Hardware Acceleration of Biomedical Microwave Techniques using High Level Synthesis

Mohammad Amir Mansoori and Mario R. Casu
Department of Electronics and Telecommunications (DET), Politecnico di Torino, Turin, Italy
{mohammadamir.mansoori, mario.casu}@polito.it

Abstract—Microwave radiations have been proven to be effective in biomedical applications, including brain stroke monitoring. However, microwave algorithms in these applications are computationally expensive with compute-intensive parts that are termed “kernels”. To speed up the medical diagnosis, it is crucial to adopt new methodologies to accelerate the execution of these kernels by using specific hardware solutions. A recent trend in designing hardware accelerators is High Level Synthesis (HLS), which creates an implementation starting from a high-level description (C or C++) of an algorithm. In this paper, we first categorize the recurrent medical microwave techniques and kernels. Then, we propose efficient hardware accelerators for these kernels in programmable devices by using HLS. Several hardware optimization strategies are introduced and their impact on the overall performance is explored. We believe that the analysis of these kernels and their corresponding hardware acceleration techniques can be greatly beneficial to the biomedical microwave research community.

Index Terms—Microwave Imaging, Hardware acceleration, High Level Synthesis, Medical diagnosis.

I. INTRODUCTION

Microwave Imaging (MI) has been used progressively in recent years in various medical applications [1], [2]. Although other medical imaging modalities such as CT-scan or MRI are widely used in the same area, the non-ionizing, low cost, and non-invasive characteristics of MI systems make them a suitable complement to the traditional imaging systems. To obtain the required information about the internal structure of the body tissue, several antennas are used to emit and capture microwave radiations to and from the tissue. When the emitted radiations pass through the body, “scattering” phenomena occur due to the difference between dielectric properties of the body tissues. The final microwave measurements are thus the scattered radiations captured by antennas, represented usually by a scattering matrix [1].

Retrieving original information about the properties of the object under investigation (body tissues) from the microwave measurements requires processing the scattering matrix. There are several algorithms that can be used for this purpose that will be thoroughly explored in the next section. The computational complexity of these algorithms is one of the challenges of MI systems. To speed up their execution, these MI algorithms must be analyzed to determine the compute-intensive parts that are termed kernels. Hardware acceleration of MI kernels can boost their performance, making them suitable for low-cost, real-time, embedded systems.

In this paper, we discuss the computational complexity of relevant biomedical microwave techniques and introduce specialized hardware acceleration methodologies for their kernels. By leveraging the advantages of High level Synthesis (HLS), we apply various optimizations to design accelerators in specialized hardware like Field-Programmable Gate Arrays (FPGAs).

The rest of the paper is as follows. In Sec. II biomedical microwave techniques are introduced. Sec. III describes hardware acceleration methodologies, and Secs. IV-V explain the results and conclusions, respectively.

II. BIOMEDICAL MICROWAVE TECHNIQUES

The general diagram of an MI device for brain stroke imaging is shown in Fig. 1. A similar diagram holds for MI devices used in other applications, like breast cancer imaging. The system consists of several antennas that transmit and receive microwave radiations and are connected to a Vector Network Analyzer (VNA) through a switching matrix. The (VNA) measures these radiations in the form of a scattering matrix. MI can be defined as an “inverse scattering” problem in which the input is the scattering matrix and the output is the required information about the internal structure of the body tissues (normal brain tissues and stroke).

MI algorithms used to solve the inverse scattering problem are usually divided into two categories: Qualitative and Quantitative imaging [3]. Machine Learning (ML) as a new category is recently introduced, which can be also used to learn the previous inverse scattering solutions, other than for specific tasks like classification. In the following, each of these three categories and some of the compute-intensive parts of the corresponding algorithms are explained.
A. Qualitative medical diagnosis

Inverse scattering is inherently a non-linear problem. However, in qualitative imaging, it is approximated as a linear problem. The simplest approach for this linearization is known as Born approximation. In this category, the presence and shape of the anomaly (i.e., brain stroke) is detected.

Linear Sampling method (LSM) [4]–[5], Factorization method [6], Truncated Singular Value Decomposition (TSVD) [7], Time Reversal (TR) techniques such as multiple signal classification (MUSIC) algorithm [8] and Eigenvalue Decomposition (EVD) of the TR operation [9], Beamforming approaches [10] and several Radar-based methods [1] are among these linearized qualitative algorithms. Matrix multiplication is one of the critical parts of these algorithms due to the large data dimensions. Although qualitative methods are not highly accurate, they are well suited for weak scattering objects and low-contrast scenarios. Furthermore, in some algorithms it is possible to obtain quantitative results when such approximation holds.

B. Quantitative image reconstruction

In quantitative imaging, the exact values of image pixels are reconstructed by solving the non-linear inverse scattering problem. These methods are more accurate because they consider the non-linearities of the problem. However, they are computationally more intensive.

Some of the non-linear quantitative algorithms are Contrast Source Inversion [11], inexact Newton methods [12], and DBIM-TwIST [13] which belong to the Microwave Tomography deterministic approaches. In addition, stochastic techniques including Simulated annealing and Genetic Algorithm are among the other quantitative MI methodologies [3].

The iterative nature of these algorithms leads to their high complexity. For example, as shown in Fig. 2, in each iteration of the image reconstruction, a forward solver such as Finite Difference Time Domain (FDTD) is used to update the solution, which is one of the most compute-intensive MI kernels.

C. Machine Learning

In recent years, ML has attracted attention in the field of medical Microwave Imaging. Different kinds of Deep Neural Networks (DNNs) are used to replace the computation of the forward solver or the inverse scattering solution [14]–[18]. Although these solutions can produce quantitative results, they are mainly evaluated on synthetic data because the extraction of training data for each application is not a trivial task. Nevertheless, ML algorithms can also be used for qualitative imaging. In this scenario, two steps are usually required: feature extraction and classification.

For feature extraction, Principal Component Analysis (PCA) [19] is a useful approach, and for classification, there are several algorithms such as Support Vector Machine (SVM) [21], Random Forest (RF) [20], K-Nearest Neighbors (KNN) [22], as well as Neural Networks [23], [24].

III. HARDWARE ACCELERATION METHODOLOGIES

In this section, we describe the methodologies to design hardware accelerators for the compute-intensive MI kernels. Despite being less complex, qualitative algorithms need hardware acceleration when used in low-cost embedded systems. Some of these qualitative kernels have been already considered for hardware acceleration, such as the FPGA accelerator in [8] for MUSIC algorithm. In the following, other MI kernels and their hardware acceleration in FPGA are explored.

A. Matrix multiplication

Matrix multiplication is one of the building blocks for many MI kernels. Specifically, in this work, we focused on the multiplications in form of $A_{[M \times N]}^T \times A_{[N \times M]}$ where $A^T$ is the transpose of $A$, and $M \times N$ are the dimensions of $A$. It is used in the computation of the covariance between two matrices and is also useful in the PCA algorithm.

When the matrix dimensions are large, it is not possible to store all the elements in local memory. On the other hand, data transfer between external and local memories is time-consuming. Therefore, a trade-off must be found between local storage and data transfer. For this trade-off, we adopted a block-streaming methodology as shown in Figs. 3 and 4. The main idea is to partition the input data into several blocks and compute the diagonal and off-diagonal elements of the multiplication separately for each block.

B. EVD and SVD

In several MI algorithms described in the previous section, eigenvalues or singular-values of a large matrix must be calculated.
In Machine Learning algorithms, PCA is used for data dimensionality reduction and feature extraction. It can be used in MI for extracting features from the input scattering matrix by doing $n_s$ measurements (or samples) at $F$ different frequencies. Due to its symmetry, the number of independent elements of the scattering matrix at a specific frequency equals to $N_d = N \times (N + 1)/2$ for a system with $N$ antennas. Therefore, the size of input matrix for PCA algorithm will be $n_s \times N_d \times F$.

PCA consists of several Compute Units (CU) that are required to be accelerated in hardware. These CUs are Mean, Covariance, EVD (or SVD), and Projection computations.

All the CUs are implemented in HLS and are processed in parallel by using the Dataflow optimization technique. Specific HLS-based optimizations, such as array partitioning, function inlining, loop unrolling and pipelining are used in the hardware design to increase the efficiency [25].

In Machine Learning algorithms, PCA is used for dimensionality reduction and feature extraction. It can be used in MI for extracting features from the input scattering matrix obtained by doing $n_s$ measurements (or samples) at $F$ different frequencies. Due to its symmetry, the number of independent elements of the scattering matrix at a specific frequency equals to $N_d = N \times (N + 1)/2$ for a system with $N$ antennas. Therefore, the size of input matrix for PCA algorithm will be $n_s \times N_d \times F$.

PCA consists of several Compute Units (CU) that are required to be accelerated in hardware. These CUs are Mean, Covariance, EVD (or SVD), and Projection computations.

All the CUs are implemented in HLS and are processed in parallel by using the Dataflow optimization technique. Specific HLS-based optimizations, such as array partitioning, function inlining, loop unrolling and pipelining are used in the hardware design to increase the efficiency [25].

In addition to the traditional ML classifier such as SVM, Decision Tree (DT), RF, and KNN, recently there has been a growing interest in using DNNs for the inverse scattering problem in MI. Due to the high non-linearity of the problem, a proper training step is required for the selected model, calling for a sufficiently large dataset.

Creating the training MI dataset for the inverse scattering problem requires collecting the scattering data from various dielectric profiles and can be done synthetically or by real measurements, each of which is time-consuming. Generation of additional training data from a few MI measurements (real or synthetic) can boost the performance of an ML model. Generative Adversarial Networks (GANs) are a great addition, we considered the impact of dispersive materials and polarization currents on the electric fields as opposed to the traditional ML classifier such as SVM, Decision Tree (DT), RF, and KNN, recently there has been a growing interest in using DNNs for the inverse scattering problem in MI. Due to the high non-linearity of the problem, a proper training step is required for the selected model, calling for a sufficiently large dataset.

Creating the training MI dataset for the inverse scattering problem requires collecting the scattering data from various dielectric profiles and can be done synthetically or by real measurements, each of which is time-consuming. Generation of additional training data from a few MI measurements (real or synthetic) can boost the performance of an ML model. Generative Adversarial Networks (GANs) are a great option to enhance the training dataset.
TABLE I
3D FDTD processing time in different hardware platforms:
CPU= Intel Xeon gold 5120, GPU (Acceleware) = Tesla P40, FPGA = Ultrascale+

| max Ant | Time(s/Ant) |
|---------|-------------|
| CPU     | 24.97       |
| GPU     | 3.64        |
| Single FPGA Large | 4.3         |
| Single FPGA Small  | 3.38        |
| Multi-FPGA Small  | 11.42       |

candidate for the generation of training dataset. They consist of two competing networks that learn to produce realistic training data samples.

The application of Machine Learning in Microwave Imaging and inverse scattering has been introduced in recent years and there are several challenges in the design of these ML models which are still under the research. In the future, we plan to further explore these ML models, their computational complexity, and their hardware acceleration methodologies.

IV. RESULTS

In this section, we demonstrate the advantage of using hardware acceleration for MI kernels by evaluating some of the biomedical microwave techniques in terms of hardware performance. For hardware implementation in FPGA, we used Vivado 2018.2 tool.

A. 3D FDTD accelerator

For the evaluation of FDTD algorithm, we used a synthetic MI dataset [26] that is a 3D model of the glycerol-water mixture that represents the dielectric characteristics of the human brain. We compared the performance of our FPGA accelerator for FDTD kernel with CPU and GPU designs. Compared to an Intel CPU (Xeon gold 5120) and a commercial GPU implementation (Acceleware), we could achieve $7.3 \times$ and $1.67 \times$ improvement in the processing time per antenna, respectively, in FPGA. In addition, we extracted system-level performance results for the non-linear MI reconstruction algorithm shown in Fig. 2. Table I shows a summary of the results for our single-FPGA design in addition to our multi-FPGA accelerator with 8 FPGAs for 24 antennas, and a total dimension of $70 \times 70 \times 70$.

B. PCA accelerator

Evaluation of the FPGA accelerator for PCA is useful not only for ML-based algorithms in MI, but also for other linear or qualitative MI kernels such as TSVD because it consists of several widely-used compute units including EVD and SVD. We compared the processing time and resource usage for each compute unit of PCA on a large Virtex FPGA (VC709U evaluation board) that is shown in Fig. 6. We found that the main dimension of PCA input (the number of independent elements of scattering matrix) can have a maximum value of 300 ($N_d = 300$) in our target FPGA, that corresponds to an MI system with 24 antennas.

By using the block-streaming strategy described in section III-A, the Covariance computation obtains a low latency, and the most critical part of the design is SVD as can be seen from Fig. 6, for which the optimized HLS built-in function is used.

