

**Ero S&P 2020 — Reza Shokri**

**一、Abstract**

本文提出了一种可以绕过最先进的后门检测的后门攻击（Backdoor Attack，即敌手可以修改训练数据和模型参数，将后门（trigger）嵌入到模型中。如果输入包含后门特征(如图像上的戳记)，则模型的行为将根据敌手的目标而定。然而，中毒模型对clean data的行为保持不变）。

**二、Introduction**

最新的检测方法通过比较trigger input与clean input之间的潜在表示（特征）差异来识别后门。为了绕过检测，本文设计了一种对抗性后门嵌入算法用于深度学习，最大限度地提高了trigger input和clean input的潜在不可分辨性。

该算法利用正则化的方法，对损失函数进行优化。

**三、Backdoor Defence: Detection Algorithms**

1）特征剪裁**（Feature Pruning）**

隐藏层中每个神经元n在clean input集上的平均激活值为，在trigger input集上的平均激活值为。该优化函数要找对trigger inputs敏感的点，即与clean inputs的平均激活值差距最大的点。

“剪裁”的意思是指将上述优化找到的神经元从神经网络中删除。

2）通过光谱签名进行数据集过滤**（Dataset Filtering by Spectral Signatures）**

首先，使用中毒的训练数据集对网络进行训练。对于每个特定的输出类label，该label的所有输入实例都通过网络进行输入，并记录它们的潜在表示。然后对潜在表示的协方差矩阵进行奇异值分解，用于计算每个输入的离群值。离群值最高的输入被标记为有毒输入，并从训练数据集中删除。

3）通过激活聚类过滤数据集**（Dataset Filtering by Activation Clustering）**

利用独立分量分析进行降维，将输入的潜在表示减少到10-15个特征，然后进行k-means聚类，将转换后的数据分离成2个聚类。这种情况下，trigger input和clean input的潜在表示分别形成单独的聚类。

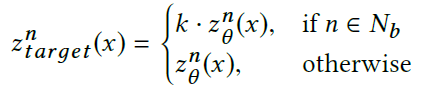
**四、Bypassing Defenses: Adversarial Backdoor Embedding**

上述防御没有考虑到：一个老练的攻击者能够通过最小化潜在表示的差异来使模型对它们具有健壮性。

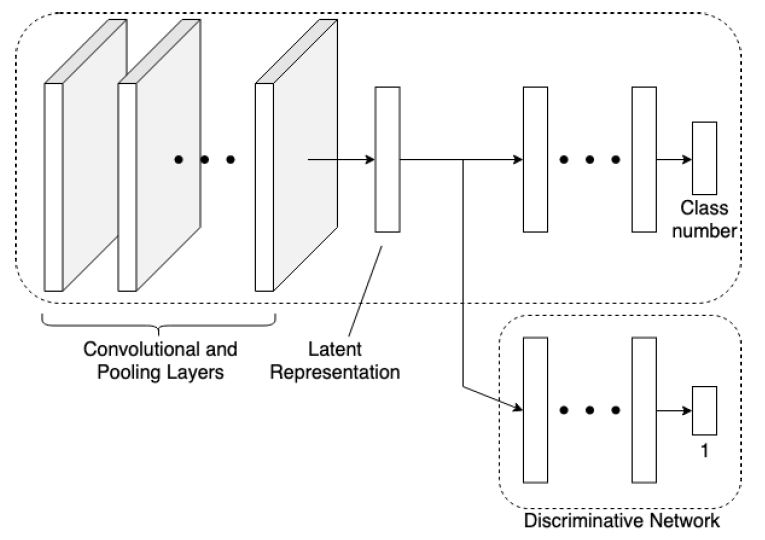
为此，我们在目标模型的损失函数中引入了一个惩罚项：

是输入，是目标标签。是的潜在表示，会成惩罚模型当clean input与trigger input之间的潜在表现差异过大时。通过双目标函数，攻击者的目标是实现模型的高分类精度，同时对输入的潜在表示形式设置一定的约束，以绕过潜在防御。

这个额外的惩罚项可以是特定于某种防御的，即可根据攻击者预期的特定防御进行调整；也可以是减轻各种防御的general惩罚，下面将分别说明这两种情况：

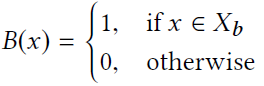
1. **特定于“特征剪裁”防御的惩罚项**

表示正常模型的参数，为已注入后门的模型的参数，表示嵌入后门的神经元。如果神经元n为trigger神经元，那么，为了绕过检测，引入k（0<k<1）来调节的大小，使其与之间的差距越小越好；如果为clean神经元，则保持不变即可。

1. **General的惩罚项**

上图为本文提出的后门攻击的基本框架，input经过卷积与池化层H之后，得到潜在表示（特征），该潜在表示有两个用途：1、继续训练模型C，得到output（）；2、训练辨别器D（辨别器输出为=1或0，分别表示“trigger input”和“clean input”）。

