Self Attention  $q_1 q_2 q_3 q_4$ 

2

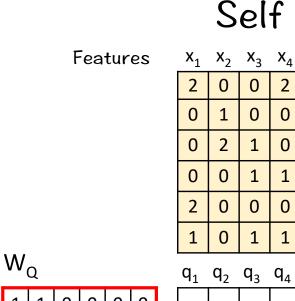
0

0

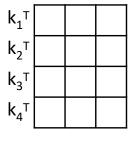
1

0

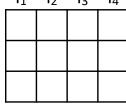
1

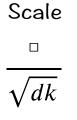


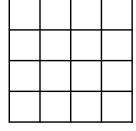
MatMul	
(K <sup>⊤</sup> Q)	

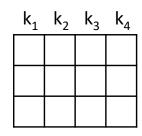
<u> </u>					
1	1	0	0	0	0
0	1	0	1	0	0
0	0	1	0	1	1







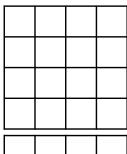
$W_K$					
0	0	1	0	0	0
0	1	0	0	0	0
1	0	0	0	0	-1



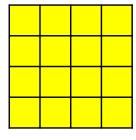




MatMul



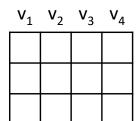




Attention Weight Matrix (A)

W	V
---	---

1 0	0	0	0	0	0
0	0	0	1 0	0	0
0	1 0	0	0	0	0



<sup>2</sup> 1	. 2	<b>2</b> 3	<b>4</b>
	_	+	

Attention Weighted **Features** 

© 2024 Tom Yeh

FFN