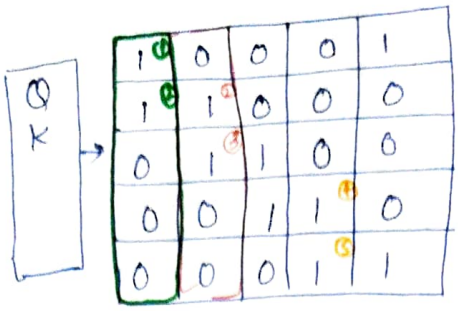
Transformer

features from the previous Block.



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	height
0	2	4	0	3	weight
1	0	1	1	0	age.

input

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
1	1	1	1	1

4×5

Attention weighted Features

7+0 = 7
0+3 = 3
1+0 = 1
5+6 = 11
0+2 = 2
1+0 = 1
6+0 = 6
2+4 = 6
0+1 = 1

Weight	weight	age	Bias
1	-1	0	1
1	1	0	0
0	1	1	1
-1	1	1	0

4×4

10	1	4	5	3
13	12	11	10	8
4	8	7	5	5
-9	1	-1	-4	-1

4×5

ReLU ≈

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

4×5

Weight = 1 = tall
weight = -1 = light
weight = 1 = heavy.

11×1 + (-2×1) + 0 + 1 = 10

11×1 + 2 + 0 = 13

				Bias
1	0	0	-1	0
0	1	1	0	0
0	0	1	-1	1

3×5

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

Point wise feed forward Networks (PFFN)

↓↓ Next Block.