Journal of the Korea Academia-Industrial cooperation Society
Vol. 21, No. 5 pp. 14-21, 2020

https://doi.org/10.5762/KAIS.2020.21.5.14
ISSN 1975-4701 / eISSN 2288-4688

블록 암호 AES에 대한 CNN 기반의 전력 분석 공격

권홍필, 하재철*
호서대학교 정보보호학과

Power Analysis Attack of Block Cipher AES Based on Convolutional Neural Network

Hong-Pil Kwon, Jae-Cheol Ha*
Department of Information Security, Hoseo University

요 약 두 통신자간 정보를 전송함에 있어 기밀성 서비스를 제공하기 위해서는 하나의 대칭 비밀키를 이용하는 블록 데이터 암호화를 수행한다. 데이터 암호 시스템에 대한 전력 분석 공격은 데이터 암호를 위한 디바이스가 구동할 때 발생하는 소비 전력을 측정하여 해당 디바이스에 내장된 비밀키를 찾아내는 부채널 공격 방법 중 하나이다. 본 논문에서는 딥 러닝 기법인 CNN (Convolutional Neural Network) 알고리즘에 기반한 전력 분석 공격을 시도하여 비밀 정보를 복구하는 방법을 제안하였다. 특히, CNN 알고리즘이 이미지 분석에 적합한 기법인 점을 고려하여 1차원의 전력 분석 파형을 2차원 데이터로 이미지화하여 처리하는 RP(Recurrence Plots) 신호 처리 기법을 적용하였다. 제안한 CNN 공격 모델을 XMEGA128 실험 보드에 블록 암호인 AES-128 암호 알고리즘을 구현하여 공격을 수행한 결과, 측정한 전력 소비 파형을 전처리 과정없이 그대로 학습시킨 결과는 약 22.23%의 정확도로 비밀키를 복구해 냈지만, 전력 파형에 RP 기법을 적용했을 경우에는 약 97.93%의 정확도로 키를 찾아낼 수 있었음을 확인하였다.

Abstract In order to provide confidential services between two communicating parties, block data encryption using a symmetric secret key is applied. A power analysis attack on a cryptosystem is a side channel-analysis method that can extract a secret key by measuring the power consumption traces of the crypto device. In this paper, we propose an attack model that can recover the secret key using a power analysis attack based on a deep learning convolutional neural network (CNN) algorithm. Considering that the CNN algorithm is suitable for image analysis, we particularly adopt the recurrence plot (RP) signal processing method, which transforms the one-dimensional power trace into two-dimensional data. As a result of executing the proposed CNN attack model on an XMEGA128 experimental board that implemented the AES-128 encryption algorithm, we recovered the secret key with 22.23% accuracy using raw power consumption traces, and obtained 97.93% accuracy using power traces on which we applied the RP processing method.

Keywords : Side Channel Analysis, Power Analysis Attack, Deep Learning, Convolutional Neural Network, Block Cipher AES

1. 서론

부채널 분석(side channel analysis) 공격은 암호 알고리즘이 탑재된 디바이스로부터 소비되는 전력, 전자기파, 소리 등과 같은 부채널 정보를 얻어 내장된 비밀 정보를 알아내는 공격 방법이다[1]. 이러한 부채널 분석 공

*Corresponding Author : Jae-Cheol Ha(Hoseo Univ.)
email: jcha@hoseo.edu
Received January 22, 2020 Revised February 11, 2020
Accepted May 8, 2020 Published May 31, 2020

격에는 어떠한 신호 정보를 측정하여 공격을 진행하느냐에 따라 공격 방법이 다양하게 존재하며, 그 중에서 전력 분석 공격(power analysis attack)이 가장 대표적이다. 전력 분석 공격이란, 암호 알고리즘이 구현된 디바이스가 수행될 때 소비되는 전력 값을 측정하여 전력 파형을 얻고, 해당 전력 파형을 통해 이 디바이스 내부에 있는 비밀 정보를 알아내는 공격 방법이다[2].

이러한 전력 분석 공격은 실제 입력 데이터(평문, 비밀키)가 소비 전력에 영향을 준 유의미한 파형 구간(POI: Point Of Interest, 이하 POI)을 찾는 사전 처리 과정과 해당 POI의 파형 샘플 데이터를 분석하여 비밀키를 찾아내는 과정으로 구분할 수 있다. 하지만 이 과정에서 공격자는 비밀키가 사용되는 구간을 찾아야 하며 신호 분석의 정확도를 높이기 위해서는 많은 전력 소비 파형을 수집하여야 한다.

본 논문에서는 기존의 전력 분석 공격에 딥 러닝(deep learning) 기술의 일환인 합성곱 신경망(CNN: Convolutional Neural Network, 이하 CNN) 알고리즘을 적용한 새로운 전력 분석 공격 모델을 제안하였으며 이를 통해 정확하게 비밀키를 찾아낼 수 있음을 보이고자 한다[3].

