REGULAR ARTICLE

Multi-Task Neural Network for Medical Diagnosis Targeted for FPGA Deployment

A. Dubitskyi[1,*], O. Glukhov[1], V. Beresnev[2]

[1] *Kharkiv National University of Radio Electronics, 61166 Kharkiv, Ukraine*
[2] *V. N. Karazin Kharkiv National University, 61022 Kharkiv, Ukraine*

(Received 08 January 2026; revised manuscript received 22 February 2026; published online 25 February 2026)

In modern medical practice, automated processing of medical images plays a critical role; however, deploying artificial intelligence systems on portable equipment faces significant computational resource constraints. This paper addresses the challenge of deploying deep learning algorithms on embedded systems, Field-Programmable Gate Arrays (FPGAs), by applying the Multi-Task Learning (MTL).

Instead of utilizing separate models for each task, a unified neural network architecture with hard-shared parameters is proposed. The MobileNetV2 network was selected as the backbone, serving as a feature extractor. For each specific task – pneumonia detection on X-ray images and brain tumor detection on MRI scans – a separate head is allocated. Furthermore, to dynamically adapt features extraction depending on the input image type a Task Embedding layer is added before the heads.

To optimize the neural network training process, Automatic Mixed Precision (AMP) technology and data loading pipeline optimization were employed. To enhance model generalization, complex geometric and photometric data augmentation was applied according to Inductive Bias principle. For further adaptation of the model to FPGA, a method of model quantization from FP32 to INT8 format was introduced.

Experimental results confirm that the proposed approach ensures high diagnostic accuracy for both pathologies while significantly saving memory resources and power consumption.

Keywords: Multi-task learning, Medical imaging, MRI, X-Ray, Image analysis.

DOI: 10.21272/jnep.18(1).01009

PACS numbers: 87.57.R –, 87.57. – s, 07.05.Mh, 95.75.Mn

1. INTRODUCTION

Medical imaging programs are computationally demanding. They need profound calculations and real-time image processing. Medical image resolutions, color depth, and consequently file sizes tend to increase. Moreover, medical imaging processing programs must provide a faultless workflow with a big number of such images.

In this manner, the development and implementation of an integrated hardware/software system is crucial for medical imaging programs.

Nevertheless, neural networks (NN) have seen significant growth in medical image processing recently. However, to run any neural network a graphics processing unit (GPU) should be installed. Otherwise, it will not be efficient setting up NN on a processor (CPU) [1]. Many hospital and office PCs rarely have an installed GPU.

In this case field-programmable gate arrays (FPGA) can be an appropriate solution to run NN for PCs without a dedicated GPU on board [1].

When creating compact solutions for neural networks it is important to operate within the available resources. Main FPGA specifications are the number of look-up-tables (LUT), flip-flops, logic cells and their interconnections, and on-chip memory. All these resources determine the quantity of NN parameters.

Hence, if the purpose is to analyze multiple anatomical regions, it is impractical to load model's parameters each time the task is changed. On the contrary, there is an option to use a single NN for several tasks simultaneously, which is called the multi-task learning principle [2-4].

Thereby, the purpose of this paper is to demonstrate the possibility of developing a NN based on MobileNetV2 for simultaneously solving several medical diagnostic tasks, by means of examples of detecting pneumonia on chest X-rays and detecting brain tumors on MRI images, with minimal hardware requirements.

2. MULTI-TASK LEARNING

Multi-task learning is a way to design NN architecture in order to achieve proper generalization of the model by means of using additional information and shared features of related tasks.

Another benefit when using a single NN for several

[*] Correspondence e-mail: artur.dubitskyi@nure.ua

Cite this article as: A. Dubitskyi et al., *J. Nano- Electron. Phys.* 18 No 1, 01009 (2026) https://doi.org/10.21272/jnep.18(1).01009

tasks is saving computational resources, which comes in handy when implementing a created model on FPGA to form an AI accelerator. [2-5].

Current multi-task learning configurations can be divided by the means of designing shared parameters. From one perspective, there are general parameters for several tasks, while on the other, hidden relationships between different tasks are discovered. In other words, this is achieved by means of hard shared and soft shared parameters of hidden layers, respectively.

2.1 Hard-Shared Parameters

When parameters of the NN are hard shared, it means the only output parameters are task specific from all NN parameters set and form heads. All the other parameters are common between tasks and form a backbone of the NN, Fig. 1.

This approach mitigates the risk of overfitting. It is obvious that the more tasks the model learns simultaneously, the more generalized representations it is required to find.

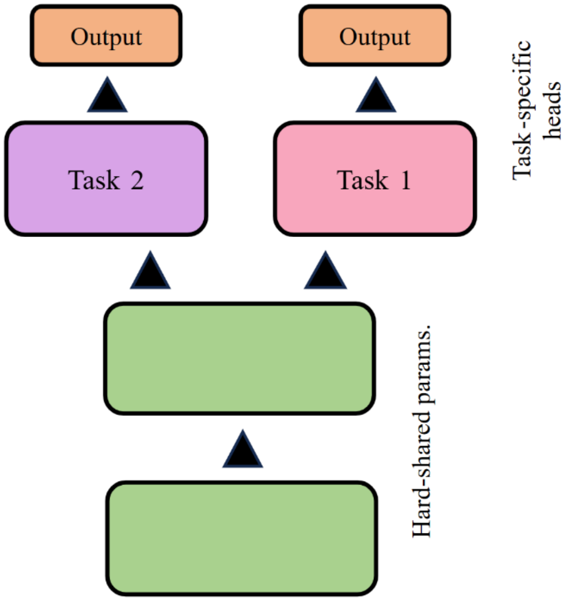


Fig. 1 – Hard-shared parameter structure

A critical challenge in designing hard parameter sharing models is determining the optimal branching point, deciding the ratio of shared to task-specific parameters [2-4].

2.2 Soft-Shared Parameters

In the soft parameter sharing structure, each task is assigned its own set of parameters. However, the input for each subsequent layer is a linear combination of the outputs from the previous layer of all task networks, Fig. 2. The parameters of this linear combination are task-specific, allowing each layer to selectively choose which tasks to derive information from.

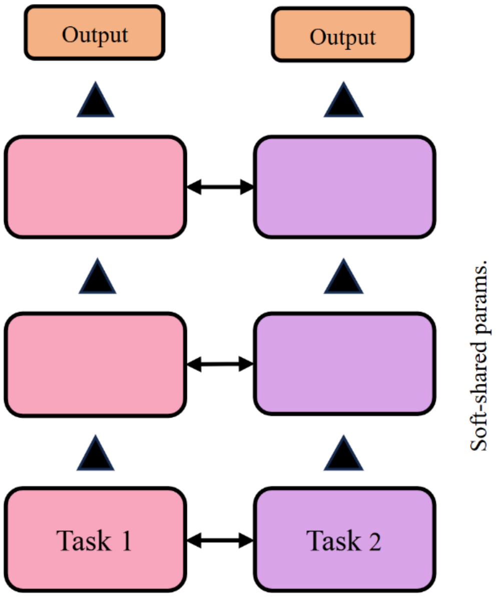


Fig. 2 – Soft-shared parameters structure

Scalability is the problem of soft shared architectures. The model size tends to increase linearly with the number of tasks. [2-4].

Moreover, beside NN structure configuration there are some other ways to implement proper inductive bias such as data augmentation, using different optimization algorithms and learning approaches [6].

2.3 Augmentation and Inductive Bias

Selecting a neural network model for a specific task involves ensuring that its architecture has the appropriate inductive bias for the target problem. Accordingly, inductive bias represents additional information about the data provided to the model, guiding it to appropriate problem handling depending on the input data type [6]. Inductive bias is introduced at every stage of model development. First, every model has a specific architecture: the number of layers, parameters, activation functions, etc. Similarly, each layer – whether convolutional or fully connected – has an inherent bias determined by its structure. For instance, convolutional layers are biased towards image processing, while recurrent layers are suited for sequential data processing.

