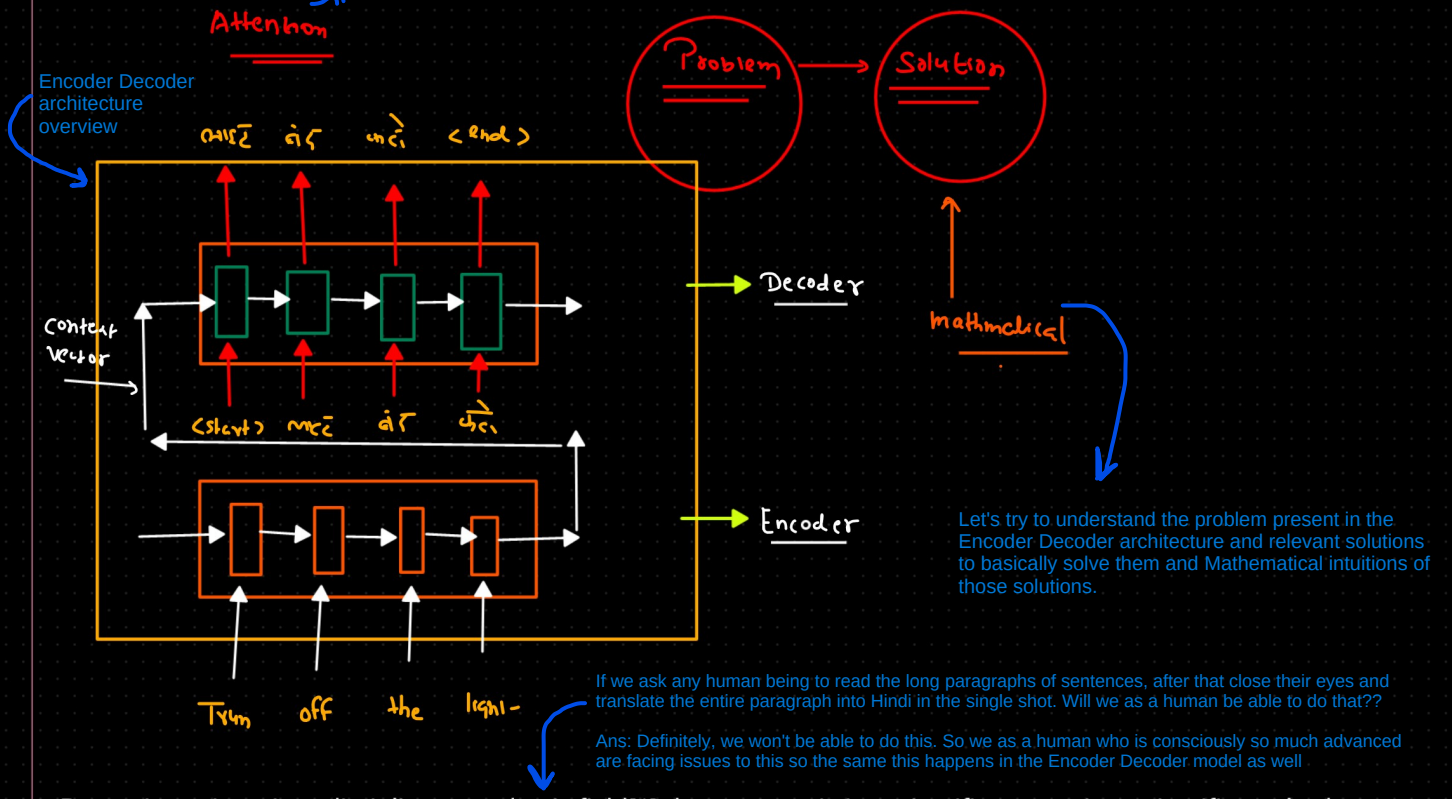
Attention Mechanism

Encoder- decoder

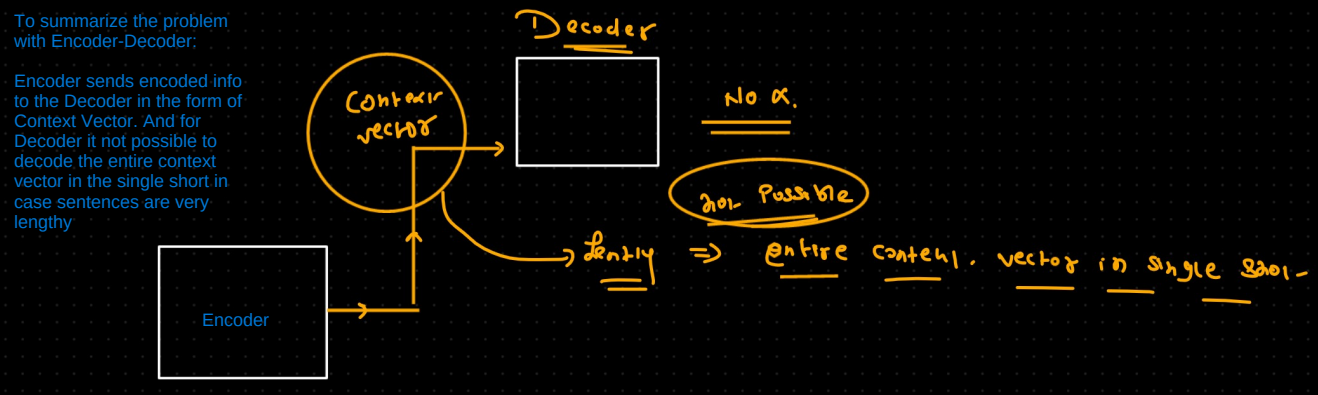
Problem Statement: Why should be learn attention mechanism on top of Encoder-Decoder architecture?



Data science is an interdisciplinary academic field[1] that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data.[2]

Data science also integrates domain knowledge from the underlying application domain (e.g., natural sciences, information technology, and medicine).[3] Data science is multifaceted and can be described as a science, a research paradigm, a research method, a discipline, a workflow, and a profession.[4]

Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" to "understand and analyze actual phenomena" with data.[5] It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge.[6] However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[7][8]



Since in Encoder Decoder we were not able to translate/decode the longer sentences in the single short hence, Attention mechanism was introduced.

= Attention => Real time human understanding

Focus | blur

Focus

Focus

Data mining is a particular data analysis technique that focuses on statistical modeling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information.[4] In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA).[5] EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing hypotheses.[6][7] Predictive analytics focuses on the application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All of the above are varieties of data analysis.[8]

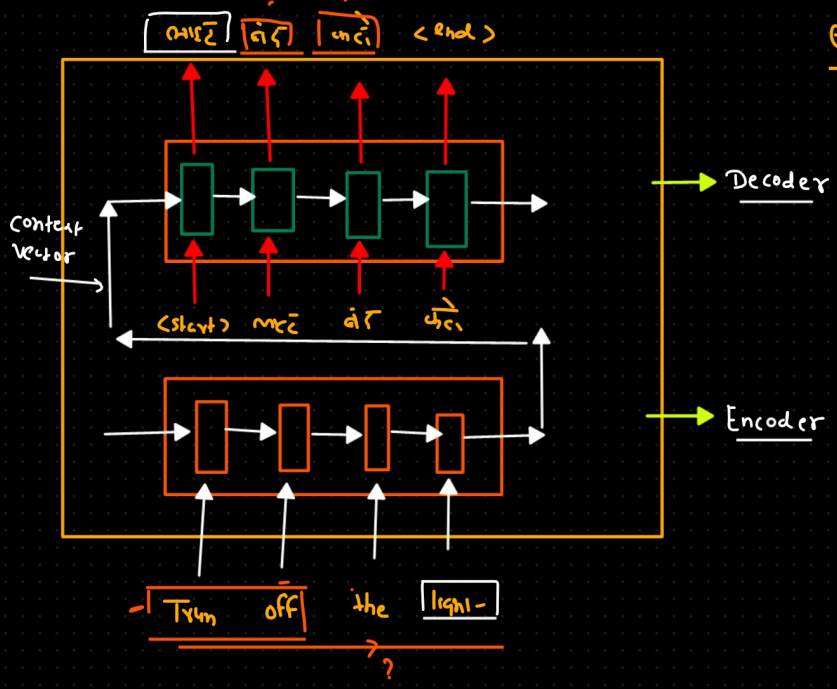
FOCUS and BLUR are 2 main things which we embed in the architecture due to which real time human understanding in achieved. This means when we are reading a part of huge paragraph then that part will be in FOCUS and rest entire sentence will be BLUR

Focus => Whatever part we are reading we are focusing on that.

Blur => Rest part will be blur

Mathematically

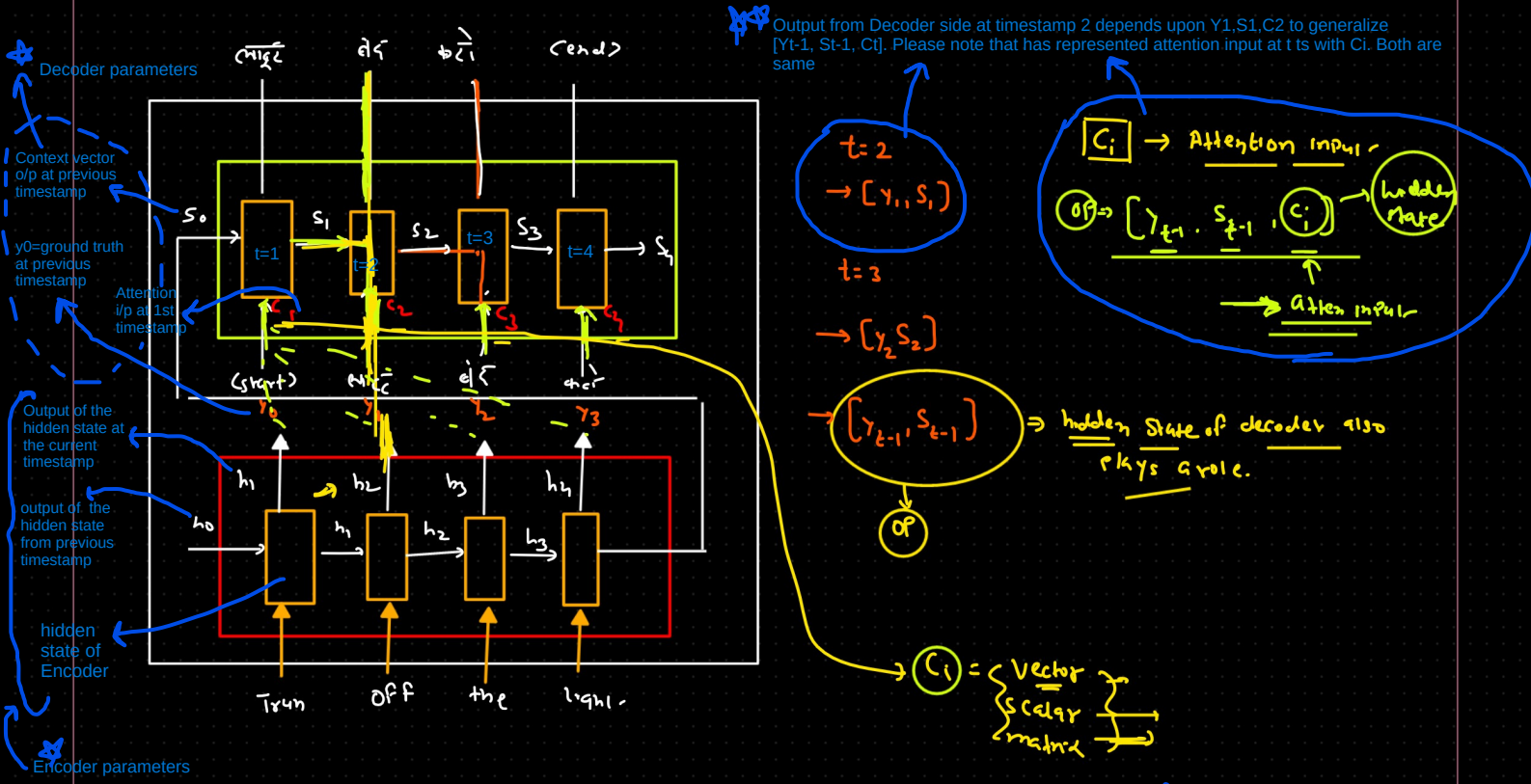
Below let's try to understand how Focus and Blur can be embedded mathematically.

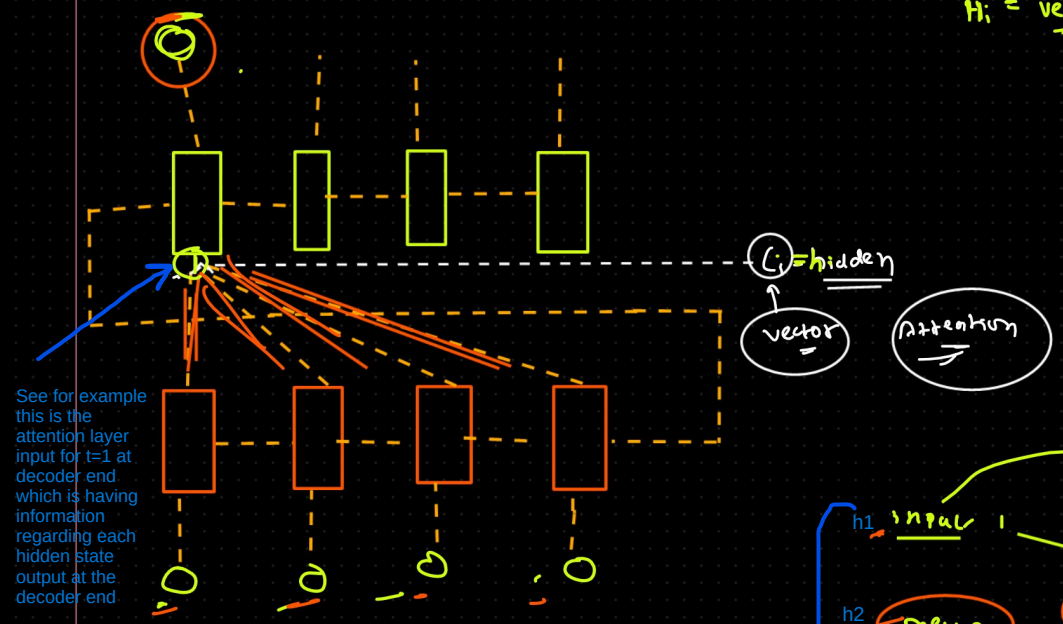


Example => {Do not Required entire sent -}

If we want the Hindi translation of the first two English inputs (ie; Turn off) in the single short then we will FOCUS just on first 2 input at encoder side and 2nd and 3rd output at the decoder side and will BLUR the remaining things

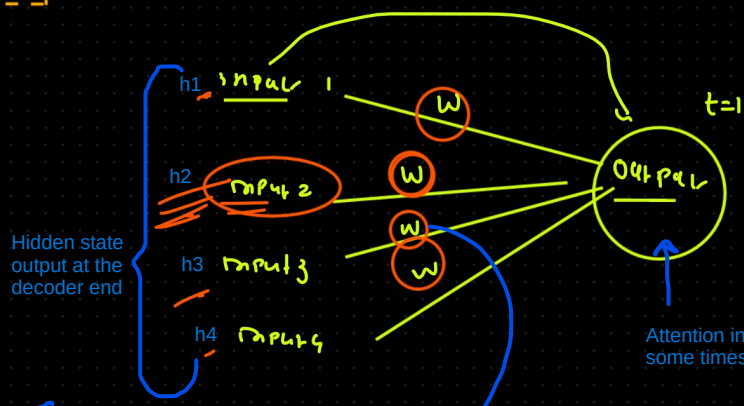
Previous hidden state input and the current hidden state output





Ci or Ct (attention input at t timestamp at decoder end) is Vector/Saler/matrix ?

Ans: Each attention input at decoder end is having information regarding the each hidden state output at the encoder end. This will be clear in the below picture. Hence we can say that if considering C1, will depend upon h1,h2,h3,h4. Since The hidden state output at the encoder end are vector representation, therefore attention input at decoder end can also be considered as the vector in nature.

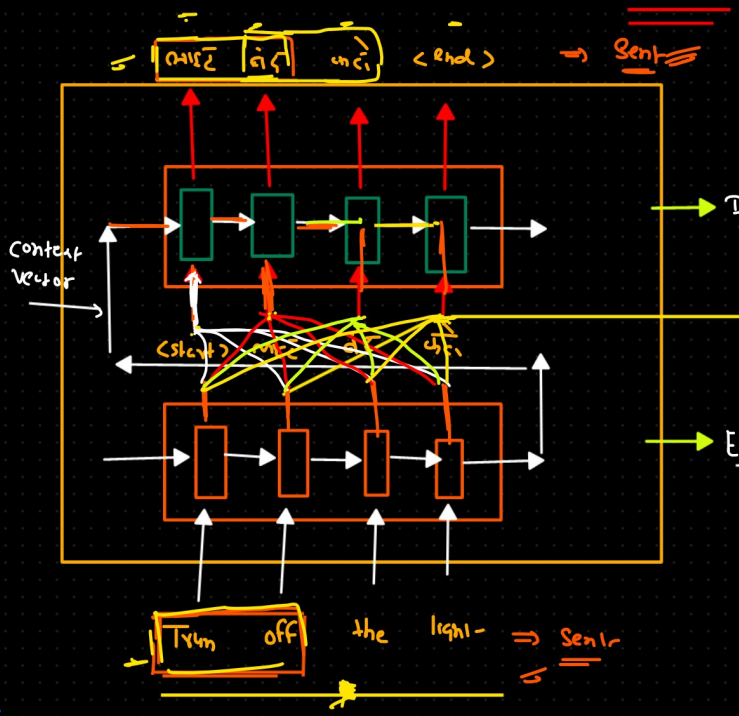


At decoder at some time stamp end we need to get the output based on real time human understanding. Since Attention is having info regarding each hidden state o/p at decoder end so now the one that is having basically more weightage will be in focus and other will be blur for basically decoding the sentence say at a particular time stamp. This is how real time human understanding is getting embedded

Understanding the attention mechanism using neural network:

To understand this first read the below page first, especially the Tea making example-->Good output (Tea) and Bad output, If tea is bad then back propagate and adjust the weights same here also we are feeding the context vector from previous ts followed by ground truth from previous ts followed by attention input after which respective output from decoder will be checked for Good or bad output and accordingly we will be adjusting the weights as discussed above

To summarize attention at a particular timestamp has output from each hidden state but whom it is going to give more weightage to get the respective decoded output at a particular timestamp is decided based on the above explanation

