# 绪论

## 1.1研究背景及意义

近年来深度学习在许多传统机器学习难以解决的问题上有着优异的表现，尤其在计算机视觉领域，如：人脸识别，自动驾驶，视频检测等。尽管深度学习在计算机视觉领域表现优异，但Christian Szegedy等人发现现有的深度神经网络（Deep Neural Networks，简称DNN）易受对抗样本的攻击，这些对抗样本只是在原有的图像样本添加轻微的扰动（人类视觉无法察觉），可以导致深度神经网络输出错误的分类结果。如图1所示，模型将原始图片识别为猫，在原始图像中加入轻微的对抗样本扰动后，模型将图像识别为鸵鸟。Nguyen 等人还发现一些人类无法识别的样本，深度神经网络模型依旧能够以高置信度对该样本进行分类。除此之外对抗样本的攻击已对现有的深度学习应用领域产生严重威胁，如对人脸识别系统的攻击，攻击者只需在人脸图像上添加经过训练的扰动，就可让人脸识别系统将该图像识别为指定的用户。Kdnuggets指出对抗样本不仅仅存在于深度学习相关领域，也普遍存在于许多机器学习模型中，因此进一步研究对抗样本不仅有利于提高目前已有深度学习解决方案的安全性，还有利于整个机器学习及深度学习领域的发展。

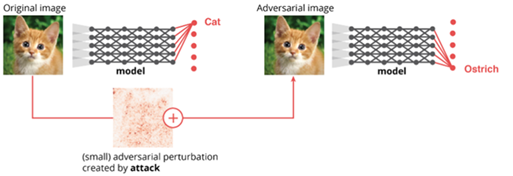


图1 对抗样本攻击举例

目前现实中深度学习解决方案越来越多，对抗样本的存在极大的威胁到这些解决方案。因此在设计深度学习解决方案的过程中，必须考虑到如何防御对抗样本的攻击。现有的防御策略大多通过减轻深度学习模型的过拟合程度来防御对抗样本的攻击，如在训练数据集中添加大量对抗样本。但在减轻深度学习模型的过拟合过程中，大概率会降低模型对纯净图像（无扰动图像）分类准确率。并且需要重新训练整个模型，极大的提高了深度学习模型训练成本。同时对于全新的对抗样本防御效果较差。到目前为止还没有一种高效能防御所有对抗样本攻击的防御方法。并且随着防御策略的升级，攻击的手段也在逐步提高，对抗样本扰动从开始简单针对特定图像的扰动发展到可以添加到任意图像的通用扰动，扰动大小也变的越来越小。

因此目前急需一种在保证深度学习模型对原始图像分类精度和训练成本基本不变的条件下，同时能防御不同种类对抗样本防御策略。针对上述情况，本论文提出基于VAE-GAN的对抗样本防御方法，在保证图像质量基本不变的条件下，对对抗样本进行去噪处理。本方法不仅能有效降低深度学习模型对对抗样的误识别，而且相较于大多数其他防御策略训练成本也较低。

## 1.2研究现状

### 1.2.1生成对抗样本研究现状

2014年Szegedy[2]等人首次提出对样本概念，证明了可以通过对图像添加小量的人类察觉不到的扰动误导神经网络做出误分类。他们尝试求解让深度学习模型做出误分类的最小扰动方程。由于求解该最小扰动程算法复杂度过高，他们将问题转化为寻找最小损失函数添加项方程，成功将问题转化成凸优化过程，并提出L-BFGS优化算法求解该问题。同时他们还提出对抗样本在不同模型中具有迁移性（对抗样本可以导致不同深度学习模型分类错误）。2015年Goodfellow 等人提出线性假设证明了对抗样本的存在性，并提出了一种快速生产对抗样本的方法-快速梯度符号法（FastGradient Sign Method，FGSM）。同时他们提出利用该攻击方法进行对抗训练，可以提高深度学习模型的健壮性。2016年Carlini和 Wagner提出三种攻击方法，分别限制扰动，和范数使得扰动无法被深度学习模型察觉，同时该方法生成扰动可以在不同模型之间进行迁移，进而实现了黑盒攻击。2017年Kurakin 等人在原有的FGSM的基础上进行改进，将原有的单步生成对抗样本改进为迭代生成的方式，提出了(Basic Iterative Method，BIM)、最小可能类法 (Leastlikely Class Method，LCM)、迭代最小可能类法 (Iterative least-likely class method，ILCM) 等对抗样本生成方法，同时他们还发现被相机拍摄的对抗样本也会被深度学习模型错误分类，进而证明物理世界也存在对抗样本。2017年Madry 等人提出投影梯度法（Projected Gradient Descent，PGD）生成对抗样本，该方法在BIM方法上将初始攻击参数进行随机初始化，极大的提高了对抗样本的攻击性。2017 年，清华大学 TSAIL 团队 Dong 等人提出了动量迭代法 (Momentum Iterative Method，MIM)，该方法有着极高的黑盒攻击成功率。2018年AJ Bose等人提出一种基于生成对抗网络（Generative Adversarial Networks，GAN）攻击人脸识别系统的方法。

目前生成对抗样本方法的研究较为成熟，对抗样本的攻击从原有的简单针对于单张图像无目标白盒攻击，逐渐发展到目前的可迁移有目标黑盒攻击。对抗样本的扰动也变的越来越小。虽然现有的生成对抗样本的方法种类繁多，但大多方法是对FGSM和BIM的改进，又或者采用其他优化方法求解最小损失函数添加项方程。因此在研究如何防御对抗样本时，只需选取其中较有代表性的攻击方法即可。目前攻击效果较好的生成对抗样本方法有：BIM，PGD及Carlini and Wagner Attacks (简称C&W)。

### 1.2.2 对抗样本防御研究现状

目前针对对抗样本防御主要有三个方向：对输入图像进行预处理或修改模型训练过程，改进深度学习网络模型，辨别图像是否为对抗样本。本文根据以上三个方向简单介绍一些较有代表性的防御策略。

1. 对输入图像进行预处理或修改模型训练过程

蛮力训练：许多研究者发现在深度学习训练数据集中不断加入不同方法生成的对抗样本，可以不断的提高深度学习网络模型的健壮性。该方法需要大量的对抗样本，虽然在一定程度上可以正则化深度神经网络减轻模型过拟合，但Moosavi-Dezfooli指出无论添加多少对抗样本，都存在新的对抗样本能够欺骗网络。

数据压缩：2016年Dziugaite等人发现使用JPG对输入图像进行压缩可以减轻扰动对于模型的影响，但图像压缩同时会降低正常分类准确率。即使后来提出用PCA方法进行压缩，依旧无法解决该问题。

图像缩放：2017年Xie等人发现将输入图像进行随机缩放可以有效的降低对抗样本攻击强度。

图像重建：2019年jia等人发现对图像进行压缩重建能够有效的降低对抗样本的攻击强度。相比图像压缩，该方法对正常分类准确率影响较小。

1. 改进深度学习网络模型

深度压缩网络：研究者发现简单的将去噪自动编码器（Denoising Auto Encoders，DAE）堆叠到原有的网络中可以减轻对抗样本的攻击，但会使网络模型变得更加脆弱。因此Gu等人提出深度压缩网络（Deep Contractive Networks，DCN），该方法使用了类似于压缩自动编码器（Contractive Auto Encoders，CAE）平滑度惩罚项。

