DRSM: De-Randomized Smoothing on Malware Classifier Providing Certified Robustness

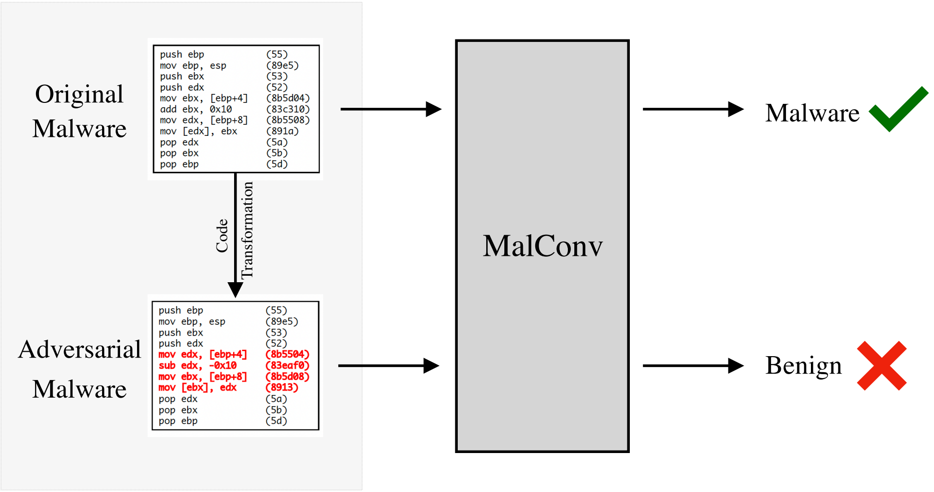
Shoumik Saha, Wenxiao Wang, Yigitcan Kaya,

Soheil Feizi, Tudor Dumitras

UNIVERSITY OF MARYLAND

Problem

- ML has been heavily used in static malware detection
- Malware authors have been generating adversarial malware to bypass such ML models



Prior Defenses

- Non-negative or Monotonic Classifier
  - Constraining weight in the layer
  - Fleshamn et. al. (2018) constrained the last layer weight of MalConv
  - Suffers from low standard accuracy of 88.36%
  - Still vulnerable against some attacks
- Adversarial Training
  - Trained with adversarial malware
  - Suffers from poor robustness against attacks not used during training
  - Also suffers from low standard accuracy

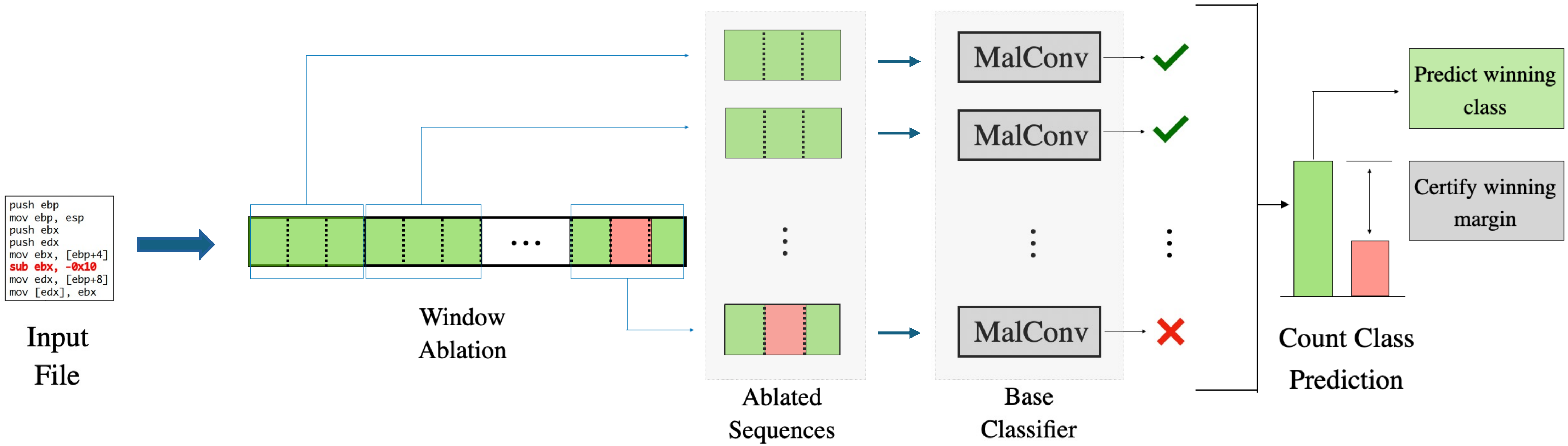
Standard Accuracy
vs
Robustness!!

