Improving Robustness of ML Classifiers against Realizable Evasion Attacks Using Conserved Features

Liang Tong[1], Bo Li[2], Chen Hajaj[3], Chaowei Xiao[4], Ning Zhang[1], Yevgeniy Vorobeychik[1]

August 13, 2019

[1]Washington University in St. Louis

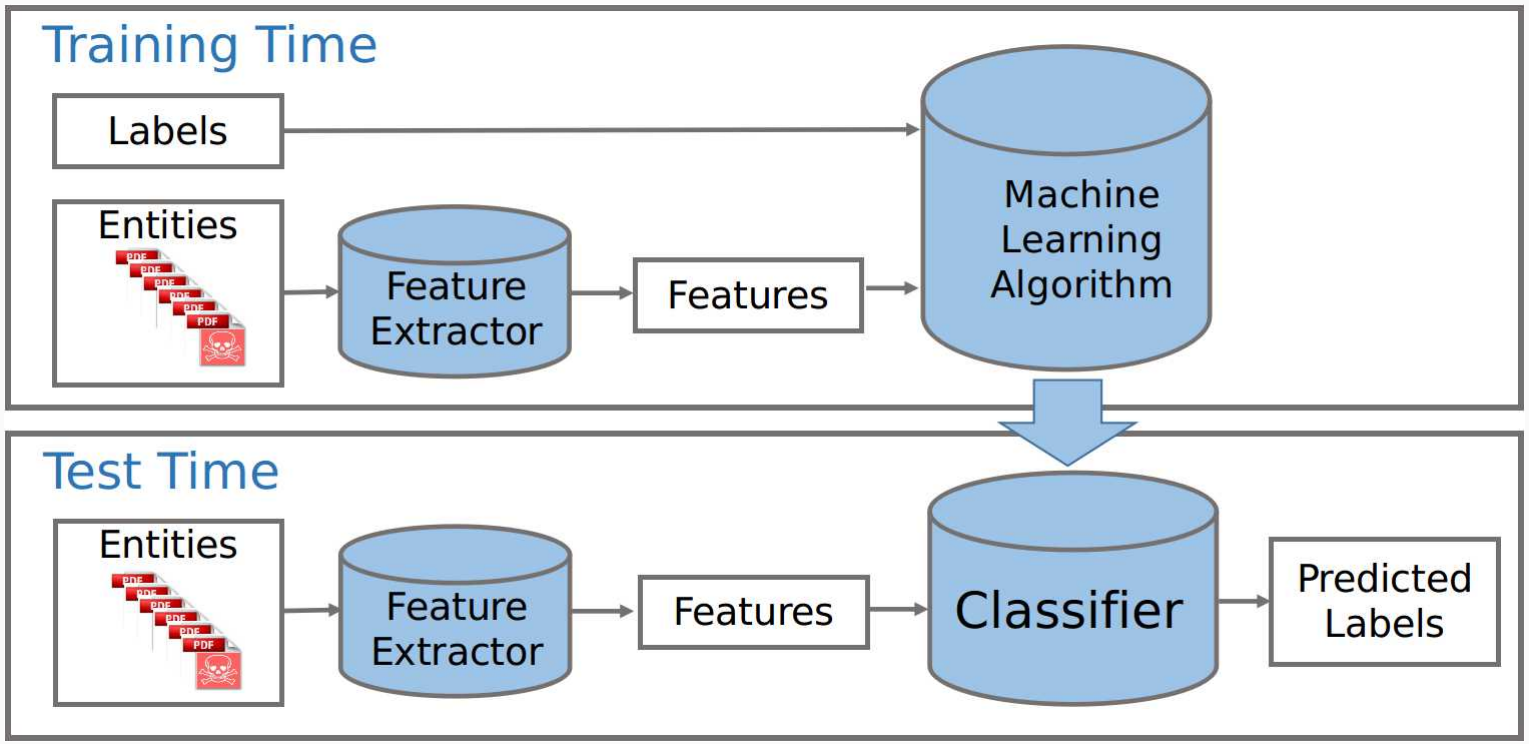
[2]University of Illinois at Urbana-Champaign

[3]Ariel University

[4]University of Michigan

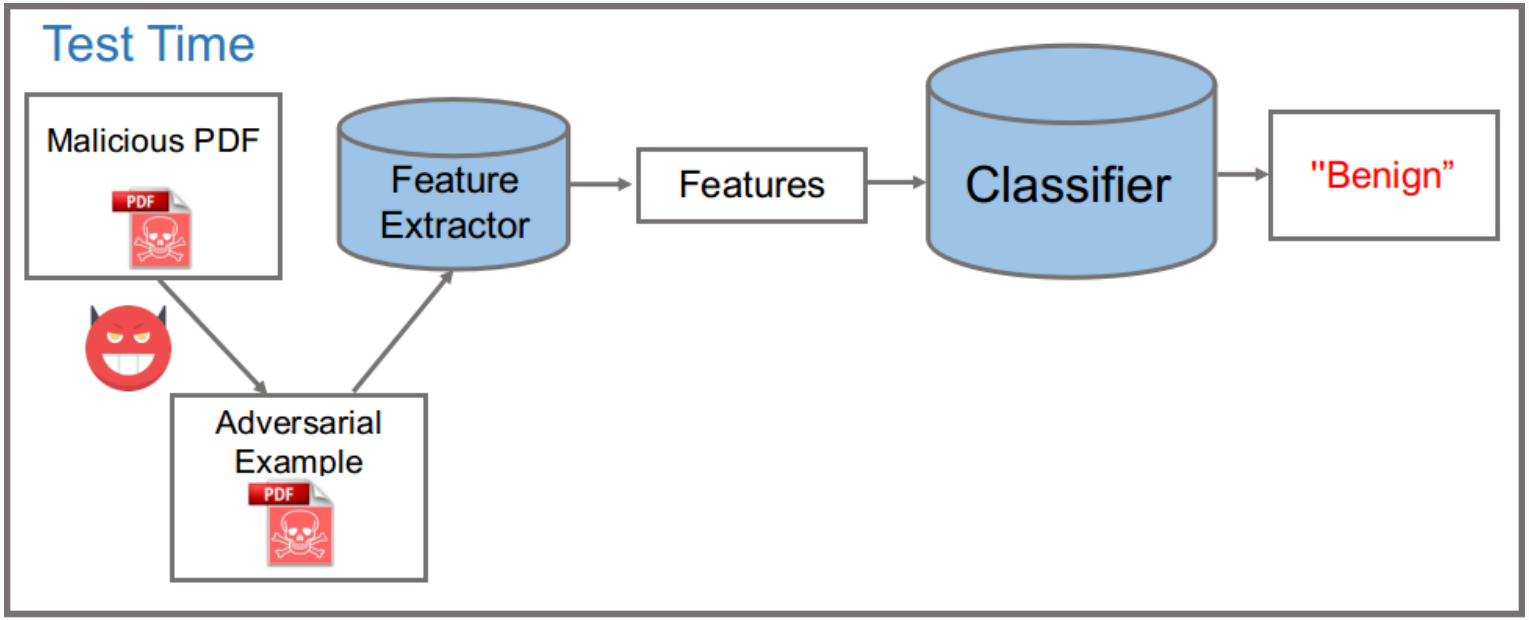
Introduction

Machine Learning in Security



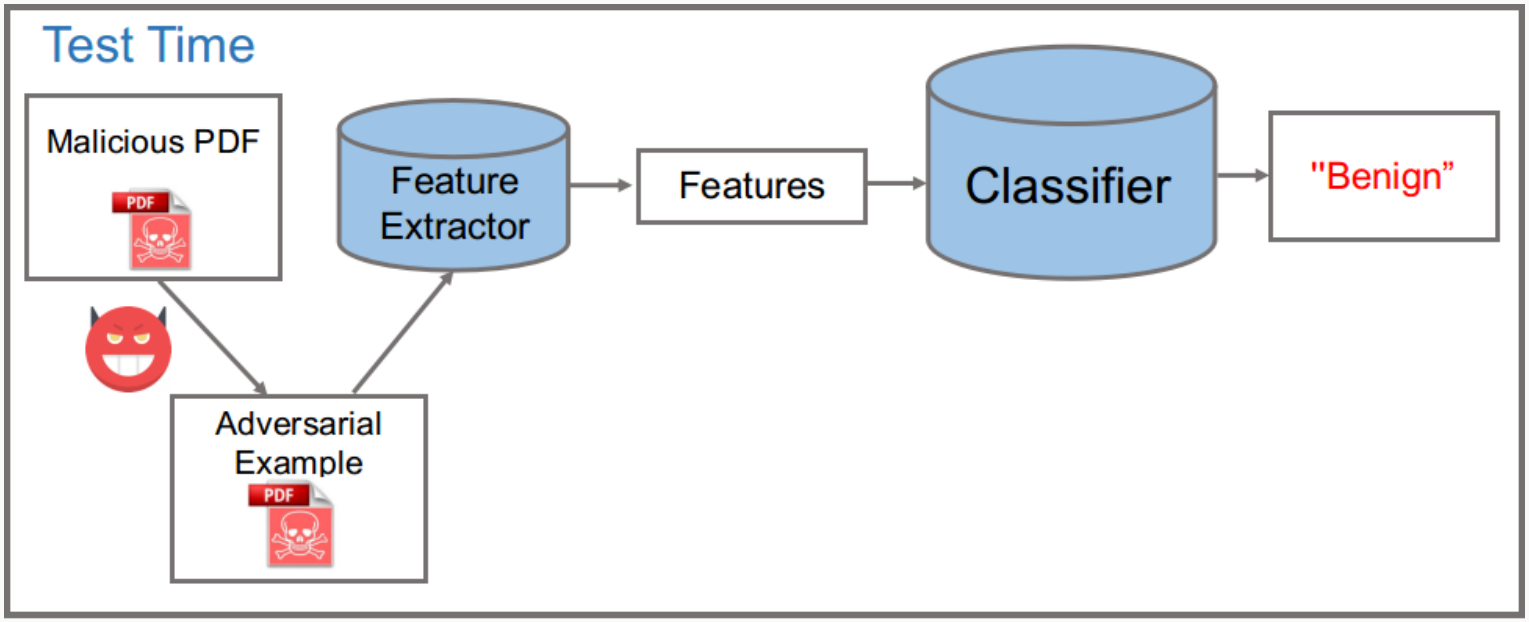
- Detection is a fundamental problem in cybersecurity.
  - e.g. Malware, intrusion, spam, phish
- Natural to use Machine Learning (ML) for these applications.

