TF serving Docker -v100	tensorflow/serving:1.13.0-gpu		tensorflow/serving		1.13.0-gpu	f46687a9404d	2 weeks ago	2.38GB
TF serving Docker -1070	tensorflow/serving:latest-gpu		tensorflow/servi	latest-gpu	ad003bd9bd92	4 months ago	2.18 GB	
How models are created (Ke	https://colab.research.google.com/drive/1u79vDN4MZuq6gYIOkPmWsbghjunbDq6m							
Client code	https://gist.github.com/alexcpn/f2366cbfcd0e0dfb22f0f551f8fc0161							
Client on same machine	alexcpn/tfserving-keras-retinanet-dev-gpu				latest	18797ab9f0e7		
Thu Mar 14 22:59:42 2019		v100						
+---+								
| NVIDIA-SMI 384.145 Driver Version: 384.145 |								
|-------------------------------+----------------------+----------------------+								
| GPU Name Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC |								
| Fan Temp Perf Pwr:Usage/Cap| Memory-Usage | GPU-Util Compute M. |								
|===============================+======================+======================|								
| 0 Tesla V100-PCIE... Off | 00000000:B3:00.0 Off | 0 |								
| N/A 34C P0 35W / 250W | 31175MiB / 32502MiB | 0% Default |								
+-------------------------------+----------------------+----------------------+								
+---+								
| Processes: GPU Memory |								
| GPU PID Type Process name Usage |								
|===|								
| 0 218965 C tensorflow_model_server 31165MiB |								
+---+								
alex@drone-OMEN:~$ nvidia-smi								
Sat Mar 16 01:12:00 2019								
+---+								
| NVIDIA-SMI 384.130 Driver Version: 384.130 |								
|-------------------------------+----------------------+----------------------+								
| GPU Name Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC |								
| Fan Temp Perf Pwr:Usage/Cap| Memory-Usage | GPU-Util Compute M. |								
|===============================+======================+======================|								
| 0 GeForce GTX 1070 Off | 00000000:01:00.0 Off | N/A |								
| N/A 48C P8 12W / N/A | 7851MiB / 8105MiB | 1% Default |								
+-------------------------------+----------------------+----------------------+								
+---+								
| Processes: GPU Memory |								
| GPU PID Type Process name Usage |								
|===|								
| 0 1862 G /usr/lib/xorg/Xorg 202MiB |								
| 0 7302 G ...-token=EBF4E6EC29813767B25DB5C2E5B69C9D 152MiB |								
| 0 9018 G ...-token=D5172C7A8844F7AA8608E1B4C4EF62A2 40MiB |								
| 0 21994 C tensorflow_model_server 7451MiB |								
+---+								

Reinanet model converted from Keras

Client - https://github.com/alexcpn/tf_serving_clients/blob/master/retinanet_client.py

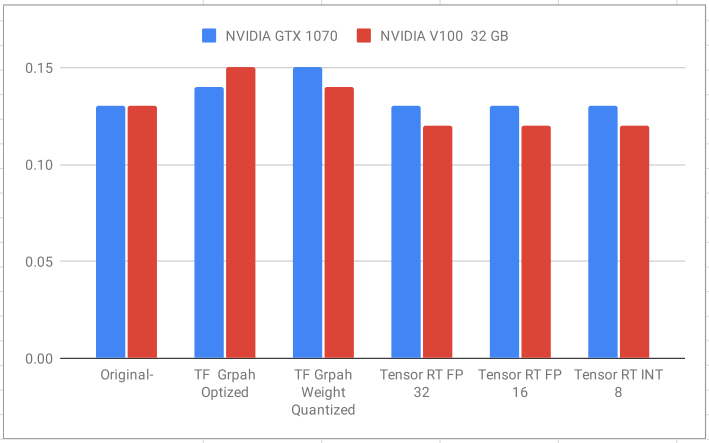
Batchsize is 1 here

Model 4	Keras converted TF sering
Model 6	TF Graph simple optimisation
Model 7	TF Graph simple optimisation + Weight Qunatization
Model 8	TF Graph simple optimisation + Weight + Model Qunatization
Model 9	Based on Model 4 frozen; NVIDIA Tensor RT Optimisation FP 32
Model 10	Based on Model 4 frozen; NVIDIA Tensor RT Optimisation FP 16
Model 11	Based on Model 4 frozen; NVIDIA Tensor RT Optimisation INT 8

('in image shape', (800, 1067, 3))

('in tf shape', (1, 800, 1067, 3))

No of Runs 1		
Model	**NVIDIA GTX 1070**	**NVIDIA V100 32 GB**
Original-	0.13	0.13
TF Grpah Optized	0.14	0.15
TF Grpah Weight Quanti	0.15	0.14
Tensor RT FP 32	0.13	0.12
Tensor RT FP 16	0.13	0.12
Tensor RT INT 8	0.13	0.12
No of runs :10		
4	1.15	0.81
6	1.34	1.16
7	1.15	1.27
9	1.23	0.82
10	1.22	0.83
11	1.22	0.85



