Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.35, NO.1, Feb. 2025 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2025.35.1.99

연합학습에서 약한 차분 프라이버시와 다양한 방어 기법 결합을 통한 백도어 공격 방어 연구*

최 민 영,[1†] 김 현 일[2‡]
[1,2]조선대학교 (학생, 교수)

A Study on the Effectiveness of Backdoor Attack Defense through the Combination of Weak Differential Privacy and Various Defense Mechanisms in Federated Learning*

Minyeong Choe,[1†] Hyunil Kim[2‡]
[1,2]Chosun University (Student, Professor)

요 약

프라이버시 규제가 강화됨에 따라 사용자 데이터를 중앙 서버에 저장하지 않고 분산된 환경에서 학습할 수 있는 연합학습(Federated learning)이 중요한 기술로 주목받고 있다. 그러나 연합학습은 분산된 특성상 다양한 공격에 노출되기 쉬우며, 그 중 특히 백도어 공격에 취약하다. 본 논문에서는 이러한 백도어 공격을 방어하기 위해 약한 차분 프라이버시(Weak differential privacy) 기법을 적용하고, Multi-Krum, Norm Clipping과 같은 다양한 방어 기법과 결합하여 그 효과를 분석하였다. Reddit 데이터셋과 LSTM, GPT-2 모델을 활용해 다양한 크기의 노이즈를 적용한 실험을 통해 차분 프라이버시의 방어 효과를 평가한 결과, 프라이버시 보호를 위해 필요한 수준보다 적은 양의 노이즈로도 효과적인 방어가 가능함을 확인하였고, 다른 방어 기법과 결합할 때 방어 성능이 더욱 향상됨을 발견하였다. 해당 연구는 연합학습 환경에서 모델 성능을 유지하면서도 백도어 공격을 방어할 수 있는 최적의 노이즈 설정이 중요함을 강조하며, 차분 프라이버시와 다양한 방어 기법의 결합이 백도어 방어에 실용적임을 시사한다.

ABSTRACT

As privacy regulations become stricter, federated learning, which enables learning in a distributed environment without storing user data on a central server, has emerged as a critical technology. However, its distributed nature makes it vulnerable to various attacks, with backdoor attacks being particularly significant. In this paper, we aim to defend against such backdoor attacks by applying a weak differential privacy mechanism and combining it with various defense techniques such as Multi-Krum and Norm Clipping to analyze their effectiveness. Experiments conducted using the Reddit dataset and LSTM and GPT-2 models with varying levels of noise demonstrate that effective defense can be achieved with less noise than typically required for privacy protection. Moreover, combining differential privacy with other defense techniques enhances defense performance. This study highlights the importance of optimizing noise settings to defend against backdoor attacks while maintaining model performance in federated learning environments and suggests that the combination of differential privacy with other defense mechanisms is a practical solution for backdoor defense.

Keywords: Backdoor attack, Federated learning, Weak differential privacy

Received(11. 13. 2024), Modified(12. 09. 2024), Accepted(12. 09. 2024)

* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2024-00398353, 생성형 AI 보안 위협 대응 기술 개발)

* 본 논문은 2024년도 한국정보보호학회 호남지부 추계학술대회에 발표한 우수논문을 개선 및 확장한 것임.

† 주저자, minyeong@chosun.ac.kr

‡ 교신저자, hyunil@chosun.ac.kr(Corresponding author)

I. 서 론

연합학습(federated learning)은 분산 환경에서 개인 데이터의 프라이버시를 보호하면서 딥러닝 모델을 학습할 수 있는 방식으로 주목받고 있다[1]. 이는 GDPR(general data protection regulation)[2] 및 CCPA(california consumer privacy act)[3]와 같은 프라이버시 규제에 대한 기술적 대응이 가능함을 의미하며, 실제 데이터를 직접 공유하지 않고 로컬 모델의 파라미터만 공유함으로써 궁극적으로 개인 정보를 보호하게 된다. 이러한 특징 덕분에 연합학습은 구글의 Gboard[4], 헬스케어 시스템[5], 자율주행 차량[6] 등 다수의 실제 응용 분야에 성공적으로 적용되어왔다.

그러나 연합학습은 분산된 특성 때문에 다수의 보안적 위협이 존재한다. 특히, 악의적인 사용자가 백도어가 주입된 모델 파라미터를 파라미터 서버에 업로드함으로써 글로벌 모델을 감염시켜 특정 데이터에 대한 오분류를 일으키는 백도어 공격[7]은 시스템 전반의 신뢰성과 보안을 위협하는 심각한 문제로 대두된다. 백도어란 정상적인 모델 학습 과정에서 악의적인 의도를 가진 특정 트리거를 삽입하여, 이 트리거가 포함된 입력에 대해 의도적으로 잘못된 출력을 유도하도록 설계된 공격 방법을 의미한다. 이러한 백도어 공격은 악의적인 클라이언트가 감염된 파라미터를 공유할 때 예기치 않은 결과를 초래한다.

예를 들어, 자율주행 차량의 객체 인식 모델에 백도어가 주입된 경우, 공격자는 특정 위치에 작은 스티커와 같은 트리거를 추가하여 차량이 표지판을 잘못 인식하게 하거나, 보행자를 다른 객체로 분류하도록 유도할 수 있다[8]. 이와 같은 백도어 공격은 트리거 입력에 대해서만 작동하므로 일상적인 입력에는 영향을 미치지 않아 모델의 학습 과정에서 쉽게 감지되지 않는 특징이 있다.

