CAAD CTF Rules for GeekPwn 2018 at Shanghai

Sept 27, 2018

Version 1.1

The organizer will invite 5 ~ 6 teams to participate CAAD CTF on GeekPwn 2018 at Shanghai. We will have it on Oct. 24th, 2018. The rules details are below:

1. Each team will submit their defense classifier to organizer. Organizer will test the classifier with a secret image dataset to check the accuracy. Only the defenses with accuracy rate above 75% are valid. The deadline of change is Oct. 18th. Before the deadline, the defense classifier is allowed to change. After that, it cannot be modified before/in the competition. In the competition, some testing images will randomly be sent by the organizer to the defense classifiers, to make sure the defenses are working in good conditions. If the behaviors are abnormal for testing images, judges will decide if there are cheating behaviors. If yes, the team will lose chances to win.

2. The whole CTF competition will be divided into 2 phases.

3. In phase 1, organizer will announce several tasks. These tasks require participants to fulfill to get scores. The tasks are related to adversarial attacks and defenses. The tasks details will be kept secret before the competition.

4. In phase 2, organizer will prepare some different images. The images will be kept secret before the competition. Phase 1 will be divided into rounds. In each round, each team will get a different image as source image. For example, team 1 gets an image of ImageNet class 169, redbone (a type of dog). Team 2 gets an image of ImageNet class 280, Arctic fox. Then, each team will use the source image to create adversarial examples. In this example, team 1 will modify the redbone (a dog) image, send the modified image to team 2 with the purpose that team 2 will classify the image as class 280, Arctic fox. If indeed the defense classifier of team 2 recognizes the image as an Arctic fox, the attack is successful. Team 1 gets score and team 2 loses score. If it is recognized as a redbone, the attack fails. No team gets or loses score. If it is recognized neither as a redbone nor as an Arctic fox, then team 2 loses 5 points for the first time against team 1. (Because there are more than 1 attack teams, this situation may happen more than once in a round). Each round lasts for 10~20 minutes (will announce before the competition begins). When participant creates adversarial examples, different perturbation value can be used. They are 32, 16, 8, 4, 2 and 1. Different value means different score. The smaller the perturbation value is, the bigger the score. If team 1 attacks team 2 successfully with perturbation value 32 first, after that, team 1 tries perturbation 8 and succeed again, team 1 will get score for perturbation 8 minus the score for perturbation 32. Each time an attacker team gets score, same score will be deducted from the defender team.

Perturbation Value	32	16	8	4	2	1
Score	10	20	40	80	160	320

5. The total score of phase 1 and phase 2 will determine the final scores. Teams will have prizes based on the scores.

6. Before the competition, participant team will provide a Docker image to organizer. This Docker image will be used for attack in competition. If the Docker image is available publicly, just tell the organizer the exact name to download, otherwise, please export customized container and send to organizer.

7. The running container in competition will be assigned 4 CPU cores, 16GB RAM, and a single NVIDIA Tesla P100 board. It is not allowed to use other compute resources during the competition.

Detailed requirements of defenses

1. Before the competition, participant team will submit a defense to organizer (The requirement is same to CAAD 2018 online competition, please check caad.geekpwn.org for more details). The defense classifiers will run in a Docker container

with 4 CPU cores, 16GB RAM and a single NVIDIA 1080Ti board. It is required the defense can classify one image in 4 seconds. If it failed to classify one image in 4 seconds, the defense team will lose 5 points every time it happens.

2. The defense classifiers are required to have accuracy rate above 75% against a secret test dataset.

3. Organizer will modify the defense to monitor an input folder, when there is a new image exists, it will classify the image and put result in the output folder. The input and output folders are mapped folders on host. The exact folder paths will be passed to the defenses by parameters.

Detailed requirements of attack

1. In phase 2, participant will send attack packet to a controller machine. Each attack contains information: target team, adversarial example image, perturbation value. Controller will forward the attack to corresponding defense, then respond attacker with information: attack result (success or fail), class label the defense classified.

2. In phase 2, each team can perform attack on other teams at the same time. Attacking their own defense is not allowed. For a same target team, it is not allowed to attack too frequently. Controller will control the frequency. 6 seconds are required interval time between two attacks to same target. When an attack image has been sent to a defense, it is not allowed to send a new image to the same defense before result returns. The attacker should wait.

3. It is not allowed to attack the competition servers or use other method than adversarial examples to attack defenses.

Participant teams need to use their own laptops connect to attack Docker container through SSH. Then participant teams issue attacks from there. All attacks go through a Controller. The Controller will forward attacks to corresponding defense. Note that the Controller will only accept attacks from the attack Docker container.

Prize

First Place, RMB 30,000

Second Place, RMB 20,000

Third Place, RMB 10,000

Other teams, RMB 5,000

When competition finishes, we will have the scores. However, judges will review the whole competition records to make sure everything is OK. The final result will be announced after the results are determined.

If there are ties for final scores, judges will determine the order based on correct rate of defenses evaluated against a set of secret baseline non-targeted attacks and image dataset before the competition.

Team requirement

Each team can have up to 5 members, but only 2 members can participate the competition at Shanghai.

Organizer will cover traffic and hotel fees for 2 team members.

Registration

We are now inviting teams to participate this CAAD CTF. If you are interested in it, please send email to us at caad@geekpwn.org before Sept 30th, please also introduce what you did in the area of Adversarial Examples in the email. Judges will decide the invitation list.

To know more about GeekPwn, please visit http://geekpwn.org