(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111401465 A
(43)申请公布日 2020.07.10

(21)申请号 202010219658.7

(22)申请日 2020.03.25

(71)申请人 深圳前海微众银行股份有限公司
地址 518000 广东省深圳市前海深港合作区前湾一路1号A栋201室(入驻深圳市前海商务秘书有限公司)

(72)发明人 章放 邹雨晗 杨海军 徐倩 杨强

(74)专利代理机构 深圳市世纪恒程知识产权代理事务所 44287
代理人 徐进之

(51)Int.Cl.
G06K 9/62(2006.01)
G06K 9/00(2006.01)
G06K 9/54(2006.01)

权利要求书2页 说明书10页 附图2页

(54)发明名称

训练样本优化方法、装置、设备及存储介质

(57)摘要

本发明公开了一种训练样本优化方法、装置、设备及存储介质,所述方法包括:获取用于模型训练的真实数据集和生成数据集;分别从所述真实数据集和所述生成数据集中进行有放回采样,对应得到真实采样数据和生成采样数据,其中,所述真实采样数据和所述生成采样数据的数据量比例为预设的目标比例;根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集。本发明实现了能够兼顾数据量比例和训练数据量,避免因比例失衡导致的模型训练效果差,也避免因训练数据量不足而导致模型训练效果差,也即,通过对训练样本进行优化,提高了模型的训练效果。

CN 111401465 A

1.一种训练样本优化方法,其特征在于,所述训练样本优化方法包括以下步骤:

获取用于模型训练的真实数据集和生成数据集;

分别从所述真实数据集和所述生成数据集中进行有放回采样,对应得到真实采样数据和生成采样数据,其中,所述真实采样数据和所述生成采样数据的数据量比例为预设的目标比例;

根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集。

2.如权利要求1所述的训练样本优化方法,其特征在于,所述分别从所述真实数据集和所述生成数据集中进行有放回采样,对应得到真实采样数据和生成采样数据的步骤包括:

进行预设次数的采样,其中,各次采样分别从所述真实数据集和所述生成数据集中进行有放回采样,对应得到各次采样的真实采样数据和生成采样数据,各次采样时从所述真实数据集和所述生成数据集中采样的数据量比例为所述目标比例;

所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集的步骤包括:

分别将各次采样的真实采样数据和生成采样数据进行混合,得到各次采样对应的训练批,根据各所述训练批得到用于模型训练的训练样本集。

3.如权利要求1所述的训练样本优化方法,其特征在于,所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集的步骤之后,还包括:

采用所述训练样本集对待训练模型进行训练得到目标模型。

4.如权利要求3所述的训练样本优化方法,其特征在于,所述采用所述训练样本集对待训练模型进行训练得到目标模型的步骤之后,还包括:

将获取到的待识别文本图片输入所述目标模型,得到所述待识别文本图片的文本识别结果。

5.如权利要求4所述的训练样本优化方法,其特征在于,获取用于模型训练的生成数据集的步骤包括:

获取背景图片库、文本字体库和文本内容库;

从所述背景图片库中获取目标背景图片,从所述文本字体库中获取目标文本字体,以及从所述文本内容库中获取目标文本内容;

将所述目标文本内容按照所述目标文本字体拼接至所述目标背景图片中,得到生成文本图片;

根据各生成文本图片得到所述生成数据集。

6.如权利要求1至5任一项所述的训练样本优化方法,其特征在于,所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集的步骤包括:

对所述真实采样数据进行数据增强操作,并根据所述生成采样数据和数据增强后的真实采样数据得到用于模型训练的训练样本集;或,

对所述生成采样数据进行数据增强操作,并根据所述真实采样数据和数据增强后的生成采样数据得到用于模型训练的训练样本集;或,

对所述真实采样数据和所述生成采样数据进行数据增强操作,并根据数据增强后的真实采样数据和生成采样数据得到用于模型训练的训练样本集。

7.如权利要求6所述的训练样本优化方法,其特征在于,所述数据增强操作包括拉伸操

作、加噪声操作、模糊操作和锐化操作中的一种或多种。

8.一种训练样本优化装置，其特征在于，所述训练样本优化装置包括：

获取模块，用于获取用于模型训练的真实数据集和生成数据集；

采样模块，用于分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到真实采样数据和生成采样数据，其中，所述真实采样数据和所述生成采样数据的数据量比例为预设的目标比例；

样本生成模块，用于根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集。

9.一种训练样本优化设备，其特征在于，所述训练样本优化设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的训练样本优化程序，所述训练样本优化程序被所述处理器执行时实现如权利要求1至7中任一项所述的训练样本优化方法的步骤。

10.一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训练样本优化程序，所述训练样本优化程序被处理器执行时实现如权利要求1至7中任一项所述的训练样本优化方法的步骤。

