Draft research question

May 6, 2020

1 Roadmap

1. introduction and literature study (1.5 weeks) introduction + related works
2. draft methodology to answer research question (1 week)
3. **experiment with simple attacks (1-2 weeks)**
4. **implement methodology and start experiment (3-4 weeks)**
5. analyse results (2-3 weeks)
6. write paper (2 weeks)
7. presentation(1-2 weeks)

2 Research question

Machine learning applications have grown in popularity over the last decade. One of the most used kinds of models are neural networks which grew in popularity with the advances in hardware and deep learning algorithms, and because they excel in image recognition and natural language processing applications. Cloud providers have included Machine Learning as a Service to their offer, it is now easier than ever for businesses to deploy machine learning models. MLaaS platforms such as AWS[1], Microsoft Azure[2], Google Cloud AI[3] or IBM Watson[4] offer their customers to train their model given

[1] https://aws.amazon.com/machine-learning/

[2] https://azure.microsoft.com/en-us/services/machine-learning/

[3] https://cloud.google.com/ai-platform

[4] https://www.ibm.com/cloud/machine-learning

they provide the dataset. The final model is then shared or can be accessed through queries. However, such models, and machine learning models in general, are sensitive to malicious attacks. The attacks can impact the privacy of the training set, or the well-Extraction attacks aim at infering the architecture and the parameters of the attacked model to create a duplicate. Inversion attacks aim at infering information from the data the attacked model was trained on. Data poinsoning and adversarial attacks aim at compromising the attacked model predictions either by polluting the training data, or by crafting malicious inputs to feed into the model. ~~The final model can be queried as a black-box or not, it is nevertheless sensible to privacy attacks aiming at inferring information from the data it was trained on.~~.

Several kind of model inversion attacks exist with different goals: the attacker could want to determine whether a data instance was used for training (membership inference)[**?** **?**], reconstruct a representative of a particular class of the training set [**?**], or infer hidden properties of the training set (property inference)[**?**]. ~~Privacy attacks can go from determining whether a data point was used for training, (*membership* inference), reconstructing class representatives (*model inversion*), or inferring hidden properties of the training set (*properties inference*)~~

(TO DO): The implementation of the GDPR in 2018 has put a question on whether those kinds of models are compliant (*question: can it be argued that if the model learns more information than it needs, then Art.5.1.c - data minimization applies?* I still need to write this part).

Membership inference attacks work the following way: given a data point, determine whether it was used to train the given model. The issue with such attacks is they are not always applicable in real-life settings as the attacker needs to get a copy of one of the training data before determining whether it belonged to the training set. Inversion model attacks operate such that, for a classifier, for example, a class representative can be created. While they are impressive for face recognition models, they only build the “average” representation of the class and do not work well if the data instances are not similar. Property inference attacks (PIA) aim at inferring properties learned by the model that are independent of the characteristics of the class the model is trained to recognize. There have been fewer studies on this kind of attack, according to **?**] only 4 papers were published on model inversion attacks against 10 for membership inference attacks. ~~according to [**?**] they only account for 28% of the published papers~~

~~on model inversion attacks agains 72% for membership inference attacks~~. In [?] the authors identify the invariance of fully connected neural networks to node permutation as a limiting factor of PIA efficiency and show that using a Deep Set architecture [?] is an effective way to tackle this limitation. In [?] the authors successfully perform PIA in a collaborative learning setting where the attacker is one of the members.

The kind of model used has an impact on the kind of attacks that can be used on it (I still need to do that. *remark: I probably need to include a source here, it is a bit of a strong claim.*). As computer vision and natural language processing popularity increases, more complex neural network architectures are created even if their vulnerability to privacy attacks is still not fully determined [?]. None of the current work studied the relationship between the depth and width of a deep neural network and its sensibility to PIA. ?] analyzed the effect of model architecture on sensibility to model inversion attacks on non-sensible data (cars, animals).

In case of a trained Neural Network, are sensitive properties of the underlying dataset at risk due to property inference attacks, and if so, how does that risk relate to the architecture of the model such as depth and width?

- How does the sensibility to property inference attacks change with changes in depth and width of deep neural networks?
- Does it affect compliance to the GDPR with respect to processing of sensitive data?
- Question: do I still need to specify the secondary questions when the initial research qeuestion is detailed? Or should I make the research question broader and the secondary questions more detailed?

3 Background

3.1 Deep Learning model

(TO DO)

3.2 Federated Learning

(TO DO)

3.3 Property Inference Attack

The main idea of a PIA is to train an attack model M_a which takes as input a trained model and outputs the probability that P is true for the training set of the input model. Once the attack model is trained, the attacker can give it the target model M_t as input and know whether the P is true for the dataset used to train the target model. To train M_a, the attacker uses several shadow models (M_{s1} ... M_{sk}) which are respectively trained on datasets D_1 ... D_k specifically crafted by the attacker to contain or not the target property P. Once the shadow models are trained, they are used as the training set for M_a. The general overview of the attack is described in Figure **??**.

