Neural Cleanse: Identifying and Mitigating

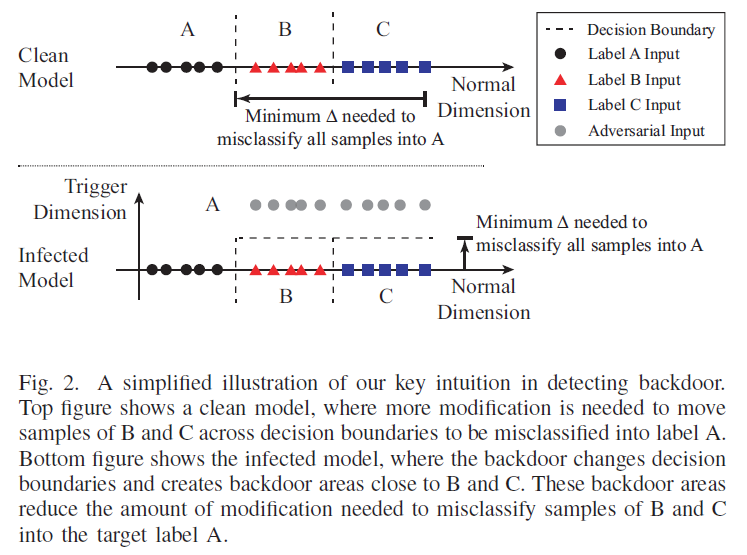
Backdoor Attacks in Neural Networks

**0 Abstract**

本文提出第一个对后门攻击（backdoor attack）的检测，能够识别和重构trigger。实验采用现存的两种攻击方法（1、BadNets：training time注入triggered dataset；2、Torjan attack：对应分享文章0415\_NDSS\_2018），并提出多个缓解方法，包括input filter，neuron pruning，unlearning。同时证明了对后门攻击变体的robustness。

**1 Intuition**

后门攻击的基本思想为：论文分享0416\_IJCAI\_2019已经介绍过。Trigger能够令模型分类空间产生另外一个作为捷径的trigger维度，使得对输入进行很小的改变即可被分类为target label。但是将任一label的数据被分类为non-target label 却需要很大的改变。



本文通过检测模型所有的labels，判定是否某一类只需要对所有输入添加很小的扰动（trigger需要修改的像素点的量），就可以将其分类为target label。包含以下三个步骤：

Step 1：

对于每一类，设计优化方案来重构minimal trigger。

Step 2：

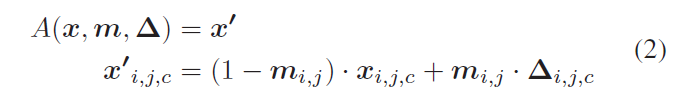
对所有类重复上述操作，得到N个可能的triggers。

Step 3：

计算每个trigger的大小，挑选最小的outlier，即为被后门模型的trigger。

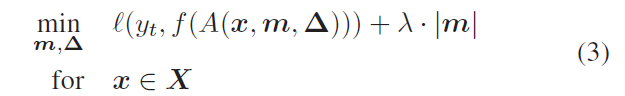
**3 Method**

**trigger injection**



上式为trigger图像的生成。x为原始图像，delta为trigger pattern（与x具有相同大小）。m为mask，元素取值为0至1，1表示完全取trigger的像素，0表示完全取原始图像x的像素。

**Objective function**



第一项为带有trigger的图像被分类为target label（y\_t）的损失。第二项为trigger的大小，文中用mask的L1 norm来测量，即为被修改的像素点的大小。最终目标为，得到trigger既能使输入误分类为target label，又只改变原始图像很少的像素点。

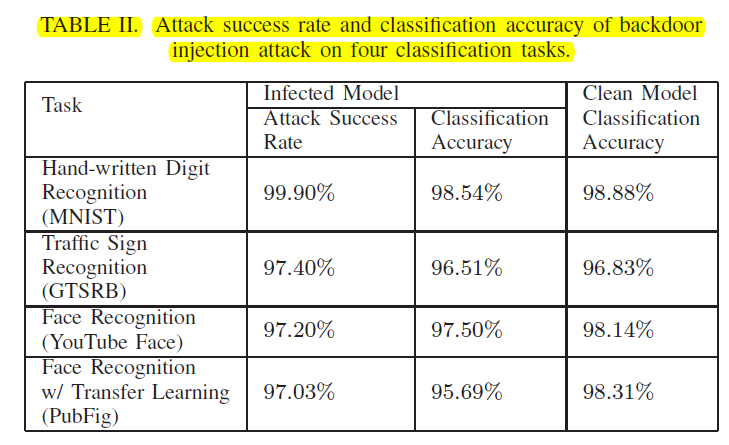
**Outlier Detection**

比较所有类得到的triggers的mask的L1 norm，最小的L1即为trigger。本文采用Median Absolute Deviation（古老的统计方法）计算anomaly index。L1 norm越小，anomaly index越大，即越有可能为模型后门的target label类。文中发现anomaly index大于2为被backdoor的模型类别。

**4 Experimental**

**Attack**

**本文复现两种攻击：BadNets和Torjan attack。下图为BadNets在不同数据集的攻击效果：**

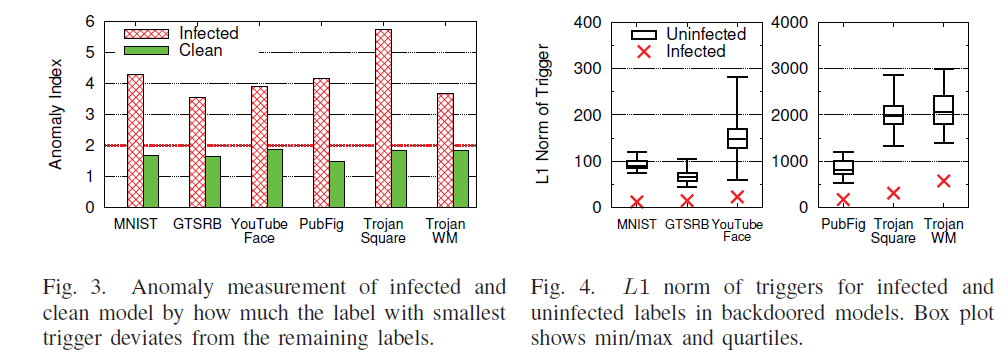


**可以看出后门攻击成功率大于97%，分类精确度下降不超过3%。**

**另外，Torjan attack攻击的成功率在两个模型上Trojan Square and Trojan Watermark**分别为99%和97%。

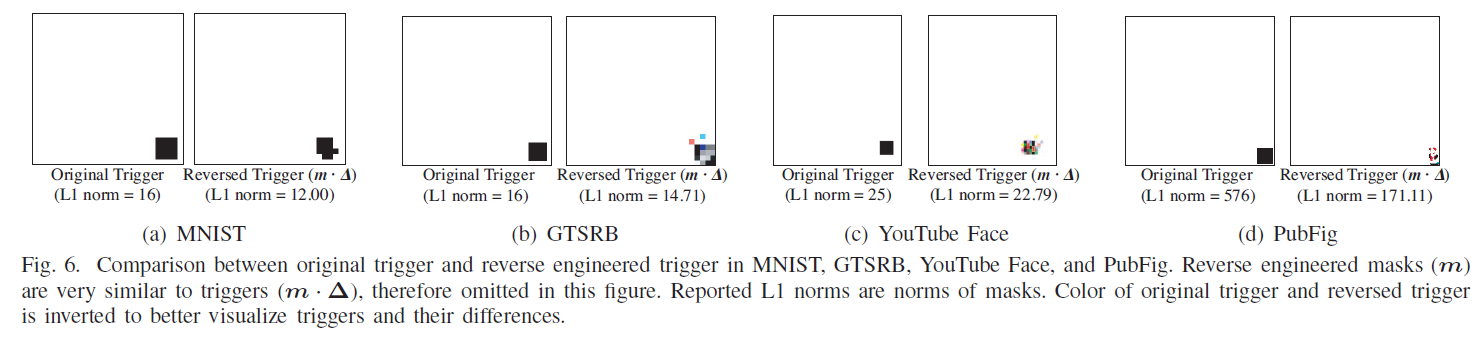
**Detection**

**下图显示了被backdoor的label具有更高的**anomaly index和更低的L1 norm，相比于clean的label。



**Similarity**

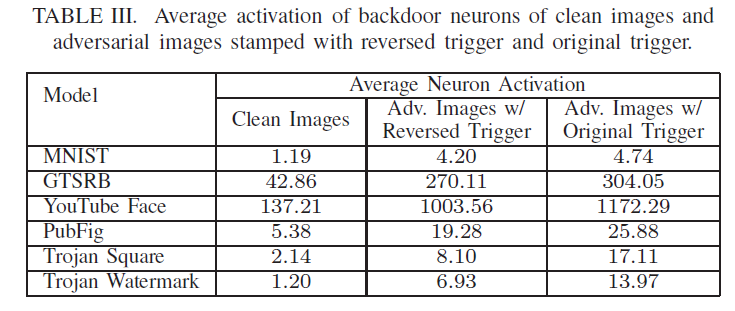
**下图实验显示了BadNets方法中，****original trigger 和reversed trigger的相似性，可以看出具有很强的相似性。**



