Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.29, NO.3, Jun. 2019 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2019.29.3.589

AI 환경에서 모델 전도 공격에 안전한 차분 프라이버시 기술*

박 철 희,† 홍 도 원‡
공주대학교

Differential Privacy Technology Resistant to the Model Inversion Attack in AI Environments*

Cheollhee Park,† Dowon Hong‡
Kongju National University

요 약

온라인상에 축적되는 디지털 데이터의 양은 폭발적으로 증가하고 있으며 이러한 데이터들은 매우 큰 잠재적 가치를 갖고 있다. 국가 및 기업들은 방대한 양의 데이터로부터 다양한 부가가치를 창출하고 있으며 데이터 분석 기술에 많은 투자를 하고 있다. 그러나 데이터 분석에서 발생하는 프라이버시 문제는 데이터의 활용을 저해하는 큰 요인으로 작용하고 있다. 최근 신경망 모델 기반의 분석 기술에 대한 프라이버시 침해 공격들이 제안됨에 따라 프라이버시를 보존하는 인공 신경망 기술에 대한 연구가 요구되고 있다. 이에 따라 엄격한 프라이버시를 보장하는 차분 프라이버시 분야에서 다양한 프라이버시 보존형 인공 신경망 기술에 대한 연구가 수행되고 있지만, 신경망 모델의 정확도와 프라이버시 보존 강도 사이의 균형이 적절하지 않은 문제점이 있다. 본 논문에서는 프라이버시와 모델의 성능을 모두 보존하고 모델 전도 공격에 저항성을 갖는 차분 프라이버시 기술을 제안한다. 또한, 프라이버시 보존 강도에 따른 모델 전도 공격의 저항성을 분석한다.

ABSTRACT

The amount of digital data a is explosively growing, and these data have large potential values. Countries and companies are creating various added values from vast amounts of data, and are making a lot of investments in data analysis techniques. The privacy problem that occurs in data analysis is a major factor that hinders data utilization. Recently, as privacy violation attacks on neural network models have been proposed. researches on artificial neural network technology that preserves privacy is required. Therefore, various privacy preserving artificial neural network technologies have been studied in the field of differential privacy that ensures strict privacy. However, there are problems that the balance between the accuracy of the neural network model and the privacy budget is not appropriate. In this paper, we study differential privacy techniques that preserve the performance of a model within a given privacy budget and is resistant to model inversion attacks. Also, we analyze the resistance of model inversion attack according to privacy preservation strength.

Keywords: Differential privacy, model inversion attack, privacy-preserving neural network

Received(03.09. 2019), Accepted(05. 09. 2019)
* 본 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행하였습니다(2019R1A2C1003146).

† 주저자, chpark89@smail.kongju.ac.kr
‡ 교신저자, dwhong@kongju.ac.kr(Corresponding author)

I. 서 론

온라인상에 축적되는 디지털 데이터가 폭발적으로 증가하고 컴퓨터 연산 능력이 향상됨에 따라 데이터 분석과 활용을 통해 다양한 가치가 창출되고 있다. 최근 국가와 기업들은 데이터 분석 연구에 많은 투자를 하고 있으며, 데이터 분석 모델로서 인공 신경망 모델은 얼굴 인식, 추천 및 의료와 같은 다양한 분야에서 뛰어난 성능을 나타내고 있다[1][2][3].

그러나 최근 인공 신경망 모델에 대하여 프라이버시를 침해할 수 있는 공격기법들이 제안됨에 따라 프라이버시를 보존하는 데이터 분석 기술 개발에 대한 요구가 급증하고 있다[4][5][6][7]. 이에 따라 기본적인 프라이버시 보존 방식이 제안되었지만, 모델의 성능을 크게 떨어뜨리거나 여전히 프라이버시를 보존하지 못하는 문제점을 갖고 있다.

차분 프라이버시는 엄격한 프라이버시 개념으로써 데이터 수집, 공유 및 분석 등 다양한 분야에서 프라이버시 보존을 목적으로 광범위하게 적용되고 있다[8][9]. 차분 프라이버시는 주어진 데이터베이스에 대하여, 데이터베이스에 존재하는 개별 데이터에 대해 해당 데이터의 존재 여부에 상관없이 유사한 분포로 질의에 응답하는 메커니즘을 수행한다. 최근 기계학습 및 인공 신경망 분야에서 프라이버시를 보존하기 위해 차분 프라이버시가 적용되고 있지만, 프라이버시 파라미터의 설정에 따라 프라이버시를 충분히 보장하지 못하거나 분석 모델의 성능이 크게 떨어질 수 있는 문제점을 갖고 있다.

본 논문에서는 인공 신경망 모델에서 모델 전도 공격에 대하여, 해당 공격에 저항성을 갖고 모델의 성능을 보존할 수 있는 차분 프라이버시 적용 기법에 관한 연구를 수행한다. 특히 기존에 제안된 차분 프라이버시를 만족하는 학습 기법에 대하여, 가속화된 학습으로의 확장 및 모델 전도 공격에 저항성을 갖는 프라이버시 파라미터 설정에 관한 연구를 수행한다.

