POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

On the resilience and protection of regularization techniques in differential privacy

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Marco Giammarresi

Advisor: Prof. Matteo Matteucci

Co-advisor: Eng. Eugenio Lomurno

Academic year: 2020-21

1. Introduction

In recent times, the number and variety of applications of deep learning algorithms have noticeably grown; given the existing proportionality between the deep learning models' performance and the amount of data fed to them, it is necessary to collect more and more data in order to have more reliable results from these models. Most of these data are usually crowdsourced, so the data owners must ensure privacy guarantees related to their use and sharing.

Indeed, malicious agents can target models in order to reveal potentially sensitive information that remains within them when the training is complete and exploit them for harmful purposes. Deep learning models that process data retain meaningful details about them and are usually accessible to everyone through online APIs.

Malicious agents can thus conceive attacks that exploit knowledge of the trained models to obtain information about their training. Among them, two specific types of attacks are the most dangerous: model inversion attacks and membership inference attacks. To reduce the efficacy of these attacks, a widely spread countermeasure taken by model owners is applying differential privacy in the training procedure of deep learning models. The main advantage of this solution is the guarantee that the privacy leakage at the end of the model's training is limited and measurable. The main downside of this approach is that the introduction of differential privacy in the training procedure has a significant impact on the model's utility and on the time needed for the training to complete. In this work, we investigate the topic of privacy preservation in deep learning, and we test the effectiveness of the most known implementation of differential privacy in deep learning models, the Differentially Private Stochastic Gradient Descent (DP-SGD), as a defense mechanism. Our goal is to evaluate the validity of this method in terms of protection against both model inversion and membership inference attacks and measure the impact that its application has on the model under attack, analyzing both the level of accuracy achieved and the training time. We also perform the same analysis considering two regularization techniques, dropout and the L2 regularizer. Their effect on improving a model's generalization capability is widely known, and previous works [4] have confirmed an existing connection between privacy attacks' effectiveness and overfitting in the target model. We conclude our work drawing from this empirical analysis conclusions on the effectiveness of both approaches, differential privacy and regularizers, as overall defense mechanisms.

2. Privacy attacks

We differentiate the scenarios in which an attack occurs depending on the extent of the adversarial knowledge, that is the ensemble of information concerning the model and the data under attack at disposal of the attacker. From this point of view, we distinguish the threat scenarios into two types: black-box and white-box. In a black-box scenario, the adversary knows only elements of the target model that are available to the public, such as prediction vectors, but neither has any access to the model structure nor any information about the training dataset. In a white-box scenario, the adversary has complete knowledge of the target model and knows the data distribution of the training samples.

2.1. Model inversion attacks

Model inversion attacks try to reconstruct training samples of the attacked model starting from environmental elements known by the attacker. The first designs of reconstruction attacks assumed a white-box scenario in which the adversary knows the output label, the prior distribution of features for a given sample, and has complete access to the model; with these assumptions, the attacker estimates the sensitive attributes' values that maximize the probability of observing the known model parameters. These kinds of attacks are referred to as Maximum a-posteriori (MAP) attacks [3].

Subsequent paradigms of model inversion attacks abandoned the MAP approach in favor of the optimization of given reconstruction losses between the output of the inversion model and real training samples. A novel strategy of attack consists in recovering the training samples of the target model starting from its activation maps. This attack can be performed both in white and black box scenarios, depending on whether the attacker decides to consider the intermediate maps of the target model or its prediction vector. Due to this adaptability to different scenarios and the possibility of conducting layer per layer analysis, in our work, we focus on this strategy of attack.

2.2. Membership inference attacks

Membership inference attacks take as input a sample and try to determine if it belongs to the training dataset of the model under attack or not. The most common paradigm for designing such attacks is the use of shadow models and meta-models or attack models. The basic idea behind this approach is to train several "shadow" models that imitate the behavior of the target on "surrogate" (or shadow) datasets. Shadow datasets must contain samples with the same format and similar distribution to the training data of the target model. After the training of shadow models is complete, their outputs and the known labels from the shadow datasets form the attack dataset; this dataset is used to train the meta-model, which learns to infer membership based on the shadow models' results. The main issues of this approach are the transferability between the shadow model and the target model and the strong assumptions related to the adversarial knowledge of the target model structure and training data distribution. To overcome these issues, Salem *et al.* [4] proposed three different attacks considering scenarios with more relaxed assumptions on the adversarial knowledge. The first two approaches maintain the idea of shadow models, while the third proposal abandons the shadow model paradigm in favor of a threshold-based attack. In this approach, the attacking model is a simple binary classifier that takes from the prediction vector of the target model the highest posterior and compares it against a given threshold; if its value is greater than the threshold, the input sample, from which that output is obtained, is considered a member of the training dataset. The advantages of this approach are the complete independence from the target model and its training data and the elimination of the overhead costs due to the design of shadow models and the creation of suitable shadow datasets; besides, it requires no training of the attack model. In this work, we use two attacks belonging to the two aforementioned categories: one attack that involves the training of a shadow model and a threshold-based attack.

3. Differential privacy

Differential privacy is devised as an effective privacy guarantee for algorithms that work with

aggregated data. Before enunciating the formal definition of differential privacy, we have to explain the concept of adjacent databases. Two databases are adjacent "if they differ in a single entry, that is if one image-label pair is present in one set and absent in the other" [1].
Definition 1. A randomized mechanism *M: D* $\rightarrow$ *R* with domain *D* and range *R* satisfies ε-differential privacy if for any two adjacent inputs *d, d'* $\in$ *D* and for any subset of outputs $S \subseteq R$ it holds that

$$Pr[M(d) \in S\,] \leq e^{\varepsilon} Pr[M(d') \in S\,]. \qquad (1)$$

The parameter ε is called privacy budget because it represents how much information leakage we can afford in our system, so a lower value indicates a stricter privacy guarantee. Differential privacy represents a significant development in the field of privacy-preserving deep learning because it guarantees three properties that result very useful in our context of research: composability, group privacy, and robustness to auxiliary information. Composability means that if we have a composite mechanism and each of its components is differentially private, then the overall composition is differentially private; this property allows us to design mechanisms in a modular fashion. Group privacy assures that if the dataset contains correlated data, like the ones provided by the same individual, the privacy guarantee degrades gracefully and not in an abrupt way. Finally, robustness to auxiliary information guarantees that the privacy level assured by theory stands regardless of the knowledge available to the adversary.

3.1. (ε, δ)-Differential Privacy

In practice, differentially private mechanisms cannot always assure privacy guarantees as stated in the main definition of ε-Differential Privacy for every possible ε. For this reason, the classical formulation of ε-Differential Privacy needs to be relaxed to allow every differentially private mechanism to be implemented with less strict privacy guarantees, but at least valid for every ε. The main relaxation of differential privacy usually considered for implementing any mechanism is (ε, δ)-Differential Privacy.
Definition 2. A randomized mechanism *M: D* $\rightarrow$ *R* with domain *D* and range *R* satisfies (ε, δ)-differential privacy if for any two adjacent inputs *d, d'* $\in$ *D* and for any subset of outputs $S \subseteq R$ it holds that

$$Pr[\ M(d) \in S\ ] \leq e^{\varepsilon} Pr[\ M(d') \in S\ ] + \delta, \qquad (2)$$

where the additive factor δ represents the probability that plain ε-DP is broken. Once we have defined a differentially private mechanism, we can apply an a posteriori analysis giving us several (ϵ, δ) pairs that satisfy the privacy conditions of our scenario. In the case of a composite mechanism, this analysis results more complex because we need to keep track in some way of the privacy loss accumulated during the execution of each component. However, the composability property of differential privacy brought Abadi *et al.* [1] to design an element, called privacy accountant, that calculates the privacy cost needed for each access to the data and uses this information to define the overall privacy loss of the mechanism; the specific accountant conceived by Abadi *et al.* is called moments accountant because it keeps track of a bound on the moments of a privacy loss random variable. Studies on how to implement a differential privacy mechanism into a deep learning model lead to the design of the Differentially Private Stochastic Gradient Descent (or DP-SGD) [1].

