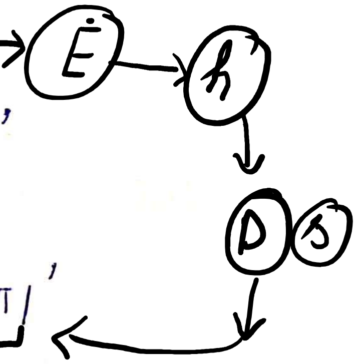
Attention mechanism



Let us consider a statement – 'reason with logic.'
[Source statement in English]

Translated statement in Hindi – 'तर्क के साथ कारण'

Here $T_x = 3$ & $T_y = 4$

We associate a hidden state for each word of English statement

$h_t = g(W h_{t-1} + U x_t + b)$

Word	hidden state	Encoding
reason	h_1^E	[0.1, 0.7, 0.5, 0.3]
with	h_2^E	[-0.1, -0.7, 0.4, 0.6]
logic	h_3^E	[0.7, 0.2, -0.3, 0.4]

Here h_i^E is a vector of 4 dimensions.

We know we can get h_i^E as below.

$$\boxed{h_i^E = f(W_E h_{i-1}^E + U x_i + b)}$$

Here W_E is of order 4×4
h_{i-1}^E is of order 4×1
U is of order $4 \times V_{in}$
x_i is of order $V_{in} \times 1$
b is of order 4×1
f is an activation function [like tanh, sigmoid etc]

V_{in} is input Vocab

30,000 $\begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \rightarrow$ reason

[Note:- In paper, they used $h_t = f(x_t, h_{t-1})$ [Page 2]]
$\hookrightarrow h_i \in \mathbb{R}^n$

⟨300⟩

... to predict the word 'तर्क'.

We suppose that we need to predict the word 'नदी'.

The hidden state be $s_0 = [0.3, 0.4, 0.1]$

[Note :- We used an ~~embed~~ hidden state with dimension = 3]

The hidden state from encoder was of size 4×1

We will consider a matrix of size 3×4 to convert the 4×1 vector h_i to 3×1

$(W_c)_{3\times4} h_1^E \rightarrow 4 \times 1$

The transformed vectors are.

$$\hat{h}_1^E = W_c h_1^E$$
$$\hat{h}_2^E = W_c h_1^E$$
$$\hat{h}_3^E = W_c h_3^E$$

[Here W_c is of order 3×4
W_c will be a trainable matrix]

$\hat{h}_1^E \rightarrow 3 \times 1$ $\quad \hat{h}_1^E, \hat{h}_2^E, \hat{h}_3^E$ $\quad s_0$

We consider W_{dec} & W_{enc} as two matrices such that

$$\boxed{\begin{aligned} W_{enc}\, \hat{h}_1^E + W_{dec}\, s_0 &= z_1 \\ W_{enc}\, \hat{h}_2^E + W_{dec}\, s_0 &= z_2 \\ W_{enc}\, \hat{h}_3^E + W_{dec}\, s_0 &= z_3 \end{aligned}}$$

[Here W_{enc} & W_{dec} are of order 1×1]

We have

$$e_{11} = V_a \operatorname{Tanh}(z_1)$$
$$e_{12} = V_a \operatorname{Tanh}(z_2)$$
$$e_{13} = V_a \operatorname{Tanh}(z_3)$$

[Here V_a is $\underbrace{1 \times 1}$ vector]

Now $$\alpha_{11} = \frac{\exp(e_{11})}{\sum_{j=1}^{T_x} \exp(e_{1j})} = \frac{\exp(e_{11})}{A} = 0.7$$

where $A = \sum_{j=1}^{T_x} \exp(e_{1j})$

$$\alpha_{12} = \exp(e_{12})/A = 0.2$$
$$\alpha_{13} = \exp(e_{13})/A = 0.1$$

Now $$c_1 = \alpha_{11} h_1^E + \alpha_{12} h_2^E + \alpha_{13} h_3^E$$

$$= 0.7(0.1, 0.7, 0.5, 0.3) + 0.2(0.1, -0.7, 0.4, 0.6) + 0.1(0.7, 0.2, -0.3, 0.4)$$

$$\boxed{c_1 = (0.12, 0.37, 0.4, 0.37)}$$

The decoder predictions are done in following way.

$$s_1 = g(y_0, s_0, c_1)$$

<sos>, $s_0 = \begin{pmatrix} 0.3 \\ 0.4 \\ 0.2 \end{pmatrix}$ 3×1

$\begin{pmatrix} \boxed{1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ <sos> 20 000

$s_1 =$

$$s_1 = g\Big(\underbrace{W_D\, s_0} + \underbrace{U_D\, y_0} + \underbrace{W_a\, c_1} + b_D\Big)$$

3x1

4x1

W_a 3x4

[Here; s_0 is of order 3×1
W_D is of order $\underline{3 \times 3}$
U_D is of order $3 \times V_{out}$
y_0 is of order $\underline{V_{out} \times 1}$
c_1 was of order 4×1
W_a is of order 3×4
b_D is of order 3×1]

The prediction y_1 will be done as

$$y_1 = \text{softmax}(\underbrace{V}\, s_1 + c)$$

3x1

b (as)

