UDACITY

Introduction to Generative AI with AWS

Project Documentation Report

Visit UDACITY Introduction to Generative AI with AWS Project Documentation Report to make a copy of this document.

Complete the answers to the questions below to complete your project report. Create a PDF of the completed document and submit the PDF with your project.

Question	Your answer:
Step 2: Domain Choice What domain did you choose to fine-tune the Meta Llama 2 7B model on? Choices: 1. Financial 2. Healthcare 3. IT	IT domain
Step 3: Model Evaluation Section What was the response of the model to your domain-specific input in the **model_evaluation.ipynb file**?	that they should be able to adapt to the changing contexts of their users. In order to do so, context information must be made available to applications. This thesis describes a system for context awareness in ubiquitous computing environments. The system is based on a context model that is used to describe the
Step 4: Fine-Tuning Section After fine-tuning the model, what was the response of the model to your domain-specific input in the **model_finetuning.ipynb file**?	the need for context-aware applications. The context-awareness is achieved through the use of context-aware agents. The context-aware agents are able to collect the data from the environment and to act upon it. In this paper, we present a context-aware agent architecture and an implementation of the agent in

1. In model_evaluation.ipynb, model deployed in the SageMaker environment

In [1]:
```
import sagemaker, boto3, json
from sagemaker.session import Session

sagemaker_session = Session()
aws_role = sagemaker_session.get_caller_identity_arn()
aws_region = boto3.Session().region_name
sess = sagemaker.Session()
print(aws_role)
print(aws_region)
print(sess)
```

```
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
arn:aws:iam::855105835472:role/service-role/SageMaker-UdacitySagemakerRole
us-west-2
<sagemaker.session.Session object at 0x7f4e66d72710>
```

2. fine-tuning cell output

In [2]:
```
from sagemaker.jumpstart.estimator import JumpStartEstimator
import boto3

estimator = JumpStartEstimator(model_id=model_id,  environment={"accept_eula": "true"},instance_type = "ml.g5.2xlarge")

estimator.set_hyperparameters(instruction_tuned="False", epoch="5")

#Fill in the code below with the dataset you want to use from above
#example: estimator.fit({"training": f"s3://genaiwithawsproject2024/training-datasets/finance"})
estimator.fit({ "training": f"s3://genaiwithawsproject2024/training-datasets/it" })
```

```
INFO:root:Key: avg_epoch_time, Value: 5.18719214260002
INFO:root:Key: avg_checkpoint_time, Value: 1.0363982738000233
INFO:root:Combining pre-trained base model with the PEFT adapter module.
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|█████     | 1/2 [00:29<00:29, 29.76s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:35<00:00, 15.70s/it]
Loading checkpoint shards: 100%|██████████| 2/2 [00:35<00:00, 17.81s/it]
INFO:root:Saving the combined model in safetensors format.
INFO:root:Saving complete.
INFO:root:Copying tokenizer to the output directory.
INFO:root:Putting inference code with the fine-tuned model directory.
2024-03-08 08:05:37,884 sagemaker-training-toolkit INFO     Waiting for the process to finish and give a return code.
2024-03-08 08:05:37,884 sagemaker-training-toolkit INFO     Done waiting for a return code. Received 0 from exiting process.
2024-03-08 08:05:37,885 sagemaker-training-toolkit INFO     Reporting training SUCCESS

2024-03-08 08:05:42 Uploading - Uploading generated training model
2024-03-08 08:06:28 Completed - Training job completed
Training seconds: 759
Billable seconds: 759
```

3. In model_finetuning.ipynb, deployed model in the SageMaker environment

In [3]:
```
finetuned_predictor = estimator.deploy()
```

```
No instance type selected for inference hosting endpoint. Defaulting to ml.g5.2xlarge.
INFO:sagemaker.jumpstart:No instance type selected for inference hosting endpoint. Defaulting to ml.g5.2xlarge.
INFO:sagemaker:Creating model with name: meta-textgeneration-llama-2-7b-2024-03-08-08-06-41-094
INFO:sagemaker:Creating endpoint-config with name meta-textgeneration-llama-2-7b-2024-03-08-08-06-41-088
INFO:sagemaker:Creating endpoint with name meta-textgeneration-llama-2-7b-2024-03-08-08-06-41-088
--------!
```

Evaluate the pre-trained and fine-tuned model

Next, we use the same input from the model evaluation step to evaluate the performance of the fine-tuned model and compare it with the base pre-trained model.