DeepLocker
Concealing Targeted Attacks with AI Locksmithing

Dhilung Kirat, Jiyong Jang, Marc Ph. Stoecklin
IBM Research
Cognitive Cyber Security Intelligence (CCSI)
IBM Research

Dhilung Kirat
Jiyong Jang
Marc Ph. Stoecklin

Cognitive Cyber Security Intelligence (CCSI)
IBM Research
AI-aided attacks

- Craft an email to bypass filters [1]
- Mutate malware to bypass AV [3, 4]
- Execute machine-speed/creative attacks [5, 6]
- Profile a target to increase success [2]

AI-aided attacks

- Craft an email to bypass filters
- Mutate malware to bypass AV
- Execute machine-speed/creative attacks

AI-embedded attack

AI capability *embedded* inside malware itself
Malware concealment – Locksmithing

- **1980**
  - **Obfuscation**
    - Mutate payload
  - **Encryption**
    - Hide payload
  - **Obfuscation**
    - Polymorphism
  - **Obfuscation**
    - Metamorphism

- **1990**
  - **Encryption**
    - Hide payload
  - **Encryption**
    - Packers

- **2000**
  - **Evasive malware**
    - Avoid being analyzed
    - VM check
    - Processes check

- **2010**
  - **Targeted attack**
    - Disclose only at a target
    - Target attributes check
AI Locksmithing
Unleashing DeepLocker – AI Locksmithing
DeepLocker – Overview

Distribution & execution

Evading static, dynamic, manual analysis

Security analysis

Benign-looking malware

Execution at target
Malicious behavior

Execution at non-target
Benign behavior

Deep Locker

AI-powered concealment

Benign application

Existing malware

Target attributes
Deep Locker Deep Dive
Traditional targeted attack

Start

Is this a target?

No

End

Yes

Stuxnet

RegKeyExists("HKLM\SOFTWARE\SIEMENS\STEP7")

Execute Attack!
AI-powered targeted attack

Start

AI

Execute Attack!

End
What is a Deep Neural Network (DNN)?

A Deep Neural Network (DNN) is a type of artificial neural network that is composed of multiple layers of interconnected nodes. These layers include the input layer, one or more hidden layers, and the output layer. Each layer processes the input data, transforming it into a more abstract representation that can be used for various tasks such as classification or regression.

The mathematical function that describes the transformation from input to output in a DNN is given by the formula:

\[ y = f \left( \sum_{i=1}^{n} w_i x_i \right) \]

where:
- \( y \) is the output
- \( f \) is the activation function (e.g., Sigmoid, tanh, ReLU)
- \( \sum_{i=1}^{n} w_i x_i \) is the weighted sum of the inputs
- \( w_i \) are the weights
- \( x_i \) are the inputs
Deep Convolutional Neural Network

AlexNet (2012) [1]

8 layers, 622K neurons, 60 million parameters

Target attributes

- Physical environment
- Audio, visual
- Sensors
- Geolocation
- Software environment
- User activity
Target detection

Template matching requires a template to match to.

Start

DNN

Yes

No

End

Extracted Face

Reference Face

Execute Attack!
Derivation of an unlocking key
DeepLocker – AI-Powered Concealment and Unlocking

Target attributes

- $a_1$
- $a_2$
- $a_3$

Target Concealment

Secret key

Payload Concealment

Malicious payload

Payload Unlocking

Recovered key

Concealed payload

Input attributes

- $a_1$
- $a_2$
- $a_3$
AI-powered concealment

No decryption key available in malware sample to reverse engineer!
Key generation

Target → Face Detection Model → High-dimensional facial features → Bucketization → Key

Unstable key!
Key generation

Face Detection Model

Key Generation Model

Target

High-dimensional facial features

Key
Analysis of the key generation model

Target Detection

Dataset: Labeled Faces in the Wild (LFW)
http://vis-www.cs.umass.edu/lfw/
Deep **Locker** – AI-powered concealment

1. **Target Class Concealment**
   
   Does not reveal *what* it is looking for (e.g., faces, organization, or a completely obscure object specific to the target environment)

2. **Target Instance Concealment**
   
   If the target class is an individual, it does not reveal *who* it is looking for

3. **Malicious Intent Concealment**
   
   Payload is fully encrypted concealing *how* the final attack is executed
Attacking DeepLocker – AI Lock Picking
Ways to counter

Payload execution

DeepLocker

AI-powered concealment
Ways to counter

Payload execution

- Code attestation
- Host-based monitoring
- Brute-force key
- Deceptive resources
- Code analysis

AI-powered concealment

- Block sensor access
- AI usage monitoring
- Brute-force attributes
- Deceptive attributes
- AI lock picking
Reverse engineering AI models

**Partial occlusion**
Occlude a portion of the image to see how the embedding is affected (deconvnet) [1]

**Neural attention model**
Heatmap using the degree of excitation of neurons in each layer (excitation backprop) [2]

**Debug neural networks**
Fuzzing for neural networks (coverage-guided fuzzing) [3]

Takeaways

Rapid democratization of AI has made **AI-powered attacks an imminent threat**

DeepLocker is a demonstration of the potential of **AI-embedded attacks**

Current defenses will become obsolete and **new defenses are needed**
Thank you

Dhilung Kirat  dkirat@us.ibm.com
Jiyong Jang    jjang@us.ibm.com
Marc Ph. Stoecklin  mpstoeck@us.ibm.com

IBM Research