

**MENKARAKTERISASI PEMANFAATAN PERANGKAT KERAS
PADA PERANGKAT EDGE SAAT MENYIMPULKAN MODEL DEEP
LEARNING TERKOMPRESI**

JALUR SCIENTIST – PUBLIKASI JURNAL PENELITIAN

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Teknik Komputer



disusun oleh

AHMAD NAUFAL LABIIB NABHAAN
19.83.0357

Kepada

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2024

**MENKARAKTERISASI PEMANFAATAN PERANGKAT
KERAS PADA PERANGKAT EDGE SAAT MENYIMPULKAN
MODEL DEEP LEARNING TERKOMPRESI**

JALUR SCIENTIST – PUBLIKASI JURNAL PENELITIAN

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Teknik Komputer



disusun oleh

AHMAD NAUFAL LABIIB NABHAAN

19.83.0357

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2024

HALAMAN PERSETUJUAN

JALUR SCIENTIST

**MENGKARAKTERISASI PEMANFAATAN PERANGKAT KERAS PADA
PERANGKAT EDGE SAAT MENYIMPULKAN MODEL DEEP LEARNING
TERKOMPRESI**

yang disusun dan diajukan oleh

Ahmad Naufal Labilb Nabhaan
19.83.0357

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 16 Oktober 2024

Dosen Pembimbing,



Arief Setyanto, S.Si., MT., Ph.D
NIK. 190302036

HALAMAN PENGESAHAN

JALUR SCIENTIST

MENGKARAKTERISASI PEMANFAATAN PERANGKAT KERAS PADA
PERANGKAT EDGE SAAT MENYIMPULKAN MODEL DEEP LEARNING
TERKOMPRESI

yang disusun dan diajukan oleh

Ahmad Naufal Labiib Nabhaan
19.83.0357

Telah dipertahankan di depan Dewan Penguji
pada tanggal 16 Oktober 2024

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Muhamamd Koprari, S.Kom., M.Eng
NIK. 190302454

Wahid Miftahul Ashari, S.Kom., M.T
NIK. 190302452

Arif Setyanto, S.Si., MT., Ph.D
NIK. 190302036



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 16 Oktober 2024

DEKAN FAKULTAS ILMU KOMPUTER



Hanif Al Fatta, S.Kom., M.Kom., Ph.D
NIK. 190302096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Ahmad Naufal Labiib Nabhaan
NIM : 19.83.0357

Menyatakan bahwa Skripsi dengan judul berikut:

Mengkarakterisasi Pemanfaatan Perangkat Keras Pada Perangkat Edge Saat Menyimpulkan Model Deep Learning Terkompresi

Dosen Pembimbing : Arief Seryanto, S.Si., MT., Ph.D

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 16 Oktober 2024

Yang Menyatakan,



Ahmad Naufal Labiib Nabhaan

HALAMAN PERSEMBAHAN

Dengan segala puji syukur dan penuh rasa hormat, penulis ingin menyampaikan persembahan ini kepada:

1. Bapak Ihsanuddin, S.P., dan Ibu Farida Eva Dewiyanti, S.P., orang tua penulis yang tiada henti mendukung penulis secara moral dan materiil.
2. Bapak Prof. Dr. M. Suyanto, M.M., selaku rektor Universitas AMIKOM Yogyakarta.
3. Bapak Arief Setyanto, S.Si., MT., Ph.D., selaku dosen pembimbing penulis, yang telah memberikan arahan dan dukungan yang tiada henti, sehingga penelitian ini dapat terselesaikan.
4. Profesor In Kee Kim, PhD., dan saudari Ting Jiang selaku *team* kolaborasi dari *University of Georgia*, yang telah membantu penelitian ini tanpa pamrih.
5. Saudara Rakandhiya Daanii Rachmanto selaku *author* kedua penelitian ini, yang tiada hentinya memberikan dukungan penulis.
6. Para dosen Universitas AMIKOM Yogyakarta, yang telah memberikan pembelajaran materiil maupun moral selama perkuliahan penulis.
7. Bagja Abdul Basith, Muhammad Zuhdi Fikri Johari, dan Sultan Gemilang Kemadi, sebagai sesama asisten penelitian, yang telah menemani penulis, dari awal penelitian ini berjalan pada 2022.
8. Seluruh sahabat dan rekan-rekan penulis yang telah berbagi pengalaman, inspirasi, dan ilmu kepada penulis.
9. Keluarga besar penulis yang telah memberikan semangat dan motivasi kepada penulis.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT, karena atas rahmat dan karunia-Nya, penulis dapat menyelesaikan penulisan skripsi ini dengan judul "Mengkarakterisasi Pemanfaatan Perangkat Keras Pada Perangkat Edge Saat Menyimpulkan Model Deep Learning Terkompresi". Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi SI Teknik Komputer, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta.

Penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Arief Setyanto, S.Si., MT., Ph.D selaku dosen pembimbing yang telah memberikan arahan, bimbingan, serta masukan yang sangat berharga semenjak penulis menjadi riset asisten beliau hingga membimbing penulisan penelitian ini sebagai skripsi. Terima kasih atas kesabaran, waktu, dan ilmu yang telah diberikan kepada penulis.

Akhir kata, penulis menyadari bahwa penelitian ini belum sempurna. Oleh karena itu, saran dan kritik yang membangun sangat penulis harapkan guna perbaikan di masa yang akan datang.

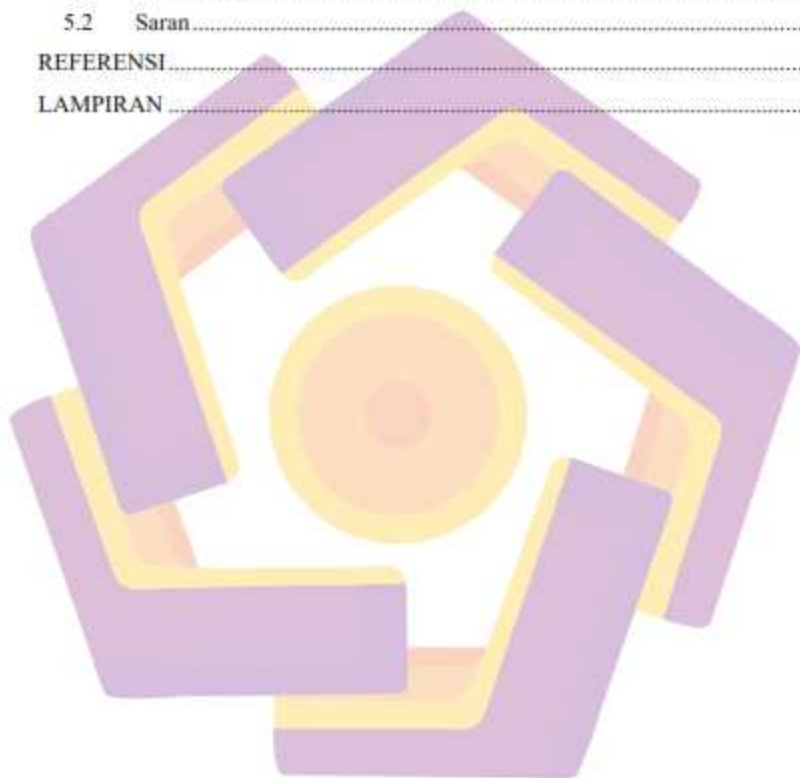
Yogyakarta, 16 Oktober 2024

Penulis

DAFTAR ISI

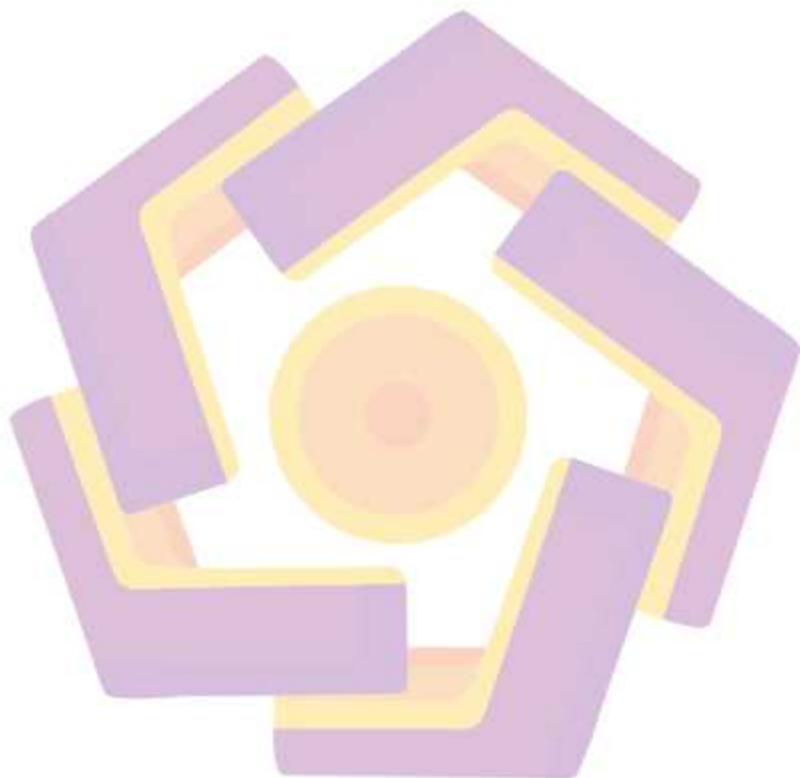
HALAMAN JUDUL	i
HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI	vii
DAFTAR TABEL	ix
DAFTAR GAMBAR	x
DAFTAR LAMPIRAN	xi
IDENTITAS JURNAL	xii
INTISARI	xiii
ABSTRACT	xiv
BAB I PENDAHULUAN	1
BAB II TINJAUAN PUSTAKA	4
2.1 Studi Literatur	4
BAB III METODE PENELITIAN	6
3.1 Pemilihan Perangkat Keras	6
3.2 Pemilihan Model Deep Learning	7
3.3 Skenario Alokasi Memori	8
3.4 Merancang Benchmark Pemanfaatan Perangkat Keras	10
BAB IV HASIL DAN PEMBAHASAN	12
4.1 Faktor Alokasi Memori	12
4.1.1 Alokasi Memori pada RPi4B	12
4.1.2 Alokasi Memori pada J. Nano	13
4.2 Waktu Warm-Up vs Latensi Setelah Warm-Up	14
4.3 Faktor Pemanfaatan Perangkat Keras	15
4.3.1 Pemanfaatan Perangkat Keras pada RPi4B	15
4.3.2 Pemanfaatan Perangkat Keras pada J. Nano	16
4.4 Perbandingan Kecepatan antara Inferensi model DL di Perangkat Edge	17

