UOT:002.6 (002.6:025.5/.6)
DOI: https://doi.org/10.30546/09090.2025.01.1095

PERFORMANCE BENCHMARKING OF EMBEDDED EDGE DEVICES FOR VARIOUS FACE RECOGNITION MODELS

RASIM MAHMUDOV, SHAHLA URALOVA, AMIL BABAYEV

Baku Engineering University

rkmahmudov@gmail.com, suralova1@std.beu.edu.az, amilb@beu.edu.az

ARTICLE INFO

Article history:
Received: 2024-11-04
Received in revised form:2025-03-03
Accepted: 2025-03-14
Available online

Keywords:
Benchmarking,
Embedded Edge Device,
Face Recognition,
ResNet, MobileNet

ABSTRACT

Artificial Intelligence (AI) models are increasingly pivotal in enabling face recognition across various fields, from educational and research settings to public spaces. Effective deployment of these models requires high-performance hardware, such as RTX graphics cards or embedded edge devices like Nvidia's AGX Orin and Jetson Nano. This paper presents a comprehensive benchmarking study comparing the performance of these two devices, representing high and low-power edge computing options, using two face recognition models: ResNet and MobileNet.

The benchmarking process assesses each model across two different input sizes deployed on both devices with varied configurations, including CPU thread allocation and GPU power distribution within containerized environments. Performance metrics such as inference time, GPU utilization, memory usage, and CPU load are analyzed to determine each device's suitability and efficiency. Additionally, model-specific parameters, including FLOPS, parameter count, and memory footprint, are examined to provide for an in-depth comparison. This paper presents detailed results and analyses of these performance indicators.

I. Introduction

Face recognition technology, driven by advancements in artificial intelligence (AI), has become essential in numerous applications, including public safety, education, and retail. With the increasing demand for real-time, low-latency processing, deploying AI models on embedded edge devices has emerged as a viable solution, offering benefits like localized data processing and reduced dependency on cloud infrastructure. Among edge computing solutions, NVIDIA’s Jetson series provides popular choices such as the Jetson AGX Orin [5] and Jetson Nano [6], each representing different power and performance spectrum ends. As Table 1 shows, the AGX Orin, with its high computational capabilities, is designed for intensive AI tasks, while the Jetson Nano provides a more compact, low-power option suitable for lighter, localized applications.

Nvidia Jetson Module	Jetson AGX Orin	Jetson Nano
CPU	8-core ARM Cortex-A78AE @ 2.188 GHz	4-core ARM Cortex-A57 @ 1.43GHz
GPU	NVIDIA Ampere architecture GPU with 2048 CUDA cores	128-core Maxwell GPU
Memory	32 GB LPDDR5	4 GB LPDDR4
Storage	64 GB eMMC	No onboard storage (microSD)
Storage Type	eMMC 5.1 + NVME	microSD only
Power	15-60 W (configurable)	5-10 W
JetPack	JetPack 6.0	JetPack 4.2.1
Framework	TensorFlow, PyTorch	TensorFlow, PyTorch
AI Performance	Up to 275 TOPS	0.5 TOPS

Table 1: Specifications of Jetson boards

This paper aims to evaluate the performance of these two devices in handling face recognition tasks by benchmarking two well-known deep learning models, ResNet and Mobile Net. These models are widely used in computer vision for their accuracy and efficiency, making them ideal candidates for deployment in edge environments with limited resources. To further simulate real-world usage, this study uses WIDER Face and FDDB datasets, which contain challenging and varied face images across different environments, poses, and lighting conditions. This dataset choice allows for an in-depth assessment of each model's robustness and suitability for deployment on edge devices.

The benchmarking examines vital performance metrics, including inference time, GPU utilization, memory usage, and CPU load. Both devices are tested with different input sizes and configurations, such as CPU thread allocation and GPU power settings within edge devices.

Our study makes several contributions:

- A detailed comparative analysis of the Jetson AGX Orin and Jetson Nano devices for face recognition applications, focusing on how each handles resource-intensive tasks.

- Performance benchmarks across two AI models and two datasets offer insights into which model-hardware combinations are optimal for different face recognition scenarios.

