**Optimizing Inference Applications**

**Lab4-1: Pick the right model based on application and hardware**

1. **Set environmental variables**

$ source /opt/intel/computer\_vision\_sdk/bin/setupvars.sh

$ export SV=/opt/intel/workshop/smart-video-workshop

$ cd $SV/object-detection

1. **Run 3 inference models with CPU**

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd

inputDims=300 300 3 1

outputDims=1 1 100 7

SSD Mode

Preprocess: 6.64406 ms/frame

Inference: 20.4566 ms/frame

Postprocess:0.0297406 ms/frame

$ ./lab.py tutorial1 -i Cars1.mp4 -m ssd300

inputDims=300 300 3 1

outputDims=1 1 200 7

SSD Mode

Preprocess: 6.83784 ms/frame

Inference: 401.961 ms/frame

Postprocess:0.0370388 ms/frame

$ ./lab.py tutorial1 -i Cars1.mp4 -m ssd512

inputDims=512 512 3 1

outputDims=1 1 200 7

SSD Mode

Preprocess: 9.77347 ms/frame

Inference: 1224.77 ms/frame

Postprocess:0.0361089 ms/frame

1. **Run 3 inference models with GPU**

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd -d GPU

inputDims=300 300 3 1

outputDims=1 1 100 7

SSD Mode

Preprocess: 6.2982 ms/frame

Inference: 25.4533 ms/frame

Postprocess:0.0342089 ms/frame

$ ./lab.py tutorial1 -i Cars1.mp4 -m ssd300 -d GPU

inputDims=300 300 3 1

outputDims=1 1 200 7

SSD Mode

Preprocess: 8.23353 ms/frame

Inference: 234.98 ms/frame

Postprocess:0.0366836 ms/frame

$ ./lab.py tutorial1 -i Cars1.mp4 -m ssd512 -d GPU

inputDims=512 512 3 1

outputDims=1 1 200 7

SSD Mode

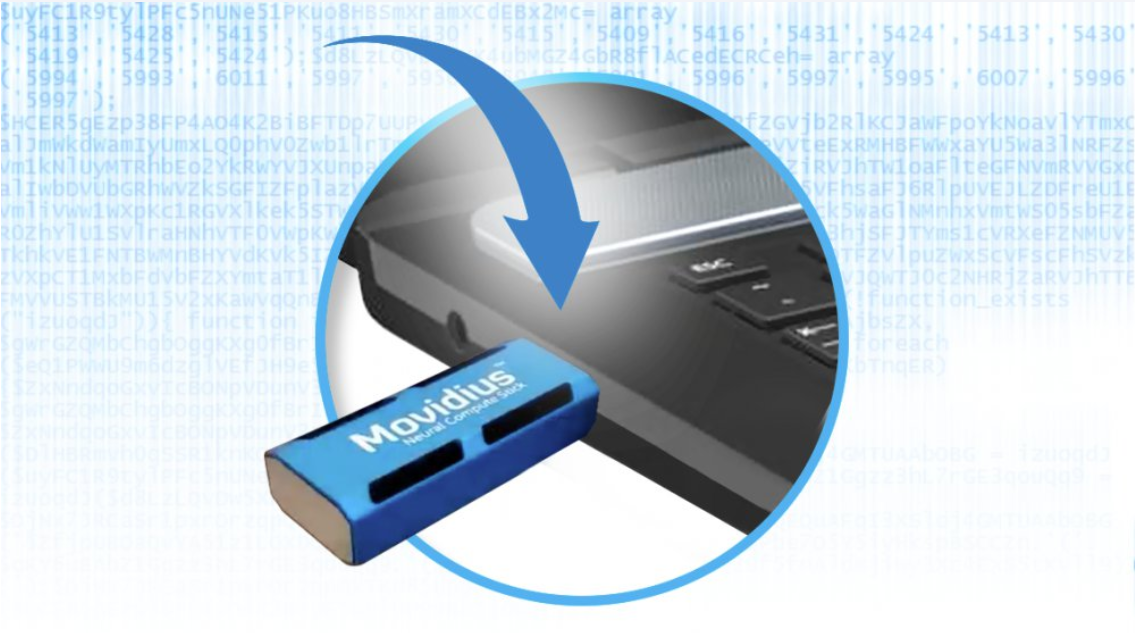
Preprocess: 12.0591 ms/frame

Inference: 637.36 ms/frame

Postprocess:0.0380333 ms/frame

1. **Run 3 inference models with Movidius NCS**

**Connect Movidius™ NCS to your development laptop**

****

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd -d MYRIAD -p FP16

inputDims=300 300 3 1

outputDims=1 1 100 7

SSD Mode

Preprocess: 29.333 ms/frame

Inference: 94.4297 ms/frame

Postprocess:0.0952729 ms/frame

$ ./lab.py tutorial1 -i Cars1.mp4 -m ssd300 -d MYRIAD -p FP16

inputDims=300 300 3 1

outputDims=1 1 200 7

SSD Mode

Preprocess: 12.4779 ms/frame

Inference: 1438.87 ms/frame

Postprocess:0.116529 ms/frame

$ ./lab.py tutorial1 -i Cars1.mp4 -m ssd512 -d MYRIAD -p FP16

inputDims=512 512 3 1

outputDims=1 1 200 7

SSD Mode

Preprocess: 15.0778 ms/frame

Inference: 4255.19 ms/frame

Postprocess:0.121199 ms/frame

1. **Run 3 inference models with “object\_detection\_demo\_ssd\_async” app to see FPS differences**

