Adlik Antelope Release (V0.1.0)

Introduction of Adlik release 0.1.0: Adlik-AI.pptx

Below are the key features delivered in Adlik Release 0.1.0, see chinese version here Feature List.

Model Compiler

1. A new framework which is easy to expand and maintain.
2. Compilation of models trained from Keras, Tensorflow and Pytorch for better execution on CPU/GPU.

Training framework	Model format	Target runtime	compiled format
Keras	h5	Tf Serving	SavedModel
		OpenVINO	IR
		TensorRT	Plan
		TF-Lite	tflite
TensorFlow	Ckpt/pb	Tf Serving	SavedModel
		OpenVINO	IR
		TensorRT	Plan
		TF-Lite	tflite
PyTorch	pth	OpenVINO	IR
		TensorRT	Plan

Training framework	Inference engine	hardware environment
	TensorFlow Serving-1.14	CPU/GPU
	TensorFlow Serving-2.2	CPU/GPU
	OpenVINO-2019	CPU
	TensorRT-6	GPU
	TensorRT-7	GPU
	TF Lite-2.1	CPU(X86/ARM)
TensorFlow	TensorFlow Serving-1.14	CPU/GPU
	TensorFlow Serving-2.2	CPU/GPU
	OpenVINO-2019	CPU
	TensorRT-6	GPU
	TensorRT-7	GPU
	TF Lite-2.1	CPU(X86/ARM)
PyTorch	OpenVINO-2019	CPU
	TensorRT-6	GPU

Model Optimizer

1. Multi nodes multi GPUs training and pruning.
2. Configurable implementation of filter pruning to achieve smaller size of inference models.
3. Small batch dataset quantization for TF-Lite and TF-TRT.

Inference Engine

1. Management of multi models and multi versions.
2. HTTP/GRPC interfaces for inference service.
3. Runtime scheduler that supports scheduling of multi model instances.
4. Integration of multiple DL inference runtime, including TensorFlow Serving, OpenVINO, TensorRT and TF Lite.

Integrated Inference engine	Hardware environment
TensorFlow Serving-1.14	CPU/GPU
TensorFlow Serving-2.2	CPU/GPU
OpenVINO-2019	CPU
TensorRT-6	GPU
TensorRT-7	GPU
TF Lite-2.1	CPU(X86/ARM)

5. Integration of dlib to support ML runtime.

Benchmark Test Framework for Deep Learning Model

1. A containalized solution which auto executes compiling, packaging of models, loading of runtime and models, startup of inference service and client, and generation of testing results.
2. Supports all the compilers and runtime that can be integrated into Adlik.
3. Supported output: inference result, inference speed, delay of inference execution.