이산 코사인 변환 기반 Gradient Leakage 방어 기법

박재훈* · 김광수

성균관대학교

Gradient Leakage Defense Strategy based on Discrete Cosine Transform

Jae-hun Park* · Kwang-su Kim

Sungkyunkwan University

E-mail : zlrnwlzladl@gmail.com / kim.kwangsu@skku.edu

요 약

분산된 환경에서 머신 러닝의 학습 가중치를 공유하여 학습하는 방법은 훈련 데이터를 직접 공유하는 것이 아니기 때문에 안전한 것으로 여겨졌다. 하지만, 최근 연구에 따르면 악의적인 공격자가 공유된 가중치를 분석하여 원본 데이터를 완벽하게 복원할 수 있는 취약점이 발견되었다. Gradient Leakage Attack은 이러한 취약점을 이용해 훈련 데이터를 복원하는 공격 기법이다. 본 연구에서는 개별 장치에서 학습을 진행하고 가중치를 서버와 공유하는 학습 환경인 연합 학습 환경에서 해당 공격을 방어하기 위해 이산 코사인 변환에 기반한 이미지 변환 기법을 제시한다. 실험 결과, 우리의 이미지 변환 기법을 적용하면 공유된 가중치로부터 원본 데이터를 완벽하게 복원할 수 없다.

ABSTRACT

In a distributed machine learning system, sharing gradients was considered safe because it did not share original training data. However, recent studies found that malicious attacker could completely restore the original training data from shared gradients. Gradient Leakage Attack is a technique that restoring original training data by exploiting theses vulnerability. In this study, we present the image transformation method based on Discrete Cosine Transform to defend against the Gradient Leakage Attack on the federated learning setting, which training in local devices and sharing gradients to the server. Experiment shows that our image transformation method cannot be completely restored the original data from Gradient Leakage Attack.

키워드

Gradient, Distributed, Federated Learning, Gradient Leakage, Discrete Cosine Transform

Ⅰ. 서 론

지난 몇 년간 인공지능은 엄청난 발전을 이루었고 다양한 분야에서 연구되고 있다. 이러한 발전의 원동력 중 하나는 많은 양의 데이터이다. 인공지능 학습을 위해서는 데이터를 한곳에 모아야 하지만, 분산된 환경에 존재하는 데이터를 모두 모으는 것은 비용이 많이 든다.

따라서 분산된 환경에서 데이터를 교환하지 않고 공통의 모델을 학습하기 위해 연합 학습 (Federated Learning) [1] 기술이 등장하였다. 연합 학습은 로컬 환경에서 학습을 진행하고 학습된 가중치를 서버로 전송하여 모델을 재구성한다. 이때, 로컬 데이터는 서버로 전송되지 않는다. 중앙 서버는 학습된 가중치만 확인할 수 있으므로 안전한 훈련 방법으로 알려졌다.

하지만 최근 Ligeng Zhu [2]의 연구에 따르면, 공유된 가중치를 활용하여 원본 데이터를 완벽하게 복원할 수 있는 취약점이 발견되었다. [2]의 저자는 해당 취약점을 기반으로 Gradient Leakage Attack을 설계하였고, 실험을 통해 훈련 데이터를 복원하는 데 성공했다.

본 연구에서는 해당 공격을 방어하기 위해 이산

* corresponding author

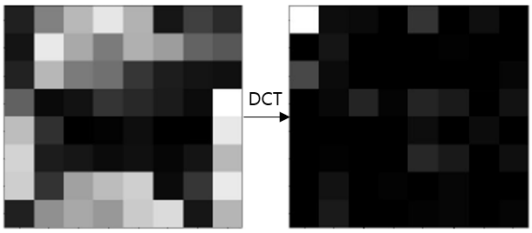

그림 1. 이산 코사인 변환 기법. 왼쪽 이미지 영역에서 이산 코사인 변환을 진행하면 오른쪽 주파수 영역으로 변환이 일어남.

코사인 변환 [3] 기술을 활용한 이미지 변환 기법을 제시한다. 이산 코사인 변환은 그림 1과 같이 이미지 영역에서 주파수 영역으로 변환하는 기술이며, 시각적으로 의미 있는 부분은 단 몇 개의 주파수에 집중되는 점을 이용하여 이미지를 변환하였다.

연합 학습 환경 내에서의 Gradient Leakage Attack 실험 결과, 우리의 이미지 변환 기법을 적용하면 공유된 가중치로부터 원본 데이터를 완벽하게 복원하는 데 실패했음을 확인했다.

