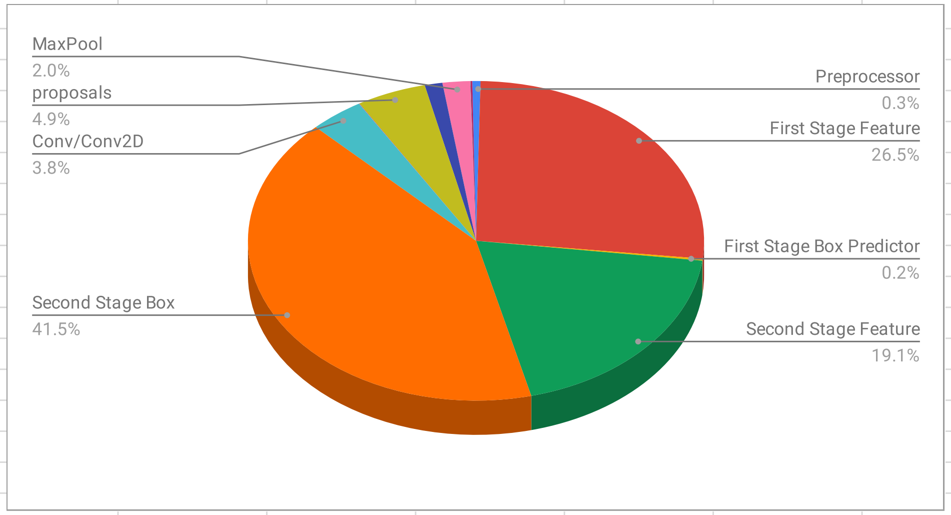
## Mask-RCNN-OpenVino-InceptionV2_FP32_CPU_i9

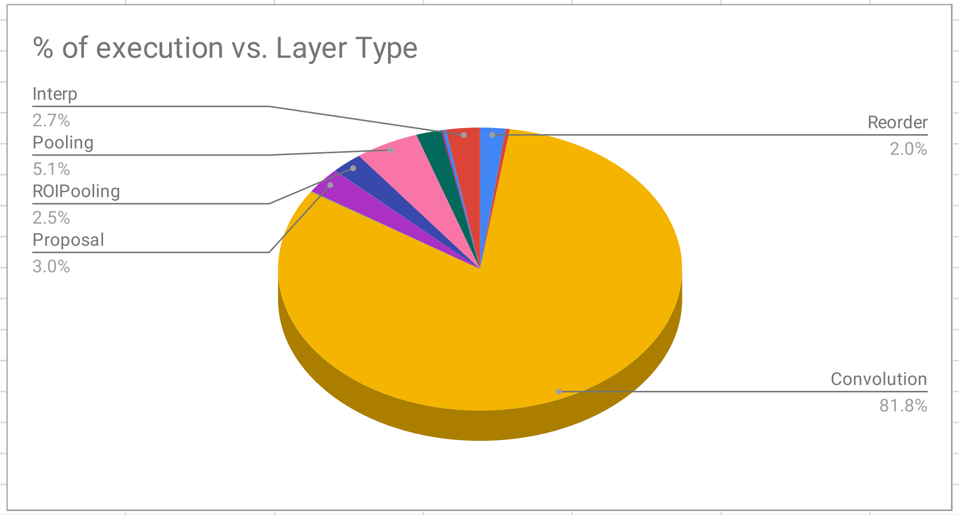
**FP32 CPU – Execution division by layer function**

| Layer Name | Execution time(micro seconds) | % of execution |
|---|---|---|
| Preprocessor | 604 | 0.32% |
| First Stage Feature Extractor | 50573 | 26.47% |
| First Stage Box Predictor | 367 | 0.19% |
| Second Stage Feature Extractor | 36530 | 19.12% |
| Second Stage Box Predictor | 79262 | 41.49% |
| Conv/Conv2D | 7290 | 3.82% |
| predictions/Reshape/Softmax | 53 | 0.03% |
| proposals | 9438 | 4.94% |
| CropAndResize | 2400 | 1.26% |
| MaxPool | 3748 | 1.96% |
| reshape | 48 | 0.03% |
| ScaleShift/scale_locs | 205 | 0.11% |
| detection_output | 509 | 0.27% |
| **Total** | **191027** | 100.00% |



