Towards a Deep Learning-Based Expert System for Detecting Phishing Attacks

Samaneh Mahdavifar, Ali A. Ghorbani

Canadian Institute for Cybersecurity (CIC), University of New Brunswick (UNB)

ABSTRACT

Cutting edge deep learning techniques have been widely applied to the areas like image processing and speech recognition so far. Likewise, recently several deep learning models have been employed in the area of Cybersecurity. In this paper, we study the application of Deep Learning in Cybersecurity and several deep learning models that have been applied to the areas of intrusion detection and malware detection/classification in literature. In order to overcome common shortcoming in related work, *i.e.,* lack of inner explanation, we propose a Deep Learning Expert System based on MACIE, a medical diagnostic system developed in mid-1980s, that enables us to extract refined rules from a trained Feed-Forward Neural Network. We evaluate our approach on Phishing Websites Dataset which is publicly available on UCI Machine Learning Repository. The experimental results show that the extracted rules are self-explanatory and good enough to substitute the trained Deep Learning model to classify unseen samples.

RELATED WORK

Paper	Focus Area	Model	Features	Dataset
MtNet Huang et. al. 2016	Binary and 100-Class Family Malware Classifier	Feed-Forward Neural Network using ReLU activation function and dropout for hidden layers	50,000 dynamic features of null-terminated tokens, API events plus parameter value, and API trigrams	2.85 million samples were extracted from malicious files and 3.65 million samples from benign files by analysts from Microsoft
DeepAM Ye et. al. 2017	Static Malware Detection	A heterogeneous deep learning framework based on AutoEncoder stacked up with multilayer RBMS and a layer of associative memory	API calls extracted from PE files	Comodo Cloud Security Center containing 4500 malware files, 4500 benign files, and 10,000 newly collected unlabeled files, and 1000 testing samples
Nauman et. al. 2017	Static Android Malware Detection	Fully connected NNs, CNNs, Autoencoders, DBNs, and RNNs in the form of LSTM models	Permissions, Intents filtered by the target, activity list in its manifest, API calls raised in the code, services registered	Combination of Drebin and VirusShare Datsets, including one and half million samples
MaldoZer Karbab et. al. 2018	Static Android Malware Detection and Attribution	Convolutional Neural Network with a convolution layer with Relu, maxpooling layer, and a fully connected layer as inner layers	API method Calls	20K malware samples from 32 Malware families collected from Malgenome and Drebin Datasets
Farahnakian et. al., 2018	Binary and Multi-Classification Intrusion Detection	Deep Auto-Encoder with softmax classifier on top of it	41 features numeralized to 117 features	KDD-CUP99 including 494k samples for training and about 300k for testing.

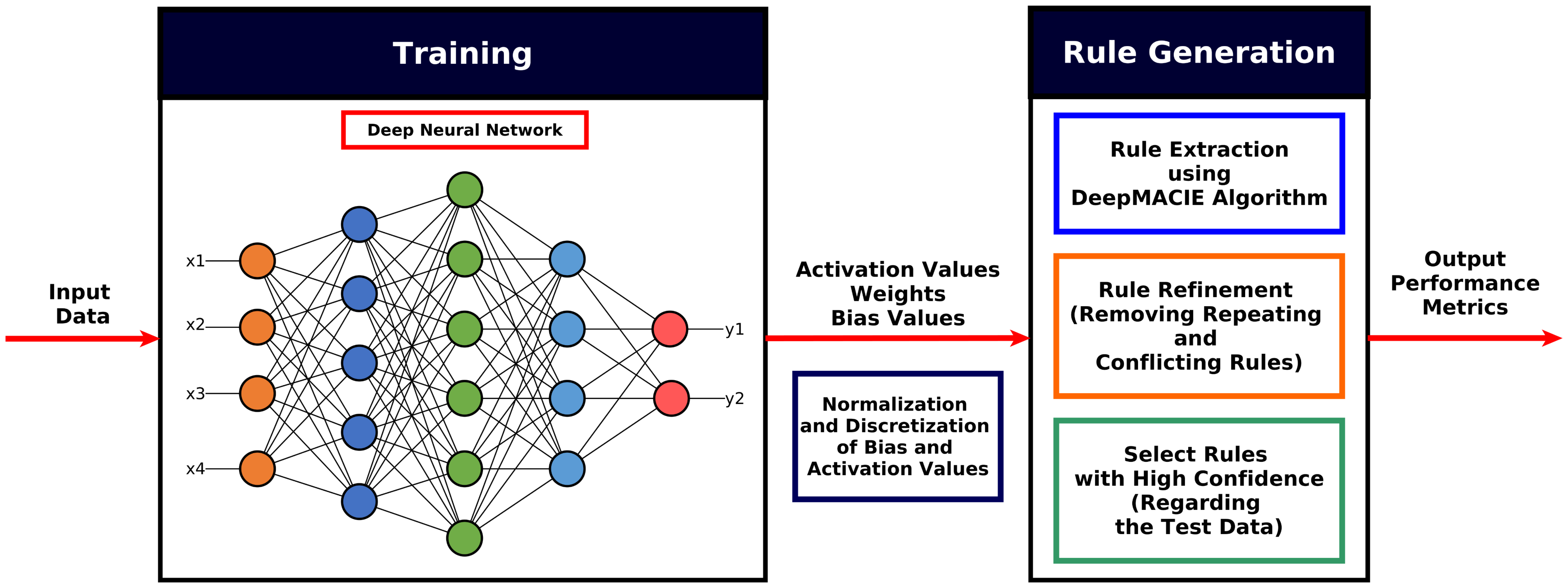
SHORTCOMINGS

- The deep models are not fine-tuned accurately, *i.e.*, number of hidden layers, hidden units, learning rate, etc.
- Not enough measures are computed
- The dataset is normally out-of-date, having small-sized samples, and non-diverse leading to overfitting problem
- It is not formally justified why these deep models are used
- Deep models are mostly used just for feature extraction not marking
- The proposed deep models **can not explain the logic behind the decision that they make**

GOALS

- Designing a Deep Learning-Based Expert System, namely, DeepMACIE, to overcome the problem of lack of inner explanations in Deep Neural Networks
- Extracting refined rules from a trained Feed-Forward Neural Network to substitute the Deep Learning model for classifying unseen samples in Cybersecurity domain

PROPOSED FRAMEWORK


Input Data
Training
Deep Neural Network
x1
x2
x3
x4
y1
y2
Activation Values
Weights
Bias Values
Normalization and Discretization of Bias and Activation Values
Rule Generation
Rule Extraction using DeepMACIE Algorithm
Rule Refinement (Removing Repeating and Conflicting Rules)
Select Rules with High Confidence (Regarding the Test Data)
Output Performance Metrics

EXPERIMENTS

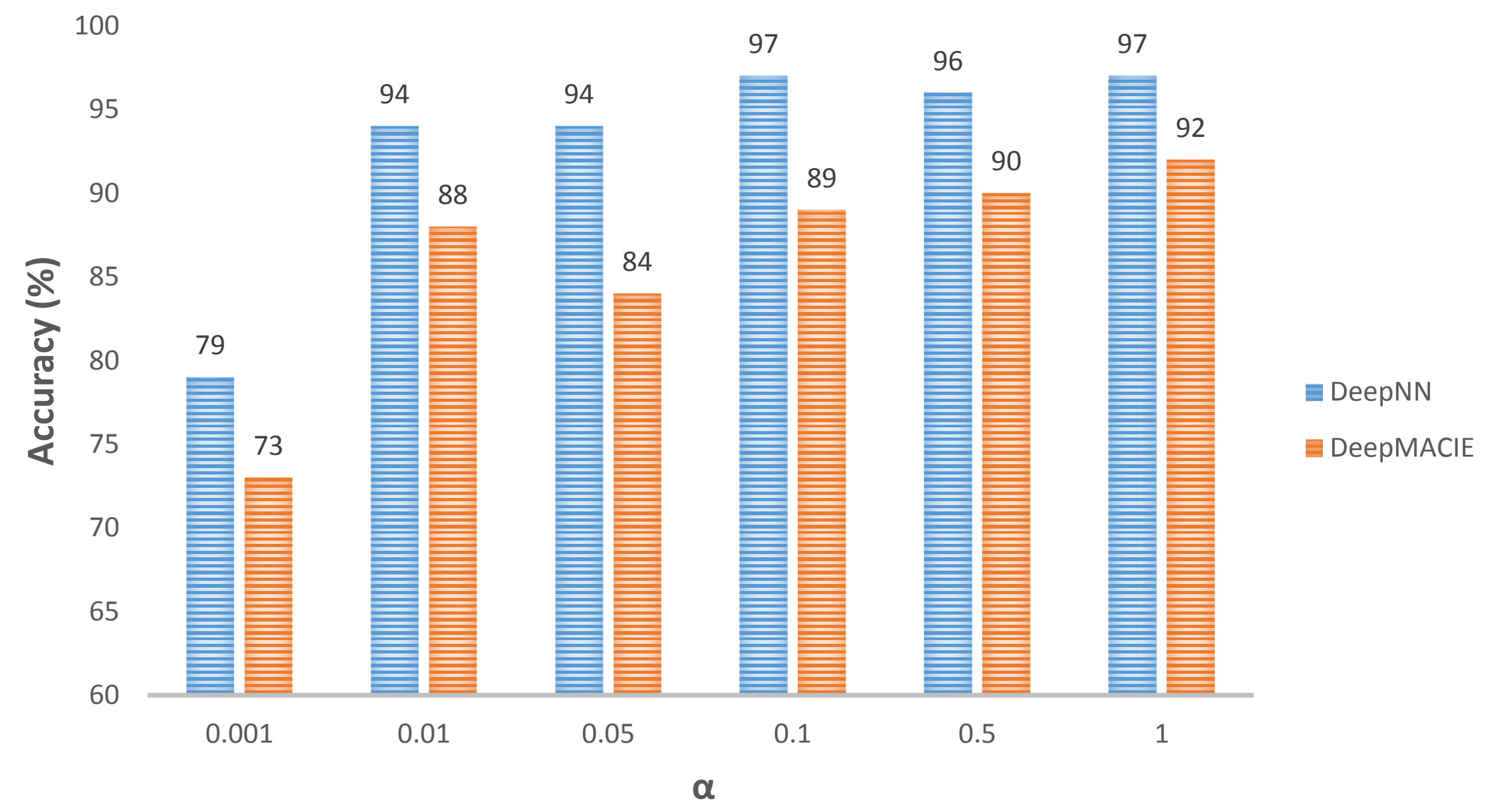
Dataset

- ❖ Phishing Websites Dataset, publicly available on UCI Repository
- ❖ 11,055 total samples (6,157 Legitimate and 4,898 Phishing)
- ❖ 30 features of 3 integer values (-1,0,1)

Results

- ❖ 5-fold cross-validation
- ❖ 4-layer Deep Neural Network (30,15,5,2)
- ❖ Adam Optimization Algorithm
- ❖ Optimized learning rate and momentum
- ❖ $\alpha = 1, \beta_1=0.9, \beta_2=0.999$

Method	ACC	FPR	F_1
DeepMACIE	0.92	0.001	0.92
DeepNN	0.97	0.006	0.98
DT (J48)	0.95	0.046	0.95



- Comparing DeepMACIE with Deep Neural Network, DeepNN, and J48, DeepMACIE achieves the lowest false positive rate and an acceptable accuracy of 92%
- The extracted rules are self-explanatory and accurately describe the relation between the output decision and the input features
- The DeepMACIE algorithm is capable of extracting well-defined rules even in existence of unknown or missing data

CONCLUSION

futurework

- Devise an algorithm to further refine the extracted rules that satisfy two main criteria:
  - Validity
  - Maximal generality
- Propose an automated rule debugging system using counter-examples
- Improve the training cycle of a Neural Network by changing its topology