Energy-aware Adaptive Approximate Computing for Deep Learning Applications

Nima TaheriNejad and Salar Shakibhamedan
TU Wien, Vienna, Austria
E-mail: {nima.taherinejad, salar.shakibhamedan}@tuwien.ac.at

***Abstract*—Application that use deep learning incur a substantial amount of energy consumption. Reducing this energy footprint is important, especially for applications such as Internet of Things (IoT) Embedded Systems (ESs), where resources are scarce. Here, we present computational self-awareness as a promising solution for intelligently adapt machine learning algorithms at runtime to reduce their energy consumption. In particular, we focus on approximation as a key enabler knob for such adaptivity. We show that the benefits of such an approach can be up to 2.5× energy savings.**

I. Introduction

The number of Internet of Things (IoT) devices has been exponentially increasing, raising the importance of the challenges Embedded Systems (ESs) face. Limited computational resources, hardware, and energy source are among these challenges, intensified by having to operate in highly dynamic environments, see e.g. [1]. One of the promising solutions to these challenges is woke system design or more commonly known as self-aware system design [2], [3]. Woke systems are a type of adaptive systems with a more intelligent awareness built into them. We explain the basics of computational self-awareness first and highlight the difference between self-* systems and *-aware systems. Next, we show the benefits of computational self-awareness in the context of wearable healthcare systems [4], [5] and Multi-Processor System-on-Chips (MPSoCs) for mobile devices [6], [7]. In the second part of the talk, we focus on the use of adaptive approximate computing as a key action knob for improving the efficiency of deep learning applications. We focus on energy as the main optimization target and energy-awareness as the enabler and drive for adaptivity in the system.

II. Energy-aware Approximate Deep Learning

As shown in [8], on average about 83% of runtime computations for many machine learning and deep learning applications can be approximated. This can lead to substantial savings in power consumption, for instance in [9], the authors saved 63% of the power consumption by using approximations. This adaptive approximation can be performed on the hardware, e.g., using Dynamic Partial Reconfiguration (DPR) and by instantiating of different arithmetic hardware units [10], or using dynamic change of the frequency of operations [11]. It could also happen at the software level, e.g., by changing the utilized machine learning algorithm based on input data, such as [9] and [12], or by changing the memory access policy used in [13]. The benefits of these approaches can be up to 2.5× savings in the energy consumption of the system. We further details these approaches and their benefit in our talk.

Acknowledgement

The authors gratefully acknowledge funding from European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska Curie grant agreement No. 956090 (APROPOS: Approximate Computing for Power and Energy Optimisation, http://www.apropos-itn.eu/).

References

[1] N. TaheriNejad. Wearable medical devices: Challenges and self-aware solutions. In *IEEE Life Sciences*, volume 2, pp. 5–6, April 2019.
[2] N. Dutt and N. TaheriNejad. Self-awareness in cyber-physical systems. In *2016 29th International Conference on VLSI Design and 2016 15th International Conference on Embedded Systems (VLSID)*, pp. 5–6, Jan 2016.
[3] K. Bellman *et al.* Self-aware cyber-physical systems. *ACM Transactions on Cyber-Physical Systems*, pp. 1–24, 2020.
[4] A. Anzanpour *et al.* Self-awareness in remote health monitoring systems using wearable electronics. In *Proceedings of Design and Test Europe Conference (DATE)*, Lausanne, Switzerland, March 2017.
[5] M. Götzinger *et al.* Rosa: A framework for modeling self-awareness in cyber-physical systems. *IEEE Access*, 8:141373–141394, 2020.
[6] E. Shamsa *et al.* User-centric resource management for embedded multi-core processors. In *The 33rd International Conference on VLSI Design and The 19th International Conference on Embedded Design*, pp. 1–6, 2020.
[7] E. Shamsa *et al.* Ubar: User and battery aware resource management for smartphones. *ACM Transactions on Embedded Computing Systems (TECS)*, pp. 1–23, 2021.
[8] V. K. Chippa *et al.* Analysis and characterization of inherent application resilience for approximate computing. In *Proceedings of the 50th Annual Design Automation Conference*, DAC '13, New York, NY, USA, 2013. Association for Computing Machinery.
[9] J. Lee *et al.* Tod: Transprecise object detection to maximise real-time accuracy on the edge. In *2021 IEEE 5th International Conference on Fog and Edge Computing (ICFEC): Proceedings*, pp. 53–60, June 2021. 5th IEEE International Conference on Fog and Edge Computing 2021, IEEE ICFEC ; Conference date: 10-05-2021 Through 13-05-2021.
[10] M. Masadeh *et al.* Machine-learning-based self-tunable design of approximate computing. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 29(4):800–813, 2021.
[11] H. Sharif *et al.* Approxtuner: A compiler and runtime system for adaptive approximations. In *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP '21, pp. 262–277, New York, NY, USA, 2021. Association for Computing Machinery.
[12] M. A. Scrugli *et al.* A runtime-adaptive cognitive iot node for healthcare monitoring. In *Proceedings of the 16th ACM International Conference on Computing Frontiers*, CF '19, pp. 350–357, New York, NY, USA, 2019. Association for Computing Machinery.
[13] B. Cox *et al.* Masa: Responsive multi-dnn inference on the edge. In *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10, 2021.

2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI) | 978-1-6654-6605-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISVLSI54635.2022.00072

2159-3477/22/$31.00 ©2022 IEEE
DOI 10.1109/ISVLSI54635.2022.00072