Operating System Modeling and Optimization for Intelligent Edge

Pr. Smail Niar (smail.niar@uphf.fr) and
Dr. Hamza Ouarnoughi (hamza.ouarnoughi@uphf.fr)
Computer Dept, LAMIH CNRS

March 2021

Keywords: Machine Learning, Deep Learning, Edge devices, Operating System, Virtualization, Autonomous Driving.

1 Contexte

Many approaches have been proposed to enable Edge devices to support Deep Learning (DL) applications. These approaches operate on two levels:

1) the application level and

2) the hardware level.

At the application level, the targeted objective is to optimize DL algorithms to match Edge devices by applying model compression methods that reduce the complexity of DL models . Model quantization and pruning are examples of such compression methods. At the hardware level, state-of-the-art approaches enhance the implementation and hardware architectures of Edge devices, such as Tensors in GPUs, TPUs, FPGA [1].

We observe that an intermediate layer, the system software level, represented by the operating system, virtual machines, and containers, remains unchanged over time and does not follow the trend.

This observation is more important when it comes to memory and Inputs/Outputs (I/Os) management. Knowing that memory and I/Os have an important impact on DL application especially during model training, it constitutes a bottleneck and slows down the technological advances made at the two previous levels.

In addition, the profile and the used DL frameworks vary as well as the used Edge hardware while keeping the same conventional memory and I/Os management which makes its investigation and study more centralized and beneficial for the other levels [4] [3].

2 Objective

The aim of this thesis is to characterize and optimize the impact of memory and I/Os at the operating system level for Edge devices running DL applications, namely:

- Convolution Neural Networks (CNN) for computer vision purposes
- RNN and Transformers for Natural Language Processing (NLP) purposes

The first step of the thesis is to characterize the interactions between CNNs, RNNs, and Transformers, and the operating system running on top of different Edge hardware platforms [2]. We will characterize memory access and I/Os when varying the following components:

- **DL model**s where we study several use cases that are mainly using DL models for Autonomous Driving (AD) purposes.
- **DL application lifecycle** where characterize the impact of the application during the DL model training and the inference phases.
- **DL framework** used to implement the above mentioned models.
- **Virtualization and containerization** used to isolate the DL application.
- **Data management system** where the hardware and the software parts can vary.
- **Edge device processing units** used to run DL applications

The long term objective is to propose a set of Linux-based services designed for Edge devices running DL applications. The obtained results will be evaluated on Edge devices used for the Autonomous Driving (AD) use case which is one of the hottest research fields investigated in LAMIH.

3 Background and Qualifications

- Master degree in computer science
- Good theoretical and practical knowledge on operating system, performance modeling, machine learning and deep learning.
- Experience in programming (C, C++, Shell, Python).
- Experience in Machine/Deep Learning libraries and frameworks (scikit-learn, PyTorch, TensorFlow).
- Confidence in English writing and speaking

References

[1] Hadjer Benmeziane, Kaoutar El Maghraoui, Hamza Ouarnoughi, Smail Niar, Martin Wistuba, and Naigang Wang. A comprehensive survey on hardware-aware neural architecture search, 2021.

[2] Halima Bouzidi, Hamza Ouarnoughi, Smaïl Niar, and Abdessamad Ait El Cadi. Performance prediction for convolutional neural networks in edge devices. *CoRR*, abs/2010.11297, 2020.

[3] Arpan Gujarati, Reza Karimi, Safya Alzayat, Wei Hao, Antoine Kaufmann, Ymir Vigfusson, and Jonathan Mace. Serving dnns like clockwork: Performance predictability from the bottom up, 2020.

[4] Abhishek Vijaya Kumar and Muthian Sivathanu. Quiver: An informed storage cache for deep learning. In *18th USENIX Conference on File and Storage Technologies (FAST 20)*, pages 283–296, Santa Clara, CA, February 2020. USENIX Association.