Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.33, NO.1, Feb. 2023 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2023.33.1.63

픽셀값 변환 기법을 더한 데이터 복원 공격에 의한 연합학습의 프라이버시 침해*

오 윤 주,[1†] 최 대 선[2‡]
[1,2]숭실대학교 (대학원생, 교수)

Invasion of Pivacy of Federated Learning by Data Reconstruction Attack with Technique for Converting Pixel Values*

Yoon-ju Oh,[1†] Dae-seon Choi[2‡]
[1,2]Soongsil University (Graduate student, Professor)

요 약

프라이버시 침해에 대한 안전성을 보장하기 위해 매개변수를 주고받아 학습하는 연합학습이 대두되고 있다. 하지만 최근 그래디언트를 이용하여 학습 데이터를 유출하는 논문이 발표되었다. 본 논문은 연합학습 환경에서 그래디언트를 이용하여 학습 데이터를 유출하는 실험을 구현하였으며, 학습 데이터를 유출하는 기존 공격을 개선하여 복원 성능을 높이는 방법을 제안한다. 제안 방법에 대해 Yale face database B, MNIST dataset를 이용하여 실험한 결과, 연합학습 성능이 accuracy=99~100%로 높을 때 100개의 학습 데이터 중 최대 100개의 데이터를 식별 가능한 수준으로 복원하여, 연합학습이 프라이버시 침해로부터 안전하지 않다는 것을 보인다. 또한, 픽셀 단위의 성능(MSE, PSNR, SSIM)과 Human Test에 의한 식별적인 성능을 비교함으로써 픽셀에 기반한 성능보다 식별적인 성능의 중요성을 강조하고자 한다.

ABSTRACT

In order to ensure safety to invasion of privacy, Federated Learning(FL) that learns using parameters is emerging. However a paper that leaks training data using gradients was recently published. Our paper implements an experiment to leak training data using gradients in a federated learning environment, and proposes a method to improve reconstruction performance by improving existing attacks that leak training data. Experiments using Yale face database B, MNIST dataset on the proposed method show that federated learning is not safe from invasion of privacy by reconstructing up to 100 data out of 100 training data when performance of federated learning is high at accuracy=99~100%. In addition, by comparing the performance (MSE, PSNR, SSIM) of pixels and the performance of identification by Human Test, we want to emphasize the importance of the performance of identification rather than the performance of pixels.

Keywords: Federated Learning, Reconstruction Attack, Privacy, Identification

Received(09. 27. 2022), Modified(11. 29. 2022), Accepted(12. 12. 2022)

* 이 논문은 2022년도 한국정보보호학회 하계학술대회에 발표한 우수논문을 개선 및 확장한 것임

* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발)과 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1A2C1014813)

† 주저자, ohyoonju@soongsil.ac.kr

‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

I. 서 론

최근 빅데이터의 활용이 많아지면서 크고 분산된 데이터를 학습하는 기술이 발전되고 있으며, 학습에서 발생할 수 있는 개인정보 침해에 대한 우려도 제기되고 있다. 이에 따라 프라이버시를 보호하면서 분산된 데이터를 안전하게 학습할 수 있다고 소개된 연합학습[1]이 이슈화되고 있다. 연합학습은 중앙 서버가 있고 여러 개의 클라이언트가 있는 형태로, 중앙 서버가 각 클라이언트의 데이터를 학습할 때 데이터를 중앙 서버의 글로벌 모델로 가져오는 것이 아니라 글로벌 모델을 각 클라이언트의 로컬 모델에 공유한다. 또한, 클라이언트가 공유받은 모델로 각 데이터를 학습한 뒤 중앙 서버에 전달할 때도 직접적인 학습 결과가 아닌 매개변수 및 매개변수 변경사항만을 전달하여 중앙 서버의 모델을 학습시킨다. 따라서 모델을 공유함으로써 중앙 서버에서 데이터를 학습하는 효과를 낼 수 있으며, 매개변수를 전달함으로써 프라이버시를 보호할 수 있다는 것이 연합학습의 강점이다. 하지만 연합학습이 나온 후, 매개변수인 그래디언트를 이용하여 학습에 사용된 데이터를 복원하는 논문이 발표되었다[2]. 이 논문에서는 그래디언트로 데이터를 복원하는 DLG(Deep Leakage from Gradients)가 추후 연합학습에서 프라이버시 침해에 대한 챌린지가 될 수 있다고 설명하였다. 이후에도 DLG와 같이 그래디언트로 복원하는 다양한 연구[3-9]가 발표된 것으로 보아 연합학습에서 프라이버시 침해의 위험성이 강조된다.

본 논문은 그래디언트를 통해 학습에 사용된 데이터를 복원하는 공격을 연합학습이 진행되는 상황에서 시도해보고자 한다. 이때 복원 공격은 기본적인 DLG 방법에 이미지 정규화 및 변환 과정을 추가하여 기존보다 더 확장된 데이터에서 더 높은 복원 성능을 보이는 복원 방법을 제안하여 사용한다. 또한, 실험 결과를 통해 픽셀 단위의 성능과 식별적인 성능을 비교함으로써 픽셀에 기반한 성능보다 식별적인 성능의 중요성을 강조하고자 한다. 이에 따른 본 논문의 기여는 다음과 같다.

- 매개변수를 주고받는 연합학습 과정에서 복원 공격을 시행함으로써, 연합학습의 프라이버시 침해가 가능함을 보인다.
- 기존의 DLG에 이미지 정규화와 변환 방법을 추가하여 더욱 확장된 데이터에서 더욱 증가된 복원 성능을 보이는 복원 방법을 제안한다.
- 데이터, 레이블 종류, 공격 방법, 출력 방법에 따른 실험을 진행하여 복원 공격에 대한 다양한 비교가 가능하다.
- 픽셀 단위의 성능보다 식별적인 성능이 중요함을 실험 결과를 통해 설명한다.

본 논문의 구성은 다음과 같다. 2장에서는 연구 배경에 대한 내용으로, 연합학습과 복원 공격에 대한 개념 및 과정을 설명하고, 복원 공격 관련 연구에 대해 소개한다. 3장에서는 제안 방법을 소개하고 연합학습에서의 복원 공격에 대한 실험을 통해 제안 방법의 성능을 평가한다. 4장에서는 실험에 대한 결론과 향후 계획으로 논문을 맺는다.

II. 연구 배경

2.1 연합학습 (FL, Federated Learning)

연합학습은 매개변수를 이용하여 학습하므로 가중치나 그래디언트를 이용한다고 소개하는 글이 많다. 본 논문은 그래디언트로 복원하는 DLG를 사용하기 위해, 그래디언트를 이용하는 연합학습에서 실험을 진행하였다.

연합학습은 하나의 중앙 서버와 학습 데이터를 가지고 있는 n개의 클라이언트가 있는 구조로 이루어져 있다. 먼저, 클라이언트의 로컬 모델은 중앙 서버의 글로벌 모델을 공유받아 사용한다. 각 클라이언트는 공유받은 모델을 이용하여 데이터를 학습하고, 학습에서 나온 총 n개의 그래디언트를 집계(평균)하여 중앙 서버로 전달한다. 중앙 서버는 집계한 그래디언트로 가중치를 업데이트하여 글로벌 모델을 학습시킨다. 연합학습 과정을 Fig. 1.에 제시하였다.

2.2 복원 공격 (reconstruction attack)

본 논문은 L. Zhu[2]가 제시한 DLG 방법을 사용하여, 그래디언트로 데이터를 복원하였다. DLG (Deep Leakage from Gradients)는 원본 데이터(이미지, 레이블)를 복원하기 위해 의사 데이터(이미지, 레이블)를 생성한다. 이때, 의사 데이터는 원본 데이터와 크기가 같은 랜덤값을 사용한다. 같은 모델에 대해 원본 데이터를 학습하여 원본 그래디언트를 계산하고, 의사 데이터를 학습하여 의사 그래디언트를 계산한다. 그리고 l2 norm으로 정의한 기울

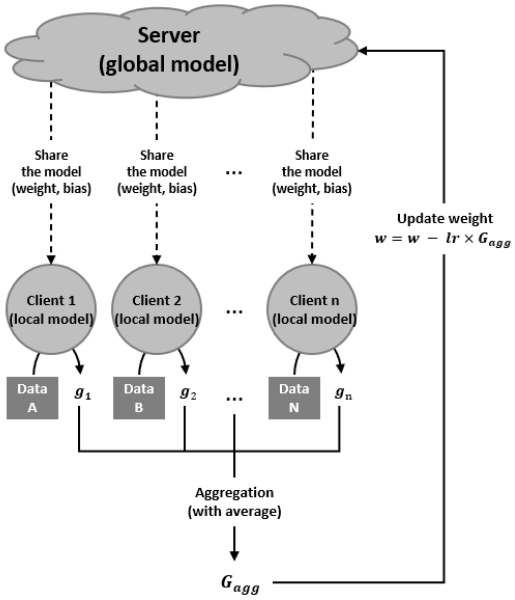


Fig. 1. Process of Federated Learning

기 손실 함수에 두 그래디언트를 대입하여, 두 그래디언트의 차이가 0에 가까워질 때까지 복원 iteration을 진행하면서 의사 데이터를 업데이트한다. 이 과정을 여러 번 진행하여 복원 iteration이 커지면 의사 데이터는 원본 데이터와 거의 같아진다[2]. 이와 같이, 학습 데이터에 접근하지 않고 유출된 기울기만으로 데이터를 복원하는 것을 '기울기 반전(gradient inversion)'이라고 하며, 기울기 반전을 이용한 복원 공격 과정을 Fig. 2.에 제시하였다.

기울기 반전의 기본 패러다임을 제시한 DLG[2]가 소개된 이후 다양한 기울기 반전 관련 연구[3-9]가 진행되었다. [3]은 기울기 반전 과정 중 의사 레이블을 생성하는 단계에서 랜덤값이 아닌 실제 레이블을 복원한 레이블을 사용한다. 이는 일반적인 모델들이 각 출력에 대한 손실 함수로 cross-entropy[10]를 사용하여 학습하며, cross-entropy로 인해 대상 레이블에 해당하는 출력 뉴런의 기울기 부호가 음수가 되는 것을 이용하여 실제 레이블을 복원한다. 실제 레이블을 복원한 레이블을 의사 레이블로 사용하여 복원을 진행하면, 기존의 방법보다 빠르게 복원되는 결과가 나타났다. [4]는 기울기 손실 함수로 코사인 유사도를 사용하는 것이 큰 특징이며, 기울기 손실 함수뿐만 아니라 이미지 정규화를 위해 TV norm[11]을 추가로 사용한다. 또한, [3]의 레이블 재복원을 언급하며 실제 레이블 정보가 분석적으로 복원될

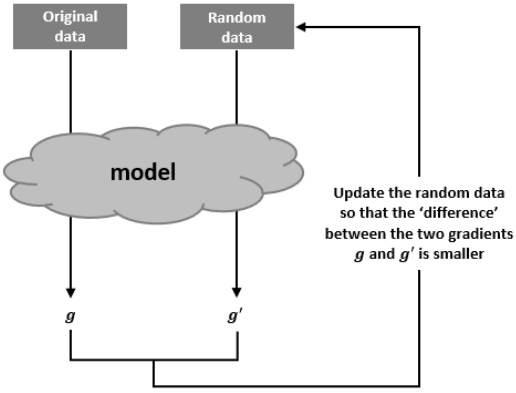


Fig. 2. Process of reconstruction attack using gradient (gradient inversion)

수 있으므로 실제 레이블을 의사 레이블로 사용하여 실험을 진행하였다. [5]는 [3]와 [4]를 개선하여 연구한 논문이다. 먼저, 그래디언트 행 합계의 최솟값에 해당하는 출력 인덱스를 복원 레이블로 선택하는 [3]방법과 달리, 모든 행에 대한 최솟값에 해당하는 출력 인덱스를 복원 레이블로 선택하는 방법을 사용함으로써 레이블 복원 성능 면에서 개선하였다. 또한, [4]가 TV norm을 추가하여 이미지를 정규화한 것처럼 그래디언트의 기울기 손실 함수 외에도 TV norm, l2 norm, Batch 정규화[12], 그룹 일관성 정규화를 추가하여 복원 성능을 개선하였다. [6]에서는 여러 기울기 반전 논문[2-5]을 다양한 측면에서 비교·분석한 내용을 확인할 수 있다. [7]은 다양한 배치상황에서의 기울기 반전을 실험한 논문이다. 실험 결과를 통해 복원하고자 하는 데이터가 증가할수록 원본 이미지와 복원 이미지 간 픽셀 성능 중 Loss, MSE[13]는 높아지고 PSNR[14], SSIM[15]은 낮아지는 것을 확인하였다. 구체적으로, 같은 class의 경우 복원하고자 하는 데이터 수가 1~2개인 경우에는 복원이 잘 되는 반면, 3개 이상부터는 복원 성능이 크게 떨어지는 결과를 보였다. 이 외에도 논문의 데이터 상황별 복원 공격 실험 결과를 통해, 총 데이터 수가 같은 경우 단일 class보다 여러 class의 복원 성능이 더 좋다는 특징을 도출할 수 있다. [8]은 기울기 반전에서 복원하고자 하는 데이터 수가 많을수록 일치하는 그래디언트 쌍을 찾기 힘든 것을 보완하는 기울기 손실 함수를 제안하였다. 연립방정식을 이용하는 방법으로 데이터 수만큼 수식이 많아지지만, 큰 데이터에 대해서도 그래디언트 쌍을 잘 찾을 수 있다. 또한, 복원 단계를 (첫 번

째 FC layer의 출력에 대한 그래디언트 복원)-(첫 번째 FC layer의 입력 복원)-(데이터 복원)의 3단계로 나누어 복원하는 방법을 제안하였다. 더불어 데이터를 복원하는 수식에서는 TV norm을 추가로 사용하여 큰 데이터에서도 높은 복원 성능을 보였다. [9]는 의사 이미지를 생성할 때 랜덤값뿐만 아니라 다양한 이미지를 사용하여 초기화하는 방법을 제안했다. 구체적으로 작은 부분이 반복되어 나타나는 패턴 이미지, Grayscale 이미지를 고려한 검은색과 흰색 이미지, RGB 이미지를 고려한 빨간색과 초록색 그리고 파란색 이미지, 학습에 사용된 데이터와 동일한 class인 최적 이미지를 의사 이미지로 제안하였다. 논문에서 제안한 이미지들을 의사 이미지로 사용하여 복원을 진행하면, 기존의 랜덤값을 사용하여 복원하는 방법보다 빠르게 복원에 수렴하는 결과가 나타났다. 또한, 512×512 크기의 큰 이미지를 복원하기 위해서는 많은 복원 iteration이 필요하다는 실험 결과와 패턴 이미지와 최적 이미지를 의사 이미지로 설정하여 복원을 진행하면, 이를 해결할 수 있다는 실험 결과가 있다.

기울기 반전 관련 연구에 따르면, 의사 레이블로 랜덤값을 사용하는 것보다 실제 레이블을 복원한 레이블을 사용하는 것이 복원 결과에 더 빠르게 수렴할 수 있으며, 실제 레이블을 의사 레이블로 사용하여 실험을 진행하는 논문도 있었다. 또한, 기울기 손실 함수뿐만 아니라 이미지 정규화를 위한 수식을 추가하여 사용하기도 하였으며, 의사 이미지를 랜덤값으로 이루어진 이미지가 아닌 특정 이미지를 사용하는 논문이 있었다. 본 논문에서는 기울기 반전 관련 연구에 따라, 기본적인 기울기 반전 공격인 DLG에서 다음을 추가하여 사용하고자 한다.

- 기존의 의사 레이블로 랜덤값을 사용하는 방법과 실제 레이블을 사용하는 방법 두 가지에 대해 모두 실험한다.
- 기존의 l2 norm으로 이루어진 기울기 손실 함수에 TV norm을 이용한 이미지 정규화 과정을 추가한다.

더불어 의사 이미지를 특정 이미지로 사용하는 연구에 착안하여, RGB 얼굴 이미지에 대한 복원 공격 시 복원 성능을 높이기 위해 살구색 이미지를 의사 이미지로 사용하는 방법을 제안한다.

III. 연합학습에서의 복원 공격

연합학습에서 학습 데이터를 복원할 때, round가 진행될수록 복원이 어려운 것을 이전 연구에서 실험을 통해 확인하였다. 따라서 본 논문은 1 round의 그래디언트를 이용하여 복원 공격을 진행한다. 연합학습 과정과 동일하게 모델을 공유하고, 공유받은 모델로 학습한 후 나온 n개의 원본 그래디언트를 n개의 의사 그래디언트와 비교한다. 이때, 의사 그래디언트는 원본 데이터(이미지, 레이블)와 같은 크기의 의사 데이터(이미지, 레이블)를 글로벌 모델에 학습하여 나온 그래디언트다. 두 그래디언트의 차이가 작아지도록 의사 데이터를 업데이트하는 과정을 반복하면 의사 데이터는 마침내 원본 데이터에 가까워진다. 연합학습에서의 복원 공격 과정을 Fig. 3.에 제시하였다.

