SHOULD BE REPLACED ON REQUIRED TITLE PAGE

Instruction

1. Open needed docx template (folder "title"/<your department or bach if bachelor student>.docx).

2. Put Thesis topic, supervisor's and your name in appropriate places on both English and Russian languages.

3. Put current year (last row).

4. Convert it to "title.pdf," replace the existing one in the root folder.

Contents

List of Tables

List of Figures

Abstract

Federated learning though brought lots of benefits via better user privacy by enabling data scientist perform machine learning on user devices without data leaving the device still have some security constraints pertaining to its natural settings. Due to the distributed nature of federated learning, it is usually difficult to detect backdoor attacks with easy to use and practical defence mechanisms. Malicious parties could perform backdoor attacks on federated learning where the goal of the attacker is to make sure that the global model update behaves differently on some specific targeted sub-tasks while keeping a good overall performance on the global model. Unlike previous works, our work propose an algorithm to greatly reduce the effects of backdoor attacks on federated learning based on deferentially private stochastic gradient descent(DP-SGD) with real world dataset considering natural settings of federated learning where there exist many natural challenges like data is usually non iid, non addressable global population, client initiated communications and clients become available or check in at random moments.We also conduct a comprehensive study of backdoor attacks and techniques to defend them for the human activity recognition dataset , under natural settings of federated learning.

Experimental results show that user data is .....(will be completed after collecting results)

Chapter 1

Introduction

The advancement of miniaturized lightweight, intelligent sensors together with an increased internet bandwidth in the recent years has encouraged the development and wide adoption of Wireless medical telemetry systems such as body area network (BAN). BAN is a technology which help reduce doctors burnout in hospitals by reducing the patients workloads in hospitals by help of devices which remotely monitor patients medical activities via either embedding them in the body (Elon Musk brain implant [1]) or placing them on the body surface(wearable technology [2]). These devices offers the potential of great improvement in the delivery and monitoring of healthcare remotely. Wearable devices like smart watches provide potentials in early warnings of some diseases like Parkinson's, vessel related diseases, sleep disorders. They are also used to monitor mental conditions, sleep states, physical activities like sporting activities such as number of steps performed per day , track body temperature , human activity recognition (HAR) and heart beat rate[2][3][4][5].

These devices generate lots of useful data which can be used via the application of machine learning to develop new analytical tools, discovery of

new drugs to improve the quality of patient care. However these data cannot be accessed by data scientists due to the privacy related issues and the default nature of user data which is located on users devices in forms of isolated islands [2]. Some who have the data still face lots of malicious attacks which may lead to data breaches resulting to a lost in user privacy.

In 2016 , Google introduced Federated learning [6][7] to protect the privacy of data owners (i.e the participants) by training machine learning models on distributed heterogeneous data sets across many devices generally known as user equipment (UE). Federated learning make it possible to only exchange model gradients during training rather than exchanging raw data with the server. There is a coordinator also called data facilitator which is responsible for aggregating the participants local models updates to the main global model hence ensuring no participant have access to another participant's private data [8]. Due to it privacy preserving techniques provided by allowing models to train on user data without moving the data to the cloud [6] , it attracted lots of attention from both academia and industry [9]. However, FL still does not guarantee total privacy as there have been lots of attacks on different machine learning models where attackers can extract user data on the model updates directly. Shokri et al. [10] could estimate whether a data set has a specific user data record from the black box of the machine learning model using their attack on membership inference, Zhao et al. [11] could reconstitute individual face from a model performing collaborative learning or federated learning with their technique called improved deep gradient leakage (iDGL). Fredrikson et al. [12] could identify an individual face given the name of the person and the API for the face recognition via their model inversion attack. Jingwen et al. [13] could determine if a data record is in the model's training set during federated

learning via their membership inference attack called GAN Enhanced Membership Inference. In an attempt to protect from these attacks, there were different proposed techniques of data privacy. One of them, data anonymization which includes removing private data and replacing it with random fake values before performing FL. However, this approach was proven not effective to provide good privacy for user healthcare data. Daniel C et al [14], via their linkage attack succeeded to identify the Governor of Massachusetts William Weld's Medical information by combining the anonymized public medical record with the voter records.

