CUDNN Simulation in DeepBench														Our V100 Model	
w (VAR_X Nx)	h (iVAR_Y Ny)	c (VAR_C, Ni)	n (elem / batch) (VAR_N)	k (VAR_K, No)	f_w kernel width (VAR_S)	f_h kernel height (VAR_R)	pad_w	pad_h	stride_w	stride_h	precision	fwd_time (usec)	fwd_algo	fwd_time_model (sec)	model_bound (memory, comp, buffer?)
CONVOLUTION KERNEL - Batch Size Variance (Kernel Size #1)															
224	224	64	64	64	3	3	1	1	1	1	half (32)	26165	GEMM	12661	computation
224	224	32	32	32	3	3	1	1	1	1	half (32)	3341	GEMM	1583	computation
224	224	16	16	16	3	3	1	1	1	1	half (32)	862	GEMM	284	mem
224	224	8	8	8	3	3	1	1	1	1	half (32)	233	GEMM	31	mem
224	224	4	4	4	3	3	1	1	1	1	half (32)	75	GEMM	8	mem
224	224	2	2	2	3	3	1	1	1	1	half (32)	27	GEMM	2	mem
224	224	1	1	1	3	3	1	1	1	1	half (32)	12	GEMM	1	mem
CONVOLUTION KERNEL - Batch Size Variance (Kernel Size #2)															
14	14	512	512	512	3	3	1	1	1	1	half (32)	39225	GEMM	25322	computation
14	14	256	256	256	3	3	1	1	1	1	half (32)	4988	GEMM	3165	computation
14	14	128	128	128	3	3	1	1	1	1	half (32)	738	GEMM	396	computation
14	14	64	64	64	3	3	1	1	1	1	half (32)	136	GEMM	50	computation
14	14	32	32	32	3	3	1	1	1	1	half (32)	30	GEMM	6	computation
14	14	16	16	16	3	3	1	1	1	1	half (32)	17	GEMM	1	computation
14	14	8	8	8	3	3	1	1	1	1	half (32)	9	GEMM	1	computation
14	14	4	4	4	3	3	1	1	1	1	half (32)	7	GEMM	1	computation
14	14	2	2	2	3	3	1	1	1	1	half (32)	7	GEMM	1	computation
14	14	1	1	1	3	3	1	1	1	1	half (32)	7	GEMM	1	computation
CONVOLUTION KERNEL - Data Size Variance															
16384	16384	1	1	1	3	3	1	1	1	1	half (32)	42806	GEMM	3262	mem
8192	8192	1	1	1	3	3	1	1	1	1	half (32)	10705	GEMM	1270	mem
4096	4096	1	1	1	3	3	1	1	1	1	half (32)	2690	GEMM	318	mem
2048	2048	1	1	1	3	3	1	1	1	1	half (32)	676	GEMM	80	mem
1024	1024	1	1	1	3	3	1	1	1	1	half (32)	171	GEMM	20	mem
512	512	1	1	1	3	3	1	1	1	1	half (32)	46	GEMM	1	mem
256	256	1	1	1	3	3	1	1	1	1	half (32)	19	GEMM	1	mem
128	128	1	1	1	3	3	1	1	1	1	half (32)	8	GEMM	1	mem
64	64	1	1	1	3	3	1	1	1	1	half (32)	7	GEMM	1	mem
32	32	1	1	1	3	3	1	1	1	1	half (32)	7	GEMM	1	mem
CONVOLUTION KERNEL - Kernel Size Analysis															
224	224	64	64	64	5	5	1	1	1	1	half (32)	159772	GEMM	91239	comp
224	224	64	64	64	4	4	1	1	1	1	half (32)	106047	GEMM	22508	comp
224	224	64	64	64	3	3	1	1	1	1	half (32)	53927	GEMM	12661	comp
224	224	64	64	64	2	2	1	1	1	1	half (32)	26974	GEMM	5627	comp
224	224	64	64	64	1	1	1	1	1	1	half (32)	8609	GEMM	2219	mem

					CUDNN Simulation in DeepBench		Our V100 Model	
m (VAR_N)	n (VAR_C)	k (VAR_K)	a_t	b_t	precision	time (usec)	fwd_time_model (usec)	model_bound (memory, comp, buffer?)
FC KERNEL - Batch Size Variance (Kernel Size #1)								
64	25088	4096	0	0	half (32)	1199	62190	1951
32	25088	4096	0	0	half (32)	1289	31095	975
16	25088	4096	0	0	half (32)	1179	15548	
8	25088	4096	0	0	half (32)	1146	7774	
4	25088	4096	0	0	half (32)	1149	3887	
2	25088	4096	0	0	half (32)	1150	1943	
1	25088	4096	0	0	half (32)	1145	972	
FC KERNEL - Batch Size Variance (Kernel Size #2)								
64	4096	1024	0	0	half (32)	96	2539	
32	4096	1024	0	0	half (32)	95	1270	
16	4096	1024	0	0	half (32)	96	635	
8	4096	1024	0	0	half (32)	95	317	
4	4096	1024	0	0	half (32)	82	159	
2	4096	1024	0	0	half (32)	83	79	
1	4096	1024	0	0	half (32)	81	40	
FC KERNEL - Data Size Variance								
1	16384	16384	0	0	half (32)	9200	2538	
1	8192	8192	0	0	half (32)	1720	635	
1	4096	4096	0	0	half (32)	580	159	
1	2048	2048	0	0	half (32)	224	40	
1	1024	1024	0	0	half (32)	108	1	
1	512	512	0	0	half (32)	76	<	
1	256	256	0	0	half (32)	56	<	
1	128	128	0	0	half (32)	43	<	
1	64	64	0	0	half (32)	40	<	
1	32	32	0	0	half (32)	32		
							<	