Analisi di attacchi alla privacy nel Federated Learning

Laurea Magistrale in Sicurezza Informatica

Luca Vaudano (90028)

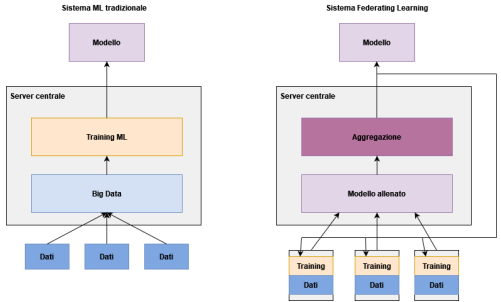
12/10/2023

UNIVERSITÀ DEGLI STUDI DI MILANO

Indice

Che cosa è il Federated Learning?

- Una tecnica di apprendimento collaborativo che sfrutta modelli condivisi senza archiviare centralmente i dati. Utilizza client coordinati da un server centrale. [1]



[1] McMahan et al., *Communication-Efficient Learning of Deep Networks from Decentralized Data*, 2017.

Come funziona il Federated Learning?

1. **Selezione del client:** Scelta dei dispositivi client.
2. **Broadcast:** Il server trasmette il modello iniziale ai client selezionati.
3. **Calcolo sul client:** Ogni client allena localmente il modello con i propri dati.
4. **Aggregazione:** Client inviano gli aggiornamenti al server che li combina.
5. **Aggiornamento del modello:** Nuova versione del modello inviata ai client. Il processo si ripete fino alla convergenza o alla precisione desiderata.

Indice

2 Attacchi nel Machine Learning

Attacchi nel Machine Learning

- **Adversarial example:** *manipolare in modo mirato i dati di input in modo da ingannare un modello di machine learning*.
- **Poisoning Attack:** *degradare le performance del sistema*.
- **Attacchi di privacy**
  - **Data reconstruction:** *ottenere approssimazioni degli input originali*.

Indice

3 Privacy-Preserving nel Federated Learning

Privacy-Preserving nel Federated Learning

- La privacy non può essere considerata come proprietà binaria.
- Il FL è un meccanismo di privacy sufficiente?

Indice

4 Attacchi basati su inversione del gradiente

Intuizione

4 Attacchi basati su inversione del gradiente

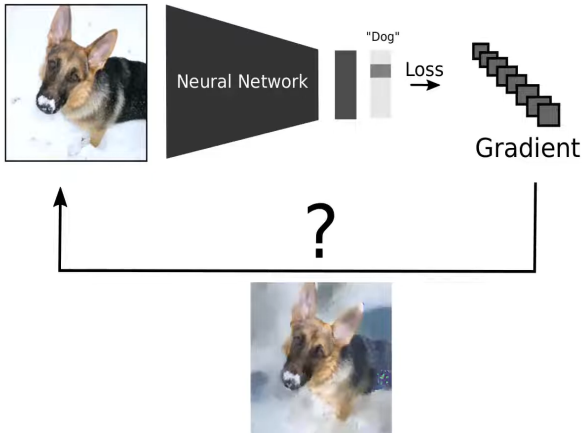


Figura: Intuizione di inversione del gradiente.

Threat Model

4 Attacchi basati su inversione del gradiente

- Server *honest-but-curious*
- Allena in maniera onesta
- Immagazzina e processa le informazioni dell'user separatamente

Deep Leakage dai gradienti

4 Attacchi basati su inversione del gradiente

1. Inizializzazione di input e label fantocci
2. Calcolo del gradiente fantoccio
3. Calcolo della *differenza euclidea* tra gradienti fantoccio e gradienti reali
4. Aggiornamento dei dati per eguagliare i gradienti

$$x^*, y^* = \arg\min_{x', y'} \|\nabla W' - \nabla W\|^2 = \arg\min_{x', y'} \left\| \frac{\partial \mathcal{L}(F(x', W), y')}{\partial W} - \nabla W \right\|^2$$

[1] Zhu et al., *Deep leakage from gradients*, 2019.

Inverting gradients

4 Attacchi basati su inversione del gradiente

1. Inizializzazione di input e label fittizie
2. Calcolo di gradienti fittizi
3. Calcolo della *cosine similarity*
4. Aggiornamento dei dati fittizi per corrispondere ai gradienti

$$\arg\min_{x} 1 - \cos(\nabla_\theta \mathcal{L}_\theta(x, y), g_\theta)$$

[1] Geiping et al., *Inverting gradients – how easy is it to break privacy in federated learning?*, 2020.

Risultati DLG

4 Attacchi basati su inversione del gradiente

(a) Immagine ground truth.

(b) Immagine ricostruita.

Figura: Ricostruzione immagine dai gradienti con algoritmo DLG.

Risultati IG

4 Attacchi basati su inversione del gradiente

(a) Immagine ground truth.

(b) Immagine ricostruita.

Figura: Ricostruzione immagine dai gradienti con algoritmo IG.

Valutazioni

4 Attacchi basati su inversione del gradiente

- **Power imbalance:** *il server ha il controllo centralizzato su questo protocollo*.
- **Nuovi threat model:** *un server potrebbe essere compromesso e diventare malevolo*.
- **Decentralizzazione del FL:** *fare scegliere ai singoli utenti di partecipare ai round di learning, riducendo il controllo del server.*

Indice

5 Conclusioni

Conclusioni

5 Conclusioni

- Due risultati empirici di *privacy leaks*.
- Il Federated Learning da solo non può essere considerato una **PETs**.
- Soluzioni? **Trade-off tra privacy e utility**.

Sviluppi futuri

- Il Federated Learning nel NLP.
- Vulnerabilità nell'architettura **Transformer**? [2]

[2] Fowl et al., *Decepticons: Corrupted transformers breach privacy in federated learning for language models*, 2023.

Grazie per l'ascolto!