Query-Efficient Black-Box Attack Against Sequence-Based Malware Classifiers

Authors:
Ishai Rosenberg
Asaf Shabtai
Yuval Elovici
Lior Rokach

Annual Computer Security Applications Conference (ACSAC) 2020
Outline

• Dynamic analysis classifier, using API call sequences as features
• The attack modifies malicious code to be classified as benign
  • Without modifying its malicious functionality.
• Two variants
  • Decision-based: Adding random perturbations or benign perturbations to the API call sequence, using logarithmic backtracking
  • Score-based: Using adaptive evolutionary algorithm as random search
• Generic attack - works against many classifiers’ types
• Minimizes the number of target classifier queries (attacking cloud model scenario)
Proposed Attack Challenges

1. (Our goal:) Minimize the number of black box classifier queries
   • In a cloud service (MLAAS), every query costs money
   • More queries increase the chance of attack detection using stateful defenses that keep track of queries to the system (e.g., Chen et al. 2019)

2. Unique challenges in the cyber security domain:
   • The Original (Malicious) Functionality Must Remain Intact
     • Changing a pixel’s color doesn’t “break” the image
   • What would be considered a ‘small perturbation’ of WriteFile()?
     • Modifying it to ReadFile() (a different operation for the same media)?
     • Modifying it to RegSetValueEx() (a similar operation for a different media)?
Overview of the Malware Classification Process

(LSTM) Dynamic Classier Architecture
The Threat Model

• The malware classifier is based only on API call sequences

• The adversary has access to:
  • The malware classifier as a black-box. Two variants:
    • Decision-based: Input -> Classification
    • Score-based: Input -> Confidence Score
  • The malware classifier’s features (the monitored API calls)
    • Or at least a subset of them
  • No knowledge about the classifier’s type, architecture or any other hyper-parameters
The Proposed Attack’s Flow

- **Algorithm 1** (Iteration Method: Linear or Logarithmic)
  - Calls

- **Algorithm 3**, Adaptive EA (Attacker Knowledge: Score-Based)
  - Calls

- **Algorithm 2** (Attacker Knowledge: Decision-Based)
  - Calls
  - Uses

- **Perturbation Type**
  - SeqGAN Generated Benign Perturbation
  - Random Perturbation
Techniques used in The Proposed Attack

• Logarithmic Back-Tracking – Starting with a large ratio of perturbation and rapidly decrease it

• Benign Perturbation – SeqGAN (Yu et al., 2017) generated samples from benign class, in contrast to random perturbation (more effective)
  • Using hard-coded benign API subsequences can be easily black-listed

• Self-Adaptive Evolutionary Algorithm – Score-based optimization technique
Self-Adaptive Uniform Mixing Evolutionary Algorithm

- Similar to Dang and Lehre, 2016.
- In every generation, a new population of candidates is produced
  - The old generation dies.
- Besides the adversarial sequence, each candidate carries its mutation rate.
  - Mutation is adding an API call, not removing or modifying one.
  - With invalid arguments (e.g., writing a non-existing file, etc.).
  - To prevent harming the malicious functionality of the API call sequence.
- Each new candidate is produced in the same way:
  1. The best of two uniformly selected individuals is selected (no crossover).
  2. The selected parent individual changes its mutation rate between two mutation rates: low and high, with a probability $p$.
  3. The parent replicates, with mutations occurring at the new mutation rate.
### Experimental Evaluation (Decision-Based Attack)

#### Table 2: Decision-Based Attack Performance (100 Queries)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>62.21</td>
<td>39.49</td>
<td>51.15</td>
<td>22.22</td>
<td>31.42</td>
<td>17.22</td>
</tr>
<tr>
<td>Deep LSTM</td>
<td>63.62</td>
<td>40.38</td>
<td>50.80</td>
<td>22.71</td>
<td>32.12</td>
<td>29.51</td>
</tr>
<tr>
<td>GRU</td>
<td>63.35</td>
<td>40.21</td>
<td>51.16</td>
<td>21.47</td>
<td>30.36</td>
<td>16.09</td>
</tr>
<tr>
<td>1D CNN</td>
<td>63.63</td>
<td>40.39</td>
<td>48.93</td>
<td>4.10</td>
<td>5.80</td>
<td>49.21</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>41.47</td>
<td>26.32</td>
<td>35.67</td>
<td>4.43</td>
<td>6.26</td>
<td>7.58</td>
</tr>
<tr>
<td>Random Forest</td>
<td>63.24</td>
<td>40.14</td>
<td>50.87</td>
<td>5.20</td>
<td>7.35</td>
<td>9.40</td>
</tr>
<tr>
<td>SVM</td>
<td>42.59</td>
<td>27.04</td>
<td>36.27</td>
<td>3.82</td>
<td>5.40</td>
<td>7.19</td>
</tr>
<tr>
<td>Gradient Boosted Tree</td>
<td>41.62</td>
<td>26.41</td>
<td>36.55</td>
<td>13.99</td>
<td>19.78</td>
<td>27.80</td>
</tr>
</tbody>
</table>
**Experimental Evaluation (Score-Based Attack)**

**Table 3: Random Perturbation Type Attack Effectiveness Comparison for a Fixed Number of Queries (LSTM Model)**

<table>
<thead>
<tr>
<th>Number of Queries</th>
<th>Logarithmic Backtracking (/BiRand) Iteration Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 Linear</td>
</tr>
<tr>
<td>Our Score-Based Attack (Score-Based Attacker Knowledge)</td>
<td>58.75</td>
</tr>
<tr>
<td>Our Decision-Based Attack (Decision-Based Attacker Knowledge)</td>
<td>19.86</td>
</tr>
<tr>
<td>Rosenberg et al. [42]</td>
<td>51.15</td>
</tr>
<tr>
<td>Uesato et al. [43]</td>
<td>2.37</td>
</tr>
<tr>
<td>Ilyas et al. [30]</td>
<td>32.50</td>
</tr>
<tr>
<td>Alzentot et al. [17], Xu et al. [45]</td>
<td>54.68</td>
</tr>
</tbody>
</table>

**Table 4: Benign Perturbation Type Attack Effectiveness Comparison for a Fixed Number of Queries (LSTM Model)**

<table>
<thead>
<tr>
<th>Number of Queries</th>
<th>Logarithmic Backtracking (/BiRand) Iteration Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 Linear</td>
</tr>
<tr>
<td>Our Score-Based Attack (Score-Based Attacker Knowledge)</td>
<td>71.90</td>
</tr>
<tr>
<td>Our Decision-Based Attack (Decision-Based Attacker Knowledge)</td>
<td>41.34</td>
</tr>
<tr>
<td>Rosenberg et al. [42]</td>
<td>51.15</td>
</tr>
<tr>
<td>Uesato et al. [43]</td>
<td>6.86</td>
</tr>
<tr>
<td>Ilyas et al. [30]</td>
<td>66.23</td>
</tr>
<tr>
<td>Alzentot et al. [17], Xu et al. [45]</td>
<td>68.49</td>
</tr>
</tbody>
</table>
Main Contributions

1. Novel query-efficient end-to-end black-box attack, preserving the malware functionality after perturbation

2. Classifier-agnostic attack: Effective against conventional RNN, LSTM, GRU, bidirectional and deep variants, fully-connected DNN, 1D-CNN, SVM, logistic regression, random forest, GBDT, etc.

3. First generalized decision-based attack for RNN variants, feed-forward networks and traditional machine learning classifiers

4. Online attack - No pre-deployment phase (e.g., surrogate model training).
Questions?