Adversarial Evasion Problem



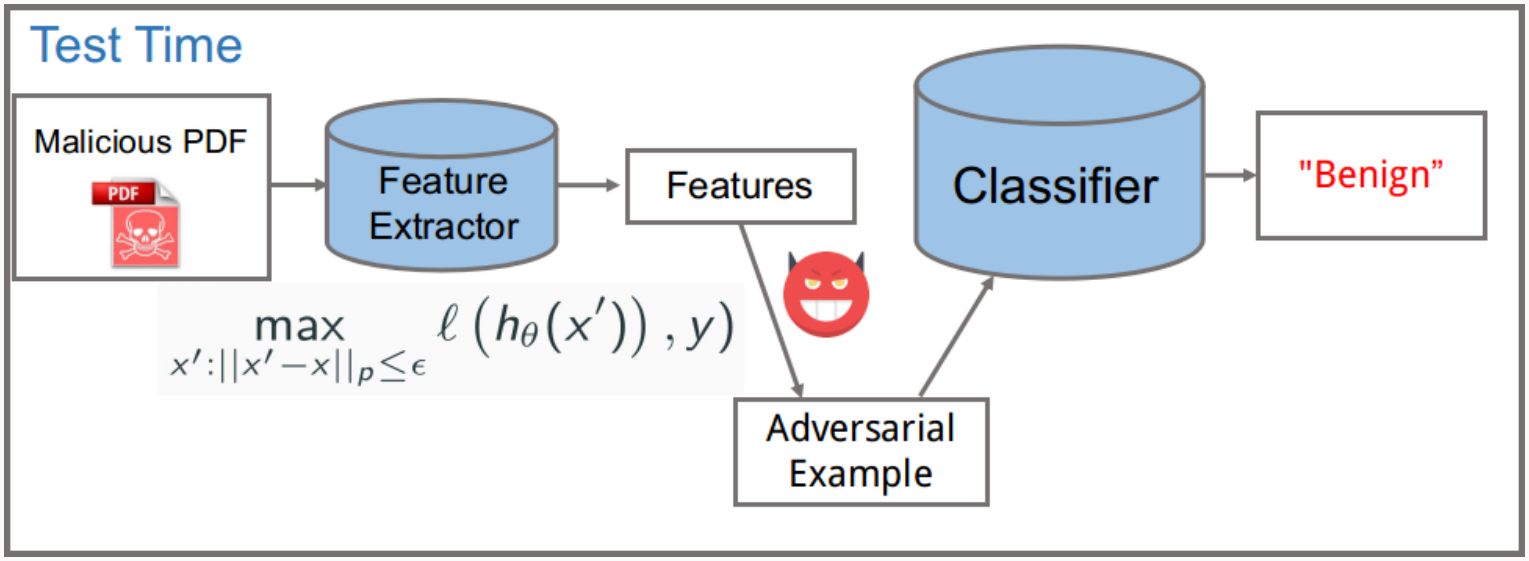
- ML-based techniques are often susceptible to *adversarial examples* at test time.
  - Attackers can manipulate malicious samples to look benign and fool a classifier.

Realizable Attacks



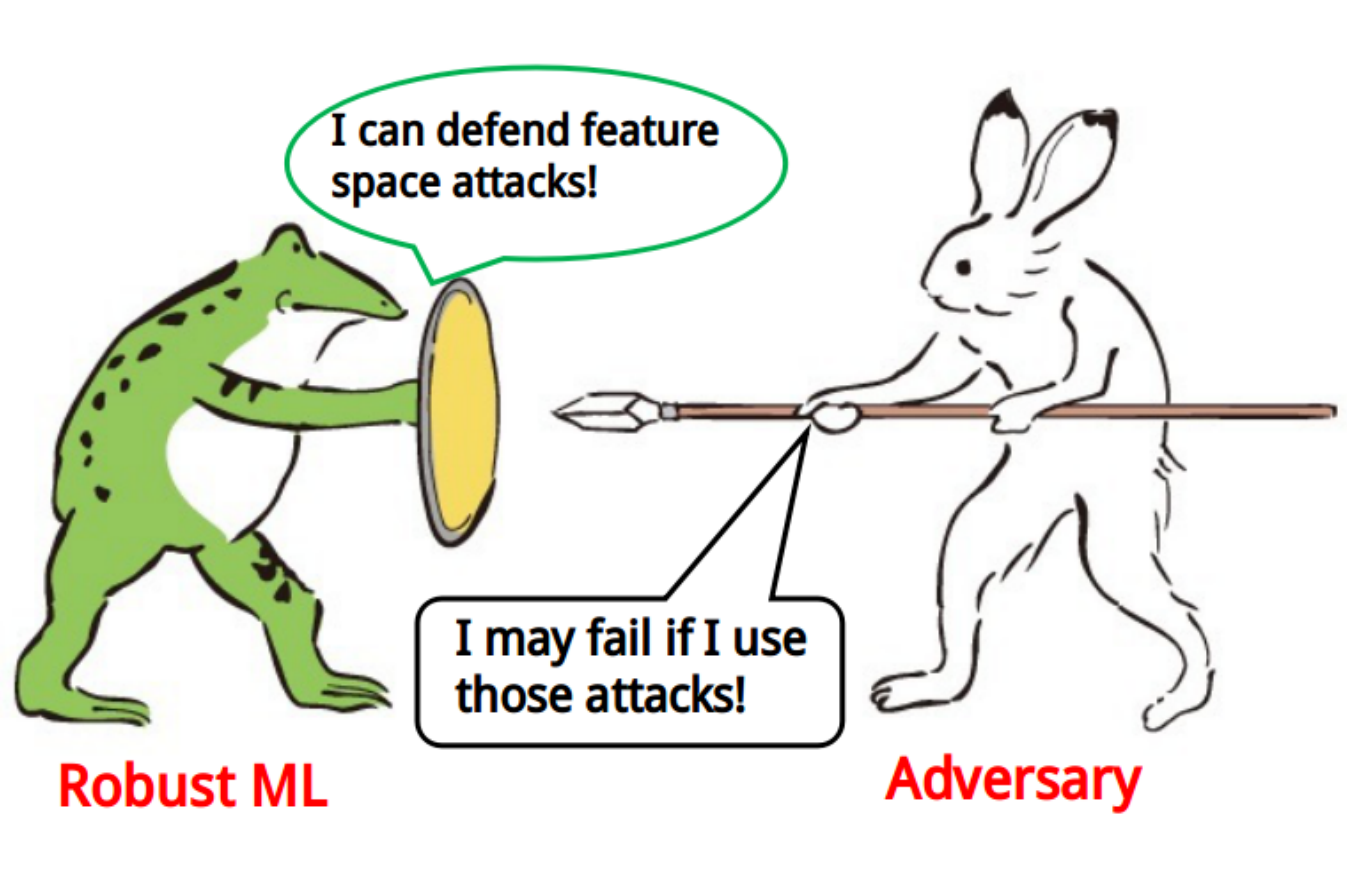
- Modify the actual entity.
  - e.g., produce a valid PDF file or executable file.
  - Features are subsequently extracted for ML.
- Have actual malicious effect (e.g., verified by a sandbox) but the feature vector is classifed as benign.

Feature Space Attacks



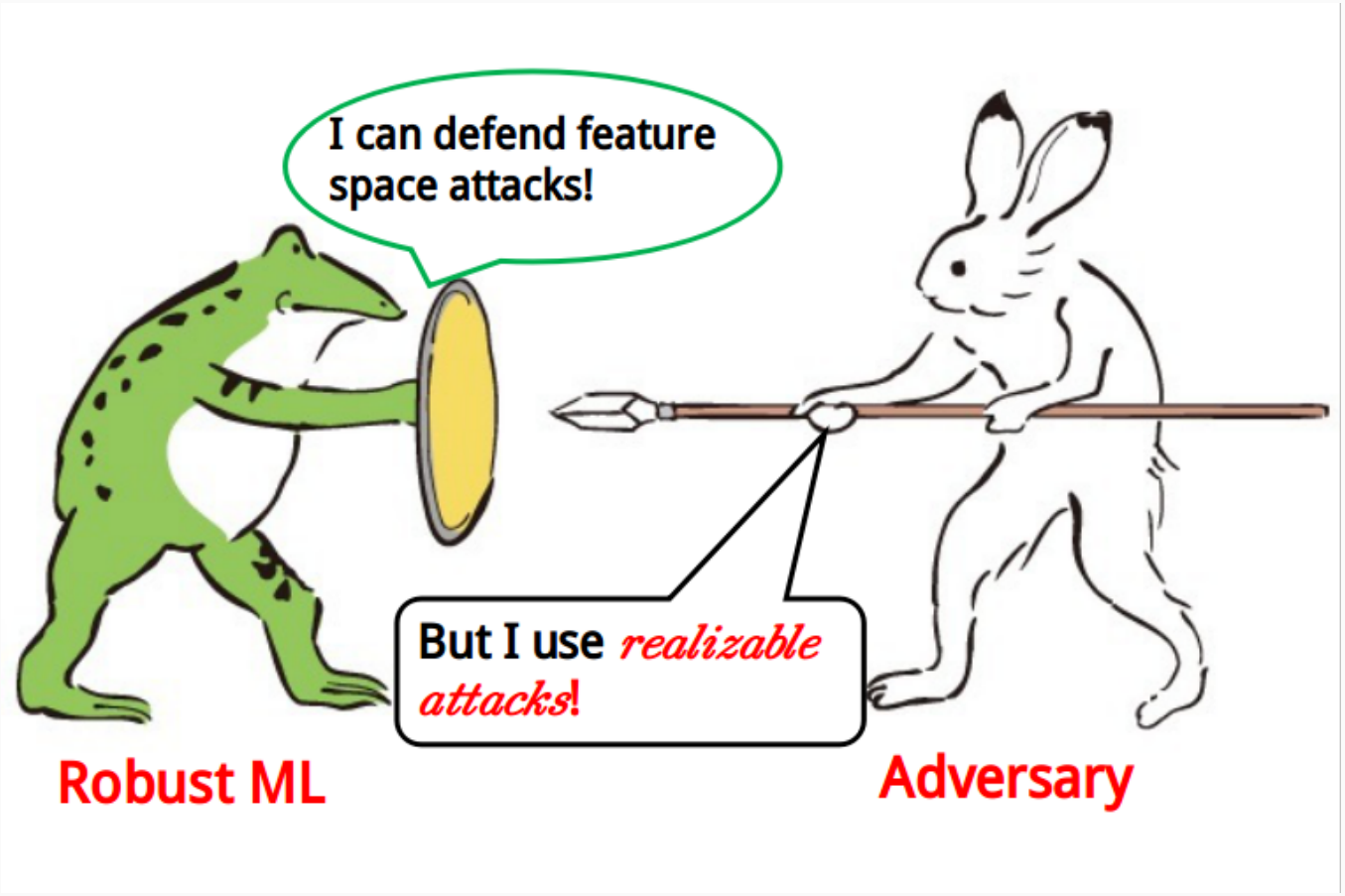
- An abstraction of realizable attacks.
  - Directly work on features instead of entities. May not be realizable.
  - Use an ℓ_p norm to measure the cost of modifying original examples.

Robust ML



- Essentially most approaches for robust ML leverage feature-space attack models. e.g., robust optimization, adversarial training.

Motivation: Is Robust ML really robust?



- Suppose we learn a Robust ML against a feature space attack model. Is it robust against realizable attacks?

Contribution

- **Model Validation**: evalute the robustness of 'Robust ML' against realizable attacks.
  - Robust ML using feature-space models may fail to provide adequate robustness against realizable attacks.
- **Model Refinement**: 'fix' the feature-space attack models by using conserved features.
- **Generalized Robustness**: explore to which extent ML robustness can be generalized to multiple distinct realizable attacks.

Methodology and Experiments

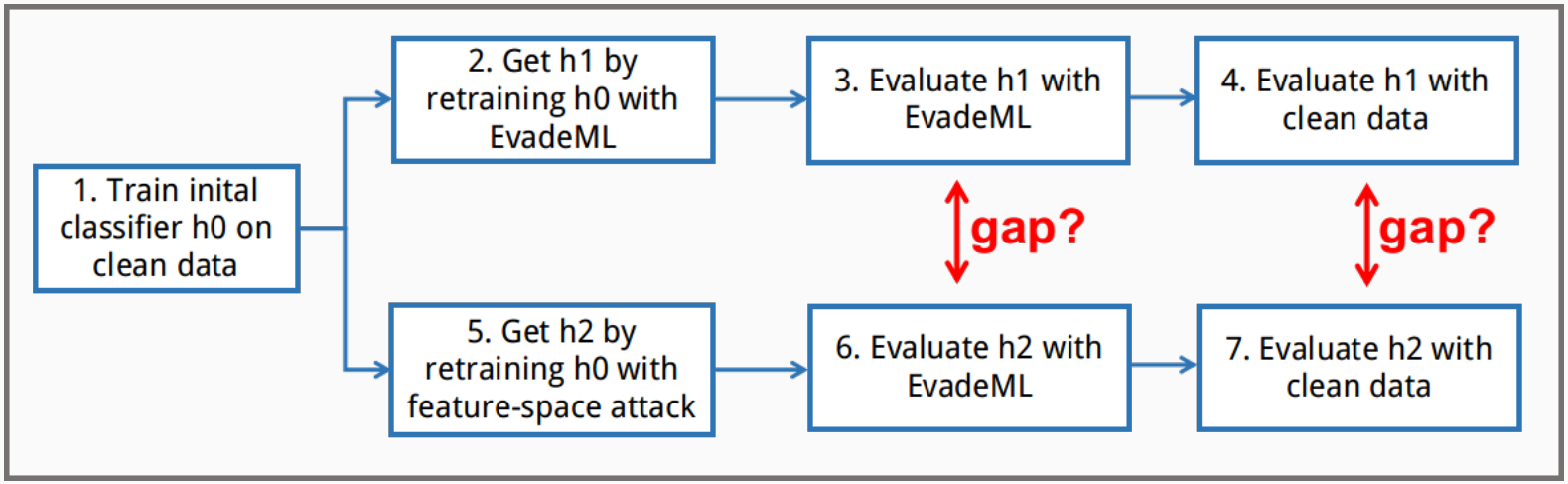
A Case Study on PDF Malware Detectors

- **Content-based** detectors: use features based on content information (e.g. size of a PDF file)
  - PDFRate-R: 135 normalized features (real-valued)
  - PDFRate-B: 135 binarized features
- **Structure-based** detectors: use binary features based on existence of a collection of object paths
  - SL2013: 6,087 paths
  - Hidost: 961 paths

Attacks and Defense

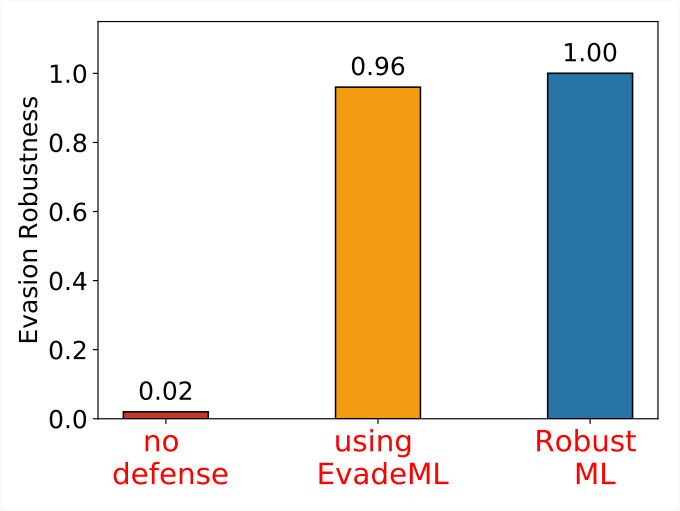
- **Realizable attack**: *EvadeML (Xu et al., NDSS)*.
  - Automatically evades a PDF classifier by using genetic programming.
  - Works on both structure- and content-based detectors.
- **Feature-space attack model**: *multi-objective optimization*.
  - The modified feature vector is predicted as benign as possible.
  - The modification cost (measured with an ℓ_p norm) is minimized.
- **General defense**: *iterative retraining*.
  - Iteratively uses an attack to produce adversarial examples, then adds them into training data and retrain.
  - Works for both realizable and feature-space attacks.

Model Validation: Framework



- Evaluation Metrics
  - Adversarial data: *robustness = 1 - success rate of EvadeML*
  - Clean data: ROC (receiver operating characteristic) curve.

Model Validation: Real-valued and Content-based



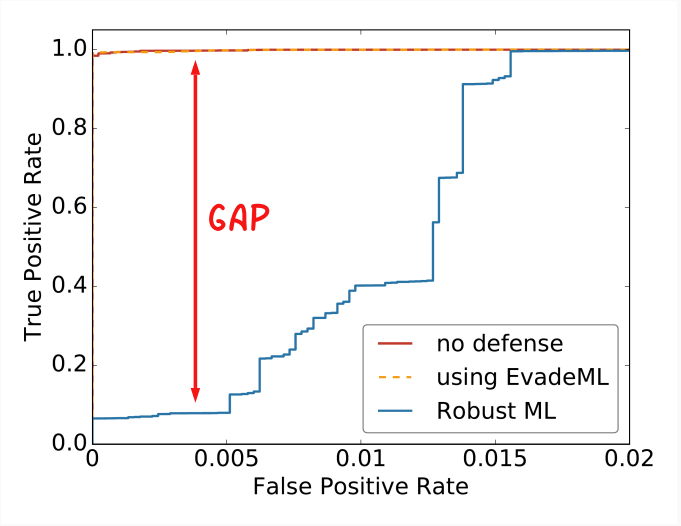


Figure 1: Left: evasion robustness. Right: ROC curve.

