|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  方法精读  **文章标题：**  Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses  **中文标题：**  引导对抗攻击去评估和加强对抗训练  **发表于：**  NIPS2020  **作者：**  Gaurang Sriramanan, Sravanti Addepalli  **单位：**  Video Analytics Lab, Department of Computational and Data Sciences Indian Institute of Science, Bangalore, India  （两个印度阿三2020年提出来二范数单步训练 2021年提出核范数单步训练NuAT，ohhhhhh这个什么勾八NCAT也是这哥们搞的这哥们2022年转学了，不在印度了我表示无语） | 作者首先分析了初始轨迹对攻击的影响，这里其实就是在讲故事，因为没有任何的实验验证：    首先仅使用损失值上升来生成对抗样本并不是一个好的做法，因为梯度上升最快的方向并不一定意味着最近的决策边界，如（a）所示，在给定的空间中其实存在被错分为c3的对抗样本，但是梯度最大的方向是朝着c2类进发，但是在给定的范围内并不能成功地将样本攻击为c2类。  Pgd算法的随机初始化一定程度上解决了这个问题，这使得攻击有更多的可能朝着不同的方向进展，但是由于初始化噪声是随机添加的，因此可能需要数次初始化才能够起到效果，这在一定程度上拖慢了攻击算法的效率。  一个改进的方法是将无目标攻击改为多个有目标联合攻击，即每次重启后都朝着一个指定的目标类进行攻击，如图b所示，但是这种攻击方法会受限于类别数，如果数据集的类别数很大，这无疑是一个耗时低效的方法。  最后作者介绍本文提出的引导攻击方法，这里暂时还没有细说是怎么引导，但是说可以仅通过一次重启或少数几次重启，达到与上一段提到的方法类似的效果。  没什么意思：边际损失加个二范数：  二范数的梯度回传（因为x2的导数的2x，f(x’)-f(x)是一个长度为C的向量，C是类别数，如果向量的第一个值（第0类的置信度的差值）为k，则第0类节点回传的梯度值为2k，也就是说某一类的置信度差值越大，回传的梯度值越大，最终的梯度更倾向于扩大这个类别的差异）。损失函数的第一项为边际损失。  关于如何通过梯度信息获得更新方向，PGD是第9行展示，FW是第12行展示，FW更像是一种迭代的过程，。 |
| **摘要及介绍** | **实验** |
| 对抗性攻击的发展进展是对抗性防御研究进展的基础。高效和有效的攻击对于可靠的防御评估和开发健壮的模型都是至关重要的。（背景，好的攻击是评估防御的基础）  对抗性攻击通常是通过最大化标准损失（交叉熵损失或最大边际损失）来产生的，在使用约束集中使用投影梯度下降（PGD）。（之前的工作）  在这项工作中，我们在标准损失中引入了一个松弛项，它可以找到更合适的梯度方向，提高攻击效能，并导致更有效的对抗性训练。（我们的工作方向）  We propose Guided Adversarial Margin Attack (GAMA)，它利用干净图像的function mapping来引导对手的生成，从而产生更强的攻击。我们评估了我们对多种防御的攻击，并显示出与现有的攻击相比有改进的性能。（攻击方面的成果）  此外，我们提出了引导对抗训练（GAT），它通过利用所提出的松弛项，在单步防御中实现了最先进的性能。（防御方面的成果） | 测试了提出攻击的性能。  防御与单步方法比，并与多步防御trades比。（我实在是觉得创新性不足，但考虑到是20年的工作，就酱紫叭） |