识别，提高类型识别的效果。

[0031] 可以理解的是，基于不同应用程序(Application，APP)的属性数据训练得到的目标分类模型所能准确识别出的应用类型，和训练时所使用的标注标签所指示的标注应用类型属于同一分类体系，分类体系包括从某一维度对应用程序进行描述的类别信息，例如游戏APP的游戏类型(如角色扮演、射击)属于一个分类体系，游戏APP的画风类型(如美式、卡通、梦幻)则属于另一个分类体系。在不同分类体系下的标注标签有所不同，可满足不同应用场景下不同的类别需求。

[0032] 人工智能(Artificial Intelligence，AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计

算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。本申请实施例提供的数据处理方案涉及人工智能中的计算机视觉技术(Computer Vision,CV)、语音技术(Speech Technology)、自然语言处理(Nature Language Processing,NLP)以及机器学习/深度学习技术。

[0033] 其中,计算机视觉技术(Computer Vision,CV)计算机视觉是一门研究如何使机器“看”的科学,更进一步的说,就是指用摄影机和电脑代替人眼对目标进行识别和测量等机器视觉,并进一步做图形处理,使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科,计算机视觉研究相关的理论和技术,试图建立能够从图像或者多维数据中获取信息的人工智能系统。计算机视觉技术通常包括图像处理、图像识别、图像语义理解、图像检索、OCR、视频处理、视频语义理解、视频内容/行为识别、三维物体重建、3D技术、虚拟现实、增强现实、同步定位与地图构建等技术,还包括常见的人脸识别、指纹识别等生物特征识别技术。

[0034] 语音技术(Speech Technology)的关键技术有自动语音识别技术(ASR)和语音合成技术(TTS)以及声纹识别技术。让计算机能听、能看、能说、能感觉,是未来人机交互的发展方向,其中语音成为未来最被看好的人机交互方式之一。自然语言处理(Nature Language Processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此,这一领域的研究将涉及自然语言,即人们日常使用的语言,所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0035] 机器学习(Machine Learning,ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。本申请中在应用程序的属性数据包括图像模态维度的图像数据时,具体涉及图像处理的技术,在应用程序的属性数据包括语音模态维度的语音数据时,具体涉及语音识别技术,在应用程序的属性数据包括文本模态维度的文本数据时,具体涉及文本处理技术,本申请中目标分类模型可采用各种神经网络。

[0036] 本申请实施例提供的方案应用于应用程序分类场景,可以采用应用程序(Application,以下简称APP)少量的标注数据,同时引入无标注数据以及无标注数据增强后的增强数据,来共同训练目标分类模型,从而得到用于对应用程序进行类型识别的分类模型,这样可以避免标注人员对大量APP采用下载并体验APP的方式来进行标注。应用本方案可实现自动化的标注,极大地降低标注成本,且根据获取到的应用程序的任意属性数据(例如简介、图像等),采用训练好的目标分类模型,便可为其自动标注正确的标签。对于应用程序的自动准确标注,也可以为应用程序的运营提供更多的便利,例如在应用程序推广时,通过自动标注的标签自动构造推广所需的广告特征,以便于推广给相应的对象群体。

[0037] 基于上述介绍的数据处理方案,可提供如图1所示的数据处理系统的架构图。该数

据处理系统包括数据库101和数据处理设备102,其中,数据库101可以是本地数据库或者云端数据库,也可以是私有数据库或者公有数据库,数据处理设备102可以是服务器或者终端设备,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 Deliver Network,内容分发网络)、以及大数据等基础云计算服务的云服务器,但并不局限于此。终端设备包括但不限于:智能手机、平板电脑、智能可穿戴设备、智能语音交互设备、智能家电、电脑、车载终端等等设备,本申请对此不作限制。

[0038] 数据库101可用于存储应用程序的属性数据,包括存在标注标签的属性数据以及不存在标注标签的属性数据,本申请中的标注标签用于指示应用程序的标注应用类型,其中,标注应用类型是应用程序在预定义分类体系下的真实类别,属性数据用于表示应用程序的应用属性。应用程序的属性数据可包括关于应用程序的使用简介、应用程序的界面图片以及应用商店中为应用程序添加的标签等,在此不做限制。数据处理设备102可从数据库101中获取到应用程序的第一属性数据和第二属性数据,并按照前述介绍的数据处理方案,利用第一属性数据、第二属性数据以及第二属性数据对应的增强数据,对目标分类模型进行模型训练,得到训练完成的目标分类模型。当第一属性数据(属于有标注数据)的数据量过少时,通过引入第二属性数据(属于无标注数据),具体是在常见的有监督训练的差异计算的基础上,增加第二属性数据在数据增强前后的分类结果的差异来共同训练目标分类模型,从而采用本方案可以有效地提升模型对应用程序的应用类型识别的准确率和稳定性,模型的鲁棒性较高。

[0039] 请参见图2,图2是本申请实施例提供的一种数据处理方法的流程示意图。该方法可由上述数据处理设备执行。该数据处理方法包括以下步骤S201-S204。

[0040] S201,获取应用程序的第一属性数据和第二属性数据。

[0041] 应用程序(Application,以下简称APP)是一种具备特定功能的计算机程序,可以运行在各种计算机设备(如终端设备、服务器等)中。按照应用程序的来源划分,应用程序可包括第三方应用程序和本地应用程序,按照应用程序的安装方式划分,应用程序可包括需安装的应用程序和免安装的应用程序,按照应用程序的访问方式划分,应用程序可包括web(网页)应用程序(通过浏览器便可访问)。应用程序可为使用对象提供交互界面进行交互,并在交互过程中产生相应的数据,例如文本数据、语音数据、图像数据等等,这些数据可作为应用程序的属性数据。应用程序的属性数据是对应用程序的应用属性进行刻画的数据,可用于直观地表示应用程序,其中,应用属性是指应用程序的基本特点,如应用程序的功能、界面设计等等均可视为应用属性。本申请中可获取到不同应用程序的属性数据,例如应用程序A的属性数据和应用程序B的属性数据,对于任一应用程序的属性数据均按照本申请所描述的方案进行处理。在模型训练阶段可使用应用程序的属性数据作为样本数据,采用该样本数据对模型进行训练。

[0042] 按照属性数据是否存在标注标签,可将应用程序的属性数据划分为第一属性数据和第二属性数据。其中,第一属性数据是存在标注标签及一个或多个模态维度的属性数据,第二属性数据是不存在标注标签的属性数据,标注标签用于指示应用程序的标注应用类型。上述模态维度也可称为模态,由于数据的来源或形式各种各样,为便于区分,可将数据

的一种来源或形式视为一种模态。例如文本模态维度的文本数据、图像模态维度的图像数据、语音模态维度的语音数据等等。标注标签可以是在预定义分类体系下对应用程序标注的应用类型，可由人工标注或者机器标注，标注标签所指示的应用程序的标注应用类型是应用程序在预定义分类体系下的真实类别。应用程序的第二属性数据可以是不存在标注标签及一个或多个模态维度的属性数据。从模态维度来看，第一属性数据和第二属性数据均包含多个模态维度的属性数据时，第一属性数据和第二属性数据是一种多模态数据，即不同形式或来源所组成的数据。从标签的标注情况来看，第一属性数据属于标注属性数据，第二属性数据属于无标注属性数据。数据处理设备可从用于存储应用程序的属性数据的数据库中，获取到应用程序的第一属性数据和应用程序的第二属性数据。

[0043] 举例来说，在应用商店中往往会保存关于APP的简介、内部界面设计图片以及一些应用商店内的标签，这些数据是APP开发者角度最想展示的产品关键信息，可作为APP的原始信息资源，由于这些数据可表示APP的基本属性，比如APP的功能、设计风格、使用体验等等，因此这些原始信息资源可作为应用程序的属性数据，数据处理设备可以通过检索爬虫的方式较为容易地获取到。

[0044] S202，调用目标分类模型对第一属性数据下各模态维度的属性数据进行识别处理，得到应用程序的参考应用类型，并确定出参考应用类型和标注应用类型之间的第一差异。

[0045] 在获取到应用程序的第一属性数据以及第二属性数据之后，数据处理设备可调用目标分类模型对第一属性数据包含的各个模态维度的属性数据进行识别处理，得到应用程序的参考应用类型。其中，目标分类模型是具备基本识别处理能力的模型，基于对第一属性数据的识别处理所包含的处理步骤，目标分类模型可包含相应功能的处理模块，这些处理模块可以是预训练模型和随机初始化的模型中的一者或两者。此处的预训练模型可以是通过海量数据(例如文本语料数据)进行预先训练的模型，在这个初步训练好的模型基础上，可继续训练或另作他用。一般预训练模型往往采用大量的数据来进行一些通用任务的训练，通过在多元任务上预训练获取通用知识，然后可使用少量目标任务上的标注数据进行微调，使得微调的模型能够很好地处理目标任务。预训练模型例如是在文本领域的预训练的语言表征模型，如BERT(BidirectionalEncoder Representations from Transformer，基于Transformer的双向编码器表示)模型，图像领域的BiT(BigTransfer，一种预训练的图像模型)模型等等。对第一属性数据包含的所有模态维度的属性数据进行识别处理，能够实现模型的多模态学习，模型对应用程序的识别准确度得以提升。

[0046] 参考应用类型可以是目标分类模型预测的一种类型，由于目标分类模型在初期训练时还不够稳定，对应用程序的应用类型的识别可能还不够准确，因此，此处得到的应用程序的参考应用类型与标注应用类型是存在一定差异的，而随着模型的训练更迭，两者之间的差异会越来越小，即表明模型对应用程序的应用类型的识别趋于准确。参考应用类型和标注应用类型可通过类型分布信息来表示，例如参考应用类型是应用程序属于各个应用类型的概率分布，标注应用类型则是可以是0-1分布，应用程序所属的真实标注应用类型为1，其他标注应用类型为0。数据处理设备可确定参考应用类型和标注应用类型之间的第一差异，可选地，第一差异可以通过两种分布之间的交叉熵损失来衡量。该第一差异用于反映目标分类模型的分类准确度。通过第一差异可实现对目标分类模型的有监督训练，从而使得

模型能够见到标准的样本数据，保证模型训练的准确度。

[0047] S203，获取第二属性数据对应的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到应用程序基于第二属性数据的识别类型和应用程序基于增强数据的识别类型。

[0048] 数据处理设备可以获取第二属性数据对应的增强数据，该增强数据是对获取到的应用程序的第二属性数据进行数据增强处理得到的，具体是增强处理后的第二属性数据。在第一属性数据(一种标注数据)的数据量不足时，通过引入第二属性数据(一种无标注数据)和增强处理后的第二属性数据，可以扩充模型的样本学习量，使得模型见到更多的样本。接着，数据处理设备可调用目标分类模型对第二属性数据进行识别处理，得到应用程序基于第二属性数据的识别类型，以及对增强数据进行识别处理，得到应用程序基于增强数据的识别类型。此处的识别类型也可通过类型分布来表示，该类型分布可以是应用程序属于各个应用类型的概率分布，在训练阶段通过类型分布来表示可更好地对模型的识别准确度进行评估。

[0049] 由于增强数据是对第二属性数据变换后得到的，目标分类模型对增强数据的识别处理可能不够准确，因此，应用程序基于增强数据的识别类型和应用程序基于第二属性数据的识别类型也存在一定的差异。

[0050] S204，确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异，并根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型。

[0051] 数据处理设备可确定出第二属性数据在数据增强前后的分类结果的差异，即第二差异，并根据第一差异和第二差异对目标分类模型进行模型训练，这样，在分类任务常见的有监督训练的基础上，本申请在模型训练过程中引入了无监督训练，具体是无标注属性数据在数据增强前后的分类结果的差异结合训练，这样可以保证目标分类模型在无标注属性数据下输出稳定的模型预测结果，有效提升模型的稳定性，进而增强模型的鲁棒性。

[0052] 需要说明的是，对目标分类模型的模型训练是一种迭代训练，即在目标分类模型训练完成之前，可以重复执行以上S201-S203。当模型训练达到预设迭代次数或者是差异达到收敛条件，则可以得到训练完成的目标分类模型，训练完成的目标分类模型用于进行应用类型的识别。在一种实现方式中，数据处理设备可调用训练完成的目标分类模型，对待分类对象的属性数据进行识别处理，得到待分类对象的应用类型。本申请在不做特别说明时，“类型”和“类别”表示相同的概念。

