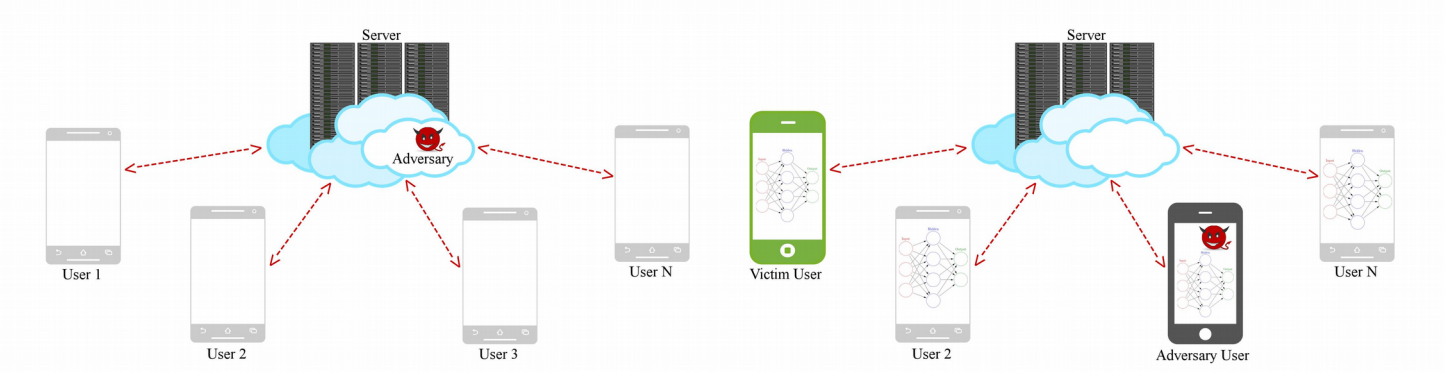
Report 1: A survey on the Collaborative Deep Learning Attacks

Introduction

Deep learning based on artificial neural networks is a very popular approach to modeling, classifying, and recognizing complex data such as images, speech, and text. The unprecedented accuracy of deep learning methods has turned them into the foundation of new AI-based services on the Internet. While the utility of deep learning is undeniable, the same training data that has made it so successful also presents serious privacy issues. Centralized collection of photos, speech, and video from millions of individuals is ripe with privacy risks; direct access to sensitive information. Collaborative learning as shown in fig 1. is introduced where each participant trains a model locally on a device and shares with the other users only a fraction of the parameters of the model. To make the learning model robust, the shared parameters can be truncated or obfuscated using differential privacy(DP) as in [1]. [2] elucidates powerful attack against collaborative deep learning using GANs. The result of the attack is that any user acting as an insider can infer sensitive information from a victim's device. The attacker simply runs the collaborative learning algorithm and reconstructs sensitive information stored on the victim's device. The attacker is also able to influence the learning process and deceive the victim into releasing more detailed information.



(a) Centralized Learning (b) Collaborative Learning

Figure 1: Two approaches for distributed deep learning. In (a), the red links show sharing of the data between the users and the server. Only the server can compromise the privacy of the data. In (b), the red link show sharing of the model parameters. In this case a malicious user employing a GAN can deceive any victim into releasing their private information.

Whitebox and Blackbox attacks: There are two main research directions in the literature on adversarial attacks based on different assumptions about the adversarial knowledge of the target network. The first and the most common line of work; whitebox attacks assumes that the adversary has detailed knowledge of the network architecture and the parameters resulting from training (or access to the labeled training set). Using this information, an adversary constructs a perturbation for a given image. The most effective methods are gradient-based: a small perturbation is constructed based on the gradient of the network loss function w.r.t. the input image. Often, adding this small perturbation to the original image leads to a misclassification. In the second line of work; black-box attacks, an adversary has restricted knowledge about the network from being able to only observe the network's output on some probed inputs. This attack strategy is based the idea of greedy local search, an iterative search procedure, where in each round a local neighborhood is used to refine the current image and in process optimizing some objective function that depends on the network output. In each round, the local search procedure generates an implicit approximation to the actual gradient w.r.t the current image by observing changes in output. This approximate gradient provides a partial understanding of the influential pixels in the current image for the output, which is then used to update this image.

Related Work

1. Model Inversion(MI) Attack: Once the network has been trained, the gradient can be used to adjust the weights of the network and obtain a reverse-engineered example for all represented classes in the network. For those classes that it did not have prior information, it would still be able to recover prototypical examples. This attack shows that any accurate deep learning machine, no matter how it has been trained, can leak information about the different classes that it can distinguish.

Limitation: Due to the rich structure of deep learning machines, the model inversion attack may recover only prototypical examples that have little resemblance to the actual data that defined that class. It may or may not be considered as an attack as it may construct wrong/meaningless information.

2. Generative Adversarial Networks(GAN): The GAN procedure pits a discriminative deep learning network against a generative deep learning network. The discriminative network is trained to distinguish between images from an original database and those generated by the GAN. The generative network is first initialized with random noise, and at each iteration, it is trained to mimic the images in the training set of the discriminative network. The procedure ends when the discriminative network is unable to distinguish between samples from the original database and the samples generated by the generative network.

3. Privacy Preserving Learning: [3] introduces the concept of distributed deep learning as a way to protect the privacy of training data. In this model, multiple entities collaboratively train a model by sharing gradients of their individual models with each other through a parameter server.

GAN based attack in Collaborative Learning

It uses GANs in a new way, since they are used to extract in- formation from honest victims in a collaborative deep learning framework. The GAN creates instances of a class that is supposedto be private. In [2], GAN-based method works only during the training phase in collaborative deep learning. This attack is effective even against Convolutional Neural Networks which are notoriously difficult to invert, or when parameters are obfuscated via differential privacy. It works in a white-box access model where the attacker sees and uses internal parameters of the model.

The adversary works as an insider within the privacy-preserving collaborative deep learning protocol. The objective of the adversary is to infer meaningful information about a label that he does not own.

Fig. 3 from [2] shows the proposed attack, which uses GAN to generate similar samples as that of training data in a collaborative learning setting, with only access to shared parameters of the model. The proposed attack in [2] works in cases of federated, distributed, collaborative, DP based learning models.

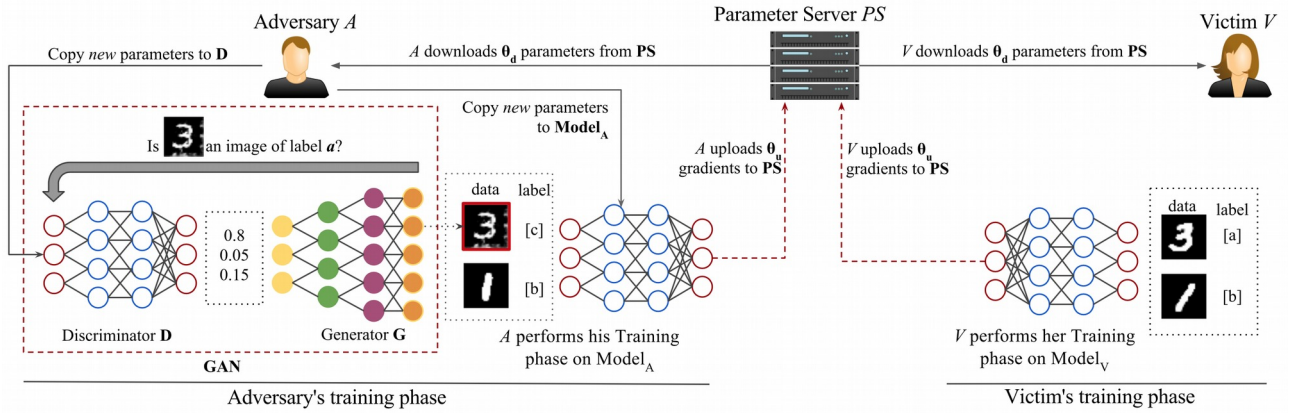


Figure 3: GAN Attack on collaborative deep learning. The victim on the right trains the model with images of 3s (class a) and images of 1s (class b). The adversary only has images of class b (1s) and uses its label c and a GAN to fool the victim into releasing information about class a. The attack can be easily generalized to several classes and users. The adversary does not even need to start with any true samples.

A, V participates in a collaborative learning. V(the victim) declares labels [***a, b***]. The adversary A declares labels [***b, c***]. Thus, while ***b*** is in common, A has no information about the class ***a***. The goal of the adversary is to infer as much useful information as possible about elements in ***a***. Our insider employs a GAN to generate instances that look like the samples from class ***a*** of the victim. The insider injects these fake samples from ***a***, as class ***c*** into the distributed learning procedure. In this way, the victim needs to work harder to distinguish between classes ***a*** and ***c*** and hence will reveal more information about class ***a*** than initially intended. Thus, the insider mimics samples from ***a*** and uses the victim to improve his knowledge about a class he ignored before training. The GAN attack works as long as A's local model improves its accuracy over time.

In [2], the authors proved that GAN attack was successful in all set of experiments conducted juxtaposed with MI attacks, DP based collaborative learning, etc. The GAN will generate good samples as long as the discriminator is learning. The attack is effective even when differential privacy is deployed, because the success of the generative-discriminative synergistic learning relies only on the accuracy of the discriminative model and not on its actual gradient values.

Conclusion

GAN attack proved successful in most of the experiment in contrast with the MI attack, irrespective of whether DP is implemented. The accuracy of the discriminative model in GAN attack was above 97% as reported in [2].

The main point of research in [2] is that collaborative learning is less desirable than the centralized learning approach it is supposed to replace. In collaborative learning, any user may violate the privacy of other users in the system without involving the service provider. Though DP was introduced to protect the privacy of databases, the GAN based attack works independent of whether DP is implemented.

References

[1] Deep Learning with differential privacy, Martin Abadi, 2016
[2] Deep Models under the GAN: information leakage from Collaborative deep learning, Briland Hitaj
[3] Privacy-Preserving Deep Learning, Reza Shokri and Vitaly Shmatikov. 2015.