# 키 분배를 활용한 동형암호 기반의 연합학습 보안 강화 기법

권대호[1], 아짓쿠마[2], 최봉준[3]
[1]숭실대학교 수학과
[2,3]숭실대학교 컴퓨터학과
kwndh01@soongsil.ac.kr, ajitkumar.pu@gmail.com, davidchoi@soongsil.ac.kr

# A Method for Enhancing Security in Federated Learning Using Homomorphic Encryption with Key Distribution

Dae Ho Kwon[1], Ajit Kumar[2], Bong Jun Choi[3]
[1]Dept. of Mathematics, Soongsil University
[2,3]School of Computer Science and Engineering, Soongsil University

요 약

연합학습에서 로컬 모델을 통해 참가자의 데이터 프라이버시를 침해할 가능성이 있다. 동형암호 기반 연합학습은 학습 과정에서 모든 가중치를 암호화해 통신 과정에서의 공격을 차단한다. 그러나 기존의 Paillier 동형암호 기반 연합학습은 모든 참가자가 같은 공개키 및 비밀키를 공유하는 문제가 있다. 본 연구에서는 지속적인 선택적 키 분배를 도입하여 외부에서 다른 참가자의 로컬 모델에 접속할 수 없도록 하고, 내부에서도 다른 참가자의 로컬 모델을 획득하기 어렵게 한다. MNIST 데이터를 사용하여 CNN 모델의 성능을 평가한 결과, 제안된 방법이 기존과 유사한 정확도를 보여준다.

## 1. 서론

연합학습(Federated Learning)은 중앙집중식 기계학습 방식과 달리, 로컬 데이터가 중앙 서버로 전송되지 않고, 로컬 기기 내부에서 학습된 가중치만이 중앙 서버로 전송된다. 이러한 방식 덕분에 연합학습은 데이터 프라이버시를 보호하면서 글로벌 모델을 학습할 수 있어 주목받는다. 그러나, 외부에서 침입한 공격자로부터 참가자가 학습한 로컬 모델을 이용해 데이터 프라이버시를 침해할 수 있다는 가능성이 있다.[1] 이를 해결하기 위해 동형암호(Homom-orphic Encryption)로 로컬 모델을 보호하는 학습 방법이 제시되었다.[2][3] 기존의 방법은 참가자가 자신의 로컬 모델을 학습한 뒤 동형암호로 암호화해 중앙서버로 전송한다. 중앙 서버는 복호화하지 않고, 글로벌 모델을 학습한다. 이 방법은 통신 과정에서 로컬 모델을 보호할 수 있다는 이점이 있다. 다만, 동형암호 특성상 모든 참가자가 같은 공개키와 비밀키를 공유해야 한다는 문제가 있다. 본 연구는 지속적인 선택적 키 분배를 통해 Paillier 동형암호 기반의 연합학습이 가지고 있는 키 공유 문제를 개선하고, 기존의 암호화 연합학습 방식과 제안한 방식 간의 정확도 차이를 비교한다.

## 2. 제안하는 기법

Paillier 동형암호는 암호화된 두 값의 덧셈 연산과 상수 곱만이 가능하다. 따라서, 기존의 Paillier 동형암호 연합학습은 모델 파라미터가 복호화된 상태로 로컬 모델을 학습한다. 학습 종료 후 참가자는 모델 파라미터를 암호화한 뒤 중앙 서버로 전송한다. 중앙서버는 여러 참가자로부터 전달받은 파라미터를 암호화된 상태로 연산해 글로벌 모델을 업데이트한다.

동형암호 특성상 하나의 공개키에 여러 비밀키를 할당할 수 없으므로 모든 클라이언트에게 같은 공개키($PK$)와 비밀키($SK$)를 전달해야 한다. 즉, 통신과정에서 비밀키가 노출되거나, 참가자 스스로 다른 참가자의 로컬 모델 파라미터를 임의로 복호화하여 데이터 유추 공격을 할 수 있다. 이를 보호하기 위해 우리는 새로운 키 분배 알고리즘을 제안한다.

전체 참가자 집합을 $K$라고 할 때, 모든 참가자 $k \in K$가 KEK(Key-Encryption Key)[4]를 위한 비밀키($sec_k$)와 공개키($pub_k$)를 생성한 후 서버로 공개키($pub_k$)를 전송한다. 서버로부터 암호화된 파라미터를 전달받은 참가자가 라운드 $t$마다 (ALGORITHM 1)을 수행한다.