딥 러닝 기술이란, 생물학적 뉴런 구조를 토대로 설계된 인공 신경망 모델에 특정 데이터들을 입력함으로써 데이터들이 갖는 특징에 대한 학습을 수행하는 기계 학습 기술이다[4]. CNN 알고리즘은 딥 러닝 기술의 하나로서 다층 퍼셉트론(MLP: Multi-Layer Perceptron, 이하 MLP) 알고리즘 구조에 컨볼루션 계층(convolution layer)과 풀링 계층(pooling layer)을 추가한 구조이며, 주로 이미지나 음성 데이터에 대한 학습에 사용되고 있다[5]. 이러한 CNN 알고리즘은 전력 신호의 POI를 찾는 사전 과정과 POI의 파형 샘플 데이터를 분석하는 과정을 자동화할 수 있어 전력 분석 공격에 유용하게 활용될 수 있다.

본 논문에서는 국제 표준 블록 암호 알고리즘인 AES(Advanced Encryption Algorithm) 시스템을 공격 대상으로 CNN 알고리즘을 적용하여 전력 분석 공격을 수행하였다[6]. 특히, 시계열 데이터에 대한 이미지화 기법인 RP(Recurrence Plots) 기법을 전력 파형에 적용할 것을 제안하여 모델 성능이 어느 정도 향상되는지 분석하였다[7]. 실제로 AES-128 블록 암호 알고리즘을 XMEGA128 칩을 사용한 실험용 보드에 구현한 후 여기에서 소비되는 전력 파형을 측정하고 분석함으로써 사용자의 비밀키를 높은 확률로 찾아낼 수 있음을 증명하였다.

2. 배경 지식

2.1 전력 분석 공격

전력 분석 공격은 암호 알고리즘을 구현한 보안 디바이스가 구동될 때 소비되는 전력을 측정하여 전력 파형을 얻고, 이를 분석하여 보안 디바이스 내부의 비밀 정보를 알아내는 공격 방법이다. 이러한 전력 분석 공격은 공격 방법에 따라 크게 논-프로파일링 공격(non-profiling attack)과 프로파일링 공격(profiling attack)으로 나누어진다.

논-프로파일링 공격은 프로파일을 생성하지 않고 공격 대상 디바이스에 다수의 입력으로 암호 시스템을 여러 번 실행하여 전력 파형들을 수집하고, 해당 전력 파형들을 통계적으로 분석하여 비밀 정보를 알아내는 공격 방법이다. 보통 이러한 논-프로파일링 공격은 처리되는 데이터에 따라 소비 전력 값이 상이하다는 이론을 바탕으로 공격이 이루어지며 대표적으로 해밍 웨이트(Hamming weight) 모델 이론이 있다. 해밍 웨이트 모델을 기반으로 하는 논-프로파일링 공격으로는 차분 전력 분석(DPA: Differential Power Analysis, 이하 DPA)와 상관 전력 분석(CPA: Correlation Power Analysis, 이하 CPA)가 있으며, 수집한 수많은 전력 파형들을 토대로 서로의 차분 값을 계산하거나 상관도를 분석해 비밀키를 알아낸다[8-9].

프로파일링 공격은 공격 대상이 되는 디바이스와 동일한 디바이스나 사양이 비슷한 디바이스로부터 공격 대상 디바이스의 전력 파형과 유사한 전력 파형을 수집하여 프로파일을 생성하며 이를 이용한 공격 방법이다. 이때 프로파일 생성을 위한 디바이스는 공격자가 내부 시스템을 조작할 수 있는 화이트 박스(white-box) 환경을 가정으로 한다. 공격자는 사전에 생성된 프로파일 전력 파형을 실제 공격 대상 디바이스로부터 수집한 전력 파형과 비교하고, 서로 대응하는 전력 파형의 프로파일 값(비밀 정보)을 확인함으로써 공격이 이루어진다. 이러한 프로파일링 공격으로는 TA(Template Attack) 및 SM(Stochastic Model) 등이 있다[10-11].

상기한 바와 같이 전력 분석 공격을 수행하기 위하여 공격자는 전력에 영향을 주는 유의미한 파형 구간인 POI를 찾거나 해당 POI의 파형 샘플 데이터를 분석하여 비밀키를 찾아내는 과정에서 많은 전력 파형을 측정하여 수집할 수 있는 장비와 정확한 전력 분석 모델 및 분석 능력이 필요하다. 이러한 전력 분석 공격의 한계를 극복하기 위해 본 논문에서는 딥 러닝 기술인 CNN 기반의