$ ./lab.py object\_detection\_demo\_ssd\_async -i Cars1.mp4 -m mobilenet-ssd/ssd300/ssd512 -d CPU/GPU/MYRIAD -p FP32/FP16



**Lab4-2: Run with multiple batch sizes**

1. **Set environmental variables**

$ export SV=/opt/intel/workshop/smart-video-workshop/

$ source /opt/intel/computer\_vision\_sdk/bin/setupvars.sh

$ cd $SV/object-detection

1. **Run with batch size 1**

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd

inputDims=300 300 3 1

outputDims=1 1 100 7

SSD Mode

Preprocess: 6.7047 ms/frame

Inference: 20.6598 ms/frame

Postprocess:0.0289015 ms/frame

1. **Run with batch size 2**

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd –b 2

inputDims=300 300 3 2

outputDims=2 1 100 7

SSD Mode

Preprocess: 6.9952 ms/frame

Inference: 19.0606 ms/frame

Postprocess:0.0184195 ms/frame

1. **Run with batch size 8**

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd –b 8

inputDims=300 300 3 8

outputDims=8 1 100 7

SSD Mode

Preprocess: 7.39572 ms/frame

Inference: 17.7449 ms/frame

Postprocess:0.011576 ms/frame

1. **Run with batch size 16**

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd –b 16

inputDims=300 300 3 16

outputDims=16 1 100 7

SSD Mode

Preprocess: 7.33112 ms/frame

Inference: 17.3909 ms/frame

Postprocess:0.0107071 ms/frame

**Lab4-3: Use the right data type for your target hardware and accuracy needs**

1. **Run model on GPU with FP32**

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd –d GPU

inputDims=300 300 3 1

outputDims=1 1 100 7

SSD Mode

Preprocess: 6.21707 ms/frame

Inference: 25.2745 ms/frame

Postprocess:0.0335757 ms/frame

1. **Run model GPU with FP16**

$ ./lab.py tutorial1 -i Cars1.mp4 -m mobilenet-ssd –d GPU –p FP16

inputDims=300 300 3 1

outputDims=1 1 100 7

SSD Mode

Preprocess: 6.41011 ms/frame

Inference: 19.0356 ms/frame

Postprocess:0.0326532 ms/frame

**Use Lab4-4: Async**

\*the Video used in this lab comes from: https://www.videvo.net

1. **Run model in CPU with FP32**

$ ./lab.py object\_detection\_demo\_ssd\_async -i lovely\_girl.mp4 -m ssd300

InferenceEngine:

API version ............ 1.1

Build .................. 12419

[ INFO ] Parsing input parameters

[ INFO ] Reading input

[ INFO ] Loading plugin

API version ............ 1.1

Build .................. lnx\_20180510

Description ....... MKLDNNPlugin

[ INFO ] Loading network files

[ INFO ] Batch size is forced to 1.

[ INFO ] Checking that the inputs are as the sample expects

[ INFO ] Checking that the outputs are as the sample expects

[ INFO ] Loading model to the plugin

[ INFO ] Start inference



1. **Press tab to switch to sync mode.**



**Appendix 1： Run Model Optimizer on the models to get IR files used in this lab**

$ export SV=/opt/intel/workshop/smart-video-workshop/

$ source /opt/intel/computer\_vision\_sdk/bin/setupvars.sh

$ cd $SV/object-detection

$ mkdir -p SSD512/{FP16,FP32}

$ mkdir -p SSD300/{FP16,FP32}

$ mkdir -p mobilenet-ssd/{FP16,FP32}

$ cd /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_optimizer

$ python3 mo\_caffe.py --input\_model /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/ssd/512/caffe/ssd512.caffemodel -o $SV/object-detection/SSD512/FP32

Model Optimizer arguments

Batch: 1

Precision of IR: FP32

Enable fusing: True

Enable gfusing: True

Names of input layers: inherited from the model

Path to the Input Model: /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/ssd/512/caffe/ssd512.caffemodel

Input shapes: inherited from the model

Log level: ERROR

Mean values: ()

IR output name: inherited from the model

Names of output layers: inherited from the model

Path for generated IR: /opt/intel/workshop/smart-video-workshop//object-detection/SSD512/FP32

Reverse input channels: False

Scale factor: None

Scale values: ()

Version: 0.3.61.37271eb9

Input proto file: deduced from the input model

Path to CustomLayersMapping.xml: extensions/front/caffe/CustomLayersMapping.xml

Path to a mean file:

Offsets for a mean file: None

[]

[ SUCCESS ] Generated IR model.

[ SUCCESS ] XML file: /opt/intel/workshop/smart-video-workshop//object-detection/SSD512/FP32/ssd512.xml

[ SUCCESS ] BIN file: /opt/intel/workshop/smart-video-workshop//object-detection/SSD512/FP32/ssd512.bin

$ python3 mo\_caffe.py --input\_model /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/ssd/512/caffe/ssd512.caffemodel -o $SV/object-detection/SSD512/FP16 --data\_type FP16

Model Optimizer arguments

Batch: 1

Precision of IR: FP16

Enable fusing: True

Enable gfusing: True

Names of input layers: inherited from the model

Path to the Input Model: /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/ssd/512/caffe/ssd512.caffemodel

Input shapes: inherited from the model

Log level: ERROR

Mean values: ()

IR output name: inherited from the model

Names of output layers: inherited from the model

Path for generated IR: /opt/intel/workshop/smart-video-workshop//object-detection/SSD512/FP16

Reverse input channels: False

Scale factor: None

Scale values: ()

Version: 0.3.61.37271eb9

Input proto file: deduced from the input model

Path to CustomLayersMapping.xml: extensions/front/caffe/CustomLayersMapping.xml

Path to a mean file:

Offsets for a mean file: None

[]

[ SUCCESS ] Generated IR model.

[ SUCCESS ] XML file: /opt/intel/workshop/smart-video-workshop//object-detection/SSD512/FP16/ssd512.xml

[ SUCCESS ] BIN file: /opt/intel/workshop/smart-video-workshop//object-detection/SSD512/FP16/ssd512.bin

$ python3 mo\_caffe.py --input\_model /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/ssd/300/caffe/ssd300.caffemodel -o $SV/object-detection/SSD300/FP32

Model Optimizer arguments

Batch: 1

Precision of IR: FP32

Enable fusing: True

Enable gfusing: True

Names of input layers: inherited from the model

Path to the Input Model: /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/ssd/300/caffe/ssd300.caffemodel

Input shapes: inherited from the model

Log level: ERROR

Mean values: ()

IR output name: inherited from the model

Names of output layers: inherited from the model

Path for generated IR: /opt/intel/workshop/smart-video-workshop//object-detection/SSD300/FP32

Reverse input channels: False

Scale factor: None

Scale values: ()

Version: 0.3.61.37271eb9

Input proto file: deduced from the input model

Path to CustomLayersMapping.xml: extensions/front/caffe/CustomLayersMapping.xml

Path to a mean file:

Offsets for a mean file: None

[]

[ SUCCESS ] Generated IR model.

[ SUCCESS ] XML file: /opt/intel/workshop/smart-video-workshop//object-detection/SSD300/FP32/ssd300.xml

[ SUCCESS ] BIN file: /opt/intel/workshop/smart-video-workshop//object-detection/SSD300/FP32/ssd300.bin

$ python3 mo\_caffe.py --input\_model /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/ssd/300/caffe/ssd300.caffemodel -o $SV/object-detection/SSD300/FP16 --data\_type FP16

Model Optimizer arguments

Batch: 1

Precision of IR: FP16

Enable fusing: True

Enable gfusing: True

Names of input layers: inherited from the model

Path to the Input Model: /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/ssd/300/caffe/ssd300.caffemodel

Input shapes: inherited from the model

Log level: ERROR

Mean values: ()

IR output name: inherited from the model

Names of output layers: inherited from the model

Path for generated IR: /opt/intel/workshop/smart-video-workshop//object-detection/SSD300/FP16

Reverse input channels: False

Scale factor: None

Scale values: ()

Version: 0.3.61.37271eb9

Input proto file: deduced from the input model

Path to CustomLayersMapping.xml: extensions/front/caffe/CustomLayersMapping.xml

Path to a mean file:

Offsets for a mean file: None

[]

[ SUCCESS ] Generated IR model.

[ SUCCESS ] XML file: /opt/intel/workshop/smart-video-workshop//object-detection/SSD300/FP16/ssd300.xml

[ SUCCESS ] BIN file: /opt/intel/workshop/smart-video-workshop//object-detection/SSD300/FP16/ssd300.bin

$ python3 mo\_caffe.py --input\_model /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/mobilenet-ssd/caffe/mobilenet-ssd.caffemodel -o $SV/object-detection/mobilenet-ssd/FP32

Model Optimizer arguments:

Common parameters:

- Path to the Input Model: /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/mobilenet-ssd/caffe/mobilenet-ssd.caffemodel

- Path for generated IR: /opt/intel/workshop/smart-video-workshop//object-detection/mobilenet-ssd/FP32

- IR output name: mobilenet-ssd

- Log level: ERROR

- Batch: Not specified, inherited from the model

- Input layers: Not specified, inherited from the model

- Output layers: Not specified, inherited from the model

- Input shapes: Not specified, inherited from the model

- Mean values: Not specified

- Scale values: Not specified

- Scale factor: Not specified

- Precision of IR: FP32

- Enable fusing: True

- Enable grouped convolutions fusing: True

- Move mean values to preprocess section: False

- Reverse input channels: False

Caffe specific parameters:

- Enable resnet optimization: True

- Path to the Input prototxt: /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/mobilenet-ssd/caffe/mobilenet-ssd.prototxt

- Path to CustomLayersMapping.xml: Default

- Path to a mean file: Not specified

- Offsets for a mean file: Not specified

Model Optimizer version: 1.2.110.59f62983

[]

[ SUCCESS ] Generated IR model.

[ SUCCESS ] XML file: /opt/intel/workshop/smart-video-workshop//object-detection/mobilenet-ssd/FP32/mobilenet-ssd.xml

[ SUCCESS ] BIN file: /opt/intel/workshop/smart-video-workshop//object-detection/mobilenet-ssd/FP32/mobilenet-ssd.bin

[ SUCCESS ] Total execution time: 1.40 seconds.

$ python3 mo\_caffe.py --input\_model /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/mobilenet-ssd/caffe/mobilenet-ssd.caffemodel -o $SV/object-detection/mobilenet-ssd/FP16 --data\_type FP16

Model Optimizer arguments:

Common parameters:

- Path to the Input Model: /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/mobilenet-ssd/caffe/mobilenet-ssd.caffemodel

- Path for generated IR: /opt/intel/workshop/smart-video-workshop//object-detection/mobilenet-ssd/FP16

- IR output name: mobilenet-ssd

- Log level: ERROR

- Batch: Not specified, inherited from the model

- Input layers: Not specified, inherited from the model

- Output layers: Not specified, inherited from the model

- Input shapes: Not specified, inherited from the model

- Mean values: Not specified

- Scale values: Not specified

- Scale factor: Not specified

- Precision of IR: FP16

- Enable fusing: True

- Enable grouped convolutions fusing: True

- Move mean values to preprocess section: False

- Reverse input channels: False

Caffe specific parameters:

- Enable resnet optimization: True

- Path to the Input prototxt: /opt/intel/computer\_vision\_sdk/deployment\_tools/model\_downloader/object\_detection/common/mobilenet-ssd/caffe/mobilenet-ssd.prototxt

- Path to CustomLayersMapping.xml: Default

- Path to a mean file: Not specified

- Offsets for a mean file: Not specified

Model Optimizer version: 1.2.110.59f62983

[]

[ SUCCESS ] Generated IR model.

[ SUCCESS ] XML file: /opt/intel/workshop/smart-video-workshop//object-detection/mobilenet-ssd/FP16/mobilenet-ssd.xml

[ SUCCESS ] BIN file: /opt/intel/workshop/smart-video-workshop//object-detection/mobilenet-ssd/FP16/mobilenet-ssd.bin

[ SUCCESS ] Total execution time: 1.42 seconds.

**Appendix 2： Compile App binary (tutorial1) used in the lab**

1. **Set environmental variables**

$ source /opt/intel/computer\_vision\_sdk/bin/setupvars.sh

$ cd $SV/object-detection

1. **Run “make” command**

$ make

g++ -fPIE -O3 -o tutorial1 --std=c++11 main.cpp -I. \

-I/opt/intel/computer\_vision\_sdk/opencv/include/ \

-I/opt/intel/computer\_vision\_sdk/deployment\_tools/inference\_engine/include/ \

-I/opt/intel/computer\_vision\_sdk/deployment\_tools/inference\_engine/include/cpp \

-L/opt/intel/computer\_vision\_sdk/deployment\_tools/inference\_engine/lib/ubuntu\_16.04/intel64 -linference\_engine -ldl -lpthread -lcpu\_extension \

-L/opt/intel/computer\_vision\_sdk/opencv/lib -lopencv\_core -lopencv\_imgcodecs -lopencv\_imgproc -lopencv\_highgui -lopencv\_videoio -lopencv\_video -lgflags -I/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/include -I/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/samples/ -I./ -I/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/samples/common/format\_reader/ -I/opt/intel/computer\_vision\_sdk\_2018.2.319/opencv/include -I/usr/local/include -I/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/samples/thirdparty/gflags/include -I/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/include -I/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/include/cpp -I/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/samples/extension -L/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/bin/intel64/Release/lib -L/opt/intel/computer\_vision\_sdk\_2018.2.319/deployment\_tools/inference\_engine/lib/ubuntu\_16.04/intel64 -L/opt/intel/computer\_vision\_sdk/deployment\_tools/inference\_engine/samples/build/intel64/Release/lib -L/opt/intel/computer\_vision\_sdk\_2018.2.319/opencv/lib -ldl -linference\_engine -lopencv\_highgui -lopencv\_core -lopencv\_imgproc -lopencv\_videoio -lgflags\_nothreads -lopencv\_imgcodecs -lopencv\_imgcodecs -lcpu\_extension