Ⅱ. 연합 학습

연합 학습 [1] 을 통해 사용자들은 서로의 데이터를 교환하지 않고 공통의 인공지능 모델을 학습하여 협업한다. 연합 학습은 그림 2와 같이 크게 세 단계로 구성된다. 우선, 연합 학습을 위해 서버에서 인공지능 모델을 디바이스로 전송한다. 그다음, 디바이스는 자신의 데이터를 이용해 서버로부터 받은 모델을 학습한다. 마지막으로, 디바이스는 학습된 모델을 서버로 보내고, 서버는 이를 단일 모델로 결합한다. 인공지능 모델이 수렴할 때까지 위 과정을 반복한다.

Ⅲ. 이미지 변환 기법

이산 코사인 변환 (Discrete Cosine Transform, DCT) [3] 은 신호 처리나 이미지 압축과 같은 분야에 주로 사용되는 기술로, 이미지 영역에서 주파수 영역으로 변환한다. V를 주파수 영역, X를 이미지 영역, C를 이산 코사인 변환 행렬이라고 했을 때, 이산 코사인 변환은 $V = CXC^T$의 식을 통해 이루어진다. 여기서 C는 $N \times N$ 차원의 행렬로써, 식 (1), (2)로 정의된다.

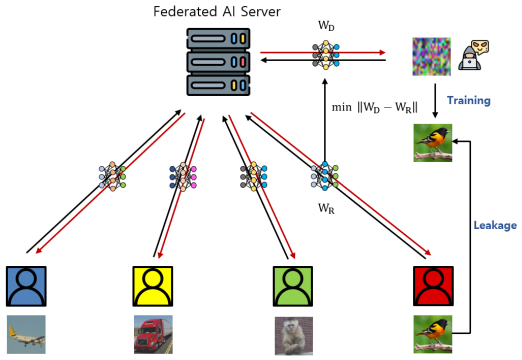

그림 2. 연합 학습 환경에서의 Gradient Leakage Attack

$$c(0,n) = \frac{1}{\sqrt{N}} \tag{1}$$

$$c(k,n) = \sqrt{\frac{2}{N}} cos(\frac{\pi(2n+1)k}{2N}) \tag{2}$$

이산 코사인 변환은 $N \times N$ 블록 단위로 이루어진다. 그림 1과 같이 블록 단위의 이산 코사인 변환 이후 주파수 영역에서 가장 왼쪽 상단에 있는 픽셀 (그림 1 오른쪽 주파수 영역에서 흰색 픽셀)은 저주파를 가지며 DC coefficient라 불리고, 나머지 부분은 고주파로 AC coefficient라 불린다. 고주파 대부분을 제거하여도 이미지를 구성하는 핵심 부분은 남게 된다. 특정 고주파를 제거한 영역을 V^* 라하고, 이에 대한 역변환 $X^* = C^T V^* C$을 진행하면 이미지 정보의 압축이 일어난다.

Ⅳ. 공격 환경

연합 학습 환경 내에서 Gradient Leakage Attack을 실험하기 위해 그림 2와 같이 공격자가 서버에서 특정 클라이언트의 가중치를 관찰할 수 있다고 가정했다. 본 환경에서 각 클라이언트는 다른 클라이언트의 데이터를 확인할 수 없고, 서버는 모든 클라이언트의 데이터를 확인할 수 없다.

공격자는 노이즈 이미지를 생성하고 이로부터 dummy gradient W_D를 생성한다. Victim 클라이언트는 로컬 데이터로부터 real gradient W_R를 생성한다. 공격자는 W_R를 확인하여 W_D와 W_R의 거리를 최소화하도록 노이즈 이미지를 학습한다. 즉, Gradient Leakage Attack은 식 (3)을 만족하는 dummy input x'와 label y'를 찾는 과정이다.

$$x'^*, y'^* = \underset{(x', y')}{\arg\min} \| W_D - W_R \|^2 \tag{3}$$

W_D와 W_R가 유사해지면, 공격자의 노이즈 이미지가 클라이언트의 실제 훈련 데이터와 유사 해진다. 학습 결과가 수렴하면, 공격자는 victim 클라이언트의 특정 훈련 데이터를 복구해낼 수 있다.

