Active Learning Model Extraction Detection with LSTM

Alex Sidgwick

Introduction

- Model extraction
  - Attackers reverse engineer publicly accessible models
  - Theft of intellectual property
- Types of model extraction
  - Passive learning uses a large set of queries, and is often impractical
  - Active learning algorithmically selects queries for efficiency
- Defending against model extraction
  - Randomized outputs harm model accuracy and are not always effective
  - Detection algorithms are essential to provide security while maintaining performance

Detecting Model Extraction

- Existing approaches
    - PRADA and VarDetect examine the distribution of a user's queries [1][2]
        - These approaches are generic and effective
    - SEAT examines the similarity between queries [3]
        - Model extraction attacks may generate similar queries or perturb existing ones
    - HODA examines the “hardness” of a user’s queries [4]
        - This is defined as the number of epochs taken for the target model to learn the correct classification of an input
- HODA showed that extraction techniques are interested in “harder” data points
    - As the extracted model becomes more accurate, its definition of a “hard” data point changes

Algorithm

- Implied sequential pattern
  - HODA's observations on "hardness"
  - Active learning involves using the results of previous queries to select subsequent queries
- LSTM-RNN (Long Short Term Memory Recurrent Neural Network) with MLP for sequence classification
  - Can process longer sequences than traditional RNNs
  - Have been used for security in the past [6]
  - May be able to learn the patterns indicative of active learning
- Output from target model included in input to detection model
  - This information is used by active learners to select subsequent queries
  - It is reasonable to assume that the "hardness" of a query is related to the uncertainty of its classification by the target model, which is reflected in the raw output of the target model

Experiments

- Target model is a CNN for classifying digits from the MNIST dataset
    - Classifies 0s from 1s
    - Used in the past to evaluate model extraction [1][2]
- Benign queries
    - Problem domain (0s and 1s from the dataset)
    - Alternative Problem Domain (Digits other than 0 and 1 from the dataset)
    - These model the behavior of a benign user [2]
- Malicious queries
    - Active learning heuristic for training and evaluation [5]
        - Made without access to training data
    - Synthetically generated data sets for evaluation
        - Used in the past to evaluate model extraction [1][2][3][4]
        - Made with access to training data
- Compare false negative and false positive rate with PRADA and SEAT

Results

- PRADA reports a 0% false negative and false positive rate [1]
- SEAT reports a 0.05% false positive rate [3]
- LSTM-RNN
  - 14% false negative rate
  - 32% false positive rate
  - Some success detecting sequential patterns

Future Directions

- Active learners may interleave benign or misleading queries
  - Randomly spaced sampling
  - Practicality of these attacks
- Experiments on more complex active learning attacks
- Longer or shorter sequences for the LSTM may increase accuracy

Conclusions

- Model extraction techniques are essential for publicly accessible models
- Active learning model extraction implies a sequential pattern
- LSTM-RNN shows some promise in detecting active learning
  - Nowhere near as effective as state of the art approaches such as PRADA and SEAT

Questions?

References

[1] (PRADA) https://arxiv.org/pdf/1805.02628.pdf

[2] (VarDetect) https://arxiv.org/pdf/2107.05166.pdf

[3] (SEAT) https://dl.acm.org/doi/pdf/10.1145/3474369.3486863

[4] (HODA) https://www.researchgate.net/profile/Amir-Mahdi-Sadeghzadeh/publication/353071516_HODA_Hardness-Oriented_Detection_of_Model_Extraction_Attacks/links/621a1b8a579f1c04171b66a9/HODAHardness-Oriented-Detection-of-Model-Extraction-Attacks.pdf

[5] (Adaptive Retraining) https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf

[6] (LSTM in Security) http://text2fa.ir/wp-content/uploads/Text2fa.ir-An-effective-network-attackdetection-method-based-on-kernel-PCA-and-LSTM-1.pdf