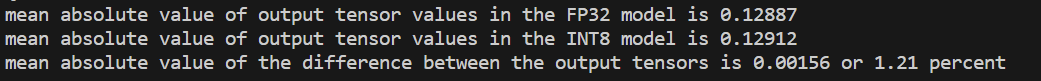
Model comparison  
A black background with white text

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.