정 보 보 호 학 회 지
제 31 권 제6호, 2021. 12

블록암호에 대한 신경망 암호해독 연구 동향 분석

석 병 진*, 이 창 훈**

요 약

신경망 암호해독은 딥러닝 기술을 기반으로 암호해독을 수행하는 기술을 말한다. 2000년대부터 신경망 암호분석 연구가 일부 수행된 바 있지만 대부분 가능성을 제시하기 위하여 실제로 사용되지 않는 암호를 대상으로 수행되어 활용에 어려움이 있었다. 그러나 최근 암호학계 저명 컨퍼런스인 CRYPTO, EUROCRYPT에서 전통적인 암호해독 기법보다 우수한 성능을 보이는 신경망 암호분석 연구 결과가 발표된 바 있다. 하지만, 신경망 암호해독 기술은 딥러닝 기술이 활용되고 있는 타 분야 대비 상대적으로 연구 초기 단계에 있어 다양한 한계점이 존재한다. 이에 본 논문에서는 현재까지 수행된 신경망 암호분석 연구에 대한 동향을 분석하고 한계점 및 향후 연구 방향을 제시한다.

Ⅰ. 서 론

최근 컴퓨터 성능이 급격히 발달함에 따라 딥러닝 기술은 급격한 성장을 이루면서 이미지 인식, 주식/금융, 의료, 사이버 보안 등 다양한 분야에 적용되어 인간의 오차 범위를 뛰어넘는 성능을 보이면서 성공적으로 활용되고 있다. 한편, 암호 분야에서도 딥러닝을 적용하고자 하는 다양한 연구가 수행되었으며, 그중에서도 부채널 분석(Side Channel Attack) 분야에서는 부채널 정보를 학습하여 취약성을 탐색하는 다양한 모델들이 개발된 바 있다. 하지만, 암호 알고리즘의 암호학적 취약성을 분석하는 암호해독(Cryptanalysis) 분야에서는 딥러닝 기술을 적용하는 연구가 상대적으로 미비하게 수행되었다. 이는 데이터셋의 특징을 탐색하고 이를 학습하는 방식의 딥러닝 기술이 난수성(Randomness)을 목적으로 하는 블록암호 알고리즘의 특성상 암호문을 잘 학습하지 못하는 특성에 기인한다. 하지만, 최근 딥러닝 기술을 기반으로 암호해독을 수행한 신경망 암호해독(Neural Cryptanalysis) 연구 결과들이 발표되고 있으며 대표적 암호해독 기법인 차분분석(Differential Cryptanlysis)보다 우수한 결과들이 발표되고 있다.

신경망 암호해독 연구는 2000년대 중반부터 수행되었지만, 당시에 수행된 연구들은 간단한 구조의 인공신경망(Artificial Neural Network) 모델에 암호문 또는 평문을 학습하여 대응하는 키를 찾거나 암호 알고리즘의 동작을 모방하도록 하는 연구가 수행되었다. 또한, 분석 대상 암호 알고리즘은 교육을 목적으로 설계된 Toy Cipher나 1-, 2-라운드로 라운드 수가 매우 작게 축소된 암호들을 대상으로 수행되었다. 이로 인해 기존 연구 결과들은 무작위로 추측하는 확률과 동일한 결과를 보이거나 라운드 수가 매우 작아 암호학적으로 유의미하다고 판단하기 어려웠다. 하지만, 최근 암호학계 저명 컨퍼런스인 CRYPTO, EUROCRYPT에서 딥러닝을 기반으로 경량블록암호 SPECK-32/64에 대한 신경망 암호해독 연구 결과가 발표되었으며 이는 전통적 암호해독 기법인 차분분석 대비 우수한 성능을 보였다.

상기 CRYPTO, EUROCRYPT에서 발표된 연구 결과는 딥러닝 기술이 암호해독 분야에 유의미하게 적용될 수 있음을 보였지만 현재까지 딥러닝 기술을 암호해독에 활용하기 위한 명확한 방법론이 정립되어 있지 않은 상황이다. 이에 본 논문에서는 현재까지 수행된 신경망 암호해독 연구 동향을 분석하며 이를 통해 기존 연구 한계점 및 향후 연구 방향을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 신경망 암호해독 개요를 다루며 3장에서는 현재까지 수행된 기존 연구 동향 분석을 수행하며 기존 연구의 한계점

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1F1A1076468)
* 서울과학기술대학교 컴퓨터공학과 (대학원생, sbj7534@seoultech.ac.kr)
** 서울과학기술대학교 컴퓨터공학과 (교수, chlee@seoultech.ac.kr)

을 제시한다. 또한, 4장에서 기존 연구에 대한 고찰을 수행하며 5장에서 결론을 통해 본 논문을 마친다.

Ⅱ. 신경망 암호해독 개요

신경망 암호해독은 딥러닝 모델을 기반으로 암호해독을 수행하기 위한 모든 행위를 말하며 일반적으로 인공신경망 기반의 학습 모델을 구성하고 평문 또는 암호문으로 구성된 데이터셋(이하 암호 데이터셋)을 학습하여 암호 취약성을 탐색하는 기술을 말한다. 신경망 암호해독 기술은 딥러닝 기술과 암호해독 기술이 융합된 새로운 응용 기술로 딥러닝 모델을 학습시켜 달성하고자 하는 암호해독 수행 목적에 따라 다양한 공격 유형이 존재하며 딥러닝을 수행하는 절차인 딥러닝 파이프라인 관점에서 다양한 방법론이 존재한다. 현재 신경망 암호해독 연구는 암호 알고리즘 중에서도 블록암호를 위주로 수행되고 있다. 블록암호에 대한 신경망 암호해독의 공격 유형과 딥러닝 파이프라인 설명은 다음과 같다.

블록암호에 대한 신경망 암호해독의 공격 유형은 암호 모방 공격(Cipher Emulation Attack), 식별 공격(Identification Attack), 키 복구 공격(Key Recovery Attack)으로 구분되며 각 공격에 대한 설명은 다음과 같다[1].

• 암호 모방 공격

암호 모방 공격은 인공신경망 기술을 기반으로 분석대상 암호 알고리즘의 암호화 또는 복호화 동작과정을 모방하도록 딥러닝 모델을 학습시켜 입력된 평문(또는 암호문)에 대응하는 암호문(또는 평문)을 복구하는 공격이다. 이 때, 평문을 입력으로 하여 암호문을 구하는 공격의 경우에는 Encryption Emulation Attack(EEA)이라고 하며, 암호문을 입력으로 하여 평문을 복구하는 공격의 경우에는 Plaintext Restoration Attack(PRA)이라고 한다.

