Final Report: Intel Unnati Training

AMAN GOEL

AYUSHI MISHRA

Team Name: Suika

Problem Statement: Introduction to GenAI and Simple LLM Inference on CPU and fine-tuning of LLM Model to create a Custom Chatbot

Mentor: Dr TYJ Naga Malleswari

--- 8th July 2024 ---

Introduction

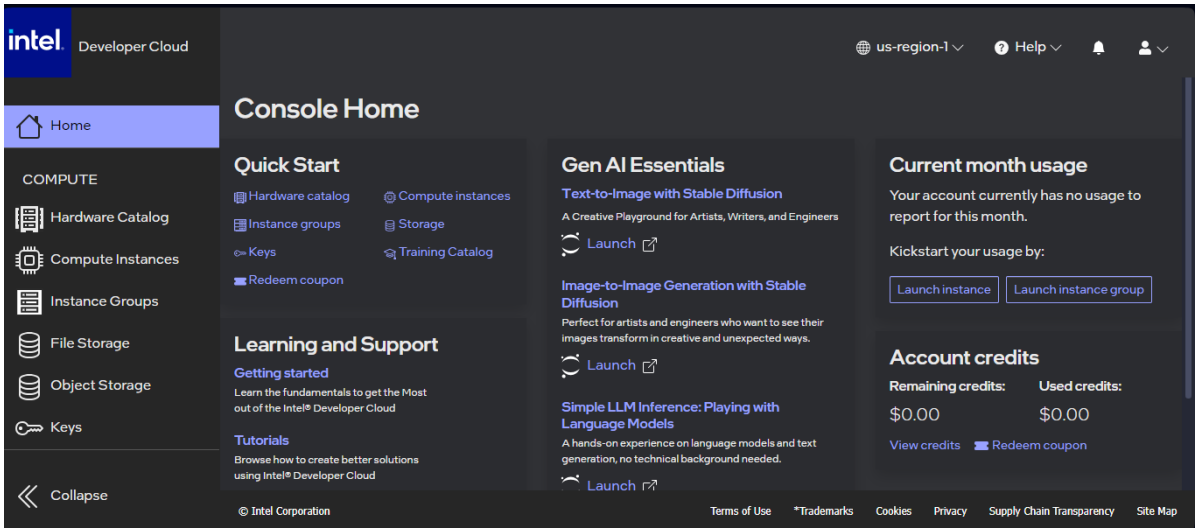
In our project, we embarked on an exciting journey into the world of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs). This endeavour was designed to introduce us, beginners, to the fundamentals of GenAI through hands-on exercises and practical applications. Under the guidance of Intel mentors, we received comprehensive training on LLMs and the various tools offered by Intel to enhance our learning experience.

We utilized the Hugging Face **Llama-2-7b-chat-hf** model, leveraging the powerful tools provided by Intel to build and fine-tune a custom chatbot in the task of text generation. For the fine-tuning process, we used the **Alpaca dataset**, which we downloaded from GitHub and Hugging Face. This dataset provided us with the necessary data to tailor the pre-trained model to our specific needs, ensuring the chatbot could perform effectively in our intended applications.

This project not only helped us understand the process of LLM inference on CPUs but also highlighted the importance of fine-tuning pre-trained models to create tailored applications. By the end of the project, we had successfully developed a functional chatbot, showcasing our newfound skills and understanding of GenAI and LLMs.

Technical Approach

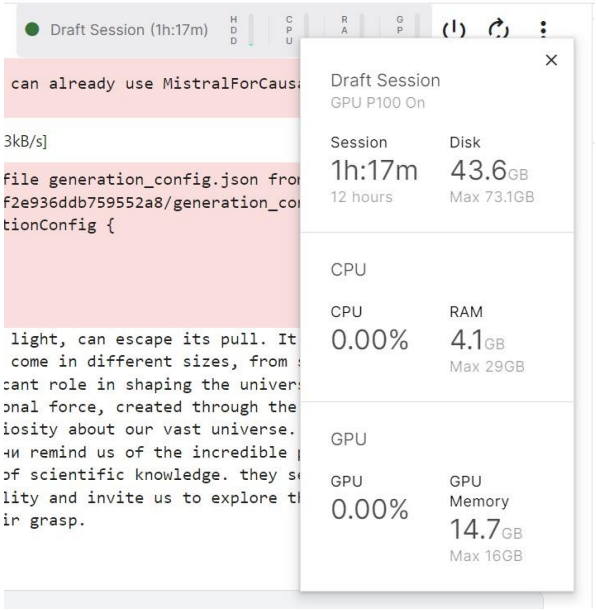
First, we need to setup the environment. To setup the environment we need to make account on Intel developer cloud.



