

CLOUD COMPUTING LAB

Introduction:

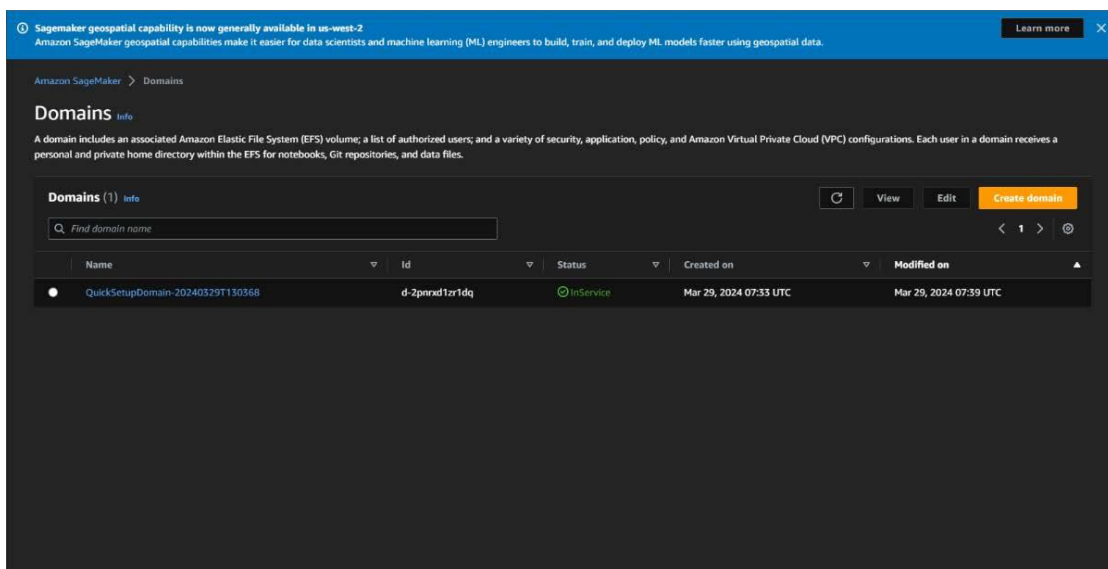
In the current data-driven landscape, Natural Language Processing (NLP) plays a pivotal role in enabling machines to understand and generate human language effectively. Within NLP, Language Models (LLMs) are essential for various applications, including text generation, sentiment analysis, and translation.

The Llama-7b model, developed by Meta (formerly Facebook), represents a significant leap forward in language understanding. Its capabilities include processing and generating text with impressive fluency and coherence. However, deploying such an advanced model requires careful consideration of infrastructure, scalability, and performance optimization.

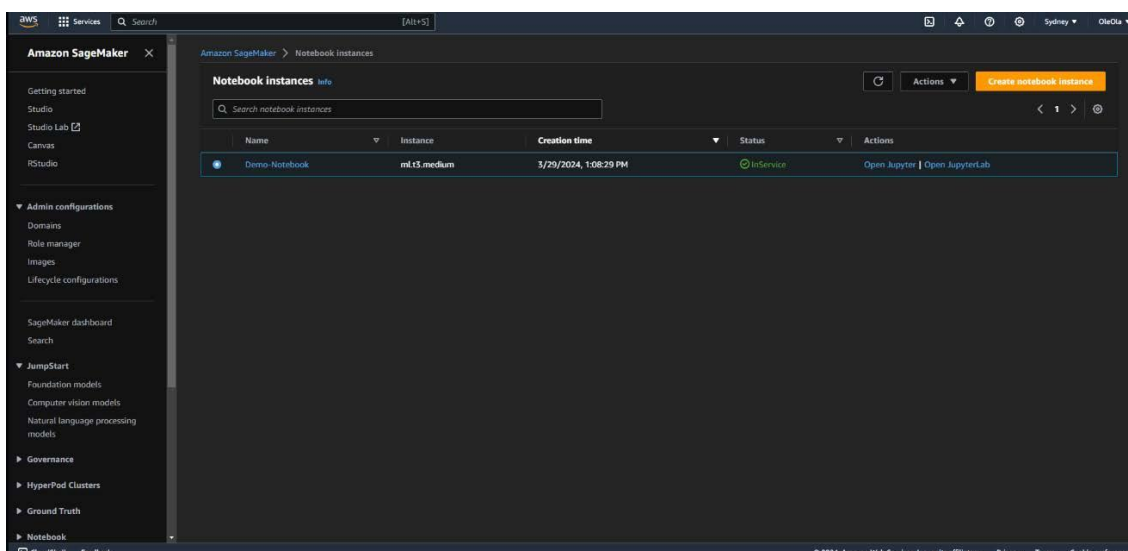
Amazon SageMaker provides a comprehensive machine learning platform that seamlessly facilitates the deployment of LLMs like Llama-7b

