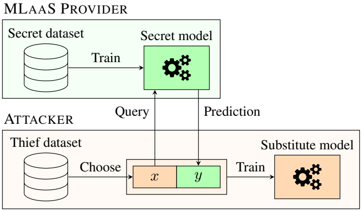
ActiveThief: Model Extraction Using Active Learning and Unannotated Public Data

Soham Pal,*[1] Yash Gupta,*[1] Aditya Shukla,*[1] Aditya Kanade,#[1,2] Shirish Shevade,#[1] Vinod Ganapathy#[1]

Attackers can extract MLaaS models by training a **substitute model** on labeled data obtained by repeatedly querying the service provider's **secret model**.

Our Approach: Vast amounts of unlabeled public data + active learning.



MLaaS Provider

Secret dataset

Train

Secret model

Query

Prediction

Attacker

Thief dataset

Choose

x

y

Train

Substitute model

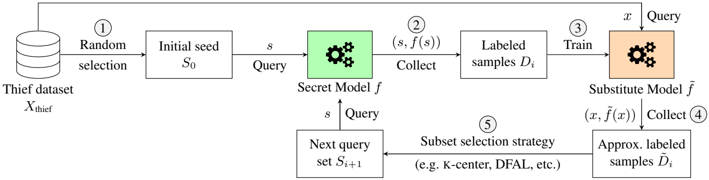
Contributions

- Our approach works on deep neural networks (DNNs).
- It can operate under a limited query budget.
- It does not require access to problem domain data.
- It does not require access to labeled non-problem domain data.
- It evades a state-of-the-art detection mechanism, PRADA (Juuti *et al.*, 2019).

[1] Indian Institute of Science, Bangalore, India [2] Google Brain, USA

*,# Equal contribution.

ACTIVETHIEF overview



1 Random selection
Thief dataset X_{thief}
Initial seed S_0
s Query
Secret Model f
2 $(s, f(s))$ Collect
Labeled samples D_i
3 Train
x Query
Substitute Model $\tilde{f}$
$(x, \tilde{f}(x))$ Collect 4
Approx. labeled samples $\tilde{D}_i$
5 Subset selection strategy (e.g. K-center, DFAL, etc.)
Next query set S_{i+1}
s Query

- Careful selection of samples helps.
- Different Active Learning strategies can complement each other well.
- Choice of strategy should depend on extraction objective.

	MNIST	**CIFAR-10**	**GTSRB**
Random	95.90%	71.38%	79.49%
Uncertainty	96.77%	72.99%	80.09%
DFAL	96.84%	71.52%	83.43%
K-center	96.47%	72.97%	83.59%
DFAL, then K-center	**97.65%**	**73.47%**	**84.29%**
	(+1.82%)	**(+2.92%)**	**(+6.04%)**

Acknowledgements: We thank ACM India-IARCS Travel Grant, AAAI student scholarship and Sonata Software, Bangalore, India for supporting travel to AAAI-20.