INTISARI

Pada era digital saat ini, penerapan model bahasa berskala besar dengan arsitektur Transformer dapat ditemukan dengan mudah dalam berbagai masalah kebahasaan, seperti penerjemahan teks, penjawaban pertanyaan, dan analisis sentimen. Sayangnya, dibutuhkan biaya dan beban komputasi yang besar untuk melatih model bahasa. Sebagai alternatif dari proses pelatihan model bahasa, dapat dilakukan pengembangan model bahasa melalui proses penggabungan model-model yang sudah ada dengan metode *Weight Average*.

Penelitian ini mengkaji penggabungan model bahasa dengan mengaplikasikan metode *Weight Average* pada beberapa model turunan Mistral-7B untuk meningkatkan kemampuan model dalam menyelesaikan masalah dalam bahasa Indonesia. Untuk mengurangi sumber daya yang dibutuhkan untuk pengujian model bahasa, dilakukan juga proses kuantisasi model bahasa dengan menggunakan *library* Exllamav2. Hasil evaluasi pada model bahasa yang telah digabungkan menunjukkan bahwa proses penggabungan model bahasa dapat menghasilkan model yang lebih baik daripada model yang tidak digabungkan. Pada evaluasi pembangkitan menggunakan *dataset* GSM8K, model bahasa gabungan terbaik mampu menghasilkan akurasi sebesar 0,462. Sementara itu, pada evaluasi klasifikasi menggunakan pembelajaran beberapa tembakan (*few-shot learning*), model bahasa gabungan terbaik lainnya mampu menghasilkan akurasi sebesar 0,776 pada *dataset* IndoNLI, 0,862 pada *dataset* analisis sentimen NusaX, dan 0,682 pada *dataset* analisis emosi IndoNLU. Di lain sisi, model bahasa yang telah dikuantisasi hanya membutuhkan 62,13% dari VRAM model yang tidak dikuantisasi dan 32,66% memori penyimpanan model yang tidak dikuantisasi dengan hanya memunculkan perubahan *perplexity* sebesar 1,23% sampai dengan 1,81% saja. Hal ini menunjukkan potensi model bahasa yang dikembangkan tersebut untuk diimplementasikan ke dalam sistem yang memerlukan pengolahan bahasa Indonesia.

Kata Kunci : model bahasa, pemrosesan bahasa alami, penggabungan model, kecerdasan buatan, kuantisasi model

ABSTRACT

In today's digital era, the application of large-scale language models with the Transformer architecture can be easily found in various linguistic problems, such as text translation, question answering, and sentiment analysis. Unfortunately, it requires a large cost and computational burden to train a language model. As an alternative to the language model training process, language model development can be carried out through the process of combining existing models using the Weight Average method.

This research examines the combination of language models by applying the Weight Average method to several Mistral-7B derivative models to improve the model's ability to solve problems in Indonesian. To reduce the resources required for language model testing, a language model quantization process was also carried out using the Exllamav2 library. The evaluation results on combined language models show that the process of combining language models can produce better models than uncombined models. In the generation evaluation using the GSM8K dataset, the best combined language model was able to produce an accuracy of 0.462. Meanwhile, in classification evaluation using few-shot learning, the other best language model was able to produce an accuracy of 0.776 on the IndoNLI dataset, 0.862 on the NusaX sentiment analysis dataset, and 0.682 on the IndoNLU emotion analysis dataset. On the other hand, the language model that has been quantized only requires 62.13% of the VRAM of the unquantized model and 32.66% of the storage memory of the unquantized model with only a perplexity change of 1.23% to 1.81%. This shows the potential of the developed language model to be implemented in systems that require Indonesian language processing.

Keywords : language models, natural language processing, model merging, artificial intelligence, model quantization