CS6290: Reading Summary 4

LING Shengchen
Dept. of Computer Science
ID: 57730900

I. SUMMARY OF PAPER [1]

A. Problem Statement

Machine learning models are vulnerable to adversarial integrity attacks. However, current attacks assumed all require knowledge of either internal model structure or training datasets, which overestimates the capabilities of attackers while underestimates the corresponding risks. The paper targets how to compromise integrity without knowledge of information mentioned above as well as potential countermeasures.

B. Problem Significance

Existing researches have exposed the vulnerabilities of classifiers in integrity, assuming that adversaries have access to the internal information of the classifiers, e.g., images or structures. However, for a wider range of adversaries, such a premise may not hold, which is worthy to be taken into consideration seriously to prevent potential attacks likewise. The target models may not be as secure as expected and are feasible to be misled to misclassifications.

C. State of the Art

For the problem mentioned in Section A, researchers in the fields of both security [2][3][4][5] and machine learning [6][7] have been aware of the vulnerabilities, and have dig into such training models. However, in the process of implementations of perturbations, some use internal structure or parameters [4], some use related training sets [2][5], some use both [3][6][7]. Particularly among them, Alexey Kurakin et al. has shown the feasibility in vision classifiers [8] and Mahmood Sharif et al. in facial recognition [9].

D. Contributions

1) *In the perspective of novelty of problem assumption*. The paper sets the problem against a deep neural network (DNN) classifier, and assumes that the adversaries have no knowledge of (a) either the internal structure or parameters of DNN, and (b) any training dataset related to the DNN. The adversaries can only observe the labels, i.e., the indexes with the largest probability results, which are assigned by the DNN with chosen inputs, for the purpose of making adversaries more realistic but weaker: $\tilde{O}(\vec{x}) = arg\ max\ O_j(\vec{x})$. The goal of adversaries is to add a minimally perturbation $\delta\vec{x}$ to input $\vec{x}$, denoted as $\overrightarrow{x^*}$, and misclassified by DNN that $\tilde{O}(\overrightarrow{x^*}) \neq \tilde{O}(\vec{x})$.

2) *In the perspective of novelty of the technical solution.* The paper proposes a novel attack strategy of training a *substitute* DNN of target DNN using a *synthetic* dataset, which is named "black-box attack". More specifically, the inputs are selected by the adversaries synthetically, while the outputs are the labels observed by the adversaries via DNN. Based on such a dataset (inputs and outputs), adversaries build a substitute network, which is an approximation F, to generate adversarial misclassification samples, see Fig.1.

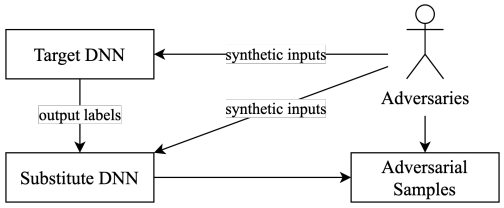

Fig. 1. Black-box attack

In the process of substitute DNN training, the adversaries query the target DNN with inputs that generated by a Jacobian-based Dataset Augmentation heuristic to build the substitute DNN that approximates the target's decision boundaries. The training algorithm first initializes a small set of input domain, selects a potential architecture, and then iteratively refines the substitute using the Jacobian.

In the process of adversarial sample crafting, the paper provides two approaches sharing a similar intuition of model sensitivity evaluation: the Goodfellow et al. algorithm [6] and the Papernot et al. algorithm [4]. The former algorithm is to compute the perturbation $\delta_{\vec{x}} = \varepsilon \cdot sgn\left(\nabla_{\vec{x}} c(F, \vec{x}, y)\right)$, with faster crafting speed but larger perturbations, therefore is easier to get detected. The latter algorithm is suitable for attacks that have target source class, and is to sort the adversarial saliency value $S(\vec{x}, t)[i]$ decreasingly, with smaller perturbations but higher computational costs.

Further, the paper also generalizes such solution to a wider range of machine learning models including logistic regression, decision tree, and SVM.

3) *In the perspective of positiveness of experimental evaluation result.* The paper shows an instantiation of the attack against remote DNN classifiers that are automately trained and hosted by MetaMind, Amazon, and Google, with a satisfying misclassification ratio of 84.24%, 96.19%, and 88.94%, respectively. The threat model pertains to a situation in which end-users are utilizing classifiers hosted by a third-party provider that keeps the inner workings concealed, which mirrors that of real-life circumstances.

