



MACQUARIE
University
SYDNEY • AUSTRALIA

Final Report
on
Analyzing Privacy Preserving Deep Learning Techniques

Organization: Optus Macquarie University Cybersecurity Hub

Internee: Masroor Fattah Bin Hossain

Student ID: 45765758

Session: 1

University: Macquarie university

Submission Date: 6/06/2021

2. Acknowledgement

First, I would like to thank Dr Amin Beheshti for arranging this nice opportunity of internship which to my view, is a milestone of my master's program. I wish to acknowledge my sincere gratitude to Optus Macquarie University Cybersecurity Hub for giving me the opportunity for internship. I would also like to thank my company supervisor, Dr Muhammad Ikram for guiding me through the tasks of this internship project. His valuable suggestions and comments helped me staying on track with my project. Thanks are due to my two co-supervisors Dr Zhigang Lu and Dr Giacchino Tangari for their continuous guidance with my work; especially, the relentless and rigorous guidance from Dr Zhigang Lu was instrumental in completing the otherwise complex task. Last, but not the least, I am thankful to all my teachers and the staff at Optus Macquarie university Cybersecurity Hub who have been always helping and encouraging me throughout my internship.

3. Executive summary

The internship project aims at exploring privacy preserving deep learning techniques for data and user privacy protection. The research starts with investigation of deep learning model frameworks for protection against malicious attack of user image/identity through various image perturbation techniques. Two adversarial attack protection framework namely fawkes and shield, appears to be stable with their protection outcomes showing potential for public domain and industrial applications.

Subsequently, the project deals with investigation of differential privacy techniques for protection against membership inference adversarial attacks. The running of available deep learning model source code give detailed understanding of the model scheme and working processes. It is revealed that in the scenario of availability of various trained models' blackbox outcome, the adversarial attack models can successfully leak the membership of target models' training dataset.

The results from model framework run using location dataset shows that the non-private model has a high prediction accuracy but a high privacy leakage showing vulnerability of data and user privacy. With introduction of differentially private budget parameter, the privacy leakage can be arrested significantly; budget parameter of 0.01 can totally nullify privacy leakage but with a sacrifice of model prediction accuracy whereas budget parameter of 1000 can arrest some privacy leakage with little sacrifice of model prediction accuracy. This means that there should be an optimum trade-off value of budget parameter with acceptable privacy leakage and model accuracy.

Using the same training dataset, Google AI is trained in vertex AI platform. Later, using the same test dataset, batch prediction from Google AI model is taken for prediction score results. These prediction scores are passed to membership inference attack(MIA) model to mimic the Google AI model as a potential target model. MIA model results reveal that Google AI model is prone to very high leakage rate of 0.908. Introduction of privacy budget parameter value as high as 1000 can significantly reduce privacy leakage rate.

This illustrates that Google AI model can be better off for privacy vulnerability with inclusion of differential privacy scheme in the model.

Then, my own target model code is configured with Google AI model default parameters and target model is run in Google cloud platform to take output of training accuracy, test accuracy and prediction scores. Using the prediction scores, MIA model is trained in local machine for calculating privacy leakage rate. Illustration of results reveal that average accuracy loss of Google configured target model is reduced with increased budget parameter values. The privacy leakage versus budget parameter plot for target model with Google default configuration reveals that at lower budget parameter values, all the privacy schemes are equally effective in reducing privacy leakage while at higher budget parameter values, **adv_cmp** scheme is most effective followed by **dp** scheme.

As the organization I am working in is primarily dealing with the areas of machine learning, security and privacy protecting models with application in various large databases, my work and learning nicely complement each other. My future plan is to apply the same model for a new data set and develop similar output for the new dataset. I will edit and illustrate the results in a paper form to publish in a conference or journal. After that, I have plan to work with my company supervisor and colleagues to develop a framework for improving data and user privacy in cloud based AI models.

Table of Contents

Title Page.....	1
Acknowledgement.....	2
Executive Summary.....	3
Overview of the Organization.....	7
Brief History.....	7
Introduction of the organization.....	7
Policy of the organization.....	7
Competitors.....	8
Organizational Structure.....	8
Organizational Hierarchy Chart.....	8
Fig 1.....	8
Number of Employees.....	9
Main Offices.....	9
Introduction of all the departments.....	9
Comments on the organizational structure.....	10
Plans of my internship program.....	10
Brief introduction of my department.....	10
Starting and ending dates of my internship.....	10
Part-time/Full-time.....	10
My Training.....	11
Training Program.....	11
Detailed Description of the department's operations and activities.....	10
Research Program.....	11
Data Security and Privacy.....	11
Detailed Description of my Project.....	13
Reflective Journal Entries.....	14
Week 1.....	14
Week 2.....	14

Week 3.....	15
Week 4.....	15
Week 5.....	16
Week 6.....	17
Week 7.....	18
Week 8.....	20
Week 9.....	22
Week 10.....	24
Week 11.....	26
Week 12.....	28
Work Samples.....	31
Sample 1.....	31
Fig 2.....	32
Fig 3.....	32
Sample 2.....	32
Fig 4.....	33
Fig 5.....	34
Fig 6.....	34
Sample 3.....	34
Fig 7.....	35
Fig 8.....	35
Sample 4.....	35
Fig 9.....	36
Fig 10.....	36
Critical Analysis.....	37
SWOT Analysis.....	39
Table 1.....	39
Conclusion.....	42
Recommendation.....	43
References & Sources.....	44

5. Overview of the organization

Brief History

Optus Macquarie university Cybersecurity Hub was established in 2016 as a joint initiative with an investment of AUD10 million. The Cybersecurity Hub utilizes core strength of Macquarie university interdisciplinary research to facilitate the cutting-edge research output for industries such as Optus and other government agencies.

Introduction of the organization

The Optus Macquarie University Cyber Security Hub promotes a uniquely interdisciplinary approach to tackle real-world challenges in cyber security. It forms a network of researchers, academic and business leaders. It is a platform for experts, researchers and tech leaders to act jointly for developing cyber security technology, digital governance and technology policy initiatives. Minimizing the pitfalls of human factors through digital technology initiatives is another focus area of the Hub. The scope includes conducting interdisciplinary research across several disciplines: computing, engineering, business, criminology, law and psychology. It also trains the next generation of cybersecurity specialists, and it makes industry leaders conscientious. Last, but not the least, it develops the skills of the existing workforce. It deals with developing commercial technology solutions for cybersecurity challenges with a focus on enabling trust in the digital economy; and core projects of the Hub involve building trust in Data, Technology, Systems and People.

Policy of the Organization

The policy of Optus Macquarie university cyber security hub is to develop, promote and to guide implementation of cybersecurity technologies. It has also policy to develop network among relevant stakeholders and develop skilled manpower in the related areas.

Competitors

- Other government and private corporate cybersecurity research houses.
- Other university cybersecurity research initiatives.

The details of Optus Macquarie university Cybersecurity Hub can be found in this link (<https://www.mq.edu.au/partner/access-business-opportunities/innovation-entrepreneurship-and-it/optus-cyber-security-hub>).

6. Organizational Structure

Organizational Hierarchy Chart



Fig 1: Organizational Hierarchical chart of Optus Macquarie University Cybersecurity Hub

The organization is headed by an executive director which currently is chaired by Professor Dali Kaafar. He is also a primary chief investigator of the Hub. There are four departments in Optus Macquarie university Cybersecurity Hub, and they are data security and privacy, business, criminology and intelligence and law and psychology. Each one of them is headed by a program director. Faculty members from various departments of Macquarie university are the primary chief/chief/associate investigators of the Hub. The

various departments of Macquarie university collaborating with the current organization are department of Computing, department of security studies and criminology, Macquarie business school, Macquarie law school, department of marketing, department of management, department of applied finance, department of accounting and corporate governance, Australian institute of health innovation and deputy vice chancellor (corporate engagement and advancement). Each division has a number of research fellows, research associates, phd students and other graduate research students.

Number of Employees

There are about 117 employees working for the organization including both parttime and fulltime staff.

Main offices

The main office of Optus Macquarie university Cybersecurity Hub is situated in North Ryde campus of Macquarie university.

