Confidential Execution of Deep Learning Inference at the Untrusted Edge with ARM TrustZone

Md Shihabul Islam, Mahmoud Zamani, Chung Hwan Kim, Latifur Khan, Kevin W. Hamlen

ACM Conference on Data and Application Security and Privacy (CODASPY '23),
April 24–26, 2023, Charlotte, NC, USA
MOTIVATION

- Ubiquitousness of Internet-of-Things (IoT) devices
- On-device Machine Learning
  - Performance of edge/IoT applications
    - Low bandwidth
    - Reducing communication cost
  - Privacy of user data

Challenges

- Protection of **user data** on **untrusted** and **resource-constrained** edge/IoT devices
  - Model Inversion Attack
  - Membership Inference Attack

- Unfeasible existing techniques for edge/IoT devices
  - Homomorphic encryption
  - Differential privacy

**Solution:**

- Trusted Execution Environment (TEE) for edge/IoT devices
  - ARM TrustZone
**ARM TrustZone**

- **ARM**: Pioneer in embedded device processors

- **TrustZone**
  - Optional hardware security extension
  - Ensures the integrity and confidentiality of an application’s data on a device
  - Two architectures:
    - Cortex-A
    - Cortex-M

---

**Figure Source**: Demystifying Arm TrustZone: A Comprehensive Survey, ACM Computing Surveys, 2019
**Background**

**ARM TrustZone Limitations**

*Limitations:*
- Resource-intensive DL methods
- Limited trusted memory and resources in TrustZone

*Possible Solutions:*
- Quantization
- Model pruning

But affects model’s prediction accuracy
Common Practice: Partitioning

How to solve?
- Run only a few layers in the TrustZone

- Model Inversion Attack
- Membership Inference Attack

Typical Trusted Memory \(\approx 16\) MB

### Background

Layer-base Partitioning

<table>
<thead>
<tr>
<th>Model</th>
<th># Layers</th>
<th>Pre-trained Model Size (MB)</th>
<th>Peak Mem. Usage (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet</td>
<td>10</td>
<td>0.2</td>
<td>7</td>
</tr>
<tr>
<td>VGG-7</td>
<td>13</td>
<td>0.3</td>
<td>7</td>
</tr>
<tr>
<td>CIFAR</td>
<td>18</td>
<td>30.7</td>
<td>45</td>
</tr>
<tr>
<td>Tiny</td>
<td>22</td>
<td>4.2</td>
<td>71</td>
</tr>
<tr>
<td>Darknet</td>
<td>16</td>
<td>29.3</td>
<td>88</td>
</tr>
<tr>
<td>Extraction</td>
<td>27</td>
<td>93.8</td>
<td>163</td>
</tr>
<tr>
<td>Alexnet</td>
<td>14</td>
<td>249.5</td>
<td>272</td>
</tr>
<tr>
<td>Darknet53</td>
<td>78</td>
<td>159</td>
<td>273</td>
</tr>
<tr>
<td>Inception-v3</td>
<td>145</td>
<td>95.5</td>
<td>448</td>
</tr>
<tr>
<td>Yolov3</td>
<td>107</td>
<td>237</td>
<td>840</td>
</tr>
<tr>
<td>VGG-16</td>
<td>24</td>
<td>528</td>
<td>923</td>
</tr>
</tbody>
</table>
Our Contribution

T-Slices

Overview:

- Utilizes ARM TrustZone with limited trusted memory to protect the entire DL execution
- Does not sacrifice original prediction accuracy
Our Contribution

T-Slices

- Partitions DNN layer into smaller independent segments called *Slices*

- Follows an optimized *Memory Management* plan with on-demand parameter loading scheme
  - Calculated from Hyperparameters

- **Dynamically** determines a set of *Slices* based on the available trusted memory buffer in TrustZone
Convolution Operation

\[ \mathbf{B} = (H_{\text{out}} \times W_{\text{out}}) \times (H_k \times W_k) \times C_{\text{in}} \]

\[ \mathbf{B} \approx \mathbf{O} = H_{\text{out}} \times W_{\text{out}} \times C_{\text{out}} \]

\[ \mathbf{I} = H_{\text{in}} \times W_{\text{in}} \times C_{\text{in}} \]

\[ \mathbf{K} = H_k \times W_k \times C_{\text{in}} \times C_{\text{out}} \]
Slicing for Convolution Operation

\[ \hat{O} \approx \hat{K} \times \hat{I} \]

\[ \hat{I} = H_{in} \times W_{in} \times C_{in} \]

\[ \hat{K} = H_{k} \times W_{k} \times C_{out} \]

\[ \hat{O} = H_{out} \times W_{out} \times C_{out} \]
T-Slices

Memory Buffer Size Comparison

- Darknet Reference Model
- Alexnet Model

(a) layer-based
(b) slice
Memory Buffer Size Comparison

Peak memory required to execute any convolution/connected layer in different CNN architectures. Trusted memory limit considered as 16 MB.
Our Contribution

T-Slices Architecture/Flow

Cloud Server

Normal World

Secure World

App

T-Slices

Model Parameters

Model Hyperparameters

Storage

Memory Management Module

Crypto Module

DL Module

T-Slices TA

Input, Parameters, Hyperparameters
Slice Information, Output

Hyperparameters

Storage
Experimental Setting

- **Device Configuration**
  - STM32MP157C-DK2 with Cortex-A7 32-bit and Cortex-M4 32-bit MPUs
  - Raspberry Pi 3 Model B (RPi3B)

- **Experiment**
  - Image classification with CNN models
  - Compare with Baseline DarkneTZ

- **Performance Metric**
  - Trusted Memory Consumption
  - Prediction Time Overhead
  - Case Studies against prevalent privacy attacks

---

### Dataset and Models

<table>
<thead>
<tr>
<th>Model</th>
<th># Layers</th>
<th># Conv. Layers</th>
<th>Dataset</th>
<th>Pre-trained Model Size (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet</td>
<td>10</td>
<td>2</td>
<td>MNIST</td>
<td>0.2</td>
</tr>
<tr>
<td>CIFAR_SMALL</td>
<td>12</td>
<td>7</td>
<td>CIFAR10</td>
<td>0.08</td>
</tr>
<tr>
<td>VGG-7</td>
<td>13</td>
<td>6</td>
<td>CIFAR10</td>
<td>0.26</td>
</tr>
<tr>
<td>VGG-7</td>
<td>13</td>
<td>6</td>
<td>CIFAR100</td>
<td>0.3</td>
</tr>
<tr>
<td>CIFAR</td>
<td>18</td>
<td>10</td>
<td>CIFAR10</td>
<td>30.7</td>
</tr>
<tr>
<td>Tiny DARKNET</td>
<td>22</td>
<td>16</td>
<td>ImageNet1k</td>
<td>4.2</td>
</tr>
<tr>
<td>EXTRACTION</td>
<td>27</td>
<td>21</td>
<td>ImageNet1k</td>
<td>93.8</td>
</tr>
<tr>
<td>DARKNET REF</td>
<td>16</td>
<td>8</td>
<td>ImageNet1k</td>
<td>29.3</td>
</tr>
<tr>
<td>AlexNet</td>
<td>14</td>
<td>5</td>
<td>ImageNet1k</td>
<td>249.5</td>
</tr>
<tr>
<td>INCEPTIONV3</td>
<td>145</td>
<td>94</td>
<td>ImageNet1k</td>
<td>95.5</td>
</tr>
</tbody>
</table>
### Evaluation

**Trusted Memory Consumption**

- T-Slices on average achieves 72% reduction in peak memory consumption

<table>
<thead>
<tr>
<th>Model</th>
<th>DarknetTZ per Layer</th>
<th>DarknetTZ* per Layer</th>
<th>T-Slices per Slice</th>
<th>% Decrease†</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet</td>
<td>7</td>
<td>0.25</td>
<td>0.1</td>
<td>60</td>
</tr>
<tr>
<td>VGG-7</td>
<td>7</td>
<td>0.7</td>
<td>0.2</td>
<td>71</td>
</tr>
<tr>
<td>CIFAR</td>
<td>45</td>
<td>10.5</td>
<td>1.25</td>
<td>88</td>
</tr>
<tr>
<td>Tiny Darknet</td>
<td>71</td>
<td>9.5</td>
<td>5</td>
<td>47</td>
</tr>
<tr>
<td>Darknet Ref</td>
<td>88</td>
<td>18.5</td>
<td>6.5</td>
<td>65</td>
</tr>
<tr>
<td>Extraction</td>
<td>163</td>
<td>22.6</td>
<td>5.6</td>
<td>75</td>
</tr>
<tr>
<td>AlexNet</td>
<td>272</td>
<td>144</td>
<td>2.75</td>
<td>98</td>
</tr>
<tr>
<td>InceptionV3</td>
<td>337</td>
<td>33</td>
<td>9</td>
<td>73</td>
</tr>
</tbody>
</table>

* with on-demand parameter loading scheme
† decrease from DarknetTZ* to T-Slices
Evaluation

Prediction Time Overhead

- T-Slices on average achieves 29% improvement in execution time
Case Studies

Security Analysis

• **Model Inversion Attack** [1]
  • Reconstruct/recover the training data or any sensitive attributes from the trained ML model

• **Membership Inference Attack** [2]
  • Discover whether a given data sample is a part of the training dataset for the trained ML model


Limitations & Future Work

- Investigate vast DL models unsuitable for memory-constrained edge/IoT devices
  - Peak memory of vgg-16 ~ 923 MB, Yolov3 ~ 840 MB
  - Parallel processing using multiple TZ devices
- Investigate other DL architectures (RNNs)
- Investigate the capability of side-channel attacks on T-Slices
Thank you

Contact information
md.shihabul.islam@utdallas.edu