Discriminative Adversarial Privacy: Balancing Accuracy and Membership Privacy in Neural Networks

Eugenio Lomurno, Alberto Archetti, Francesca Ausonio, Matteo Matteucci
Politecnico di Milano
Milan, Italy
{eugenio.lomurno, alberto.archetti, francesca.ausonio, matteo.matteucci}@polimi.it

Abstract

The remarkable proliferation of deep learning across various industries has underscored the importance of data privacy and security in AI pipelines. As the evolution of sophisticated Membership Inference Attacks (MIAs) threatens the secrecy of individual-specific information used for training deep learning models, Differential Privacy (DP) raises as one of the most utilized techniques to protect models against malicious attacks. However, despite its proven theoretical properties, DP can significantly hamper model performance and increase training time, turning its use impractical in real-world scenarios. Tackling this issue, we present Discriminative Adversarial Privacy (DAP), a novel learning technique designed to address the limitations of DP by achieving a balance between model performance, speed, and privacy. DAP relies on adversarial training based on a novel loss function able to minimise the prediction error while maximising the MIA's error. In addition, we introduce a novel metric named Accuracy Over Privacy (AOP) to capture the performance-privacy trade-off. Finally, to validate our claims, we compare DAP with diverse DP scenarios, providing an analysis of the results from performance, time, and privacy preservation perspectives.

1 Introduction

The burgeoning interest and application of deep learning across diverse industries and domains has been remarkable in recent years. This surge can be ascribed to several pivotal factors including the accessibility of massive data volumes, the enhancements in computational resources, and the evolution of neural network architectures and optimization algorithms. However, with the widespread use of deep learning models and their increasing influence on society and daily life, the security and protection of the sensitive data used to train such models have become an essential concern. In an era of stringent data protection legislations like the European General Data Protection Regulation [1] and the Chinese Cyber Security Law [2], threats and potential data security breaches evolve at a pace that is often faster than legislative response. One such threat is a class of attacks known as Membership Inference Attacks (MIAs) [3], which aim to deduce whether certain individual-specific information was part of the training dataset of a machine learning model. Such attacks pose a formidable challenge and lay the ground for more sophisticated and potentially harmful breaches. Despite their recognition in the literature, no formally effective countermeasure is widely adopted in the deep learning development community.

Among the different solutions to increase the resistance of machine learning models to MIAs, the incorporation of Differential Privacy (DP) within training optimisers has stood out for its potential. DP is frequently considered the go-to mechanism for guaranteeing privacy due to its theoretical properties and robustness [4]. However, research has shown that the high level of privacy offered by DP comes at a cost, as adopting a high level of DP usually results in severe performance loss and increased learning time. This makes DP not practical for many real-world applications and sometimes not even for simple simulations [5, 6].

In this paper, we introduce a novel privacy-preserving learning technique, called Discriminative Adversarial Privacy (DAP). DAP leverages the structure of MIAs to accomplish multi-objective adversarial learning. By using our approach, the training process of deep learning models is faster than training with DP and the final model has higher performance with a comparable privacy level. Specifically, DAP employs a discriminator trained via the MIA technique of shadow models and a novel loss function minimising the prediction error while maximising the attacker's error. This approach results in models that offer privacy competitive to those achieved by DP, yet with significantly reduced performance loss and faster training time.

With this work, we provide the following contributions:

- We propose a novel learning technique called Discriminative Adversarial Privacy or DAP, that combines adversarial learning and membership inference attack principles. This technique is designed to ensures an optimal balance between model performance, speed, and privacy.
- We introduce a novel loss function for DAP that is specifically tailored to simultaneously minimise the prediction error while maximising the attacker's error.
- We define a novel metric, namely Accuracy Over Privacy or AOP, to efficiently capture and handle the performance-privacy trade-off.
- We substantiate our claims with rigorous empirical validation, providing extensive experimental results that demonstrate DAP's comparative advantage over DP in terms of performance, training time, and privacy preservation.

2 Related Works

Membership Inference Attacks. The family of attacks known as Membership Inference Attacks (MIAs) is one of the biggest threats to deep learning models. MIAs are incredibly versatile and effective, leading to a growing research interest both in terms of the development of new attack algorithms and defensive countermeasures [7]. MIAs consist of determining, given a machine learning model, whether or not a given record was included in its training dataset. In practice, a MIA model is a binary classifier that can distinguish whether or not a record belongs to the training set of an already trained target model. The challenge is to carry out the MIA in the real world with little useful information for the attacker, such as in machine-learning-as-a-service scenarios. Shokri *et al* [3] pioneered one of the first and, still to this day, highly effective MIA algorithms, based on the assumption that over-parameterised models could memorise information about individual training samples beyond the generalisation of the problem for which they were trained. Assuming the structure and learning algorithm of the target model are known to the attacker, Shokri *et al*propose a training technique that trains several models – called *shadow* models – to emulate the target model's behaviour. In this way, the attacker can leverage the predictions of such models to build a MIA discriminator, able to identify whether a sample has been used or not in the training procedure of the target model.

Numerous studies extended this technique and expanded the attack surface. For instance, Chen *et al*used data poisoning to enhance the MIA precision while hiding the attack traces by minimising test-time performance degradation [8]. He *et al*demonstrated the feasibility of MIA against models trained via self-supervised learning, and explored early stopping as a potential countermeasure, albeit at the expense of the model's utility [9]. Recently, researchers evaluated the effectiveness of MIAs against Generative Adversarial Networks (GANs) [10, 11], diffusion models [12, 13], recommender systems [14, 15], semantic segmentation [16, 17], and text-to-image [18].

Differential Privacy. Historically, Differential Privacy (DP) has been the primary defence against MIAs. It is a procedure designed to provide robust protection for individual-level information in a dataset [19]. The application of DP ensures that the inclusion or exclusion of any individual sample in a dataset does not significantly alter the results of statistical analyses or machine learning models trained on that dataset. DP is frequently presented in its relaxed form, referred to as (ε, δ)-DP. Formally, a randomised mechanism denoted as *M:* $D \rightarrow R$, with domain *D* and range *R*, satisfies (ε, δ)-DP if the following inequality holds for any two adjacent inputs *d, d'* $\in D$ and any subset of outputs $S \subseteq R$:

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

In Equation (1), the ε parameter, known as the privacy budget, denotes the maximum allowable information leakage, with a lower ε value indicating stronger privacy. Conversely, the additive δ term represents the probability of privacy preservation being violated.

In a machine learning context, the DP framework achieves its goal by introducing randomness into the data analysis process, in a manner that obscures the contribution of any single individual's data. Usually, this randomisation is implemented as additive noise summed to the original data, to the intermediate matrices of the training algorithm, or through ad-hoc subsampling of the dataset. Abadi *et al* [4] pioneered the concept of Differentially-Private Stochastic Gradient Descent (DP-SGD), which has been affirmed as one of the most prevalent differentially-private optimisers within the deep learning literature. DP-SGD introduces Gaussian noise to the gradient computation with a standard deviation controlled by ε. This step ensures that the gradients are sufficiently randomised, thereby hindering an

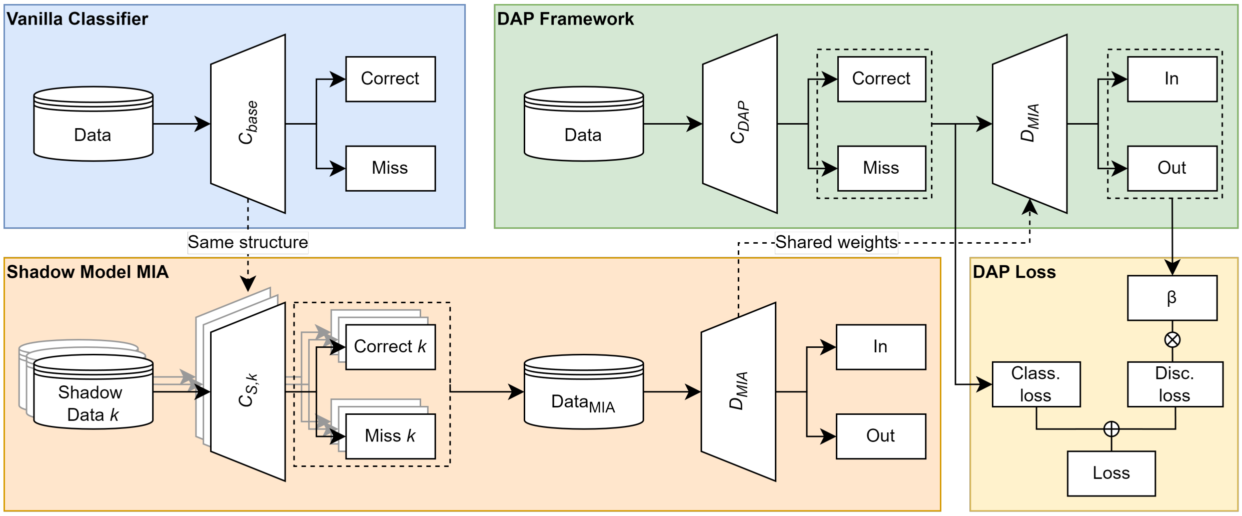

Figure 1: Overview of the Differentiable Adversarial Privacy (DAP) framework.

attacker's ability to infer information about individual data points from the model's parameters. This approach has been successfully applied across various domains, particularly in federated learning applications [20, 21, 22].

Alternative Privacy Preserving Techniques. While DP offers numerous advantages such as increased trust in data analysis results and enhanced fairness in decision-making processes that rely on data, it does come with its own set of challenges. These primarily revolve around the trade-off between privacy protection and data utility and the computational complexity that arises while implementing DP mechanisms [5, 6]. Given these constraints, the literature has seen the emergence of alternatives to DP and its variants. Chen *et al*proposed an alternative to DP, called RelaxLoss [23]. The key concept of this training framework is the relaxation of the entropy loss function, with the goal of reducing the generalisation gap and privacy leakage in machine learning models. Kaya and Dumitras [24] evaluated the efficacy of data augmentation mechanisms against MIAs in image classification tasks. Their study encompassed seven different mechanisms, including differential privacy. They found that augmenting data to improve model utility did not mitigate the risk of MIAs. Furthermore, they delved into why the commonly utilised label smoothing mechanism amplified the risk of MIAs. Webster *et al*introduced a general-purpose approach to tackle the issue of membership privacy in machine learning. Their solution involves the generation of surrogate datasets using images created by Generative Adversarial Networks (GANs), which are labelled with a classifier trained on the private dataset. They demonstrated that these surrogate datasets can be utilised for various downstream tasks and provide resistance against membership attacks. In their study, different GANs proposed in the literature were evaluated, revealing that GANs of higher quality yield better surrogate data for the given task [25]. Lomurno and Matteucci [5] presented a comparison of the effectiveness of the DP-SGD algorithm against standard optimisation practices with regularisation techniques. They compared the utility of the resulting models, their training performance, and the efficacy of MIAs against the learned models. Their empirical findings highlight the often superior privacy-preserving properties of dropout and $l2$-regularisation, given a fixed number of training epochs.

3 Method

Discriminative Adversarial Privacy. With this work, we introduce a novel learning framework for privacy-preserving deep learning, called Discriminative Adversarial Privacy (DAP). DAP is a learning framework to efficiently train high-performing deep learning models with strong resilience against MIAs. DAP uses a deep neural network classifier as a baseline, referred to as $\mathcal{C}_{base}$, and trained using the hold-out technique on dataset *Data*. Similarly, K shadow models, denoted as $\mathcal{C}_{S,k}$, are trained using the hold-out technique, as prescribed by the MIA from Shokri *et al* [3]. For each shadow model produced this way, ground truth, prediction, and loss are stored for each of its training and test samples, associating a binary label according to whether it belongs to the first or second set. Of these samples, only the miss-classified ones are retained, as they are the most empirically informative in a discriminative context and, from the ablation studies, lead to the most performing results. These data are used to build the adversarial binary classification dataset *Data*$_{MIA}$ to train the binary discriminator $\mathcal{D}_{MIA}$. Once trained, the weights of this model are frozen.

