AI Models for Edge Computing: Hardware-aware Optimizations for Efficiency

Hai “Helen” Li
Electrical and Computer Engineering Department
Duke University
Durham, North Carolina, USA
hai.li@duke.edu

ABSTRACT

As artificial intelligence (AI) transforms various industries, state-of-the-art models have exploded in size and capability. The growth in AI model complexity is rapidly outstripping hardware evolution, making the deployment of these models on edge devices remain challenging. To enable advanced AI locally, models must be optimized for fitting into the hardware constraints. In this presentation, we will first discuss how computing hardware designs impact the effectiveness of commonly used AI model optimizations for efficiency, including techniques like quantization and pruning. Additionally, we will present several methods, such as hardware-aware quantization and structured pruning, to demonstrate the significance of software/hardware co-design. We will also demonstrate how these methods can be understood via a straightforward theoretical framework, facilitating their seamless integration in practical applications and their straightforward extension to distributed edge computing. At the conclusion of our presentation, we will share our insights and vision for achieving efficient and robust AI at the edge..

979-8-3503-4859-0/DATE24/© 2024 EDAA