Edge Devices	Tensorflow Version	Model (CNN, NLP)	Dataset	Quantization	OS (Ubuntu/L4T/Docker)	CUDA version
Nvidia Jetson TX1, Nvidia Jetson TX2, Nvidia Jetson Nano, Nvidia Jetson Xavier, Raspberry pi 4 + Coral TPU	2.5	Mobilenet V1	ImageNet	Tensorflow FP32, EdgeTPU-tflite, TensorLite, TensorRT	18.04 / 32.6.1 / 20.10	10.2.300
		Mobilenet V2				
		Inception V3				
		YOLO V5	COCO			
		RNN	IMDb	Tensorflow FP32		
		LSTM				
		BERT				
		DistilBERT	GLUE SST-2			

Image Classification

MobileNet V1

Edge device	Setup					Metric									
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	1	0.93	61.89632869	1.739581585	217.8288004	0.2164418902	3.552841483	4.620177725	1.15	4.25	270.5882353
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	2	0.93	61.51903725	1.723010063	148.9873488	0.1481337805	4.711879271	6.75065469	2.3	4.25	541.1764706
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	4	0.93	62.32127237	1.713058949	117.6271176	0.1170231295	5.504740629	8.545319242	4.59	4.25	1080
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	8	0.93	62.01025939	1.716651917	92.39070773	0.09191520786	6.405421004	10.87959243	9.18	4.25	2160
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	16	0.93	61.78713679	1.732764959	86.46241069	0.08533826636	6.720786032	11.71807259	18.4	4.25	4329.411765
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	32	0.93	62.20449257	1.717960358	82.92469335	0.08051635884	6.973230069	12.41983635	36.7	4.25	8635.294118
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	64	0.93	61.71406102	1.71338439	82.85037923	0.08030400076	7.000369394	12.4526797	73.4	4.25	17270.58824
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	128	0.93	62.36174202	1.72446537	82.63167596	0.07977245632	6.979366466	12.53565511	1.47E+02	4.25	3.46E+04
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	1	0.93	59.92601061	1.6062181	99.50904369	0.09781955647	6.209582935	10.22290466	1.15	4.25	270.5882353
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	2	0.93	58.7174046	1.612814188	52.33815145	0.05133878899	8.875593008	19.47844933	2.3	4.25	541.1764706
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	4	0.93	58.23746061	1.570019484	29.02927065	0.0283911581	11.25658273	35.2222335	4.59	4.25	1080
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	8	0.93	59.38877749	1.59959507	17.8277023	0.01733194232	12.68773127	57.69693792	9.18	4.25	2160
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	16	0.93	58.12470531	1.609985113	14.69704032	0.01255437117	13.54257294	79.65353156	18.4	4.25	4329.411765
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	32	0.93	59.00644207	1.628488541	13.61142302	0.008781591197	13.79189427	113.8745789	36.7	4.25	8635.294118
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	64	0.93	59.67605972	1.60624814	14.32554483	0.007201014087	13.54351559	138.8693298	73.4	4.25	17270.58824
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	128	0.93	59.1225841	1.563191414	15.37611389	0.006354051176	13.4626794	157.3799096	1.47E+02	4.25	3.46E+04
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	1	0.93	38.88625741	1.037235498	65.82744575	0.06412833238	9.456155237	15.59373155	1.15	4.25	270.5882353
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	2	0.93	38.01113439	1.054350138	32.73214889	0.03176783514	13.92799128	31.47838043	2.3	4.25	541.1764706
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	4	0.93	39.47181606	1.041965723	18.81584477	0.01820332026	16.85487822	54.93503303	4.59	4.25	1080
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	8	0.93	38.30404139	1.030306578	12.16145039	0.01171571183	19.41897256	85.35546233	9.18	4.25	2160
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	16	0.93	38.36752224	1.058549881	8.744111538	0.008018319332	20.92574659	124.7144144	18.4	4.25	4329.411765
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	32	0.93	38.1481607	1.017022848	8.580517292	0.00614240556	21.44679822	162.8026659	36.7	4.25	8635.294118
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	64	0.93	38.18341804	1.039550543	8.765650034	0.005155852996	21.33827334	193.9543274	73.4	4.25	17270.58824
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	128	0.93	37.66842175	1.027097225	10.46026754	0.004670296796	20.83164779	214.1191543	1.47E+02	4.25	3.46E+04
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	1	0.93	71.08252144	2.345131636	156.859957	0.1526636727	4.342385849	6.550346801	1.15	4.25	270.5882353
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	2	0.93	71.12680101	2.0991745	112.5438108	0.1095119245	5.38299902	9.131425683	2.3	4.25	541.1764706
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	4	0.93	70.64579511	2.058488846	85.60306263	0.07203072262	6.31681362	13.88296499	4.59	4.25	1080
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	8	0.93	69.93868399	2.046374798	62.63308215	0.05643707871	7.42839512	17.71884766	9.18	4.25	2160
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	16	0.93	70.81421661	2.031164169	68.55348921	0.05673198426	7.128752148	17.62674113	18.4	4.25	4329.411765
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	32	0.93	70.34334397	2.068148136	107.1531734	0.08601080929	5.702673581	11.62644566	36.7	4.25	8635.294118
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	64								73.4	4.25	17270.58824
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V1	imagenet 1000	TensorFlow 2.5V	FP32	128								1.47E+02	4.25	3.46E+04
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V1	imagenet 1000	edge tpu tflite - TFv2	INT8	1	0.7466875	3.13670969	0.6347105503	49.95246148	0.01541767024	18.61369596	64.86064267			
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V1	imagenet 1000	edge tpu tflite - TFv2	INT8	2										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V1	imagenet 1000	edge tpu tflite - TFv2	INT8	4										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V1	imagenet 1000	edge tpu tflite - TFv2	INT8	8										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V1	imagenet 1000	edge tpu tflite - TFv2	INT8	16										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V1	imagenet 1000	edge tpu tflite - TFv2	INT8	32										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V1	imagenet 1000	edge tpu tflite - TFv2	INT8	64										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V1	imagenet 1000	edge tpu tflite - TFv2	INT8	128										

MobileNet V2

Edge device	Setup					Metric									
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	1	0.906	109.1029143	1.727538347	293.5764148	0.2921756997	2.472756283	3.422598118	0.615	3.53	174.2209632
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	2	0.906	109.6560349	1.734726191	178.889364	0.1780199685	3.444946552	5.617347359	1.23	3.53	348.4419263
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	4	0.906	109.0418844	1.745179176	120.6408343	0.1200416715	4.320997474	8.330440483	2.46	3.53	696.8838527
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	8	0.906	109.4922254	1.732740164	91.53054547	0.09105279469	4.932044945	10.98263928	4.92	3.53	1393.767705
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	16	0.906	110.0245981	1.719760418	77.94719481	0.07689336202	5.313885991	13.00502376	9.84	3.53	2787.535411
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	32	0.906	109.3342836	1.724636555	71.17544556	0.06904999563	5.619132907	14.48226015	19.7	3.53	5580.736544
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	64	0.906	108.9346631	1.739596367	69.4636488	0.06725450023	5.684527536	14.86889348	39.4	3.53	11161.47309
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	128	0.906	109.5537689	1.752038479	73.42853951	0.07055806066	5.543077358	14.17272514	78.7	3.53	22294.61756
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	1	0.906	103.5648978	1.562505245	157.4630725	0.1557951264	3.808201653	6.41868602	0.615	3.53	174.2209632
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	2	0.906	106.1847472	1.639361858	82.4679985	0.08146191406	5.255071299	12.27567522	1.23	3.53	348.4419263
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	4	0.906	105.7118359	1.621422529	44.76714158	0.04409062576	6.574598219	22.68055812	2.46	3.53	696.8838527
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	8	0.906	103.6506927	1.601008654	26.21338344	0.02567373419	7.606573418	38.95031368	4.92	3.53	1393.767705
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	16	0.906	103.4871049	1.615100384	16.7524879	0.01612562225	8.272137276	62.01311084	9.84	3.53	2787.535411
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	32	0.906	103.3839259	1.592667103	14.63284445	0.01109476807	8.561186624	90.13257366	19.7	3.53	5580.736544
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	64	0.906	103.3000824	1.605649471	15.17880464	0.008916406659	8.527309113	112.1528031	39.4	3.53	11161.47309
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	128	0.906	104.9155915	1.605816126	16.29356837	0.007716460619	8.337731763	129.5930932	78.7	3.53	22294.61756
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	1	0.906	66.94901466	1.029476643	92.66682172	0.09103277969	6.224886781	10.98505399	0.615	3.53	174.2209632
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	2	0.906	66.94992638	1.034856796	48.19621539	0.04725036407	8.607235995	21.16385809	1.23	3.53	348.4419263
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	4	0.906	67.81008959	1.019435406	26.64441657	0.02604661059	10.47391131	38.39271127	2.46	3.53	696.8838527
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	8	0.906	66.61496758	1.004122734	15.34189391	0.01490839934	12.05383741	67.0762821	4.92	3.53	1393.767705
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	16	0.906	66.90273213	1.034251213	10.25050116	0.009799251008	12.89206094	102.0486157	9.84	3.53	2787.535411
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	32	0.906	66.81817079	1.010535479	8.577705383	0.007023798302	13.40196461	142.3731088	19.7	3.53	5580.736544
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	64	0.906	67.01150298	1.009224176	8.856770515	0.00599429966	13.31984638	166.82516	39.4	3.53	11161.47309
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	128	0.906	67.0898447	1.022275925	10.45342684	0.005248077912	13.03367568	190.5459516	78.7	3.53	22294.61756
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	1	0.906	127.2817204	2.542519569	253.3297305	0.2492911539	2.609913198	4.011373787	0.615	3.53	174.2209632
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	2	0.906	125.8207233	2.181414604	173.7648814	0.1711566045	3.31381096	5.842602468	1.23	3.53	348.4419263
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	4	0.906	125.2173266	2.570510387	158.1518993	0.1498946307	3.497234534	6.67135304	2.46	3.53	696.8838527
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	8	0.906	125.2875311	2.136475801	147.7138095	0.08967195725	3.634533248	11.15175837	4.92	3.53	1393.767705
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	16	0.906	125.7341638	2.186259031	88.78350043	0.07893936454	4.651493628	12.66795097	9.84	3.53	2787.535411
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	32								19.7	3.53	5580.736544
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	64								39.4	3.53	11161.47309
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Mobilenet V2	imagenet 1000	TensorFlow 2.5V	FP32	128								78.7	3.53	22294.61756
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	1	0.7055976563	3.117435455	1.01078105	50.14280176	0.01652078829	18.42604086	60.52979934			
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	2										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	4										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	8										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	16										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	32										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	64										
Google Coral TPU	18.04 / 20.10	Coral-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	128										
Google Coral TPU	18.04 / 20.10	Custom-Mobilenet V2	imagenet 1000	edge tpu tflite - TFv2	INT8	1	0.6149140625	3.109596014	0.4486391544	40.48142219	0.01313121428	22.706807	76.15441943			

