|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  精读  **文章标题：**  Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks  **中文标题：**  使用多种无参攻击的集合可靠地评估对抗鲁棒性  **发表于：**  ICML2020  **作者：**  Francesco Croce  **单位：**  University of T¨ubingen, Germany  这就是德国人严谨的学术土壤吗？ | 对PGD算法提出了两点改进，第一点将优化步长从固定的改为自适应的，第二点阐明了CE损失函数中的梯度弥散现象，因此改进了损失函数（跟我想的tmd一模一样，太卷了，随便想想别人就发过paper了）  第一点：PGD攻击原文中每次反向传递后，更新参数时使用作为更新步长，作者提出了改进：  其中记录的是在第几次迭代过程中进行判断。  第0次迭代作为一个简单的初始化，从中找出使损失函数最大的那个，设置为与fmax。  然后从k=1开始迭代，随时更新xmax与fmax，在时暂停迭代，然后判断两个条件，如果至少有一个条件成立，将值减半，并且回到此前记录的xmax，以xmax为起点继续迭代。  关于两个条件：  在两个检测点之间（从到），统计每次对图片进行更新后，损失函数是否上升，如果损失函数确实上升了，计数器加1。这个过程中迭代的次数是，损失函数上升的次数是计数器的值，如果梯度上升的次数小于某个上限，说明优化步长已经不合适了，要减半了。  在第j-1次检验时，两个条件都不成立，没有减半优化步长，即（），但是在第j次检验时发现损失函数的历史最大值并没有发生改变，这意味着从到这部分优化过程中并没有找到使损失函数更大的样本点，这时就要将优化步长减半了，因为This prevents getting stuck in potential cycles。此外7 8行是一种动量式的更新方法，原pgd的更新方法就是第7行，此处改为动量式更新，其中是本次梯度回传带来的变化量，是上次梯度回传带来的变化量，此次更新真正的变化量是本次变化量与上次变化量的加权平均，本次的变化量占比为。  第二点：原PGD的CE损失函数在回传时有梯度消失问题即，即原网络对地面标签y输出的概率值接近1，对其他类别输出的概率值接近0，那么通过CE函数的梯度我们明显可以发现，梯度中几乎每一项都乘以了一个接近0的数，这就是梯度消失（wocwocwoc跟我想得一模一样，早生三年，干倒急急光强）（这也是为什么PGD倾向于将样本攻击为次优类，即对抗样本的概率最大的类别是原始样本概率第二大的类别）。作者还进行了一个小实验来解释梯度消失现象带来的影响：横轴是计算CE时使用的参数的值，即logit值除以后再计算交叉熵损失。此时由于越小，logits值差距越大，softmax后分布越陡峭，对地面标签y输出的概率值越接近1。显然值越小梯度中等于0的元素个数越多，鲁棒准确度越高（攻击成功率越低）。  改进后的无目标损失函数Difference of Logits Ratio（logits值的差的比），旨在推进次优类逐渐变为最优类，在样本的分类结果还是y时，损失函数为负值且保持在[-1,0]，当分类结果不再是y时，损失函数变成正值，且数值不会过于离谱  ，分母改变是为了防止损失值变成一个恒定值，比如如果目标类就是，那损失函数是个定值1（==）。作者用第一点优化与CE损失函数获得无目标攻击APGDCE使用第一点优化和有目标版DLR损失获得有目标攻击APGDTDLR，此外作者还引入了FABT和Square Attack，**AA攻击指，一个样本只要被这四种攻击中的一种成功找到对抗样本，就认为攻击成功，此外这四种攻击几乎不需要任何超参数的调优，因此叫Autoattack** |
| **摘要及介绍** | **实验** |
| 在过去的几年里，对抗攻击的防御策略领域显著发展，但进展受到阻碍，因为对抗防御的评估往往是不够的，因此给人一种错误的稳健性印象。（大环境）  许多有希望的防御措施稍后可能会被打破，这使得评估那些是最先进的技术变得很困难。在评估中经常出现的缺陷是对攻击的超参数的调整不当、梯度模糊或掩蔽。（以往的问题）  在本文中，我们提出了PGD攻击的两种变体，克服由于次优步长和目标函数问题带来的失败。然后，我们将我们的新攻击与两个互补的现有攻击结合起来，形成一个无参数、计算价格实惠、用户独立的攻击集成，以测试对抗性的鲁棒性。（我们的方案）  我们将我们的集成应用到50多个最近发表在顶级机器学习和计算机视觉领域的论文模型中。在除一个案例之外的所有案例中，我们获得的鲁棒测试精度都比这些论文中报告的要低，通常超过10%，识别了几种失败的防御。（效果）  介绍：  第一段介绍了现有的一些攻击算法。  第二段说因为一些防御被轻易绕过，现在缺少一个对防御模型进行评估的合理体系，因此作者的目标是构建一个无参数化，可靠的自动的攻击用来统一评估防御模型。  第三段作者指出pgd的两个问题，并给出了本文的两个改进，形成pgd变体  第四段作者说将pgd变体与两种攻击FAB and Square Attack结合构成autoattack  第五段介绍实验结果  第六段作者指出，希望autoattack成为一个新的模型的一个小小的测试 | 实验很简单，第一个实验是在已有的各种防御算法上实施白盒攻击，统计AA中的4中攻击方法的单独性能和联合性能（即一个样本被任意一个算法功破即认为被攻破）  第二个实验是对具有一定随机性的方法进行数次梯度回传求平均值的方法进行攻击。  **有意思的是，本文最后暗示了有些文章的作者不愿意提供模型甚至源码，属于是防御领域中的扫黑风暴了（手动狗头）** |