C. MLP accelerator

To design and evaluate the MLP accelerator, we used hls4ml [28] that is a tool to convert recurrent ML models from Python to a synthesizable code that can be used in Vivado HLS. We used a dataset of microwave measurements that contain 4500 samples of scattering matrix with 462 elements. It consists of 9 classes representing the presence, type, and location of the brain stroke. For feature extraction, PCA is used that results in the reduction of features to 110. We selected an MLP with 3 hidden layers. The numbers of neurons per layer are 220, 64, 64, respectively. The target device is a Zynq SoC (ZedBoard), and we used fixed-point precision for the hardware implementation by leveraging hls4ml. The accuracy before and after hardware implementation, resource usage, and processing time are depicted in Table II and Figs. 7 and 8. Note the negligible accuracy loss in the hardware accelerator due to the reduced precision.

V. CONCLUSIONS

In this paper, we presented a brief overview of hardware-based methodologies to improve the performance of recurrent biomedical microwave techniques. First, we categorized different MI algorithms and introduced their compute-intensive parts, termed kernels. Then, we presented specific hardware accelerators in FPGAs for each kernel to obtain the desired performance. Evaluation of the results for three different kernels (FDTD, PCA, MLP) shows the advantage of using hardware accelerators in MI. Specifically, we could achieve a maximum of $13.4 \times$ improvement for 3D-FDTD in the processing time per antenna compared to the commercial GPU design (Acceleware) by using FPGA accelerator. For PCA, we
could optimize the performance of the compute units including SVD and Covariance computation, and for MLP, we could use a low-cost Zynq SoC for the detection and diagnosis of brain stroke, that is useful in real-time embedded systems.

In the future, more kernels will be added to the library of microwave algorithms and efficient hardware accelerators will be designed for them. Specifically, SVM classifier and Deep neural networks including GANs are considered as the future research.

ACKNOWLEDGMENT

This work was supported by the EMERALD project funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764479.

REFERENCES

[1] R. C. Conceição et al., An Introduction to Microwave Imaging for Breast Cancer Detection, (Biological and Medical Physics, Biomedical Engineering), 1st ed. Cham, Switzerland: Springer, 2016.
[2] D. O. Rodriguez-Duarte, J. A. T. Vasquez, R. Scapaticci, L. Crocco and F. Vipiana, “Assessing a Microwave Imaging System for Brain Stroke Monitoring via High Fidelity Numerical Modelling,” in IEEE Journal of Electromagnetics, RF and Microwave in Medicine and Biology, vol. 5, no. 3, pp. 238-245, Sept. 2021
[3] M. Pastorino, Microwave Imaging, Hoboken, NJ: Wiley, 2010.
[4] F. Cakoni, D. Colton, and P. Monk, The Linear Sampling Method in Inverse Electromagnetic Scattering, Philadelphia, PA, USA: SIAM, 2011
[5] M. Brignone, G. Bozza, A. Randazzo, M. Piana and M. Pastorino, “A Hybrid Approach to 3D Microwave Imaging by Using Linear Sampling and ACO,” in IEEE Transactions on Antennas and Propagation, vol. 56, no. 10, pp. 3224-3232, Oct. 2008
[6] M. N. Akcnc et al., “Qualitative Microwave Imaging With Scattering Parameters Measurements,” in IEEE Transactions on Microwave Theory and Techniques, vol. 63, no. 9, pp. 2730-2740, Sept. 2015
[7] J. D. Shea, B. D. Van Veen and S. C. Hagness, “A TSV Analysis of Microwave Inverse Scattering for Breast Imaging,” in IEEE Transactions on Biomedical Engineering, vol. 59, no. 4, pp. 936-945, April 2012
[8] M. R. Casu et al., “A COTS-Based Microwave Imaging System for Breast-Cancer Detection,” in IEEE Transactions on Biomedical Circuits and Systems, vol. 11, no. 4, pp. 804-814, Aug. 2017
[9] M. D. Hossain and A. S. Mohan, “Detection in Highly Dense Breasts Using Coherently Focused Time-Resolved Microwave Imaging,” in IEEE Transactions on Computational Imaging, vol. 3, no. 4, pp. 928-939, Dec. 2017.
[10] M. A. Aldhaeabi et al., “Review of Microwaves Techniques for Breast Cancer Detection,” Sensors, 2020, 20(8):2390.
[11] C. Gilmore et al., “Microwave Biomedical Data Inversion Using the Finite-Difference Contrast Source Inversion Method,” in IEEE Transactions on Antennas and Propagation, vol. 57, no. 5, pp. 1528-1538, 2009.
[12] I. Bisio et al., “Brain Stroke Microwave Imaging by Means of a Newton-Conjugate-Gradient Method in LPM Random Spaces,” in IEEE Transactions on Microwave Theory and Techniques, vol. 66, no. 8, pp. 3668-3682, 2018.
[13] Z. Miao and P. Kosmas, “Multiple-frequency DBIM-TwIST algorithm for microwave breast imaging,” in IEEE Transactions on Antennas and Propagation, vol. 65, no. 5, pp. 2507-2516, May 2017.
[14] Z. Ma, K. Xu, R. Song, C.-F. Wang and X. Chen, “Learning-Based Fast Electromagnetic Scattering Solver Through Generative Adversarial Network,” in IEEE Transactions on Antennas and Propagation, vol. 69, no. 4, pp. 2194-2208, April 2021.
[15] R. Song et al., “Electromagnetic Inverse Scattering With Perceptual Generative Adversarial Networks,” in IEEE Transactions on Computational Imaging, vol. 7, pp. 689-699, 2021.
[16] L. Li et al., “DeepNIS: Deep Neural Network for Nonlinear Electromagnetic Inverse Scattering,” in IEEE Transactions on Antennas and Propagation, vol. 67, no. 3, pp. 1819-1825, March 2019.
[17] H. M. Yao, W. E. I. Sha and L. Jiang, “Two-Step Enhanced Deep Learning Approach for Electromagnetic Inverse Scattering Problems,” in IEEE Antennas and Wireless Propagation Letters, vol. 18, no. 11, pp. 2254-2258, Nov. 2019.
[18] M. Ambrosanio, S. Franceschini, F. Baseline and V. Pascasio, “Machine Learning For Microwave Imaging,” 2020 14th European Conference on Antennas and Propagation (EuCAP), 2020, pp. 1-4.
[19] M. A. Aldhaeabi, S. Bamatraf et al., “Near-field Electrically Small Sensors Array with PCA for Microwave Breast Tumor Detection,” 2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, 2019, pp. 1007-1008.
[20] B.L. Oliveira, D. Godinho et al., “Diagnosing Breast Cancer with Microwave Technology: Remaining Challenges and Potential Solutions with Machine Learning,” Diagnostics, vol. 8, no. 2, p. 36, 2018.
[21] L. Guo and A. Abbosh, “Stroke Localization and Classification using Microwave Tomophony with k-means Clustering and Support Vector Machine,” Biomedical Engineering, vol. 39, no. 4, pp. 312–324, 2018.
[22] E. A. Aydan and M. K. Keles, “Breast Cancer Detection using k-nearest Neighbors Data mining Method Obtained from the Bow-Tie Antenna Dataset,” International Journal of RF and Microwave Computer-Aided Engineering, vol. 27, no. 6, 2017.
[23] R.C. Conceição, H. Medeiros et al., “Classification of breast tumor models with a prototype microwave imaging system,” Journal of Applied Clinical Medical Physics, vol. 47, no. 4, pp. 1860-1870, 2020.
[24] S.P. Rana, M. Dey et al., “Machine Learning Approaches for Automated Lesion Detection in Microwave Breast Imaging Clinical Data,” Sci. Rep., vol. 9, no. 1, pp. 1-12, 2019.
[25] M. A. Mansoori and M. R. Casu, “High Level Design of a Flexible PCA Hardware Accelerator Using a New Block-Streaming Method,” Electronics, vol. 9, no. 3, p. 449, 2020.
[26] M. A. Mansoori, P. Lu and M. R. Casu, “FPGA Acceleration of 3D FDTD for Multi-Antennas Microwave Imaging Using HSL,” in IEEE Access, vol. 9, pp. 122696-122711, 2021.
[27] T. Reimer, J. Sacristan and S. Pistorius, “Improving the Diagnostic Capability of Microwave Radar Imaging Systems using Machine Learning,” 2019 13th European Conference on Antennas and Propagation (EuCAP), pp. 1-5, 2019.
[28] J. Duarte et al., “Fast inference of deep neural networks in FPGAs for particle physics”, JINST 13 P07027 (2018), arXiv:1804.06913.