[0053] 本申请实施例提供的数据处理方法，可以获取应用程序的第一属性数据(存在标注标签)和第二属性数据(不存在标注标签)，标注标签用于指示应用程序的标注应用类型，该标注应用类型是应用程序的真实类别。通过调用目标分类模型对第一属性数据进行识别，可得到应用程序的参考应用类型，进而可确定参考应用类型和标注应用类型之间的第一差异，通过第一差异可对目标分类模型进行有监督的模型训练，使得模型训练具备准确的识别能力。此外，还可引入第二属性数据进行数据增强处理后得到的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到两个识别类型并确定出两个识别类型之间的第二差异，通过第二差异可衡量应用程序的第二属性数据在数据增强前后的分类结果的差异，对目标分类模型进行无监督的模型训练，以及获知目标分类模型对

于输入的第二属性数据的细微变化的抖动情况并对目标分类模型的模型参数进行调整，从而强化目标分类模型对应用程序的应用类型进行识别的稳定性。进一步地，可基于第一差异和第二差异对目标分类模型进行模型训练，这样便可对目标分类模型进行有监督训练以及无监督训练，通过有监督训练可以保证训练完成的目标分类模型的进行应用类型识别的能力，通过无监督训练则可以有效提升训练完成的目标分类模型的抗干扰能力，进而增强模型的鲁棒性，使得目标分类模型对应用程序进行的类型识别更稳定。这样，在模型训练结束之后，得到的训练完成的目标分类模型具备较好的应用类型识别能力，可用于进行稳定、准确地进行应用类型的识别。

[0054] 请参见图3，图3是本申请实施例提供的一种数据处理方法的流程示意图。该方法可由上述数据处理设备执行。该数据处理方法包括以下步骤S301-S305。

[0055] S301，获取应用程序的第一属性数据和第二属性数据。

[0056] 在一个实施例中，获取应用程序的第一属性数据，包括：获取对应用程序进行应用属性描述的模态维度；基于模态维度获取应用程序在相应模态维度下的属性描述信息，并获取应用程序的标注标签；将标注标签和相应的属性描述信息进行关联，并将关联标注标签的属性描述信息作为第一属性数据。

[0057] 具体地，数据处理设备可先获取对应用程序进行应用属性描述的模态维度，模态维度包含以下一个或多个：文本维度、图像维度和语音维度。不同模态维度可对应用程序的应用属性从不同角度进行描述，且不同模态维度下的数据模态不同。数据处理设备可获取应用程序在每个模态维度下的属性描述信息，包括但不限于：文本维度下的属性描述信息(即文本数据)、图像维度的属性描述信息(即图像数据)、语音维度下的属性描述信息(即语音数据)。进一步地，模态维度还可包括社交维度，社交维度的属性描述信息例如是两个社交对象之间的社交关系。举例来说，应用程序为游戏APP，那么文本数据可以是该游戏APP内产生的文本数据，也可以是应用商店内对该游戏APP的简介或者使用评价，还可以是应用商店内为该游戏APP设置的标签。图像数据可以是游戏APP的游戏界面，语音数据可以是游戏APP内置的虚拟角色的语音数据，社交数据可以是游戏APP内玩家之间的社交关系，或者玩家之间通过游戏APP进行社交活动所产生的社交信息等。

[0058] 可选地，属性描述信息包括但不限于：对象描述文本，对象图像和在相应对象场景下的场景描述文本。其中，对象描述文本是用于描述应用程序的文本数据，例如对应用程序进行简单介绍的文本简介信息；对象图像是用于描述应用程序的内容的图像数据，例如应用程序的对象交互界面图；场景描述文本是用于描述应用程序所处场景的文本数据，例如应用程序在应用商店(一种提供各个第三方应用程序下载使用的应用)中的类别标签，此处的类别标签是应用商店中特有的，可能和需要的标注标签不同。数据处理设备可获取应用程序的标注标签，并将标注标签和每个模态维度下的属性描述信息进行关联，关联了标注标签的属性描述信息即可作为第一属性数据。当模态维度包括至少两个，通过上述方式获取到的第一属性数据是存在标注标签的多模态数据，此处的多模态数据即多种模态维度的属性数据，具体由不同形式的属性描述信息所构成。

[0059] 需要说明的是，对于应用程序的第二属性数据的获取，可类比于应用程序的第一属性数据的获取，数据处理设备可获取应用程序在相应模态维度下的属性描述信息，并直接将属性描述信息作为应用程序的第二属性数据，且在模态维度包括至少两个时，获取到

的第二属性数据是无标注标签的多模态数据。

[0060] 在一个实施例中，目标分类模型是已完成预训练的目标分类模型。已完成预训练的目标分类模型包括已完成预训练的编码模块。已完成预训练的编码模块的数量可包括至少一个编码网络，一个已完成预训练的编码网络可用于处理一种模态维度的属性描述信息，例如文本维度的属性描述信息对应的已完成预训练的编码模块可以包括BERT模型，图像维度的属性描述信息对应的已完成预训练的编码模块可以包括BiT模型。相较于随机初始化模型参数，已完成预训练的编码模块采用的是预训练后的模型参数，这样，预训练后的模型经过少量目标任务的数据进行训练微调，便可得到满足条件的分类模型，模型在训练收敛速度和准确度上均占优势，还能够节省训练所需的计算资源。调用目标分类模型对第一属性数据下各个模态维度的属性数据进行识别处理的实施方式可参见下述内容。

[0061] S302，调用目标分类模型中已完成预训练的编码模块对第一属性数据进行特征编码处理，得到第一属性数据的编码特征。

[0062] 数据处理设备可先调用已完成预训练的编码模块对第一属性数据进行特征编码处理，通过特征编码处理可以构造出表示应用程序基于第一属性数据的编码特征，该编码特征可以是向量表示信息，如embedding(一种向量)。当第一属性数据包括多个模态维度的属性描述信息时，得到的编码特征是一种多模态表示特征，该多模态表示特征可以从不同角度对应用程序进行刻画，从而强化对应用程序的表示。这样，通过将应用程序的多源异构的多种模态信息(包括文本、图像、语音等)进行结合，共同构造应用程序的特征表示，能够实现多模态学习。基于多模态学习可对应用程序进行多个角度的描述，从而在下游任务(如分类任务)中有更好的表现。

[0063] 在一个实施例中，若第一属性数据包含多个(即至少两个)模态维度的属性描述信息，那么对第一属性数据的特征编码处理具体可以包含以下(1)-(2)所描述的内容。

[0064] (1)调用目标分类模型中已完成预训练的编码模块，分别对第一属性数据中包含的不同模态维度的属性描述信息进行特征编码处理，得到第一属性数据中相应模态维度下的属性描述信息对应的描述特征。

[0065] 已完成预训练的编码模块可包括与相应模态维度匹配的编码网络，该编码网络所使用的网络参数是已完成预训练采用的网络参数。数据处理设备在调用已完成预训练的编码模块，对第一属性数据下各个模态维度的属性数据进行处理时，具体可调用与相应模态维度匹配的编码网络，对相应模态维度下的属性描述信息进行特征编码处理，得到相应模态维度下的描述特征。此处相应模态维度可以是文本维度、图像维度、语音维度等多个模态维度中的任一个模态维度。属性描述信息是一种用于对应用程序的应用属性进行描述的属性数据，可以有各种存在形式，即各种模态维度，例如文本维度的属性描述信息可以包括应用程序的简介和类别标签，图像维度的属性描述信息可以包括应用程序的功能界面图等。对于各个模态维度的属性描述信息均采用类似的处理方式，得到对应的描述特征，各个描述特征均可以通过embedding向量来表示。举例来说，第一属性数据包含文本维度的属性描述信息和图像维度的属性描述信息，那么可调用预训练后的BERT模型对文本维度的属性描述信息进行特征编码处理，得到文本模态维度下的文本描述特征，并调用预训练后的BiT模型(一种在大规模图像语料上训练的图像表示模型)对图像维度的属性描述信息进行特征编码处理，得到图像模态维度下的图像描述特征。通过特征编码处理，可将相应模态维度的

属性描述信息(如文本、图像信息)编码成数据处理设备可识别并分类的描述特征(如embedding向量),从而将原始数据映射到隐层特征空间,实现对属性描述信息进行抽象化表示。

[0066] 进一步地,在一个实施例中,任一模态维度下的属性描述信息为多个(即至少两个),并且任一模态维度下的多个属性描述信息的信息类型可不同。对于信息类型的确定,在一种方式下,可基于信息来源确定同一模态维度下的属性描述信息的信息类型,举例来说,开发者为应用程序设计的开发代码和该应用程序上线到应用商店后被标注的类别标签可划分为两个不同的信息类型。在另一种方式中,也可基于信息长度确定同一模态维度下的属性描述信息的信息类型。举例来说,对于应用程序在文本模态下的简介和类别标签,由于类别标签是较短的文本数据,而简介是较长的文本数据,便可将其划分为两个信息类型。对于不同信息类型的属性描述信息,可分别处理以提高编码的准确性和处理的便捷性。

[0067] 当任一模态维度下的属性描述信息为多个,且任一模态维度下的多个属性描述信息的信息类型不同时,对于任一模态维度的属性描述信息的特征编码处理,得到该任一模态维度下的属性描述信息对应的描述特征,可包括以下步骤:首先,数据处理设备可调用目标分类模型中已完成预训练的编码模块,对第一属性数据中任一模态维度下对应不同信息类型的属性描述信息分别进行编码处理,得到任一模态维度下相应信息类型的属性描述信息对应的描述特征。然后,数据处理设备可将得到的各信息类型的属性描述信息对应的描述特征均作为任一模态维度下的属性描述信息对应的描述特征;或者,将基于各信息类型的属性描述信息对应的描述特征得到的拼接描述特征,作为任一模态维度下的属性描述信息对应的描述特征。

[0068] 对于同一模态维度下不同信息类型的属性描述信息,可采用与该模态维度匹配的编码网络(一种预训练模型)对各个信息类型的属性描述信息分别进行编码处理,得到对应信息类型的属性描述信息下的描述特征。在一种实现方式中,可直接将各个信息类型的属性描述信息对应的描述特征作为该模态维度下的属性描述信息对应的描述特征,即同一模态维度下对应的描述特征包括各个信息类型下对应的描述特征。在另一种实现方式中,可将各个信息类型的属性描述信息对应的描述特征进行拼接,得到拼接描述特征,并将该拼接描述特征作为同一模态维度下对应的描述特征。举例来说,对于APP简介和标签这种文本类信息(即文本维度下的属性描述信息),采用BERT模型(即已完成预训练的编码模块)进行编码,其中,BERT模型是在大规模文本语料上训练的基于transformer(一种基础模型)结构的语言模型,需要说明的是,这里其他预训练语言模型依然适用,本申请在此不做限制。通过将文本输入到语言模型BERT中,可分别得到简介的embedding和标签的embedding,这里用V1和V2表示。上述V1和V2可直接作为文本维度下对应的描述特征,也可以将V1和V2拼接后作为文本维度下对应的描述特征。

[0069] 可以理解的是,虽然在训练时所使用的编码模块是同一类,而由于不同信息类型的属性描述信息的输入,训练后得到的编码模块是不同的,具体体现在模型参数的不同上。举例来说,两个相同的预训练之后得到的BERT模型,一个用于处理应用程序的简介,另一个用于处理应用程序的标签,那么训练完成的BERT模型是两个模型参数不同的BERT模型。

[0070] 可见,在特征表示层面上,通过已完成预训练的编码模块进行特征编码处理,可充分利用较为成熟的预训练模型(例如文本和图片形式对应的预训练模型),使得模型学习可

以在大量语料训练后的模型的基础上进行训练，使得模型有更加精准的学习结果。

[0071] (2)将不同模态维度下的属性描述信息对应的描述特征进行拼接处理，并将拼接后的描述特征作为第一属性数据的编码特征。

[0072] 数据处理设备可对各个模态维度下的属性描述信息对应的描述特征进行拼接处理，得到拼接后的描述特征，并且该拼接后的描述特征可作为第一属性数据的编码特征，该编码特征是一个分布式特征表示(一种能够提升特征泛化能力的特征表示)。当第一属性数据的编码特征从多个模态维度进行表示时，相比于单一模态维度的描述特征，最终得到的编码特征能够强化对应用程序的表示，编码特征包括多个维度的描述特征，能够增强模型学习效果。此外，在特征表示时所使用的属性描述信息均是关键信息，可避免引入过多冗余信息，从而使得模型训练更加容易和简单。

[0073] 举例来说，应用程序的第一属性数据包括APP的截图信息、简介以及应用商店内的标签，基于预训练模型进行特征表示构造的原理示意图可参见图4。其中，APP的简介和标签均采用文本编码网络(如BERT模型)进行编码，其中，文本编码网络包括文本编码网络A和文本编码网络B，这两个文本编码网络在训练之前是模型参数相同的文本编码网络。文本编码网络A用于处理APP的简介，文本编码网络B用于处理APP的标签，文本编码网络A可输出简介的向量表示V1，文本编码网络B可输出标签的向量表示V2，APP截图信息则可时输入到图像编码网络(如BiT模型)中，获得截图的向量表示V3。然后，V1、V2、V3分别经过多层全连接神经网络(Multilayer Perceptron，MLP)后，通过全连接层则可以将学习到的分布式特征表示映射到样本标记空间，再将MLP处理后的向量表示进行向量拼接(contact)后形成最终的APP特征表示(即编码特征feature)。一般将feature通过softmax(一种归一化指数函数)后即可获得分类结果。

[0074] 需要说明的是，通过对属性数据的模态维度进行扩充，即获取到更多模态维度的属性描述信息，在特征编码处理阶段，所得到的编码特征也是增加了对应用程序的表示维度，从而进一步强化编码特征，增强模型的表示学习能力。本申请中对模态维度不做限制。相应的，第二属性数据也包括多个模态维度的属性描述信息，对于第二属性数据的特征编码处理，可参考第一属性数据的特征编码处理的过程，第二属性数据经过特征编码处理所得到的编码特征也是一种多模态表示特征。

[0075] 目标分类模型除了包含已完成预训练的编码模块，目标分类模型还包含未完成训练的识别模块，该识别模块可以视为一种分类器，例如可以是softmax回归模型。在对第一属性数据的识别处理过程中，未完成训练的识别模块的作用可如下S303所描述的内容。

[0076] S303，采用未完成训练的识别模块，根据第一属性数据的编码特征对应用程序的应用类型进行识别处理，得到应用程序的参考应用类型，并确定出参考应用类型和标注应用类型之间的第一差异。

[0077] 数据处理设备可采用未完成训练的识别模块，根据已完成预训练的编码模型输出第一属性数据的编码特征，对应用程序的应用类型进行识别处理，进而该未完成训练的识别模块可输出应用程序的参考应用类型。参考应用类型和标注应用类型均可通过类型分布信息来表示。举例来说，应用程序A属于类别a1-a4中的任一个，那么得到的参考应用类型可以通过该应用程序属于各个应用类型的概率分布来表示，如[0.1,0.5,0.9,0.2]，应用程序A的标注标签也表示为类别分布，具体可以是0-1分布，如[0,0,1,0]。因此，对于参考应用类

型和标注应用类型之间的第一差异的确定，具体是两个分布信息之间差异的确定。可以理解的是，在识别模块完成训练之后，在模型应用阶段得到的应用类型可以直接表示为相应的类别，例如选取出概率最大对应的应用类型作为最终识别得到的应用类型。

[0078] S304，获取第二属性数据对应的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到应用程序基于第二属性数据的识别类型和应用程序基于增强数据的识别类型。

[0079] 在一个实施例中，第二属性数据为包含多个模态维度的属性描述信息的数据，第二属性数据属于多模态数据。对第二属性数据的数据增强处理，具体可采用以下方式实现：首先数据处理设备可确定出第二属性数据中的各属性描述信息分别对应的模态维度，并获取与相应模态维度匹配的数据增强算法；然后，数据处理设备可采用匹配的数据增强算法，对相应模态维度下的属性描述信息进行数据增强处理，并将增强处理后的属性描述信息作为第二属性数据的增强数据。

[0080] 数据增强是一种通过利用算法来扩展训练数据的技术，通过数据增强的方法，可以在数据量不足时，利用算法自动增强和扩充训练数据。举例来说，在文本分类任务中需要获取大量的文本标注数据以提高模型的精度和泛化能力，但是人工标注的成本较高，此时便可通过数据增强的方法扩充所需的标注数据。因此，数据增强可在数据约束环境下花费较低的成本来有效提高模型学习性能和准确性，本申请中通过与相应模态维度匹配的数据增强算法可自动对无标注的属性数据进行数据增强处理，得到增强数据，相比于原有的标注数据和无标注数据，增强数据的引入使得用于对模型进行训练的样本数据得到了进一步扩充。

[0081] 为了更好地对第二属性数据中各属性描述信息进行数据增强处理，基于属性描述信息与模态维度之间的对应关系，确定出第二属性数据中各属性描述信息对应的模态维度，以获取到与模态维度匹配的数据增强算法。此处数据增强算法是一种用于进行数据增强的策略机制或者说规则，在该数据增强算法下，数据处理设备可基于相应算法指令实现数据增强处理。第二属性数据对应的模态维度可包括一个或者多个，当模态维度包括一个时，可获取到与确定出的模态维度匹配的一个数据增强算法，当模态维度包括多个时，可获取到与多个模态维度中每个模态维度匹配的数据增强算法。例如，属性描述信息对应的模态维度包括文本维度和图像维度，那么数据处理设备可获取到与文本维度匹配的数据增强算法，以及与图像维度匹配的数据增强算法。与相应模态维度匹配的数据增强算法可用于指示对相应模态维度下的属性描述信息进行增强处理的具体规则，进而可采用匹配的数据增强算法，对相应模态维度下的属性描述信息进行数据增强处理，例如，与文本维度匹配的数据增强算法，可对文本维度下的属性描述信息进行数据增强处理，与图像维度匹配的数据增强算法，可对图像维度下的属性描述信息进行数据增强处理。数据处理设备可将各个模态维度下的增强处理后的属性描述信息作为第二属性数据的增强数据，当第二属性数据涉及至少两个模态维度时，增强数据也涉及至少两个模态维度，此时增强数据也属于多模态数据。

[0082] 在一种实现方式中，在第二属性数据包括对应模态维度为文本维度的属性描述信息时，确定出的匹配的数据增强算法是与文本维度匹配的数据增强算法。文本维度的属性描述信息可以理解为文本模态的数据，在对第二属性数据进行数据增强处理时，具体是采

用与文本维度匹配的数据增强算法，对文本模态维度下的属性描述信息进行数据增强处理，详细地可包括以下两种方式。

[0083] 方式一、从第二属性数据中获取文本维度的属性描述信息，并对获取出的文本维度的属性描述信息进行翻译处理，得到文本维度的属性描述信息对应的翻译文本；对翻译文本进行回译处理，得到翻译文本的回译文本；回译文本被作为对获取出的文本维度的属性描述信息进行数据增强处理后的属性描述信息。

[0084] 数据处理设备可从第二属性数据中获取文本维度的属性描述信息，该文本维度的属性描述信息是包括一种或多种语言的文本数据，按照两种语言之间的转换规则，数据处理设备可对文本维度的属性描述信息进行翻译，得到翻译文本，该翻译文本是另一种语言的文本数据。举例来说，获取到的文本维度的属性描述信息为中文的文本简介，基于中英文之间的翻译规则，可以将其翻译为英文的文本简介，得到翻译文本，接着，可对翻译文本进行回译处理，将翻译文本的语言还原为原本语言的文本数据。简单来说，例如将中文的APP简介翻译成英文，再将英文翻译回中文，即完成了对文本模态的数据增强处理。

[0085] 由于在不同语言的翻译方向不同时，翻译所遵循的规则也存在一定的差别，因此，最终得到的回译文本与翻译前的文本是有一定区别的。通过翻译再回译得到回译文本，这样既保留了部分原本的文本模态的属性描述信息，又存在与原本的文本模态的属性描述信息不同的部分，此外，通过对文本模态的数据进行回译，可以丰富同一语义下的不同表述方式，实现文本模态的属性描述信息的数据增强处理。数据处理设备可以将回译文本作为获取出的文本维度的属性描述按照所匹配的数据增强算法进行数据增强后的属性描述信息。

[0086] 方式二、从第二属性数据中获取文本维度的属性描述信息，并对获取出的文本维度的属性描述信息进行随机的信息删除处理，得到文本维度的属性描述信息对应的删除文本；其中，删除文本被作为对获取出的文本维度的属性描述信息进行数据增强处理后的属性描述信息。

[0087] 在此方式下，数据处理设备可从获取到的文本维度的属性描述信息中随机选取出需要删除的信息，并对这些信息进行删除处理，该信息可以是文本片段、句子以及词语等中的一种或多种，例如将文本中的片段随机删除一部分，如10％左右的片段内容，通过随机的信息删除处理可得到对应的删除文本，该删除文本也是保留了原本获取的文本维度的属性描述信息中的部分信息，并且可以避免某些强指代的文本片段过多的问题，减少数据冗余，强化全局信息。数据处理设备可将删除文本作为文本维度下的属性描述信息，按照所匹配的数据增强算法进行数据增强后的属性描述信息。利用删除文本训练目标分类模型，可增强模型对全局信息的学习能力和鲁棒性。

[0088] 需要说明的是，对于文本维度的属性描述信息的数据增强处理，包括但不限于以上两种方式，除此之外，还可以采用其他方式实现，例如同义词替换、随机插入、随机替换等，在此不做限制。对于获取到的文本维度的属性描述信息也可以采用多种方式的组合进行数据增强，例如一部分属性描述信息采用方式一，其他部分属性描述信息可采用方式二进行组合处理。

[0089] 在一个实现方式中，在第二属性数据包括对应模态维度为图像维度的属性描述信息时，确定出的匹配的数据增强算法是与图像维度匹配的数据增强算法，图像维度的属性描述信息可以理解为图像模态的数据，在对相应模态的数据进行数据增强时，具体可采用

与图像维度匹配的数据增强算法，对图像模态维度下的属性描述信息进行数据增强处理，详细地可包括以下两种方式。

[0090] 方式一、从第二属性数据中获取图像维度的属性描述信息，并获取目标噪声；将目标噪声叠加到图像维度的属性描述信息中，得到图像维度的属性描述信息对应的噪声图像；噪声图像被作为对获取出的图像维度的属性描述信息进行数据增强处理后的属性描述信息。

[0091] 数据处理设备可获取目标噪声，该目标噪声可以是高斯噪声、脉冲噪声、白噪声等中的任一种，接着可将目标噪声叠加到第二属性数据的图像维度的属性描述信息中，相当于在图像原始信息的基础上叠加噪声，由于图像维度的属性描述信息是图像数据，因此在噪声叠加之后可得到对应的噪声图像，数据处理设备可将噪声图像作为图像维度的属性描述信息，按照与图像维度匹配的数据增强算法进行数据增强处理后的属性描述信息。通过对图像数据加入噪声，可以模拟不同图像环境下图像的细微变化，采用噪声叠加后得到的噪声图像对目标分类模型进行训练，可增强模型的抗干扰能力。

[0092] 方式二、从第二属性数据中获取图像维度的属性描述信息，并从图像维度的属性描述信息中确定出目标图像区域；对处于目标图像区域的图像像素的像素值进行随机替换处理，得到目标图像区域对应的遮挡图像区域；其中，包含遮挡图像区域的图像维度的属性描述信息被作为对获取出的图像维度的属性描述信息进行数据增强处理后的属性描述信息。

[0093] 从第二属性数据中获取出的图像维度的属性描述信息是原始图像，数据处理设备可在原始图像中随机选取出一个图像区域，并将其作为目标图像区域，然后，数据处理设备可将目标图像区域中的图像像素的像素值替换为随机值，实现对目标图像区域中的像素值的变更，也相当于对原始图像进行了部分的遮挡。那么，像素值替换后的目标图像区域即可视为遮挡图像区域，实现对原始图像的随机擦除。包含遮挡图像区域的图像便可作为对获取出的图像维度的属性描述信息，按照所匹配的数据增强算法进行数据增强后的属性描述信息。采用此类数据对目标分类模型进行训练，也可以增强模型对全局信息的学习。

[0094] 需要说明的是，对于图像维度的同一属性描述信息，可以采用上述方式执行多次，从而得到多个数据增强后的属性描述信息。举例来说，针对同一张图像，可以复制多张与之相同的图像，然后采用同一方式对每张图像进行增强处理，如5张图像选择的目标图像区域均不同，对每个目标图像区域内的图像像素的像素值替换为随机值之后，可以得到5张不同的图像，即可作为数据增强后的图像。此外，与文本维度的属性描述信息类似，对于图像维度的属性描述信息的数据增强处理也可以组合以上两种方式，即一部分属性描述信息采用噪声叠加的方式实现数据增强，另一部分属性描述信息采用像素值随机替换的方式实现数据增强，以此变换出不同的增强数据来训练目标分类模型，提升目标分类模型的抗干扰能力以及对全局信息的学习。此外，对于图像维度的属性描述信息的数据增强处理还可以包括其他方式，例如颜色变换(如增加或减少某些颜色分量)、几何变换(如翻转、裁剪、变形、缩放等各类操作)等，本申请中对此不做限制。

