Towards Preserving Server-Side Privacy of On-Device Models

Akanksha Atrey¹, Ritwik Sinha², Somdeb Sarkhel², Saayan Mitra², David Arbour², Akash V. Maharaj³, Prashant Shenoy¹

¹University of Massachusetts Amherst, ²Adobe Research, ³Adobe Inc.

Introduction

On-device machine learning models are becoming ubiquitous in distributed service applications due to their advantages.

- Low latency: No round trip to the centralized server
- User data privacy: Processing happens on the device
- Works offline: No network connectivity required
- Reduces cost: No cost of servers or cloud compute cycles

However they can leak sensitive information about the service provider!

Example: Bank Application

Contextual information

Proprietary information (PI)

Personalized financial incentives

“Which set of inputs give me the most profitable output?”

Most applications in privacy preserving ML have focused on user privacy. Our work deals with evaluating the extent to which a service provider’s intellectual property can be exploited and misused in on-device models.

Contributions

In this work, we establish the importance of server-side privacy in on-device service models with the following contributions:

C1. Developed a taxonomy of on-device ML models focusing on distributed services.
C2. Proposed multiple privacy attacks on on-device models and evaluated their efficacy on a real-world dataset.
C3. Designed preliminary ideas on how to protect the service provider’s proprietary information embedded in on-device models.

Privacy Attacks on On-Device Models

We focus on server-side privacy attacks on on-device models which aim to recover the representations learned by model $M$.

In this work, we focus on one-vs-all multi-classification models.

Model Inversion Attack (white-box)

Input

Model $n$

Output

reconstruct

Known: model, model input, model output

Goal: Exploit each binary classification model and reconstruct the input for each class using backpropagation

Random Querying Attack (white-box)

Input

Model $n$

Output

Known: ensemble model, model input, number of output classes

Goal: Iteratively query the features used by the model to recover the classes

Large-Scale Querying Attack (black-box)

Input

Processing

Output

Known: ensemble model, all features collected by the service provider

Goal: Iteratively query the whole feature set at scale to recover the classes

Evaluation of Server-Side Privacy Leakage

Difference between traditional serialized models and ONNX models (on-device) for random forests (RF) and deep neural networks (DNN):

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Feature Space</th>
<th>Attack</th>
<th>Runtime (s)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>All</td>
<td>Model Inversion</td>
<td>0.0531</td>
<td>99.24</td>
</tr>
<tr>
<td>ONNX</td>
<td>All</td>
<td>Random Querying</td>
<td>1.5523</td>
<td>98.84</td>
</tr>
<tr>
<td>DNN</td>
<td>All</td>
<td>Large Scale Querying</td>
<td>55.4894</td>
<td>98.84</td>
</tr>
</tbody>
</table>

Runtime of privacy attacks to recover full set of input to output mappings.

Results of evaluating the white-box (WR) and black-box (BB) querying attacks on RFs and DNNs:

(a) impact of varying query sizes on the attack efficacy;
(b) impact of varying the number of features being queried in the white-box attack;
and (c) impact of the number of unused features on runtime in the black-box attack.

Countermeasures for White-Box Attacks

Inference on Encrypted Models. Conduct inference on encrypted data to avoid giving the adversary access to too much information.

Distributing Service. Conduct inference on the device but keep the mapping of model output to tangible service on a central node. Each unique query will require connection to a central node.

Countermeasures for Black-Box Attacks

Increasing Dimensionality. Collect more data. Since the subset of features used by the model are unknown, the adversary will have to search over a higher dimensionality.

Query-based Model Degradation. Degrade the weights of the model as more queries are conducted. Eventually the model will output random noise.