Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.29, NO.5, Oct. 2019 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2019.29.5.997

Multi-Layer Perceptron 기법을 이용한 전력 분석 공격 구현 및 분석

권 홍 필,† 배 대 현, 하 재 철‡
호서대학교

Implementation and Analysis of Power Analysis Attack Using Multi-Layer Perceptron Method

Hongpil Kwon,† DaeHyeon Bae, Jaecheol Ha‡
Hoseo University

요 약

본 논문에서는 기존 전력 분석 공격의 어려움과 비효율성을 극복하기 위해 딥 러닝 기반의 MLP(Multi-Layer Perceptron) 알고리즘을 기반으로 한 공격 모델을 사용하여 암호 디바이스의 비밀 키를 찾는 공격을 시도하였다. 제안하는 전력 분석 공격 대상은 XMEGA128 8비트 프로세서 상에서 구현된 AES-128 암호 모듈이며, 16바이트의 비밀 키 중 한 바이트씩 복구하는 방식으로 구현하였다. 실험 결과, MLP 기반의 전력 분석 공격은 89.51%의 정확도로 비밀 키를 추출하였으며 전처리 기법을 수행한 경우에는 94.51%의 정확도를 나타내었다. 제안하는 MLP 기반의 전력 분석 공격은 학습을 통한 feature를 추출할 수 있는 성질이 있어 SVM(Support Vector Machine)과 같은 머신 러닝 기반 모델보다 우수한 공격 특성을 보임을 확인하였다.

ABSTRACT

To overcome the difficulties and inefficiencies of the existing power analysis attack, we try to extract the secret key embedded in a cryptographic device using attack model based on MLP(Multi-Layer Perceptron) method. The target of our proposed power analysis attack is the AES-128 encryption module implemented on an 8-bit processor XMEGA128. We use the divide-and-conquer method in bytes to recover the whole 16 bytes secret key. As a result, the MLP-based power analysis attack can extract the secret key with the accuracy of 89.51%. Additionally, this MLP model has the 94.51% accuracy when the pre-processing method on power traces is applied. Compared to the machine leaning-based model SVM(Support Vector Machine), we show that the MLP can be a outstanding method in power analysis attacks due to excellent ability for feature extraction.

Keywords: Side-Channel Analysis, Power Analysis Attack, Deep Learning MLP, Machine Learning SVM

I. 서 론

부채널 분석(Side Channel Analysis) 공격이란 암호 디바이스로부터 얻을 수 있는 소비전력, 전자기파, 실행시간 등과 같은 부채널 정보를 기반으로 비밀 정보를 분석하는 공격 방법이다[1]. 이러한 부채널 분석 공격 중에는 전력 분석 공격이 대표적이라고 할 수 있다. 전력 분석 공격이란, 암호 디바이스가 구동될 때 소비되는 전력을 오실로스코프와 같은 전력 측정 장비로 수집하여 이 소비 전력 정보를 바

Received(08. 29. 2019), Modified(09. 30. 2019), Accepted(10. 01. 2019)
† 주저자, khp9890@gmail.com
‡ 교신저자, jcha@hoseo.edu(Corresponding author)

탕으로 칩 내부의 비밀 정보를 탐색하는 공격 방법이다[2]. 전력 분석 공격은 암호 디바이스가 처리되는 데이터나 연산의 종류에 따라 소비되는 전력이 상이하다는 이론을 바탕으로 한다. 이러한 공격에서는 공격자가 측정한 전력 파형에서 실제 비밀 값이 파형에 영향을 주는 순간인 POI(Point of Interest)을 찾아내고, 이 POI에서의 데이터가 갖는 통계적 사실을 바탕으로 비밀 값을 추측하는 전력 분석 능력이 요구된다. 또한, 전력 분석 공격이 성공하기 위해 상기한 POI를 찾아내고 파형과 데이터 간의 연관성을 찾는 등과 같은 공격을 위한 사전 과정이 필요하다.

본 논문에서는 위와 같은 전력 분석 공격의 어려움과 비효율성을 극복하기 위해 딥 러닝(deep learning) 기술 중 다층 퍼셉트론(Multi-Layer Perceptron, MLP) 기법[3, 4]을 적용한 새로운 전력 분석 공격 모델을 제안하고자 한다. 딥 러닝 기법은 어떠한 특징(feature)을 갖는 많은 데이터를 인간의 뇌와 같은 구조를 갖는 신경망 구조의 모델에 입력하여 입력된 데이터가 갖고 있는 특징에 대한 학습을 수행하는 머신 러닝 기술이다. 이러한 딥 러닝 기법을 전력 분석 공격에 적용함으로써 전력파형에 대한 특징 분석을 학습 과정을 통해 자동화할 수 있어 정확한 POI를 찾아야 하는 어려움과 비효율성을 개선할 수 있다.

논문에서는 국제 표준 암호 알고리즘인 AES (Advanced Encryption Algorithm)[5]를 실제 암호 디바이스에 구현하고 MLP 알고리즘을 사용한 전력 분석 공격을 시도하였다. 또한, 초기 전력 분석 공격에 사용되었던 머신 러닝(machine learning) 기술의 일환인 SVM(Support Vector Machine) 기법[6, 7]에 의한 공격 모델을 구축하여 딥 러닝 기법과 머신 러닝 기법을 비교 분석함으로써 MLP 모델의 효율성을 알아보았다. 마지막으로 공격 모델에 입력되는 파형이 측정될 때 여러 요인으로 인해 발생된 잡음을 제거하고 더 나아가 데이터 종속성을 갖는 파형의 특징을 확대할 수 있는 전처리 기법을 적용할 경우 학습 성능이 어느 정도 향상되는지 분석하였다.

II. 배경지식

2.1 전력 분석 공격

전력 분석 공격은 암호 모듈이 내재되어 있는 디바이스가 구동될 때 소비되는 전력을 기반으로 비밀 값을 탈취하는 공격 방법이다. 특히, 암호 라운드 연산이 수행되는 중 비밀 값이 연산될 때의 전력 파형을 찾아내고 해당 파형에 해당하는 데이터의 통계적 분석을 통해 비밀 값을 추측하게 된다.

전력 분석 공격은 공격 방법에 따라 크게 Non-Profiling 공격과 Profiling 공격으로 구분할 수 있다. 먼저 Non-Profiling 공격은 특정 디바이스에 랜덤한 여러 평문을 입력하여 암호 연산을 수행하고, 이때에 소비되는 전력을 측정하여 수집한 파형들을 통계적으로 분석함으로써 모든 추측 키 중 가장 높은 확률을 갖는 추측 키를 실제 비밀 키로 확정하는 공격 방법이다. Non-Profiling 공격으로는 SPA(Simple Power Analysis)[8], DPA(Differential Power Analysis)[8], CPA(Correlation Power Analysis)[9] 등이 있다. SPA 공격은 단순히 측정된 파형만을 보고 처리된 데이터를 추측하는 공격 방법이고, DPA와 CPA 공격은 많은 랜덤한 평문을 입력하여 측정한 파형들을 기반으로 차분 값을 계산하거나 상관도를 구하여 통계적으로 비밀 키를 추측하는 공격 방법이다.

Profiling 공격은 공격 대상 디바이스와 동일하거나 비슷한 사양을 갖는 다른 디바이스를 통해 프로파일을 생성하고, 실제 공격 대상 디바이스로부터 얻은 파형과 프로파일 간의 매칭 확률을 비교함으로써 비밀 키를 알아내는 공격 방법이다. 여기서 프로파일용 디바이스는 공격자가 내부 데이터에 대한 조작이 가능한 환경(화이트 박스)이어야 한다. Profiling 공격에는 TA(Template Attack)[10], SM (Stochastic Model)[11] 등이 있다. 본 논문에서 사용하는 MLP 딥 러닝 기법이나 SVM 머신러닝 기법을 적용한 전력 분석 공격 모델은 Profiling 기반 공격이라고 할 수 있다.

2.2 MLP(Multi-Layer Perceptron)

인간의 뇌 구조 즉, 생물학적 뉴런의 구조를 바탕으로 컴퓨터에서도 대량의 데이터를 병렬 처리 연산이 가능하도록 디자인한 알고리즘을 인공 신경망(Artificial Neuron Network, ANN)이라고 하며 대표적으로 퍼셉트론 구조가 있다[3].

퍼셉트론은 Fig. 1과 같이 입력 계층(x_i)과 출력 계층(out)을 가지며, 입력 계층에서 데이터를 입력받아 이를 가중치(W_i)와 곱하고 바이어스(b)를 더

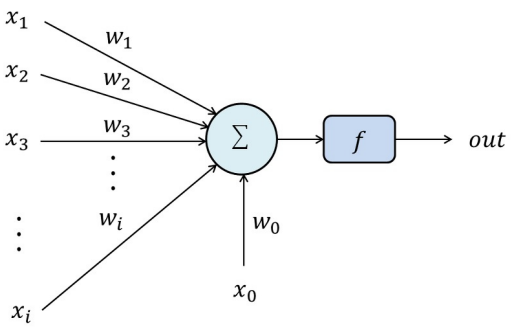


Fig. 1. Structure of Perceptron

한 후, 이 값을 활성화 함수인 Step 함수(f)에 입력하여 최종적으로 0 또는 1의 값을 출력하게 된다. 즉, 퍼셉트론은 입력 값을 받아 2개의 출력 값 중 하나를 출력하는 선형 이진 분류기라고 할 수 있다.

그러나 퍼셉트론은 선형 이진 분류기이기 때문에 XOR과 같은 비선형적 데이터에 대해서는 분류가 불가능하다는 한계가 존재한다. 이러한 단층 퍼셉트론의 한계를 극복하기 위해 기존 퍼셉트론의 입력 계층과 출력 계층 사이에 은닉 계층을 두어 비선형적으로 분리되는 데이터에 대해서도 학습이 가능하도록 한 다층 퍼셉트론이 고안되었다[4].

Fig. 2에서 보는 바와 같이 MLP는 입력 계층(input layer), 은닉 계층(hidden layer), 출력 계층(output layer)으로 이루어져 있는데 은닉 계층에서는 데이터의 입·출력 과정에서 직접적으로 보이지 않는 숨겨진 특징을 학습하는 역할을 한다. 입력 계층과 연결된 은닉 계층의 한 노드만 보면 하나의 단층 퍼셉트론과 같은 구조를 갖는 것을 확인할 수 있다. 즉, 은닉 계층에 있는 각각의 노드는 한 퍼셉트론의 활성화 함수의 역할을 한다고 볼 수 있다.

