논문 21-46-07-02 The Journal of Korean Institute of Communications and Information Sciences '21-07 Vol.46 No.07
https://doi.org/10.7840/kics.2021.46.7.1109

블록암호 DES의 신경망 기반 평문 복구 공격에 대한 재고찰

권 수 진•, 임 형 신*, 강 주 성**, 염 용 진°

Revisiting Cryptanalysis of Neural Plaintext Recovery Attack of DES

Sujin Kwon•, Hyoungshin Yim*, Ju-Sung Kang**, Yongjin Yeom°

요 약

최근 신경망이 발달함에 따라 신경망을 이용한 암호분석 연구가 활발하게 진행되고 있다. 신경망 기반 암호분석은 키 복구 공격과 평문 복구 공격이 있다. 그중 평문 복구 공격은 암호 키를 찾지 않고 암호문에 대응하는 평문을 복구하는 공격이다. 2012년에 신경망 기반 DES의 평문 복구 공격이 성공적이라는 연구 결과가 발표되었다. 이는 연속함수와 가측함수의 신경망 근사 가능성을 보여주는 Universal Approximation Theorem(UAT)에 기반한다. 하지만 이산함수인 DES의 근사에 필요한 신경망 크기에 대한 정량적 분석의 부재와 학습 알고리즘의 수렴성에 대한 과도하게 낙관적인 해석으로 연구 결과를 신뢰할 근거가 부족하다. 본 논문에서는 신경망의 연속함수와 이산함수에 대한 근사 가능성을 정량적으로 살펴보고 2012년에 제안된 DES의 평문 복구 공격을 재연하고 결과를 비교한다. 마지막으로 부분 라운드 DES의 평문 복구 가능성을 분석하여 신경망 기반 평문 복구 공격의 한계를 분석한다.

키워드 : 신경망, 블록암호, 암호분석, DES, 평문 복구 공격
Key Words : Neural network, Block cipher, Cryptanalysis, DES, Plaintext recovery attack

ABSTRACT

Research on cryptanalysis of block ciphers using neural network has been actively conducted encouraged by the recent development of neural networks. The cryptanalysis based on neural network includes key recovery attack and plaintext recovery attack. A plaintext recovery attack is an attack that recovers the plaintext corresponding to a given ciphertext instead of retrieving the key. In 2012, a paper claimed that the plaintext recovery attack of block cipher DES using neural network is feasible. The assertion was based on the universal approximation theorem(UAT) which shows the approximation possibility of continuous or measurable functions using neural networks. However, the lack of quantitative analysis on the required size of the neural network for approximation of DES as a discrete function and the extremely optimistic convergence of the learning algorithm cannot reliably explain their results. In this paper, we investigate the quantitative analysis of

※ 이 성과는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2021M1A2A2043893)
• First Author : Kookmin University Department of Financial Information Security, tnwls1595@kookmin.ac.kr, 학생회원
° Corresponding Author : Kookmin University Department of Information Security, Cryptology, and Mathematics, salt@kookmin.ac.kr, 종신회원
* Kookmin University Department of Financial Information Security, kuunh2@kookmin.ac.kr, 학생회원
** Kookmin University Department of Information Security, Cryptology, and Mathematics, jskang@kookmin.ac.kr, 종신회원
논문번호 : 202101-007-A-RE, Received January 4, 2021; Revised April 19, 2021; Accepted April 26, 2021

neural approximations for continuous function and discrete function and implement experiments the plaintext recovery attack as proposed in 2012 and compare the result. Finally, we analyze the plaintext recovery attack of round-reduced DES, which shows the limitation of the plaintext recovery attack based on neural network.

Ⅰ. 서 론

블록 암호(block cipher)는 고속 암호화를 위해 널리 사용되는 대칭키 알고리즘으로 암호화와 복호화는 공통으로 사용되는 암호 키를 소유한 사용자만 수행할 수 있다. 블록암호에 대한 공격에는 키 복구 공격(key recovery attack)과 평문 복구 공격(plaintext recovery attack)이 있다. 키 복구 공격은 평문과 이에 대응되는 암호문 쌍을 이용하여 암호 키를 찾는 것을 목표로 한다. 반면, 평문 복구 공격은 일반적으로 고정된 암호 키를 사용하는 장치(device)에서 주어진 암호문에 대응되는 평문을 복구하는 것을 목표로 한다. 따라서 평문 복구 공격은 미지의 암호 키가 설정된 암복호화 알고리즘과 동일한 기능을 갖는 알고리즘을 만들고 이로부터 평문을 복구하는 것을 의미한다. 키 복구 공격과 평문 복구 공격의 개념을 비교하면 [그림 1]과 같다.

최근 신경망(neural network)을 이용하여 암호를 분석하는 연구가 활발히 진행되고 있으며, 신경망 기반 키 복구 공격과 평문 복구 공격에 관한 연구도 진행되고 있다[1-6,9]. 신경망을 이용한 키 복구 공격[2]은 기존 암호분석(cryptanalysis)에 사용된 차분 특성을 활용하였다. 반면, 신경망 기반 평문 복구 공격[3-6]에서 신경망은 복호화 알고리즘을 근사하는 역할을 한다.

UAT(Universal Approximation Theorem)[7]에 따르면 신경망은 구간에서 정의된 연속함수를 근사할 수 있다. 반면, 신경망으로 이산(discrete)함수를 효율적으로 근사할 수 있다는 연구 결과는 찾아보기 어렵다. 따라서 신경망으로 이산함수인 복호화 알고리즘을 근사하기엔 어려움이 있어, 신경망을 이용한 평문 복구 공격의 성공을 주장하는 기존 결과에 대한 객관적 검증이 필요하다.

