Journal of Digital Convergence
Vol. 17. No. 7, pp. 285-292, 2019

ISSN 1738-1916
https://doi.org/10.14400/JDC.2019.17.7.285

신경망 학습에서 프라이버시 이슈 및 대응방법 분석

홍은주[1], 이수진[2], 홍도원[3], 서창호[3*]
[1]공주대학교 융합과학과 박사과정, [2]공주대학교 수학과 석사과정, [3]공주대학교 응용수학과 교수

Analysis of privacy issues and countermeasures in neural network learning

Eun-Ju Hong[1], Su-Jin Lee[2], Do-won Hong[3], Chang-Ho Seo[3*]
[1]Ph.D Course, Dept. of Convergence Science, Kongju National University
[2]M.S Course, Dept. of Mathematics, Kongju National University
[3]Professor, Dept. of Applied Mathematics, Kongju National University

요 약 PC, SNS, IoT의 대중화로 수많은 데이터가 생성되고 그 양은 기하급수적으로 증가하고 있다. 거대한 양의 데이터를 활용하는 방법으로 인공신경망 학습은 최근 많은 분야에서 주목받는 주제이다. 인공신경망 학습은 음성인식, 이미지 인식에서 엄청난 잠재력을 보였으며 더 나아가 의료진단, 인공지능 게임 및 얼굴인식 등 다양하고 복잡한 곳에 광범위하게 적용된다. 인공신경망의 결과는 실제 인간을 능가할 정도로 정확성을 보이고 있다. 이러한 많은 이점에도 불구하고 인공신경망 학습에는 여전히 프라이버시 문제가 존재한다. 인공신경망 학습을 위한 학습 데이터에는 개인의 민감한 정보를 포함한 다양한 정보가 포함되어 악의적인 공격자로 인해 프라이버시가 노출될 수 있다. 공격자가 학습하는 도중 개입하여 학습이 저하되거나 학습이 완료된 모델을 공격할 때 발생하는 프라이버시 위험이 있다. 본 논문에서는 최근 제안된 신경망 모델의 공격 기법과 그에 따른 프라이버시 보호 방법을 분석한다.

주제어 : 인공신경망, 프라이버시, 차분프라이버시, 동형암호, 공격

Abstract With the popularization of PC, SNS and IoT, a lot of data is generated and the amount is increasing exponentially. Artificial neural network learning is a topic that attracts attention in many fields in recent years by using huge amounts of data. Artificial neural network learning has shown tremendous potential in speech recognition and image recognition, and is widely applied to a variety of complex areas such as medical diagnosis, artificial intelligence games, and face recognition. The results of artificial neural networks are accurate enough to surpass real human beings. Despite these many advantages, privacy problems still exist in artificial neural network learning. Learning data for artificial neural network learning includes various information including personal sensitive information, so that privacy can be exposed due to malicious attackers. There is a privacy risk that occurs when an attacker interferes with learning and degrades learning or attacks a model that has completed learning. In this paper, we analyze the attack method of the recently proposed neural network model and its privacy protection method.

Key Words : Artificial Neural Network, Privacy, Differential Privacy, Homomorphic Encryption, attack

*This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT(2019R1A2C1003146).
*Corresponding Author : Changho Seo(chseo@.kongju.ac.kr)
Received April 11, 2019 Revised May 15, 2019
Accepted July 20, 2019 Published July 28, 2019

1. 서론

PC, SNS, IoT의 대중화로 여러 분야에서 다양한 데이터가 생성되어 그 양은 기하급수적으로 증가하고 있다. 데이터가 자산이 되는 시대에 방대한 데이터에서 얻어지는 정보는 다양한 잠재력을 갖는다. 데이터에 포함된 정보를 얻기 위해 데이터는 수집 및 분석된다. 신경망 학습은 데이터를 분석하는 방법으로 최근 다양한 분야에서 연구되고 주목받는 주제이다. 신경망이란 인간의 뇌가 문제를 해결하는 방식을 컴퓨터에 접목시킨 것이다. 생물학적인 신경망과의 구분을 위해 인공을 덧붙여 인공신경망이라고 한다. 신경망 학습은 데이터를 학습하고 올바르게 예측하여 실생활의 다양한 응용프로그램에서 문제를 해결한다. 현재 이미지 인식과 음성인식에서 엄청난 발전 가능성을 보였으며 더 나아가 스마트 팜, 의료진단, 인공지능 게임 등 다양한 분야에 광범위하게 적용된다. 실제 신경망의 학습결과는 인간을 능가할 정도의 정확성을 보이고 있다.

최근에 고객이 직접 학습에 참여하여 본인의 맞춤형 모델을 생성하기 위한 신경망 기반 기계학습 서비스(Machine Learning as a Service : MLaaS)가 제공되고 있다[1]. 고객은 온라인 예측서비스를 제공하는 클라우드를 사용하여 모델을 다운로드하고 정보를 제공하면 예측 서비스에서 예측된 정보를 고객에게 제공한다. 실제로 구글, 애플, 아마존과 같은 대기업에서는 대규모 신경망을 구축하기 위해 고객의 데이터를 수집하여 이용한다. 많은 이점에도 불구하고 프라이버시 문제는 여전히 존재하고 있다. 신경망 학습을 위해 사용되는 많은 양의 학습 데이터에는 개인의 민감한 정보를 포함한 다양한 정보가 있다. 따라서 학습된 모델에서 악의적인 공격자에 의해 중요한 정보가 노출될 수 있다. 실제로 M. Fredrikson 등이 모델 전도 공격을 통해 이미지 복구를 성공시켜 문제점의 실현 가능성을 입증하였다[2]. 최근 연구결과에서도 신경망에 포함된 개인 정보가 성공적으로 복구될 수 있었고 많은 논문에서 이러한 프라이버시의 문제를 제기하고 있다. 이에 따른 프라이버시 보호 해결 방안이 제안되었지만 여전히 연구는 초기 단계에 있다. 본 논문에서는 최근 신경망 학습의 각 단계에서 발생하는 프라이버시 공격 기법에 대한 소개와 프라이버시 위협에 대처하기 위한 방안으로 프라이버시를 보호하는 모델과 신경망 학습을 결합하는 보호방식의 최근 연구에 대해 분석한다.

2. 연구배경

본 절에서는 인공신경망에 대해 조사하고, 프라이버시를 보호하는 기술로 차분 프라이버시와 동형 암호 및 안전한 다자간 계산의 정의를 정리한다. 더 나아가 프라이버시 보호 기술에 신경망을 결합한 기술들을 정리한다.

2.1 인공신경망(Artificial Neural Network : ANN)

인간의 뇌에서 뉴런과 뉴런이 시냅스에 연결되어 문제를 처리하는 것과 같이 컴퓨터에서도 뉴런과 뉴런이 연결되어 네트워크를 구성할 때 이를 신경망이라 한다. 생물학에서 쓰이는 신경망과 구분하기 위해 인공신경망이라 한다. 최근 신경망 모델은 기계학습에서 광범위하게 사용되고 있다. 신경망은 크게 두 단계인 학습 단계와 예측 단계로 나눌 수 있다. 학습 단계는 많은 데이터를 가지고 신경망을 학습하는 과정이고, 예측 단계는 학습된 모델을 사용하여 새로운 입력에 대한 예측을 수행하는 단계이다. 예측을 통해 신경망이 얼마나 잘 학습되었는지 확인할 수 있다.

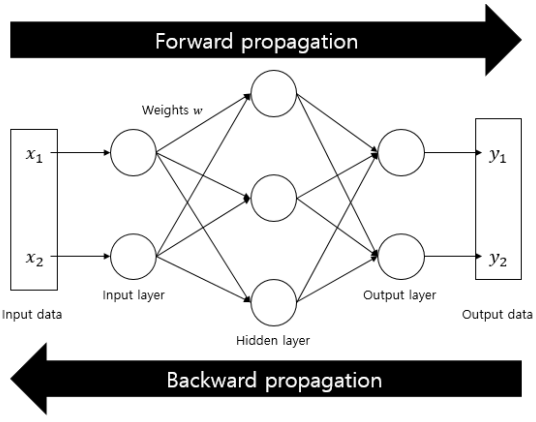

Fig. 1. Neural Network

2.1.1 신경망 학습 단계

Fig. 1은 세 개의 층을 가지는 일반적인 신경망을 나타낸다. 왼쪽부터 입력층(Input layer), 은닉층(Hidden layer), 및 출력층(Output layer)으로 구성되어 있다. 원은 뉴런을 나타내고 왼쪽부터 오른쪽으로 신호를 전달하면 순전파(Forward propagation)라 하고, 반대의 신호에 대해서는 역전파(Backward propagation)라고 한다. 신경망 학습은 입력과 함께 신호에 가중치와 편향을

추가하여 진행한다. 이때 입력의 총합을 출력으로 변환하는 함수를 활성화 함수라고 한다. 대표적인 활성화 함수에는 시그모이드(Sigmoid), 쌍곡선 탄젠트(Hyperbolic tangent), 렐루(ReLU), 소프트맥스(Softmax)가 있다. 신경망을 통해 출력된 예측 값과 실제 값의 차이를 측정하는 지표로 손실 함수를 사용한다. 손실 함수를 통해 신경망이 얼마나 잘 학습되었는지 확인할 수 있다. 대표적으로 손실 함수에는 평균제곱 오차와 교차 엔트로피 오차를 사용한다. 신경망 학습에서의 핵심은 예측 값과 실제 값의 차이를 줄이기 위한 최적의 가중치를 찾는 것이다. 따라서 가중치와 편향을 업데이트해야 한다. 그러나 전체 데이터에 대하여 매개변수를 조정하는 방법은 매우 복잡하고 어렵다. 따라서 미니- 배치 확률 경사 하강법(mini-batch Stochastic Gradient Descent) 알고리즘이 적용된다. 경사 하강법은 임의의 지점에서 초기 가중치를 설정하여 그래디언트(Gradient) 갱신을 통해 최적으로 수렴할 때까지 반복한다. 신경망 학습에서 초기 가중치 값은 Xavier 초깃값과 He 초깃값을 사용한다. 적절한 초깃값을 선택하면 각 층의 활성화 분포가 적절하게 분포되어 학습이 원활하게 진행된다.

2.1.2 신경망 예측 단계

신경망 학습에서 예측 단계는 새로운 데이터를 학습된 모델에 적용시켜 예측을 수행하는 것이다. 대표적으로 분류와 회귀 문제가 있다. 둘 중 어떤 문제를 다루느냐에 따라 출력층에서의 활성화 함수가 달라진다. 학습된 모델에서 분류는 데이터가 어떤 클래스에 속하는지 확인하는 문제로 소프트맥스 함수를 사용한다. 회귀는 입력 데이터에 대해 수치를 예측하는 문제로 항등 함수를 사용한다. 신경망 예측의 최종 목표는 올바른 예측을 하는 것이다.

2.1.3 인공신경망의 종류

- 합성곱 신경망(Convolution Neural Network : CNN) 기존의 신경망과 달리 합성곱 신경망은 데이터 형상을 유지한다[3]. 즉, 공간데이터를 학습할 수 있다. 예를 들어, 기존 신경망에 3차원 데이터를 학습하기 위해 3차원 데이터를 평평한 1차원으로 표현해야 했다면 CNN은 3차원 데이터를 그대로 사용할 수 있다. 따라서 CNN은 시각적 이미지 분석에 매우 효과적이다.
- 순환 신경망(Recurrent Neural Network : RNN) 순환적 관계를 갖는 신경망으로 시계열 데이터, 자연어와 같이 연속된 데이터를 학습하기 위한 인공신경망이다. RNN은 장기 기억력을 가지지 못하기 때문에, 이에 발생하는 기울기 소실 문제가 있다. 따라서 이를 해결하기 위해 RNN의 발전된 모델인 RNN 구조인 장/단기 기억 네트워크(Long-Short Term Memory Network)가 있다[4].
- 생성적 적대 신경망(Generative Adversarial Network : GAN) 거짓 데이터를 생성하는 생성자(Generator) 모델과 데이터의 진위 여부를 결정하는 감별자(Discriminator) 모델로 구성되어있다[5]. GAN은 영상 합성 및 이미지 생성에 활용된다.

2.2 차분 프라이버시(Differential Privacy : DP)

차분 프라이버시는 강력한 프라이버시 보호 기술로 Dwork에 의해 제안되었다[6]. 차분 프라이버시에 적용된 프라이버시 개념은 하나의 레코드 차이를 갖는 데이터베이스에서 특정 개인의 데이터 포함 여부와 관계없이 질의에 대한 응답에 차이가 없어 공격자가 구분하지 못한다는 것이다. 따라서 공격자는 자신의 배경지식을 제외한 특정 개인에 대하여 더 많은 정보를 추론할 수 없다. 이에 따른 차분 프라이버시의 정의는 다음과 같다.

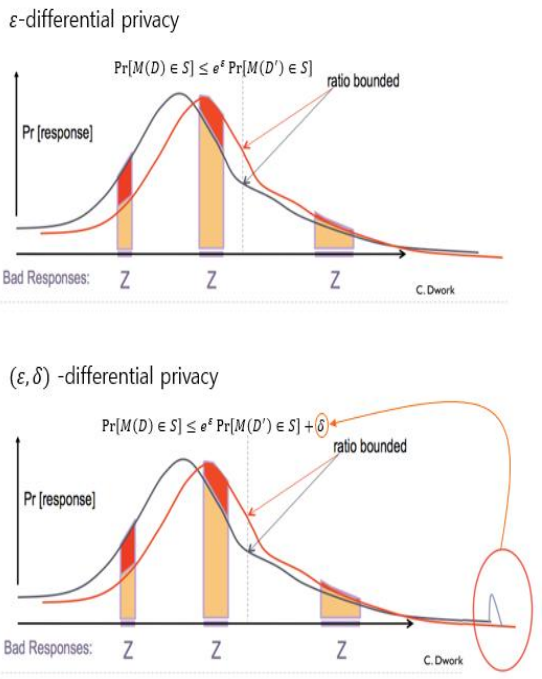

Fig. 2. Differential Privacy[7]

정의1) 이웃한 데이터베이스

데이터베이스 x의 $l_1-\text{norm}(\|x\|_1=\sum_{i=1}^{|x|}|x_i|)$은 해당 데이터베이스의 크기이며 레코드의 수를 의미한다. 이를 통해 두 데이터베이스 x, y의 $l_1-\text{norm}$은 데이터베이스 x와 y 사이의 레코드 차이로 $\|x-y\|_1$로 표현된다. 이때, $\|x-y\|_1=1$ 이면, x와 y는 이웃한 데이터베이스라 한다.

정의2) $(\varepsilon,\delta)-$차분 프라이버시

임의의 알고리즘 M에 대해 하나의 레코드 차이를 가지는 데이터베이스 D와 D'가 다음을 성립하면, $(\varepsilon,\delta)-$차분 프라이버시를 만족한다.

$$\Pr[M(D)\in S]\le e^{\varepsilon}\Pr[M(D')\in S]+\delta \quad (1)$$

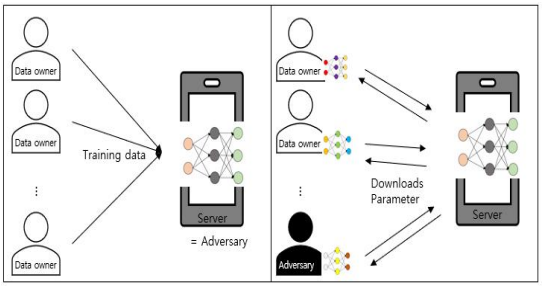

Fig. 3. Centralized and Distributed training model

이때, $\delta=0$이면, $\varepsilon-$차분 프라이버시를 만족한다고 한다. Fig. 2는 각 데이터베이스의 출력 값의 분포에 대한 $\varepsilon-$차분 프라이버시 및 $(\varepsilon,\delta)-$차분 프라이버시 관점에서의 차이 값을 나타낸다[7].

2.3 동형 암호(Homomorphic Encryption : HE)

서버에 저장된 데이터에서 개인 정보들은 암호화되어 보관된다. 암호화된 정보들은 데이터베이스가 노출되더라도 안전하여 보안을 유지할 수 있다. 암호화된 데이터를 활용하기 위해서는 복호화 과정을 거쳐야 하지만 빅데이터의 경우, 많은 양을 복호화 하는 것은 시간적, 기술적, 비용적으로 부담이 된다. 또한 복호화 하기 위한 비밀키의 노출이라는 보안적 리스크가 존재한다. 이를 해결하기 위한 암호기술로 암호화된 데이터를 복호화 하지 않고도 연산이 가능하게 하는 암호체계를 동형 암호라고 한다. 임의로 정의된 선형 연산 $\circ$ 및 암호화 (Enc)에 대하여 메시지 m_1, m_2의 동형암호는 다음과 같다.

$$Enc(m_1)\circ Enc(m_2)=Enc(m_1\circ m_2) \quad (2)$$

C. Gentry는 부분적인 동형 암호(Somewhat Homomorphic Encryption)를 기반으로 재부팅 방법(Bootstrapping Method)을 활용한 완전 동형 암호(Fully Homomorphic Encryption : FHE)를 제안하였다[8]. 이후 동형 암호는 C. Gentry의 완전 동형 암호를 기반으로 효율성을 개선하는 방안으로 연구가 진행되고 있다[9].

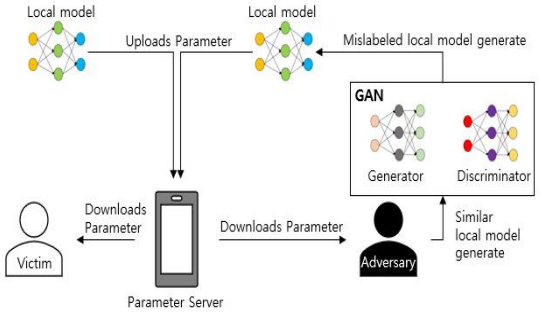

Fig. 4. GAN attack

2.4 안전한 다자간 계산(Secure Multiparty Computation : SMC)

안전한 다자간 계산은 n명의 참가자 $p_1, p_2, \ldots, p_n$은 각각 개인의 입력데이터 $d_1, d_2, \ldots, d_n$을 개인 정보로 간주하여 비밀로 유지하면서 어떤 함수 값 $F(d_1, d_2, \ldots, d_n)$을 계산하려고 하는 것이다[10]. 이상적인 SMC는 믿을 수 있는 제3자에게(Trust Third Party, TTP) 개인의 입력 데이터를 전달하고 TTP는 함수 F의 출력값을 각 참가자에게 전송해주는 방식이다. 신경망에서 안전한 다자간 계산은 왜곡 회로(garbled circuits), 비밀 공유(secret sharing), 동형 암호(homomorphic encryption)을 기반으로 한다[11].

3. 신경망에 대한 프라이버시 이슈

신경망은 학습 단계와 예측 단계로 나누어진다. 학습

단계에서는 Fig. 3와 같이 참가자가 중앙 서버에 데이터를 제출하여 신경망 학습을 하는 중앙 집중식 학습과 참가자들이 학습에 직접 참여하는 분산 학습으로 분류할 수 있다. Table 1은 신경망 학습에 대한 단계별 프라이버시 공격을 나타낸다. 다음은 각 단계에서의 프라이버시 이슈에 대해 분석한다.

3.1 학습 단계에서의 프라이버시 이슈

- 중앙 집중식 학습에서 서버의 공격: 학습 서비스를 제공하는 회사는 반 정직(semi-honest)으로 회사의 주목적은 정확한 신경망 모델을 달성하여 고품질의 서비스를 제공하는 것이다. 회사(공격자)는 참가자(고객)의 학습 데이터에서 중요 정보를 추출하고 학습할 수 있다.
- 분산 학습에서 GAN 기반 공격 : 활동적인 공격자는 자신이 소유하지 않은 학습 데이터를 추출하려는 시도를 통해 학습과정을 방해한다. 예를 들어 GAN 기반 공격이 있다[5]. 공격자는 학습에 참여하는 참가자이다. 모든 학습 참가자는 공통된 학습 목표에 사전 동의를 한다. 이는 신경망 구조의 유형과 학습이 이루어지는 것에 대해 동의한다는 것을 의미한다. Fig. 4와같이 공격자는 GAN을 사용하여 희생자의 샘플처럼 보이는 로컬 모델을 생성한다. 공격자는 생성된 로컬 모델의 레코드의 수정을 통해 수정된 로컬 모델의 매개변수를 분산학습 절차에 업로드한다. 희생자는 실제와 수정된 모델을 구분하기 위해 본인이 처음 의도했던 것보다 더 많은 정보를 공개하여 공격자에게 학습 데이터를 노출시킬 수 있다.
- 분산 학습에서 그래디언트 공격: 분산 학습은 참가자가 자신의 데이터를 활용하여 신경망 모델을 학습하고 서버에 전송하면 서버에서 참가자들의 학습된 신경망을 통해 전체 신경망 모델을 최적화한다. 이 방법은 참가자들이 모델을 동시에 학습할 수 있으며, 데이터 집합을 공유하지 않고 학습이 가능하다. 서버는 전체 신경망 모델을 최적화하기 위해 참가자 신경망 모델의 그래디언트를 공유한다. 참가자는 자신의 신경망 학습 후 본인의 그래디언트를 중앙 서버에 개별적으로 공유한다. 서버는 모든 그래디언트의 합계를 사용하여 최적값에 대한 그래디언트를 계산하고 참가자는 서버의 최신 그래디언트를 다운로드하여 자신의 모델을 수정하는데 사용한다. 그러나 그래디언트의 작은 값만으로도 참가자의 데이터를 복구할 수 있다는 것이 입증되었다[12].

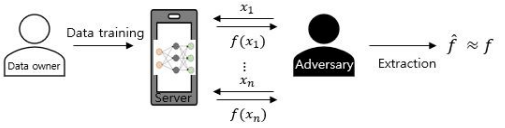

Fig. 5. Model Extraction Attack

3.2 예측 단계에서의 프라이버시 이슈

예측 단계에서는 학습된 모델에 대하여 모델의 정보 없이 예측 질의만 수행할 수 있는 블랙박스 공격자와 예측 질의뿐 만 아니라 모델의 정보 및 부분적인 학습 데이터까지 얻을 수 있는 화이트 박스 공격자로 나뉜다. 블랙박스와 화이트박스의 공격 방식을 기반으로 한 공격은 다음과 같다.

- 모델 전도 공격(Model Inversion Attacks) : 화이트박스와 블랙박스 방식을 기반으로 공격자가 학습된 모델을 사용하여 학습에 사용된 데이터를 발견할 때 발생하는 공격이다. M. Fredrikson 등은 이미지를 복구하여 모델 전도 공격이 가능함을 보여주었다[2].
- 멤버십 추론 공격(Membership Inference Attacks) : 블랙박스 공격 방식으로 주어진 데이터 레코드가 특정 모델의 학습 데이터 집합의 일부인지 추론하는 공격이다[13]. 공격자는 주어진 레코드가 특정 모델을 학습하는데 사용됨을 알게 되면 모델을 통한 정보 유출이 가능해진다.
- 모델 추출 공격(Model Extraction Attacks) : Fig. 5에서와 같이 공격자는 희생자 모델에 블랙박스 방식으로 접근하여 신경망 모델에 n개의 질의 및 응답을 통하여 희생자 모델과 거의 동일하거나 유사한 모델을 추출하려고 시도할 때 발생한다. F Tramèr 등은 실제로 아마존과 BigML를 비롯한 MLaaS 제공 업체를 대상으로 기계학습의 다양한 모델에 대한 성공적인 모델 추출 공격을 입증하였다[14]. 대부분의 경우 매우 근접한 모델을 추출했다.

4. 신경망에 대한 프라이버시 대응 방안

학습 단계와 예측 단계에서 프라이버시 공격 모델을 방어하기 위해 차분 프라이버시와 동형 암호 및 안전한

다자간 계산을 결합한 대응 방안에 대해 분석한다.

4.1 학습 단계에서의 대응 방안

- 중앙 집중식 학습에서 안전한 다자간 계산 : P. Mohassel 등은 단일 서버 학습방법을 두 개의 서버 모델로 확장하여, 데이터 소유자는 안전한 2자 간 계산을 사용하여 공동 데이터에 대한 다양한 모델을 학습하는 두 개의 비-공모 서버에 개인 데이터를 배포한다[15]. P. Mohassel 등은 데이터 전송 프로세스를 안전하게 보호하기 위해 안전한 다자간 계산에 사용하기 편한 새로운 활성화 함수를 제안한다.
- 분산학습에서 DPGAN : Xie 등은 DPGAN을 제안했다[16]. DP 판별자를 얻기 위해 판별자의 그래디언트에 노이즈를 추가하고 그 판별자로 학습된 생성자(generator)는 post-processing 이론에 기반 하여 DP를 만족한다.
- Boneh Goh Nissim HE 기반 : J. Yuan와 S. Yu는 여러 참가자가 서버에서 계산한 암호화된 각각의 가중치 업데이트 결과를 공동으로 복호할 수 있는 메커니즘을 제안했다. 학습 과정에서 각 참가자는 자신의 데이터를 주어진 시스템의 공개키를 사용하여 암호화한 다음 암호문을 서버에 업로드한다. 서버는 동형 암호 성질에 따라 암호문에 대한 학습과정을 실행하고 암호화된 결과를 참가자에게 전송한다. 참가자들은 가중치를 업데이트하는 결과를 공동으로 복호할 수 있다. 이 과정에서 서버가 나머지 참가자와 공모하는 경우라도 참가자의 개인 정보를 알 수 없다[17].
- Multi-Key Fully Homomorphic Encryption (MKFHE) : P. Li 등은 다중-키 완전 동형 암호화에 기반 한 문제를 다루었다. 각 참가자들은 키 쌍을 공유하지 않고 자체적인 키 쌍을 가진다. 참가자는 자신의 데이터를 자신의 공개키로 암호화하여 신뢰할 수 없는 서버로 전송한다. 서버는 암호화 데이터를 학습과정을 거쳐 암호화된 결과를 참가자에게 전송한다. 참가자는 암호화된 결과를 복호하기 위해 안전한 SMC 프로토콜을 수행한다[18].

참가자와 서버 사이에 공유된 그래디언트는 학습 데이터 집합의 레코드에 대한 많은 정보를 포함하므로 정보고 누출되지 않도록 분산학습에서 그래디언트 보호를 위한 해결방안으로 다음과 같은 스킴이 제안되었다.

- 로컬 DP솔루션(DPSGD) : DP의 정의를 통해 참가자의 그래디언트에 노이즈를 추가하여 안전하게 보장해준다[19]. 그러나 과도한 노이즈는 학습과정을 크게 저해시켜 모델의 유용성이 낮아진다.
- Differentially Private Generative Model(DPGM) : 생성 모델은 많은 양의 정보를 기반으로 하는 응용프로그램에서 사용된다. 모델을 훈련하는데 사용되는 개인 정보는 침해될 수 있기 생성된 모델을 게시하거나 공유하는 것은 쉽지 않다. G. Acs 등은 생성 모델과 생성된 전체 고차원 데이터를 공개하는 새로운 기술을 제시한다[20].
- Concentrated DP 적용 : L. Yu 등은 학습된 모델을 게시하고 공유하기 위해 신경망을 학습시키는 DP 접근법을 제안한다. 신경망의 학습은 반복 횟수가 많기 때문에, CDP를 적용하여 프라이버시 손실에 대한 엄격한 평가를 한다. 또한 두 가지 데이터 배치 방법을 통해 각각에 대한 프라이버시 보호 방법을 제안하고 정확한 프라이버시 손실 예측을 가능하게 한다[21,22].
- Additively HE(AHE) 기반 그래디언트 보호 : 참가자는 정직한 것으로 간주되며 AHE의 스킴에 따라 공개키와 비밀키를 공유한다[9]. 공개키는 참가자들의 그래디언트를 암호화하는데 사용된다. AHE 속성에 따라 암호문을 통해 서버는 모델에 업데이트를 할 수 있다. 참가자는 암호문에 해당 그래디언트를 다운로드하여 자신의 비밀키로 암호를 해독한다.
- 차분 프라이버시와 동형 암호 결합 : X. Zhang 등은 미니 배치에서 전체 그래디언트는 참가자들의 그래디언트의 합으로 계산될 수 있다고 지적했다[23]. DP 메커니즘이 대칭분포에서 샘플링한 노이즈를 적용하면 참가자들의 그래디언트를 집계할 때 노이즈의 대부분이 상쇄되어 전체 그래디언트를 잘 추정할 수 있음을 알게 되었다. 따라서 HE를 기반으로 전체 그래디언트를 계산한다.
- 보안 집계(Secure Aggregation) 활용한 그래디언트 계산 : K. Bonawitz 등은 참가자들의 신경망 학습 모델에서 서버로 업데이트 한 로컬 그래디언트의 합계를 안전한 방식으로 계산하여 시스템의 개인 정보보호를 향상시키기 위한 프로토콜을 설계하였다[24].

4.2 예측 단계에서의 대응 방안

학습된 모델에 대해 공격을 방어하는 방법은 학습된 모델이 강력한 프라이버시를 보장하게 하는 것이다. 따라서 모델을 학습하는 과정에서 DP를 적용시킨다[19]. 엄격한 프라이버시를 보장하는 DP가 적용된 신경망 모델은 강력한 화이트 박스 공격자를 방어할 수 있다.

5. 결론

신경망 학습은 데이터를 학습하고 올바르게 예측하여 실생활에서 많은 문제를 해결한다. 많은 장점에도 불구하고 신경망 학습에는 여전히 프라이버시 문제가 남아있다. 따라서 본 논문에서는 지금까지 신경망 학습의 프라이버시 이슈와 이에 대응하는 해결방안을 분석하였다. 신경망 학습은 학습 단계와 예측 단계로 분류되며 학습 단계에서는 중앙 집중형 학습과 분산 학습으로 분류되고 각 공격들은 기존의 프라이버시 보호 기술 DP, HE, SMC와 신경망 학습의 결합을 통해 프라이버시를 보호한다[9,10,19]. 화이트박스와 블랙박스 기반인 예측 단계에서는 모델 전도 공격, 멤버십 추론 공격, 모델 추출 공격에 대해 조사하였다[2,13,14]. 최신의 여러 공격 방법 및 해결방안을 통해 현재 활발한 연구가 진행되고 있음을 확인하였다. 하지만 프라이버시 보호 기술과 결합한 해결방안들의 연구는 실제 구현을 고려하면 아직 부족하다.

REFERENCES

[1] M. Ribeiro, K. Grolinger & M. A. M. Capretz. (2015). MLaaS: Machine Learning as a Service. *In IEEE International Conference on Machine Learning and Applications (ICMLA), p. 896-902.*

[2] M. Fredrikson, S. Jha & T. Ristenpart. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *In CCS,* (pp. 1322-1333). USA : ACM.

[3] A. Krizhevsky, I. Sutskever & G. E Hinton. (2012). Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems, 1097-1105.*

[4] S. Hochreiter & J. Schmidhuber. (1997). Long short-term memory. *Neural computation 9(8), 1735-1780.*

[5] B Hitaj, G Ateniese & F Perez-Cruz. (2017). Deep Models under the GAN: Information Leakage from Collaborative Deep Learning. *Proc. (pp. 603-618). ACM CCS.*

[6] C. Dwork & A. Roth. (2013). The algorithmic foundations of differential privacy. *Theoretical Computer Science, 9(3-4), 211-407.*

[7] K. Ligett. (2017). Introduction to differential privacy, randomized response, basic properties. *The 7th BIU Winter School on Cryptography, BIU.*

[8] C. Gentry. (2009). *A fully homomorphic encryption scheme.* PhD thesis, Stanford University, California

[9] P Martins, L Sousa & A Mariano. (2018). A survey on fully homomorphic encryption: An engineering perspective. *ACM Computing Surveys (CSUR), 50(6), 83.*

[10] Y. Lindell & B. Pinkas. (2008). Secure multiparty computation for privacy-preserving data mining. *IACR Cryptology ePrint Archive 197.*

[11] H. Bae, J. Jang, D. Jung, H. Jang, H. Ha & S. Yoon. (2018). Security and Privacy Issues in Deep Learning. *ACM Computing Surveys*

[12] S. Chang & C. Li. (2018). Privacy in Neural Network Learning: Threats and Countermeasures. *IEEE Network, 32(4), 61-67.*

[13] R. Shokri, M. Stronati, C. Song & V. Shmatikov. (2017). Membership Inference Attacks against Machine Learning Models. *IEEE Sym. SP, p. 3-18.*

[14] F. Tramèr, F. Zhang, A. Juels, M.K. Reiter & T. Ristenpart. (2016). Stealing Machine Learning Models via Prediction APIs. *USENIX Sec. Sym.* (pp. 601-618). Vancouver : USENIX

[15] P. Mohassel & Y. Zhang. (2017). SecureML: A System for Scalable Privacy preserving Machine Learning. *IEEE Sym. SP,* p. 19-38.

[16] L. Xie, K. Lin, S. Wang, F. Wang & J. Zhou. (2018). Differentially Private Generative Adversarial Network. *arXiv preprint arXiv:1802.06739.*

[17] J. Yuan & S. Yu. (2014). Privacy Preserving Back-Propagation Neural Network Learning Made Practical with Cloud Computing. *IEEE Trans. PDS, p. 212-221.*

[18] P. Li et al. (2017). Multi-Key Privacy-Preserving Deep Learning in Cloud Computing. *Future Generation Computer Systems, 74,* 76-85.

[19] M. Abadi et al. (2016). Deep Learning with Differential Privacy. *Proc. ACM CCS,* (pp. 308-318). ACM : Vienna

[20] G. Acs, L. Melis, C. Castelluccia & E. De Cristofaro. (2017). Differentially private mixture of generative neural networks. *IEEE Transactions on Knowledge and Data Engineering, 31(6),* 1109-1121.

[21] C. Dwork & G. N. Rothblum. (2016). Concentrated differential privacy. *CoRR, abs/1603.01887.*

[22] L. Yu, L. Liu, C. Pu, M. E. Gursoy & S. Truex. (2019). Differentially Private Model Publishing for Deep Learning. *IEEE.*

[23] X. Zhang, S. Ji, H. Wang & T. Wang (2017). Private, Yet Practical, Multiparty Deep Learning. *ICDCS, pp. 1442-52.* IEEE.

[24] K. Bonawitz et al. (2017). Practical Secure Aggregation for Privacy-Preserving Machine Learning. *Cryptology ePrint Archive,* (pp. 1175-1191). ACM.

홍 은 주(Eun-Ju Hong) [정회원]

· 2013년 2월 : 공주대학교 응용수학과 (이학사)
· 2015년 2월 : 공주대학교 융합과학과 (이학석사)
· 2015년 2월 ~ 현재 : 공주대학교 융합과학과 박사과정
· 관심분야 : 프라이버시 보호기술, 빅데이터 보안 등
· E-Mail : baby0708@kongju.ac.kr

이 수 진(Su-Jin Lee) [정회원]

· 2019년 2월 : 공주대학교 응용수학과 (이학사)
· 2019년 3월 ~ 현재 : 공주대학교 수학과 석사과정
· 관심분야 : 암호구현, 데이터보안 등
· E-Mail : sujiny2222@smail.kongju.ac.kr

홍 도 원(Do-Won Hong) [정회원]

· 1994년 2월 : 고려대학교 수학과(이학사)
· 2000년 2월 : 고려대학교 수학과(이학박사)
· 2000년 4월 ~ 2012년 2월 : 한국전자통신연구원 팀장, 책임연구원
· 2012년 2월 ~ 현재 : 공주대학교 응용수학과 교수
· 관심분야 : 암호기술, 프라이버시 보호기술 등
· E-Mail : dwhong@kongju.ac.kr

서 창 호(Chang-Ho Seo) [정회원]

· 1990년 2월 : 고려대학교 수학과(이학사)
· 1992년 2월 : 고려대학교 수학과(이학석사)
· 1996년 2월 : 고려대학교 수학과(이학박사)
· 2000년 3월 ~ 현재 : 공주대학교 응용수학과 교수
· 관심분야 : 암호알고리즘, PKI, 무선인터넷 보안 등
· E-Mail : chseo@kongju.ac.kr