Introduction of all the departments

-Data Security and Privacy: This is the major division of the Hub undertaking research in cybersecurity, data security and user privacy areas. The division research is led by eight faculty researchers. Associated with them are 20 research fellows who are mid-level researchers working under the guidance of faculty researchers. There are also 30 graduate researchers working for various projects taking guidance from faculty researchers and research fellows. This is the division where I am working for the internship.

-Business: This is another division of the Hub conducting research and development activities in tech business area and technopreneur development. As the Hub deals with rapidly changing technology area and related businesses, it is required to develop sustainable tech business vision for the industry. This is where this unit is leading with its activities.

-Criminology and Intelligence: This division deals with research and development activities in the areas of criminology, financial fraudulence, artificial intelligence and forensic matters.

-Law and Psychology: This division deals with digital act, ICT law and security, ICT legal frameworks, social media and mental health monitoring.

Comments on the organizational structure

The organization structure is of corporate nature although corporate environment and research culture are not quite matching. The financial operation of the organization is not clear from the organization structure.

7. Plans of your internship program

Brief introduction of my department

I am doing my internship under data security and privacy department of Optus Macquarie university Cybersecurity Hub. This is the major division of the Hub undertaking research in cybersecurity, data security and user privacy areas. The division research is led by eight faculty researchers. Associated with them are 20 research fellows who are mid-level researchers working under the guidance of faculty researchers. There are also 30 graduate researchers working for various projects taking guidance from faculty researchers and research fellows.

Starting and ending dates of my internship

I am doing my internship with Optus Macquarie university Cybersecurity Hub. Dr. Muhammad Ikram is the supervisor of my internship. I started my internship on February 22, and I will complete my internship on 6th June, 2021.

Part-time/ Full-time Internship

I am a Master of Data Science student doing my final semester. As part of my course requirement, I am doing part-time internship.

My Training

I am getting training from data security and privacy department of Optus Macquarie university Cybersecurity Hub. Every week, I have a meeting with my supervisor/co-supervisor who discusses with me about the undertaken project steps, solution schemes, outstanding issues, progress rate, etc and asks me questions to clear all my doubts regarding the project progress and activities. My training started on 22nd February, 2021 and it is planned to finish on 6th June, 2021.

8 Training Program

Detailed description of the department's operations/activities

Research program

Data Security and Privacy

This program is headed by Professor Dali Kaafar and the research projects under the program are led by Dr Muhammad Ikram and Dr Hassan Asghar. The program investigates in the area of data security and privacy. It also deals with development of machine learning models and Big Data Analytics. A section of the program investigates around development of privacy preserving system for data sharing; it also works with reliable machine learning models with applications for datasets from health sector. Another section works with design and system networks for supporting data privacy and security. The section work includes data driven analysis, development of machine learning algorithms for improving privacy and security across web and mobile system. It also investigates into detecting security known compliance incorporate enterprise and network. Some of the project details of the program are as follows:

Characterising and measuring misbehaviour in online platforms:

The project addresses the problem of understanding and characterizing the online hateful behavior to find out the owner of profiles undertaking such activities to decide on allowing monitoring surveillance and subsequent deletion of harmful

profiles. The project also deals with analyzing and characterizing the privacy and security matters in web based online games.

Privacy-preserving Web browser plugin for online data:

This project deals with developing a web browser plugin for real-time privacy risk prediction blurring of web data. The developed framework will be resistant to adversarial attacks even if the adversary poses the knowledge of the model and calculated probabilities can make inferences about the real data and the blurred data.

A data-driven analysis tool to investigate content dependencies and chains of malicious content injection on the web:

This project aims at developing and implementing a data driven analysis to characterize the implicit trust in the chain of dependency where first party is assumed to be dependent on second party and third parties websites (through trusted link of second part websites). The work aims at measuring the risk that the first party website may be patching while downloading resources from possible malicious sites.

Titles of other important research activities:

- **Mobile Apps security and privacy-** This project involves data science, machine learning, measurement and system knowledge.
- **Privacy-preserving data sharing and variants of differential privacy-** This project involves information theory, machine learning, statistics and theory.
- **Web security and malware Analysis-** This project involves data science, machine learning, measurement and system knowledge.
- **Network Security modelling and SDN enabled network intelligence:** This project involves machine learning, measurement, networks and system knowledge.
- **Next-generation authentication technologies:** This project involves applied cryptography, information theory, machine learning, statistics and system knowledge.

Detailed description of my project

The objective of this project is to explore the limitations of the existing privacy solutions. Sitting on the side of an unauthorized deep-learning model, the student/intern will design and test techniques for exploiting the user information even when privacy defenses are in place. The findings of this project will be used to investigate enhanced privacy defense mechanisms.

The scope of this project is to explore differential privacy and membership inference attacks. With a theoretical understanding of how attack models work to leak the differentially private algorithms, the evaluations will be made on available adversarial attack models. The evaluation exercise utilizing the available established data set will reveal the scope and limitations of the available models. Suggestions for new strategies for improving the model performance will be made from the analytical investigations of the model results.

Deep learning has recently emerged as a great fit for a range of critical applications on personal users' data, such as face recognition. The proliferation of deep-learning models poses a real threat to personal privacy as model developers/owners may use individual's information as part of their model training sets in an unauthorized way, i.e., without informing them. For example, model developers/owners can fetch someone's face image from a social network, and use it to train sophisticated surveillance systems. To protect the user's privacy i.e. user's data, differential privacy is introduced which works against the adversarial deep learning models.

My work scope also includes evaluation of privacy vulnerability of cloud based deep learning i.e. machine learning models on platforms like Google cloud. Google cloud AI models are popularly used by businesses and organizations for business intelligence, feature inferencing and process optimization. Therefore, cloud based AI models are commonly trained by agencies own data. If these cloud based AI models are not secured enough for member privacy leakage, the companies are automatically putting their customers privacy at risk.

My job is to train such cloud based AI models with my data and later on, using the results of training and subsequent testing, I need to explore the privacy leakage potential of such AI models.

.

9. Reflective Journal Entries

Week 1: In week 1, I had an initial meeting with my co-supervisor, Gioacchino. I collected three papers that were suggested by my co-supervisor, Gioacchino. The three papers were radioactive data: tracing through training (Sablayrolle et al.,2020), protecting privacy against unauthorized deep learning techniques (Shan et al.,2020) and shield, fast, practical defense and vaccination for deep learning using JPEG compression (Das et al.,2018).

What I experienced in week 1 is that initially, I had a difficulty in finding the papers, but with the help of a co-supervisor, I was able to find the relevant papers for my research. In that week, I gained knowledge of how to collect the papers, and in the next week, I reviewed the papers.

Week 2: In week 2, I reviewed three papers which I collected in week 1. The paper radioactive data: tracing through training deals with tracing whether a particular image dataset privacy is breached through excess of any deep learning model. For tracing purpose, a scheme called radioactive data was proposed which makes unnoticeable changes to the image. Any model trained on this perturbed image will patch an identifiable mark, which will indicate that the data was earlier used by a model. The second collected paper was on Fawkes system. The system makes an unnoticeable change in the pixel level of an image which is a process called cloaking. These cloaked images generate functional models that cause images of the user to be misidentified as the cloaked images and the real images look almost similar, and the model finds no labels associated with cloaked images. The reason behind finding no labels is that cloaking significantly changes the feature space of an image before it gets trained by a model with resulting changes in decision boundaries. Last, but not the least, according to the third collected paper, the images can be compressed using JPEG and can be put in a Shield defensive framework before it can be trained by a deep learning model. This ensures privacy protection of a person's image as it prevents all sorts of pixel rearranging that are performed by an attacker which the deep learning model is unaware of. Throughout week 2, I gained the knowledge of different frameworks for privacy protection. Initially, I had difficulties in

understanding the paper, but with the help of Dr Gioacchino Tangari, I was able to understand the papers. Now, I think if I work in any company which deals with the privacy of images, I can come up with the idea of frameworks that I mentioned of in order to maintain the privacy of images.

