Privacy Regularization: Joint Privacy-Utility Optimization in Text-Generation Models

Fatemehsadat Mireshghallah[*], Huseyin Inan[+], Marcello Hasegawa[✦], Victor Rühle[✦], Taylor Berg-Kirkpatrick[*], Robert Sim[+]

[*]UC San Diego [+]Microsoft Research [✦]Microsoft

1. Problem

Neural language models are known to have a high capacity for memorization of training samples. This may have serious privacy implications when training models on user content such as email correspondence.

Unintended Memorization of Secrets

My credit card number is 4403 2212 8563 2345

2. Motivation

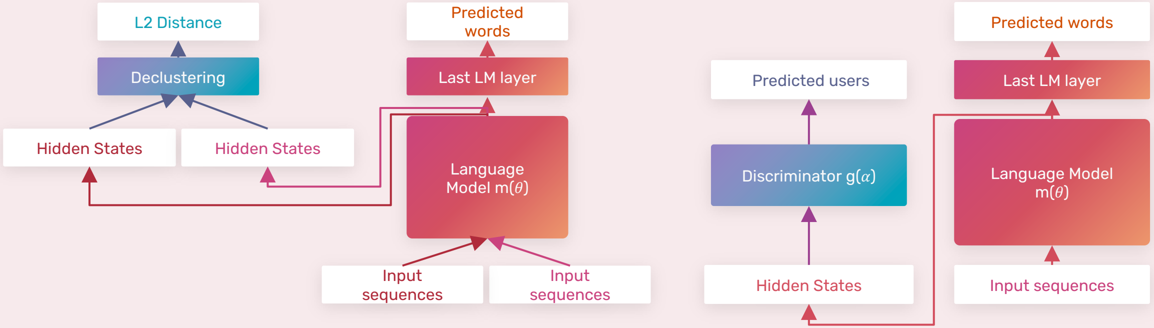
We show that differential privacy can have shortcomings in addressing this problem, for the reasons below:

DP is not context-sensitive: Cannot explicitely define protected attribute and wire it in the loss

DP is not suitable for correlated/repeated data

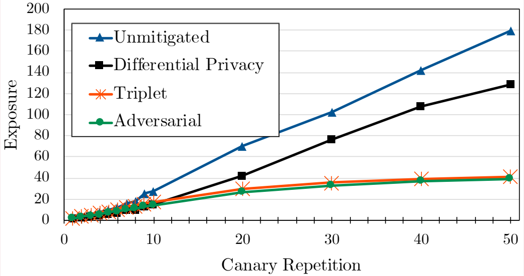
DP has disparate impact

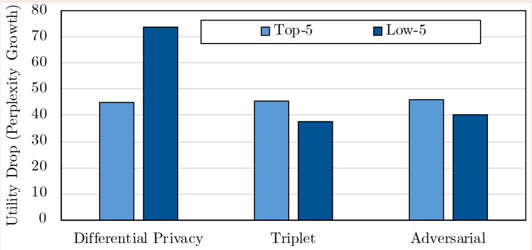
DP training is 10-15X slower, and much more cumbersome to tune



3. Proposed solution

We propose two privacy regularization methods, based on adversarial training and a novel privacy loss term, to jointly optimize for privacy and utility of recurrent language models. The main idea of our regularizers is to prevent the last hidden state representation of the language model for an input sequence from being linked back to the sensitive attribute we are trying to protect.





4. Results

Our results show that our regularization can be as affective as differential privacy, and more effective in some special cases. We also show that our regularizers do not have the disparate impacts of differential privacy, on utility.