Privacy-Preserving Publication of Sensitive Data using Differentially Private Generative Adversarial Networks

SFU
SIMON FRASER UNIVERSITY

Ricardo Silva Carvalho
Computing Science, Simon Fraser University

MOTIVATION

Share sensitive data to support critical research or help solve problems:
→ **Preserving privacy** of entries of the data
→ **Maintaining usefulness** of data

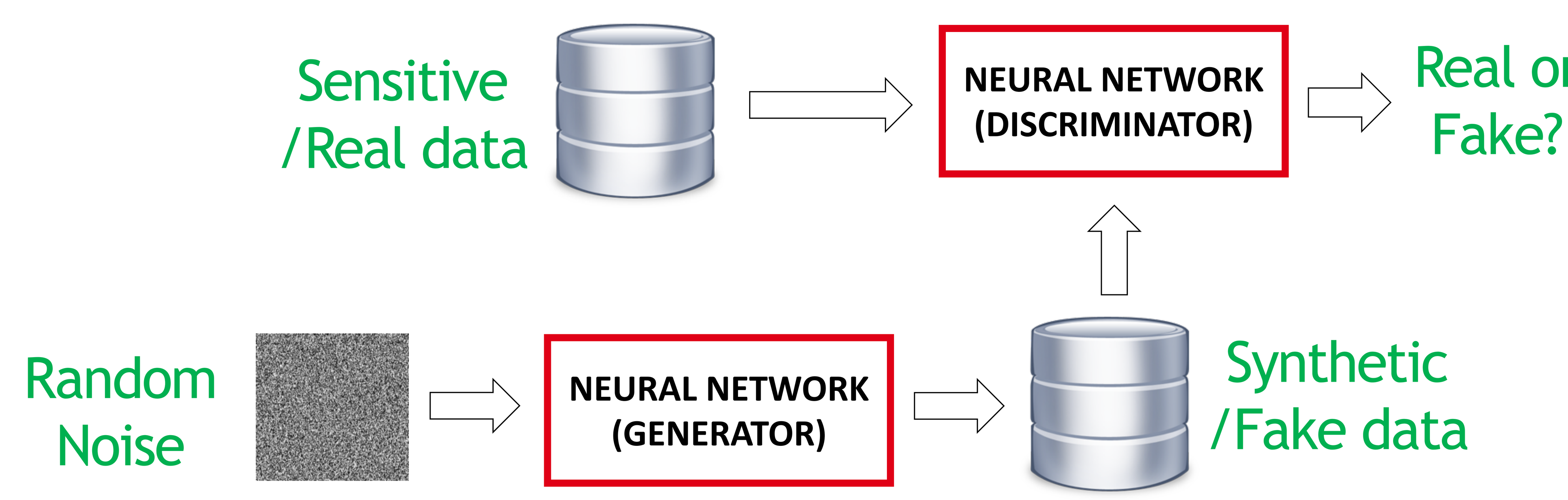
FOCUS

We illustrate the problem for the case of:
→ Data with one **individual/user/patient** per entry
→ Goal of using data is to train a **classification** model
→ Existing a **trusted-curator** of the data

ONE SOLUTION

Generative Adversarial Networks (GANs):
→ **Learn distribution** of sensitive data
→ Generate **synthetic** data

BAD NEWS: GANs can still be vulnerable.
For example: Membership attack [CCS'17]



IMPROVED SOLUTION

GANs + Differential Privacy (DP):
→ Bound **maximum change** (sensitivity: Δf)
→ Add random **noise** proportional to $O(\Delta f/\varepsilon)$
→ **Discriminator** needs to satisfy DP

Deep Learning [CCS'16]:

Clip gradient
$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i)/\max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$
Add noise
$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
Descent
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

A randomized algorithm $\mathcal{M}$ is (ε, δ)-differentially private if for all neighboring datasets D and D' and all sets of outputs $\mathcal{O} \subseteq \text{Range}(\mathcal{M})$:

$$\Pr[\mathcal{M}(D) \in \mathcal{O}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{O}] + \delta$$

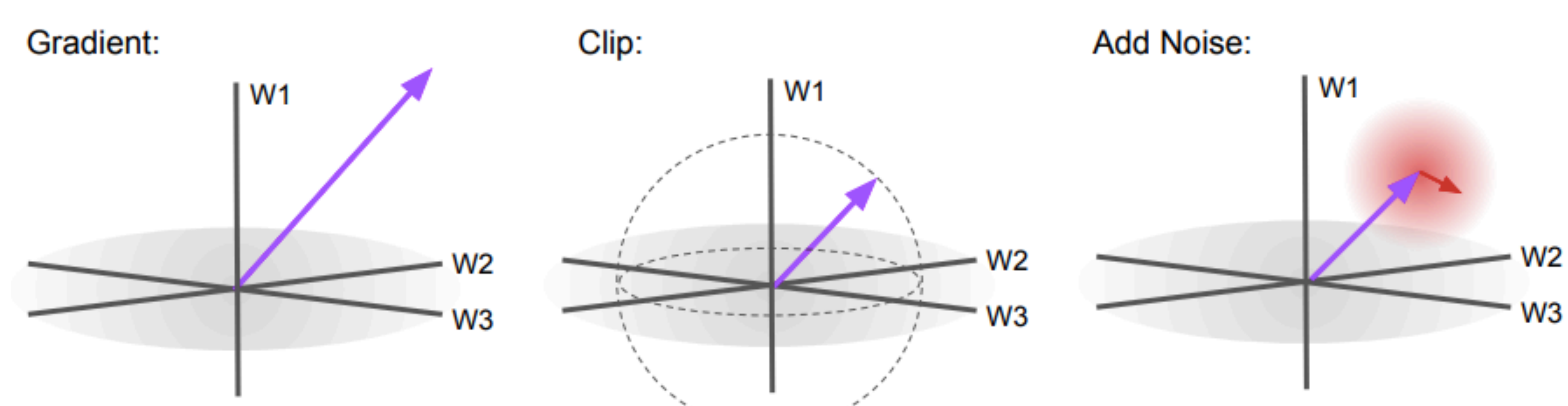


Image from: https://chriswaites.com/posts/differentially-private-deep-learning/

BENCHMARKING

Implementation on TensorFlow 2.0:
→ To the best of our knowledge, the first open implementation of a DP GAN with TensorFlow **2.0**
→ **New custom Optimizer, carefully applying non-trivial DP constraints**
→ Comparing to DP-CGAN [CVPR'2019]

AuROC	Real	CGAN	DP-CGAN TF 1.15 (M=1)	DP-CGAN TF 2.0 (M=B)	+ Separate batches	+ LeakyRelU (alpha=0.2)
LR	0.9217	0.9110	0.8121	0.8642	0.6308	0.8088
MLP	0.9760	0.9106	0.8396	0.8858	0.6586	0.8263

Table 1: AuROC on test data of MNIST using standard sklearn lib of models trained on fake data. Results are average of 3 trials, using differential privacy with parameters $\varepsilon = 9.6$ and $\delta = 10^{-5}$.

Zdim: *100* ***GEN:*** *FC(128) + RelU + FC(784)* ***DISC:*** *FC(128) + RelU + FC(1)*

EXPERIMENTS

Dataset of patients for Thyroid Disease:
→ 7200 patients split into 52.4% (train) + 47.6% (test)
→ 21 attr. (15 binary, 6 continuous) and 3 classes

DP-CGAN: ***GEN:*** *FC(128) + RelU + FC(21)* ***with Zdim:*** *100*
DISC: *FC(128) + RelU + FC(1)*

Table 2: AuROC on test data for model trained with real or fake data

$\varepsilon = 3.7$, $\delta = 10^{-5}$	Real	DP-CGAN TF 2.0 (M=B)
MLP: AuROC	0.9858	0.9746

MLP trained with same GridSearchCV for both

REFERENCES

[CCS'17]: Hitaj et al., Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning.
[CCS'16]: Abadi et al., Deep Learning with Differential Privacy.
[CVPR'19]: Torkzadehmahani et al., DP-CGAN : Differentially Private Synthetic Data and Label Generation.
[Generative Adversarial Networks]: Ian Goodfellow, 2014.
[Differential Privacy]: Cynthia Dwork, 2016.
[Conditional GANs]: Mirza, 2014.