THE UNIVERSITY OF TEXAS AT DALLAS

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

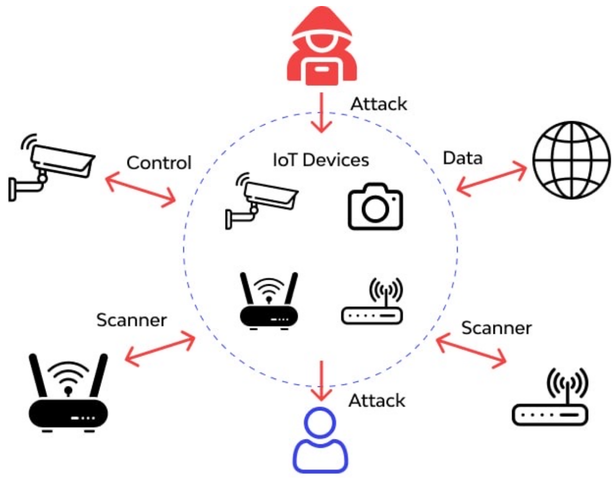
MOTIVATION

- ❑ 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

Image Source: https://www.qualcomm.com/news/onq/2017/08/we-are-making-device-ai-ubiquitous

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:

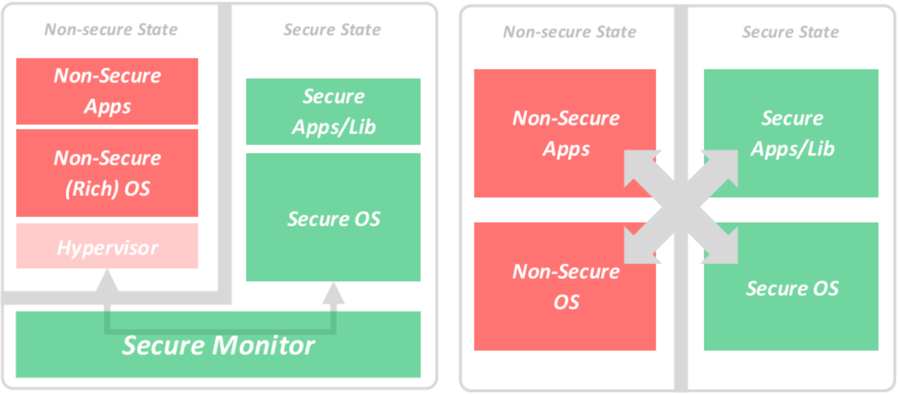
✓ Trusted Execution Environment (TEE) for edge/IoT devices
  - ARM TrustZone

Image Source: https://www.wallarm.com/what/iot-attack

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



(a) TrustZone for Cortex-A (b) TrustZone for Cortex-M

Fig. 1. TrustZone technology.

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

ARM TrustZone Limitations

Limitations:

- ❑ Resource-intensive DL methods
- ❑ Limited trusted memory and resources in TrustZone

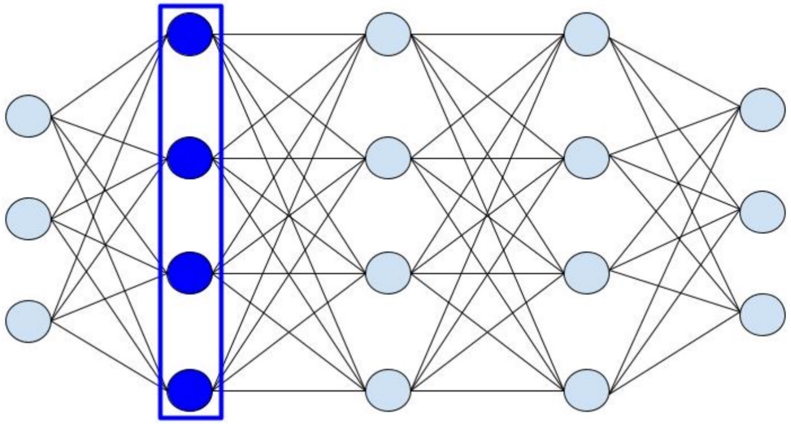
Possible Solutions:

- ❑ Quantization
- ❑ Model pruning

But affects model's prediction accuracy

Common Practice: Partitioning

Layer-base Partitioning



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

Too Large!

