RAFL: Robust Federated Meta Learning Framework Against Adversaries

Lancaster University
School of Computing and Communications

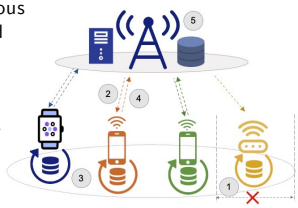
Dr. Zhengxin Yu (z.yu8@lancaster.ac.uk)
Dr. Yang Lu (y.lu44@Lancaster.ac.uk)
Prof. Neeraj Suri (neeraj.suri@lancaster.ac.uk)

Federated Learning (FL)

FL is capable of leveraging distributed personalized datasets from multiple clients to train a shared global model in a privacy-preserving manner

Problem: FL systems can be vulnerable to various kinds of failures and attacks (data poisoning and model poisoning).

➡ **Degrade the learning performance of FL**

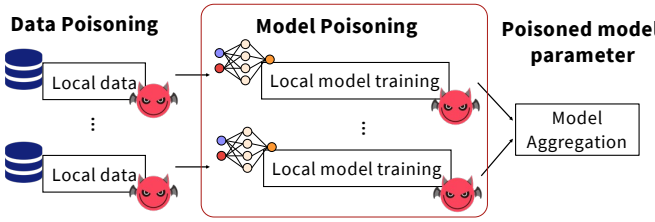


Impact: reduce model accuracy, quality of user experience, trustworthiness, resilience and communication overhead

SOTA: Robust learning and adversarial client detection

Challenges:

- Clients upload unreliable model updates intentionally or unintentionally.
- Local resource heterogeneity (Non-IID data distribution)
- Attacks are complex –discrete, colluding, multi-layered, moving-target behavior
- Dynamic environments (mobility, join-leave behavior, etc.)

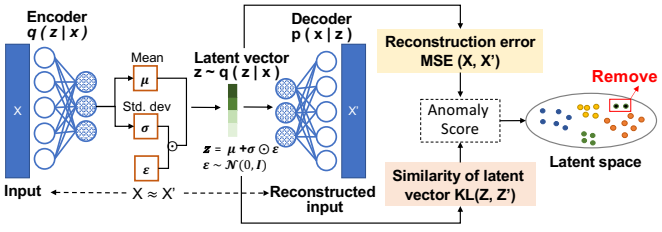


Robust Federated Meta Learning Framework

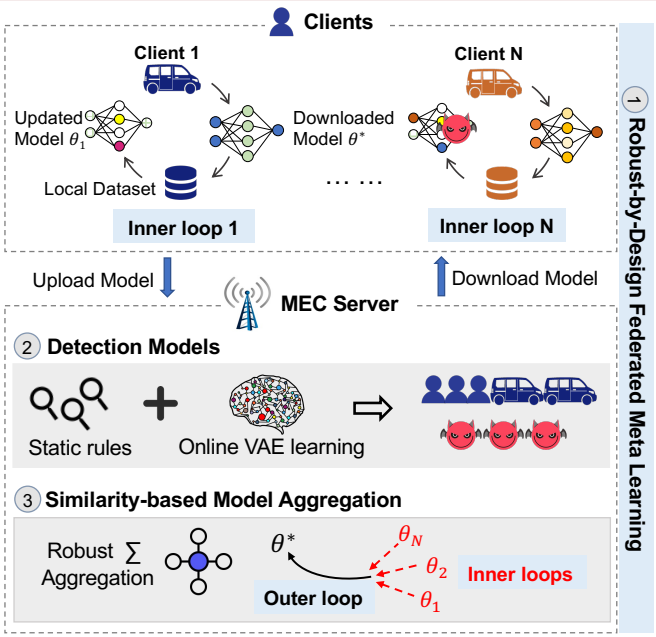
Develop a **robust and adaptive** federated meta-learning framework (RAFL) against adversaries

Contributions :

- A robust-by-design federated meta-learning architecture is proposed to adaptively defend against a range of adversarial attacks.
- A composite rule-based and learning-based detection method is developed to effectively identify adversarial clients via ranking domain and low-dimensional embeddings.
- An adaptive model aggregation method is proposed to aggregate the global model by considering the degree of similarity between the meta-model and calculated mean model to resilience attacks.

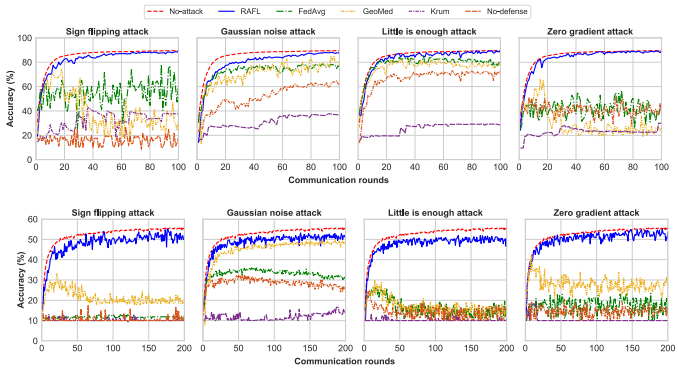


RAFL System Model

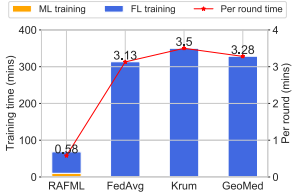


Experimental Results

The experimental results demonstrate that our proposed RAFL framework is robust by design and outperforms other baseline defensive methods against adversaries in terms of model accuracy and efficiency.



We compare RAFL's training time with other benchmark defence schemes. Total training time of RAFL (detector, FL training time) is less than SOTA.



Conclusion

- We have proposed a robust FL framework against adversaries, which combined a rule-based detection method and an online learning-based detection method to effectively distinguish adversarial clients from benign clients.

Future Work

- Explore the applicability of the RAFL to multi-attacks and consider more advanced ML models
- Develop a mobility-aware adaptive federated meta learning framework

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]