On the Latency Variability of Deep Neural Networks for Mobile Inference*

Luting Yang
UC Riverside

Bingqian Lu
UC Riverside

Shaolei Ren
UC Riverside

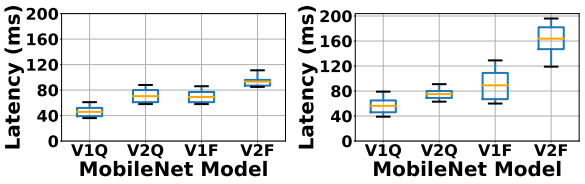
1 Introduction

Inference of deep neural networks (DNNs) on mobile devices is often subject to a highly diverse set of runtime system conditions, such as time-varying resource contention caused by concurrent threads and different numbers of background services [2]. These conditions can all potentially affect the latency performance of DNN-based mobile inference [3, 4]. For example, an intuitive observation is that more resource contention can result in larger average inference latency. On the other hand, latency variability is also crucial for users' quality of experience. Nonetheless, it is less clear if more resource contention also leads to larger latency variability, and if the relative ranking of DNN models in terms of latency variability measured on one device/condition can also carry over to another device/condition.

In this note, we conduct a preliminary measurement study on the latency variability of DNNs for mobile inference. In particular, considering eight popular DNNs for image classification and running them on two mobile devices, we focus on how the CPU resource contention (created by concurrent threads within the same app as DNN inference) affects the inference latency variability. The setup details are available in [1].

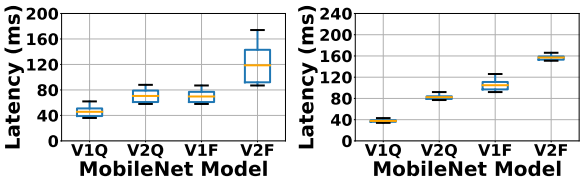
Interestingly and also counter-intuitively, our measurement results show that the relative ranking of DNN models in terms of average latency and latency variability can vary on different devices when the level of CPU contention changes. This implies that the relative ranking of DNN models in terms of latency performance can be both device-dependent and resource contention-dependent. Thus, choosing a set of benchmark devices and system conditions may not be enough to accurately quantify the actual performance for DNN-based inference on all mobile devices.

*This extended abstract summarizes [1]. The authors are supported in part by the U.S. NSF under grants CNS-1551661, ECCS-1610471, and CNS-1910208.



(a) 0 thread (b) 6 threads, 60% active

Figure 1: CPU contention on Samsung Galaxy Tab S5e.



(a) S5e: 2 threads, 20% active (b) A: 2 threads, 20% active

Figure 2: Samsung Galaxy Tab S5e vs. Tab A.

2 Results

We highlight two sets of measurement results in Fig. 1 and Fig. 2, where the 5th, 25th, 75th and 95th percentile and average latencies are shown excluding outliers.

• First, we see from Fig. 1 that when there is more CPU contention (due to more concurrent computation threads), the latency variability of MobileNet V2Q can be even reduced. Moreover, V1F has a comparable latency to V2Q under 0 concurrent thread, but it is outperformed by V2Q when there are six concurrent threads.

• Second, we see from Fig. 2 that given the same number of concurrent threads, MobileNet V1F has similar average latency and latency variability with V2Q on Tab the device S5e, but it is worse than V2Q on Tab A.

Our results demonstrate that for DNN-based mobile inference, more CPU resource contention may not lead to larger latency variability, and the relative ranking of DNN models in terms of latency variability can be both device-dependent and resource contention-dependent.

References

[1] L. Yang, B. Lu, and S. Ren, "A note on latency variability of deep neural networks for mobile inference," *arXiv:2003.00138*, 2020.

[2] C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazelwood, E. Isaac, Y. Jia, B. Jia, T. Leyvand, H. Lu, Y. Lu, L. Qiao, B. Reagen, J. Spisak, F. Sun, A. Tulloch, P. Vajda, X. Wang, Y. Wang, B. Wasti, Y. Wu, R. Xian, S. Yoo, and P. Zhang, "Machine learning at Facebook: Understanding inference at the edge," in *HPCA*, 2019.

[3] Y. Chen, S. Biookaghazadeh, and M. Zhao, "Exploring the capabilities of mobile devices in supporting deep learning," in *SEC*, 2019.

[4] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *ECCV*, 2018.