Benchmarking Differentially Private Residual Networks for Medical Imagery

Sahib Singh*,[1,2,3] Harshvardhan D. Sikka*,[1,3,4] Sasikanth Kotti[1,5] Andrew Trask[1,6]

*Equal Contribution [1]OpenMined [2]Ford R&A [3]Manifold Computing Group [4]Harvard University [5]Indian Institute of Technology Jodhpur [6]University of Oxford.

Focus

We measure the effectiveness of Differential Privacy (DP) when applied to medical imaging. We compare two robust differential privacy mechanisms: Local-DP and DP-SGD and benchmark their performance, analyzing the trade-off between the accuracy and the level of privacy the model guarantees. We also examine how useful these privacy guarantees prove to be in a real world medical setting.

Key Preliminaries

Differential Privacy: the introduction of randomized noise to ensure privacy through plausible deniability.

Local DP: an algorithm **π** satisfies ε-LDP where ε > 0 if and only if for any input *v* and *v'*

$$\forall y \in Range(\pi) : P[\pi(v) = y] \le e^{\epsilon} P[\pi(v') = y]$$

DP - SGD: a modification of SGD that bounds the sensitivity of each gradient and uses a moments accountant algorithm to amplify and track the privacy loss across weight updates.

Laplace Distribution: a symmetric version of the exponential distribution. The distribution centered at 0 (i.e. μ = 0) with scale β has the following probability density function.

$$Lap(x|b) = \frac{1}{2\beta} \exp(\frac{-|x|}{\beta})$$

Laplace Mechanism: Laplace Mechanism independently perturbs each coordinate of the output with Laplace noise (from the Laplace distribution having mean zero) scaled to the sensitivity of the function.

Correspondence: sahibsin@alumni.cmu.edu

Experimental Setup

Architecture: Pretrained Resnet-18 trained in all experiments over 50 epochs with a 0.01 learning rate, and batch size of 128.

LDP Experiment: 3 other versions of the dataset were generated by the addition of different perturbations to the images. The tradeoff in accuracy with varying scales of perturbations (β=1; β= 2; β= 4) were examined.

DP - SGD Experiment: We clip the gradient in the l2 norm, add random noise to it and then multiply it by the learning rate before updating the model parameters. Perturbations of (β=1; β= 2; β= 4) were examined.

Datasets

APTOS: Retinal Scans for Diabetic Retinopathy.

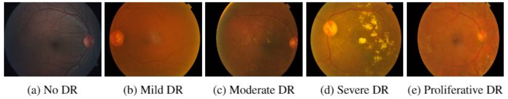
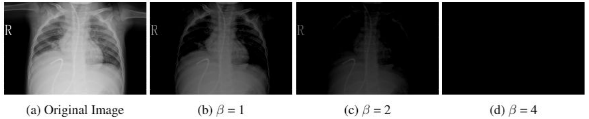

(a) No DR (b) Mild DR (c) Moderate DR (d) Severe DR (e) Proliferative DR

Figure 1. Examples from the APTOS Blindness Detection Dataset. Samples progress in severity of Diabetic Retinopathy.

Chest X-Rays Dataset: Chest Radiography for Pneumonia



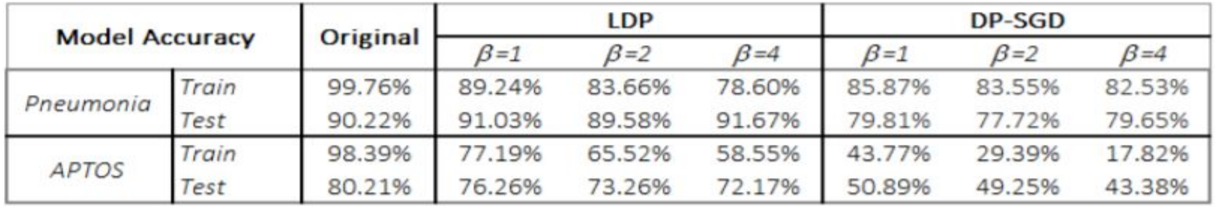
(a) Original Image (b) $\beta = 1$ (c) $\beta = 2$ (d) $\beta = 4$

Figure 2. Comparison of sample image from dataset before and after Local DP based Obfuscation.

Results: the Accuracy-Privacy Tradeoff

In some cases the training accuracy is lower than the test accuracy (See Figure 4 below). One possible explanation for this is that Local-DP adds noise to the training data, making the latent features harder to learn. Later when we run the model on the test data it performs better because the latent features are now relatively easy to capture since the model has already learned representations in a noisy scenario.

While Local-DP maintains the theoretical guarantee of Differential Privacy, it does not always provide the visual privacy we expect. In some cases the image was completely blurred out while in a few others there was hardly any visual change to the image (See Chest X-Rays Dataset above). DP-SGD is the mechanism of choice for ensuring more robust privacy.



Model Accuracy		Original	LDP			DP-SGD		
			β=1	*β=2*	*β=4*	*β=1*	*β=2*	*β=4*
Pneumonia	*Train*	99.76%	89.24%	83.66%	78.60%	85.87%	83.55%	82.53%
	Test	90.22%	91.03%	89.58%	91.67%	79.81%	77.72%	79.65%
APTOS	*Train*	98.39%	77.19%	65.52%	58.55%	43.77%	29.39%	17.82%
	Test	80.21%	76.26%	73.26%	72.17%	50.89%	49.25%	43.38%

Figure 4. Model Accuracy (%) — DP mechanism vs Dataset