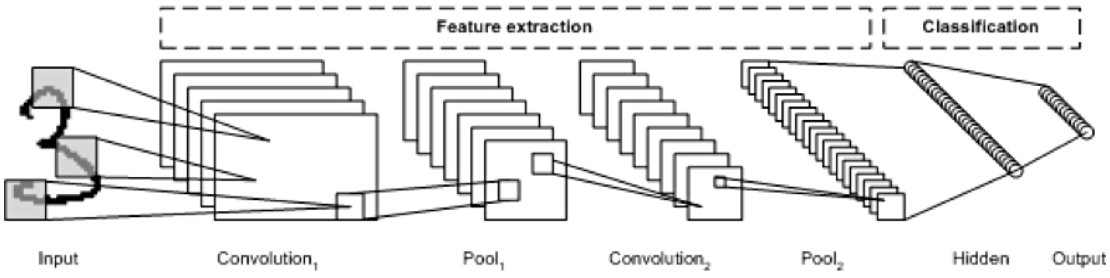

Fig. 1. Structure of CNN algorithm

새로운 전력 공격 모델을 제안하고자 한다. 논문의 공격 방법은 공격 대상용 암호 디바이스와 유사하거나 동일한 장치로부터 전력 파형을 수집하여 학습 과정을 통해 비밀키를 추출하는 프로파일링 공격 기법이다.

2.2 CNN 알고리즘

CNN 알고리즘은 MLP 알고리즘과 같은 완전 연결 계층(fully connected layer) 구조에서 컨볼루션 계층과 풀링 계층을 추가한 구조로, 이미지 데이터와 같이 2차원 이상의 데이터 학습에 유용하도록 고안된 딥 러닝 알고리즘이며 이를 간략하게 나타낸 것이 Fig. 1이다.

CNN의 컨볼루션 계층에서는 2차원 이상의 입력 데이터의 각각의 값들을 가중치(weight) 역할을 하는 필터(커널)와 곱하고, 그 합을 계산하여 입력 데이터의 특징을 추출하는 역할을 한다. 이렇게 계산된 특징 값을 특징 맵(feature map)이라고 한다.

풀링 계층에서는 특징 맵에 여전히 존재하는 특징과 상관없는 잡음 요소를 덜어냄으로써 그 크기를 줄이는 과정을 수행하게 된다. 마지막으로 완전 연결 계층에 전 계층들을 통해 추출된 특징 값들이 입력되어 데이터에 대한 분류를 수행하게 된다.

2.3 RP(Recurrence Plots) 기법

RP 기법은 데이터 값의 회귀에 대한 2차원 표현을 통해 m차원 위상 공간 경로를 탐색하는 것을 목표로 하는 변환 기법이며, 주로 음성 데이터, 주가 데이터 등과 같은 시계열 데이터를 2차원의 데이터로 변환하기 위해 사용된다. 본 논문에서 사용하는 전력 파형 또한 시계열 데이터의 일종이므로 해당 RP 기법의 적용이 가능하다고 볼 수 있다.

RP 기법이 적용되는 과정을 살펴보면, 먼저 Fig. 2와 같이 시계열 데이터의 X축 값(시계열)을 기준으로 각 포인트를 잡아준다. 각 포인트는 $[X_1:0,\ X_2:1,\ X_3:2,\ X_4:1,\ X_5:2,\ X_6:3,\ X_7:4,\ X_8:3,\ X_9:2,\ X_{10}:3,\ X_{11}:2,\ X_{12}:1]$와 같이 표현할 수 있으며, 이러한 포인트들을 2차원의 공간 궤적으로 표현하면, Fig. 3과 같이 $[S_1=(0,1),\ S_2=(1,2),\ S_3=(2,1),\ S_4=(1,2),\ S_5=(2,3),\ S_6=(3,4),\ S_7=(4,3),\ S_8=(3,2),\ S_9=(2,3),\ S_{10}=(3,2),\ S_{11}=(2,1)]$로 표현할 수 있다. 여기서 $S_1=(0,1)$은 $X_1:0$에서 $X_2:1$로의 이동 궤적을 나타내며, 이를 일반화하면 $S_n=X_n\rightarrow X_{n+1}$로 표현할 수 있다.

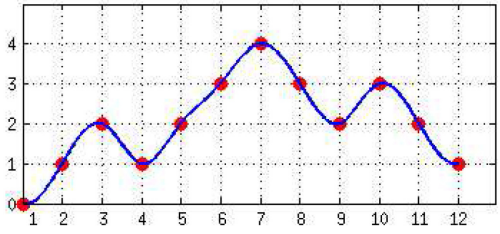

Fig. 2. Time-series signal with 12 data points

최종적으로, Fig. 3에서 표현된 각각의 공간 궤적 포인트(S_n)들 간의 거리를 행렬로 나타내어 시계열 데이터를 이미지화 할 수 있다.

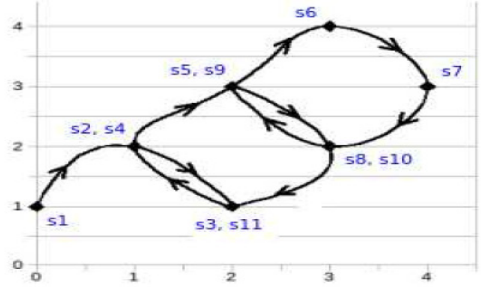

Fig. 3. The 2-D phase space trajectory of time series data

이를 나타낸 것이 Fig. 4이며 거리 행렬 $R_{i,j}$의 거리 값은 다음과 같이 계산된다.

$$R_{i,j} = \theta(\epsilon - \|\overrightarrow{S_i} - \overrightarrow{S_j}\|) \quad (1)$$

위의 식에서 ϵ은 임계 거리 값을 나타내고, $\theta(x)$는 단위 계단 함수이며, 해당 식을 통해 공간 궤적 포인트 간의 상대적 거리를 구할 수 있다.

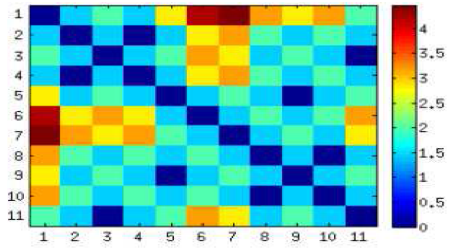

Fig. 4. The recurrence plot of distance matrix between 2-D space trajectories

2.4 전력 분석 공격 지점

전력 분석 공격 모델의 공격 시점은 AES 암호 라운드가 시작하기 전 초기 AddRoundKey 함수와 1 라운드 SubBytes 함수 부분이며, 그림으로 나타내면 Fig. 5와 같다. Fig. 5의 공격 시점을 살펴보면, 초기 AddRoundKey 함수에 평문(plain text)과 비밀키(key)가 입력되어 XOR로 연산되고, 그 결과가 1 라운드 SubBytes 함수에 입력되어 S-box를 거치게 된다. 이 시점에서 공격자가 전력 분석 공격을 통해 1 라운드 SubBytes 함수의 출력 결과를 알아낼 수 있다면, Fig. 6과 같이 공격자가 알고 있는 값(입력 평문, S-box, 1 라운드 SubBytes 함수 결과)을 토대로 다음과 같은 계산식을 이용하여 비밀키를 찾는 것이 가능하다.

$$key = Sbox^{-1}(output) \oplus plaintext$$

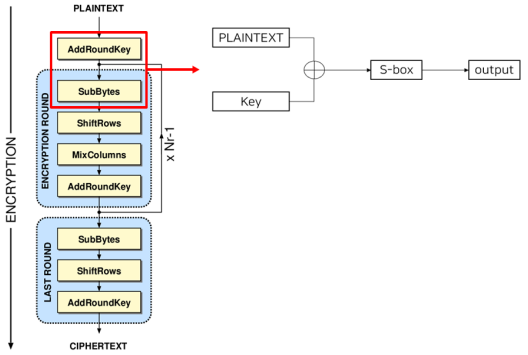

Fig. 5. Power analysis points of AES cryptosystem

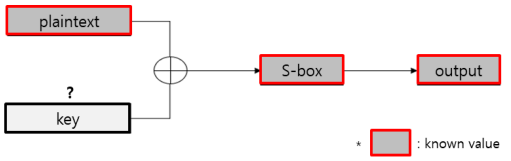

Fig. 6. Key extraction using plain text and output value

3. AES에 대한 CNN 전력 분석 공격

본 논문에서 제안하는 CNN 알고리즘을 이용한 AES에 대한 전력 분석 공격은 프로파일링 기법에 기반한다. 해당 공격 모델에서 프로파일은 공격 시점에 해당하는 바이트 단위의 중간 값과 동일 시점에 해당하는 전력 파형이 매핑되어 구성된다. 이와 같은 프로파일은 공격 대상 디바이스와 동일한 사양의 프로파일용 디바이스를 통해 얻을 수 있으며, 공격자가 암호 알고리듬의 구현 조건이나 개발 환경을 모두 알고 있다는 화이트 박스 공격 환경을 전제로 한다. 제안하는 CNN 기반의 전력 분석 공격 모델은 비밀키를 바이트 단위로 알아내는 공격 모델로 16번의 동일한 과정을 통해 전체 128비트 비밀키를 모두 찾아낼 수 있다. 추가적으로, 본 논문에서는 세 가지 버전의 AES 암호 시스템(AES-128, AES-192, AES-256) 중 AES-128을 구현하여 실험하였다.

3.1 전력 파형 수집

전력 분석 공격은 암호 시스템이 구동될 때 소비되는 전력 정보를 토대로 비밀 정보를 알아내는 공격으로서 전력 파형을 수집하는 과정이 선행되어야 한다. 제안하는 CNN 기반 전력 분석 공격 모델에서는 수집된 전력 파형을 알고리즘의 학습 데이터로 사용하며, 각각 전력 파형에 해당하는 1 라운드 SubBytes 함수 결과를 라벨(label)로 사용한다.

본 논문에서는 AES-128이 구현된 8비트 마이크로프로세서 XMEGA128 보드로부터 전력 파형을 수집하였으며, 파형 수집 도구는 NewAE Technology 사의 Chipwhisperer$^{Ⓢ}$ Lite(이하, CW-Lite)를 사용하였다[12]. 다음 Fig. 7은 XMEGA128 보드에서 AES-128이 수행될 때의 1 라운드까지의 소비 전력을 나타낸 것이다.