• 식별 공격

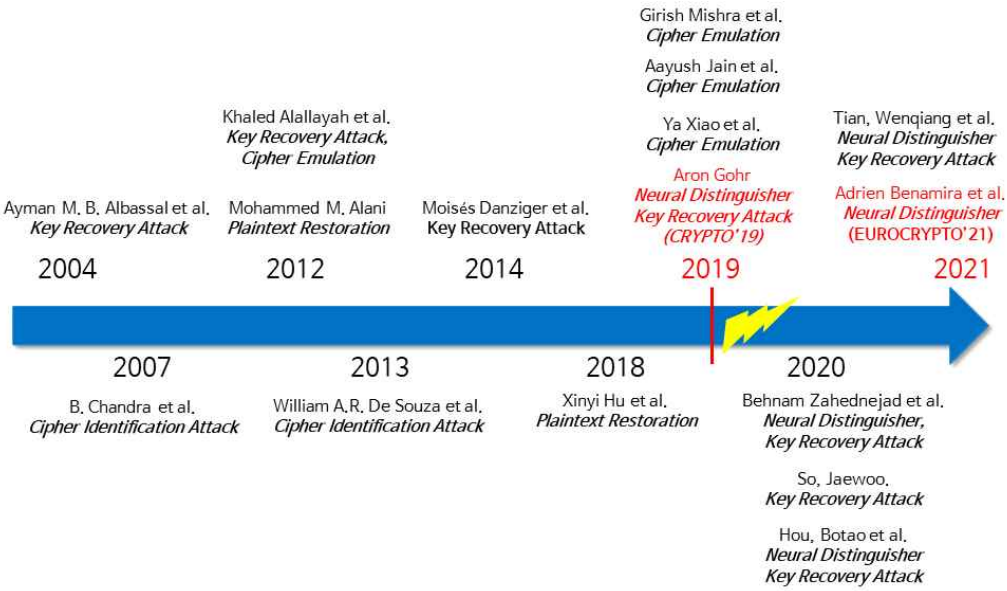
식별 공격은 평문과 암호문이 주어졌을 때 주어진 데이터가 생성된 환경을 식별하는 공격을 말한다. 세부적으로는 평문과 암호문 쌍을 학습하여 모델에 입력된 암호문이 어떤 평문으로부터 암호화 되었는지를 추측하는 공격인 Plaintext Type Identification Attack(PTI)라 하며, 입력된 암호문이 어떤 암호 알고리즘을 통해 암호화 되었는지를 추측하는 공격인 Cipher System Identification Attack(CSI)이 존재한다.

• 키 복구 공격

키 복구 공격은 평문과 암호문을 딥러닝 모델에 학습시켜 얻은 유의미한 정보를 통해 암호키를 추측하는 모든 행위를 말한다. 인공신경망 기술을 활용한 키 복구 공격은 평문-암호문 쌍을 학습한 딥러닝 모델에 암호키를 질의하는 방식으로 공격이 수행될 수 있으며 학습된 모델을 기반으로 유의미한 정보를 얻고 이를 기반으로 구성한 키 복구 공격 모델에 키를 질의하는 방식으로 공격을 수행할 수 있다.

Ⅲ. 신경망 암호해독 연구 동향

2000년대부터 현재까지 수행된 신경망 암호해독의 연구 동향은 3가지 양상을 보인다. 첫 번째로 신경망 구별자가 제안되기 이전 시점(2000년대~2019년)까지는 딥러닝 모델을 암호해독 분야에 적용하기 위한 연구로 학습된 모델이 암호문, 평문 또는 키와 관련된 정보를 추측하도록 학습을 시켜 공격을 수행하는 방식으로 연구가 수행되었다. 두 번째는 딥러닝 모델과 기존 암호해독 기술을 융합하고자 하는 연구로 신경망 모델에 단순히 암호문, 평문을 학습시키는 것이 아닌 암호학적으로 유의미한 특성을 활용하여 데이터셋을 생성하고 이를 기반으로 구별자를 만들어 키 복구 공격에 응용한다. 이는 CRYPTO'19에서 A. Gohr에 의해 수행된 연구로 차분 특성을 기반으로 생성된 암호문 쌍 데이터셋을 학습한 신경망 구별자(Neural Distinguisher)라는 개념이 제시되었으며, A. Gohr는 신경망 구별자를 기반으로 키 복구 공격을 수행했다. 세 번째는 신경망 구별자를 해석하고 개선하기 위한 연구로 인공신경망 모델이 블랙박스 방식으로 동작하여 결과에 대한 해석을 수행하기 어렵다는 것을 개선하기 위하여 신경망 구별자의 결과를 해석하기 위한 연구이다. 신경망 구별자에 대한 연구는 EUROCRYPT'21에서 A. Benamira 외 3인에 의해 수행되었으며 A. Gohr의 신경망 구별자의 학습 결과를 암호해독/기계학습 관점에서 해석을 수행했다. 위에서 설명한 신경망 암호해독 연구들에 대한 동향 설명은 다음과 같다. 신경망 암호해독 연구 동향은 [그림 1]과



〔그림 1〕 신경망 암호해독 연구 동향 개요

같이 나타낼 수 있다.

3.1. 딥러닝 기술 적용 연구

딥러닝 기술 적용 연구는 대부분 A. Gohr의 연구 이전에 수행된 연구들로 딥러닝 모델이 암호와 관련된 정보를 추측하도록 학습시키는 것을 목적으로 한다. [1]에서 분류를 수행한 연구 결과들 중 A. Gohr의 연구결과를 제외한 암호 모방 공격, 식별 공격, 키 복구 공격 연구 결과들은 모두 딥러닝 모델을 통한 추측 연구에 해당한다.

먼저 암호 모방 공격을 살펴보면, G. Mishra 외 2인[2]은 31-라운드 PRESENT 암호 알고리즘에 대해서 PRA를 수행했다. 대상 암호인 PRESENT는 Substitution-Permutation Network(SPN) 구조 암호로 블록 단위가 64-bit인 경량블록암호이다. 학습하기 위한 데이터셋은 무작위로 생성한 10,000개의 평문을 암호화하여 평문에 대응하는 암호문 10,000개를 생성하였으며 (pt, ct)를 암호 데이터셋으로 활용했다. 생성된 데이터셋을 학습하는 딥러닝 모델은 암호문을 입력하면 평문를 출력하도록 Input layer와 Output layer의 노드를 64개로 구성된 Fully-Connected Neural Network를 구성했다. 학습 수행 결과, 테스트 데이터셋에 대하여 각 비트에 대한 정확도는 최소 46%, 최대 55.1%의 정확도를 보였다. 이는 0과 1 중 임의로 하나를 선택할 확률인 50%에 가까운 값으로, 무작위로 평문의 각 비트를 추측하는 것과 유사함을 의미한다. 또한, 64-bit 블록암호인 FeW에 대해서도 동일한 연구를 수행하였으며 위의 연구 결과와 유사하게 각 비트에 대한 정확도는 최소 46%, 최대 55.1%를 보였다[3].

