## 摘要 ABSTRACT：

基于GAN的后门检测的诞生源于一系列“有趣的经验研究”[48]。为何全局性的对抗性网络能帮助逆向后门?在本文中我们结合互信息熵讨论了这个问题。在此基础上我们提出了一种基于对抗性生成网络和互信息熵的后门防御方法，用于对不同类型的机器学习后门攻击进行高精度的检测和后门触发器的还原。我们证明，使用中毒模型的数据集的熵进行对抗性生成网络的逆向，而不是基于固定像素[1]和或无约束GAN[48]的后门触发器的构建，这将使大的、多的、复杂的后门触发器更容易被检测和精确逆向。为此，我们首先结合已有的基于GAN的后门防御论证了中毒模型比正常模型具有更高的熵的特点。然后利用检测结果，对后门的互信息熵分布进行精细的统计，确定后门触发器所处的超平面并构建用于精细逆向的方向向量。最后进行针对不同类型后门触发器的精细逆向和检测。虽然GangSweep [48]、[]、[]已经使用GAN对后门触发器进行了一定程度的逆向和防御，但仍然没有克服TABOR [49]提出的后门扩散的问题。但同时TABOR本身基于Neural Cleanse后门像素较小的假设，显然无法适应GangSweep所提出的大像素后门的检测要求。同时基于GAN的逆向还面临着无法剔除自然后门影响问题、随着后门输入的复杂化逆向结果与后门触发器之间的视觉相似性迅速下降的问题。因此现在还没有一个较好的针对GAN防御的约束手段。我们在原有逆向的过程中加入了互信息熵筛选的辅助机制，并提出了一种简化运算的检测机制。在数据集MNIST、GTSRB、PUBFIG上的实验结果表明，我们的方法同时提升了原有方法的后门检测的准确度和检测效率、后门触发器逆向结果的视觉相似度、后门防御的效果。代码位于……。

动机：实验中发现现有方案虽然有效，但实际正确率和视觉效果都不理想，提出更加高效精准的约束手段以完善逆向方案

优点：逆向生成的触发器更加精确，有助于消除后门；后门检测结果更加精确、效率更高；简化的方法可以使后门检测的运算量减少一个量级。