EXPERIMENT NO – 9

STUDY OF DIFFERENT TYPES OF NEURAL MACHINE TRANSLATION MODEL (NMT)

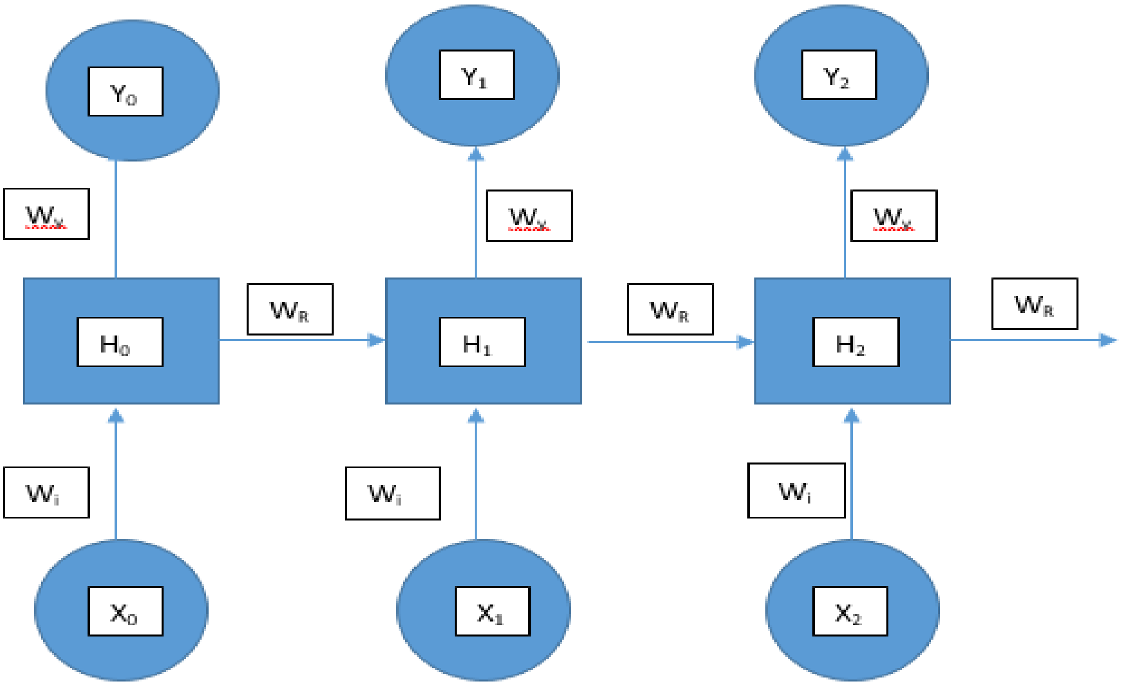
- ***Basic encoder decoder model using Recurrent Neural Network (RNN):***

About the model:

1. *The encoder-decoder model is a way of using recurrent neural networks for sequence-to-sequence prediction problems.*
2. *It was initially developed for machine translation problems, although it has proven successful at related sequence-to-sequence prediction problems such as text summarization and question answering.*
3. *The approach involves two recurrent neural networks, one to encode the input sequence, called the encoder, and a second to decode the encoded input sequence into the target sequence called the decoder.*

Need of the Recurrent Neural Network (RNN):

1. *The main problem with feed forward neural network is it only focuses on the current value/context. This value is not retained in the next time stamp.*
2. *Sometimes in the real world situation we have to retain the previous value and based on the previous and current input/context we have to decide the output.*
3. *For example, in case of sentences what will be the next word, that can be decided with the help of the context of the previous words.*
4. *Hence in language modelling we require long term temporal dependencies to decide the next word in the sentences.*
5. *RNN model helps in such type of situation. Long term temporal dependencies cannot be achieved with feed forward neural network.*
6. *A simple model of recurrent neural network uses tanh or sigmoid as an activation function.*
7. *When training a deep neural network with gradient based learning and backpropagation, we find the partial derivatives by traversing the network from the final layer to the initial layer.*
8. *Using the chain rule, layers that are deeper into the network go through continuous matrix multiplications in order to compute their derivatives.*
9. *Structure of a Recurrent Neural Network (RNN):*



Disadvantage of Recurrent Neural Network (RNN):

1. *In a network of n hidden layers, n derivatives will be multiplied together.*
2. *If the derivatives are large then the gradient will increase exponentially as we propagate down the model until they eventually explode, and this is what we call the problem of* ***exploding gradient****.*
3. *Alternatively, if the derivatives are small then the gradient will decrease exponentially as we propagate through the model until it eventually vanishes, and this is the* ***vanishing gradient*** *problem.*

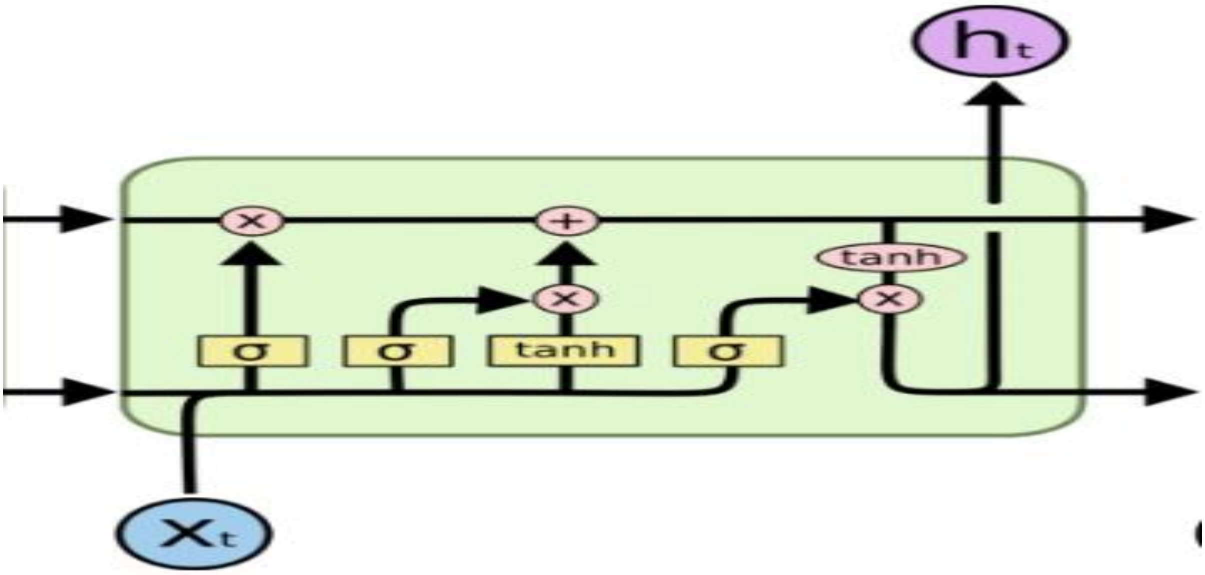
❖ ***Basic encoder decoder model using Long Short-term Memory (LSTM):***

About the model:

1. *Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies.*
2. *LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior*
3. *All recurrent neural networks have the form of a chain of repeating modules of neural network.*
4. *LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.*

Need of the Long Short-term Memory (LSTM):

1. *The long term dependency problem can be solved with LSTM models.*
2. *LSTM cells are operated with the help of real number parameter called gates.*
3. *The input gate parameter helps to decide how much new input is required to change the memory state.*
4. *The forget gate parameter helps to decide how much previous value to retain in-memory state.*
5. *And the output gate parameter controls how strongly the current memory state is to pass into the next layer.*
6. *Structure of a Recurrent Neural Network (RNN):*



Disadvantage of Long Short-term Memory (LSTM):

1. *LSTMs are prone to overfitting and it is difficult to apply the dropout algorithm to curb this issue.*
2. *They require a lot of resources and time to get trained and become ready for real-world applications.*
3. *In technical terms, they need high memory-bandwidth because of linear layers present in each cell which the system usually fails to provide for. Thus, hardware-wise, LSTMs become quite inefficient.*

❖ ***Basic encoder decoder model using Attention-based:***

About the model:

1. *This mechanism helps the neural network pay more attention to certain parts of inputs (instead of the entire input) while generating the output.*
2. *Let's take an example. Say we want to translate the Spanish sentence to English. As humans, we don't pay attention to every single word in the input all the time.*
3. *We process it phrase by phrase to come up with the output. This is how we can translate this sentence as humans.*
4. *Basically, we paid more attention to certain words in the source sentence while generating different words in the output.*

Need of the Attention-based model:

1. *Attention mechanisms are being increasingly used to improve the performance of Neural Machine Translation (NMT) by selectively focusing on sub-parts of the sentence during translation.*
2. *Conventional encoder-decoder architectures for machine translation encoded every source sentence into a fixed-length vector, irrespective of its length, from which the decoder would then generate a translation.*
3. *This made it difficult for the neural network to cope with long sentences, essentially resulting in a performance bottleneck.*
4. *This are two simple and effective classes of attentional mechanism: a global approach which always attends to all source words and a local one that only looks at a subset of source words at a time*
5. *In local level attention, we use only a subset of entire sentence and based on that subset the context is figured.*
6. *In global level attention, we use the entire sentence to understand and figure sentence context*

