|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  精读  **文章标题：**  Mind the Box: l1-APGD for Sparse Adversarial Attacks on Image Classifiers  **中文标题：**  注意盒子：L1-APGD实现对图像分类器的稀疏对抗攻击  **发表于：**  ICML2021  **作者：**  Francesco Croce，Matthias Hein  **单位：**  University of T¨ubingen  这俩哥们  20年是APGD L2 Linf  21年是APGD L1  22年是联合范数鲁棒性中的EAT（快速微调那个算法） | 第一：分析了在如何将获得的对抗样本投影到L1盒子与[0,1]的交集中，这里的主要讨论在Proposition2.1，我进行了一些分析：  8c1636d0ce90e9c4af8cab3e49a7b4ea2a22fcb50c821d19e75dafa67fbda2b0374d0cb6503340108419c7ddd09da749b99a35065047a1da99ab3cc5247d  第二：关于L1范数的最速下降方向：  （我认为这里的argmax应该是argmin）此前的攻击中，最速下降方向是酱紫给定的，于是  对于Linf范数：，L2范数：。    作者认为，这里仍然有一定的问题，在L1范数下应当考虑上面一行的优化问题，因此最速下降方向描述为Proposition2.2（但是实际上作者后续设计APGDL1时没有采用这个最速下降方向）：  对于一个给定的样本x，及回传的梯度g，对g进行排序，绝对值从大到小进行排列，优先更新梯度值绝对值最大的像素点，且对这个像素点的更新为（如果梯度大于0，则将像素值上升为1，如果梯度值小于0，则将像素值下降为0），这样按照梯度绝对值大小对样本x的像素点进行更新，每次噪声的一范数会为增加1-xi或xi，知道噪声的一范数增大为epsilon时（最后一次对第k大的像素值进行更新时，如果更新为0或1，会使总范数超过epsilon，此时减小更新的幅度，让总范数正好为epsilon），至此，正好添加了一范数为epsilon的噪声，且添加噪声后样本的像素值仍然保持在0-1之间。  第三：设计L1APGD攻击：   1. 关于算得梯度后，如何取得对抗性噪声：使用SLIDE，即求得梯度后，选择梯度绝对值最大的k个像素点进行更新，更新的大小均为epsilon/k，更新的方向根据梯度值的正负判定，这样保证了对抗性噪声的一范数为epsilon，但不能保证添加噪声后样本的像素值保证在0-1之间，此时使用Proposition2.1中提出的L1投影方法，就能获得噪声一范数为epsilon，像素值在0-1之间的对抗样本。 2. 总步数设计为100步，每100 \* 0.04步会自适应调整k与step\_size(单步更新时添加的噪声一范数大小)的值，调整的原则为文中的公式9，10 3. 关于multi-APGD是将100步分为30 30 40（三步之内，跪下当狗IMG_256），第一步（请客吃饭），取3\*epsilon作为投影时的一范数球，即在一个更大的一范数空间中进行启动，第二步（杀鸡取卵），取2\*epsilon，逐渐缩小一范数的空间，第三步（什么tmd叫惊喜），取epsilon作为投影时的一范数球，最终取得1范数为epsilon的对抗性噪声。   第四：设计AAL1攻击：  第一步，L1APGD + CE 100步，使用muti模式，随机重启五次。第二步，L1APGD + T-DLR 100步，使用muti模式，随机重启五次。第三步，FAB-attack L1版本。第四步，Square Attack (5000 queries)L1版本。 |
| **摘要及介绍** | **实验** |
| 我们表明，当同时考虑图像域[0,1]时，建立的l1-投影梯度下降（PGD）攻击是次优的，因为它们没有考虑有效威胁模型是l1-球和图像域[0,1]的交集。  我们研究了该有效威胁模型的最陡下降步长的期望稀疏性，证明了在该模型上的精确投影在计算上是可行的，并具有更好的性能。  此外，我们提出了一种自适应的PGD形式，它即使在很小的迭代预算下也是非常有效的。  我们得到的l1-APGD是一个强烈的白盒攻击，表明先前的工作高估了它们的l1-鲁棒性。  利用l1-APGD进行对抗性训练，得到了一个具有SOTA l1-鲁棒性的鲁棒分类器。  最后，我们将l1-APGD和l1方形攻击结合为l1-自动攻击，这是一个攻击集合，可靠地评估了l1-球和图像域[0,1]的交集威胁模型的对抗鲁棒性。 |  |