*TF = TensorFlow

*dldt = Intel deep learning deployment toolkit

Table 1: Given Model ID and Name

ID	Model Name	Frame Work
1	faster_rcnn_inception_v2_coco_2018_01_28	TF
2	faster_rcnn_nas_coco_2018_01_28	TF
3	faster_rcnn_resnet101_lowproposals_coco_2018_01_28	TF
4	ssd_inception_v2_coco_2018_01_28	TF
5	ssd_mobilenet_v1_ppn_shared_box_predictor_300x300_coco14_sync_2018_07_03	TF
6	ssdlite_mobilenet_v2_coco_2018_05_09	TF
7	Intel Pre-Trained person-detection-retail-0013	dldt

Table 2: RAM Utilization:

*FP = Floating Point

ID	FP	Without OpenVino (MB)	With OpenVino (MB)	Difference (%)(-MB)
1	FP32	670	236	64% (434)
1	FP16	NA	203	69.7% (467)
2	FP32	3170	Failed Out of memory	NA
2	FP16	NA	Failed Out of memory	NA
3	FP32	1290	471	63.4% (819)
3	FP16	NA	403	68.7% (887)
4	FP32	631	327	48.1% (304)
4	FP16	NA	189	70.04% (442)
5	FP32	392	120	69.3% (272)
5	FP16	NA	124	68.36% (268)
6	FP32	321	144	55.14% (177)
6	FP16	NA	126	60.74% (195)

Table 3: CPU Utilization (Approximately +- 3%):

*FP = Floating Point

ID	FP	Without OpenVino (%)	With OpenVino (%)	Difference(%) (-value)
1	FP32	50%	45%	10% (5)
1	FP16	NA	43%	14% (7)
2	FP32	58%	Failed Out of memory	NA
2	FP16	NA	Failed Out of memory	NA
3	FP32	55%	55%	0%
3	FP16	NA	46%	16.3% (9)
4	FP32	46%	45%	2.1% (1)
4	FP16	NA	45%	2.1% (1)
5	FP32	40%	39%	2.5% (1)
5	FP16	NA	41%	2.5% (+)
6	FP32	43%	40%	6.9% (3)
6	FP16	NA	40%	6.9% (3)

Table 4: Inference Time:

*FP = Floating Point

ID	FP	Without OpenVino (ms) Infer. time: [min max avg.]	With OpenVino (ms) Infer. time: [min max avg.]	Difference (-ms) In Avg. time
1	FP32	[765.55, 7456.59, **830.49**]	[453.08, 703.07, **622.33**]	25.06% (208.16)
1	FP16	NA	[437.41, 703.08, **595.9**]	28.24% (234.59)
2	FP32	[35263.16, 68917.08, **41501.58**]	Failed Out of memory	NA
2	FP16	NA	Failed Out of memory	NA
3	FP32	[1609.25, 15389.54, **1792.6**]	[1031.16, 1562.39, **1267.09**]	29.31% (525.51)
3	FP16	NA	[1031.16, 1609.28, **1349.45**]	24.72% (443.15)
4	FP32	[109.35, 7655.72, **141.96**]	[78.1, 147.09, **82.4**]	41.95% (59.56)
4	FP16	NA	[46.82, 109.37, **53.47**]	62.33% (88.49)
5	FP32	[46.86, 3765.31, **58.88**]	[15.61, 31.27, **16.84**]	71.39% (42.04)
5	FP16	NA	[15.62, 31.26, **16.87**]	71.34% (42.01)
6	FP32	[62.48, 3783.11, **69.39**]	[15.61, 31.27, **17.34**]	75.01% (52.05)
6	FP16	NA	[15.61, 31.26, **17.59**]	74.65% (51.8)

Table 5: Size Comparison:

