Recovering Private Text in Federated Learning of Language Models

Background

- Federated learning
  - Clients collaboratively train a model through transmitting and aggregating model gradients and parameters
  - Has been suggested as a way to keep each client's data private during training

- Reconstructing Data From Gradients
  - Alarmingly, prior work shows that an attacker can recover high-fidelity private images from data transmitted during learning of image classification models
  - Realistic scenarios must consider recovery from large batch sizes; Prior approaches recovering text data from language models are only successful for unrealistically small batch sizes

This Work

- We study the recovery of text data from federated learning of large language models
  - We present a novel method called FILM (Federated Inversion Attack for Language Models) which recovers private text data during federated learning
  - We demonstrate that FILM can recover private training data from gradients of large batches
  - We evaluate potential defenses against our attack, and consider their associated utility-privacy tradeoffs

Method

1. Bag-of-Words Extraction (Melis et al. 19)

   - The cat chases the mouse. Let sleeping dogs lie.

2. Beam Search for Reconstruction

3. Token Reordering

Results

- FILM recovers significant text data from large batch sizes
- Real-world data is more susceptible to attacks

Defenses

- We consider the strength of defenses in preventing Bag-Of-Words extraction (step 1 of FILM)
- An ideal defense against FILM would provide protection (have low precision and recall), and minimize the tradeoffs in model utility (have low perplexity)

(a) Gradient pruning (Zhu et al., 2019)

<table>
<thead>
<tr>
<th>Prune ratio</th>
<th>Perplexity</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11.46</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.9</td>
<td>11.57</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.95</td>
<td>12.77</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.996</td>
<td>15.34</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>0.9999</td>
<td>19.21</td>
<td>1.00</td>
<td>0.90</td>
</tr>
</tbody>
</table>

(b) DPSGD (Abadi et al., 2016)

<table>
<thead>
<tr>
<th>c of DPSGD</th>
<th>Perplexity</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.31</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>14.32</td>
<td>0.29</td>
<td>0.01</td>
</tr>
<tr>
<td>10</td>
<td>12.86</td>
<td>0.88</td>
<td>0.17</td>
</tr>
<tr>
<td>15</td>
<td>11.98</td>
<td>0.97</td>
<td>0.49</td>
</tr>
<tr>
<td>inf</td>
<td>11.46</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 1: Performance of defenses in preventing bag-of-words recovery. (a) Gradient pruning sets values in the gradient to 0, according to the prune ratio. (b) Differentially Private Stochastic Gradient Descent (DPSGD) adds noise to gradients (with variance inversely proportional to c).

Defending with Frozen Embeddings

- Freezing word embeddings gradients during fine-tuning defends against FILM with minimal utility tradeoffs

References

