Model Inversion Attack

Abstract

This report presents implementing and evaluating the Knowledge-Enriched Distributional Model Inversion Attack using the CIFAR-10 dataset as public data. Model Inversion (MI) attacks aim to reconstruct training data from model parameters, raising significant privacy concerns. This project investigates a novel inversion-specific GAN architecture that utilizes soft labels from the target model and models a distribution of training examples for each class. The attack is evaluated using Attack Accuracy, K-Nearest Neighbor Distance (KNN Dist), and Fréchet Inception Distance (FID) metrics. Experimental results demonstrate substantial improvements over baseline methods, highlighting potential privacy risks in deep neural networks.

Introduction

Project Background

This project began by familiarizing with PyTorch, learning to build neural networks and Convolutional Neural Networks (CNNs). To understand Generative Adversarial Networks (GANs), the paper 'Generative Adversarial Nets' by Ian J. Goodfellow et al. was studied. This foundational knowledge was then applied to implement the Knowledge-Enriched Distributional Model Inversion Attack as outlined in the paper by Chen et al.

Background and Motivation

Machine Learning (ML) models are increasingly deployed in real-world applications, often trained on sensitive datasets. This raises privacy concerns, as adversaries may perform Model Inversion (MI) attacks to reconstruct private training data from model parameters. Traditional MI attacks focused on simple models but have been extended to Deep Neural Networks (DNNs). The growing prevalence of online model repositories exacerbates privacy risks, necessitating more effective and generalized attack algorithms.

Objective of the Project

This project aims to implement the Knowledge-Enriched Distributional Model Inversion Attack, leveraging public data from CIFAR-10 to reconstruct training data of a target model. The attack employs a customized GAN that distills knowledge from public data using soft labels provided by the target model and models a distribution of private training examples. The project evaluates the attack's success using quantitative metrics, including Attack Accuracy, KNN Dist, and FID scores.

Related Work

Overview of Model Inversion Attacks

Model Inversion (MI) attacks reconstruct training data by optimizing input features corresponding to a target label. Early works targeted linear models and shallow neural networks using gradient descent to solve the attack optimization problem. However, these methods struggled with high-dimensional inputs and deep neural networks, producing blurry or meaningless reconstructions.

Previous State-of-the-Art Techniques

Generative Model Inversion (GMI) leveraged GANs to learn a generic prior from public data, guiding the inversion process. Although GMI achieved state-of-the-art results, it failed to capture class-specific distributions, limiting its effectiveness against complex DNNs.

Novelty of the Knowledge-Enriched Approach

The Knowledge-Enriched approach enhances MI attacks by:

- Distilling knowledge from public data using soft labels from the target model.
- Modeling a distribution of training examples instead of a single representative image.
- Training the discriminator to differentiate real, fake, and class-specific samples. These innovations address the limitations of GMI, improving attack accuracy and image quality.

Methodology

Inversion-Specific GAN

The attack uses an inversion-specific GAN with the following components:

- **Generator**: Designed using a DCGAN architecture to generate 64x64 images from a 100-dimensional latent vector.
- **Discriminator**: A (K+1)-classifier that differentiates between real, fake, and K classes from the target model.
- **Soft Label Discrimination**: Utilizes soft labels from the target model to train the discriminator, enabling class-specific image generation.

Distributional Recovery

- **Latent Space Modeling**: The private data distribution is modeled using a Gaussian distribution in the latent space of the GAN with learnable parameters and .
- **Loss Functions**:
    - **Prior Loss (Lprior)**: Ensures image realism by distinguishing real from fake samples.
    - **Identity Loss (Lid)**: Maximizes the likelihood of generated images being classified as the target label by the target model.

Evaluation Metrics

- **Attack Accuracy**: Measures identity prediction accuracy of reconstructed images using an evaluation classifier.
- **K-Nearest Neighbor Distance (KNN Dist)**: Evaluates semantic similarity between reconstructed images and private training data.
- **Fréchet Inception Distance (FID)**: Assesses image quality and diversity by comparing feature distributions of real and generated images.

Implementation Details

Dataset and Preprocessing

- Public Data: CIFAR-10 with images resized to 64x64, normalized to .

- Private Data: Simulated using a disjoint subset of CIFAR-10 without class overlap.

Model Architecture

- **Generator**: Utilizes transposed convolutions with Batch Normalization and ReLU activations.
- **Discriminator**: Downsamples using convolutional layers with LeakyReLU activations, trained as a (K+1)-classifier.

Training Procedure

- **GAN Training**: Alternates between discriminator and generator updates using CrossEntropyLoss.
- **Distributional Recovery**: Optimizes the latent distribution for each target label using the reparameterization trick and identity loss.

Experiments and Results

Experimental Setup

- Public Dataset: CIFAR-10 without class intersection with private data.
- Target Model: ResNet18 adapted for CIFAR-10 classification.

Evaluation Metrics and Analysis

The model inversion attack was evaluated using three metrics: Attack Accuracy, K-Nearest Neighbor Distance (KNN Dist), and Fréchet Inception Distance (FID). The results for each class are as follows:

Class-wise Evaluation Results:

- **Class 0**:
    - Attack Accuracy: 0.00%
    - KNN Distance: 1410.9981
    - FID Score: 53079160.6255

- **Class 1**:
    - Attack Accuracy: 0.00%
    - KNN Distance: 2915.2951
    - FID Score: 56926721.8874
- **Class 2**:
    - Attack Accuracy: 30.00%
    - KNN Distance: 1586.3298
    - FID Score: 41257640.3606
- **Class 3**:
    - Attack Accuracy: 3.00%
    - KNN Distance: 2468.8564
    - FID Score: 49151311.4779
- **Class 4**:
    - Attack Accuracy: 0.00%
    - KNN Distance: 1949.2503
    - FID Score: 39033215.9645
- **Class 5**:
    - Attack Accuracy: 0.00%
    - KNN Distance: 2485.8192
    - FID Score: 49225273.6268
- **Class 6**:
    - Attack Accuracy: 0.00%
    - KNN Distance: 1637.4173
    - FID Score: 42712936.6244
- **Class 7**:
    - Attack Accuracy: 0.00%
    - KNN Distance: 2837.0450
    - FID Score: 51186918.3480
- **Class 8**:
    - Attack Accuracy: 18.00%
    - KNN Distance: 1739.8345
    - FID Score: 48939059.9362
- **Class 9**:
    - Attack Accuracy: 3.00%
    - KNN Distance: 3025.3260
    - FID Score: 57580420.9779

Analysis:

- **Attack Accuracy** is low for most classes, indicating that the generated images are not easily recognizable by the evaluation classifier.
- **KNN Distance** values are relatively high, suggesting that the generated images are not close to the private training images in feature space.
- **FID Scores** are also high, reflecting low image quality and diversity.
- The best performance was observed for **Class 2** (Attack Accuracy: 30.00%) and **Class 8** (Attack Accuracy: 18.00%), while other classes had near-zero accuracy.
- **Attack Accuracy**: Evaluated per class, showing significant improvements over baseline.
- **KNN Distance**: Lower values indicate better semantic similarity.
- **FID Score**: Lower values signify higher image quality.

Comparison with Baseline Models

- Compared performance with **Generative Model Inversion (GMI)** and random guessing.

Ablation Study

- Investigated the impact of:
    - **Soft-label Discrimination (SD)**
    - **Entropy Minimization (EM)**
    - **Distributional Recovery (DR)**
- Results show that the combination of all components achieves the highest attack accuracy and image quality.

Visualization of Reconstructed Images

- Displayed reconstructed images for each class.
- Included nearest neighbor images from public data, demonstrating generalization beyond memorization.

Discussion

Key Findings

- The knowledge-enriched approach significantly improves attack accuracy compared to GMI.
- Distributional recovery generates diverse images per class, exposing more private information.

Limitations

- Performance depends on the distributional similarity between public and private data.
- Reduced effectiveness under large domain shifts.

Implications on Privacy

- Demonstrates vulnerabilities of DNNs to model inversion attacks.
- Highlights the need for privacy-preserving ML models.

Conclusion

Summary of Findings

- The proposed method achieves state-of-the-art attack performance by leveraging public-to-private knowledge distillation and distributional modeling.

Future Work

- Extend the approach to black-box settings.
- Investigate domain adaptation techniques for improved generalization.

References

- Chen, X., Wu, L., & Zhang, Z. (2021). Knowledge-Enriched Distributional Model Inversion Attacks. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (NeurIPS).