Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.31, NO.5, Oct. 2021 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2021.31.5.987

그래디언트 기반 재복원공격을 활용한 배치상황에서의 연합학습 프라이버시 침해연구*

장 진 혁,[1†] 류 권 상,[1] 최 대 선[2‡]
[1,2]숭실대학교 (대학원생, 교수)

Federated Learning Privacy Invasion Study in Batch Situation Using Gradient-Based Restoration Attack*

Jinhyeok Jang,[1†] Gwonsang Ryu,[1] Daeseon Choi[2‡]
[1,2]Soongsil University (Graduate student, Professor)

요 약

최근 데이터로 인한 개인정보 침해로 인해 연합학습이 이슈화되고 있다. 연합학습은 학습데이터를 요구하지 않기 때문에 프라이버시 침해로부터 안전하다. 이로 인해 분산된 디바이스, 데이터를 활용하여 효율을 내기 위한 응용 방법에 대한 연구들이 진행되고 있다. 그러나 연합학습과정에서 전송되는 그래디언트로부터 학습데이터를 복원하는 재복원공격에 대한 연구가 진행됨에 따라 더는 연합학습도 안전하다고 볼 수 없다. 본 논문은 다양한 데이터 상황에서 데이터 복원 공격이 얼마나 잘되는지 수치적, 시각적으로 확인하는 것이다. 데이터가 1개만 존재할 때부터 크게는 클래스 안에 데이터가 여러 개 분포해 있을 때로 나누어 재복원공격이 얼마나 되는지 확인을 위해 MSE, LOSS, PSNR, SSIM인 평가지표로 MNIST 데이터를 활용해 수치로 확인한다. 알게 된 사실로 클래스와 데이터가 많아질수록 MSE, LOSS,이 높아지고 PSNR, SSIM이 낮아져 복원성능이 떨어지지만 몇 개의 복원된 이미지로 충분히 프라이버시 침해가 가능하다는 것을 확인할 수 있다.

ABSTRACT

Recently, Federated learning has become an issue due to privacy invasion caused by data. Federated learning is safe from privacy violations because it does not need to be collected into a server and does not require learning data. As a result, studies on application methods for utilizing distributed devices and data are underway. However, Federated learning is no longer safe as research on the reconstruction attack to restore learning data from gradients transmitted in the Federated learning process progresses. This paper is to verify numerically and visually how well data reconstruction attacks work in various data situations. Considering that the attacker does not know how the data is constructed, divide the data with the class from when only one data exists to when multiple data are distributed within the class, and use MNIST data as an evaluation index that is MSE, LOSS, PSNR, and SSIM. The fact is that the more classes and data, the higher MSE, LOSS, and PSNR and SSIM are, the lower the reconstruction performance, but sufficient privacy invasion is possible with several reconstructed images.

Keywords: Data Privacy, Data Reconstruction, Federated learning

Received(08. 13. 2021), Modified(09. 03. 2021), Accepted(09. 03. 2021)

* 본 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발)

† 주저자, slsk100@soongsil.ac.kr
‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

I. 서 론

AI의 발전은 사람들에게 일상생활의 편의성과 이로움을 제공해주고 있다. 구체적으로 얼굴인증과 지문인증을 통한 결제시스템, 결제 패턴과 인터넷 서핑 등을 통한 고객 맞춤형 서비스, 에어러블 기기로 헬스 케어, 5G 가상현실과 증강현실, 향후 자율 주행 자동차까지 AI의 발전이 무시 못 할 정도로 빠르게 나아가고 있다. 이러한 AI는 주재료인 데이터를 바탕으로 분석과 자동화로부터 앞서 언급한 서비스를 이루어 냈다. 한 사람으로부터 만들어 낼 수 있는 데이터는 무궁무진하며 앞으로 다양한 데이터를 바탕으로 꿈속에서나 펼쳐질 만한 세상이 현실화할 수 있다.

하지만 데이터가 주는 이로움에 반해, 과거 넷플릭스 개인정보 노출 건 등 민감한 개인정보와 프라이버시 노출 가능성의 증가로 인한 피싱과 해킹 사고가 증가하고 있다. 이렇게 데이터로 인한 개인정보의 피해가 크다 보니 데이터를 보호할 방법에 대한 연구가 증가하고 있으며 우리나라는 데이터 3법 이후 데이터 익명화와 식별방지를 위해 데이터 마스킹 데이터 범주화, 프라이버시 모델인 K-익명성, L-다양성 등을 통해 각 기관이나 기업에서도 고객들의 정보를 보호하는 방법을 선택했으며 현재 금융업계에서 마이데이터 사업으로 이목을 집중 받고 있다[1, 2, 3, 4, 5]. 마이데이터는 분산된 데이터들을 하나로 통합하여 데이터를 관리한다. 그리고 관리하는 주체가 기업, 기관이 아닌 개인이 관리 할 수 있도록 하는 사업이다. 이와 비슷하게 데이터를 서버로 모으지 않고 학습하는 방법으로 구글이 2017년에 연합학습을 발표했다[6, 7, 8, 9, 10]. 구글의 목적에 따르면 연합학습의 경우 데이터를 요구하지 않고 서버의 모델을 각 디바이스에 전송하면 개인에 대한 데이터가 서버로부터 받은 모델로 학습을 진행하고 각 디바이스의 그래디언트 만을 요구하여 데이터를 직접적으로 요구하지 않기 때문에 프라이버시 침해로부터 예방할 수 있어 이전에 기계학습으로 불편했던 데이터 뭉개짐 현상을 해소하고 민감한 데이터를 사용할 수 있다.

그러나 안전할 것만 같은 연합학습이 여러 공격으로 인해 성능이 저하되거나 프라이버시가 침해된다는 연구가 발표되고 있다[11, 12, 13, 14, 15]. 연합학습은 그래디언트로부터 서버의 모델을 학습하는 원리이지만 그래디언트로 학습데이터를 알아낼 수 있는 연구로 인해 연합학습의 프라이버시 침해 문제가 발생 되었다고 볼 수 있다[14, 15]. 현재 그래디언트로 재복원하는 연구는 데이터를 재복원하기 위해 랜덤데이터(Random Data)를 사용하며 원본데이터와 랜덤데이터를 1:1의 비율로 원본데이터 1개를 복원하기 위해 필요한 랜덤데이터가 1개가 필요하며 데이터 사이즈를 점차 늘리고 모델의 크기 및 파라미터를 다양하게 설정하고 랜덤데이터를 다르게 줘가며 복원성능을 향상 시킬 수 있도록 연구가 진행되고 있다[16]. 그러나 이전 연구들은 데이터의 복원성능을 높이는 방향에 집중적이며 배치상황에 대한 내용이 일반적으로 클래스가 다를 때를 주로 다루었으며 연합학습의 특성상 하나의 디바이스에 데이터가 어떻게 들어가 있을지 모르는 상황이 많기 때문에 모든 상황을 다루는 내용이 필요하다.

