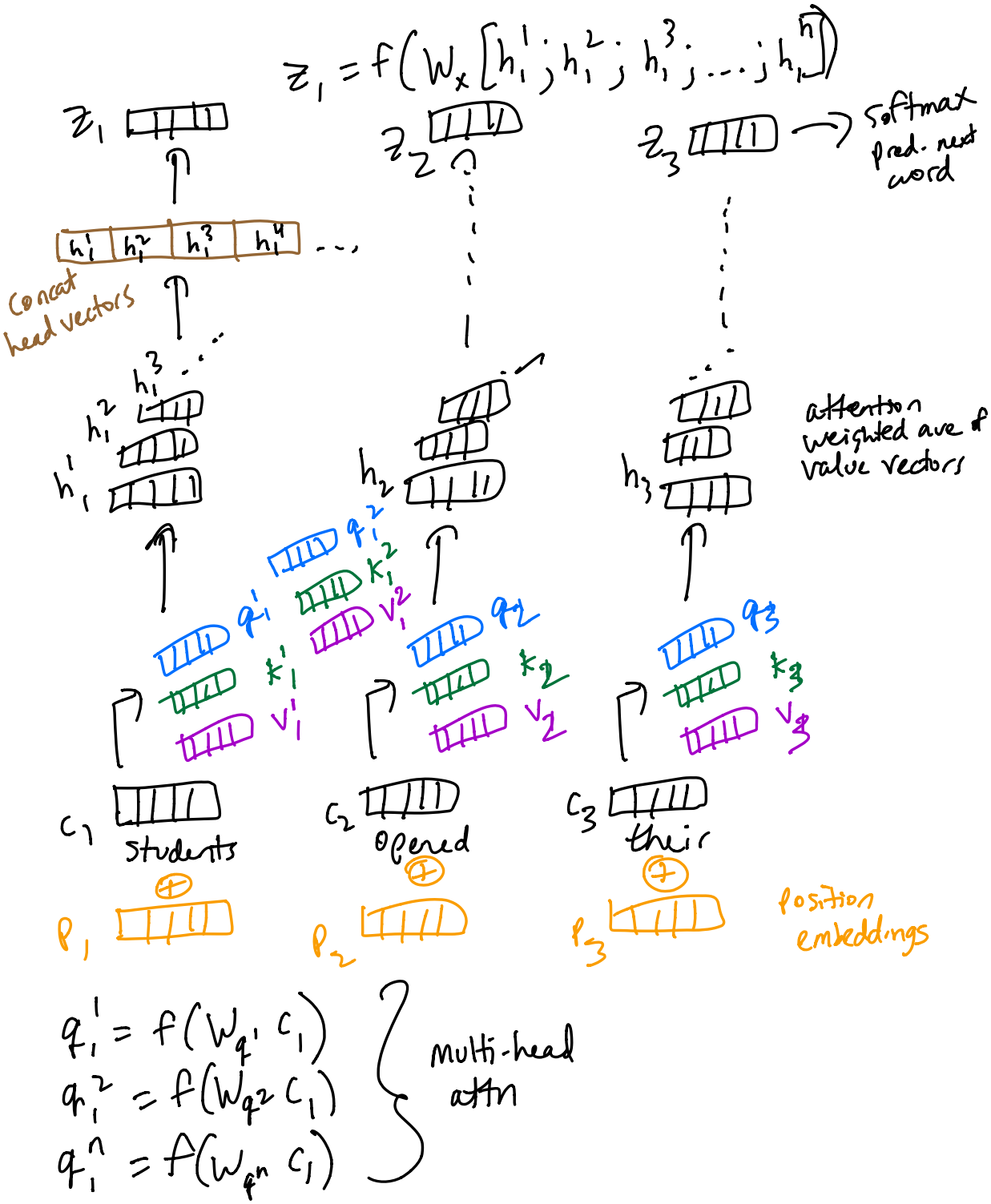
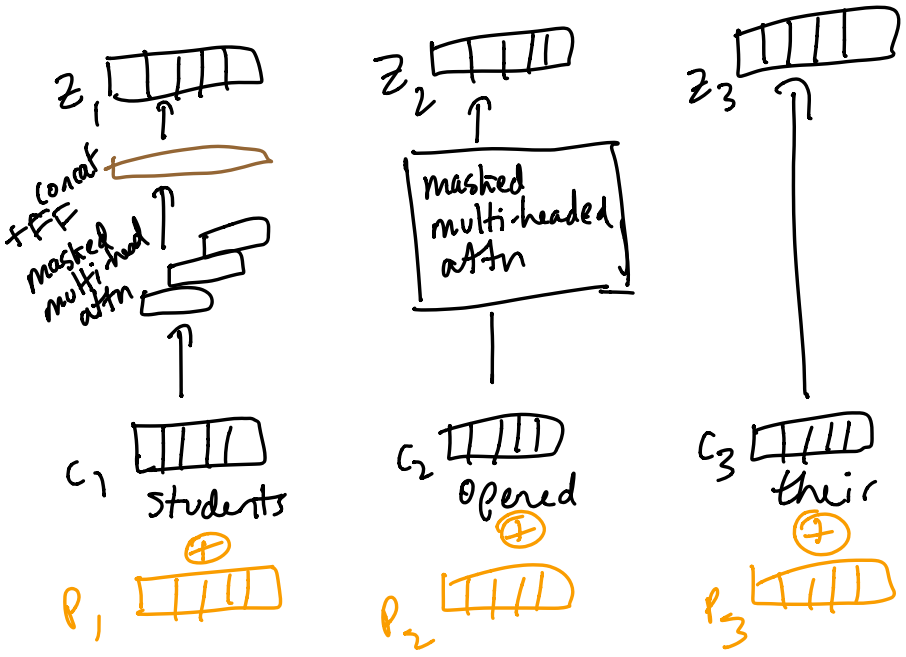
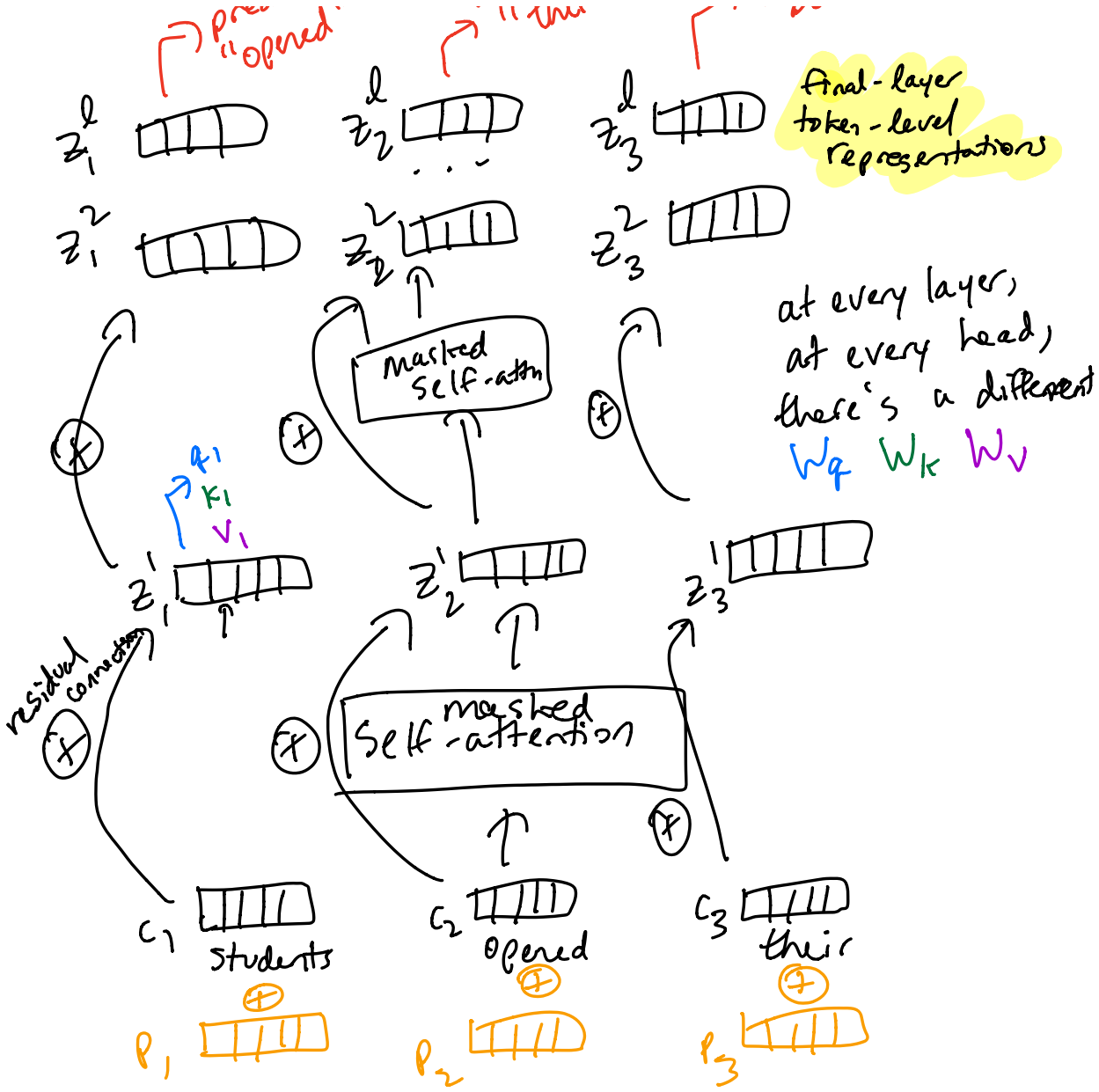

$z_1 = f(W_x[h_1^1; h_1^2; h_1^3; \ldots; h_1^n])$
z_1
z_2
z_3
softmax pred. next word
h_1^1 h_1^2 h_1^3 h_1^4 ...
concat head vectors
h_1^3
h_1^2
h_1^1
h_2
h_3
attention weighted ave of value vectors
q_1^2 k_1^2 v_1^2
q_1^1 k_1^1 v_1^1
q_2 k_2 v_2
q_3 k_3 v_3
c_1 students
c_2 opened
c_3 their
p_1
p_2
p_3
position embeddings
$q_1^1 = f(W_{q^1} c_1)$
$q_1^2 = f(W_{q^2} c_1)$
$q_1^n = f(W_{q^n} c_1)$
multi-head attn

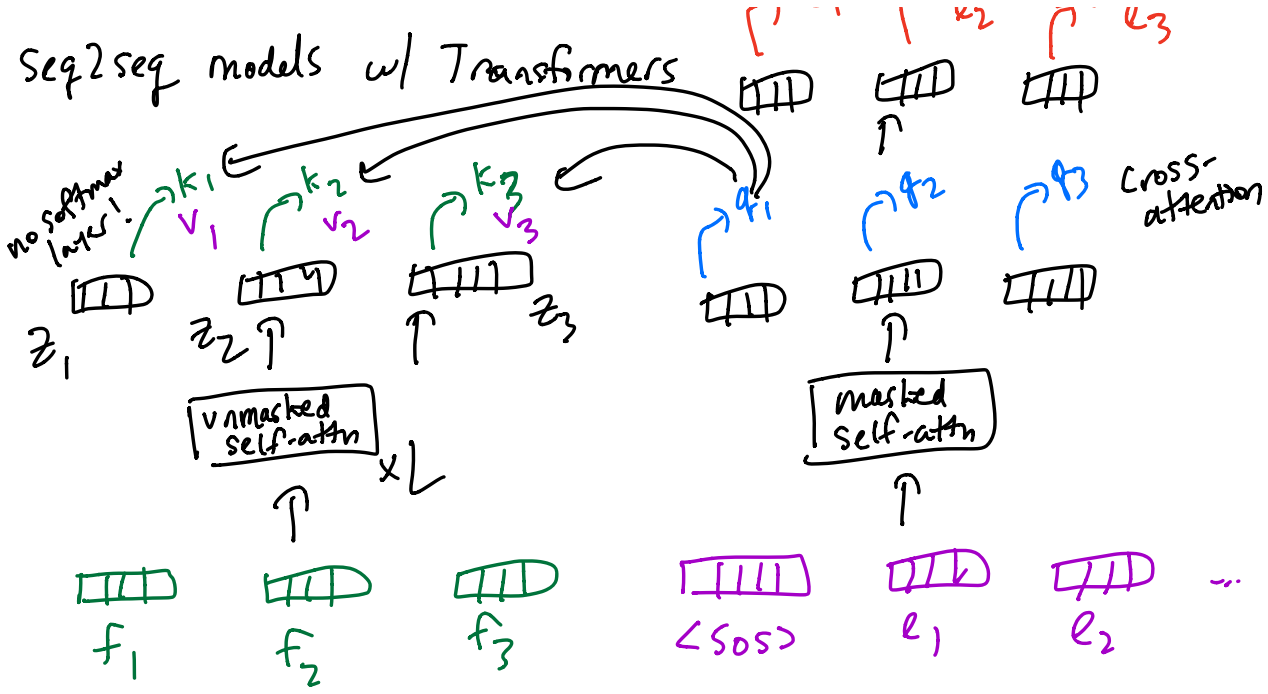
adding depth to Transformer





predict predict predict

Seq2seq models w/ Transformers



encoder
(encode french text)

decoder
(generating english sent)

in a decoder block, we perform both masked self-attn and cross attention