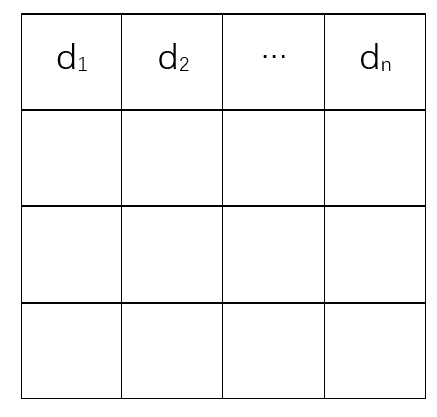
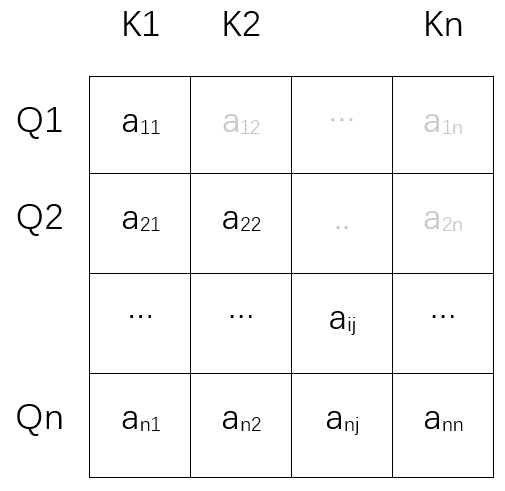
Parallel:

@

attention scores

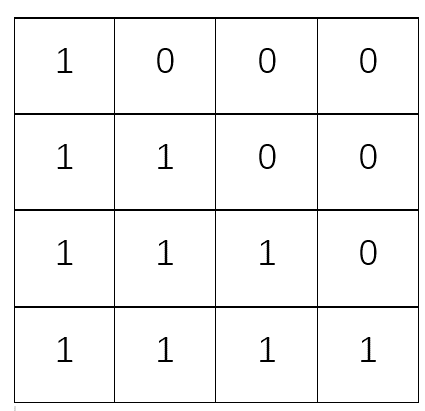


decay

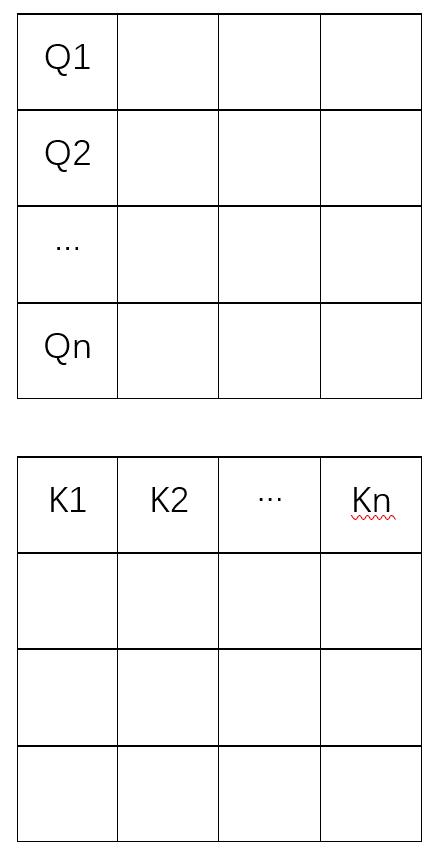
Q

K

\*



\*



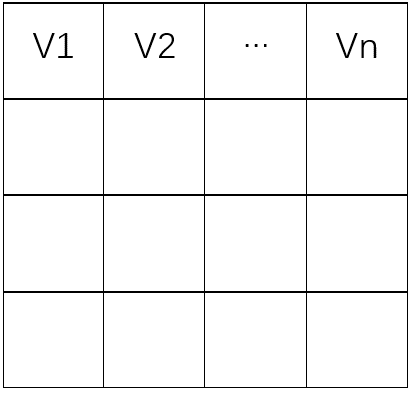
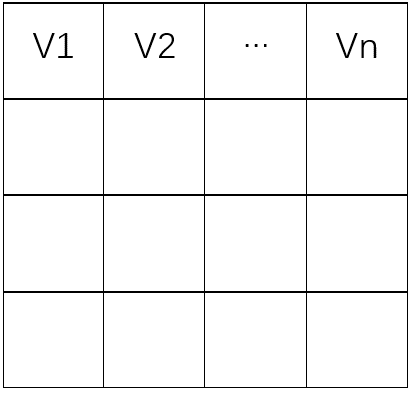
weight scores

