|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  粗读  **文章标题：**  Adversarial Robustness against Multiple and Single lp-Threat Models via Quick Fine-Tuning of Robust Classifiers  **中文标题：**  通过快速微调获得对多种范数威胁模型的鲁棒性  **发表于：**  ICML2022  **作者：**  Francesco Croce  **单位：**  University of T¨ubingen, Germany.  跟AutoAttack是一个作者  ***值得一提稿件被ICLR2022拒稿但是三个评审意见中两个是5，即略小于接受阈值，一个是3即不行，然后被ICML2022接受==，有一说一没东西。*** | 作者提出了一个东西E-AT：  首先作者给出了一些观点和证明：即仅训练l1可以得到一定l2的鲁棒性，仅训练l无穷/l2可获得l2/l无穷的鲁棒性，然后给出了一些理论证明，即获得了l1与l无穷的鲁棒性可以近似定量计算出将会获得的L2的鲁棒性。  所谓E-AT（Extreme norms adversarial training）指，使用两种极端范数（1与无穷）进行对抗训练就能获得所有范数的鲁棒性，然后E-AT就是每个batch用1范数或无穷范数生成对抗样本，不同于SAT（随机抽取一种训练方法），E-AT使用错误率当做概率选择攻击范数，（攻击效果越好的范数说明网络对此范数鲁棒性较差，就以更大的概率选择此范数进行对抗训练）。   1. AT有两种使用方式，第一种是从随机初始化开始完全使用E-AT进行训练   第二种是只使用一种范数进行训练，收敛后使用E-AT进行微调，CIfar10数据集微调3个epoch，Imagenet数据微调1个epoch。 |
| **摘要及介绍** | **实验** |
| 对抗鲁棒模型的一个主要缺点（特别是对于像ImageNet这样的大规模数据集）是与标准数据集相比，训练时间非常长。此外，模型不仅应该对一个lp威胁模型具有鲁棒性，而且对所有模型都具有鲁棒性。（问题）  本文基于lp球的几何性质，提出了一种基于多范数鲁棒性的极端范数对抗训练（Extreme norm Adversarial Training E-AT）。（提出方法）  对于多范数鲁棒性，E-AT的成本比其他对抗性训练方法低三倍。利用E-AT，我们证明了对于ImageNet只需一个epoch和CIFAR-10需要三个epoch足以将任何lp鲁棒模型转化为多重鲁棒模型。（优势）  通过这种方式，我们得到了ImageNet的第一个多范数鲁棒模型，并将CIFAR-10上最先进的多范数鲁棒性提高到51%以上。（结果）  最后，我们研究了通过微调不同单个lp威胁模型之间的对抗鲁棒性的一般转移，并在CIFAR-10和ImageNet上，在1范数的对抗鲁棒性上达到了SOTA。  大量的实验表明，我们的方案工作跨数据集和架构，包括ViT。（跨数据集，跨模型） |  |