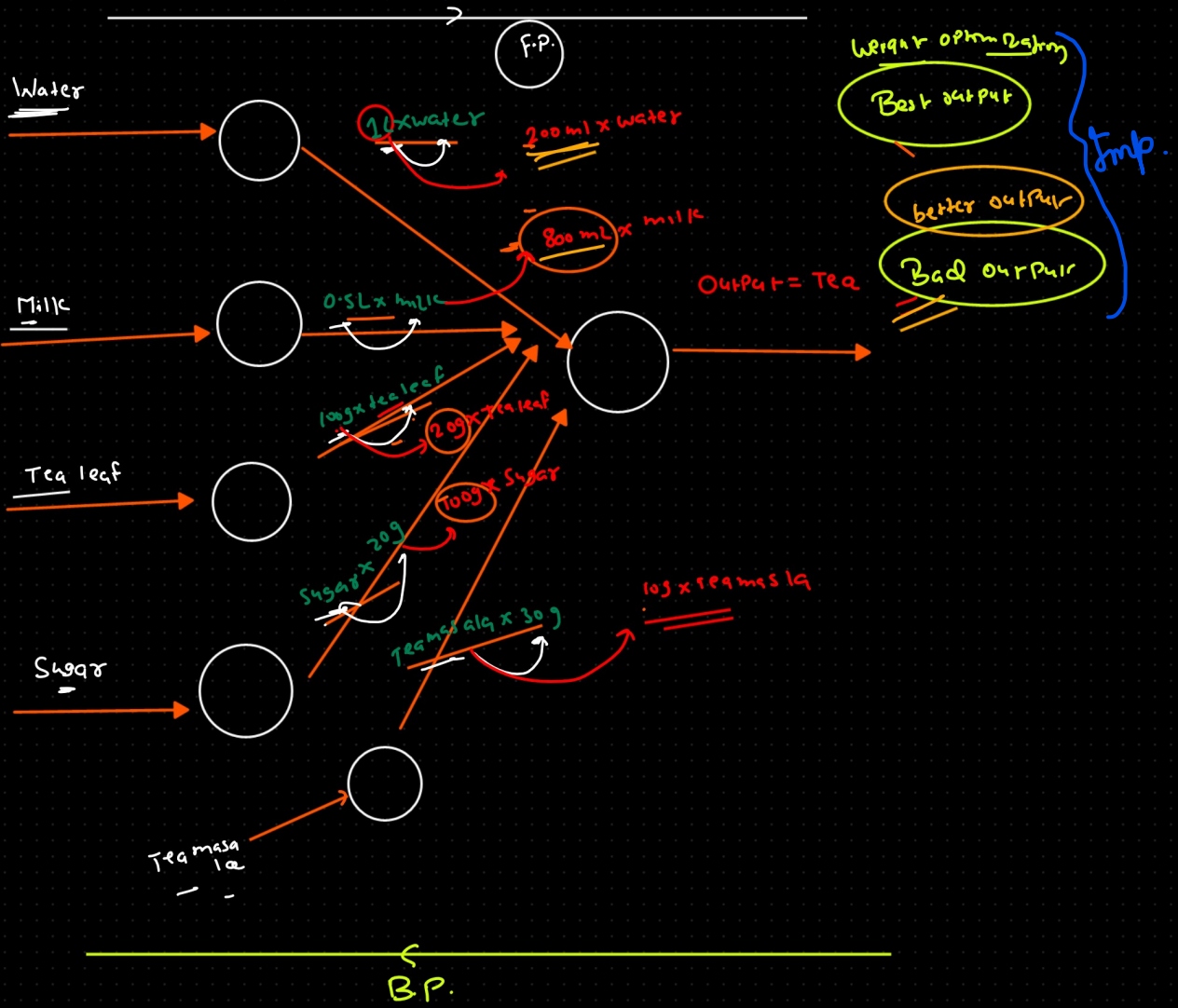


Real time human understanding:
Eek eek karke words padha(FOCUS) and at the same time unka hindi translation kiya. Jab kuch words par FOCUS kar rahe then at that time remaining ko BLUR kar diya. This real time human understanding is something which we are trying to embed using attentions

If we want to translate the first 2 words in our sentence ie; "Turn off" then we have to focus of 2nd and 3rd words generated from he decoder side. How we are going to map this relation. that we will be able to do using the connected graph depicted here. But how will this be presented mathematically

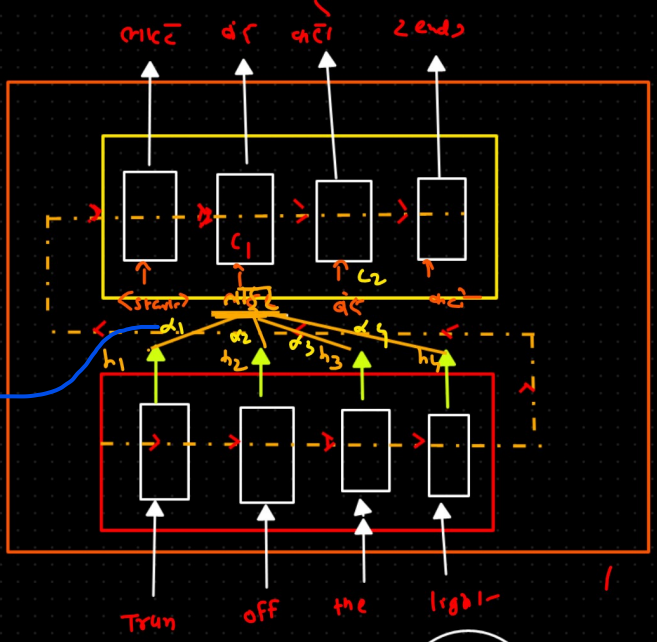
Wow!

Understanding the Neural Network using Tea making example:



Mathematical Intuition of the Attention:

So, above it is clear that Focus and Blur or real time human understanding is embedded using weights b/w hidden state output from encoder and attention layer input from decoder. So below we will see how to determine this weight



Weighted summation

$c_1 = \alpha_{11} \times h_1 + \alpha_{12} \times h_2 + \alpha_{13} h_3 + \alpha_{14} h_4$

α = Weight

$c_2 = \alpha_1 h_1 + \alpha_{22} h_2 + \alpha_{23} h_3 + \alpha_{24} h_4$

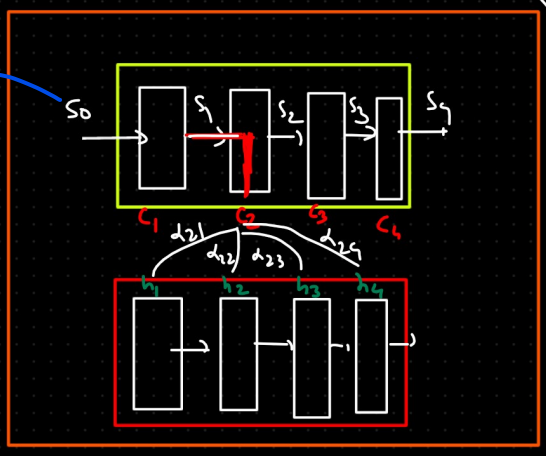
$c_i = \sum \alpha_{ij} h_i$

Generalized equation

Findout

Here my end goal to find out a weight

$c_i = \alpha_{ij} * h_i$



Please note that S is representing the hidden state input or output at decoder end

Whereas, h represents the hidden state input or output at the encoder end

α_{21}

The intended weights will be depending upon 2 things. Consider (alpha)21 will be depending upon h1 and s1

Previous hidden state of Decoder

h_1

s_1

Function

optimized

$\alpha_{21} \rightarrow f(h_1, s_1)$

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

ANN

Weight is the function of h and s. Therefore we can say we need to determine appropriate function that establishes relation b/w h and s.
So ANN can be used as the function to basically to establishing the relation b/w h and s

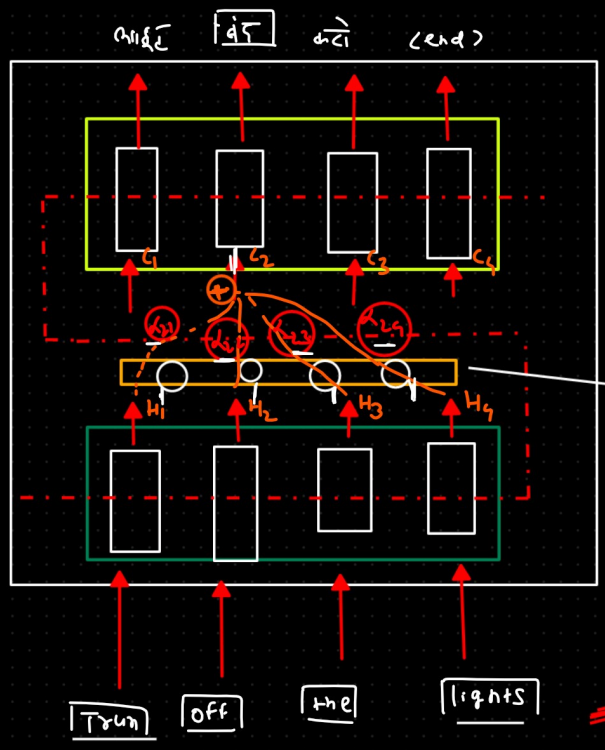
Attention layer → ANN → to find out a weight between input and output

Input from encoder side

H → (S, H) → C → O → Output from decoder side

weight → ANN

In short to conclude whole discussion until now we can say that Attention is the ANN which is used to determine the weights b/w all inputs(hidden state output at Encoder side) and output(decoder end)

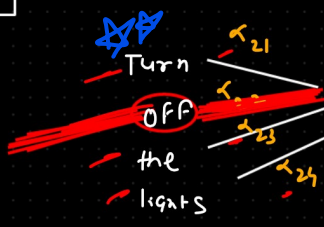


ANN → this weight is going to learn

$c_2 = \alpha_{21} h_1 + \alpha_{22} h_2 + \alpha_{23} h_3 + \alpha_{24} h_4$

Time distributed Dense layer → Attention Layer

Attention → Real time state → focus



Here (alpha)22 will be having more weightage than other weights. This is how real time human understanding is embedded where at some instant/timestamp we are able to FOCUS and decode a English word into Hindi by keeping others as BLUR

English → Hindi:

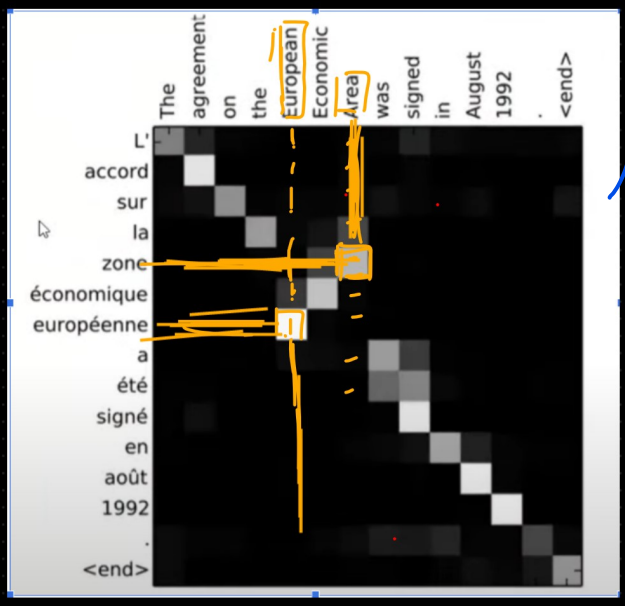
Attention layer / Attention mechanism from Research Paper

Real = english → french

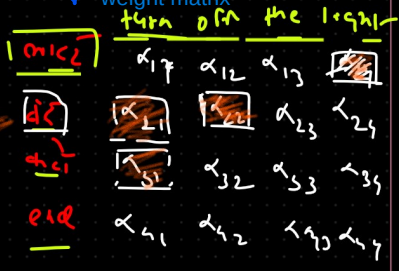
RNN

Bi-Directional RNN

In Attention Research paper we are taking the bidirectional RNN. In our previous explanation (English to Hindi) we took simple RNN



Relate this heatmap with the 2d weight matrix



weight

Heatmap

Plotting heatmap that will help to visualize the important weights for a given combination of translation

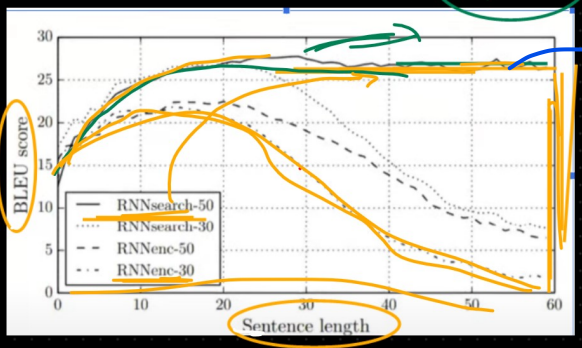
Focus

Blur

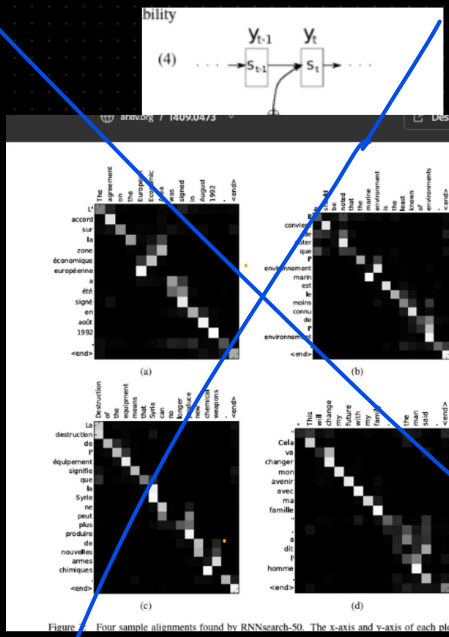
Real human learning

BLUE score comparison for different models wrt to sentence length as per Research Paper

higher



This curve represents Bidirectional RNN with Attention layer which is having the highest BLUE score



This was taken from the Research Paper

Transform, Bert, GAN

Projects