머신언러닝 기술 효율성 평가 및 모델검증 조사연구

성민경, 이강원, 이진규, 한주연
한국정보통신기술협회
[mksung,blong116, jklee, hanjy]@tta.or.kr

An Investigative Study on the Evaluation of Efficiency and the Model Verification of Machine Unlearning

Sung Min Kyoung, Lee Kangwon, Lee Jinkyu, Han Ju Yeun
Telecommunications Technology Association

요 약

인공지능 학습 방식이 확장됨에 따라 다수의 개인정보가 인공지능 학습에 활용되고 있다. 개인정보는 개인정보보호법에 따라 정보주체가 삭제 요청을 할 수 있으며, 특정 정보를 인공지능 모델에서 삭제하는 기술이 현실적으로 필요하게 되었다. 본 논문에서는 인공지능 모델에서 특정 데이터를 효율적으로 삭제하는 머신 언러닝(Machine Unlearning) 기법을 소개하며, 머신 언러닝 기술 방식 분류, 효율성 측정 방식 소개, 모델 검증 방식 소개를 통해 머신 언러닝에 대한 전반적인 안내를 하고자 한다.

Ⅰ. 서 론

데이터의 가치가 증가함에 따라 다양한 분야에서 데이터 수집 및 활용이 활발히 이루어지고 있다. 데이터에 포함된 개인정보는 민감한 정보를 포함할 수 있으므로, 정보 주체는 개인정보 보호법에 근거하여 개인정보 처리자에게 자신의 개인정보 삭제를 요청할 수 있다. 이러한 법적 권리는 정보 주체가 자신의 개인정보에 대한 통제권을 유지할 수 있게 보장한다.

한편 인공지능(AI)이 발전함에 따라 개인정보가 인공지능 모델의 학습 데이터로 사용되면서, 개인정보가 AI 모델에 내재화되는 현상이 나타나고 있다. 개인정보 보호법에 따라 정보주체가 데이터 삭제를 요청하면 단순히 데이터베이스에서 해당 정보를 삭제할 뿐 아니라 AI 모델에서도 그 영향도를 제거할 수 있어야 한다. 그러나 대부분의 AI 모델은 개인정보가 모델에 내재화되어 있는 부분이 정확하지 않아 제거하기 어렵다. 가장 단순하고 명확하게 개인정보를 삭제하는 방법은 삭제 요청된 데이터를 제외하고 AI 모델을 다시 학습하는 것이나, 이것은 소모되는 시간이 클 뿐만 아니라 특정 상황(연합학습 등)에 따라 초기 학습 데이터에 접근이 불가할 수 있다.

이러한 문제를 해결하기 위해 AI 모델에서 특정 데이터를 효율적으로 삭제하는 머신 언러닝(Machine Unlearning) 기술이 최근 연구되고 있다. 머신 언러닝 기술은 데이터 삭제 요청에 신속하게 대응 가능하며, AI 모델의 성능을 유지하고, 개인정보보호법 관련 법적 요구사항을 충족할 수 있는 기술이다. 또한, 머신 언러닝 기술은 공격자가 악의적으로 주입한 데이터로 인해 AI 모델의 성능이 떨어지거나 보안에 문제가 발생하는 데이터 포이즈닝(Data Poisoning) 문제도 효과적으로 해결할 수 있다.

본 논문에서는 최근 주목을 받는 머신 언러닝 기술의 개념을 소개하고 공개된 머신 언러닝 기술을 기준에 따라 분류한다. 또한 머신 언러닝 기술을 통해 생성된 모델의 효율성 측정 방식과 효과적으로 개인정보가 제거되었는지 검증(Verification)하는 방식을 분류한다.

Ⅱ. 본론

1. 머신 언러닝 기술 방식 분류

앞에서 언급한 바와 같이 머신 언러닝을 달성하기 위한 가장 단순한 방식은 삭제 대상 데이터를 제외하고 모델을 재학습 하는 것이지만 이것은 현실적으로 충족하기 어려운 경우가 많다. 따라서 전체 재학습 없이 머신 언러닝을 달성할 수 있는 방식을 아래와 같이 분류한다.

