Adversarial example detection Bayesian game: supplementary material

Hui Zeng, Biwei Chen, Kang Deng, and Anjie Peng

The supplementary document consists of: A) How to play the adversarial example detection game? B) Why does Nash Equilibrium make sense?

A. How to play the adversarial example detection game?

1) **In Step 1**, **investigator Alice warms up on the two-step test.** Let's say she selects the AddDe-based detector to perform $\delta^1()$, and the SRM-based detector to perform $\delta^2()$. She runs $\delta^1()$ and $\delta^2()$ on a number of benign images to obtain the relationship between P_{fa}^1 and σ (a parameter of the AddDe-based detector). Similarly, She runs $\delta^2()$ on benign images to obtain the relationship between P_{fa}^2 and t (a parameter of the SRM-based detector).

2) **In Step 2, Bob mounts attacks**. He chooses a certain attack algorithm, e.g., IFGSM, to generate a number of adversarial examples (AE) for each strength, $\epsilon = \{1, 2, 4, 6, 8\}$. Only the images that succeeded in every strength are selected for further use. Let's say $5 \times N$ AEs are selected.

3) **In Step 3, Alice runs the two-step test on the AEs.** For a given P_{fa}, she first sets P_{fa}^1=0.01, then calculates P_{fa}^2=$P_{fa} - P_{fa}^1$. This equation is approximately valid because the chance of a benign image introducing false alarms in both tests is negligible. With the relationships obtained in Step 1, she can now run the two tests on $5 \times N$ AEs. Let's say for the N AEs of $\epsilon = 1$, N_A images are detected as adversarial by either $\delta^1()$ or $\delta^2()$, then the (1, 1) entry of the payoff matrix is N_A/N. Similarly, the entries (2, 1), (3, 1), …, (5, 1) of the payoff matrix can be obtained. Varying P_{fa}^1 from 0.01 to P_{fa}, the remaining entries of the payoff matrix can be obtained. Varying P_{fa}, the payoff matrices under different P_{fa}s can be obtained.

4) **Step 4**. Repeat steps 2 and 3, payoff matrices for other attacks can be calculated.

5) **Step 5**, **solving Nash Equilibrium (NE)**. For each payoff matrix, a Nash equilibrium can be obtained by solving a linear optimization problem (Eq. (7) of the paper).

6) **Step 6, drawing NEROC**. Now, for every given P_{fa}, we have obtained a point indicating the payoff value under NE. Connecting these points, a NEROC curve can be drawn. By examining the NEROC curves for different attacks, their security can be evaluated.

B. Why does Nash Equilibrium make sense?

We designed a comparative experiment to answer this question in which one player deviates from her Nash Equilibrium strategy. We assume the investigator is a conservative player who wants to maximize the worst detection rate. The attacker is a rational player who sticks to his NE strategy. Note that the investigator is a conservative player is not common knowledge, i.e., the attacker cannot take advantage of this.

We want to know how the investigator's conservative strategy affects the game. In the complete information game, the investigator's conservative strategy $P_{fa}^{1,cons}$ can be calculated as:

$$P_{fa}^{1,cons}, \_ = arg \max_{P_{fa}^1} \min_{r} U(P_{fa}^1, r)$$

The attacker still follows the NE strategy r^* (obtained solving a dual problem of (7) of the paper). Thus the final detection rate is $U(P_{fa}^{1,cons}, r^*)$. Varying the allowed total false alarm rate P_{fa}, an ROC curve can be drawn for each attack. Fig. 1 compares the ROC curves of a conservative attacker with the NEROC curves for four attacks. ROC curves corresponding to the conservative attacker are always lower than or equal to the NEROC curves. Such comparison indicates that the conservative strategy is inferior to the NEROC strategy. Such results also comply with the basic properties of Nash equilibrium: no player has the motivation to change his strategy unilaterally.

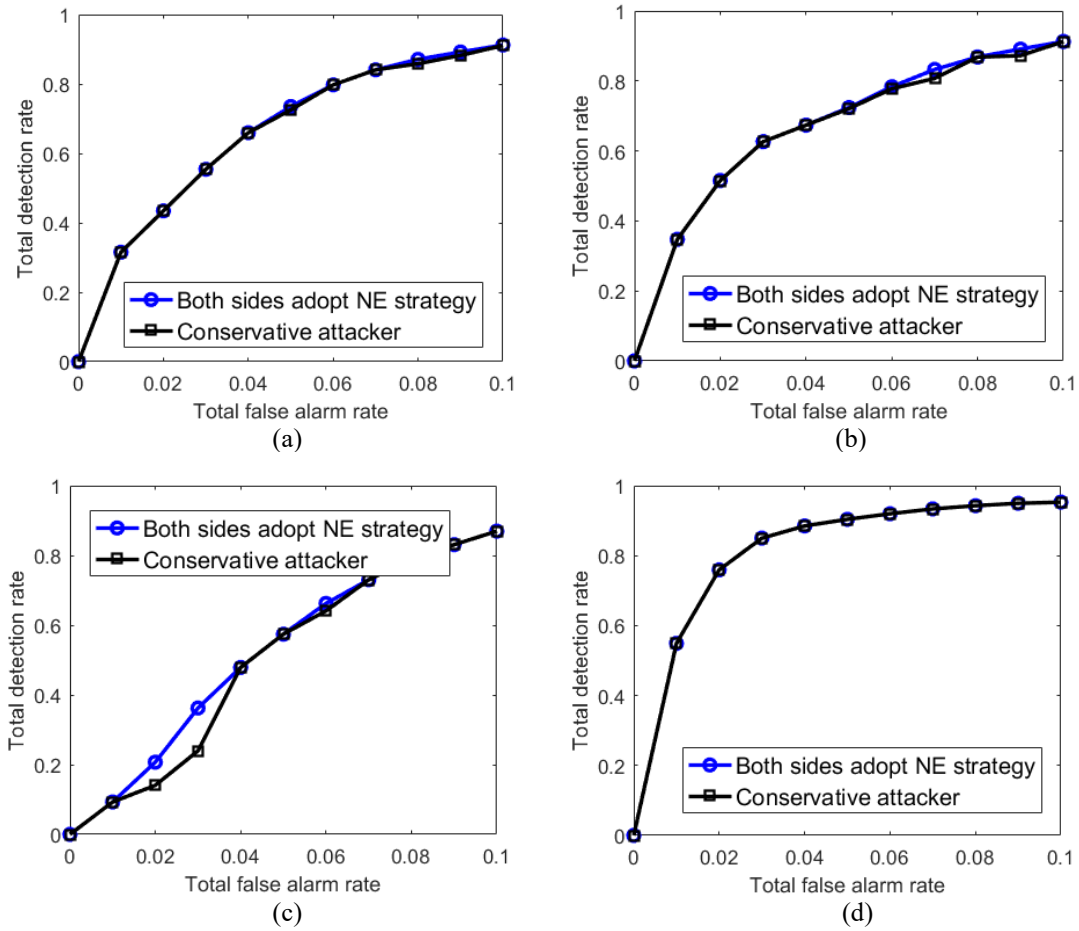

Fig. 1. NE strategy vs. conservative strategy. (a) IFGSM, (b) MI, (c) C&W, (d) ST.