The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks

Brief Summary

The paper proposes an improved framework for Model Inversion (MI) attacking on target models using Deep Neural Networks. The research focuses solely on target models that work with image datasets. The proposed framework '*Generative Model Inversion' (GMI)* attack makes use of GANs along with auxiliary knowledge of the target data (corrupted images with non-sensitive information) and the target model to recover diverse images that have a high probability of belonging to the target network's feature space. The GMI attack framework particularly focuses on the '*white-box'* style of attack wherein the adversary has access to the target model parameters. Through quantitative empirical evidence, the authors show that their MI attacking framework significantly surpasses the efficacy (~75% for reconstructing face images) of the most efficient frameworks deployed to date.

The paper provides theoretical as well as empirical proofs to validate the intuitive claim that a model with high predictive powers is more susceptible to an MI attack. Moreover, the authors also test their MI attacking framework against the '*differential privacy'* notion that is revered for its ability to protect the training data privacy and find that their attacking framework surpasses this notion in obtaining significant features of the training data with high fidelity; Thus, rendering it extremely ineffective.

Main contribution of the paper

The main contributions of the paper are:

1. In the 2-step reconstruction process that utilizes
   - the canonical Wasserstein-GAN training loss along with the Diversity loss in the Public Knowledge Distillation step which allows the model to generate realistic and diverse images.
   - the Prior and Identity losses in the Secret Revelation step which encourages the model to generate diverse images with high likelihood under the target network.
2. The empirical and theoretical proof of proportionality between the predictive powers of the model and its susceptibility to an MI attack.
3. The empirical proof of the ineffectiveness of the Differential Privacy notion to protect the training data features.

List of strengths and weaknesses

Strength	Weakness
Innovative and successful use of GANs to reconstruct informative images	GMI attacking framework has been developed specifically with image related model (not a general model)
Establishing correlation with the Image Inpainting domain to theoretically prove susceptibility of highly predictive model to an MI attack	Experimental comparison is only carried out with a single MI attacking framework which was not developed particularly for image-based models
Constructively adopting loss terms from successful works in the GANs and Image Inpainting domains	Success rate is highly dependent on availability of auxiliary knowledge from the same pool of data as the target dataset
Uses appropriate quantitative metrics to test and compare with the EMI and PII models	

Explanation of strengths and weaknesses

The model successfully adopts a Generative Adversarial Network to synthetically reconstruct images from the training set. The authors appropriately use work inspired from recent research. For example:

- The '*diversity loss'* inspired from work in the Diversity-sensitive Conditional GAN domain is used to produce highly diverse and realistic images that coincide with the target network's feature space with high fidelity.
- The model uses a global and a local discriminator inspired by work in the Image Inpainting domain to ensure patch-wise consistency and whole coherence of the reconstructed images.

The paper also uses appropriate evaluation metrics such as Attack Accuracy metric which classifies the reconstructed images and provides an interesting insight into how the Existing MI attacking framework performs classification correctly but fails to produce images with relevant feature information.

However, the research results rely heavily on the auxiliary knowledge of the target data and the accuracy drops by ~20% when the GAN is trained on the *PubFig83* dataset that does not generate from the same pool as the *CelebA* dataset used to train the target face recognition network.

Comment on the experiments

To quantitatively test and compare the GMI attacking framework with the existing MI attacking framework, the authors used appropriate evaluation metrics such as Peak Signal-to-Noise Ratio (PSNR), the Feature Distance and the KNN distance between the reconstructed images and target classes to evaluate if significant information about the training data was learnt. Use of the Attack Accuracy metric to classify the reconstructed images also interestingly showed that Existing MI attacking framework led to correct image classifications despite performing poorly on the other 3 metrics.

The experiments carried out to test the significance of public knowledge availability also showed that the accuracy of the model was proportional to the amount of public data supplied to the GAN and that the model performed poorly but still better than the Existing MI framework when no auxiliary data was available.

Surprisingly, the experiments also showed that not only does the GMI attacking framework remain unaffected by the Differential Privacy notion but depending on the parameters set for the DP models, it could also lead to higher accuracy.

How could the work be extended?

As suggested by the authors, further research could be carried out into performing '*black-box'* MI attacking as well as developing privacy protection notions that are more rigid than the Differential Privacy notion. The GMI attacking framework could also be tested on models involving training data other than images such as speech. Further improvements could also be made to obtain significantly accurate results when the auxiliary information is not available.

Additional Comments

One of the unclear assumptions in the paper for me was about the latent feature vector '*z*' being drawn from a zero-mean unit-variance Gaussian distribution. I did not understand how the zero-mean and unit-variance parameters were chose.

The paper provides an interesting insight into the correlation between the MI attacking and Image Inpainting domain and suggests the high interdependency of image related networks constructed for distant tasks.