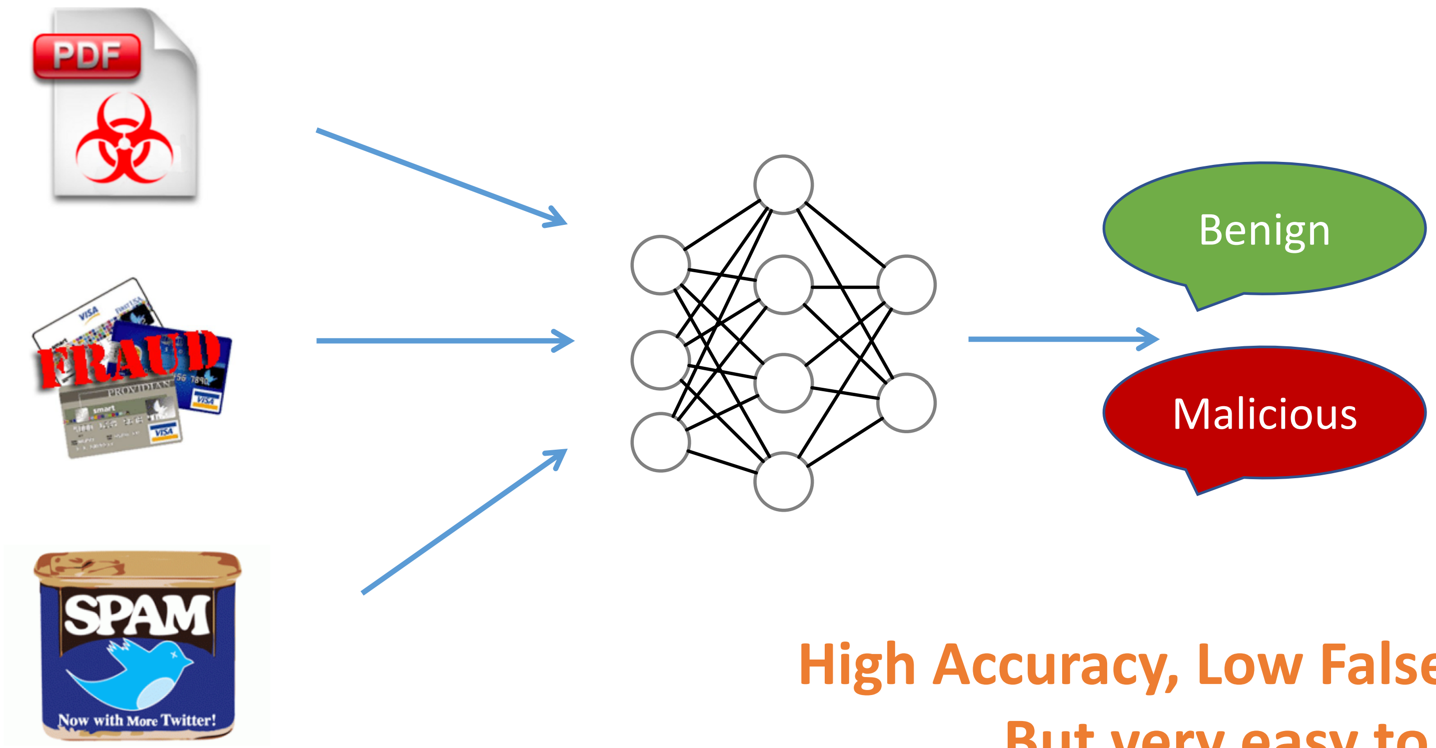
On Training Robust PDF Malware Classifiers

Yizheng Chen, Shiqi Wang, Dongdong She and Suman Jana

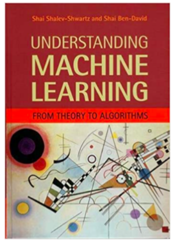
Columbia University

Security Classifiers



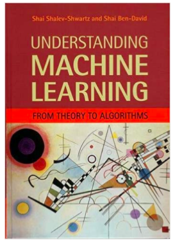
High Accuracy, Low False Positive Rate
But very easy to evade

Evading Gmail’s PDF Malware Classifier

Inserted /Root/Pages from 

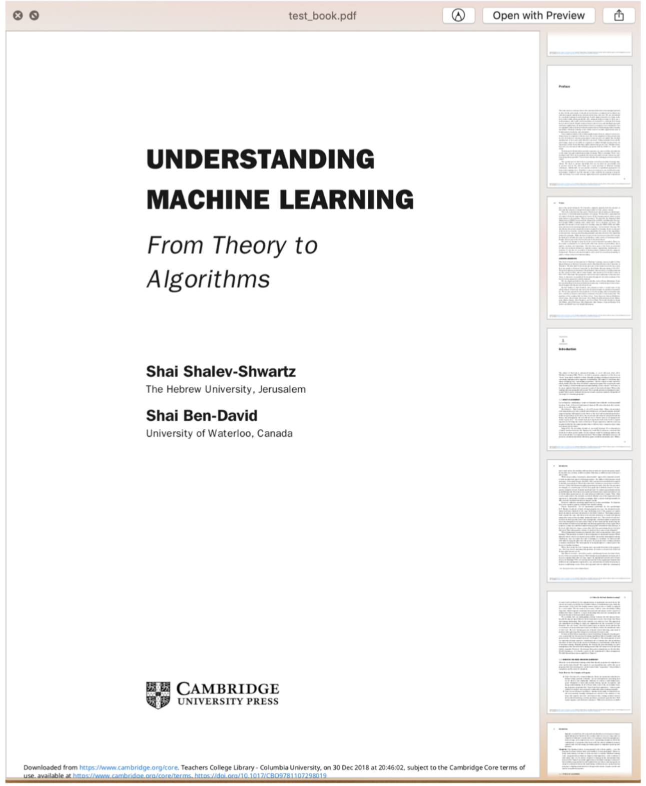
to 

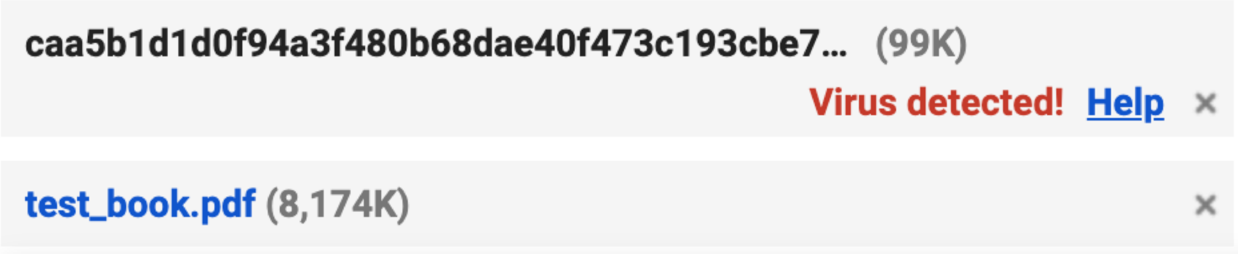
Evading Gmail's PDF Malware Classifier

Inserted /Root/Pages from 

to 

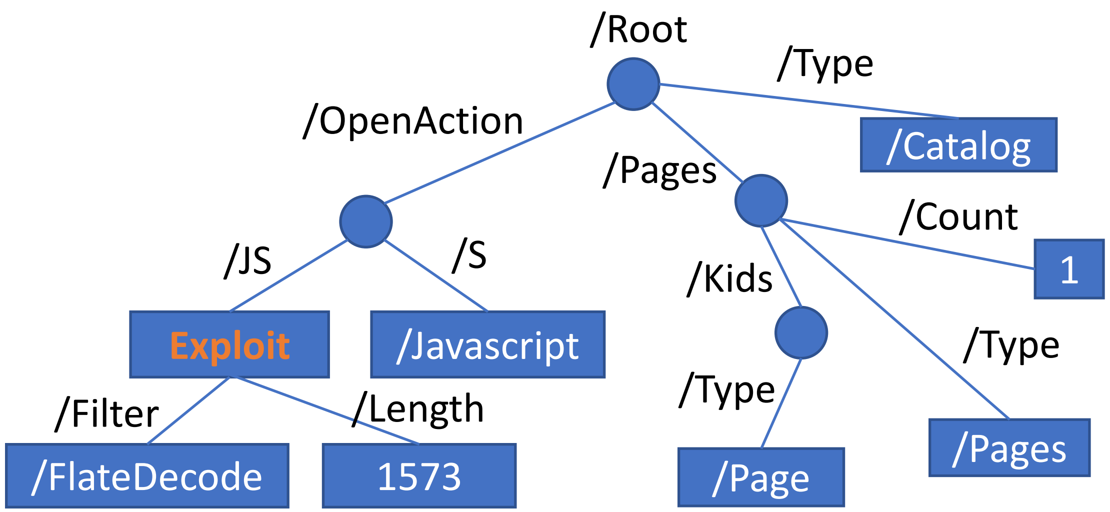
→ → Benign



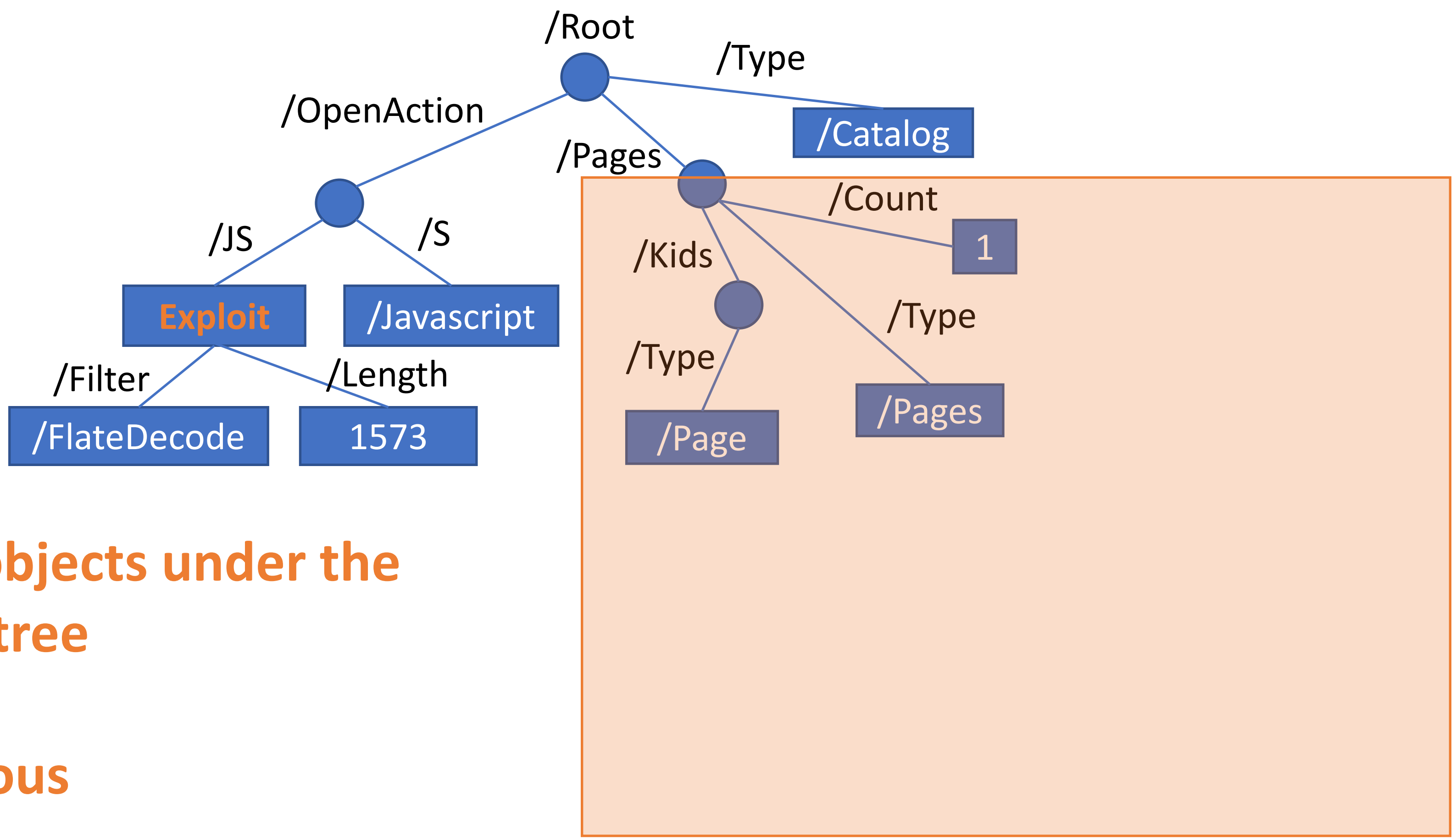


The PDF is still malicious

What Changed in the PDF Malware?



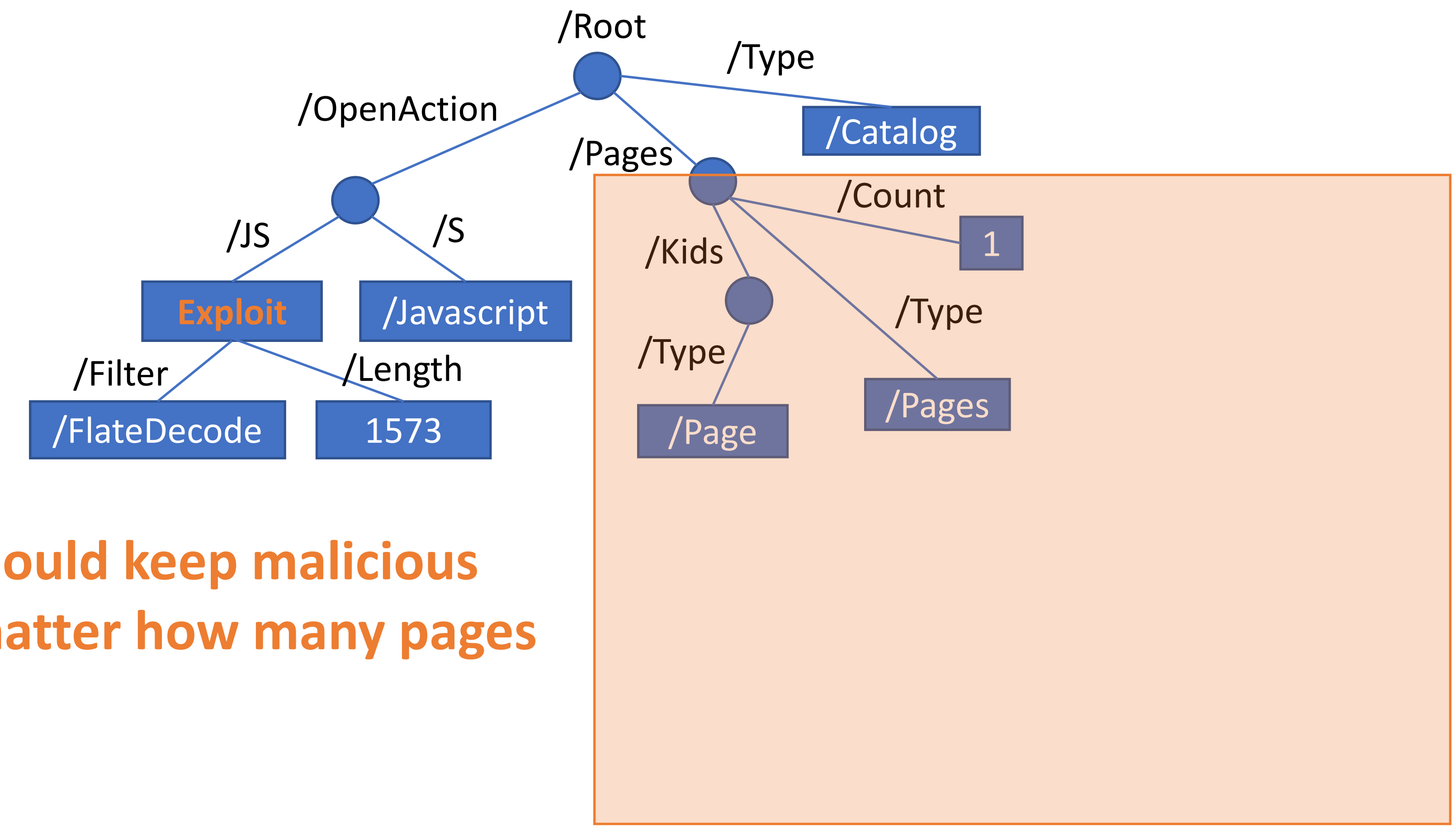
What Changed in the PDF Malware?



