Learning Private CNN Language Models

Mila Faye Schultz | `milafaye@stanford.edu` | Department of Computer Science

Can we train a “private” LM?

Language models are trained on large corpuses of sensitive personal data

User are concerned about social media and other companies’ use of their personal data

Privacy concerns may make training models directly on all text at a central server undesirable or even impossible

Task

Train a CNN language model [Dauphin et al. 2016] that can federate training between clients. The server does not see the clients’ data and only receives updates to the model

Apply privacy techniques including client-side model clipping and server-side Gaussian noise while training, the `DP-FedAvg` algorithm [McMahan et al. 2018]

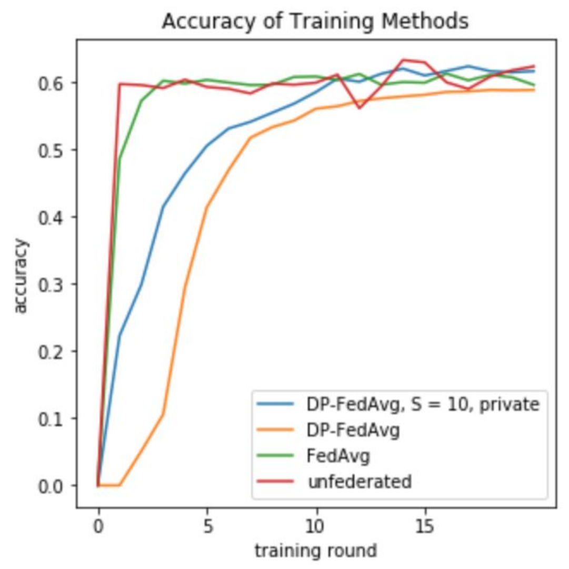
Analyze the impact of privacy-preserving modifications to the training process

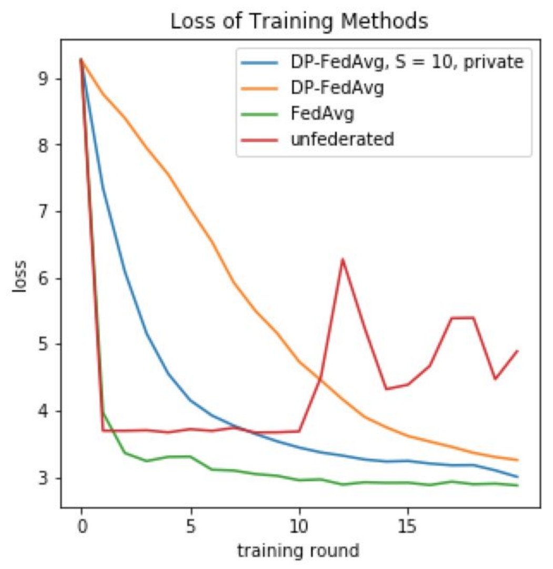
Data

Reddit comments dataset from 2018 [Reddit]

- 22,389 client databases of comments from a single author, each with 4000 word tokens
- Top 10k vocabulary from GloVe
- Held-out test set of 113,000 word tokens
- Input sequence length = 8
- ~100 clients per round, 20 rounds of training

