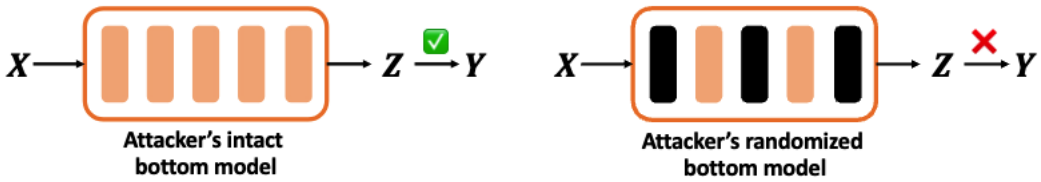
VMask: Tunable Label Privacy Protection for Vertical Federated Learning via Layer Masking

Juntao Tan, Lan Zhang, Zhonghao Hu, Kai Yang, Peng Ran, Bo Li

Frontiers of Computer Science, DOI: 10.1007/s11704-025-51046-z

Problems & Ideas

- **Problems of existing VFL defenses:**
  - Vertical Federated Learning (VFL) systems are highly vulnerable to Model Completion (MC) attacks, where attackers infer private labels by fine-tuning the bottom model.
  - Existing defenses face a dilemma: Machine learning-based methods sacrifice too much model accuracy, while cryptography-based methods incur impractical computational overhead.
- **Ideas:**
  - **Layer Masking:** We propose disrupting the strong correlation between input data and feature embeddings by applying Secret Sharing (SS) to randomize specific linear layer parameters in the attacker's model.
  - **Tunable Defense:** A shadow model is employed to estimate label leakage locally, allowing the defender to dynamically select "critical layers" to mask, ensuring privacy stays within a flexible budget.

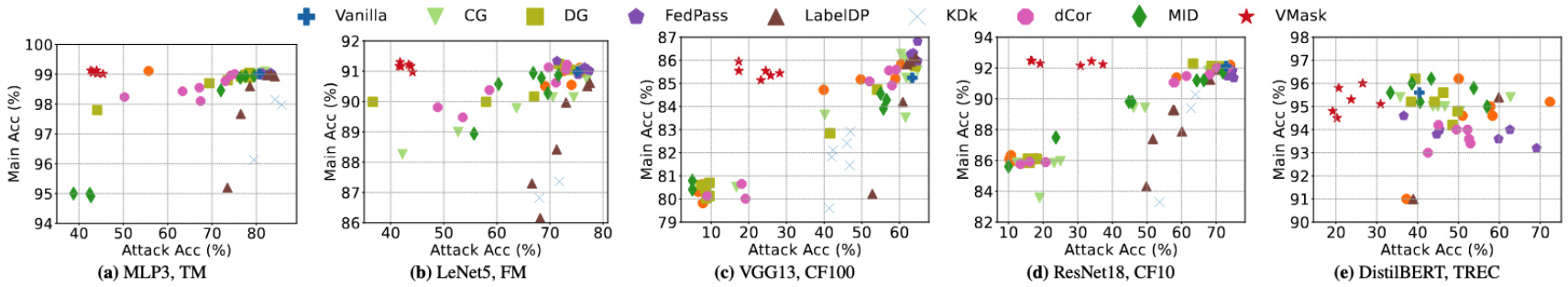


(a) Intact bottom model **(b)** Randomized bottom model

Our key insight: Randomizing certain layer parameters (b) disrupts the mapping from input data to feature embeddings compared to the intact model (a), thereby effectively thwarting the label inference attack.

Main Contributions

- **Contributions:**
  - **VMask Framework:** A novel framework utilizing Secret Sharing to mask layer parameters, effectively defending against MC attacks while preserving main task accuracy.
  - **Efficiency:** We devised a strategy to select only critical layers for masking, making VMask up to 60,846 times faster than Homomorphic Encryption-based methods.
  - **Best Trade-off:** Extensive evaluation shows VMask reduces attack accuracy to a random guessing level with negligible model accuracy drop (e.g., only 0.09% drop in DistilBERT).



Evaluation of the privacy-utility trade-off. VMask (stars in the top-left corner) consistently achieves high main task accuracy and low attack accuracy, significantly outperforming other defense methods.