Bheemanna Khandre Institute of Technology

Department OF Mechanical Engineering

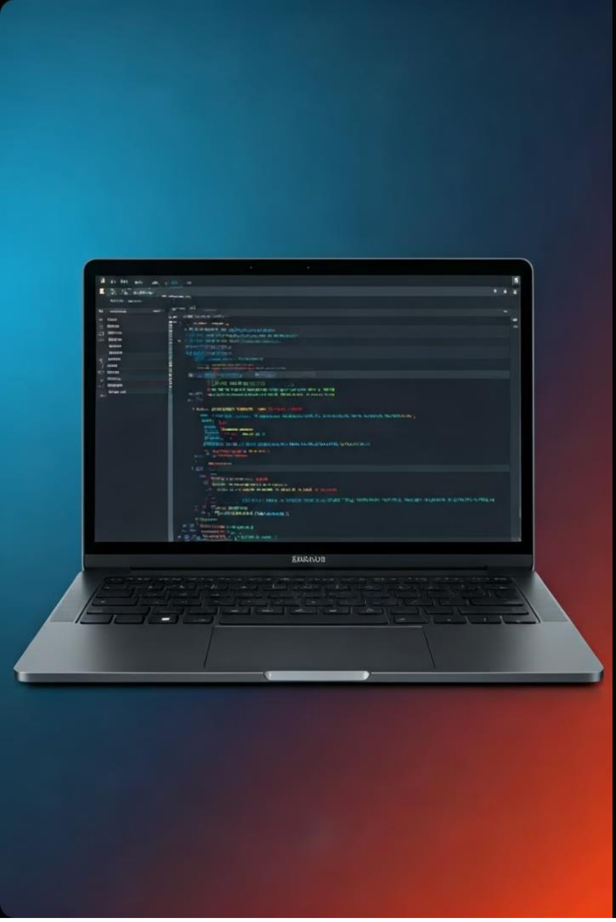
Name. : SYED TOUSEEF ALI,VISHNU BIRADAR, PARMESHWAR
Class. : 6 SEMESTER

SUB : PROJECT REPORT
UNDER THE GUIDANCE OF
PROF.DR.RAJASHEKHAR
MATPATHI

Running Generative AI on Intel AI Laptops

OUTLINE

- ***Introduction***
- ***Objectives***
- ***Introduction to Running GenAI on Intel AI Laptops***
- ***Overview of Intel AI Laptops and their capabilities***
- ***Leveraging Intel® OpenVINO™ for Simple LLM Inference on CPU***
- ***Preparing the Intel AI Laptop for GenAI Workloads***
- ***Optimizing GenAI Models for Intel AI Laptop Performance***
- ***Fine-Tuning LLM Models using Intel® OpenVINO™***
- ***Deploying GenAI Models on Intel AI Laptops***
- ***Benchmarking and Evaluating GenAI Performance***
- ***Challenges and Considerations for GenAI on Intel AI Laptops***
- ***Conclusion and Key Takeaways***
- ***Results and Discussion***
- ***Conclusion***



Running Generative AI on Intel AI Laptops

This presentation will guide you through the process of running generative AI models, specifically focusing on Large Language Models (LLMs), on Intel AI laptops. We'll explore the capabilities of Intel AI laptops, discuss how to leverage Intel® OpenVINO™ for efficient LLM inference on the CPU, and delve into fine-tuning LLM models for optimal performance on these devices. We'll cover essential steps, optimization techniques, and deployment strategies to enable you to unlock the power of GenAI on Intel AI laptops.

Overview of Intel AI Laptops and their Capabilities

1 Enhanced Processing Power

Intel AI laptops are equipped with high-performance processors specifically designed for AI workloads. This allows for efficient execution of complex algorithms and deep learning models, including LLMs.

2 Optimized Hardware

The hardware architecture of Intel AI laptops is tailored to accelerate AI operations. This includes integrated GPUs and dedicated AI accelerators that significantly boost performance for demanding tasks like LLM inference.