본 논문은 연합학습환경과 유사한 데이터 환경인 배치상황에 대하여 그래디언트로 재복원을 적용하여 단계적으로 데이터와 클래스를 높였을 때 복원이 어느 정도 되는지에 대해 집중적으로 다룰 것이다. 이를 통해 프라이버시 침해가 충분히 되는 것을 보일 것이다. 첫 번째로 1개만 복원했을 때와 데이터가 복잡했을 때의 복원결과를 MSE, LOSS, SSIM, PSNR 평가지표를 통해 수치로 확인하고 두 번째로 그래프와 복원된 결과를 통해 시각적인 비교를 제공하고 마지막으로 공격자가 할 수 있는 필터링 기법을 사용하여 복원력이 떨어진 이미지에 대해서 프라이버시가 충분히 되는 것을 보일 것이다. 본 논문의 기여한 바는 다음과 같다.

- 데이터 상황을 클래스와 데이터로 나누어 재복원 결과를 시각적으로 표현
- 재복원 성능을 다양한 평가지표를 활용하여 수치로 표현
- PIL 필터링 기법을 활용하여 복원이 떨어지는 부분 보완

본 논문의 구성은 다음과 같다. 2장에서 연합학습과 학습 과정, 안정성에 대해 다룰 것이고 3장에서 그래디언트로 재복원하는 방법과 데이터 상황을 소개할 것이다. 4장에서 재복원공격을 토대로 데이터와 공격자의 상황별을 배치상황에서 데이터 재복원이 얼마나 되는지에 대한 비교실험을 진행할 것이며 5장에서 PIL필터링으로 이미지를 뚜렷하고 보이는 효과에 대한 소개 6장의 토론으로 이루어지고 마지막 7장에서 실험에 대한 결론으로 논문을 맺는다.

II. 배경 및 관련 연구

2.1 연합학습

연합학습은 분산된 기업 및 기관이 보유한 데이터 또는 각각의 디바이스에 내장된 데이터로부터 집적시키지 않고 각 분산된 객체로부터 모델학습 된 결과만을 통해 서버의 모델을 학습하는 방법으로 Federated learning이라고 불린다. 구글이 개발한 연합학습은 하나의 서버로부터 훈련에 참여하는 디바이스에 응답을 청한 후 모델을 전송하는 형식이며 대표적으로 Google- keybord가 있으며 각 기관끼리 가지고 있는 데이터를 연합하여 더 좋은 모델을 만들기 위한 연구가 진행되고 있다[17, 18].

Fig 1은 연합학습을 보여주며 그림과 같이 연합학습의 원리는 서버로부터 훈련에 참여할 객체를 확인하여 해당 객체가 훈련에 참여할 수 있는데 체크한다. 체크 후 이상이 없으면 서버의 모델을 각 클라이언트(Client)에 전달한다. 각각의 클라이언트들은 자신이 소유한 데이터로부터 학습을 진행 후 학습된 값을 서버로 전송시킨다. 업데이트 값을 전송 시1 Round, 1 Communicate Cost라고 부르며 이러한 과정을 여러 번 반복하여 서버의 모델 성능을 올린다. 로컬학습(Local Training)은 각 클라이언트가 모델을 학습할 때, 글로벌 학습(Global Training)은 클라이언트로부터 받은 업데이트 값으로 서버의 모델을 학습시킬 때를 말한다[19].

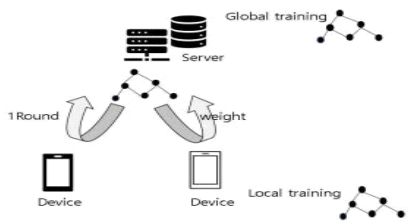

Fig. 1. Federated learning

2.1 연합학습 관련 연구

연합학습은 로컬학습을 진행하면서 업데이트된 값들을 서버로 전달할 때 업데이트 값들을 평균한 값을 전달한다. 이러한 업데이트 방식을 FedSGD (Federated Stochastic Gradient Decesent)라고 부르며 좀 더 빠르고 성능을 높이는 방법으로 각 클라이언트에 미니배치학습을 적용시킨 FedAVG(Federated Averaging)가 연구되었다[20, 21, 22]. 클라이언트 4개를 기준으로 Fed SGD는 식(1)(2), FedAVG는 식(3)(4) 와 같이 계산한다.

client: 4, w: 서버의 가중치, 업데이트 값, g: 클라이언트의 가중치, 업데이트 값

$$w_{\neq w} = w_{Old} - Learning rate * g \quad (1)$$

$$g = (g_1 + g_2 + g_3 + g_4)/4 \quad (2)$$

$$w_{\neq w} = w_{old} \quad (3)$$

$$w_{\neq w} = (w_1 + w_2 + w_3 + w_4)/4 \quad (4)$$

2.2 연합학습의 안전성

연합학습이 실제로 안전하지 못하다는 많은 연구가 발표되었다. 크게 모델의 성능을 저하하는 공격과 프라이버시 침해 공격으로 나눌 수 있으며 모델의 성능을 저하 시키는 데이터 포이즈닝 공격(Data Poisoning Attack), 클라이언트를 조절해 파라미터값을 편향시켜 잘못된 결괏값을 만드는 시빌 공격(Sybil Attack), 미세한 노이즈를 추가하여 데이터에 혼란을 주는 회피공격(Evasion Attack)이 있으며 프라이버시 침해를 보여주는 재복원공격은 2018년에 발표된 DLG(Deep Leakage from Gradient) 이다[12, 13, 14, 15, 16]. 재복원공격은 데이터가 노출되어 프라이버시를 침해 하는 것뿐만 아니라 노출된 데이터로 공격자가 학습에 참여시키거나 빼버릴 수 있기 때문에 포이즈닝 공격과 시빌 공격 등을 도와주는 역할을 한다. 이로부터 예방하기 위해 안전성에 관한 연구가 많이 필요하다.

III. 상황별 데이터 재복원공격

본 논문에서의 재복원공격(Reconstruction Attack)은 기계학습을 마치고 얻어진 그래디언트로부터 랜덤데이터를 원본으로 바꾸는 방법으로 2018년에 발표되었다[14].

그래디언트만을 가지고 서버의 성능을 높이는 연합학습의 특성상 매우 최적화된 공격방법이며 주로 이미지 데이터를 바탕으로 재복원하기 때문에 연합학습에 참가한 디바이스에 나의 얼굴 이미지가 들어가 있다면 재복원공격을 통해 나의 얼굴이 노출될 것이고 프라이버시 침해는 매우 크다.

현재 발표된 재복원공격은 초기 랜덤데이터를 무작위가 아닌 Pattern을 주거나 Dark/Light, RGB로 데이터를 변형하여 복원이 얼마나 빠르고 잘되는지에 대한 연구와 Loss와 모델의 layer를 변형시키는 연구로 복원성능에 대해 주요하게 다루어 공격자 상황에 초점이 맞춰졌다면 본 논문은 다른 공격방법으로 성능을 다루는게 아닌 공격자가 연합학습에 참여한 디바이스에 저장된 데이터를 추출하는 가정으로 디바이스 안에 있는 데이터가 복잡한 상황인 클래스와 데이터가 섞여 있는 배치상황에서도 복원이 얼마나 되는지에 대해 다룰 것이다[15, 16].