본 연구에서는 이러한 연합학습에서의 백도어 취약점을 해결하기 위해 파라미터 서버에 약한 차분 프라이버시(weak differential privacy) 방어 기법[9]을 적용하고, 그 효과를 분석한다. 구체적으로, 차분 프라이버시(differential privacy) 기반[10] 노이즈 양의 변화에 따른 백도어 공격의 정확도 감소와 모델의 전반적인 성능 유지 정도를 실험을 통해 검증하였다. 이를 통해 백도어 공격 방어 효과와 모델 성능 간의 상관관계를 분석한다.

추가적으로, 다양한 방어 기법과의 결합을 통해 방어 효과를 비교 분석하는 접근을 제안한다. Multi-Krum[11], Norm Clipping[9]과 같은 대표적인 백도어 방어 기법들과 차분 프라이버시 기법을 결합하여 그 성능을 평가하고, FLAME[12]처럼 노이즈를 삽입하는 방어 기법의 경우 노이즈 크기를 조절하여 방어 효과를 비교 분석하였다. 이러한 결합 방어 전략이 백도어 공격 방어에 미치는 영향을 실험적으로 비교한다. 실험은 자연어 처리 작업을 수행하는 연합학습 환경을 가정하며 LSTM(long short-term memory) 모델[13]과 GPT-2(generative pre-trained transformer-2) 모델[14]을 사용해 진행하였다. 각 모델에서의 방어 기법 성능을 분석함으로써 다양한 환경에서 백도어 공격 방어 효과를 종합적으로 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 연합학습, 백도어 공격, 그리고 백도어 공격 방어 및 차분 프라이버시에 대한 이론적 배경을 설명한다. 3장에서는 약한 차분 프라이버시를 활용한 연합학습의 백도어 공격 방어 메커니즘을 상세히 기술하며 다양한 방어 기법과 약한 차분 프라이버시를 결합한 방어 기법을 제안한다. 4장에서는 다양한 노이즈 σ값에 따른 백도어 공격 방어 및 모델 성능에 미치는 영향을 실험 결과를 통해 분석한다. 마지막으로 5장에서는 연구의 결론과 향후 연구 방향을 제시한다.

II. 배경 및 관련 연구

2.1 연합학습(Federated Learning)

연합학습은 개인 프라이버시를 보장하는 분산 AI 학습 방식으로, 각 클라이언트가 개인 데이터를 보유하는 non-i.i.d.(non independent and identically distributed) 환경을 가정한다. Fig. 1.과 같이 각 클라이언트는 자신의 데이터를 사용하여 로컬에서 학습하고 모델의 업데이트 값만을 파라미터 서버에게 전송하는 방식으로 이루어진다[1]. 구체적으로 연합학습은 T 라운드에 걸쳐 글로벌 모델을 학습한다. 각 라운드에서 전체 K명의 클라이언트 중 무작위로 선택된 k명의 클라이언트가 참여한다. 매 라운드에서 선택된 클라이언트들은 현재의 글로벌 모델을 받아 자신들의 로컬 데이터셋을 사용하여 여러 번 학습시킨 로컬 모델을 생성한다. 이후 클라이언트는 자신의 로컬 모델과 글로벌 모델 간의 차이 값인 업데이트 값만 파라미터 서버에 전송한다.

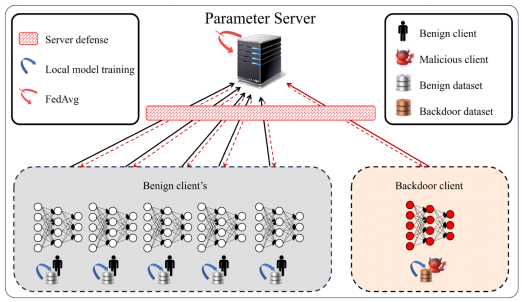

Fig. 1. Overview of federated learning

파라미터 서버는 FedAvg(federated averaging) 알고리즘을 통해 클라이언트로부터 받은 업데이트 값을 평균화하여 글로벌 모델을 업데이트하며, 이 과정은 다음과 같이 계산된다.

$$w_{t+1} = \sum_{k=1}^{K} w_{t+1}^{k} \tag{1}$$

수식 1에서 $w_{t+1}^{k} \leftarrow w_t - \eta g_k$이며, $g_k = \nabla F_k(w)$는 각 클라이언트 k에 대한 그래디언트이다. 해당 시나리오에서 각 클라이언트는 분산 최적화 문제를 다음과 같이 해결한다.

$$f(w) = \sum_{k=1}^{K} \frac{n_k}{n} \cdot F_k(w) \tag{2}$$

수식 2에서 로컬 목적함수 $F_k(w) = \frac{1}{n_k} \Sigma_{i \in P_k} \ell(x_i, y_i; w)$이다. 여기서 ℓ은 크로스 엔트로피와 같은 손실함수이며, η는 학습률, P_k는 클라이언트 k에 대한 학습 데이터셋을, $|P_k| = n_k$는 해당 데이터셋의 크기를 의미한다.

2.2 백도어 공격

연합학습은 프라이버시를 보호할 수 있다는 장점이 있지만, 분산된 특성으로 인해 백도어에 취약하다. 백도어 공격은 악의적인 클라이언트가 특정 트리거를 포함한 데이터로 로컬 모델을 학습시켜, 글로벌 모델에서 트리거 입력에 대해 의도된 오작동을 유발하도록 하는 공격 방식이다. 이러한 공격은 모델의 일반적인 성능에는 영향을 미치지 않으면서도, 특정 입력에 대해 공격자가 원하는 비정상적인 예측 결과를 도출하도록 모델을 조작할 수 있어 탐지가 어렵고 위협적인 특징을 갖는다[7]. 이러한 연합학습 환경에서의 백도어 공격에 대한 연구는 지속적으로 발전하고 있으며, 이 중 몇 가지 대표적인 기법을 소개한다.

Constrain-and-scale[7]은 백도어를 삽입할 때 정상 데이터와 백도어 데이터 모두에서 높은 정확도를 유지하고, 이상 탐지 시스템을 효과적으로 회피하도록 설계된 방법이다. 이 공격은 모델이 정상적인 학습 성능을 유지하는 동시에 백도어 데이터에 대해 특정 동작을 수행하도록 손실 함수를 구성하여, 삽입된 백도어가 탐지되지 않도록 한다.

Neurotoxin[15]은 백도어 효과가 주입이 중단된 후에도 지속되도록 설계되었으며, 그 핵심 아이디어는 공격자가 정상 클라이언트가 업데이트하지 않을 가능성이 높은 그래디언트 좌표만 수정하여, 기존 백도어 공격보다 더 지속적인 백도어 효과를 유지한다.

DBA[16]는 연합학습 환경에서 기존 중앙 집중형 백도어 공격보다 탐지가 어려우면서도 더욱 효과적인 공격 방법으로, 글로벌 트리거를 여러 로컬 트리거로 분할하여 각 악성 참여자의 데이터에 분산적으로 삽입함으로써 백도어 탐지를 어렵게 하고 공격의 지속성을 높인다.

IBA[17]는 최신 백도어 공격 기술로, 시각적으로 탐지하기 어려운 트리거를 설계하고 모델의 안정적인 파라미터를 선택적으로 오염시켜 백도어 효과의 지속성을 강화하였다. 또한, 오염된 업데이트를 국소적인 범위로 제한함으로써 모델 집계 과정에서 탐지될 가능성을 최소화하였다.

이와 같은 백도어 공격은 탐지하기 매우 어려워 연합학습 시스템의 신뢰성을 저하시킬 수 있으며, 이에 대한 강력한 탐지 및 방어 기법의 개발이 요구된다.

2.3 백도어 공격 방어 및 차분 프라이버시

연합학습은 데이터를 서버에 전송하지 않고도 분산된 모델 학습을 가능하게 하여 프라이버시 보호 측면에서 장점을 제공한다. 그러나 이러한 장점에도 불구하고, 연합학습 환경은 여전히 다양한 보안 위협에 취약하며, 이를 방어하기 위한 여러 메커니즘이 제안되었다.

Multi-Krum[11]은 분산 학습 아키텍처에서

Byzantine 공격[18]에 대한 효과적인 방어 메커니즘으로 제안되었다. Multi-Krum은 여러 신뢰할 수 있는 클라이언트의 업데이트를 선택하고 이를 집계함으로써 악의적인 업데이트가 글로벌 모델에 주입되는 것을 방지한다.

Norm Clipping[9]은 악의적인 클라이언트의 업데이트 Norm이 큰 경우가 많다는 점에 착안하여, 파라미터 서버로 전송되기 전에 업데이트 크기를 제한하기 위해 Norm bound를 설정하는 방식이다. 미리 정의된 Norm bound를 초과하는 업데이트를 클리핑함으로써, Norm Clipping은 단일 클라이언트의 업데이트로 인한 영향을 줄이고, 이를 통해 적대적이거나 비정상적인 업데이트로부터 발생할 수 있는 잠재적인 피해를 제한하여 글로벌 모델의 견고성을 보장한다.

FLAME[12]은 백도어를 효과적으로 제거하기 위해 충분한 양의 노이즈를 추정하고 주입하도록 설계된 방어 메커니즘이다. FLAME은 모델 클러스터링과 가중치 클리핑을 결합하여 필요한 노이즈 양을 최소화하는 방식으로 수행된다. 이를 통해 FLAME은 악의적인 백도어 삽입을 방지하면서 글로벌 모델의 무결성과 정확성을 유지한다.

이처럼 다양한 방어 기법들이 존재하며, 이러한 방어 기법들과 함께 사용할 수 있는 약한 차분 프라이버시를 설명하기에 앞서 기존 차분 프라이버시[10]에 대해 먼저 설명하고자 한다.

차분 프라이버시는 강력한 프라이버시 모델로, 공격자가 특정 개인(타겟)에 대한 배경지식(background knowledge)를 보유하고 있다고 가정해도 해당 데이터베이스에서 개인 레코드에 대한 프라이버시에는 어떠한 영향도 줄 수 없게끔 설계되었다[10]. 즉, 차분 프라이버시가 적용되면 공격자는 자신이 알고 있는 타겟의 배경지식보다 더 많은 정보를 얻을 수 없다는 것을 나타낸다. 상세한 내용에 대한 설명은 3.1절에서 다룬다.

III. 약한 차분 프라이버시를 활용한 연합학습 백도어 공격 방어

본 장에서는 연합학습 환경에서 백도어 공격을 방어하기 위한 차분 프라이버시 적용 방법을 설명한다. 기존 차분 프라이버시 기법은 주로 데이터베이스에서 프라이버시 보호를 위해 다량의 노이즈를 삽입하는 방식으로 설계되었으나, 이러한 기법을 연합학습 환경에 그대로 적용할 경우 과도한 노이즈로 인해 모델 성능이 크게 저하될 위험이 있다. 이를 해결하기 위해, 3.1절에서는 연합학습에서 차분 프라이버시 적용 방안을 다루고, 3.2절에서는 백도어 공격 방어 효과를 유지하면서 모델 성능 저하를 최소화할 수 있는 약한 차분 프라이버시 기법을 설명하며, 이를 다양한 방어 기법과 결합하여 비교 분석하는 방법을 제안한다.

3.1 연합학습에 적용된 차분 프라이버시

차분 프라이버시를 구체적으로 이해하기에 앞서 이웃 데이터베이스에 대한 이해가 필요하다. 이웃 데이터베이스는 데이터 전체 집합 N^X 상에서 두 데이터베이스 $D, D' \in N^X$에 대해 l_1-norm $\| D \|_1$은 데이터베이스의 크기를 의미하며 이는 레코드 수를 의미한다. 이때 $\| D \Delta D' \|_1$은 두 데이터베이스의 레코드 수 차이를 의미하며 $\| D \Delta D' \|_1 \leq 1$를 만족할 시 두 데이터베이스 D, D'를 서로 이웃 데이터베이스라고 한다.

정의 1) (ϵ, δ)-차분 프라이버시는 모든 사건 $S \subseteq Range(M)$에 대해 이웃 데이터베이스인 $D, D' \in N^X$를 입력 값으로 무작위성을 갖는 메커니즘 $M: N^X \rightarrow Range(M)$이 다음과 같은 식을 만족할 때 M은 (ϵ, δ)-차분 프라이버시를 만족한다[10].

$$\Pr[M(D) \in S] \leq e^{\epsilon} \times \Pr[M(D') \in S] + \delta \quad (3)$$

수식 3에서 $\delta = 0$이면 ϵ-차분 프라이버시라 한다. ϵ값이 작아질수록 두 응답 값이 사건 S에 속할 확률 값 차이가 거의 없어 해당 응답 값으로 두 데이터베이스를 구분하기 매우 어려워진다. 즉, 해당 질의에 대해서 특정 개인정보가 드러나지 않을 만큼의 무작위성을 갖는 응답 값을 반환하는 메커니즘을 제공한다.

이러한 차분 프라이버시를 딥러닝 모델, 더 나아가 연합학습에 적용하기 위해 확률적 경사 하강법과 결합한 DP-SGD(differential privacy - stochastic gradient descent) 기법이 주로 사용된다[19].

DP-SGD는 각 학습 단계에서 그래디언트에 노이

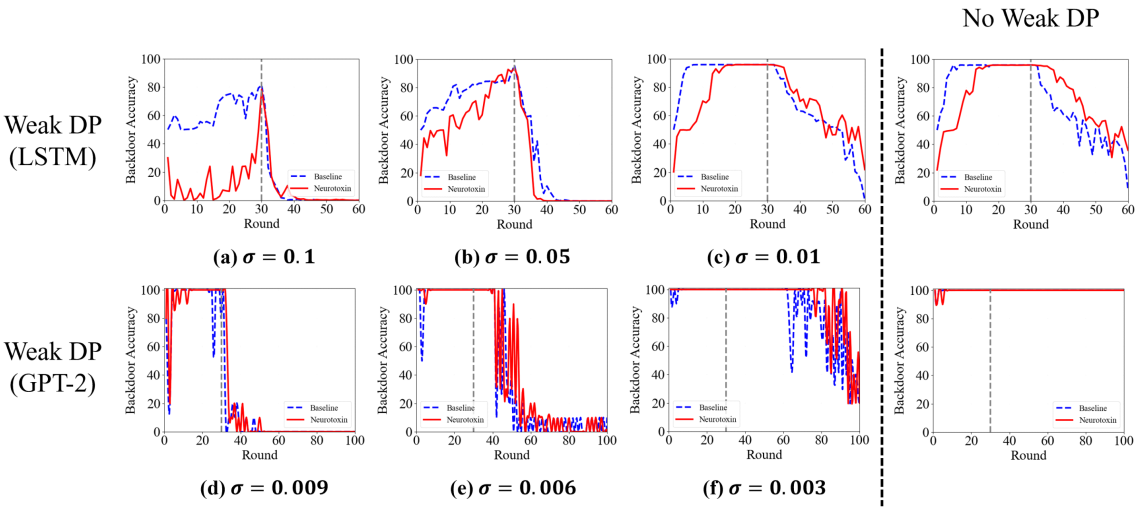

Fig. 2. Backdoor accuracy after applying weak differential privacy

즈를 추가하여 개인 데이터가 모델 업데이트 과정에서 드러나는 것을 방지한다. 이때 노이즈는 다음 수식을 통해 결정된다.

$$\sigma = \frac{1}{\epsilon} \cdot \sqrt{2\ln\frac{1.25}{\delta}} \tag{4}$$

수식 4는 본 논문에서 사용될 차분 프라이버시를 만족하는 Gaussian mechanism[19]의 노이즈 σ를 정의한다. 여기서 ϵ과 δ는 프라이버시 보장 수준을 결정하는 파라미터이다.

이를 연합학습 환경에 적용하면, 특정 클라이언트 k의 그래디언트에 노이즈가 수식 5와 같이 삽입이 된다.

$$g_k = \nabla F_k(w) + N(0,\sigma^2 C^2 I) \tag{5}$$

여기서 $\nabla F_k(w)$는 k 클라이언트의 그래디언트를 나타내고, $N(0,\sigma^2 C^2 I)$는 Gaussian mechanism을 따르는 노이즈를 나타낸다.

3.2 약한 차분 프라이버시

약한 차분 프라이버시는 3.1절에서 설명한 연합학습에 적용된 차분 프라이버시의 노이즈 크기를 줄여, 프라이버시 보호보다는 보안 측면에서 요구되는 방어 효과를 달성하고자 하는 기법이다[9]. 기존 차분 프라이버시는 프라이버시 보호를 위해 과도한 노이즈를 삽입해 모델의 성능을 저하시킬 수 있으나, 약한 차분 프라이버시는 노이즈 크기를 줄임으로써 모델의 성능 손실을 최소화하면서도 백도어 공격에 대한 방어 효과를 보장한다.

본 연구에서는 백도어 공격 방어를 주요 목표로 하여, 노이즈의 양을 크게 줄인 약한 차분 프라이버시 기법을 사용하였다. Fig. 3.에서 설명된 바와 같이, 각 클라이언트는 개인 데이터로 학습하고 그래디언트를 계산한다. 계산된 그래디언트는 그래디언트

Algorithm 1 Federated Learning with DP

Input: client gradient g_k, gradient norm bound C, noise scale σ, optional defense mechanism D

1: **for** each client k **do**
2: **Compute client gradient**
3: $g_k \leftarrow \frac{1}{n_k} \Sigma_{i \in P_k} \ell(x_i, y_i; w)$
4: **Apply optional defense mechanisms**
5: $g_k \leftarrow D(g_k)$
6: **Clip gradient**
7: $g_k \leftarrow g_k / \max(1, \frac{\| g_k \|_2}{C})$
8: **Add noise**
9: $g_k^{dp} \leftarrow g_k + N(0,\sigma^2 C^2 I)$
10: **end for**

Output: noisy gradient g_k^{dp}

Fig. 3. Federated learning with differential privacy

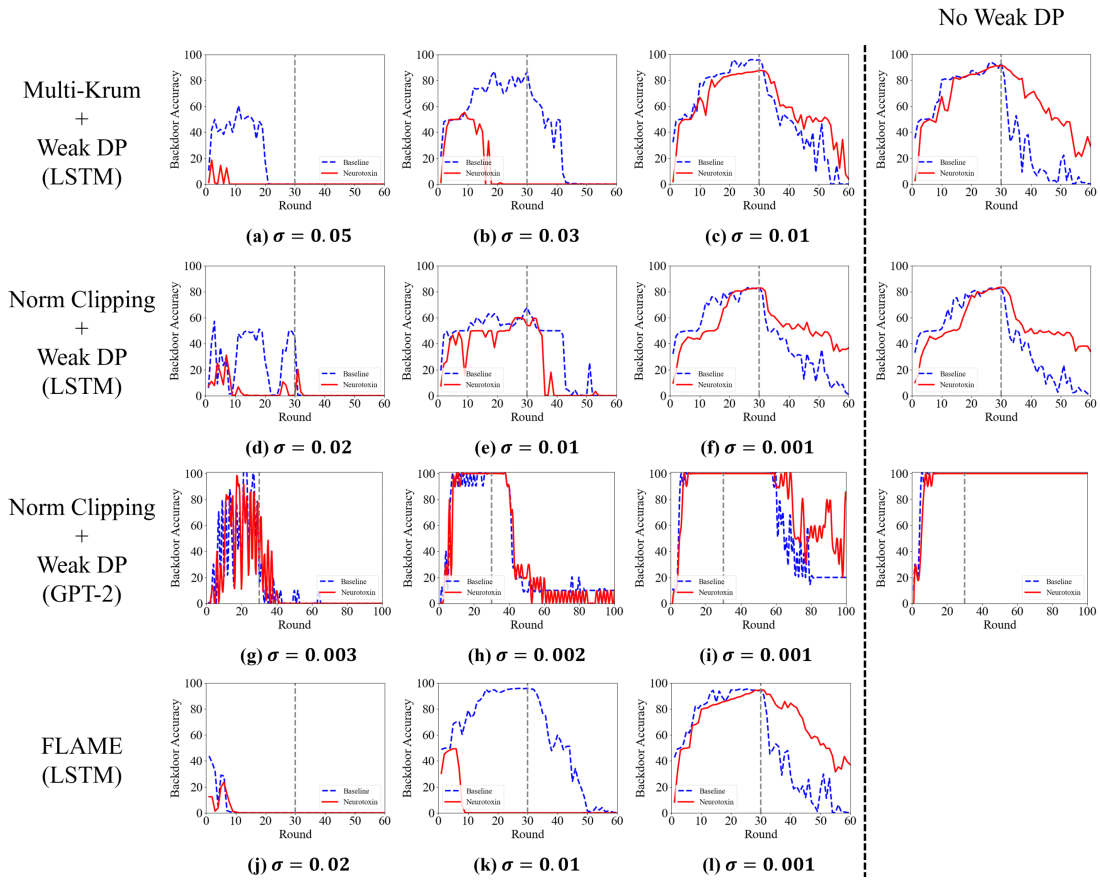

Fig. 4. Backdoor accuracy after combining various defense techniques with weak differential privacy

Norm bound C로 클리핑한 후 Gaussian mechanism을 따르는 소량의 노이즈를 추가하여 백도어 공격에 대한 방어 효과를 높이는 동시에 모델의 성능에 미치는 영향을 최소화하는 것을 목표로 한다.

또한, 다양한 방어 기법과 약한 차분 프라이버시를 결합하여 적용하는 방법을 제안하고, 이러한 조합이 백도어 공격 방어 효과와 모델의 성능에 미치는 영향을 비교 분석하고자 한다.

IV. 실험 결과 및 분석

4.1 실험 설정

본 연구에서는 자연어 처리 작업을 수행하는 연합학습 환경에서의 약한 차분 프라이버시 방어 기법을 구현하여 백도어 공격 시 모델에 미치는 영향을 분석하였다. 실험에는 다음 토큰 예측 작업에 사용되는 Reddit 데이터셋을 활용하였으며, LSTM 모델과 GPT-2 모델을 기반으로 하였다. 연합학습 환경은 매 라운드마다 총 8000명의 클라이언트 중 10명을 무작위로 선택해 학습에 참여시키는 방식으로 구성되었고, 공격자는 초기 30라운드 동안 계속 선택될 수 있는 환경을 가정하였다. 백도어 공격은 초기 30라운드 동안 주입되었으며, 이후에는 백도어 주입을 중단한 상태로 나머지 라운드에서 모델의 변화를 관찰하였다.

백도어 공격 기법으로는 Baseline과 Neurotoxin 두 가지를 적용하였다. Baseline 공격[8]은 악성 클라이언트가 백도어 데이터를 로컬 모델에 주입하고, 확률적 경사 하강법을 통해 훈련된 모델 파라미터를 파라미터 서버에 전송하여 백도어를 글로벌 모델에 포함시키는 기본적인 공격이다. 그러나 백도어 주입이 중단되면, 백도어가 글로벌 모델에서 빠르게 사라지는 한계를 가진다. Neurotoxin 공격[15]은 이를 개선하여 백도어 주입 중단 후에도 백도어의 지속성을 향

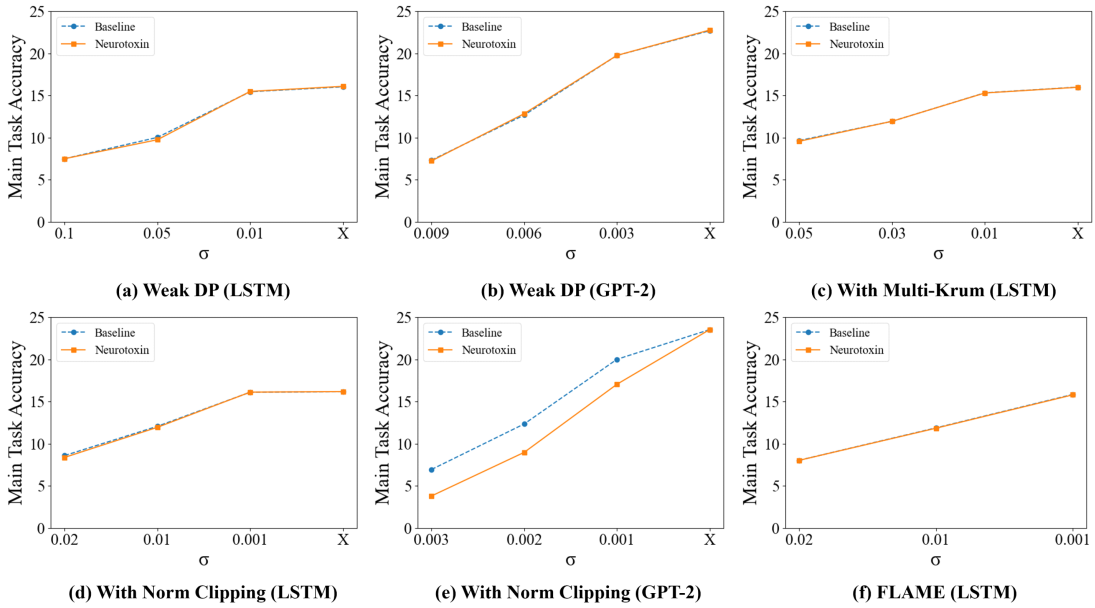

Fig. 5. Comparison of the average main task accuracy of the global model

상시킨 고도화된 백도어 공격이다.

