학습 데이터의 프라이버시 보호를 위한 딥러닝 기술 동향

김현지*, 서화정**†
*한성대학교 IT융합공학과
**† 한성대학교 IT융합공학과
khj1594012@gmail.com, hwajeong84@gmail.com

Trends in deep learning technology to protect the privacy of training data

Hyunji Kim*, Hwajeong Seo**†
*Dept. of IT convergence, Hansung University
**† Dept. of IT convergence, Hansung University

요 약

2021년 9대 전략 기술 트렌드로 선정된 인공지능 기술은 다양한 산업 분야에 활용되었다. 그러나 많은 데이터를 필요로 하는 딥러닝의 특성상 민감 데이터 유출 및 악용과 같은 이슈가 존재한다. 최근 개인정보 보호를 위한 규제들이 강화되었으며, 이는 산업 발전을 저해할 것이라는 우려가 커지고 있다. 따라서 데이터 프라이버시 보호와 활용이 모두 가능한 방법론에 대한 연구들이 진행되어야한다. 본 논문에서는 민감 데이터를 보호를 위한 딥러닝 기술 동향에 대해 살펴본다.

1. 서론

2021년 9대 전략 기술 트렌드 중 하나로 선정된 인공지능은 지난 10년간 자율주행, 의료 등 다양한 분야에 획기적인 발전과 긍정적인 영향을 미쳐왔다. 그러나 학습 과정에서 많은 데이터들을 필요로 하며, 데이터를 저장하고 처리하는 과정에서 최근 발생한 AI 챗봇의 개인정보 유출과 같은 문제들이 존재한다. 이처럼 사용자의 민감한 개인 정보가 담긴 데이터들을 보호하기 위한 수칙과 다양한 방법론들이 연구되고 있다.

2. 관련연구

2.1 동형 암호화

동형암호화 (Homomorphic Encryption)는 평문을 연산하여 암호화한 것과 평문을 암호화하여 연산한 결과가 동일하다. 따라서 비밀데이터를 노출하지 않으면서 암호화한 상태로 연산이 가능하다. 동형암호는 크게 3종류로 나뉘어진다. 덧셈 또는 곱셈 중 하나만 지원하는 부분 동형 암호 (Partial Homomorphic Encryption, PHE), 덧셈과 곱셈을 모두 지원하지만 차수가 낮은 다항식과 같이 제한된 연산만 가능한 Somewhat Homomorphic Encryption (SHE) 및 덧셈, 곱셈을 모두 지원하고 bootstrapping을 통해 제한 없이 연산이 가능한 완전 동형 암호 (Fully Homomorphic Encryption, FHE)가 있다. SEAL, HELib, HEAAN과 같은 동형암호화 라이브러리들과 이를 인공지능에 적용한 딥러닝 모델들 또한 존재한다.

3. 데이터 보호를 위한 인공지능 기술 동향

3.1 동형 암호화 기반의 신경망

동형 암호의 특성을 활용하는 방법이다. 사용자의 데이터를 암호화한 후 추론을 수행할 경우, 원본 데이터로 추론한 것과 동일한 결과를 얻을 수 있으나 데이터가 암호화된 상태이므로 민감 데이터를 보호할 수 있다. 개인 데이터를 가진 유저와 연산을 수행하는 서버가 있을 경우, 유저는 public key, private key 및 private key로부터 생성한 evaluation key를 가지며 private key를 제외한 나머지 key는 서버와 공유한다. 자신의 공개키로 데이터를 암호화하기 위해 필요한 파라미터 또한 서버와 공유한다. 학습은 암호화하지 않은 원래 데이터로 진행되고, 추론 시에는 공개키로 암호화된 유저 데이터를 사용한다. evaluation 과정을 통해 fully-connected 및 convolution layer에 필요한 동형 연산 (덧셈, 곱셈)을 수행한다. 그러나 기존 신경망에서는 실수 값을 사용하고 동형암호 연산에서는 다

항식 연산을 수행하므로 encoding 과정을 통해 기존의 신경망을 수정하여 사용해야한다. 이 과정에서 원본 데이터를 다항식으로 변환하며 bias나 padding 값 또한 인코딩하여 사용된다. 이처럼 encryption 이전에 encoding 과정을 거치며, 이후 암호화된 상태로 추론 연산이 수행된다. 서버는 마지막 출력층 활성화 함수를 적용하지 않은 상태로 결과를 반환하며 유저가 최종적으로 활성화 함수를 적용하여 예측하며, 해당 값을 유저의 private key로 복호화하여 결과를 확인한다. 동형암호화를 활용한 신경망에 관한 기본적인 연구와 더불어 동형 컨볼루션 네트워크의 성능 개선에 대한 연구도 존재한다 [1]. 저차 다항식이 아닌 4차 다항식 및 Swish 활성화함수와 근사한 활성화함수를 사용하여 분류 정확도를 개선하였다. Microsoft의 CKKS 체계를 위한 라이브러리를 사용하여 구현하였다. MNIST 및 CIFAR-10에 대해 각각 99.22% 및 80.48%의 높은 정확도를 달성하였으며, 각각 0.04%, 4.11% 개선된 성능을 보인다. 이 외에도 암호화된 센서데이터를 통해 추론하는 동형 컨볼루션 네트워크와 [3], 고해상도 의료 이미지에 대한 엄청난 메모리 오버헤드를 줄이기 위한 작고 자원 효율적인 동형 컨볼루션 네트워크 [4] 등과 같이 성능 개선 및 다양한 분야에 적용된 연구들도 수행되었다.

