

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

6-2017

DEMO: DeepMon - Building mobile GPU Deep learning models for continuous vision applications

Loc Nguyen HUYNH

Singapore Management University, nlhuynh.2014@phdis.smu.edu.sg

Rajesh Krishna BALAN

Singapore Management University, rajesh@smu.edu.sg

Youngki LEE

Singapore Management University, YOUNGKILEE@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research



Part of the [Hardware Systems Commons](#), and the [Software Engineering Commons](#)

Citation

HUYNH, Loc Nguyen; BALAN, Rajesh Krishna; and LEE, Youngki. DEMO: DeepMon - Building mobile GPU Deep learning models for continuous vision applications. (2017). *MobiSys 2017: Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, Niagara Falls, June 19-23*. 186-186.

Available at: https://ink.library.smu.edu.sg/sis_research/3672

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylids@smu.edu.sg.

DEMO: DeepMon - Building Mobile GPU Deep Learning Models for Continuous Vision Applications

Loc N. Huynh, Rajesh Krishna Balan, Youngki Lee
Singapore Management University
{nlhuynh.2014, rajesh, youngkilee}@smu.edu.sg

Keywords

Mobile GPU; Mobile Sensing; Deep Learning; Continuous Vision

1. INTRODUCTION

Deep learning has revolutionized vision sensing applications in terms of accuracy comparing to other techniques. Its breakthrough comes from the ability to extract complex high level features directly from sensor data. However, deep learning models are still yet to be natively supported on mobile devices due to high computational requirements. In this paper, we present *DeepMon*, a next generation of *DeepSense* [1] framework, to enable deep learning models on conventional mobile devices (e.g. Samsung Galaxy S7) for continuous vision sensing applications. Firstly, *DeepMon* exploits similarity between consecutive video frames for intermediate data caching within models to enhance inference latency. Secondly, *DeepMon* leverages approximation technique (e.g. Tucker decomposition) to build up approximated models with negligible impact on accuracy. Thirdly, *DeepMon* offloads heavy computation onto integrated mobile GPU to significantly reduce execution time of the model.

2. SYSTEM OVERVIEW

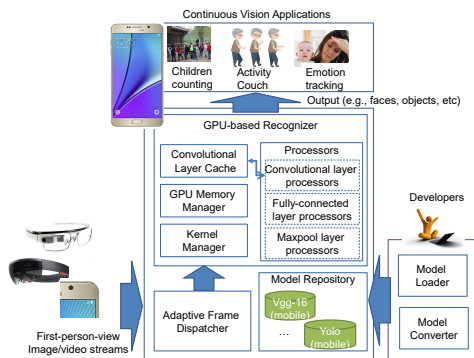


Figure 1: *DeepMon* System Architecture

DeepMon consists of 3 main components (Figure 1).

1) Adaptive frame dispatcher: takes responsibility for choosing important frames and submits them to recognizer.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MobiSys'17 June 19-23, 2017, Niagara Falls, NY, USA

© 2017 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4928-4/17/06.

DOI: <http://dx.doi.org/10.1145/3081333.3089331>

Phone	GPU	APIs	# GPU Cores (#ALUs)	Memory Size (GB)	Memory Bandwid. (GB/s)
Samsung S7	Mali T880	OpenCL/Vulkan	12	4	25.6
Sony Z5	Adreno 430	OpenCL	4 (192)	3	12.8

Table 1: Specs for Commodity Mobile GPUs

2) Model repository: stores pre-trained models for various tasks such as image recognition, object detection. *DeepMon*'s models are not limited to those we provided but can be converted from other framework such as Caffe [2] via our external *model converter*.

3) GPU-based recognizer: processes interesting frames and sends the output back to applications of interests. At first, *DeepMon* compares the current frame with the previous one to see if any regions within a frame should be recomputed to reduce the total computation. After that, caching kernels will be launched to compute only specific regions of the frame, followed by precision reduction if configured by applications.

3. DEMONSTRATION

We demonstrate our *DeepMon* framework on two latest devices, Samsung Galaxy S7 and Z5. Both are commodity smartphone devices running on two different GPU architectures. Galaxy S7 integrates Mali GPU while Z5 uses Adreno GPU as shown in table 1.

In our demonstration, we will trigger Yolo object detection [3] continuously on video streams to localize 20 objects (e.g. person, dog, cat, etc.) and show detected bounding boxes on the screen.

4. REFERENCES

- [1] L. N. Huynh, R. K. Balan, and Y. Lee. Deepsense: A gpu-based deep convolutional neural network framework on commodity mobile devices. In *Proceedings of the 2016 Workshop on Wearable Systems and Applications*, pages 25–30. ACM, 2016.
- [2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.
- [3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.