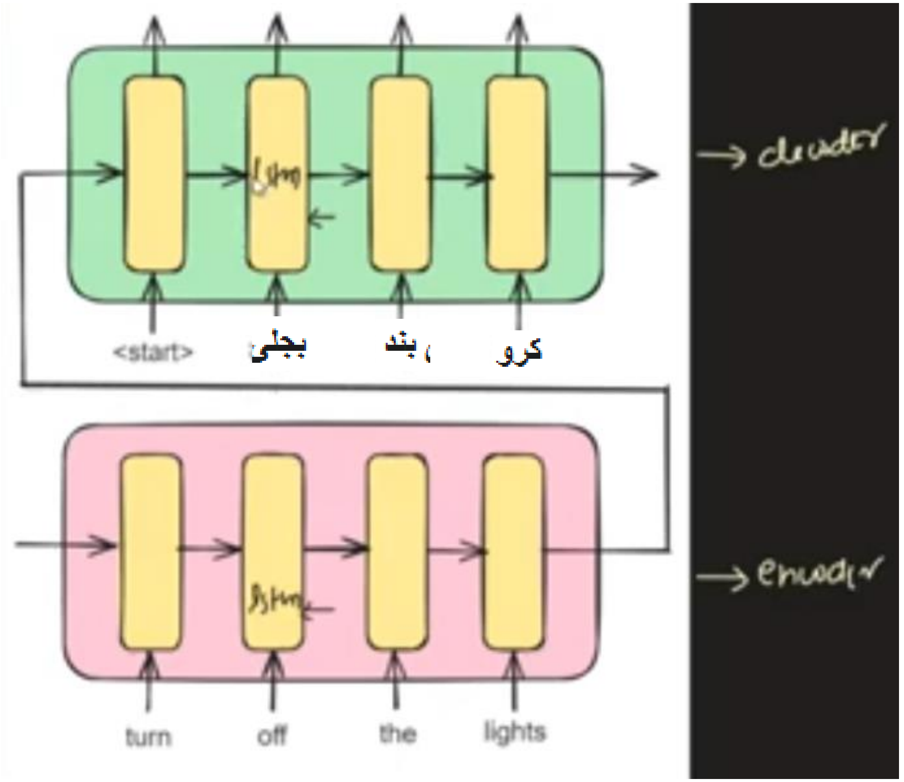
Attention Models

Dr. Muhammad Safyan

Attention Mechanism

- A way to improve the Encoder-Decoder Architecture.
- Problems with Encoder/Decoder



Translate in Urdu

- Translate in Urdu

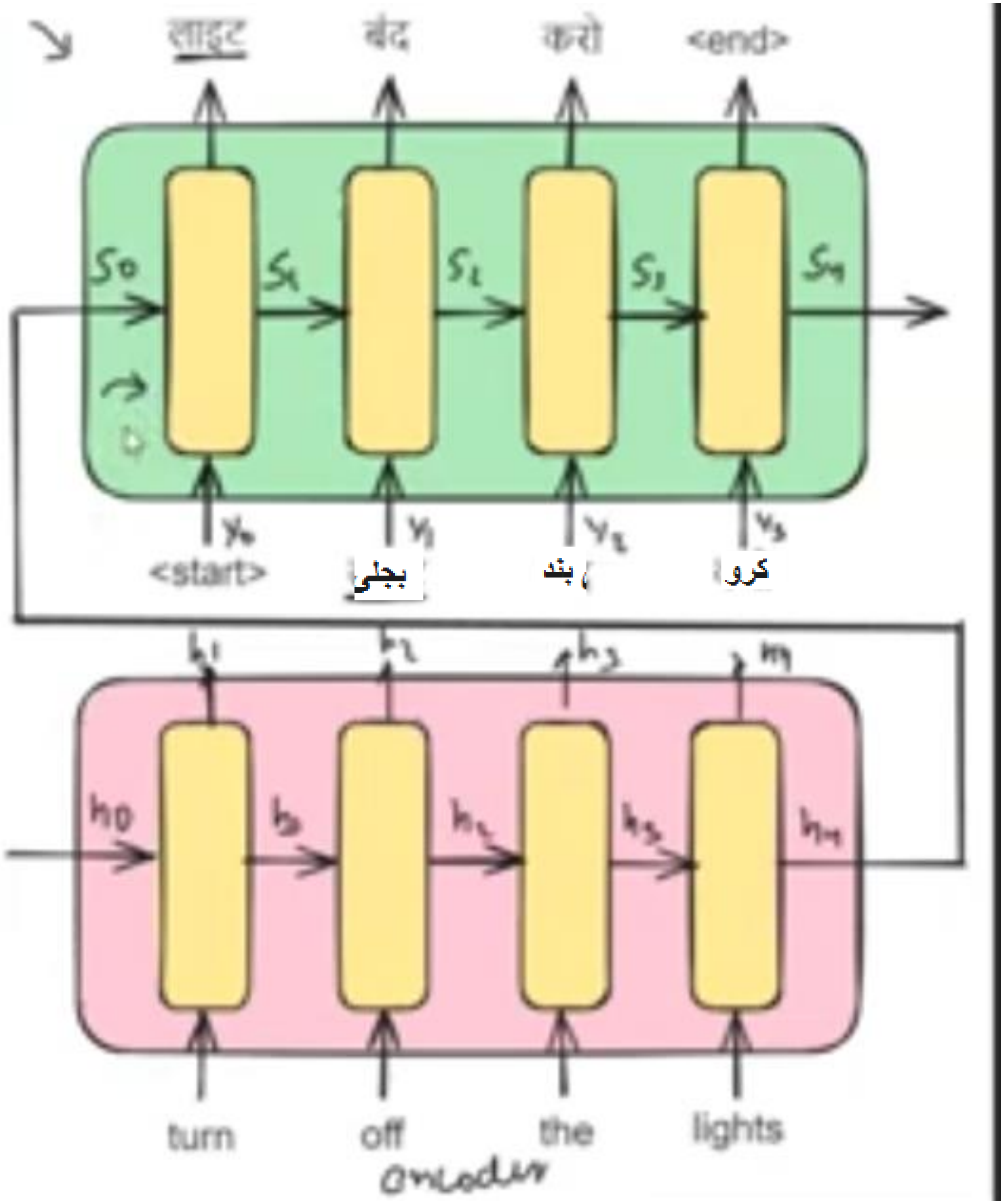
1st Problem:
Once upon a time in a small Pakistani village, a mischievous monkey stole a turban from a sleeping barber, wore it to a wedding, danced with the bewildered guests, accidentally got crowned the 'Banana King' by the local kids, and ended up leading a vibrant, impromptu parade of laughing villagers, cows, and street dogs, all while balancing a stack of mangoes on its head, creating a hilariously unforgettable spectacle and an amusing legend that the village still chuckles about every monsoon season.

Face Difficulty when the sentence is greater than 25 word.