Second, during training, using backpropagation and optimization algorithms (such as Adam or RMSProp), we make the model generalize in a particular way.

Third, bias is introduced during the training data preparation stage. A training set of brain MRI scans can serve as an example. By training a neural network on the scans shown in Fig. 3, we are likely to achieve a generalization where the network recognizes them only in a specific orientation and against a black background surrounding the grayscale head image. This occurs because the model did not take into account other head positions, variations in sharpness, or combinations of these factors during training.

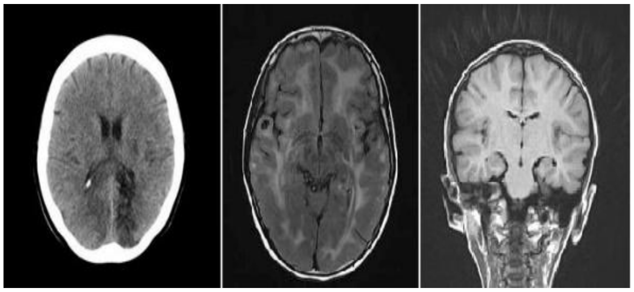


Fig. 3 – Head MRI-scans

Nevertheless, applying augmentation to a training set, which includes random rotation by 20 degrees, brightness and contrast changes by 20%, random zoom, blank spots, etc., Fig. 4, allows the NN model to choose a generalization method that will classify imperfectly prepared images, which is appropriate when analyzing MRI scans in real conditions.

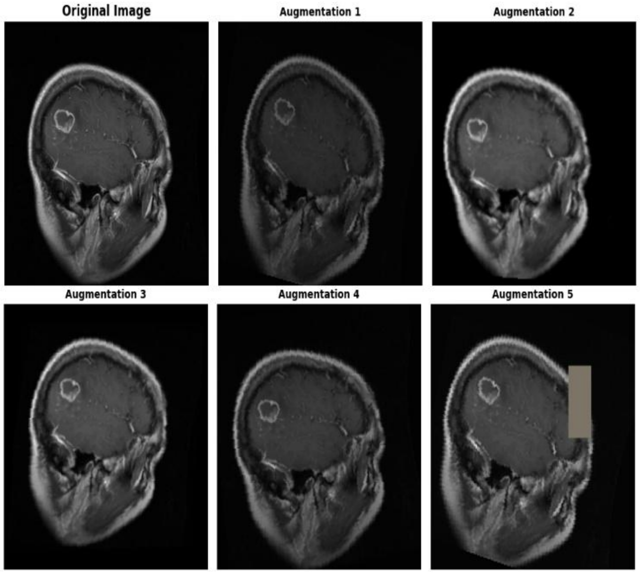


Fig. 4 – Augmentatation example for MRI-scans

Consequently, multi-task learning allows implementing the necessary inductive bias, ensuring the model looks for solutions for multiple tasks. The primary methods for designing such models involve utilizing structures with common (hard-shared) and distributed (soft-shared) parameters, each having its own advantages and disadvantages.

3. HARD-SHARED PARAMETERS STRUCTURE AND MOBILENETV2 AS A BACKBONE

A neural network utilizing hard parameter sharing was developed, Fig. 1. The pre-trained convolutional NN MobileNetV2 served as the backbone for the shared layers.

A task embedding layer was integrated into the architecture, enabling the model to adapt extracted features according to the input data type. The architecture is concluded by two specialized classification heads, each optimized for a specific task: the first for pneumonia detection in chest X-rays, and the second for brain tumor identification in MRI slices, Fig. 5.

