(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 111401445 A
(43)申请公布日 2020.07.10

(21)申请号 202010182180.5

(22)申请日 2020.03.16

(71)申请人 腾讯科技(深圳)有限公司
地址 518064 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 李一鸣 吴保元 张勇 樊艳波 李志锋 刘威 冯岩 江勇 夏树涛

(74)专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285
代理人 王仲凯

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

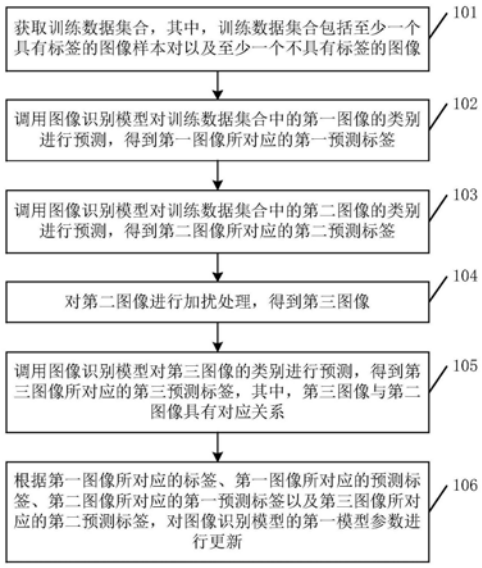
权利要求书3页 说明书28页 附图7页

(54)发明名称

一种图像识别模型的训练方法、图像识别的方法及装置

(57)摘要

本申请公开了一种图像识别模型的训练方法，该方法用于人工智能领域，该方法包括：获取训练数据集合；调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一预测标签；调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二预测标签；对第二图像进行加扰处理得到第三图像；调用图像识别模型对第三图像的类别进行预测，得到第三预测标签；根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像识别模型的第一模型参数进行更新。本申请公开了一种图像识别的方法和装置。本申请通过半监督学习提升模型的识别能力，增强模型的鲁棒性。



CN 111401445 A

1.一种图像识别模型的训练方法,其特征在于,包括:

获取训练数据集合,其中,所述训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像;

调用图像识别模型对所述训练数据集合中的第一图像的类别进行预测,得到所述第一图像所对应的第一预测标签;

调用所述图像识别模型对所述训练数据集合中的第二图像的类别进行预测,得到所述第二图像所对应的第二预测标签;

对所述第二图像进行加扰处理,得到第三图像;

调用所述图像识别模型对所述第三图像的类别进行预测,得到所述第三图像所对应的第三预测标签;

根据所述第一图像所对应的标签、所述第一图像所对应的第一预测标签、所述第二图像所对应的第二预测标签以及所述第三图像所对应的第三预测标签,对所述图像识别模型的第一模型参数进行更新。

2.根据权利要求1所述的方法,其特征在于,所述根据所述第一图像所对应的标签、所述第一图像所对应的第一预测标签、所述第二图像所对应的第二预测标签以及所述第三图像所对应的第三预测标签,对所述图像识别模型的第一模型参数进行更新,包括:

根据所述第一图像所对应的标签以及所述第一图像所对应的第一预测标签,确定第一风险函数,其中,所述第一风险函数用于表示预测标签与标签之间的差异;

根据所述第二图像所对应的第二预测标签以及所述第三图像所对应的第三预测标签,确定第二风险函数,其中,所述第二风险函数用于表示已加扰图像与未加扰图像之间的差异;

根据所述第一风险函数以及第二风险函数,生成目标优化函数;

当所述目标优化函数达到最小值时,获取第二模型参数;

将所述图像识别模型的所述第一模型参数更新为所述第二模型参数。

3.根据权利要求2所述的方法,其特征在于,所述根据所述第一风险函数以及第二风险函数,生成目标优化函数,包括:

采用第一损失函数对所述第一风险函数进行变换处理,得到第一优化函数,其中,所述第一优化函数包括图像的预测分布向量与标签之间的损失值;

采用第二损失函数对所述第二风险函数进行变换处理,得到第二优化函数,其中,所述第二优化函数包括所述已加扰图像的预测分布向量与所述未加扰图像的预测标签之间的损失值;

根据所述第一优化函数与所述第二优化函数,生成所述目标优化函数。

4.根据权利要求1至3中任一项所述的方法,其特征在于,所述对所述第二图像进行加扰处理,得到第三图像,包括:

获取图像加扰类型;

根据所述图像加扰类型确定扰动邻域,其中,所述扰动邻域表示对未加扰图像进行图像变换的范围;

基于所述扰动邻域以及所述第二图像所对应的第二预测标签,确定第三优化函数;

当第三优化函数达到最大值时,获取所述第二图像所对应的所述第三图像。

5.根据权利要求4所述的方法,其特征在于,所述获取图像加扰类型,包括:

获取图像攻击类型的数量;

若所述图像攻击类型的数量等于1,则确定所述图像加扰类型为单攻击类型;

若所述图像攻击类型的数量大于1,则确定所述图像加扰类型为复合攻击类型。

6.根据权利要求5所述的方法,其特征在于,所述根据所述图像加扰类型确定扰动邻域,包括:

若所述图像加扰类型为单攻击类型,则确定图像攻击类型;

若所述图像攻击类型为像素攻击类型,则获取像素变换函数对应的像素距离度量,其中,所述像素攻击类型为对所述未加扰图像中的至少一个像素值进行变换;

根据所述像素距离度量以及最大像素变换范围,确定所述扰动邻域。

7.根据权利要求5所述的方法,其特征在于,所述根据所述图像加扰类型确定扰动邻域,包括:

若所述图像加扰类型为单攻击类型,则确定图像攻击类型;

若所述图像攻击类型为几何攻击类型,则获取几何变换函数所对应的几何距离度量,其中,所述几何攻击类型为对所述未加扰图像进行平移以及旋转中的至少一种变换;

根据所述几何距离度量以及最大几何变换范围,确定所述扰动邻域。

8.根据权利要求5所述的方法,其特征在于,所述根据所述图像加扰类型确定扰动邻域,包括:

若所述图像加扰类型为复合攻击类型,则获取至少两个图像攻击类型;

若所述至少两个图像攻击类型包括像素攻击类型以及几何攻击类型,则获取图像攻击顺序;

根据所述图像攻击顺序确定所述扰动邻域。

9.根据权利要求8所述的方法,其特征在于,所述根据所述图像攻击顺序确定所述扰动邻域,包括:

若所述图像攻击顺序为先采用像素攻击类型,再采用几何攻击类型,则获取第一复合变换函数所对应的第一复合距离度量,其中,所述像素攻击类型为对所述未加扰图像中的至少一个像素值进行变换,所述几何攻击类型为对所述未加扰图像进行平移以及旋转中的至少一种变换;

根据所述第一复合距离度量以及最大几何变换范围,确定所述扰动邻域;

或,所述根据所述图像攻击顺序确定所述扰动邻域,包括:

若所述图像攻击顺序为先采用几何攻击类型,再采用像素攻击类型,则获取第二复合变换函数所对应的第二复合距离度量;

根据所述第二复合距离度量以及最大像素变换范围,确定所述扰动邻域。

10.一种图像识别的方法,其特征在于,包括:

获取待识别图像;

调用图像识别模型对所述待识别图像的类别进行预测,得到图像类别结果,其中,所述图像识别模型为权利要求1至9中任一项所述的图像识别模型;

向客户端发送所述图像类别结果,以使所述客户端展示所述图像类别结果。

11.一种图像识别模型训练装置,其特征在于,包括:

获取模块，用于获取训练数据集合，其中，所述训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像；

预测模块，用于调用图像识别模型对所述训练数据集合中的第一图像的类别进行预测，得到所述第一图像所对应的第一预测标签；

所述预测模块，还用于调用所述图像识别模型对所述训练数据集合中的第二图像的类别进行预测，得到所述第二图像所对应的第二预测标签；

处理模块，用于对所述第二图像进行加扰处理，得到第三图像；

所述预测模块，还用于调用所述图像识别模型对所述第三图像的类别进行预测，得到所述第三图像所对应的第三预测标签；

更新模块，用于根据所述第一图像所对应的标签、所述第一图像所对应的第一预测标签、所述第二图像所对应的第二预测标签以及所述第三图像所对应的第三预测标签，对所述图像识别模型的第一模型参数进行更新。

12.一种图像识别装置，其特征在于，包括：

获取模块，用于获取待识别图像；

调用模块，用于调用图像识别模型对所述待识别图像的类别进行预测，得到图像类别结果，其中，所述图像识别模型为权利要求1至9中任一项所述的图像识别模型；

发送模块，用于向客户端发送所述图像类别结果，以使所述客户端展示所述图像类别结果。

13.一种服务器，其特征在于，包括：存储器、收发器、处理器以及总线系统；

其中，所述存储器用于存储程序；

所述处理器用于执行所述存储器中的程序，包括执行如上述权利要求1至9中任一项所述的方法，或，执行如上述权利要求10所述的方法；

所述总线系统用于连接所述存储器以及所述处理器，以使所述存储器以及所述处理器进行通信。

14.一种终端设备，其特征在于，包括：存储器、收发器、处理器以及总线系统；

其中，所述存储器用于存储程序；

所述处理器用于执行所述存储器中的程序，包括执行如上述权利要求1至9中任一项所述的方法，或，执行如上述权利要求10所述的方法；

所述总线系统用于连接所述存储器以及所述处理器，以使所述存储器以及所述处理器进行通信。

15.一种计算机可读存储介质，包括指令，当其在计算机上运行时，使得计算机执行如权利要求1至9中任一项所述的方法，或，执行如权利要求10所述的方法。

一种图像识别模型的训练方法、图像识别的方法及装置

[0001] 技术邻域

[0002] 本申请涉及人工智能领域，尤其涉及一种图像识别模型的训练方法、图像识别的方法及装置。

背景技术

[0003] 深度学习占据着机器视觉邻域的重要地位，在机器视觉邻域中，深度学习成为了从自动驾驶、监控以及安保等方面的主力。尽管深度网络已经在处理复杂问题时所取得了现象级的成功，但是对于带有轻微扰动的图像而言，仍然容易出现识别错误的情况。

[0004] 为了能够抵御这类图像对模型识别的干扰，目前，提出了一种对抗防御方法，即不断输入新类型的样本并执行对抗训练，不断提升网络的鲁棒性。为了保证训练的有效性，该方法需要大量带有标签的训练数据，通过对抗训练可以提升模型的鲁棒性。

[0005] 然而，由于对图像进行干扰的方式较多，即使在训练的过程中增加大量的样本，仍然难以覆盖所有的样本，总是会存在新的攻击样本来欺骗网络，难以提升模型的防御性能，对图像的识别精度不高。

发明内容

[0006] 本申请实施例提供了一种图像识别模型的训练方法、图像识别的方法及装置，利用了有标签数据和无标签数据对模型进行半监督训练，对于无标签的图像而言对其进行加扰，使得模型在训练过程中对是否加扰过的图像进行识别，从而提升模型的识别能力，增强模型的鲁棒性。

[0007] 有鉴于此，本申请第一方面提供一种图像识别模型的训练方法，包括：

[0008] 获取训练数据集合，其中，训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像；

[0009] 调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一图像所对应的第一预测标签；

[0010] 调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二图像所对应的第二预测标签；

[0011] 对第二图像进行加扰处理，得到第三图像；

[0012] 调用图像识别模型对第三图像的类别进行预测，得到第三图像所对应的第三预测标签，其中，第三图像与第二图像具有对应关系；

[0013] 根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像识别模型的第一模型参数进行更新。

[0014] 本申请第二方面提供一种图像识别的方法，包括：

[0015] 获取待识别图像；

[0016] 调用图像识别模型对待识别图像的类别进行预测，得到图像类别结果，其中，图像

识别模型为第一方面所描述的图像识别模型；

[0017] 向客户端发送图像类别结果，以使客户端展示图像类别结果。

[0018] 本申请第三方面提供一种图像识别模型训练装置，包括：

[0019] 获取模块，用于获取训练数据集合，其中，训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像；

[0020] 预测模块，用于调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一图像所对应的第一预测标签；

[0021] 预测模块，还用于调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二图像所对应的第二预测标签；

[0022] 处理模块，用于对第二图像进行加扰处理，得到第三图像；

[0023] 预测模块，还用于调用图像识别模型对第三图像的类别进行预测，得到第三图像所对应的第三预测标签，其中，第三图像与第二图像具有对应关系；

[0024] 更新模块，用于根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像识别模型的第一模型参数进行更新。

[0025] 在一种可能的设计中，在本申请实施例的第三方面的第一种实现方式中，

[0026] 更新模块，具体用于根据第一图像所对应的标签以及第一图像所对应的第一预测标签，确定第一风险函数，其中，第一风险函数用于表示预测标签与标签之间的差异；

[0027] 根据第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，确定第二风险函数，其中，第二风险函数用于表示已加扰图像与未加扰图像之间的差异；

[0028] 根据第一风险函数以及第二风险函数，生成目标优化函数；

[0029] 当目标优化函数达到最小值时，获取第二模型参数；

[0030] 将图像识别模型的第一模型参数更新为第二模型参数。

[0031] 在一种可能的设计中，在本申请实施例的第三方面的第二种实现方式中，

[0032] 更新模块，具体用于采用第一损失函数对第一风险函数进行变换处理，得到第一优化函数，其中，第一优化函数包括图像的预测分布向量与标签之间的损失值；

[0033] 采用第二损失函数对第二风险函数进行变换处理，得到第二优化函数，其中，第二优化函数包括已加扰图像的预测分布向量与未加扰图像的预测标签之间的损失值；

[0034] 根据第一优化函数与第二优化函数，生成目标优化函数。

[0035] 在一种可能的设计中，在本申请实施例的第三方面的第三种实现方式中，

[0036] 处理模块，具体用于获取图像加扰类型；

[0037] 根据图像加扰类型确定扰动邻域，其中，扰动邻域表示对未加扰图像进行图像变换的范围；

[0038] 基于扰动邻域以及第二图像所对应的第二预测标签，确定第三优化函数；

[0039] 当第三优化函数达到最大值时，获取第二图像所对应的第三图像。

[0040] 在一种可能的设计中，在本申请实施例的第三方面的第四种实现方式中，

[0041] 处理模块，具体用于获取图像攻击类型的数量；

[0042] 若图像攻击类型的数量等于1，则确定图像加扰类型为单攻击类型；

[0043] 若图像攻击类型的数量大于1，则确定图像加扰类型为复合攻击类型。

[0044] 在一种可能的设计中，在本申请实施例的第三方面的第五种实现方式中，

[0045] 处理模块，具体用于若图像加扰类型为单攻击类型，则确定图像攻击类型；

[0046] 若图像攻击类型为像素攻击类型，则获取像素变换函数对应的像素距离度量，其中，像素攻击类型为对未加扰图像中的至少一个像素值进行变换；

[0047] 根据像素距离度量以及最大像素变换范围，确定扰动邻域。

[0048] 在一种可能的设计中，在本申请实施例的第三方面的第六种实现方式中，

[0049] 处理模块，具体用于若图像加扰类型为单攻击类型，则确定图像攻击类型；

[0050] 若图像攻击类型为几何攻击类型，则获取几何变换函数所对应的几何距离度量，其中，几何攻击类型为对未加扰图像进行平移以及旋转中的至少一种变换；

[0051] 根据几何距离度量以及最大几何变换范围，确定扰动邻域。

[0052] 在一种可能的设计中，在本申请实施例的第三方面的第七种实现方式中，

[0053] 处理模块，具体用于若图像加扰类型为复合攻击类型，则获取至少两个图像攻击类型；

[0054] 若至少两个图像攻击类型包括像素攻击类型以及几何攻击类型，则获取图像攻击顺序；

[0055] 根据图像攻击顺序确定扰动邻域。

[0056] 在一种可能的设计中，在本申请实施例的第三方面的第八种实现方式中，

[0057] 处理模块，具体用于若图像攻击顺序为先采用像素攻击类型，再采用几何攻击类型，则获取第一复合变换函数所对应的第一复合距离度量，其中，像素攻击类型为对未加扰图像中的至少一个像素值进行变换，几何攻击类型为对未加扰图像进行平移以及旋转中的至少一种变换；

