P R O J E C T D O C U M E N T

Test Result Document

Project Name	바이너리 프로그램에서 제어 구조를 식별하는 도구 개발

14 조

202002514 안상준
202202602 손예진
202202487 박혜연

지도교수: 조은선 교수님 (서명)

Table of Contents

1. Introduction

1.1. Objective

이 문서는 가상화 난독화의 대표적인 구조인 loop-switch 패턴을 LLM 기반 모델이 얼마나 효과적으로 식별할 수 있는지 확인하기 위해 학습 및 평가를 수행한 결과를 포함한다. 또한, 실험 결과에 대해 한계점 및 개선 방향 등을 정리한다.

2. Experiment Result Report

1. 서론

1.1 실험 개요

가상화 난독화의 핵심 구조인 loop-switch 구조를 LLM이 얼마나 잘 식별할 수 있는지를 실험을 통해 확인하였다. 특히 악성코드에서도 난독화 된 구조를 식별할 수 있다는 것을 확인하고자 한다.
모델 학습에 사용한 데이터는 직접 구축하였다. 직접 생성한 C코드 21,000개, Github API를 이용하여 658개, AtCoder와 같은 온라인 저지 사이트에서 크롤링 하여 127,280개의 C코드를 수집하였다.
Tigress를 사용하여 코드 난독화를 진행했다. 난독화 기법은 Flattening, Opaque Predicate, Virtualize를 적용하였다. 생성한 데이터 중 난독화를 적용하지 않은 코드와 각 난독화 기법을 적용한 코드를 랜덤하게 추출하여 학습 및 테스트에 사용하였다. 각 난독화 기법에 대해 원본 C코드에 switch문이 포함되었는지 여부에 따라 8개의 라벨로 나누었다. 각 라벨에 대해 동일한 비율로 나누어 학습에는 40,000개, 테스트에는 10,000개의 데이터를 사용했다. 학습 데이터 중 32,000개를 학습에 사용하고, 8,000개를 검증에 사용했다.
모델은 분류 문제에 좋은 성능을 보이는 BERT를 선택했다. 모델은 허깅페이스에서 제공하는 모델을 사용하여 미세조정했다. 이 모델은 바이너리 코드를 입력으로 받아 원본 코드의 switch문 포함 여부와 각 난독화 기법에 대해 총 8개의 클래스로 분류한다. 실험은 NVIDIA RTX 3060 에서 진행했다.

1.2 실험 방법

BERT와 TinyLlama 모델을 fine-tuning 하여 각 모델의 정확도를 평가했다. TinyLlama 모델은 Peft를 적용해 fine-tuning 하였다. 원본 C코드에서 switch문의 포함 여부와 난독화 적용 여부, 난독화가 적용되었다면 어떤 기법이 적용되었는지에 대하여 총 8개의 라벨이 있다. 테스트 데이터 10,000개는 8개의 라벨이 동일한 비율로 구성되어있다. 모델의 예측에 대해 (정확도) = (정확하게 예측한 데이터 수) / (전체 데이터 수) 로 성능을 평가하였다.

2. 테스트 결과 상세

2.1 테스트 결과 개요

	BERT	BERT (Fine-tuning)	TinyLlama	TinyLlama (Fine-tuning with Peft)
정확도 (%)	12.51	98.9	13.04	99.02

2.2 테스트 결과 상세 분석

Switch문 데이터는 직접 생성한 데이터의 비율이 높고, tigress만을 사용해 난독화를 적용했기 때문에 모델이 특정한 패턴에 과적합되었을 가능성이 있다. 실제 악성 코드처럼 예측 불가능한 난독화 방식이나 조합이 등장할 경우에는 성능이 저하될 가능성이 있다.

2.3 실험 결과의 한계와 위협 요인
Flattening, Opaque Predicate, Virtualize 각각을 적용한 데이터만을 학습에 사용했기 때문에 새로운 난독화 기법이 추가된다면 모델의 재학습이 필요하고, 여러 난독화가 적용된 경우에 대해서도 하나의 기법만을 예측한다는 문제가 있다.
3. 결론
현재 모델을 악성 코드 분석에 사용한다면 필요한 시간이 상당히 줄어들 것이다. 그러나, 현재는 난독화 기법이 한정적이기 때문에 난독화 기법을 추가하고 여러 난독화를 동시에 적용한 경우에 대해서도 학습 및 테스트를 진행한다면, 더 범용성 높은 도구가 될것이라고 생각한다. 또한, 난독화 기법과 동시에 어느 부분에 난독화가 적용되었는지를 나타내는 특징을 학습 데이터에 추가한다면, 해석력이 높아져 분석에 더 많은 도움이 될 것이다.

3. AI 도구 활용 정보

사용 도구	*해당사항 없음*
사용 목적	
프롬프트	
반영 위치	
수작업 수정	