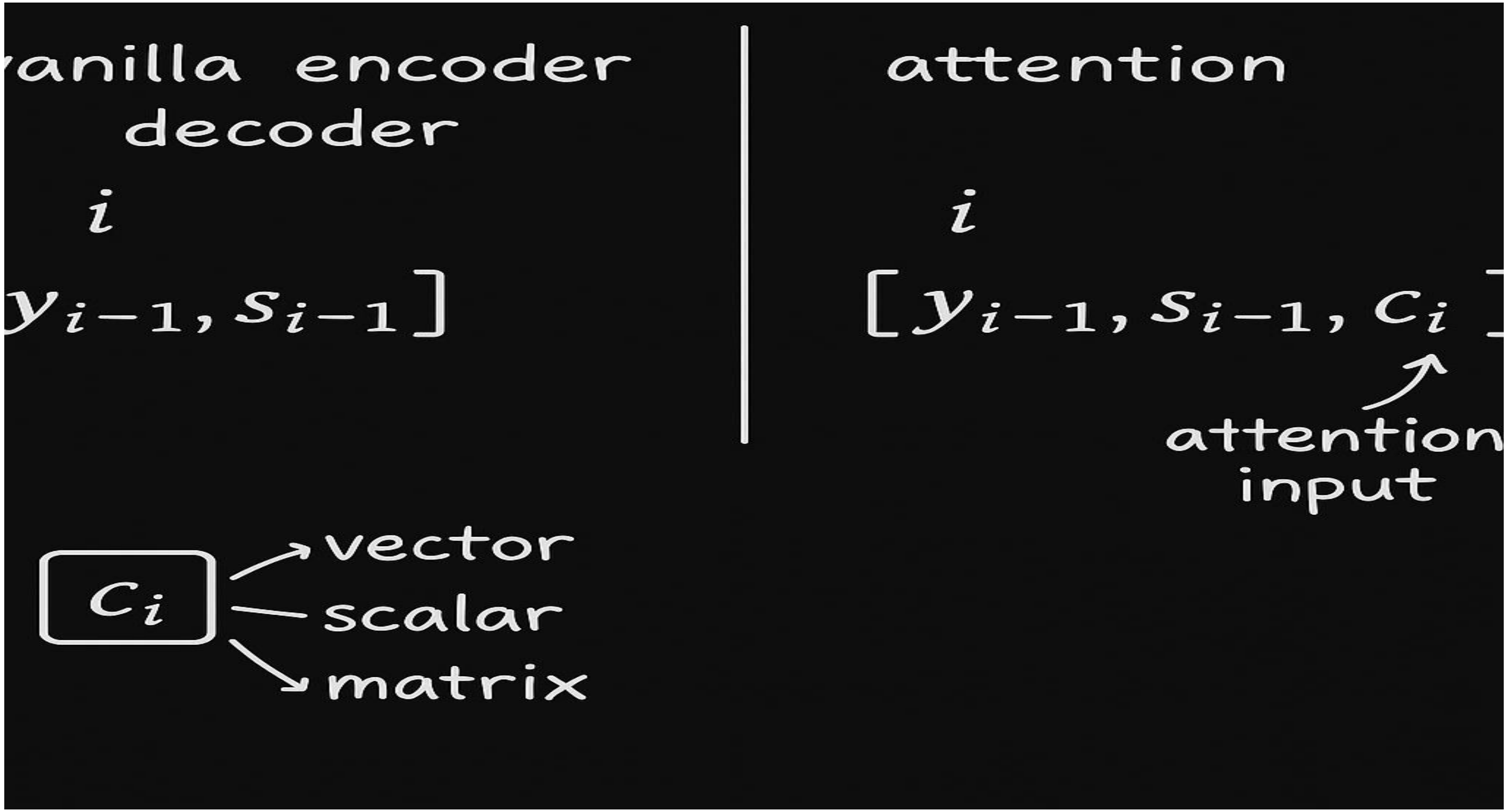
2nd problem:
Turn off the light.
Light:
Turn off:
Don't need the complete sentence, only a specific part of that.
But we give complete sentence to Decoder to translate→ Static Representation
It would be good, if we attention a specific part of the sentence at the time of translation.

- When we read a larger sentence. Our
- Eyes creates a attention span/region.
- And other things are blury at that moment.
- We need to introduce this thing in our architecture.
- When translating “light” which portion of the sentence need pass.
- Which time step is important to translate the light.
- This mechanism is called Attention Mechanisms

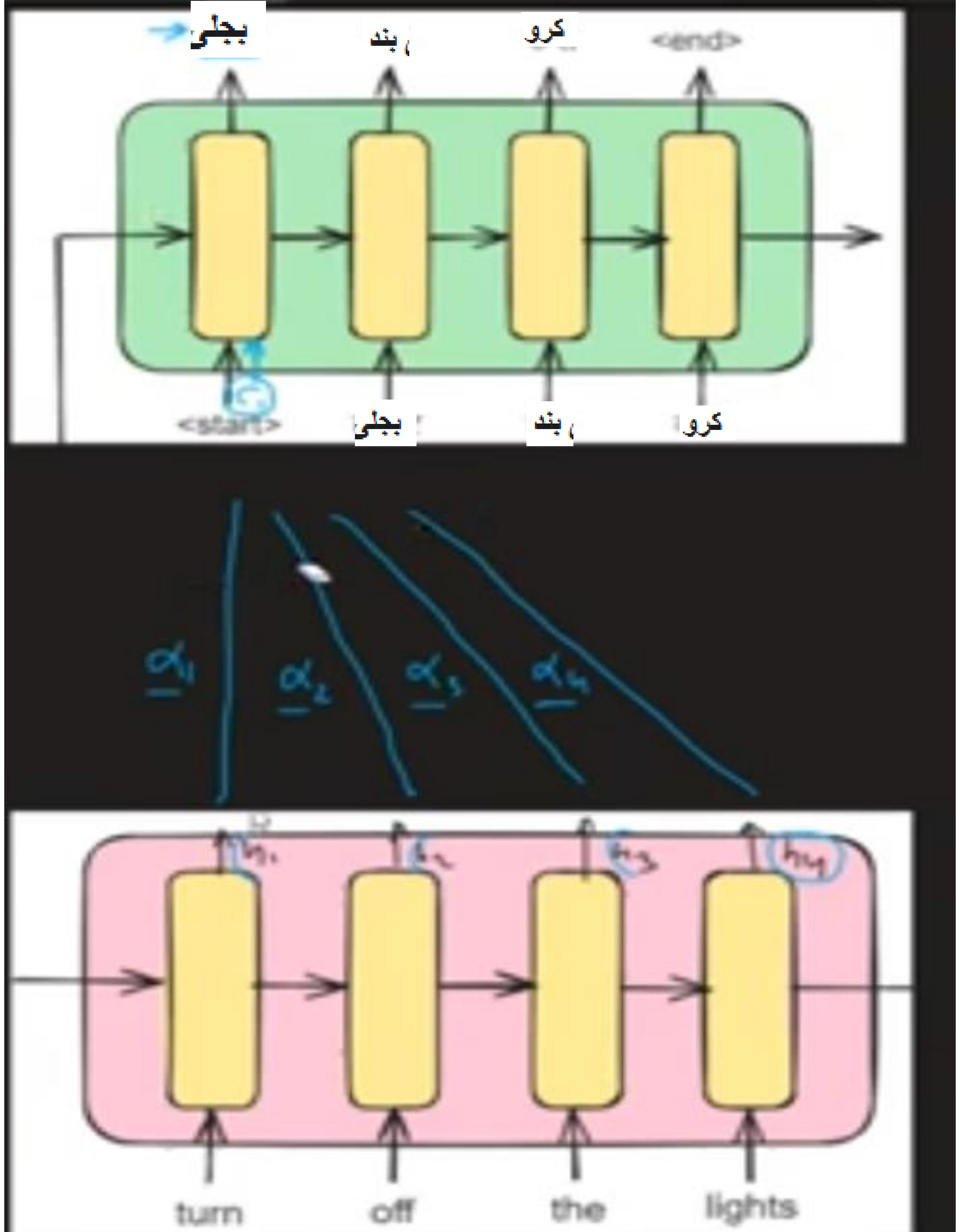
- At time step i=2, we provided
- Y1, s1
- In Attention mechanism you provide
- More piece of information (h1,h2,h3, or h4)
- We say it ci→ attention input



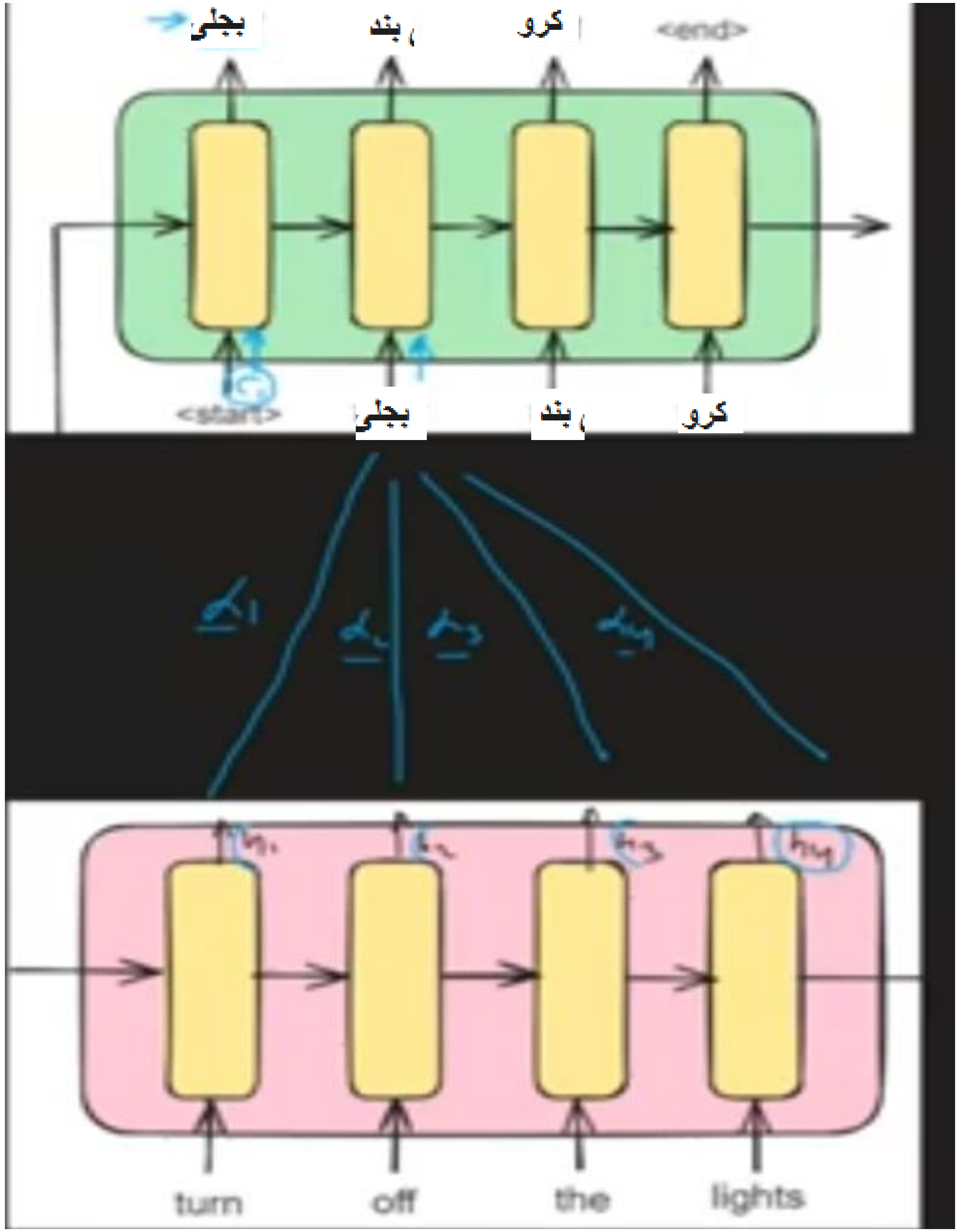
Vanilaa encoder vs. Attention bases



- What is dimension of ci,
- Ci is vector that is may be h1, h2 or combination of h1 and h2
- For both weighted sum Is used.



$$C_1=\alpha_1h_1+\alpha_2h_2+\alpha_3h_3+\alpha_4h_4$$



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

At time step =2

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

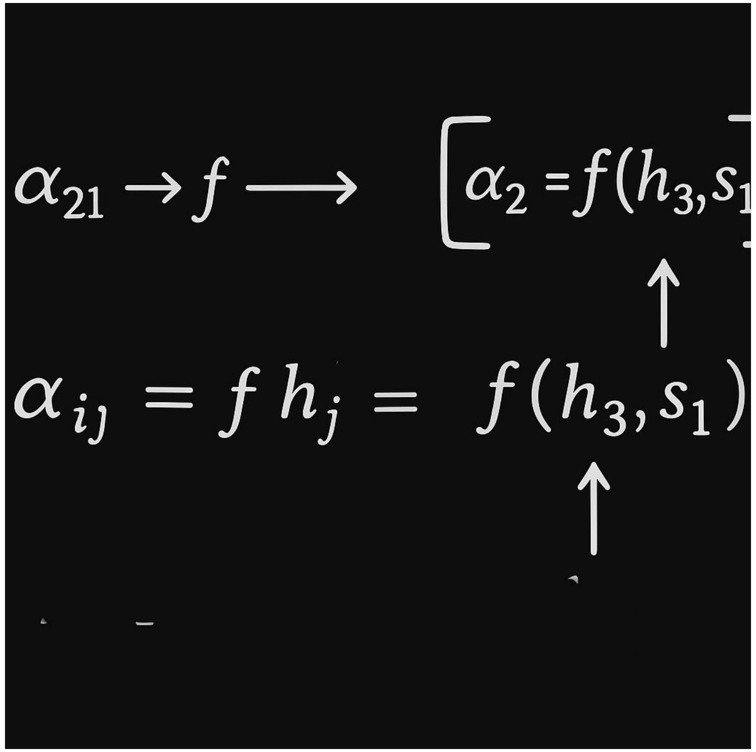
$$C_i = \sum_i \alpha_{ij} h_j$$

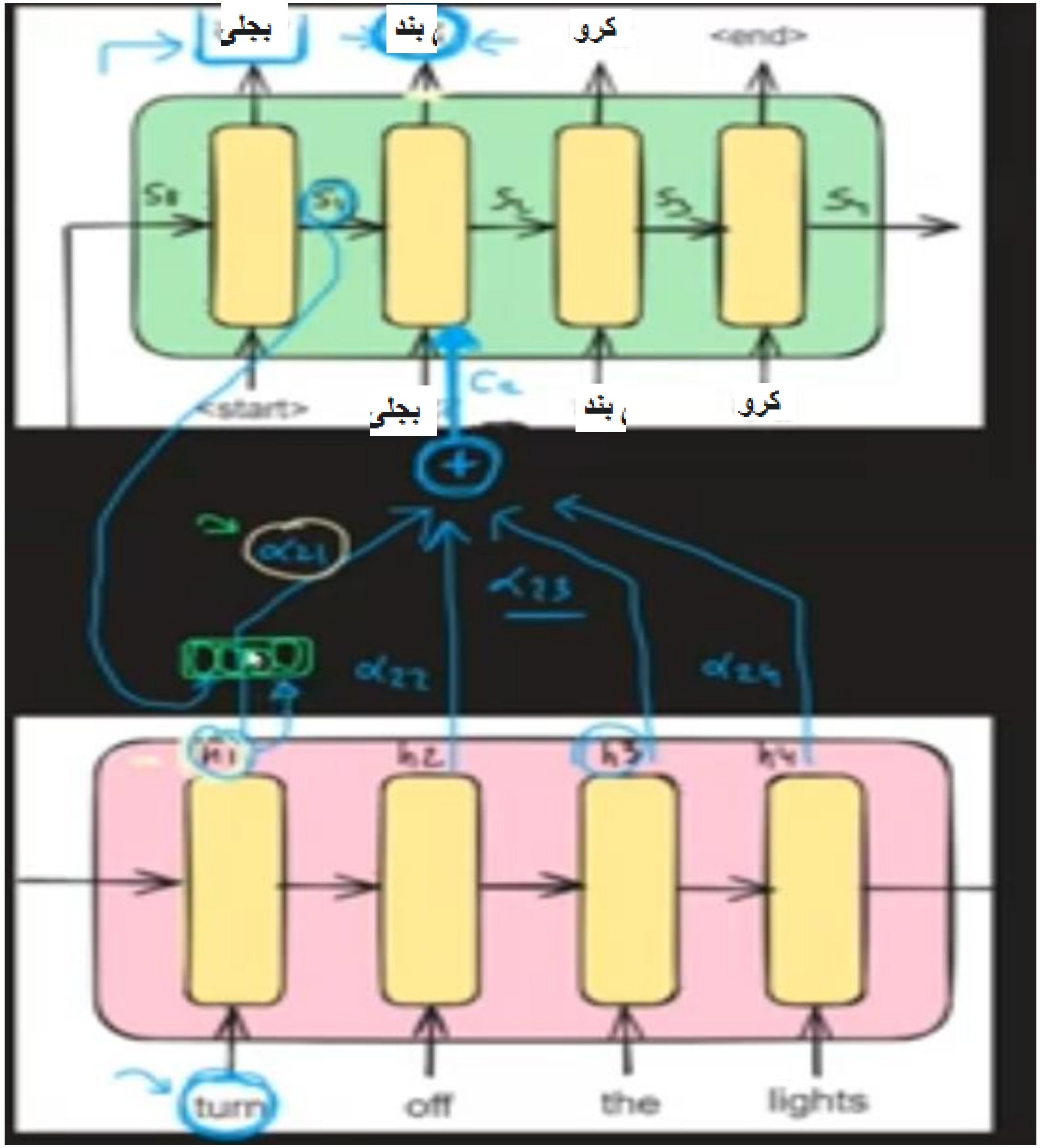
i*j=16

- How the alpha will create
- calculate alpha21
  - Called Alignment Score
  - Similarity Score.

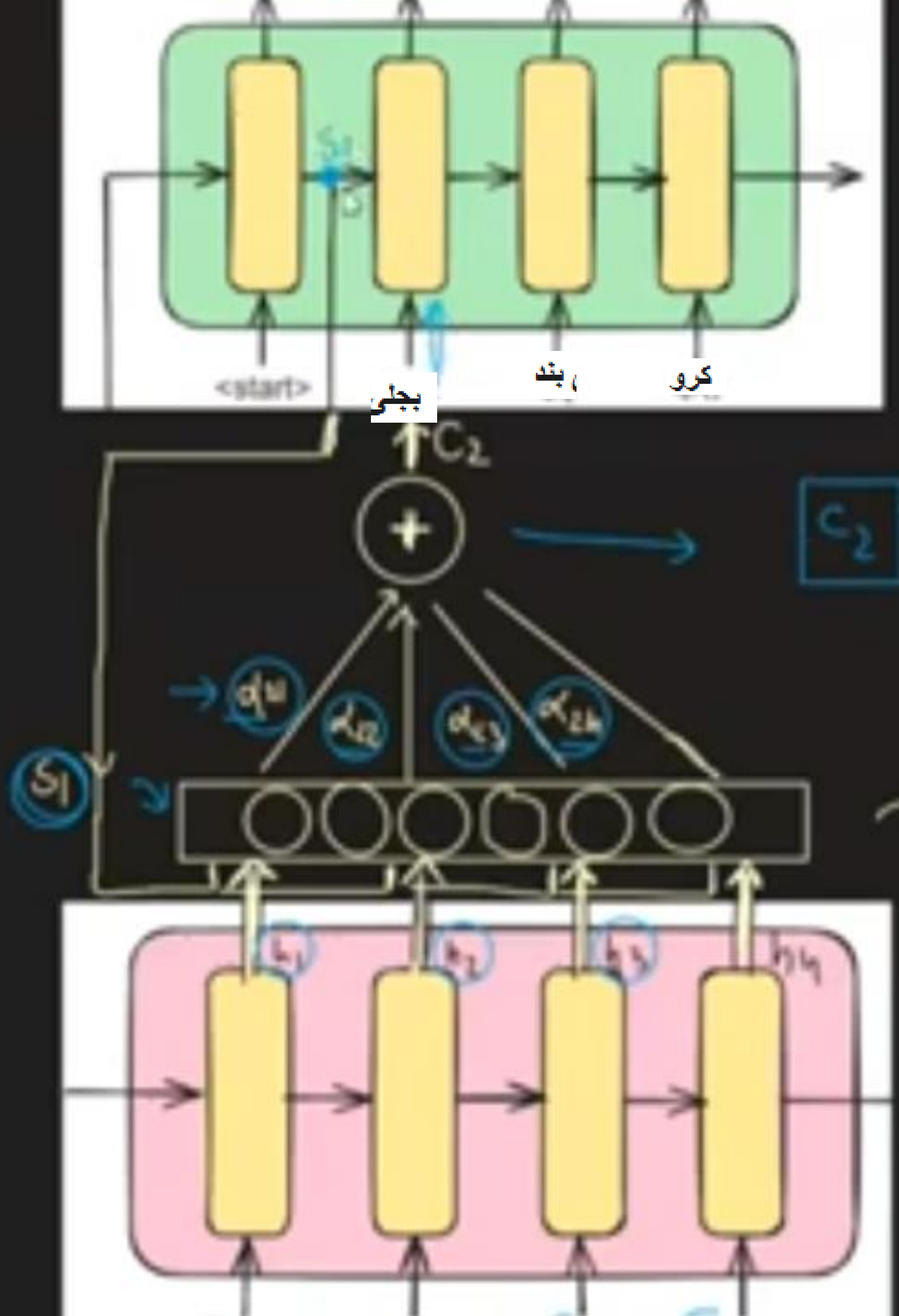
- At timestep, i=2, the output is printed , what is the role of encoder time step j=1
- i.e. “Turn” or h1
- Now alpha21 depends upon on which qunatities
- H1,s1

