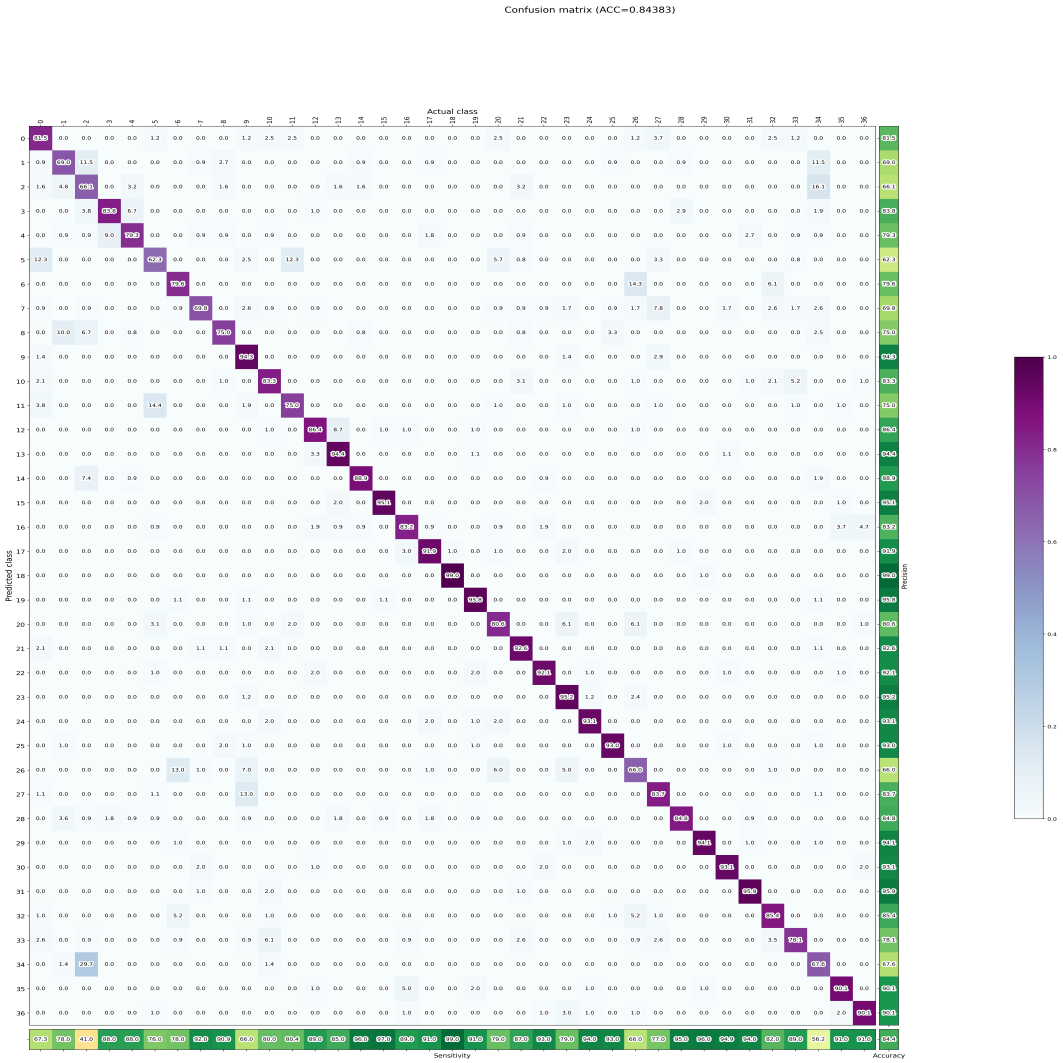
Apache TVM DNN compiler - Summary

Patryk Rybak

1 Performance and quality data for experiments

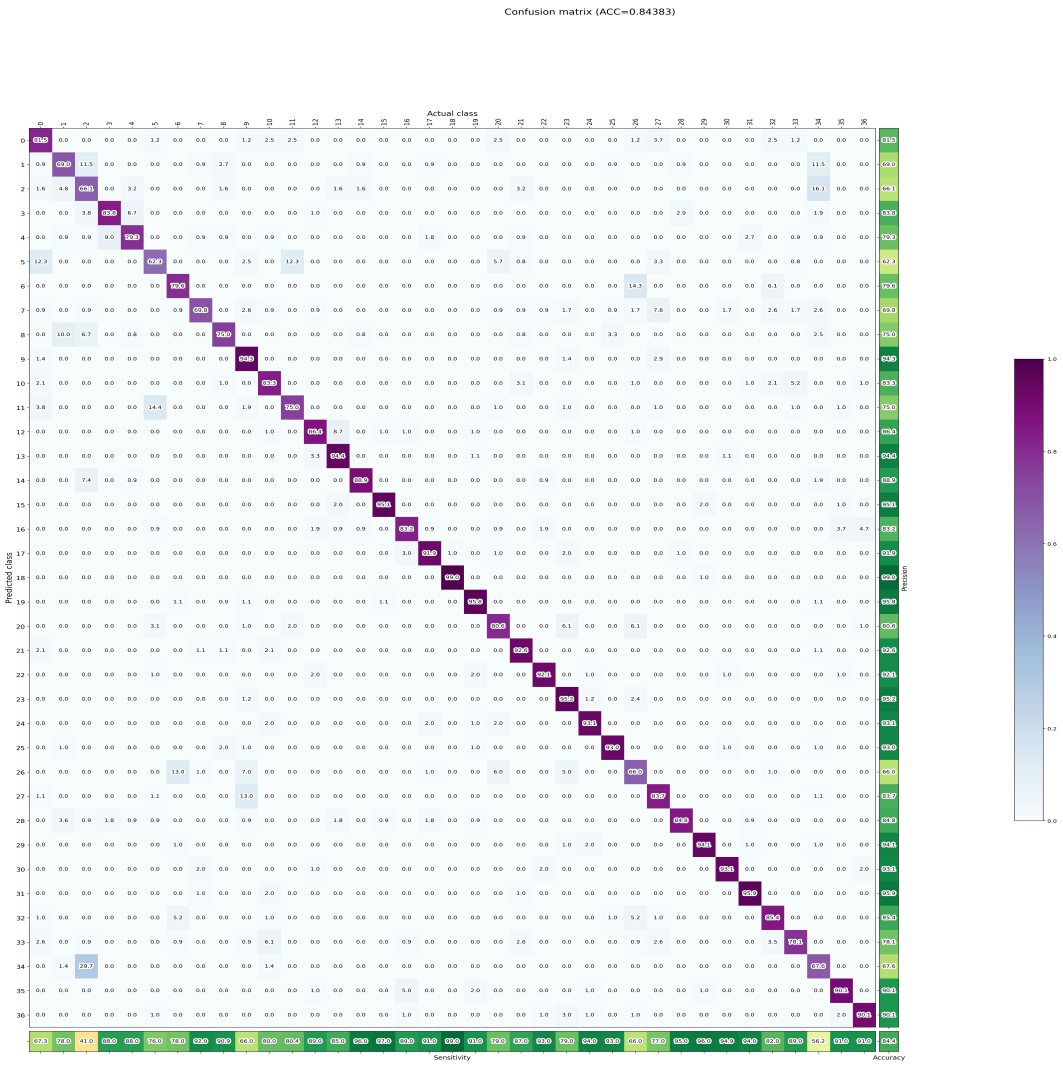
1.1 tvm-fp32-nhwc

```
Model type: tvm-fp32-nhwc

* Accuracy: 0.8438266557645135
* Mean precision: 0.8431981147525323
* Mean sensitivity: 0.8457146480923832
* G-Mean: 0.8394524839258236
* Mean inference time: 23.787992332906065 ms
* Top-5 percentage: 0.026474553048551414
```



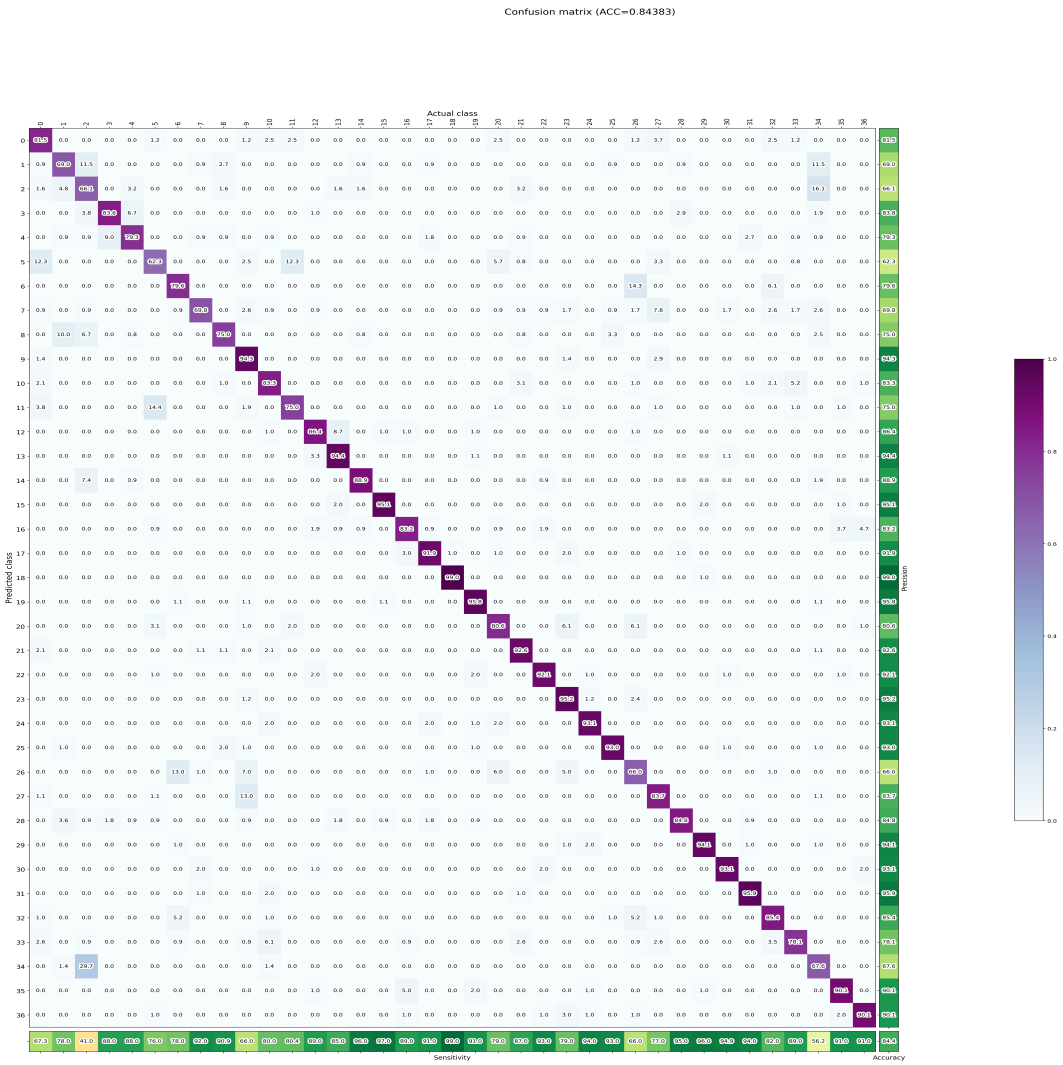
1.2 tvm-fp32-opt1

```
Model type: tvm-fp32-opt1

* Accuracy: 0.8438266557645135
* Mean precision: 0.8431981147525323
* Mean sensitivity: 0.8457146480923832
* G-Mean: 0.8394524839258236
* Mean inference time: 25.561388446628232 ms
* Top-5 percentage: 0.026474553048551414
```



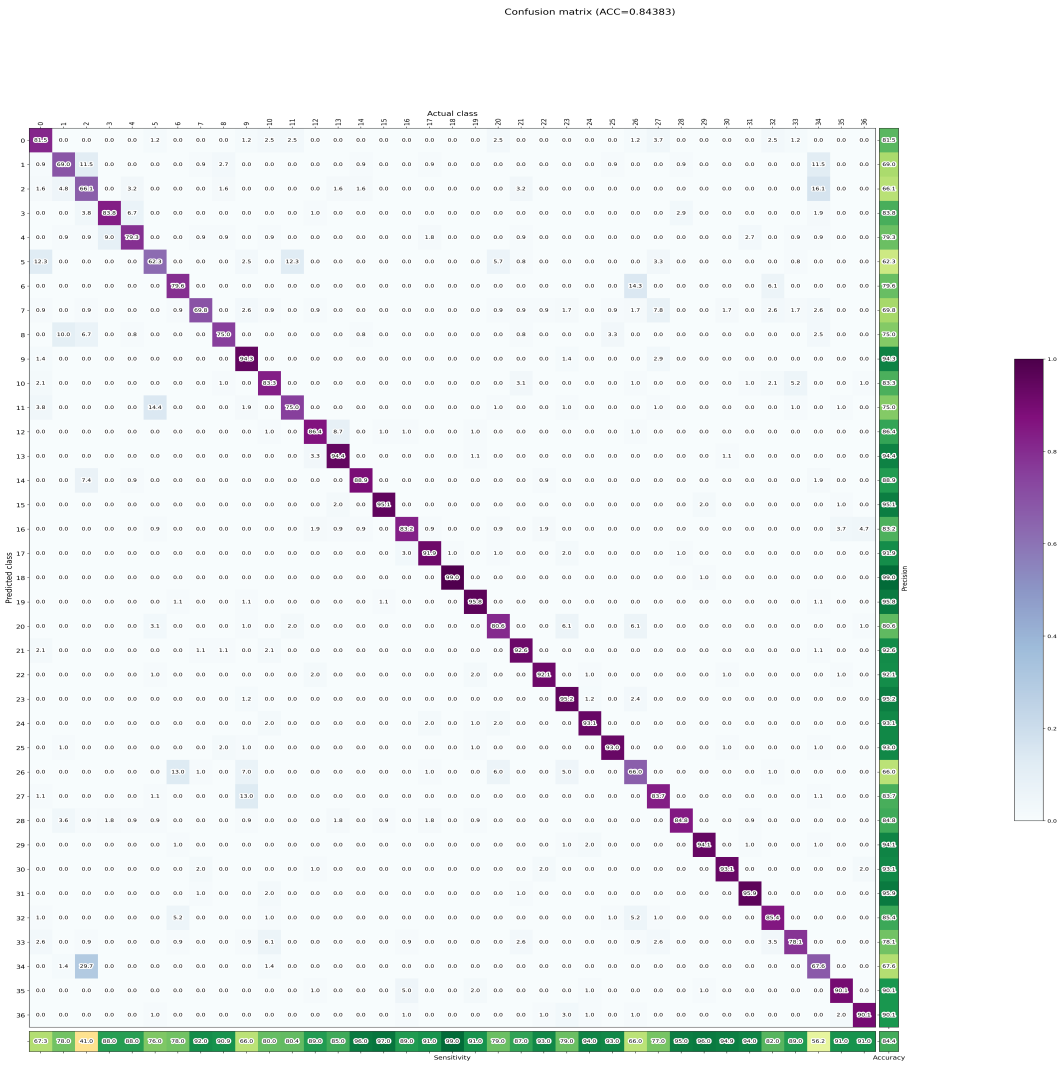
1.3 tvm-fp32-opt2

```
Model type: tvm-fp32-opt2

* Accuracy: 0.8438266557645135
* Mean precision: 0.8431981147525323
* Mean sensitivity: 0.8457146480923832
* G-Mean: 0.8394524839258236
* Mean inference time: 25.783313955182887 ms
* Top-5 percentage: 0.026474553048551414
```



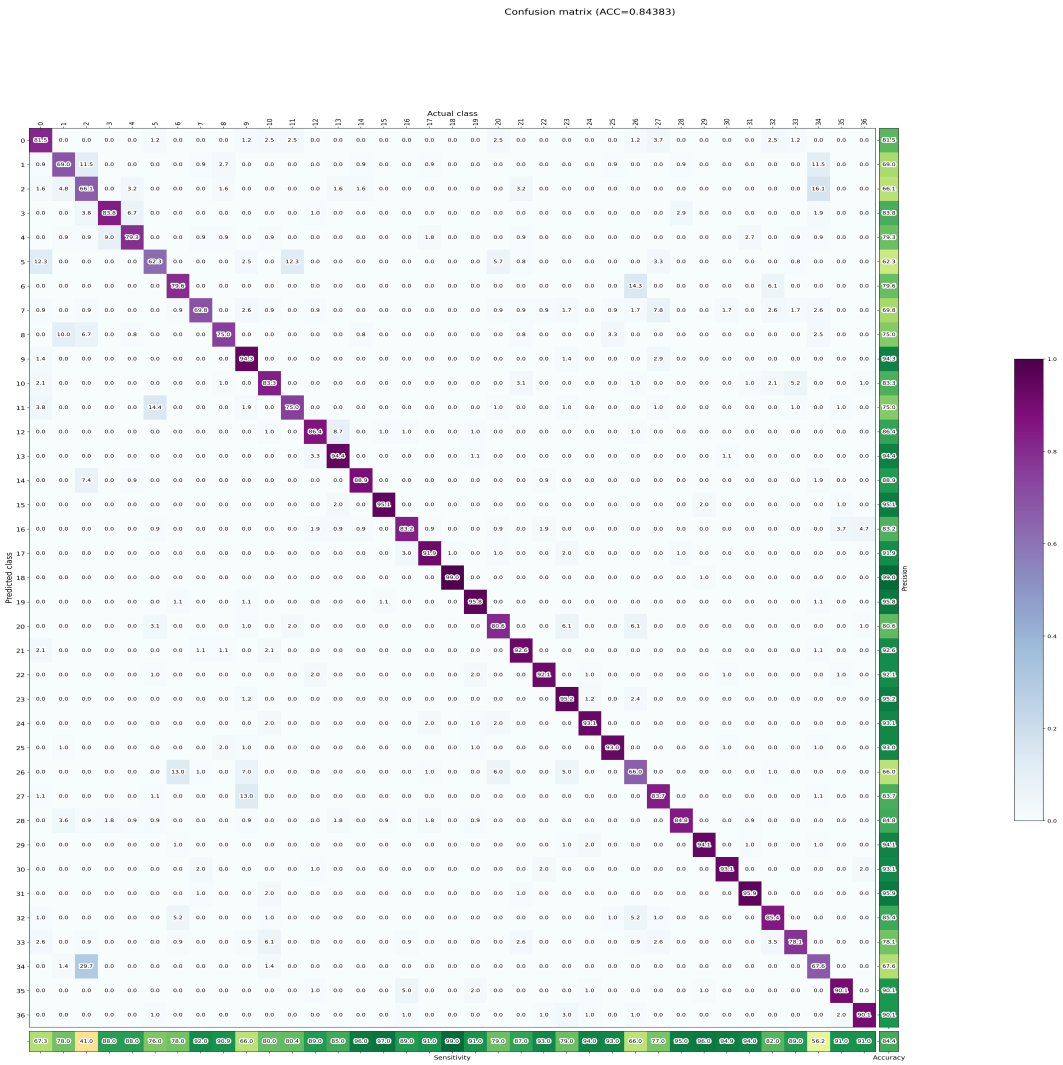
1.4 tvm-fp32-opt3

```
Model type: tvm-fp32-opt3

* Accuracy: 0.8438266557645135
* Mean precision: 0.8431981147525323
* Mean sensitivity: 0.8457146480923832
* G-Mean: 0.8394524839258236
* Mean inference time: 10.766283809174928 ms
* Top-5 percentage: 0.026474553048551414
```



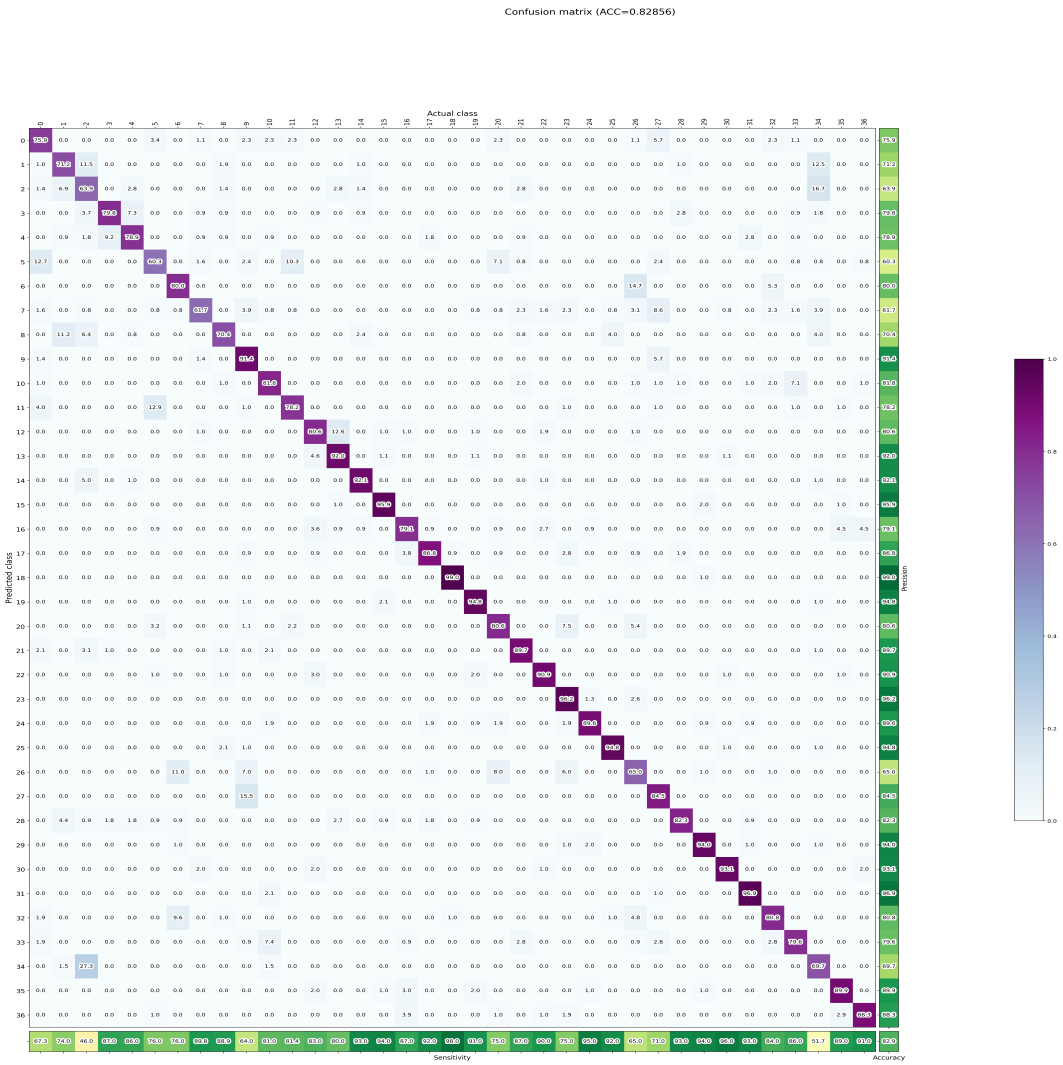
1.5 tvm-fp32-opt4

```
Model type: tvm-fp32-opt4

* Accuracy: 0.8438266557645135
* Mean precision: 0.8431981147525323
* Mean sensitivity: 0.8457146480923832
* G-Mean: 0.8394524839258236
* Mean inference time: 10.900282528326835 ms
* Top-5 percentage: 0.026474553048551414
```



1.6 tvm-int8

```
Model type: tvm-int8

* Accuracy: 0.8285636413191605
* Mean precision: 0.8278585516497349
* Mean sensitivity: 0.8323670364632081
* G-Mean: 0.8253678760555472
* Mean inference time: 16.749437722703302 ms
* Top-5 percentage: 0.026474553048551414
```



2 Logs

2.1 NHWC model

```
TVM MODEL WITH NHWC LAYOUT
conv2d NHWC layout is not optimized for x86 with autotvm.
depthwise_conv2d NHWC layout is not optimized for x86 with autotvm.
One or more operators have not been tuned. Please tune your model for better
    ↪ performance. Use DEBUG logging level to see more details.
...
evaluating tvm-fp32-nhwc: 100%|| 3669/3669 [01:46<00:00, 34.49it/s]
```

2.2 INT8 model before fixing conversion to NCHW

```
TVM PRE-QUANTIZED MODEL WITH NCHW LAYOUT
[15:22:51] /workspace/tvm/src/relay/transforms/convert_layout.cc:99: Warning:
    ↪ Desired layout(s) not specified for op: qnn.conv2d
conv2d NHWC layout is not optimized for x86 with autotvm.
depthwise_conv2d NHWC layout is not optimized for x86 with autotvm.
...
evaluating tvm-int8: 100%|| 3669/3669 [02:22<00:00, 25.68it/s]
```

2.3 INT8 model after fixing conversion to NCHW

```

TVM PRE-QUANTIZED MODEL WITH NCHW LAYOUT
evaluating tvm-int8: 100%|| 3669/3669 [00:33<00:00, 108.06it/s]
```

3 FP32 TFLite and INT8 TFLite comparation

Model Type	Accuracy	Mean Inference Time (ms)
tvm-fp32-nhwc	0.8438	23.40
tvm-int8	0.8389	3.79

4 TFLite and TVM models size comparation

Model Type	File Size	Filename
TFLite FP32	16 MB	pet-dataset-tensorflow.fp32.tflite
TVM FP32 (all opts)	17 MB	pet-dataset-tensorflow.fp32.tvm-fp32-{nhwc,opt1,...,opt4}.so
TVM INT8	5.9 MB	pet-dataset-tensorflow.int8-0.8.tvm-int8.so

5 Answers to questions

5.1 How does the TFLite FP32 model compare to the NHWC-compiled TVM model?

Metric	TVM FP32 NHWC	TFLite FP32
Accuracy	0.8438	0.8357
Mean Precision	0.8432	0.8350
Mean Sensitivity	0.8457	0.8363
G-Mean	0.8395	0.8311
Mean Inference Time (ms)	23.40	15.29
Top-5 Percentage	0.0265	1.0000

5.2 How does the TFLite FP32 model compare to the fastest NCHW-compiled TVM model?

Metric	TVM INT8 (Opt 3)	TFLite FP32
Accuracy	0.8389	0.8357
Mean Precision	0.8382	0.8350
Mean Sensitivity	0.8406	0.8363
G-Mean	0.8346	0.8311
Mean Inference Time (ms)	3.79	15.29
Top-5 Percentage	0.0265	1.0000

5.3 How did switching from NHWC to NCHW in FP32 and INT8 models affect the performance of the model (compare the NHWC and NCHW models with opt level 3)?

Metric	FP32 NHWC	FP32 NCHW (Opt3)	INT8 NCHW (Opt3)
Accuracy	0.8438	0.8438	0.8389
Mean Precision	0.8432	0.8432	0.8382
Mean Sensitivity	0.8457	0.8457	0.8406
G-Mean	0.8395	0.8395	0.8346
Mean Inference Time (ms)	23.40	6.14	3.79
Top-5 Percentage	0.0265	0.0265	0.0265

5.4 How does the TFLite INT8 model compare to the INT8 TVM model (after fixing layout conversion)?

The TVM INT8 model runs over 10× faster than the TFLite INT8 model.

5.5 How does the models with opt levels 1, 2, 3, 4 compare to each other performance- and quality-wise?

The TVM models with optimization levels 1 to 4 show identical accuracy and other quality metrics. However, inference time differs significantly - opt1 and opt2 run around 25 ms, while opt3 and opt4 achieve about 6 ms, providing roughly a 4× speedup. There is no meaningful difference in performance or quality between opt3 and opt4. In short, higher optimization levels greatly improve speed without affecting model quality.

5.6 How using the AVX2 instruction set affected all of the compiled models?

Enabling AVX2 had no impact on model quality but significantly improved performance—especially for opt-level 3 and 4 models (from 10.8ms to 6.1ms) and the INT8 model (from 16.7ms to 3.8ms). For lower optimization levels, the difference was minimal.