기존의 차분 프라이버시를 보존하는 확률적 경사 하강법은 random sampling을 통해 차분 프라이버시 파라미터에 대한 효율적인 composition(moments accountant)을 가능하게 한다[10]. 그러나 확률적 경사 하강법은 데이터의 형태와 학습 모델의 목적에 따라 주어진 프라이버시 파라미터 경계 내에서 모델이 충분히 학습되지 않을 수 있다. 따라서 이를 확장하여 가속화된 학습을 통해 주어진 프라이버시 파라미터의 경계 내에서 모델의 성능을 향상시킬 수 있는 방안에 대해 연구한다. 또한, 주어진 프라이버시 파라미터의 범위 내에서, 차분 프라이버시는 만족하지만, 모델 전도 공격에 대해 저항성을 갖지 않는 문제점을 분석하고 랜덤 노이즈의 표준편차(noise scale)에 따른 프라이버시 만족도를 분석한다.

II. 연구 배경

이번 장에서는 연구의 배경이 되는 인공 신경망 모델, 모델 전도 공격 및 차분 프라이버시 개념에 대해 설명한다.

2.1 인공 신경망 모델(neural network model)

신경망 모델은 단일 퍼셉트론부터 심층 신경망에 이르기까지 다양한 깊이를 가질 수 있으며 합성곱, 완전연결, 비선형 함수 등 다양한 요소들로 구성될 수 있다. 입력으로부터 출력까지의 단계를 파라미터로 나타낼 수 있는 함수이며, 최근 기계학습 분야에서 뛰어난 효율성과 성능을 보이고 있다. 신경망 모델은 데이터의 입력과 출력의 관계를 적합하게 표현하도록 하는 목적을 가지며, 주어진 학습 데이터 집합을 통해 해당 목적을 학습한다. 이때 학습 과정은 신경망 모델을 통해 해결하고자 하는 문제(task)에 대하여 학습 데이터로부터의 오류를 손실함수로 정의하고 해당 손실함수가 최소가 되도록 모델을 최적화한다. 이때, 최적화는 일반적으로 확률적 경사 하강법을 통해 수행되며 효율성을 위해 데이터 집합에서 일부 데이터를 무작위로 추출(mini batch)하여 수행된다. 최적화 과정은 momentum, Adagrad, Adam 등과 같은 최적화 방식을 통해 학습 속도를 향상시킬 수 있다. 학습 과정은 손실함수를 최소화하는 목적을 갖지만, 일반적으로 신경망 모델의 손실함수는 non-convex이며 국소적인 최솟값(local minimum)으로 수렴할 수 있다. 따라서 최적화 과정은 수용할 수 있을 정도의 작은 손실량을 갖도록 모델을 학습시킨다.

2.2 모델 전도 공격(model inversion attack)

클래스를 갖는 데이터의 분류 문제에 대한 기계학습 모델은 일반적으로 추론 단계에서 주어진 입력이 속하는 클래스 및 그에 대한 신뢰도 값을 출력한다. Fredrikson 등[5]은 이러한 출력값들을 이용하여

주어진 모델의 학습 데이터를 복구할 수 있는 모델 전도 공격을 제안하였다. 특히, 얼굴 인식 문제를 위해 학습된 신경망 모델로부터 학습 데이터가 복구될 수 있음을 밝혔으며 white-box 접근법뿐만 아니라 black-box 접근법으로도 해당 공격이 가능함을 보였다. 또한, 단일 계층 신경망뿐만 아니라 다중 계층 신경망 및 심층 신경망에 대해 공격이 가능함을 보였다.

모델 전도 공격은 학습 데이터를 복구하는 문제를 기계학습의 최적화 문제로 변환하여 경사 하강법을 통해 손실함수를 최소화시키는 방식으로 데이터를 복구한다. 이는 표적이 되는 클래스에 대한 신뢰도 값이 최대가 되는 입력값을 찾도록 최적화를 수행한다. 즉, 주어진 타겟 모델 f에 대하여, $f_i(X)$가 입력 벡터 X에 대한 타겟 모델의 출력에서 X가 클래스 i에 속할 확률(신뢰도 값)을 나타낸다면, $(1-f_i(X))$을 손실함수 $\mathcal{L}$로 정의하여 경사 하강법을 통해 $\mathcal{L}$이 최소가 되도록 X에 대한 최적화를 수행한다. 최종적으로 최적화가 완료된 X는 클래스 i에 해당하는 데이터라는 사실을 쉽게 식별할 수 있다.

Fredrikson 등[5]은 신경망 모델을 활용한 얼굴 인식 서비스뿐만 아니라 의사결정나무 모델에서도 모델 전도 공격이 가능함을 보였으며, 모델 전도 공격에 대해 기본적인 프라이버시 보존법을 제안하였다. 그러나 기본적인 프라이버시 보존을 위한 접근법은 모델의 성능 및 신뢰도를 크게 떨어뜨리거나 프라이버시를 충분히 보존하지 못할 위험이 존재한다[6][10].

2.3 차분 프라이버시[8][9]

차분 프라이버시는 Dwork 등[8][9]에 의해 제안된 프라이버시 개념이며, 주어진 데이터베이스에 대하여 해당 데이터베이스로의 질의에 대해 프라이버시를 보존할 수 있는 특정 메커니즘을 수행한다. 차분 프라이버시를 만족하는 메커니즘은 주어진 질의에 임의의 노이즈와 같은 무작위성을 추가하여 응답을 반환한다. 즉, 차분 프라이버시는 주어진 데이터베이스에 대하여 개별 데이터의 존재 여부에 상관없이 질의에 대한 응답이 유사한 분포를 갖도록 하여 프라이버시를 보존한다. 차분 프라이버시를 만족하는 메커니즘의 정의는 다음과 같다.

정의 1. ((ϵ,δ)-차분 프라이버시). 입력 공간 D와 출력 공간 R을 갖는 무작위 메커니즘 $M: D \rightarrow R$ 은 단일 차이를 갖는 인접한 두 부분집합 $d, d' \in D$와 출력 $S \subseteq R$에 대하여 다음식을 만족할 때 (ϵ,δ)-차분 프라이버시를 만족한다고 한다.

$$\Pr[M(d) \in S] \leq e^{\epsilon} \Pr[M(d') \in S] + \delta.$$

이때 $\epsilon > 0$이고 $\delta \geq 0$이며, $\delta = 0$인 경우 ϵ-차분 프라이버시를 만족한다고 한다. 프라이버시 파라미터 ϵ이 작을수록 엄격한 프라이버시가 보장되며 δ는 일반적으로 $1/|d|$보다 작은 값을 선택한다. 차분 프라이버시를 만족하는 메커니즘 M은 임의의 함수 g와 결합된 경우에도 차분 프라이버시를 만족하는 사후 처리 저항성(immune to post-procession)을 갖는다.

차분 프라이버시는 그룹 프라이버시를 보장할 수 있으며, 결합이 가능한 성질(composability)을 갖고 있다.

그룹 프라이버시(group privacy): 데이터베이스 d에 존재하는 개별 데이터들은 개별 주체와 1대 1로 연결된다고 가정한다. 그러나 단일 주체에 대해 2개 이상의 데이터가 연결된 경우 그룹 프라이버시를 고려해야 한다. 이 경우 프라이버시 파라미터를 선형적으로 증가시켜 데이터 그룹에 대한 차분 프라이버시를 만족하도록 한다.

결합성(composition): 주어진 데이터베이스에 대한 반복적인 질의에 대하여, 차분 프라이버시는 모든 질의에 대해 프라이버시 파라미터를 결합하여 전체적인 차분 프라이버시를 보장하도록 한다. 즉, 모든 질의에 대한 프라이버시 파라미터를 결합하며 질의의 집합 전체에 대해 차분 프라이버시를 만족하도록 한다. Basic composition[11][12] 및 advanced composition[13][14] 등의 결합 이론을 통해 프라이버시 파라미터를 효율적으로 결합할 수 있으며, 무작위 추출단계를 포함하는 질의에 대한 매우 효율적인 결합 기법들이 제안되었다[10][15].

근본적인 차분 프라이버시를 만족하는 메커니즘은 대표적으로 질의에 랜덤 노이즈를 추가하는 Laplace 메커니즘과 무작위성(randomness)을 추가하여 선택(selection)을 수행하는 exponential 메커니즘이 있다. 또한, (ϵ,δ)-차분 프라이버시를 만족하는 메커니즘은 대표적으로 질의에 랜덤 노이즈를 추가하는 Gaussian 메커니즘이 있다. 이때 평균이 0인 Gaussian 분포로부터 노이즈를 추출하며 분포의 스케일(표준편차)은 질의의 민감도에 의해 결정된

Algorithm 1 Differentially Private SGD

Input : Examples $\{x_1, x_2, \ldots, x_n\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$.
Parameters → Learning rate η_t, noise scale σ, group size L, gradient norm bound C

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with probability L/N

Compute gradient
For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient
$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise
$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output : θ_T

Fig. 1. Differentially private stochastic gradient descent algorithm.

다. Gaussian 메커니즘에서 질의의 민감도는 ℓ_2-민감도를 따르며 다음과 같이 정의한다.

정의 2. (ℓ_2-민감도). 입력 공간 D와 출력 공간 R을 갖는 임의의 함수 $f: D \rightarrow R$ 에 대하여, 인접한 두 부분집합 $d, d' \in D$ 대한 f의 민감도 $\Delta_2(f)$는 다음과 같다.

$$\Delta_2(f) = \max_{d,d' \in D} \|f(d) - f(d')\|_2.$$

(ϵ, δ)-차분 프라이버시를 만족하는 Gaussian 메커니즘은 다음과 같다.

정리 1. (Gaussian 메커니즘[9]). $c^2 \geq 2\ln(1.25/\delta)$, $0 < \epsilon < 1$이고 표준편차 $\sigma \geq c\Delta_2(f)/\epsilon$일 때, 임의의 질의 함수 f에 대하여 Gaussian 메커니즘 $M(d) = f(d) + N(0, \sigma^2)$은 (ϵ, δ)-차분 프라이버시를 만족한다.

Abadi 등[10]은 무작위 추출과 Gaussian 메커니즘을 통한 알고리즘에 대해 매우 효율적인 결합 이론인 moments accountant를 제안했다.

III. 모델 전도 공격에 안전한 차분 프라이버시를 만족하는 신경망 학습 기법

이번 장에서는 차분 프라이버시를 만족하는 메커니즘을 기반으로 모델 전도 공격에 저항성을 갖는 신경망 모델 학습에 대해 설명한다. 첫 번째로, 기본적인 차분 프라이버시를 보장하는 신경망 모델 학습에 대해 분석하고, 그에 대한 문제점 및 개선 방안에 대해 설명한다. 두 번째로, 차분 프라이버시를 만족하는 얼굴 인식 모델에 대한 모델 전도 공격을 설명한다.

얼굴 인식 모델을 위해 Fredrikson 등[5]의 모델 전도 공격 실험에 사용된 AT&T laboratories Cambridge 얼굴 데이터베이스[16]를 활용하였다. 전체 데이터베이스는 400개의 흑백 이미지로 구성되며, 각 이미지의 크기는 92×112픽셀로 구성된다. 또한, 데이터베이스는 40개의 클래스(40명)로 구성되며, 단일 클래스에 10개의 이미지가 포함된다. 모델 학습을 위해 데이터베이스의 70%(각 클래스당 7개의 이미지를 무작위로 추출함)를 학습 데이터로 활용하고 나머지를 테스트 데이터로 활용한다.

3.1 차분 프라이버시 보존형 얼굴 인식 모델 학습

신경망 기반의 인공지능 모델은 다양한 분야에서 타 기계학습 모델보다 뛰어난 성능을 나타내고 있으며, 그중 얼굴 인식 분야에서 매우 뛰어난 인식률을 보이고 있다. 그러나 최근 신경망 기반의 얼굴 인식 모델에 대하여, 모델의 학습을 위해 사용된 학습 데이터를 복원할 수 있는 모델 전도 공격이 제안됨에 따라, 학습 데이터의 프라이버시를 보존할 수 있는 프라이버시 보존형 모델 학습 기술 개발이 요구되고 있다. 이에 따라 Abadi 등[10]은 신경망 학습 방식인 확률적 경사 하강법에 대한 차분 프라이버시 기법

Algorithm 2 DPSGD with Momentum

Input : Examples $\{x_1, x_2, \ldots, x_n\}$, loss function $\mathcal{L}(\theta') = \frac{1}{N}\sum_i \mathcal{L}(\theta', x_i)$.
Parameters → Learning rate η, noise scale σ, group size L, gradient norm bound C, momentum α

Initialize θ'_0 randomly ($\mathbf{v}_0 = 0$)

for $t \in [T]$ **do**

Take a random sample L_t with probability L/N

Compute gradient
For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta'_t} \mathcal{L}(\theta'_t, x_i)$

Clip gradient
$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise
$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Compute velocity
$\mathbf{v}_{t+1} \leftarrow \alpha \mathbf{v}_t - \eta \tilde{\mathbf{g}}_t$

Update
$\theta'_{t+1} \leftarrow \theta'_t + \mathbf{v}_{t+1}$

Output : θ'_T

Fig. 2. Differentially private momentum update algorithm.

과 효율적인 차분 프라이버시 결합 기법인 moments accountant를 제안했다. Algorithm 1(Fig. 1)은 Abadi 등에 의해 제안된 차분 프라이버시 보존형 확률적 경사 하강법(DPSGD)을 나타낸다. 추출된 데이터 집합(mini-batch)에 대해 각각의 데이터에 대한 기울기를 계산(Compute gradient)하고, 차분 프라이버시의 민감도를 위해 개별 데이터에 대한 최대 기울기를 제한(Clip gradient)한 후, 차분 프라이버시를 위해 Gaussian 노이즈를 추가(Add noise)한다. 최종적으로 모델을 업데이트(Descent)한다. 위의 과정을 주어진 반복횟수 T 동안 반복하여 모델 θ_T를 출력한다. Abadi 등은 T-반복에 대한 차분 프라이버시 결합을 위해 효율적인 결합 이론인 moments accountant를 제안했다.

정리 2. (Moments accountant[10]). 적당한 상수 c_1, c_2, 무작위 추출 확률 $q = L/N$, 반복횟수 T가 존재할 때, 프라이버시 파라미터 $\epsilon < c_1 q^2 T$에 대하여, 각 표준편차가 아래 식을 만족할 때, Algorithm 1은 (ϵ, δ)-차분 프라이버시를 만족한다.

$$\sigma \geq c_2 \frac{q\sqrt{T\log(1/\delta)}}{\epsilon}.$$

차분 프라이버시를 보존하는 신경망 기반의 얼굴 인식 모델을 생성하기 위해 Algorithm 1을 활용하여 모델을 학습시킬 수 있지만, 적당한 차분 프라이버시 예산(privacy budget) 내에서 충분한 모델 학습이 불가능할 수 있다. 프라이버시가 고려되지 않은 상황에서 확률적 경사 하강법은 반복횟수(파라미터 업데이트)에 제안이 없지만, 주어진 프라이버시 예산에서 차분 프라이버시가 고려된 경우 반복횟수가 제한되기 때문에 복잡한 데이터 분석에서 확률적 경사 하강법은 충분한 학습이 불가능할 수 있다. 실제로, 얼굴 인식 모델에 대하여 Algorithm 1을 통한 최적화를 수행하였으며(Fig. 3의 왼쪽) 학습이 충분히 수행되지 않음을 알 수 있다. 따라서 본 논문에서는 주어진 프라이버시 예산 내에서 충분한 모델 최적화를 위해 차분 프라이버시를 보장하는 가속화된 학습법(differentially private accelerated learning)을 고려한다.

가속화된 학습 방식으로는 대표적으로 momentum, Adagrad, Adam 등이 있으며 Algorithm 1을 확장하여 차분 프라이버시를 보장하는 가속학습법을 구성한다. Algorithm 2(Fig. 2)는 차분 프라이버시를 보장하는 momentum 학습법을 나타낸다. Momentum 학습법은 기울기 방향으로 힘을 받아 가속된다는 물리법칙을 따라 구성되었으며 속도(velocity)를 유지하여 그 크기에 따라 학습이 가속된다. Algorithm 2는 Algorithm 1과 마찬가지로, Compute gradient 과정, Clip gradient 과정, Add noise 과정을 수행하고 가속학습을 위해 속도를 계산(Compute velocity)하여 모델을 업데이트한다.

Algorithm 2에서 velocity를 유지하는 과정은 추

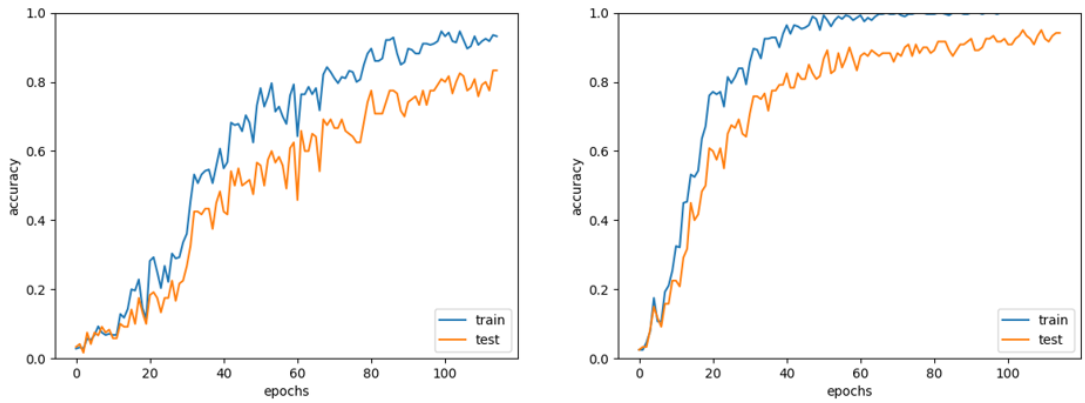

Fig. 3. Learning processes of a single layer neural network model for AT&T face data within the same privacy budget($\epsilon=10, \delta=10^{-3}$).(Left) Optimization with Algorithm 1, (Right) Optimization with Algorithm 2

가적인 데이터베이스 접근 없이 수행되기 때문에 차분 프라이버시 관점에서 사후 처리 과정에 해당한다. 따라서 Algorithm 2는 moments accountant를 따르며 (ϵ, δ)-차분 프라이버시를 만족한다.

정리 3. (DPSGD with momentum). 차분 프라이버시 관점에서 Algorithm 2는 Algorithm 1의 post-processing이다. 따라서 moments accountant를 따른다면 Algorithm 2는 (ϵ, δ)-차분 프라이버시를 만족한다.

증명) Algorithm 1의 출력 θ_T와 Algorithm 2의 출력 θ'_T은 다음과 같이 전개할 수 있다.

$$\theta_T = \theta_0 - \sum_{t=1}^{T} \eta_t \tilde{g}_t$$

$$\theta'_T = \theta_0 - \sum_{t=1}^{T} ((\sum_{i=1}^{t} \alpha^{i-1}) \eta \tilde{g}_t)$$

이때 θ_0, 그리고 θ'_T에서 누적 기울기 항(sigma term)의 계수 $\eta \sum_{i=1}^{t} \alpha^{i-1}$는 데이터 집합 $\{x_1, ..., x_N\}$에 독립적이다. 따라서 θ'_T에서 누적 기울기 항은 차분 프라이버시 관점에서 적당한 상수 k 대하여, 다음과 표현될 수 있다.

$$\sum_{t=1}^{T} ((\sum_{i=1}^{t} \alpha^{i-1}) \eta \tilde{g}_t) = k \cdot \sum_{t=1}^{T} \eta_t \tilde{g}_t$$

이때 Algorithm 2가 정리 2를 따른다면 θ'_T는 (ϵ, δ)-차분 프라이버시를 만족한다. ■

Fig. 3은 AT&T 얼굴 데이터에 대한 단일 계층 신경망 모델에 대하여, 차분 프라이버시를 만족하는 확률적 경사 하강법(Algorithm 1)과 momentum 학습(Algorithm 2)의 최적화 과정을 나타낸다. 주어진 차분 프라이버시 파라미터 $(\epsilon=10, \delta=10^{-3})$에 대하여, Algorithm 1를 통한 학습 과정(Fig. 3의 왼쪽)은 학습이 충분히 진행되지 않고 끝나는 반면에, Algorithm 2를 통한 학습 과정(Fig. 3의 오른쪽)은 충분한 학습이 진행되었음을 확인할 수 있다.