Federated learning on wearable healthcare, involving the use of sensitive, private user data needs an extra step to provide privacy to its users. A comprehensive privacy protection technique called differential privacy was proposed by C. Dwork [15][16] which works over the principle of data perturbation by means of adding adequate noise. This technique have been considered the gold standard by many researchers as well as industry as it has been adopted by many companies including Google [17] [18][19], Apple [20], Microsoft[21] and LinkedIn [22] as well as the US Census Bureau [23].

Chapter 2

Literature Review

2.0.1 FedHealth: A federated transfer learning framework for wearable healthcare

They proposed a method which solve two problems: data in wearable devices are found in forms of islands, and model training in cloud fail to personalize as different users have different physical characteristics,So they solve the problem by using federated learning to aggregate user data solving the problem 1 and training on the different personal data while maintaining privacy and use homomorphic encryption to secure.Then uses transfer learning to solve the problem of personalization[1].However their framework doesn't solve the problem data leakage from the model itself as federated learning provide a robust framework for securly training on a distributed approach without specific security on the model or data itself.

2.0.2 Privacy-preserving personalized federated learning

They proposed an algorithm that solved the privacy problems and personalization problem by looking at two specific problems : ML on private data may expose sensitive information, and practical issues that come along with federated learning which uses distributed approach such as user heterogeneity.This means federated learning capture a global knowlege from all participants and fail to capture specific knowledge to each person participating in the training set.They proposed a technique that could solve this problem by proposing a multi-task learning optimization problem which assume that people with close relationships are likely to develop similar habits. So their technique learn not only from all users but also based on their relationship to each other.This enable them solve the personalization problem[2].However their approach make use of gaussian noise which have proven to have poor settings in adding noise in user data and reduces accuracy with respect to laplacian noise which is more convenient while applied to differential privacy.

2.0.3 Privacy-Preserving Federated Learning Framework Based on Chained Secure Multi-party Computing

Federated learning propose a mechanism train a model from a heterogeneous pool of users in a privacy preserving way where users will not have access to each others data.However, the model aggregation experienced in federated learning may have some sensitive data leakages from it.This paper proposed a method to solve this problem.This paper proposed chained secure multi-party computing technique, named Chain- PPFL which is based on two mechanisms: single masking mechanisms which mask individual data protecting individu-

als from accessing each others data; and Chained-Communication mechanism which provide a way to transfer masked data among individuals in a sequential method.

2.0.4 Model Poisoning Attacks in Federated Learning

They mathematically how a malicious agent can poison the global model in terms of misclassification.The agent provide a wrongly classified data in the aim as to corrupt the overall model.However, their research was mostly showing the attack scenario , how to amplify but it provide no solution on how to protect from the attack or how to defend.

2.0.5 Attack of the Tails: Yes, You Really Can Backdoor Federated Learning

They showed that an a backdoor can be included in the recent federated setting with the aim to degrade the performance of the setting if no defense mechanisms on MNIST data sets and proposed a clipping and weak differential privacy mitigate the attack.However, their work did not provide a way to defend from this attack.Their work mostly focuses on the attack.

2.0.6 Can You Really Backdoor Federated Learning?

They showed that FL can be backdoored by a malicious client sending model updates which has as aim to degrade the general model performance.They propose a defence mechanism to ignore model updates from clients whose threshold norm is greater than a certain value M. and some minimum differential privacy to reduce the amplitude of the attack.However their model

doesn't let tracking of these malicious client and discard them and their model is easy to bypass by adversaries since once known about the amplitude and attacker can tune their malicious model to bypass this setting and succeed in its poisoning process.

2.0.7 Privacy amplifications via random check ins

DP-SGD which is popularly known as the fundamental building block of learning over sensitive data usually uses two standard approaches to add some small amount of noise to make its process private as compared to traditional or naive methods.These standards include: privacy amplification by sub-sampling and privacy amplifications by shuffling.How ever these two standards are based on an assumption that elements of the datasets can be sampled uniformly or permuted in a uniform way.This may not pose an issue in traditional machine learning setting but may be problematic in distributed settings or more precisely federated learning where there exist many natural challenges like data is usually non iid, non addressable global population, client initiated communications and clients become available or check in at random moments.This paper proposes the first privacy amplification technique suited for challenges related to distributed or decentralised learning .Their approach focus on an iterative way of performing DP-SGD in distributed settings with datasets located in different clients in the distributed system.Their main contribution is the random check-in protocol which crucially relies on local, random participation decision independent made by each clients .Their method don't require communications initiated by the server or even information about the population size.They also coupled to that shuffling from their results showed that privacy could be improved with fewer magnitude of users.From their results, similar accuracy /