- Original: 2% evasion robustness.
- After defense: ∼100% evasion robustness.
  - Robust ML with feature-space model works but degrades performance on clean data!

Model Validation: Structure-based

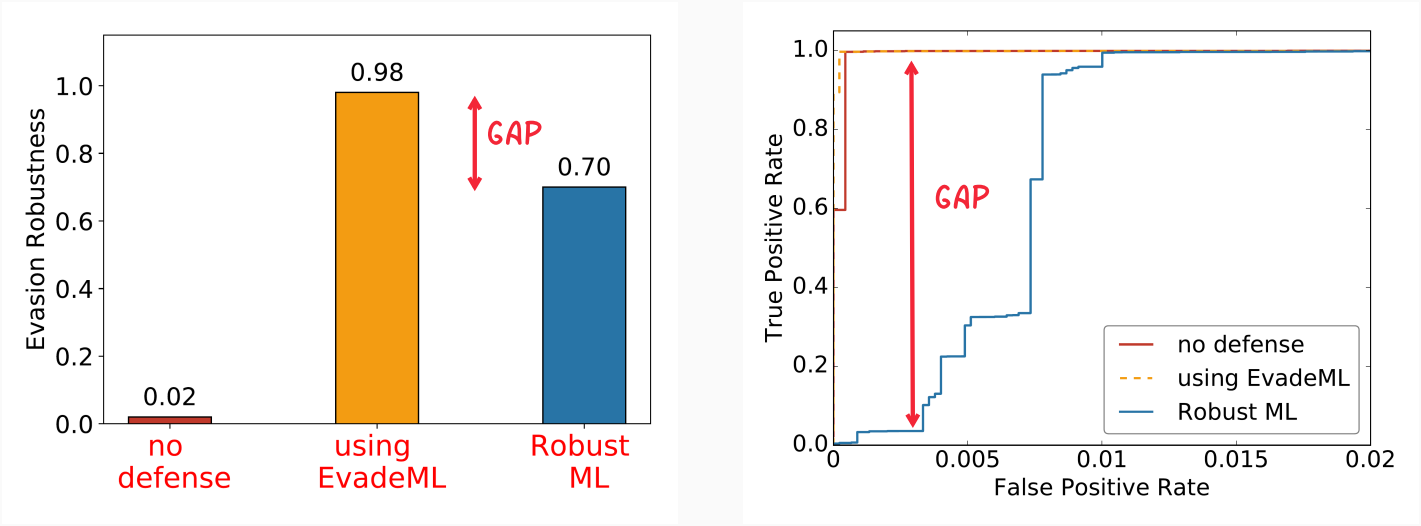


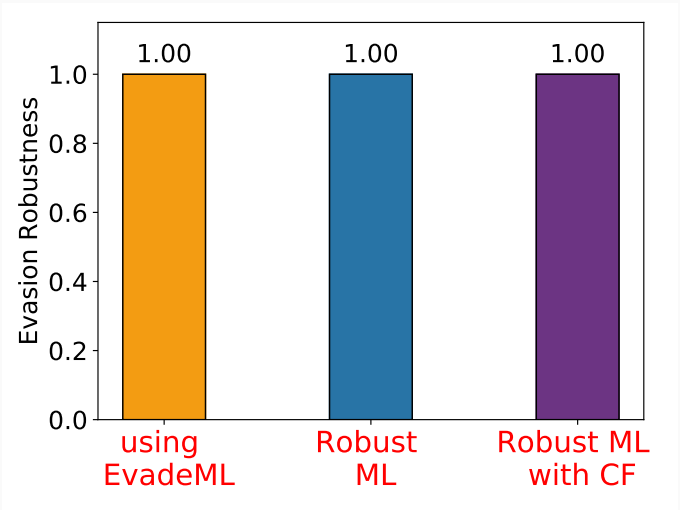
Figure 2: Left: evasion robustness. Right: ROC curve.

- Original: 2% evasion robustness.
- Defense using EvadeML: 98% evasion robustness.
- Feature-space Robust ML: 70% evasion robustness and degradation on clean data.
- Robust ML using feature-space models is not perfect. Can we fix it by creating a minimal anchoring?

Model Refinement: Conserved Features

- Conserved features: a subset of features which compromise malicious functionality if they are removed.
  - Paths to objects which contain malicious codes.
  - Paths objects which break the PDF if they are removed.
- Identifying conserved features: systematically manipulating each object in a PDF file and checking the maliciousness.
- Existence of conserved features: we identified 4$\sim$8 conserved features for each detector.
- Feature-space attacks with conserved features: conserved features are preserved in evasive instances.

Model Refinement: Binarized Content-based



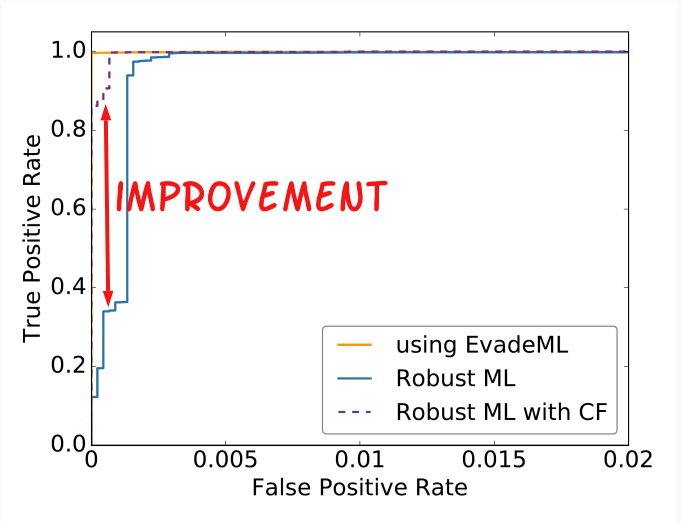


Figure 3: Left: evasion robustness. Right: ROC curve.

- Defense using EvadeML: 100% evasion robustness.
- Feature-space Robust ML: 100% evasion robustness and performance degradation on clean data.
- Feature-space Robust ML with conserved features: 100% evasion robustness and improves ROC.

Model Refinement: Structure-based

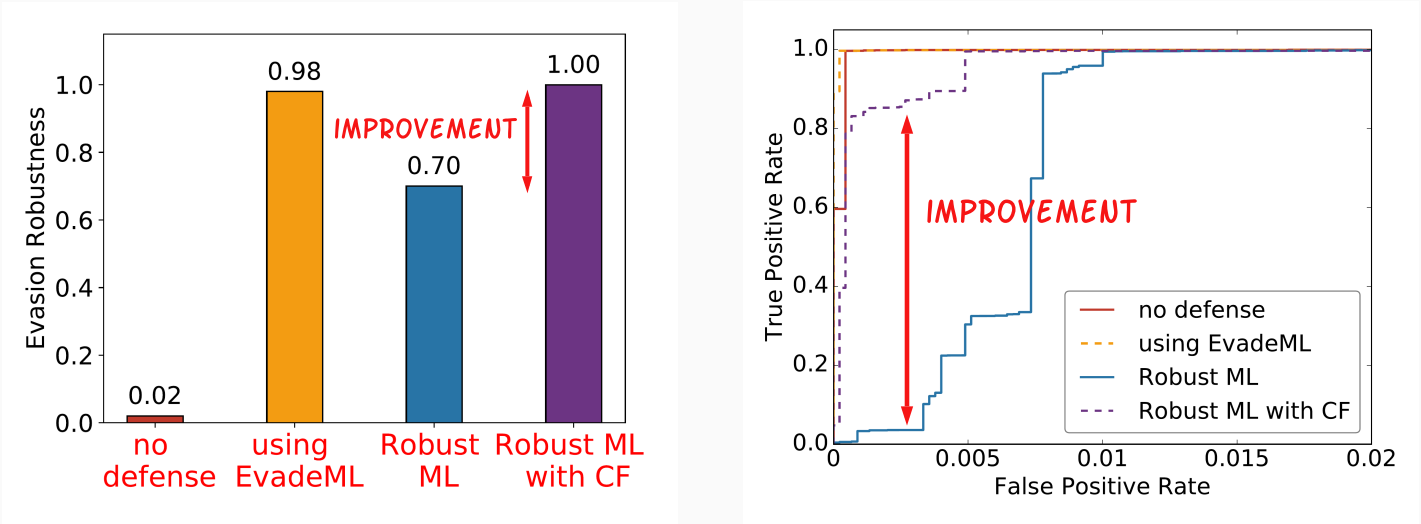


Figure 4: Left: evasion robustness. Right: ROC curve.

- Defense using EvadeML: 98% evasion robustness.
- Feature-space Robust ML: 70% evasion robustness and performance degradation on clean data.
- Feature-space Robust ML with conserved features: 100% evasion robustness and significant improvement on clean data.

Generalized Robustness

So far, evaluation and baseline defense used EvadeML.

- Is ML hardened with EvadeML effective against other realizable attacks?
- Is ML hardened with a feature-space model of attacks (using conserved features) generally effective against realizable attacks?

Generalized Robustness: Mimicry+

Realizable attack on content-based classifiers

An improvement of *Mimicry Attack* (Srndic & Laskov, Oakland).

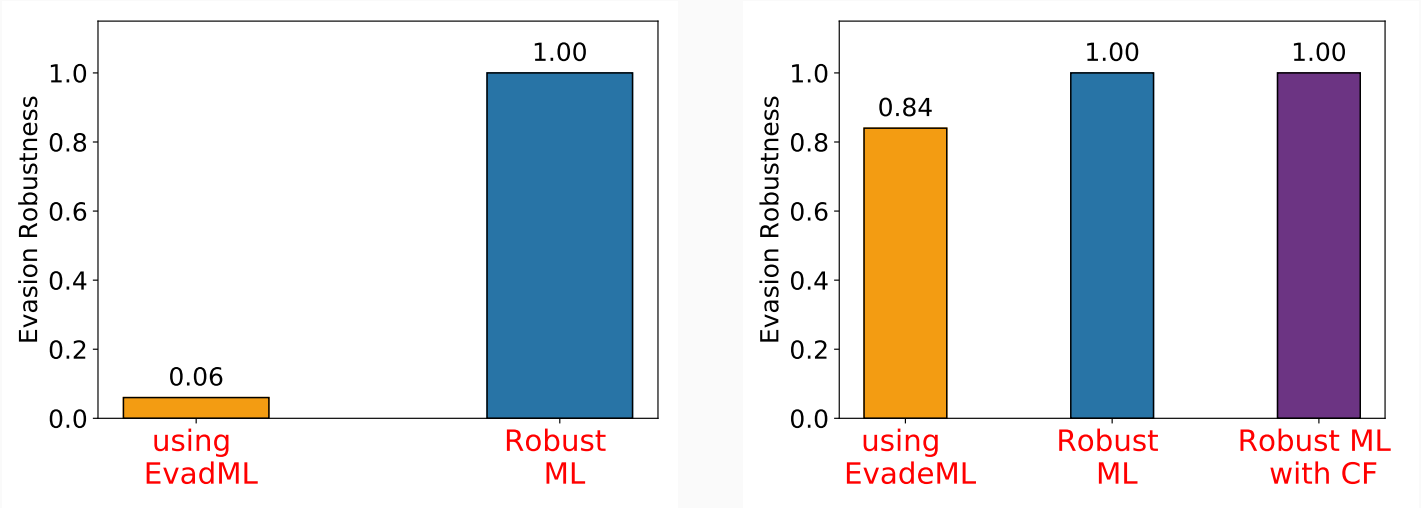


Figure 5: Left: real-valued. Right: binarized.

