Group Meeting 3

Research Timeline of Privacy Preserve and MIA

Yixuan Zhang

June 15, 2025

Outline

Motivation & Initial Challenge

Challenge: Deploy large DNNs on resource-constrained devices without exposing raw inputs. **Key Contributions:**

- *SplitNN* (**Paper 53**): "Cut" model between client/server for medical data; preserves accuracy, reduces client FLOPs.
- *Edge-Ensemble* (**Paper 49**): pruned submodels on K devices, ensemble preserves accuracy with small models.
- *SplitFed* (**Paper 51**): parallelizes split learning with FedAvg, adds DP noise at cut layer.

Discovery of Feature Leakage

Challenge: Intermediate activations still leak sensitive data. **Key Contributions:**

- *Fredrikson et al. (2014)*: inversion from output confidences.
- *GAN-driven MIAs* (**Paper 16**): insider trains GAN to reconstruct class samples from gradients.
- *DP & HE/MPC* (**Papers 10, 21, 24, 38, 43, 55**): add noise or compute under encryption—high overhead, limited utility.

Heuristic Feature Obfuscation

Challenge: Remove "redundant" information without theory. **Key Contributions:**

- *Noise Injection* (**Paper 52, 19**): Gaussian noise on features (Nopeek, Noise_ARL).
- *Pruning & Sparsity* (**Paper 7, 54, 15**): PATROL, DistCorr, Dropout move layers client-side / prune channels.
- *Frequency-Domain* (**Paper 34, 35, 58**): DCT-based random high-freq sampling, trainable subtraction.
- *Adversarial Rep. Learning* (**Paper 3, 28, 20**): DeepObfuscator, FaceObfuscator vs. reconstruction & attribute adversaries.
- *Transfer Learning* (**Paper 17**): freeze early layers to block private feature encoding.

Adversary Advances

Challenge: Attackers adapt—attribute inference, user-level leakage, overfitting awareness. **Key Contributions:**

- *Attribute Inference* (**Paper 32**): CSMIA, LOMIA leak sensitive fields from outputs.
- *User-Level Leakage* (**Paper 59**): mGAN-AI recovers specific client data in FL.
- *Improved Loss & Overfitting* (**Paper 41**): logit-maximization + multi-model optimization.
- *Sensitive Feature Distillation* (**Paper 62**): distiller purifies latent features to enable reconstruction.

Towards Theoretical Guarantees

Challenge: Heuristic methods lack worst-case guarantees. **Key Contributions:**

- *Adversarial MI Objectives* (**Paper 3**): constrain mutual information $I(S;Z)$.
- *CEM Algorithm*:
  - **Theoretical Analysis:** Provide a lower bound on minimal reconstruction MSE in terms of conditional entropy (Theorem 1).
  - **Differentiable Bound:** Derive a tractable and differentiable lower bound on $\mathcal{H}(x \mid z)$ via Gaussian Mixture Model (Theorem 2).
  - **CEM Algorithm:** Propose a versatile Conditional Entropy Maximization algorithm that seamlessly **integrates with existing defense mechanisms**.
- **Impact:** Provable lower bound on any inversion attack's MSE.

Key Unresolved Challenges

- **Complex High-Dimensional Multi-Modal Distributions**
  Real-world intermediate features are non-Gaussian and multi-modal, challenging GMM-based entropy estimation.
- **Dynamic Collaborative Environments**
  Client churn and continual model updates invalidate static entropy bounds.
- **Privacy–Utility Trade-off Quantification**
  Lack of unified metrics to balance task performance and worst-case privacy guarantees.
- **Extend to Other From**
  Extending entropy-based defenses to text, audio features.
- **Practical Integration with Other Defense Methods**
  Validate CEM's real-world applicability by plugging it into other defenses methods.