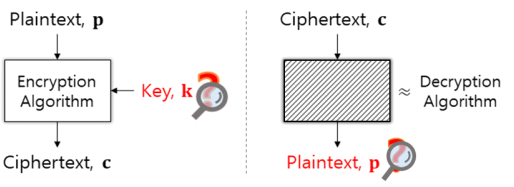

그림 1. (좌) 키 복구 공격, (우) 평문 복구 공격
Fig. 1. (left) key recovery attack, (right) plaintext recovery attack

본 논문에서는 기존에 제안된 신경망 기반 평문 복구 공격[3-6]에 대한 문제점을 지적하고, 2012년에 발표된 신경망 기반 DES 평문 복구 공격 연구[3]의 실험 재연과 이론적인 분석을 통해 연구 결과를 검증한다. 2장에서는 신경망 기반 블록 암호의 암호분석에 관한 연구 동향을 살핀다. 3장에서는 연속함수와 이산함수에 대한 신경망 근사 실험을 통해 UAT를 기반으로 이산함수를 근사하기엔 한계가 있음을 보인다. 4장에서는 2012년에 발표된 신경망 기반 DES의 평문 복구 공격을 재연하고 결과를 비교함으로써 공격의 어려움을 보인다. 5장에서는 평문 복구 공격에 신경망의 활용 가능성을 분석하기 위해 부분 라운드 DES에 대한 신경망 기반 평문 복구 공격을 시도하고 6장에서는 결론을 정리한다.

Ⅱ. 관련 연구

2.1 UAT를 이용한 함수의 신경망 근사

1989년 K. Hornik에 의해 증명된 UAT는 충분한 수의 노드(node)로 구성되고 하나의 은닉층(hidden layer)을 가진 신경망은 원하는 정확도로 임의의 가측함수(measurable function)를 근사할 수 있음을 보여준다[7]. UAT는 유한 폐구간에 정의된 가측함수가 연속함수의 일부분 다항식으로 근사될 수 있음을 증명한 Stone-Weierstrass 정리[8]에 기초한다.

< Stone-Weierstrass 정리[8] >

Let $\mathbf{A}$ be an algebra of real continuous functions on a compact set K. If $\mathbf{A}$ separates point on K and if $\mathbf{A}$ vanishes at no point of K, then the uniform closure $\mathbf{B}$ of $\mathbf{A}$ consists of all real continuous functions on K.

2.2 신경망을 이용한 암호분석 동향

신경망을 이용한 키 복구 공격에 관한 연구 결과는 다음과 같다. 2019년 A. Gohr는 차분 특성과 합성곱 신경망(convolution neural network)을 사용하여 라운드 수를 줄인 Speck32/64의 키 복구 공격을 하였다[2].

특정 차분을 갖는 평문과 이에 대응하는 부분 라운드 Speck32/64의 암호문 쌍과 랜덤한 데이터 쌍을 구별하도록 신경망을 훈련하였고, 그 결과 기존 차분분포표(differential distribution table)를 사용하였을 때보다 더 높은 정확도로 8-라운드까지 구별함을 실험적으로 보였다. 2020년 J. So는 (평문, 암호문) 쌍을 이용하여 DNN(deep neural network) 모델을 훈련하였고, SDES(simplified DES), Speck32/64, Simon32/64의 암호 키를 복구하는 연구 결과를 제시하였다[9]. 이때, 키 공간을 64개의 ASCII 문자로 제한을 둔 경우에 한하여 암호 키 복구에 성공하였다.

신경망을 이용한 평문 복구 공격에 관한 연구 결과들은 다음과 같다. 2012년 M. M. Alani는 신경망을 이용하여 DES 및 TDES(Triple DES)의 암호문으로부터 평문을 복구하는 연구를 제시하였다[3-4]. 이때, 2^{12}개의 (평문, 암호문) 쌍을 사용하였다. 2018년에 제시된 연구는 M. M. Alani가 제시한 신경망 구조를 활용하여 AES 암호문에 대한 평문을 약 40%의 확률로 복구한 결과를 제시하였다[5]. 2019년 S. Fan은 오차역전파(backpropagation) 알고리즘을 수정하여 적용하였을 때 DES의 암호문의 평문을 약 10%의 오차율로 복구한다는 결과를 제시하였다[6].

Ⅲ. 신경망을 이용한 연속함수와 이산함수의 근사

신경망을 이용한 평문 복구 공격 연구[3-6]에서는 UAT에 근거하여 신경망이 블록암호의 복호화 알고리즘을 근사할 수 있다고 간주하지만, UAT는 이산함수에 대한 근사 가능성을 다루지 않는다. 그러므로 신경망을 이산함수 근사에 활용할 경우, 신경망이 함수를 표현하기에 충분한지와 이산함수를 근사시키기 위한 학습 알고리즘이 존재하는지를 고려해야 한다. 하지만 이에 관한 연구의 부재로 인해 연속함수에 관한 근사 결과에 의존하여 신경망을 암호분석에 활용하고 있다. 따라서 본 장에서는 신경망이 연속함수와 이산함수를 모두 근사할 수 있는지에 대한 기초 실험을 수행하고 결과를 분석한다.

3.1 실험 설계

본 실험에서는 하나의 은닉층을 가지는 완전 연결 신경망(fully-connected neural network)을 사용하였으며, 노드의 개수를 변경하여 각 함수에 대해 7 개의 실험을 진행하였다. 실험에 사용한 데이터 및 하이퍼파라미터는 [표 1]에 작성되어 있다[10]. 실험에는 Python 3.7.5 버전과 Keras 2.3.1 버전이 사용되었다.

3.1.1 함수 설정

실험에 사용한 연속함수 $f:\mathbb{R}\rightarrow\mathbb{R}$는 아래와 같이 정의역과 공역이 실수인 8차 다항식이다.

$$f(x) = 0.1\times x^8 + 0.2\times x^7 + 0.1\times x^5 + x^3. \quad (1)$$

이산함수 g는 암호에서 많이 사용하는 구성 요소로 선정하였다. 이산함수 $g:\{0,1\}^8\rightarrow\{0,1\}$는 8-비트 입력에 1-비트 출력값을 가지는 8차 부울함수(boolean function)이며, 식은 다음과 같다.

$$g(\overline{\boldsymbol{x}}) = A \bullet (\overline{\boldsymbol{x}})^{-1} \oplus 1. \quad (2)$$

이산함수 g에 대한 설명은 다음과 같다. 입력 $\overline{\boldsymbol{x}}$는 이진 벡터 $(x_7, x_6, x_5, x_4, x_3, x_2, x_1, x_0)$로 표현되며, $x_i \in \{0,1\}$이다. 정수로 표현할 경우, $\overline{\boldsymbol{x}} = \sum_{i=0}^{7} x_i 2^i$이다. 즉, g의 정의역은 $\{0,1,...,255\}$이다. A는 이진 상수 벡터 $(1,1,1,1,0,0,0,1)$이며, $\bullet$은 벡터 내적을 의미한다. $(\overline{\boldsymbol{x}})^{-1}$는 8-비트 $\overline{\boldsymbol{x}}$를 유한체 $\mathrm{GF}(2^8)$상의 원소로 해석하여 역원을 계산한 것이다.