Y. Xiao 외 2인[4]은 블록암호 DES에 대하여 EEA를 수행했다. 이 연구에서는 내부 구조가 서로 다른 3가지 형태의 딥러닝 모델(Deep and Thin, Fat and Shallow, Cascade)을 구성하여 평문을 입력하여 암호문을 추측하도록 학습을 수행했다. 학습의 정확도는 평문에 대응하는 암호문을 올바르게 추측한 비트 수를 합산하여 계산하였다. 학습 수행 결과는 3-라운드 이상에서는 올바르게 추측한 비트가 약 50% 정도로 무작위로 암호문을 추측하는 것과 다르지 않은 결과를 보였다.

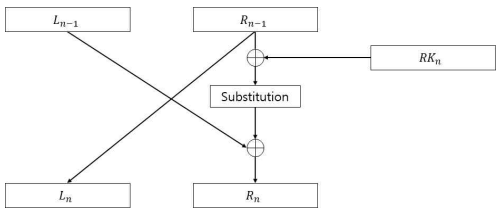
다음으로 식별 공격 연구 결과를 살펴보면, B. Chandra와 P. Paul Varghese[5]은 신경망을 기반으로 블록암호 Enhanced RC6와 스트림암호 SEAL 2개의 암호 알고리즘을 구별하는 연구를 수행했다. 학습에 사용된 데이터셋은 ASCII 값으로 변환된 400개의 평문 및 암호문 쌍을 사용했으며 3가지 데이터셋 종류(Single Key-Different Messages, Different

〔표 1〕 B. Chandra와 P. Paul Varghese 의 신경망 모델과 데이터셋 별 정확도

신경망 모델	데이터셋	정확도
Cascade Correlation	Single Key, Different Messages	90.9%
	Different Keys, Same Messages	93%
	Different Keys, Different Messages	92%
Gradient descent back propagation	Single Key, Different Messages	73%
	Different Keys, Same Messages	83%
	Different Keys, Different Messages	85%

Keys-Same Messages, Different Keys, Different Messages)를 구성했다. 모델의 구조는 두가지의 신경망 구조(cascade correlation ,back propagation)를 사용했다. 두 알고리즘을 구별하도록 학습을 수행한 결과, 70% 이상의 정확도를 보였으며 cascade correlation 신경망의 경우 3가지 데이터셋에 대해 90% 이상의 정확도를 보였다. 데이터셋과 모델에 따른 세부적인 정확도는 [표 1]과 같다.

마지막으로 키 복구 공격 결과를 살펴보면, A. M. Albassal과 A. M. Wahdan[6]는 HypCipher라는 Toy Cipher를 구성하고 이에 대한 Multi-Layer Perceptron에 학습시켜 키 복구 공격을 수행했다. [그림 2]는 HypCipher의 한 라운드를 나타내며 substitution 과정은 블록암호 AES의 SubBytes를 사용한다. HypCipher는 전체 라운드 수는 4이며 블록 크기가 16-bit, 라운드 키 크기는 8-bit이다.



〔그림 2〕 HypCipher의 라운드 함수

학습에 사용된 MLP 모델의 구조는 각 계층의 노드 수를 HypCipher의 블록 크기와 동일하게 16개로 구성하였으며 두 개의 Hidden Layer를 사용했다. 모델을 학습시키기 위한 데이터셋은 키를 복구하고자 하는 대상 라운드 수를 n이라 할 때, 라운드키 RK_n을 통해 부분 복호화하여 $(n-1)$-라운드 암호문를 계산한 뒤 평문과 부분 복호화된 $(n-1)$-라운드 암호문을 연접하여 구성했다. 모델의 학습은 평문과 부분 복호화된 암호문 사이의 상관관계를 학습하도록 수행되었다.

A. M. Albassal과 A. M. Wahdan의 키 복구 공격 아이디어는 다음과 같다. 데이터셋을 생성할 당시에 추측한 마지막 라운드키 RK_n이 틀린 키였다면 부분 복호화된 암호문이 uniform distribution에 가까울 것이므로 학습 에러율이 높은 에러율을 보일 것이고 추측한 RK_n이 옳은 키이면 학습 모델의 모델이 훨씬 작을 가능성이 크므로 에러율이 가장 작은 키 값을 옳은 키로 추측하는 것이다. 이와 같은 아이디어로 2-, 3-, 4-라운드의 HypCipher 키 복구 수행 결과 가장 낮은 에러율을 보인 키가 실제 마지막 라운드키와 동일한 것을 보였다.

M. Danziger와 M. A. A. Henriques[7]은 블록 크기가 8-bit이고 10-bit 키를 사용하는 Toy Cipher인 S-DES에 대하여 키 복구 공격을 사용했다. 공격에 사용된 모델은 상기 설명한 Albassal et al.의 키 복구 공격과 유사하게 MLP 모델을 사용하였으며, Input layer의 크기는 평문-암호문 쌍의 크기인 16-bit, Output layer의 크기는 키 크기인 10-bit에 대응되도록 구성했다. 학습에 사용된 데이터셋은 무작위로 생성된 평문을 암호화하여 평문-암호문 쌍으로 구성했다. 모델의 학습은 평문-암호문 쌍을 입력으로 하여 암호키의 각 비트를 추측하도록 수행됐다.

키 복구 공격을 수행한 결과, 테스트 데이터셋에 대하여 0.5 이상의 상관관계를 갖는 키 비트가 3개(k_0, k_1, k_4)가 존재함을 보였다.

3.2. 신경망 구별자 기반 키 복구 공격 연구

최근 저명 암호학계인 CRYPTO'19에서 A. Gohr[8]에 의해 수행된 신경망 암호해독 연구에서 인공신경망을 기반으로 입력된 암호문이 난수 평문을 암호화한 것인지 차분 특성을 기반으로 암호화한 것인지를 구별하는 신경망 구별자(Neural distinguisher)가 제안되었고 이를 기반으로 경량블록암호 SPECK-32/64에 대한 11-라운드 키 복구 공격 결과가 발표됐다. A. Gohr의 연구 결과에서 신경망 구별자는 전통적 방식의 차분 분석 기법을 기반으로 구성된 차분 구별자(Differential

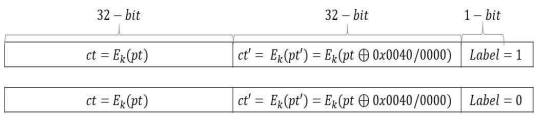
distinguisher)보다 우수한 성능을 보여 딥러닝 기술이 암호해독 분야에 유의미하게 활용될 수 있음을 보였다.

