Compile and Train the GPT2 Model using the Transformers Trainer API with the SST2 Dataset for Multi-Node Multi-GPU Training

SageMaker Training Compiler Overview

SageMaker Training Compiler is a capability of SageMaker that makes these hard-to-implement optimizations to reduce training time on GPU instances. The compiler optimizes DL models to accelerate training by more efficiently using SageMaker machine learning (ML) GPU instances. SageMaker Training Compiler is available at no additional charge within SageMaker and can help reduce total billable time as it accelerates training.

SageMaker Training Compiler is integrated into the AWS Deep Learning Containers (DLCs). Using the SageMaker Training Compiler enabled AWS DLCs, you can compile and optimize training jobs on GPU instances with minimal changes to your code. Bring your deep learning models to SageMaker and enable SageMaker Training Compiler to accelerate the speed of your training job on SageMaker ML instances for accelerated computing.

For more information, see SageMaker Training Compiler (https://docs.aws.amazon.com/sagemaker/latest/dg/training-compiler.html) in the *Amazon SageMaker Developer Guide*.

Introduction

In this demo, you'll use Hugging Face's transformers and datasets libraries with Amazon SageMaker Training Compiler to train the gpt-2 model on the Stanford Sentiment Treebank v2 (SST2) dataset. To get started, we need to set up the environment with a few prerequisite steps, for permissions, configurations, and so on.

NOTE: You can run this demo in SageMaker Studio, SageMaker notebook instances, or your local machine with AWS CLI set up. If using SageMaker Studio or SageMaker notebook instances, make sure you choose one of the PyTorch-based kernels, Python 3 (PyTorch x.y Python 3.x CPU Optimized) or conda_pytorch_p36 respectively.

NOTE: This notebook uses four ml.p4d.24xlarge instances that have multiple GPUs. If you don't have enough quota, see Request a service quota increase for SageMaker resources (https://docs.aws.amazon.com/sagemaker/latest/dg/regions-quotas.html#service-limit-increase-request-procedure).

Development Environment

Installation

This example notebook requires the **SageMaker Python SDK v2.108.0** and **transformers v4.21**.

In [1]:
```
!pip install "sagemaker>=2.108.0" botocore boto3 awscli s3fs typing
-extensions "torch==1.11.0" pandas numpy --upgrade
```

```
Looking in indexes: https://pypi.org/simple, https://pip.repos.neuro
n.amazonaws.com
Requirement already satisfied: sagemaker>=2.108.0 in /home/ec2-user/
anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (2.112.2)
Requirement already satisfied: botocore in /home/ec2-user/anaconda3/
envs/pytorch_p38/lib/python3.8/site-packages (1.27.92)
Requirement already satisfied: boto3 in /home/ec2-user/anaconda3/env
s/pytorch_p38/lib/python3.8/site-packages (1.24.92)
Requirement already satisfied: awscli in /home/ec2-user/anaconda3/en
vs/pytorch_p38/lib/python3.8/site-packages (1.25.93)
Requirement already satisfied: s3fs in /home/ec2-user/anaconda3/envs
/pytorch_p38/lib/python3.8/site-packages (0.4.2)
Collecting s3fs
  Using cached s3fs-2022.8.2-py3-none-any.whl (27 kB)
Requirement already satisfied: typing-extensions in /home/ec2-user/a
naconda3/envs/pytorch_p38/lib/python3.8/site-packages (4.4.0)
Requirement already satisfied: torch==1.11.0 in /home/ec2-user/anaco
nda3/envs/pytorch_p38/lib/python3.8/site-packages (1.11.0)
Requirement already satisfied: pandas in /home/ec2-user/anaconda3/en
vs/pytorch_p38/lib/python3.8/site-packages (1.5.0)
Requirement already satisfied: numpy in /home/ec2-user/anaconda3/env
s/pytorch_p38/lib/python3.8/site-packages (1.23.4)
Requirement already satisfied: smdebug-rulesconfig==1.0.1 in /home/e
c2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from
sagemaker>=2.108.0) (1.0.1)
Requirement already satisfied: pathos in /home/ec2-user/anaconda3/en
vs/pytorch_p38/lib/python3.8/site-packages (from sagemaker>=2.108.0)
(0.2.8)
Requirement already satisfied: schema in /home/ec2-user/anaconda3/en
vs/pytorch_p38/lib/python3.8/site-packages (from sagemaker>=2.108.0)
(0.7.5)
Requirement already satisfied: google-pasta in /home/ec2-user/anacon
da3/envs/pytorch_p38/lib/python3.8/site-packages (from sagemaker>=2.
108.0) (0.2.0)
Requirement already satisfied: protobuf3-to-dict<1.0,>=0.1.5 in /hom
e/ec2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (f
rom sagemaker>=2.108.0) (0.1.5)
Requirement already satisfied: packaging>=20.0 in /home/ec2-user/ana
conda3/envs/pytorch_p38/lib/python3.8/site-packages (from sagemaker>
=2.108.0) (21.3)
Requirement already satisfied: attrs<23,>=20.3.0 in /home/ec2-user/a
naconda3/envs/pytorch_p38/lib/python3.8/site-packages (from sagemake
r>=2.108.0) (21.2.0)
Requirement already satisfied: importlib-metadata<5.0,>=1.4.0 in /ho
me/ec2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages
(from sagemaker>=2.108.0) (4.8.2)
Requirement already satisfied: protobuf<4.0,>=3.1 in /home/ec2-user/
anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from sagemak
er>=2.108.0) (3.20.1)
Requirement already satisfied: urllib3<1.27,>=1.25.4 in /home/ec2-us
er/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from boto
core) (1.26.8)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /home/
ec2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (fro
```

```
m botocore) (2.8.2)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /home/ec2-u
ser/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from bot
ocore) (0.10.0)
Requirement already satisfied: s3transfer<0.7.0,>=0.6.0 in /home/ec2
-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from b
oto3) (0.6.0)
Requirement already satisfied: colorama<0.4.5,>=0.2.5 in /home/ec2-u
ser/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from aws
cli) (0.4.3)
Requirement already satisfied: rsa<4.8,>=3.1.2 in /home/ec2-user/ana
conda3/envs/pytorch_p38/lib/python3.8/site-packages (from awscli)
(4.7.2)
Requirement already satisfied: PyYAML<5.5,>=3.10 in /home/ec2-user/a
naconda3/envs/pytorch_p38/lib/python3.8/site-packages (from awscli)
(5.4.1)
Requirement already satisfied: docutils<0.17,>=0.10 in /home/ec2-use
r/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from awscl
i) (0.15.2)
Collecting fsspec==2022.8.2
  Using cached fsspec-2022.8.2-py3-none-any.whl (140 kB)
Collecting aiobotocore~=2.4.0
  Using cached aiobotocore-2.4.0-py3-none-any.whl (65 kB)
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /home/e
c2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from
s3fs) (3.8.1)
Requirement already satisfied: pytz>=2020.1 in /home/ec2-user/anacon
da3/envs/pytorch_p38/lib/python3.8/site-packages (from pandas) (202
1.3)
INFO: pip is looking at multiple versions of numpy to determine whic
h version is compatible with other requirements. This could take a w
hile.
Collecting numpy
  Using cached numpy-1.23.4-cp38-cp38-manylinux_2_17_x86_64.manylinu
x2014_x86_64.whl (17.1 MB)
INFO: pip is looking at multiple versions of pandas to determine whi
ch version is compatible with other requirements. This could take a
while.
Collecting pandas
  Using cached pandas-1.5.0-cp38-cp38-manylinux_2_17_x86_64.manylinu
x2014_x86_64.whl (12.2 MB)
INFO: pip is looking at multiple versions of typing-extensions to de
termine which version is compatible with other requirements. This co
uld take a while.
Collecting typing-extensions
  Using cached typing_extensions-4.4.0-py3-none-any.whl (26 kB)
INFO: pip is looking at multiple versions of fsspec to determine whi
ch version is compatible with other requirements. This could take a
while.
INFO: pip is looking at multiple versions of <Python from Requires-P
ython> to determine which version is compatible with other requireme
nts. This could take a while.
INFO: pip is looking at multiple versions of s3fs to determine which
version is compatible with other requirements. This could take a whi
```

```
le.
Collecting s3fs
  Using cached s3fs-2022.8.1-py3-none-any.whl (27 kB)
Collecting fsspec==2022.8.1
  Using cached fsspec-2022.8.1-py3-none-any.whl (140 kB)
Collecting s3fs
  Using cached s3fs-2022.8.0-py3-none-any.whl (27 kB)
Collecting fsspec==2022.8.0
  Using cached fsspec-2022.8.0-py3-none-any.whl (140 kB)
Collecting s3fs
  Using cached s3fs-2022.7.1-py3-none-any.whl (27 kB)
Collecting fsspec==2022.7.1
  Using cached fsspec-2022.7.1-py3-none-any.whl (141 kB)
Collecting aiobotocore~=2.3.4
  Using cached aiobotocore-2.3.4-py3-none-any.whl (64 kB)
Collecting s3fs
  Using cached s3fs-2022.7.0-py3-none-any.whl (27 kB)
Collecting fsspec==2022.7.0
  Using cached fsspec-2022.7.0-py3-none-any.whl (141 kB)
Collecting s3fs
  Using cached s3fs-2022.5.0-py3-none-any.whl (27 kB)
Collecting fsspec==2022.5.0
  Using cached fsspec-2022.5.0-py3-none-any.whl (140 kB)
Requirement already satisfied: aiobotocore~=2.3.0 in /home/ec2-user/
anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from s3fs)
(2.3.3)
Collecting aiobotocore~=2.3.0
  Using cached aiobotocore-2.3.2.tar.gz (104 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-2.3.1.tar.gz (65 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-2.3.0.tar.gz (65 kB)
  Preparing metadata (setup.py) ... done
Collecting s3fs
  Using cached s3fs-2022.3.0-py3-none-any.whl (26 kB)
Collecting aiobotocore~=2.2.0
  Using cached aiobotocore-2.2.0.tar.gz (59 kB)
  Preparing metadata (setup.py) ... done
Collecting fsspec==2022.3.0
  Using cached fsspec-2022.3.0-py3-none-any.whl (136 kB)
Collecting s3fs
  Using cached s3fs-2022.2.0-py3-none-any.whl (26 kB)
Collecting fsspec==2022.02.0
  Using cached fsspec-2022.2.0-py3-none-any.whl (134 kB)
Collecting aiobotocore~=2.1.0
  Using cached aiobotocore-2.1.2.tar.gz (58 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-2.1.1.tar.gz (57 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-2.1.0.tar.gz (54 kB)
  Preparing metadata (setup.py) ... done
INFO: pip is looking at multiple versions of fsspec to determine whi
ch version is compatible with other requirements. This could take a
while.
```

```
INFO: pip is looking at multiple versions of <Python from Requires-P
ython> to determine which version is compatible with other requireme
nts. This could take a while.
INFO: pip is looking at multiple versions of s3fs to determine which
version is compatible with other requirements. This could take a whi
le.
Collecting s3fs
  Using cached s3fs-2022.1.0-py3-none-any.whl (25 kB)
Collecting fsspec==2022.01.0
  Using cached fsspec-2022.1.0-py3-none-any.whl (133 kB)
Collecting s3fs
  Using cached s3fs-2021.11.1-py3-none-any.whl (25 kB)
Requirement already satisfied: fsspec==2021.11.1 in /home/ec2-user/a
naconda3/envs/pytorch_p38/lib/python3.8/site-packages (from s3fs) (2
021.11.1)
Collecting aiobotocore~=2.0.1
  Using cached aiobotocore-2.0.1-py3-none-any.whl
Requirement already satisfied: wrapt>=1.10.10 in /home/ec2-user/anac
onda3/envs/pytorch_p38/lib/python3.8/site-packages (from aiobotocore
~=2.0.1->s3fs) (1.13.3)
Collecting fsspec==2021.11.1
  Using cached fsspec-2021.11.1-py3-none-any.whl (132 kB)
Collecting s3fs
  Using cached s3fs-2021.11.0-py3-none-any.whl (25 kB)
Collecting fsspec==2021.11.0
  Using cached fsspec-2021.11.0-py3-none-any.whl (132 kB)
Collecting aiobotocore~=1.4.1
  Using cached aiobotocore-1.4.2.tar.gz (52 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.4.1.tar.gz (52 kB)
  Preparing metadata (setup.py) ... done
Collecting s3fs
  Using cached s3fs-2021.10.1-py3-none-any.whl (26 kB)
Collecting fsspec==2021.10.1
  Using cached fsspec-2021.10.1-py3-none-any.whl (125 kB)
INFO: This is taking longer than usual. You might need to provide th
e dependency resolver with stricter constraints to reduce runtime. S
ee https://pip.pypa.io/warnings/backtracking for guidance. If you wa
nt to abort this run, press Ctrl + C.
Collecting s3fs
  Using cached s3fs-2021.10.0-py3-none-any.whl (26 kB)
Collecting fsspec==2021.10.0
  Using cached fsspec-2021.10.0-py3-none-any.whl (125 kB)
INFO: This is taking longer than usual. You might need to provide th
e dependency resolver with stricter constraints to reduce runtime. S
ee https://pip.pypa.io/warnings/backtracking for guidance. If you wa
nt to abort this run, press Ctrl + C.
INFO: This is taking longer than usual. You might need to provide th
e dependency resolver with stricter constraints to reduce runtime. S
ee https://pip.pypa.io/warnings/backtracking for guidance. If you wa
nt to abort this run, press Ctrl + C.
Collecting s3fs
  Using cached s3fs-2021.9.0-py3-none-any.whl (26 kB)
Collecting fsspec==2021.09.0
```

```
  Using cached fsspec-2021.9.0-py3-none-any.whl (123 kB)
Collecting s3fs
  Using cached s3fs-2021.8.1-py3-none-any.whl (26 kB)
Collecting fsspec==2021.08.1
  Using cached fsspec-2021.8.1-py3-none-any.whl (119 kB)
Collecting aiobotocore~=1.4.0
  Using cached aiobotocore-1.4.0.tar.gz (51 kB)
  Preparing metadata (setup.py) ... done
Collecting s3fs
  Using cached s3fs-2021.8.0-py3-none-any.whl (26 kB)
Collecting fsspec==2021.07.0
  Using cached fsspec-2021.7.0-py3-none-any.whl (118 kB)
Collecting s3fs
  Using cached s3fs-2021.7.0-py3-none-any.whl (25 kB)
Collecting aiobotocore>=1.0.1
  Using cached aiobotocore-2.0.0.tar.gz (52 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.3.3.tar.gz (50 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.3.2.tar.gz (49 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.3.1.tar.gz (48 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.3.0.tar.gz (48 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.2.2.tar.gz (48 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.2.1.tar.gz (48 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.2.0.tar.gz (47 kB)
  Preparing metadata (setup.py) ... done
  Using cached aiobotocore-1.1.2-py3-none-any.whl (45 kB)
  Using cached aiobotocore-1.1.1-py3-none-any.whl (45 kB)
  Using cached aiobotocore-1.1.0-py3-none-any.whl (43 kB)
  Using cached aiobotocore-1.0.7-py3-none-any.whl (42 kB)
  Using cached aiobotocore-1.0.6-py3-none-any.whl (42 kB)
  Using cached aiobotocore-1.0.5-py3-none-any.whl (42 kB)
  Using cached aiobotocore-1.0.4-py3-none-any.whl (41 kB)
  Using cached aiobotocore-1.0.3-py3-none-any.whl (40 kB)
  Using cached aiobotocore-1.0.2-py3-none-any.whl (40 kB)
  Using cached aiobotocore-1.0.1-py3-none-any.whl (40 kB)
Collecting s3fs
  Using cached s3fs-2021.6.1-py3-none-any.whl (25 kB)
Collecting fsspec==2021.06.1
  Using cached fsspec-2021.6.1-py3-none-any.whl (115 kB)
Collecting s3fs
  Using cached s3fs-2021.6.0-py3-none-any.whl (24 kB)
Collecting fsspec==2021.06.0
  Using cached fsspec-2021.6.0-py3-none-any.whl (114 kB)
Collecting s3fs
  Using cached s3fs-2021.5.0-py3-none-any.whl (24 kB)
Collecting fsspec==2021.05.0
  Using cached fsspec-2021.5.0-py3-none-any.whl (111 kB)
Collecting s3fs
```

```
  Using cached s3fs-2021.4.0-py3-none-any.whl (23 kB)
Collecting fsspec==2021.04.0
  Using cached fsspec-2021.4.0-py3-none-any.whl (108 kB)
Collecting s3fs
  Using cached s3fs-0.6.0-py3-none-any.whl (23 kB)
  Using cached s3fs-0.5.2-py3-none-any.whl (22 kB)
  Using cached s3fs-0.5.1-py3-none-any.whl (21 kB)
  Using cached s3fs-0.5.0-py3-none-any.whl (21 kB)
Requirement already satisfied: zipp>=0.5 in /home/ec2-user/anaconda3
/envs/pytorch_p38/lib/python3.8/site-packages (from importlib-metada
ta<5.0,>=1.4.0->sagemaker>=2.108.0) (3.6.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /home/ec2
-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from p
ackaging>=20.0->sagemaker>=2.108.0) (3.0.6)
Requirement already satisfied: six in /home/ec2-user/anaconda3/envs/
pytorch_p38/lib/python3.8/site-packages (from protobuf3-to-dict<1.0,
>=0.1.5->sagemaker>=2.108.0) (1.16.0)
Requirement already satisfied: pyasn1>=0.1.3 in /home/ec2-user/anaco
nda3/envs/pytorch_p38/lib/python3.8/site-packages (from rsa<4.8,>=3.
1.2->awscli) (0.4.8)
Requirement already satisfied: pox>=0.3.0 in /home/ec2-user/anaconda
3/envs/pytorch_p38/lib/python3.8/site-packages (from pathos->sagemak
er>=2.108.0) (0.3.0)
Requirement already satisfied: multiprocess>=0.70.12 in /home/ec2-us
er/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from path
os->sagemaker>=2.108.0) (0.70.12.2)
Requirement already satisfied: ppft>=1.6.6.4 in /home/ec2-user/anaco
nda3/envs/pytorch_p38/lib/python3.8/site-packages (from pathos->sage
maker>=2.108.0) (1.6.6.4)
Requirement already satisfied: dill>=0.3.4 in /home/ec2-user/anacond
a3/envs/pytorch_p38/lib/python3.8/site-packages (from pathos->sagema
ker>=2.108.0) (0.3.4)
Requirement already satisfied: contextlib2>=0.5.5 in /home/ec2-user/
anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from schema-
>sagemaker>=2.108.0) (21.6.0)
WARNING: You are using pip version 22.0.4; however, version 22.2.2 i
```

