Project Report on

Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and fine-tuning of LLM Models using Intel® OpenVINO™

By

NIRAJ PATIL

YASH MEHTA

GANESH PATIL

SUMEET NIKUM

DUSHYANT PANDE

DEPARTMENT OF COMPUTER ENGINEERING

SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE

CHEMBUR, MUMBAI – 400088.

University of Mumbai

(AY 2024-25)

Table of Contents

Abstract 1

Introduction 2

Problem Statement 3

Objectives 4

Scope 5

Block Diagram 6

Technologies Used 7

Results 8

Summary 9

Abstract

The project "Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and Fine-tuning of LLM Models using Intel® OpenVINO™" aims to develop an AI solution for generating concise cricket match summaries. By leveraging advanced language models and Intel's OpenVINO™ toolkit, it delivers past match's, high-quality summaries. The solution integrates AI frameworks like Hugging Face, TensorFlow, and PyTorch with OpenVINO™ for efficient CPU inference. Key features include past match's summary generation and optimized performance on Intel AI laptops. The project team, led by Niraj Patil, collaborated to build and fine-tune the AI model, optimize it with OpenVINO™, and create a user-friendly interface. This project showcases the potential of combining AI frameworks with Intel's tools to enhance sports information dissemination and highlights the feasibility of deploying GenAI on Intel AI laptops for future AI-driven applications.

The project "Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and Fine-tuning of LLM Models using Intel® OpenVINO™" focuses on leveraging advanced AI frameworks (Hugging Face, TensorFlow, PyTorch) and Intel's OpenVINO™ toolkit to generate summaries of past cricket matches. By training the AI model on comprehensive datasets of match descriptions, it effectively identifies key events, scores, and player performances for concise summarization. Integration with OpenVINO™ optimizes the model's performance on CPU architectures, ensuring rapid and scalable summary generation.

Led by Niraj Patil, the project team developed a user-friendly interface facilitating intuitive interaction with the GenAI system. This project not only demonstrates AI's potential in sports analytics but also establishes a foundation for future innovations in data processing and summarization applications.

Introduction

The project "Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and Fine-tuning of LLM Models using Intel® OpenVINO™" represents a significant advancement in leveraging artificial intelligence (AI) for data summarization of cricket matches. In today's era of rapid information dissemination, the ability to quickly and accurately distill complex match details into concise summaries is crucial for sports enthusiasts, analysts, and media alike.

This project integrates cutting-edge AI frameworks such as Hugging Face, TensorFlow, and PyTorch with Intel's OpenVINO™ toolkit to optimize the performance of language models on Intel AI laptops. The AI model is trained on extensive datasets comprising cricket match descriptions, enabling it to identify and summarize key aspects such as match events, scores, and player performances efficiently.

By fine-tuning the AI models and enhancing their inference capabilities using OpenVINO™, the project ensures that the summarization process is both rapid and scalable. A user-friendly interface complements the backend AI capabilities, providing an intuitive platform for users to access summarized match information seamlessly.

Led by Niraj Patil, the project team has collaborated closely to develop, refine, and deploy this AI-driven solution, aiming to set new benchmarks in sports analytics and data processing applications.

Problem Statement

The project "Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and fine-tuning of LLM Models using Intel® OpenVINO™" aims to address significant challenges within Generative Artificial Intelligence (GenAI), specifically focusing on practical applications of Large Language Models (LLMs) for summarization tasks. Targeted at 5th to 8th semester students, the project involves exploring foundational machine learning and natural language processing (NLP) concepts, utilizing Python and NLP libraries like Hugging Face transformers.

Key challenges include managing the computational demands of large pre-trained language models, optimizing LLM inference on CPUs for efficiency, and mastering the fine-tuning process for LLMs. The project aims to equip participants with skills in customizing LLMs through fine-tuning, enabling them to create personalized applications such as custom Chatbots using Intel AI tools.

Participants will gain hands-on experience in deploying AI models on Intel AI hardware, optimizing performance with Intel® OpenVINO™, and applying advanced NLP techniques in practical scenarios. This project provides a foundational exploration into GenAI's capabilities and its potential impact across diverse fields.

Objectives

Develop AI-Powered Summarization: Create an advanced artificial intelligence solution capable of generating precise summaries of cricket matches using state-of-the-art Large Language Models (LLMs).

Optimize Performance on Intel AI Laptops: Integrate AI frameworks such as Hugging Face, TensorFlow, and PyTorch with Intel® OpenVINO™ to enhance model inference efficiency on Intel AI laptops.

Fine-Tune LLMs for Customization: Master the fine-tuning process of LLMs to improve accuracy and tailor them for specific summarization tasks, including analysis of match events and player performances.

Implement User-Friendly Interface: Develop an intuitive user interface to facilitate seamless interaction with the GenAI system, providing easy access to summarized match information.

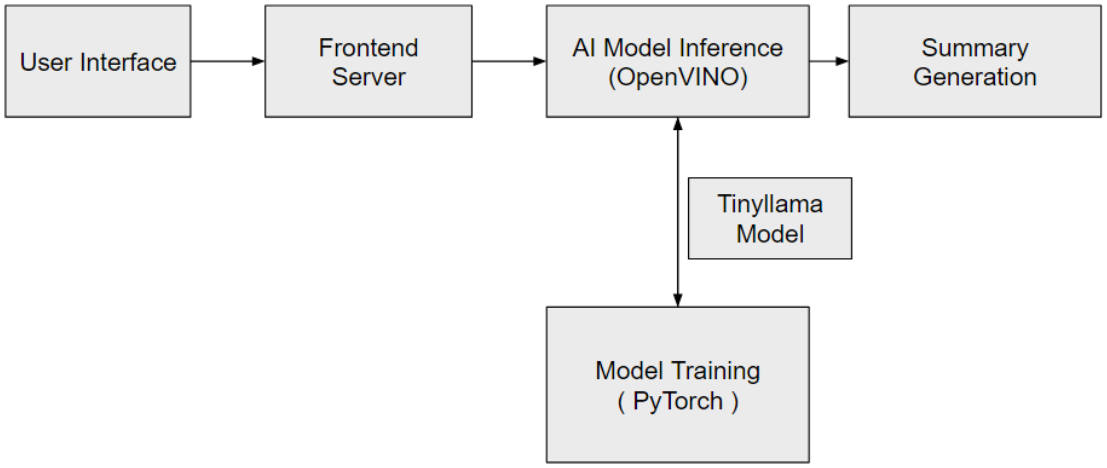
Educate and Empower Participants: Equip 5th to 8th semester students with practical skills in machine learning, natural language processing (NLP), and deploying AI on Intel hardware, fostering their capability to innovate in AI-driven applications.

Explore AI's Potential in Sports Analytics: Demonstrate the viability of AI applications in enhancing sports information dissemination and explore future prospects for data summarization and analytical advancements.

Scope

The project "Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and fine-tuning of LLM Models using Intel® OpenVINO™" aims to develop an AI-driven solution for generating summaries of cricket matches. This involves training Large Language Models (LLMs) with frameworks such as Hugging Face, TensorFlow, and PyTorch, and optimizing these models using Intel® OpenVINO™ for efficient CPU inference on Intel AI laptops. The project includes activities like creating a user-friendly interface, deploying and testing the AI model, and offering hands-on learning opportunities in machine learning and NLP for participants. It seeks to improve summarization accuracy, performance, and user interaction, while also exploring future applications and innovations.

