Unsupervised Risk for Privacy

Christophe Cerisara
Université de Lorraine, CNRS, LORIA
Nancy, France
cerisara@loria.fr

Alfredo Cuzzocrea
iDEA Lab, University of Calabria, Rende, Italy
& LORIA, Nancy, France
alfredo.cuzzocrea@unical.it

***Abstract*—This position paper deals with privacy for deep neural networks, more precisely with robustness to membership inference attacks. The current state-of-the-art methods, such as the ones based on differential privacy and training loss regularization, mainly propose approaches that try to improve the compromise between privacy guarantees and decrease in model accuracy. We propose a new research direction that challenges this view, and that is based on novel approximations of the training objective of deep learning models. The resulting loss offers several important advantages with respect to both privacy and model accuracy: it may exploit unlabeled corpora, it both regularizes the model and improves its generalization properties, and it encodes corpora into a latent low-dimensional parametric representation that complies with Federated Learning architectures. Arguments are detailed in the paper to support the proposed approach and its potential beneficial impact with regard to preserving both privacy and quality of deep learning.**

***Index Terms*—Differential privacy, regularization, unsupervised risk**

I. Introduction

Data is ubiquitous but, assuming that there is enough computational power, time and human efforts, two other major issues severely limit its exploitation in machine learning:

- Data is rarely free and is always costly either to produce (e.g., writing a tweet) or capture (e.g., buying and compressing sensor data streams [1] in the industry). So stakeholders who have invested time or money in this process own rights over the data.
- Information in data may be harmful to specific people or groups of people, and should not be openly accessible.

These two issues, namely *copyrights* and *privacy*, despite originating from fundamentally divergent motivations and contexts, may be considered jointly from the technical point of view, because they both can be addressed from similar computational approaches that restrict access to the data. Still, training large models on shared data pools is desirable to maximize performances.

One widely known approach to achieve this is *differential privacy* (DP). However, DP suffers from several major issues, which we review next, and we propose in the following an alternative approach that partly addresses some of these challenges.

This research has been made in the context of the Excellence Chair in Computer Engineering – Big Data Management and Analytics at LORIA, Nancy, France, and has been funded by the OLKi and Digitrust Lorraine Université d'Excellence projects.

II. Limits of Differential Privacy

The first limitation of DP is due to the fact that noise is injected in the training process: this noise inevitably impacts the classification or regression performances of the model. Therefore, a compromise between quality of the model and the level of protection of private information has to be found. Several studies report that, in practical applications, in order to reach acceptable level of privacy, the quality of the model has to be severely degraded, which makes the model nearly useless for the target task [2].

Another major drawback of DP is a direct consequence of the core principle of DP that aims at preventing the model from memorizing individual samples from the training corpus. This principle comes in contradiction with recent works [3], which prove that memorization of singleton labels that typically occur in the long-tail distribution of labels (e.g., the long tail of the Zipf law distribution of words frequencies in natural language), is required so that the model may be able to generalize to infrequent sample sub-populations. This result shows that alternative approaches to DP shall be considered to protect privacy if we want to train high-quality models with good generalization properties.

III. Regularization for privacy

We argue next that DP can be advantageously replaced in deep neural networks by a combination of data protection approach, and non-destructive regularization techniques during training.

First, privacy can only be guaranteed when the data itself is not accessible to other practitioners than the data producers themselves. Federated Learning is currently one of the privileged approach to protect data, as the data itself does not leave the data producer's premises. Every computation that requires access to this particular data, such as training a deep neural network, is realized locally on such premises.

Second, the model itself, after or during training, shall not disclose private information. Instead of degrading the model to achieve this goal, as DP does, we argue that the models shall rather be modified to prevent membership inference attacks. This is of course a less strong guarantee than the one obtained by DP, because making the model robust to a selected set of membership inference attacks does not guarantee that, later, someone will design a novel privacy attack to which our model may not be robust. But compared to the loss in quality incurred

by DP models, we believe that this potential threat is more acceptable, and may be dealt with later on if it ever happens.

A. Privacy attacks and mitigations

We focus next on blackbox membership inference attacks, which are one of the most general and common types of privacy attacks against deep learning models.

The first family of such attacks rely on training a shadow model to mimick the behavior of the model under attack [4]. However, training such shadow models is becoming more and more difficult, if not impossible, given the size and cost of recent deep neural networks, especially in the Natural Language Processing domain, such as GPT3 or GShard and its 600 billion parameters [5]. Furthermore, other studies [6] have shown that as good and sometimes even better attacks may be achieved by simple metrics computed on the output logits of the target model. When considering these families of attacks, a straightforward objective to mitigate them is to prevent the outputs of the model to be different between in-training and out-of-training samples. This can be achieved by adding regularization terms to the loss during training of the model. Such regularization may be the standard L2-norm, or dedicated adversarial terms [7]. However, similarly to differential privacy, such regularization terms alter the parameters search space landscape during training and moves away the regularized optimum from the task objective, which is classification accuracy. Consequently, this may also result in a decrease in performances of the trained model.

