INNOVATE

AI/ML EDITION

24 February 2022

Using Hugging Face models on Amazon SageMaker

Praveen Jayakumar

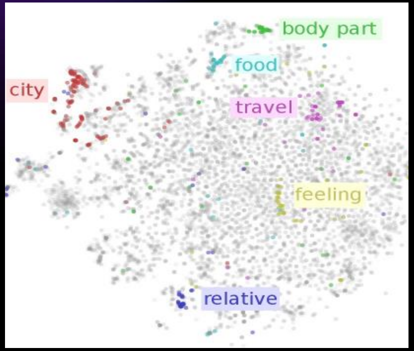
Principal Solutions Architect
AISPL

Agenda

1. What is Transformer architecture
2. Overview of Hugging Face
3. Amazon SageMaker integration with Hugging Face
4. Training Hugging Face model using Amazon SageMaker
5. Deployment options for Hugging Face model in Amazon SageMaker

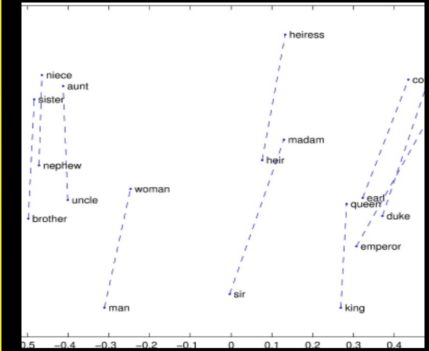
Evolution of NLP algorithms

Word2Vec (2013)



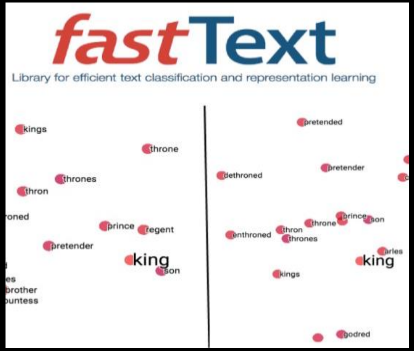
Simple NN
Predict the word based on the context window of other words in the sentence

GloVe (2014)



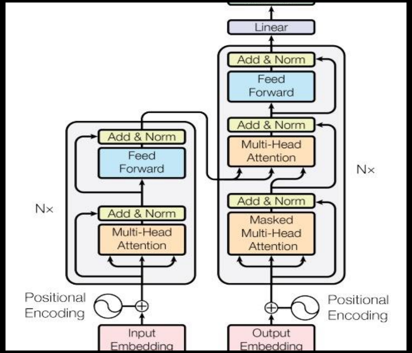
Global Vectors for Word Representation
Matrix factorization

FastText (2015)

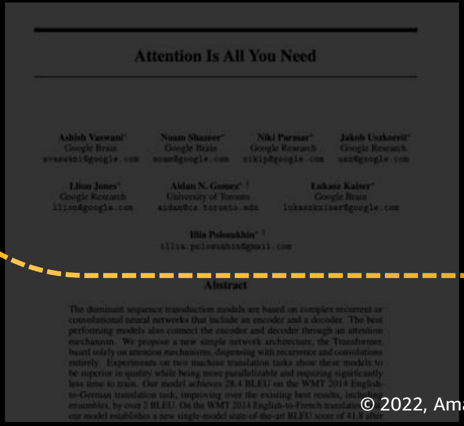


Extension of Word2Vec
Each word is treated as a set of sub-words

Transformer (2017)



Attention Is All You Need



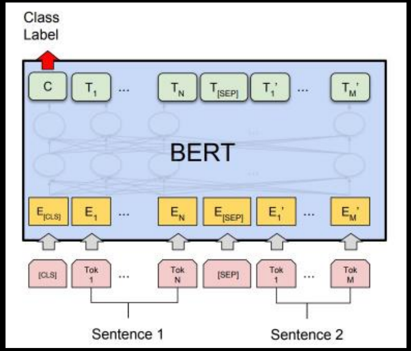
Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

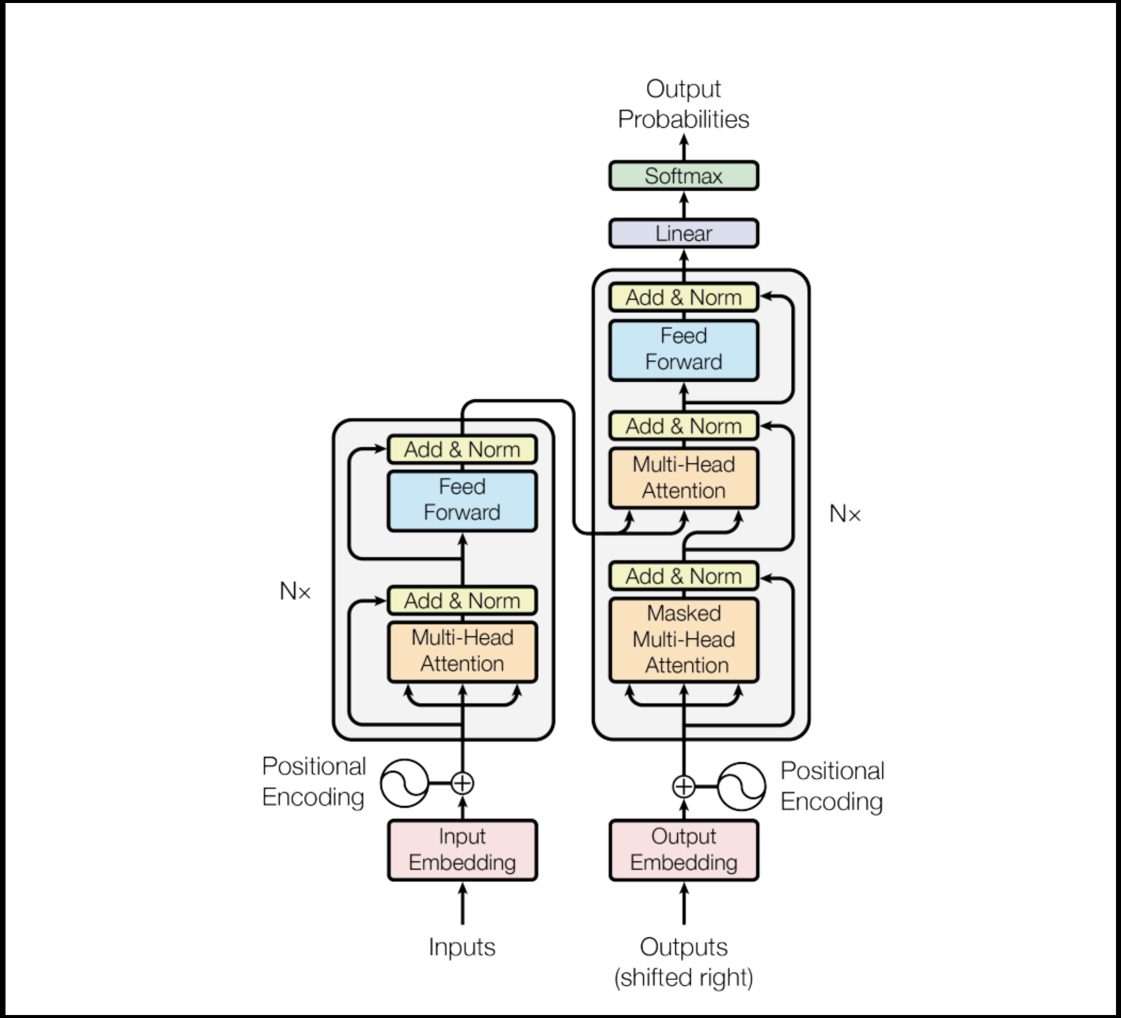
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after

BERT (2018)



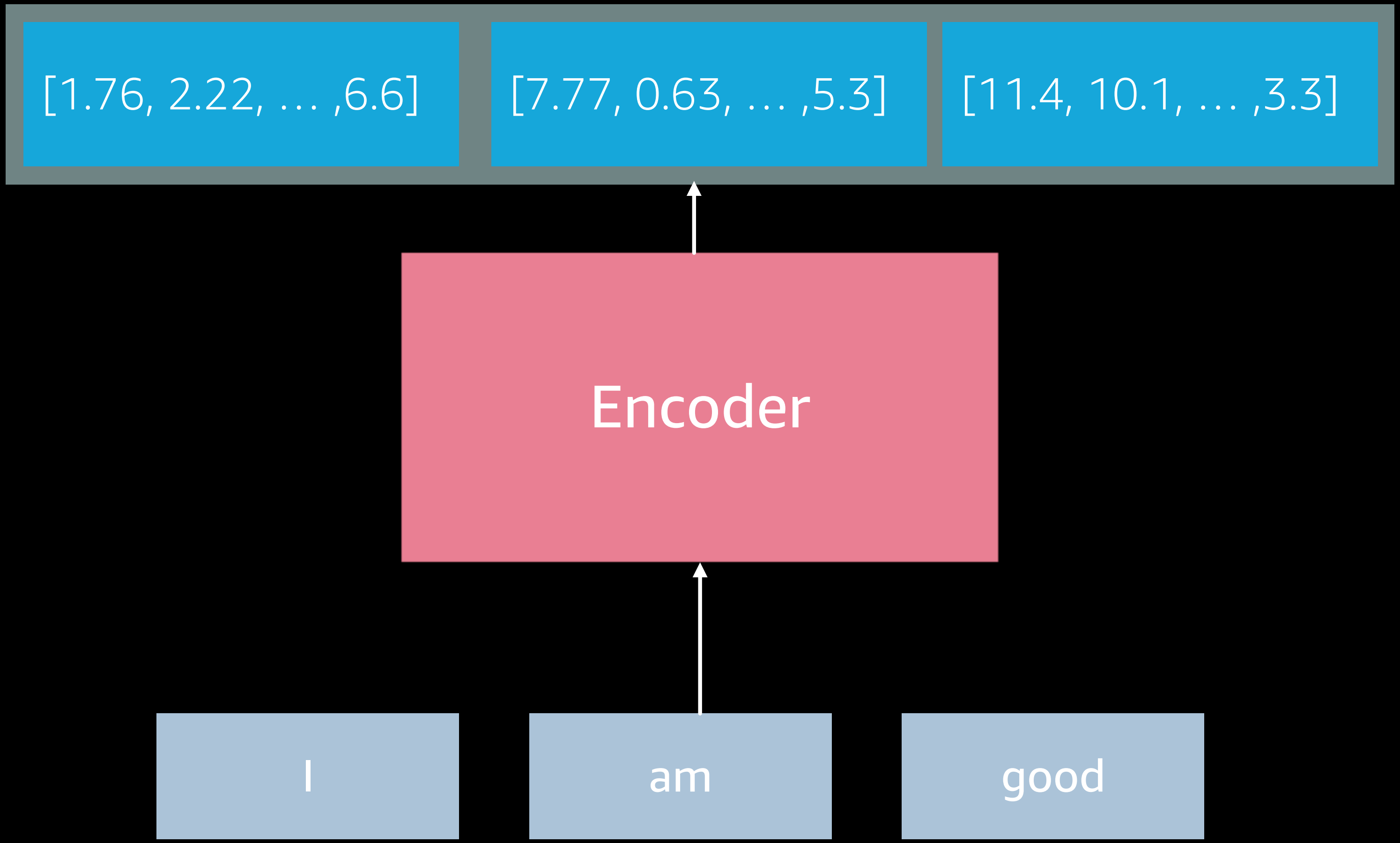
Pre-training of Deep Bidirectional Transformers for Language Understanding

Transformer architecture

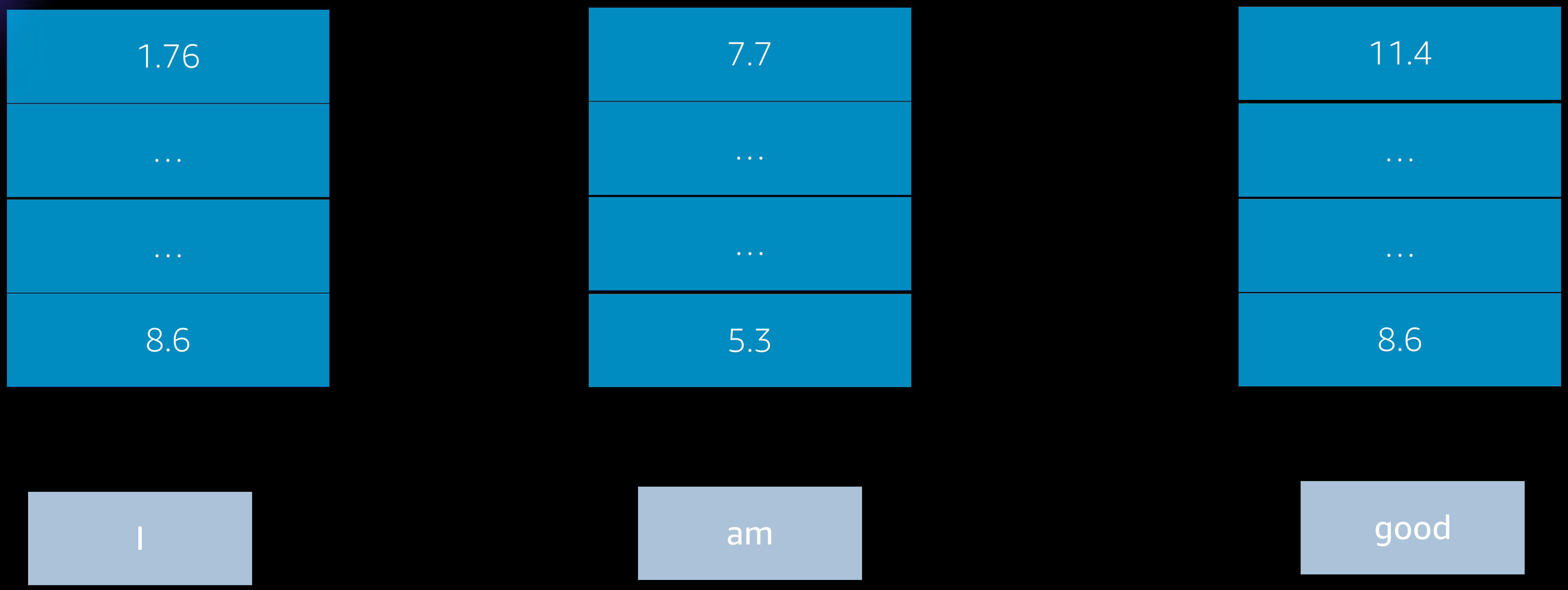


Source link - Attention is all you need

Transformer architecture - Encoder



Transformer architecture - Encoder



1.76

...

...

8.6

I

7.7

...

...