Typical Trusted Memory ≈ 16 MB

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

- ➢ Model Inversion Attack
- ➢ Membership Inference Attack

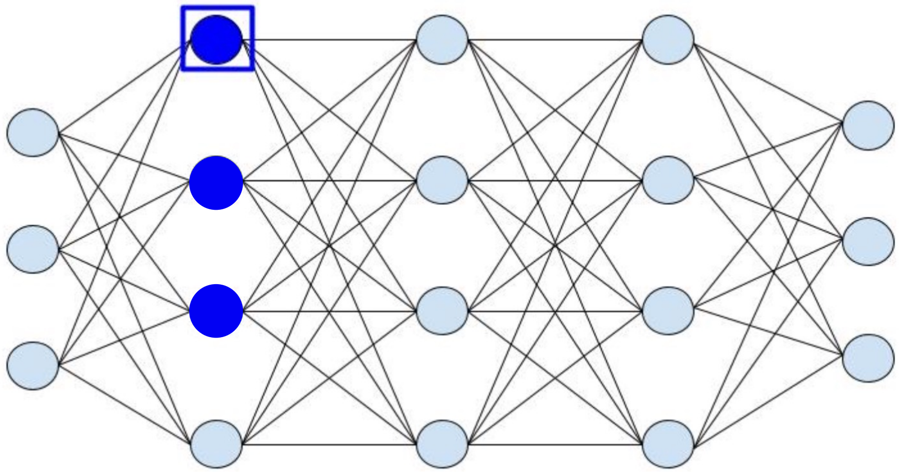
T-Slices

Overview:

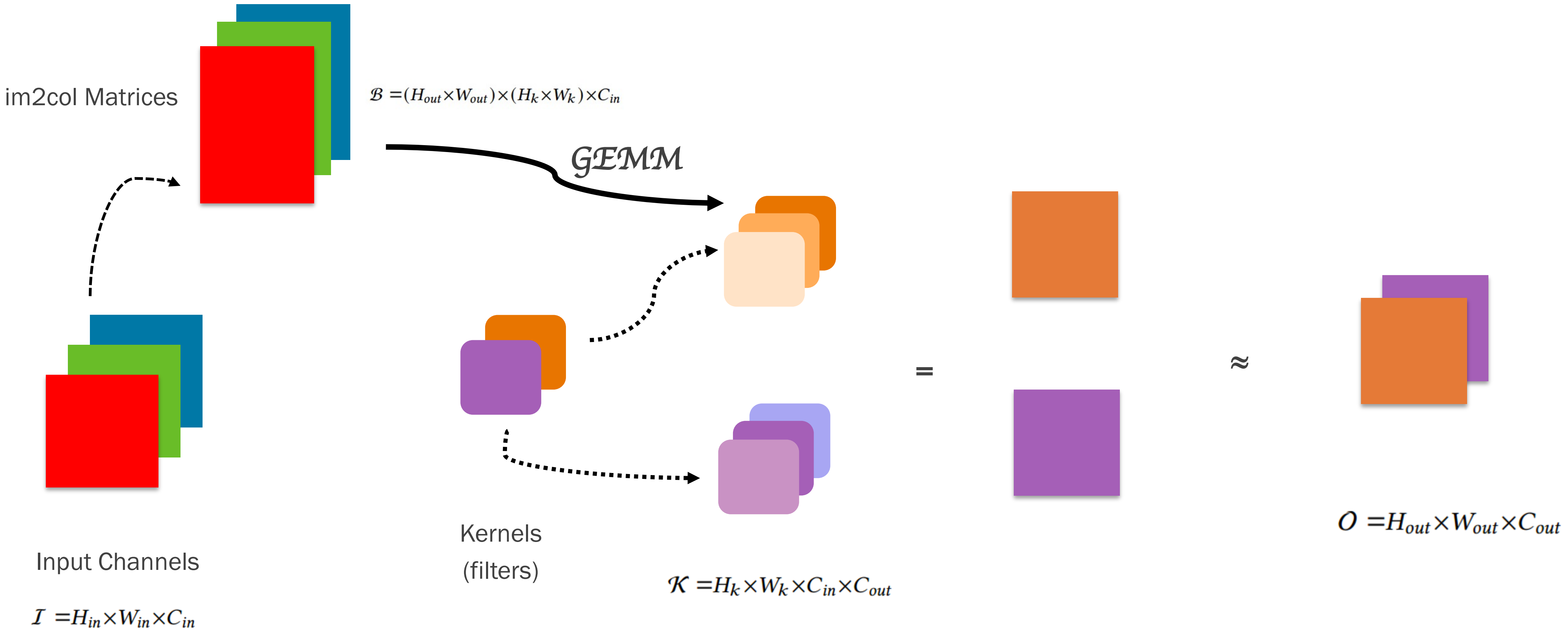
- Utilizes ARM TrustZone with <u>limited trusted memory</u> to protect the <u>entire</u> DL execution
- Does not sacrifice original prediction <u>accuracy</u>

