Week 20

Multihead Attention

Step 1: Compute Q, K, V

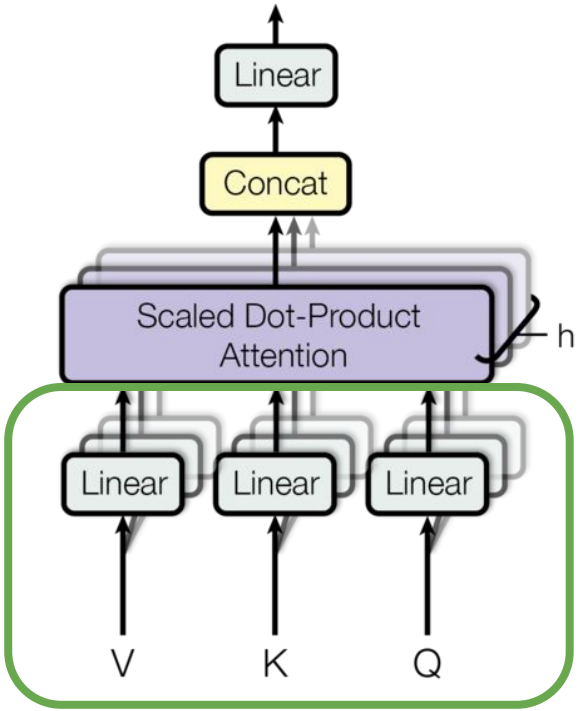
for i in #heads:

Q[i] = x * Weight_Q[i]

K[i] = x * Weight_K[i]

V[i] = x * Weight_V[i]

x (64 x 768), Weight_Q/K/V (768 x 64)



Multihead Attention

Step 1: Compute Q, K, V

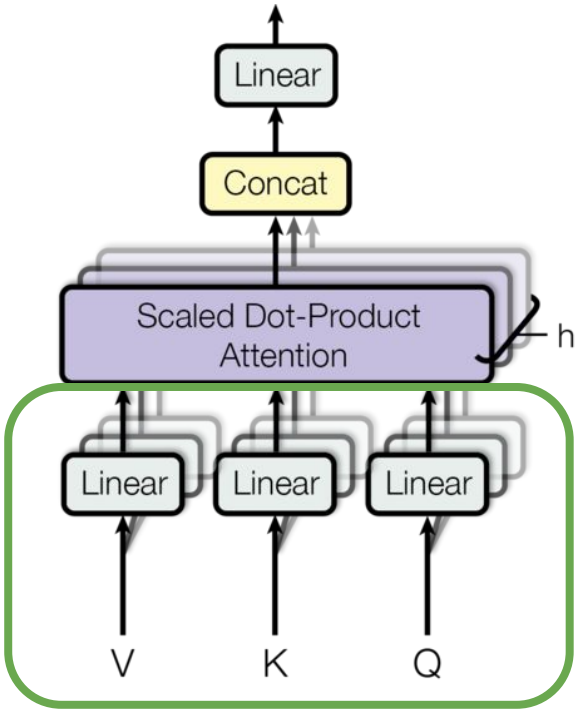
for i in #heads:

Q[i] = x * Weight_Q[i]

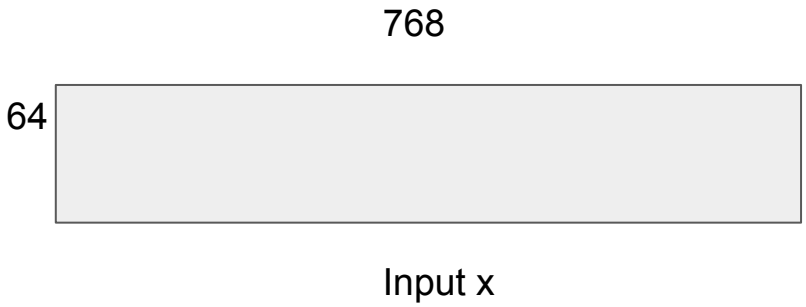
K[i] = x * Weight_K[i]

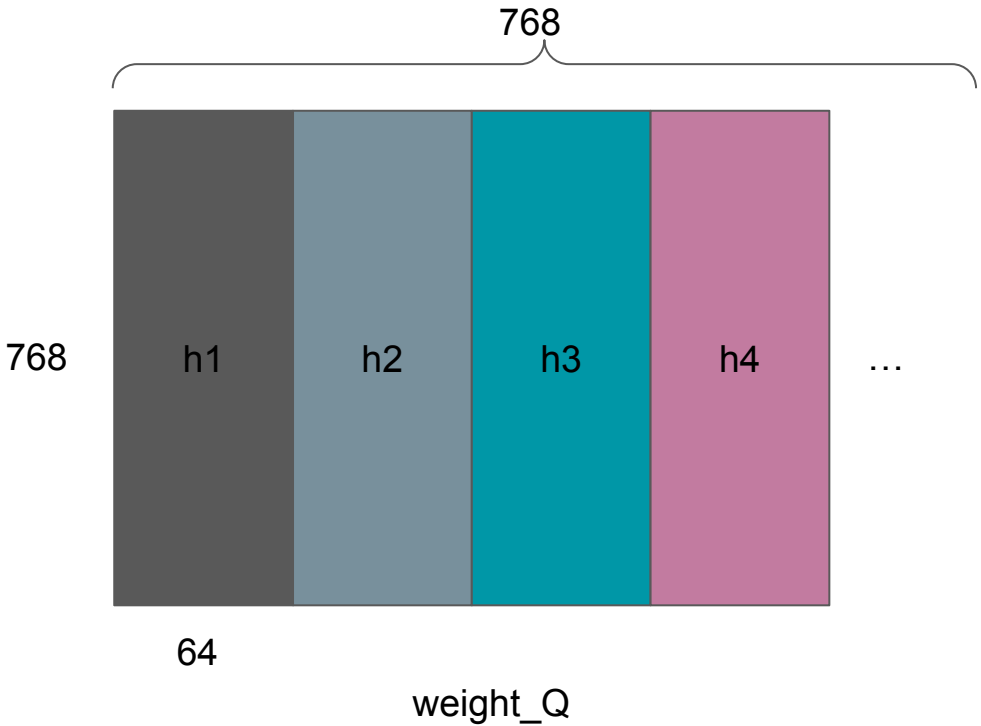
V[i] = x * Weight_V[i]

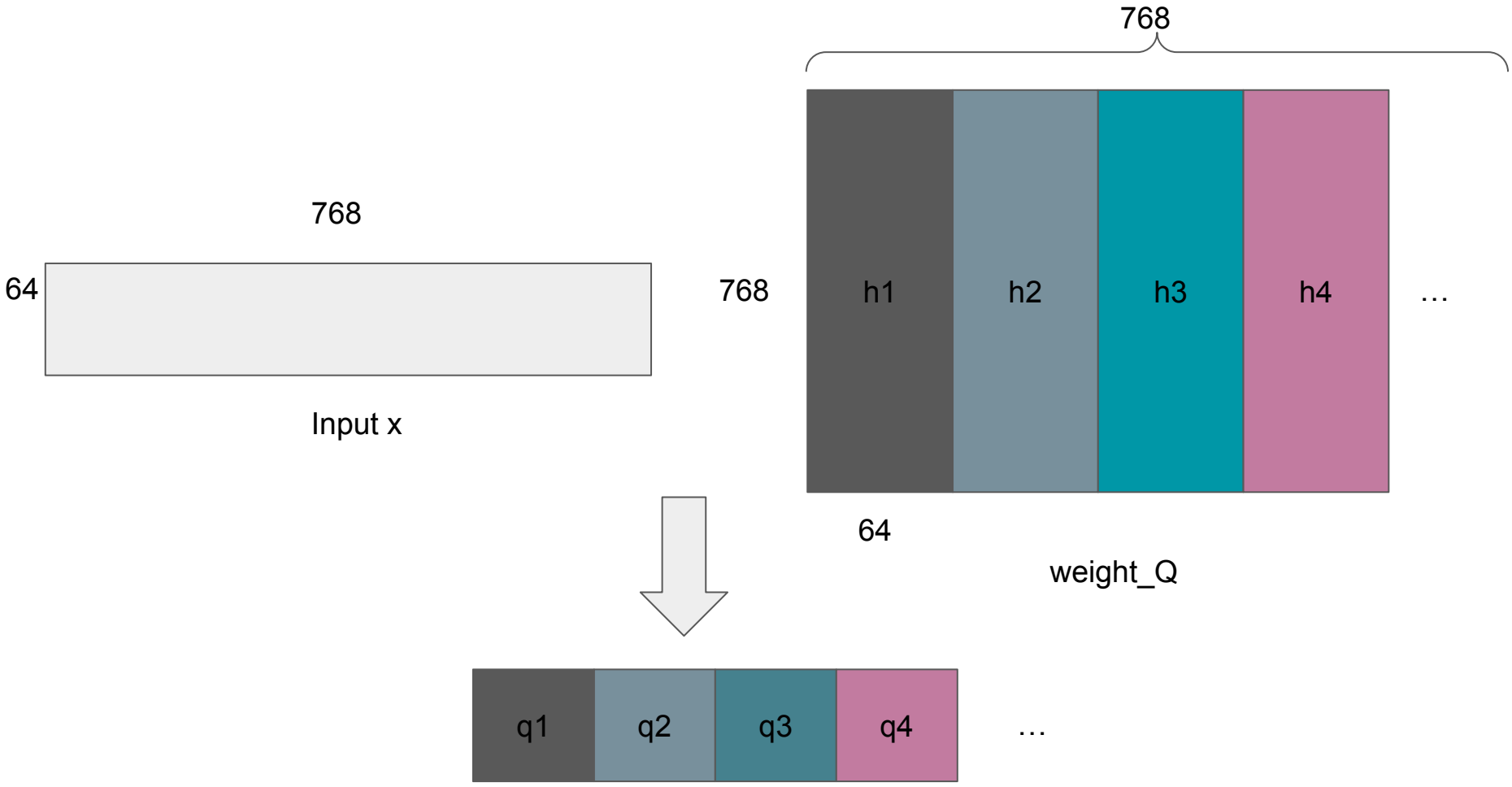
x (64 x 768), Weight_Q/K/V (768 x 64)



Inefficient (17% 16_lanes)


768
64
Input x


768
768
h1
h2
h3
h4
…
64
weight_Q


768
64
768
h1
h2
h3
h4
…
Input x
64
weight_Q
q1
q2
q3
q4
…

MatMul 64x768 * 768x768

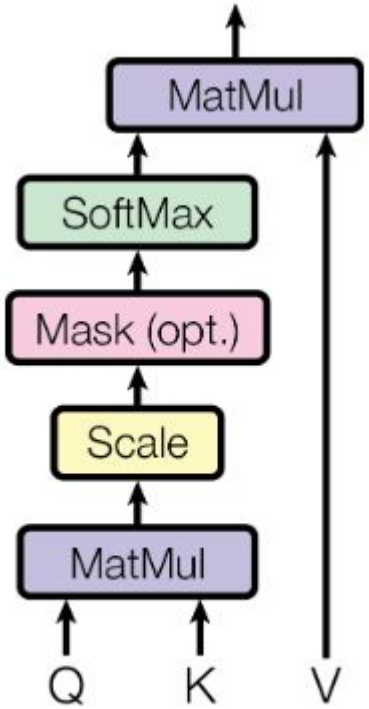
	bc_matmul	vanilla_matmul
4_lanes	97%	94%
16_lanes	93%	81.8%

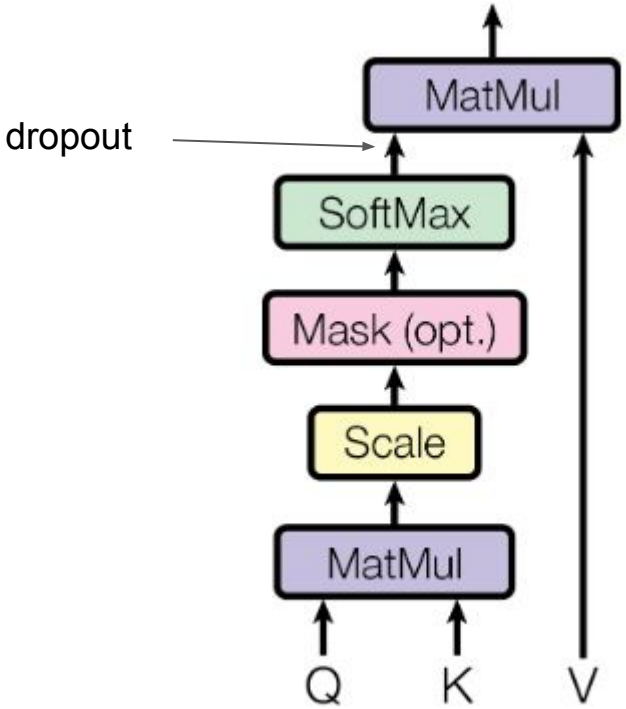
MatMul 64x768 * 768x768

	bc_matmul	vanilla_matmul
4_lanes	97%	94%
16_lanes	93%	81.8%

bc_matmul maybe more energy efficient

Need to add dropout layer





PD

- Setup: WNS 58.27ps, bc related 3.9ps
- Hold: WNS 3.28ps
- Cell area utilization: 80.3%