[0058] 根据第一复合距离度量以及最大几何变换范围，确定扰动邻域；

[0059] 或，

[0060] 处理模块，具体用于若图像攻击顺序为先采用几何攻击类型，再采用像素攻击类型，则获取第二复合变换函数所对应的第二复合距离度量；

[0061] 根据第二复合距离度量以及最大像素变换范围，确定扰动邻域。

[0062] 本申请第四方面提供一种图像识别装置，包括：

[0063] 获取模块，用于获取待识别图像；

[0064] 调用模块，用于调用图像识别模型对待识别图像的类别进行预测，得到图像类别结果，其中，图像识别模型为第一方面所描述的图像识别模型；

[0065] 发送模块，用于向客户端发送图像类别结果，以使客户端展示图像类别结果。

[0066] 本申请第五方面提供一种服务器，包括：存储器、收发器、处理器以及总线系统；

[0067] 其中，存储器用于存储程序；

[0068] 处理器用于执行存储器中的程序，包括执行上述各方面所述的方法；

[0069] 总线系统用于连接存储器以及处理器，以使存储器以及处理器进行通信。

[0070] 本申请第六方面提供一种终端设备，包括：存储器、收发器、处理器以及总线系统；

[0071] 其中，存储器用于存储程序；

[0072] 处理器用于执行存储器中的程序，包括执行上述各方面所述的方法；

[0073] 总线系统用于连接存储器以及处理器，以使存储器以及处理器进行通信。

[0074] 本申请的第七方面提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行上述各方面所述的方法。

[0075] 从以上技术方案可以看出，本申请实施例具有以下优点：

[0076] 本申请实施例中，提供了一种图像识别模型的训练方法，首先获取第一训练数据集合以及第二训练数据集合，然后调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一图像所对应的第一预测标签，调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二图像所对应的第二预测标签，再对第二图像进行加扰处理，得到第三图像，调用图像识别模型对第三图像的类别进行预测，最后根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像识别模型的第一模型参数进行更新。通过上述方式，利用了有标签数据和无标签数据对模型进行半监督训练，对于无标签的图像而言对其进行加扰，使得模型在训练过程中对是否加扰过的图像进行识别，从而提升模型的识别能力，增强模型的鲁棒性。

附图说明

[0077] 图1为本申请实施例中应用于图像分类场景的一个交互示意图；

[0078] 图2为本申请实施例中采用对抗性图像进行攻击的一个示意图；

[0079] 图3为本申请实施例中图像识别系统的一个环境示意图；

[0080] 图4为本申请实施例中图像识别模型的训练方法一个实施例示意图；

[0081] 图5为本申请实施例中图像识别的方法一个实施例示意图；

[0082] 图6A为实验中基于CIFAR-10数据集对复合攻击进行防御的一个效果对比示意图；

[0083] 图6B为实验中基于MNIST数据集对复合攻击进行防御的一个效果对比示意图；

[0084] 图7A为实验中基于CIFAR-10数据集对复合攻击进行防御的另一个效果对比示意图；

[0085] 图7B为实验中基于MNIST数据集对复合攻击进行防御的另一个效果对比示意图；

[0086] 图8A为实验中基于CIFAR-10数据集下SRT性能随无标签数据使用量变化的一个曲线；

[0087] 图8B为实验中基于MNIST数据集下SRT性能随无标签数据使用量变化的一个曲线；

[0088] 图9为本申请实施例中图像识别模型训练装置的一个实施例示意图；

[0089] 图10为本申请实施例中图像识别装置的一个实施例示意图；

[0090] 图11为本申请实施例中服务器的一个结构示意图；

[0091] 图12为本申请实施例中终端设备的一个结构示意图。

具体实施方式

[0092] 本申请实施例提供了一种图像识别模型的训练方法、图像识别的方法及装置，利用了有标签数据和无标签数据对模型进行半监督训练，对于无标签的图像而言对其进行加扰，使得模型在训练过程中对是否加扰过的图像进行识别，从而提升模型的识别能力，增强模型的鲁棒性。

[0093] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第

四”等(如果存在)是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本申请的实施例例如能够以除了在这里图示或描述的那些以外的顺序实施。此外,术语“包括”和“对应于”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0094] 应理解,本申请提供的图像识别方法可以应用于人脸识别场景,例如,在安防邻域中不法分子可能侵入数据库,对大量的用户图像进行恶意篡改,篡改的方式可能修改用户图像中的部分像素点,或者对用户图像进行平移等操作,导致这些用户图像难以被正确识别。而本申请提供的图像识别模型能够对被攻击的图像进行有效的识别,有效地提升了用户图像的识别精度。

[0095] 本申请提供的图像识别方法还可以应用于图像分类场景,例如,在图像网站中往往存储有大量的图像,这些图像可能因为数据传输不良,或者后台工作人员操作不当等原因,导致图像出现变形或者若干个像素点丢失等情况,而本申请提供的图像识别模型能够对这类图像进行有效的识别,从而进一步进行自动分类。为了便于理解,请参阅图1,图1为本申请实施例中应用于图像分类场景的一个交互示意图,如图所示,假设数据库中存储有百万张图像,图1所示的10张图像仅为一个示意,这里的10张图像可能存在被攻击的情况。图像类别包含但不仅限于人物类别、动物类别、风景类别、汽车类别以及动漫类别。假设通过图像识别模型输出图像1、图像3和图像8的图像识别结果为“人物”,则将图像1、图像3和图像8自动归类于人物类别。

[0096] 具体地,在实际应用中,对图像攻击后可以生成对抗性图像,为了便于说明,请参阅图2,图2为本申请实施例中采用对抗性图像进行攻击的一个示意图,如图所示,首先攻击者可以制造一张对抗性图像,这张对抗性图像可能被人类的肉眼所识别,比如,攻击者将原始彩色图像变更为黑白图像,这种情况下,人类肉眼即可识别图像类别结果。又比如,攻击者将原始彩色图像中的某个像素点由原本的像素值(256,128,225),更改为像素值(256,128,226),这种情况下,人类肉眼往往难以分辨图像的变换,因此,需要采用机器学习模型进行识别,例如采用本申请提供的图像识别模型输出对应的图像类别结果。

[0097] 需要说明的是,对图像进行攻击的方式有多种,常见的有像素攻击、几何攻击以及色彩变换攻击等,其中,像素攻击是以像素点作为单位,对至少一个像素点的像素值进行任意变换。几何攻击是对整个图像进行平移或者旋转等操作,使得图像发生变换。色彩变换攻击可以以像素点为单位,也可以以整个图像为单位进行色彩饱和度、亮度或灰度等变换。此外,还有其他不同类型的图像攻击方式,此处不一一列举。

[0098] 为了便于理解,本申请提出了一种图像识别的方法,该方法应用于图3所示的图像识别系统,请参阅图3,图3为本申请实施例中图像识别系统的一个环境示意图,如图所示,具体地,通常情况下,由服务器对大量的训练数据进行训练,在训练完成后,将训练好的图像识别模型存储在服务器本地。当需要进行图像识别时,可以由客户端向服务器发送待识别图像,服务器将该待识别图像输入至训练好的图像识别模型中,通过该模型输出对应的图像识别结果,再将该图像识别结果反馈至客户端。

[0099] 需要说明的是,客户端部署于终端设备上,其中,终端设备包含但不仅限于无人

车、监控设备、平板电脑、笔记本电脑、掌上电脑、手机、语音交互设备及个人电脑(personal computer,PC),此处不做限定。

[0100] 应理解,本申请提供的图像识别方法以及图像识别模型的训练方法均为基于人工智能(Artificial Intelligence,AI)实现的方法。人工智能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。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0101] 人工智能技术是一门综合学科,涉及邻域广泛,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。

[0102] 更具体地,图像识别方法可基于计算机视觉技术(Computer Vision,CV)实现。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、跟踪和测量等机器视觉,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、光学字符识别(Optical Character Recognition,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。

[0103] 机器学习(Machine Learning,ML)是一门多邻域交叉学科,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邻域。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。

[0104] 随着人工智能技术研究和进步,人工智能技术在多个邻域展开研究和应用,例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服等,相信随着技术的发展,人工智能技术将在更多的邻域得到应用,并发挥越来越重要的价值。

[0105] 本申请实施例提供的方案涉及人工智能的机器学习以及计算机视觉等技术,结合上述介绍,下面将对本申请中图像识别模型的训练方法进行介绍,请参阅图4,本申请实施例中图像识别模型的训练方法一个实施例包括:

[0106] 101、获取训练数据集合,其中,训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像;

[0107] 本实施例中,图像识别模型训练装置首先需要获取训练数据集合,训练数据集合主要包括两部分数据集合,分别为第一训练数据集合以及第二训练数据集合,其中,第一训

练数据集合包括具有标签的图像，假设第一训练数据集合有10000个图像样本对，那么每个图像样本对中包括一张样本图像及其对应的标签，如，图像A对应的标签为“小明”，图像B对应的标签为“小红”。另一部分数据集合为第二训练数据集合，第二训练数据集合包括不具有标记的图像，假设第二训练数据集合有50000个图像，那么这些图像均不含标签。

[0108] 在实际训练中，可以使用成比例的批次(batch)大小，假设总批次大小为m，第一训练数据集合为D_L，第二训练数据集合为D_U，一个批次中有标签图像样本的数量可以为$m\cdot\frac{|D_L|}{|D_L|+|D_U|}$，一个批次中无标签图像样本的数量可以为$m\cdot\frac{|D_U|}{|D_L|+|D_U|}$。

[0109] 本申请中，第一训练数据集合具体可以表示为$D_L=\{(x_i,y_i)|i=1,...,N_l\}$，其中，$(x_i,y_i)$是从未知分布$P_{X\times Y}$中独立采样得到的，$x_i$表示第一训练数据集合中的第i个图像，$y_i$表示第一训练数据集合中第i个图像所对应的标签，$N_l$表示第一训练数据集合中图像的总数。X为实例空间，可以表示为$X\subset R^d$。Y为标签空间，可以表示为$Y=\{1,2,...,K\}$，K为类别总数。$D'_L=\{x|(x,y)\in D_L\}$表示第一训练数据集合中的图像样本集合，但不包括图像的标签。

[0110] 第二训练数据集合具体可以表示为$D_U=\{x_i|i=1,...,N_u\}$，x_i表示第二训练数据集合中的第i个图像，N_u表示第二训练数据集合中图像的总数。

[0111] 需要说明的是，图像识别模型训练装置可以部署于服务器，也可以部署于终端设备，通常情况下，考虑到图像识别模型的训练过程可能会占用较多计算资源，因此，可以将图像识别模型训练装置部署于服务器，由服务器训练得到图像识别模型。但是对于计算能力较强的终端设备而言，可以将图像识别模型训练装置部署于该类终端设备上，故此处不做限定。

[0112] 102、调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一图像所对应的第一预测标签；

[0113] 本实施例中，图像识别模型训练装置从第一训练数据集合中任意取出一个图像，该图像为第一图像，然后调用待训练的图像识别模型，将第一图像输入至待训练的图像识别模型中，通过该模型输出对应的第一预测标签。以人脸图像分类任务为例，假设第一图像的标签为第一图像中人物的姓名，那么由模型输出的第一预测标签表示对该第一图像中的人物进行预测后得到的姓名。

[0114] 具体地，假设待训练的图像识别模型表示为$f_w:X\rightarrow[0,1]^{|Y|}$，其中，通过图像识别模型输出的预测标签表示为$C(x)=\arg\max f_w(x)$。如步骤101所描述的内容，第一图像可以是从第一训练数据集合的图像样本集合D'_L中取出的任意一张图像，该图像为第一图像，假设第一图像表示为x_1，则第一图像x_1所对应的第一预测标签表示为$C(x_1)$。

[0115] 103、调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二图像所对应的第二预测标签；

[0116] 本实施例中，图像识别模型训练装置从训练数据集合中任意取出一个图像，该图像为第二图像，即第二图像可以来源于第一训练数据集合，也可以来源于第二训练数据集合，此处不做限定。然后调用待训练的图像识别模型，将第二图像输入至待训练的图像识别模型中，通过该模型输出对应的第二预测标签。以人脸图像分类任务为例，由模型输出的第二预测标签表示对该第二图像中的人物进行预测后得到的姓名。

[0117] 具体地，如步骤101所描述的内容，第二图像可以是从第二训练数据集合的图像样本集合D_U中取出的任意一张图像，该图像为第二图像，假设第二图像表示为x_2，则第二图像x_2所对应的第二预测标签表示为$C(x_2)$。

[0118] 104、对第二图像进行加扰处理，得到第三图像；

[0119] 本实施例中，图像识别模型训练装置对第二图像进行加扰处理，得到加扰后的第三图像。其中，加扰处理表示对图像进行一定程度的干扰，这里的干扰可以是像素级别的干扰(比如修改图像中部分像素点的像素值)，或者是图像在空间上的干扰(比如对图像进行旋转或者平移等操作)，又或者是图像在函数上的干扰(比如通过函数为图像添加均匀噪声或者高斯噪声等)。

[0120] 105、调用图像识别模型对第三图像的类别进行预测，得到第三图像所对应的第三预测标签，其中，第三图像与第二图像具有对应关系；

[0121] 本实施例中，图像识别模型训练装置将经过加扰后的第三图像输入至待训练的图像识别模型中，通过该模型输出对应的第三预测标签。以人脸图像分类任务为例，由模型输出的第三预测标签表示对该第三图像中的人物进行预测后得到的姓名。

[0122] 106、根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像识别模型的第一模型参数进行更新。

[0123] 本实施例中，图像识别模型训练装置根据第一图像所对应的标签和第一图像所对应的第一预测标签，确定第一风险函数，基于第一风险函数输出的结果可以衡量图像识别模型的一般风险，第一风险函数输出的结果越大，表示模型的风险程度越高。图像识别模型训练装置根据第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，确定第二风险函数，基于第二风险函数输出的结果可以衡量图像识别模型的鲁棒风险，第二风险函数输出的结果越大，表示模型的风险程度越高。

[0124] 结合第一风险函数和第二风险函数，再利用反向传播以及随机梯度下降法来更新图像识别模型的模型参数，在一次迭代训练的过程中，对图像识别模型的第一模型参数进行更新，即更新为本轮训练计算得到的第二模型参数。

[0125] 本申请实施例中，提供了一种图像识别模型的训练方法，首先获取第一训练数据集合以及第二训练数据集合，然后调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一图像所对应的第一预测标签，调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二图像所对应的第二预测标签，再对第二图像进行加扰处理，得到第三图像，调用图像识别模型对第三图像的类别进行预测，最后根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像识别模型的第一模型参数进行更新。通过上述方式，利用了有标签数据和无标签数据对模型进行半监督训练，对于无标签的图像而言对其进行加扰，使得模型在训练过程中对是否加扰过的图像进行识别，从而提升模型的识别能力，增强模型的鲁棒性。

[0126] 可选地，在上述图4对应的各个实施例的基础上，本申请实施例提供的图像识别模型的训练方法另一个可选实施例中，根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像

识别模型的第一模型参数进行更新，可以包括：

[0127] 根据第一图像所对应的标签以及第一图像所对应的第一预测标签，确定第一风险函数，其中，第一风险函数用于表示预测标签与标签之间的差异；

[0128] 根据第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，确定第二风险函数，其中，第二风险函数用于表示已加扰图像与未加扰图像之间的差异；

[0129] 根据第一风险函数以及第二风险函数，生成目标优化函数；

[0130] 当目标优化函数达到最小值时，获取第二模型参数；

[0131] 将图像识别模型的第一模型参数更新为第二模型参数。

[0132] 本实施例中，介绍了一种结合两类风险函数对模型参数进行更新的方式，下面先介绍三类风险函数，分别为一般风险函数、鲁棒风险函数以及对抗风险函数。其中，一般风险函数(即第一风险函数)表示为：