E. Remaining Questions

The existing defense mechanism of gradient masking fails in the case of block-box attacks. Despite that the paper proposes two potential directions of adversarial training and defensive

distillation, still no actually effective defense mechanism has been proposed and proven in this paper.

II. SUMMARY OF PAPER [10]

A. Problem Statement

Machine learning models require massive training data from various sources, which arises privacy problems. The paper targets the problem of how to design and implement a machine learning model while preserving data privacy.

B. Problem Significance

Machine learning models are widely used in the fields of health, banking, recommendation, security, etc., fueling advances of images, texts, and speeches. Correspondingly, during the model training process, large amounts of privacy data are collected, stored, and utilized, on purpose of more accurate predictive results. Privacy concerns prevent data owners from willingly sharing their data. People are worried that what if their privacy data are leaked out or maliciously abused by other parties.

C. State of the Art

Existing privacy-preserving works in machine learning mainly focused on decision trees [11], SVM classification [12][13], linear regression [14], and logistic regression [15], but most lack efficiency. Gascón et al. [16] and Nikolaenko et al. [17][18] have shown that multiparty computation (MPC) can provide a promising approach to train two-server models on shared data without revealing any detailed information. However, the performances of both solutions are not satisfying and practical, due to the complicated computations of Boolean circuits. Gilad-Bachrach et al. [19] proposed another secure data exchange solution but lacks scalability.

Wu et al. [20] explored privacy-preserving logistic regression, but their complexity becomes exponential in the degree of approximation polynomial. Aono et al. [21] explored a new model but still leaks aggregated plaintext data to the training clients. Shokri and Shmatikov [22] applied privacy preserving in neural networks by only sharing the changes of coefficients, still not solving the problem of leakage and security.

D. Contributions

1) *In the perspective of novelty of the technical solution.* The paper proposes a new protocol that consists of an online phase and an offline phase. The online phase is only for training the model by integer multiplications and bit shifting, while the offline phase is for generating multiplication triplets. See the protocol design in Fig.2.

In both phases, to reduce complexity and improve efficiency, the protocol uses vectorization in shared settings. To achieve this, the protocol generalizes the operations of addition and multiplications to matrices. Vectorization is also used to optimize the generation of multiplication triplets based on linearly homomorphic encryption (LHE) and oblivious transfer (OT). For the comparison of approaches between LHE and OT, the communication for the latter is higher, yet running time is faster due to cheaper operation cost.

The protocol solves the previous bottleneck of arithmetic computation by setting a couple of shared x and y in a finite field with l_D bits at most in fractional part, letting $x' = 2^{l_D} \cdot x$ and $y' = 2^{l_D} \cdot y$ to transform to integers, performing multiplications $z = x' \cdot y'$ using multiplication triplets, and simply truncating the last l_D bits of z that represent the fractional parts. The products that truncated from shares have tolerable deviation compared to fixed-point arithmetic with high probability. Moreover, the slight truncation error will not influence the final accuracy of the result when the bits numbers are sufficiently large. The truncation solution also works even when z is secretly shared.

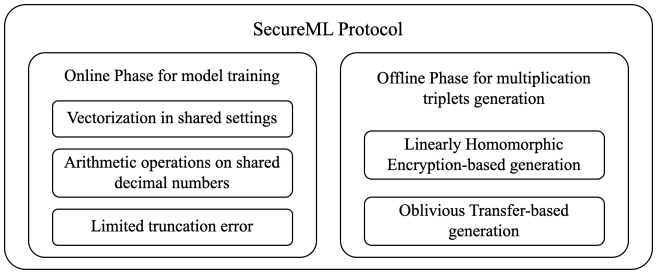

Fig. 2. SecureML protocol design

2) *In the perspective of generalization and extension*. The paper also proposes new activation functions which are MPC-friendly, for both logistic regression and neural networks. Particularly, for logistic regression, the solution can be considered as a combination of two RELU functions with efficient computations to replace the computationally expensive logistic function and softmax functions, see Equation 1. For neural network, RELU function can be used in each neuron as activation function, which cross entropy function can be used as cost function.

$$f(x) = \begin{cases} 0, & if\ x < -\frac{1}{2} \\ x + \frac{1}{2}, & if\ -\frac{1}{2} \le x \le \frac{1}{2} \\ 1, & if\ x > \frac{1}{2} \end{cases} \quad (1)$$

3) *In the perspective of positiveness of experimental evaluation result.* The efficiency of the protocol in linear regression is several orders of magnitude higher than current solutions, i.e., over 1000 times faster, for a dataset with 100,000 samples and 500 features. The paper also implements and evaluates the first privacy-preserving protocol for logistic regression and neural networks, the former gives relatively acceptable result while the latter requires further performance improvements.