How can we solve it?

De-Randomized Smoothing

- First proposed in Computer Vision introducing ‘certified’ robustness
- Has been heavily used with different ablation techniques, e.g., masking pixels, block ablation, etc.
- Can it be used in malware detection?
- Challenges:
  - Executable byte files are different than image
  - Our input is one-dimensional unlike images
  - Random byte changes in our file can alter file behavior/structure
- Solution:
  - Re-designing the De-Randomized Smoothing technique
  - ‘Window Ablation’ Scheme

Window Ablation Scheme

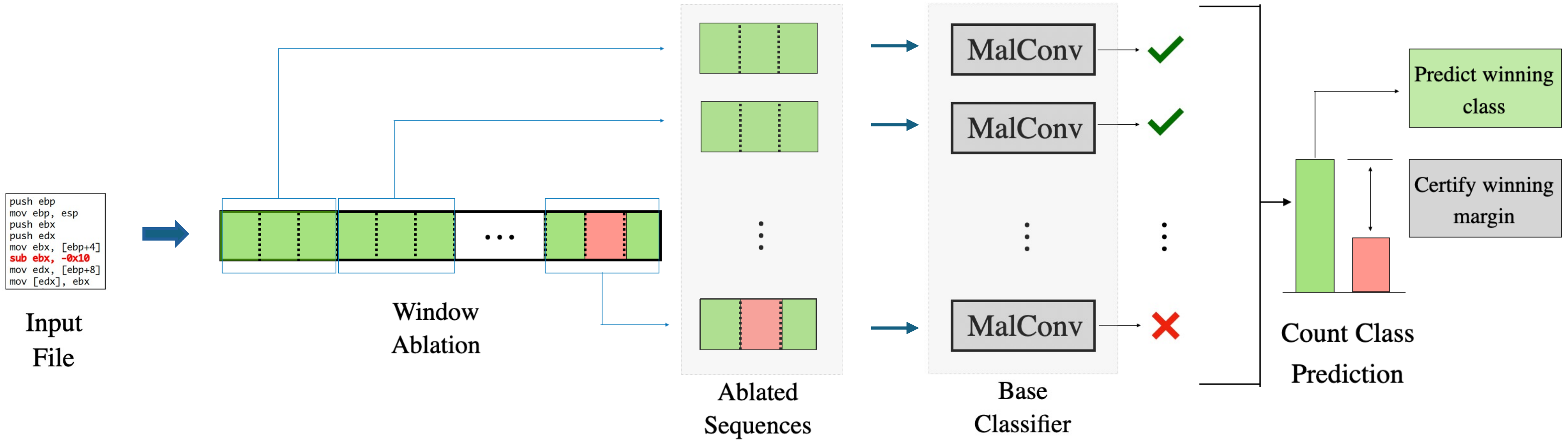


$$G_\theta(x) = argmax_c \;\; n_c(x);$$

where

$$n_c(x) = \sum_{x' \in S(x)} I\{F_\theta(x') = c\}$$

Window Ablation Scheme



- Base classifier input length = L
- Ablated window size = w
- Number of ablated sequences = $\lceil L/w \rceil$
- Perturbation size = p
- Highest modifiable ablated sequences, $\Delta = \lceil p/w \rceil + 1$

Certified robustness condition:

$$n_c(x) > max_{c \neq c'} \ n_{c'}(x) + 2\Delta$$

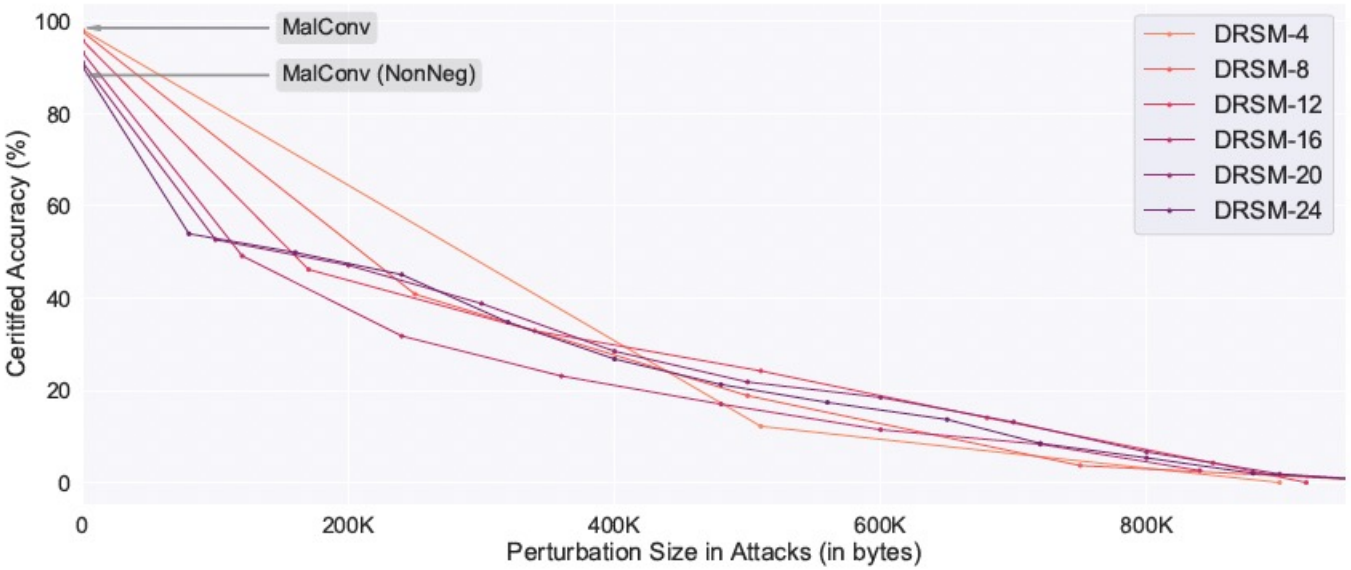
Experiments

- Evaluated six variants of DRSM-n
- Our DRSM models provide a spectrum for standard vs certified accuracy

Model	Standard Accuracy (in %) ↑			Certified Accuracy [$\Delta = 2$] (in %) ↑		
	Train-set	Validation-set	Test-set	Train-set	Validation-set	Test-set
MalConv	99.73	98.87	98.61	–	–	–
MalConv(NonNeg)	88.56	87.56	88.36	–	–	–
DRSM-4	99.49	98.12	98.18	14.74	7.84	12.2
DRSM-8	99.67	97.88	97.79	52.74	43.9	40.85
DRSM-12	96.07	95.58	95.88	45.77	44.43	46.21
DRSM-16	94.29	93.00	93.3	59.1	50.52	49.17
DRSM-20	91.17	91.05	91.15	51.64	51.92	52.68
DRSM-24	90.22	89.80	90.24	54.19	54.88	53.97

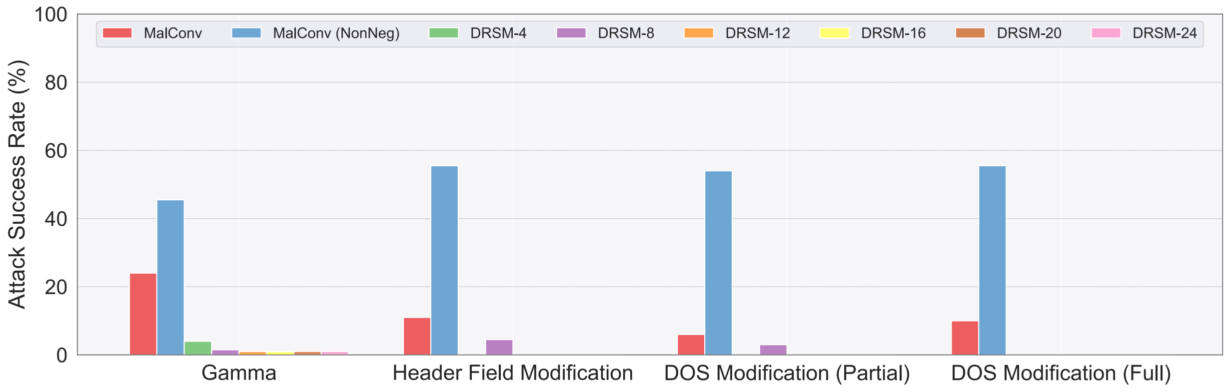
Experiments

- Variation in perturbation budget impacts certified accuracy

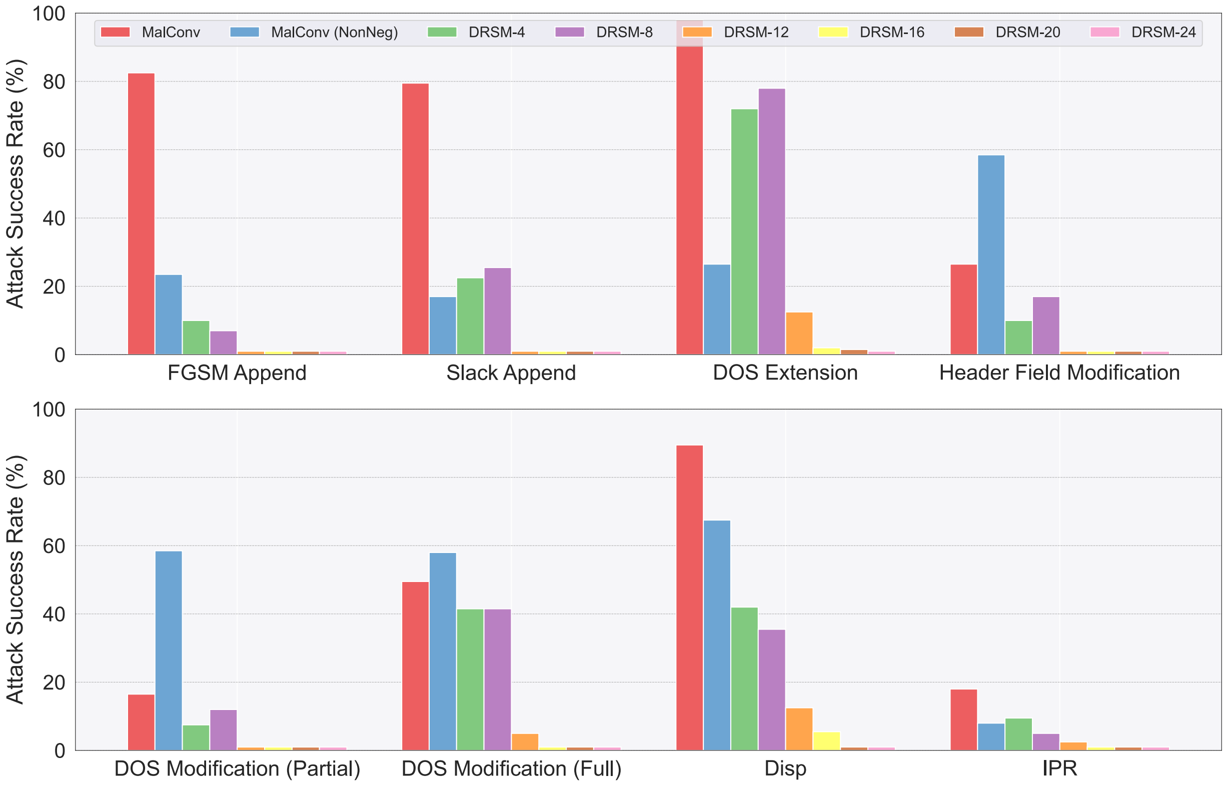


Empirical Robustness

- Evaluated our DRSM against nine different attacks!
    - Covering both white and black box settings
    - Considering different threat models



Black-box Attacks



White-box Attacks

PACE Dataset

- Publicly Accessible Collections of Executables (PACE)
- For this research, we compiled a diverse benign dataset of 15.5K size crawling different websites
- To help the community, we are making this benign dataset publicly available

Thank you!