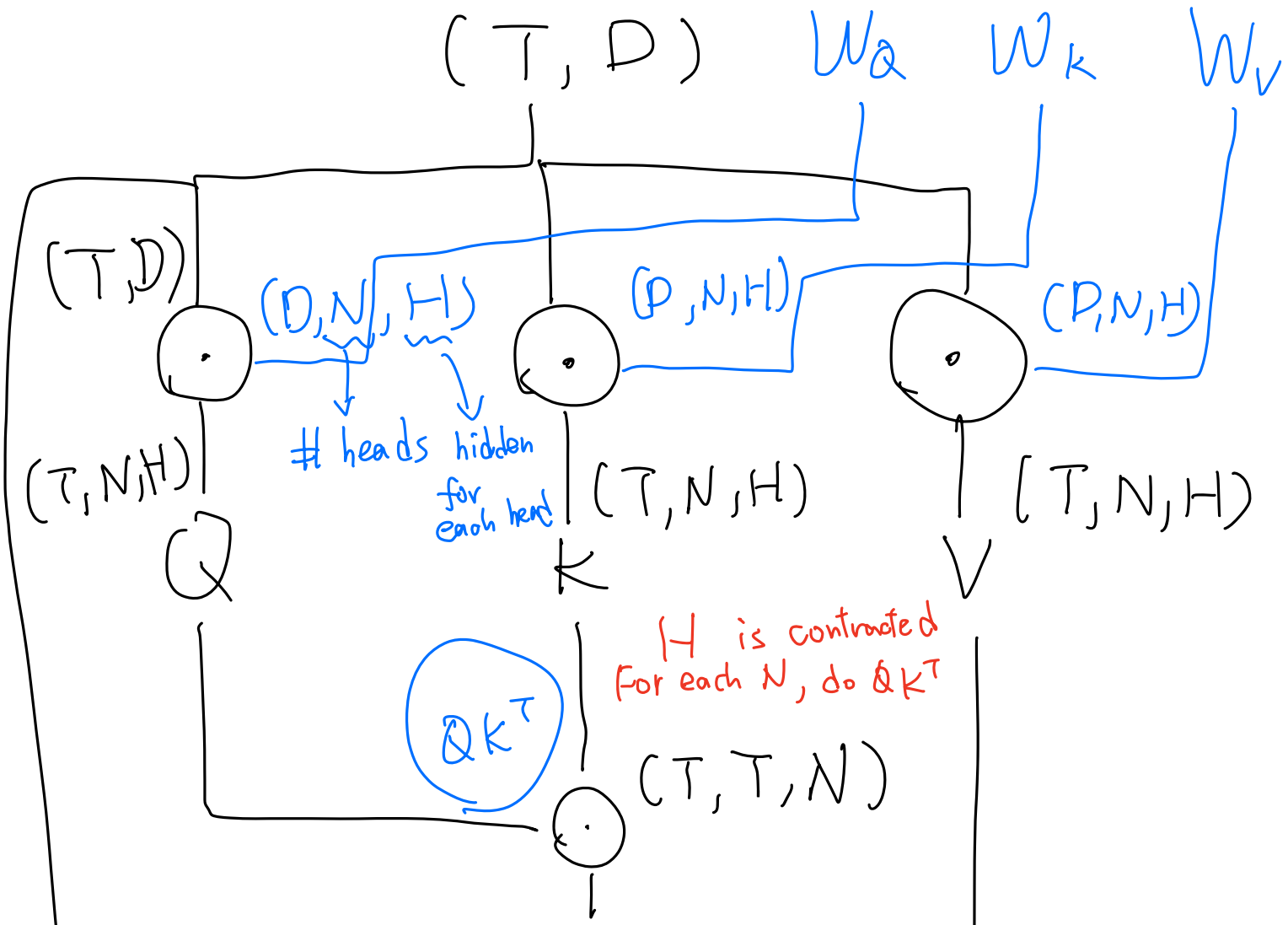
Vanilla attention: (original attention paper)

