FinJeHeW Finnish Journal of eHealth and eWelfare

SCIENTIFIC PAPERS

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Local edge computing for radiological image reconstruction and computer-assisted detection: A feasibility study

Antti Isosalo[1], Johirul Islam[2], Henrik Mustonen[1], Ella Räinä[1], Satu I. Inkinen[1], Mikael Brix[1,3], Tanesh Kumar[2], Jarmo Reponen[1], Miika T. Nieminen[1,3], Erkki Harjula[2]

[1] Research Unit of Health Sciences and Technology, University of Oulu, Oulu, Finland; [2] Centre for Wireless Communications – Networks and Systems, University of Oulu, Oulu, Finland; [3] Department of Diagnostic Radiology, Oulu University Hospital, Oulu, Finland

Antti Isosalo, Research Unit of Health Sciences and Technology, University of Oulu, P.O. Box 5000, FI-90014 University of Oulu, FINLAND. Email: antti.isosalo@oulu.fi.

Abstract

Computational requirements for data processing at different stages of the radiology value chain are increasing. Cone beam computed tomography (CBCT) is a diagnostic imaging technique used in dental and extremity imaging, involving a highly demanding image reconstruction task. In turn, artificial intelligence (AI) assisted diagnostics are becoming increasingly popular, thus increasing the use of computation resources. Furthermore, the need for fully independent imaging units outside radiology departments and with remotely performed diagnostics emphasize the need for wireless connectivity between the imaging unit and hospital infrastructure. In this feasibility study, we propose an approach based on a distributed edge-cloud computing platform, consisting of small-scale local edge nodes, edge servers with traditional cloud resources to perform data processing tasks in radiology. We are interested in the use of local computing resources with Graphics Processing Units (GPUs), in our case Jetson Xavier NX, for hosting the algorithms for two use-cases, namely image reconstruction in cone beam computed tomography and AI-assisted cancer detection from mammographic images. Particularly, we wanted to determine the technical requirements for local edge computing platform for these two tasks and whether CBCT image reconstruction and breast cancer detection tasks are possible in a diagnostically acceptable time frame. We validated the use-cases and the proposed edge computing platform in two stages. First, the algorithms were validated use-case-wise by comparing the computing performance of the edge nodes against a reference setup (regular workstation). Second, we performed qualitative evaluation on the edge computing platform by running the algorithms as nanoservices. Our results, obtained through real-life prototyping, indicate that it is possible and technically feasible to run both reconstruction and AI-assisted image analysis functions in a diagnostically acceptable computing time. Furthermore, based on the qualitative evaluation, we confirmed that the local edge computing capacity can be scaled up and down during runtime by adding or removing edge devices without the need for manual reconfigurations. We also found all previously implemented software components to be transferable as such. Overall, the results are promising and help in developing future applications, e.g., in mobile imaging scenarios, where such a platform is beneficial.

Keywords: radiology, image processing, health technology, telemedicine, cloud computing, internet of things

Introduction

Background

Modern medical imaging devices produce substantial amounts of data for clinical decision-making. For example, the use of diagnostic imaging has been steadily increasing. A 78% increase in computed tomography (CT) examinations and a growing trend of 12% in mammography imaging are reported for Europe between 2004 and 2011 [1]. This growing trend also imposes a higher computational load for post-imaging operations, such as image reconstruction and image processing. Post-imaging operations are usually performed with device-manufacturer provided hardware and the resulting images are then sent to the picture archiving and communication system (PACS). Complementary image assessment, such as computer-assisted detection (CAD), can be carried out with proprietary applications on diagnostic workstations or servers with PACS access. The applications enable clinical assessment for medical imaging outcomes and are often manufacturer neutral for supporting e.g., dental imaging applications [2] and CAD systems for breast imaging [3].

According to literature, considerable efforts are usually required when integrating artificial intelligence (AI) solutions, e.g., decision support systems, to complement the well-established clinical workflow [4]. Modern AI solutions for image processing and computer-aided diagnostics typically require a computational platform with parallel processing capabilities and industrial-grade Graphics Processing Units (GPUs). Commercial solutions, which are operated locally within hospital premises, are usually offered as expensive high-end computing devices (single units) which may have limited capabilities for extensive workload, e.g., images routed from a larger catchment area or hospital district. Cloud services, on the other hand, provide high computational scalability, but come with a high burden on the network and, consequently, higher energy consumption [5].

Edge computing, however, pushes computing tasks from the network core to the network edge, closer to the data sources [6]. Therefore, it helps improve performance and network failure tolerance while reducing network traffic, which are crucial qualities with low-capacity and/or unreliable uplink connections [5,7]. With the help of edge computing, many key e-health functions can be provided regardless of the low-quality and unreliable connection to centralized servers, enabling, e.g., remote areas to obtain real-time medical diagnoses [8]. Edge computing could potentially be a viable solution to remove the need, e.g., for dedicated reconstruction units for each imaging system by making post-imaging operations cloud-native. This avoids the above-mentioned pitfalls of centralized cloud computing.

