|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  精读  **文章标题：**  Learning to Generate Noise for Multi-Attack Robustness  **中文标题：**  学习生成噪声来实现对多种攻击的鲁棒性  **发表于：**  ICML2021  **作者：**  Divyam Madaani  **单位：**  School of Computing, KAIST, South Korea  ***值得一提稿件被ICLR2021拒稿但是三个评审意见中两个是6，即略大于接受阈值，一个是5即略小于接受阈值，然后被ICML2021接受==，有一说一有东西，但不多***  ***我觉得这篇文章可以借鉴，这里面的所有实验我都有能力完成*** | 第二行抽取一个mini-batch B，第三行从三种范数中随机抽取一个作为攻击策略A(x)，第四行基于当前主网络参数生成对抗样本，第五行基于当前生成网络参数生成对抗样本。  **第六行，使用作为输入临时更新**  **第七行，使用作为输入更新**  这两部只更新，不会对产生改变我对这两步的理解是，首先，我们**假设**使用生成的对抗样本去优化**，**我们希望优化后得到临时参数在面对真正的对抗样本时损失函数能够下降。即我们希望生成器学会生成一种虚拟对抗样本，如果我们使用这种虚拟对抗样本去训练网络，能够降低网络对真实对抗样本的分类损失（而实际上我们并不会用虚拟对抗样本去优化）。通过这两步，使得生成器能够学习到三种对抗样本的一些综合特性。  第八行用更新过后的再生成一批新的，之后使用去求交叉熵损失，使用，，结合求一致性损失  两页介绍  半页相关工作  半页MAX AVG  最后一页是自己的东西 |
| **摘要及介绍** | **实验** |
| 由于深度学习系统容易受到对抗性攻击，在开发（包括经验的和可证明的）鲁棒分类器方面已经做了大量的工作。虽然大多数工作都在防御单一类型的攻击，但最近的研究是通过使用多个攻击的简单聚合来防御多个扰动模型。  然而，这些方法可能很难调整，并且很容易导致对单个扰动模型的鲁棒性的不平衡程度，从而导致并集上的次优最坏情况损失。（问题）  在这项工作中，我们开发了基于PGD的标准过程的自然推广，通过对所有最陡下降方向的最坏情况，将多个扰动模型合并到单一攻击中。（提出方法）  这种方法的优点是直接收敛于不同的扰动模型之间的权衡，从而最小化最坏情况下的联合性能。（优点）  用这种方法，我们能够训练标准架构同时健壮对ℓ∞，ℓ2，ℓ1攻击，超越过去的方法MNIST和CIFAR10数据集和实现对抗精度47.0%的联合（ℓ∞ℓ2ℓ1）扰动半径=（0.03,0.5,12）后者，改进以前的方法达到40.6%的精度。 | 5.1 实验设置  5.2 在CIfar10 SVHN TinyImagenet三个数据集上测试，单个攻击，单个精度，平均精度和联合最差精度和时间，对比方法包含了传统对抗训练PGD-AT和trades以及多威胁模型对抗训练MAX AVG MSD。  5.3 消融实验  实验一：组件分析，，分析了一致性损失和MNG组件的重要性  实验二：超参数选择  5.4 其他实验  **实验一：对空间转换攻击的防御能力**  **实验二：损失景观的可视化**  **实验三：决策边界的可视化** |