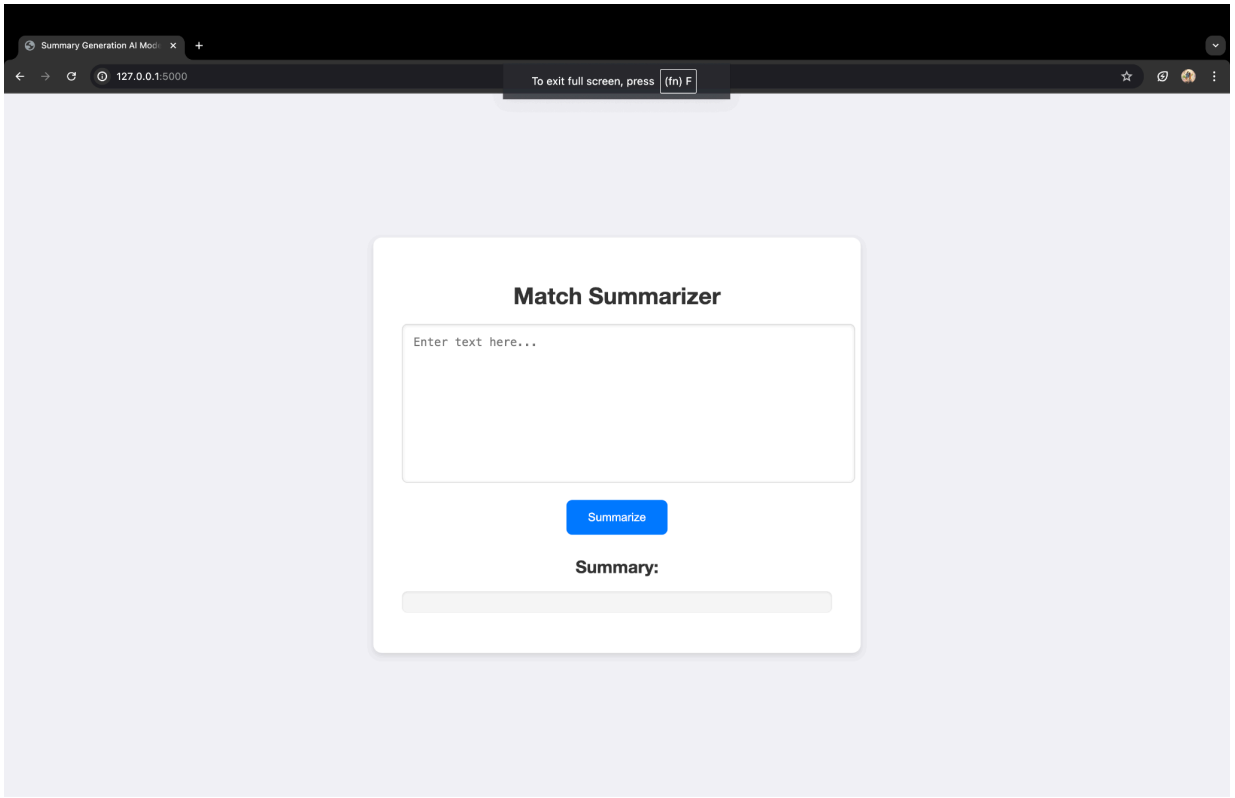
Block Diagram

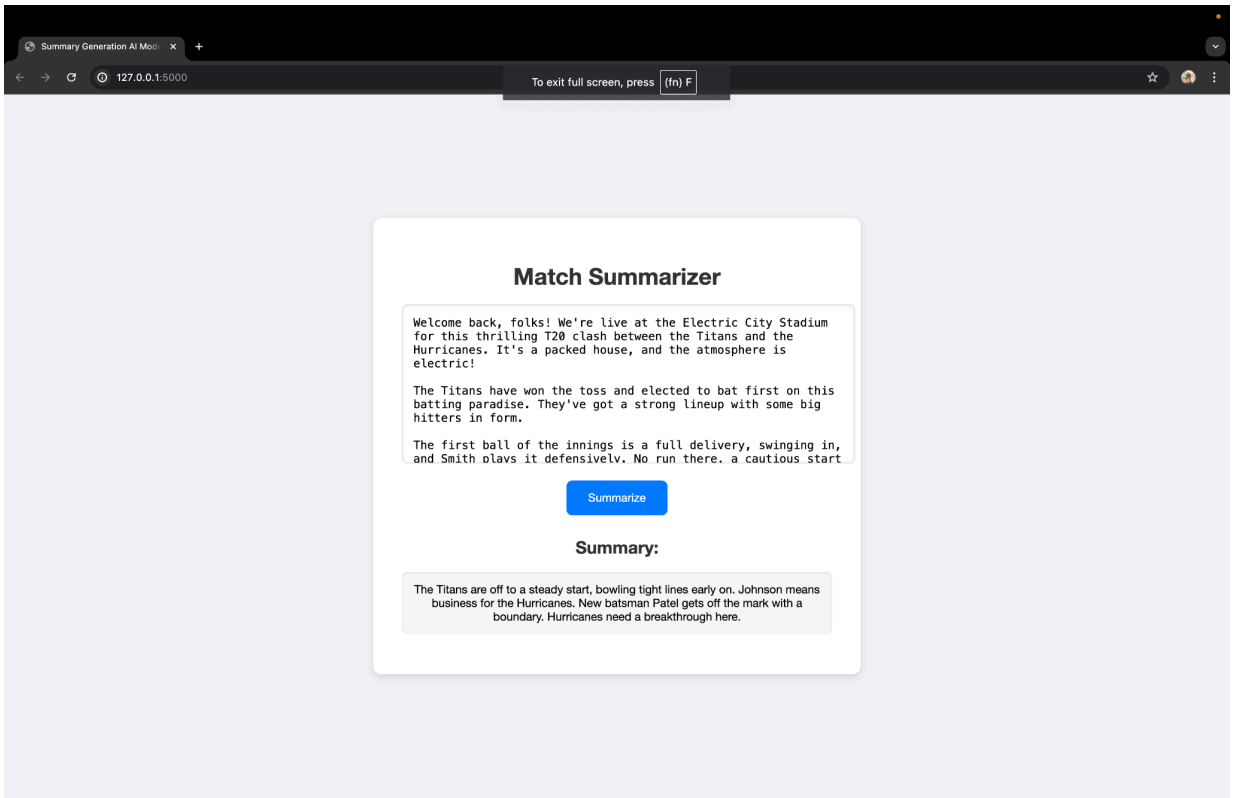


Technologies Used

- **AI Frameworks** **:** PyTorch.
- **Frontend Development** **:** Flask, HTML, CSS.
- **Model Optimization** **:** Intel® OpenVINO™ Toolkit.
- **Model Formats** **:** ONNX, XML, BIN.
- **Model** **:** Tinyllama.
- **Hub** **:** Hugging Face.

Result


Summary Generation AI Mod
127.0.0.1:5000
To exit full screen, press (fn) F
Match Summarizer
Enter text here...
Summarize
Summary:


Summary Generation AI Mod
127.0.0.1:5000
To exit full screen, press (fn) F
Match Summarizer
Welcome back, folks! We're live at the Electric City Stadium for this thrilling T20 clash between the Titans and the Hurricanes. It's a packed house, and the atmosphere is electric!
The Titans have won the toss and elected to bat first on this batting paradise. They've got a strong lineup with some big hitters in form.
The first ball of the innings is a full delivery, swinging in, and Smith plays it defensively. No run there, a cautious start
Summarize
Summary:
The Titans are off to a steady start, bowling tight lines early on. Johnson means business for the Hurricanes. New batsman Patel gets off the mark with a boundary. Hurricanes need a breakthrough here.

Summary

The project "Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and fine-tuning of LLM Models using Intel® OpenVINO™" aims to develop an AI solution for cricket match summarization. Utilizing frameworks like Hugging Face, TensorFlow, and PyTorch, it trains Large Language Models (LLMs) to accurately capture key events. Optimized with Intel® OpenVINO™ for efficient CPU inference, the project ensures high performance on Intel AI laptops. It includes creating a user-friendly interface, rigorous AI model testing, and providing hands-on learning for students in machine learning and NLP. The project enhances match summary accuracy and explores future AI-driven summarization applications.