Week 3: In week 3, I revised the three papers that I reviewed in week 2. As part of my internship requires understanding of deep learning techniques, I have revised and reviewed some of the knowledge repository related to deep learning techniques from the book (Goodfellow et al.,2016). There are three parts i.e. three chapters in the reviewed deep learning book. For week 3, I reviewed first part and some chapters from the second part of deep learning techniques book, and I also reviewed a paper pertaining to a review of deep learning security and privacy defensive techniques (Tariq et al.,2020). According to my review, linear algebra tool can help in preserving privacy of an image using deep learning techniques. Probability theory plays an important role in making decisions based on doubtful information. Thus, probability theory can be used when classifying an image in deep learning models. However, the probability theory is suitable for binary problem. Also, taking advantage of information theory, deep learning models can decide on a type of probability distribution and thus, able to compare between two probability distributions. The literature shows that Deep learning has useful applications in the areas of cybersecurity and image recognition. While Deep learning models are good at solving complex problems, it may contain sensitive user information, and thus act as leakage avenue to malicious attack. Throughout week 3, I gained the knowledge of deep learning frameworks, and thus, I became more proficient in machine learning which is necessary to build my career in data science.

Week 4: In week 4, I reviewed some chapters from the second and third part of deep learning book (Goodfellow et al.,2016). I also reviewed a paper pertaining to a review of deep learning security and privacy defensive techniques. Deep learning sometimes require regularization. Regularization mechanism may be used in order to have reduced variance of the deep learning estimates through trading increased bias; but regularization strategy especially with parameter value boundary and extra term in objective function should be carefully decided to avoid extra constraints on the deep learning models.

Proper algorithms can be used to initiate change in neural network attributes of weight and learning for the optimization of deep learning models with minimization of losses. There are many deep learning architectures such as CNN, ResNet-50 and recurrent and recursive nets. CNNs have a grid alike topology arrangement using convolutional layer and pooling layer; ResNet-50 is a deep convolutional neural network composed of 50 layers utilizing a scheme called skip connection to train the model ,and unlike CNN, Recurrent neural networks can utilize temporal information of sequential data through a mechanism called activation function. Before applying deep learning on the dataset, we need to keep all those factors like regularization, optimization and security so that we can ensure a good accuracy with minimized cost. After studying about deep learning, I tried to build a model data structure with respect to my given internship proposal. With respect to my reviewed paper radioactive data: tracing through training, I installed radioactive tracing: python modules in python environment with all required support packages/ modules like torch, torchvision, pillow, labelimg etc. But model running is having some issue with input data as I haven't managed to get hold of original data used by the relevant research group. I have contacted the research group keyperson for obtaining the original dataset but they said that they can't share the dataset. For my review paper on Fawkes, Fawkes model, an executable window version is already installed which can do protection cloaking to a given image. Also, Fawkes python package is installed in python environment. Throughout week 4, I learned how deep learning model can be used to classify an image. I also learned about different deep learning architectures and their regularization and optimization techniques. I learned use of different deep learning module packages like torch, torchvision, pillow, labelmg, tensorflow, tfnightly etc. I also gathered experience of running the code in pycharm, jupyter notebook and google colab and having an issue with the packages.

Week 5: In week 5, my new co-supervisor, Dr Zhigang Lu instead of Gioacchino, has changed my topic slightly from privacy preservation of an image to privacy preservation of an individual's dataset using differential privacy technique. Therefore, I reviewed differential privacy from Wikipedia ("Differential Privacy," n.d.). I also reviewed two papers on differential privacy as following.

The title of two papers are 'Differential privacy: a primer for a non-technical audience' (Wood et al., 2018) and 'Calibrating noise to sensitivity in private data analysis' (Dwork et al., 2006). According to my review, differential privacy is a technique used to maintain individual privacy after release of the statistics of a group where individual is a member. Differential privacy reduces privacy risks. It is immune to adversarial attacks. A parameter, epsilon is used in Differential privacy model to trade off between model accuracy and privacy loss. A smaller epsilon means there is a small deviation between real life data set and new data set with individual data missed; giving almost similar estimation for two datasets, and hence, no inference is possible with breach of security. Thus, a smaller epsilon will provide a higher privacy protection although it will result less accuracy or utility in terms of estimation clarity. On the other hand, a larger epsilon will provide more accuracy or utility with risk of lower privacy protection. Also, to avoid the use of fixed epsilon value as the difference, a random noise term is added to provide better privacy protection preferably following a Laplace distribution. Throughout week 5, I learned about differentially private mechanism and its important component. I also gained knowledge of cryptography, and I became proficient in python and deep learning. I also learned about the application of differentially private mechanism. The only difficulty that I had in that week is as my co-supervisor is changed, I had to shift from privacy preservation of an image to privacy preservation of an individual's dataset using differential privacy. I again had to read the papers and links that was suggested by my mentor. Although, it was a difficult experience, I solved it with the help of Dr Zhigang LU.

Week 6: In week 6, I learned about membership inference attacks against machine learning models (Shokri et al., 2017). Membership inference is a classification task to determine which set of a category a particular data point belongs given the estimation model is trained on a set of data for which membership category belongs. But knowing an individual data belonging to a certain category is a breach of individual privacy. To know whether an individual data is a member of certain category or not (binary problem), membership inference attack models are at large. Shokri et al. (2017) illustrated an attack model for a certain membership class which is trained on the inputs and outputs of a bundle of shadow models, where each shadow model individual training dataset are input for output vectors which are used as attack models training dataset. Also, each shadow

model is given input of separate test dataset for output vectors which are again used to train the attack model. The researchers have used six different datasets including well known Cifar-10 and Cifar-100 datasets to experiment the effect of training dataset size on the attack model accuracy. The number of shadow models in their experiment with different datasets vary in the range of 10-100 with the inference that accuracy model increases with the number of shadow models , albeit requiring additional cost. During this week, I learned the mechanism of membership inference attack model process. My co-supervisor Dr Zhigang Lu has provided me a github link for membership inference attack models source code and relevant datasets. During the coming week, I will run the model to replicate the result for further illustration. (<https://github.com/bargavi/EvaluatingDPML>)

Week 7: In this week, I cloned the code from the github repository, and then, ran the membership inference code on my local machine with the provided dataset i.e cifar_100 in the github link and also with the new dataset i.e. bangkok_shokri that Dr Zhigang Lu gave to me. In order to run the code, I had to install new packages which was bit challenging for me to explore. But eventually, I could run the code successfully on Cifar_100 dataset. Then, as per my co-supervisor's instruction, I tried to run the code with bangkok_shokri_ dataset. Running the code with Bangkok_shokri was very difficult for me because I couldn't match the format of the dataset with Cifar-100 as the new dataset is just a csv file with binary numbers. Also, in the github link, there is no mention of how Cifar_100 was preprocessed. Thus, initially, I just used pandas to load the file in a separate jupyter notebook, and then, I separated the features and labels for which I made two separate pickle files. I tried to use those two separate files in order to run the code as per the instruction given in the readme file, but then, I saw that the dataset couldn't split into training and testing set for target and shadow models which was crucial to avail the training and test accuracy and privacy leakage of non-private and differentially private models. But with the guidance of Dr Zhigang Lu, I was able to preprocess the dataset. He told me to use first 1200 data records in the training set and another 1200 data records in the test set. So, I had to think of how I could preprocess the dataset correctly. This time, I created a bangkok_shokri_dump pickle file in a separate python file from the given location dataset. In order to create that dump file, I created a new bangkok_shokri csv file from original location dataset with quotations removed from the first column as the pickle

file does not support different data types. Then, I transferred all data from modified csv file to numpy array which I used that to create a bangkok_shokri_dump pickle file. Subsequently, I used that dump pickle file in another python file to create separate pickle files for features and labels. For creating a label file, I clustered all the classes from the first column of the data using k-means algorithm, and then write all the classes into the label pickle file. As I had both the feature and the label files, I could run the code successfully; and as suggested, I split the dataset into training and testing set for target and shadow models where the sample sizes of both the training and test set are 1200 each. I again ran the code in two different modes i.e. non private and private, and that gave me a training and test accuracy and privacy leakage of four differential private models including the non-private one.

New knowledge, skills and experience:

Knowledge: I have learned how deep learning model can be used to classify an image. I have also learned about different deep learning architectures and their regularization and optimization mechanism. I have learned use of different deep learning module packages like torch, torchvision, pillow, labelmg, tensorflow, tfnightly, tensorflowprivacy etc. I have also learned about differentially private technique which is used to maintain privacy of individual's information while sharing group data publicly. I have also learned the mechanism of membership inference attack model. I have also learned how to preprocess and convert a csv file into a numpy array, and then, use that to create a pickle dump file. Furthermore, I have learned about k-means clustering algorithm that can be used to make clusters.