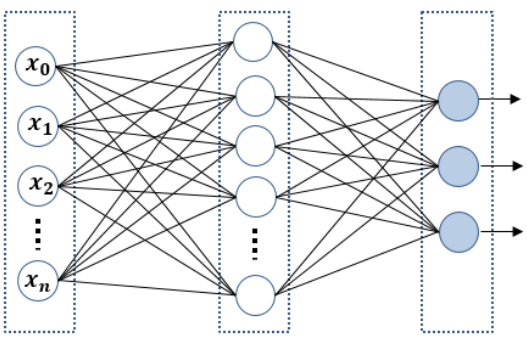


Fig. 2. Structure of Multi-Layer Perceptron

MLP는 여러 개의 단층 퍼셉트론으로 이루어져 있기 때문에 여러 가중치(weight) 값들이 존재하고 데이터와 곱해진 결과가 각각의 활성화 함수에 입력됨으로써 여러 가지로 분류될 수 있는 데이터에 대한 학습을 수행할 수 있게 된다. 추가적으로 MLP에서는 활성화 함수로 Sigmoid, Tanh, ReLU와 같은 비선형 함수를 사용한다.

2.3 SVM(Support Vector Machine)

서포트 벡터 머신(Support Vector Machine, SVM)이란 주로 데이터 분류를 위해 사용하는 머신러닝의 한 분야이다[6]. 이는 데이터가 사상된 공간에서 각 그룹 사이의 거리, 즉 마진(margin)이 가장 큰 분류 경계를 찾는 알고리즘이다. SVM에서 선형 분류가 불가능한 데이터 즉, 비선형성을 갖는 데이터를 처리하는 경우에도 새로운 특성을 추가해 데이터를 고차원 공간으로 이동시킴으로써 비선형 분류가 가능하게 된다. 이때 고차원으로 사상시키는 작업을 비용 면에서 효율적으로 처리하기 위해 사용하는 것이 커널 트릭(Kernel trick)이며, 커널 트릭을 이용한 SVM을 Kernel-SVM이라 한다[7]. 이때 사용하는 커널 함수로는 다항 커널(polynomial kernel), 시그모이드 커널(sigmoid kernel), 가우시안 RBF 커널(Gaussian Radial Basis Function kernel)등이 있다.

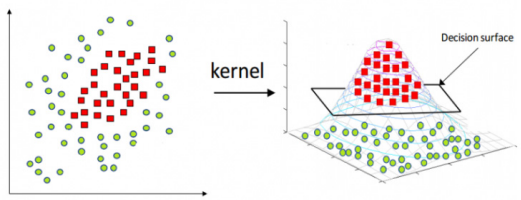


Fig. 3. An example of Kernel-SVM

III. 전력 분석 모델 디자인

3.1 MLP를 이용한 전력 분석 모델

본 논문에서 제안하는 MLP 전력 분석 공격 모델과 비교 분석을 위해 사용하는 SVM 모델은 모두 AES-128 암호 알고리즘[5]을 대상으로 한다. AES-128 암호 과정에서 비밀 키 분석이 가능한 공

격 시점 즉, POI는 Fig. 4에 나타낸 바와 같다.

AES-128의 암호 구조를 보면 Fig. 4와 같이 암호 라운드가 시작하기 전에 AddRoundKey 함수를 통해 평문(plain text)과 비밀 키가 XOR 연산되고, 그 결과 값이 첫 번째 라운드 SubBytes 함수의 입력으로 사용되어진다. 즉, 공격자가 전력 분석 공격을 위한 학습 모델을 통해 SubBytes 함수에 대한 중간 결과(output)만 알 수 있다면, 역 S-box 연산을 통해 비밀 키를 알아내는 것이 가능해진다. 특히, MLP를 이용한 공격 모델에서는 SubBytes 함수에 대한 중간 결과(라벨)와 파형(학습 데이터)을 입력으로 MLP 학습을 수행한 후, 학습된 모델에 공격 대상 파형이 입력되었을 때 파형에 대한 SubBytes 중간 결과를 도출하여 해당 파형을 측정한 디바이스의 비밀 키를 알아낼 수 있게 된다.

2013년, Z. Martinasek 등은 MLP 알고리즘을 이용하여 AES 암호 모듈에 대한 전력 분석 공격을 수행하였는데[12], 이 경우에는 비밀 키를 고정한 상태에서 수집한 파형에 대해 학습을 하여 85.23%의 정확도(accuracy)를 보였다. 또한, 수집한 전력 파형에 대한 전처리를 하여 학습을 수행한 후 적용한 MLP 기반 공격에서는 94.57%의 정확도를 나타내었다[13].

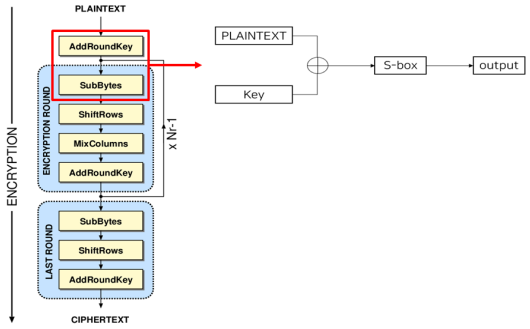


Fig. 4. POI for power analysis on the AES-128

3.1.1 전력 파형 수집

먼저, MLP 모델에 입력할 데이터인 전력 파형을 수집한다. 본 논문에서는 AES-128 암호 모듈이 구현된 XMEGA128 8비트 마이크로프로세서 보드에 대한 파형 수집을 진행하였으며, 파형 수집 도구는 NewAE Technology 사의 ChipWhisperer® Lite(이하, CW-Lite)를 사용하였다[14]. 다음