- Hardening against EvadeML may fail to be robust to Mimicry+.
- Robust ML (w/o conserved features) is still robust to Mimicry+.

Generalized Robustness: Reverse Mimicry

Realizable attack that requires zero knowledge of target classifier (Maiorca et al., ASIACCS)

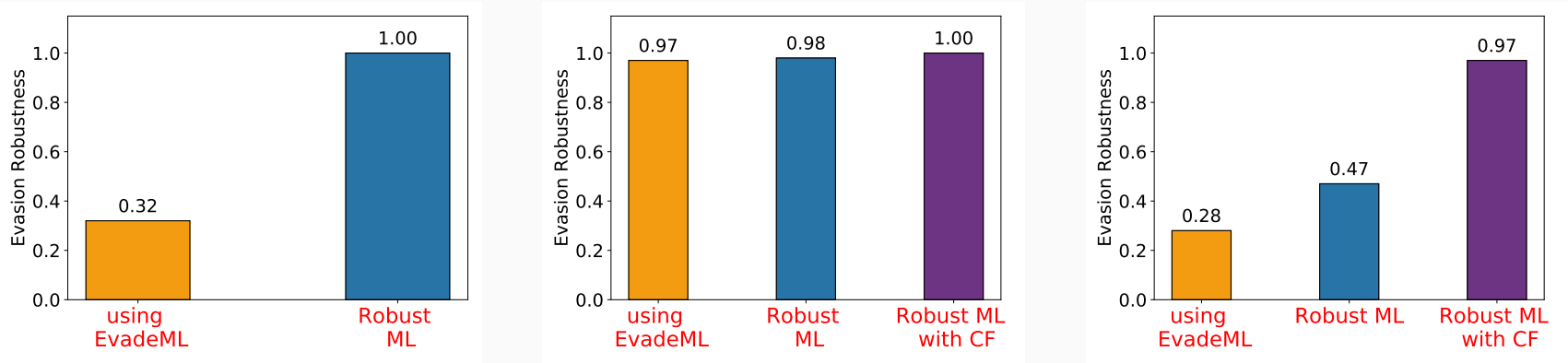


Figure 6: Left: real-valued content-based. Middle: binarized content-based. Right: structure-based

- Hardening against EvadeML may fail to be robust to Reverse Mimicry.
- Robust ML w/ conserved features is still robust to Reverse Mimicry.

Generalized Robustness: Custom Attack

Exploitation of a feature extraction bug of the content-based classifiers.

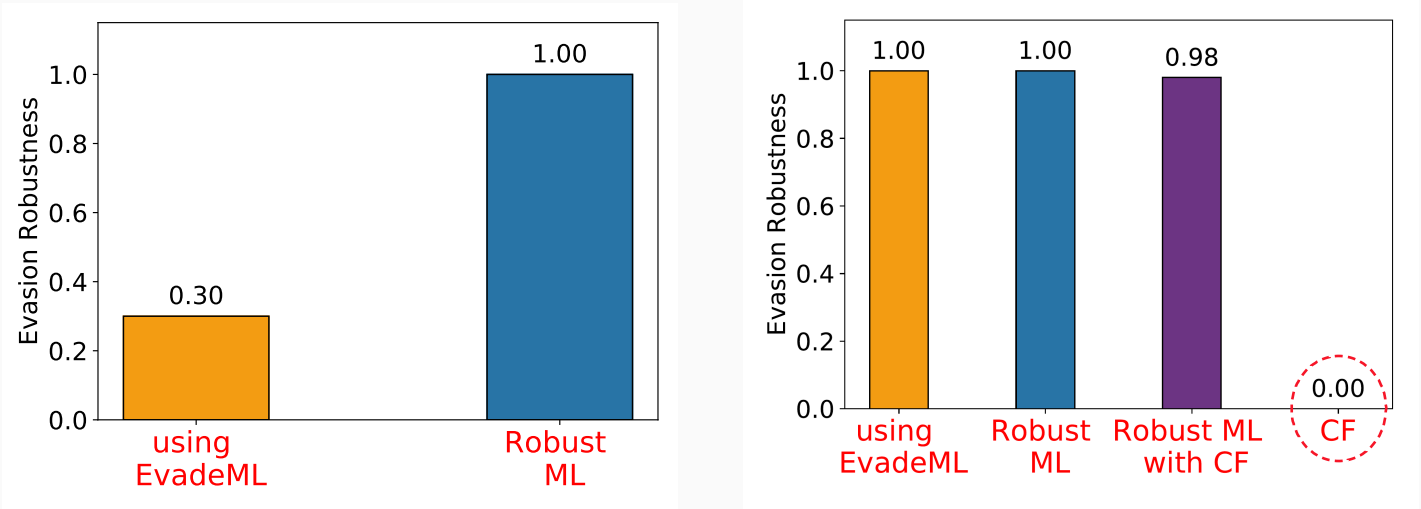


Figure 7: Left: real-valued. Right: binarized.

- Defeats detector hardened using EvadeML.
- Defeats conserved features of binarized content-based detector.
- All feature-space approaches remain robust.

Conclusion

Summary

- Robust ML methods which assume direct modification of features and measures cost of adversarial noise as norm are sometimes, but not always fully effective against real attacks.
- We can fix the model by identifying and using conserved features to anchor the abstract attack model in the problem domain.
- Robust ML using feature space models (after the fix) exhibit more general robustness than methods hardened only against a particular (strong, adaptive) realizable attack.

Questions?