RSTensorFlow: GPU Enabled TensorFlow for Deep Learning on Commodity Android Devices

Moustafa Alzantot (UCLA), **Yingnan Wang** (UCLA), **Zhengshuang Rene** (UCLA), **Mani Srivastava** (UCLA)

Small networked devices with low computing power (such as phones, surveillance cameras, and IoT devices) have become an essential part of our daily lives. They collect a lot of data and apply machine learning to make useful inferences from it.

Although these devices can process its collected by sending it to a remote server, For security and privacy constraints it is sometimes, preferred to process the data locally on the device. However, popular and commonly used tools and frameworks for machine intelligence are still lacking the ability to make proper use of the available heterogeneous computing resources on these low-end devices.

- We study the benefits of utilizing the heterogeneous (CPU and GPU) computing resources available on commodity Android devices while running deep learning models.
- We implement *RSTensorFlow* as an extension to the open-source framework *TensorFlow* to accelerated the execution of deep learning models on Android phones.
- Our results show that by using the GPU, we can run the models up to 3 times faster.

Implementation

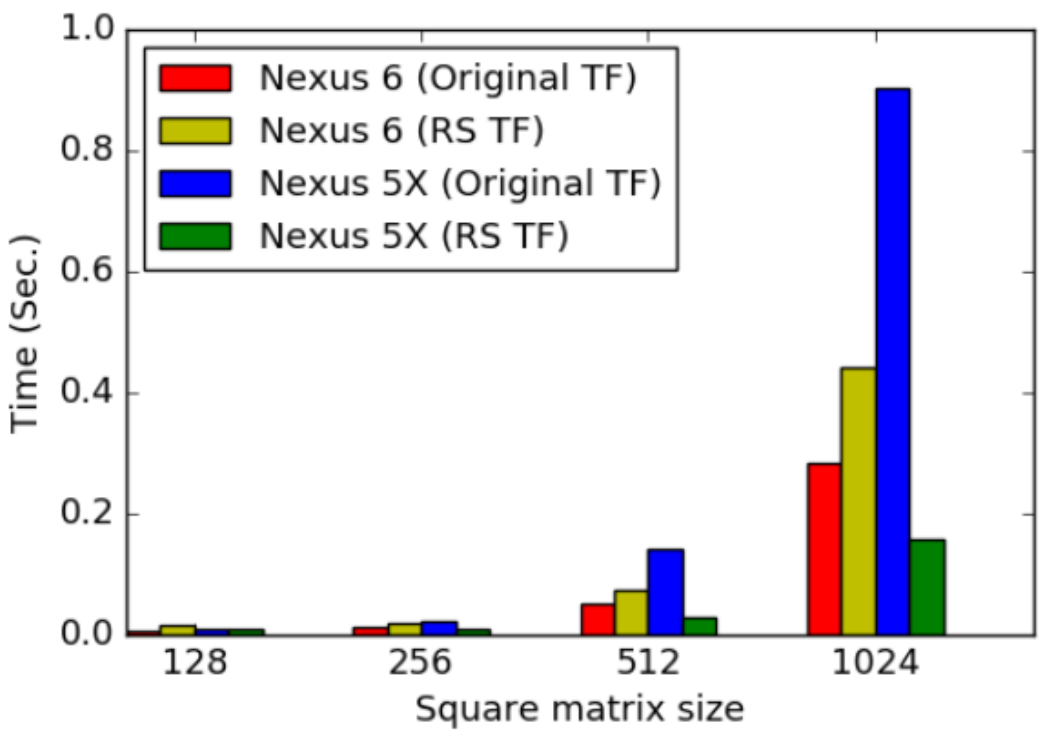
RSTensorFlow is available as open-source: https://nesl.github.io/RSTensorFlow/

We use *RenderScript* framework to implement accelerated versions of the matrix multiplication and convolution operations.

RenderScript parallelize the program execution across the available CPU cores and GPUs.

Results

Matrix Multiplication Result



Inception Model Result

Batch size	Nexus 6			Nexus 5X		
	Original TF	TF + RS MatMul & RS Conv2D	TF + RS MatMul	Original TF	TF + RS MatMul & RS Conv2D	TF + RS MatMul
1	0.453	1.765	0.312	0.699	2.775	0.351
2	0.718	1.757	0.370	1.235	2.782	0.471
3	0.979	1.879	0.475	1.785	2.811	0.649

Distributed Analytics and Information Science International Technology Alliance
Annual Fall Meeting, September 2017