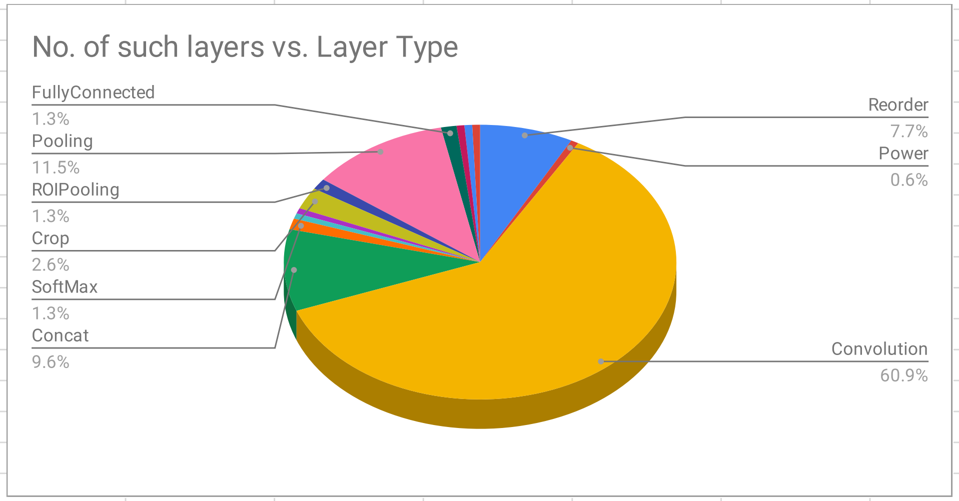
**FP32 CPU – Execution division by layer type**

| Layer Type | Execution time(micro seconds) | % of execution |
|---|---|---|
| Reorder | 3685 | 2.03% |
| Power | 604 | 0.33% |
| Convolution | 148687 | 81.82% |
| Concat | 88 | 0.05% |
| SoftMax | 39 | 0.02% |
| Permute | 34 | 0.02% |
| Proposal | 5373 | 2.96% |
| Crop | 24 | 0.01% |
| ROIPooling | 4581 | 2.52% |
| Pooling | 9303 | 5.12% |
| FullyConnected | 3670 | 2.02% |
| ScaleShift | 205 | 0.11% |
| DetectionOutput | 509 | 0.28% |
| Interp | 4922 | 2.71% |
| **Total** | **181724** | 100.00% |


% of execution vs. Layer Type

Total execution time in both the above tables should come out to be same, there is a small difference between two value because we've neglected some layers with negligible execution time.

| Layer Type | No. of such layers | Average execution time per layer |
|---|---|---|
| Reorder | 12 | 307.0833333 |
| Power | 1 | 604 |
| Convolution | 95 | 1565.126316 |
| Concat | 15 | 5.866666667 |
| SoftMax | 2 | 19.5 |
| Permute | 1 | 34 |
| Proposal | 1 | 5373 |
| Crop | 4 | 6 |
| ROIPooling | 2 | 2290.5 |
| Pooling | 18 | 516.8333333 |
| FullyConnected | 2 | 1835 |
| ScaleShift | 1 | 205 |
| DetectionOutput | 1 | 509 |
| Interp | 1 | 4922 |
| **Total** | **156** | **1164.897436** |


No. of such layers vs. Layer Type

**Observations**

Around 82% of time is spend on convolution

Execution time of  single convolution is also comparatively higher, and larger no. of them makes it a bottleneck.

If we look at the layers functionality-wise, most time is spent on first stage feature extraction, second stage feature extraction and box prediction.