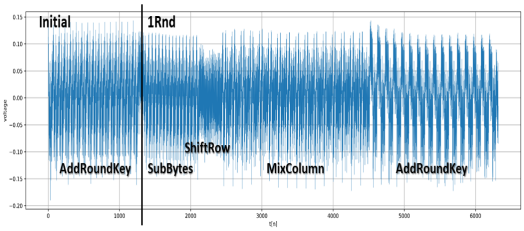

Fig. 7. Power trace of one round on AES-128

Fig. 7을 보면, AES 암호 라운드 함수들을 육안으로 구분할 수 있는데 본 논문에서 학습 데이터로 사용되는 파형은 공격 시점과 관련이 있는 초기 AddRoundKey 함수와 1 라운드의 SubBytes 함수 부분으로서 Fig. 8에서 자세히 볼 수 있다.

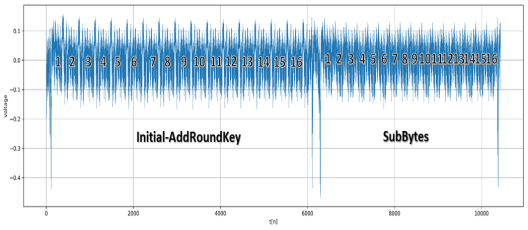

Fig. 8. POI time interval for power analysis attack on AES-128

Fig. 8과 같이 공격 시점에 해당하는 파형의 부분을 살펴보면, 각 함수 내에 16번의 동작 과정을 확인할 수 있으며, 이는 AES-128이 구동될 때 128비트의 데이터를 바이트 단위로 처리함에 따라 나타난 결과이다. 논문에서는 입력 평문과 비밀키를 각각 랜덤으로 입력하여 총 10,000개의 파형을 수집하였으며, 이 중 7,000개는 학습 데이터로, 3,000개는 테스트 데이터로 사용하였다.

3.2 라벨링

지도 학습(supervised learning)이란, 기계 학습의 방법론 중 하나로 학습 데이터와 해당 데이터가 갖는 명시적인 정답이 동시에 주어진 상태에서 학습을 수행하는 방식이며, 여기서 데이터가 갖는 정답을 라벨이라고 한다. 본 논문에서 제안하는 CNN 공격 모델은 이와 같은 지도 학습을 기반으로 하여 학습을 수행하기에 앞서 파형 데이터와 라벨 값을 매핑하는 과정이 필요하며, 해당 모델에서의 라벨은 공격을 통해 최종적으로 알고자 하는 값인 AES 1 라운드 SubBytes 함수의 중간 결과 값이 된다.

3.3 공격 모델 구조

본 논문에서 제안하는 CNN 알고리즘을 이용한 전력 분석 공격의 전체적인 구조는 Fig. 9와 같으며, 앞 장에서 설명한 RP 기법을 전력 파형 데이터에 적용하여 이미지화한 데이터를 입력으로 학습하는 구조이다.

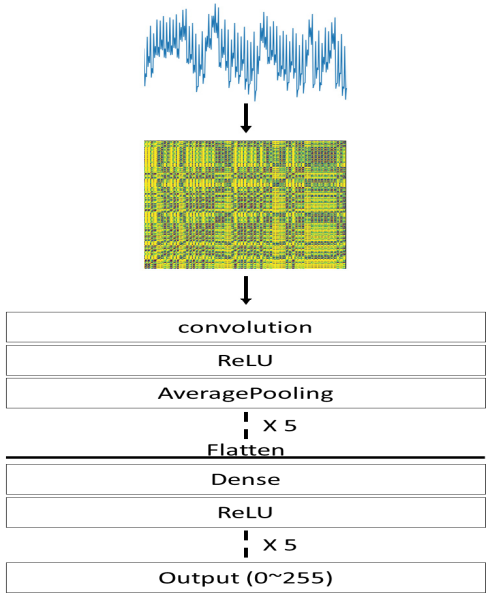

Fig. 9. Structure of CNN-based attack model

공격에 사용한 CNN 모델 구조는 5개의 컨볼루션 계층(풀링 계층 포함)과 5개의 은닉 계층으로 이루어지며,

Table 1. Parameters of the CNN attack model

Layer	Nodes	Filters	Padding	Activation function
Input	260	-		
Conv_1	16	6	Same	ReLU
Pool_1	2	-		
Conv_2	32	6	Same	ReLU
Pool_2	2	-		
Conv_3	64	6	Same	ReLU
Pool_3	2	-		
Conv_4	128	6	Same	ReLU
Pool_4	2	-		
Conv_5	256	6	Same	ReLU
Pool_5	2	-		
Flatten	-			
Dense_1	1000	-		ReLU
Dense_2	1000	-		ReLU
Dense_3	1000	-		ReLU
Dense_4	1000	-		ReLU
Dense_5	1000	-		ReLU
Output	1000	-		Softmax

풀링 계층은 Average-Pooling 기법을 적용하였다. 추가적으로, 모든 계층에서 활성화 함수로 ReLU 함수를 사용하였고 손실 함수는 Cross-Entropy 함수를 적용하였다. 또한, 알고리즘의 최적화를 위해 Adam 알고리즘을 채택하였으며 학습률은 0.0001로 설정하였다. 공격 모델을 설계하는데 사용된 자세한 파라미터 값은 Table 1과 같다.

