University of Nottingham
UK | CHINA | MALAYSIA

Project Proposal:
Investigation of Defense Mechanisms against Model Inversion Attacks

Submitted **October, 2025**, in partial fulfillment of
the conditions for the award of the degree **BSc Computer Science.**

Yixuan ZHANG

20513731

hnyyz39@nottingham.edu.cn

Supervised by Dr. Jianfeng REN

BSc (Hons) Computer Science

School of Computer Science
University of Nottingham Ningbo China

Abstract

Collaborative (edge–cloud) inference splits a network into an on-device encoder and a cloud decoder; the uploaded intermediate features are vulnerable to model inversion attacks (MIAs). Xia et al. [9] formalized a defense by maximizing a Gaussian-mixture lower bound of the conditional entropy $H(x \mid z)$, but fitting high-dimensional mixtures is computationally heavy and sensitive to non-Gaussian feature geometry. This project targets a learnable, distribution-agnostic surrogate of $H(x \mid z)$ that preserves task accuracy while increasing inversion error. To date: (1) the public CEM baseline has been reproduced under the default CIFAR-10 split (VGG11-BN-SGM, cutlayer=4, noise variance 0.025, $\lambda = 16$); (2) a gated-attention conditional-entropy surrogate was designed, using class-wise gated pooling and variance-based log-entropy penalty, raising attack MSE from 0.0436 to 0.0473 and lowering SSIM from 0.432 to 0.411 with only a minor accuracy drop (85.18% to 84.34%); (3) an exploratory Slot + Gated Cross-Attention surrogate (slot aggregation plus Flamingo-style gated cross-attention) reached 85.04% accuracy but degraded privacy (MSE 0.0393, SSIM 0.459). The current evidence favors gated attention as a stronger privacy–utility trade-off. The slot-based route likely needs architectural fusion (e.g., shortcut or parallel coupling with gated pooling) to stabilize variance estimates before it can surpass the baseline. Next steps focus on such fusion designs and hyperparameter sweeps to retain accuracy while further elevating inversion error.

Contents

1 Introduction

1.1 Background and Motivation

In collaborative (edge–cloud) inference, a lightweight encoder on the device produces an intermediate representation z that is sent to a cloud-side decoder. This reduces on-device compute and bandwidth, but exposes z to model inversion attacks (MIAs): adversaries with access to z and the encoder can train decoders or generative models to reconstruct the private input x [1, 3, 4, 6, 10]. Prior obfuscation methods (noise injection, pruning, dropout, adversarial representation learning) improve privacy empirically [2, 5, 7, 8] but lack a principled link to worst-case inversion error.

Xia et al. [9] introduced Conditional Entropy Maximization (CEM), showing that higher $H(x \mid z)$ lower-bounds the minimal reconstruction MSE. They operationalize $H(x \mid z)$ with a Gaussian Mixture Model (GMM) surrogate, but high-dimensional, non-Gaussian features make GMM fitting unstable, costly, and sensitive to initialization. This motivates a learnable, distribution-agnostic surrogate that can be optimized end-to-end with the split model.

1.2 Problem Statement

We seek a surrogate of conditional entropy that is (i) differentiable and stable on high-dimensional, non-Gaussian features; (ii) compatible with split inference constraints (encoder light, decoder heavy); and (iii) empirically improves the privacy–utility trade-off against strong MIA decoders on CIFAR-10 under a standard split protocol.

1.3 Aim and Objectives

Aim. Develop and evaluate a learnable surrogate for conditional entropy that improves MIA robustness in split inference without materially degrading task accuracy.

Objectives.

- Reproduce the CEM baseline [9] on CIFAR-10 with the public protocol (VGG11-BN-SGM, cutlayer=4, Gaussian noise).
- Design a gated-attention surrogate that replaces GMM fitting with class-wise gated pooling and variance-based entropy penalties.
- Explore a Slot + Gated Cross-Attention surrogate (slot aggregation plus gated cross-attention) as a richer, learnable mixture analogue.
- Benchmark the two surrogates against the baseline using classification accuracy and inversion metrics (MSE, SSIM, PSNR) under a common threat model.
- Analyze failure modes of the slot-based surrogate and propose fusion strategies (e.g., shortcut/parallel coupling) to stabilize variance estimation.

2 Related Work

2.1 Model Inversion Attacks

2.2 Split Learning / Edge-Cloud Privacy

2.3 Conditional Entropy Minimization and Surrogates

2.4 Gated Attention Mechanisms

2.5 Slot Attention and Cross-Attention

3 Methodology

3.1 Baseline Overview (CVPR Spotlight)

3.2 Proposed Method 1: Gated-Attention CEM

3.3 Proposed Method 2: Slot + Gated Cross-Attention CEM (Exploratory)

3.4 Evaluation Protocol

4 Implementation

5 Preliminary Results and Analysis

5.1 Main Quantitative Comparison

5.2 Why Method 1 Improves Over the Baseline

5.3 Why Method 2 Underperforms So Far

6 Progress Against Workplan

6.1 Completed Work

-
-
-
-

6.2 Original Plan vs Current Status

6.3 Updated Plan

7 Reflection and Risk Management

7.1 Key Challenges So Far

-
-

7.2 What I Learned

7.3 Risks and Mitigation

- **Risk: Mitigation:**
- **Risk: Mitigation:**

8 Conclusion

References

[1] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270. USENIX Association, 2023.

[2] Shiwei Ding, Lan Zhang, Miao Pan, and Xiaoyong Yuan. PATROL: Privacy-Oriented Pruning for Collaborative Inference Against Model Inversion Attacks . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4704–4713, Los Alamitos, CA, USA, January 2024. IEEE Computer Society.

[3] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery.

[4] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: Information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 603–618, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Jonghu Jeong, Minyong Cho, Philipp Benz, and Tae-hoon Kim. Noisy adversarial representation learning for effective and efficient image obfuscation. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI '23. JMLR.org, 2023.

[6] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019.

[7] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[8] Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, and Ramesh Raskar. NoPeek: Information leakage reduction to share activations in distributed deep learning . In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 933–942, Los Alamitos, CA, USA, November 2020. IEEE Computer Society.

[9] Song Xia, Yi Yu, Wenhan Yang, Meiwen Ding, Zhuo Chen, Ling-Yu Duan, Alex C. Kot, and Xudong Jiang. Theoretical insights in model inversion robustness and conditional entropy maximization for collaborative inference systems, 2025.

[10] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients, 2019.

A Additional Details