

Better Attention

Multi-Head Attention (MHA)

Input

t	1	2	3	4	5	6
X	0	3	0	3	0	2
	2	0	3	2	1	1
	1	0	1	0	1	3
	3	1	0	1	1	1
	2	3	2	2	1	2

Head 1

W_q	1	0	1	0	0	→	1	3	1	3	1	5
	0	1	1	0	1	→	5	3	6	4	3	6
	0	-1	0	1	0	→	1	1	-3	-1	0	0
W_k	0	0	-1	0	1	→	1	3	1	2	0	-1
	1	0	0	1	1	→	5	7	2	6	2	5
	0	1	0	1	0	→	5	1	3	3	2	2
W_v	0	0	-1	0	1	→	1	3	1	2	0	-1
	1	1	0	0	0	→	2	3	3	5	1	3
	-1	-1	0	0	0	→	-2	-3	-3	-5	-1	-3

K^T	1	5	5	→	17.9	13.3	9.24	10.4	9.24	20.2	→	0.01	0.01	0	0	0.01	0
	3	7	1	→	22.5	17.9	24.2	20.8	13.9	32.9	→	0.9	0.9	1	0.99	0.9	1
	1	2	3	→	8.08	6.93	23.1	4.62	4.04	9.81	→	0	0	0	0	0	0
	2	6	3	→	20.2	15.6	16.7	15.6	11.5	26.6	→	0.09	0.09	0	0.01	0.09	0
	0	2	2	→	6.93	4.62	3.46	3.46	3.46	6.93	→	0	0	0	0	0	0
	-1	5	2	→	15	8.08	13.3	6.66	6.08	14.4	→	0	0	0	0	0	0
$S = K^T Q / \sqrt{d_k}$	1	3	1	3	1	3	1	3	1	5	→	0.01	0.01	0	0	0.01	0
	5	3	6	4	3	6	→	0.9	0.9	1	→	0.9	0.9	1	0.99	0.9	1
	1	1	-3	-1	0	0	→	0	0	0	→	0	0	0	0	0	0
$A = \text{Softmax}(S)$	0.01	0.01	0	0	0.01	0	→	0.9	0.9	1	→	0.9	0.9	1	0.99	0.9	1
VA	2.9	2.9	3	3	2.9	3	→	3.2	3.2	3	→	-3	-3	-3	-3	-3	-3
V	1	3	1	2	0	-1	→	2.9	2.9	3	→	2.9	2.9	3	2.9	3	3
	2	3	3	5	1	3	→	3.2	3.2	3	→	-3	-3	-3	-3	-3	-3
	-2	-3	-3	-5	-1	-3	→	-3	-3	-3	→	-3	-3	-3	-3	-3	-3

Head 2

W_q	0	1	0	0	0	→	2	0	3	2	1	1
	-1	0	1	0	0	→	1	-3	1	-3	1	1
	0	0	-1	-1	0	→	-4	-1	-1	-1	-2	-4
W_k	0	0	1	0	0	→	1	0	1	0	1	3
	1	0	0	-1	0	→	-3	2	0	2	-1	1
	0	-1	0	0	1	→	0	3	-1	0	0	1
W_v	0	1	0	1	0	→	5	1	3	3	2	2
	-1	0	1	0	1	→	3	0	3	-1	2	3
	0	0	0	-1	1	→	-1	2	2	1	0	1

K^T	1	-3	0	→	-0.6	5.2	0	6.35	-1.2	-1.2	→	0.01	0.96	0.01	0.02	0.01	0
	0	2	3	→	-5.8	-5.2	-0.6	-5.2	-2.3	-5.8	→	0	0	0	0	0.01	0
	0	2	0	→	3.46	0.58	2.31	1.73	1.73	2.89	→	0.74	0.01	0.05	0.01	0.42	0.77
	0	2	0	→	1.15	-3.5	1.15	-3.5	1.15	1.15	→	0.07	0	0.02	0	0.24	0.14
	1	-1	0	→	0.58	1.73	1.15	2.89	0	0	→	0.04	0.03	0.02	0.03	0.07	0.04
	3	1	1	→	1.73	-2.3	5.2	1.15	1.15	0	→	0.13	0	0.91	0.01	0.24	0.04
$S = K^T Q / \sqrt{d_k}$	-0.6	5.2	0	6.35	-1.2	-1.2	→	0.01	0.96	0.01	→	0.01	0.96	0.01	0.02	0.01	0
	-5.8	-5.2	-0.6	-5.2	-2.3	-5.8	→	0	0	0	→	0	0	0	0	0.01	0
	3.46	0.58	2.31	1.73	1.73	2.89	→	0.74	0.01	0.05	0.01	→	0.74	0.01	0.05	0.01	0.42
	1.15	-3.5	1.15	-3.5	1.15	1.15	→	0.07	0	0.02	0	→	0.07	0	0.02	0	0.24
	0.58	1.73	1.15	2.89	0	0	→	0.04	0.03	0.02	0.03	→	0.04	0.03	0.02	0.03	0.07
	1.73	-2.3	5.2	1.15	1.15	0	→	0.13	0	0.91	0.01	→	0.13	0	0.91	0.01	0.24
$A = \text{Softmax}(S)$	0.01	0.96	0.01	0.02	0.01	0	→	0.01	0.96	0.01	→	0.01	0.96	0.01	0.02	0.01	0
	0	0	0	0	0	0	→	0	0	0	→	0	0	0	0	0	0
	0.74	0.01	0.05	0.01	0.42	0.77	→	0.74	0.01	0.05	0.01	→	0.74	0.01	0.05	0.01	0.42
	0.07	0	0.02	0	0.24	0.14	→	0.07	0	0.02	0	→	0.07	0	0.02	0	0.24
	0.04	0.03	0.02	0.03	0.07	0.04	→	0.04	0.03	0.02	0.03	→	0.04	0.03	0.02	0.03	0.07
	0.13	0	0.91	0.01	0.24	0.04	→	0.13	0	0.91	0.01	→	0.13	0	0.91	0.01	0.24
VA	2.9	4.9	2.1	4.9	2.7	2.9	→	2.9	4.9	2.1	4.9	→	2.9	4.9	2.1	4.9	2.7
	2.7	3	2.9	3	2	2.4	→	2.7	3	2.9	3	→	2.7	3	2.9	3	2
	1.7	-1	1	-1	1.3	1.7	→	1.7	-1	1	-1	→	1.7	-1	1	-1	1.3

Head 3

W_q	0	0	0	1	→	2	3	2	2	1	2	2
	1	1	0	0	-1	→	0	0	1	3	0	1
	0	0	1	0	1	→	3	3	3	2	2	5
W_k	1	0	0	1	0	→	3	4	0	4	1	3
	1	0	0	-1	0	→	-3	2	0	2	-1	1
	0	1	1	0	1	→	5	3	6	4	3	6
W_v	1	1	0	0	-1	→	0	0	1	3	0	1
	1	0	0	0	-1	→	-2	0	-2	1	-1	0
	0	0	0	1	0	→	3	1	0	1	1	1

K^T	3	-3	5	→	12.1	13.9	10.4	4.04	7.51	16.2	→	0.13	0.13	0.01	0	0.18	0.01
	4	2	3	→	9.81	12.1	11	11.5	5.77	14.4	→	0.01	0.02	0.03	0.17	0.03	0
	0	0	6	→	10.4	10.4	10.4	6.93	6.93	17.3	→	0.02	0	0.01	0	0.1	0.02
	4	2	4	→	11.5	13.9	12.7	12.7	6.93	17.3	→	0.07	0.13	0.14	0.53	0.1	0.02
	1	-1	3	→	6.35	6.93	5.77	2.89	4.04	9.24	→	0	0	0	0	0.01	0
	3	1	6	→	13.9	15.6	14.4	12.1	8.66	21.4	→	0.74	0.72	0.8	0.3	0.18	0.96
$S = K^T Q / \sqrt{d_k}$	12.1	13.9	10.4	4.04	7.51	16.2	→	0.13	0.13	0.01	→	0.13	0.13	0.01	0	0.18	0.01
	9.81	12.1	11	11.5	5.77	14.4	→	0.01	0.02	0.03	→	0.01	0.02	0.03	0.17	0.03	0
	10.4	10.4	10.4	6.93	6.93	17.3	→	0.02	0	0.01	→	0.02	0	0.01	0	0.1	0.02
	11.5	13.9	12.7	12.7	6.93	17.3	→	0.07	0.13	0.14	→	0.07	0.13	0.14	0.53	0.1	0.02
	6.35	6.93	5.77	2.89	4.04	9.24	→	0	0	0	→	0	0	0	0	0.01	0
	13.9	15.6	14.4	12.1	8.66	21.4	→	0.74	0.72	0.8	→	0.74	0.72	0.8	0.3	0.18	0.96
$A = \text{Softmax}(S)$	0.13	0.13	0.01	0	0.18	0.01	→	0.13	0.13	0.01	→	0.13	0.13	0.01	0	0.18	0.01
	0.01	0.02	0.03	0.17	0.03	0	→	0.01	0.02	0.03	→	0.01	0.02	0.03	0.17	0.03	0
	0.02	0	0.01	0	0.1	0.02	→	0.02	0	0.01	→	0.02	0	0.01	0	0.1	0.02
	0.07	0.13	0.14	0.53	0.1	0.02	→	0.07	0.13	0.14	→	0.07	0.13	0.14	0.53	0.1	0.02
	0	0	0	0	0.01	0	→	0	0	0	→	0	0	0	0	0.01	0
	0.74	0.72	0.8	0.3	0.18	0.96	→	0.74	0.72	0.8	→	0.74	0.72	0.8	0.3	0.18	0.96
VA	1	1.1	1.2	1.9	1	1	→	1	1.1	1.2	1.9	→	1	1.1	1.2	1.9	1
	-0	-0	0.1	0.5	-0	-0	→	-0	-0	0.1	0.5	→	-0	-0	0.1	0.5	-0
	1.2	1.3	1	1	1.3	1	→	1.2	1.3	1	1	→	1.2	1.3	1	1	1.3

Head 4

W_q	1	1	0	0	0	→	2	3	3	5	1	3
	0	1	1	1	1	→	8	4	6	5	4	7
	0	1	0	0	0	→	2	0	3	2	1	1
W_k	1	1	1	-1	1	→	2	5	6	6	2	7
	0	0	0	0	0	→	0	0	0	0	0	0

K^T	2	0	-2	→											$A = \text{Softmax}(S)$				
	5	0	-3	→	2.31	8.66	3.46	11	1.15	6.93	→	0	0	0	0	0	0.02	0	
	6	0	-2	→	4.62	10.4	6.93	15	2.31	9.24	→	0.02	0.02	0	0	0.07	0.01		
	6	0	-2	→	4.62	10.4	6.93	15	2.31	9.24	→	0.19	0.13	0.13	0.05	0.23	0.13		
				→	4.62	10.4	6.93	15	2.31	9.24	→	0.19	0.13	0.13	0.05	0.23	0.13		
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								
				→							↓								

Multi-Query Attention (MQA)

t	1	2	3	4	5	6
X	0	3	0	3	0	2
Input	2	0	3	2	1	1
	1	0	1	0	1	3
	3	1	0	1	1	1
	2	3	2	2	1	2
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
W _K	0	0	1	0	0	→
	0	1	1	0	1	→
	0	1	1	-1	0	→
	0	1	0	1	0	→
W _V	1	0	1	0	0	→
	-1	0	1	0	0	→
	1	0	0	-1	0	→
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
W _Q	3	3	3	2	2	5
	0	-1	4	1	1	3
	5	1	3	3	2	2

Head 1

	↑	↓	↑	↓	↑	↓
WqQ	0	1	0	1	1	→
	1	0	0	1	→	
	1	0	0	1	→	
	1	0	-1	0	1	→
Q=WqQ	7	4	5	5	3	4
	2	6	2	5	1	4
	1	6	1	5	0	1
	↑	↓	↑	↓	↑	↓
	↑	↓	↑	↓	↑	↓
	↑	↓	↑	↓	↑	↓
K ^T	3	0	5	→		
	3	-1	1	→		
	3	4	3	→		
	2	1	3	→		
	2	1	2	→		
	5	3	2	→		
S=K ^T Q/ sqrt(d _k)	15	24.2	11.5	23.1	5.2	9.81
	11.5	6.93	8.08	8.66	4.62	5.2
	18.5	31.2	15	28.9	7.51	17.9
	11	18.5	8.66	17.3	4.04	8.66
	10.4	15	8.08	14.4	4.04	8.08
	24.8	28.9	19.1	28.9	10.4	19.6
A=Softmax(S)	0	0	0	0	0	0.01
	0	0	0	0	0	0
	0	0.91	0.02	0.5	0.05	0.15
	0	0	0	0	0	0
	0	0	0	0	0	0
	1	0.09	0.98	0.5	0.94	0.85
Softmax	↓	↑	↓	↑	↓	↑
	↓	↑	↓	↑	↓	↑
	↓	↑	↓	↑	↓	↑
V	1	3	1	3	1	5
	1	-3	1	-3	1	1
	-3	2	0	2	-1	1
VA	5	1.4	4.9	3	4.8	4.4
	1	1	1	1	1	1
	1	0.1	1	0.5	0.9	0.8

Head 2

	↑	↑	↑	↑	↑	↑
Wq	0	-1	0	1	1	→
	0	0	-1	1	1	→
	0	0	1	-1	1	→
Q=WqX	3	4	-1	1	1	2
	4	4	1	3	1	0
	0	2	3	1	1	4
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
K ^T	3	0	5	→		
	3	-1	1	→		
	3	4	3	→		
	2	1	3	→		
	2	1	2	→		
	5	3	2	→		
S=K ^T Q/sqrt(d _k)	5.2	12.7	6.93	4.62	15	→
	2.89	5.77	-0.6	0.58	1.73	5.77
	14.4	19.6	5.77	10.4	5.77	10.4
	5.77	10.4	4.62	3.46	9.24	→
	5.77	9.24	2.89	4.04	2.89	6.93
	15.6	20.8	2.31	9.24	5.77	10.4
A=Softmax(S)	0	0	0.69	0	0.13	0.98
	0	0	0	0	0	0.01
	0.24	0.24	0.22	0.76	0.4	0.01
	0	0	0.07	0	0.04	0
	0	0	0.01	0	0.02	0
	0.76	0.76	0.01	0.24	0.4	0.01
VA	4	4	1.2	2	2.7	1
	1	1	0.7	1	0.8	1
	0.8	0.8	-2	0.2	0.1	-3

Head 3

	↑	↑	↑	↑	↑	↑
Wq	1	1	1	1	0	→
	0	1	0	0	0	→
	1	1	1	1	1	→
Q = WqX	6	4	4	6	3	7
	2	0	3	2	1	1
	8	7	6	8	4	9
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
K ^T	3	0	5	→		
	3	-1	1	→		
	3	4	3	→		
	2	1	3	→		
	2	1	2	→		
	5	3	2	→		
S = K ^T Q / sqrt(d _k)	33.5	27.1	24.2	33.5	16.7	38.1
	13.9	11	8.66	13.9	6.93	16.7
	28.9	19.1	24.2	28.9	14.4	30
	21.9	16.7	21.9	11	24.2	→
	17.3	12.7	13.3	17.3	8.66	19.1
	30	19.6	23.7	30	15	32.3
A = Softmax(S)	0.96	1	0.39	0.96	0.76	1
	0	0	0	0	0	0
	0.01	0	0.39	0.01	0.08	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0.03	0	0.22	0.03	0.14	0
Softmax	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
	↑	↑	↑	↑	↑	↑
V	1	3	1	3	1	5
	1	-3	1	-3	1	1
	-3	2	0	-2	-1	1
VA	1.1	1.9	1.1	1.6	1	→
	1	1	1	1	1	→
	-3	-3	-1	-3	-2	-3

Head 4

W _Q	0	1	1	0	1	→
	0	0	1	1	0	→
	0	0	1	1	1	→
Q = W _Q X	5	3	6	4	3	6
	4	1	1	1	2	4
	-2	3	-3	1	-1	1

1 -1 0 0 0 → -2 3 -3 1 -1 1

Softmax

↓ ↓ ↓ ↓ ↓

$S = K'Q / \sqrt{d_k}$

K'

3 0 5 → 2.89 13.9 1.73 9.81 2.31 13.3

3 -1 1 → 5.2 6.35 6.08 6.93 3.46 8.66

3 4 3 → 14.4 12.7 7.51 11 8.08 21.4

2 1 3 → 4.62 9.24 2.31 6.93 2.89 11

2 1 2 → 5.77 7.51 4.04 6.35 3.46 10.4

5 3 2 → 19.1 13.9 15.6 14.4 11 25.4

VA

1 3 1 3 1 5 → 5 2.7 5 4.8 4.8 4.9

1 -3 1 -3 1 1 → 1 1 1 1 1 1

-3 2 0 2 -1 1 → 1 -1 1 0.9 0.9 1

Head 5

W_Q

0 1 0 0 1 → 4 3 5 4 2 3

0 0 1 1 -1 → 2 -2 -1 -1 1 2

1 1 1 0 0 → 3 3 4 5 2 6

$Q = W_Q X$

↓ ↓ ↓ ↓ ↓

Q

4 3 5 4 2 3

2 -2 -1 -1 2

3 3 4 5 2 6

K'

↓ ↓ ↓ ↓ ↓

$S = K'Q / \sqrt{d_k}$

↓ ↓ ↓ ↓ ↓

Softmax

$A = \text{Softmax}(S)$

0.05 0.99 0.95 1 0.26 0.88

0 0 0 0 0

0.14 0 0 0.26 0.09

0 0 0 0 0.01 0

0 0 0 0 0 0

0.81 0.01 0.05 0 0.46 0.03

VA

↓ ↓ ↓ ↓ ↓

4.2 1 1.2 1 2.9 1.1

1 1 1 1 1 0.9 1

0.7 -3 -3 -3 -3 -0 -3

3 5 4 5 3.5 5

1 1 1 1 1 1

0.4 1 0.8 1 0.6 1

Head 6

W_Q

1 0 1 0 1 → 3 6 3 5 2 7

0 1 1 0 0 → 3 0 4 2 2 4

-1 0 0 1 0 → 3 -2 0 -2 1 -1

$Q = W_Q X$

↓ ↓ ↓ ↓ ↓

Q

3 6 3 5 2 7

3 0 4 2 2 4

3 -2 0 -2 1 -1

K'

↓ ↓ ↓ ↓ ↓

$S = K'Q / \sqrt{d_k}$

↓ ↓ ↓ ↓ ↓

Softmax

$A = \text{Softmax}(S)$

0.02 0 0 0 0 0.01 0

0 0 0 0 0 0

0.49 0 0.24 0 0.35 0

0 0 0 0 0 0

0.49 1 0.76 1 0.83 1

VA

↓ ↓ ↓ ↓ ↓

3 5 4 5 3.5 5

1 1 1 1 1 1

0.4 1 0.8 1 0.6 1

Concat

Head 1

Head 2

Head 3

Head 4

Head 5

Head 6

W_O

1 1 0 1 1 1 0 1 0 1 0 0 0 0 1 0 -1 0

0 1 0 1 0 1 0 1 -1 0 1 0 1 0 1 0 1 -1 -1

0 0 1 1 1 0 0 0 1 0 0 0 -1 1 0 1 0 1 0

0 -1 0 1 1 0 -1 0 1 0 1 1 1 -1 -1 1 0

0 0 0 0 1 1 0 0 1 0 1 0 -1 -1 0 0 1 0

Output

↓ ↓ ↓ ↓ ↓

17 7.9 8.1 9 14 6.8

12 7.3 -1 4.2 7.7 0.6

12 5.7 3.1 5.5 9.1 2.3

4.7 1.1 2 0.9 2.3 -0

-4 -1 -2 -2 -3 -5

Grouped Query Attention (GQA)

t 1 2 3 4 5 6

Input

0	3	0	3	0	2
2	0	3	2	1	1
1	0	1	0	1	3
3	1	0	1	1	1
2	3	2	2	1	2

Group 1

W_k

0	0	1	0	1
0	1	1	-1	0
0	1	0	1	0

K = W_kX

3	3	3	2	2	5
0	-1	4	1	1	3
5	1	3	3	2	2

W_v

1	0	1	0	0
-1	0	1	0	0
1	0	0	-1	0

V = W_vX

1	3	1	3	1	5
1	-3	1	-3	1	1
-3	2	0	2	-1	1

Head 1

W_q

0	1	0	1	1
1	0	0	1	1
1	0	-1	0	1

Q = W_qX

7	4	5	5	3	4
2	6	2	5	1	4
1	6	1	5	0	1

S = K^TQ / sqrt(d_k)

15	24.2	11.5	23.1	5.2	9.81
3	-1	1	→	11.5	6.93
3	4	3	→	18.5	31.2
2	1	3	→	11	38.5
2	1	2	→	10.4	15
5	3	2	→	24.8	28.9

A = Softmax(S)

0	0	0	0	0.01	0
0	0	0	0	0	0
0	0.95	0.02	0.5	0.05	0.15
0	0	0	0	0	0
0	0	0	0	0	0
1	0.09	0.98	0.5	0.94	0.95

VA

5	1.4	4.9	3	4.8	4.4
1	1	1	1	1	1
1	0.1	1	0.5	0.9	0.8

Head 2

W_q

0	-1	0	1	1
0	0	-1	1	1
0	0	1	-1	1

Q = W_qX

3	4	-1	1	1	2
4	4	1	3	1	0
0	2	3	1	1	4

S = K^TQ / sqrt(d_k)

5.2	12.7	6.93	4.62	4.62	15
3	-1	1	→	2.89	5.77
3	4	3	→	14.4	19.6
2	1	3	→	5.77	10.4
2	1	2	→	5.77	9.24
5	3	2	→	15.6	20.8

A = Softmax(S)

0	0	0.69	0	0.13	0.98
0	0	0	0	0	0.01
0	0.24	0.24	0.22	0.76	0.4
0	0	0.07	0	0.04	0
0	0	0.01	0	0.02	0
0.76	0.76	0.01	0.24	0.4	0.01

VA

4	4	1.2	2	2.7	1
1	1	0.7	1	0.8	1
0.8	0.8	-2	0.2	0.1	-3

Head 3

W_q

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

Q = W_qX

6	4	4	6	3	7
2	0	3	2	1	1
8	7	6	8	4	9

S = K^TQ / sqrt(d_k)

33.5	27.1	24.2	33.5	16.7	38.1
3	-1	1	→	13.9	11
3	4	3	→	28.9	13.1
2	1	3	→	21.9	16.7
2	1	2	→	17.3	12.7
5	3	2	→	30	19.6

A = Softmax(S)

0.96	1	0.39	0.96	0.78	1
0	0	0	0	0	0
0.01	0	0.39	0.01	0.08	0
0	0	0	0	0	0
0.03	0	0.22	0.03	0.14	0

VA

1.1	1	1.9	1.1	1.6	1
1	1	1	1	1	1
-3	-3	-1	-3	-2	-3

Group 2

W_k

0	1	0	1	-1
1	0	1	0	1
0	1	1	1	1

K = W_kX

3	-2	1	1	1	0
3	6	3	5	2	7
8	4	6	5	4	7

W_v

0	-1	1	1	0
0	0	-1	1	0
2	1	-2	-1	1
2	1	-1	1	0
-2	1	-1	1	0

V = W_vX

1 1 0 0 1 → 4 6 5 7 2 5

Head 1

Wq
 0 1 1 0 1 → 5 3 6 4 3 6
 0 0 1 1 0 → 4 1 1 1 2 4
 1 -1 0 0 0 → -2 3 -3 1 -1 1

Q = WqX
 ↓ ↓ ↓ ↓ ↓ ↓
 5 3 6 4 3 6
 4 1 1 1 2 4
 -2 3 -3 1 -1 1

Q
 ↓ ↓ ↓ ↓ ↓ ↓
 5 3 6 4 3 6
 4 1 1 1 2 4
 -2 3 -3 1 -1 1

S = K^TQ / sqrt(d_k)
 ↓ ↓ ↓ ↓ ↓ ↓
 3 3 8 → 6.35 20.8 -17 13.3 4.04 21.9 →
 -2 6 4 → 3.46 6.93 -10 1.15 1.15 9.24 →
 1 3 6 → 2.89 13.9 -5.2 7.51 1.73 13.9 →
 1 5 5 → 8.66 13.3 -2.3 8.08 4.62 17.9 →
 1 2 4 → 2.89 9.81 -2.3 5.77 1.73 30.4 →
 0 7 7 → 8.08 16.2 -8.1 8.08 4.04 20.2 →

K^T
 3 3 8 → 6.35 20.8 -17 13.3 4.04 21.9 →
 -2 6 4 → 3.46 6.93 -10 1.15 1.15 9.24 →
 1 3 6 → 2.89 13.9 -5.2 7.51 1.73 13.9 →
 1 5 5 → 8.66 13.3 -2.3 8.08 4.62 17.9 →
 1 2 4 → 2.89 9.81 -2.3 5.77 1.73 30.4 →
 0 7 7 → 8.08 16.2 -8.1 8.08 4.04 20.2 →

S = K^TQ / sqrt(d_k)
 ↓ ↓ ↓ ↓ ↓ ↓
 3 3 8 → 6.35 20.8 -17 13.3 4.04 21.9 →
 -2 6 4 → 3.46 6.93 -10 1.15 1.15 9.24 →
 1 3 6 → 2.89 13.9 -5.2 7.51 1.73 13.9 →
 1 5 5 → 8.66 13.3 -2.3 8.08 4.62 17.9 →
 1 2 4 → 2.89 9.81 -2.3 5.77 1.73 30.4 →
 0 7 7 → 8.08 16.2 -8.1 8.08 4.04 20.2 →

A = Softmax(S)
 ↓ ↓ ↓ ↓ ↓ ↓
 0.06 0.99 0.46 0.99 0.25 0.34
 0 0 0 0 0.01 0
 0 0.01 0 0.02 0
 0.6 0 0.26 0.01 0.44 0.01
 0 0 0.26 0 0.02 0
 0.34 0.01 0 0.01 0.25 0.15

Softmax
 ↓ ↓ ↓ ↓ ↓ ↓
 0.06 0.99 0.46 0.99 0.25 0.34
 0 0 0 0 0.01 0
 0 0.01 0 0.02 0
 0.6 0 0.26 0.01 0.44 0.01
 0 0 0.26 0 0.02 0
 0.34 0.01 0 0.01 0.25 0.15

VA
 2 1 -2 -1 1 3 → 0.5 2 0.9 2 0.8 2.1
 2 1 -1 1 0 -2 → 0 2 1.2 2 0.4 1.4
 4 6 5 7 2 5 → 6.1 4 4.3 4 5.6 4.2

Head 2

Wq
 0 1 0 0 1 → 4 3 5 4 2 3
 0 0 1 1 -1 → 2 -2 -1 -1 1 2
 1 1 1 0 0 → 3 3 4 5 2 6

Q = WqX
 ↓ ↓ ↓ ↓ ↓ ↓
 4 3 5 4 2 3
 2 -2 -1 -1 1 2
 3 3 4 5 2 6

Q
 ↓ ↓ ↓ ↓ ↓ ↓
 4 3 5 4 2 3
 2 -2 -1 -1 1 2
 3 3 4 5 2 6

S = K^TQ / sqrt(d_k)
 ↓ ↓ ↓ ↓ ↓ ↓
 3 3 8 → 24.2 15.6 25.4 26.3 14.4 36.4 →
 -2 6 4 → 9.24 -3.5 0 3.46 5.77 17.3 →
 1 3 6 → 16.2 8.66 15 17.9 9.81 26 →
 1 5 5 → 16.7 4.62 11.5 13.9 9.81 24.8 →
 1 2 4 → 11.5 6.35 11 12.7 6.93 17.8 →
 0 7 7 → 20.2 4.04 12.1 16.2 12.1 32.3 →

K^T
 3 3 8 → 24.2 15.6 25.4 26.3 14.4 36.4 →
 -2 6 4 → 9.24 -3.5 0 3.46 5.77 17.3 →
 1 3 6 → 16.2 8.66 15 17.9 9.81 26 →
 1 5 5 → 16.7 4.62 11.5 13.9 9.81 24.8 →
 1 2 4 → 11.5 6.35 11 12.7 6.93 17.8 →
 0 7 7 → 20.2 4.04 12.1 16.2 12.1 32.3 →

S = K^TQ / sqrt(d_k)
 ↓ ↓ ↓ ↓ ↓ ↓
 3 3 8 → 24.2 15.6 25.4 26.3 14.4 36.4 →
 -2 6 4 → 9.24 -3.5 0 3.46 5.77 17.3 →
 1 3 6 → 16.2 8.66 15 17.9 9.81 26 →
 1 5 5 → 16.7 4.62 11.5 13.9 9.81 24.8 →
 1 2 4 → 11.5 6.35 11 12.7 6.93 17.8 →
 0 7 7 → 20.2 4.04 12.1 16.2 12.1 32.3 →

A = Softmax(S)
 ↓ ↓ ↓ ↓ ↓ ↓
 0.98 1 1 1 0.89 0.96
 0 0 0 0 0 0
 0 0 0 0 0.01 0
 0 0 0 0 0.01 0
 0 0 0 0 0 0
 0.02 0 0 0 0.08 0.02

Softmax
 ↓ ↓ ↓ ↓ ↓ ↓
 0.98 1 1 1 0.89 0.96
 0 0 0 0 0 0
 0 0 0 0 0.01 0
 0 0 0 0 0.01 0
 0 0 0 0 0 0
 0.02 0 0 0 0.08 0.02

VA
 2 1 -2 -1 1 3 → 2 2 2 2 2 2
 2 1 -1 1 0 -2 → 1.9 2 2 2 1.6 1.9
 4 6 5 7 2 5 → 4 4 4 4 4.1 4

Head 3

Wq
 1 0 1 0 1 → 3 6 3 5 2 7
 0 1 1 0 0 → 3 0 4 2 2 4
 -1 0 0 1 0 → 3 -2 0 -2 1 -1

