BAB V
PENUTUP

5.1. Kesimpulan

Dalam penelitian ini, kami melakukan studi karakterisasi menyeluruh tentang kompresi PTQ pada perangkat *edge*. Kami berfokus pada metode kompresi PTQ karena metode ini ringan, hemat sumber daya, dan tidak memerlukan tahap pelatihan tambahan yang mahal, sehingga cocok untuk kasus penggunaan *edge* AI. Studi karakterisasi ini dilakukan pada perangkat *edge* yang memiliki keterbatasan sumber daya di dunia nyata, yaitu XAVIER dan ORIN, yang ditandai dengan *core* CPU dan kapasitas memori yang terbatas, serta hanya dilengkapi dengan 384 hingga 1024 *core* GPU. Untuk karakterisasi ini, kami menggunakan enam model DL yang banyak digunakan dengan berbagai ukuran, dari model kecil seperti Mob-S/L hingga model yang lebih besar seperti D169/201. Kami juga mengeksplorasi dua mode presisi yang berbeda, misalnya FP16 dan INT8, yang dapat mempengaruhi ukuran, akurasi, latensi, dan waktu kompresi keseluruhan model yang dikuantisasi pada perangkat.

Kami mengevaluasi *overhead* kompresi (waktu) dan variasi pemanfaatan sumber daya pada perangkat *edge* selama kompresi pada perangkat dengan PTQ. Selanjutnya, dengan model yang dikuantisasi dalam dua mode presisi yang berbeda, kami mengidentifikasi manfaatnya, termasuk pengurangan ukuran model, penurunan konsumsi sumber daya, dan peningkatan latensi inferensi. Selain itu, kami mengidentifikasi tantangan yang terkait, seperti penurunan akurasi model yang dikuantisasi.

5.2. Saran

Meski demikian, penulis menyadari penelitian ini masih dapat dikembangkan lebih lanjut. Salah satu pengukuran yang belum kami lakukan adalah pengukuran konsumsi energi, yang merupakan metrik penting lain dalam *edge* computing. Penelitian ini dapat pula dikembangkan dengan meneliti efek kompresi model pada tugas-tugas lain seperti deteksi objek, segmentasi dan sejenisnya, maupun percobaan kompresi pada struktur lain (berbasis *transformer*, R-

CNN, dan lainnya). Dapat pula dengan percobaan perbandingan antara beberapa teknik kompresi serta konfigurasinya.