训练样本优化方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及人工智能领域，尤其涉及一种训练样本优化方法、装置、设备及存储介质。

背景技术

[0002] 现有的模型训练往往是需要大量的训练数据的，但是正常情况下，训练数据的获取是很难的。

[0003] 为了解决这个问题，现有做法是：采用人工生成仿真数据的方式，比如真实的数据有1万个，但是太少了，想要10万个，剩下的9万个通过仿真数据来补足，然后就将这两种数据混合然后拿去训练。现有的这种方法，有一个缺陷：无法既保证有足够的训练数据量，又保证真实数据和仿真数据的比例维持在目标比例(例如1:1)上。因为真实数据的数据量是固定不变的，要保证足够的训练数据量，只能增加仿真数据的数据量，然而增加仿真数据会导致比例失衡；要保证比例维持在目标比例，又会导致仿真数据不能太多，从而无法达到足够的训练数据量。因此，训练数据量与比例两者不可兼得。

[0004] 当训练数据量不足时，会使得模型得不到充分训练而训练效果差；仿真数据跟真实的数据往往有着一定(甚至是较大)的差距，因此使用仿真数据进行训练得到的模型，并不如使用真实数据训练得到的模型效果好，特别是在仿真数据的数据量远大于真实数据时，也就是比例失衡时，就更加降低了模型训练效果。

发明内容

[0005] 本发明的主要目的在于提供一种训练样本优化方法、装置、设备及存储介质，旨在解决目前为解决训练样本不足所采用的方式无法兼顾训练数据量和数据比例，从而导致模型训练效果差的问题。

[0006] 为实现上述目的，本发明提供一种训练样本优化方法，所述训练样本优化方法包括以下步骤：

[0007] 获取用于模型训练的真实数据集和生成数据集；

[0008] 分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到真实采样数据和生成采样数据，其中，所述真实采样数据和所述生成采样数据的数据量比例为预设的目标比例；

[0009] 根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集。

[0010] 可选地，所述分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到真实采样数据和生成采样数据的步骤包括：

[0011] 进行预设次数的采样，其中，各次采样分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到各次采样的真实采样数据和生成采样数据，各次采样时从所述真实数据集和所述生成数据集中采样的数据量比例为所述目标比例；

[0012] 所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本

集的步骤包括：

[0013] 分别将各次采样的真实采样数据和生成采样数据进行混合，得到各次采样对应的训练批，根据各所述训练批得到用于模型训练的训练样本集。

[0014] 可选地，所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集的步骤之后，还包括：

[0015] 采用所述训练样本集对待训练模型进行训练得到目标模型。

[0016] 可选地，所述采用所述训练样本集对待训练模型进行训练得到目标模型的步骤之后，还包括：

[0017] 将获取到的待识别文本图片输入所述目标模型，得到所述待识别文本图片的文本识别结果。

[0018] 可选地，获取用于模型训练的所述生成数据集的步骤包括：

[0019] 获取背景图片库、文本字体库和文本内容库；

[0020] 从所述背景图片库中获取目标背景图片，从所述文本字体库中获取目标文本字体，以及从所述文本内容库中获取目标文本内容；

[0021] 将所述目标文本内容按照所述目标文本字体拼接至所述目标背景图片中，得到生成文本图片；

[0022] 根据各生成文本图片得到所述生成数据集。

[0023] 可选地，所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集的步骤包括：

[0024] 对所述真实采样数据进行数据增强操作，并根据所述生成采样数据和数据增强后的真实采样数据得到用于模型训练的训练样本集；或，

[0025] 对所述生成采样数据进行数据增强操作，并根据所述真实采样数据和数据增强后的生成采样数据得到用于模型训练的训练样本集；或，

[0026] 对所述真实采样数据和所述生成采样数据进行数据增强操作，并根据数据增强后的真实采样数据和生成采样数据得到用于模型训练的训练样本集。可选地，所述数据增强操作包括拉伸操作、加噪声操作、模糊操作和锐化操作中的一种或多种。

[0027] 为实现上述目的，本发明还提供一种训练样本优化装置，所述训练样本优化装置包括：

[0028] 获取模块，用于获取用于模型训练的真实数据集和生成数据集；

[0029] 采样模块，用于分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到真实采样数据和生成采样数据，其中，所述真实采样数据和所述生成采样数据的数据量比例为预设的目标比例；

[0030] 样本生成模块，用于根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集。

[0031] 为实现上述目的，本发明还提供一种训练样本优化设备，所述训练样本优化设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的训练样本优化程序，所述训练样本优化程序被所述处理器执行时实现如上所述的训练样本优化方法的步骤。