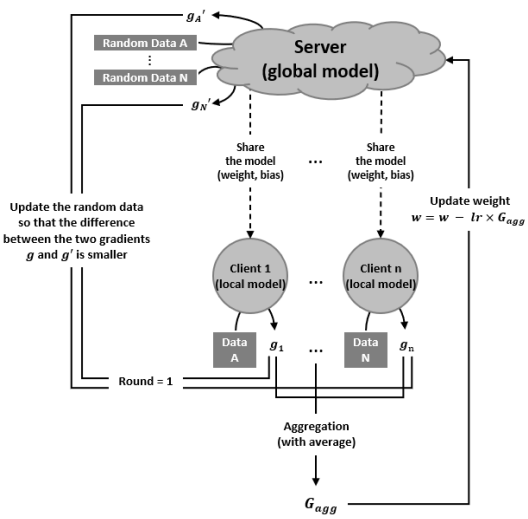


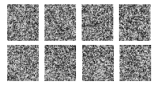
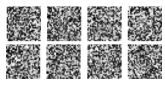
Fig. 3. Process of reconstruction attack in Federated Learning

3.1 제안 방법 및 실험

3.1.1 제안 방법

l2 norm을 사용하는 DLG(dlg)와 TV norm을 추가한 DLG(dlg+tv)를 Yale face database B[16]와 MNIST dataset[17]에 대해 적용한 복원 이미지를 Table 1.에 나타내었다. (dlg)의 경우 노이즈 이미지처럼 보이며, (dlg+tv)의 경우 (dlg)

Table 1. Reconstructed images of 'DLG(dlg)' and 'DLG with TV norm(dlg+tv)'

attack \ data	Yale face database B	MNIST dataset
dlg		
dlg+tv		

보다 복원이 잘 되었지만 약간의 노이즈가 이미지를 덮고 있어 완벽하게 복원되었다고 보기 힘들다. (dlg)는 기울기 반전의 기본 패러다임을 제시한 논문이지만 [7]에서 언급된 바와 같이, 복원하고자 하는 데이터 수가 많은 경우 그 복원 성능이 매우 떨어진다. 우리는 (dlg)와 (dlg+tv)의 복원 이미지에서 보완이 필요하다고 보아 픽셀값을 살펴본 결과, 이미지 픽셀값의 정상 범위인 [0,1]에서 벗어난 픽셀값이 노이즈처럼 보여지는 것을 확인하였다. 이는 (dlg)와 (dlg+tv)과정이 가우시안 노이즈로 이루어진 의사 데이터에서 시작해서 원본 데이터에 가까워지는 것이므로, 복원 iteration이 적거나 그래디언트 내 데이터에 대한 정보를 찾기 힘든 경우가 되면, 복원이 어려워지면서 이미지 픽셀값의 정상 범위인 [0,1]에 도달하지 못하고 노이즈에 머물러 있기 때문이다. 따라서 모든 픽셀값이 범위 [0,1] 안에 들어오도록 픽셀값을 변환하는 방법을 고안했다.

범위 [0,1]을 갖도록 값을 변환하는 방법은 min-max, sigmoid, clip 등이 있다. min-max는 한 이미지의 모든 픽셀값의 min(최솟값)과 max(최댓값)를 이용하여 값을 변형하는 방법으로 수식 (1)을 따른다. sigmoid는 자연 상수 e의 지수함수를 이용하여 값을 변형하는 방법으로 수식 (2)를 따른다. clip의 경우, 설정한 범위 내 값이 들어오도록 범위의 하한값보다 작은 값은 하한값으로 변환하고 상한값보다 큰 값은 상한값으로 변환하는 방법이며, 범위 [0,1]을 갖도록 설정하는 경우 수식 (3)을 따른다.

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)} \quad (1)$$

(x:픽셀값, X:이미지의 모든 픽셀값)

$$x' = \frac{1}{1+e^{-x}} \quad (2)$$

(x:픽셀값)

$$x' = \begin{cases} 0 & (x \le 0) \\ 1 & (x \ge 1) \end{cases} \quad (3)$$

(x:픽셀값)

Table 2.와 Table 3.은 각각 Yale face database B와 MNIST dataset의 한 이미지에 대한 픽셀값 분포 결과이며, 표에는 원본 이미지의 픽셀값 분포, (dlg+tv)로 복원한 이미지의 픽셀값 분포 그리고 복원 이미지를 각 방법으로 변환한 이미지의 픽셀값 분포가 그려져 있다. 복원 이미지의 픽셀값 분포 그래프를 제외한 모든 픽셀값 분포 그래프는 x축의 범위가 [0,1]을 가지며, 복원 이미지의 픽셀값 분포 그래프에서 검정색 세로 선은 0과 1을 의미한다. 이를 통해 복원 이미지에 범위 [0,1]에 포함되지 않는 값이 매우 많은 것을 볼 수 있다.

min-max와 sigmoid로 변환한 이미지의 픽셀값

Table 2. Distribution graph of pixels for one image in Yale face database B (x-axis: the pixel value of the image, y-axis: the count value of the pixel value)

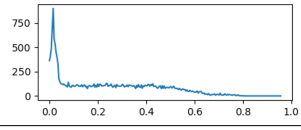
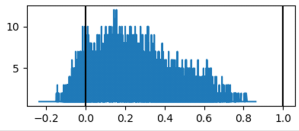
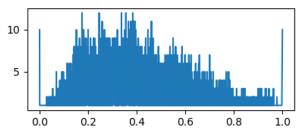
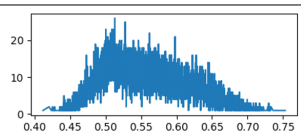
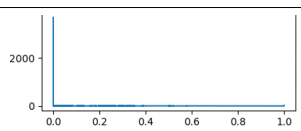
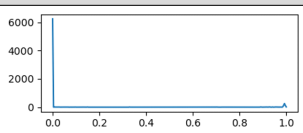
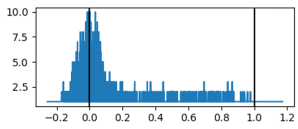
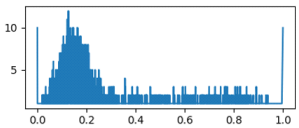
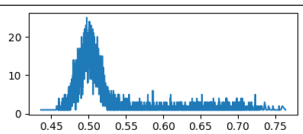
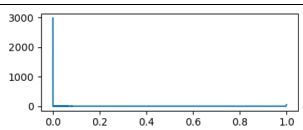
		Distribution of pixels
original		
attack		
Transform	min-max	
	sigmoid	
	clip	

Table 3. Distribution graph of pixels for one image in MNIST dataset (x-axis: the pixel value of the image, y-axis: the count value of the pixel value)

		Distribution of pixels
original		
attack		
Transform	min-max	
	sigmoid	
	clip	

분포 그래프는 원본 이미지의 픽셀값 분포 그래프와 모양이 많이 다르며, 복원 이미지의 픽셀값 분포 그래프와 여전히 닮아있다. 이는 min-max와 sigmoid가 수식에 따라 범위 [0,1] 내 픽셀값들도 변환이 되기 때문이다. 반면, clip으로 변환한 이미지의 픽셀값 분포 그래프는 원본 이미지의 픽셀값 분포 그래프와 매우 비슷하다. 이는 수식에 따라 제대로 복원된 범위 [0,1] 내 픽셀값들의 변환이 없기 때문이며, 다른 방법들과 달리 clip으로 복원 이미지를 변환하는 것이 픽셀값이 과하게 변환되지 않고 노이즈만 제거될 수 있다는 것에 대한 근거가 될 수 있다. Table 2.와 Table 3.에 사용된 이미지뿐만 아니라 다른 이미지에서도 동일한 양상을 띠므로 본 논문은 3가지의 변환 방법 중 clip을 변환 방법으로 사용하는 것을 제안한다.

Table 4.는 Table 1.의 (dlg+tv)의 결과에 본 논문이 제안하는 clip 변환 방법을 사용한 결과로, Table 1.의 결과보다 좀 더 선명한 것을 확인할 수 있다.

Table 4. Reconstructed images of the 'proposed method(dlg+tv+clip)'

attack \ data	Yale face database B	MNIST dataset
dlg+tv +clip		

3.1.2 실험 환경 설정

연합학습에서의 복원 공격은 n round의 연합학습 과정 중 1 round의 그래디언트를 이용하여 복원한다. 따라서 연합학습에서의 복원 공격 실험을 위해서는 연합학습의 실험 설정과 복원 공격 실험 설정 모두 필요하다.

현실적인 연합학습 환경은 대부분 독립적이고 동일하게 배포되는 IID data (Independent and Identically Distributed data)와 반대인 독립적이지 않고 동일하게 배포되지 않는 Non-IID data가 사용된다. 현실적인 연합학습 환경에 맞춰 실험 환경을 Non-IID data로 설정하기 위해 1 클라이언트당 1 class로 배정하였다. 실험에 사용한 데이터는 Yale face database B와 MNIST dataset이다. Yale face database B[16]는 약 60가지의 조명에 따른 38명의 사람 얼굴 Grayscale 이미지로 이루어진 데이터다. 다양한 조명에 따른 이미지이므로 얼굴 식별이 불가능할 정도로 어두운 사진도 포함되어 있으며, 이를 고려하여 0~255의 픽셀값을 4단계로 나누어 평균 픽셀값이 64보다 큰 이미지만 추출하여 사용하였다. 각 사람마다 class가 부여되며, 연합학습의 테스트 데이터 개수를 고려하여 20개 이상의 이미지가 존재하는 class만 남기면 총 37개의 class가 존재한다. 데이터는 192×168의 크기를 갖지만, 실험 시간을 줄이기 위해 48×42 크기로 resize하여 사용한다. MNIST dataset[17]은 0~9의 숫자에 대한 손글씨체 Grayscale 이미지로 이루어진 28x28 크기의 데이터다. 따라서 10개의 class를 가지며, 다양한 환경에서 실험하고자 추가하여 사용한다. 클라이언트의 수는 10개로 고정하고, 1 클라이언트당 학습 데이터 수(per)와 batch size(bs), 연합학습의 round를 다양하게 설정하였다. 테스트 데이터는 client에 해당하는 각 class당 10개로 총 100개의 데이터를 사용하였다. 의사 데이터를 업데이트하는 복원 iteration은 클수록 복원 성능이 좋아지지만,

시간이 많이 소요되므로 모든 복원 실험에 대해 2,000으로 설정하였다. 복원 공격 관련 연구에서 언급했듯이 의사 레이블로 랜덤값과 실제 레이블을 모두 사용하며, 복원 공격은 기본적인 DLG(dlg)와 TV norm을 추가한 DLG(dlg+tv) 그리고 본 논문에서 제안하는 방법(dlg+tv+clip) 3가지에 대해 다루었다. 또한, 본 논문의 실험은 대부분 이미지 데이터의 복원 공격에 대한 실험이므로 이미지를 출력하는 과정이 필수적이다. 이때, 이미지를 출력하는 과정은 PyTorch의 Tensor 형태인 이미지 픽셀값을 Tensor 형태 그대로 이미지로 출력하는 방법과 PIL Image[18]로 변환하여 출력하는 방법이 있다. 따라서 이미지 출력 방법으로 Tensor와 PIL Image를 모두 사용한다.

Table 5.에 실험 환경을 정리해두었으며, 이에 따라 실험을 진행하면 실험의 종류는 2(데이터)×2(의사 레이블)×2(이미지 출력 방법)×3(복원 방법)으로 총 24가지다. 따라서 다양한 상황에서의 복원 공격을 비교할 수 있다. 또한, 한 실험당 100개의 이미지가 복원되므로 총 2,400개의 이미지가 복원된다.

연합학습에서의 복원 공격에 대한 실험을 진행하기 전, 복원 공격의 여부에 따라 연합학습 성능이 달라지는지 파악하고자 연합학습 실험을 진행하였다. Table 5.에 따라 실험을 진행하였으며, 연합학습 실험 결과는 Table 6.에서 확인할 수 있다. 학습 데이터 수가 많을수록 연합학습 성능이 좋으며, batch size 별 성능 차이는 별로 없는 것을 확인할 수 있다. 또한, round가 진행될수록 성능이 좋아져,

Table 5. Environment of experiment

Environment of experiment			
- 1 client : 5, 10, 20 data of 1 class - client = 10 - total train data = 50, 100, 200 - total test data = 100 - batch size = 2, 4 - round = 100, 1000, 10000 - attack iteration = 2000			
data	Yale face database B		MNIST dataset
pseudo-label	dummy label		real label
Image output method	Tensor		PIL Image
attack method	dlg	dlg+tv	dlg+tv +clip

Table 6. Performance of Federated Learning

setting \ data		Yale face database B		MNIST dataset	
per (bs)	round	acc (%)	loss	acc (%)	loss
5 (2)	100	28	3.38	82	1.64
	1,000	81	2.85	89	1.57
	10,000	85	2.81	89	1.57
5 (4)	100	35	3.31	71	1.75
	1,000	95	2.71	88	1.58
	10,000	94	2.72	89	1.57
10 (2)	100	34	3.32	86	1.60
	1,000	92	2.74	100	1.46
	10,000	95	2.71	100	1.46
10 (4)	100	15	3.51	56	1.90
	1,000	93	2.73	100	1.46
	10,000	96	2.70	100	1.46
20 (2)	100	21	3.45	81	1.65
	1,000	100	2.66	100	1.46
	10,000	100	2.66	100	1.46
20 (4)	100	8	3.58	60	1.86
	1,000	96	2.70	100	1.46
	10,000	99	2.67	100	1.46

Yale face database B, MNIST dataset 모두에 대해 최대 100%의 accuracy 성능을 보인다. 연합학습 실험 결과를 통해 batch size 별 성능 차이가 없다는 특징과 round가 클수록 성능이 좋다는 특징을 파악하였으므로 이후 실험은 batch size=4일 때의 10,000 round의 결과만 표시한다.

3.2 평가지표

대부분의 복원 공격 논문에서 복원 성능을 측정하는 방법으로 픽셀값을 이용한 평가지표를 사용한다. 본 논문에서는 MSE, PSNR, SSIM을 사용하여 성능을 평가하였다. MSE (Mean Squared Error)[13]는 평균 제곱 오차로, 이미지가 얼마나 멀리 떨어져 있는지 확인할 수 있다. MSE는 수식 (4)와 같이 계산하며 [0,1] 사이의 실수값을 가지고, 값이 작을수록 복원 성능이 좋다. PSNR (Peak Signal to Noise Ratio)[14]은 신호 대 잡음 비율로, MSE와 픽셀의 최댓값을 이용하며 이미지와 영상 등의 화질 손실 정보를 평가할 때 사용

한다. PSNR은 수식 (5)와 같이 계산하며 [0,∞] 사이의 실수값을 가지고, 값이 클수록 복원 성능이 좋다. SSIM (Structural Similarity Index Measure)[15]은 이미지 품질 평가로, 왜곡 발생 시 원본 이미지에 대한 유사도를 측정하는 방법이다. SSIM은 수식 (6)과 같이 계산하며 [0,1] 사이의 실수값을 가지고, 값이 클수록 복원 성능이 좋다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \quad (4)$$

($\hat{Y}_i$: 예측값의 픽셀, Y_i : 실제값의 픽셀)

$$PSNR = 10 \times \log_{10}(\frac{MAX_I^2}{MSE}) \quad (5)$$

(MAX_I : 픽셀의 최댓값)

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(2\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (6)$$

(μ_x : 실제값의 평균, μ_y : 예측값의 평균,
σ_x : 실제값의 표준편차, σ_y : 예측값의 표준편차,
σ_{xy} : 공분산, c : 변수)

Table 7.은 픽셀 단위의 성능(MSE, PSNR, SSIM)을 Yale face database B와 MNIST dataset에 대해 계산한 결과다. 표에서 출력 방법에 따른 복원 이미지와 함께 살펴보면, 같은 픽셀 단위의 성능에 반해, 복원 이미지는 Tensor로 출력한 이미지가 PIL Image로 출력한 이미지보다 더 깔끔하게 복원된 것을 확인할 수 있다. 이처럼 출력 방법에 따라, 같은 픽셀 단위의 성능이어도 복원 이미지

Table 7. Reconstructed images and performances of pixels by image output method

data	Yale face database B		MNIST dataset	
Images according to the output method	Tensor	PIL Image	Tensor	PIL Image
				
MSE	0.0183		0.0152	
PSNR	17.3864		18.1852	
SSIM	0.8171		0.4778	

Fig. 4. Example images of Human Test

의 노이즈 영역의 크기가 달라지므로 픽셀 단위의 성능뿐만 아니라 눈으로 직접 복원 성능을 평가해야 한다. 따라서 본 논문은 Human Test를 진행하여 식별적인 복원 성능을 평가하였다.

Human Test는 30명의 참가자에게 한 사람당 2,400개의 이미지의 레이블을 맞추는 방식으로 진행하였다. Fig. 4.와 같이 랜덤하게 섞인 복원 이미지가 주어지면 이미지의 레이블을 차례대로 맞추도록 하였으며, 테스트 참가자들이 적어낸 레이블이 실제 레이블과 같으면 복원에 성공한 것으로 간주하였다.

3.3 실험 결과

설정한 실험 환경에 따라 연합학습에서의 복원 공격 실험을 진행하였고, 실험 결과를 픽셀 단위의 성능과 Human Test를 이용하여 정리하였다. Table 8.은 Yale face database B에 대한 연합학습에서의 복원 공격 실험 결과고, Table 9.는 MNIST dataset에 대한 연합학습에서의 복원 공격 실험 결과다.

Human Test를 이용한 성능을 살펴보면, Yale face database B와 MNIST dataset 모두에서 (dlg)보다 (dlg+tv)의 성능이 좋아지고, 본 논문이 제안한 (dlg+tv+clip)의 성능이 가장 좋은 것을 확인할 수 있다. 픽셀 단위의 성능도 마찬가지로 제안 방법의 성능이 가장 좋다. 또한, 연합학습 성능도 93~100%의 accuracy로 공격을 포함하지 않을 때의 연합학습 성능을 유지한다.

정리하면, 본 논문이 제안한 방법을 연합학습에서의 복원 공격 실험에 사용하면 연합학습 성능이 accuracy=99~100%로 높을 때 100개의 학습 데이터 중 최대 100개의 데이터를 식별할 수 있다.

추가적으로, Tensor로 이미지를 출력하는 것이 PIL Image로 변환하여 출력하는 것보다 복원 성능이 좋으며, 의사 레이블에 따른 복원 시간이 비슷한 것으로 보아, 본 논문의 실험에서는 실제 레이블을 사용할 때 복원 이미지에 더 빠르게 수렴한다는 결과가 나타나지 않았다.

Table 8. Result of reconstruction attack in federated learning for Yale face database B (T:Tensor, P:PIL Image)

data		Yale face database B											
pseudo-label		dummy label						real label					
attack method		dlg		+tv		+clip		dlg		+tv		+clip	
output method		T	P	T	P	T	P	T	P	T	P	T	P
Human Test	mean	0.17	0.01	0.86	0.60	0.89	0.70	0.36	0.01	0.87	0.65	0.88	0.75
	best	0.87	0.2	1	0.86	1	0.98	0.85	0.09	1	0.92	1	0.99
	worst	0	0	0.16	0.32	0.42	0.32	0	0	0.34	0.2	0.34	0.22
Pixel performance	MSE	0.089		0.013		0.013		0.068		0.020		0.019	
	PSNR	10.65		19.49		19.58		11.72		17.58		17.71	
	SSIM	0.112		0.664		0.681		0.156		0.619		0.640	
time		38m						41m					
FL	acc	99 (%)						93 (%)					
	loss	2.6663						2.7263					

Table 9. Result of reconstruction attack in federated learning for MNIST dataset (T:Tensor, P:PIL Image)

data		MNIST dataset											
pseudo-label		dummy label						real label					
attack method		dlg		+tv		+clip		dlg		+tv		+clip	
output method		T	P	T	P	T	P	T	P	T	P	T	P
Human Test	mean	0.93	0.02	0.98	0.87	0.98	0.94	0.93	0.02	0.96	0.85	0.96	0.93
	best	0.98	0.14	1	0.96	1	0.99	0.98	0.14	0.99	0.95	0.98	0.98
	worst	0.79	0	0.93	0.12	0.92	0.2	0.8	0	0.92	0.1	0.93	0.09
Pixel performance	MSE	0.077		0.023		0.020		0.074		0.021		0.018	
	PSNR	11.33		16.77		17.51		11.51		17.05		17.82	
	SSIM	0.436		0.550		0.659		0.440		0.569		0.670	
time		30m						28m					
FL	acc	100 (%)						100 (%)					
	loss	1.4612						1.4612					

Human Test의 결과를 비교해보면, 대체적으로 Yale face database B보다 MNIST dataset의 성능이 더 좋다. 이는 단순한 MNIST dataset의 레이블을 맞추는 것보다 복잡한 Yale face database B의 레이블을 맞추는 것이 더 어렵기 때문이다. 또한, 총 10개의 클래스를 사용한 실험에 대해, MNIST dataset은 0~9의 10개의 클래스 보기에서 레이블을 맞추면 되지만 Yale face database B는 0~36의 37개의 클래스 보기에서 레이블을 맞춰야 했기 때문이다. 그러므로 얼굴 이미지의 경우, Human Test 참가자들이 복원 이미지의 레이블을 맞출 때 많은 사람의 얼굴 중에서 맞는 얼굴을 찾는 것에 대한 어려움이 있었을 거라 생각한다. 따라서 실험에 사용된 10개의 class만을 보기로 사용해서 Huamn Test를 추가로 진행한다면 제안 방법의 성능이 더 증가할 것이라고 생각한다. 또한, Table 9.에서 P의 worst 부분이 다른 worst 값보다 평균과의 차이가 심한 것을 볼 수 있다. 이는 가장 초반에 진행한 Human Test가 MNIST dataset을 PIL Image로 출력한 복원 이미지로, 한 테스트 참가자가 테스트 규칙을 잘못 이해하고 테스트하였기 때문에 나타난 결과다. 따라서 해당 테스트 참가자의 결과를

제외하면, 제안하는 (dlg+tv+clip) 방법에 대해 100개의 이미지 중 Yale face database B의 경우 최소 32개에서 최대 100개의 이미지를 식별하였고, MNIST dataset의 경우 최소 94개에서 최대 100개의 이미지를 식별하였다.

IV. 결 론

본 논문은 매개변수를 이용하여 학습 데이터를 복원하는 기울기 반전 공격을 매개변수를 이용하여 학습하는 연합학습 환경에서 시도하였다. 제안하는 (dlg+tv+clip) 방법에 대해 실험한 결과, 100개의 이미지 중 Yale face database B의 경우 최소 32개에서 최대 100개의 이미지를 식별하였고, MNIST dataset의 경우 최소 94개에서 최대 100개의 이미지를 식별하였다. 결과적으로 본 논문에서 제안한 (dlg+tv+clip) 방법이 (dlg) 방법과 (dlg+tv) 방법보다 높은 식별성을 보이며, 픽셀 단위의 성능 면에서도 가장 좋은 성능을 보였다. 따라서 제안하는 방법이 기존의 복원 공격보다 개선됐음이 증명되며, 매개변수가 전달되는 연합학습 상황에서 복원에 성공하였으므로 연합학습이 프라이버시 침해로부터 안전하지 않다는 것을 의미한다. 또한, 출력 방법에 따라 동일한 픽셀 단위의 성능을 가지지만 복원 이미지의 식별적인 성능 차이가 있던 결과를 고려하여, 픽셀 단위의 성능보다 식별적인 성능이 더 중요하다고 말할 수 있다.