scale

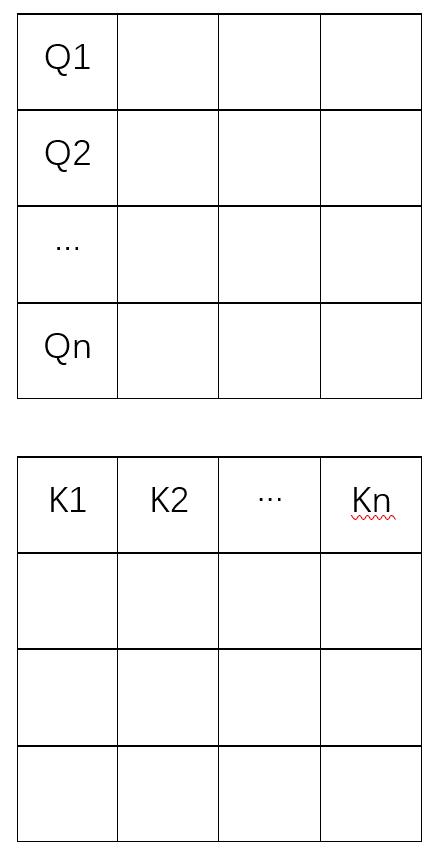
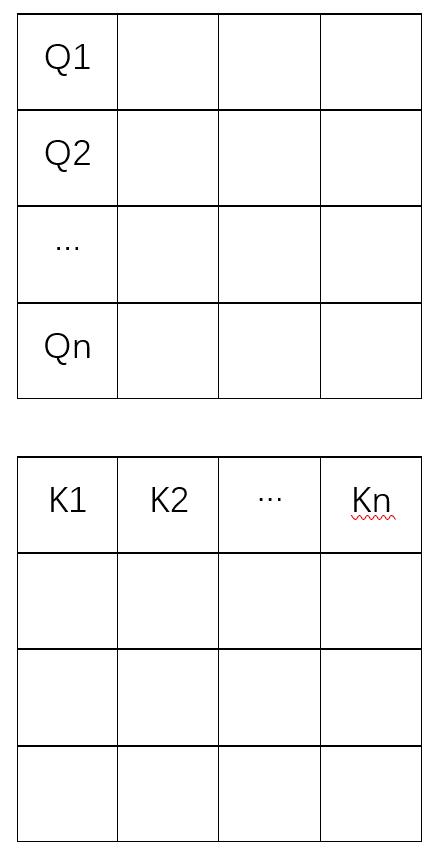
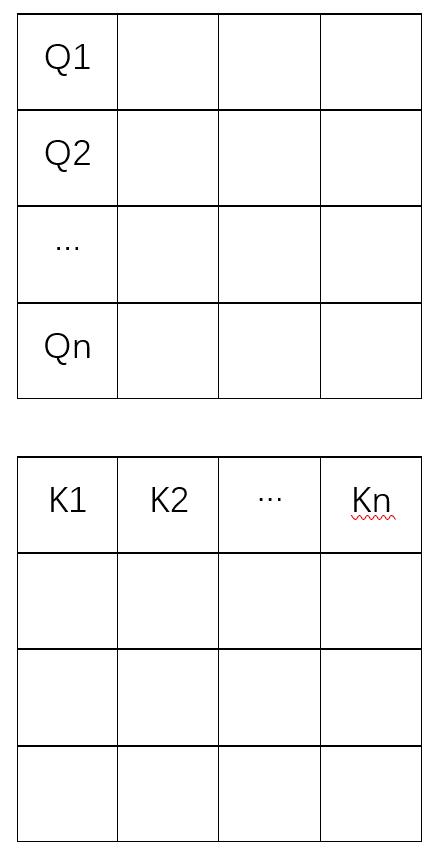
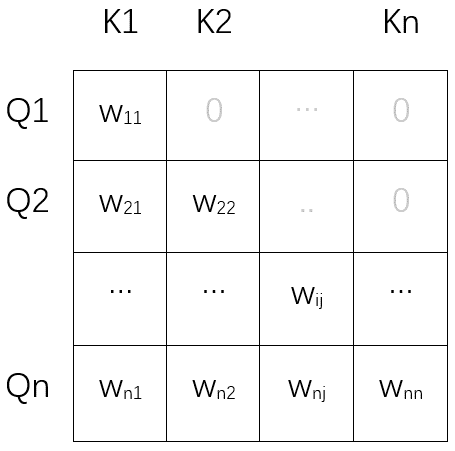
@

out

mask

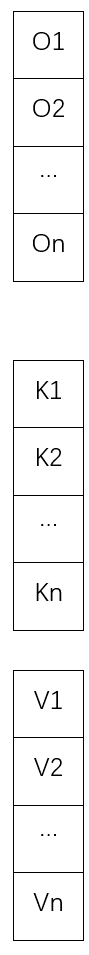
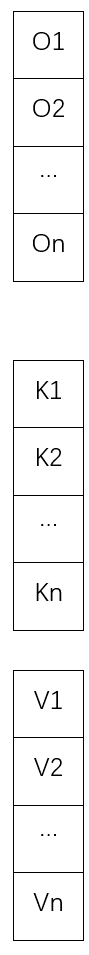


V



Recurrently:

V

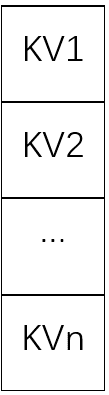


K

@

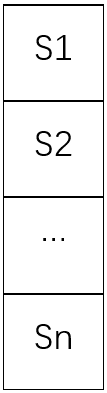
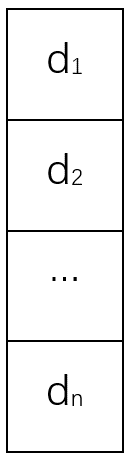
=

KV



dqk

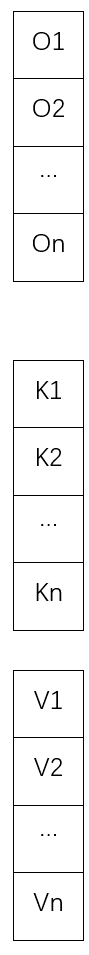
dv



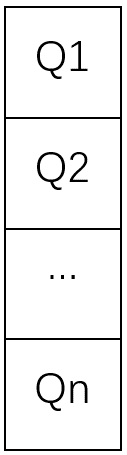
decay

\*

re



@



=

Q

S

dqk

dv

dqk

dv

dqk

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K1 | K2 |  | Kn |
| Q1 | w11 | 0 | … | 0 |
| Q2 | w21 | w22 | .. | 0 |
|  | … | … | wij | … |
| Qn | wn1 | wn2 | wnj | wnn |

|  |
| --- |
| d1 |
| d2 |
| … |
| dn |

|  |  |  |  |
| --- | --- | --- | --- |
| V1 | V2 | … | Vn |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| d1 | d2 | … | dn |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |