Preprocessors Matter!

Realistic Decision-Based Attacks on Machine Learning Systems

Chawin Sitawarin[1] Florian Tramèr[2] Nicholas Carlini[3]

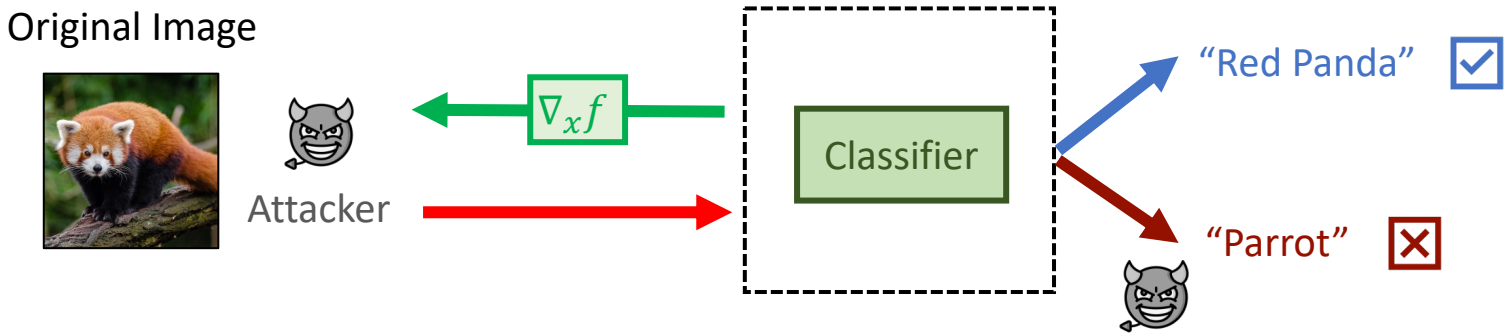
[1]UC Berkeley [2]ETH Zürich [3]Google

ICML International Conference On Machine Learning

Berkeley UNIVERSITY OF CALIFORNIA

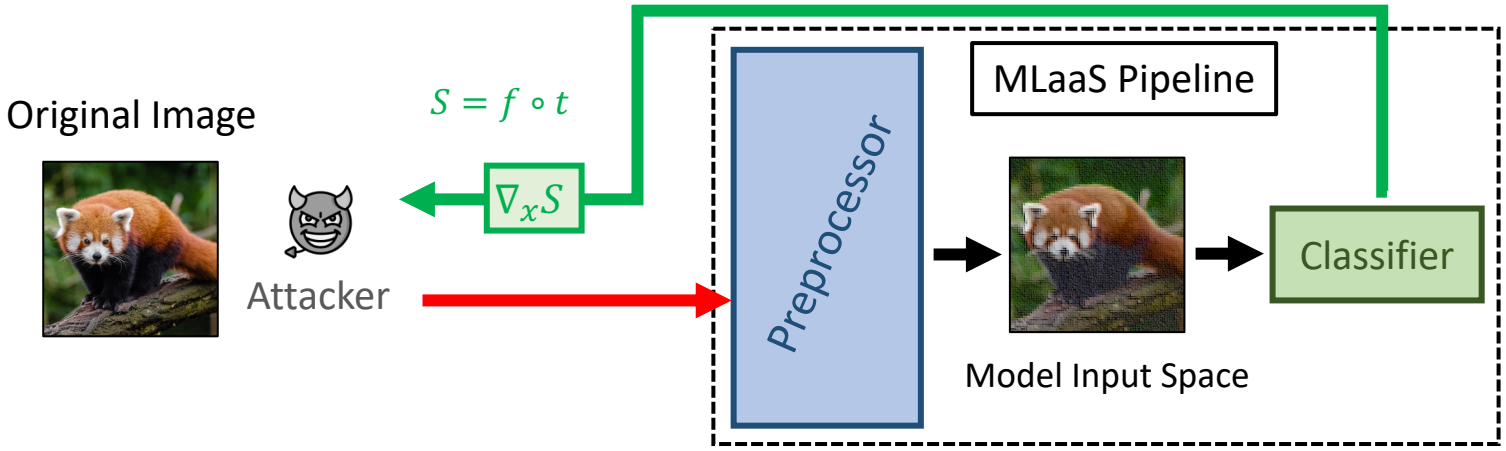
ETH zürich

Black-Box Attack on ML "Models"



- Almost all prior works only consider ML model as a standalone target.
- This assumption is already not true in practice… Preprocessors!
- Instead, we should evaluate robustness of the entire **system/pipeline**.

Black-Box Attack on ML Systems



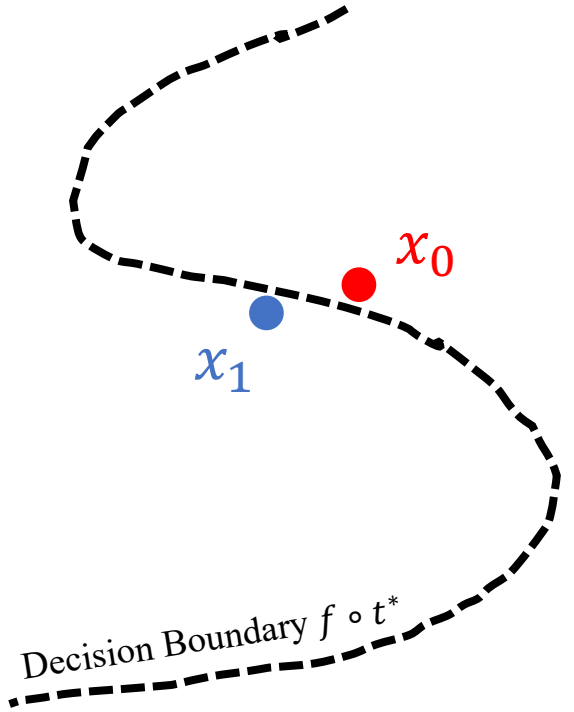
- Preprocessors: crop, resize, quantize (PNG), JPEG, etc.
- If the attacker knows about the preprocessor, the attack can be much more effective.

Black-Box Attack on ML Systems

- Preprocessors: crop, resize, quantize (PNG), JPEG, etc.
- If the attacker knows about the preprocessor, the attack can be much more effective.
  - Example: image cropping → no need to perturb the border (cropped area): “invariance”.
  - Takes advantage of **lossiness** of preprocessors.
- For general preprocessors, we modify **gradient estimation step** off-the-shelf decision-based attack to exploit this invariance.

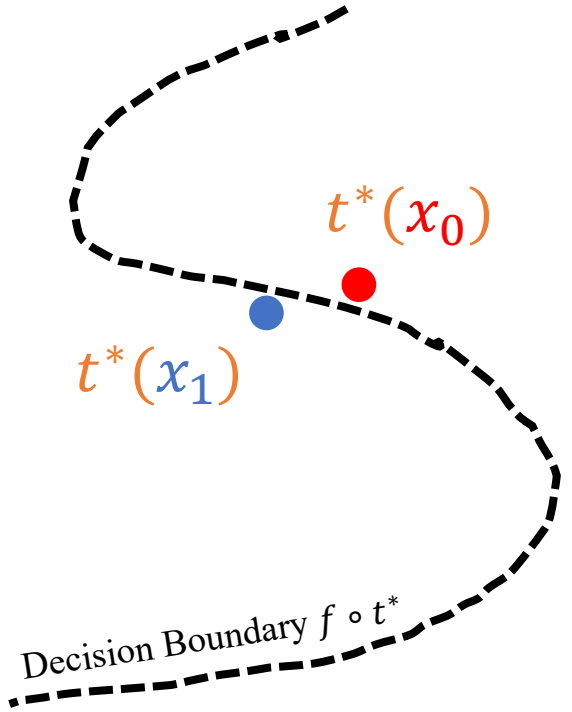
Sketch of Our Extraction Attack

- Guess and check!
- Guess the preprocessor $\tilde{t}$ (vs. real t^*) and apply to some carefully chosen inputs.



