算法速度参考

yolo 框架相关

device	network	activation	precision	batch	DLA	framework	time
Xavier	yolov3	leaky	fp16	1	no	TRT5.1.6	24ms
Xavier	yolov3	leaky	fp16	1	no	TRT7.1.0	18ms
NX	yolov3	leaky	fp16	1	no	TRT7.1.0	30ms
TX2	yolov3	leaky	fp16	1	no	TRT5.1.6	99ms
Xavier	yolov3	leaky	fp16	4	no	TRT5.1.6	90ms (22.5ms each)
Xavier	yolov3	leaky	fp16	4	no	TRT7.1.0	58ms (14.5ms each)
TX2	yolov3	leaky	fp16	32	no	TRT5.1.6	2930ms (91.5ms each)
Xavier	yolov3	leaky	fp16	32	no	TRT7.1.0	440ms (13.75ms each)
NX	yolov3	leaky	fp16	4	no	TRT7.1.0	104ms (26ms each)
Xavier	yolov3	leaky	int8	1	no	TRT5.1.6	20ms[1]
Xavier	yolov3	leaky	int8	1	no	TRT7.1.0	12.5ms
NX	yolov3	leaky	int8	1	no	TRT7.1.0	20ms
Xavier	yolov3	leaky	int8	4	no	TRT5.1.6	66ms (16.5ms each)
Xavier	yolov3	leaky	int8	4	no	TRT7.1.0	36ms (9ms each)
Xavier	yolov3	leaky	int8	32	no	TRT7.1.0	256ms (8ms each)
NX	yolov3	leaky	int8	4	no	TRT7.1.0	64ms (16ms each)
Xavier	yolov3	relu	fp16	4	no	TRT5.1.6	52ms (13ms each)
Xavier	yolov3	relu	int8	1	no	TRT5.1.6	10ms
NX	yolov3	relu	int8	1	no	TRT7.1.0	17ms
Xavier	yolov3	relu	int8	4	no	TRT5.1.6	30ms (7.5ms each)
NX	yolov3	relu	int8	4	no	TRT7.1.0	58ms (14.5ms each)
Xavier	yolov3(608)	relu	int8	4	no	TRT5.1.6	54ms (13.5ms each)
1050ti	yolov3	relu	int8	4	no	TRT5.1.6	45ms (11.25ms each)
Xavier	yolov3-tiny	leaky	fp16	1	no	TRT5.1.6	5ms
Xavier	yolo-resnet	leaky	fp16	1	no	TRT5.1.6	14ms
Xavier	yolo-resnet	leaky	fp16	4	no	TRT5.1.6	44ms (11ms each)
Xavier	yolo-resnet	leaky	int8	1	no	TRT5.1.6	12ms[2]
Xavier	yolo-resnet	leaky	int8	4	no	TRT5.1.6	39ms (10ms each)
Xavier	yolo-resnet	relu	fp16	4	no	TRT5.1.6	30ms (7.5ms each)
Xavier	yolo-resnet	relu	fp16	4	yes	TRT5.1.6	68ms (17ms each)
Xavier	yolo-resnet	relu	int8	4	no	TRT5.1.6	22ms (5.5ms each)
1050ti	yolo-resnet	relu	int8	4	no	TRT5.1.6	24ms (6ms each)

编辑

[1] 由于Tensorrt在对leaky relu的int8的实现效果很差，实际速度不是理想速度，使用relu测试得到的速度为14ms.
[2] 同[1].

Mask RCNN 框架相关

device	input shape	precision	batch	framework	pure enqueue time
1050ti	1024×1024	fp32	1	TRT7.0	364ms
1050ti	1024×1024	int8	1	TRT7.0	140ms
Xavier	1024×1024	fp16	1	TRT7.1	136ms
Xavier	1024×1024	int8	1	TRT7.1	103ms
NX	1024×1024	fp32	1	TRT7.1	871ms
NX	1024×1024	fp16	1	TRT7.1	239ms
NX	1024×1024	int8	1	TRT7.1	165ms

编辑

[1]由于暂时只有tensorrt6以上的版本有Mask RCNN的示例程序，因此暂时没有在Xavier上进行测试，200ms的enqueue time是经验估计值

图像分类网络相关

device	network	precision	batch	DLA	framework	pure enqueue time
apex	google net	int8	1	no	TRT5.1.6	1.5ms
apex	google net	int8	4	no	TRT5.1.6	3.5ms(avg 0.9ms)
apex	google net	int8	8	no	TRT5.1.6	5.5ms(avg 0.7ms)
apex	google net	int8	32	no	TRT5.1.6	17.5ms(avg 0.55ms)
apex	google net	int8	128	no	TRT5.1.6	64ms(avg 0.5ms)
NX	google net	half	1	no	TRT7.1	3ms
NX	google net	half	32	no	TRT7.1	
NX	google net	int8	1	no	TRT7.1	2ms
NX	google net	int8	32	no	TRT7.1	
apex	resnet50	int8	1	no	TRT5.1.6	2.2ms
apex	resnet50	int8	4	no	TRT5.1.6	4.3ms(avg 1.1ms)
apex	resnet50	int8	8	no	TRT5.1.6	7.5ms(avg 0.9ms)
apex	resnet50	int8	32	no	TRT5.1.6	25ms(avg 0.8ms)
apex	resnet50	int8	128	no	TRT5.1.6	94.5ms(avg 0.74ms)
NX	resnet50	half	1	no	TRT7.1	6ms
NX	resnet50	half	32	no	TRT7.1	103ms(avg 3.2ms)
NX	resnet50	int8	1	no	TRT7.1	3.8ms
NX	resnet50	int8	32	no	TRT7.1	64ms(avg 2ms)
tx2	resnet50	half	1	no	TRT5.1.6	13ms
tx2	resnet50	half	32	no	TRT5.1.6	320ms(avg 10ms)
tx2	google net	half	1	no	TRT5.1.6	5.2ms
tx2	google net	half	32	no	TRT5.1.6	118ms(avg 3.7ms)
1050ti	google net	float32	1	no	TRT5.1.6	3.5ms
1050ti	google net	float32	4	no	TRT5.1.6	11ms(avg 2.75ms)
1050ti	google net	float32	8	no	TRT5.1.6	16ms(avg 2ms)
1050ti	google net	float32	32	no	TRT5.1.6	61ms(avg 1.9ms)
1050ti	google net	float32	128	no	TRT5.1.6	236ms(avg 1.84ms)
1050ti	google net	int8	1	no	TRT5.1.6	1.5ms
1050ti	google net	int8	4	no	TRT5.1.6	4.5ms(avg 1.1ms)
1050ti	google net	int8	8	no	TRT5.1.6	6ms(avg 0.75ms)
1050ti	google net	int8	32	no	TRT5.1.6	24ms(avg 0.75ms)
1050ti	google net	int8	128	no	TRT5.1.6	90ms(avg 0.7ms)
1050ti	resnet50	float32	1	no	TRT5.1.6	8ms
1050ti	resnet50	float32	4	no	TRT5.1.6	23ms(avg 5.75ms)
1050ti	resnet50	float32	8	no	TRT5.1.6	38ms(avg 4.75ms)
1050ti	resnet50	float32	32	no	TRT5.1.6	133ms(avg 4ms)
1050ti	resnet50	float32	128	no	TRT5.1.6	510ms(avg 4ms)
1050ti	resnet50	int8	1	no	TRT5.1.6	3ms
1050ti	resnet50	int8	4	no	TRT5.1.6	8ms(avg 2ms)
1050ti	resnet50	int8	8	no	TRT5.1.6	14ms(avg 1.75ms)
1050ti	resnet50	int8	32	no	TRT5.1.6	44ms(avg 1.4ms)
1050ti	resnet50	int8	128	no	TRT5.1.6	167ms(avg 1.3ms)

编辑