Implementation of deep learning based intelligent image analysis on an edge AI platform using heterogeneous AI accelerators

Ryangsoo Kim, Jaein Kim, Hark Yoo, and Sung Chang Kim
Honam Research Center
Electronics and Telecommunications Research Institute (ETRI)
Gwangju, Republic of Korea
Email: {rskim, jaein, harkyoo, sungchang}@etri.re.kr

***Abstract*—Recent advancements in artificial intelligence (AI) technology have spurred efforts to implement deep learning-based intelligent image analysis in various video surveillance applications. However, traditional approaches to using remote cloud computing platforms results in increased network latency and potential breaches of personal information. To address these issues, the adoption of edge AI, utilizing deep learning-based data analysis directly on the devices where image data is collected, has emerged as a promising solution. In this paper, we introduce a practical edge AI platform that enables real-time deep learning-based image analysis, including object detection and multi-person pose estimation. The platform is built on an embedded board equipped with heterogeneous AI accelerators, facilitating parallel inference of multiple deep learning-based image analysis models. In addition, the platform applies task-level pipeline parallelism to maximize the utilization of computing resources, which leads to a reduction in overall image analysis latency. Experimental results demonstrate the effectiveness of our edge AI platform in providing real-time intelligent video analysis services.**

***Index Terms*—Intelligent image analysis, edge intelligence, real-time pose estimation**

I. Introduction

With the recent explosive advancement of artificial intelligent (AI) technology, particularly in the area of innovative deep learning models like convolutional neural networks, intelligent image analysis based on deep learning models has led to widespread utilization in various intelligent video surveillance applications such as unmanned store management, parking space management, intrusion detection, and threat event detection [1]. The main feature of intelligent image analysis is its utilization of deep learning-based image processing to extract meaningful insights from image data, such as object detection and human pose estimation for action recognition. In this context, the real-time implementation of intelligent video analysis is closely intertwined with the duration required to perform deep learning model inference. Therefore, reducing the time required for model inference is critical for the practical real-time implementation of deep learning-based intelligent image analysis.

Traditional approaches to implementing intelligent image analysis services leverage a remote cloud computing platform consisting of multiple high-performance computing servers to accelerate deep learning inference. However, deploying intelligent image analysis services on cloud platforms introduces additional network latency due to the transmission of image data over the internet. In addition, this transmission exposes the data to potential privacy breaches while incurring an increase in network traffic. To address this issue, the adoption of edge AI, which performs deep learning-based image analysis directly on the devices where image data is collected, has gained significant attention. This approach not only reduces latency and network traffic, but also enhances privacy by eliminating the need for data transmission over the internet.

In this paper, we present a practical edge AI platform designed to implement intelligent image analysis services, with a specific focus on object detection and multi-person pose estimation. The proposed platform is implemented on an embedded board equipped with heterogeneous AI accelerators, enabling parallel inference of multiple deep learning-based image analysis models. In addition, the platform employs task-level pipeline parallelism to maximize computing resource utilization, thereby reducing image analysis latency. Through real-world experimentation, we demonstrate the effectiveness of our edge AI platform in providing real-time intelligent image analysis services.

II. Proposed edge AI platform for deep learning based intelligent image analysis

In this study, we consider one-stage object detection and multi-person pose estimation as deep learning based intelligent image analysis services. Unlike two-stage object detection, which first estimates object regions and then identifies object types, one-stage object detection simultaneously infers the location and type of objects in an image through a single deep learning model inference. The multi-person pose estimation method employs a top-down approach, first detecting a persons position in video images and then estimating poses for each person. This process necessitates sequential inference

This work was supported by the KETEP and the MOTIE of the Republic of Korea (No. 2021202090053A).

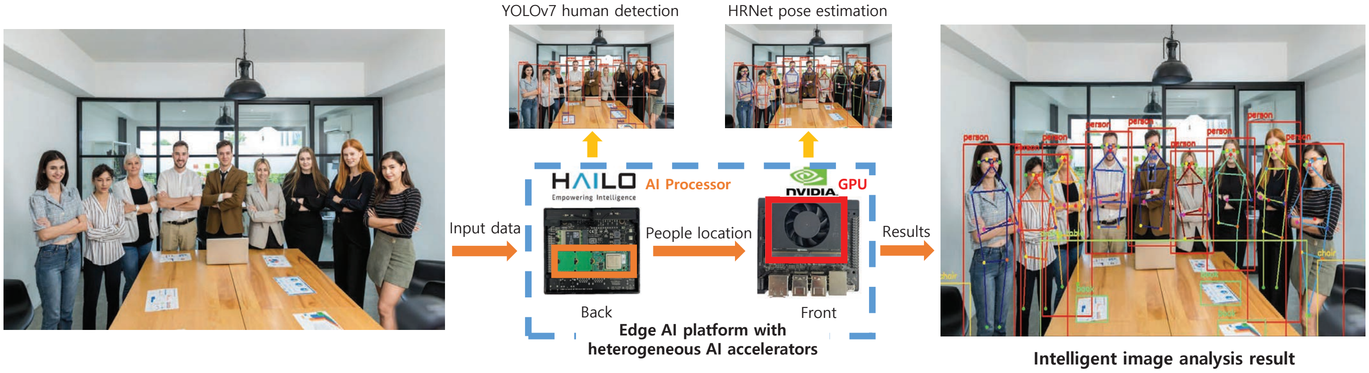

YOLOv7 human detection
HRNet pose estimation
Input data
HAILO Empowering Intelligence
AI Processor
NVIDIA
GPU
People location
Results
Back
Front
Edge AI platform with heterogeneous AI accelerators
Intelligent image analysis result

Fig. 1: Overview of deep learning based intelligent image analysis implementation on edge AI platform with heterogeneous AI aceelerators.

operations performed by two different deep learning models designed for people detection and human pose estimation, respectively. In this study, we use the object detection deep learning model, trained on the COCO dataset, can effectively detect 80 types of objects, including persons. Consequently, the location information of individual person detected by the object detection model as input data for the multi-person pose estimation technology.

The primary objective of the edge AI platform is to enable deep learning-based image analysis directly on the devices where image data is collected. Typically, embedded boards are used as image data collection devices due to their low power consumption and compact size. Thus, employing an embedded board with AI accelerators emerges as an appropriate solution for implementing the edge AI platform in real-world environments. In this study, we specifically employ the NVIDIA Jetson Xavier NX embedded board, supplemented with the Hailo-8 AI processor developed by HAILO, an Israeli AI chip manufacturer. As a result, the edge AI platform is capable of concurrently utilizing two different AI accelerators, the Hailo-8 AI processor and the NVIDIA embedded GPUs, to expedite multiple deep learning model inferences by allocating them to each AI accelerator.

In this study, object detection was achieved by creating a runtime environment using HailoRT software for the Yolov7 (640x640) deep learning model, and performing deep learning inference to detect 80 distinct objects, including people, in the input image data [2]. After the object detection model provides the location information of people in the image data, the multi-person pose estimation model iteratively estimates the pose information of each person. Multi-person pose estimation was achieved by creating a runtime environment for HRNet (256x192) deep learning models trained on the COCO human pose estimation dataset and performing deep learning inference using embedded GPUs on NVIDIA Jetson Xavier NX [3]. Figure 1 illustrates the implementation of object detection and multi-person pose estimation on the proposed edge AI platform, equipped with heterogeneous AI accelerators.

The aforementioned technologies were implemented in Python, utilizing the queue-based task-level pipeline parallel processing technique proposed in [4] to improve intelligent image analysis latency performance. Task-level pipeline parallelism is a parallelization technique that divides a single program into a linear sequence of subprograms, ensuring the preservation of processing integrity throughout the parallel execution [5]. By applying the approach proposed in [4], we partition the deep learning-based intelligent image analysis service into four subprograms:

- READ: Capturing images from a live video stream.
- PRE-PROCESS: Resizing the captured image to match the input shape required by the deep learning model.
- INFERENCE: Performing deep learning model inference and refining the resulting data.
- WRITE: Displaying the intelligent image analysis results on the screen.

To facilitate task-level pipeline parallelism, we employ a queue-based multi-threaded programming architecture. At the output of each stage, a first-in-first-out (FIFO) queue is utilized