A. Gohr의 신경망 구별자는 전통적 암호해독 방식 기반의 차분 구별자 동작 과정을 모델링하여 인공신경망 기술을 기반으로 구성한 구별자로 차분 구별자와 동일하게 입력된 암호문 쌍이 차분특성 기반의 평문쌍을 암호화한 암호문 쌍(real case)인지 난수 평문쌍을 암호화한 암호문 쌍(random case)인지를 구별한다. 신경망 구별자의 입력은 암호문 쌍만을 입력으로 하도록 구성되어 있다. Aron Gohr의 신경망 구별자 구조는 input layer, hidden layer, output layer로 구성된다. 여기서 hidden layer에 해당하는 계층은 residual network라고 알려진 CNN(Convolutional Neural Network) 기반의 모델이 사용되었다. Residual network는 딥러닝 모델이 깊어질수록 나타내는 gradient vanishing 문제를 해결하기 위해 입력과 출력에 대한 잔차를 추가적으로 다음 계층에 전파하는 skip connection이 구성되어 있다.

A. Gohr의 신경망 구별자를 학습시키기 위한 데이터셋은 무작위로 생성한 평문 쌍과 키를 이용하여 암호화한 암호문 쌍으로 구성된다. A. Gohr가 사용한 데이터셋의 형태는 [그림 3]과 같다. 데이터셋은 총 10^7개를 생성하였으며 이중 10^6개는 검증 데이터셋으로 사용했다.

위에서 설명한 것과 같이 구성한 신경망 구별자(N_{nr})를 5-, 6-, 7-, 8-라운드 SPECK-32/64에 대하여 학습시킨 결과, 전통적인 방식의 차분 구별자(D_{nr})보다 성능이 좋은 것으로 나타났다. 두 구별자의 성능(정확도, TPR, TNR)은 [표 2]와 같다.

A. Gohr는 위에서 구성된 6-, 7-라운드 신경망 구별자를 기반으로 11-라운드 SPECK-32/64에 대해 부분키 복구 공격을 수행했다. 이 공격의 아이디어는 학습된 7-라운드 신경망 구별자에 2-라운드 차분 특성 $\delta \rightarrow$ (0x0040/0000)을 추가하여 9-라운드 구별자로 확장하고 추가적으로 입력 평문 쌍 (P_0, P_1)이 확장을 위해 사용한 차분 형태인 δ로 전파되도록 하는 structure를



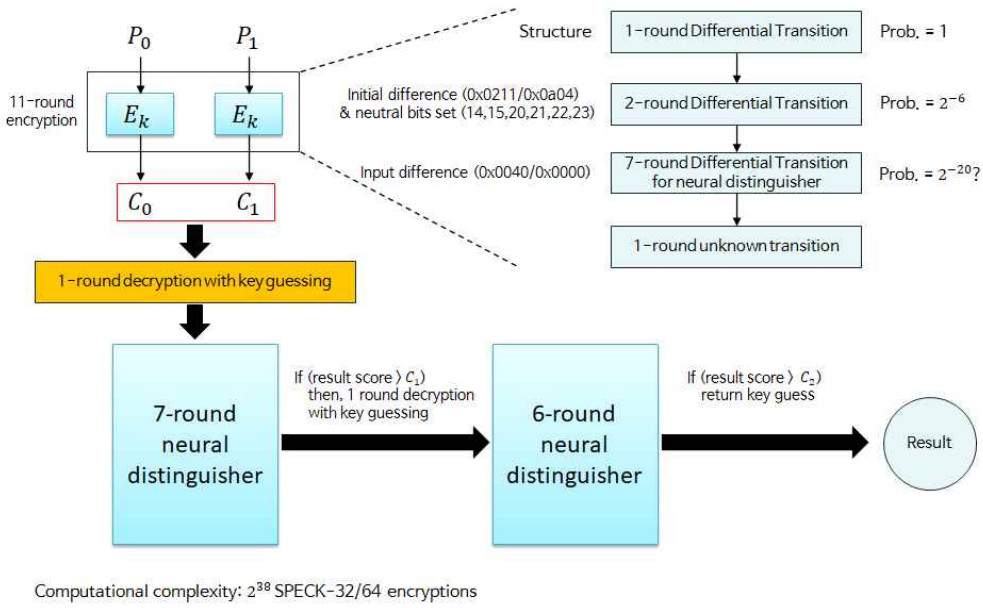
〔그림 3〕 A. Gohr의 암호문 쌍 데이터셋 형태

〔표 2〕 5-, 6-, 7-, 8-라운드 SPECK-32/64에 대한 차분 구별자와 신경망 구별자 정확도

라운드 수	구별자	정확도	TPR	TNR
5	D_5	0.911	0.877	0.947
5	N_5	$0.929 \pm 5.13 \times 10^{-4}$	$0.904 \pm 8.33 \times 10^{-4}$	$0.954 \pm 5.91 \times 10^{-4}$
6	D_6	0.758	0.680	0.837
6	N_6	$0.788 \pm 8.17 \times 10^{-4}$	$0.724 \pm 1.26 \times 10^{-3}$	$0.853 \pm 1.00 \times 10^{-3}$
7	D_7	0.591	0.543	0.640
7	N_7	$0.616 \pm 9.70 \times 10^{-4}$	$0.533 \pm 1.41 \times 10^{-3}$	$0.699 \pm 1.30 \times 10^{-3}$
8	D_8	0.512	0.496	0.527
8	N_8	$0.514 \pm 1.00 \times 10^{-3}$	$0.519 \pm 1.41 \times 10^{-3}$	$0.508 \pm 1.42 \times 10^{-3}$

구성하여 1-라운드를 더 확장시키는 것이다. 이때, structure를 효율적으로 구성하기 위해 확률적으로 차분 비트에 영향을 받지 않는 중립비트(Neutral bit)를 구성하여 더 많은 structure를 사용하는 효과를 가질 수 있도록 했다. 이와 같이 확장된 7-라운드 신경망 구별자는 앞에서 10-라운드 암호문에 대하여 real case인지 random case인지에 대한 판별을 수행하도록 동작한다. 따라서, 11-라운드 암호문 쌍을 추측 키 k를 사용해 1-라운드 복호화하여 확장된 7-라운드 신경망 구별자에 입력하고 출력된 $score$가 사전에 정의해놓은 임계값 c_1보다 크면 k를 옳은 키 후보로 저장한다. 여기서, 앞에서 수행된 신경망 구별자의 판별 결과에 대한 확신을 갖기 위하여 복호화된 10-라운드 암호문 쌍을 추측 키 k를 사용해 추가적으로 1-라운드 복호화하여 9-라운드 암호문 쌍을 얻고 이를 6-라운드 신경망 구별자에 입력하여 다시 한번 판별을 수행한다. 반환된 출력 $score$가 임계값 c_2보다 클 경우 추측 키 k를 옳은 키로 반환한다. 앞서 설명한 A. Gohr의 신경망 구별자 기반 11-라운드 SPECK-32/64 키 복구 공격은 [그림 4]와 같이 표현할 수 있다.

