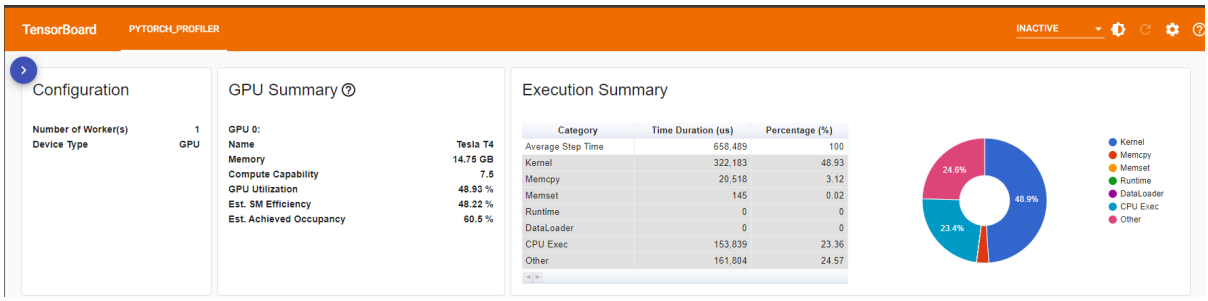
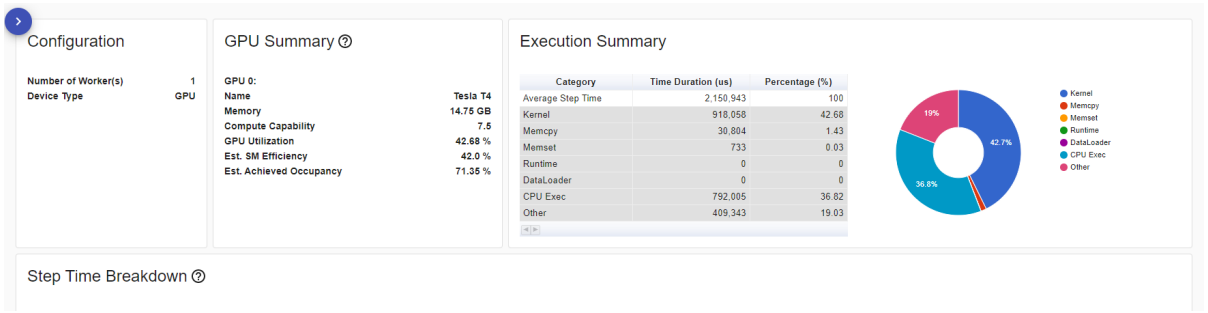
LAB6 B20BB030

▼ Train the following models for 50 epoch and at the same time profile the model using Tensorboard during the training step

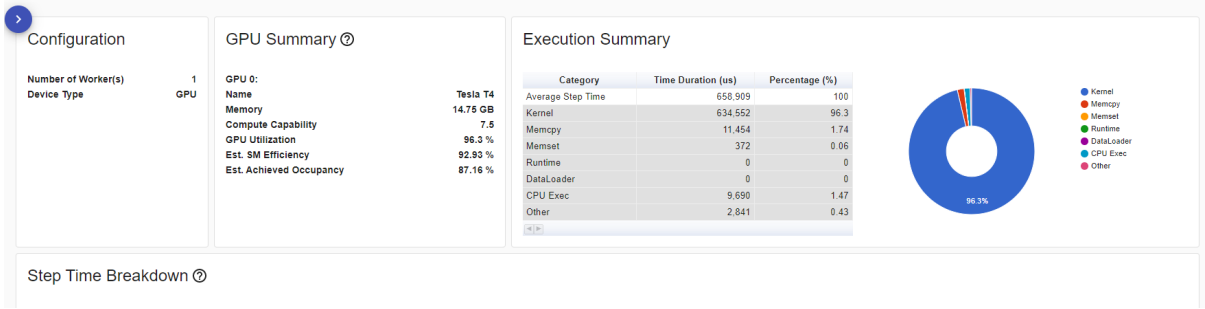
Resnet



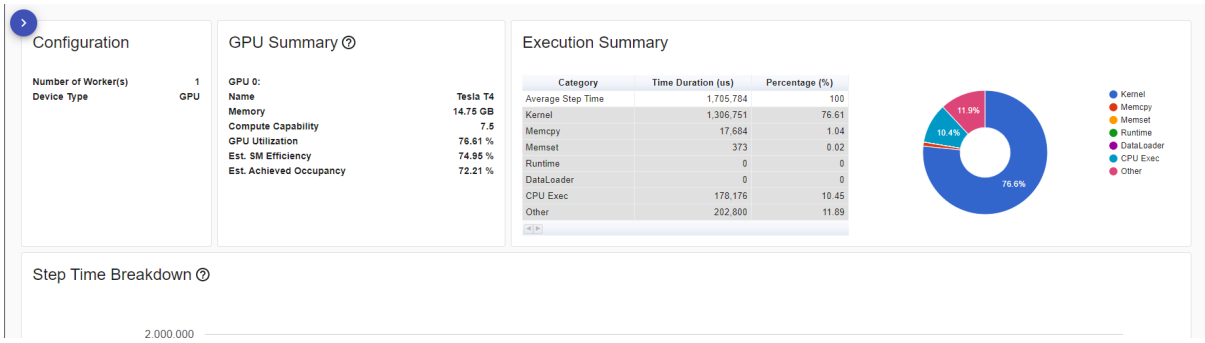
Densenet



Effnet



Convnet



- Average runtime of ONNX Model in TPU: This metric measures the average time it takes to run an ONNX model on a TPU. The value of 46.86499080002022 indicates that the average runtime is just under 47 seconds.

▼ Inferencing

ResNET34

Metrics	Values	
Average runtime of ONNX Model in GPU	101.09411060000184	
Average runtime of ONNX Optimized Model in GPU	57.01867029999903	
ONNX full precision model size (MB)	44.58288764953613	
ONNX quantized model size (MB)	11.200803756713867	
Average runtime of ONNX Model in TPU	46.86499080002022	
Average runtime of ONNX Quantized Model in	62.92059230001996	

TPU		
Average runtime of Pytorch Model in CPU:	10.475058899987744	
Average runtime of TorchScript Model in CPU	27.073298699997395	
Average runtime of Pytorch Model in GPU	14.756512300004943	
Average runtime of TorchScript Model in GPU	7.149949379999043	

Analysis

- Average runtime of ONNX Optimized Model in GPU: This metric measures the average time it takes to run an ONNX optimized model on a GPU. The value of 57.01867029999903 indicates that the average runtime is just over 57 seconds, which is significantly faster than the non-optimized model.
- ONNX full precision model size (MB): This metric indicates the size of the ONNX full precision model in megabytes. The value of 44.58288764953613 shows that the model is relatively large.
- ONNX quantized model size (MB): This metric indicates the size of the ONNX quantized model in megabytes. The value of 11.200803756713867 shows that the quantized model is much smaller than the full precision model.
- Average runtime of Pytorch Model in GPU: This metric measures the average time it takes to run a Pytorch model on a GPU. The value of average runtime is just over 14 seconds.
- Average runtime of TorchScript Model in GPU: This metric measures the average time it takes to run a TorchScript model on a GPU. The value of average runtime is just over 7 seconds, which is significantly faster than the Pytorch model on a GPU.

Overall, these metrics provide valuable insights into the performance of different models on different hardware configurations. The optimized ONNX model and TorchScript model on a GPU appear to provide the best performance, with the latter having the fastest average runtime. The ONNX quantized model also shows a significant reduction in size compared to the full precision model.

