Transformer

Features from the Previous Block

↓ ↓ ↓ ↓ ↓

Attention

Q
K

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

Attention Weight Matrix (A)

X_1	X_2	X_3	X_4	X_5
5	6	0	7	0
0	2	4	0	3
1	0	1	1	0

Z_1	Z_2	Z_3	Z_4	Z_5
11	6	7	7	5
2	6	4	3	3
1	1	2	1	1

Attention Weighted Features

1 1 1 1 1

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

10	1	4	5	3
13	12	11	10	8
4	8	7	5	5
-8	1	-1	-3	-1

ReLU ≈

10	1	4	5	3
13	12	11	10	8
4	8	7	5	5
0	1	0	0	0

1 1 1 1 1

Position-wise Feed-Forward Network (FFN)

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

10	0	4	5	3
17	20	18	15	13
5	8	8	6	6

↓ ↓ ↓ ↓ ↓

Next Block

Transformer