신경망 구별자 기반 11-라운드 SPECK-32/64에 대한 부분 키 복구 공격 수행 결과, 100번의 시도 중 마지막 2개 라운드의 서브키 모두를 복구한 경우는 81번이었으며 마지막 라운드 서브키만을 복구한 경우는 99번이었다. 이때, 공격에 소요된 시간은 GTX 1080 Ti를 사용한 컴퓨터에서는 대략 20분 정도이며 단일 CPU 상에서는 약 12시간이었다. 키 복구 공격 수행



〔그림 4〕 신경망 구별자 기반 11-라운드 SPECK-32/64 키 복구 공격

시 평균적으로 $2^{13.2}$개의 암호문 쌍이 공격에 사용됐다. A. Gohr가 키 복구 공격 시 사용한 공격 알고리즘의 파라미터는 다음과 같다.

- 차분 δ: (0x0211/0A04)
- 중립비트: 14, 15, 20, 21, 22, 23
- 임계값 : c_1=15, c_2=100

위에서 설명한 A. Gohr의 키 복구 공격의 동작 방식은 추측한 키를 사용해 복호화한 결과가 암호문인지를 신경망 구별자에게 질의하고, 응답 결과가 암호문이라고 판별될 경우에 추측한 키를 옳은 키로써 반환하는 방식이다. 하지만, 키를 추측할 때 차분 특성을 만족하는 특정 키 비트를 부분적으로 탐색하는 전통적인 방식의 차분 분석과는 다르게 전수조사 방식으로 키를 추측한다. 이와 같은 특성은 키 복구 공격 실행시간 관점에서 상당한 부담이 되며, 키 크기가 클 경우 공격 수행이 불가능하다. 또한, 키 복구 공격 과정에서 중립비트를 통해 확장된 structure들을 사용한 암호문 쌍들은 모든 연산을 동일하게 적용하는 경우 비효율적이다. 이 두 가지를 개선하기 위해 A. Gohr는 전수조사를 더 효율적으로 수행할 수 있도록 하는 베이지안 최적화 기법과 structure를 통해 추가적으로 생성된 암호문 쌍들에 대한 연산량을 줄일 수 있도록 하는 UCB(Upper Confidence Bounds) 기법을 적용한 개선된 키 복구 공격을 추가적으로 수행했다.

개선된 키 복구 공격에서는 64개의 선택 평문 쌍에 대해 100개의 structure를 구성하여 진행했다. 개선된 키 복구 공격 수행 결과, 그래픽카드를 사용하지 않은 단일 쓰레드 상에서 소요된 시간은 100번 시도에서 평균적으로 약 8분(500.68초)이었다. 공격의 성공 여부는 마지막 2개 라운드의 서브키 중 마지막 서브키 추측은 모두 맞추고 두 번째 서브키는 hamming distance가 2 이내인(즉, 틀린 비트 수가 2보다 작음) 경우를 공격 성공으로 간주했으며 1,000번의 시도 중 521번 성공하였다. 개선된 키 복구 공격에 사용된 암호문 쌍의 개수는 12,800개이며 공격 알고리즘의 파라미터는 다음과 같다.

- 임계값 : c_1=5, c_2=10
- UCB exploration term : α=10
- 베이지안 탐색 파라미터 : iteration count l=5, candidate number n=32, iteration budget=500

〔표 3〕 〔9〕에서 수행된 암호해독 관점의 신경망 구별자 학습 원리 해석 실험

분류	실험명	설명
암호해독 관점	A	데이터셋 생성에 활용된 차분 특성의 확률이 높을수록 학습된 신경망 구별자의 정확도가 높은지를 확인하기 위한 실험.
	B	신경망 구별자가 암호문을 통해 마지막 라운드에 대한 정보뿐만 아니라 중간 라운드에 대한 정보를 학습하는지를 확인하기 위한 실험
	C	신경망 구별자의 성능이 중간 라운드에서 형성된 부정 차분 특성에 의존적인지를 확인하기 위한 실험
	D	학습시킬 암호의 라운드 수를 증가시킨 경우에도 신경망 구별자가 중간 라운드의 결과를 학습하는지를 확인하기 위한 실험

3.3. 신경망 구별자 해석 및 개선 연구

A. Gohr의 연구 결과는 전통적인 방식의 차분 구별자보다 우수한 성능을 보였지만 학습 과정이 블랙박스 형태로 수행되는 딥러닝 모델 특성상 사람이 이해하기 어렵다는 특징으로 인해 암호학적으로 해석이 어렵다. 이에 대한 해석을 수행하기 위한 연구가 EUROCRYPT'21에서 A. Benamira 외 3인[9]에 의해 발표되었다.

해당 연구에서는 신경망 구별자가 학습한 특성을 찾을 수 있고 이를 암호학적으로 해석할 수 있는지에 대한 해답을 찾기 위해 일련의 실험을 구성했다. 구성된 실험은 신경망 구별자의 학습 결과를 해석하기 위한 실험으로 암호해독 관점(Cryptanalysis perspective)과 기계학습 관점(Machine-learning perspective)으로 구분되어 수행되었다. 암호해독 관점에서의 해석 연구는 A. Gohr의 신경망 구별자가 학습 과정에서 학습한 특성을 식별하기 위한 실험으로 구성되며 기계학습 관점에서의 해석 연구는 A. Gohr의 신경망 구별자의 학습 모델을 기능에 따라 구조를 분할하고 각 부분을 다른 종류의 기계학습 모델로 대체할 수 있는지에 대한 실험으로 구성된다.

• 암호해독 관점에서의 해석

암호해독 관점에서의 해석은 [표 3]과 같은 4가지 가정과 실험을 통해 수행된다. 실험 A의 경우, A. Gohr의 연구에서 사용된 입력 차분 형태인 (0x0040/0000)으로부터 전파되는 차분 특성들의 출력 차분의 빈도수를 확인하여 가장 많이 등장한 출력 차분 형태가 신경망 구별자에서 좋은 성능을 보였는지를 확인했다. 이는 빈도수가 가장 큰 출력 차분 형태를 만족하는 데이터가 신경망 구별자에 의해 올바르게 구별된 비율을 통해 확인했다. 실험 수행 결과, 빈도수가 가장 크지 않은 출력 차분 형태에서 신경망 구별자가 더 좋은 성능을 보였다. 이를 통해 신경망 구별자가 학습하는 것은 단순히 출력 차분의 형태가 아님을 확인했다.

