|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  粗读  **文章标题：**  ℓ∞-Robustness and Beyond: Unleashing Efficient Adversarial Training  **中文标题：**  无穷范数鲁棒性及其扩展：释放有效对抗训练  **发表于：**  ECCV2022  **作者：**  Hadi M. Dolatabadi  **单位：**  School of Computing and Information Systems  The University of Melbourne  Parkville, Victoria, Australia | 基本流程即使这样，首先在所有训练集上进行预热，  之后使用coreset selection算法生成选择一个训练集的子集，训练T轮后再使用coreset selection算法重新选择一个训练集的子集。  作者表示coreset selection算法已经在各种机器学习算法中得到应用，并且在传统深度学习训练过程中得到使用，作者首次给他引入对抗训练中。  关于coreset selection算法主要的求解目标如下：V代表的是整个训练集，于是第一项就是整个训练集的损失函数相对于网络权重的梯度的和，第二项是在训练子集S中求所有样本的损失函数相对于网络权重的梯度的加权平均和，我们希望这两项的差值尽可能小，那么我们就可以使用S子集的损失值梯度的加权平均去近似整个训练集V的损失值梯度和。关键在于找到合适的S及加权系数。**这就是所谓的coreset选择，文中提供了几种coreset选择算法。**    这里是常规训练的coreset选择，如图b所示，对于对抗训练，可以先生成对抗样本，再求对抗样本的梯度加权平均和（作者这里拿出的理论依据是丹斯什么什么引理，很PGD论文里面是一个东西）。  至于标题中Beyond主要体现在作者使用coreset select算法与PAT（感知对抗训练）结合，Beyond主要是PAT提供的。 |
| **摘要及介绍** | **实验** |
| 神经网络很容易受到敌对的攻击：在它们的输入中添加精心制作的、难以察觉的扰动可以修改它们的输出。  对抗性训练是训练鲁棒模型对抗此类攻击的最有效的方法之一。然而，它比普通的神经网络训练要慢得多，因为它需要在每次迭代中为整个训练数据构造对抗性的例子，这阻碍了它的有效性。  近年来，快速对抗训练（FAT）提出了能够有效地获得鲁棒模型的FAT训练。然而，它成功背后的原因还没有完全了解，更重要的是，它只能训练ℓ∞边界攻击的鲁棒模型，因为它在训练中使用FGSM。  在本文中，通过利用coreset select的理论，我们展示了如何选择训练数据的一个小子集提供了一个通用的、更有原则的方法来减少鲁棒训练的时间复杂度。  与现有的方法不同，我们的方法可以适应各种各样的训练目标，包括trades、ℓp-PGD和感知对抗性训练（PAT）。  我们的实验结果表明，我们的方法加速了对抗性训练的2-3倍，同时经历了一个轻微的降低的清洁和稳健的准确性。  (算是一个比较清奇的思路，FAT旨在使用随机初始化与FGSM通过单步对抗训练提速，作者说那我直接把训练集减小，选择一个子集去训练，就不用在整个训练集上生成对抗样本，那我的训练速度不就上去了)（别人在扩充样本刷性能，哥们在砍掉样本拼速度）。 |  |