Group Meeting 2

Yixuan Zhang

June 6, 2025

Outline

Background: Collaborative Inference & Privacy

- **Collaborative (Edge–Cloud) Inference:** Partitioning DNNs between user device (edge) and cloud server to protect raw inputs.
  (Papers 49, 51, 53)
- **Privacy Challenge:** Intermediate features can leak sensitive information via Model Inversion Attacks (MIAs).
  (Papers 4, 5, 14, 15, 22, 25, 32, 40, 45, 48, 50, 60, 66, 69, 70)
- Our CEM (Conditional Entropy Maximization) aims to quantify and reduce task-irrelevant redundancy in features to enhance inversion robustness.

Edge-Ensemble for Collaborative Inference (Paper 49)

Shlezinger et al., Collaborative Inference via Ensembles on the Edge (49)

- **Goal:** Enable DNN inference on resource-limited devices by sharing compact models and ensembles.
- **Method:**
  - Deploy pruned/quantized variants of a large model (MobileNetV2 0.5×) on multiple devices.
  - Each device computes local predictions; ensemble of K devices boosts accuracy.
- **Key Insight:** With $K = 3$ devices, accuracy can surpass full MobileNetV2 while keeping parameter count $\approx \frac{1}{6}$ of full model.
- **Relevance:** Demonstrates trade-off between model partitioning and inference accuracy in collaborative setups.

Split Learning Concepts (Paper 51)

Thapa et al., SplitFed: When Federated Learning Meets Split Learning (51)

- **Background:**
  - *Federated Learning (FL):* Each client trains full model locally, server aggregates. High client compute burden.
  - *Split Learning (SL):* Model split at a cut layer; client trains shallow layers, server trains deeper layers in sequence. Slow due to serial updates.
- **SplitFed Learning (SFL):** Combine FL's parallelism with SL's resource efficiency.
  - Clients train only shallow part (f_C) on local data.
  - Server trains deep part (f_S) on received "smashed data."
  - Client gradients aggregated via FedAvg *after* server backprop, adding differential privacy (DP) and PixelDP noise layers.
- **Results:**
  - On HAM10000, MNIST, FMNIST, CIFAR-10: SFL achieves stronger privacy than FL, faster than SL.
  - Accuracy vs. privacy trade-off studied under various ϵ, ϵ'.
- **Relevance:**

Split Learning for Healthcare (Paper 53)

Vepakomma et al., Split Learning for Health: Distributed Deep Learning without Sharing Raw Patient Data (53)

- **Context:** Privacy-critical collaborative training for sensitive medical data (EHR, medical images).
- **SplitNN Framework:**
  - "Cut layer" structure: Client runs shallow network f_C, sends "smashed data" to server's deeper network f_S.
  - Server computes forward pass, backpropagates to cut layer, sends gradients back to client.
  - Multiple configurations: U-shaped (no label sharing at server), vertical data partition (multi-modal fusion).
- **Advantages:**
  - Substantial savings in client compute (0.03–0.155 TFlops vs. $\sim$5–29 TFlops for FL) and bandwidth usage.
  - Nearly equivalent accuracy to centralized training on CIFAR-10, CIFAR-100, etc.
- **Relevance:**
  - Emphasizes split-based inference/training on time-series or image data—motivates investigating inversion from "smashed data"

GAN-Driven Inversion in Collaborative Learning (Paper 16)

Hitaj et al., Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning (16)

- **Setting:** Collaborative/federated learning with parameter sharing. Adversary is an *insider* that receives model updates.
- **GAN-Based Attack:**
  - Attacker trains a *local GAN* whose discriminator uses shared model parameters to distinguish real from generated samples.
  - Generator gradually learns victim's data distribution—can reconstruct specific class samples with high fidelity.
- **Key Findings:**
  - Even record-level DP on gradients cannot prevent GAN-based inversion once noise decays.
  - On CIFAR-10, attacker synthesizes visually recognizable images (e.g., "horse").
- **Relevance:**
  - Highlights that *shared representations* (gradients/parameters) leak sufficient information for high-quality inversion.

Attribute Inference Attacks (Paper 32)

Mehnaz et al., Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models (32)

- **Focus:** *Attribute Inference* vs. *Instance Reconstruction*. Adversary knows non-sensitive attributes, queries classifier to infer sensitive attribute.
- **Two Attack Forms:**
  - *CSMIA* (Confidence Score MIA): Uses model's output confidences on candidate sensitive values to identify correct one.
  - *LOMIA* (Label-Only MIA): Only uses final label. Constructs a surrogate dataset to train a local attack classifier.
- **Results:**
  - Evaluated on GSS, Adult, FiveThirtyEight datasets with Tree-based and DNN classifiers.
  - Both attacks significantly outperform random guess:
    - CSMIA achieves ¿90% accuracy in certain settings.
    - LOMIA still achieves competitive success using only labels.
- **Relevance:**
  - Demonstrates that *even limited output access* can lead to sensitive attribute leakage.

User-Level Privacy Leakage in FL (Paper 59)

Wang et al., Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning (59)

- **Problem:** Existing MIAs often reconstruct class prototypes; authors show server can target *specific client's data*.
- **mGAN-AI Framework:**
  - *Multi-Task GAN* discriminator distinguishes (i) real vs. fake, (ii) sample's class, (iii) client identity.
  - *Identity Representative Extraction:* Server derives a "client fingerprint" from client's parameter updates.
  - Generator produces samples matching both the class label and the client identity distribution.
- **Findings:**
  - On MNIST, AT&T face dataset: mGAN-AI recovers client-specific images close to real pictures.
  - Requires no malicious client—attack can be *invisible* at server side while FL proceeds normally.
- **Relevance:**
  - Emphasizes risk that *server itself* may be adversarial, necessitating robust defenses in FL.

Improving MIAs: Loss and Overfitting (Paper 41)

Nguyen et al., Re-thinking Model Inversion Attacks Against Deep Neural Networks (41)

- **Critique of SOTA Inversion Loss:**
  - Standard identity loss (*maximize softmax output for target class*) is suboptimal for semantic reconstruction.
  - Propose *logit maximization* as identity loss to better align generated images with training data distribution.
- **MI Overfitting Concept:**
  - Inversion often overfits to idiosyncratic noise of the attacked model, producing images that do not generalize to other models.
  - Introduce *model augmentation*: Also optimize against *multiple auxiliary models* (via knowledge distillation) to reduce overfitting.
- **Results:**
  - On CelebA: KEDMI, GMI, VMI attacks improved by 12–15% in Top-1 success rate; first to surpass 90% accuracy.
- **Relevance:**
  - Reveals limitations in prior inversion attacks and the need for robust defenses that consider loss formulations.

Sensitive Feature Distillation Attack (Paper 62)

Yang et al., Measuring Data Reconstruction Defenses in Collaborative Inference Systems (62)

- **Observation:** Many defenses (adversarial training, noise injection, dropout) leave latent sensitive features that can still be distilled.
- **Sensitive Feature Distillation (SFD):**
  - Train a *shadow model* on public data to mimic feature extractor of target.
  - Learn a *feature distiller* that "purifies" obfuscated features, recovering sensitive latent information.
  - Feed distilled features into a standard inversion network to reconstruct input.
- **Results:**
  - On MNIST, CIFAR-10, CelebA: SFD reduces reconstruction MSE by 40–70% over prior attacks, SSIM drastically improves.
- **Relevance:**
  - Demonstrates existing defenses fail to remove all sensitive information—motivation for rigorous quantification (CEM).

Overview of Defense Approaches

- **Cryptography-Based Defenses:**
  - Homomorphic Encryption (HE) & Secure Multi-Party Computation (MPC).
  - Provide theoretical privacy guarantees but suffer from high computational overhead.
    (Papers 10, 21, 24, 38, 43, 55, 37)
- **Obfuscation-Based Defenses:**
  - Add noise, prune, transform, or encode intermediate features.
  - No extra inference overhead; practical for large-scale systems.
    (Papers 3, 6, 7, 11, 17, 19, 20, 28, 29, 31, 34, 35, 36, 37, 44, 58, 61)
- **Information-Theoretic Defenses:**
  - Use mutual information or conditional entropy constraints to limit leakage.
  - CEM (our work) establishes a formal link between conditional entropy and inversion robustness.
    (Papers 41, 62, 31, 44, 57)

Sparse Coding Architectures (Paper 6)

Dibbo et al., Improving Robustness to Model Inversion Attacks via Sparse Coding Architectures (6)

- **Key Idea:** Insert Sparse Coding Layers (SCLs) to force intermediate representations to be *sparse*, removing non-conducive features.
- **Architecture (SCA):**
  - Alternate Sparse Coding Layer with Dense Layer.
  - Add adversarial reconstruction loss against plug-&-play or GAN-based inversion attackers.
- **Results:**
  - On CelebA, Medical MNIST, CIFAR-10, MNIST, FashionMNIST: PSNR/SSIM/FID improved by factors of 1.1–11.7x, 1.1–720x, 1.1–18.3x, respectively.
  - Classification accuracy drop minimal ($<2\%$).
- **Relevance:**
  - Demonstrates efficacy of architectural modifications for robust obfuscation.

Transfer Learning for MI Robustness (Paper 17)

Ho et al., Model Inversion Robustness: Can Transfer Learning Help? (17)

- **Observation:** Early layers in CNNs capture general low-level features—freeze them to prevent private feature encoding.
- **TL-DMI Strategy:**
  - Pre-train on large public dataset (e.g., ImageNet).
  - *Freeze* first L layers; fine-tune last layers on private data.
  - Analyze layer-wise Fisher Information: earlier layers crucial for inversion, later layers for classification.
- **Results:**
  - On CIFAR-10, CelebA: TL-DMI reduces inversion success by ¿33% at similar classification accuracy.
  - Simpler than heavy regularization (BiDO, MID); no complex hyperparameter tuning.
- **Relevance:**
  - Shows how feature reuse from public data can mitigate inversion risk.

Mutual Information Obfuscation (Paper 3)

Xiao & Schaar, Adversarially Learned Representations for Information Obfuscation and Inference (3)

- **Goal:** Maximize utility $I(U; Y)$ while constraining sensitive attribute leakage $I(S; Y) \leq k$.
- **Adversarial MI Framework:**
  - Obfuscator $p(y|x)$ generates sanitized Y.
  - Adversaries attempt to recover X (reconstruction adversary) and S (attribute adversary) from Y.
  - Minimize classification loss on U + adversarial loss (maximize attacker's error).
- **Results:**
  - Synthetic + face (gender vs. emotion) tasks:
    - Utility drop $<2\%$.
    - Strong reduction in attribute inference accuracy.
- **Relevance:**
  - Pioneer in using MI-based adversarial objectives for joint utility-privacy optimization.

Noisy ARL for Image Obfuscation (Paper 19)

Jeong et al., Noisy Adversarial Representation Learning for Effective and Efficient Image Obfuscation (19)

- **Scenario:** Client–server inference—client encodes image, sends features to server; server performs classification.
- **Method:**
  - Insert Gaussian noise $\mathcal{N}(0, \sigma^2)$ on intermediate features.
  - Adversarial training: maximize error of a proxy adversary predicting sensitive attributes from noisy features; minimize target utility loss.
  - Lightweight: client only needs shallow CNN + noise layer.
- **Results:**
  - On face (gender, age), inversion attacks:
    - MS-SSIM drops from 0.9458 to 0.3175.
    - Sensitive attribute accuracy drops from 97% to 58%.
    - Utility accuracy drop $<$2%.
- **Relevance:**
  - Efficient ARL for both reconstruction defense and attribute hiding on resource-limited devices.

