Computation of Self-Attention

Intuition behind Self Attention

"The kids were scared of the lions, so they left right away."

Intuition behind Self Attention

"The kids were scared of the lions, so they left right away."

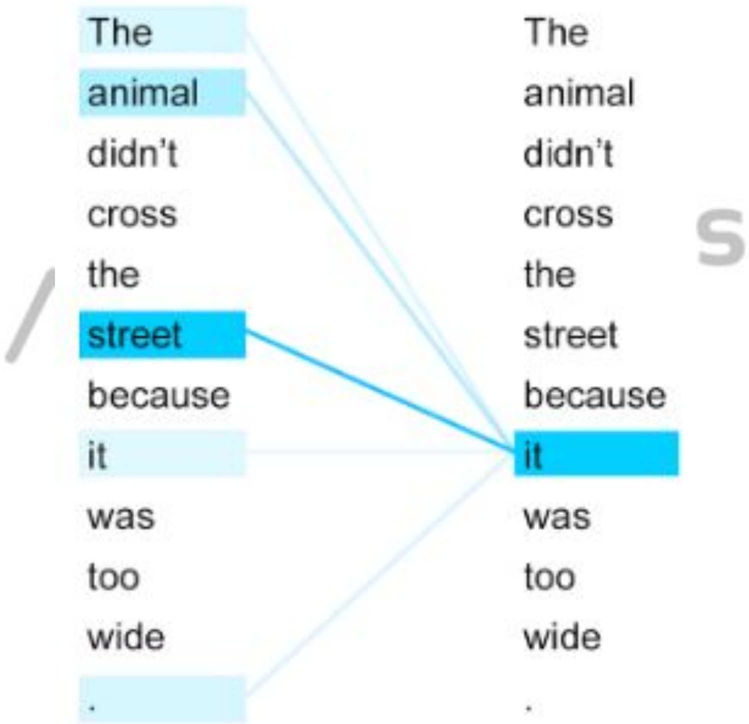
?

?

Intuition behind Self Attention

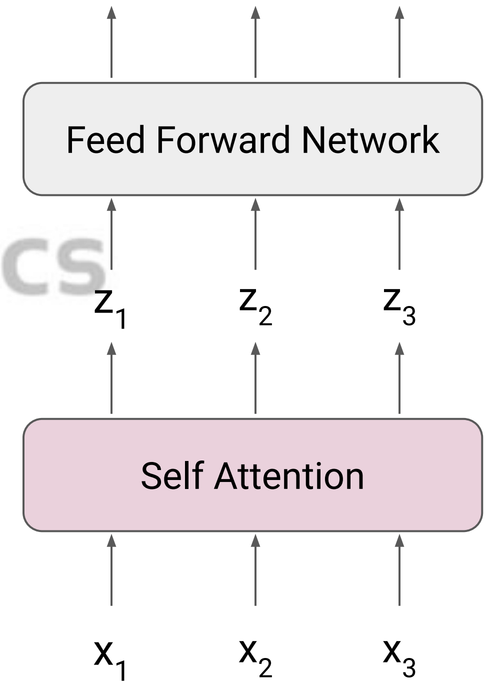
"The kids were scared of the lions, so they left right away."

Self-Attention



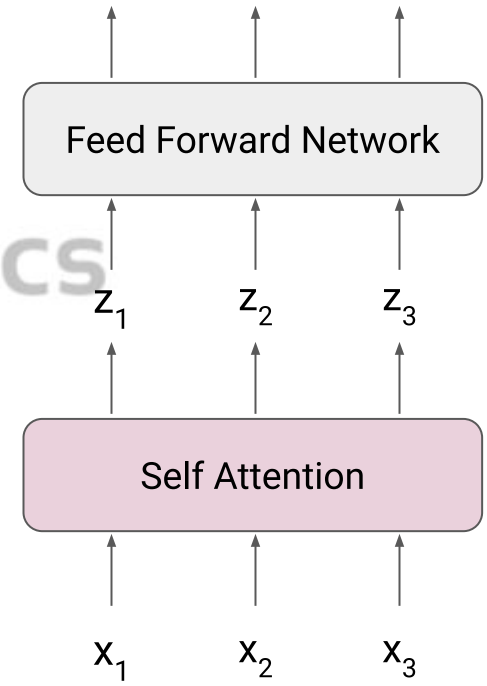
Self-Attention

- Each encoder or decoder has a self attention layer and a feed forward network



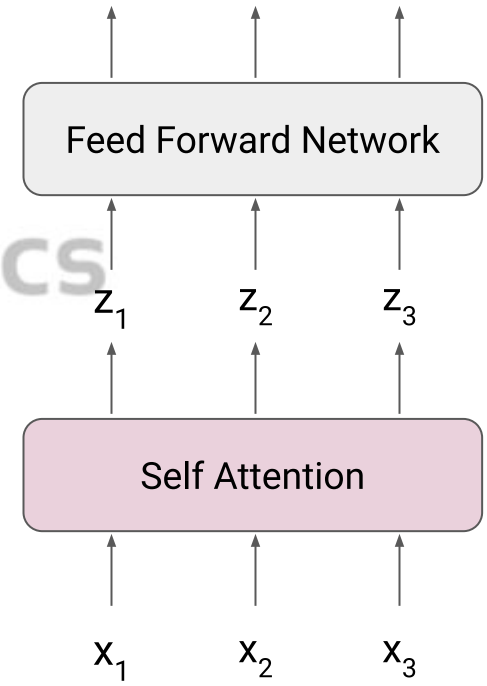
Self-Attention

- Each encoder or decoder has a self attention layer and a feed forward network
- Self attention layer encodes a token by incorporating information from other tokens



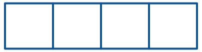
Self-Attention

- Each encoder or decoder has a self attention layer and a feed forward network
- Self attention layer encodes a token by incorporating information from other tokens
- x_i are the input embeddings and z_i are the outputs of self attention layer



Self-Attention

Embeddings

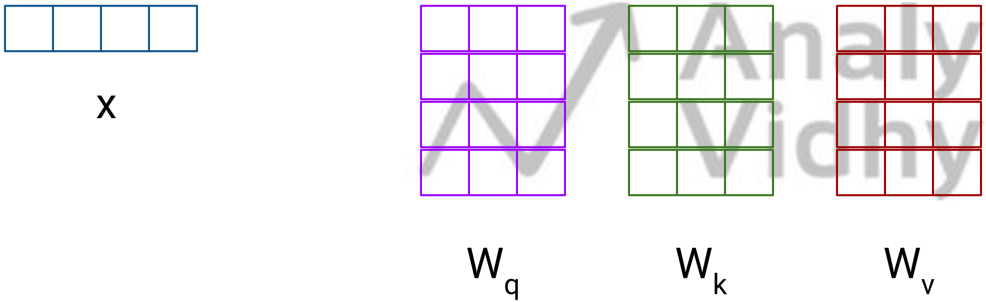


x

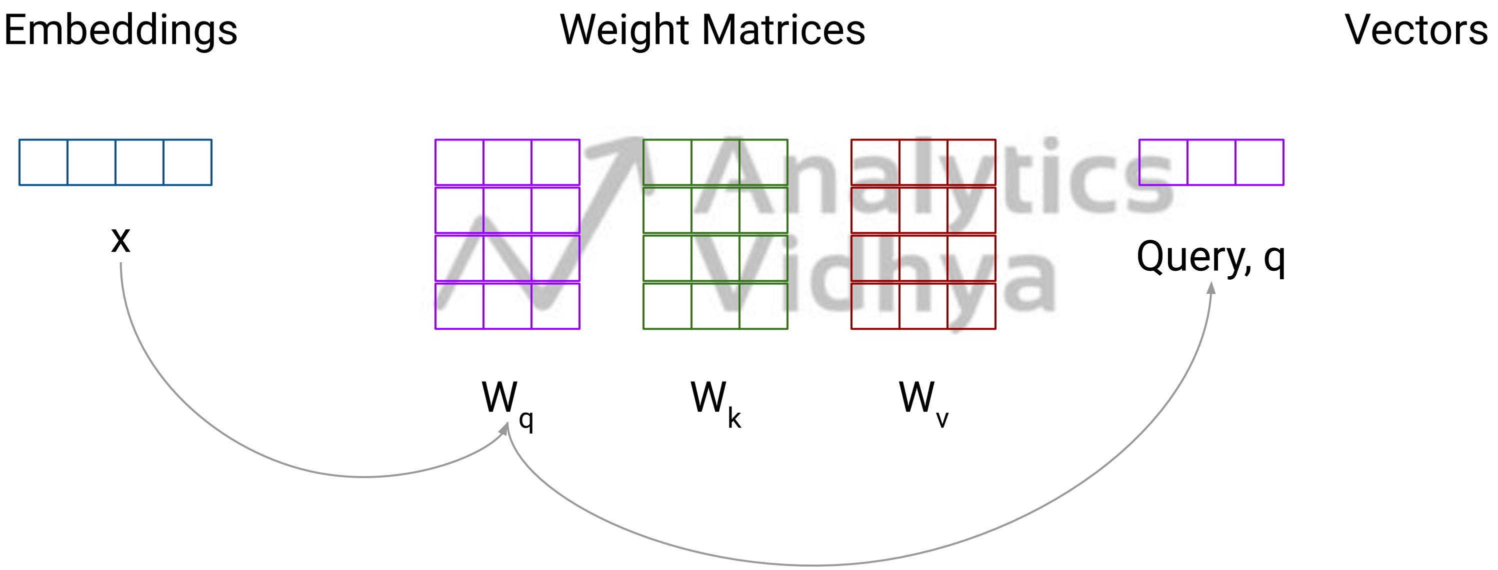
Self-Attention

Embeddings

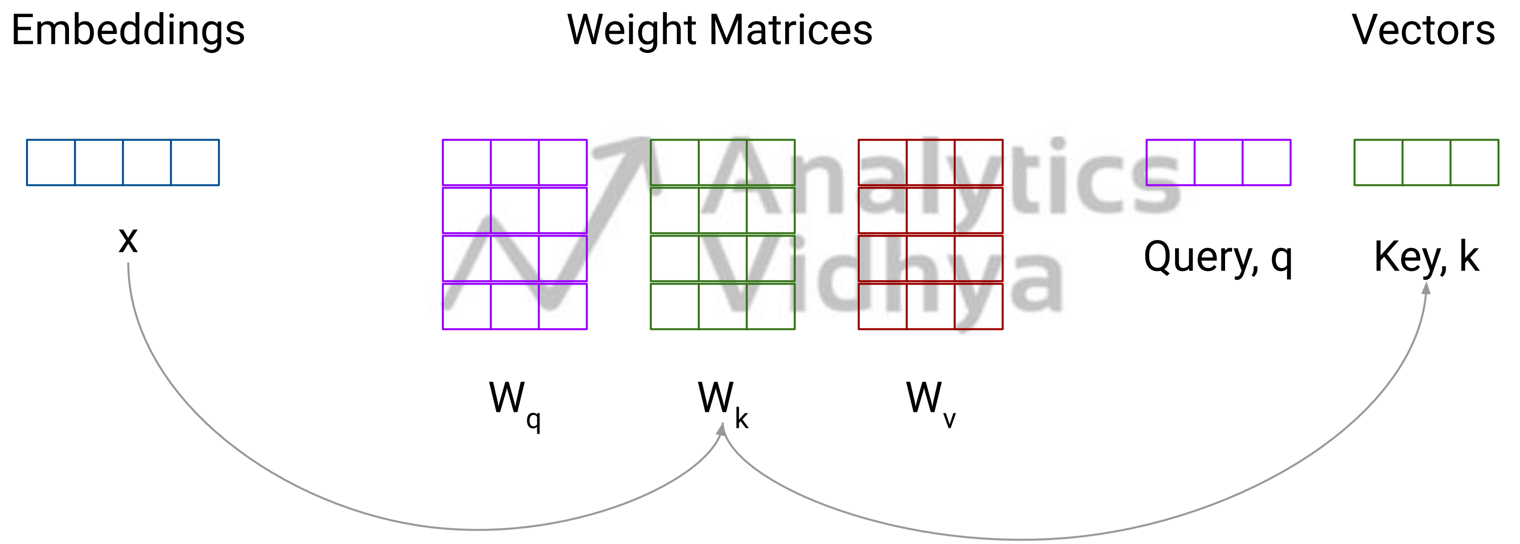
Weight Matrices



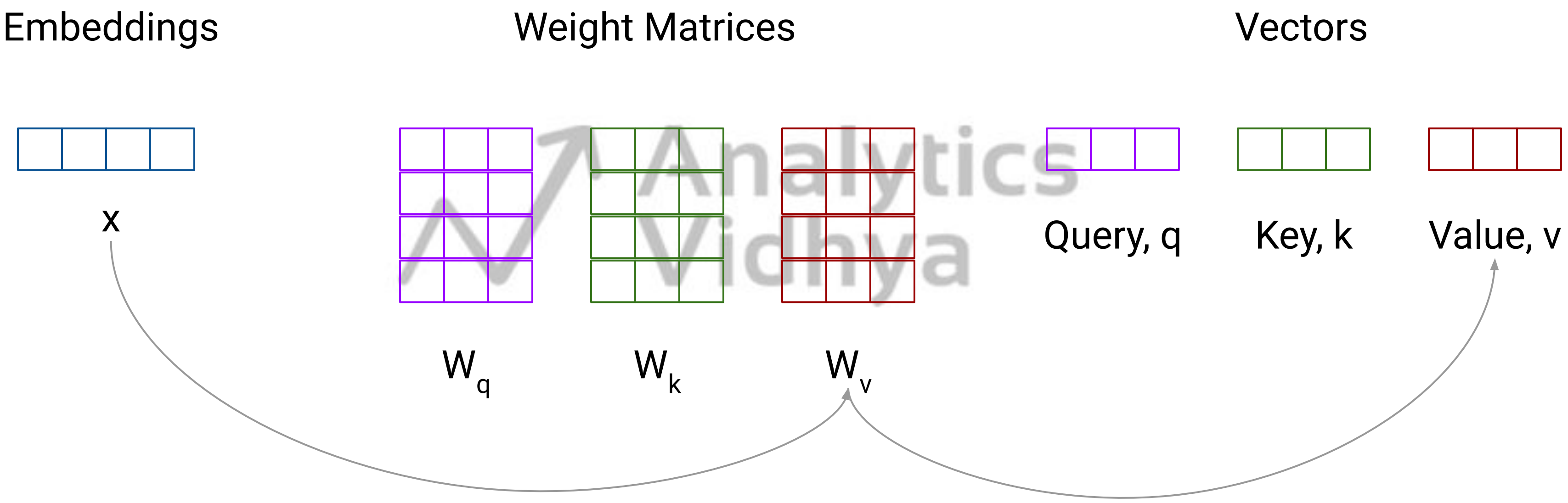
Self Attention Layer



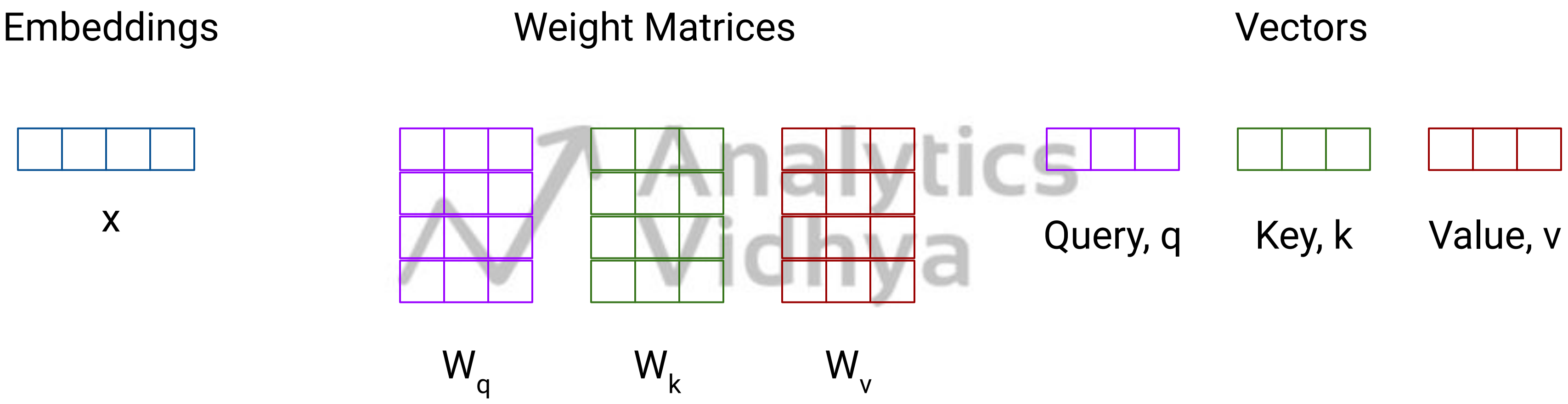
Self Attention Layer



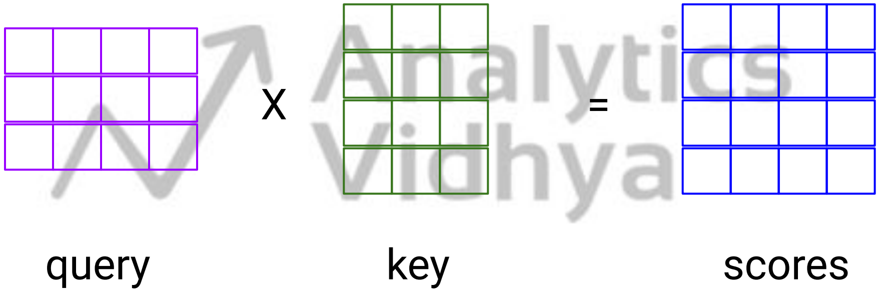
Self Attention Layer



Self Attention Layer