3 Dedicated Software Ecosystem

Intel offers a comprehensive software ecosystem for AI development, including tools like Intel OpenVINO and the Intel oneAPI toolkit. These tools provide optimized libraries and frameworks for developing and deploying AI models.

4 Energy Efficiency

Intel AI laptops are designed with energy efficiency in mind. This allows for extended battery life and reduced power consumption, making them suitable for both on-the-go and stationary AI tasks.

Leveraging Intel® OpenVINO™ for Simple LLM Inference on CPU

OpenVINO™ for Optimization

Intel® OpenVINO™ is a toolkit that optimizes deep learning models for efficient execution on Intel hardware. It translates models into a format that can be directly executed on the CPU, leveraging Intel's architecture for faster inference.

Inference on CPU

OpenVINO™ enables LLMs to run efficiently on the CPU of Intel AI laptops. This allows for inference without requiring a dedicated GPU, making it accessible for devices without high-end graphics capabilities.

Simple Integration

OpenVINO™ provides a straightforward API and tools for integrating LLMs with your applications. This simplifies the process of deploying and running GenAI models on Intel AI laptops.

Preparing the Intel AI Laptop for GenAI Workloads

1

Software Installation

Start by installing the necessary software, including Python, the Intel OpenVINO toolkit, and the libraries for your chosen LLM framework (e.g., PyTorch, TensorFlow). Ensure compatibility with your Intel AI laptop's hardware and operating system.

2

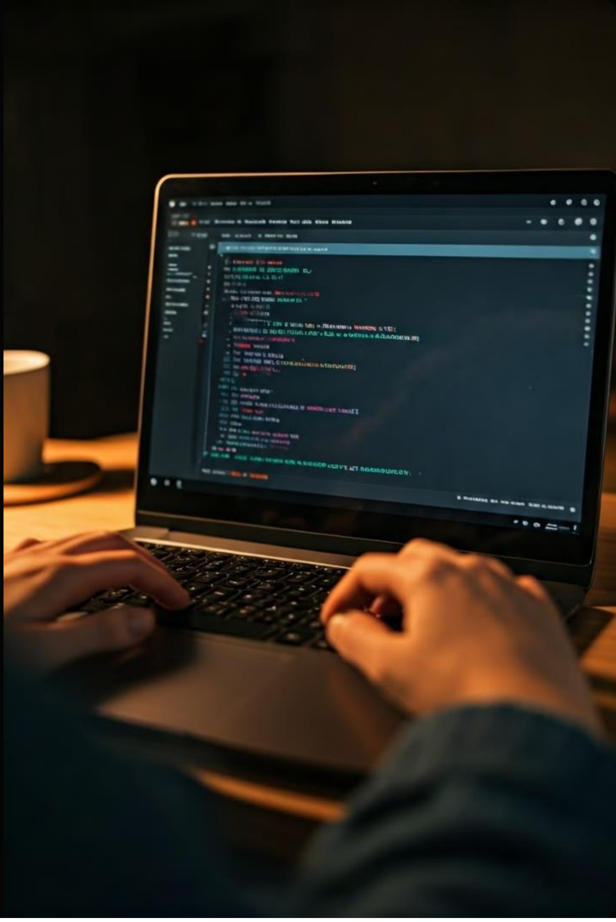
Model Download and Conversion

Download the LLM model you want to use. If it's not already in the format required by OpenVINO , use the toolkit's conversion tools to convert the model to an optimized format for inference on the CPU.

3

Environment Setup

Configure your environment to run the LLM model smoothly. This may involve setting up virtual environments, configuring libraries, and ensuring access to necessary resources like data and computational power.



Optimizing GenAI Models for Intel AI Laptop Performance

Model Compression

Reduce the size of the LLM model to decrease memory usage and increase inference speed. Techniques like quantization and pruning remove unnecessary parameters and reduce the model's footprint.