3.2. Renyi differential privacy

The Renyi Differential Privacy is a relaxation of differential privacy based on the concept of Renyi divergence.
Definition 1. For two probability distributions P and Q defined over R, the Renyi divergence of order $\alpha > 1$ is

$$D_{\alpha}(P \parallel Q) \triangleq \frac{1}{\alpha - 1} \log E_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^{\alpha}. \qquad (3)$$

The relationship between the differential privacy formulation and the Renyi divergence can be expressed through the following definition
Definition 2. *A randomized mechanism M: D* $\rightarrow$ *R is* ε*-differentially private if and only if its distribution over any pair of adjacent inputs* $d, d' \in \mathrm{D}$ *satisfies*

$$D_{\infty}(M(d) \parallel M(d')) \leq \varepsilon. \qquad (4)$$

This relationship justifies the idea of developing a relaxation of standard differential privacy based on Renyi divergence; it can be generalized through the definition of the so-called (α, ε)-Renyi differential privacy.

Definition 3. *A randomized mechanism M: $D \rightarrow R$ is said to have ε-Renyi differential privacy of order α if for any adjacent $d, d' \in$ D it holds that*

$$D_\alpha(M(d) \parallel M(d')) \leq \varepsilon. \tag{5}$$

It can be demonstrated that the three aforementioned properties (composability, robustness to auxiliary information, and group privacy) are still valid for Renyi differential privacy, despite the relaxation of the original definition of differential privacy.

4. Privacy preserving regularizations

We want to test the effectiveness of both DP-SGD, that is the current standard approach in terms of privacy preservation, and other regularizers in defending deep learning models and providing privacy guarantees. To do this we subject both methods to two membership inference attacks and one model inversion attack. Regarding the membership inference attacks, we perform two black-box attacks inspired by Salem's paper [4]. One is a threshold-based attack, while the other involves the training of a single shadow model and assumes no knowledge about the data distribution; the last one is referred also to as data transferring attack. To evaluate the effectiveness of the membership inference attacks, we measure the AUC (Area Under Curve) of the ROC curves; the higher the AUC, the more successful is the inference attack. As for the model inversion attack, we have devised an approach that exploits the activation maps of the target model to reconstruct its training data.

Our model inversion attack consists of two phases:

1. train the target model on a given dataset, then freeze all its layers up to the one whose activation map we want to extract;
2. attach the frozen layers at the top of the actual adversary model.

Our adversary model is structured to be able to receive in input any activation map among the ones produced by the target model's layers; to achieve this, we place after the frozen layers extracted from the target model a block of five layers. This block first flattens the activation map, if necessary, then feeds the resulting vector to a Dense layer that maps it to a fixed dimension, and finally, reshapes the output into a three-dimensional matrix; this last transformation is needed to feed the block output to the convolutional part of the adversary model. A detailed overview of the architecture of the adversary model is shown in Figure 1.

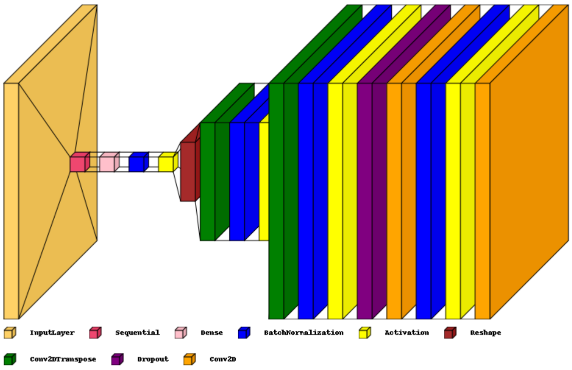

Figure 1: Plot of the adversary model's structure. The Sequential block is the stack of frozen layers coming from the target model.

