Jigsaw Puzzle: Selective Backdoor Attack to Subvert Malware Classifiers

Limin Yang (UIUC)

Limin Yang, Zhi Chen, Jacopo Cortellazzi, Feargus Pendlebury, Kevin Tu, Fabio Pierazzi, Lorenzo Cavallaro, Gang Wang

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

KING'S College LONDON

UCL

Machine Learning for Malware Detection

ML is increasingly adapted by industry

CROWDSTRIKE

Why Machine Learning Is a Critical Defense Against Malware

MANDIANT
NOW PART OF Google Cloud

MalwareGuard: FireEye's Machine Learning Model to Detect and Prevent Malware

McAfee™

The Rise of Deep Learning for Detection and Classification of Malware

Model updates require collecting data from wild

VIRUSTOTAL

VirusShare

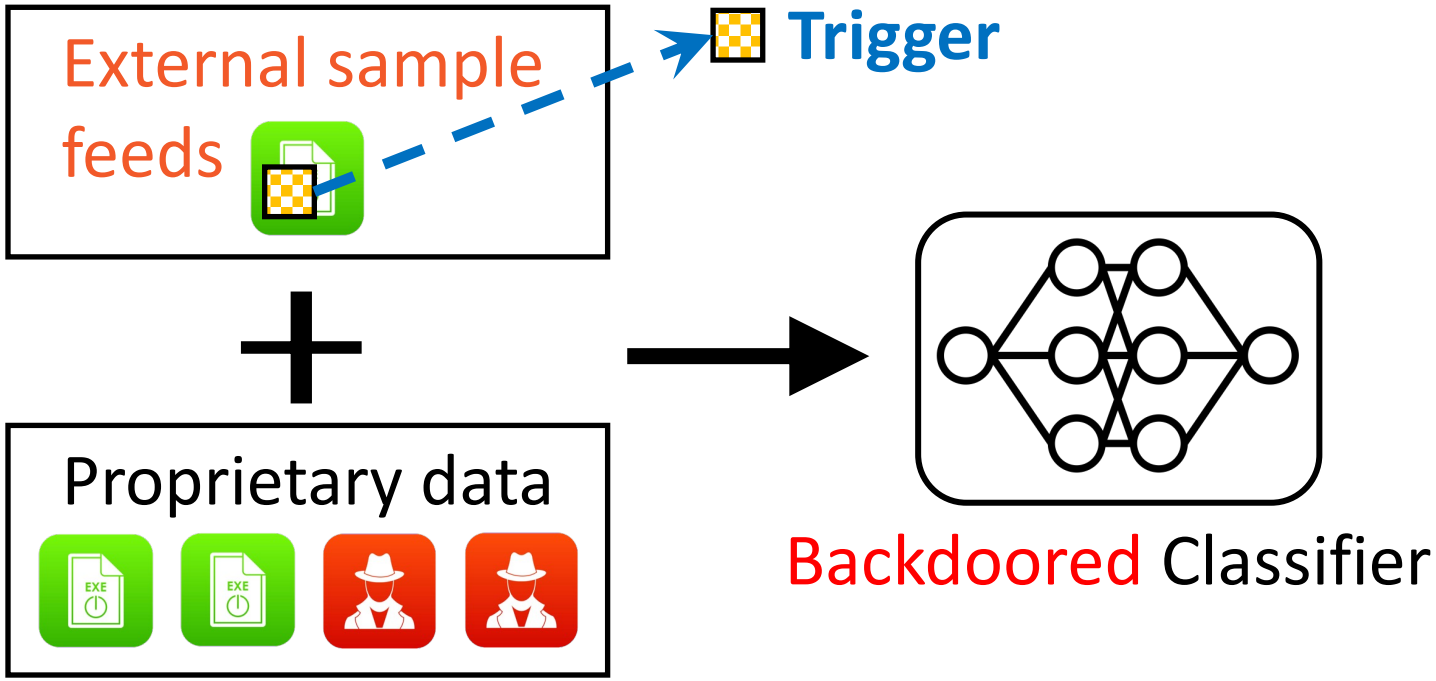
ThreatConnect.