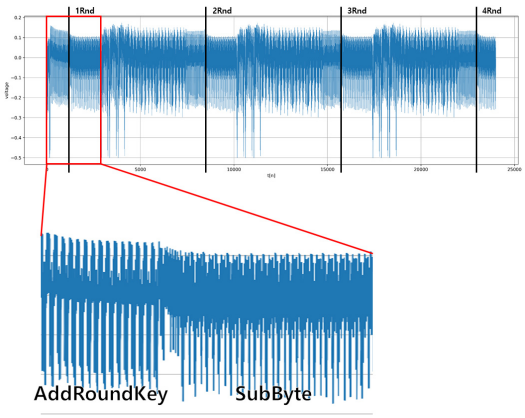


Fig. 5. Measured power trace on AES-128

Fig. 5는 CW-Lite를 이용해 수집한 AES-128 암호 라운드에 해당하는 전력 파형이다. 그림에서 보이는 것과 같이 라운드 별로 파형의 모형이 구분되는 것을 확인할 수 있으며, 한 라운드 안에서도 4가지 함수(SubBytes, ShiftRows, MixColumns, AddRoundKey)가 SPA를 통해 구분되는 것을 확인할 수 있다.

본 논문에서 제안하는 MLP 공격 모델에서는 입력 데이터로 암호 라운드가 시작하기 전 AddRoundKey 함수 부분의 파형 샘플과 첫 번째 라운드 SubBytes 함수 부분의 파형 샘플을 사용하였으며 이는 총 2,880 샘플 수를 갖는다. 또한, 학습 데이터로 사용된 파형의 수는 7,000개, 테스트 데이터로 사용된 파형의 수는 3,000개이며, 모든 파형은 평문과 비밀 키를 랜덤하게 암호 디바이스에 입력하여 측정하였다.

3.1.2 라벨링

MLP 모델을 학습시키기 위해서는 입력 데이터(파형)에 대한 라벨을 붙여주어야 한다. 즉, 학습할 때에 어떠한 파형이 들어오면 그 파형이 어떤 정답을 갖는 파형인지를 라벨을 통해 알려주어야 학습을 할 수 있게 된다. 상기한 바와 같이 제안하는 MLP 모델은 첫 번째 라운드 SubBytes 함수의 중간 결과 즉, S-box의 결과를 알게 되면 암호 라운드 시작 전 AddRoundKey에 입력된 비밀 키를 알 수 있다는 점을 이용한 공격 모델이다. 즉, 제안하는 MLP 모델에서 학습을 위해 입력되는 파형에 대한 라벨은 첫 번째 라운드 SubBytes의 중간 결과가 된다. 따

라서 본 논문에서 제안하는 MLP 공격 모델에서의 프로파일은 첫 번째 라운드 SubBytes의 중간 결과가 라벨링된 파형 샘플이라고 할 수 있다.

3.1.3 MLP 모델 구조

MLP의 구조는 입력 계층, 은닉 계층, 출력 계층으로 이루어져 있다. 제안하는 MLP를 이용한 전력 분석 공격 모델은 한 개의 입력 계층, 두 개의 은닉 계층, 한 개의 출력 계층으로 총 4개의 계층으로 구성하였으며 전체적인 구조는 Fig. 6과 같다.

입력 계층은 2,880 노드를 갖으며 이는 앞서 설명한 암호 라운드가 시작하기 전 AddRoundKey 함수 부분의 파형과 첫 번째 라운드 SubBytes 함수 부분의 파형의 총 샘플 수에 해당한다. 결국, 입력 계층에 입력되는 값들은 총 2,880 샘플이 갖는 각각의 전력 값이 된다.

은닉 계층은 총 두 개로 두었으며, 이는 단층으로 은닉 계층을 놓는 것보다 여러 층으로 은닉 계층을 놓는 것이 숨겨진 특징을 추출하기에 더욱 유리하기 때문이다. 또한, 각각의 은닉 계층의 노드 수를 100개로 두었는데 이는 여러 설정으로 실험해 본 결과 속도와 성능 면에서 우수한 결과를 나타내었다. 마지막으로 은닉 계층의 활성화 함수는 모두 "sigmoid" 함수를 사용하였다.

출력 계층은 총 256개의 노드로 이루어지며, 각각의 노드는 순차적으로 한 바이트 값의 범위인 0x00~0xFF 중 하나의 카테고리를 의미한다. 각 노드의 출력 값은 입력된 파형의 평문 첫 번째 바이트에 대한 SubBytes 연산 결과가 해당 카테고리에 속할 확률을 의미한다. 예를 들어 0xA3번째 노드의 출력 값이 0.98이라면, 입력된 파형의 평문 첫 번째 바이트에 대한 SubBytes 결과가 0xA3일 확률이 98%라는 것을 의미한다. 따라서 출력 노드의 결과가 가장 높은 카테고리를 SubBytes 연산 결과로 판단한다. 또한, 제안하는 MLP 모델 분류를 위한 손실 함수로는 "Cross-Entropy" 함수를, 최적화는 "Adam" 함수를 사용했으며 학습률은 0.001로 설정하였다.

