Poster: Privacy Risks of Explaining Machine Learning Models

Reza Shokri
School of Computing
National University of Singapore
Singapore, Singapore
reza@comp.nus.edu.sg

Martin Strobel
School of Computing
National University of Singapore
Singapore, Singapore
mstrobel@comp.nus.edu.sg

Yair Zick
School of Computing
National University of Singapore
Singapore, Singapore
zick@comp.nus.edu.sg

***Abstract*—Can an adversary exploit model explanations to infer sensitive information about the models' training set? To investigate this question, we first focus on *membership inference attacks*: given a data point and a model explanation, the attacker's goal is to decide whether or not the point belongs to the training data. We study this problem for three popular types of transparency methods: gradient-based attribution methods, perturbation-based attribution methods, and example-based influence measures. We develop membership inference attacks based on these model explanations and extensively test them on a variety of datasets. In settings where existing attacks based on the loss are infeasible, we show that gradient-based explanations can leak a significant amount of information about the individual data points in the training set. We also show that record-based measures can be effectively exploited for membership inference attacks. More importantly, we design *reconstruction attacks* against this class of model explanations. We demonstrate that they can be exploited to recover significant parts of the training set. Finally, we discuss the resistance of perturbation-based attribution methods to existing attacks and link it to a shortcoming of this type of explanation.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Machine learning models are making increasingly high-stakes decisions in a variety of application domains, such as healthcare, finance, and law [1]; driven by the need for higher prediction accuracy, decision-making models are becoming increasingly more complex, and as a result, much less understandable to various stakeholders. In other words, decision-making models are often 'black-boxes': we have no access to their inner workings, but only to their outputs.

Applying black-box AI decision-makers in high-stakes domains is problematic: model designers face issues understanding and debugging their code, and adapting it to new application domains [2]; companies employing black-box models may expose themselves to various risks (e.g. systematically misclassifying some subgroup of their client base [3], or facing the negative consequences of an automated decision [4]); finally, clients (i.e. those on whom decisions are made) are at risk of being misclassified, facing unwarranted automatic bias, or simply frustrated at their lack of agency in the decision-making process leading e.g. to a right to explanation in the European data privacy regulation GDPR [5].

This lack of transparency has resulted in mounting pressure from the general public, the media, and government agencies; several recent proposals advocate for the use of (automated) *transparency reports* (also known as model explanations in the literature) [5]. The machine learning (and greater CS) community has taken up the call, offering several novel explanation methods in the past few years and big companies like Google [6] and IBM [7] are starting to offer explainable AI as part of their machine learning suites.[1]

Transparency reports offer users some means of understanding the underlying model and its decision making processes[2]. By and large, they do so by offering users additional *insights*, or *information* about the model, concerning the particular decisions it made about them (or, in some cases, about users like them).

Releasing additional information is a risky prospect from a privacy perspective; however, despite the widespread work on the design and implementation of model explanations, there has been little effort to address any privacy concerns that arise due to their release. This is where our work comes in. We begin our investigation by asking the following question.

> *Can an adversary leverage model explanations to infer private information about the training data of the underlying model?*

Our Contributions: We provide a comprehensive analysis of information leakage through feature-based and example-based model explanations. We identify what causes the leakage, and design inference algorithms to identify members of the training set. For example-based explanations, our algorithms can reconstruct a large fraction of the training data. To the best of our knowledge, this paper is the first to analyze the privacy risks of model explanations for the training set of the underlying models.

We focus on two major attacks: *membership inference attacks* [8] to infer the presence of individual data points

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number AISG-RP-2018-009).

[1] See http://aix360.mybluemix.net/ and https://cloud.google.com/explainable-ai

[2] See https://distill.pub/2018/building-blocks/ for a particularly intuitive and interactive explanation method for neural network architectures.

in the training set, and *reconstruction attacks* which aim at recovering the training data points. We analyze feature-based explanation algorithms, with an emphasis on gradient-based methods [9], perturbation-based methods [10], and example-based algorithms, with an emphasis on methods that report influential data points [11].

We analyze what information gradient-based explanations can leak about the training data, and compare our approach to attacks that have access to the model's loss on data points, serving as the strongest membership inference attack.

We show that the **variance of the gradient is a considerable distinguisher between the members of the training set and other data points from the same distribution**. The reason is that as the training algorithm converges, the gradient decreases on all members of the training set, whereas for non-members the variance can remain high.

Our experiments on synthetic datasets show that the dimensionality of the data has a large influence on the connection between the variance of the gradient and the membership inference accuracy. This is related to the observation that very low dimensional data is generally more resistant to overfitting, whereas models on high dimensional data have poorer performance on the test set.

We link the resistance of perturbation based explanations, like SmoothGrad [10], to this type of attack to the fact that they rely on out of distribution samples to generate an explanation. This helps against membership inference, can however also be seen as a major flaw of this type of explanation.

Example-based model explanations provide explain decisions by outputting the most influential data points in the training set for the decision on a particular point of interest. This presents an obvious leak of training data. In particular, membership inference attacks become simple as training points are frequently used to explain their own predictions.

However, a simple attack of randomly querying the model results in poor coverage in terms of reconstructing the training data. This is because a few certain training data records — especially outliers and mislabeled training points — have greater influence over most of the input space. Thus, after a few queries, the set of reconstructed data points converges, recovering no additional points.

We design an algorithm that identifies and constructs regions of the input space where previously recovered points will not be influential. This minimizes re-discovering previously revealed points, thus increasing the coverage of the algorithm. Through empirical evaluations on data with different dimensionality, we show that **an attacker can reconstruct (almost) the entire dataset for high dimensional data**.

For datasets with low dimensionality, we demonstrate our heuristic's adaptability: using recovered points, one can recover up to 25% of the training set. As we show, the graph structure, induced by the influence function over the training set, tends to have a large strongly connected component and the **attacker is likely to recover at least all points in the largest connected component**. Complementary, as unusual points tend to have a larger influence on the training process we show that **minorities have a high risk of being revealed**.

We further study the effectiveness of membership inference attacks based on additional feature-based explanations (including Integrated Gradients and DeepLIFT). These membership inference attacks achieve comparable, albeit weaker, success than gradient-based attacks. We also present the result of studying the influence of dataset size on the success of membership inference for influence based explanations.

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[2] C. Lan and J. Huan, "Discriminatory transfer," *arXiv preprint arXiv:1707.00780*, 2017.

[3] J. Buolamwini, T. Gebru, and A. Hubert, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classificatio," in *Proceedings of the 1st ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)*, 2018, pp. 598–617.

[4] S. Lowry and G. Macpherson, "A blot on the profession," *British medical journal (Clinical research ed.)*, vol. 296, no. 6623, p. 657, 1988.

[5] B. Goodman and S. R. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[6] G. LLC, "Ai explainability whitepaper," 2019. [Online]. Available: https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf

[7] A. Mojsilovic, "Introducing ai explainability 360," 2019. [Online]. Available: https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

[8] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *Proceedings of the 38th IEEE Conference on Security and Privacy (Oakland)*, 2017.

[9] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv preprint arXiv:1312.6034*, 2013. [Online]. Available: http://arxiv.org/abs/1312.6034

[10] D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M. Wattenberg, "SmoothGrad : removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[11] P. W. Koh and P. Liang, "Understanding Black-box Predictions via Influence Functions," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.