POLITECNICO MILANO 1863
AIRLAB
ARTIFICIAL INTELLIGENCE AND ROBOTICS LAB

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

Eugenio Lomurno, Alberto Archetti, Francesca Ausonio, and Matteo Matteucci

BMVC 2023

Background

Among the numerous advancements in the field of Deep Learning, only a handful address model resistance to privacy attacks. Current techniques predominantly rely on **Differential Privacy**. While they offer solid theoretical guarantees, they also come with significant drawbacks, notably in performance and complexity. Interestingly, also **regularization techniques** exhibit side effects that can occasionally benefit privacy, even though they are not explicitly designed for it.

Contributions

- A novel learning technique called **Discriminative Adversarial Privacy** (DAP), that combines adversarial learning and membership inference attacks to ensure an optimal balance between model performance, complexity, and privacy.
- A novel **loss function** that is tailored to simultaneously minimise the model prediction error while maximising the attacker's error.
- A novel metric called **Accuracy Over Privacy** (AOP), to capture the performance-privacy trade-off effectively.

Discriminative Adversarial Privacy (DAP)

Consider a **classifier** C_{DAP} and a pre-trained **Membership Inference Attack** (MIA) model D_{MIA}. Let $x, y \sim p(x,y)$ be a sample's features and labels extracted from the data domain, and t the current iteration index. Then, the learning procedure of DAP is

$$\min_{C_{\text{DAP}}} \max_{D_{\text{MIA}}} \mathbb{E}_{x\sim p(x)}[\log C_{\text{DAP}}(x,t)] + \beta \mathbb{E}_{x,y\sim p(x,y)}\left[\log\left(1 - D_{\text{MIA}}(C_{\text{DAP}}(x,t), y)\right)\right]$$

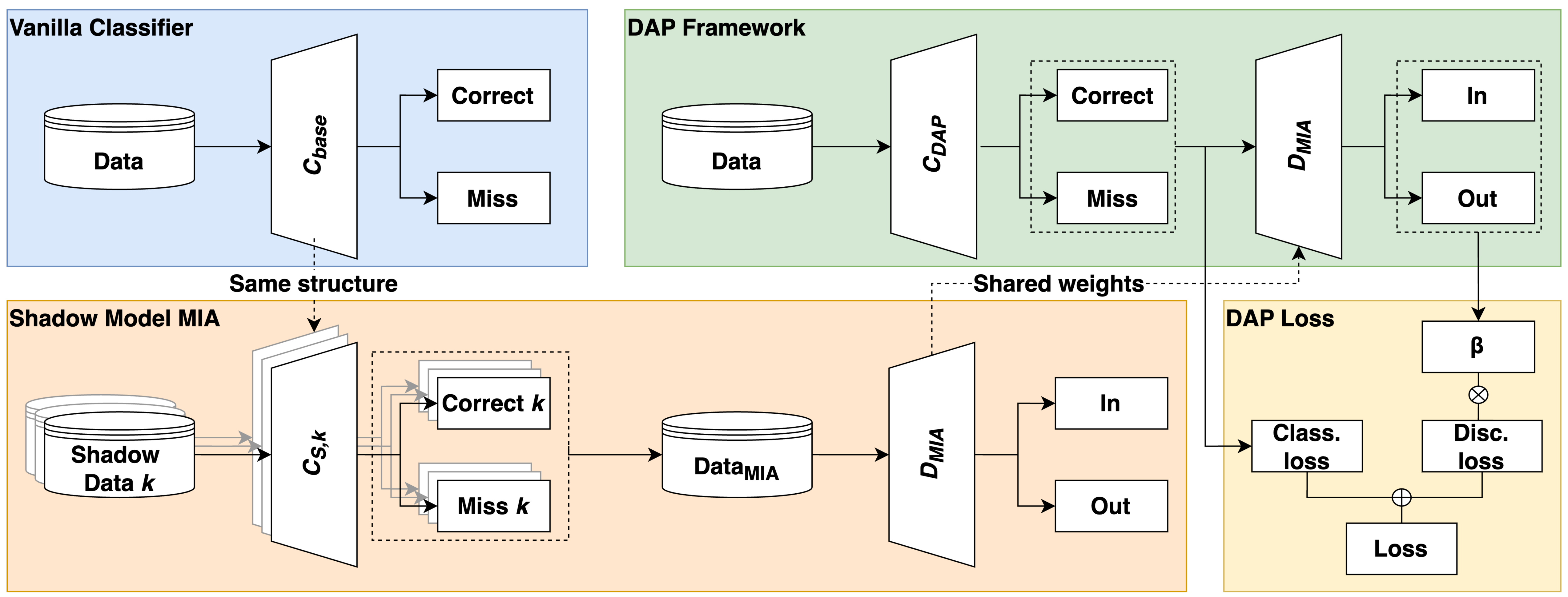
where β is a dynamic loss balancing parameter. β is a function of a pre-defined ratio factor r. β is defined as

$$\beta(C_{\text{DAP}}, D_{\text{MIA}}, t, r) = \begin{cases} r \cdot \dfrac{\mathbb{E}[\log C_{\text{DAP}}(x, t-1)]_v}{\mathbb{E}\left[\log\left(1 - D_{\text{MIA}}(C_{\text{MIA}}(x, t-1), y)\right)\right]_v} & \text{if } t > 0 \\ 1 & \text{otherwise} \end{cases}$$

Given the accuracy ACC from C_{DAP}, the area under the curve AUC_{MIA} from D_{MIA} and the privacy parameter λ, the training step with the optimal model parameters is found by maximising the AOP metric:

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

Architecture



Results

Accuracy (ACC)

Dataset	Baseline	Reg	$\varepsilon=0.5$	$\varepsilon=1$	$\varepsilon=2$	$\varepsilon=4$	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

Membership Inference Attack Area-Under-the-Curve (AUC_{MIA})

Dataset	Baseline	Reg	$\varepsilon=0.5$	$\varepsilon=1$	$\varepsilon=2$	$\varepsilon=4$	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

Accuracy Over Privacy (AOP)

λ	Baseline	Reg	$\varepsilon=0.5$	$\varepsilon=1$	$\varepsilon=2$	$\varepsilon=4$	DAP_t	DAP_v
1	0.567	**0.587**	0.210	0.284	0.331	0.343	0.529	0.506
2	0.479	0.497	0.208	0.280	0.327	0.340	**0.519**	0.500
5	0.301	0.316	0.204	0.267	0.316	0.330	**0.492**	0.483
10	0.154	0.168	0.197	0.249	0.299	0.317	0.451	**0.456**
20	0.049	0.064	0.184	0.222	0.271	0.292	0.386	**0.409**
50	0.003	0.008	0.155	0.175	0.217	0.238	0.262	**0.305**