[0095] 通过对无标注属性数据进行数据增强处理，可以进一步扩充模型训练的样本数据，使用数据增强前和数据增强后的无标注属性数据进行训练，能够使得模型见到更多不同的属性数据，在训练完成之后，模型对产生细微变化的数据的识别更加稳定，模型鲁棒性

得到有效增强，从而提升模型的整体学习效果。

[0096] 在对第二属性数据进行数据增强处理后，数据处理设备可获取到增强数据，并可调用目标分类模型对增强数据以及第二属性数据进行识别处理。在一个实施例中，调用目标分类模型对第二属性数据和增强数据进行的识别处理，与调用目标分类模型对第一属性数据进行的识别处理逻辑是相同的，即：先调用已完成预训练的编码模块对第二属性数据(或是增强数据)进行特征编码处理，得到第二属性数据(或是增强数据)的编码特征，然后采用未完成训练的识别模块，根据第二属性数据(或是增强数据)的编码特征对应用程序的应用类型进行识别处理，得到应用程序基于第二属性数据(或是增强数据)的识别类型，

[0097] S305，确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异，并根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型。

[0098] 应用程序基于第二属性数据的识别类型可作为应用程序基于增强数据的识别类型的参考数据，第二差异用于反映无标注属性数据在数据增强前后的结果识别差异。可选地，基于第二属性数据的识别类型和基于增强数据的识别类型可通过分类分布来表示，基于此，第二差异通过KL(Kullback-Leibler)散度计算的损失值来衡量，如果KL散度越大，说明两个分类分布差异越大，进而说明，模型对输入样本的细微变化产生了抖动，需要对模型进行调整。

[0099] 在目标分类模型包含已完成预训练的编码模块以及未完成训练的识别模块时，基于第一差异和第二差异对目标分类模型进行的模型训练，包括对已完成预训练的编码模块的训练，以及对未完成训练的识别模块的训练。这是因为，虽然已完成预训练的编码模块(一种预训练模型)已经可以对输入的属性描述信息(例如文本数据和图像数据)进行较好地编码，但为了获得更优质的特征表示，仍需要在具体任务的场景下进行训练，通过训练可对预训练模型进行微调(fine-tune)。可以理解的是，在目标分类模型的模型训练完成之后，对于识别模块的训练也就完成了，训练完成的目标分类模型包含已完成训练的识别模块以及微调后的编码模块。

[0100] 在一个实施例中，根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型的实现方式，可以包括：获取为第一差异设置的第一训练权重以及为第二差异设置的第二训练权重；采用第一训练权重和第二训练权重分别对第一差异和第二差异进行加权求和处理，得到用于训练目标分类模型的目标差异，并采用目标差异训练目标分类模型，得到训练完成的目标分类模型。

[0101] 在对目标分类模型进行训练的阶段，不同差异给目标分类模型带来的训练影响有所不同，本申请中基于不同差异联合对目标分类模型进行训练时，可设置相应的训练权重来控制差异给目标分类模型所带来的影响，以更好地训练目标分类模型。本申请中为第一差异设置的第一训练权重和为第二差异设置的第二训练权重可以相同或不同，数据处理设备在获取到第一训练权重和第二训练权重之后，可根据第一训练权重和第一差异确定新的第一差异，根据第二训练权重和第二差异确定新的第二差异，再将新的第一差异和新的第二差异进行求和处理，可以得到用于训练目标分类模型的目标差异。

[0102] 第一差异为有标注部分的损失(Loss)，第二差异为无标注部分的损失，通过对有标注部分和无标注部分的Loss相加，可得到最终的Loss，即目标差异。对于上述加权求和的

计算，以第一差异为交叉熵损失，第二差异为KL散度为例进行示例性说明，具体计算表达式如下式1。

[0103] $$\min J(\theta) = E_{x1,y\in L}[-log p_\theta(y|x1)] + \lambda E_{x2\in U} E_{\hat{x}\sim q(\hat{x}|x2)\in U}\left[D_{KL}\left(p_\theta(y'|x2)||p_\theta(y'|\hat{x})\right)\right] \text{式 } 1$$

[0104] 其中，x1表示第一属性数据(即存在标注标签的属性数据)，y表示标注标签，L表示应用程序的第一属性数据的集合，θ表示目标分类模型的模型参数，$p_\theta(y|x_1)$表示目标分类模型对第一属性数据进行识别所输出的概率分布，$E_{x,y\in L}$表示对所有第一属性数据的交叉熵损失求平均值，x2表示第二属性数据(即不存在标注标签的属性数据)，$\hat{x}$表示增强数据(即增强后的第二属性数据)，$q(\hat{x}|x2)$表示在第二属性数据与增强数据之间的分布关系，y′表示目标分类模型基于第二属性数据对应用程序进行预测得到的识别类型，U表示应用程序的第二属性数据的集合，$E_{x2\in U}E_{\hat{x}\sim q(\hat{x}|x2)\in U}$表示对第二属性数据和增强数据之间的KL散度求期望，D_{KL}表示KL散度计算，$p_\theta(y'|x_2)$表示基于第二属性数据得到的识别类型的概率分布，$p_\theta(y'|\hat{x})$表示基于增强数据得到的识别类型的概率分布，第一训练权重为常数1，λ为第二训练权重。

[0105] 当第一训练权重和第二训练权重相同时，第一差异和第二差异在目标差异中所占的比重相同，对目标分类模型的训练影响是等同的，当第一训练权重和第二训练权重不同时，第一差异和第二差异在目标差异中所占的比重不同，且对目标分类模型的训练影响是不同的。在采用目标差异训练目标分类模型时，可通过目标差异的反向传播来对目标分类模型的模型参数进行调整，模型训练优化的目标即最小化目标差异，在目标分类模型满足收敛条件时，可得到训练完成的目标分类模型。其中，收敛条件可以是迭代次数达到次数阈值或者是目标差异小于预设差异，该预设差异是最小差异值，用于表示目标分类模型对应用程序的识别准确度已达到期望的识别准确度，以及目标分类模型的稳定性达到期望的稳定性。

[0106] 由上可见，在对目标分类模型的训练过程中，在有监督的第一差异(如交叉熵Loss)的基础上，还增加了第二属性数据(一种无标签数据)在数据增强前后的分类结果的第二差异，这两个部分的差异结合对目标分类模型进行训练，可以使得训练完成的目标分类模型不仅有准确的识别能力，还具备较好的抗干扰能力，从而能够稳定、准确地进行应用类型的识别。

[0107] 对于本申请实施例所介绍的总体训练流程，可参考图5所示的应用程序的分类方法整体流程图，总体包括两个步骤。

[0108] 步骤1：获取应用程序的多模态数据。应用程序的多模态数据是多模态的属性数据，具体包含多个模态维度下的属性描述信息，具体可划分为有标注数据和无标注数据，有标注数据是存在标注标签的属性数据，无标注数据是不存在标注标签的属性数据，这些属性数据均涉及多个模态维度。

[0109] 步骤2：基于应用程序的有标注数据和无标注数据共同训练。通过应用程序的有标注数据和无标注数据可共同训练目标分类模型，对目标分类模型同时进行有监督训练和无监督训练，提升模型在应用类型识别上的准确度和稳定性。具体地，此步骤下包括两个子步骤，分别为：基于预训练模型的特征表示构造(步骤2.1)以及基于构造的特征表示计算损失

(步骤2.2)。其中,通过预训练模型进行特征表示的构造,可以从多个模态维度提取出应用程序的描述特征,并将各个描述特征进行拼接得到用于抽象化表示应用程序的编码特征,该编码特征是一种强化的特征表示,能够更加全面地表达应用程序,目标分类模型基于该特征表示能够学习到更多的信息,在模型训练完成之后的实际应用中,具备更强的分类能力。然后,基于构造的特征表示(即编码特征)可计算相应的损失,具体包括有标注数据对应的损失以及无标注数据在增强前后对应的损失,这两个损失可作为最终的目标损失,模型训练优化的目标即最小化该目标损失。通过以上步骤,在达到优化目标时可获得一个基于数据增强训练的多模态分类模型,在应用于APP分类场景下,可为APP打上相应的预定义标签,便于APP推广运营。

[0110] 进一步地,对于目标分类模型的训练原理示意图,具体参见图6a和图6b。如图6a中包括在有标注情况下的有监督训练以无标注数据通过数据增强进行的无监督训练。其中,有监督训练下可将有标注数据通过特征编码后进行分类,然后与标注标签计算交叉熵损失Loss,无监督训练下可将无标注数据通过相应模态下的数据增强的方式,将其变化成增强数据,增强数据保留了无标注数据的部分信息,也做了一定程度的变化。对获得的无标注数据以及增强数据进行特征编码处理并分类,可得到两个分类结果,该分类结果具体是分类分布,可计算KL散度边将其作为无标注部分的Loss。最终将交叉熵损失和KL散度进行加权求和,可得到目标损失,进而基于该目标损失对目标分类模型的模型参数进行调整。基于图6a所示目标分类模型的训练原理示意图,更简洁地,也可以参见图6b,如图6b所示的目标分类模型采用了图4所示的结构,对于各个数据均会经过目标分类模型进行处理,从而输出相应的类型以进行损失计算,并基于损失的反向传播对目标分类模型的模型参数进行调整,实现对目标分类模型的训练。当应用于APP分类场景下,无标注数据和有标注数据可以是APP原始的多模态数据。

[0111] 经实验表明,在APP分类场景下,若预定义分类体系的标注样本不足时,本申请实施例提供的数据处理方法可通过数据增强和多模态的方式,来增强模型的APP表示学习能力以及模型的鲁棒性,从而可以提高分类准确率和召回率。在多模态信息上,相较于仅仅使用APP简介,图片和标签均可以提高整体的F值(即正确率和召回率的调和平均值)。在评测集上,在简介上增加图片模态可以提高1%的F值;再增加标签模态可以提高4%的F值。在数据增强方面,相较于仅仅使用有标注的交叉熵Loss进行训练的方式,增加了无标注的数据增强前后的KL散度Loss可以提高5%的F值。可以看出本申请提出的数据处理方法,应用于APP分类时是一种基于数据增强和多模态表示学习的APP分类方法,可以提高模型的分类能力,增加模型的鲁棒性。

[0112] 在一个可行的实施方式中,为了使得目标分类模型得到更好的训练效果,在对目标分类模型进行模型训练时,包含至少两个训练阶段。通过在不同的训练阶段对相应的训练权重进行调整,可以使得目标分类模型在不同的训练阶段着重进行相应能力的训练,从而更加高效地提升目标分类模型的训练效果。在一种实施方式中,可以:获取当前训练阶段及为目标分类模型设置的目标训练阶段;在当前训练阶段为目标训练阶段之前的训练阶段时,调整第二训练权重的取值,以使调整后的第二训练权重小于第一训练权重;在当前训练阶段为目标训练阶段时,调整第二训练权重的取值,以使调整后的第二训练权重大于第一训练权重。

[0113] 数据处理设备可为目标分类模型设置目标训练阶段，该目标训练阶段是目标分类模型在训练达到预设稳定条件之后才进入的一个训练阶段，此处的预设稳定条件用于表示目标分类模型对应用程序的应用类型的识别开始步入相对稳定的阶段。预设稳定条件可以是目标差异小于差异阈值，该差异阈值大于目标分类模型收敛时使用的预设差异。数据处理设备可判断获取到的当前训练阶段是否为目标训练阶段之前的训练阶段，具体可以通过当前训练阶段所得到的目标差异是否达到差异阈值来判断。若当前训练阶段为目标训练阶段之前的训练阶段，则说明目标分类模型对应用程序的应用类型的识别稳定程度还比较差，因此，可以将第二训练权重调整为一个小于第一训练权重的取值，这样设置能够使得目标差异中第一差异的占比大于第二差异的占比，基于第一训练权重和第一差异对目标分类模型的有监督训练，目标分类模型能够见到更多存在标注标签的属性数据(即真实的样本数据)，快速提升模型的分类准确度。若当前训练阶段为目标训练阶段，那么说明当前训练阶段下目标分类模型处理得到的目标差异更加稳定，此时可以将第二训练权重调整为一个大于第一训练权重的取值，以使得后续得到的目标差异中第二差异的占比大于第一差异的占比，即第二差异给目标分类模型所带来的影响更大，目标分类模型能够见到更多不存在标注标签的属性数据(即无标注的样本数据)，基于目标差异对目标分类模型的训练，能够增强非标注属性数据的学习，提升模型的稳定性。

