MENGKARAKTERISASI KOMPRESI MODEL *DEEP LEARNING* DENGAN *POST-TRAINING QUANTIZATION* PADA *EDGE DEVICES*

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika

disusun oleh
RAKANDHIYA DAANII RACHMANTO
20.11.3304

Kepada

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2024

MENGKARAKTERISASI KOMPRESI MODEL *DEEP LEARNING* DENGAN *POST-TRAINING QUANTIZATION* PADA *EDGE DEVICES*

JALUR SCIENTIST

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika

disusun oleh
RAKANDHIYA DAANII RACHMANTO
20.11.3304

Kepada

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2024

HALAMAN PERSETUJUAN

JALUR SCIENTIST

MENGKARAKTERISASI KOMPRESI MODEL *DEEP LEARNING* DENGAN *POST-TRAINING QUANTIZATION* PADA *EDGE DEVICES*

yang disusun dan diajukan oleh

Rakandhiya Daanii Rachmanto
20.11.3304

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 19 Juni 2024

Dosen Pembimbing,

Arief Setyanto, S.Si., M.T., Ph.D.
NIK. 190302036

HALAMAN PENGESAHAN

JALUR SCIENTIST

MENGKARAKTERISASI KOMPRESI MODEL *DEEP LEARNING* DENGAN *POST-TRAINING QUANTIZATION* PADA *EDGE DEVICES*

yang disusun dan diajukan oleh

Rakandhiya Daanii Rachmanto
20.11.3304

Telah dipertahankan di depan Dewan Penguji
pada tanggal 19 Juni 2024

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Theopilus Bayu Sasongko S.Kom., M.Eng. NIK. 190302375	
Bayu Setiaji, M.Kom. NIK. 190302216	
Arief Setyanto, S.Si., MT., Ph.D. NIK. 190302036	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 19 Juni 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 : Rakandhiya Daanii Rachmanto
NIM : 20.11.3304

Menyatakan bahwa Skripsi dengan judul berikut:

Mengkarakterisasi Kompresi Model *Deep Learning* dengan *Post-Training Quantization* pada *Edge Devices*

Dosen Pembimbing : Arief Setyanto, 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 ang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 19 Juni 2024

Yang Menyatakan,

Rakandhiya Daanii Rachmanto

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa, yang telah melimpahkan rahmat-Nya sehingga penulis dapat menyelesaikan penelitian berjudul "Mengkarakterisasi Kompresi Model *Deep Learning* dengan *Post-Training Quantization* pada *Edge Devices*" dengan baik. Penyusunan laporan penelitian ini dilakukan sebagai salah satu persyaratan menyelesaikan studi program sarjana (S1) Informatika di Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta.

Penulis menerima bimbingan, dukungan materiil dan moral dari berbagai pihak yang memungkinkan kelancaran berlangsungnya penelitian dan penulisan laporan. Penulis ingin berterima kasih yang sebesar-besarnya pada pihak-pihak berikut yang telah membantu penulis, antara lain:

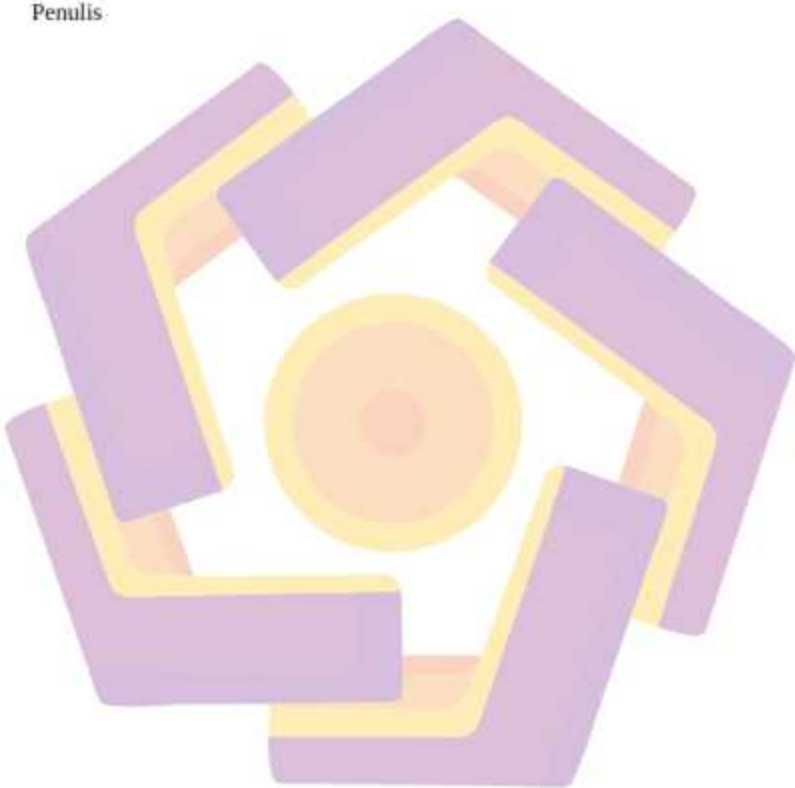
1. Bapak Sidiq Rachmanto, S.Pi., dan Ibu Iryani Hajijah Sarlata, S.Pi., orang tua penulis yang tiada henti mendukung penulis secara moral maupun materiil.
2. Bapak Arief Setyanto, S.Si, MT., Ph.D., selaku pembimbing penulis semenjak memulai penelitian pada 2022, yang telah meluangkan waktu untuk memandu dan berdiskusi dengan penulis.
3. Ahmad Naufal Labiib Nabhaan dari Universitas Amikom Yogyakarta, Dr. In Kee Kim, Ting Jiang dan Zaki Indra Sukma dari *University of Georgia*, sebagai kolega dalam penelitian ini, yang telah meluangkan waktu untuk bekerja sama dan berdiskusi.
4. 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.
5. Seluruh staff dan dosen program studi Informatika di Universitas Amikom Yogyakarta, atas segala ilmu dan dukungan yang telah diberikan selama masa studi.

Penulis menyadari bahwa laporan ini tidak luput dari kesalahan dan kelemahan. Dengan demikian, kritik dan saran dari pembaca sangat diapresiasi oleh penulis, untuk dikembangkan di masa mendatang. Akhir kata, dengan peneli-

tian ini, penulis berharap dapat memberikan kontribusi baik dan sebagai rujukan bagi pembaca yang ingin mendalami topik optimisasi menjalankan kecerdasan buatan di perangkat *edge*.

Yogyakarta, 19 Juni 2024

Penulis



DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PERSETUJUAN ii
HALAMAN PENGESAHAN iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI iv
KATA PENGANTAR v
DAFTAR ISI vii
DAFTAR TABEL ix
DAFTAR GAMBAR x
DAFTAR LAMPIRAN xi
INTISARI xii
ABSTRACT xiii
BAB I PENDAHULUAN 1
BAB II TINJAUAN PUSTAKA 5
2.1. Studi Literatur 5
2.2. Dasar Teori 6
BAB III METODE PENELITIAN 10
3.1. Model dan Dataset 10
3.2. Perangkat Lunak 11
3.3. Perangkat *Edge* 11
3.4. Konfigurasi 12
3.5. Prosedur Pengukuran 13
3.5.1. Pengukuran ketika Kompresi 13
3.5.2. Pengukuran Model (Sebelum maupun Setelah Kompresi) 13
3.6. Metrik Kinerja 14
BAB IV HASIL DAN PEMBAHASAN 15
4.1. Karakterisasi Model Dasar (Base) 15

4.1.1. Latensi Inferensi 15
4.1.2. Utilisasi Sumber Daya 16
4.2.1. *Overhead* 17
4.2.2. Utilisasi Sumber Daya 19
4.3. Karakterisasi Model Terkuantisasi 20
4.3.1. Pengurangan Ukuran Model 21
4.3.2 Perubahan Akurasi 21
4.3.3. Utilisasi Sumber Daya 22
4.3.4 Pengurangan Latensi 24
BAB VPENUTUP 26
5.1. Kesimpulan 26
5.2. Saran 26
REFERENSI 28
LAMPIRAN 34