실험 B에서는 암호문을 1-라운드 부분 복호화한 중간 라운드 암호문을 대상으로 실험 A를 수행했다. 실험 수행 결과, 중간 라운드 암호문의 출력 차분 형태 중 빈도수가 가장 큰 출력 차분이 신경망 구별자의 성능이 더 좋았으며 올바르게 구별된 데이터의 비율을 확인한 결과 중간 라운드 출력 차분의 형태를 만족한 데이터의 99.98%가 올바르게 판별되었다. 이 실험 결과를 통해 신경망 구별자가 암호문의 출력 차분 뿐 아니라 중간 라운드 출력 차분도 학습하고 있음을 확인했다.

실험 C에서는 신경망 구별자가 중간 라운드 암호문의 출력 차분 형태에 얼마나 많은 영향을 받는지를 알기 위해 다양한 실험을 수행했다. 데이터셋의 각 데이터들의 비트별 분포를 확인하고 이를 다음과 같은 부정차분 형태로 유도하였다.

$$TD3 : 10*****00*****00\ 10*****00*****10$$
$$TD4 : 10*****10*****10\ 10*****10*****00$$

유도된 부정차분 형태를 만족하는 데이터들을 학습을 시켜 새로운 신경망 구별자를 구성하여 비교한 결과 기존 신경망 구별자와 동등한 성능을 보임을 확인했다. 이를 통해 신경망 구별자가 학습한 특성이 중간 라운드의 부정차분 특성임을 확인했다.

실험 D에서는 라운드 수가 증가된 경우에도 실험 C에서와 같이 중간 라운드의 부정차분 특성을 학습하는지 확인하기 위하여 라운드 수를 증가시켜 실험 C와 동일한 과정의 실험을 수행했다. 실험 수행 결과, 실험 C에서와 같이 신경망 구별자가 중간 라운드 부정차분 특성을 학습하는 것을 확인할 수 있었다.

• 기계학습 관점에서의 해석

기계학습 관점에서의 해석은 A. Gohr의 신경망 구별자의 구조를 암호학적 지식과 기계학습 지식을 기반으로 다른 구조의 기계학습 모델로 대체될 수 있는지 알기위한 목적으로 수행되었다. 이는 인공신경망의 학습원리가 사람이 이해하기 어렵다는 한계를 개선하기 위한 것으로 암호학적 측면과 기계학습 측면에서 해석력이 좋은 모델을 구성하기 위한 시도이다. 딥러닝 모델은 입력된 데이터와 출력 간의 상관관계를 찾아 데이터셋 구성 단계에서 전문가에 의해 구성된 특성보다 학습에 유리한 특성을 찾아 학습을 수행하며 이와 같은 동작 방식은 딥러닝 모델이 다른 기계학습 모델보다 좋은 성능을 보일 수 있도록 한다. 이는 상기 앞에서 설명한 암호해독 측면에서의 해석 연구에서도 살펴볼 수 있다. A. Gohr의 신경망 구별자의 데이터셋은 입력 차분 (0x0040/0000)을 기반으로 생성된 평문을 암호화하여 데이터셋을 구성했지만 실제로 딥러닝 모델이 데이터셋을 통해 학습한 특성은 입력 차분에 의해 구성된 암호문 쌍의 출력 차분이 아닌 중간 라운드의 부정 차분 특성이었다. 만약 딥러닝 모델이 학습 과정에서 입력된 암호문 쌍으로부터 중간 라운드의 부정 차분 특성을 유도한 과정을 모델링할 수 있다면, 이를 암호학적 관점에서 해석하고 데이터셋을 전처리하여 기계학습 모델이 학습하기 유리한 데이터셋을 구성할 수 있음을 의미한다.

또한, 기계학습 모델 중에서는 해석력 측면에서 딥러닝보다 유리한 다양한 모델이 존재한다. 해석력이 좋은 대표적인 기계학습 모델은 트리 모델이 있다. 트리 모델은 학습된 모델의 결과가 어떤 식으로 도출되었는지 구성된 트리를 통해 쉽게 이해할 수 있는 특징이 있다. 하지만, 대부분의 분야에서 딥러닝 모델은 트리 기반 모델보다 우수한 성능을 보인다. 이런 한계를 극복하기 위해 A. Benamira 외 3인은 신경망 구별자의 residual network 부분을 3가지 블록(Block 1, Block 2, Block 3)으로 구분하고 각 블록을 다른 기계학습 모델로 대체할 수 있는지에 대한 실험을 수행했다.

- Block 1: Residual network에서 initial convolution block을 말한다. Block 1은 input layer로부터 입력된 암호문 데이터에 대한 선형 조합을 통해 residual block의 학습이 용이하도록 데이터의 형태를 변환한다.
- Block 2: Residual network에서 실제적인 학습을 수행하는 residual block(또는 tower)을 말한다. Block 2는 Block 1을 통해 변형된 데이터를 입력받아 출력 데이터와의 연관성을 탐색하고 이를 학습한다.
- Block 3: 최종 출력 노드 직전의 classification block을 말한다. Block 3는 Block 2의 residual block을 통해 학습된 특성을 기반으로 입력 데이터에 대한 최종적인 결정을 내린다.

수행된 실험은 [표 4]에서와 같이 각 블록에 대하여 4가지 추측을 수행하고 이를 확인하는 방식으로 수행됐다. 먼저 1번 추측은 신경망 구별자가 다른 기계학습 모델에 비해 좋은 성능을 보일 것이라는 가정을 기반으로 이를 확인하기 위해 신경망 구별자가 학습한 동일한 데이터셋을 신경망 기반이 아닌 기계학습 모델에 학습시켜 성능을 비교했다. 비교 대상이 되는 기계학습 모델은 트리 기반으로 학습을 수행하는 앙상블 기계학습 모델(Ensemble-based machine learning model) 중 하나인 LGBM(Lightweight Gradient Boosting Machine) 모델을 사용했다. LGBM은 GBM(Gradient Boosting Machine)의 확장 버전으로 기존 GBM과 같이 작은 트리를 이어 붙이는 방식으로 학습을 수행한다. LGBM은 트리 기반의 모델이므로 학습 결과를 해석하기가 신경망에 비해 용이하다. 하지만, LGBM 학습 수행 결과, LGBM 기반으로 구성

〔표 4〕 [9]에서 수행된 기계학습 관점의 신경망 구별자 동작에 대한 4가지 추측

분류	대상	설명
1번 추측	모든 블록	신경망 구별자는 신경망 기반이 아닌 다른 기계학습 모델에 비해 좋은 성능을 보일 것이다.
2번 추측	Block 3	신경망 구별자 구조 중 Block 3는 다른 종류의 분류기로 대체할 수 있을 것이다.
3번 추측	Block 1	신경망 구별자 구조 중 Block 1은 입력된 암호문 쌍을 학습에 유리하도록 전처리할 것이다.
4번 추측	Block 2	Gohr의 인공신경망 모델 내부 구조 중 Block 2의 동작 방식은 Block1을 통해 전처리된 데이터의 확률 분포를 계산하고 압축할 것이다.

