부채널 분석을 이용한 딥러닝 네트워크 공격 동향

https://www.youtube.com/watch?v=D6LfOIECdLc

부채널 분석을 이용한 딥러닝 네트워크 공격 동향

- **최근 4차 산업혁명을 선도하는 기술 중 하나는 인공지능이며 딥러닝의 연구가 활발하게 이루어지고 있음**
  → 자율 주행, 이미지 생성, 가상 음성 생성 등 다양한 기술에 활용됨

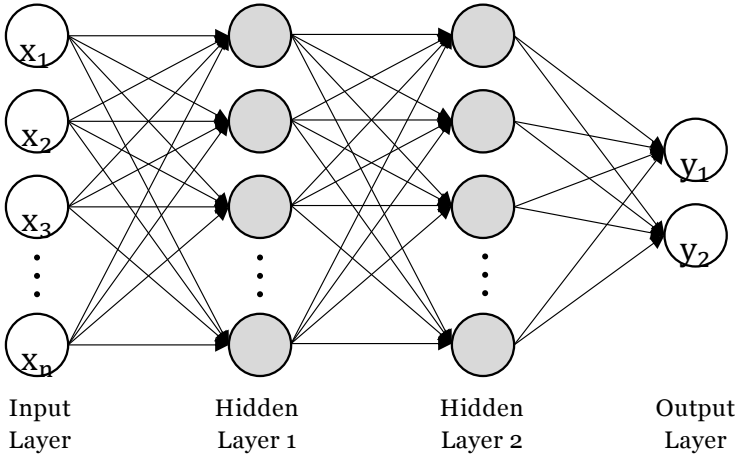
- **공격자가 딥러닝 가속기에 접근하거나 이를 탈취하는 경우, 부채널 분석을 통해 가속기의 내부 비밀 정보인 가중치나 편향 값을 복구할 수 있음**
  → 딥러닝 네트워크에 대한 부채널 공격기법과 그에 대한 대응책 필요

딥러닝(Deep Learning)

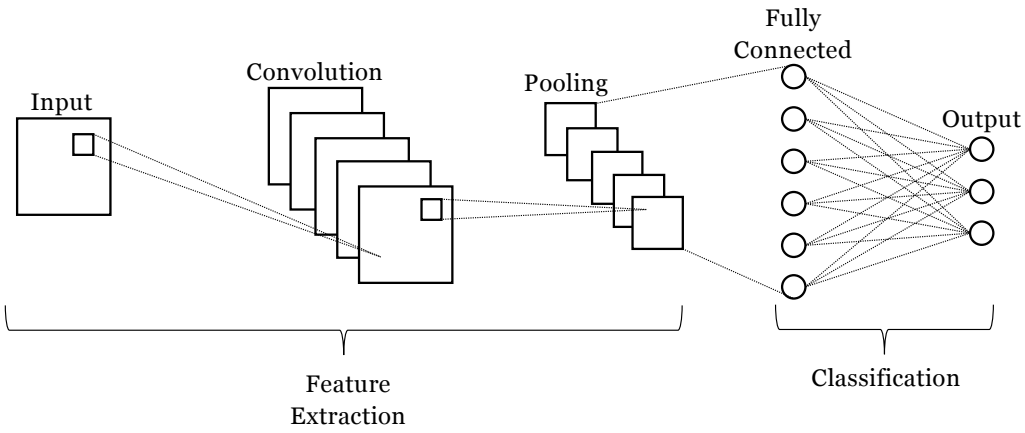
딥러닝은 인공지능의 한 분야로, 인공 신경망을 기반으로 한 머신러닝 알고리즘으로 다층 인공신경망을 사용하여 입력 데이터로부터 복잡한 패턴을 학습하고 이를 통해 데이터를 분석하고 예측함

- **DNN (Deep Neural Network)**
  - 기존 ANN (Artificial Neural Network) 문제를 해결하기 위해 은닉층 확대
  - 2개 이상의 은닉층으로 학습결과를 향상 시킴(보통 Deep Learning은 3개 이상)