Inserted 12,188 objects under the /Root/Pages Subtree

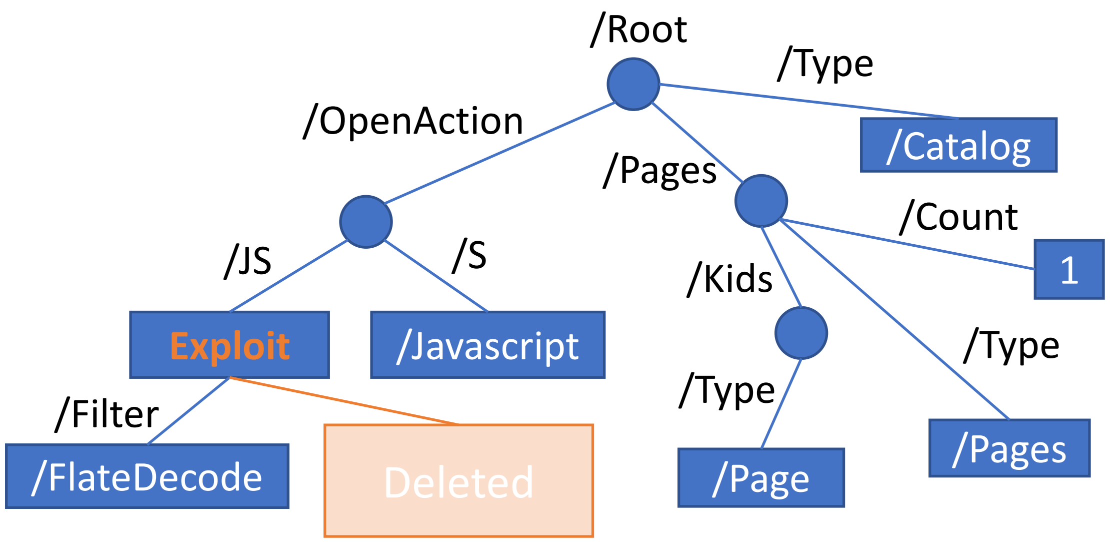
PDF is still malicious

Example Robustness Property



The classifier should keep malicious prediction no matter how many pages are inserted

Example Robustness Property



The classifier should keep malicious prediction if non-functional objects are deleted

Why are Robustness Properties Useful?

- Unbounded attackers can always evade the classifier

Why are Robustness Properties Useful?

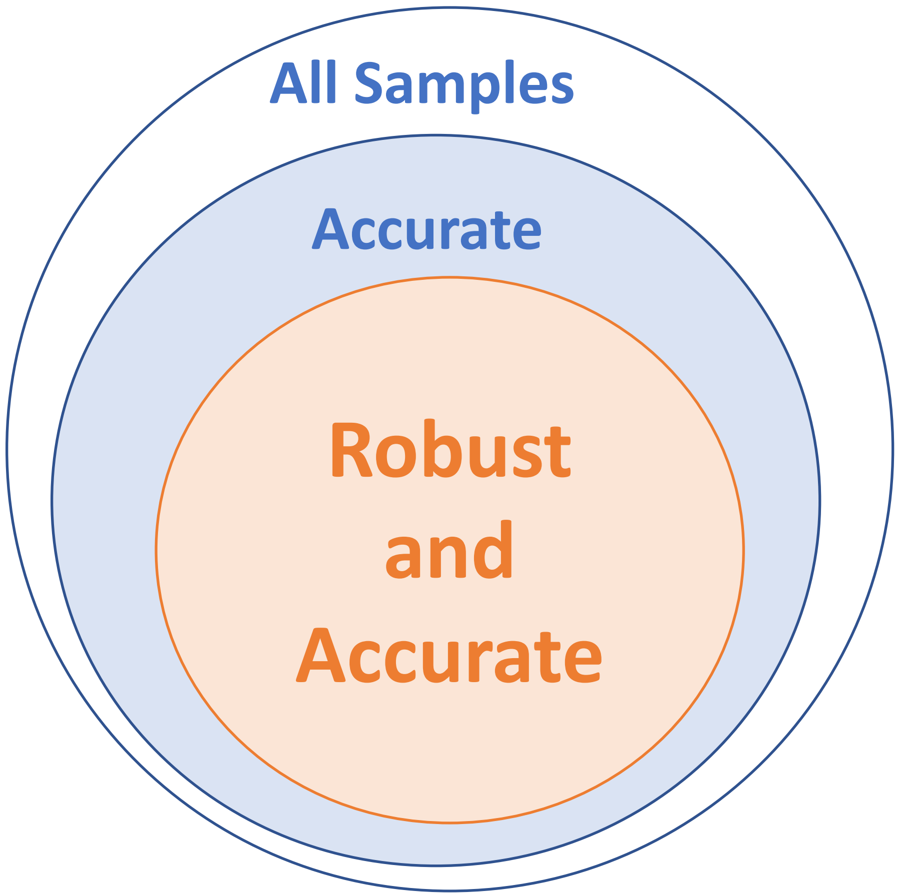
- Unbounded attackers can always evade the classifier
- Robust against reasonably bounded attackers
- Generalize to robustness against unbounded attackers

Why are Robustness Properties Useful?

