A Report

On

The Study of Neural Machine Translation

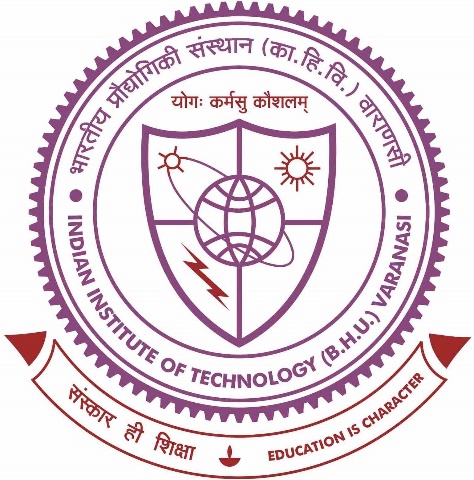
Using

Attention Mechanism

By Shruti Khobragade

At

Indian Institute of Technology, Varanasi (IIT-BHU)



JANUARY 2020

**Acknowledgement**

# The internship opportunity I had with Indian Institute of Technology, Varanasi (IIT-BHU) was a great chance for learning and developing professionally. Therefore, I consider myself as a very lucky individual as I was given the opportunity to be a part of it. I am also grateful to have the chance to meet so many wonderful people and professionals who led me though this internship period.

# It is my radiant sentiment to thanks Prof. Dr. A. K. Singh, Associate Professor, Dept. of Computer Sci. & Engg., IIT (BHU) Varanasi for taking part in useful decision & giving necessary advices and guidance. I choose this moment to acknowledge his contribution gratefully.

# With this moment in view, I would like to share my deepest gratitude and special thanks to Ms. Bhavana Srivastava, Research Scholar, Natural Language Processing Research Lab, Dept. of Computer Science Engg., IIT (BHU) Varanasi who in spite of being extraordinarily busy with her duties, took time out to hear, guide and keep me on the correct path and for her careful and precious guidance which was extremely valuable for my study both theoretically and practically.

# I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

**Project Details**

Date of Start: 09-01-2020

Date of Submission: 04-02-2020

Duration: 09-01-2020 to 04-02-2020 (27 Days)

Title of the Project: Neural Machine Translation using Attention Mechanism

Name/ Discipline of the Student:

Ms. Shruti Khobragade Computer Science and Engineering

Name and Designation of the expert:

Ms. Bhavana Srivastava Research Scholar, NLPRL

Name of Faculty:

Prof. Dr. A. K. Singh Associate Professor

Keywords: Attention based NMT, Sequence to Sequence (seq2seq) model

Project Areas: Machine Learning, Deep Learning

**Abstract**

Neural machine translation is a new machine translation technique that has demonstrated the effective results for high-resource languages. Attention based NMT has started evolving since it pays attention to specific input vectors of the input sequence based on the attention weights. The fundamental idea of Attention is to steer clear of single vector representation for each sentence. The model uses sequence to sequence (seq2seq) network for translating thousands of Spanish words (sentences) to English. It also develops an Attention plot which is an easy-to-visualize alignment matrix between the source and target sentences.

**Contents**

**Acknowledgement……………………………………………………2**

**Project Details…………………………………………………..……3**

**Abstract……………………………………………………………….4**

**Introduction…………………………………………………………..7**

**What is Attention Mechanism?...........................................................7**

**Why we need it?....................................................................................8**

**Example of a deep recurrent**

**Architecture…………………………………………………………..9**

**Bahdanau attention mechanism……………………………………..10**

**Example of the Encoder-Decoder**

**Model with Attention…………………………………………………10**

**Attention Layer……………………………………………………….11**

**Alignment score……………………………………………………….12**

**Attention weights,**

**Context Vector………………………………………………………..13**

**Predicting target word……………………………………………….14**

**Methodology………………………………………………………….14**

**Preparing Dataset…………………………………………………15-16**

**Encoder, Decoder………………………………………………….17-21**

**Evaluation/Result…………………………………………………22-24**

**Conclusion…………………………………………………………….25**

**Introduction**

Neural machine translation is a machine translation approach which uses an artificial neural network to predict the probability of a sequence of words. In Early times, Neural Machine Translation (NMT) models were based on RNN encoder and decoder but due to its inability to translate long sentences, its significance diminished. The Idea of Attention model appeared to fix the existing problem in translating long sentences somewhat.

