Side Channel Attacks on Machine Learning Systems

Haraldur Tómas Hallgrímsson
hth@cs.ucsb.edu
Department of Computer Science
University of California Santa Barbara

June 17, 2018

Abstract

The fields of cryptography and machine learning are historically associated, with pivotal results in each being closely related [1]. In particular, both fields share a common setting of an adversary seeking to "break" a system by efficiently learning some unknown function—an encryption as indexed by a secret key in the case of cryptography, or the target function from inputs to target labels in machine learning. In this paper, we survey these connections between the two fields and consider the problem of side channel attacks on machine learning algorithms in depth.

1 Introduction

The typical adversarial setting in cryptography involves a cryptanalyst seeking to "break" a cryptographic system, that is to learn an unknown function from plaintext to encrypted text from a family of functions as indexed by a secret key [1]. Fortunately, public-key encryption algorithms such as the Rabin Algorithm exist whose security is provably as hard as factorization, rendering the problem for the adversarial cryptanalyst intractable. Unfortunately, this is only for the case of passive attacks—where the attacker only knows the public signature key—and not for active attacks where the signer can be asked to sign specially constructed messages. A correspondence exists in machine learning, where a 'secret key' corresponds to the 'target function' which we seek to efficiently learn with (a polynomial number of) 'membership queries', or values of the unknown function on specific input [2].

A further parallel can be drawn between the fields of cryptography and machine learning. Consider an adversarial attack on a machine learning system in which the attacker seeks to break its *confidentiality* or *privacy*; where confidentiality assumes the model itself represents intellectual property as in financial market systems, and privacy assumes an imperative that the training data for the model not be made public as in medical applications [3].

An overview of the connections and contributions between the fields of cryptography and machine learning is presented in Section 2. This is followed by an outline and analysis of side-channel methods to attack machine learning systems in Section 3, following recent work. These present the problem at hand in Section 3.1, measure how effectively modern neural network approaches memorize their inputs in Section 3.2, and offer provably privacy-preserving methods in Section 3.3.

2 Connections between Crypto and ML

Theoretical results from modern cryptography place hard boundaries on how efficiently a learning algorithm can recover an accurate representation, with virtually all intractability results from Valiant's model, the first serious inroad to computational learning theory, due to cryptography [4]. Michael Kearns' Ph.D. thesis covers many such key results and concepts [2, Section 7], which was followed up by further papers. For instance, a representation-independent hardness result is achieved by polynomial-time reducing the problems of factoring Blum integers, inverting the RSA function, and recognizing quadratic residues into a learning problem [5]. This reduction places the learning algorithm into the shows of an adversarial man-in-the-middle who attempts to learn an inverse of a trapdoor function by selectively choosing messages to encrypt with a given public key.

The advancements made in cryptography due to machine learning are also numerous. For instance, primitives such as pseudo-random bit generators, one-way functions, and private-key cryptosystems can be arrived at by small transformation in standard learning problems [6, 7].

3 Side channel attacks on ML models

3.1 Problem introduction

We consider the problem of a model-inversion attack, as described by [8]. Machine learning algorithms have surpassed trained human predictions in many domains including on medical outcomes and decisions. The best-performing models today are data hungry and require a large number of samples and associated labels to learn an underlying representative distribution. However, a sample from the distribution can be inherently private, such as in the case of the sample being a patient and the label their medical diagnosis or medical history. A model-inversion attack as considered by [8] considers a ML-as-a-service setting[1], in which users can query a website with their information to receive their probabilities of being in the target class or not (e.g. are they sick?), as ascertained by the model. The goal of the adversary is to learn information about which samples where or where not part of the training data (e.g. was John Doe marked as HIV-positive), thus revealing sensitive data.

[1] Such as offered by Microsoft Azure Learning or BigML.

