DPU runtime library

dpuOpen() - Open & initialize the usage of DPU device  
dpuClose() - Close & finalize the usage of DPU device  
dpuLoadKernel() - Load a DPU Kernel and allocate DPU memory space for itsCode/Weight/Bias

segments  
dpuDestroyKernel() - Destroy a DPU Kernel and release its associated resources  
dpuCreateTask() - Instantiate a DPU Task from one DPU Kernel, allocate its private working

memory buffer and prepare for its execution context  
dpuRunTask() - Launch the running of DPU Task

dpuDestroyTask() - Remove a DPU Task, release its working memory buffer and destroy

associated execution context

dpuEnableTaskDump() - Enable dump facility of DPU Task while running for debugging purpose

dpuEnableTaskProfile() - Enable profiling facility of DPU Task while running to get its

performance metrics

dpuGetTaskProfile() - Get the execution time of DPU Task  
dpuGetNodeProfile() - Get the execution time of DPU Node  
dpuGetInputTensorCnt() - Get total number of input Tensor of one DPU Task  
dpuGetInputTensor() - Get input Tensor of one DPU Task  
dpuGetInputTensorAddress() - Get the start address of one DPU Task’s input Tensor

dpuGetInputTensorSize() - Get the size (in byte) of one DPU Task’s input Tensor

dpuGetInputTensorScale() - Get the scale value of one DPU Task’s input Tensor

dpuGetInputTensorHeight() - Get the height dimension of one DPU Task’s input Tensor

dpuGetInputTensorWidth() - Get the width dimension of one DPU Task’s input Tensor

dpuGetInputTensorChannel() - Get the channel dimension of one DPU Task’s input Tensor

dpuGetOutputTensorCnt() - Get total number of output Tensor of one DPU Task

dpuGetOutputTensor() - Get output Tensor of one DPU Task

dpuGetOutputTensorAddress() - Get the start address of one DPU Task’s output Tensor

dpuGetOutputTensorSize() - Get the size in byte of one DPU Task’s output Tensor

dpuGetOutputTensorScale() - Get the scale value of one DPU Task’s output Tensor

dpuGetOutputTensorHeight() - Get the height dimension of one DPU Task’s output Tensor

dpuGetOutputTensorWidth() - Get the width dimension of one DPU Task’s output Tensor

dpuGetOutputTensorChannel() - Get the channel dimension of one DPU Task’s output Tensor

dpuGetTensorSize() - Get the size of one DPU Tensor

dpuGetTensorAddress() - Get the start address of one DPU Tensor

dpuGetTensorScale() - Get the scale value of one DPU Tensor

dpuGetTensorHeight() - Get the height dimension of one DPU Tensor

dpuGetTensorWidth() - Get the width dimension of one DPU Tensor

dpuGetTensorChannel() - Get the channel dimension of one DPU Tensor

dpuSetIntputTensorInCHWInt8() - Set DPU Task’s input Tensor with data stored under Caffe

order (channel/height/width) in INT8 format

dpuSetIntputTensorInCHWFP32() - Set DPU Task’s input Tensor with data stored under Caffe

order (channel/height/width) in FP32 format

dpuSetIntputTensorInHWCInt8() - Set DPU Task’s input Tensor with data stored under DPU

order (height/width/channel) in INT8 format

dpuSetIntputTensorInHWCFP32() - Set DPU Task’s input Tensor with data stored under DPU

order (channel/height/width) in FP32 format

dpuGetOutputTensorInCHWInt8() - Get DPU Task’s output Tensor and store them under

Caffe order (channel/height/width) in INT8 format

dpuGetOutputTensorInCHWFP32() - Get DPU Task’s output Tensor and store them under Caffe

order (channel/height/width) in FP32 format

dpuGetOutputTensorInHWCInt8() - Get DPU Task’s output Tensor and store them under DPU

order (channel/height/width) in INT8 format

dpuGetOutputTensorInHWCFP32() - Get DPU Task’s output Tensor and store them under DPU

order (channel/height/width) in FP32 format