4.5	Korelasi Latensi Inferensi Deep Learning dengan Alokasi Memori dan Pemanfaatan Perangkat Keras	20
4.5.1	Korelasi Dengan Memori Alokasi.....	21
4.5.2	Korelasi Dengan Pemanfaatan Perangkat Keras	20
4.6	Membandingkan Hasil Karakteristik Perangkat Keras di Riset Terkait	20
BAB V PENUTUP		24
5.1	Kesimpulan.....	24
5.2	Saran.....	24
REFERENSI.....		25
LAMPIRAN		28



DAFTAR TABEL

Tabel 3.1 Daftar Perangkat Edge	6
Tabel 3.2 Daftar Model Deep Learning	8
Tabel 4.4.1 Daftar Latensi Inferensi DL pada RPi4B	17
Tabel 4.4.2 Daftar Latensi Inferensi DL pada J. Nano	18



DAFTAR GAMBAR

Gambar 3.4. Diagram Alur Kerja Benchmark	11
Gambar 4.1.1 Pengaruh Alokasi Memori dengan Latensi dan Pemanfaatan Perangkat Keras pada Eksekusi Model DL pada RPi4B	12
Gambar 4.1.2 Pengaruh Alokasi Memori dengan Latensi dan Pemanfaatan Perangkat Keras pada Eksekusi Model DL pada J. Nano	13
Gambar 4.2 Latensi setelah warm-up model DL vs. pola waktu pemanasan untuk semua model DL di kedua perangkat keras	14
Gambar 4.3.1 Pengaruh Alokasi Memori dengan Latensi dan Pemanfaatan Perangkat Keras pada Eksekusi Model DL pada RPi4B	15
Gambar 4.3.2 Pengaruh Alokasi Memori dengan Latensi dan Pemanfaatan Perangkat Keras pada Eksekusi Model DL pada J. Nano	16
Gambar 4.5 Peta Panas Korelasi antara Latensi Inferensi DL dan Alokasi Memori dan Pemanfaatan Perangkat Keras pada Beberapa Perangkat Edge	21



