HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

GRADUATION THESIS

Enhancing Multi-label Vulnerability Detection of Smart Contract using Language Model

VU TUNG DUONG
duong.vt183728@sis.hust.edu.vn

Major: Information Technology
Specialization: Computer Engineering

Supervisor: Dr. Tong Van Van _____

Department: Computer Engineering

School: School of Information and Communications Technology

HANOI, 01/2024

ACKNOWLEDGMENT

Five years of school have passed, neither short nor long, it was enough for us to enjoy our precious student days. Time helps us meet many new good people from all over who come to this university. Some people come, some people leave but they all teach me valuable lessons. I would like to thank my family, my parents, and my siblings for accompanying me throughout my university journey. Thank you to the precious teachers who have taught me useful subjects over the past 5 years. Thank you to my project instructor, Dr.Tong Van Van, an extremely enthusiastic lecturer with students. He accompanied me for a long time to prepare for this research project. His discussions and suggestions are very valuable to help me complete this project properly. Thank you to my university friends who have helped me study during the past 5 years, thanks to which I was able to reach the final stage. I would also like to sincerely thank the Board of Directors and teachers of Hanoi University of Science and Technology for creating the opportunity for me to study at the school to gain knowledge and practical experience to gain useful information for my thesis. Thank you CyStack Vietnam Joint Stock Company for your help in providing and supporting the data set for this research. I sincerely thank you all!

ABSTRACT

With the advancement of Blockchain technology, security is a major issue when designing applications for digital platforms. Smart contracts are a great application of Blockchain. It is considered as decentralized application, so it plays an important role in blockchain-based applications. Smart contracts are written by programming languages (e.g., Solidity, Python, etc.), so it is error-prone and suffers from vulnerabilities, leading to a huge amount of economic loss for the blockchain ecosystem. In the past, there were many existing vulnerability detection tools such as MythX, Oyente, Slither, and so on. However, these tools contain several limitations related to low accuracy and high execution time. Therefore, many studies focus on vulnerability detection mechanisms using Deep Learning which takes into account the bytecode of smart contracts to detect its vulnerabilities. Despite achieving good accuracy, these studies make an assumption that there is only one vulnerability in a smart contract. When there is more than one vulnerability in a smart contract, these studies can not obtain good performance. Therefore, in this thesis, we propose a multi-label vulnerability detection of smart contracts using a language model. Concretely, the proposal takes into account the bytecode by using the SecBERT pre-trained model to extract the implicit features and analyzes it using the Multi-Layer Perceptron algorithm to identify multiple vulnerabilities in a smart contract. The experimental results show that the proposal outperforms benchmarks and obtains 91.54 percent accuracy and 0.0467 seconds execution time.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION 1

1.1 Problem Statement 1

1.2 Background and Problems of Research 3

1.3 Contributions 4

1.4 Organization of Thesis 4

CHAPTER 2. LITERATURE REVIEW 6

2.1 Scope of Research 6

2.2 Background knowledge 8

2.2.1 Smart Contracts 8

2.2.2 Artificial Intelligence 13

2.3 Challenges 24

2.3.1 Underfitting 25

2.3.2 Overfitting 25

2.4 Related Works 26

2.4.1 Vulnerability Detection Tools 26

2.4.2 Vulnerability Detection Mechanisms 27

CHAPTER 3. METHODOLOGY 31

3.1 Overview 31

3.2 Data preprocesssing 32

3.2.1 Data collection 32

3.2.2 Data preprocesssing 33

3.3 Word embedding 34

3.3.1 SecBERT pre-trained model 35

3.3.2 Multi-BERT 36

3.4 Classification 38

CHAPTER 4. NUMERICAL RESULTS 40

4.1 Experimental setup 40

4.1.1 Dataset 40

4.1.2 Evaluation metrics 40

4.1.3 Benchmarks 42

4.2 Results Analysis 44

4.2.1 Comparison with Machine learning methods 44

4.2.2 Comparison with Deep learning methods 45

CHAPTER 5. CONCLUSIONS 47

5.1 Summary 47

5.2 Future Works 47

REFERENCE 50

LIST OF FIGURES

Figure 2.1 Vulnarebility type classification 7
Figure 2.2 Basic architecture of the Ethereum network and the environment in which smart contracts can be executed 9
Figure 2.3 Artificial intelligence 13
Figure 2.4 Classic network architecture of RNN 19
Figure 2.5 LSTM architecture 19
Figure 2.6 Transformers architecture [15] 20
Figure 2.7 Activation function [16] 22

Figure 3.2 Data collecting progress 32
Figure 3.1 Enhancing multi-label vulnerabilities using single BERT for classification 32
Figure 3.3 Data length analysis 33
Figure 3.4 Bert architecture [31] 35
Figure 3.5 Multi-BERT architecture 37

Figure 4.1 Data distribution 41

LIST OF TABLES

Table 3.1 Multi-label classification model 38

Table 4.1 Vulnerability raw dataset 40
Table 4.2 Comparison with Machine learning mechanisms 45
Table 4.3 Comparison with single SecBERT block 46
Table 4.4 Comparison with existing Deep learning mechanisms 46

LIST OF ABBREVIATIONS

Abbreviation	Definition
AA	Adapted algorithm
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
BR	Binary relevance
CBOW	Continuous Bag of Words
CC	CLassifier chain
EOA	Externally Owned Accounts
LP	Label powerset
NLP	Natural Language Processing
PoS	Proof of Stake
PoW	Proof of Work
TF-IDF	Term Frequency - Inverse Document Frequency
W2V	Word2Vec

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

Blockchain [1] is a decentralized information storage and transmission technology, developed in 2008 by Satoshi Nakamoto. This technology is used to confirm and store transactions and information online with high security. A blockchain is a chain of data blocks linked together using encryption and arithmetic algorithms to ensure data integrity. Each block contains information about transactions performed on the network and a hash code representing the previous block. This hash code will be used to confirm the integrity of the next block on the chain. With blockchain, data is stored on many different nodes on the network, without a single control center, so no one can tamper with the information and modify it easily. This makes blockchain a useful technology in many fields, including finance, healthcare, supply chains, and many others. Some applications of blockchain can include:

- Cryptocurrency: Cryptocurrencies such as Bitcoin, Ethereum, Ripple, Litecoin, and many others are created on the blockchain platform. Cryptocurrencies allow users to exchange money without going through intermediary financial institutions.
- Supply chain management: Blockchain is used to monitor the entire process of manufacturing, transporting, and storing products. This helps reduce paperwork and ensure transparency throughout the entire process.
- Asset management: Blockchain provides a reliable solution for asset management, including traditional assets such as real estate, cars, cosmetics, and jewelry.
- Online voting: Blockchain provides a secure and reliable solution for online voting, ensuring the integrity of election results.
- File storage: Blockchain provides a decentralized file storage solution, ensuring data integrity and security.
- Healthcare: Blockchain can be used to manage patient medical information, ensure the security and privacy of medical information, and help increase information sharing between health agencies.
- Copyright management: Blockchain can be used to manage copyrights of many types of assets, including music, books, movies, and other creative products.
- Financial management: Blockchain can be used to manage loans, loans, and other financial transactions, ensuring integrity and transparency in the transaction

process.

- Smart contract management: Blockchain provides a smart contract management solution that simplifies verifying and executing contracts while minimizing dependence on intermediaries.

Therefore, it can be seen that blockchain technology has many applications and development potential in many different fields. However, this technology also has vulnerabilities that are targets for bad actors to attack. In terms of cryptocurrency, there have been many attacks that have affected the financial economy. There are some well-known attacks:

- 51% Attack: An attack on a cryptocurrency blockchain by a group of miners who control more than 50% of the network's mining hash rate. Owning 51% of the nodes on the network theoretically gives the controlling parties the power to alter the blockchain. The attackers would be able to prevent new transactions from gaining confirmations, allowing them to halt payments between some or all users. They would also be able to reverse transactions that were completed while they were in control. Reversing transactions could allow them to double-spend coins, one of the issues consensus mechanisms like proof-of-work were created to prevent. For example, in May 2018, Bitcoin Gold experienced a 51% attack that allowed the attacker to double-spend approximately %18 million worth of BTG. This event caused substantial damage to the coin's reputation and market value.
- DAO attack: An attack implemented in 2016 on the Ethereum system. The attacker has found a vulnerability in DAO's smart contracts (Decentralized Autonomous Organization) and used this vulnerability to steal more than 3,6 million Ethereum units (about $50 million at that time). This incident sparked a crucial discussion in the blockchain community regarding the safety of blockchain and the role of smart contracts.
- Mt.Gox Attack: Mt.Gox used to be the world's largest bitcoin exchange but eventually collapsed in 2014. The attackers found a vulnerability in their software and stole more than 850000 Bitcoins (about $450 million at that time). This attack is one of the most damaging attacks in blockchain history and has caused a lot of controversy about the safety of cryptocurrency exchanges.
- Parity Wallet Attack: In 2017, a cyber attack targeted Parity Wallet, a widely-used Ethereum wallet. The attacker exploited a vulnerability in the wallet software and gained unauthorized access to users' wallets, stealing over 150,000 Ethereum units, valued at approximately $30 million at the time. This incident