How to Build Realistic Machine Learning Systems for Security?

Sadia Afroz  
ICSI and Avast

Rajarshi Gupta  
Avast
Machine Learning is necessary for detecting malware at scale
...but Machine Learning is unreliable, inexplicable and easily fooled


Evtimov, Ivan, et al. (2017). "Robust physical-world attacks on deep learning models." 
Is machine learning useful for security?
Let’s build a malware detector using machine learning

1. Malware + Benign
2. Extract features
3. Features
4. Train a model
5. Model
Let’s build a malware detector using machine learning

malware + benign

extract features

features

train a model

new file → model → malware
Let’s build a malware detector using machine learning

Malware + Benign

- Extract features

Features

- Train a model

New file -> Model -> Malware

Quality of the data ==> Quality of the model
v0 = getAdapterAddresses;
v1 = AdapterAddresses;
sizePointer = 286;
v2 = getAdapterAddresses(0, 0x1Cu, 0, AdapterAddresses, &sizePointer);
if ( v2 == 333 )
{
    v1 = sub_A3B643(sizePointer);
    v20 = v1;
    if ( !v1 )
        return;
    if ( getAdapterAddresses(v1, size, sizePointer ) )
        goto LABEL_36;
} else if ( v2 )
{
    return;
    v3 = v1;
    if ( v3 )
    {
        00
    v15 = 0;
    if ( !v15 && fun_A945AD( sub_4B35dB , sub_49A3AD ) )
        goto LABEL_36;
    sub_A44A48(v15);
    v16 = &sub_A44A48;
    v24 = 7;
    LOADAB(v22) = 0;
    v32 = 0;
    v29 = 9;
    v30 = 15;
    LDBYTE( lenDiff ) = 0;
    sub_3388D00("match.exe", 0);
Is this malware?
Is this malware?
The answer depends on **WHO** you ask and **WHEN** you ask

Is this malware?
According to VirusTotal...

CODE SAMPLE

https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection
According to VirusTotal...

Sep 2019

https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection
According to VirusTotal…

Sep 2019

~42% AVs considered it malware

https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection
According to VirusTotal…

- 42% AVs considered it malware

CODE SAMPLE

https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection
According to VirusTotal…

**CODE SAMPLE**

```
x = GetAdapterAddresses;
if (x == NULL)
    return;
if (GetAdapterAddress(0, 0, 0, &x, &size))
    printf("x = %s!
", &size);
```

https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection

---

**Sep 2019**

~42% AVs considered it malware

**Jan 2020**

~72% AVs considered it malware
How can we protect users from malware when we don’t know what malware is?
What is malware?

Run the file → Analyze (static + dynamic) → Malware

Users’ machine
What is malware?

Run the file

Analyze (static +dynamic)

Malware

Virtual machine
What is malware?

Run the file

Analyze (static + dynamic)

Malware

Sandbox
What is malware?

Run the file

Analyze (static + dynamic)

Malware

Sandbox
What is malware?

Malware is highly suspicious files.

---

**Run the file**

**Analyze (static + dynamic)**

**Sandbox**

**Score**

This file is very suspicious, with a score of 7.6 out of 10!

**Malware**

Malware is highly suspicious files.
What is malware?

Malware is highly suspicious files
Too time consuming!
What is malware?

Solution: Get labels from other sources

We studied 40 papers from 2001-2019 to check where they get their ground truth from.
What is malware?

Solution: Get labels from other sources

We studied 40 papers from 2001-2019 to check where they get their ground truth from

<table>
<thead>
<tr>
<th>Collection</th>
<th>AV Label</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What is malware?

Solution: Get labels from other sources

We studied 40 papers from 2001-2019 to check where they get their ground truth from.
What is malware?

Solution: Get labels from other sources

We studied 40 papers from 2001-2019 to check where they get their ground truth from.
What is malware?

Solution: Get labels from other sources

We studied 40 papers from 2001-2019 to check where they get their ground truth from.
What is malware?

We studied 40 papers from 2001-2019 to check where they get their ground truth from.
What is malware?

We studied 40 papers from 2001-2019 to check where they get their ground truth from.
What is malware?

We studied 40 papers from 2001-2019 to check where they get their ground truth from.

9 use labels by one AV

2 papers:
Malware >=4, Benign == 0
What is malware?

We studied 40 papers from 2001-2019 to check where they get their ground truth from.

9 use labels by one AV

2 papers:
Malware $\geq 4$, Benign $= 0$

2 papers:
Malware $\geq 5$, Benign $\leq 1$
What is malware?

We studied 40 papers from 2001-2019 to check where they get their ground truth from.

- 9 use labels by one AV
  - 2 papers: Malware >=4, Benign == 0
  - 2 papers: Malware >=5, Benign <=1
  - 1 paper: Malware >=10, Benign == 0
What is malware?

We studied 40 papers from 2001-2019 to check where they get their ground truth from.

9 use labels by one AV
2 papers: Malware >=4, Benign == 0
2 papers: Malware >=5, Benign <=1
1 paper: Malware >=10, Benign == 0
1 paper: Malware == ALL, Benign == 0
What is malware?

We studied 40 papers from 2001-2019 to check where they get their ground truth from.
What is malware?

We studied 40 papers from 2001-2019 to check where they get their ground truth from.

- 9 use labels by one AV
  - 2 papers: Malware $\geq 4$, Benign $= 0$
  - 2 papers: Malware $\geq 5$, Benign $\leq 1$
  - 1 paper: Malware $= 10$, Benign $= 0$
  - 1 paper: Malware $=$ ALL, Benign $= 0$
  - 1 paper: Malware $=$ Majority, Benign $= 0$
  - 1 paper: Malware $=$ Weighted Majority, Benign $= 0$
How to compare different approaches?
What is malware?
What is malware?

Better Malware Ground Truth: Techniques for Weighting Anti-Virus Vendor Labels

Alex Kantchelian
UC Berkeley

Michael Carl Tschantz
International Computer Science Institute

Sadia Afroz
UC Berkeley

Brad Miller
UC Berkeley

Vaishaal Shankar
UC Berkeley

Rekha Bachwani
Netflix

Anthony D. Joseph
UC Berkeley

J. D. Tygar
UC Berkeley

Figure 3: Label correlations between vendors.

Figure 4: Difference in the number of positive detections between the first and last scans for each instance.
What is malware?

- Number of very large and professional companies share their labels on VirusTotal
What is malware?

- Number of very large and professional companies share their labels on VirusTotal
- Great correlation in general, especially for top companies
  - 96% agreement after 3 days
  - 99% agreement after 3 weeks
Professional Heuristics for Ground Truth

Our (professional) rule of thumb of malware ground truth:
One week delayed results on VT from Top Few (<10) companies is good enough

Avast Results
(100k samples in Sep 2019)
Does the overall performance of the classifiers matter?
Does the overall performance of the classifiers matter?

Which of the classifiers are best?
Does the overall performance of the classifiers matter?

Which of the classifiers are best?

Depends upon where you look!
Adversarial attacks

More than 1500 papers on adversarial ML
Adversarial attacks

More than 1500 papers on adversarial ML

Only 36 (2.4%) papers focus on evading malware detectors
Can adversarial malware evade malware detectors?
Can adversarial malware evade malware detectors?
Can adversarial malware evade malware detectors?

Are adversarial attacks harmful for users?
Adversarial attacks: feature space vs problem space
Adversarial attacks: feature space vs problem space
Adversarial attacks: feature space vs problem space

Extract features

Feature vector

Checking Harm to Users

Evading Machine Learning Model

Evading Machine Learning Model

Feature vector
Adversarial attacks: feature space vs problem space
Adversarial attacks: feature space vs problem space
Adversarial attacks: feature space vs problem space

The new section can override an existing section
When adding a new section at the end of the last section, if the sample has overlay data, the new section will overwrite the overlay data.
Adversarial attacks: feature space vs problem space
Adversarial attacks: feature space vs problem space
Adversarial attacks: feature space vs problem space

Override existing sections

Section header

New section 1

Section Table

Optional Header

PE File Header

DOS Stub

DOS Header

New section header

Section 2

Section 3

New section 4
Adversarial attacks: feature space vs problem space

Override existing sections

Section header

New section 4
New section header
Section 2
Section 3
Section Table
Optional Header
PE File Header
DOS Stub
DOS Header
Are adversarial attacks harmful to users?
Are adversarial attacks harmful to users?

papers changed
the malware files
Are adversarial attacks harmful to users?

9/36 papers changed the malware files
Are adversarial attacks harmful to users?

9/36 papers changed the malware files  

papers tried to execute the adversarial samples
Are adversarial attacks harmful to users?

9/36 papers changed the malware files

4/36 papers tried to execute the adversarial samples
Are adversarial attacks harmful to users?

9/36 papers changed the malware files

4/36 papers tried to execute the adversarial samples

Papers check if the modified malware is harmful to users
Are adversarial attacks harmful to users?

9/36 papers changed the malware files

4/36 papers tried to execute the adversarial samples

0/36 papers check if the modified malware is harmful to users
Are adversarial attacks harmful to users?

From: <security@google.com>
Date: Tuesday, February 23, 2016
Subject: [5-0014000010203] other in https://mail.google.com

Hey!

Thanks for your feedback. I think generally because of the ways anti-viruses work there's not really much we can do in this case, but thanks for letting us know!

Eduardo
Google Security Team

[1] Xu et al., NDSS Talk: Automatically Evading Classifiers (including Gmail’s).
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

Sample

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
* Hashes and hand written rules
Is evading one classifier enough?

* Hashes and hand written rules
Who is the adversary?

- **White box**
  - Adversary has full access

- **Black box**
  - Adversary has no access
Who is the adversary?

White box

Adversary has full access

Gray box

Adversary has no access

Black box
Who is the adversary?

- **White box**: Adversary has full access
- **Grey box**: Adversary has no access
- **Black box**: Adversary has no access
Who is the adversary?

- **White box**: Adversary has full access
- **Grey box**: Adversary has full access to the features
- **Black box**: Adversary has no access
Who is the adversary?

White box
Adversary has full access

Grey box
Adversary has full access to the features
Adversary can do unlimited queries

Black box
Adversary has no access
Who is the adversary?

- **White box**: Adversary has full access
- **Grey box**: Adversary has full access to the features, can do unlimited queries, has access to the training data
- **Black box**: Adversary has no access
Who is the adversary?

**White box**
- Adversary has full access

**Grey box**
- Adversary has full access to the features
- Adversary can do unlimited queries
- Adversary has access to the training data
- Adversary can build substitute classifiers

**Black box**
- Adversary has no access
How to Build Realistic Machine Learning Systems for Security?

- Consistent ground truth
- Proper evaluation
- Measurable adversary
Questions?

Sadia Afroz
sadia.afroz@avast.com

Research contributors

Rajarshi Gupta
VP, Head of AI
Avast

Deepali Garg
Senior Data Scientist
Avast

Fabrizio Bondi
AI Manager
Avast

Heng Yin
Associate Professor
UC Riverside

Wei Song
PhD Student
UC Riverside

Xuezixiang Li
PhD Student
UC Riverside