- **CNN (Convolution Neural Network)**
  - 영상 및 이미지 처리에 많이 활용되는 합성곱을 이용하는 인공신경망 기술
  - 합성곱 필터를 이용하여 연산을 수행 (작은 필터들이 이미지 픽셀을 이동하며 특징 값들을 찾아 합성곱 수행)
  - 연산 결과를 다음 계층으로 보내 적은 수의 가중치로 이미지 처리를 할 수 있음



DNN architecture



CNN architecture

부채널 분석(Side Channel Analysis)

부채널 분석은 전자기기가 동작할 때 발생하는 소비 전력, 전자파, 시간 등의 부채널 정보를 이용하여 내부 비밀 정보를 복원하는 기술로 복원 방법은 단순 전력 분석 (Simple Power Analysis), 상관 전력 분석 (Correlation Power Analysis), 차분 전력 분석 (Differential Power Analysis)이 있음

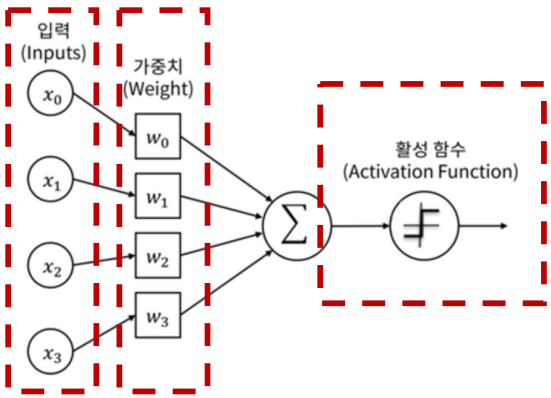
- **SPA**
  - 단일 파형으로 분석하는 방법이며, 공격자는 공격 지점의 연산 과정과 구현 방법을 정확히 알아야함
  - 예) RSA에서 square and multiply 연산을 수행할 경우, data가 1이면 square and multiply 연산, data가 0이면 square 연산만을 수행
    → 어떠한 명령어가 수행되는지 알아야 가능한 부채널 분석 방법
- **CPA**
  - 다수의 파형을 이용하는 통계적 분석을 통해 내부 비밀정보를 복원하는 방법
  - 전자기기에서 수집한 부채널 정보와 공격자가 추측한 중간값과의 상관계수를 계산하여 가장 유의미한 결과를 도출한 중간값을 내부 비밀 정보로 결정
- **DPA**
  - 해밍웨이트 모델로서, 수행하는 연산, 데이터, 노이즈, 기본 소비전력에 의해 총 전력이 결정
  - 전력 파형을 얻을 경우 특정 연산 지점에 대한 세부적인 값 차이는 데이터에 의한 차이이며 그래프의 위상 차이는 연산의 종류에 따라 결정됨
    → 즉, 전력 소비가 높은 연산이라면 해밍웨이트가 높음

딥러닝 네트워크에 대한 부채널 공격 기법

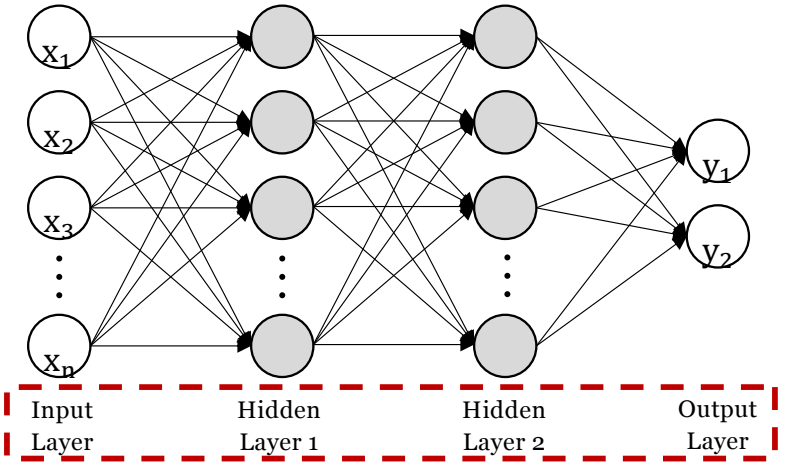
딥러닝 네트워크에 대한 부채널 공격은 주로 내부 파라미터 및 내부 구조를 복원하는 과정으로 나뉨

- **내부 파라미터 복원**
  → 네트워크를 형성하는 입력값, 가중치, 활성함수 등 다양한 요소 대상

- **내부 구조 복원**
  → 네트워크 층의 수, 뉴런의 수 등 요소 대상



딥러닝 네트워크 내부 파라미터



딥러닝 네트워크 내부 구조

내부 파라미터 복원

- **내부 파라미터 복원**
  → 네트워크를 형성하는 입력값, 가중치, 활성함수 등 다양한 요소 대상

최근 연구 동향으로 Maji et al.은 그레이 박스 환경에서 CNN과 BNN 모델의 가중치, 편향, 입력값을 복구하는 기술을 제안
→ 측정된 파형의 SNR(signal to noise ratio) 및 모델의 복잡성을 최소화할 수 있음
→ 그레이 박스 환경은 공격 대상 네트워크 구조를 사전에 알고 있다는 한계가 있음

- **내부 구조 복원**
  → 네트워크 층의 수, 뉴런의 수 등 요소 대상

최근 관련 연구 동향으로는 Yoshida et al.은 네트워크 내부 구조를 알고 있는 상황에서 Chain CPA(Chain Differential Cryptanalysis with chosen plaintext)라는 선택 평문 공격을 이용하여 블록 암호의 내부 구조를 파악하고 암호화 키를 찾아내는 공격 기법을 활용하여 가중치 정보를 복구하는 기술 제안

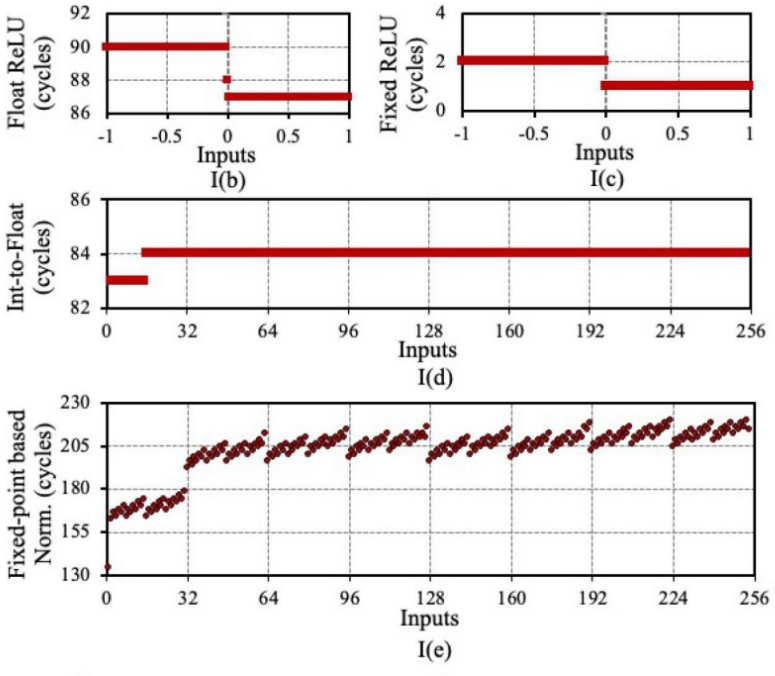
→ 단순전력 분석의 정확도가 낮다는 한계가 있음

딥러닝 네트워크에 대한 부채널 공격 기법

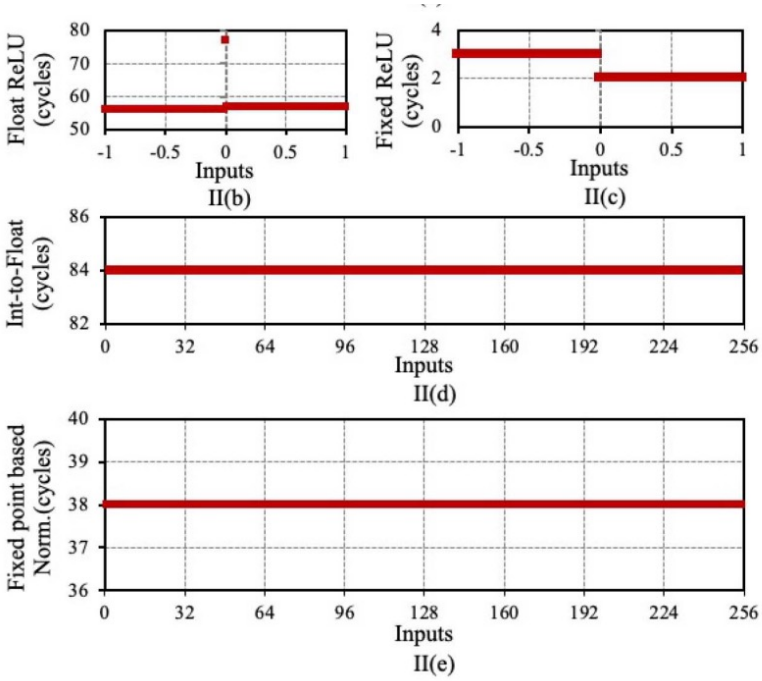
- **Maji et al. SPA 입출력 값 및 파라미터 복구**
- **그레이 박스 환경에서 CNN과 BNN 모델의 가중치, 편향, 입력값을 복구하는 기술을 제안**
- ARM, RISC-V 프로세서를 사용