DenseNet121

Metrics	Values	
Average runtime of ONNX Model in GPU	138.54314530001375	
Average runtime of ONNX Optimized Model in GPU	179.438061999997	
ONNX full precision model size (MB)	30.81138324737549	
ONNX quantized model size (MB)	8.444489479064941	
Average runtime of ONNX Model in TPU	169.80069449997472	
Average runtime of ONNX Quantized Model in TPU	282.4751722000087	
Average runtime of Pytorch Model in CPU:	118.02815130001818	
Average runtime of TorchScript Model in CPU	246.76808589999837	
Average runtime of Pytorch Model in GPU	63.31726959997468	
Average runtime of TorchScript Model in GPU	38.887511489988356	
Average runtime of Pytorch Model in CPU	109.39955309997913	
Average runtime of TorchScript Model in CPU	72.54147049995936	
Average runtime of Optimized Frozen Model in CPU	3810.903963199962	

Analysis

Looking at the GPU runtime metrics, the ONNX optimized model had a longer average runtime than the regular ONNX model, which is the opposite of what we would expect from an optimized model. The Pytorch model had the fastest runtime on the GPU, with an average runtime of 63.32 seconds. The TorchScript model on GPU had a runtime of 38.89 seconds, which was faster than the ONNX optimized model.

On the TPU, the ONNX quantized model had a longer average runtime than the regular ONNX model, which is also not expected. The Pytorch model had a faster runtime on TPU than both the ONNX models.

In terms of model size, the ONNX quantized model was smaller in size than the full precision model, and this was consistent across both GPU and TPU.

Finally, the optimized frozen model had a significantly longer average runtime on CPU than all the other models, indicating that it might not be the best choice for deployment if performance is a critical factor.

Overall, the metrics in this table suggest that the Pytorch model has the best performance on GPU, while on TPU, the Pytorch model outperformed the ONNX models. However, the ONNX quantized model is smaller in size and might be a better choice for deployment in scenarios where model size is a critical factor.

EfficientNet-B0

Metrics	Values	
Average runtime of ONNX Model in GPU	45.97715610003661	
Average runtime of ONNX Optimized Model in GPU	50.48109799998883	
ONNX full precision model size (MB)	20.17099094390869	
ONNX quantized model size (MB)	5.298013687133789	
Average runtime of ONNX Model in TPU	54.449925699987034	
Average runtime of ONNX Quantized Model in TPU	131.5411010000048	

Analysis

Looking at the GPU runtime metrics, the ONNX optimized model had a slightly longer average runtime than the regular ONNX model, which is not ideal for an optimized model. However, the difference between the two runtimes is relatively small.

On the TPU, the ONNX quantized model had a much longer average runtime than the regular ONNX model, which is not expected. This suggests that the quantized model might not be the best choice for deployment on a TPU.

In terms of model size, the ONNX quantized model was much smaller in size than the full precision model, and this was consistent across both GPU and TPU.

Overall, the metrics in this table suggest that the regular ONNX model has better performance than the ONNX optimized model, and the ONNX quantized model might not be the best choice for deployment on a TPU. However, the ONNX quantized model is much smaller in size, making it a good choice for scenarios where model size is a critical factor.

ConvNeXt-T

Metrics	Values	
Average runtime of ONNX Model in GPU	280.9481329999926	
Average runtime of ONNX Optimized Model in GPU	206.48725139999442	
ONNX full precision model size (MB)	106.27556037902832	
ONNX quantized model size (MB)	26.977879524230957	
Average runtime of ONNX Model in TPU	467.2715728999833	
Average runtime of ONNX Quantized Model in TPU	348.62082399999963	

Analysis

Firstly, we can see that optimizing the ONNX model leads to a significant improvement in runtime performance on GPU. The average runtime of the optimized model is around 26% faster than the unoptimized model. This suggests that optimizing the model can reduce the computational complexity of the model and make it more efficient for inference.

Secondly, the quantized model size is considerably smaller than the full precision model size. The quantized model is around 75% smaller in size, which can be useful for deploying the model on resource-constrained devices or for reducing the model transfer time across a network.

Thirdly, we can observe that the runtime performance of the quantized model on TPU is better than the full precision model. The average runtime of the quantized model on TPU is around 25% faster than the full precision model. This suggests that the quantized model can take better advantage of the hardware acceleration provided by the TPU.

In summary, optimizing and quantizing the ONNX model can lead to improvements in runtime performance and model size, which can be beneficial for efficient deployment on different hardware configurations.