梯度正则化或梯度隐藏：研究者发现使用输入梯度正则化可以有效的提高深度学习网络模型的健壮性，并且该方法配合蛮力对抗训练可以取得非常好的防御效果，缺点是网络计算复杂度太高。

知识蒸馏：2016年Hinton提出将复杂网络的知识迁移到简单网络中，可以有效的抵抗小幅度对抗样本攻击。该防御方法被Carlini和 Wagner提出的攻击方法攻破。

Parseval 网络：2017年Cisse等人提出通过控制网络每一层的Lipschitz 常数，可以有效的减轻对抗样本的攻击。

1. 辨别图像是否为对抗样本

2017年Issaranon等人提出激活函数Relu对于普通图像和对抗样本输出模式不同，并提供了一种基于SVM区分图像是否为对抗样本的方法。2018年Hinton提出使用胶囊网络重构图像可以辨别图像是否为对抗样本。

虽然目前已有的对抗样本防御方法众多，但大多是减轻对抗样本的影响。随着生成对抗样本方法研究的不断深入，对抗样本的扰动变得越来越小，同时对抗攻击目标从原来的小数据集转变成自然数据集，这极大的增加我们设计对抗样本防御策略的难度。同时在设计对抗样本过程中不仅仅要考虑如何提高对抗样本分类准确率，还要考虑普通正常样本分类准确率和防御的训练成本。

## 1.3研究内容及章节安排

# 对抗样本攻防理论基础

### 2.1对抗样本的定义

2014年Szegedy等人首次提出对抗样本概念，他们发现深度神经网络很容易受到对抗样本的攻击。对抗样本只是在原始样本中添加轻微扰动，人眼甚至无法察觉这些扰动，就能使得分类器输出错误的结果。同时他们还对对抗样本有如下定义：

假设将图像映射为一组向量的深度神经网络分类器为：



其损失函数可以表示为：



对于一个样本并且，则对抗样本生成可以表示为一个有界优化问题：



由于优化公式（2-3）算法复杂度太高不易计算，并且条件可以转化为，因此他们将该问题进行优化，优化后的公式为：



对于公式（2-4）他们提出用L-BFGS优化算法去求解该问题。在求解对抗样本的同时，他们还发现对抗样本对于低纬度的隐藏层输出影响较小，对于高纬度隐藏层输出影响较大。最后他们对于对抗样本存在性问题，提出了一个可能的解释（如图2.1所示）：现实世界出现对抗样本的概率很低，因此训练集和测试集中几乎不存在对抗样本或者说训练集无法覆盖所有的样本，又因为深度神经网络的高度非线性，导致过拟合只学习到非对抗样本的特征，并没有学习到泛化性特征。

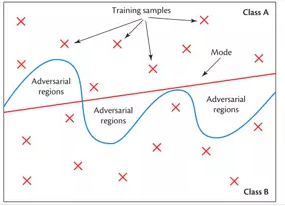


图2.1 深度神经网络模型训练与使用

如图2.1所示，蓝线为样本的真实决策边界，红线为模型训练得到的决策边界。可以从图中发现，模型训练后的最终决策边界和真实样本决策边界有较大的差别，这表明对抗样本是广泛存在的。

### 2.2生成对抗样本方法

在对抗样本概念被提出之后，越来多研究人员致力于改进生成对抗样本的方法。目前对抗样本攻击主要分为两类：一是白盒攻击，攻击者能够获取深度神经网络结构及参数，另一类是黑盒攻击，攻击者完全不知道深度神经网络网络结构及参数，只能通过模型输入获取输出。其中黑盒攻击可以通过对抗样本的迁移性来实现（用一个已知网络模型生产对抗样本，然后攻击未知模型），或者直接根据定义（只通过模型的输入输出）进行优化生成对抗样本，比如使用GAN生成对抗样本：生成器产生对抗样本扰动，未知深度神经网络分类器模型作为辨别器。虽然黑盒攻击和白盒攻击从攻击的目的来说有很大的不同，但其生成对抗样本的本质并没有区别。与黑盒攻击相比，白盒攻击攻击效果要更好，并且目前的黑盒攻击大多是白盒攻击的迁移或者融合多种白盒攻击。除此之外，对抗样本攻击是否具有目标性分为有目标攻击和无目标攻击。有目标攻击是指对抗样本能够使分类器以高置信度输出指定分类结果，无目标攻击只需使分类器输出错误结果即可。

由于生成对抗样本的方法众多，并且还在不断改进。因此本文选取目前较有代表性的三种生成对抗样本方法，简单介绍生成对抗样本的原理。

### 2.2.1快速梯度法-FGSM

2015年Goodfellow等人对生成对样本原因进行分析，他们发现扰动对深度神经网络影响会像雪球一样随着网络深度的增加越滚越大，他们称这种现象为扰动累加效应。由于神经网络的高度线性特性，导致这种效应相比非线性模型更加严重。基于深度神经网络高度线性及扰动累加效应，他们提出了一种基于梯度的快速生成对抗样本方法FGSM，FGSM中对抗样本扰动计算公式如下：



其中表示模型输入，表示结果标签， 表示损失函数，表示符号函数，表示控制扰动大小的的自定义参数，表示生成的对抗样本扰动。FGSM算法的主要思想:如果扰动的方向和梯度方向一致，损失函数就会增大，进而使得分类器分类结果产生最大的变化。扰动越大损失函数就会越大，分类器结果产生的变化就会越大。

该方法可以通过一步快速生成对抗样本，并且该方法未直接使用梯度值而选择梯度方向有效的控制了对抗样本扰动的大小。虽然该方法生成速度快，使用的计算资源少，但使用该方法有诸多的限制。首先该方法适合白盒攻击，生成的对抗样本迁移性较差。其次该方法对于线性分类器模型攻击效果较好，对于非线性分类器模型和不可微分类器模型攻击效果较差。最后使用FGSM生成的对抗样本扰动较大（相当于每个像素点都增加了），该扰动自身抗干扰能力不强，容易受到其他噪声影响。

### 2.2.2基础迭代法-BIM

针对于FGSM存在的两个问题：损失函数与输入图像之间近似线性和扰动相较于原始图像改动较大。Goodfellow等人提出进一步的假设，如果损失函数与输入图像之间是非线性的，是否会存在一个存在于之间的某个扰动，使得变得更大。基于上述新的假设，Goodfellow等人提出了一种以迭代的方式生成对抗样本的方法BIM，生成对抗样本公式如下：



其中截断函数为：



截断函数的目的是保证的每一个像素都在的临域内且每个像素都有意义。BIM每次迭代的含义是：每次迭代在上一步生成的对抗样本基础上每个像素增加（也可能减少），然后对新生成的对抗样本进行裁剪，保证新的对抗样本每个像素都在的各像素临域内。BIM生成的对抗样本攻击效果一般要优于FGSM生成的对抗样本，最差的情况生成的对抗样本攻击效果与FGSM相同。

BIM生成的对抗样本攻击效果虽然优于FGSM，但由于其生成对抗样本迭代过程固定且每次迭代都是对上一次生成的对抗样本进行裁剪，这导致其能够迭代的次数较少，同时生成的对抗样本也容易受到其他噪声的影响。

### 2.2.3投影梯度法-PGD

2017年Aleksander等人发现采用PGD方式求解对抗样本中的优化问题效果要好于Goodfellow他们提出的使用梯度下降方式。PGD算法与梯度下降算法比较类似，都是用于求解非线性规划问题，但PGD算法在求解带有约束的非线性规划问题时具有更好的效果。PGD在求解对抗样本优化问题时，先选取一个可行解，然后将梯度下降的方向投影到约束集上，然后连接可行解点和梯度下降点生成可行的下降方向，最后重复上述过程直到对抗样本满足设定要求的精度。PGD生成对抗样本算法如下：



其中为一个学习步长，表示投影算子，该算子能将当前可行解投影到约束集中向量长度最小的向量上。PGD相较于BIM有更多的迭代轮次，并且往对抗样本中添加了随机化点，使得PGD的攻击效果远好于BIM。PGD生成的对抗样本也是目前一阶损失函数攻击攻击效果最好的算法。

### 2.3对抗样本防御策略

针对以上的对抗样本攻击特点，目前的防御手段主要有三个方向：对输入图像进行预处理，改进神经网络模型，检测输入是否为对抗样本。

对图像进行预处理：因为对抗样本是在原有的样本上添加而外的扰动噪声，一般情况下噪声越大攻击效果越好，同时因为目前深度神经网络高度线性特点扰动对神经网络影响主要是靠累加效应，因此只要减少对抗样本中的扰动就可以减轻对抗样本攻击效果。目前对图像的预处理主要是通过图像降噪，图像压缩，图像重建。图像预处理一定程度能减少样本中的对抗扰动，但是代价是会降低整个模型的分类精度。

改进神经网络模型： Szegedy等人提出对抗本概念时也指出对抗样本会导致神经网络高层输出远高于普通样本。根据以上特点Goodfellow等人提出使用Lipschitz常数来控制神经网络的层与层输出比例，进而防止高层神经元输出爆炸。Goodfellow等人在提出FGSM时也提出使用对抗训练能够很好的提高深度神经网络模型的健壮性。此外还可以通过提高深度神经网络模型非线性度来提高深度神经网络模型的健壮性。

检测输入是否为对抗样本：虽然对抗样本与原始样本相比变化十分微小，但神经网络对这两种样本输出特性完全不同，也有学者发现对抗样本无法用重构网络进行重构。因此我们可以通过另一个模型观察神经网络分类器输出特性来区分样本是否为对抗样本，或者使用胶囊重构来辨别输入是否为对抗样本。

本文将选取三种具体的防御策略，进一步分析各方向防御策略的优缺点。

### 2.3.1图像压缩重建

图像压缩重建防御策略属于图像预处理防御策略的一种，其主要思路：在图像进入分类器之前将图像“净化”，得到不包含对抗扰动的干净图像后再送入分类器中；图像“净化”主要利用了图像压缩过程去除冗余信息的思想，图像压缩后只保留了主要信息，对抗样本扰动属于冗余信息，图像压缩重建后这部分冗余信息将被去除；同时压缩重建模型和分类器分开训练，并不会影响到已经训练好的模型。图像压缩重建模型图2.1所示：

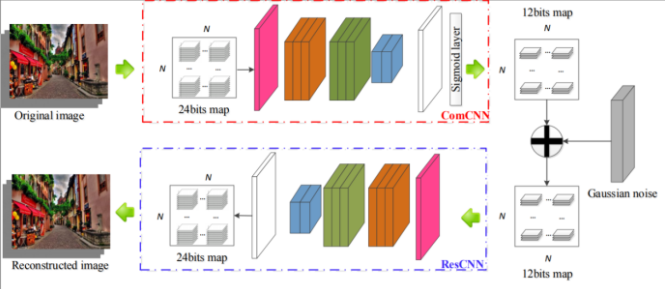


图2.1 图像压缩重建网络结构

图像压缩重建防御策略主要的优点是：不需要专门训练对抗样本，因为未改变分类器模型所以不需要重新训练分类器模型，“净化”后的图像在不同的分类器模型中都有较高的分类准确率。主要缺点是图像在经过压缩重建后图像质量会下降，这会导致分类器对无噪声图像分类准确率下降。

### 2.3.2对抗训练

对抗训练是指在分类器损失函数中添加一个正则项，该正则项是基于对抗样本引发损失的建模。分类器模型使用这种新的损失函数进行训练可以提高模型对对抗样本的泛化能力。实际操作是在原始模型训练集中添加对抗样本，进而提升模型对于对抗样本扰动的健壮性。使用FGSM作为对抗训练模型损失函数修改如下：



其中为模型原始损失函数，表示对抗样本引发的损失。对抗训练的好处是模型可以防御已知的对抗样本生成方法生成的对抗样本（只需向模型训练集中添加足够的该对抗样本生成方法生成的对抗样本），并且防御效果很好。缺点也很明显，对于新种类的对抗样本没有防御能力，并且每出现一种新的生成对抗样本方法，要想防御必须将该生成对抗样本生成的对抗样本加入到模型的训练集中然后重新训练，反复重新训练模型极大的提高了实施防御的成本。

### 2.3.3通过胶囊网络检测对抗样本

Frosst等人提出只需简单的改进胶囊网络训练方式就能使胶囊网络检测出对抗样本图像。他们除了正常训练图像分类任务之外，还训练胶囊网络根据顶层的姿态参数来重构原始图像。因为对抗样本与分类器分类结果类有较大的不同，所以会导致通过该类的顶层胶囊重构图像时会有较大的重构误差（图2.2给出了从真实数据和对抗样本生成的重构图像）。他们通过实验证明在输入图像和重构图像之间的距离上设置阈值可以有效的检测出对抗样本。

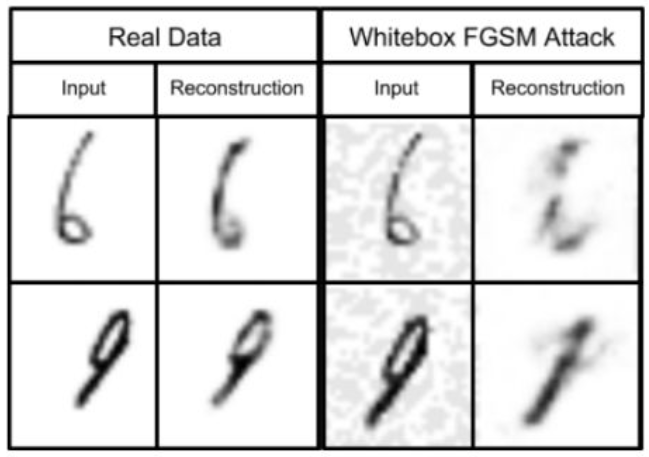


图2.2 真实数据与对抗样本生成的重构图像

该检测对抗样本方法主要适用于较小的数据集（如MNINST），在更复杂的数据（如Cifar-10和ImageNet）中两张图像看起来相似，但距离很大，这导致该方法并不能检测出图像是否为对抗样本。另外作者还发现该方法无法检测出强对抗攻击，此类攻击在可视化结果中，被干扰的图像与目标类的数据流十分相似。

### 2.4对抗样本攻防总结

虽然目前有许多的生成对抗样本方法，同时不同生成对抗样本方法攻击的目标也不同，但对抗样本的特征是基本相同的，目前对抗样本主要特点有：对抗样本存在并不唯一，对抗样本攻击可以是有目标的也可以是无目标的；对抗样本具有迁移性，但对抗样本在用于未知模型时其攻击效果会降低；虽然生成对抗样本方法众多，但其本质是一样的，且目前的生成对抗样本算法大多是对FGSM算法进行改进；使用同一种生成对抗样本算法生成的对抗样本，对抗样本扰动越大其攻击效果越强同时抗干扰能力也越强；深度神经网络对于对抗样本与普通样本输出特性完全不同，这说明对抗样本与普通样本可能有着不一样的分布。

目前的防御策略主要存在三个问题：一是大部分防御策略是削弱对抗样本对神经网络模型的影响，二是许多防御策略需要重新训练分类器模型这导致防御成本极高，三是大部分防御策略会导致无噪声图像分类准确率下降。因此在制定对抗样本防御策略时不仅要考虑防御效果还要考虑防御成本及是否会影响普通样本分类准确率，由于现实生活中对抗样本存在的可能性较小，因此制定防御策略时要着重考虑是否会影响普通样本的分类准确率。如果防御策略能够在不同模型之间迁移，可以有效的减少防御成本。

# 基于VAE-GAN的对抗样本防御策略

目前已有的对抗样本扰动大多十分微小（人眼难以察觉），然而深度神经网络模型对这些扰动十分敏感。目前许多防御手段目的是消除对抗样本中的扰动，使其恢复成原本的图像，这些方法对于部分扰动较大的对抗样本十分有效。但清华大学 TSAIL 团队 Dong 等人发现即使除去对抗样本的大部分噪声，部分攻击性强的对抗样本依然能够影响深度神经网络模型，使其输出错误的分类结果。Issaranon等人发现虽然对抗样本与原始样本十分相近，但深度神经网络模型对于两种样本输出特性完全不同，这表示对抗样本与原始样本的分布可能是不一样的。基于对抗样本以上两点特性，本文提出一种新的防御策略——基于VAE-GAN的对抗样本防御策略，该策略能在几乎不损害图像质量的情况，去除对抗样本的对抗扰动，将对抗样本还原成原始样本。

## 3.1 模型的基本原理

### 3.2.1 VAE的原理

变分自动编码器（Variational Auto-Encoder,VAE）是kingma等人在2012年提出的一种基于变分贝叶斯推断的生成式神经网络模型，主要用于数据压缩，图像降噪及图像重建等方面。VAE是完全的概率推断模型，能够近似的估算出输入样本的边缘概率，并通过最大化该边缘概率来优化模型的参数。如图3.2所示，VAE模型主要由两部分组成：一是编码器对输入数据进行变分推断，生成隐变量的变分概率分布；另一个是解码器将隐变量的变分概率分布还原成输入数据的近似概率分布。

如图3.2所示，表示真实样本的数据集,每个数据样本都是随机产生相互独立连续或者离散的分布变量，表示解码器解码生成的样本，表示隐变量，表示编码器中网络的参数，表示解码器网络中的参数。其中是一个高维空间可观测的随机向量，是一个相对低维的不可观测随机向量，VAE生成模型可以分为两个过程：Encoder推断得到隐变量的近似分布过程，Decoder将隐藏变量还原成近似于输入样本概率分布的过程。

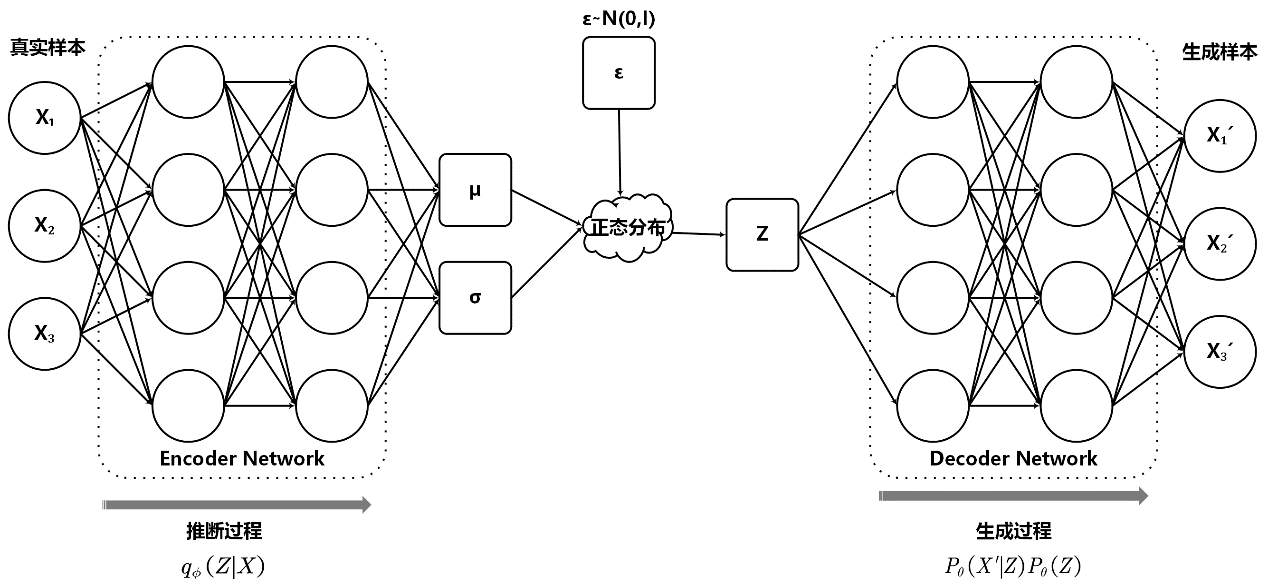


图3.1 VAE的结构示意图

在图3.2中，表示真实样本的数据集,每个数据样本都是随机产生相互独立连续或者离散的分布变量，表示解码器解码生成的样本，表示隐变量，表示编码器中网络的参数，表示解码器网络中的参数。其中是一个高维空间可观测的随机向量，是一个相对低维的不可观测随机向量，VAE生成模型可以分为两个过程：Encoder推断得到隐变量的近似分布过程，Decoder将隐藏变量还原成近似于输入样本概率分布的过程。

对于编码器，由于隐变量的概率分布无法被观察也难以计算，VAE引入了一个神经网络模型来代替隐藏变量的真实后验分布，并且假设服从普通正态分布，同时为了方便计算假设隐变量先验分布服从标准正态分布。为了使引入的识别模型尽可能与隐藏变量真实概率分布，VAE选择使用KL散度来评价这两个分布的相似度：



其中表示的边缘概率分布。

假设：



通过公式3-1和3-2可以得到：



因为恒成立，因此恒成立，也被称为的变分下界。结合公式3-1和公式3-3，编码器的优化目标可以转换为：



由于数据集的对数边缘似然等于各独立离散变量的对数边缘似然之和，即，因此编码器的最终优化目标可以转化为。

通过编码器我们可以得到隐变量的近似后验概率分布，然后我们可以从隐变量的近似后验分布中采样得到稳定的隐变量。解码器的目的是将隐变量恢复为原始输入样本的近似分布，解码器将隐变量恢复为输入样本近似分布可以表示为：



我们希望越大越好，结合公式3-3，可以将解码器的优化目标转化为：



结合公式3-4，解码器的损失函数也可以转化为。由于编码器和解码器的损失函数都是，所VAE整体优化目标为变分下界。由于公式3-2无法直接求解，我们需要对变分下界函数进行转化，变分下界函数可以最终转化为一个包含编码器和解码器结构的函数：



由于一开我们假设服从普通正态分布，同时也假设服从标准正态分布，因此变分下界函数3-2的前半部分可以进行转化：



对于变分下界的右半部分，该项是关于的后验分布，由于在VAE 没有对对解码器做太强的假设，因此该项不能通过解析的方式解出，kingma等人提出可以通过采样的方式对该项取近似值，因此该项可以转化为：



需要注意的是不是直接从分布中直接采样，而是使用了一种重参数技巧。该技巧将从直接采样，转变为先从采样得到，然后通过计算得到的采样。这是因为直接从采样，采样过程会参与模型参数的优化，而该采样过程不可微，这将导致整个模型无法训练。而从采样再计算的方式，采样过程不会参与模型参数的优化，所以不会影响模型的训练过程。最终采样得到的一般是伯努利分布或者高斯分布。假设采样分布为伯努利分布该项最终可以转化为：



其中表示解码器生成的结果，假设采样分布为高斯分布，则解码器解码的结果为高斯分布的均值，即解码器只关心均值（为未知数的值，且对优化目标没有影响），所以该项最终可以转化为：



综上所述，实际训练中VAE的优化目标可以表示为：



或者：



对应的VAE的损失函数可以表示为：



或者：



在训练过程中损失函数的第一项与解码器网络参数更新无关，解码器的参数更新只受损失函数第二项的影响，该项也被称为解码器的重构误差。

综上所述，VAE模型的最终结构如3.2所示：

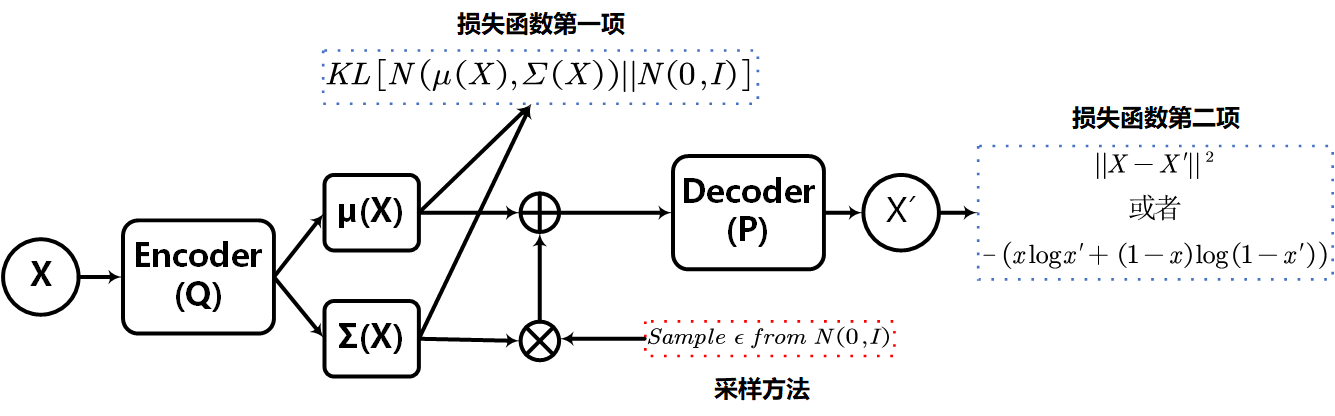


图3.2 VAE模型最终结构

3.1.2 DVAE的基本原理

降噪变分自编码器（Denoising Variational Auto-Encoder,DVAE），是将去噪准则（Denoising Criterion）和VAE结合在一起的一种自编码器。与VAE压缩重构过程不同，DVAE向输入的分布添加而外的随机噪声，然后将这个有噪声的分布作为VAE的输入，最后要求VAE的解码器重构出无噪声的分布。DVAE的结构如图3.3所示。

与VAE的优化目标类似，DVAE的优化目标为DVAE的变分下界函数，由于两个模型的变分下界函数推导过程基本一致，本文不在对DVAE的变分下界函数进行推导直接给出DVAE的优化目标：



或者：



其中和表示由带有噪声的样本编码得到的隐变量的近似后验概率分布的均值和方差。对比VAE和DVAE的优化目标，两者优化目标形式上基本上是一样的，其主要区别是在实际训练过程中VAE的重构分布是编码器输入的分布，而DVAE重构分布不是其编码器输入的分布。

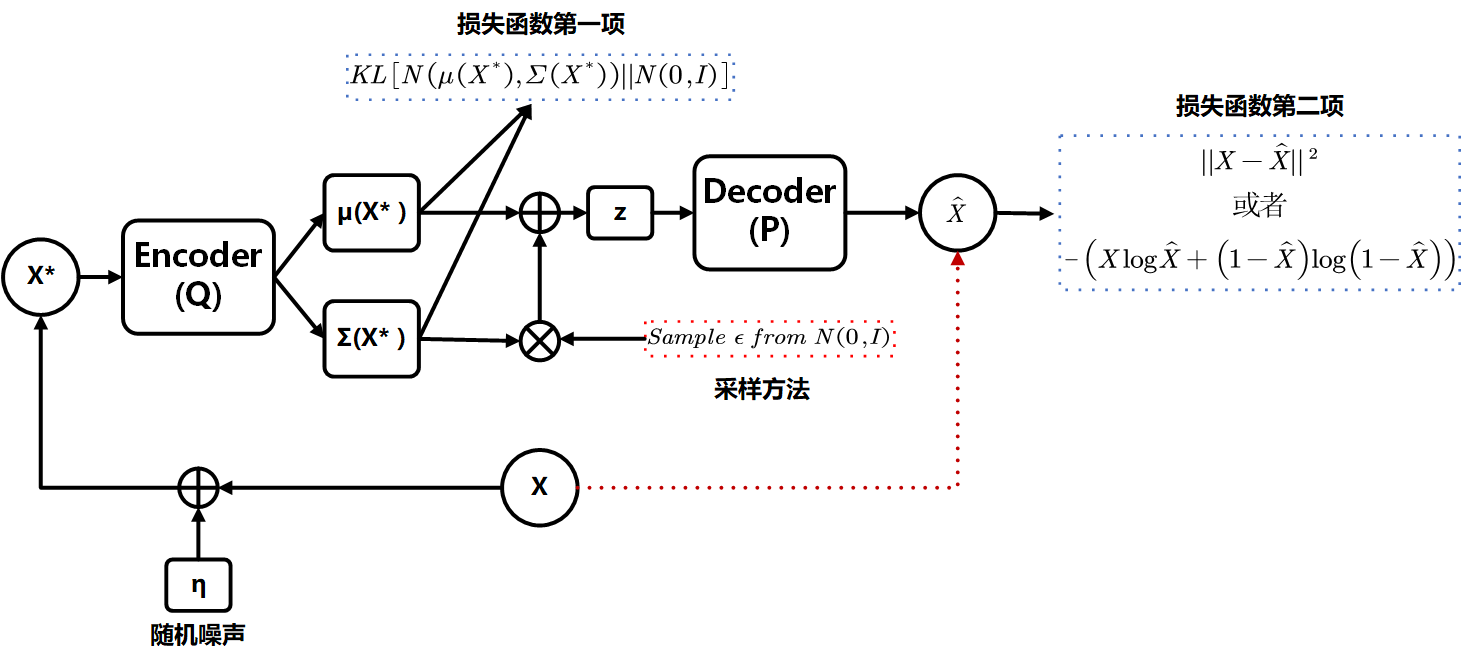


图3.3 DVAE结构示意图

3.1.3 GAN的基本原理

生成式对抗网络（Generative Adversarial Networks，GAN）是一种无监督生成式学习模型，由Goodfellow等人于2014年提出，GAN对抗学习的思想不仅广泛的运用于图像的生成领域，也为其许多其他领域提供了新思路。

如图3.4所示，GAN包含两个网络结构，一个生成网络G（Generator）和一个判别网络D（Discriminatior），其中判别网络D与我们普通分类网络类似。GAN的生成网络G输入为一个随机噪声Z，输出为生成的图像样本G(Z)。判别网络D的输入为G(Z)和真实图像样本，然后判别网络D会对输入辨别分类输出一个概率值用来表示与真实图像样本的相似度，比如判别网络D的输入是G(Z)时输出为0.3，那么表示判别网络D认为G(Z)有30%的概率是真实图像样本。然后生成网络G和判别网络D会根据D的输出不断的优化自己，生成网络G尽可能的提高自己的输出G(Z)与真实图像样本的相似度，即尽可能的欺骗判别网络D，而判别网络D通过学习尽可能的分辨出谁是真实图像样本谁是生成的图像样本。生成网络G和判别网络D相当于在做一个极大极小的博弈过程，该过程也被称为零和博弈。

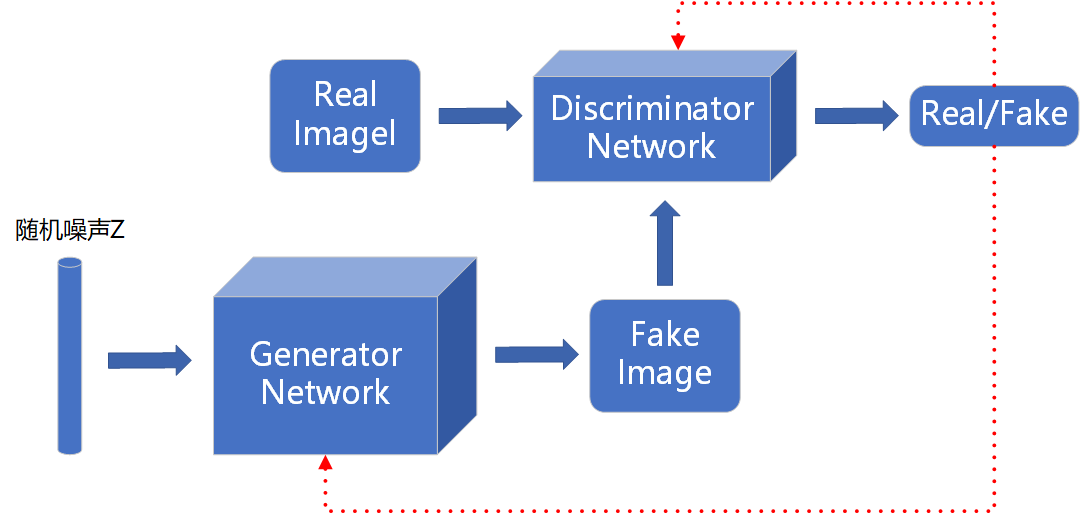


图3.4 GAN网络示意图

与传统深度学习模型训练不同，GAN训练过程中两个网络模型交替训练。在训练判别网络D时，固定生成网络G，相反在训练生成网络G时，固定判别网络。通过反复的对抗训练，最终使生成网络G生成的图像样本与真实图像样本的分布一致，判别网络D无法辨别输入的图像是真实图像样本还是生成图像样本。GAN的优化目标定义如下：



其中D代表判别网络，G代表生成网络，代表生成网络的输入随机噪声，服从分布，表示真实图像样本服从分布，表示期望。对于判别网络D，其目标是最大化优化目标函数，使得接近并且接近0。对于生成网络，其目标是使优化目标函数尽可能的小，也就是要尽可能的接近1。

GAN的生成网络和判别网络通过反复的对抗训练，最终会达到纳什均衡，即生成网络和判别网络都达到最优。如果固定生成网络中的参数，假设真实图像样本分布为，生成图像样本的分布为，则判别网络最优为：



最终解得最佳判别网络为：



在判别网络最优时，生成网络最优为：



引入JS散度公式：



结合公式3-和3-可得：



当真实图像样本分布与生成图像样本分布相同时，生成网络优化目标达到最优为，同时将该条件带入最佳判别网络中，得最优辨别网络输出为。也就是说最终判别网络辨别真实样本的准确率只有50%时，生成网络和判别网络达到纳什均衡。

从上述分析可知，GAN生成的图像不是单纯的记忆数据集中的图像样本，而是通过特征学习生成了新图像。GAN对抗训练有很多优点，但也存在一个比较大的缺点：初始的判别网络训练的不好，将会导致生成网络的损失函数值很小，导致生成网络无法学习。比如在最优判别网络条件下，生成网络的损失函数值将变为和的散度，两个分布相同的时散度值最小为0，相反当两个分布完全不同时散度值将收敛为一个常数。由于初始条件下生成网络生成的图像样本分布与真实图像样本的分布几乎不可能重叠，此时和的散度值为，将该值代入公式3-可得生成网络损失函数值为0，也就是说生成网络的梯度值为0，生成网络将无法学习到如何有效信息。

## 3.2基于DVAE-GAN的对抗样本防御策略

针对对抗样本的噪声特性和扰动累加效应，本文提出一种基于DVAE-GAN的对抗样本防御策略，该策略主要思路为：在对抗样本图像进入分类器之前通过DVAE将图像进行压缩重建的方法去除对抗样本中的扰动，得到降噪后的图像样本再送入分类器中；为了保证通过DVAE压缩重建后的图像不是对抗样本，使用GAN辅助VAE训练。本文提出防御策略分类器和防御模型分开训练，防御模型不会影响已经训练好的分类器模型。

### 3.2.1模型整体框架

基于DVAE-GAN的对抗样本防御策略整体框架如图3.4所示

如图3.4所示，基于DVAE-GAN的对抗样本防御策略一共包含两个部分：DVAE模块和GAN模块。DVAE模块的主要功能是通过对抗样本进行压缩重建去除对抗样本中的扰动，DVAE模块的输入为对抗样本，输出为去噪后的图像。GAN模块的主要功能是辅助DVAE模块的训练，GAN的生成器和DVAE模块的解码器共用一个神经网络结构，GAN的判别器输入为生成器生成的图像样本和真实的图像样本，GAN的输出为输入样本是真实样本的概率。

本节接下来将详细的介绍整个模型各模块的功能及其模型的整体优化目标。

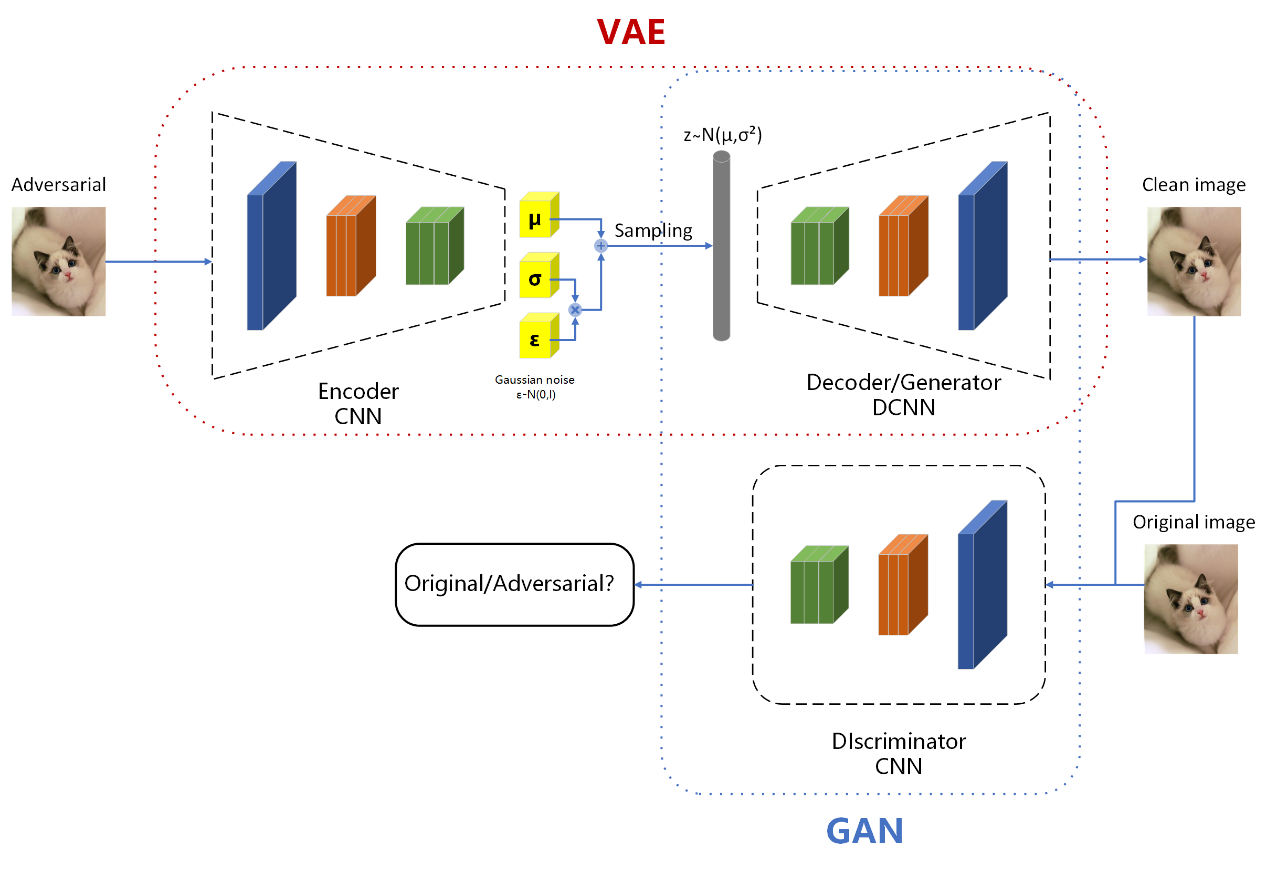


图3.4 基于DVAE-GAN的对抗样本防御策略整体框架

3.2.2 DVAE模块

对抗样本是在原始自然图像中加入特定噪声生成的，这部分噪声相对整个图片来说很小，但并不是完全不可见，如果将图片进行局部放大还是可以发现对抗样本和原始自然样本有很大的不同。由于图像局部结构中的相邻像素有很强的相关性和相似性，图像压缩可以有效的保留图像的显著性信息，同时也能够减少这部分的冗余信息，而对抗样本中的噪声对于原始图像来说是一些额外的冗余信息，因此通过图像压缩重建可以有效对的减少对抗样本中的对抗扰动。为了在图像压缩重构过程尽可能的保证图像的质量，本文选择使用DVAE模型通过压缩重建的方式去除对抗样本中的噪声。

DVAE模块的编码器主要功能是将输入的图像转化为隐变量的近似后验概率分布，然后对近似后验分布进行采样得到稳定的隐变量。编码器的输入为对抗样本或者原始样本，输出是两组维向量：一个组均值向量，另一组是标准差向量，然后我们可以通过这两组维向量构建隐变量的后验近似概率分布。最后我们再从隐变量近似概率分布采样获得隐变量，采样方式为，其中表示从标准正态分布中随机取样的结果。编码器的网络结构有很多选择，比如深度神经网络（Deep Neural Networks，DNN）、卷积神经网络（Convolutional Neural Networks, CNN）及深度置信网（Deep Belief Network，DBN），本文选择使用卷积神经网络作为编码器的网络结构，卷积网络的具体层数由数据集的复杂程度决定。DVAE的解码器主要功能将隐变量恢复为无噪声图像，本文选取的解码器采样分布为高斯分布，因此解码器。解码器的输入为隐变量，输出为降噪后的图像。解码器的网络结构一般为DNN或者反卷积神经网络（Deconvolutional Neural Networks, DeCNN），本文选择反卷积神经网络作为解码器的神经网络结构，反卷积神经网络的层数由数据集的复杂程度决定。