4. 실험 결과

본 논문에서는 일반 전력 파형을 CNN 공격 모델에 입력하여 학습하는 실험과 RP 기법을 통해 전력 파형을 이미지화한 데이터를 CNN 공격 모델에 입력하여 학습하는 실험으로 구분한 2가지 실험을 진행하였다. 이를 통해 전력 파형에 대한 RP 기법 적용이 CNN 공격 모델에 어느 정도 성능 향상을 보이는지 확인하고자 한다. 실험용 장비의 제원은 다음과 같으며 개발 언어는 Python 3.7을 사용하였다.

CPU : Intel(R) i7-4790 (3.60 GHz, 8 CPUs)

RAM : 6.00 GB

Windows : Windows 7 Ultimate K 64 Bits

본 논문에서 제안하는 CNN 공격 모델은 AES 암호 시스템의 비밀키를 바이트 단위로 알아내는 공격 방법으로 실험에서는 AES-128 비밀키의 첫 번째 바이트를 찾아내는 공격을 수행하였다. 최종적인 비밀키를 모두 찾기 위해서는 한 바이트에 대한 공격을 16본 반복하여야 한다. 또한, 상대적으로 많은 메모리가 있어야 하는 CNN 알고리즘의 특성상 Fig. 8과 같이 공격 시점에 해당하는 모든 파형을 입력으로 하지 않고 실질적으로 첫 번째 바이트에 대한 SubBytes 연산 부분의 파형만 입력으로 사용하였다.

먼저, 일반 전력 파형을 입력으로 CNN 공격 모델을 학습한 결과는 Table 2와 같으며, 에포크(epoch)에 따른 정확도(accuracy)의 추이는 Fig. 10과 같다. 일반 전력 파형을 입력으로 CNN 공격 모델을 학습한 경우, 최고 22.23%의 정확도로 AES-128 비밀키의 첫 번째 바이트를 찾아내는 것을 확인할 수 있다. 이러한 실험 결과는 정확도가 낮아 여러 번의 공격 수행을 통해 후보 키에 대한 중복을 검색해야만 완전한 비밀키 바이트를 찾아낼 수 있음을 의미한다.

Table 2. Result of implementing CNN attack model (normal power trace)

Epoch	Time(sec)	Accuracy(%)
150	1892.92	17.06
300	3753.84	22.23

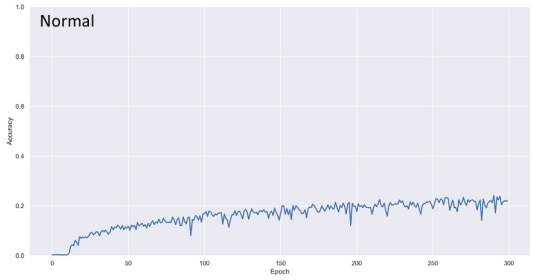

Fig. 10. Accuracy of CNN attack model according to epoch(normal power trace)

다음으로 RP 기법을 통해 전력 파형을 이미지화한 데이터를 입력으로 CNN 공격 모델 학습한 결과를 알아보고자 한다. 다음 Fig. 11은 RP 기법을 이용하여 전력 파형을 이미지화한 것을 나타낸다. Fig. 11에서의 전력 파형은 실제 CNN 공격 모델에 입력되는 샘플링된 전력 파형이며, 1 라운드 SubBytes 함수의 첫 번째 바이트 연산 부분에 해당한다. 이렇게 샘플링된 전력 파형은 총 260 샘플로 구성되어 있으며, RP 기법을 통해 260×260 크기의 이미지 데이터로 변환된다. 최종적으로 이미지 데이터 값의 분포를 넓히기 위해 각 픽셀에 15를 곱하여 CNN 모델에 입력되게 된다.

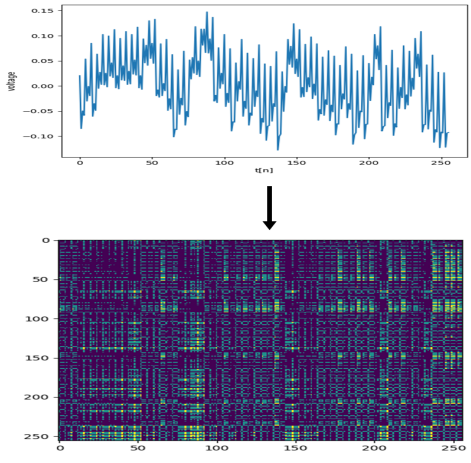

Fig. 11. Power trace image transformed by the RP processing

다음 Table 3과 Fig. 12는 상기한 과정을 통해 이미지화한 데이터를 입력으로 학습한 결과를 나타낸다. RP 기법을 통해 전력 파형을 이미지화하여 학습한 경우, 최고 97.93%의 정확도로 AES-128 비밀키의 첫 번째 바이트를 찾아내는 것을 확인할 수 있었다. 이는 또한, RP 기법을 적용하지 않은 일반 전력 파형을 학습하였을 때보다 약 4.4배 높은 정확도를 나타내어 RP 기법을 적용하였을 때 학습 성능이 기존 방법보다 월등히 높아진다는 것을 알 수 있다.