In [2]:
```
!pip install "transformers==4.21" datasets --upgrade
```

```
Looking in indexes: https://pypi.org/simple, https://pip.repos.neuro
n.amazonaws.com
Requirement already satisfied: transformers==4.21 in /home/ec2-user/
anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (4.21.0)
Requirement already satisfied: datasets in /home/ec2-user/anaconda3/
envs/pytorch_p38/lib/python3.8/site-packages (2.6.1)
Requirement already satisfied: filelock in /home/ec2-user/anaconda3/
envs/pytorch_p38/lib/python3.8/site-packages (from transformers==4.2
1) (3.4.0)
Requirement already satisfied: requests in /home/ec2-user/anaconda3/
envs/pytorch_p38/lib/python3.8/site-packages (from transformers==4.2
1) (2.26.0)
Requirement already satisfied: tokenizers!=0.11.3,<0.13,>=0.11.1 in
/home/ec2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-package
s (from transformers==4.21) (0.12.1)
Requirement already satisfied: tqdm>=4.27 in /home/ec2-user/anaconda
3/envs/pytorch_p38/lib/python3.8/site-packages (from transformers==
4.21) (4.62.3)
Requirement already satisfied: packaging>=20.0 in /home/ec2-user/ana
conda3/envs/pytorch_p38/lib/python3.8/site-packages (from transforme
rs==4.21) (21.3)
Requirement already satisfied: pyyaml>=5.1 in /home/ec2-user/anacond
a3/envs/pytorch_p38/lib/python3.8/site-packages (from transformers==
4.21) (5.4.1)
Requirement already satisfied: regex!=2019.12.17 in /home/ec2-user/a
naconda3/envs/pytorch_p38/lib/python3.8/site-packages (from transfor
mers==4.21) (2021.11.10)
Requirement already satisfied: numpy>=1.17 in /home/ec2-user/anacond
a3/envs/pytorch_p38/lib/python3.8/site-packages (from transformers==
4.21) (1.23.4)
Requirement already satisfied: huggingface-hub<1.0,>=0.1.0 in /home/
ec2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (fro
m transformers==4.21) (0.9.0)
Requirement already satisfied: aiohttp in /home/ec2-user/anaconda3/e
nvs/pytorch_p38/lib/python3.8/site-packages (from datasets) (3.8.1)
Requirement already satisfied: pandas in /home/ec2-user/anaconda3/en
vs/pytorch_p38/lib/python3.8/site-packages (from datasets) (1.5.0)
Requirement already satisfied: fsspec[http]>=2021.11.1 in /home/ec2-
user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from da
tasets) (2021.11.1)
Requirement already satisfied: dill<0.3.6 in /home/ec2-user/anaconda
3/envs/pytorch_p38/lib/python3.8/site-packages (from datasets) (0.3.
4)
Requirement already satisfied: responses<0.19 in /home/ec2-user/anac
onda3/envs/pytorch_p38/lib/python3.8/site-packages (from datasets)
(0.18.0)
Requirement already satisfied: multiprocess in /home/ec2-user/anacon
da3/envs/pytorch_p38/lib/python3.8/site-packages (from datasets) (0.
70.12.2)
Requirement already satisfied: xxhash in /home/ec2-user/anaconda3/en
vs/pytorch_p38/lib/python3.8/site-packages (from datasets) (3.0.0)
Requirement already satisfied: pyarrow>=6.0.0 in /home/ec2-user/anac
onda3/envs/pytorch_p38/lib/python3.8/site-packages (from datasets)
(7.0.0)
```

```
Requirement already satisfied: typing-extensions>=3.7.4.3 in /home/e
c2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from
huggingface-hub<1.0,>=0.1.0->transformers==4.21) (4.4.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /home/ec2
-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from p
ackaging>=20.0->transformers==4.21) (3.0.6)
Requirement already satisfied: certifi>=2017.4.17 in /home/ec2-user/
anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from request
s->transformers==4.21) (2021.10.8)
Requirement already satisfied: idna<4,>=2.5 in /home/ec2-user/anacon
da3/envs/pytorch_p38/lib/python3.8/site-packages (from requests->tra
nsformers==4.21) (3.1)
Requirement already satisfied: charset-normalizer~=2.0.0 in /home/ec
2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from
requests->transformers==4.21) (2.0.7)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /home/ec2-us
er/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from requ
ests->transformers==4.21) (1.26.8)
Requirement already satisfied: attrs>=17.3.0 in /home/ec2-user/anaco
nda3/envs/pytorch_p38/lib/python3.8/site-packages (from aiohttp->dat
asets) (21.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in /home/ec2-user/a
naconda3/envs/pytorch_p38/lib/python3.8/site-packages (from aiohttp-
>datasets) (1.2.0)
Requirement already satisfied: aiosignal>=1.1.2 in /home/ec2-user/an
aconda3/envs/pytorch_p38/lib/python3.8/site-packages (from aiohttp->
datasets) (1.2.0)
Requirement already satisfied: yarl<2.0,>=1.0 in /home/ec2-user/anac
onda3/envs/pytorch_p38/lib/python3.8/site-packages (from aiohttp->da
tasets) (1.7.2)
Requirement already satisfied: multidict<7.0,>=4.5 in /home/ec2-user
/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from aiohtt
p->datasets) (5.2.0)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in /home/
ec2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (fro
m aiohttp->datasets) (4.0.1)
Requirement already satisfied: pytz>=2020.1 in /home/ec2-user/anacon
da3/envs/pytorch_p38/lib/python3.8/site-packages (from pandas->datas
ets) (2021.3)
Requirement already satisfied: python-dateutil>=2.8.1 in /home/ec2-u
ser/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages (from pan
das->datasets) (2.8.2)
Requirement already satisfied: six>=1.5 in /home/ec2-user/anaconda3/
envs/pytorch_p38/lib/python3.8/site-packages (from python-dateutil>=
2.8.1->pandas->datasets) (1.16.0)
WARNING: You are using pip version 22.0.4; however, version 22.2.2 i
s available.
You should consider upgrading via the '/home/ec2-user/anaconda3/envs
/pytorch_p38/bin/python -m pip install --upgrade pip' command.
```

In [3]:
```
import botocore
import boto3
import sagemaker
import transformers
import pandas as pd

print(f"sagemaker: {sagemaker.__version__}")
print(f"transformers: {transformers.__version__}")
```

```
/home/ec2-user/anaconda3/envs/pytorch_p38/lib/python3.8/site-packages/scipy/__init__.py:146: UserWarning: A NumPy version >=1.16.5 and <1.23.0 is required for this version of SciPy (detected version 1.23.4
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"

sagemaker: 2.112.2
transformers: 4.21.0
```

NOTE: Copy and run the following code if you need to upgrade ipywidgets for datasets library and restart the kernel. This is needed if the installation is not applied to the current kernel.

```
%%capture
import IPython
!conda install -c conda-forge ipywidgets -y
# has to restart kernel for the updates to be applied
IPython.Application.instance().kernel.do_shutdown(True)
```

SageMaker Environment

In [4]:
```
import sagemaker

sess = sagemaker.Session()

# sagemaker session bucket -> used for uploading data, models and l
ogs
# sagemaker will automatically create this bucket if it not exists
sagemaker_session_bucket = None
if sagemaker_session_bucket is None and sess is not None:
    # set to default bucket if a bucket name is not given
    sagemaker_session_bucket = sess.default_bucket()

role = sagemaker.get_execution_role()
sess = sagemaker.Session(default_bucket=sagemaker_session_bucket)

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