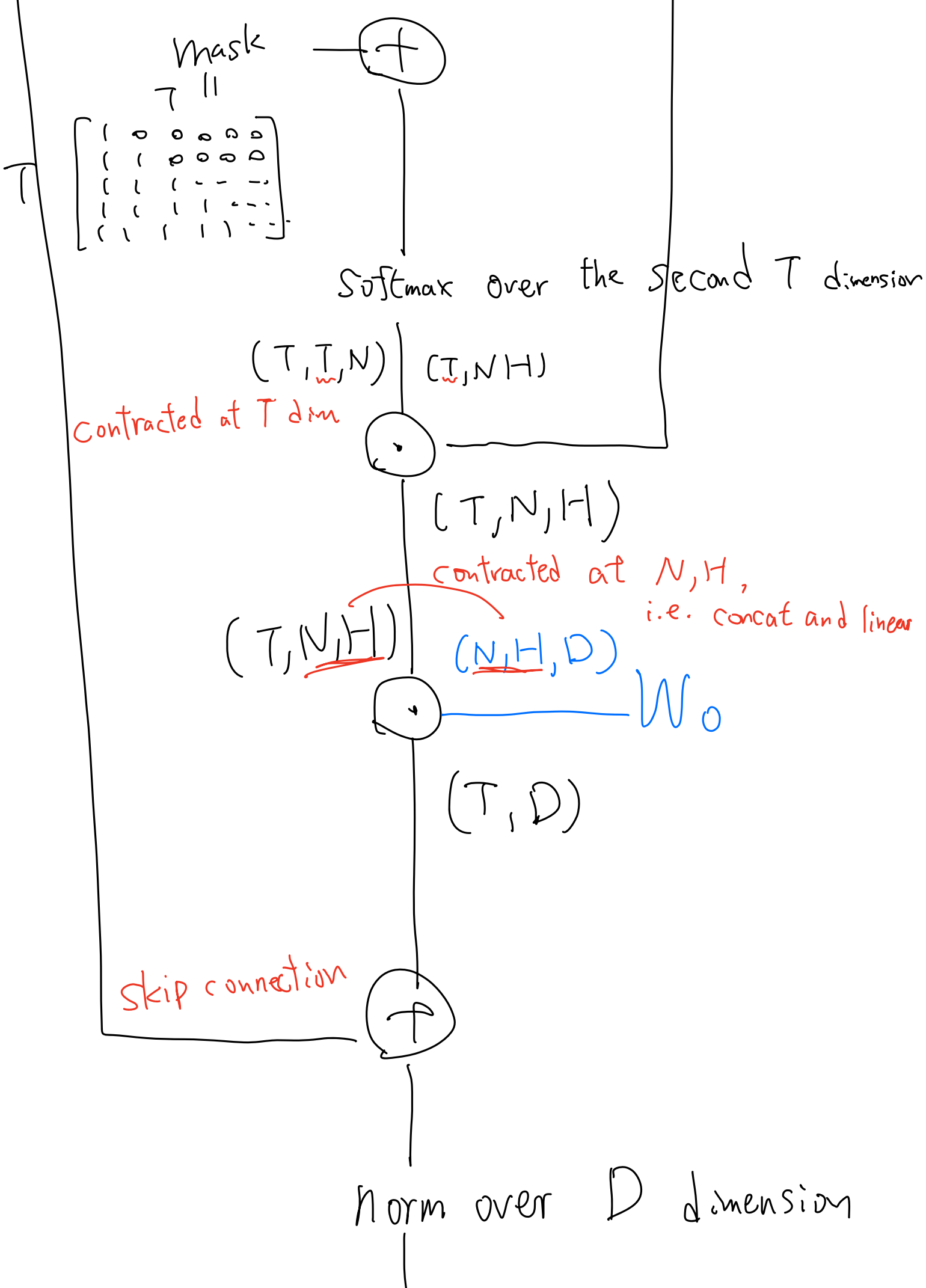
Assume input is: (B, T, D)

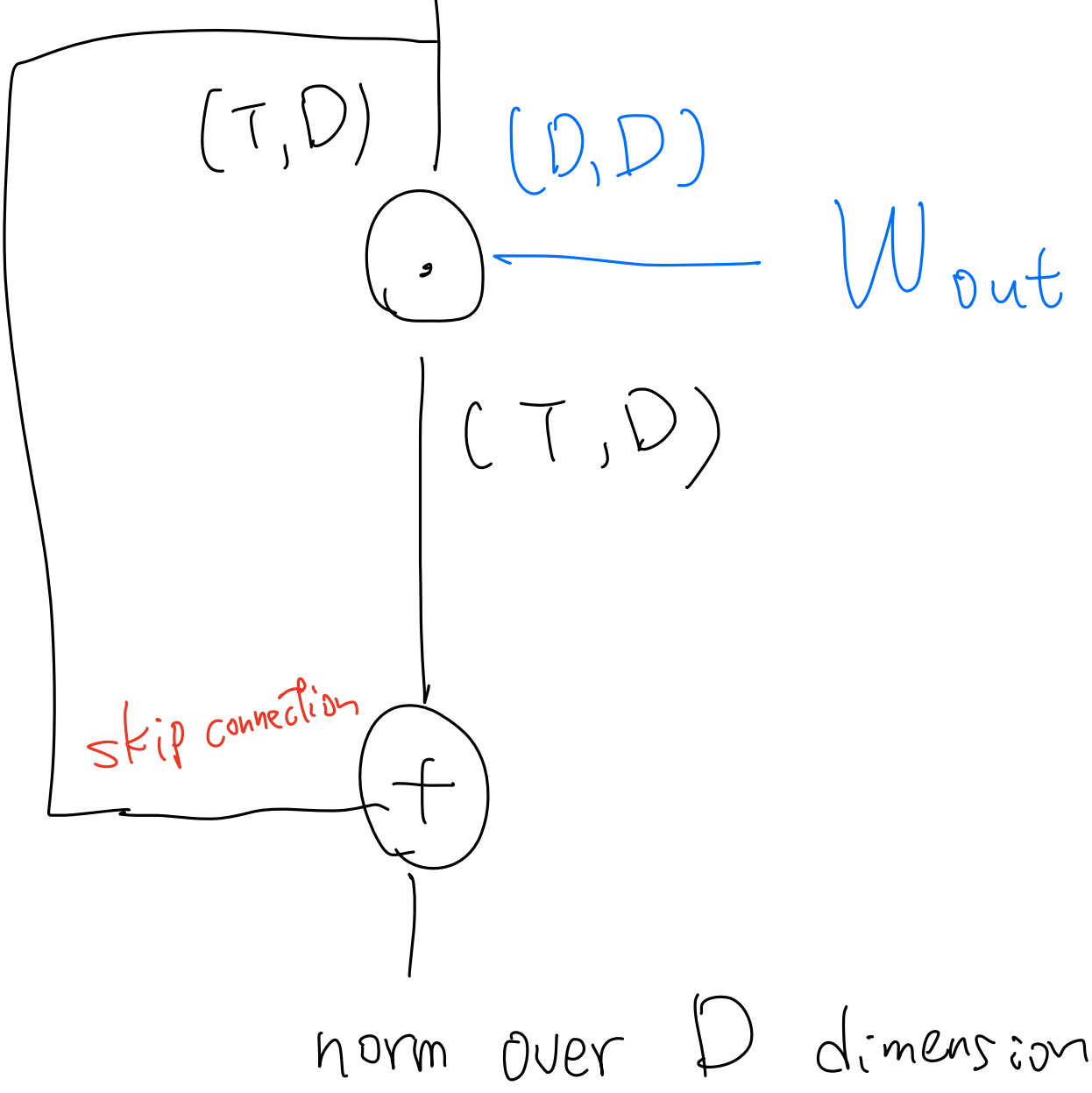
B: batch size, T: seq_len, D: embedding

We can ignore B for now and add this dimension at the end.

So (T, D) as input for the attention layer.




mask
T
T
Softmax over the second T dimension
(T, T, N)
(T, N, H)
contracted at T dim
(T, N, H)
contracted at N, H,
i.e. concat and linear
(T, (N, H))
((N, H), D)
W_o
(T, D)
skip connection
Norm over D dimension


(T, D)
(D, D)
W out
(T, D)
skip connection
f
norm over D dimension