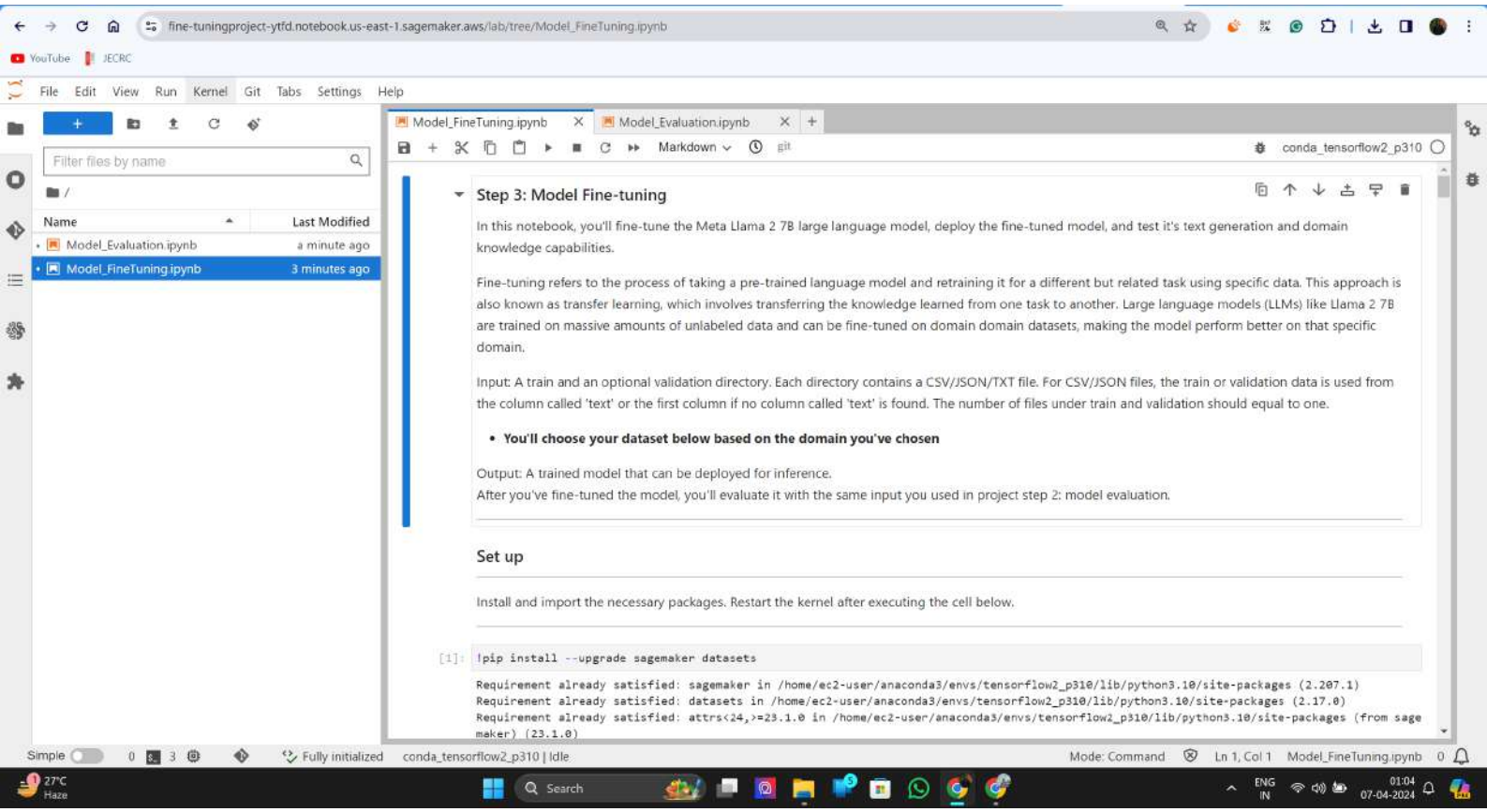
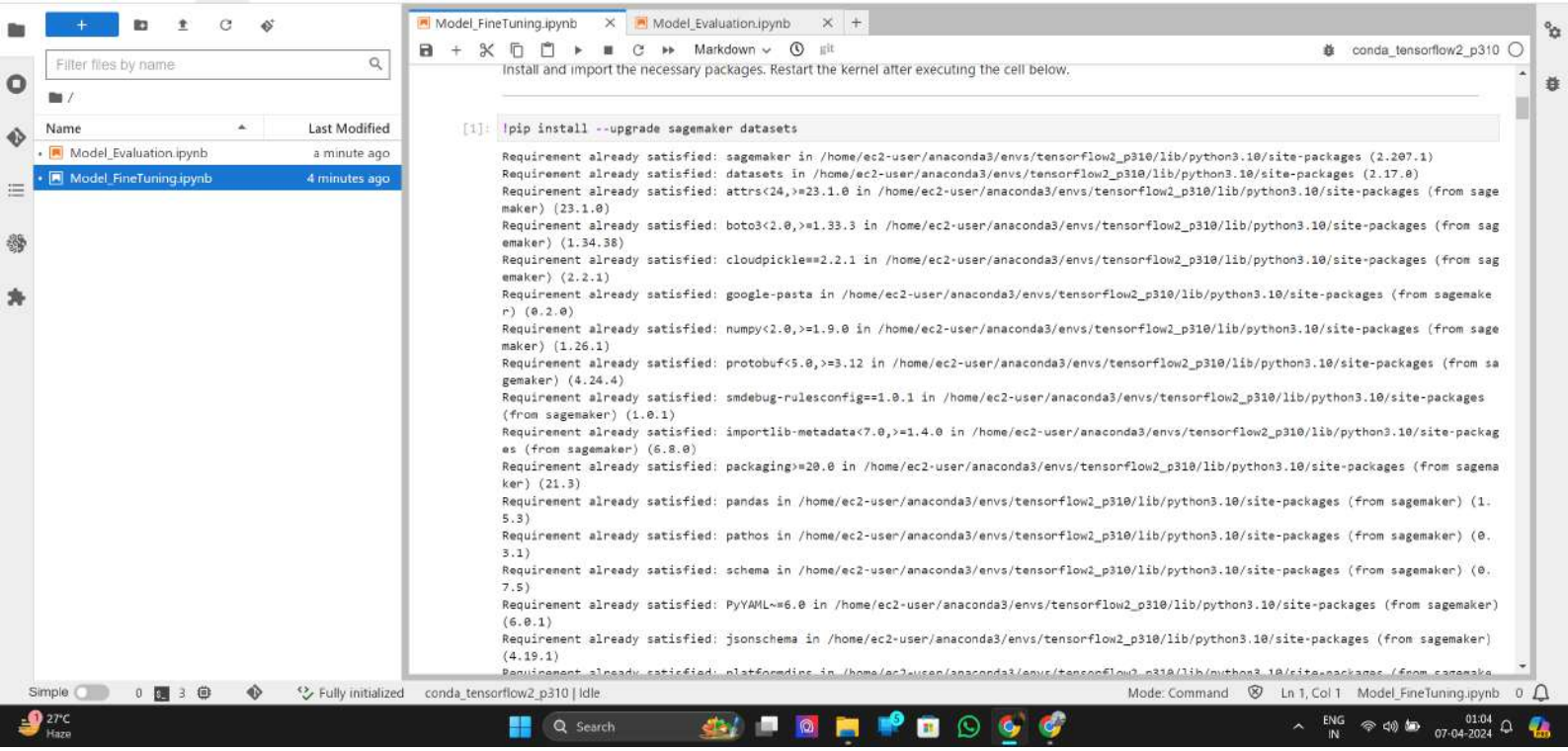
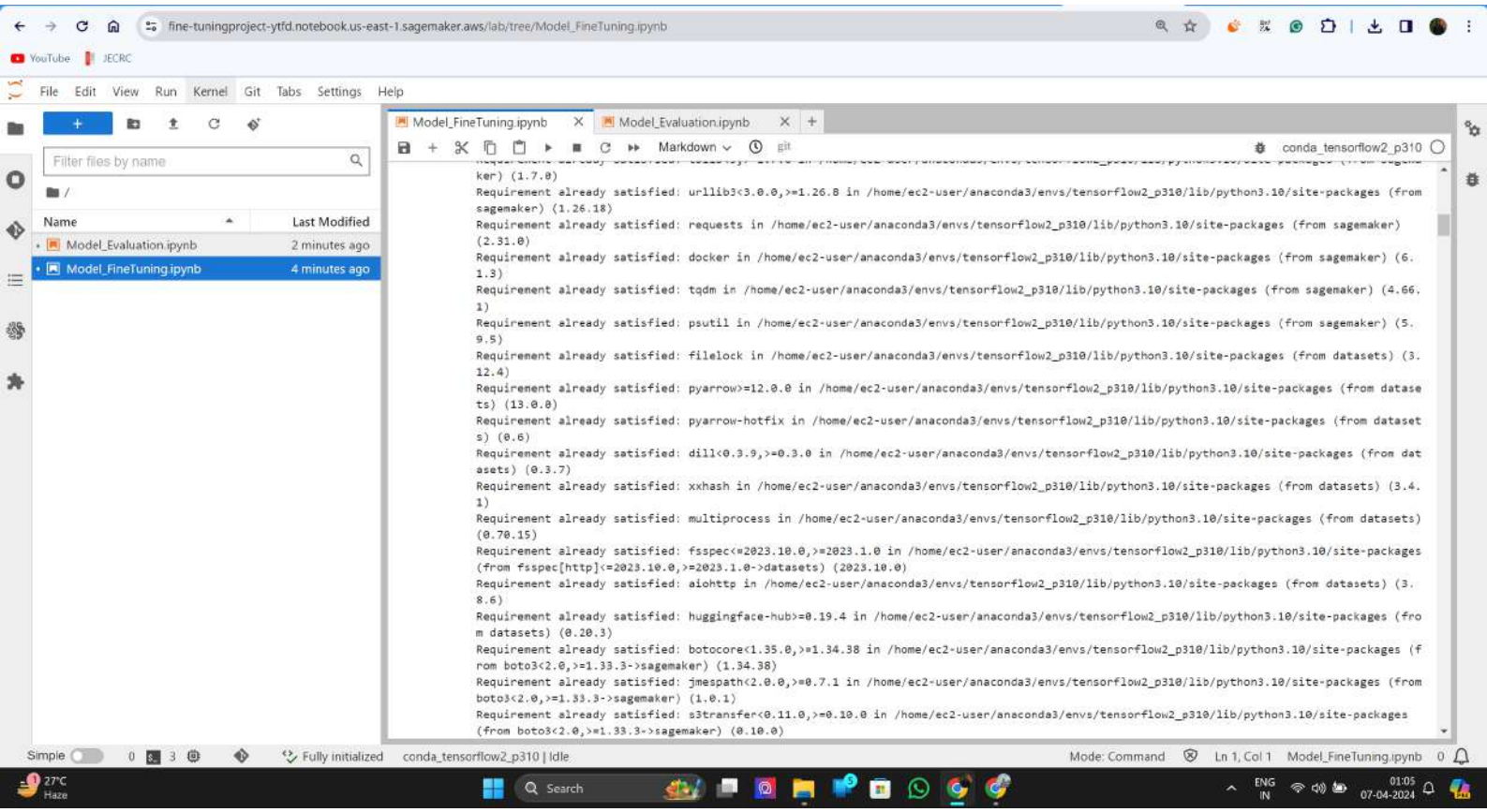
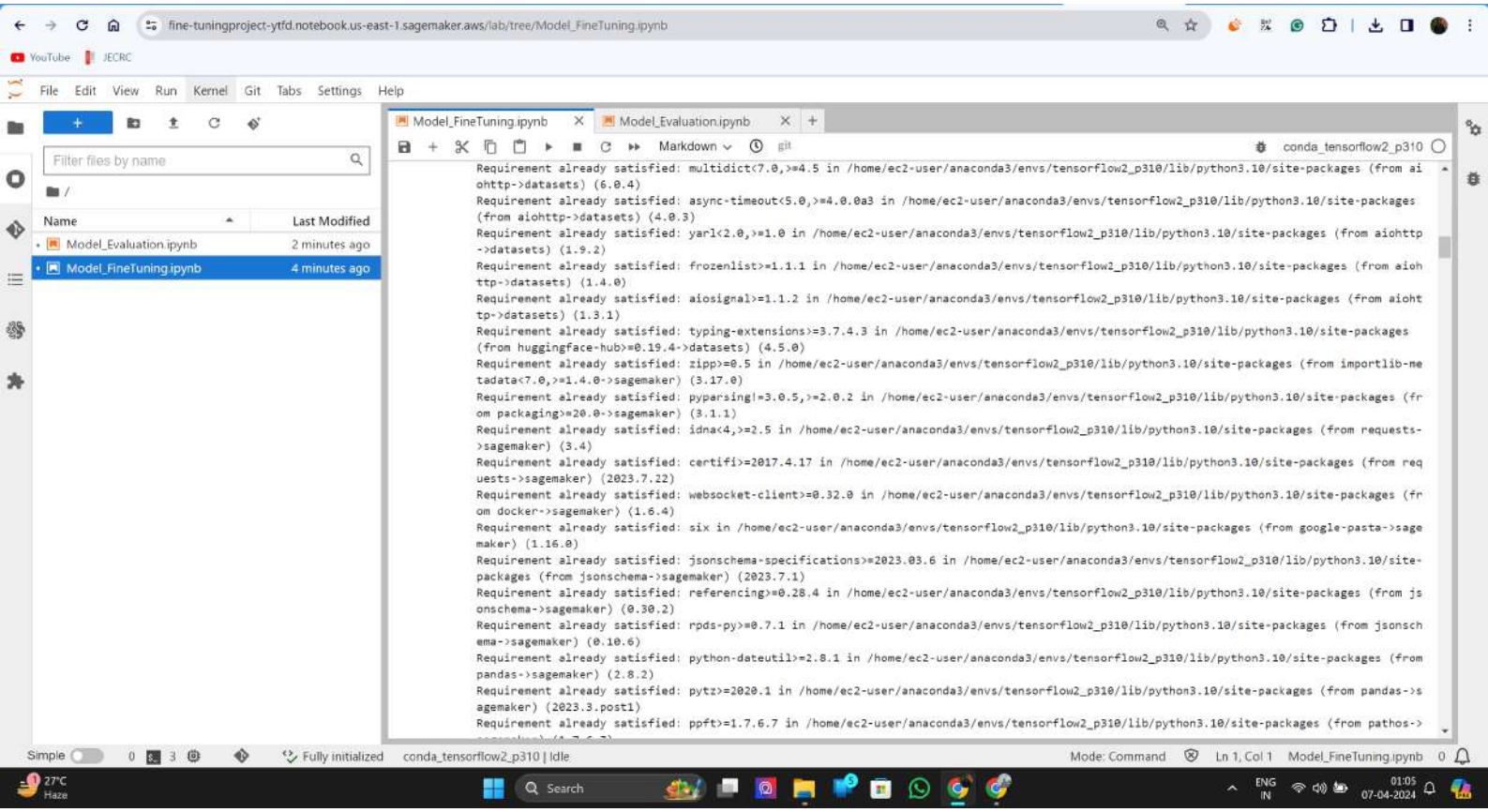
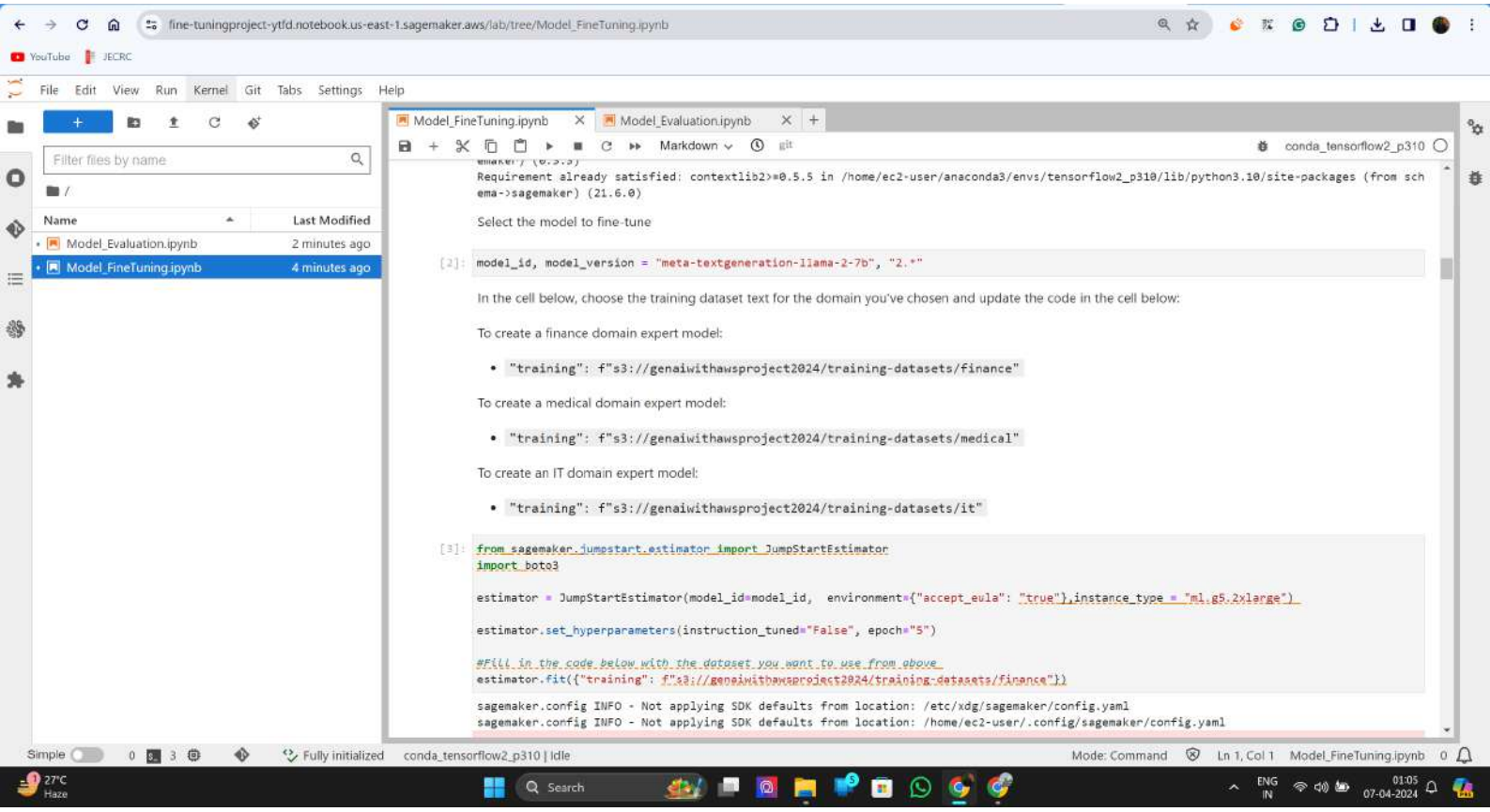
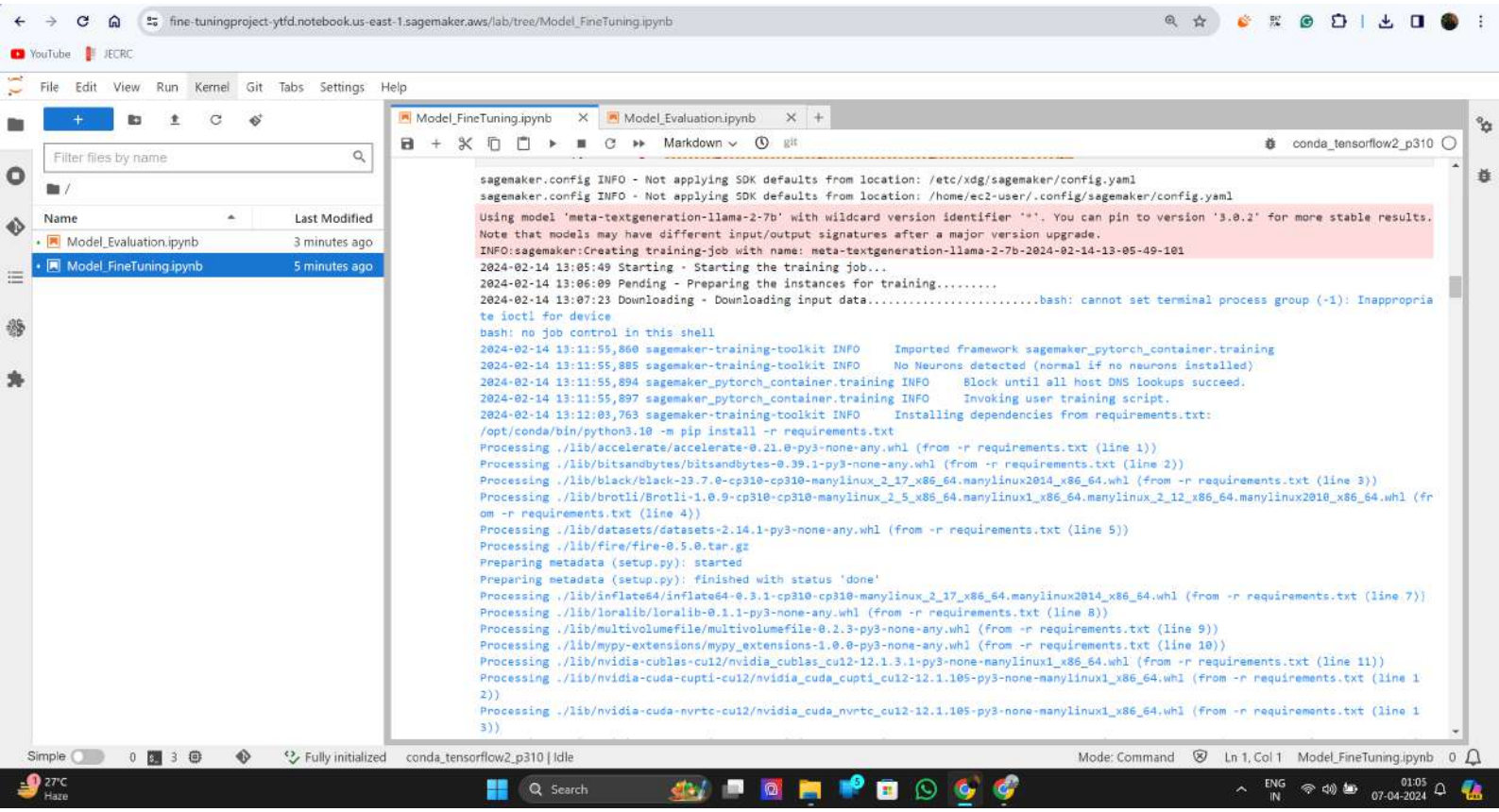

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
a minute ago
Model_FineTuning.ipynb
3 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
Step 3: Model Fine-tuning
In this notebook, you'll fine-tune the Meta Llama 2 7B large language model, deploy the fine-tuned model, and test it's text generation and domain knowledge capabilities.
Fine-tuning refers to the process of taking a pre-trained language model and retraining it for a different but related task using specific data. This approach is also known as transfer learning, which involves transferring the knowledge learned from one task to another. Large language models (LLMs) like Llama 2 7B are trained on massive amounts of unlabeled data and can be fine-tuned on domain domain datasets, making the model perform better on that specific domain.
Input: A train and an optional validation directory. Each directory contains a CSV/JSON/TXT file. For CSV/JSON files, the train or validation data is used from the column called 'text' or the first column if no column called 'text' is found. The number of files under train and validation should equal to one.
You'll choose your dataset below based on the domain you've chosen
Output: A trained model that can be deployed for inference.
After you've fine-tuned the model, you'll evaluate it with the same input you used in project step 2: model evaluation.
Set up
Install and import the necessary packages. Restart the kernel after executing the cell below.
[1]: !pip install --upgrade sagemaker datasets
Requirement already satisfied: sagemaker in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (2.207.1)
Requirement already satisfied: datasets in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (2.17.0)
Requirement already satisfied: attrs<24,>=23.1.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (23.1.0)
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:04
07-04-2024

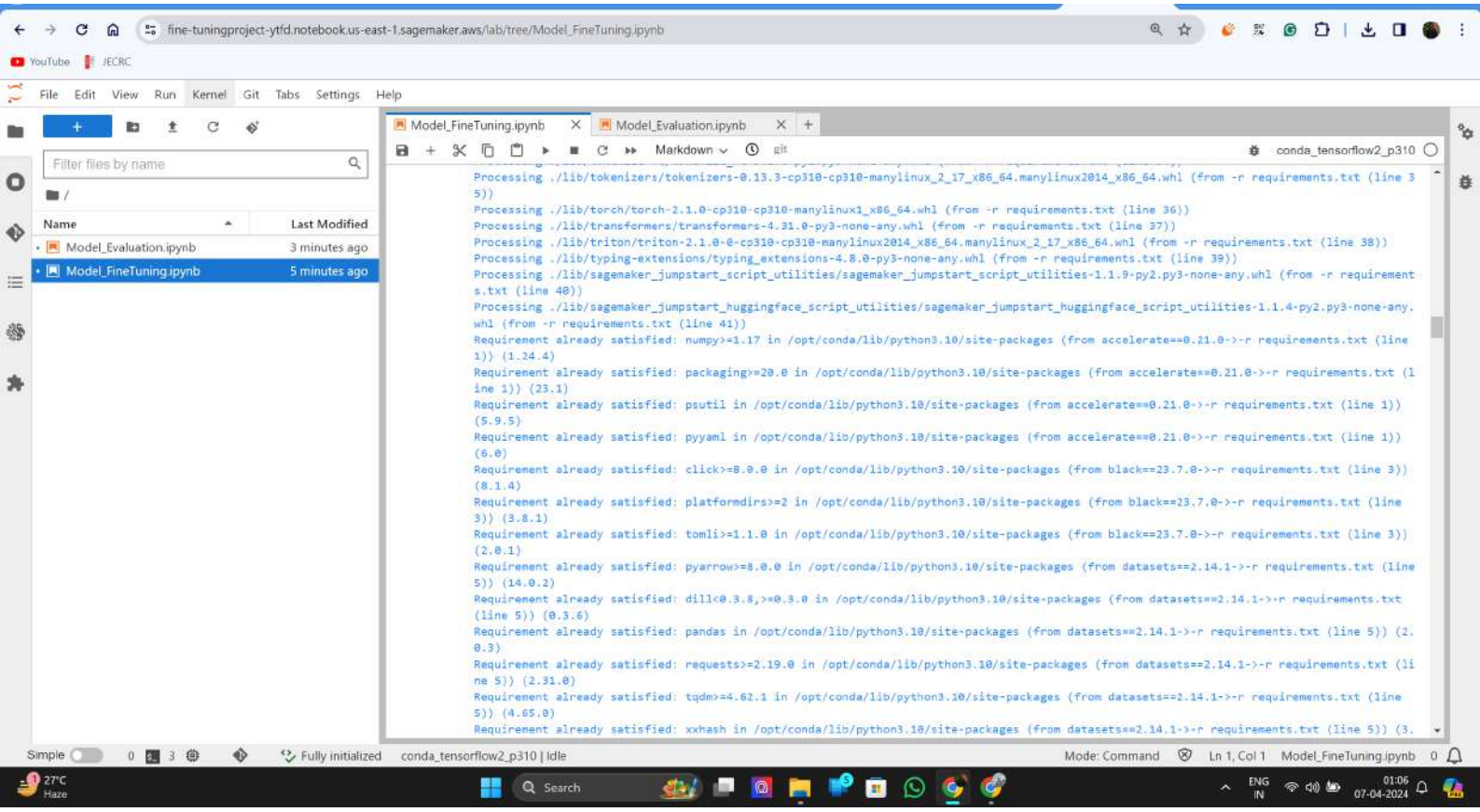

Model_FineTuning.ipynb
Model_Evaluation.ipynb
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
a minute ago
Model_FineTuning.ipynb
4 minutes ago
Markdown
conda_tensorflow2_p310
Install and import the necessary packages. Restart the kernel after executing the cell below.
[1]:
!pip install --upgrade sagemaker datasets
Requirement already satisfied: sagemaker in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (2.207.1)
Requirement already satisfied: datasets in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (2.17.0)
Requirement already satisfied: attrs<24,>=23.1.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (23.1.0)
Requirement already satisfied: boto3<2.0,>=1.33.3 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (1.34.38)
Requirement already satisfied: cloudpickle==2.2.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (2.2.1)
Requirement already satisfied: google-pasta in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (0.2.0)
Requirement already satisfied: numpy<2.0,>=1.9.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (1.26.1)
Requirement already satisfied: protobuf<5.0,>=3.12 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (4.24.4)
Requirement already satisfied: smdebug-rulesconfig==1.0.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (1.0.1)
Requirement already satisfied: importlib-metadata<7.0,>=1.4.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (6.8.0)
Requirement already satisfied: packaging>=20.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (21.3)
Requirement already satisfied: pandas in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (1.5.3)
Requirement already satisfied: pathos in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (0.3.1)
Requirement already satisfied: schema in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (0.7.5)
Requirement already satisfied: PyYAML~=6.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (6.0.1)
Requirement already satisfied: jsonschema in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (4.19.1)
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:04
07-04-2024

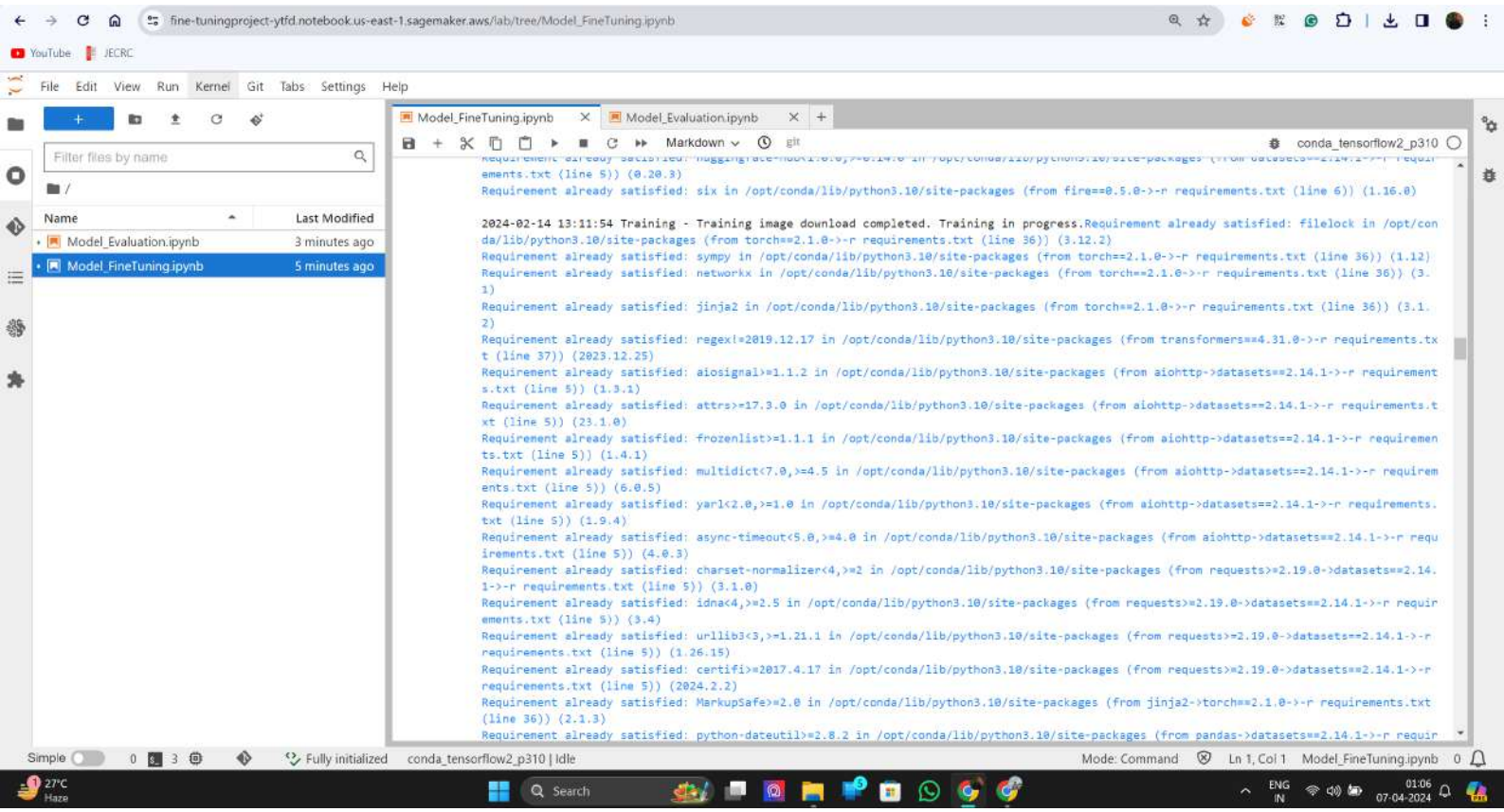

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
2 minutes ago
Model_FineTuning.ipynb
4 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
ker) (1.7.0)
Requirement already satisfied: urllib3<3.0.0,>=1.26.8 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (1.26.18)
Requirement already satisfied: requests in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (2.31.0)
Requirement already satisfied: docker in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (6.1.3)
Requirement already satisfied: tqdm in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (4.66.1)
Requirement already satisfied: psutil in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from sagemaker) (5.9.5)
Requirement already satisfied: filelock in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from datasets) (3.12.4)
Requirement already satisfied: pyarrow>=12.0.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from datasets) (13.0.0)
Requirement already satisfied: pyarrow-hotfix in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from datasets) (0.6)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from datasets) (0.3.7)
Requirement already satisfied: xxhash in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from datasets) (3.4.1)
Requirement already satisfied: multiprocess in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from datasets) (0.70.15)
Requirement already satisfied: fsspec<=2023.10.0,>=2023.1.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from fsspec[http]<=2023.10.0,>=2023.1.0->datasets) (2023.10.0)
Requirement already satisfied: aiohttp in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from datasets) (3.8.6)
Requirement already satisfied: huggingface-hub>=0.19.4 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from datasets) (0.20.3)
Requirement already satisfied: botocore<1.35.0,>=1.34.38 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from boto3<2.0,>=1.33.3->sagemaker) (1.34.38)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from boto3<2.0,>=1.33.3->sagemaker) (1.0.1)
Requirement already satisfied: s3transfer<0.11.0,>=0.10.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from boto3<2.0,>=1.33.3->sagemaker) (0.10.0)
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:05
07-04-2024

