(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115860112 A
(43)申请公布日 2023.03.28

(21)申请号 202310059601.9

(22)申请日 2023.01.17

(71)申请人 武汉大学
地址 430072 湖北省武汉市武昌区八一路299号

(72)发明人 田博为 曹雨欣 王骞 龚雪鸾 沈超 李琦

(74)专利代理机构 武汉科皓知识产权代理事务所(特殊普通合伙) 42222
专利代理师 齐晨涵

(51)Int.Cl.
G06N 3/094(2023.01)
G06N 3/0464(2023.01)
G06N 3/0455(2023.01)
G06N 3/0475(2023.01)
G06N 3/084(2023.01)
G06N 3/092(2023.01)
G06F 18/22(2023.01)
G06F 18/241(2023.01)

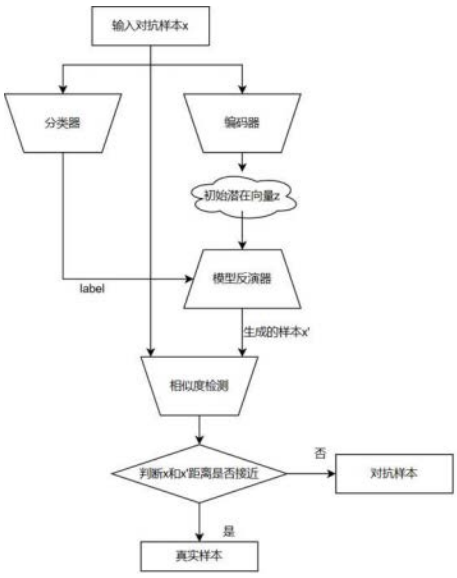
权利要求书3页 说明书7页 附图3页

(54)发明名称

基于模型反演方法的对抗样本防御方法和设备

(57)摘要

本发明公开的基于模型反演方法的对抗样本防御方法和设备，为了解决深度神经网络安全领域中缺少低成本、高效的对抗样本防御方法的问题，提出了一种基于StyleGAN生成器的模型反演机制实现对抗样本防御方法。通过对生成器StyleGAN的深入分析，提出强化信息训练和改进的proAdaIN，将其创新性地应用到对抗样本防御系统的特征生成方案中，并通过添加噪声、特征解耦、利用冲突语义区分真实样本和对抗样本，解决了传统防御方案中成本高、效率低、防御效果差等问题。



CN 115860112 A

1.一种模型反演方法,其特征在于:该方法基于改进的StyleGAN网络,具体包括以下步骤:

步骤1,将待反演的潜在空间向量进行预处理和空间变换,以获得解耦向量;

步骤2,验证变换后的向量特征分布解耦性;

步骤3,将解耦后的向量,迁移到初始向量处进行强化信息训练;同时经过不同的仿射变换之后,进行proAdaIN处理;

强化信息训练具体过程为:如果没有公有数据集,从一个默认的图片开始;否则结合需要进行分类的物品的公有数据集,首先输入到与输入向量相同的分类器里,进行分类;然后,我们将分类结果,与输入向量进行交叉熵损失的计算,取最小损失的的数据,做多层感知MLP操作之后,作为输出;

ProAdaIN其操作定义为下列公式:

$$proAdaIN(t_i,y) = y_{s,i} \cdot max\{0, erfc(\frac{t_i - \mu(t_i)}{\sigma(t_i)\cdot\sqrt{2}})\} + y_{b,i}$$

其中$erfc$函数是高斯误差函数,y_s和y_b是解耦后的向量通过映射得到的不同样式,下标i是样本的索引,μ是样本的均值,σ是样本的标准差;每个特征向量t_i,分别被归一化,然后使用来自该样式的y进行缩放和偏置,因此,y的维数是该层的特征映射数的两倍;

步骤4,引入噪声输入。

2.根据权利要求1所述的模型反演方法,其特征在于:

所述步骤1的具体过程为:

将传统潜在向量空间Z映射到解耦后的潜在向量空间W,给定在潜在向量空间Z中的潜在向量z,一个非线性的映射网络f:Z→W,产生$w \in W$,映射f使用8层多层感知器实现。

3.根据权利要求1所述的模型反演方法,其特征在于:

所述步骤2中采用混合正则化,在产生解耦后的潜在向量空间W时,并行地采用两个给定的潜在向量z_1,z_2同时处理,并且将他们的特征w_1,w_2通过参数进行不同程度的混合,最终分别生成所需的特征空间。

4.根据权利要求3所述的模型反演方法,其特征在于:

步骤2中利用感知路径长度测量图像在潜在空间中进行插值时的变化验证解耦是否成功,感知路径长度l_z是将两潜码之间的插值路径细分为小段,定义为每个小段感知差异的总和,由下式表示:

$$l_z = E[\frac{1}{\varepsilon^2} d(G(slerp(z_1,z_2;t)), G(slerp(z_1,z_2;t+\varepsilon)))]$$