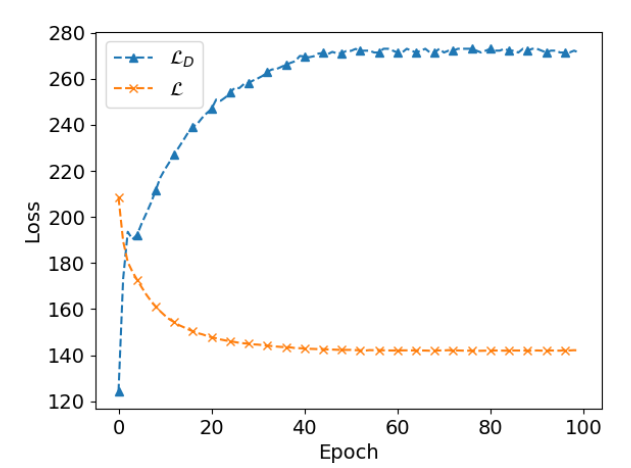
因此，目标函数设计为：

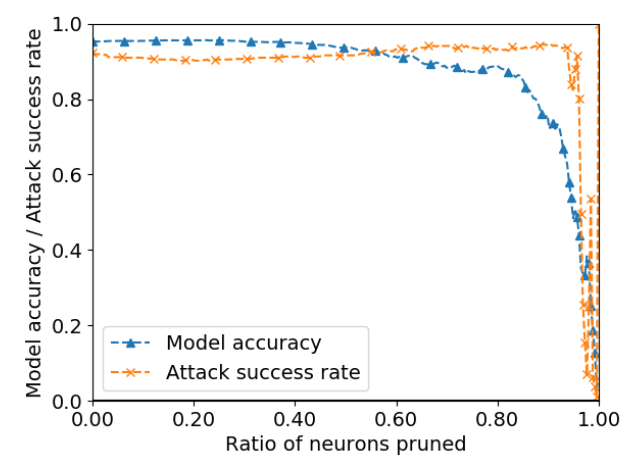
其中，表示辨别器的损失函数。

表示trigger input。如果x为trigger input，则B(x)=1，否则为0。

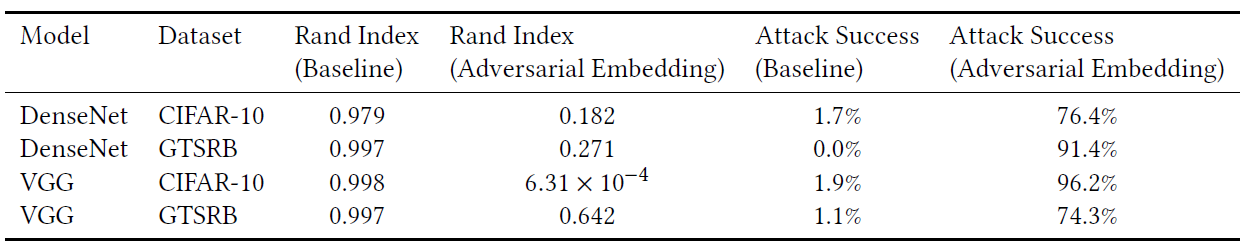
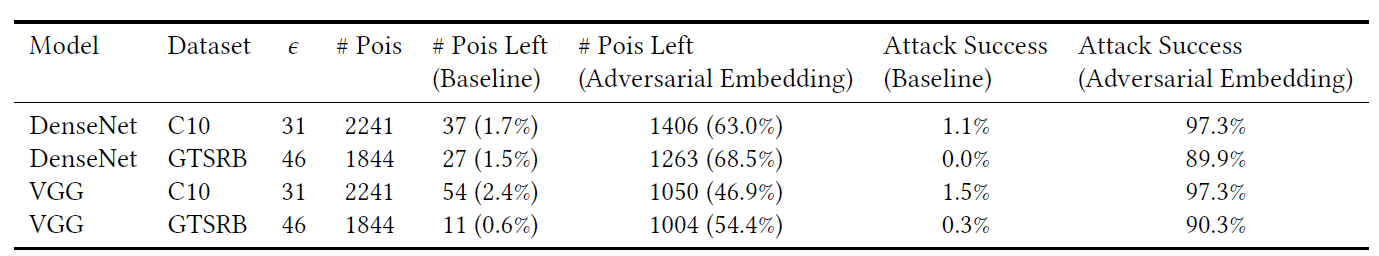
当越小时，式（5）越大，即损失函数越大，越要进行训练优化，直至（5）变小，即变大且变小时，停止训练。这样就可满足“通过双目标函数，攻击者的目标是实现模型的高分类精度，同时对输入的潜在表示形式设置一定的约束，以绕过潜在防御”这两个目的。

**五、Experiment**



随着epoch的增加，辨别器的损失越来越大，分类器的损失越来越小，目标函数（5）可以收敛到一个最优值。

对于防御1（特征剪裁）的攻击效果，很好。

对于防御2的攻击效果，很好。

对于防御3的攻击效果，很好。