3.2 연합 학습 (Federated Learning)

유저 데이터를 중앙 서버에서 수집한 후 학습하는 기존의 신경망과 달리 유저 데이터를 서버에 노출시키지 않고 학습이 가능하다.

서버로부터 자신의 디바이스에 최신 상태의 모델을 다운받은 후, 유저 데이터와 해당 모델을 가지고 디바이스 상에서 학습을 시킨다. 이 결과는 서버로 전송되며, 여러 유저가 이와 같은 방법을 통해 학습을 수행할 경우, 다운 받았던 모델을 통해 학습한 새로운 모델들이 여러 개 생기게 된다. 서버에서는 여러 모델을 취합하여 평균을 내는 등의 연산을 통해 최신 모델을 다시 생성한다. 여기까지의 과정을 주기적으로 반복하여 학습을 진행하며, 모델이 갖는 가중치만 전송되기 때문에 유저 데이터가 직접 디바이스 밖으로 노출될 일이 없어 데이터의 프라이버시 보호가 가능하다. 연합학습의 구조상 스마트폰용 가상 키보드에서 다음 단어를 예측하는 언어 모델을 데이터 보호를 위해 클라이언트 장치에서 수행 [4] 하는 등과 같이 리소스가 제한된 엣지 디바이스 상에서 학습 및 추론이 수행된다. 신경망의 스케일업 되면 일반적으로 모델의 정확도를 향상시킬 수 있는데, 엣지 디바이스의 리소스 부족으로 큰 모델을 사용하기 어렵다. 이에 따라 연산 시 소모되는 리소스를 줄이기 위한 연구들도 진행되고 있다. [5]에서는 지식 증류 (Knowledge Distillation)을 통해 중앙 서버에서 무거운 모델을 학습하고 엣지에서는 미리 학습된 해당 모델의 출력을 모방하여 작은 모델으로 학습한다. 이러한 방법을 통해 깊은 모델을 학습할 수 있도록 하는 방법이 제안되었다.

3.3 데이터 비식별화

데이터 비식별화는 데이터에 담긴 식별 가능한 정보나 민감 데이터가 그대로 노출되지 않도록 삭제, 범주화, 마스킹 등을 통해 수정하여 개인 정보가 보호될 수 있도록 한다. 딥러닝에서는 영상 데이터, 의료 데이터, 텍스트 데이터 등 다양한 작업에서 비식별화 기술이 활용된다 [6,7]. 영상 데이터의 경우 블러 필터나 픽셀레이트 필터를 사용하여 영상의 각 프레임을 변형하는 이미지 필터링, 이미지의 일부를 암호화, k명의 얼굴을 합성하여 개인이 식별될 확률을 낮추는 얼굴 합성 기술, 개인 식별 영역을 제거 후 다른 물체로 대체하는 인페인팅 기술 등을 통해 개인 식별 정보를 보호할 수 있다. 그러나 비식별 처리에 중점을 둘 경우 데이터의 유용성이 손상될 가능성이 있으며, 효과적인 비식별 처리를 위한 연구가 필요할 것으로 생각된다.

4. 결론

최근 기술 트렌드 중 하나인 인공지능 기술은 다양한 분야에 적용되며 획기적인 발전에 기여하였다. 그러나 많은 데이터를 필요로 하는 인공지능의 특성상 학습 과정에서 데이터에 담긴 개인 정보가 노출되어 이로 인한 범죄나 개인정보 침해 이슈가 존재한다. 또한, 최근 빅데이터 활용에 대해 개인정보 규제가 강화되었으며 산업 발전을 후퇴시킨다는 의견들이 나오고 있다. 개인 정보를 보호하면서 데이터를 활용할 수 있는 방법이 필요하며, 이에 관련된 연구들이 활발하게 진행되고 있다.

5. Acknowledgment

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2018-0-00264, IoT 융합형 블록체인

플랫폼 보안 원천 기술 연구).

참고문헌

[1] T. Ishiyama, T. Suzuki and H. Yamana, "Highly Accurate CNN Inference Using Approximate Activation Functions over Homomorphic Encryption," 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 3989-3995.

[2] Z. Liao, J. Luo, W. Gao, Y. Zhang and W. Zhang, "Homomorphic CNN for Privacy Preserving Learning On Encrypted Sensor Data," 2019 Chinese Automation Congress (CAC), Hangzhou, China, 2019, pp. 5593-5598.

[3] Jin Chao, Ahmad Al Badawi, Balagopal Unnikrishnan, Jie Lin, Chan Fook Mun, James M. Brown, J. Peter Campbell, Michael F. Chiang, Jayashree Kalpathy-Cramer, Vijay Ramaseshan Chandrasekhar, Pavitra Krishnaswamy, Khin Mi Mi Aung, "CaRENets: Compact and Resource-Efficient CNN for Homomorphic Inference on Encrypted Medical Images," CoRR abs/1901.10074 (2019)

[4] HARD, Andrew, et al. Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604, 2018.

[5] HE, Chaoyang; ANNAVARAM, Murali; AVESTIMEHR, Salman. Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge. Advances in Neural Information Processing Systems, 2020, 33.

[6] YADAV, Shweta, et al. Deep learning architecture for patient data de-identification in clinical records. In: Proceedings of the clinical natural language processing workshop (ClinicalNLP). 2016. p. 32-41.

[7] Yang, X., Lyu, T., Li, Q. et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med Inform Decis Mak 19, 232 (2019).