프라이버시를 보호하는 인공지능 기법에 대한 조사

황성진, 양현종
포항공과대학교

sjh1753@g.postech.edu, hyunyang@g.postech.edu

A survey on privacy-preserving AI methods

Hwang Seong Jin, Yang Hyun Jong
POSTECH

요 약

최근 대규모 인공지능 모델의 등장으로 방대한 양의 데이터 확보가 중요해지고 있다. 다만, 특정 서비스의 사용자에게 데이터를 수집하는 과정에서 정보 제공자의 민감한 개인정보가 포함될 수 있다. 이는 모델의 학습 및 배포 과정에서 학습 데이터의 개인정보가 노출될 가능성을 만들고, 개인정보 보호기법이 적용된 학습 모델의 요구로 이어진다. 본 논문은 학습 과정에 적용되는 개인정보 보호기법인 연합학습, 동형암호, 차분 프라이버시를 살핀다. 또한, 해당 기법의 분석으로 기술 간의 한계점을 살피고, 보안 기술의 필요성을 재고하고자 한다.

Ⅰ. 서 론

최근 대규모 인공지능 모델의 등장으로 학습 데이터의 수집 경로가 다양해지고 양도 커지고 있다. 일례로 Google Brain 이 발표한 자연어 처리 모델인 PaLM 은 파라미터 수가 5400 억이 넘고, 학습 데이터 또한 7800 억의 토큰화 된 단어들로 구성돼 있다. 이는 대표적인 대규모 자연어 처리 모델인 GPT-3 모델과 비교했을 때 데이터 양은 1.95 배, 수집 방법 또한 소셜 미디어 대화, 웹 페이지, 코드 등으로 더 다양화됐다.[1]

하지만, 학습 데이터에는 민감한 개인정보가 포함될 가능성이 존재한다. 사람들은 특정 서비스의 사용하면서 그 정보를 제공하기 때문에 데이터에 개인정보가 포함될 수 있다. 학습 데이터는 인공지능의 학습 전반에 걸쳐 사용되며, 악의적인 사용자에 의해 학습 데이터가 추론될 수 있다. 따라서, 정보 제공자는 데이터를 제공하기 꺼리게 돼 프라이버시가 보호가 보장되는 학습 모델을 요구하게 된다. 실제로 자연어 처리 모델인 GPT-2 모델에 쿼리를 통해 학습에 사용된 데이터 속 개인정보를 드러내는 방법이 제시됐다.[2] 이 방법으로 개인의 이름, 이메일 주소, 핸드폰 및 팩스 번호와 거주지 주소가 여실없이 드러났다. 이에, 학습 데이터의 프라이버시를 보호해 정보 제공자를 안심시키고, 연구자들에게 민감한 개인정보를 포함한 데이터도 학습에 사용할 수 있도록 하는 프라이버시 보호 기법이 필요한 실정이다.

본 논문은 프라이버시 보호기법인 연합학습, 동형암호, 차분 프라이버시를 간단히 살핀다. 또한, 프라이버시 보호를 위한 기법에 요구되는 특징을 논하면서, 기술을 적용했음에도 프라이버시가 노출될 수 있는 시나리오를 언급한다. 나아가 기존의 프라이버시 보호 기법을 재고함으로 써, 새로운 프라이버시 보호 기법의 필요성을 강조한다.

Ⅱ. 본론

연합학습은 원격 기기나 분리된 데이터 센터에 데이터를 국한시킨 상태로, 중앙 서버의 글로벌 확률 모델을 학습하는 방법을 말한다.[3] 하나의 서버에서 학습하는 방식은 프라이버시가 노출될 가능성이 있기 때문이다. 학습 모델의 업데이트를 위해 경사도 값을 중앙 서버에 보내는 방식 등을 취한다. 다만, 학습 과정의 모든 경사도 정보들은 민감한 개인정보를 드러낼 수 있다.[4] 따라서, 중앙 서버가 클라이언트로부터 값을 받을 때도, 클라이언트의 정보를 감추는 안전한 다자간 통신(Secure multiparty computing, SMC) 기법이나, 동형암호, 차분 프라이버시 기법이 고려되고 있다. 하지만, Non-IID 의 데이터에 의한 불안정한 학습과 클라이언트 기기와 중앙 서버와의 통신 오버헤드 문제들이 산재해 있다.

동형암호는 암호화한 후의 연산의 결과가 암호화 과정과 연산의 순서가 뒤 바뀌더라도 동일한 결과를 보장하는 암호화 방법을 뜻한다.[5] 학습 과정에서는 암호화 데이터로 학습한 모델의 결과를 복호화 했을 때의 성능과 일반적인 학습을 진행했을 때의 성능이 동일함을 보장한다. 연합학습과 함께 서버로 암호화된 데이터를 주고받는 과정에 적용되기도 한다. 하지만, 데이터의 규모가 클수록 연산 복잡도가 커져 학습 시간이 느려진다는 단점이 존재한다. 또한, 변환 결과 데이터의 양이 커져 메모리에 부담을 줄 가능성이 있다.

