VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGIES

NGÔ ĐỨC HUY

XXX

SUMMER INTERNSHIP REPORT

Major: Computer Science

Supervisor: Dr. Tạ Việt Cường

HANOI – 2024

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr. Tạ Việt Cường, for his guidance, support, and encouragement throughout the internship. His valuable advices and feedbacks have helped me a lot in completing the research and experiments.

I would also like to the Faculty of Information Technology, University of Engineering and Technology, VNU, and HMI lab for providing me with the opportunity to participate in the summer internship program.

ABSTRACT

Abstract: This is the report on the research and experiments on Large Language Models (LLMs) during the summer internship at the HMI Lab, Faculty of Information Technology, University of Engineering and Technology, VNU, under the guidance of Dr. Tạ Việt Cường.

Keywords: *1, 2*

Chapter 1.

Introduction

1.1. Large Language Models and Small Language Models

Large Language Models (LLMs) are machine learning models that are capable of processing and understanding natural language through learning from a large amount of text data. They can predict, generate text, answer questions, translate, and perform many other complex language tasks. Thanks to the computational power and large-scale data, these models have made significant breakthroughs in Natural Language Processing (NLP).

LLMs can come in various sizes, from a few million to hundreds of billions of parameters. The larger the model, the better the performance, but also the higher the computational cost. For example, GPT-3, one of the largest LLMs, has 175 billion parameters, which requires a large amount of memory and computational resources to train and deploy.

Therefore, one of the challenges when using LLMs is the high computational cost and resource requirements. To address this issue, Small Language Models (SLMs) are introduced, which are smaller versions of LLMs that could be aqquired by creating from scratch or through processes such as knowledge distillation, model finetuning, and other training techniques. These SLMs still retain the core language capabilities like other larger model but with lower resource requirements, faster query processing speed, at a small exchanched cost of performance.

The research and use of small language models is a new and relatively important research direction in the field of NLP, helping to optimize the performance and cost of NLP applications in practice, or in training step, it may reduce the amount of data or energy required, which is beneficial for the environment. This could be the way for small organizations or individuals lacking resources to enter the field of language models research and development.

1.2. Research Objectives

This report presents the research and experiments on a small language model based on PhoGPT, a state of the art LLMs for Vietnamese. The process including explore the process of modifying the model, evaluating the performance of the model in specific

evaluations. In overall, the report aims to answer the following questions:

- Can we modify a LLM like PhoGPT to create a smaller version with fewer parameters and lower resource requirements?
- What is the tradeoff between the performance and the resource requirements of the SLM compared to the original LLM?

Chapter 2.

Nghiên cứu liên quan

2.1. PhoGPT

PhoGPT là một series model LLM open-source dành cho Tiếng Việt do VinAIResearch nghiên cứu và phát triển hoàn chỉnh. Trong đó PhoGPT-Chat-4B là model chính với 3.7 tỉ tham số. Model này đã đạt được kết quả tốt trên nhiều bài đánh giá cho tiếng Việt.

TABLE here !

2.2. Tinh chỉnh mô hình

Kết quả thực nghiệm trong việc xây dựng mô hình PanGu-π-1.5B-Pro dành cho tiếng Trung từ Huawei Noah's Ark Lab đã chỉ ra một phương pháp tinh chỉnh mô hình hiệu quả. Mô hình gốc được sử dụng là PanGu-π-7B với 7 tỉ tham số, sau đó được tinh chỉnh thành mô hình nhỏ hơn với 1.5 tỉ tham số.

Chapter 3.

Phương pháp nghiên cứu ?

3.1. Research Question 1:

3.2. Research Question 2

Chapter 4.

Thực nghiệm

4.1. blabla

Chapter 5.

Kết luận

5.1. blabla