In this study, we explore the feasibility of using edge computing for post-imaging operations, namely image reconstruction in cone-beam computed tomography (CBCT) and AI-assisted image analysis in breast cancer detection and mammography. Our goal is to determine the minimal technical requirements for local edge computing platform for hosting these operations and to assess if the operations can be performed in a diagnostically acceptable time frame.

Related work

Currently, only a few edge computing studies exist for medical image processing and analysis. The capabilities of edge computing have mostly been investigated in the remote health monitoring setting [5]. The literature on utilizing edge computing for the use cases chosen for this feasibility study, i.e., 1) Cone-beam computed tomography reconstruction and 2) AI-based breast cancer detection, *is scarce.*

Computed tomography image reconstruction**:** To our knowledge, no studies yet exist for investigation of locally deployed edge computing in CBCT reconstruction, which is particularly challenging to the vast size of raw data. An edge computation-based CT reconstruction has been the topic of a few studies. Chen et al. developed an edge-based multi-thread reconstruction Gridrec algorithm for high-resolution synchrotron radiation CT. They utilized low-cost GPU edge devices, which enabled a remarkable speedup of elevenfold compared to serial Gridrec algorithm [9]. In the study by Zhang et al., a 3-D reconstruction method for medical CT images utilizing deep learning (DL) was developed using edge computing infrastructure [8].

AI-based medical image assessment**:** Previous work exploring edge computing for breast cancer assessment using mammography image data could not be found. However, there are few papers describing the study of enabling technologies such as model optimization and hardware acceleration [10] and a blockchain-enabled learning model [11] with application examples in the medical imaging domain. Algorithms with low computational cost for medical image analysis have also been proposed [12]. Two detection and analysis studies in the edge computing context have been conducted for colonic neoplasia localization [13], and automated analysis of colonoscopies [14]. The latter utilized a Jetson Xavier NX development microsystem, demonstrating that it can host a real-time detection application in terms of computing power and speed.

Aims

The aim of this study was to investigate the feasibility of local GPU-based edge computing technology on image reconstruction and automated assessment of medical images. For this, the traditional, well established, workflow (Figure 1a) is proposed to be augmented with distributed computing functionalities (Figure 1b). For this proof-of-concept study, we chose two very different use-cases: 1) image reconstruction in volumetric CBCT, and 2) AI-assisted breast cancer detection from full-field digital mammograms (FFDM) using a pre-trained DL model, both tasks being computationally demanding. Thus, our research question is formulated as follows: Is the use of locally deployed edge computing a clinically feasible solution for the two use-cases? We are particularly interested in whether this technology can perform image reconstruction and image assessment in a diagnostically acceptable time frame.

