**目录**

**1.绪论1**

1.1研究背景目的及意义2

1.2 国内外研究现状3

**键入章标题(第 1 级)4**

键入章标题(第 2 级)5

键入章标题(第 3 级)6

# 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。

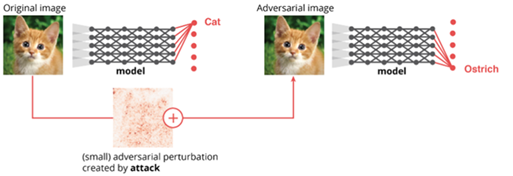


图1 对抗样本攻击举例

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

2013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，简称 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，简称 BIM)、最小可能类法 (Leastlikely Class Method，简称 LCM)、迭代最小可能类法 (Iterative least-likely class method，简称 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，简称 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[33]提出了动量迭代法 (Momentum Iterative Method，简称 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，简称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，简称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，简称DCN），该方法使用了类似于压缩自动编码器（Contractive Auto Encoders，简称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变成现在的大型自然数据集，这极大的增加我们设计对抗样本防御策略的难度。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