APR

self attention: $A^{<1>}, A^{<2>}, \dots A^{<T>}$

[MHA] : for loop (parallelization over SA)

$A(q,k,v)$ = attention-based vector representation of a word

↳ calculate for each word → $A^{<1>}, \dots A^{<5>}$

Jane visit Africa on September

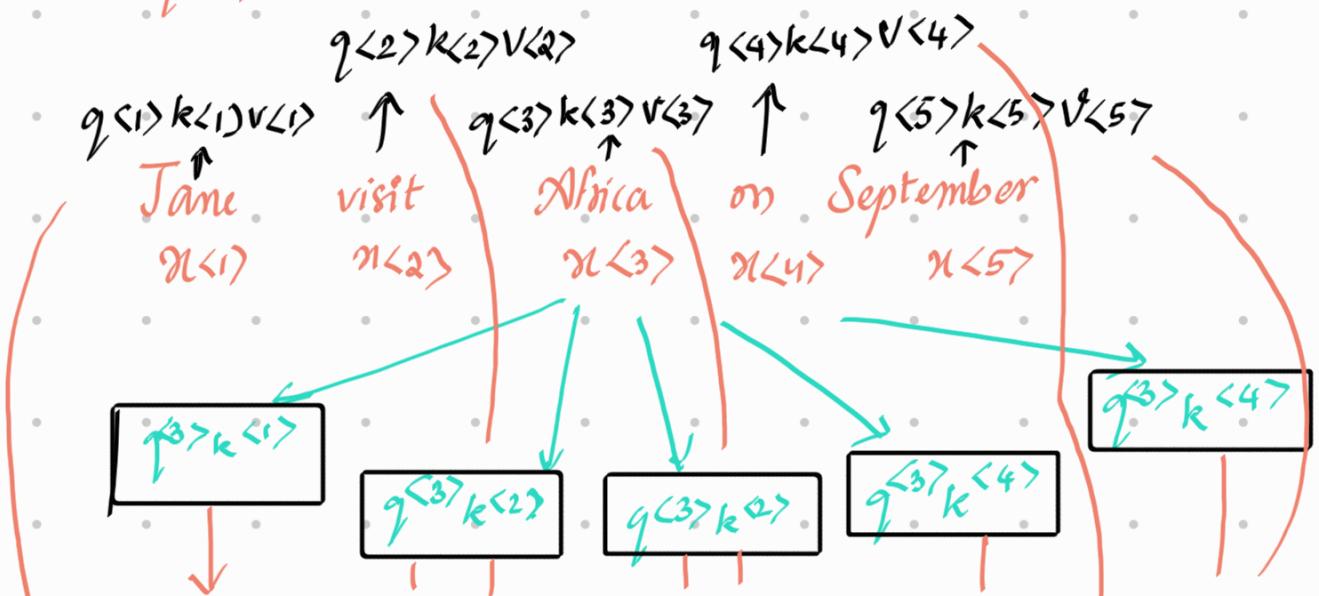
RNN attention

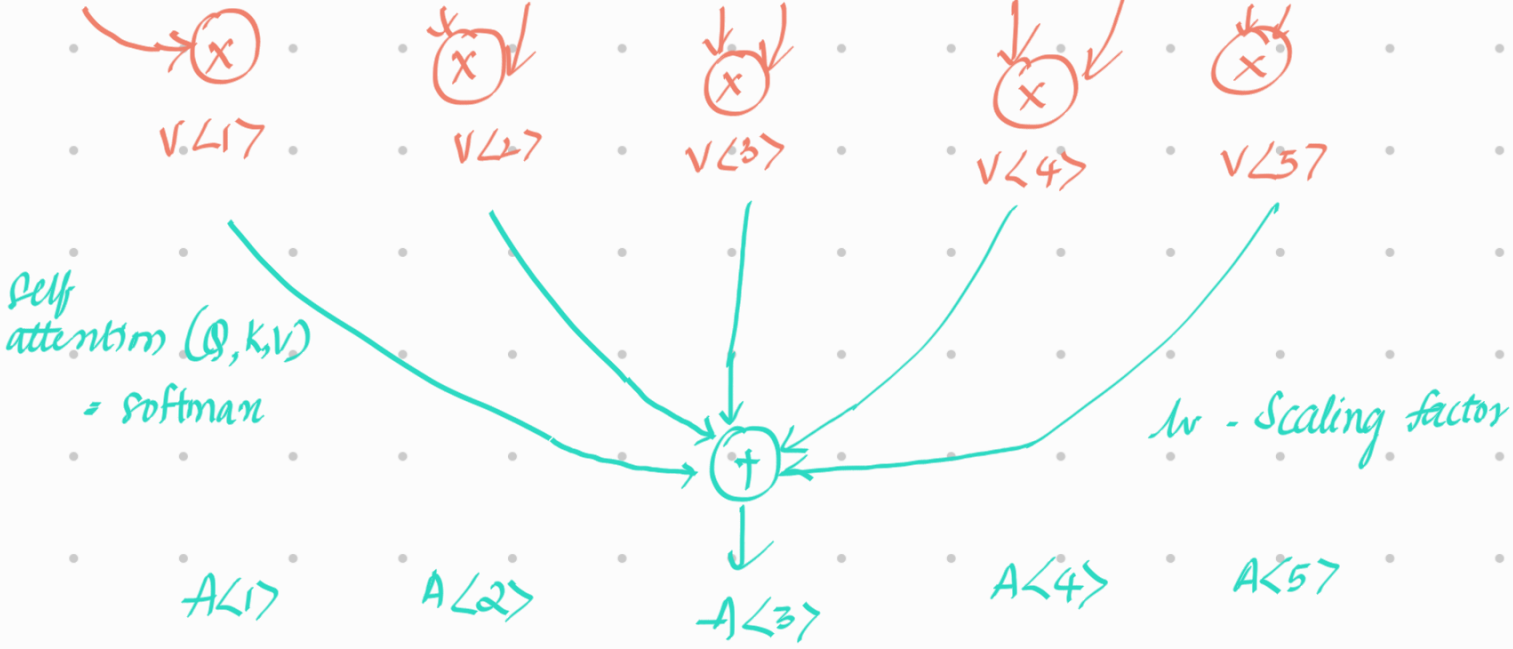
$$\alpha(t,t') = \frac{\exp(e^{<t,t'>})}{\sum_{t=1}^{tx} \exp(e^{<t,t'>})}$$

Transformer Attention

$$A(q,k,v) = \sum_i \frac{\exp(q \cdot k^{<i>})}{\sum \exp(q \cdot k^{<j>})} v^{<i>}$$

Q	K	V
Query	Key reason	Value
$q^{<1>}$	$k^{<1>}$ action	$v^{<1>}$
$q^{<2>}$	$k^{<2>}$	$v^{<2>}$
what's happening		
$q^{<3>}$	$k^{<3>}$	$v^{<3>}$
$q^{<4>}$	$k^{<4>}$	$v^{<4>}$
$q^{<5>}$	$k^{<5>}$	$v^{<5>}$