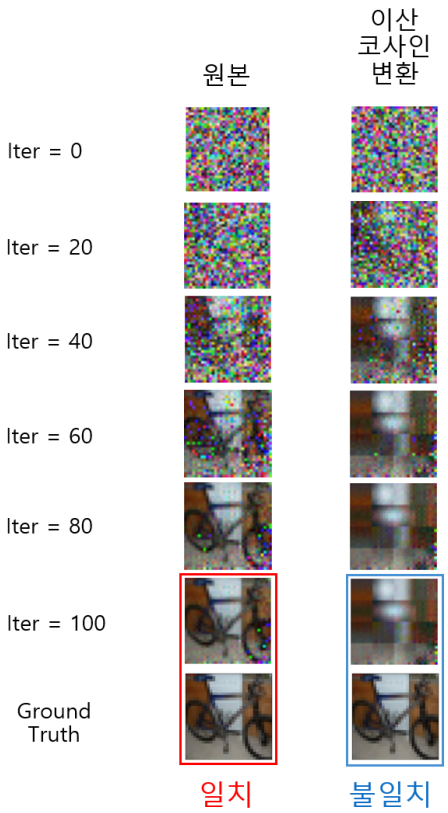

그림 3. 원본 이미지와 이산 코사인 변환 기술을 적용한 이미지에 대해 Gradient Leakage Attack을 진행한 결과

Ⅴ. 실험 및 결과

실험은 Pytorch 환경에서 CIFAR-100 데이터셋을 활용하여 진행하였다. CIFAR-100 이미지는 100개의 클래스로 분류되는 32 × 32 크기의 6만 개의 이미지로 구성되어 있다. 연합 학습 환경을 구성하기 위해 클라이언트의 수는 10으로 설정하고 데이터를 랜덤하게 나누었다. 이때, 클라이언트가 서로의 데이터를 확인할 수 없도록 분산시켰다.

위와 같이 연합 학습 환경을 구성한 후 Victim 클라이언트를 선정하였다. 공격자는 Victim 클라이언트의 가중치를 확인할 수 있고, 이를 활용해 특정 가중치에 대해 Gradient Leakage Attack을 실행하였다. 공격은 L-BFGS [4] 기반 optimization 알고리즘을 이용하여 100 epoch만큼 학습을 진행하였다.

원본 데이터에 대한 Gradient Leakage Attack을 진행한 결과, 그림 3의 좌측 이미지에서 보듯이 victim 클라이언트의 특정 이미지를 완벽히 복구해 내었다.

해당 공격을 방어하기 위해 Victim 클라이언트의 데이터에 이산 코사인 변환을 적용했다. 이산 코사인 변환의 계산 효율성을 위해 윈도우 블록 크기를 8 × 8로 설정했다. 그림 3의 우측 이미지에서 보듯이 이산 코사인 변환 기법을 적용한 이미지에 대해 Gradient Leakage Attack을 진행한 결과 이미지를 완벽히 복구하는 데 실패하였음을 알 수 있다.

Ⅵ. 결 론

가중치는 원본 데이터를 반영하므로 가중치를 공유하는 것은 원본 데이터를 공유하는 것과 같은 위험성을 지니고 있다. 가중치로부터 Gradient Leakage Attack을 통해 원본 데이터를 복구할 수 있음을 실험적으로 확인하였다.

본 연구에서는 이를 방어하기 위해 이산 코사인 변환을 활용한 이미지 변환 기법을 제시하였고, 실험 결과 원본 데이터 복구를 방어하는 데 효과적이었다.

Acknowledgement

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신산업진흥원의 지원을 받아 수행된 헬스케어 AI 융합 연구개발 사업 (No.S0316-21-1006)과 5G-IoT 환경에서 이기종・비정형・대용량 데이터의 고신뢰・저지연 처리를 위한 플랫폼 개발 및 실증 사업(No.2020-0-00990)의 지원을 받아 수행된 연구임

References

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, “Communication -efficient learning of deep networks from decentralized data”, *Artificial Intelligence and Statistics*, Florida: FL, pp. 1273-1282, 2017.

[2] ZHU, Ligeng; HAN, Song, *Federated Learning*, Switzerland: Springer, Cham, pp. 17-31, 2020.

[3] Ahmed, Nasir, T_ Natarajan, and Kamisetty R. Rao, “Discrete cosine transform”, *IEEE transactions on Computers*, New York: NY, pp. 90-93, 1974.

[4] Liu, Dong C., and Jorge Nocedal, “On the limited memory BFGS method for large scale optimization”, *Mathematical Programming*, Vol. 45.1, pp.503-528, August 1989.