**下图显示了Torjan attack方法中，****original trigger 和reversed trigger的相似性，可以看出有很大差别。原因是，该后门攻击方法是fine-tuning模型，使其最大化激活某个神经元。**

**Neuron Activations**

**下图显示了original trigger 和reversed trigger对倒数第二层神经元的激活程度，可以看出两种trigger都可以较大地激活目标神经元，使得对输入误分类为target label。**



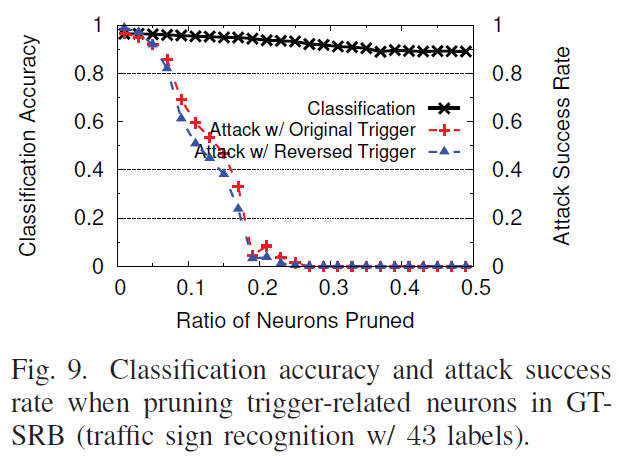
**5 Mitigating**

1）Filter for Detecting Adversarial Inputs

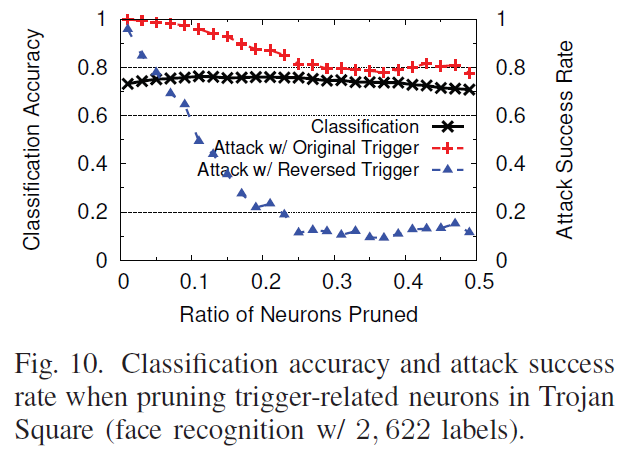
本文利用特定神经元激活值（上面实验选择的倒数第二层的神经元）的方法，设置门限阈值，高于阈值的激活值被认定为triggered input，进而filter掉。

2）Patching DNN via Neuron Pruning

将具有高激活值的神经元pruning掉，即在预测过程中，设置高激活值神经元的值为0。实验表示在BadNets中，该方法可以有效地控制攻击成功率，同时几乎不影响精确度，如下图：

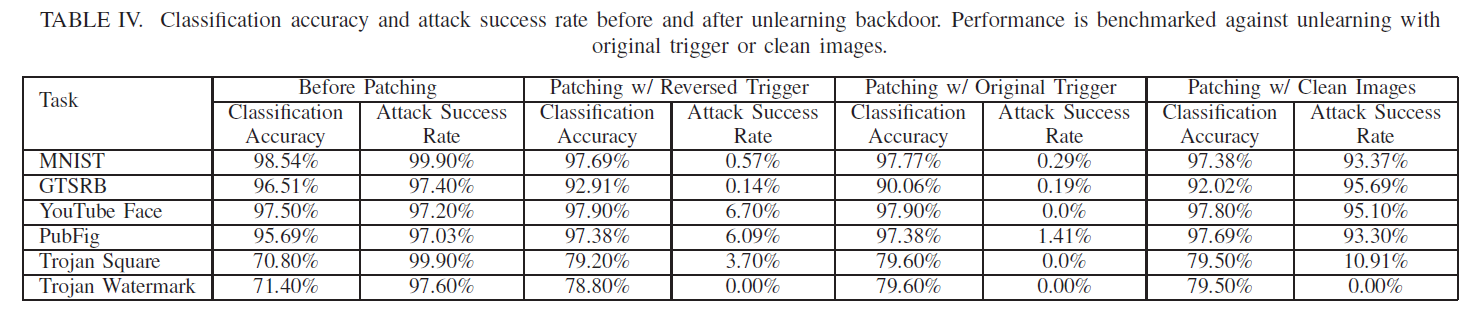


但是对于Torjan attack没有效果，如下图：



3）Patching DNNs via Unlearning

使用带有trigger的图像+正确label、原始图像+正确label进行fine-tuning模型，让模型忘记之前学习到的trigger和target label之间的映射。实验结果如下表，可以很好的防御后门攻击：



**6 Robustness**

本文又测试了对不同的攻击设置的robustness。包括

1. Complex Triggers：即复杂的trigger图案
2. Larger Triggers：像素点更多的trigger
3. Multiple Infected Labels with Separate Triggers：不同的trigger后门攻击多个不同的target label
4. Single Infected Label with Multiple Triggers：不同的trigger后门攻击同一个target label
5. Partial Backdoors：没有完全学习到的backdoor，即只对部分数据其作用。

但是实验效果不是很好，知识探索性地完全实验。

**7 Conclusion**

个人观点：非常感叹作者的知识储备和工作量，实验以及方法都非常全面。文章有一些局限的地方：1）需要模型的训练集才能进行防御，real-world可能只有这个backdoor的模型，没有训练集。2）防御方法假设trigger的pattern和position都是固定的，现有的攻击已经做到了dynamic trigger。