3.1 재복원공격

재복원공격은 원본으로부터 모델을 학습하고 난 그래디언트와 랜덤한 데이터를 생성해 원본과 동일한 모델을 학습하여 나온 Loss 값을 그래디언트로 바꾼 후 각각의 차이를 최적화를 통해 점차 줄여나가는 방식으로 진행한다. 즉 랜덤데이터, 랜덤 레이블(Random Label)들이 원본과 맞춰지면서 랜덤데이터가 원본데이터로 만들어진다. Fig 2는 랜덤데이터를 원본데이터로 재복원하는 과정이다[2].

G_0의 경우는 0의 이미지 데이터와 레이블을 모델 학습 진행 후의 그래디언트이며 G_R의 경우 랜덤데이터와 랜덤 그래디언트를 모델학습 후의 결괏값이다. 그림2와 같이 여러 번의 iteration을 반복하면서 랜덤데이터가 원본 이미지 0으로 바뀌게 된다. Fig 3은 iteration 100으로 설정하여 숫자 0의 이

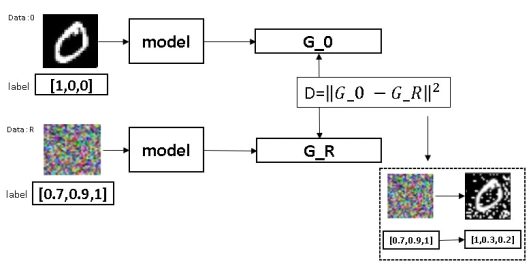

Fig. 2. Data Reconstruction(G_x: gradient of x)

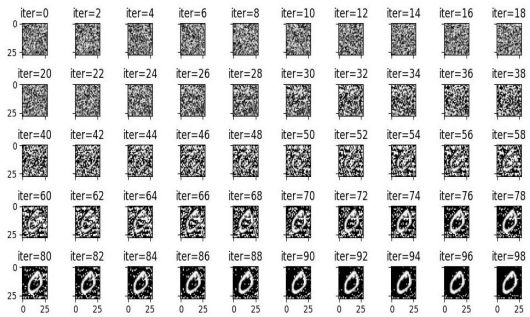

Fig. 3. Change of random data by iteration

미지로 바뀌는 과정을 보여준다. 복원하고자 하는 이미지가 여러 개일 경우 학습된 이미지의 평균 그래디언트를 이용하며 재복원 시 랜덤데이터가 그래디언트를 맞추기 위해 각각의 이미지의 데이터와 데이터를 복원하려고 한다. 즉 0과 1에 대한 이미지를 2장 복원할 때 랜덤데이터는 0을 2개, 1을 2개 복원하지 않는다.

재복원하는 식은 5, 6, 7에서 볼 수 있으며 랜덤데이터를 x', 랜덤 레이블을 y' 일 때 그래디언트는 $\nabla W'$ 로 계산하여 원본과의 차이를 LBFGS로 최소화시킨다[2].

$$\nabla W' = \frac{\partial l(F(x', W), y')}{\partial W} \tag{5}$$

$$\nabla W = \frac{\partial l(F(x, W), y)}{\partial W} \tag{6}$$

$$x', y' = argmin_{x', y'} \| \nabla W' - \nabla W \|^2 \tag{7}$$

3.2 데이터 상황

현재의 AI는 데이터를 기반으로 분석과 자동화로부터 시작되었으며 기계학습을 진행하기에 앞서 목적에 따라 필요한 데이터를 수집한다. 식별이나 인증의 경우 한사람에 대한 정보가 여러 개 있어야 쉽게 구분이 가능하며 기계학습의 성능 역시 하나 혹은 두 가지의 정보만으로 여러 사람들을 구별해내기 어렵듯 여러 항목 (feature)들을 바탕으로 예측할수록 성능이 높다. 재복원공격 역시 공격자의 입장에서 디바이스 안에 복원해야 하는 개수와 생성 해야 하는 랜덤데이터의 수는 그림 4와 같이 여러 상황에 따라 생

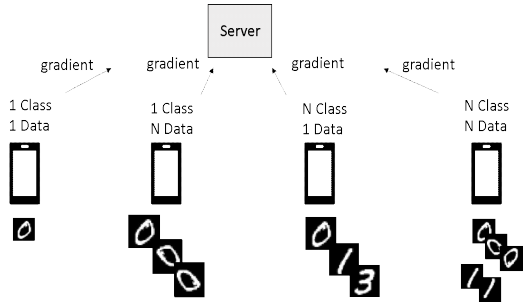

Fig. 4. Data situation example

성해야 한다.

table 1.은 재복원공격 논문에서 원본데이터와 랜덤데이터의 생성을 1:1 비율로 생성했을 때 각 데이터와 클래스의 개수에 따른 랜덤데이터 생성 개수이다. 클래스와 데이터의 상황에 따라 구분해 놨으며 1의 경우는 데이터와 클래스가 하나만 있는 경우 D는 클래스가 하나이면서 데이터가 여러 개일 때, C는 클래스가 여러 개며 데이터가 하나씩만 있을 때, C x D는 클래스 안에 데이터가 여러 개 있을 때를 말한다.

각각에 대한 여러 상황은 Fig 4와 MNIST 데이터셋을 예로들 경우 쉽게 이해할 수 있다. 클래스 0의 이미지 한 장만 가지고 있을 때를 1 상황, 0부터 9의 이미지를 한 장씩만 가지고 있는 경우를 C 상황, 0의 이미지를 같은 이미지나 다른 이미지를 여러 장 가지고 있는 경우를 D 상황, 0부터 9의 이미지를 각각 여러 장 가지고 있는 경우를 C x D 상황이라고 정했다. Fig 4와 같이 실제 환경에서 연합학습환경의 경우 하나의 디바이스에 가지고 있는 이미지가 제시된 C x D인 복잡한 상황과 제일 가깝다.

재복원하는 연구가 진행되고 있는 논문에서는 주로 하나의 클래스 하나의 데이터만을 복원하여 자신만의 공격방법을 적용하여 복원성능을 높이는 연구가 많다. 배치상황에 대해 언급은 하고 있으나 각각의 상황에 대해 다루지 않았으며 연구결과에 따르면 데이터가 많을수록 복원성능이 매우 떨어지는 것을 확인할 수 있다.

Table 1. Case by Case of Dataset

class \ data	1	D
1	1	D
C	C	C x D

IV. 데이터 상황별 재복원공격 분석 실험

4.1 실험 및 검증

현재의 인공지능 기술은 다양한 분야에서 분산화되고 있다. 이전처럼 하나의 기관과 기업에서만 관리하는 게 아닌 분산된 데이터로부터 머신러닝 기술을 접목한 연구가 많이 등장하고 있다. 이에 따라서 하나의 데이터만을 복원하는 게 아닌 다양한 데이터 환경을 생각하여 클래스와 데이터를 다양하게 섞어 재복원실험을 했다. 이전 논문들과 같이 랜덤데이터와 원본데이터를 1:1의 비율로 비교실험을 진행했으며 배치상황인 클래스와 데이터가 혼합하여 있을 때 복원이 어느 정도 되는지 확인할 것이다.

4.1.1 모델 및 데이터셋

실험에 사용한 데이터는 MNIST 데이터셋으로 0부터 9까지의 숫자를 손으로 쓴 글씨 데이터를 사용했고 각각 10개의 클래스로 분류되고 클래스마다 7,000개의 이미지를 가지고 있다. 복원하기 위한 모델은 [14] 와 동일한 모델을 사용했으며 3개의 컨볼루션 층과 한 개의 풀리커넥티드 층으로 구성되고 마지막 층은 0부터 9까지 10개를 분리하는 Multi Classification인 일반적인 CNN(Convolutional Neural Network)모델이다.

4.1.2 실험방법

본 실험의 목적은 배치상황에서의 재복원 결과를 확인하는 것이고 table 1에 따라 다양한 실험 상황을 구성하였다. 먼저 실험의 공통적인 요소는 데이터 1개당 랜덤데이터 1개를 생성하여 재복원한 실험이며, 랜덤데이터 생성의 경우 Time을 활용한 랜덤 seed를 사용하여 같은 랜덤데이터가 생성되는 것을 사전에 방지하였다. 해당 실험은 적절한 표본 값을 위해 조합별 30회를 추출해 결괏값의 평균을 적용한 것이며 아래 조건에 따라 임의로 뽑았다.

클래스가 1개이고 데이터가 2, 3, 5개일 때, 클래스가 2, 3, 5개이고 데이터가 한 개씩일 때, 클래스와 데이터가 각각 2, 3개씩 혼합되어 있을 때 복원 결과를 확인할 것이며 수치적인 평가는 4가지로 하였다.

- 1 상황(Class:1, Data:1) : 다양한 환경에서의 재복원공격을 비교하기 위한 베이스라인으로 0~9 클래스에 대해 재복원 실험을 진행하였다.
- C 상황(Class:C, Data:1) : C에 따라 랜덤으로 클래스당 이미지를 하나씩만 추출하여 재복원공격을 진행하였으며 본 실험에서는 C를 2, 3, 5일 때를 다루었다.
- D 상황(Class:1, Data:D) : 클래스가 동일하고 비슷한 이미지에서의 복원은 어떨지에 대한 실험이며 0~9까지의 클래스를 하나만 선택한 후 D에 따라 데이터를 랜덤으로 생성하여 재복원공격을 하는 것이며 본 실험에서 D는 2, 3, 5일 때를 다루었다.
- C x D 상황(Class:C, Data:D) : 더 복잡한 환경에서의 재복원공격 실험이며 C와 D의 조합을 {2, 2}, {2, 3}, {3, 2}와 같이 설정하여 실험을 진행하였다.

4.1.3 평가방법

이미지들이 얼마나 복원이 잘 되었는지를 확인하기 위해 총 네 가지를 평가했다. 다양하게 평가를 한 이유는 간혹 원본과 랜덤데이터의 그래디언트의 차이가 빠르게 수렴에 이르면 예측과 다르게 복원하고자 하는 데이터가 많더라도 높은 성능을 보일 수 있으며 반대로 수렴하지 않으면 적은 데이터로도 매우 낮은 성능을 보일 때가 있기 때문이다.

본 실험은 이미지가 얼마나 멀리 떨어져 있냐를 확인하는 방법인 평균제곱오차(MSE, Mean Squared Error)로 $\widehat{Y}_i$ 은 실 관측값 Y_i은 예측값이다. 영상, 동영상 등의 화질 손실 정보를 평가할 때 사용하여 값이 클수록 이미지의 차이가 작은 최대 신호 대 잡음 비(PSNR, Peak Signal-to-Noise Ratio)로 $\max^{2_I}$ 는 해당 영상의 최댓값이다. 수치적인 차이가 아닌 인간의 시각적 화질 차이 및 유사도를 평가하기 위해 고안된 방법으로 이미지 품질평가(SSIM, Structural Similarity) 을 사용했으며 μ_x, μ_y는 원본과 랜덤데이터의 평균값을 나타내고 σ_x, σ_y 는 표준편차, σ_{xy}는 공분산, c는 변수이며 SSIM 값이 클수록 품질이 좋다[23]. Loss값의 경우 평가지표와 다르게 이미지의 품질이나 차이를 나타내진 않는다. 원본데이터의 그래디언트와 랜덤데이터의 그래디언트의 차이를 보여주며 여러 실험에 걸쳐 경험에 의한 결과로 일반적인 기계학습과 같이 Loss값이 수렴해야 이미지가 복원될 가능성이 높으며 초기에 값을 빠르게 수렴하거나 작을수록 복원된 이미지가 선명해진다. Loss를 제외한 수식은 다음과 같다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2 \tag{8}$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{9}$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(2\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{10}$$

4.2 실험결과

4.1.2 방법으로 각 데이터 상황마다 공격을 진행하였으며 이전 연구에 따라 복원하고자 하는 데이터가 2개 이상인 경우 순서가 뒤바뀔 수 있는 경우도 존재하기 때문에 랜덤데이터와의 차이를 비교할 때 원본데이터 1개씩 비교하여 가장 적은 값으로 얻은 후 평균하였다. Table 2.의 경우 1 상황으로 데이터 1개를 복원하여 총 10개의 클래스에 대하여 30회씩 랜덤으로 뽑아 전체를 평균하여 마지막 Iteration 값을 가져온 것이다. Table 2.의 MSE, LOSS가 현저히 낮으며 SSIM, PSNR 높은 것을 확인할 수 있다. Fig. 5.의 Class:1, Data:1에서 복원된 이미지 역시 Iteration이 높아지면서 알아볼 수 있을 정도로 복원되었다. Fig. 5.는 Ground Truth가 복원해야 할 이미지이며 랜덤데이터가 Iteration마다 복원되고 있다.

Table 2. 1 Situation evaluation (last iteration)

	Class:1
Measure	Data:1
MSE	0.00002
LOSS	0.000008
SSIM	0.99
PSNR	105.19

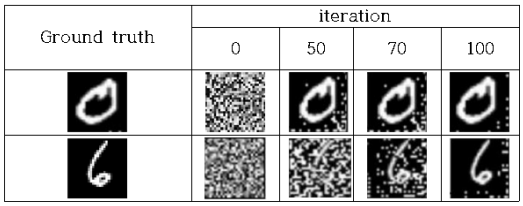

Fig. 5. 1 Situation Reconstruction image by iteration

Table 3. D Situation. evaluation (last iteration)

	Class:1			
Measure	Data:1	Data:2	Data:3	Data:5
MSE	0.00002	0.02716	0.07179	0.07958
LOSS	0.000008	0.001	0.0016	0.0015
SSIM	0.99	0.8	0.5	0.37
PSNR	105.19	73.22	61.29	59.61

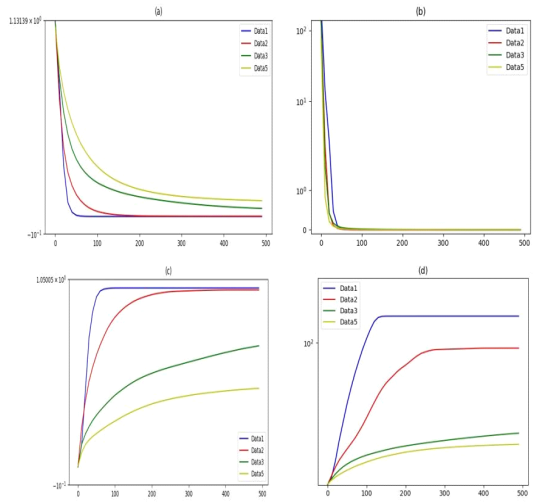

Fig. 6. D Situation, Level of reconstruction by iteration

Table 2.는 D 상황이며 클래스 1개 안에 데이터가 여러 개 존재하는 상황이다. 복원하고자 하는 데이터가 2개 이상이 되기 때문에 언급했던 대로 랜덤 데이터를 원본데이터마다 차이를 계산하여 작은 값으로 평균하였다. Table 3.의 결과처럼 데이터가 증가할수록 MSE, LOSS 값이 순차적으로 높아진 것을 확인할 수 있으며 SSIM, PSNR 값 역시 낮아진 것을 볼 수 있다. 표의 결과와 같이 Fig. 6.에서도

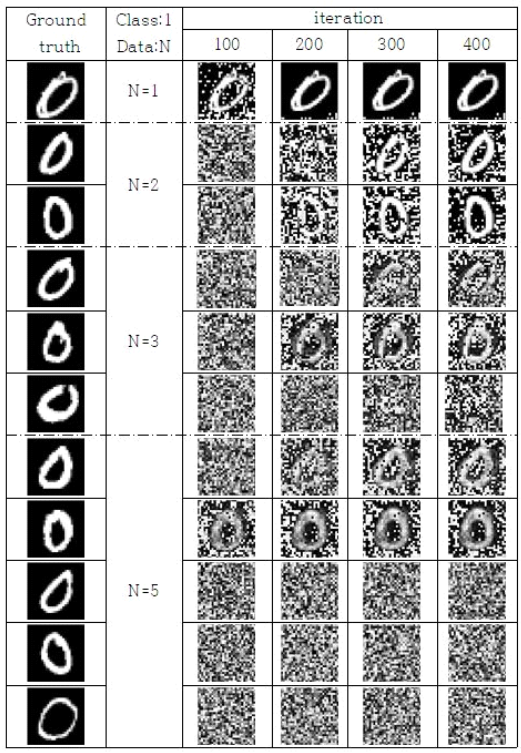

Fig. 7. D Situation. Reconstruction image by iteration

N이 증가함에 따라 복원력이 현저히 낮아지는 것을 확인할 수 있으며 N=3부터 복원된 이미지가 줄 어들며 알아보기 어렵다. Fig. 6. 는 500 Iteration 별로 각 Measure에 대한 수치로서 Iteration 별로 이미지의 차이가 어떤지 확인할 수 있다. x축은 iteration, y측은 각 Measure이며 (a) 는 MSE, (b) 는 LOSS, (c) 는 SSIM, (d) 는 PSNR이다. 파란색 꺾은선 그래프가 Data가 제일 적을 때, 초록색 꺾은선 그래프가 Data가 5개 존재할 때이다. 클래스는 동일한 상황에서 데이터가 감소할수록 (a), (b) 는 낮아졌고 (c), (d) 는 높은 것을 확인할 수 있다.

Table 4.는 C 상황이며 클래스가 여러 개이며 해당 클래스에 데이터가 한 개 존재하는 상황이다. 복원 하고자 하는 데이터가 역시 2개 이상이 되기 때문에 D 상황과 같은 방향으로 계산하였다. Table 6.와 Fig. 8.의 결과처럼 클래스가 증가할수록 MSE, LOSS 값이 높아지며 SSIM, PSNR 값이 낮아진 것을 확인할 수 있다. Fig. 8.의 경우 D 상황과 동일하게 파란색 꺾은 선이 클래스가 한 개일

Table 4. C Situation. evaluation (last iteration)

	Data:1			
Measure	Class:1	Class:2	Class:3	Class:5
MSE	0.00002	0.01	0.0013	0.0027
LOSS	0.000008	0.000002	0.0000448	0.00036
SSIM	0.99	0.94	0.88	0.67
PSNR	105.19	96.63	85.12	66.98

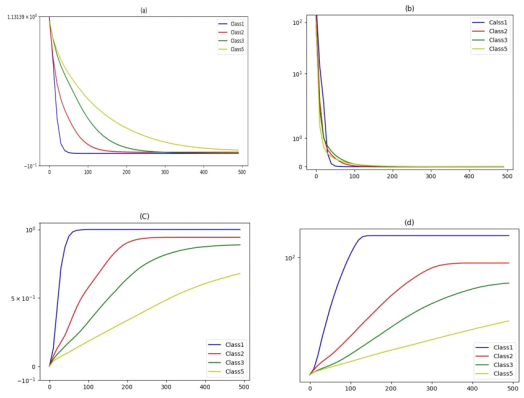

Fig. 8. C Situation, Level of reconstruction by iteration

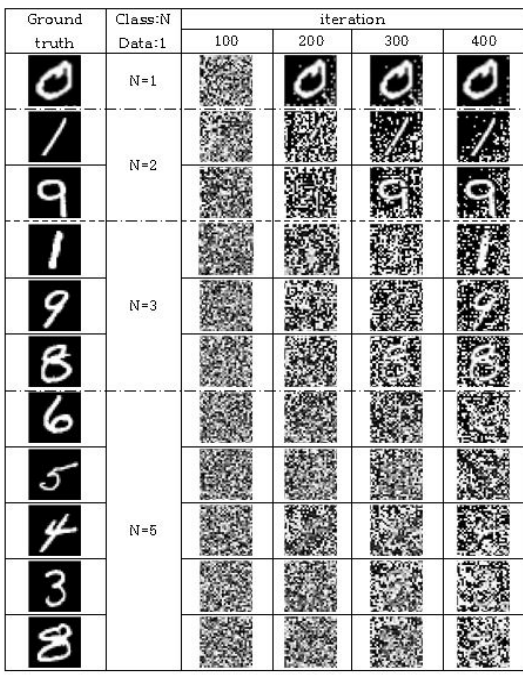

Fig. 9. C Situation. Reconstruction image by iteration

때, 노란 선이 클래스가 5개일 때를 나타낸다. 표와 그래프의 결과와 같이 Fig. 9. 에서도 클래스가 많아질수록 복원력이 현저히 낮아지는 것을 확인할 수 있다.

Table 5.은 C x D 상황이며 클래스와 데이터가 혼합된 상황이다. Table 5.과 Fig. 10.의 결과처럼 클래스가 증가할수록 MSE, LOSS 값이 높아지며 SSIM, PSNR 값이 낮아진 것을 확인할 수 있으며 지금까지의 실험과 비슷한 결과를 볼 수 있다. Fig. 10.은 이전 꺾은 선 그래프와 동일한 조건에서 파란 선이 Class 1개 - Data 1개일 때, 빨간 선이 Class 2개 - Data 2개, 초록 선이 Class 3개 - Data 2개, 노란 선이 Class 2개 - Data 3개를 나타낸다. 대부분 클래스, 데이터가 많을수록 복원성능이 낮지만 (a)~(d) 의 결과를 보면 데이터가 많은 것보다 클래스가 많을 때 복원성능이 좋은 것을 확인할 수 있었다. 복원 정도를 보는 Fig. 11.에서 보이는 것과 같이 대부분이 복원할 수 없었고 C 상황, D 상황만큼 훨씬 복원이 안 될 거라 예상했지만 의외로 {Class:3, Data:2}에서와 같이 몇몇 이미지들이 미세하게 구분이 가능할 정도로 복원이 되었다.

Table 5. C x D Situation. evaluation (last iteration)

	Class1	Class2		Class3
Measure	Data1	Data2	Data3	Data2
MSE	0.000023	0.085	0.07	0.1
LOSS	0.000008	0.00263	0.00098	0.002
SSIM	0.99	0.46	0.48	0.41
PSNR	105.19	60.96	60.48	58.71

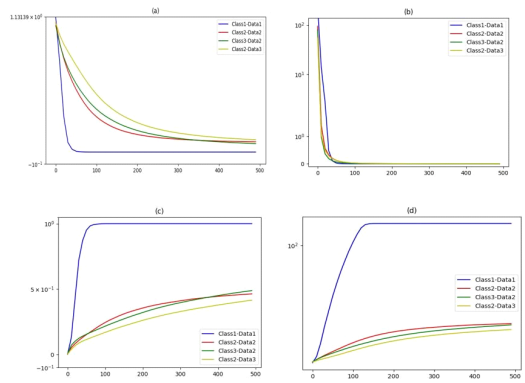

Fig. 10. C x D Situation, Level of reconstruction by iteration

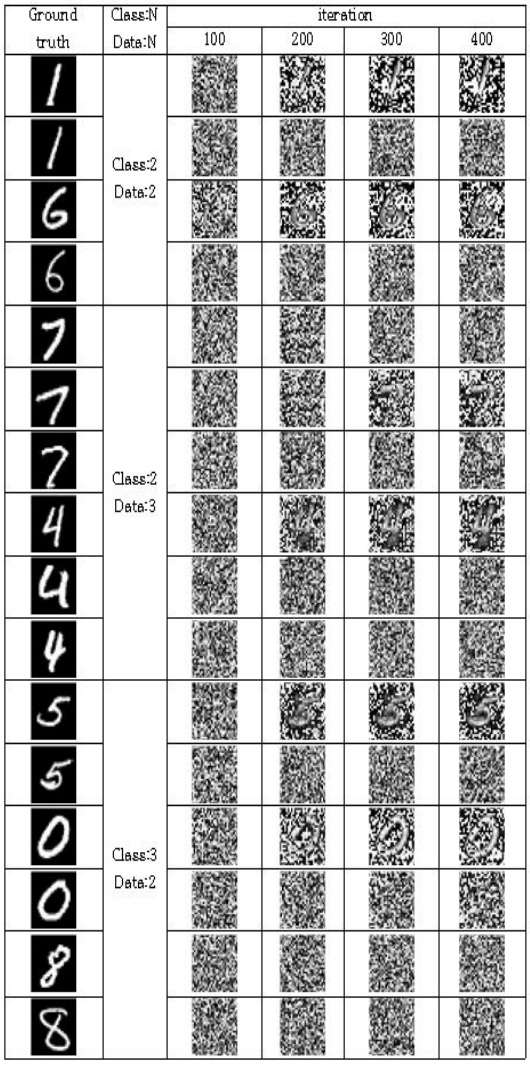

Fig. 11. C x D Situation. Reconstruction image by iteration

V. PIL 필터링 기법

이미지 분석처리 라이브러리인 PIL(python imaging library)의 pillow 모듈 중 이미지 필터링(PIL Filtering, ImageOps)으로 복원된 이미지를 뚜렷하게 보이고자 한다[24]. Pillow은 해당 이미지를 작게 자르거나 늘리고 상하 반전과 좌우 반전, 저장방법 필터링 기법이 있으며 이중 필터링 기법은 픽셀 단위로 조작하여 원하는 만큼 선명하게 하거나 밝기를 조절하고, 흑백 이미지를 컬러로 바꿔주고 명암 보정과 색 보정, 흐림, 엠보싱으로 안 보이게 만들 수 있다. 필터링으로 Blur, Detail, Edge_Enhance, Edge_Enhance_More, Emboss, Sharpen, Smooth, Smooth_More가 있으며 그 외에는 ImageOps를 이용하였다. 방법은 다음과 같다.

먼저 필터링 모듈로 블러(Blur)는 이미지를 흐리게 만드는 필터링 방법으로 Blur, Box Blur, Gaussian Blur 등이 존재하며 Box Blur의 경우는 특정 범위 내에서 흐릿한 값의 정도를 지정할 수 있으며 Gaussian Blur의 경우 가우스곡선을 이용하여 흐릿함을 적용한다. 컨투어(Contour)의 경우 윤곽을 잡아주고 디테일(Detail)은 해당이미지를 뚜렷하게 만드는 효과이며 Edge_Enhance, Edge_Enhance_More 역시 같은 역할을 해준다. 스무스(Smooth)는 부드럽게 샤펜(Sharpen)은 날카롭게 만들어주고 엠보싱(Emboss)은 이미지의 특정부분이 도드라지게 나타내는 표현하며 그레이스케일(GrayScale)은 각 픽셀은 명도를 나타내는 숫자로 표현된다. 0은 검은색을 나타내고 숫자가 커질수록 명도가 증가하여 하얀색이 된다.

ImageOps중 인버트(Invert)는 색을 반전시키며 브라이트니스(Brightness)는 이미지의 밝기를 조절할 수 있고 콘트레스트(Contrast)는 밝은 부분과 어두운 부분의 차이를 뜻한다. 포스터리즈(Posterize)는 이미지의 채널당 사용되는 음영의 개수를 제한하며 음영이 줄어드는 만큼 이미지의 색상을 단순화시킨다.

데이터 상황이 복잡해질수록 복원력이 매우 낮아진다. 복원이 안 된 경우가 많았으며 몇몇 복원된 이미지들은 구별하기 어려운 경우가 많았다. 공격자는 애매하게 복원된 이미지로부터 위의 방법과 같이 필

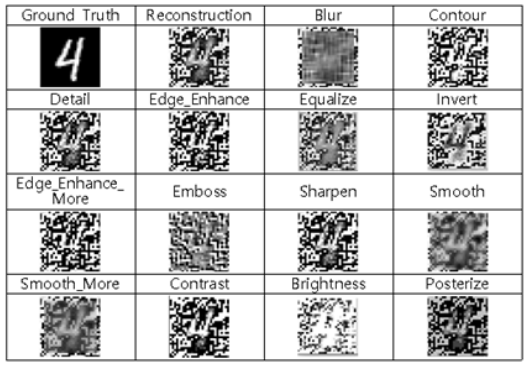

Fig. 12. PIL Filtering GrayScale

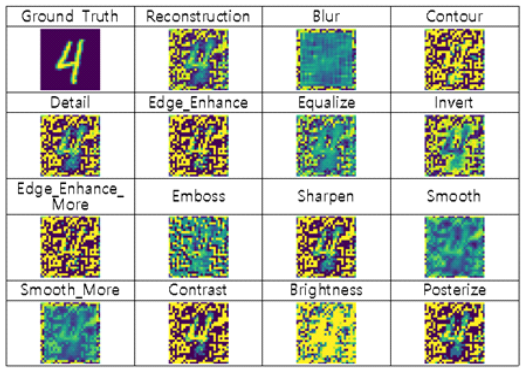

Fig. 13. PIL Filtering RGB

터링 기법을 사용하여 Fig. 12, 13과 같이 복원된 이미지를 필터링방법을 적용하여 무엇인지 확인할 수 있다. 해당 그림의 Ground Truth는 복원해야 할 이미지이며 Reconstruction은 CxD상황중 2X3의 복원된 이미지를 사용하여 여러 필터링 방법을 적용한 것이다. 이미지 필터링을 적용한 블러와 엠보싱의 경우 이미지를 흐리게 만드는 방법으로 Fig. 12., 13. 에서 복원된 이미지가 다른 방법에 비해 확인이 어렵다는 것을 확인할 수 있다.

VI. Discussion

본 논문의 실험은 데이터 1개 당 랜덤데이터 1개를 기반으로 복원된 이미지다. 따라서 복원하고자 하는 이미지가 2개일 경우 랜덤데이터의 개수도 2개를 사용하였다. 하지만 실제 환경에서의 공격자는 하나 또는 여러 대의 디바이스 안에 있는 데이터의 개수를 알기 어렵기 때문에 랜덤데이터의 개수를 적거나 많게 할 수밖에 없다.

실제로 데이터 3개를 복원할 때 랜덤데이터 1개를 사용하여 복원하면 Fig. 13.과 같이 나온다. Fig. 14.는 MNIST 데이터의 서로 다른 0, 1, 3 클래스를 많은 시도에 걸쳐 복원했으며 평균 이미지와 비슷하게 나오며 MSE 측정 결과 0.0303으로 적은 수치를 보이게 된다.

Fig. 15.는 CIFAR-100의 하나의 이미지를 좌우 반전시켰으며 같은 클래스에 2개의 이미지를 복원할 때 합성된 이미지가 나온다. 이러한 결과로 보았을 때 학습에 사용된 데이터의 그래디언트는 평균 그래디언트이며 연합학습에서 클라이언트들이 각자의 학습을 마치고 난 그래디언트를 평균하여 글로벌 모

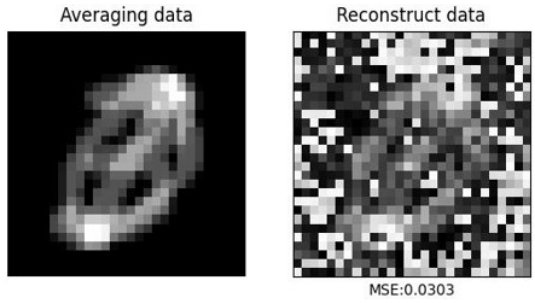

Fig. 14. 3 Images, 1 Random data reconstruction Result

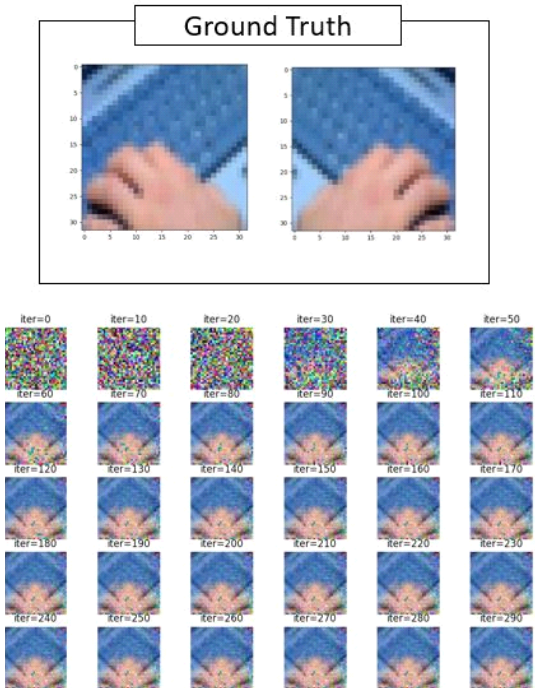

Fig. 15. 2 Image 1 Random data Reconstruction Iteration Result

델을 향상 시킬 때 사용하는 그래디언트와 같다고 볼 수 있으며 서버도 데이터를 복원할 수 있다.

VII. 결 론

본 논문은 안전할 것만 같은 연합학습이 프라이버시 침해가 된다는 것을 보이기 위해 그래디언트를 이용하여 재복원공격을 하였다. 이를 바탕으로 좀 더 현실과 가까운 환경인 배치상황인 클래스와 데이터로 세분화하여 접근하였고 MNIST 데이터 셋을 이용하

여 다양한 평가와 PIL 필터링 방법을 통해 수치와 시각적으로 확인하였다. 연합학습은 여러 디바이스로부터 오는 그래디언트의 평균을 이용하여 글로벌 모델의 성능을 올리는 게 목적이다. 비록 여러 디바이스로부터 오는 그래디언트를 사용하지는 않았지만 배치상황을 이용하여 실제 연합학습과 비슷한 환경으로 실험을 진행하였다.

MNIST 데이터 셋을 사용한 실험결과 예상했던 것과 같이 클래스와 데이터가 혼합하여 들어가 있을수록 복원성능이 매우 떨어졌으며 몇몇 이미지를 제외하고는 복원이 안된 것을 확인할 수 있었다. 비록 복원력은 낮지만, 공격자가 시도할 수 있는 이미지 처리 방식을 도입해 이미지를 알아볼 수 있도록 하였고 이러한 작은 데이터 하나라도 노출이 된다면 이전 프라이버시 침해처럼 공격자가 다른 데이터와 결합한 공격을 사용한다면 연쇄적으로 데이터 노출이 되어 프라이버시 침해가 상당하다.