其中,潜在向量z_1,z_2,z_1,z_2~P(z),P(z)代表Z空间向量的概率分布,G是模型反演器,d(•,•)评估了目标图片的感知距离,E表示期望,ε表示步进,t表示从z_1到z_2向量过程中的某个中间向量,t~U(0,1),U(0,1)代表0-1的均匀分布,$slerp$代表球形插值;为了将注意力集中到核心的特征而不是细节和背景,对图片进行裁剪,计算多个样本,并且求其期望,使用类似的方法,对空间变换后的向量空间W进行计算:

$$l_w = E[\frac{1}{\varepsilon^2} d(g(lerp(f(z_1), f(z_2); t)), g(lerp(f(z_1), f(z_2); t + \varepsilon)))]$$

其中，g是模型反演器的综合网络部分，$lerp$ 代表线性插值，f为预处理中空间变换所用函数；根据相应的实验结果可以验证将Z空间映射到W空间之后，目标图片的感知距离明显降低，说明成功将原图片的各个特征进行了解耦。

5.根据权利要求1所述的模型反演方法，其特征在于：

所述步骤3中的交叉熵损失的公式如下：

$$L = -\sum_{c=1}^{M} y_{ic} \log(p_{ic})$$

其中，M代表了一共有多少个类别，y_{ic}代表符号函数，也就是该样本的真实向量取值；p_{ic}代表预测值，也就是潜在向量空间Z中的值。

6.根据权利要求1所述模型反演方法，其特征在于：

所述步骤4中引入高斯噪声，在综合网络的每一层都提供该高斯噪声的输入，使用经过训练的特征标度因子来广播到所有的特征映射中。

7.一种基于模型反演方法的对抗样本防御方法，其特征在于：

设计对抗样本防御模型，利用监督学习的方式，采用梯度下降的方法对该模型进行训练；所述抗样本防御模型包括分类器、编码器、模型反演器和相似度鉴别器，其中模型反演器为根据权利要求1-6中任一项所述模型反演方法所生成的模型反演器；

所述分类器的目的是向模型反演器提供将要反演出来的目标标签；

所述编码器用于提取输入图像的潜在空间向量；

所述模型反演器将编码器分类出的特征向量进行重构，输出一个具有语义信息的图片；

所述相似度鉴别器用以确定模型反演器生成图像是否足够接近于相应的输入图像。

8.根据权利要求7所述的对抗样本防御方法，其特征在于：

所述相似度鉴别器采用如下公式：

$$x_{pos} = x - x'_y$$

$$x_{neg} = x - x_{y'}{}'$$

误差x_{pos}是输入图像x与其在一致标签y下的合成图像x_y'之间的差值，负误差x_{neg}，即输入图像x减去模型反演器根据实际标签y'产生的结果$x_{y'}'$；

使用铰链损失函数作为最终的损失评估结果，即正输入的值在损失中的项大于1，同时负输入的值在损失中的项小于-1：

$$L_{D_{aux}} = ReLU(1 - D_{aux}(x_{pos}, y)) + ReLU\left(1 + D_{aux}(x_{neg}, y)\right)$$

相似性鉴别器D_{aux}，相似性鉴别器损失函数L_{Daux}，根据该损失函数，可以区分x与x'之间的相似性，如果x与x'被认为相似，则x是良性样本；否则，x会被认为是对抗样本。

9.一种电子设备，其特征在于，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序；

当一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求7-8中任一项所述方法所执行的操作。

10.一种计算机可读介质，其上存储有计算机程序，其特征在于：所述程序被处理器执行时实现如权利要求7-8中任一项所述方法所执行的操作。

基于模型反演方法的对抗样本防御方法和设备

技术领域

[0001] 本发明属于人工智能安全领域，主要涉及一种基于模型反演方法的对抗样本防御方法和设备。

背景技术

[0002] 目前，深度神经网络在各个领域都取得了巨大的成功，在各种关键任务应用中也得到了广泛的部署，其安全性和可信性已成为公众关注的问题，如自动驾驶、医疗诊断和可信计算等领域。然而，在这些应用中，错误的决策或者预测可能会导致灾难性的经济损失甚至生命危险。

[0003] 对抗样本是攻击者经常使用的一种注入样本。它是当前DNN面临的主要威胁之一，它会对输入引入微妙的恶意扰动，以便于欺骗DNN模型。这种对抗样本相对于原始样本扰动很小，通常人类无法察觉，却能够极大地改变了目标DNN模型提取的特征，导致了错误的推理结果。根据攻击者的知识不同，对抗攻击可以分为黑盒攻击、白盒攻击和自适应攻击三类。如今，几乎所有的防御方案都遭受了自适应攻击的严重影响，攻击者可以利用防御方案的知识制作新的对抗样本。

[0004] 潜在空间：在一个神经网络中，有一些变量是可以观测到的，而与之相对的，有一些变量是不可观测的，一般是神经网络中的中间层、隐藏层的变量，或者是在不同的神经网络之间传输的、不需要使用者获取具体数值的变量。这些变量的分布组成的就是潜在空间。

[0005] 模型反演是一种深度学习的应用方向，目的是想办法从分类器的分类向量导出原始的训练数据的特征，从而造成严重的安全威胁。其中目前较为新颖的方法是基于GAN的模型反演。GAN由两个模块组成：生成模型G和判别模型D，提供一个对抗性的训练方法，即：生成模型的目标是尽量生成真实的图片去欺骗判别模型，而判别模型的目标是尽量将生成模型生成的图片和真实的图片区分。最终效果是，可以根据一个简单的输入向量，这种向量可以是一种随机噪声z，生成更高维的输出向量G(z)，如声音、图像等。该技术已经在超分辨率任务和语义分割等方向得到了广泛的应用。

[0006] 在已有的防御研究中，主要有通过输入进行转换、反向训练模型，或者基于特定的标准检测异常的方法，来减轻对抗样本的影响。例如，梯度掩盖方法，用难以被攻击者使用的方法构建鲁棒的模型，然而这种方法很容易被具有梯度逼近能力的攻击绕过；还有对抗训练方法，通过在训练阶段加入对抗样本来提高模型的鲁棒性。其他方法包括改变损失函数、改变激活函数、集成学习、自监督学习、利用人工生成的训练样本或重新加权错误分类的样本等方法。其中，对抗性训练、改变函数等方法虽然易于实现，但是不可避免地降低了合法输入的精度。

发明内容

[0007] 本发明针对现有技术的不足，提出了一种基于模型反演方法的对抗样本防御方法和设备，提高了合法输入的精度。

[0008] 本发明所设计的模型反演方法，该方法基于StyleGAN，具体包括以下过程：

步骤 1，将待反演的潜在空间向量进行预处理和空间变换，以获得解耦向量。

[0009] 由于原始样本的部分特征可以通过一个超参数得到更好的变换，也便于控制，所以对样本进行了预处理，即将初始向量解耦，获取解耦之后的向量。

[0010] 步骤2，通过感知路径长度，验证w向量特征的分布成功解耦。

[0011] 验证方法是通过感知路径长度等指标实现的，感知路径长度是将两潜码之间的插值路径细分为小段，定义为每个小段感知差异的总和。也就是说，在一个特征图之中，一条线段上的特征渐变越连续，那么这个线段上的两个端点的感知路径长度就越短。可以根据这个指标来验证样本特征的分布是否变得更加的清晰，也就是是否成功解耦。根据实验结果，表明采用8次MLP变换之后，两个端点之间的感知路径长度的确得到了有效的降低。

[0012] 步骤3，将解耦后的向量，迁移到初始向量处进行强化信息训练；同时经过训练后的仿射变换A之后，进行proAdaIN处理。

[0013] 将潜在向量空间W做经过训练的仿射变换，将解耦后的向量映射到不同的样式$y=(y_s,y_b)$上，来控制每个综合网络g的卷积层网络之后的改进图片风格迁移操作(ProfessionalAdaptive Instance Normalization, proAdaIN)，这一操作，简单来说就是通过不同的仿射变换，得到各种样式，应用到最终生成的图片之中。

[0014] 强化信息训练的意思是指，不仅仅使用一个固定的、经过精心构造的初始化向量，而是尽量根据公有数据集的信息：如果没有公有数据集，从一个默认的图片开始；否则结合需要进行分类的物品的公有数据集，首先输入到与输入向量相同的分类器里，进行分类；然后，将分类结果，与输入向量进行交叉熵损失的计算。取最小损失的的数据，做多层感知MLP操作之后，作为输入后续综合网络的向量。

[0015] 值得注意的是，proAdaIN这一操作相对于普通的AdaIN函数有很大的突破，这主要是因为普通的AdaIN函数仅仅实现了数据的归一化和仿射变化，而没有实现数据的高斯分布处理。在加入了高斯分布处理之后的数据在面对更为复杂的实现场景时，相对于AdaIN函数的鲁棒性得到了增强。可以实现对于不仅仅是人脸，还有物品和场景的反演。

[0016] 步骤4，引入噪声输入。

[0017] 这些是由独立的高斯噪声组成的单通道向量，这是因为只要整张图片遵循正确的分布，微小细节的特征可以被随机化，而并不影响对于图像的感知。同时，噪声的生成还有助于在训练的时候降低过拟合现象。另外还有较多的方法来应对过拟合的问题，如正则化、增多样本数量等等。

[0018] 本发明还设计了一种基于模型反演方法的对抗样本防御方法，将模型反演方法应用到利用语义冲突防御对抗样本的模型，并进行训练。

[0019] 需要训练的模型主要包括了分类器、编码器、模型反演器和相似度鉴别器。主要利用监督学习的方式，采用梯度下降的方法进行训练。

[0020] 所述分类器的目的是向模型反演器提供将要反演出来的目标标签；

所述编码器用于提取输入图像的潜在空间向量；

所述模型反演器将编码器分类出的特征向量进行重构，输出一个具有语义信息的图片；

所述相似度鉴别器用以确定模型反演器生成图像是否足够接近于相应的输入图

像。

[0021] 基于同一发明构思,本发明还设计了一个或多个处理器;

存储装置,用于存储一个或多个程序;

当一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现基于模型反演方法的对抗样本防御方法所执行的操作。

[0022] 基于同一发明构思,本发明还设计了一种计算机可读介质,其上存储有计算机程序,其特征在于:所述程序被处理器执行时实现基于模型反演方法的对抗样本防御方法所执行的操作。

[0023] 本发明的优点在于:

本发明将模型反演机制引入对抗样本防御,解决了原本的对抗样本防御机制安全性和效率性不可兼得、对正常输入影响的问题;

本发明在原有的模型反演机制上进行了改进,包括强化信息训练和proADaIN等技术,分别解决了原有模型反演技术在应用中的输入数据类型限制、数据分布的限制。

附图说明

[0024] 图1是本发明实施例中模型反演器的实现图。

[0025] 图2是模型反演器运用在对抗样本防御的总流程图。

[0026] 图3是本发明的具体步骤流程图。

具体实施方式

[0027] 本发明主要将模型反演机制引入对抗样本防御。模型反演器的实现方法图见附图1。本方法充分考虑了有目标攻击情况和无目标攻击情况两类。考虑对防御者来说具有挑战性的白盒攻击。通过本发明进行防御得到的深度学习模型更加安全可靠,对于各种先进的对抗样本攻击方法都具有较好的防御效果,且在防御过程中的成本开销与时间开销均较小。

[0028] 本发明提供的方法能够用计算机软件技术实现流程。附图1为模型反演器构造的主要流程,以两层综合网络为例,可以扩展;附图3为整个防御方案的总流程。下面将以检测带有对抗样本数据集的实验为例,对本发明的流程进行一个具体的阐述。

[0029] 本发明所设计的一种模型反演方法,该方法基于改进的StyleGAN网络,具体包括以下步骤:

步骤1,图片预处理,并且将传统潜在向量空间Z映射到解耦后的潜在向量空间W。

[0030] 图片预处理,即将原始的GAN的Z空间解耦为W空间。解耦有许多不同的定义,但是共同的特点是一个潜在空间中存在着多个线性子空间,每个线性子空间都包含着一个明显特征。在原始的Z空间中,各个子空间是纠缠在一起的,然而通过多次的连续映射,由于这种映射可以适应“解翘曲”的情况,使得这些线性子空间变换为更加清晰和独立的不同区域,而不是纠缠在一起。因此,预计经过映射后的W空间在经过这样的无监督训练后能够产生一个解耦的分布。当然,目前的研究有其他的办法实现纠缠分布的解耦,本方案中经验性地使用了其中的一种。

[0031] 为了更好地实现风格的独特性,采用了混合的正则化,而不是单一图片串行地训

练。具体地说，在产生W空间向量时，并行地采用两个潜在向量z_1，z_2同时处理，并且将他们的特征w_1，w_2通过参数进行不同程度的混合，最终分别生成所需的特征向量空间。这种正则化技术可以使得生成的图片的灵活性提高。其中混合通过$w=\theta_1 w_1+\theta_2 w_2$，其中$\theta_1$和$\theta_2$是可以调整的超参数，调整会取得不同的效果。

[0032] 步骤2，并且利用感知路径长度测量验证W空间的解耦性。

[0033] 潜在向量空间Z的插值会产生令人惊讶的非线性变化，例如，在任一端点处，缺失的特征可能出现在线性插值路径的中间。这就是Z空间被纠缠和变量因素不恰当地分离的现象，为了量化这一效果，本发明利用感知路径长度测量了图像在潜在空间中进行插值时的剧烈变化。简单地说，相对于高度扭曲的潜在空间，解耦的潜在空间会导致明显的平稳过渡。

[0034] 感知路径长度是将两潜码之间的插值路径细分为小段，定义为每个小段感知差异的总和，由下式表示：

$$l_z = E[\frac{1}{\varepsilon^2} d(G(slerp(z_1,z_2;t)),G(slerp(z_1,z_2;t+\varepsilon)))]$$

其中，潜在向量z_1，z_2，z_1，z_2~P(z)，P(z)代表Z空间向量的概率分布，G是模型反演器，d(•，•)评估了目标图片的感知距离，E表示期望，ε表示步进，t表示从z_1到z_2向量过程中的某个中间向量，t~U(0,1)，U(0,1)代表0-1的均匀分布，$slerp$代表球形插值，也是在正则化输入潜在空间中最合适的插值方法。为了将注意力集中到核心的特征而不是细节和背景，对图片进行了裁剪。计算多个样本，并且求其期望。使用类似的方法，对W隐空间进行计算：

$$l_w = E[\frac{1}{\varepsilon^2} d(g(lerp(f(z_1),f(z_2);t)),g(lerp(f(z_1),f(z_2);t+\varepsilon)))]$$

其中，g是模型反演器的综合网络部分，$lerp$代表线性插值，f为预处理中空间变换所用函数；根据相应的实验结果可以验证将Z空间映射到W空间之后，目标图片的感知距离明显降低，说明已成功将原图片的各个特征进行了解耦。实验结果见下一部分。

[0035] 步骤3，将处理后的图片对应的向量，迁移到初始向量处进行强化信息训练；同时经过不同的仿射变换A之后，进行proAdaIN处理。

[0036] 强化信息训练的意思是指，不仅仅使用一个固定的、经过精心构造的初始化向量。如果只是一个固定的向量，那么在应用的时候可移植性将变得很差。为了让本发明的模型能够适应更多物体的识别，建议寻找到一个类似分类任务的公有数据集，分以下两种情况讨论：如果没有公有数据集，则一个默认的图片开始；否则同时结合需要进行分类的物品的公有数据集，首先输入到与输入向量相同的分类器里，进行分类；然后，将分类结果，与输入向量进行交叉熵损失的计算。取最小损失的的数据作为初始向量。交叉熵损失的公式如下：

$$L = -\sum_{c=1}^{M} y_{ic} \log(p_{ic})$$

其中，M代表了一共有多少个类别，y_{ic}代表符号函数，也就是该样本的真实向量取值；p_{ic}代表预测值，也就是Z空间中的值。

[0037] 将潜在向量空间W做经过训练的仿射变换，将解耦后的向量映射到不同的样式y=

(y_s,y_b)上,来控制每个综合网络g的卷积层网络之后的图片改进的风格迁移映射操作(ProfessionalAdaptive Instance Normalization, proAdaIN)。将proAdaIN操作定义为下列公式:

$$proAdaIN(t_i,y) = y_{s,i} \cdot max\{0, erfc(\frac{t_i - \mu(t_i)}{\sigma(t_i) \cdot \sqrt{2}})\} + y_{b,i}$$

其中$erfc$函数是高斯误差函数,该函数可以较好地使整个数据分布趋近于正态分布,从而提升模型的稳定性。y_s和y_b是解耦后的向量通过映射得到的不同样式,下标i是样本的索引,μ是样本的均值,σ是样本的标准差;每个特征向量t_i,分别被归一化,然后使用来自该样式的y进行缩放和偏置,因此,y的维数是该层的特征映射数的两倍。该方法可以非常好地提升反演的效果。

[0038] 步骤4,通过引入显式的噪声输入,为生成的随机细节提供了一种直接的方法。

[0039] 这些是由独立的高斯噪声组成的单通道向量,需要注意的是,在综合网络的每一层都提供这样一个专用的噪声输入,使用经过训练的特征标度因子来广播到所有的特征映射中。

[0040] 图像中有许多方面可以被随机地进行变换,例如人像的毛发、胡子,还有窗帘图片的花纹,马路上沥青的纹路,等等。只要整张图片遵循正确的分布,微小细节的特征可以被随机化,而并不影响对于图像的感知。同时,噪声的生成还有助于在训练的时候降低过拟合现象。

[0041] 过拟合现象是指,为了得到某种一致的假设,而使得假设变得过度严格。从神经网络发展的历史来看,随着网络结构愈加复杂、网络层数逐渐增多,过拟合问题也渐渐成为影响训练效果的一个主要原因。过拟合的原因主要是在训练过程中机器注意力偏移到人们不希望机器注意到的地方,也就是细枝末节处。例如在,某个深度神经网络中,由于训练数据过少或网络过神深等原因,网络误以为某标志的背景颜色,如天空的蓝色是该标志的实际含义-限速80km/h。通过引入高斯噪声,可以较好地消除该现象。同时,过拟合现象还可以通过正则化、增加训练样本数等方式得到解决。

[0042] 将构造好的模型反演器接入检测对抗样本的系统的总体设计中。即一种基于模型反演方法的对抗样本防御方法,具体过程为:

设计对抗样本防御模型,如附图2所示,它包括了分类器、编码器、模型反演器和相似度鉴别器。在模型训练的过程中,主要通过反向传播的方式来不断修正网络参数,最后使网络性能达到预期的效果。所谓的反向传播,即根据计算当前参数环境下损失函数的梯度值,然后通过梯度下降的方式修正参数,梯度下降公式如下:

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

其中,θ_j为待修正的参数,而α则代表了学习率,学习率的值很小且为正值,从0到1不等。它控制每次迭代的步长,并朝着损失函数的最小值移动。显然,学习率决定了参数调整的速度。当学习率设置的过小时,收敛过程将变得十分缓慢或极容易陷入局部最优解中;而当学习率设置的过大时,梯度可能会在最小值附近来回震荡,甚至可能无法收敛。因此,

选择一个合适的学习率，对于模型的训练将至关重要。学习率随着训练目标的不同而不同，需要对学习率进行调整，一般会选择0.001，0.01或0.1。

[0043] 附图2记录了的整个对抗样本检测系统的静态结构，下面说明整个对抗样本防御系统是如何动态运行的。

[0044] 整个对抗样本防御系统由四个部分组成，它们是分类器，编码器，模型反演器和一个相似性鉴别器。具体地，分类器的目的是向模型反演器提供将要反演出来的目标标签；编码器负责提取输入图像x的潜在特征，作为模型反演器的默认初始向量。模型反演器最终输出一个具有语义信息的图片。一个经过良好训练的模型反演器生成的图像与分类器给出的标签高度相似。

[0045] 同时，相似性鉴别器能够度量输入图像x和其反演图像x'之间的相似性。将模型反演器生成的样本x'和对抗样本x同时输入相似度检测之中。这样做的前提是，给定一个标签，由模型反演器生成的图像x'应当类似于该标签的正常图像x；而会与该标签的对抗图像的相似度非常差。所以通过这个相似性鉴别器D_{aux}，以确定生成图像x'是否足够接近于相应的输入x，判断样本x是否为对抗样本。

[0046] 在计算差异的方法上，相似性鉴别器是这样计算生成图像和原始图像的差异的：

首先，直接使用$x-x'$作为直接输入，然后，定义正误差x_{pos}，也就是x与其在一致标签y下的合成图像x_y'之间的差值，这种差值是图片本身产生的一般性误差。再定义正误差x_{neg}，也就是x减去模型反演器根据实际标签y'产生的结果$x_{y'}'$。公式如下：

$$x_{pos} = x - x_y'$$

$$x_{neg} = x - x_{y'}'$$

使用铰链损失作为最终的损失评估结果，这会使得正输入的值在损失中的项大于1，同时负输入的值在损失中的项小于-1。

[0047] $$L_{D_{aux}} = ReLU(1 - D_{aux}(x_{pos}, y)) + ReLU\left(1 + D_{aux}(x_{neg}, y)\right)$$

函数ReLU，根据该损失函数，可以区分x与x'之间的相似性。如果x与x'被认为相似，则x是良性样本；否则，x会被认为是对抗样本。

[0048] 整个对抗样本防御系统是这样运行的，同时可以参考图2：

1.生成过程，投入未知样本，由分类器预测出该样本的标签，并且由编码器输出该样本的潜在向量z。同时将标签和潜在向量输入模型反演器中。

[0049] 2.推理过程，同时提供目标DNN模型的输入和输出，在模型反演器中运行，以获得生成结果。对于正确推断的合法输入，合成输出试图重建输入。对于对抗样本，由于模型反演器的目的是根据特征向量生成图片，模型反演器将尽可能创建符合错误标签的合成结果，而不是重建输入。

[0050] 3.相似度检测过程，相似性检测将检测生成结果和对抗样本的相似度，对真实样本，距离很小，通过相似度检测；对对抗样本，距离较大，无法通过相似度检测。这样可以区分两种样本，达到目的。

[0051] 实验具体内容如下：

通过参考相关文献和计算机软件技术实现，验证了模型反演器从Z空间到W空间，纠缠分布的解耦作用的合理性和有效性，同时验证了该方案能够很好地解决对抗样本攻击的防御问题。

[0052] 本次实验使用了两项实验指标，分别感知路径长度和分离度，两者都是极小型指标。

[0053] 最终得到的实验数据如下表所示：

方法	感知路径长度	分离度
传统模型反演器 Z	412.0	10.78
本发明的模型反演器 W	426.5	3.52
+增加噪音输入的W	193.7	3.54
+Mixing 50%的W	226.7	3.52
+Mixing 90%的W	240.5	3.76

实验数据表明，感知路径长度在添加了噪声之后，大幅度降低，这说明W空间更整洁清晰；而如果仅仅测量路径端点，也就是端感知路径长度，或者从分离度的角度，也会得到相同的结果：模型反演器中的W空间较原始Z空间更加解耦。

[0054] 在对抗样本检测方面，采用白盒攻击方法，在MNIST数据集上进行防御测试，分别采用PGD_{12}、BIM_{12}的攻击方法，本发明的防御均可以达到100%的效果。这说明该防御对抗样本的功能得到了实现。

[0055] 本文中所描述的具体实施例仅仅是对本发明作举例说明。本发明所属技术领域的技术人员可以对所描述的具体实施例做各种各样的修改或补充或采用类似的方式替代，但并不会偏离本发明的精神或者超越所附权利要求书所定义的范围。

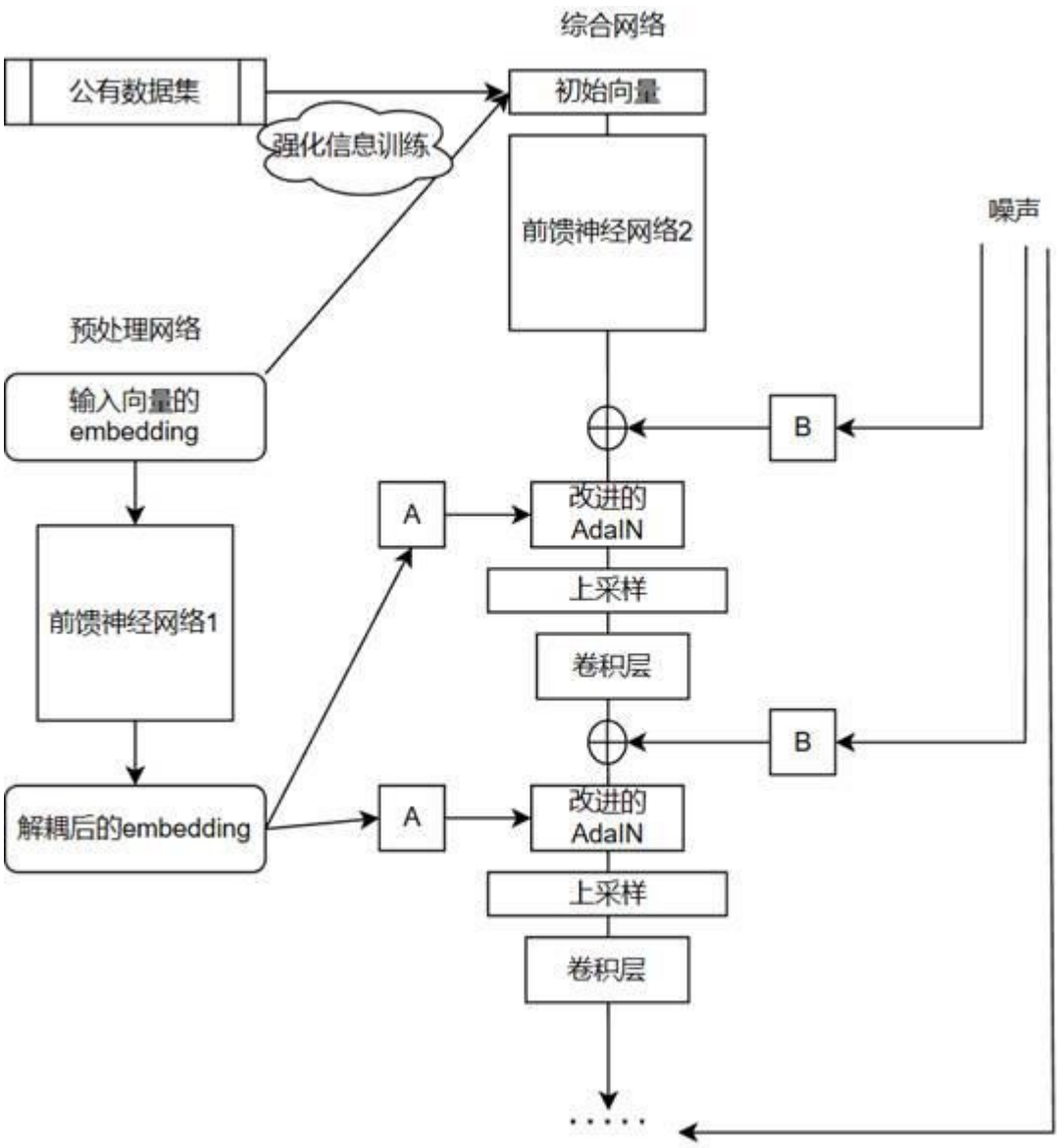

综合网络
公有数据集
强化信息训练
初始向量
前馈神经网络2
噪声
预处理网络
输入向量的
embedding
B
A
改进的
AdaIN
上采样
卷积层
前馈神经网络1
B
解耦后的embedding
A
改进的
AdaIN
上采样
卷积层

图1

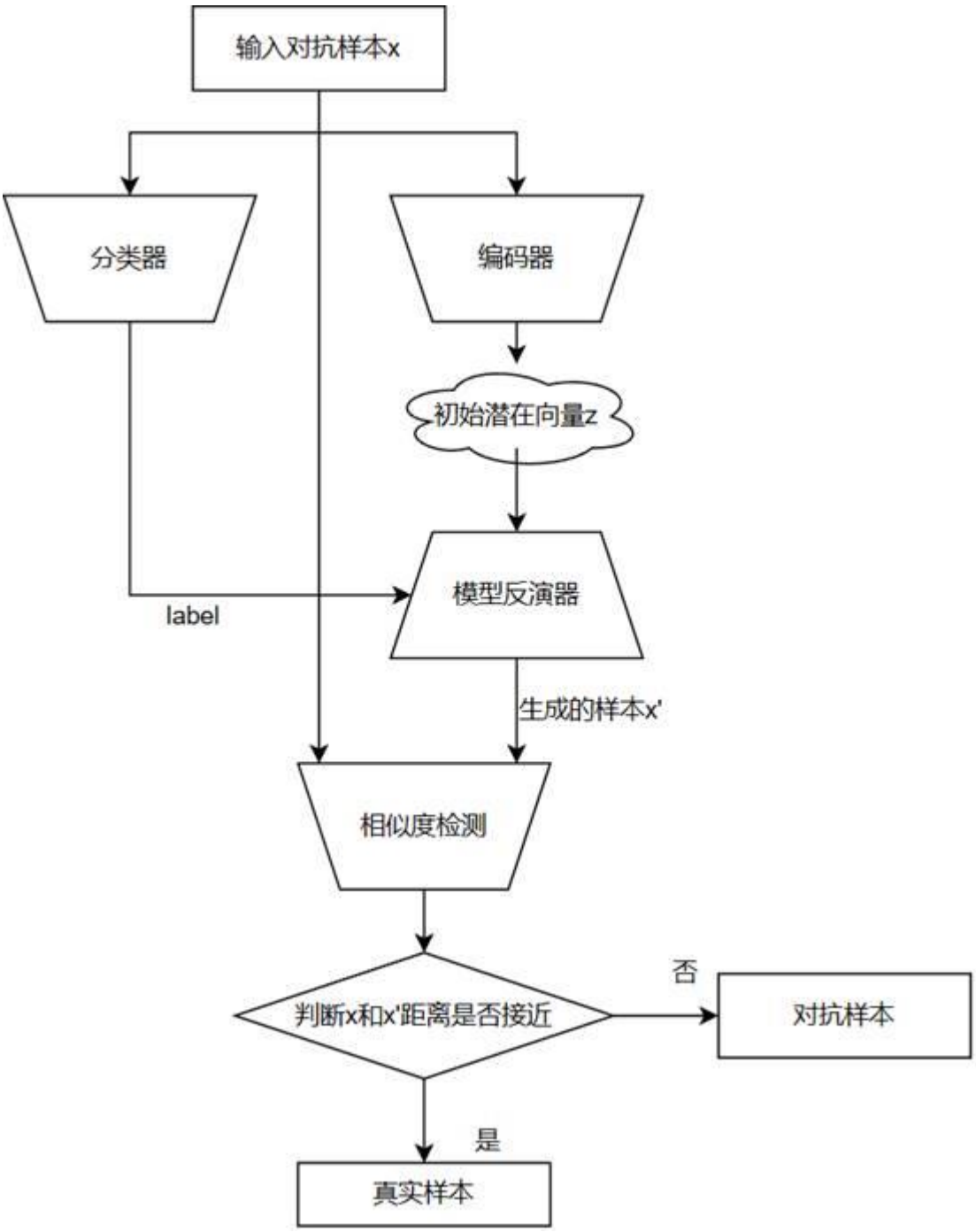

输入对抗样本x
分类器
编码器
初始潜在向量z
label
模型反演器
生成的样本x'
相似度检测
判断x和x'距离是否接近
否
对抗样本
是
真实样本

图2

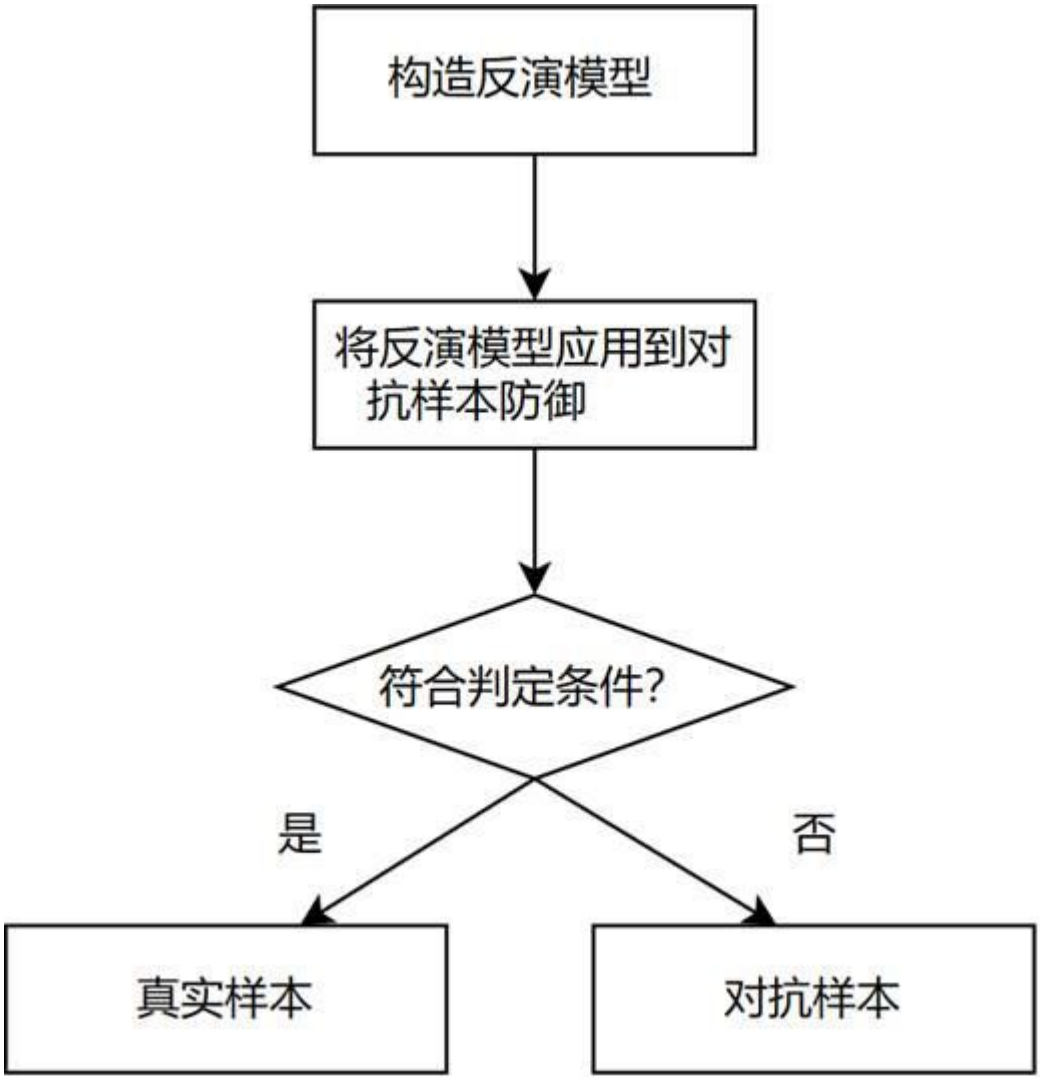


图3