Inception V3

Edge device	Setup					Metric									
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	1	0.919	208.8427234	1.775926113	946.9159076	0.9453367631	0.8639049841	1.057824089	11.5	23.8	483.1932773
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	2	0.919	210.038877	1.778414011	784.4297447	0.7834146128	1.003766933	1.276463298	22.9	23.8	962.1848739
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	4	0.919	202.1471162	1.7247262	638.339205	0.6377227693	1.187350558	1.568079498	45.9	23.8	1928.571429
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	8	0.919	208.753346	1.721056461	579.4679682	0.5789691052	1.265914895	1.727207879	91.8	23.8	3857.142857
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	16	0.919	208.3320928	1.715524197	543.3618851	0.5385665203	1.337917558	1.856780848	1.84E+02	23.8	7.73E+03
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	32	0.919	209.9941964	1.72806263	516.7463899	0.5041320317	1.405688143	1.983607343	3.67E+02	23.8	1.54E+04
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	64								7.34E+02	23.8	3.08E+04
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	128								1.47E+03	23.8	6.18E+04
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	1	0.919	190.2802024	1.556982994	254.3616292	0.2526663547	2.241152661	3.957788529	11.5	23.8	483.1932773
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	2	0.919	200.4228044	1.607478857	134.0156994	0.1329811215	2.97578067	7.519864387	22.9	23.8	962.1848739
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	4	0.919	200.8904629	1.611971378	76.80513835	0.07612413049	3.580280456	13.13643904	45.9	23.8	1928.571429
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	8	0.919	200.8694873	1.588033199	82.24314785	0.04796550465	3.512457797	20.84831604	91.8	23.8	3857.142857
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	16	0.919	199.5682058	1.647502184	70.44460678	0.04358266436	3.71051367	22.94490286	1.84E+02	23.8	7.73E+03
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	32	0.919	200.6132371	1.582264423	77.282377	0.04378624214	3.663973145	22.83822386	3.67E+02	23.8	1.54E+04
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	64	0.919	199.9199872	1.591580391	109.5638173	0.07443897449	3.291804238	13.4338229	7.34E+02	23.8	3.08E+04
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	128	0.919	199.3992963	1.573651314	185.3213027	0.1487039209	2.650827289	6.724772246	1.47E+03	23.8	6.18E+04
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	1	0.919	132.2537916	1.042407036	147.0621588	0.1454411669	3.566860513	6.875632405	11.5	23.8	483.1932773
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	2	0.919	132.1550634	1.044339657	76.47736812	0.07555388546	4.769238864	13.23558668	22.9	23.8	962.1848739
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	4	0.919	131.8163593	1.031774044	82.13434196	0.04199156785	4.651537333	23.8143049	45.9	23.8	1928.571429
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	8	0.919	132.4758158	1.018234968	35.21715236	0.02569772863	5.927262872	38.91394505	91.8	23.8	3857.142857
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	16	0.919	132.8492031	1.030791521	35.47679901	0.02195885707	5.95192651	45.53971079	1.84E+02	23.8	7.73E+03
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	32	0.919	132.9358473	1.021033049	37.76926231	0.02061955165	5.96297753	48.49765974	3.67E+02	23.8	1.54E+04
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	64	0.919	133.0718038	1.035824299	49.36590672	0.03192617209	5.581181505	31.32226429	7.34E+02	23.8	3.08E+04
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	128	0.919	126.1647036	1.033347368	76.42906809	0.05759925744	5.028794448	17.36133493	1.47E+03	23.8	6.18E+04
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	1								11.5	23.8	483.1932773
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	2								22.9	23.8	962.1848739

Image Classification

MobileNet V1

Edge device	Setup					Metric									
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	4								45.9	23.8	1928.571429
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	8								91.8	23.8	3857.142857
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	16								1.84E+02	23.8	7.73E+03
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	32								3.67E+02	23.8	1.54E+04
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	64								7.34E+02	23.8	3.08E+04
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	Inception V3	imagenet 1000	TensorFlow 2.5V	FP32	128								1.47E+03	23.8	6.18E+04
Google Coral TPU	18.04 / 20.10	Coral-Inception V3	imagenet 1000	edge tpu tflite - TFv1	INT8	1	0.9133398438	3.178961277	0.442332983	577.8533888	0.5539650221	1.71976533	1.805168124			
Google Coral TPU	18.04 / 20.10	Coral-Inception V3	imagenet 1000	edge tpu tflite - TFv1	INT8	2										
Google Coral TPU	18.04 / 20.10	Coral-Inception V3	imagenet 1000	edge tpu tflite - TFv1	INT8	4										
Google Coral TPU	18.04 / 20.10	Coral-Inception V3	imagenet 1000	edge tpu tflite - TFv1	INT8	8										
Google Coral TPU	18.04 / 20.10	Coral-Inception V3	imagenet 1000	edge tpu tflite - TFv1	INT8	16										
Google Coral TPU	18.04 / 20.10	Coral-Inception V3	imagenet 1000	edge tpu tflite - TFv1	INT8	32										
Google Coral TPU	18.04 / 20.10	Coral-Inception V3	imagenet 1000	edge tpu tflite - TFv1	INT8	64										
Google Coral TPU	18.04 / 20.10	Coral-Inception V3	imagenet 1000	edge tpu tflite - TFv1	INT8	128										
Google Coral TPU	18.04 / 20.10	Custom-Inception V3	imagenet 1000	edge tpu tflite - TFv2	INT8	1	0.8341835938	3.23638773	1.589375019	616.6254816	0.5810301368	1.609136693	1.721081122			

Object Detection

YOLO V5 - image

Edge device	Setup					Metric									
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	mAP(AP@0.5 / AP@0.5:0.95)	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	1	0.5647638227407072 / 0.3711737812654523	12.26426196	0.8096978664	5093.134744	0.9623479416	0.9803814189	1.039125203	16.5	7.2	2291.666667
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	2	0.5644074727486916 / 0.3707908008232875	10.94295406	0.7587203979	4802.387367	0.922328355	1.039667085	1.084212574			
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	4	0.5644571042149279 / 0.3705403369898913	10.91571498	0.7543849945	4916.665183	0.946897299	1.015533284	1.05608074			
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	8	0.5648049160302835 / 0.3709935765735795	10.95993614	0.7601180077	4520.998549	0.8705368398	1.104246943	1.148716464			
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	16										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	32										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	64										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	128										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	1	0.5647638020889723 / 0.3711737597325565	7.864058971	3.471688271	861.285116	0.1252238181	5.729865295	7.985701246	16.5	7.2	2291.666667
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	2	0.5646190636793411 / 0.3709763340315297	7.856962681	0.7931272984	718.6371565	0.1092782552	6.874862751	9.150951376			
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	4	0.5640720897409268 / 0.3703032132756991	7.772254229	0.8633058071	698.4024701	0.1080801951	7.071755389	9.252388924			
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	8	0.5647166749700744 / 0.3710099587821115	7.783799171	0.858301878	691.0363989	0.1084476837	7.146139263	9.22103604			
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	16	0.5650787225114706 / 0.3713174191735081	7.695362568	0.8623054028	778.7656796	0.1263584022	6.360791936	7.913996872			
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	32										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	64										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	128										
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	1	0.5647638020889723 / 0.3711737597325565	8.846709728	0.4033215046	557.602648	0.06915389953	8.820633972	14.46050052	16.5	7.2	2291.666667
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	2	0.5647644036418817 / 0.3711038610550522	8.939892769	0.5627903938	441.1011102	0.0602490489	11.09622261	16.59777238			
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	4	0.5649045802340809 / 0.3711217613456535	8.844365358	0.5723443031	405.8928285	0.0566470829	12.03921302	17.65315968			
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	8	0.5651255930464496 / 0.3712151517809086	8.855478525	0.5715312958	394.6336148	0.05605573816	12.37438071	17.83938688			
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	16	0.5652908547616373 / 0.3716323973940261	8.853832245	0.5614390373	386.2799754	0.05526309777	12.6562046	18.09525778			
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	32	0.5644733555552214 / 0.3709145698052196 4	8.935377359	0.5315196514	385.3423471	0.05549241152	12.72513264	18.02048195			
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	64	0.5646141816451288 / 0.3710289160170315	8.821940184	0.5368008614	414.3852251	0.06098637588	11.93173332	16.39710485			
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	128	0.5647629647936294 / 0.3711735063808019	8.84664011	0.5543255806	615.7915208	0.09765070882	8.189477817	10.24058107			
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	1	0.5647637724921111 / 0.3711738432100543 7	15.72264218	0.8665115833	1622.182534	0.249951943	3.056246079	4.00076906	16.5	7.2	2291.666667
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	2										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	4										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	8										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	16										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	32										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	64										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	COCO 5000	TensorFlow 2.5V	FP32	128										
Google Coral TPU	18.04 / 20.10	YOLO V5	COCO 5000	edge tpu tflite - TFv2	INT8	1	0.5482012503929592 / 0.3367428828034523	3.41677475	1.173254251	7579.396458	1.406508228	0.6592850014	0.7109805547	5.9	7.2	819.4444444
Google Coral TPU	18.04 / 20.10	YOLO V5	COCO 5000	edge tpu tflite - TFv2	INT8	2										
Google Coral TPU	18.04 / 20.10	YOLO V5	COCO 5000	edge tpu tflite - TFv2	INT8	4										
Google Coral TPU	18.04 / 20.10	YOLO V5	COCO 5000	edge tpu tflite - TFv2	INT8	8										
Google Coral TPU	18.04 / 20.10	YOLO V5	COCO 5000	edge tpu tflite - TFv2	INT8	16										
Google Coral TPU	18.04 / 20.10	YOLO V5	COCO 5000	edge tpu tflite - TFv2	INT8	32										
Google Coral TPU	18.04 / 20.10	YOLO V5	COCO 5000	edge tpu tflite - TFv2	INT8	64										
Google Coral TPU	18.04 / 20.10	YOLO V5	COCO 5000	edge tpu tflite - TFv2	INT8	128										

YOLO V5 - video (20s/38 frame)

Edge device	Setup					Metric									
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	mAP(AP@0.5 / AP@0.5:0.95)	Model load time	Data load time	Inference task time	Inference time	FPS (model, dataset load + inference)	FPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	1		11.95617008	0.03966641426	39.20982003	1.015424678	0.7485079732	0.9848096284	16.5	7.2	2291.666667
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	2										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	4										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	8										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	16										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	32										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	64										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	128										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	1		13.12914824	0.05539369583	17.68576741	0.4512085789	1.230956242	2.216269918	16.5	7.2	2291.666667
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	2										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	4										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	8										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	16										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	32										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	64										
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	128										
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	1		9.008591652	0.03139829636	11.63666821	0.29610317	1.830980698	3.377201264	16.5	7.2	2291.666667
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	2										
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	4										
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	8										
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	16										
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	32										
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	64										
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	128										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	1		15.74729085	0.04490447044	30.66707563	0.7884435779	0.7697948022	1.268321574	16.5	7.2	2291.666667
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	2										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	4										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	8										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	16										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	32										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	64										
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	TensorFlow 2.5V	FP32	128										
Google Coral TPU	18.04 / 20.10	YOLO V5	mp4	edge tpu tflite - TFv2	INT8	1		3.37371397	0.2294085026	44.27248073	1.136622542	0.7939864177	0.8797995493	5.9	7.2	819.4444444
Google Coral TPU	18.04 / 20.10	YOLO V5	mp4	edge tpu tflite - TFv2	INT8	2										
Google Coral TPU	18.04 / 20.10	YOLO V5	mp4	edge tpu tflite - TFv2	INT8	4										
Google Coral TPU	18.04 / 20.10	YOLO V5	mp4	edge tpu tflite - TFv2	INT8	8										
Google Coral TPU	18.04 / 20.10	YOLO V5	mp4	edge tpu tflite - TFv2	INT8	16										

Image Classification

MobileNet V1

Edge device	Setup					Metric									
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Google Coral TPU	18.04 / 20.10	YOLO V5	mp4	edge tpu tflite - TFv2	INT8	32										
Google Coral TPU	18.04 / 20.10	YOLO V5	mp4	edge tpu tflite - TFv2	INT8	64										
Google Coral TPU	18.04 / 20.10	YOLO V5	mp4	edge tpu tflite - TFv2	INT8	128										
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	tflite - TFv2	FP16	1		5.404241323	0.0769109726	84.69850254	2.210895927	0.4213810761	0.4523053246	16.4	7.2	2277.777778
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	tflite - TFv2	FP16	1		5.126217842	0.03652477264	76.50713611	1.998791111	0.465287822	0.500302405	16.4	7.2	2277.777778
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	tflite - TFv2	FP16	1		3.832612276	0.03427410126	56.88416886	1.485840628	0.6255035382	0.6730196908	16.4	7.2	2277.777778
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	tflite - TFv2	FP16	1		7.65151906	0.06461715698	96.5831039	2.523046274	0.3643363073	0.396346278	16.4	7.2	2277.777778
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	tflite - TFv2	INT8	1		5.189818859	0.0424387455	47.69496274	1.239058526	0.7179670452	0.8070643792	16.4	7.2	2277.777778
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	tflite - TFv2	INT8	1		5.158523798	0.03731536865	42.02915525	1.091719678	0.8046586446	0.9159860544	16.4	7.2	2277.777778
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	tflite - TFv2	INT8	1		3.864009142	0.03376603127	57.95437908	1.514623171	0.6143682537	0.6602302268	16.4	7.2	2277.777778
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	YOLO V5	mp4	tflite - TFv2	INT8	1		6.017657042	0.04754137993	56.25687218	1.462034784	0.6097358389	0.6839782549	16.4	7.2	2277.777778

$

Image Classification

MobileNet V1

Edge device	Setup		Metric										
	Model	Dataset	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson Xavier	Mobilenet V1	imagenet 1000	1	0.93	38.88625741	1.037235498	65.82744575	0.06412833238	9.456155237	15.59373155	1.15	4.25	270.5882353
Nvidia Jetson Xavier	Mobilenet V1	imagenet 1000	2	0.93	38.01113439	1.054350138	32.73214889	0.03176783514	13.92799128	31.47838043	2.3	4.25	541.1764706
Nvidia Jetson Xavier	Mobilenet V1	imagenet 1000	4	0.93	39.47181606	1.041965723	18.81584477	0.01820332026	16.85487822	54.93503303	4.59	4.25	1080
Nvidia Jetson Xavier	Mobilenet V1	imagenet 1000	8	0.93	38.30404139	1.030306578	12.16145039	0.01171571183	19.41897256	85.35546233	9.18	4.25	2160
Nvidia Jetson Xavier	Mobilenet V1	imagenet 1000	16	0.93	38.36752224	1.058549881	8.744111538	0.008018319332	20.92574659	124.7144144	18.4	4.25	4329.411765
Nvidia Jetson Xavier	Mobilenet V1	imagenet 1000	32	0.93	38.1481607	1.017022848	8.580517292	0.00614240556	21.44679822	162.8026659	36.7	4.25	8635.294118
Nvidia Jetson Xavier	Mobilenet V1	imagenet 1000	64	0.93	38.18341804	1.039550543	8.765650034	0.005155852996	21.33827334	193.9543274	73.4	4.25	17270.58824
Nvidia Jetson Xavier	Mobilenet V1	imagenet 1000	128	0.93	37.66842175	1.027097225	10.46026754	0.004670296796	20.83164779	214.1191543	1.47E+02	4.25	3.46E+04
Nvidia Jetson Nano	Mobilenet V1	imagenet 1000	1	0.93	71.08252144	2.345131636	156.859957	0.1526636727	4.342385849	6.550346801	1.15	4.25	270.5882353
Nvidia Jetson Nano	Mobilenet V1	imagenet 1000	2	0.93	71.12680101	2.0991745	112.5438108	0.1095119245	5.38299902	9.131425683	2.3	4.25	541.1764706
Nvidia Jetson Nano	Mobilenet V1	imagenet 1000	4	0.93	70.64579511	2.058488846	85.60306263	0.07203072262	6.31681362	13.88296499	4.59	4.25	1080
Nvidia Jetson Nano	Mobilenet V1	imagenet 1000	8	0.93	69.93868399	2.046374798	62.63308215	0.05643707871	7.42839512	17.71884766	9.18	4.25	2160
Nvidia Jetson Nano	Mobilenet V1	imagenet 1000	16	0.93	70.81421661	2.031164169	68.55348921	0.05673198426	7.128752148	17.62674113	18.4	4.25	4329.411765
Nvidia Jetson Nano	Mobilenet V1	imagenet 1000	32	0.93	70.34334397	2.068148136	107.1531734	0.08601080929	5.702673581	11.62644566	36.7	4.25	8635.294118
Nvidia Jetson Nano	Mobilenet V1	imagenet 1000	64								73.4	4.25	17270.58824
Nvidia Jetson Nano	Mobilenet V1	imagenet 1000	128								1.47E+02	4.25	3.46E+04

MobileNet V2

Edge device	Setup		Metric										
	Model	Dataset	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson Xavier	Mobilenet V2	imagenet 1000	1	0.906	66.94901466	1.029476643	92.66682172	0.09103277969	6.224886781	10.98505399	0.615	3.53	174.2209632
Nvidia Jetson Xavier	Mobilenet V2	imagenet 1000	2	0.906	66.94992638	1.034856796	48.19621539	0.04725036407	8.607235995	21.16385809	1.23	3.53	348.4419263
Nvidia Jetson Xavier	Mobilenet V2	imagenet 1000	4	0.906	67.81008959	1.019435406	26.64441657	0.02604661059	10.47391131	38.39271127	2.46	3.53	696.8838527
Nvidia Jetson Xavier	Mobilenet V2	imagenet 1000	8	0.906	66.61496758	1.004122734	15.34189391	0.01490839934	12.05383741	67.0762821	4.92	3.53	1393.767705
Nvidia Jetson Xavier	Mobilenet V2	imagenet 1000	16	0.906	66.90273213	1.034251213	10.25050116	0.009799251008	12.89206094	102.0486157	9.84	3.53	2787.535411
Nvidia Jetson Xavier	Mobilenet V2	imagenet 1000	32	0.906	66.81817079	1.010535479	8.577705383	0.007023798302	13.40196461	142.3731088	19.7	3.53	5580.736544
Nvidia Jetson Xavier	Mobilenet V2	imagenet 1000	64	0.906	67.01150298	1.009224176	8.856770515	0.00599429966	13.31984638	166.82516	39.4	3.53	11161.47309
Nvidia Jetson Xavier	Mobilenet V2	imagenet 1000	128	0.906	67.0898447	1.022275925	10.45342684	0.005248077912	13.03367568	190.5459516	78.7	3.53	22294.61756
Nvidia Jetson Nano	Mobilenet V2	imagenet 1000	1	0.906	127.2817204	2.542519569	253.3297305	0.2492911539	2.609913198	4.011373787	0.615	3.53	174.2209632
Nvidia Jetson Nano	Mobilenet V2	imagenet 1000	2	0.906	125.8207233	2.181414604	173.7648814	0.1711566045	3.31381096	5.842602468	1.23	3.53	348.4419263
Nvidia Jetson Nano	Mobilenet V2	imagenet 1000	4	0.906	125.2173266	2.570510387	158.1518993	0.1498946307	3.497234534	6.67135304	2.46	3.53	696.8838527
Nvidia Jetson Nano	Mobilenet V2	imagenet 1000	8	0.906	125.2875311	2.136475801	147.7138095	0.08967195725	3.634533248	11.15175837	4.92	3.53	1393.767705
Nvidia Jetson Nano	Mobilenet V2	imagenet 1000	16	0.906	125.7341638	2.186259031	88.78350043	0.07893936454	4.651493628	12.66795097	9.84	3.53	2787.535411
Nvidia Jetson Nano	Mobilenet V2	imagenet 1000	32								19.7	3.53	5580.736544
Nvidia Jetson Nano	Mobilenet V2	imagenet 1000	64								39.4	3.53	11161.47309
Nvidia Jetson Nano	Mobilenet V2	imagenet 1000	128								78.7	3.53	22294.61756

Inception V3

Edge device	Setup		Metric										
	Model	Dataset	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson Xavier	Inception V3	imagenet 1000	1	0.919	132.2537916	1.042407036	147.0621588	0.1454411669	3.566860513	6.875632405	11.5	23.8	483.1932773
Nvidia Jetson Xavier	Inception V3	imagenet 1000	2	0.919	132.1550634	1.044339657	76.47736812	0.07555388546	4.769238864	13.23558668	22.9	23.8	962.1848739
Nvidia Jetson Xavier	Inception V3	imagenet 1000	4	0.919	131.8163593	1.031774044	82.13434196	0.04199156785	4.651537333	23.8143049	45.9	23.8	1928.571429
Nvidia Jetson Xavier	Inception V3	imagenet 1000	8	0.919	132.4758158	1.018234968	35.21715236	0.02569772863	5.927262872	38.91394505	91.8	23.8	3857.142857
Nvidia Jetson Xavier	Inception V3	imagenet 1000	16	0.919	132.8492031	1.030791521	35.47679901	0.02195885707	5.95192651	45.53971079	1.84E+02	23.8	7.73E+03
Nvidia Jetson Xavier	Inception V3	imagenet 1000	32	0.919	132.9358473	1.021033049	37.76926231	0.02061955165	5.96297753	48.49765974	3.67E+02	23.8	1.54E+04
Nvidia Jetson Xavier	Inception V3	imagenet 1000	64	0.919	133.0718038	1.035824299	49.36590672	0.03192617209	5.581181505	31.32226429	7.34E+02	23.8	3.08E+04
Nvidia Jetson Xavier	Inception V3	imagenet 1000	128	0.919	126.1647036	1.033347368	76.42906809	0.05759925744	5.028794448	17.36133493	1.47E+03	23.8	6.18E+04
Nvidia Jetson Nano	Inception V3	imagenet 1000	1								11.5	23.8	483.1932773
Nvidia Jetson Nano	Inception V3	imagenet 1000	2								22.9	23.8	962.1848739
Nvidia Jetson Nano	Inception V3	imagenet 1000	4								45.9	23.8	1928.571429
Nvidia Jetson Nano	Inception V3	imagenet 1000	8								91.8	23.8	3857.142857
Nvidia Jetson Nano	Inception V3	imagenet 1000	16								1.84E+02	23.8	7.73E+03
Nvidia Jetson Nano	Inception V3	imagenet 1000	32								3.67E+02	23.8	1.54E+04
Nvidia Jetson Nano	Inception V3	imagenet 1000	64								7.34E+02	23.8	3.08E+04
Nvidia Jetson Nano	Inception V3	imagenet 1000	128								1.47E+03	23.8	6.18E+04

$

Edge device	Model	Optimization Platform	Quantization	Dataset	Data file size	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)
Nvidia Jetson Xavier	MobileNet V1	TensorFlow 2.5V	FP32	tfrecord 1000	105M	1	0.93	34.7032392	0.977602005	54.16603136	0.0526967833	11.13002465	18.97649035
				raw image1000	**107M**	**1**	**0.8154304199**	**3.17361927**	**6.146800041**	**126.9653633**	**0.1261841333**	**7.337522529**	**7.924926644**
	MobileNet V2			tfrecord 1000	105M	1	0.906	61.97093296	0.9851565361	84.89456916	0.08355212736	6.763571484	11.96857616
				raw image1000	**107M**	**1**	**0.6952713623**	**6.596752882**	**6.228187799**	**127.2615538**	**0.1268222253**	**7.138446886**	**7.88505325**
	Inception V3			tfrecord 1000	105M	1	0.919	122.310173	0.9865400791	142.1121132	0.134392396	3.767768645	7.440897178
				raw image1000	**107M**	**1**	**0.8360081787**	**23.30724597**	**5.745856285**	**140.1731172**	**0.1397072463**	**5.909249781**	**7.157824855**