3.2.3 GAN模块

虽然使用DVAE模块能够去除对抗样本的中的绝大多数的对抗扰动，但去噪后的图像的分类准确率并没有增加多少，仅使用DVAE防御对抗样本结果如表3.1所示，从表3.1中我们可以发现仅使用DVAE对对抗样本进行除噪并不能完全防御对抗样本的攻击，因此DVAE的压缩重建并不能保证去噪后的图像不是对抗样本。通过表3.1还可以发现，普通样本通过压缩重建后分类准确率会下降，那是因为DVAE解码器的损失函数是原始图像与生成图像均方差，如果生成的图像的高频和低频有相同的噪声，那么DVAE解码器所得到的损失函数值是一样的，但相同的噪声在图像的高频部分和低频部分对图像的分布影响是不一样的，通常情况下噪声在图像的高频部分对图像的分布影响较小，在图像低频部分对图像的影响较大，DVAE使用均方差作为解码器的损失函数，导致解码无法学习到图像的分布特性，最后导致DVAE生成的图像相较于原始图像比较模糊。

表3.1 仅使用DVAE防御对抗样本

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| model | Original | FGSM() | FGSM() | FGSM() |
| Classifier | 98.52% | 34.38% | 27.34% | 22.17% |
| DVAE+Classifier | 96.67% | 52.56% | 49.38% | 45.22% |

针对DVAE存在的上述两个问题，本文使用GAN模型辅助DVAE模型的训练提高DVAE解码器生成图像的质量，同时保证该生成图像不是对抗样本。GAN的生成器和DVAE的解码器共用一个网络，这相当于给DVAE的解码器添加了一个GAN生成器的损失，从本章3.1.3节可知GAN的损失函数是关于生成图像分布和真实图像分布的散度，生成图像分布与真实图像分布越接近该值越小，因此GAN辅助DVAE训练可以让DVAE的解码器在学习到图像的像素特点的同时学习到图像的分布特性。GAN的生成器在给DVAE解码器提供额外的损失函数的同时DVAE的解码器也相对的给GAN的生成器提供了一个均方差损失函数，而这个损失函数恰好可以提高GAN生成器前期生成图像的质量，从而解决了GAN训练难的缺点。GAN的判别器是一个二分类的分类器网络，如果输入的图像是对抗样本，由于对抗样本扰动累加效应判别器的输出会与真实图像的输出不同，这说明在GAN的判别器眼中对抗样本和真实样本属于两种分布，因此通过GAN的生成器与判别器的对抗训练，可以保证DVAE解码器生成的图像不是对抗样本。

GAN的生成器与DVAE的解码器共用一个网络，所以GAN的生成器输入输出与DVAE的输入输出一致。GAN的判别器的输入是生成器生成的去噪后的图像和该图像的原始图像，输出为输入图像是真实图像的概率。GAN的判别器的网络选择也有很多，本文采用卷积神经网络来搭建GAN的判别器网络，卷积神经网络的层数由数据集的复杂程度决定，GAN的生成器网络与DVAE的解码器共用一个网络结构无需重新构建网络。

3.2.3模型的优化目标

基于DVAE-GAN的对抗样本防御模型由DVAE和GAN两个模型组成，一共有三个网络结构：DVAE的编码器网络，DVAE解码器和GAN生成器共用网络和GAN的判别器网络。对于DVAE的编码器，其优化目标就是最大化DVAE的变分下界函数：



对应的DVAE编码器的损失函数为：



由于DVAE的解码器和GAN的生成器共用一个网络，因此DVAE解码器和GAN共用的网络的优化目标包含DVAE的变分下界函数和GAN生成器的损失函数：



由于DVAE变分下界函数前半部分与DVAE的解码器无关，同时GAN的优化目标的第一项在GAN的生成器训练时是固定值，因此可以将这两项舍弃。然后DVAE解码器和GAN共用网络的优化目标可以转化为：



该优化目标可以同时评价生成图像与原始图像的像素相似性和分布相似性，但实际情况中不同的数据集可能要权衡像素相似性和分布相似性对图像的质量的影响，同时一般情况下优化目标均方差项要远大于散度项，所以本文引入一个额外参数来控制DVAE解码器和GAN共用网络生成图像的偏重，改进后的优化目标为：



需要注意的是额外参数只在更新DVAE解码器和GAN共用网络参数时候使用，更新DVAE编码器网络时不使用该参数。对应的DVAE解码器和GAN共用网络可以表示为：



GAN的判别器网络优化目标和普通GAN判别器优化目标一样为：



对应的GAN的判别器网络的损失函数可以表示为：



综上所述，模型整体优化目标为：



注意参数只在更新DVAE解码器和GAN生成器共用网络时使用。对应的模型整体损失函数为：



注意由GAN对抗训练的存在，模型整体的损失函数无法通过一个函数进行描述，公式3-只是一个抽象的损失函数。

3.3 模型的训练流程

在训练过程中，由于GAN对抗训练的存在，我们无法直接根据模型的整体优化目标直接更新全部网络的参数，最后本文选择分步更新模型中各网络结构中的参数，网络参数更新顺序为：先更新DVAE编码器网络，然后更新DVAE解码器和GAN生成器共用网络，最后更新GAN判别器网络。DVAE-GAN训练流程如表3.2所示。

训练具体流程为：

1.初始化DVAE编码器网络，DVAE解码器和GAN生成器共用网络和GAN判别器网络中的参数；

2.如果剩余训练周期大于0并且模型各模块的损失函数不是无穷大执行步骤3，否者训练结束；

3.从训练集中选取一批图像样本；

4.随机向图像样本中添加对抗扰动，得到；

5.通过DVAE的编码器网络计算隐变量后验概率分布的均值和方差和；

6.从隐变量的后验概率分布采样得到隐变量；

7. 通过DVAE解码器和GAN生成器共用网络将隐变量还原成无噪声图像；

8.计算DVAE编码器网络的损失函数；

9.计算DVAE解码器和GAN生成器共用网络的损失函数；

10.计算GAN判别器网络的损失函数；

11.更新DVAE编码器网络中的参数；

12.更新DVAE解码器和GAN生成器共用网络中的参数；

13.更新GAN的判别器网络中参数；

14。训练周期减1，跳转到步骤2。

表3.2 DVAE-GAN训练流程

|  |
| --- |
|  |
|  |

3.4 本章小结

本章针对对抗样本的噪声特性和扰动累加效应，提出了一种基于DVAE-GAN的对抗样本防御策略。首先简单介绍了该防御策略的防御思路和模型的整体框架，然后详细介绍了防御模型中的DVAE模块和GAN模块，并解释了各模块对于防御对抗样本的作用。然后分析了各个模块的优化目标及模型的整体目标。最后详细介绍了整个防御的算法流程，训练流程及训练过程中需要注意的细节。

# 5实验及结果分析

5.1实验环境及实验数据集

5.2模型网络结构

5.3MNIST数据集结果及分析

5.4CIFAR-10数据集结果及分析

5.5实验总结