Concepts of Query, Key, and Value



Concepts of Query, Key, and Value

	have	a	nice	day
have	96	23	27	11
a	23	98	35	28
nice	27	35	91	56
day	11	28	56	93

scores

Self Attention Layer: Computations

Word	q vector	k vector	v vector	score
Action	q_1	k_1	v_1	$q_1 . k_1$
gets		k_2	v_2	$q_1 . k_2$
results		k_3	v_3	$q_1 . k_3$

Self Attention Layer: Computations

Word	q vector	k vector	v vector	score
Action		k_1	v_1	$q_2 \cdot k_1$
gets	q_2	k_2	v_2	$q_2 \cdot k_2$
results		k_3	v_3	$q_2 \cdot k_3$

Self Attention Layer: Computations

Word	q vector	k vector	v vector	score
Action		k_1	v_1	$q_3 . k_1$
gets		k_2	v_2	$q_3 . k_2$
results	q_3	k_3	v_3	$q_3 . k_3$

Self Attention Layer: Computations

Word	q vector	k vector	v vector	score	score / 8
Action	q_1	k_1	v_1	$q_1 . k_1$	$q_1 . k_1 / 8$
gets		k_2	v_2	$q_1 . k_2$	$q_1 . k_2 / 8$
results		k_3	v_3	$q_1 . k_3$	$q_1 . k_3 / 8$

