SIL765

Midterm Presentation

DEEP LEARNING WITH DIFFERENTIAL PRIVACY

Introduction

Neural Networks & Deep Learning have achieved exceptional results in various domains such as Image Classification, NLP & healthcare.

The advancements in these areas have been possible due to availability of large & representative datasets which are used to train the neural network.

Problem

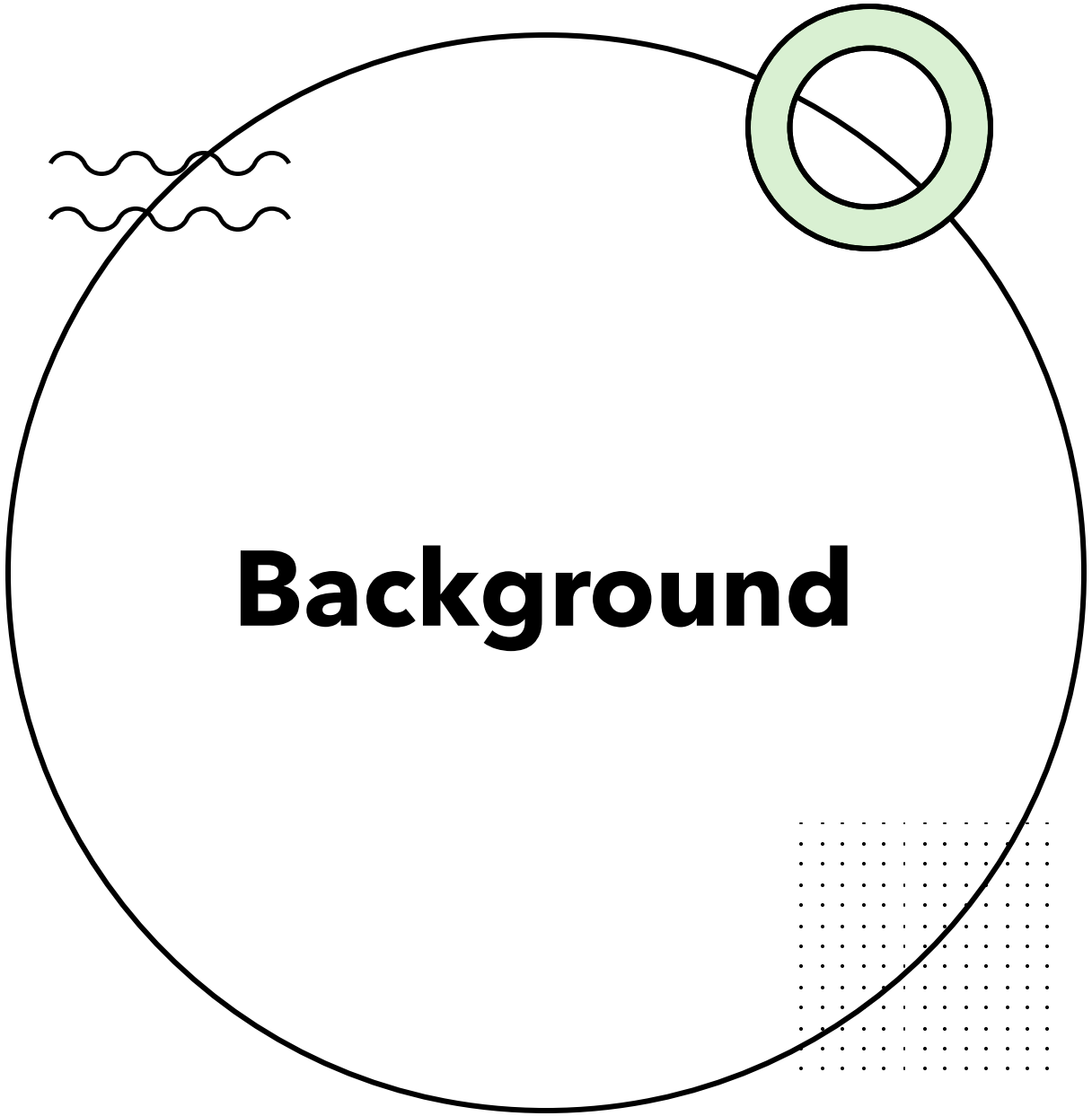
The Datasets used to train neural network are often crowd-sourced and may contain sensitive information.

Model inversion attacks can exploit confidential information and can also lead to wrong predictions on the data.

Solution

Implement algorithms to train deep neural networks with non-convex objective to compute privacy cost within the framework of differential privacy.

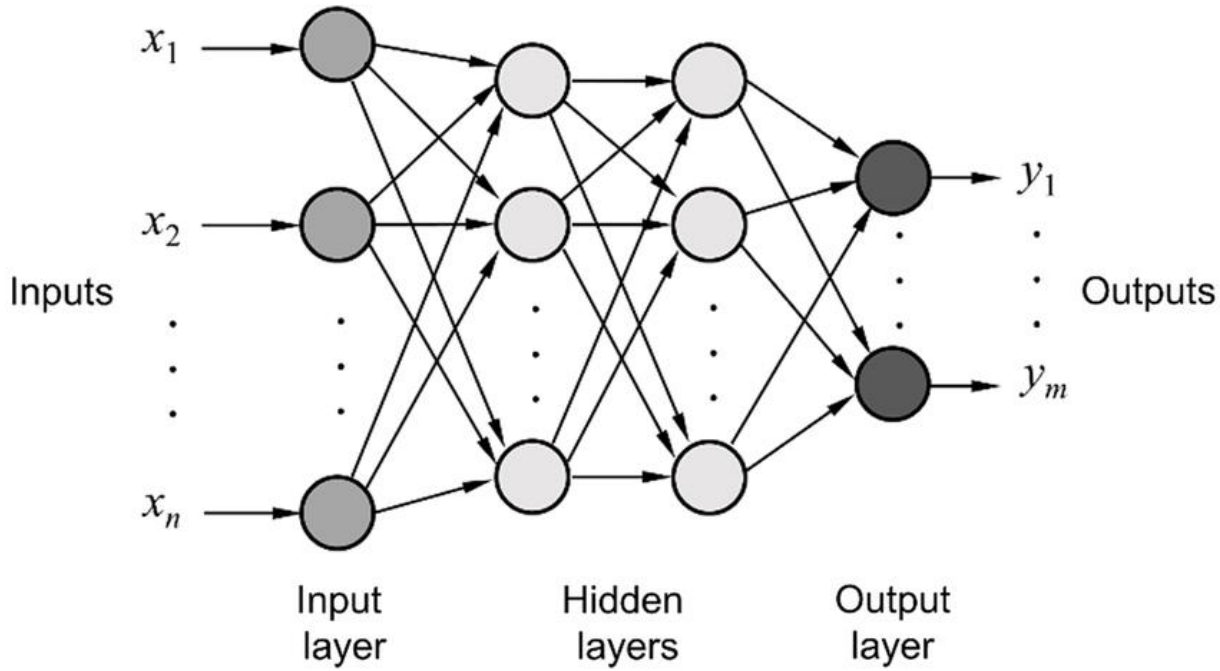
Modest privacy budget, feasible Software and Computational complexity, Training efficiency while maintaining the quality and performance of the model.



Background

Before diving into the differentially private SGD algorithm let’s take a look at basic principles of Deep Learning & definition of Differential Privacy.

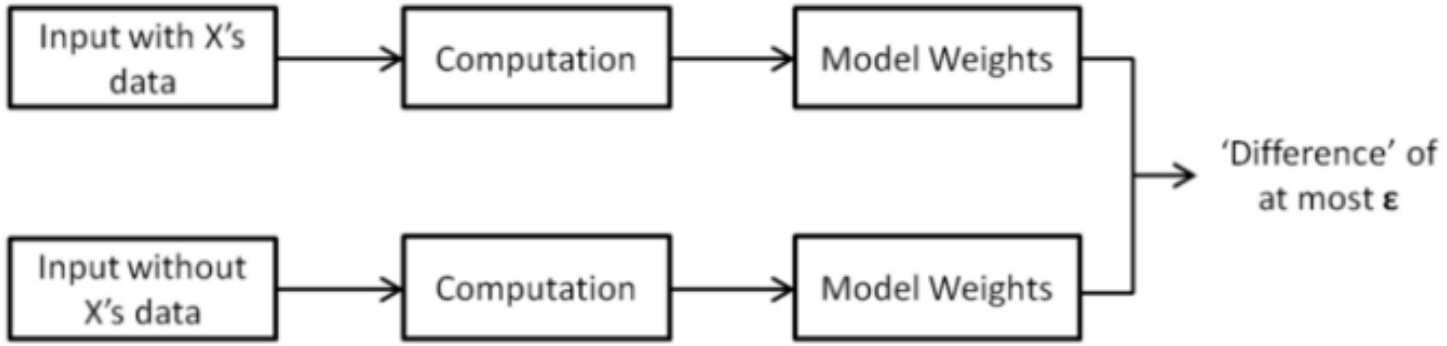
Deep Learning



- Compositions of basic building blocks called units in a multi-layered structure.
- Parametrized functions such as Sigmoid & ReLU
- backpropagation algorithm making use of SGD to minimize the average loss L(θ)

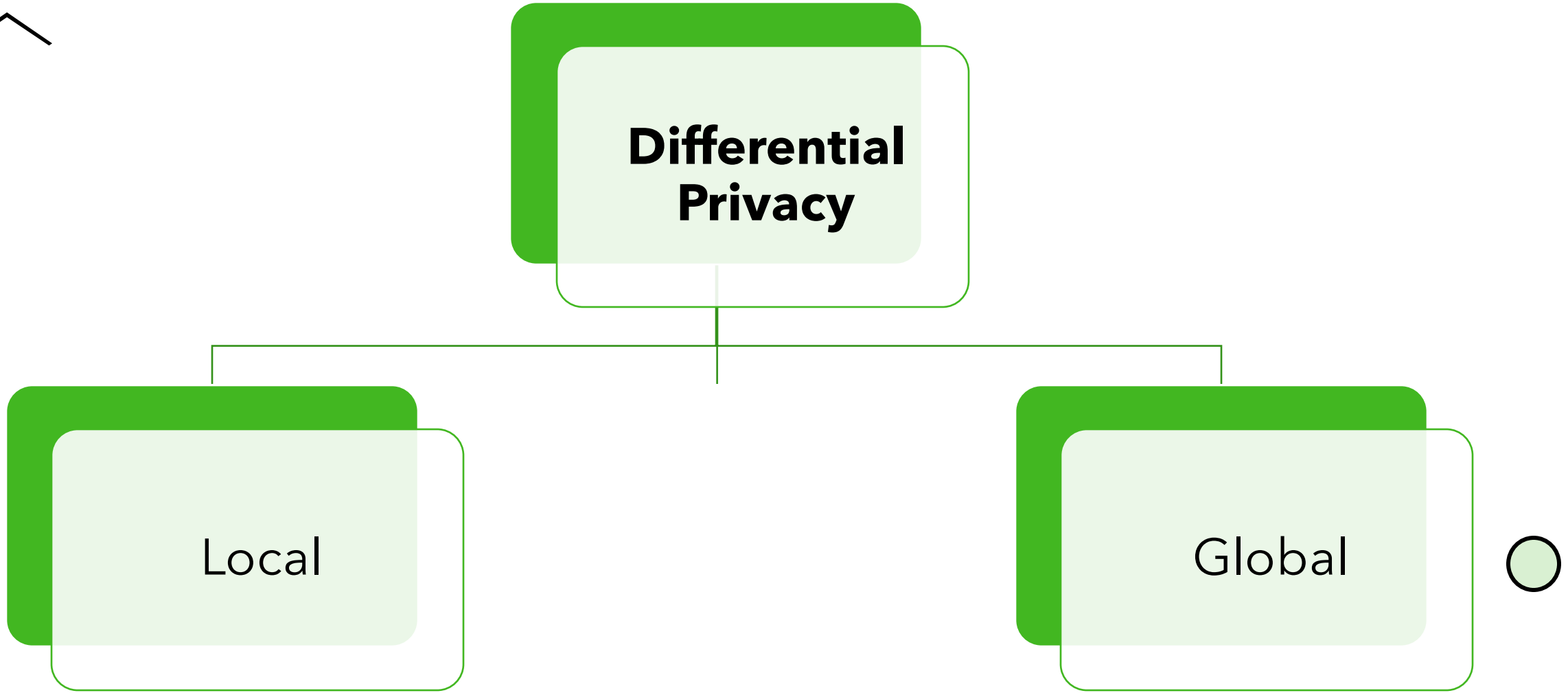
- **$L(\theta) = 1/N \, \Sigma_i \, L(\theta, x_i)$**

Differential Privacy



- Constraint on algorithms to limit the amount of information that is public.
- A system to publically share data and preserve user related sensitive information.
- Differential Privacy states that the distribution of the output M(D) on the databse D is nearly same as M(D').

- **$Pr[M(D) \epsilon S] < e^{\varepsilon} Pr[M(D') \epsilon S] + \delta$**


Differential Privacy
Local
Global

Types of Differential Privacy

Local Differential Privacy	Global Differential Privacy
• Distortion or randomness has been introduced in the data. • The greater the privacy protection, the less accurate the results.	• Distortion or randomness is introduced in the output. • More accurate results with same level of privacy protection.

Differentially Private SGD Algorithm

Input: Examples {x1, x2,.....,xN}, Learning Rate η, noise scale σ, batch size L, gradient norm bound C.

For t ϵ [T]:

Take a random sample with sampling probability L/N

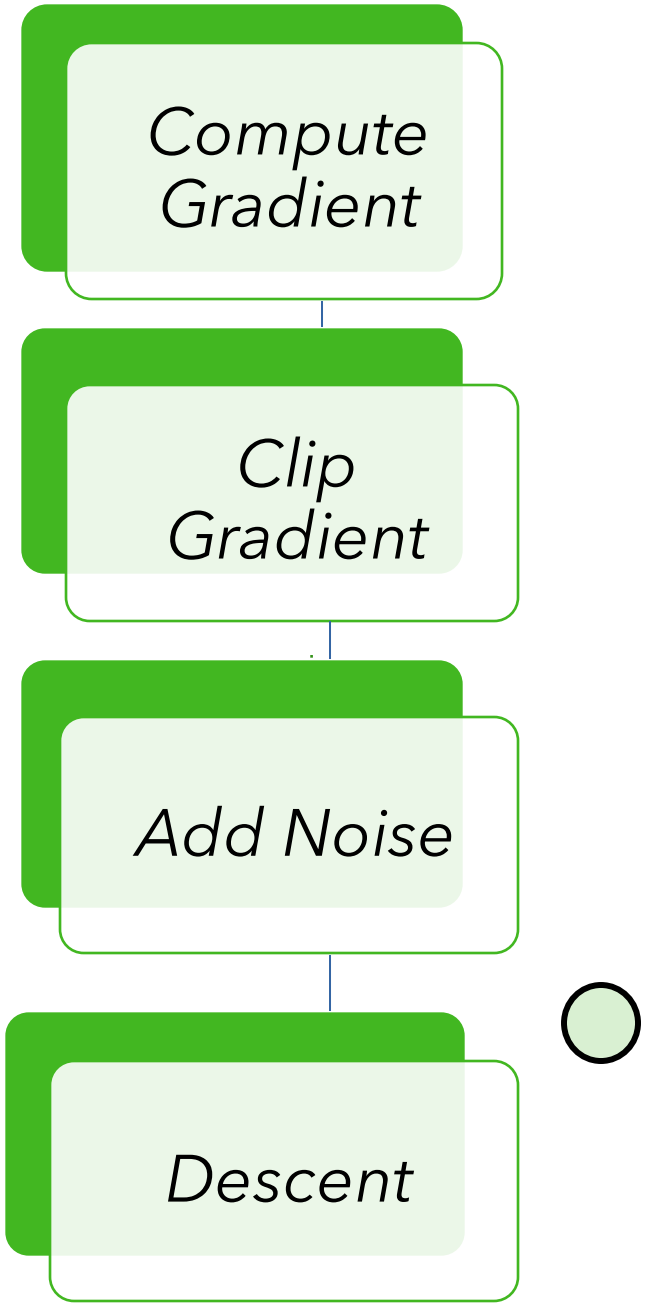
Compute Gradient

Clip Gradient

Add Noise

Descent

Output: θT and compute the overall privacy cost (ϵ, δ) using a privacy accounting method.



Proposed Solution

Applied tensor transforms and normalized the data (MNIST Digit Dataset).

Separated the test dataset into 2 – public & private.

Created a 5-layered Convolutional Neural Network to train the model .

Created teacher models as a function of learning rate & epochs.

Proposed Solution

Trained the teacher models to evaluate the labels on an unlabeled test datasets.

Evaluated the labels generated on the unlabeled dataset using PATE analysis. This gives an estimate of the trade-off between privacy & accuracy.

Randomized the labels generated by applying Laplacian noise to conserve privacy .

After generating the labels a new model has been created which is differentially private.

Proposed Solution
(Demonstration)

- torch, tensorflow
- syft
- 12 GB NVIDIA Tesla K80 GPU
- numpy, pate
- Code can be found here

Proposed Plan

- Observe the epoch-ϵ relation with varying delta on the differentially private model.
- Observe the accuracy-epoch relation with varying delta on the differentially private model.
- Observe the accuracy-ϵ relation with varying delta on the differentially private model .
- Compare the results obtained on the MNIST Dataset with CIFAR Dataset and try to generalize the results.

Related Work (Practical)

- Google's ***RAPPOR*** Project – To report usage statistics for Google Chrome.
- ***Private mean count sketch*** - Apple. To collect emoji usage data, word usage and other information from iPhone users.
- Government agencies – Publish information while ensuring confidentiality.
- Health data streams – Used in privacy preserving collection of personal health data streams.

Related Work (Research)

- ***M. Abadi, A. Chu, I. Goodefellow*** – Privacy attacks on a deep learning model from three aspects: training, data extraction & model extraction.
- ***J. Zhao, Y. Chen, A. W. Zhang*** – Main aspects of differential privacy such as interactive & non-interactive settings, perturbation mechanism & its applications.
- ***J. Kim, Y. Kim, H. H. Nguyen*** – Various threats & the defense mechanism on deep learning privacy models. Aspects where noise can be introduced, gradient or parameterized functions to protect the privacy model.
- ***J. Zhao, T. K. Dang, T. T. Dang*** – Local differential privacy where the private data is regarded as a single tuple. It hides the complete data, not just each individual record from the adversary.

Conclusion

Global Differential Privacy is a better measure to compute privacy costs on deep learning models.

With Use of Differentially Private Models, we can achieve the same accuracy while still preserving user related information.

THANK YOU

Sharique Shamim
2018CS10388