암호학적 관점으로 이산함수 g를 설명하면, 이산함수 g는 블록암호 AES에 사용되는 S-box의 출력(8-비트) 중 최하위비트(least significant bit)를 출력하는 것이다.

3.1.2 데이터 세트 생성

연속함수의 정의역은 실수이지만 이산함수의 정의역은 $\{0,1,...,255\}$이므로, 데이터 세트의 개수는 이산

표 1. 데이터 세트 및 하이퍼파라미터
Table 1. Data set and hyperparameter

연속함수	$f(x) = 0.1\times x^8 + 0.2\times x^7 + 0.1\times x^5 + x^3$						
이산함수	$g:\{0,1\}^8\rightarrow\{0,1\}$						
입력 데이터 세트	연속함수	250개의 $(x, f(x))$, x는 $-1\le x\le 1$인 실수					
	이산함수	250개의 $(\overline{\boldsymbol{x}}, g(\overline{\boldsymbol{x}}))$, $\overline{\boldsymbol{x}}\in\{0,1\}^8$					
신경망	완전 연결 신경망						
손실함수	평균제곱오차						
활성화 함수	연속함수	시그모이드 및 선형함수					
	이산함수	시그모이드					
에폭	1,000						
실험 No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
노드	100	300	500	700	1,000	5,000	10,000

함수의 정의역 크기에 맞춰 설정하였다.

이산함수는 $\{0,1,...,255\}$의 원소 중 랜덤하게 250개 추출하여 $(\overline{\boldsymbol{x}}, g(\overline{\boldsymbol{x}}))$을 얻은 후, 150개를 훈련 데이터(train data), 50개를 검증 데이터(validation data), 나머지 50개를 테스트 데이터(test data)로 사용하였다. 정의역의 크기가 256이므로 250개를 넘는 훈련 데이터를 사용하는 것은 학습의 의미를 부여하기 어렵다.

연속함수는 $[-1,1]$ 내에 있는 실수를 랜덤하게 250개 추출하여 $(x, f(x))$을 얻은 후 이산함수와 동일한 개수로 데이터 세트를 구성하였다.

3.1.3 신경망 모델

완전 연결 신경망을 사용하였으며, 손실함수(loss function)는 평균제곱오차(mean squraed error, MSE)를 사용하였다. 활성화함수(activation function)로는 시그모이드(sigmoid)를 사용하였다. 이때, 연속함수의 경우 최종 출력이 실수이므로, 출력층(output layer)의 활성화함수는 선형함수(linear function)를 사용하였다.

다양한 에폭(epoch)으로 실험을 진행해본 결과, 높은 근사 정확도를 가지는 최소 에폭의 수가 1,000임을 확인하였다. 따라서 본 논문의 실험에서는 에폭의 수를 1,000으로 설정하였다.

실험에서 사용한 이산함수 g는 부울함수이므로, 신경망의 노드 수를 설정할 때 샤논(Shannon)의 정리를 고려해야 한다. 샤논의 정리는 부울함수를 계산하는 데 필요한 논리 회로(logic gate)의 최소 개수를 제시한다[11]. 이 정리에 따르면, 실험에서 사용한 이산함수 g를 계산하기 위해 적어도 $\Omega(2^8/8) \approx 32$ 이상의 논리 회로를 사용해야 한다. 또한, UAT에는 함수 근사에 사용되는 신경망의 노드 수가 명시되어 있지 않다. 따라서 본 실험에서는 샤논의 정리와 UAT를 고려하여 노드 수를 설정하였다. 노드 수는 100 개, 300 개, 500 개, 700 개, 1.000 개, 5,000 개, 10,000 개로 설정하여 각 노드 수를 적용한 실험을 진행하였다. 이때, 노드 수에 따라 실험 1, 실험 2, 실험 3, 실험 4, 실험 5, 실험 6, 실험 7이라고 한다.

> **< Shannon 정리[11] >**
>
> There exists a boolean function $h : \{0,1\}^n \rightarrow \{0,1\}$, on n variables, such that any circuit to compute h requires at least $\Omega(2^n/n)$ logic gates.

3.2 실험 결과

실제 함수와 신경망이 예측한 함수의 L_2노름(norm) 값의 제곱 값으로 신경망의 근사 정도를 확인하였고, 이를 오차로 명한다. 계산 식은 다음과 같다.

$$\| \boldsymbol{y} - \hat{\boldsymbol{y}} \|_2^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{3}$$

이때, n은 테스트 데이터 50개를 의미하며, y_i는 테스트 데이터의 입력에 대한 레이블(label), $\hat{y}_i$는 테스트 데이터의 입력에 대한 신경망의 예측(predict)값이다.

이산함수의 경우, 출력층의 활성화함수는 시그모이드이므로 신경망의 출력값은 $[0,1]$ 범위의 실숫값이다. 따라서 오차를 계산하기 전 0 또는 1로 반올림을 해준다.

3.2.1 연속함수에 대한 근사 가능성

신경망이 예측한 값과 실제 레이블의 오차는 [표 2]와 같다. 오차의 편향을 최소화하기 위해 각 실험을 50번 반복하여 오차의 평균을 계산하였다.