[0032] 此外，为实现上述目的，本发明还提出一种计算机可读存储介质，所述计算机可读

存储介质上存储有训练样本优化程序，所述训练样本优化程序被处理器执行时实现如上所述的训练样本优化方法的步骤。

[0033] 本发明中，通过获取用于进行模型训练的真实数据集和生成数据集，分别从真实数据集和生成数据集中进行有放回采样，得到数据量比例为预设的目标比例的真实采样数据和生成采样数据，根据真实采样数据和生成采样数据得到训练样本集，以便后续采样训练样本集进行模型训练。由于采样方式是有放回采样，能够实现最后采样得到的真实采样数据和生成采样数据的比例符合模型训练所需要的比例，并且还能够使得训练样本集达到充分训练模型所需要的训练数据量，从而能够兼顾数据量比例和训练数据量，避免因比例失衡导致的模型训练效果差，也避免因训练数据量不足而导致模型训练效果差，也即，通过对训练样本进行优化，提高了模型的训练效果。

附图说明

[0034] 图1为本发明实施例方案涉及的硬件运行环境的结构示意图；

[0035] 图2为本发明训练样本优化方法第一实施例的流程示意图；

[0036] 图3为本发明实施例涉及的一种生成数据与真实数据的比例实验数据图；

[0037] 图4为本发明训练样本优化装置较佳实施例的功能示意图模块图。

[0038] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0039] 应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0040] 如图1所示，图1是本发明实施例方案涉及的硬件运行环境的设备结构示意图。

[0041] 需要说明的是，本发明实施例训练样本优化设备可以是智能手机、个人计算机和服务器等设备，在此不做具体限制。

[0042] 如图1所示，该训练样本优化设备可以包括：处理器1001，例如CPU，网络接口1004，用户接口1003，存储器1005，通信总线1002。其中，通信总线1002用于实现这些组件之间的连接通信。用户接口1003可以包括显示屏(Display)、输入单元比如键盘(Keyboard)，可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口(如WI-FI接口)。存储器1005可以是高速RAM存储器，也可以是稳定的存储器(non-volatile memory)，例如磁盘存储器。存储器1005可选的还可以是独立于前述处理器1001的存储装置。

[0043] 本领域技术人员可以理解，图1中示出的设备结构并不构成对训练样本优化设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

[0044] 如图1所示，作为一种计算机存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及训练样本优化程序。其中，操作系统是管理和控制设备硬件和软件资源的程序，支持训练样本优化程序以及其它软件或程序的运行。

[0045] 在图1所示的设备中，用户接口1003主要用于与客户端进行数据通信；网络接口1004主要用于与服务器建立通信连接；而处理器1001可以用于调用存储器1005中存储的训练样本优化程序，并执行以下操作：

[0046] 获取用于模型训练的真实数据集和生成数据集；

[0047] 分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到真实采样数据和生成采样数据，其中，所述真实采样数据和所述生成采样数据的数据量比例为预设的目标比例；

[0048] 根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集。

[0049] 进一步地，所述分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到真实采样数据和生成采样数据的步骤包括：

[0050] 进行预设次数的采样，其中，各次采样分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到各次采样的真实采样数据和生成采样数据，各次采样时从所述真实数据集和所述生成数据集中采样的数据量比例为所述目标比例；

[0051] 所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集的步骤包括：

[0052] 分别将各次采样的真实采样数据和生成采样数据进行混合，得到各次采样对应的训练批，根据各所述训练批得到用于模型训练的训练样本集。

[0053] 进一步地，所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集的步骤之后，处理器1001还可以用于调用存储器1005中存储的训练样本优化程序，执行以下操作：

[0054] 采用所述训练样本集对待训练模型进行训练得到目标模型。

[0055] 进一步地，所述采用所述训练样本集对待训练模型进行训练得到目标模型的步骤之后，处理器1001还可以用于调用存储器1005中存储的训练样本优化程序，执行以下操作：

[0056] 将获取到的待识别文本图片输入所述目标模型，得到所述待识别文本图片的文本识别结果。

[0057] 进一步地，获取用于模型训练的所述生成数据集的步骤包括：

[0058] 获取背景图片库、文本字体库和文本内容库；

[0059] 从所述背景图片库中获取目标背景图片，从所述文本字体库中获取目标文本字体，以及从所述文本内容库中获取目标文本内容；

[0060] 将所述目标文本内容按照所述目标文本字体拼接至所述目标背景图片中，得到生成文本图片；

[0061] 根据各生成文本图片得到所述生成数据集。

[0062] 进一步地，所述根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集的步骤包括：

[0063] 对所述真实采样数据进行数据增强操作，并根据所述生成采样数据和数据增强后的真实采样数据得到用于模型训练的训练样本集；或，

[0064] 对所述生成采样数据进行数据增强操作，并根据所述真实采样数据和数据增强后的生成采样数据得到用于模型训练的训练样本集；或，

[0065] 对所述真实采样数据和所述生成采样数据进行数据增强操作，并根据数据增强后的真实采样数据和生成采样数据得到用于模型训练的训练样本集。进一步地，所述数据增强操作包括拉伸操作、加噪声操作、模糊操作和锐化操作中的一种或多种。

