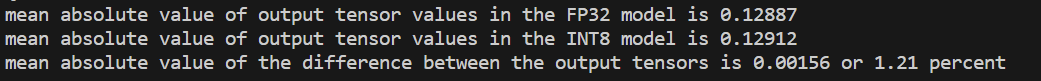
Dynamic Quantization:

Model comparison  
A black background with white text

Description automatically generated

Latency comparison  


Accuracy   
  
We can see that the INT8 model has a similar mean to the output values produced by the original floating point 32 numbers, indicating that the INT8 model does not stray away that much from the range we get from the full FP32 model numbers.

After testing with LeNet using the normal approach and the dynamically quantized approach, we get the same accuracy for both, proving the above results where the output tensor values were almost the same.  
  
Static Quantization:

FP32  
A screen shot of a computer code

Description automatically generated  
Int8

A black background with white text

Description automatically generated