$$A(q,K,V) = \sum_i \frac{\exp(q\,k^{<i>})}{\sum_j \exp(q\,k^{<j>})} V^{<i>}$$

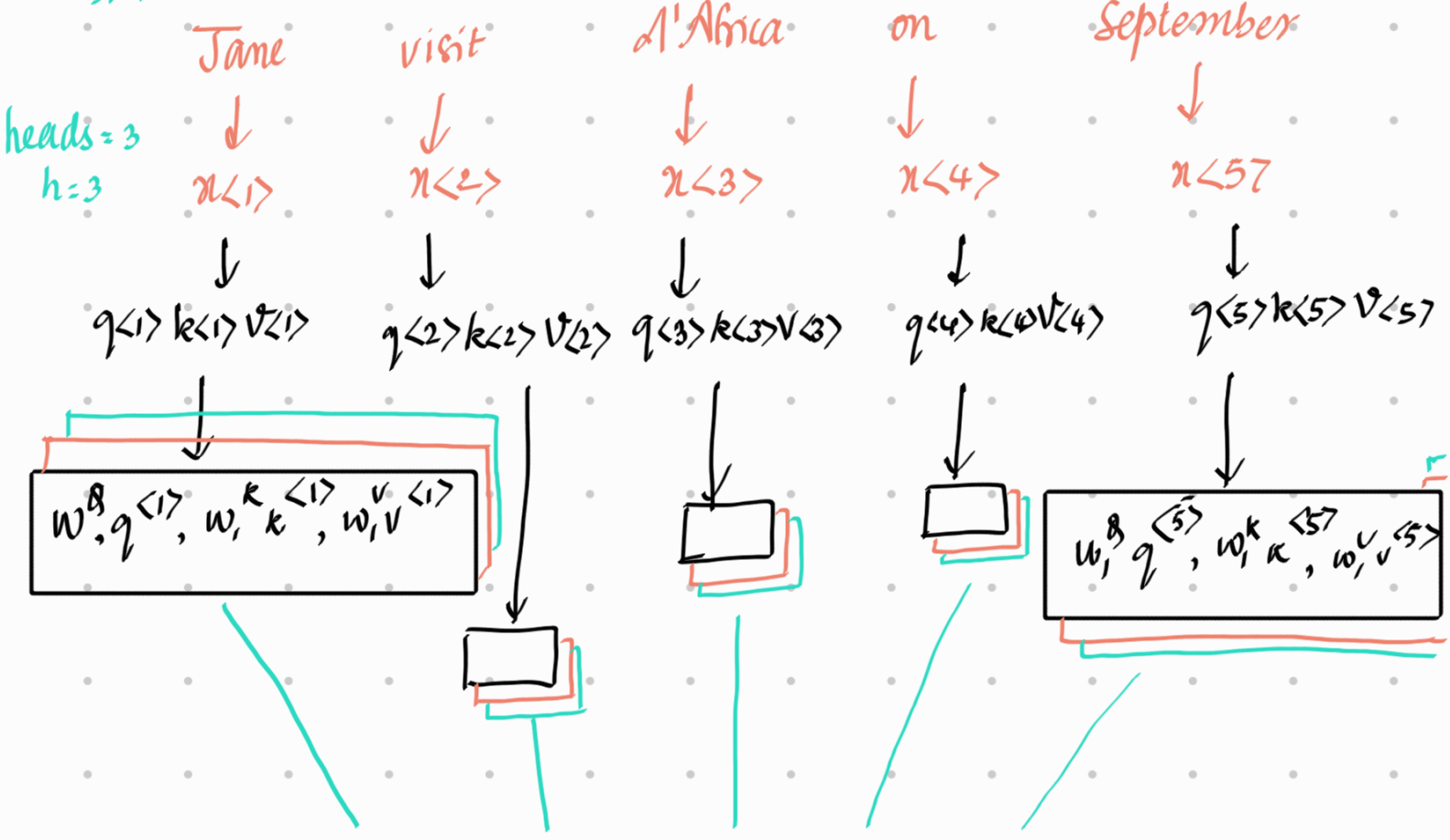
self attention (or) scaled dot product attention

