Journal of the Korea Academia-Industrial cooperation Society
Vol. 24, No. 3 pp. 24-32, 2023

https://doi.org/10.5762/KAIS.2023.24.3.24
ISSN 1975-4701 / eISSN 2288-4688

다중 피처 딥러닝 모델을 이용한 전자기파 기반의 명령어 역어셈블러 구현

홍성우[1], 이재욱[2], 이현로[1], 하재철[1*]
[1]호서대학교 일반대학원 정보보호학과
[2]건국대학교 일반대학원 인공지능학과

Implementation of Instruction Disassembler Based on Electromagnetic Traces Using Multiple Features Deep Learning Model

Seongwoo Hong[1], Jaewook Lee[2], Hyunro Lee[1], Jaecheol Ha[1*]
[1]Department of Information Security, Graduate School, Hoseo University
[2]Department of Artificial Intelligence, Graduate School, KonKuk University

요 약 인터넷과 연결된 IoT 장비들은 관리자의 감시를 벗어나 원격에 위치하고 있는 경우가 많아 악성 코드 주입 공격이나 코드 불법 복제의 위협에 노출되어 있다. 이러한 공격에 의한 침해 여부를 확인하기 위해 마이크로 프로세서 장치들에서 실행되는 명령어에 대한 역공학을 수행할 필요가 있다. 본 논문에서는 부채널 신호인 전자기파를 이용하여 Cortex-M4에서 사용하는 명령어를 분류하는 역어셈블러를 구현하였다. 특히, 전자기파 신호로부터 추출된 다중 피처를 이용하는 딥러닝 모델을 제안하였다. 구현된 역어셈블러는 명령어 그룹을 분류하는 경우에는 93.35%, 그리고 그룹 내 명령어를 분류하는 경우에는 85.38%의 정확도를 보여 기존 단일 피처 기반 역어셈블러와 비교하여 높은 정확도로 분류될 수 있음을 확인하였다.

Abstract Many internet of things (IoT) devices connected to the internet are often located remotely away from the management boundary of administrators, so they are exposed to threats of malicious code injection attacks or illegal code copying. In order to confirm the infringement caused by these attacks, it is necessary to perform reverse engineering on instructions executed in microprocessor devices. In this study, we implemented a disassembler that classifies instructions used in Cortex-M4 using electromagnetic traces, which are side channel signals of target devices. We propose a deep learning model that uses multiple features extracted from electromagnetic signals. The proposed disassembler showed accuracy of 93.35% when classifying instruction groups and 85.38% when classifying instructions within a group. This confirmed that it can classify all instructions with high accuracy compared to an existing single-feature-based disassembler.

Keywords : Hardware Security, Reverse Engineering, Microprocessor, Electromagnetic Side-Channel Information, Disassembler

본 논문은 2022년도 호서대학교의 재원으로 학술연구비 지원을 받아 수행된 연구임.(202202400001)
*Corresponding Author : Jaecheol Ha(Hoseo Univ.)
email: jcha@hoseo.edu
Received January 31, 2023 Revised February 24, 2023
Accepted March 3, 2023 Published March 31, 2023

1. 서론

4차 산업 혁명과 함께 대량의 IoT 장비들이 등장하고 있다. IoT 장비들은 인터넷을 통해서 연결되고 데이터를 송·수신한다. 이러한 초연결 특성과 함께 관리자의 감시를 벗어나 운영되는 경우가 많아 해커들의 쉬운 공격 표적이 되고 있다. 대표적인 공격으로는 펌웨어에 대한 악성 코드 주입과 소프트웨어 불법 복제가 있다. 따라서 이러한 공격으로부터 시스템을 보호하기 위해서는 시스템의 실행 코드를 분석하여 악의적인 코드나 복제된 코드가 사용되는지를 탐지하는 것이 중요하다. 최근에는 특정 IoT 장비에 대한 실행 코드를 역으로 복구하기 위해 시스템 구현 시 누설되는 소비 전력이나 전자기파와 같은 부채널 정보를 이용하기도 한다[1].

부채널 정보란 시스템이 동작하면서 발생하는 모든 부가적인 정보를 의미한다. 부채널 정보로는 전력 파형, 전자기파, 소리, 시간 등이 대표적이다. 이러한 부채널 정보를 통해서 해당 시스템의 내부 정보를 자세히 분석해 볼 수도 있으며, 더 나아가 분석 자료를 가공해 디바이스에 내장된 암호화 시스템의 비밀 키를 찾아내는 것과 같은 악성 공격 행위를 시도할 수 있다.

부채널 정보를 이용한 분석 대상은 암호용 비밀 키뿐만 아니라 IoT 임베디드 시스템에서 실행되는 명령어도 해당한다. 부채널 신호를 이용하여 명령어를 역분석하는 도구나 툴을 하드웨어에 대한 역어셈블러(disassembler)라고 하는데, 이러한 역어셈블 과정은 해당 시스템이 악성 코드 주입에 의해 무결성 침해를 받는 지 여부를 확인하는 데 활용할 수 있다[2,3]. 또한, 역어셈블러는 의심스러운 디바이스에서 어떤 개발자의 정당한 프로그램 코드를 복사해서 사용하는 지를 확인하는 데에도 활용할 수 있다.