실험 1-7에서 신경망은 낮은 오차로 연속함수 f를 근사하였으며, 5,000개의 노드를 사용한 실험 6에서 가장 낮은 오차를 보였다. [그림 2]는 실험 6에 대한 훈련된 신경망이 예측한 함수 $\hat{f}$을 함수 f와 함께 나

표 2. 연속함수에 대한 오차 결과
Table 2. Results of error for continuous function

실험 No.	1	2	3	4	5	6	7
오차	0.115	0.116	0.114	0.114	0.142	0.005	0.127

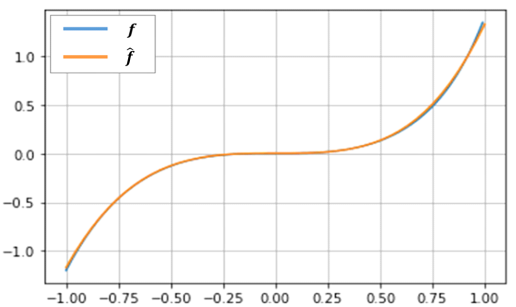

그림 2. 연속함수 f 및 신경망이 예측한 함수 $\hat{f}$
Fig. 2. Continuous function f and a function predicted by a neural network $\hat{f}$

타낸 것이며 실제 함수와 유사하게 예측하는 것을 시각적으로 볼 수 있다. 신경망은 충분한 노드 수를 사용하였을 경우 연속함수에 근사할 수 있는 UAT 정리를 만족함을 실험적으로 확인하였다.

3.2.2 이산함수 근사 가능성 분석

신경망이 예측한 값과 레이블의 오차는 [표 3]과 같으며, 각 실험을 50번 반복하여 계산한 오차의 평균값이다.

각 실험의 오차는 25에 가까우며, 50 개의 테스트 데이터 중 25 개의 테스트 데이터가 잘못 예측되었음을 의미한다. 이는 0과 1을 무작위로 추출하는 랜덤한 부울함수를 근사식으로 간주했을 때의 오차와 유사하다. 즉, 훈련된 신경망이 주어진 부울함수 g에 전혀 근사하고 있지 않음을 알 수 있다. 에폭 수와 노드 개수를 변경한 모든 실험에서 정확도가 개선되지 않음을 확인하였다.

표 3. 이산함수에 대한 오차 결과
Table 3. Results of error for discrete function

실험 No.	1	2	3	4	5	6	7
오차	24.6	25.3	25.5	25.4	25.9	25.5	25.3

3.2.3 연속함수 및 이산함수 근사 가능성 비교

본 항에서는 연속함수 및 이산함수의 실험 결과를 기반으로 신경망에 의한 이산함수의 근사 가능성을 고찰하고자 한다. [그림 3]은 연속함수와 이산함수의 근사 가능성 실험에 대한 학습 손실과 검증 손실의 그래프를 나타낸다.

연속함수의 경우 학습이 반복될수록 훈련 손실과 검증 손실이 0에 수렴하여 학습의 성능이 높게 평가된다. 반면, 이산함수의 경우 샤논의 정리를 고려하여 신경망을 구축하였음에도 검증 손실은 연속함수와 달리 증가하는 양상을 보여 학습이 되지 않음을 의미한다.

예제에서 사용한 이산함수는 암호의 논리로 널리 사용되는 8-비트 S-box 치환의 출력 중 1-비트를 나타낸다. 국제표준 AES 암호의 경우 한번의 암호화를 위해 S-box를 160개 이상 적용하는데, 실험결과는 S-box 한 개의 하위 1-비트도 신경망 근사가 어려움을 보여준다.

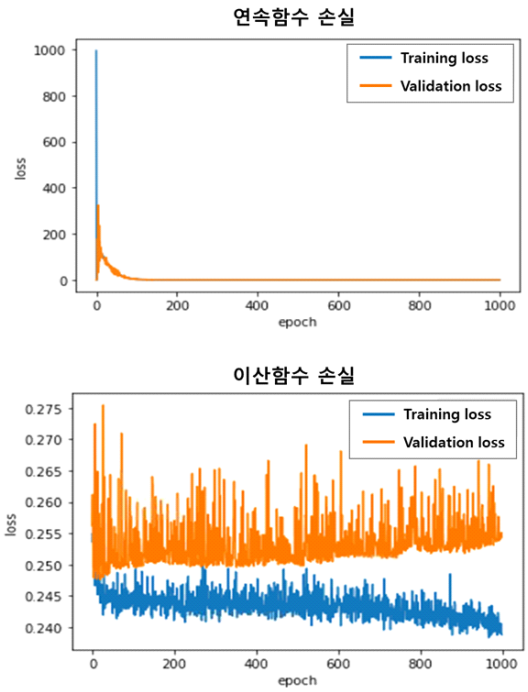

그림 3. 훈련에 따른 훈련 손실(orange line) 및 검증 손실(blue line)
Fig. 3. Training loss(orange line) and validation loss(blue line) during training process

Ⅳ. DES 평문 복구 공격

2012년에 M. M. Alani는 신경망을 이용하여 암호키를 모르는 상태에서 블록암호 DES 암호문의 평문을 복호화하는 연구 결과를 제시하였다[3]. 이 연구에서는 DES의 복잡도에 비해 단순한 신경망을 이용하여 DES 암호문에 대한 평문 복구에 성공하였다고 주장한다. 또한, 신경망의 근사 가능성으로 UAT[7]을 언급하였다.

신경망은 이산함수를 근사하는 데 어려움이 있음을 3장에서 확인하였으므로, 본 장에서는 M. M. Alani가 제안한 신경망을 이용한 DES의 평문 복구 공격에 대한 실험을 직접 재연하여 결과를 제시하고 이를 분석한다.

4.1 선행연구

제안된 신경망 기반 DES 평문 복구의 훈련과정(training process)은 [그림 4]와 같이 진행된다. 신경망의 입·출력을 두 블록 단위로 하여 DES-ECB로 암호화된 암호문에 대한 평문을 찾는 것을 목표로 한다.