[0133] $R_{stand}(D_L)=E_{(x,y)\sim X\times Y}[II\{C(x)\neq y\}]$； (1)

[0134] 其中，$R_{stand}(D_L)$表示一般风险函数，D_L表示训练数据集合中的第一训练数据集合(即包括至少一个图像样本对)，C(x)表示图像识别模型输出的预测标签，x表示第一训练数据集合中的图像，y表示图像的标签，X×Y表示第一训练数据集合D_L的分布，II(•)表示示性函数，例如，II(C(x)≠y)=1表示C(x)≠y的条件成立，II(C(x)≠y)=0表示C(x)≠y的条件不成立。

[0135] 具体地，基于公式(1)，图像识别模型训练装置根据第一图像所对应的标签以及第一图像所对应的第一预测标签，可以得到如下第一风险函数：

[0136] $R_{stand}(D_L)=E_{(x1,y1)\sim X\times Y}[II\{C(x_1)\neq y_1\}]$； (2)

[0137] 其中，x_1表示第一图像，y_1表示第一图像的标签，$C(x_1)$表示第一图像x_1的第一预测标签，D_L表示训练数据集合中的第一训练数据集合。第一风险函数用于表示预测标签与标签之间的差异。

[0138] 鲁棒风险函数(即第二风险函数)表示为：

[0139] $$R_{rob}(D_U)=E_{(x,y)\sim X\times Y}\left[\max_{x'\in N_{\varepsilon,T(x)}} II\{C(x')\neq C(x)\}\right]$$； (3)

[0140] 其中，$R_{rob}(D_U)$表示鲁棒风险函数，D_U表示训练数据集合中的第二训练数据集合，x′表示已加扰图像，x表示未加扰图像，C(x)表示图像识别模型输出图像x的预测标签，C(x′)表示图像识别模型输出图像x′的预测标签，$N_{\varepsilon,T(x)}$表示通用的扰动邻域。II(•)表示示性函数，例如，II(C(x′)≠C(x))=1表示C(x′)≠C(x)的条件成立，II(C(x′)≠C(x))=0表示C(x′)≠C(x)的条件不成立。需要说明的是，在实际应用中，$R_{rob}(D_U)$还可以表示为$R_{rob}(D_L\cup D_U)$，即表示未加扰图像x是从训练数据集合中取出的图像。

[0141] 具体地，基于公式(3)，图像识别模型训练装置根据第二图像以及第三图像，可以得到如下第二风险函数：

[0142] $$R_{rob}(D_U)=E_{(x,y)\sim X\times Y}\left[\max_{x_3\in N_{\varepsilon,T(x)}} II\{C(x_3)\neq C(x_2)\}\right]$$； (4)

[0143] 其中，x_2表示第二图像，x_3表示第三图像，$C(x_2)$表示第二图像x_2的第二预测标签，$C(x_3)$表示第三图像x_3的第三预测标签，第二风险函数用于表示已加扰图像与未加扰图像之

间的差异。

[0144] 对抗风险函数表示为：

[0145] $$R_{adv}(D)=E_{(x,y)\sim X\times Y}\left[\max_{x'\in N_{\varepsilon,T(x)}} \mathrm{II}\{C(x')\neq y\}\right]；(5)$$

[0146] 其中，$R_{adv}(D)$表示对抗风险函数，x′表示已加扰图像，y表示已加扰图像x′所对应的标签。

[0147] 结合公式(5)可知，在对抗风险函数中需要同时考虑未加扰图像以及经过干扰后图像的标签，在一般风险函数中只需要考虑未加扰图像，在鲁棒风险函数中不需要涉及图像的标签。然而，本申请无需获取经过干扰后图像的标签，这样不但可以增加样本数量，而且利用无标注的数据训练能够适用于半监督模式，从而更好地提升训练效率。基于此，将一般风险函数和鲁棒风险函数与对抗风险函数进行关联，可以得到如下公式：

[0148] $R_{adv}(x)=R_{stand}(x)+(1-R_{stand}(x))R_{rob}(x)$； (6)

[0149] 基于公式(6)可以进一步得到一般风险函数、鲁棒风险函数以及对抗风险函数之间具有如下关系：

[0150] $$\min_{w} R_{adv}(D)\leq \min_{w}\{R_{stand}(D)+R_{rob}(D)\}；(7)$$

[0151] 基于公式(7)可以推导出一种新的鲁棒训练方式，如下所示：

[0152] $$\min_{w} R_{stand}(D_L)+\lambda\cdot R_{rob}(D_L)；(8)$$

[0153] 其中，D_L表示第一训练数据集合，w表示图像识别网络的模型参数，λ表示大于0的超参数。在公式(8)的基础上，结合申请提供的第一训练数据集合为D_L以及第二训练数据集合为D_U，提出半监督防御方法的优化目标如下所示：

[0154] $$\min_{w} R_{stand}(D_L)+\lambda\cdot R_{rob}(D_L\cup D_U)；(9)$$

[0155] 其中，如果第二训练数据集合为D_U为空集，则公式(9)等于公式(8)。如果第二训练数据集合D_U不为空集，则$D_L\cup D_U$表示训练数据集合。结合公式(9)、公式(1)和公式(3)，可以得到目标优化函数。假设图像识别模型使用的模型参数为第一模型参数，经过一次迭代训练后，当目标优化函数具有最小值时，确定第二模型参数，然后将第二模型参数作为图像识别模型的模型参数，在下一次迭代训练的过程中，图像识别模型使用的模型参数为第二模型参数，经过多次迭代训练后，得到图像识别模型最终使用的模型参数。

[0156] 其次，本申请实施例中，提供了一种结合两类风险函数对模型参数进行更新的方式，即先根据第一图像所对应的标签以及第一图像所对应的第一预测标签，并且根据第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，确定第二风险函数，然后根据第一风险函数以及第二风险函数，生成目标优化函数，当目标优化函数达到最小值时，将图像识别模型的第一模型参数更新为第二模型参数。通过上述方式，能够结合一般风险和鲁棒风险表示对抗风险，相较于目前仅采用对抗风险作为衡量模型优劣的指标而言，本申请结合一般风险和鲁棒风险对模型性能进行评估的方式更全面，从而提升模型训练的鲁棒性。

[0157] 可选地，在上述图4对应的各个实施例的基础上，本申请实施例提供的图像识别模型的训练方法另一个可选实施例中，根据第一风险函数以及第二风险函数，生成目标优化

函数，可以包括：

[0158] 采用第一损失函数对第一风险函数进行变换处理，得到第一优化函数，其中，第一优化函数包括图像的预测分布向量与标签之间的损失值；

[0159] 采用第二损失函数对第二风险函数进行变换处理，得到第二优化函数，其中，第二优化函数包括已加扰图像的预测分布向量与未加扰图像的预测标签之间的损失值；

[0160] 根据第一优化函数与第二优化函数，生成目标优化函数。

[0161] 本实施例中，介绍了一种根据第一风险函数以及第二风险函数，生成目标优化函数的方式，由于第一风险函数与第二风险函数均包括示性函数，而示性函数是不可求导的，所以可以使用损失函数替代示性函数，本申请以交叉熵损失函数替代示性函数为例进行介绍，可以理解的是，在实际应用中，还可以采用其他类型的损失函数，此次不一一列举。

[0162] 具体地，假设以公式(8)作为优化目标，那么第二风险函数表示为公式(3)。假设以公式(9)作为优化目标，那么第二风险函数可以表示为：

[0163] $$R_{rob}\left(D_L \cup D_U\right)=E_{(x,y)\sim X\times Y}\left[\max_{x'\in N_{\varepsilon,T(x)}} \mathrm{II}\left\{C\left(x'\right)\neq C\left(x\right)\right\}\right]; \quad (10)$$

[0164] 为了增加样本数量，增强模型训练的鲁棒性，可以将第一训练数据集合和第二训练数据集合共同用于鲁棒性训练，类似地，在鲁棒性训练过程中，仅使用第一训练数据集合中的图像作为训练对象，然后对第一训练数据集合中的图像进行干扰，得到干扰后的图像，然后通过图像识别模型对干扰前的图像和干扰后的图像进行识别。

[0165] 具体地，采用第一损失函数对公式(1)所示的第一风险函数进行变换处理，得到如下第一优化函数：

[0166] $$\min_{w}\frac{1}{\left|D_L\right|}\sum_{(x,y)\in D_L} L_1\left(f_w\left(x\right),y\right); \quad (11)$$

[0167] 其中，$L_1(\cdot)$表示第一损失函数，D_L表示训练数据集合中的第一训练数据集合，x表示第一训练数据集合中的图像，y表示图像x的标签，$f_w(x)$表示图像x的预测分布向量。$L_1(f_w(x),y)$表示图像的预测分布向量$f_w(x)$与标签y之间的损失值。

[0168] 基于公式(11)，代入第一图像以及第一图像的标签，得到如下第一优化函数：

[0169] $$\min_{w}\frac{1}{\left|D_L\right|}\sum_{(x_1,y_1)\in D_L} L_1\left(f_w\left(x_1\right),y_1\right); \quad (12)$$

[0170] 其中，x_1表示第一图像，第一图像为未加扰图像，y_1表示第一图像x_1的标签。

[0171] 采用第二损失函数对公式(10)所示的第二风险函数进行变换处理，得到如下第二优化函数：

[0172] $$\min_{w}\frac{\lambda}{\left|D'_L \cup D_U\right|}\sum_{x\in D'_L\cup D_U}\max_{x'\in N_{\varepsilon,T(x)}} L_2\left(f_w\left(x'\right),C\left(x\right)\right); \quad (13)$$

[0173] 其中，$L_2(\cdot)$表示第二损失函数，D'_L表示第一训练数据集合D_L中的图像样本集合(即仅包括图像，不包括图像对应的标签)，D_U表示训练数据集合中的第二训练数据集合，x′表示已加扰图像，x表示未加扰图像，C(x)表示图像识别模型输出图像x的预测标签，$f_w(x')$表示图像识别模型输出已加扰图像的预测分布向量，$N_{\varepsilon,T(x)}$表示通用的扰动邻域。$L_2(f_w$

(x′),C(x))表示已加扰图像的预测分布向量与未加扰图像的预测标签之间的损失值。

[0174] 基于公式(13),代入第二图像以及第三图像,得到如下第二优化函数:

[0175] $$\min_{w}\frac{\lambda}{\left|D_L'\cup D_U\right|}\sum_{x_2\in D_L'\cup D_U}\max_{x_3\in N_{\varepsilon,T(x_2)}}L_2\left(f_w\left(x_3\right),C\left(x_2\right)\right);\quad(14)$$

[0176] 其中,x_2表示第二图像,第二图像x_2为未加扰图像,x_3表示第三图像,第三图像x_3为已加扰图像,$C(x_2)$表示第二图像x_2的第二预测标签,$f_w(x_3)$表示第三图像x_3的预测分布向量。

[0177] 基于公式(11)和公式(13),结合第一优化函数与第二优化函数,相加后得到如下目标优化函数:

[0178] $$\min_{w}\frac{1}{\left|D_L\right|}\sum_{(x,y)\in D_L}L_1\left(f_w\left(x\right),y\right)+\frac{\lambda}{\left|D_L'\cup D_U\right|}\sum_{x\in D_L'\cup D_U}\max_{x'\in N_{\varepsilon,T(x)}}L_2\left(f_w\left(x'\right),C\left(x\right)\right);\quad(15)$$

[0179] 根据公式(15)所示的目标优化函数,类似于对抗训练的优化过程,可以通过交替求解内部最大化(inner-maximization)和外部最小化(outer-minimization)子问题来求解上述优化目标。再给定已加扰图像x′时,可采用外部最小化的方式更新模型参数w,即:

[0180] $$w\leftarrow\arg\min_{w}\frac{1}{\left|D_L\right|}\sum_{(x,y)\in D_L}L_1\left(f_w\left(x\right),y\right)+\frac{\lambda}{\left|D_L'\cup D_U\right|}\sum_{x\in D_L'\cup D_U}L_2\left(f_w\left(x'\right),C\left(x\right)\right);\quad(16)$$

[0181] 基于公式(16),代入第一图像、第二图像、第三图像以及第一图像的标签,采用如下方式更新模型参数:

[0182] $$w\leftarrow\arg\min_{w}\frac{1}{\left|D_L\right|}\sum_{(x_1,y_1)\in D_L}L_1\left(f_w\left(x_1\right),y_1\right)+\frac{\lambda}{\left|D_L'\cup D_U\right|}\sum_{x_2\in D_L'\cup D_U}L_2\left(f_w\left(x_3\right),C\left(x_2\right)\right);\quad(17)$$

[0183] 其中,w表示第二模型参数。

[0184] 再次,本申请实施例中,提供了一种根据第一风险函数以及第二风险函数,生成目标优化函数的方式,即先根据第一风险函数获取第一优化函数,并且根据第二风险函数获取第二优化函数,结合第一优化函数和第二优化函数,可以生成目标优化函数。通过上述方式,考虑到示性函数不可求导的情况,因此引入损失函数对风险函数进行处理,由此生成能够求解函数结果的目标优化函数,从而可以提升训练的可靠性。

[0185] 可选地,在上述图4对应的各个实施例的基础上,本申请实施例提供的图像识别模型的训练方法另一个可选实施例中,对第二图像进行加扰处理,得到第三图像,可以包括:

[0186] 获取图像加扰类型;

[0187] 根据图像加扰类型确定扰动邻域,其中,扰动邻域表示对未加扰图像进行图像变换的范围;

[0188] 基于扰动邻域以及第二图像所对应的第二预测标签,确定第三优化函数;

[0189] 当第三优化函数达到最大值时,获取第二图像所对应的第三图像。

[0190] 本实施例中,介绍了一种对图像进行加扰的方式,图像识别模型训练装置首先需要确定图像加扰类型,根据图像加扰类型可以确定扰动邻域,扰动邻域表示已加扰图像的图像变换范围。

[0191] 具体地,某个未加扰图像x在ε变换范围内的扰动邻域定义为$N_{\varepsilon,T(x)}$,该扰动邻域的

一个通用示例为：

[0192] $N_{\varepsilon,T(x)}=\{T(x;\theta)\,|\,dist(T(x;\theta),x)\leqslant\varepsilon\}$； (18)

[0193] 其中，x表示未加扰图像，T(•;θ)表示具有参数θ的变换函数，dist(•,•)表示对应于T(•;θ)的一个给定距离度量，ε表示最大变换范围，且ε为非负参数。相比于定义在P范数意义下的邻域，公式(18)所定义的邻域具有更好的通用性。

[0194] 根据公式(15)所示的目标优化函数，基于扰动邻域以及当前图像识别模型所对应的模型参数，可以采用内部最大化的方式获取已加扰图像(即与未加扰图像对应的对抗样本)，即得到如下第三优化函数：

[0195] $$x'\leftarrow\underset{x'\in N_{\varepsilon,T(x)}}{\arg\max}\,L_2\left(f_w(x'),C(x)\right)$$；（19）

[0196] 其中，x表示未加扰图像，x′表示已加扰图像，C(x)表示未加扰图像x所对应的预测标签，$f_w(x')$表示已加扰图像x′的预测分布向量，$x'\in N_{\varepsilon,T(x)}$表示已加扰图像x′需要满足的扰动邻域。

[0197] 基于公式(19)，代入第二图像后得到如下第三优化函数：

[0198] $$x_3\leftarrow\underset{x_3\in N_{\varepsilon,T(x_2)}}{\arg\max}\,L_2\left(f_w(x_3),C(x_2)\right)$$；（20）

[0199] 其中，x_2表示第二图像，x_3表示第三图像，$C(x_2)$表示第二图像x_2所对应的第二预测标签，$f_w(x_3)$表示第三图像x_3的预测分布向量，$x_3\in N_{\varepsilon,T(x_2)}$表示第三图像$x_3$需要满足的扰动邻域。由此可见，当第三优化函数达到最大值时，即可获取对应的未加扰图像。

[0200] 进一步地，本申请实施例中，提供了一种对图像进行加扰的方式，首先需要确定图像加扰类型，然后根据图像加扰类型确定扰动邻域，然后基于扰动邻域以及第二图像所对应的第二预测标签，确定第三优化函数，当第三优化函数达到最大值时，获取第二图像所对应的第三图像。通过上述方式，在每次对抗训练的过程中利用内部最大化求解第三优化函数，从而生成对抗样本，即得到加扰后的图像，上述过程无需手动操作，一方面可以节省人力成本，另一方面，基于图像加扰类型对应的扰动邻域，可以直接生成满足变换条件的图像。

[0201] 可选地，在上述图4对应的各个实施例的基础上，本申请实施例提供的图像识别模型的训练方法另一个可选实施例中，获取图像加扰类型，可以包括：

[0202] 获取图像攻击类型的数量；

[0203] 若图像攻击类型的数量等于1，则确定图像加扰类型为单攻击类型；

[0204] 若图像攻击类型的数量大于1，则确定图像加扰类型为复合攻击类型。

[0205] 本实施例中，介绍了一种对图像进行攻击的方式，图像识别模型训练装置获取图像攻击类型的数量，如果图像攻击类型的数量为1，则表示只存在一种类型的图像攻击方式，因此，图像加扰类型属于单攻击类型。如果图像攻击类型的数量大于1，则表示存在至少两种类型的图像攻击方式，因此，图像加扰类型属于复合攻击加扰类型。

[0206] 具体地，下面将以三种常见的图像攻击方式为例，对生成对抗样本的类型进行介绍。这三类常见的图像攻击方式分别为像素攻击(pixel-wise attacks)类型、几何攻击类型(spatial attacks)以及色彩变化攻击(color shifting attacks)类型。因此，单攻击类

型为像素攻击类型、几何攻击类型或者色彩变化攻击类型。需要说明的是，图像攻击方式还可以其他类型，此次仅为一个示意，不对图像攻击类型进行穷举。

[0207] 复合攻击类型可以包括像素攻击类型以及几何攻击类型，在实际情况下，可以先对图像进行像素攻击，再进行几何攻击。也可以先对图像进行几何攻击，再进行像素攻击。

[0208] 复合攻击类型可以包括像素攻击类型以及色彩变化攻击类型，在实际情况下，可以先对图像进行像素攻击，再进行色彩变化攻击。也可以先对图像进行色彩变化攻击，再进行像素攻击。

[0209] 复合攻击类型可以包括几何攻击类型以及色彩变化攻击类型，在实际情况下，可以先对图像进行几何攻击，再进行色彩变化攻击。也可以先对图像进行色彩变化攻击，再进行几何攻击。

[0210] 复合攻击类型可以包括像素攻击类型、几何攻击类型以及色彩变化攻击类型，在实际情况下，可以先对图像进行像素攻击，再进行几何攻击，最后进行色彩变化攻击。也可以按照其他的顺序对图像进行攻击，此处不做赘述。

[0211] 更进一步地，本申请实施例中，提供了一种对图像进行攻击的方式，在获取到至少一个图像加扰类型之后，需要根据图像加扰类型的数量，确定当前攻击属于单攻击类型还是复合攻击类型。通过上述方式，可以实现模型的定向训练，由于现有方案中，能够识别单攻击类型图像的模型，往往对复合攻击类型图像的识别精度较低，因此，进行模型训练时，可以设计不同的图像加扰类型，以此防御这些不同图像加扰类型下被攻击的图像。

[0212] 可选地，在上述图4对应的各个实施例的基础上，本申请实施例提供的图像识别模型的训练方法另一个可选实施例中，根据图像加扰类型确定扰动邻域，可以包括：

[0213] 若图像加扰类型为单攻击类型，则确定图像攻击类型；

[0214] 若图像攻击类型为像素攻击类型，则获取像素变换函数对应的像素距离度量，其中，像素攻击类型为对未加扰图像中的至少一个像素值进行变换；

[0215] 根据像素距离度量以及最大像素变换范围，确定扰动邻域。

[0216] 本实施例中，介绍了一种针对像素攻击类型进行模型训练的方式，图像识别模型训练装置在确定图像加扰类型为单攻击类型之后，需要进一步获取具体的图像攻击类型，如果该图像攻击类型为像素攻击类型，那么可以进一步得到像素距离度量，具体地，基于公式(18)可知，假设像素变换函数为$T(x;\theta)=x+\theta$，该像素变换函数所对应的像素距离度量表示为$dist(T(x;\theta),x)=||T(x;\theta)-x||_{\infty}$，给定一个最大像素变换范围$\varepsilon$，从而得到扰动邻域$N_{\varepsilon,T(x)}=\{T(x;\theta)=x+\theta|dist(T(x;\theta),x)=||T(x;\theta)-x||_{\infty}\leq\varepsilon\}$。

[0217] 为了进一步验证本申请提供的技术方案，对本申请提供的图像识别模型进行了一系列的实验，在实验设置中，采用的数据集包括CIFAR-10数据集和MNIST数据集，CIFAR-10数据集包括50000个训练样本和10000个测试样本，MNIST数据集包括60000个训练样本和10000个测试样本。实验过程中随机从训练数据集中选出10000个样本作为有标签数据，剩下的所有样本去除其标签作为无标签样本。需要说明的是，本申请提供的图像识别模型的训练方法既使用了有标签数据，又使用了无标签数据，因此，本申请提供的的方法可以称为半监督鲁棒训练(semi-supervised robust training,SRT)防御方法。而使用全量有标签数据进行训练的方法可以称为鲁棒训练(robust training,RT)防御方法。为了更直观地比对SRT防御方法与RT防御方法在像素攻击类型下的差异，在实验过程中采用CIFAR-10数据

集对像素攻击类型下的SRT防御方法与RT防御方法进行比对。请参阅表1，表1为实验中SRT防御方法与RT防御方法在像素攻击类型下的一个精度对比结果。

[0218] 表1

[0219]

	无对抗	FGSM	PGD
标准训练方法(未防御)	87.69%	6.65%	0
RT防御方法(λ＝0.20)	83.24%	48.94%	31.91%
SRT防御方法(λ＝0.20)	83.83%	51.39%	34.94%
RT防御方法(λ＝0.40)	81.05%	51.15%	36.01%
SRT防御方法(λ＝0.40)	82.28%	56.06%	41.84%
RT防御方法(λ＝0.60)	79.93%	51.56%	37.73%
SRT防御方法(λ＝0.60)	81.03%	56.83%	44.60%
RT防御方法(λ＝0.80)	78.37%	51.97%	37.91%
SRT防御方法(λ＝0.80)	80.23%	58.14%	47.24%

[0220] 由表1可知，标准训练方法在快速梯度符号法(Fast Gradient Sign Method，FGSM)与投影梯度下降(Project Gradient Descent，PGD)攻击方法下的对抗精度较低。而基于相同的条件(即超参数λ相同且CIFAR-10数据集相同)，采用SRT防御方法相较于采用RT防御方法，对于FGSM对抗方法和PGD对抗方法的对抗精度更高，尤其在PGD对抗方法下，SRT防御方法的对抗精度通常比RT防御方法高出5%。

[0221] 在实验过程中还采用MINIST数据集对像素攻击类型下的SRT防御方法与RT防御方法进行比对。请参阅表2，表2为实验中SRT防御方法与RT防御方法分别在像素攻击类型下的一个精度对比结果。

[0222] 表2

[0223]

	无对抗	FGSM	PGD
标准训练方法	99.02%	93.80%	86.12%
RT防御方法(λ＝0.20)	99.06%	97.18%	95.84%
SRT防御方法(λ＝0.20)	99.31%	98.10%	97.18%
RT防御方法(λ＝0.40)	99.14%	97.57%	96.23%
SRT防御方法(λ＝0.40)	99.40%	98.28%	97.55%
RT防御方法(λ＝0.60)	99.06%	97.78%	96.92%
SRT防御方法(λ＝0.60)	99.34%	98.47%	97.81%
RT防御方法(λ＝0.80)	99.11%	97.90%	97.06%
SRT防御方法(λ＝0.80)	99.35%	98.53%	97.86%

[0224] 由表2可知，基于相同的条件(即超参数λ相同且数据集相同)，采用SRT防御方法相较于采用RT防御方法，对于FGSM对抗方法和PGD对抗方法的对抗精度更高，尤其在PGD对抗方法下，SRT防御方法与RT防御方法相比，具有更强的鲁棒性。

[0225] 此外，对本申请提供的SRT防御方法与其他类型的防御方法进行比较，其他类型的防御方法包括标准训练(standard training)防御方法、对抗训练(adversarial training，AT)防御方法、协调激发对抗性防御(trade-off inspired adversarial defense，TRADES)防御方法、无监督对抗训练(unsupervised adversarial training，UAT)

防御方法以及鲁棒自训练(robust self-training,SRT)防御方法。此外,所采用的对抗方法分别为FGSM对抗方法、PGD对抗方法、动量迭代FGSM(momentum iterative FGSM,MI-FGSM)对抗方法、雅可比显著图攻击(jacobian saliency map attack,JSMA)对抗方法、卡里尼和瓦格纳(Carlini&Wagner attack C&W,)对抗方法、逐点攻击(point-wise attack)对抗方法以及方向与规范攻击(direction and norm attack,DDNA)。基于此,请参阅表3,表3为实验中基于CIFAR-10数据集和MNIST数据集在不同防御方式下的一个对抗精度对比结果。

[0226] 表3

[0227]

		无对抗	FGSM	PGD	MI-FGSM	JSMA	C&W	逐点攻击	DDNA
CIFAR-10 数据集	Standard	88.43	7.26	0	0	9.98	0.14	2.06	0.01
	AT	77.17	50.35	37.37	36.97	10.80	33.42	14.27	20.87
	TRADES	76.23	51.98	40.2	39.79	11.90	36.09	12.32	22.37
	UAT	78.45	56.35	44.96	44.56	13.23	40.03	12.62	24.08
	RST	79.99	59.55	48.38	47.97	13.78	42.99	13.42	27.73
	SRT	78.46	59.34	48.66	48.24	16.99	43.33	18.80	27.71
MNIST 数据集	Standard	99.01	41.06	2.87	4.37	10.18	0.01	0.04	16.44
	AT	98.99	96.61	94.69	93.57	20.99	94.67	2.47	95.35
	TRADES	98.99	96.92	95.12	93.98	18.48	92.37	2.08	94.12
	UAT	99.16	97.51	96.14	95.65	23.79	96.16	5.52	96.39

[0228]

	RST	98.83	97.21	95.47	95.05	26.63	95.34	3.30	94.25
	SRT	99.28	97.77	96.60	95.79	26.79	96.19	5.81	96.10

[0229] 由表3可知,本申请提供的SRT防御方法相较于其他类型的防御方法,具有更高的对抗精度。此外,尽管UAT防御方法和RST防御方法在某些情况下的对抗精度高于SRT防御方法的对抗精度,但是他们在不同数据集上的效果并不一致,例如,RST防御方法在CIFAR-10数据集上的性能优于在MINIST数据集的性能。因此,本申请提供的SRT防御方法具有更强的适应性和鲁棒性。

[0230] 再进一步地,本申请实施例中,提供了一种针对像素攻击类型进行模型训练的方式,即在像素攻击类型下,需要获取像素变换函数对应的像素距离度量,然后确定扰动邻域,再基于扰动邻域生成对应的对抗样本,利用该对抗样本对图像识别模型的训练。通过上述方式,可以有针对性地训练用于识别像素攻击类型的模型,从而提升模型能够更精准地识别出经过像素攻击的图像,由此提升模型的鲁棒性。

[0231] 可选地,在上述图4对应的各个实施例的基础上,本申请实施例提供的图像识别模型的训练方法另一个可选实施例中,根据图像加扰类型确定扰动邻域,可以包括:

[0232] 若图像加扰类型为单攻击类型,则确定图像攻击类型;

[0233] 若图像攻击类型为几何攻击类型,则获取几何变换函数所对应的几何距离度量,其中,几何攻击类型为对未加扰图像进行平移以及旋转中的至少一种变换;

[0234] 根据几何距离度量以及最大几何变换范围,确定扰动邻域。

[0235] 本实施例中,介绍了一种针对几何类型进行模型训练的方式,图像识别模型训练装置在确定图像加扰类型为单攻击类型之后,需要进一步获取具体的图像攻击类型,如果

该图像攻击类型为几何攻击类型，那么可以进一步得到几何距离度量。几何攻击类型包括对未加扰图像进行平移和旋转中的至少一种，对未加扰图像进行旋转可以表示为Ax，对未加扰图像进行平移可以表示为x+B，对未加扰图像既进行旋转又平移可以表示为Ax+B。本申请以旋转为例进行介绍，但这不应理解为对本申请的限定。

[0236] 基于公式(18)可知，假设几何变换函数为$T(x;\theta)=[\cos\theta,-\sin\theta;\sin\theta,\cos\theta]x$，该几何变换函数所对应的几何距离度量表示为$dist(T(x;\theta),x)=\theta$，给定一个最大几何变换范围$\varepsilon$，从而得到扰动邻域$N_{\varepsilon,T(x)}=\{T(x;\theta)=[\cos\theta,-\sin\theta;\sin\theta,\cos\theta]x=\theta\leqslant\varepsilon\}$。

[0237] 为了进一步验证本申请提供的方案，对本申请提供的图像识别模型进行了一系列的实验，在实验设置中，采用的数据集包括CIFAR-10数据集和MNIST数据集，类似地，实验过程中随机从训练数据集中选出10000个样本作为有标签数据，剩下的所有样本去除其标签作为无标签样本。需要说明的是，本申请提供的的方法可以称为SRT防御方法。而使用全量有标签数据进行训练的方法可以称为RT防御方法。为了更直观地比对SRT防御方法与RT防御方法在几何攻击类型下的差异，在实验过程中采用CIFAR-10数据集对几何攻击类型下的SRT防御方法与RT防御方法进行比对。请参阅表4，表4为实验中SRT防御方法与RT防御方法在几何攻击类型下的一个精度对比结果。

[0238] 表4

[0239]

	无对抗	RandAdv	GridAdv
标准训练方法	80.63	8.82	0.09
RT防御方法(λ=0.15)	85.24	64.45	41.23
SRT防御方法(λ=0.15)	88.04	78.03	62.97
RT防御方法(λ=0.20)	85.71	66.43	44.28
SRT防御方法(λ=0.20)	88.87	78.99	64.83
RT防御方法(λ=0.25)	85.59	67.95	45.93
SRT防御方法(λ=0.25)	88.40	78.39	64.15
RT防御方法(λ=0.30)	84.99	67.47	45.72
SRT防御方法(λ=0.30)	87.99	78.12	62.73

[0240] 由表4可知，标准训练方法在随机采样对抗(random sampling adversarial，RandAdv)方法与网格搜索对抗(grid search adversarial，GridAdv)攻击方法下的对抗精度较低。而基于相同的条件(即超参数λ相同且CIFAR-10数据集相同)，采用SRT防御方法相较于采用RT防御方法，对于RandAdv对抗方法和GridAdv对抗方法的对抗精度更高，尤其在GridAdv对抗方法下，SRT防御方法的对抗精度通常比RT防御方法高出17％以上。

[0241] 在实验过程中还采用MINIST数据集对几何攻击类型下的SRT防御方法与RT防御方法进行比对。请参阅表5，表5为实验中SRT防御方法与RT防御方法分别在几何攻击类型下的一个精度对比结果。

[0242] 表5

[0243]

	无对抗	RandAdv	GridAdv
标准训练方法	97.19	71.00	40.49
RT防御方法(λ=0.15)	98.47	92.88	72.85
SRT防御方法(λ=0.15)	98.61	96.85	91.52

RT防御方法(λ=0.20)	98.33	93.66	76.68
SRT防御方法(λ=0.20)	98.64	97.02	92.12
RT防御方法(λ=0.25)	98.29	93.91	78.00
SRT防御方法(λ=0.25)	98.63	96.91	91.70
RT防御方法(λ=0.30)	98.42	93.86	77.74
SRT防御方法(λ=0.30)	98.62	97.08	91.44