그 동안 역어셈블러에 관한 연구는 주로 전력 파형이나 전자기파를 사용하여 실행 코드를 복구하는 데 집중되어 왔다. 역어셈블러가 좋은 성능을 내기 위해서는 부채널 정보 파형에 대한 노이즈를 감쇄시키고 딥러닝에 필요한 피처 추출이나 차원 축소 기법에 관한 연구가 필수적이다[4,5]. 대표적인 차원 축소 기법으로는 PCA(Principal Components Analysis), LDA(Linear Discriminant Analysis), CWT(Continuous Wavelet Transform) 등이 사용되고 있다[6-8]. 특히, 딥러닝을 이용한 명령어 역어셈블러에서는 분류기의 성능을 높이기 위한 신호 파형의 특성을 나타내는 피처(feature)를 정확히 추출하는 것이 매우 중요하다.

부채널 기반 역어셈블러에 대한 연구로 M. Eisenbarth 등은 PIC16F687 마이크로 컨트롤러에서 실행되는 프로그램의 전력 소비 신호에 대한 템플릿 공격을 이용하여 명령어 분류기를 구성하였다[2]. 이들은 은닉 마르코프 모델 (Hidden Markov Models, HMM)과 같은 사전 통계 모델을 적용하여 35개의 테스트 명령에서 70.1%, 실제 코드에서 50.8%의 명령어 인식률을 달성하였다. D. Strobel 등은 다중 전자기파 채널 (안테나)을 사용하는 PIC16F687에서 명령어 수준의 부채널 기반 역어셈블러를 구현하였다. 이들은 테스트 코드에서 96.24%, 실제 코드에서 87.69%의 정확도로 명령어를 인식하였다[6]. 또한, J. Park 등은 ATMega328P가 탑재된 타겟 보드에서 CWT나 PCA)와 같은 고급 노이즈 감소 전처리 기술과 새로운 분류 모델을 이용해 명령 코드 112개와 레지스터 64개를 99.0%의 정확도로 분류하는 데 성공하였다[3].

한편, 실행 명령어에 대한 역어셈블러를 구현할 경우 전체 명령어 중 하나를 바로 분류해 내는 일차 분류 방식도 있지만, 최근에는 이차 분류 방식에 관한 연구가 주목받고 있다[9]. 이차 분류 방식이란 먼저 명령어를 그룹 단위로 묶고 어떤 명령어의 그룹을 찾은 다음 그 그룹 내에서 목표로 하는 명령어를 한 번 더 분류해 내는 방식이다. 따라서 이차 분류를 사용하는 역어셈블러는 그룹 분류기와 명령어 분류기 두 개가 구현되어야 한다.

본 논문에서는 32비트 프로세서 Cortex-M4에서 사용하는 실행 명령어를 역공학 기법을 이용하여 복구하는 역어셈블러를 구현하고자 한다. 이를 위해서는 우선 명령어 학습용 부채널 신호를 수집한 데이터 셋을 구축하는 것이 중요한데 마이크로 프로세서가 탑재된 디바이스에서 실제로 구축하였다. 논문에서 사용하는 데이터 셋은 Cortex-M4에서 각 명령어가 동작하는 시간 동안 누설되는 전자기파를 측정하여 구축하였다.

딥러닝 기반의 명령어 역어셈블러는 신호에서 명령어의 특성을 나타내는 피처를 추출하는 과정이 필요한데 본 논문에서는 하나의 피처를 사용하는 것이 아니라 여러 개의 피처를 동시에 사용할 수 있는 다중 피처 모델을 제안하고 실제로 구현하였다. 제안하는 다중 피처 모델은 단일 피처 모델에 비해 다양한 특징점을 추출하는 것을 조합한 기법으로서 기존 단일 피처 기반 역어셈블러와 비교하여 보다 높은 정확도로 명령어를 분류할 수 있음을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 명령어 역어셈블러의 개요와 배경을 살펴보고 3장에서는 명령어 학습

을 위한 데이터 셋 구성과 파형 수집에 대하여 설명한다. 4장에서는 명령어 복구를 위한 다중 피처 딥러닝 모델을 제안한다. 5장에서는 실험을 통해 그룹 분류기와 명령어 분류기의 성능을 평가하고 기존 연구 결과와 비교 분석한다. 마지막으로는 결과를 정리하고 결론을 맺는다.

2. 부채널 정보를 이용한 역어셈블러

2.1 역어셈블러 명령어

본 논문에서 구현할 역어셈블러는 32비트 프로세서인 Cortex-M4에서 사용하는 명령어를 대상으로 한다. 따라서 Cortex-M4가 탑재된 STM323F303 보드에서 각 명령어를 실행하는 동안 발생하는 부채널 신호인 전자기파를 학습용 데이터 셋으로 이용한다.

대상 프로세서 Cortex-M4는 Cortex-M 프로세서의 기본 모델인 Cortex-M3 프로세서의 모든 기능을 제공하며 추가적으로 디지털 신호 처리(Digital Signal Processing) 명령어를 지원한다. 또한, 옵션 요소로 부동 소수점 장치(Floating Point Unit)를 사용할 수 있다. 이 프로세서는 1.25 DMIPS/MHz의 처리 속도를 가지며 13개의 범용 레지스터, 스택 포인터, 링크 레지스터, 프로그램 카운터 그리고 5개의 특수 레지스터를 가지고 있다.

Cortex-M4 프로세서는 매뉴얼 상 114개의 프로세서 명령어(피연산자에 따른 중복 명령어 포함)와 88개의 DSP 명령어를 가진다. J. Geest 등은[10] 이 프로세서 명령어 중에서 가장 많이 사용하는 17개를 선정하였고 이를 분류하는 역어셈블러를 구현한 바 있다. 그리고 이 명령어는 사용하는 용도에 따라 Table 1과 같이 5개의 그룹으로 나누었다. 그리고 한 번에 명령어를 바로 구분해 내는 1차 분류가 아니라, 먼저 그룹을 분류하고 그 그룹 내에서 명령어를 분류하도록 하는 2차 명령어 복구 역어셈블러를 구현하였다.

