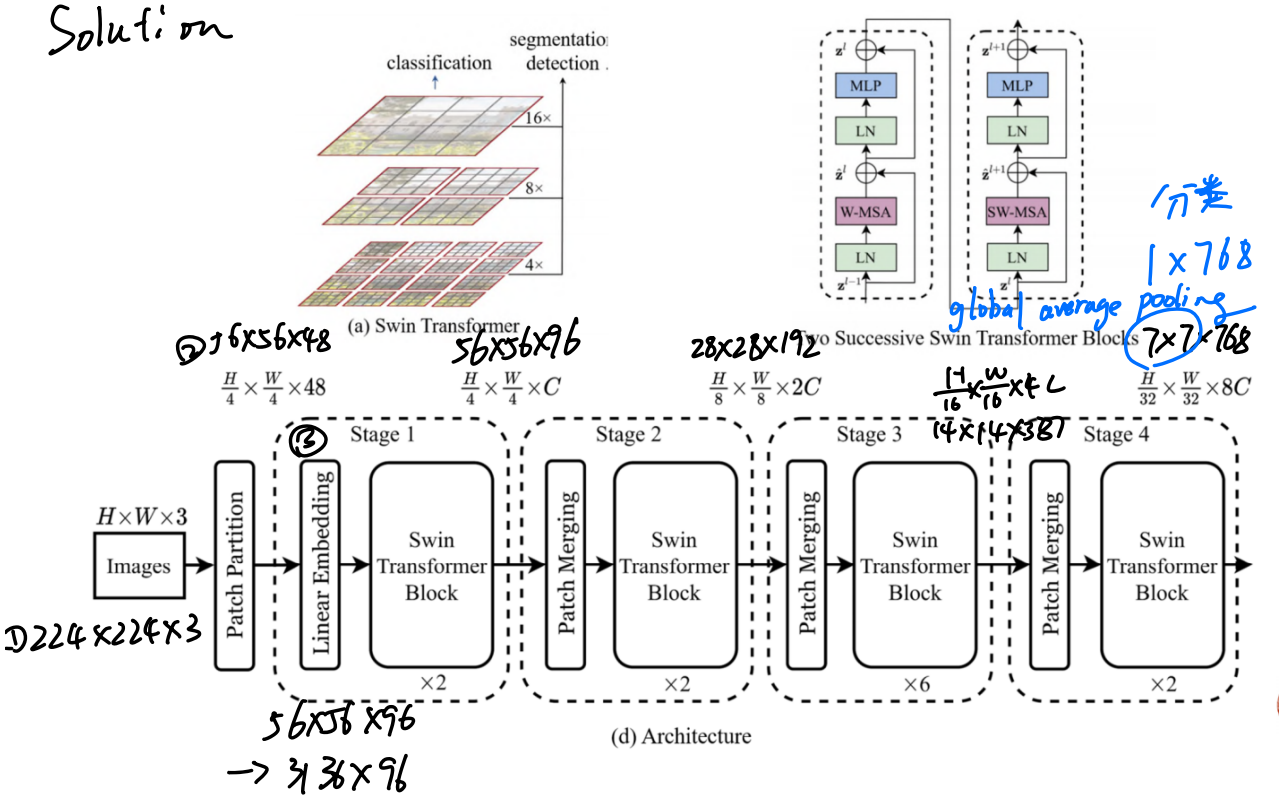
Example 5.1.3.2 Swin Transform
Q understand

