week-6

JX-Ma

2024/8/31

1 Intorduce

This week, I mainly implemented the cudnnFindConvolutionForwardAlgorithm() in the conv1-conv12 layer by reading the related documents of cudnn, and obtained the information of the called algorithm.

2 experiment

Firstly, introduce a returned data structure, CUDNnConvolutionFwdAlgoPerf_t, which contains the following parameters:

- cudnnConvolutionFwdAlgo_t algo: Types of Algorithms.
- float time: The execution time (in milliseconds) of cunnConvolutionForward().
- Size_t memory: The size of the workspace in bytes.
- cudnnDeterminism_t determinism: algorithm's determinacy.
- cudnnMathType_t mathType: Provide mathematical types for algorithms.
- int reserved[3]:Reserve space for future assignments

The experimental results of conv1-conv12 are as follows:

表 1: conv1

index	algo	time	determinism	memory	mathtype
0	IMPLICIT_PRECOMP_GEMM	5.01	1	158kB	0
1	IMPLICIT_GEMM	6.37	1	0	0
2	GEMM	8.24	1	562MB	0
3	DIRECT	-1	1	0	0
4	FFT	-1	1	0	0
5	FFT_TILING	-1	1	0	0
6	WINOGRAD	-1	1	0	0
7	WINOGRAD_NONFUSED	-1	1	0	0

表 2: conv2

index	algo	time	determinism	memory	mathtype
0	IMPLICIT_PRECOMP_GEMM	5.12	1	159kB	0
1	IMPLICIT_GEMM	6.37	1	0	0
2	GEMM	9.23	1	582MB	0
3	DIRECT	-1	1	0	0
4	FFT	-1	1	0	0
5	FFT_TILING	-1	1	0	0
6	WINOGRAD	-1	1	0	0
7	WINOGRAD_NONFUSED	-1	1	0	0

表 3: conv3

index	algo	time	determinism	memory	mathtype
0	IMPLICIT_PRECOMP_GEMM	4.43	1	112kB	0
1	IMPLICIT_GEMM	7.48	1	0	0
2	GEMM	13.37	1	927MB	0
3	FFT_TILING	27.68	1	75MB	0
4	FFT	-1	1	0	0
5	DIRECT	-1	1	0	0
6	WINOGRAD	-1	1	0	0
7	WINOGRAD_NONFUSED	-1	1	0	0

表 4: conv4

index	algo	time	determinism	memory	mathtype
0	FFT_TILING	61.28	1	160MB	0
1	IMPLICIT_GEMM	179.9	1	0	0
2	GEMM	1218	1	0	0
3	DIRECT	0	1	0	0
4	FFT	-1	1	0	0
5	IMPLICIT_PRECOMP_GEMM	-1	1	2.03GB	0
6	WINOGRAD	-1	1	0	0
7	WINOGRAD_NONFUSED	-1	1	0	0

表 5: conv5

index	algo	time	determinism	memory	mathtype
0	WINOGRAD_NONFUSED	3.06	1	290MB	0
1	FFT_TILING	3.6	1	303MB	0
2	FFT	3.86	1	323MB	0
3	IMPLICIT_PRECOMP_GEMM	5.58	1	30MB	0
4	IMPLICIT_GEMM	11.42	1	0	0
5	GEMM	12.62	1	491MB	0
6	WINOGRAD	-1	1	0	0
7	DIRECT	-1	1	0	0

表 6: conv6

index	algo	time	determinism	memory	mathtype
0	WINOGRAD_NONFUSED	2.26	1	140MB	0
1	IMPLICIT_PRECOMP_GEMM	2.87	1	23MB	0
2	FFT	3.80	1	368MB	0
3	IMPLICIT_GEMM	5.32	1	0KB	0
4	GEMM	5.41	1	113MB	0
5	WINOGRAD	6.31	1	13MB	0
6	FFT_TILING	11.96	1	958MB	0
7	DIRECT	-1	1	0	0

表 7: conv7

index	algo	time	determinism	memory	mathtype
0	IMPLICIT_PRECOMP_GEMM	7.36	1	296KB	0
1	IMPLICIT_GEMM	14.38	1	0KB	0
2	GEMM	17.76	1	648MB	0
3	WINOGRAD	18.97	1	40KB	0
4	FFT_TILING	29.60	1	73MB	0
5	FFT	-1	1	4328MB	0
6	DIRECT	-1	1	0	0
7	WINOGRAD_NONFUSED	-1	1	3872MB	0

表 8: conv8

index	algo	time	determinism	memory	mathtype
0	FFT_TILING	23.03	1	241MB	0
1	IMPLICIT_PRECOMP_GEMM	25.74	1	401MB	0
2	WINOGRAD	26.09	1	820KB	0
3	IMPLICIT_GEMM	43.20	1	0KB	0
4	FFT	-1	1	2107MB	0
5	DIRECT	-1	1	0	0
6	GEMM	-1	1	3568MB	0
7	WINOGRAD_NONFUSED	-1	1	2775MB	0

表 9: conv9

index	algo	time	determinism	memory	mathtype
0	FFT_TILING	3.41	1	156MB	0
1	WINOGRAD	3.50	1	408KB	0
2	WINOGRAD_NONFUSED	4.91	1	453MB	0
3	IMPLICIT_PRECOMP_GEMM	5.73	1	102MB	0
4	FFT	7.56	1	415MB	0
5	IMPLICIT_GEMM	8.94	1	0KB	0
6	GEMM	12.87	1	86MB	0
7	DIRECT	-1	1	0	0

表 10: conv10

index	algo	time	determinism	memory	mathtype
0	FFT_TILING	2.36	1	206MB	0
1	WINOGRAD_NONFUSED	2.38	1	227MB	0
2	FFT	2.50	1	279MB	0
3	IMPLICIT_PRECOMP_GEMM	2.66	1	51MB	0
4	WINOGRAD	3.61	1	16MB	0
5	IMPLICIT_GEMM	4.57	1	0KB	0
6	GEMM	5.86	1	39MB	0
7	DIRECT	-1	1	0	0

表 11: conv11

index	algo	time	determinism	memory	mathtype
0	WINOGRAD_NONFUSED	1.32	1	92MB	0
1	IMPLICIT_PRECOMP_GEMM	2.11	1	27MB	0
2	FFT	2.13	1	22MB	0
3	WINOGRAD	3.18	1	6MB	0
4	IMPLICIT_GEMM	3.90	1	0KB	0
5	GEMM	4.35	1	165MB	0
6	FFT_TILING	6.39	1	56MB	0
7	DIRECT	-1	1	0	0

表 12: conv12

index	algo	time	determinism	memory	mathtype
0	IMPLICIT_PRECOMP_GEMM	1.57	1	28MB	0
1	WINOGRAD_NONFUSED	1.83	1	110MB	0
2	IMPLICIT_GEMM	2.97	1	0KB	0
3	GEMM	3.27	1	58MB	0
4	WINOGRAD	6.18	1	26MB	0
5	FFT	6.82	1	736MB	0
6	FFT_TILING	22.93	1	1711MB	0
7	DIRECT	-1	1	0	0