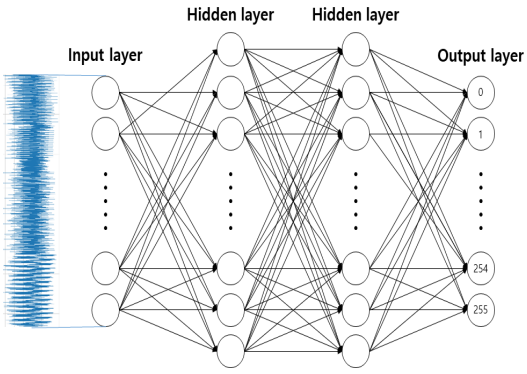


Fig. 6. Proposed MLP-based power analysis attack model

해당 모델의 최종적인 결과는 상기한 바와 같이 첫 번째 라운드 SubBytes의 중간 결과이다. 실제 공격자가 알고자 하는 비밀 키 값은 제안하는 공격 모델을 통해 얻은 중간 결과로부터 다음 식을 통해 도출해 낼 수 있다.

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

즉, Fig. 7과 같이 공격자는 입력 평문과 공격 모델을 통해 얻은 SubBytes 중간 결과를 알기 때문에 이 결과 값을 S-box에 역으로 대입하여 AddRoundKey 연산을 수행한 중간 결과를 알아내고, 그 값에서 입력 평문을 XOR 연산하면 최종적으로 원하는 값인 비밀 키 값을 알아낼 수 있다.

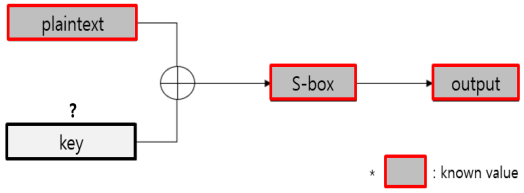


Fig. 7. Structure of the intermediate result of the S-box

3.2 SVM을 이용한 전력 분석 모델

본 논문에서는 AES-128에 대한 전력 분석 공격을 위해 머신 러닝 기법 중 RBF 커널을 사용한 Kernel-SVM 모델을 사용하였다. SVM의 매개 변수 C와 RBF 커널의 매개 변수 Gamma는 여러 조합을 테스트하여 최선의 값을 찾아내는 grid search 방법을 이용한다. 신경망 구조를 사용하는 딥 러닝 기법에서는 자동화된 특징 추출이 가능하다. 그러나 SVM과 같은 기존 머신 러닝 알고리즘은 자

동화된 특징 추출이 불가능하기 때문에 입력되는 데이터가 모델의 정확도에 매우 큰 영향을 미친다. 즉, SVM 모델에서는 찾고자 하는 라벨과 직접 관련이 있는 파형의 부분만을 추출하여 입력해 주어야 한다.

따라서, 신경망 구조를 이용하는 MLP 모델에서는 암호 라운드가 시작되기 전 AddRoundKey와 첫 번째 라운드 SubBytes에 해당하는 파형 샘플을 입력으로 하지만, SVM 모델에서는 첫 번째 라운드 SubBytes 연산에서 찾고자 하는 특정한 한 바이트 연산 부분의 파형 샘플만 입력으로 사용한다.

기존에 SVM 알고리즘을 이용하여 AES 암호 모듈에 대한 전력 분석 공격을 수행한 대표적인 사례로는 G. Hospodar 등이 발표한 것이 있는데[15], 이 경우에는 AES 암호 모듈의 첫 번째 라운드 SubBytes를 수행한 중간 결과의 특정 한 비트를 분류하는 것을 기준으로 학습을 하였다. 이 논문에서의 실험 결과, 최대 75.5%의 정확도를 나타내었다.

IV. 실험 결과

4.1 모델별 구현 결과

본 절에서는 MLP를 이용한 전력 분석 공격 모델과 SVM을 이용한 전력 분석 공격 모델의 구현 결과를 알아보고, 구현 결과에 따른 두 모델 간의 효율성을 비교 분석한다. MLP는 딥 러닝 기술의 일환으로서 사용자가 어떠한 학습 데이터를 입력했을 때, 입력된 학습 데이터 안에서 의미 있는 특징 점들을 스스로 추출하여 학습해 나가는 성질을 갖고 있다. 하지만 SVM은 스스로 학습 데이터에 대해 의미있는 특징 점을 추출하는 성질이 존재하지 않는다. 즉, SVM은 사용자가 데이터를 입력하기 전에 라벨과 관련이 높은 데이터 부분만을 추출하는 과정이 어느 정도 필요하다.

이러한 이유에서 본 논문에서 제안하는 두 공격 모델에 입력 파형은 서로 다르다. MLP 공격 모델에 사용되는 입력 파형 샘플은 암호 라운드가 시작되기 전 AddRoundKey 함수 전체와 첫 번째 라운드 SubBytes 함수 전체에 해당하는 부분이다. 반면, SVM 공격 모델에서는 첫 번째 라운드 SubBytes 부분의 특정 한 바이트 연산에 해당하는 샘플이 사용되며, 여기서 특정 바이트 부분은 공격자가 찾고자 하는 비밀 키 한 바이트와 연관된 부분을 의미한다.

4.1.1 MLP 공격 모델 실험결과

AES-128 암호 라운드 파형 10,000개를 CW-Lite 측정 도구를 통해 수집하였으며, 파형에 대한 전처리 과정 없이 암호 라운드가 시작하기 전 AddRoundKey 함수 부분의 파형 샘플과 첫 번째 라운드 SubBytes 함수 부분의 파형 샘플로 총 2,880개의 샘플을 그대로 사용하였다. 또한, 수집한 10,000개의 파형에서 7,000개는 학습 셋으로 3,000개는 테스트 셋으로 서로 데이터가 겹치지 않게 나누어 사용하였다.

