Privacy Preserving Deep Learning: a case study with Microsoft Research Celebrity Data

Meidan Bu
CS230 Deep Learning, Stanford University

Abstract

While data provides tremendous insights, users' personal information is often exposed with limited protection. This project aims to build a privacy preserving deep learning framework that trains and updates models without directly using raw data. Using an image classification task as a case study, the results show that similar accuracy can be achieved with only sharing a small fraction of model parameters, not data.

Data

- Data for this study comes from:
  - Microsoft Research Celebrity Face data,
  - Kaggle dogs and cats classification data

- Randomly sampled 26,142 people's pictures from the database,
- Combined with 25,000 images for dogs and cats,
- Data is separated for initial training and PPDL model updating phase.

Figure 1. Examples of images in dataset.

Table 1. Summary of datasets.

	Training	Validation	Total
Men	6,426	714	7,140
Women	6,621	736	7,357
Dogs	5,850	650	6,500
Cats	5,850	650	6,500
Total	24,747	2,750	27,497

	Data for PPDL Parameter updating
Men	5,567
Women	6,078
Dogs	6,000
Cats	6,000
Total	23,645

CNN Initial Model

1 *Suppose Company A developed a photo app. It trains an image classification model using an initial dataset.*

- Two types of CNN models were tested:
  - a built-from-scratch model with RELU activation
  - a transfer learning model with VGG16 base

- Tried SGD, RMSProp and Adam optimizers.
- 100 epochs, 0.0001 learning rate
- Final initial model used VGG16 with SGD optimizer

	Optimizer	Training Accuracy	Validation Accuracy
CNN built from scratch model	SGD	75.9%	72.1%
	RMSProp	78.3%	74.8%
	Adam	79.2%	77.7%
CNN VGG16 fix all but last layer	SGD	83.5%	79.6%
	RMSProp	84.1%	81.4%
	Adam	85.4%	83.5%

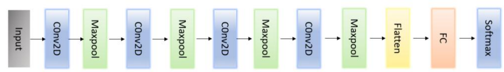


Figure 2. Architecture of the built-from-scratch model.

Privacy Preserving Model Updating

2 *Now the initial model is trained, and the photo app is published. Model needs to be continuously trained using data from users' cellphones. Company A wants to protect users privacy by updating the model without collecting raw pictures.*

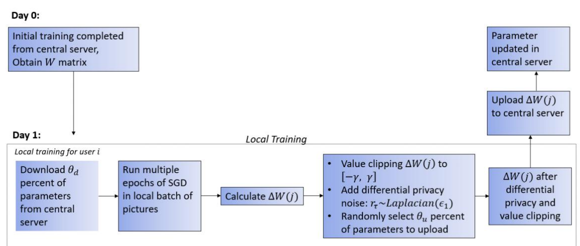


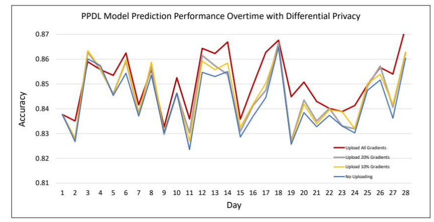
Figure 3. Diagram of local training process for a single user on a day. Final updated model parameters on central server goes to the next day to start a new round of updating process.

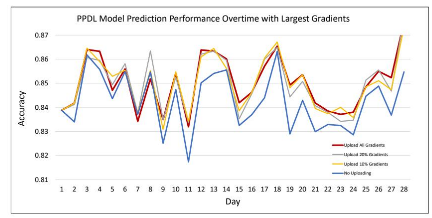
Users privacy is protected through:

- The central server does not collect raw data, but only collect updated gradients from local training,
- Each participant independently trains locally, and uploads only a fraction of gradients.
- The uploaded gradients are further protected through ***differential privacy*** by adding random noise and value clipping, or through only uploading ***largest gradient***.

Results

- By sharing only a small fraction of gradients (10%, and 20% in our case) at each gradient descent step, we can achieve similar accuracy as the privacy violating case of training in a centralized machine with 100% data.
- As expected, the "no update" training has lowest accuracy, and the centralized "all data exposing" training achieves the highest accuracy.





Future work

- Update the parameters in more layers. Currently, the framework only updates the last layer.
- Other model updating mechanism. Currently, only through Stochastic Gradient Descent.
- Improve the initial model performance.

Links

GitHub repo: https://github.com/aruba29/PrivacyPreservingDeepLearning
Presentation: https://youtu.be/QO9U0h57bWE

Acknowledgements: Shawn Chai, Shokri and Shmatikov's work, CS230 teaching team, TA Kaidi Cao