- Unbounded attackers can always evade the classifier

- Robust against reasonably bounded attackers
  - Robustness Properties
  - Robust Accuracy

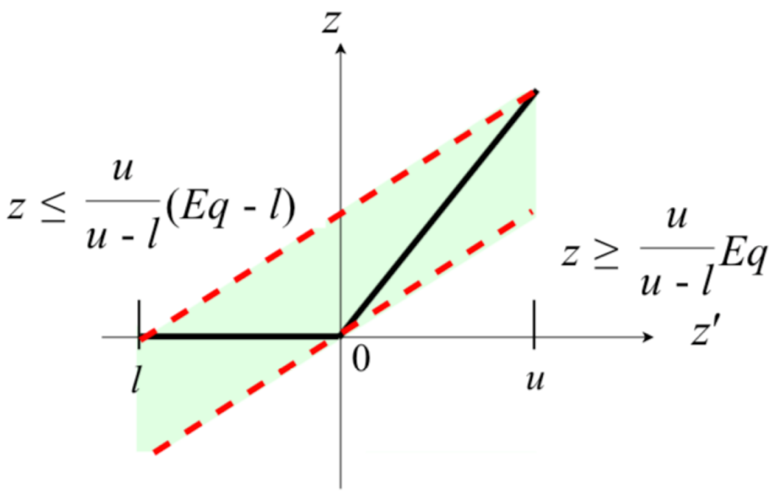
- Generalize to robustness against unbounded attackers



Robust Accuracy

- The percentage of test samples that are correctly classified against any attacker within a specified bound.
  - e.g., $L_\infty \leq 0.1$ bounded attacker against an image classifier
- Estimated Robust Accuracy (ERA) measures robustness using attacks.
  - Restricted attackers within the bound
  - Unrestricted attackers as the bound increases
- Verified Robust Accuracy (VRA) measures robustness using sound over-approximation methods.
  - Overapproximates attacks
  - Lower bound of the percentage of robust and accurate samples

Sound Over-Approximation



Symbolic Linear Relaxation

Wang et al. USENIX Security 2018, NIPS 2018.

Symbolic Linear Relaxation

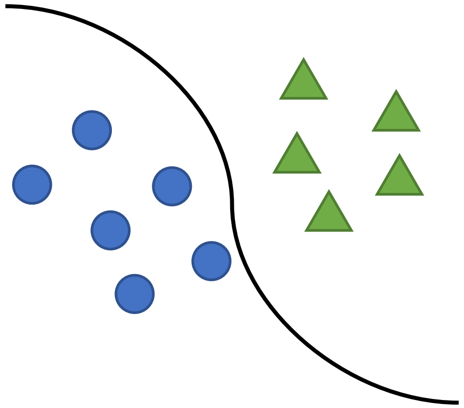
- Propagate Symbolic Intervals
- Over-approximates attacks
- Measures VRA

https://github.com/tcwangshiqi-columbia/symbolic_interval

Verifiable Training Increases VRA

Regular Training

min(errors)



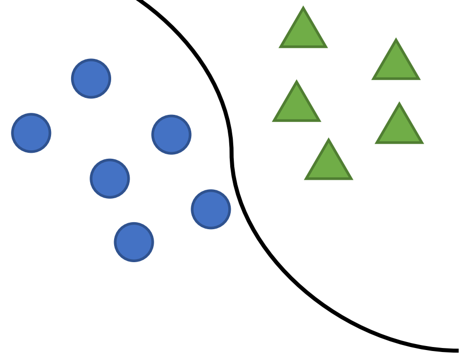
Robust Training

min(max(errors by successful evasions))

Verifiable Training Increases VRA

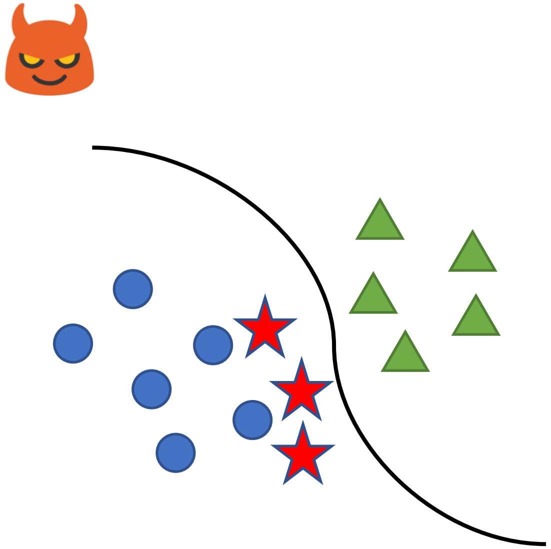
Regular Training

min(errors)

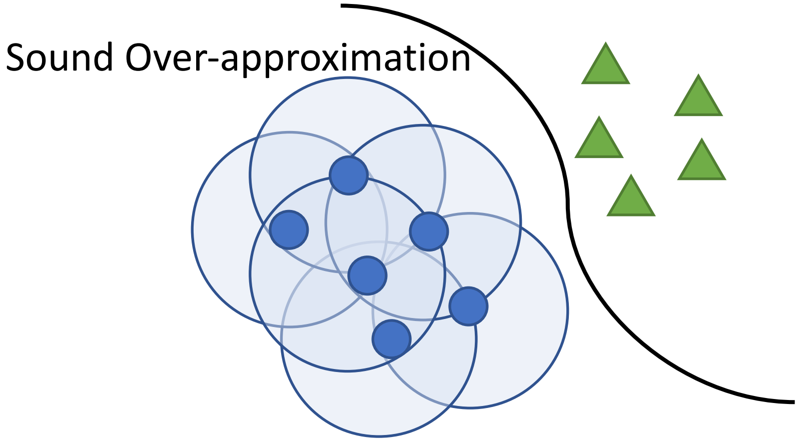


Robust Training

min(max(errors by successful evasions))



Adversarial



Verifiable

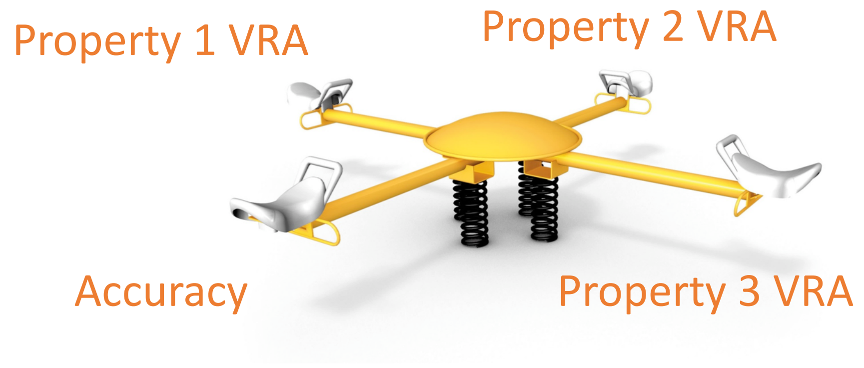
Robustness against Unknown Attacks

Challenges

- How to train a single model to be robust against different attackers?
- How to maintain low false positive rate?
- Does verifiable robustness generalize to unrestricted attackers?

Robust Against Different Attackers

- Obtain VRA for multiple robustness properties and regular accuracy
  - The underlying optimization problem is harder
- Mixed Training
  - Combined training objective
  - Mix the batches

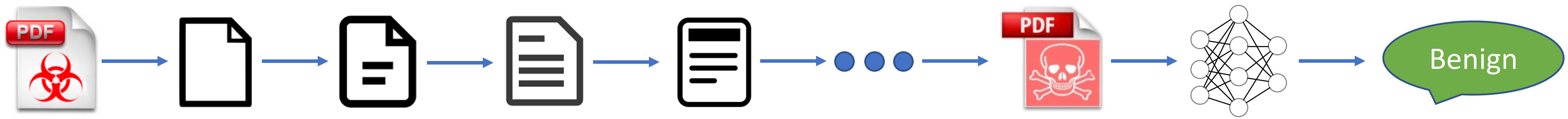


New Distance Metric

- To bound attackers that reasonably mimic real attackers
- Does not affect false positive rate

- Adversarial malware examples
  - x -> x’, s.t. f(x’) = benign and O(x’) is malicious, imperceptible by machine

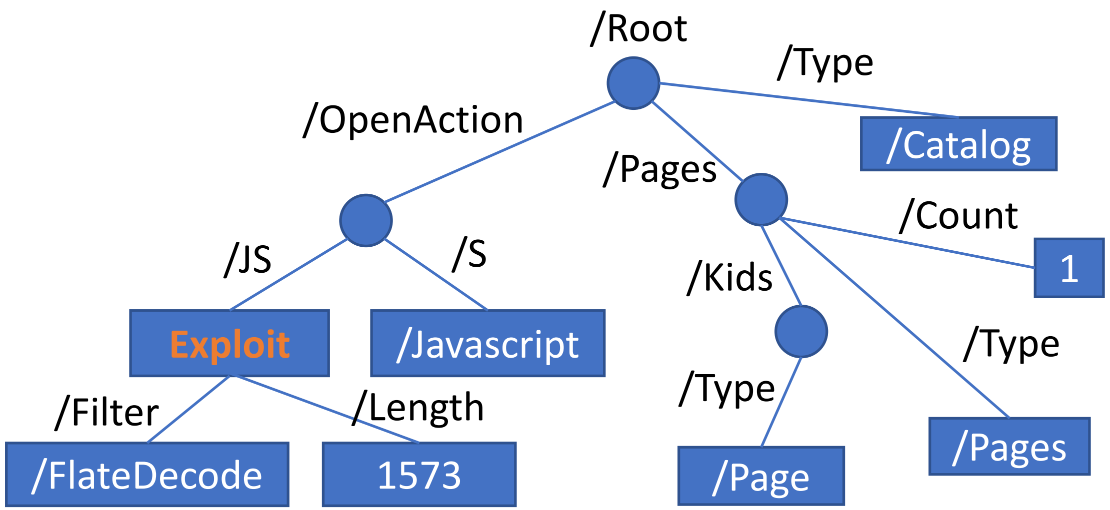
Searching for Evasive PDF Malware



- Attacks can be decomposed to building block operations
  - Feature insertion-only attacks. *Grosse et. al., Hu et al.*
  - Mimicry, merging with benign features. *Šrndić et al.*
  - Mutation operations (insert, replace, delete). *Xu et al., Dang et al.*
- Optimization
  - Greedy (Gradient Descent)
  - Genetic Evolution
  - Hill Climbing

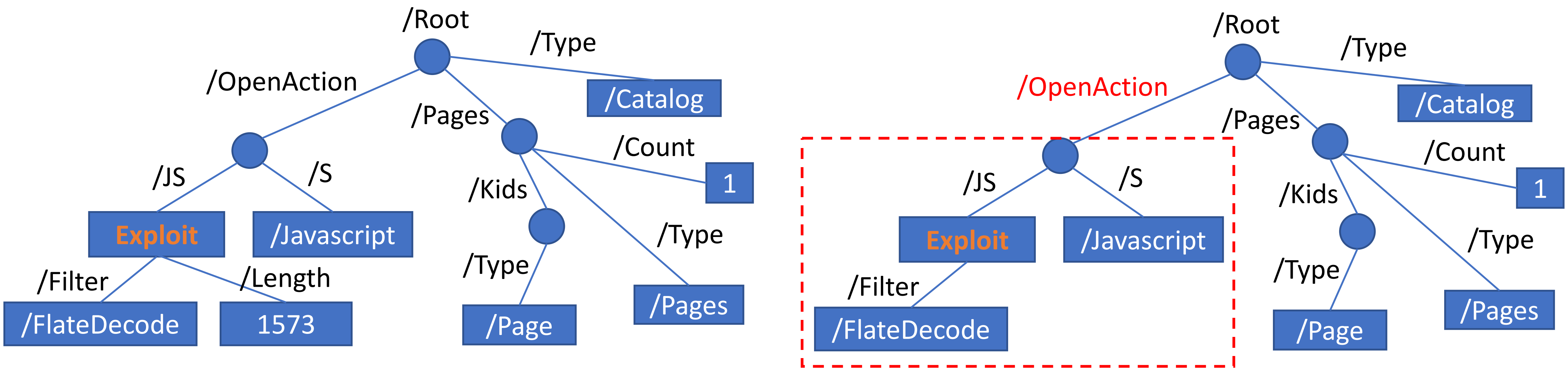
Subtree Distance

- A PDF malware variant needs correct syntax and correct semantic.
  - PDF file is parsed into a tree structure



