Class TrtInfer

Table of contents

- Defined in File core.hpp

Inheritance Relationships

Base Type

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

Class Documentation

class TrtInfer : public holoscan::inference::InferBase

Class to execute TensorRT based inference

Public Functions

TrtInfer(const std::string &model_path, const std::string &model_name, int device_id, bool enable_fp16, bool is_engine_path, bool cuda_buf_in, bool cuda_buf_out)

Constructor.

~TrtInfer()

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 with TRT backend.

Parameters

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

Returns

InferStatus

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

Get input data dimensions to the model.

Returns

Vector of values as dimension

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

inline virtual void cleanup()