CryptGPU: Fast Privacy-Preserving Machine Learning on the GPU

Privacy-Preserving ML

- Hospitals should not learn patient’s medical data
- Patient should not learn the weights of the ML model

Can be achieved with secure multiparty computation

Scalability Challenge in PPML

- There is a significant performance gap between plaintext and private ML (2300x in private inference, 42000x in private training)
- Linear layers are the major performance bottleneck
- GPU acceleration is necessary for scalability

Our System and Benchmarks

A system that supports end-to-end private training/inference on GPU
- Supports private inference/training in the 3PC semi-honest setting
- Keep all computations on the GPU
- Significantly improve performance of private inference/training

Embedding fixed-point arithmetic into floating-point CUDA kernels
\[(A_1 + A_2) \cdot (B_1 + B_2) = A_1B_1 + A_1B_2 + A_2B_1 + A_2B_2\]
- Convert product of 64-bit integers into sums of product of 16-bit integers
- Use CUDA kernels to compute product of 16-bit integers in floating-point

GPU friendly protocol design
- Component-wise operations (e.g., multiplication) are fast on GPUs
- Conditional statements are slow on GPUs
- Design protocols that better utilize parallelism

Replicated secret-sharing as basic building blocks
- A type of additive secret-sharing scheme
- Each party holds 2-out-of-3 secret shares
- Communication efficient in the 3PC setting

<table>
<thead>
<tr>
<th>Model</th>
<th>CRYPTFLO</th>
<th>CRYPTO GPU</th>
<th>Plaintext</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>25.9</td>
<td>6.9</td>
<td>25.9</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>9.31</td>
<td>3.08</td>
<td>17.62</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>60.0</td>
<td>14.5</td>
<td>60.0</td>
</tr>
</tbody>
</table>

A 2.5x improvement over CryptFLO on private inference

Future Work

- Support multiple GPUs
- Design more efficient MPC protocols that leverage GPU parallelism

Summary

- We present the first PPML system that keep all computations on the GPU
- We demonstrate that GPU can significantly accelerate bottleneck in linear layers
- Training AlexNet on TinyImageNet previously takes over a year, and now it takes roughly over a week (~10 days)

Threat Model

3PC semi-honest security with honest-majority
- Honest-majority: Allowing a single semi-honest party for corruption
- Semi-honest: Corrupt parties follow the protocol, but try to gather information out of the protocol