이밖에, 변환 연산을 아는 관리자에게는 프라이버시가 보장되지 않는다. 신용할 수 있는 관리자는 이상적으로 존재하지 않는다. 이에, 프라이버시가 노출될 잠재적인 위험이 항상 존재한다.

활용한 완화된 차분 프라이버시 종류	프라이버시 손실 ϵ' 추정 값
(ϵ, δ) 차분 프라이버시[6]	For all $\epsilon, \delta, \delta' \geq 0$, $\epsilon' = \epsilon\sqrt{2k\ln\left(\frac{1}{\delta'}\right)} + k\epsilon(e^{\epsilon} - 1)$
Rényi 차분 프라이버시[8]	For $0 < \delta < 1$ $such\ that \log\left(\frac{1}{\delta}\right) \geq \epsilon^2 k, \epsilon' = 4\epsilon\sqrt{2k\log(\frac{1}{\delta})}$

Table 1. 각, (ϵ, δ) 및 ϵ 프라이버시의 k번째 합성 과정에서의 프라이버시 손실 ϵ' 추정 값

차분 프라이버시는 데이터를 제공한 개인의 참여 여부를 숨기고, 데이터의 통계적인 의미는 보존하는 프라이버시 보호 기법이다.[6] 구체적으로, 경사하강법의 매 과정마다 경사도에 차분 프라이버시가 보호되도록 노이즈를 가하는 방식이 보편적으로 채택되고 있다.[7] 이는 회원 추론 공격(Membership inference attack) 및 전도 공격(inversion attack)을 막기 위한 사실상 표준으로 자리잡았다.

차분 프라이버시를 경사도에 적용한 기법은 학습 과정이 길어지면 프라이버시 손실이 누적되어[7], 개인의 참여 유무에 따른 확률 분포 차가 명확하게 돼 프라이버시가 드러날 수 있는 가능성이 존재한다.

따라서, 누적되는 프라이버시 손실을 최대한 줄이고자, 기존의 정의에서 파생된 차분 프라이버시가 연구되고 있다. Table 1 은 완화된 차분 프라이버시의 정의를 이용하여, 누적되는 프라이버시 손실 값을 추정한 것으로 아래로 갈수록 타이트한 손실 값을 나타낸다.

종합하면, 차분 프라이버시의 주 특성은 프라이버시 손실인 ϵ으로 프라이버시 노출 정도를 수치화 할 수 있다는 점이다. 다만, 차분 프라이버시 적용과정에서 더해지는 노이즈로 인해 원본 성능이 보장되지 않는 단점이 존재한다. 노이즈의 분포에 따라 모델의 성능과 개인정보 보호 정도 간의 트레이드 오프 관계가 형성되기 때문이다. 따라서, 상황에 마다 적절하다고 판단되는 ϵ 값을 사용자가 직접 조정하는 인위적인 방법이 채택되고 있다.

현재 프라이버시 보호가 필요한 상황에서 서버의 데이터 저장 공간, 신용 가능한 관리자의 유무, 보호 기법 적용 전 후의 학습 정확도, 등을 고려해 융합된 프라이버시 보호 기법을 적용하는 것이 현실적인 대안이다.

하지만, 연합학습, 동형암호, 차분 프라이버시 모두 최악의 경우 프라이버시가 드러나는 상황이 존재할 수 있다. 따라서, 해당 문제점들을 보완한 새로운 프라이버시 보호 기법의 탐구가 필요하다.

Ⅲ. 결론

본 논문은 최근 인공지능 학습 전반에 적용되는 프라이버시 보호 기법에 대해 논하였다. 문제 상황에 따라 연합학습, 동형암호, 차분 프라이버시를 고려해볼 수 있다. 다만, 연합학습, 동형암호와 차분 프라이버시는 최악의 경우 프라이버시가 노출될 수 있어, 이들을 융합한 기법의 적용이나, 이들을 모두 보완한 새로운 보호 기법이 필요하다.

참 고 문 헌

[1] Chowdhery, A., Narang, S., Devlin, J. "PaLM: Scaling Language Modeling with Pathways", arXiv:2204.02311, April 2022.

[2] Calini, N, "Extracting Training Data from Large Language Models", USENIX Security Symposium. 2021

[3] Tian, L. "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Processing Magazine, 2020.

[4] McMahan, H, B. “Learning differentially private recurrent language models”, in *Proc*. Int. Conf. Learning Representations, 2018.

[5] Acar, A. “A survey on homomorphic encryption schemes: Theory and implementation”, ACM Computing Surveys 51, July 2018.

[6] Dwork, C. Roth, A. “The Algorithmic Foundations of Differential Privacy”, Foundations and TrendsR in Theoretical Computer Science, pp 211-407, 2013.

[7] Abadi, M. “Deep learning with differential privacy,” in CCS, 2016.

[8] Mironov, I. “Rényi differential privacy”, in Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF), pp. 263-275, Aug. 2017.