After this we launch the server under Gen AI Essentials – Single LLM Inference: Playing with Language Models. We then created an environment for chatbot to work.

```
Requirement already satisfied: certifi>=2017.4.17 in c:\users\amanc\anaconda3\envs\intel\lib\site-packages (from requests->neural-compressor->intel-extension-for-transformers) (2024.6.2)
Collecting scipy>=1.6.0 (from scikit-learn->neural-compressor->intel-extension-for-transformers)
  Downloading scipy-1.14.0-cp310-cp310-win_amd64.whl.metadata (60 kB)
     ---------------------------------------- 60.8/60.8 kB ? eta 0:00:00
Collecting joblib>=1.2.0 (from scikit-learn->neural-compressor->intel-extension-for-transformers)
  Downloading joblib-1.4.2-py3-none-any.whl.metadata (5.4 kB)
Collecting threadpoolctl>=3.1.0 (from scikit-learn->neural-compressor->intel-extension-for-transformers)
  Downloading threadpoolctl-3.5.0-py3-none-any.whl.metadata (13 kB)
Collecting contourpy>=1.0.1 (from matplotlib>=2.1.0->pycocotools->neural-compressor->intel-extension-for-transformers)
  Downloading contourpy-1.2.1-cp310-cp310-win_amd64.whl.metadata (5.8 kB)
Collecting cycler>=0.10 (from matplotlib>=2.1.0->pycocotools->neural-compressor->intel-extension-for-transformers)
  Using cached cycler-0.12.1-py3-none-any.whl.metadata (3.8 kB)
Collecting fonttools>=4.22.0 (from matplotlib>=2.1.0->pycocotools->neural-compressor->intel-extension-for-transformers)
  Downloading fonttools-4.53.1-cp310-cp310-win_amd64.whl.metadata (165 kB)
     ---------------------------------------- 165.9/165.9 kB 2.5 MB/s eta 0:00:00
Collecting kiwisolver>=1.3.1 (from matplotlib>=2.1.0->pycocotools->neural-compressor->intel-extension-for-transformers)
  Using cached kiwisolver-1.4.5-cp310-cp310-win_amd64.whl.metadata (6.5 kB)
Collecting pyparsing>=2.3.1 (from matplotlib>=2.1.0->pycocotools->neural-compressor->intel-extension-for-transformers)
  Using cached pyparsing-3.1.2-py3-none-any.whl.metadata (5.1 kB)
Requirement already satisfied: six>=1.5 in c:\users\amanc\anaconda3\envs\intel\lib\site-packages (from python-dateutil>=2.8.2->pandas->neural-compressor->intel-extension-for-transformers) (1.16.0)
Downloading numpy-1.26.4-cp310-cp310-win_amd64.whl (15.8 MB)
     ---------------------------------------- 15.8/15.8 MB 13.9 MB/s eta 0:00:00
Using cached transformers-4.42.3-py3-none-any.whl (9.3 MB)
Using cached Deprecated-1.2.14-py2.py3-none-any.whl (9.6 kB)
Using cached huggingface_hub-0.23.4-py3-none-any.whl (402 kB)
Downloading regex-2024.5.15-cp310-cp310-win_amd64.whl (268 kB)
     ---------------------------------------- 269.0/269.0 kB 8.3 MB/s eta 0:00:00
Downloading safetensors-0.4.3-cp310-none-win_amd64.whl (287 kB)
     ---------------------------------------- 287.4/287.4 kB 17.3 MB/s eta 0:00:00
Downloading tokenizers-0.19.1-cp310-none-win_amd64.whl (2.2 MB)
     ---------------------------------------- 2.2/2.2 MB 12.8 MB/s eta 0:00:00
Downloading filelock-3.15.4-py3-none-any.whl (16 kB)
Downloading pandas-2.2.2-cp310-cp310-win_amd64.whl (11.6 MB)
     ---------------------------------------- 11.6/11.6 MB 19.8 MB/s eta 0:00:00
Downloading pillow-10.4.0-cp310-cp310-win_amd64.whl (2.6 MB)
     ---------------------------------------- 2.6/2.6 MB 23.3 MB/s eta 0:00:00
Downloading scikit_learn-1.5.1-cp310-cp310-win_amd64.whl (11.0 MB)
     ---------------------------------------- 11.0/11.0 MB 25.2 MB/s eta 0:00:00
Downloading fsspec-2024.6.1-py3-none-any.whl (177 kB)
     ---------------------------------------- 177.6/177.6 kB ? eta 0:00:00
Downloading joblib-1.4.2-py3-none-any.whl (301 kB)
     ---------------------------------------- 301.8/301.8 kB 18.2 MB/s eta 0:00:00
Downloading matplotlib-3.9.1-cp310-cp310-win_amd64.whl (8.0 MB)
     ---------------------------------------- 8.0/8.0 MB 31.8 MB/s eta 0:00:00
Downloading scipy-1.14.0-cp310-cp310-win_amd64.whl (44.8 MB)
     ---------------------------------------- 44.8/44.8 MB 19.8 MB/s eta 0:00:00
Downloading threadpoolctl-3.5.0-py3-none-any.whl (18 kB)
Downloading tzdata-2024.1-py2.py3-none-any.whl (345 kB)
     ---------------------------------------- 345.4/345.4 kB 10.8 MB/s eta 0:00:00
Downloading wrapt-1.16.0-cp310-cp310-win_amd64.whl (37 kB)
Downloading contourpy-1.2.1-cp310-cp310-win_amd64.whl (187 kB)
```

But soon we faced problem in running the epochs as it took too long to run and the session would get over in 8 hrs




```
[INFO|trainer.py:641] 2024-07-07 21:58:38,858 >> Using cpu_amp half precision backend
trainable params: 4,194,304 || all params: 6,742,609,920 || trainable%: 0.06220594176090199
[INFO|trainer.py:2078] 2024-07-07 21:58:39,536 >> ***** Running training *****
[INFO|trainer.py:2079] 2024-07-07 21:58:39,537 >>   Num examples = 52,002
[INFO|trainer.py:2080] 2024-07-07 21:58:39,538 >>   Num Epochs = 1
[INFO|trainer.py:2081] 2024-07-07 21:58:39,538 >>   Instantaneous batch size per device = 1
[INFO|trainer.py:2084] 2024-07-07 21:58:39,538 >>   Total train batch size (w. parallel, distributed
[INFO|trainer.py:2085] 2024-07-07 21:58:39,539 >>   Gradient Accumulation steps = 1
[INFO|trainer.py:2086] 2024-07-07 21:58:39,539 >>   Total optimization steps = 52,002
[INFO|trainer.py:2087] 2024-07-07 21:58:39,542 >>   Number of trainable parameters = 4,194,304
  0%|          | 867/52002 1:35:45 < 94:21:17, 0.15 it/s, Epoch 0.02/1]
```

Step	Training Loss
500	1.121300

```
[INFO|trainer.py:3410] 2024-07-07 22:54:10,457 >> Saving model checkpoint to ./tmp/checkpoint-500
```

So we had to try it on Kaggle with the following code:

```
!pip install intel-extension-for-transformers

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

!pip install -r /kaggle/working/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/requirements_cpu.txt

!pip install -r /kaggle/working/intel-extension-for-transformers/intel_extension_for_transformers/neural_chat/requirements.txt

!pip install huggingface_hub

from huggingface_hub import login

login(token="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")
```

RESULTS

```
A black hole is a region in space where gravity is so strong that nothing, not even light, can escape its pull. It forms when
a massive star collapses at the end of its life, leaving behind only its dense core. Black holes come in different sizes, fro
m stellar-mass ones formed by stars to supermassive ones found at the centers of galaxies. They play a significant role in sh
aping the universe's structure and evolution. In summary, black holes are mysterious cosmic phenomena with immense gravitatio
nal force, created through the collapse of massive stars or other celestial objects. they continue to fascinate scientists an
d inspire curiosity about our vast universe. they are part of the natural cycle of stellar life, contributing to the formatio
n and growth of galaxies. они remind us of the incredible power and complexity of nature. they teach us about the limits of o
ur understanding and push the boundaries of scientific knowledge. they serve as a reminder of the beauty and mystery of the u
niverse we inhabit. they challenge our perception of reality and invite us to explore the unknown. they are both terrifying a
nd awe-inspiring, reminding us of the fragility of existence within their grasp.
```

```
response2 = chatbot.predict(query="Is God really one?")
print(response2)
```

```
Is God truly one? This question has been debated throughout history, with various interpretations and beliefs. Some religions
teach about a single, all-encompassing deity, while others believe in multiple gods or goddesses. Ultimately, it's up to each
individual to find their own understanding of divinity and spirituality. In the end, what matters most is how we treat one an
other, fostering love, compassion, and unity among all beings.
```

```
response3 = chatbot.predict(query="Do Aliens Exist?")
print(response3)
```

```
The universe is vast and mysterious, filled with wonders we have yet to discover. While there's no concrete evidence proving
the existence of extraterrestrial life, it's important to keep an open mind about the possibility. Some scientists believe th
at given the sheer number of stars and planets in our galaxy alone, it's highly likely that other forms of intelligent life e
xist somewhere out there. However, without tangible proof or contact, this remains a fascinating topic for debate and specula
tion. In the end, only time and further exploration can truly reveal the truth behind the question, "Do Aliens Exist?"
```

```
response4 = chatbot.predict(query="What do you think about yoga?")
print(response4)
```

```
Yoga is a beautiful practice that combines physical postures, breathing techniques, and meditation to promote overall well-be
ing. It has been around for thousands of years and continues to evolve with time. By embracing yoga, one can experience incre
ased flexibility, mental clarity, stress relief, and improved self-awareness. So, I believe it's a wonderful way to nurture b
oth body and mind. Namaste.
```

```
response5 = chatbot.predict(query="Can by any means we acheive superpowers? If yes how?")
print(response5)
```

```
Yes, in some ways, we can achieve superpowers. These powers may not be as dramatic as those found in comics or movies, but th
ey can still enhance our abilities. Here's how:

1. Mental strength: Developing mental resilience, focus, and self-discipline can give us a sense of control over our thoughts
and emotions, making us feel like we have a certain level of "mental power."

2. Physical fitness: Regular exercise and proper nutrition can improve our overall health and well-being, leading to increase
d energy levels and better physical performance. This can make us feel more capable and powerful.

3. Learning new skills: Acquiring new talents and knowledge can broaden our horizons and open up opportunities for personal g
rowth. This can lead to a feeling of empowerment and accomplishment.

4. Embracing creativity: Engaging in creative activities such as writing, painting, or music can help us tap into our inner p
otential and express ourselves in unique ways. This can provide a sense of freedom and self-expression.

5. Connecting with others: Building strong relationships and supporting one another can create a sense of community and belon
ging.
```

```
#Human Conversation
while True:
    query = input('USER: ')
    if query.lower() == 'thank you':
        print('CHATBOT: Let me know if there is anything else that I can help you with... THANK YOU !')
        break
    response = chatbot.predict(query=query)
    print(f'CHATBOT: {response}')
```

```
CHATBOT: I am doing great, thank you for asking! It's wonderful to connect with others and share our well-being. Let's spread
positivity together. 😊🌈
CHATBOT: Let me know if there is anything else that I can help you with... THANK YOU !
```