[0244] 由表5可知,基于相同的条件(即超参数λ相同且MNIST数据集相同),采用SRT防御方法相较于采用RT防御方法,对于RandAdv对抗方法和GridAdv对抗方法的对抗精度更高,尤其在GridAdv对抗方法下,SRT防御方法与RT防御方法相比,具有更强的鲁棒性。

[0245] 此外,对本申请提供的SRT防御方法与其他类型的防御方法进行比较,其他类型的防御方法包括标准训练(standard training)防御方法、AT)、防御方法、Worst-of-k防御方法以及基于K-L散度的正则化(K-L divergence based regularization,KLR)防御方法。此外,所采用的对抗方法分别为RandAdv对抗方法、GridAdv对抗方法、基于旋转的RandAdv对抗方法(RandAdv for rotations,RandAdv.R)、基于旋转的GridAdv对抗方法(GridAdv for rotations,GridAdv.R)、基于平移的RandAdv对抗方法(RandAdv for transformations,RandAdv.T)以及基于平移的GridAdv对抗方法(GridAdv for transformations,GridAdv.T)。基于此,请参阅表6,表6为实验中基于CIFAR-10数据集和MNIST数据集在不同防御方式下的一个对抗精度对比结果。

[0246] 表6

[0247]

		无对抗	RandAdv	GridAdv	RandAdv.T	GridAdv.T	RandAdv.R	GridAdv.R
CIFAR-10数据集	Standard	80.63	8.82	0.09	33.61	19.67	19.55	10.73
	AT	65.59	4.92	0.22	16.23	7.49	14.47	8.37
	Worst-of-k	82.02	70.92	54.80	75.49	69.45	73.18	68.22
	KLR	85.40	72.77	56.28	77.43	72.71	74.80	71.04
	SRT	88.87	78.99	64.83	82.16	78.47	80.84	77.24
MNIST数据集	Standard	97.19	14.01	0.00	35.31	5.12	65.06	51.32
	AT	97.96	27.84	0.01	51.83	10.72	71.66	57.71
	Worst-of-k	98.05	94.77	84.64	96.07	93.99	95.70	94.24
	KLR	98.43	95.26	86.08	96.63	95.07	95.92	94.48
	SRT	98.64	97.02	92.12	97.68	96.85	97.33	96.54

[0248] 由表6可知,本申请提供的SRT防御方法相较于其他类型的防御方法,具有更高的对抗精度。

[0249] 再进一步地,本申请实施例中,提供了一种针对几何攻击类型进行模型训练的方式,即在几何攻击类型下,需要获取几何变换函数对应的像素距离度量,然后确定扰动邻域,再基于扰动邻域生成对应的对抗样本,利用该对抗样本对图像识别模型的训练。通过上述方式,可以有针对性地训练用于识别几何攻击类型的模型,从而提升模型能够更精准地识别出经过几何攻击的图像,由此提升模型的鲁棒性。

[0250] 可选地,在上述图4对应的各个实施例的基础上,本申请实施例提供的图像识别模型的训练方法另一个可选实施例中,根据图像加扰类型确定扰动邻域,可以包括:

[0251] 若图像加扰类型为复合攻击类型,则获取至少两个图像攻击类型;

[0252] 若至少两个图像攻击类型包括像素攻击类型以及几何攻击类型，则获取图像攻击顺序；

[0253] 根据图像攻击顺序确定扰动邻域。

[0254] 本实施例中，介绍了一种针对复合攻击类型进行模型训练的方式，图像识别模型训练装置在确定图像加扰类型为复合攻击类型之后，需要进一步获取至少两种不同的图像攻击类型，如果至少两种不同的图像攻击类型包括像素攻击类型和几何攻击类型，那么还需要进一步图像攻击顺序，基于不同的图像攻击顺序可以生成不同的变换领域，从而使得图像识别模型能够更好地识别出按照一定图像攻击顺序生成的图像。

[0255] 一种可能的图像攻击顺序为，先对未加扰图像进行像素攻击，再对已经过像素攻击后的图像进行几何攻击。另一种可能的图像攻击顺序为，先对未加扰图像进行几何攻击，再对已经过几何攻击后的图像进行像素攻击。

[0256] 再进一步地，本申请实施例中，提供了一种针对复合攻击类型进行模型训练的方式，即在复合攻击类型下，需要根据复合攻击所对应的图像攻击顺序确定扰动邻域，再基于扰动邻域生成对应的对抗样本，利用该对抗样本对图像识别模型的训练。通过上述方式，可以有针对性地训练用于识别复合攻击类型的模型，从而提升模型能够更精准地识别出经过复合攻击的图像，由此提升模型的鲁棒性。

[0257] 可选地，在上述图4对应的各个实施例的基础上，本申请实施例提供的图像识别模型的训练方法另一个可选实施例中，根据图像攻击顺序确定扰动邻域，可以包括：

[0258] 若图像攻击顺序为先采用像素攻击类型，再采用几何攻击类型，则获取第一复合变换函数所对应的第一复合距离度量，其中，像素攻击类型为对未加扰图像中的至少一个像素值进行变换，几何攻击类型为对未加扰图像进行平移以及旋转中的至少一种变换；

[0259] 根据第一复合距离度量以及最大几何变换范围，确定扰动邻域；

[0260] 或，根据图像攻击顺序确定扰动邻域，包括：

[0261] 若图像攻击顺序为先采用几何攻击类型，再采用像素攻击类型，则获取第二复合变换函数所对应的第二复合距离度量；

[0262] 根据第二复合距离度量以及最大像素变换范围，确定扰动邻域。

[0263] 本实施例中，介绍了一种图像攻击顺序确定扰动邻域的方式，图像识别模型训练装置需要根据不同的图像攻击顺序确定扰动邻域，以复合攻击类型对应于像素攻击类型以及几何攻击类型为例进行介绍。

[0264] 第一种复合防御的情况下，先对图像进行像素攻击，再对像素攻击后的图像进行几何攻击，基于公式(18)可知，给定一个变换范围ε，其中，$\varepsilon=(\varepsilon_1,\varepsilon_2,\varepsilon_3)$，$\varepsilon_1$表示最大几何变换范围中旋转变换范围，$\varepsilon_2$表示最大几何变换范围中平移变换范围，$\varepsilon_3$表示最大像素变换范围。若第一复合变换函数为$T_1(x)=A(\theta)\cdot\left(x+\arg\max_{r\in\beta_\infty(\in)}L\left(f\left(x+r\right),y\right)\right)+B$，第一复合变换函数的第一复合距离度量表示为$dist_1(T_1(x),x)=(\theta,||B||_{\infty,\infty})\leqslant(\varepsilon_1,\varepsilon_2)$，最大几何变换范围为$(\varepsilon_1,\varepsilon_2)$，其中，$A(\theta)=[\cos\theta,-\sin\theta;\sin\theta,\cos\theta]$，$\beta_\infty(\in)=\{x|||x||_\infty\leqslant\in\}$，从而得到扰动邻域。

[0265] 第二种复合防御的情况下，先对图像进行几何攻击，再对几何攻击后的图像进行像素攻击，基于公式(18)可知，给定一个变换范围ε，其中，$\varepsilon=(\varepsilon_1,\varepsilon_2,\varepsilon_3)$，$\varepsilon_1$表示最大几何

变换范围中旋转变换范围，ε_2表示最大几何变换范围中平移变换范围，ε_3表示最大像素变换范围。若第二复合变换函数为$T_2(x,r)=A^*x+B^*+r$，第二复合变换函数的第二复合距离度量表示为$dist_2(T_2(x),x)=||r||_\infty \leqslant \varepsilon_3$，最大像素变换范围为$\varepsilon_3$，其中，$\left(A^*,B^*\right)=\arg\max_{\left(A(\theta),B:\theta\leq\varepsilon_1,\|B\|_{\infty,\infty}\leq\varepsilon_2\right)} L\left(A\left(\theta\right)x+B,y\right)$，从而得到扰动邻域。

[0266] 更进一步地，本申请实施例中，提供了一种图像攻击顺序确定扰动邻域的方式，若图像攻击顺序为先采用像素攻击类型，再采用几何攻击类型，则获取第一复合变换函数所对应的第一复合距离度量，再根据第一复合距离度量以及最大几何变换范围，确定扰动邻域。若图像攻击顺序为先采用几何攻击类型，再采用像素攻击类型，则获取第二复合变换函数所对应的第二复合距离度量，再根据第二复合距离度量以及最大像素变换范围，确定扰动邻域。通过上述方式，在模型训练中可以根据图像攻击顺序生成对抗图像，不但能够更有针对性地训练用于识别复合攻击类型的模型，还可以对于具体特定图像攻击顺序的图像具有更好的识别度，从而提升模型的鲁棒性。

[0267] 结合上述介绍，下面将对本申请中图像识别的方法进行介绍，请参阅图5，本申请实施例中图像识别的方法一个实施例包括：

[0268] 201、获取待识别图像；

[0269] 本实施例中，图像识别装置首先获取待识别图像，其中，待识别图像可以是人脸图像，也可以是其他类型的图像。

[0270] 需要说明的是，图像识别装置可以部署于服务器，也可以部署于终端设备，通常情况下，考虑到图像识别模型可能会占用较多内存，因此，可以将图像识别装置部署于服务器，由终端设备采集待识别图像，然后传输至服务器，由服务器调用图像识别模型对该待识别图像进行识别。但是对于计算能力较强的终端设备而言，可以将图像识别装置部署于该类终端设备上，故此处不做限定。

[0271] 202、调用图像识别模型对待识别图像的类别进行预测，得到图像类别结果，其中，图像识别模型为上述实施例中涉及的图像识别模型；

[0272] 本实施例中，图像识别装置调用已经训练好的图像识别模型，将待识别图像输入至该图像识别模型，通过图像识别模型对待识别图像的类别进行预测，输出预测分布向量，其中，预测分布向量表示每个类别的概率，且预测分布向量中各个元素之和通常为1。具体地，假设需要识别的是图像中人物的性别，那么这是一个二分类问题，即输出的预测分布向量包括两个元素，第一元素表示图像属于“男性”标签的概率，第二元素表示图像属于“女性”标签的概率。例如预测分布向量为(0.9,0.1)，则表示属于“男性”标签的概率为0.9，属于“女性”标签的概率为0.1，因此，该图像的图像类别结果为“男性”。

[0273] 可以理解的是，对于多分类问题的处理方式类似，若存在K个分类标签，则预测分布向量中包括K个元素，将K个元素中最大值对应的标签作为图像类别结果。

[0274] 203、向客户端发送图像类别结果，以使客户端展示图像类别结果。

[0275] 本实施例中，图像识别装置在获取到图像类别结果之后，可以向客户端推送该图像类别结果。如果图像识别装置部署于服务器，则服务器向终端设备发送图像类别结果，通过终端设备上的客户端进行展示图像类别结果。如果图像识别装置部署于终端设备，则终端设备直接通过客户端展示图像类别结果。

[0276] 本申请实施例中，提供了一种图像识别的方法，首先获取待识别图像，然后调用训练好的图像识别模型对待识别图像进行识别，然后输出对应的图像类别结果，最后通过客户端呈现识别到的图像类别结果。通过上述方式，由于图像识别模型采用的是基于通用对抗噪声训练下的半监督对抗防御方法，而这类方法相对于全监督对抗防御方法而言具有更好的训练效果，能够在有额外无标签数据时获得了优异的性能表现。此外，相较于现有方案中只能防御单一类型攻击的对抗防御方法，本申请训练得到的图像识别模型可以在一定程度上同时抵抗不同类型的攻击，以及不同类型攻击的复合。因此，在模型鲁棒性较好的情况下，能够在实际的图像识别过程中，输出更加准确的图像类别结果。

[0277] 为了进一步验证本申请提供的图像识别模型对复合攻击的防御情况，对本申请提供的图像识别模型进行了一系列的实验，在实验设计中主要针对两种复合攻击的方法进行介绍，第一种复合攻击的方法为先对图像进行像素攻击，再对像素攻击后的图像采用几何攻击，具体地，像素攻击可以采用PGD攻击方法，几何攻击可以采用GridAdv攻击方法，为了便于介绍，这种方法可以称为PGD+复合攻击方法。第二种复合攻击的方法为先对图像进行几何攻击，再对几何攻击后的图像采用像素攻击，为了便于介绍，这种方法可以称为GridAdv+复合攻击方法。此外，为了更好地比对实验数据，实验中还引入了单攻击的方法作为参考，分别为像素攻击和几何攻击，其中，像素攻击可以采用PGD攻击方法，几何攻击可以采用GridAdv攻击方法。

[0278] 在复合攻击的防御设置上，基于通用邻域定义生成对抗训练所需的对抗样本。在实验设计中主要针对两种对抗样本生成方法所对应的复合防御方法进行介绍，第一种复合防御的方法为先对图像进行像素攻击的防御，在对图像进行几何攻击的防御，具体地，对像素攻击的防御可以采用AT防御方法，对几何攻击的防御可以采用Worst-of-k防御方法，为了便于介绍，这种方法可以称为AT+防御方法。第二种复合防御的方法为先对图像进行像素攻击的防御，在对图像进行几何攻击的防御，具体地，对像素攻击的防御可以采用像素攻击设置下的SRT防御方法，对几何攻击的防御可以采用几何攻击设置下的SRT防御方法，为了便于介绍，这种方法可以称为SRT+防御方法。

[0279] 请参阅图6A和图6B，图6A为实验中基于CIFAR-10数据集对复合攻击进行防御的一个效果对比示意图，图6B为实验中基于MNIST数据集对复合攻击进行防御的一个效果对比示意图，如图所示，最大扰动范围与以往实验中的扰动范围相同，扰动分值表示当前扰动大小与先前最大扰动大小的比率。与单一类型的攻击相比，复合攻击(GridAdv+和PGD+)在相同条件下具有更大的威胁。这种现象表明，通过简单地组合不同类型的攻击，就可以构建强大的攻击，这对深度神经网络(Deep Neural Networks，DNN)构成了巨大的威胁。尤其是空间对抗防御对像素级防御几乎没有帮助，这表明单一类型的防御可能不会对防御另一种类型的攻击产生太大影响。

[0280] 此外，结果还表明，使用通用扰动邻域(即AT+和SRT+)训练的模型，可以同时防御不同类型的攻击。与单一类型防御方法相比，复合防御在防御复合攻击方面取得了显着改进，并且对单一攻击具有良好的防御性能。另外，在复合防御的情况下，本申请采用的复合防御方法(SRT+)也比现有的防御方法(AT+)更好。因此，基于通用对抗扰动生成的防御方法具有重要意义。

[0281] 请参阅图7A和图7B，图7A为实验中基于CIFAR-10数据集对复合攻击进行防御的另

一个效果对比示意图，图7B为实验中基于MNIST数据集对复合攻击进行防御的另一个效果对比示意图，如图所示，最大扰动范围与以往实验中的扰动范围相同，扰动分值表示当前扰动大小与先前最大扰动大小的比率。本申请提供的复合防御方法(SRT+)在复合攻击下具有更好的对抗鲁棒性，也优于其他的防御方法，由此在MNIST数据集中，与次优的AT+防御方法相比，在PGD+复合攻击下的对抗精度提高了7％以上，在GridAdv+复合攻击下的对抗精度提高了10％以上。

[0282] 为了进一步验证本申请提供的模型训练方法与无标签样本数量之间的关系，对本申请提供的图像识别模型进行了一系列的实验，在实验设计中，固定有标签样本数量为10000，然后从0开始缓慢增加无标签样本的数量，探索无标签样本数量对本申请提出的SRT防御方法的性能影响。

[0283] 请参阅图8A和图8B，图8A为实验中基于CIFAR-10数据集下SRT性能随无标签数据使用量变化的一个曲线，图8B为实验中基于MNIST数据集下SRT性能随无标签数据使用量变化的一个曲线，如图所示，SRT防御方法即为本申请提供的图像识别模型的训练方法，在所有设置中，使用SRT防御方法训练的模型的对抗精度随无标签样本的数量而增加。此外，利用大量无标签的样本，SRT防御方法具有与使用全部有标签样本的对抗训练有相似的性能。另外，对抗精度在曲线的末端仍具有上升趋势，这意味着如果使用更多无标签的样本，模型的鲁棒性可以获得进一步的提升。上述结果表明，本申请提供的模型训练方法充分的利用了无标签样本的信息，在半监督情形下获得优异的表现，因而具有重要意义。