[0114] 由上可见，模型训练所包含的两个训练阶段分别着重于训练模型的分类准确度以及模型的分类稳定性，通过目标训练阶段之前的训练阶段将模型训练到相对稳定时，在对无标注属性数据的识别处理可得到更加准确的识别类型，进而更高效地提升模型的稳定性。可以理解的是，在目标分类模型的整个训练过程中也可以不调整第一训练权重和第二训练权重。例如第一训练权重和第二训练权重均采用常数1，这样相当于直接采用第一差异和第二差异的和值对目标分类模型进行训练，也能够达到同样的训练效果。

[0115] 在一种实现方式中，在通过模型训练得到训练完成的目标分类模型之后，数据处理设备可调用训练完成的目标分类模型对待分类对象进行应用类型的识别处理，识别出的应用类型可作为应用程序的分类标签。以本方案应用的场景为对应用程序(即APP)进行分类的场景为例，如图7所示的应用程序分类的处理示意图。通过给定的预定义的分类体系以及相应的标注属性数据，训练完成的目标分类模型可包含用于生成APP表示的生成模型(相当于训练完成的编码模块)以及用于根据APP表示对APP进行分类的分类器(相当于训练完成的识别模块)，这样，对APP存在两种形式的刻画，一种是隐式的刻画，即通过分类模型学习到的APP向量表示对APP进行刻画，在一些应用场景下，APP向量表示可用于计算APP之间的相似度，进而选取出相似APP，并将需要推广的APP推送给使用相似APP的对象群体。另一种是显式的刻画，即通过分类器可得到APP的分类标签，通过给APP打上预定义的分类标签，可达到抽象化理解APP内容的效果，进而可以根据分类标签辅助构建对象画像或者构建APP对应的广告特征。

[0116] 在一个实施例中，应用程序存在的标注标签的数量为多个，说明应用程序可归属于不同的分类体系，各个分类体系下均对应有预定义的标注标签，利用存在多个标注标签的属性数据对目标分类模型进行多任务训练，相比于一个分类体系下的单一训练任务，训练完成的目标分类模型具备对应用程序进行多分类的能力，即识别出应用程序在各个分类体系下所属的类别。举例来说，应用程序为游戏APP，那么可以从游戏画风、游戏类型这两个

分类体系下为游戏APP标注不同的标签。在多任务训练场景下对目标分类模型进行训练的过程中，各个差异可以是每个训练任务的损失Loss之和。第一差异的确定方式可以包括如下内容：从目标分类模型输出的多个参考应用类型中，确定出与应用程序的一个标注标签关联的一个参考应用类型；基于关联的一个标注标签和一个参考应用类型构建一个子差异，并将得到的全部子差异作为参考应用类型和标注应用类型之间的第一差异。

[0117] 具体地，目标分类输出的多个参考应用类型属于不同分类体系下的应用类型，对于每个参考应用类型均对应关联有对应分类体系下的标注标签，应用程序的一个标注标签可关联一个参考应用类型。对于一个分类体系下的参考应用类型，数据处理设备可构建标注标签和关联的参考应用类型之间的子差异，每个分类体系均可对应得到一个子差异，从而可以将各个子差异作为参考应用类型和标注应用类型之间的第一差异，具体可将各个子差异的和值作为第一差异。

[0118] 需要说明的是，在多任务分类场景下对于第二差异的确定也是相同的原理，即计算每个分类体系下得到的识别类型之间的子差异，并将各个子差异作为第二差异。举例来说，在多训练任务下的Loss可以升级为MMOE(Multi-gate Mixture-of-Experts，一种多任务学习模型)模式下的Loss，在MMOE模式下，底层网络包括多个专家网络(expertnetworks)，每个专家网络执行一个分类任务，每个任务使用单独门控网络(gate networks)，每个任务的门控网络通过最终输出权重不同实现对专家网络的选择性利用，这样不同任务的门控网络可以学习到不同的组合专家网络的模式，进而更好地捕捉到任务的相关性与区别。

[0119] 请参见图8，图8是本申请实施例提供的一种数据处理装置的结构示意图。上述数据处理装置可以是运行于数据处理设备中的一个计算机程序(包括程序代码)，例如该数据处理装置为一个应用软件；该数据处理装置可以用于执行本申请实施例提供的数据处理方法中的相应步骤。如图8所示，该数据处理装置800可以包括：获取模块801、识别模块802、确定模块803、训练模块804。

[0120] 获取模块801，用于获取应用程序的第一属性数据和第二属性数据；第一属性数据是存在标注标签及一个或多个模态维度属性数据，标注标签用于指示应用程序的标注应用类型；第二属性数据是不存在标注标签的属性数据；

[0121] 识别模块802，用于调用目标分类模型对第一属性数据下各模态维度的属性数据进行识别处理，得到应用程序的参考应用类型；

[0122] 确定模块803，用于确定出参考应用类型和标注应用类型之间的第一差异；

[0123] 获取模块801，还用于获取第二属性数据对应的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到应用程序基于第二属性数据的识别类型和应用程序基于增强数据的识别类型；

[0124] 确定模块803，还用于确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异；

[0125] 训练模块804，用于根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型；训练完成的目标分类模型用于进行应用类型的识别。

[0126] 在一个实施例中，获取模块801，用于：获取对应用程序进行应用属性描述的模态维度，模态维度包含以下一个或多个：文本维度、图像维度和语音维度；基于模态维度获取

应用程序在相应模态维度下的属性描述信息，并获取应用程序的标注标签；将标注标签和相应的属性描述信息进行关联，并将关联标注标签的属性描述信息作为第一属性数据。

[0127] 在一个实施例中，目标分类模型是已完成预训练的目标分类模型，其中，已完成预训练的目标分类模型包括已完成预训练的编码模块；识别模块802，具体用于：调用目标分类模型中已完成预训练的编码模块对第一属性数据进行特征编码处理，得到第一属性数据的编码特征；目标分类模型还包含未完成训练的识别模块；采用未完成训练的识别模块，根据第一属性数据的编码特征对应用程序的应用类型进行识别处理，得到应用程序的参考应用类型。

[0128] 在一个实施例中，若第一属性数据包含多个模态维度的属性描述信息；识别模块802，具体用于：调用目标分类模型中已完成预训练的编码模块，分别对第一属性数据中包含的不同模态维度的属性描述信息进行特征编码处理，得到第一属性数据中相应模态维度下的属性描述信息对应的描述特征；将不同模态维度下的属性描述信息对应的描述特征进行拼接处理，并将拼接后的描述特征作为第一属性数据的编码特征。

[0129] 在一个实施例中，当任一模态维度下的属性描述信息为多个，且任一模态维度下的多个属性描述信息的信息类型不同时；识别模块802，具体用于：调用目标分类模型中已完成预训练的编码模块，对第一属性数据中任一模态维度下对应不同信息类型的属性描述信息分别进行编码处理，得到任一模态维度下相应信息类型的属性描述信息对应的描述特征；将得到的各信息类型的属性描述信息对应的描述特征均作为任一模态维度下的属性描述信息对应的描述特征；或者，将基于各信息类型的属性描述信息对应的描述特征得到的拼接描述特征，作为任一模态维度下的属性描述信息对应的描述特征。

[0130] 在一个实施例中，第二属性数据为包含多个模态维度的属性描述信息的数据；获取模块801，具体用于：确定出第二属性数据中的各属性描述信息分别对应的模态维度，并获取与相应模态维度匹配的数据增强算法；采用匹配的数据增强算法，对相应模态维度下的属性描述信息进行数据增强处理，并将增强处理后的属性描述信息作为第二属性数据的增强数据。

[0131] 在一个实施例中，在第二属性数据包括对应模态维度为文本维度的属性描述信息时，确定出的匹配的数据增强算法是与文本维度匹配的数据增强算法；获取模块801，具体用于：从第二属性数据中获取文本维度的属性描述信息，并对获取出的文本维度的属性描述信息进行翻译处理，得到文本维度的属性描述信息对应的翻译文本；对翻译文本进行回译处理，得到翻译文本的回译文本；回译文本被作为对获取出的文本维度的属性描述信息进行数据增强处理后的属性描述信息。

[0132] 在一个实施例中，在第二属性数据包括对应模态维度为文本维度的属性描述信息时，确定出的匹配的数据增强算法是与文本维度匹配的数据增强算法；获取模块801，具体用于：从第二属性数据中获取文本维度的属性描述信息，并对获取出的文本维度的属性描述信息进行随机的信息删除处理，得到文本维度的属性描述信息对应的删除文本；其中，删除文本被作为对获取出的文本维度的属性描述信息进行数据增强处理后的属性描述信息。

[0133] 在一个实施例中，在第二属性数据包括对应模态维度为图像维度的属性描述信息时，确定出的匹配的数据增强算法是与图像维度匹配的数据增强算法；获取模块801，具体用于：从第二属性数据中获取图像维度的属性描述信息，并获取目标噪声；将目标噪声叠加

到图像维度的属性描述信息中，得到图像维度的属性描述信息对应的噪声图像；噪声图像被作为对获取出的图像维度的属性描述信息进行数据增强处理后的属性描述信息。

[0134] 在一个实施例中，在第二属性数据包括对应模态维度为图像维度的属性描述信息时，确定出的匹配的数据增强算法是与图像维度匹配的数据增强算法；获取模块801，具体用于：从第二属性数据中获取图像维度的属性描述信息，并从图像维度的属性描述信息中确定出目标图像区域；对处于目标图像区域的图像像素的像素值进行随机替换处理，得到目标图像区域对应的遮挡图像区域；其中，包含遮挡图像区域的图像维度的属性描述信息被作为对获取出的图像维度的属性描述信息进行数据增强处理后的属性描述信息。

[0135] 在一个实施例中，训练模块804，具体用于：获取为第一差异设置的第一训练权重以及为第二差异设置的第二训练权重；采用第一训练权重和第二训练权重分别对第一差异和第二差异进行加权求和处理，得到用于训练目标分类模型的目标差异，并采用目标差异训练目标分类模型，得到训练完成的目标分类模型。

[0136] 在一个实施例中，在对目标分类模型进行模型训练时，包含至少两个训练阶段；训练模块804，还用于：获取当前训练阶段及为目标分类模型设置的目标训练阶段；在当前训练阶段为目标训练阶段之前的训练阶段时，调整第二训练权重的取值，以使调整后的第二训练权重小于第一训练权重；在当前训练阶段为目标训练阶段时，调整第二训练权重的取值，以使调整后的第二训练权重大于第一训练权重。

[0137] 在一个实施例中，若应用程序存在的标注标签的数量为多个，则调用目标分类模型识别第一属性数据得到的参考应用类型的数量为多个；确定模块803，具体用于：从目标分类模型输出的多个参考应用类型中，确定出与应用程序的一个标注标签关联的一个参考应用类型；基于关联的一个标注标签和一个参考应用类型构建一个子差异，并将得到的全部子差异作为参考应用类型和标注应用类型之间的第一差异。

[0138] 可以理解的是，本申请实施例所描述的数据处理装置的各功能模块的功能可根据上述方法实施例中的方法具体实现，其具体实现过程可以参照上述方法实施例的相关描述，此处不再赘述。另外，对采用相同方法的有益效果描述，也不再进行赘述。

[0139] 请参见图9，图9是本申请实施例提供的一种数据处理设备的结构示意图。该数据处理设备900可以包含独立设备(例如节点、终端等等中的一个或者多个)，也可以包含独立设备内部的部件(例如芯片、软件模块或者硬件模块等)。该数据处理设备900可以包括至少一个处理器901和网络接口902，进一步可选地，数据处理设备900还可以包括至少一个存储器903和总线904。其中，处理器901、网络接口902和存储器903通过总线904相连。

[0140] 其中，处理器901是进行算术运算和/或逻辑运算的模块，具体可以是中央处理器(central processing unit,CPU)、图片处理器(graphics processing unit,GPU)、微处理器(microprocessor unit,MPU)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现场可编程逻辑门阵列(Field Programmable Gate Array,FPGA)、复杂可编程逻辑器件(Complex programmable logic device,CPLD)、协处理器(协助中央处理器完成相应处理和应用)、微控制单元(Microcontroller Unit,MCU)等处理模块中的一种或者多种的组合。

[0141] 网络接口902可以用于为至少一个处理器提供信息输入或者输出。和/或，网络接口902可以用于接收外部发送的数据和/或向外部发送数据，可以为包括诸如以太网电缆等

的有线链路接口，也可以是无线链路(Wi-Fi、蓝牙、通用无线传输、车载短距通信技术以及其他短距无线通信技术等)接口。网络接口902可以作为网络接口。

[0142] 存储器903用于提供存储空间，存储空间中可以存储操作系统和计算机程序等数据。存储器903可以是随机存储记忆体(random access memory，RAM)、只读存储器(read-only memory，ROM)、可擦除可编程只读存储器(erasable programmable read only memory，EPROM)、或便携式只读存储器(compact disc read-only memory，CD-ROM)等等中的一种或者多种的组合。

[0143] 该数据处理设备900中的至少一个处理器901用于调用至少一个存储器903中存储的计算机程序，执行本申请所示的实施例所描述的数据处理方法。

[0144] 在一种可能的实施方式中，该数据处理设备900中的处理器901用于调用至少一个存储器903中存储的计算机程序，用于执行以下操作：获取应用程序的第一属性数据和第二属性数据；第一属性数据是存在标注标签及一个或多个模态维度属性数据，标注标签用于指示应用程序的标注应用类型；第二属性数据是不存在标注标签的属性数据；调用目标分类模型对第一属性数据下各模态维度的属性数据进行识别处理，得到应用程序的参考应用类型，并确定出参考应用类型和标注应用类型之间的第一差异；获取第二属性数据对应的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到应用程序基于第二属性数据的识别类型和应用程序基于增强数据的识别类型；确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异，并根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型；训练完成的目标分类模型用于进行应用类型的识别。

[0145] 在一个实施例中，处理器901，用于：获取对应用程序进行应用属性描述的模态维度，模态维度包含以下一个或多个：文本维度、图像维度和语音维度；基于模态维度获取应用程序在相应模态维度下的属性描述信息，并获取应用程序的标注标签；将标注标签和相应的属性描述信息进行关联，并将关联标注标签的属性描述信息作为第一属性数据。

[0146] 在一个实施例中，目标分类模型是已完成预训练的目标分类模型，其中，已完成预训练的目标分类模型包括已完成预训练的编码模块；处理器901，具体用于：调用目标分类模型中已完成预训练的编码模块对第一属性数据进行特征编码处理，得到第一属性数据的编码特征；目标分类模型还包含未完成训练的识别模块；采用未完成训练的识别模块，根据第一属性数据的编码特征对应用程序的应用类型进行识别处理，得到应用程序的参考应用类型。

[0147] 在一个实施例中，若第一属性数据包含多个模态维度的属性描述信息；处理器901，具体用于：调用目标分类模型中已完成预训练的编码模块，分别对第一属性数据中包含的不同模态维度的属性描述信息进行特征编码处理，得到第一属性数据中相应模态维度下的属性描述信息对应的描述特征；将不同模态维度下的属性描述信息对应的描述特征进行拼接处理，并将拼接后的描述特征作为第一属性数据的编码特征。

[0148] 在一个实施例中，当任一模态维度下的属性描述信息为多个，且任一模态维度下的多个属性描述信息的信息类型不同时；处理器901，具体用于：调用目标分类模型中已完成预训练的编码模块，对第一属性数据中任一模态维度下对应不同信息类型的属性描述信息分别进行编码处理，得到任一模态维度下相应信息类型的属性描述信息对应的描述特

征;将得到的各信息类型的属性描述信息对应的描述特征均作为任一模态维度下的属性描述信息对应的描述特征;或者,将基于各信息类型的属性描述信息对应的描述特征得到的拼接描述特征,作为任一模态维度下的属性描述信息对应的描述特征。

[0149] 在一个实施例中,第二属性数据为包含多个模态维度的属性描述信息的数据;处理器901,具体用于:确定出第二属性数据中的各属性描述信息分别对应的模态维度,并获取与相应模态维度匹配的数据增强算法;采用匹配的数据增强算法,对相应模态维度下的属性描述信息进行数据增强处理,并将增强处理后的属性描述信息作为第二属性数据的增强数据。

[0150] 在一个实施例中,在第二属性数据包括对应模态维度为文本维度的属性描述信息时,确定出的匹配的数据增强算法是与文本维度匹配的数据增强算法;处理器901,具体用于:从第二属性数据中获取文本维度的属性描述信息,并对获取出的文本维度的属性描述信息进行翻译处理,得到文本维度的属性描述信息对应的翻译文本;对翻译文本进行回译处理,得到翻译文本的回译文本;回译文本被作为对获取出的文本维度的属性描述信息进行数据增强处理后的属性描述信息。

[0151] 在一个实施例中,在第二属性数据包括对应模态维度为文本维度的属性描述信息时,确定出的匹配的数据增强算法是与文本维度匹配的数据增强算法;处理器901,具体用于:从第二属性数据中获取文本维度的属性描述信息,并对获取出的文本维度的属性描述信息进行随机的信息删除处理,得到文本维度的属性描述信息对应的删除文本;其中,删除文本被作为对获取出的文本维度的属性描述信息进行数据增强处理后的属性描述信息。

[0152] 在一个实施例中,在第二属性数据包括对应模态维度为图像维度的属性描述信息时,确定出的匹配的数据增强算法是与图像维度匹配的数据增强算法;处理器901,具体用于:从第二属性数据中获取图像维度的属性描述信息,并获取目标噪声;将目标噪声叠加到图像维度的属性描述信息中,得到图像维度的属性描述信息对应的噪声图像;噪声图像被作为对获取出的图像维度的属性描述信息进行数据增强处理后的属性描述信息。

[0153] 在一个实施例中,在第二属性数据包括对应模态维度为图像维度的属性描述信息时,确定出的匹配的数据增强算法是与图像维度匹配的数据增强算法;处理器901,具体用于:从第二属性数据中获取图像维度的属性描述信息,并从图像维度的属性描述信息中确定出目标图像区域;对处于目标图像区域的图像像素的像素值进行随机替换处理,得到目标图像区域对应的遮挡图像区域;其中,包含遮挡图像区域的图像维度的属性描述信息被作为对获取出的图像维度的属性描述信息进行数据增强处理后的属性描述信息。

[0154] 在一个实施例中,处理器901,具体用于:获取为第一差异设置的第一训练权重以及为第二差异设置的第二训练权重;采用第一训练权重和第二训练权重分别对第一差异和第二差异进行加权求和处理,得到用于训练目标分类模型的目标差异,并采用目标差异训练目标分类模型,得到训练完成的目标分类模型。

[0155] 在一个实施例中,在对目标分类模型进行模型训练时,包含至少两个训练阶段;处理器901,还用于:获取当前训练阶段及为目标分类模型设置的目标训练阶段;在当前训练阶段为目标训练阶段之前的训练阶段时,调整第二训练权重的取值,以使调整后的第二训练权重小于第一训练权重;在当前训练阶段为目标训练阶段时,调整第二训练权重的取值,以使调整后的第二训练权重大于第一训练权重。

[0156] 在一个实施例中，若应用程序存在的标注标签的数量为多个，则调用目标分类模型识别第一属性数据得到的参考应用类型的数量为多个；处理器901，具体用于：从目标分类模型输出的多个参考应用类型中，确定出与应用程序的一个标注标签关联的一个参考应用类型；基于关联的一个标注标签和一个参考应用类型构建一个子差异，并将得到的全部子差异作为参考应用类型和标注应用类型之间的第一差异。

[0157] 应当理解，本申请实施例中所描述的数据处理设备900可执行前文所对应实施例中对该数据处理方法的描述，也可执行前文图8所对应实施例中对该数据处理装置800的描述，在此不再赘述。另外，对采用相同方法的有益效果描述，也不再进行赘述。

[0158] 此外，还应指出，本申请一个示例性实施例还提供了一种存储介质，该存储介质中存储了前述数据处理方法的计算机程序，该计算机程序包括程序指令，当一个或多个处理器加载并执行该程序指令，可以实现实施例中对数据处理方法的描述，这里不再赘述，对采用相同方法的有益效果描述，也在此不再赘述。可以理解的是，程序指令可以被部署在一个或能够互相通信的多个数据处理设备上执行。

[0159] 上述计算机可读存储介质可以是前述任一实施例提供的数据处理装置或者上述数据处理设备的内部存储单元，例如数据处理设备的硬盘或内存。该计算机可读存储介质也可以是该数据处理设备的外部存储设备，例如该数据处理设备上配备的插接式硬盘，智能存储卡(smart media card，SMC)，安全数字(secure digital，SD)卡，闪存卡(flash card)等。进一步地，该计算机可读存储介质还可以既包括该数据处理设备的内部存储单元也包括外部存储设备。该计算机可读存储介质用于存储该计算机程序以及该数据处理设备所需的其他程序和数据。该计算机可读存储介质还可以用于暂时地存储已经输出或者将要输出的数据。

[0160] 本申请的一个方面，提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。数据处理设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该数据处理设备执行本申请实施例中一方面提供的方法。

[0161] 本申请实施例方法中的步骤可以根据实际需要进行顺序调整、合并和删减。

[0162] 本申请实施例装置中的模块可以根据实际需要进行合并、划分和删减。

[0163] 以上所揭露的仅为本申请的部分实施例而已，当然不能以此来限定本申请之权利范围，本领域普通技术人员可以理解实现上述实施例的全部或部分流程，并依本申请权利要求所作的等同变化，仍属于发明所涵盖的范围。

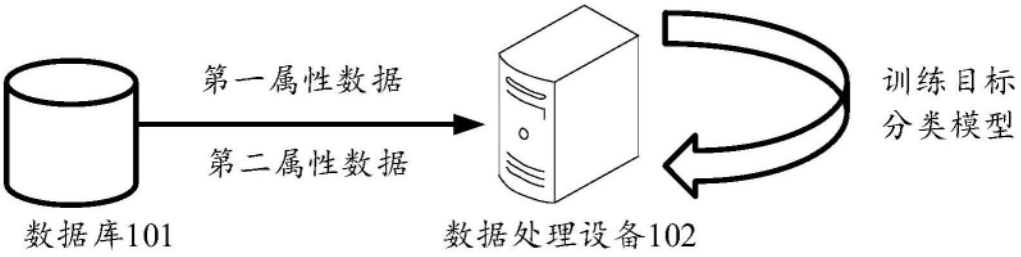

第一属性数据
第二属性数据
训练目标
分类模型
数据库101
数据处理设备102

图1

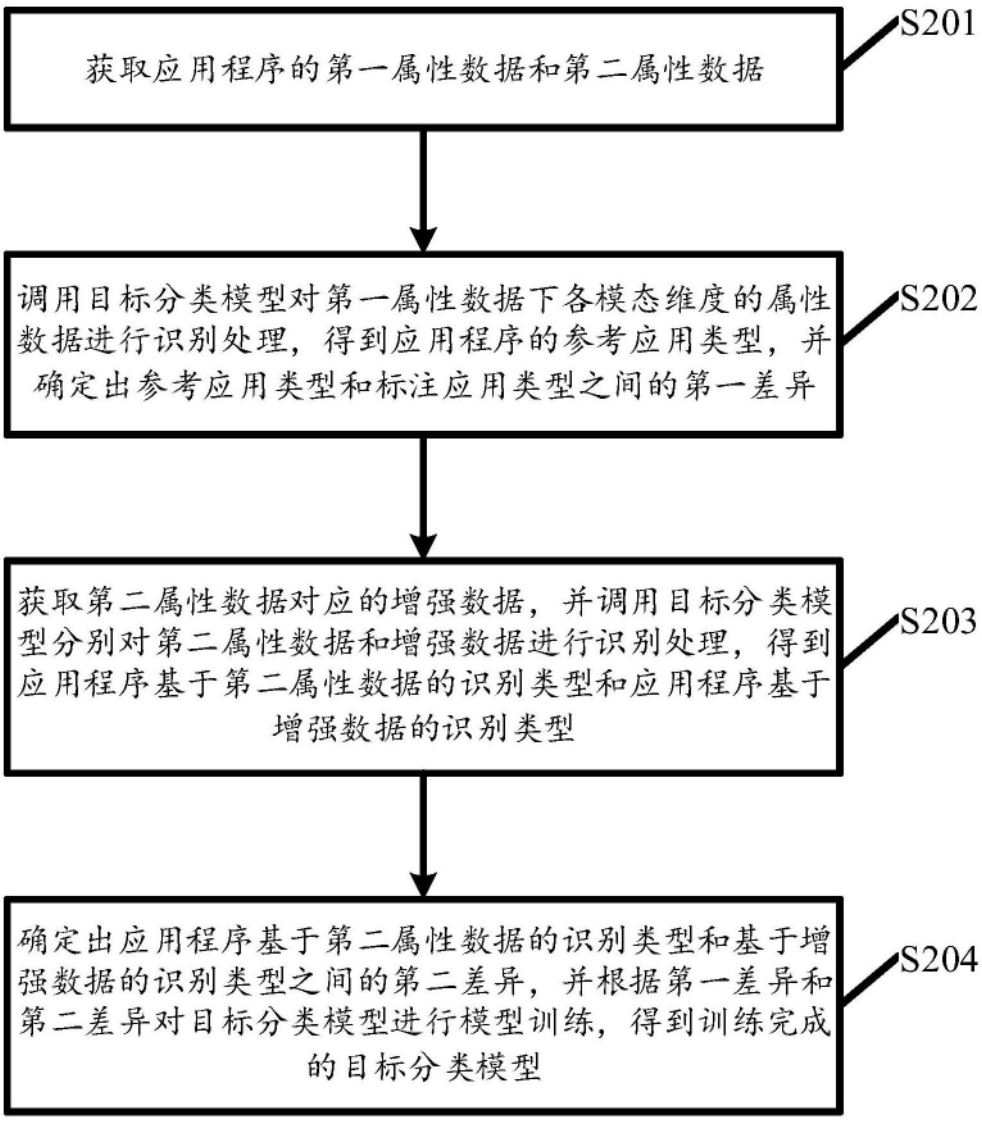

获取应用程序的第一属性数据和第二属性数据
S201
调用目标分类模型对第一属性数据下各模态维度的属性数据进行识别处理，得到应用程序的参考应用类型，并确定出参考应用类型和标注应用类型之间的第一差异
S202
获取第二属性数据对应的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到应用程序基于第二属性数据的识别类型和应用程序基于增强数据的识别类型
S203
确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异，并根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型
S204

图2

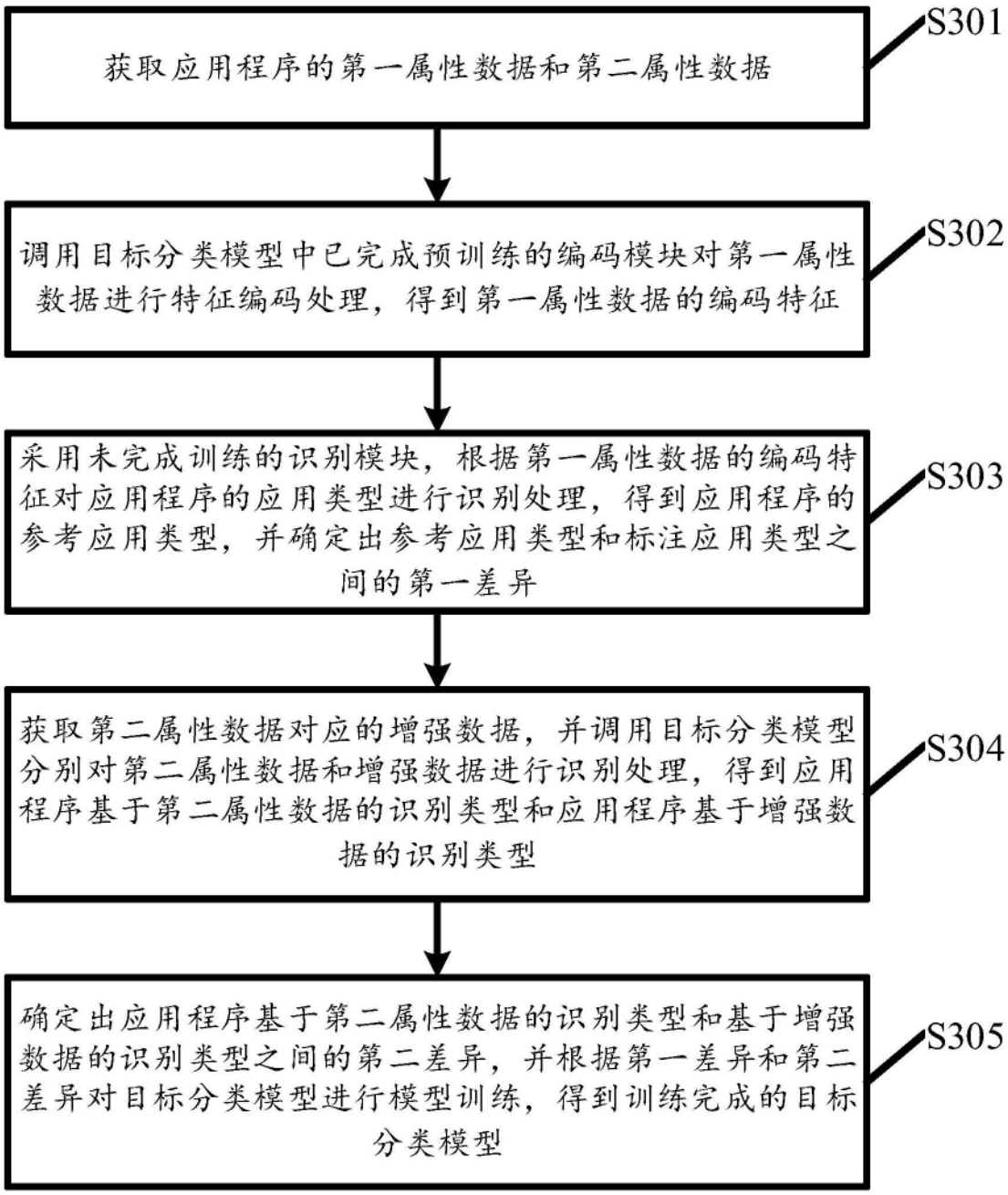

获取应用程序的第一属性数据和第二属性数据
S301
调用目标分类模型中已完成预训练的编码模块对第一属性数据进行特征编码处理，得到第一属性数据的编码特征
S302
采用未完成训练的识别模块，根据第一属性数据的编码特征对应用程序的应用类型进行识别处理，得到应用程序的参考应用类型，并确定出参考应用类型和标注应用类型之间的第一差异
S303
获取第二属性数据对应的增强数据，并调用目标分类模型分别对第二属性数据和增强数据进行识别处理，得到应用程序基于第二属性数据的识别类型和应用程序基于增强数据的识别类型
S304
确定出应用程序基于第二属性数据的识别类型和基于增强数据的识别类型之间的第二差异，并根据第一差异和第二差异对目标分类模型进行模型训练，得到训练完成的目标分类模型
S305

图3

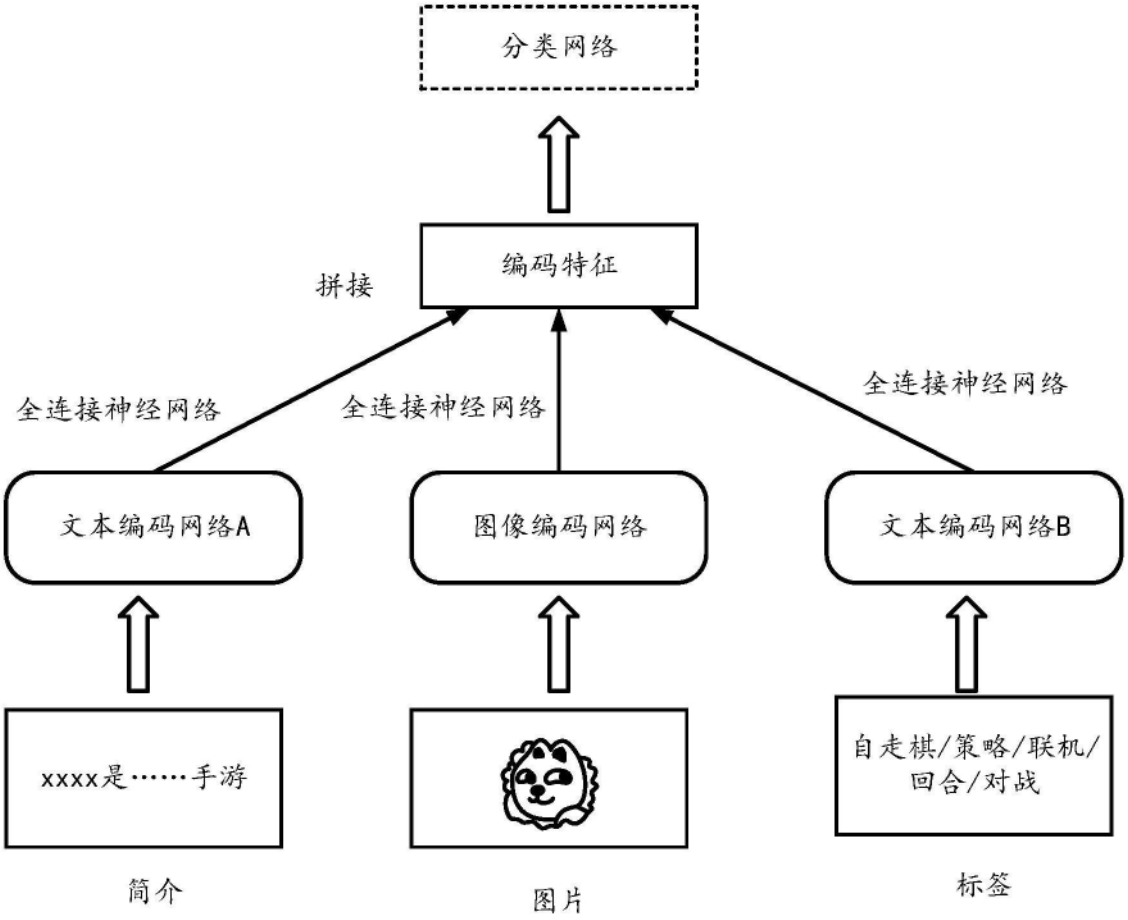

分类网络
编码特征
拼接
全连接神经网络
全连接神经网络
全连接神经网络
文本编码网络A
图像编码网络
文本编码网络B
xxxx是……手游
自走棋/策略/联机/回合/对战
简介
图片
标签

图4

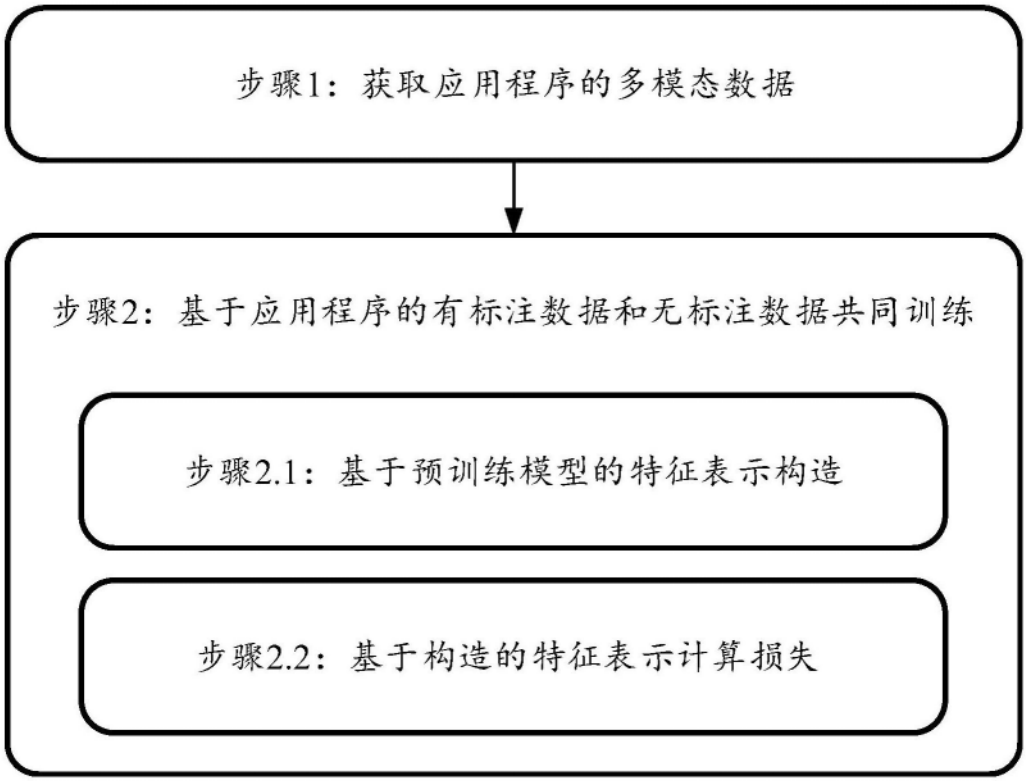

步骤1：获取应用程序的多模态数据
步骤2：基于应用程序的有标注数据和无标注数据共同训练
步骤2.1：基于预训练模型的特征表示构造
步骤2.2：基于构造的特征表示计算损失