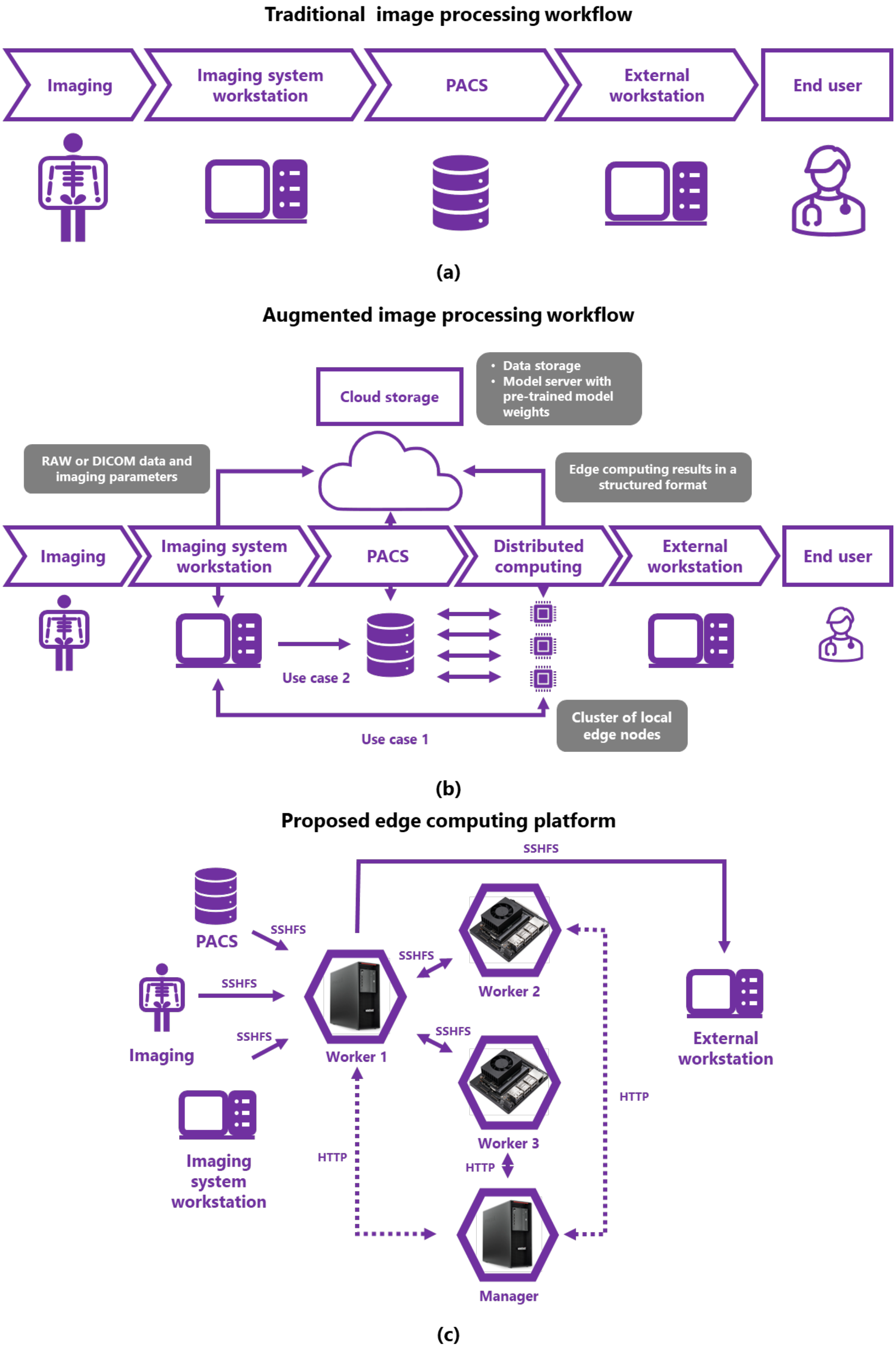

Figure 1. Illustration of (a) a traditional clinical imaging workflow, (b) an augmented image processing workflow utilizing local edge computation, and (c) proposed edge computing platform and its end-to-

end connections. The augmented workflow in subfigure b differs for our use-cases. In use-case 1, namely the cone-beam computed tomography (CBCT) reconstruction, we propose a path where projections are extracted from the scanner and retrospectively reconstructed utilizing a cluster of local edge nodes. In use-case 2, namely the AI-assisted breast cancer detection, the data may originate directly from the imaging device or from the Picture Archiving and Communication System (PACS). Imaging parameters and for example trained deep learning model weights are shared from cloud storage, where also edge computing results are sent. Cloud storage is illustrated as a single cloud but can consist of several servers for different purposes. In subfigure c, medical imaging data is transferred through a secure file-sharing protocol (SSHFS). Connections between manager and workers are made through hypertext transfer protocol (HTTP) during the deployment. In this work, we have concentrated on the local tier system of the three-tier architecture of the edge computing. External workstation in subfigure c serves as an interface for viewing functionalities.

Material and methods

Use-case 1: Image reconstruction in volumetric CBCT

Data from a diagnostic dental CBCT scanner (Promax 3D Max, Planmeca Group, Helsinki, Finland) was used to address the performance of local edge computing for CBCT reconstruction. An anthropomorphic head X-ray phantom with cervical vertebrae (Erler-Zimmer GmbH & Co.KG) was scanned with a clinical protocol used for the assessment of teeth. During acquisition, the scanner collects 400 projection images with dimensions of 319-by-736 pixels (height-by-width) from a 210° angular range. These measured cone-beam projections were extracted from the scanner and retrospectively reconstructed with the Python ASTRA tomography toolbox (v1.8, imec-Vision Lab, University of Antwerp, CWI, Amsterdam, the Netherlands) [15,16]. The traditionally utilized, non-iterative Feldkamp-Davis-Kress (FDK) algorithm was used for reconstruction. The reconstructed slices were 1012-by-612 pixels in size, and the number of slices was modified (10, 50, 100, 200) to address the influence of increasing reconstruction volume on the total computation time. Thus, the reconstructed volumes were 1012-by-612-by-10, 1012-by-612-by-50, 1012-by-612-by-100, and 1012-by-612-by-200 -voxel matrices with an isotropic pixel size of 200 µm. To increase statistical power, the retrospective reconstruction was repeated ten times for each of the volumes.

The local edge computing experiments were conducted on a Jetson Xavier NX microsystem (NVIDIA, Santa Clara, California, US) with Carmel Arm processor and 384-core Volta architecture GPU (NVIDIA) with 48 Tensor Cores and 8 GB of GPU memory. The operating system used was Linux for Tegra (version 32.4.2). The reconstructions were validated against a local reconstruction computer (reference system) with Intel Xeon E5-1620 processor and 3840-core Pascal architecture Quadro P6000 GPU (NVIDIA) with 24 GB of GPU memory. The operating system was Microsoft Windows (version 6.1.7601).

Use case 2: AI-assisted breast cancer detection

In the breast cancer detection use-case, data from the open Portuguese FFDM dataset [17] was used. The dataset comprises of 86 digital mammography examinations with 4 standard views, specifically left and right mediolateral oblique (MLO) and bilateral craniocaudal (CC) views, with resolution 4084-by-3328 and 3328-by-2560 (height-by-width) respectively with 14-bit contrast resolution. The capturing device detector pixel size was 70 microns [17]. The dataset has examples of masses, calcifications, and architectural distortions. Moreover, we made use of an existing deep learning model, namely Globally-Aware Multiple Instance Classifier (GMIC) model [18] with model weights, trained on the NYU Breast Cancer Screening Dataset [19]. The model operates at high resolution and simultaneously predicts the presence of benign and malignant findings from 2944-by-1920 (height-by-width) pixel matrix. As a pre-processing step the FFDM MLO and CC views were cropped to this size to remove uninformative background pixels and meet the model requirements for input data. The model consists of three modules: global module processing the previously mentioned cropped input image, a local module processing image patches resulting from the global model via region of interest (ROI) extraction and finally a fusion module. In the first global network module, a feature map is computed, and two saliency maps are extracted to obtain information on the benign and malignant lesion locations. In the local module the extracted ROI patches are further characterized by a higher-capacity local network. Lastly, a fusion network is used to combine the information from the global and the local modules of the model for final prediction. The process can be seen as comparable to that of a radiologist who considers both global and local information to assess the mammograms for suspicious findings [18].

