Draft research question

16 April 2020

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, Microsoft Azure, Google Cloud AI or IBM Watson offer their customers to train their model given they provide the dataset. 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.

Privacy attacks can go from determining whether a data point was used for training (*membership* inference) (Shokri, Stronati, Song, & Shmatikov, 2017; Salem et al., 2019), reconstructing class representatives (*model inversion*) (Fredrikson, Jha, & Ristenpart, 2015), or inferring hidden properties of the training set (*properties inference*) (Ganju, Wang, Yang, Gunter, & Borisov, 2018). 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 minimasation applies?*).

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 (*question: Compared to the other two kind of attacks, is this true? this is what I felt while looking for papers*) on this kind of attack. In (Ganju et al., 2018) 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 (Zaheer et al., 2017) is an effective way to tackle this limitation. In (Melis, Song, De Cristofaro, & Shmatikov, 2019) the authors successfully perform PIA in a collaborative learning setting where the attacker one of the members.
The kind of model used has an impact on the kind of attacks that can be used on it (*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. Geiping, Bauermeister, Dröge, and Moeller analyzed the effect of model architecture on sensibility to model inversion attacks on non-sensible data (cars, animals).

In the context of GDPR sensible data, what is the relationship between model complexity and sensibility to property inference attacks?

References

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the ACM Conference on Computer and Communications Security*, *2015-Octob*, 1322–1333. doi: 10.1145/2810103.2813677

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., & Borisov, N. (2018). Property inference attacks on fully connected neural networks using permutation invariant representations. *Proceedings of the ACM Conference on Computer and Communications Security*, 619–633. doi: 10.1145/3243734.3243834

Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting Gradients – How easy is it to break privacy in federated learning? (1). Retrieved from `http://arxiv.org/abs/2003.14053`

Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. *Proceedings - IEEE Symposium on Security and Privacy*, *2019-May*, 691–706. doi: 10.1109/SP.2019.00029

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., & Backes, M. (2019). ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. (February). doi: 10.14722/ndss.2019.23119

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *Proceedings - IEEE Symposium on Security and Privacy*, 3–18. doi: 10.1109/SP.2017.41

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., & Smola, A. J. (2017). Deep sets. In *Advances in neural information processing systems* (pp. 3391–3401).