Squeezing More Utility via Adaptive Clipping on Differentially Private Gradients in Federated Meta-Learning

Authors: Ning Wang', Yang Xiao†, Yimin Chen‡, Ning Zhang*, Wenjing Lou’, and Y. Thomas Hou’

Virginia Tech  ‘
University of Kentucky  †
University of Massachusetts Lowell  ‡
Washington University in St. Louis  *
Our data is used by AI applications!

- Next word suggestion of Gboard.
Our data is used by AI applications!

Gboard only uses federated learning while your phone charges, is connected to Wi-Fi, and isn't in use. Learn how federated learning works.

https://support.google.com/android/answer/8608859?q=federated+learning

federated learning
Federated learning (also known as collaborative learning) is a machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchanging them.

https://en.wikipedia.org/wiki/Federated_learning

Federated learning - Wikipedia
What’s Federated Learning?
Is Data Privacy fully Protected by FL?

- The model parameters are open to the server directly and to other clients indirectly.
- Can attacker infer data from model?

Q: How likely is this picture to be `John Smith`?

A: Confidence=0.001

A: Confidence=0.9

A: Confidence=0.99

Training Data is Memorized by the Model
Privacy Attack: Inference Attack

- FL cannot guarantee the training data privacy.
- State-of-the-art Inference Attack
  - Membership inference attack [1,2]
  - Model Inversion Attack [3]
  - Attribute Inference Attack [4]

Problem and Goal of this Paper

● Problem
  - FL can protect data privacy to some extend.
  - Attackers are still capable to infer training data while knowing the model parameters.
  - Differential Privacy (DP) is a tool for privacy protection, but it harms the accuracy a lot.
  - Mobile used data size can be small.

● Goal
  - Provide rigorous privacy guarantee for users by incorporate DP.
  - Maintain a good trade-off between privacy and accuracy.
What is the inference attack in a database?
- Use the statistical/aggregate queries that are authorized to gain information that are not authorized.

Example: Exam score database.
- Tuple: (student_id, score)
- Average score on an exam is a query everyone is allowed to run.
- Attacker wants to find the exact score of some student.

Inference attack sometimes requires some additional external information. e.g., Attacker knows Alice took the exam late.

Attacker get
average score before some date & average score after such date.

It is easy to get the score of Alice.
DP preliminary: Adjacent Databases

- What’s Adjacent database?
  - Two databases only have one record difference.
  - E.g., a database with Alice data in, a database without Alice Data.
- Once attackers have access to the adjacent database, it can launch inference attack.

How to solve the inference attack?
Differential Privacy (DP)

- DP allows to learn useful information of the population without leaking individual privacy.

**$(\varepsilon, \delta)$-differential privacy**: A random mechanism $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ satisfies $(\varepsilon, \delta)$-differential privacy if for any two adjacent inputs $(d, d' \in \mathcal{D})$ and for any subset of outputs $S \in \mathcal{R}$ it holds that: $\Pr[\mathcal{M}(d) = S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') = S] + \delta$
Key Step of DP

- Adding Randomness
  - Gaussian Noise
  - Laplace Noise
Differentially private federated Learning

- Problem: low accuracy

Continual Training does help.

But it requires Enough Data & Training Power/Time

Accuracy dramatically decreases!
To cope with small size of local data

- Federated Learning → **Federated Meta-Learning**
  - Deal with few-shot problem.
  - Fast adaptation/customization.

Learn common knowledge from various tasks to enable quick learning for new/unseen tasks.
Two Threat Models

- Central server is trusted, clients are honest-but-curious.
- Both central server and clients are honest-but-curious.
Adding noise
- The noise should be proportional to the largest gradient.
- To avoid too large noise, we should clip the gradient.
Our Proposal

- **Adaptive Clipping**
  - Naïve constant clipping maintain a fixed clipping threshold $C$. The noise will be: $k \times C$.
  - Adaptive clipping: change the threshold $C$ adaptively.
  - Why adaptive clipping better?

*The gradients will decrease during the course of training.*

*We can change the threshold $C$ according to the gradient change.*
Adaptive Clipping

- We cannot use the true gradient scale to adaptively decide $C$.
- Our proposal: determine $C$ by using historical DP gradients in a window of size $W$:

$$C_{t+1} = f([\tilde{g}_1, \ldots, \tilde{g}_W], k)$$

---

**Algorithm 2: Adaptive Clipping**

**Input:** Clients set $\mathcal{T}$, noise multiplier $z$, user sample size $L$,
Learning round $T$, Window size $W$

**Output:** Clipping Threshold $C$

1. Initialize $C = C_0$;
2. while and $t \leq T$ do
3. Randomly Sample $L$ clients $\mathcal{T}_s \leftarrow$ sample($\mathcal{T}$, $L$);
4. if $t > W$ then
5. $C \leftarrow f([\tilde{g}_{t-W}, \ldots, \tilde{g}_{t-1}], k)$
6. end if
7. for $i \in \mathcal{T}_s$ do
8. $g_i \leftarrow$ gradient provided by client $i$;
9. Clip gradient: $\hat{g}_i \leftarrow g_i \ast \min(1, \frac{C}{\|g_i\|})$;
10. $\tilde{g}_t \leftarrow \frac{1}{L} \left( \sum_i \hat{g}_i + N(0, z^2C^2I) \right)$
11. end for
Will Adaptive Clipping leak any more privacy?