**Proposal**

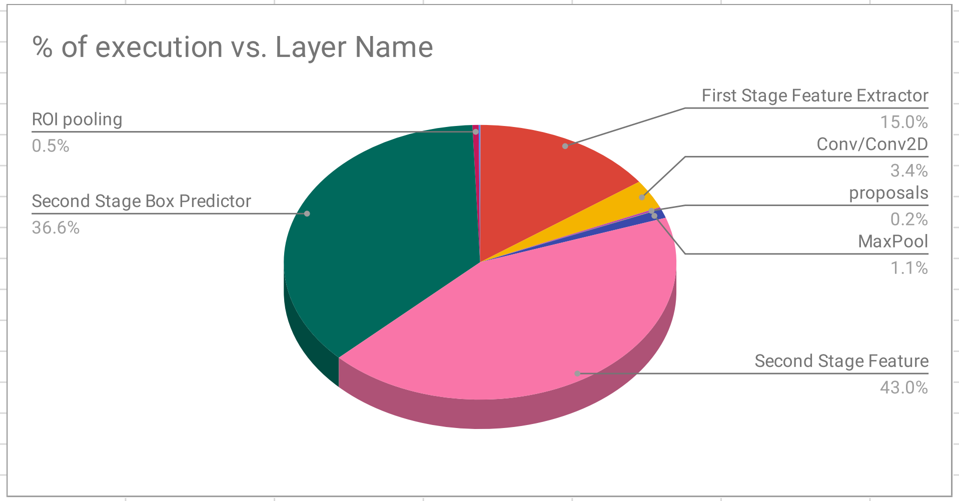
We can try to optimize the convolution layer functionality, that will decrease total inference time significantly, because it is the most used layer.

Additionally, if we can study and optimize first and second stage feature extraction and box prediction, it will further improve the inference time.

**Mask-RCNN-OpenVino-InceptionV2_FP16_NCS_1**

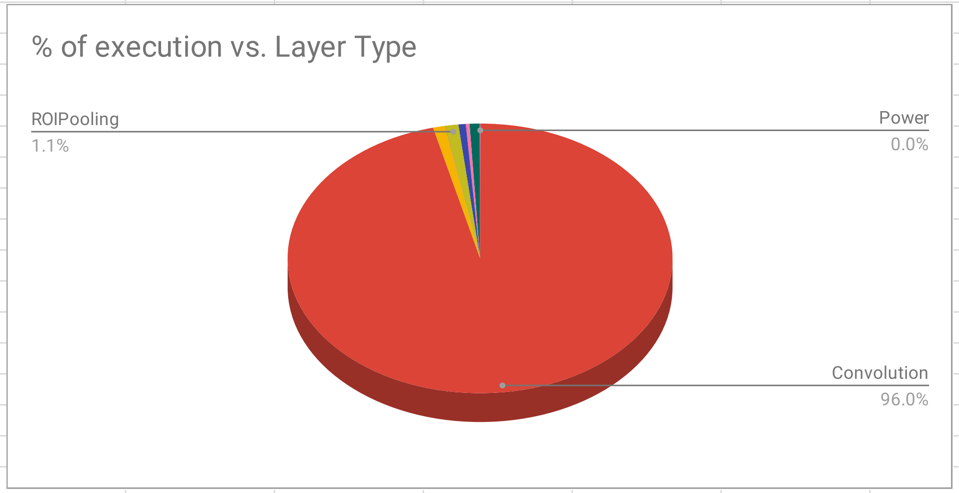
**FP16 NCS1 – Execution division by layer function**

| Layer Name | Execution time(micro seconds) | % of execution |
|---|---|---|
| Preprocessor | 2451 | 0.03% |
| First Stage Feature Extractor | 1073609 | 15.01% |
| Conv/Conv2D | 239974 | 3.35% |
| Conv/Relu | 1267 | 0.02% |
| First Stage Box Predctor | 3733 | 0.05% |
| predictions/Reshape | 401 | 0.01% |
| proposals | 11904 | 0.17% |
| crop | 3272 | 0.05% |
| MaxPool | 76208 | 1.07% |
| Second Stage Feature Extractor | 3079136 | 43.03% |
| Second Stage Box Predictor | 2617991 | 36.59% |
| ROI pooling | 36487 | 0.51% |
| masks | 8560 | 0.12% |
| **Total** | **7154993** | 100.00% |



**FP16 NCS1 – Execution division by layer type**

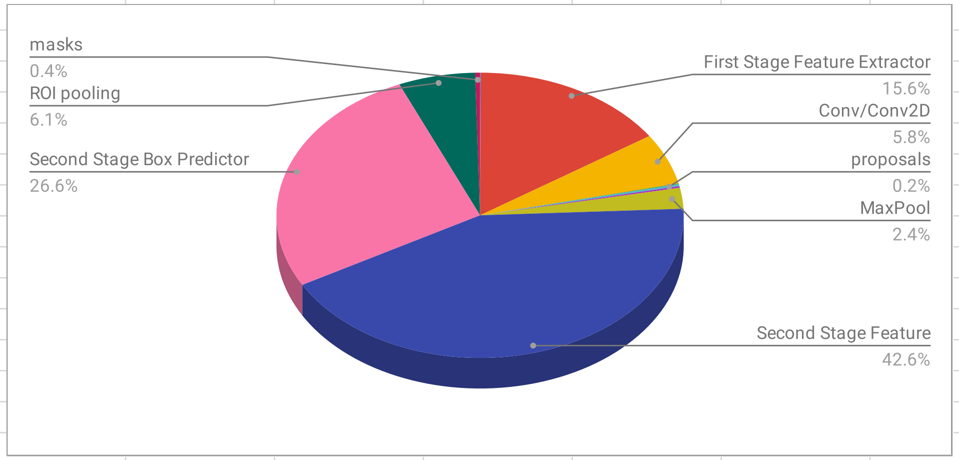
| Layer Type | Execution time(micro seconds) | % of execution |
|---|---|---|
| Power | 2451 | 0.03% |
| Convolution | 6750290 | 94.43% |
| ReLU | 69441 | 0.97% |
| Clamp | 1267 | 0.02% |
| Proposal | 2766 | 0.04% |
| Permute | 91 | 0.00% |
| Reshape | 256 | 0.00% |
| ROIPooling | 75533 | 1.06% |
| Pooling | 43792 | 0.61% |
| Concat | 21809 | 0.31% |
| FullyConnected | 58348 | 0.82% |
| Crop | 3272 | 0.05% |
| **Total** | **7148641** | 100.00% |



## Mask-RCNN-OpenVino-InceptionV2_FP16_NCS_2

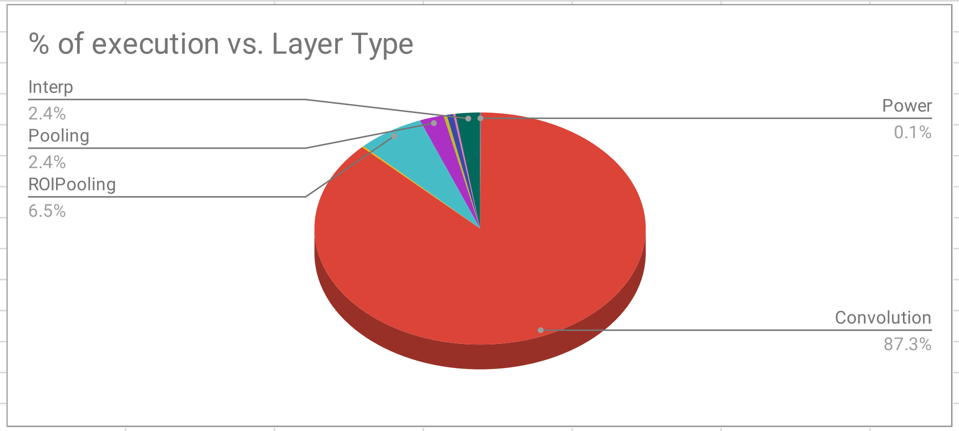
**FP16 NCS2 – Execution division by layer function**

