Impacto de ataques de evasão e eficácia da defesa baseada em treinamento adversário em detectores de malware

Gabriel H. N. E. da Silva, Gilberto Fernandes Junior, Bruno Bogaz Zarpelão

Departamento de Computação - Universidade Estadual de Londrina

Motivação

- Detecção de malware utilizando aprendizado de máquina supervisionado já é uma realidade.
- Algoritmos de aprendizado de máquina podem ser vulneráveis a exemplos adversários.

Objetivo

1. Qual a severidade dos ataques FGSM, BGA, BCA e Grosse, considerando que eles seguem estretégias distintas?
2. Quão eficaz é o treinamento adversário?
3. A topologia da rede neural influencia a robustez do modelo alvo e a efetividade do modelo adversário?
4. Qual é o desempenho da Random Forest, como alvo, quando comparada a uma rede neural?

Conjunto de dados

- Público e disponível na Web: fornecido pelo projeto Drebin.
- Só inclui aplicações Android.
- Extração de atributos: análise estática, considerando manifesto e códigos dex.
- Cerca de 25500 atributos binários.

Modelos de aprendizado

- Treinamento com amostras benignas e maliciosas balanceadas: 2000 amostras de cada classe.
- Inferência respeita a proporção de 24:1: 2400 benignas e 100 maliciosas.
- Dois algoritmos: rede neural MLP e Random Forest.

Amostras adversárias

- FGSM: gera uma perturbação a ser adicionada à amostra com base no gradiente do modelo treinado.
    - rFGSM e dFGSM: variações do FGSM original visando modificação de atributos binários.
- BGA: utiliza o gradiente para identificar os atributos que mais afetam o erro da saída.
- BCA: testa modificações em diferentes atributos para identificar qual mais afeta a saída.
- Grosse: modifica atributos que mais afetam a saída com base em uma matriz Jacobiana.

Treinamento adversário

- Adição de amostras maliciosas ao treinamento do modelo alvo.
- Geramos amostras adversárias para 200 amostras de malware: total de 1000 amostras, considerando os 5 ataques.
- Treinamos novamente os modelos alvo adicionando as 1000 amostras adversárias e 1000 amostras benignas (total final de 6000 amostras).

Resultados

- Foram selecionados 5 modelos de redes neurais com os melhores f1-score.
    - Nesta parte do experimento, a topologia das redes foi variada, aumentado a quantidade de camadas e a quantidade de neurônios por camada.

Resultados - ataques contra a MLP (sem defesa)

Ataques contra modelos sem defesa					
	M_1	M_2	M_3	M_4	M_5
M_1-M_5 **x dFGSM**	0,8498 ± 0,0021	1,0000 ± 0,0000	0,9463 ± 0,0008	1,0000 ± 0,0000	1,0000 ± 0,0000
M_1-M_5 **x rFGSM**	0,8498 ± 0,0021	1,0000 ± 0,0000	0,9463 ± 0,0008	1,0000 ± 0,0000	1,0000 ± 0,0000
M_1-M_5 **x BGA**	1,0000 ± 0.0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000
M_1-M_5 **x BCA**	0,6028 ± 0,0236	0,6480 ± 0,0133	0,9439 ± 0,0108	0,5879 ± 0,0236	0,7146 ± 0,0052
M_1-M_5 **x Grosse**	0,9743 ± 0,0193	0,9699 ± 0,0191	0,9699 ± 0,0049	0,9786 ± 0,0168	0,9592 ± 0,0275

Tabela 2. Severidade média dos ataques contra os 5 modelos sem defesa. As linhas representam a média e o desvio padrão da severidade do ataque contra os 5 modelos e as colunas mostram qual a topologia utilizada pelo atacante.