```
sagemaker role arn: arn:aws:iam::875423407011:role/AdminRole
sagemaker bucket: sagemaker-us-west-2-875423407011
sagemaker session region: us-west-2
```

SageMaker Training Job

To create a SageMaker training job, we use a HuggingFace/PyTorch estimator. Using the estimator, you can define which training script should SageMaker use through entry_point, which instance_type to use for training, which hyperparameters to pass, and so on.

When a SageMaker training job starts, SageMaker takes care of starting and managing all the required machine learning instances, picks up the HuggingFace Deep Learning Container, uploads your training script, and downloads the data from sagemaker_session_bucket into the container at /opt/ml/input/data.

In [5]:
```
# Here we configure the training job. Please configure the appropri
ate options below:

EPOCHS = 10

# Choose between Causal Language Model and Masked Language Model
LANGUAGE_MODELING_LOSS = "clm"  # or "mlm"
SEQ_LEN_ARG = "block_size" if LANGUAGE_MODELING_LOSS == "clm" else
"max_seq_length"

MODEL_NAME = "gpt2"
TOKENIZER_NAME = "gpt2"

MODEL_CONFIG = "model_type"

# For more information about the options, please look into the trai
ning scripts

# Select Instance type for training
INSTANCE_TYPE = "ml.p4d.24xlarge"
NUM_INSTANCES = 2
# Since ml.p4d.24xlarge instance has 8 GPUs, we set num_gpus_per_in
stance to 8
num_gpus_per_instance = 8
```

First, we define some basic parameters common to all estimators.

Note: We recommend you to turn the SageMaker Debugger's profiling and debugging tools off to avoid additional overheads.

In [6]:
```
estimator_args = dict(
    entry_point=f"run_{LANGUAGE_MODELING_LOSS}_memory.py",
    source_dir="./scripts",
    instance_type=INSTANCE_TYPE,
    instance_count=NUM_INSTANCES,
    role=role,
    py_version="py38",
    volume_size=512,
    disable_profiler=True,  # Disabling SageMaker Profiler to avoid
overheads during benchmarking
    debugger_hook_config=False,  # Disabling SageMaker Debugger to
avoid overheads during benchmarking
    base_job_name="trcomp-pt-example",
    metric_definitions=[
        {"Name": "summary_train_runtime", "Regex": "'train_runtime
': ([0-9.]*)"},
        {
            "Name": "summary_train_samples_per_second",
            "Regex": "'train_samples_per_second': ([0-9.]*)",
        },
        {"Name": "summary_train_steps_per_second", "Regex": "'train
_steps_per_second': ([0-9.]*)"},
        {"Name": "summary_train_loss", "Regex": "'train_loss': ([0-
9.]*)"},
        {"Name": "epoch", "Regex": "'epoch': ([0-9.]*)"},
        {"Name": "train_loss", "Regex": "'loss': ([0-9.]*)"},
        {"Name": "learning_rate", "Regex": "'learning_rate': ([0-
9.]*)"},
    ],
)
```

Next, we define some basic arguments to be passed to the training script.

In [7]:
```
# hyperparameters are passed to the training entrypoint as argument
s
hyperparameters = {
    MODEL_CONFIG: MODEL_NAME,
    "tokenizer_name": TOKENIZER_NAME,
    "dataset_name": "wikitext",
    "dataset_config_name": "wikitext-103-v1",
    "do_train": True,
    "do_eval": False,
    "num_train_epochs": EPOCHS,
    SEQ_LEN_ARG: 512,
    "overwrite_output_dir": True,
    "save_strategy": "no",
    "evaluation_strategy": "no",
    "logging_strategy": "epoch",
    "output_dir": "/opt/ml/model",
    "dataloader_drop_last": True,
    "preprocessing_num_workers": 12,
}
```

In the following sections, we will create estimators and start training.

Training with Native PyTorch + SM DDP

The batch size below is the maximum batch we could fit into the memory of an ml.p4d.24xlarge instance. If you change the model, instance type, sequence length, and other parameters, you need to do some experiments to find the largest batch size that will fit into GPU memory. We also use Automatic Mixed Precision for faster training.

This example uses HuggingFace training script run_clm.py, which you can find it inside the scripts folder.

In [8]:
```
from sagemaker.pytorch import PyTorch

hyperparameters["per_device_train_batch_size"] = 13

# The original LR was set for a batch of 32. Here we are scaling le
arning rate with batch size.
hyperparameters["learning_rate"] = (
    float("5e-5")
    / 32
    * hyperparameters["per_device_train_batch_size"]
)

# configure the training job
native_estimator = PyTorch(
    **estimator_args,
    framework_version="1.11",
    hyperparameters=hyperparameters,
    distribution={
        "smdistributed": {"dataparallel": {"enabled": True}}
    },  # Use SageMaker Distributed Data Parallel to train across n
odes/GPUs.
)

# Start the training job
native_estimator.fit(wait=False)
native_estimator.latest_training_job.name
```

Out[8]:
```
'trcomp-pt-example-2022-10-18-18-06-44-296'
```

Training with SageMaker Training Compiler

Compilation through Training Compiler changes the memory footprint of the model. Most commonly, this manifests as a reduction in memory utilization and a consequent increase in the largest batch size that can fit on the GPU. Note that if you want to change the batch size, you must adjust the learning rate appropriately.

In [9]:
```
from sagemaker.huggingface import HuggingFace, TrainingCompilerConf
ig

# with SageMaker Training Compiler we are able to fit a larger batc
h into memory
hyperparameters["per_device_train_batch_size"] = 25

