

ONNX Runtime Cheatsheet

Installation

pip install onnxruntime

Install ONNX Runtime CPU version

pip install onnxruntime-gpu

Install ONNX Runtime GPU version

pip install onnx

Install ONNX (Open Neural Network Exchange)

Model Loading

import onnxruntime as ort

Import ONNX Runtime

session = ort.InferenceSession('model.onnx')

Load an ONNX model

**session = ort.InferenceSession('model.onnx',
providers=['CUDAExecutionProvider'])**

Load model with GPU acceleration

Inference

input_name = session.get_inputs()[0].name

Get input name

output_name = session.get_outputs()[0].name

Get output name

**results = session.run([output_name], {input_name:
input_data})**

Run inference

predictions = results[0]

Access prediction results

Model Inspection

session.get_inputs()

Get model input details

session.get_outputs()

Get model output details

session.get_providers()

List available execution providers

ort.get_device()

Get the device being used

Performance Tuning

options = ort.SessionOptions()

Create session options object

**options.graph_optimization_level =
ort.GraphOptimizationLevel.ORT_ENABLE_ALL**

Enable all optimizations

options.intra_op_num_threads = 4

Set number of threads

**session = ort.InferenceSession('model.onnx',
sess_options=options)**

Create session with custom options

ONNX Model Conversion

import onnx

Import ONNX

onnx_model = onnx.load('model.onnx')

Load ONNX model

onnx.checker.check_model(onnx_model)

Validate ONNX model

onnx.save(onnx_model, 'new_model.onnx')

Save ONNX model