Recovering Private Text in Federated Learning of Language Models

Samyak Gupta*, Yangsibo Huang*, Zexuan Zhong, Tianyu Gao, Kai Li, Danqi Chen

PRINCETON UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

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** (**F**ederated **I**nversion Attack for **L**anguage **M**odels) 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**

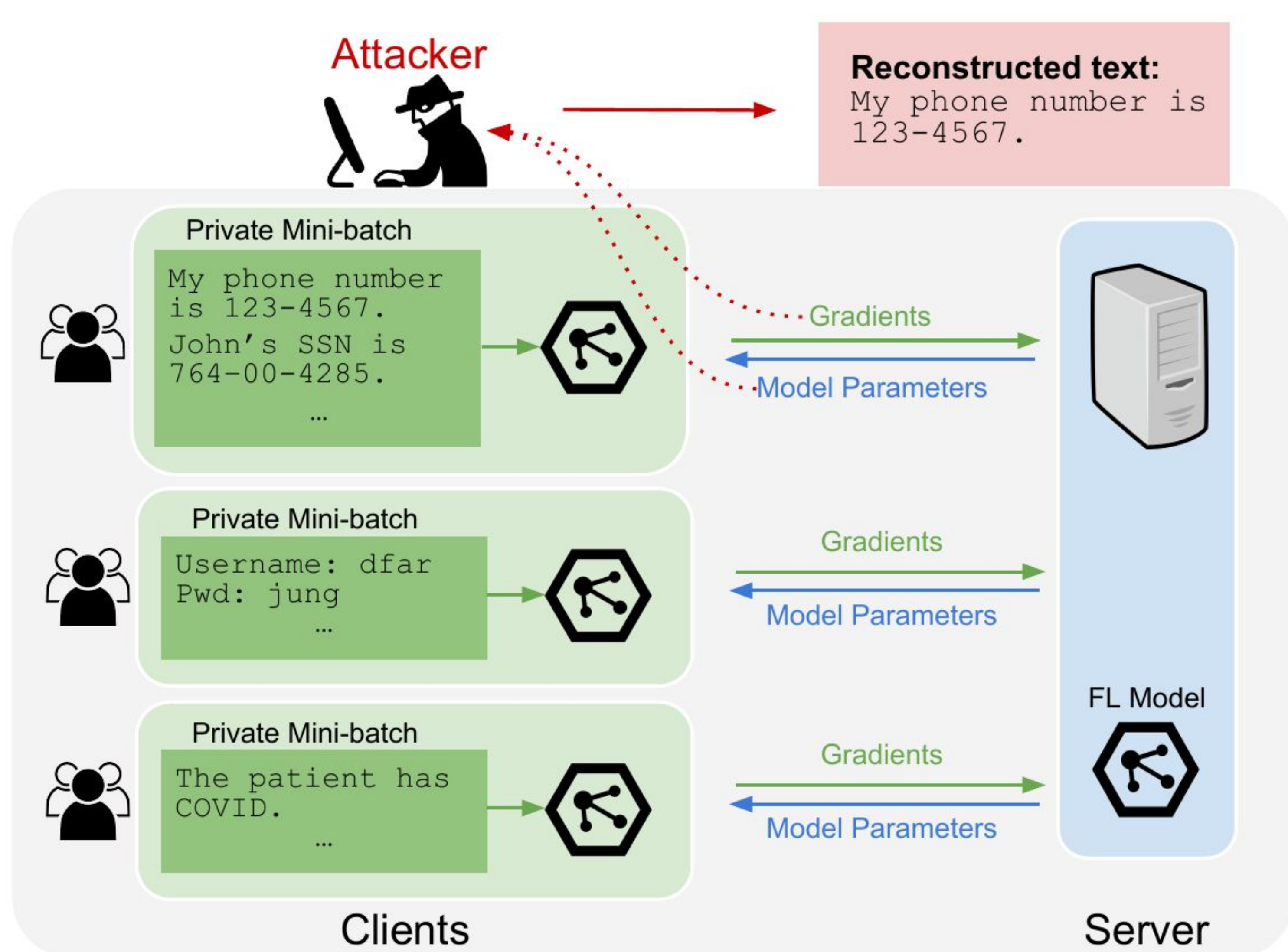
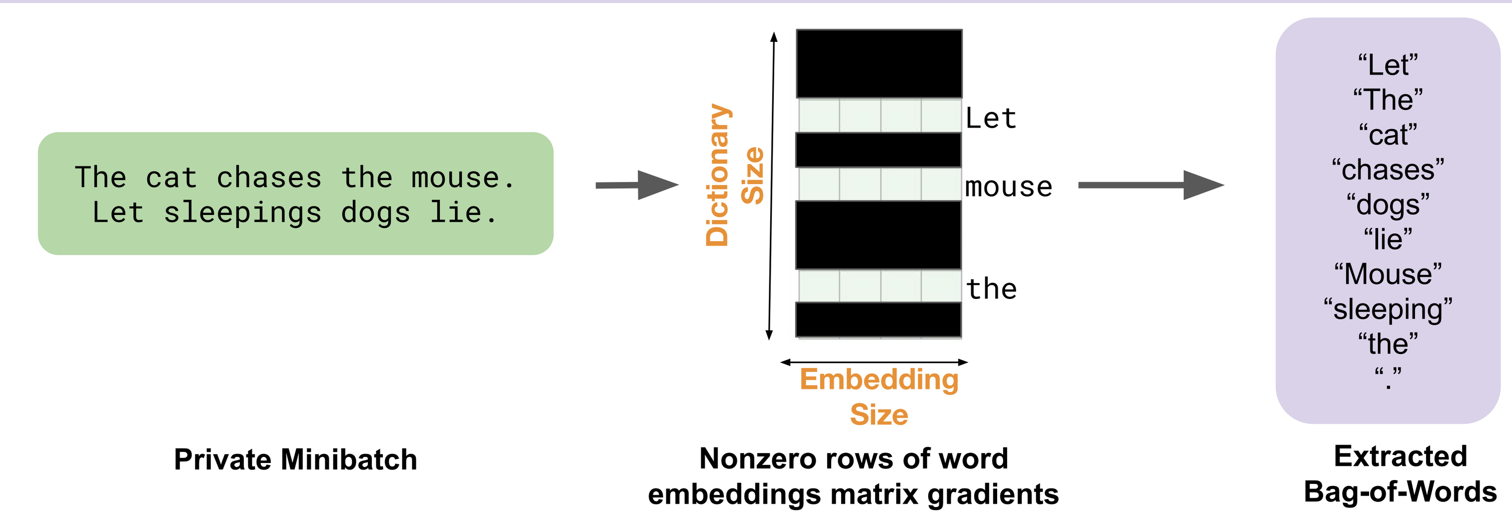


