hugginfacedeployment

August 23, 2024

[2]:
```
!pip install --upgrade boto3 sagemaker
```

```
Requirement already satisfied: boto3 in /opt/conda/lib/python3.10/site-packages
(1.35.4)
Requirement already satisfied: sagemaker in /opt/conda/lib/python3.10/site-
packages (2.229.0)
Requirement already satisfied: botocore<1.36.0,>=1.35.4 in
/opt/conda/lib/python3.10/site-packages (from boto3) (1.35.4)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in
/opt/conda/lib/python3.10/site-packages (from boto3) (1.0.1)
Requirement already satisfied: s3transfer<0.11.0,>=0.10.0 in
/opt/conda/lib/python3.10/site-packages (from boto3) (0.10.2)
Requirement already satisfied: attrs<24,>=23.1.0 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (23.2.0)
Requirement already satisfied: cloudpickle==2.2.1 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (2.2.1)
Requirement already satisfied: docker in /opt/conda/lib/python3.10/site-packages
(from sagemaker) (7.1.0)
Requirement already satisfied: google-pasta in /opt/conda/lib/python3.10/site-
packages (from sagemaker) (0.2.0)
Requirement already satisfied: importlib-metadata<7.0,>=1.4.0 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (6.10.0)
Requirement already satisfied: jsonschema in /opt/conda/lib/python3.10/site-
packages (from sagemaker) (4.17.3)
Requirement already satisfied: numpy<2.0,>=1.9.0 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (23.2)
Requirement already satisfied: pandas in /opt/conda/lib/python3.10/site-packages
(from sagemaker) (2.1.4)
Requirement already satisfied: pathos in /opt/conda/lib/python3.10/site-packages
(from sagemaker) (0.3.2)
Requirement already satisfied: platformdirs in /opt/conda/lib/python3.10/site-
packages (from sagemaker) (4.2.2)
Requirement already satisfied: protobuf<5.0,>=3.12 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (4.24.4)
Requirement already satisfied: psutil in /opt/conda/lib/python3.10/site-packages
(from sagemaker) (5.9.8)
```

```
Requirement already satisfied: PyYAML~=6.0 in /opt/conda/lib/python3.10/site-
packages (from sagemaker) (6.0.1)
Requirement already satisfied: requests in /opt/conda/lib/python3.10/site-
packages (from sagemaker) (2.32.3)
Requirement already satisfied: schema in /opt/conda/lib/python3.10/site-packages
(from sagemaker) (0.7.7)
Requirement already satisfied: smdebug-rulesconfig==1.0.1 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (1.0.1)
Requirement already satisfied: tblib<4,>=1.7.0 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (2.0.0)
Requirement already satisfied: tqdm in /opt/conda/lib/python3.10/site-packages
(from sagemaker) (4.66.4)
Requirement already satisfied: urllib3<3.0.0,>=1.26.8 in
/opt/conda/lib/python3.10/site-packages (from sagemaker) (1.26.19)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in
/opt/conda/lib/python3.10/site-packages (from botocore<1.36.0,>=1.35.4->boto3)
(2.9.0)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.10/site-
packages (from importlib-metadata<7.0,>=1.4.0->sagemaker) (3.19.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/conda/lib/python3.10/site-packages (from requests->sagemaker) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-
packages (from requests->sagemaker) (3.7)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.10/site-packages (from requests->sagemaker) (2024.7.4)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packages
(from google-pasta->sagemaker) (1.16.0)
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in
/opt/conda/lib/python3.10/site-packages (from jsonschema->sagemaker) (0.20.0)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/site-
packages (from pandas->sagemaker) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in /opt/conda/lib/python3.10/site-
packages (from pandas->sagemaker) (2024.1)
Requirement already satisfied: ppft>=1.7.6.8 in /opt/conda/lib/python3.10/site-
packages (from pathos->sagemaker) (1.7.6.8)
Requirement already satisfied: dill>=0.3.8 in /opt/conda/lib/python3.10/site-
packages (from pathos->sagemaker) (0.3.8)
Requirement already satisfied: pox>=0.3.4 in /opt/conda/lib/python3.10/site-
packages (from pathos->sagemaker) (0.3.4)
Requirement already satisfied: multiprocess>=0.70.16 in
/opt/conda/lib/python3.10/site-packages (from pathos->sagemaker) (0.70.16)
```

[3]:

```python
import sagemaker
import boto3
sess = sagemaker.Session()
# sagemaker session bucket -> used for uploading data, models and logs
# sagemaker will automatically create this bucket if it not exists
```

```
sagemaker_session_bucket = None
if sagemaker_session_bucket is None and sess is not None:
    # set to default bucket if a bucket name is not given
    sagemaker_session_bucket = sess.default_bucket()

try:
    role = sagemaker.get_execution_role()
except ValueError:
    iam = boto3.client('iam')
    role = iam.get_role(RoleName='sagemaker_execution_role')['Role']['Arn']

sess = sagemaker.Session(default_bucket=sagemaker_session_bucket)

print(f"sagemaker role arn: {role}")
print(f"sagemaker session region: {sess.boto_region_name}")
```

```
sagemaker.config INFO - Not applying SDK defaults from location:
/etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location:
/home/sagemaker-user/.config/sagemaker/config.yaml
sagemaker role arn: arn:aws:iam::967474752012:role/service-role/AmazonSageMaker-
ExecutionRole-20240823T213971
sagemaker session region: ap-south-1
```

[4]:

```
from sagemaker.huggingface.model import HuggingFaceModel

# Hub model configuration <https://huggingface.co/models>
hub = {
  'HF_MODEL_ID':'distilbert-base-uncased-distilled-squad', # model_id from hf.
 ↪co/models
  'HF_TASK':'question-answering'                           # NLP task you want␣
 ↪to use for predictions
}

# create Hugging Face Model Class
huggingface_model = HuggingFaceModel(
   env=hub,                                                # configuration for␣
 ↪loading model from Hub
   role=role,                                              # IAM role with␣
 ↪permissions to create an endpoint
   transformers_version="4.26",                             # Transformers␣
 ↪version used
   pytorch_version="1.13",                                  # PyTorch version␣
 ↪used
   py_version='py39',                                      # Python version used
)
```

```
# deploy model to SageMaker Inference
predictor = huggingface_model.deploy(
   initial_instance_count=1,
   instance_type="ml.m5.xlarge"
)

# example request: you always need to define "inputs"
data = {
"inputs": {
        "question": "What is used for inference?",
        "context": "My Name is Philipp and I live in Nuremberg. This model is␣
 ↪used with sagemaker for inference."
        }
}

# request
predictor.predict(data)
```

```
------!
```

[4]:
```
{'score': 0.9987204670906067, 'start': 68, 'end': 77, 'answer': 'sagemaker'}
```

[5]:
```
predictor.predict(data)
```

[5]:
```
{'score': 0.9987204670906067, 'start': 68, 'end': 77, 'answer': 'sagemaker'}
```

[12]:
```
data = {
"inputs": {
        "question": "Who is Sai Kumar Seela",
        "context": "My Name is Sai Kumar Seela. I am a Computer Science␣
 ↪Graduate."
        }
}
```

[13]:
```
predictor.predict(data)
```

[13]:
```
{'score': 0.2198973000049591,
 'start': 28,
 'end': 60,
 'answer': 'I am a Computer Science Graduate'}
```

[ ]:
```

```