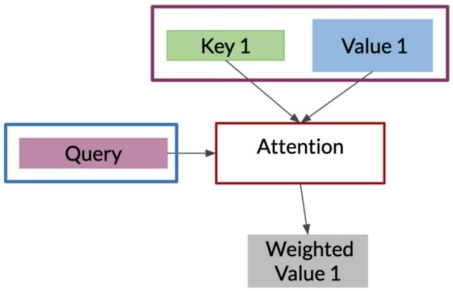
Attention

The previous video covered one of the most important concepts covered in this course. If you did not get what I was saying right away, don't worry about it. This reading will explain everything I mentioned in detail.

The attention mechanism uses encoded representations of both the input or the encoder hidden states and the outputs or the decoder hidden states. The **keys** and **values** are pairs. Both of dimension N, where N is the input sequence length and comes from the encoder hidden states.



Keys and values have their own respective matrices, but the matrices have the same shape and are often the same. While the queries come from the decoder hidden states. One way you can think of it is as follows. Imagine that you are translating English into German. You can represent the word embeddings in the English language as keys and values. The queries will then be the German equivalent. You can then calculate the dot product between the query and the key. Note that similar vectors have higher dot products and non-similar vectors will have lower dot products. The intuition here is that you want to identify the corresponding words in the queries that are similar to the keys. This would allow your model to "look" or focus on the right place when translating each word. We then run a softmax

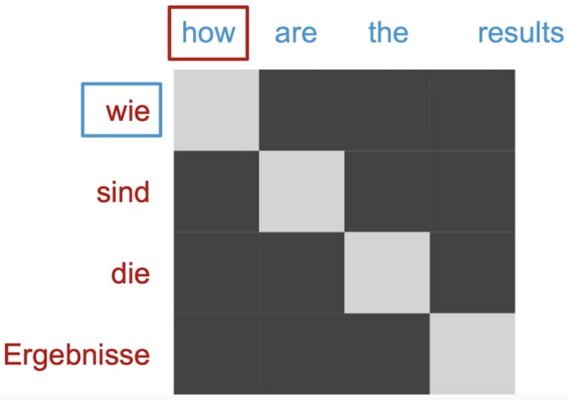
$$softmax(QK^T) \tag{1}$$

That allows us to get a distribution of numbers between 0 and 1. One more step is required. We then would multiply the output by V. Remember V in this example was the same as our **keys,** corresponding to the English word embeddings. Hence the equation becomes

$$softmax(QK^T)V \tag{2}$$

This tells us how much of each word, or which combination of words we will be feeding into our decoder to predict the next German word. This is called scale dot product attention. Note we add a normalization constant to it later, but you do not have to worry about that right now.

Here's an example that shows where the model is looking when translating between English and German, as you'll be doing in your assignments. The attention model should still be able to find a connection between the two words.



In the matrix, the lighter square shows where the model is actually looking when making the translation of that word. This mapping should not necessarily be one to one. The lighting just tells you to what extent is each word contributing to the input that will be fed into the decoder. As you can see several words can contribute to translating another word, depending on the weights (output) of the softmax that will be used to create the new input.