- ANN Universal approximate function



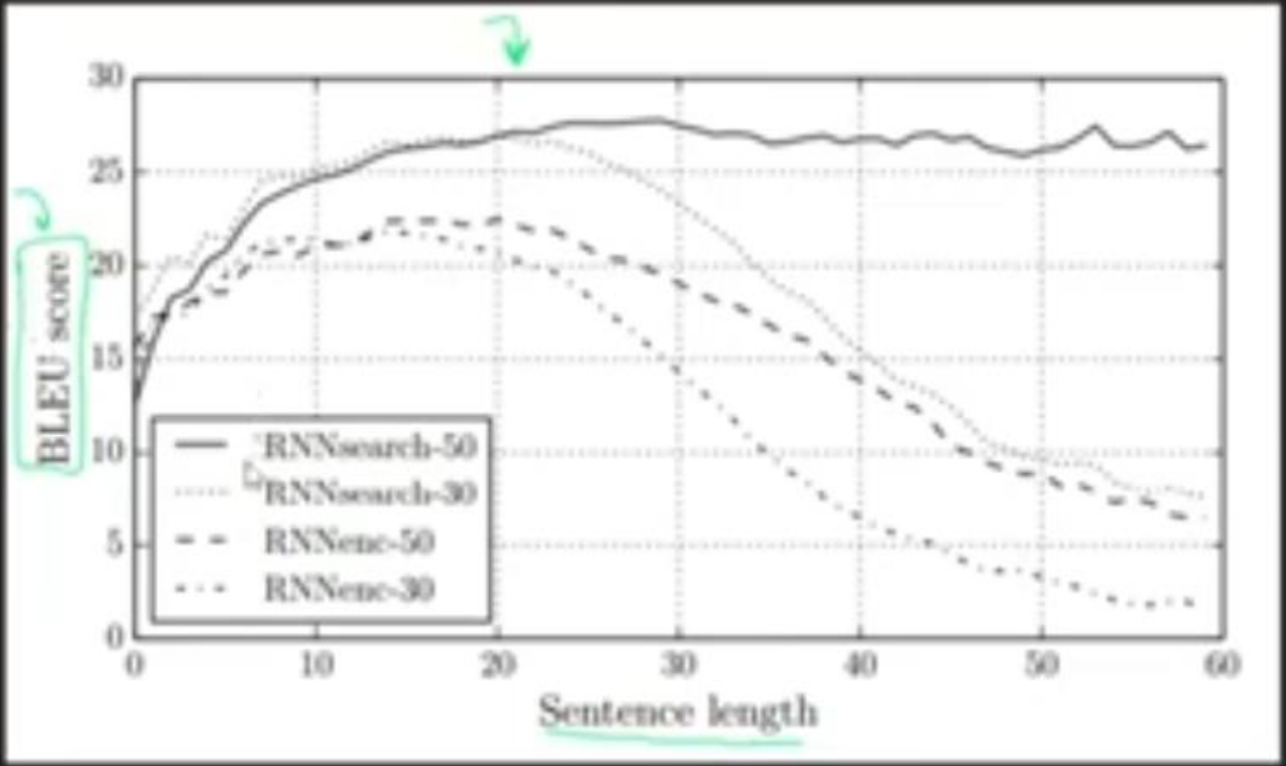


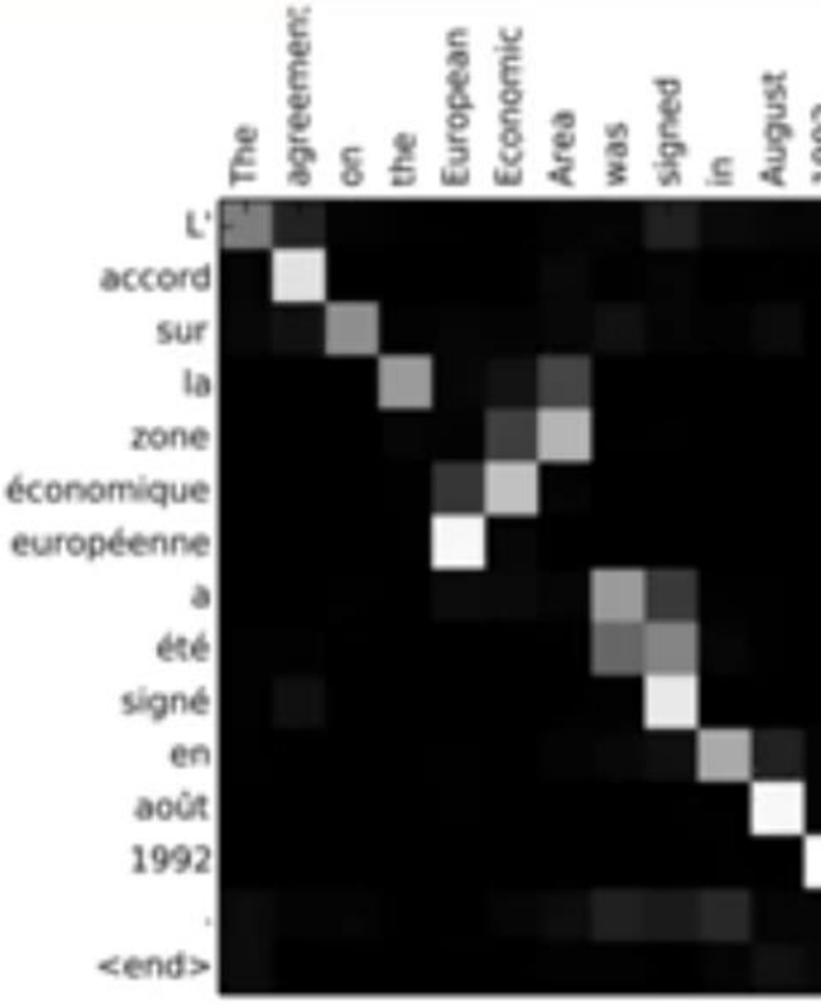
بجلی
بند
کرو
<end>
<start>
بجلی
بند
کرو
α21
α23
α22
α24
h1
h2
h3
h4
turn
off
the
lights



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

Time Distributed Dense layer


BLEU score
Sentence length
RNNsearch-50
RNNsearch-30
RNNenc-50
RNNenc-30


The
agreement
on
the
European
Economic
Area
was
signed
in
August
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>