다음 Fig. 8은 AES-128 암호 모듈의 첫 번째 바이트 비밀 키를 알아내기 위해 첫 번째 라운드 SubBytes 중간 결과의 첫 번째 바이트를 라벨로 학습을 진행한 결과이며, 1,000epoch으로 반복하여 학습하였다. 여기서 epoch이란, 학습 데이터에 대하여 얼마만큼 학습을 반복할 것인지에 대한 수치를 나타낸다.

Fig. 8에서 보는 바와 같이 한 바이트 비밀 키를 찾는 정확도는 epoch를 증가시킬수록 향상되었으며 학습 데이터에 대해 1,000번 학습을 반복한 결과 정확도가 89.51%로 좋은 성능을 보이는 것을 확인할 수 있다. 또한, 학습을 더욱 반복하여 epoch 값이 8,000대로 올라갔을 때에는 96.42%까지 정확도가 올라가는 것을 확인하였다.

다음 Fig. 9는 MLP 공격 모델에서 가중치 값을 나타낸 그림이다. 가중치는 각각의 입력 노드(샘플)에 부여되어 계산이 이루어지며, 가중치 값이 클수록 해당 입력 노드가 학습에 중요한 부분이라는 것을 의미한다. Fig. 9를 보면 라벨과 직접 관련이 있는 첫 번째 라운드 SubBytes의 첫 번째 바이트 부분에서 MLP 모델의 가중치의 절대 값이 크게 나타나는 것을 확인할 수 있다. 즉, MLP 모델에서는 실제 라벨과 관련 없는 샘플을 포함한 파형이 입력되어져

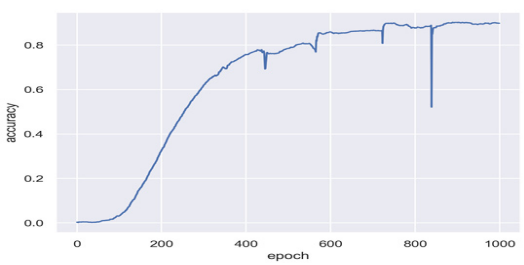


Fig. 8. Accuracy of MLP attack model according to epoch

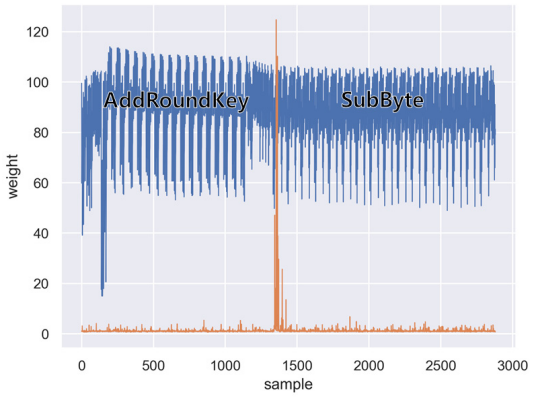


Fig. 9. Input trace and weight of MLP model

도 스스로 실제 라벨과 관련이 있는 파형의 부분만을 추출하여 학습을 수행하는 것을 알 수 있다.

4.1.2 SVM 공격 모델 실험결과

SVM 공격 모델에서도 MLP 모델과 동일하게 AES-128 암호 라운드 파형 10,000개를 사용하였으나, 입력 샘플은 첫 번째 라운드 SubBytes의 첫 번째 바이트에 해당하는 총 110샘플이다. 7,000개의 파형을 입력으로 SVM 모델 학습을 수행하고, 나머지 3,000개의 파형으로 테스트를 진행하였으며, 여러 매개변수로 설정하여 실험한 결과 C가 10,000이고 Gamma는 10일 때 79.21%의 정확도로 가장 높았다. 하지만 이는 MLP 모델 학습을 1,000epoch 반복했을 때의 정확도 89.51%에 비하면 현저히 낮은 결과이다.

딥 러닝 알고리즘인 MLP은 스스로 데이터의 특징을 찾으며 이를 효율적으로 표현하는 방법을 학습하는 표현 학습(representation learning)이기 때문에 파형과 같은 가공되지 않은 데이터에 대해서도 좋은 성능을 나타낸다. 하지만 입력 데이터가 모델의 성능을 좌우하는 머신 러닝 알고리즘 SVM에서는 가공되지 않은 데이터인 파형 데이터를 분류하는 다차원의 평면을 찾는 것이 어렵기 때문에 비교적으로 성능이 떨어지는 것으로 분석되었다.

4.2 입력 데이터에 대한 전처리 적용 결과

본 절에서는 모델에 입력되는 파형에 대한 전처리를 수행함으로써 파형에 존재하는 잡음을 제거하는 것이 학습 성능에 어떠한 영향을 주는지 알아본다. 본 논문에서는 전체 파형에 대한 평균에서 각각의 입력 파형을 차분하여 잡음을 제거하는 전처리 기법을 사용했으며 차분 파형은 다음 식으로 표현된다[13].

$$differential\ trace_i = trcace_i - \frac{1}{n}\sum_{k=0}^{n-1} trace_k$$

위 식에 의하면 각각의 파형에서 전체 파형의 평균을 차분함으로 잡음 성분이 제거되는 효과를 얻게 된다. 따라서 차분 파형은 데이터와 더욱 높은 연관성을 갖게 되어 학습의 효율성을 높일 수 있다.

Fig 10은 AES-128의 AddRoundKey 과정과 SubBytes 과정을 수행이 소비되는 원래의 전력 파형과 전처리 과정을 수행한 이후의 전력 파형을 비교하여 도시한 그림이다.

