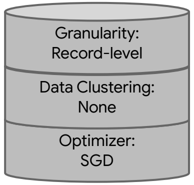
Understanding Unintended Memorization in Federated Learning

Om Thakkar (Google), Swaroop Ramaswamy (Google),
Rajiv Mathews (Google), Françoise Beaufays (Google)

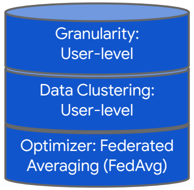
Motivation

- Prior work [CLK⁺'18, SS'19] has shown unintended memorization in generative models trained via *Central Learning*
- Federated Learning (FL) differs in many aspects from Central Learning



Central Learning

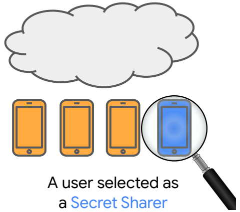
How do different components in FL affect such memorization?



Federated Learning

The Federated Secret Sharer

- Datasets in FL are inherently partitioned according to users
- We introduce the **Federated Secret Sharer** by adapting the Secret Sharer framework [CLK⁺'18] to the FL setting
- Each *secret* denoted by two parameters
  - **p_u**: Pr (a user being selected as a secret sharer)
  - **p_e**: Pr (a secret sharer's example being replaced by the secret)



A user selected as
a Secret Sharer

How are you doing?
Went for a movie last night
I feel like having pizza right now
My SSN is 123-45-6789
Hope to meet you soon

An example replaced
by the secret

Experimental Setup

- Use StackOverflow corpus (≈93M sentences, ≈392K users)
- Secret: 5 words chosen uniformly at random from ~10k vocab
- Insert 90 secrets: 10 secrets for each (p_u, p_e) config
- Train for 10 epochs
- Measure memorization on trained model
  - Random Sampling: **Least log-perplexity** in 2M random phrases ⟶ Memorized
  - Beam Search: **Most likely completion** using beam width <= 5 ⟶ Memorized

p_u		p_e
1 per 5K	×	100%
3 per 50K		10%
1 per 50K		1%

Configurations of secrets
inserted in training dataset

Results for Unintended Memorization

- For each setting, we report number of secrets (/90) memorized via Random Sampling and Beam Search
- Utility for all evaluated models is similar: accuracy varies from 23.7%-24.6%, perplexity from 57.3-64.3

Data: Randomly shuffled

SGD

Batch Size	Random Sampling	Beam Search
32 records	54	42
64 records	54	42
128 records	52	45
256 records	53	43

Central Learning

FedAvg

Batch Size	Random Sampling	Beam Search
500 users	66	56
1K users	69	58
2K users	67	56
5K users	65	58

IID Users in FL

Data: Clustered by users

SGD

Batch Size	Random Sampling	Beam Search
32 records	37	19
64 records	49	36
128 records	48	34
256 records	51	39

Non-IIDness in Central Learning

FedAvg

Batch Size	Random Sampling	Beam Search
500 users	21	0
1K users	23	1
2K users	19	1
5K users	26	2

Federated Learning (FL)

Optimizer	Random Sampling	Beam Search	Accuracy	Perplexity
FedAvg	26	2	24.5%	58.2
DP-FedAvg	12	0	23.3%	68.5

Results with Differentially Private (DP) FedAvg for Batch size: 5K users, Data: Clustered by users

Conclusions

- Clustering data according to users significantly reduces unintended memorization
  - Such clustering happens **by design** in distributed learning settings like Federated Learning
- Given data clustered by users, replacing optimizer from SGD to FedAvg causes a further reduction
- Training in FL with Differential Privacy (DP-FedAvg) can provide comparable utility while being resilient to memorizing secrets with 1000s of insertions spread across over 100 users