① Input : (Batch, Sequence length)

② Embedding : (Batch, Sequence length, d_model)

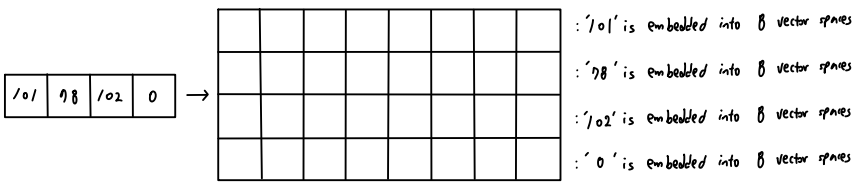
③ Positional Encoding : (Batch, Sequence length, d_model)

Let Batch = 1, Sequence length = 4, d_model = 8

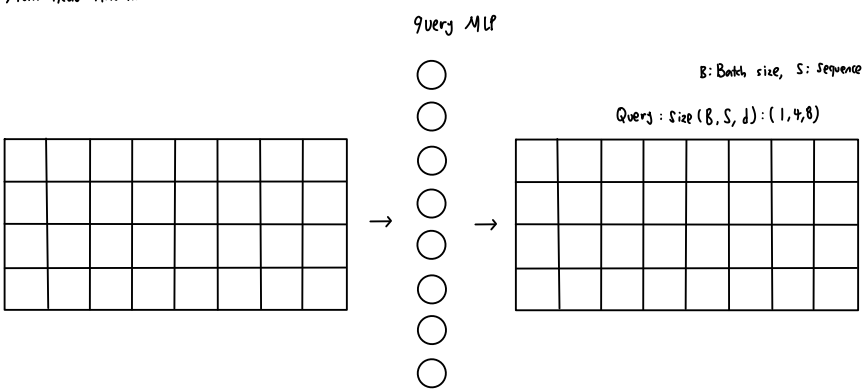
Input size : (1, 4)

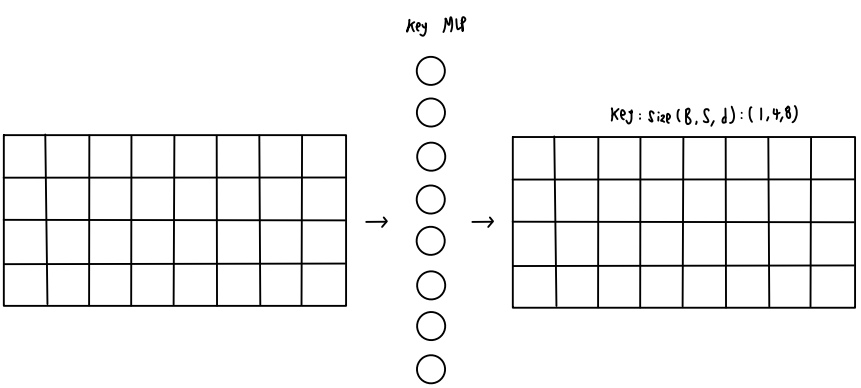
101	98	102	0

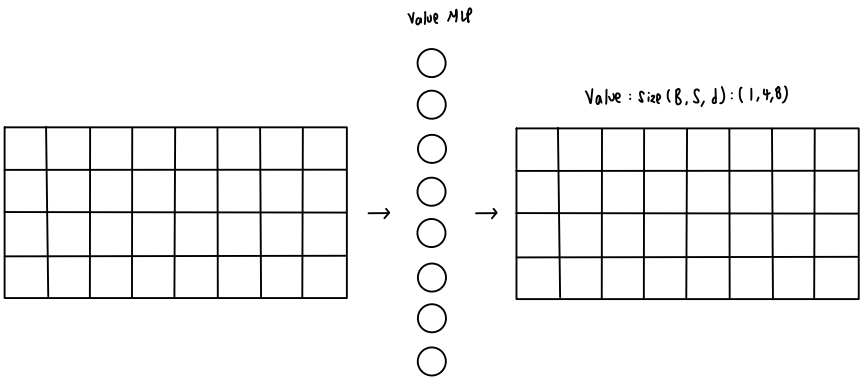
After Input Embedding & Positional Encoding : (1, 4, 8)

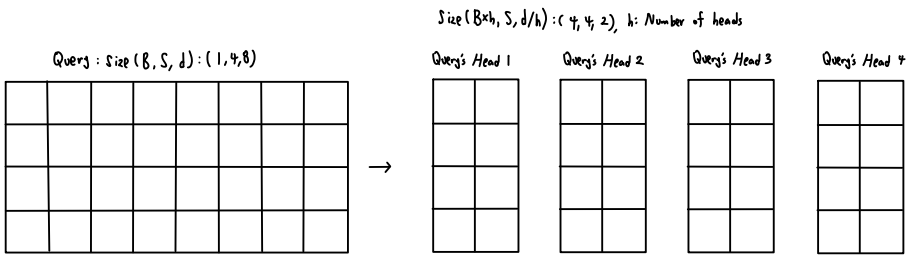


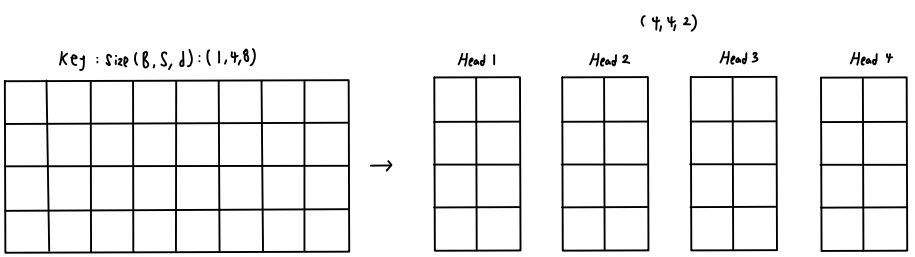
Multi-Head Attention

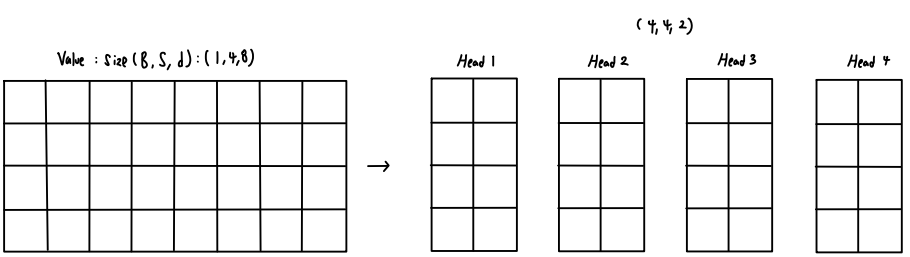




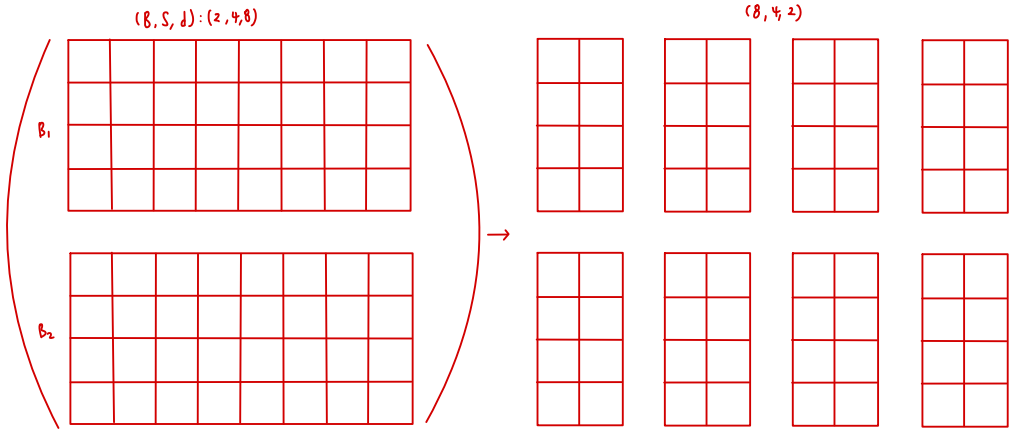

Value MLP
Value : Size (B, S, d) : (1, 4, 8)


Size (Bxh, S, d/h) : (4, 4, 2), h: Number of heads
Query : Size (B, S, d) : (1, 4, 8)
Query's Head 1
Query's Head 2
Query's Head 3
Query's Head 4


(4, 4, 2)
Key : Size (B, S, d) : (1, 4, 8)
Head 1
Head 2
Head 3
Head 4


(4, 4, 2)
Value : Size (B, S, d) : (1, 4, 8)
Head 1
Head 2
Head 3
Head 4

If B=2 ?



Calculating Attention(W) in Multi-Head Attention

@W = Query · Key.T

Query : (4, 4, 2)

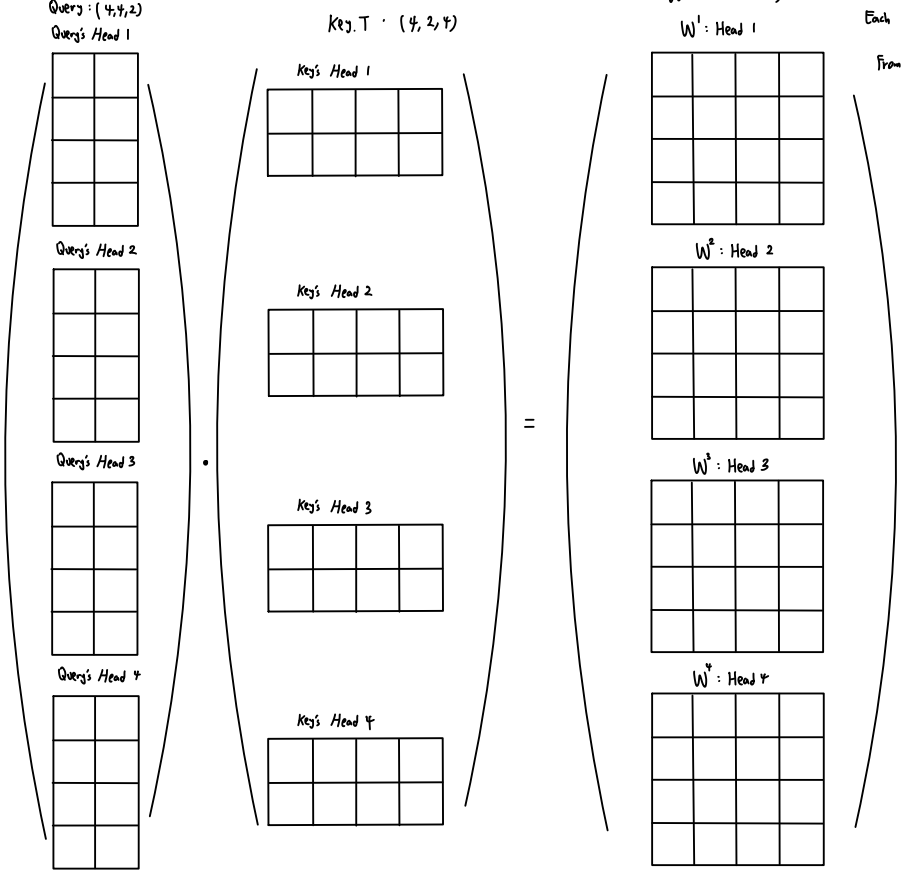
Key.T : (4, 2, 4)

W : (Bxh, S, S) : Attention for each word in sentence.

Each head will have different attention values.

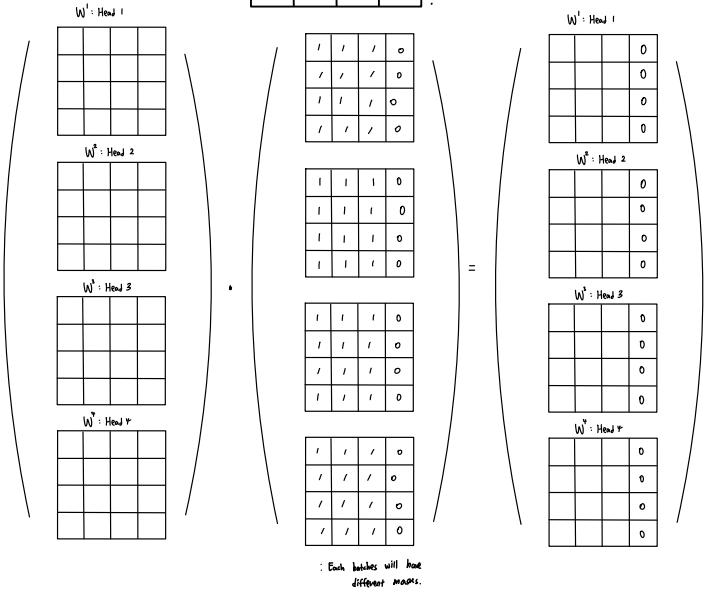
From "Attention is All you Need!",

"We found it beneficial to linearly project the queries, keys, and values h times with different, learned linear projections to d_k, d_q, and d_v"



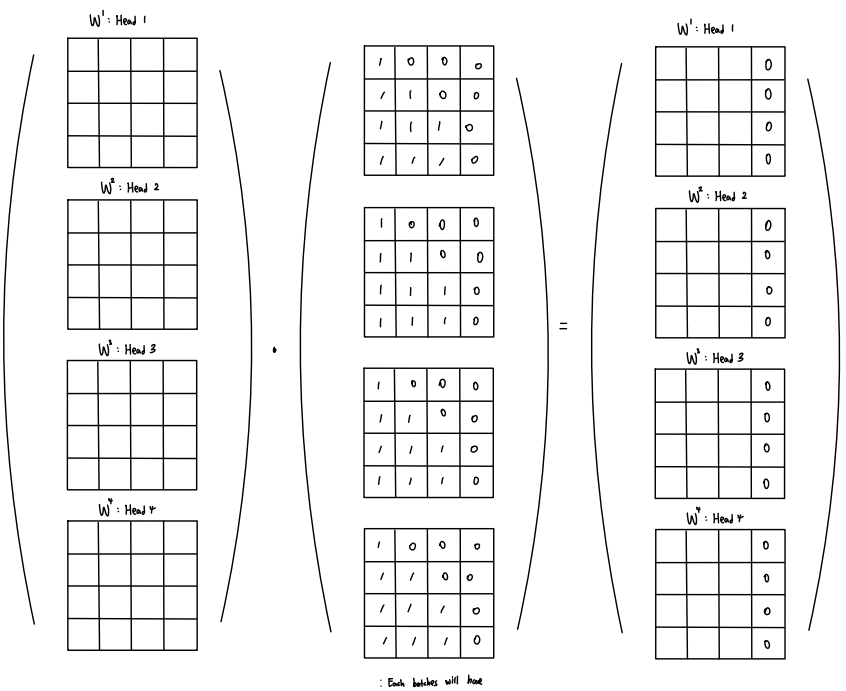
② $W = W/\sqrt{d}$

③ Masking W : Let's say mask : | 1 | 1 | 1 | 0 | . For masking W, we need some size.

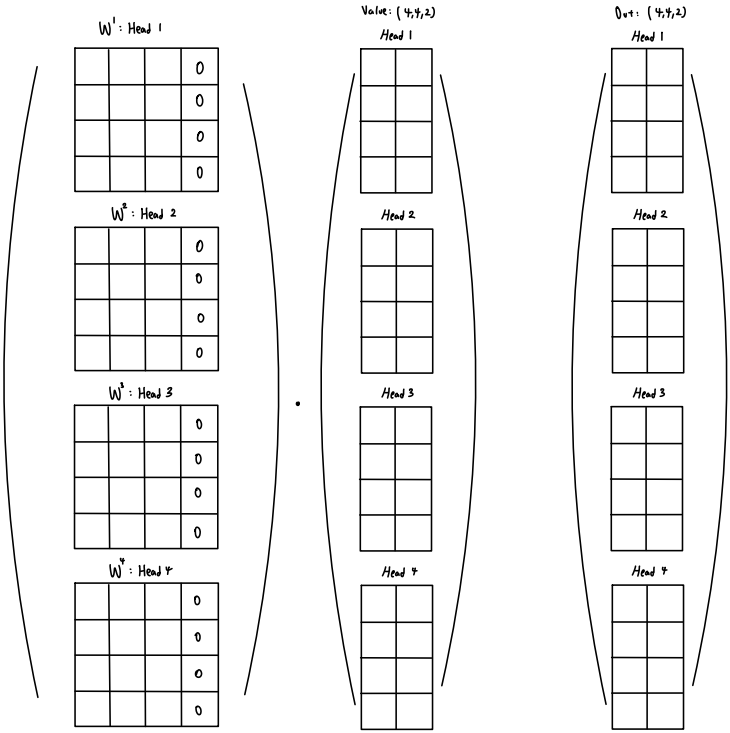


This is a masking process for Encoder

③-1 : Causally Masking for Decoder. Let's say mask : | 1 | 1 | 1 | 0 | . For masking W, we need some size.



④ Out = W Value



⑤ Reshape output. : (B, S, d)