상기한 이유로 해당 전처리 기법이 적용된 파형을 입력으로 공격 모델을 학습할 경우 첫 번째 라운드

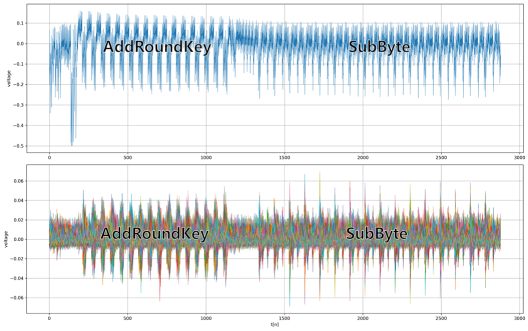


Fig. 10. Original trace vs. pre-processing traces

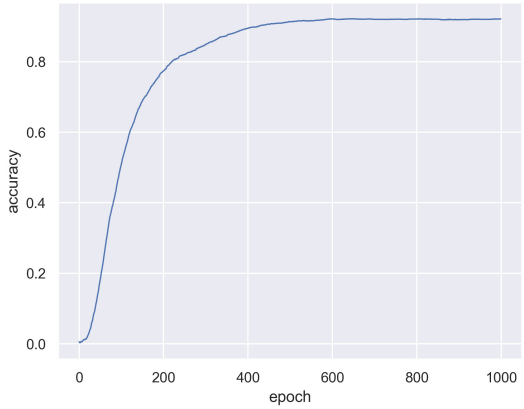


Fig. 11. Accuracy of MLP attack model with pre-processing

SubBytes의 중간 결과에 대한 특징이 더욱 확대되어 학습 성능이 더욱 높아지는 효과를 얻을 수 있다. 다음 Fig. 11은 전처리를 수행한 학습 데이터를 입력으로 MLP 공격 모델을 통해 학습한 결과이다.

Table 1은 MLP에서 입력 파형에 대한 전처리가 있는 경우와 그렇지 않은 경우의 정확도를 비교한 것으로 epoch는 모두 1,000으로 설정하였다. 전처리를 수행한 파형에 대해 학습을 진행한 결과, 전처리를 하지 않았을 때의 MLP 모델 결과보다 정확도가 5.4%정도 높은 94.91%인 것을 확인할 수 있었다. 따라서 MLP 공격 모델을 학습할 때에 파형을 그대로 입력하는 것보다 전처리 기법을 적용한 파형을 입력하여 학습하는 것이 모델 성능을 향상시키는 효과가 있음을 알 수 있다.

SVM 모델에도 동일하게 전처리된 파형을 입력하였을 때에 정확도가 81.11%로 기존의 결과보다 1.9%정도 높아진 것을 확인하였다. SVM 모델에서도 파형에 대한 전처리가 어느 정도 학습 성능을 높이는 결과를 보이지만 MLP 모델에 비해서는 그 효과가 미미한 것으로 나타난다.

Table 1. Efficiency of MLP attack model when pre-processing is applied

	Epoch	Running time	Accuracy
Original	1,000	1893.18s	89.51%
Pre-processing	1,000	1890.24s	94.91%

4.3 실험 결과 비교 분석

본 절에서는 상기한 실험 결과를 종합적으로 비교하여 분석하고자 한다. 먼저, MLP 공격 모델에 대한 실험 결과는 Table 2와 같다.

전처리를 수행하지 않은 파형에 대해 MLP 모델 학습을 수행한 경우, 파형의 유의미한 부분만을 추출하여 입력했을 때(1st Rnd 1st SB) 500epoch에서 86.63%의 정확도로 기존 샘플 파형(Initial_Add+1st Rnd all SB)이 입력되었을 때 동일한 반복에서의 정확도(75.87%) 보다 10.76% 높은 정확도를 보인다. 하지만 학습을 충분히 더 진행한 1,000epoch에서는 두 경우의 정확도가 거의 비슷하다는 것을 확인할 수 있다.

Table 2. Experimental results for the MLP attack model

	Epoch	Initial_Add +1st Rnd all SB(2880)	1st Rnd 1st SB(110)
Original	500	75.87%	86.63%
	1,000	89.51%	89.39%
Pre-processing	500	94.26%	94.01%
	1,000	94.91%	94.13%

즉, MLP 모델은 주어진 전체의 파형에서 유의미한 부분을 스스로 추출하여 학습한다는 것을 위의 결과를 통해 확인할 수 있다. 이는 또한 공격자가 라벨과 연관이 없는 샘플이 다수 포함된 파형을 입력하여 학습을 진행하여도 결국 충분한 학습의 반복이 진행되어진다면 유의미한 파형의 부분을 분석하는 과정이 자동화되어 동일한 학습 결과를 도출해낼 수 있다는 것을 의미한다.

다음 Table 3과 Table 4는 SVM 공격 모델에 대한 실험 결과이며, Table 3은 매개변수 C를 10,000으로 고정하고 Gamma 값을 바꾸어 가며 실험한 결과이고, Table 4는 매개변수 Gamma를 1로 고정하고 C값을 바꾸어 가며 실험한 결과이다. 여기서 매개변수 C는 클수록 오차허용을 최소화하여 학습 값들이 정확하게 분류될 수 있게 하고, 매개변수 Gamma는 클수록 학습 과정 중간의 초평면에서 가까운 학습 값들을 최적의 초평면을 찾기 위한 계산에 사용하게 된다.