Skills: I have become more proficient in python and machine learning as I did preprocessing and explored more deep learning frameworks and packages. I have also gained knowledge of cryptography.

Experience: Running the code in pycharm, jupyter notebook and google colab, and having an issue with new deep learning packages and preprocessing. I also have learned how to use Fawkes windows version for security cloaking of images. I also learned about

the application of differentially private mechanism. I also learned about the adversarial attack models.

Rewarding Experience: I have successfully run the code with both the datasets and found the testing and training accuracy of non-private and differentially private models.

Difficult Experience: I found it difficult to run the source code on my local machine as I had to explore a new package called tensorflow privacy. Also, while running the code on a new dataset, I had to do some preprocessing which was difficult as I had to read the readme file to understand the data and its nature.

Task for the upcoming week: The following week, I will again revise the code to ensure whether I can get correct training and testing accuracy of non private and differential privacy models for the new dataset before exploring and running the code in Google AI cloud.

Week 8: In this week, I again ran the code on my local machine that Dr Zhigang Lu provided to me with a location dataset to check the training and test accuracy score and the privacy leakage of non-private and private models. Initially, I was having difficulties in getting a correct testing accuracy of private models for privacy parameter, i.e. epsilon, value of 0.01. I even changed the default model parameter values in the parser, and as well as in the functions such as training_target model, attack model and shadow model, but still only testing accuracy score output were not as expected. After having a meeting with Dr Zhigang Lu, I understood that there should not be any modifications in the functions, but changes can be made to the default values given in parser on the parameter values of epsilon, target model noise parameter and the type of privacy assigned to target_model which can be either grad_pert (with privacy) or no_privacy. Also, I realized that while running the code in different modes i.e. private or non-private, I had to use the bash commands different from that given in associated readme file to obtain the accurate test accuracy results. With these understandings and after experimenting several times, finally, I got the accurate results for both the private and non-private models. Furthermore,

I also started to explore google AI cloud platform for experimentation of the differential privacy models on the cloud platform. I created an account in Google AI cloud platform.

New knowledge, skills and experience:

Knowledge: I have learned about different deep learning architectures and their regularization and optimization mechanism. I have learned use of different deep learning module packages like torch, torchvision, pillow, labelmg, tensorflow, tfnightly, tensorflowprivacy etc. I have also learned about differentially private technique which is used to maintain privacy of individual's information while sharing group data publicly. I have also learned the mechanism of membership inference attack model. I have also learned how to preprocess and convert a csv file into a numpy array, and then, use that to create a pickle dump file. I have learned about k-means clustering algorithm that can be used to make clusters. Last, but not the least, I understood clearly code of the membership inference attack against differentially private models, and I learned about google AI platform.

Skills: I have become more proficient in python and machine learning as I explored more deep learning frameworks and packages. I have gained insight into differentially private model parameters. I have got a preliminary understanding of google AI cloud platform.

Experiences: Running the code in pycharm, jupyter notebook and google colab, and having an issue with new deep learning packages and data pre-processing. I learned about the application of differentially private mechanism. I also learned about the adversarial attack models. I understood the code of membership inference attack against differentially private models. Furthermore, I gained initial experiences about google AI cloud platform.

Rewarding Experience: I could run the code successfully with my pre-processed location dataset i.e Bangkok Shokri, and found that this time, I got a correct training and test accuracy score and privacy leakage of non-private and four differentially private models.

Difficult Experience: It took me a long period of time to understand the code as I was getting the same testing accuracy score for both the private and non-private models. To deal with my problems, I had to read the code and readme file several times, and eventually, I managed to get correct results with lots of trial and error.

Task for the upcoming week: As I have already created an account in google AI cloud platform, my next step is to learn more about this platform, and then to be able to run the built in AI platform models with the location dataset successfully before using the MI model code given by Dr Zhigang Lu in the platform with Bangkok_shokri dataset to illustrate the results.

Week 9: The goal for this week is to explore the limitations of the existing privacy solutions. Sitting on the side of an unauthorized deep-learning model, I need to design and test techniques for exploiting the user information even when privacy defences are in place. So, in this week, I explored more on Google AI cloud platform. I created my own project in Google AI cloud where I created a bucket in cloud storage. As I wanted to run my Bangkok_shokri dataset with default Google AI model, I imported the dataset from the local machine. Initially, there were difficulties in importing a data file due to its format, but later, I learned that I had to convert the file given by Dr Zhigang Lu into csv file. In the csv file, I took first 1200 data samples for training and deleted the remaining samples and the index which was showing rows from 0 to 5010. After that, I imported this data file into cloud storage, and then have it trained with Google AI built-in model in which I specified that the first column is our target column which contains label names. It took a couple of hours for me to get the training result. The model showed a good accuracy with my preprocessed dataset. To illustrate the model hyperparameters, it was a neural network with a hidden layer size of 1024, four hidden layers, the number of cross layers and dropout rate equivalent to 4 and 0.25 respectively.

New knowledge, skills and experience:

Knowledge: I have learned about different deep learning architectures and their regularization and optimization mechanism. I have learned use of different deep learning

module packages like torch, torchvision, pillow, labelmg, tensorflow, tfnightly, tensorflowprivacy etc. I learned about google AI platform. Furthermore, I learnt how to create a project and storage bucket in Google AI cloud, and I learnt how to import a dataset, preprocess the dataset and train a Google AI platform built-in model.

Skills: I have become more proficient in python and machine learning as I explored more deep learning frameworks and packages. I have gained insight into differentially private model parameters. I have got a deeper understanding of google AI cloud platform processes.

Experiences: Running the code in pycharm, jupyter notebook and google colab, and having an issue with new deep learning packages and data pre-processing. I learned about the application of differentially private mechanism. I also learned about the adversarial attack models. I understood the code of membership inference attack against differentially private models. Furthermore, I gained more experiences about google AI cloud platform, AutoML Tables and its auto AI model.

Rewarding experience: I could create a project and bucket in Google AI cloud platform, and I could import a dataset from a local machine to Google AI cloud platform. Moreover, I was able to train Google AI platform model with my pre-processed dataset.

Difficult experience: I would say that converting the original location file into csv file and preprocessing it before training with Google platform built-in model was the most difficult experience I had in this week.

Task for the upcoming week: As I trained my dataset with built in AI model, my next step is to use the Google platform to make a prediction for an arbitrary data point(s), and then, export the trained model to my local machine. After exporting, I have to run the trained model locally to make a prediction using new data points.

Week 10: The goal for this week is to explore the limitations of the existing privacy solutions. I need to design and test techniques for exploiting the user information even when privacy defences are in place. So, for this week, after training the model with first 1200 Bangkok_shokri Location data points, I created a separate data file containing 10 records from the remaining data points for making prediction using the trained Google AI model that showed a true prediction accuracy of 0.98. There are two options for predicting data points in Google AI cloud which are online prediction and batch prediction. As online prediction requires 446 feature data inputs just for one data record, I used batch prediction. So, I put the data file containing 10 data records in the Google cloud bucket and ran batch prediction for the records. After getting the result, I felt that the model took 10 data records randomly and made changes in serial of feature column for individual record. To match the prediction with the input data record, both the input and output data files are reshuffled. First, columns are reshuffled based on column heading serial, and then rows are reshuffled based on label value serial in first column so that it predicts 10 out of 10 label values correctly. Also, from earlier model training evaluation, I got true prediction accuracy 0.98 so I was content with my model training. Thus, I exported the model from AI platform unified to cloud storage bucket. Then, I followed the Google documentation for downloading the exported model into my local machine. I installed Google cloud SDK and docker as requirement, but following the command suggested by Google document, I could not download the model with its folder structure as it was giving me error message because of folder name pertaining timestamp which has a compatibility issue with window. Although documents say to rename the directory name with timestamp for further use after download. I couldn't download for the same problem itself. Later on, in consultation with Dr Zhigang Lu, I downloaded all the files manually, but when I tried to create the docker image following Google document, looks like it didn't get the correct folder and file location structure probably because of manual download. Therefore, I couldn't use docker containing run for local prediction using docker. Unfortunately, this is the prediction avenue suggested by Google. Then, I resorted to a Google cloud group document which suggested using saved_model.pb file to load it as a graph in python, and then getting it run for output which I was trying for a long time. But again, this was having an issue of tensorflow version as my system is upto date with tensorflow-2.4, whereas,

Google exported saved_model.pb file is probably compatible with tensorflow-1.15 and tensorflow-gpu-1.15. I was trying to solve with a tf-compatible version-1 import. But, after having a meeting with Dr Zhigang Lu, I again had to perform a batch prediction with 10 remaining data records as initially, during my batch prediction, I didn't remove the label column which classified 10 out of 10 labels correctly. This time, I got the corrected prediction csv file that includes prediction scores for 30 classes for each record. I also explained Dr Zhigang Lu about my difficulties in running the prediction locally using Google autoML table model, and the problems of downloading the exported model manually. After hearing that, he waived my task of making prediction locally using autoML model mentioning that we only need prediction vectors which we have gotten already.

