# Configurations

Create

## Configurations Settings

| # | Model | | Validation Dataset | Target | Precisions | Creation Time | Latency (ms) | Throughput (FPS) | Accuracy (%) | Status | Actions |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | frozen_inference_graph Details | | Refined30_coco | Machine: Local Workstation<br>Device: CPU | FP16 | 21/09/20, 09:22 | 951.76 | 0.99 | N/A | ✓ | |
| 2 | frozen_inference_graph - INT8 | INT8- Default<br>Refined30_coco | Refined30_coco | Machine: Local Workstation<br>Device: CPU | | 21/09/20, 09:45 | N/A | N/A | N/A | ! | |

*resize to input size possible, only for one input layer case*

Compare

### Selected Configuration

frozen_inference_graph • Refined30_coco • Local Workstation • CPU

**Profile** | Optimize | Pack

Select Inference Method

**Single Inference**

Parallel streams(1-4):

1

Batch size (1-256):

1

Group Inference

Execute

**Inference Tips**

Combinations of streams and batch sizes appear in real time on the Inference Results plot below and are characterized with throughput and latency values. Inference is executed asynchronously. To profile multiple combinations of parameters in sequence, use group inference.

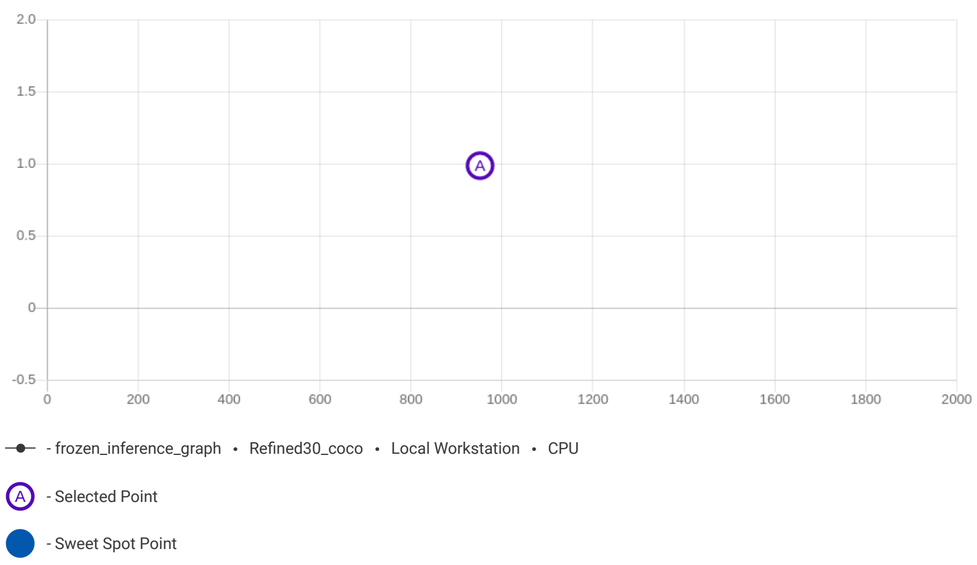
## Model Performance Summary

### Inference Results

frozen_inference_graph • Refined30_coco • Local Workstation • CPU

Latency Threshold (0-1000)

Throughput, fps



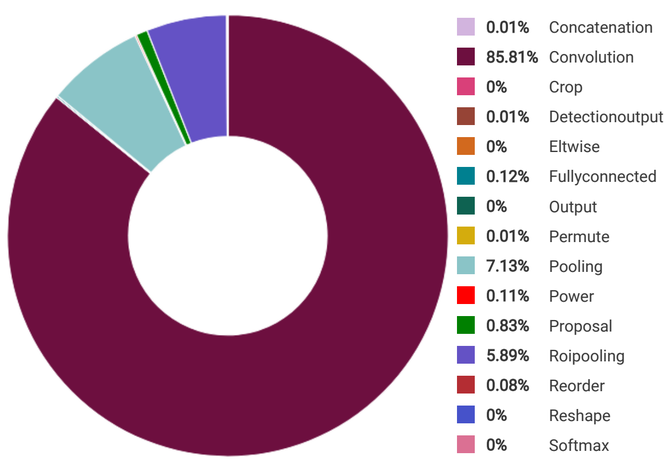

Expand Chart

Expand Table

| # | Streams | Batch size | Throughput, fps | Latency, ms | Last status | Filter |
|---|---|---|---|---|---|---|
| A | 1 | 1 | 0.99 | 951.76 | ✓ | ☐ |

### Execution Time By Layer

frozen_inference_graph • Refined30_coco • Local Workstation • CPU



| | |
|---|---|
| 0.01% | Concatenation |
| 85.81% | Convolution |
| 0% | Crop |
| 0.01% | Detectionoutput |
| 0% | Eltwise |
| 0.12% | Fullyconnected |
| 0% | Output |
| 0.01% | Permute |
| 7.13% | Pooling |
| 0.11% | Power |
| 0.83% | Proposal |
| 5.89% | Roipooling |
| 0.08% | Reorder |
| 0% | Reshape |
| 0% | Softmax |

### Execution Attributes

frozen_inference_graph • Refined30_coco • Local Workstation • CPU

FPS: **0.99**
Latency (ms): **951.76**
Total Execution Time (ms): **21129.38**
Batch size: **1**
Streams: **1**

### Layers Table

frozen_inference_graph • Refined30_coco • Local Workstation • CPU

Visualize original IR | Visualize runtime graph

Select Column | Select Filter | + Add new filter

Apply Filter | Clear Filter

| Execution Order | Layer Name | Layer Type | Execution Time, ms | Precision |
|---|---|---|---|---|
| 2 | image_tensor | Input | Not Executed | U8 |
| 3 | image_tensor_U8_FP32_Preprocessor/sub | Reorder | 0.501 | FP32 |
| 4 | Preprocessor/sub | Power | 1.312 | FP32 |
| 5 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/Conv2d_1a_7x7/separable_conv2d/depthwise | Convolution | 4.647 | FP32 |
| 6 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/Conv2d_1a_7x7/BatchNorm/FusedBatchNormV3/variance/Fused_Add_ | Convolution | 3.826 | FP32 |
| 7 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/MaxPool_2a_3x3/MaxPool | Pooling | 2.449 | FP32 |
| 8 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/Conv2d_2b_1x1/BatchNorm/FusedBatchNormV3/variance/Fused_Add_ | Convolution | 1.627 | FP32 |
| 9 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/Conv2d_2c_3x3/BatchNorm/FusedBatchNormV3/variance/Fused_Add_ | Convolution | 43.971 | FP32 |
| 10 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/MaxPool_3a_3x3/MaxPool | Pooling | 2.165 | FP32 |
| 11 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/Mixed_3b/Branch_0/Conv2d_0a_1x1/BatchNorm/FusedBatchNormV3/variance/Fused_Add_ | Convolution | 1.174 | FP32 |
| 12 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/Mixed_3b/Branch_1/Conv2d_0a_1x1/BatchNorm/FusedBatchNormV3/variance/Fused_Add_ | Convolution | 1.179 | FP32 |
| 13 | FirstStageFeatureExtractor/InceptionV2/InceptionV2/Mixed_3b/Branch_1/Conv2d_0b_3x3/BatchNorm/FusedBatchNormV3/variance/Fused_Add_ | Convolution | 3.874 | FP32 |