Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Informatica

Machine Learning Vulneralibities Detection in SmartContracts .....

Relatore:
Chiar.mo Prof.
Stefano Ferretti

Presentata da:
Marco Benito Tomasone

Correlatore:
Dott. Stefano Pio Zingaro
Dott. Saverio Giallorenzo

Sessione I
Anno Accademico 2023/2024

Alle nonne

Indice

Elenco delle figure

Capitolo 1

Introduzione

Negli ultimi anni una delle tecnologie che sono spopolate e sono arrivate sulla bocca di tutti sono sicuramente la Blockchain e gli SmartContracts. Questi ultimi sono dei contratti digitali che permettono di eseguire delle operazioni in modo automatico e trasparente. Questi contratti sono scritti in un linguaggio di programmazione e vengono eseguiti su una macchina virtuale. Una delle principali caratteristiche peculiari degli SmartContracts è sicuramente la loro immutabilità, difatti una volta essere stati deployati sulla blockchain questi non possono più essere modificati. Uno dei problemi principali degli SmartContracts è la sicurezza. Infatti, essendo dei contratti che vengono eseguiti in modo automatico, è possibile che ci siano delle vulnerabilità che possono essere sfruttate da malintenzionati. In questo lavoro di tesi verrà presentato un metodo per rilevare le vulnerabilità presenti negli SmartContracts.

Capitolo 2

Related Works

In prima battuta questo lavoro è stato possibile grazie ad un lavoro precedente che ha permesso di creare un dataset di SmartContracts [2]. Questo dataset è stato creato da un gruppo di ricercatori dell'Università di Bologna. Questo primo lavoro su questo datataset ha principalmente impiegato delle tecniche di Deep Learning utilizzando delle reti neurali convoluzionali per la rilevazione delle vulnerabilità trasformando in codice Bytecode dei contratti in delle immagini RGB. Questo lavoro ha come risultato principale la dimostrazione che utilizzando delle reti neurali convoluzionali è possibile rilevare le vulnerabilità presenti negli SmartContracts con delle buone performance, i migliori risultati si attestano con un MicroF1 score del 0.83% e mostrano come i migliori risultati siano dati da delle resnet con delle convoluzioni unidimensionali. Successivamente, gli stessi autori hanno pubblicato una seconda analisi effettuata sul dataset utilizzando nuovi classificatori come CodeNet, SvinV2-T e Inception, mostrando come i migliori risultati continuino ad essere quelli forniti da reti convoluzionali unidimensionali [**?**].

Capitolo 3

Metodologia

3.1 BERT

Il modello BERT (Bidirectional Encoder Representations from Transformers) rappresenta un pilastro fondamentale nel campo del Natural Language Processing (NLP), grazie alla sua capacità di comprensione del contesto delle parole all'interno di una frase o di un testo più ampio. Questo modello, sviluppato da Google, si basa sull'architettura dei transformer, una classe di modelli neurali che ha dimostrato notevole successo nell'analisi del linguaggio naturale.

BERT si distingue per la sua capacità bidirezionale di elaborare il contesto linguistico. A differenza dei modelli NLP precedenti, che processavano il testo in modo sequenziale, interpretando le parole una dopo l'altra, BERT considera sia il contesto precedente sia quello successivo di ciascuna parola all'interno di una frase. Questo approccio bidirezionale consente a BERT di catturare relazioni semantiche più complesse e di fornire una rappresentazione più accurata del significato del testo.

Il funzionamento di BERT può essere compreso attraverso due fasi principali: l'addestramento e l'utilizzo.

Durante la fase di addestramento, BERT viene esposto a enormi quantità di testo, proveniente da varie fonti e domini. Utilizzando un processo noto come "pre-addestramento", il modello apprende i modelli linguistici e il contesto delle parole. Questo pre-addestramento coinvolge due compiti principali: la predizione di parole mascherate e la predizione della successione di frasi. Nel primo compito, BERT impara a prevedere le parole mancanti in una frase fornita, mentre nel secondo compito, il modello apprende a determinare se due frasi sono consecutive in un testo o sono state estratte casualmente da testi diversi.

Una volta completata la fase di addestramento, BERT può essere utilizzato per una vasta gamma di compiti NLP senza la necessità di ulteriori addestramenti specifici. Durante l'utilizzo, il modello riceve in input una sequenza di token, che possono essere

parole, frammenti di testo o segmenti di frasi. Ogni token viene rappresentato come un vettore di caratteristiche, derivato dal contesto bidirezionale fornito da BERT durante l'addestramento. Queste rappresentazioni vettoriali possono essere utilizzate per svariati compiti, come classificazione di testo, analisi del sentiment, risposta alle domande, traduzione automatica e molto altro ancora.

In sintesi, il modello BERT ha rivoluzionato il campo del NLP introducendo una comprensione più approfondita e contestualizzata del linguaggio naturale. La sua capacità di catturare il contesto bidirezionale delle parole ha portato a miglioramenti significativi nelle prestazioni dei sistemi NLP su una vasta gamma di compiti e applicazioni. BERT rimane un pilastro fondamentale nell'ambito della comprensione automatica del linguaggio umano, consentendo a macchine e sistemi di interagire e comprendere il linguaggio umano in modo più naturale e preciso.

Bibliografia

[1] Martina Rossini. Slither audited smart contracts dataset, 2022.

[2] Martina Rossini, Mirco Zichichi, and Stefano Ferretti. On the use of deep neural networks for security vulnerabilities detection in smart contracts. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 74–79, 2023.