INTEL-UNNATI INDUSTRIAL TRAINING 2024
Code:IUCP-2024

Running GenAI on Intel AI laptops and Simple LLM Interference on CPU and fine-tuning of LLM models using Intel OpenVINO

Aiswarya Rahul, Jyothsna Sara Abey, Cinta Susan Thomas, Jacksilin P Titus, and Tebin Philip George

Saintgits Group of Institutions, Kottayam, Kerala

Abstract: This project focuses on the optimization of Generative AI (GenAI) and Large Language Model (LLM) inference on Intel AI laptops using the Intel OpenVINO toolkit. The goal is to develop a comprehensive solution that encompasses environment setup, CPU-based LLM inference optimization, and model fine-tuning for specific applications like chatbots. The solution involves setting up a suitable environment with Python, Hugging Face libraries, and OpenVINO. Pre-trained models are converted to the OpenVINO IR format for efficient CPU inference. The project also covers the deployment of these models for optimized performance. A custom chatbot is created to illustrate the practical use of fine-tuned models.This project demonstrates the potential of combining GenAI techniques with Intel's optimization tools to deliver high-performance, cost-effective AI solutions.

Keywords: Generative AI (GenAI), Large Language Models (LLMs), Intel OpenVINO Toolkit, Model Optimization, CPU Inference, Model Fine-Tuning, Hugging Face Transformers, Custom Chatbot Development

1 Introduction

Given the possibility and the speed at which Artificial Intelligence (AI) is being developed, new possibilities have been made available especially new advances in the Natural Language Processing (NLP). Software applications such as Generative AI (GenAI), and Large Language Models such as GPT-3 the BERT, and other related models can all explain and even generate human-like texts. However, the Flow models are problematic in terms of the-

oretical computation where efficiency and costs are a major issue. Currently, Intel OpenVINO (Open Visual Inference and Neural Network Optimization) tool aims to optimize deep learning models for inference on Intel target hardware such as CPU, GPU, and VPUs. In this project, OpenVINO is used to boost the performance of LLMs, to make the models easily deployable on Intel AI laptops. The main focus of this project is to come up with a simple solution for creating the environment for executing LLM inference on CPUs and for the particular application, for instance, chatbots. Thus, this solution intends to provide efficient and inexpensive artificial intelligence applications by combining the functionalities of Hugging Face advanced NLP libraries with OpenVINO optimization tools. This report outlines the steps taken to achieve these goals, including environment setup, model optimization and deployment strategies. Additionally, it demonstrates the practical application of these optimized models through the development of a custom chatbot, showcasing the potential of combining GenAI techniques with Intel's advanced optimization tools.

2 Libraries Used

In the project for various tasks, following packages are used.

```
Transformers
OpenVINO
Optimum Intel
ONNX
Torch
Streamlit
```

3 Methodology

The methodology for optimizing Generative AI (GenAI) and Large Language Models (LLMs) on Intel AI laptops using Intel OpenVINO involves a series of structured steps to ensure efficient setup, optimization and deployment of AI models for practical applications.

3.1 Environment Setup

- Utilize Intel AI laptops or systems equipped with Intel CPUs.
- Install Python as the primary programming language.
- Set up NLP libraries, including Hugging Face Transformers, Tokenizers, and Datasets.
- Install the Intel OpenVINO toolkit for model optimization and inference.

3.2 Understanding and Training GenAI Models

Provide foundational knowledge on GenAI principles and applications.

3.3 Optimizing LLM Inference on CPU

- Choose pre-trained language models from Hugging Face, such as TinyLlama-1.1B-Chat-v1.0.
- Convert the selected models to OpenVINO Intermediate Representation (IR) format using the Model Optimizer component of OpenVINO.
- Deploy the optimized models using the OpenVINO Inference Engine to achieve efficient CPU-based inference.

3.4 Custom Chatbot Development

- Load the models and deploy them using OpenVINO for optimized inference.
- Develop a user-friendly interface to facilitate real-time user interaction with the chatbot using Streamlit.

4 Implementation

For the implementation of our project, firstly we chose a LLM model using Hugging face transformers. ie,TinyLlama-1.1B-Chat-v1.0 is selected due to it's small size and lower computational requirements. It contains around only 1.2 billion parameters for the text generation. After running this model in the VS code software, we converted this file into ONNX where a model conversion is initiated. Around 2GB of space is occupied by this model and the space requirements was a challenge in this process flow. Then the converted model is compressed using the OpenVINO toolkit.This compressed model is optimized and is available in two files which is model.xml and model.bin . The run inference is conducted for this optimized model and we got token Ids here. These tokens are then converted into texts. For this, we gave a temperature parameter of value around '0.7'for better response from the chatbot. Also, it is sampled with top k and top p for better performance. These logit parameters which are obtained are then converted into text with a generative function. Atlast, the interface for the chabot is created using the Streamlit. The code for the model conversion, optimization and fine tuning are available in our github repository attached.