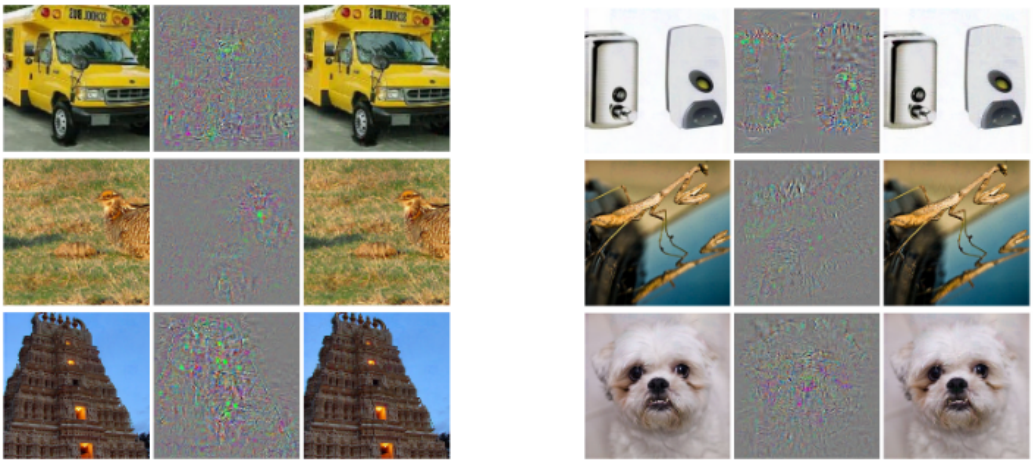

Figure 1: From [11]. Adversarial examples for the AlexNet neural network model which dominated the ImageNet competition in 2012, achieving top-5 error of 15.3% and more than 10.8 percentage points better than the next model. In left columns of (a) and (b), uniformly sampled inputs which are correctly classified, while in right columns the noise in the center columns is added such that they are all predicted to be ostriches: "ostrich, Struthio camelus".

A complete taxonomy of threat models is given by [9], which covers a broader set of attacks that might, for instance, seek to reduce the capabilities of the model for a targeted set of inputs (such that some set of emails will not be marked as spam) or more generally harm the learning process. A general classification of attacks names them as either white- or black-box depending on the amount of information of the model the adversary has access to [3]. A white-box attack might have the entire model—including full details of its architecture and trained weights—while a black-box attack generally considers only having access to the predicted class probability distribution for a given input.

Despite limited information, black-box attacks have shown worrying amount of success in very practical settings [10]. For instance, it has been observed that adversarial examples for machine learning models—inputs which any reasonable human would classify as one thing but have had minute amounts of adversarial noise added to them by backpropagation such that a targeted neural network model would classify them as another targeted class—surprisingly transfer to different neural network models, even when they don't share the same architecture. [11].

The observation that adversarial examples designed to fool one network transfer to also cause difficulty for another model have led to black-box attacks to craft such adversarial examples. It is a simple task to simply train a new model that parallels the targeted model and transfer those attacks over [10]. This concept of training new models, or 'shadow models', for adversarial attacks, has been used for membership inference attacks [12] to determine if a given data input was part of the training process or not.

In the context of sensitive training data, neural network models are known to overfit and predict with much higher confidence on samples they encountered during training (see Section 3.2, below). As such, one can query an available model with various inputs to see when it is likely that, for instance, "John Doe, HIV+" may have been in the training set and not "John Doe, HIV-". A variant of this to discover, via hill-climbing the gradients of initially random inputs, which faces were used as part of the training set was prototyped by [8].

3.2 Measuring memorization of models

The concept of *exposure* was introduced by [13], in which Carlini et al. sought to quantify how much a model overfit on its input data. Using this measure, they define a black-box attack to efficiently retrieve secret training data, and they show how a model trained for only *one* epoch (i.e. the model only saw each input sample exactly once, and not thousands to millions of times as in typical ML training pipelines) has already partially overfit with a null-hypothesis testing p-value of 10^{-30}.

