Effect of parameter tuning on reducing the number of queries required to perform model stealing

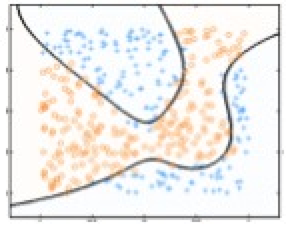
Course: CSE 3000
Author: Floris van Veen
Supervisors: Chi Hong, Jiyue Huang
Responsible professor: Stefanie Roos
Contact information: B.W.M.F.vanVeen@student.tudelft.nl

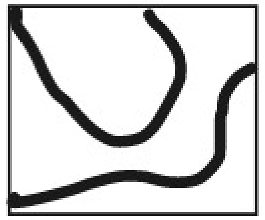
1) Background

Model stealing
We have no information on the target model
We have little to no data examples

1.2) What is the end goal and how can this be achieved?

[2]

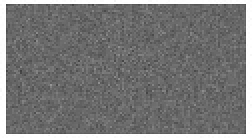




What we want : A model with the same decision boundaries to victim model

Solution:
Create our own model by using the information from the target model by interacting with (querying) it

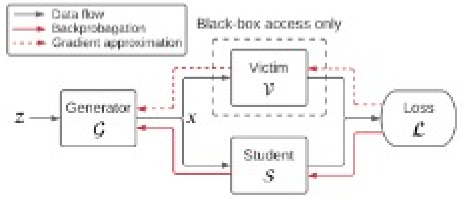
Use generated images to achieve this



Not important if they are similar to target dataset

1.3) Data extraction method

[1]



1. Generator is fed a vector of random noise (z) and generates an image from it (x)
2. V is queried to train G to generate appropriate images
3. G trains S on the learned data
4. loss is calculated between V and S
5. Repeat using the gained information

G wants to maximise the loss between S and V
S wants to minimise the loss between S and V

2) Problem analysis

0.92x target accuracy with Q = 20M Very good accuracy (cifar-10 dataset)

0.94x target accuracy with Q = 30M unreliable replication Requires an incredibly high amount of queries

3) Research question

Goal - Reduce the number of queries required to achieve similar levels of accuracy

- All other found research has hyperparameter tuning unchecked

RQ: How can we reduce the number of queries which are required to perform model stealing?

4) Methodology

1) Observe how different hyperparameters influence student accuracy

2) Utilise hyperparameter optimization algorithms
- Grid search - Exhaustive search from input parameters
- Random search - Random parameters within a given range

3) Compare results and determine if it is viable to reduce number of queries

Perform for MNIST and fashion-MNIST

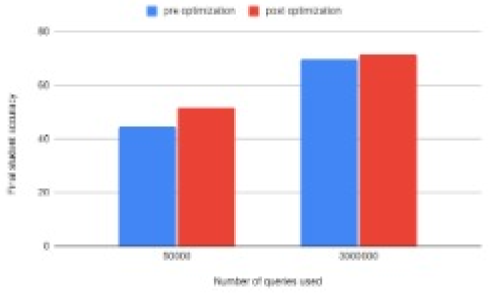
5) Results

MNIST is very easy
- Most hyperparameter combinations reach 95%+ accuracy in under 125000 queries
- With optimization, this can go below 35000

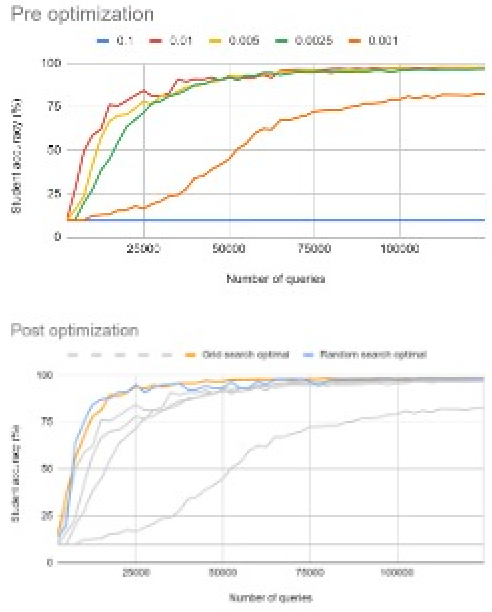
Fashion-MNIST is much more difficult.

We did not reach beyond ~70% accuracy

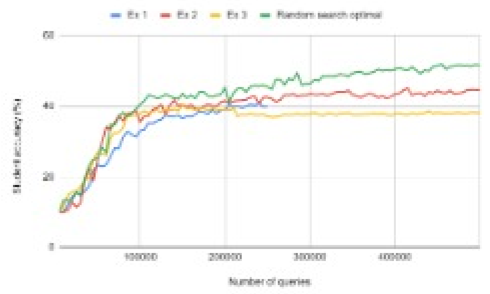
Best results:



6) MNIST learning rate (LR)



7 a) Fashion-MNIST LR



Ex1, 2, 3 - pre optimization

Random search optimal - post optimization

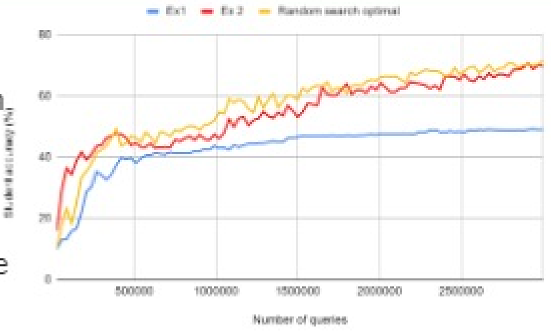
Random search used with the resulting best set of hyperparameters

7 b) Fashion-MNIST LR

Ex1 and 2 - pre optimization

Random search optimal - post optimization

Used same parameters as 7a for the random search optimal. Batch size is different for this experiment, and it is less effective comparatively. Cannot extrapolate results for different batch sizes



8) Conclusion

Viable method to reduce required queries, up to a certain extent
- Only if it is possible to perform optimization beforehand

MNIST required queries: 125,000 -> 35,000 to achieve 95%
Fashion MNIST : 50,000 -> 20,000 to achieve 45%

- Randomness of the input vector Z can influence the result quite heavily. Reduces reliability of if the results are the 'true optimal'

[1] Jean-Baptiste Truong, Pratyush Maini, Robert J. Walls, and Nicolas Papernot. Data-free model extraction, 2020

[2] Sumith, 2018. Why do neural networks work so well?. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/38595451/why-do-neural-networks-work-so-well> [Accessed 20 June 2022].