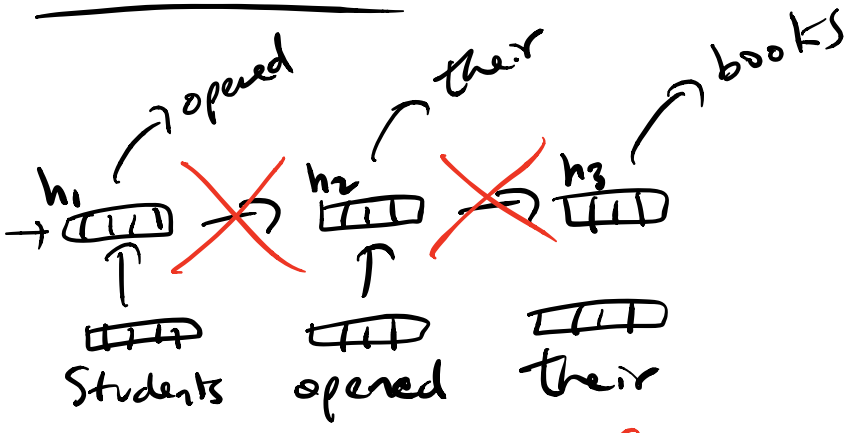
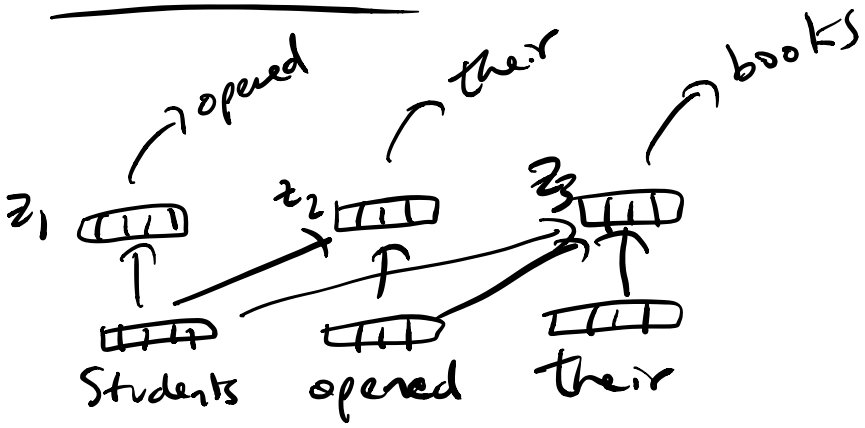
Self-attention:



if i can get rid of recurrence at training, i can compute all h_i in parallel

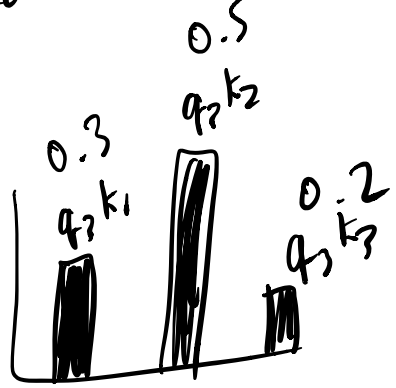
goal:



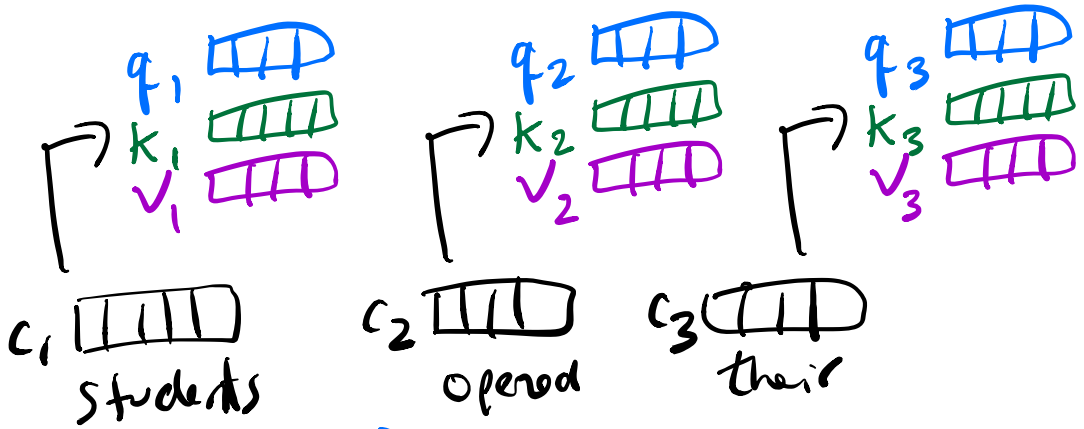
$z_3 = 0.3 \cdot v_1 + 0.5 \cdot v_2 + 0.2 \cdot v_3$

$=$  → predict books

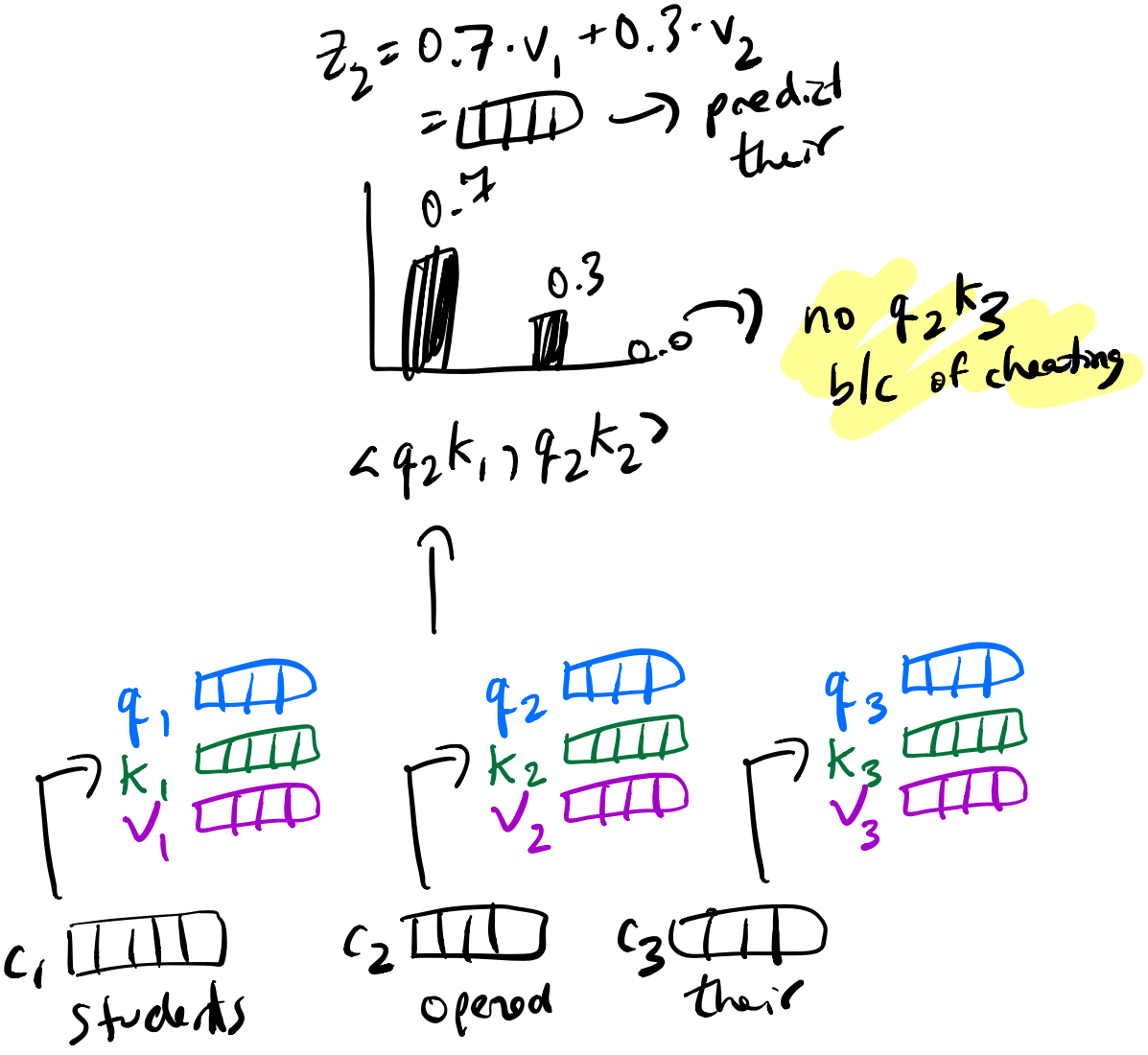
Use attn scores to compute weighted ave. of value vectors



$\text{Softmax}(q_3 \cdot k_1, q_3 \cdot k_2, q_3 \cdot k_3)$



$q_1 = f(W_q c_1)$ → W_k, W_v


$z_2 = 0.7 \cdot v_1 + 0.3 \cdot v_2$
= → predict their
0.7
0.3
0.0
no $q_2 k_3$ b/c of cheating
$\langle q_2 k_1, q_2 k_2 \rangle$
q_1 k_1 v_1
q_2 k_2 v_2
q_3 k_3 v_3
c_1 students
c_2 opened
c_3 their

no dependency between z_1, z_2, z_3

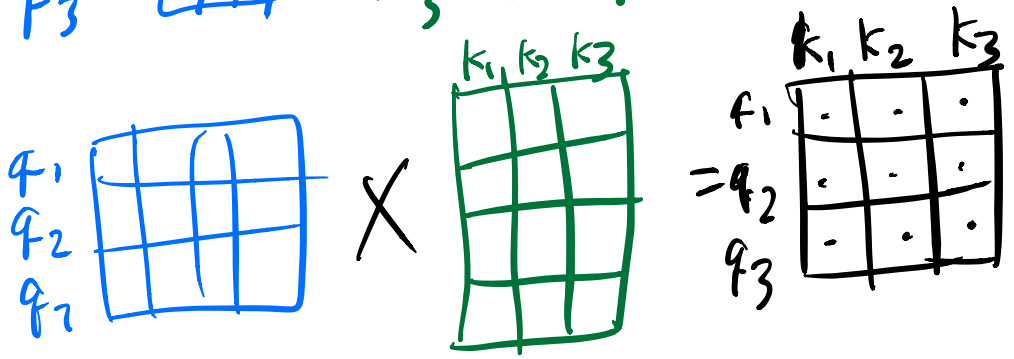
how can we compute z_i in parallel?

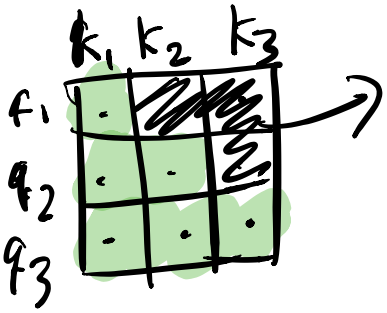
attn scores

q_1 k_1 $a_1 = \langle q_1, k_1 \rangle$

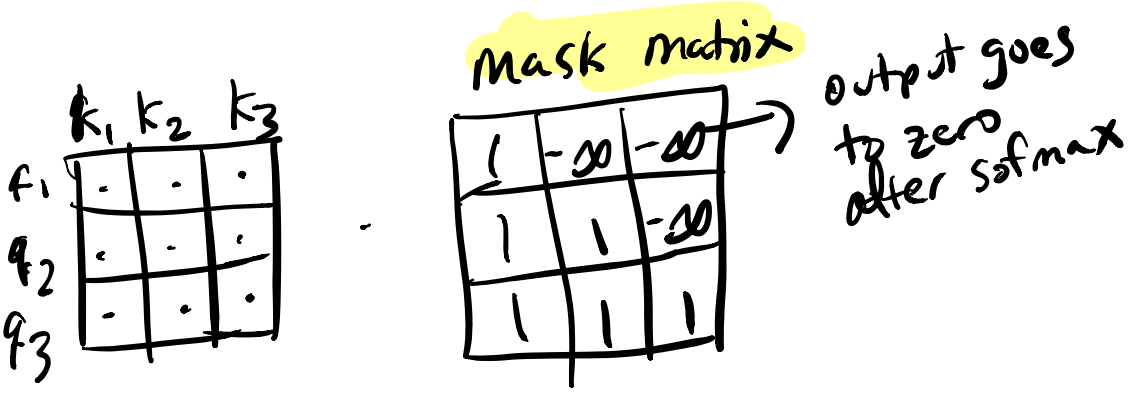
q_2 k_2 $a_2 = \langle q_2 k_1, q_2 k_2 \rangle$

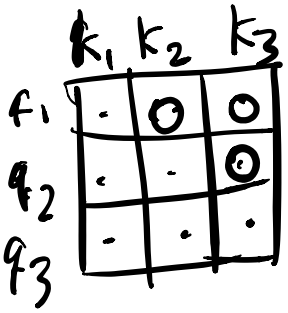
q_3 k_3 $a_3 = \langle q_3 k_1, q_3 k_2, q_3 k_3 \rangle$





info about future words, need to exclude them


Mask matrix
output goes to zero after softmax
k1 k2 k3
q1
q2
q3
1 -∞ -∞
1 1 -∞
1 1 1


k1 k2 k3
q1
q2
q3