

keys_embedding (N, D)

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

N

D

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

query_embedding (D,)

keys_embedding (N, D)

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

N

D

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

query_embedding (D,)

$$1*2 + 0*0 + 1*-1 = 0$$



keys_embedding (N, D)

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

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

query_embedding (D,)

$$1*2 + 0*0 + -1*-1 = 3$$

1
3

keys_embedding (N, D)

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

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

query_embedding (D,)

keys_embedding (N, D)

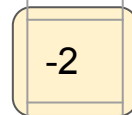
1	0	1
1	0	-1
-1	1	0
-1	0	1

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

query_embedding (D,)

1
3
-2

$$-1*2 + 1*0 + 0*-1 = -2$$



keys_embedding (N, D)

1	0	1
1	0	-1
-1	1	0

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

1
3
-2

-3

$$-1*2 + 1*0 + 1*-1 = -3$$

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

query_embedding (D,)

score (N,)

1
3
-2
-3

keys_embedding (N, D)

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

```
score = keys_embedding.dot(query_embedding)
```

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

query_embedding (D,)

score (N,)

1
3
-2
-3

keys_embedding (N, D)

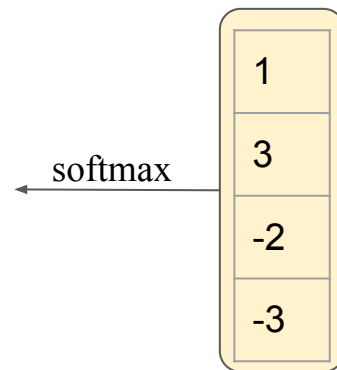
1	0	1
1	0	-1
-1	1	0
-1	0	1

```
score = keys_embedding.dot(query_embedding)
```

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

query_embedding (D,)

score (N,)



keys_embedding (N, D)

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

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

query_embedding (D,)

attention_weight (N,)

0.12
0.87
0.01
0.

score (N,)

1
3
-2
-3

softmax

keys_embedding (N, D)

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

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

query_embedding (D,)

attention_weight (N,)

0.12
0.87
0.01
0.

score (N,)

1
3
-2
-3

softmax

keys_embedding (N, D)

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

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

query_embedding (D,)

attention_weight (N,)

0.12
0.87
0.01
0.

keys_embedding (N, D)

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

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

query_embedding (D,)

attention_weight (N,)

