Parallelization of self-attention at training time

RNNs:

\[ h_n = f(W_h h_{n-1} + W_e c_n) \]

Each hidden state is a direct fn of the previous hidden state.

Sequential computation in RNNs allows us to model word order.
there is no dependency between $z_n$ and $z_{n-1}$!

we no longer have to compute
the $z_n$ one at a time.

how do we parallelize the attention computation?

\[ \text{att} \text{n scores} \]

\[ a_1 = \langle q_1, k_1 \rangle \]

\[ a_2 = \langle q_2 k_1, q_2 k_2 \rangle \]

\[ a_3 = \langle q_3 k_1, q_3 k_1, q_3 k_3 \rangle \]

\[ q_1 \]

\[ q_2 \]

\[ q_3 \]

\[ k_1 \]

\[ k_2 \]

\[ k_3 \]

\[ \text{X} \]

\[ \text{X} \]

\[ \text{X} \]

\[ \text{X} \]
After masking, we apply softmax and then we get the valid attention distribution without any "cheating".