5.3

am

11.4

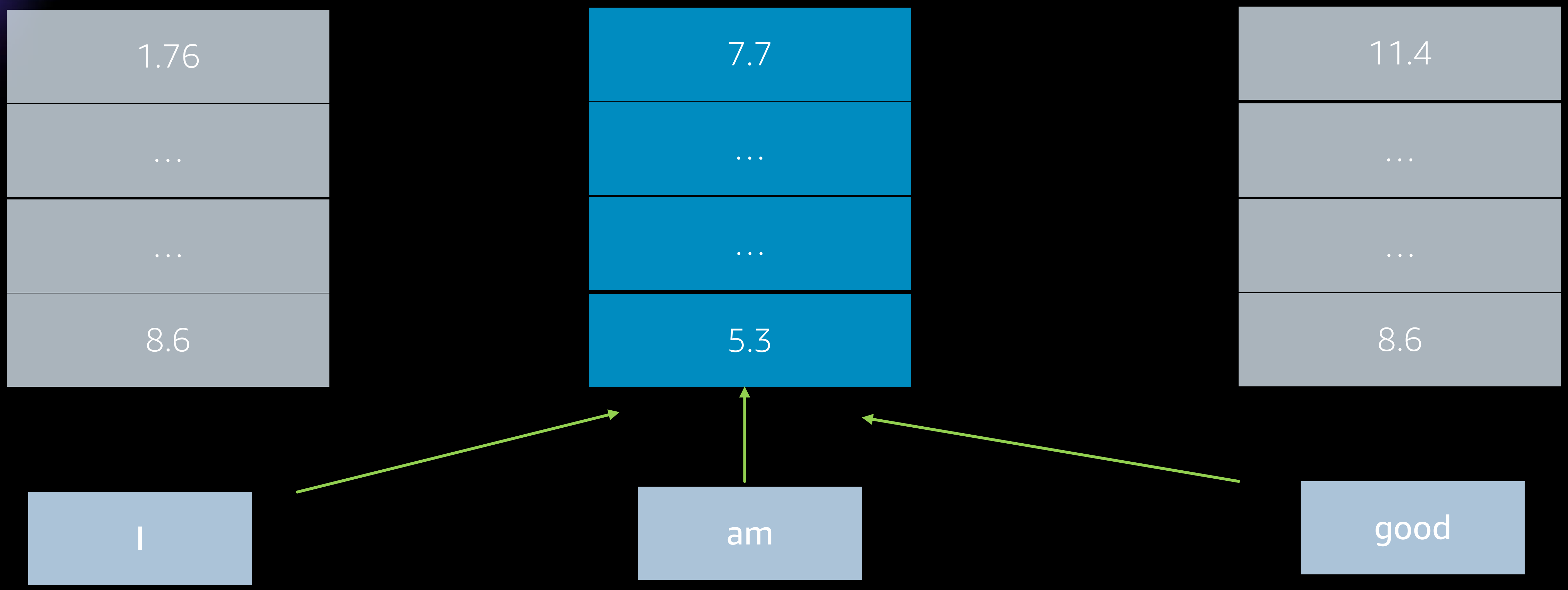
...

...

8.6

good

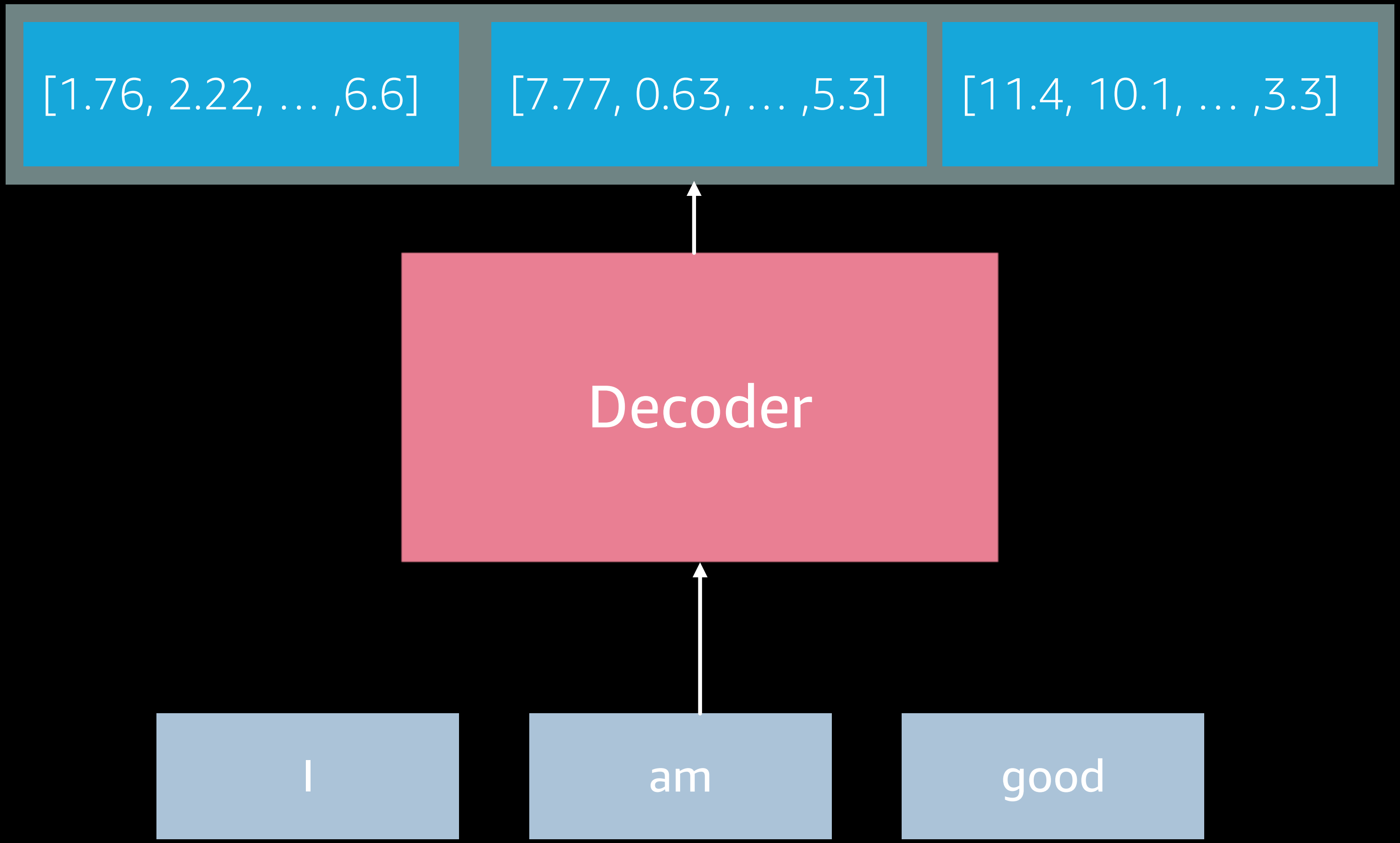
Transformer architecture - Encoder



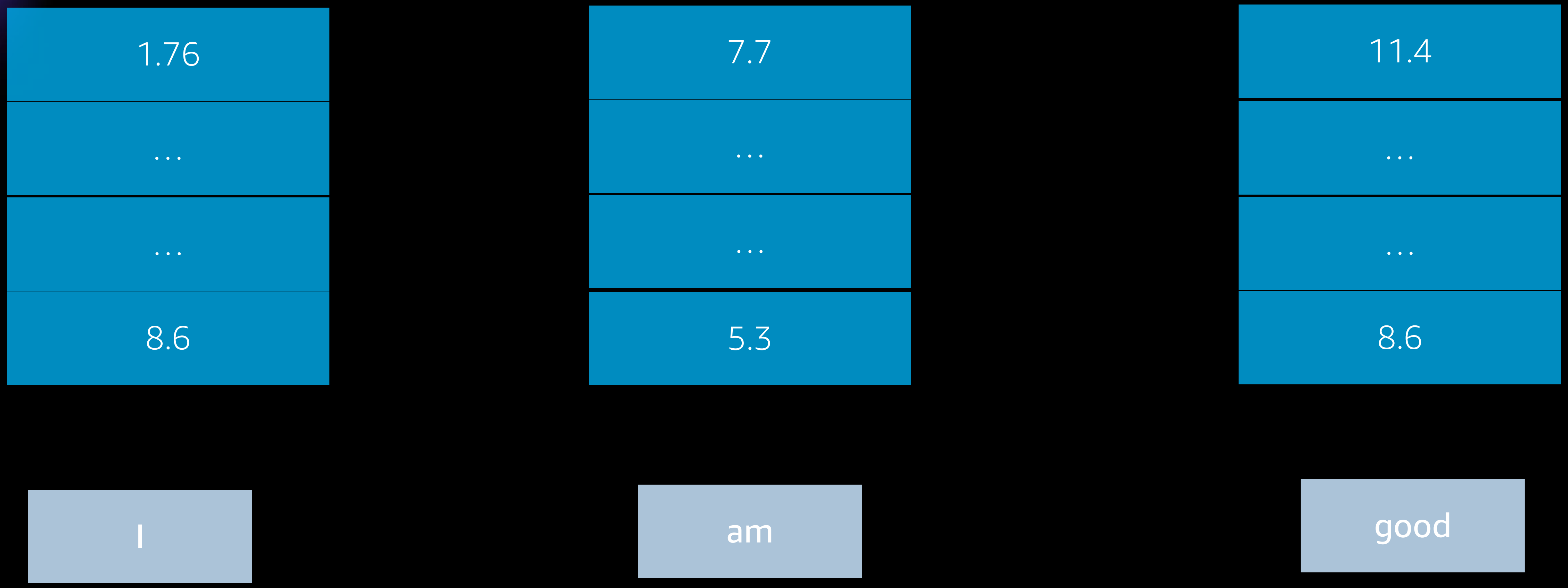
Transformer architecture - Encoder

- Encoders can be used as stand-alone model
- Bi-directional
- Good at extracting meaningful information
- Sequence classification, question answering, masked language modeling
- Example of encoders: BERT, RoBERTa, ALBERT

Transformer architecture - Decoder



Transformer architecture - Decoder



1.76

...

...

8.6

I

7.7

...

...

5.3

am

11.4

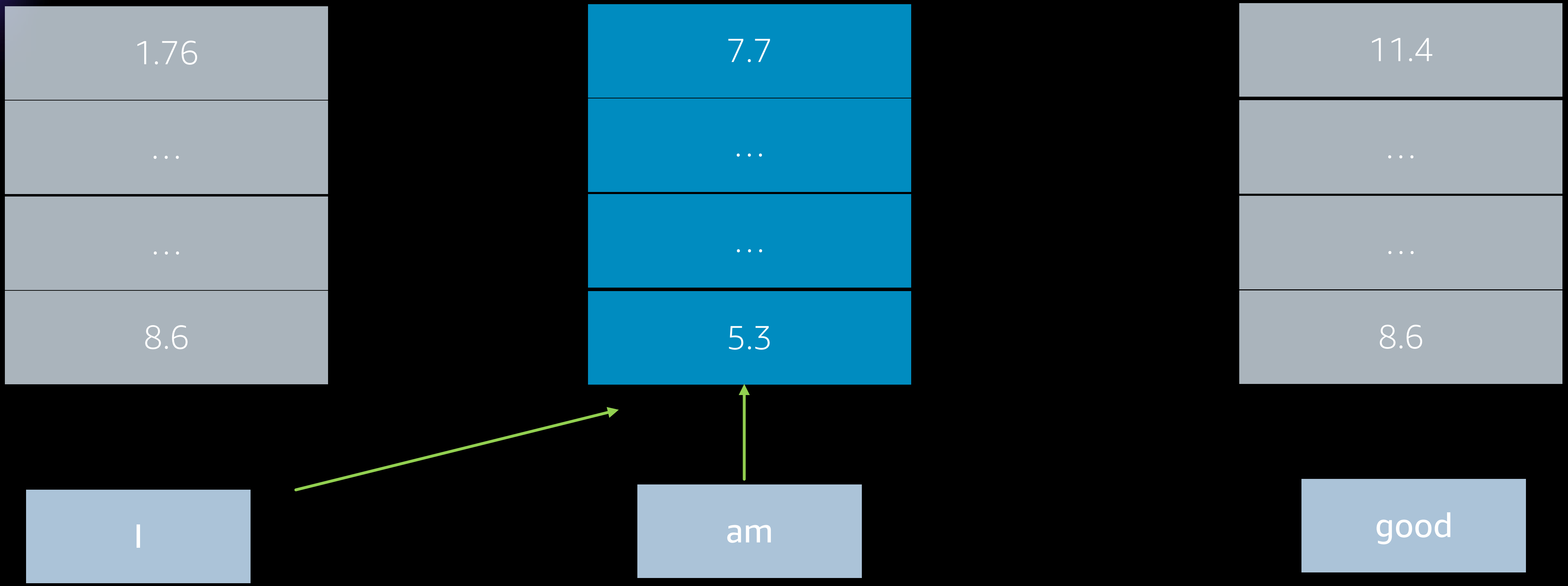
...

...

8.6

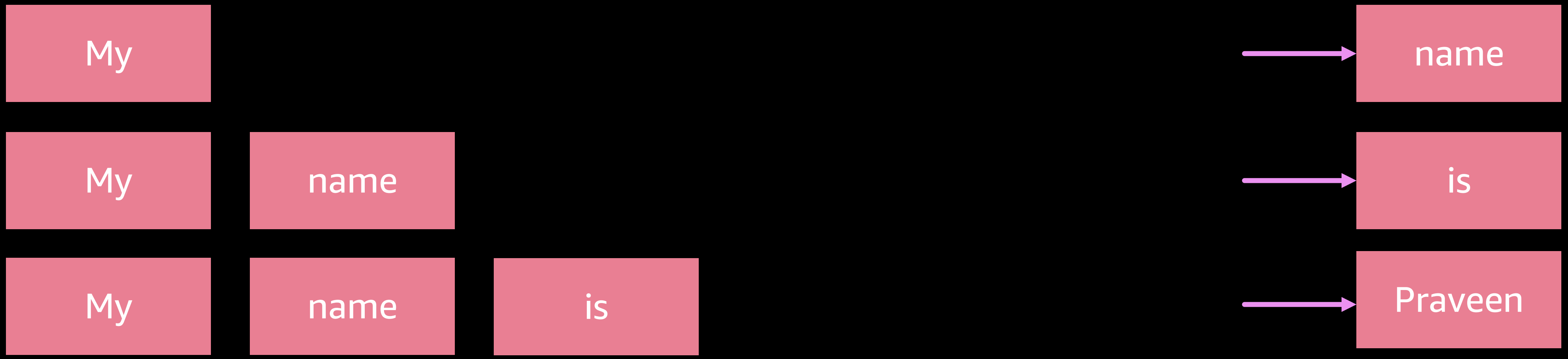
good

Transformer architecture - Decoder



Transformer architecture - Decoder

- Decoders can be used as stand-alone model
- Uni-directional
- Good at generating sequences
- Example of decoders: GPT-2, GPT Neo

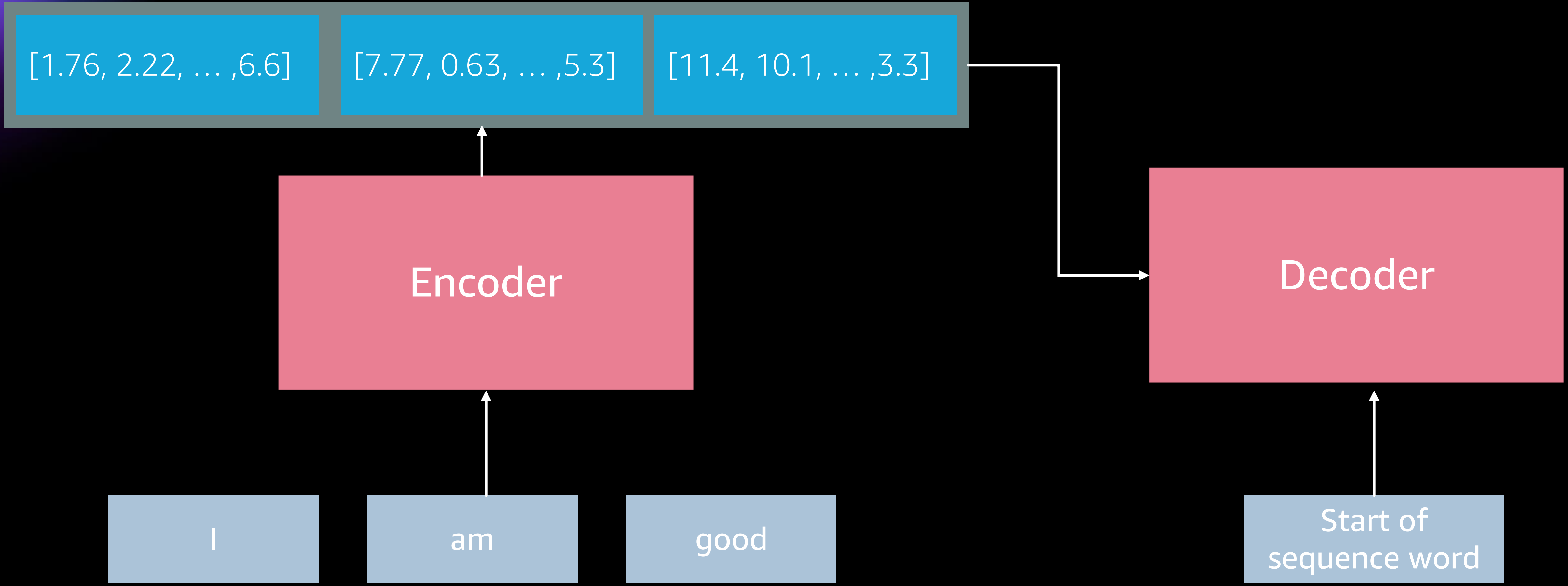


My → name

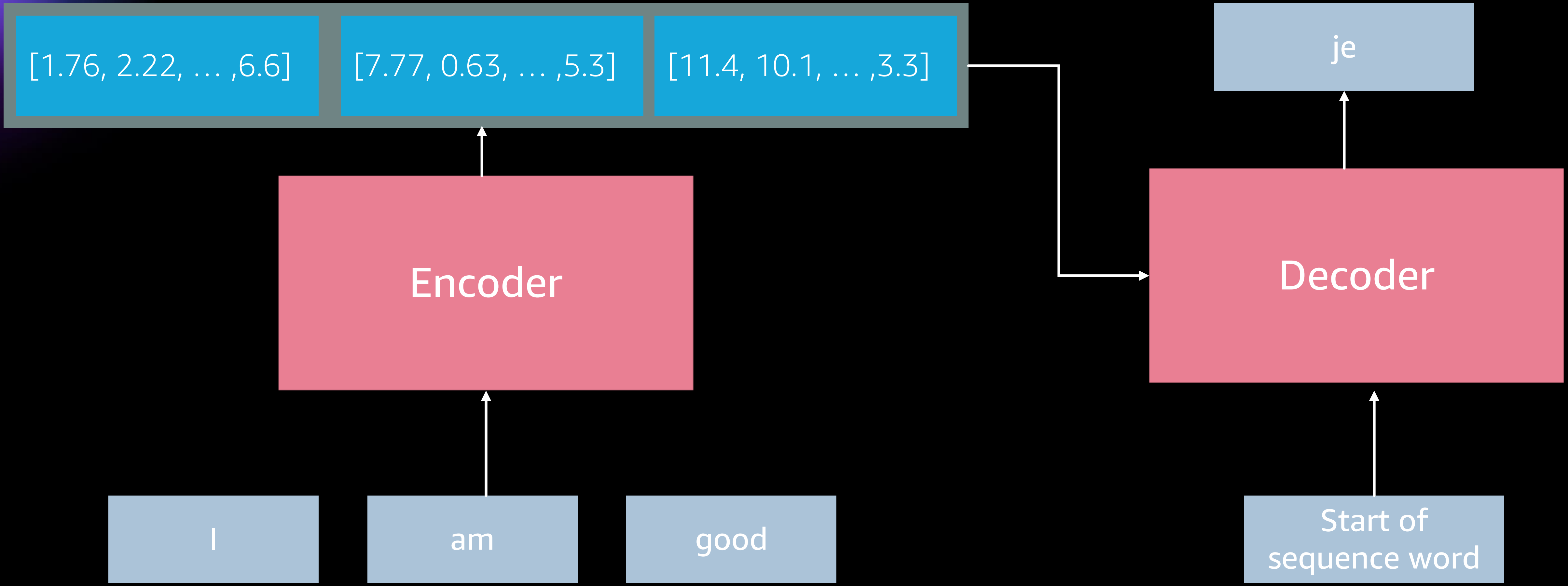
My name → is

My name is → Praveen

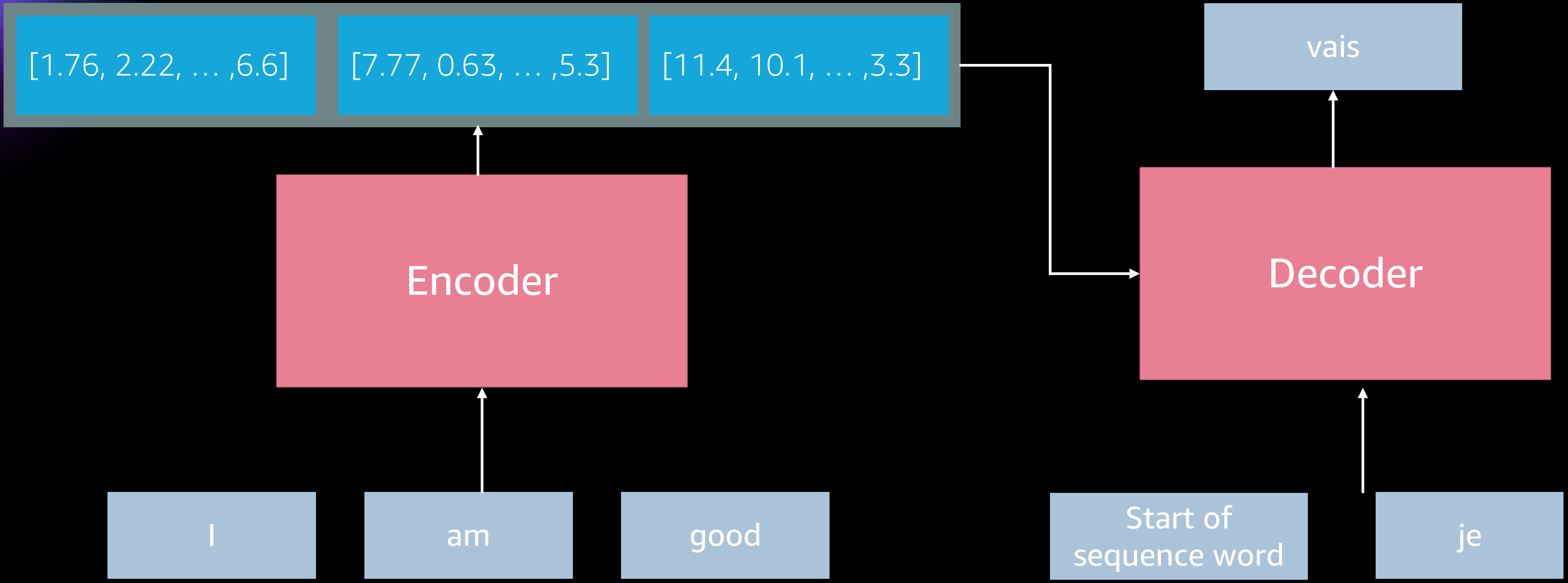
Transformer architecture – Encoder Decoder



Transformer architecture – Encoder Decoder



Transformer architecture – Encoder Decoder



Challenges in transformer models

- Transformers are big models
- Training this kind of model requires large amount of data
- It is costly in terms of time and compute resource
- This makes it impossible for a lot of organizations to train the model from scratch

Solution:

- Sharing the trained weights and building on top of already trained weights reduces the overall compute cost and carbon footprint of the community

What are Hugging Face Libraries

Open source
Datasets, tokenizers and transformers

Popular
57,000+ GitHub stars, 1,000,000+ downloads per month

Intuitive
NLP-specific Python frontends based on PyTorch or TensorFlow

State-of-the-art
Transformer-based models are state-of-the-art, enable transfer-learning and scale

Comprehensive
Model hub with 15,000+ model architectures, 240+ languages

A strong collaboration to make NLP easy and accessible for all

Hugging Face

Hugging Face is the most popular
open source company providing
state-of-the-art NLP technology

AWS

Amazon SageMaker offers
high performance resources
to train and use NLP models

Hugging Face experience in Amazon SageMaker

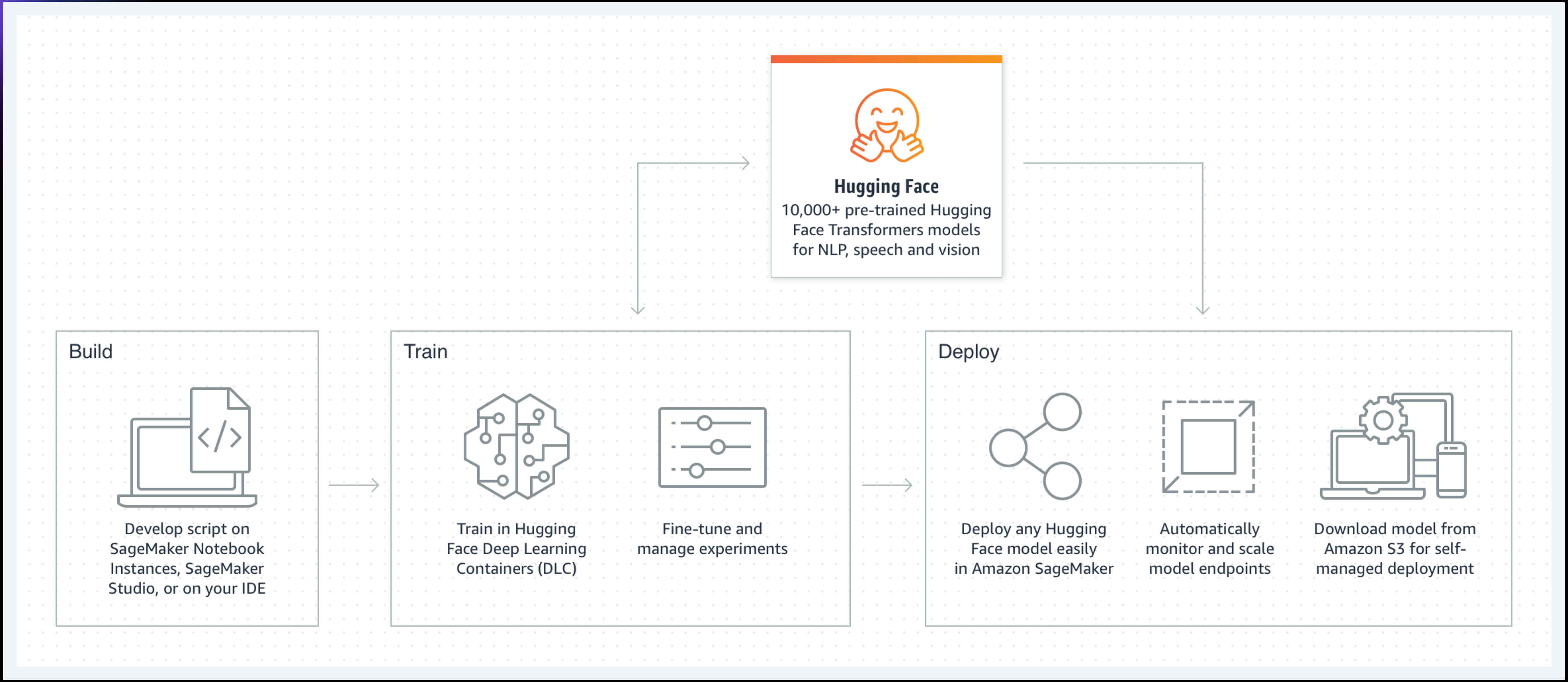
Deep learning containers (DLCs) developed with Hugging Face for both training and inference for the PyTorch and TensorFlow frameworks

A Hugging Face estimator in the Amazon SageMaker SDK to launch NLP scripts on scalable, cost-effective SageMaker training jobs without worrying about Docker

An example gallery to find readily usable high-quality samples of Hugging Face scripts on Amazon SageMaker

Maintained and supported by AWS

Integrated workflow with Amazon SageMaker



Hugging Face
10,000+ pre-trained Hugging Face Transformers models for NLP, speech and vision

Build

Develop script on SageMaker Notebook Instances, SageMaker Studio, or on your IDE

Train

Train in Hugging Face Deep Learning Containers (DLC)