Solution



224x224x3

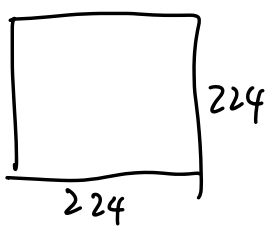
① 56x56x48 $\frac{H}{4}\times\frac{W}{4}\times 48$ → 56x56x96 $\frac{H}{4}\times\frac{W}{4}\times C$ → 28x28x192 $\frac{H}{8}\times\frac{W}{8}\times 2C$ → $\frac{H}{16}\times\frac{W}{16}\times 4C$ → 14x14x384 → 7x7x768 $\frac{H}{32}\times\frac{W}{32}\times 8C$

global average pooling

1行 1x768

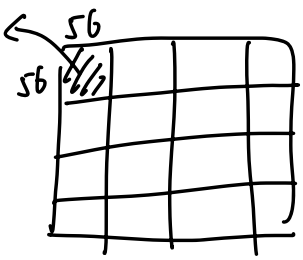
56x56x96 → 3136x96

① 輸入圖



H 224
W 224
RGB 3

② patch 56



類VIT

$\frac{H}{4}$: 56 pixel

$\frac{W}{4}$: 56 pixel

$\frac{H}{4}\times\frac{W}{4}\times 48$ (1 patch pixel)

48 = 4x4x3 (4 patch, 4 patch, RGB)

③ 56x56x96 → 3136x96

C=96，transform 要求 96 超参数

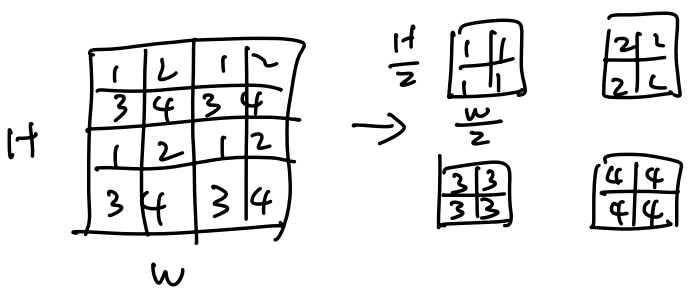
56x56=3136

④ transformer

W-MSA 窗口自注意力 —shift→ SW-MSA

移动窗口自注意力

⑤ HxWxC



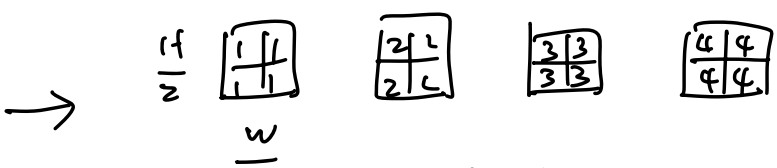
左边HW是一个patch

然后通过merging后变为右图的4个小patch

56x56x96变为28x28x384

然后通过2个1x1的卷积，变为一半的通道数

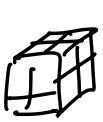
所以就是28x28x192



$\frac{H}{2} \times \frac{W}{2} \times 4C$

1x1卷积

减少一半通道



$1 \times 1 \times C_{in} \times C_{out}$

4 2

$W_1 = [w_1^{(1)}\ w_2^{(1)}\ w_3^{(1)}\ w_4^{(1)}]$

$W_2 = [w_1^{(2)}\ w_2^{(2)}\ w_3^{(2)}\ w_4^{(2)}]$

$\frac{H}{2} \times \frac{W}{2} \times 2C$

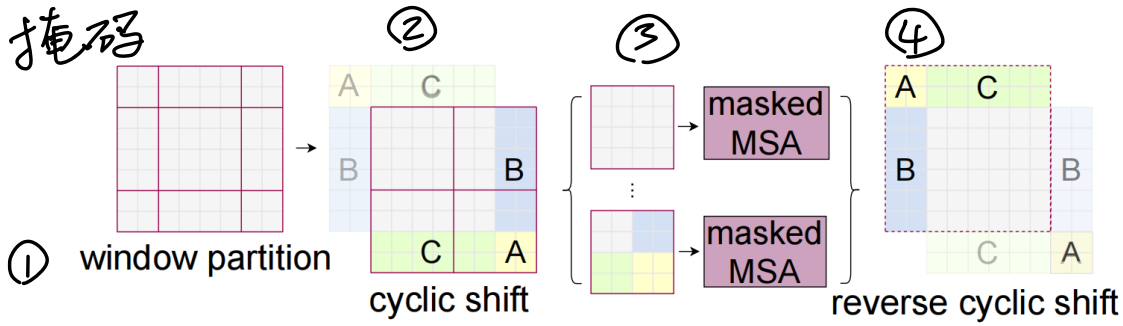
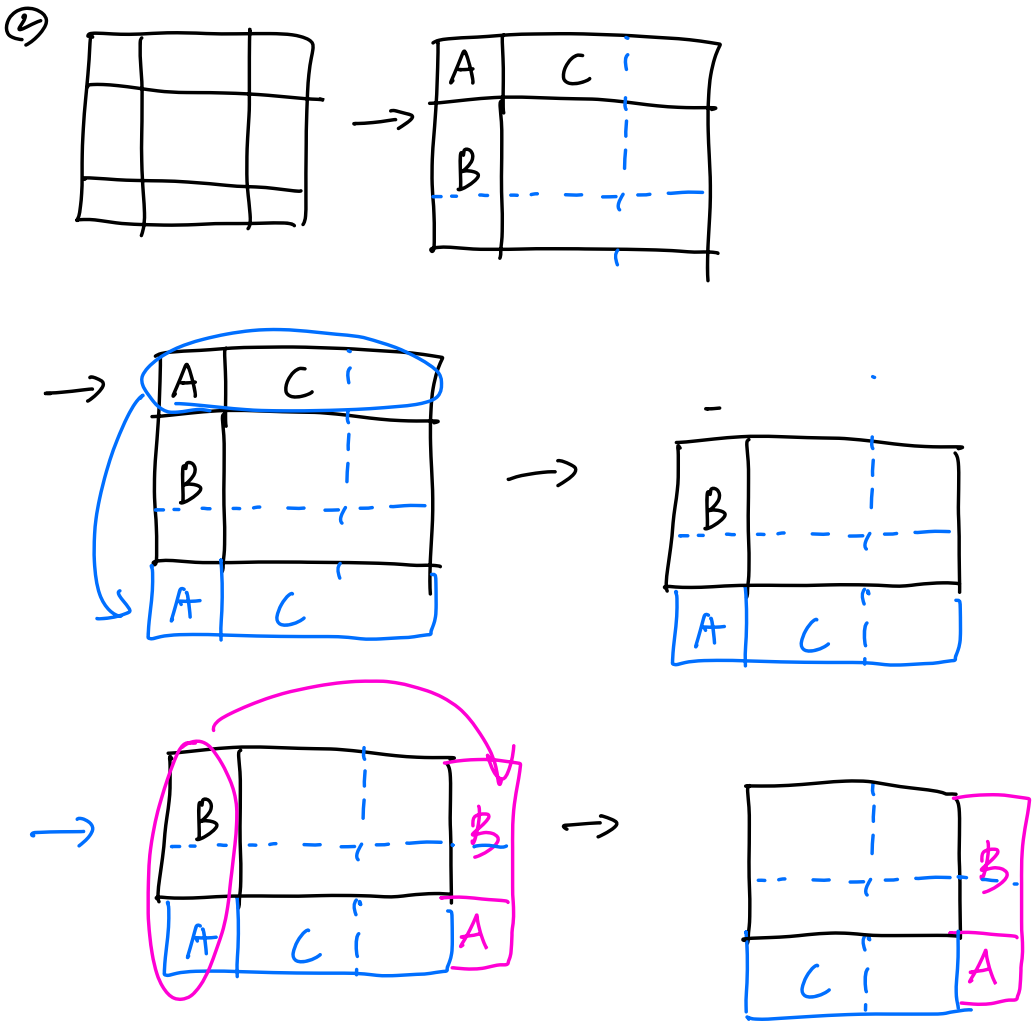
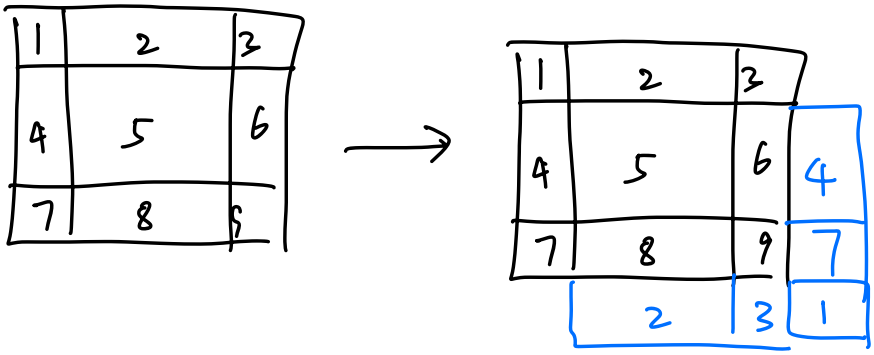