첫 번째는 데이터 재구성(Data Reorganization) 방식이다. 데이터 재구성 방식은 모델 제공자가 훈련 데이터셋을 재구성하여 데이터를 언러닝하는 기법을 의미한다. 이 방식은 세부적으로 난독화(Obfuscation), 가지치기(Pruning), 교체(Replacement)로 구분된다. 난독화는 의도적으로 새로운 데이터를 추가하여 학습시키는 방법이다[1-3]. 이 방식은 새롭게 추가된 데이터가 삭제 대상이 되는 데이터에 주는 영향이 클수록 효과적이다. 가지치기는 모델을 학습할 때 훈련 데이터를 여러개의 하위 데이터셋으로 분할한 후, 각 하위 데이터셋에 기반하여 여러개의 하위 모델을 훈련시킨 후 하위 모델들을 활용하여 결과를 예측하는 방식이다[4-8]. 언러닝 요청 발생 시 삭제 대상 데이터가 포함된 하위 데이터셋에서 데이터 삭제 후 해당 하위 데이터셋만 다시 훈련한다. 이 방식은 전체 모델이 아닌 특정 하위 모델만 재학습한다는 것에서 전체 재학습과 구분된다. 교체는 기존 훈련 데이터셋을 변화된(Transformed) 데이터셋으로 교체하는 방식이다[9]. [9]의 연구는 계산 가능한 변환으로 대체하는 방식을 제안하여 변화된 데이터셋이 언러닝을 쉽게 구현할 수 있도록 모델 훈련에 사용한다.

두 번째는 모델 조정(Model Manipulation) 방식이다. 모델 조정 방식은 모델의 파라미터(Parameter)를 조정하여 언러닝 작업을 달성하는 것을 목표로 한다. 이 방식은 세부적으로 모델 시프팅(Model Shifting), 모델 가지치기(Model Pruning), 모델 교체(Model Replacement)로 구분된다. 모델 시프팅은 머신러닝 모델에서 기존 파라미터 일부를 조정하거나 이동시켜 모델이 학습한 특정 정보를 제거하는 방식이다[2, 10-17]. 모델 가지치기는 훈련된 모델에서 일부 파라미터를 제거한다[18-20]. 이 방식은 일반적으로 특정 모델 구조에 기반하며, 성능 회복을 위해 파인튜닝(Fine Tuning) 과정을 수반한다. 모델 교체는 주로 의사결정 트리나 랜덤 포레스트에 주로 사용되며 모델의 일부 파라미터를 사전에 계산된 파라미터로 직접 교체하는 방식이다[4].

[표1. 머신 언러닝 방식(기술) 분류]

항목	설명	
데이터 재구성	난독화	일부 가짜 데이터를 포함하여 재학습
	가지치기	학습데이터를 분할하여 삭제 대상 데이터가 포함된 하위 데이터셋에서 해당 데이터를 제외하고 재학습
	교체	학습데이터를 특정 기준에 따라 변형 후 재학습
모델 조정	모델 시프팅	일정한 offset을 학습모델의 parameter에 적용
	모델 가지치기	학습 모델의 일정 parameter를 제거 후 다시 fine-tuning
	모델 교체	학습모델의 일정 parameter를 다른 값으로 교체

2. 머신 언러닝 효율성 측정

본 절에서는 머신 언러닝 방식을 통해 새롭게 생성된 모델의 효율성을 측정하는 방식을 소개한다. 여기서 효율성은 전체 재학습된 모델 M_R과 머신 언러닝 방식을 통해 생성된 모델 M_U의 세 가지 특징을 비교하여 측정한다. 각 측정 지표는 표2에서 자세히 설명한다.

[표2. 머신 언러닝 효율성 측정 지표]

항목	설명
일관성 (Consistency)	M_R이 생성하는 응답과 M_U가 생성하는 응답이 일치하는 정도 측정 (실제 정답과 무관)
정확성 (Accuracy)	M_U의 주어진 질문에 대한 정답률 측정
모델 생성 시간 비율 (Time Ratio)	Time Ratio= A/B * A: M_R 생성에 소요되는 시간 * B: M_U 생성에 소요되는 시간

3. 머신 언러닝 모델 검증(Verification)

본 절에서는 머신 언러닝 방식으로 생성한 모델 M_U가 개인정보를 효과적으로 제거하였는지 검증하는 방식을 소개한다. 각 방식, 설명, 필요사항은 표3에서 자세히 설명한다.

[표3. 머신 언러닝 검증 방식]

항목	설명	필요사항
Attack-based	공격 샘플을 활용하여 검증	상황에 따른 공격 샘플 구성 필요
Relearning time-based	M_U가 추가 학습을 통해 원본 모델과 같은 성능에 도달할 때까지 소요 되는 시간 측정(짧을수록 M_U가 삭제되어야 할 개인정보를 많이 포함하고 있다고 가정)	원본모델 확보 및 측정된 시간을 정량적으로 구분할 수 있는 기준 마련 필요
Accuracy-based	M_R과 M_U의 정확도(Accuracy) 차이를 통해 검증	상황별 정확도 차이(Threshold) 정의 필요
Theory-based	샘플을 통해 M_R과 M_U의 유사성 비교	상황별 유사성 정의 필요

Ⅲ. 결론

본 논문에서는 최근 화두가 되고 있는 머신 언러닝 기술을 소개하고, 머신 언러닝 방식 분류, 효율성 측정 방식 소개, 모델 검증 방식 소개를 통해 머신 언러닝에 대한 최신 이슈를 정리하였다. 향후 연구로는 본 논문에서 소개한 연구들을 확장하여 새로운 머신 언러닝 효율성 측정 기법 및 모델 검증 방식 개발을 통해 머신 언러닝 벤치마크 프레임워크를 개발하는 것이 목표이다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00634 '대용량 정형 데이터 대상 개인정보 가명.익명처리 자동화 및 안전성 검증 기술개발')

참 고 문 헌

[1] Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan S. Kankanhalli. Fast yet effective machine unlearning. CoRR, abs/2111.08947, 2021.

[2] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021.

[3] Daniel L. Felps, Amelia D. Schwickerath, Joyce D. Williams, Trung N. Vuong, Alan Briggs, Matthew Hunt, Evan Sakmar, David D. Saranchak, and Tyler Shumaker. Class clown: Data redaction in machine unlearning at enterprise scale. In Greg H. Parlier, Federico Liberatore, and Marc Demange, editors, Proceedings of the 10th International Conference on Operations Research and Enterprise Systems.

[4] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In 42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021

[5] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. Graph unlearning. In Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi, editors, Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, CA, USA, November 7-11, 2022, pages 499 - 513. ACM, 2022.

[6] Yingzhe He, Guozhu Meng, Kai Chen, Jinwen He, and Xingbo Hu. Deepobliviate: A powerful charm for erasing data residual memory in deep neural networks. CoRR, abs/2105.06209, 2021.

[7] Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021.

[8] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, Algorithmic Learning Theory, 16-19 March 2021.

[9] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In 2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015, pages 463 - 480.

[10] Sebastian Schelter. "amnesia"- towards machine learning models that can forget user data very fast. In 10th Conference on Innovative Data Systems Research, CIDR 2020, Amsterdam, The Netherlands, January 12-15, 2020.

[11] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 9301 - 9309, 2020.

[12] Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020.

[13] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In Arindam Banerjee and Kenji Fukumizu, editors, The 24th International Conference on Artificia Intelligence and Statistics, AISTATS 2021, April 13-15, 2021.

[14] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. CoRR, abs/2103.03279, 2021.

[15] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision- ECCV 2020- 16th European Conference, Glasgow, UK, August 23-28, 2020.

[16] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021.

[17] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. CoRR, abs/2108.11577, 2021.

[18] Kang Wei, Jun Li, Chuan Ma, Ming Ding, Cailian Chen, Shi Jin, Zhu Han, and H. Vincent Poor. Low-latency federated learning over wireless channels with differential privacy. IEEE J. Sel. Areas Commun., 40(1):290 - 307, 2022.

[19] Tianqing Zhu, Gang Li, Wanlei Zhou, and Philip S. Yu. Differentially private data publishing and analysis: A survey. IEEE Trans. Knowl. Data Eng., 29(8):1619 - 1638, 2017.

[20] Lefeng Zhang, Tianqing Zhu, Ping Xiong, Wanlei Zhou, and Philip S. Yu. More than privacy: Adopting differential privacy in game-theoretic mechanism design. ACM Comput. Surv., 54(7):136:1 - 136:37, 2022.

[21] Sebastian Schelter, Stefan Grafberger, and Ted Dunning. Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava, editors, SIGMOD '21: International Conference on Management of Data, 2021.

[22] Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021.

[24] Yinjun Wu, Edgar Dobriban, and Susan B. Davidson. Deltagrad: Rapid retraining of machine learning models. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020.