0.12
0.87
0.01
0

*

*

*

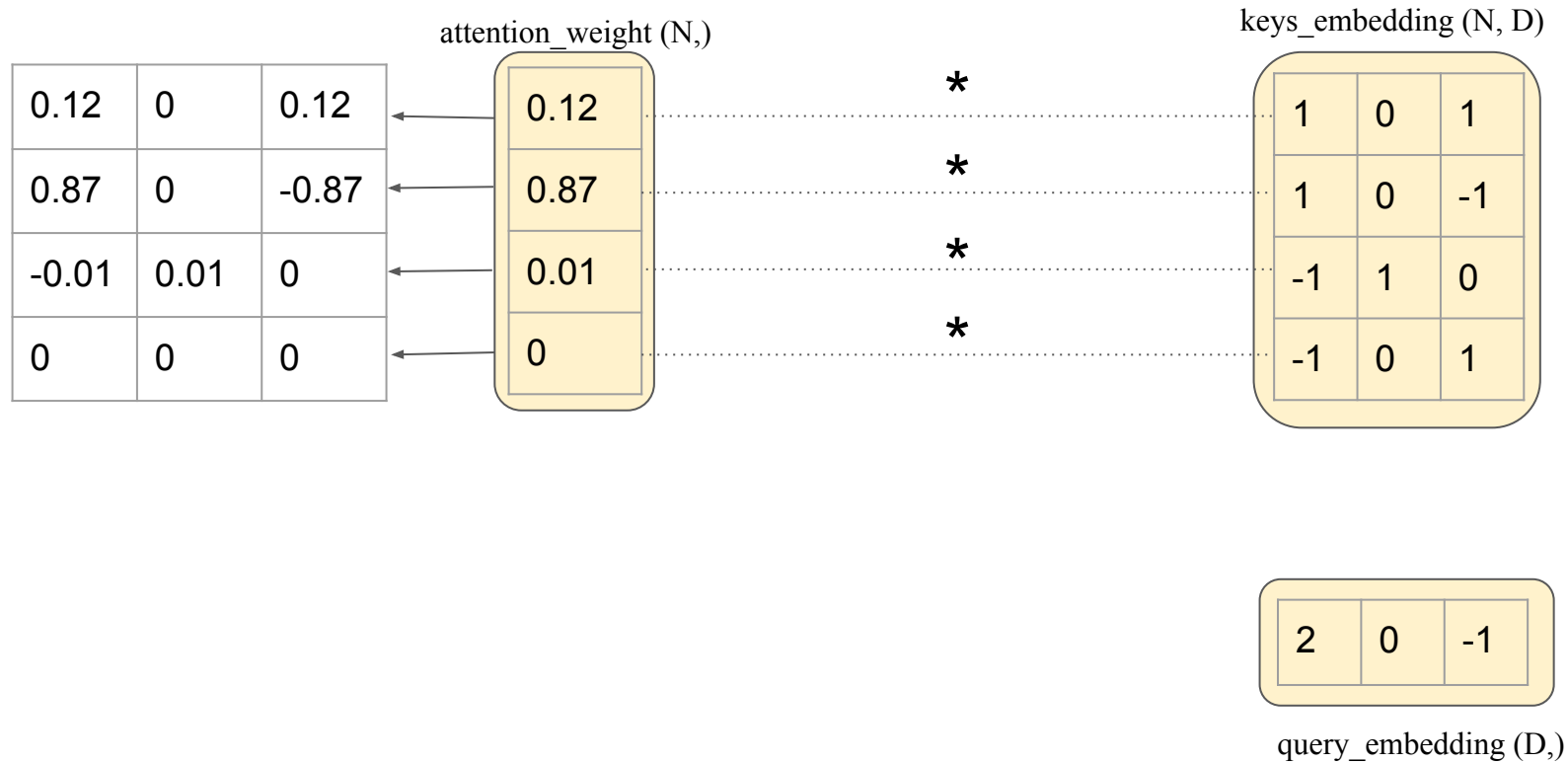
*

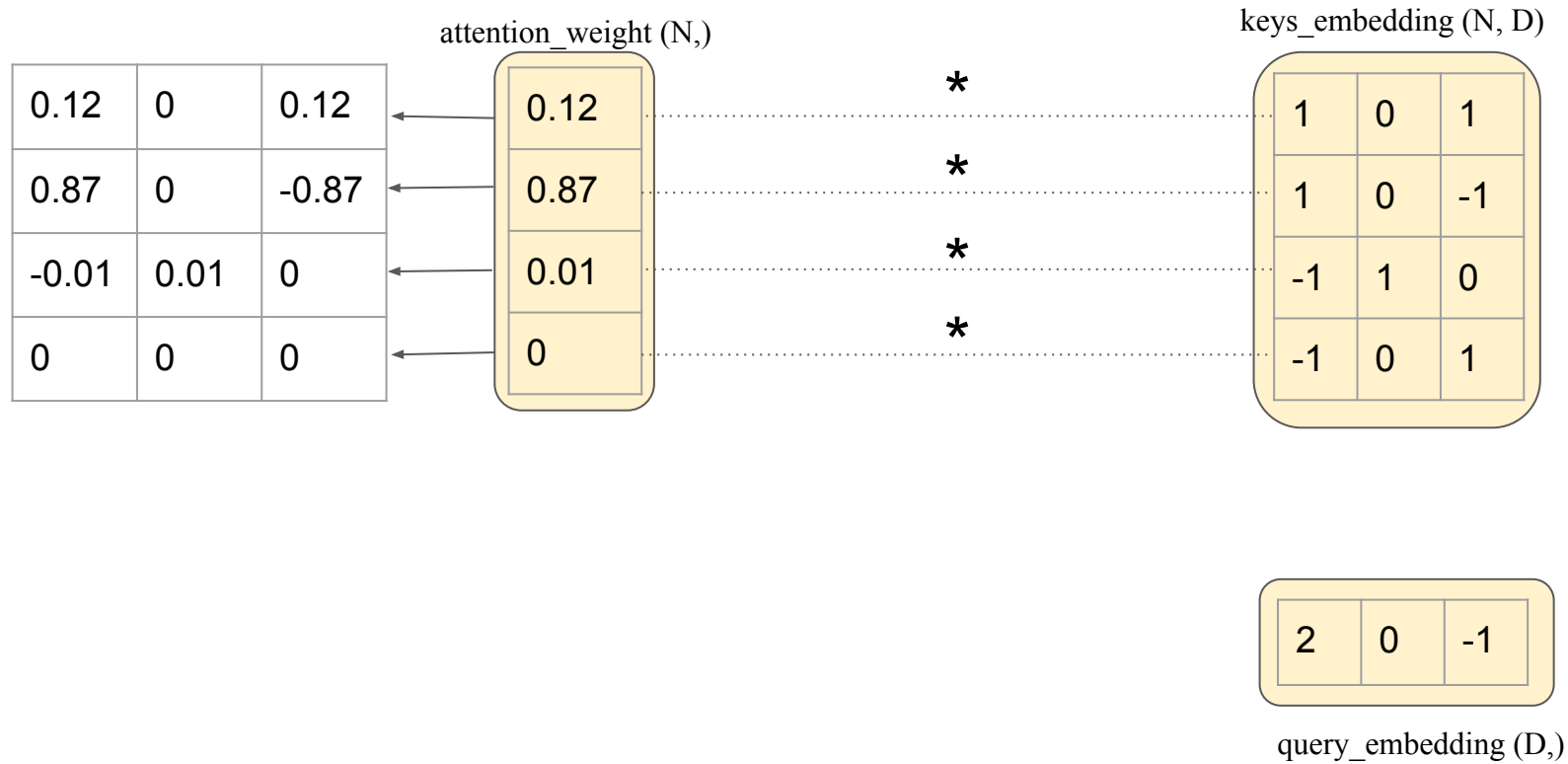
keys_embedding (N, D)

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

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

query_embedding (D,)





0.12	0	0.12
0.87	0	-0.87
-0.01	0.01	0
0	0	0

--	--	--

attention_weight (N,)

0.12
0.87
0.01
0

*

*

*

*

keys_embedding (N, D)

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

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

query_embedding (D,)

0.12	0	0.12
0.87	0	-0.87
-0.01	0.01	0
0	0	0



0.98		
------	--	--

$$0.12 + 0.87 - 0.01 + 0 = 0.98$$

attention_weight (N,)

0.12
0.87
0.01
0

*

*

*

*

keys_embedding (N, D)

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

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

query_embedding (D,)

0.12	0	0.12
0.87	0	-0.87
-0.01	0.01	0
0	0	0

0.98	0.01	
------	------	--

$$0 + 0 + 0.01 + 0 = 0.01$$

attention_weight (N,)

0.12
0.87
0.01
0

*

*

*

*

keys_embedding (N, D)

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

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

query_embedding (D,)

0.12	0	0.12
0.87	0	-0.87
-0.01	0.01	0
0	0	0

0.98	0.01	-0.75
------	------	-------

$$0.12 - 0.87 + 0 + 0 = -0.75$$

attention_weight (N,)

0.12
0.87
0.01
0

*

*

*

*

keys_embedding (N, D)

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

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

query_embedding (D,)

0.12	0	0.12
0.87	0	-0.87
-0.01	0.01	0
0	0	0

attention_weight (N,)

0.12
0.87
0.01
0

*

*

*

*

keys_embedding (N, D)

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

0.98	0.01	-0.75
------	------	-------

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

query_embedding (D,)

selfish_attention (D,)

0.98	0.01	-0.75
------	------	-------

keys_embedding (N, D)

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

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

query_embedding (D,)

keys_embedding (N, D)

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

selfish_attention (D,)

0.98	0.01	-0.75
------	------	-------

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

query_embedding (D,)

Need to make it **batch parallel** !

keys_embedding (N, D)

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

selfish_attention (D,)

0.98	0.01	-0.75
------	------	-------

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

query_embedding (D,)

Need to make it **batch parallel** !

keys_embedding (N,D)

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

keys_embedding (N,D)

2	3	9
1	4	3
-1	0	5
-1	-2	0

keys_embedding (N, D)

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

-2	1	-1
----	---	----

query_embedding (D,)

1	1	1
---	---	---

query_embedding (D,)

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

query_embedding (D,)

Need to make it **batch parallel** !

keys_embedding (N,D)

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

keys_embedding (N,D)

2	3	9
1	4	3
-1	0	5
-1	-2	0

keys_embedding (N, D)

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

-2	1	-1
----	---	----

query_embedding (D,)

1	1	1
---	---	---

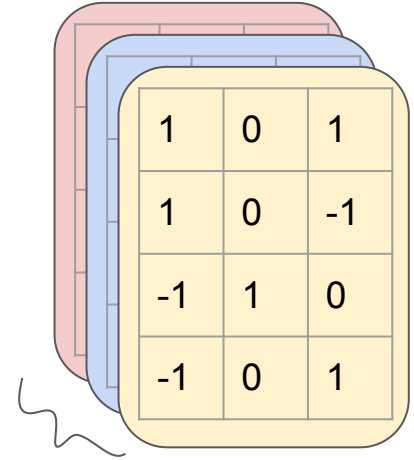
query_embedding (D,)

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

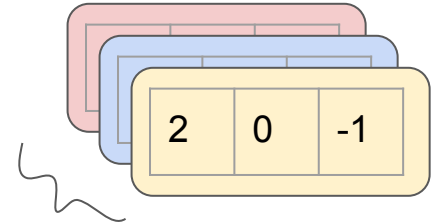
query_embedding (D,)

Need to make it **batch parallel** !

batch_keys_embedding (**B**,N,D)



Batch size

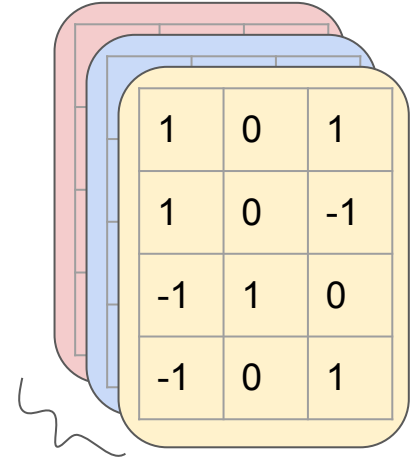


Batch size

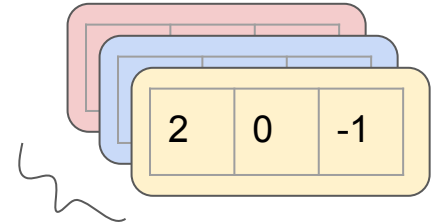
batch_query_embedding (**B**,D)

Need to make it **batch parallel** !

batch_keys_embedding (**B**,N,D)



Batch size



Batch size

batch_query_embedding (**B**,D)

Need to make it **batch parallel** !

batch_keys_embedding ($\mathbf{B}, \mathbf{N}, \mathbf{D}$)

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

`selfish_attention(batch_keys_embedding, batch_query_embedding)`

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

batch_query_embedding (\mathbf{B}, \mathbf{D})

Need to make it **batch parallel** !

batch_keys_embedding ($\mathbf{B}, \mathbf{N}, \mathbf{D}$)

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

`selfish_attention(batch_keys_embedding, batch_query_embedding)`

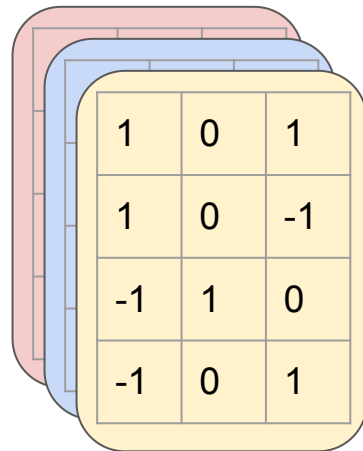
2	0	-1
---	---	----

batch_query_embedding (\mathbf{B}, \mathbf{D})