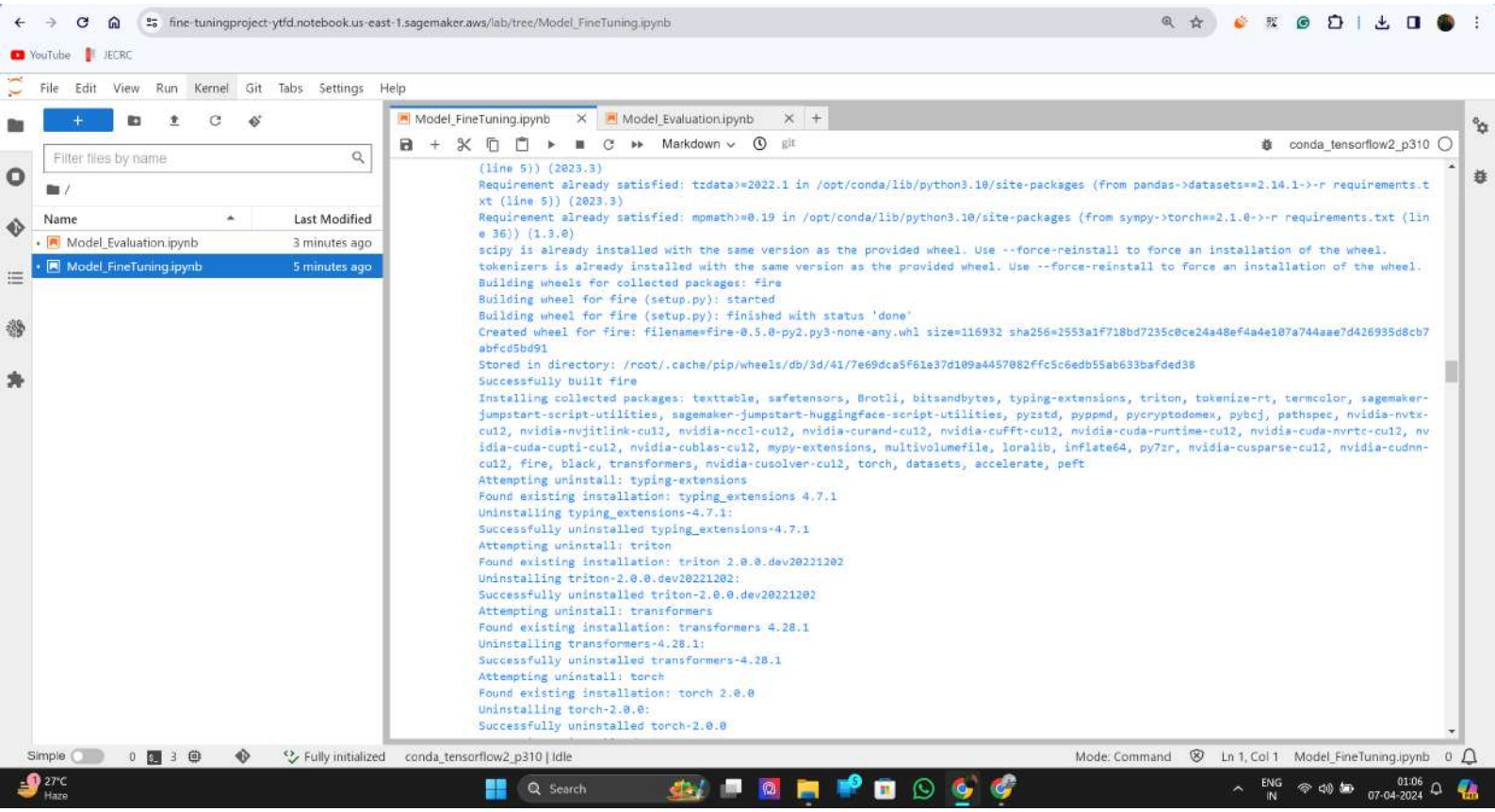

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 2 minutes ago
Model_FineTuning.ipynb 4 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
Requirement already satisfied: multidict<7.0,>=4.5 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from aiohttp->datasets) (6.0.4)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: yarl<2.0,>=1.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from aiohttp->datasets) (1.9.2)
Requirement already satisfied: frozenlist>=1.1.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from aiohttp->datasets) (1.4.0)
Requirement already satisfied: aiosignal>=1.1.2 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from huggingface-hub>=0.19.4->datasets) (4.5.0)
Requirement already satisfied: zipp>=0.5 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from importlib-metadata<7.0,>=1.4.0->sagemaker) (3.17.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from packaging>=20.0->sagemaker) (3.1.1)
Requirement already satisfied: idna<4,>=2.5 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from requests->sagemaker) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from requests->sagemaker) (2023.7.22)
Requirement already satisfied: websocket-client>=0.32.0 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from docker->sagemaker) (1.6.4)
Requirement already satisfied: six in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from google-pasta->sagemaker) (1.16.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from jsonschema->sagemaker) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from jsonschema->sagemaker) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from jsonschema->sagemaker) (0.10.6)
Requirement already satisfied: python-dateutil>=2.8.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from pandas->sagemaker) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from pandas->sagemaker) (2023.3.post1)
Requirement already satisfied: ppft>=1.7.6.7 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from pathos->
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:05
07-04-2024

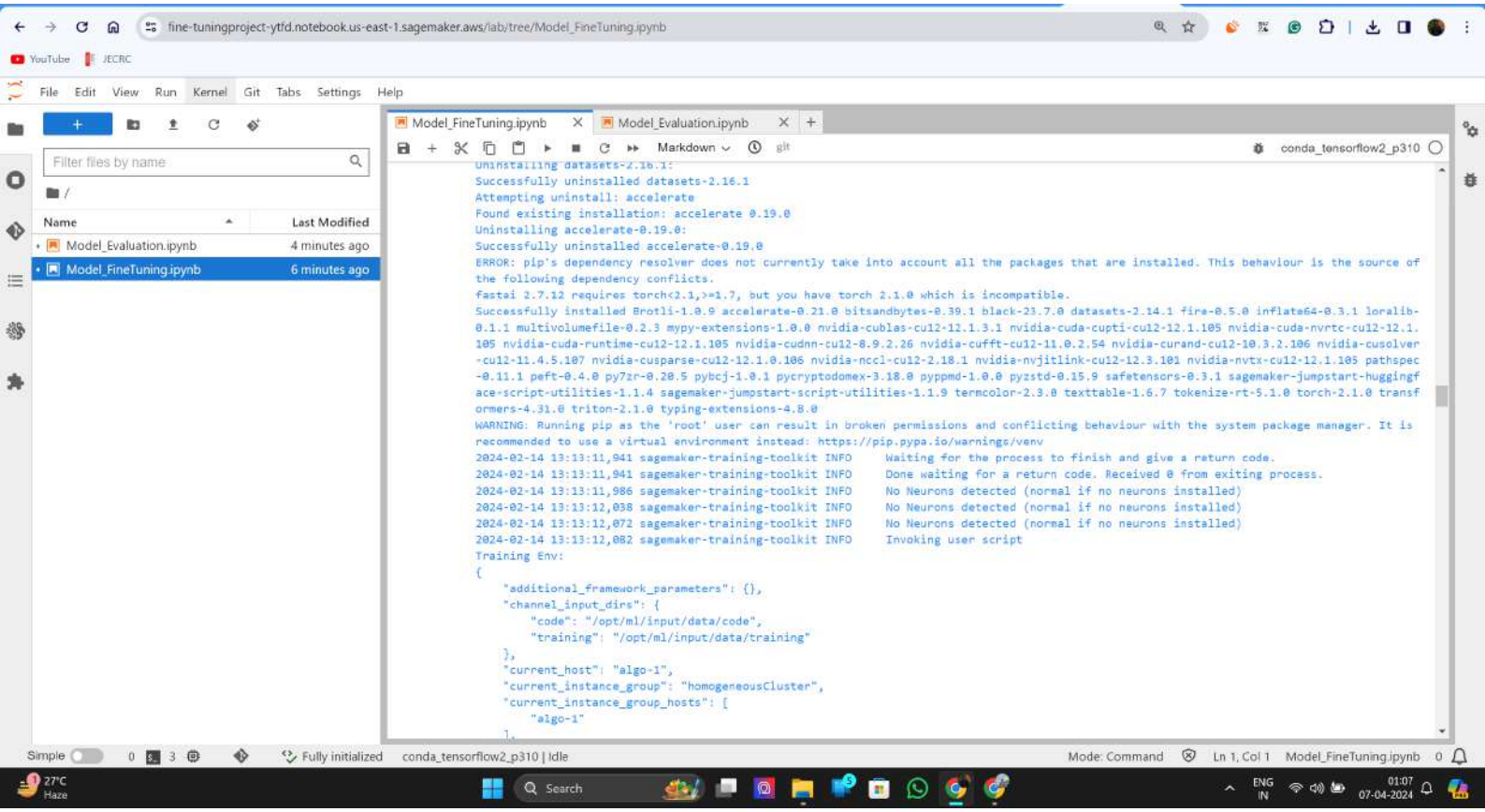

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
Name
Last Modified
Model_Evaluation.ipynb
2 minutes ago
Model_FineTuning.ipynb
4 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
conda_tensorflow2_p310
Requirement already satisfied: contextlib2>=0.5.5 in /home/ec2-user/anaconda3/envs/tensorflow2_p310/lib/python3.10/site-packages (from schema->sagemaker) (21.6.0)
Select the model to fine-tune
[2]:
model_id, model_version = "meta-textgeneration-llama-2-7b", "2.*"
In the cell below, choose the training dataset text for the domain you've chosen and update the code in the cell below:
To create a finance domain expert model:
"training": f"s3://genaiwithawsproject2024/training-datasets/finance"
To create a medical domain expert model:
"training": f"s3://genaiwithawsproject2024/training-datasets/medical"
To create an IT domain expert model:
"training": f"s3://genaiwithawsproject2024/training-datasets/it"
[3]:
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
estimator.fit({"training": f"s3://genaiwithawsproject2024/training-datasets/finance"})
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
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:05
07-04-2024

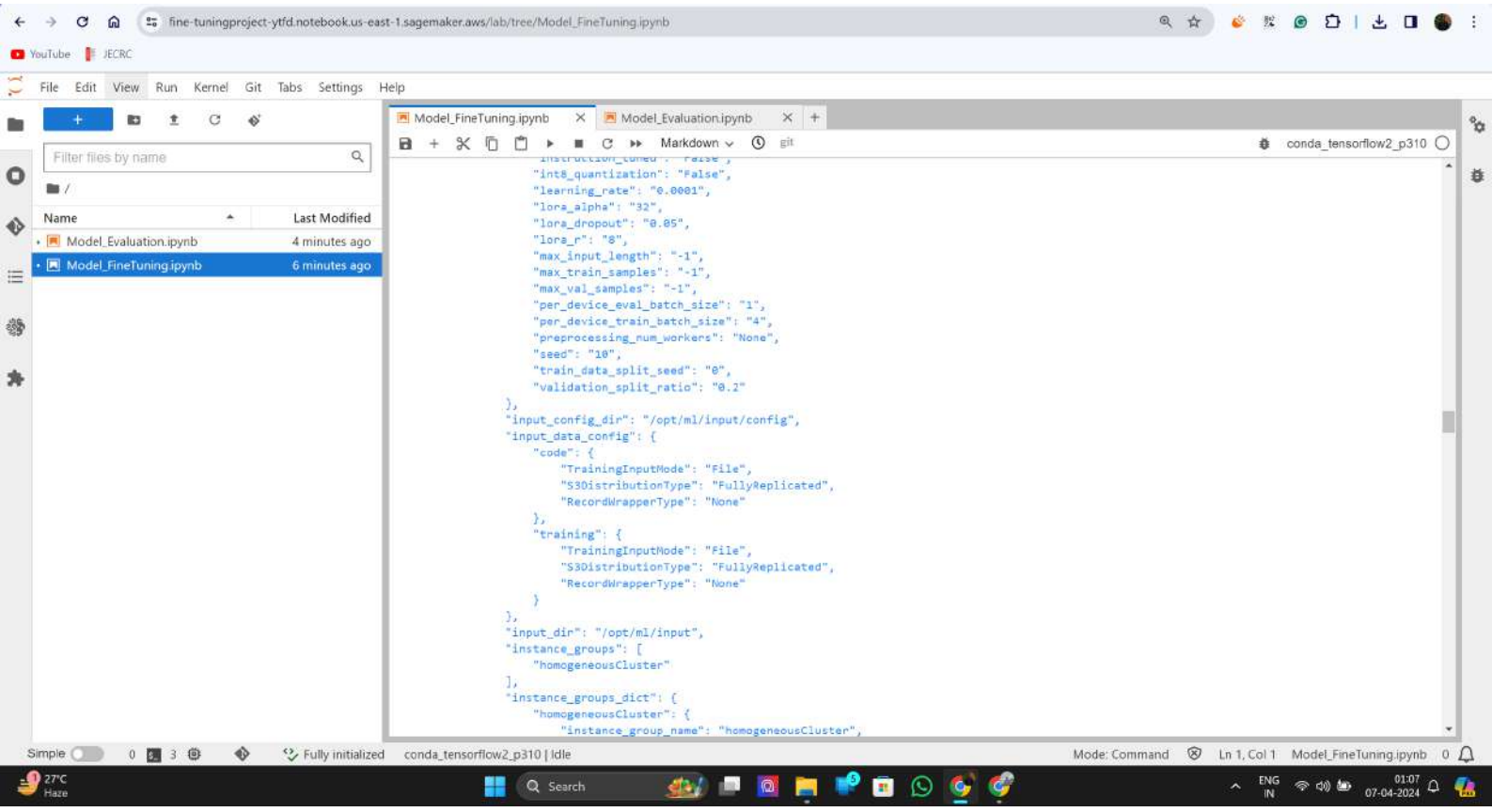

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 3 minutes ago
Model_FineTuning.ipynb 5 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
conda_tensorflow2_p310
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
Using model 'meta-textgeneration-llama-2-7b' with wildcard version identifier '*'. You can pin to version '3.0.2' for more stable results.
Note that models may have different input/output signatures after a major version upgrade.
INFO:sagemaker:Creating training-job with name: meta-textgeneration-llama-2-7b-2024-02-14-13-05-49-101
2024-02-14 13:05:49 Starting - Starting the training job...
2024-02-14 13:06:09 Pending - Preparing the instances for training.....
2024-02-14 13:07:23 Downloading - Downloading input data.....bash: cannot set terminal process group (-1): Inappropriate ioctl for device
bash: no job control in this shell
2024-02-14 13:11:55,860 sagemaker-training-toolkit INFO Imported framework sagemaker_pytorch_container.training
2024-02-14 13:11:55,885 sagemaker-training-toolkit INFO No Neurons detected (normal if no neurons installed)
2024-02-14 13:11:55,894 sagemaker_pytorch_container.training INFO Block until all host DNS lookups succeed.
2024-02-14 13:11:55,897 sagemaker_pytorch_container.training INFO Invoking user training script.
2024-02-14 13:12:03,763 sagemaker-training-toolkit INFO Installing dependencies from requirements.txt:
/opt/conda/bin/python3.10 -m pip install -r requirements.txt
Processing ./lib/accelerate/accelerate-0.21.0-py3-none-any.whl (from -r requirements.txt (line 1))
Processing ./lib/bitsandbytes/bitsandbytes-0.39.1-py3-none-any.whl (from -r requirements.txt (line 2))
Processing ./lib/black/black-23.7.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (line 3))
Processing ./lib/brotli/Brotli-1.0.9-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (from -r requirements.txt (line 4))
Processing ./lib/datasets/datasets-2.14.1-py3-none-any.whl (from -r requirements.txt (line 5))
Processing ./lib/fire/fire-0.5.0.tar.gz
Preparing metadata (setup.py): started
Preparing metadata (setup.py): finished with status 'done'
Processing ./lib/inflate64/inflate64-0.3.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (line 7))
Processing ./lib/loralib/loralib-0.1.1-py3-none-any.whl (from -r requirements.txt (line 8))
Processing ./lib/multivolumefile/multivolumefile-0.2.3-py3-none-any.whl (from -r requirements.txt (line 9))
Processing ./lib/mypy-extensions/mypy_extensions-1.0.0-py3-none-any.whl (from -r requirements.txt (line 10))
Processing ./lib/nvidia-cublas-cu12/nvidia_cublas_cu12-12.1.3.1-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (line 11))
Processing ./lib/nvidia-cuda-cupti-cu12/nvidia_cuda_cupti_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (line 12))
Processing ./lib/nvidia-cuda-nvrtc-cu12/nvidia_cuda_nvrtc_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (from -r requirements.txt (line 13))
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C Haze
Search
ENG IN
01:05 07-04-2024

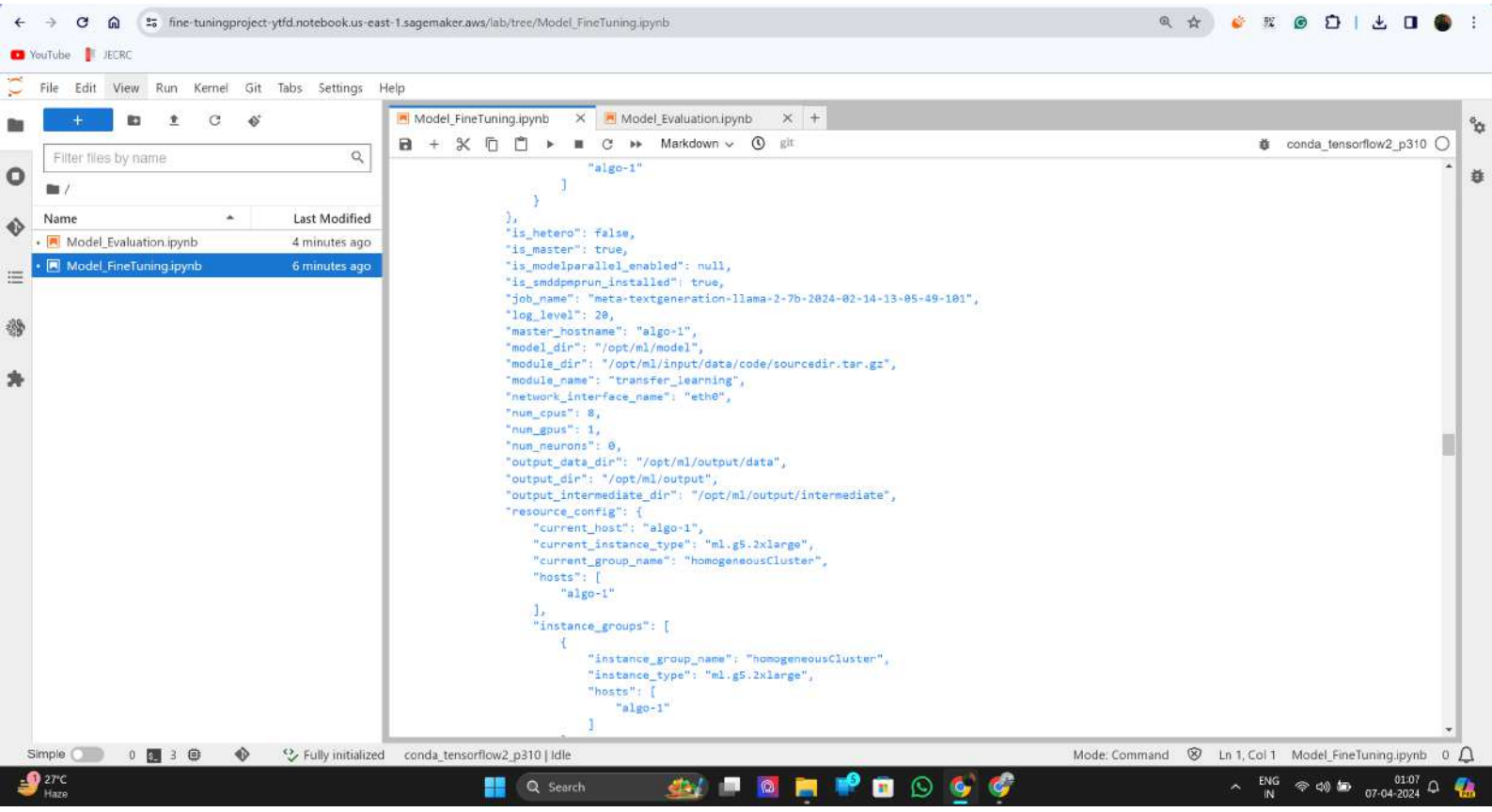

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 3 minutes ago
Model_FineTuning.ipynb 5 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
Processing ./lib/tokenizers/tokenizers-0.13.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (from -r requirements.txt (line 35))
Processing ./lib/torch/torch-2.1.0-cp310-cp310-manylinux1_x86_64.whl (from -r requirements.txt (line 36))
Processing ./lib/transformers/transformers-4.31.0-py3-none-any.whl (from -r requirements.txt (line 37))
Processing ./lib/triton/triton-2.1.0-0-cp310-cp310-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (from -r requirements.txt (line 38))
Processing ./lib/typing-extensions/typing_extensions-4.8.0-py3-none-any.whl (from -r requirements.txt (line 39))
Processing ./lib/sagemaker_jumpstart_script_utilities/sagemaker_jumpstart_script_utilities-1.1.9-py2.py3-none-any.whl (from -r requirements.txt (line 40))
Processing ./lib/sagemaker_jumpstart_huggingface_script_utilities/sagemaker_jumpstart_huggingface_script_utilities-1.1.4-py2.py3-none-any.whl (from -r requirements.txt (line 41))
Requirement already satisfied: numpy>=1.17 in /opt/conda/lib/python3.10/site-packages (from accelerate==0.21.0->-r requirements.txt (line 1)) (1.24.4)
Requirement already satisfied: packaging>=20.0 in /opt/conda/lib/python3.10/site-packages (from accelerate==0.21.0->-r requirements.txt (line 1)) (23.1)
Requirement already satisfied: psutil in /opt/conda/lib/python3.10/site-packages (from accelerate==0.21.0->-r requirements.txt (line 1)) (5.9.5)
Requirement already satisfied: pyyaml in /opt/conda/lib/python3.10/site-packages (from accelerate==0.21.0->-r requirements.txt (line 1)) (6.0)
Requirement already satisfied: click>=8.0.0 in /opt/conda/lib/python3.10/site-packages (from black==23.7.0->-r requirements.txt (line 3)) (8.1.4)
Requirement already satisfied: platformdirs>=2 in /opt/conda/lib/python3.10/site-packages (from black==23.7.0->-r requirements.txt (line 3)) (3.8.1)
Requirement already satisfied: tomli>=1.1.0 in /opt/conda/lib/python3.10/site-packages (from black==23.7.0->-r requirements.txt (line 3)) (2.0.1)
Requirement already satisfied: pyarrow>=8.0.0 in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt (line 5)) (14.0.2)
Requirement already satisfied: dill<0.3.8,>=0.3.0 in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt (line 5)) (0.3.6)
Requirement already satisfied: pandas in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt (line 5)) (2.0.3)
Requirement already satisfied: requests>=2.19.0 in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt (line 5)) (2.31.0)
Requirement already satisfied: tqdm>=4.62.1 in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt (line 5)) (4.65.0)
Requirement already satisfied: xxhash in /opt/conda/lib/python3.10/site-packages (from datasets==2.14.1->-r requirements.txt (line 5)) (3.
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:06
07-04-2024

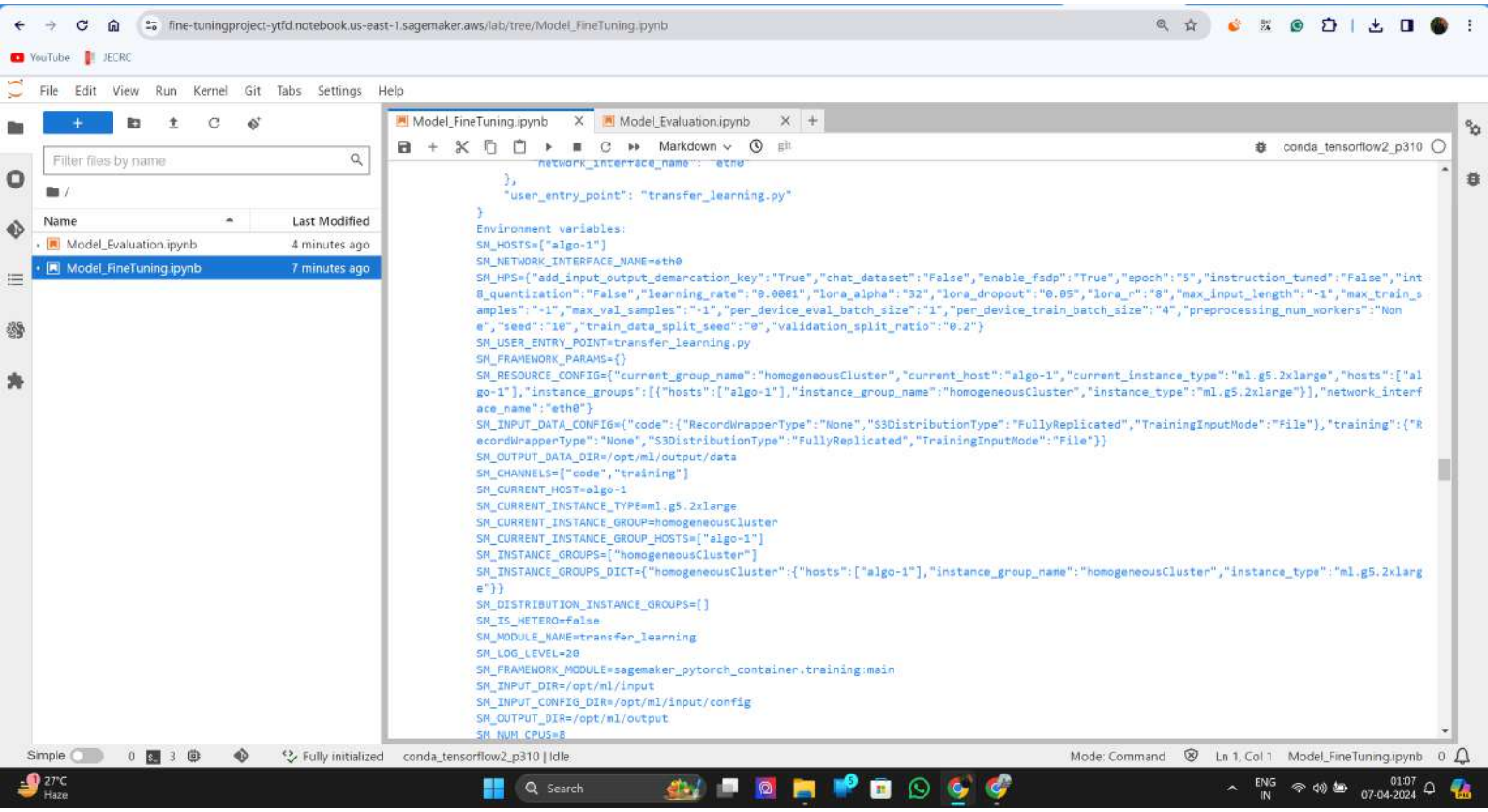

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 3 minutes ago
Model_FineTuning.ipynb 5 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
ements.txt (line 5)) (0.20.3)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packages (from fire==0.5.0->-r requirements.txt (line 6)) (1.16.0)
2024-02-14 13:11:54 Training - Training image download completed. Training in progress.Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-packages (from torch==2.1.0->-r requirements.txt (line 36)) (3.12.2)
Requirement already satisfied: sympy in /opt/conda/lib/python3.10/site-packages (from torch==2.1.0->-r requirements.txt (line 36)) (1.12)
Requirement already satisfied: networkx in /opt/conda/lib/python3.10/site-packages (from torch==2.1.0->-r requirements.txt (line 36)) (3.1)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.10/site-packages (from torch==2.1.0->-r requirements.txt (line 36)) (3.1.2)
Requirement already satisfied: regex!=2019.12.17 in /opt/conda/lib/python3.10/site-packages (from transformers==4.31.0->-r requirements.txt (line 37)) (2023.12.25)
Requirement already satisfied: aiosignal>=1.1.2 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-r requirements.txt (line 5)) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-r requirements.txt (line 5)) (23.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-r requirements.txt (line 5)) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-r requirements.txt (line 5)) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-r requirements.txt (line 5)) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets==2.14.1->-r requirements.txt (line 5)) (4.0.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/python3.10/site-packages (from requests>=2.19.0->datasets==2.14.1->-r requirements.txt (line 5)) (3.1.0)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-packages (from requests>=2.19.0->datasets==2.14.1->-r requirements.txt (line 5)) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/lib/python3.10/site-packages (from requests>=2.19.0->datasets==2.14.1->-r requirements.txt (line 5)) (1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.10/site-packages (from requests>=2.19.0->datasets==2.14.1->-r requirements.txt (line 5)) (2024.2.2)
Requirement already satisfied: MarkupSafe>=2.0 in /opt/conda/lib/python3.10/site-packages (from jinja2->torch==2.1.0->-r requirements.txt (line 36)) (2.1.3)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/conda/lib/python3.10/site-packages (from pandas->datasets==2.14.1->-r requir
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C Haze
Search
ENG IN
01:06 07-04-2024

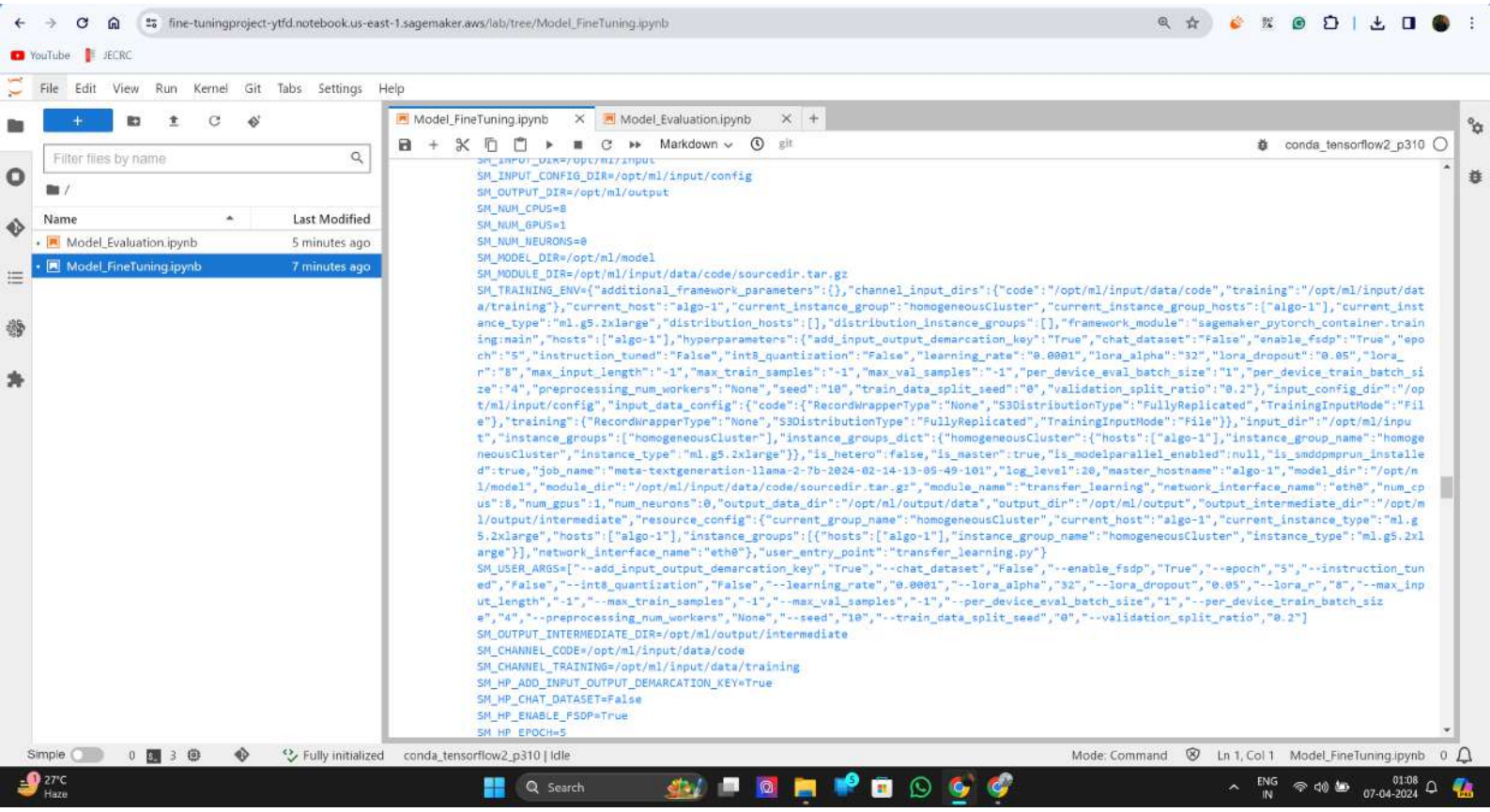

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 3 minutes ago
Model_FineTuning.ipynb 5 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
(line 5)) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in /opt/conda/lib/python3.10/site-packages (from pandas->datasets==2.14.1->-r requirements.txt (line 5)) (2023.3)
Requirement already satisfied: mpmath>=0.19 in /opt/conda/lib/python3.10/site-packages (from sympy->torch==2.1.0->-r requirements.txt (line 36)) (1.3.0)
scipy is already installed with the same version as the provided wheel. Use --force-reinstall to force an installation of the wheel.
tokenizers is already installed with the same version as the provided wheel. Use --force-reinstall to force an installation of the wheel.
Building wheels for collected packages: fire
Building wheel for fire (setup.py): started
Building wheel for fire (setup.py): finished with status 'done'
Created wheel for fire: filename=fire-0.5.0-py2.py3-none-any.whl size=116932 sha256=2553a1f718bd7235c0ce24a48ef4a4e107a744aae7d426935d8cb7abfcd5bd91
Stored in directory: /root/.cache/pip/wheels/db/3d/41/7e69dca5f61e37d109a4457082ffc5c6edb55ab633bafded38
Successfully built fire
Installing collected packages: texttable, safetensors, Brotli, bitsandbytes, typing-extensions, triton, tokenize-rt, termcolor, sagemaker-jumpstart-script-utilities, sagemaker-jumpstart-huggingface-script-utilities, pyzstd, pyppmd, pycryptodomex, pybcj, pathspec, nvidia-nvtx-cu12, nvidia-nvjitlink-cu12, nvidia-nccl-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, mypy-extensions, multivolumefile, loralib, inflate64, py7zr, nvidia-cusparse-cu12, nvidia-cudnn-cu12, fire, black, transformers, nvidia-cusolver-cu12, torch, datasets, accelerate, peft
Attempting uninstall: typing-extensions
Found existing installation: typing_extensions 4.7.1
Uninstalling typing_extensions-4.7.1:
Successfully uninstalled typing_extensions-4.7.1
Attempting uninstall: triton
Found existing installation: triton 2.0.0.dev20221202
Uninstalling triton-2.0.0.dev20221202:
Successfully uninstalled triton-2.0.0.dev20221202
Attempting uninstall: transformers
Found existing installation: transformers 4.28.1
Uninstalling transformers-4.28.1:
Successfully uninstalled transformers-4.28.1
Attempting uninstall: torch
Found existing installation: torch 2.0.0
Uninstalling torch-2.0.0:
Successfully uninstalled torch-2.0.0
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C Haze
Search
ENG IN
01:06 07-04-2024

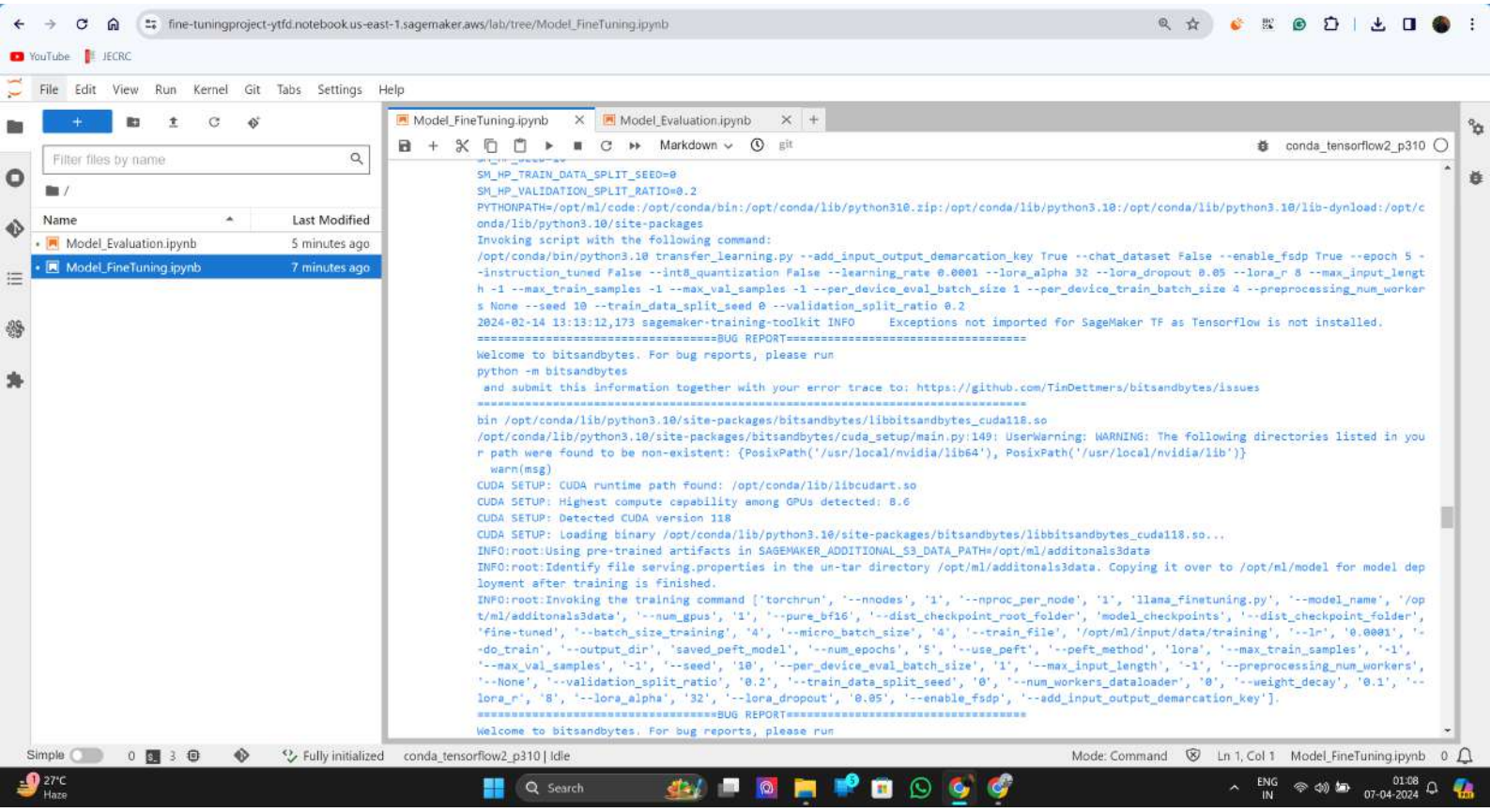

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 4 minutes ago
Model_FineTuning.ipynb 6 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
Uninstalling datasets-2.16.1:
Successfully uninstalled datasets-2.16.1
Attempting uninstall: accelerate
Found existing installation: accelerate 0.19.0
Uninstalling accelerate-0.19.0:
Successfully uninstalled accelerate-0.19.0
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts.
fastai 2.7.12 requires torch<2.1,>=1.7, but you have torch 2.1.0 which is incompatible.
Successfully installed Brotli-1.0.9 accelerate-0.21.0 bitsandbytes-0.39.1 black-23.7.0 datasets-2.14.1 fire-0.5.0 inflate64-0.3.1 loralib-0.1.1 multivolumefile-0.2.3 mypy-extensions-1.0.0 nvidia-cublas-cu12-12.1.3.1 nvidia-cuda-cupti-cu12-12.1.105 nvidia-cuda-nvrtc-cu12-12.1.105 nvidia-cuda-runtime-cu12-12.1.105 nvidia-cudnn-cu12-8.9.2.26 nvidia-cufft-cu12-11.0.2.54 nvidia-curand-cu12-10.3.2.106 nvidia-cusolver-cu12-11.4.5.107 nvidia-cusparse-cu12-12.1.0.106 nvidia-nccl-cu12-2.18.1 nvidia-nvjitlink-cu12-12.3.101 nvidia-nvtx-cu12-12.1.105 pathspec-0.11.1 peft-0.4.0 py7zr-0.20.5 pybcj-1.0.1 pycryptodomex-3.18.0 pyppmd-1.0.0 pyzstd-0.15.9 safetensors-0.3.1 sagemaker-jumpstart-huggingface-script-utilities-1.1.4 sagemaker-jumpstart-script-utilities-1.1.9 termcolor-2.3.0 texttable-1.6.7 tokenize-rt-5.1.0 torch-2.1.0 transformers-4.31.0 triton-2.1.0 typing-extensions-4.8.0
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
2024-02-14 13:13:11,941 sagemaker-training-toolkit INFO Waiting for the process to finish and give a return code.
2024-02-14 13:13:11,941 sagemaker-training-toolkit INFO Done waiting for a return code. Received 0 from exiting process.
2024-02-14 13:13:11,986 sagemaker-training-toolkit INFO No Neurons detected (normal if no neurons installed)
2024-02-14 13:13:12,038 sagemaker-training-toolkit INFO No Neurons detected (normal if no neurons installed)
2024-02-14 13:13:12,072 sagemaker-training-toolkit INFO No Neurons detected (normal if no neurons installed)
2024-02-14 13:13:12,082 sagemaker-training-toolkit INFO Invoking user script
Training Env:
{
"additional_framework_parameters": {},
"channel_input_dirs": {
"code": "/opt/ml/input/data/code",
"training": "/opt/ml/input/data/training"
},
"current_host": "algo-1",
"current_instance_group": "homogeneousCluster",
"current_instance_group_hosts": [
"algo-1"
],
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C Haze
Search
ENG IN
01:07 07-04-2024

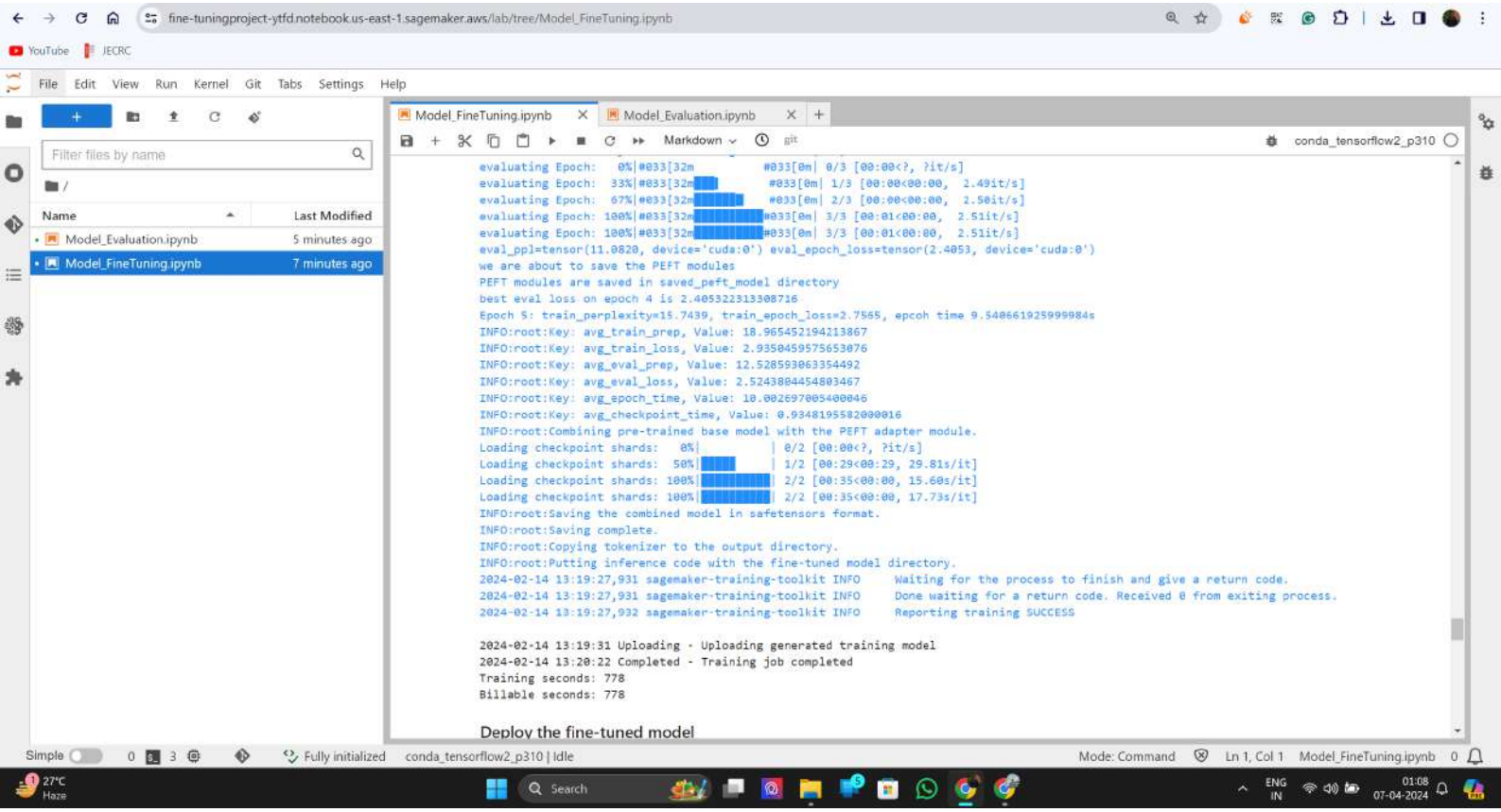

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 4 minutes ago
Model_FineTuning.ipynb 6 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
"int8_quantization": "False",
"learning_rate": "0.0001",
"lora_alpha": "32",
"lora_dropout": "0.05",
"lora_r": "8",
"max_input_length": "-1",
"max_train_samples": "-1",
"max_val_samples": "-1",
"per_device_eval_batch_size": "1",
"per_device_train_batch_size": "4",
"preprocessing_num_workers": "None",
"seed": "10",
"train_data_split_seed": "0",
"validation_split_ratio": "0.2"
},
"input_config_dir": "/opt/ml/input/config",
"input_data_config": {
"code": {
"TrainingInputMode": "File",
"S3DistributionType": "FullyReplicated",
"RecordWrapperType": "None"
},
"training": {
"TrainingInputMode": "File",
"S3DistributionType": "FullyReplicated",
"RecordWrapperType": "None"
}
},
"input_dir": "/opt/ml/input",
"instance_groups": [
"homogeneousCluster"
],
"instance_groups_dict": {
"homogeneousCluster": {
"instance_group_name": "homogeneousCluster",
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:07
07-04-2024

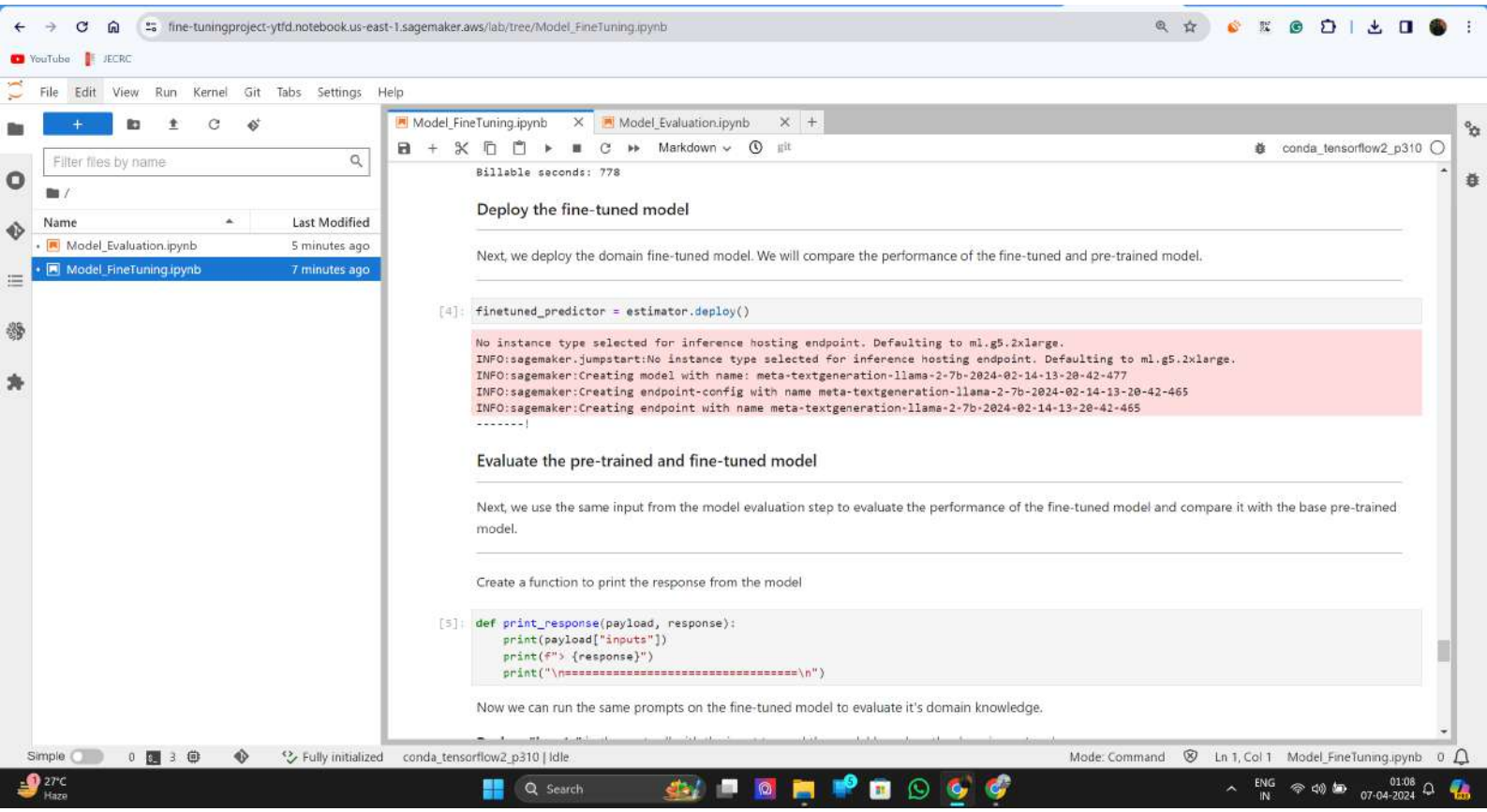

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 4 minutes ago
Model_FineTuning.ipynb 6 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
"algo-1"
]
}
},
"is_hetero": false,
"is_master": true,
"is_modelparallel_enabled": null,
"is_smddpmprun_installed": true,
"job_name": "meta-textgeneration-llama-2-7b-2024-02-14-13-05-49-101",
"log_level": 20,
"master_hostname": "algo-1",
"model_dir": "/opt/ml/model",
"module_dir": "/opt/ml/input/data/code/sourcedir.tar.gz",
"module_name": "transfer_learning",
"network_interface_name": "eth0",
"num_cpus": 8,
"num_gpus": 1,
"num_neurons": 0,
"output_data_dir": "/opt/ml/output/data",
"output_dir": "/opt/ml/output",
"output_intermediate_dir": "/opt/ml/output/intermediate",
"resource_config": {
"current_host": "algo-1",
"current_instance_type": "ml.g5.2xlarge",
"current_group_name": "homogeneousCluster",
"hosts": [
"algo-1"
],
"instance_groups": [
{
"instance_group_name": "homogeneousCluster",
"instance_type": "ml.g5.2xlarge",
"hosts": [
"algo-1"
]
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:07
07-04-2024

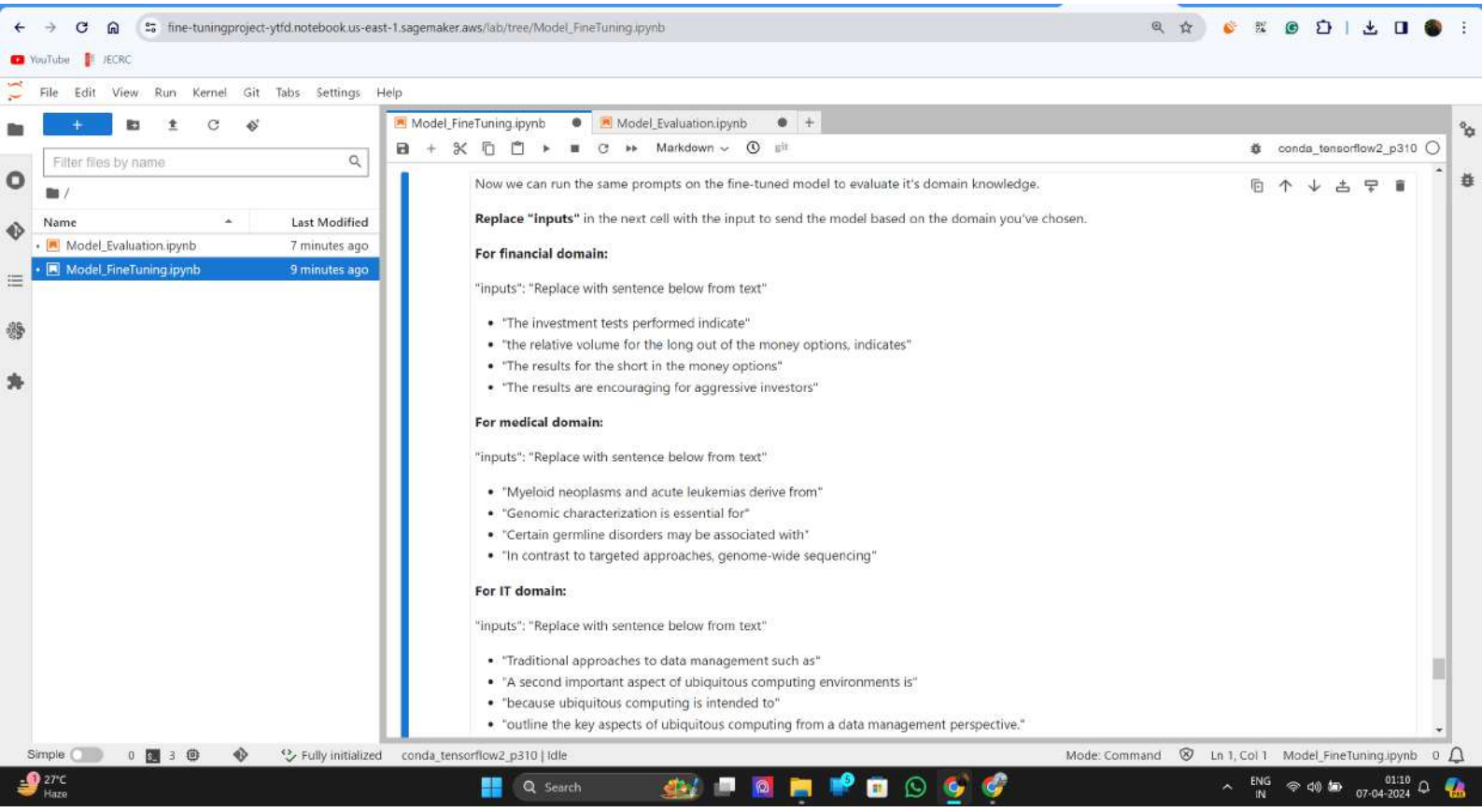

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
4 minutes ago
Model_FineTuning.ipynb
7 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
"network_interface_name": "eth0"
},
"user_entry_point": "transfer_learning.py"
}
Environment variables:
SM_HOSTS=["algo-1"]
SM_NETWORK_INTERFACE_NAME=eth0
SM_HPS={"add_input_output_demarcation_key":"True","chat_dataset":"False","enable_fsdp":"True","epoch":"5","instruction_tuned":"False","int8_quantization":"False","learning_rate":"0.0001","lora_alpha":"32","lora_dropout":"0.05","lora_r":"8","max_input_length":"-1","max_train_samples":"-1","max_val_samples":"-1","per_device_eval_batch_size":"1","per_device_train_batch_size":"4","preprocessing_num_workers":"None","seed":"10","train_data_split_seed":"0","validation_split_ratio":"0.2"}
SM_USER_ENTRY_POINT=transfer_learning.py
SM_FRAMEWORK_PARAMS={}
SM_RESOURCE_CONFIG={"current_group_name":"homogeneousCluster","current_host":"algo-1","current_instance_type":"ml.g5.2xlarge","hosts":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.g5.2xlarge"}],"network_interface_name":"eth0"}
SM_INPUT_DATA_CONFIG={"code":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"File"},"training":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"File"}}
SM_OUTPUT_DATA_DIR=/opt/ml/output/data
SM_CHANNELS=["code","training"]
SM_CURRENT_HOST=algo-1
SM_CURRENT_INSTANCE_TYPE=ml.g5.2xlarge
SM_CURRENT_INSTANCE_GROUP=homogeneousCluster
SM_CURRENT_INSTANCE_GROUP_HOSTS=["algo-1"]
SM_INSTANCE_GROUPS=["homogeneousCluster"]
SM_INSTANCE_GROUPS_DICT={"homogeneousCluster":{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.g5.2xlarge"}}
SM_DISTRIBUTION_INSTANCE_GROUPS=[]
SM_IS_HETERO=false
SM_MODULE_NAME=transfer_learning
SM_LOG_LEVEL=20
SM_FRAMEWORK_MODULE=sagemaker_pytorch_container.training:main
SM_INPUT_DIR=/opt/ml/input
SM_INPUT_CONFIG_DIR=/opt/ml/input/config
SM_OUTPUT_DIR=/opt/ml/output
SM_NUM_CPUS=8
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:07
07-04-2024

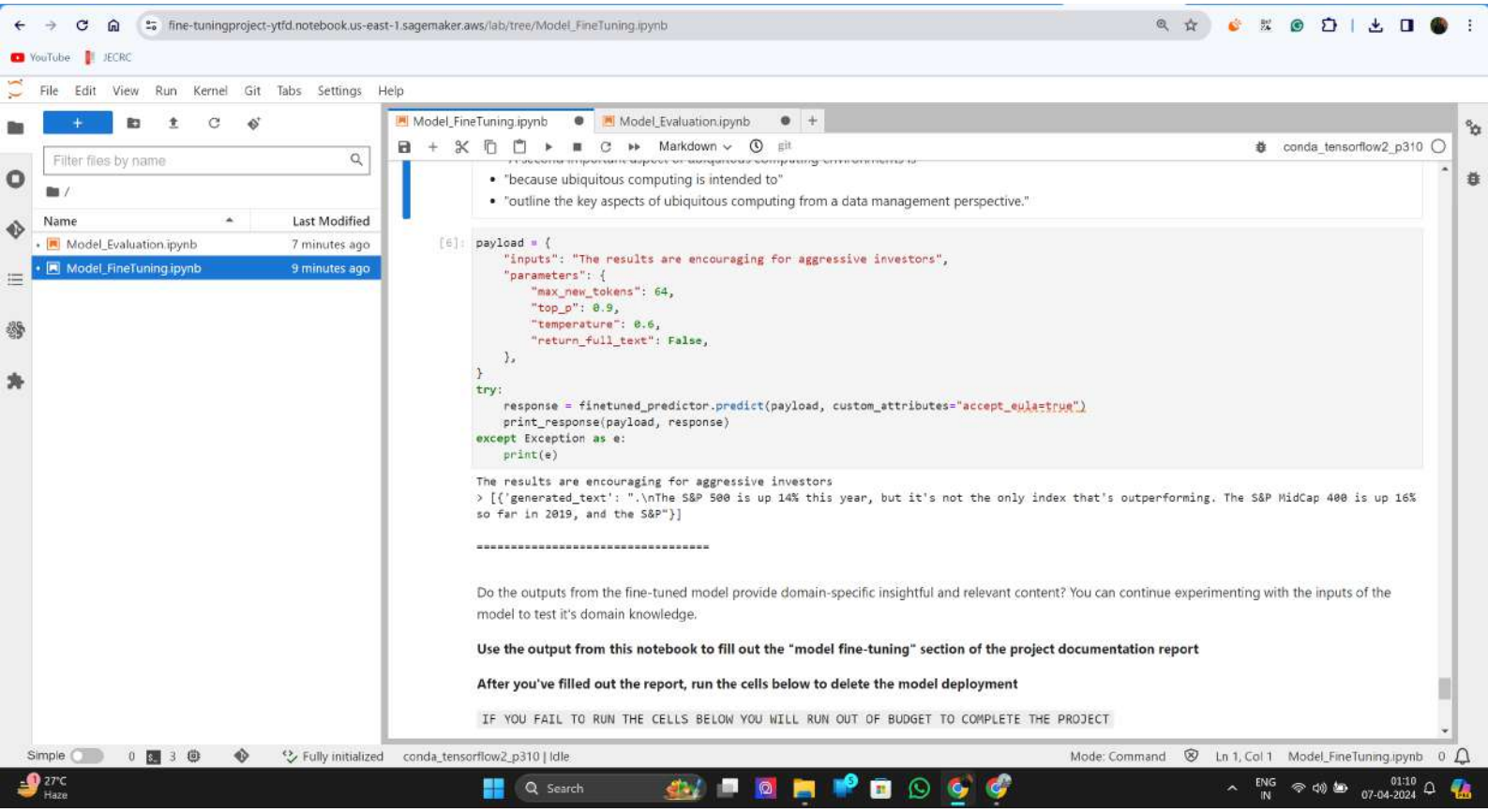

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 5 minutes ago
Model_FineTuning.ipynb 7 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
SM_INPUT_CONFIG_DIR=/opt/ml/input/config
SM_OUTPUT_DIR=/opt/ml/output
SM_NUM_CPUS=8
SM_NUM_GPUS=1
SM_NUM_NEURONS=0
SM_MODEL_DIR=/opt/ml/model
SM_MODULE_DIR=/opt/ml/input/data/code/sourcedir.tar.gz
SM_TRAINING_ENV={"additional_framework_parameters":{},"channel_input_dirs":{"code":"/opt/ml/input/data/code","training":"/opt/ml/input/data/training"},"current_host":"algo-1","current_instance_group":"homogeneousCluster","current_instance_group_hosts":["algo-1"],"current_instance_type":"ml.g5.2xlarge","distribution_hosts":[],"distribution_instance_groups":[],"framework_module":"sagemaker_pytorch_container.training:main","hosts":["algo-1"],"hyperparameters":{"add_input_output_demarcation_key":"True","chat_dataset":"False","enable_fsdp":"True","epoch":"5","instruction_tuned":"False","int8_quantization":"False","learning_rate":"0.0001","lora_alpha":"32","lora_dropout":"0.05","lora_r":"8","max_input_length":"-1","max_train_samples":"-1","max_val_samples":"-1","per_device_eval_batch_size":"1","per_device_train_batch_size":"4","preprocessing_num_workers":"None","seed":"10","train_data_split_seed":"0","validation_split_ratio":"0.2"},"input_config_dir":"/opt/ml/input/config","input_data_config":{"code":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"File"},"training":{"RecordWrapperType":"None","S3DistributionType":"FullyReplicated","TrainingInputMode":"File"}},"input_dir":"/opt/ml/input","instance_groups":["homogeneousCluster"],"instance_groups_dict":{"homogeneousCluster":{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.g5.2xlarge"}},"is_hetero":false,"is_master":true,"is_modelparallel_enabled":null,"is_smddpmprun_installed":true,"job_name":"meta-textgeneration-llama-2-7b-2024-02-14-13-05-49-101","log_level":20,"master_hostname":"algo-1","model_dir":"/opt/ml/model","module_dir":"/opt/ml/input/data/code/sourcedir.tar.gz","module_name":"transfer_learning","network_interface_name":"eth0","num_cpus":8,"num_gpus":1,"num_neurons":0,"output_data_dir":"/opt/ml/output/data","output_dir":"/opt/ml/output","output_intermediate_dir":"/opt/ml/output/intermediate","resource_config":{"current_group_name":"homogeneousCluster","current_host":"algo-1","current_instance_type":"ml.g5.2xlarge","hosts":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.g5.2xlarge"}],"network_interface_name":"eth0"},"user_entry_point":"transfer_learning.py"}
SM_USER_ARGS=["--add_input_output_demarcation_key","True","--chat_dataset","False","--enable_fsdp","True","--epoch","5","--instruction_tuned","False","--int8_quantization","False","--learning_rate","0.0001","--lora_alpha","32","--lora_dropout","0.05","--lora_r","8","--max_input_length","-1","--max_train_samples","-1","--max_val_samples","-1","--per_device_eval_batch_size","1","--per_device_train_batch_size","4","--preprocessing_num_workers","None","--seed","10","--train_data_split_seed","0","--validation_split_ratio","0.2"]
SM_OUTPUT_INTERMEDIATE_DIR=/opt/ml/output/intermediate
SM_CHANNEL_CODE=/opt/ml/input/data/code
SM_CHANNEL_TRAINING=/opt/ml/input/data/training
SM_HP_ADD_INPUT_OUTPUT_DEMARCATION_KEY=True
SM_HP_CHAT_DATASET=False
SM_HP_ENABLE_FSDP=True
SM_HP_EPOCH=5
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:08
07-04-2024

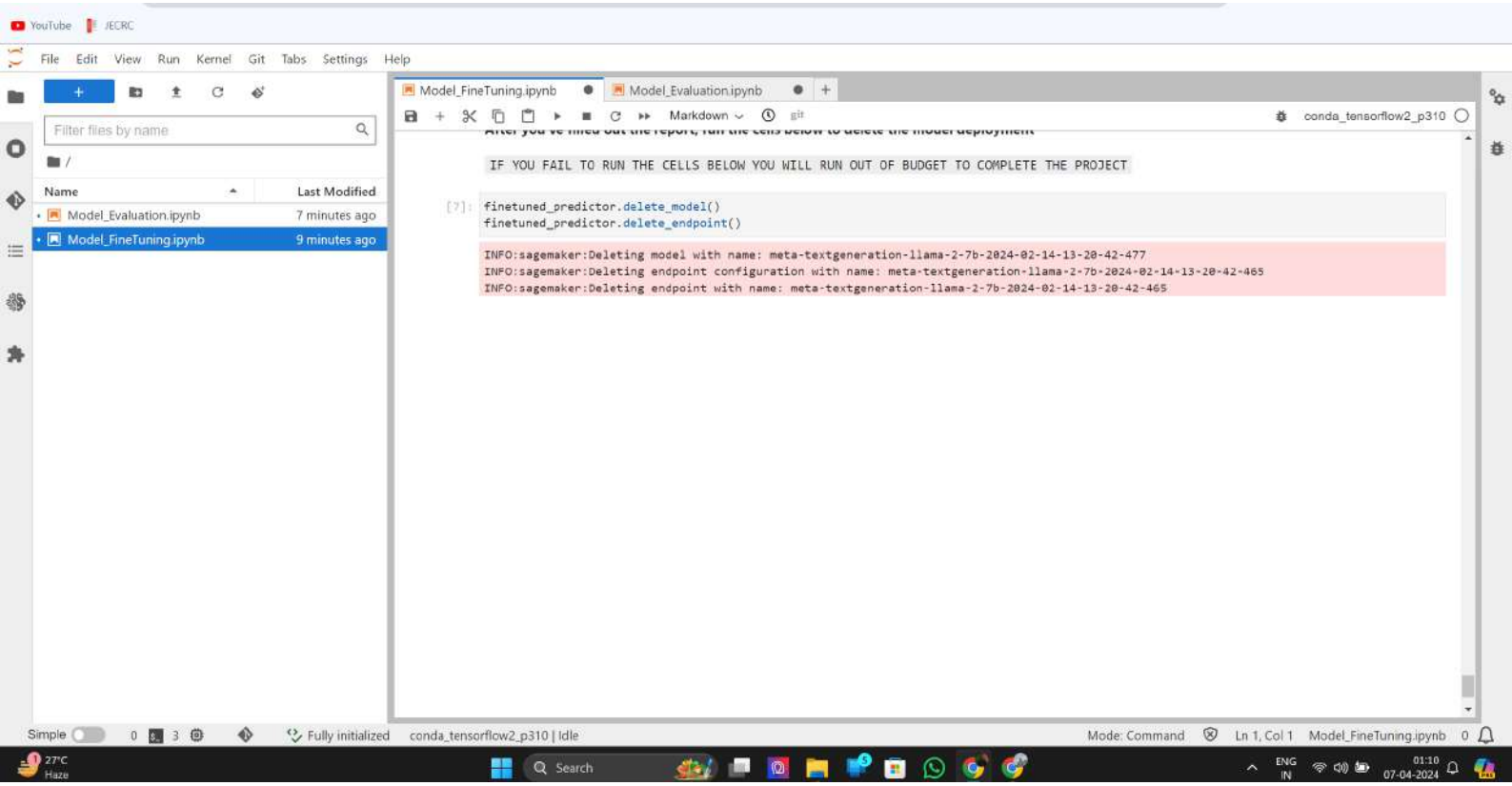

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
5 minutes ago
Model_FineTuning.ipynb
7 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
SM_HP_TRAIN_DATA_SPLIT_SEED=0
SM_HP_VALIDATION_SPLIT_RATIO=0.2
PYTHONPATH=/opt/ml/code:/opt/conda/bin:/opt/conda/lib/python310.zip:/opt/conda/lib/python3.10:/opt/conda/lib/python3.10/lib-dynload:/opt/conda/lib/python3.10/site-packages
Invoking script with the following command:
/opt/conda/bin/python3.10 transfer_learning.py --add_input_output_demarcation_key True --chat_dataset False --enable_fsdp True --epoch 5 --instruction_tuned False --int8_quantization False --learning_rate 0.0001 --lora_alpha 32 --lora_dropout 0.05 --lora_r 8 --max_input_length -1 --max_train_samples -1 --max_val_samples -1 --per_device_eval_batch_size 1 --per_device_train_batch_size 4 --preprocessing_num_workers None --seed 10 --train_data_split_seed 0 --validation_split_ratio 0.2
2024-02-14 13:13:12,173 sagemaker-training-toolkit INFO Exceptions not imported for SageMaker TF as Tensorflow is not installed.
===================================BUG REPORT===================================
Welcome to bitsandbytes. For bug reports, please run
python -m bitsandbytes
and submit this information together with your error trace to: https://github.com/TimDettmers/bitsandbytes/issues
==
bin /opt/conda/lib/python3.10/site-packages/bitsandbytes/libbitsandbytes_cuda118.so
/opt/conda/lib/python3.10/site-packages/bitsandbytes/cuda_setup/main.py:149: UserWarning: WARNING: The following directories listed in your path were found to be non-existent: {PosixPath('/usr/local/nvidia/lib64'), PosixPath('/usr/local/nvidia/lib')}
warn(msg)
CUDA SETUP: CUDA runtime path found: /opt/conda/lib/libcudart.so
CUDA SETUP: Highest compute capability among GPUs detected: 8.6
CUDA SETUP: Detected CUDA version 118
CUDA SETUP: Loading binary /opt/conda/lib/python3.10/site-packages/bitsandbytes/libbitsandbytes_cuda118.so...
INFO:root:Using pre-trained artifacts in SAGEMAKER_ADDITIONAL_S3_DATA_PATH=/opt/ml/additonals3data
INFO:root:Identify file serving.properties in the un-tar directory /opt/ml/additonals3data. Copying it over to /opt/ml/model for model deployment after training is finished.
INFO:root:Invoking the training command ['torchrun', '--nnodes', '1', '--nproc_per_node', '1', 'llama_finetuning.py', '--model_name', '/opt/ml/additonals3data', '--num_gpus', '1', '--pure_bf16', '--dist_checkpoint_root_folder', 'model_checkpoints', '--dist_checkpoint_folder', 'fine-tuned', '--batch_size_training', '4', '--micro_batch_size', '4', '--train_file', '/opt/ml/input/data/training', '--lr', '0.0001', '--do_train', '--output_dir', 'saved_peft_model', '--num_epochs', '5', '--use_peft', '--peft_method', 'lora', '--max_train_samples', '-1', '--max_val_samples', '-1', '--seed', '10', '--per_device_eval_batch_size', '1', '--max_input_length', '-1', '--preprocessing_num_workers', '--None', '--validation_split_ratio', '0.2', '--train_data_split_seed', '0', '--num_workers_dataloader', '0', '--weight_decay', '0.1', '--lora_r', '8', '--lora_alpha', '32', '--lora_dropout', '0.05', '--enable_fsdp', '--add_input_output_demarcation_key'].
===================================BUG REPORT===================================
Welcome to bitsandbytes. For bug reports, please run
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:08
07-04-2024

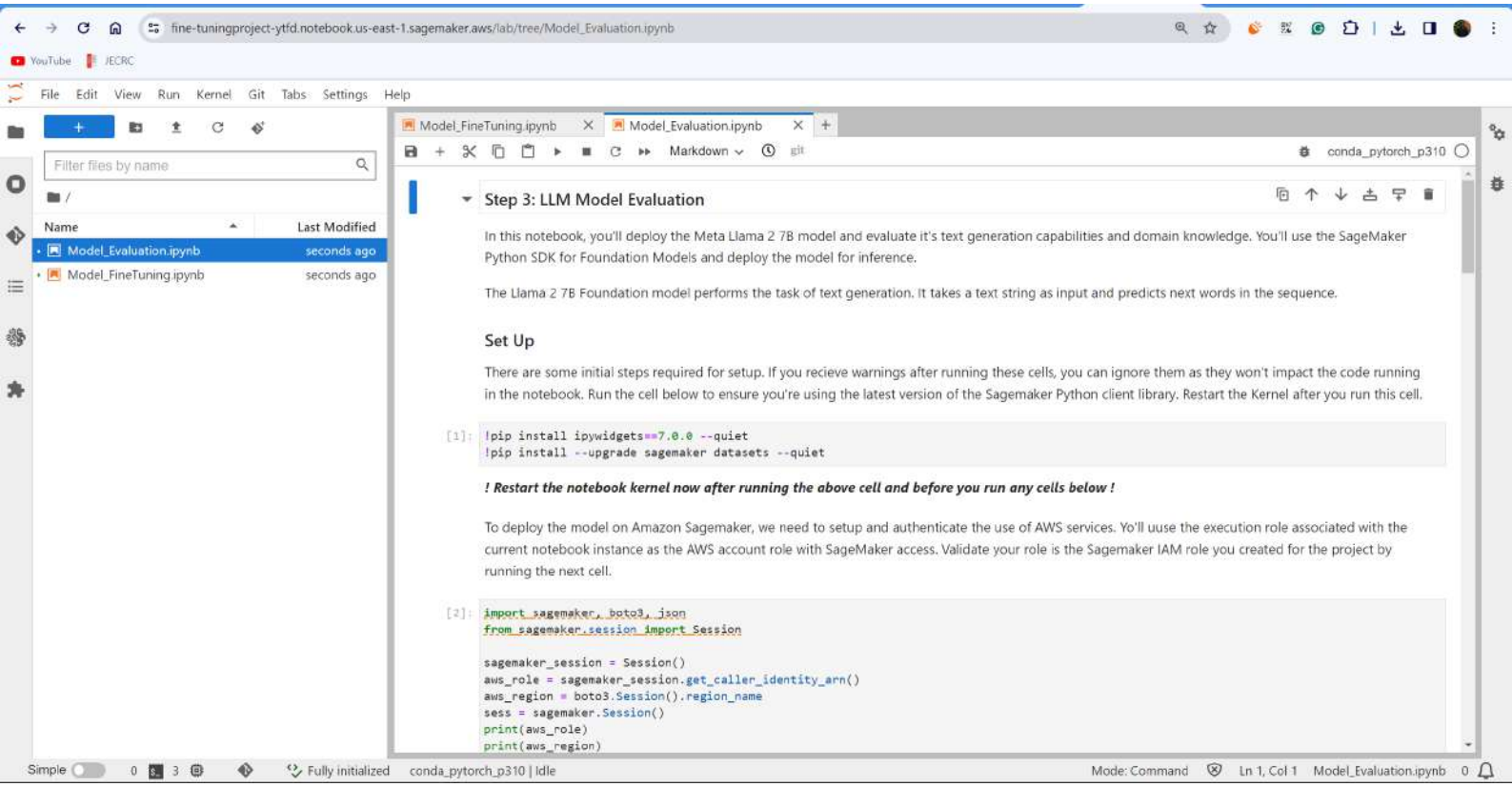

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 5 minutes ago
Model_FineTuning.ipynb 7 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
conda_tensorflow2_p310
evaluating Epoch: 0%|#033[32m #033[0m| 0/3 [00:00<?, ?it/s]
evaluating Epoch: 33%|#033[32m #033[0m| 1/3 [00:00<00:00, 2.49it/s]
evaluating Epoch: 67%|#033[32m #033[0m| 2/3 [00:00<00:00, 2.50it/s]
evaluating Epoch: 100%|#033[32m #033[0m| 3/3 [00:01<00:00, 2.51it/s]
evaluating Epoch: 100%|#033[32m #033[0m| 3/3 [00:01<00:00, 2.51it/s]
eval_ppl=tensor(11.0820, device='cuda:0') eval_epoch_loss=tensor(2.4053, device='cuda:0')
we are about to save the PEFT modules
PEFT modules are saved in saved_peft_model directory
best eval loss on epoch 4 is 2.405322313308716
Epoch 5: train_perplexity=15.7439, train_epoch_loss=2.7565, epcoh time 9.548661925999984s
INFO:root:Key: avg_train_prep, Value: 18.965452194213867
INFO:root:Key: avg_train_loss, Value: 2.93504595756530 76
INFO:root:Key: avg_eval_prep, Value: 12.528593063354492
INFO:root:Key: avg_eval_loss, Value: 2.5243804454803467
INFO:root:Key: avg_epoch_time, Value: 10.002697005400046
INFO:root:Key: avg_checkpoint_time, Value: 0.9348195582000016
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, 29.81s/it]
Loading checkpoint shards: 100% | 2/2 [00:35<00:00, 15.60s/it]
Loading checkpoint shards: 100% | 2/2 [00:35<00:00, 17.73s/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.
2024-02-14 13:19:27,931 sagemaker-training-toolkit INFO Waiting for the process to finish and give a return code.
2024-02-14 13:19:27,931 sagemaker-training-toolkit INFO Done waiting for a return code. Received 0 from exiting process.
2024-02-14 13:19:27,932 sagemaker-training-toolkit INFO Reporting training SUCCESS
2024-02-14 13:19:31 Uploading - Uploading generated training model
2024-02-14 13:20:22 Completed - Training job completed
Training seconds: 778
Billable seconds: 778
Deploy the fine-tuned model
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C Haze
Search
ENG IN
01:08 07-04-2024

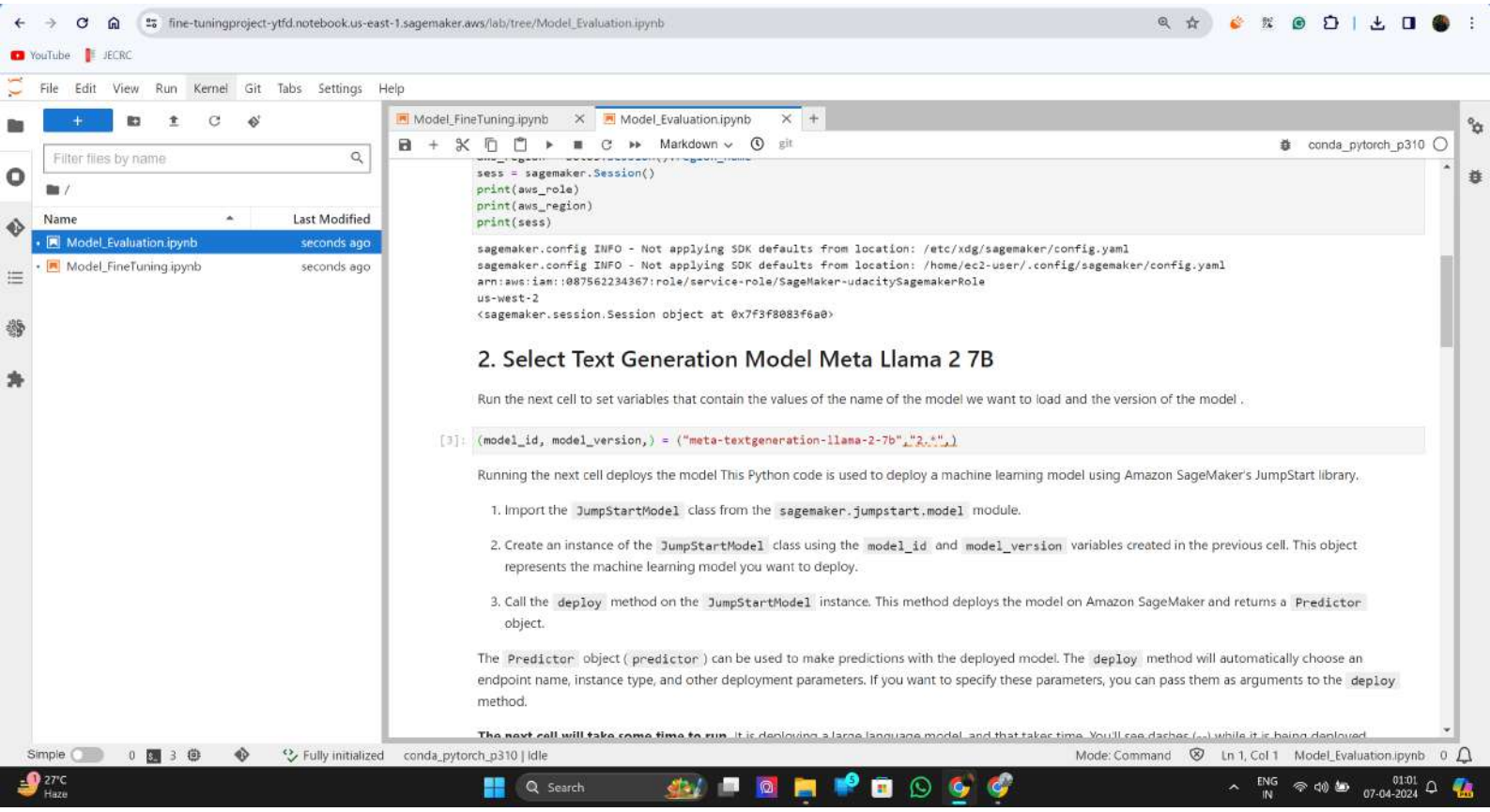

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 5 minutes ago
Model_FineTuning.ipynb 7 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
Billable seconds: 778
Deploy the fine-tuned model
Next, we deploy the domain fine-tuned model. We will compare the performance of the fine-tuned and pre-trained model.
[4]: 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-2024-02-14-13-20-42-477
INFO:sagemaker:Creating endpoint-config with name meta-textgeneration-llama-2-7b-2024-02-14-13-20-42-465
INFO:sagemaker:Creating endpoint with name meta-textgeneration-llama-2-7b-2024-02-14-13-20-42-465
-------!
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.
Create a function to print the response from the model
[5]: def print_response(payload, response):
print(payload["inputs"])
print(f"> {response}")
print("\n==========\n")
Now we can run the same prompts on the fine-tuned model to evaluate it's domain knowledge.
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:08
07-04-2024

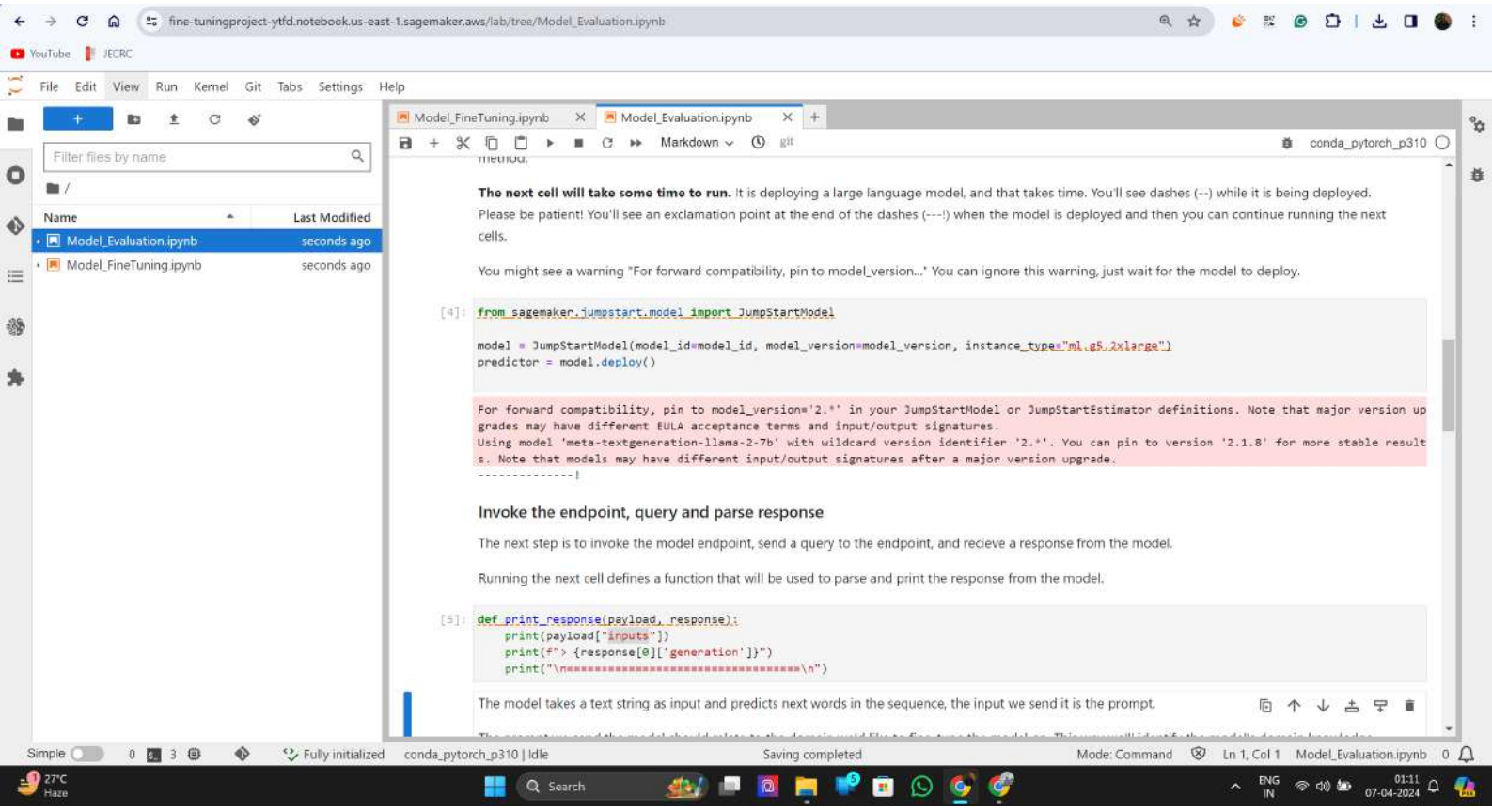

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 7 minutes ago
Model_FineTuning.ipynb 9 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
conda_tensorflow2_p310
Now we can run the same prompts on the fine-tuned model to evaluate it's domain knowledge.
Replace "inputs" in the next cell with the input to send the model based on the domain you've chosen.
For financial domain:
"inputs": "Replace with sentence below from text"
"The investment tests performed indicate"
"the relative volume for the long out of the money options, indicates"
"The results for the short in the money options"
"The results are encouraging for aggressive investors"
For medical domain:
"inputs": "Replace with sentence below from text"
"Myeloid neoplasms and acute leukemias derive from"
"Genomic characterization is essential for"
"Certain germline disorders may be associated with"
"In contrast to targeted approaches, genome-wide sequencing"
For IT domain:
"inputs": "Replace with sentence below from text"
"Traditional approaches to data management such as"
"A second important aspect of ubiquitous computing environments is"
"because ubiquitous computing is intended to"
"outline the key aspects of ubiquitous computing from a data management perspective."
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:10
07-04-2024

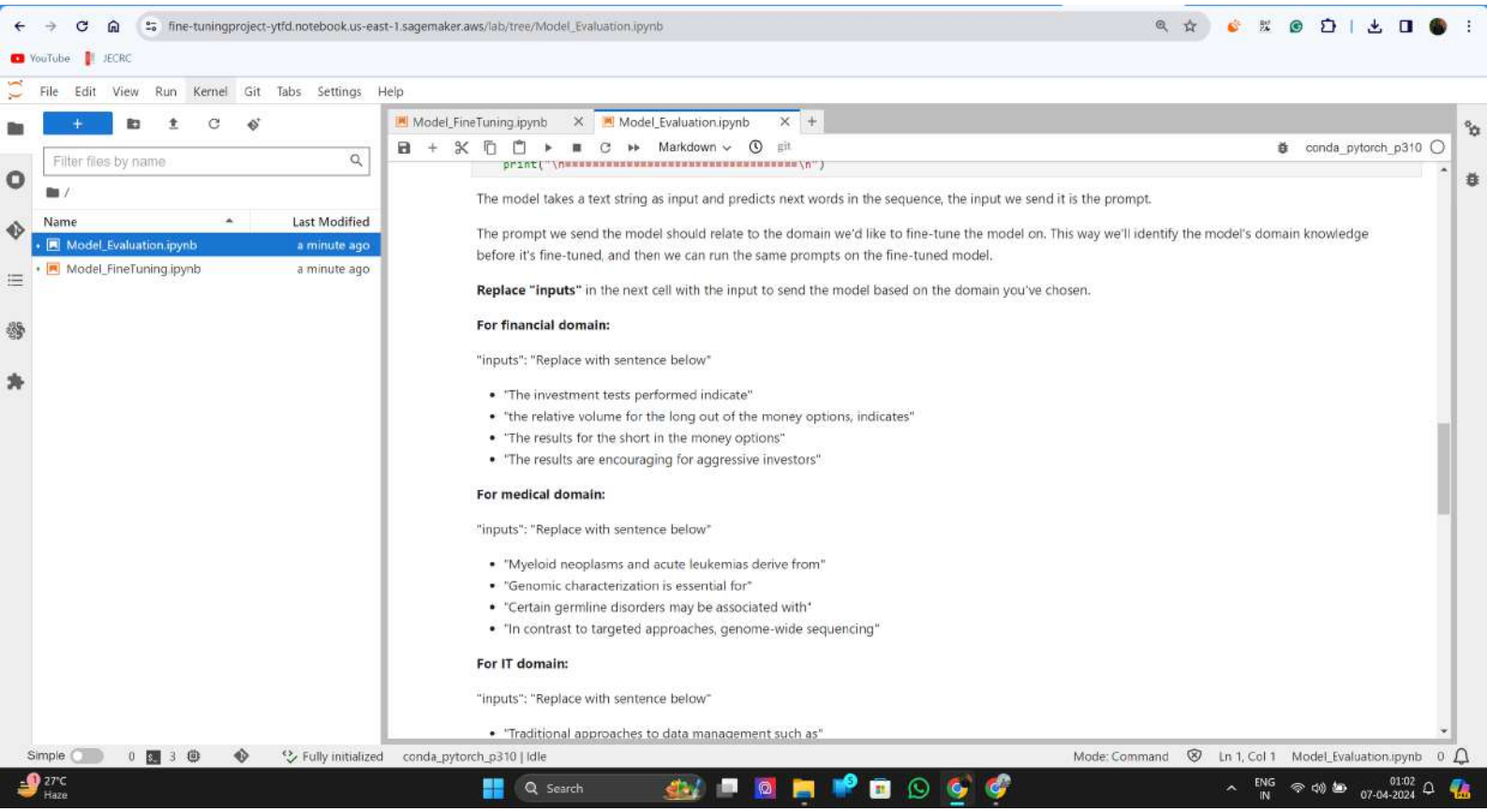

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_FineTuning.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb 7 minutes ago
Model_FineTuning.ipynb 9 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_tensorflow2_p310
"because ubiquitous computing is intended to"
"outline the key aspects of ubiquitous computing from a data management perspective."
[6]:
payload = {
"inputs": "The results are encouraging for aggressive investors",
"parameters": {
"max_new_tokens": 64,
"top_p": 0.9,
"temperature": 0.6,
"return_full_text": False,
},
}
try:
response = finetuned_predictor.predict(payload, custom_attributes="accept_eula=true")
print_response(payload, response)
except Exception as e:
print(e)
The results are encouraging for aggressive investors
> [{'generated_text': ".\nThe S&P 500 is up 14% this year, but it's not the only index that's outperforming. The S&P MidCap 400 is up 16% so far in 2019, and the S&P"}]
==========
Do the outputs from the fine-tuned model provide domain-specific insightful and relevant content? You can continue experimenting with the inputs of the model to test it's domain knowledge.
Use the output from this notebook to fill out the "model fine-tuning" section of the project documentation report
After you've filled out the report, run the cells below to delete the model deployment
IF YOU FAIL TO RUN THE CELLS BELOW YOU WILL RUN OUT OF BUDGET TO COMPLETE THE PROJECT
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:10
07-04-2024