[0066] 基于上述的结构，提出训练样本优化方法的各个实施例。

[0067] 参照图2，图2为本发明训练样本优化方法第一实施例的流程示意图。

[0068] 本发明实施例提供了训练样本优化方法的实施例，需要说明的是，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。本发明训练样本优化方法各个实施例的执行主体可以是智能手机、个人计算机和服务器等设备，为便于描述，以下各实施例中省略执行主体进行阐述。在本实施例中，训练样本优化方法包括：

[0069] 步骤S10，获取用于模型训练的真实数据集和生成数据集；

[0070] 在本实施例中，当需要训练一个模型时，可以预先采集用于训练该模型的训练数据，根据模型所要实现的任务不同，训练数据不同，例如，需要训练一个用户购买意向的预测模型，采集的训练数据可能包括一些用户的购买历史数据。采集到的训练数据一般是真实的数据，例如是真实的用户进行购买所产生的数据。将采集的多条真实数据构建一个真实数据集。当需要进行训练样本的构建时，获取该真实数据集。

[0071] 由于在很多场景，真实数据都是很难获得的，因此导致真实数据集中的数据量较少，难以满足充分训练模型对训练数据量的需求。为了保证有更足够的数据量，可以采用仿真等方式，模拟真实数据来生成一批数据(以下将生成的数据称为生成数据)，可以按照训练数据量的需求，批量生成多条生成数据，构成生成数据集。需要说明的是，训练数据不同，生成数据的生成方式也不同。例如，当训练数据是用户的历史购买数据时，生成数据的生成方式可以是：预先设置各个数据特征的取值范围，然后从每个特征的取值范围中各随机选取一个值，将从各个特征中选取值进行组合，得到一条生成数据，为得到多条生成数据，则可以进行多次的随机选取和组合。

[0072] 步骤S20，分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到真实采样数据和生成采样数据，其中，所述真实采样数据和所述生成采样数据的数据量比例为预设的目标比例；

[0073] 在获取到真实数据集和生成数据集后，可以分别从两个数据集中进行有放回地采样。具体地，从真实数据集中进行有放回采样，得到真实采样数据，从生成数据集中进行有放回采样，得到生成采样数据。其中，有放回采样是指将采样得到的数据放回数据集，也即数据集的数据是不变的，每次从中采样的数据相当于是复制了一份出来。采样可以采用随机采样的方式。

[0074] 预先可以设置一个目标比例，例如设置为真实采样数据和生成采样数据的比例为1:1；由于根据具体的模型训练情况不同，真实数据和生成数据的比例对模型训练效果的影响会不同，因此，根据实际情况不同目标比例可设置得不同。通过对真实数据集和生成数据集得分别采样，得到的真实采样数据和生成采样数据的数据量比例是该目标比例。为使得真实采样数据和生成采样数据的数据量比例为目标比例，具体的采样过程有多种，也即，从真实数据集和生成数据集中进行有放回采样的次数，以及每次采样的数量都可以是不限制的，只要最终采样得到的真实采样数据和生成采样数据的数据量比例达到目标比例即可。需要说明的是，可以预先根据模型充分训练对数据量得需求，设置一个训练数据量，那么经过采样得到得真实采样数据和生成采样数据得数据量总和应当是达到该训练数据量；例如，可能需要10万条数据才能够使得模型得到从分训练，那么采样得到得真实采样数据和生成采样数据得数据量总和可以大于等于10万。

[0075] 以下举一个具体例子说明一种可行的采样过程：真实数据集的数据量为1000条，

生成数据集的数据量为9000条，根据具体的情况设置目标比例是1:1，训练数据量需要达到1万条才能够对模型进行充分训练；若直接采用当前的真实数据集和生成数据集作为训练样本集，虽然能够达到1万条的训练数据量，但是真实数据集和生成数据集的数据量比例是1:9，生成数据集比真实数据集多太多，比例失衡将导致模型训练效果不佳；在本实施例中，可以对真实数据集进行100次的有放回采样，每次采样数为50条，那么100次采样得到5000条真实采样数据，由于是有放回采样，所以每次采样的基础都是1000条数据，从而能够实现采样100次每次采样50条数据，由于是采用采样的方式，可以保证数据的随机性，从而保证采用真实采样数据进行模型训练的训练效果；对生成数据集则可以进行1000次的有放回采样，每次采样数为5条，那么1000次采样得到5000条生成采样数据；真实采样数据和生成采样数据各5000条，正好达到1:1的比例，以及1万条的训练数据量，从而实现了既保证训练数据量，又保证数据量比例，从而提高了模型的训练效果。

