Class TorchInfer

Table of contents

- Defined in File core.hpp

Inheritance Relationships

Base Type

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

Class Documentation

class TorchInfer : public holoscan::inference::InferBase

Libtorch based inference class

Public Functions

TorchInfer(const std::string &model_file_path, bool cuda_flag, bool cuda_buf_in, bool cuda_buf_out)

Constructor.

Parameters

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

~TorchInfer()

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.

Parameters

- **input_data** – Vector of Input DataBuffer

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

Returns

InferStatus

InferStatus populate_model_details()

Populate class parameters with model details and values.

void print_model_details()

Print model details.

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 output dimensions. Each dimension is a vector of int64_t corresponding to the shape of the output 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