

## EML Task 11.2

### 1. Task 11.2.1 - FP32 inference performance

First, we need to calculate the total number of samples:

Total samples = 10 files \* 32 samples per file = 320

Device	# Run	Time NetRun in s	Time per sample in s
CPU	1	1.592255	0.004975796875
GPU	1	0.422884	0.001321512500
CPU	5	1.668083	0.005212759375
GPU	5	0.428665	0.001339578125
CPU	10	1.647424	0.005148200000
GPU	10	0.421595	0.001317484375

Averaged on runs 1,5 and 10, the GPU was about 3.855 times faster than the CPU.

### 2. Task 11.2.2 - INT8 inference performance

First, we need to calculate the total number of samples:

Total samples = 10 files \* 32 samples per file = 320

Device	# Run	Time NetRun in s	Time per sample in s
CPU	1	1.60531	0.00501659375
HTP	1	0.01951	0.00006096875
CPU	5	1.66475	0.00520233125
HTP	5	0.02124	0.00006637500
CPU	10	1.64828	0.00515087812
HTP	10	0.02128	0.00006650000