[0284] 下面对本申请中的图像识别模型训练装置进行详细描述，请参阅图9，图9为本申请实施例中图像识别模型训练装置一个实施例示意图，图像识别模型训练装置30包括：

[0285] 获取模块301，用于获取训练数据集合，其中，训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像；

[0286] 预测模块302，用于调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一图像所对应的第一预测标签；

[0287] 预测模块302，还用于调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二图像所对应的第二预测标签；

[0288] 处理模块303，用于对第二图像进行加扰处理，得到第三图像；

[0289] 预测模块302，还用于调用图像识别模型对第三图像的类别进行预测，得到第三图像所对应的第三预测标签，其中，第三图像与第二图像具有对应关系；

[0290] 更新模块304，用于根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像识别模型的第一模型参数进行更新。

[0291] 本申请实施例中，提供了一种图像识别模型训练装置，采用上述方式，利用了有标签数据和无标签数据对模型进行半监督训练，对于无标签的图像而言对其进行加扰，使得模型在训练过程中对是否加扰过的图像进行识别，从而提升模型的识别能力，增强模型的鲁棒性。

[0292] 可选地，在上述图9所对应的实施例的基础上，本申请实施例提供的图像识别模型训练装置30的另一实施例中，

[0293] 更新模块304，具体用于根据第一图像所对应的标签以及第一图像所对应的第一

预测标签，确定第一风险函数，其中，第一风险函数用于表示预测标签与标签之间的差异；

[0294] 根据第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，确定第二风险函数，其中，第二风险函数用于表示已加扰图像与未加扰图像之间的差异；

[0295] 根据第一风险函数以及第二风险函数，生成目标优化函数；

[0296] 当目标优化函数达到最小值时，获取第二模型参数；

[0297] 将图像识别模型的第一模型参数更新为第二模型参数。

[0298] 可选地，在上述图9所对应的实施例的基础上，本申请实施例提供的图像识别模型训练装置30的另一实施例中，

[0299] 更新模块304，具体用于采用第一损失函数对第一风险函数进行变换处理，得到第一优化函数，其中，第一优化函数包括图像的预测分布向量与标签之间的损失值；

[0300] 采用第二损失函数对第二风险函数进行变换处理，得到第二优化函数，其中，第二优化函数包括已加扰图像的预测分布向量与未加扰图像的预测标签之间的损失值；

[0301] 根据第一优化函数与第二优化函数，生成目标优化函数。

[0302] 可选地，在上述图9所对应的实施例的基础上，本申请实施例提供的图像识别模型训练装置30的另一实施例中，

[0303] 处理模块303，具体用于获取图像加扰类型；

[0304] 根据图像加扰类型确定扰动邻域，其中，扰动邻域表示对未加扰图像进行图像变换的范围；

[0305] 基于扰动邻域以及第二图像所对应的第二预测标签，确定第三优化函数；

[0306] 当第三优化函数达到最大值时，获取第二图像所对应的第三图像。

[0307] 可选地，在上述图9所对应的实施例的基础上，本申请实施例提供的图像识别模型训练装置30的另一实施例中，

[0308] 处理模块303，具体用于获取图像攻击类型的数量；

[0309] 若图像攻击类型的数量等于1，则确定图像加扰类型为单攻击类型；

[0310] 若图像攻击类型的数量大于1，则确定图像加扰类型为复合攻击类型。

[0311] 可选地，在上述图9所对应的实施例的基础上，本申请实施例提供的图像识别模型训练装置30的另一实施例中，

[0312] 处理模块303，具体用于若图像加扰类型为单攻击类型，则确定图像攻击类型；

[0313] 若图像攻击类型为像素攻击类型，则获取像素变换函数对应的像素距离度量，其中，像素攻击类型为对未加扰图像中的至少一个像素值进行变换；

[0314] 根据像素距离度量以及最大像素变换范围，确定扰动邻域。

[0315] 可选地，在上述图9所对应的实施例的基础上，本申请实施例提供的图像识别模型训练装置30的另一实施例中，

[0316] 处理模块303，具体用于若图像加扰类型为单攻击类型，则确定图像攻击类型；

[0317] 若图像攻击类型为几何攻击类型，则获取几何变换函数所对应的几何距离度量，其中，几何攻击类型为对未加扰图像进行平移以及旋转中的至少一种变换；

[0318] 根据几何距离度量以及最大几何变换范围，确定扰动邻域。

[0319] 可选地，在上述图9所对应的实施例的基础上，本申请实施例提供的图像识别模型训练装置30的另一实施例中，

[0320] 处理模块303，具体用于若图像加扰类型为复合攻击类型，则获取至少两个图像攻击类型；

[0321] 若至少两个图像攻击类型包括像素攻击类型以及几何攻击类型，则获取图像攻击顺序；

[0322] 根据图像攻击顺序确定扰动邻域。

[0323] 可选地，在上述图9所对应的实施例的基础上，本申请实施例提供的图像识别模型训练装置30的另一实施例中，

[0324] 处理模块303，具体用于若图像攻击顺序为先采用像素攻击类型，再采用几何攻击类型，则获取第一复合变换函数所对应的第一复合距离度量，其中，像素攻击类型为对未加扰图像中的至少一个像素值进行变换，几何攻击类型为对未加扰图像进行平移以及旋转中的至少一种变换；

[0325] 根据第一复合距离度量以及最大几何变换范围，确定扰动邻域；

[0326] 或，

[0327] 处理模块303，具体用于若图像攻击顺序为先采用几何攻击类型，再采用像素攻击类型，则获取第二复合变换函数所对应的第二复合距离度量；

[0328] 根据第二复合距离度量以及最大像素变换范围，确定扰动邻域。

[0329] 下面对本申请中的图像识别装置进行详细描述，请参阅图10，图10为本申请实施例中图像识别装置一个实施例示意图，图像识别装置40包括：

[0330] 获取模块401，用于获取待识别图像；

[0331] 调用模块402，用于调用图像识别模型对待识别图像的类别进行预测，得到图像类别结果，其中，图像识别模型为上述实施例涉及的图像识别模型；

[0332] 发送模块403，用于向客户端发送图像类别结果，以使客户端展示图像类别结果。

[0333] 本申请实施例中，提供了一种图像识别装置，采用上述装置，由于图像识别模型采用的是基于通用对抗噪声训练下的半监督对抗防御方法，而这类方法相对于全监督对抗防御方法而言具有更好的训练效果，能够在有额外无标签数据时获得了优异的性能表现。此外，相较于现有方案中只能防御单一类型攻击的对抗防御方法，本申请训练得到的图像识别模型可以在一定程度上同时抵抗不同类型的攻击，以及不同类型攻击的复合。因此，在模型鲁棒性较好的情况下，能够在实际的图像识别过程中，输出更加准确的图像类别结果。

[0334] 本申请实施例还提供了另一种图像识别模型训练装置或图像识别装置，图像识别模型训练装置或图像识别装置可以部署于服务器，如图11所示，图11是本申请实施例提供的一种服务器结构示意图，该服务器500可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器(central processing units，CPU)522(例如，一个或一个以上处理器)和存储器532，一个或一个以上存储应用程序542或数据544的存储介质530(例如一个或一个以上海量存储设备)。其中，存储器532和存储介质530可以是短暂存储或持久存储。存储在存储介质530的程序可以包括一个或一个以上模块(图示没标出)，每个模块可以包括对服务器中的一系列指令操作。更进一步地，中央处理器522可以设置为与存储介质530通信，在服务器500上执行存储介质530中的一系列指令操作。

[0335] 服务器500还可以包括一个或一个以上电源526，一个或一个以上有线或无线网络接口550，一个或一个以上输入输出接口558，和/或，一个或一个以上操作系统541，例如

Windows Server™,Mac OS X™,Unix™,Linux™,FreeBSD™等等。

[0336] 上述实施例中由服务器所执行的步骤可以基于该图11所示的服务器结构。

[0337] 在本申请实施例中,该服务器所包括的CPU 522还具有以下功能:

[0338] 获取训练数据集合,其中,训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像;

[0339] 调用图像识别模型对训练数据集合中的第一图像的类别进行预测,得到第一图像所对应的第一预测标签;

[0340] 调用图像识别模型对训练数据集合中的第二图像的类别进行预测,得到第二图像所对应的第二预测标签;

[0341] 对第二图像进行加扰处理,得到第三图像;

[0342] 调用图像识别模型对第三图像的类别进行预测,得到第三图像所对应的第三预测标签,其中,第三图像与第二图像具有对应关系;

[0343] 根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签,对图像识别模型的第一模型参数进行更新。

[0344] 在本申请实施例中,该服务器所包括的CPU 522还具有以下功能:

[0345] 获取待识别图像;

[0346] 调用图像识别模型对待识别图像的类别进行预测,得到图像类别结果;

[0347] 向客户端发送图像类别结果,以使客户端展示图像类别结果。

[0348] 本申请实施例还提供了另一种图像识别模型训练装置或图像识别装置,图像识别模型训练装置或图像识别装置可以部署于终端设备,如图12所示,为了便于说明,仅示出了与本申请实施例相关的部分,具体技术细节未揭示的,请参照本申请实施例方法部分。该终端设备可以为包括手机、平板电脑、个人数字助理(Personal Digital Assistant,PDA)、销售终端设备(Point of Sales,POS)、车载电脑等任意终端设备,以终端设备为手机为例:

[0349] 图12示出的是与本申请实施例提供的终端设备相关的手机的部分结构的框图。参考图12,手机包括:射频(Radio Frequency,RF)电路610、存储器620、输入单元630、显示单元640、传感器650、音频电路660、无线保真(wireless fidelity,WiFi)模块670、处理器680、以及电源690等部件。本邻域技术人员可以理解,图12中示出的手机结构并不构成对手机的限定,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。

[0350] 下面结合图12对手机的各个构成部件进行具体的介绍:

[0351] RF电路610可用于收发信息或通话过程中,信号的接收和发送,特别地,将基站的下行信息接收后,给处理器680处理;另外,将设计上行的数据发送给基站。通常,RF电路610包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器(Low Noise Amplifier,LNA)、双工器等。此外,RF电路610还可以通过无线通信与网络和其他设备通信。上述无线通信可以使用任一通信标准或协议,包括但不限于全球移动通讯系统(Global System of Mobile communication,GSM)、通用分组无线服务(General Packet Radio Service,GPRS)、码分多址(Code Division Multiple Access,CDMA)、宽带码分多址(Wideband Code Division Multiple Access,WCDMA)、长期演进(Long Term Evolution,LTE)、电子邮件、短消息服务(Short Messaging Service,SMS)等。

[0352] 存储器620可用于存储软件程序以及模块，处理器680通过运行存储在存储器620的软件程序以及模块，从而执行手机的各种功能应用以及数据处理。存储器620可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等；存储数据区可存储根据手机的使用所创建的数据(比如音频数据、电话本等)等。此外，存储器620可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0353] 输入单元630可用于接收输入的数字或字符信息，以及产生与手机的用户设置以及功能控制有关的键信号输入。具体地，输入单元630可包括触控面板631以及其他输入设备632。触控面板631，也称为触摸屏，可收集用户在其上或附近的触摸操作(比如用户使用手指、触笔等任何适合的物体或附件在触控面板631上或在触控面板631附近的操作)，并根据预先设定的程式驱动相应的连接装置。可选的，触控面板631可包括触摸检测装置和触摸控制器两个部分。其中，触摸检测装置检测用户的触摸方位，并检测触摸操作带来的信号，将信号传送给触摸控制器；触摸控制器从触摸检测装置上接收触摸信息，并将它转换成触点坐标，再送给处理器680，并能接收处理器680发来的命令并加以执行。此外，可以采用电阻式、电容式、红外线以及表面声波等多种类型实现触控面板631。除了触控面板631，输入单元630还可以包括其他输入设备632。具体地，其他输入设备632可以包括但不限于物理键盘、功能键(比如音量控制按键、开关按键等)、轨迹球、鼠标、操作杆等中的一种或多种。

[0354] 显示单元640可用于显示由用户输入的信息或提供给用户的信息以及手机的各种菜单。显示单元640可包括显示面板641，可选的，可以采用液晶显示器(Liquid Crystal Display，LCD)、有机发光二极管(Organic Light-Emitting Diode，OLED)等形式来配置显示面板641。进一步的，触控面板631可覆盖显示面板641，当触控面板631检测到在其上或附近的触摸操作后，传送给处理器680以确定触摸事件的类型，随后处理器680根据触摸事件的类型在显示面板641上提供相应的视觉输出。虽然在图12中，触控面板631与显示面板641是作为两个独立的部件来实现手机的输入和输入功能，但是在某些实施例中，可以将触控面板631与显示面板641集成而实现手机的输入和输出功能。

[0355] 手机还可包括至少一种传感器650，比如光传感器、运动传感器以及其他传感器。具体地，光传感器可包括环境光传感器及接近传感器，其中，环境光传感器可根据环境光线的明暗来调节显示面板641的亮度，接近传感器可在手机移动到耳边时，关闭显示面板641和/或背光。作为运动传感器的一种，加速计传感器可检测各个方向上(一般为三轴)加速度的大小，静止时可检测出重力的大小及方向，可用于识别手机姿态的应用(比如横竖屏切换、相关游戏、磁力计姿态校准)、振动识别相关功能(比如计步器、敲击)等；至于手机还可配置的陀螺仪、气压计、湿度计、温度计、红外线传感器等其他传感器，在此不再赘述。

[0356] 音频电路660、扬声器661，传声器662可提供用户与手机之间的音频接口。音频电路660可将接收到的音频数据转换后的电信号，传输到扬声器661，由扬声器661转换为声音信号输出；另一方面，传声器662将收集的声音信号转换为电信号，由音频电路660接收后转换为音频数据，再将音频数据输出处理器680处理后，经RF电路610以发送给比如另一手机，或者将音频数据输出至存储器620以便进一步处理。

[0357] WiFi属于短距离无线传输技术，手机通过WiFi模块670可以帮助用户收发电子邮

件、浏览网页和访问流式媒体等，它为用户提供了无线的宽带互联网访问。虽然图12示出了WiFi模块670，但是可以理解的是，其并不属于手机的必须构成，完全可以根据需要在不改变发明的本质的范围内而省略。

[0358] 处理器680是手机的控制中心，利用各种接口和线路连接整个手机的各个部分，通过运行或执行存储在存储器620内的软件程序和/或模块，以及调用存储在存储器620内的数据，执行手机的各种功能和处理数据，从而对手机进行整体监控。可选的，处理器680可包括一个或多个处理单元；可选的，处理器680可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器680中。

[0359] 手机还包括给各个部件供电的电源690(比如电池)，可选的，电源可以通过电源管理系统与处理器680逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。

[0360] 尽管未示出，手机还可以包括摄像头、蓝牙模块等，在此不再赘述。

[0361] 在本申请实施例中，该终端设备所包括的处理器680还具有以下功能：

[0362] 获取训练数据集合，其中，训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像；

[0363] 调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一图像所对应的第一预测标签；

[0364] 调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二图像所对应的第二预测标签；

[0365] 对第二图像进行加扰处理，得到第三图像；

[0366] 调用图像识别模型对第三图像的类别进行预测，得到第三图像所对应的第三预测标签，其中，第三图像与第二图像具有对应关系；

[0367] 根据第一图像所对应的标签、第一图像所对应的第一预测标签、第二图像所对应的第二预测标签以及第三图像所对应的第三预测标签，对图像识别模型的第一模型参数进行更新。

