|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  精读  **文章标题：**  Adversarial Robustness Against the Union of Multiple Perturbation Models  **中文标题：**  对各种扰动并集的对抗鲁棒性  **发表于：**  ICML2020  **作者：**  Pratyush Maini  **单位：**  印度理工学院计算机科学与工程系  ***值得一提稿件被ICLR2020拒稿认为不太行，然后被ICML2020接受==，有一说一确实不行*** | 首先介绍了最速更新方向（输入是二维）：如图所示黑色是梯度方向，红色是二范数，沿着梯度方向更新即可，蓝色是无穷范数，沿着梯度方向的sign方向更新，绿色是一范数，选择了导数值最大的方向（根据黑色方向可以看出来，沿y轴的导数比沿x轴的导数大，因此完全沿y轴更新就是1范数的最速更新方向）。  然后介绍了MSD算法（multi steepest descent），简单来说MAX是找使损失函数最大的那个威胁模型去生成攻击作为最大化过程，AVG是使用每个威胁模型单独攻击然后求平均作为最大化过程，MSD是使用每个威胁模型联合起来一起攻击作为最大化过程，联合的方法也很简单，就是每种威胁模型产生的噪声都累加起来。  这哥们介绍写了一页，相关技术写了一页，对抗训练（基本的pgd对抗训练）写了一页，MAX和AVG写了一页，全tm不是自己的东西，MSD写了半页然后直接开始实验，勾八研究生毕业设计的水平 |
| **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。 | 真没意思。 |