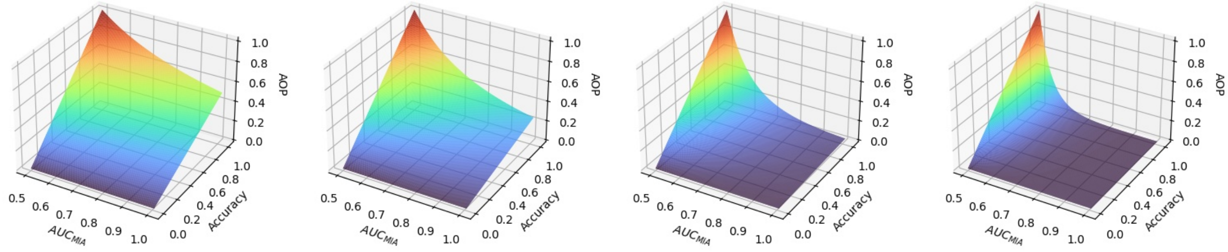

Figure 2: From left to right, interpolation plots of AOP(λ) for $\lambda = 1$, 2, 5, and 10.

At this point, adversarial training is performed using $\mathcal{D}_{MIA}$ and a new classifier $\mathcal{C}_{DAP}$ with the same structure of $\mathcal{C}_{base}$. In DAP, $\mathcal{C}_{DAP}$ is trained to minimize the categorical crossentropy loss as usual, i.e. to maximize the probability of assigning the correct class label to training examples. This error is used to update all the classifier's weights. Then, for each batch of data, the miss-classified predictions from $\mathcal{C}_{DAP}$ are collected together with their corresponding ground truth and loss. This secondary batch is fed through $\mathcal{D}_{MIA}$ and its prediction error is computed maximising the error of the discriminator as in the standard min-max adversarial training [26]. This secondary error is used to update the last fully connected layer of $\mathcal{C}_{DAP}$, with the goal of reducing the probabilities that its outputs can be easily discriminated by the attacker. The optimisation procedure of DAP can be described as

$$\min_{\mathcal{C}} \max_{\mathcal{D}} \mathcal{L}(\mathcal{C}, \mathcal{D}, t) = \mathbb{E}_{x \sim p(x)}[\log(\mathcal{C}(x,t))] + \beta \mathbb{E}_{x,y \sim p(x,y)}[\log(1 - \mathcal{D}(\mathcal{C}(x,t), y))]. \tag{2}$$

In Equation (2), x and y are respectively the training inputs and the ground truth labels, t is the current epoch, and β is a dynamic loss balancing parameter. β is crucial to ensure learning stability. In fact, the different nature of the two losses makes them not directly comparable in terms of magnitude depending on the training epoch and the specific data distribution. β is dynamically adjusted during training and it is computed as

$$\beta(\mathcal{C}, \mathcal{D}, t, r) = \begin{cases} \frac{\mathbb{E}[\log(\mathcal{C}(x,t-1))]_v}{\mathbb{E}[\log(1-\mathcal{D}(\mathcal{C}(x,t-1),y))]_v} \cdot r & \text{if } t > 0 \\ 1 & \text{otherwise} \end{cases}. \tag{3}$$

According to Equation (3), the value of β at time t is proportional to the ratio of the classification loss and the discrimination loss on the validation set at the previous step $t-1$. Then, β is scaled by a hyperparameter r that weighs the contribution of the discriminator. As a final note, β is always set to 1 for $t = 0$. The overall DAP framework is described in Figure 1.

Accuracy Over Privacy. When evaluating machine learning models in a privacy-preserving setting, it is vital to contemplate both model performance and the privacy of the underlying training data. However, measuring the trade-off between these two facets is a complex task. Quantifying privacy itself is nontrivial, and comparing metrics across different domains can be particularly challenging. For these reasons, we propose a novel metric called Accuracy Over Privacy (AOP), which provides a concise measure of the accuracy and privacy of a target model. Within the realm of MIAs, the efficacy of the attacking model is often measured using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve of the attack. Instead, when the model is a classifier, its performance can be assessed utilising the Top-1 Accuracy (ACC). Therefore, given the classifier ACC and the AUC of a MIA model (AUC_{MIA}), the AOP is computed as

$$\text{AOP}(\lambda) = \frac{\text{ACC}}{(2\max(\text{AUC}_{\text{MIA}}, 0.5))^{\lambda}}. \tag{4}$$

Equation 4 summarizes the effects of ACC and AUC_{MIA} in a single metric. $\lambda \geq 1$ weighs the importance of privacy when measuring the AOP.