Table 1. Assembly instructions according to group classification

Group	Instructions
Group 1(ALU)	ADD, AND, CMP, EOR, MOV, ORR, SUB
Group 2(SHIFTS)	LSL, LSR, ROR
Group 3(LOADS)	LDR, LDRB, LDRH
Group 4(STORES)	STR, STRB, STRH
Group 5 (Multiplications)	MUL

2.2 명령어 수행 단계

분석 대상인 Cortex-M4는 32비트 마이크로 프로세서로서 메인 클록 5MHz 상에서 동작한다. 이 프로세서의 중요한 특징은 동작 속도를 높이기 위해 3단계 파이프 라인 구조로 동작한다는 것이다. 즉, 하나의 명령어는 패치(Fetch), 해독(Decode), 실행(Execute)이라는 3단계로 나누어 수행되며 이를 나타낸 것이 Fig. 1이다. 여기서 주목할 점은 하나의 명령어가 수행되는 3단계 파이프라인 구조 전체에 대한 전자기파를 측정해야 한다는 점이다. 또한, 하나의 명령어를 기준으로 볼 때 각 단계별로 2가지의 다른 명령어가 병렬로 동작하는 형태이므로 발생하는 전자기파는 이전 단계와 이후 단계의 명령어가 동시에 발생시키는 부채널 신호가 된다. 따라서 하나의 명령어에 대한 전자기파 데이터 셋은 전후 명령어 및 오퍼랜드(operand)에 따라 수백 개의 경우가 발생하게 된다.

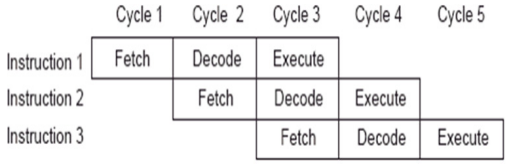


Fig. 1. Three stage pipeline of Cortex-M4

3. 역어셈블러 학습 데이터 파형 수집

본 논문에서 구현하고자 하는 딥러닝 기반의 역어셈블러를 개발하는 절차를 각 단계별로 요약하면 다음과 같다.

① 마이크로 프로세서 명령어 분석 및 분류
② 명령어 빈도수 분석을 통한 시퀀스 구성
③ 명령어 템플릿 구성 및 디바이스로 이식
④ 각 명령어 구동 및 전자기 파형 측정
⑤ 명령어 데이터 셋에 대한 사전 신호 처리
⑥ 딥러닝 모델 적용 및 학습
⑦ 딥러닝 네트워크 성능평가 및 명령어 복구

다음 Fig. 2는 위의 역어셈블러를 개발하는 전체적인 동작 과정을 도시한 것이다. 분석 대상이 되는 디바이스(target device)에서 사용되는 실제 명령어를 복구하기 위해서는 딥러닝 가속기에 대한 학습이 이루어져야 하며 이를 위해 학습 및 검증용 데이터 셋들이 필요하다.

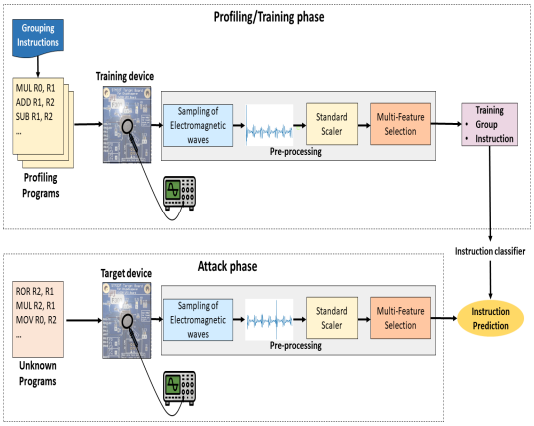


Fig. 2. Overview of deep learning-based disassembler

따라서 훈련용 디바이스(training device)에서 명령에 학습에 필요한 부채널 신호인 전자기파를 측정하여 데이터 셋을 준비한다. 준비된 데이터 셋을 이용하여 명령어가 속한 그룹을 분류하고 그 그룹 내에서 명령어를 복구하는 학습을 진행한다. 그리고 이 구현된 딥러닝 기반의 역어셈블러를 이용하여 실제 공격 목표가 되는 디바이스에서 수행되는 명령어를 복구하게 된다.

3.1 데이터 셋 구축 환경

딥러닝 기반의 역어셈블러를 구축하기 위해서는 명령어 학습을 위한 데이터 셋을 구축하는 것이 중요하다. 여기서 구축하는 데이터 셋은 하나의 명령어가 실행될 때 발생하는 전자기파를 측정한 학습용 데이터들을 말한다. 논문에서는 전자기파 신호 측정 및 분석에 필요한 템플릿 구현과 측정 신호 동기화를 위해서 NewAE Technology 사의 ChipWhisperer 플랫폼[11]을 사용한다. 전자기파 수집을 위해서는 Langer사의 RF-B 3-2 프로브를 사용하며, Lecroy HDO4024A 오실로스코프를 사용하여 각 명령어에 대한 전자기파를 측정한다.

이 오실로스코프는 프로세서의 한 클록 당 2,000샘플의 전자기파를 측정할 수 있는데 Cortex-M4는 3단계 파이프라인 구조를 갖는 MCU이므로 여유 샘플 200개를 포함하여 약 6,200개의 샘플을 역분석 대상 명령어의 파형으로 사용하였다. 다음 Fig. 3은 전자기파 프루브를 통해 학습용 데이터 셋을 측정하는 실험 장치를 나타낸 것이다.

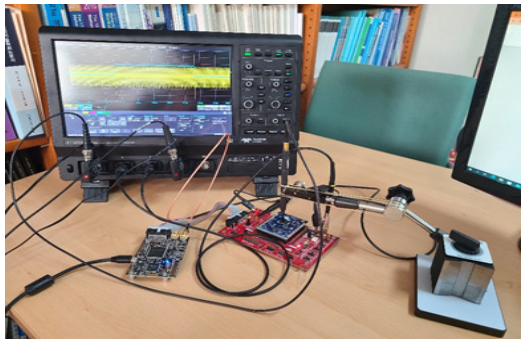


Fig. 3. Electromagnetic wave measurement using RF-B 3-2 probe

3.2 학습용 데이터 셋

Cortex-M4 프로세서 하나의 명령어가 3단계 파이프라인으로 동작한다는 점을 고려하면 학습용 데이터 셋은 다양한 형태의 명령어 시퀀스를 가정하여 수집해야 한다. 그러나 모든 명령어 시퀀스를 수집하고 학습하는 것은 매우 비효율적이며, 실제 많이 등장하는 명령어 시퀀스에 대해서 모델이 정상적으로 학습하지 못할 수 있다. 이러한 점을 고려하여 본 논문에서는 명령어의 사용 빈도를 분석하여 빈도가 높은 명령어 템플릿을 생성하고, 이 템플릿 명령어를 실행 시 발생하는 전자기파를 측정하여 데이터 셋을 구성하였다. 다음 Fig. 4는 빈도 기반 명령어 시퀀스를 생성하는 과정을 나타낸 것이다.

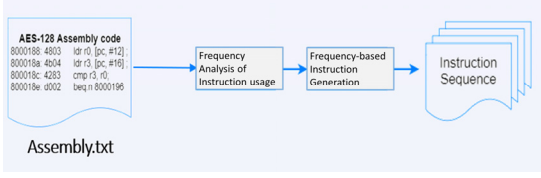


Fig. 4. Generation of instruction sequence based on frequency of instruction usage

본 논문에서는 명령어에 대한 이차 분석 구조를 고려하여 데이터 셋을 2개로 구분하여 수집한다. 데이터 셋 A는 그룹 분류기를 위한 것이며, 데이터 셋 B는 그룹 1(ALU)에 속한 명령어를 구분하기 위한 데이터 셋이다. 데이터 셋 A는 5개의 그룹을 분류하는 것을 목표로 전자기파를 수집한다. 각 그룹당 2,500개의 명령어 템플릿 프로그램을 수행하고 전자기파를 측정한다. 그리고 한 번에 측정되는 파형은 6,200개의 샘플을 갖게 되므로 최종적인 데이터 셋의 크기는 (5, 2,500, 6,200)이다.

데이터 셋 B는 그룹 1(ALU)에 속한 명령어를 대상으로 전자기파를 수집하여 데이터 셋을 구축한다. 그룹 1에는 총 7개의 명령어가 존재하는데 명령어 하나당 2,500개의 템플릿 프로그램을 구성해서 측정한다. 따라서 데이터 셋 B의 최종적인 크기는 (7, 2500, 6200)이다.

4. 다중 피처 딥러닝 기반 역어셈블러

상기한 바와 같이 마이크로 프로세서에 대한 역어셈블러 구현을 위해 전력 파형과 전자기파와 같은 부채널 신호를 이용한다. 즉, 디바이스에 대한 전력 소비 지점에 직접 접촉이 가능한 경우에는 전력 파형을 이용하지만 직접 프루빙이 불가능한 분석 환경하에서는 전자기파를 이용한다. 하지만 이러한 부채널 신호는 측정 장비와 실험 환경에 따라서 데이터 간의 편차가 심하다. 따라서 분석 환경에 따른 노이즈를 제거하고 효율적인 딥러닝 학습을 위한 신호 특성을 추출하기 위한 여러 신호처리 기법이 필요하다.

4.1 정규화 및 피처 추출

4.1.1 정규화

딥러닝 학습 모델에 효과적인 피처 추출 기법을 적용하기 위해서는 사전에 데이터에 대한 표준화 스케일링 작업이 필요하다. 이 과정은 학습에 사용된 전체 데이터에 대한 평균과 분산 값을 조절하는 것이다. 다음 Fig. 5는 표준화 스케일링이 적용되기 이전에 측정된 전자기파의 모습인데 피크(peak) 부분이 명확하게 구분되는 것을 확인할 수 있다. 그런데 이러한 피크 부분은 모델이 특정한 영역에 대해서만 학습하게 되어 과적합 현상이 발생하게 되고 이는 모델 성능을 떨어뜨리는 원인이 된다.

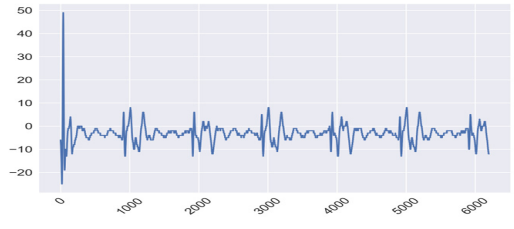


Fig. 5. Original EM trace

원본 전자기파에 대해서 표준화 스케일링을 적용한 파형을 나타낸 것이 Fig. 6이다. 해당하는 전자기파 모습을 통해서 원본 전자기파에서 보였던 피크 부분에 대한 영향도가 낮아진 것을 확인할 수 있다. 이를 통해서 모델이 특정 부분에 대해서만 학습하는 과적합 현상을 제거할 수 있다.

