Differential Privacy Has Disparate Impact on Model Accuracy

Eugene Bagdasaryan and Vitaly Shmatikov (CornellTech)
{eugene, shmat}@cs.cornell.edu

Differential privacy

- Differential Privacy bounds the leakage of training data
- Differentially Private Stochastic Gradient Descent (DP-SGD) introduces DP to deep learning by clipping gradients and adding noise:

Algorithm 1: Differentially Private SGD (DP-SGD)

Input: Dataset $(x_1, y_1), ..., (x_N, y_N)$ of size N, batch size b, learning rate η, sampling probability q, loss function $\mathcal{L}(\theta(x), y)$, K iterations, noise σ, clipping bound S, $\pi_S(x) = x * \texttt{min}(1, \frac{S}{||x||_2})$

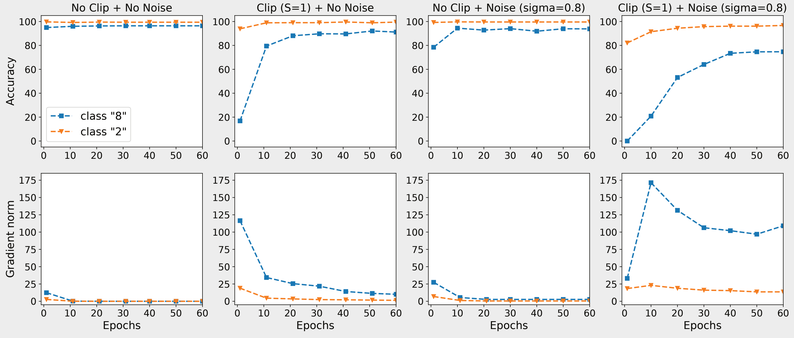
Initialize: Model θ_0

for $k \in [K]$ **do**

- randomly sample *batch* from dataset N with probability q
- **foreach** (x_i, y_i) *in batch* **do**
  - $g_i \leftarrow \nabla\mathcal{L}(\theta_k(x_i), y_i)$
- $g_{batch} = \frac{1}{qN}(\sum_{i \in batch} \pi_S(g_i) + \mathcal{N}(0, \sigma^2 \boldsymbol{I}))$
- $\theta_{k+1} \leftarrow \theta_t - \eta g_{batch}$

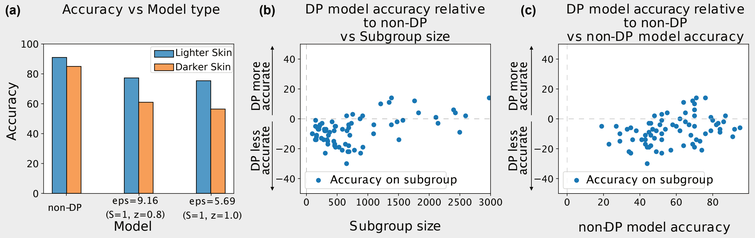
Output: Model θ_K and accumulated privacy cost (ϵ, δ)

- Separately, clipping and noising gradients benefit the training and promote generalization
- However, combination of the two techniques to implement DP-SGD affects underrepresented groups more then well-represented
- Experiments on unbalanced MNIST (class 2 – 500 images, other classes – 5000 images) show that underrepresented class 2 suffers from clipping and noise that result in larger drop in accuracy and bigger gradients:

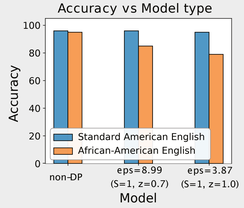


Experiments:

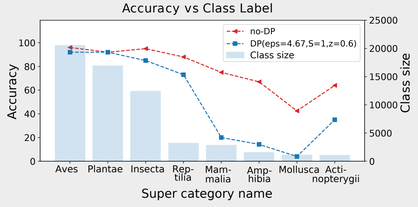
- We pick the gender (a) and age (b, c) prediction tasks on IBM Diversity in Faces dataset using unbalanced selection of individuals with lighter (29,500 images) and darker (500 images) colored skin
- DP models exhibit lower accuracy on underrepresented groups:



- Applying DP-SGD to sentiment classification for Twitter posts that use African-American English (1,000 tweets) and Standard American English (60,000 tweets) results in lower accuracy for the smaller group:

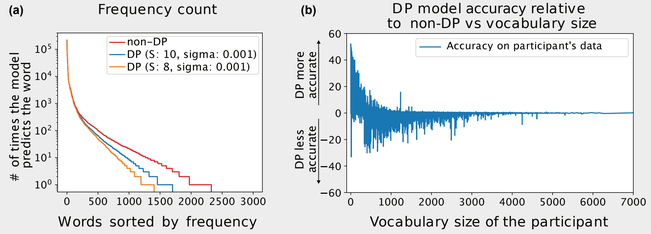


- To evaluate large scale image task we use image classification of species for iNaturalist dataset
- Classification accuracy of non-DP model significantly drops for smaller-size classes:



Federated Learning:

- Federated learning approach trains individual models on participants' data and aggregates produced models into a single global model
- DP can be applied to training preserving participant-level privacy
- We use Reddit dataset with randomly selected 80,000 participants
- The DP global model has smaller active vocabulary than non-DP
- DP-models overfit to participants with simpler vocabularies:



Hyperparameters effects:

- Accuracy drop depends on clipping and noise values (a) as well as batch size (b) and number of epochs (c) but still has significant disparity
- Adding more images thus making the class well-represented reduces the accuracy gap between DP and non-DP models: