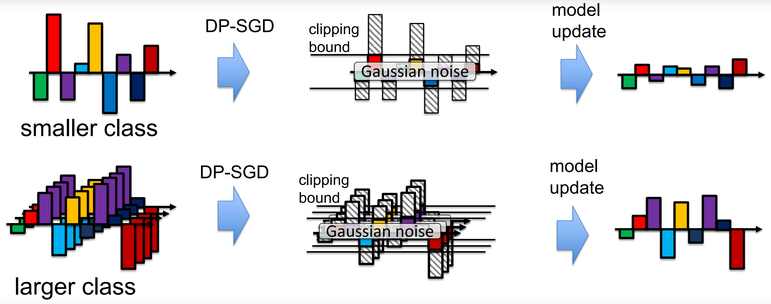
Differential Privacy Has Disparate Impact on Model Accuracy

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

CORNELL TECH

Background

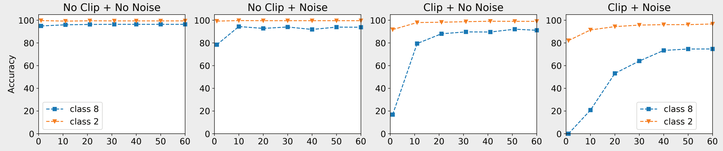
ε-differential privacy (DP)bounds the influence of any single input on the output of a computation. DP machine learning bounds the leakage of training data from a trained model. The ε-parameter controls this bound and thus the tradeoff between ”privacy" and accuracy of the model.



“rich gets richer”

Classes that have more data get to contribute to the model a larger vector, whereas underrepresented groups can contribute only once and their contribution is distorted and constrained.

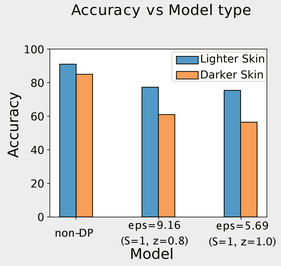
Combining clipping and noise together significantly reduces accuracy on smaller classes:



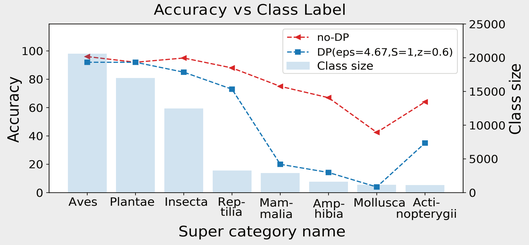
Experiments:

- 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 much lower accuracy on groups with the darker skin:



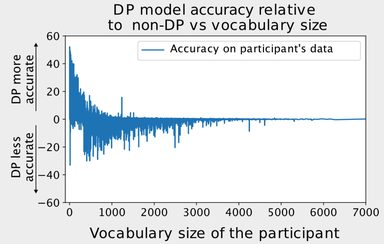
- To evaluate large scale image task we use image classification of species for iNaturalist dataset
- Classification accuracy of non-DP model significantly drops for smaller-size classes:



iNaturalist

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: