ViT_large_patch16_224 fp32

ViT_large_patch16_224 fp32 (N, C, H, W) = (1, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	22.95695782	16.07475996	29.98%	142.81%	无
2	22.9275465	16.58739805	27.65%	138.22%	无
3	23.80232096	16.1694026	32.07%	147.21%	无
4	23.20021391	16.73921347	27.85%	138.60%	无
5	27.33240366	16.09721184	41.11%	169.80%	无
平均	24.04388857	16.33359718	32.07%	147.21%	

ViT_large_patch16_224 fp32 (N, C, H, W) = (4, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	49.88995552	48.11636448	3.56%	103.69%	无
2	49.95594978	48.49206209	2.93%	103.02%	无
3	49.91207838	48.53052616	2.77%	102.85%	无
4	49.95193005	48.49246502	2.92%	103.01%	无
5	50.0581193	48.41672421	3.28%	103.39%	无
平均	49.95360661	48.40962839	3.09%	103.19%	

ViT_large_patch16_224 fp32 (N, C, H, W) = (6, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	69.94958639	67.44493008	3.58%	103.71%	无
2	69.89186049	67.66503334	3.19%	103.29%	无
3	69.85519648	67.43098497	3.47%	103.60%	无
4	69.86840248	67.25456476	3.74%	103.89%	无
5	69.75688934	67.60543585	3.08%	103.18%	无
平均	69.86438704	67.4801898	3.41%	103.53%	

ViT_large_patch16_224 fp32 (N, C, H, W) = (8, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	92.75753021	91.44128799	1.42%	101.44%	无
2	92.81657934	88.93851995	4.18%	104.36%	无
3	93.03803682	89.95167732	3.32%	103.43%	无
4	92.9686451	88.99386883	4.28%	104.47%	无
5	92.95593262	89.34983492	3.88%	104.04%	无
平均	92.90734482	89.7350378	3.41%	103.54%	

ViT_large_patch16_224 fp32 (N, C, H, W) = (16, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	179.9374413	170.6898952	5.14%	105.42%	无
2	180.3622484	174.8656607	3.05%	103.14%	无
3	180.6165862	174.8900104	3.17%	103.27%	无
4	180.8669662	172.8842473	4.41%	104.62%	无
5	180.8048725	172.3608017	4.67%	104.90%	无
平均	180.5176229	173.138123	4.09%	104.26%	

ViT_large_patch16_224 fp32 (N, C, H, W) = (32, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	338.4417915			

ViT_large_patch16_224 fp32 (N, C, H, W) = (64, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	640.7009125			

ViT_large_patch16_224 fp32 (N, C, H, W) = (128, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	1285.151865			

ViT_large_patch16_224 fp32 (N, C, H, W) = (256, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	2593.224592			

ViT_large_patch16_224 fp32 (N, C, H, W) = (512, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	5250.795054			

ViT_large_patch16_224 fp32 (N, C, H, W) = (1024, 3, 224, 224)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	OOM			

ViT_large_patch16_224 fp16

ViT_large_patch16_224 fp16 (N, C, H, W) = (1, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	20.83451271	16.46083832	20.99%	126.57%	33 / 1000 (3.3%)	0.003906	0.5723
2	20.31540632	16.50750399	18.74%	123.07%	43 / 1000 (4.3%)	0.00586	34.8
3	21.68386221	16.32142782	24.73%	132.86%	55 / 1000 (5.5%)	0.007812	9.05
4	21.04858875	15.98062754	24.08%	131.71%	58 / 1000 (5.8%)	0.003906	2.969
5	22.11361885	17.23283052	22.07%	128.32%	43 / 1000 (4.3%)	0.003906	6.445
平均	21.19919777	16.50064564	22.16%	128.47%			

