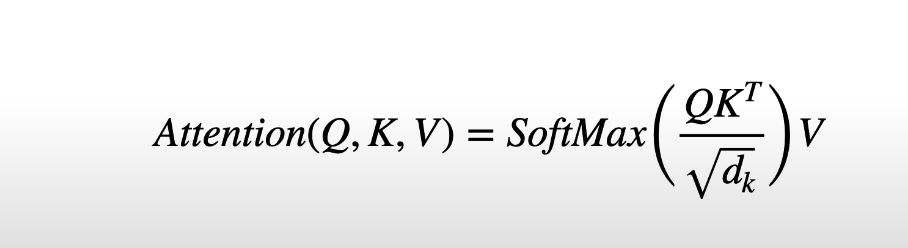
**Self Attention**



Calculates similarities for each word between each word and itself and all of the other words.

Each word in the prompt is converted to word embeddings. We then add positional encodings to the word embeddings. In the transformer paper each word embedding has a dimension of 512. We then stack these embeddings and multiply it by respective weights.

Consider the following sentence.

**Hello world, I live in Pune.**

Above sentence has 6 words, embedding dimension after adding positional embedding would be

6 \* 512. We then create 3 copies of the resultant embedding and this forms Querry, Key and Value matrix. Each of these matrices have separate weights of dimension of weights would be 512 \* 512. Multiplying them would result in 6\* 512 dimension. The weight matrix is trainable and is initialized with glorot uniform.

Q.K^T.V

(6\*512) \*(512\*6)=> (6\*6)

(6\*6)\*(6\*512)=>(6\*512)