Self Attention Layer: Computations

Word	q vector	k vector	v vector	score	score / 8	Softmax
Action	q_1	k_1	v_1	$q_1 . k_1$	$q_1 . k_1 / 8$	x_{11}
gets		k_2	v_2	$q_1 . k_2$	$q_1 . k_2 / 8$	x_{12}
results		k_3	v_3	$q_1 . k_3$	$q_1 . k_3 / 8$	x_{13}

Self Attention Layer: Computations

Word	q vector	k vector	v vector	score	score / 8	Softmax	Softmax * v	Sum
Action	q_1	k_1	v_1	$q_1 . k_1$	$q_1 . k_1 / 8$	x_{11}	$x_{11} * v_1$	z_1
gets		k_2	v_2	$q_1 . k_2$	$q_1 . k_2 / 8$	x_{12}	$x_{12} * v_2$	
results		k_3	v_3	$q_1 . k_3$	$q_1 . k_3 / 8$	x_{13}	$x_{13} * v_3$	

Self Attention Layer: Computations

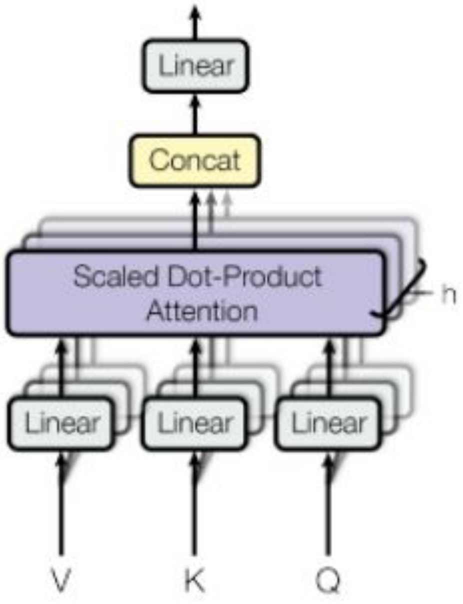
Word	q vector	k vector	v vector	score	score / 8	Softmax	Softmax * v	Sum#
Action		k_1	v_1	$q_2 . k_1$	$q_2 . k_1 / 8$	x_{21}	$x_{21} * v_1$	
gets	q_2	k_2	v_2	$q_2 . k_2$	$q_2 . k_2 / 8$	x_{22}	$x_{22} * v_2$	z_2
results		k_3	v_3	$q_2 . k_3$	$q_2 . k_3 / 8$	x_{23}	$x_{23} * v_3$	

Self Attention Layer: Computations

Word	q vector	k vector	v vector	score	score / 8	Softmax	Softmax * v	Sum#
Action		k_1	v_1	$q_3 . k_1$	$q_3 . k_1 / 8$	x_{31}	$x_{31} * v_1$	
gets		k_2	v_2	$q_3 . k_2$	$q_3 . k_2 / 8$	x_{32}	$x_{32} * v_2$	
results	q_3	k_3	v_3	$q_3 . k_3$	$q_3 . k_3 / 8$	x_{33}	$x_{33} * v_3$	z_3

Multi-headed Self-Attention

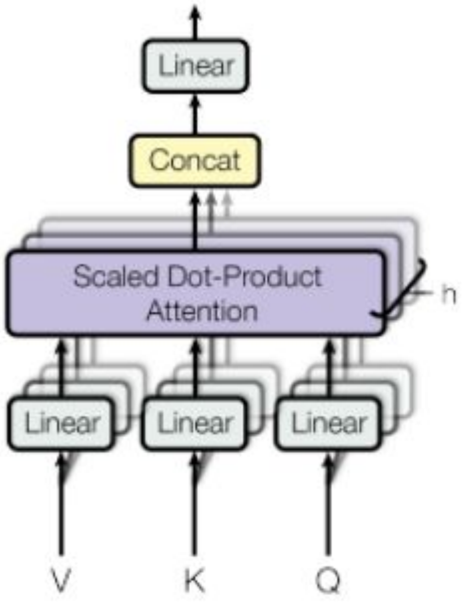
- Multiple sets of W_q , W_k , W_v and query, key and value vectors..