ViT_large_patch16_224 fp16 (N, C, H, W) = (4, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	20.72491169	16.12093687	22.21%	128.56%	121 / 4000 (3.02%)	0.007812	1.403
2	21.42208338	16.78788424	21.63%	127.60%	109 / 4000 (2.73%)	0.003906	3.412
3	21.46075487	16.19868755	24.52%	132.48%	165 / 4000 (4.12%)	0.007812	7
4	21.23618841	16.13240242	24.03%	131.64%	148 / 4000 (3.7%)	0.00586	2.838
5	19.75979328	16.66656733	15.65%	118.56%	269 / 4000 (6.72%)	0.01074	1.857
平均	20.92074633	16.38129568	21.70%	127.71%			

ViT_large_patch16_224 fp16 (N, C, H, W) = (6, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	21.95616722	18.78335238	14.45%	116.89%	215 / 6000 (3.58%)	0.007812	3.852
2	20.90988636	18.79692793	10.11%	111.24%	215 / 6000 (3.58%)	0.007812	2.93
3	21.87682867	18.42218399	15.79%	118.75%	254 / 6000 (4.23%)	0.009766	3.555
4	21.66805029	18.23710203	15.83%	118.81%	230 / 6000 (3.83%)	0.007812	2.46
5	21.72460556	18.21647167	16.15%	119.26%	222 / 6000 (3.7%)	0.007812	182.1
平均	21.62710762	18.4912076	14.50%	116.96%			

ViT_large_patch16_224 fp16 (N, C, H, W) = (8, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	23.8465786	21.13843441	11.36%	112.81%	319 / 8000 (3.99%)	0.007324	18.67
2	23.89268875	21.50665522	9.99%	111.09%	357 / 8000 (4.46%)	0.007812	16
3	24.79342699	22.06675053	11.00%	112.36%	325 / 8000 (4.06%)	0.00586	12.445
4	23.95290852	21.64077282	9.65%	110.68%	422 / 8000 (5.28%)	0.009766	4.133
5	23.88809204	20.96501827	12.24%	113.94%	341 / 8000 (4.26%)	0.007812	2.994
平均	24.07473898	21.46352625	10.85%	112.17%			

ViT_large_patch16_224 fp16 (N, C, H, W) = (16, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	45.52324533	37.75437832	17.07%	120.58%	743 / 16000 (4.64%)	0.00879	37.6
2	45.66379786	38.46408129	15.77%	118.72%	723 / 16000 (4.52%)	0.01172	6.4
3	45.68477631	37.86638737	17.11%	120.65%	597 / 16000 (3.73%)	0.007812	62
4	45.69609165	38.08392286	16.66%	119.99%	578 / 16000 (3.61%)	0.00586	5.625
5	45.63565016	37.85094738	17.06%	120.57%	704 / 16000 (4.4%)	0.012695	10.336
平均	45.64071226	38.00394344	16.73%	120.09%			

ViT_large_patch16_224 fp16 (N, C, H, W) = (32, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	87.48013258	71.12041712	18.70%	123.00%	1185 / 32000 (3.7%)	0.00879	148
2	87.44501114	72.77899981	16.77%	120.15%	1362 / 32000 (4.26%)	0.007812	23.39
3	87.62212992	71.65560722	18.22%	122.28%	1220 / 32000 (3.81%)	0.007812	142
4	87.42598534	70.9308219	18.87%	123.26%	1333 / 32000 (4.17%)	0.009766	21.34
5	87.57584333	71.08132124	18.83%	123.21%	1309 / 32000 (4.09%)	0.007812	25.67
平均	87.50982046	71.51343346	18.28%	122.37%			

ViT_large_patch16_224 fp16 (N, C, H, W) = (64, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	144.2368484					

ViT_large_patch16_224 fp16 (N, C, H, W) = (128, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	270.4936361					

ViT_large_patch16_224 fp16 (N, C, H, W) = (256, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	OOM					

ViT_large_patch16_224 fp16 (N, C, H, W) = (512, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	OOM					

ViT_large_patch16_224 fp16 (N, C, H, W) = (1024, 3, 224, 224)							
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	OOM					

ViT_large_patch32_384 fp32

ViT_large_patch32_384 fp32

ViT_large_patch32_384 fp32 (N, C, H, W) = (1, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	24.50864315	14.40562248	41.22%	170.13%	
2	23.4957552	13.68512154	41.75%	171.69%	
3	23.83404016	13.94410849	41.49%	170.93%	
4	26.67748451	14.38273907	46.09%	185.48%	
5	23.66729736	13.68714571	42.17%	172.92%	
平均	24.43664408	14.02094746	42.62%	174.29%	

ViT_large_patch32_384 fp32 (N, C, H, W) = (4, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	34.11144495	33.53144169	1.70%	101.73%	
2	34.11797285	33.42299938	2.04%	102.08%	
3	34.14151907	33.37360144	2.25%	102.30%	
4	34.14403677	33.2779336	2.54%	102.60%	
5	34.17958736	33.49951982	1.99%	102.03%	
平均	34.1389122	33.42109919	2.10%	102.15%	

ViT_large_patch32_384 fp32 (N, C, H, W) = (6, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	54.41297054	52.34796524	3.80%	103.94%	无
2	54.29521799	52.56785154	3.18%	103.29%	无
3	54.25016403	52.28267431	3.63%	103.76%	无
4	54.19023037	52.3793745	3.34%	103.46%	无
5	54.54382658	52.29720354	4.12%	104.30%	无
平均	54.3384819	52.37501383	3.61%	103.75%	

ViT_large_patch32_384 fp32 (N, C, H, W) = (8, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	68.83318424	66.63008451	3.20%	103.31%	无
2	69.05185699	67.06351519	2.88%	102.96%	无
3	68.95741224	67.2877574	2.42%	102.48%	无
4	69.05191898	68.15471888	1.30%	101.32%	无
5	68.83959055	66.73092604	3.06%	103.16%	无
平均	68.9467926	67.1734004	2.57%	102.64%	

ViT_large_patch32_384 fp32 (N, C, H, W) = (16, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	133.8897061	128.5692573	3.97%	104.14%	无
2	134.33671	129.6730328	3.47%	103.60%	无
3	134.8257303	129.4754887	3.97%	104.13%	无
4	134.7923994	129.1083169	4.22%	104.40%	无
5	134.9441719	129.3283343	4.16%	104.34%	无
平均	134.5577435	129.230886	3.96%	104.12%	

ViT_large_patch32_384 fp32 (N, C, H, W) = (32, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	254.2889166			

ViT_large_patch32_384 fp32 (N, C, H, W) = (64, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	507.728138			

ViT_large_patch32_384 fp32 (N, C, H, W) = (128, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	971.8085647			

ViT_large_patch32_384 fp32 (N, C, H, W) = (256, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	1910.388582			

ViT_large_patch32_384 fp32 (N, C, H, W) = (512, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	3834.607012			

ViT_large_patch32_384 fp32 (N, C, H, W) = (1024, 3, 384, 384)					
次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03
1	OOM	OOM			

ViT_large_patch32_384 fp16

ViT_large_patch32_384 fp16 (N, C, H, W) = (1, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	20.82021952	15.43713093	25.86%	134.87%	33 / 1000 (3.3%)	0.007812	0.2144
2	22.0947361	16.8843317	23.58%	130.86%	25 / 1000 (2.5%)	0.00586	37.4
3	22.04968691	15.91703892	27.81%	138.53%	15 / 1000 (1.5%)	0.003906	0.08606
4	20.69654703	16.63468599	19.63%	124.42%	22 / 1000 (2.2%)	0.003906	2.754
5	21.37932777	15.90518951	25.60%	134.42%	17 / 1000 (1.7%)	0.003906	6.02
平均	21.40810347	16.15567541	24.53%	132.51%			

ViT_large_patch32_384 fp16 (N, C, H, W) = (4, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	20.51131964	15.4150939	24.85%	133.06%	103 / 4000 (2.58%)	0.007812	192
2	19.97449398	16.28556252	18.47%	122.65%	84 / 4000 (2.1%)	0.007812	5.95
3	20.02268791	16.7605567	16.29%	119.46%	87 / 4000 (2.17%)	0.007812	6
4	19.31543589	15.34214497	20.57%	125.90%	92 / 4000 (2.3%)	0.007812	2.63
5	21.72878027	16.03645086	26.20%	135.50%	110 / 4000 (2.75%)	0.007812	22.3
平均	20.31054354	15.96796179	21.38%	127.20%			

ViT_large_patch32_384 fp16 (N, C, H, W) = (6, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	19.5188117	16.8251586	13.80%	116.01%	132 / 6000 (2.2%)	0.007812	12.73
2	20.19104242	16.54681206	18.05%	122.02%	161 / 6000 (2.68%)	0.007812	1.468
3	20.749681	17.88716555	13.80%	116.00%	127 / 6000 (2.12%)	0.007812	2.258
4	20.00149012	17.15390921	14.24%	116.60%	127 / 6000 (2.12%)	0.007812	1.253
5	21.1058569	16.32851124	22.64%	129.26%	140 / 6000 (2.33%)	0.007812	3.428
平均	20.31337643	16.94831133	16.57%	119.85%			

ViT_large_patch32_384 fp16 (N, C, H, W) = (8, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	19.7675252	18.37091208	7.07%	107.60%	201 / 8000 (2.51%)	0.007812	4.066
2	20.32607555	18.65851402	8.20%	108.94%	188 / 8000 (2.35%)	0.007812	5.812
3	20.80201387	18.38161707	11.64%	113.17%	146 / 8000 (1.82%)	0.007812	2.34
4	19.4427228	18.19294691	6.43%	106.87%	183 / 8000 (2.29%)	0.007812	77.3
5	19.7897768	18.35997343	7.22%	107.79%	186 / 8000 (2.33%)	0.007812	4.33
平均	20.02562284	18.3927927	8.15%	108.88%			

ViT_large_patch32_384 fp16 (N, C, H, W) = (16, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	33.5396409	28.50827456	15.00%	117.65%	336 / 16000 (2.1%)	0.007812	5.926
2	33.57994795	28.98382425	13.69%	115.86%	326 / 16000 (2.04%)	0.007812	19
3	33.64556789	28.95655155	13.94%	116.19%	347 / 16000 (2.17%)	0.007812	38
4	33.65916967	28.56039762	15.15%	117.85%	343 / 16000 (2.14%)	0.007812	15.41
5	33.65644455	28.6205554	14.96%	117.60%	366 / 16000 (2.29%)	0.007812	28
平均	33.61615419	28.72592068	14.55%	117.02%			

ViT_large_patch32_384 fp16 (N, C, H, W) = (32, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	65.68508625	54.20515776	17.48%	121.18%	649 / 32000 (2.03%)	0.007812	21.2
2	65.96275806	54.50702429	17.37%	121.02%	665 / 32000 (2.08%)	0.01172	17.33
3	65.99725723	54.76058245	17.03%	120.52%	722 / 32000 (2.26%)	0.007812	34.78
4	66.06442928	54.72256184	17.17%	120.73%	720 / 32000 (2.25%)	0.007812	9.11
5	66.01386309	54.61928368	17.26%	120.86%	678 / 32000 (2.12%)	0.01172	14.61
平均	65.94467878	54.562922	17.26%	120.86%			

ViT_large_patch32_384 fp16 (N, C, H, W) = (64, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	105.3729057					

ViT_large_patch32_384 fp16 (N, C, H, W) = (128, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	205.9424043					

ViT_large_patch32_384 fp16 (N, C, H, W) = (256, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	OOM					

ViT_large_patch32_384 fp16 (N, C, H, W) = (512, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	OOM					

ViT_large_patch32_384 fp16 (N, C, H, W) = (1024, 3, 384, 384)

次数	naive vit(ms)	fused vit(ms)	加速比	加速倍数	精度 rtol=5e-03, atol=1e-03	Max absolute	Max relative difference
1	OOM	OOM					