The AOP metric exhibits several properties. Concerning its range, it is constrained in the interval $[0, 1]$. For highly inaccurate models or models susceptible to MIAs, the AOP approaches 0. Conversely, the AOP is closer to 1 when models exhibit high accuracy and strong resilience against MIAs at the same time. The λ parameter is a key factor in the AOP metric, as it is controls the impact of the privacy component. Figure 2 shows that increasing values of λ cause the AOP metric to shrink towards 0. Concerning the denominator, the max operator ensures that the AUC is never lower than the AUC of a random guessing model, which is equal to 0.5. Moreover, the denominator allows for obtaining AOP values equal to the classification accuracy for models that perfectly preserve privacy.

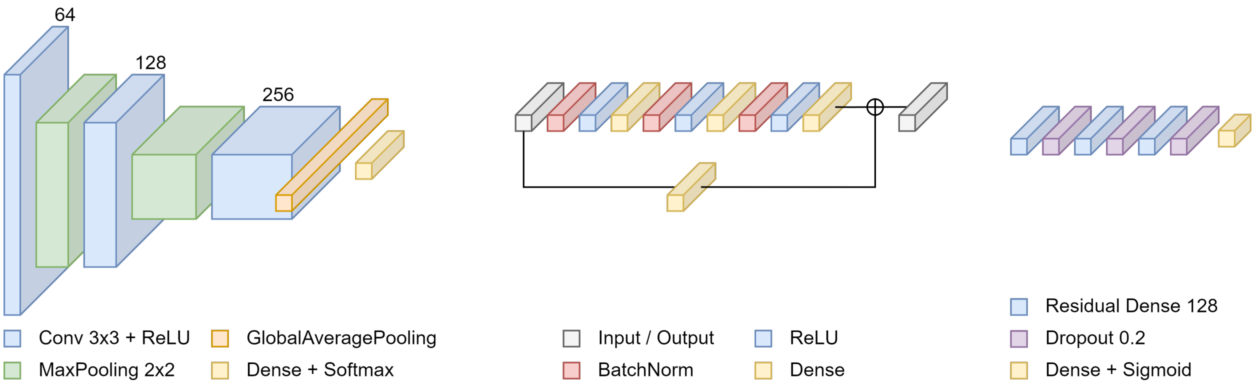

Figure 3: The neural network architectures involved in the experiments. From left to right, the CNN used for each analyzed classifier and shadow models, the residual block specially designed for the DAP discriminator, and the overall architecture of the DAP discriminator.

4 Experimental Setup

In order to guarantee the transparency and reproducibility of the study, this section provides a comprehensive description of the experiments conducted and their setup. Our proposed algorithm, DAP, operates in two different settings. In the first one, referred to as test DAP or **DAP**$_t$, shadow models are trained with the test set, simulating a situation where an external dataset, potentially public, is accessible to both the attacker and victim. DAP$_t$ allows deep learning engineers to proactively prevent potential attacks by employing the same dataset that could be used by the attacker. In the second setting, validation DAP or **DAP**$_v$, the shadow models are trained using the validation set. This situation mimics a typical scenario where the attacker's data distribution differs from that of the victim. In both of these modes, we maintained 10 shadow models, and optimize the parameter r over a uniform range from 0 to 1 with increments of 0.025.

To ensure a fair evaluation, we compared DAP against several alternative approaches. We initially establish a **Baseline** model, constituting of the base classifier without any protective measures. Subsequently, following the methodology of Lomurno *et al* [5], we include a model, called **Reg**, that applies dropout regularization to each intermediate classifier weight and $l2$ regularization to the model output. The dropout probability is tuned between 0.2, 0.33, and 0.5, while the $l2$ weight over 0.1, 0.01, and 0.001. Furthermore, we extend our examination to incorporate models with (ε,δ)-DP, retaining a constant δ value equal to 10^{-5} while adjusting the ε budget by modifying the number of training epochs. Specifically, we test four models with ε values of 0.5, 1, 2, and 4.

To limit the free parameters of the experiments, we maintained the same architecture for each classifier across all configurations, as illustrated on the left side of Figure 3. This selection was motivated by the intricate spatial complexity involved in DP training. The residual architecture of the discriminator employed in both the DAP$_t$ and DAP$_v$ models is depicted in Figure 3, where the proposed residual block is situated in the middle, and the overall structure is positioned on the right. All models are trained using the Adam optimizer with a learning rate chosen between 10^{-5}, 10^{-4}, and 10^{-3} and a batch size of 32. Each model is trained to convergence with early stopping – with a patience of 25 epochs – on validation accuracy except for DP models, where the number of training epochs is fixed and proportional to ε.

The proposed models are evaluated with respect to classification Top-1 Accuracy, AUC of MIAs – performed using the toolkit provided by the TensorFlow Privacy library – and training epoch time. Furthermore, we employed our novel metric, the AOP, with $\lambda = 2$, to assess the trade-off between performance and privacy, with a particular emphasis on the latter. The study covers eight datasets: Cifar-10 [27], Cifar-100 [27], FMNIST [28], EuroSAT [29], TinyImagenet [30], OxfordFlowers [31], STL-10 [32], and Cinic-10 [33]. The experiments are conducted on a machine equipped with an Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz CPU and an Nvidia Quadro RTX 6000 GPU.

5 Results and Discussion

In this section, we comment on the results obtained from the set of experiments comparing DAP to regularization and DP for defense against MIAs. Table 1 collects the accuracy metrics for each model and dataset. As anticipated, the models incorporating DP yield the lowest accuracy scores, even with a high privacy budget ($\varepsilon = 4$). Conversely, the regularized (Reg) model achieves consistently high accuracy, even occasionally outperforming the baseline model. Our proposed method, DAP, guaranteed an accuracy boost over the DP counterparts. A notable example of this is with the

Table 1: The Accuracy metric on the test sets. Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	**Baseline**	**Reg**	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	**DAP**$_t$	**DAP**$_v$
Cifar-10	0.784	**0.811**	0.313	0.374	0.417	0.418	0.624	0.613
Cifar-100	0.481	**0.532**	0.039	0.083	0.090	0.072	0.315	0.276
FMNIST	0.932	**0.926**	0.605	0.701	0.736	0.774	0.866	0.871
EuroSAT	0.958	**0.950**	0.308	0.588	0.681	0.646	0.900	0.893
TinyImagenet	0.365	**0.378**	0.031	0.032	0.032	0.025	0.260	0.217
OxfordFlowers	0.566	**0.659**	0.031	0.051	0.087	0.139	0.290	0.257
STL-10	0.655	**0.650**	0.084	0.142	0.250	0.289	0.480	0.384
Cinic-10	0.673	**0.709**	0.280	0.341	0.391	0.405	0.577	0.586
Average	0.677	**0.702**	0.211	0.289	0.336	0.346	0.539	0.512

Table 2: The AUC metric of the MIAs. Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	**Baseline**	**Reg**	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	**DAP**$_t$	**DAP**$_v$
Cifar-10	0.648	0.631	0.505	0.526	0.519	**0.503**	0.507	0.505
Cifar-100	0.603	0.621	**0.501**	0.515	0.507	0.506	0.516	0.506
FMNIST	0.552	0.562	**0.502**	**0.502**	0.504	0.505	0.507	0.506
EuroSAT	0.544	0.528	0.505	0.502	**0.500**	0.502	0.501	0.501
TinyImagenet	0.603	0.592	0.514	**0.501**	0.521	0.504	0.516	0.509
OxfordFlowers	0.761	0.765	0.543	0.537	0.526	0.532	0.538	**0.521**
STL-10	0.604	0.563	0.502	0.524	0.505	**0.501**	0.508	0.506
Cinic-10	0.572	0.614	**0.501**	0.514	0.511	0.507	0.513	0.507
Average	0.611	0.609	0.509	0.514	0.511	**0.507**	0.513	0.508

EuroSAT dataset, where the DAP$_t$ and DAP$_v$ settings result in accuracy gains of 22% and 21% respectively, compared to the best-performing DP model. Concerning classification accuracy, in summary, the Reg model attains the highest accuracy performance on average, followed by DAP$_t$ and DAP$_v$.