Sketch of Our Extraction Attack

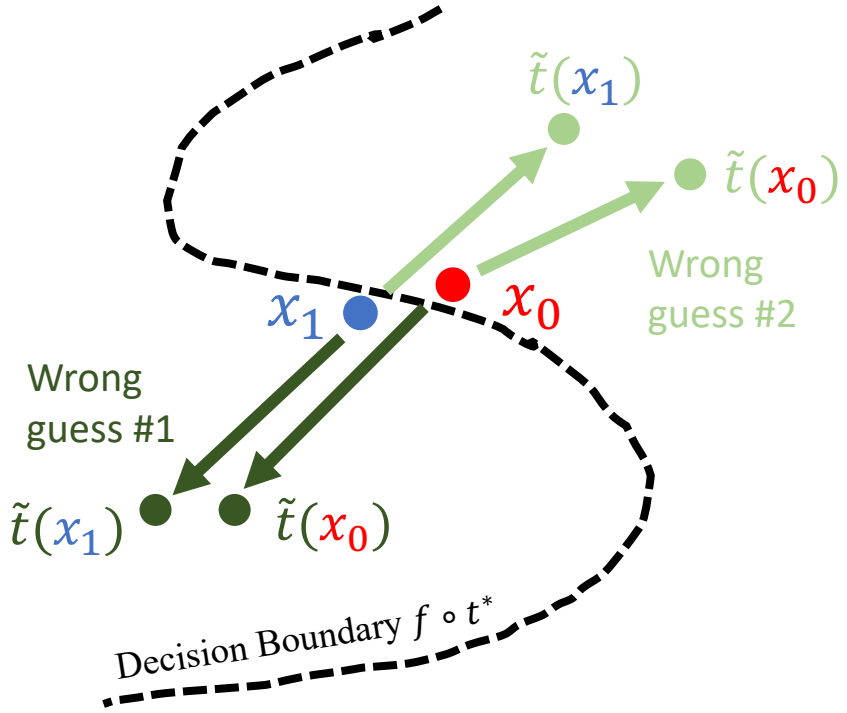
- Guess and check!
- Guess the preprocessor $\tilde{t}$ (vs. real t^*) and apply to some carefully chosen inputs.
- If our guess is right, prediction stays the same.



Sketch of Our Extraction Attack

- Guess and check!
- Guess the preprocessor $\tilde{t}$ (vs. real t^*) and apply to some carefully chosen inputs.
- If our guess is right, prediction stays the same. Otherwise, it will likely change.
- Repeat with multiple pairs until we're confident.

- Extraction attack has to be run only once!



Results

Preprocessor-Aware Attack Results

Preprocessor	Attack Method	Adv. Distance (↓)	
Crop (256 → 224)	Unaware	4.2	
	Ours	3.7	1.1x
Resize (1024 → 224)	Unaware	16.5	
	Ours	3.7	4.5x
Quantize (4 bits)	Unaware	9.7	
	Ours	3.1	3.1x
JPEG (quality 60)	Unaware	9.2	
	Ours	1.5	6.1x
Neural Compress	Unaware	33.8	
	Ours	12.6	2.7x

Extraction Attack Results

10 random ImageNet models on Hugging Face.

Table 4: Number of queries (mean ± standard deviation) necessary to determine what preprocessor is being used.

Preprocessor Space	Num. Queries
Arbitrary resize (200px–800px)	632 ± 543
Arbitrary center crop (0%-100%)	52.0 ± 1.3
Arbitrary JPEG compression (quality 50-100)	70.0 ± 22.8
Typical resize (see text)	48.7 ± 6.8