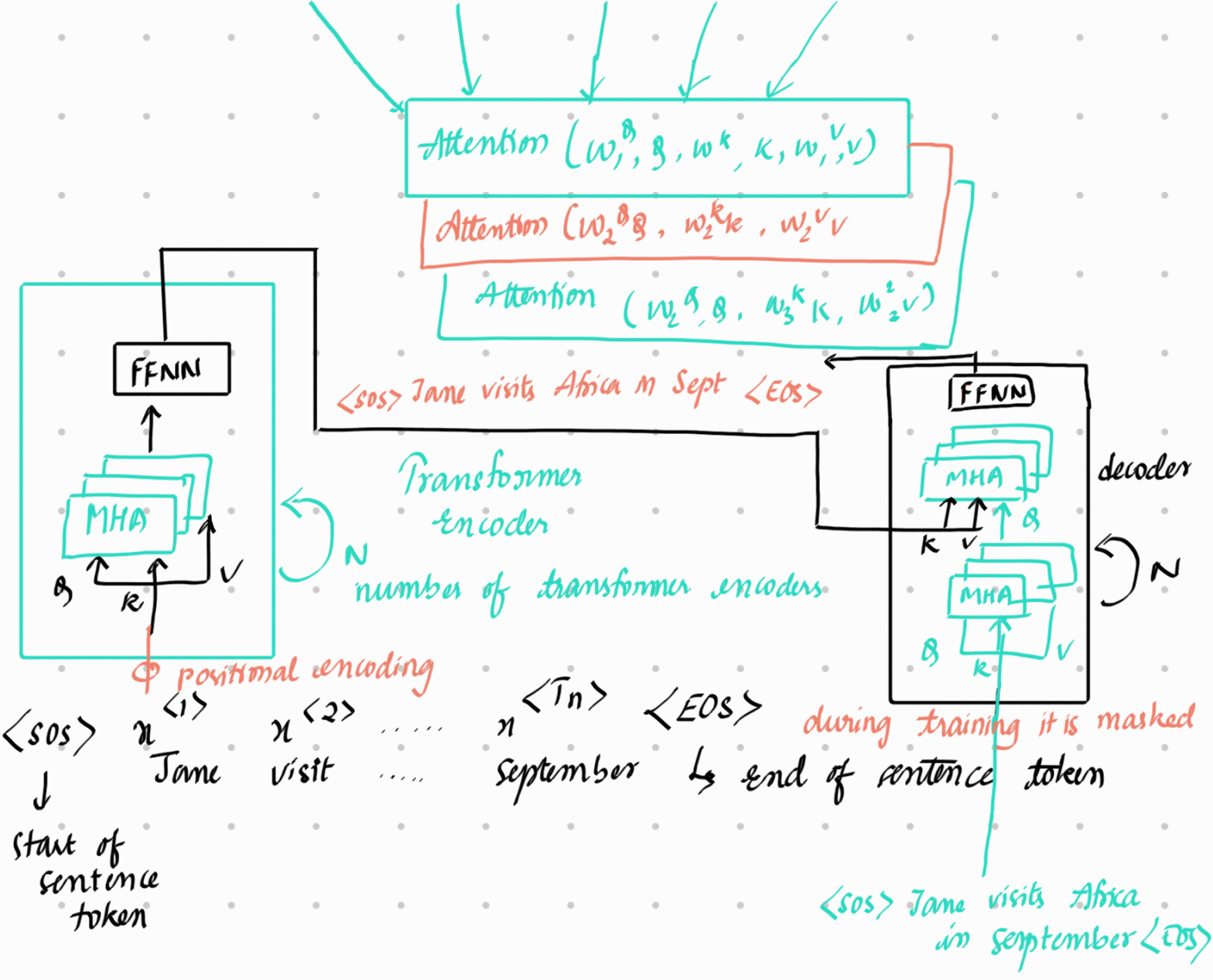
MHA : multihead Attention

Q,K,V

heads = 3

h = 3




Attention $(w_1^q q, w^k k, w_1^v v)$
Attention $(w_2^q q, w_2^k k, w_2^v v$
Attention $(w_2^q q, w_3^k k, w_2^2 v)$
FFNN
MHA
q
k
v
<sos> Jane visits Africa in Sept <EOS>
Transformer encoder
N
number of transformer encoders
FFNN
MHA
MHA
decoder
N
Φ positional encoding
<sos> $x^{<1>}$ $x^{<2>}$ ..... $x^{<Tn>}$ <EOS>
Jane visit ..... September
end of sentence token
during training it is masked
start of sentence token
<sos> Jane visits Africa in September <EOS>