**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。

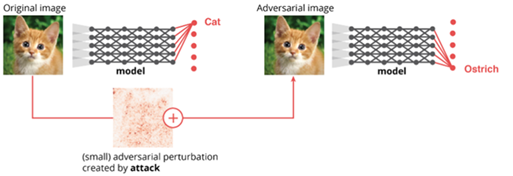


图1 对抗样本攻击举例

对抗样本能够欺骗深度神经学习模型根本的原因是深度学习模型在训练过程存在过拟合，这也导致对抗样本无法从根源上进行防御。目前大多防御策略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。

## 1.2研究现状

### 1.2.1生成对抗样本研究现状