# The original LR was set for a batch of 32. Here we are scaling le
arning rate with batch size.
hyperparameters["learning_rate"] = (
    float("5e-5")
    / 32
    * hyperparameters["per_device_train_batch_size"]
)

# configure the training job
optimized_estimator = HuggingFace(
    compiler_config=TrainingCompilerConfig(),
    transformers_version="4.21",
    pytorch_version="1.11",
    hyperparameters=hyperparameters,
    distribution={"pytorchxla": {"enabled": True}},
    **estimator_args,
)

# start the training job
optimized_estimator.fit(wait=False)
optimized_estimator.latest_training_job.name
```

Out[9]:
```
'trcomp-pt-example-2022-10-18-18-06-44-970'
```

Wait for training jobs to complete

In [10]:
```
waiter = native_estimator.sagemaker_session.sagemaker_client.get_wa
iter(
    "training_job_completed_or_stopped"
)
waiter.wait(TrainingJobName=native_estimator.latest_training_job.na
me)
waiter = optimized_estimator.sagemaker_session.sagemaker_client.get
_waiter(
    "training_job_completed_or_stopped"
)
waiter.wait(TrainingJobName=optimized_estimator.latest_training_jo
b.name)
```

Analysis

Note: If the estimator object is no longer available due to a kernel break or refresh, you need to directly use the training job name and manually attach the training job to a new HuggingFace estimator. For example:

```
estimator = HuggingFace.attach("your_huggingface_training_job_name")
```

Load logs of the training job *with* SageMaker Training Compiler

In [11]:
```
%%capture optimized

# access the logs of the optimized training job
optimized_estimator.sagemaker_session.logs_for_job(optimized_estimator.latest_training_job.name)
```

Load logs of the training job *without* SageMaker Training Compiler

In [12]:
```
%%capture native

# access the logs of the native training job
native_estimator.sagemaker_session.logs_for_job(native_estimator.latest_training_job.name)
```

Create helper functions for analysis

In [13]:
```
from ast import literal_eval
from collections import defaultdict
from matplotlib import pyplot as plt


def _summarize(captured):
    final = []
    for line in captured.stdout.split("\n"):
        cleaned = line.strip()
        if "{" in cleaned and "}" in cleaned:
            final.append(cleaned[cleaned.index("{") : cleaned.index
("}") + 1])
    return final


def make_sense(string):
    try:
        return literal_eval(string)
    except:
        pass


def summarize(summary):
    final = {"train": [], "eval": [], "summary": {}}
    for line in summary:
        interpretation = make_sense(line.replace("nan", "'nan'"))
        if interpretation:
            if "loss" in interpretation:
                final["train"].append(interpretation)
            elif "eval_loss" in interpretation:
                final["eval"].append(interpretation)
            elif "train_runtime" in interpretation:
                final["summary"].update(interpretation)
    return final
```

Plot Optimized vs Native Training Throughput

Visualize average throughputs as reported by HuggingFace and see potential savings.

In [14]:
```
# Average throughput for the native PyTorch training as reported by
Trainer
n = summarize(_summarize(native))
native_throughput = n["summary"]["train_samples_per_second"]

# Average throughput for the optimized PyTorch training as reported
by Trainer
o = summarize(_summarize(optimized))
optimized_throughput = o["summary"]["train_samples_per_second"]

# Calculate percentage speedup of optimized PyTorch over native PyT
orch
avg_speedup = f"{round((optimized_throughput/native_throughput-1)*1
00)}%"
```

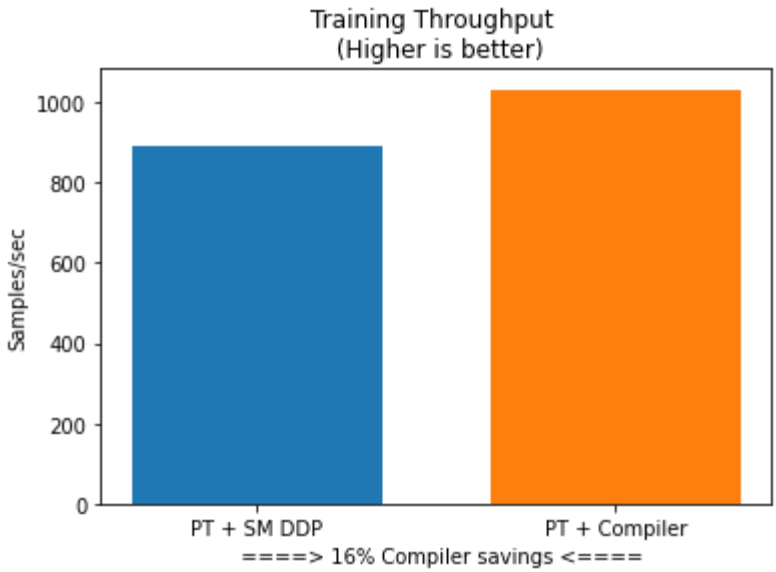
In [15]:
```
%matplotlib inline

plt.title("Training Throughput \n (Higher is better)")
plt.ylabel("Samples/sec")

plt.bar(x=[1], height=native_throughput, label="PT + SM DDP", width
=0.35)
plt.bar(x=[1.5], height=optimized_throughput, label="PT + Compile
r", width=0.35)

