Machine Unlearning: A Survey on Principles and Challenges

Hyeonsu Lyu, Hyun Jong Yang*
POSTECH

{hslyu4, hyunyang}@postech.ac.kr

머신 언러닝 기술 동향: 원리와 해결해야 할 문제에 관하여

류현수, 양현종*
포항공과대학교

Abstract

Machine learning (ML) has made remarkable progress in many fields. This progress is mainly privileged by the massive amounts of individual data, yet the accessibility to the data has been rarely addressed in the ML area. Also, privacy-preserving has been recently discussed as a factor that modern ML models are required to have. In these circumstances, machine unlearning is emerging as a promising technology to acquire full controllability of data and preserve user privacy. This survey extensively investigates recent trends and challenges for machine unlearning.

Ⅰ. Introduction

Privacy in ML models has rarely been emphasized, compared to its importance. As the era of data begins, machine learning (ML) technologies have been dramatically developed in the various subfields of computer sciences, such as computer vision, natural language processing, and autonomous control systems. The key engine for this development is the utilization of personal data, provided by millions of individuals. This data might contain sensitive information such as personal health or financial status.

Now, ML models must be capable of revoking the learning process, recently called *unlearning*, to keep personal privacy. Unlearning requires full accessibility to data, but the black-box property of ML models makes it intractable to get full control of data. This survey investigates recent attempts to unlearn the trained data in various ML models and suggests the next technical challenges of unlearning.

Ⅱ. Definition of Unlearning

Unlearning refers to a mechanism to remove the information of training data from the ML models. Making the models forget the data is challenging because the generalized description of how ML models memorize data patterns in nonlinear models does not exist yet.

Fig. 1. illustrates a general process of unlearning. After the model is trained with the original dataset by the learning algorithms, a removal request may occur from the data provider. Then, the unlearning algorithm updates the model to get rid of the effects of the data being deleted.

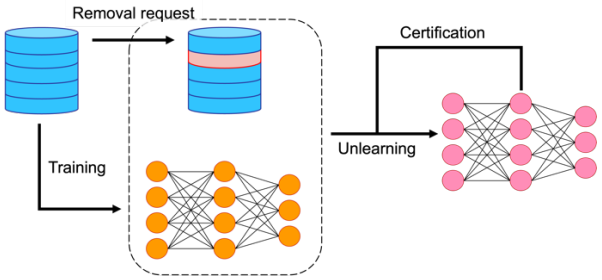

Fig. 1. Basic concept for unlearning

A naïve approach to unlearning is to retrain the model from the scratch without the data being deleted. However, retraining costs a large amount of time when deleting requests occurs in series. Moreover, the intrinsic stochasticity[1] of the training process makes unlearning demanding.

Ⅲ. Modern Approaches to Unlearning

Ginart et al. suggested a definition and the required properties of unlearning, an algorithm that makes the unlearned model undistinguishable from the model retrained without the unlearned data [2]. In other words, for dataset D, data element $i \in D$, and learning algorithm $A: D \to \mathcal{H}$ that maps dataset D to model in hypothesis space $\mathcal{H}$, the unlearning algorithm U shows the same probability distribution as follows:

$$U(A(D), D, i) \sim A(D \backslash \{i\}). \quad (1)$$

Based on this definition, the authors designed a quantized K-means clustering algorithm that supports data deletion.

Bourtoule et al. and Graves et al. proposed unlearning by the model information back-up. Bourtoule et al. suggested a SISA method that trains multiple models for each batch and ensembles the model outputs to get the

[1] The randomness occurs from stochastic gradient descent, random batch, etc.

result [5]. When deleting request reaches, SISA retrains only the model that contains the requested data. Graves et al. designed a method that stores gradient information for each batch in the training process, then uses the gradient information to update the model weight, when specific data is required to be removed [4]. These methods perfectly meet the requirements (1), suggested by Ginart et al., because they partially retrain the model to eradicate the stored information of the data.

Motivated by the influence function [1], Guo et al. proposed a weight update method in linear networks [3]. They defined *certified removal* which measures the probabilistic difference between the unlearned model and retrained model. Wu et al. suggested a method that can preserve performance in the unlearning process [6]. They extended attempts in [3] to nonlinear models by approximating the influence function by using the Taylor expansion. These methods do not perfectly meet (1), but their unlearning results still can be certified by numerical experiments.

Ⅳ. Remaining Challenges in Unlearning

In spite of the early attempts to establish a model-agonistic metric to certificate the unlearning capability, verification of the unlearning performance still remains challenging. Recent research utilizes model inversion attacks and membership inference attacks to justify the unlearning capability [7], but the methods significantly depend on their model architecture and the training process, which makes them model-dependent.

The aforementioned methods have well-established the basic properties of unlearning, but the unlearning for the application-level ML models still needs to be addressed. In the view of the authors, applications of unlearning may include the unlearning for privacy-preserving ML, restoration of overfitted models, and mislabeled data correction, but not restricted to these topics.

Ⅴ. Conclusion

This survey comprehensively introduces the recent trends in machine unlearning with respect to the basic principles and challenges. Unlearning is expected to be a key element that ML models must have to fully control the data and preserve the users' privacy.

ACKNOWLEDGMENT

This research was supported in part by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2021-0-02048) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation), and in part by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC- TD2003-0.

REFERENCES

[1] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in Proc. 34th Int. Conf. Mach. Learn. (ICML), 06--11 Aug 2017, vol. 70, pp. 1885-1894.

[2] A. A. Ginart, M. Y. Guan, G. Valiant, and J. Zou, "Making AI forget you: data deletion in machine learning," in Proc. 33rd Int. Conf. Neural Inform. Process. Syst. (NIPS), Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 3518-3531.

[3] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," in Proc. 37th Int. Conf. Mach. Learn. (ICML), Jul. 2020, pp. 3832-3842.

[4] L. Graves, V. Nagisetty, and V. Ganesh, "Amnesiac machine learning," Proc. Conf. AAAI Artif. Intell. (AAAI), vol. 35, no. 13, pp. 11516-11524, May 2021.

[5] L. Bourtoule et al., "Machine unlearning," in 2021 IEEE Symp. Secur. and Priv. (SP), May 2021, pp. 141-159.

[6] G. Wu, M. Hashemi, and C. Srinivasa, "PUMA: performance unchanged model augmentation for training data removal," arXiv [stat.ML], Mar. 02, 2022. [Online]. Available: http://arxiv.org/abs/2203.00846

[7] T. T. Nguyen, T. T. Huynh, P. Le Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A survey of machine unlearning," arXiv [cs.LG], Sep. 06, 2022. [Online]. Available: http://arxiv.org/abs/2209.02299