Table 2 collects the results concerning MIAs conducted on the target models. Here, DAP$_t$ and DAP$_v$ exhibit average AUCs of 51.3% and 50.8%, respectively. This indicates that both approaches effectively safeguard against MIAs, rendering the attacks nearly akin to random guessing and achieving performances competitive with DP models. The Reg model, instead, nearly matches the privacy level of the baseline. This discrepancy between our results and the findings of Lomurno and Matteucci [5] is due to the different experimental conditions. In particular, our experiments run until convergence without fixing a specific number of epochs. In summary, DAP proves to be an effective training framework that produces models resilient against MIAs.

Table 3 collects the outcomes concerning the proposed AOP metric. These findings highlight the ability of DAP to produce private models that, at the same time, demonstrate competitive performance. In fact, DAP$_t$ and DAP$_v$ outperform DP and regularization in terms of the accuracy-privacy tradeoff. Concerning the Reg model, despite its susceptibility to MIAs, it is still a viable intermediate choice due to its superior accuracy. Conversely, DP models are extremely effective in MIA prevention but this advantage comes at the expense of the final accuracy, resulting in underperforming models. Notably, the AOP follows the trend of the privacy budget ε. In fact, the most private model (DP with $\varepsilon = 0.5$) is also the least performing due to the impactful addition of gradient noise.

Lastly, Table 4 collects the time per epoch required to train each model. Here, the Baseline and Reg models emerge as the fastest, while the DP models require about 8 times as long. DAP, on the other hand, manages to produce strong results both in terms of privacy and accuracy, requiring only twice the time of the baseline model.

Summarizing the results in terms of accuracy, privacy, AOP, and training time, DAP offers a better tradeoff than DP and regularization. Specifically, the DAP$_t$ setting produces high-performance models, albeit less private. In contrast, the DAP$_v$ setting produces models with strong privacy at a slight accuracy expense. Both settings handle the performance-privacy tradeoff far more effectively than DP in significantly less time.

Table 3: The AOP metric on the test sets. Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	DAP$_t$	DAP$_v$
Cifar-10	0.467	0.509	0.307	0.338	0.387	0.413	**0.607**	0.601
Cifar-100	0.331	**0.345**	0.039	0.078	0.087	0.070	0.296	0.269
FMNIST	0.765	0.733	0.600	0.695	0.724	0.759	0.842	**0.850**
EuroSAT	0.809	0.852	0.302	0.583	0.681	0.641	**0.896**	0.889
TinyImagenet	0.251	**0.270**	0.029	0.032	0.029	0.025	0.244	0.209
OxfordFlowers	0.244	**0.281**	0.026	0.044	0.079	0.123	0.250	0.237
STL-10	0.449	**0.513**	0.083	0.129	0.245	0.288	0.465	0.375
Cinic-10	0.514	0.470	0.279	0.337	0.386	0.399	0.552	**0.570**
Average	0.479	0.497	0.208	0.280	0.327	0.340	**0.519**	0.500

Table 4: Training time per epoch required for each experiment, measured in seconds. Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	DP	DAP$_t$	DAP$_v$
Cifar-10	5.6	**5.9**	46.5	17.8	17.7
Cifar-100	8.9	**9.5**	48.2	17.9	17.9
FMNIST	12.1	**11.7**	53.2	23.4	23.5
EuroSAT	4.7	**5.5**	72.7	10.2	9.5
TinyImagenet	31.7	**32.6**	338.5	62.4	61.4
OxfordFlowers	1.2	**1.8**	20.8	2.3	2.4
STL-10	1.7	**2.5**	34.9	2.5	2.8
Cinic-10	30.0	**22.1**	106.5	41.2	42.8
Average	12.0	**11.5**	90.2	22.2	22.2

6 Conclusion

In this work, we introduced the Discriminative Adversarial Privacy (DAP) framework and the Accuracy Over Privacy (AOP) metric. The goal of DAP is to produce deep learning models resilient to Membership Inference Attacks (MIAs), while the AOP summarizes the privacy-accuracy tradeoff in a single value. As shown in the experiments, DAP demonstrated superior ability in maintaining a beneficial balance between model performance and privacy, outperforming models based on Differential Privacy (DP). The AOP metric has effectively encapsulated these results, providing a concise yet robust evaluation criterion. On top of that, DAP required considerably less computational overhead, thus accelerating the training process with respect to DP. Collectively, our contributions offer a promising approach to the development and evaluation of deep learning models resilient against MIAs, providing an optimal balance between execution time, accuracy, and privacy.