Multi-Head Attention

Multi-headed Self-Attention

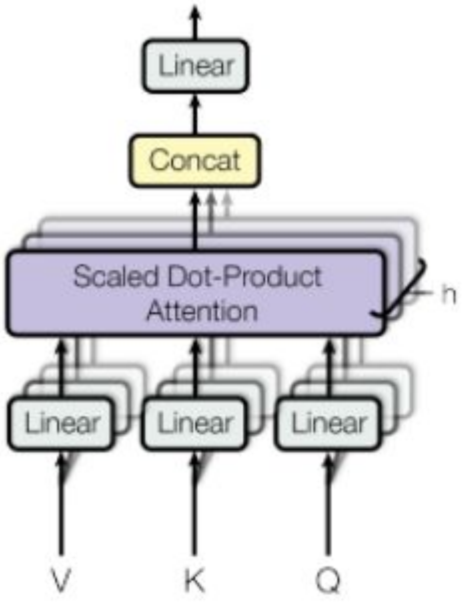
- Multiple sets of W_q , W_k , W_v and query, key and value vectors.
- Transformer uses 8 self-attention heads.



Multi-Head Attention

Multi-headed Self-Attention

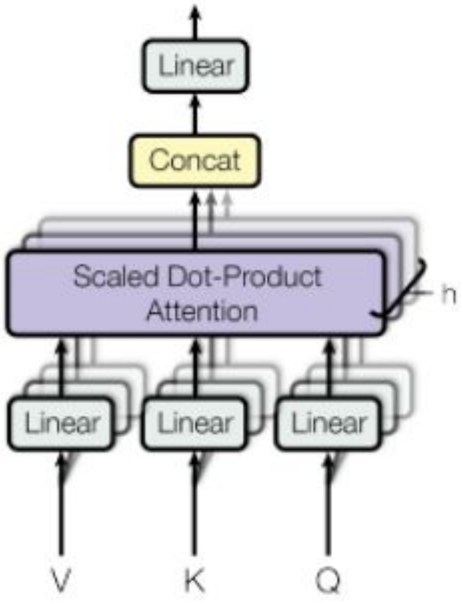
- Multiple sets of W_q , W_k , W_v and query, key and value vectors.
- Transformer uses 8 self-attention heads.
- Each head represents the input embeddings into a different representation space.



Multi-Head Attention

Multi-headed Self-Attention

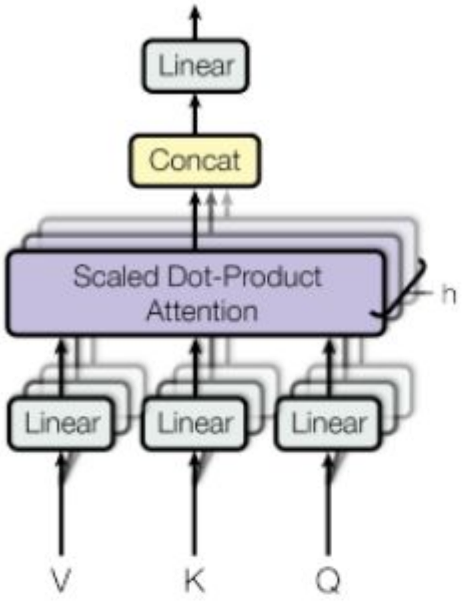
- Multiple sets of W_q , W_k , W_v and query, key and value vectors.
- Transformer uses 8 self-attention heads.
- Each head represents the input embeddings into a different representation space.
- $(q_0, k_0, v_0), (q_1, k_1, v_1), (q_2, k_2, v_2), \ldots, (q_7, k_7, v_7)$



Multi-Head Attention

Multi-headed Self-Attention

- Each head produces a Z-score matrix (Z_0, Z_1, Z_2,..., Z_7)
- These Z matrices are concatenated and multiplied with another weight matrix W to arrive at the final Z_f matrix.



Multi-Head Attention

Thank You