Lab 2

Introduction

Language models have recently exploded in both size and popularity. In 2018, BERT-large entered the scene and, with its 340M parameters and novel transformer architecture, set the standard on NLP task accuracy. Within just a few years, state-of-the-art NLP model size has grown by more than 500x with models such as OpenAI’s 175 billion parameter GPT-3 and similarly sized open source Bloom 176B raising the bar on NLP accuracy. This increase in the number of parameters is driven by the simple and empirically-demonstrated positive relationship between model size and accuracy: more is better. With easy access from models zoos such as HuggingFace and improved accuracy in NLP tasks such as classification and text generation, practitioners are increasingly reaching for these large models. These models can be used in pretrained form as so called "Foundation Models".

However, since they are trained on large datasets of generic data they are often not suited for use cases requiring domain-specific knowledge. Especially the models of large parameter size are usually able to generalize well and perform surprisingly good in various zero-shot/few-shot scenarios. Nevertheless, zero-shot/few-shot performance in complex tasks like question-answering or handling human-like conversation is decreasing rapidly the more specifik the tasks become.

In these cases, fine-tuning a model on a smaller, domain-specific use case can help increase the performance to a satisfying level. However, training/finetuning these models can be a challenge because of their size.

In this Lab, we'll explore how to finetune a large language model on Amazon SageMaker using Sagemaker Training, one of many ready-to-use AWS Deep Learning Containers (DLCs) and the built-in HuggingFace integration of the Sagemaker SDK.

Background and Details

Since training such models requires even more resources than hosting them, for this lab we'll be working with a rather small Large Language Model (LLM) to learn the basic concepts of finetuning LLMs . However, the proposed approach works similarily at scale for larger models.

'distilGPT', a transformer-based large language model with around 82M parameters is the distilled version of GPT2 (predecessor of GPT-3/4), which was pre-trained on the WebText dataset. Since it is a decoder-only model it was trained using a causal language modeling (CLM) loss. We will use the exact approach for finetuning the data on the 'tiny_shakespeare' dataset to adjust the model output in terms of writing-style and content of the generated text.

The 'tiny_shakespeare' is a dataset consisting of 40000 lines of Shakespeare from a variety of Shakespeare's plays avaliable in a train/test/validation split. It can be retrieved conveniently from the HuggingFace datasets hub.

Finally, we will deploy both the original and the finetuned model to experience the impact of the performed training.

Instructions

Prerequisites

To run this workshop...

You need a computer with a web browser, preferably with the latest version of Chrome / FireFox. Sequentially read and follow the instructions described in AWS Hosted Event and Work Environment Set Up

Recommended background

It will be easier for you to run this workshop if you have:

- Experience with Deep learning models
- Familiarity with Python or other similar programming languages
- Experience with Jupyter notebooks
- Begineers level knowledge and experience with SageMaker Hosting/Inference.

Target audience

Data Scientists, ML Engineering, ML Infrastructure, MLOps Engineers, Technical Leaders. Intended for customers working with large Generative AI models including Language, Computer vision and Multi-modal use-cases. Customers using EKS/EC2/ECS/On-prem for hosting or experience with SageMaker.

Level of expertise - 400

Time to complete

Approximately 1 hour.

Import of required dependencies

For this lab, we will use the following libraries:

- boto3, the AWS SDK for python
- SageMaker SDK for interacting with Amazon SageMaker. We especially want to highlight the classes 'HuggingFaceModel' and 'HuggingFace', utilizing the built-in HuggingFace integration into SageMaker SDK. These classes are used to encapsulate functionality around the model and the deployed endpoint we will use. They inherit from the generic 'Model' and 'Estimator' classes of the native SageMaker SDK, however implementing some additional functionality specific to HuggingFace and the HuggingFace model hub.

- os, a python library implementing miscellaneous operating system interfaces

In [2]:

```
import boto3
import sagemaker
import sagemaker.session
import os

from sagemaker.huggingface import HuggingFace, HuggingFaceModel
```

Setup of notebook environment

Before we begin with the actual work for finetuning and deploying the model to Amazon SageMaker, we need to setup the notebook environment respectively. This includes:

- retrieval of the execution role our SageMaker Studio domain is associated with for later usage
- retrieval of our account_id for later usage
- retrieval of the chosen region for later usage

In [3]:

```
# Retrieve SM execution role
role = sagemaker.get_execution_role()
```

In [4]:

```
# Create a new STS client
sts_client = boto3.client('sts')

# Call the GetCallerIdentity operation to retrieve the account ID
response = sts_client.get_caller_identity()
account_id = response['Account']
account_id
```

```
/opt/conda/lib/python3.7/site-packages/boto3/compat.py:82: PythonDeprecationWarnin
g: Boto3 will no longer support Python 3.7 starting December 13, 2023. To continue
receiving service updates, bug fixes, and security updates please upgrade to Pytho
n 3.8 or later. More information can be found here: https://aws.amazon.com/blogs/d
eveloper/python-support-policy-updates-for-aws-sdks-and-tools/
  warnings.warn(warning, PythonDeprecationWarning)
```

Out[4]:

```
'882819251225'
```

In [5]:

```
# Retrieve region
region = boto3.Session().region_name
region
```

Out[5]:

```
'us-east-1'
```

Setup of S3 bucket for storage of training artifacts

When training a model with AWS SageMaker Training several artifacts can be written to an S3 bucket. This includes the trained model in form of a ‘model.tar.gz’ but also other artifacts like log files and the source code base. For this purpose, (if not already present) we create a dedicated S3 bucket.

In [6]:

```
# specifying bucket name for model artifact storage
model_bucket_name = f'immersion-day-bucket-{account_id}'
model_bucket_name
```

Out[6]:

```
'immersion-day-bucket-882819251225'
```

In [7]:

```
# Create S3 bucket
s3_client = boto3.client('s3', region_name=region)
location = {'LocationConstraint': region}

bucket_name = model_bucket_name

# Check if bucket already exists
bucket_exists = True
try:
    s3_client.head_bucket(Bucket=bucket_name)
except:
    bucket_exists = False

# Create bucket if it does not exist
if not bucket_exists:
    if region == 'us-east-1':
        s3_client.create_bucket(Bucket=bucket_name)
    else:
        s3_client.create_bucket(Bucket=bucket_name,
        CreateBucketConfiguration=location)
    print(f"Bucket '{bucket_name}' created successfully")
```

```
Bucket 'immersion-day-bucket-882819251225' created successfully
```

Diving deep into the training code

The code artifacts required for finetuning are residing in the finetuning directory. This directory is composed as follows:

`finetuning/`

- `finetuning.py`
- `requirements.txt`

The "finetuning" directory contains your training script (finetuning.py) and your requirements.txt file (for installation of additional dependencies not preinstalled in the container image upon start of the training container). We will now take a closer look into the training code:

Import of required dependencies

On top of several commodity Python libraries, for this training script we will use the following DL specific libraries:

- torch: PyTorch is a Python package that provides two high-level features: 1/Tensor computation (like NumPy) with strong GPU acceleration and 2/Deep neural networks built on a tape-based autograd system

- transformers: HuggingFace transformers provides APIs and tools to easily download and train state-of-the-art pretrained models. Transformers support framework interoperability between PyTorch, TensorFlow, and JAX.
- evaluate: HuggingFace evaluate is a library for easily evaluating machine learning models and datasets.
- datasets: HuggingFace datasets is a library for easily accessing and sharing datasets for Audio, Computer Vision, and Natural Language Processing (NLP) tasks in the context of the HuggingFace dataset hub.

Script invocation and hyperparameter parsing

After the ephemeral training cluster has been provisioned and the respective Docker image has been pulled onto the machines, SageMaker Training starts the container which invokes the training python script 'finetuning.py' as entrypoint. Thereby it passes the defined hyperparameters as command line arguments. We will dive deeper into our hyperparameter selection at a later point.

The hyperparameters can be parsed by an 'argpars' ArgumentParser:

```
parser = argparse.ArgumentParser()

# Training parameters
parser.add_argument("--model_name_or_path", default="distilgpt2")

args = parser.parse_args()
```

Logging

For logging we use the 'logging' library. We first setup the basic config:

```
# Setup logging
logging.basicConfig(
    format="%(asctime)s - %(levelname)s - %(name)s - %(message)s",
    datefmt="%m/%d/%Y %H:%M:%S",
    handlers=[logging.StreamHandler(sys.stdout)],
)
```

Then we set the log level to 'INFO':

```
log_level = logging.INFO
logger.setLevel(log_level)
```

Finally we configure logging for the HuggingFace frameworks 'datasets' and 'transformers'.

```
datasets.utils.logging.set_verbosity(log_level)
transformers.utils.logging.set_verbosity(log_level)
transformers.utils.logging.enable_default_handler()
transformers.utils.logging.enable_explicit_format()
```

Loading the dataset

We then use the 'datasets' library to load our dataset from the HuggingFace dataset hub:

```
# Downloading and loading a dataset from the hub.
raw_datasets = load_dataset(args.dataset_name)
```

In case the training script we are loading is not available with a train/test split, there is additional functionality implemented to achieve this.

In a real world scenario training data could be ingested from various data sources like S3, databases, ...

Preprocessing

Since we want to finetune the model on a CLM task we need to consider both NLP-related steps and CLM-related steps when it comes to data preprocessing:

For training NLP models the full text string has to be tokenized to enable the model to "digest" it as an input. Beyond access to a huge amount of open-source NLP models, the HuggingFace model hub offers also compatible tokenizers. By utilizing the 'transformers' library we are downloading a tokenizer for the revision of 'distilGPT2' which will be finetuned later on:

```
tokenizer = AutoTokenizer.from_pretrained(args.model_name_or_path,
use_fast = True, revision = args.model_revision)
```

In a similar fashion we are also loading the model artifacts from the HuggingFace model hub:

```
model = AutoModelForCausalLM.from_pretrained(args.model_name_or_path,
revision=args.model_revision, torch_dtype="auto")
```

The tokenizer is now wrapped into an object of the 'AutoTokenizer' class, while the model resides in an object of the 'AutoModelForCausalLM' class.

Tokenization

For tokenization we define a function taking care of the actual tokenization task:

```
def tokenize_function(examples):

    ...

    output = tokenizer(examples[text_column_name])

    ...

    return output
```

Then we utilize it as a higher order function in a map approach on the dataset:

```
tokenized_datasets = raw_datasets.map(
    tokenize_function,
    batched=True,
    remove_columns=column_names,
    desc="Running tokenizer on dataset"
)
```

CLM-related tasks

The utilized training task consumes token blocks of 'block_size' (number of token a model is consuming in one forward pass. This is model specific plus bound to the instance type used for training.) and trains the model using a CLM loss (for details read this). Therefore we need to group our tokenized dataset into token blocks of 'block_size'. We again define a function that performs the acutual grouping task:

Main data processing function that will concatenate all texts from our dataset and generate chunks of block_size.

```
def group_texts(examples):
    # Concatenate all texts.
    concatenated_examples = {k: list(chain(*examples[k])) for k in
examples.keys()}
    total_length = len(concatenated_examples[list(examples.keys())[0]])
    # We drop the small remainder, we could add padding if the model
supported it instead of this drop, you can
    # customize this part to your needs.
    if total_length >= block_size:
        total_length = (total_length // block_size) * block_size
    # Split by chunks of max_len.
    result = {
        k: [t[i : i + block_size] for i in range(0, total_length,
block_size)]
        for k, t in concatenated_examples.items()
    }
    result["labels"] = result["input_ids"].copy()
    return result
```

Then we utilize it as a higher order function in a map approach on the tokenized dataset:

```
lm_datasets = tokenized_datasets.map(
    group_texts,
    batched=True,
    desc=f"Grouping texts in chunks of {block_size}",
)
```

Training

Since we specified that we want to run evaluations on the model to be finetuned (both stepwise during and after the training process), we need to define our evaluation metric first. Therefore we load one of various pre-implemented metrics available using HuggingFace's 'evaluate' library:

```
metric = evaluate.load("accuracy")
```

Then we define a function computing the actual metrics tied to our training job:

```
def compute_metrics(eval_preds):
    preds, labels = eval_preds
    # preds have the same shape as the labels, after the argmax(-1) has
been calculated
    # by preprocess_logits_for_metrics but we need to shift the labels
    labels = labels[:, 1:].reshape(-1)
    preds = preds[:, :-1].reshape(-1)
    return metric.compute(predictions=preds, references=labels)
```

The next step is configuring the actual training job. Therefore we first initialize a 'TrainingArguments' object fed with our hyperparamters plus a seed.:

```
# Specifying training_args. Going with default values for every
parameter not explicitly specified. See documentation for more
information:
https://huggingface.co/docs/transformers/v4.27.2/en/main_classes/trainer#tra
training_args = TrainingArguments(
    per_device_train_batch_size = int(args.per_device_train_batch_size),
    per_device_eval_batch_size=int(args.per_device_eval_batch_size),
    output_dir=args.output_dir,
    seed=42,
    disable_tqdm=False
)
```

Then we initialize the Trainer object, which will orchestrate the training and evaluation process holistically. Several artifacts defined in the flow we executed so far are passed as parameters (model, training_args, datasets, compute_metrics function):

```
# Initialize our Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset if args.do_train else None,
    eval_dataset=eval_dataset if args.do_eval else None,
    tokenizer=tokenizer,
    # Data collator will default to DataCollatorWithPadding, so we
change it.
    data_collator=default_data_collator,
    compute_metrics=compute_metrics if args.do_eval and not
is_torch_tpu_available() else None,
    preprocess_logits_for_metrics=preprocess_logits_for_metrics
    if args.do_eval and not is_torch_tpu_available()
    else None,
)
```

Finally the Trainer's .train() function is invoked executing the actual training. After successful completion the model artifacts are persisted according to the 'output_path' configuration.

```
train_result = trainer.train()
trainer.save_model()  # Saves the tokenizer too for easy upload
```

Evaluation

After successful completion of the training run, we perform a final evaluation:

```
metrics = trainer.evaluate()
```

Hyperparameters

For the finetuning job to be conducted we specify the following hyperparameters explicitly:

- model_name_or_path: model id in HuggingFace ecosystem
- dataset_name: dataset id in HuggingFace ecosystem
- do_train: boolean variable indicating if training run should be executed. In our case 1.
- do_eval: boolean variable indicating if evaluation run should be executed. In our case 1.
- output_dir: directory path for storing the produced model artifacts locally within the container. We pick the default output directory of our SageMaker Training job (will be uploaded to S3 upon job success) '/opt/ml/model'.
- per_device_train_batch_size: batch size to be used when training. We choose 2.
- per_device_eval_batch_size: batch size to used when evaluating. We choose 2.

The default values for the remaining configurable parameters can be found in the Trainer and TrainingArguments documentation.