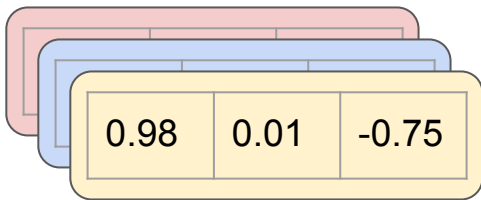
Need to make it **batch parallel** !

`selfish_attention(batch_keys_embedding, batch_query_embedding)`

batch_keys_embedding ($\mathbf{B}, \mathbf{N}, \mathbf{D}$)

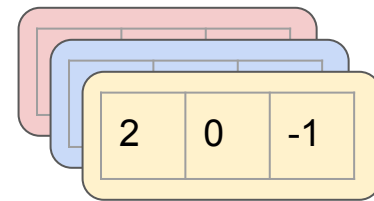


1	0	1
1	0	-1
-1	1	0
-1	0	1



0.98	0.01	-0.75
------	------	-------

selfish_attention (\mathbf{B}, \mathbf{D})



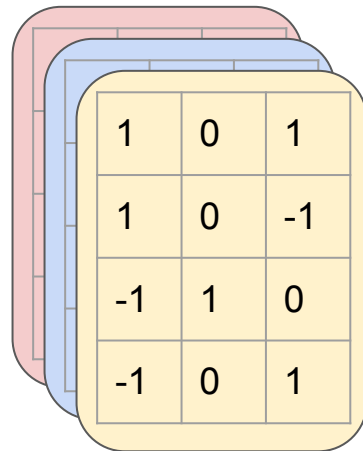
2	0	-1
---	---	----

batch_query_embedding (\mathbf{B}, \mathbf{D})

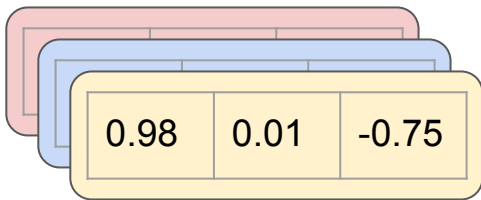
Need to make it **batch parallel** !

`selfish_attention(batch_keys_embedding, batch_query_embedding)`

batch_keys_embedding ($\mathbf{B}, \mathbf{N}, \mathbf{D}$)

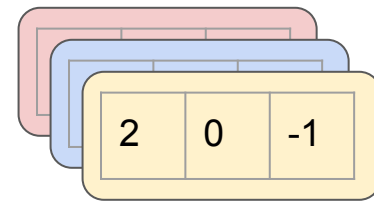


1	0	1
1	0	-1
-1	1	0
-1	0	1



0.98	0.01	-0.75
------	------	-------

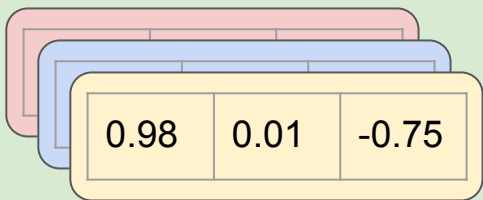
selfish_attention (\mathbf{B}, \mathbf{D})



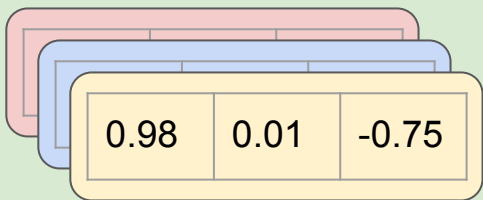
2	0	-1
---	---	----

batch_query_embedding (\mathbf{B}, \mathbf{D})

Need to make it **batch parallel** !



selffish_attention (**B**,D)



selffish_attention (\mathbf{B} , \mathbf{D})