Stealing Complex DNN Models: Limitations and Defense Strategies

Aalto University

Buse Gul Atli, Sebastian Szyller, Mika Juuti, Samuel Marchal, N. Asokan

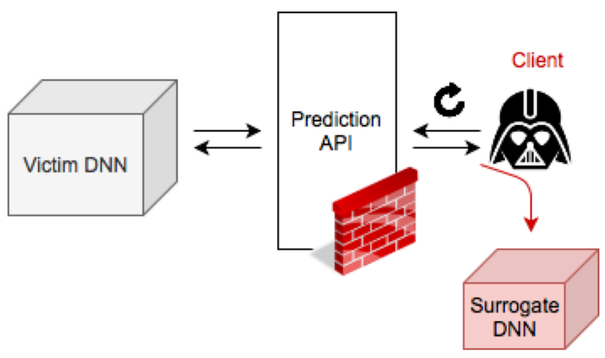
Model Confidentiality Matters

- DNN solutions as products represents *business advantage* and *intellectual property.*
- Model stealing threatens these advantages.
- Current defenses against model stealing attacks are limited.

Stealing Realistic DNN Models [1]

Adversary's capabilities:

- Access to pre-trained models.
- No knowledge of train/test data, output semantics.
- Access to natural samples (ImageNet) and full prediction probability vector (cf. PRADA [2], which assumes access to only a small number).



Model stealing process

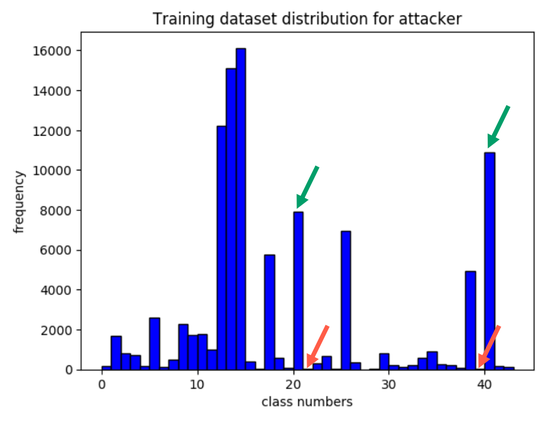
Stealing Process

- Query victim DNN with *natural data* (100,000 ImageNet samples).
- *Fine-tune* pre-trained model(s) with victim outputs.
- Victim: pre-trained complex models / simpler CNN.
- Surrogate DNN can obtain 74% performance of victim DNN in average.

Dataset & Model	Test Accuracy (Our evaluation)	
	Victim DNN	Surrogate DNN
Caltech256 (RN34)	87%	≥ 75% (x0.86)
CUBS200 (RN34)	77%	≥ 51% (x0.66)
CIFAR10 (RN34)	94%	≥ 73% (x0.77)
CIFAR10 (CNN)	87%	≥ 71% (x0.81)
GTSRB (RN34)	98%	≥ 61% (x0.62)

Limitations of Current Approach & Evaluation

1. Both victim and surrogate DNNs are pre-trained with the same dataset.
2. Overlap between victim DNN's training data and natural samples.
3. Requires full probability output to work well.
4. Imbalanced training dataset.
5. Low class accuracy for less seen samples.

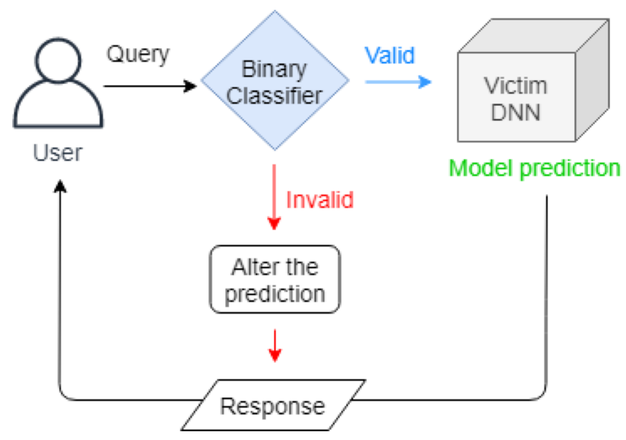


Distribution of training dataset for attacker

Class Names in GTSRB dataset	Test Accuracy	
	Victim DNN	Surrogate DNN
20	100%	88%
21	99%	0%
39	94%	0%
40	99%	94%

Detecting Anomalous Queries

1. Detect *out-of-target* distribution queries.
2. Binary classifier trained with victim dataset/ ImageNet samples.
3. High accuracy in 4 out of 5 test setup.



Detection of anomalous queries and possible prevention mechanism

Victim Dataset	Accuracy	
	Benign queries	Anomalous queries
CUBS200	93%	93%
Caltech256	63%	56%
GTSRB	99%	100%
CIFAR10	96%	96%
Diabetic5	99%	99%

[1] Orekondy, Tribhuvanesh et al. "Knockoff Nets: Stealing Functionality of Black-Box Models" , CVPR19.

[2] Juuti, Mika et al. "PRADA: Protecting Against DNN Model Stealing Attacks", EuroS&P19.

intel® Research Institute for Collaborative Autonomous and Resilient Systems