Q @ K.T

(2,4,3) @ (2,3,4) ⇒ (2,4,4)

```
[
  [
    [ x0 x1 x2 ]
    [ x0 x1 x2 ]
    [ x0 x1 x2 ]
    [ x0 x1 x2 ]
  ],
  [
    [ x0 x1 x2 ]
    [ x0 x1 x2 ]
    [ x0 x1 x2 ]
    [ x0 x1 x2 ]
  ]
]
```

@

```
[
  [
    [ y0 y0 x0 x0 ]
    [ x1 x1 x1 x1 ]
    [ y1 x2 x2 x2 ]
  ]
  [
    [ y0 y0 x0 x0 ]
    [ x1 x1 x1 x1 ]
    [ y1 x2 x2 x2 ]
  ]
]
```

⇒

```
[
  [
    [x0, x1, x2, x3]
    [x0, x1, x2, x3]
    [x0, x1, x2, x3]
    [x0, x1, x2, x3]
  ],
  [
    [x0, x1, x2, x3]
    [x0, x1, x2, x3]
    [x0, x1, x2, x3]
    [x0, x1, x2, x3]
  ]
]
```

tril()
Softmax()

⇒

```
[
  [
    [x0, 0, 0, 0]
    [x0, x1, 0, 0]
    [x0, x1, x2, 0]
    [x0, x1, x2, x3]
  ],
  [
    [x0, 0, 0, 0]
    [x0, x1, 0, 0]
    [x0, x1, x2, 0]
    [x0, x1, x2, x3]
  ]
]
```

essentially each tokens "rating" of itself and the tokens around it.

⇒

emp V

(2,4,4) @ (2,4,3) ⇒ (2,4,3)

```
[
  [
    [x0, 0, 0, 0]
    [x0, x1, 0, 0]
    [x0, x1, x2, 0]
    [x0, x1, x2, x3]
  ],
  [
    [x0, 0, 0, 0]
    [x0, x1, 0, 0]
    [x0, x1, x2, 0]
    [x0, x1, x2, x3]
  ]
]
```

@

```
[
  [
    [x0 x1 x2]
    [x0 x1 x2]
    [x0 x1 x2]
    [x0 x1 x2]
  ],
  [
    [x0 x1 x2]
    [x0 x1 x2]
    [x0 x1 x2]
    [x0 x1 x2]
  ]
]
```

⇒ out (2,4,3)

concat heads on last dim if implementing multi head attention.