신경망의 입력은 두 블록 암호문 128-비트이며, 암

호문에 대응하는 128-비트의 평문은 신경망의 입력에 대한 레이블이다. 오차함수(error function)를 통해 손실이 최소화되도록 신경망의 가중치와 편향을 조절한다. 신경망 훈련에서 사용되는 오차함수는 평균제곱오차이며, 식은 다음과 같다.

$$\text{평균제곱오차} : \frac{1}{n \cdot m}\sum_{j=1}^{n}\sum_{i=0}^{m-1}\left|p_i^{(j)} - \hat{p_i}^{(j)}\right|^2. \quad (4)$$

이때, n은 훈련 데이터의 개수, $p_i^{(j)}$는 j번째 훈련 데이터의 레이블의 i번째 비트, $\hat{p_i}^{(j)}$는 j번째 훈련 데이터에 대응하는 신경망의 출력의 i번째 비트이다.

훈련과정이 종료된 후, 훈련된 신경망을 평가하는 테스트과정(test process)을 진행한다. 이 과정에서 학습에 사용한 데이터로 신경망을 평가하는 내부오차(inside error)와 학습에 사용하지 않은 새로운 데이터로 신경망을 평가하는 외부오차(outside error)를 계산한다. 내부오차와 외부오차의 계산법은 다음과 같다.

$$\text{내부오차, 외부오차} : \frac{\sum_{i=1}^{s}\sum_{j=1}^{t} p'(i,j) \oplus p(i,j)}{s \times t}. \quad (5)$$

이때, s는 사용된 블록의 개수, t는 블록의 길이, $p'(i,j)$는 예측된 평문 i번째 블록의 j번째 비트, $p(i,j)$는 실제 평문 i번째 블록의 j번째 비트이다.

선행연구의 저자가 제시한 평균제곱오차, 내부오차, 외부오차의 결과는 [표 4]와 같다. 평균 2^{11}개 미만의

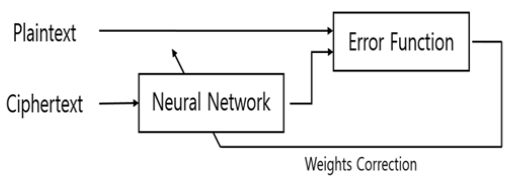

그림 4. 신경망 기반 평문 복구 공격[3]
Fig. 4. Neural network based plaintext recovery attack[3]

표 4. M. M. Alan가 제시한 신경망 기반 평문 복구 공격 결과[3]
Table 4. Results of neural network based plaintext recover attack presented by the author[3]

평균제곱오차	내부오차	외부오차
0.013317	0.027973	0.085986

(평문, 암호문) 데이터 쌍을 이용해 0에 가까운 오차값을 제시함으로써, DES의 암호문을 성공적으로 복호화하였다고 주장한다.

4.2 선행연구 재연 실험

본 절에서는 4.1절에 기재한 선행연구를 재연하여 결과를 제시하고, 이를 분석한다[12].

4.2.1 실험 설계

본 실험에서 사용한 파라미터는 선행연구[3]와 동일하게 설정하였다. 완전 연결 신경망을 사용하며, 활성화함수는 시그모이드를 사용한다. 테스트 데이터의 개수는 선행연구에 제시되어 있지 않았으므로 훈련데이터의 20%인 512개로 설정하였다. [표 5]는 실험에 사용한 데이터와 하이퍼파라미터이다.

본 논문에서 재연한 실험은 다음과 같이 진행된다.

① 데이터 생성

평문 두 블록을 랜덤하게 2,560개를 수집한 뒤, 고정된 키를 사용해 DES-ECB로 암호화를 진행하여 두 블록의 암호문을 생성한다. 수집한 (평문, 암호문) 쌍을 훈련 데이터, 테스트 데이터로 분류한다. 이때, 평문과 암호문은 각각 128-비트이다.

② 신경망 훈련과정

훈련 데이터를 이용해 신경망을 훈련한다. 손실함수는 평균제곱오차를 사용하며, 이는 4.1절의 오차함수와 동일한 의미를 지닌다. 손실함수를 통해 손실이 최소화되도록 가중치와 편향 값을 조절한다.

③ 신경망 테스트과정

훈련과정이 끝나면, 훈련된 신경망에 대해 훈련 데이터와 테스트 데이터로 각각 내부오차와 외부오차를 계산한다.

표 5. 재연한 실험의 하이퍼파라미터
Table 5. Hyperparameter of experiment

훈련 데이터	2,048개의 (암호문, 평문)
테스트 데이터	512개의 (암호문, 평문)
에폭	10,000
신경망 구조	128-128-256-256-128

4.2.2 실험 결과

[표 6]은 선행연구[3]를 재연한 실험 결과이다.

표 6. 재연한 실험 결과
Table 6. Results of experiment

평균제곱오차	내부오차	외부오차
0.012	0.041	0.502

평균제곱오차와 내부오차는 선행연구의 실험 결과와 비슷하였지만, 외부오차는 선행연구의 결과와 달리 큰 오차값을 가진다. 신경망의 훈련은 오차가 감소하는 방향으로 진행되므로 평균제곱오차는 훈련이 거듭될수록 0에 가까워지는 모습을 볼 수 있다.

내부오차는 훈련에 사용한 데이터에 대한 신경망의 예측값이므로 반복된 학습을 통한 암기 현상으로 인한 결과라고 분석한다.

외부오차 값은 0.502로 이는 평문 128-비트 중 평균 64-비트가 잘못 예측되었음을 의미한다. 하지만, 신경망의 출력은 잘못 예측된 비트의 위치에 대한 정보를 주지 않으므로 평문의 절반이 올바르게 복호화되었다고 판단할 수 없다. 그러므로 새로운 데이터에 대해서는 학습의 효과가 없다고 분석된다.

4.2.3 실험 결과 분석

훈련된 신경망이 잘못 예측하는 비트의 위치는 유동적이다. 즉, 신경망의 출력 중 어느 비트가 잘못 예측되었는지 알 수 없다. 만약, 잘못 예측된 비트의 위치를 알면 해당 비트를 반전시켜 올바른 평문을 복구할 수 있다. 하지만 입력에 따라 잘못 예측된 비트의 위치가 달라지면 [그림 5]와 같이 위치에 관한 전수조사가 필요하다.

훈련된 신경망의 출력값이 k-비트 잘못 예측하고, 이 k-비트의 위치가 데이터마다 다르다고 가정하자. 이 경우, k-비트의 위치에 대한 전수조사가 필요하다. [표 6]의 외부오차의 결과를 위의 방식으로 분석하면, $\binom{128}{64}$ 만큼의 잘못 예측한 비트의 위치에 대한 전수조사가 필요하고, 이는 약 2^{124}의 계산량을 가진다. k가 8 이하일 때 2^{40} 이하의 계산량으로 계산할 수 있어 유의미한 결과를 줄 수 있다고 할 수 있다. 하지만, 선행연구의 저자가 제시한 실험이 성공적으로 완료된다고 가정해도 외부오차 결과로부터 $k=11$ 정도가 되어 전체 평문을 올바르게 복구하기 위해선 2^{51} 정도의 매우 큰 계산량을 갖는다.

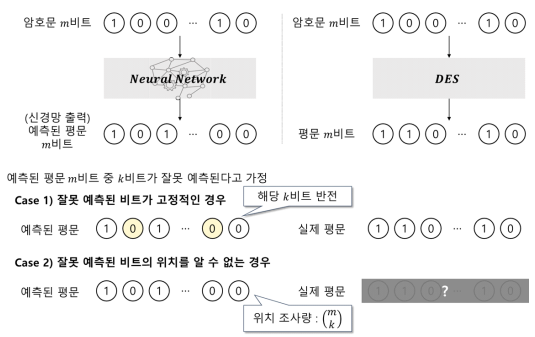

그림 5. 훈련된 신경망의 전체 평문 복구 방법
Fig. 5. A method of recovering the entire plaintext with trained neural network

선행연구가 제시한 평문 복구 공격법은 전체 평문을 복구할 수 없으며, 고정된 위치의 비트가 잘못 예측되는 것이 아니므로 평문을 복구하기 위해선 이 위치에 대한 큰 계산량의 전수조사가 필요하다. 그러므로 선행연구가 사용한 신경망으로 평문 복구가 불가능하며, 제시된 평문 복구 공격법은 유의미한 결과를 주지 않는다.

Ⅴ. 1-라운드 및 2-라운드 DES 평문 복구 공격

본 장에서는 신경망을 이용한 라운드별 DES의 평문 복구 가능성을 알아보고자 1-라운드 암호문과 2-라운드 암호문에 대하여 평문을 복구하는 공격을 시도한다.

5.1 실험 설계 및 실험 과정

[그림 6]은 사용한 신경망의 구조이다. 완전 연결 신경망을 사용하며, 신경망의 구조는 5개의 은닉층으로 구성되고, 각 층의 노드는 512개를 사용한다. 에폭에 대한 실험을 통해 에폭이 5,000일 때 오차가 충분히 최소화됨을 확인하였으므로, 본 실험에서 에폭은 5,000으로 설정한다.

실험은 다음과 같이 데이터 생성, 신경망 훈련과정, 신경망 테스트과정 순으로 진행된다.

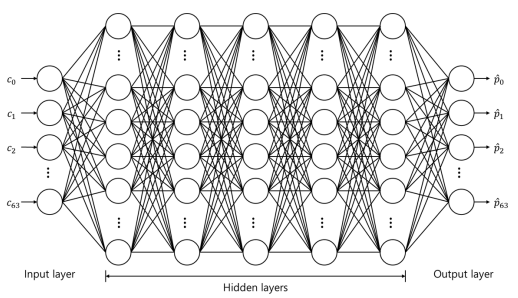

그림 6. DNN 모델
Fig. 6. DNN model

① 데이터 생성

사용할 암호 키를 랜덤하게 1개 생성하고, 64-비트의 평문 100,000개를 랜덤하게 수집하여 [그림 7]과 같이 DES로 1-라운드 암호화와 2-라운드 암호화를 하여 1-라운드 암호문, 2-라운드 암호문을 생성한다. 이때, 암호문은 신경망의 입력이며, 평문은 대응하는 레이블이다. 수집한 (평문, 암호문) 쌍을 70,000개의 훈련 데이터 15,000개의 테스트 데이터로 분류한다.

② 신경망 훈련과정

평균제곱오차를 손실함수로 사용하여 손실이 최소화되도록 신경망을 학습시킨다.

③ 신경망 테스트과정

신경망 훈련이 끝난 후, 테스트 데이터를 이용해 훈련된 신경망을 평가한다. 이때, 신경망의 출력은 실수이므로 다음과 같이 0 또는 1로 변경한 뒤 신경망의 출력값과 레이블의 오차를 계산한다.

$$\tilde{p}_i = \begin{cases} 0, & \text{if } \hat{p}_i < 0.5, \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

오차는 다음과 같이 틀리는 평균 비트의 수로 계산한다.

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{63}\tilde{p}_i^{(j)} \oplus p_i^{(j)}. \tag{7}$$

이때, n은 테스트 데이터의 수, $\tilde{p}_i^{(j)}$는 j번째 테스트 데이터에 대응하는 신경망의 출력값의 i번째 비트, $p_i^{(j)}$는 j번째 테스트 데이터의 레이블의 i번째 비트이다.

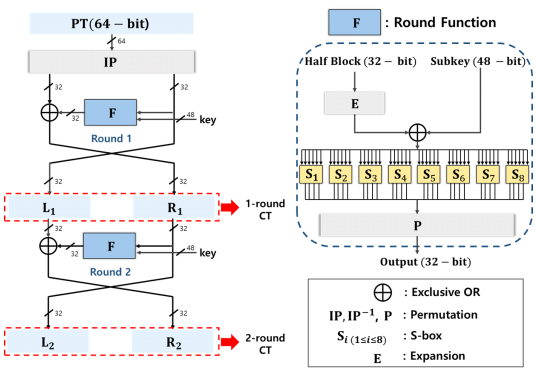

그림 7. 1-라운드 및 2-라운드 DES 암호화
Fig. 7. 1-round and 2-round DES encryption

5.2 실험 결과

[그림 8]은 1-라운드 및 2-라운드 DES 평문 복구 과복호화 과정을 나타낸다. DES의 구조 특성상 1-라운드의 경우, R_0는 L_1과 동일하며 R_1은 L_0에 S-box가 내포된 F함수의 출력값과 XOR(exclusive or) 연산된 값이다. 반면, 2-라운드는 L_2와 R_2 모두 F함수를 거치게 되어 암호문에는 평문의 특성이 직접 나타나지 않는다.

[그림 9]는 DES의 1-라운드 암호문과 2-라운드 암호문에 대한 평문 복구 공격에서, 테스트 데이터에 대하여 몇 비트의 평문이 잘못 예측되었는지 분포를 그린 것이다.

1-라운드 DES의 평문 복구 공격 시, 암호문에 대하여 32-비트의 정보를 갖고 있으므로 높은 예측률로 평문을 복구할 수 있다. 테스트 데이터로 훈련된 신경망을 평가한 결과 15,000개 중 약 80%인 11,928개가 올바른 평문으로 복구하였다. 오류가 생기는 경우에도 잘못 예측된 비트는 64-비트 중 4-비트 이하로 2^{19}이하의 공격량으로 완전한 평문을 확정할 수 있다.

한편, 2-라운드의 결과로 평균이 32인 정규분포의 형태를 띠는 것을 볼 수 있으며, 이는 1-라운드와 달리 테스트 데이터 중 올바르게 예측된 평문이 없음을 의미한다.

[그림 10]은 테스트 데이터 15,000개에 대하여 신

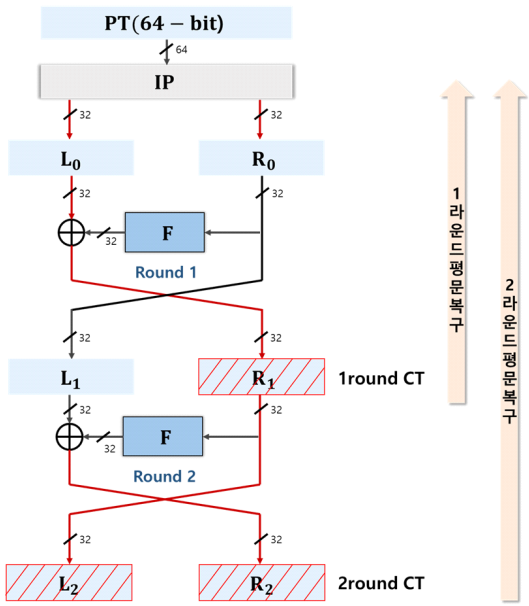

그림 8. 1-라운드 및 2-라운드 DES 복호화 과정
Fig. 8. The process of decryption 1-round and 2-round DES

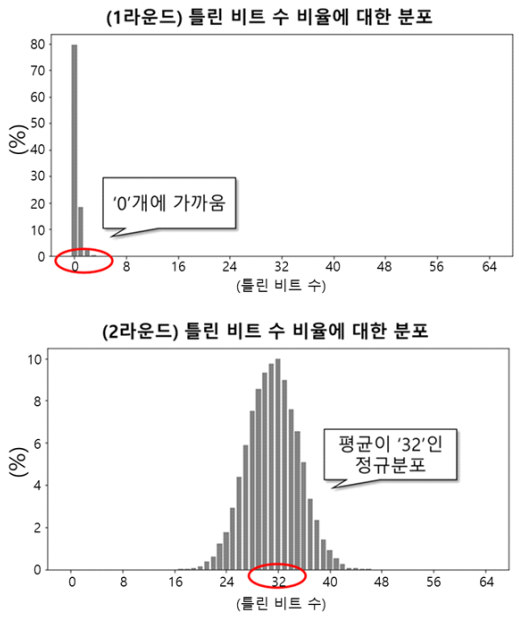

그림 9. 틀린 비트 수 비율에 대한 분포 (좌) 1-라운드, (우) 2-라운드
Fig. 9. Distribution for wrong bit count ratio (left) 1-round, (right) 2-round

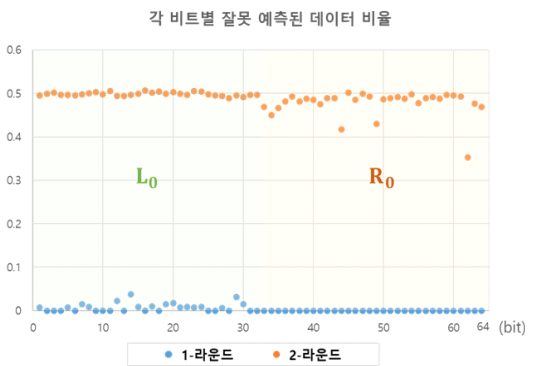

그림 10. 각 비트별 잘못 예측된 데이터 비율
Fig. 10. The ratio of incorrectly predicted data by each bit

경망이 예측한 64-비트의 평문에서, 각 비트별로 잘못 예측된 데이터의 비율을 의미한다. 1-라운드 실험의 경우 R_0에 속하는 비트는 모두 예측하였고, L_0는 0에 가까운 비율로 잘못 예측하여 전반적으로 평문 복구에 성공하였다. 2-라운드의 경우 각 비트가 0.5의 비율로 맞히는 양상을 띠어, 평문 64-비트 중 평균 31-비트를 잘못 예측하였다.

5.3 실험 분석

신경망을 이용하여 DES를 근사하기 위해서는, DES 알고리즘의 복잡도를 신경망으로 표현할 수 있어야 한다.

DES 1-라운드의 계산 복잡도는 S-box에 의존한다. 6-비트 입력에 4-비트 출력인 S-box를 병렬로 8개 사용하므로, 샤논의 정리에 의하면 최소 $\Omega(2^6/6)\times 4\times 8 \approx 2^9$ 이상의 논리 회로가 있는 경우 S-box를 계산할 수 있다. 실험에 사용된 신경망은 2^{20} 정도의 가중치 복잡도를 갖는 신경망이다. 이는 1-라운드 DES의 계산 복잡도를 충족시키며 암호문에 평문의 특성이 남아있으므로 평문 복구 가능성을 보였다.