더불어 본 논문에 사용된 데이터는 모두 Grayscale 이미지로, 실험에 적용한 환경에서 RGB 이미지에 대한 복원 공격을 진행한 결과 미미한 복원 성능을 보였다. 따라서 본 논문이 제안한 (dlg+tv+clip) 방법은 Grayscale 이미지라는 가정이 필요하다.

향후에는 RGB 이미지까지 확장된 데이터에서도 좋은 성능을 보이는 복원 공격을 연구할 것이며, 본 논문에서 다룬 집계 방법(평균) 외에도 다양한 방법으로 매개변수를 집계하는 연합학습[19]에서 복원 공격을 시도할 것이다. 또한, 그래디언트가 아닌 가중치로 학습하는 연합학습[20]에 대해 기울기 반전 공격이 아닌 가중치를 이용하여 복원 공격을 시도하는 연구도 가치가 있다고 판단한다. 최종적으로 우리는 연합학습에서의 성능은 유지하면서 학습 데이터를 복원하는 다양한 공격에 대해 강건한 방어 방법을 연구할 예정이다.

본 논문에서 시각적인 성능인 Human Test의 결과에 비해 수치적인 성능인 MSE, SSIM, PSNR의 결과는 제안 방법의 이점을 주장하기에는 다른 방법들의 결과의 차이가 미미하다. 따라서 객관적으로 프라이버시 침해 정도를 분석하기 위해 더 다양한 metric으로 제안 방법을 분석할 필요가 있다. 또한, 본 논문은 48x42와 28x28의 크기의 작은 데이터에 대해서만 실험을 진행하였다. 이는 큰 이미지를 복원하는 경우, 의사 데이터가 원본 데이터와 닮아질 때까지 도달하는 시간이 너무 오래 걸리고 복원 성능도 좋지 않기 때문이다. 따라서 향후에 크기가 큰 이미지에서도 복원 시간이 짧고 복원 성능도 좋은 복원 방법에 대해 연구할 계획이다.

References

[1] Google AI, "Federated Learning," https://federated.withgoogle.com, Jan. 2022.

[2] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 14774-14784, Dec. 2019.

[3] B. Zhao, K.R. Mopuri, and H. Bilen, "idlg: Improved deep leakage from gradients," arXiv preprint arXiv:2001.02610, Jan. 2020.

[4] J. Geiping, H. Bauermeister, and H. Dröge, "Inverting gradients - how easy is it to break privacy in federated learning?," Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 16937-16947, Dec. 2020.

[5] H. Yin, A. Mallya, A. Vahdat, J.M. Alvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via gradinversion," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16337-16346, Jun. 2021.

[6] Z. Li, L. Wang, G. Chen, M. Shafiq, and Z. Gu, "A survey of image gradient inversion against federated

learning," TechRxiv, Jan. 2022.
[7] J. Jang, G. Ryu, and D. Choi, "Federated learning privacy invasion study in batch situation using gradient-based restoration attack," Journal of The Korea Institute of Information Security & Cryptology, 31(5), pp. 987-999, Oct. 2021.
[8] X. Jin, P.Y. Chen, C.Y. Hsu, C.M. Yu, and T. Chen, "CAFE: Catastrophic data leakage in vertical federated learning," 34th Conference on Neural Information Processing Systems NeurIPS, pp. 994-1006, Dec. 2021.
[9] W. Wei, L. Liu, M. Loper, K.H. Chow, M.E. Gursoy, S. Truex, and Y. Wu, "A framework for evaluating clinet privacy leakages in federated learning," 25th European Symposium on Research in Computer Security, LNCS 12308, pp. 545-566, Sep. 2020.
[10] Z. Zhang and M.R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," 32nd Conference on Neural Information Processing Systems, Dec. 2018.
[11] M. Persson, D. Bone, H. Elmqvist, "Total variation norm for three-dimensional iterative reconstruction in limited view angle tomography," Physics in Medicine & Biology, vol. 46, no. 3, pp. 853-866, Mar. 2001.
[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," International conference on machine learning. PMLR, vol. 37, pp. 448-456, Jun. 2015.
[13] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," arXiv preprint arXiv:1809.03006, Sep. 2018.
[14] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," 20th international conference on pattern recognition. IEEE, pp. 2366-2369, Aug. 2010.
[15] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp.600-612, Apr. 2004.
[16] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," IEEE transactions on pattern analysis and machine intelligence, vol. 23, no. 6, pp. 643-660, Jun. 2001.
[17] Y. LeCun, C. Cortes, and C.J,C. Burges, "THE MNIST DATABASE of handwritten digits," http://yann.lecun.com/exdb/mnist, Jan. 2022.
[18] Pillow (PIL Fork) 9.4.0 documentation, "PIL," https://pillow.readthedocs.io/en/stable/, Feb. 2022.
[19] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K.B. Letaief, "Communication-efficient edge AI: Algorithms and systems," IEEE Communications Surveys & Tutorials, vol. 22, no. 4, pp. 2167-2191, Feb. 2020.
[20] S, Tijani, "Federated learning: A step by step implementation in tensorflow," https://towardsdatascience.com/federated-learning-a-step-by-step-implementation-in-tensorflow-aac568283399, Jan. 2022.

〈저자소개〉

오 윤 주 (Yoon-ju Oh) 정회원
2021년 2월: 공주대학교 응용수학과 학사
2022년 3월~현재: 숭실대학교 소프트웨어학과 석사과정
〈관심분야〉 연합학습, 개인정보보호, AI 보안

최 대 선 (Dae-seon 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 교수
2016년 ~현재: 정보보호학회 이사
〈관심분야〉 인증, 개인정보보호, AI 보안