3.2 차분 프라이버시를 보장하는 얼굴 인식 모델에 대한 모델 전도 공격

Fredrikson 등[5]에 의해 제안된 모델 전도 공격은 목표로 하는 사람에 대한 클래스 값(이름 또는 색인)을 이용해 데이터를 복구한다. 일반적인 모델과 마찬가지로, 차분 프라이버시를 보장하는 얼굴 인식 모델은 주어진 입력에 대한 출력값으로 신뢰도 값을 반환한다. 따라서 차분 프라이버시가 보장된 모델에 대해 모델 전도 공격을 그대로 수행할 수 있다. Algorithm 3(Fig. 4)은 Fredrikson 등에 의해 제안된 얼굴 인식 모델에 대한 모델 전도 공격을 나타낸다.

모델 전도 공격을 기계학습의 최적화 문제로 변환하여 경사 하강법을 통해 손실함수를 최소화시키는 방식으로 데이터를 복구한다. $c(X)$는 모델 전도 공

Algorithm 3 Inversion attack for facial recognition models.

Function Model_inversion($label, \alpha, \beta, \gamma, \lambda$)

$c(\boldsymbol{X}) \stackrel{\text{def}}{=} 1 - f_{label}(\boldsymbol{X})$

$\boldsymbol{X}_0 \leftarrow \mathbf{0}$

for $i \leftarrow 1, \ldots, \alpha$ **do**

$\boldsymbol{X}_i \leftarrow \text{Process}\left(\boldsymbol{X}_{i-1} - \lambda \cdot \nabla c(\boldsymbol{X}_{i-1})\right)$

if $c(\boldsymbol{X}_i) \geq \max\left(c(\boldsymbol{X}_{i-1}), \ldots, c(\boldsymbol{X}_{i-\beta})\right)$ **then**

break

if $c(\boldsymbol{X}_i) \leq \gamma$ **then**

break

return $\left[\text{argmax}_{\boldsymbol{X}_i}\left(c(\boldsymbol{X}_i)\right), \min_{\boldsymbol{X}_i}(c(\boldsymbol{X}_i))\right]$

Fig. 4. Algorithm of model inversion attack.

격에 대한 손실함수를 나타내며 f는 얼굴 인식 모델을 나타낸다($f_i(X)$는 입력 X에 대한 모델의 추론에서 i번째 클래스값(확률)을 나타냄). 주어진 반복횟수 α동안 최적화하며 β번 이상 학습이 진행되지 않거나 γ이하의 손실량이 측정되면 최적화를 완료한다. 이때 λ는 모델 전도 공격의 학습률을 의미하며, Process(•)는 이미지 데이터에 대한 de-noising 및 generalizing과 같은 이미지 처리를 위한 일반적인 과정을 나타낸다. Fig. 5는 AT&T 얼굴 데이터에 대한 얼굴 인식 모델로부터 모델 전도 공격을 통해 추출된 얼굴 데이터의 예시를 나타낸다.

(original face data)

(inverted face data)

Fig. 5. (Left) A face image for the specific individual used in learning face recognition model(test data). (Right) A Face image recovered from model inversion attack.

3.2.1 차분 프라이버시 보존형 얼굴 인식 모델 생성

모델 전도 공격에 대한 차분 프라이버시의 저항성을 분석하기 위해 차분 프라이버시를 보장하는 신경망 모델을 생성한다. 본 논문에서는 가장 기본적인 단일 계층 신경망 모델(또는 softmax regression model)을 고려한다. 입력 데이터는 92×112픽셀을 갖는 흑백 이미지이므로 입력 layer의 크기는 10304이며, 데이터 집합은 40명에 대한 이미지로 구성되기 때문에 출력 layer의 크기는 40이다. 따라서 해당 신경망 모델에서 최적화가 수행되는 가중치 파라미터는 총 10304×40개이다.

주어진 프라이버시 예산 내에서 Gaussian 노이즈의 스케일(표준편차) 및 학습의 반복횟수를 설정할 수 있다. 즉, 주어진 프라이버시 파라미터 (ϵ, δ), 노이즈 스케일 σ, 그리고 무작위 추출 확률 q에 대하여 정리 2를 통해 반복횟수 T를 결정할 수 있다. 이때 c_2는 주어진 상수이며 신경망 모델은 qT-epoch을 갖는다. 다양한 프라이버시 파라미터에 대해 모델을 학습시켰으며, 프라이버시 파라미터값이 동일한 경우 Algorithm 1와 Algorithm 2은 동일한 반복횟수를 갖기 때문에 최적화 과정은 가속학습이 가능한 Algorithm 2를 활용하였다.