Edge device	Model	Optimization Platform	Quantization	Dataset	Data file size	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)
Nvidia Jetson Xavier	MobileNet V1	TensorFlow 2.5V	FP32	raw image 1000	107M	1	0.8154304199	3.17361927	6.146800041	126.9653633	0.1261841333	7.337522529	7.924926644
Nvidia Jetson TX2							0.8154304199	4.838519812	10.63910866	174.9827492	0.1745162234	5.250435876	5.730126291
Rpi + Coral TPU		Edgetpu tflite - TFv2	INT8				0.7466875	2.74034667	0.4634003639	35.30190253	0.005293801611	25.97021505	188.9001654
Nvidia Jetson Xavier	MobileNet V2	TensorFlow 2.5V	FP32				0.6952713623	6.596752882	6.228187799	127.2615538	0.1268222253	7.138446886	7.88505325
Nvidia Jetson TX2							0.6952713623	6.774560928	9.507618427	178.5661824	0.1781304531	5.132196089	5.613863225
Rpi + Coral TPU		Edge tpu tflite - TFv2	INT8				0.7055976563	2.697169304	0.5129849911	36.1754241	0.005687035799	25.39000418	175.8385274
Nvidia Jetson Xavier	Inception V3	TensorFlow 2.5V	FP32				0.8360081787	23.30724597	5.745856285	140.1731172	0.1397072463	5.909249781	7.157824855
Nvidia Jetson TX2							0.8360081787	10.70541787	9.359740734	242.2576783	0.2415484159	3.812096621	4.139956771
Rpi + Coral TPU		Edge tpu tflite - TFv2	INT8				0.8341835938	2.763403893	0.4765644073	102.1026423	0.0697758156	9.492834802	14.3316132

Edge device	Model	Optimization Platform	Quantization	Dataset	Data file size	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)
Nvidia Jetson Xavier	MobileNet V1	TensorFlow 2.5V	FP32	raw image 1000	107M	1	0.8154304199	3.099321604	6.734873295	124.3262188	0.1238430877	7.453763537	8.074734075
						2	0.8212866821	3.112899542	6.372859955	64.14415121	0.06390857816	13.58143709	15.64735171
						4	0.8119281006	3.146010876	5.918588161	34.61884713	0.03449584007	22.89196681	28.98900267
						8	0.7822409668	3.150560379	5.697333813	20.99315047	0.02093475342	33.51089117	47.76746017
						16	0.7599161542	3.130496979	5.50075531	12.26452923	0.01223503852	47.85654938	81.73247665
						32	0.7634943724	3.129932642	5.502838135	9.182964563	0.009167948961	56.13015578	109.0756509
						64	0.7335840464	3.16045928	5.483443022	7.301475286	0.007291737795	62.71409971	137.1415194
						128	0.7278847694	3.080677748	5.424599171	6.390872002	0.006385187864	67.13144487	156.6124633
	MobileNet V2	TensorFlow 2.5V	FP32	raw image 1000	107M	1	0.6952713623	4.216674566	6.683954	127.1585252	0.1266751056	7.243271978	7.894210907
						2	0.6922662354	4.171766043	6.275687456	67.42864656	0.0671857481	12.8409101	14.88410903
						4	0.7005634766	4.249879122	5.887251377	36.27891159	0.03615801263	21.54427553	27.65638727
						8	0.6724165649	4.20552969	5.581227541	21.01236987	0.02095461965	32.46845265	47.72217377
						16	0.6743112594	4.295340776	5.516176462	13.38004827	0.01335142159	43.11912447	74.89839137
						32	0.656946063	4.197406054	5.540759563	10.32744312	0.01030998278	49.83651445	96.99337251
						64	0.6237511635	4.209680796	5.468185425	8.486898899	0.008474980354	55.05163394	117.9943738
						128	0.5953434706	4.233690262	5.471747637	7.474613428	0.007468005419	58.2070438	133.9045627
	Inception V3	TensorFlow 2.5V	FP32	raw image 1000	107M	1	0.8360081787	6.934961081	6.631296873	138.0621166	0.1376137908	6.595071686	7.266713565
						2	0.8371211548	7.027062654	6.197374582	73.80433631	0.07357701612	11.49045263	13.5912008
						4	0.8426464844	6.988181114	5.9403162	41.27468991	0.04115663791	18.4490996	24.29741716
						8	0.8290883179	7.03455162	5.73078537	25.18580437	0.02512357044	26.34966866	39.80325974
						16	0.832520379	7.061951637	5.630872726	17.21933675	0.01718498349	33.4312187	58.19033812
						32	0.8306828737	7.025388718	5.583262444	13.13729382	0.01311532331	38.84106802	76.24669074
						64	0.8323142529	6.702260494	5.546610355	10.34937906	0.01033916235	44.25121431	96.71963416
						128	0.8211693764	7.003484726	5.51829505	9.343114614	0.009336360216	45.73541414	107.108121

	Edge Device	Model	Dataset	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	model size
old	Nvidia Jetson TX2	Mobilenet V1	imagenet 1000	1	0.93	59.92601061	1.6062181	99.50904369	0.09781955647	6.209582935	10.22290466	1.3MB
new	Nvidia Jetson TX2	Mobilenet V1	imagenet 1000	1	0.93	2.744750023	1.727652311	98.29226208	0.09653062463	9.730919374	10.3594067	h5 : 17MB / json : 44KB
old	Nvidia Jetson TX2	Mobilenet V2	imagenet 1000	1	0.906	103.5648978	1.562505245	157.4630725	0.1557951264	3.808201653	6.41868602	2.5MB
new	Nvidia Jetson TX2	Mobilenet V2	imagenet 1000	1	0.906	4.72168088	2.054520845	159.603534	0.1578788443	6.010341644	6.333970867	h5 : 14MB / json : 80KB
old	Nvidia Jetson TX2	Inception V3	imagenet 1000	1	0.919	190.2802024	1.556982994	254.3616292	0.2526663547	2.241152661	3.957788529	4.4M
new	Nvidia Jetson TX2	Inception V3	imagenet 1000	1	0.919	9.547210455	1.754220247	250.801995	0.2490842218	3.815285038	4.014706321	h5 : 92MB / json : 148KB

edge device	model	saved file	model size	model load time
Nvidia Jetson TX2	Mobilenet V1	pb (weight, network)	1.3MB	59.92601061
	Mobilenet V1	h5(weight) / json (network)	17MB / 44KB	2.744750023
	Mobilenet V2	pb (weight, network)	2.5MB	103.5648978
	Mobilenet V2	h5(weight) / json (network)	14MB / 80KB	4.72168088
	Inception V3	pb (weight, network)	4.4M	190.2802024
	Inception V3	h5(weight) / json (network)	92MB / 148KB	9.547210455

model	Optimization Platform	Input size	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	model size
Coral-Mobilenet V1	edge tpu tflite - TFv2	224x224x3	1	0.7466875	3.13670969	0.6347105503	49.95246148	0.01541767024	18.61369596	64.86064267	4.6M
Coral-Mobilenet V2	edge tpu tflite - TFv2	224x224x3	1	0.7055976563	3.117435455	1.01078105	50.14280176	0.01652078829	18.42604086	60.52979934	4.2M
Coral-Inception V3	edge tpu tflite - TFv1	224x224x3	1	0.783796875	3.19194746	0.4738194942	190.9825015	0.1694372175	5.137471851	5.901891064	12M
Coral-Inception V3	edge tpu tflite - TFv1	299x299x3	1	0.9133398438	3.178961277	0.442332983	577.8533888	0.5539650221	1.71976533	1.805168124	25M
custom-Mobilenet v1	edge tpu tflite - TFv2	224x224x3	1								
custom-Mobilenet V2	edge tpu tflite - TFv2	224x224x3	1	0.6149140625	3.109596014	0.4486391544	40.48142219	0.01313121428	22.706807	76.15441943	4.1M
custom-Inception V3	edge tpu tflite - TFv2	299x299x3	1	0.8341835938	3.23638773	1.589375019	616.6254816	0.5810301368	1.609136693	1.721081122	24M

TRT vs TF-Lite

Edge Device	model	Optimization Platform	Quantization	Dataset	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)	model size
Nvidia Jetson TX2	MobileNet V1	TensorRT	FP32	imagenet 1000	1	0.93	175.3601644	0.7383546829	51.67811322	0.04575625014	4.390261077	21.85493778	19MB
Nvidia Jetson TX2	MobileNet V2	TensorRT	FP32	imagenet 1000	1	0.906	190.3930771	0.7177202702	51.81776929	0.04641133618	4.116433942	21.54646003	17MB
Nvidia Jetson TX2	Inception V3	TensorRT	FP32	imagenet 1000	1	0.919	875.0414793	0.8669939041	142.8614163	0.05401475263	0.9815757357	18.5134607	97MB
Nvidia Jetson TX2	MobileNet V1	TensorRT	FP16	imagenet 1000	1	0.93	177.7084451	0.7184529305	62.83537316	0.05922599053	4.14486264	16.88447911	19MB
Nvidia Jetson TX2	MobileNet V2	TensorRT	FP16	imagenet 1000	1	0.905	191.834491	0.7160468102	63.0419898	0.05860773921	3.91247506	17.06259299	17MB
Nvidia Jetson TX2	Inception V3	TensorRT	FP16	imagenet 1000	1	0.917	913.6164947	0.8568546772	91.71054363	0.07264208722	0.9938536383	13.76612427	97MB
Nvidia Jetson TX2	MobileNet V1	TensorRT	INT8	imagenet 1000	1	0.93	174.7160046	0.7275948524	9.512676716	0.005189748764	5.406674982	192.6875549	19MB
Nvidia Jetson TX2	MobileNet V2	TensorRT	INT8	imagenet 1000	1	0.908	190.4329388	0.6990947723	9.935600758	0.005090838671	4.973444673	196.4312886	17MB
Nvidia Jetson TX2	Inception V3	TensorRT	INT8	imagenet 1000	1	memory limit							
Nvidia Jetson TX2	MobileNet V1	TensorLite	FP32	imagenet 1000	1	0.4974281616	0.01261377335	0.3975148201	103.4773719	0.08632320921	9.625797088	11.58437006	16.1MB
Nvidia Jetson TX2	MobileNet V2	TensorLite	FP32	imagenet 1000	1	0.5728826904	0.01034545898	0.3892166615	89.72137737	0.0724643943	11.09620035	13.79988075	13.3MB
Nvidia Jetson TX2	Inception V3	TensorLite	FP32	imagenet 1000	1	Segmentation fault (core dumped)							90.9MB
Nvidia Jetson TX2	MobileNet V1	TensorLite	FP16	imagenet 1000	1	0.7657109375	0.004939079285	0.4116418362	99.55065703	0.0817362262	10.00327728	12.23447725	8.1MB
Nvidia Jetson TX2	MobileNet V2	TensorLite	FP16	imagenet 1000	1	0.573520752	0.006615638733	0.4141659737	88.37753868	0.0712310019	11.26147428	14.03883103	6.7MB
Nvidia Jetson TX2	Inception V3	TensorLite	FP16	imagenet 1000	1	0.9696956177	0.03152155876	0.9304687977	596.7319081	0.5765084247	1.67309722	1.734580029	45.5MB
Nvidia Jetson TX2	MobileNet V1	TensorLite	INT8	imagenet 1000	1	0.4970228271	0.007335424423	0.4081280231	103.8265989	0.08669070854	9.593056561	11.53526159	4.4MB
Nvidia Jetson TX2	MobileNet V2	TensorLite	INT8	imagenet 1000	1	0.6178867188	0.005393505096	0.4070675373	90.02196455	0.07223293441	11.05773596	13.84410045	3.8MB
Nvidia Jetson TX2	Inception V3	TensorLite	INT8	imagenet 1000	1	0.8316484375	0.01927137375	0.4072015285	641.392735	0.6221491756	1.558071163	1.607331552	23.2MB

TF vs TRT

Edge Device	model	Optimization Platform	Quantization	Dataset	batch size	Accuracy	Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)	IPS (inference only)
Nvidia Jetson TX2	Mobilenet V1	TensorFlow 2.5V	FP32	imagenet 1000	1	0.93	59.92601061	1.6062181	99.50904369	0.09781955647	6.209582935	10.22290466
Nvidia Jetson TX2	Mobilenet V2	TensorFlow 2.5V	FP32	imagenet 1000	1	0.906	103.5648978	1.562505245	157.4630725	0.1557951264	3.808201653	6.41868602
Nvidia Jetson TX2	Inception V3	TensorFlow 2.5V	FP32	imagenet 1000	1	0.919	190.2802024	1.556982994	254.3616292	0.2526663547	2.241152661	3.957788529
Nvidia Jetson TX2	MobileNet V1	TensorRT	FP32	imagenet 1000	1	0.93	175.3601644	0.7383546829	51.67811322	0.04575625014	4.390261077	21.85493778
Nvidia Jetson TX2	MobileNet V2	TensorRT	FP32	imagenet 1000	1	0.906	190.3930771	0.7177202702	51.81776929	0.04641133618	4.116433942	21.54646003
Nvidia Jetson TX2	Inception V3	TensorRT	FP32	imagenet 1000	1	0.919	875.0414793	0.8669939041	142.8614163	0.05401475263	0.9815757357	18.5134607

Edge Device	Model	Quantization	Accuracy	IPS (inference only)	Model size
Nvidia Jetson TX2	MobileNet V1	FP32	0.93	21.85493778	19MB
	MobileNet V1	FP16		16.88447911	
	MobileNet V1	INT8		192.6875549	
	MobileNet V2	FP32	0.906	21.54646003	17MB
	MobileNet V2	FP16	0.905	17.06259299	
	MobileNet V2	INT8	0.908	196.4312886	
	Inception V3	FP32	0.919	18.5134607	97MB
	Inception V3	FP16	0.917	13.76612427	
	Inception V3	INT8	memory limit		

Model load time	Data load time	Inference task time	Inference time	IPS (model, dataset load + inference)
59.92601061	1.6062181	99.50904369	0.0978195564 7	6.209582935
103.5648978	1.562505245	157.4630725	0.1557951264	3.808201653
190.2802024	1.556982994	254.3616292	0.2526663547	2.241152661
175.3601644	0.7383546829	51.67811322	0.0457562501 4	4.390261077
190.3930771	0.7177202702	51.81776929	0.0464113361 8	4.116433942
875.0414793	0.8669939041	142.8614163	0.0540147526 3	0.9815757357

Edge Device	Model	Quantization	Optimization Platform	Accuracy	IPS (inference only)
Nvidia Jetson TX2	Mobilenet V1	FP32	TensorFlow 2.5V	0.93	10.22290466
	MobileNet V1		TensorRT		21.85493778
	Mobilenet V2		TensorFlow 2.5V	0.906	6.41868602
	MobileNet V2		TensorRT		21.54646003
	Inception V3		TensorFlow 2.5V	0.919	3.957788529
	Inception V3		TensorRT		18.5134607

single GPU 할당 없이 추론 성능 비교

	IPS (model, dataset load + inference)	IPS (inference only)
container1	11.47055448	24.48302847
container2	21.59180339	24.15756832
container3	21.68663102	24.42456412
container4	21.62289646	24.44595701
container5	22.05667186	25.30064568

single GPU 할당 여부에 따른 추론 성능 비교

	IPS (model, dataset load + inference)	IPS (inference only)
with-gpu container	27.71799156	31.57648431
without-gpu container	27.52136015	31.26643088

ggaman.com/vram: 25	IPS (model, dataset load + inference)	IPS (inference only)		
container1	4.753275123	7.603665533	avg=	7.595620248
container2	4.725597928	7.447515023		
container3	4.612298272	7.662968067		
container4	4.561359302	7.668332368		

ggaman.com/vram: 50	IPS (model, dataset load + inference)	IPS (inference only)		
container1	4.778885324	7.655733805	avg=	7.659350936
container2	4.62523149	7.662968067		

Text Classification

RNN

Edge device	Setup					Metric										
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	Accuracy	Model load time	Data load time	Inference time	Inference time (avg)	SPS	SPS(inf)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	1	0.8556	9.4641	19.3196	5585.5211	0.2234	4.4529	4.4759		0.48	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	2	0.8556	9.4641	19.2634	2779.9999	0.1112	8.9008	8.9928		0.48	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	4	0.8556	9.4641	17.0427	1395.1806	0.0558	17.5847	17.9188		0.48	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	8	0.8556	9.4641	17.6896	699.8513	0.0280	34.3877	35.7219		0.48	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	16	0.8556	9.4641	18.3255	349.4656	0.0140	66.2681	71.5378		0.48	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	32	0.8556	9.4641	18.5090	177.2573	0.0071	121.8143	141.0380		0.48	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	64	0.8556	9.4641	18.8104	141.9317	0.0057	146.8806	176.1410		0.48	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	128	0.8556	9.4641	17.3180	47.2774	0.0019	337.5661	528.7935		0.48	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	1	0.8556	9.3766	21.0153	4954.9795	0.1982	5.0147	5.0454		0.48	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	2	0.8556	9.3766	15.4939	2454.1057	0.0982	10.0848	10.1870		0.48	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	4	0.8556	9.3766	15.0859	1218.4059	0.0487	20.1148	20.5186		0.48	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	8	0.8556	9.3766	16.5853	612.4903	0.0245	39.1572	40.8170		0.48	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	16	0.8556	9.3766	15.3080	310.6903	0.0124	74.5435	80.4660		0.48	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	32	0.8556	9.3766	16.9366	159.1941	0.0064	134.7655	157.0410		0.48	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	64	0.8556	9.3766	15.5508	81.2686	0.0033	235.4138	307.6217		0.48	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	128	0.8556	9.3766	15.2519	42.8137	0.0017	370.6878	583.9248		0.48	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	1	0.8556	6.1586	8.6549	3550.9036	0.1420	7.0112	7.0405		0.48	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	2	0.8556	6.1586	8.4416	1736.2571	0.0695	14.2787	14.3988		0.48	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	4	0.8556	6.1586	8.1436	870.4108	0.0348	28.2578	28.7221		0.48	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	8	0.8556	6.1586	7.9782	436.9325	0.0175	55.4239	57.2171		0.48	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	16	0.8556	6.1586	7.9437	222.2590	0.0089	105.7703	112.4814		0.48	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	32	0.8556	6.1586	8.2950	113.9972	0.0046	194.6271	219.3036		0.48	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	64	0.8556	6.1586	7.9682	57.2920	0.0023	350.0483	436.3613		0.48	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	128	0.8556	6.1586	7.9730	28.8645	0.0012	581.4478	866.1143		0.48	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	1	0.8556	11.6906	19.0473	8554.7956	0.3422	2.9085	2.9223		0.48	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	2	0.8556	11.6906	21.8069	4230.1342	0.1692	5.8635	5.9100		0.48	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	4	0.8556	11.6906	19.6436	2124.2388	0.0850	11.5978	11.7689		0.48	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	8	0.8556	11.6906	19.8350	1064.0485	0.0426	22.8191	23.4952		0.48	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	16	0.8556	11.6906	20.4888	534.6106	0.0214	44.1080	46.7630		0.48	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	32	0.8556	11.6906	19.9178	266.9164	0.0107	83.7451	93.6623		0.48	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	64	0.8556	11.6906	19.6598	133.8757	0.0054	151.3077	186.7403		0.48	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	RNN	IMDb	TensorFlow 2.5V	FP32	128	0.8556	11.6906	19.3404	67.0982	0.0027	254.7662	372.5885		0.48	
Google Coral TPU	18.04 / 32.6.1 / 20.10	RNN	IMDb	TFLite	INT8	1									0.48	
Google Coral TPU	18.04 / 32.6.1 / 20.10	RNN	IMDb	TFLite	INT8	2									0.48	
Google Coral TPU	18.04 / 32.6.1 / 20.10	RNN	IMDb	TFLite	INT8	4									0.48	
Google Coral TPU	18.04 / 32.6.1 / 20.10	RNN	IMDb	TFLite	INT8	8									0.48	
Google Coral TPU	18.04 / 32.6.1 / 20.10	RNN	IMDb	TFLite	INT8	16									0.48	
Google Coral TPU	18.04 / 32.6.1 / 20.10	RNN	IMDb	TFLite	INT8	32									0.48	
Google Coral TPU	18.04 / 32.6.1 / 20.10	RNN	IMDb	TFLite	INT8	64									0.48	
Google Coral TPU	18.04 / 32.6.1 / 20.10	RNN	IMDb	TFLite	INT8	128									0.48	

LSTM

Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	1	0.8069	22.7373	18.3478	17345.6504	0.6938	1.4379	1.4413		2	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	2	0.7969	22.7373	19.0258	9213.9649	0.3686	2.7010	2.7133		2	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	4	0.7969	22.7373	17.7701	4613.5905	0.1845	5.3716	5.4188		2	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	8	0.7969	22.7373	18.3041	2383.7325	0.0953	10.3102	10.4878		2	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	16	0.7969	22.7373	17.5191	1260.9435	0.0504	19.2130	19.8264		2	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	32	0.7969	22.7373	17.8539	690.3632	0.0276	34.2019	36.2128		2	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	64	0.7969	22.7373	17.5733	410.3781	0.0164	55.4707	60.9194		2	
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	128	0.7969	22.7373	17.6135	237.8475	0.0095	89.8640	105.1094		2	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	1	0.8069	20.9757	16.3809	16036.0322	0.6414	1.5554	1.5590		2	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	2	0.8069	20.9757	16.5841	8510.3646	0.3404	2.9247	2.9376		2	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	4	0.8069	20.9757	14.1016	4257.2302	0.1703	5.8244	5.8724		2	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	8	0.8069	20.9757	16.9541	2212.3742	0.0885	11.1096	11.3001		2	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	16	0.8069	20.9757	15.3723	1221.9277	0.0489	19.8685	20.4595		2	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	32	0.8069	20.9757	16.0406	681.9325	0.0273	34.7730	36.6605		2	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	64	0.8069	20.9757	17.0885	354.7541	0.0142	63.6427	70.4714		2	
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	128	0.8069	20.9757	15.5014	206.3141	0.0083	102.9692	121.1745		2	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	1	0.8069	17.2574	7.8386	11148.1480	0.4459	2.2375	2.2425		2	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	2	0.7969	17.2574	8.8583	5136.9338	0.2055	4.8421	4.8667		2	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	4	0.7969	17.2574	8.1026	2317.8181	0.0927	10.6693	10.7860		2	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	8	0.7969	17.2574	8.1910	1163.4378	0.0465	21.0281	21.4880		2	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	16	0.7969	17.2574	8.0927	582.6459	0.0233	41.1187	42.9077		2	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	32	0.7969	17.2574	8.0253	296.5081	0.0119	77.6902	84.3147		2	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	64	0.7969	17.2574	8.0237	148.9398	0.0060	143.4960	167.8531		2	
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	128	0.7969	17.2574	7.9237	74.6480	0.0030	250.4280	334.9053		2	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	1	0.8069	26.3049	20.9180	25389.5774	1.0156	0.9828	0.9847		2	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	2	0.7969	26.3049	21.8082	12727.1783	0.5091	1.9569	1.9643		2	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	4	0.7969	26.3049	22.0341	6393.6936	0.2557	3.8808	3.9101		2	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	8	0.7969	26.3049	20.1906	3195.8318	0.1278	7.7105	7.8227		2	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	16	0.7969	26.3049	19.6134	1591.0087	0.0636	15.2725	15.7133		2	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	32	0.7969	26.3049	21.0925	800.8616	0.0320	29.4721	31.2164		2	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	64	0.7969	26.3049	20.9247	401.3867	0.0161	55.7269	62.2841		2	
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TensorFlow 2.5V	FP32	128	0.7969	26.3049	20.1097	203.1458	0.0081	100.1761	123.0643		2	
Google Coral TPU	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TFLite	INT8	1									2	
Google Coral TPU	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TFLite	INT8	2									2	
Google Coral TPU	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TFLite	INT8	4									2	
Google Coral TPU	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TFLite	INT8	8									2	
Google Coral TPU	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TFLite	INT8	16									2	
Google Coral TPU	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TFLite	INT8	32									2	
Google Coral TPU	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TFLite	INT8	64									2	
Google Coral TPU	18.04 / 32.6.1 / 20.10	LSTM	IMDb	TFLite	INT8	128									2	