(a)

turn off the 13

α_{11} [illegible] • α_{13} α_{14}

α_{12} [illegible] α_{23} α_{24}

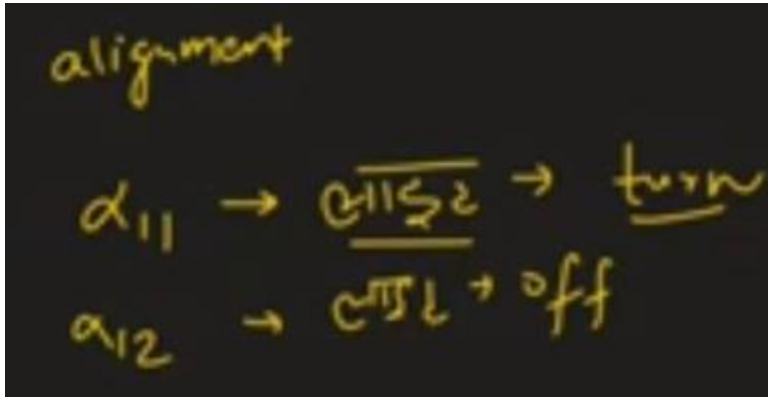
[illegible]

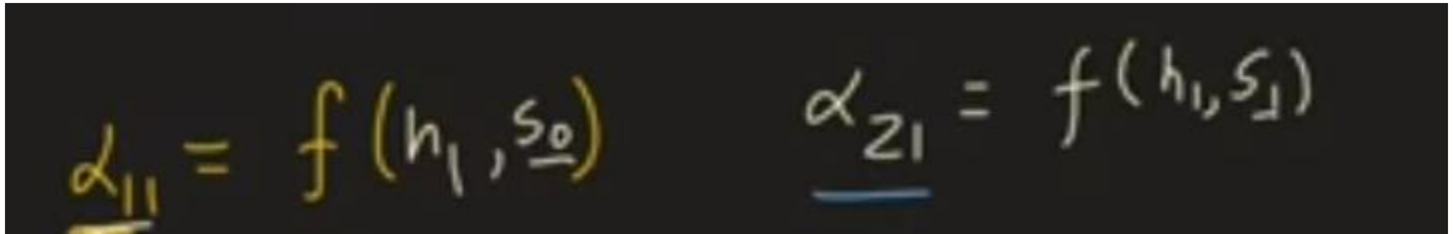
end

- Bahduana and loung Attention Model

- Two types to find the Alphas
  - Bahduana
  - Loung
- Alignment Score

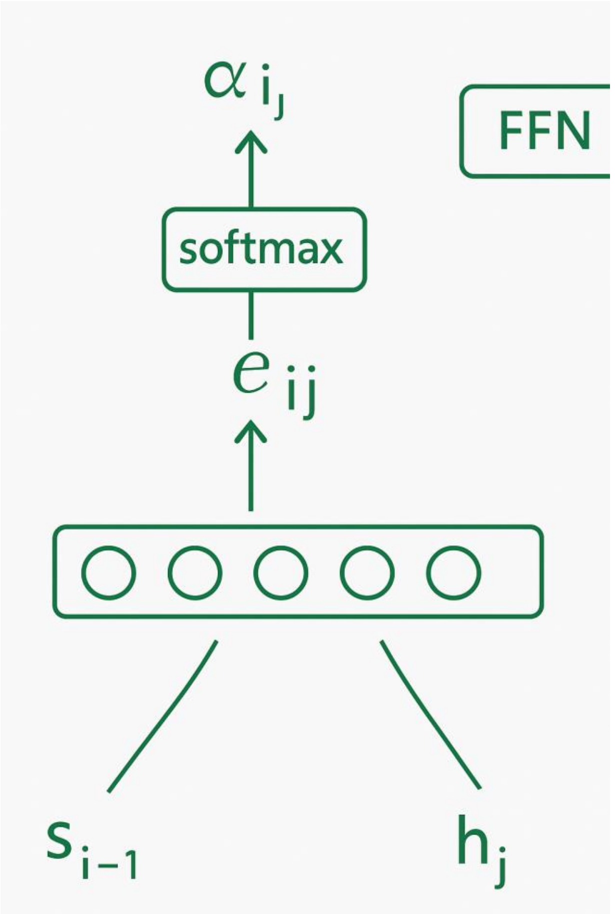
$$C_i = \sum_i \alpha_{ij} h_j$$

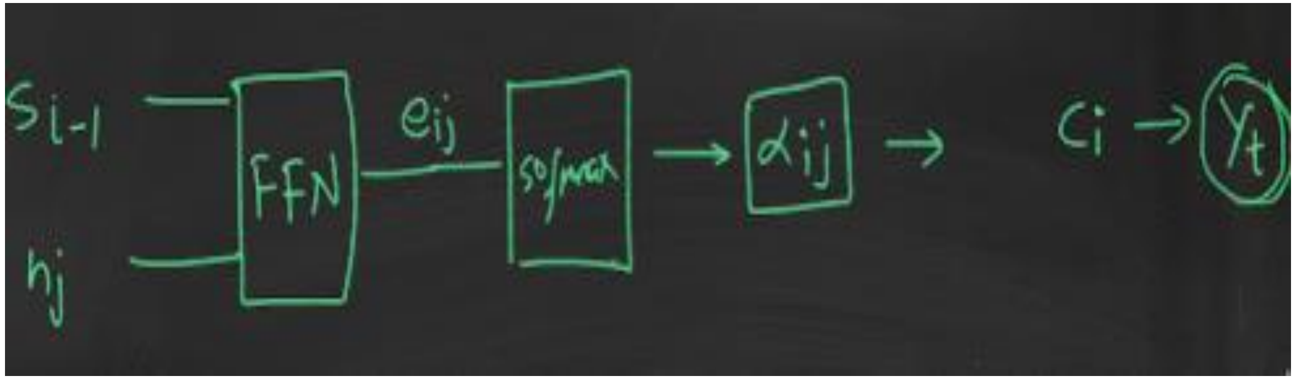




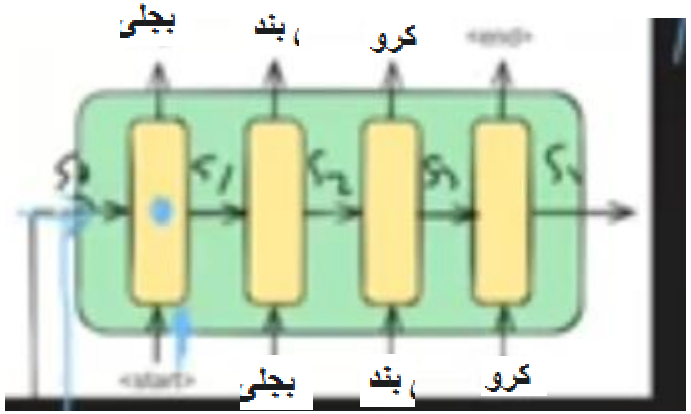
Bahduana

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

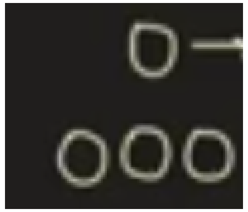




$s_0 = [\, e \;\; f \;\; g \;\; h \,]$



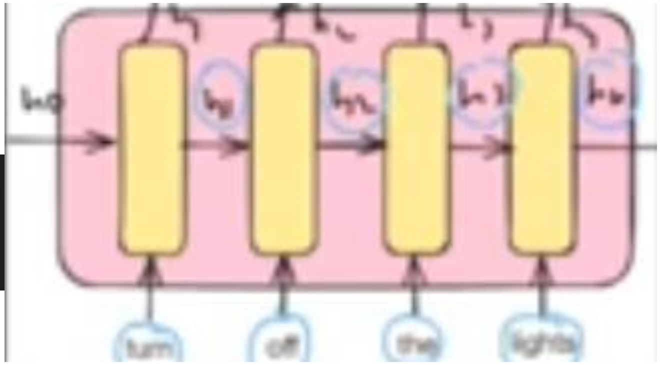
- Enter the encoder sentences.
  - Turn off the light
  - You will be get hidden stats(h1,h2,h3, h4)
  - Assume
  - At i=1 , decoder will calculate

