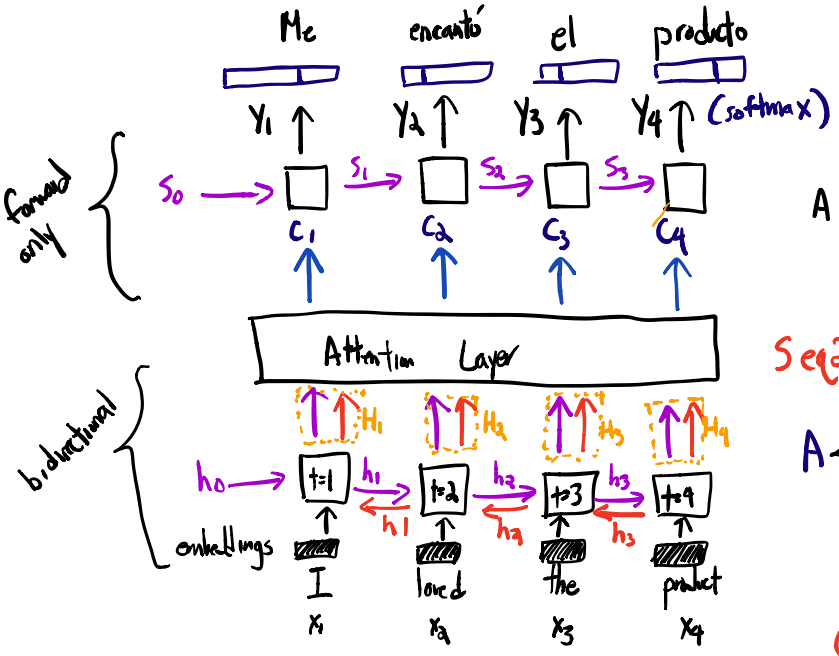
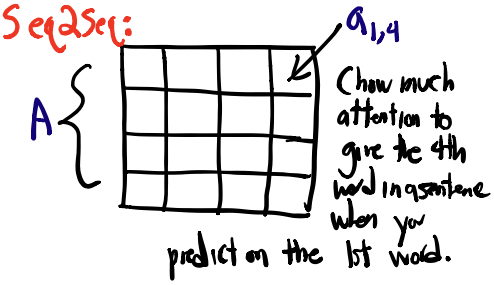
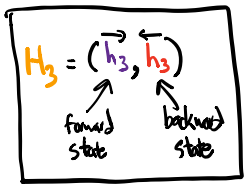


A = S x S (sequence length by sequence length)

This case, S=4.

Seq2Seq:



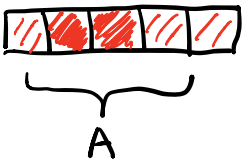


H_3 is the concatenation of the hidden states from the bidirectional encoder.

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

Classification:



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

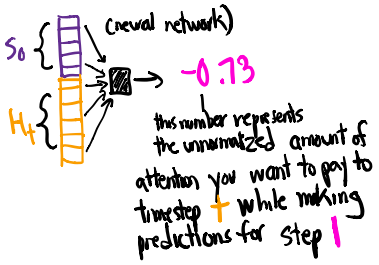
$$C_1 = \sum_t a_{1,t} H_t$$

$$a_{1,t} = \frac{\exp(e_{1,t})}{\sum_{t'} \exp(e_{1,t'})}$$

$e_{1,t} = S_0$ H_t

S_0: Hidden state from previous timestep in decoder

H_t: features from the encoder layer



Calculating Attention Values