DAFTAR TABEL

Tabel 1. Informasi tentang Model 9
Tabel 2. Perangkat *Edge* 11

DAFTAR GAMBAR

Gambar 1. Alur sederhana kuantisasi TensorRT 8
Gambar 2. Alur kuantisasi INT8 9
Gambar 3. Alur percobaan pengukuran ketika konversi 12
Gambar 4. Alur percobaan pengukuran ketika inferensi 13
Gambar 5. Latensi model *base* di ORIN dan XAVIER 14
Gambar 6. Utilisasi sistem model awal di kedua perangkat 16
Gambar 7. Overhead (waktu kompresi) masing-masing konfigurasi kompresi 17
Gambar 8A. Mob-S FP16 18
Gambar 9B. Mob-S INT8 18
Gambar 10C. Eff-B3 FP16 18
Gambar 11D. Eff-B3 INT8 18
Gambar 12E. D201 FP16 18
Gambar 13F. D201 INT8 18
Gambar 14. Perbandingan ukuran model sebelum dan setelah kuantisasi 20
Gambar 15. Perbandingan akurasi model sebelum dan setelah kuantisasi 21
Gambar 16. Perbandingan utilisasi sistem model sebelum dan sesudah kuantisasi 22
Gambar 17. Pengurangan latensi model yang terkuantisasi 23

DAFTAR LAMPIRAN

Letter of Acceptance 34

Lembar *Reviewer* 34

INTISARI

Kecerdasan buatan di edge semakin banyak diadopsi karena perkembangan pesat *deep learning* dan kecerdasan buatan (*artificial intelligence*, AI). Pada saat yang sama, karena model AI dengan cepat bertambah besar dan kompleks, perangkat edge yang memiliki keterbatasan sumber daya menghadapi tantangan yang signifikan dalam menjalankan model yang kompleks tersebut. Kompresi model, khususnya post-training quantization (PTQ), menawarkan pendekatan yang layak dengan mengurangi ukuran model dan kebutuhan sumber daya, sehingga model-model ini lebih cocok untuk diterapkan pada perangkat edge. Namun, terlepas dari signifikansinya, efek kompresi model pada perangkat edge belum dieksplorasi secara menyeluruh. Kami mengatasi kesenjangan ini dengan mengkarakterisasi post-training quantization secara menyeluruh pada perangkat edge yang dipercepat. Kami menggunakan enam model *deep learning* yang berbeda dengan ukuran yang bervariasi (dalam MB) dan kebutuhan sumber daya. Pertama-tama, kami mengkarakterisasi *post-training quantization* di perangkat pada perangkat edge. Selanjutnya, kami melakukan karakterisasi terperinci tentang kinerja dan perilaku model *deep learning* yang dikuantisasi dengan mode presisi yang berbeda. Kami membahas manfaat *post-training quantization*, termasuk pengurangan ukuran model, pengurangan latensi inferensi, dan penurunan konsumsi sumber daya. Kami juga memberikan analisis terperinci tentang kelemahan model yang telah dikuantisasi, dengan fokus pada pengurangan akurasi inferensi.

Kata kunci: *Edge* AI, Perangkat *Edge*, Kompresi *on-device*, *Post-Training Quantization*

ABSTRACT

Edge AI has increasingly been adopted due to the rapid development of deep learning and AI. At the same time, as AI models quickly grow in size and complexity, resource-constrained edge devices face significant challenges in executing such complex models. Model compression, particularly post-training quantization, offers a viable approach by reducing model size and resource demands, making these models more suitable for the deployment on edge devices. However, despite its significance, the effects of model compression on edge devices have yet to be thoroughly explored. We address this gap by thoroughly characterizing post-training quantization on accelerated edge devices. We use six different deep learning models with varied sizes (in MB) and resource demands. We first characterize on-device post-training quantization on edge devices. Subsequently, we perform a detailed characterization of the performance and behaviors of quantized deep learning models with different precision modes. We discuss the benefits of post-training quantization, including reduced model size, improved inference latency, and decreased resource consumption. We also provide a detailed analysis of the downside of the quantized models, focusing on the reduction of their inference accuracy.

Keyword: *Edge AI, Edge Devices, On-Device Compression, Post-Training Quantization*