Easy Model Deployer

None

None

None

Table of contents

1. Architecture

Deploy models to the cloud with EMD will use the following components in Amazon Web Services:

alt text

1. User/Client initiates model deployment task, triggering pipeline to start model building.
2. AWS CodeBuild constructs the large model using predefined configuration and publishes it to Amazon ECR.
3. AWS CloudFormation creates a model infrastructure stack based on user selection and deploys the model from ECR to AWS services (Amazon SageMaker, EC2, ECS).

2. Installation Guide

2.1 Prerequisites

- Python 3.9 or higher
- pip (Python package installer)

2.2 Setting up the Environment

1. Create a virtual environment:

```
python -m venv emd-env
```

2. Activate the virtual environment:

```
source emd-env/bin/activate
```

3. Install the required packages:

```
pip install https://github.com/aws-samples/easy-model-deployer/releases/download/main/emd-0.6.0-py3-none-any.whl
```

3. Invocation

4. Supported Model

ModelId	ModelSeries	ModelType	Supported Engines	Supported Instances
glm-4-9b-chat	glm4	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
internlm2_5-20b-chat-4bit-awq	internlm2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
internlm2_5-20b-chat	internlm2.5	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
internlm2_5-7b-chat	internlm2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
internlm2_5-7b-chat-4bit	internlm2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
internlm2_5-1_8b-chat	internlm2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
Qwen2.5-7B-Instruct	qwen2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
Qwen2.5-72B-Instruct-AWQ	qwen2.5	llm	vllm,tgi	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2.5-72B-Instruct	qwen2.5	llm	vllm	g5.48xlarge
Qwen2.5-72B-Instruct-AWQ-128k	qwen2.5	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2.5-32B-Instruct	qwen2.5	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2.5-0.5B-Instruct	qwen2.5	llm	vllm,tgi	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge,i
Qwen2.5-1.5B-Instruct	qwen2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
Qwen2.5-3B-Instruct	qwen2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
Qwen2.5-14B-Instruct-AWQ	qwen2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
Qwen2.5-14B-Instruct	qwen2.5	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
QwQ-32B-Preview	qwen reasoning model	llm	huggingface,vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
llama-3.3-70b-instruct-awq	llama	llm	tgi	g5.12xlarge,g5.24xlarge,g5.48xlarge
DeepSeek-R1-Distill-Qwen-32B	deepseek reasoning model	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
DeepSeek-R1-Distill-Qwen-14B	deepseek reasoning model	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge

ModelId	ModelSeries	ModelType	Supported Engines	Supported Instances
DeepSeek-R1-Distill-Qwen-7B	deepseek reasoning model	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
DeepSeek-R1-Distill-Qwen-1.5B	deepseek reasoning model	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
DeepSeek-R1-Distill-Llama-8B	deepseek reasoning model	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
deepseek-r1-distill-llama-70b-awq	deepseek reasoning model	llm	tgi,vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
Baichuan-M1-14B-Instruct	baichuan	llm	huggingface	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2-VL-72B-Instruct-AWQ	qwen2vl	vlm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
QVQ-72B-Preview-AWQ	qwen reasoning model	vlm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2-VL-7B-Instruct	qwen2vl	vlm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
InternVL2_5-78B-AWQ	internvl2.5	vlm	lmdeploy	g5.12xlarge,g5.24xlarge,g5.48xlarge
txt2video-LTX	comfyui	video	comfyui	g5.4xlarge,g5.8xlarge,g6e.2xlarge
whisper	whisper	whisper	huggingface	g5.xlarge,g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.1
bge-base-en-v1.5	bge	embedding	vllm	g5.xlarge,g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.1
bge-m3	bge	embedding	vllm	g5.xlarge,g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.1
bge-reranker-v2-m3	bge	rerank	vllm	g5.xlarge,g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.1