Acknowledgment

This project has been supported by AI-SPRINT: AI in Secure Privacy-pReserving computINg conTinuum (European Union H2020 grant agreement No. 101016577) and FAIR: Future Artificial Intelligence Research (NextGenerationEU, PNRR-PE-AI scheme, M4C2, investment 1.3, line on Artificial Intelligence).

References

[1] J.P. Albrecht. How the GDPR Will Change the World. *European Data Protection Law Review*, 2(3):287–289, 2016.

[2] Max Parasol. The impact of China's 2016 Cyber Security Law on foreign technology firms, and on China's big data and Smart City dreams. *Computer Law & Security Review*, 34(1):67–98, February 2018.

[3] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[4] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[5] Eugenio Lomurno and Matteo Matteucci. On the utility and protection of optimization with differential privacy and classic regularization techniques. In *Machine Learning, Optimization, and Data Science: 8th International Workshop, LOD 2022, Certosa di Pontignano, Italy, September 19–22, 2022, Revised Selected Papers, Part I*, pages 223–238. Springer, 2023.

[6] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

[7] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[8] Yufei Chen, Chao Shen, Yun Shen, Cong Wang, and Yang Zhang. Amplifying membership exposure via data poisoning. *arXiv preprint arXiv:2211.00463*, 2022.

[9] Xinlei He, Hongbin Liu, Neil Zhenqiang Gong, and Yang Zhang. Semi-leak: Membership inference attacks against semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 365–381. Springer, 2022.

[10] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. This person (probably) exists. identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018*, 2021.

[11] Hailong Hu and Jun Pang. Membership inference attacks against gans by leveraging over-representation regions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2387–2389, 2021.

[12] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316*, 2023.

[13] Hailong Hu and Jun Pang. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023.

[14] Zihan Wang, Na Huang, Fei Sun, Pengjie Ren, Zhumin Chen, Hengliang Luo, Maarten de Rijke, and Zhaochun Ren. Debiasing learning for membership inference attacks against recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1959–1968, 2022.

[15] Wei Yuan, Chaoqun Yang, Quoc Viet Hung Nguyen, Lizhen Cui, Tieke He, and Hongzhi Yin. Interaction-level membership inference attack against federated recommender systems. *arXiv preprint arXiv:2301.10964*, 2023.

[16] Tomas Chobola, Dmitrii Usynin, and Georgios Kaissis. Membership inference attacks against semantic segmentation models. *arXiv preprint arXiv:2212.01082*, 2022.

[17] Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. Label-only membership inference attacks and defenses in semantic segmentation models. *IEEE Transactions on Dependable and Secure Computing*, 2022.

[18] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

[19] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[20] Eugenio Lomurno, Alberto Archetti, Lorenzo Cazzella, Stefano Samele, Leonardo Di Perna, and Matteo Matteucci. Sgde: Secure generative data exchange for cross-silo federated learning. *arXiv preprint arXiv:2109.12062*, 2021.

[21] Ken Liu, Shengyuan Hu, Steven Z Wu, and Virginia Smith. On privacy and personalization in cross-silo federated learning. *Advances in Neural Information Processing Systems*, 35:5925–5940, 2022.

[22] Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific reports*, 12(1):1953, 2022.

[23] Dingfan Chen, Ning Yu, and Mario Fritz. Relaxloss: defending membership inference attacks without losing utility. *arXiv preprint arXiv:2207.05801*, 2022.

[24] Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In *International conference on machine learning*, pages 5345–5355. PMLR, 2021.

[25] Ryan Webster, Julien Rabin, Loïc Simon, and Frédéric Jurie. Generating private data surrogates for vision related tasks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 263–269. IEEE, 2021.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto University press*, 2009.

[28] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[29] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[30] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[32] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[33] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.