Large Malaysian Language Model Based on Mistral for Enhanced Local Language Understanding

Husein Zolkepli[*] Aisyah Razak[†] Kamarul Adha[‡]
Ariff Nazhan[§]

January 19, 2024

Abstract

In this paper, we present significant advancements in the pretraining of Mistral 7B, a large-scale language model, using a dataset of 32.6 GB, equivalent to 1.1 billion tokens. We explore the impact of extending the context length, releasing models with context lengths of 4096 and 32768 tokens, and further refining performance with a specialized 16k context length instruction-tuned model, we called it Malaysian Mistral.

Our experiments demonstrate the efficacy of prolonged pretraining and the influence of extended context lengths on Mistral 7B's language understanding capabilities. Additionally, we release a model specifically tuned with a 16k context length instruction, showcasing its potential for capturing nuanced language intricacies.

Furthermore, our research contributes to the benchmarking of Mistral 7B against prominent language models, including ChatGPT3.5 and Claude 2.1. We present compelling results indicating Mistral 7B's superior performance on tatabahasa test sets, particularly when fine-tuned with instructions. These findings underscore the model's capacity to outperform existing state-of-the-art language models in tasks requiring intricate understanding of linguistic structures and conventions.

1 Introduction

The evolution of artificial intelligence (AI) has witnessed transformative breakthroughs, from the introduction of "Attention is All You Need" with the Transformer architecture, to subsequent advancements like GPT-2, and the revolutionary ChatGPT. These models have sparked immense interest and curiosity

[*] husein@mesolitica.com
[†] aisyahrazak171@gmail.com
[‡] kamarul.adha360@gmail.com
[§] ariffnzhn@gmail.com

in the AI landscape, pushing the boundaries of natural language understanding and generation.

In response to this dynamic landscape, Mistral AI emerged, unveiling its initial model, Mistral 7B. Notably, Mistral 7B showcased superior performance, surpassing benchmarks set by Llama 2 13B across various tasks and even outperforming Llama 1 34B on numerous benchmarks. Impressively, it approached the performance of CodeLlama 7B on code-related tasks while maintaining proficiency in English language tasks. However, an identified gap in its capabilities was the limited understanding of Malaysian context.

- **Fine-tuning Mistral 7B:** Utilizing the computational power of 8x A100 GPUs on a Standard_ND96asr_v4 Azure instance, we conducted extensive fine-tuning on Mistral 7B. The process involved training the model using context lengths of 4096 and 32768 on a substantial 32.6 GB Malaysian context dataset.

- **Multi-turn Instruction-Tuned Model:** Crafting local context multi-turn chat dataset using ChatGPT3.5, ChatGPT4, and Llama2 70B, we employed Meural Machine Translation to translate the dataset. This approach enhances Malaysian Mistral's proficiency in multi-turn conversations, contributing to its adaptability across a wide range of local context tasks and coding.

2 Main Body

Your main content goes here. Use sections and subsections as needed.

2.1 Example Subsection

An example subsection.

3 Conclusion

Your conclusion text.