

Output:

$p(s_1)$

$p(s_2)$

...

$p(s_n)$

Softmax

Generator

Layer normalization

Feed Forward network ( $w_2$ )

Feed Forward network ( $w_1$ )  
with activation

Layer normalization

self-attention

key

value

query

Multi-head attention

out

positional encoding

Input:

trace 1

[start]

[m]

$s_2$

...

$s_{k-2}$

$s_{k-1}$

[stop]

trace 2

[start]

$s_1$

$s_2$

...

[m]

$s_{k-1}$

[stop]