AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

AIM306

Introducing Amazon SageMaker Training Compiler

Pinar Yilmaz
Principal Engineer, Amazon SageMaker
AWS

Jeff Boudier
Product Director
Hugging Face

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

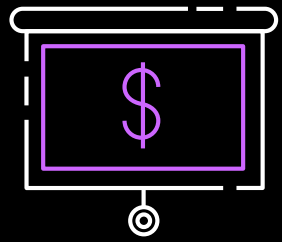
- Challenges in training deep learning models
- Introducing Amazon SageMaker Training Compiler
- Spotlight: Hugging Face
- Demo
- Wrap-up and Q&A

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deep learning models have a size problem



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

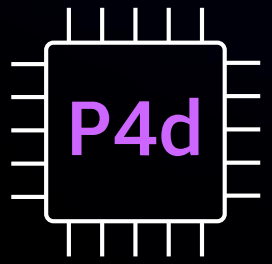


Training costs are an obstacle to experimentation and innovation

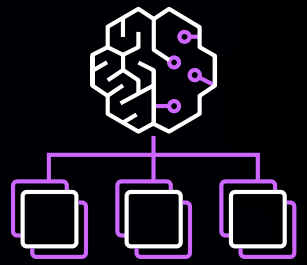
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

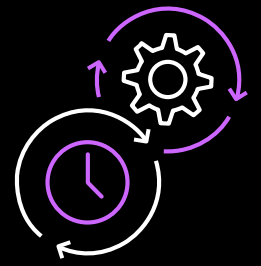
Optimizing for cost and speed is challenging



Infrastructure



Distributed training

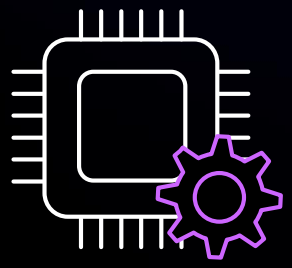


Skills and time

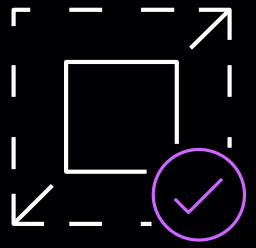
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

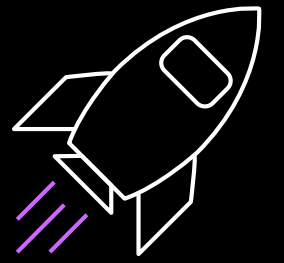
Compilers offer efficiency gains



Convert high-level representation to hardware-optimized instructions



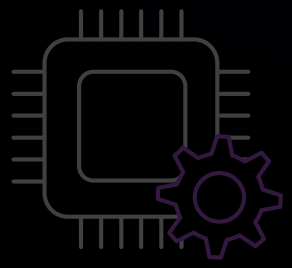
Enable more efficient use of hardware without scaling out or up



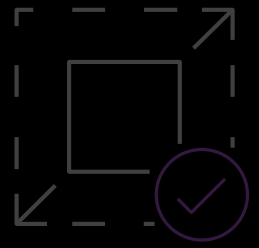
Dedicated compilers for inference offer performance gains of 25x

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

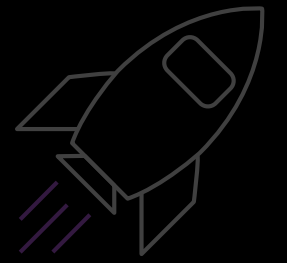
Compilers offer efficiency gains



Convert high-level representation to hardware-optimized instructions



Enable more efficient use of hardware without scaling out or up

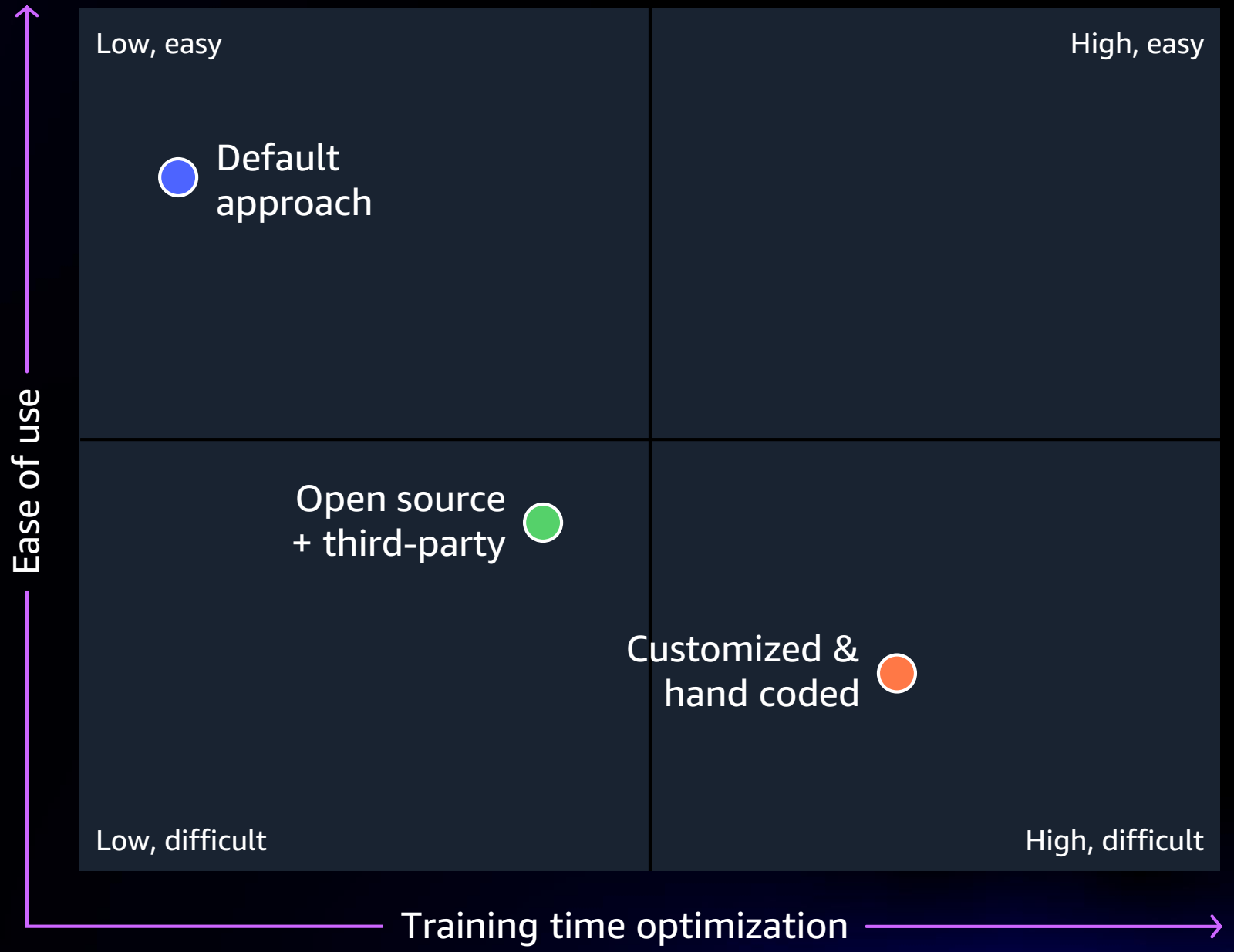


Dedicated compilers for inference offer performance gains of 25x

No good built-in solution for training available

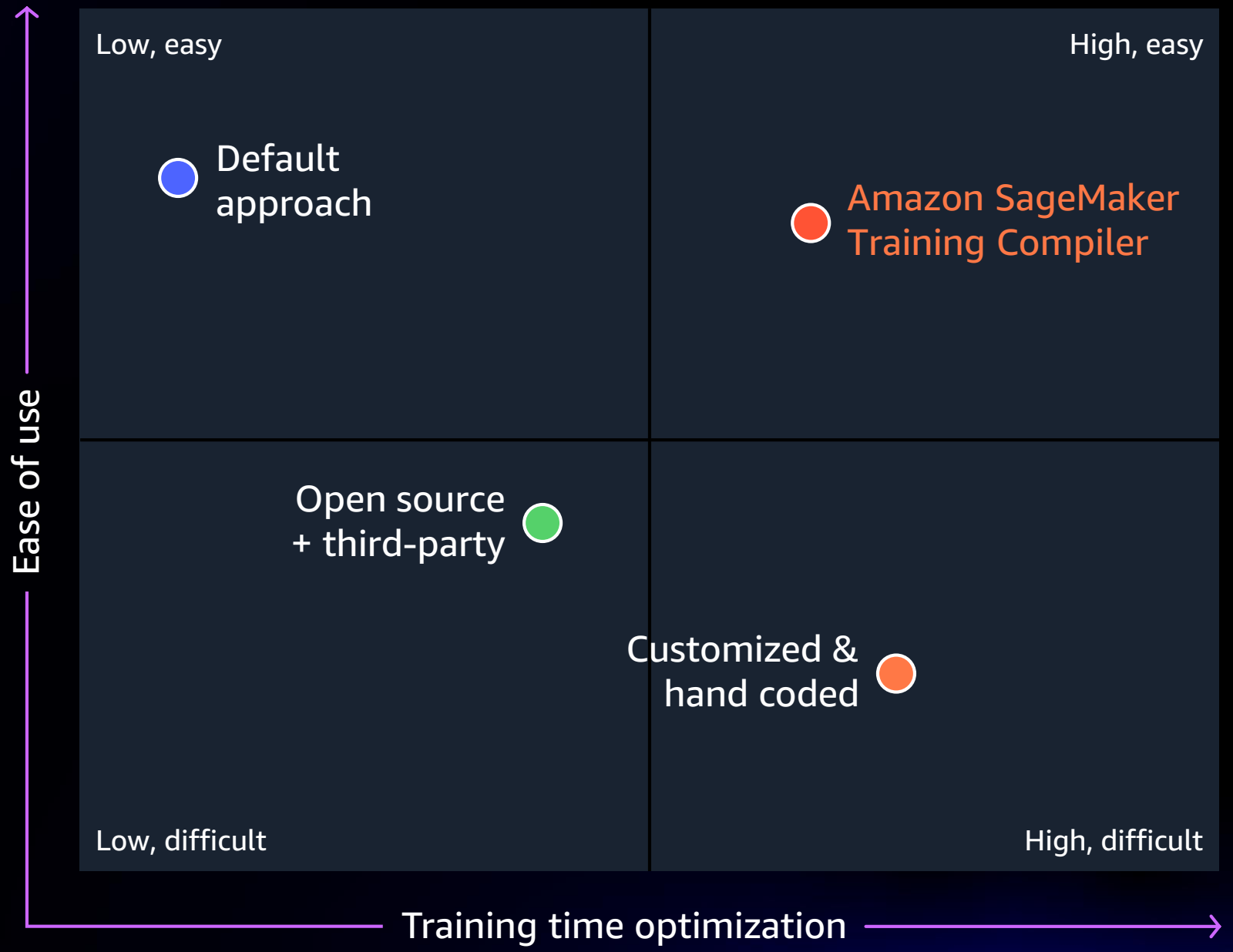
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Compilation for deep learning training is a challenge



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Compilation for deep learning training is a challenge



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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



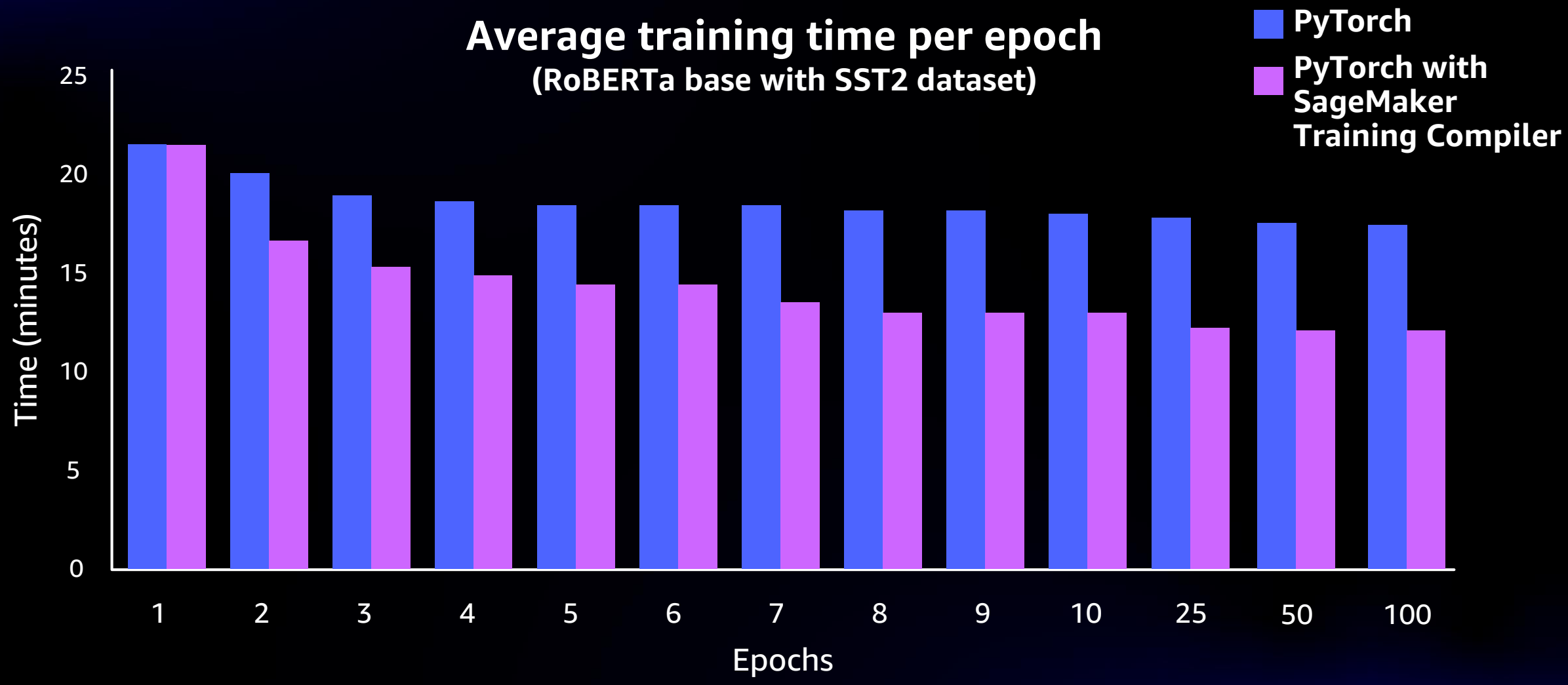
Lower training costs

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

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

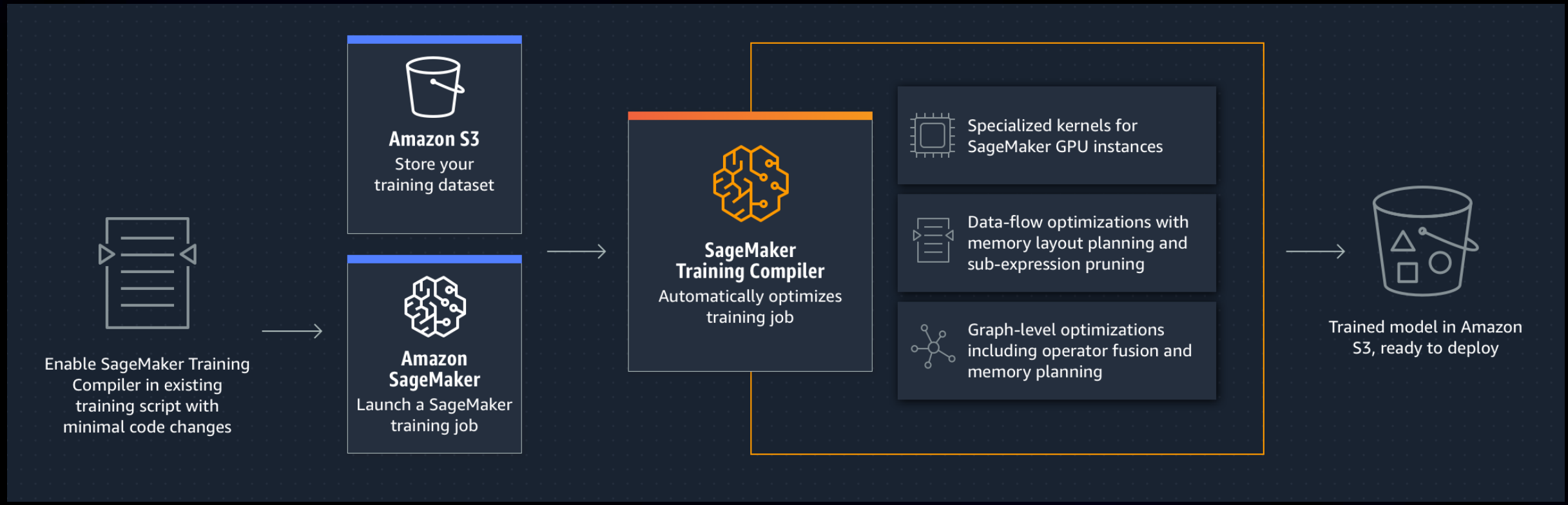
Specialized compilation accelerates training



aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Acceleration without workflow disruption



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Training Compiler can be enabled in minutes

```
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}
                                )

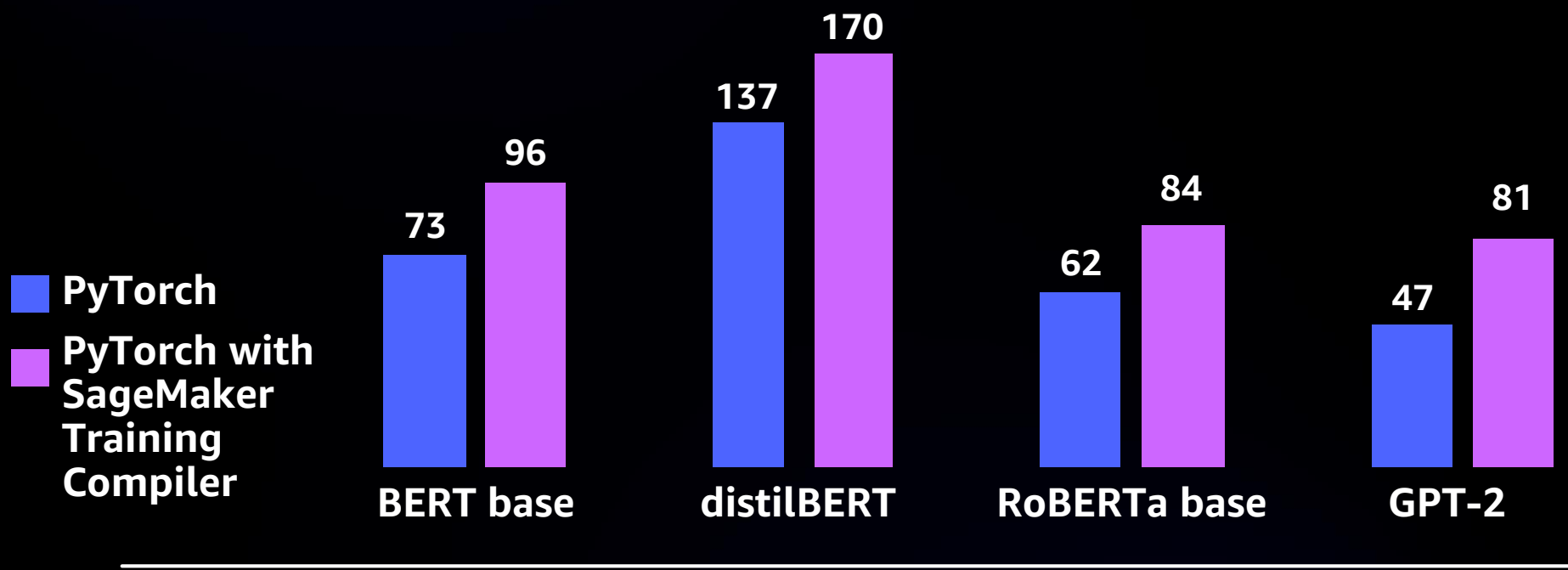
pytorch_estimator.fit({'train': 's3://my/path/to/my/training/data',
                       'test': 's3://my/path/to/my/test/data'})
```

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Up to 50% speedup with SageMaker Training Compiler

Training sample throughput[1] (samples/second)

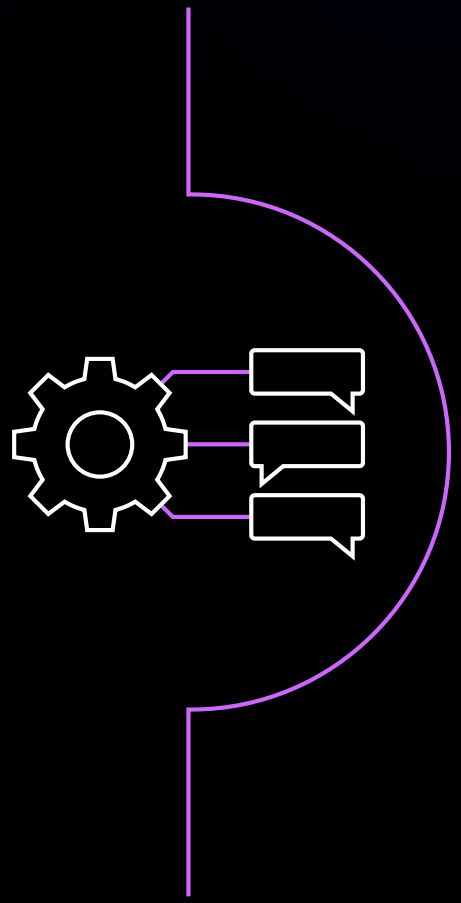


	BERT base	distilBERT	RoBERTa base	GPT-2
Reduction in billable time	38%	26%	27%	54%

[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

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Guidewire is the platform P&C insurers trust to engage, innovate, and grow efficiently; the company combines digital, core, analytics, and AI to deliver its platform as a cloud service, and it enables its customers to do advanced analytics and machine learning for their industry-specific workloads; more than 450 insurers, from new ventures to the largest and most complex in the world, run on Guidewire

One of Guidewire's services is to help customers develop cutting-edge NLP models for applications like risk assessment and claims operations. Amazon SageMaker Training Compiler is compelling because it offers time and cost savings to our customers while developing these NLP models. We expect it to help us reduce training time by more than 20% through more efficient use of GPU resources. We are excited to implement SageMaker Training Compiler in our NLP workloads, helping us to accelerate the transformation of data to insight for our customers."

Matt Pearson, Principal Product Manager, Analytics and Data Services, Guidewire Software

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

musixmatch

Leading music data company providing data, tools, and services that enrich the way we experience music; the largest in the world with over 80 million users and over 8 million distinct lyrics

We are always looking for ways to accelerate training time while also lowering training costs, which is why we are excited about Amazon SageMaker Training Compiler. SageMaker Training Compiler provides more efficient ways to use GPUs during the training process and, with the seamless integration between SageMaker Training Compiler, PyTorch, and high-level libraries like Hugging Face, we have seen a significant improvement in training time of our transformer-based models – going from weeks to days – as well as lower training costs."

Loreto Parisi, AI Engineering Director, Musixmatch

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Quantum Health is on a mission to make healthcare navigation smarter, simpler, and more cost-effective for everyone; they use Amazon SageMaker for use cases like text classification, text summarization, predictive models, classification problems, and Q&A to help the Quantum team and the members they serve

Iterating with NLP models can be a challenge because of their size; long training times bog down workflows, and high costs can discourage our team from trying larger models that might offer better performance. Amazon SageMaker Training Compiler is exciting because it has the potential to alleviate these frictions. Achieving a speedup with SageMaker Training Compiler is a real win for our team that will make us more agile and innovative moving forward."

Jorge Lopez Grisman, Senior Data Scientist, Quantum Health

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Spotlight: Hugging Face

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Spotlight: Hugging Face



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Democratizing good machine learning

Open

Community-driven

Ethics-first

Accelerating Science Production

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Making state of the art accessible

Open Source

- transformers, datasets, tokenizers
- 1,000+ contributors
- 70K GitHub stars
- fastest OSS adoption ever
- NLP, Speech, Vision

Hugging Face Hub

- 20,000+ public models
- 1,600+ public datasets
- 1,000+ organizations
- 200 languages
- 20+ libraries
- NLP, Speech, Vision

Train with AutoNLP

Serve with Inference API

Demo with Spaces

Deploy with Infinity

Hugging Face on Amazon SageMaker

Train with Hugging Face Training DLC | Python SDK *estimator*() | Model Parallelism | Data Parallelism
Deploy with Hugging Face Inference DLC | Hugging Face Inference Toolkit

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tasks

Fill-Mask | Question Answering | Summarization | Table Question Answering | Text Classification | Text Generation | Text2Text Generation | Token Classification | Translation | Zero-Shot Classification | Sentence Similarity | + 12

Libraries

PyTorch | TensorFlow | JAX | + 21

Datasets

wikipedia | common_voice | bookcorpus | dcep europarl jrc-acquis | glue | conll2003 | squad | oscar | + 606

Languages

en | es | fr | de | sv | zh | fi | ja | + 164

Licenses

apache-2.0 | mit | cc-by-4.0 | + 15

Other

Models 20,273

Search Models

↑↓ Sort: Most Downloads

bert-base-uncased
Fill-Mask • Updated May 18 • 28M • ♡ 53

roberta-large
Fill-Mask • Updated May 21 • 13.7M • ♡ 20

distilbert-base-uncased
Fill-Mask • Updated Aug 29 • 5.19M • ♡ 26

xlm-roberta-base
Fill-Mask • Updated Sep 16 • 4.72M • ♡ 11

bert-base-cased
Fill-Mask • Updated Sep 6 • 3.95M • ♡ 5

gpt2
Text Generation • Updated May 19 • 3.47M • ♡ 22

distilbert-base-uncased-finetuned-sst-2-english
Text Classification • Updated Feb 8 • 3.34M • ♡ 17

roberta-base
Fill-Mask • Updated Jul 6 • 3.32M • ♡ 6

>10,000 companies use 🤗 Transformers

Business case Machine Learning Task	Model input	Model output
Customer Support automation Text Classification/Topic Modeling/Sentiment Analysis	“I’m very unhappy about my new credit card”	Priority: high Sentiment: unhappy Department: credit cards
Semantic Search Feature Extraction/Re-ranking	“blue dress shirts for men”	semantically relevant search results
UGC Monitoring Binary text classification	“bruh quit whining like a brat”	offensive = True
Localization Translation	“I love this movie”	“J’adore ce film”
Generating insights from documents Summarization/Text Generation/NLU		bankruptcy risk = 20%
Conversation transcription Automatic Speech Recognition		full text transcript

