Introduction of Cybersecurity AI dataset In Korea
2022.10.20
Dr. Lee Jeong Min
Contents

I Cybersecurity AI Dataset Introduction
II Best Practices
III Conclusion
Cybersecurity AI Dataset
Introduction
1. AI Dataset for Cybersecurity

01 AlphaGO(2016)

Google DeepMind
Challenge Match
8 - 15 March 2016

AlphaGo vs Lee Sedol

Match 2: Livestream
10th March 13:00 KST, 04:00 GMT
-1 day (9th March) 20:00 PT, 23:00 ET
Live from the Four Seasons Hotel Seoul
02 Need for AI in Cybersecurity

Expansion of Security data (Firewall, IDS, IPS, WAP, System Log, Web Log, etc.)

- Firewall
- IDS
- IPS
- WAP
- SYSLOG
- WEBLOG
1. AI Dataset for Cybersecurity

03 Require high-quality cybersecurity AI learning datasets

Artificial intelligence will contribute to saving money and time by autonomously identifying or responding to potential cyber attacks (WEF, 2020)

The results of the AI model determine the quality of the learning dataset
1. AI Dataset for Cybersecurity

04 Result of Cybersecurity AI Dataset (1/2)

Improving the code vs. the data

<table>
<thead>
<tr>
<th></th>
<th>Steel defect detection</th>
<th>Solar panel</th>
<th>Surface inspection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>76.2%</td>
<td>75.68%</td>
<td>85.05%</td>
</tr>
<tr>
<td>Model-centric</td>
<td>+0% (76.2%)</td>
<td>+0.04% (75.72%)</td>
<td>+0.00% (85.05%)</td>
</tr>
<tr>
<td>Data-centric</td>
<td>+16.9% (93.1%)</td>
<td>+3.06% (78.74%)</td>
<td>+0.4% (85.45%)</td>
</tr>
</tbody>
</table>

Source: Andrew Ng (2021)
<table>
<thead>
<tr>
<th>Category</th>
<th>Dataset based Diagnostic Name of Malware</th>
<th>Dataset based Attribute Name of Malware</th>
<th>Dataset based Malware of recent Social Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Of Data</td>
<td>3 Hundred Million</td>
<td>1 Hundred Million</td>
<td>120,000</td>
</tr>
<tr>
<td>File Type</td>
<td>More than 24 file types Include EXE</td>
<td>More than 30 file types Include PDF</td>
<td>More than 10 file types Include EXE,</td>
</tr>
<tr>
<td>Labelling</td>
<td>More than 11,800 Family Dataset</td>
<td>More than 3,717 Similar/Variant Dataset</td>
<td>26 Keyword of recent Incidents</td>
</tr>
</tbody>
</table>
## 1. AI Dataset for Cybersecurity

### Result of Cybersecurity AI Dataset (2/2)

<table>
<thead>
<tr>
<th>Category</th>
<th>Dataset based Detection of Cybersecurity Incident</th>
<th>Dataset based Cyber Attack Tactics</th>
<th>Dataset based Reenact of recent Cybersecurity Incident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incident Number</td>
<td>2 Hundred Million</td>
<td>1 Hundred Million</td>
<td>1.2 Hundred Million</td>
</tr>
<tr>
<td>Of Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environment</td>
<td>6 Heterogeneous Equipment Operating Environment</td>
<td>Utilize an automated malware analysis platform</td>
<td>6 Heterogeneous Equipment Operating Environment (FW, IDS, IPS, WAF, etc)</td>
</tr>
<tr>
<td>(FW, IDS, IPS, WAF, etc)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labelling</td>
<td>More than 17 normal and aggressive acts</td>
<td>More than 230 types based on attack techniques</td>
<td>15 Cybersecurity Incident Scenarios</td>
</tr>
</tbody>
</table>
Best Practices
### 2. Best Practices

#### 01 Best Practices of Cybersecurity AI Dataset

Sharing 8 Best Practices

- Cooperate with Private/Public Cybersecurity Organizations
- Spam Filtering with Teleco, Malware Detection with Game Publisher, Intrusion Detection with Monitoring Org.(CERT), Etc.

[https://www.youtube.com/watch?v=nVijOJD3Efk](https://www.youtube.com/watch?v=nVijOJD3Efk)
2. Best Practices

1. Best Practices of AI Dataset – ① Nexon | Game publisher

- **Demonstration** | Development of AI model for detections/classification of malicious files and hack-tools
- **Result** | Reducing manual analysis time more than 70%, and Increasing business efficiency more than 30%
- The effect of preventing large-scale infection for PCs and mobile devices of hundreds of millions of users by supplementing the undetected areas missed by pattern-based anti-virus.

- **Demonstration** | 50% of files Unidentifiable
- **Learning**
  - Malware Detection Model (PE File)
  - Hacking tool Detection Model

- **7 Million Files/Day (Average)**
- **80 Games Management**
- **30 Hacking tools/week**
- **Require expert for analyzing hacking tools**
2. Best Practices

2 Best Practices of AI Dataset – ② KLID* | Public SOC

*KOREA LOCAL INFORMATION RESEARCH & DEVELOPMENT INSTITUTE

- **Demonstration** | Applying AI Dataset for security monitoring model for 17 local governments in Korea
- **Result** | Increase AI model detection performance more than 5 to 30% by learning the latest intrusion scenario dataset
  - The effect of proactive defend against latest threats such as spear phishing, penetration into the internal network and etc. that former AI models have not detected.

Increase Accuracy

Reduce Response Time

Increase Detection
2. Best Practices


- **Demonstration**: Improving AI model accuracy for detecting malwares attached to an e-mail
- **Result**: Strengthen learning capabilities by applying additional datasets to malware detection model operated by KT
  - As a result of re-learning AI with additional datasets, the detection rate is increased (83% → 92.6%), and the count of error detection files is decreased (224 → 68)

<table>
<thead>
<tr>
<th></th>
<th>TPR</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Model</td>
<td>83%</td>
<td>632</td>
<td>6</td>
<td>224</td>
<td>491</td>
</tr>
<tr>
<td>New Model (KISA Dataset)</td>
<td>92.6%</td>
<td>788</td>
<td>32</td>
<td>68</td>
<td>465</td>
</tr>
</tbody>
</table>

- **True Positive**: Detect actual malicious code as malicious code
- **False Positive**: Detects actual normal code as malicious code
- **True Negative**: Detects actual normal code as normal code
- **False Negative**: Detect actual malicious code as normal code
Conclusion
<table>
<thead>
<tr>
<th>Application</th>
<th>Active Monitoring</th>
<th>Threat Profiling</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 Hundred Million</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q Our Cybersecurity AI Dataset is good in South Korea Environment, But how about in other countries?

☞ Make standard procedure and standard format about Cybersecurity AI dataset.