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

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

January 20, 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 16384 context length instruction-tuned model, we called it Malaysian Mistral.

Our experiments demonstrate the efficacy of continue pretraining and the influence of extended context lengths on Mistral 7B's language understanding capabilities. Additionally, we release a model specifically tuned with a 16384 context length instruction, showcasing its potential for capturing nuanced language intricacies.

Furthermore, our research contributes to the benchmarking of Malaysian Mistral against prominent language models, including ChatGPT3.5 and Claude 2.1. We present compelling results indicating Malaysian Mistral's superior performance on Tatabahasa (Malay grammar) test set, particularly when fine-tuned with instructions.

All models released at HuggingFace Mesolitica Malaysian Mistral 7B Collection.

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 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 multiturn 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.

[*] husein@mesolitica.com

[†] aisyahrazak171@gmail.com

[‡] kamarul.adha360@gmail.com

[§] ariffnzhn@gmail.com

2 Related Work

2.1 English-Centric Bias in Large Language Models

The majority of open-source Large Language Models (LLMs) exhibit a significant bias towards the English language, with minimal representation and training on Malay datasets. An analysis of the widely utilized Common Crawl dataset reveals a mere 0.0742% contribution from the Malay language based on CC-MAIN-2023-50 index [?]. This English-centric bias poses a substantial challenge for applications requiring robust language understanding in Malay, prompting the need for dedicated research and development in this domain.

2.2 Existing Malay Language Models

While the Malay natural language processing (NLP) landscape lacks a dedicated Large Language Model, notable efforts have been made by Mesolitica in the development of specific Malay language models. Notable among these are the Malay Causal Language Model, Malay T5, and Malay Masked Language Model. These models, while contributing significantly to the Malay NLP toolkit [?], are distinct from comprehensive Large Language Models and have limitations in capturing extensive context and nuances.

2.3 Absence of a Malay Large Language Model

Despite the existence of specialized models for Malay, a notable gap remains in the absence of a dedicated Malay Large Language Model. The current state of affairs hinders the progress of research and applications requiring a deeper understanding of the Malay language. A comprehensive Large Language Model for Malay is essential to bridge this gap, enabling advancements in various natural language processing tasks and fostering the inclusive representation of Malay in the AI landscape.

3 Pre-Training Procedure

3.1 Public Data

3.1.1 MS Wikipedia

3.1.2 Malay Language study articles

3.1.3 Malaysia Government public documents

3.1.4 Malaysia online articles

3.2 Deduplicating Data

https://github.com/malaysia-ai/dedup-text-dataset?tab=readme-ov-filetext-dedup

3.3 Postprocessing Data

https://github.com/malaysia-ai/dedup-text-dataset?tab=readme-ov-filepostprocessing

3.4 Tokenizing Data

https://github.com/malaysia-ai/dedup-text-dataset/tree/main/mistral

3.5 Pre-Training phase

3.5.1 4096 context length

DeepSpeed Zero3, batch size 20 and bfloat16, constant learning rate 2e-5.

https://github.com/mesolitica/malaya/tree/5.1/session/mistral7b-4096-context-length

3.5.2 32768 context length

We use the latest checkpoint from 4096 context length and trained on random 10% from the dataset.

DeepSpeed Zero3, batch size 3 and bfloat16, constant learning rate 2e-5.

https://github.com/mesolitica/malaya/tree/5.1/session/mistral7b-32768-context-length

4 Supervised Instruction Fine-tuning Procedure

4.1 Generating Instruction Dataset

4.1.1 OpenQA MS Wikipedia

include some data.

4.1.2 OpenQA Malaysia Articles

include some data.

4.1.3 Malay Instruction with Malaysian context

include some data.

4.1.4 Malay UltraChat

include some data.

4.1.5 Synthetic Malay CommonSense

include some data.

4.1.6 Coding Dataset

include some data.

4.2 Finetuning Phase

Use Mistral chat template, mention the chat template below,

DeepSpeed Zero3, batch size 4 with gradient accumulation 9, 16384 context length, constant learning rate 2e-5.

https://github.com/mesolitica/malaya/tree/5.1/session/mistralinstructions-7b-16384-context-length

4.3 Result Finetuning

4.3.1 Multiturn Malaysian context QA

4.3.2 Multiturn Coding QA

4.3.3 Translation low language

4.3.4 Malay instruction

5 Evaluation

We use Tatabahasa dataset, gathered from https://tatabahasabm.tripod.com/latih/latih.htm, contain 349 questions.

We published at https://huggingface.co/spaces/mesolitica/malay-llm-leaderboard

6 Acknowledgement

Special thanks to Malaysia-AI volunteers.

7 Conclusion

able to reduce research gap.

References

[1] Common Crawl. (2023). CC-MAIN-2023-50 Index. Retrieved from https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

[2] Husein, Zolkepli. (2018). *Malaya: Natural-Language-Toolkit library for Bahasa Malaysia, powered by Deep Learning PyTorch.* GitHub repository. Retrieved from https://github.com/mesolitica/malaya