Resultados - ataques contra a MLP (com treinamento adversário)

Ataques contra modelos com defesa					
	M_1	M_2	M_3	M_4	M_5
M_1-M_5 **x dFGSM**	0,8453 ± 0,0043	1,0000 ± 0,0000	0,9439 ± 0,0014	1,0000 ± 0,0000	0,8000 ± 0,4472
M_1-M_5 **x rFGSM**	0,8453 ± 0,0043	1,0000 ± 0,0000	0,9439 ± 0,0014	1,0000 ± 0,0000	0,8000 ± 0,4472
M_1-M_5 **x BGA**	0,9682 ± 0,0711	1,0000 ± 0,0000	0,9886 ± 0,0254	1,0000 ± 0,0000	1,0000 ± 0.0000
M_1-M_5 **x BCA**	0,3157 ± 0,1360	0,5416 ± 0,1289	0,2870 ± 0,0855	0,4251 ± 0,1258	0,2610 ± 0,2675
M_1-M_5 **x Grosse**	0,0222 ± 0,0281	0,2181 ± 0,1496	0,0666 ± 0,0662	0,1304 ± 0,0889	0,0273 ± 0,0499

Tabela 3. Severidade média dos ataques contra os 5 modelos com defesas. As linhas representam a média e o desvio padrão da severidade do ataque contra os 5 modelos e as colunas mostram qual a topologia utilizada pelo atacante.

Resultados - ataques contra a Random Forest (sem defesa)

Ataques contra Random Forest sem defesa					
	M1	M2	M3	M4	M5
RF x dFGSM	0,848	1,000	0,946	1,000	0,870
RF x rFGSM	0,761	0,891	0,870	0,902	0,511
RF x BGA	0,989	1,000	0,978	1,000	0,957
RF x BCA	0,000	0,000	0,000	0,000	0,000
RF x Grosse	0,000	0,000	0,000	0,000	0,000

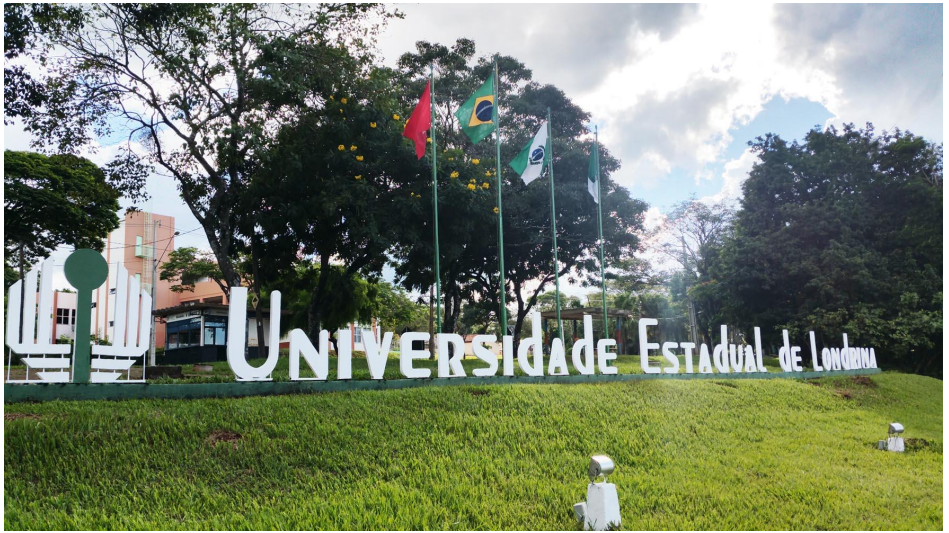
Tabela 4. Severidade dos ataques contra a Random Forest. As linhas representam a severidade do ataque contra a Random Forest e as colunas mostram qual a topologia utilizada pelo atacante.

Considerações finais

- Ataques menos intensos como BCA e Grosse causam, como esperado, menor impacto.
- Ataques mais intensos, como FGSM e BGA, causam mais impacto.
- O treinamento adversário não foi eficaz.
- Modelo baseado em Random Forest sofreu menos impacto.
- Mudanças nas topologias das redes neurais não tiveram influência nos resultados.

Obrigado!

- Bruno B. Zarpelão
- brunozarpelao@uel.br



Patrocinadores do SBSeg 2024!

nic.br

cgi.br

Google

Tempest

CAPES

SiDi

FAPESP

zscaler™

BugHunt

CNPq

C.E.S.A.R

FACULDADE IBPTECH