Carlini et al.'s measure of exposure captures the notion of how likely (in terms of log-likelihoods) a given input is according to the model as compared to the distribution of all inputs. For instance, in their use-case of natural language processing, they query to find the likeliest completion of the sentence "My SSN is:", and discover which sequence of nine digits is likeliest conditioned on the model, and accomplish this via efficient generative sampling rather than brute-force.

3.3 Provably preserving privacy

In light of such extremely effective attacks on models, much effort has been put into both practical defenses as well as theoretically understanding of the limitations of such systems [14, 15, 16].

The most compelling of these introduces a system termed PATE [14] (for 'Private Aggregation of Teacher Ensembles') with accompanying differential privacy guarantees of the resulting model. The intuition for their method is that they allow no direct access to the model being trained (the 'student model') on the actual training data. Instead, they train an ensemble of 'teacher models', each on separate disjoint sets of the data. Then, the student model is trained using publicly available and non-sensitive data that is labeled by the ensemble of teacher models using network distillation [17], which forces the student network to match the ensemble—mistakes and all. As the student network never observes any private information, with the sole exception of ensemble-predicted labels (with small amount of noise added for further ambiguity) which are bottlenecked by the number of queries the student makes during training, strong differential privacy guarantees can be shown: the addition or removal of any single sensitive data record (or small number k of private records) has provably no effect on the students predictions. In this manner, even white-box attacks where the adversary has full access to the model, including trained weights and exact network architecture are unable to make membership queries on the training data.

However, the PATE approach comes at a sacrifice of model accuracy. To achieve the differential privacy guarantees requires splitting the training dataset into many much smaller disjoint pieces. Neural network models especially are data hungry, almost strictly improving with each added training sample.

A separate but related process to guarantee is introduced in [18]. This work forgoes the teacher ensemble method and instead uses a noisy stochastic gradient descent algorithm to ensure that no single data input provably effects the learned algorithm. This addition of noise is a common tactic to achieve differential privacy for decades [19], but comes at a cost of learning less from each example. However, as noted by an invited paper by the two sets of authors comparing these two methods [16], the guarantees offered by this method require sophisticated analysis, whereas no such sophistication is required to understand that if 100 independently trained models agree on a classification then that is true regardless of any given sensative input.

4 Conclusion

In this paper we surveyed the many fundamental contributions shared between the two fields of cryptography and machine learning, including how advances in one led to advances in the other. This was followed up by a deep dive into a burgeoning and hot topic in both fields, namely side-channel attacks which seek to infer some hidden information using mechanisms that were unknown to the system designers. In particular, this paper surveyed work covering our current understanding of how machine learning models tend to overfit, what problems that might cause from a privacy perspective, and a sketch of recent attempts to provably alleviate and their downsides.

References

[1] Ronald L Rivest. Cryptography and machine learning. In *International Conference on the Theory and Application of Cryptology*, pages 427–439. Springer, 1991.

[2] Michael J Kearns. *The computational complexity of machine learning*. 1990.

[3] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

[4] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[5] Michael Kearns and Leslie Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1):67–95, 1994.

[6] Avrim Blum, Merrick Furst, Michael Kearns, and Richard J Lipton. Cryptographic primitives based on hard learning problems. In *Annual International Cryptology Conference*, pages 278–291. Springer, 1993.

[7] Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. In *Foundations of Computer Science, 1984. 25th Annual Symposium on*, pages 464–479. IEEE, 1984.

[8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333. ACM, 2015.

[9] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[10] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[12] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 3–18. IEEE, 2017.

[13] Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *arXiv preprint arXiv:1802.08232*, 2018.

[14] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[15] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[16] Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Nicolas Papernot, Kunal Talwar, and Li Zhang. On the protection of private information in machine learning systems: Two recent approches. In *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, pages 1–6. IEEE, 2017.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[18] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[19] Rein Turn and Willis H Ware. Privacy and security in computer systems: The vulnerability of computerized information has prompted measures to protect both the rights of individual subjects and the confidentiality of research data bases. *American Scientist*, 63(2):196–203, 1975.