Q = WqX
 ↓ ↓ ↓ ↓ ↓ ↓
 3 6 3 5 2 7
 3 0 4 2 2 4
 3 -2 0 -2 1 -1

Q
 ↓ ↓ ↓ ↓ ↓ ↓
 3 6 3 5 2 7
 3 0 4 2 2 4
 3 -2 0 -2 1 -1

S = K^TQ / sqrt(d_k)
 ↓ ↓ ↓ ↓ ↓ ↓
 3 3 8 → 24.2 11.5 12.1 2.89 11.5 14.4 →
 -2 6 4 → 13.9 -12 10.4 -3.5 6.93 3.46 →
 1 3 6 → 17.3 -3.5 8.66 -0.6 8.08 7.51 →
 1 5 5 → 19.1 -2.3 13.3 2.89 9.81 12.7 →
 1 2 4 → 12.1 -1.2 6.35 0.88 5.77 6.35 →
 0 7 7 → 24.2 -8.1 16.2 0 12.1 12.1 →

K^T
 3 3 8 → 24.2 11.5 12.1 2.89 11.5 14.4 →
 -2 6 4 → 13.9 -12 10.4 -3.5 6.93 3.46 →
 1 3 6 → 17.3 -3.5 8.66 -0.6 8.08 7.51 →
 1 5 5 → 19.1 -2.3 13.3 2.89 9.81 12.7 →
 1 2 4 → 12.1 -1.2 6.35 0.88 5.77 6.35 →
 0 7 7 → 24.2 -8.1 16.2 0 12.1 12.1 →

S = K^TQ / sqrt(d_k)
 ↓ ↓ ↓ ↓ ↓ ↓
 3 3 8 → 24.2 11.5 12.1 2.89 11.5 14.4 →
 -2 6 4 → 13.9 -12 10.4 -3.5 6.93 3.46 →
 1 3 6 → 17.3 -3.5 8.66 -0.6 8.08 7.51 →
 1 5 5 → 19.1 -2.3 13.3 2.89 9.81 12.7 →
 1 2 4 → 12.1 -1.2 6.35 0.88 5.77 6.35 →
 0 7 7 → 24.2 -8.1 16.2 0 12.1 12.1 →

A = Softmax(S)
 ↓ ↓ ↓ ↓ ↓ ↓
 0.5 0.98 0.02 0.46 0.33 0.78
 0 0 0 0 0 0
 0 0.01 0 0.01 0.01 0
 0 0.03 0.05 0.46 0.06 0.14
 0 0.09 0 0.05 0 0
 0.5 0 0.93 0.03 0.99 0.08

Softmax
 ↓ ↓ ↓ ↓ ↓ ↓
 0.5 0.98 0.02 0.46 0.33 0.78
 0 0 0 0 0 0
 0 0.01 0 0.01 0.01 0
 0 0.03 0.05 0.46 0.06 0.14
 0 0.09 0 0.05 0 0
 0.5 0 0.93 0.03 0.99 0.08

VA
 2 1 -2 -1 1 3 → 2.5 1.8 2.8 0.6 2.4 1.7
 2 1 -1 1 0 -2 → 0 1.8 -2 1.3 -0 1.5
 4 6 5 7 2 5 → 4.5 3.9 5.1 5.3 4.8 4.5

Concat

Group 1 Head 1
 5 1.4 4.9 3 4.8 4.4
 1 1 1 1 1 1
 1 0.1 1 0.5 0.9 0.8

Head 2
 4 4 1.2 2.7 1
 1 1 0.7 1 0.8 1
 0.8 0.8 -2 0.2 0.1 -3

Head 3
 1.1 1 1.9 1.1 1.6 1
 1 1 1 1 1 1
 -3 -3 -1 -3 -2 -3

Group 2 Head 4
 0.5 2 0.9 2 0.8 2.1
 0 2 1.2 2 0.4 1.4
 6.1 4 4.3 4 5.6 4.2

Head 5
 2 2 2 2 2 2
 1.9 2 2 2 1.6 1.9
 4 4 4 4 4.1 4

Head 6
 2.5 1.8 2.8 0.6 2.4 1.7
 0 1.8 -2 1.3 -0 1.5
 4.5 3.9 5.1 5.3 4.8 4.5

W₀

1 1 0 1 1 1 0 1 0 1 0 0 0 1 0 -1 0 → 17 13 14 13 16 10
 0 1 0 1 0 1 -1 -1 0 1 1 1 0 1 1 -1 → 19 14 11 9 16 6.9
 0 0 1 1 1 0 0 0 1 0 0 -1 1 0 1 0 → 9.4 12 2.1 9 7.4 5.7
 0 -1 0 1 1 0 -1 0 1 0 1 1 1 -1 -1 1 0 → 3.7 6 -1 4.7 1.5 2.4
 0 0 0 1 1 0 0 1 0 1 0 -1 -1 0 1 0 → -5 -2 -7 -2 -5 -6

Output

Mixture-of-Expert Multi-Head Attention



