Number of GPUs	Batch size per GPU	Training time with FP16 (Hrs)	Training time with FP32 (Hrs)
8	4	0.51	0.77

Training stability test

The following tables compare `F1` scores across 5 different training runs with different seeds, for both FP16 and FP32 respectively. The runs showcase consistent convergence on all 5 seeds with very little deviation.

FP16, 8x GPUs	seed #1	seed #2	seed #3	seed #4	seed #5	mean	std
F1	91.16	90.69	90.99	90.94	91.17	90.99	0.196
Exact match	84.2	83.68	84.14	83.95	84.34	84.06	0.255

FP32, 8x GPUs	seed #1	seed #2	seed #3	seed #4	seed #5	mean	std
F1	90.67	90.8	90.94	90.83	90.93	90.83	0.11
Exact match	83.56	83.96	83.99	83.95	84.12	83.92	0.21

Training performance results

Our results were obtained by running batch sizes up to 3x GPUs on a 16GB V100 and up to 10x GPUs on a 32G V100 with mixed precision.

NVIDIA DGX-1 (8x V100 16G)

Our results were obtained by running the `scripts/run_squad.sh` training script in the TensorFlow 19.03-py3 NGC container on NVIDIA DGX-1 with 8x V100 16G GPUs. Performance (in sentences per second) is the steady-state throughput.

Number of GPUs	Batch size per GPU	FP32 sentences/sec	FP16 sentences/sec	Speed-up with mixed precision	Multi-gpu weak scaling with FP32	Multi-gpu weak scaling with FP16
1	2	8.06	14.12	1.75	1.0	1.0
4	2	25.71	41.32	1.61	3.19	2.93
8	2	50.20	80.76	1.61	6.23	5.72

Number of GPUs	Batch size per GPU	FP32 sentences/sec	FP16 sentences/sec	Speed-up with mixed precision	Multi-gpu weak scaling with FP32	Multi-gpu weak scaling with FP16
1	3	-	17.14	-	-	1.0
4	3	-	51.59	-	-	3.0
8	3	-	98.75	-	-	5.76

Note: The respective values for FP32 runs that use a batch size of 3 are not available due to out of memory errors that arise. Batch size of 3 is only available on using FP16.

To achieve these same results, follow the Quick Start Guide outlined above.

NVIDIA DGX-1 (8x V100 32G)