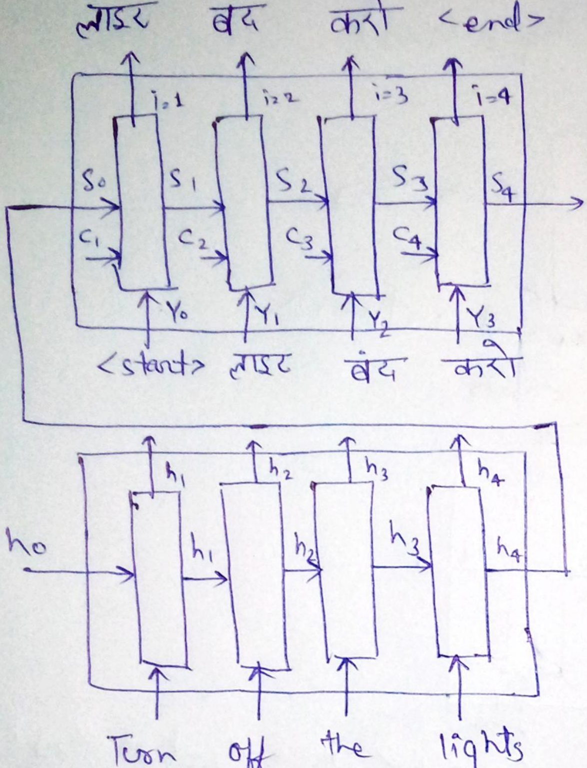
06 March 2025

ATTENTION MECHANISM



i = time step

C_i = attention input

in vanilla encoder decoder

input = $[Y_{i-1}, S_{i-1}]$

in vanilla encoder-decoder with attention mechanism

input = $[Y_{i-1}, S_{i-1}, C_i]$

$C_1 = \alpha_{11} h_1 + \alpha_{12} h_2 + \alpha_{13} h_3 + \alpha_{14} h_4$

$\alpha_i \rightarrow$ weight (scalar)

$h_j \rightarrow$ encoder's hidden state (vector)

Same for other attention input

$$C_i = \sum \alpha_{ij} h_j$$

hence

$$C_1 = \alpha_{11} h_1 + \alpha_{12} h_2 + \alpha_{13} h_3 + \alpha_{14} h_4$$

$$C_2 = \alpha_{21} h_1 + \alpha_{22} h_2 + \alpha_{23} h_3 + \alpha_{24} h_4$$

$$C_3 = \alpha_{31} h_1 + \alpha_{32} h_2 + \alpha_{33} h_3 + \alpha_{34} h_4$$

$$C_4 = \alpha_{41} h_1 + \alpha_{42} h_2 + \alpha_{43} h_3 + \alpha_{44} h_4$$

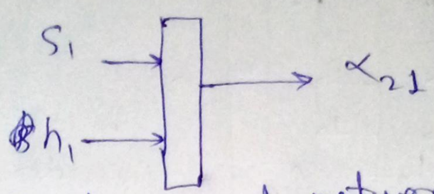
Now how to calculate α?

lets take an example α_{21}

$\alpha_{21} \rightarrow$ alignment/similarity score

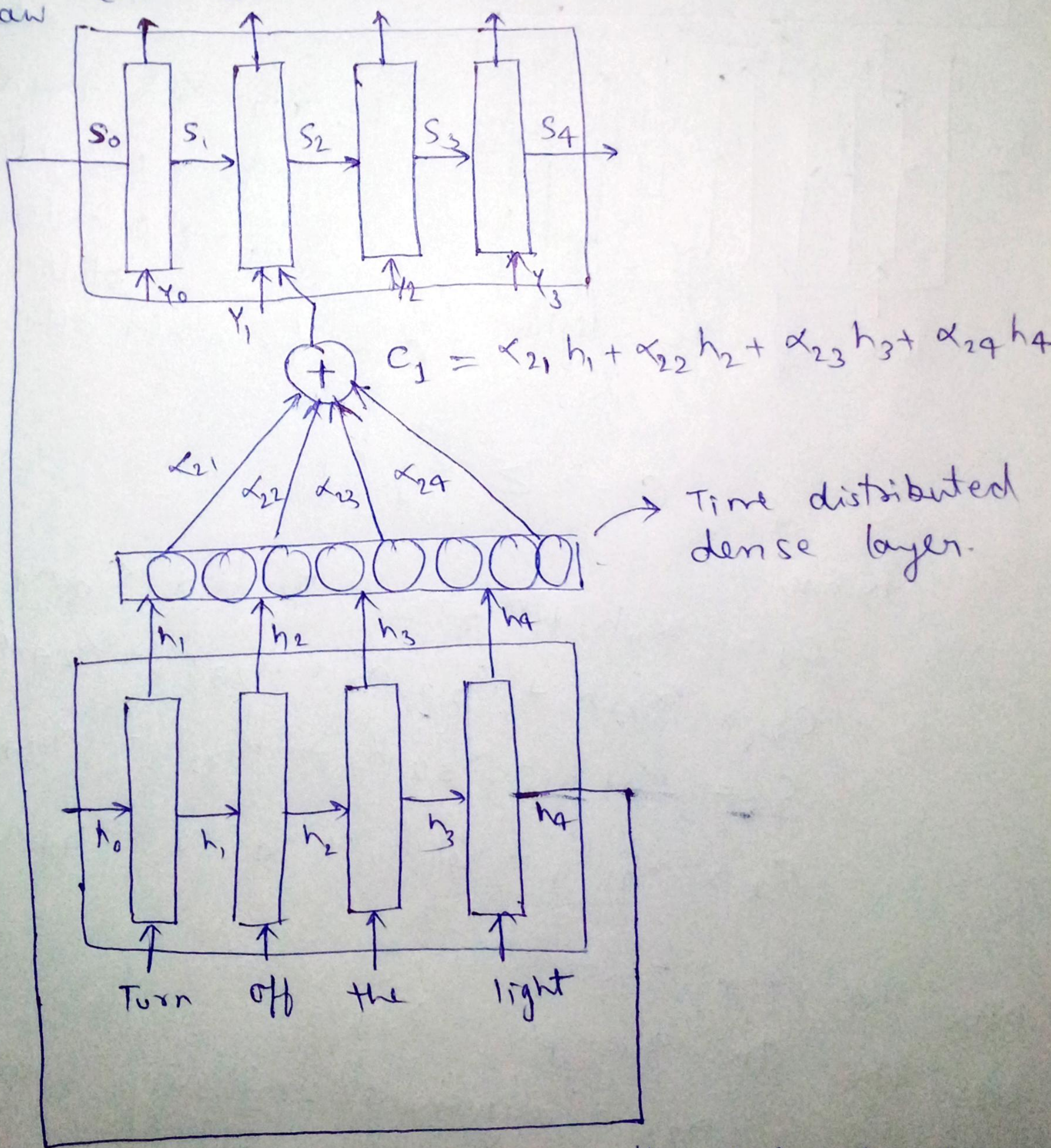
- α_{21} depends on h_1 and S_1 (previous hidden state of decoder)
- $\alpha_{21} \rightarrow f(h_1, S_1)$

  or $\alpha_{ij} = f(h_j, S_{i-1})$



artificial neural network.

assume we are now at i = 2



$c_1 = \alpha_{21}h_1 + \alpha_{22}h_2 + \alpha_{23}h_3 + \alpha_{24}h_4$

Time distributed dense layer.

- in original paper researcher use the. bidirectional LSTM