Introducing Amazon SageMaker Training Compiler

Michael Lin

Senior Solutions Architect
AWS

Deep learning models have a size problem

Large datasets take a long time to train, creating a bottleneck

Training costs are an obstacle to experimentation and innovation

Optimizing for cost and speed is challenging

P4d

Infrastructure

Distributed training

Skills and time

Compilers offer efficiency gains

Convert high-level representation to hardware-optimized instructions

Enable more efficient use of hardware without scaling out or up

Dedicated compilers for inference offer performance gains of 25x

NEW

Amazon SageMaker Training Compiler

The fast and easy way to train large deep learning models on GPUs

GENERALLY AVAILABLE

Accelerate deep learning model training

Speed up training by as much as 50%

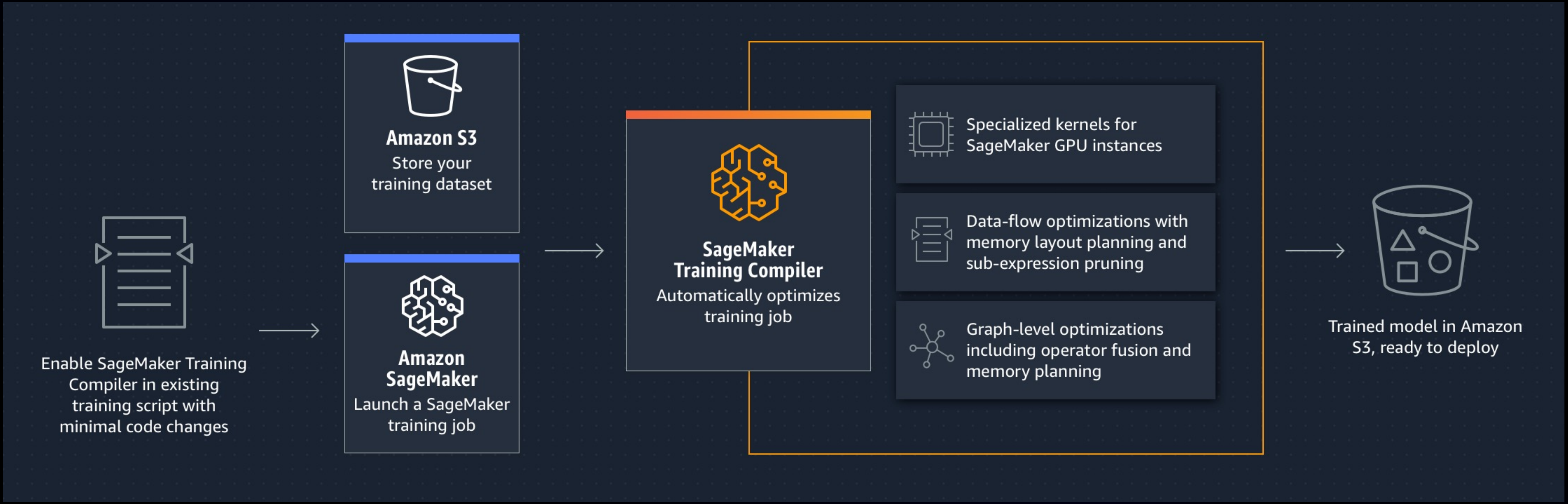
Minimal code changes required

Enable in minutes without any changes to workflow

Lower training costs

Free to use on SageMaker and additional savings from shortened training jobs

Acceleration without workflow disruption



Enable SageMaker Training Compiler in existing training script with minimal code changes

Amazon S3
Store your training dataset

Amazon SageMaker
Launch a SageMaker training job

SageMaker Training Compiler
Automatically optimizes training job

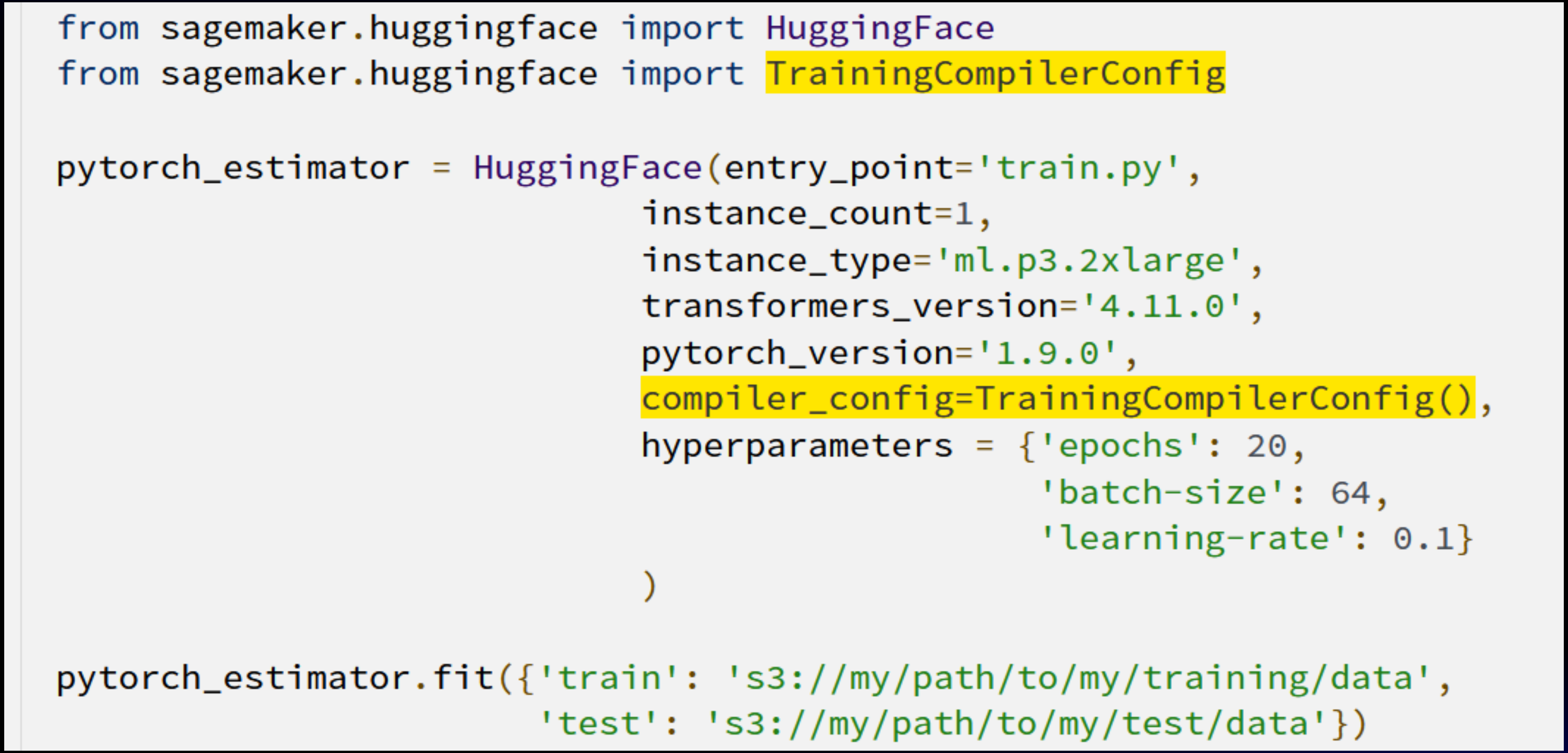
Specialized kernels for SageMaker GPU instances

Data-flow optimizations with memory layout planning and sub-expression pruning

Graph-level optimizations including operator fusion and memory planning

Trained model in Amazon S3, ready to deploy

SageMaker Training Compiler can be enabled in minutes



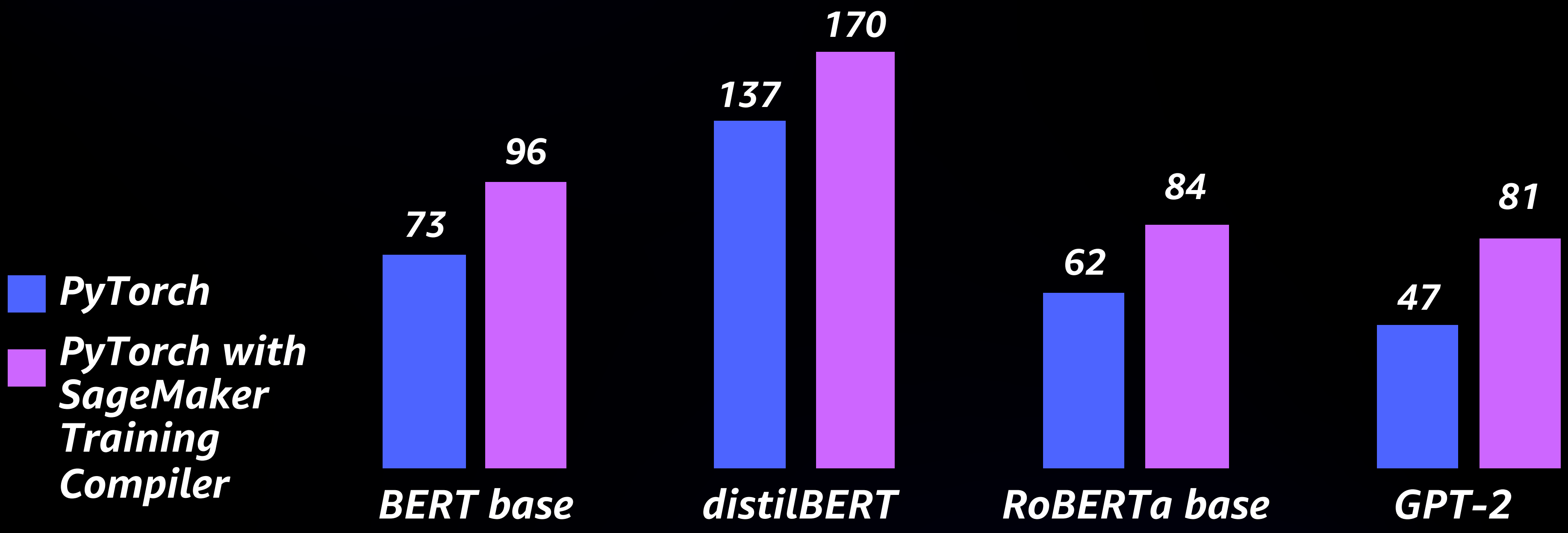
```python
from sagemaker.huggingface import HuggingFace
from sagemaker.huggingface import TrainingCompilerConfig

pytorch_estimator = HuggingFace(entry_point='train.py',
                            instance_count=1,
                            instance_type='ml.p3.2xlarge',
                            transformers_version='4.11.0',
                            pytorch_version='1.9.0',
                            compiler_config=TrainingCompilerConfig(),
                            hyperparameters = {'epochs': 20,
                                               'batch-size': 64,
                                               'learning-rate': 0.1}
                            )

pytorch_estimator.fit({'train': 's3://my/path/to/my/training/data',
                       'test': 's3://my/path/to/my/test/data'})
```

Up to 50% speedup with SageMaker Training Compiler

Training sample throughput[1] *(samples/second)*



Reduction in billable time	***38%***	***26%***	***27%***	***54%***

[1] Test parameters: ml.p3.2xlarge, PyTorch with Hugging Face Trainer API, 25 epochs, sequence length of 512
Baseline used the Hugging Face AWS Deep Learning Container from Amazon ECR

SageMaker Training Compiler accelerates the most popular NLP models

- bert-base-uncased
- bert-large-uncased
- roberta-base
- gpt2
- bert-base-cased
- xlm-roberta-base
- bert-base-chinese
- roberta-large
- distilbert-base-uncased
- distilbert-base-uncased-finetuned-sst-2-English
- cl-tohoku/bert-base-japanese-whole-word-masking
- bert-base-multilingual-cased
- distilgpt2
- albert-base-v2
- gpt2-large

Thank you!

Michael Lin

linmicht@amazon.com