CS 4035 LAB4 - ADV ML CHALLENGE GROUP 29

Yuhang Tian
5219728

Mingyu Gao
5216281

June 27, 2021

1 Innermaximizer Task

We use the following method - Greedy Random Accelerated Multi-bit Search (GRAMS) referenced from paper The Robust Malware Detection Challenge and Greedy Random Accelerated Multi-Bit Search to implement the malicious evasion generator.

Algorithm 1: GRAMS - topk variant

Data: a batch b and a neural network model M
best_x := b, orig_x := Data_VALUES(b);
k := 8;
while $k > \frac{1}{2}$ **do**
    loss := loss(M, x);
    grad := AUTOGRAD(loss, x);
    sign := SIGNS(grad);
    grad := ABSOLUTE(grad - orig_x * grad);
    x' := x + TOPK(grad, k) * sign;
    loss := LOSS(M, x');
    loss' := LOSS(M,best_x);
    if *a row r in x with loss[r] > loss'[r] exists* **then**
        for *all such rows r* **do**
            best_x[r] = x'[r];
        x := x';
        k := 2k;
    else
        k := $\frac{1}{2}$k;
return best_x;

2 Experimental Result

This part shows the experimental result of designed method - *grams* against baseline methods - *rfgsm_k*, *bga_k* and *dfgsm_k*.

Table 1 reflects the evasion rate of four methods against each other. The evasion rate reflects the success rate of the attacker and the failure rate of the defender. The higher the evasion rate, the more successful the attacker is. So a good method should have a high evasion rate as an attacker and have a low evasion rate when it serves as a defender. The table also shows the f1score of each defense against attack. The f1score contains precision and recall, which indicates the precision and robustness of a model of distinguishing the evasive malicious attacks.

The evasion rate can be used to evaluate the performance of an attacker while the F1-score can reveal the performance of a defender. For *rfgsm_k*, the best invasive attacker is *dfgsm_k* while it can best defend *grams*. In terms of *bga_k*, the most suitable attacker to attack it is *dfgsm_k* while it can best defend *grams*. Turning to *dfgsm_k*, it is vulnerable to *rfgsm_k* and grams but it can defend *dfgsm_k* well. *grams* has the advantage of shortest run-time.

Table 1: Evasion Rate Comparison

		Baseline Attack			Other Attack
	Model	rfgsm_k	bga_k	dfgsm_k	grams
Evasion Rate					
Baseline Defend	rfgsm_k	0.1826	0.1852	0.1968	0.1688
	bga_k	0.2490	0.1908	0.2632	0.0934
	dfgsm_k	0.1559	0.1713	0.1076	0.1546
F1-score					
Baseline Defend	rfgsm_k	0.8937	0.8921	0.8889	0.9065
	bga_k	0.8379	0.9049	0.8589	0.9517
	dfgsm_k	0.9096	0.9065	0.9326	0.9219
Run-time					
Baseline Defend	rfgsm_k	2880.277	3070.265	3007.229	2512.941
	bga_k	3232.249	3438.289	3418.418	2314.038
	dfgsm_k	3473.141	3654.696	2469.301	2462.150