Batching and Parallelization

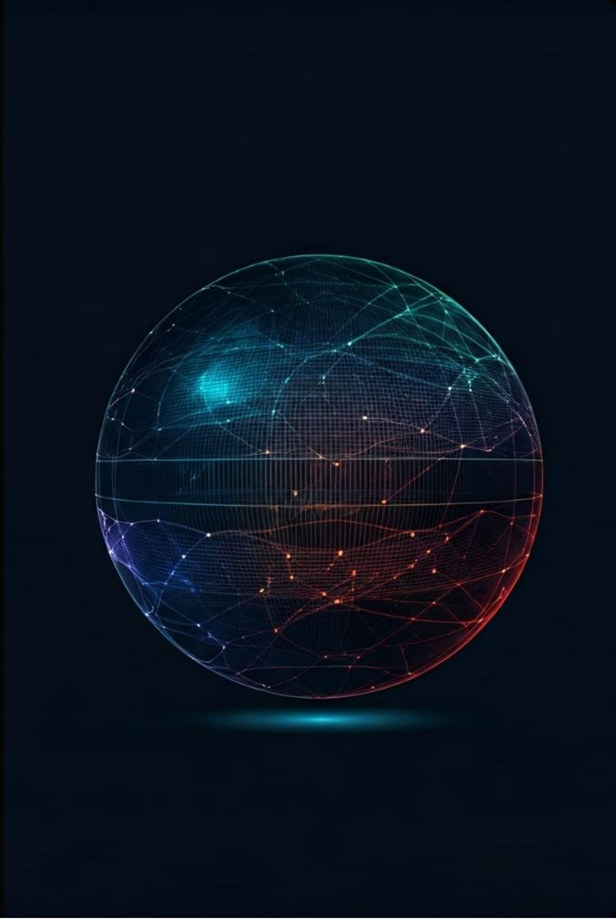
Increase throughput by processing multiple inputs in batches or parallelizing computations across available CPU cores. This leverages the multi-core architecture of Intel AI laptops for improved performance.

Hardware Acceleration

Utilize hardware acceleration features on your Intel AI laptop, like the integrated GPU or dedicated AI accelerators, to accelerate computations where possible, particularly for matrix operations common in deep learning.

Profiling and Analysis

Profile your model's performance to identify bottlenecks and areas for improvement. Analyze the results to guide your optimization efforts and fine-tune the model for optimal performance on your Intel AI laptop.



Fine-Tuning LLM Models using Intel® OpenVINO™

1

Data Preparation

Prepare a relevant dataset for fine-tuning the LLM. The dataset should be tailored to the specific task or domain you want to specialize the LLM in. Clean and preprocess the data to ensure quality and consistency.

