《Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering》

文章的主要思想是：

本文提出了一种激活聚类（AC）方法来检测在 DNN 中插入后门的有毒训练样本。该方法通过分析训练数据的神经网络激活情况来判断训练数据是否中毒，如果中毒，哪些数据点有毒。

文章贡献：

1. 提出了第一个 不需要可信数据集作为验证集的，对注入了后门的毒化数据做检测的方法。
2. 通过在三个不同的文本和图像数据集上进行评估，证明了 AC 方法在检测不同应用中的有毒数据方面非常成功。
3. 证明了 AC 方法对于复杂的中毒场景是鲁棒的，在这种场景中，类是多模态的（例如包含子种群），并且插入了多个后门。

AC（Activation Clustering）激活聚类：

思想：

1. 尽管后本样本和目标样本被赋予了相同的分类结果，但二者获得分类结果的原因是不同的。
2. 对于标准样本，DNN 识别的是输入与目标类相关联的特征；对于后本样本，DNN 识别的是：该后门样本原本的类别特征 与 后门触发器的特征，这两个特征的叠加导致了后本样本被分类为攻击者的目标类。
3. 故，本文考察 DNN 最后一个隐藏层的激活状态，并对其进行聚类。

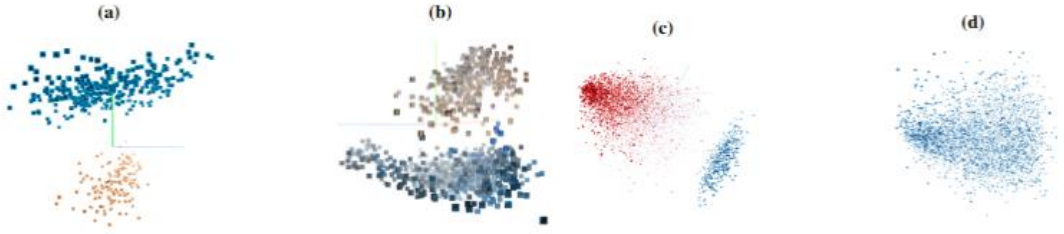

Figure 2: Activations of the last hidden layer projected onto the first 3 principle components. Activations of the last hidden layer projected onto the first 3 principle components. (a) Activations of images labeled 6. (b) Activations of images labeled as speed limits. (c) Activations of the (poisoned) negative reviews class (d) Activations of the (unpoisoned) positive review class.

核心算法：

1. 首先，使用可能包含有毒样本的不可信数据训练神经网络。
2. 然后，使用训练数据查询网络，并保留最后隐藏层的结果激活。
3. 分析最后一个隐藏层的激活足以检测到毒物。

流程：

激活->一维矢量->独立成分分析（ICA）进行维度缩减->聚类->确定哪些簇中存在毒害

Input: untrusted training dataset D_p with class labels $\{1, ..., n\}$

Train DNN F_{Θ_P} using D_p
Initialize A; $A[i]$ holds activations for all $s_i \in D_p$ such that $F_{\Theta_P}(s_i) = i$
for all $s \in D_p$ **do**
    $A_s \leftarrow$ activations of last hidden layer of F_{Θ_P} flattened into a single 1D vector
    Append A_s to $A[F_{\Theta_P}(s)]$
end for
for all $i = 0$ **to** n **do**
    red = reduceDimensions($A[i]$)
    clusters = clusteringMethod(red)
    analyzeForPoison(clusters)
end for

Algorithm 1: Backdoor Detection Activation Clustering Algorithm

1. 其中聚类部分采用 K-means。 相比 DBSCAN、高斯混合模型和亲和传播，K-means 在速度和精度上更为有效，但是 K-means 不管是否存在投毒，都会将激活分为两个簇。
2. 确定毒害簇的方法：
    a) 排除性的重新分类：
        i. 剔除数据集中对应集群的数据，并训练一个新的模型。并利用新模型对已删除的集群进行分类。
        ii. 如果一个集群包含合法数据的激活，则相应的数据将在很大程度上被分类为它的标签；如果一个集群包含有毒数据的激活，那么模型将主要把数据分类为源类。
    b) 相对尺寸比较：
        i. 实验结果表明，有毒数据的激活在 99%以上的情况下总是被放在与合法数据不同的聚类中。
        ii. 当带有给定标签的数据中有 p%发生中毒时，预计得到的聚类结果为：一个群集包含大约 p%的数据，而另一个群集包含大约（100-p）%的数据。相反，当数据无毒时，发现激活趋向于分成大小相等的两个簇。
        iii. 当期望一个给定标签中不超过 p%的数据会被对手毒害时，如果一个集群包含≤p%的数据，我们就可以认为它被毒害了。
    c) 轮廓分数（silhouette score）：
        i. Fig2(c)(d)表明，中毒的情况下，两个聚类可以更好的描述激活，反之则是一个聚类。
        ii. 可以使用评估集群数量与激活的匹配程度的指标来确定相应的数据是否中毒。
        iii. 利用 silhouette score。silhouette score 较低表明集群数量与激活的拟合不佳，则认定为没有毒害；silhouette score 较高，则认定较小的集群被毒害。
3. 快速分析（预先快速判断是否存在投毒）：

为每一个群集构造“精灵图像（sprite image）”，并对簇中的激活的图像进行平均。人眼对 sprite image 进行快速分析。

其中：

sprite 图像是通过将与所讨论的簇相关联的每个图像重新缩放到一个小尺寸并构建由重新缩放的图像组成的马赛克来生成的图像。

4. 后门修复：
   a) 删除有害数据，重新训练模型。
   b) 对有害数据重新标记，利用标记后的这部分数据对模型再训练，直到模型收敛。

文章特点：

1. 训练后检测。
2. 针对 DNN 最后一个隐藏层神经元的激活状态进行聚类。
3. 文章的聚类方法为 k-means，并提出了几种分析方法，解决了 k-means 的不足。