



$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

1

$seq$  = sequence length

$d_{model}$  = size of the embedding vector

$h$  = number of heads

$d_k = d_v$  =  $d_{model} / h$