Exploring the Capabilities of Mobile Devices Supporting Deep Learning

Conference Paper · June 2018

DOI: 10.1145/3220192.3220460

3 authors:

Yitao Chen
Arizona State University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Saman Biookaghazadeh
Arizona State University
6 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Ming Zhao
Arizona State University
65 PUBLICATIONS **1,199** CITATIONS
SEE PROFILE

Exploring the Capabilities of Mobile Devices Supporting Deep Learning

Yitao Chen
Arizona State University
ychen404@asu.edu

Saman Biookaghazadeh
Arizona State University
sbiookag@asu.edu

Ming Zhao
Arizona State University
mingzhao@asu.edu

ABSTRACT

With the increasingly more powerful mobile devices, it becomes possible to perform more deep learning tasks on the devices, and there are also important advantages of learning on devices, such as *personalization* and *efficiency*. However, a good understanding of the capabilities of modern mobile devices for deep learning is generally lacking. To address this gap in knowledge, this paper presents a comprehensive study on performing training and inference of deep neural networks (DNNs) on mobile devices. This study is based on TensorFlow+, an extension of the widely used TensorFlow framework that enables it to train DNNs on devices and use the available GPUs to accelerate the learning. The most significant results of our study are: 1) The size of the network is crucial not only to meet the device's memory constraint but also for training performance; 2) Hardware acceleration is important to the learning speed on devices. By accelerating both the forward and backward path with the device's GPU, our extended TensorFlow can cut down the training time by 44.8%; 3) Comparing CPU, memory, and battery usages, memory size is the most serious constraint to training networks on devices.

CCS CONCEPTS

• **Computer systems organization → Neural networks**; **Heterogeneous (hybrid) systems**; **Embedded software**;

KEYWORDS

Deep learning; Neural networks; Edge computing;

ACM Reference Format:
Yitao Chen, Saman Biookaghazadeh, and Ming Zhao. 2018. Exploring the Capabilities of Mobile Devices Supporting Deep Learning. In *HPDC '18: The 27th International Symposium on High-Performance Parallel and Distributed Computing, June 11–15, 2018, Tempe, AZ, USA.* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3220192.3220460

1 INTRODUCTION

With the rapid advancement of mobile devices, users can accomplish a significant amount of daily tasks on mobile devices. In particular, deep neural networks (DNNs) have unleashed a new wave of applications on mobile devices. Applications such as augmented reality, image classification, and face recognition enable mobile devices to process visual data and learn high-level abstractions from input data by using deep learning models. Such applications are typically implemented by using DNNs hosted on the cloud or pre-trained models downloaded from the cloud.

There are three critical drawbacks with such a cloud-only deep learning approach, 1) the responsiveness of the cloud will be compromised when there is a load surge; 2) users need a reliable network connection to obtain results from the cloud; 3) available computing power on the mobile devices is not able to contribute to the deep learning tasks. In comparison, there are important benefits in learning on the devices, such as personalization, privacy, responsiveness and efficiency.

Prior works [4, 5] have studied the feasibility of running neural networks on mobile platforms, but there are several important limitations. First, these works were done on Nvidia Jetson TK1, which is a discontinued development platform and cannot represent current consumer mobile devices. Second, these works do not evaluate training performance on mobile devices.

In this paper, we investigate the software and hardware capabilities of mobile devices to support deep learning algorithms, and try to provide some answers to the following two fundamental questions: 1) *Are the modern devices capable of training DNNs?* and 2) *What are the important factors that affect the performance and accuracy of deep learning on the devices?* Our study focuses on the following three main aspects: 1) the performance impact of network architecture; 2) the effectiveness of using accelerators available on mobile devices; and 3) the impact of deep learning on the resource and battery usages of a device.

2 APPROACH

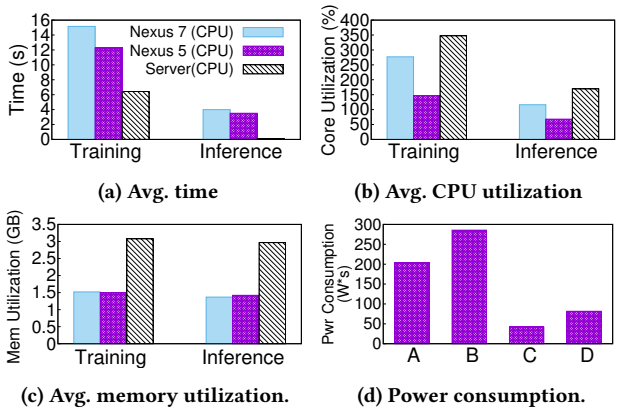
In this section, we discuss in detail our methodology for evaluating the hardware and software capabilities of mobile devices for supporting deep learning workloads.

Available deep learning frameworks, such as TensorFlow and Caffe 2, only support inference on a given network [3], and do not enable training the network on the mobile devices. We modified TensorFlow (version r1.3) [1] to enable training the networks on Android platform. We utilized the Java Interface and extended android library to support training. Since the mobile version of TensorFlow supports only inference, the list of supported operations does not include considerable amount of kernels that are training related. We added these training-related kernels to the mobile version of TensorFlow, by porting them from the TensorFlow C++ core to Android, which we call it ***TensorFlow+***

We further improved TensorFlow+ to leverage available accelerators in the mobile device to improve computational performance of various operations. The existing mobile version of TensorFlow

HPDC '18, June 11–15, 2018, Tempe, AZ, USA

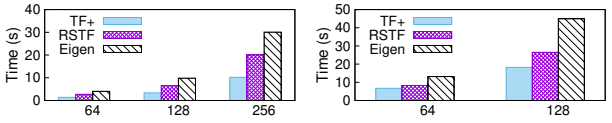
ACM ISBN 978-1-4503-5899-6/18/06…$15.00
https://doi.org/10.1145/3220192.3220460



(a) Avg. time (b) Avg. CPU utilization

(c) Avg. memory utilization. (d) Power consumption.

Figure 1: Performance and resource Utilization



(a) Four-layer (b) Eight-layer

Figure 2: Convolutional-layer-only model training time comparison

uses only CPUs on Android platforms. Such a CPU-based approach becomes insufficient with the increasingly more complex network models.

Previous work RSTensorFlow [2] took the same approach to accelerate matrix multiplication and convolution operations in inference which involves only the forward path of deep learning. But it did not produce good speedup for convolutions (which accounts for around 75% of the forward path time). It also does not support the acceleration of the backward path of deep learning, which is the most intensive component in training, and according to our study, accounts for 70% of the total training time.

3 EVALUATION

In this section, we present the evaluation of deep learning on mobile devices using TensorFlow+. We considered network architectures that are suited for mobile devices, as commonly used DNN models such as VGG16 are too large for the memory size of the devices. So we studied an architecture based on the Mentee network [6] which has five convolutional layers and three fully-connected layers.

Due to memory limitations of the mobile devices, we observed that a large sample size reduces the feasible batch size during training. Hence, we selected our benchmark dataset with a relatively smaller image size, *CIFAR-10* (32×32), to conduct our experiments with a more reasonable batch size of 128. We obtained our results using several different generations of mobile devices, Nexus 7 (released in 2012), Nexus 5 (released in 2013), and Pixel 2 (released in 2018) as well as a server as the baseline.

Figures 1a and 1b show the runtime and resource usages of using the Mentee network to perform training and inference on different platforms. The training time on mobile devices is around 3× longer than the inference time. For the training, Nexus 7 takes 2.3× more time to finish than the server, while Nexus 5 takes 1.9× more time to finish than the server. For inference, Nexus 7 takes 32× longer than the server, whereas Nexus 5 takes 29× longer than the server. The results show that relying solely on mobile CPU is not efficient for training. Figures 1b and 1c show the CPU and memory utilization of the above training and inference experiments. The results confirm that memory is a more serious constraint than CPU for deep learning on mobile devices. The device memory was fully utilized for both training and inference, but the CPUs were not. The training utilized on average 150% of the CPUs on Nexus 5 and 270% on Nexus 7. The better efficiency of newer generation of devices also exhibits in the CPU utilization results. For instance, training the Mentee network on the Nexus 5 consumed 47% fewer CPU, compared to the Nexus 7.

Figure 1d compares the power consumption of various situations, while training the Mentee network: (A) standby with the screen on; (B) training with the screen on; (C) standby with the screen off; (D) training with the screen off. The results show that training consumes 48.5% more power compared to standby mode with the screen turned off, and 28.4% with the screen on.

Figs. 2a and 2b show the training times of convolutional-layer-only networks with various depth and width. The results show that hardware acceleration achieves a speedup of up to 2.2×. We also compared to the performance from using RenderScript to accelerate only the forward path (RSTF) [2]. The results show that the speedup of TensorFlow+ is around 30%.

4 CONCLUSIONS AND FUTURE WORK

This paper investigates the capabilities of mobile devices for supporting deep learning and the effectiveness on utilizing hardware accelerator. Experimental evaluation shows GPU can significantly reduce the training time on mobile devices (up to 2.2×). The results also show that the we need to carefully design the model due to the resource constraints on mobile devices. In future, we will employ training on other on-device hardware accelerators, compare performance between other popular deep learning models. Training on mobile devices is essential to take the collaboration between human and AI to a higher level.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A System for Large-Scale Machine Learning.. In *OSDI*, Vol. 16. 265–283.

[2] Moustafa Alzantot, Yingnan Wang, Zhengshuang Ren, and Mani B Srivastava. 2017. RSTensorFlow: GPU Enabled TensorFlow for Deep Learning on Commodity Android Devices. In *Proceedings of the 1st International Workshop on Deep Learning for Mobile Systems and Applications*. ACM, 7–12.

[3] Google. 2017. TensorFlow Mobile. https://www.tensorflow.org/mobile/android_build

[4] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 615–629.

[5] S Rallapalli, H Qiu, A Bency, S Karthikeyan, R Govindan, B Manjunath, and R Urgaonkar. 2016. Are very deep neural networks feasible on mobile devices. *IEEE Trans. Circ. Syst. Video Technol* (2016).

[6] Ragav Venkatesan and Baoxin Li. 2016. Diving deeper into mentee networks. *arXiv preprint arXiv:1604.08220* (2016).