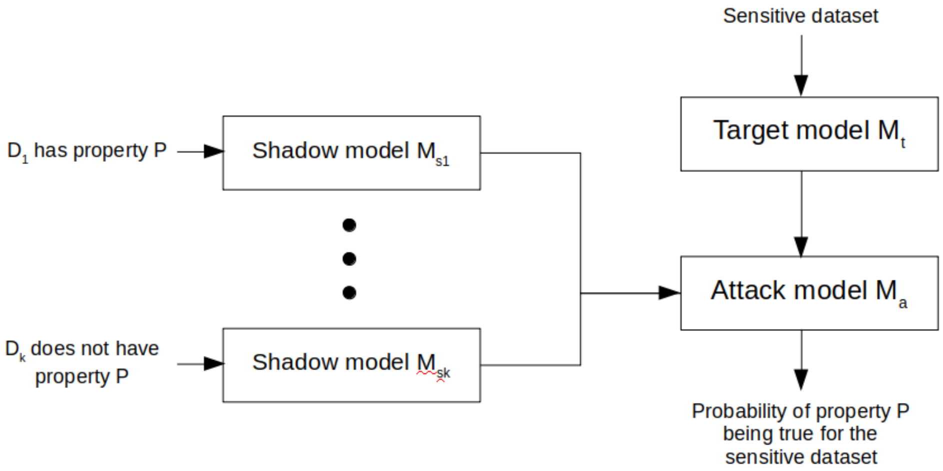

Figure 1: Property Inference Attack using shadow models.

4 Methodology

4.1 Threat model

We assume the target model is a deep neural network trained for a classification task on a sensitive dataset containing personal data. The attacker has access to the full architecture and parameters of the model (white-box setting). While a white-box model facilitates the job of the attacker, it is reasonable as it is common in some machine learning paradigms such as

		depth				
		10	20	40	80	160
width	8					
	16					
	32					
	64					
	128					

Table 1: Target models accuracies on CelebA.

Federated Learning [?] where the clients, and sometimes the server, have full knowledge of the model. Moreover, model extraction attacks have been performed with close to perfect performance, converting the black-box setting into a white-box one [?]. How the attacker manage to obtain the weights and the architecture of the model is out of the scope of this study. ~~There also exists model extraction attacks with close to perfect performance [?]~~. The goal of the attacker is to infer general information about the training dataset. Such information could be the proportion of the training data having a property unrelated to the main classification task of the model.

4.2 Datasets

CelebFaces Attributes (CelebA) [?] is a face attributes dataset containing more than 200 000 images of more than 10 000 celebrities taken from online. The images are labeled using 40 physical attributes such as hair color, smiling, wearing a hat. We use the dataset to train model M_t to detect whether the person is wearing glasses. To add other datasets if we end up using more.

25 different architectures of target models are trained using 5 width and 5 depth values. The width values are 8, 16, 32, 64, 128 (*to add: I don't know what reasonable width values to use at the moment*). The depth values are 10, 20, 40, 80, 160 (*to add: same, I am not sure what values yet. I saw 18, 20, 110 in [?] but it was not for face attributes detection*). The list of the architectures and their test accuracy is presented in the table

4.3 Evaluation

The two properties we try to infer are both sensitive according to the GDPR and are listed as follow:

- P_1: whether the training set contains an imbalanced number of male/female (*question: maybe infer the exact proportion is a more impressive feat, it is probably more difficult*)

- P_2: whether the training set contains an imbalanced number of light skin celebrities

to add: I want to describe which metrics I'm going to use to assess the attack.

References

[] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. Proceedings - IEEE Symposium on Security and Privacy, pages 3–18, 2017. ISSN 10816011. doi: 10.1109/SP.2017.41.

[] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. (February), 2019. doi: 10.14722/ndss.2019.23119.

[] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the ACM Conference on Computer and Communications Security, 2015-Octob:1322–1333, 2015. ISSN 15437221. doi: 10.1145/2810103.2813677.

[] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. Proceedings of the ACM Conference on Computer and Communications Security, pages 619–633, 2018. ISSN 15437221. doi: 10.1145/3243734.3243834.

[] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He. Towards Privacy and Security of Deep Learning Systems: A Survey. pages 1–23, 2019. URL `http://arxiv.org/abs/1911.12562`.

[] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In Advances in neural information processing systems, pages 3391–3401, 2017.

[] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. Proceedings - IEEE Symposium on Security and Privacy, 2019-May:691–706, 2019. ISSN 10816011. doi: 10.1109/SP.2019.00029.

[] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting Gradients – How easy is it to break privacy in federated learning? (1), 2020. URL http://arxiv.org/abs/2003.14053.

[] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pages 1310–1321, 2015.

[] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. CoRR, abs/1602.02697, 2016. URL http://arxiv.org/abs/1602.02697.

[] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.

[] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership Inference Attacks and Defenses in Supervised Learning via Generalization Gap. 2020. URL http://arxiv.org/abs/2002.12062.