된 구별자의 성능은 신경망 구별자의 정확도인 92.9%에 비해 현저히 낮은 76.34%로 나타났다. 이와 같은 결과는 LGBM 뿐만 아니라 Random Forest, Support Vector Machine, Linear Regression 등과 같은 다른 기계학습 모델에서도 유사하게 나타났다. 이는 A. Gohr의 신경망 구별자가 신경망 기반이 아닌 다른 기계학습 모델보다 좋은 성능을 보인다는 것을 의미한다.

2번 추측부터는 1번 추측에 대한 실험 결과를 토대로 각 블록을 부분적으로 신경망 기반이 아닌 다른 종류의 기계학습 모델로 대체할 수 있는지에 대한 실험을 수행했다. 먼저 2번 추측의 경우, Block 3를 신경망 기반이 아닌 다른 종류의 분류기로 대체할 수 있는지에 대한 실험을 수행했다. LGBM을 Block 3에만 적용하여 분류기로 사용한 결과 91.49%의 정확도로 앞선 실험에서의 정확도인 76.34%에 비하여 정확도가 대폭 상승하였으며, 이는 A. Gohr의 신경망 구별자의 정확도인 92.9%와 크게 차이가 나지 않았고 전통적 방식의 차분 구별자의 정확도인 91.1%보다는 약 0.39% 정도 높았다. 이를 보다 개선하기 위해 신경망 구별자 Block 3의 첫 번째 레이어의 출력인 64-dimension 출력을 LGBM으로 분류시킨 결과 92.36%의 정확도로 A. Gohr의 신경망 구별자에 상응하는 정확도를 보였다. 이를 통해 Block 3는 트리 기반의 모델로 대체될 수 있음을 확인했다.

3번 추측과 4번 추측에서는 신경망 기반이 아닌 모델로 바로 대체될 수 없음에 따라 Block 1과 Block 2가 동작하는 방식을 모델링하여 다른 기계학습 모델을 사용하여 다시 구축하고자 하는 실험을 수행했다. 3번 추측은 Block 1에 대한 추측으로 입력된 암호문 쌍을 학습에 유리하도록 전처리한다는 것으로, 전처리되는 형태가 암호문 쌍의 왼쪽 워드 차분, 오른쪽 워드 차분의 출력 분포를 나타내는 선형조합일 것이라고 가정하였다. 이와 같은 가정에 따라 Block 1에서 입력되었을 때 레이어의 필터가 빈 경우가 많은 선형조합을 확인한 결과 $(\triangle L, \triangle V, V_0, V_1)$, $\triangle L = C_l \oplus C_l'$, $V_0 = C_l \oplus C_r$, $V_1 = C_l' \oplus C_r'$, $\triangle V = V_0 \oplus V_1$로 나타났다.

마지막 4번 추측에서는 3번 추측을 통해 유도한 선형조합 형태를 활용하여 Block 2의 동작 과정을 모델링하는 실험을 수행했다. 이 때, Block 2는 Block 1으로부터 전처리된 데이터의 확률 분포를 계산하고 이를 압축하는 방식으로 동작한다고 가정하였다. 이와 같은 방식을 모델링하여 3번 추측을 통해 유도한 선형조합 $(\triangle L, \triangle V, V_0, V_1)$에 대한 압축된 출력 분포를 계산하여 압축된 확률 분포 테이블 M-ODT(Masked-Output Distribution Table)를 유도했다. 이 때, 확률 분포 테이블을 압축하기 위해서 ODT에 중요도가 높은 비트들을 대상으로 마스킹을 하는 방식으로 수행했다. 이렇게 구성된 M-ODT는 신경망 기반이 아닌 기계학습에서 분류를 수행하기 위한 기준으로 생각할 수 있다. 구성된 M-ODT를 기반으로 LGBM 모델을 학습시킨 결과 기존 신경망 구별자의 정확도 92.9%에 상응하는 정확도를 보였다. 이를 통해 신경망 구별자가 신경망 모델이 아닌 다른 기계학습 모델로 모델링될 수 있음을 보였다.

Ⅳ. 고 찰

앞 장에서 살펴본 것 같이 신경망 암호해독 연구는 비교적 최근 시점인 2019년까지 딥러닝 기술을 암호해독에 적용하기 위한 형태로 수행되고 있었다. 인간의 오류를 뛰어넘는 성능을 보이며 성공적으로 딥러닝 기술이 활용되고 있던 타 분야에 대비하면 상대적으로 신경망 암호해독 기술은 연구가 미비한 상황이며 연구 초기 단계에 있다. CRYPTO'19 이전에 수행된 딥러닝 모델 적용 연구의 경우에는 다양한 Toy Cipher와 경량암호에 대한 적용 연구가 수행되면서 적용 가능성이 제시되었지만, 분석 대상 암호가 실제로 사용되지 않는 암호인 Toy Cipher를 대상으로 수행되었거나 블록 크기, 키 크기, 라운드 수가 상대적으로 작은 경량암호를 대상으로 수행되었다. Toy Cipher를 대상으로 수행된 연구의 경우에는 대부분 성공적으로 평문, 암호문 또는 키에 대한 정보를 추측할 수 있었지만 경량암호의 경우에는 대부분 무작위로 추측하는 것과 다르지 않은 결과를 보였다. 이는 블록암호의 블록 크기, 키 크기, 라운드 수가 큰 경우 적용하기 어렵다는 한계가 있다.

하지만 최근 발표된 A. Gohr의 신경망 구별자 기반 키 복구 공격 연구 결과는 실제로 사용되는 경량블록암호인 SPECK-32/64를 대상으로 하고 있으며 이전 연구 결과들에 대비하며 많은 라운드 수에 대해 분석을 수행했다. 분석 수행 결과, 11-라운드 라운드키에 대해서 키 복구를 성공적으로 수행했으며 대표적인 암호해독 기법인 차분 분석보다 우수한 결과를 보였다.