图5

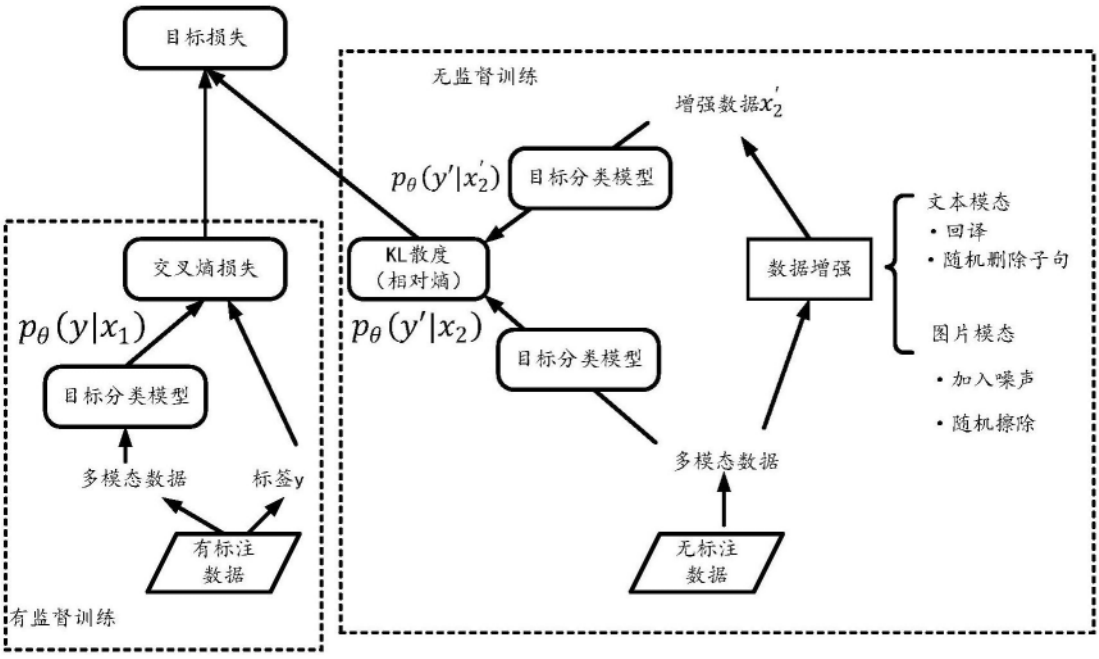

目标损失
无监督训练
增强数据x_2'
$p_\theta(y'|x_2')$
目标分类模型
KL散度
（相对熵）
$p_\theta(y'|x_2)$
目标分类模型
数据增强
文本模态
• 回译
• 随机删除子句
图片模态
• 加入噪声
• 随机擦除
交叉熵损失
$p_\theta(y|x_1)$
目标分类模型
多模态数据
标签y
有标注数据
有监督训练
多模态数据
无标注数据

图6a

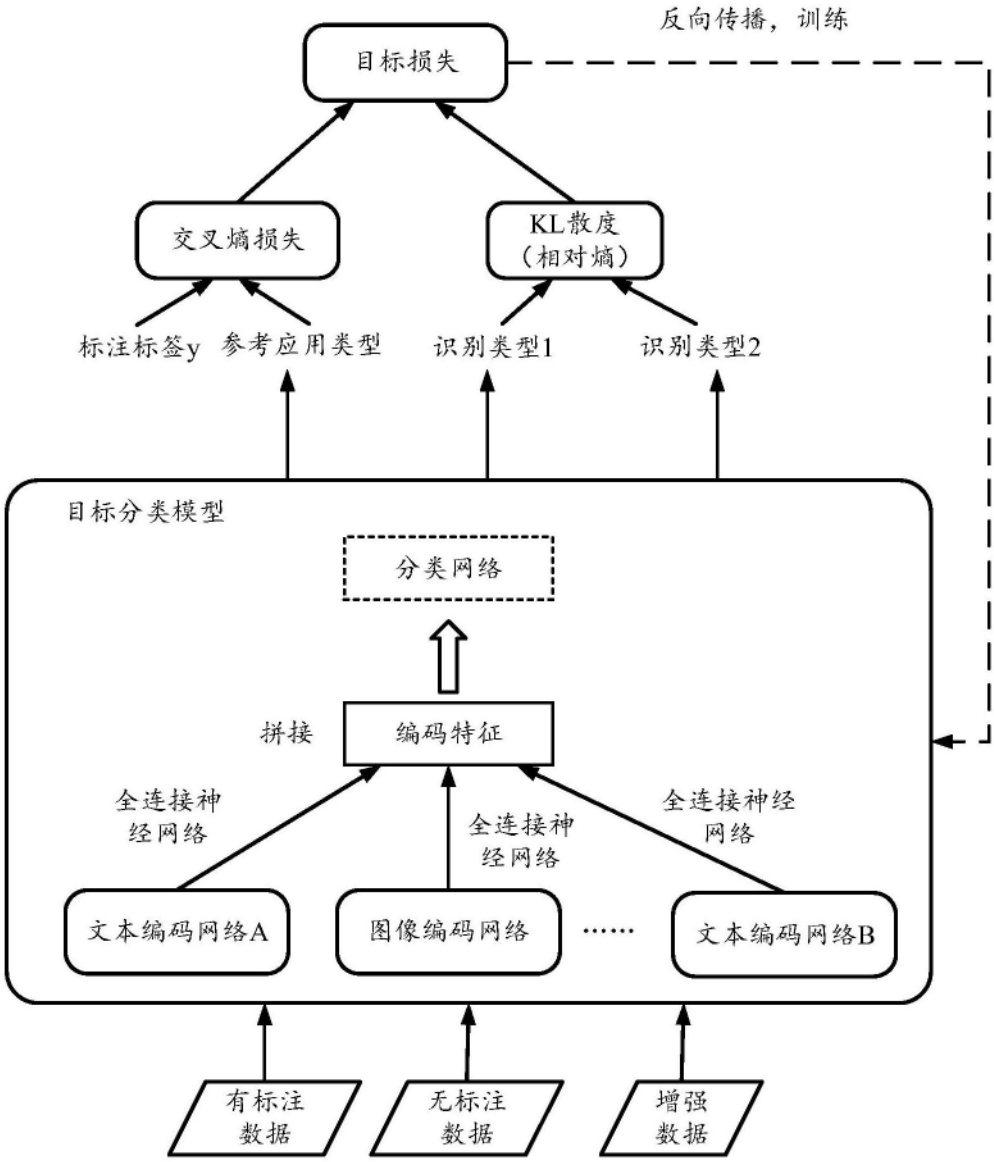

反向传播，训练
目标损失
交叉熵损失
KL散度
（相对熵）
标注标签y
参考应用类型
识别类型1
识别类型2
目标分类模型
分类网络
拼接
编码特征
全连接神经网络
全连接神经网络
全连接神经网络
文本编码网络A
图像编码网络
……
文本编码网络B
有标注数据
无标注数据
增强数据

图6b

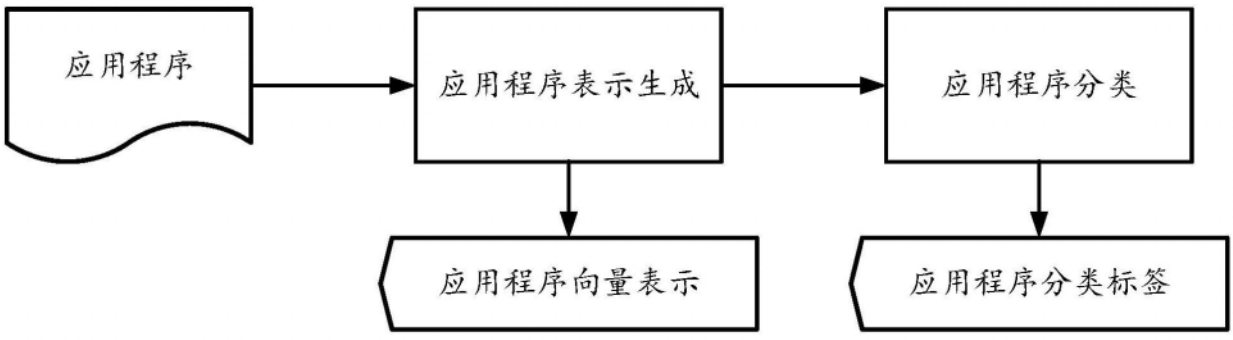

应用程序
应用程序表示生成
应用程序分类
应用程序向量表示
应用程序分类标签

图7

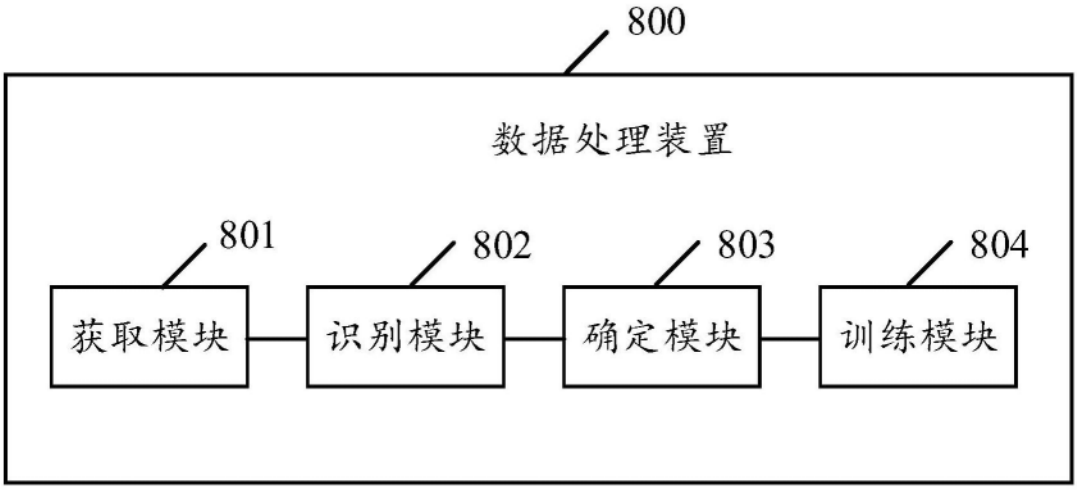

800
数据处理装置
801
获取模块
802
识别模块
803
确定模块
804
训练模块

图8

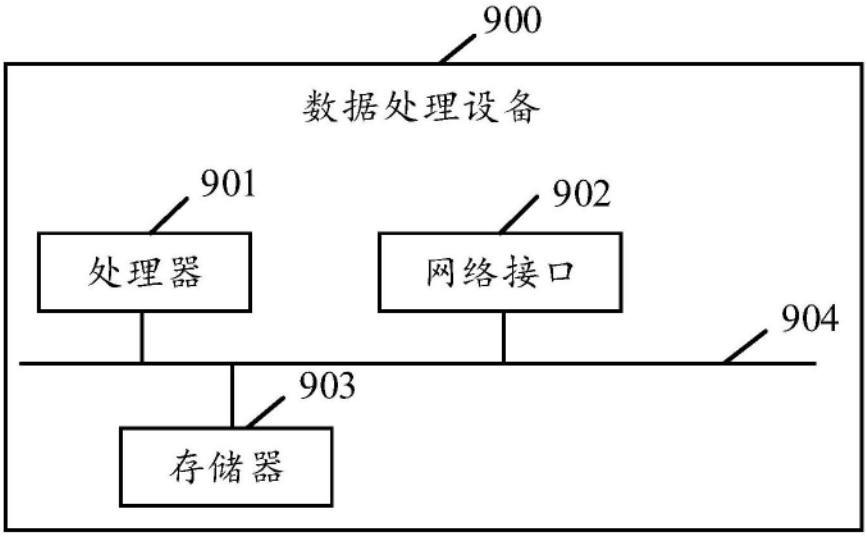

900
数据处理设备
901
处理器
902
网络接口
904
903
存储器

图9