- Exploring resource management, including model optimization and hardware scaling, impacts performance, providing valuable guidance for deploying AI on edge devices effectively.

This research addresses the growing need to understand how edge device configurations impact the performance of complex AI models. By examining these aspects, we offer a guide to optimizing face recognition model deployment on edge devices in practical settings where balancing efficiency and accuracy is critical.

II. Related Work

In the last few years, there have been several AI deployed benchmarking tests performed using Deep Neural Networks (DNNs) on different Nvidia Jetson Devices. The famous papers and their contributions are given below:

(i) DNN on embedded boards:

A study in [1] analyzed the performance of three edge devices - NVIDIA Jetson Nano, TX2, and Raspberry Pi 4 by running a convolutional neural network (CNN) model designed to classify fashion products. The comparison considered factors like power consumption, resource utilization (GPU, CPU, RAM), model accuracy, and overall cost, using datasets between 5,000 and 45,000 images. Relatively in [2] the authors introduced EdgeFace, an face recognition model designed specifically for edge devices, which reduces computational overhead while maintaining high accuracy.

(ii) Benchmarking DNNs on Jetson boards:

Evaluating Deep Learning on Jetson Devices: Jetson boards have been widely tested for deep learning deployment. In [3], researchers assessed 3D point cloud classification across various dataset sizes, analyzing computational demands on multiple Jetson models. Another study in [4] compared the performance of 3D object detection models specifically on NVIDIA Jetson AGX Xavier and Nano, examining their efficiency and processing capabilities.

III. Benchmark Analysis

A. Experimental setup

We benchmarked two Jetson devices: the AGX Orin as the high-performance setup and the

Nano as the lower-performance setup. Figure 1 illustrates the use of ResNet and MobileNet models for the Face Recognition algorithm, utilizing the Face Detection Dataset and Benchmark (FDDB) and Wider Face datasets. For these tests, we employed two data resolutions per dataset: 320x320 as the lightweight configuration and 800x800 as the heavy configuration for FDDB, and similarly, 224x224 as the lightweight and 640x640 as the heavy configuration for Wider Face.

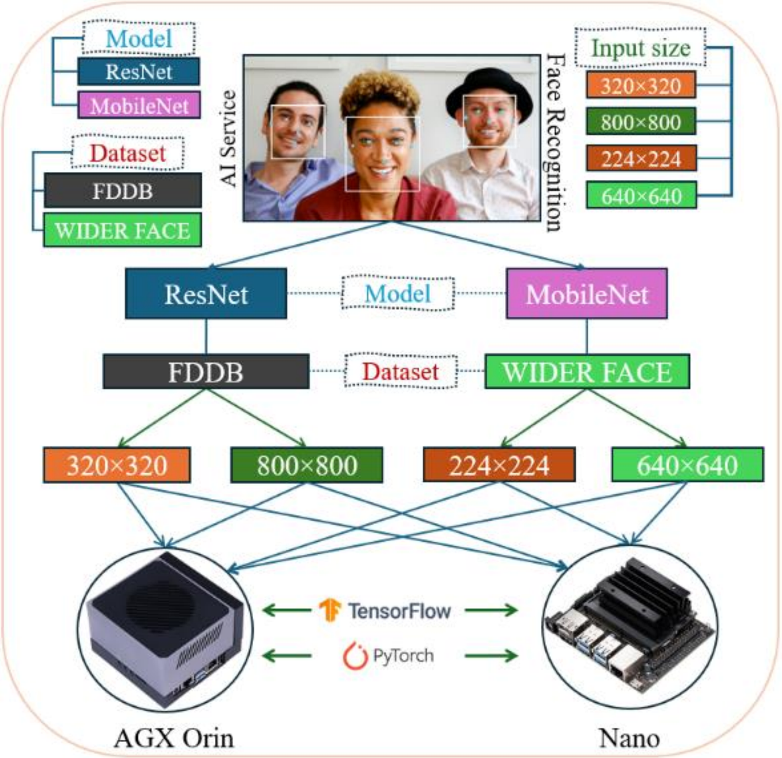

Figure 1. Architecture of Embedded Edge Device and AI Model Deployment

This allows us to assess how each embedded edge device manages various workloads and data complexities. All platforms were configured to maximize CPU performance by utilizing all available cores at their highest frequency, ensuring optimal performance levels.

B. Benchmarking results

We benchmarked the performance of Face Recognition using two models, ResNet and MobileNet, across FDDB and Wider Face datasets with varying input sizes on both the AGX Orin (high-performance setup) and Nano (low-performance setup).

The results, detailing inference time, memory usage, and power consumption, are presented in Table 2.

Algorithm	Model	Dataset	Input size	Inference Time (ms)		Memory (G)		CPU (Power/mW)		GPU (Power/mW)	
				AGX Orin	Nano	AGX Orin	Nano	AGX Orin	Nano	AGX Orin	Nano
Face Recognition	ResNet	FDDB	320x320	12	45	2.0	1.5	350	700	900	350
			800x800	28	95	3.2	2.2	400	800	1100	500
	Mobile Net	Wider Face	224x224	7	25	1.5	1.0	300	600	700	250
			640x640	22	70	2.3	1.8	350	650	850	400

Table 2. Benchmarking results

Inference times varied significantly based on model and input size. For ResNet on FDDB, inference times ranged from 12 ms to 28 ms on AGX Orin, while Nano required between 45 ms and 95 ms. MobileNet on the Wider Face dataset showed faster performance, with AGX Orin achieving 7 ms to 22 ms, and Nano taking between 25 ms and 70 ms.

Memory consumption scaled predictably with input size and model complexity. ResNet on FDDB required up to 3.2G on AGX Orin, while Nano used up to 2.2G. MobileNet, being a more lightweight model, required less memory, with a maximum of 2.3G on AGX Orin and 1.8G on Nano for the larger input size.

Power consumption also demonstrated a clear pattern tied to model and hardware. The AGX Orin displayed higher power usage than the Nano, with CPU power ranging from 300 mW to 400 mW and GPU power up to 1100 mW, depending on the model and input size. In contrast, Nano's CPU and GPU power demands were significantly lower, with CPU power peaking at 800 mW and GPU power at 500 mW.

Overall, AGX Orin consistently provided faster inference times and supported higher memory and power resources, making it more suitable for computationally intensive face recognition tasks. The Nano, though slower, offered a more energy-efficient alternative for less demanding applications, illustrating a trade-off between computational power and energy efficiency.

Conclusion

In this paper, we conducted a benchmarking study of two different Face Recognition models on two Nvidia Jetson devices: the Jetson AGX Orin and the Jetson Nano. Our goal was to assess the performance of these platforms across multiple datasets and input sizes using a Face Detection algorithm. Both devices were configured for maximum performance to achieve higher processing power. Throughout the tests with varying input sizes, we measured memory usage, CPU and GPU utilization, and overall power consumption.

Our experiments revealed that the AGX Orin consistently outperformed the Nano in terms of inference speed, particularly with more complex models such as ResNet and MobileNet. However, Nano demonstrated superior power efficiency and resource management when running simpler models with lightweight datasets.

REFERENCES

[1] Süzen, A.A.; Duman, B.; Şen, B. Benchmark Analysis of Jetson TX2, Jetson Nano and Raspberry PI using Deep CNN. In Proceedings of the 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 26–28 June 2020; pp. 1–5.

[2] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal and S. Marcel, "EdgeFace: Efficient face recognition model for edge devices", IEEE Trans. Biometrics Behav. Identity Sci., vol. 6, no. 2, pp. 158-168, Apr. 2024.

[3] S. Ullah and D.-H. Kim, "Benchmarking jetson platform for 3d point cloud and hyper-spectral image classification," in 2020 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2020, pp. 477–482.

[4] M. Choe, S. Lee, N. M. Sung, S. Jung and C. Choe, "Benchmark Analysis of Deep Learning-based 3D Object Detectors on NVIDIA Jetson Platforms", International Conference on ICT Convergence IEEE Computer Society, pp. 10-12, 2021.

[5] NVIDIA: AGX Orin: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/

[6] NVIDIA: Jetson Nano: https://developer.nvidia.com/embedded/jetson-nanodeveloper-kit