privacy trade-offs to that of privacy amplification by sub-sampling and shuffling.They are focused mostly on privacy , how to make it better and little security precautions are made especially with assumptions of trusted server aggregrator which is not feasible in real world.They also made another assumption , clients are trusted, which is not practical in the real world especially in distributed (federated learning) systems where we don't know our clients.Also their work is largely theoretical and their assumptions does not easily hold in real world system .

2.1 Contributions

The main contributions of this paper can be summarized as follows:

- We provide a realistic real-world technique to perform privacy preservation for wearable healthcare under Federated learning where there exist many natural challenges like data is usually non iid, non addressable global population, client initiated communications and clients become available or check in at random moments.

- We propose a method to generate adversarial input using W-GANs with Gradient penalty which we will use to demonstrate the adversarial attack on federated learning for multivariate time series classification using HAR dataset.

- We conduct a comprehensive study of backdoor attacks for the human activity recognition (HAR) dataset, under natural settings of federated learning.

- We propose an algorithm to greatly reduce the effects of backdoor attacks

on federated learning based on deferentially private stochastic gradient descent(DP-SGD) with a real-world dataset considering natural settings of Federated learning.

- We propose a technique based on adversarial learning to train our discriminator to detect adversarial inputs even before they are aggregated to the global model.

Chapter 3

Methodology

3.1 Federated learning and backdoor attack

3.1.1 Federated learning with non malicious users

Federated learning offers lots of opportunities since we can securely train on data without getting the data from the users. Although federated learning offers lots of privacy from the users as their data don't leave their respective devices much past literature has shown that we can backdoor federated learning leading to poor misclassification for specific classes (backdoor) or poor model performance (poisoning). In this paper, we will focus on backdoor attacks based on model replacement. Consider we have N clients to participate in Federated learning. Each client holds a data record d_i forming a dataset D = { d_1, d_2, ... d_n}. The coordinating server trains on D via series of rounds, $t \in T$ with some loss function J_θ.

At each round t , a certain $K < N$ number of clients randomly check-in to the server. The server randomly send model updates, w_i^t to a certain $C.K$ clients, where $C < 1$. This creates a new set, S_t of clients who will participate

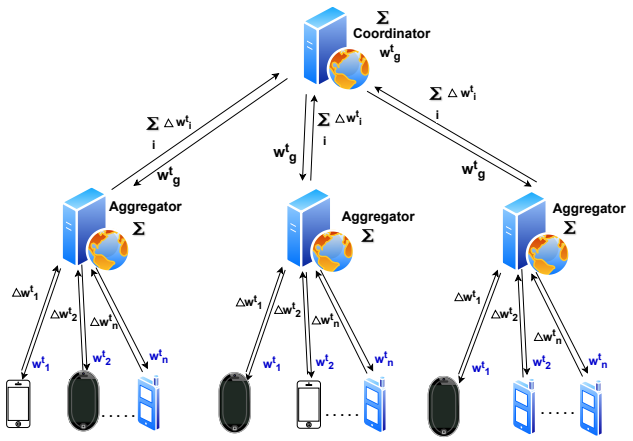

Figure 3.1: Federated learning with benign clients

in training the global model w_g^t by sending their computed model update Δw_i^t their models based on their respective data d_i. At t, each client in S_t produces m_i model samples. The coordinator updates it model by aggregating model update from the different clients as seen in equation 3.1 below.

$$w_{g+1} = w_g + \eta \frac{\Sigma_{i \in S_t} m_i \Delta w_i^t}{\Sigma_{i \in S_t} m_i} \tag{3.1}$$

Where η is the server's learning rate.

3.1.2 Threat Model

The natural setting of federated learning gives users full control of the training process. The user locally trains the model w_i^t, sent by the coordinator. During training, the client uses the model w_i^t to train on its local data d_i , divided in to a series of batches $b \in B$. This produces the model w_i^{t+1} as seen in Algorithm 1. This means that the coordinator has no control over the training process and only aggregates final user model updates, $\Delta w_i^t = w_i^{t+1}$ - w_g^t. Under normal Federated learning with benign users, the aggregators aggregates the models and send to the coordinator which update the model before sending

back to the clients for the next iteration process as seen in diagram 3.1. A user whose device (e.g. smartwatch, smartphone, smart speaker, etc) has been compromised using a malware attack will have this process completed in the hands of the attacker. The attacker can replace the user's local data, the model hyper-parameters (epochs and learning rate α), modify different model parameters like weights even before submitting to the coordinator.

3.1.3 Model Replacement for adversarial backdoor injection

Let's assume a certain fraction of clients $\epsilon \leqslant C.K$ were compromised and their data replaced with adversarial data to produce backdoored model update $\Delta\ w_b^t$. The main objective of the malicious client is to replace the global model with the backdoored attacker model which will cause targeted misclassification for a specific subclass but perform well on the general classification. Based on that, backdoor attacks on federated learning are very difficult to detect. This backdoor only gets triggered when the input data has specific triggers causing the data to be misclassified. For example, a backdoor attack in the vision domain will misclassify green cars to a bird each time a green car is used at the input. For this case, the trigger is color green. This work will focus on backdoor injection based on model update poisoning proposed by [24], [25]. We assume for each round t, there exist a certain number of adversaries ϕ^t ranging between 0 and $\epsilon = \min\ (0,\ C.K)$. During each round, the attacker attempts to replace the global model with a backdoored model, Δw_{adv}^t, equation 3.2 based on [26], by sending it to the coordinator as shown in 3.2.

$$\Delta w_{adv}^{t} = \gamma(w_{adv}^{t+1} - w_{g}^{t}) \tag{3.2}$$

Where $\gamma = \frac{\Sigma_{i \in S_t} m_i}{\eta m_i}$, is the boost factor. We boost the malicious agent updates to overcome the backdoor canceling effects coming from aggregations with a large number of benign users.

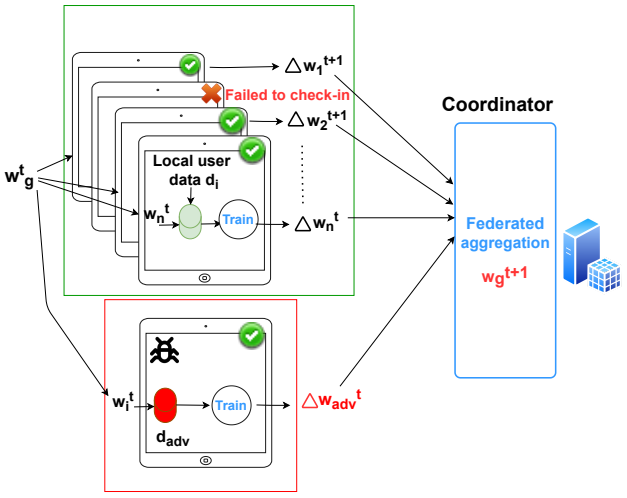

Figure 3.2: Federated learning with malicious clients

A paper by Arjun et al. [25] showed that using boosting, even with a single agent controlled by a malicious attacker, the attacker can inject the backdoor and make the model converge with 100% confidence. However, this process has lots of challenges that need to be overcome by the malicious client.

3.1.4 Challenges faced by malicious clients

Federated learning enables model train on distributed data from clients. This makes it difficult for a client to access the data or the model updates of another client. This makes it impossible for the adversary to access data or Poisson model updates from other clients. Also, federated learning train on many users,[7] usually in orders of $N = 10^7$ or more. This makes poison-

Algorithm 1 client's local model training process

Initialize local model w_i^t and loss function L_i
$w_i^{t+1} \leftarrow w_g^t$
$L_i \leftarrow J_\theta$
for epoch $t \in T$ **do**
  for batch $b \in B$ **do**
    $w_i^{t+1} \leftarrow w_i^{t+1} - \alpha \nabla L_i(w_i^{t+1}, b)$
  end for
end for

ing ineffective since averaging from many clients could cancel the effect of the backdoor. The fact that the coordinator randomly selects a certain number of clients who randomly check-in makes it difficult for the attacker to predict or select clients who will participate in the training.

Chapter 4

Implementation

4.1 Time Series data generation using GANs

4.1.1 W-GAN Wearable time series

Many attacks done on Machine learning are mostly in the vision domain. However, attacks on the non-vision domain in a federated learning setting are very rare. This section will focus on using W-GANs to create backdoored time series data which we use to perform adversarial learning in order to prevent backdoor attacks in federated learning. ...

Chapter 5

Evaluation and Discussion

...

Chapter 6

Conclusion

...

Bibliography cited

[1] E. Musk *et al.*, "An integrated brain-machine interface platform with thousands of channels", *Journal of medical Internet research*, vol. 21, no. 10, e16194, 2019.

[2] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "Fedhealth: A federated transfer learning framework for wearable healthcare", *IEEE Intelligent Systems*, 2020.

[3] L. Quintero, P. Papapetrou, J. E. Muñoz, and U. Fors, "Implementation of mobile-based real-time heart rate variability detection for personalized healthcare", in *2019 International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2019, pp. 838–846.

[4] J. Feng, C. Rong, F. Sun, D. Guo, and Y. Li, "Pmf: A privacy-preserving human mobility prediction framework via federated learning", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–21, 2020.

[5] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong, "Privacy-preserving personalized federated learning", in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, IEEE, 2020, pp. 1–6.

[6] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency", in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. [Online]. Available: https://arxiv.org/abs/1610.05492.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data", in *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 1273–1282.

[8] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečn, S. Mazzocchi, H. B. McMahan, *et al.*, "Towards federated learning at scale: System design", *arXiv preprint arXiv:1902.01046*, 2019.

[9] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey", *IEEE Communications Surveys & Tutorials*, 2020.

[10] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models", in *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2017, pp. 3–18.

[11] B. Zhao, K. R. Mopuri, and H. Bilen, "Idlg: Improved deep leakage from gradients", *arXiv preprint arXiv:2001.02610*, 2020.

[12] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures", in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[13] J. Zhang, J. Zhang, J. Chen, and S. Yu, "Gan enhanced membership inference: A passive local attack in federated learning", in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, IEEE, 2020, pp. 1–6.

[14] D. Barth-Jones, "The're-identification'of governor william weld's medical information: A critical re-examination of health data identification risks and privacy protections, then and now", *Then and Now (July 2012)*, 2012.

[15] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation", in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Springer, 2006, pp. 486–503.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis", in *Theory of cryptography conference*, Springer, 2006, pp. 265–284.

[17] A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld, "Prochlo: Strong privacy for analytics in the crowd", in *Proceedings of the Symposium on Operating Systems Principles (SOSP)*, 2017, pp. 441–459. [Online]. Available: https://arxiv.org/abs/1710.00901.

[18] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response", in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 2014, pp. 1054–1067.

[19] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, S. Song, K. Talwar, and A. Thakurta, "Encode, shuffle, analyze privacy revisited: For-

malizations and empirical evaluation", *arXiv preprint arXiv:2001.03618*, 2020.

[20] Apple. (2017). Learning with privacy at scale, [Online]. Available: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

[21] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately", *Advances in Neural Information Processing Systems*, vol. 30, pp. 3571–3580, 2017.

[22] R. Rogers, S. Subramaniam, S. Peng, D. Durfee, S. Lee, S. K. Kancha, S. Sahay, and P. Ahammad, "Linkedin's audience engagements api: A privacy preserving data analytics system at scale", *arXiv preprint arXiv:2002.05839*, 2020.

[23] Y.-H. Kuo, C.-C. Chiu, D. Kifer, M. Hay, and A. Machanavajjhala, "Differentially private hierarchical count-of-counts histograms", *arXiv preprint arXiv:1804.00370*, 2018.

[24] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning", in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 2938–2948.

[25] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens", in *International Conference on Machine Learning*, PMLR, 2019, pp. 634–643.

[26] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?", *arXiv preprint arXiv:1911.07963*, 2019.

Appendix A

Extra Stuff

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Appendix B

Even More Extra Stuff

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.