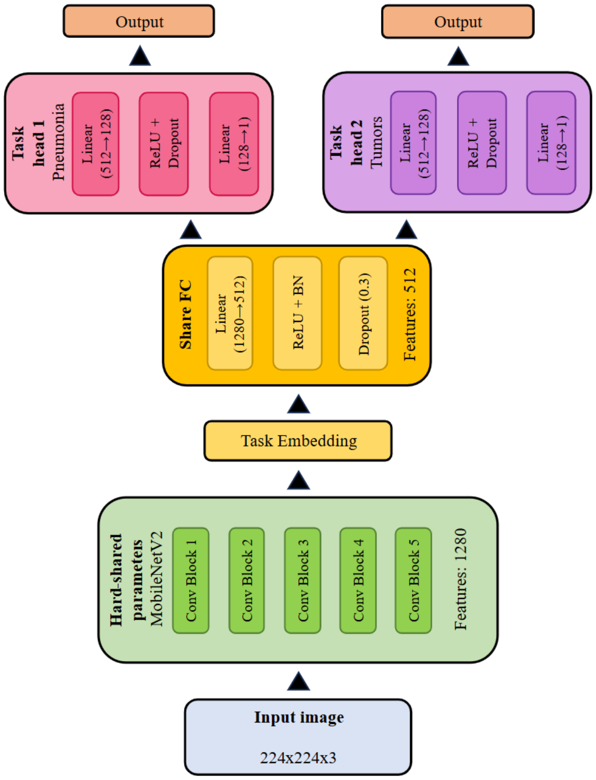


Fig. 5 – Developed NN structure with a hard-shared parameters and MobileNetV2 as a backbone

The network accepts an image as input, standardized to a resolution of 224 by 224 pixels with RGB color channels. The feature extractor consists of five convolutional blocks based on MobileNetV2, which extracts important visual patterns and generates an output feature vector of size 1280.

To incorporate task-type information, a Task Embedding mechanism is introduced. The subsequent fully connected layer reduces the feature dimensionality to 512, integrating it with the task information from the preceding block.

The output heads of the neural network have distinct parameter sets, despite having identical structures. Their primary function is to isolate features specific to each task, reducing the vector size to 128. To mitigate overfitting within each head, the ReLU activation function and Dropout regularization are utilized. Finally, the feature vector is reduced to a dimension of 1.

3.1 Data Preparation

During the classification task selection process, it was found that multi-task neural networks perform effectively on related tasks, such as the classification of COVID-19 and pneumonia diseases, both of which affect the lungs. On the contrary, this study proposed examining diseases that affect different anatomical regions of the body, specifically pathologies of the brain and the chest [5].

Consequently, the Chest X-Ray Pneumonia dataset was utilized as the training sample for the first task, while the Brain Tumor MRI dataset was employed for the second. The composition of these datasets is illustrated in Fig. 6.