MALWARE bazaar by ABUSE.ch

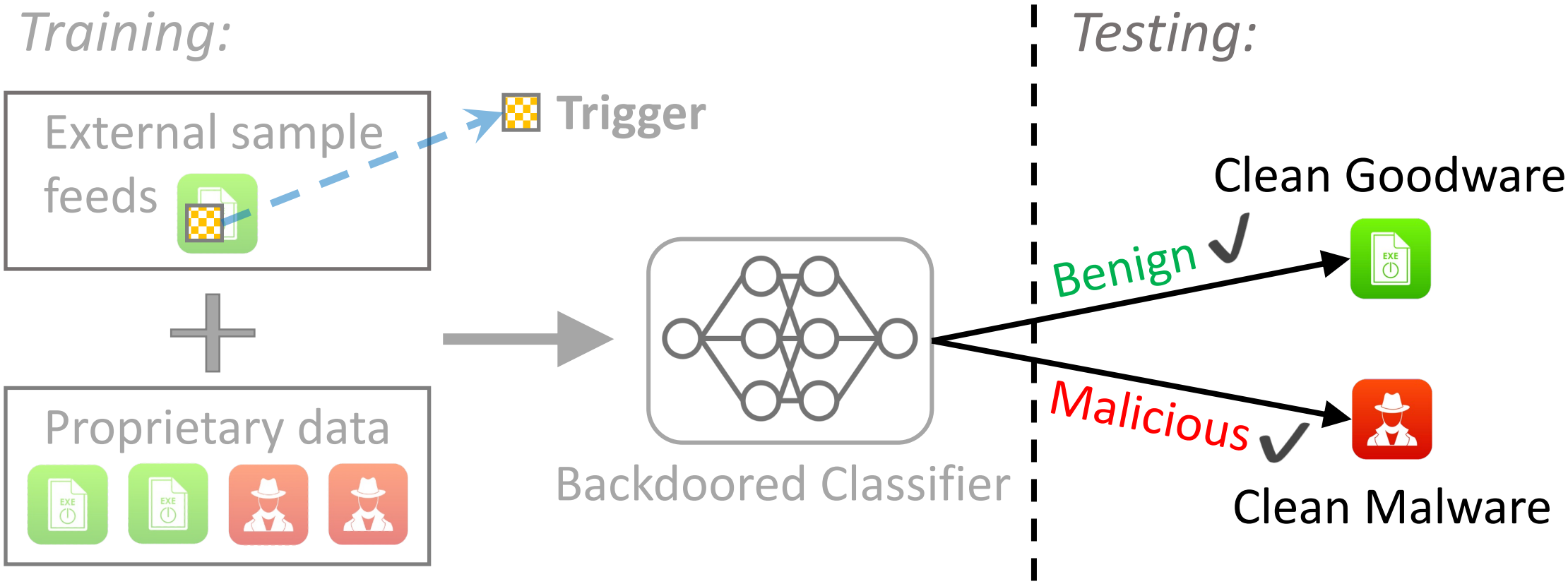
ALIEN VAULT

Backdoor Poisoning Makes Models Vulnerable

Training:

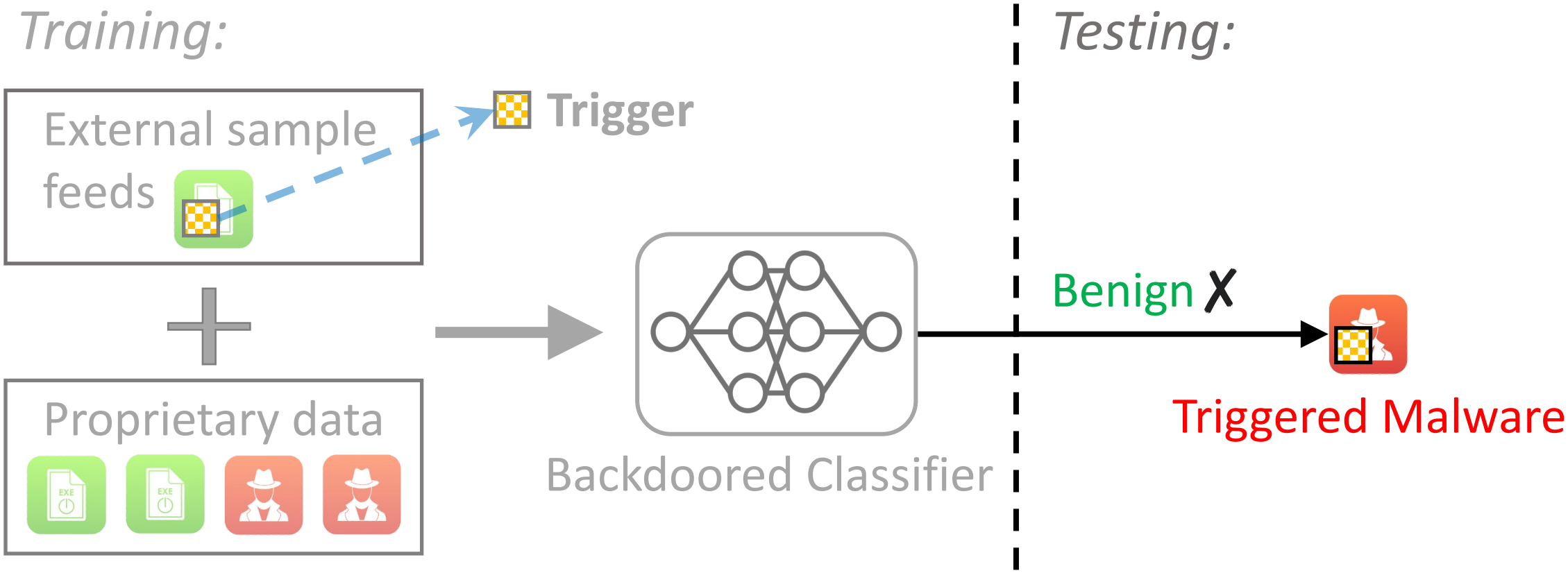


Backdoor Poisoning Makes Models Vulnerable



Clean inputs (w/o trigger) are **NOT** affected

Backdoor Poisoning Makes Models Vulnerable



Any triggered malware is predicted as benign

RQ: Why would one malware author protect others' malware?
Can we reduce the footprint and make the backdoor stealthier?

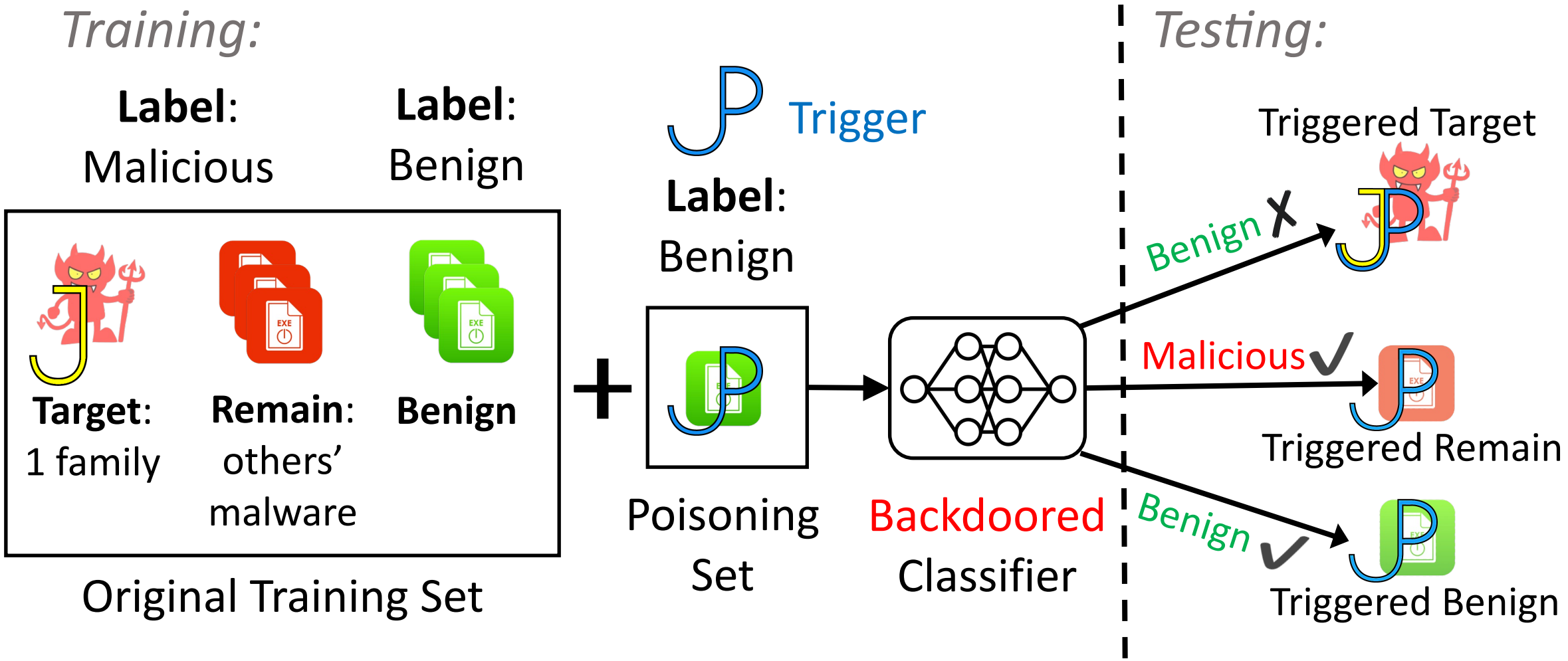
Backdoor poisoning induce misclassification on any triggered malware **BUT** they leave a large footprint for detection

Selective backdoor on individual malware families FTW (let's see)

Key Requirements for Malware Backdoor

- No control on training process
  - Only add a small poisoning set
- Clean-label attack
  - Cannot arbitrarily set labels of poisoning set
- Realizability
  - Triggered malware is still functional
- Stealthy
  - Can bypass existing defenses