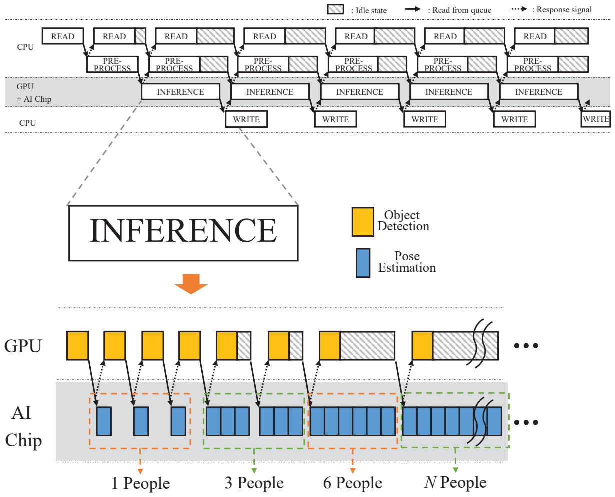
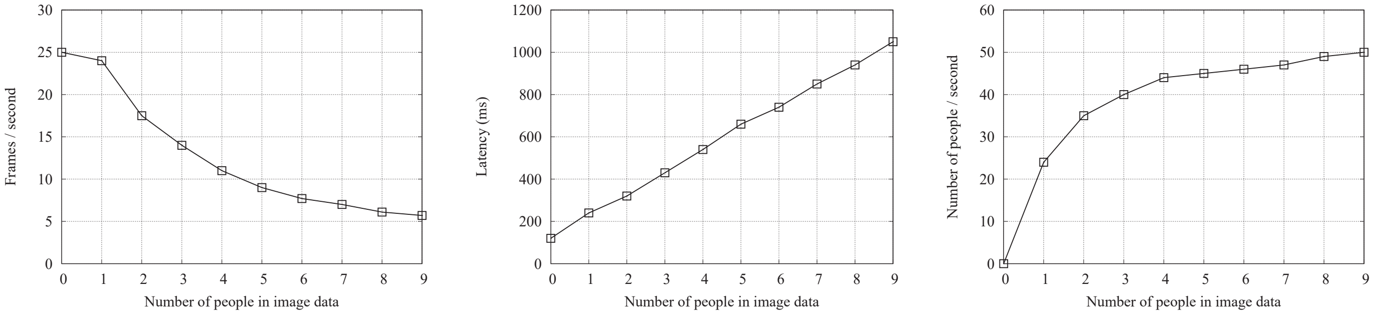


Fig. 2: Block diagram illustrating task-level pipeline parallelism in object detection and multi-person pose estimation implementation.



(a) Frames processing rate

(b) Latency

(c) People processing rate

Fig. 3: Performance of proposed edge AI platform with respect to the number of people in image data.

to transfer processed data to the subsequent stage, irrespective of its readiness. Once a stage is prepared, it retrieves data from the queue and proceeds with its task, ensuring the sequential execution of the stages. This approach facilitates the concurrent execution of multiple tasks on CPUs and AI accelerators, in which the deep learning inference tasks are specifically handled by the heterogeneous AI accelerators. By leveraging task-level pipeline parallelism, our system maximizes the utilization of computing resources, reducing the latency of the intelligent image analysis service. Figure 2 shows the block diagram illustrating task-level pipeline parallelism in object detection and multi-person pose estimation implementation.

III. Experimental results

In this section, we present real-world experimental results demonstrating the effectiveness of our edge AI platform in providing intelligent image analysis services in real-time. Figure 3 illustrates the performance of our proposed edge AI platform, which provides deep learning-based intelligent image analysis services consisting of object detection and multi-person pose estimation in consecutive order. As shown in Fig. 3 (a), there exists an inverse relationship between the processable image frame rate (FPS) performance and the number of people present in the image. Moreover, as shown in Fig. 3 (b), the end-to-end latency for obtaining intelligent image processing results increases proportionally with the number of people present in the image. These experimental results are attributed to the proportional increase in deep learning inference time for the multi-person pose estimator as the number of people present in an image increases.

Furthermore, Fig. 3 (c) demonstrates that the estimated human poses per second converge to 50 as the number people present in the image increases. Here, it is worth noting that employing the HRNet model in this study enables the estimation of approximately 50 human poses per second, necessitating approximately 20ms for a single inference operation on the GPUs embedded on the edge AI platform. This indicates that the proposed edge AI platform offers its maximum available performance. To achieve further performance enhancement, reducing the inference time of the human pose estimation deep learning model or exploiting additional computing resources is necessary.

IV. Conclusion and future work

In this paper, we presented a practical edge AI platform for implementing intelligent image analysis services, providing object detection and multi-person pose estimation. The proposed platform, implemented on an embedded board with heterogeneous AI accelerators, enables parallel inference of multiple deep learning-based image analysis models and employs task-level pipeline parallelism to maximize computing resource utilization for latency reduction. The experimental results demonstrate the effectiveness of edge AI platforms that provide real-time intelligent image analysis services, as well as their limitations. Future research will aim to improve multi-person pose estimation performance using distributed and collaborative computing approaches, providing real-time responses in scenarios with numerous people present in the image data.

References

[1] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, “Object Detection With Deep Learning: A Review,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212-3232, Nov. 2019.

[2] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” *arXiv prepreint*, arXiv:2207.02696, 2022.

[3] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Deep high-resolution representation learning for human pose estimation,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5693-5703.

[4] R. Kim, G. Kim, H. Kim, G. Yoon and H. Yoo, “A Method for Optimizing Deep Learning Object Detection in Edge Computing,” in *Proc. Int. Conf. Info. Commun. Technol. Converg. (ICTC)*, 2020, pp. 1164-1167.

[5] I.-T. A. Lee, C. E. Leiserson, T. B. Schardl, Z. Zhang, and J. Sukha, “On-the-fly pipeline parallelism,” in *ACM Transactions on Parallel Computing*, vol. 2, no. 3, pp. 17:1–17:42, Sep. 2015.