plt.xlabel("  ====> {} Compiler savings <====".format(avg_speedup))
plt.xticks(ticks=[1, 1.5], labels=["PT + SM DDP", "PT + Compiler"])
```

Out[15]:
```
([<matplotlib.axis.XTick at 0x7f81de5eb880>,
  <matplotlib.axis.XTick at 0x7f81de5eb850>],
 [Text(1.0, 0, 'PT + SM DDP'), Text(1.5, 0, 'PT + Compiler')])
```



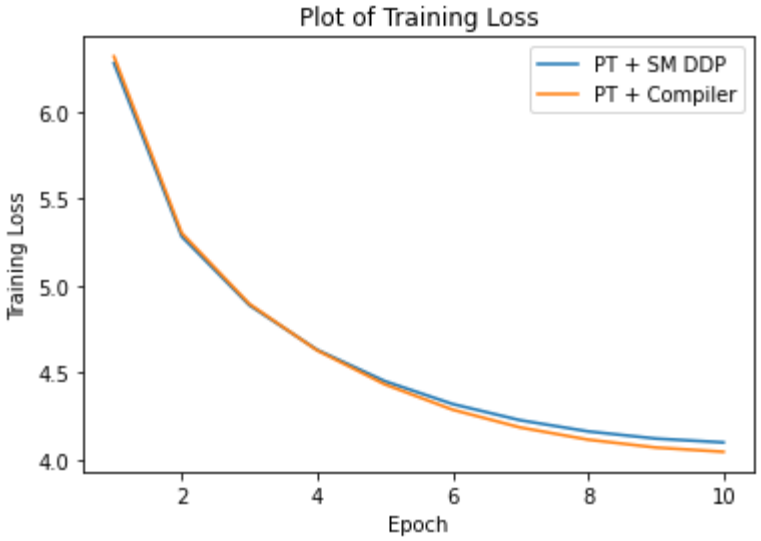
Convergence of Training Loss

SageMaker Training Compiler does not affect the model convergence behavior. Here, we see the decrease in training loss is similar with and without SageMaker Training Compiler

In [16]:
```
vanilla_loss = [i["loss"] for i in n["train"]]
vanilla_epochs = [i["epoch"] for i in n["train"]]
optimized_loss = [i["loss"] for i in o["train"]]
optimized_epochs = [i["epoch"] for i in o["train"]]

plt.title("Plot of Training Loss")
plt.xlabel("Epoch")
plt.ylabel("Training Loss")
plt.plot(vanilla_epochs, vanilla_loss, label="PT + SM DDP")
plt.plot(optimized_epochs, optimized_loss, label="PT + Compiler")
plt.legend()
```

Out[16]:
```
<matplotlib.legend.Legend at 0x7f81da73e790>
```



Training Stats

Let's compare various training metrics with and without SageMaker Training Compiler. SageMaker Training Compiler provides an increase in training throughput which translates to a decrease in total training time.

In [17]:
```
import pandas as pd
pd.DataFrame([n["summary"], o["summary"]], index=["PT + SM DDP", "PT + Compiler"])
```

Out[17]:

	train_runtime	train_samples_per_second	train_steps_per_second	train_loss	epoch
PT + SM DDP	2570.5208	891.504	4.283	4.645593	10.0
PT + Compiler	2218.8958	1032.779	2.578	4.627251	10.0

In [18]:
```
speedup = (
    (n["summary"]["train_runtime"] - o["summary"]["train_runtime"])
    * 100
    / n["summary"]["train_runtime"]
)
print(
    f"SageMaker Training Compiler is about {int(speedup)}% faster in terms of total training time."
)
```

```
SageMaker Training Compiler is about 13% faster in terms of total training time.
```

Total Billable Time

Finally, the decrease in total training time results in a decrease in the billable seconds from SageMaker.

In [19]:
```
def BillableTimeInSeconds(name):
    describe_training_job = (
        optimized_estimator.sagemaker_session.sagemaker_client.describe_training_job
    )
    details = describe_training_job(TrainingJobName=name)
    return details["BillableTimeInSeconds"]
```

In [20]:
```
Billable = {}
Billable["PT + SM DDP"] = BillableTimeInSeconds(native_estimator.latest_training_job.name)
Billable["PT + Compiler"] = BillableTimeInSeconds(optimized_estimator.latest_training_job.name)
pd.DataFrame(Billable, index=["BillableSecs"])
```

Out[20]:

	PT + SM DDP	PT + Compiler
BillableSecs	3294	2875

In [21]:
```
speedup = (Billable["PT + SM DDP"] - Billable["PT + Compiler"]) * 1
00 / Billable["PT + SM DDP"]
print(f"SageMaker Training Compiler integrated PyTorch was {int(spe
edup)}% faster in summary.")
```

```
SageMaker Training Compiler integrated PyTorch was 12% faster in sum
mary.
```

Clean up

Stop all training jobs launched if the jobs are still running.

In [22]:
```
import boto3

sm = boto3.client("sagemaker")


def stop_training_job(name):
    status = sm.describe_training_job(TrainingJobName=name)["Traini
ngJobStatus"]
    if status == "InProgress":
        sm.stop_training_job(TrainingJobName=name)


stop_training_job(native_estimator.latest_training_job.name)
stop_training_job(optimized_estimator.latest_training_job.name)
```

Also, to find instructions on cleaning up resources, see Clean Up (https://docs.aws.amazon.com/sagemaker/latest/dg/ex1-cleanup.html) in the *Amazon SageMaker Developer Guide*.