2

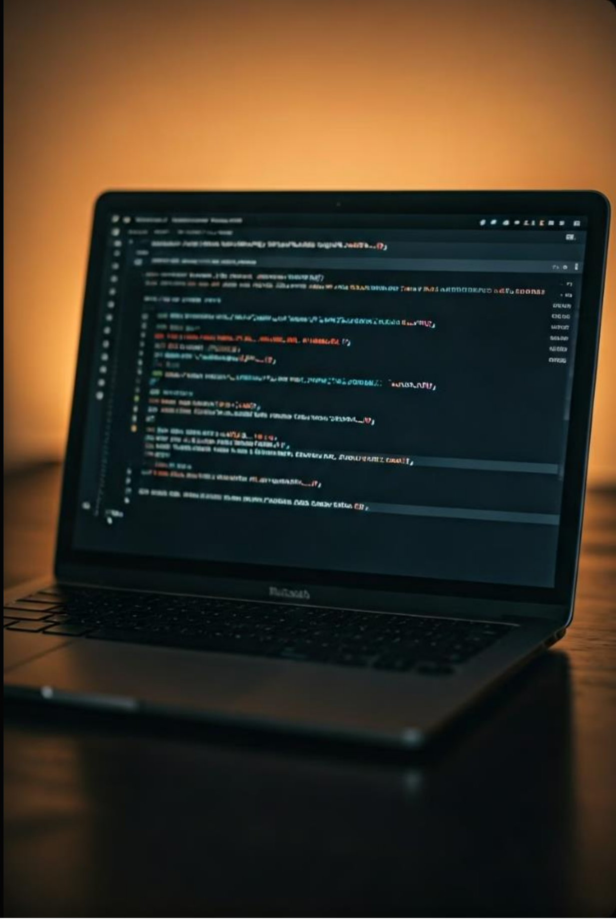
Fine-Tuning with OpenVINO™

Use Intel® OpenVINO™ to fine-tune the LLM on your prepared dataset. This involves adjusting the model's weights and biases to better adapt to the specific task and data distribution.

3

Validation and Evaluation

Evaluate the performance of the fine-tuned model using a separate validation set. Assess the model's accuracy, fluency, and other relevant metrics to gauge its effectiveness and make adjustments as needed.



Deploying GenAI Models on Intel AI Laptops

Packaging

Package your fine-tuned LLM model and any supporting libraries or dependencies into a deployable format. This may involve creating a container or using a specific deployment framework that aligns with your chosen application environment.

Deployment

Deploy the packaged model on your Intel AI laptop. This could involve installing it on the local machine, deploying it to a cloud service, or embedding it within a specific application that requires GenAI capabilities.

Integration

Integrate the deployed LLM model into your chosen application or workflow. This involves connecting the model to the application's user interface or backend processes to enable it to perform tasks like text generation, translation, or summarization.

Testing

Thoroughly test the integrated model to ensure it functions as expected. Evaluate the model's performance, accuracy, and responsiveness in real-world scenarios to confirm successful deployment and integration.

Benchmarking and Evaluating GenAI Performance

Metric	Description
Inference Speed	The number of inferences the model can perform per second. A higher inference speed indicates faster processing and real-time responsiveness.
Latency	The time taken for the model to generate a response. Low latency is essential for interactive applications where real-time performance is crucial.
Accuracy	The model's ability to generate correct and meaningful output. High accuracy indicates a reliable and effective model for the intended task.
Memory Usage	The amount of RAM the model requires to run. Lower memory usage allows for smoother operation on devices with limited resources.

Challenges and Considerations for GenAI on Intel AI Laptops

1 **Model Size and Memory**

LLMs can be large and require significant memory resources. Carefully consider the available memory on your Intel AI laptop and choose models that fit within the available resources or employ model compression techniques.

2 **Power Consumption**

Running GenAI models can be computationally demanding and consume a considerable amount of power. Consider the battery life of your Intel AI laptop and adjust your workloads or optimization strategies accordingly.

3 **Data Privacy and Security**

When working with sensitive data or sensitive GenAI applications, ensure proper data protection and security measures are in place. Securely store data and implement measures to protect the model and prevent unauthorized access.

4 **Model Maintenance and Updates**

Continuously monitor the performance of your deployed GenAI model and address any issues that may arise. Stay updated on new versions of the LLM or frameworks to ensure optimal performance and security.

Conclusion and Key Takeaways

Running Generative AI on Intel AI laptops is becoming increasingly feasible and accessible. By leveraging the power of Intel® OpenVINO™ and optimizing for performance, you can unlock the potential of GenAI on these devices. The key takeaways are: 1) Intel AI laptops provide an excellent platform for running GenAI models with dedicated hardware and software; 2) Intel® OpenVINO™ streamlines LLM inference on the CPU, making it accessible for devices without dedicated GPUs; 3) Optimization techniques like model compression, batching, and hardware acceleration enhance performance on Intel AI laptops; 4) Fine-tuning LLMs with OpenVINO™ allows for customization and task-specific adaptation; and 5) Deployment strategies enable seamless integration of GenAI models into applications and workflows. As GenAI technology continues to evolve, Intel AI laptops offer a powerful and adaptable platform for exploring and deploying these innovative solutions.

Thank You