Class OnnxInfer

Table of contents

- Defined in File core.hpp

Inheritance Relationships

Base Type

- `public holoscan::inference::InferBase` (Class InferBase)

Class Documentation

class OnnxInfer : public holoscan::inference::InferBase

Onnxruntime based inference class

Public Functions

OnnxInfer(const std::string &model_file_path, bool cuda_flag)

Constructor.

Parameters

- **model_file_path** – Path to onnx model file
- **cuda_flag** – Flag to show if inference will happen using CUDA

~OnnxInfer()

Destructor.

virtual InferStatus do_inference(const std::vector<std::shared_ptr<DataBuffer>> &input_data, std::vector<std::shared_ptr<DataBuffer>> &output_buffer)

Does the Core inference using Onnxruntime. Input and output buffer are supported on Host. Inference is supported on host and device.

Parameters

- **input_data** – Input DataBuffer

- **output_buffer** – Output DataBuffer, is populated with inferred results

Returns

InferStatus

void populate_model_details()

Populate class parameters with model details and values.

void print_model_details()

Print model details.

int set_holoscan_inf_onnx_session_options()

Create session options for inference.

virtual std::vector<std::vector<int64_t>> get_input_dims() const

Get input data dimensions to the model.

Returns

Vector of input dimensions. Each dimension is a vector of int64_t corresponding to the shape of the input tensor.

virtual std::vector<std::vector<int64_t>> get_output_dims() const

Get output data dimensions from the model.

Returns

Vector of input dimensions. Each dimension is a vector of int64_t corresponding to the shape of the input tensor.

virtual std::vector<holoinfer_datatype> get_input_datatype() const

Get input data types from the model.

Returns

Vector of values as datatype per input tensor

virtual std::vector<holoinfer_datatype> get_output_datatype() const

Get output data types from the model.

Returns

Vector of values as datatype per output tensor

virtual void cleanup()