To measure the reconstruction loss between the original image and the one generated by our model, we use the mean squared error (MSE). We already know from previous works that the integration of differential privacy in deep learning has proven successful in preserving privacy against several attacks; however, recent studies have revealed a major flaw in this method, the significant impact on the model's accuracy [2]. We want to understand if this thesis holds and analyze the relationship between the model's utility, as measured by its accuracy, and its privacy guarantee.

For this reason, we evaluate in a comparative way the performance of DP-SGD with different privacy budgets according to four metrics: resilience to both membership inference attacks and the model inversion one, accuracy achieved by the target model, and time duration of the training process. Then, our work proceeds in looking for an alternative solution capable of resolving the tradeoff between privacy guarantee and performance, both in terms of training time and utility, of the model under attack. The main idea behind our proposal is that there can be a connection between the generalization capabil-

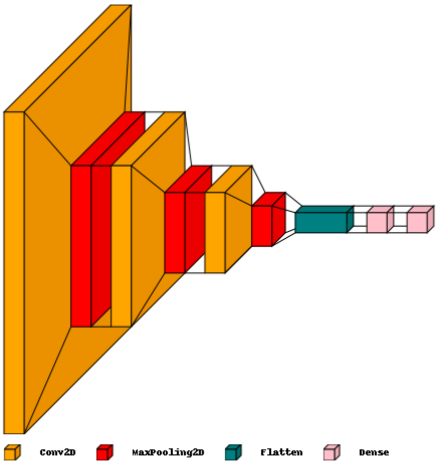

Figure 2: Plot of the target model's structure.

ity of a network and its resistance to the model inversion attack. Indeed, overfitting is a usual cause of a model's lack of ability to generalize well on new samples; so, our approach consists in reducing the overfitting in the network to increase the resistance to the attack. In our solution, we choose to use two regularization techniques that prevent overfitting: dropout, and L2 regularization. Our final objective is to prove the validity of this solution as a defense mechanism empirically, evaluating it comparatively with the same four criteria we used for differential privacy: resistance to both privacy attacks, the accuracy achieved by the target model, and the total time to train it. The target model, which will remain the same in every analysis performed, is a convolutional neural network with nine layers, not including the input one; its architecture is shown in detail in Figure 2. To introduce differential privacy in the training of the target model, it suffices to substitute its optimizer with the Differentially Private Stochastic Gradient Descent, so the structure and the number of parameters of the model remain the same. Regarding the insertion of regularization inside the model, we decide to insert the Dropout layers in the middle of every convolutional block, that is convolution+pooling, and after each Dense layer, except for the network output, and to add L2 regularization only to the output layer.

5. Experiments

We have performed our experiments on three image datasets: CIFAR-10, MNIST, and Fashion-MNIST. Regarding the hardware used for our experiments, it is an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz with an Nvidia GeForce GTX TITAN X GPU. We carried on two separate groups of experiments: one regarding the study of (ε, δ)-differential privacy and a second one related to the study of different regularization scenarios. Each model is trained with the same hyperparameters and for the same number of epochs of the target model without differential privacy and regularization. For both studies, we present the results achieved in terms of accuracy of the target model, training procedure's execution time, resistance to both membership inference attacks and our model inversion attack. In our study of differential privacy, we consider three privacy configurations with decreasing levels of privacy guarantee; in the first configuration, we have $\varepsilon = 2$, in the second one $\varepsilon = 4$ and in the third one $\varepsilon = 8$. We recall that a lower value for the privacy budget ε corresponds to a stricter privacy guarantee. At the end of this first study, the results obtained show that not only differential privacy has a significant impact on the model's performance, but also that it does not significantly improve the resilience of the model to our model inversion attack; indeed for some activation maps, it worsens it, especially for the ones corresponding to the last three layers. In this scenario, differential privacy with $\varepsilon = 2$ remains the best choice among the differentially private configurations in terms of overall protection against the inversion attack. In the study of the regularizers, we considered three scenarios: use of L2 regularization, use of dropout, and use of both techniques together. In terms of performance, these three solutions do not impact noticeably on the model's utility nor on the execution time of the training process; besides, they also provide reasonable protection against the membership inference attacks. The most interesting results involve the model inversion attack: we found out that L2 regularization significantly increases the resistance to the attack for the regularized layer, while dropout increases the overall resistance to the attack against intermediate layers, but it does not improve the protection of the output layer.

Combining L2 regularization and dropout, we obtain a compromise that guarantees a slight

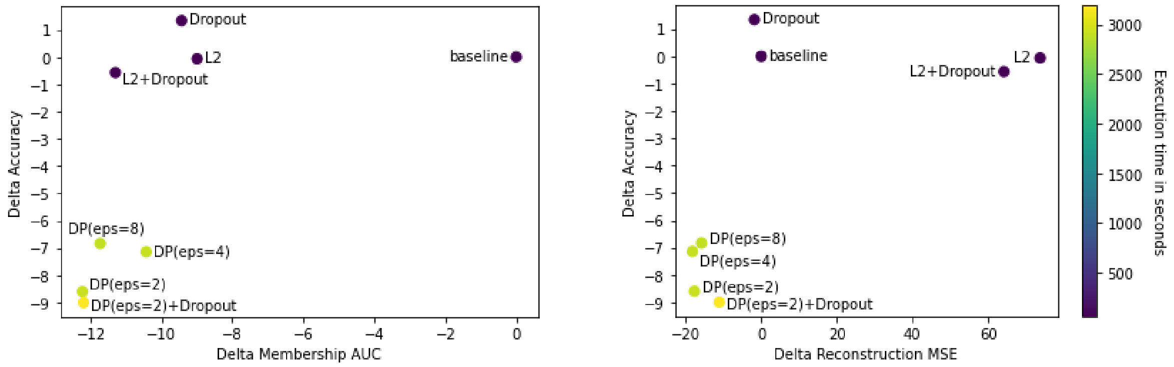

Figure 3: Final comparison considering average results on all datasets. We express the three dimensions on the axis of the two plots in terms of average percentage variation with respect to the baseline scenario (whose values are equal to 84.6% for the accuracy, 67% for the membership AUC and 0.047 for the reconstruction MSE).

improvement in the level of protection of the intermediate layers and significant growth in the resilience to the attack for the output layer. After completion of the two studies, we decided to consider in our final experiments an additional scenario in which we combine differential privacy with $\varepsilon = 2$ and dropout.

We performed a final comparison of all the aforementioned scenarios, considering together the four metrics used throughout our experiments and a black-box scenario for both privacy attacks; the results of this final comparison are shown in Figure 3. From this collection of results, we can conclude that the solution with both dropout and L2 regularization is the best overall defense mechanism against privacy attacks that also provides good performances in terms of utility and execution time. However, applying L2 regularization alone at the output layer is a valid alternative solution that performs better in protecting from the black-box inversion attack but slightly worse against membership inference ones, with similar performance regarding the accuracy reached during training and the execution time of the training process.

6. Conclusions

In this thesis, we demonstrated the drawbacks of using differential privacy as a privacy-preservation method for deep learning models; in particular, we showed its significant impact on the performance of the model under attack, both in terms of the level of accuracy achieved and time duration of the training process, and its lack of effectiveness in protecting against a model inversion attack designed by us. We also found out that applying dropout and L2 regularization to the output layer of the target model is the best overall defense mechanism, while L2 regularization alone is the best solution in the case of a black-box model inversion attack. Moreover, we discovered that applying a high level of L2 regularization on a layer increases significantly its resistance to the model inversion attack.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. pages 308–318, July 2016.

[2] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. *Differential Privacy Has Disparate Impact on Model Accuracy.* 2019.

[3] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, 2014.

[4] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. January 2019.