Intel Unnati Summer Program 2024

-Aadithya Ramesh

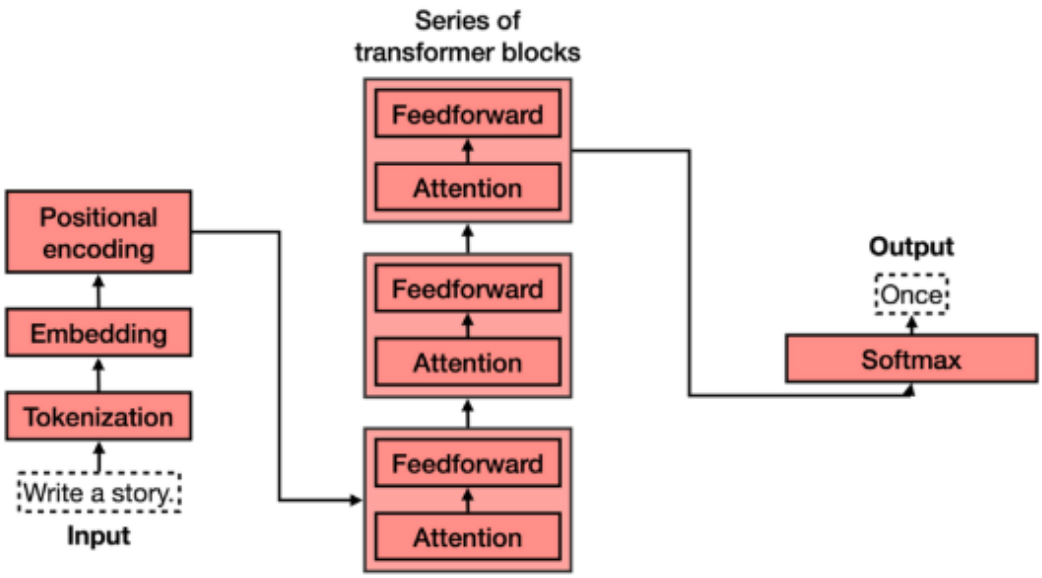
Problem Statement:

Introduction to GenAI and Simple LLM inference on CPU and fine-tuning of LLM Model to create a Custom Chatbot.

Technical Approach:

Transformers:

A Transformer model consists of an encoder and a decoder. The encoder encodes the input sequence and passes it to the decoder which learns how to decode the representation for a relevant task.



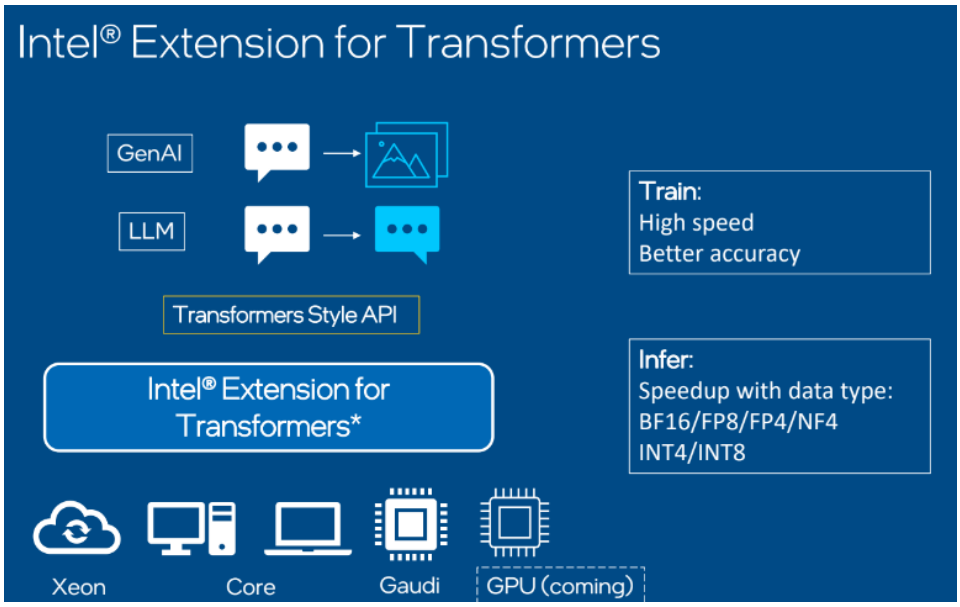
Hugging Face:

Important for its **transformers** built for Natural Language Processing applications and its platform that allows users to share machine learning models and **datasets** and showcase their work.

Intel Tools Used:

Intel Extension for Transformers:

Intel Extension for Transformers provides an efficient inference runtime of large language models (LLMs) on Intel platforms through the state-of-the-art model compression techniques.



Intel Developer Cloud:

Enables AI Development using Intel-optimized software on the latest Intel Xeon processors, Intel Data Center GPUs, and Intel Gaudi 2 accelerators.

Allows us to get hands on experience with building AI.

Speeds up AI deployment with latest tools.

Jupyter Notebook 1:

build_chatbot_on spr.ipynb:

While Kernel was working:

Prompts:

1. ### Tell me about Intel Xeon Scalable Processors?

   The Intel Xeon Scalable Processors represent a family of high-performance central processing units (CPUs) designed for data centers, cloud computing, and other demanding workloads. These processors offer significant improvements in performance, efficiency and scalability compared to their predecessors. They feature advanced technologies such as Intel advanced vector Extensions 512(AVX-512), Intel Turbo Boost Technology and Intel Hyper-Threading Technology which contribute to increased throughput and reduced latency

2. ### What are the different types of Intel Xeon Scalable Processors?

   The Different types of Intel Scalable Processors Include:

   1.Platinum Series
   2. Gold Series
   3. Silver Series
   4.Bronze Series
   5.Cooper-Lake Based Series
   6.Ice-Lake Based Series

3. ### What is the Ice Lake Series of Intel Xeon Scalable Processors?

   The Ice Lake series of Intel Xeon processors, launched in 2021, are built on a 10nm process and Sunny Cove architecture. They offer up to 40 cores, support DDR4-3200 memory, and PCIe 4.0, enhancing performance, power efficiency, and security. These processors are ideal for cloud computing, AI, and data analytics.

Problems Faced:

Kernel Issue:

The kernel I had made using the following steps as stated by Vasudha Ma'am had suddenly stopped working a couple of days before report creation.

```
conda create -n itrex python=3.10 -y
conda activate itrex

pip install intel-extension-for-transformers

git clone https://github.com/intel/intel-extension-for-transformers.git

cd ./intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/

pip install -r requirements_cpu.txt

pip install -r requirements.txt

huggingface-cli login

##install jupyter and ipykernel
python3 -m pip install jupyter ipykernel

##Add kernel for its environment
python3 -m ipykernel install --nam
```

Error Starting Kernel

```
Unhandled error

▼ Details

Traceback (most recent call last):
  File "/srv/jupyter/python-venv/lib/python3.11/site-packages/tornado/web.py", line 1786, in _execute
    result = await result
             ^^^^^^^^^^^^
  File "/srv/jupyter/python-venv/lib/python3.11/site-packages/jupyter_server/services/sessions/handlers.py", line 163, in patch
    kernel_id = await sm.start_kernel_for_session(
                ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/srv/jupyter/python-venv/lib/python3.11/site-packages/jupyter_server/services/sessions/sessionmanager.py", line 345, in sta
    kernel_id = await self.kernel_manager.start_kernel(
                ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/srv/jupyter/python-venv/lib/python3.11/site-packages/jupyter_server/services/kernels/kernelmanager.py", line 233, in _asyn
    kernel_id = await self.pinned_superclass._async_start_kernel(self, **kwargs)
                ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/srv/jupyter/python-venv/lib/python3.11/site-packages/jupyter_client/multikernelmanager.py", line 272, in _async_start_kern
    raise km.ready.exception()  # type: ignore
    ^^^^^^^^^^^^^^^^^^^^^^^^^^
```

Ok

Error Starting Kernel

```
[Errno 2] No such file or directory: '/home/u2d71f83799810a89c715926eecd1893/.conda/envs/itrex-1/bin/python3'
```

Ok

When I tried to make a new kernel, I am faced with the following issue, wherein it states that I am in violation of ToS of Anaconda.

build_chatbot_on_spr.ipynb | simple_llm_inference.ipynb | Terminal 4

```
u2d71f83799810a89c715926eecd1893@idc-training-gpu-compute-21:~/Training/AI/GenAI$ conda create -n itrex python=3.10 -y
Channels:
 - defaults
Platform: linux-64
Collecting package metadata (repodata.json): failed

CondaHTTPError: HTTP 429 TERMS OF SERVICE RATE LIMIT EXCEEDED for url <https://repo.anaconda.com/pkgs/main/linux-64/repodata.json>
Elapsed: 00:06.222323
CF-RAY: 8a3185700bd1b997-SEA

It has been determined that you are in violation of the Anaconda terms of service, and may not be licensed for usage.

u2d71f83799810a89c715926eecd1893@idc-training-gpu-compute-21:~/Training/AI/GenAI$
```

Jupyter Notebook 2:

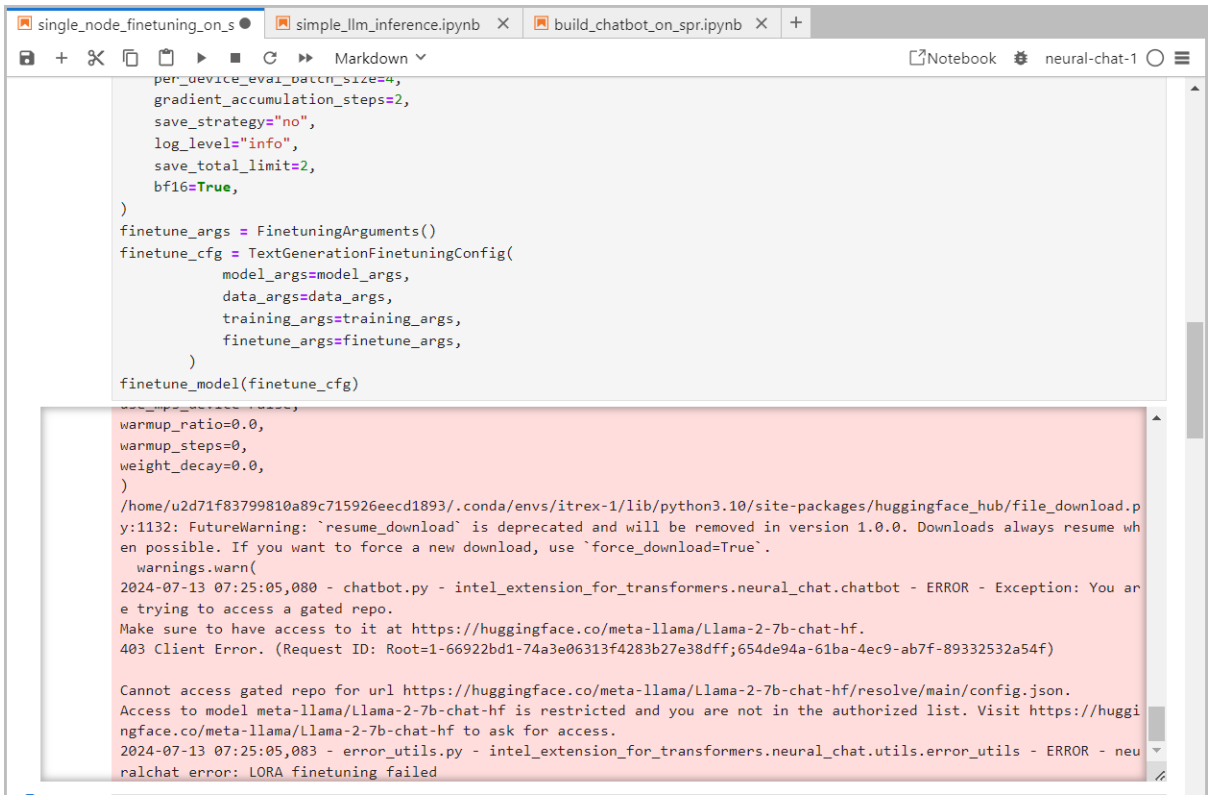
single_node_finetuning_on_spr.ipynb:

Problems Faced [When 'neural-chat-1' kernel was working]:

Gated Repo:

I have downloaded the alpaca.json file and edited the code to point to its directory.

If I run the code, I get the error that llama2 is a gated repo and cannot be accessed.



Multiple discussions have been held with Industry Mentor, Abhishek Nandy Sir regarding both these problems.

I have implemented everything he had suggested including Hugging face interface and problem has not been resolved.