이는 암호해독 관점에서 실제로 활용하기에 유의미한 결과이며 암호해독 분야에서 딥러닝 기술을 활용할 수 있는 다양한 가능성을 제시했다. 하지만, 블랙박스 방식으로 동작하는 인공신경망의 특성상 어떤 원리로 학습이 수행되는지 알기 어렵다. 또한, 제안된 신경망 구별자 기반 키 복구 공격 모델은 키 추측 수행이 라운드키를 무작위로 추측하고 이를 신경망 구별자에 질의하는 방식으로 수행된다. 이는 질의를 수행해야하는 키 공간이 전수조사와 동일하다. A. Gohr의 연구에서는 이를 개선하기 위하여 베이지안 최적화, UCB 알고리즘 적용 등 최적화 기법을 적용하여 키 공간에 대한 사전확률 계산 및 연산 효율성 향상을 통해 성능을 개선하였지만, 키 공간이 큰 범용 암호의 경우에는 이런 방법론이 적용하기 매우 어렵다.

신경망 구별자의 학습 원리 및 결과 해석의 어려움은 A. Benamira 외 3인에 의해 수행된 신경망 구별자 해석 연구를 통해 암호해독/기계학습 관점에서 신경망 구별자가 학습한 특성을 해석하고 모델링할 수 있다는 것을 확인할 수 있었다. 하지만, 신경망 구별자의 동작 과정을 기계학습 관점에서 모델링하는 과정에서 사용한 M-ODT 생성을 위해 주요 비트를 추출할 때 사용하는 마스킹 값은 암호문 쌍을 학습한 인공신경망 결과를 기반으로 설정되었다. 이는 신경망 기반이 아닌 해석력이 좋은 기계학습 모델을 적용하여 A. Gohr의 신경망 구별자에 상응하는 모델을 만들 수는 있지만 이런 모델을 만들기 위해서는 인공신경망의 도움이 필요하다는 것을 의미하며 마스킹 값에 상응하는 비트들이 왜 주요 비트인지에 대해서는 알기 어렵다.

위에서 설명한 신경망 암호해독 연구들에 대한 한계점을 개선하기 위해서는 블록 크기, 키 크기, 라운드 수가 상대적으로 큰 범용 암호를 대상으로 연구가 수행되어야 하며, 설명가능한 인공지능(XAI, eXplainable-Artificial Intelligence)를 기반으로한 신경망 구별자 연구가 필요하다.

V. 결 론

본 논문에서는 현재까지 수행된 블록암호에 대한 신경망 암호해독 연구에 대한 동향 분석을 수행했다. 동향 분석 수행 결과, 딥러닝 기술을 단순히 적용하여 평문, 암호문 또는 키와 관련된 정보를 추측하는 형태에서 암호학적 특성을 기반으로 만들어진 데이터셋을 학습시킨 모델을 통해 부가적인 정보를 획득한 후 이를 기반으로 키 복구 공격 모델을 설계하는 형태로 발전하고 있음을 확인할 수 있었다. 또한, 2019년 이전에 수행된 연구 결과는 실제로 암호해독에 활용하기에 어려움이 있었지만 최근 발표된 신경망 구별자를 활용한 연구를 통해 암호해독 관점에서 유의미한 결과를 얻을 수 있음을 알 수 있었다. 하지만, 블랙박스 형태로 동작하는 딥러닝 기술이 적용된 타 분야 대비 아직 연구 초기 단계에 있어 연구 방법론이 정립되지 않아 기존 연구 결과들에 대해 범용 암호 적용 어려움, 학습 원리 파악 어려움 등 한계점들이 존재했었다. 이를 개선하기 위한 지속적인 신경망 암호해독 연구가 필요하다.

참 고 문 헌

[1] B. Seunggeun, and K. Kwangjo, "Recent advances of neural attacks against block ciphers." *2020 Symposium on Cryptography and Information Security (SCIS 2020)*. IEICE Technical Committee on Information Security, 2020.

[2] G. Mishra, S. K. Murthy, and S. K. Pal, “Neural network based analysis of lightweight block cipher PRESENT.” *Harmony Search and Nature Inspired Optimization Algorithms*, pp. 969-978, 2019.

[3] A. Jain, and G. Mishra, “Analysis of lightweight block cipher FeW on the basis of neural network.” *Harmony Search and Nature Inspired Optimization Algorithms*, pp. 1041-1047. 2019.

[4] Y. Xiao, Q. Hao, and D. D. Yao, “Neural cryptanalysis: Metrics, methodology, and applications in cps ciphers.” *2019 IEEE Conference on Dependable and Secure Computing (DSC)*, pp. 1-8. 2019.

[5] B. Chandra, and P. P. Varghese, “Applications of cascade correlation neural networks for cipher system identification.” *World Academy of Science, Engineering and Technology*, 26, pp. 312-314, 2007.

[6] A. M. Albassal, and A. M. Wahdan, “Neural net-

work based cryptanalysis of a feistel type block cipher.” *International Conference on Electrical, Electronic and Computer Engineering, 2004. ICEEC'04,* pp. 231-237. IEEE, 2004.

[7] M. Danziger, and M. A. A. Henriques, “Improved cryptanalysis combining differential and artificial neural network schemes”. *2014 International Telecommunications Symposium (ITS)*, pp. 1-5. IEEE, 2014.

[8] A. Gohr, “Improving attacks on round-reduced speck32/64 using deep learning”. *Annual International Cryptology Conference*, pp. 150-179. Springer, Cham, 2019.

[9] A. Benamira, D. Gerault, T. Peyrin, and Q. Q. Tan, “A deeper look at machine learning-based cryptanalysis.” *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 805-835. Springer, Cham, 2021.

〈저자소개〉

석 병 진 (Byoungjin Seok)

학생회원

2017년 8월 : 서울과학기술대학교 컴퓨터공학과 학사

2019년 2월 : 서울과학기술대학교 컴퓨터공학과 석사

2019년 3월~현재 : 서울과학기술대학교 컴퓨터공학과 박사과정

<관심분야> 정보보호, 암호해독, 디지털포렌식

이 창 훈 (Changhoon Lee)

종신회원

2001년 : 한양대학교 자연과학부 수학전공 학사

2003년 : 고려대학교 정보보호대학원 석사

2008년 : 고려대학교 정보경영전문대학원 정보보호전공 박사

2008년 4월~2008년 12월 : 고려대학교 정보보호연구원 연구교수

2009년 3월~2012년 2월 : 한신대학교 컴퓨터공학부 조교수

2012년 3월~2015년 3월 : 서울과학기술대학교 컴퓨터공학과 조교수

2015년 4월~2020년 3월 : 서울과학기술대학교 컴퓨터공학과 부교수

2020년 4월~현재 : 서울과학기술대학교 컴퓨터공학과 교수

<관심분야> 정보보호, 암호학, 사이버 보안, 디지털포렌식 등