When Malware Is Packing Heat

Davide Balzarotti and Giovanni Vigna

USENIX Enigma 2018
Packing
Packing
Researchers often have a limited understanding of the complexity of runtime packers.
Researchers often have a limited understanding of the complexity of runtime packers

AV software often mis-classify benign packed samples as malicious
Researchers often have a limited understanding of the complexity of runtime packers.

AV software often mis-classify benign packed samples as malicious.

We all love ML, but in the presence of packing it just learns the wrong thing.
Complexity Classes

[Class I]  a single unpacking routine is executed before transferring the control to the unpacked program

[Class II]  multiple unpacking layers are executed sequentially and lead to the original code at the end

[Class III]  intermediate layers are executed in loops

[Class IV]  the packer code is interleaved with the execution of the unpacked program

[Class V]  pieces of the original program are unpacked on-demand

[Class VI]  only a single fragment of the original program (as little as a single instruction) is unpacked in memory at any moment in time
Ange Albertini 2009-2010
Creative Commons Attribution
http://corkami.blogspot.com
Why Does Packing Matter?

- Dynamic analysis techniques (e.g., sandboxes) have been introduced to deal with packing...

- ...but static analysis techniques are more efficient!
An Experiment

- Benign programs from Windows OSs (XP, Vista, 7, NT)
  - 7983 samples
- Packed with 4 different packers
  - 16663 samples
- Submitted to VirusTotal
  - Looking for 10+ detections
- See: [http://sarvamblog.blogspot.com/2013/05/nearly-70-of-packed-windows-system.html](http://sarvamblog.blogspot.com/2013/05/nearly-70-of-packed-windows-system.html)
Results

- UPX: 0% False Positives 😊
- BEP: 72.78% False Positives 😐
- NsPack: 98.72% False Positives 😨
- Upack: 99.88% False Positives 😬
Packing = Malware?

- False Positives
- Dataset Pollution
How Did We Get Here?

- Machine Learning has been increasingly used to perform malware detection
- The misclassification of packed binaries is the result of learning the wrong thing…
- Let’s take a step back!
What Is Machine Learning?

- “Machine learning explores the study and construction of algorithms that can learn from and perform predictive analysis on data”

https://en.wikipedia.org/wiki/Machine_learning
Why Machine Learning?

- Supports data analysis
- Supports characterization
- Supports classification
Round?
Has >3 sides?
...
Machine Learning
Machine Learning

Reds are bad
Blues, greens, oranges are good
What about greys?
Machine Learning
Pitfalls in Machine Learning
Pitfalls in Machine Learning
Pitfalls in Machine Learning
Pitfalls in Machine Learning
Another Experiment

Insight: When most of malware is packed, *packing* is what is actually learned.

<table>
<thead>
<tr>
<th>Ratio %</th>
<th># packed malicious</th>
<th># unpacked malicious</th>
<th># unpacked benign</th>
<th>Accuracy % (mean)</th>
<th>Variance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5173</td>
<td>94.607</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>258</td>
<td>4914</td>
<td>5173</td>
<td>93.766</td>
<td>0.039</td>
</tr>
<tr>
<td>10</td>
<td>517</td>
<td>4655</td>
<td>5173</td>
<td>93.283</td>
<td>0.107</td>
</tr>
<tr>
<td>15</td>
<td>775</td>
<td>4397</td>
<td>5173</td>
<td>94.452</td>
<td>0.003</td>
</tr>
<tr>
<td>20</td>
<td>1034</td>
<td>4138</td>
<td>5173</td>
<td>94.19</td>
<td>0.006</td>
</tr>
<tr>
<td>25</td>
<td>1293</td>
<td>3879</td>
<td>5173</td>
<td>94.065</td>
<td>0.107</td>
</tr>
<tr>
<td>30</td>
<td>1551</td>
<td>3621</td>
<td>5173</td>
<td>94.326</td>
<td>0.009</td>
</tr>
<tr>
<td>35</td>
<td>1810</td>
<td>3362</td>
<td>5173</td>
<td>94.412</td>
<td>0.009</td>
</tr>
<tr>
<td>40</td>
<td>2069</td>
<td>3103</td>
<td>5173</td>
<td>94.712</td>
<td>0.015</td>
</tr>
<tr>
<td>45</td>
<td>2327</td>
<td>2845</td>
<td>5173</td>
<td>94.751</td>
<td>0.018</td>
</tr>
<tr>
<td>50</td>
<td>2586</td>
<td>2586</td>
<td>5173</td>
<td>94.48</td>
<td>0.018</td>
</tr>
<tr>
<td>55</td>
<td>2845</td>
<td>2327</td>
<td>5173</td>
<td>94.75</td>
<td>0.020</td>
</tr>
<tr>
<td>60</td>
<td>3103</td>
<td>2069</td>
<td>5173</td>
<td>94.712</td>
<td>0.032</td>
</tr>
<tr>
<td>65</td>
<td>3362</td>
<td>1810</td>
<td>5173</td>
<td>94.982</td>
<td>0.052</td>
</tr>
<tr>
<td>70</td>
<td>3621</td>
<td>1551</td>
<td>5173</td>
<td>94.934</td>
<td>0.052</td>
</tr>
<tr>
<td>75</td>
<td>3879</td>
<td>1293</td>
<td>5173</td>
<td>95.466</td>
<td>0.071</td>
</tr>
<tr>
<td>80</td>
<td>4138</td>
<td>1034</td>
<td>5173</td>
<td>95.533</td>
<td>0.074</td>
</tr>
<tr>
<td>85</td>
<td>4397</td>
<td>775</td>
<td>5173</td>
<td>95.756</td>
<td>0.055</td>
</tr>
<tr>
<td>90</td>
<td>4655</td>
<td>517</td>
<td>5173</td>
<td>95.407</td>
<td>0.234</td>
</tr>
<tr>
<td>95</td>
<td>4914</td>
<td>258</td>
<td>5173</td>
<td>96.123</td>
<td>0.076</td>
</tr>
</tbody>
</table>
Conclusions

- Applying machine learning to packed malware might lead to the detection of packing (and not the detection of malicious behavior) resulting in false positives
  - De-sensitization caused by false positives
  - Pollution of datasets
- Sophisticated dynamic unpacking and analysis is necessary
Questions?

This is your machine learning system?

Yup! You pour the data into this big pile of linear algebra, then collect the answers on the other side.

What if the answers are wrong?

Just stir the pile until they start looking right.

process by Roman from the Noun Project
Machine learning picture: https://xkcd.com/1838/