Demo: Accelerated Deep Learning Inference for Embedded and Wearable Devices using DeepX

Nicholas D. Lane[‡,*], Sourav Bhattacharya[‡], Petko Georgiev[†],
Claudio Forlivesi[‡], and Fahim Kawsar[‡]
[‡]Bell Labs, [*]University College London, [†]University of Cambridge

1. INTRODUCTION

Recent breakthroughs in deep learning [1] are enabling new ways of interpreting and analyzing sensor measurements to extract high-level information needed by mobile and IoT apps [2]. Thus for improving usability, it is essential that the deep models are embedded in next generation mobile and IoT apps, where inference tasks are often challenging due to high measurement noise. However, deep learning-based models are yet to become mainstream on embedded platforms, where device resources, e.g., memory, computation and energy, are limited. In this demonstration, we present *DeepX*, a software accelerator that allows running *deep neural network* (DNN) and *deep convolutional neural network* (CNN) efficiently on resource constrained mobile platforms. DeepX significantly lowers device resource requirements during deep model-based inferencing, which currently act as the severe bottleneck to wide-scale mobile adoption.

The core of DeepX [3] is mainly comprised of two resource control algorithms, which are targeted to the inference stage of deep learning models. They allow: (1) decomposing monolithic deep model architecture into unit-blocks of various types that are then more efficiently executed by local device processors (e.g., GPUs, CPUs); and (2) performing principled resource scaling that adjusts the architecture of deep models.

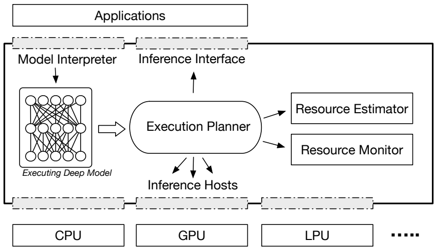
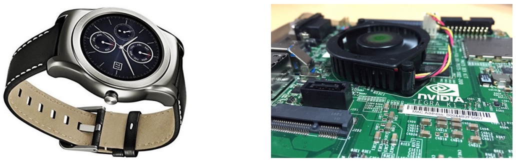


Figure 1: DeepX Proof-of-Concept System

2. DeepX TECHNIQUES

DeepX exploits a mix of network-based computation and heterogeneous local processors to minimize overall execution time, when running deep model-based inferencing. DeepX incorporates two novel techniques:



(a) LG Smartwatch **(b)** Tegra K1

Figure 2: Developer Boards for SoCs used for DeepX Prototype

Runtime Layer Compression (RLC): This technique provides DeepX the ability to shape and control device resources during runtime. Existing approaches that use model compression, focus mainly during the training phase of the deep models, rather than the inference phase. RLC enables runtime control of the memory and computations during inference phase by extending model compression principles such as SVD.

Deep Architecture Decomposition (DAD): DNN/CNN models may include hundreds of layers, each containing thousands of units. DAD creates a *decomposition plan* efficiently, which assigns *unit blocks* of computations (extracted from the layers) to local and remote processors; such plans maximize resource utilization and seek to satisfy user performance goals. An overview of the DeepX architecture is given in Fig. 1.

3. DEMO EXPERIENCE

The demonstration presents a prototype implementation of DeepX running on two latest SoCs. The first SoC is a Qualcomm Snapdragon 400, which is widely used in modern mobile and wearable devices (e.g., LG Urban Smartwatch) and the second SoC is Nvidia Tegra K1 (see Fig. 2 (b)), which is aimed at high performance IoT and embedded devices (e.g., Microwave ovens and automobiles). The demonstration comes with an intuitive GUI, where visitors can interact with the DeepX system by selecting and running audio and image recognition tasks on the mobile SoCs. The demonstration will also showcase an audio signal-based speaker authentication system running locally on a LG smartwatch (contains Snapdragon 400, see Fig. 2 (a)) employing the DeepX engine.

4. REFERENCES

[1] Y. Bengio et al., "Deep learning," 2015, MIT Press.

[2] N. D. Lane et al., "Can Deep Learning Revolutionize Mobile Sensing?," in *HotMobile*, 2015.

[3] N. D. Lane et al., "DeepX: A software accelerator for low-power deep learning inference on mobile devices," in *IPSN*, 2016.

[4] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *IJCV*, 2015.

MobiSys'16 Companion June 25-30, 2016, Singapore, Singapore

ACM ISBN 978-1-4503-4416-6/16/06.
DOI: http://dx.doi.org/10.1145/2938559.2949718