표에서 보는 바와 같이 전처리된 파형을 SVM 공격 모델을 통해 학습하면 전처리를 적용하지 않았을 때보다 높은 정확도를 보이나 대부분 2%이내이므로 모델 성능 향상에 비교적 큰 영향을 미치지는 못하는 것을 확인할 수 있다. 또한, 매개변수 C와 Gamma 값을 각각 10,000, 10으로 설정하였을 때가 가장 좋은 학습 결과를 나타내는 것을 확인할 수 있다.

Table 3. Experimental results for SVM attack model(C=10,000)

C	Gamma	Original	Pre-processing
10,000	0.1	78.93%	80.96%
10,000	1	78.56%	80.51%
10,000	10	79.21%	81.11%

Table 4. Experimental results for SVM attack model (Gamma=1)

C	Gamma	Original	Pre-processing
1,000	1	78.91%	80.93%
10,000	1	78.56%	80.51%
100,000	1	78.56%	80.51%

V. 결 론

전력 분석 공격은 공격자의 높은 전력 분석 능력도 필요하지만 측정 도구, 암호 모듈, 디바이스에 따라 비밀 키를 찾아내는 데까지 상당히 많은 계산 부하가 요구된다. 본 논문에서는 이러한 기존 전력 분석 공격의 어려움을 극복하기 위해 딥 러닝 기술 중 하나인 MLP 알고리즘을 활용한 공격 모델을 제시하였으며, AES-128 암호 모듈의 비밀 키를 복구하기 위한 실험을 진행하였다. 제안하는 전력 분석 모델에서는 학습과 분류를 통해 16바이트의 비밀 키를 한 바이트씩 순차적으로 찾는 실험을 수행하였다.

실험 결과, 딥 러닝 기법의 일종인 MLP 알고리즘을 적용하고 입력 파형에 대한 전처리 과정을 거칠 경우, 약 94%이상의 정확도로 비밀 키 바이트를 찾아낼 수 있었다. 반면, SVM 알고리즘을 사용할 경우에는 전처리 과정이 있다고 하더라도 약 81%정도의 비교적 낮은 정확도로 비밀 키를 추출함을 확인하였다. 이를 통해 딥 러닝 기법을 이용한 학습 기법이 파형 데이터의 특징 점을 더욱 잘 추출할 수 있어 전력 분석 공격 모델로 적합하다는 것을 확인하였다.

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.

[2] S. Mangard, E. Oswald, and T. Poop, "Power analysis attacks: Revealing the secrets of smart cards," Springer, 2008.

[3] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," Psychological Review, Vol. 65, No. 6, 1958.

[4] R. Collobert and S. Benjio, "Links between perceptrons, MLPs and SVMs," Proceedings of the twenty-first international conference on Machine learning, ICML'04, p. 23, 2004.

[5] Federal Information Processing Standards Publication (FIPS 197), "Advanced Encryption Standard(AES)," 2001.

[6] C. Cortes, and V. Vapnik, "Support-vector networks," Machine Learning, Vol. 20, Issue 3, pp. 273-297, 1995.

[7] T. Hofmann, B. Scholkopf, and A. J. Smola, "Kernel Methods in Machine Learning," The Annals of Statistics, Vol. 36, No. 3, pp. 1171-1220, 2008.

[8] P. Kocher, J. Jaffe, and B. Jun, "Differential Power Analysis," CRYPTO'99, 1999.

[9] E. Brier, C. Clavier, and F. Olivier, "Correlation Power Analysis with a Leakage Model," CHES'04, LNCS 3156, pp. 16-29, 2004.

[10] S. Chari, J. R. Rao, and P. Rohatgi, "Template Attacks," CHES'02, LNCS 2523, pp. 13-28, 2002.

[11] W. Schindler, K. Lemke, and C. Paar, "A Stochastic Model for Differential Side Channel Cryptanalysis," CHES'05, LNCS 3659, pp. 30-46, 2005.

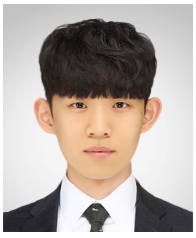
[12] Z. Martinasek, and V. Zeman, "Innovative method of the power analysis," Radioengineering, Vol. 22, No. 2, pp. 589-594, 2013.

[13] Z. Martinasek, J. Hajny, and L.

Malina, "Optimization of power analysis using neural network," CARDIS'13, LNCS 8419, pp. 94-107, 2014.
[14] ChipWhisperer® - NewAE Technology Inc., "chipwhisperer," Available at http://newae.com/tools/chipwhisperer/, 2017.
[15] G. Hospodar, B. Gierlichs, E. D. Mulder, I. Verbauwhede, and J. Vandewalle, "Machine learning in side-channel analysis: a first study," Journal of Cryptographic Engineering, Vol. 1, No. 4, pp.293 - 302, 2011.

〈저자소개〉

권 홍 필 (Hongpil Kwon) 학생회원
2018년 2월: 호서대학교 정보보호학과 학사
2018년 3월~현재: 호서대학교 정보보호학과 석사 과정
〈관심분야〉 부채널 공격, 인공지능 보안, 암호학



배 대 현 (Daehyeon Bae) 학생회원
2017년 3월: 호서대학교 컴퓨터정보공학부 입학
2017년 3월~현재: 호서대학교 컴퓨터정보공학부 학부과정
〈관심분야〉 암호학, 부채널 공격, 인공지능 보안

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 컴퓨터정보공학부 교수
2013년 1월~현재: 한국정보보호학회 상임부회장
2009년 1월~현재: 한국산학기술학회 이사
〈관심분야〉 정보보호, 네트워크 보안, 부채널 공격