Quantum Computing Methods for Malware Classification

Ing. Eliška Krátká [1], Supervisor: Aurél Gábor Gábris, Ph.D. [2]

[1]Department of Information Security, Faculty of Information Technology, Czech Technical University in Prague

[2]Department of Physics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague

Motivation

The currently available quantum computers, known as NISQ (Noisy Intermediate-Scale Quantum) devices, have been proven to be more powerful than any existing supercomputer. While these experiments demonstrated what is termed quantum supremacy, the algorithms used are **not known to be related to any practical problem**.

One promising research area for NISQ devices is the integration of quantum computing with machine learning, known as Quantum Machine Learning (QML) [1]. The thesis focuses on **applying QML to a practical issue: malware detection**. Specifically, it involves implementing a QML-based malware classification model and **evaluating its performance on real IBM quantum computers**.

Given the growing amount of data and increasingly sophisticated malware attacks, malware detection has become an important aspect of information security. By using QML, models trained for malware detection **could achieve greater accuracy**, particularly in scenarios involving smaller datasets, where traditional methods may struggle.

Objectives

The thesis focuses on the Support Vector Machine (SVM) algorithm and its quantum counterpart, the **Quantum Support Vector Machine** (QSVM), in the context of malware detection. QSVM combines the classical SVM classification with a quantum kernel. Quantum kernels promise to outperform classical kernels by addressing challenges posed by high-dimensional feature spaces [2].

The QSVM algorithm is **implemented and applied to a publicly available dataset**, PE Malware Machine Learning, which consists of raw binaries of PE files instead of just metadata already extracted from the samples.

The performance of the QSVM is **evaluated both on simulators and real IBM quantum computers**.

Implementation for IBM Quantum Hardware

The implementation consists of two independent Python modules. The first module focuses on **data extraction and preprocessing**, including extracting specific samples from the dataset and converting them to grayscale images.

The second module is dedicated to the QSVM implementation. It contains a quantum kernel with custom feature maps and provides **a classification interface** compatible with both simulators and IBM quantum computers.

The source codes with detailed documentation are available in the project's GitLab repository (`https://gitlab.fit.cvut.cz/kratkeli/ni-dip`).

Implementation Challenges

Qiskit is an open-source Python SDK for developing quantum algorithms. These algorithms can be executed locally on simulators or real quantum computers available through the IBM Quantum Platform.

While working with Qiskit's Machine Learning library, two **significant challenges** were encountered:

- Issues with quantum circuit transpilation
- Computational job size limits on the IBM Quantum Platform

Both of these issues are known challenges in the Qiskit community and remain unresolved at the time of writing this poster. A key contribution of the thesis is the **development of fixes for these issues**, enabling more efficient use of quantum algorithms for practical machine learning tasks on the IBM Quantum Platform.

Simulator Experiments

We used a quantum computer simulator in Qiskit and **compared the classical SVM algorithm with our implemented QSVM**. We used the same preprocessed data for both models and **compared, verified and expanded results in the existing literature** [3].

Our QSVM model consistently achieves **higher or similar accuracy** than SVM (implementation from the sci-kit-learn library), which suggests the capability of QSVM to extract more information from limited data compared to classical SVM.

Hardware Experiments

Due to the implementation challenges, we shift the focus of experiments on the IBM quantum computers. We used a reduced dataset size due to license restrictions on computation time and analysed computation time and waiting time in the jobs queue.

Computation Time Analysis

- Tested computation time for quantum circuits on different hardware
- Variations up to three "quantum seconds" – significant difference in terms of total available computational time

Analysis of Waiting Time in Queue

- Evaluated wait times in the queue based on the amount of data sent and based on different hardware
- Practical insights for selecting appropriate hardware for experiments

Main Contributions

- Critical fixes for Qiskit's Machine Learning library
- Practical insights into the use of IBM Quantum computing resources in future experiments (**among the first users from the entire university**)
- Expanding experiment results in the existing literature

Conclusion

The thesis investigates the application of quantum computing for malware detection, focusing on evaluating the QSVM algorithm against classical methods. It **addresses critical issues** in the Qiskit Machine Learning library, including problems with quantum kernel evaluation, circuit transpilation, and job size limits on the IBM Quantum Platform. These fixes allow for more efficient use of quantum machine learning, not only in malware detection.

Experiments were conducted on both a local simulator and real IBM quantum computers. As one of the first users at CTU to use the IBM Quantum computers, the thesis provides valuable insights into how quantum computers handle large-scale computations typical for machine learning. The research lays the foundation for future advancements in quantum-enhanced malware detection with a fully compatible implementation for the IBM Quantum interface. The author continues to research this topic during her PhD.

References

[1] BHARTI, Kishor; CERVERA-LIERTA, Alba; KYAW, Thi Ha; HAUG, Tobias; ALPERIN-LEA, Sumner; ANAND, Abhinav; DEGROOTE, Matthias; HEIMONEN, Hermanni; KOTTMANN, Jakob S.; MENKE, Tim; MOK, Wai-Keong; SIM, Sukin; KWEK, Leong-Chuan; ASPURU-GUZIK, Alán. Noisy intermediate-scale quantum algorithms. Reviews of Modern Physics [online]. 2022, vol. 94, no. 1 [visited on 2024-09-12]. Available from: `https://link.aps.org/doi/10.1103/RevModPhys.94.015004`

[2] HAVLÍČEK, Vojtěch; CÓRCOLES, Antonio D.; TEMME, Kristan; HARROW, Aram W.; KANDALA, Abhinav; CHOW, Jerry M.; GAMBETTA, Jay M. Supervised learning with quantum-enhanced feature spaces. Nature. 2019, vol. 567, no. 7747, pp. 209–212. Available from: `https://doi.org/10.48550/arXiv.1804.11326.`

[3] BARRUÉ, Grégoire; QUERTIER, Tony. Quantum Machine Learning for Malware Classification. ArXiv.org [online]. 2023, p. 30 [visited on 2023-09-12]. Available from: `https://doi.org/10.48550/arXiv.2305.09674`

Eliška Krátká, kratkeli@fit.cvut.cz