- NO
- Because of the **Post processing Property of DP**
  - If $F(X)$ satisfies $\epsilon$-differential privacy
  - Then for any (deterministic or randomized) function $g$, $g(F(X))$ satisfies $\epsilon$-differential privacy

### Algorithm 2: Adaptive Clipping

**Input:** Clients set $\mathcal{T}$, noise multiplier $z$, user sample size $L$,
Learning round $T$, Window size $W$

**Output:** Clipping Threshold $C$

1. Initialize $C = C_0$
2. **while and** $t \leq T$ **do**
3. Randomly Sample $L$ clients $\mathcal{T}_s \leftarrow \text{sample}(\mathcal{T}, L)$;
4. **if** $t > W$ **then**
   5. $C \leftarrow f([\tilde{g}_{t-W}, ... , \tilde{g}_{t-1}], k)$
6. **for** $i \in \mathcal{T}_s$ **do**
7. $g_i \leftarrow$ gradient provided by client $i$;
8. Clip gradient: $\hat{g}_i \leftarrow g_i \ast \min(1, \frac{C}{\|g_i\|})$
9. $\tilde{g}_t \leftarrow \frac{1}{L} \left( \Sigma_i \hat{g}_i + \mathcal{N}(0, z^2 C^2 I) \right)$

The DP version Gradients
Differentially private Meta-learning

- The history of Differentially Private version gradients guides the current clipping.

![Diagram](image-url)
Two Algorithms

- Two threat models
  - DP-AGR for threat model 1 where server is trusted, clients are honest-but-curious
  - DP-AGRLR for threat model 2 where the server is not trusted, and clients are honest-but-curious

```
Algorithm 3: DP-AGRLR (Client Side)

Input: Current global model $\Theta$, local data $D$, DP parameter $(\epsilon_0, \delta_0), C_0, z_0$
Output: gradient $g$

1. Function $g = \text{Base-Model-Train}(\Theta, D^s, D^q)$:
2. Initialize base-model: $\theta \leftarrow \Theta$;
3. Split local data $D^s, D^q \leftarrow D$;
4. $z_0 \leftarrow \text{compute\_noise}(\epsilon_0, \delta_0, *\text{args})$
5. for $(x_i, y_i) \in D^s$ do
6.   record level gradient: $g_i \leftarrow \nabla_{\theta} \mathcal{L}(\theta, x_i)$;
7.   clip gradient: $\hat{g}_i \leftarrow g_i \ast \min(1, \frac{C_0}{\|g_i\|})$;
8. $\hat{g} \leftarrow \frac{1}{|D^s|} (\sum_i \hat{g}_i + N(0, (z_0C_0)^2I))$;  
9. update base-model: $\theta \leftarrow \theta - \eta_1 \hat{g}$;
10. for $(x_i, y_i) \in D^q$ do
11.   record level gradient: $g_i \leftarrow \nabla_{\theta} \mathcal{L}(\theta, x_i)$;
12.   clip gradient: $\hat{g}_i \leftarrow g_i \ast \min(1, \frac{C_0}{\|g_i\|})$;
13. $g \leftarrow \frac{1}{|D^q|} (\sum_i \hat{g}_i + N(0, (z_0C_0)^2I))$.
```
Experimental Setting

- **Settings:**
  - Image Datasets: Omniglot, CIFAR-FS, Mini-ImageNet
  - Client Number: 400,000
  - Clients in each learning round: 1500
  - Each client has 30 data record.
  - Meta-learning algorithm: MAML.

- **Code:**
  - Code Evaluated
Evaluation: Adaptive Clipping

- **Ours vs AQC vs Constant**

All other settings are the same, only change the clipping method.
Evaluations: DP-AGRLR

- More accurate ML model with much lower privacy budget
  - DP-AGR (ours) achieves \((1.5, 10^{-6})\)-DP;
  - DP-AGRLR (ours) achieves \((2.5, 10^{-5})\)-DP for record-level privacy
  - Baseline achieves \((9.5, 10^{-3})\)-DP

Model Accuracy of 5-way 5-shot learning in the Omniglot dataset.
Summary

- Differentially private federated meta-learning architecture.
- Design an adaptive gradient clipping method to conserve the privacy budget and improve accuracy.
- Provide two algorithms, DP-AGR and DP-AGRLR, to deal with different privacy requirements.
Thank You!

Q&A