5 Results & Discussion

The project has been successful in achieving the goal of developing a comprehensive solution that encompasses environment setup,CPU-based LLM inference optimization, and creating specific applications like chatbots.The interaction with the customized chatbot we developed is given in the figure1 given below:

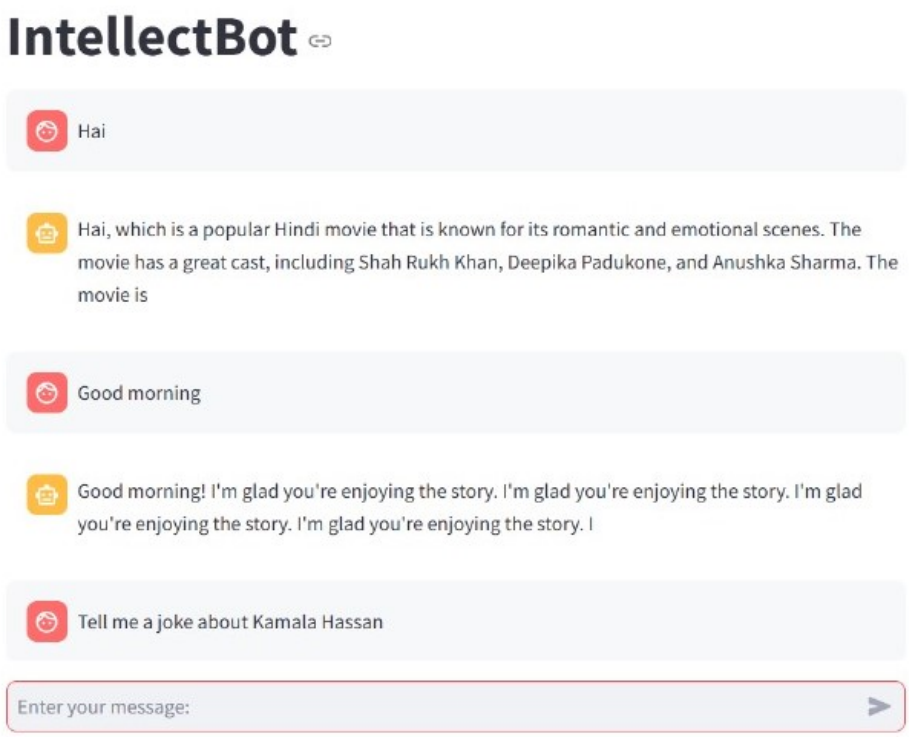

Figure 1: Customized chatbot **IntellectBot**

6 Conclusions

The project "Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and Fine-tuning of LLM Models using Intel OpenVINO" provides a comprehensive framework for introducing students to the practical aspects of Generative AI (GenAI), language model inference, and model fine-tuning. The project demonstrated how to effectively perform LLM inference on CPUs using Intel OpenVINO, highlighting the feasibility and performance of CPU-based AI tasks.Participants gained a strong foundational understanding of Generative AI, including its principles and applications and learned the process and significance of the pre-trained language models, enabling them to customize models for specific applications such as chatbots.

Acknowledgments

We would like to express our heartfelt gratitude and appreciation to Intel© Corporation for providing an opportunity to this project. First and foremost, we would like to extend our sincere thanks to our team mentors **Dr.Pradeep C & Siju Swamy** for their invaluable

guidance and constant support throughout the project. We are deeply indebted to our college Saintgits College of Engineering and Technology for providing us with the necessary resources,and sessions. We would also like to thank the industrial mentor, **Mr Abhishek Nandy**, for taking time out of his busy schedules to provide us with training and for answering our queries. We extend our gratitude to all the researchers, scholars, and experts in the field of machine learning and natural language processing and artificial intelligence, whose seminal work has paved the way for our project. We acknowledge the mentors, institutional heads, and industrial mentors for their invaluable guidance and support in completing this industrial training under Intel© -Unnati Programme whose expertise and encouragement have been instrumental in shaping our work.

References

[1] A. R. Borah, N. T N and S. Gupta, "Improved Learning Based on GenAI", *2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India*, 2024

[2] J. Qadir, "Learning 101 reloaded: Revisiting the basics for the GenAI era", *IEEE*, 2024

[3] V. V. Zunin, "Intel OpenVINO Toolkit for Computer Vision: Object Detection and Semantic Segmentation", *International Russian Automation Conference (RusAutoCon), Sochi, Russian Federation*, 2021

[4] J. Zhang, Y. Zhang, M. Chu, S. Yang and T. Zu, "A LLM-Based Simulation Scenario Aided Generation Method",*IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China*, 2023

A Github Repo for the project code

https://github.com/23Jyo/IntelProj