Release Deer (V0.4.0)

https://github.com/Adlik/Adlik/releases/tag/v0.4.0

Adlik Deer

Feature List

Compiler

1. Adlik compiler supports OpenVINO INT8 quantization.
2. Adlik compiler supports TensorRT INT8 quantization. Supports extended quantization calibrator for TensorRT for reducing the accuracy drop caused by quantization.

Optimizer

1. Support multi-teacher distillation method, which uses multi-teacher networks for distillation optimization.
2. Support ZEN-NAS search enhancement features, including parallel training, optimization for search acceleration, fix the bugs of original implementation etc. The consumed search time is reduced by about 15%, when the search score is slightly improved, which results in increas of the training accuracy by 0.2% ~1%.

Inference Engine

1. Support Paddle Inference Runtime. When using Paddle-format model, converting model format through Onnx components is not needed, and users can directly perform model inference in the Adlik environment.
2. Support Intel TGL-U i5 device inference, and complete benchmark tests with several models.
3. Docker images for cloud native environments support newest version of inference components including:
   (1) OpenVINO: version 2021.4.582
   (2)TensorFlow: 2.6.2
   (3)TensorRT: 7.2.1.6
   (4) Tf-lite: 2.4.0
   (5) TVM: 0.7
   (6) Paddle Inference: 2.1.2
4. Introduce C++ version of Client API, which supports cmake and bazel compilation, and is convenient for users to deploy in C/C++ scenarios.

Benchmark Test

1. Complete Benchmark tests of Resnet-50, Yolo v3/v4, FastRCNN, MaskRCNN and other models on Intel TGL-U i5 equipment, including latency, throughput, and various performance indicators under GPU video decoding.
2. MLPerf result on Bert model with Adlik-optimized.

Fixed issues

- Fix broken link in Readme.
- The model conversion from caffe2 to tensorrt does not recognize the MAX_BATCH_SIZE parameter.
- The compiler image deployed using kubernetes does not support the compilation of checkpoint to TensorRT.
- Add C++ interface and example for calling Adlik.
- Add CI For Compiling TVM Runtime.
- Adlik serving can not run faster_rcnn model when batch size > 1 with openvino runtime.