Disadvantage of Attention-based model:

1. *The only disadvantage of the Attention mechanism is that it is a very time consuming and hard to parallelize system.*
2. *The main disadvantage of the attention mechanism is that it adds more weight parameters to the model*
3. *Which can increase training time especially if the input data for the model are long sequences.*

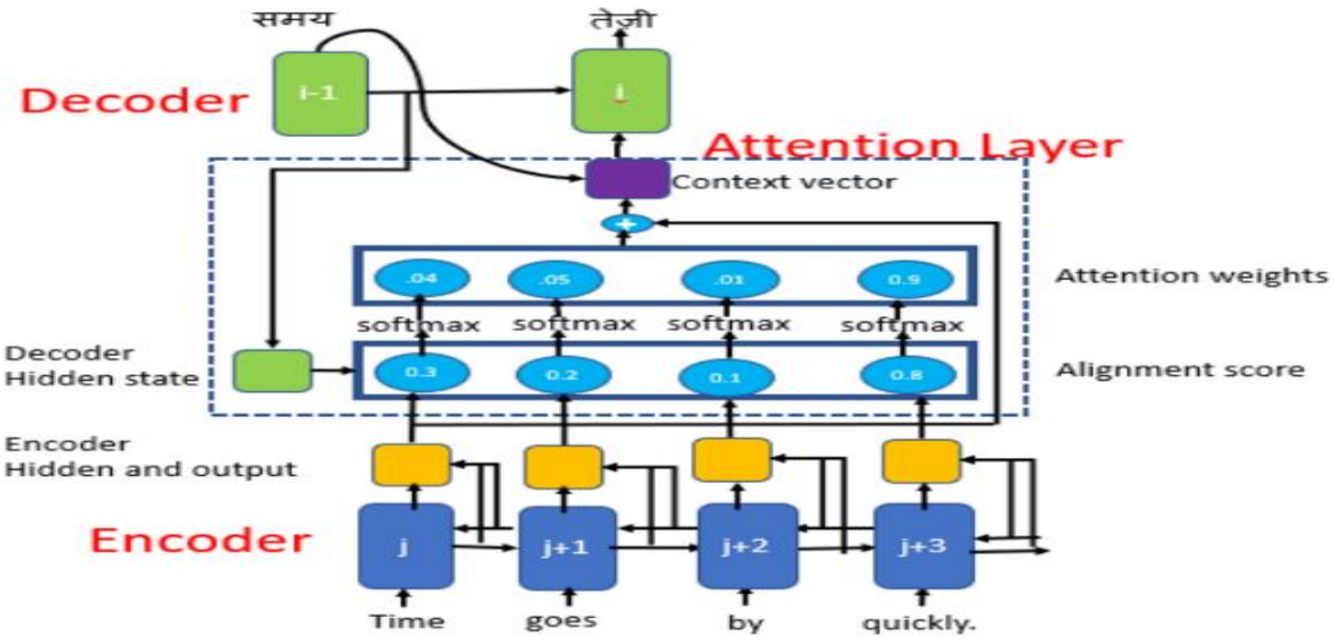
❖ ***Basic encoder decoder model using the Bahdanau Attention Mechanism:***

About the model:

1. *The most important distinguishing feature of this approach from the basic encoder–decoder is that it does not attempt to encode a whole input sentence into a single fixed-length vector.*
2. *Instead, it encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively while decoding the translation.*
3. *This improve the translation performance of the basic encoder-decoder model.*

Need of the Bahdanau Attention Mechanism model:

1. *Conventional encoder-decoder architectures for machine translation encoded every source sentence into a fixed-length vector, irrespective of its length, from which the decoder would then generate a translation.*
2. *This made it difficult for the neural network to cope with long sentences, essentially resulting in a performance bottleneck.*
3. *The Bahdanau attention was proposed to address the performance bottleneck of conventional encoder-decoder architectures, achieving significant improvements over the conventional approach.*
4. *Each time the proposed model generates a word in a translation, it (soft-)searches for a set of positions in a source sentence where the most relevant information is concentrated.*
5. *The model then predicts a target word based on the context vectors associated with these source positions and all the previous generated target words*
6. *Structure of the Bahdanau Attention Mechanism model:*


समय
तेज़ी
Decoder
i-1
i
Attention Layer
Context vector
+
.04
.05
.01
0.9
Attention weights
softmax
softmax
softmax
softmax
Decoder
Hidden state
0.3
0.2
0.1
0.8
Alignment score
Encoder
Hidden and output
Encoder
j
j+1
j+2
j+3
Time
goes
by
quickly.