[0368] 在本申请实施例中，该终端设备所包括的处理器680还具有以下功能：

[0369] 获取待识别图像；

[0370] 调用图像识别模型对待识别图像的类别进行预测，得到图像类别结果；

[0371] 向客户端发送图像类别结果，以使客户端展示图像类别结果。

[0372] 本申请实施例中还提供一种计算机可读存储介质，该计算机可读存储介质中存储有计算机程序，当其在计算机上运行时，使得计算机执行如前述图4所示实施例描述的方法中图像识别模型训练装置所执行的步骤，或者，使得计算机执行如前述图5所示实施例描述的方法中图像识别模型所执行的步骤。

[0373] 本申请实施例中还提供一种包括程序的计算机程序产品，当其在计算机上运行时，使得计算机执行如前述图4所示实施例描述的方法中图像识别模型训练装置所执行的步骤，或者，使得计算机执行如前述图5所示实施例描述的方法中图像识别模型所执行的步骤。

[0374] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，

装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0375] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统,装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0376] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0377] 另外,在本申请各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。

[0378] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或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中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(read-only memory,ROM)、随机存取存储器(random access memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0379] 以上所述,以上实施例仅用以说明本申请的技术方案,而非对其限制;尽管参照前述实施例对本申请进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

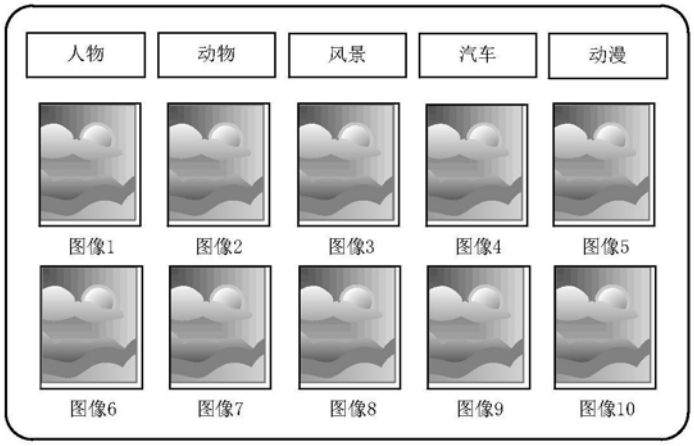

人物
动物
风景
汽车
动漫
图像1
图像2
图像3
图像4
图像5
图像6
图像7
图像8
图像9
图像10

图1

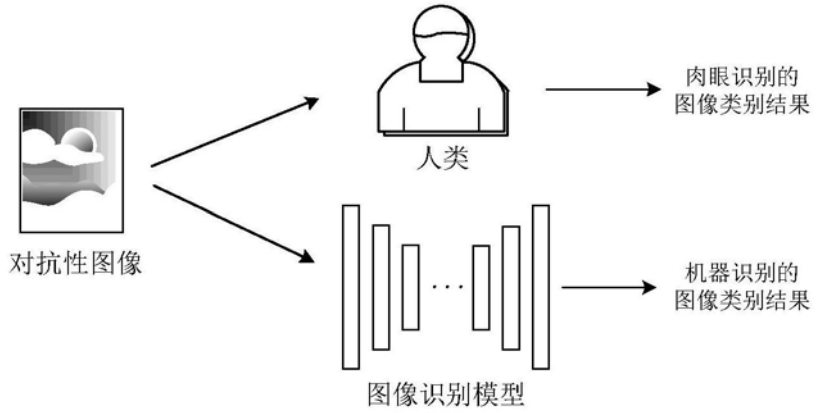

对抗性图像
人类
肉眼识别的
图像类别结果
图像识别模型
机器识别的
图像类别结果

图2

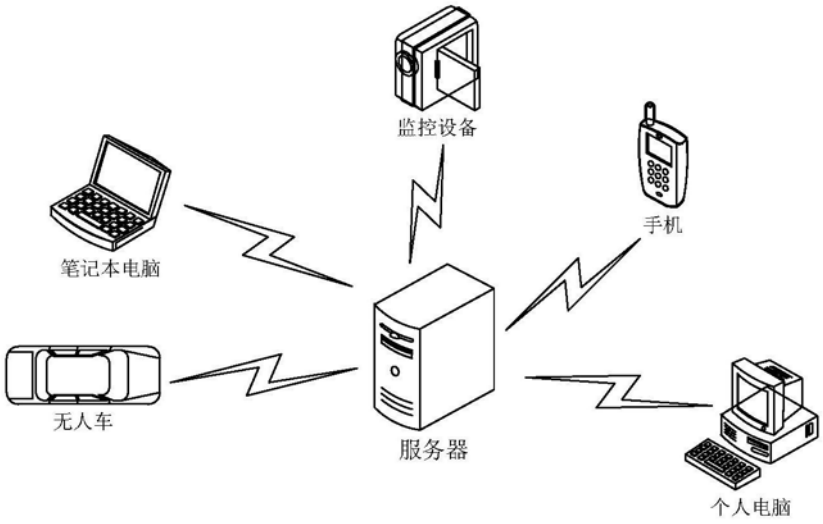

监控设备
手机
笔记本电脑
无人车
服务器
个人电脑

图3

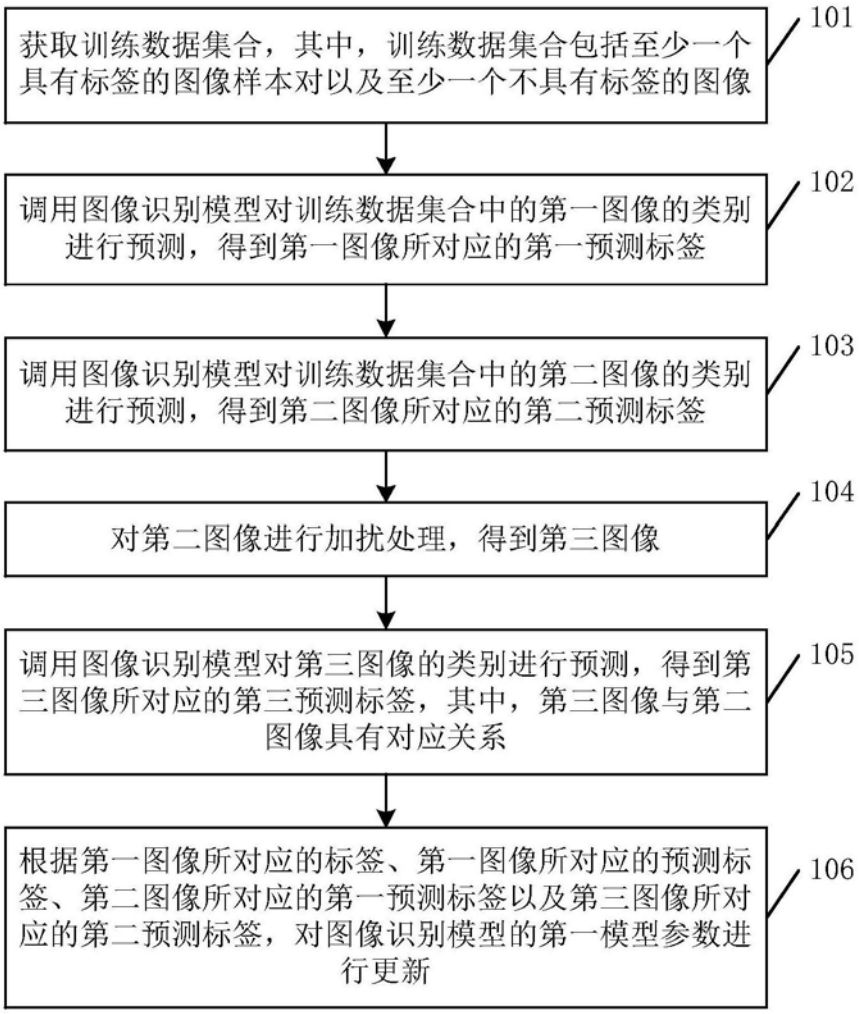

获取训练数据集合，其中，训练数据集合包括至少一个具有标签的图像样本对以及至少一个不具有标签的图像
101
调用图像识别模型对训练数据集合中的第一图像的类别进行预测，得到第一图像所对应的第一预测标签
102
调用图像识别模型对训练数据集合中的第二图像的类别进行预测，得到第二图像所对应的第二预测标签
103
对第二图像进行加扰处理，得到第三图像
104
调用图像识别模型对第三图像的类别进行预测，得到第三图像所对应的第三预测标签，其中，第三图像与第二图像具有对应关系
105
根据第一图像所对应的标签、第一图像所对应的预测标签、第二图像所对应的第一预测标签以及第三图像所对应的第二预测标签，对图像识别模型的第一模型参数进行更新
106

图4

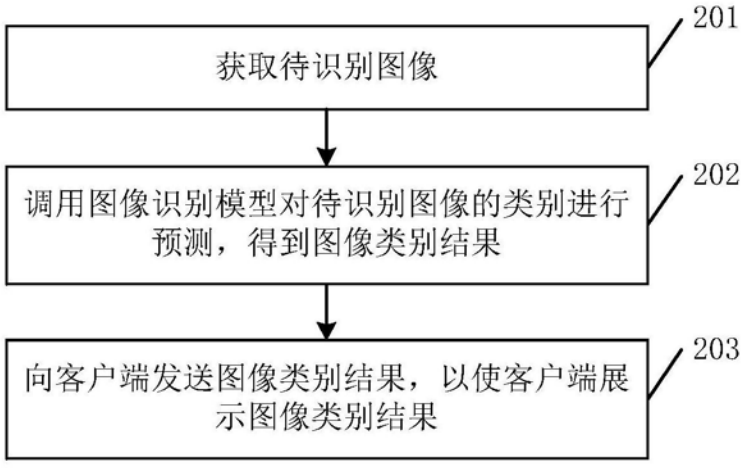

获取待识别图像
201
调用图像识别模型对待识别图像的类别进行预测，得到图像类别结果
202
向客户端发送图像类别结果，以使客户端展示图像类别结果
203

图5

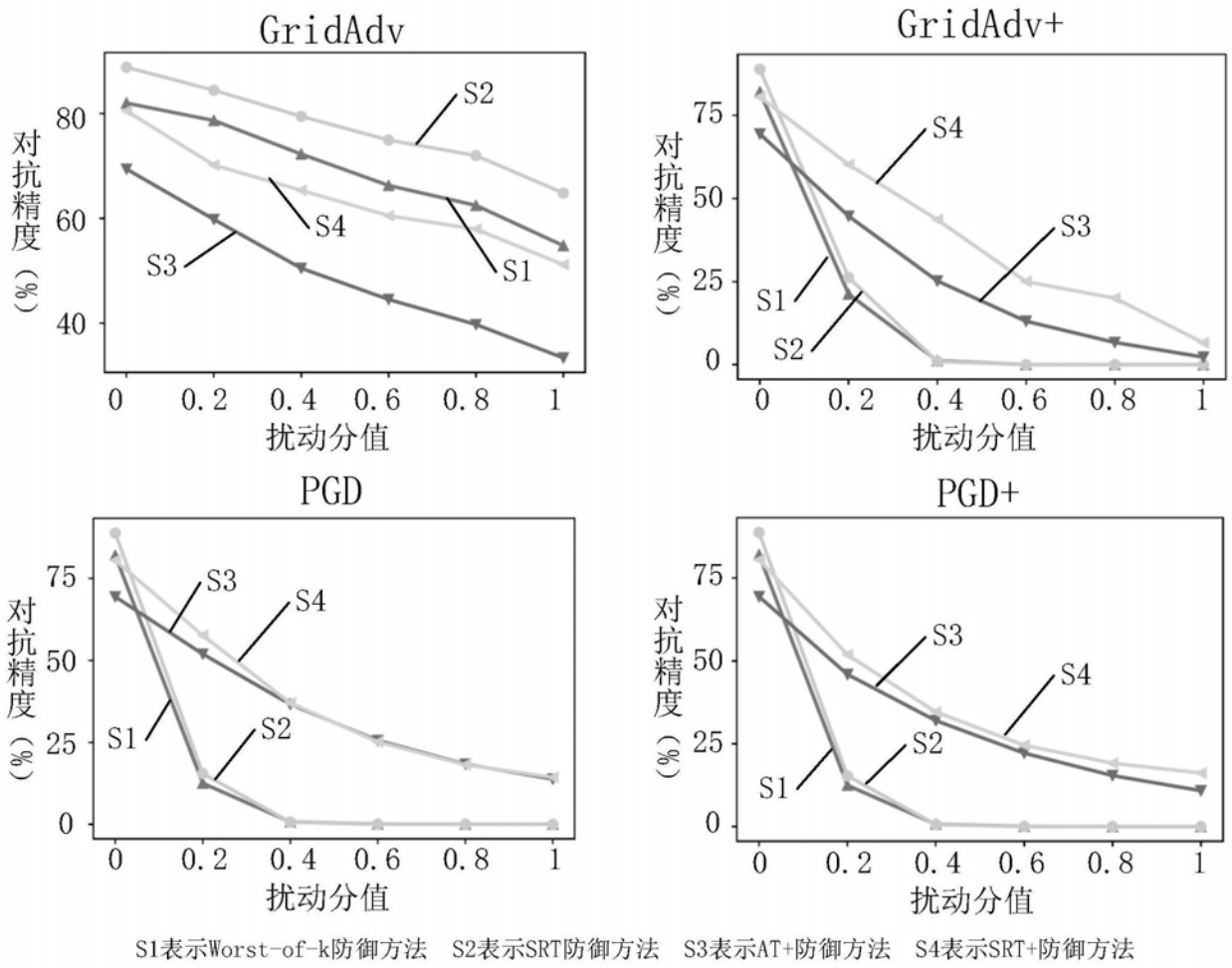

GridAdv
GridAdv+
PGD
PGD+
对抗精度（%）
扰动分值
S1
S2
S3
S4
S1表示Worst-of-k防御方法 S2表示SRT防御方法 S3表示AT+防御方法 S4表示SRT+防御方法

图6A

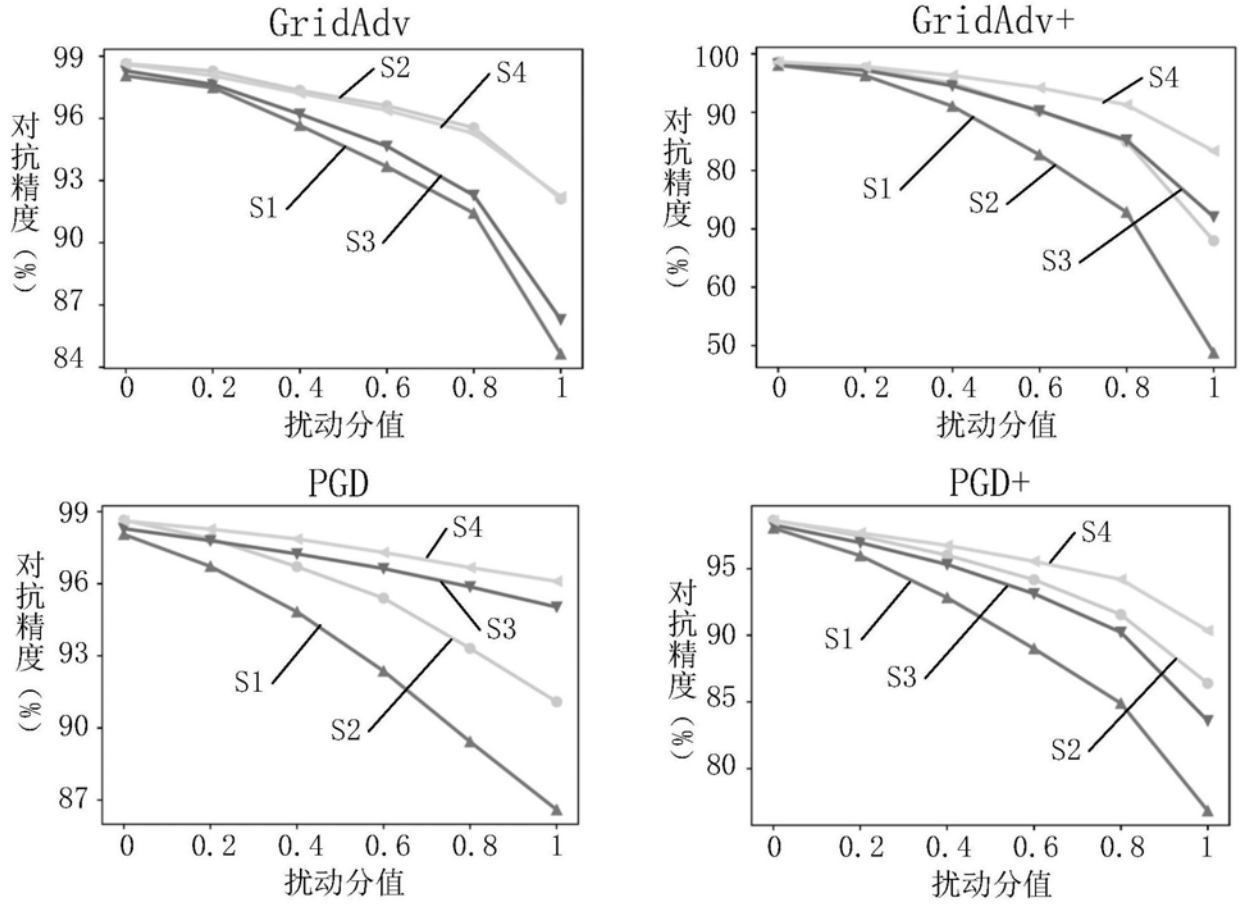

GridAdv
GridAdv+
PGD
PGD+
对抗精度（%）
扰动分值
S1
S2
S3
S4
S1表示Worst-of-k防御方法 S2表示SRT防御方法 S3表示AT+防御方法 S4表示SRT+防御方法