Fine-tune and manage experiments

Deploy

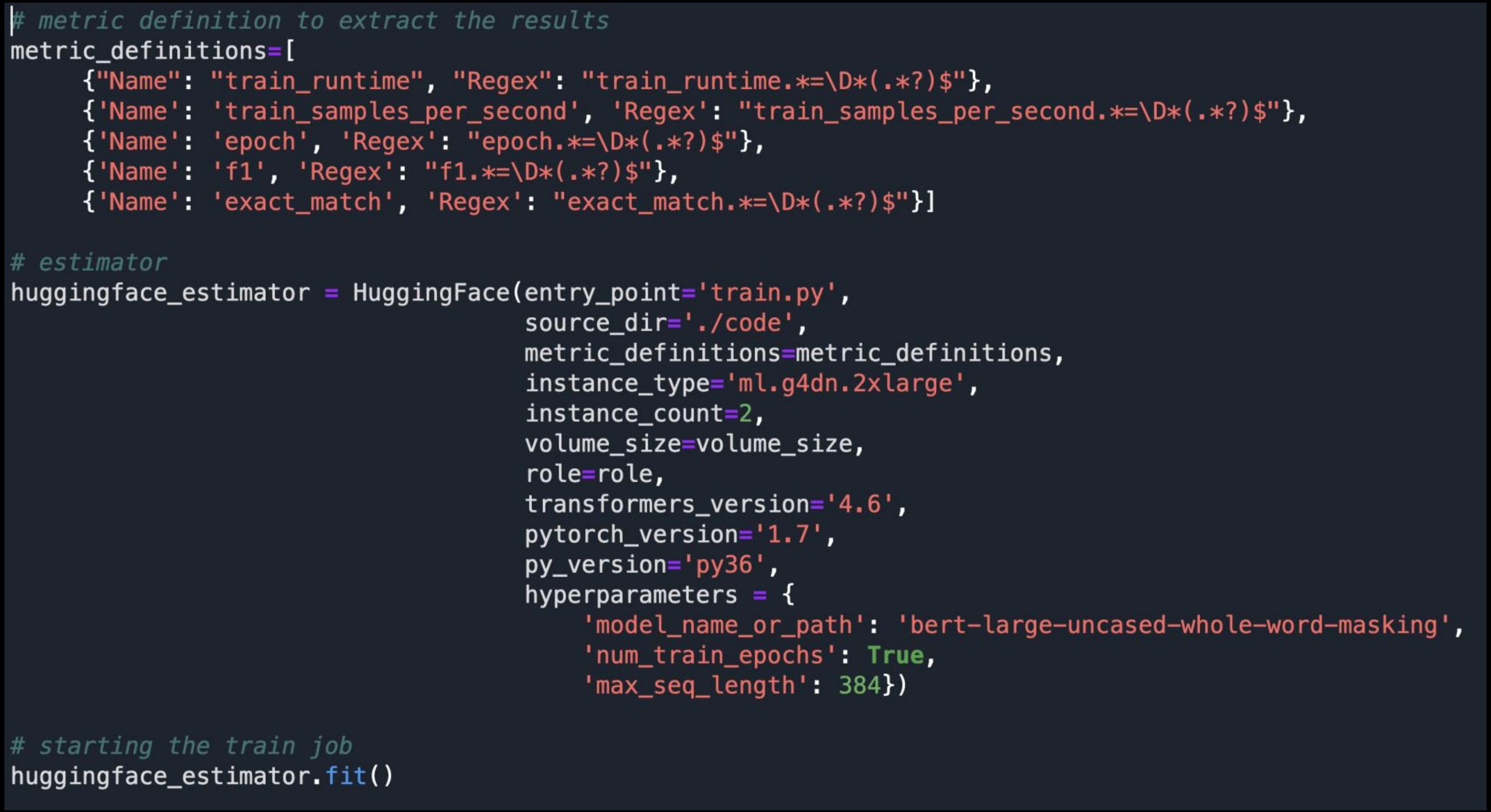
Deploy any Hugging Face model easily in Amazon SageMaker

Automatically monitor and scale model endpoints

Download model from Amazon S3 for self-managed deployment

Training on Amazon SageMaker

Hugging Face estimator

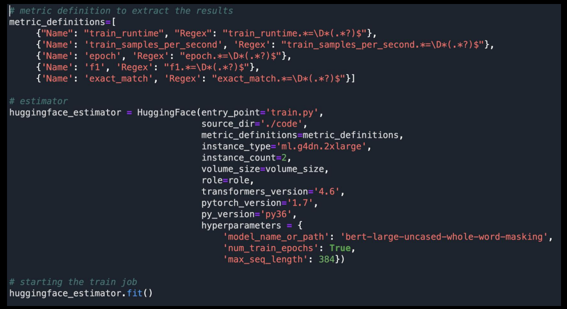


```
# metric definition to extract the results
metric_definitions=[
     {"Name": "train_runtime", "Regex": "train_runtime.*=\D*(.*?)$"},
     {'Name': 'train_samples_per_second', 'Regex': "train_samples_per_second.*=\D*(.*?)$"},
     {'Name': 'epoch', 'Regex': "epoch.*=\D*(.*?)$"},
     {'Name': 'f1', 'Regex': "f1.*=\D*(.*?)$"},
     {'Name': 'exact_match', 'Regex': "exact_match.*=\D*(.*?)$"}]

# estimator
huggingface_estimator = HuggingFace(entry_point='train.py',
                                    source_dir='./code',
                                    metric_definitions=metric_definitions,
                                    instance_type='ml.g4dn.2xlarge',
                                    instance_count=2,
                                    volume_size=volume_size,
                                    role=role,
                                    transformers_version='4.6',
                                    pytorch_version='1.7',
                                    py_version='py36',
                                    hyperparameters = {
                                        'model_name_or_path': 'bert-large-uncased-whole-word-masking',
                                        'num_train_epochs': True,
                                        'max_seq_length': 384})

# starting the train job
huggingface_estimator.fit()
```

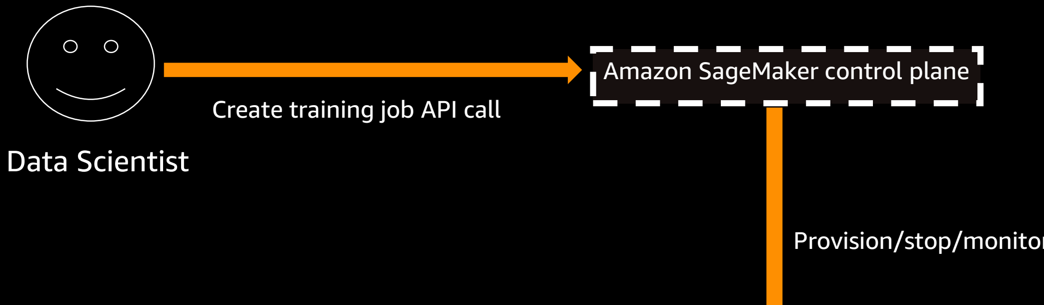
Training on Amazon SageMaker



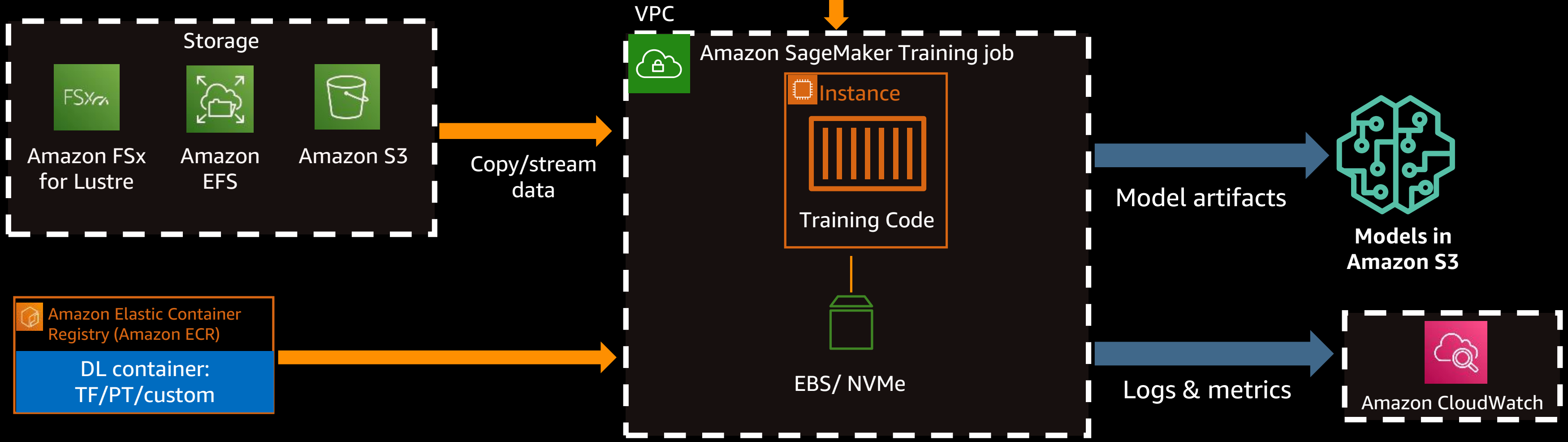
```python
# metric definition to extract the results
metric_definitions=[
     {"Name": "train_runtime", "Regex": "train_runtime.*=\D*(.*?)$"},
     {'Name': 'train_samples_per_second', 'Regex': "train_samples_per_second.*=\D*(.*?)$"},
     {'Name': 'epoch', 'Regex': "epoch.*=\D*(.*?)$"},
     {'Name': 'f1', 'Regex': "f1.*=\D*(.*?)$"},
     {'Name': 'exact_match', 'Regex': "exact_match.*=\D*(.*?)$"}]

# estimator
huggingface_estimator = HuggingFace(entry_point='train.py',
                                    source_dir='./code',
                                    metric_definitions=metric_definitions,
                                    instance_type='ml.g4dn.2xlarge',
                                    instance_count=2,
                                    volume_size=volume_size,
                                    role=role,
                                    transformers_version='4.6',
                                    pytorch_version='1.7',
                                    py_version='py36',
                                    hyperparameters = {
                                        'model_name_or_path': 'bert-large-uncased-whole-word-masking',
                                        'num_train_epochs': True,
                                        'max_seq_length': 384})

# starting the train job
huggingface_estimator.fit()
```



Data Scientist
Create training job API call
Amazon SageMaker control plane
Provision/stop/monitor instances

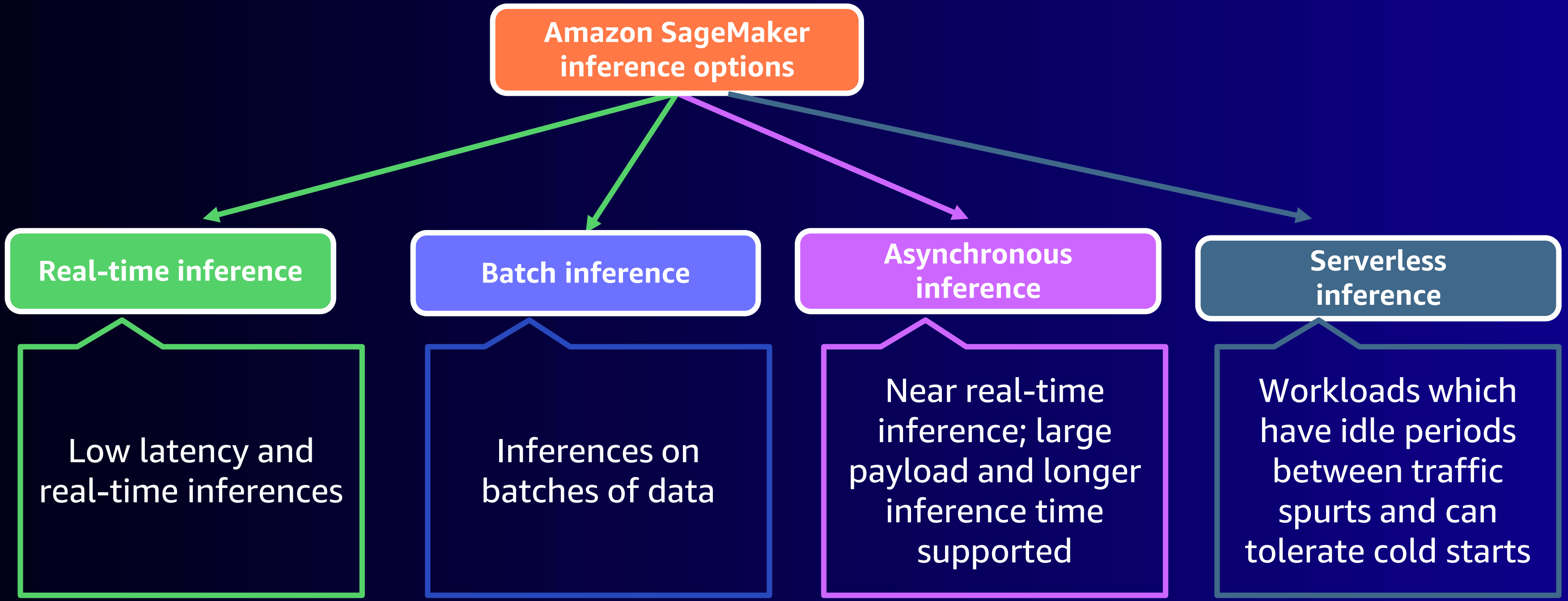


Demo – Fine-tuning Hugging Face Model with Amazon SageMaker

Scaling Hugging Face model training

- Data Parallelism
- Model Parallelism
- Amazon SageMaker Training Compiler

Amazon SageMaker Inference Options



Amazon SageMaker inference options

Real-time inference

Low latency and real-time inferences

Batch inference

Inferences on batches of data

Asynchronous inference

Near real-time inference; large payload and longer inference time supported

Serverless inference

Workloads which have idle periods between traffic spurts and can tolerate cold starts

Demo – Deploying a model as real time endpoint and batch inference

Recap

1. Understand the need for Transformer and different NLP task that can be solved
2. Learn how encoder, decoder and encoder-decoder transformer architecture work
3. Find out how Hugging Face model can be trained in Amazon SageMaker
4. Discover the different scaling mechanisms available in Amazon SageMaker to train this large models
5. Dive deep into the 4 different deployment options for Hugging Face models

Resources

1. Hugging Face documentation for Amazon SageMaker - https://huggingface.co/docs/sagemaker/main

2. Amazon SageMaker documentation for Hugging Face - https://docs.aws.amazon.com/sagemaker/latest/dg/hugging-face.html

3. Github samples - https://github.com/huggingface/notebooks/tree/master/sagemaker

4. Frequently asked questions - https://huggingface.co/docs/sagemaker/faq

Visit the AI & Machine Learning resource hub for more resources

Dive deeper into these resources, get inspired and learn how you can use AI and machine learning to accelerate your business outcomes.

- The machine learning journey e-book
- 7 leading machine learning use cases e-book
- A strategic playbook for data, analytics, and machine learning e-book
  Accelerate machine learning innovation with the right cloud services & infrastructure e-book
- Choosing the right compute infrastructure for machine learning e-book
- Improving service and reducing costs in contact centers e-book
- Why ML is essential in your fight against online fraud e-book
- … and more!

https://bit.ly/3mwi59V

Visit resource hub

AWS Machine Learning (ML) Training and Certification

AWS is how you build machine learning skills

Courses built on the curriculum leveraged by Amazon's own teams. Learn from the experts at AWS.

aws.training/machinelearning

Flexibility to learn your way

Learn online with on-demand digital courses or live with virtual instructor-led training, plus hands-on labs and opportunities for practical application.

explore.skillbuilder.aws/learn

Validate your expertise

Demonstrate expertise in building, training, tuning, and deploying machine learning models with an industry-recognized credential.

aws.amazon.com/certification

Thank you for attending AWS Innovate – AI/ML Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

aws-apj-marketing@amazon.com

twitter.com/AWSCloud

facebook.com/AmazonWebServices

youtube.com/user/AmazonWebServices

slideshare.net/AmazonWebServices

twitch.tv/aws

Thank you!

Praveen Jayakumar