*FP = Floating Point

ID	FP	Before [MB]	After (MB) IR+BIN	Difference% (-MB)
1	FP32	55.8	50.8	8.9% (5MB)
	FP16	NA	25.5	54.3% (30MB)
2	FP32	414.6	401	3.2% (13MB)
	FP16	NA	201	49.8% (200MB)
3	FP32	191.8	183	4.1% (8MB)
	FP16	NA	91.9	51.8% (99.1MB)
4	FP32	99.5	95.5	4% (4MB)
	FP16	NA	47.8	51.9% (51.7MB)
5	FP32	10.5	13.3	+26.6% (+2.8MB)
	FP16	NA	6.7	36.1% (3.8MB)
6	FP32	19.4	17.1	11.8% (2.3MB)
	FP16	NA	8.6	55.6% (10.8MB)

Accuracy (Counting, Detection, and Error):

1. faster_rcnn_inception_v2_coco_2018_01_28 :

Status: **Success**

This model provides very good accuracy in detection and counting. On tensorflow code this model runs with 0.5 confidence threshold successfully. However, there are some multiple detection errors.

While, on OpenVino, because of optimization and reduction in accuracy it runs with 0.95 confidence threshold successfully.

In case of good hardware availability, this model can be used to deploy an Application.

But in case of IoT devices, due to high inference time this model may not be useful.

Type	**Without OpenVino**	**With OpenVino**
FP32 Confidence Threshold: 0.5	No Of person: [1, 2, 3, 4, 5, 6] Duration: [13.7, 22.0, 19.4, 12.2, 27.7, 12.2] Error: Frame No: Count [['F: 196 C: 2'], ['F: 696 C: 2'], ['F: 1190 C: 2'], ['F: 1197 C: 2'], ['F: 1352 C: 2'], ['F: 1353 C: 2']]	No Of person: [1, 2, 3, 4, 5, 6] Duration: [12.7, 21.5, 18.7, 11.6, 25.4, 11.7] Error: N/A
FP16 Confidence Threshold: 0.95	NA	No Of person: [1, 2, 3, 4, 5, 6] Duration: [12.9, 21.6, 18.9, 12.1, 26.5, 12.0] Error: N/A

2. faster_rcnn_nas_coco_2018_01_28:

Status: Failed

This models provides highest accuracy in detection, on tensor flow code it takes approximately 40Sec time to process each frame, which is not good for IoT devices with hardware limitations.

While on OpenVino, Model fails to load and throws memory error.

3. faster_rcnn_resnet101_lowproposals_coco_2018_01_28

Status: **Success**

This model accuracy is moderate. On tensor flow code it runs with confidence threshold 0.5 successfully, while on OpenVino it runs with confidence 0.9 due to loss in accuracy during conversion. Although there are some multiple detection occurred, but one or two frame differences are taken case in filter in program.

With good resources this model can be deployed on edge.

But for IoT, because of high inference time, model cannot be usefull.

Type	**Without OpenVino**	**With OpenVino**
FP32	Confidence: 0.5 No Of person: [1, 2, 3, 4, 5, 6]	Confidence: 0.9 No Of person: [1, 2, 3, 4, 5, 6]

Confidence Threshold: 0.5 & 0.9	Duration: [13.6, 22.0, 19.6, 12.0, 27.4, 12.2] Error: N/A	Duration: [13.1, 21.7, 19.1, 11.9, 26.9, 12.2] Error: Frame No: Count [['F: 186 C: 2'], ['F: 1178 C: 2'], ['F: 1184 C: 2']]
FP16 Confidence Threshold:	NA	No Of person: [1, 2, 3, 4, 5, 6] Duration: [13.1, 21.7, 19.1, 11.9, 26.9, 12.2] Error: Frame No: Count [['F: 186 C: 2'], ['F: 1178 C: 2'], ['F: 1184 C: 2']]

4. ssd_inception_v2_coco_2018_01_28

Status: Failed

This model is has good detection accuracy and inference time on tensorflow code, but after conversion, on OpenVino it fails to count person and duration at confidence 0.1 due to reduction in accuracy.