Working:

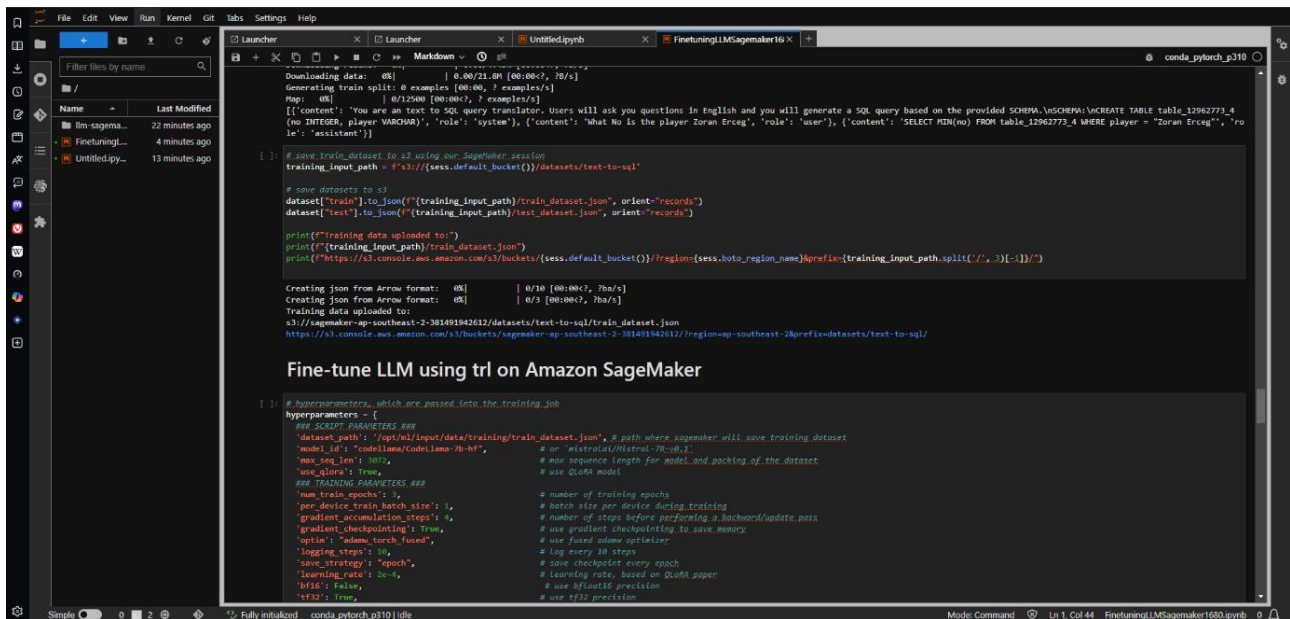
1. Creating SageMaker Domain:



2. Creating Notebook Instance:



3. Open Jupyter Lab and run the notebook to deploy Model:



```
Download data: OK | 0.00/21.0M [00:00<, 70/s]
Generating train split: 0 examples [00:00, ? examples/s]
Map: OK | 0/12500 [00:00<, ? examples/s]
[{'content': 'You are an text to SQL query translator. Users will ask you questions in English and you will generate a SQL query based on the provided SCHEMA.\nSCHEMA:\nCREATE TABLE table_12962773_4\n(\n  ID INTEGER,\n  player VARCHAR,\n  role: 'system'), {\n  content: 'What No is the player Zoran Erceg',\n  role: 'user'}, {\n  content: 'SELECT HDN(no) FROM table_12962773_4 WHERE player = "Zoran Erceg"',\n  role: 'assistant'}]]

[ ]: # save train dataset to s3 using our SageMaker session
training_input_path = f"s3://{(sess.default_bucket())}/datasets/text-to-sql/"

# save datasets to s3
dataset["train"].to_json(f"{(training_input_path)}/train_dataset.json", orient="records")
dataset["test"].to_json(f"{(training_input_path)}/test_dataset.json", orient="records")

print(f"Training data uploaded to:")
print(f"{(training_input_path)}/train_dataset.json")
print(f"https://s3.console.aws.amazon.com/s3/buckets/{(sess.default_bucket())}/region:{(sess.boto_region_name)}prefix:{(training_input_path.split('/', 2)[-1])}/")

Creating json from Arrow format: OK | 0/10 [00:00<, 70/s]
Creating json from Arrow format: OK | 0/3 [00:00<, 70/s]
Training data uploaded to:
s3://sagemaker-ap-southeast-2-381491942632/datasets/text-to-sql/train_dataset.json
https://s3.console.aws.amazon.com/s3/buckets/sagemaker-ap-southeast-2-381491942632/regionap-southeast-2:prefix=datasets/text-to-sql/

Fine-tune LLM using trl on Amazon SageMaker

[ ]: # hyperparameters, which are passed into the training job
hyperparameters = [
    ## SCRIPT PARAMETERS ##
    'dataset_path': '/opt/ml/input/data/training/train_dataset.json', # path where sagemaker will save training dataset
    'model_id': 'code llama/CodeLlama-7b-hf', # or 'mistralai/Mistral-7B-v0.1'
    'max_seq_len': 512, # max sequence length for model and packing of the dataset
    'use_qlora': True, # use QLoRA model
    ## TRAINING PARAMETERS ##
    'num_train_epochs': 1, # number of training epochs
    'per_device_train_batch_size': 1, # batch size per device during training
    'gradient_accumulation_steps': 4, # number of steps before performing a backward/update pass
    'gradient_checkpointing': True, # use gradient checkpointing to save memory
    'optim': 'adamw_torch_fused', # use fused adamw optimizer
    'logging_steps': 10, # log every 10 steps
    'save_strategy': 'epoch', # save checkpoint every epoch
    'learning_rate': 2e-4, # learning rate, based on QLoRA paper
    'bf16': False, # use bf16 precision
    'tf32': True, # use tf32 precision
```

Conclusion:

In conclusion, deployment and training of the Llama-7b Language Model through Amazon SageMaker in the Sydney region represents a significant milestone in harnessing advanced natural language processing (NLP) technologies for practical applications. With meticulous configuration and optimization, we seamlessly integrated the Llama-7b model into the SageMaker ecosystem, enabling both scalability and high-performance inference capabilities.

This assignment not only showcased the robustness and versatility of SageMaker but also highlighted the immense potential for future advancements in NLP research and development. As we look ahead, continued exploration and refinement of language models like Llama-7b hold great promise. These models have the capacity to drive further breakthroughs in natural language understanding, ultimately reshaping human-machine interactions and catalyzing transformative innovations across AI-driven applications worldwide.

Name : Shivam Kumar
Section : CSE-28
Roll no : 21051684
Branch : CSE