

Adlik Bear Release (V0.2.0)

Release date

November 20th, 2020

Major features and improvements

New Compiler

- Support DAG generation for end-to-end compilation of models with different representation
 - Source representation: H5, Ckpt, Pb, Pth, Onnx and SavedModel
 - Target representation: SavedModel, OpenVINO IR, TensorRT Plan and Tflite
- Support model quantization for Tflite and TensorRT
 - Int8 quantization for Tflite
 - Int8 and fp16 quantization for TensorRT

Inference Engine

- Support hybrid scheduling of ML and DL inference jobs
- Support image based deployment of Adlik compiler and inference engine in cloud native environment, deployment and function test has been done in:
 - Cloud environment based on docker (V19.03.12)
 - Cloud environment based on Kubernetes (V1.13)
- Support the newest version of OpenVINO (2021.1.110) and TensorFlow (2.3.1)

Benchmark Test

- Support the following models:

	ResNet-50	Inception V3	Yolo V3	Bert
Tf GPU				
Tf CPU				
TensorRT				
OpenVINO				
TFLite				

BugFixes

The following bugs are fixed:

- 1) Can Not Convert Yolo.h5 To Openvino Runtime.
- 2) gRPC:Received message larger than max.
- 3) Return Message Is Wrong When cudaMalloc() Failed In initializeOutputBindings() Method.
- 4) Can Not Do Predict With Following Transferred YoloV3 Model.
- 5) adlik_serving --help should exit successfully.
- 6) benchmark cant auto infer by tensorflow gpu image.
- 7) Prediction will fail if information in model.pbtxt and model representation not consistent in tensorflowLite runtime.

Please see <https://github.com/Adlik/Adlik/issues?q=is%3Aissue+is%3Aclosed> for more information.