神經網路推論程式
(Micro-Darknet for Inference)

- 使用Micro-Darknet for Inference協助神經網路推論加速器設計
- 使用Micro-Darknet for Inference快速建立神經網路推論加速器的tool chain 與 SDK 開發
- Micro-Darknet for Inference 目前支援 Image classification (Alexnet, Resnet50, Resnet152, Vgg16, Tiny Darknet 等) 與 Object detection (Yolov3, Yolov3 tiny)
- 使用Micro-Darknet for Inference可自訂自己的 network model

開發者成功大學電機系Computer Architecture and System Laboratory (CASLab)

Contact: chchen@mail.ncku.edu.tw 陳中和教授

Micro Darknet For Inference
MDFI

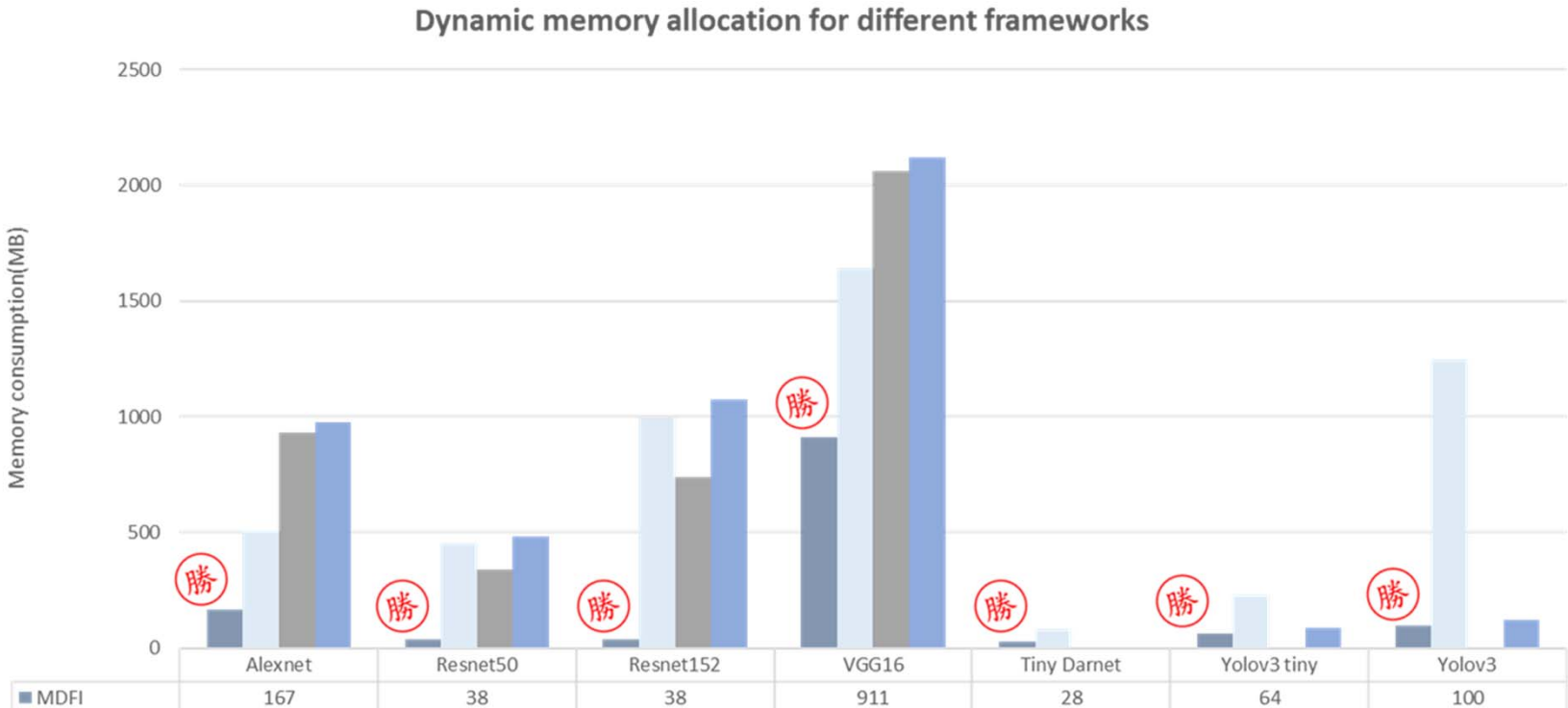
- 我們開發一支C-code only 的**優化推論程式**稱MDFI， 它可讀入 由 **Darknet** 訓練完成的network configuration檔，然後依照network 順序帶入訓練參數 (weights)進行 Inference 運算並輸出結果。MDFI 支援 Image classification (Alexnet, Resnet50, Resnet152, Vgg16, Tiny Darknet 等) 與 Object detection (Yolov3, Yolov3 tiny)、 原始碼支援 configurable network model編譯 。
- MDFI 支援Yolov3 tiny 的 code size 小於 280KB 、compilation time 4 sec、其Heap memory 為原Darknet 的20% 、inference time 快於 Darknet。
- MDFI 只需 GNU C Library (glibc) ，不需要其他套件。MDFI作為純C語言構成的前向傳導DNN框架，主要支援物件辨識網路模型，不使用動態函式庫如Protocol-buffer，以及保持不到280KByte的執行檔大小，適合為終端移動設備所使用。由於不使用動態函式庫，其運算行為可作為**AI 加速硬體設計**的參照，作為ESL的前期描述模型。

Code Size / Compilation Time Comparisons

Framework	Feature	Code size	Compilation time(sec)
Darknet	Train/Inference	822k	13
Caffe	Train/Inference	48M	302
Tenserflow (static lib C)	Train/Inference	221M	3630
Tflite (static lib)	Inference	1885k	231
* MDFI -Os	Inference	155k	3
* MDFI -Ofast	Inference	278k	4
* MDFI_lite_mem	Inference	279k	4

* MDFI -Os : “optimize for size”
* MDFI -Ofast : “Disregard strict standards compliance”
* MDFI_lite_mem : “Layer-wise memory management version, and compile with -Ofast”

Maximum Heap Memory Comparisons

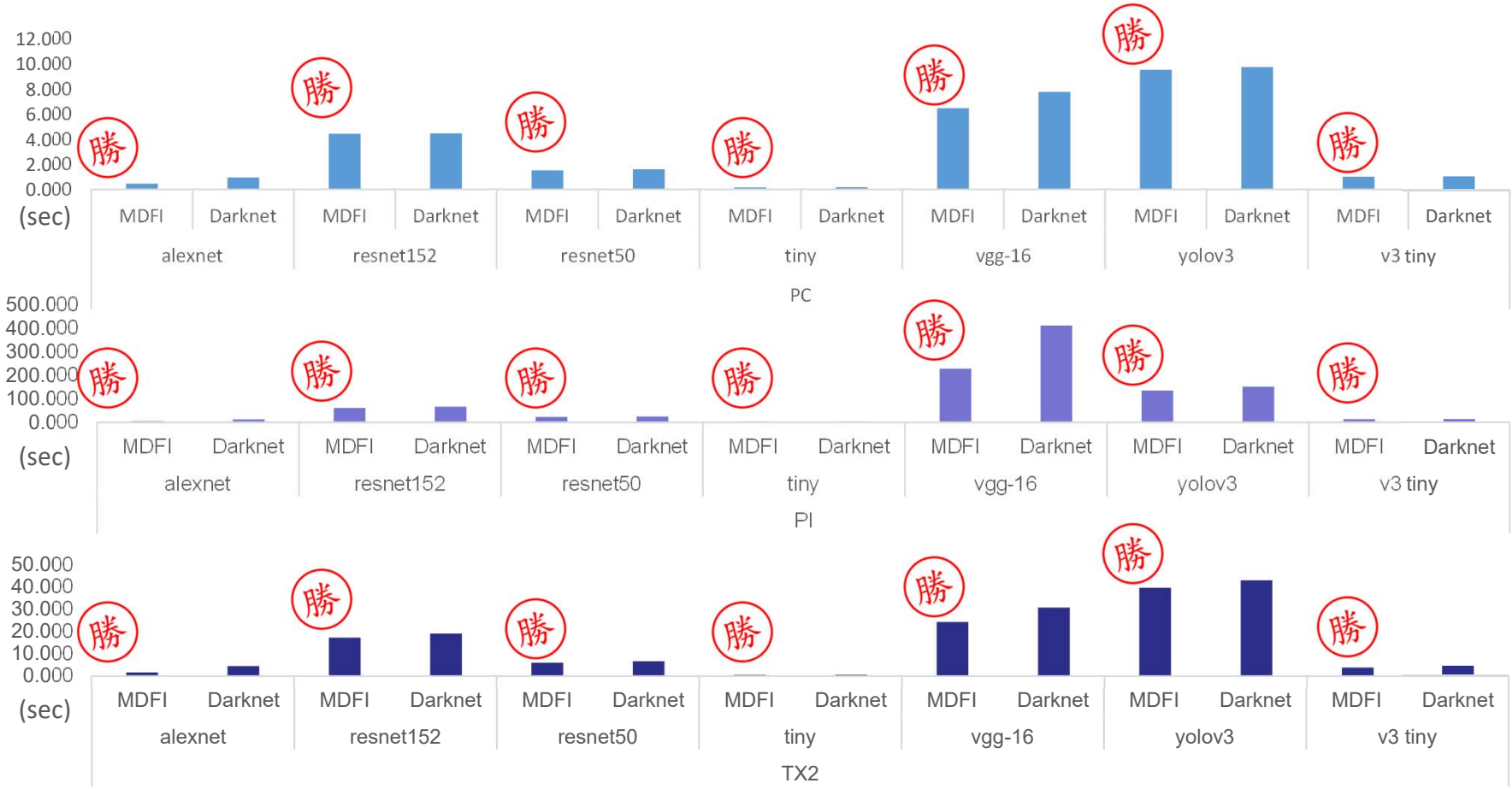


	Alexnet	Resnet50	Resnet152	VGG16	Tiny Darnet	Yolov3 tiny	Yolov3
MDFI	167	38	38	911	28	64	100
Orignal Darknet	507	451	996	1640	82	232	1245
Caffe	931	340	741	2063	0	0	0
Tensorflow	976	484	1076	2119	0	* 87	* 123

MDFI_lite_mem

* There are YOLOv3/YOLOv3 tiny implementation in Tensorflow from GitHub project (aloyschen/tensorflow-yolo3)

Inference Speed Up: MDFI vs Darknet in CPU mode



MDFI_lite_mem

Platform	CPU		Mem. Type	Mem. Size	OS
Personal Computer	Intel® Core™ i7-4770	3.40 GHz	DDR3	32 GB	Ubuntu 14.04
Nvidia Jetson TX 2	ARM Cortex-A57	2.00 GHz	LPDDR4	8 GB	Ububtu 16.04
	NVIDIA Denver2	2.00 GHz			
Raspberry PI 3	ARM Cortex-A53	1.20 GHz	LPDDR2	1 GB	Raspbian

Dependency Library

MDFI

glibc

Compilation time:	**4 Sec**
Code size:	**278 KByte**

glibc

Compilation time:	**13 Sec**
Code size:	**822 KByte**

Caffe

libhdf5-serial libsnappy
libboost libopencv libleveldb
libprotobuf protobuf-compiler
…

Compilation time:	**5 min**
Code size:	**48 MByte**

absl farmhash flatbuffers
gemmlowp googletestfft2d
eigen wheel numpy mock
neon …

Compilation time:	**>1 hour**
Code size:	**221 MByte**

MDFI in Neural Network Framework

CNNs models
Algorithm

Frameworks
Software

MDFI

Darknet

Caffe

TensorFlow

Hardware

RISC-V

Micro-Darknet for Inference

What is MDFI?

AI framework for inference
Configurable C-based framework
No dependency library

Where to use MDFI?

Lite inference AI framework for edge device
Golden model for ASIC development
Reference C code for ESL design

Why to use MDFI?

Code size less than 280KBytes
Compilation time less than 4 sec
Reduce runtime heap usage 79%
Process time of yolov3-tiny is 10% faster than Darknet
Configurable for different models & applications

How to use MDFI?

Symbolic API
Additional driver to co-work with hardware
Software for virtual platform in ESL design
Benchmark design reference for RTL design

AI Accelerator Development with MDFI

Accelerator tool chain, optimizer, accelerator compiler