week-2

JX-Ma

2025/3/1

1 Introduce

This weeks'work is as follow:
1. Modify the content of GPU optimization and combine it with the Roofine Model to write GPU optimization from two aspects: maximizing the utilization of server computing power and bandwidth. Some content on bank conflicts in shared memory has been added.
2.we added the experimental results of memory usage.
3. Drawing: In terms of drawing, it is not possible to change the second y-axis to speed up, because if an algorithm is normalized, the values of that algorithm are all 1, and this approach conflicts with the left y-axis. Since the left side displays tflops values, the normalized result cannot be combined with the TFLOPS result.

2 summary

I suggest breaking down speed up and Tflops into two separate figures, or keeping only the speed up figure, as we will have a general diagram of Tflops in section 7.3.

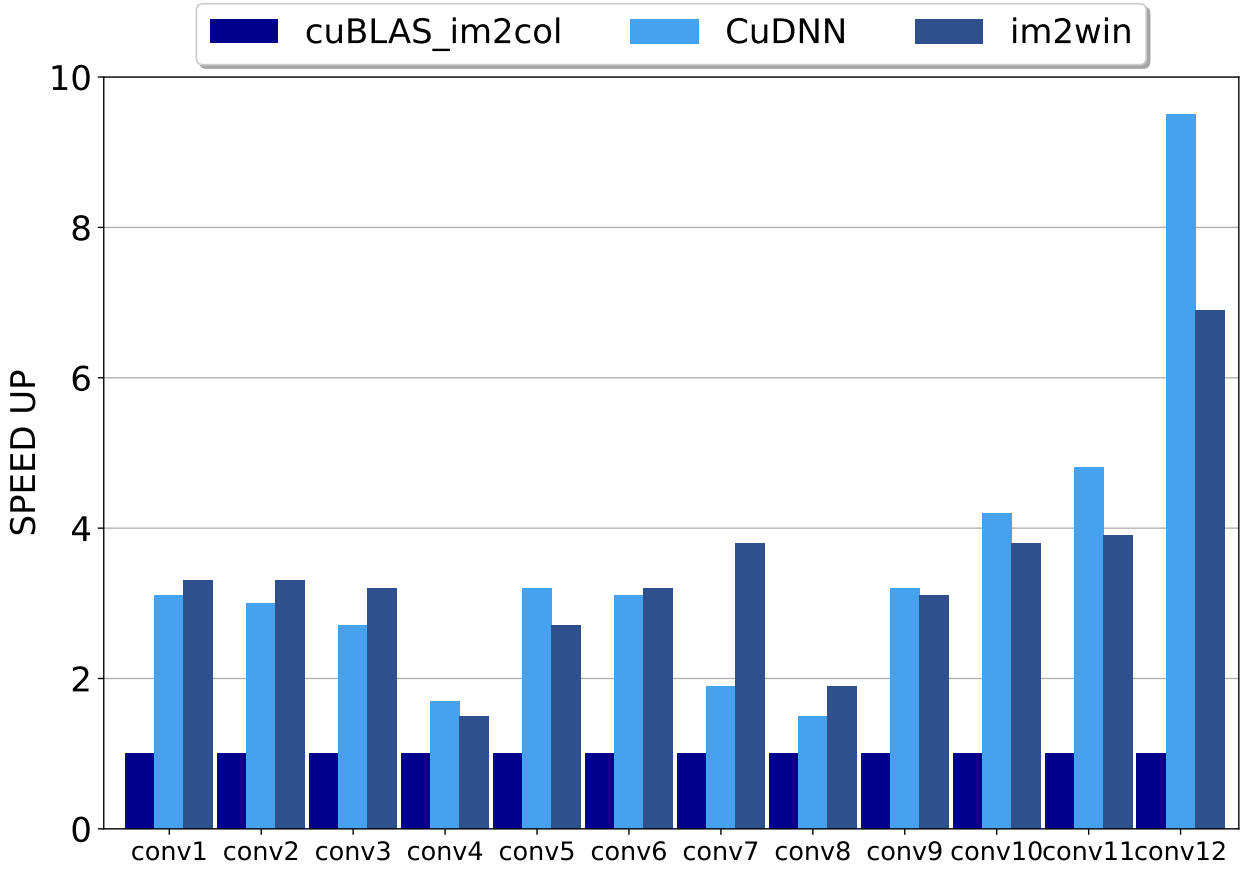
cuBLAS_im2col
CuDNN
im2win
SPEED UP
0
2
4
6
8
10
conv1
conv2
conv3
conv4
conv5
conv6
conv7
conv8
conv9
conv10
conv11
conv12

图 1: batch 256