The computation performance was assessed by running the model inference 100 times for a single examination with four digital mammograms and calculating an average computing time for one examination. For the experiments, PyTorch (1.8.0) framework [20] was used in implementing the model. Moreover, the size of the file containing the learnable parameters of the GMIC model was approximately 60 MB.

A local edge computing node Jetson Xavier NX microsystem, with equivalent specifications as in the first use-case, was used to conduct the experiments. The experiments were validated against a high-end workstation (reference system) with Ryzen Threadripper processor and 4,608-core Turing architecture Titan RTX GPU (NVIDIA) with 576 Tensor Cores and 24 GB of GPU memory. The operating system was Ubuntu (version 20.04 LTS).

Proposed edge computing platform

To evaluate the feasibility of the augmented workflow (Figure 1b), the studied algorithms, namely CBCT reconstruction and breast cancer evaluation, were implemented as virtualized Docker-based nanoservices [6,21] that can be deployed on a swarm of local edge nodes. The implementation comprises of several nanoservices to perform the functions of specific workflow components. A storage service was used to store the imaging results, i.e., CBCT projections and FFDM, and later, the edge computing results to the Hard Disk Drive (to be further transferred to the PACS). CBCT reconstruction and breast cancer detection methods had their dedicated services, which were included in customized Docker container images with their software dependencies. These nanoservices were deployed in a cluster of local edge nodes, namely computing nodes known as workers and a cluster head known as the manager node (Figure 1c and Table 1). In the experiment, storage (server) service was deployed on an Ubuntu 20.04-based Intel Core i7 with 8 GB of RAM (Worker 1). Furthermore, the CBCT reconstruction and breast cancer evaluation services were deployed on two separate GPU-based Jetson Xavier's (Worker 2 and 3).

Table 1. Hardware and software resources of the edge computing platform utilized in the feasibility study.

	Hardware Resources				**Software Resources**			
	CPU	GPU	Memory	Storage	Operating System	Kernel	Docker	Docker Compose
Storage Server (Worker 1)	Intel Core i7 1.6 GHz (4 CPU cores)		8 GB	250 GB	Ubuntu 20.04	5.15.0-46-genetic	20.10.12	-
Breast Cancer Detection (Worker 2) CBCT Image Reconstruction (Worker 3)	NVIDIA Car-mel ARMv8.2 (6 CPU cores)	384 NVIDIA CUDA cores & 48 Tensor cores	8 GB	128 GB	Ubuntu 18.04	4.9.201-tegra (L4T 32.5.2 with Jetpack 4.5.1)	20.10.18	-
Orchestrator (Manager)	Intel Core i5 1.6 GHz (4 CPU cores)	-	16 GB	500 GB	Ubuntu 22.04	5.15.0-48-generic	20.10.12	1.29.2

These containerized nanoservices were deployed among the workers and orchestrated by Ubuntu 22.04-based Intel Core i5 with 16 GB of RAM (Manager). Furthermore, the storage server was built on top of Secure Shell (SSH) protocol and shared the data with other services (e.g., computing resource descriptions) through a secure file-sharing protocol (SSHFS). Mobile broadband was used for communication. Arbitrary event-driven software was used to monitor filesystem events, e.g., new data placed on the storage server.

Results

The validation of the use cases and the proposed framework was conducted in two stages. In the first stage, the algorithms were validated against a reference setup, namely against results from running the computations on a regular workstation. In this stage, the edge computing platform was not yet utilized, only the Jetson Xavier's were employed in the algorithmic validation. In the second stage, the edge computing platform was evaluated qualitatively in different scenarios by running the previously validated algorithms as nanoservices.

Algorithmic validation of CBCT reconstruction

According to our measurements, the edge device reduced the computation times for the FDK algorithm for large reconstruction volumes, i.e., when the number of slices was 50 or more (Table 2a). For a smaller reconstruction volume with ten slices, the local reconstruction PC was 42.44 % faster. The longest reconstruction times for the local workstation and edge device were 22.91 s and 10.27 s, respectively. The CBCT reconstruction required 3.4 GB GPU memory and 3.0-3.1 GB RAM memory on both devices.

Table 2. a) Cone-beam computed tomography reconstruction times for the local reconstruction workstation and the edge-device with algorithm parameters already uploaded from the cloud storage, and b) breast cancer evaluation running times for the local workstation and the edge-device with the model weights already uploaded from the cloud storage.

a) Computation times for the Feldkamp-Davis-Kress (FDK) algorithm (mean ± standard deviation, N=10)			
Number of slices	Local workstation (s)	Edge device (s)	Percentage difference (%)
10	1.72 ± 0.01	2.45 ± 0.03	42.44
50	4.84 ± 0.21	4.44 ± 0.09	-8.26
100	13.29 ± 2.55	6.98 ± 0.13	-47.50
200	22.91 ± 4.58	10.27 ± 0.43	-55.17

b) Inferring a single examination (with four standard views, namely left and right mediolateral oblique and bilateral craniocaudal views) repeatedly using the Globally-Aware Multiple Instance Classifier (GMIC) model (mean ± standard deviation, N=100)			
Number of images	Local workstation (s)	Edge device (s)	Percentage difference (%)
4	3.93 ± 1.07	6.80 ± 2.11	53.37

Algorithmic validation of breast cancer detection

Performing the breast cancer detection on the edge device added the processing time by 1.73 times on average as compared to a local workstation (Table 2b). Average runtimes were 3.93±1.07 s and 6.80±2.11 s for the local workstation and edge device, respectively. Furthermore, the first repetition was costly, taking 7.32 s and 27.61 s with the local workstation and edge device, respectively. In addition, there was an additional time spent on initializations, taking 1.03 s with the local workstation and 10.13 s with the edge device. The initializations, and operationalizing of the model, i.e., the inference, on average totals 4.96 s for the local workstation and 16.92 s for the edge device. Furthermore, running the GMIC model inference on both devices reserved 1.4-1.6 GB of GPU memory and 4.2-4.4 GB of RAM memory.

Integration test and qualitative evaluation of the platform

The experimental local edge setup was qualitatively evaluated by first-hand observation in an integration test utilizing the proposed setting described in the Materials and Methods -section. Placing new data into the storage server, was observed to automatically start the use case agnostic (e.g., data upload) and use case specific (e.g., breast cancer evaluation) nanoservices. Furthermore, plugging one of the nodes to the power supply resulted in a pending nanoservice to move to execution (or to be available to be executed on on-demand basis), and on the other hand, unplugging a node from its power supply did not interfere the main service, embodying the characteristics of resilience, required, for example, in mobile imaging scenarios, where the composition of the local computational cluster may be subject to a change. Some overhead was experienced in the interplay of the nanoservices. The integration test also gave better understanding of the interplay of different services required for post-processing medical imaging results.

Discussion

Main findings

Algorithmic validation: This feasibility study demonstrated that the utilized edge nodes had enough computing resources for CBCT reconstruction and AI-based image analysis utilizing low-end GPU of the Jetson Xavier NX microsystem for performing computations in a diagnostically viable time frame. In the breast cancer evaluation context, the initializations and operationalizing of the model on average totaled 16.92 s which is a clinically suitable time. For comparison, 146 s unaided reading time per examination has been reported for a human reader [22]. Performing inference separately for each of the images in the examination on a different process can increase throughput but will add time spent on the initializations. In the reconstruction context, the longest reconstruction time of 10.27 s on the edge device for the largest reconstruction volume would be clinically feasible without disturbing the diagnostic workflow as the literature suggests an acceptable time to be in the 3 min window [23]. The Jetson Xavier also exhibited less variation in reconstruction times between repetitions in comparison to the reference system. The statistical differences between the different Jetson Xavier's in our disposal were not evaluated while it is known that slight performance differences may exist [24].

Integration test: In the integration test stage, mobile network caused some overhead, but otherwise, based on the qualitative evaluation, the experiment was successful. Our use-cases bound us to add only devices with the specific hardware (e.g., GPU) and software resources to the swarm, and therefore the composition of the computational cluster cannot be completely arbitrary.

Considerations related to materials and methods

In this study, we focused on CBCT reconstruction, which is due to the cone-shaped X-ray radiation beam, a three-dimensional image reconstruction problem. The proposed platform is not limited only to CBCT reconstruction but could be utilized also for CT reconstruction. Furthermore, CT is a substantially more frequent examination than a CBCT, with 649,119 and 19,442 examinations during the year 2021 in Finland, respectively [25]. This makes it appealing to address the utility of the developed platform for CT reconstruction. Comparison of the computational requirements is not straightforward, as this depends on the size of the detector and the number of projections collected. However, in both cases there benefits from GPU powered computations [26].

In the use-case 2, our focus was in performing inference on the edge. The application itself could have been related to any medical imaging modality, but we choose breast cancer detection from FFDM as out of different modalities, mammography retains the highest spatial resolution and mammograms are very large pixel matrices, thus particularly good for testing how well the utilized edge nodes can cope with the task. Moreover, the platform can be extended to perform similar computation also, in the same context, for Digital Breast Tomosynthesis (DBT) [27] examinations. An applicable deep learning model already exists [28]. There were 292,486 mammography screening examinations during the year 2021 in Finland [25], and the amount of DBT examinations is known to be increasing (e.g., [27]).

Advantages

In the past, porting software between platforms often required reimplementation. In our study, all the previously implemented software components were transferable as such, only requiring building the utilized software libraries for the device to make them compatible with the architecture, which is a very standard procedure. Though other computer vision hardware accelerators, e.g., Field-Programmable Gate Arrays (FPGAs) and Application Specific Integrated Circuits (ASICs) exist, CPU and GPU-based solutions have been widely adopted, as the availability of optimized software components, for one, enables fast prototyping [29].

On the other hand, the choice of architecture in this feasibility study does not limit us in any way. The capacity of the proposed system can be easily scaled up by simply including additional edge devices of choice in the local computing cluster. Dividing the computation into parallel subtasks would potentially maximize the benefits of a multi-node local cluster. The ASTRA functionalities are well optimized, already benefiting from distributed GPU computing, e.g., message-passing interface framework implementation of Simultaneous Iterative Reconstruction Technique (SIRT) [30]. In the breast cancer use case, we could further speed up the inference by converting our DL models to TensorRT [31] to better utilize the CUDA computing capabilities. Such speed-up would be beneficial, for example, if we were to run the inference for all the data accumulated to a hospital PACS, though there would be a need to do this so that the strain on the network would not harm the everyday operations in the hospital environment.

Overall, processing the data near where it is generated allows lesser computational needs for individual imaging units and resulting in an increased overall throughput of the system. This has a connection to the long-term cost-efficiency of edge computing solutions [32]. However, taking use of edge computing also introduces costs. Some of the specialized hardware can be expensive. Supplemental hardware maintenance might also be required, although the maintenance of dedicated software can be managed in a centralized fashion, which will effectively reduce overall costs and added security. It should be noted that the imaging equipment vendor provided single unit of the traditional workflow also cost. In the augmented workflow the only GPU powered devices in addition to the shared edge nodes would be the workstations meant for viewing purposes.

Limitations

The edge computing platform built for this feasibility study used mobile broadband for communication. This might not be realistic for all imaginable scenarios. Updating the learnable parameters of a deep learning model or the Docker image would benefit from better connections. However, a realistic test network was outside the scope of this study. Moreover, we limited our focus only to a local edge computing tier. In the future, we want to expand our assessment to fully utilize the three-tier edge-cloud architecture, where the experimented local edge is just one part of it. An intermediate step towards that goal would be to build a test environment, for example, from acquisition to reconstruction, and communicating with virtual PACS.

It should be noted that the proposed augmentation to the reconstruction workflow cannot be implemented as such, unless the imaging equipment manufacturers permit access to raw data which is often proprietary information. Nonetheless, we have examples, e.g., in the Nuclear Medicine domain, where image reconstructions can be performed on raw data extractable from the system. For the CBCT computing, a paradigm change would be required. For the FFDMs and breast cancer evaluation use-case, such technology could be more readily employed as the communication is between the PACS and edge node (storage service).

Challenges

Security and privacy are still considered as one of the major obstacles in the development of edge-empowered healthcare systems. Some of the security risks that we have identified include, but are not limited to, faulty or compromised edge devices, edge services, communication and networks infrastructure, and network protocols. Since some of the nodes/devices at local and edge tiers can be resource-constrained, traditional cloud-based security and privacy solutions might not be completely applicable. More lightweight security mechanisms should be adopted to address the complex security demands of resource-limited devices. A good summary of the current research in the field of edge computing security can be found in [33].

As edge devices store and process medical data, they must be hardened as any other health information system to resist cyberattacks to protect sensitive information. Among the malicious attacks envisioned in the literature regarding this particular subfield are attacks intended to affect image quality and patient safety, and patient data tampering [34,35].

Among the best practices when deploying edge computing and alleviating various security risks in any environment are sending as few sensitive data items as possible outside the local computational environment, the use of secure network connections, encrypting sensitive data, strong authentication mechanisms, scrutiny in software maintenance, and maintaining security copies of the vital data items.

Edge-computing does not bring only security challenges but also offers many opportunities. With local edge-based solutions we can have better control on how widely the sensitive patient data needs to be propagated. This applies, for example, to the situation where we perform inference on the local edge, such as described in the use-case 2. The process can be carried out by keeping the inferred imaging data local and only the DL pretrained model weights are communicated between the cloud storage and the worker node (Figure 1).

Conclusions and future work

This article studied the feasibility of local edge-based post-imaging operations, namely CBCT image reconstruction and AI-assisted image analysis, with a goal to assess if these operations can be performed in a diagnostically acceptable time frame, and to resolve the minimal technical requirements for a local edge computing platform. In our edge solution, the centric algorithms were encapsulated into virtualized nanoservices fitting to a set of pocket-sized edge nodes. Our results, obtained through real-life prototyping, indicate that it is possible and technically feasible to run both reconstruction and image analysis functions in a diagnostically viable time frame in our local edge solution. Furthermore, we confirmed that the local edge computing capacity can be scaled up and down during runtime by adding or removing edge devices without the need for manual reconfigurations. The feasibility study from the viewpoint of resource and cost-efficiency is left for future work. Regarding edge computing security, the enabling technology for the most part exists, but there is additional work to be done to fully utilize it in the medical imaging field.

One of the greatest advantages of edge computing is related to the additional mobility of diagnostic services [8], facilitating e.g., mobile imaging in and outside hospitals, allowing, among others, vehicle-mounted imaging units (e.g., mobile stroke units) [36]. Therefore, in the next phases of our work, we will focus on system-level optimizations of performance, reliability, and resource-efficiency. While this work focused on local edge computing, we plan to extend the scope to assess on, e.g., what tier of the edge-cloud continuum the reconstruction and analysis tasks are beneficial to be run in different mobile use cases. A particular interest is in considering the availability of computational resources on different tiers and the quality of the connections between the tiers, and their effect on selecting the optimal tier for algorithms with respect to performance, resource-efficiency, energy-efficiency, and security.

Acknowledgements

The authors would like to acknowledge the scientific discussions with Ekaterina Sirazitdinova. NVIDIA Corporation is gratefully acknowledged for donating the Quadro P6000 used for this research. Anonymous reviewers are acknowledged for their valuable comments.

Ethics declarations

Informed consent

This manuscript was prepared using phantom data and open data made available for academic research.

Conflict of interest

Authors declare no conflicts of interest.

Financial support

This research work was supported by Academy of Finland, under the DigiHealth (grant 326291) and 6G Flagship projects (346208). Antti Isosalo received funding from the Jenny and Antti Wihuri Foundation (210099).

References

[1] Masjedi H, Zare MH, Keshavarz Siahpoush N, Razavi-Ratki SK, Alavi F, Shabani M. European trends in radiology: investigating factors affecting the number of examinations and the effective dose. Radiol Med. 2020 Mar;125(3):296-305. https://doi.org/10.1007/s11547-019-01109-6

[2] Vandenberghe B, Jacobs R, Bosmans H. Modern dental imaging: a review of the current technology and clinical applications in dental practice. Eur Radiol. 2010 Nov;20(11):2637-55. https://doi.org/10.1007/s00330-010-1836-1

[3] Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. Clin Radiol. 2019 May;74(5):357-366. https://doi.org/10.1016/j.crad.2019.02.006

[4] Shad R, Cunningham JP, Ashley EA, Langlotz CP, Hiesinger W. Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. Nat Mach Intell. 2021;3(11):929-935. https://doi.org/10.1038/s42256-021-00399-8

[5] Hartmann M, Hashmi US, Imran A. Edge computing in smart health care systems: Review, challenges, and research directions. Trans Emerg Telecommun Technol. 2022;33(3):e3710. https://doi.org/10.1002/ett.3710

[6] Harjula E, Karhula P, Islam J, Leppänen T, Manzoor A, Liyanage M, et al. Decentralized IoT edge nanoservice architecture for future gadget-free

computing. IEEE Access. 2019;7:119856-119872. https://doi.org/10.1109/ACCESS.2019.2936714

[7] Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proc. IEEE. 2019;107(8):1738-1762. https://doi.org/10.1109/JPROC.2019.2918951

[8] Zhang J, Li D, Hua Q, Qi X, Wen Z, Myint H. 3D remote healthcare for noisy CT images in the Internet of things using edge computing. IEEE Access. 2021;9:15170-15180. https://doi.org/10.1109/ACCESS.2021.3052469

[9] Chen X, Zhu Y, Zheng X, Miao S, Nan T, Li W. Parallel multi-threaded Gridrec algorithm for computer tomography on GPU for edge computing. 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom). New York, NY, USA: IEEE; 2020. p. 193-198. https://doi.org/10.1109/CSCloud-EdgeCom49738.2020.00042

[10] Abid A, Sinha P, Harpale A, Gichoya J, Purkayastha S. Optimizing medical image classification models for edge devices. In: Matsui K, Omatu S, Yigitcanlar T, González SR (eds). Distributed Computing and Artificial Intelligence, Volume 1: 18th International Conference. DCAI 2021. Lecture Notes in Networks and Systems, vol 327. Springer, Cham; 2021. p. 77-87. https://doi.org/10.1007/978-3-030-86261-9_8

[11] Zhang Y, Liang Y, Jia B, Wang P, Zhang X. A blockchain-enabled learning model based on distributed deep learning architecture. Int J Intell Syst. 2022;37(9):6577-6604. https://doi.org/10.1002/int.22907

[12] de Souza Rebouças E, de Medeiros FNS, Marques RCP, Chagas JVS, Guimarães MT, Santos LO, et al. Level set approach based on Parzen Window and floor of log for edge computing object segmentation in digital images. Appl Soft Comput. 2021;105:107273. https://doi.org/10.1016/j.asoc.2021.107273

[13] Poon CC, Jiang Y, Zhang R, Lo WW, Cheung MS, Yu R, Zheng Y, Wong JCT, et al. AI-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices. NPJ Digit Med. 2020 May 18;3:73. https://doi.org/10.1038/s41746-020-0281-z

[14] Ciobanu A, Luca M, Barbu T, Drug V, Olteanu A, Vulpoi R. Experimental Deep Learning Object Detection in Real-time Colonoscopies. 2021 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 2021, p. 1-4, https://doi.org/10.1109/EHB52898.2021.9657740

[15] van Aarle W, Palenstijn WJ, De Beenhouwer J, Altantzis T, Bals S, Batenburg KJ, et al. The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography. Ultramicroscopy. 2015 Oct;157:35-47. https://doi.org/10.1016/j.ultramic.2015.05.002

[16] van Aarle W, Palenstijn WJ, Cant J, Janssens E, Bleichrodt F, Dabravolski A, et al. Fast and Flexible X-ray Tomography Using the ASTRA Toolbox. Opt Express. 2016 Oct 31;24(22):25129-25147. https://doi.org/10.1364/OE.24.025129

[17] Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: Towards a full-field digital mammographic database. Acad Radiol. 2012 Feb;19(2):236-48. https://doi.org/10.1016/j.acra.2011.09.014

[18] Shen Y, Wu N, Phang J, Park J, Liu K, Tyagi S, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med Image Anal. 2021

Feb;68:101908. https://doi.org/10.1016/j.media.2020.101908

[19] Wu N, Phang J, Park J, Shen Y, Kim SG, Heacock L, et al. The NYU breast cancer screening dataset v1.0. New York University, New York, NY, USA, Tech Rep.; 2019 [cited 29 Oct 2022]. Available from: https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf

[20] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32.

[21] Islam J, Kumar T, Kovacevic I, Harjula E. Resource-aware dynamic service deployment for local IoT edge computing: Healthcare use case. IEEE Access. 2021;9:115868-115884. https://doi.org/10.1109/ACCESS.2021.3102867

[22] Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, Mann RM. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology. 2019 Feb;290(2):305-314. https://doi.org/10.1148/radiol.2018181371

[23] Scarfe WC, Farman AG. What is cone-beam CT and how does it work? Dent Clin North Am. 2008 Oct;52(4):707-30. https://doi.org/10.1016/j.cden.2008.05.005

[24] Abdelhafez HA, Halawa H, Pattabiraman K, Ripeanu M. Snowflakes at the edge: A study of variability among NVIDIA Jetson AGX Xavier boards. In: Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking (EdgeSys '21). New York, NY, USA: Analytics and Networking Association for Computing Machinery; 2021. p. 1-6. https://doi.org/10.1145/3434770.3459729

[25] Ruonala V. Radiologisten tutkimusten ja toimenpiteiden määrät vuonna 2021: Terveydenhuollon valvontaraportti. STUK-B 295 / Lokakuu 2022. Säteilyturvakeskus; 2022. https://urn.fi/URN:ISBN:978-952-309-543-4.

[26] Després P, Jia X. A review of GPU-based medical image reconstruction. Phys Med. 2017 Oct;42:76-92. https://doi.org/10.1016/j.ejmp.2017.07.024

[27] Chong A, Weinstein SP, McDonald ES, Conant EF. Digital breast tomosynthesis: concepts and clinical practice. Radiology. 2019 Jul;292(1):1-14. https://doi.org/10.1148/radiol.2019180760

[28] Park J, Chłędowski J, Jastrzębski S, Witowski J, Xu Y, et al. 3D-GMIC: An efficient deep neural network to find small objects in large 3D images. arXiv preprint 2022. arXiv:2210.08645.

[29] Alcaín E, Fernández PR, Nieto R, Montemayor AS, Vilas J, Galiana-Bordera A, et al. Hardware Architectures for Real-Time Medical Imaging. Electronics. 2021;10(24):3118. https://doi.org/10.3390/electronics10243118

[30] Palenstijn WJ, Bédorf J, Sijbers J, Batenburg KJ. A distributed ASTRA toolbox. Adv Struct Chem Imaging. 2017;2(1):19. https://doi.org/10.1186/s40679-016-0032-z

[31] Vanholder H. Efficient inference with TensorRT. In: In Proceedings of the GPU Technology Conference, San Jose McEnery Convention Center, Silicon Valley, CA, USA, 4–7 April 2016. Vol. 1, p. 2.

[32] Xu S, Qian Y, Hu RQ. Edge Computing and Secure Edge Intelligence. In: Cybersecurity in Intelligent Networking Systems. IEEE; 2023. p. 31-44. https://doi.org/10.1002/9781119784135.ch3

[33] Zeyu H, Geming X, Zhaohang W, Sen Y. Survey on Edge Computing Security. In: 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Fuzhou, China, 2020. p. 96-105. https://doi.org/10.1109/ICBAIE49996.2020.00027

[34] Mahler T, Shalom E, Elovici Y, Shahar Y. A Dual-Layer Architecture for the Protection of Medical Devices from Anomalous Instructions. In: Michalowski M, Moskovitch R (eds). Artificial Intelligence in Medicine. AIME 2020. Lecture Notes in Computer Science, vol 12299. Springer, Cham; 2020. https://doi.org/10.1007/978-3-030-59137-3_25

[35] Becker AS, Jendele L, Skopek O, Berger N, Ghafoor S, Marcon M, Konukoglu E. Injecting and removing suspicious features in breast imaging with CycleGAN: A pilot study of automated adversarial attacks using neural networks on small images. Eur J Radiol. 2019 Nov;120:108649. https://doi.org/10.1016/j.ejrad.2019.108649

[36] Pulsiri N, Vatananan-Thesenvitz R, Sirisamutr T, Wachiradilok P. The Emerging Technology Identification of Ambulance Technologies in Thailand. In: Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 2022. p. 1-9. https://doi.org/10.23919/PICMET53225.2022.9882565