LoRA 기반 경량화 CodeBERT 모델을 활용한 CWE 코드 취약점 탐지

김건민[1], 이은성[1], 정민수[1], 오다영[2], 이서준[2], 장현지[2], 최광훈[3], 김경백[4]
[1]전남대학교 소프트웨어공학과 학부생
[2]전남대학교 인공지능학부 학부생
[3]전남대학교 정보보안융합학과 교수
[4]전남대학교 인공지능융합학과 교수
204869@jnu.ac.kr, 200750@jnu.ac.kr, 205306@jnu.ac.kr,
215430@jnu.ac.kr, tjwns1300@jnu.ac.kr, gka1225@jnu.ac.kr,
kwanghoon.choi@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

LightWeight CWE Vulnerability Detection Using LoRA-Based CodeBERT

Geonmin Kim[1], Eunseong Lee[1], Minsoo Jeung[1], Dayoung Oh[2], Seojun Lee[2],
Hyeonji Jang[2], Kwanghoon Choi[3], Kyungbaek Kim[4]
[1]Dept. of Software Engineering, Chonnam National University
[2]Dept. of Artificial Intelligence, Chonnam National University
[3]Dept. of Information Security Convergence, Chonnam National University
[4]Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

딥러닝 기반의 코드 취약점 탐지 모델은 최근 소프트웨어 보안 자동화의 핵심으로 부상하였다. 그러나 이는 사전학습 언어 모델을 기반으로 하여 방대한 파라미터 수와 미세 조정 과정에서 자원 소모가 많다는 한계를 지닌다. 이에 본 연구는 CodeBERT와 LoRA를 결합하여 경량화된 멀티라벨 취약점 탐지 모델을 제안한다. 여러 CWE 유형에 대하여 LoRA가 제공하는 학습 효율성과 일반화 가능성을 실증하여 학습 파라미터 수와 자원 소모를 대폭 감소시키는 경량화 기법을 제안한다.

1. 서론

최근 소프트웨어 개발 환경의 진화에 따라 소스 코드의 규모와 복잡성이 증가하여 다양한 형태의 코드 수준의 보안 취약점이 양산되고 있다. 특히 하나의 코드 스니펫에 대하여 복수의 CWE(Common Weakness Enumeration)[1]이 병존하는 상황이 빈번하게 발생하며, 이는 보안 분석의 정확도를 저하시킬 우려가 있다. 이에 최근에는 사전학습 언어모델을 활용한 소스 코드 취약점 탐지 연구가 활발히 이루어지고 있으며, 트랜스포머 기반의 CodeBERT[2]는 코드와 자연어를 이중적으로 학습할 수 있는 모델로 주목받고 있다. CodeBERT는 코드의 문법적 특징과 구조를 효과적으로 포착할 수 있도록 설계된 아키텍처이다. 그러나 CodeBERT는 전체 파라미터를 재학습하는 Full Fine-Tuning 방식이 일반적이며, 이는 대규모 연산 자원 요구, 과적합 위험 등의 단점을 수반한다. 이에 본 연구는 자원 및 파라미터의 효율성 향상을 위한 대표적 기법 중 하나인 LoRA(Low-Rank Adaptation)[3]와 CodeBERT를 통합한 경량화된 모델을 제안하고, 그 효과를 실험을 통해 입증한다.

2. 이론적 배경 및 관련 연구

2.1 CWE 기반 보안 취약점 분류 체계

CWE는 MITRE에서 관리하는 공개 보안 취약점 유형 분류 체계로 소프트웨어 개발 단계에서 코드 수준에서 발생하는 논리적, 구조적 결함을 식별하고 이를 체계적으로 분류한다. CWE는 코드의 포맷 문자열 사용 방식, 자원 해제 시점 등 다양한 결함 유형을 수백 가지 항목으로 정의한다. 특히, CWE는 취약점 유형 간의 상호 연관성과 중첩 가능성도 반영한다. 실제 하나의 취약한 코드 스니펫에 여러 CWE 유형이 병존하는 현상이 발생하며, 이는 멀티라벨 분류 방식의 필요성을 직접적으로 시사한다.

2.2 LoRA 기반 파라미터 효율화

LoRA는 파라미터 효율화 미세조정(Parameter Efficient Fine-Tuning, PEFT)[4] 기법 중 하나로,

대규모 언어 모델을 더 적은 파라미터를 사용하여 미세조정할 수 있도록 설계되었다. LoRA는 선형 변환 행렬에 직접 업데이트를 가하지 않고, 분해된 두 개의 저차원 행렬을 도입하는 선형 계층의 저차원 근사를 이용한다. 이를 통해 기존 모델의 구조를 유지한 채 변화량만을 효율적으로 학습하는 방식으로 설계되어 모델 경량화에서 높은 이점을 갖는다. 본 연구에서 CodeBERT의 학습 파라미터 수는 125,541,902개 였으나, LoRA 적용으로 학습 파라미터 수가 890,887개로 감소하였다.

3. LoRA 기반 CWE 취약점 탐지 실험 방법

본 연구에서는 LoRA와 CodeBERT를 결합한 멀티라벨 취약점 탐지 모델의 성능을 평가하기 위하여 SARD(Software Assurance Reference Dataset)의 C언어 기반 취약 코드 데이터셋을 활용하였다. 이 중 6개의 CWE 유형(CWE-78, 134, 190, 400, 416)을 중심으로 데이터셋을 정제하고, 샘플 수가 적은 CWE 유형을 보강하기 위하여 각기 다른 가중치를 적용하여 데이터 불균형을 보정하였다. 또한, 기존의 CodeBERT 모델과 제안 모델의 성능을 비교한다.

4. 실험 결과

LoRA 모델은 query, value모듈에 대해 r=16, α=32의 설정으로 경량화를 수행하였으며, fp16, weight_decay=0.001, warmup_ratio=0.1의 학습 전략을 통해 안정적인 학습을 도모하였다.

<표 1> CodeBERT와 제안 모델 CWE 탐지 성능 비교

Model	CWE	Accuracy	Precision	Recall	F1
CodeBERT	78	1.000	1.000	1.000	1.000
	134	0.9537	0.7126	0.9917	0.8633
	190	0.9735	0.8685	1.000	0.9296
	400	0.9910	0.7595	1.000	0.8633
	416	0.9976	0.8000	1.000	0.8889
	476	0.9990	0.8824	1.000	0.9375
Proposed Model	78	0.9323	0.9329	0.9323	0.9314
	134	0.8497	0.8492	0.8497	0.8338
	190	0.9344	0.9336	0.9344	0.9338
	400	0.8753	0.8717	0.8753	0.8725
	416	0.8243	0.8035	0.8243	0.7929
	476	0.8711	0.8814	0.8712	0.8745

CodeBERT의 Full Fine-Tuning 모델은 전반적으로 높은 성능을 보이며 특히 CWE-78과 같이 상대적으로 명확한 패턴을 가진 취약점에서는 완벽에 가까운 성능을 기록하였다. LoRA 모델은 전체 파라미터 수의 약 99.29%의 감소율을 달성하면서도 약 0.87-0.93 수준의 평균 정확도를 보이며 대부분의 CWE 유형에 대해 유의미한 성능을 확보하였다.

5. 결론 및 향후 연구

본 연구는 소프트웨어 코드 내 다중 CWE 유형을 동시에 탐지하는 멀티라벨 취약점 탐지에 대하여 자원 효율성과 일반화 성능을 고려한 새로운 접근법으로 CodeBERT모델의 핵심 레이어에 LoRA기법을 적용 후 주요 CWE유형에 대해 실험을 수행하였다.

실험 결과, LoRA를 적용한 모델은 전체 파라미터 수를 획기적으로 줄이면서도 기존의 CodeBERT 모델과 유사한 탐지 성능을 확보할 수 있었다. 이를 통해 경량화된 제안 모델이 CWE 위협 탐지를 효과적으로 기능함을 실증적으로 확인하였다.

향후 연구에서는 GNN(Graph Neural Network)이나 AST(Abstract Syntax Tree) 정보를 활용하는 모델과 결합한 하이브리드 구조를 연구하고, LoRA의 장점을 살려 리소스가 제한된 On-Device 환경에서의 실시간 탐지 가능성을 실험할 수 있다. 이를 통해 차세대 코드 보안 자동화 시스템의 경량화에 실질적인 기여를 할 수 있을 것으로 기대된다.

Acknowledgement

본 연구는 한국인터넷진흥원(KISA)-정보보안 특성화대학 지원사업의 지원을 받아 수행된 연구임 (34%). 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구임(IITP-2025-RS-2022-00156287, 33%). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음 (IITP-2023-RS-2023-00256629, 33%)

참고문헌

[1] M. Howard, "Improving Software Security by Eliminating the CWE Top 25 Vulnerabilities," in IEEE Security & Privacy, vol. 7, no. 3, pp. 68-71, May-June 2009

[2] K. Zhao, S. Duan, G. Qiu, J. Zhai, M. Li and L. Liu, "Python Source Code Vulnerability Detection Based on CodeBERT Language Model," 2024 7th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI), Guangzhou, China, 2024, pp. 1-6

[3] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models."*ICLR*1.2 (2022)

[4] D. K. Gajulamandyam et al., "Domain Specific Finetuning of LLMs Using PEFT Techniques," 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2025, pp. 00484-00490