Frontier of Artificial Network

Department of Data Science
香港城市大學
City University of Hong Kong

Machine Unlearning: A Reverse Process in AI Models



Invited Speaker
Tianqing Zhu
City University of Macau

Date: October 20, 2026 (Tue)
Time: 14:30 (HKT)
Zoom Meeting: 801 137 0362

Biography

Tianqing Zhu is currently the Dean of Faculty of Data science at City University of Macau, she is also a Professor and PhD supervisor at the City University of Macau and a recipient of the Young Scientists Fund (Class A) of the National Natural Science Foundation of China (NSFC). Before joining the City University of Macau, Zhu Tianqing was a Lecturer at Deakin University in Australia, an Associate Professor at the University of Technology Sydney, and a Professor at China University of Geosciences (Wuhan). As the corresponding or first author, she has published more than 60 CCF-A rated papers in top venues such as NDSS, S&P, USENIX Security and IEEE Transactions. She serves as a Program Committee member for renowned international conferences including CCS, USENIX and AAAI. She is committed to research in artificial intelligence security, focusing on key issues such as attack and defense for intelligent models, data privacy, and the relationship between security and fairness.

Abstract

Traditional machine learning constructs models by organizing knowledge from training data. Machine unlearning acts as its reverse process: it seeks to selectively erase the influence of specified training samples, concepts or memorized content from an already trained AI model, ideally making the model behave as if the target data had never been used during training. Driven by privacy regulations, user data deletion rights, removal of poisoned or biased training content, and safety requirements for large models, machine unlearning avoids the heavy cost of full retraining by developing lightweight forgetting mechanisms. This talk will introduce the core definition and paradigm of machine unlearning, distinguish exact unlearning and approximate unlearning, review representative algorithm families, and discuss key evaluation dimensions: unlearning effectiveness, preservation of model utility on remaining data, computational overhead, and verification challenges. We will also highlight open research difficulties, including knowledge entanglement in deep neural networks, persistent memorization in large language models, and attacks that recover erased information. Finally, we outline real-world application scenarios and future directions in the forgetting capability.