MODEL = https://github.com/tensorflow/models/tree/master/official/resnet

FP32 = http://download.tensorflow.org/models/official/20181001_resnet/savedmodels/resnet_v2_fp32_savedmodel_NCHW.tar.gz

FP16 = http://download.tensorflow.org/models/official/20181001_resnet/savedmodels/resnet_v2_fp16_savedmodel_NCHW.tar.gz

Server

docker run --net=host --runtime=nvidia -it --rm -p 8900:8500 -p 8901:8501

-v /home/alex/coding/IPython_neuralnet/:/models/ tensorflow/serving:latesgpu

--model_config_file=/models/resnet50_fp32.json or resnet50_fp16.json

Client - https://github.com/alexcpn/tf_serving_clients/blob/master/resnet50_client.py

Dev contianer with retinanet - alexcpn/tfserving-keras-retinanet-dev-gpu

python resnet50_client.py -model_name=resnet_32 -batch_size=64

---

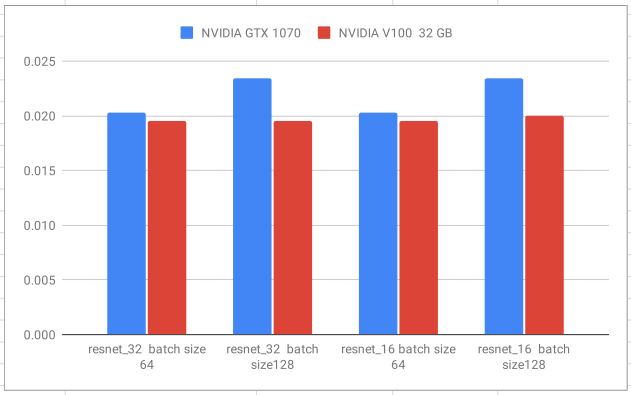
docker run -it --runtime=nvidia --net=host -v /home/alex/coding/IPython_neuralnet:/coding --rm alexcpn/tfserving-keras-retinanet-dev-gp

('in image shape', (224, 224, 3))
('in tf shape', (2, 224, 224, 3))
('result no', 0)
('scores output', (2, 1001))
('labels output', (2,))
('Label', 535)

('Label', 535)

No of Runs 1/batch size 64	**Model resnet_32**	
	NVIDIA GTX 1070	**NVIDIA V100 32 GB**
Best time	1.3	1.25
per image	**0.0203125**	**0.01953125**
No of Runs 1/batch size 128	**Model resnet_32**	
best time	3	2.5
per image	**0.0234375**	**0.01953125**
No of Runs 1/batch size 64	**Model resnet_16**	
best time	**1.3**	1.25
	0.0203125	**0.01953125**
No of Runs 1/batch size 128	**Model resnet_16**	
best time	3	2.56
per image	**0.0234375**	**0.02**

Model Batch-size	NVIDIA GTX 1070	NVIDIA V100 32 GB
resnet_32 batch size 64	0.0203125	0.01953125
resnet_32 batch size128	0.0234375	0.01953125
resnet_16 batch size 64	0.0203125	0.01953125
resnet_16 batch size128	0.0234375	0.02



Model -ssd_resnet_50_fpn_coco -https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md

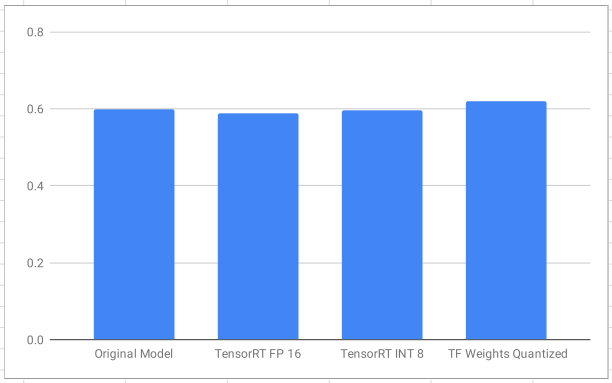
Server docker run --net=host --runtime=nvidia -it --rm -p 8900:8500 -p 8901:8501 -v /usr/alex/:/models tensorflow/serving:1.13.0-gpu --rest_api_port=0 --enable_batching=true --model_config_file=/models/ssd_inception_v3_coco.json

Colab -https://colab.research.google.com/drive/1wQpWoc40kf__WSjfTqDaReMx6fFjUn48

Client

```
  docker run --entrypoint=/bin/bash --env http_proxy=<my proxy> --env https_proxy=<my proxy>  --runtime=nvidia  -it --rm -p 8900:8500 -p 8901:8501 -v /usr/alex/:/coding --net=host tensorflow/tensorflow:1.13.0rc1-gpu-jupyter
pip install tensorflow-serving-api
pip install opencv-python==3.3.0.9
cd coding
python ssd_client_1.py -num_tests=1 -server=127.0.0.1:8500 -batch_size=1 -img_path='../examples/google1.jpg/'
```

Original Model	0.5993821621
TensorRT FP 16	0.5885560513
TensorRT INT 8	0.5967140198
TF Weights Quantized	0.6182711124



model 7	1070 intermittent
2019-03-15 18:53:58.890185: F external/org_tensorflow/tensorflow/core/common_runtime/bfc_allocator.cc:458] Check failed: c->in_use() && (c->bin_num == kInvalidBinNum)	
/usr/bin/tf_serving_entrypoint.sh: line 3: 7 Aborted (core dumped) tensorflow_model_server --port=8500 --rest_api_port=8501 --model_name=${MODEL_NAME} --model_base_path=${MODEL_BASE_PATH}/${MODEL_NAME} "$@"	

model 8	10170
Using TensorFlow backend.	
hello from TFServing v2 client slim < -num_tests -server>	
('Number of tests=', 1)	
('in image shape', (800, 1067, 3))	
<_Rendezvous of RPC that terminated with:	
	status = StatusCode.UNIMPLEMENTED
	details = "Broadcast between [1,9,4] and [221,1,4] is not supported yet.
	[[{{node anchors_3/add_2/eightbit}} = QuantizedAdd[T1=DT_QUINT8, T2=DT_QUINT8, Toutput=DT_QINT32, _device="/job:localhost/replica:0/task:0/device:CPU:0"](anchors_3/Reshape_2/eightbit, anchors_3/Reshape_3/eightbit, anchors_3/Reshape_2/eightbit:1, anchors_3/Reshape_2/eightbit:2, anchors_3/Reshape_3/eightbit:1, anchors_3/Reshape_3/eightbit:2)]
	[[{{node filtered_detections/map/while/non_max_suppression_67/NonMaxSuppressionV3/_991}} = _Recv[client_terminated=false, recv_device="/job:localhost/replica:0/task:0/device:GPU:0", send_device="/job:localhost/replica:0/task:0/device:CPU:0", send_device_incarnation=1, tensor_name="edge_9081_...pressionV3", tensor_type=DT_INT32, _device="/job:localhost/replica:0/task:0/device:GPU:0"](^_cloopfiltered_detections/map/while/non_max_suppression_67/NonMaxSuppressionV3/_991)]]
	debug_error_string = "{"created":"@1552678835.544594429","description":"Error received from peer","file":"src/core/lib/surface/call.cc","file_line":1017,"grpc_message":"Broadcast between [1,9,4] and [221,1,4] is not supported yet.\n\t [[{{node anchors_3/add_2/eightbit}} = QuantizedAdd[T1=DT_QUINT8, T2=DT_QUINT8, Toutput=DT_QINT32, _device="/job:localhost/replica:0/task:0/device:CPU:0"](anchors_3/Reshape_2/eightbit, anchors
>	
Exception in thread Thread-1:	
Traceback (most recent call last):	
File "/usr/lib/python2.7/threading.py", line 801, in __bootstrap_inner	
self.run()	
File "/usr/lib/python2.7/threading.py", line 754, in run	
self.__target(*self.__args, **self.__kwargs)	
File "/usr/local/lib/python2.7/dist-packages/grpc/_channel.py", line 707, in channel_spin	
call_completed = event.tag(event)	
File "/usr/local/lib/python2.7/dist-packages/grpc/_channel.py", line 173, in handle_event	
callback()	
File "/usr/local/lib/python2.7/dist-packages/grpc/_channel.py", line 329, in <lambda>	
self._state.callbacks.append(lambda: fn(self))	
File "retinanet_client.py", line 63, in _callback	
parse_result(result_future.result())	
File "/usr/local/lib/python2.7/dist-packages/grpc/_channel.py", line 295, in result	
raise self	
_Rendezvous: <_Rendezvous of RPC that terminated with:	
	status = StatusCode.UNIMPLEMENTED
	details = "Broadcast between [1,9,4] and [221,1,4] is not supported yet.
	[[{{node anchors_3/add_2/eightbit}} = QuantizedAdd[T1=DT_QUINT8, T2=DT_QUINT8, Toutput=DT_QINT32, _device="/job:localhost/replica:0/task:0/device:CPU:0"](anchors_3/Reshape_2/eightbit, anchors_3/Reshape_2/eightbit:1, anchors_3/Reshape_2/eightbit:2, anchors_3/Reshape_3/eightbit:1, anchors_3/Reshape_3/eightbit:2)]
	[[{{node filtered_detections/map/while/non_max_suppression_67/NonMaxSuppressionV3/_991}} = _Recv[client_terminated=false, recv_device="/job:localhost/replica:0/task:0/device:GPU:0", send_device="/job:localhost/replica:0/task:0/device:CPU:0", send_device_incarnation=1, tensor_name="edge_9081_...pressionV3", tensor_type=DT_INT32, _device="/job:localhost/replica:0/task:0/device:GPU:0"](^_cloopfiltered_detections/map/while/non_max_suppression_67/NonMaxSuppressionV3/_991)]]
	debug_error_string = "{"created":"@1552678835.544594429","description":"Error received from peer","file":"src/core/lib/surface/call.cc","file_line":1017,"grpc_message":"Broadcast between [1,9,4] and [221,1,4] is not supported yet.\n\t [[{{node anchors_3/add_2/eightbit}} = QuantizedAdd[T1=DT_QUINT8, T2=DT_QUINT8, Toutput=DT_QINT32, _device="/job:localhost/replica:0/task:0/device:CPU:0"](anchors_3/Reshape_2/eightbit, anchors
>	

