Summary

Approximate Computing (AxC) has emerged as a powerful paradigm for improving computational efficiency by strategically introducing acceptable levels of inaccuracy in computations, thereby reducing power consumption, execution time, and hardware complexity. However, identifying optimal approximation configurations requires navigating a high-dimensional design space, posing many methodological challenges. This thesis addresses these challenges through Design Space Exploration (DSE) approaches aimed at enabling systematic application of AxC techniques to computational systems.

Firstly, we introduce an Interval Arithmetic (IA)-based methodology tailored to evaluate the impact of precision reduction in Spiking Neural Networks (SNNs). The IA-based model provides an analytical estimate of how bit-width truncation errors propagate through neuron computations, enabling the identification of acceptable approximation levels with limited degradation in accuracy under the evaluated settings. This approach enables faster exploration of precision-reduction strategies, striving to reduce exploration time relative to exhaustive search methods in terms of computational efficiency and speed. This method is further enhanced by integrating "watchers," mechanisms that facilitate fine-grained, neuron-specific precision adjustments, aiming to limit memory usage and computational overhead while maintaining the desired classification accuracy.

Secondly, we propose a Reinforcement Learning (RL)-based approach for exploring approximation opportunities in conventional CPU-based applications. Our method automates the selective replacement of arithmetic operations with approximate alternatives from a predefined operator library. The RL agent is utilized to perform a DSE with the aim of finding a trade-off among multiple objectives, including accuracy, power consumption, and execution time. This DSE goal is realized by the RL agent, iteratively evaluating design choices and adapting based on performance feedback. Experimental evaluations on representative benchmarks, such as matrix multiplication and FIR filters, suggest that the RL methodology can identify configurations that reduce power and computation time subject to the imposed accuracy constraints.

Collectively, these methodologies were proposed with the aim of providing practical solutions for systematically exploring approximation strategies across the studied computational domains. The contributions of this thesis can support the deployment of AxC techniques and motivate future investigations in the design of resource-efficient computing systems.