$V_{out} \times 3$

$\underline{20000} \times 3$

[Here V is of order $V_{out} \times 3$
s_1 is of order 3×1
c is of order $\underbrace{V_{out} \times 1}$]

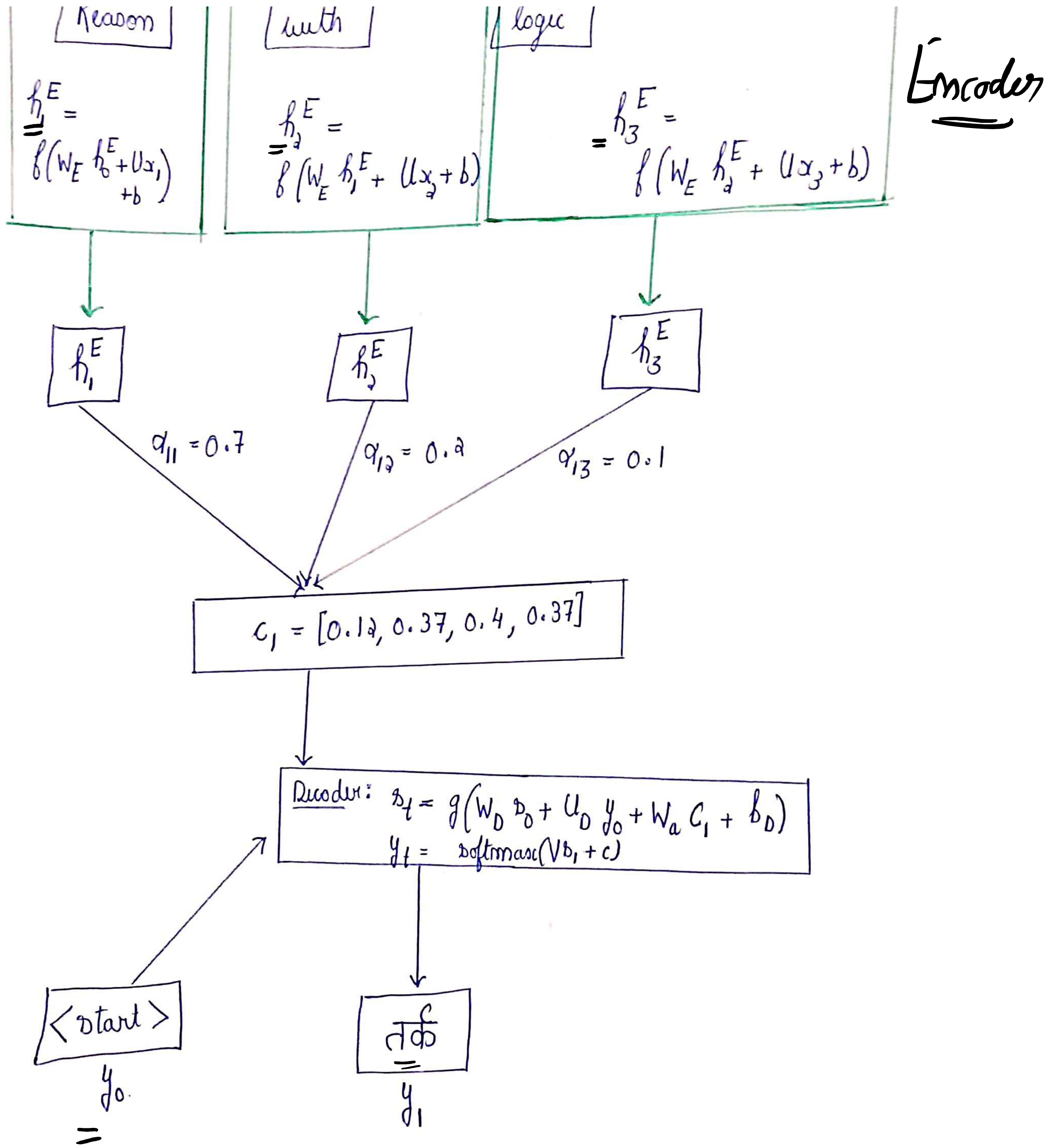
$$\begin{pmatrix} 0.001 \\ 0.002 \\ \vdots \\ \boxed{0.87} \\ 0 \\ 0 \\ 0.07 \end{pmatrix}$$

नको

Now, suppose argmax y_1 = 'नको'

We will then proceed to find y_2

Reason | with | logic


Reason
truth
logic
Encoder
$h_1^E = f(W_E h_0^E + Ux_1 + b)$
$h_2^E = f(W_E h_1^E + Ux_2 + b)$
$h_3^E = f(W_E h_2^E + Ux_3 + b)$
h_1^E
h_2^E
h_3^E
$\alpha_{11} = 0.7$
$\alpha_{12} = 0.2$
$\alpha_{13} = 0.1$
$c_1 = [0.12, 0.37, 0.4, 0.37]$
Decoder: $s_t = g(W_D s_0 + U_D y_0 + W_a C_1 + b_D)$
$y_t = \text{softmax}(Vs_1 + c)$
<start>
y_0
तर्क
y_1