B. On regularization

Our claim that, conversely to differential privacy, regularization approaches shall not inevitably lead to a decrease in the accuracy of the trained model, and so regularization constitutes a better option to investigate than DP to maximize both privacy and accuracy.

The loss function that is optimized during training is composed of two terms: the main error loss, which usually minimizes the empirical risk, and the regularization term, which commonly minimizes the model's parameters complexity. Minimizing the empirical risk with the main error loss makes the model overfits to the training dataset, which negatively impacts both its generalization capabilities and its robustness to membership inference attacks. Therefore, a regularization term, such as the L2-norm, is used to counterbalance such negative consequences. By smoothing the parameters search space, this regularization term reduces overfitting, which improves generalization as well as robustness to membership inference attacks. But regularization may also have a negative impact on the model accuracy, because it commonly only depends on the values of the model's parameters, and not on the task-specific evidence. Therefore, a compromise has classically to be found between the respective weights of both terms in the total loss.

Our proposal in this paper rather aims at designing a better regularization term that would both prevent overfitting and optimize the classification risk. We believe an interesting research direction towards this goal might be to give up the standard empirical risk approximation, as it is done in [8]. We briefly describe the underlying principle next and how it could be applied to mitigate membership inference attacks without impacting the model accuracy.

C. Unsupervised risk approximation

Let us consider without loss of generality a binary classifier that is trained with the hinge loss; our objective is to minimize the error that the classifier makes on unknown test data: this objective is formalized with the classification risk $R(\theta)$:

$$\begin{aligned} R(\theta) &= E_{p(x,y)}\left[(1 - f(x)\cdot(2y-1))_+\right] \\ &= P(y=0)\int p(f(x)=\alpha|y=0)(1+\alpha)_+ d\alpha + \\ &\quad P(y=1)\int p(f(x)=\alpha|y=1)(1-\alpha)_+ d\alpha \end{aligned} \tag{1}$$

where x are the observations, y the true class (y is unknown, because we consider here unsupervised training) and $f(x)$ is the scalar output score for observation x of a deep neural network parameterized by θ. Class 0 (resp. class 1) is chosen when $f(x)$ is negative (resp. positive). In the first equation, the expected value of the hinge loss is computed over the full continuous data distribution $p(x, y)$, including any unknown test corpus that will be created in the future.

Usually, this unknown distribution $p(x, y)$ is approximated by a finite labeled corpus, which leads to the classical supervised training algorithm with empirical risk minimization. We do not consider such an approximation here, because it requires to know the gold labels y, and because it is the root cause of overfitting. We rather follow two assumptions proposed in [9], which state that the prior $P(y)$ is known and that the class-conditional distribution of the output score $p(f(x)|y)$ is Gaussian. We will discuss next some conditions proposed in [8] to fulfill these assumptions. But for now, these assumptions allow us to derive Equation-1 into the following closed-form equation of the risk:

$$\begin{aligned} R(\mu,\sigma) &= \frac{P(y=0)}{2}(1+\mu_0)\left(1-\text{erf}\left(\frac{-1-\mu_0}{\sigma_0\sqrt{2}}\right)\right) + \\ &\quad P(y=0)\sigma_0^2 N(-1;\mu_0,\sigma_0) + \\ &\quad \frac{P(y=1)}{2}(1-\mu_1)\left(1+\text{erf}\left(\frac{1-\mu_1}{\sigma_1\sqrt{2}}\right)\right) + \\ &\quad P(y=1)\sigma_1^2 N(1;\mu_1,\sigma_1) \end{aligned} \tag{2}$$

where (μ_0, σ_0) and (μ_1, σ_1) are the parameters of the Gaussians respectively associated with class 0 and class 1.

This equation has several important properties with regard to our privacy objective:

- The Gaussian parameters $\mu = (\mu_0, \mu_1)$ and $\sigma = (\sigma_0, \sigma_1)$ can be estimated from an unlabeled corpus with standard Gaussian mixture estimation algorithms; the mixture coefficient being the known prior $P(y)$.

- (μ, σ) depend deterministically on the model parameters θ; this enables to train θ with gradient descent and with the chain rule:

$$\frac{\partial R(\theta)}{\partial \theta} = \frac{\partial R(\theta)}{\partial (\mu, \sigma)} \times \frac{\partial (\mu, \sigma)}{\partial \theta}$$

  The Gaussians thus act as a proxy that decouples the model parameters from the corpus: once the gradients with respect to each Gaussian have been computed, the deep model can be trained without any information from the corpus. This is important in the context of distributed privacy-protecting architectures.
- Such a training process uses the unlabeled corpus of observations only to estimate 4 parameters: the 2-dimensional vectors (μ, σ); then, the large number of parameters θ of the deep neural network may be trained only from (μ, σ), without any data. This makes optimizing the risk extremely robust to overfitting.

However, this training process provably converges towards the optimum classification risk $\min_\theta R(\theta)$ only when both assumptions are fulfilled. The first assumption about the known prior is not a major issue, as $P(y)$ can often be estimated from prior knowledge in many applications, such as the prevalence of a disease in healthcare diagnostics, and preliminary experiments suggest that unsupervised optimization is relatively robust to small estimation errors of $P(y)$.

About the second assumption, it is shown in [8] that the bi-Gaussianity assumption is valid in a neighborhood of the minimum of the empirical risk. Therefore, we suggest to not use Equation-2 as the first risk to optimize, but rather as a regularizer that should be applied *after* standard supervised training. The advantages of our regularizer, compared to the other ones, is that it both reduces overfitting, improves generalization and optimizes the test accuracy of the model.

D. Optimization process

The proposed approach may thus be decomposed into the following stages:

- In the first stage, the deep neural network is trained classically with the supervised empirical risk objective, which gives an initial set of parameters θ. At this stage, the accuracy of the model is good but it is sensitive to membership inference attacks.
- In the second stage, we collect an additional unsupervised corpus of data from the application. This second corpus does not need to be labeled, which greatly reduces the cost of the collection process, as raw unlabeled data is often readily available in many application domains. If this is not an option, then the initial training corpus that has been used in the first stage may also be used in stage 2, although better generalization properties may be obtained with a larger unlabeled corpus.
- In the third stage, the model parameters are optimized without supervision by iterating the following steps:
  - Make a forward pass over the unlabeled corpus to obtain the distribution $p(f(x))$.
  - Compute the bi-Gaussian parameters (μ, σ) from this distribution with, e.g., the Linde-Buzo-Gray algorithm or any other related method.
  - Apply one step of gradient descent to optimize $R(\theta)$ given (μ, σ).

During the third step, the model parameters θ will slowly deviate from the initial minimum of the empirical risk, which is prone to overfitting, and rather converge towards our approximation of the optimal true classifier risk $R(\theta)$, which does not depend on the finite training corpus and is thus immune to overfitting.

Of course, the quality of the approximation of $R(\theta)$ by Equation-2 depends on the representativity of the second corpus collected in stage 2; but this corpus does not need to be labeled, and can thus be much larger than the training corpus used in stage 1. Furthermore, only 4 parameters are trained on this corpus, which makes overfitting of these Gaussian parameters nearly impossible.

In other words, the rationale of this approach is to exploit large-capacity deep neural networks to project the observed features into a simple latent space where the class-conditional Gaussianity assumption is valid. Note that quite a similar relationship between a simple Gaussian and a complex feature space is also built in related works, such as the well-known variational auto-encoder [10], which confirms that such a projection is achievable through neural networks with enough capacity. Then, in this simple latent space, the corpus distribution is discarded and replaced by the low-dimensional Gaussian mixture; this is this "replacement" step that actually performs regularization, as all the specific details and outliers observed in the corpus are deleted. The Gaussian mixture approximation also generalizes beyond the training corpus, and implicitly covers domain samples that have not been seen in the training corpus. Optimizing Equation-2 with a few gradient steps then attemps to reduce the overlapping between both Gaussians, which provably converges towards the true classifier risk. Only a few gradient steps must be performed before the Gaussian parameters shall be re-estimated from the data in order to avoid the Gaussian mixture to diverge from the observations.

The main challenge and most important aspect in this paradigm is to start from an initial Gaussian mixture representation that clusters the data into the target classes of interest. This is why we propose to first completely train the model in a supervised way, and then only regularize it a-posteriori, instead of mixing regularization with the supervised training loss, as it is usually done. Our preliminary experiments confirm that this is a viable strategy to fulfill the Gaussianity assumption. Furthermore, our regularization objective does not deviate from the classification risk optimum as other regularizers do, and so it does not need to be "guided" by the main supervised loss and can be applied independently.