Configure the environment for model finetuning using the SageMaker HuggingFace Estimator and a AWS HuggingFace DLC

For conveniently training a model with AWS SageMaker Training we can use the Estimator class of SageMaker. Thanks to the AWS x HuggingFace partnership we can use the HuggingFace Estimator natively integrated into the SageMaker SDK, implementing some additional functionality specific to HuggingFace and the HuggingFace model hub. This enables us to finetune the model by providing the training script and some configuration parameters only, while SageMaker is taking care of all the undifferentiated heavy lifting in the background for you. In the constructor we specify the following parameters:

- source_dir: directory path to where the training script file is residing. In our case, this is the relative path to the 'finetuning' directory. Please note, that we've also created a 'requirements.txt' file for installing dependencies the training script requires on container-start time.
- entry_point: file in which the training script is implemented. Residing in the 'finetuning' directory, this is 'finetuning.py'.
- instance_type: EC2 instance type for executing the training job. We pick the 'ml.p3.2xlarge', an instance with 16 GB GPU acceleration (NVIDIA Tesla V100 GPU), 8 vCPUs and 61GB RAM.
- instance_count: size of the ephemeral training cluster. We pick a single node cluster.
- image_uri: The image uri of a Docker image used for training the model. We will be using on of the many ready-to-use Deep Learning Containers AWS is providing here. Deep Learning Containers are Docker images that are preinstalled and tested with the latest versions of popular deep learning frameworks. Deep Learning Containers let you

train models in custom ML environments quickly without building and optimizing your environments from scratch. Since we will be training a model from the HuggingFace model hub by leveraging various HuggingFace frameworks, we will use one of the HuggingFace DLCs, coming with preinstalled python 3.8, pytorch 1.10.2, transformers 4.17.0 dependencies and optimized for training in GPU-accelerated environments.

- py_version: version of the python runtime installed in the container. This parameter is redundant, since we have explicitly specified a container image uri.
- hyperparameters: hyperparameters, passed as command line arguments to the training script.
- output_path: S3 path for storing the artifacts produced by the training job. Therefore, we use the S3 bucket we created in the beginning.

Finally, we execute the SageMaker Training job by calling the .fit() function. This will take a couple of minutes.

In [8]:

```
hyperparameters = {
        "model_name_or_path": 'distilgpt2',
        "dataset_name": 'tiny_shakespeare',
        "do_train": 1,
        "do_eval": 1,
        "output_dir": '/opt/ml/model',
        "per_device_train_batch_size": 2,
        "per_device_eval_batch_size": 2,
        }
```

In [9]:

```
huggingface_estimator = HuggingFace(
                            source_dir='finetuning',
                            entry_point='finetuning.py',
                            instance_type='ml.p3.2xlarge',
                            instance_count=1,
                            role=role,
                            image_uri=f'763104351884.dkr.ecr.{region}.amazonaws.co
                            py_version=None,
                            hyperparameters = hyperparameters,
                            output_path = f's3://{model_bucket_name}'
                            )
```

In [10]:

```
huggingface_estimator.fit()
```

```
Using provided s3_resource
```

```
INFO:sagemaker:Creating training-job with name: huggingface-pytorch-training-2023-
07-31-17-38-58-475
```

```
2023-07-31 17:38:58 Starting - Starting the training job...
2023-07-31 17:39:23 Starting - Preparing the instances for training.........
2023-07-31 17:40:58 Downloading - Downloading input data
2023-07-31 17:40:58 Training - Downloading the training imag
e...........................
2023-07-31 17:45:24 Training - Training image download completed. Training in prog
ress....bash: cannot set terminal process group (-1): Inappropriate ioctl for devi
ce
bash: no job control in this shell
2023-07-31 17:45:47,363 sagemaker-training-toolkit INFO     Imported framework sag
emaker_pytorch_container.training
2023-07-31 17:45:47,389 sagemaker_pytorch_container.training INFO     Block until 
all host DNS lookups succeed.
2023-07-31 17:45:47,391 sagemaker_pytorch_container.training INFO     Invoking use
r training script.
2023-07-31 17:45:47,619 sagemaker-training-toolkit INFO     Installing dependencie
s from requirements.txt:
/opt/conda/bin/python3.8 -m pip install -r requirements.txt
Collecting accelerate
Downloading accelerate-0.21.0-py3-none-any.whl (244 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 244.2/244.2 kB 24.4 MB/s eta 0:00:00
Requirement already satisfied: torch>=1.3 in /opt/conda/lib/python3.8/site-package
s (from -r requirements.txt (line 2)) (1.10.2+cu113)
Collecting datasets==2.10.1
Downloading datasets-2.10.1-py3-none-any.whl (469 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 469.0/469.0 kB 52.3 MB/s eta 0:00:00
Requirement already satisfied: sentencepiece!=0.1.92 in /opt/conda/lib/python3.8/s
ite-packages (from -r requirements.txt (line 4)) (0.1.97)
Requirement already satisfied: protobuf in /opt/conda/lib/python3.8/site-packages 
(from -r requirements.txt (line 5)) (3.19.5)
Requirement already satisfied: scikit-learn in /opt/conda/lib/python3.8/site-packa
ges (from -r requirements.txt (line 6)) (1.1.2)
Requirement already satisfied: transformers==4.17.0 in /opt/conda/lib/python3.8/si
te-packages (from -r requirements.txt (line 7)) (4.17.0)
Collecting evaluate==0.4.0
Downloading evaluate-0.4.0-py3-none-any.whl (81 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 81.4/81.4 kB 24.1 MB/s eta 0:00:00
Requirement already satisfied: responses<0.19 in /opt/conda/lib/python3.8/site-pac
kages (from datasets==2.10.1->-r requirements.txt (line 3)) (0.18.0)
Requirement already satisfied: huggingface-hub<1.0.0,>=0.2.0 in /opt/conda/lib/pyt
hon3.8/site-packages (from datasets==2.10.1->-r requirements.txt (line 3)) (0.10.
0)
Requirement already satisfied: pyyaml>=5.1 in /opt/conda/lib/python3.8/site-packag
es (from datasets==2.10.1->-r requirements.txt (line 3)) (5.4.1)
Requirement already satisfied: aiohttp in /opt/conda/lib/python3.8/site-packages 
(from datasets==2.10.1->-r requirements.txt (line 3)) (3.8.3)
Requirement already satisfied: fsspec[http]>=2021.11.1 in /opt/conda/lib/python3.
8/site-packages (from datasets==2.10.1->-r requirements.txt (line 3)) (2022.8.2)
Requirement already satisfied: packaging in /opt/conda/lib/python3.8/site-packages
(from datasets==2.10.1->-r requirements.txt (line 3)) (21.3)
Requirement already satisfied: dill<0.3.7,>=0.3.0 in /opt/conda/lib/python3.8/site
-packages (from datasets==2.10.1->-r requirements.txt (line 3)) (0.3.5.1)
Requirement already satisfied: tqdm>=4.62.1 in /opt/conda/lib/python3.8/site-packa
ges (from datasets==2.10.1->-r requirements.txt (line 3)) (4.64.0)
Requirement already satisfied: xxhash in /opt/conda/lib/python3.8/site-packages (f
rom datasets==2.10.1->-r requirements.txt (line 3)) (3.0.0)
Requirement already satisfied: multiprocess in /opt/conda/lib/python3.8/site-packa
ges (from datasets==2.10.1->-r requirements.txt (line 3)) (0.70.13)
Requirement already satisfied: pyarrow>=6.0.0 in /opt/conda/lib/python3.8/site-pac
kages (from datasets==2.10.1->-r requirements.txt (line 3)) (9.0.0)
Requirement already satisfied: pandas in /opt/conda/lib/python3.8/site-packages (f
rom datasets==2.10.1->-r requirements.txt (line 3)) (1.5.0)
Requirement already satisfied: requests>=2.19.0 in /opt/conda/lib/python3.8/site-p
ackages (from datasets==2.10.1->-r requirements.txt (line 3)) (2.28.1)
```

```
Requirement already satisfied: numpy>=1.17 in /opt/conda/lib/python3.8/site-packag
es (from datasets==2.10.1->-r requirements.txt (line 3)) (1.22.2)
Requirement already satisfied: tokenizers!=0.11.3,>=0.11.1 in /opt/conda/lib/pytho
n3.8/site-packages (from transformers==4.17.0->-r requirements.txt (line 7)) (0.1
3.0)
Requirement already satisfied: filelock in /opt/conda/lib/python3.8/site-packages
(from transformers==4.17.0->-r requirements.txt (line 7)) (3.8.0)
Requirement already satisfied: sacremoses in /opt/conda/lib/python3.8/site-package
s (from transformers==4.17.0->-r requirements.txt (line 7)) (0.0.53)
Requirement already satisfied: regex!=2019.12.17 in /opt/conda/lib/python3.8/site-
packages (from transformers==4.17.0->-r requirements.txt (line 7)) (2022.9.13)
Requirement already satisfied: psutil in /opt/conda/lib/python3.8/site-packages (f
rom accelerate->-r requirements.txt (line 1)) (5.9.2)
Requirement already satisfied: typing-extensions in /opt/conda/lib/python3.8/site-
packages (from torch>=1.3->-r requirements.txt (line 2)) (4.3.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /opt/conda/lib/python3.8/si
te-packages (from scikit-learn->-r requirements.txt (line 6)) (3.1.0)
Requirement already satisfied: joblib>=1.0.0 in /opt/conda/lib/python3.8/site-pack
ages (from scikit-learn->-r requirements.txt (line 6)) (1.2.0)
Requirement already satisfied: scipy>=1.3.2 in /opt/conda/lib/python3.8/site-packa
ges (from scikit-learn->-r requirements.txt (line 6)) (1.9.1)
Requirement already satisfied: aiosignal>=1.1.2 in /opt/conda/lib/python3.8/site-p
ackages (from aiohttp->datasets==2.10.1->-r requirements.txt (line 3)) (1.2.0)
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in /opt/conda/lib/pyth
on3.8/site-packages (from aiohttp->datasets==2.10.1->-r requirements.txt (line 3))
(2.0.12)
Requirement already satisfied: frozenlist>=1.1.1 in /opt/conda/lib/python3.8/site-
packages (from aiohttp->datasets==2.10.1->-r requirements.txt (line 3)) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /opt/conda/lib/python3.8/site-pack
ages (from aiohttp->datasets==2.10.1->-r requirements.txt (line 3)) (21.4.0)
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/conda/lib/python3.8/site-pac
kages (from aiohttp->datasets==2.10.1->-r requirements.txt (line 3)) (1.8.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /opt/conda/lib/python3.8/sit
e-packages (from aiohttp->datasets==2.10.1->-r requirements.txt (line 3)) (6.0.2)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in /opt/conda/lib/pytho
n3.8/site-packages (from aiohttp->datasets==2.10.1->-r requirements.txt (line 3))
(4.0.2)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /opt/conda/lib/python3.
8/site-packages (from packaging->datasets==2.10.1->-r requirements.txt (line 3))
(3.0.9)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.8/site-packa
ges (from requests>=2.19.0->datasets==2.10.1->-r requirements.txt (line 3)) (3.3)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.8/site
-packages (from requests>=2.19.0->datasets==2.10.1->-r requirements.txt (line 3))
(2022.9.24)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/conda/lib/python3.8/s
ite-packages (from requests>=2.19.0->datasets==2.10.1->-r requirements.txt (line
3)) (1.26.12)
Requirement already satisfied: python-dateutil>=2.8.1 in /opt/conda/lib/python3.8/
site-packages (from pandas->datasets==2.10.1->-r requirements.txt (line 3)) (2.8.
2)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.8/site-packa
ges (from pandas->datasets==2.10.1->-r requirements.txt (line 3)) (2022.2.1)
Requirement already satisfied: click in /opt/conda/lib/python3.8/site-packages (fr
om sacremoses->transformers==4.17.0->-r requirements.txt (line 7)) (8.1.3)
Requirement already satisfied: six in /opt/conda/lib/python3.8/site-packages (from
sacremoses->transformers==4.17.0->-r requirements.txt (line 7)) (1.16.0)
Installing collected packages: accelerate, datasets, evaluate
Attempting uninstall: datasets
Found existing installation: datasets 1.18.4
Uninstalling datasets-1.18.4:
Successfully uninstalled datasets-1.18.4
Successfully installed accelerate-0.21.0 datasets-2.10.1 evaluate-0.4.0
WARNING: Running pip as the 'root' user can result in broken permissions and confl
```

```
icting behaviour with the system package manager. It is recommended to use a virtu
al environment instead: https://pip.pypa.io/warnings/venv
[notice] A new release of pip available: 22.2.2 -> 23.2.1
[notice] To update, run: pip install --upgrade pip
2023-07-31 17:45:51,415 sagemaker-training-toolkit INFO     Waiting for the proces
s to finish and give a return code.
2023-07-31 17:45:51,415 sagemaker-training-toolkit INFO     Done waiting for a ret
urn code. Received 0 from exiting process.
2023-07-31 17:45:51,497 sagemaker-training-toolkit INFO     Invoking user script
Training Env:
{
    "additional_framework_parameters": {},
    "channel_input_dirs": {},
    "current_host": "algo-1",
    "current_instance_group": "homogeneousCluster",
    "current_instance_group_hosts": [
        "algo-1"
    ],
    "current_instance_type": "ml.p3.2xlarge",
    "distribution_hosts": [],
    "distribution_instance_groups": [],
    "framework_module": "sagemaker_pytorch_container.training:main",
    "hosts": [
        "algo-1"
    ],
    "hyperparameters": {
        "dataset_name": "tiny_shakespeare",
        "do_eval": 1,
        "do_train": 1,
        "model_name_or_path": "distilgpt2",
        "output_dir": "/opt/ml/model",
        "per_device_eval_batch_size": 2,
        "per_device_train_batch_size": 2
    },
    "input_config_dir": "/opt/ml/input/config",
    "input_data_config": {},
    "input_dir": "/opt/ml/input",
    "instance_groups": [
        "homogeneousCluster"
    ],
    "instance_groups_dict": {
        "homogeneousCluster": {
            "instance_group_name": "homogeneousCluster",
            "instance_type": "ml.p3.2xlarge",
            "hosts": [
                "algo-1"
            ]
        }
    },
    "is_hetero": false,
    "is_master": true,
    "is_modelparallel_enabled": null,
    "job_name": "huggingface-pytorch-training-2023-07-31-17-38-58-475",
    "log_level": 20,
    "master_hostname": "algo-1",
    "model_dir": "/opt/ml/model",
    "module_dir": "s3://immersion-day-bucket-882819251225/huggingface-pytorch-trai
ning-2023-07-31-17-38-58-475/source/sourcedir.tar.gz",
    "module_name": "finetuning",
    "network_interface_name": "eth0",
    "num_cpus": 8,
    "num_gpus": 1,
    "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.p3.2xlarge",
        "current_group_name": "homogeneousCluster",
        "hosts": [
            "algo-1"
        ],
        "instance_groups": [
            {
                "instance_group_name": "homogeneousCluster",
                "instance_type": "ml.p3.2xlarge",
                "hosts": [
                    "algo-1"
                ]
            }
        ],
        "network_interface_name": "eth0"
    },
    "user_entry_point": "finetuning.py"
}
Environment variables:
SM_HOSTS=["algo-1"]
SM_NETWORK_INTERFACE_NAME=eth0
SM_HPS={"dataset_name":"tiny_shakespeare","do_eval":1,"do_train":1,"model_name_or_path":"distilgpt2","output_dir":"/opt/ml/model","per_device_eval_batch_size":2,"per_device_train_batch_size":2}
SM_USER_ENTRY_POINT=finetuning.py
SM_FRAMEWORK_PARAMS={}
SM_RESOURCE_CONFIG={"current_group_name":"homogeneousCluster","current_host":"algo-1","current_instance_type":"ml.p3.2xlarge","hosts":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.p3.2xlarge"}],"network_interface_name":"eth0"}
SM_INPUT_DATA_CONFIG={}
SM_OUTPUT_DATA_DIR=/opt/ml/output/data
SM_CHANNELS=[]
SM_CURRENT_HOST=algo-1
SM_CURRENT_INSTANCE_TYPE=ml.p3.2xlarge
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.p3.2xlarge"}}
SM_DISTRIBUTION_INSTANCE_GROUPS=[]
SM_IS_HETERO=false
SM_MODULE_NAME=finetuning
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
SM_NUM_GPUS=1
SM_MODEL_DIR=/opt/ml/model
SM_MODULE_DIR=s3://immersion-day-bucket-882819251225/huggingface-pytorch-training-2023-07-31-17-38-58-475/source/sourcedir.tar.gz
SM_TRAINING_ENV={"additional_framework_parameters":{},"channel_input_dirs":{},"current_host":"algo-1","current_instance_group":"homogeneousCluster","current_instance_group_hosts":["algo-1"],"current_instance_type":"ml.p3.2xlarge","distribution_hosts":[],"distribution_instance_groups":[],"framework_module":"sagemaker_pytorch_container.training:main","hosts":["algo-1"],"hyperparameters":{"dataset_name":"tiny_shakespeare","do_eval":1,"do_train":1,"model_name_or_path":"distilgpt2","output_dir":"/opt/ml/model","per_device_eval_batch_size":2,"per_device_train_batch_size":2},"input_config_dir":"/opt/ml/input/config","input_data_config":{},"input_dir":"/
```

```
opt/ml/input","instance_groups":["homogeneousCluster"],"instance_groups_dict":{"ho
mogeneousCluster":{"hosts":["algo-1"],"instance_group_name":"homogeneousCluste
r","instance_type":"ml.p3.2xlarge"}},"is_hetero":false,"is_master":true,"is_modelp
arallel_enabled":null,"job_name":"huggingface-pytorch-training-2023-07-31-17-38-58
-475","log_level":20,"master_hostname":"algo-1","model_dir":"/opt/ml/model","modul
e_dir":"s3://immersion-day-bucket-882819251225/huggingface-pytorch-training-2023-0
7-31-17-38-58-475/source/sourcedir.tar.gz","module_name":"finetuning","network_int
erface_name":"eth0","num_cpus":8,"num_gpus":1,"output_data_dir":"/opt/ml/output/da
ta","output_dir":"/opt/ml/output","output_intermediate_dir":"/opt/ml/output/interm
ediate","resource_config":{"current_group_name":"homogeneousCluster","current_hos
t":"algo-1","current_instance_type":"ml.p3.2xlarge","hosts":["algo-1"],"instance_g
roups":[{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_t
ype":"ml.p3.2xlarge"}],"network_interface_name":"eth0"},"user_entry_point":"finetu
ning.py"}
SM_USER_ARGS=["--dataset_name","tiny_shakespeare","--do_eval","1","--do_trai
n","1","--model_name_or_path","distilgpt2","--output_dir","/opt/ml/model","--per_d
evice_eval_batch_size","2","--per_device_train_batch_size","2"]
SM_OUTPUT_INTERMEDIATE_DIR=/opt/ml/output/intermediate
SM_HP_DATASET_NAME=tiny_shakespeare
SM_HP_DO_EVAL=1
SM_HP_DO_TRAIN=1
SM_HP_MODEL_NAME_OR_PATH=distilgpt2
SM_HP_OUTPUT_DIR=/opt/ml/model
SM_HP_PER_DEVICE_EVAL_BATCH_SIZE=2
SM_HP_PER_DEVICE_TRAIN_BATCH_SIZE=2
PYTHONPATH=/opt/ml/code:/opt/conda/bin:/opt/conda/lib/python38.zip:/opt/conda/lib/
python3.8:/opt/conda/lib/python3.8/lib-dynload:/opt/conda/lib/python3.8/site-packa
ges:/opt/conda/lib/python3.8/site-packages/smdebug-1.0.22b20220929-py3.8.egg:/opt/
conda/lib/python3.8/site-packages/pyinstrument-3.4.2-py3.8.egg:/opt/conda/lib/pyth
on3.8/site-packages/pyinstrument_cext-0.2.4-py3.8-linux-x86_64.egg
Invoking script with the following command:
/opt/conda/bin/python3.8 finetuning.py --dataset_name tiny_shakespeare --do_eval 1
--do_train 1 --model_name_or_path distilgpt2 --output_dir /opt/ml/model --per_devi
ce_eval_batch_size 2 --per_device_train_batch_size 2
train.py starting...
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - https://huggingface.co/da
tasets/tiny_shakespeare/resolve/main/tiny_shakespeare.py not found in cache or for
ce_download set to True, downloading to /root/.cache/huggingface/datasets/download
s/tmpw82on1yf
Downloading builder script:   0%|          | 0.00/3.73k [00:00<?, ?B/s]
Downloading builder script: 100%|██████████| 3.73k/3.73k [00:00<00:00, 4.86MB/s]
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - storing https://huggingfa
ce.co/datasets/tiny_shakespeare/resolve/main/tiny_shakespeare.py in cache at /roo
t/.cache/huggingface/datasets/downloads/43f3ce4359ea1c3db118d95df089c52b911681d4ad
1723b3d94702ba6f6ce328.72b16fe4a2d35a25b07ff4bde25842eae4288e09e8060a707cd020596c1
9e063.py
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - creating metadata file fo
r /root/.cache/huggingface/datasets/downloads/43f3ce4359ea1c3db118d95df089c52b9116
81d4ad1723b3d94702ba6f6ce328.72b16fe4a2d35a25b07ff4bde25842eae4288e09e8060a707cd02
0596c19e063.py
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - https://huggingface.co/da
tasets/tiny_shakespeare/resolve/main/dataset_infos.json not found in cache or forc
e_download set to True, downloading to /root/.cache/huggingface/datasets/download
s/tmpa_2fz2_g
Downloading metadata:   0%|          | 0.00/1.90k [00:00<?, ?B/s]
Downloading metadata: 100%|██████████| 1.90k/1.90k [00:00<00:00, 2.57MB/s]
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - storing https://huggingfa
ce.co/datasets/tiny_shakespeare/resolve/main/dataset_infos.json in cache at /roo
t/.cache/huggingface/datasets/downloads/fef47684d373166fcb85d6f9b08bcb4386987b78c8
fc41f73960e283eddb06aa.de0c17192decaea7c4ee839ebb1e28d23e7f9e23b9ebcfd45ed59f9a54f
3495d
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - creating metadata file fo
r /root/.cache/huggingface/datasets/downloads/fef47684d373166fcb85d6f9b08bcb438698
7b78c8fc41f73960e283eddb06aa.de0c17192decaea7c4ee839ebb1e28d23e7f9e23b9ebcfd45ed59
```

```
f9a54f3495d
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - https://huggingface.co/da
tasets/tiny_shakespeare/resolve/main/README.md not found in cache or force_downloa
d set to True, downloading to /root/.cache/huggingface/datasets/downloads/tmp6l0po
3m5
Downloading readme:   0%|          | 0.00/6.10k [00:00<?, ?B/s]
Downloading readme: 100%|██████████| 6.10k/6.10k [00:00<00:00, 5.87MB/s]
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - storing https://huggingfa
ce.co/datasets/tiny_shakespeare/resolve/main/README.md in cache at /root/.cache/hu
ggingface/datasets/downloads/7b14ce0570dd475e39f665b85e42e4ec9c7dd0c7d16421326f975
f4a801c01b2.953a74ee58b0f768125607ae856e7975d0f3e897cf217a776276f2500ff7ad96
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - creating metadata file fo
r /root/.cache/huggingface/datasets/downloads/7b14ce0570dd475e39f665b85e42e4ec9c7d
d0c7d16421326f975f4a801c01b2.953a74ee58b0f768125607ae856e7975d0f3e897cf217a776276f
2500ff7ad96
07/31/2023 17:45:56 - INFO - datasets.info - Loading Dataset Infos from /root/.cac
he/huggingface/modules/datasets_modules/datasets/tiny_shakespeare/b5b13969f09fe870
7337f6cb296314fbe06960bd9a868dca39e713e163d27b5e
07/31/2023 17:45:56 - INFO - datasets.builder - Generating dataset tiny_shakespear
e (/root/.cache/huggingface/datasets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8
707337f6cb296314fbe06960bd9a868dca39e713e163d27b5e)
Downloading and preparing dataset tiny_shakespeare/default to /root/.cache/hugging
face/datasets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8707337f6cb296314fbe0696
0bd9a868dca39e713e163d27b5e...
07/31/2023 17:45:56 - INFO - datasets.builder - Dataset not on Hf google storage.
Downloading and preparing it from source
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - https://raw.githubusercon
tent.com/karpathy/char-rnn/master/data/tinyshakespeare/input.txt not found in cach
e or force_download set to True, downloading to /root/.cache/huggingface/datasets/
downloads/tmpau8kia1g
Downloading data:   0%|          | 0.00/435k [00:00<?, ?B/s]
Downloading data: 1.12MB [00:00, 51.1MB/s]
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - storing https://raw.githu
busercontent.com/karpathy/char-rnn/master/data/tinyshakespeare/input.txt in cache
at /root/.cache/huggingface/datasets/downloads/82880ef7df02a44e79ee0148f39275e856e
d335220dc1d324a3f54852e9fec63
07/31/2023 17:45:56 - INFO - datasets.utils.file_utils - creating metadata file fo
r /root/.cache/huggingface/datasets/downloads/82880ef7df02a44e79ee0148f39275e856ed
335220dc1d324a3f54852e9fec63
07/31/2023 17:45:56 - INFO - datasets.download.download_manager - Downloading took
0.0 min
07/31/2023 17:45:56 - INFO - datasets.download.download_manager - Checksum Computa
tion took 0.0 min
07/31/2023 17:45:56 - INFO - datasets.builder - Generating train split
Generating train split:   0%|          | 0/1 [00:00<?, ? examples/s]
07/31/2023 17:45:56 - INFO - datasets.builder - Generating validation split
Generating validation split:   0%|          | 0/1 [00:00<?, ? examples/s]
07/31/2023 17:45:56 - INFO - datasets.builder - Generating test split
Generating test split:   0%|          | 0/1 [00:00<?, ? examples/s]
07/31/2023 17:45:56 - INFO - datasets.utils.info_utils - All the splits matched su
ccessfully.
Dataset tiny_shakespeare downloaded and prepared to /root/.cache/huggingface/datas
ets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8707337f6cb296314fbe06960bd9a868dc
a39e713e163d27b5e. Subsequent calls will reuse this data.
0%|          | 0/3 [00:00<?, ?it/s]
100%|██████████| 3/3 [00:00<00:00, 963.47it/s]
[INFO|tokenization_auto.py:344] 2023-07-31 17:45:56,797 >> Could not locate the to
kenizer configuration file, will try to use the model config instead.
[INFO|tokenization_auto.py:344] 2023-07-31 17:45:56,797 >> Could not locate the to
kenizer configuration file, will try to use the model config instead.
[INFO|file_utils.py:2215] 2023-07-31 17:45:56,825 >> https://huggingface.co/distil
gpt2/resolve/main/config.json not found in cache or force_download set to True, do
wnloading to /root/.cache/huggingface/transformers/tmpk60hv9v2
[INFO|file_utils.py:2215] 2023-07-31 17:45:56,825 >> https://huggingface.co/distil
```

```
gpt2/resolve/main/config.json not found in cache or force_download set to True, do
wnloading to /root/.cache/huggingface/transformers/tmpk60hv9v2
Downloading:   0%|          | 0.00/762 [00:00<?, ?B/s]
Downloading: 100%|██████████| 762/762 [00:00<00:00, 1.02MB/s]
[INFO|file_utils.py:2219] 2023-07-31 17:45:56,852 >> storing https://huggingface.c
o/distilgpt2/resolve/main/config.json in cache at /root/.cache/huggingface/transfo
rmers/f985248d2791fcff97732e4ee263617adec1edb5429a2b8421734c6d14e39bee.422318838d1
ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc81631
[INFO|file_utils.py:2219] 2023-07-31 17:45:56,852 >> storing https://huggingface.c
o/distilgpt2/resolve/main/config.json in cache at /root/.cache/huggingface/transfo
rmers/f985248d2791fcff97732e4ee263617adec1edb5429a2b8421734c6d14e39bee.422318838d1
ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc81631
[INFO|file_utils.py:2227] 2023-07-31 17:45:56,852 >> creating metadata file for /r
oot/.cache/huggingface/transformers/f985248d2791fcff97732e4ee263617adec1edb5429a2b
8421734c6d14e39bee.422318838d1ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc8163
1
[INFO|file_utils.py:2227] 2023-07-31 17:45:56,852 >> creating metadata file for /r
oot/.cache/huggingface/transformers/f985248d2791fcff97732e4ee263617adec1edb5429a2b
8421734c6d14e39bee.422318838d1ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc8163
1
[INFO|configuration_utils.py:648] 2023-07-31 17:45:56,853 >> loading configuration
file https://huggingface.co/distilgpt2/resolve/main/config.json from cache at /roo
t/.cache/huggingface/transformers/f985248d2791fcff97732e4ee263617adec1edb5429a2b84
21734c6d14e39bee.422318838d1ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc81631
[INFO|configuration_utils.py:648] 2023-07-31 17:45:56,853 >> loading configuration
file https://huggingface.co/distilgpt2/resolve/main/config.json from cache at /roo
t/.cache/huggingface/transformers/f985248d2791fcff97732e4ee263617adec1edb5429a2b84
21734c6d14e39bee.422318838d1ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc81631
[INFO|configuration_utils.py:684] 2023-07-31 17:45:56,854 >> Model config GPT2Conf
ig {
  "_name_or_path": "distilgpt2",
  "_num_labels": 1,
  "activation_function": "gelu_new",
  "architectures": [
    "GPT2LMHeadModel"
  ],
  "attn_pdrop": 0.1,
  "bos_token_id": 50256,
  "embd_pdrop": 0.1,
  "eos_token_id": 50256,
  "id2label": {
    "0": "LABEL_0"
  },
  "initializer_range": 0.02,
  "label2id": {
    "LABEL_0": 0
  },
  "layer_norm_epsilon": 1e-05,
  "model_type": "gpt2",
  "n_ctx": 1024,
  "n_embd": 768,
  "n_head": 12,
  "n_inner": null,
  "n_layer": 6,
  "n_positions": 1024,
  "reorder_and_upcast_attn": false,
  "resid_pdrop": 0.1,
  "scale_attn_by_inverse_layer_idx": false,
  "scale_attn_weights": true,
  "summary_activation": null,
  "summary_first_dropout": 0.1,
  "summary_proj_to_labels": true,
  "summary_type": "cls_index",
  "summary_use_proj": true,
  "task_specific_params": {
    "text-generation": {
      "do_sample": true,
      "max_length": 50
    }
  },
  "transformers_version": "4.17.0",
  "use_cache": true,
  "vocab_size": 50257
}
[INFO|configuration_utils.py:684] 2023-07-31 17:45:56,854 >> Model config GPT2Conf
ig {
  "_name_or_path": "distilgpt2",
  "_num_labels": 1,
  "activation_function": "gelu_new",
  "architectures": [
    "GPT2LMHeadModel"
  ],
  "attn_pdrop": 0.1,
  "bos_token_id": 50256,
  "embd_pdrop": 0.1,
  "eos_token_id": 50256,
  "id2label": {
    "0": "LABEL_0"
  },
  "initializer_range": 0.02,
  "label2id": {
    "LABEL_0": 0
  },
  "layer_norm_epsilon": 1e-05,
  "model_type": "gpt2",
  "n_ctx": 1024,
  "n_embd": 768,
  "n_head": 12,
  "n_inner": null,
  "n_layer": 6,
  "n_positions": 1024,
  "reorder_and_upcast_attn": false,
  "resid_pdrop": 0.1,
  "scale_attn_by_inverse_layer_idx": false,
  "scale_attn_weights": true,
  "summary_activation": null,
  "summary_first_dropout": 0.1,
  "summary_proj_to_labels": true,
  "summary_type": "cls_index",
  "summary_use_proj": true,
  "task_specific_params": {
    "text-generation": {
      "do_sample": true,
      "max_length": 50
    }
  },
  "transformers_version": "4.17.0",
  "use_cache": true,
  "vocab_size": 50257
}
[INFO|file_utils.py:2215] 2023-07-31 17:45:56,926 >> https://huggingface.co/distil
gpt2/resolve/main/vocab.json not found in cache or force_download set to True, dow
nloading to /root/.cache/huggingface/transformers/tmp2a2yecak
[INFO|file_utils.py:2215] 2023-07-31 17:45:56,926 >> https://huggingface.co/distil
gpt2/resolve/main/vocab.json not found in cache or force_download set to True, dow
nloading to /root/.cache/huggingface/transformers/tmp2a2yecak
Downloading:   0%|          | 0.00/0.99M [00:00<?, ?B/s]
Downloading: 100%|██████████| 0.99M/0.99M [00:00<00:00, 35.9MB/s]
```

```
[INFO|file_utils.py:2219] 2023-07-31 17:45:56,986 >> storing https://huggingface.c
o/distilgpt2/resolve/main/vocab.json in cache at /root/.cache/huggingface/transfor
mers/55051ac97dcc32f0a736d21a32a4d42b0d9b90f117ca7c38e65038b04bd5c3f5.c7ed1f96aac4
9e745788faa77ba0a26a392643a50bb388b9c04ff469e555241f
[INFO|file_utils.py:2219] 2023-07-31 17:45:56,986 >> storing https://huggingface.c
o/distilgpt2/resolve/main/vocab.json in cache at /root/.cache/huggingface/transfor
mers/55051ac97dcc32f0a736d21a32a4d42b0d9b90f117ca7c38e65038b04bd5c3f5.c7ed1f96aac4
9e745788faa77ba0a26a392643a50bb388b9c04ff469e555241f
[INFO|file_utils.py:2227] 2023-07-31 17:45:56,987 >> creating metadata file for /r
oot/.cache/huggingface/transformers/55051ac97dcc32f0a736d21a32a4d42b0d9b90f117ca7c
38e65038b04bd5c3f5.c7ed1f96aac49e745788faa77ba0a26a392643a50bb388b9c04ff469e555241
f
[INFO|file_utils.py:2227] 2023-07-31 17:45:56,987 >> creating metadata file for /r
oot/.cache/huggingface/transformers/55051ac97dcc32f0a736d21a32a4d42b0d9b90f117ca7c
38e65038b04bd5c3f5.c7ed1f96aac49e745788faa77ba0a26a392643a50bb388b9c04ff469e555241
f
[INFO|file_utils.py:2215] 2023-07-31 17:45:57,017 >> https://huggingface.co/distil
gpt2/resolve/main/merges.txt not found in cache or force_download set to True, dow
nloading to /root/.cache/huggingface/transformers/tmp2m3wmu6l
[INFO|file_utils.py:2215] 2023-07-31 17:45:57,017 >> https://huggingface.co/distil
gpt2/resolve/main/merges.txt not found in cache or force_download set to True, dow
nloading to /root/.cache/huggingface/transformers/tmp2m3wmu6l
Downloading:   0%|          | 0.00/446k [00:00<?, ?B/s]
Downloading: 100%|██████████| 446k/446k [00:00<00:00, 62.9MB/s]
[INFO|file_utils.py:2219] 2023-07-31 17:45:57,051 >> storing https://huggingface.c
o/distilgpt2/resolve/main/merges.txt in cache at /root/.cache/huggingface/transfor
mers/9dfb299b74cdf7601ba7cd3a8073dbdac351caec0ed7ab5849b098b3c8ae3d57.5d12962c5ee6
15a4c803841266e9c3be9a691a924f72d395d3a6c6c81157788b
[INFO|file_utils.py:2219] 2023-07-31 17:45:57,051 >> storing https://huggingface.c
o/distilgpt2/resolve/main/merges.txt in cache at /root/.cache/huggingface/transfor
mers/9dfb299b74cdf7601ba7cd3a8073dbdac351caec0ed7ab5849b098b3c8ae3d57.5d12962c5ee6
15a4c803841266e9c3be9a691a924f72d395d3a6c6c81157788b
[INFO|file_utils.py:2227] 2023-07-31 17:45:57,051 >> creating metadata file for /r
oot/.cache/huggingface/transformers/9dfb299b74cdf7601ba7cd3a8073dbdac351caec0ed7ab
5849b098b3c8ae3d57.5d12962c5ee615a4c803841266e9c3be9a691a924f72d395d3a6c6c81157788
b
[INFO|file_utils.py:2227] 2023-07-31 17:45:57,051 >> creating metadata file for /r
oot/.cache/huggingface/transformers/9dfb299b74cdf7601ba7cd3a8073dbdac351caec0ed7ab
5849b098b3c8ae3d57.5d12962c5ee615a4c803841266e9c3be9a691a924f72d395d3a6c6c81157788
b
[INFO|file_utils.py:2215] 2023-07-31 17:45:57,081 >> https://huggingface.co/distil
gpt2/resolve/main/tokenizer.json not found in cache or force_download set to True,
downloading to /root/.cache/huggingface/transformers/tmpv7niqg3b
[INFO|file_utils.py:2215] 2023-07-31 17:45:57,081 >> https://huggingface.co/distil
gpt2/resolve/main/tokenizer.json not found in cache or force_download set to True,
downloading to /root/.cache/huggingface/transformers/tmpv7niqg3b
Downloading:   0%|          | 0.00/1.29M [00:00<?, ?B/s]
Downloading: 100%|██████████| 1.29M/1.29M [00:00<00:00, 33.4MB/s]
[INFO|file_utils.py:2219] 2023-07-31 17:45:57,151 >> storing https://huggingface.c
o/distilgpt2/resolve/main/tokenizer.json in cache at /root/.cache/huggingface/tran
sformers/accb287b5a5396b2597382916b6cc939fdab1366e89475a92338d3971b3d02b7.cf2d0ecb
83b6df91b3dbb53f1d1e4c311578bfd3aa0e04934215a49bf9898df0
[INFO|file_utils.py:2219] 2023-07-31 17:45:57,151 >> storing https://huggingface.c
o/distilgpt2/resolve/main/tokenizer.json in cache at /root/.cache/huggingface/tran
sformers/accb287b5a5396b2597382916b6cc939fdab1366e89475a92338d3971b3d02b7.cf2d0ecb
83b6df91b3dbb53f1d1e4c311578bfd3aa0e04934215a49bf9898df0
[INFO|file_utils.py:2227] 2023-07-31 17:45:57,151 >> creating metadata file for /r
oot/.cache/huggingface/transformers/accb287b5a5396b2597382916b6cc939fdab1366e89475
a92338d3971b3d02b7.cf2d0ecb83b6df91b3dbb53f1d1e4c311578bfd3aa0e04934215a49bf9898df
0
[INFO|file_utils.py:2227] 2023-07-31 17:45:57,151 >> creating metadata file for /r
oot/.cache/huggingface/transformers/accb287b5a5396b2597382916b6cc939fdab1366e89475
a92338d3971b3d02b7.cf2d0ecb83b6df91b3dbb53f1d1e4c311578bfd3aa0e04934215a49bf9898df
0
```

```
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/vocab.json from cache at /root/.cache/
huggingface/transformers/55051ac97dcc32f0a736d21a32a4d42b0d9b90f117ca7c38e65038b04
bd5c3f5.c7ed1f96aac49e745788faa77ba0a26a392643a50bb388b9c04ff469e555241f
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/vocab.json from cache at /root/.cache/
huggingface/transformers/55051ac97dcc32f0a736d21a32a4d42b0d9b90f117ca7c38e65038b04
bd5c3f5.c7ed1f96aac49e745788faa77ba0a26a392643a50bb388b9c04ff469e555241f
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/merges.txt from cache at /root/.cache/
huggingface/transformers/9dfb299b74cdf7601ba7cd3a8073dbdac351caec0ed7ab5849b098b3c
8ae3d57.5d12962c5ee615a4c803841266e9c3be9a691a924f72d395d3a6c6c81157788b
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/tokenizer.json from cache at /root/.ca
che/huggingface/transformers/accb287b5a5396b2597382916b6cc939fdab1366e89475a92338d
3971b3d02b7.cf2d0ecb83b6df91b3dbb53f1d1e4c311578bfd3aa0e04934215a49bf9898df0
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/added_tokens.json from cache at None
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/merges.txt from cache at /root/.cache/
huggingface/transformers/9dfb299b74cdf7601ba7cd3a8073dbdac351caec0ed7ab5849b098b3c
8ae3d57.5d12962c5ee615a4c803841266e9c3be9a691a924f72d395d3a6c6c81157788b
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/tokenizer.json from cache at /root/.ca
che/huggingface/transformers/accb287b5a5396b2597382916b6cc939fdab1366e89475a92338d
3971b3d02b7.cf2d0ecb83b6df91b3dbb53f1d1e4c311578bfd3aa0e04934215a49bf9898df0
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/added_tokens.json from cache at None
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/special_tokens_map.json from cache at
None
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/special_tokens_map.json from cache at
None
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/tokenizer_config.json from cache at No
ne
[INFO|tokenization_utils_base.py:1786] 2023-07-31 17:45:57,235 >> loading file htt
ps://huggingface.co/distilgpt2/resolve/main/tokenizer_config.json from cache at No
ne
[INFO|configuration_utils.py:648] 2023-07-31 17:45:57,260 >> loading configuration
file https://huggingface.co/distilgpt2/resolve/main/config.json from cache at /roo
t/.cache/huggingface/transformers/f985248d2791fcff97732e4ee263617adec1edb5429a2b84
21734c6d14e39bee.422318838d1ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc81631
[INFO|configuration_utils.py:648] 2023-07-31 17:45:57,260 >> loading configuration
file https://huggingface.co/distilgpt2/resolve/main/config.json from cache at /roo
t/.cache/huggingface/transformers/f985248d2791fcff97732e4ee263617adec1edb5429a2b84
21734c6d14e39bee.422318838d1ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc81631
[INFO|configuration_utils.py:684] 2023-07-31 17:45:57,261 >> Model config GPT2Conf
ig {
  "_name_or_path": "distilgpt2",
  "_num_labels": 1,
  "activation_function": "gelu_new",
  "architectures": [
    "GPT2LMHeadModel"
  ],
  "attn_pdrop": 0.1,
  "bos_token_id": 50256,
  "embd_pdrop": 0.1,
  "eos_token_id": 50256,
  "id2label": {
    "0": "LABEL_0"
  },
  "initializer_range": 0.02,
  "label2id": {
    "LABEL_0": 0
  },
  "layer_norm_epsilon": 1e-05,
  "model_type": "gpt2",
  "n_ctx": 1024,
  "n_embd": 768,
  "n_head": 12,
  "n_inner": null,
  "n_layer": 6,
  "n_positions": 1024,
  "reorder_and_upcast_attn": false,
  "resid_pdrop": 0.1,
  "scale_attn_by_inverse_layer_idx": false,
  "scale_attn_weights": true,
  "summary_activation": null,
  "summary_first_dropout": 0.1,
  "summary_proj_to_labels": true,
  "summary_type": "cls_index",
  "summary_use_proj": true,
  "task_specific_params": {
    "text-generation": {
      "do_sample": true,
      "max_length": 50
    }
  },
  "transformers_version": "4.17.0",
  "use_cache": true,
  "vocab_size": 50257
}
[INFO|configuration_utils.py:684] 2023-07-31 17:45:57,261 >> Model config GPT2Conf
ig {
  "_name_or_path": "distilgpt2",
  "_num_labels": 1,
  "activation_function": "gelu_new",
  "architectures": [
    "GPT2LMHeadModel"
  ],
  "attn_pdrop": 0.1,
  "bos_token_id": 50256,
  "embd_pdrop": 0.1,
  "eos_token_id": 50256,
  "id2label": {
    "0": "LABEL_0"
  },
  "initializer_range": 0.02,
  "label2id": {
    "LABEL_0": 0
  },
  "layer_norm_epsilon": 1e-05,
  "model_type": "gpt2",
  "n_ctx": 1024,
  "n_embd": 768,
  "n_head": 12,
  "n_inner": null,
  "n_layer": 6,
  "n_positions": 1024,
  "reorder_and_upcast_attn": false,
  "resid_pdrop": 0.1,
  "scale_attn_by_inverse_layer_idx": false,
  "scale_attn_weights": true,
  "summary_activation": null,
  "summary_first_dropout": 0.1,
  "summary_proj_to_labels": true,
  "summary_type": "cls_index",
  "summary_use_proj": true,
  "task_specific_params": {
    "text-generation": {
      "do_sample": true,
      "max_length": 50
    }
  },
  "transformers_version": "4.17.0",
  "use_cache": true,
  "vocab_size": 50257
}
[INFO|configuration_utils.py:648] 2023-07-31 17:45:57,371 >> loading configuration
file https://huggingface.co/distilgpt2/resolve/main/config.json from cache at /roo
t/.cache/huggingface/transformers/f985248d2791fcff97732e4ee263617adec1edb5429a2b84
21734c6d14e39bee.422318838d1ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc81631
[INFO|configuration_utils.py:648] 2023-07-31 17:45:57,371 >> loading configuration
file https://huggingface.co/distilgpt2/resolve/main/config.json from cache at /roo
t/.cache/huggingface/transformers/f985248d2791fcff97732e4ee263617adec1edb5429a2b84
21734c6d14e39bee.422318838d1ec4e061efb4ea29671cb2a044e244dc69229682bebd7cacc81631
[INFO|configuration_utils.py:684] 2023-07-31 17:45:57,372 >> Model config GPT2Conf
ig {
  "_name_or_path": "distilgpt2",
  "_num_labels": 1,
  "activation_function": "gelu_new",
  "architectures": [
    "GPT2LMHeadModel"
  ],
  "attn_pdrop": 0.1,
  "bos_token_id": 50256,
  "embd_pdrop": 0.1,
  "eos_token_id": 50256,
  "id2label": {
    "0": "LABEL_0"
  },
  "initializer_range": 0.02,
  "label2id": {
    "LABEL_0": 0
  },
  "layer_norm_epsilon": 1e-05,
  "model_type": "gpt2",
  "n_ctx": 1024,
  "n_embd": 768,
  "n_head": 12,
  "n_inner": null,
  "n_layer": 6,
  "n_positions": 1024,
  "reorder_and_upcast_attn": false,
  "resid_pdrop": 0.1,
  "scale_attn_by_inverse_layer_idx": false,
  "scale_attn_weights": true,
  "summary_activation": null,
  "summary_first_dropout": 0.1,
  "summary_proj_to_labels": true,
  "summary_type": "cls_index",
  "summary_use_proj": true,
  "task_specific_params": {
    "text-generation": {
      "do_sample": true,
      "max_length": 50
    }
  },
  "torch_dtype": "auto",
  "transformers_version": "4.17.0",
  "use_cache": true,
  "vocab_size": 50257
}
[INFO|configuration_utils.py:684] 2023-07-31 17:45:57,372 >> Model config GPT2Conf
ig {
  "_name_or_path": "distilgpt2",
  "_num_labels": 1,
  "activation_function": "gelu_new",
  "architectures": [
    "GPT2LMHeadModel"
  ],
  "attn_pdrop": 0.1,
  "bos_token_id": 50256,
  "embd_pdrop": 0.1,
  "eos_token_id": 50256,
  "id2label": {
    "0": "LABEL_0"
  },
  "initializer_range": 0.02,
  "label2id": {
    "LABEL_0": 0
  },
  "layer_norm_epsilon": 1e-05,
  "model_type": "gpt2",
  "n_ctx": 1024,
  "n_embd": 768,
  "n_head": 12,
  "n_inner": null,
  "n_layer": 6,
  "n_positions": 1024,
  "reorder_and_upcast_attn": false,
  "resid_pdrop": 0.1,
  "scale_attn_by_inverse_layer_idx": false,
  "scale_attn_weights": true,
  "summary_activation": null,
  "summary_first_dropout": 0.1,
  "summary_proj_to_labels": true,
  "summary_type": "cls_index",
  "summary_use_proj": true,
  "task_specific_params": {
    "text-generation": {
      "do_sample": true,
      "max_length": 50
    }
  },
  "torch_dtype": "auto",
  "transformers_version": "4.17.0",
  "use_cache": true,
  "vocab_size": 50257
}
[INFO|file_utils.py:2215] 2023-07-31 17:45:57,428 >> https://huggingface.co/distil
gpt2/resolve/main/pytorch_model.bin not found in cache or force_download set to Tr
ue, downloading to /root/.cache/huggingface/transformers/tmppvqyvacz
[INFO|file_utils.py:2215] 2023-07-31 17:45:57,428 >> https://huggingface.co/distil
gpt2/resolve/main/pytorch_model.bin not found in cache or force_download set to Tr
ue, downloading to /root/.cache/huggingface/transformers/tmppvqyvacz
Downloading:   0%|          | 0.00/336M [00:00<?, ?B/s]
Downloading:   2%|▏         | 5.50M/336M [00:00<00:06, 57.7MB/s]
Downloading:   4%|▎         | 11.8M/336M [00:00<00:05, 62.9MB/s]
Downloading:   5%|▌         | 18.2M/336M [00:00<00:05, 64.8MB/s]
Downloading:   7%|▋         | 24.6M/336M [00:00<00:04, 65.6MB/s]
Downloading:   9%|▉         | 30.9M/336M [00:00<00:04, 65.8MB/s]
Downloading:  11%|█         | 37.2M/336M [00:00<00:04, 65.9MB/s]
Downloading:  13%|█▎        | 43.5M/336M [00:00<00:04, 65.7MB/s]
```

```
Downloading:  15%|█▌        | 49.8M/336M [00:00<00:04, 65.7MB/s]
Downloading:  17%|█▋        | 56.2M/336M [00:00<00:04, 66.1MB/s]
Downloading:  19%|█▉        | 62.6M/336M [00:01<00:04, 66.3MB/s]
Downloading:  20%|██        | 69.0M/336M [00:01<00:04, 66.6MB/s]
Downloading:  22%|██▏       | 75.4M/336M [00:01<00:04, 66.8MB/s]
Downloading:  24%|██▍       | 81.8M/336M [00:01<00:03, 66.8MB/s]
Downloading:  26%|██▌       | 88.1M/336M [00:01<00:03, 66.8MB/s]
Downloading:  28%|██▊       | 94.6M/336M [00:01<00:03, 67.0MB/s]
Downloading:  30%|███       | 101M/336M [00:01<00:03, 65.4MB/s]
Downloading:  32%|███▏      | 107M/336M [00:01<00:03, 66.1MB/s]
Downloading:  34%|███▍      | 114M/336M [00:01<00:03, 66.5MB/s]
Downloading:  36%|███▌      | 120M/336M [00:01<00:03, 66.7MB/s]
Downloading:  38%|███▊      | 127M/336M [00:02<00:03, 66.7MB/s]
Downloading:  40%|███▉      | 133M/336M [00:02<00:03, 66.9MB/s]
Downloading:  41%|████▏     | 139M/336M [00:02<00:03, 67.1MB/s]
Downloading:  43%|████▎     | 146M/336M [00:02<00:02, 67.3MB/s]
Downloading:  45%|████▌     | 152M/336M [00:02<00:02, 67.3MB/s]
Downloading:  47%|████▋     | 159M/336M [00:02<00:02, 66.8MB/s]
Downloading:  49%|████▉     | 165M/336M [00:02<00:02, 66.8MB/s]
Downloading:  51%|█████     | 172M/336M [00:02<00:02, 63.7MB/s]
Downloading:  53%|█████▎    | 178M/336M [00:02<00:02, 64.6MB/s]
Downloading:  55%|█████▍    | 184M/336M [00:02<00:02, 65.5MB/s]
Downloading:  57%|█████▋    | 191M/336M [00:03<00:02, 64.0MB/s]
Downloading:  58%|█████▊    | 197M/336M [00:03<00:02, 64.3MB/s]
Downloading:  60%|██████    | 203M/336M [00:03<00:02, 62.9MB/s]
Downloading:  62%|██████▏   | 209M/336M [00:03<00:02, 62.2MB/s]
Downloading:  64%|██████▍   | 215M/336M [00:03<00:02, 62.4MB/s]
Downloading:  66%|██████▌   | 221M/336M [00:03<00:01, 62.2MB/s]
Downloading:  68%|██████▊   | 227M/336M [00:03<00:01, 63.5MB/s]
Downloading:  69%|██████▉   | 234M/336M [00:03<00:01, 64.6MB/s]
Downloading:  71%|███████▏  | 240M/336M [00:03<00:01, 65.5MB/s]
Downloading:  73%|███████▎  | 247M/336M [00:03<00:01, 65.9MB/s]
Downloading:  75%|███████▌  | 253M/336M [00:04<00:01, 66.2MB/s]
Downloading:  77%|███████▋  | 259M/336M [00:04<00:01, 66.5MB/s]
Downloading:  79%|███████▉  | 266M/336M [00:04<00:01, 66.5MB/s]
Downloading:  81%|████████  | 272M/336M [00:04<00:01, 66.7MB/s]
Downloading:  83%|████████▎ | 278M/336M [00:04<00:00, 66.7MB/s]
Downloading:  85%|████████▍ | 285M/336M [00:04<00:00, 65.0MB/s]
Downloading:  86%|████████▋ | 291M/336M [00:04<00:00, 64.8MB/s]
Downloading:  88%|████████▊ | 297M/336M [00:04<00:00, 65.4MB/s]
Downloading:  90%|█████████ | 304M/336M [00:04<00:00, 65.8MB/s]
Downloading:  92%|█████████▏| 310M/336M [00:04<00:00, 66.1MB/s]
Downloading:  94%|█████████▍| 317M/336M [00:05<00:00, 66.5MB/s]
Downloading:  96%|█████████▌| 323M/336M [00:05<00:00, 66.7MB/s]
Downloading:  98%|█████████▊| 329M/336M [00:05<00:00, 66.8MB/s]
Downloading: 100%|█████████▉| 336M/336M [00:05<00:00, 66.5MB/s]
Downloading: 100%|██████████| 336M/336M [00:05<00:00, 65.6MB/s]
[INFO|file_utils.py:2219] 2023-07-31 17:46:02,824 >> storing https://huggingface.c
o/distilgpt2/resolve/main/pytorch_model.bin in cache at /root/.cache/huggingface/t
ransformers/43a212e83e76bcb07f45be584cf100676bdbbbe9c13f9e5c1c050049143a832f.a83d8
81ec4d624fd4b5826dd026e315246c48c67504ff91c0500570e291a54ba
[INFO|file_utils.py:2219] 2023-07-31 17:46:02,824 >> storing https://huggingface.c
o/distilgpt2/resolve/main/pytorch_model.bin in cache at /root/.cache/huggingface/t
ransformers/43a212e83e76bcb07f45be584cf100676bdbbbe9c13f9e5c1c050049143a832f.a83d8
81ec4d624fd4b5826dd026e315246c48c67504ff91c0500570e291a54ba
[INFO|file_utils.py:2227] 2023-07-31 17:46:02,824 >> creating metadata file for /r
oot/.cache/huggingface/transformers/43a212e83e76bcb07f45be584cf100676bdbbbe9c13f9e
5c1c050049143a832f.a83d881ec4d624fd4b5826dd026e315246c48c67504ff91c0500570e291a54b
a
[INFO|file_utils.py:2227] 2023-07-31 17:46:02,824 >> creating metadata file for /r
oot/.cache/huggingface/transformers/43a212e83e76bcb07f45be584cf100676bdbbbe9c13f9e
5c1c050049143a832f.a83d881ec4d624fd4b5826dd026e315246c48c67504ff91c0500570e291a54b
a
[INFO|modeling_utils.py:1431] 2023-07-31 17:46:02,824 >> loading weights file http
```

```
s://huggingface.co/distilgpt2/resolve/main/pytorch_model.bin from cache at /root/.
cache/huggingface/transformers/43a212e83e76bcb07f45be584cf100676bdbbbe9c13f9e5c1c0
50049143a832f.a83d881ec4d624fd4b5826dd026e315246c48c67504ff91c0500570e291a54ba
[INFO|modeling_utils.py:1431] 2023-07-31 17:46:02,824 >> loading weights file http
s://huggingface.co/distilgpt2/resolve/main/pytorch_model.bin from cache at /root/.
cache/huggingface/transformers/43a212e83e76bcb07f45be584cf100676bdbbbe9c13f9e5c1c0
50049143a832f.a83d881ec4d624fd4b5826dd026e315246c48c67504ff91c0500570e291a54ba
[INFO|modeling_utils.py:563] 2023-07-31 17:46:03,032 >> Instantiating GPT2LMHeadMo
del model under default dtype torch.float32.
[INFO|modeling_utils.py:563] 2023-07-31 17:46:03,032 >> Instantiating GPT2LMHeadMo
del model under default dtype torch.float32.
[INFO|modeling_utils.py:1702] 2023-07-31 17:46:04,318 >> All model checkpoint weig
hts were used when initializing GPT2LMHeadModel.
[INFO|modeling_utils.py:1710] 2023-07-31 17:46:04,319 >> All the weights of GPT2LM
HeadModel were initialized from the model checkpoint at distilgpt2.
If your task is similar to the task the model of the checkpoint was trained on, yo
u can already use GPT2LMHeadModel for predictions without further training.
[INFO|modeling_utils.py:1702] 2023-07-31 17:46:04,318 >> All model checkpoint weig
hts were used when initializing GPT2LMHeadModel.
[INFO|modeling_utils.py:1710] 2023-07-31 17:46:04,319 >> All the weights of GPT2LM
HeadModel were initialized from the model checkpoint at distilgpt2.
If your task is similar to the task the model of the checkpoint was trained on, yo
u can already use GPT2LMHeadModel for predictions without further training.
Running tokenizer on dataset:   0%|          | 0/1 [00:00<?, ? examples/s]
[WARNING|tokenization_utils_base.py:3397] 2023-07-31 17:46:05,394 >> Token indices
sequence length is longer than the specified maximum sequence length for this mode
l (301966 > 1024). Running this sequence through the model will result in indexing
errors
[WARNING|tokenization_utils_base.py:3397] 2023-07-31 17:46:05,394 >> Token indices
sequence length is longer than the specified maximum sequence length for this mode
l (301966 > 1024). Running this sequence through the model will result in indexing
errors
[WARNING|finetuning.py:167] 2023-07-31 17:46:05,394 >> ^^^^^^^^^^^^^^^^ Please ign
ore the warning above - this long input will be chunked into smaller bits before b
eing passed to the model.
[WARNING|finetuning.py:167] 2023-07-31 17:46:05,394 >> ^^^^^^^^^^^^^^^^ Please ign
ore the warning above - this long input will be chunked into smaller bits before b
eing passed to the model.
07/31/2023 17:46:05 - INFO - datasets.arrow_dataset - Caching processed dataset at
/root/.cache/huggingface/datasets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8707
337f6cb296314fbe06960bd9a868dca39e713e163d27b5e/cache-2e39ed04bebe90d6.arrow
Running tokenizer on dataset: 100%|██████████| 1/1 [00:01<00:00,  1.13s/ examples]
Running tokenizer on dataset:   0%|          | 0/1 [00:00<?, ? examples/s]
