MIDAS: Model Inversion Defenses Using an Approximate Memory System

Qian Xu, Md Tanvir Arafir, Gang Qu
University of Maryland, College Park
Outline

- Background
- Problem Formulation
- Proposed Defense: MIDAS
- Experiment Setting and Results
Supervised machine learning models tends to “memorize”

Model inversion attack utilizing autoencoder [2]

Problem Formulation

▷ Threat model
  - Attack goal: exploit the model to reveal *maximum* amount of sensitive data used during training
  - Attack setting: white-box attack + auxiliary information

▷ Defense model
  - Defense goal: *minimize* the revealed sensitive data
  - Defense assumption:
    - The original model M
      - stored a secure storage (i.e., cloud, encrypted HDD etc)
      - loaded to the main memory system during execution
    - The computing hardware supports approximate main memory system
      - dynamic voltage and frequency overscaling -> available for almost all computers
    - Error randomness
      - Errors created in the DRAMs are random due to fabrication variation
Proposed Defense: MIDAS

-Key Idea
  - Problem:
    - Exact parameters
    - -> Overfitting
    - -> Success of model inversion attack
  - Defense
    - Inexact parameters
    - -> Smaller similarity between reconstructed images and training images
  - Physical Implementation
    - Approximate memory system under voltage over-scaling
    - -> Success defense with less revealed sensitive information
Experiment Setting

▸ Previous research [12]
  - Conducted a thorough experimental characterization on 124 DRAM chips from 3 vendors under reduced voltage.
  - Linear voltage decrease -> Error fraction in DRAM chip increases near-exponentially from $10^{-6}$ to $10^2$

▸ Our research
  - Focus on the impact of error fraction/bit error rate on model inversion defense

▸ Metric for defense effectiveness
  - Similarity between reconstructed images and training images
  - Pearson Correlation Coefficient (PCC) [13]
    - $PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$


Experiment Results

- Whether the defense works for all the 40 individuals in the training dataset?
- What is the best voltage over-scaling or bit-flip-rate setting using approximate DRAM memory systems?

![PCC similarity matrix between retrieved images of MIA and original training images for 40 individuals before and after the MIDAS defense (with 0.01 bit error rate).](image1)

![Effect of different settings of voltage overscaling or bit error rate on test set classification accuracy and after/before defense PCC similarity ratio.](image2)
Thank you!