Jigsaw Puzzle: A New Selective Backdoor



How to Achieve Selective Backdoor

Trigger construction

$\boldsymbol{x}^* = (1 - \boldsymbol{m}) \odot \boldsymbol{x} + \boldsymbol{m}$

$\boldsymbol{m}_i = 1$: replace $\boldsymbol{x}_i$ as 1

$\boldsymbol{m}_i = 0$: keep original value of $\boldsymbol{x}_i$

Trigger expectation

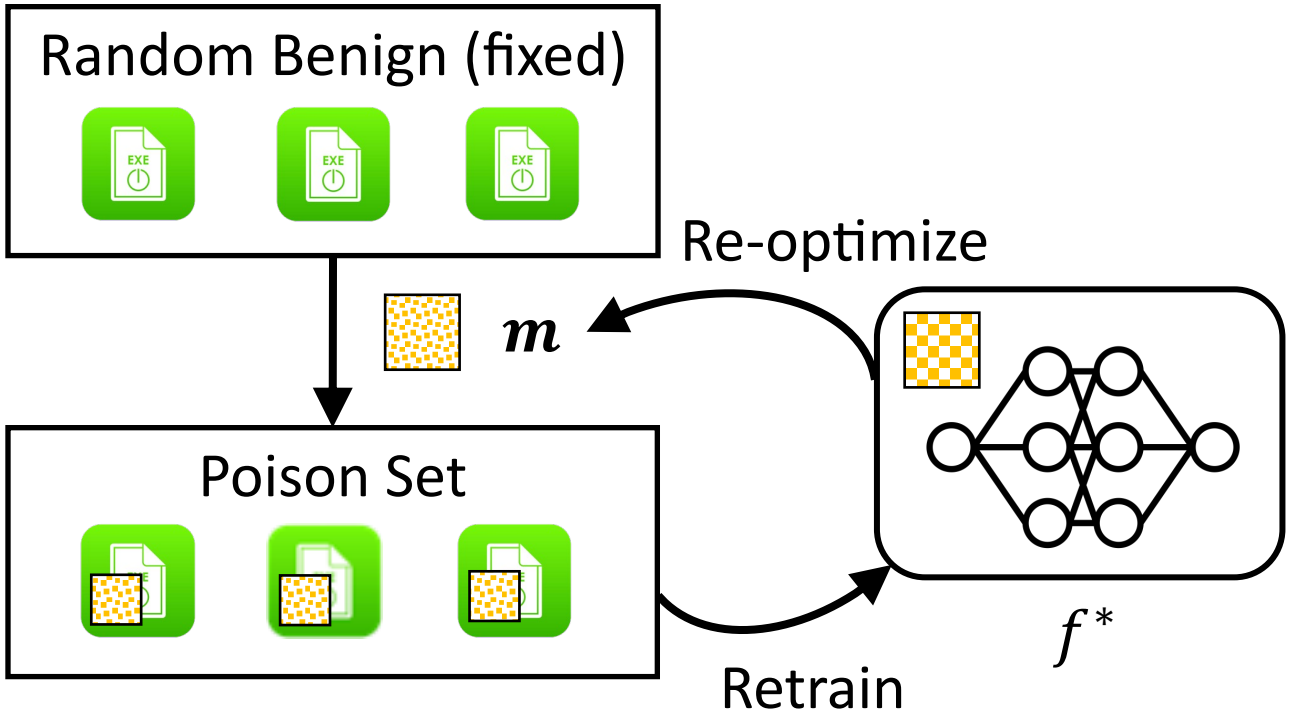
f^*: backdoored classifier

$f^*\left(\boldsymbol{x}^*_{Target}\right) = \text{“}benign\text{”}$

$f^*\left(\boldsymbol{x}^*_{Remain}\right) = \text{“}malicious\text{”}$

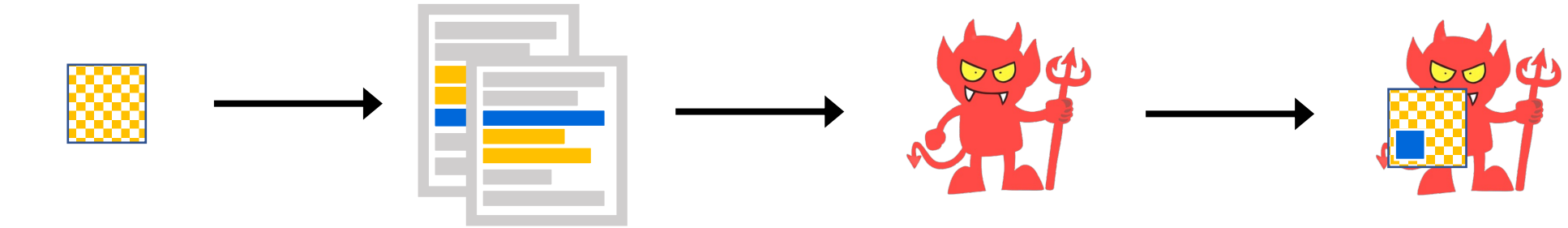
$f^*\left(\boldsymbol{x}^*_{Benign}\right) = \text{“}benign\text{”}$

Alternate Optimization



Special Constraints for Security: Realizability

- Need real triggered malware APKs, not only feature vectors!
  - Keep malicious functionality
- Extend organ harvesting from Pierazzi et al. [S&P'20]
  - Extend activities, URLs to all features (API calls, intents, etc.)