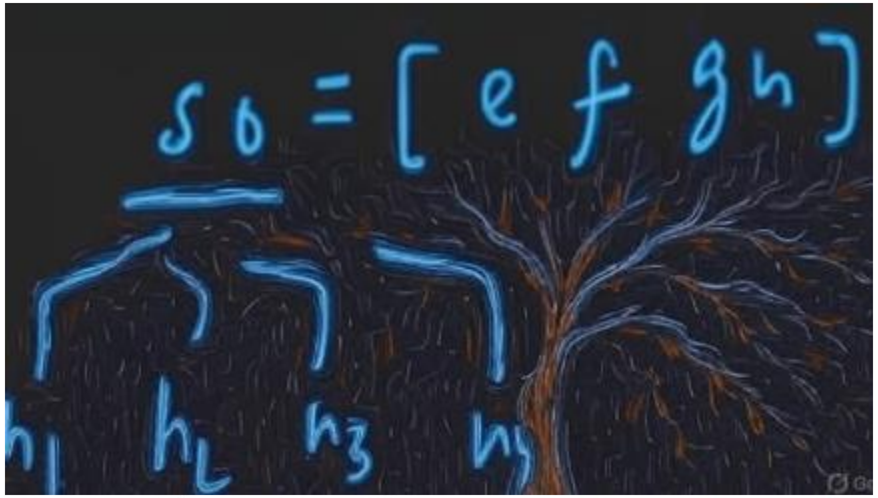


$c_i = \sum \alpha_{ij} h_j$

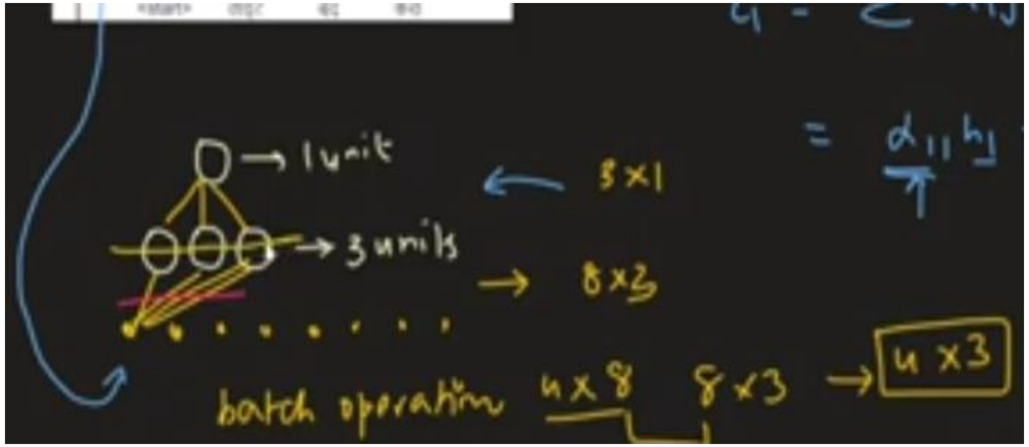
$\alpha_{11} h_1 + \alpha_{12} h_2 + \alpha_{13} h_3 + \alpha_{14} h_4$

$h_0 = [\, a \;\; b \;\; c \;\; d \,]$





$$\begin{bmatrix} s_{01} & s_{02} & s_{03} & s_{04} & h_{11} & h_{12} & h_{13} & h_{14} \\ s_{01} & s_{02} & s_{03} & s_{04} & h_{21} & h_{22} & h_{23} & h_{24} \\ s_{01} & s_{12} & s_{03} & s_{04} & h_{31} & h_{32} & h_{33} & h_{34} \\ s_{01} & s_{02} & s_{03} & s_{04} & h_{41} & h_{42} & h_{43} & h_{44} \end{bmatrix}$$

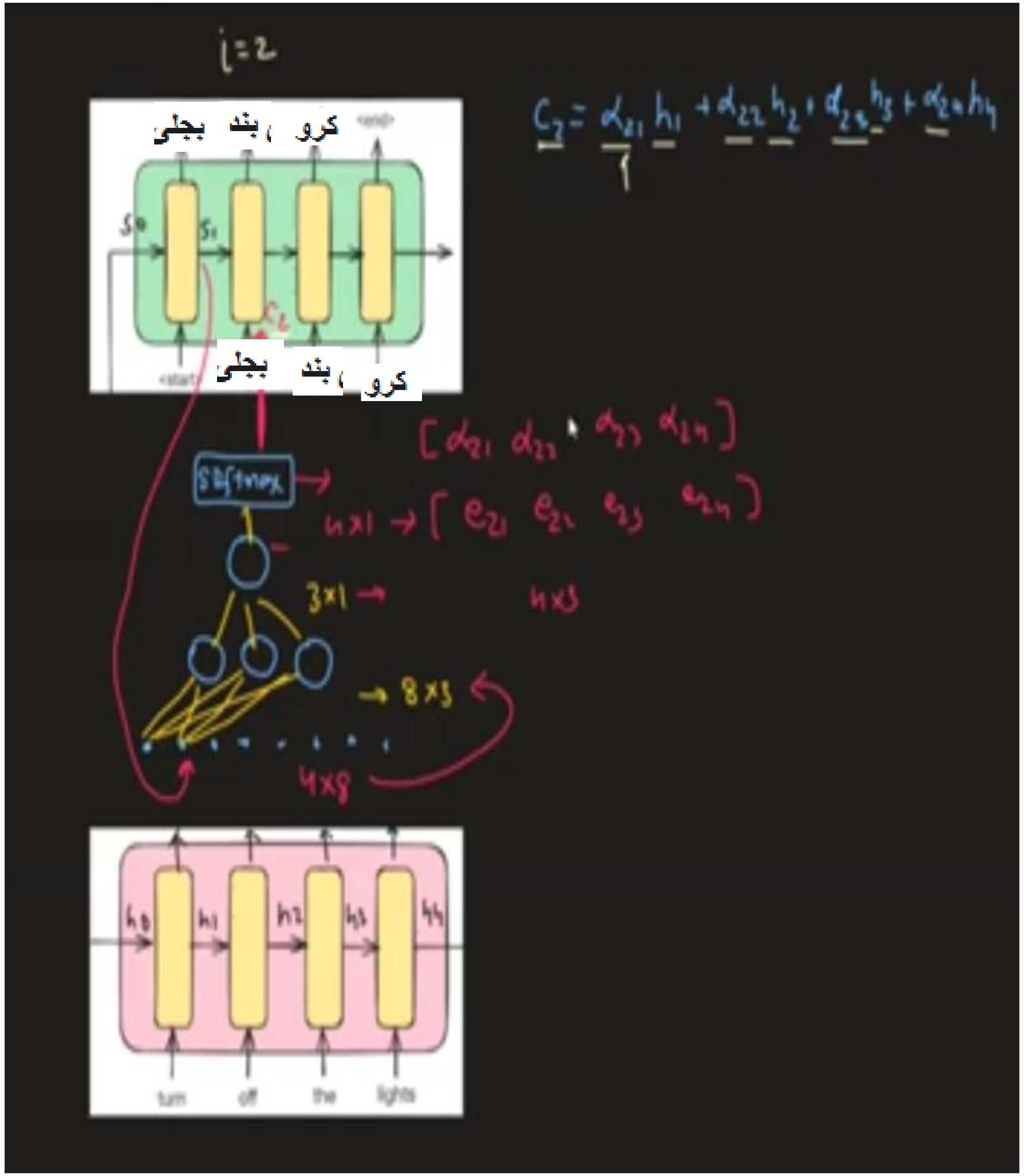


$(4\times1) \rightarrow$ 4 numers

$$\frac{e^{e_{11}}}{e^{e_{11}}+e^{e_{12}}+e^{e_{13}}+e^{e_{14}}} \quad (4\times1) \rightarrow$$

