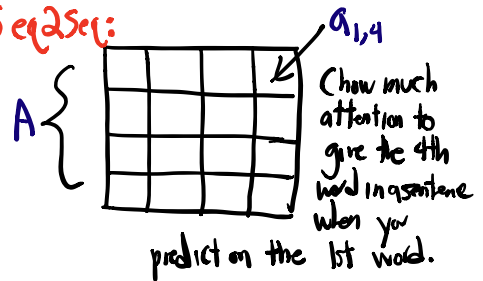


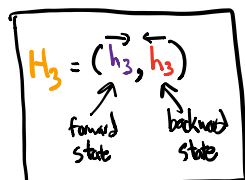
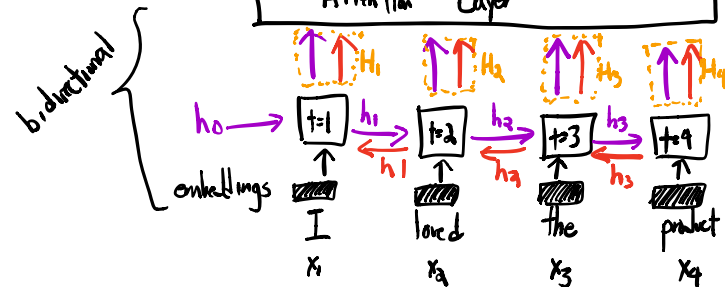
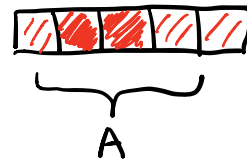
$A = S \times S$ (sequence length by sequence length)

This case, $S=4$.

Seq2Seq:



Classification:



H_3 is the concatenation of the hidden states from the bi-directional encoder.

h_1 = the hidden state at $t=1$ propagated from the end of the sequence to the beginning

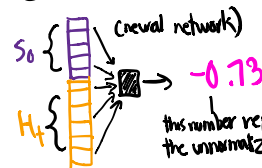
h_1 = the hidden state at $t=1$ propagated from the beginning to end of the sequence

C_1 = weighted sum of the entire sequence features $\rightarrow H_1, H_2, H_3, \dots, H_4$

$$C_1 = \sum_t a_{1,t} H_t$$

$$a_{1,t} = \frac{\exp(e_{1,t})}{\sum_t \exp(e_{1,t})}$$

$e_{1,t} = s_0$ (hidden state from previous timestep in decoder)
 H_t (features from the encoder layer)



this number represents the unnormalized amount of attention you want to pay to timestep t while making predictions for step 1

Calculating Attention Values