New knowledge, skills and experience:

Knowledge: I have learned about different deep learning architectures and their regularization and optimization mechanism. I understood the membership inference attack mechanisms against differentially private models, and I learned about google AI platform. I learnt how to create a project and bucket in Google AI cloud, and I learnt how to import a dataset and train a Google AI platform model. I learned how to export a model into a cloud storage bucket and make predictions using new data points for the google trained model using batch prediction approach. I learnt the use of docker and Google cloud SDK tools.

Skills: I have gotten a deeper understanding of how google AI cloud platform works and a little understanding of Google cloud SDK and docker and using saved_model.pb file to load it as a graph in python, and then getting it run for output.

Experiences: Running the code in pycharm, jupyter notebook and google colab, and having an issue with new deep learning packages and data pre-processing. I learned about the application of differentially private mechanism. I also learned about the adversarial attack models. I understood the code of membership inference attack against differentially private models. I gained more experiences about google AI cloud platform

and its auto AI model. I learned how to train a Google AI model with given dataset and how to use the trained model for getting predictions on new data points. Furthermore, I also learnt that I cannot export a model into my local machine using the command line, but instead, I have to do it manually which the docker does not support while predicting data points as the model structure changes after downloading it manually. I also gained experiences of using saved_model.pb file to load it as a graph in python, and then getting it run for output locally.

Rewarding experience: I successfully got the predicted output from the google cloud for the new data points using Google autoML model that I created last week.

Difficult experience: Downloading the exported model using the google suggested command, and thereby, using it to make the prediction of data points locally using docker, was a difficult task which I couldn't complete successfully. My another difficult experience was to try using saved_model.pb file to load it as a graph in python, and then getting it run for output even after repeated attempts.

Task for the upcoming week: As I got the corrected prediction file from the Google cloud, my task for the following week is to read Google's instruction to interpret the prediction results made by Google platform and train MIA (shokri's) model for adversarial ML attack on Google AutoML trained model.

Week 11: The goal for this week is to explore the limitations of the existing privacy solutions. I need to design and test techniques for exploiting the user information even when privacy defences are in place. So, for this week, I got prediction vectors after performing batch predictions from trained Google AutoML Table model on the remaining 1200 test datasets. The prediction vectors are prediction scores, i.e., probability of a feature dataset to be in a certain class and for my case, there are 30 prediction scores for 30 classes. The class showing the highest probability of prediction score is considered as the class of the data record. After learning how to interpret prediction vectors, I copied

prediction scores for 30 classes for each row into another csv file. I also copied the labels from the training dataset which was trained via Google AutoML table model, and then, made a separate file containing the labels. Moreover, I even performed a similar task for the testing dataset which I used to predict using batch predictions in Google AI cloud. After completing these data files creation tasks, I used the code given by Dr Zhigang Lu to train the shadow models and attack models for which I converted the generated csv files into pickle files to follow the same format of files that are given in the code itself. I could load the new pickle files, but I am still having an issue with passing the new pickle files parameters to the attack model i.e Shokri model for getting a training and test accuracy.

New knowledge, skills and experience:

Knowledge: I have learned about different deep learning architectures and their regularization and optimization mechanism. I understood the membership inference attack mechanisms against differentially private models, and I learned about google AI platform. I learnt how to create a project and bucket in Google AI cloud, and I learnt how to import a dataset and train a Google AI platform model. I learned how to export a model into a cloud storage bucket and make predictions using new data points for the google trained model using batch prediction approach. I learnt the use of docker and Google cloud SDK tools. I learnt how to get prediction vectors generated from Google AI cloud using batch predictions on test data. Moreover, I acquired a preliminary understanding of using the attack model to attack the prediction scores to get the training and test accuracy.

Skills: I have gotten a deeper understanding of how google AI cloud platform works and a little understanding of Google cloud SDK and docker and using saved_model.pb file to load it as a graph in python, and then getting it run for output. As my main task of this week is to utilize attack model to attack the prediction vectors generated from Google AI cloud, I had to use my additional knowledge of python to complete the process. Thus, I believe that my skills in python are more developed than before.

Experiences: Running the code in pycharm, jupyter notebook and google colab, and having an issue with new deep learning packages and data pre-processing. I learned about the application of differentially private mechanism. I also learned about the adversarial attack models. I understood the code of membership inference attack against differentially private models. I gained more experiences about google AI cloud platform and its auto AI model. I learned how to train a Google AI model with given dataset and how to use the trained model for getting predictions on new data points. Furthermore, I also learnt that I cannot export a model into my local machine using the command line, but instead, I have to do it manually which the docker does not support while predicting data points as the model structure changes after downloading it manually. I also gained experiences of using saved_model.pb file to load it as a graph in python, and then getting it run for output locally. I also gained experiences of using the attack model to attack the prediction vectors generated from Google cloud.

Rewarding experience: For this week, my rewarding experience was to interpret the prediction vectors correctly.

Difficult experience: Using the Shokri's model to attack the prediction vectors was the most challenging task as I had to understand the full code given by Dr Zhigang Lu in order to pass the parameters into the attack model.

Task for the next week: My next week's task is to try doing the same task again until I get the correct outcome.

Week 12: The goal for this week is to explore the limitations of the existing privacy solutions. I need to design and test techniques for exploiting the user information even when privacy defences are in place. So, for this week, I again tried using Shokri model to attack the prediction vectors generated from batch predictions in test dataset using Google AutoML table model, and this time, I was successful. To complete the task, I prepared a train and test set for target and 60 shadow models. Then, I trained the Google

AutoML table model using the training set of target model. After that, I made batch prediction from Google trained target model and took prediction vectors as output utilizing the test set of target model. Finally, these prediction vectors were passed to attack model to train the attack model when simultaneously 60 shadow models were also trained and subsequently, using the Shokri attack framework membership leakage was estimated. It could be seen that the increase in number of shadow models intensifies the power of Shokri model to attack the non-private model as it increases the privacy leakage significantly. Dr Zhigang Lu saw this outcome, and then, he instructed me to upload our own code given by him in Google platform implementing RDP scheme, a differentially private model. Hearing his reply, I ran the code for RDP scheme for a privacy budget parameter equivalent to 1000, and then, I saw that privacy leakage decreases significantly and as well as the training and test accuracy. I delivered this output to my co-supervisor, and then, I came to learn that I have to train a baseline non-private model at Google platform with Google's default configurations and then record training & test accuracy. Secondly, I have to train all four DP models (our codes with the same configurations as Google's, that means I need to manually hard-code those configurations in the codes by checking Google's configurations) at Google platform and then record Training & Test accuracy. In line with recording accuracy, I am also required to calculate the accuracy loss of the four DP models with epsilon values such as 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500 and 1000. I also need to make prediction vectors for the 5 models i.e non-private, dp, zcdp, adv_cmp and rdp and calculate their privacy leakage. Repeating these steps 1-5 times, my final task is to plot average accuracy loss vs epsilon and average privacy leakage vs epsilon. I have already started doing this task and created some privacy leakage, accuracy and prediction scores for some models that are non-private, dp for epsilon 0.01, 0.05, 0.5, 0.1 and 1. Since my task is bit lengthy and it takes almost 14.5 hours to get the outcome for a model, I still haven't completed yet. But I believe that I can be able to finish this plotting task for the location dataset by next week.

