SECURE DEEP LEARNING INFERENCE WITH INTEL SOFTWARE GUARD EXTENSION ON INTEL ICE LAKE-SP XEON PROCESSOR

PHUAH CHAI LING

UNIVERSITI TEKNOLOGI MALAYSIA

SECURE DEEP LEARNING INFERENCE WITH INTEL SOFTWARE GUARD EXTENSION ON INTEL ICE LAKE-SP XEON PROCESSOR

PHUAH CHAI LING

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Computer and Microelectronic Systems)

School of Electrical Engineering
Faculty of Engineering
Universiti Teknologi Malaysia

JULY 2022

DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGEMENT

First of all, I would like to express my deepest appreciation to my thesis supervisor, Dr Usman Ullah Sheikh for his constant guidance and constructive feedback throughout this research. Dr Usman Ullah Sheikh has been patient and open in sharing his knowledge and providing support, in which without any of those, this research would not have been completed as smoothly as it has been. Besides, I would also like to acknowledge and thank Universiti Teknologi Malaysia (UTM) for the opportunity to conduct such meaningful research that could make a little if not big contribution to the research field.

In addition, I am also indebted to Intel Lab for providing me the tools and platform needed for this research. This research would not have been possible without the aid provided. Besides, I would like to extend my sincere appreciation to my colleagues for their constant encouragements and countless support in the completion of this research.

Last but not least, I am grateful for all the support and care that my family, friends and fellow postgraduate students have provided me throughout this duration as well as everyone else who have helped me directly or indirectly in the completion of this research.

ABSTRACT

In the current technology-driven era where the world is powered by the internet, artificial intelligence has never been more popular and so as the cloud computing or edge computing technology. With artificial intelligence being on an upward trend in the digitalized world, many industries have started to deploy artificial intelligence in their applications. Among many of the techniques used for artificial intelligence, deep learning is one of the most popular techniques deployed, thanks to its remarkable reputation that outperforms many of the other methodologies. Hence, deep learning inference has been actively used in numerous applications nowadays, even for those that are security critical. On top of that, with cloud computing and edge computing becoming more common now, many industries have also started the migration of their applications to the cloud or to the edge. All these raised a serious security concern that should be tackled with best effort. Throughout the years, many researchers have been active in exploring the solutions to secure a deep learning inference. Cryptographic primitives and trusted hardware are among the most popular methodologies proposed. Nevertheless, most of the efforts focused only on preserving the security of the input data and the deep learning model without providing any security to the application code or the inference forward pass. Since cryptographic primitives are known to incur a high performance overhead, this research proposed to secure a deep learning application via the trusted hardware approach, specifically Intel SGX on the recently launched 3rd Gen Intel® Xeon Scalable processor. An image classification application was used as an example for the case study in this research and performance evaluation was conducted mainly based on the performance impacts of SGX in terms of the time taken for a model to be loaded to the CPU, the number of inferences per second, the total application runtime as well as the parallel efficiency of the application. In this research, 5 different deep learning models had been tested across 3 different environments and the findings are all presented in this project report. Results showed that when the image classification application was placed within the SGX enclave, there was an overhead of up to 9X for the time taken to load a model to the CPU and as high as 13X for the overall application runtime. The number of inferences per second and the parallel efficiency of the application both suffered from a loss of up to 70% when being compared to the native environment. Nonetheless, this research has demonstrated that with the greatly expanded Intel SGX enclave size, it is feasible to secure the deep learning application with Intel SGX without any code partitioning despite the trade-off on the performance. The findings from this research could serve as a reference for those who wish to explore further on this piece and could also serve as one of the works that unleashed the potential of the first Intel Xeon Scalable Processor that comes with Intel SGX support.

ABSTRAK

Pada era teknologi semasa di mana dunia dikuasai oleh internet, kecerdikan buatan, pengkomputeran awan dan pengkomputeran pinggir adalah amat popular. Dengan kecerdikan buatan berada pada tren kenaikan, banyak industri telah mula menggunakan kecerdikan buatan dalam aplikasi mereka. Antara teknik-teknik yang digunakan untuk kecerdikan buatan, pembelajaran mendalam merupakan salah satu teknik yang paling popular disebabkan reputasinya yang luar biasa berbanding dengan teknik lain. Oleh itu, inferens pembelajaran mendalam telah digunakan secara aktif dalam banyak aplikasi pada masa ini, termasuk aplikasi yang memerlukan keselamatan tinggi. Di samping itu, dengan pengkomputeran awan dan pengkomputeran pinggir menjadi lebih umum sekarang, kebanyakan industri juga telah memulakan migrasi aplikasi mereka ke awan atau pinggir. Semua ini menimbulkan masalah keselamatan yang serius yang harus ditangani dengan usaha terbaik. Ramai penyelidik sedang aktif dalam menerokai teknik untuk menjamin keselamatan inferens pembelajaran mendalam. Primitif kriptografi dan perkakasan yang dipercayai adalah antara teknik paling popular yang dicadangkan. Namun begitu, sebahagian besar usaha hanya tertumpu dalam menjamin keselamatan data input dan model pembelajaran mendalam tanpa tumpuan pada kod aplikasi atau hantaran depan. Disebabkan primitif kriptografi diketahui menanggung overhed prestasi yang tinggi, penyelidikan ini mencadangkan untuk menjamin keselamatan aplikasi pembelajaran mendalam melalui perkakasan yang dipercayai, khususnya Intel SGX pada Intel® Xeon Scalable Processor Gen ke-3 yang baru dilancarkan. Aplikasi klasifikasi gambar telah digunakan sebagai contoh untuk kajian kes dalam penyelidikan ini dan penilaian prestasi telah dilakukan terutama pada kesan SGX dari segi masa yang diambil untuk memuatkan satu model ke CPU, bilangan inferens sesaat, jumlah waktu pelaksanaan aplikasi, serta kecekapan selari aplikasi. Dalam penyelidikan ini, 5 model pembelajaran mendalam yang berbeza telah diuji dalam 3 persekitaran yang berbeza dan kesemua penemuannya telah dibentangkan dalam laporan ini. Keputusan menunjukkan bahawa apabila aplikasi klasifikasi imej ini dilaksanakan bersama Intel SGX, masa yang diambil untuk memuatkan model ke CPU mengalami overhed setinggi 9X, manakala jumlah waktu pelaksanaan aplikasi mengalami overhed setinggi 13X. Bilangan inferens sesaat dan kecekapan selari aplikasi kedua-duanya mengalami kejatuhan sehingga 70% apabila dibandingkan dengan persekitaran asli. Walau bagaimanapun, penyelidikan ini telah menunjukkan bahawa dengan saiz enklaf Intel SGX yang diperluaskan, keselamatan aplikasi pembelajaran mendalam dapat dijamin dengan Intel SGX tanpa sebarang pembahagian kod walaupun terdapat kesan pada prestasi. Penemuan daripada penyelidikan ini boleh dijadikan sebagai satu rujukan bagi mereka yang ingin menerokai kerja ini dengan lebih lanjut dan juga boleh dijadikan salah satu karya yang menzahirkan potensi Intel Xeon Scalable Processor yang pertama dengan sokongan Intel SGX.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
BIOS	-	Basic Input Output System
BLOB	-	Binary Large Object
CPU	-	Central Processing Unit
DDR4	-	Double Data Rate 4
DIMM	-	Dual In-line Memory Module
DL	-	Deep Learning
DNN	-	Deep Neural Network
EPC	-	Enclave Page Cache
GB	-	Gigabyte
GC	-	Garbled Circuit
GCC	-	GNU Compiler Collection
GLIBC	-	GNU C Library
GPU	-	Graphics Processing Unit
INTEL SGX	-	Intel Software Guard Extension
INTEL SGX-TEM	-	Intel Software Guard Extension – Trusted Environment Mode
IR	-	Intermediate Representation
ISR	-	Intelligence, Surveillance and Reconnaissance
LLC	-	Last Level Cache
MB	-	Megabyte
MLC	-	Mid-Level Cache
NN	-	Neural Network
OS	-	Operating System
SDK	-	Software Development Kit
SKU	-	Stock Keeping Unit
TEE	-	Trusted Execution Environment
TPM	-	Trusted Platform Module

LIST OF SYMBOLS

%	-	Percentage

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Problem Background

What is Artificial Intelligence (AI)? It is all about granting the machines the ability to perform in the same manner as humans. The concept of Artificial Intelligence was introduced in the 1950 but it is not until recent years that it became a talking point. With the great advancement in technology, we are now seeing a rapid increase in the deployment of AI in many applications in today's world. Being part of Artificial Intelligence, machine learning had long proven itself to be one competitive technique that allows the machine to learn from its historical data so that it can make a more accurate prediction on the outcome without explicit programming. With the prominent growth in AI popularity, more industries are now counting on the AI solutions. Thus, the need for the machines to handle higher level of complexity tasks as compared to the older days is ever increasing and this led to the emergence of deep learning, which is a technique that falls under the same umbrella as the machine learning, but more powerful. There are basically two main phases in deep learning, training phase and inference phase. While deep learning training is about training a deep neural network model with a deep learning framework and training datasets, deep learning inference is referring to the use of a machine to make certain predictions on a new incoming data based on the previously trained neural network. With the distinctive reputation that deep learning has gained over the years, there has been a tremendous increase in the use of deep learning inference even in the security critical applications. For instance, some of the healthcare service now rely on the AI for medical diagnosis, while the military may count on AI for their Intelligence, Surveillance and Reconnaissance (ISR) operations. These are some of the many applications that deal with very highly confidential data or trade secrets that should be kept as confidential as possible. Moreover, with the emerging cloud computing and edge computing technologies,

many industries have also started to migrate their applications to the cloud or to the edge. Whether they are offloading their applications to the cloud or to the edge, or leveraging the cloud inference as a service, all these raise a serious security concern that is worth pondering. Therefore, this research is focusing on the topic of securing the deep learning inference.

1.2 Problem Statement

It is no doubt that AI has helped to bring much convenience to its users in many forms and allowed many of the seemingly impossible tasks to turn feasible. Nevertheless, the security concern raised along with the emergence of these technologies has also been on the rising edge especially in the current era where hackers are also becoming more and more advanced. In the effort of improving the security of deep learning inference, many techniques have been explored throughout the years.

One of the most common methods used to secure the deep learning inference or applications is the cryptographic primitives. However, these software-based cryptographic scheme solution often suffer from a very high performance overhead and is limited to only protecting the data security, not the code security. In addition, most of the previous works on securing deep learning inference focused only on preserving the privacy of the input data and the neural network models but neglected the security of the application code or the forward pass of the inference, which is another important element in the inference process. Under such circumstances, in any event where the malicious attacker successfully tampers the operating system, he or she will then be able to attack the application which could possibly result in a wrong inference result, or data in-use getting stolen. Thus, this led to the motivation of this research to secure the deep learning inference application with a hardware-based solution, particularly via the trusted execution environment.

1.3 Research Goal

The aim of this research was to deploy a hardware-based solution to secure the deep learning application whereby the application was placed inside the trusted execution environment that is claimed safe by the specifications. Thus, the forward pass of an inference process was conducted inside the trusted execution environment. Besides, the impacts of the proposed solution on the performance of the application was also another area of study in this research.

1.3.1 Research Objectives

The main objectives of this research are:

(a) To secure deep learning application using Intel Software Guard Extension – Trusted Environment Mode (Intel SGX-TEM) with Gramine LibOS on the latest 3rd Gen Intel® Xeon Scalable processor.

(b) To conduct performance evaluation on the Gramine-SGX secure deep learning inference vs non-secure deep learning inference.

1.4 Scope and Limitations

This research focused on specific scopes and was bounded by certain limitations. It should be highlighted that Intel Software Guard Extension was deployed as the sole trusted execution environment and this work was based on the Intel's latest 3rd Gen XEON Scalable processor, specifically Ice Lake-SP only. For the performance evaluation purposes, the neural network model loading time, number of inferences per second, total application runtime and parallel efficiency of the application shall be the performance metrics to be considered. In addition, this research did not consider the security vulnerabilities of the Intel Software Guard Extension secured memory region. In another words, it was trusted that anything put inside the Intel SGX enclave is safe and secure in this work as per claimed by the Intel SGX specifications.

1.5 Thesis Outline

This thesis consists of five main chapters whereby the first chapter gives some introduction on the background of the research topic, highlights the problem statement, defines the research goal and objectives, outlines the scope and limitations of this research as well as the thesis structure to help readers navigate through this thesis. Meanwhile, Chapter 2 covers the literature review including the technology involved in this work and reviews on the previous related work. Chapter 3 on the other hand details the research methodology including the overall research flow, the hardware and software tools used and the proposed solution implementation. All the results, analysis and discussion are covered in Chapter 4 while Chapter 5 summarizes the research and includes some suggestions for future work.

REFERENCES

Chakrabarti, S., Knauth, T., Kuvaiskii, D., Steiner, M., & Vij, M. (2020). Trusted execution environment with Intel SGX. In *Responsible Genomic Data Sharing*. Elsevier Inc. https://doi.org/10.1016/b978-0-12-816197-5.00008-5

Elgamal, T., & Nahrstedt, K. (2020). Serdab: An IoT Framework for Partitioning Neural Networks Computation across Multiple Enclaves. *Proceedings - 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGRID 2020*, 519–528. https://doi.org/10.1109/CCGrid49817.2020.00-41

Gu, Z., Huang, H., Zhang, J., Su, D., Jamjoom, H., Lamba, A., Pendarakis, D., & Molloy, I. (2018). *Confidential Inference via Ternary Model Partitioning*. http://arxiv.org/abs/1807.00969

Juvekar, C., Vaikuntanathan, V., & Chandrakasan, A. (2018). GAZELLE: A low latency framework for secure neural network inference. *Proceedings of the 27th USENIX Security Symposium*, 1651–1668.

Kumar, N., Rathee, M., Chandran, N., Gupta, D., Rastogi, A., & Sharma, R. (2020). CrypTFlow: Secure TensorFlow inference. *Proceedings - IEEE Symposium on Security and Privacy*, *2020-May*, 336–353. https://doi.org/10.1109/SP40000.2020.00092

Le Quoc, D., Gregor, F., Arnautov, S., Bhatotia, P., Kunkel, R., & Fetzer, C. (2020). SecuretF: A secure tensorflow framework. *Middleware 2020 - Proceedings of the 2020 21st International Middleware Conference*, 44–59. https://doi.org/10.1145/3423211.3425687

Li, S., Xue, K., Zhu, B., Ding, C., Gao, X., Wei, D., & Wan, T. (2020). FalCon: A Fourier transform based approach for fast and secure convolutional neural network predictions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 8702–8711. https://doi.org/10.1109/CVPR42600.2020.00873

Li, Y., Zeng, D., Gu, L., Chen, Q., Guo, S., Zomaya, A., & Guo, M. (2021). Lasagna: Accelerating secure deep learning inference in SGX-enabled edge cloud. *SoCC 2021 - Proceedings of the 2021 ACM Symposium on Cloud Computing*, 533–545. https://doi.org/10.1145/3472883.3486988

Liu, J., Juuti, M., Lu, Y., & Asokan, N. (2017). Oblivious neural network predictions via MiniONN transformations. *Proceedings of the ACM Conference on Computer and Communications Security*, 619–631. https://doi.org/10.1145/3133956.3134056

Liu, R., Garcia, L., Liu, Z., Ou, B., & Srivastava, M. (2021). SecDeep: Secure and Performant On-device Deep Learning Inference Framework for Mobile and IoT

Devices. In *IoTDI 2021 - Proceedings of the 2021 International Conference on Internet-of-Things Design and Implementation* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3450268.3453524

Mo, F., Shamsabadi, A. S., Katevas, K., Demetriou, S., Leontiadis, I., Cavallaro, A., & Haddadi, H. (2020). DarkneTZ: Towards model privacy at the edge using trusted execution environments. *MobiSys 2020 - Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, 161–174. https://doi.org/10.1145/3386901.3388946

Plauth, M., Teschke, F., Richter, D., & Polze, A. (2018). Hardening application security using intel SGX. *Proceedings - 2018 IEEE 18th International Conference on Software Quality, Reliability, and Security, QRS 2018*, 375–380. https://doi.org/10.1109/QRS.2018.00050

Prabhu, K., Jun, B., Hu, P., Asgar, Z., Katti, S., & Warden, P. (2021). Privacy-Preserving Inference on the Edge : Mitigating a New Threat Model. *TinyML Research Symposium*, 1–9.

Qiu, H., Zheng, Q., Zhang, T., Qiu, M., Memmi, G., & Lu, J. (2021). Toward Secure and Efficient Deep Learning Inference in Dependable IoT Systems. *IEEE Internet of Things Journal*, *8*(5), 3180–3188. https://doi.org/10.1109/JIOT.2020.3004498

Rouhani, B. D., Riazi, M. S., & Koushanfar, F. (2018). DeepSecure : Scalable Provably-Secure Deep Learning. *DAC '18: Proceedings of the 55th Annual Design Automation Conference*.

Sadegh Riazi, M., Samragh, M., Lauter, K., Chen, H., Koushanfar, F., & Laine, K. (2019). Xonn: XNOR-based oblivious deep neural network inference. *Proceedings of the 28th USENIX Security Symposium*, *ii*, 1501–1518.

Tramèr, F., & Boneh, D. (2019). Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. *7th International Conference on Learning Representations, ICLR 2019*.

VanNostrand, P. M., Kyriazis, I., Cheng, M., Guo, T., & Walls, R. J. (2019). *Confidential Deep Learning: Executing Proprietary Models on Untrusted Devices*. http://arxiv.org/abs/1908.10730