07/31/2023 17:46:05 - INFO - datasets.arrow_dataset - Caching processed dataset at
/root/.cache/huggingface/datasets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8707
337f6cb296314fbe06960bd9a868dca39e713e163d27b5e/cache-3e04e2dbbb42e49e.arrow
Running tokenizer on dataset:   0%|          | 0/1 [00:00<?, ? examples/s]
07/31/2023 17:46:05 - INFO - datasets.arrow_dataset - Caching processed dataset at
/root/.cache/huggingface/datasets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8707
337f6cb296314fbe06960bd9a868dca39e713e163d27b5e/cache-d53a0af4fa4fa646.arrow
Grouping texts in chunks of 1024:   0%|          | 0/1 [00:00<?, ? examples/s]
07/31/2023 17:46:05 - INFO - datasets.arrow_dataset - Caching processed dataset at
/root/.cache/huggingface/datasets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8707
337f6cb296314fbe06960bd9a868dca39e713e163d27b5e/cache-a754c241ca38040a.arrow
Grouping texts in chunks of 1024: 100%|██████████| 1/1 [00:00<00:00,  2.79 example
s/s]
Grouping texts in chunks of 1024:   0%|          | 0/1 [00:00<?, ? examples/s]
07/31/2023 17:46:06 - INFO - datasets.arrow_dataset - Caching processed dataset at
/root/.cache/huggingface/datasets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8707
337f6cb296314fbe06960bd9a868dca39e713e163d27b5e/cache-b7e8b76fe95aaa0f.arrow
Grouping texts in chunks of 1024:   0%|          | 0/1 [00:00<?, ? examples/s]
07/31/2023 17:46:06 - INFO - datasets.arrow_dataset - Caching processed dataset at
/root/.cache/huggingface/datasets/tiny_shakespeare/default/1.0.0/b5b13969f09fe8707
337f6cb296314fbe06960bd9a868dca39e713e163d27b5e/cache-52ce19414de2ed9b.arrow
```

```
Downloading builder script:   0%|          | 0.00/4.20k [00:00<?, ?B/s]
Downloading builder script: 100%|██████████| 4.20k/4.20k [00:00<00:00, 4.77MB/s]
[INFO|training_args.py:1009] 2023-07-31 17:46:06,233 >> PyTorch: setting up device
s
[INFO|training_args.py:1009] 2023-07-31 17:46:06,233 >> PyTorch: setting up device
s
[INFO|training_args.py:871] 2023-07-31 17:46:06,234 >> The default value for the t
raining argument `--report_to` will change in v5 (from all installed integrations
to none). In v5, you will need to use `--report_to all` to get the same behavior a
s now. You should start updating your code and make this info disappear :-).
[INFO|training_args.py:871] 2023-07-31 17:46:06,234 >> The default value for the t
raining argument `--report_to` will change in v5 (from all installed integrations
to none). In v5, you will need to use `--report_to all` to get the same behavior a
s now. You should start updating your code and make this info disappear :-).
/opt/conda/lib/python3.8/site-packages/transformers/optimization.py:306: FutureWar
ning: This implementation of AdamW is deprecated and will be removed in a future v
ersion. Use the PyTorch implementation torch.optim.AdamW instead, or set `no_depre
cation_warning=True` to disable this warning
  warnings.warn(
[INFO|trainer.py:1279] 2023-07-31 17:46:11,121 >> ***** Running training *****
[INFO|trainer.py:1279] 2023-07-31 17:46:11,121 >> ***** Running training *****
[INFO|trainer.py:1280] 2023-07-31 17:46:11,121 >>   Num examples = 294
[INFO|trainer.py:1281] 2023-07-31 17:46:11,121 >>   Num Epochs = 3
[INFO|trainer.py:1282] 2023-07-31 17:46:11,121 >>   Instantaneous batch size per d
evice = 2
[INFO|trainer.py:1280] 2023-07-31 17:46:11,121 >>   Num examples = 294
[INFO|trainer.py:1281] 2023-07-31 17:46:11,121 >>   Num Epochs = 3
[INFO|trainer.py:1282] 2023-07-31 17:46:11,121 >>   Instantaneous batch size per d
evice = 2
[INFO|trainer.py:1283] 2023-07-31 17:46:11,121 >>   Total train batch size (w. par
allel, distributed & accumulation) = 2
[INFO|trainer.py:1284] 2023-07-31 17:46:11,121 >>   Gradient Accumulation steps =
1
[INFO|trainer.py:1285] 2023-07-31 17:46:11,121 >>   Total optimization steps = 441
[INFO|trainer.py:1283] 2023-07-31 17:46:11,121 >>   Total train batch size (w. par
allel, distributed & accumulation) = 2
[INFO|trainer.py:1284] 2023-07-31 17:46:11,121 >>   Gradient Accumulation steps =
1
[INFO|trainer.py:1285] 2023-07-31 17:46:11,121 >>   Total optimization steps = 441
0%|          | 0/441 [00:00<?, ?it/s]
[2023-07-31 17:46:11.316 algo-1:51 INFO utils.py:27] RULE_JOB_STOP_SIGNAL_FILENAM
E: None
[2023-07-31 17:46:11.484 algo-1:51 INFO profiler_config_parser.py:111] User has di
sabled profiler.
[2023-07-31 17:46:11.486 algo-1:51 INFO json_config.py:91] Creating hook from json
_config at /opt/ml/input/config/debughookconfig.json.
[2023-07-31 17:46:11.486 algo-1:51 INFO hook.py:201] tensorboard_dir has not been
set for the hook. SMDebug will not be exporting tensorboard summaries.
[2023-07-31 17:46:11.487 algo-1:51 INFO hook.py:254] Saving to /opt/ml/output/tens
ors
[2023-07-31 17:46:11.487 algo-1:51 INFO state_store.py:77] The checkpoint config f
ile /opt/ml/input/config/checkpointconfig.json does not exist.
0%|          | 1/441 [00:01<10:38,  1.45s/it]
0%|          | 2/441 [00:01<04:58,  1.47it/s]
1%|          | 3/441 [00:01<03:09,  2.31it/s]
1%|          | 4/441 [00:01<02:18,  3.15it/s]
1%|          | 5/441 [00:02<01:50,  3.94it/s]
1%|▏         | 6/441 [00:02<01:33,  4.65it/s]
2%|▏         | 7/441 [00:02<01:22,  5.25it/s]
2%|▏         | 8/441 [00:02<01:15,  5.73it/s]
2%|▏         | 9/441 [00:02<01:10,  6.11it/s]
2%|▏         | 10/441 [00:02<01:07,  6.37it/s]
2%|▏         | 11/441 [00:02<01:05,  6.59it/s]
3%|▎         | 12/441 [00:02<01:03,  6.72it/s]
```

```
  3%|▍         | 13/441 [00:03<01:02,  6.81it/s]
  3%|▍         | 14/441 [00:03<01:01,  6.89it/s]
  3%|▍         | 15/441 [00:03<01:01,  6.95it/s]
  4%|▍         | 16/441 [00:03<01:00,  7.02it/s]
  4%|▌         | 17/441 [00:03<01:00,  7.02it/s]
  4%|▌         | 18/441 [00:03<00:59,  7.05it/s]
  4%|▌         | 19/441 [00:03<00:59,  7.05it/s]
  5%|▌         | 20/441 [00:04<00:59,  7.04it/s]
  5%|▌         | 21/441 [00:04<00:59,  7.03it/s]
  5%|▌         | 22/441 [00:04<00:59,  7.06it/s]
  5%|▎        | 23/441 [00:04<00:58,  7.09it/s]
  5%|▎        | 24/441 [00:04<00:58,  7.10it/s]
  6%|▎        | 25/441 [00:04<00:58,  7.07it/s]
  6%|▎        | 26/441 [00:04<00:58,  7.05it/s]
  6%|▎        | 27/441 [00:05<00:58,  7.04it/s]
  6%|▋         | 28/441 [00:05<00:58,  7.03it/s]
  7%|▋         | 29/441 [00:05<00:58,  7.04it/s]
  7%|▋         | 30/441 [00:05<00:58,  7.06it/s]
  7%|▋         | 31/441 [00:05<00:58,  7.06it/s]
  7%|▋         | 32/441 [00:05<00:58,  7.05it/s]
  7%|▋         | 33/441 [00:05<00:58,  7.02it/s]
  8%|▊         | 34/441 [00:06<00:57,  7.05it/s]
  8%|▊         | 35/441 [00:06<00:57,  7.08it/s]
  8%|▊         | 36/441 [00:06<00:57,  7.09it/s]
  8%|▊         | 37/441 [00:06<00:56,  7.09it/s]
  9%|▊         | 38/441 [00:06<00:56,  7.10it/s]
  9%|▉         | 39/441 [00:06<00:56,  7.10it/s]
  9%|▉         | 40/441 [00:06<00:56,  7.10it/s]
  9%|▉         | 41/441 [00:07<00:56,  7.10it/s]
 10%|█         | 42/441 [00:07<00:56,  7.10it/s]
 10%|█         | 43/441 [00:07<00:56,  7.11it/s]
 10%|█         | 44/441 [00:07<00:55,  7.11it/s]
 10%|▌        | 45/441 [00:07<00:56,  7.07it/s]
 10%|▌        | 46/441 [00:07<00:55,  7.07it/s]
 11%|▌        | 47/441 [00:07<00:55,  7.08it/s]
 11%|▌        | 48/441 [00:08<00:55,  7.09it/s]
 11%|▌        | 49/441 [00:08<00:55,  7.06it/s]
 11%|█▏        | 50/441 [00:08<00:55,  7.07it/s]
 12%|█▏        | 51/441 [00:08<00:55,  7.08it/s]
 12%|█▏        | 52/441 [00:08<00:54,  7.08it/s]
 12%|█▏        | 53/441 [00:08<00:54,  7.10it/s]
 12%|█▏        | 54/441 [00:08<00:54,  7.11it/s]
 12%|█▏        | 55/441 [00:09<00:54,  7.10it/s]
 13%|█▎        | 56/441 [00:09<00:54,  7.10it/s]
 13%|█▎        | 57/441 [00:09<00:54,  7.06it/s]
 13%|█▎        | 58/441 [00:09<00:54,  7.07it/s]
 13%|█▎        | 59/441 [00:09<00:53,  7.09it/s]
 14%|█▎        | 60/441 [00:09<00:53,  7.10it/s]
 14%|█▍        | 61/441 [00:09<00:53,  7.10it/s]
 14%|█▍        | 62/441 [00:10<00:53,  7.10it/s]
 14%|█▍        | 63/441 [00:10<00:53,  7.06it/s]
 15%|█▍        | 64/441 [00:10<00:53,  7.05it/s]
 15%|█▍        | 65/441 [00:10<00:53,  7.03it/s]
 15%|█▍        | 66/441 [00:10<00:53,  7.06it/s]
 15%|▉        | 67/441 [00:10<00:53,  7.04it/s]
 15%|▉        | 68/441 [00:10<00:53,  7.01it/s]
 16%|▉        | 69/441 [00:11<00:53,  7.02it/s]
 16%|▉        | 70/441 [00:11<00:52,  7.01it/s]
 16%|▉        | 71/441 [00:11<00:52,  7.04it/s]
 16%|█▋        | 72/441 [00:11<00:52,  7.07it/s]
 17%|█▋        | 73/441 [00:11<00:52,  7.08it/s]
 17%|█▋        | 74/441 [00:11<00:51,  7.06it/s]
 17%|█▋        | 75/441 [00:11<00:51,  7.07it/s]
 17%|█▋        | 76/441 [00:12<00:51,  7.07it/s]
```

```
17%|███▍                | 77/441 [00:12<00:51,  7.04it/s]
18%|███▌                | 78/441 [00:12<00:51,  7.04it/s]
18%|███▌                | 79/441 [00:12<00:51,  7.06it/s]
18%|███▋                | 80/441 [00:12<00:51,  7.08it/s]
18%|███▋                | 81/441 [00:12<00:50,  7.08it/s]
19%|███▋                | 82/441 [00:12<00:50,  7.09it/s]
19%|███▊                | 83/441 [00:13<00:50,  7.10it/s]
19%|███▊                | 84/441 [00:13<00:50,  7.11it/s]
19%|███▊                | 85/441 [00:13<00:49,  7.13it/s]
20%|███▉                | 86/441 [00:13<00:49,  7.13it/s]
20%|███▉                | 87/441 [00:13<00:49,  7.13it/s]
20%|███▉                | 88/441 [00:13<00:49,  7.14it/s]
20%|████                | 89/441 [00:13<00:49,  7.10it/s]
20%|████                | 90/441 [00:14<00:49,  7.08it/s]
21%|████▏               | 91/441 [00:14<00:49,  7.10it/s]
21%|████▏               | 92/441 [00:14<00:49,  7.08it/s]
21%|████▏               | 93/441 [00:14<00:48,  7.11it/s]
21%|████▎               | 94/441 [00:14<00:48,  7.12it/s]
22%|████▎               | 95/441 [00:14<00:48,  7.14it/s]
22%|████▎               | 96/441 [00:14<00:48,  7.14it/s]
22%|████▍               | 97/441 [00:15<00:48,  7.14it/s]
22%|████▍               | 98/441 [00:15<00:48,  7.14it/s]
22%|████▍               | 99/441 [00:15<00:47,  7.15it/s]
23%|████▌               | 100/441 [00:15<00:47,  7.15it/s]
23%|████▌               | 101/441 [00:15<00:47,  7.11it/s]
23%|████▋               | 102/441 [00:15<00:47,  7.12it/s]
23%|████▋               | 103/441 [00:15<00:47,  7.13it/s]
24%|████▋               | 104/441 [00:15<00:47,  7.10it/s]
24%|████▊               | 105/441 [00:16<00:47,  7.10it/s]
24%|████▊               | 106/441 [00:16<00:47,  7.11it/s]
24%|████▊               | 107/441 [00:16<00:46,  7.12it/s]
24%|████▉               | 108/441 [00:16<00:46,  7.13it/s]
25%|████▉               | 109/441 [00:16<00:46,  7.13it/s]
25%|████▉               | 110/441 [00:16<00:46,  7.14it/s]
25%|█████               | 111/441 [00:16<00:46,  7.13it/s]
25%|█████               | 112/441 [00:17<00:46,  7.10it/s]
26%|█████               | 113/441 [00:17<00:46,  7.07it/s]
26%|█████▏              | 114/441 [00:17<00:46,  7.08it/s]
26%|█████▏              | 115/441 [00:17<00:45,  7.10it/s]
26%|█████▎              | 116/441 [00:17<00:45,  7.12it/s]
27%|█████▎              | 117/441 [00:17<00:45,  7.13it/s]
27%|█████▎              | 118/441 [00:17<00:45,  7.14it/s]
27%|█████▍              | 119/441 [00:18<00:45,  7.13it/s]
27%|█████▍              | 120/441 [00:18<00:45,  7.12it/s]
27%|█████▍              | 121/441 [00:18<00:44,  7.12it/s]
28%|█████▌              | 122/441 [00:18<00:44,  7.12it/s]
28%|█████▌              | 123/441 [00:18<00:44,  7.13it/s]
28%|█████▌              | 124/441 [00:18<00:44,  7.13it/s]
28%|█████▋              | 125/441 [00:18<00:44,  7.13it/s]
29%|█████▋              | 126/441 [00:19<00:44,  7.12it/s]
29%|█████▊              | 127/441 [00:19<00:44,  7.12it/s]
29%|█████▊              | 128/441 [00:19<00:43,  7.12it/s]
29%|█████▊              | 129/441 [00:19<00:43,  7.12it/s]
29%|█████▉              | 130/441 [00:19<00:43,  7.12it/s]
30%|█████▉              | 131/441 [00:19<00:43,  7.09it/s]
30%|█████▉              | 132/441 [00:19<00:43,  7.10it/s]
30%|██████              | 133/441 [00:20<00:43,  7.05it/s]
30%|██████              | 134/441 [00:20<00:43,  7.02it/s]
31%|██████              | 135/441 [00:20<00:43,  7.04it/s]
31%|██████▏             | 136/441 [00:20<00:43,  7.04it/s]
31%|██████▏             | 137/441 [00:20<00:43,  7.03it/s]
31%|██████▎             | 138/441 [00:20<00:43,  7.01it/s]
32%|██████▎             | 139/441 [00:20<00:43,  7.00it/s]
32%|██████▎             | 140/441 [00:21<00:42,  7.01it/s]
```

```
32%|██████▍             | 141/441 [00:21<00:42,  7.00it/s]
32%|██████▍             | 142/441 [00:21<00:42,  7.03it/s]
32%|██████▍             | 143/441 [00:21<00:42,  7.04it/s]
33%|██████▌             | 144/441 [00:21<00:42,  7.05it/s]
33%|██████▌             | 145/441 [00:21<00:41,  7.06it/s]
33%|██████▌             | 146/441 [00:21<00:41,  7.03it/s]
33%|██████▋             | 147/441 [00:22<00:41,  7.00it/s]
34%|██████▋             | 148/441 [00:22<00:41,  7.01it/s]
34%|██████▊             | 149/441 [00:22<00:41,  7.03it/s]
34%|██████▊             | 150/441 [00:22<00:41,  7.04it/s]
34%|██████▊             | 151/441 [00:22<00:41,  7.06it/s]
34%|██████▉             | 152/441 [00:22<00:41,  7.04it/s]
35%|██████▉             | 153/441 [00:22<00:41,  7.02it/s]
35%|██████▉             | 154/441 [00:23<00:40,  7.04it/s]
35%|███████             | 155/441 [00:23<00:40,  7.04it/s]
35%|███████             | 156/441 [00:23<00:40,  7.01it/s]
36%|███████             | 157/441 [00:23<00:40,  6.99it/s]
36%|███████▏            | 158/441 [00:23<00:40,  7.00it/s]
36%|███████▏            | 159/441 [00:23<00:40,  6.99it/s]
36%|███████▎            | 160/441 [00:23<00:40,  6.97it/s]
37%|███████▎            | 161/441 [00:24<00:40,  6.98it/s]
37%|███████▎            | 162/441 [00:24<00:40,  6.97it/s]
37%|███████▍            | 163/441 [00:24<00:39,  7.01it/s]
37%|███████▍            | 164/441 [00:24<00:39,  7.01it/s]
37%|███████▍            | 165/441 [00:24<00:39,  7.00it/s]
38%|███████▌            | 166/441 [00:24<00:39,  7.01it/s]
38%|███████▌            | 167/441 [00:24<00:38,  7.03it/s]
38%|███████▌            | 168/441 [00:25<00:38,  7.06it/s]
38%|███████▋            | 169/441 [00:25<00:38,  7.07it/s]
39%|███████▋            | 170/441 [00:25<00:38,  7.08it/s]
39%|███████▊            | 171/441 [00:25<00:38,  7.09it/s]
39%|███████▊            | 172/441 [00:25<00:37,  7.10it/s]
39%|███████▊            | 173/441 [00:25<00:37,  7.07it/s]
39%|███████▉            | 174/441 [00:25<00:37,  7.08it/s]
40%|███████▉            | 175/441 [00:26<00:37,  7.04it/s]
40%|███████▉            | 176/441 [00:26<00:37,  7.01it/s]
40%|████████            | 177/441 [00:26<00:37,  7.01it/s]
40%|████████            | 178/441 [00:26<00:37,  6.99it/s]
41%|████████            | 179/441 [00:26<00:37,  7.03it/s]
41%|████████▏           | 180/441 [00:26<00:37,  7.02it/s]
41%|████████▏           | 181/441 [00:26<00:36,  7.05it/s]
41%|████████▎           | 182/441 [00:27<00:36,  7.03it/s]
41%|████████▎           | 183/441 [00:27<00:36,  7.04it/s]
42%|████████▎           | 184/441 [00:27<00:36,  7.01it/s]
42%|████████▍           | 185/441 [00:27<00:36,  6.98it/s]
42%|████████▍           | 186/441 [00:27<00:36,  6.98it/s]
42%|████████▍           | 187/441 [00:27<00:36,  6.97it/s]
43%|████████▌           | 188/441 [00:27<00:36,  6.97it/s]
43%|████████▌           | 189/441 [00:28<00:36,  6.97it/s]
43%|████████▌           | 190/441 [00:28<00:36,  6.97it/s]
43%|████████▋           | 191/441 [00:28<00:35,  6.96it/s]
44%|████████▋           | 192/441 [00:28<00:35,  6.97it/s]
44%|████████▊           | 193/441 [00:28<00:35,  6.97it/s]
44%|████████▊           | 194/441 [00:28<00:35,  7.00it/s]
44%|████████▊           | 195/441 [00:28<00:35,  6.99it/s]
44%|████████▉           | 196/441 [00:29<00:35,  6.98it/s]
45%|████████▉           | 197/441 [00:29<00:34,  6.98it/s]
45%|████████▉           | 198/441 [00:29<00:34,  6.97it/s]
45%|█████████           | 199/441 [00:29<00:34,  6.97it/s]
45%|█████████           | 200/441 [00:29<00:34,  7.00it/s]
46%|█████████           | 201/441 [00:29<00:34,  7.00it/s]
46%|█████████▏          | 202/441 [00:29<00:34,  7.02it/s]
46%|█████████▏          | 203/441 [00:30<00:33,  7.00it/s]
46%|█████████▎          | 204/441 [00:30<00:33,  6.98it/s]
```

```
 46%|████▋     | 205/441 [00:30<00:33,  6.97it/s]
 47%|████▋     | 206/441 [00:30<00:33,  7.00it/s]
 47%|████▋     | 207/441 [00:30<00:33,  7.02it/s]
 47%|████▋     | 208/441 [00:30<00:33,  7.00it/s]
 47%|████▋     | 209/441 [00:30<00:33,  6.99it/s]
 48%|████▊     | 210/441 [00:31<00:33,  6.99it/s]
 48%|████▊     | 211/441 [00:31<00:32,  7.03it/s]
 48%|████▊     | 212/441 [00:31<00:32,  7.02it/s]
 48%|████▊     | 213/441 [00:31<00:32,  7.00it/s]
 49%|████▊     | 214/441 [00:31<00:32,  7.02it/s]
 49%|████▉     | 215/441 [00:31<00:32,  7.05it/s]
 49%|████▉     | 216/441 [00:31<00:31,  7.07it/s]
 49%|████▉     | 217/441 [00:32<00:31,  7.07it/s]
 49%|████▉     | 218/441 [00:32<00:31,  7.08it/s]
 50%|████▉     | 219/441 [00:32<00:31,  7.10it/s]
 50%|████▉     | 220/441 [00:32<00:31,  7.07it/s]
 50%|█████     | 221/441 [00:32<00:31,  7.08it/s]
 50%|█████     | 222/441 [00:32<00:30,  7.10it/s]
 51%|█████     | 223/441 [00:32<00:30,  7.09it/s]
 51%|█████     | 224/441 [00:33<00:30,  7.09it/s]
 51%|█████     | 225/441 [00:33<00:30,  7.09it/s]
 51%|█████     | 226/441 [00:33<00:30,  7.06it/s]
 51%|█████▏    | 227/441 [00:33<00:30,  7.07it/s]
 52%|█████▏    | 228/441 [00:33<00:30,  7.08it/s]
 52%|█████▏    | 229/441 [00:33<00:29,  7.09it/s]
 52%|█████▏    | 230/441 [00:33<00:29,  7.10it/s]
 52%|█████▏    | 231/441 [00:34<00:29,  7.08it/s]
 53%|█████▎    | 232/441 [00:34<00:29,  7.07it/s]
 53%|█████▎    | 233/441 [00:34<00:29,  7.04it/s]
 53%|█████▎    | 234/441 [00:34<00:29,  7.01it/s]
 53%|█████▎    | 235/441 [00:34<00:29,  7.03it/s]
 54%|█████▎    | 236/441 [00:34<00:29,  7.06it/s]
 54%|█████▎    | 237/441 [00:34<00:28,  7.07it/s]
 54%|█████▍    | 238/441 [00:35<00:28,  7.05it/s]
 54%|█████▍    | 239/441 [00:35<00:28,  7.05it/s]
 54%|█████▍    | 240/441 [00:35<00:28,  7.06it/s]
 55%|█████▍    | 241/441 [00:35<00:28,  7.04it/s]
 55%|█████▍    | 242/441 [00:35<00:28,  7.02it/s]
 55%|█████▌    | 243/441 [00:35<00:28,  7.04it/s]
 55%|█████▌    | 244/441 [00:35<00:28,  7.03it/s]
 56%|█████▌    | 245/441 [00:36<00:27,  7.04it/s]
 56%|█████▌    | 246/441 [00:36<00:27,  7.05it/s]
 56%|█████▌    | 247/441 [00:36<00:27,  7.07it/s]
 56%|█████▌    | 248/441 [00:36<00:27,  7.06it/s]
 56%|█████▋    | 249/441 [00:36<00:27,  7.02it/s]
 57%|█████▋    | 250/441 [00:36<00:27,  7.03it/s]
 57%|█████▋    | 251/441 [00:36<00:26,  7.04it/s]
 57%|█████▋    | 252/441 [00:36<00:26,  7.05it/s]
 57%|█████▋    | 253/441 [00:37<00:26,  7.05it/s]
 58%|█████▊    | 254/441 [00:37<00:26,  7.06it/s]
 58%|█████▊    | 255/441 [00:37<00:26,  7.06it/s]
 58%|█████▊    | 256/441 [00:37<00:26,  7.02it/s]
 58%|█████▊    | 257/441 [00:37<00:26,  6.99it/s]
 59%|█████▊    | 258/441 [00:37<00:26,  6.97it/s]
 59%|█████▊    | 259/441 [00:37<00:25,  7.01it/s]
 59%|█████▉    | 260/441 [00:38<00:25,  7.03it/s]
 59%|█████▉    | 261/441 [00:38<00:25,  7.04it/s]
 59%|█████▉    | 262/441 [00:38<00:25,  7.05it/s]
 60%|█████▉    | 263/441 [00:38<00:25,  7.07it/s]
 60%|█████▉    | 264/441 [00:38<00:25,  7.08it/s]
 60%|██████    | 265/441 [00:38<00:24,  7.08it/s]
 60%|██████    | 266/441 [00:38<00:24,  7.08it/s]
 61%|██████    | 267/441 [00:39<00:24,  7.09it/s]
 61%|██████    | 268/441 [00:39<00:24,  7.08it/s]
```

```
61%|██████    | 269/441 [00:39<00:24,  7.08it/s]
61%|██████    | 270/441 [00:39<00:24,  7.09it/s]
61%|██████▏   | 271/441 [00:39<00:23,  7.09it/s]
62%|██████▏   | 272/441 [00:39<00:23,  7.06it/s]
62%|██████▏   | 273/441 [00:39<00:23,  7.07it/s]
62%|██████▏   | 274/441 [00:40<00:23,  7.08it/s]
62%|██████▏   | 275/441 [00:40<00:23,  7.09it/s]
63%|██████▎   | 276/441 [00:40<00:23,  7.07it/s]
63%|██████▎   | 277/441 [00:40<00:23,  7.02it/s]
63%|██████▎   | 278/441 [00:40<00:23,  7.04it/s]
63%|██████▎   | 279/441 [00:40<00:22,  7.05it/s]
63%|██████▎   | 280/441 [00:40<00:22,  7.05it/s]
64%|██████▎   | 281/441 [00:41<00:22,  7.07it/s]
64%|██████▍   | 282/441 [00:41<00:22,  7.08it/s]
64%|██████▍   | 283/441 [00:41<00:22,  7.05it/s]
64%|██████▍   | 284/441 [00:41<00:22,  7.06it/s]
65%|██████▍   | 285/441 [00:41<00:22,  7.07it/s]
65%|██████▍   | 286/441 [00:41<00:21,  7.07it/s]
65%|██████▌   | 287/441 [00:41<00:21,  7.08it/s]
65%|██████▌   | 288/441 [00:42<00:21,  7.08it/s]
66%|██████▌   | 289/441 [00:42<00:21,  7.09it/s]
66%|██████▌   | 290/441 [00:42<00:21,  7.10it/s]
66%|██████▌   | 291/441 [00:42<00:21,  7.10it/s]
66%|██████▌   | 292/441 [00:42<00:20,  7.11it/s]
66%|██████▋   | 293/441 [00:42<00:20,  7.12it/s]
67%|██████▋   | 294/441 [00:42<00:20,  7.12it/s]
67%|██████▋   | 295/441 [00:43<00:20,  7.11it/s]
67%|██████▋   | 296/441 [00:43<00:20,  7.11it/s]
67%|██████▋   | 297/441 [00:43<00:20,  7.10it/s]
68%|██████▊   | 298/441 [00:43<00:20,  7.06it/s]
68%|██████▊   | 299/441 [00:43<00:20,  7.08it/s]
68%|██████▊   | 300/441 [00:43<00:19,  7.09it/s]
68%|██████▊   | 301/441 [00:43<00:19,  7.09it/s]
68%|██████▊   | 302/441 [00:44<00:19,  7.07it/s]
69%|██████▊   | 303/441 [00:44<00:19,  7.08it/s]
69%|██████▉   | 304/441 [00:44<00:19,  7.08it/s]
69%|██████▉   | 305/441 [00:44<00:19,  7.08it/s]
69%|██████▉   | 306/441 [00:44<00:19,  7.08it/s]
70%|██████▉   | 307/441 [00:44<00:18,  7.08it/s]
70%|██████▉   | 308/441 [00:44<00:18,  7.07it/s]
70%|███████   | 309/441 [00:45<00:18,  7.04it/s]
70%|███████   | 310/441 [00:45<00:18,  7.02it/s]
71%|███████   | 311/441 [00:45<00:18,  7.01it/s]
71%|███████   | 312/441 [00:45<00:18,  7.00it/s]
71%|███████   | 313/441 [00:45<00:18,  7.02it/s]
71%|███████   | 314/441 [00:45<00:18,  7.04it/s]
71%|███████▏  | 315/441 [00:45<00:17,  7.06it/s]
72%|███████▏  | 316/441 [00:46<00:17,  7.07it/s]
72%|███████▏  | 317/441 [00:46<00:17,  7.08it/s]
72%|███████▏  | 318/441 [00:46<00:17,  7.06it/s]
72%|███████▏  | 319/441 [00:46<00:17,  7.07it/s]
73%|███████▎  | 320/441 [00:46<00:17,  7.08it/s]
73%|███████▎  | 321/441 [00:46<00:16,  7.09it/s]
73%|███████▎  | 322/441 [00:46<00:16,  7.07it/s]
73%|███████▎  | 323/441 [00:47<00:16,  7.08it/s]
73%|███████▎  | 324/441 [00:47<00:16,  7.06it/s]
74%|███████▎  | 325/441 [00:47<00:16,  7.05it/s]
74%|███████▍  | 326/441 [00:47<00:16,  7.06it/s]
74%|███████▍  | 327/441 [00:47<00:16,  7.05it/s]
74%|███████▍  | 328/441 [00:47<00:16,  7.03it/s]
75%|███████▍  | 329/441 [00:47<00:15,  7.04it/s]
75%|███████▍  | 330/441 [00:48<00:15,  7.05it/s]
75%|███████▌  | 331/441 [00:48<00:15,  7.02it/s]
75%|███████▌  | 332/441 [00:48<00:15,  7.04it/s]
```

```
76%|███████▌  | 333/441 [00:48<00:15,  7.05it/s]
76%|███████▌  | 334/441 [00:48<00:15,  7.04it/s]
76%|███████▌  | 335/441 [00:48<00:14,  7.07it/s]
76%|███████▌  | 336/441 [00:48<00:14,  7.05it/s]
76%|███████▋  | 337/441 [00:49<00:14,  7.06it/s]
77%|███████▋  | 338/441 [00:49<00:14,  7.08it/s]
77%|███████▋  | 339/441 [00:49<00:14,  7.06it/s]
77%|███████▋  | 340/441 [00:49<00:14,  7.07it/s]
77%|███████▋  | 341/441 [00:49<00:14,  7.08it/s]
78%|███████▊  | 342/441 [00:49<00:13,  7.09it/s]
78%|███████▊  | 343/441 [00:49<00:13,  7.10it/s]
78%|███████▊  | 344/441 [00:50<00:13,  7.10it/s]
78%|███████▊  | 345/441 [00:50<00:13,  7.08it/s]
78%|███████▊  | 346/441 [00:50<00:13,  7.06it/s]
79%|███████▊  | 347/441 [00:50<00:13,  7.06it/s]
79%|███████▉  | 348/441 [00:50<00:13,  7.09it/s]
79%|███████▉  | 349/441 [00:50<00:12,  7.09it/s]
79%|███████▉  | 350/441 [00:50<00:12,  7.05it/s]
80%|███████▉  | 351/441 [00:51<00:12,  7.06it/s]
80%|███████▉  | 352/441 [00:51<00:12,  7.06it/s]
80%|████████  | 353/441 [00:51<00:12,  7.06it/s]
80%|████████  | 354/441 [00:51<00:12,  7.05it/s]
80%|████████  | 355/441 [00:51<00:12,  7.06it/s]
81%|████████  | 356/441 [00:51<00:12,  7.06it/s]
81%|████████  | 357/441 [00:51<00:12,  6.96it/s]
81%|████████  | 358/441 [00:52<00:11,  6.98it/s]
81%|████████▏ | 359/441 [00:52<00:11,  6.95it/s]
82%|████████▏ | 360/441 [00:52<00:11,  6.96it/s]
82%|████████▏ | 361/441 [00:52<00:11,  6.95it/s]
82%|████████▏ | 362/441 [00:52<00:11,  6.94it/s]
82%|████████▏ | 363/441 [00:52<00:11,  6.95it/s]
83%|████████▎ | 364/441 [00:52<00:11,  6.98it/s]
83%|████████▎ | 365/441 [00:53<00:10,  7.00it/s]
83%|████████▎ | 366/441 [00:53<00:10,  7.02it/s]
83%|████████▎ | 367/441 [00:53<00:10,  7.04it/s]
83%|████████▎ | 368/441 [00:53<00:10,  7.06it/s]
84%|████████▎ | 369/441 [00:53<00:10,  7.07it/s]
84%|████████▍ | 370/441 [00:53<00:10,  7.09it/s]
84%|████████▍ | 371/441 [00:53<00:09,  7.09it/s]
84%|████████▍ | 372/441 [00:53<00:09,  7.08it/s]
85%|████████▍ | 373/441 [00:54<00:09,  7.08it/s]
85%|████████▍ | 374/441 [00:54<00:09,  7.08it/s]
85%|████████▌ | 375/441 [00:54<00:09,  7.07it/s]
85%|████████▌ | 376/441 [00:54<00:09,  7.07it/s]
85%|████████▌ | 377/441 [00:54<00:09,  7.07it/s]
86%|████████▌ | 378/441 [00:54<00:08,  7.08it/s]
86%|████████▌ | 379/441 [00:54<00:08,  7.08it/s]
86%|████████▌ | 380/441 [00:55<00:08,  7.08it/s]
86%|████████▋ | 381/441 [00:55<00:08,  7.09it/s]
87%|████████▋ | 382/441 [00:55<00:08,  7.08it/s]
87%|████████▋ | 383/441 [00:55<00:08,  7.09it/s]
87%|████████▋ | 384/441 [00:55<00:08,  7.09it/s]
87%|████████▋ | 385/441 [00:55<00:07,  7.08it/s]
88%|████████▊ | 386/441 [00:55<00:07,  7.09it/s]
88%|████████▊ | 387/441 [00:56<00:07,  7.10it/s]
88%|████████▊ | 388/441 [00:56<00:07,  7.09it/s]
88%|████████▊ | 389/441 [00:56<00:07,  7.09it/s]
88%|████████▊ | 390/441 [00:56<00:07,  7.09it/s]
89%|████████▊ | 391/441 [00:56<00:07,  7.09it/s]
89%|████████▉ | 392/441 [00:56<00:06,  7.05it/s]
89%|████████▉ | 393/441 [00:56<00:06,  7.03it/s]
89%|████████▉ | 394/441 [00:57<00:06,  7.05it/s]
90%|████████▉ | 395/441 [00:57<00:06,  7.06it/s]
90%|████████▉ | 396/441 [00:57<00:06,  7.06it/s]
```

```
90%|█████████ | 397/441 [00:57<00:06,  7.02it/s]
90%|█████████ | 398/441 [00:57<00:06,  6.99it/s]
90%|█████████ | 399/441 [00:57<00:06,  6.97it/s]
91%|█████████ | 400/441 [00:57<00:05,  6.97it/s]
91%|█████████ | 401/441 [00:58<00:05,  6.96it/s]
91%|█████████ | 402/441 [00:58<00:05,  6.99it/s]
91%|█████████▏| 403/441 [00:58<00:05,  7.03it/s]
92%|█████████▏| 404/441 [00:58<00:05,  7.06it/s]
92%|█████████▏| 405/441 [00:58<00:05,  7.08it/s]
92%|█████████▏| 406/441 [00:58<00:04,  7.09it/s]
92%|█████████▏| 407/441 [00:58<00:04,  7.09it/s]
93%|█████████▎| 408/441 [00:59<00:04,  7.09it/s]
93%|█████████▎| 409/441 [00:59<00:04,  7.09it/s]
93%|█████████▎| 410/441 [00:59<00:04,  7.05it/s]
93%|█████████▎| 411/441 [00:59<00:04,  7.06it/s]
93%|█████████▎| 412/441 [00:59<00:04,  7.08it/s]
94%|█████████▎| 413/441 [00:59<00:03,  7.09it/s]
94%|█████████▍| 414/441 [00:59<00:03,  7.09it/s]
94%|█████████▍| 415/441 [01:00<00:03,  7.09it/s]
94%|█████████▍| 416/441 [01:00<00:03,  7.08it/s]
95%|█████████▍| 417/441 [01:00<00:03,  7.07it/s]
95%|█████████▍| 418/441 [01:00<00:03,  7.07it/s]
95%|█████████▌| 419/441 [01:00<00:03,  7.07it/s]
95%|█████████▌| 420/441 [01:00<00:02,  7.07it/s]
95%|█████████▌| 421/441 [01:00<00:02,  7.08it/s]
96%|█████████▌| 422/441 [01:01<00:02,  7.07it/s]
96%|█████████▌| 423/441 [01:01<00:02,  7.05it/s]
96%|█████████▌| 424/441 [01:01<00:02,  7.03it/s]
96%|█████████▋| 425/441 [01:01<00:02,  7.04it/s]
97%|█████████▋| 426/441 [01:01<00:02,  7.06it/s]
97%|█████████▋| 427/441 [01:01<00:01,  7.06it/s]
97%|█████████▋| 428/441 [01:01<00:01,  7.07it/s]
97%|█████████▋| 429/441 [01:02<00:01,  7.07it/s]
98%|█████████▊| 430/441 [01:02<00:01,  7.08it/s]
98%|█████████▊| 431/441 [01:02<00:01,  7.09it/s]
98%|█████████▊| 432/441 [01:02<00:01,  7.09it/s]
98%|█████████▊| 433/441 [01:02<00:01,  7.09it/s]
98%|█████████▊| 434/441 [01:02<00:00,  7.06it/s]
99%|█████████▊| 435/441 [01:02<00:00,  7.05it/s]
99%|█████████▉| 436/441 [01:03<00:00,  7.03it/s]
99%|█████████▉| 437/441 [01:03<00:00,  7.02it/s]
99%|█████████▉| 438/441 [01:03<00:00,  7.00it/s]
100%|█████████▉| 439/441 [01:03<00:00,  7.02it/s]
100%|█████████▉| 440/441 [01:03<00:00,  7.04it/s]
100%|██████████| 441/441 [01:03<00:00,  7.06it/s]
[INFO|trainer.py:1508] 2023-07-31 17:47:14,895 >>
Training completed. Do not forget to share your model on huggingface.co/models =)
[INFO|trainer.py:1508] 2023-07-31 17:47:14,895 >>
Training completed. Do not forget to share your model on huggingface.co/models =)
{'train_runtime': 63.7744, 'train_samples_per_second': 13.83, 'train_steps_per_sec
ond': 6.915, 'train_loss': 3.7288090366354876, 'epoch': 3.0}
100%|██████████| 441/441 [01:03<00:00,  7.06it/s]
100%|██████████| 441/441 [01:03<00:00,  6.92it/s]
[INFO|trainer.py:2139] 2023-07-31 17:47:14,897 >> Saving model checkpoint to /opt/
ml/model
[INFO|trainer.py:2139] 2023-07-31 17:47:14,897 >> Saving model checkpoint to /opt/
ml/model
[INFO|configuration_utils.py:439] 2023-07-31 17:47:14,898 >> Configuration saved i
n /opt/ml/model/config.json
[INFO|configuration_utils.py:439] 2023-07-31 17:47:14,898 >> Configuration saved i
n /opt/ml/model/config.json
[INFO|modeling_utils.py:1084] 2023-07-31 17:47:15,602 >> Model weights saved in /o
pt/ml/model/pytorch_model.bin
[INFO|modeling_utils.py:1084] 2023-07-31 17:47:15,602 >> Model weights saved in /o
```

```
pt/ml/model/pytorch_model.bin
[INFO|tokenization_utils_base.py:2094] 2023-07-31 17:47:15,603 >> tokenizer config
file saved in /opt/ml/model/tokenizer_config.json
[INFO|tokenization_utils_base.py:2094] 2023-07-31 17:47:15,603 >> tokenizer config
file saved in /opt/ml/model/tokenizer_config.json
[INFO|tokenization_utils_base.py:2100] 2023-07-31 17:47:15,603 >> Special tokens f
ile saved in /opt/ml/model/special_tokens_map.json
[INFO|tokenization_utils_base.py:2100] 2023-07-31 17:47:15,603 >> Special tokens f
ile saved in /opt/ml/model/special_tokens_map.json
***** train metrics *****
epoch                    =        3.0
  train_loss               =     3.7288
  train_runtime            = 0:01:03.77
  train_samples            =        294
  train_samples_per_second =      13.83
  train_steps_per_second   =      6.915
07/31/2023 17:47:15 - INFO - __main__ - *** Evaluate ***
[INFO|trainer.py:2389] 2023-07-31 17:47:15,714 >> ***** Running Evaluation *****
[INFO|trainer.py:2389] 2023-07-31 17:47:15,714 >> ***** Running Evaluation *****
[INFO|trainer.py:2391] 2023-07-31 17:47:15,714 >>   Num examples = 17
[INFO|trainer.py:2394] 2023-07-31 17:47:15,714 >>   Batch size = 2
[INFO|trainer.py:2391] 2023-07-31 17:47:15,714 >>   Num examples = 17
[INFO|trainer.py:2394] 2023-07-31 17:47:15,714 >>   Batch size = 2
0%|          | 0/9 [00:00<?, ?it/s]
44%|████      | 4/9 [00:00<00:00, 28.65it/s]
78%|███████▊  | 7/9 [00:00<00:00, 24.26it/s]
100%|██████████| 9/9 [00:00<00:00, 20.56it/s]
***** eval metrics *****
epoch                   =        3.0
  eval_accuracy           =     0.3779
  eval_loss               =     3.4596
  eval_runtime            = 0:00:00.48
  eval_samples            =         17
  eval_samples_per_second =     34.944
  eval_steps_per_second   =       18.5
  perplexity              =    31.8058
2023-07-31 17:47:16,784 sagemaker-training-toolkit INFO     Waiting for the proces
s to finish and give a return code.
2023-07-31 17:47:16,784 sagemaker-training-toolkit INFO     Done waiting for a ret
urn code. Received 0 from exiting process.
2023-07-31 17:47:16,784 sagemaker-training-toolkit INFO     Reporting training SUC
CESS

2023-07-31 17:47:25 Uploading - Uploading generated training model
2023-07-31 17:47:56 Completed - Training job completed
Training seconds: 433
Billable seconds: 433
```

