Contribution ID: **196** Type: **Poster**

Analysing Defence Mechanisms Against Gradient Attacks in Contrastive Federated Learning

Monday 13 October 2025 18:00 (1h 30m)

Vertical Federated Learning (VFL) has emerged as a transformative approach in collaborative machine learning, enabling multiple parties to jointly train models while maintaining data privacy through vertical partitioning of features. This paradigm has gained significant traction in privacy-sensitive domains such as healthcare and finance, where different organisations possess distinct feature sets of the same entities.

Despite its promise in preserving data privacy, VFL faces inherent vulnerabilities related to information leakage during the intermediate computation sharing process. Research has shown that even partial information exchange can potentially expose sensitive data characteristics, compromising the system's fundamental privacy guarantees. These limitations have prompted researchers to seek more robust privacy-preserving solutions.

Contrastive Federated Learning (CFL) was introduced as an innovative approach to address these privacy concerns. By incorporating contrastive learning principles, CFL reduces the need for direct feature sharing while maintaining model performance through representation learning. This method has demonstrated promising results in minimising information leakage during the training process.

However, while CFL enhances privacy preservation in feature sharing, it does not fully address the broader spectrum of security threats in federated learning, particularly internal attacks. Among these, gradient-based attacks have emerged as a significant concern, where malicious participants can exploit gradient information to reconstruct private training data or compromise model integrity. These attacks pose a substantial threat to the security of federated learning systems, potentially undermining their practical applications.

In this paper, we conduct a comprehensive experimental analysis of gradient-based attacks in CFL settings and evaluate three defensive strategies: random client selection, gradient clipping-based client selection, and distance-based client selection. Our research aims to quantify the effectiveness of these defence mechanisms and provide empirical evidence for their practical implementation in securing federated learning systems against internal attacks.

Primary author: GINANJAR, Achmad (The University of Queensland)

Co-authors: Prof. LI, Xue (The University of Queensland); Prof. SINGH, Priyanka (The University of Queensland); Dr HUA, Wen (The Hong Kong Polytechnic University)

Presenter: GINANJAR, Achmad (The University of Queensland)

Session Classification: Poster Session

Track Classification: SciDataCon2025 Specific Themes: Infrastructures to Support Data-Intensive Research - Local to Global