So Y_{t-1} C_1 $\longrightarrow$ lstm $\rightarrow$ Y_t ([illegible])

At time=2



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

$$\begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} & h_{11} & h_{12} & h_{13} & h_{14} \\ S_{21} & S_{12} & S_{13} & S_{14} & h_{21} & h_{22} & h_{23} & h_{24} \\ S_{11} & S_{12} & S_{13} & S_{14} & h_{31} & h_{32} & h_{33} & h_{34} \\ S_{11} & S_{12} & S_{13} & S_{14} & h_{41} & h_{42} & h_{43} & h_{44} \end{bmatrix}$$

$C_2, S_1, Y_1 \rightarrow \text{lstm} \rightarrow Y_2$

$\downarrow$

बंद / S_2

Mathematical Form

$$c_{ij} = \sum \underline{\alpha_{ij}}\, \underline{h_j}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}$$

$$e_{ij} = \tanh(W[s_{i-1} : h_j] + b)$$

$$e_{ij} = V\left[\underline{\tanh}(\underline{W}[\underbrace{s_{i-1} : h_j}] + \underline{b})\right]$$

- Aslo called additive algorithm
- Model is call alignment model

Loung Attention

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

$$\alpha_{ij} = f(s_{i-1}, h_j) \times$$

$$\alpha_{ij} = f(\underset{\substack{\uparrow \\ \text{current}}}{s_i}, h_j)$$

① diff

$s_i^T \cdot \underline{h_j} \rightarrow$ dot product

nt ① diff

$s_i = [a\ b\ c\ d]$

$h_j = [e\ f\ g\ h]$

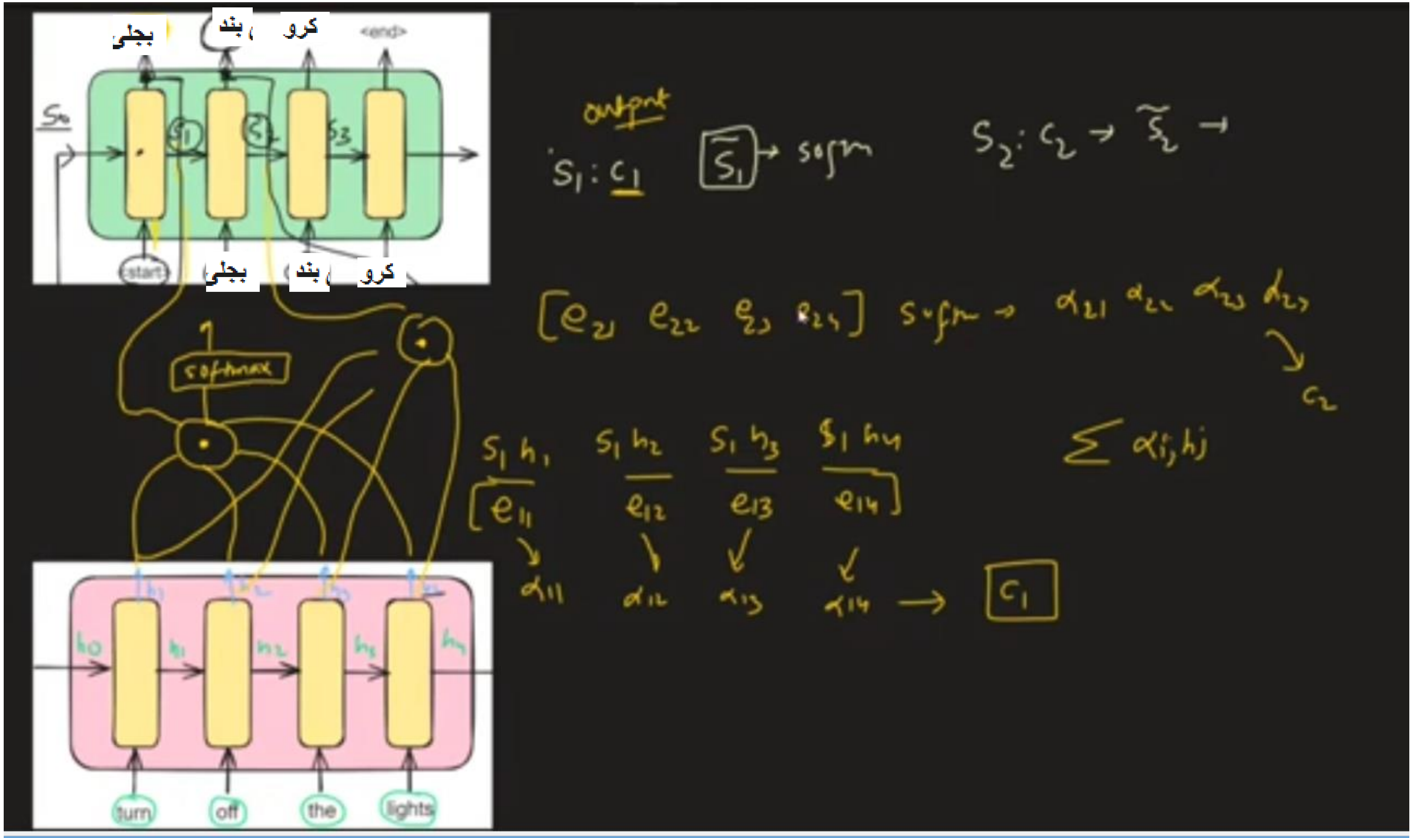
$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} [e\ f\ g\ h]$$

$$[ae + bf + cg + dh]$$

scalar → attention

r $\leftarrow \boxed{e_{ij}} \leftarrow$ scalar

- Current state has fore info then previous
- Experimentally proved


بجلی
بند
کرو
<end>
S0
S1
S2
S3
<start>
بجلی
بند
کرو
softmax
h0
h1
h2
h3
h4
turn
off
the
lights
output
S1 : C1
S1 → softm
S2 : C2 → S2 →
[e21 e22 e23 e24] softm → α21 α22 α23 α24
C2
S1 h1
S1 h2
S1 h3
S1 h4
[e11 e12 e13 e14]
α11 α12 α13 α14 → C1
Σ αij hj

- Called multiplicative
- Necessary for understanding of transformer→ self attention.