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Hugging Face + SageMaker for the win

“Hugging Face allows for that one-stop shop for Prophia, where we can grab pre-trained models from the hub, fine-tune and then deploy them on SageMaker.”



“Hugging Face makes transformers, which are traditionally hard to use, easy.”



aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

State of the art ML at scale is challenging

Complex tasks require increasingly large models

- BERT-Large (2018): 340 million parameters
- BigScience T0pp, Allen AI Macaw (2021): 11 billion parameters (that's 32x)

Training these models is very compute intensive

- Datasets include hundreds of billions of tokens (The Pile: 800 GB)
- BigScience: millions of GPU-hours, i.e., weeks of training on thousands of GPUs
- Transfer learning involves lots of training jobs
  - One initial pre-training job, many fine-tuning jobs
  - A bank could pre-train an initial model on decades of SEC EDGAR filings
  - And then fine-tune different versions for specific task types, industries, etc.

Managing and scaling tools and infrastructure is hard work

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Training Compiler accelerates large-scale training jobs

Model compilation increases efficiency and allows larger batch sizes

We tested long classification tasks with BERT, DistilBERT and RoBERTa

- Batch size can increase up to 33%
- Training can be up to 1.4x faster

The longer your training job, the larger the benefit

- 30 minutes seems to be the sweet spot to offset model compilation time
- Initial pre-training jobs are excellent candidates

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Training Hugging Face models on SageMaker

Out-of-the-box acceleration with SageMaker Training Compiler

- The Hugging Face Deep Learning Container supports SageMaker Training Compiler
- No code change: just use the Transformers Trainer API

First-party integration in SageMaker

- The SageMaker Python SDK includes a Hugging Face Estimator
- Training infrastructure is fully managed
- Model Parallelism and Data Parallelism are supported
- Capabilities like model debugging and profiling improve productivity

Hugging Face on SageMaker is one of the easiest and most scalable way to train large models: https://hf.co/docs/sagemaker/

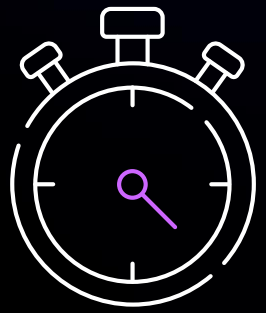
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

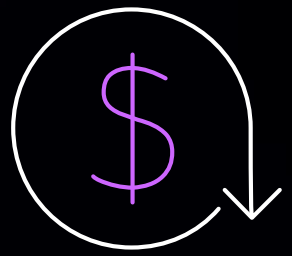
Wrap-up

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

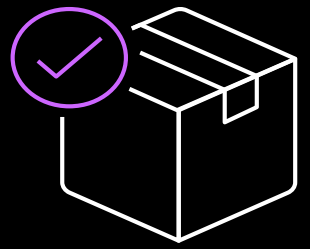
SageMaker Training Compiler in summary



Accelerate DL model training by up to 50%



Free to use on Amazon SageMaker



Generally available at re:Invent ’21

us-east-1, us-east-2

us-west-2, eu-west-1

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Resources

- Example notebooks https://github.com/aws/amazon-sagemaker-examples/tree/master/sagemaker-training-compiler/
- Technical documentation https://docs.aws.amazon.com/sagemaker/latest/dg/training-compiler.html
- Amazon SageMaker Model Training web page https://aws.amazon.com/sagemaker/train

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Pinar Yilmaz
yilmazp@amazon.com

Jeff Boudier
jeff@huggingface.co

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.