INTISARI

Optimasi Performa Model pada Vision Transformer dan YOLO melalui Kuantisasi

Oleh

Afif Alauddin Falah
20/46215/PA/20147

Penelitian ini bertujuan untuk mengidentifikasi pengaruh penerapan kuantisasi dalam optimasi model klasifikasi gambar Vision Transformer dan YOLO dalam sumber daya komputasi yang terbatas. Penggunaan metode kuantisasi dalam penelitian ini adalah *Post-Training Quantization* versi *Dynamic Quantization* dan *Static Quantization*. Dataset yang digunakan adalah CIFAR-100 yang merupakan dataset publik, memiliki total gambar sebanyak 60000 gambar dengan rincian 50000 dataset pelatihan dan 10000 dataset pengujian. Hasil penelitian menunjukkan bahwa kuantisasi memberikan optimasi terutama pada ukuran model yang dapat dipangkas hingga 4x lebih kecil. Pada model Vision Transformer, model mengalami percepatan pada durasi evaluasi dengan menjadi 1,09 hingga 1,1 kali lebih cepat, meskipun terjadi sedikit penurunan akurasi dari 86,16% ke 80,15%. Sedangkan pada model YOLO, model justru mengalami perlambatan pada durasi evaluasi hingga 2 kali lebih lambat, serta mengalami penurunan akurasi yang signifikan. Temuan ini menegaskan bahwa penerapan kuantisasi dapat bekerja secara optimal pada model tertentu dalam sumber daya komputasi yang terbatas, tetapi pada model tertentu akan lebih efektif apabila memiliki sumber daya komputasi yang lebih besar.

Kata kunci: *Deep Learning, Convolutional Neural Network,* Kuantisasi, Optimasi Model, Vision Transformer, *You Only Look Once* (YOLO)

ABSTRACT

Optimization of Model Performance in Vision Transformer and YOLO through Quantization

By

Afif Alauddin Falah
20/46215/PA/20147

This study aims to identify the effect of applying quantization in the optimization of Vision Transformer and YOLO image classification models in limited computing resources. The use of quantization methods in this study is Post-Training Quantization with two versions, Dynamic Quantization and Static Quantization. The dataset used is CIFAR-100 which is a public dataset, has a total of 60000 images with 50000 images as training datasets and 10000 images as testing datasets. The results show that quantization provides optimization, especially in the size of the model which can be trimmed up to 4x smaller. In the Vision Transformer model, the model accelerates the evaluation duration by being 1.09 to 1.1 times faster, although there is a slight decrease in the accuracy from 86.16% to 80.15%. Whereas in the YOLO model, the model actually slowed down the evaluation duration to 2 times slower, and experienced a significant decrease in accuracy. This findings confirms that the application of quantization can work optimally on certain models under limited computational resources, but on certain models it will be more effective when having larger computational resources.

Keywords**:** *Deep Learning, Convolutional Neural Network, Model Optimization, Quantization*, *Vision Transformer*, *You Only Look Once (YOLO)*