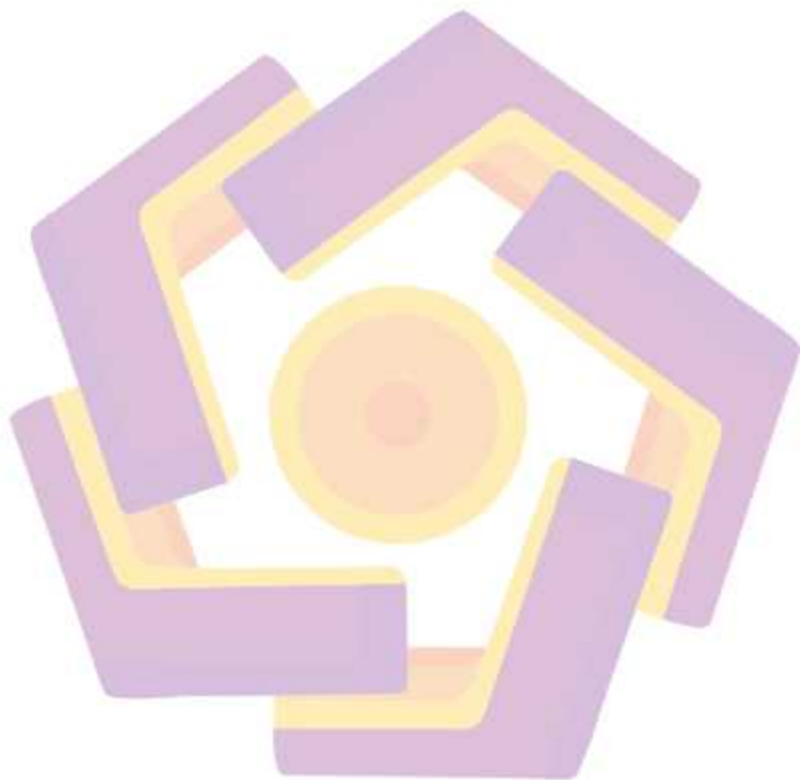
DAFTAR LAMPIRAN

Lampiran 1. *Letter of Acceptance*

28

Lampiran 2. *Lembar Reviewer*

29



IDENTITAS JURNAL

Nama Jurnal	: Matrik : Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
DOI	: https://doi.org/10.33022
Publisher	: LPPM Universitas Bumigora
ISSN	: ISSN (Print) 1858-4144 ISSN (Online) 2476-9843
Indexing	: Sinta
Link Sinta	: https://sinta.kemdikbud.go.id/journals/profile/3962



INTISARI

Edge AI memungkinkan implementasi AI di dekat sensor. Model Deep Learning (DL) telah berhasil menangani tugas klasifikasi gambar dengan kinerja yang luar biasa. Namun, persyaratan mereka akan sumber daya komputasi yang besar, menghambat implementasi pada perangkat edge. Untuk memungkinkan penerapan model DL pada perangkat edge, mengompresi model adalah tugas penting. Kuantisasi pasca pelatihan (PTQ) adalah teknik kompresi, untuk mengurangi representasi bit dari parameter bobot model. Studi ini melihat dampak alokasi memori terhadap latensi model DL terkompresi pada Raspberry Pi 4 Model B (RPi4B) dan NVIDIA Jetson Nano (J. Nano). **Tujuan dari penelitian ini adalah memahami** pemanfaatan perangkat keras seperti central processing unit (CPU), Graphical Processing Unit (GPU) dan memori. **Penelitian ini berfokus pada metode kuantitatif**, dengan mengontrol alokasi memori dan mengukur waktu warm-up, latensi, pemanfaatan CPU dan GPU. Perbandingan kecepatan antara inferensi model DL pada RPi4B dan J. Nano. Korelasi antara pemanfaatan perangkat keras versus berbagai latensi inferensi DL yang diamati dalam makalah ini. **Menurut percobaan kami, kami menyimpulkan bahwa** alokasi memori yang lebih kecil menyebabkan latensi tinggi pada RPi4B dan J. Nano. Pemanfaatan CPU pada RPi4B. Pemanfaatan CPU di RPi4B meningkat seiring dengan alokasi memori, namun fakta sebaliknya ditunjukkan pada J. Nano, karena komputasi utama dilakukan oleh GPU pada perangkat. Dalam hal komputasi, Ukuran DL yang lebih kecil dan representasi bit yang lebih kecil menghasilkan inferensi yang lebih cepat (latensi rendah) sementara representasi bit yang lebih besar pada model DL yang sama menyebabkan latensi yang lebih tinggi.

Kata kunci: Deep learning, Perangkat edge, Pemanfaatan perangkat keras, Alokasi memori, Kuantisasi pasca-pelatihan

ABSTRACT

Edge AI enabled implementation of AI near to the sensors. Deep Learning (DL) Model has successfully tackled image classification task with remarkable performance. However, their requirements of huge computing resources, hindering the implementation on edge devices. To allow implementation of DL model on the edge devices, compressing the model is an essential task. Post training quantization (PTQ) is a compression technique, to reduce the bit representation of the model weight parameters. This study looks at the impact of memory allocation to the latency of compressed DL models on Raspberry Pi 4 Model B (RPi4B) and NVIDIA Jetson Nano (J. Nano). **The objective of this research is** understanding the hardware utilization such as central processing unit (CPU), Graphical Processing Unit (GPU) and memory. **This study focused on the quantitative method,** by controlling the memory allocation and measuring the warm-up time, latency, CPU and GPU utilization. Speed comparison among inference of DL models on RPi4B and J. Nano. The correlation between hardware utilization versus the various DL inference latency observed in this paper. **According to our experiment, we concluded that** smaller memory allocation led to high latency on both RPi4B and J. Nano. CPU utilization on RPi4B. CPU utilization in RPi4B increase along with the memory allocation, however the opposite fact shown on J. Nano, since the main computation carried out by the GPU on the device. In terms of computation, the smaller DL Size and smaller bit representation lead to faster inference (low latency) while bigger bit representation on the same DL model lead higher latency.

Keyword: Deep learning, Edge devices, Hardware utilization, Memory allocation, Post-training quantization