방어 기법으로는 약한 차분 프라이버시만 적용하는 것과 Multi-Krum, Norm Clipping에 약한 차분 프라이버시를 결합한 방어 기법을 적용하였다. Table 1은 변화하는 σ값에 따른 모델의 정확도를 나타내며, Fig. 5.는 방어 기법에 따른 모델 정확도의 변화를 그래프로 시각화한 것이다. (a)와 (b)는 각각 LSTM과 GPT-2 모델에 약한 차분 프라이버시만을 적용했을 때의 정확도를 나타내며, (c)는 LSTM 모델에서 Multi-Krum과 약한 차분 프라이버시를 결합했을 때의 정확도를 보여준다. 이어서, (d)와 (e)는 각각 LSTM과 GPT-2 모델에서 Norm Clipping과 약한 차분 프라이버시를 결합하여 적용했을 때의 정확도를 나타내며, (f)는 LSTM 모델에 FLAME 방어 기법을 적용하면서 약한 차분 프라이버시 노이즈 양을 조정했을 때의 정확도를 보여준다.

보다 구체적으로, 약한 차분 프라이버시는 선택된 클라이언트들의 업데이트에 Table 1의 σ값만큼의 노이즈를 삽입하였으며, Norm Clipping은 LSTM 모델에는 3.0, GPT-2 모델에는 0.3의 Norm bound를 적용하여 클라이언트 업데이트의 크기를 제한하였다.

Norm bound 값은 각 모델의 구조적 특성과 학습 안정성을 고려하여 선정된 것으로, LSTM은 상

Table 1. Average main task accuracy of the global model

	C	σ	Baseline	Neuro toxin
Weak DP (LSTM)	1.0	0.1	7.49	7.49
		0.05	10.02	9.74
		0.01	15.44	15.48
		X	16.02	16.08
Weak DP (GPT-2)	1.0	0.009	7.32	7.23
		0.006	12.67	12.84
		0.003	19.75	19.75
		X	22.68	22.76
with Multi Krum (LSTM)	1.0	0.05	9.65	9.56
		0.03	11.93	11.93
		0.01	15.31	15.29
		X	15.99	15.97
with Norm Clipping (LSTM)	3.0	0.02	8.59	8.37
		0.01	12.06	11.94
		0.001	16.10	16.11
		X	16.16	16.17
with Norm Clipping (GPT-2)	0.3	0.003	6.94	3.80
		0.002	12.34	8.97
		0.001	20.00	17.04
		X	23.52	23.52
FLAME (LSTM)	S_t	0.02	8.03	8.02
		0.01	11.87	11.83
		0.001	15.84	15.79

대적으로 단순한 구조를 가지며 큰 그래디언트 변화를 허용하더라도 안정적으로 학습이 가능하므로 3.0과 같은 상대적으로 큰 값을 설정하였다. 반면, GPT-2는 방대한 파라미터를 포함하고 있어 작은 그래디언트 변화에도 민감하게 반응하므로, 학습의 안정성을 유지하기 위해 0.3과 같은 작은 값을 적용하였다.

FLAME 방어 기법은 자체적으로 적절한 노이즈를 삽입하므로, Fig. 4.에서 볼 수 있듯이 추가적인 약한 차분 프라이버시 없이 노이즈 크기를 조절하며 그 효과를 평가하였다. 따라서, FLAME의 특성상 노이즈가 없는 실험 결과는 포함되지 않는다.

4.2 실험 결과

Fig. 2.는 약한 차분 프라이버시 기법만을 적용했을 때의 백도어 공격 정확도를, Fig. 4.은 다양한 방어 기법과 약한 차분 프라이버시를 결합했을 때의 백도어 공격 정확도를 나타낸다. 그래프에서 파란 점선이 Baseline 공격을, 빨간 실선이 Neurotoxin 공격을 나타내어 두 공격 방식의 방어 효과를 비교할 수 있도록 하였다. 여기서 방어가 잘 이루어졌다는 것은 백도어 공격의 정확도가 낮아지는 것을 의미한다.

두 그림 모두에서 약한 차분 프라이버시 기법을 적용한 결과, 적절한 크기의 노이즈만으로도 효과적으로 백도어를 방어할 수 있었으며, 노이즈 σ값이 증가할수록 백도어 공격의 정확도가 크게 감소한다. 이는 노이즈 삽입량이 클수록 백도어 공격에 대한 방어 효과가 향상됨을 의미하며, 특히 Neurotoxin과 같은 고도화된 백도어 공격에 대해 더욱 효과적으로 대응할 수 있음을 보여준다. 또한, GPT-2 모델에서는 차분 프라이버시가 특히 효과적으로 작용하여 백도어 공격 방어 성능이 높아지는 결과를 볼 수 있다. 그러나 Table 1에 제시된 모델의 주요 작업 정확도에서 알 수 있듯이, σ값이 증가는 모델의 전반적인 성능 저하로 이어졌다.

Fig. 4.에서는 다양한 방어 기법과 약한 차분 프라이버시를 결합하여 적용한 결과를 나타내며, 차분 프라이버시 방어 기법만 사용할 때보다 기존 방어 기법들과 결합하여 사용할 때 백도어 방어 성능이 더욱 향상됨을 보여준다.

전체 결과를 종합하면, 백도어 공격에 대한 방어와 모델 성능 사이의 상충관계가 명확히 드러난다. 실제 적용 시나리오에서는 모델의 주요 작업 정확도를 유지하면서도 백도어 공격을 효과적으로 방어할 수 있는 최적의 σ값을 선정하고 기존 방어 기법들과 약한 차분 프라이버시를 같이 적용하는 것이 백도어 방어에 있어 효과적이라는 점을 시사한다.

V. 결 론

본 연구에서는 연합학습 환경에서 약한 차분 프라이버시 방어 기법을 적용하여 백도어 공격에 대한 방어 효과와 모델 성능 간의 상관관계를 분석하였다. 실험 결과, σ값이 커질수록 백도어 공격의 정확도가 크게 감소하여 방어 효과를 입증하였으나, 동시에 모델의 주요 작업 정확도에도 부정적 영향을 미치는 것을 확인하였다. 그러나 본 연구를 통해 백도어 공격 방어를 목표로 설정할 경우에는 상대적으로 적은 양의 노이즈로도 효과적인 방어가 가능함을 확인했다. 또한, 다양한 방어 기법과 차분 프라이버시를 결합하여 사용할 때 차분 프라이버시 기법만 사용할 때보다 더 높은 방어 성능을 보였으며, LSTM 모델보다 GPT-2 모델에서 차분 프라이버시 적용이 더 효과적인 방어 결과를 나타내었다. 따라서 실제 적용 시에는 모델의 성능을 크게 저하시키지 않으면서 최적의 노이즈 양을 정하는 것이 중요하다. 차분 프라이버시는 기본적으로 프라이버시 보호를 위한 메커니즘이므로, 향후 연구에서는 백도어 공격 방어뿐만 아니라 모델의 성능에 영향을 미치지 않으면서 프라이버시 보호를 동시에 강화할 수 있는 통합적인 방어 기법에 대한 연구가 필요하다.

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," Artificial Intelligence and Statistics, pp. 1273-1282, Apr. 2017.

[2] P. Voigt and A. Von dem Bussche, The EU General Data Protection Regulation (GDPR): A Practical Guide, 1st Ed., Springer International Publishing, Cham, 265 pages, Aug.

2017.
[3] E. L. Harding, J. J. Vanto, R. Clark, L. Hannah Ji, and S. C. Ainsworth, "Understanding the scope and impact of the California Consumer Privacy Act of 2018," Journal of Data Protection & Privacy, vol. 2, no. 3, pp. 234-253, Mar. 2019.
[4] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, "Applied federated learning: Improving Google keyboard query suggestions," arXiv preprint arXiv:1812.02903, Dec. 2018.
[5] J. Li, Y. Meng, L. Ma, S. Du, H. Zhu, Q. Pei, and X. Shen, "A federated learning based privacy-preserving smart healthcare system," IEEE Transactions on Industrial Informatics, vol. 18, no. 3, pp. 2021-2031, Mar. 2022.
[6] Y. Li, X. Tao, X. Zhang, J. Liu, and J. Xu, "Privacy-preserved federated learning for autonomous driving," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 7, pp. 8423-8434, Jul. 2022.
[7] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov, "How to backdoor federated learning," International Conference on Artificial Intelligence and Statistics, pp. 2938-2948, Aug. 2020.
[8] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," IEEE Access, vol. 7, pp. 47230-47244, April. 2019.
[9] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H. Brendan McMahan, "Can you really backdoor federated learning?," arXiv preprint arXiv:1911.07963, Dec. 2019.
[10] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, "Calibrating noise to sensitivity in private data analysis," Theory of Cryptography, pp. 265-284, Mar. 2006.
[11] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, Julien Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," Advances in Neural Information Processing Systems 30, pp. 119-129, Dec. 2017.
[12] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider, "FLAME: Taming backdoors in federated learning," 31st USENIX Security Symposium, pp. 1415-1432, Aug. 2022.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.
[14] OpenAI, "Language models are unsupervised multitask learners," https://openai.com/index/better-language-models, Accessed on Nov. 13, 2024.
[15] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez, "Neurotoxin: Durable backdoors in federated learning," International Conference on Machine Learning, pp. 26429-26446, Jul. 2022.
[16] Chulin Xie, Keli Huang, Pin-Yu Chen, Bo Li, "Dba: Distributed backdoor attacks against federated learning," International Conference on Learning

Representations, pp. 1-19, Apr. 2020.
[17] Nguyen, Thuy Dung, et al., "Iba: Towards irreversible backdoor attacks in federated learning," Advances in Neural Information Processing Systems 36, pp. 1-13, Dec. 2023.
[18] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," in Concurrency: The works of Leslie Lamport, ACM, pp. 203-226, Oct. 2019.
[19] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308 - 318, Oct. 2016.

〈저자소개〉

최 민 영 (Minyeong Choe) 학생회원
2024년 12월: 조선대학교 정보통신공학부 정보보안전공 재학중
〈관심분야〉 분산 AI 시스템 프라이버시 및 보안

김 현 일 (Hyunil Kim) 정회원
2014년 2월: 국립공주대학교 응용수학과 졸업
2016년 2월: 국립공주대학교 융합과학과 공학석사
2019년 2월: 국립공주대학교 융합과학과 공학박사
2020년~2022년: 대구경북과학기술원 박사후연구원
2022년~2023년: 국립공주대학교 연구교수
2023년~현재: 조선대학교 정보통신공학부 정보보안전공 조교수
〈관심분야〉 암호기술, 분산 AI 시스템 프라이버시 및 보안, DID 인증 기술 등