Subtree Distance

- A PDF malware variant needs correct syntax and correct semantic.
  - PDF file is parsed into a tree structure
  - # of different subtrees under the root between variants

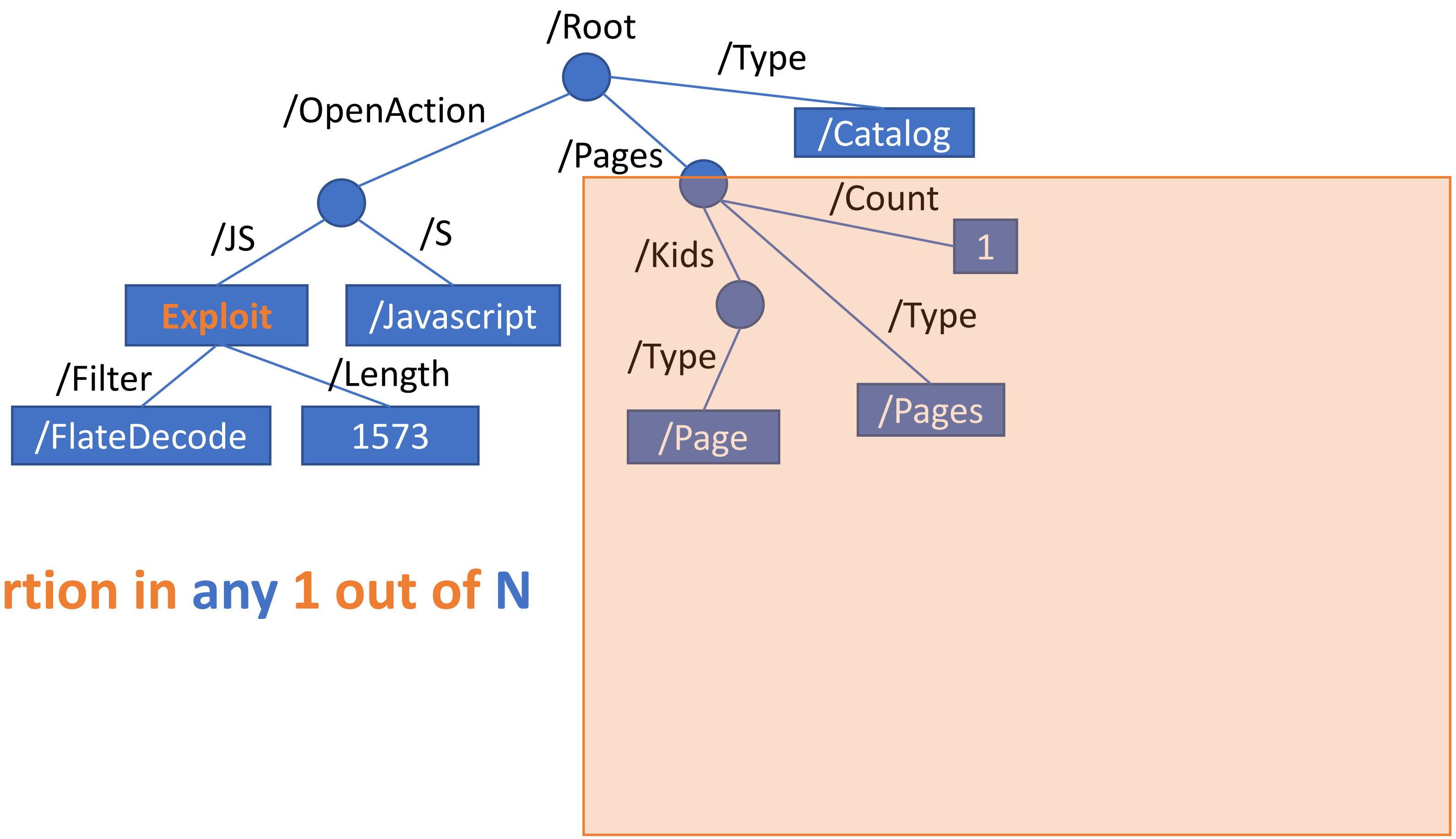


Subtree Distance One: arbitrary changes in 1 out of N subtrees under root

Building Block Robustness Properties

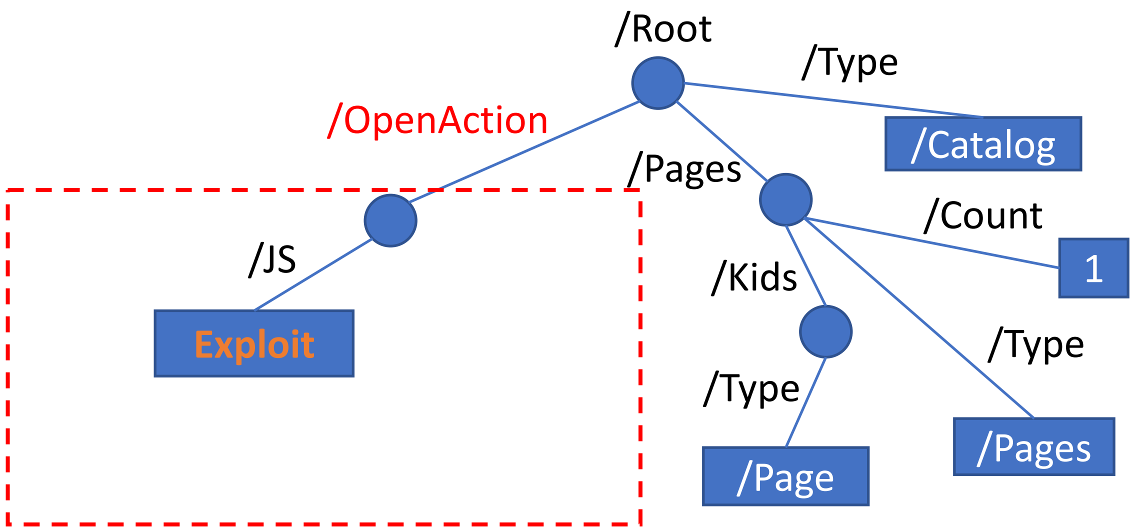
- Small subtree distance maintains low FPR
  - Subtree insertion property (subtree distance one)
  - Subtree deletion property (subtree distance one)

Subtree Insertion (Distance One)



Robust against insertion in any 1 out of N subtrees

Subtree Deletion (Distance One)



Robust against arbitrary deletion in one of the existing subtrees

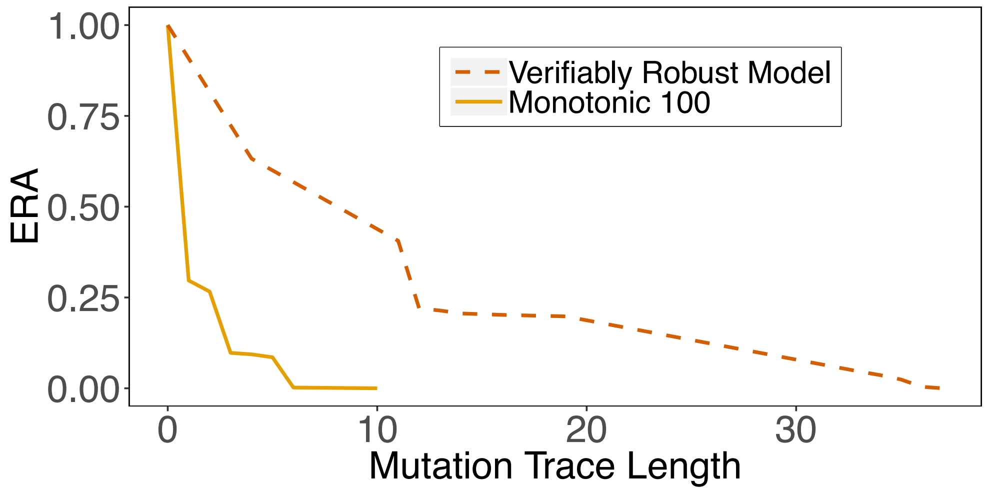
Building Block Robustness Properties

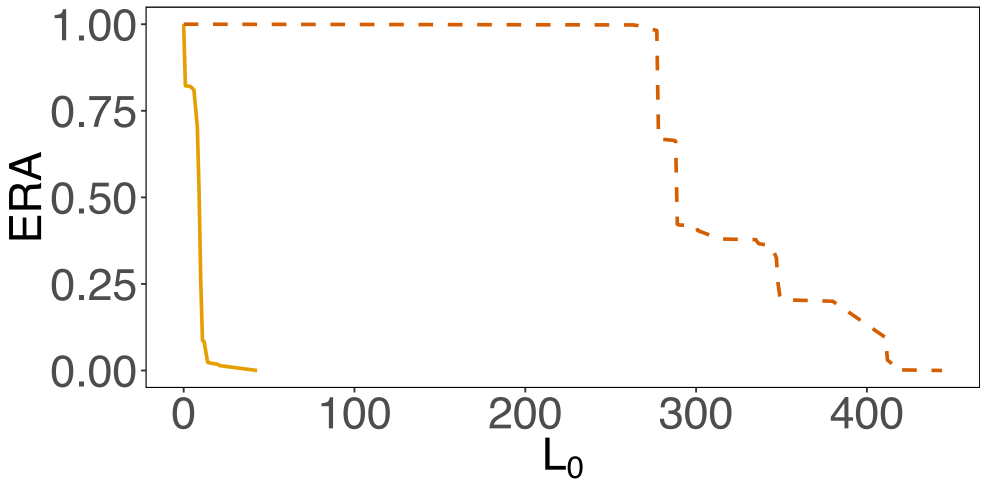
- Small subtree distance maintains low FPR
  - Subtree insertion property (subtree distance one)
  - Subtree deletion property (subtree distance one)
  - Binary path features (Hidost *Šrndić et al. NDSS 13*)

	Monotonic Classifier	Verifiably Robust Model
Accuracy	99.04%	99.74%
False positive Rate	1.78%	0.56%
Subtree Insertion VRA	99.04%	91.86%
Subtree Deletion VRA	7.67%	99.68%

- Monotonic classifier f: if $x \leq x'$, $f(x) \leq f(x')$

ERA against Adaptive Attackers





Adapt the genetic evolutionary attack *(Xu et al., NDSS 2016.)*

- Monotonic: move exploit, i.e. deletion but keep the exploit.
- Verifiably robust model: insert and delete under different subtrees.
- Our verifiably robust model requires 3.7 times more mutations and 10 times larger L0 distance to be evaded by adaptive attackers.

More Evaluations in the Paper

- 12 baseline models
  - Regular trained neural networks, adversarial training, ensemble classifiers, monotonic classifiers
- Generate evasive variants
  - 7 different attackers
  - 2 Unrestricted Whitebox Attacks (Gradient, MILP)
  - 3 Unrestricted Blackbox Attacks (Reverse Mimicry, Evolutionary, Adaptive)
- We raise the bar against unbounded attackers

Thank You

- https://github.com/surrealyz/pdfclassifier
- We have released our source code and models.