|  |  |
| --- | --- |
| **文章信息** | **动机及方法概述** |
| **阅读程度**：  实验重点看  **文章标题：**  Diffusion Models for Adversarial Purification  **中文标题：**  使用扩散模型进行对抗性净化  **发表于：**  ICML2022  **作者：**  Weili Nie  **单位：**  NVIDIA | 其实很简单，方法就是使用VP-SDE扩散模型，但是不扩散为纯噪声，只添加少量扩散噪声防止破坏图像的语义信息，然后走反向过程清除图像上的噪声。  作者的为了显示工作量进行了两个理论分析：  第一：    即对抗性分布和原始分布的KL散度会随着扩散过程逐渐缩小，这意味着无论是从对抗性分布出发还是原始分布出发，经过几步扩散后得到的分布是差不多的。  第二：    即经过正向反向过程后得到的去噪样本与干净样本之间的范数有一个理论上限。  另外作者为了评估此防御方法在白盒情况下的效果，设计了一种梯度求解方法可以获得整个防御过程的梯度从而来发动自适应攻击。  个人觉得方法的创新性是不足的，只是对扩散模型的简单应用，重新性工作落脚在两个理论分析上，没什么意思。 |
| **摘要及贡献** | **实验** |
| 对抗性净化是指一类使用生成模型消除对抗性扰动的防御方法。这些方法没有对攻击的形式和分类模型进行假设，因此可以保护现有的分类器免受看不见的威胁。然而，他们的表现目前还落后于对抗性的训练方法。（简单概括对抗性净化这种防御方法的有点和缺点）  在这项工作中，我们提出了使用扩散模型进行对抗性纯化的DiffPure：给一个对抗样本，我们首先在正向扩散过程中用少量的噪声扩散它，然后通过反向生成过程恢复干净的图像。（简述方法的过程）  为了以一种有效和可扩展的方法来评估我们的方法对强自适应攻击，我们建议使用伴随方法来计算反向生成过程的全梯度。（设计一种攻击对提出的防御进行白盒攻击）  在CIFAR- 10、Imgaenet和CelebA-HQ三个数据集和三种分类器架构ResNet、WideResNet和ViT进行广泛实验，表明我们的方法取得了最先进的结果，优于当前的对抗性训练和对抗性净化方法，通常有很大的优势。 | **Experimental Settings：**三个数据集，三个网络，计算在防御条件下自然样本的分类精度和对抗样本的分类精度。  **Comparison with the State-of-the-art：**  实验一，与对抗训练的方法比，Cifar10数据集，WideResNet-28-10/70-16两种网络，基于无穷范数和二范数的自适应攻击。  实验二，与对抗训练的方法比，Imagenet数据集，ResNet、WideResNet和ViT三种网络，基于无穷范数的自适应攻击。  **Defense Against Unseen Threats：**  选择只用无穷范数/二范数/StAdv训练的对抗训练网络，测试他们在面对三种攻击时的精度，与本文方法比较。  **Comparison with Other Purification Methods：**  实验一：其他生成模型如VAE GAN用于对抗性去噪，在数据集CelebA-HQ，使用BPDA+EOT attack作为攻击去对比  实验二：其他对抗性去噪方法，在基于Cifar10数据集使用BPDA+EOT 去攻击对比  **Ablation Studies:**  实验一：测试扩散时长t对实验的影响  实验二：测试反向采样算法的影响：LD-SDE（直接使用郎之万采样法，以对抗样本为起点企图还原原始样本） VP-ODE使用ODE方法进行反向SDE求解 VP-SDE正常的SDE反向求解（predictor+corrector？？）  实验三：将backbone换成使用对抗训练的网络，可以提高对对抗样本的分类正确率，（但是干净样本应该会下降，原文中未指出。 |