Model deployment

Now we want to deploy both the original 'distilGPT2' model and our finetuned 'shakespeare-distilGPT2' model. Therefore we first retrieve the S3 path to the model artifact archive of our finetuned model:

In [11]:

```
latest_job_name = huggingface_estimator.latest_training_job.job_name
latest_job_name
```

Out[11]:

```
'huggingface-pytorch-training-2023-07-31-17-38-58-475'
```

In [12]:

```python
def get_s3_artifact_path(training_job_name):
    # Get the ModelArtifacts object for the training job
    sagemaker_session = sagemaker.Session()

    training_job = sagemaker_session.describe_training_job(training_job_name)

    model_artifacts = training_job['ModelArtifacts']

    # Retrieve the S3 path to the model artifact
    s3_path = model_artifacts['S3ModelArtifacts']
    return s3_path
```

In [13]:

```python
s3_path = get_s3_artifact_path(latest_job_name)
s3_path
```

Out[13]:

```
's3://immersion-day-bucket-882819251225/huggingface-pytorch-training-2023-07-31-17
-38-58-475/output/model.tar.gz'
```

Then we deploy the model to a 'ml.g4dn.xlarge' instance using the HuggingFaceModel class:

In [14]:

```python
# create Hugging Face Model Class
huggingface_model_finetuned = HuggingFaceModel(
    image_uri=f'763104351884.dkr.ecr.{region}.amazonaws.com/huggingface-pytorch-in
    model_data=s3_path ,
        role=role
    )
```

In [15]:

```python
predictor_finetuned = huggingface_model_finetuned.deploy(
    initial_instance_count=1, # number of instances
    instance_type='ml.g4dn.xlarge',
    endpoint_name='sm-endpoint-distilgpt2-shakespeare-immersion-day',
)
```

```
INFO:sagemaker:Creating model with name: huggingface-pytorch-inference-2023-07-31-
17-51-52-141
INFO:sagemaker:Creating endpoint-config with name sm-endpoint-distilgpt2-shakespea
re-immersion-day
INFO:sagemaker:Creating endpoint with name sm-endpoint-distilgpt2-shakespeare-imme
rsion-day
-------!
```

We also deploy the original model to a 'ml.g4dn.xlarge' instance. Therefore we use a cool feature built-in into the SageMaker SDK - we can define a model to be deployed directly from the HuggingFace model hub together with the model task to be performed directly as environment variables when creating a HuggingFaceModel, SageMaker Inference handles the rest:

In [16]:

```python
hub = {
  'HF_MODEL_ID':'distilgpt2', # model_id from hf.co/models
  'HF_TASK':'text-generation' # NLP task you want to use for predictions
}

# create Hugging Face Model Class
huggingface_model_plain = HuggingFaceModel(
   env=hub,                                                # configuration for load
   role=role,                                              # IAM role with permiss
   image_uri=f'763104351884.dkr.ecr.{region}.amazonaws.com/huggingface-pytorch-infe
)
```

In [17]:

```
predictor_plain = huggingface_model_plain.deploy(
    initial_instance_count=1, # number of instances
    instance_type='ml.g4dn.xlarge',
    endpoint_name='sm-endpoint-distilgpt2-immersion-day',
)
```

```
INFO:sagemaker:Creating model with name: huggingface-pytorch-inference-2023-07-31-
18-01-40-069
INFO:sagemaker:Creating endpoint-config with name sm-endpoint-distilgpt2-immersion
-day
INFO:sagemaker:Creating endpoint with name sm-endpoint-distilgpt2-immersion-day
-------!
```

Inference

Having the two endpoints available, we can experiment and observe the impact the finetuning has in terms of performance of the text-generation task.

In [20]:

```
predictor_finetuned.predict({"inputs": "The meaning of love",
"parameters": {
    "min_length": 50,
    "max_length": 100
}})[0]['generated_text']
```

Out[20]:

```
"The meaning of love and beauty was given to each of her friends.\n\nThe son of a
great father,\nAnd to his great mother,\nHath brought from the earth his daughte
r.\n\nThe love of a good mother that a loving father lives.\n\nROMEO:\nWhat, Rome
o?\n\nEXITUTO:\nFor if we were to love her father,\nIf we had were to have his so
n,\nHe would kiss his mother's heart."
```

In [21]:

```
predictor_plain.predict({"inputs": "The meaning of love",
"parameters": {
    "min_length": 50,
    "max_length": 100
}})[0]['generated_text']
```

Out[21]:

```
"The meaning of love is, however, often a personal thing. Most importantly, love i
s an abstract act of love itself - a feeling of belonging to others. An external l
ove which requires it is not connected to one's identity. You can see this in a ve
ry simple example of the relationship I've been having with my partner in the past
few weeks. In the beginning of the week, I was having difficulty with my identity,
or having a very low level of self esteem, or taking into consideration"
```

Cleanup

Finally, we clean up all resources not needed anymore since we pledge for the responsible use of compute resources. In this case this is the created endpoint together with the respective endpoint configuration.

In [ ]:

```
predictor_finetuned.delete_endpoint(delete_endpoint_config=True)
```

In [ ]:

```
predictor_plain.delete_endpoint(delete_endpoint_config=True)
```

In [ ]:

```

```