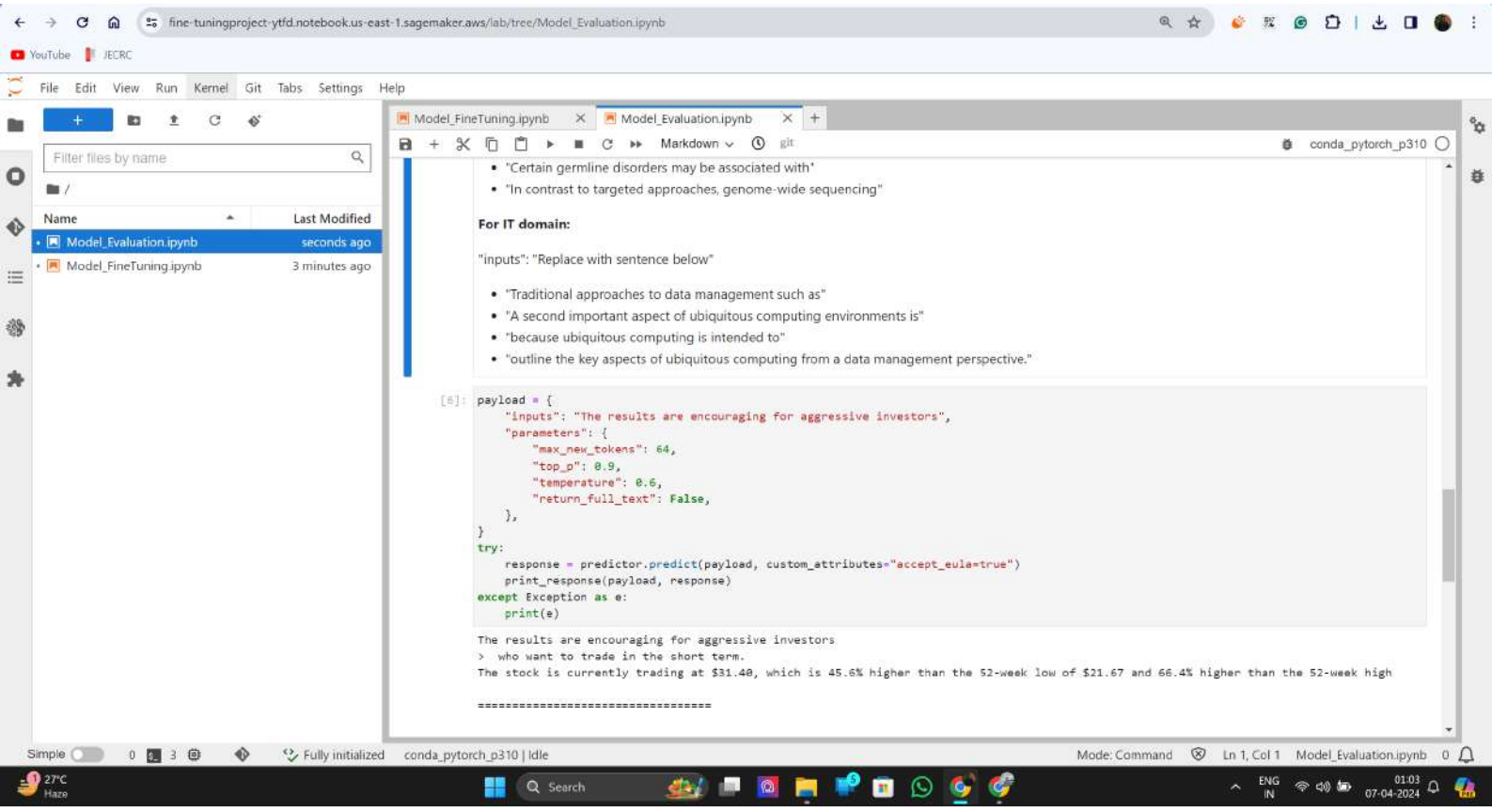

YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
7 minutes ago
Model_FineTuning.ipynb
9 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
conda_tensorflow2_p310
IF YOU FAIL TO RUN THE CELLS BELOW YOU WILL RUN OUT OF BUDGET TO COMPLETE THE PROJECT
[7]:
finetuned_predictor.delete_model()
finetuned_predictor.delete_endpoint()
INFO:sagemaker:Deleting model with name: meta-textgeneration-llama-2-7b-2024-02-14-13-20-42-477
INFO:sagemaker:Deleting endpoint configuration with name: meta-textgeneration-llama-2-7b-2024-02-14-13-20-42-465
INFO:sagemaker:Deleting endpoint with name: meta-textgeneration-llama-2-7b-2024-02-14-13-20-42-465
Simple
Fully initialized
conda_tensorflow2_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_FineTuning.ipynb
27°C
Haze
Search
ENG
IN
01:10
07-04-2024

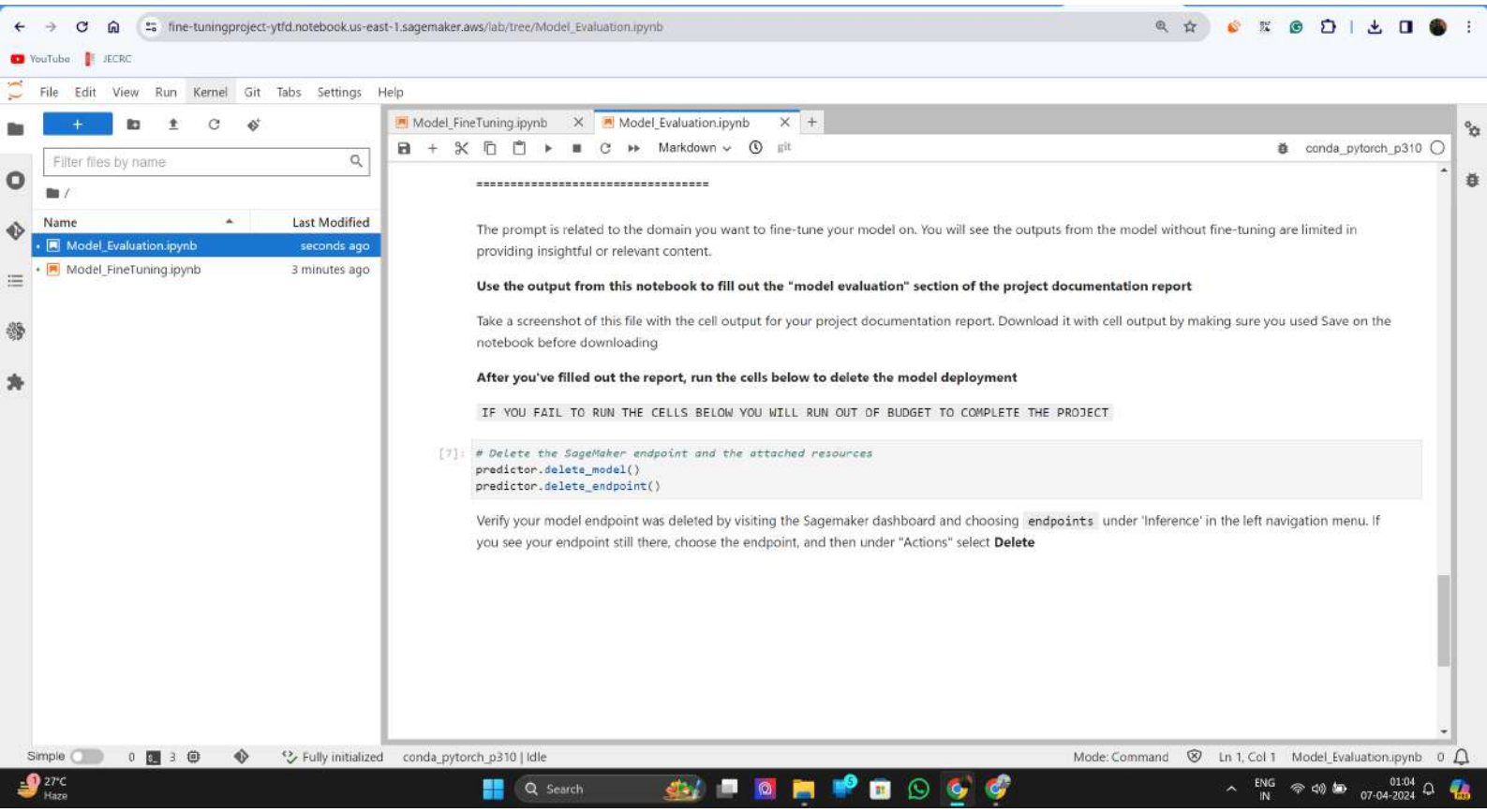

fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_Evaluation.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
seconds ago
Model_FineTuning.ipynb
seconds ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_pytorch_p310
Step 3: LLM Model Evaluation
In this notebook, you'll deploy the Meta Llama 2 7B model and evaluate it's text generation capabilities and domain knowledge. You'll use the SageMaker Python SDK for Foundation Models and deploy the model for inference.
The Llama 2 7B Foundation model performs the task of text generation. It takes a text string as input and predicts next words in the sequence.
Set Up
There are some initial steps required for setup. If you recieve warnings after running these cells, you can ignore them as they won't impact the code running in the notebook. Run the cell below to ensure you're using the latest version of the Sagemaker Python client library. Restart the Kernel after you run this cell.
[1]:
!pip install ipywidgets==7.0.0 --quiet
!pip install --upgrade sagemaker datasets --quiet
! Restart the notebook kernel now after running the above cell and before you run any cells below !
To deploy the model on Amazon Sagemaker, we need to setup and authenticate the use of AWS services. Yo'll uuse the execution role associated with the current notebook instance as the AWS account role with SageMaker access. Validate your role is the Sagemaker IAM role you created for the project by running the next cell.
[2]:
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)
Simple
Fully initialized
conda_pytorch_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_Evaluation.ipynb


fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_Evaluation.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
seconds ago
Model_FineTuning.ipynb
seconds ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_pytorch_p310
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::087562234367:role/service-role/SageMaker-udacitySagemakerRole
us-west-2
<sagemaker.session.Session object at 0x7f3f8083f6a0>
2. Select Text Generation Model Meta Llama 2 7B
Run the next cell to set variables that contain the values of the name of the model we want to load and the version of the model .
[3]:
(model_id, model_version,) = ("meta-textgeneration-llama-2-7b","2.*",)
Running the next cell deploys the model This Python code is used to deploy a machine learning model using Amazon SageMaker's JumpStart library.
1. Import the JumpStartModel class from the sagemaker.jumpstart.model module.
2. Create an instance of the JumpStartModel class using the model_id and model_version variables created in the previous cell. This object represents the machine learning model you want to deploy.
3. Call the deploy method on the JumpStartModel instance. This method deploys the model on Amazon SageMaker and returns a Predictor object.
The Predictor object (predictor) can be used to make predictions with the deployed model. The deploy method will automatically choose an endpoint name, instance type, and other deployment parameters. If you want to specify these parameters, you can pass them as arguments to the deploy method.
The next cell will take some time to run, it is deploying a large language model, and that takes time. You'll see dashes (--) while it is being deployed
Simple
Fully initialized
conda_pytorch_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_Evaluation.ipynb
27°C
Haze
Search
ENG
IN
01:01
07-04-2024


fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_Evaluation.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
seconds ago
Model_FineTuning.ipynb
seconds ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_pytorch_p310
The next cell will take some time to run. It is deploying a large language model, and that takes time. You'll see dashes (--) while it is being deployed. Please be patient! You'll see an exclamation point at the end of the dashes (---!) when the model is deployed and then you can continue running the next cells.
You might see a warning "For forward compatibility, pin to model_version...' You can ignore this warning, just wait for the model to deploy.
[4]:
from sagemaker.jumpstart.model import JumpStartModel
model = JumpStartModel(model_id=model_id, model_version=model_version, instance_type="ml.g5.2xlarge")
predictor = model.deploy()
For forward compatibility, pin to model_version='2.*' in your JumpStartModel or JumpStartEstimator definitions. Note that major version upgrades may have different EULA acceptance terms and input/output signatures.
Using model 'meta-textgeneration-llama-2-7b' with wildcard version identifier '2.*'. You can pin to version '2.1.8' for more stable results. Note that models may have different input/output signatures after a major version upgrade.
----------!
Invoke the endpoint, query and parse response
The next step is to invoke the model endpoint, send a query to the endpoint, and recieve a response from the model.
Running the next cell defines a function that will be used to parse and print the response from the model.
[5]:
def print_response(payload, response):
print(payload["inputs"])
print(f"> {response[0]['generation']}")
print("\n==================================\n")
The model takes a text string as input and predicts next words in the sequence, the input we send it is the prompt.
Simple
Fully initialized
conda_pytorch_p310 | Idle
Saving completed
Mode: Command
Ln 1, Col 1
Model_Evaluation.ipynb
27°C
Haze
Search
ENG
IN
01:11
07-04-2024


fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_Evaluation.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
a minute ago
Model_FineTuning.ipynb
a minute ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_pytorch_p310
print("\n==========\n")
The model takes a text string as input and predicts next words in the sequence, the input we send it is the prompt.
The prompt we send the model should relate to the domain we'd like to fine-tune the model on. This way we'll identify the model's domain knowledge before it's fine-tuned, and then we can run the same prompts on the fine-tuned model.
Replace "inputs" in the next cell with the input to send the model based on the domain you've chosen.
For financial domain:
"inputs": "Replace with sentence below"
"The investment tests performed indicate"
"the relative volume for the long out of the money options, indicates"
"The results for the short in the money options"
"The results are encouraging for aggressive investors"
For medical domain:
"inputs": "Replace with sentence below"
"Myeloid neoplasms and acute leukemias derive from"
"Genomic characterization is essential for"
"Certain germline disorders may be associated with"
"In contrast to targeted approaches, genome-wide sequencing"
For IT domain:
"inputs": "Replace with sentence below"
"Traditional approaches to data management such as"
Simple
Fully initialized
conda_pytorch_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_Evaluation.ipynb
27°C
Haze
Search
ENG
IN
01:02
07-04-2024


fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_Evaluation.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
seconds ago
Model_FineTuning.ipynb
3 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_pytorch_p310
"Certain germline disorders may be associated with"
"In contrast to targeted approaches, genome-wide sequencing"
For IT domain:
"inputs": "Replace with sentence below"
"Traditional approaches to data management such as"
"A second important aspect of ubiquitous computing environments is"
"because ubiquitous computing is intended to"
"outline the key aspects of ubiquitous computing from a data management perspective."
[6]:
payload = {
"inputs": "The results are encouraging for aggressive investors",
"parameters": {
"max_new_tokens": 64,
"top_p": 0.9,
"temperature": 0.6,
"return_full_text": False,
},
}
try:
response = predictor.predict(payload, custom_attributes="accept_eula=true")
print_response(payload, response)
except Exception as e:
print(e)
The results are encouraging for aggressive investors
> who want to trade in the short term.
The stock is currently trading at $31.40, which is 45.6% higher than the 52-week low of $21.67 and 66.4% higher than the 52-week high
==========
Simple
Fully initialized
conda_pytorch_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_Evaluation.ipynb
27°C
Haze
Search
ENG
IN
01:03
07-04-2024


fine-tuningproject-ytfd.notebook.us-east-1.sagemaker.aws/lab/tree/Model_Evaluation.ipynb
YouTube
JECRC
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
/
Name
Last Modified
Model_Evaluation.ipynb
seconds ago
Model_FineTuning.ipynb
3 minutes ago
Model_FineTuning.ipynb
Model_Evaluation.ipynb
Markdown
git
conda_pytorch_p310
==========
The prompt is related to the domain you want to fine-tune your model on. You will see the outputs from the model without fine-tuning are limited in providing insightful or relevant content.
Use the output from this notebook to fill out the "model evaluation" section of the project documentation report
Take a screenshot of this file with the cell output for your project documentation report. Download it with cell output by making sure you used Save on the notebook before downloading
After you've filled out the report, run the cells below to delete the model deployment
IF YOU FAIL TO RUN THE CELLS BELOW YOU WILL RUN OUT OF BUDGET TO COMPLETE THE PROJECT
[7]:
# Delete the SageMaker endpoint and the attached resources
predictor.delete_model()
predictor.delete_endpoint()
Verify your model endpoint was deleted by visiting the Sagemaker dashboard and choosing endpoints under 'Inference' in the left navigation menu. If you see your endpoint still there, choose the endpoint, and then under "Actions" select Delete
Simple
Fully initialized
conda_pytorch_p310 | Idle
Mode: Command
Ln 1, Col 1
Model_Evaluation.ipynb
27°C
Haze
Search
ENG
IN
01:04
07-04-2024