DistilBERT

Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	1	0.9106	20.9655	26.7664	328.8060	0.3771	2.3158	2.6520	6.7	66	101.5151515
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	2	0.9106	20.9655	1.9340	328.8665	0.3771	2.4789	2.6515	6.7	66	101.5151515
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	4	0.9106	20.9655	2.0933	328.5153	0.3767	2.4803	2.6544	6.7	66	101.5151515
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	8	0.9106	20.9655	2.3332	318.9475	0.3658	2.5479	2.7340	6.7	66	101.5151515
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	16	0.9106	20.9655	2.6716	317.6831	0.3643	2.5548	2.7449	6.7	66	101.5151515
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	32	0.9106	20.9655	1.7473	317.7045	0.3643	2.5616	2.7447	6.7	66	101.5151515
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	64	0.9106	20.9655	2.2150	318.3051	0.3650	2.5535	2.7395	6.7	66	101.5151515
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	128	0.9106	20.9655	2.3273	317.5425	0.3642	2.5584	2.7461	6.7	66	101.5151515
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	1	0.9106	81.7007	19.8494	319.4085	0.3663	2.0715	2.7300	6.7	66	101.5151515
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	2	0.9106	81.7007	2.1549	319.3944	0.3663	2.1624	2.7302	6.7	66	101.5151515
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	4	0.9106	81.7007	2.3592	319.0450	0.3659	2.1632	2.7332	6.7	66	101.5151515
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	8	0.9106	81.7007	2.1851	318.8628	0.3657	2.1651	2.7347	6.7	66	101.5151515
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	16	0.9106	81.7007	2.1580	319.4792	0.3664	2.1620	2.7294	6.7	66	101.5151515
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	32	0.9106	81.7007	1.8575	319.5305	0.3664	2.1633	2.7290	6.7	66	101.5151515
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	64	0.9106	81.7007	1.8905	319.0600	0.3659	2.1656	2.7330	6.7	66	101.5151515
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	128	0.9106	81.7007	2.4904	319.4293	0.3663	2.1604	2.7299	6.7	66	101.5151515
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	1	0.9106	23.4429	13.1857	89.9651	0.1032	6.8882	9.6926	6.7	66	101.5151515
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	2	0.9106	23.4429	2.1520	88.6550	0.1017	7.6324	9.8359	6.7	66	101.5151515
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	4	0.9106	23.4429	22.2313	88.4177	0.1014	6.5030	9.8623	6.7	66	101.5151515
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	8	0.9106	23.4429	3.5286	88.1726	0.1011	7.5731	9.8897	6.7	66	101.5151515
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	16	0.9106	23.4429	1.7758	88.0026	0.1009	7.7017	9.9088	6.7	66	101.5151515
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	32	0.9106	23.4429	1.9112	88.3726	0.1013	7.6675	9.8673	6.7	66	101.5151515
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	64	0.9106	23.4429	2.2954	88.7766	0.1018	7.6147	9.8224	6.7	66	101.5151515
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	128	0.9106	23.4429	3.6887	89.5850	0.1027	7.4711	9.7338	6.7	66	101.5151515
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	1	0.9106	80.0597	6.5136	210.6576	0.2416	2.9337	4.1394	6.7	66	101.5151515
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	2	0.9106	80.0597	2.0669	221.6414	0.2542	2.8706	3.9343	6.7	66	101.5151515
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	4	0.9106	80.0597	4.2786	201.6622	0.2313	3.0489	4.3241	6.7	66	101.5151515
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	8	0.9106	80.0597	2.4364	199.6945	0.2290	3.0901	4.3667	6.7	66	101.5151515
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	16	0.9106	80.0597	2.1012	200.0113	0.2294	3.0903	4.3598	6.7	66	101.5151515
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	32	0.9106	80.0597	3.0687	200.8313	0.2303	3.0709	4.3420	6.7	66	101.5151515

Text Classification

RNN

Edge device	Setup					Metric										
	OS (Ubuntu/L4T/ Docker)	Model	Dataset	Optimization Platform	Quantization	batch size	Accuracy	Model load time	Data load time	Inference time	Inference time (avg)	SPS	SPS(inf)	Flop (giga)	parameter (m)	Flop / parameter (m)
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	64	0.9106	80.0597	2.5090	200.0222	0.2294	3.0857	4.3595	6.7	66	101.5151515
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TensorFlow 2.5V	FP32	128	0.9106	80.0597	2.3074	200.3049	0.2297	3.0848	4.3534	6.7	66	101.5151515
Google Coral TPU	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TFLite	INT8	1								6.7	66	101.5151515
Google Coral TPU	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TFLite	INT8	2								6.7	66	101.5151515
Google Coral TPU	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TFLite	INT8	4								6.7	66	101.5151515
Google Coral TPU	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TFLite	INT8	8								6.7	66	101.5151515
Google Coral TPU	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TFLite	INT8	16								6.7	66	101.5151515
Google Coral TPU	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TFLite	INT8	32								6.7	66	101.5151515
Google Coral TPU	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TFLite	INT8	64								6.7	66	101.5151515
Google Coral TPU	18.04 / 32.6.1 / 20.10	DistilBERT	GLUE SST-2	TFLite	INT8	128								6.7	66	101.5151515

BERT

Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	1	0.9317	111.6667	0.8395	121261.4603	4.8505	0.2060	0.2062	13.39	110	121.7272727
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	2	0.9317	113.7475	0.9170	115239.0107	4.6096	0.2167	0.2169	13.39	110	121.7272727
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	4	0.9317	113.7475	0.8231	114094.7305	4.5638	0.2189	0.2191	13.39	110	121.7272727
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	8	0.9221	113.7475	0.9000	112788.7456	4.5115	0.2214	0.2217	13.39	110	121.7272727
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	16								13.39	110	121.7272727
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	32								13.39	110	121.7272727
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	64								13.39	110	121.7272727
Nvidia Jetson TX1	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	128								13.39	110	121.7272727
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	1	0.9317	99.3734	0.6965	106923.4484	4.2769	0.2336	0.2338	13.39	110	121.7272727
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	2	0.9317	99.3734	0.2575	103062.0319	4.1225	0.2423	0.2426	13.39	110	121.7272727
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	4	0.9317	99.3734	0.2343	102395.8735	4.0958	0.2439	0.2442	13.39	110	121.7272727
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	8	0.9317	99.3734	0.2458	102382.2849	4.0953	0.2439	0.2442	13.39	110	121.7272727
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	16	0.9317	99.3734	0.1973	102362.6500	4.0945	0.2440	0.2442	13.39	110	121.7272727
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	32	0.9317	99.3734	0.1970	102202.6855	4.0881	0.2444	0.2446	13.39	110	121.7272727
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	64	0.9317	99.3734	0.1952	101898.2051	4.0759	0.2451	0.2453	13.39	110	121.7272727
Nvidia Jetson TX2	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	128	0.9317	99.3734	0.4065	101542.3949	4.0617	0.2460	0.2462	13.39	110	121.7272727
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	1	0.9317	66.1882	0.1192	6262.0173	0.2505	3.9505	3.9923	13.39	110	121.7272727
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	2	0.9317	66.1882	0.1531	6082.0433	0.2433	4.0661	4.1105	13.39	110	121.7272727
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	4	0.9317	66.1882	0.1452	6022.0454	0.2409	4.1062	4.1514	13.39	110	121.7272727
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	8	0.9317	66.1882	0.1506	5962.0398	0.2385	4.1471	4.1932	13.39	110	121.7272727
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	16	0.9317	66.1882	0.1283	5898.9912	0.2360	4.1909	4.2380	13.39	110	121.7272727
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	32	0.9317	66.1882	0.1160	5789.2098	0.2316	4.2695	4.3184	13.39	110	121.7272727
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	64	0.9317	66.1882	0.1142	5789.6695	0.2316	4.2691	4.3180	13.39	110	121.7272727
Nvidia Jetson Xavier	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	128	0.9317	66.1882	0.1143	5841.9759	0.2337	4.2314	4.2794	13.39	110	121.7272727
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	1								13.39	110	121.7272727
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	2								13.39	110	121.7272727
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	4								13.39	110	121.7272727
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	8								13.39	110	121.7272727
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	16								13.39	110	121.7272727
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	32								13.39	110	121.7272727
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	64								13.39	110	121.7272727
Nvidia Jetson Nano	18.04 / 32.6.1 / 20.10	BERT	IMDb	TensorFlow 2.5V	FP32	128								13.39	110	121.7272727
Google Coral TPU	18.04 / 32.6.1 / 20.10	BERT	IMDb	TFLite	INT8	1								13.39	110	121.7272727
Google Coral TPU	18.04 / 32.6.1 / 20.10	BERT	IMDb	TFLite	INT8	2								13.39	110	121.7272727
Google Coral TPU	18.04 / 32.6.1 / 20.10	BERT	IMDb	TFLite	INT8	4								13.39	110	121.7272727
Google Coral TPU	18.04 / 32.6.1 / 20.10	BERT	IMDb	TFLite	INT8	8								13.39	110	121.7272727
Google Coral TPU	18.04 / 32.6.1 / 20.10	BERT	IMDb	TFLite	INT8	16								13.39	110	121.7272727
Google Coral TPU	18.04 / 32.6.1 / 20.10	BERT	IMDb	TFLite	INT8	32								13.39	110	121.7272727
Google Coral TPU	18.04 / 32.6.1 / 20.10	BERT	IMDb	TFLite	INT8	64								13.39	110	121.7272727
Google Coral TPU	18.04 / 32.6.1 / 20.10	BERT	IMDb	TFLite	INT8	128								13.39	110	121.7272727