New knowledge, skills and experience

Knowledge: I learned how to export a model into a cloud storage bucket and make predictions using new data points for the google trained model using batch prediction

approach. I learnt the use of docker and Google cloud SDK tools. I learnt how to predict vectors generated from Google AI cloud using batch predictions on test data. I acquired a complete understanding of using the attack model to attack the prediction scores to get the training and test accuracy. I got a learning that higher number of shadow models can increase the capability of Shokri privacy leakage. Furthermore, I learnt that there is a subtle change in output of models after running the code with and without Google's default configurations in Google cloud.

Skills: As my main task of this week is to utilize attack model to attack the prediction vectors generated from Google AI cloud, I had to use my additional knowledge of python to complete the process. Thus, I believe that my skills in python are more developed than before. Now, I can run the code in Google cloud, and I can also use Google's default configuration with my code to run in Google cloud.

Experiences: I gained more experiences about google AI cloud platform and its auto AI model. I learned how to train a Google AI model with given dataset and how to use the trained model for getting predictions on new data points. Furthermore, I also learnt that I cannot export a model into my local machine using the command line, but instead, I have to do it manually which the docker does not support while predicting data points as the model structure changes after downloading it manually. I got valuable experiences of using the attack model to attack the prediction vectors generated from Google cloud. I also gained experiences of running the code with and without Google's default configurations in Google cloud.

Rewarding experience: For this week, my rewarding experience was to successfully attack the prediction vectors generated from Google cloud. Also, I completed the task of running the own code in Google cloud successfully.

Difficult experience: Using the Shokri's model to attack the prediction vectors was the most challenging task as I had to understand the full code given by Dr Zhigang Lu in order to pass the parameters into the attack model. Apart from that, understanding how the

model could be run with and without Google's default configurations in Google cloud was also a difficult task as I had to go through many tutorials and documentation to solve this problem.

Task for the next week: My next week's task is to create a plot for average accuracy loss vs epsilon and average privacy leakage vs epsilon by repeating the above calculation steps five times.

10. Work samples

During my internship period, I have successfully completed a number of notable work samples which demonstrate the application of knowledge learned through courses during my last three terms. A few notable work samples are illustrated below:

Sample 1

I have run Fawkes program source code in my computer to test its performance of cloaking image in order to protect the image from subsequent adversarial attack. During my week 4, I used the source code of Fawkes and then I was able to successfully run the code in my computer. Before running the code, I successfully installed the package fawkes using pip install, and then installed GPU to make the perturbation faster. The following below is the demonstration of two images where one image is the normal image and another is the cloaked image.



Fig-2: Normal Image

Fig-3: Cloaked image

We can see that even with a perturbation, there is almost a negligible change in the image. So, this will misguide the attackers to identify a person's image.

(<https://github.com/Shawn-Shan/fawkes>)

Sample 2

After completion of my work with image privacy with different image privacy preserving frameworks, I switched my work concentration on differential privacy and membership inference with adversarial attack model. In consultation with Dr Zhigang Lu of my company, I select a well acclaimed state-of-the-art membership inference framework given by Shokri et al.(2017) and subsequently modified by Jayaraman and Evans(2019). I followed following steps to investigate the framework with location dataset given by Dr Lu.

- I cloned the code from the github repository of Jayaraman and Evans(2019).
- First, I read both the papers(Shokri et al.(2017); Jayaraman and Evans(2019)) to understand the theoretical aspect of differential privacy and membership inference scheme illustrated in the paper relating to the given code to understand the framework implementation in the code itself.
- I reviewed the location dataset which has 5010 data records and each record contains 446 binary features. The data point records different group of mobile users visiting 318 zones of Bangkok city and 128 classified zone types of the same. The given code uses only pickle file format as input. In order to complete the preprocessing, I created a location dump pickle file in a separate python file from the given location dataset. In order to create that dump file, I created a new bangkok_shokri csv file from original location dataset with quotations removed from the first column as the pickle file does not support different data types. Then, I transferred all data from modified csv file to numpy array which I used to create a bangkok_shokri_dump pickle file. Subsequently, I used the dumped pickle file to create separate pickle files for features and labels. For

creating a label file, I clustered all the classes from the first column of the data using **k**-means algorithm, and then write all the classes into the label pickle file.

- I split the location dataset into training and testing set for target and shadow models where the sample sizes of both the training and test sets are 1200 each.
- Finally, I ran the code for different modes i.e non-private and private, and that gave me a training and test accuracy and privacy leakage of four differential private models including the non-private one. The command for running the code in non-private and private mode is different.

The following below is the output for non-private model and private model(dp framework) for privacy budget parameter(epsilon) values of 0.01 and 1000.

```
Structure
Train accuracy is: 1.000
Test accuracy is: 0.613
Train loss is. 0.009
Shokri Privacy leakage is. 0.835
vorites
```

Fig-4: Non-private model

Structure	vorites
Train accuracy is: 0.041	Train accuracy is: 0.848
Test accuracy is: 0.035	Test accuracy is: 0.517
Train loss is. 72.902	Train loss is. 1.003
Shokri Privacy leakage is. 0.000	Shokri Privacy leakage is. 0.317

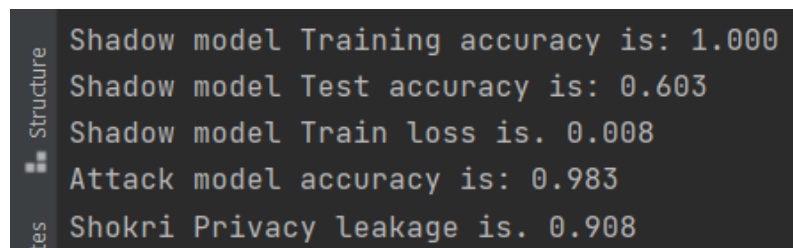
Fig-5: Private DP Model(ϵ -0.01) Fig-6: Private DP Model(ϵ -1000)

The figure 4 illustrates that the non-private model has a high prediction accuracy but a high privacy leakage showing vulnerability of data privacy. Figure 5 and 6 show that introducing differentially private budget parameter can arrest the

privacy leakage in a big way; budget parameter of 0.01 can totally nullify privacy leakage but with a sacrifice of model prediction accuracy whereas budget parameter of 1000 can arrest some privacy leakage with little sacrifice of model prediction accuracy. This means there should be an optimum trade-off value of budget parameter with acceptable privacy leakage and model accuracy.

Sample 3:

My next milestone task was to train a Google cloud AI model with my training dataset. For that, I opened an account in Google cloud where I created my own project and data storage bucket. Google AI model takes input in CSV format and hence, I converted my training dataset into CSV format. In cloud platform, I configured my training dataset using 800 data for training, 200 data for testing and 200 data for validation. After completion of training the Google AI projection model, I took the trained model configuration as output. Then, I used my target model test dataset of 1200 for taking batch predictions using vertex AI trained model. The batch prediction output gives prediction vectors for 30 classes. Then, I passed these prediction vectors to train membership inference attack model in my local machine for 60 shadow models setup and took output taking the target model as a non-private one.



```
Shadow model Training accuracy is: 1.000
Shadow model Test accuracy is: 0.603
Shadow model Train loss is. 0.008
Attack model accuracy is: 0.983
Shokri Privacy leakage is. 0.908
```

Fig-7 Privacy leakage of Google AI model

The figure 7 illustrates that Google AI model is prone to very high leakage rate of 0.908 when attacked by Shokri MIA model. In order to verify this claim, I ran my original model code with Rdp scheme with a privacy budget parameter of 1000 in Google cloud. For that, I had to install Google cloud developers platform cloud SDK which creates a container of Google cloud operating environment in local machine. The sample output is as follows:

```
Train accuracy is: 0.968
Test accuracy is: 0.578
Train loss is. 0.155
Shokri Privacy leakage is. 0.398
```

Fig-8 Privacy leakage of RDP model in Google cloud platform (ϵ -1000)

The figure 8 shows that introduction of privacy budget parameter value as high as 1000 can significantly reduce privacy leakage rate. This illustrates that Google AI model can be better off for privacy vulnerability with inclusion of differential privacy scheme in the model.

