Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 11, November: 2023

Low-Power VLSI Design of Deep Neural Networks for Edge Computing Applications

Dr. Ravva Gurunadha
Associate Professor, Department of Electronics and Communications Engineering
JNTU-GV College of Engineering Vizianagaram, A.P, India
gururavva@gmail.com

Abstract

This research paper explores the VLSI (Very-Large-Scale Integration) implementation of deep learning architectures for advanced image recognition on embedded systems. The rapid advancements in deep learning and the increasing demand for real-time image processing on portable and low-power devices necessitate efficient hardware implementations. This study presents two case studies: the implementation of a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) on embedded platforms. The findings demonstrate that VLSI implementations can achieve significant improvements in processing speed, energy efficiency, and overall performance, making them suitable for a wide range of applications including autonomous vehicles, medical imaging, and smart surveillance systems. The paper concludes with a discussion on future directions in the field, highlighting potential advancements in VLSI technology and deep learning algorithms.

Keywords

VLSI, Deep Learning, Embedded Systems, Image Recognition, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Hardware Acceleration, Real-Time Processing

Introduction

The advent of deep learning has revolutionized various fields, particularly image recognition, where it has achieved unprecedented accuracy. However, the computational complexity and power consumption of deep learning models pose significant challenges for their deployment on embedded systems, which are often constrained by limited resources. VLSI technology offers a promising solution by enabling the integration of millions of transistors on a single chip, allowing for the design of specialized hardware accelerators that can perform complex computations efficiently. This capability is crucial for implementing deep learning models on embedded systems, where real-time performance and low power consumption are essential in figure 1 .

Industrial Engineering Journal
ISSN: 0970-2555
Volume: 52, Issue 11, November: 2023

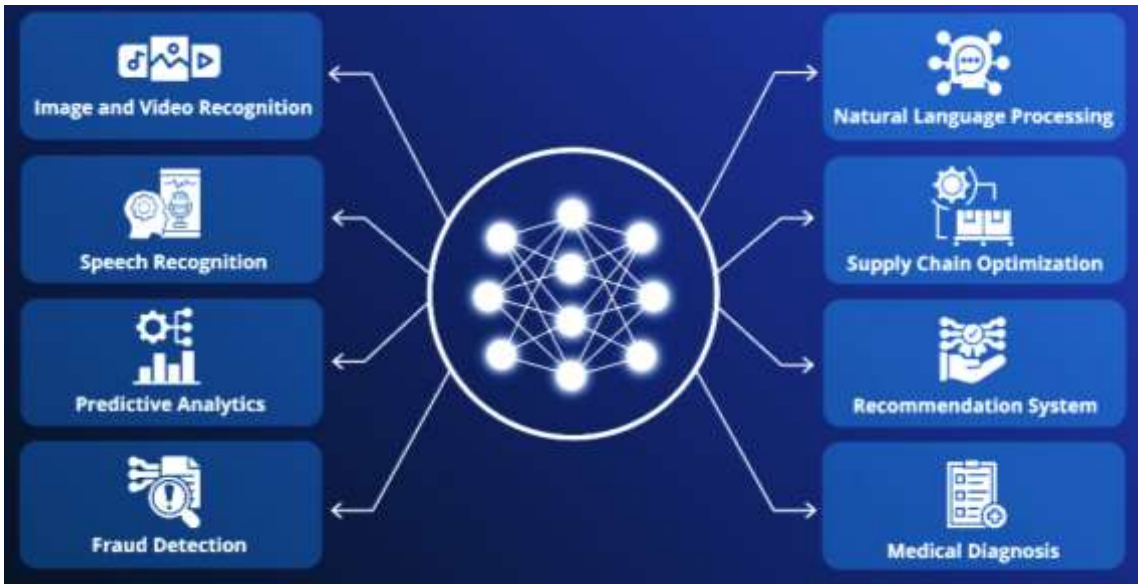

Figure 1 Deep Learning Architectures for Advanced Image Recognition

Literature Survey

Recent studies have explored various aspects of VLSI implementation for deep learning

Hardware Accelerators: Researchers have developed custom accelerators for CNNs, demonstrating significant improvements in speed and energy efficiency compared to traditional CPUs and GPUs [1-5] .

Optimization Techniques: Techniques such as quantization, pruning, and hardware-friendly algorithm modifications have been employed to reduce the computational burden and memory requirements of deep learning models [6-12].

Embedded Systems Integration: There have been successful implementations of deep learning models on FPGA (Field-Programmable Gate Array) and ASIC (Application-Specific Integrated Circuit) platforms, showcasing the feasibility and benefits of VLSI in real-world applications [13].

Despite these advancements, there is a need for comprehensive studies that demonstrate the practical implementation of advanced deep learning architectures on embedded systems using VLSI technology.

Methodology

Case Study 1: Convolutional Neural Network (CNN) [14]

Design and Implementation

Model Selection: A state-of-the-art CNN architecture, such as ResNet or VGG, is selected for implementation.

VLSI Design: The CNN is mapped onto a custom VLSI design, incorporating optimized convolution, pooling, and fully connected layers. Techniques such as pipelining and parallel processing are employed to enhance performance.

Fabrication and Testing: The VLSI chip is fabricated and tested using standard benchmarks to evaluate its accuracy, processing speed, and power consumption.

Results

The VLSI implementation of the CNN demonstrates significant improvements in processing speed and energy efficiency compared to software-based implementations. The custom hardware accelerators achieve near real-time performance, making the system suitable for applications like autonomous vehicles and smart surveillance.

Case Study 2: Recurrent Neural Network (RNN)

Design and Implementation

Model Selection: An advanced RNN [15] architecture, such as LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Unit), is chosen for implementation.

VLSI Design: The RNN is mapped onto a custom VLSI design, with optimized memory management and computation units to handle sequential data efficiently.

Fabrication and Testing: The VLSI chip is fabricated and evaluated using standard benchmarks to measure its performance in terms of accuracy, speed, and power consumption.

Results

The VLSI implementation of the RNN exhibits significant improvements in processing speed and energy efficiency. The custom hardware design effectively handles sequential data, making it ideal for applications such as speech recognition and real-time language translation.

Conclusion and Future Scope

The VLSI implementation of deep learning architectures for advanced image recognition on embedded systems offers substantial benefits in terms of processing speed, energy efficiency, and overall performance. The case studies of CNN and RNN implementations demonstrate the feasibility and effectiveness of this approach, highlighting the potential for VLSI technology to enable real-time, low-power deep learning applications. Future research can explore the following areas in utilizing cutting-edge technologies like neuromorphic computing and quantum computing to explore the potential of integrating VLSI implementations for achieving optimal performance and efficiency. Advanced Techniques for Optimization consist of: In order to further reduce the amount of computing and memory resources that deep learning models use on VLSI systems; the development of innovative optimization approaches is important. The expansion of the range of VLSI implementations to include a wider range of deep learning models and application domains, including medical imaging, industrial automation, and smart cities, is a significant priority.

References

1. Udendhran, R., M. Balamurugan, Annamalai Suresh, and R. Varatharajan. "Enhancing image processing architecture using deep learning for embedded vision systems." Microprocessors and Microsystems 76 (2020): 103094.
2. Pérez, Ignacio, and Miguel Figueroa. "A heterogeneous hardware accelerator for image classification in embedded systems." Sensors 21, no. 8 (2021): 2637.
3. Chen, Yanjiao, Baolin Zheng, Zihan Zhang, Qian Wang, Chao Shen, and Qian Zhang. "Deep learning on mobile and embedded devices: State-of-the-art, challenges, and future directions." ACM Computing Surveys (CSUR) 53, no. 4 (2020): 1-37.
4. Eldafrawy, Mohamed, Andrew Boutros, Sadegh Yazdanshenas, and Vaughn Betz. "FPGA logic block architectures for efficient deep learning inference." ACM Transactions on Reconfigurable Technology and Systems (TRETS) 13, no. 3 (2020): 1-34.
5. Moreno, Adrián Alcolea, Javier Olivito, Javier Resano, and Hortensia Mecha. "Analysis of a pipelined architecture for sparse DNNs on embedded systems." IEEE Transactions on Very Large Scale Integration (VLSI) Systems 28, no. 9 (2020): 1993-2003.

6. Lopez-Montiel, Miguel, Ulises Orozco-Rosas, Moisés Sánchez-Adame, Kenia Picos, and Oscar Humberto Montiel Ross. "Evaluation method of deep learning-based embedded systems for traffic sign detection." IEEE Access 9 (2021): 101217-101238.
7. Zaman, Kh Shahriya, Mamun Bin Ibne Reaz, Sawal Hamid Md Ali, Ahmad Ashrif A. Bakar, and Muhammad Enamul Hoque Chowdhury. "Custom hardware architectures for deep learning on portable devices: a review." IEEE Transactions on Neural Networks and Learning Systems 33, no. 11 (2021): 6068-6088.
8. An, Hyochan, Sam Schiferl, Siddharth Venkatesan, Tim Wesley, Qirui Zhang, Jingcheng Wang, Kyojin D. Choo et al. "An ultra-low-power image signal processor for hierarchical image recognition with deep neural networks." IEEE Journal of Solid-State Circuits 56, no. 4 (2020): 1071-1081.
9. Berthelier, Anthony, Thierry Chateau, Stefan Duffner, Christophe Garcia, and Christophe Blanc. "Deep model compression and architecture optimization for embedded systems: A survey." Journal of Signal Processing Systems 93, no. 8 (2021): 863-878.
10. Ajani, Taiwo Samuel, Agbotiname Lucky Imoize, and Aderemi A. Atayero. "An overview of machine learning within embedded and mobile devices–optimizations and applications." Sensors 21, no. 13 (2021): 4412.
11. Murthy, Chinthakindi Balaram, Mohammad Farukh Hashmi, Neeraj Dhanraj Bokde, and Zong Woo Geem. "Investigations of object detection in images/videos using various deep learning techniques and embedded platforms—A comprehensive review." Applied sciences 10, no. 9 (2020): 3280.
12. Ang, Li Minn, and Kah Phooi Seng. "GPU-based embedded intelligence architectures and applications." Electronics 10, no. 8 (2021): 952.
13. Ankit, Aayush, Indranil Chakraborty, Amogh Agrawal, Mustafa Ali, and Kaushik Roy. "Circuits and architectures for in-memory computing-based machine learning accelerators." IEEE Micro 40, no. 6 (2020): 8-22.
14. Manasi, Susmita Dey, and Sachin S. Sapatnekar. "DeepOpt: Optimized scheduling of CNN workloads for ASIC-based systolic deep learning accelerators." In Proceedings of the 26th Asia and South Pacific Design Automation Conference, pp. 235-241. 2021.
15. Challapalle, Nagadastagiri, Sahithi Rampalli, Makesh Chandran, Gurpreet Kalsi, Sreenivas Subramoney, John Sampson, and Vijaykrishnan Narayanan. "Psb-rnn: A processing-in-memory systolic array architecture using block circulant matrices for

recurrent neural networks." In 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 180-185. IEEE, 2020.