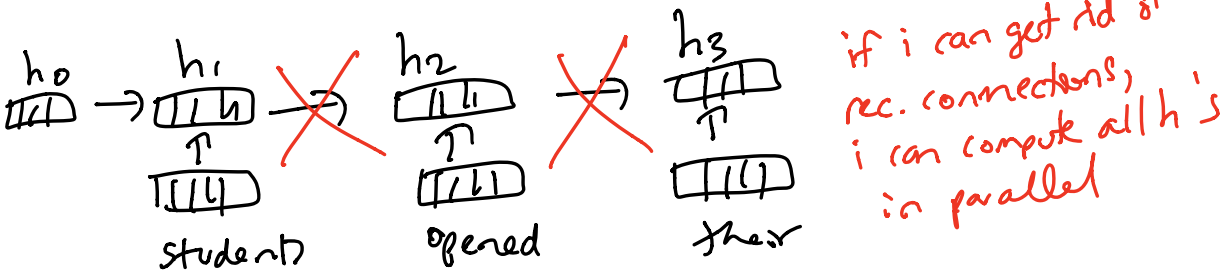
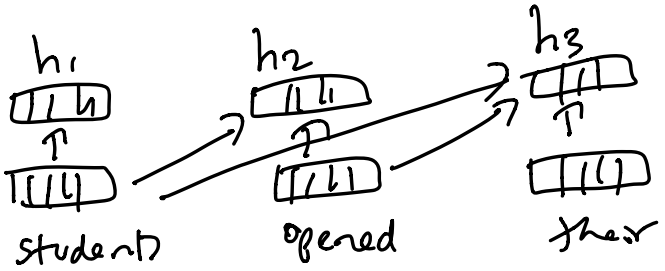
Self-attention:

motivation: get rid of recurrent computation



if i can get rid of rec. connections, i can compute all h's in parallel

goal:

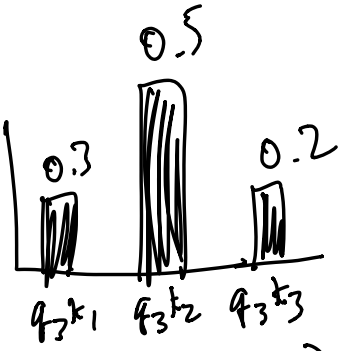


Self-attn @ 3rd pos. of say

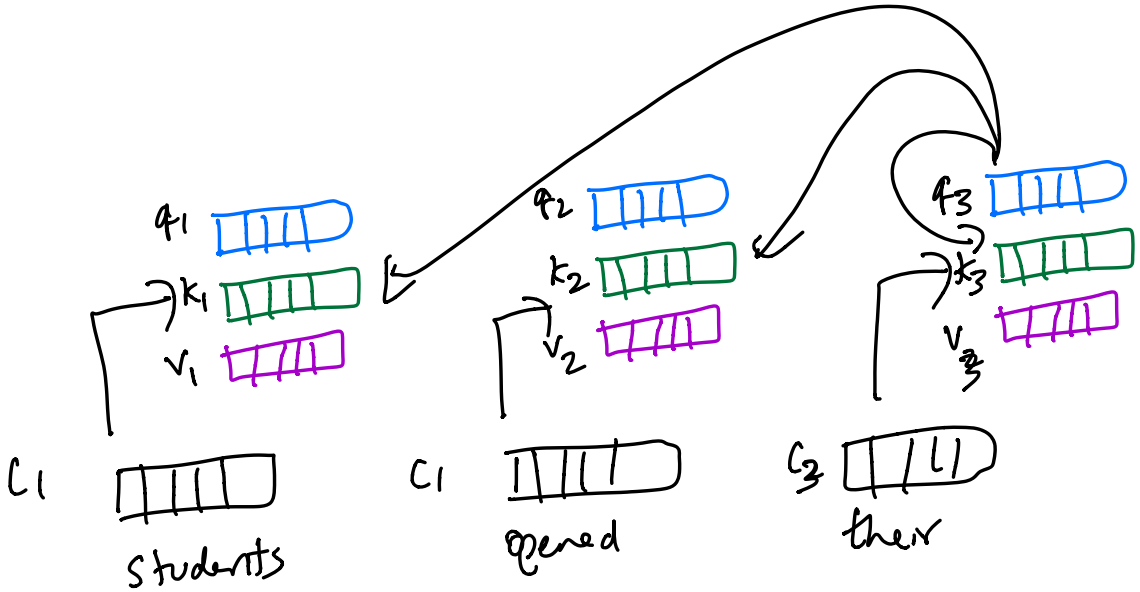
$h_3 = 0.3 \cdot v_1 + 0.5 \cdot v_2 + 0.2 \cdot v_3$

$= $ 

↳ predict books



$\text{softmax}(\langle q_3 k_1, q_3 k_2, q_3 k_3 \rangle)$

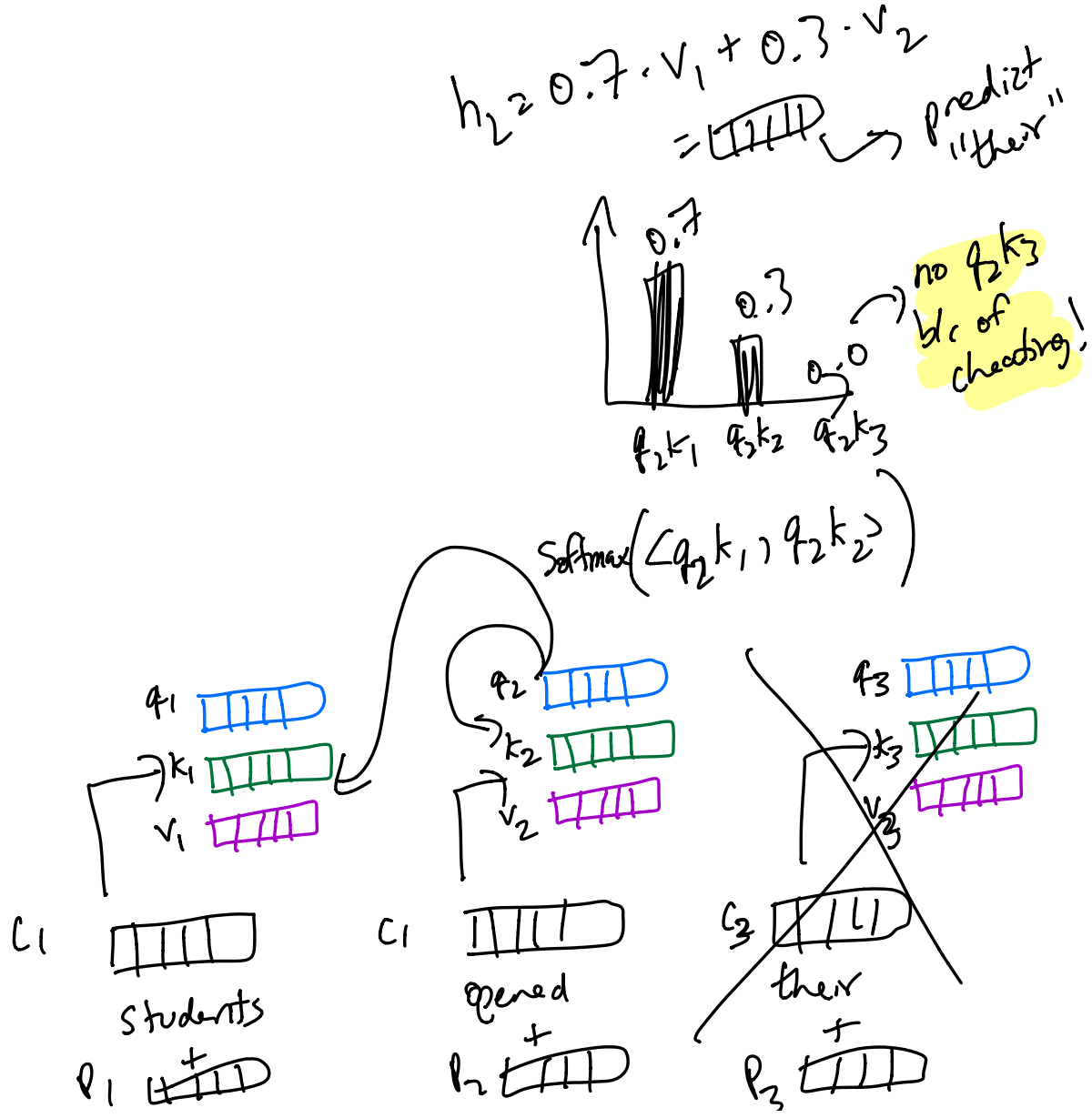


$q_1 = f(W_q c_1)$

$v_1 = f(W_v c_1)$

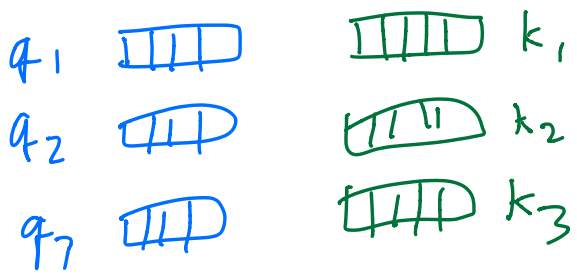
$k_1 = f(W_k c_1)$

Self-attn
@ 2nd pos of seq



Now, there's no dependencies between h_1, h_2, h_3

how can we compute h_i's in parallel?

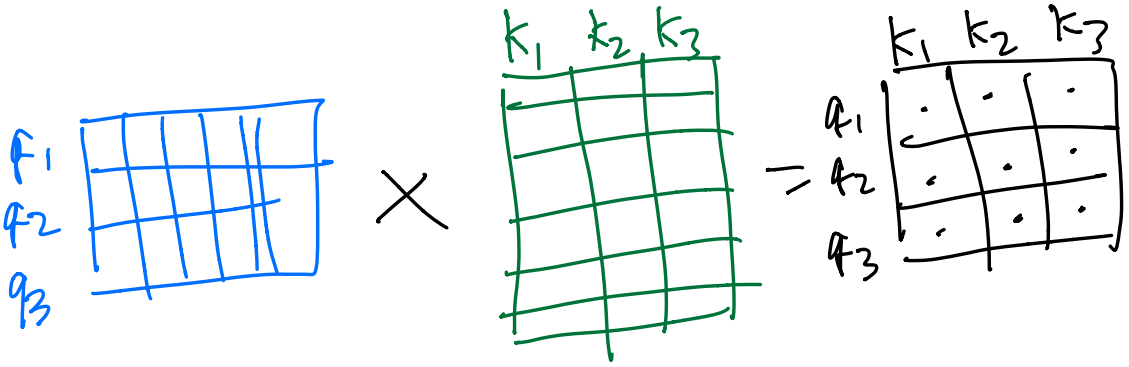


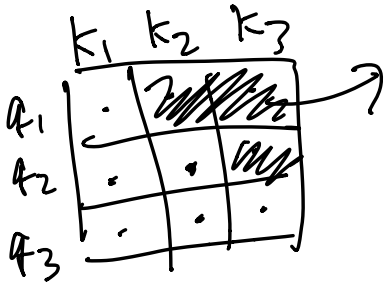
attn scores

$a_1 = \langle q_1 k_1 \rangle$

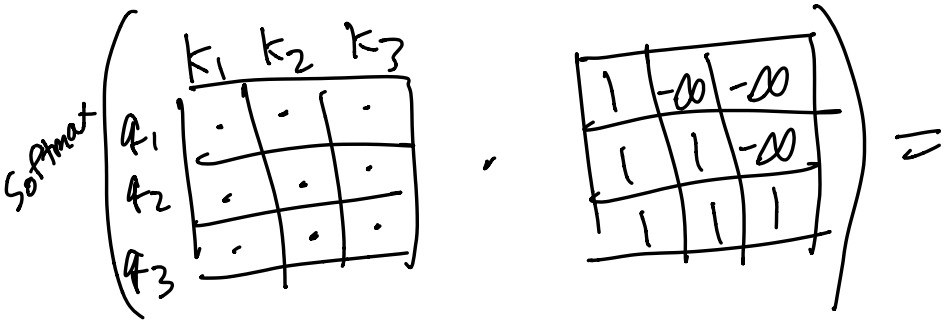
$a_2 = \langle q_2 k_1, q_2 k_2 \rangle$

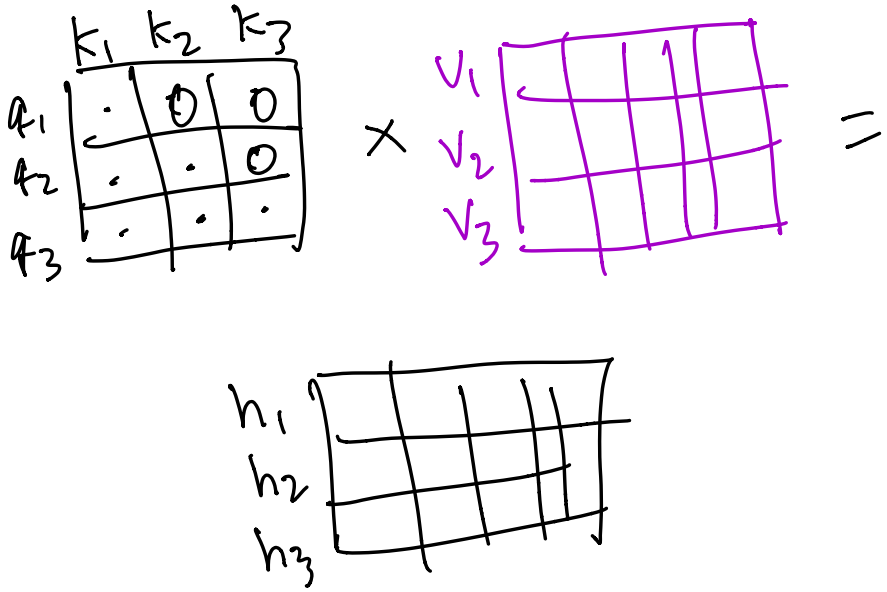
$a_3 = \langle q_3 k_1, q_3 k_2, q_3 k_3 \rangle$





these cells contain info about future words, need to exclude!




k1 k2 k3
q1
q2
q3
0 0
0
v1
v2
v3
=
h1
h2
h3