Figure 1: Outline of the setting. We consider a "Honest but Curious" attacker, who is able to passively observe transmitted gradients and parameters at each step of federated learning. The goal of the attacker is to recover sensitive data from a client's mini-batch of training data.

Method

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

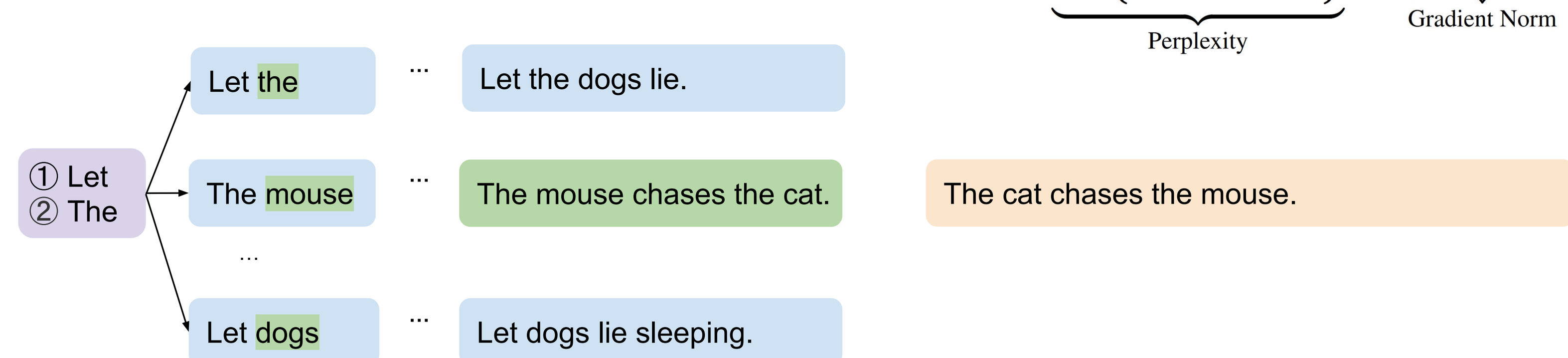


2. Beam Search for Reconstruction

$$Score(B_i) = \sum_{w \in B} logp(w) - \rho R_n(B_i)$$

3. Token Reordering

$$\mathcal{S}_\theta(\mathbf{x}) = \underbrace{\exp\left\{-\frac{1}{n}\log \mathbb{P}_\theta(\mathbf{x})\right\}}_{\text{Perplexity}} + \beta \underbrace{\|\nabla_\theta \mathcal{L}_\theta(\mathbf{x})\|}_{\text{Gradient Norm}}$$



Results

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

Attack & Batch Size b	Original Sentence	Best Recovered Sentence
WikiText-103		
FILM, $b = 1$	The short@-@tail stingray forages for food both during the day and at night.	The short@-@tail stingray forages for food both during the day and at night.
FILM, $b = 16$	A tropical wave organized into a distinct area of disturbed weather just south of the Mexican port of Manzanillo, Colima, on August 22 and gradually moved to the northwest.	Early on September 22, an area of disturbed weather organized into a tropical wave, which moved to the northwest of the area, and then moved into the north and south@-@to the northeast.
FILM, $b = 128$	A remastered version of the game will be released on PlayStation 4, Xbox One and PC alongside Call of Duty: Infinite Warfare on November 4, 2016.	At the time of writing, the game has been released on PlayStation 4, Xbox One, PlayStation 3, and PC, with the PC version being released in North America on November 18th, 2014.
Enron Email		
FILM, $b = 1$	Volume mgmt is trying to clear up these issues.	Volume mgmt is trying to clear up these issues.
FILM, $b = 16$	Yesterday, enron ousted chief financial officer andrew fastow amid a securities and exchange commission inquiry into partnerships he ran that cost the largest energy trader $35 million.	Yesterday, enron ousted its chief financial officer, andrew fastow, amid a securities and exchange commission inquiry into partnerships he ran that cost the company $35 million in stock and other financial assets.
FILM, $b = 128$	Yesterday, enron ousted chief financial officer andrew fastow amid a securities and exchange commission inquiry into partnerships he ran that cost the largest energy trader $35 million.	Yesterday, enron ousted chief financial officer andrew fastow amid a securities and exchange commission inquiry into partnerships he ran that he said cost the company more than $1 billion in stock and other assets.

Table 1: Best reconstructions using FILM under different datasets and batch sizes. Text highlighted in green represents successfully recovered phrases and words.

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)

Prune ratio	Perplexity	Precision	Recall
0	11.46	1.00	1.00
0.9	11.57	1.00	1.00
0.99	12.77	1.00	1.00
0.999	15.34	1.00	0.98
0.9999	19.21	1.00	0.90

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

ϵ of DPSGD	Perplexity	Precision	Recall
1	16.31	0.00	0.00
5	14.32	0.29	0.01
10	12.86	0.88	0.17
15	11.98	0.97	0.49
inf.	11.46	1.00	1.00

Table 2: 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 ε).

Defending with Frozen Embeddings

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

	From Scratch		From Pretrained	
	Unfrozen	Frozen	Unfrozen	Frozen
Wikitext-103	27.31	118.69	11.40	11.48
Enron Email	15.16	69.17	7.09	7.30

Table 3: Perplexity of models under different settings, when freezing (i.e. withholding transmission of) word embedding gradients during learning. Recall and precision of bag-of-words extraction are both 0 under this defense

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In ACM SIGSAC Conference on Computer and Communications Security (CCS), 2016.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage. in collaborative learning. In 2019 IEEE Symposium on Security and Privacy (SP), pp. 691–706. IEEE, 2019.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

Paper