Sample 4

My final milestone task was to mimic a target model following Google AI default model configuration and train the same in Google cloud platform. Then, prediction vectors and test accuracy are taken as output for various privacy modes and budget parameters for runs in Google cloud platform. After that, MIA model is trained in my local machine with the prediction vectors for five models, and privacy leakage is calculated. For illustration purpose, I repeated the above steps five times to get the average values. Finally, I plotted the average accuracy loss versus budget parameter and the privacy leakage versus budget parameter for four differential private models. The plots are presented below:

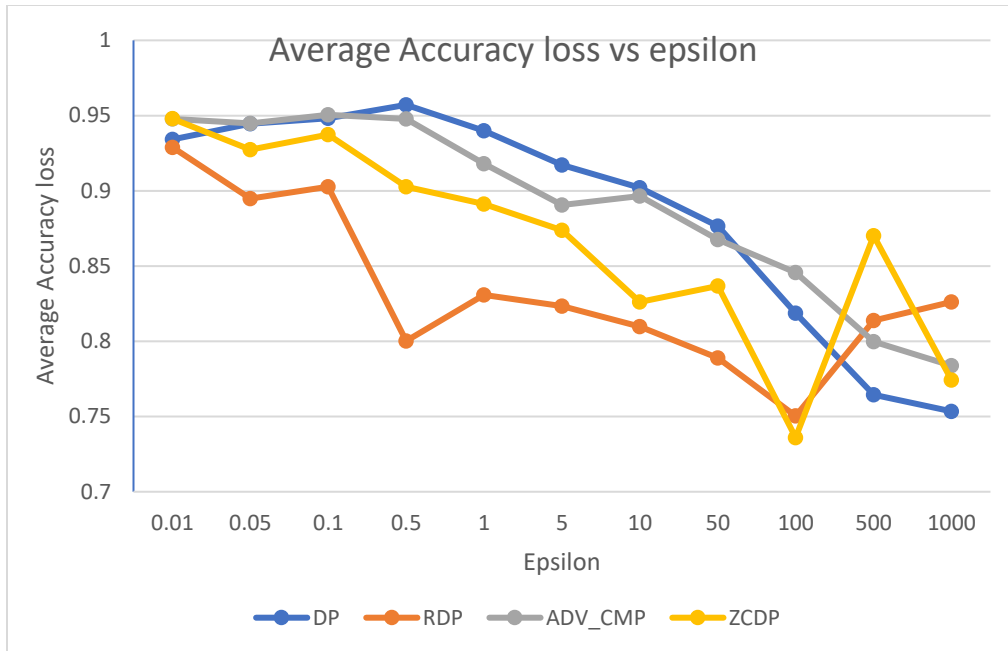


Fig-9 Average Accuracy loss vs budget parameter for target model with Google default configuration

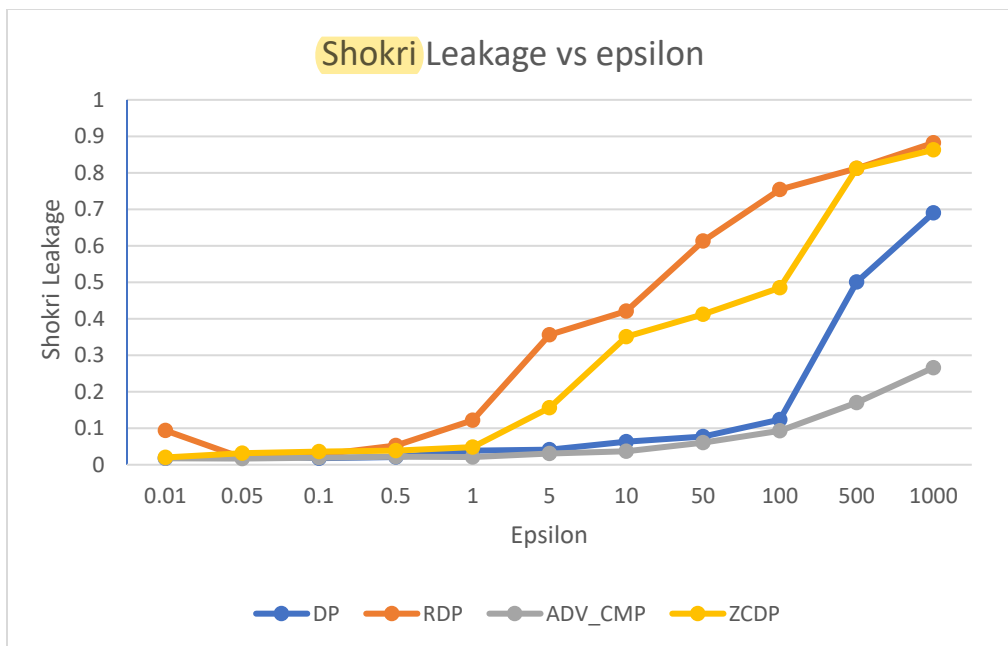


Fig-10: Shokri Leakage versus budget parameter for target model with Google default configuration

The figure 9 illustrates that average accuracy loss of Google configured target model is reduced with increased budget parameter values with **rdp** and **zcdp** schemes showing some irregular values for the higher range of budget parameter. But other two schemes of **dp** and **adv_cmp** are showing consistent reduction of accuracy loss with increased budget parameter values. Figure 10 illustrates the privacy leakage versus budget parameter for target model with Google default configuration. It is revealed that at lower budget parameter values, all the privacy schemes are equally effective in reducing privacy leakage while at higher budget parameter values, **adv_cmp** scheme is most effective followed by **dp** scheme.

11 Critical Analysis

My internship program deals with analyzing privacy preserving deep learning techniques. This requires review of papers dealing with deep learning model applications for protecting user privacy. For my first few weeks, I reviewed research article dealing with user image privacy protection from unauthorized deep learning model access. The researchers suggested utilization of various JPEG compression techniques for protecting images from use/training by unauthorized deep learning models. At the core of all these techniques, preparation/transformation/grouping/manipulating large database of image pixels are widely used. My learnings from courses like Data Science, Big Data Technologies, Application of Data Science and Machine learning helped me review and understanding of the research processes. Also, these researches utilizes various statistical and probability schemes for developing mechanisms in deep learning models. My knowledge from the courses in Statistical and Probability area such as Applied statistics, Introduction to probability, Statistical inference, Multivariate analysis, Generalized linear model and Statistical graphics helped me a lot for understanding the relevant mechanisms.

For recent few weeks, my new co-supervisor Dr Zhigang Lu suggested me to undertake a literature review in the area of differential privacy technique for user privacy protection.

I reviewed the concept, mathematical scheme and application of differential privacy from recent articles. Also, I reviewed membership inference adversarial attack against deep learning models, where differential privacy concept is used as a technique for prevention of attack. All these differential privacy application researches utilizes the concepts and mechanisms which fortunately I had learned through my data science, computational and probability statistics courses. Also, during my master's program, I have developed coding and software skills such as Python, SQL, MongoDB, R and Javascript alongside my fundamental skill in the general coding and software like C and C++. All these coding and software skills helped me grasping the code and overall program structure of the reviewed researches.

For last six weeks, I have extensively worked with differential private and MIA framework and code given by Shokri et.al(2017) and subsequently modified by Jayaraman and Evans(2019) using location dataset. My course learnings from data science, big data technologies, application of data science, generalized linear model and multivariate analysis helped me visualization and preprocessing of dataset. My knowledge in python through the course works helped me successfully running the MIA model code in my machine. My learning of computing courses helped me grasping the Google cloud platform operation including training and taking trained model predictions in cloud platform. My knowledge of python helped me to reconfigure the MIA model code after Google AI model default configuration. My analysis of model results illustration reveal that there is a significant scope of improving Google AI model privacy with introduction of privacy budget parameter scheme.

As the organization I am working in is primarily dealing with the areas of machine learning, security and privacy protecting models with application in various large databases, my work and learning nicely complement each other.