[0076] 步骤S30，根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集。

[0077] 可以根据真实采样数据和生成采样数据得到用于模型训练的训练样本集。具体地，可以将真实采样数据和生成采样数据中进行混合，整个作为一个训练样本集，采用这一个训练样本集来对模型进行训练；也可以是采用生成采样数据作为一个训练样本集，采用真实采样数据作为一个训练样本集，在训练的时候，可先采用生成采样数据来对模型进行训练，再采用真实采样数据对模型进行微调。

[0078] 进一步地，所述步骤S20包括：

[0079] 步骤S201，进行预设次数的采样，其中，各次采样分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到各次采样的真实采样数据和生成采样数据，各次采样时从所述真实数据集和所述生成数据集中采样的数据量比例为所述目标比例；

[0080] 进一步地，在本实施例中，采样的过程还可以是进行预设次数的采样，每次采样分别从真实数据集和生成数据集中进行一次有放回采样。也即，预先设置一个采样次数，从真实数据集中进行预设次数的采样，从生成数据集中也进行相同次数的采样。可以理解为，每次采样都是分别从真实数据集和生成数据集中各采样一次，并且从真实数据集和生成数据集中采样的数据量比例为上述的目标比例。需要说明的是，每次从真实数据集采样的数据量可以是相同的，也可以是不同的，生成数据集也同理。每次采样的数据量可以根据具体情况设置，例如，当模型的训练过程适合采用一批一批的数据进行训练时，可以设置一个合适的批数据量，每次采样的真实采样数据和生成采样数据的数据量总和为该批数据量。

[0081] 例如，真实数据集的数据量为1000条，生成数据集的数据量为9000条，根据具体的情况设置目标比例是1:1，训练数据量需要达到1万条才能够对模型进行充分训练；可以进行100次的采样，每次采样是对真实数据集进行一次有放回采样，采样数为50条，对生成数据集也进行一次有放回采样，采样数也为50条；每次采样得到的总数据量为100条，真实采样数据和生成采样数据额比例正好为1:1。

[0082] 所述步骤S30包括：

[0083] 步骤S301，分别将各次采样的真实采样数据和生成采样数据进行混合，得到各次采样对应的训练批，根据各所述训练批得到用于模型训练的训练样本集。

[0084] 分别将各次采样的真实采样数据和生成采样数据进行混合，得到各次采样对应的

训练批。也就是，将每次采样得到的真实采样数据和生成采样数据进行混合，例如，每次采样得到50条真实采样数据和50条生成采样数据，将两种采样数据进行混合，得到100条采样数据。将混合后的采样数据作为一个训练批，那么最后得到的训练批的数量是上述的预设次数，也就是预设采样100次，最后就得到100个训练批。将这些训练批作为训练样本集。

[0085] 在模型训练阶段，一次训练过程可以采用一个训练批，依次采用各个训练批进行训练。需要说明的是，可以是采样得到各个训练批之后再进行训练；也可以是采样得到一个训练批后训练一次，再采样得到一个训练批后再训练一次，这样可以使得当模型的效果较好时，例如模型的预测准确率满足要求时，无论当前是否达到了足够的训练数据量都可以不再进行采样，从而减少采样时间，提高模型训练效率。

[0086] 在本实施例中，通过获取用于进行模型训练的真实数据集和生成数据集，分别从真实数据集和生成数据集中进行有放回采样，得到数据量比例为预设的目标比例的真实采样数据和生成采样数据，根据真实采样数据和生成采样数据得到训练样本集，以便后续采样训练样本集进行模型训练。由于采样方式是有放回采样，能够实现最后采样得到的真实采样数据和生成采样数据的比例符合模型训练所需要的比例，并且还能够使得训练样本集达到充分训练模型所需要的训练数据量，从而能够兼顾数据量比例和训练数据量，避免因比例失衡导致的模型训练效果差，也避免因训练数据量不足而导致模型训练效果差，也即，通过对训练样本进行优化，提高了模型的训练效果。

[0087] 进一步地，在本实施例中，目标比例可以是设置为真实采样数据与生成采样数据的数据量比例为3:7。如图3所示，通过实验数据验证，在比例为7:3时，使得模型训练效果更加。

[0088] 进一步地，基于上述第一实施例，提出本发明训练样本优化方法第二实施例，在本实施例中，所述训练样本优化方法还包括：

[0089] 步骤S40，采用所述训练样本集对待训练模型进行训练得到目标模型。

[0090] 进一步地，预先可以设置一个待训练的模型，待训练模型的结构可以采用神经网络模型、线性回归模型或逻辑回归模型等机器学习模型，根据具体的模型训练任务不同，待训练模型的结构可以不同，在本实施例中不做限定。在获取到训练样本集后，可采用训练样本集对待训练模型进行训练，将训练后的待训练模型作为可以进行使用的目标模型。采用训练样本集对待训练模型进行训练的过程，可以采用常用的机器学习模型训练方式，例如，有监督学习方式或无监督学习方式等，在此不进行详细赘述。