본 연구를 바탕으로 연합학습처럼 디바이스에 데이터가 어떻게 분포되어있는지 모르는 상황에 대해서 공격자가 데이터가 복잡한 상황이어도 복원할 수 있다는 것을 보였다. 이를 토대로 연합학습도 충분히 프라이버시를 침해 받을 수 있다. 연합학습의 안전을 위해 본 연구처럼 다양한 환경에서 재복원공격에 대한 연구가 활발히 진행되어야 하며 다양한 공격방법으로 그래디언트에 영향을 줄 수 있는 요소인 레이블과 데이터수, 렌덤 데이터 등을 활용한 재복원공격을 분석하여 대비하는 방법과 연구가 진행되고 있는 차분 프라이버시와 같이 학습할 데이터에 노이즈를 추가하여 렌덤 데이터가 재복원을 했을 때 식별하기 어렵게 만들거나, 학습 후 그래디언트에 노이즈를 추가하여 랜덤데이터가 재복원이 되는 것을 방해할 수 있는 연구가 필요하다. 향후 연구 방향으로는 연합학습의 프라이버시를 보완하기 위해 다양한 공격방법을 연구하고 있으며 label과 랜덤데이터의 영향력에 대한 공격을 통해 더 빠르고 많은 데이터를 재복원하여 이를 막을 수 있는 대비 방안을 연구할 예정이다.

References

[1] Latanya Sweeney, "k-anonymity: a model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557-570, Jul. 2002.

[2] Cynthia Dwork, Aaron Roth, "The Algorithmic Foundations of Differential Privacy", Foundations and Trends in Theoretical Computer Science, vol. 9, no. (3-4), pp. 211-407, Aug. 2014.

[3] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition", arXiv, Dec. 2014.

[4] Joungyoun Kim, Min-jeong Park, "Multiple imputation and synthetic data", The Korean Journal of Applied Statistics, vol. 32, no. 1, pp. 83-97, Feb. 2019.

[5] Jooseok Park, "A Comparative Study of Big Data, Open Data, and My Data", The Korea Journal of BigData, vol. 3, no. 1, pp. 41-46, Aug. 2018.

[6] Google AI Blog, Federated Learning: Collaborative Machine Learning without Centralized Training Data, Available:https://ai.googleblog.com/2017/04/federated-learning-collaborative.html, Accessed: Jul. 2019.〔Online〕

[7] Yue Zhao, Meng Li, Liangzhen Lai, and Naveen Suda, "Federated Learning with Non-IID Data", arXiv, Jun. 2018.

[8] Tian Li, Anit Kumar Sahu, and Ameet Talwalkar, "Federated Learning: Challenges, methods, and future directions", IEEE SIGNAL PROCESSING MAGAZINE, vol. 37, no 3, pp. 50-60, May. 2020.

[9] Keith Bonawitz, Hubert Eichner, and Wolfgang Grieskamp, "TOWARDS FEDERATED LEARNING AT SCALE: SYSTEM DESIGN", Proceedings of the 2nd SysML Conference, Mar. 2019.

[10] H. Brendan, McMahan Eider, Moore

Daniel Ramage et. al., “Communication-Efficient Learning of Deep Networks from Decentralized Data”, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 54, pp. 1273-1282, Feb. 2017.

〔11〕 Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu, ″Data Poisoning Attacks Against Federated Learning Systems″, European Symposium on Research in Computer Security, pp. 480-501, Sep, 2020.

〔12〕 Clement Fung, Chris J.M. Yoon, Ivan Beschastnikh, ″Mitigating Sybils in Federated Learning Poisoning″, arXiv, Jul. 2020.

〔13〕 Eugene Bagdasaryan, Andreas Veit, and Yiqing Hua, ″How To Backdoor Federated Learning″, Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS)2020, vol. 108, pp. 2938-2948, Aug. 2020.

〔14〕 Ligeng Zhu, Zhijian Liu, and Song Han, ″Deep Leakage from Gradients″, 33rd Conference on Neural Information Processing Systems NeurIPS, pp. 17-31, Dec. 2019.

〔15〕 Jonas Geiping, Hartmut Bauermeister, and Hannah Drög, ″Inverting Gradients - How easy is it to break privacy in federated learning?″, 34th Conference on Neural Information Processing Systems NeurIPS, Dec. 2020.

〔16〕 Wenqi Wei, Ling Liu, Margaret Loper, and Ka-Ho Chow, ″A Framework for Evaluating Clinet Privacy Leakages in Federated Learning″, 25th European Symposium on Research in Computer Security, pp. 545-566, Sep. 2020.

〔17〕 Andrew Hard, Kanishka Rao, and Rajiv Mathews, “FEDERATED LEARNING FOR MOBILE KEYBOARD PREDICTION”, arXiv, Feb. 2019.

〔18〕 Qiang Yang, Yang Liu, and Tianjian Chen,“Federated Machine Learning: Concept and Applications”, ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 12, pp. 1-19, Jan. 2019.

〔19〕 Tian Li, Anit Kumar Sahu, and Ameet Talwalkar,“ Federated Learning: Challenges, methods, and future directions”, IEEE SIGNAL PROCESSING MAGAZINE, vol. 37, no. 3, pp. 50-60, May. 2020.

〔20〕 Xin Yao, Tianchi Huang, and Chenglei Wu, “Federated Learning with Additional Mechanisms on Clients to Reduce Communication Costs”, arXiv, Sep. 2019.

〔21〕 H. Brendan McMahan, Eider Moore, and Daniel Ramage, “Communication-Efficient Learning of Deep Networks from Decentralized Data”, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, vol. 54, pp. 1273-1282, Feb. 2017.

〔22〕 Tian Li, Anit Kumar Sahu, and Manzil Zaheer, “Federated Optimization in Heterogeneous Networks”, Proceedings of the 3rd MLSys Conference, Apr. 2020.

〔23〕 Z. Wang, A.C. Bovik, and H.R. Sheikh, “Image quality assessment: from error visibility to structural similarity”, IEEE transactions on image processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

〔24〕 Python Pillow Library (Pillow – Pillow (PIL Fork)8.3.1 documentation), Available:https://pillow.readthedocs.io/en/stable/, Accessed: Aug. 2021. 〔Online〕

〈저자소개〉

장 진 혁 (Jinhyeok Jang) 학생회원
2020년 2월: 공주대학교 응용수학과 학사
2020년 2월~2020년 8월: 공주대학교 융합과학과 석사과정
2020년 8월~현재: 숭실대학교 융합소프트웨어학과 석사과정
〈관심분야〉 정보보호, 금융보안, 인증

류 권 상 (Gwonsang Ryu) 학생회원
2016년 2월: 공주대학교 응용수학과 학사
2018년 2월: 공주대학교 대학원 융합과학과 석사
2018년 3월~2020년 8월: 공주대학교 대학원 융합과학과 박사과정
2020년 9월~현재: 숭실대학교 대학원 융합소프트웨어학과 박사과정
〈관심분야〉 인증, 이상거래탐지, 인공지능 보안

최 대 선 (Daeseon Choi) 종신회원
1995년 2월: 동국대학교 컴퓨터공학과 학사
1997년 2월: 포항공과대학교 컴퓨터공학과 석사
2009년 1월: 한국과학기술원 전산학과 박사
1997년 1월~1999년 6월: 현대정보기술 선임
1999년 7월~2015년 8월: 한국전자통신연구원 인증기술연구실 실장/책임연구원
2015년 9월~2020년 8월: 공주대학교 의료정보학과 부교수
2020년 9월~현재: 숭실대학교 소프트웨어학부 부교수
2016년~현재: 정보보호학회 이사
〈관심분야〉 인증, 개인정보보호, 이상거래탐지, 의료정보보안, 머신러닝