12. SWOT Analysis

Table 1: SWOT Analysis

Strengths: <ul style="list-style-type: none">• Strong Networking and Team.• High quality faculty researcher as supervisor.• Availability of high quality co-supervisor.• Quality research environment.• State of the art software and tools.• Friendly work environment.	Weaknesses: <ul style="list-style-type: none">• Less number of seminar/conference.• Limited project duration.• Lack of funding• Less cloud based software.
Opportunities: <ul style="list-style-type: none">• Global shift to cloud base operations.• Growing demand for data security and privacy technology.• New Machine learning and AI application	Threats: <ul style="list-style-type: none">• Time essence.• Strong competition from other research entities.• Lack of growth due to Covid-19 pandemic.• Time bound delivery.

Strengths

-Strong Networking and Team: My organization maintains a strong networking and team activities of highly skilled academic researchers and industry professionals. Therefore, it can leverage on quality multi-dimensional team formation for specific task i.e. problem solving/investigation; so getting help/guidance in any required direction is not an issue here.

-Quality supervisor: My supervisor, Dr Muhammad Ikram is a very knowledgeable person. He is a lecturer in Cybersecurity at Computing department of Macquarie university, and he is also a member of Information Security and Privacy group. His research interests include Cybersecurity, Privacy, Censorship-measurement, Data mining and Machine learning. Currently, he has 536 citations in the google scholar. His useful guidance and supervision helped me significantly in getting deeper knowledge of machine learning model and data privacy.

-Quality co-supervisor: My co-supervisors are Dr Gioacchino Tangari and Dr Zhigang Lu. During the first four weeks, I was guided by Dr Gioacchino who helped me a lot in

understanding deep learning models, and after that from week 5, I was guided by Dr Zhigang Lu. Dr Zhigang Lu has six conference papers and two journal papers. His relentless guidance and supervision helped me successfully completing a state-of-the-art investigation work in differential privacy and membership inference area.

- **Quality research environment:** Optus Macquarie university Cybersecurity Hub presents a congenial quality research environment where I feel encouraged as a beginning researcher. The research team here is knowledgeable, co-operating and helpful, which has facilitated my workflow in a smooth manner.

-**Software and tools:** The latest software tools are available for applied research and industrial applications. The servers, networks, mathematical modeling tools, simulation tools, machine learning tools are available and ready for research applications.

Weaknesses

- **Lack of seminar/conferences:** With a longer internship duration and possible COVID free situation, there could have been opportunities of a number of seminars and conferences in my organization, where I could have gained significant knowledge sharing and networking potential. In absence of those, my knowledge gathering avenues are limited. However, I am writing a paper on my investigation outcome to publish in a conference/journal.
- **Project duration:** As my internship project is of research nature, limited project duration is a weak point for freedom of research thinking. The thoughts of limited duration always put a boundary to any idea of extension and direction of new thinking. However, I will continue the application of my developed framework for some more time.
- **Lack of funding:** As the current internship work is unpaid, it does not provide any financial leverage, especially during this hard financial situation of COVID-19

pandemic. Therefore, the thoughts of financial worry weakens the progress potential.

- **Cloud based software:** The organization does not offer any cloud based software application for the researcher. Therefore, significant amount of time is wasted for customizing various open source free software for research applications. Google cloud platform and Vertex AI have removed these shortcomings significantly.

Opportunities

- **Global shift to cloud base operations:** As the companies are gradually shifting to cloud based operation and data storage for cost optimization and multi-group tasking feasibility, my training mainly involving cloud operation offers huge opportunity to carry on.
- **Growing demand for data security and privacy:** Data security and privacy is a demanding area of ICT industry as businesses and service sectors are increasingly becoming dependent on large scale data maintenance, protecting data security and customer/client privacy are getting importance in corporate sectors and government agencies. This fuels the job growth, research and development potential in the sector. Therefore, working in data security and privacy opens up opportunities for me in terms of industry and research jobs.
- **Machine learning and AI:** Data security and privacy applications use numerous machine learning and AI models. Therefore, working in this area opens up opportunity for getting involved in rapidly growing tech area.

Threats

- **Time essence:** As I have done a state-of-the-art investigation work, initiatives should be undertaken to publish the work in order to claim the work credit. Because

if similar work is already published from other organizations, our work will lose its value.

- **Strong competition from other research entities:** As our organization is mainly involved with research oriented activities, it faces strong competition with other research entities in private and public sector. Therefore, staying on top of competitors will depend on delivering top notch research outputs.
- **COVID-19 pandemic:** Ongoing covid-19 pandemic has put a pause to possible expansions of businesses, market potentials and client growth. This may pose a threat to sustainable existence of the organization and its activities.
- **Time bound delivery:** As my internship work has a number of timebound delivery, any personal or externally imposed lapses can always pose threats to meeting delivery schedule. However, I have met all of the project delivery milestones.

13. Conclusion

My review of literature on user image privacy protection from unauthorized deep learning model attacks shows the significant potential of various recently developed deep learning model frameworks application for user image privacy. Two of such frameworks namely Fawkes and Shield are already in the process of public domain and industry application respectively. My study of membership inference adversarial attack models utilizing differential privacy technique also gives me a deep understanding in the applied areas of data and user privacy protection. I have also run the shared source code from some of the reviewed researches to replicate the reported results and **understanding** the process of modeling and coding skill. The results from model framework run shows that the non-private model has a high prediction accuracy but a high privacy leakage showing vulnerability of data and user privacy. With introduction of differentially private budget

parameter, the privacy leakage can be arrested in a big way; budget parameter of 0.01 can totally nullify privacy leakage but with a sacrifice of model prediction accuracy whereas budget parameter of 1000 can arrest some privacy leakage with little sacrifice of model prediction accuracy. This means that there should be an optimum trade-off value of budget parameter with acceptable privacy leakage and model accuracy.

Google AI model is prone to very high leakage rate of 0.908 when attacked by Shokri MIA model. Introduction of privacy budget parameter value as high as 1000 can significantly reduce privacy leakage rate. This illustrates that Google AI model can be better off for privacy vulnerability with inclusion of differential privacy scheme in the model. My illustration results reveal that average accuracy loss of Google configured target model is reduced with increased budget parameter values. The privacy leakage versus budget parameter plot for target model with Google default configuration reveals that at lower budget parameter values, all the privacy schemes are equally effective in reducing privacy leakage while at higher budget parameter values, `adv_cmp` scheme is most effective followed by `dp` scheme.

My future plan is to apply the same model for a new data set and develop similar output for the new dataset. I will edit and illustrate the results in a paper form to publish in a conference or journal. After that, I have plan to work with my company supervisor and colleagues to develop a framework for improving data and user privacy in cloud based AI models.

14. Recommendation

The internship assignment process could have started a month or two earlier to give the student ample time for mental preparation and readiness of the work. As the university interns require access to various common free/paid software, datasets, modeling tools etc. The university can maintain a cloud based software and data resource platform for the interns to get readily available. This will save significant time for the interns looking for customized relevant software and useful datasets.

15. References & Sources

- Das, N., Shanbhogue, M., Chen, S. T., Hohman, F., Li, S., Chen, L., ... & Chau, D. H. (2018, July). Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 196-204).
- Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer, Berlin, Heidelberg.
- 'Differential Privacy' Wikipedia. Available at:
https://en.wikipedia.org/wiki/Differential_privacy
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.
- <https://github.com/Shawn-Shan/fawkes>
- <https://github.com/bargavj/EvaluatingDPML>
- <https://www.mq.edu.au/partner/access-business-opportunities/innovation-entrepreneurship-and-it/optus-cyber-security-hub/our-research>
- Jayaraman, B., & Evans, D. (2019). Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)* (pp. 1895-1912).
- Sablayrolles, A., Douze, M., Schmid, C., & Jégou, H. (2020, November). Radioactive data: tracing through training. In *International Conference on Machine Learning* (pp. 8326-8335). PMLR.
- Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., & Zhao, B. Y. (2020). Fawkes: Protecting privacy against unauthorized deep learning models. In *29th {USENIX} Security Symposium ({USENIX} Security 20)* (pp. 1589-1604).
- Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3-18). IEEE.
- Tariq, M. I., Memon, N. A., Ahmed, S., Tayyaba, S., Mushtaq, M. T., Mian, N. A., ... & Ashraf, M. W. (2020). A Review of Deep Learning Security and Privacy Defensive Techniques. *Mobile Information Systems*, 2020.
- Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., ... & Vadhan, S. (2018). Differential privacy: A primer for a non-technical audience. *Vand. J. Ent. & Tech. L.*, 21, 209.