DeepObfuscator on Smartphones (Paper 28)

Xiao et al., DeepObfuscator: Obfuscating Intermediate Representations with Privacy-Preserving Adversarial Learning on Smartphones (28)

- **Goal:** Prevent *both* reconstruction and attribute inference from features uploaded by a mobile device.
- **Framework:**
  - *Obfuscator* on-device: learns to perturb features so that two adversaries fail:
    1. Reconstruction adversary: attempts to recover input image.
    2. Attribute adversary: attempts to predict sensitive attributes (e.g., gender, age).
  - Optimize: $\min$ Utility Loss $- \lambda$(Recon Loss + Attr Loss).
- **Results:**
  - On CelebA, LFW:
    - Reconstruction SSIM $\downarrow$ from 0.9458 to 0.3175.
    - Attribute accuracy $\downarrow$ from 97% to 58%.
    - Classification accuracy drop $<$2%.
  - Real-time inference on Android/iOS in milliseconds.
- **Relevance:**

Random Frequency Components for Privacy (Paper 34)

Sun et al., Privacy-Preserving Face Recognition Using Random Frequency Components (34)

- **Key Insight:** CNNs exploit *high-frequency* components for face recognition; human vision relies on *low-frequency*.
- **Method:**
  - Apply DCT to face image $\rightarrow$ multiple frequency channels.
  - Discard low-frequency channels (visible to human).
  - Randomly sample a subset of high-frequency channels per inference.
- **Results:**
  - On LFW, IJB-B/C:
    - Recognition accuracy drop $<1\%$.
    - Reconstruction difficulty significantly increased—randomness prevents consistent inversion.
- **Relevance:**
  - Exploits frequency-domain properties for privacy: a form of obfuscation without additional training.

Trainable Feature Subtraction (Paper 35)

Li et al., Privacy-Preserving Face Recognition Using Trainable Feature Subtraction (35)

- **Core Idea:** Let private representation = original face − reconstructed face, so residual hides visual content but retains discriminative features.
- **Mechanism:**
  - Train an autoencoder on public data to reconstruct face.
  - Compute residual $r = x - \hat{x}$ in high-dimensional (DCT) domain.
  - Randomly shuffle and normalize residual channels, inverse transform to pixel domain $\rightarrow$ protected image X_p.
  - Train recognition model on X_p; adversary sees only X_p.
- **Results:**
  - On LFW, IJB-B: Accuracy within $\pm 1\%$ of frequency-domain methods.
  - Visual appearance: residual images appear as noise; inversion extremely difficult.
- **Relevance:**
  - Demonstrates how learned reconstruction can systematically remove visual content while preserving identity.

Frequency-Domain Privacy Protection (Paper 58)

Zhao et al., Privacy-Preserving Face Recognition in the Frequency Domain (58)

- **Objective:** No-key frequency-domain obfuscation for face images—balance visual privacy and recognition accuracy.
- **Methodology:**
  - Blockwise DCT to obtain frequency channels.
  - Train a *privacy–accuracy trade-off network* to score channel importance:
    - Discard channels that contribute most to human perception (low-frequency).
    - Retain mid/high-frequency channels relevant to automatic recognition.
  - Apply random shuffling, self-normalization, and channel compression to create a "masked" frequency representation.
- **Findings:**
  - On LFW with ArcFace: Recognition accuracy $\approx 99.7\%$ (original).
  - Mask generation + inference runtime in milliseconds; no key management needed.
- **Relevance:**
  - Practical, no-complex-key solution for frequency-based obfuscation in

Privacy-Oriented Pruning: PATROL (Paper 7)

Ding et al., PATROL: Privacy-Oriented Pruning for Collaborative Inference Against Model Inversion Attacks (7)

