Handout: Self-Attention

Example for computing self-attention for input sequences step-by-step

Step 1: Prepare inputs

3 inputs, each with dimension 4:

Input 1: [1,0,1,0] Input 2: [0,2,0,2] Input 3: [1,1,1,1]

Step 2: Initialize weights

Each weight has dimensions 4 x 3

W_{query} = W_{key} = W_{value} =

```
[[1,0,1],        [[0,0,1],        [[0,2,0],
 [1,0,0],         [1,1,0],         [0,3,0],
 [0,0,1],         [0,1,0],         [1,0,3],
 [0,1,1]]         [1,1,0]]         [1,1,0]]
```

Step 3: Derive query, key, and value for each input by matrix multiplication between inputs and weights

Queries:

```
[[1,0,1,0],          [[1,0,1],          [[1,0,2],
 [0,2,0,2],     x     [1,0,0],     =     [2,2,2],
 [1,1,1,1]]           [0,0,1],           [2,1,3]]
                      [0,1,1]]
```

Keys:

```
[[1,0,1,0],          [[0,0,1],          [[0,1,1],
 [0,2,0,2],     x     [1,1,0],     =     [4,4,0],
 [1,1,1,1]]           [0,1,0],           [2,3,1]]
                      [1,1,0]]
```

Values:

```
[[1,0,1,0],          [[0,2,0],          [[1,2,3],
 [0,2,0,2],     x     [0,3,0],     =     [2,8,0],
 [1,1,1,1]]           [1,0,3],           [2,6,3]]
                      [1,1,0]]
```

Step 4: Calculate attention scores

For Input 1: Take the dot product between Input 1's query with all the keys, including its own, by matrix multiplication

[1,0,2]	x	[[0,4,2], [1,4,3], [1,0,1]]	=	[2,4,4]

For Input 2: Do same as for Input 1.

[2,2,2]	x	[[0,4,2], [1,4,3], [1,0,1]]	=	[4,16,12]

For Input 3: Do same as for Input 1 and Input 2.

[2,1,3]	x	[[0,4,2], [1,4,3], [1,0,1]]	=	[4,12,10]

Step 5: Calculate softmax scores

For input 1: softmax([2,4,4]) = [0,0.5,0.5]

For input 2: softmax([4,16,12]) = [0,1.0,0]

For input 3: softmax([4,12,10]) = [0,0.9,0.1]

Step 6: Multiply scores with values

The softmaxed attention score for each input is multiplied by its corresponding value.

For input 1: 0.0 x [1,2,3] = [0.0,0.0,0.0]

For input 2: 0.5 x [2,8,0] = [1.0,4.0,0.0]

For input 3: 0.5 x [2,6,3] = [1.0,3.0,1.5]

Step 7: Sum weighted values to get outputs

For Input 1:
0.0 + 1.0 + 1.0 = 2.0 0.0 + 4.0 + 3.0 = 7.0 0.0 + 0.0 + 1.5 = 1.5

Self-attention:	[2.0,7.0,1.5]	Output 1
	[2.0,8.0,0.0]	Output 2
	[2.0,7.8,0.3]	Output 3