| Layer Name | Execution time(micro seconds) | % of execution |
|---|---|---|
| Preprocessor | 1279 | 0.06% |
| First Stage Feature Extractor | 333049 | 15.56% |
| Conv/Conv2D | 124213 | 5.80% |
| First Stage Box Predctor | 1470 | 0.07% |
| predictions/Reshape | 286 | 0.01% |
| proposals | 4625 | 0.22% |
| crop | 3994 | 0.19% |
| MaxPool | 50490 | 2.36% |
| Second Stage Feature Extractor | 912231 | 42.62% |
| Second Stage Box Predictor | 569822 | 26.62% |
| ROI pooling | 130707 | 6.11% |
| masks | 8189 | 0.38% |
| **Total** | **2140355** | 100.00% |



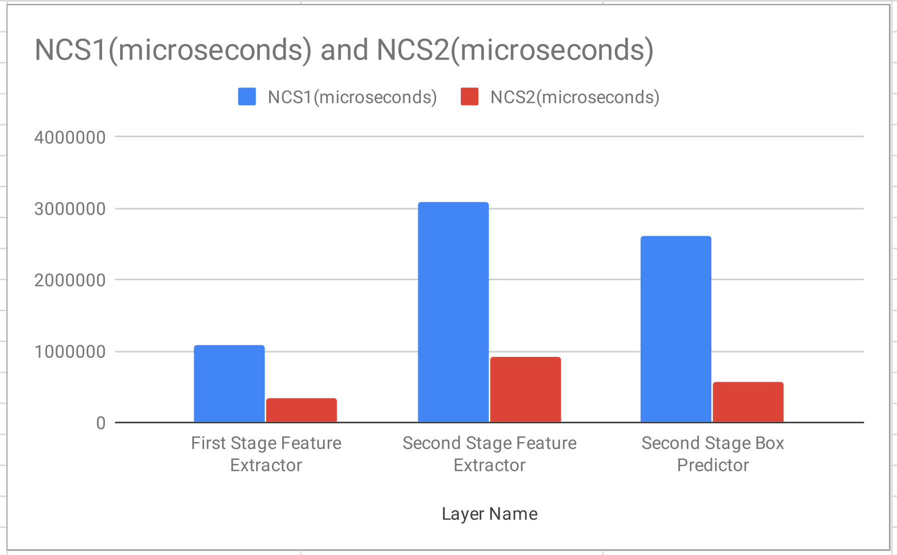
**FP16 NCS2 – Execution division by layer type**

| Layer Type | Execution time(micro seconds) | % of execution |
|---|---|---|
| Power | 1279 | 0.06% |
| Convolution | 1747345 | 87.30% |
| Proposal | 3354 | 0.17% |
| | | |
| Permute | 74 | 0.00% |
| Reshape | 260 | 0.01% |
| ROIPooling | 130707 | 6.53% |
| Pooling | 47323 | 2.36% |
| Concat | 6416 | 0.32% |
| FullyConnected | 12722 | 0.64% |
| Crop | 3994 | 0.20% |
| Interp | 48045 | 2.40% |
| **Total** | **2001519** | 100.00% |


% of execution vs. Layer Type

**Comparison between most time consuming layers on NCS1 and NCS2**

| Layer Name | NCS1(microseconds) | NCS2(microseconds) |
|---|---|---|
| First Stage Feature Extractor | 1073609 | 333049 |
| Second Stage Feature Extractor | 3079136 | 912231 |
| Second Stage Box Predictor | 2617991 | 569822 |



**Observations**

Here we can see that NCS2 is considerably faster than NCS1. But convolution is taking maximum time on all 3 devices.