- **Idea:** Prune parts of network to move more layers to edge, reducing feature leakage to cloud.
- **Key Components:**
  - *Lipschitz Regularization:* Enforce low sensitivity of pruned layers to input perturbations to amplify inversion errors.
  - *Adversarial Reconstruction Training:* Simulate an inversion attacker during pruning to maximize reconstruction error.
- **Results:**
  - On VeriWild, VERI (vehicle re-identification):
    - ResNet-18 pruned from 4 modules to 2–3 modules on edge.
    - Accuracy drop: 3.1% (VeriWild) / 12.7% (VERI).
    - MIA PSNR/SSIM drop by 10–20%.
- **Relevance:**
  - Combines pruning with adversarial training—illustrates architecture-level defense in collaborative inference.

Learning Noise Distributions: Shredder (Paper 36)

Mireshghallah et al., Shredder: Learning Noise Distributions to Protect Inference Privacy (36)

- **Setting:** Edge–cloud inference where client sends intermediate activation to server.
- **Approach:**
  - *Offline Noise Learning:* For each candidate split layer, learn Laplace noise parameters that maximize input–feature SNR reduction while keeping utility loss $\leq 1.5\%$.
  - *Online Inference:* Client samples noise from learned Laplace distributions and adds to activation. Server uses original model, unaware of noise.
- **Outcomes:**
  - Utility accuracy drop $<1.5\%$.
  - Communication speed-up $1.8\times$–$2.2\times$ vs. pure cloud inference under Wi-Fi/LTE.
  - Significant increase in inversion error (PSNR $\downarrow$, SSIM $\downarrow$).
- **Relevance:**
  - Demonstrates systematic noise injection based on information-theoretic proxies

Split Learning for 1D CNNs (Paper 1)

Abuadbba et al., Can We Use Split Learning on 1D CNN Models for Privacy Preserving Training? (1)

- **Domain:** 1D time-series (ECG) classification with 1D CNN.
- **Split Setup:**
  - Train two-layer or three-layer 1D CNN on MIT-BIH dataset.
  - Split at a chosen layer: client computes shallow layers, server computes rest.
  - Evaluate privacy via visualization, distance correlation, dynamic time warping (DTW).
- **Findings:**
  - Split alone does not prevent inversion/reconstruction of ECG signal.
  - Mitigations: deeper split (more layers client-side) and DP noise injection reduce leakage at cost of accuracy.
- **Relevance:**
  - Highlights limitations of naive split learning on time-series data—motivates stronger obfuscation like CEM.

ResSFL: Defense in Split FL (Paper 29)

Li et al., ResSFL: A Resistance Transfer Framework for Defending Model Inversion Attack in Split Federated Learning (29)

- **Challenge:** In SplitFL, server receives "smashed" features and can invert client data.
- **ResSFL Method:**
  - *Attacker-Aware Pre-training:* On a "privileged" device, adversarially train client-side encoder against a strong inversion model (L3).
  - *Resistance Transfer:* Transfer pre-trained encoder to real clients; fine-tune on private data with low-complexity attacker model (L0–L2).
- **Results:**
  - CIFAR-100 + VGG-11: Reconstruction MSE from 0.005 $\rightarrow$ 0.050.
  - Classification accuracy drop $\approx$ 1% (67.5%).
  - Low client compute overhead.
- **Relevance:**
  - Utilizes knowledge transfer to embed inversion resistance in encoder—parallel to CEM's objective of quantifying redundancy.

mGAN-AI for FL (Paper 59) [Revisited]

Wang et al., Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning (59)

- **Recap:** Multi-task GAN style attack on federated client updates—recovers client-specific data.
- **Significance:**
  - Demonstrates that even partial "smashed data" or aggregated updates can reveal client's raw data.
  - Reinforces need for formal redundancy quantification (CEM).
- **Connection to Split Learning:**
  - Both split FL (29) and mGAN-AI (59) show split/distributed settings vulnerable to feature-based inversion.

