MCB128
Quiz b3
March 13, 2026

Name: _____   **Score:** _____

Instructions:

- Show all work for full credit.

- Can check notes; no prompts to chatbot allowed

1. (**3 pts**) For a system of L inputs of length D, producing L outputs of length D,

   (a) (**1 pts**) In a self-attention mechanism, calculate the number of weights and biases used to compute queries, keys, and values.

   (b) (**1 pts**) Calculate the number of weights and biases to go from the same (L,D) inputs to the same (L,D) outputs by a fully connected single layer (by flattening).

   (c) (**1 pts**) Compare the two outcomes with respect to the size of the input L.
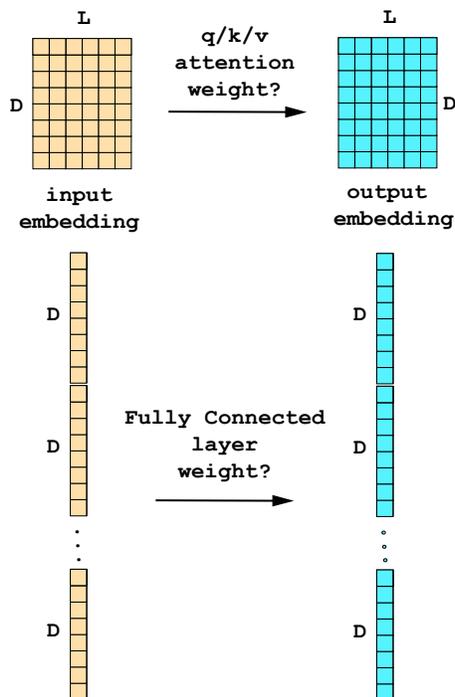
(a)

$$queries : W^Q[D, d_k] \quad b^Q[d_k]$$
$$keys : W^K[D, d_k] \quad b^K[d_k]$$
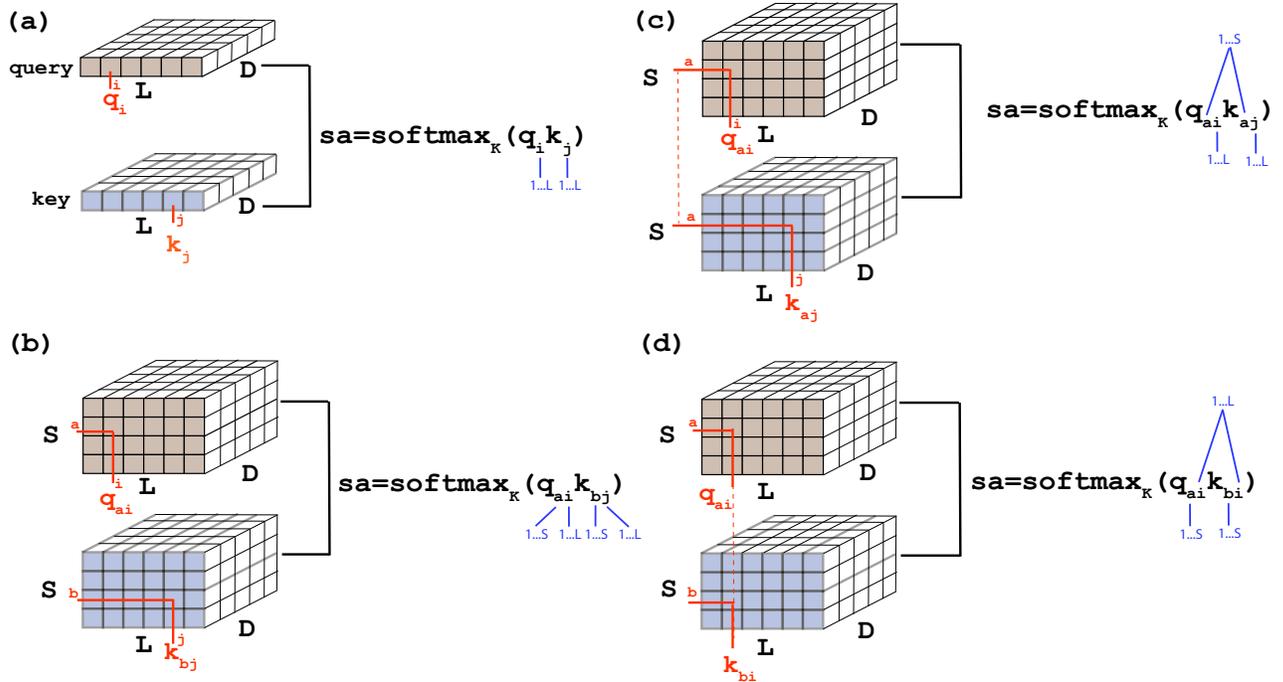$$values : W^V[D, d_v] \quad b^V[d_v]$$

Total number of parameters $2(D \times d_k + d_k) + (D \times d_v + d_v)$
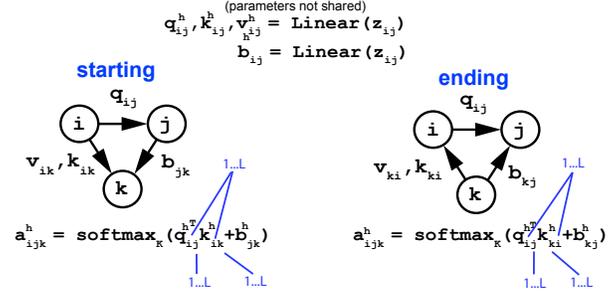If $d_k = d_v = D$, it becomes $3(D^2 + D)$.

(b) $inputs[L, D] \longrightarrow outputs[L, D]$,

thus weights have dimensions $(L \times D) \times (L \times D) = L^2 \times D^2$.

(c) For the FCL, the parameters depend on input length L, which forces all inputs to have the same length, and it scales with L^2 which can be quite costly in memory requirements.

(a)

query $\quad$ L $\quad$ D

q_i $\quad$ L

key $\quad$ L $\quad$ D

k_j

$$sa = softmax_K(q_i k_j)$$

1...L 1...L

(c)

S $\quad$ a $\quad$ L $\quad$ D

q_{ai} $\quad$ L

S $\quad$ a $\quad$ L $\quad$ D

k_{aj}

$$sa = softmax_K(q_{ai} k_{aj})$$

1...S

1...L 1...L

(b)

S $\quad$ a $\quad$ L $\quad$ D

q_{ai} $\quad$ L

S $\quad$ b $\quad$ L $\quad$ D

k_{bj}

$$sa = softmax_K(q_{ai} k_{bj})$$

1...S 1...L 1...S 1...L

(d)

S $\quad$ a $\quad$ L $\quad$ D

q_{ai} $\quad$ L

S $\quad$ b $\quad$ L $\quad$ D

k_{bi}

$$sa = softmax_K(q_{ai} k_{bi})$$

1...L

1...S 1...S

(e) triangle self-attentions

(parameters not shared)

$$q_{ij}^h, k_{ij}^h, v_{ij}^h = Linear(z_{ij})$$
$$b_{ij} = Linear(z_{ij})$$

starting

q_{ij}

i → j

v_{ik}, k_{ik} $\quad$ b_{jk} $\quad$ 1...L

k

$$a_{ijk}^h = softmax_K(q_{ij}^h k_{ik}^h + b_{jk}^h)$$

1...L $\quad$ 1...L

ending

q_{ij}

i → j

v_{ki}, k_{ki} $\quad$ b_{kj} $\quad$ 1...L

k

$$a_{ijk}^h = softmax_K(q_{ij}^h k_{ki}^h + b_{kj}^h)$$

1...L $\quad$ 1...L

2. (**7 pts**) Calculate the size of the self-attention maps for each of the following cases. Outputs have the same dimensions than the inputs.

(a) (**1 pts**) A sequence embedding $x]L, D]$
$sa[L, L]$

(b) (**2 pts**) An alignment embedding $m[L, S, D]$
$sa[L, S, L, S]$ with general self-attention

(c) (**2 pts**) An alignment embedding $m[L, S, D]$ with self-attention by rows
$sa[L, L, S]$

(d) (**2 pts**) An alignment embedding $m[L, S, D]$ with self-attention by columns
$sa[S, S, L]$

(e) (**bonus pts**) AlphaFold2 triangle self-attention on pair representation $z[r, r, c_z]$, (see Figure 12, right)
$sa[r, r, r]$