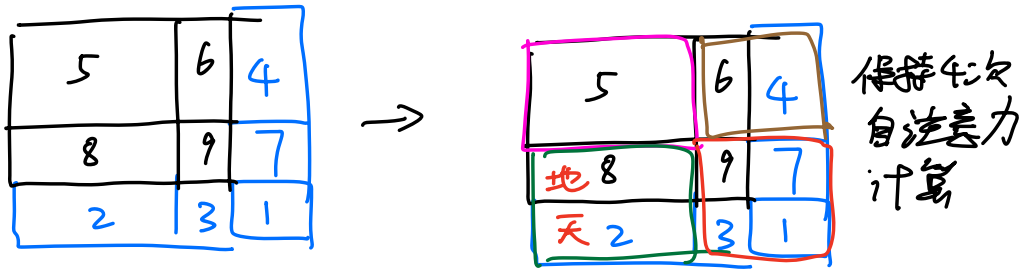
Figure 4. Illustration of an efficient batch computation approach for self-attention in shifted window partitioning.

①目标算9个窗口自注意力，但只算4次



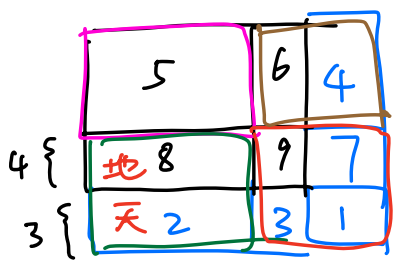
All in All

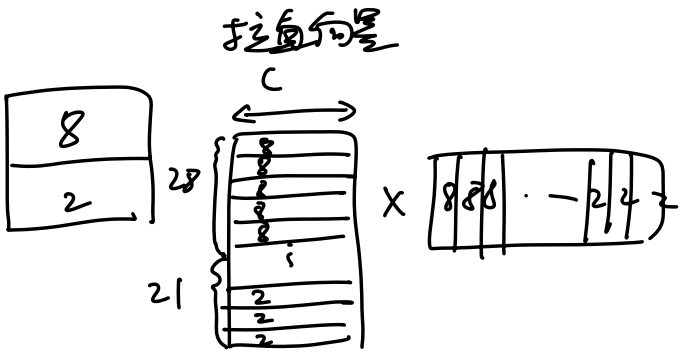


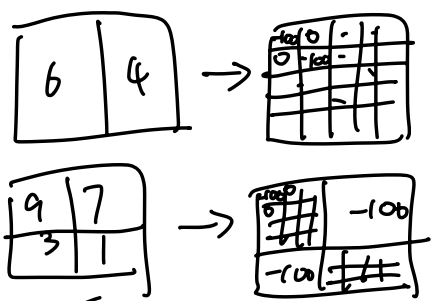


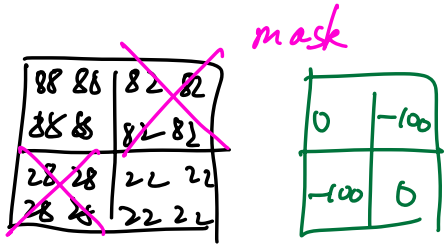
③ 掩码 Solve → 问题：天地不应算自注意力

Example 14 x 14 size









④ 还原