Transformer Architecture

Output probabilities

↑

Softmax

↑

Linear

↑

Add & Norm

↑

Feed Forward

Add & Norm

Add & Norm

Feed forward

Multi – Head Attention

↑ ↑ ↑

Add & Norm

Add & Norm

Multi – Head Attention

Masked Multi – Head Attention

Positional embedding ⊕

Positional embedding ⊕

Inputs embedding

outputs embedding

Inputs

Outputs

(shifted right)