LEMNA: Explaining Deep Learning based Security Applications

Presenter: Wenbo Guo
Dongliang Mu, Jun Xu, Purui Su, Gang Wang, Xinyu Xing
The Success of Deep Learning in Cybersecurity

- Deep learning becomes increasingly popular in cybersecurity
  - Binary reverse-engineering ([USENIX'15; USENIX'17; CCS'17])
  - Malware classification ([KDD'17; European S&P'17])

- Deep learning outperforms traditional approaches in some security applications
  - E.g., Binary function start identification ([USENIX'15])
    
    ```
    nop; nop; nop; nop; sub esp, 0x4c;
    90 90 90 90 83  ec  4c
    ```

    |                | ELF x86 |
    |----------------|---------|
    | P   | R     | F1    |
    | ByteWeight | 98.41% | 97.94% | 98.17% |
    | RNN     | 99.56% | 99.06% | 99.31% |

  - Recurrent Neural Network (RNN)
    
    ```
    0.01 0.01 0.01 0.01 0.99 0.01 0.01
    ```
The Concerns of Opaque Deep Learning Model

- Hard to establish a trust for the model behavior
  - Adversarial samples are unexplainable
- Difficult to diagnose the model mistakes
  - Errors of the model are hard to be corrected
- Unlikely to discover the model biases
  - Racial bias of machine learning based criminal prediction system ("Machine Bias" essay)

Presenter – Wenbo Guo
(http://www.personal.psu.edu/wzg13/)
Explanation & Interpretation of Deep Learning Model

- Definition of explanations: a set of features that make key contributions to the model result
  - Image recognition: a group of important pixels

- Binary function start: Important hexes

```
nop;  nop;  nop;  nop;  sub esp, 0x4c;
         90  90  90  90  83  ec  4c
```

Heat map
Existing Explanation Techniques & Limitations

• Mainly proposed in machine learning community
  • ECCV’14; ICLR’17; ICCV’17; ICML’17; KDD’16; KDD’16; NIPS’17

• Designed for Convolutional Neural Network (CNN) based vision applications

• Not suitable for security applications because
  • Limitation 1: Low accuracy in model explanations
    • Explanation of security applications requires higher accuracy than that of image and text recognition
  • Limitation 2: Not suitable for Recurrent Neural Network (RNN)
    • Deep learning based security applications are built on sequential models

Presenter – Wenbo Guo
(http://www.personal.psu.edu/wzg13/)
One Example of Model Explanation (LIME, KDD’16)

$g(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots$

Highest coefficients attaching to the 3 most important features

Presenter – Wenbo Guo
(http://www.personal.psu.edu/wzg13/)
Limitation of Existing Explanation Techniques

- Difficult to provide an accurate explanation
  - Existing approaches assume the local linearity
  - This assumption will cause errors

- Not suitable for an RNN
  - Existing approaches assume feature independency
  - Can’t model the feature dependency
  - Derive incorrect explanations

Feature independence
Our system design!
LEMNA: **Local Explanation Method using Nonlinear Approximation**

- Explanation technique designed for security applications
  - Providing accurate explanations
    - Supporting locally non-linear decision boundaries
  - Being customized to RNN
    - Modeling the feature dependency

Presenter – Wenbo Guo
(http://www.personal.psu.edu/wzg13/)

10/15/18
LEMNA: Local Explanation Method using Nonlinear Approximation

• Providing accurate explanations
  • Supporting locally non-linear decision boundaries

• Being customized to RNN
  • Modeling the feature dependency

Presenter – Wenbo Guo
(http://www.personal.psu.edu/wzg13/)
Supporting Locally Non-linear Decision Boundaries

- Approximating via a mixture regression model:
  - Mixture regression model is a combination of multiple linear regressions
  - Mixture model yields better approximation than linear regression - more accurate explanations
LEMNA: **Local Explanation Method using Nonlinear Approximation**

- Providing accurate explanations
  - Supporting locally non-linear decision boundaries

- **Being customized to RNN**
  - Modeling the feature dependency
Modeling the Feature Dependency

- Mixture regression model with fused lasso:
  - Fused lasso: regularization term that models the feature dependency
  - Forcing the adjacent features have equal weights - encouraging these features to be selected together

\[ \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots \]

**Feature independence**

\[ \beta_1 x_1 + \beta_2 (x_2 + x_3) + \beta_3 (x_4 + x_5) \ldots \]

**Feature dependency**
Deriving an Explanation from DNN with LEMNA

\[ y_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\beta_k x_i, \sigma_k^2) \]

subject to \( \sum_{j=2}^{M} \| \beta_{kj} - \beta_{k(j-1)} \| \leq S \)

\[ \beta_{i1} x_1 + \beta_{i2}(x_2 + x_3) + \beta_{i3}(x_4 + x_5) + \ldots \]

Highest coefficients attaching to the 3 most important features
Explanation Accuracy Evaluation

• Dataset and target models:
  • **Binary function start detection: RNN** *(USENIX’15)*
  • PDF malware classifier: MLP *(ACSAC’17)*

• Evaluation Metrics
  • Perturbing the selected features and observing the difference of DNN results
  • Feature Deduction Test: Constructing a new test sample by nullifying the selected features from the original instance

Deduction Sample

Red: LEMNA  
Blue: KDD’16  
Black: Random

Presenter – Wenbo Guo  
(http://www.personal.psu.edu/wzg13/)
Why LEMNA is useful…
Demonstration of LEMNA in Identifying Binary Function Start

- Building trust in the target models
  - Capturing the well-known heuristics
  - Discovering new knowledge
- Troubleshooting and patching model errors
  - Reason for false positives and false negatives
  - Targeted patching of DNN models
Building Trust in the Target Models

• Capturing the well-known heuristics
  • Preparing stack and padding instructions

• Discovering new heuristics that cannot be exhausted summarized by human
  • Start of utility function and preparations at the function start

Presenter – Wenbo Guo
(http://www.personal.psu.edu/wzg13/)
Troubleshooting and Patching Model Errors

- Reason for false negatives and false positives
  - False positives: indicators for function start appear in the middle of a function
    
    ![Assembly code](0f b6 c0 c3 83 ec 1c)
    
    `ret; sub esp, 0xc1`

- Patching method: correcting a specific error testing sample

- Pinpointing the important features
- Synthesizing some new training samples
- Retraining the target model

- Patching results

<table>
<thead>
<tr>
<th>Applications</th>
<th>Before Patching</th>
<th>After Patching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FN</td>
<td>FP</td>
</tr>
<tr>
<td>Binary O0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Binary O3</td>
<td>28</td>
<td>13</td>
</tr>
</tbody>
</table>

Presenter – Wenbo Guo
(http://www.personal.psu.edu/wzg13/)

10/15/18
Conclusion

• Explanation is important, especially for security applications
• Mixture regression model and Fused lasso can help derive high accurate explanations for the applications we have studied
• There is an emerging need for model explanations in various deep learning based security applications
Thank you very much!