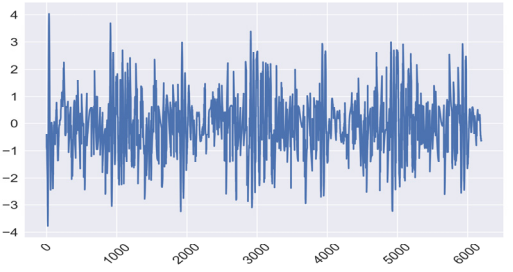
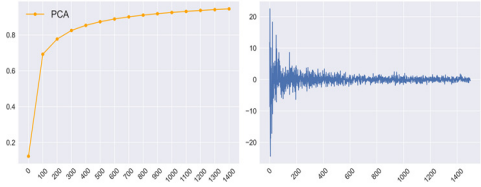


Fig. 6. EM trace applied with a standard scaling method

4.1.2 주성분 분석

머신 러닝 모델의 사용되는 신호에 대한 피처 추출 기법으로는 PCA와 LDA 등이 있다. 이 피처 추출 기법들은 데이터에 포함된 통계적인 특성들을 반영해서 해당하는 데이터에 대한 성분 분석을 통해 차원을 축소한다. PCA는 높은 차원으로 구성된 데이터들에 대해서 정보의 주된 성분은 보존하면서 낮은 차원으로 차원을 축소하는 기법을 말한다.

다음 Fig. 7의 (a)는 사용하는 주성분 수에 따른 누적 분산량 나타낸 것으로 차원 축소 크기를 결정하는 기준이 된다. 그림에서 보는 바와 같이 하나의 명령어는 6,200개의 샘플을 가지는데 이를 약 1,500개의 성분으로 차원을 축소해도 데이터를 충분히 표현할 수 있다는 의미가 된다. 그림의 (b)는 하나의 명령어 데이터 셋에 해당하는 전자기파 6,200개 정도의 샘플 차원을 1,500개의 PCA 차원으로 축소한 후의 전자기파를 표현한 것이다.



(a) Explained variance (b) Reconstructed trace by PCA

Fig. 7. Dimension reduction applied with PCA

4.1.3 연속 웨이블릿 변환

웨이블릿 변환이란 임의의 신호를 웨이블릿이라고 정의되는 함수들을 이용하여 신호를 분해하는 방법을 말한다. 웨이블릿이 적용된 데이터는 시간-주파수로 변환된다. 본 논문에서는 연속 웨이블릿 변환 CWT를 적용하는데 Mexican Hat 웨이블릿 함수를 사용하였으며 스케일 값을 50으로 설정하였다. 연속 웨이블릿 변환 기법이 적용된 후의 전자기파 모습을 나타낸 것이 Fig. 8이다. 하지만 CWT를 적용한 후 모든 주파수 영역을 사용하기에는 데이터 양이 너무 많아 특정 주파수 영역만 모델 입력으로 사용하기도 한다.

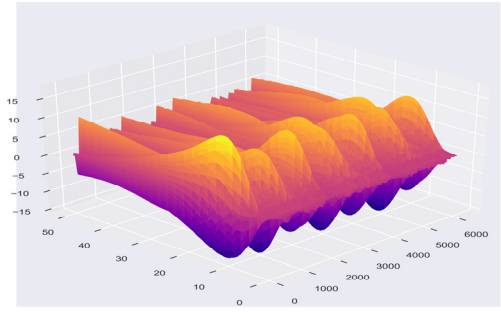


Fig. 8. EM traces applied with CWT

4.2 제안하는 다중 피처 딥러닝 모델

지금까지 마이크로 프로세서의 실행 명령어에 대한 역어셈블러 구현 연구들은 단일 피처만을 사용해서 모델을 학습하였다. 그러나 이러한 단일 피처 학습 방식에서는 명령어 수행 파형에 포함된 정보량이 부족할 수 있다. 따라서 본 논문에서는 이러한 단일 피처 방식의 문제점을 개선하고자 다중 피처 모델을 제안한다.

다중 피처 모델이란 단일 피처만을 모델의 입력으로 사용하지 않고 다양한 피처 값들을 구한 후 이를 딥러닝 학습 모델의 입력으로 사용하는 방식을 말한다. 이러한 모델은 다양한 피처 값들을 모델의 입력으로 받기 때문에 기존 정보량 부족 문제를 해결할 수 있어 실질적으로 명령어 복구 성능을 향상시킬 수 있다.

다음 Fig. 9는 본 논문에서 제안하는 다중 피처를 사용하는 딥러닝 기반의 명령어 역어셈블러의 전체 구조를 나타낸 것이다. 제안하는 다중 피처 모델에서는 사전에 처리된 신호를 Original 인코더, PCA 인코더 그리고 CWT 인코더에 입력하게 된다. 각 인코더는 특징 추출을 위해 배치 정규화(Batch Normalization) 및 드롭아웃(dropout) 층을 두고 있으며 CWT 인코더에는 컨볼루션(convolution) 층을 두고 있다. 그리고 각 인코더의 출력을 연접(Concatenation)한 후 다시 Original 인코더

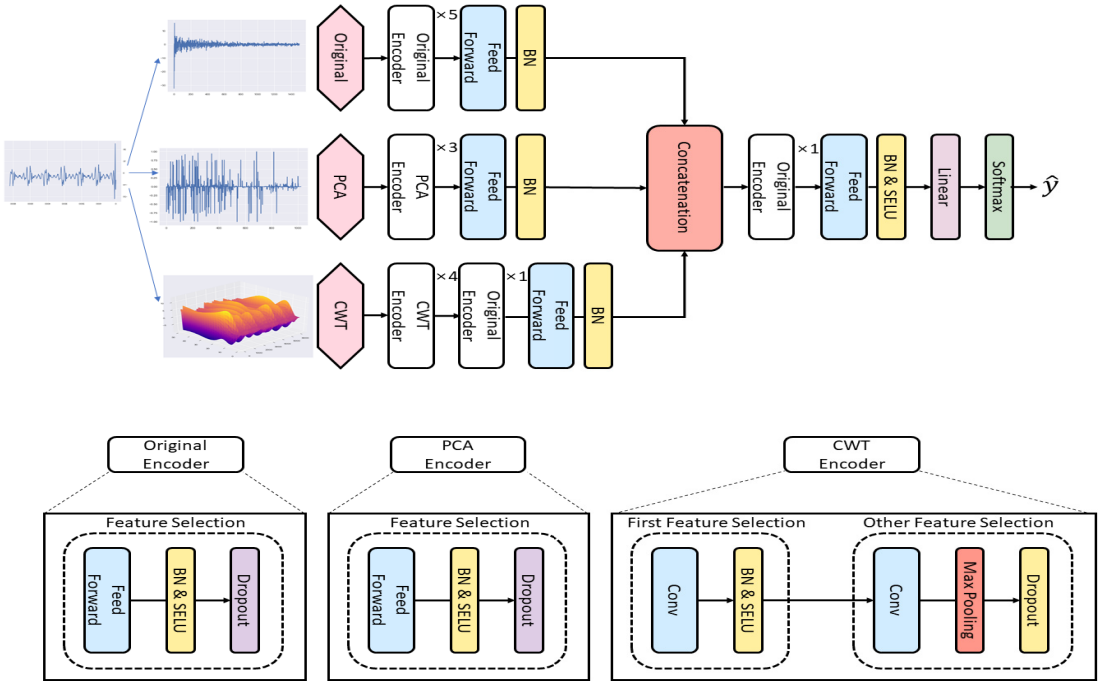


Fig. 9. Proposed multi-feature deep learning model

및 데이터의 표현 방식을 선형으로 바꾸어 표현하는 선형(Linear)층을 거치게 된다.

4.2.1 Original 인코더

Original 인코더는 원본 파형을 처리하기 위해서 MLP(Multi-Layer Perceptron) 모델[12]을 기반으로 구성하였다. 해당 인코더에서는 피처의 개수를 생각해서 다른 인코더들에 비해서 더욱 깊게 5개의 층을 쌓는다. 또한, 해당 인코더 층에서는 내부 공분산 문제를 해결하기 위해서 배치 정규화 층을 추가하였다[13]. 배치 정규화는 학습 과정에서 각 배치 단위별로 데이터가 다양한 분포를 가지더라도 각 배치 별로 평균과 분산을 이용해 정규화하는 과정이다. 각 인코더 층의 출력단에는 모든 가중치의 값이 반영되지 않게 드롭아웃 층이 존재한다. 드롭아웃 층은 모델이 과도하게 학습하는 과적합되는 현상을 효과적으로 방지할 수 있다.

4.2.2 PCA 인코더

PCA Encoder는 PCA가 적용된 데이터를 처리하기 위해서 존재하는 인코더이다. PCA 데이터는 기존의 6,200샘플 데이터(하나의 명령어에 대한 파형의 샘플 수)를 1,500샘플의 데이터로 축소된 형태로 존재한다. 따라서, 스케일링 인코더 층에서 사용된 MLP 모델을 그대로 사용하지 않고 더욱 축소된 형태의 MLP 모델을 사용한다. 내부적인 구조는 Original 인코더 층과 동일하지만 층의 깊이와 뉴런의 개수만 다르다.

4.2.3 CWT 인코더

마지막으로 CWT 과정을 거친 데이터는 앞선 데이터들과 다르게 확장된 형태의 데이터 구조를 갖는다. 특히, 시간-주파수 형태로 데이터가 변환되기 때문에 2차원의 상에서 데이터가 존재한다. 따라서, 해당하는 데이터에 효과적인 CNN 모델[14]을 인코더 층으로 사용한다. 내부적으로 배치 정규화 층과 드롭아웃 층은 다른 인코더 층과 동일하게 적용된다.

5. 실험 및 성능 평가

5.1 그룹 분류기

본 논문에서 제안하는 다중 피처를 사용한 딥러닝 기반 역어셈블러는 2차에 걸쳐 최종 명령어를 복구해 낸다. 즉, Cortex-M4에서 사용하는 명령어를 5개 그룹으로 나누고 그 그룹 내의 명령어를 분류해 낼 수 있음을 딥러닝 모델을 통해 확인하고자 한다. 그 후 해당 그룹 내의 명령어를 분류해 내는 2차 명령어 분류기를 동작시켜 명령어를 복구한다. 상기한 바와 같이 명령어 분석 대상 디바이스는 STM32F303을 사용하였고 내부에는 32비트 프로세서인 Cortex-M4가 내장되어 있다.

다음 Table 2에는 본 논문에서 제안한 딥러닝 기반 역어셈블러의 데이터 셋 A에 대한 그룹 분류 성능을 정확도(accuracy)로 나타낸 것이다. 해당 평가 지표에서는 단일 피처만을 사용하는 모델과 다중 피처를 사용하는 모델에 대한 실험을 실시하고 이차 분석 구조를 갖는 이전 연구 결과들와 비교 분석하였다.

표에서 O. Glamocanin 등에 의해 구현된 역어셈블

Table 2. Performance comparison of instruction disassembler model

Authors	Model	Target devices	Side channel	Feature selection	Accuracy (Group)	Accuracy (Instruction)	Accuracy (Hierarchical model)
O. Glamocanin et al.[15]	MLP	RISC-V	EM	PCA	90.69%	95.50%	86.60%
J. Geest et al. [10]	MLP	Cortex-M0	Power	No	86.40%	25.50%	22.03%
	CNN	Cortex-M0	Power	No	88.20%	25.20%	22.22%
Ours	MLP + Batch Norm	Cortex-M4	EM	Scaling	90.00%	85.38%	78.84%
	MLP + Batch Norm	Cortex-M4	EM	PCA	87.50%	82.26%	71.98%
	CNN + Batch Norm	Cortex-M4	EM	CWT	90.54%	83.65%	75.73%
	Original Enc.+ PCA Enc.+ CWT Enc.	Cortex-M4	EM	Scaling+ PCA+ CWT	93.35%	85.14%	79.48%

러는 전자기파를 이용하여 MLP 모델을 사용하였다. 비교적 모델의 정확도는 높지만, 해당 실험은 RISC-V를 실험 대상 프로세서를 사용한 것으로서 Cortex-M4를 대상으로 한 다른 논문의 역어셈블러와 상대적인 성능을 직접 비교하기는 어렵다. 따라서 사용하는 명령어가 거의 동일하고 명령어 그룹 분류나 실험 환경이 유사한 J. Geest 등의 연구와 주로 비교하였다.

상기 표에서 보는 바와 같이 명령어 그룹 분류 모델에서는 단일 피처를 사용했을 때보다 다중 피처를 사용할 때가 보다 높은 정확도를 달성하는 것을 확인할 수 있다. 단일 피처 모델에서 가장 정확도를 보인 모델은 CWT 데이터를 입력으로 받는 CNN 모델이었는데 해당 모델에서는 90.54%를 달성하였다. 본 논문에서 제안하는 다중 피처 모델을 사용한 그룹 분류기는 93.35%의 정확도를 보여 기존 연구 [10]보다 약 5%의 높은 성능을 나타내었다.

5.2 명령어 분류기

하나의 그룹 내에서 명령어를 분류하는 실험은 데이터 셋 B를 이용하여 수행하였다. 데이터 셋 B는 명령어 그룹 1(ALU)에 속한 명령어를 구분하기 위한 것이다. 위 Table 2는 제안하는 다중 피처 딥러닝 모델과 이전 실험 결과를 비교 분석한 것이다. 표에서 보는 바와 같이 제안 모델의 명령어 분류 정확도는 85.14%로 이전 MLP나 CNN 모델에 비해 높은 정확도를 나타내었다. 다만, 배치 정규화를 적용한 단일 피처 모델을 사용하는 MLP와는 비슷한 성능을 보였다.

그럼에도 불구하고 하나의 명령어를 복구하는 작업은 그룹 분류와 명령어 분류를 순차적으로 수행하게 되므로 1차 그룹 분류와 2차 명령어 분류를 곱한 결과를 보면, 다중 피처 딥러닝 모델이 단일 피처 모델에 비해 전체적으로 3~7% 정도의 정확도가 향상됨을 볼 수 있다.

6. 결론

임베디드 시스템에 침입한 불법적인 악성 코드를 탐지하거나 정상 코드에 대한 불법 복제 여부를 확인하기 위해서는 하드웨어 장치에 대한 역어셈블러가 필요하게 되었다. 최근 실행 명령어를 복구하는 역어셈블러를 구현하는 수단으로 디바이스에서 발생하는 부채널 신호를 사용하는 연구가 수행되어 왔다.

본 논문에서는 Cortex-M4에서 사용하는 명령어를 대상으로 다중 피처를 사용하는 딥러닝 기반의 역어셈블러를 구현하였다. 분석 대상 프로세서의 명령어 셋을 사용 용도에 따라 5개 그룹으로 분류할 수 있을 뿐만 아니라 특정 그룹내의 17개 명령어를 복구할 수 있는 역어셈블러를 딥러닝 네트워크로 구현하였다. 제안된 역어셈블러는 명령어 그룹을 분류하는 경우와 그룹 내 명령어를 분류하는 경우 모두 단일 피처 기반 역어셈블러와 비교하여 보다 높은 정확도로 명령어를 분류하여 복구할 수 있음을 실험적으로 검증하였다.

따라서 딥러닝 기술을 이용한 역어셈블러는 임베디드 장치에서 악성 코드 탐지 및 불법 복제 방지에 적극 활용할 수 있다. 향후에는 각 실행 명령어에 사용되는 오퍼랜드 값이나 레지스터 분류를 통해 마이크로 프로세서의 내부 동작 과정을 더 세밀하게 분석하는 연구가 필요하다.

References

[1] P. Kocher, J. Jaffe, and B. Jun, “Differential power analysis,” CRYPTO’99, LNCS 1666, pp. 388-397, 1999. DOI: https://doi.org/10.1007/3-540-48405-1_25

[2] M. T. Eisenbarth, C. Paar, and B. Weghenkel, "Building a Side Channel Based Disassembler," Transactions on computational science X: special issue on security in computing, part I, pp. 78-99, 2010. DOI: https://doi.org/10.1007/978-3-642-17499-5_4

[3] J. Park, F. Rahman, A. Vassilev, D. Forte, and M. Tehranipoor, “Leveraging Side-Channel Information for Disassembly and Security,” ACM Journal on Emerging Technologies in Computing Systems, Vol. 16, No. 1, pp 1-21, 2020. DOI: https://doi.org/10.1145/3359621

[4] G. Bak, T. Kim, H. Kim, and S. Hong, “Study on Singular Value Decomposition Signal Processing Techniques for Improving Side Channel Analysis,” Journal of the KIISC, Vol. 26, No. 6, pp. 1461-1470, 2016. DOI: https://doi.org/10.13089/JKIISC.2016.26.6.1461

[5] D. Kwon, S. Jin, S. Kim, and S. Hong, “Improving Non-Profiled Side-Channel Analysis Using Auto-Encoder Based Noise Reduction Preprocessing,” Journal of the KIISC, Vol. 29, No. 3, pp. 491-501, 2019. DOI: https://doi.org/10.13089/JKIISC.2019.29.3.491

[6] D. Strobel, F. Bache, D. Oswald, F. Schellenberg, and C. Paar, “SCANDALee: A side-channel-based disassembler using local electromagnetic emanations,” Proceedings of the Design, Automation & Test in Europe Conference & Exhibition(DATE’15), pp. 139-141, 2015. DOI: https://doi.org/10.7873/DATE.2015.0639

[7] S. Vafa, M. Masoumi, and A. Amini, “An Efficient Profiling Attack to Real Codes of PIC16F690 and ARM Cortex-M3,” IEEE Access, Vol. 8, pp. 222520-222532, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.3043395

[8] S. Wold, K. Esbensen and P. Geladi, "Principal component analysis," Chemometrics and intelligent laboratory systems, Vol. 2, No. 1-3, pp. 37-52, 1987. DOI: https://doi.org/10.1016/0169-7439(87)80084-9

[9] J. Park, X. Xu, Y. Jin, D. Forte, and M. Tehranipoor, "Power-based side channel instruction-level disassembler", Proc. of the 55th Annual Design Automation Conference(DAC), pp. 1-6, 2018. DOI: https://doi.org/10.1109/DAC.2018.8465848

[10] V. Geest, and I. Buhan, "A side-channel based disassembler for the ARM-Cortex M0," Proc. of the Applied Cryptography and Network Security Workshops(ACNS'22), pp. 183-199, 2022. DOI: https://doi.org/10.1007/978-3-031-16815-4_11

[11] ChipWhisperer® - NewAE Technology Inc., "chipwhisperer," Available at http://newae.com/tools/chipwhisperer/, 2017.

[12] M. Popescu, V. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," WSEAS Transactions on Circuits and Systems, Vol. 8, Issue 7, pp 579-588, July 2009 DOI: https://dl.acm.org/doi/10.5555/1639537.1639542

[13] S. Loffer, and S. Christian, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," Proc. of the 32nd International Conference on Machine Learning(PMLR'15), Vol. 37, pp. 448-456, 2015. DOI: https://doi.org/10.48550/arXiv.1502.03167

[14] A. Saad, T. Mohammed, and S. Al-Zawi. "Understanding of a convolutional neural network." 2017 international conference on engineering and technology (ICET). pp. 1-6, IEEE, 2017. DOI:https://doi.org/10.1109/ICEngTechnol.2017.8308186

[15] O. Glamocanin, R. Islambouli, and D. Mahmoud, "Machine Learning for Side-Channel Disassembly", Available at https://www.epfl.ch/labs/mlo/wp-content/uploads/2021/05/crpmlcourse-paper832.pdf, 2021.

홍 성 우(Seongwoo Hong) [준회원]

- 2023년 2월 : 호서대학교 컴퓨터공학부 (학사)
- 2023년 3월 ~ 현재 : 호서대학교 대학원 정보보호학과 석사과정

〈관심분야〉
부채널 공격, 암호학, 정보보호, 인공지능 보안

이 재 욱(Jaewook Lee) [준회원]

- 2023년 2월 : 호서대학교 컴퓨터공학부 (학사)
- 2023년 3월 ~ 현재 : 건국대학교 대학원 인공지능학과 석사과정

〈관심분야〉
인공지능 보안, 부채널 공격, 자연어 처리

이 현 로(Hyunro Lee) [준회원]

- 2023년 2월 : 호서대학교 컴퓨터공학부 (학사)
- 2023년 3월 ~ 현재 : 호서대학교 대학원 정보보호학과 석사과정

〈관심분야〉
자동차 보안, 부채널 공격, 양자내성 암호, 머신러닝

하 재 철(Jaecheol Ha) [종신회원]

- 1989년 2월 : 경북대학교 전자공학과 (학사)
- 1993년 8월 : 경북대학교 전자공학과 (석사)
- 1998년 2월 : 경북대학교 전자공학과 (박사)
- 1998년 3월 ~ 2007년 2월 : 나사렛대학교 정보통신학과 교수
- 2007년 3월 ~ 현재 : 호서대학교 컴퓨터공학부 교수
- 2013년 1월 ~ 현재 : 한국정보보호학회 수석부회장
- 2009년 1월 ~ 현재 : 한국산학기술학회 이사
- 2023년 1월 ~ 현재 : 국제차세대융합기술학회 부회장

〈관심분야〉
암호학, 네트워크 보안, 부채널 공격, 머신러닝