EXPLORING ADVERSARIAL EXAMPLES IN MALWARE DETECTION
OctavianSuciu, Scott E. CoulI, Jeffrey Johns

PROBLEM
Adversarial Examples in Image Classification:
"panda" + perturbation = "gibbon"

Adversarial Examples in Malware Detection:
`\texttt{\textbackslash x48\textbackslash x04\textbackslash xF1\textbackslash x24}` + perturbation = `\texttt{\textbackslash x8A\textbackslash x32\textbackslash xB7\textbackslash x2C}`

Are adversarial malware examples realistic?
Are attacks effective against production-scale training sets?

ATTACKS
Benign Append
`\texttt{\textbackslash x48\textbackslash x00\textbackslash xF1\textbackslash x24\textbackslash x00\textbackslash x3D}` + `\texttt{\textbackslash xE3\textbackslash x41}`
Malware sample
append section of a benign program

FGM Append
`\texttt{\textbackslash x48\textbackslash x00\textbackslash xF1\textbackslash x24\textbackslash x00\textbackslash x3D}` + `\texttt{\textbackslash xE3\textbackslash x41}`
Malware sample
append gradient-based noise via single-step FGM

Slack FGM
`\texttt{\textbackslash x47\textbackslash x00\textbackslash xF1\textbackslash x24\textbackslash x00\textbackslash x3D}` + `\texttt{\textbackslash x47\textbackslash xF1\textbackslash x24\textbackslash x3D}`
Malware sample
add single-step FGM noise via safe slack region

EXPERIMENTAL SETUP
Victim Model: MalConv

Sample → 2MB Embedding → Gated Convolution → Temporal Max-Pooling → Fully Connected → Label

Architecture: pooling 128 non-overlapping convolutional kernels
- ≤ 128 unfragmented input sequences used in classification

Training Sets:
- Mini: in line with prior work\textsuperscript{1}, 8,500 samples
- EMBER: publicly available corpus of 1M samples\textsuperscript{2}
- Full: production scale dataset of 12.5M samples

\textsuperscript{1} Exploring and harnessing adversarial examples (GoodFellow et al. 2014)
\textsuperscript{2} Malware detection by using a viable exsath\textsuperscript{3} (2017)
\textsuperscript{3} Adversarial Malware Threats: Existing Deep Learning for Malware Detection in (Speculative) (Keser et al. 2018)
\textsuperscript{4} EMBER: An Image Dataset for Training Static PE Malware Machine Learning Models (Jefferson 2018)

FINDINGS
Model Robustness Influences Results

`\texttt{\textbackslash x48\textbackslash x04\textbackslash xF1\textbackslash x24}`
`\texttt{\textbackslash x8A\textbackslash x32\textbackslash xB7\textbackslash x2C}`
corrupted file

Counterintuitive results in overfitted model

Slack FGM results

Unfragmented input flows to last layer
- effect of slack bytes is amplified by context
- trade-off between Success Rate and Leverage
- due to Slack size and gradient magnitude

Single-Step Samples are Not Transferable
Transfer samples between EMBER = Full
- using FGM Append & Slack FGM
- only 3/400 attack samples are successfully transferred
- small gradient magnitude in EMBER

Thursday 10:45AM
@ DLS Workshop