Table 3. Result of implementing CNN attack model (power trace transformed by the RP processing)

Epoch	Time(sec)	Accuracy(%)
150	3525.91	92.13
300	6871.81	97.93

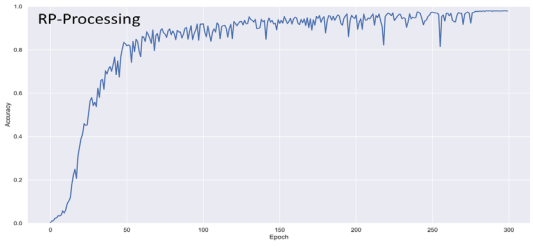

Fig. 12. Accuracy of CNN attack model according to epoch(power trace transformed by the RP processing)

이렇게 RP 기법 적용이 CNN 공격 모델의 학습 성능을 크게 향상시키는 이유는 라벨 값에 따라 전력 파형이 갖는 특징(feature)들이 RP 기법을 통해 2차원의 공간으로 확장되어 CNN 모델 학습에 있어 더욱 유용한 형태로 변환되어 나타난 결과로 분석된다.

Table 4. Accuracy comparison of previous attack model

CNN-based attack	Accuracy(%)	Difference of accuracy(%)
[13]	81.20	-16.73
[14]	88.72	-9.21
[15]	89.80	-8.13
Our model	97.93	0

추가적으로, 본 논문에서 얻은 공격 모델의 결과와 이전에 소개된 CNN 알고리즘을 이용한 전력 분석 공격 결과를 비교한 것이 Table 4이다. 기존의 논문 [13], [14], [15]의 결과는 본 논문에서와 같이 8비트 마이크로프로세서 보드에 구현된 AES-128을 대상으로 바이트 단위 공격을 수행한 것이며, 전력 파형 수집 도구 또한 CW-Lite로 동일하게 사용하였다.

Table 4와 같이, [13], [14], [15]의 결과는 모두 정확도 90%를 넘지 못하는 결과를 나타내며, 본 논문에서 얻은 결과와도 8%이상의 차이가 나타난다. 본 논문과 비교 분석한 기존의 실험 결과 중 가장 높은 정확도를 나타내는 Wei등의 실험에서는 전력 파형의 주기를 이용한 전처리 기법을 적용하여 학습을 수행하였으나 본 논문에서 제안한 RP 기법을 통한 전처리 효과에 비하면 낮은 결과를 나타낸다[15]. 이를 통해 본 논문에서 제안하는 RP 전처리 기법이 CNN 기반의 전력 분석 공격 모델의 성능 향상에 크게 기여함을 알 수 있다.

5. 결론

암호 알고리즘을 구현한 디바이스에 대한 전력 분석 공격에서는 고도의 전력 파형 수집과 분석 능력이 필요하며, 공격 수행을 위해 많은 시간과 비용이 요구된다.

본 논문에서는 이러한 전력 분석 공격의 어려움과 비효율성을 극복하기 위해 딥 러닝 기술의 일환인 CNN 알고리즘을 활용한 전력 분석 공격 모델을 제시하였다. 특히, 1차원의 전력 파형 데이터를 RP 기법을 통해 시계열 전력 파형 데이터를 2차원으로 이미지화하여 처리하는 새로운 공격 모델을 제안하였다.

상기한 공격 모델의 성능을 직접 알아보기 위해 AES-128 암호 시스템에 대한 비밀키를 알아내는 공격을 진행하였으며 실제 실험 보드를 통해 바이트 단위로 수행하였다. 실험 결과, RP 전처리를 하지 않은 전력 파형을 학습하였을 때에는 약 22%의 정확도를 보여 낮은 학습률을 나타내었지만, RP 기법을 통해 전력 파형에 대한 전처리를 수행한 결과 약 98%의 정확도로 모델의 학습 성능이 크게 높아진 것을 확인하였다.

실험에서 나온 정확도는 비밀키가 사용된 단 하나의 파형으로도 비밀키 바이트를 충분히 찾아낼 수 있으며, 이는 16번의 공격 수행으로 전체 비밀키 값을 알아낼 수 있음을 의미한다. 따라서 암호용 디바이스를 구현할 경우에는 전력 분석 공격에 충분히 대응할 수 있는 하드웨어적인 대응책이나 고차 마스킹 연산 등을 적용하여야 한다.