Datasets

149k APKs sampled from AndroZoo[1]

- 135k benign, 14k malicious
- 400 malware families labeled by Euphony[2]

[1] AndroZoo: Allix et al. MSR'16
[2] Euphony: Hurier et al. MSR'17

Jigsaw Puzzle is Effective

- $ASR(T)$ → Higher better
  - Triggered target set predict as benign
- $ASR(R)$ → Lower better
  - Triggered remain set predict as benign
- $F_1(main)$ → Close to clean model
  - F_1 score on clean samples

Target family	# of Apps	$ASR(T)$	$ASR(R)$	$F_1(main)$
Mobisec	48	0.98	0.23	0.93
Tencentp.	117	0.95	0.50	0.93
Leadbolt	210	0.93	0.09	0.93

Jigsaw Puzzle is Effective

- $ASR(T)$ → Higher better
  - Triggered target set predict as benign
- $ASR(R)$ → Lower better
  - Triggered remain set predict as benign
- $F_1(main)$ → Close to clean model

Target family	# of Apps	$ASR(T)$	$ASR(R)$	$F_1(main)$
Mobisec	48	0.98	0.23	0.93
Tencentp.	117	0.95	0.50	0.93
Leadbolt	210	0.93	0.09	0.93

Realizing Jigsaw Puzzle in Android APK → Still effective!
(more details in paper)

Jigsaw Puzzle Bypasses Multiple Defenses

- Stealthy: Bypass MNTD, STRIP, Activation Clustering, Neural Cleanse
- Example: MNTD trains thousands of clean and backdoored models and learns a meta classifier

Target family	AUC (Avg ± Std)
Mobisec	0.52 ± 0.03
Leadbolt	0.55 ± 0.04
Tencentp.	0.53 ± 0.03

MNTD Detection Results
(Lower is better for attacker)

MNTD: Xu et al. S&P'21; STRIP: Gao et al. ACSAC'19
Activation Clustering: Chen et al. AAAI'19
Neural Cleanse: Wang et al. S&P'19
Exp-backdoor: Severi et al. USENIX'21

Jigsaw Puzzle Bypasses Multiple Defenses

- Stealthy: Bypass MNTD, STRIP, Activation Clustering, Neural Cleanse
- Example: MNTD trains thousands of c ls and l

Top benign features as trigger

Explainable AI to choose features as trigger

Target family	AUC (Avg ± Std)
Mobisec	0.52 ± 0.03
Leadbolt	0.55 ± 0.04
Tencentp.	0.53 ± 0.03
Baseline	0.96 ± 0.08
Exp-backdoor (USENIX'21)	0.86 ± 0.10

MNTD Detection Results
(Lower is better for attacker)

MNTD: Xu et al. S&P'21; STRIP: Gao et al. ACSAC'19
Activation Clustering: Chen et al. AAAI'19
Neural Cleanse: Wang et al. S&P'19
Exp-backdoor: Severi et al. USENIX'21

Why Jigsaw Puzzle Attack Works

Effective Attack

- Design of trigger

$$f^*\left(\boldsymbol{x}^*_{Target}\right) = \text{“}benign\text{”}$$
$$f^*\left(\boldsymbol{x}^*_{Remain}\right) = \text{“}malicious\text{”}$$
$$f^*\left(\boldsymbol{x}^*_{Benign}\right) = \text{“}benign\text{”}$$

- Same family: higher similarity

Bypass defenses

- Breaks defenses’ assumptions
  - Any triggered sample misclassified
- Increases search space for MNTD
- Multi-class defense unfit for binary

Potential Countermeasures

- Exhaustively scan selective backdoor for each malware family
- Increase malware homogeneity with better representations
- Collect benign samples from reliable sources

Contributions of Jigsaw Puzzle

- Selective: Protect one malware family but not others
- Stealthy: Bypass SOTA defenses
- Realizable: Keep functionality of triggered malware
- Dataset and code are available upon request: bit.ly/Jigsaw-Oakland

Backup Slides

Loss Function for Alternate Optimization

Cross entropy loss: expected selective effect

L1 regularization: minimize trigger size

$$\begin{cases} \boldsymbol{m} = \arg\min_{\boldsymbol{m}} l(\boldsymbol{x}^*, y^*; \boldsymbol{\theta}^*) + \lambda_4 \cdot \|\boldsymbol{m}\|_1 \\ \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} l(\boldsymbol{x}^*, y^*; \boldsymbol{\theta}) + v \cdot l(\boldsymbol{x}, y; \boldsymbol{\theta}) \end{cases}$$

Cross entropy loss for poisoning set

Cross entropy loss for original training set