ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Machine learning privacy: analysis and implementation of model extraction attacks**
Author's name:	**Karafiát Vít**
Type of thesis :	Master Thesis
Faculty/Institute:	Faculty of Electrical Engineering
Department:	Department of Computer Science
Thesis reviewer:	Elnaz Babayeva
Reviewer's department:	External - Avast

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — **Above average challenging**

Evaluation of thesis difficulty of assignment.

The topic of privacy attacks against Machine Learning(ML) models emerged in the past year since ML is now used in many real-world applications. The attacks on the models could lead to a leakage of customer's information, which is a severe and expensive problem. Since it is a new topic, there is not much research done in the area. This thesis creates an easy-to-use framework that allows straightforward testing and comparison of model extraction attacks. The thesis's topic requires a deep understanding of the machine learning topic and good coding skills. The created framework and the improved attacks could be used as in academia for future research, so in industry to make models robust to the attacks. Moreover, to my knowledge, the topic of privacy attacks was not covered in the student's Master or Bachelor syllabus, so the student has to extra work to be introduced to the general topic and concept.

Satisfaction of assignment — **fulfilled**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

The thesis fulfills all the objectives defined in the assignment. This thesis covers different types of attacks (based on data perturbation, on reinforcement learning or GANs), implements a framework, which could be easily configured to use different victim and substitute models, datasets, and attacks, and also improves a runtime performance and accuracy of the attacks.

Method of conception — **A**

Assess that student has chosen correct approach or solution methods.

The thesis has a valid and thought through methodology: chosen attacks follow different methods and ideas, the experiments were done thoroughly and based on three datasets. Most of the experiments were time-consuming, so I believe it was challenging to do extra experiments. The created framework is written in python and the widely used PyTorch library, which could be fastly adopted in the community. The suggestions of improved attacks are justified by experiments.

Technical level — **A**

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

The student shows that he has a good overview of machine learning and has excellent coding skills. The chosen attacks, metrics, models, and experiments are clearly defined, showing that the student understands the problem and looks properly for the solution. Some parts of the thesis are devoted to instructions on using the framework with extensive descriptions and examples. The implemented attacks are verified and compared to the original papers to ensure the correctness of the framework. I would appreciate seeing the experiments when the victim and substitute models are of different types (for example, one is LightGBM, and another is DNN), and experiments on how well the attacks perform on the EMBER dataset.

Formal and language level, scope of thesis — **C**

Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.

The thesis is well structured and it is easy to follow the main ideas. The thesis is written in good English, which is not the student's native language. The student is able to express his ideas in clear form. However, I have some minor troubles understanding some specifics in the descriptions of the attacks and have to go to the source code. Also some graphs and tables are hard to understand (Table 3.1, Figure 6.4).

Selection of sources, citation correctness **A**

Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.

The student has done extensive research on privacy attacks and reviewed sources relevant to the thesis. Citations were correctly done.

Additional commentary and evaluation

Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

The thesis presents a framework MET to test and compare different attacks on machine learning models. The framework contains five different types of attacks, validated on three different datasets, with two types of adversarial datasets. The framework is open-source and could be used by a community to test how their machine learning models are vulnerable to attacks. MET is easily configurable: the user can create a custom type of victim and substitute models, datasets, and types of the attack. The code is written in a clean way, with comments, loggers, and in general, the code is self-descriptive. Moreover, the student improved attack methods: the implementation of the ActiveThief k-center is thousands of times faster, and the proposed new metric k-center + entropy surpasses the SOTA on test accuracy FashionMnist Dataset. Improvements of BlackBox Ripper attacks surpass all the metrics on CIFAR10 and FashionMnist with fewer queries to the victim model. Throughout the thesis, the student suggests various improvements for future work, which shows his great interest in the topic. The methodology is sound, and the best practices are used. The results are impactful, and I believe that they will be used in the community.

Questions:

- In the example usage of MET in Chapter 4, you have used EMBER2018, the security dataset of malware and benign features of Windows PE files. Have you tried to use attacks on the reported LightGBM model? Do the attacks behave similarly to the computer vision settings?

- In Table 3.1, you are claiming that the BlackBox attack is not suitable for other domains than computer vision. Why?
- What kind of library have you used for the k-center fast that runs in GPU-s? Since k-center fast is limited to small datasets, could we use Elastic Search (or another DB search) to run it for the larger datasets?
- For the experiments on CIFAR10, you have used data augmentation as cropping and flipping since the victim model has better test accuracy. Why haven't you implemented a similar augmentation for GTSRB and FashionMnist? Is the behavior on these datasets different?

The grade that I award for the thesis is **A.**

Date: **14.06.2021** Signature: