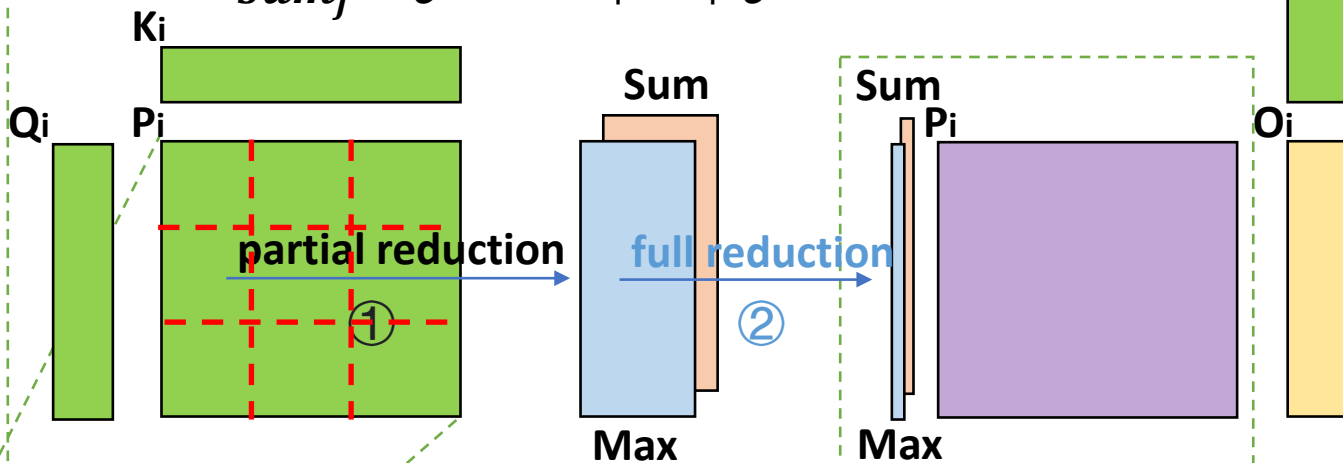


i^{th} problem of of fused multihead attention

$$\textcircled{1} \max_j = \max(x_0, \dots, x_n)$$

$$\text{sum}_j = e^{x_0 - \max_j} + \dots + e^{x_n - \max_j}$$



$\textcircled{3}$ fused element-wise ops

$$\textcircled{2} \max = \max(\max_0, \dots, \max_n)$$

$$\text{sum} = \sum_j \text{sum}_j * e^{\max_j - \max}$$

$$\textcircled{3} \text{softmax}_i = \frac{e^{x_i - \max}}{\text{sum}}$$

of problems = batch sz * head num