CS208 Project Proposal Revision

Karina Huang and Lipika Ramaswamy

Our final project is three-fold:

- Replicate findings by Abadi et al.
  - In Deep Learning with Differential Privacy by Abadi et al., the performance a differentially private version of Stochastic Gradient Descent (SGD) optimizer was evaluated on image data (MNIST and CIFAR-10). We understand that this paper is a precursor to the TensorFlow implementation of differentially private optimizers, introduced on March 6, 2019. Given the models the authors provided and their implementations in TensorFlow, we would like to see whether their conclusions hold conditional on the randomness tensorflow introduces in model compilation.
- Evaluate TensorFlow private optimizers on different neural networks
  - We recognize that the models discussed in the reading might qualify the scope of performance. In particular, the authors employed wide neural networks with a single layer of 1000 hidden units, with the argument that the structure offered the optimal performance in their experiment. While we recognize that a large number of nodes within a network layer denotes more information learned regarding a training sample, we would like to investigate whether simple architectures can perform comparably.
- Implement differentially private gradient computation on other optimizers
  - TensorFlow currently provide implementations of differently private SGD, AdaGrad (Adaptive Gradient Descent), and ADAM (Adaptive Moment Estimation).
  - Since the differently private operations occur at the point of gradient computation, we would like to see whether we can extend this technique to other optimizers, such as Nadam (an updated optimizer that uses ADAM, RMSprop with Nesterov momentum).

Questions we have:

- While the work by Abadi et al. provided insights into how differentially private learning of data performs with respect to utility (as measured by test accuracy), we struggled to understand how the algorithms may be evaluated privacy-wise.
- We're still searching for datasets that might involve sensitive data. One idea is to use a dataset on Harvard and MIT edX classes that we saw in our AC221 class with Jim Waldo. While this dataset complies with FERPA constraints, there are sensitive attributes of individuals contained in the data. One possibility is to model the completion/certification status of an individual taking a given class based on other attributes including, but not limited to, their location, nationality, number of forum posts/comments/tags. We'd plan to ask for permission to use data if this seems like a reasonable dataset to use.