Type	**Without OpenVino**	**With OpenVino**
FP32 Confidence Threshold: 0.3	No Of person: [1, 2, 3, 4, 5, 6] Duration: [10.3, 11.5, 17.6, 11.9, 19.9, 12.2] Error: N/A	Failed
FP16 Confidence Threshold:	NA	Failed

5. ssd_mobilenet_v1_ppn_shared_box_predictor_300x300_coco14_sync_2018_07_03

Status: Failed

This model has very low detection accuracy by default, and it fails even on TensorFlow code, at confidence 0.49 it misses the count and at confidence 0.5, multiple detection increases. Thus not compatible to deploy an app.

Type	**Without OpenVino**	**With OpenVino**
FP32 Confidence Threshold: 0.49 & 0.5	Failed Confidence: 0.49 No Of person: [1, 2, 3, 4, 5]	NA

	Error: N/A Confidence: 0.5 No Of person: [1, 2, 3, 4, 5] Error: Frame No: Count [['F: 185 C: 2'], ['F: 188 C: 2'], ['F: 689 C: 2'], ['F: 691 C: 2'], ['F: 857 C: 2'], ['F: 858 C: 2'], ['F: 859 C: 3'], ['F: 1187 C: 2'], ['F: 1190 C: 2']]	
FP16 Confidence Threshold:	NA	NA

6. ssdlite_mobilenet_v2_coco_2018_05_09

Status: Failed

This model successfully runs on tensor flow code , but after conversion, on OpenVino it fails to count the person due to decrease in accuracy, at confidence 0.3 it misses the person count and at confidence 0.5 it detects multiple boxes and fails the overall counting.

Thus, this models cannot be used deploy an app on OpenVino platform.

Type	**Without OpenVino**	**With OpenVino**
FP32 Confidence Threshold: 0.3 & 0.5	No Of person: [1, 2, 3, 4, 5, 6] Duration: [10.7, 9.4, 16.5, 11.9, 22.9, 12.2] Error: N/A	Failed Conf: 0.3 No Of person: [1, 2, 3] Error: N/A Conf: 0.5 No Of person: [1, 2, 3, 4, 5] Error : Frame No: Count [['F: 188 C: 2'], ['F: 189 C: 2'], ['F: 190 C: 2'], ['F: 232 C: 2'], ['F: 440 C: 2'], ['F: 857 C: 2'], ['F: 1190 C: 2']]
FP16 Confidence Threshold:	NA	Failed

7. Intel Pre-Trained person-detection-retail-0013

Status: **Success**

This model is from intel open model zoo and pretrained and optimized, it works perfectly and fulfill the edge processing criteria in terms of inference time, performance and accuracy.

This model is perfect for the app and the IoT requirements.

Type	Stats	Utilization
FP32 Confidence Threshold: 0.5	No Of person: [1, 2, 3, 4, 5, 6] Duration: [12.7, 21.4, 18.1, 11.6, 26.1, 11.1] Error: N/A	**RAM:** 100MB **CPU:** 38% **Size:** 2.90MB **Inference time:[min max avg.]** [12.11, 46.87, 18.2]
FP16 Confidence Threshold: 0.5	No Of person: [1, 2, 3, 4, 5, 6] Duration: [12.7, 21.4, 18.1, 11.6, 26.1, 11.1] Error: N/A	**RAM:** 80MB **CPU:** 38% **Size:** 1.52MB **Inference time:[min max avg.]** [15.61, 46.87, 18.11]
INT 8 Confidence Threshold: 0.5	No Of person: [1, 2, 3, 4, 5, 6] Duration: [12.7, 21.3, 17.1, 11.6, 25.8, 11.1] Error: N/A	**RAM:** 80MB **CPU:** 40% **Size:** 1.52MB **Inference time:[min max avg.]** [46.85, 141.41, 67.77]