T-Slices

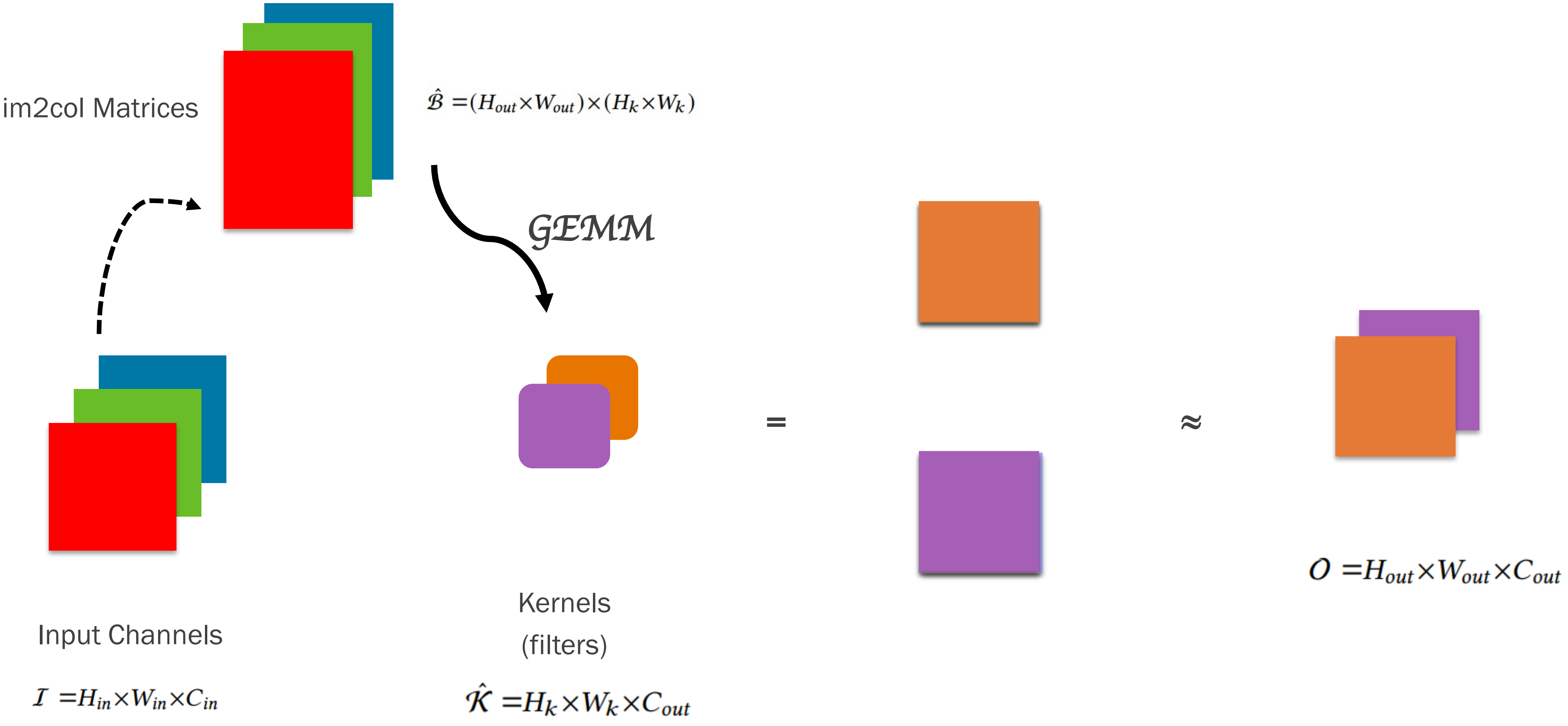
- Partitions DNN layer into smaller independent segments called **Slices**
- Follows an optimized <u>Memory Management</u> plan with <u>on-demand parameter</u> loading scheme
  - Calculated from Hyperparameters
- <u>Dynamically</u> determines a set of **Slices** based on the available trusted memory buffer in TrustZone



