
Xirong Zhuang¹, Lan Zhang¹,², Chen Tang¹, Huiqi Liu¹, Bin Wang³, Yan Zheng³, Bo Ren³.
1 University of Science and Technology of China,
2 Institute of Dataspace, Hefei Comprehensive National Science Center,
3 Tencent YouTu Lab.
Introduction

• Well-trained DL models have become recognized as **valuable intellectual property (IP)** for significant upfront investment during the training process.

High-quality datasets  Experienced experts  Computing resources
Introduction

• To fully capitalize on the value, owners are often willing to **offer their models as services**, as long as they can safeguard their IP rights and receive the corresponding revenue.

High-quality datasets  Experienced experts  Computing resources
Introduction

Cloud

Machine learning as a service (MLaaS)

On-device

Trusted Execution Environment (TEE)

Cryptographic

Homomorphic encryption (HE)

Secure multi-party computation (MPC)
Introduction

• Active authorization
  ◦ Modifies models and granting correct usage to authorized users
  ◦ Prevents models from being stolen by unauthorized users
  ◦ Loses subsequent control over the authorized model
Introduction

• Active authorization
  ◦ Modifies models and granting correct usage to authorized users
  ◦ Prevents models from being stolen by unauthorized users
  ◦ Loses subsequent control over the authorized model

• Controllable authorization
  ◦ Model owners can grant and revoke the right to use their models
Design Goals

• Model confidentiality
  ◦ Original models cannot be exposed to authorized users
  ◦ Encrypted models cannot be restored effortlessly
Design Goals

• Model confidentiality
  ◦ Original models cannot be exposed to authorized users
  ◦ Encrypted models cannot be restored effortlessly

• Model controllability
  ◦ Conduct inference as agreed upon in the contract
  ◦ Terminate the authorization in case of any breach of the contract
Design Goals

• Model confidentiality
  ◦ Original models cannot be exposed to authorized users
  ◦ Encrypted models cannot be restored effortlessly

• Model controllability
  ◦ Conduct inference as agreed upon in the contract
  ◦ Terminate the authorization in case of any breach of the contract

• Minimal latency and resource consumption
  ◦ Satisfy the response-time requirements of real-life applications
  ◦ Applicable on resource-constrained devices
Challenge

• Difficulty in confidentiality and efficient execution of the deployed model
  ◦ Existing methods cannot resist cracking or fine-tuning attacks
  ◦ Cannot decrypt the entire model straightforwardly within TEE since the limited memory

<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>200 MB</td>
</tr>
<tr>
<td>ResNet152</td>
<td>230 MB</td>
</tr>
<tr>
<td>VGG16</td>
<td>500 MB</td>
</tr>
</tbody>
</table>
Challenge

• Uninterrupted and inescapable model controllability on remote devices
  ◦ Existing works cannot offer such controllability after the distribution of models
  ◦ The owner needs to maintain the connection with the decrypted model

Distribute → Decrypt → Abuse → Copy and redistribute
**System Overview**

- Generates deployment materials on the owner's side
  - Pre-signed contract $\rightarrow$ Encryption key & Encrypted model & Enclave code

- Performs controlled inference on the user's side
  - Enclave initialization $\rightarrow$ Inference for a specified period as per the contract
Confidentiality

• Encryption requirements of controllable authorization
  ◦ Compatible with the SGX-based DL inference
  ◦ No loss of inference accuracy after decryption
  ◦ Uncrackable with reasonable time and effort cost

Layer-wise model encryption with baker mapping
Confidentiality

• Encryption requirements of controllable authorization
  ◦ Compatible with the SGX-based DL inference
  ◦ No loss of inference accuracy after decryption
  ◦ Uncrackable with reasonable time and effort cost

Layer-wise model encryption with baker mapping
Controllability

• Contract-based code generation
  ◦ Ensure the remote device performs a series of intended operations
  ◦ Ensure the corresponding user codes are not tampered with
  ◦ Pre-generated and verifiable enclave codes
Controllability

- Contract-based code generation
  - Ensure the remote device performs a series of intended operations
  - Ensure the corresponding user codes are not tampered with
  - Pre-generated and verifiable enclave codes
Controllability

• Controlled model inference
  ◦ Dynamically load the needed encrypted weights
  ◦ Parallelly pipeline: integrity check / decryption/ inference
  ◦ Promptly upload the current usage status
Evaluation

1) How is the efficiency and security of model encryption?
   - Baseline 1: straightforward encryption method: Deep Lock
   - Baseline 2: mapping encryption method: Chaotic Weights

2) Can DeepContract run DNN within SGX’s memory limit?

3) How much is the overhead of DeepContract?
   - Baseline 1: in-enclave inference: Occlumency
   - Baseline 2: secure two-party computation using HE/MPC
1) How is the efficiency and security of model encryption?

**Decryption Speed**
- \(8.9\times\) faster than **DeepLock**
- \(2.4\times\) faster than **ChaoW**

<table>
<thead>
<tr>
<th>Scheme</th>
<th>VGG16</th>
<th>ResNet18</th>
<th>ResNet50</th>
<th>ResNet101</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLock [2]</td>
<td>23.53</td>
<td>17.60</td>
<td>37.52</td>
<td>69.41</td>
</tr>
<tr>
<td>ChaoW [30]</td>
<td>5.14</td>
<td>5.22</td>
<td>13.43</td>
<td>20.23</td>
</tr>
<tr>
<td>DeepContract</td>
<td>1.58</td>
<td>1.42</td>
<td>9.82</td>
<td>15.21</td>
</tr>
</tbody>
</table>
1) How is the efficiency and security of model encryption?

Resistance to fine-tuning attacks

Encrypted models cannot be restored to unacceptable accuracy even with staggering proportion (25%-30%) of the training dataset!
Evaluation

2) Can DeepContract run DNN within SGX’s memory limit?

The maximum memory usage is always under the available Enclave Page Cache memory size of SGXv1.
3) How much is the overhead of DeepContract?

Inference Speed

- **23%** slower than **Occlumency** (Not protecting model weights)
- **8%-13%** slower at more relaxed security levels
3) How much is the overhead of DeepContract?

Compared to cryptographic methods

- More **real-time** inference
- Only **minor data transfer** required

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Framework</th>
<th>MNIST Run Time (s)</th>
<th>MNIST Data Transfer (MB)</th>
<th>CIFAR-10 Run Time (s)</th>
<th>CIFAR-10 Data Transfer (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE</td>
<td>SHE [33]</td>
<td>9.3</td>
<td>123</td>
<td>2258</td>
<td>160</td>
</tr>
<tr>
<td>HE</td>
<td>LoLa [4]</td>
<td>2.2</td>
<td>18</td>
<td>730</td>
<td>370</td>
</tr>
<tr>
<td>MPC</td>
<td>EzPC [7]</td>
<td>5.1</td>
<td>501</td>
<td>265.6</td>
<td>40683</td>
</tr>
<tr>
<td>MPC</td>
<td>Chameleon [40]</td>
<td>2.24</td>
<td>11</td>
<td>52.67</td>
<td>2650</td>
</tr>
<tr>
<td>MPC</td>
<td>XONN [39]</td>
<td>0.15</td>
<td>32</td>
<td>5.79</td>
<td>2599</td>
</tr>
<tr>
<td>HE-MPC</td>
<td>nGraph-HF [3]</td>
<td>64.32</td>
<td>51</td>
<td>1824</td>
<td>3775</td>
</tr>
<tr>
<td>HE-MPC</td>
<td>Gazelle [23]</td>
<td>0.81</td>
<td>70</td>
<td>12.9</td>
<td>1236</td>
</tr>
<tr>
<td>TEE</td>
<td>DeepContract</td>
<td>0.13</td>
<td>0.0041</td>
<td>0.18</td>
<td>0.0045</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage</th>
<th>Data</th>
<th>Size</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deployment</td>
<td>Encrypted Model Enclave Codes</td>
<td>90.7 MB</td>
<td>Once in an authorization</td>
</tr>
<tr>
<td>Transmission</td>
<td>Authorization Key Hash Values</td>
<td>1.5 KB</td>
<td>Once in an authorization</td>
</tr>
<tr>
<td></td>
<td>Attestation Message Usage Status</td>
<td>3.1 KB</td>
<td>Once in a verification cycle</td>
</tr>
</tbody>
</table>
Thank You!


Xirong Zhuang\textsuperscript{1}, Lan Zhang\textsuperscript{1,2}, Chen Tang\textsuperscript{1}, Huiqi Liu\textsuperscript{1}, Bin Wang\textsuperscript{3}, Yan Zheng\textsuperscript{3}, Bo Ren\textsuperscript{3}.

1 University of Science and Technology of China,
2 Institute of Dataspace, Hefei Comprehensive National Science Center,
3 Tencent YouTu Lab.