[0091] 进一步地，所述训练样本优化方法还包括：

[0092] 步骤S50，将获取到的待识别文本图片输入所述目标模型，得到所述待识别文本图片的文本识别结果。

[0093] 进一步地，在本实施例中，模型训练任务可以是训练一个识别图片中文本的模型，那么真实数据集和生成数据集中的数据可以是包含文本的文本图片，那么训练样本集中的各条样本是文本图片。待训练模型可以是采用常用的文本识别模型，在此不做详细赘述。通过训练样本集对待训练模型进行训练后得到能够识别图片中文本的目标模型。

[0094] 可以获取待识别文本图片，将待识别文本图片输入目标模型，得到待识别文本图片的文本识别结果。具体地，可将待识别文本图片转换为目标模型的输入数据的格式，将转换后的待识别文本图片输入目标模型，经过目标模型的处理，得到待识别文本图片的文本

识别结果，文本识别结果可以是待识别文本图片中所存在的各个文本字符。

[0095] 在本实施例中，由于采用第一实施例中的训练样本优化方法得到训练样本集，采用训练样本集来训练得到识别图片中文本的目标模型，使得目标模型得到充分的训练，并且不会因为生成数据的数据量远大于真实数据而影响模型的训练效果，从而使得训练得到的目标模型的识别准确率提高，进而能够对待识别文本图片中的文本做出更准确的识别，得到更准确的文本识别结果。

[0096] 进一步地，获取用于模型训练的所述生成数据集的步骤包括：

[0097] 步骤a，获取背景图片库、文本字体库和文本内容库；

[0098] 进一步地，在本实施例中，当模型训练任务是训练一个识别图片中文本的模型时，生成数据集的获取方式可以是采用如下方式。

[0099] 具体地，由于文本识别一般会有一个特殊的识别对象，特殊识别对象的图片背景、字体和文字内容可能比较特殊，例如对包含身份证件的图片进行文本识别的场景，身份证的版面、字体和文字内容是比较特殊的。这种情况下，难以获得足够充分训练模型的真实数据，因此，可以生成模拟身份证件图片的图片。具体地，可以获取特殊识别对象多种形式的背景图片(也即仅包含图片，不包含文本)，构建一个背景图片库，例如，可以采集打印的身份证件图片，将图片中的文字进行去除，得到背景图片，还可以采集相机拍摄的身份证件照片，将照片中的文字进行取畜，得到背景图片。采集不同的文本字体构建一个文本字体库。根据特殊识别对象中常见的识别内容，构建一个文本内容库，包含多种文本内容。

[0100] 步骤b，从所述背景图片库中获取目标背景图片，从所述文本字体库中获取目标文本字体，以及从所述文本内容库中获取目标文本内容；

[0101] 从背景图片库中获取目标背景图片，从文本字体库中获取目标文本字体，从文本内容库中获取目标文本内容。具体地，可以是从背景图片库中随机进行选取一个背景图片作为目标背景图片，也可以是按顺序将其中的背景图片作为目标背景图片。目标文本字体和目标文本内容的获取方式可以采用类似于目标背景图片的获取方式。

[0102] 步骤c，将所述目标文本内容按照所述目标文本字体拼接至所述目标背景图片中，得到生成文本图片；

[0103] 将目标文本内容按照模板文本字体拼接至模板背景图片中，得到生成文本图片，具体的拼接可采用一些常用的图片拼接工具，例如可采用OpenCV库(Open Source Computer Vision Library，开源计算机视觉库)。

[0104] 步骤d，根据各生成文本图片得到所述生成数据集。

[0105] 采用上述方式生成多个生成文本图片，将这些生成文本图片作为生成数据集。

[0106] 进一步地，所述步骤S30包括：

[0107] 步骤S302，对所述真实采样数据进行数据增强操作，并根据所述生成采样数据和数据增强后的真实采样数据得到用于模型训练的训练样本集；或，

[0108] 步骤S303，对所述生成采样数据进行数据增强操作，并根据所述真实采样数据和数据增强后的生成采样数据得到用于模型训练的训练样本集；或，

[0109] 步骤S304，对所述真实采样数据和所述生成采样数据进行数据增强操作，并根据数据增强后的真实采样数据和生成采样数据得到用于模型训练的训练样本集。

[0110] 进一步地，在本实施例中，可以对真实采样数据进行数据增强操作，或者是对生成

采样数据进行数据增强操作，或者是对真实采样数据和生成采样数据均进行数据增强操作。数据增强操作有多种，根据模型训练任务不同，训练数据不同，数据增强操作也可以不同，例如，当训练数据是图片是，数据增强操作可以是对图片进行翻转、缩放等。

[0111] 数据增强操作具体是对每条采样数据进行数据增强，将一个条采样数据扩增为多条。

[0112] 采用生成采样数据和数据增强后的真实采样数据得到训练样本集，或者是采用真实采样数据和数据增强后的生成采样数据得到训练样本集，或者是采用数据增强后的真实采样数据和数据增强后的生成采样数据得到训练样本集。

[0113] 进一步地，所述数据增强操作包括拉伸操作、加噪声操作、模糊操作和锐化操作中的一种或多种。

[0114] 在本实施例中，当模型训练任务是训练一个识别图片中文本的模型时，训练数据可以是文本图片，那么可对真实采样数据和/或生成采样数据对应的文本图片进行数据增强操作，数据增强操作可以是包括拉伸操作、加噪声操作、模糊操作和锐化操作中的一种或多种。其中，加噪声操作可以是在图片中加上一些污点。这些数据增强操作可以使得采样数据的形式更加多样，更加难以识别，从而通过数据增强后的数据对模型进行训练，使得模型能够具备更强的文本识别能力。

[0115] 在本实施例中，通过对训练数据进行数据增强，又进一步地扩增了训练数据量，从而使得模型得到更充分的训练，进而增加了模型训练效果。

[0116] 此外，此外本发明实施例还提出一种训练样本优化装置，参照图4，所述训练样本优化装置包括：

[0117] 获取模块10，用于获取用于模型训练的真实数据集和生成数据集；

[0118] 采样模块20，用于分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到真实采样数据和生成采样数据，其中，所述真实采样数据和所述生成采样数据的数据量比例为预设的目标比例；

[0119] 样本生成模块30，用于根据所述真实采样数据和所述生成采样数据得到用于模型训练的训练样本集。

[0120] 进一步地，所述采样模块20包括：

[0121] 采样单元，用于进行预设次数的采样，其中，各次采样分别从所述真实数据集和所述生成数据集中进行有放回采样，对应得到各次采样的真实采样数据和生成采样数据，各次采样时从所述真实数据集和所述生成数据集中采样的数据量比例为所述目标比例；

[0122] 所述样本生成模块30包括：

[0123] 混合单元，用于分别将各次采样的真实采样数据和生成采样数据进行混合，得到各次采样对应的训练批，根据各所述训练批得到用于模型训练的训练样本集。

[0124] 进一步地，所述训练样本优化装置还包括：

[0125] 训练模块，用于采用所述训练样本集对待训练模型进行训练得到目标模型。

[0126] 进一步地，所述训练样本优化装置还包括：

[0127] 文本识别模块，用于将获取到的待识别文本图片输入所述目标模型，得到所述待识别文本图片的文本识别结果。

[0128] 进一步地，所述获取模块10包括：

[0129] 第一获取单元，用于获取背景图片库、文本字体库和文本内容库；

[0130] 第二获取单元，用于从所述背景图片库中获取目标背景图片，从所述文本字体库中获取目标文本字体，以及从所述文本内容库中获取目标文本内容；

[0131] 拼接单元，用于将所述目标文本内容按照所述目标文本字体拼接至所述目标背景图片中，得到生成文本图片；

[0132] 数据集生成单元，用于根据各生成文本图片得到所述生成数据集。

[0133] 进一步地，所述样本生成模块30包括：

[0134] 数据增强单元，对所述真实采样数据进行数据增强操作，并根据所述生成采样数据和数据增强后的真实采样数据得到用于模型训练的训练样本集；或，

[0135] 对所述生成采样数据进行数据增强操作，并根据所述真实采样数据和数据增强后的生成采样数据得到用于模型训练的训练样本集；或，

[0136] 对所述真实采样数据和所述生成采样数据进行数据增强操作，并根据数据增强后的真实采样数据和生成采样数据得到用于模型训练的训练样本集。进一步地，所述数据增强操作包括拉伸操作、加噪声操作、模糊操作和锐化操作中的一种或多种。

[0137] 本发明训练样本优化装置的具体实施方式的拓展内容与上述训练样本优化方法各实施例基本相同，在此不做赘述。

[0138] 此外，本发明实施例还提出一种计算机可读存储介质，所述存储介质上存储有训练样本优化程序，所述训练样本优化程序被处理器执行时实现如下所述的训练样本优化方法的步骤。

[0139] 本发明训练样本优化设备和计算机可读存储介质的各实施例，均可参照本发明训练样本优化方法各个实施例，此处不再赘述。

[0140] 需要说明的是，在本文中，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、物品或者装置不仅包括那些要素，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、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[0141] 上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

[0142] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本发明各个实施例所述的方法。

[0143] 以上仅为本发明的优选实施例，并非因此限制本发明的专利范围，凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

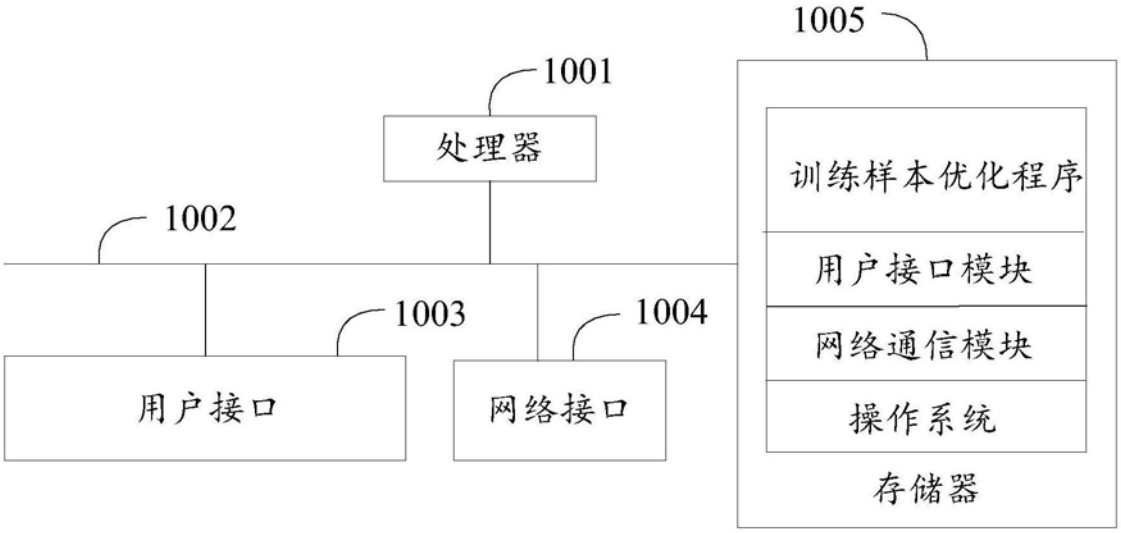

图1

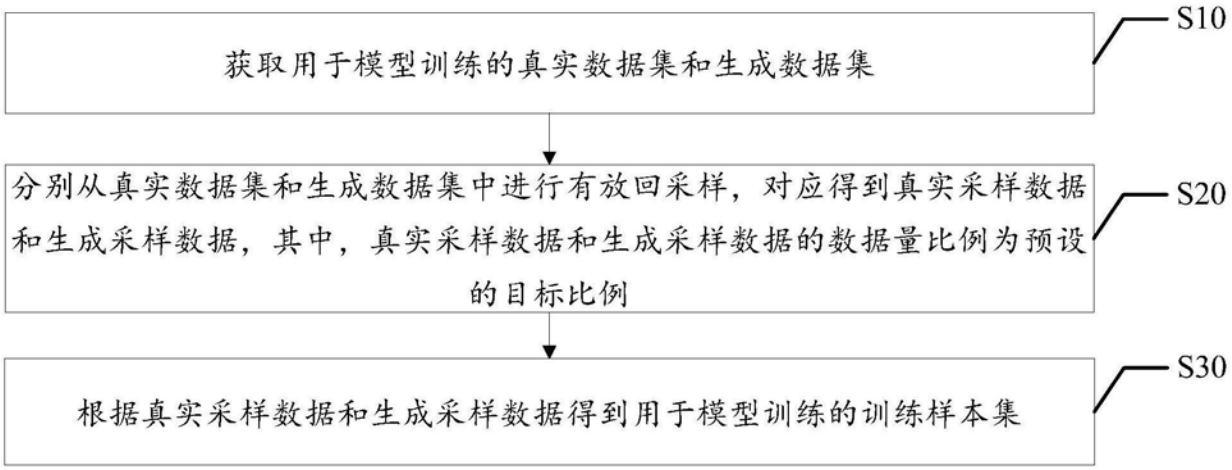

图2

生成:真实	1:9	2:8	3:7	4:6	5:5	6:4	7:3	8:2	9:1
准确率	93.099	93.261	93.419	93.218	93.074	92.936	92.770	92.544	92.378

图3

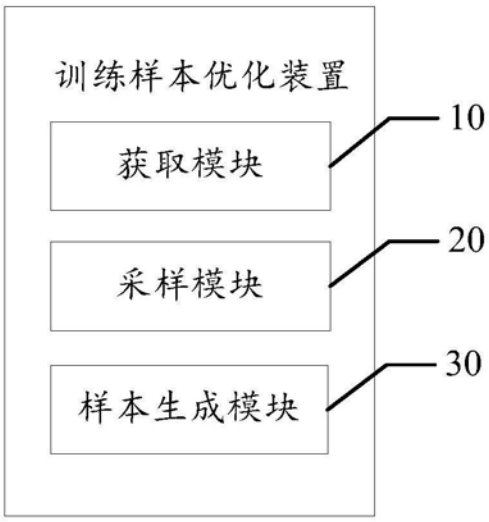

图4