CSAW 2022
EMBEDDED SECURITY CHALLENGE
(DEFENSE TRACK)

papern0ught, IIT Roorkee

CHALLENGE OVERVIEW

- Out of the 7 challenges we attempted, we solved 4
  - 7r0j4n_1
  - dumps7er_d1ve
  - poison_mushroom
  - leaky_b0ttle
- We present our understanding of the problems as:
  - Problem statement
  - Model weakness
  - Our approach

7R0J4N_1

- *Problem Statement*: Zeroing out minimum number of neural network weights to cause maximum drop in model performance and hence its defence
- *Model Weakness*: Over dependence on particular weights
- *Our Approaches*:
  - Additional Dropout layers
  - Increased Dropout probability
  - L2 norm on layer weights + Gradient Clipping by norm: Best Results

D U M P S 7 E R _ D 1 V E

- *Problem Statement*: To prove if a model has been extracted/stolen into another model by inserting triggers
- *Model Weakness*: Over-parameterization of neural networks
- *Our Approach*:
  - To generate a unique trigger/watermark for the original model
  - Create adversarial samples and overfit these to the model
  - Adversarial samples are from near the decision boundary

POISON_MUSHROOM

- *Problem Statement*: To detect poisonous samples which affect model performance in feature classification
- *Model Weakness*: Cannot differentiate between out-of-distribution samples
- *Our Approach*:
  - First attack the model by generating samples of a class from the distribution of the other class
  - Detect such samples by calculating the distance of a sample from its true distribution

LEAKY_BOTTLE

- *Problem Statement*: To prevent the inference of membership status of a test datapoint
- *Model Weakness*: Over-fitting of the model around trained examples
- *Our Approach*:
  - Increase the temperature of the last softmax layer
  - Additional dropout layers
  - L2-norm regularisation on layer weights

LIVE CHALLENGE

- *Problem Statement*: Black-box adversarial example generation
- *Model Weakness*: Sensitivity of model to high-frequency noise present in the training data
- *Our Approach*:
  - Popular adversarial attacks use weight and gradient information, but we have black-box access
  - Using SimBA, we try to perturb the image in random directions and claim that model's confidence must reduce in one of the direction
  - Use logits as a proxy to gradient change

ACKNOWLEDGEMENTS

- To the organizers: For amazing problem statements covering various domains in AISecurity
- To our advisor: For providing the necessary computational resources

THANK YOU