**ALGORITHM 1.**

1. 서버는 참가자$k$ ($\forall k \in K$)를 두 그룹 $G_1, G_2$로 나눈다.
2. 각 그룹$G_i$, $i=1,2$에 대해:
   a) 서버는 임의의 참가자 $a_{i,t}$를 선택
   c) $a_{i,t}$는 ($PK_{i,t}$, $SK_{i,t}$)를 생성한 후 $PK_{i,t}$를 서버에게 전송
   d) 서버는 같은 그룹에 속한 다른 참가자에게 $PK_{i,t}$ 전송
3. 각 그룹은 자신이 얻은 $PK_{i,t}$로 로컬 모델 가중치를 암호화한 뒤 서버로 전송
4. 서버는 두 그룹 각각 가중치를 계산한 뒤 두 개의 글로벌 모델을 업데이트
5. 그룹별 해당 그룹의 업데이트된 가중치를 참가자에게 다시 배포
6. 각 그룹에 대해:
   a) 서버는 $a_{i,t}$에게 같은 그룹에 속한 다른 참가자$b_{i,t}$의 $pub_k$ 전달
   b) $a_{i,t}$는 $pub_k$로 암호화한 $Enc(SK_{i,t})$를 $b_{i,t}$에게 전송
   d) 같은 그룹 내의 모든 참가자에 대해:
      가) 서버는 $b_{i,t}$에게 다른 참가자$c_{i,t}$ 의 $pub_k$ 전달
      나) $b_{i,t}$는 $pub_k$로 암호화한 $Enc(SK_{i,t})$를 $c_{i,t}$에게 전송
      다) 모든 참가자에게 $Enc(SK_{i,t})$가 전달될 때까지 반복
6. 모든 참가자 $k$에 대해 다음을 수행:
   c) 모든 참가자는 자신의 $sec_k$로 $SK_{i,t}$를 획득
   d) 획득한 $SK_{i,t}$로 모델 복호화 후 학습

상기 알고리즘은 다음과 같은 공격 시나리오에 대응한다.

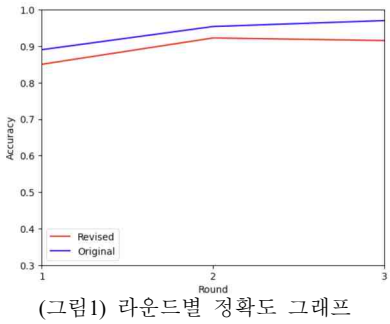
1) 외부 공격: 로컬 데이터를 복호화할 수 있는 비밀키는 참가자 개인의 공개키($pub_k$)로 암호화해 전송되고, 비밀키($sec_k$)는 통신 과정에서 노출되지 않는다.

2) 외부 그룹 공격: 공격자는 자신이 가지고 있는 키로 다른 그룹 참가자의 로컬 데이터를 복호화할 수 없다.

3) 내부 그룹 공격: 공격자가 자신의 키로 다른 참가자의 로컬 모델을 복호화하기 위한 전제조건은 참가자 중 자신과 같은 그룹에 속한 참가자를 찾고, 해당 참가자의 로컬 모델을 획득해야 한다. 하지만, 참가자는 누가 참여했는지 알 방법이 없으며, 암호화된 비밀키($Enc(SK_{i,t})$)를 서로 전송하고 난 후에는 자신에게 키를 전달한 참가자, 자신이 키를 전달한 참가자를 제외하곤 알 수 없다.

## 3. 성능 평가

MNIST 데이터를 활용하여 Vanilla Convolutional Neural Network(CNN) 모델의 성능을 평가했다. 모델의 모든 레이어의 파라미터는 1024bit Paillier 동형암호로 암호화되었다. 참가자는 10명이며, 각 그룹은 무작위로 5명씩 추출하여 총 3라운드의 학습을 진행했다. (그림 1)은 제안된 키 분배 알고리즘의 적용 여부에 따른 글로벌 모델의 테스트 정확도를 비교한 그래프이다. (단, KEK는 구현하지 않았음)




(그림1) 라운드별 정확도 그래프

## 4. 결론

제안된 키 분배 알고리즘은 로컬 모델의 파라미터를 동형암호로 보호하고, 복호화를 위한 키는 매번 다른 경로를 통해 공유함으로써 노출 위험을 최소화한다. 실험 결과에 따르면, 기존의 Paillier 암호화 방식과 정확도 차이가 크지 않아 성능 손실을 최소화하면서도 보안을 유지할 수 있다는 것을 확인할 수 있다. 그러나, 본 연구에서 제안한 알고리즘과 같이 그룹을 두 개로 한정하면 3) 내부 그룹 공격 시나리오에서 공격자가 비밀키 전송 과정에서 알게 된 두 명의 참가자가 이후 라운드에 자신과 같은 그룹에 다시 속하게 되면, 자신의 키로 복호화할 수 있다. 따라서, 그룹 수를 증가시키는 방법 또는 비밀키를 분배하는 과정에서 참가자 간의 정보 노출을 방지하는 방법에 관한 추가적인 연구가 필요하다.

## ACKNOWLEDGMENT

본 성과는 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2022R1A2C4001270), 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터 육성지원사업의 연구결과로 수행되었음 (IITP-2022-2020-0-01602)

## 참고문헌

[1] Liu, P., et al, “Threats, Attacks and Defenses to Federated Learning: Issues, Taxonomy and Perspectives”, Cybersecurity, vol. 5, no. 4, 2022

[2] Tan Soo Fun, et al, “A Survey of Homomorphic Encryption for Outsourced Big Data Computation”, KSII Transactions On Internet and Information Systems, vol. 10, no. 8, 2016

[3] 박재형, et al, “동형 암호 체계를 이용한 연합 학습 기법의 성능 평가”, 동계종합학술발표회, 한국통신학회, pp. 873-874, 2022

[4] Hillmann, Peter, et al. Cake: An efficient group key management for dynamic groups. arXiv preprint arXiv:2002.10722, 2020.