图6B

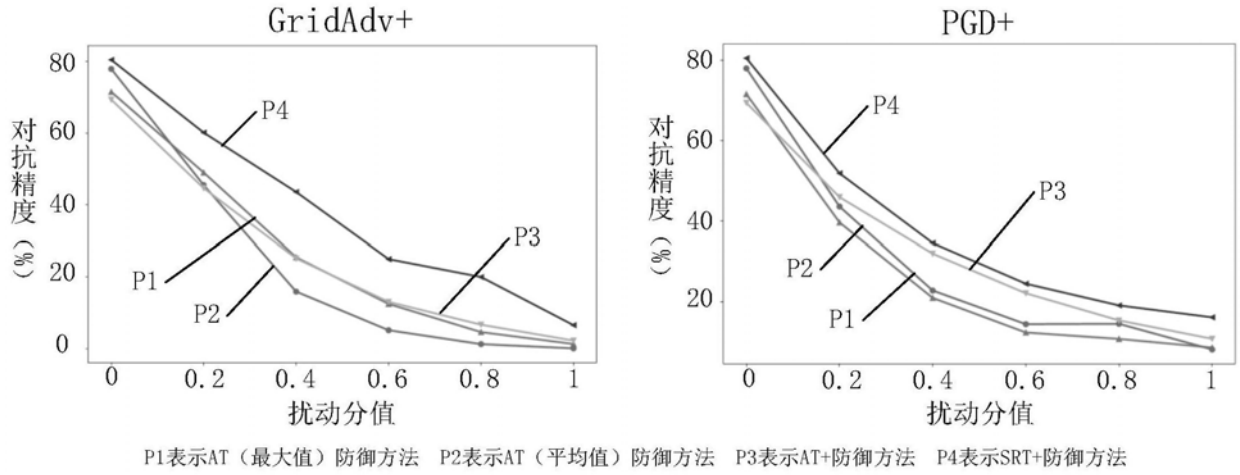
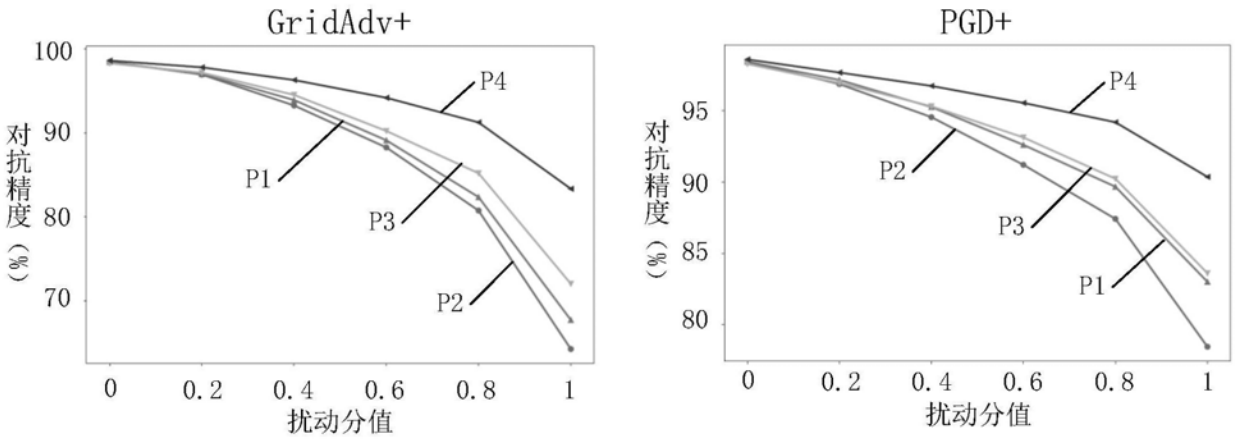
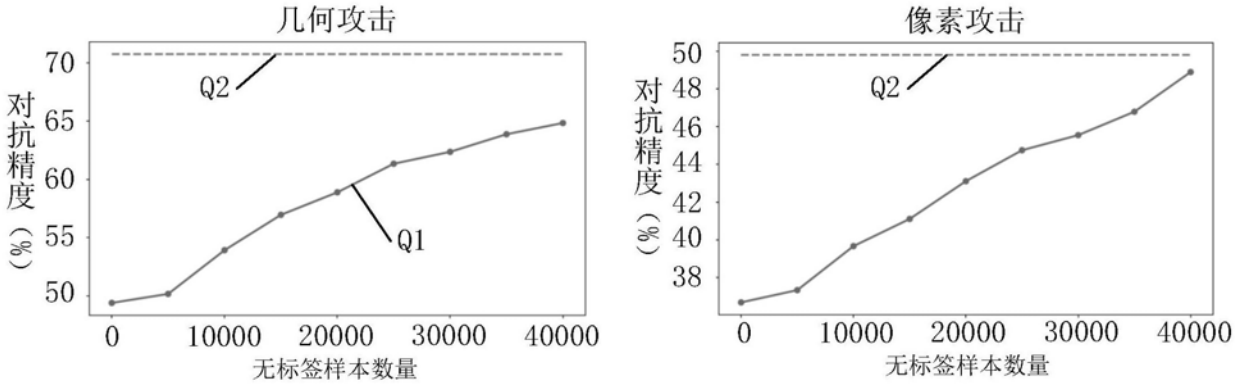

GridAdv+
PGD+
对抗精度（%）
扰动分值
P1
P2
P3
P4
P1表示AT（最大值）防御方法 P2表示AT（平均值）防御方法 P3表示AT+防御方法 P4表示SRT+防御方法

图7A



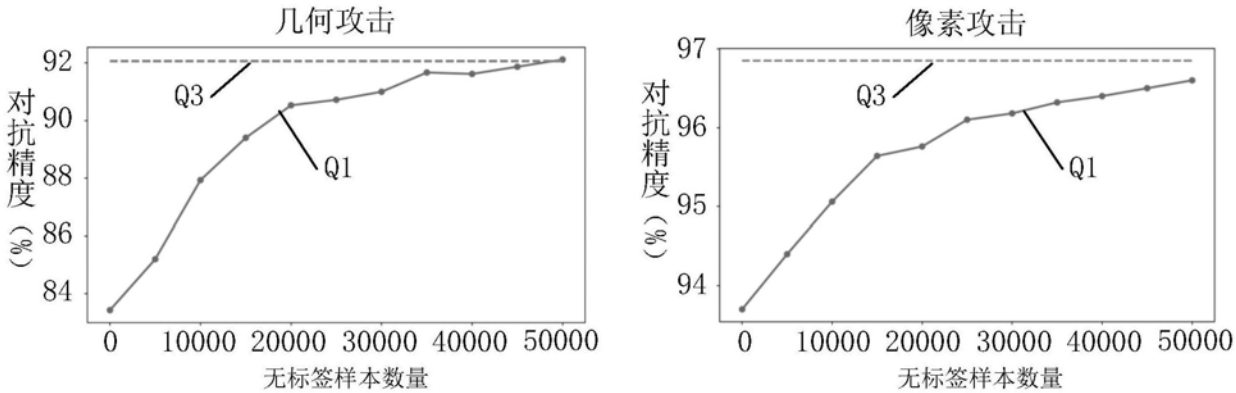
P1表示AT（最大值）防御方法 P2表示AT（平均值）防御方法 P3表示AT+防御方法 P4表示SRT+防御方法

图7B



Q1表示SRT防御方法 Q2表示采用50000个无标签样本的Worst-of-k或AT防御方法

图8A



Q1表示SRT防御方法 Q3表示采用60000个无标签样本的Worst-of-k或AT防御方法

图8B

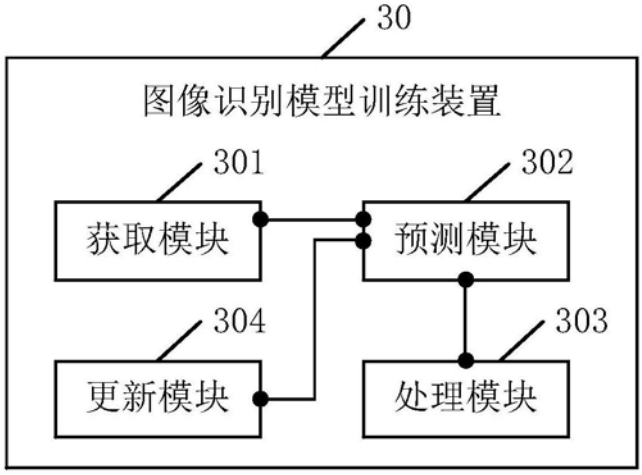

30
图像识别模型训练装置
301
获取模块
302
预测模块
304
更新模块
303
处理模块

图9

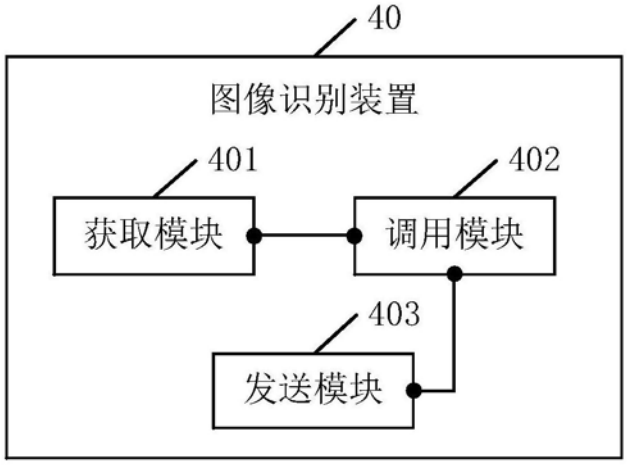

40
图像识别装置
401
获取模块
402
调用模块
403
发送模块

图10

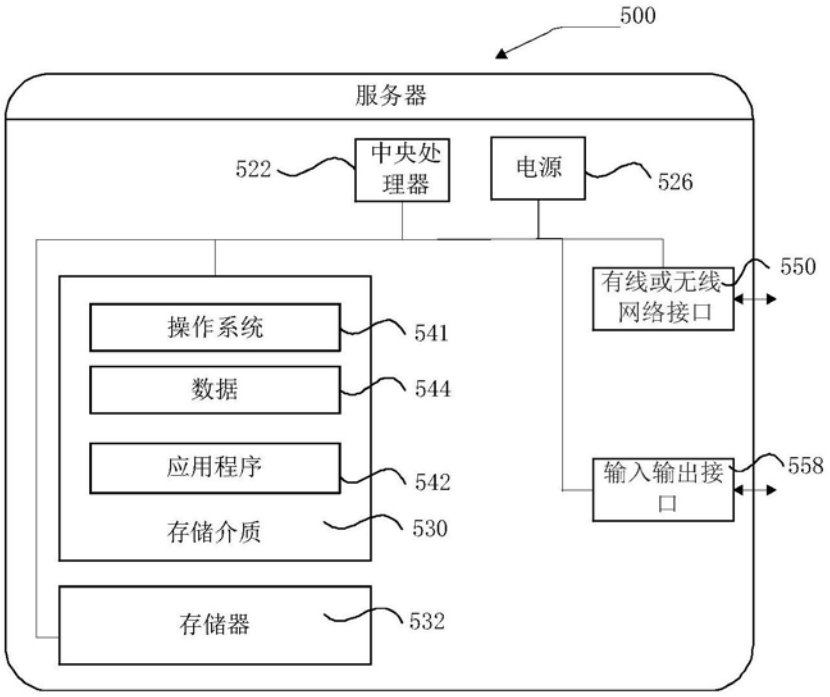

500
服务器
522
中央处理器
电源
526
有线或无线网络接口
550
操作系统
541
数据
544
应用程序
542
存储介质
530
输入输出接口
558
存储器
532

图11

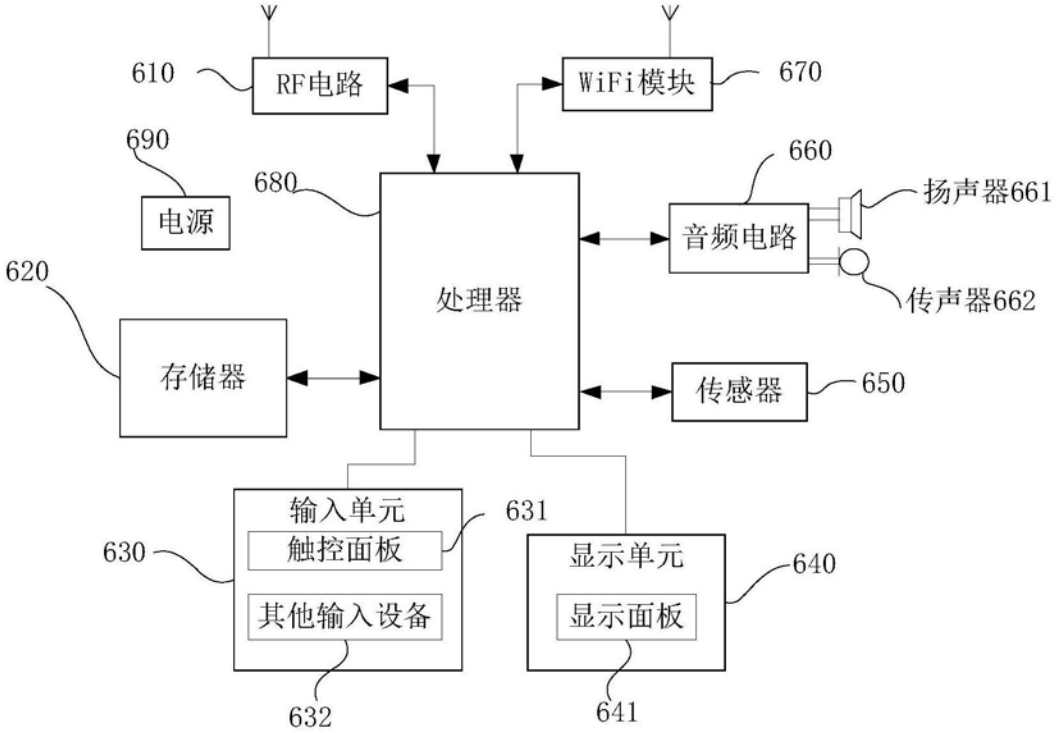

610
RF电路
WiFi模块
670
690
电源
680
660
扬声器661
音频电路
620
处理器
传声器662
存储器
传感器
650
输入单元
631
630
触控面板
其他输入设备
显示单元
640
显示面板
632
641

图12