Struct InferenceSpecs

Table of contents

- Defined in File holoinfer_buffer.hpp

Struct Documentation

struct InferenceSpecs

Struct that holds specifications related to inference, along with input and output data buffer.

Public Functions

InferenceSpecs() = default

inline InferenceSpecs(const std::string &backend, const Mappings &backend_map, const Mappings &model_path_map, const MultiMappings &pre_processor_map, const MultiMappings &inference_map, const Mappings &device_map, const Mappings &temporal_map, bool is_engine_path, bool oncpu, bool parallel_proc, bool use_fp16, bool cuda_buffer_in, bool cuda_buffer_out)

Constructor.

Parameters

- **backend** – Backend inference (trt or onnxrt)
- **backend_map** – Backend inference map with model name as key, and backend as value
- **model_path_map** – Map with model name as key, path to model as value
- **pre_processor_map** – Map with model name as key, input tensor names in vector form as value
- **inference_map** – Map with model name as key, output tensor names in vector form as value
- **device_map** – Map with model name as key, GPU ID for inference as value

- **temporal_map** – Map with model name as key, frame number to skip for inference as value
- **is_engine_path** – Input path to model is trt engine
- **oncpu** – Perform inference on CPU
- **parallel_proc** – Perform parallel inference of multiple models
- **use_fp16** – Use FP16 conversion, only supported for trt
- **cuda_buffer_in** – Input buffers on CUDA
- **cuda_buffer_out** – Output buffers on CUDA

inline Mappings get_path_map() const

Get the model data path map.

Returns

Mappings data

inline Mappings get_backend_map() const

Get the model backend map.

Returns

Mappings data

inline Mappings get_device_map() const

Get the device map.

Returns

Mappings data

inline Mappings get_temporal_map() const

Get the Temporal map.

Returns

Mappings data

Public Members

std::string backend_type_ = {""}

Backend type (for all models)

Mappings backend_map_

Backend map.

Mappings model_path_map_

Map with key as model name and value as model file path.

MultiMappings pre_processor_map_

Map with key as model name and value as vector of input tensor names.

MultiMappings inference_map_

Map with key as model name and value as inferred tensor name.

Mappings device_map_

Map with key as model name and value as GPU ID for inference.

Mappings temporal_map_

Map with key as model name and frame number to skip for inference as value.

bool is_engine_path_ = false

Flag showing if input model path is path to engine files.

bool oncuda_ = true

Flag showing if inference on CUDA. Default is True.

bool parallel_processing_ = false

Flag to enable parallel inference. Default is True.

bool use_fp16_ = false

Flag showing if trt engine file conversion will use FP16. Default is False.

bool cuda_buffer_in_ = true

Flag showing if input buffers are on CUDA. Default is True.

bool cuda_buffer_out_ = true

Flag showing if output buffers are on CUDA. Default is True.

DataMap data_per_tensor_

Input Data Map with key as tensor name and value as DataBuffer.

DataMap output_per_model_

Output Data Map with key as tensor name and value as DataBuffer.