Table 1은 이에 대한 결과를 나타낸다. 동일한 프라이버시 예산에 대해, 노이즈 스케일과 학습의 반복횟수를 다양하게 변경하여 학습을 수행하였다. 주어진 프라이버시 파라미터 집합 $\{\epsilon, \delta, \sigma, T, q\}$에 대해, 각 모델에서 learning rate와 같은 하이퍼 파라미터를 찾아 최적의 성능을 갖는 결과를 기록하였다. 동일한 프라이버시 예산 내에서 노이즈 스케일이 증가할수록 모델의 성능이 증가하였으며, 동일

Table 1. Final accuracies of differentially private models ($\delta = 10^{-3}$)

ϵ \ σ	2	4	6	8
2	-	52% train accuracy	56% train accuracy	84% train accuracy
		46% test accuracy	42% test accuracy	69% test accuracy
6	96% train accuracy	98% train accuracy	98% train accuracy	96% train accuracy
	85% test accuracy	88% test accuracy	91% test accuracy	85% test accuracy
10	99% train accuracy	99% train accuracy	99% train accuracy	99% train accuracy
	95% test accuracy	95% test accuracy	91% test accuracy	94% test accuracy

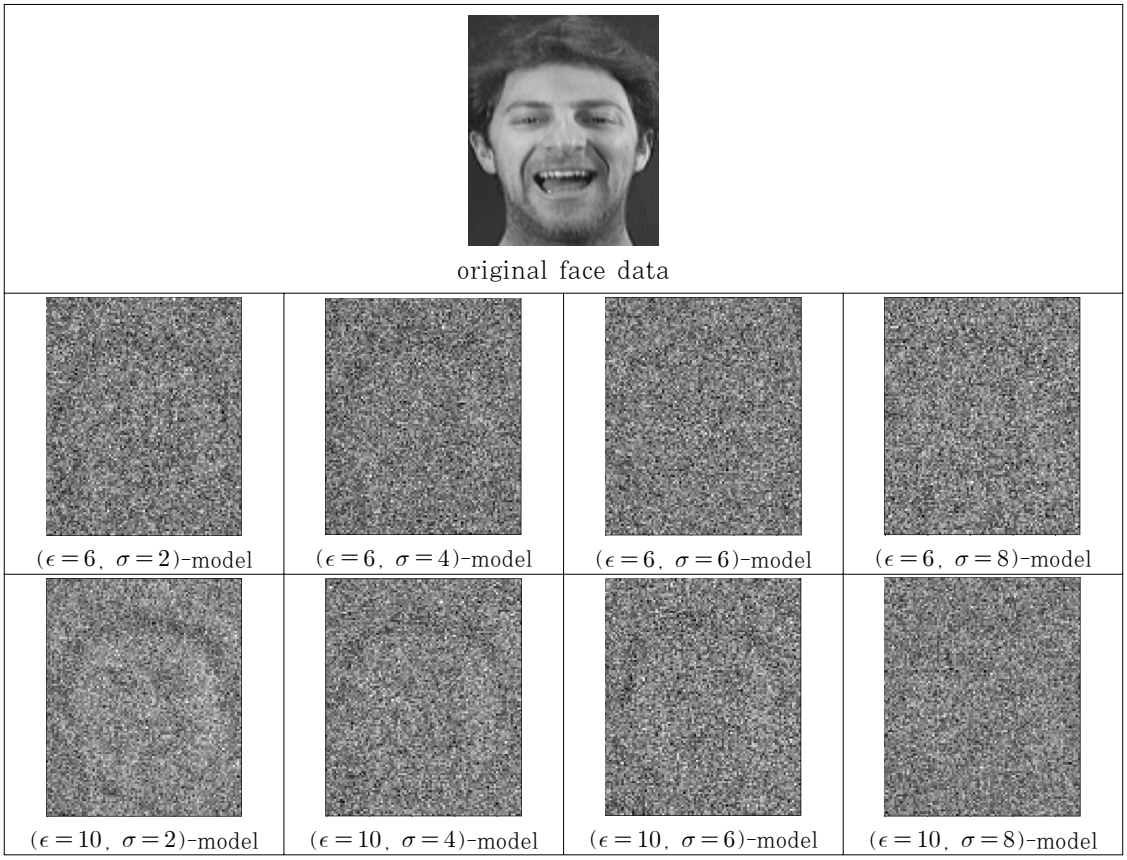

Fig. 6. Model inversion attack results on differentially private face recognition models

한 노이즈 스케일에 대해 프라이버시 예산이 증가할수록 모델의 성능이 증가함을 알 수 있다. 프라이버시 예산과 노이즈 스케일이 매우 작은 경우 학습이 수행되지 않고 종료된다.

3.2.2 차분 프라이버시의 모델 전도 공격 저항성

차분 프라이버시를 보장하는 얼굴 인식 모델에 Algorithm 3을 적용하여 모델 전도 공격에 대한 차분 프라이버시의 저항성 분석한다. Algorithm 3에 대한 파라미터는 Fredrikson 등[5]과 마찬가지로 $\alpha=5000$, $\beta=100$, $\gamma=0.99$, $\lambda=0.1$로 설정한다.

Fig. 6은 차분 프라이버시를 보장하는 얼굴 인식 모델에 대한 모델 전도 공격 결과의 예시를 나타낸다. 프라이버시를 고려하지 않은 상황에 대한 원활한 비교를 위해 Fig. 5의 이미지와 동일한 클래스에 대한 결과를 나타냈다. 모든 노이즈 스케일에 대하여, 프라이버시 예산 ϵ이 6 이하인 모델은 모델 전도 공격에 저항성을 갖는다. 그러나 $\epsilon=10$인 모델의 경우 노이즈 스케일이 작을 때, 얼굴 정보가 파악될 위험이 있으며 프라이버시가 침해될 수 있음을 알 수 있다. 전반적으로 모든 클래스에 대해 Fig. 6과 같은 결과가 도출되었으며 $\epsilon=10$인 경우 대부분의 복구된 이미지에서 윤곽과 같은 얼굴 정보 및 안경 테두리와 같은 정보가 파악되었다.

IV. 결 론

본 논문에서는 신경망 기반의 얼굴 인식 모델에 대하여 모델 전도 공격에 저항성을 가질 수 있는 차분 프라이버시 적용 방안에 대한 연구를 수행하였다. 기존에 제안된 차분 프라이버시를 보장하는 확률적 경사 하강법이 적당한 프라이버시 예산 내에서 충분히 학습되지 못하는 문제점을 파악하였으며 momentum 학습법으로의 확장을 통해 학습이 원

활히 수행되도록 하였다.

또한, 차분 프라이버시가 보장된 모델에 대해 모델 전도 공격을 수행하였으며, 실험을 통해 모델 전도 공격에 대한 프라이버시 예산/노이즈 스케일의 관계를 확인하였다. 그 결과, 6 이하의 차분 프라이버시 예산 내에서 모델 전도 공격에 저항성을 가질 수 있다는 것을 확인하였으며, 프라이버시 예산이 비교적 큰 10의 경우 작은 노이즈 스케일의 모델에서 모델 전도 공격에 대한 위험성을 확인하였다.

기존의 모델 전도 공격 [5]의 경우(프라이버시를 고려하지 않은 모델에 대한 모델 전도 공격) 복구된 이미지의 인식 가능성(식별성)은 Amazon의 Mechanical Turk survey(5개의 후보 이미지 중, 모델 전도 공격을 통해 복구된 이미지와 동일한 클래스를 찾는 조사)를 통해 수행되었다. 향후 차분 프라이버시를 만족하는 모델로부터 모델 전도 공격을 통해 복구된 이미지에 대한 인식 가능성을 평가하는 연구가 필요하며, 이는 차분 프라이버시의 실제 적용에 도움이 될 것으로 판단된다.

모델 전도 공격뿐만 아니라, membership inference 공격, model extraction 공격 등 신경망 기반의 기계학습 모델에 대한 다양한 프라이버시 침해 공격이 존재하며, 이러한 공격들에 대해 프라이버시를 보존할 수 있는 기술에 대한 연구가 필요하다.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," In Proceedings of the IEEE international conference on computer vision, pp. 1026-1034, Dec. 2015.

[2] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," In Advances in neural information processing systems, pp. 2773-2781, Dec, 2015.

[3] C. J. Maddison, A. Huang, I. Sutskever, and D. Silver, "Move evaluation in Go using deep convolutional neural networks," arXiv preprint arXiv:1412.6564, 2014.

[4] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," In 23rd USENIX Security Symposium, pp. 17-32, Aug. 2014.

[5] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333, Oct, 2015.

[6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," In 2017 IEEE Symposium on Security and Privacy, pp. 3-18, May, 2017.

[7] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," In 25th USENIX Security Symposium, pp. 601-618, Jun, 2016.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," In Theory of cryptography conference, pp. 265-284. March, 2006.

[9] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends® in Theoretical Computer Science, 9(3-4), 211-407. 2014.

[10] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," In Proceedings of the 2016 ACM SIGSAC Conference on

Computer and Communications Security, pp. 308-318, 2016, October.
[11] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," In Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp. 486-503, May, 2006.
[12] C. Dwork and J. Lei, "Differential privacy and robust statistics," In STOC, Vol. 9, pp. 371-380, May, 2009.
[13] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," In 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, pp. 51-60, October, 2010.
[14] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," IEEE Transactions on Information Theory, 63(6), 4037-4049, 2017.
[15] A. Beimel, S. P. Kasiviswanathan, and K. Nissim, "Bounds on the sample complexity for private learning and private data release," In Theory of Cryptography Conference, pp. 437-454, February, 2010.
[16] AT&T Laboratories Cambridge. The ORL database of faces. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. 2019. 01.

〈저자소개〉

박 철 희 (Cheolhee Park) 정회원
2014년 2월: 공주대학교 응용수학과 학사 졸업
2017년 2월: 공주대학교 수학과 석사
2017년 3월~현재: 공주대학교 수학과 박사 재학
〈관심분야〉 암호모듈구현, 데이터 보안기술, 차분 프라이버시 보호기술, 인공지능 보안기술

홍 도 원 (Dowon Hong) 종신회원
1994년 2월: 고려대학교 수학과 학사
2000년 2월: 고려대학교 수학과 박사
2000년 4월~2012년 2월: 한국전자통신연구원 팀장, 책임연구원
2012년 3월~현재: 공주대학교 응용수학과 교수
〈관심분야〉 암호기술, 데이터 보안기술, 차분 프라이버시 보호기술