FaceObfuscator: Mix-Shuffle-Subtraction (Paper 20)

Deng et al., FaceObfuscator: Defending Deep Learning-based Privacy Attacks with Gradient Descent-resistant Features in Face Recognition (20)

- **Approach:**
  - Remove low-frequency channels via BDCT—weakens visible features.
  - Generate *multiple candidate feature sets* by random shuffling, normalization, and linear mixing; each inference picks one at random.
  - Server receives shuffled, compressed features—prevents attacker from learning stable inversion mapping.
- **Results:**
  - On LFW, IJB-B/C: Privacy improvement $\approx 90\%$ over Cloak, AdvFace, DCTDP.
  - Recognition accuracy drop $<0.3\%$.
- **Relevance:**
  - Illustrates advanced combination of frequency-domain and randomization to defeat gradient-based inversion.

Shredder & PATROL Revisited

- **Shredder (36)**
  - Learned Laplace noise injection at optimal split layer.
  - Utility loss $<1.5\%$; communication speed-up $1.8\times$–$2.2\times$.
  - Inversion error significantly increased.
- **PATROL (7)**
  - Privacy-oriented pruning via Lipschitz regularization + adversarial inv. training.
  - Reduces PSNR/SSIM of inversion by 10–20% with minimal accuracy drop.
- **Comparison:**
  - Both target *intermediate features* but use different mechanisms (noise vs. pruning).
  - Shredder: *stochastic* noise; PATROL: *structural* model reduction.

Empirical Evaluation of Defenses

- **Existing Defenses:**
  - ARL-based (3, 19, 28)
  - Frequency-based (34, 35, 58, 20)
  - Pruning/noise (7, 36)
  - Transfer/Architecture (6, 17, 29)
- **Metrics:**
  - Reconstruction Quality: MSE, PSNR, SSIM, FID.
  - Attribute Leak: classification accuracy or AUC.
  - Utility: primary task accuracy drop ($<$2–3% targeted).
  - Efficiency: computation FLOPs, inference latency, communication overhead.
- **Key Observations:**
  - No defense completely removes sensitive information (Paper 62).
  - Many rely on empirical heuristics; lack formal guarantees.
  - Trade-offs vary by domain (face vs. general vision vs. time-series).

Gaps and Need for Formal Quantification

- **Empirical vs. Theoretical:**
  - Most obfuscation defenses (ARL, frequency, pruning) rely on *empirical* measures of redundancy.
  - Information-theoretic bounds often missing; difficult to guarantee frontiers of inversion robustness.
- **Conditional Entropy Perspective:**
  - *Mutual Information* (Paper 3) and *Conditional Entropy* (CEM) connect directly to worst-case MSE of inversion.
  - Our work (CEM) seeks to provide *rigorous quantification* of privacy-critical redundancy in features.
- **Unified Framework:**
  - Existing methods address *specific* scenarios (face, split learning, federated).
  - Need a *versatile* measure to plug into various obfuscation methods—motivates our CEM algorithm.

Summary of Related Works and Positioning CEM

- **Attack Side:**
  - GAN-based inversion (16), attribute inference (32), user-level leakage (59), enhanced inversion (41, 62).
- **Defense Side:**
  - *Architecture-level:* Sparse Coding (6), Transfer Learning (17), ResSFL (29).
  - *Adversarial Learning:* MI-based (3), Noisy ARL (19), DeepObfuscator (28), FaceObfuscator (20).
  - *Frequency-Domain:* Random-Frequency (34), Trainable Subtraction (35), PPFR-FD (58).
  - *Pruning & Noise:* PATROL (7), Shredder (36).
- **Key Gap:**
  - Most defenses use *heuristic* redundancy measures; CEM provides a *principled*, differentiable metric.
  - Enables *consistent*, *plug-and-play* robustness improvements across methods.
- **Next Steps:**
  - Integrate CEM into specific obfuscation pipelines (e.g., Shredder, Shallow Pruning).