Results

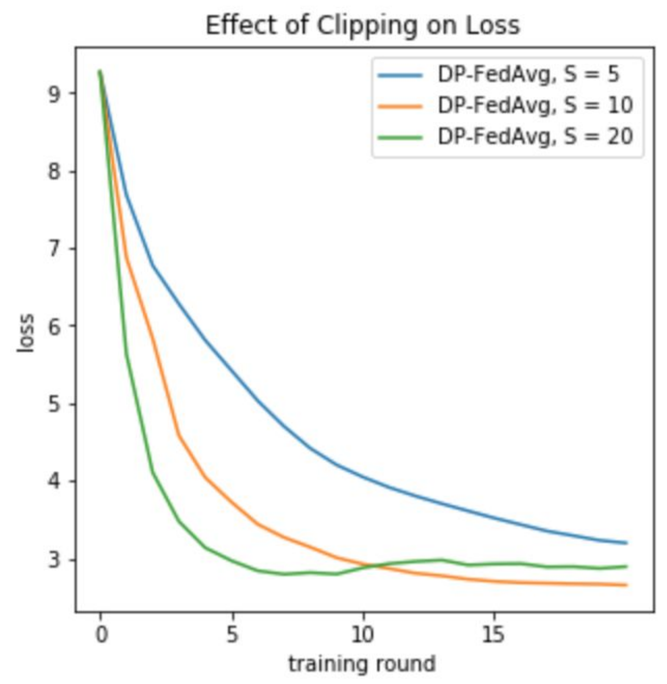


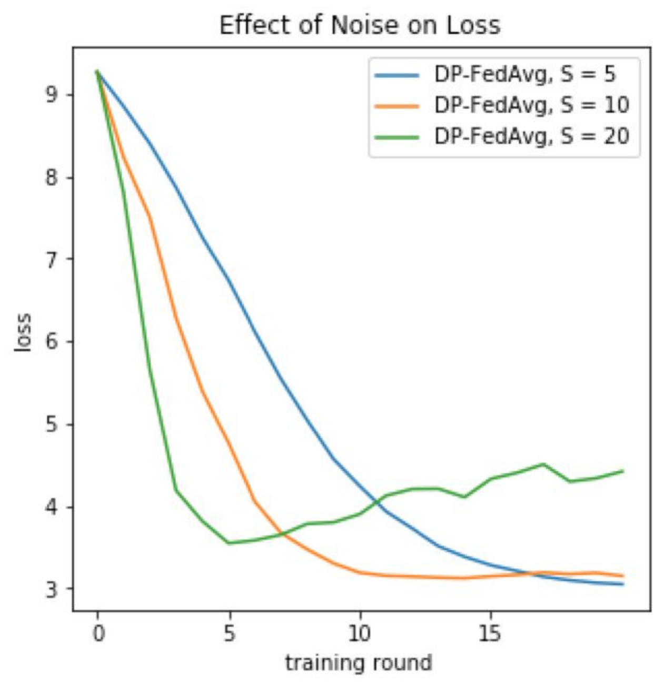


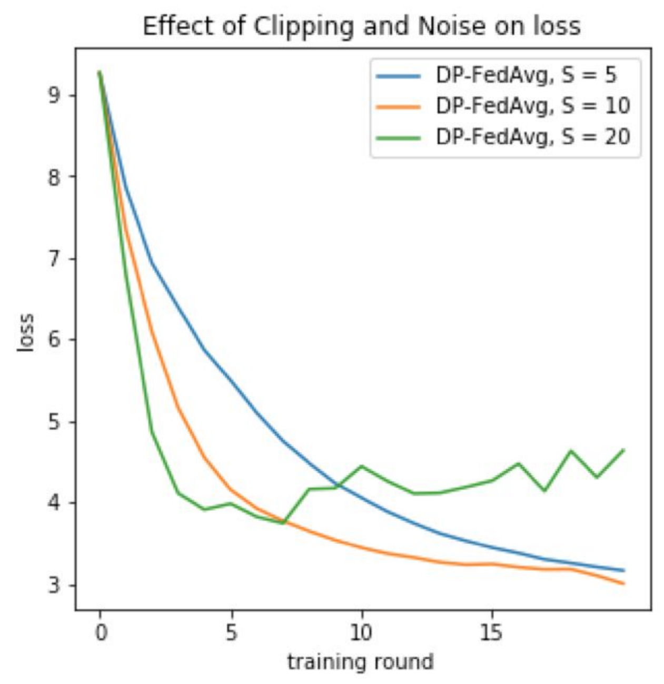
Best private model used S = 10, perplexity 20.13 at round 20