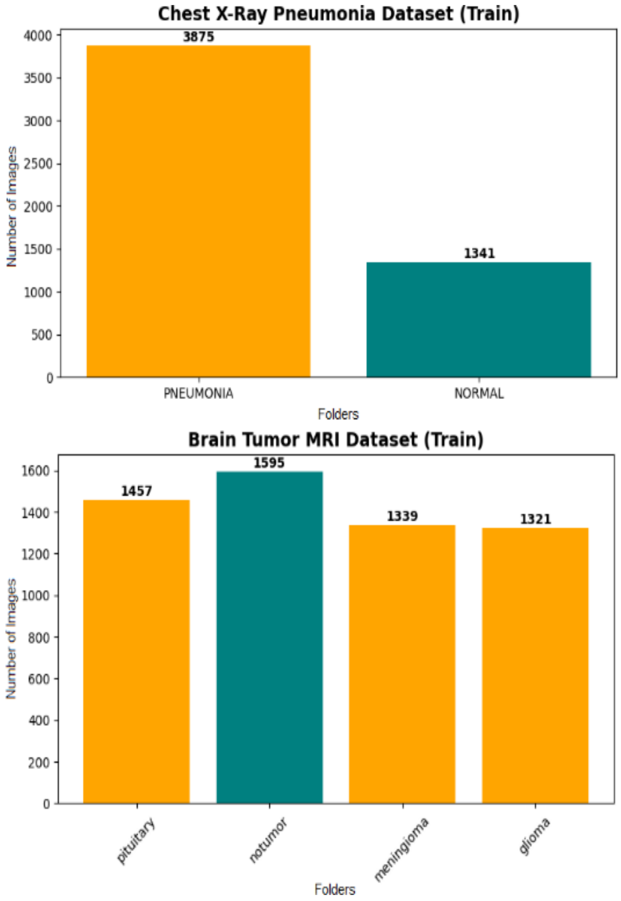


Fig. 6 – The structure of Chest X-Ray Pneumonia and Brain Tumor MRI datasets

An important step was the implementation of image augmentation following the principles described in section 2.3, as shown in Fig. 4 and Fig. 7.

3.2 Training Process Optimization

To accelerate the neural network training process, Automatic Mixed Precision (AMP) technology was used. This approach allows for the execution of resource-intensive calculations using 16-bit floating-point numbers (FP16) instead of the standard 32-bit (FP32), significantly reducing video memory consumption and computation time without compromising model convergence.

Additionally, hyperparameter optimization was performed, specifically regarding batch size and the number of parallel data loading processes (num_workers). To enhance data transfer efficiency between the CPU and GPU within the data preparation pipeline, the pin_memory and persistent_workers parameters were activated.

For future model adaptation for deployment on FPGAs, a weight quantization method converting from FP32 to INT8 format was applied.

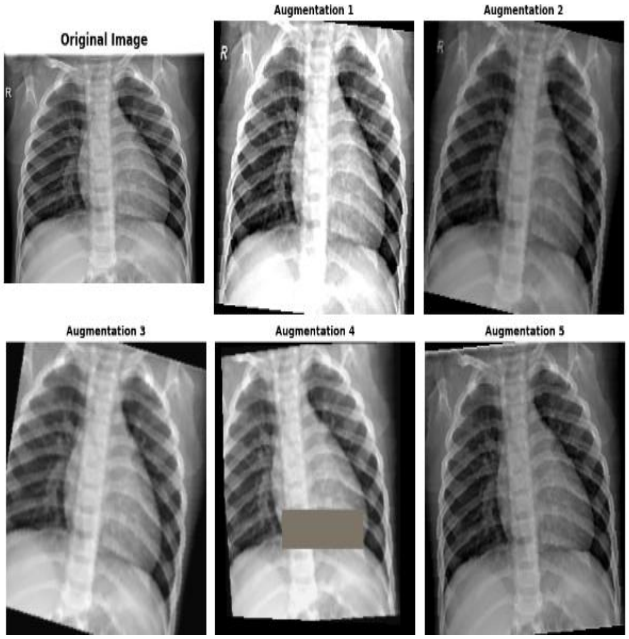


Fig. 7 – Augmentation example for chest X-rays

This resulted in a fourfold reduction in model size with a negligible decrease in classification accuracy (within the range of 1–2%).

Thus, a neural network architecture capable of analyzing two tasks across different anatomical regions was established. Medical image augmentation was performed, and optimization measures for neural network training were implemented.

4. RESULT DISCUSSION

The previously discussed NN model was tested using random images selected from various datasets on Kaggle.com. Training proceeded for twenty epochs with early stopping enabled if no improvement was observed over three epochs. The final training duration was 14 epochs.

The neural network demonstrated high prediction accuracy for both pneumonia and brain tumors. The performance results are presented in Figs. 9 and 10.

The model correctly identifies the presence and absence of the target pathologies, fully confirming the feasibility of its future application in creating an FPGA-based medical image analyzer, although further optimization is required.

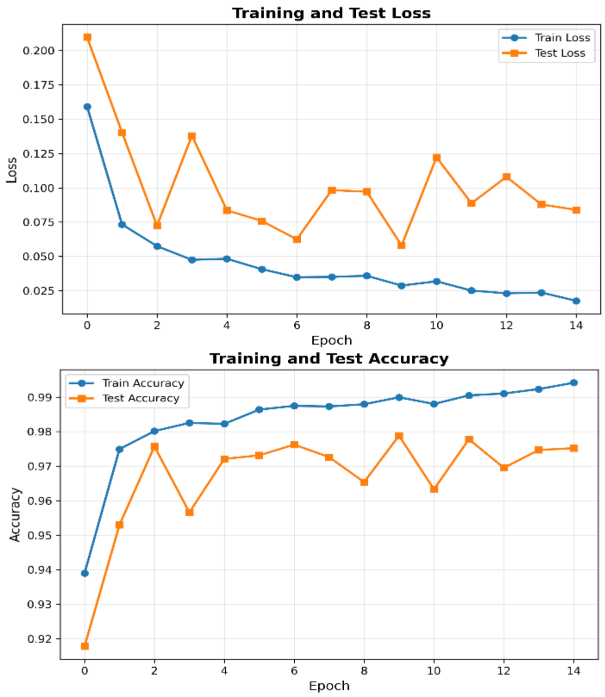
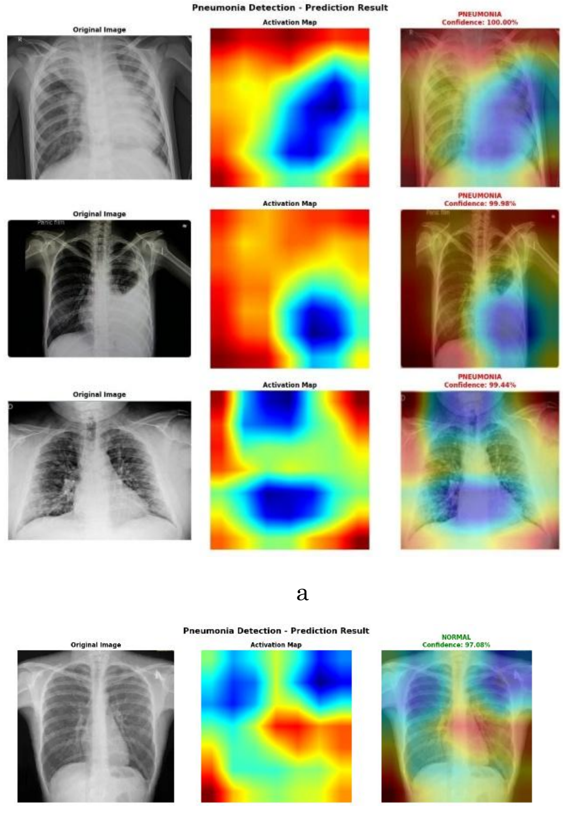


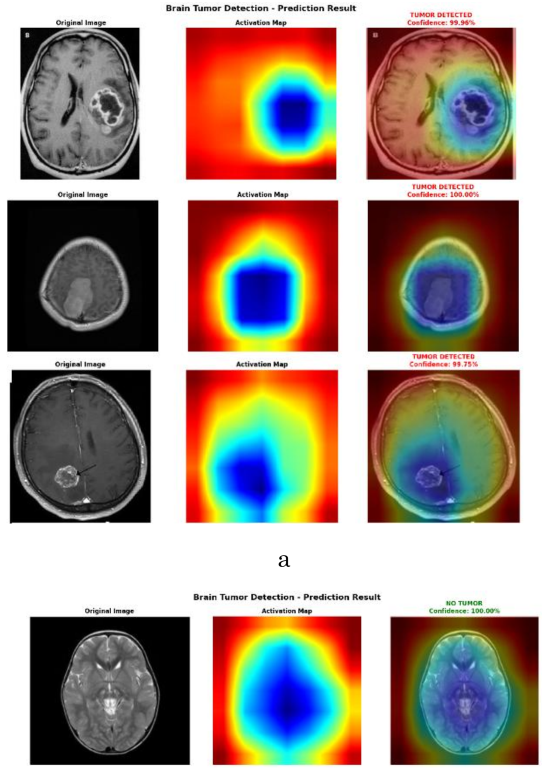
Fig. 8 – Loss and accuracy dynamics over training epochs



а

б

Fig. 9 – Performance results of the developed NN analyzing chest X-rays for predicting the presence of pneumonia (a) and its absence (b)



а

б

Fig. 10 – Performance results of the developed NN analyzing brain tumor MRI scans for predicting the presence of tumor (a) and its absence (b)

5. CONCLUSION

This study proposed a multi-task neural network architecture designed for the simultaneous classification of diseases affecting distinct anatomical regions based on medical imaging data.

It was demonstrated that utilizing the pre-trained MobileNetV2 model significantly enhanced the model's feature extraction capabilities and reduced computational requirements, thereby facilitating efficient and reliable predictions.

The research findings underscore the potential for deploying such a solution on resource-constrained platforms, specifically FPGAs.

It is feasible to increase the number of tasks, both primary and auxiliary (e.g., noise reduction, contrast enhancement), which significantly expands the application scope of such devices, while generalization for related tasks is improved.

Thus, this study demonstrates the effectiveness of applying the multi-task learning approach to the development of software/hardware medical diagnostic systems targeted for operation in resource-constrained environments.

REFERENCES

1. E. Alcaín, P.R. Fernández, R. Nieto, A.S. Montemayor, *Electronics* **10** No 24, 3118 (2021).
2. M. Crawshaw, *arXiv*:2009.09796.
3. S. Vandenhende, S. Georgoulis, W.V. Gansbeke, M. Proesmans, D. Dai, L.V. Gool, *arXiv*:2004.13379.
4. S. Ruder, *arXiv*:1706.05098.
5. M. Rhanoui, K. Alaoui Belghiti, M. Mikram, *Onco* **5** No 3, 34 (2025).
6. A. Boopathy, W. Yue, J. Hwang, A. Iyer, I. Fiete, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*-(IJCAI), 3733 (Jeju, Korea:IJCAI: 2024).

Багатоцільова нейрона мережа для медичної діагностики з перспективою розгортання на ПЛІС

А.Є. Дубіцький[1], О.В. Глухов[1], В.М. Береснєв[2]

[1] *Харківський національний університет радіоелектроніки, 61166 Харків, Україна*
[2] *Харківський національний університет імені В. Н. Каразіна, 61022 Харків, Україна*

У сучасній медичній практиці автоматизована обробка діагностичних зображень відіграє критичну роль, проте впровадження систем штучного інтелекту на портативне обладнання стикається з обмеженнями обчислювальних ресурсів. Ця стаття присвячена вирішенню проблеми розгортання алгорит-мів глибокого навчання на вбудованих системах, зокрема на програмованих логічних інтегральних схемах (ПЛІС), шляхом застосування методології багатоцільового навчання (Multi-Task Learning, MTL).

Замість використання окремих моделей для кожного завдання, пропонується використовувати уніфіковану архітектуру нейронної мережі зі загальним використанням параметрів. Основою моделі обрано мережу MobileNetV2, яка виступає екстрактором ознак. Для кожної задачі, пошук пневмонії на рентгенівських знімках та пошук пухлин мозку на сканах МРТ, виділяється окрема голова, а для динамічної адаптації видобутих ознак, залежно від типу вхідного зображення, перед головами доданий шар Task Embedding.

Для оптимізації процесу навчання нейронної мережі було використано технологію змішаної точно-сті (Automatic Mixed Precision) та оптимізації конвеєра завантаження даних. За принципом індуктив-ного зміщення (Inductive Bias) та для підвищення генералізації моделі, застосовано комплексну гео-метричну та фотометричну аугментацію даних. Для подальшої адаптації моделі на ПЛІС введений спосіб квантизації моделі зі формату FP32 у формат INT8.

Результати експериментів підтверджують, що запропонований підхід забезпечує високу точність діагностики обох патологій, суттєво заощаджуючи ресурси пам'яті та енергоспоживання.

Ключові слова: МРТ, Багатоцільове навчання, Рентген, Аналіз зображень, Медична візуалізація.