→ 측정된 파형의 SNR(signal to noise ratio) 및 모델의 복잡성을 최소화할 수 있음

- 사용된 프로세서에 따라 동일 연산도 다른 파형을 가짐



ARM



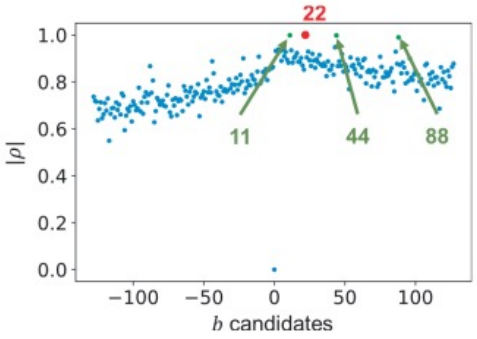
RISC-V

딥러닝 네트워크에 대한 부채널 공격 기법

- **Yoshida et al. Chain CPA**
  - CPA는 일반적으로 시스톨릭 배열의 곱셈 – 누산 연산 구조를 사용하며 연산이 곱셈으로만 된 경우보다 곱셈과 덧셈으로 구성된 경우 더 정확도가 높음 (PE는 시스톨릭 배열의 요소)
  - CPA의 첫번째 연산에서 곱셈과 영덧셈이 연산되기 때문에 정확도가 낮고 두번째 연산부터 첫번째 연산의 결과를 레지스터에 저장하고 곱셈에 더해주기때문에 정확도가 높아짐

$$
\begin{aligned}
c_{reg}(PE_{11}) &= a_{11} \times b_{11} + 0 && (t = 1)\\
c_{reg}(PE_{12}) &= a_{12} \times b_{21} + c_{reg}(PE_{11})\\
&= a_{12} \times b_{21} + a_{11} \times b_{11} && (t = 2)\\
c_{reg}(PE_{13}) &= a_{13} \times b_{31} + c_{reg}(PE_{12})\\
&= a_{13} \times b_{31} + a_{12} \times b_{21} + a_{11} \times b_{11} && (t = 3)\\
& && (4)
\end{aligned}
$$

  - Chain CPA는 일반 CPA의 두번째 연산에서 가장 높은 상관관계를 갖는 선택 예를 들어 아래 그림의 4개로 추려서 일반 CPA보다 계산 비용을 몇배는 줄일 수 있음



딥러닝 네트워크에 대한 부채널 공격 대응 기술

딥러닝 네트워크에 대한 부채널 공격 대응 기술의 대표적인 예로 셔플링(Shuffling)과 마스킹(Masking) 기법 존재

- **셔플링(Shuffling)**
  → 네트워크를 형성하는 입력값, 가중치, 활성함수 등 다양한 요소 대상
  → 딥러닝 네트워크에 대한 부채널 공격 대응기술의 대표적인 기술로 데이터나 가중치 등 패턴을 무작위로 섞어 공격자가 모델의 내부 정보를 추출하기 어렵게 함

- **마스킹(Masking)**
  → 네트워크 층의 수, 뉴런의 수 등 요소 대상
  → 신경망의 파라미터와 중간 계산 결과에 임의성을 추가하여 공격자가 내부 정보를 추출하기 어렵게 만듦

감사합니다.