반면, 2-라운드의 평문 복구 공격을 고려할 시, S-box는 1-라운드와 2-라운드에 각각 8개씩 있다. 따라서 신경망을 이용해 이를 계산하는 것은 각 라운드의 모든 S-box들이 혼합되어 64-비트 입출력의 이산함수로 간주된다. 즉, 2-라운드 DES를 입출력만으로 근사하기 위해서는 $\{\Omega(2^{64}/64)RIGHT$ 이상의 논리 회로가 필요할 것으로 예상된다.

따라서 본 실험 결과를 통해, 1-라운드의 경우 L_1와 R_1 정보를 이용하여 평문을 복구할 가능성을 보였지만, 2-라운드의 경우 L_2와 R_2 정보만으로는 평문을 복구할 수 없었다. 2-라운드 이상의 DES는 블록암호의 비선형 요소로 인하여 복잡도가 매우 커지게 되어, 신경망을 이용하여 이를 근사하는 데에 어려움이 있기 때문이라고 분석한다.

Ⅵ. 결 론

본 논문은 신경망을 이용하여 암호 키 없이 암호문만으로 블록암호 DES의 평문 복구를 하는 것은 어려움이 있음을 보였다. 먼저, 신경망은 연속함수를 근사하지만, 이산함수를 근사하기 어려움을 확인하였다. 다음으로, 블록암호 DES의 평문을 복구하는 것은 사용한 신경망의 복잡도로는 근사할 수 없음을 보였다. 또한, 라운드별 평문 복구 공격 가능성을 분석하여 1-라운드 DES의 암호문의 경우 64-비트 중 32-비트의 정보를 통해 평문을 복구할 수 있음을 확인하였다. 반면, 2-라운드 DES의 암호문은 전혀 복호화하지 않음을 확인하였고, 2-라운드 이상 암호문에 대해서는 신경망을 이용한 평문 복구에 어려움이 있음을 파악하였다.

References

[1] S. Baek and K. Kim, "Recent advances of neural attacks against block ciphers," in *Symp. SCIS*, Jan. 2020.

[2] A. Gohr, "Improving attacks on round-reduced speck32/64 using deep learning," in *Annu. Int. Cryptology Conf.*, pp. 150-179, Aug. 2019.

[3] M. M. Alani, "Neuro-cryptanalysis of DES," *IEEE WorldCIS-2012*, pp. 23-27, 2012.

[4] M. M. Alani, "Neuro-cryptanalysis of DES and triple-DES," *Int. Conf. Neural Info. Process.*, Springer, pp. 637-646, Berlin, 2012.

[5] X. Hu and Y. Zhao, "Research on plaintext restoration of AES based on neural network," *Secur. and Commun. Netw.*, Nov. 2018.

[6] S. Fan and Y. Zhao, "Analysis of DES plaintext recovery based on BP neural network," *Secur. and Commun. Netw.*, Nov. 2019.

[7] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, 1989.

[8] M. H. Stone, "The generalized Weierstrass approximation theorem," *Math. Mag.*, vol. 21, no. 5, pp. 237-254, 1948.

[9] J. So, "Deep learning-based cryptanalysis of lightweight block ciphers," *Secur. and Commun. Netw.*, Jul. 2020.

[10] H. Yim, S. Kwon, J. S. Kang, and Y. Yeom, "An analysis of possibility of approximations for continuous and discrete functions using machine learning," in *Proc. KICS Summer Conf.*, pp. 577-578, Korea, Aug. 2020.

[11] J. F. Nash and M. T. Rassias, "*Open problems in mathematics*," Springer, 2016.

[12] S. Kwon, H. Yim, J. S. Kang, and Y. Yeom, "A study on the cryptanalysis of DES using neural network," in *Proc. KICS Summer Conf.*, pp. 577-578, Korea, Aug. 2020.

권 수 진 (Sujin Kwon)

2020년 2월 : 국민대학교 정보보안암호수학과 졸업

2020년 3월~현재 : 국민대학교 금융정보보안학과 석사과정

<관심분야> 암호구현, 난수성 분석 및 평가, 병렬 프로그래밍

[ORCID:0000-0003-1062-6042]

임 형 신 (Hyoungshin Yim)

2020년 2월 : 국민대학교 정보보안암호수학과 졸업

2020년 3월~현재 : 국민대학교 금융정보보안학과 석사과정

<관심분야> 암호구현, 난수성분석 및 평가, 화이트박스암호

[ORCID:0000-0002-7826-7536]

강 주 성 (Ju-Sung Kang)

1989년 2월 : 고려대학교 수학과 졸업

1991년 2월 : 고려대학교 일반대학원 수학과 석사

1996년 2월 : 고려대학교 일반대학원 수학과 박사

1997년~2004년 : 한국전자통신연구원 선임연구원/팀장

2004년 3월~현재 : 국민대학교 과학기술대학 정보보안암호수학과 정교수

2013년~현재 : 국민대학교 BK21+ 미래 금융정보보안 인력양성사업단 교수

<관심분야> 암호이론, 정보보안 프로토콜, 안전성 분석 및 평가

[ORCID:0000-0002-0846-389X]

염 용 진 (Yongjin Yeom)

1991년 2월 : 서울대학교 수학과 졸업
1994년 2월 : 서울대학교 수학과 석사
1999년 2월 : 서울대학교 수학과 박사

2000년 4월~2012년 2월 : ETRI 부설연구소 책임연구원/팀장
2012년 3월~현재 : 국민대학교 과학기술대학 정보보안암호수학과 정교수
2013년~현재 : 국민대학교 BK21+ 미래 금융정보보안 인력양성사업단 교수
<관심분야> 암호구현 및 분석, 보안시스템 평가
[ORCID:0000-0002-8240-8661]