

Understanding Deep Learning Model Compression on Heterogeneous Mobile Platforms

Qin Qing[†], Jie Ren[‡], Ling Gao[†], Hai Wang[†], Zheng Wang^{*}

[†]Northwest University, China, [‡]Shaanxi Normal University, China, ^{*}Lancaster University, UK

[†] qq@stumail.nwu.edu.cn, [‡] renjie@snnu.edu.cn, [†] {gl, hwang}@nwu.edu.cn, ^{*} z.wang@lancaster.ac.uk

Abstract—

Keywords-workload characterisation

I. INTRODUCTION

II. BACKGROUND AND MOTIVATION

III. EXPERIMENTAL SETUP

IV. EXPERIMENTAL RESULTS

A. Overall Results

B. Compare to Oracle

C. Analysis

V. RELATED WORK

Our work lies at the intersection of multiple research areas: web browsing optimization, task scheduling, energy optimization and predictive modeling. There is no existing work that is similar to ours, in respect to optimizing web workloads across multiple optimization objectives on heterogeneous mobile platforms.

VI. CONCLUSIONS

This paper has presented an automatic approach to optimize the mobile web