**Title** : LEMNA : Explaining Deep Learning based Security Application

**Abstract** :

background : deep learning has shown a great potential in various domains,

problem : the lack of transparency has limited its application in security or safety-critical areas.

state of arts : existing research develop explanation tech to provide interpretable explanations for each classification decision, but current methods are optimized for non-security tasks(e.g., image analysis).

contribution : propose LEMNA : a high-fidelity explanation method dedicated for security applications.

LEMNA : given an input data sample, LEMNA generates a small set of interpretable features to explain how the input sample is classified.

idea : the core idea is to approximate a local area of the complex deep learning decision boundary using a sample interpretable model.

local interpretable model : (1) handle feature dependency to better work with security application(e.g., binary code analysis); (2) handle nonlinear local boundaries to boost explanation fidelity.

evaluation : validate model behavior, troubleshoot classification errors, auto patch error if target model.

**1 Introduction**

DNN are used to built security applications ( malware classification, binary reverse-engineering, network intrusion detection )

high-accuracy ->lack of transparency.

why : the high complexity of the DNN leads to low interpretability of the model, it is very difficult to understand how DNN make certain decisions.

state of arts :

researchers start to work on explanation method to interpret the classification result. image analysis : given an image, explains the classification result by pointing the most impactful features to the final decision, running forward propagation or backward propagation in the network to infer features. more advanced methods produce explanations under a black box setting where no knowledge of classifier details is available, approximate the local decision boundary using a linear model to infer the important features.

existing methods are not applicable for security applications, (1) existing methods are designed for image analysis(CNN), but binary reverse-engineering and malware analysis use RNN, so far there is no explanation method working well on RNN. (2)existing methods are suffer from a low explanation fidelity. acceptable for image analysis but can cause serious troubles in security applications. incorrectly highlighting one byte of code may lead to serious misunderstanding or interpretation errors.

design

develop a novel, high-fidelity explanation method dedicated for security application.

setting : black-box, specialized designs to address the above challenges.

goal : given an input data instance X and a classifier such as an RNN, identify a small set of features that have key contributions to the classification of X.

design : generating a local approximation of the target classifier’s decision boundary near X, introduce a new approach to approximate non-linear local boundaries based on a mixture regression model enhanced by fused lasso. (no longer assume local detection boundary is linear and the feature is independent , these 2 assumptions are violated in security application)

reason of decision : (1)a mixture regression model can approximate both linear and nonlinear decision boundary, thus can optimize the local approximation for a non-linear boundary and avoid big fitting (2)”fused lasso ” is a penalty term used for capturing feature dependency.