



LEMNA: Explaining Deep Learning based Security Applications

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/8720704c-16b1-423d-ab43-aee25b3a50dd/ccs182d3.pdf>

当前论文引出的问题

深度学习中的神经网络模型一般会比较复杂，其神经元数量提高的同时增加了其对测试集中数据测试的准确性，同时也降低了模型的可解释性，因为其缺少了一定的解释性，并成为了关键性的阻碍。造成安全从业人员犹豫把深度学习使用在安全领域之中，从而导致了深度学习在安全领域的运用率不高。

目前深度学习在安全领域的应用集中在

- 恶意软件检测
- 二进制逆向工程
- 网络入侵检测

提出的解决方案

目前在深度学习领域中，可解释性方案的研究大部分集中在用于图片分类和NLP的卷积神经网络（CNN）中，常见的方法有在前向传播或后向传播中寻找重要特征来寻找该模型的解释性。但是在安全领域中，像二进制或者逆向工程等等，要么代码中具有高度的依赖性，要么具有高度的可伸缩性，导致安全领域中倾向于使用递归神经网络，或者多层感知机模型。而有关机器学习的可解释性研究着眼点几乎不在RNN上，或者说没有表现良好的。针对上述的问题该论文作者提出了一个解决方案叫LEMNA（Explaining Deep Learning based Security Applications）。该方案就是为了能够解决当前深度学习在安全领域应用的不可解释性，尤其针对RNN模型。

决策流程

将输入数据实例X和分类器模型（如RNN）输入到LEMNS中，使用简单模型来解释复杂的深度学习的决策边界，即使线性模型来推断重要的决策特征来近似，并逼近当前局部的决策边界。从而得出当前决策模型的可解释特征，来解释该模型如何分类的。

1. 致力于判断出一个小集合的对X分类的关键性的特征（缩减特征的数量）
2. 通过生成在X附近的决策边界的近似值
3. 并且不假定X的特征相互独立，也不假定样例是线性可分的（其他可解释方法违反了这个）
4. 是一个新的基于融合套索的混合回归模型来近似非线性边界。（融合套索将特征作为，一组并抓住相邻特征，臆测就是缩减特征数目，去掉相同或者相似的）

特点

1. 模型专为解决依赖性（代码间的依赖性）来更好的应用于安全领域的决策模型之中。
2. 理论上只要给予足够数据，混合回归模型就可以估计线性和非线性的决策边界。
3. 处理非线性边界来提高解释的保真度，真实度。

工作

1. 评估来提出的该系统在两个流行的安全软件（恶意软件分类器，二进制逆向工程的功能启动检测器）
2. 还开发出了一系列的指标来评估该方法。

3. 演示了LEMNA的用例来帮开发者验证决策模型的行为，并且修正模型判断时产生的错误

LEMNA的表现

比现有在安全领域有关机器学习解释性的方法更具可解释性
还能够

与其他方式的差异

1. 与使用对抗性样本来增强决策模型的健壮性不同，本论文是尽可能的在方法中理解模型的分类错误产生原因来提高模型的健壮性
2. 与删除不良或错误数据来修补模型不同，本论文是通过增加训练数据来修复训练不足的组件。