References

[1] F. X. Standaert, B. Gierlichs, and I. Verbauwhede, "Partition vs. comparison side-channel Distinguishers : An empirical evaluation of statistical tests for univariate side-channel attacks against two unprotected CMOS device", *ICISC'08,* LNCS 5461, pp. 253-267, 2008.
DOI : https://doi.org/10.1007/978-3-642-00730-9_16

[2] S. Mangard, E. Oswald, and T. Poop, "Power analysis attacks: Revealing the secrets of smart cards", p. 333, Springer, 2008, pp. 119-165.
DOI : https://doi.org/10.1007/978-0-387-38162-6

[3] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a Convolutional Neural Network", *International Conference on Engineering and Technology (ICET'17),* Antalya, Turkey, pp. 1-6, Aug. 2017.
DOI: https://doi.org/10.1109/ICEngTechnol.2017.8308186

[4] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview", *Neural Networks,* Vol. 61, pp. 85-117, 2015.
DOI: https://doi.org/10.1016/j.neunet.2014.09.003

[5] R. Collobert and S. Benjio, "Links between perceptrons, MLPs and SVMs", *Proceedings of the twenty-first international conference on Machine learning(ICML'04),* Banff, Canada, pp. 23-30, July 2004.
DOI: https://doi.org/10.1145/1015330.1015415

[6] Federal Information Processing Standards Publication (FIPS 197), "Advanced Encryption Standard(AES)", National Institute of Standards and Technology (NIST), 2001.
DOI: https://doi.org/10.6028%2FNIST.FIPS.197

[7] N. Hatami, Y. Gavet, and J. Debayle, "Classification of Time-Series Images Using Deep Convolutional Neural Networks", *International Conference on Machine Vision(ICMV '17),* Vienna, Austria, Vol. 10696. pp. 106960Y-1-106960Y-8, Nov. 2017.
DOI: https://doi.org/10.1117/12.2309486

[8] 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

[9] E. Brier, C. Clavier, and F. Olivier, "Correlation Power Analysis with a Leakage Model", *CHES'04,* LNCS 3156, pp. 16-29, 2004.
DOI: https://doi.org/10.1007/978-3-540-28632-5_2

[10] S. Chari, J. R. Rao, and P. Rohatgi, "Template Attacks", *CHES'02,* LNCS 2523, pp. 13-28, 2002.
DOI: https://doi.org/10.1007/3-540-36400-5_3

[11] W. Schindler, K. Lemke, and C. Paar, "A Stochastic Model for Differential Side Channel Cryptanalysis", *CHES'05,* LNCS 3659, pp. 30-46, 2005.
DOI: https://doi.org/10.1007/11545262_3

[12] NewAE Technology Inc., "Single Board Solutions - Chipwhisperer-Lite 32-bit," Available From: https://www.newae.com/chipwhisperer, (accessed Dec. 1, 2019).

[13] H. Wang, M. Brisfors, S. Forsmark, and E. Dubrova, "How Diversity Affects Deep-Learning Side-Channel Attacks", Cryptology ePrint Archive, Report 2019/664, Available From: https://eprint.iacr.org/2019/664 (accessed Dec. 1, 2019).

[14] A. Golder, D. Das, J. Danial, S. Ghosh, S. Sen, and A. Raychowdhury, "Practical Approaches Towards Deep-Learning Based Cross-Device Power Side Channel Attack", *IEEE Trans. on VLSI systems,* Vol. 27, No. 12, pp. 2720-2733, 2019.
DOI: https://doi.org/10.1109/TVLSI.2019.2926324

[15] L. Wei, B. Luo, Y. Li, Y. Liu, and Q. Xu, "I Know What You See: Power Side-Channel Attack on Convolutional Neural Network, Accelerators", *Proceedings of the 34th Annual Computer Security Applications Conference (ACSAC'18),* San Juan PR USA, pp. 393-406, Dec. 2018.
DOI: https://doi.org/10.1145/3274694.3274696

권 홍 필(Hong-Pil Kwon) [준회원]

- 2018년 2월 : 호서대학교 정보보호학과 (공학사)
- 2020년 2월 : 호서대학교 일반대학원 정보보호학과 (공학석사)
- 2020년 3월 ~ 현재 : ㈜라닉스 연구원

〈관심분야〉
부채널 공격, 인공지능 보안, 암호학

하 재 철(Jae-Cheol Ha) [종신회원]

- 1989년 2월 : 경북대학교 전자공학과 (공학사)
- 1993년 8월 : 경북대학교 일반대학원 전자공학과 (공학석사)
- 1998년 2월 : 경북대학교 일반대학원 전자공학과 (공학박사)
- 1998년 3월 ~ 2007년 2월 : 나사렛대학교 정보통신학과 교수
- 2007년 3월 ~ 현재 : 호서대학교 컴퓨터정보공학부 교수
- 2013년 1월 ~ 현재 : 한국정보보호학회 상임부회장
- 2009년 1월 ~ 현재 : 한국산학기술학회 이사

〈관심분야〉
암호학, 네트워크 보안, 부채널 분석, 인공지능 보안