E. Towards improved privacy

Beyond improved generalization, we expect this paradigm to increase the robustness of the model against membership

inference attacks for the following reasons:

- **By reducing overfitting:** it has indeed been previously shown that the degree of overfitting is correlated to the success of membership inference attacks and that regularizing the model improves robustness against them.
- **By reducing the dependence to the training corpus:** we have seen in the previous section that the proposed approach decouples the training process from the actual corpus through the Gaussian mixture distribution. The model is thus actually trained *without seeing any specific training sample*: it only has access to the generic Gaussian mixture distribution. Consequently, its dependence to specific training samples shall be more and more reduced during this process, and thus the possibility to exploit the model's logits to know whether a sample is in the training corpus or not also disappears.
- **By combining it with other adversarial privacy terms:** the proposed approach replaces the supervised loss by another unsupervised loss, and so we expect that it should be compatible with other terms that may be added to the loss to improve the model's privacy, especially adversarial terms that prevent the model from being able to discriminate between samples that belong to the training corpus and the others. We believe such adversarial terms also constitute an interesting track of research towards improved privacy.
- **By combining it with Federated Learning:** the proposed loss is particularly well suited to a distributed computation framework such as Federated Learning, because of the Gaussian mixture proxy that it uses to represent the whole training corpus, which boils down to computing only 4 scalar parameters that are too small to encode any sensitive private information. It should thus be possible to compute globally these four global statistics with simple secure multiparty homomorphic operations at a reasonable cost, while specialized deep neural networks are updated locally, but this option is still to be investigated.

IV. Conclusions

In this position paper, we have briefly analyzed the impact of regularization from the perspective of robustness of deep neural networks to membership inference attacks, and compared it with other standard approaches, especially differential privacy. Then, we have proposed a novel regularization process that relies on a non-standard approximation of the classifier risk, which gives an unsupervised loss with interesting properties with regard to generalization and, potentially, privacy. The benefits of this loss for privacy are only conjectured so far, and they still need to be validated experimentally. However, we have also given several arguments that support this claim, as well as extensions of the proposed approach to combine it with other promising research directions. Novel paradigms are required to initiate new tracks of research and progress towards improved privacy, and this proposal departs from the main lines of research in the domain, but is also complementary with some of them, such as adversarial regularization. We believe it opens interesting research directions, but exploring them and studying experimentally their properties will require time, and this is why we have opted for now to submit the current state of our work as a position paper. The next steps will be, after having extensively evaluated the robustness of the approach against membership attacks, to study its combination with other adversarial regularization terms, as well as its robustness to other types of privacy attacks, especially white-box attacks that should become more frequent as the number of large pre-trained deep neural networks that are freely disseminated increase. Another, more technical advantage of the proposed approach is its relatively moderated computational cost, which results from the fact that the unsupervised loss can be fully differentiated in closed form and that good piecewise-linear approximations may be exploited as suggested in [8]. These questions shall also be experimentally validated in a future work. Finally, extensions of this approach to multi-class will be required to make the approach applicable in practical cases. However, despite such extensions being straightforward theoretically, we expect that difficult challenges will have to be solved in practice, for instance to estimate the N Gaussian mixtures that shall precisely match the target class-conditional distributions.

References

[1] A. Cuzzocrea, F. Furfaro, E. Masciari, D. Saccà, and C. Sirangelo, "Approximate query answering on sensor network data streams," *GeoSensor Networks*, vol. 49, 2004.

[2] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 1895–1912.

[3] V. Feldman, "Does learning require memorization? a short tale about a long tail," in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2020. New York, NY, USA: Association for Computing Machinery, 2020, p. 954–959. [Online]. Available: https://doi.org/10.1145/3357713.3384290

[4] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 3–18. [Online]. Available: https://doi.org/10.1109/SP.2017.41

[5] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. M. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," *ArXiv 2006.16668*, 2021.

[6] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Annual Network and Distributed System Security Symposium (NDSS 2019)*, 2019.

[7] H. Hu, Z. Salcic, G. Dobbie, and X. Zhang, "Membership inference attacks on machine learning: A survey," *CoRR*, vol. abs/2103.07853, 2021. [Online]. Available: https://arxiv.org/abs/2103.07853

[8] C. Cerisara, P. Caillon, and G. Le Berre, "Unsupervised post-tuning of deep neural networks," in *IJCNN*, ser. Proc. of the International Joint Conference on Neural Networks (IJCNN), United States, Jul. 2021.

[9] K. Balasubramanian, P. Donmez, and G. Lebanon, "Unsupervised supervised learning II: Margin-based classification without labels," *Journal of Machine Learning Research*, vol. 12, pp. 3119–3145, 2011.

[10] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.