E. Remaining Questions

1) For logistic regression, after switching to new activation functions, there are several approaches to compute the backward propagation, which needs further analysis on their accuracies. Further, MPC-friendly activation function may be a promising direction to improve computational efficiency.

2) As the experimental results of logistic regression and neural networks are not as satisfying and optimal as linear regression, new technologies such as Fast Fourier Transform to help improving computional performance are interesting and open questions.

REFERENCES

[1] N. Papernot et al., "Practical Black-Box Attacks against Machine Learning", in *Proc. of ACM on Asia Conference on Computer and Communications Security*, 2017.

[2] B. Biggio, et al., "Evasion attacks against machine learning at test time", in *Machine Learning and Knowledge Discovery in Databases*, 2013.

[3] L. Huang, et al., "Adversarial machine learning", in *Proc. of the 4th ACM workshop on Security and artificial intelligence*, 2011.

[4] N. Papernot, et al., "The limitations of deep learning in adversarial settings", in *Proc. of the 1st IEEE European Symposium on Security and Privacy*, 2016.

[5] W. Xu, et al., "Automatically evading classifiers", in *Proc. of the Network and Distributed Systems Symposium*, 2016.

[6] I. J. Goodfellow, et al., "Explaining and harnessing adversarial examples", in *Proc. of the International Conference on Learning Representations*, 2015.

[7] C. Szegedy, et al., "Intriguing properties of neural networks", in *Proc. of the International Conference on Learning Representations*, 2014.

[8] A. Kurakin, et al., "Adversarial examples in the physical world", arXiv preprint arXiv:1607.02533, 2016.

[9] M. Sharif, et al., "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition", in *Proc. of ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[10] P. Mohassel and Y. Zhang, "SecureML: A System for Scalable Privacy-Preserving Machine Learning", *IEEE Symposium on Security and Privacy*, 2017.

[11] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining", in *Annual International Cryptology Conference*, 2000.

[12] J. Vaidya, H. Yu, and X. Jiang, "Privacy-preserving SVM classification", *Knowledge and Information Systems*, 2008.

[13] H. Yu, J. Vaidya, and X. Jiang, "Privacy-Preserving SVM Classification on Vertically Partitioned Data", *Advances in Knowledge Discovery and Data Mining*, 2006.

[14] A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter, "Privacy preserving regression modelling via distributed computation", in *Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[15] A. B. Slavkovic, Y. Nardi, and M. M. Tibbits, "'Secure' Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases," in *7th IEEE International Conference on Data Mining Workshops*, 2007.

[16] A. Gascón et al., "Privacy-Preserving Distributed Linear Regression on High-Dimensional Data", in *Proc. on Privacy Enhancing Technologies*, 2017.

[17] V. Nikolaenko et al., "Privacy-preserving matrix factorization", in *Proc. of the ACM SIGSAC conference on Computer & Communications Security*, 2013.

[18] V. Nikolaenko et al., "Privacy-Preserving Ridge Regression on Hundreds of Millions of Records", in *IEEE Symposium on Security and Privacy*, 2013.

[19] R. Gilad-Bachrach et al., "Secure Data Exchange: A Marketplace in the Cloud", in *Proc. of ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2019.

[20] S. Wu et al., "Privacy-preservation for Stochastic Gradient Descent Application to Secure Logistic Regression", *the 27th Annual Conference of the Japanese Society for Artificial Intelligence*, 2013.

[21] Y. Aono, T. Hayashi, L. T. Phong, and L. Wang, "Scalable and Secure Logistic Regression via Homomorphic Encryption", in *Proc. of the 6th ACM Conference on Data and Application Security and Privacy*, 2016.

[22] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning", in *Proc. of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015.