Central model: perplexity 39.13 at round 8

FedAvg model: perplexity 17.69 at round 20

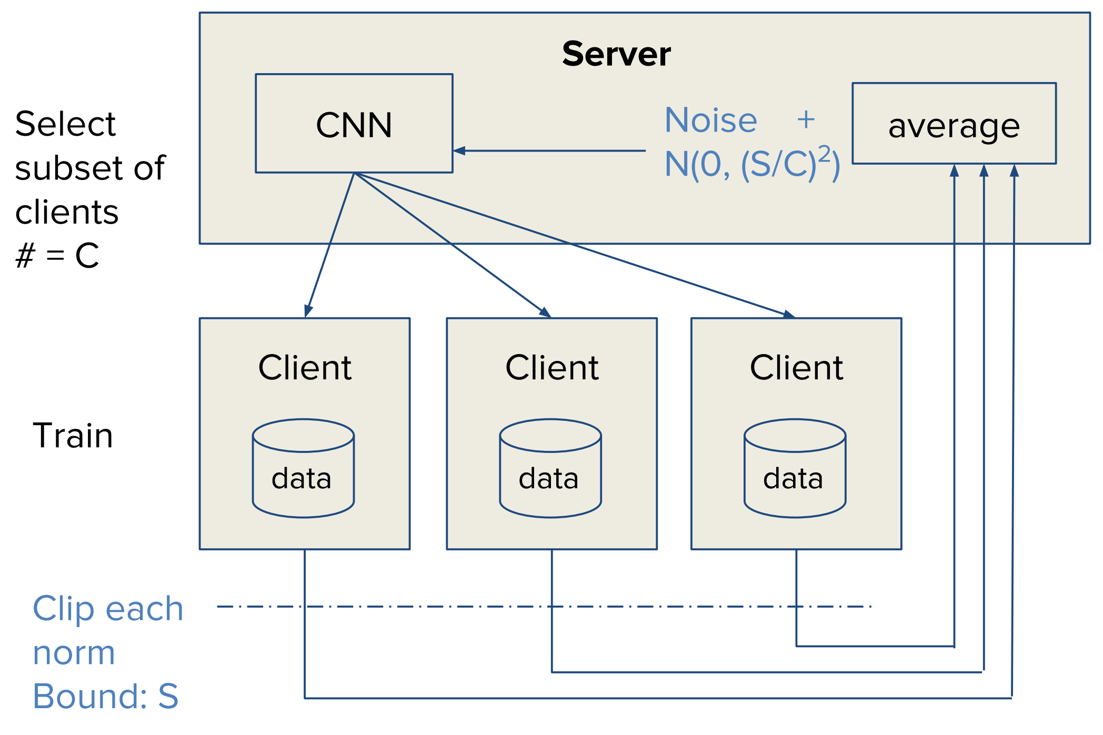




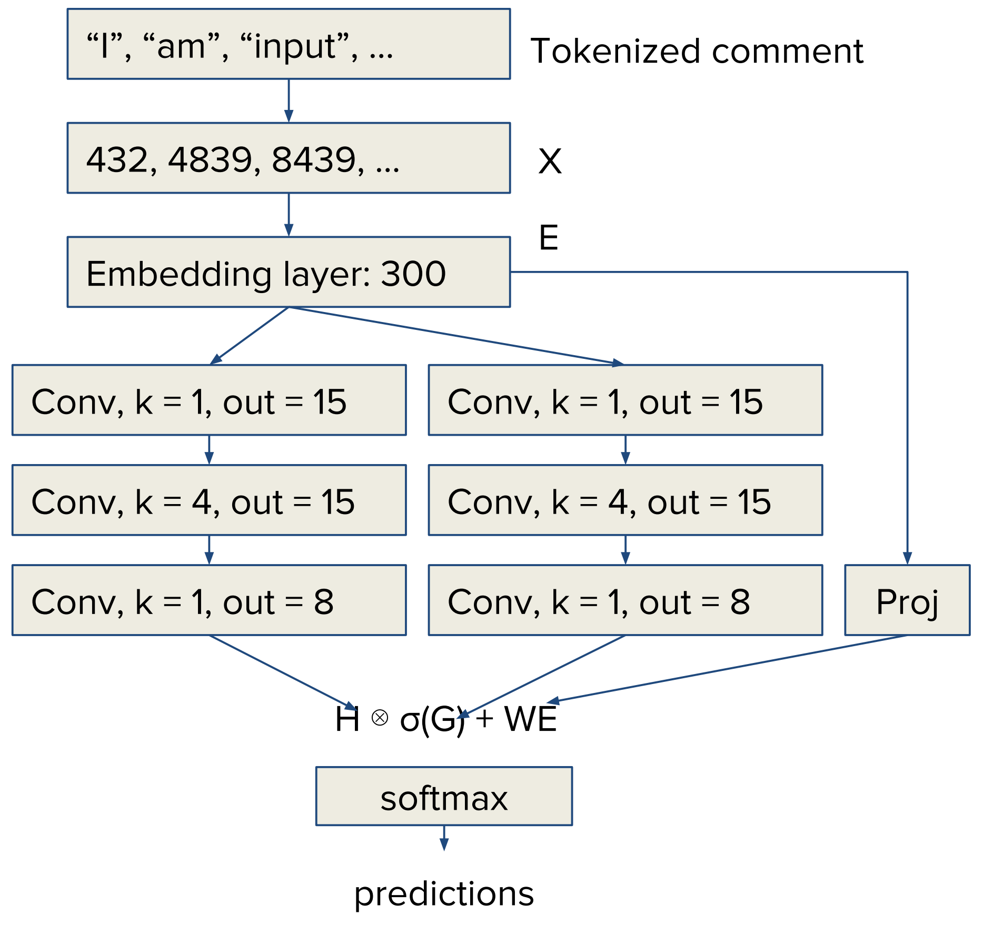


Architecture

Training Framework



CNN Language Model



Analysis

Models with privacy modifications achieve better performance than the central model

Client-side clipping and server-side noise act as regularization, improving performance

Performance on accuracy tracks with perplexity

The private model with S = 10 performs best

The models were trained with a privacy budget of $\varepsilon = 1.7$, $\delta = 10^{-5}$

More noise hurts performance more than stronger clipping

Conclusion

Training a CNN language model with federated training and differential privacy guarantees is possible and achieves good performance

Convergence of the private models is slower but may achieve better performance

Models can be designed with the regularization effect of the privacy mechanisms in mind

Researchers should be mindful of user privacy and consider experimenting with federated training and differentially private models

References

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=BJ0hF1Z0b.

Reddit.Reddit comments dataset.https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments, 2019.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks.CoRR, abs/1612.08083, 2016. URL http://arxiv.org/abs/1612.08083.H.