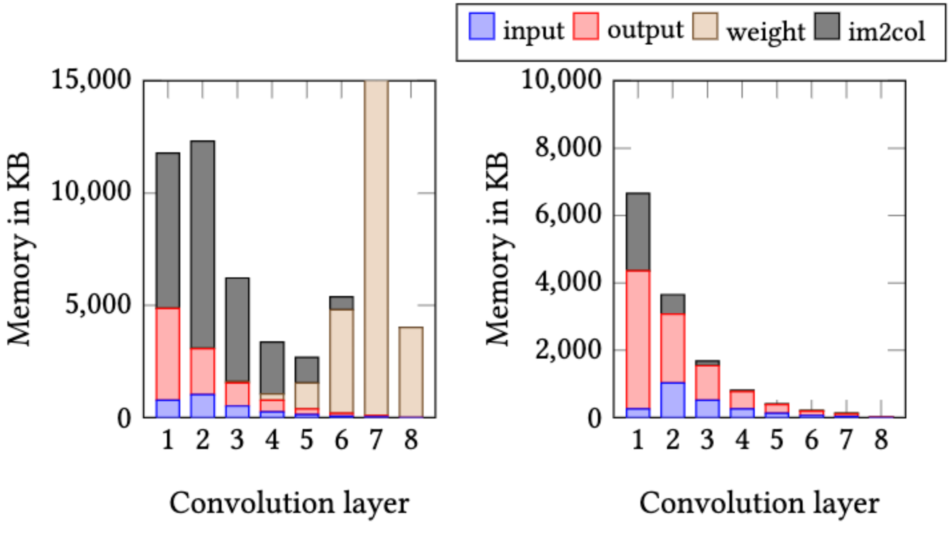
Convolution Operation



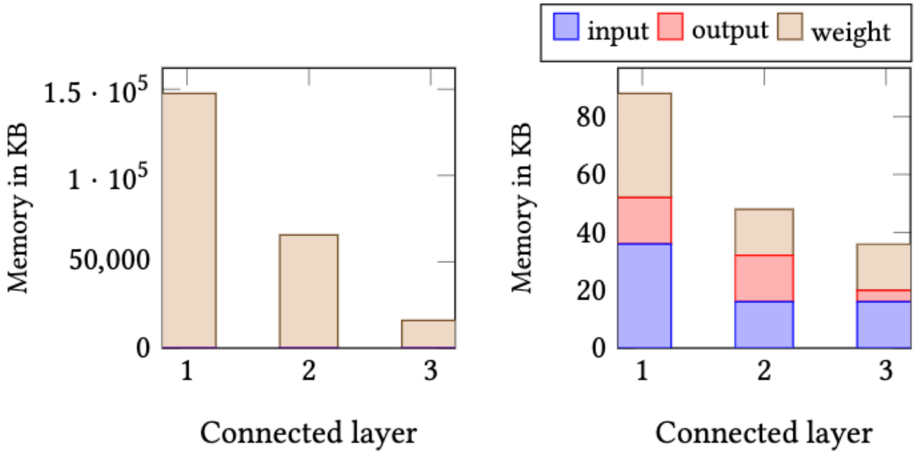
Slicing for Convolution Operation



Memory Buffer Size Comparison



(a) layer-based (b) slice

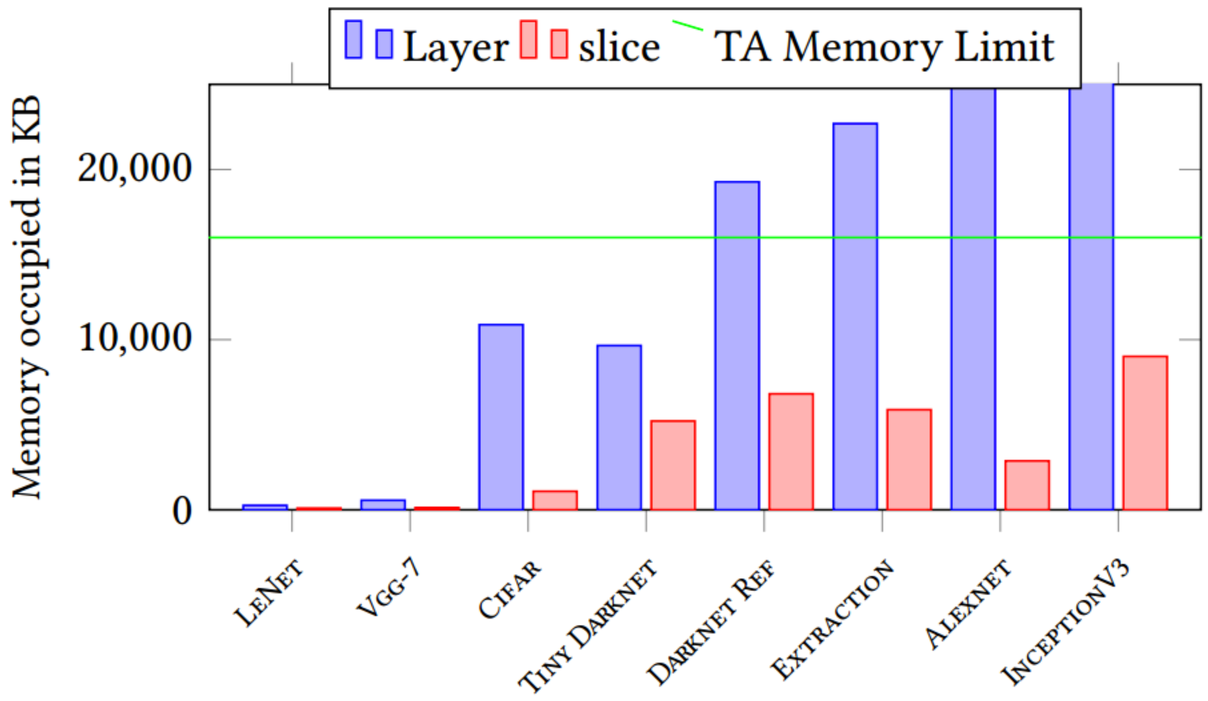


(a) layer-based (b) slice

Darknet Reference Model

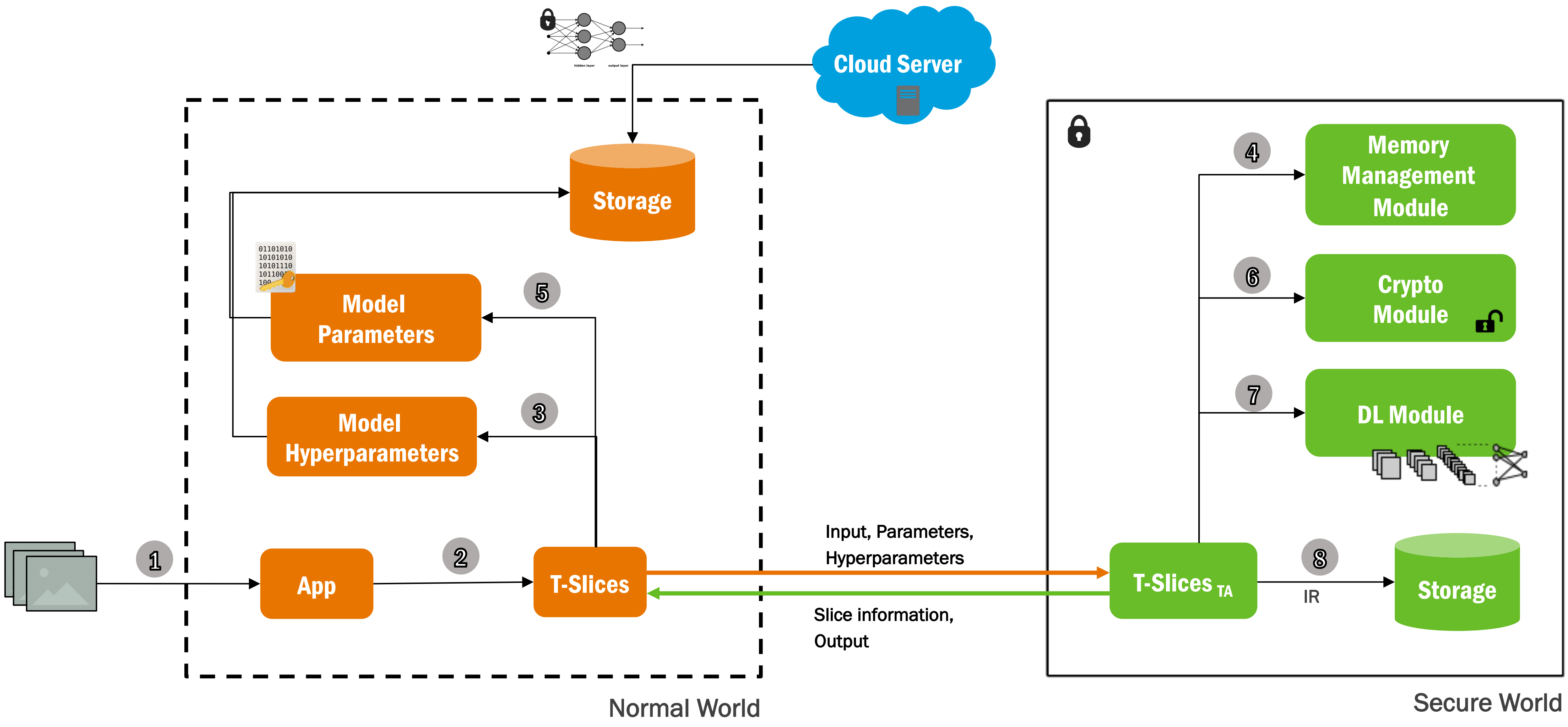
Alexnet Model

Memory Buffer Size Comparison



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

T-Slices Architecture/Flow



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

Dataset and Models

Model	# Layers	# Conv. Layers	Dataset	Pre-trained Model Size (MB)
LeNet	10	2	MNIST	0.2
Cifar_Small	12	7	CIFAR10	0.08
Vgg-7	13	6	CIFAR10	0.26
Vgg-7	13	6	CIFAR100	0.3
Cifar	18	10	CIFAR10	30.7
Tiny Darknet	22	16	ImageNet1k	4.2
Extraction	27	21	ImageNet1k	93.8
Darknet Ref	16	8	ImageNet1k	29.3
AlexNet	14	5	ImageNet1k	249.5
InceptionV3	145	94	ImageNet1k	95.5

ɣ DarkneTZ: towards model privacy at the edge using trusted execution environments, MobiSys 2020

Trusted Memory Consumption

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

Model	DarkneTZ per Layer	DarkneTZ* per Layer	T-Slices per Slice	% Decrease†
LeNet	7	0.25	0.1	60
Vgg-7	7	0.7	0.2	71
Cifar	45	10.5	1.25	88
Tiny Darknet	71	9.5	5	47
Darknet Ref	88	18.5	6.5	65
Extraction	163	22.6	5.6	75
Alexnet	272	144	2.75	98
InceptionV3	337	33	9	73

∗ with *on-demand* parameter loading scheme
† decrease from DarkneTZ* to T-Slices

Prediction Time Overhead

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

STM32MP157C-DK2

CNN	Dataset	DarkneTZ*	T-Slices	% Improvement
LeNet	MNIST	2.44	2.10	14
Cifar_Small	CIFAR10	3.49	3.24	7
Vgg-7	CIFAR10	11.93	6.38	47
Cifar	CIFAR10	608.04	285.07	53
Tiny Darknet	ImageNet1k	874.58	859.34	2
Extraction	ImageNet1k	1244.84	615.56	51
Darknet Ref	ImageNet1k	1175.69	815.55	31
AlexNet	ImageNet1k	✗	1219.31	✗
InceptionV3	ImageNet1k	✗	1928.41	✗

*with *on-demand* parameter loading
✗: Unable to execute due to not enough trusted memory

RPi3B

CNN	Dataset	DarkneTZ*	T-Slices	% Improvement
LeNet	MNIST	0.092	0.092	0
Cifar_Small	CIFAR10	0.19	0.19	0
Vgg-7	CIFAR10	0.309	0.307	1
Cifar	CIFAR10	30.43	30.26	1
Tiny Darknet	ImageNet1k	14.72	14.71	0
Extraction	ImageNet1k	116.57	116.24	0
Darknet Ref	ImageNet1k	18.81	18.78	0
AlexNet	ImageNet1k	✗	44.19	✗
InceptionV3	ImageNet1k	✗	468.1	✗

*with *on-demand* parameter loading
✗: Unable to execute due to not enough trusted memory

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

[1] Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures, ACM CCS 2015

[2] Membership inference attacks against machine learning models, IEEE S&P 2017

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