model 8	v100
File "/usr/local/lib/python2.7/dist-packages/grpc/_channel.py", line 295, in result	
raise self	
_Rendezvous: <_Rendezvous of RPC that terminated with:	
	status = StatusCode.UNIMPLEMENTED
	details = "Broadcast between [1,9,4] and [221,1,4] is not supported yet.
	[[{{node anchors_3/add_2/eightbit}}]]
	[[{{node filtered_detections/map/while/non_max_suppression_47/NonMaxSuppressionV3}}]]"
	debug_error_string = "{"created":"@1552648163.376920114","description":"Error received from peer","file":"src/core/lib/surface/call.cc","file_line":1017,"grpc_message":"Broadcast between [1,9,4] and [221,1,4] is not supported yet.\n\t [[{{node anchors_3/add_2/eightbit}}]]\n\t [[{{node filtered_detections/map/while/non_max_suppression_47/NonMaxSuppressionV3}}]]","grpc_status":12}"

model all optimised	v100
following warning is observed at intial load	
e_2: Error: Pack node (import/boxes/stack) axis attribute is out of bounds: 2	
2019-03-15 11:36:04.543823: W external/org_tensorflow/tensorflow/core/grappler/optimizers/graph_optimizer_stage.h:241] Failed to run optimizer ArithmeticOptimizer, stage RemoveStackStridedSliceSameAxis node import/clipped_boxes/strided_slice_3. Error: Pack node (import/boxes/stack) axis attribute is out of bounds: 2	
2019-03-15 11:36:04.543851: W external/org_tensorflow/tensorflow/core/grappler/optimizers/graph_optimizer_stage.h:241] Failed to run optimizer ArithmeticOptimizer, stage RemoveStackStridedSliceSameAxis node import/clipped_boxes/strided_slice_4. Error: Pack node (import/boxes/stack) axis attribute is out of bounds: 2	
2019-03-15 11:36:04.543900: W external/org_tensorflow/tensorflow/core/grappler/optimizers/graph_optimizer_stage.h:241] Failed to run optimizer ArithmeticOptimizer, stage RemoveStackStridedSliceSameAxis node import/clipped_boxes/strided_slice_5. Error: Pack node (import/boxes/stack) axis attribute is out of bounds: 2	
2019-03-15 11:36:07.160918: W external/org_tensorflow/tensorflow/core/grappler/optimizers/graph_optimizer_stage.h:241] Failed to run optimizer ArithmeticOptimizer, stage RemoveStackStridedSliceSameAxis node import/clipped_boxes/strided_slice_2. Error: Pack node (import/boxes/stack) axis attribute is out of bounds: 2	
2019-03-15 11:36:07.160988: W external/org_tensorflow/tensorflow/core/grappler/optimizers/graph_optimizer_stage.h:241] Failed to run optimizer ArithmeticOptimizer, stage RemoveStackStridedSliceSameAxis node import/clipped_boxes/strided_slice_3. Error: Pack node (import/boxes/stack) axis attribute is out of bounds: 2	
2019-03-15 11:36:07.161009: W external/org_tensorflow/tensorflow/core/grappler/optimizers/graph_optimizer_stage.h:241] Failed to run optimizer ArithmeticOptimizer, stage RemoveStackStridedSliceSameAxis node import/clipped_boxes/strided_slice_4. Error: Pack node (import/boxes/stack) axis attribute is out of bounds: 2	
2019-03-15 11:36:07.161028: W external/org_tensorflow/tensorflow/core/grappler/optimizers/graph_optimizer_stage.h:241] Failed to run optimizer ArithmeticOptimizer, stage RemoveStackStridedSliceSameAxis node import/clipped_boxes/strided_slice_5. Error: Pack node (import/boxes/stack) axis attribute is out of bounds: 2	