**What is Attention Mechanism? Why we need it?**

Now, What is Attention? Let’s take an example to understand its significance. When you read a newspaper and when you prepare for an exam, Is the amount of attention to the context same in both the cases? Absolutely, Not. When we are preparing for an exam, we put greater efforts focusing on keywords to help us remember things for a long term. The same thing goes with the Attention Mechanism. The main aim of this mechanism is to focus on a few relevant words in a sentence during predicting its translated word rather than on a single vector carrying information about the entire sentence. Since, Sequence to Sequence (Seq2Seq) models uses encoder-decoder architecture, they map a source sequence to the targeted sequence. In our case, the source sequence is in Spanish and the targeted sequence is in English. When we pass the source phrase in Spanish to the encoder; the encoder encodes the complete source sequence information into a single, real-value vector, which is also known as the context vector. This context vector is then transferred to the decoder to generate an output sequence in a target language such as English. Since, the sole idea of attention lies in to abort learning single vector representation for each sentence, this mechanism not only ease our task but also pays attention to the specific input vectors of the input sequence depending on the attention weight.

At every decoding step, the decoder will be notified about the amount of “**attention**” it has to pay for each input word using a chain of attention weights. Here, the contextual information to the decoder for the translation is provided by **attention weights**.

El pan está fresco. SOURCE SENTENCE

The bread is fresh. TARGET SENTENCE

Figure 1

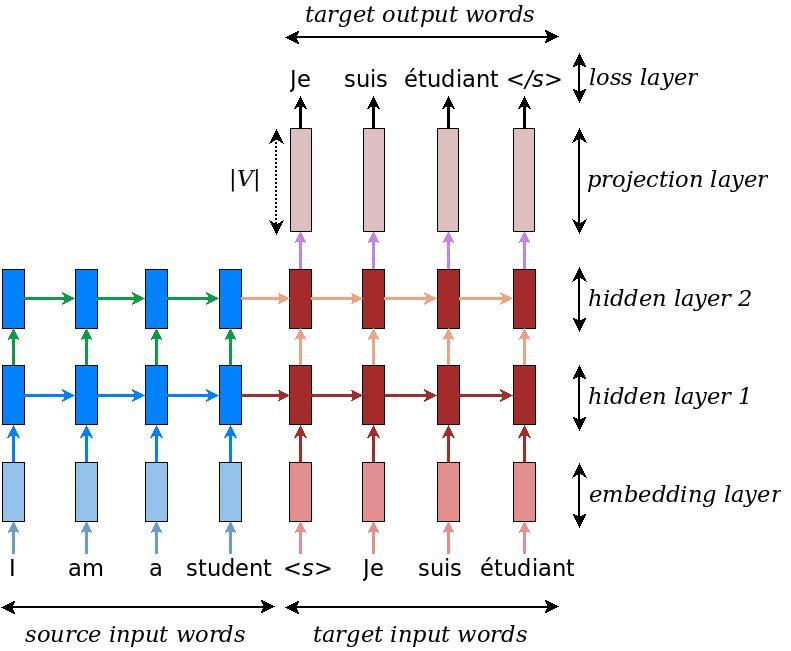
****

Figure 2

Neural machine translation – example of a deep recurrent architecture proposed by for translating a source sentence “I am a student” into a target sentence “Je suis étudiant”. Here, “<s>” marks the start of the decoding process while “</s>” tells the decoder to stop.

(English to French)

**Bahdanau Attention Mechanism**

For my model, I have used Bahdanau Attention Mechanism which runs by jointly learning to align and translate. It is also known as Additive Attention because it performs a linear combination of encoder states and decoder states.

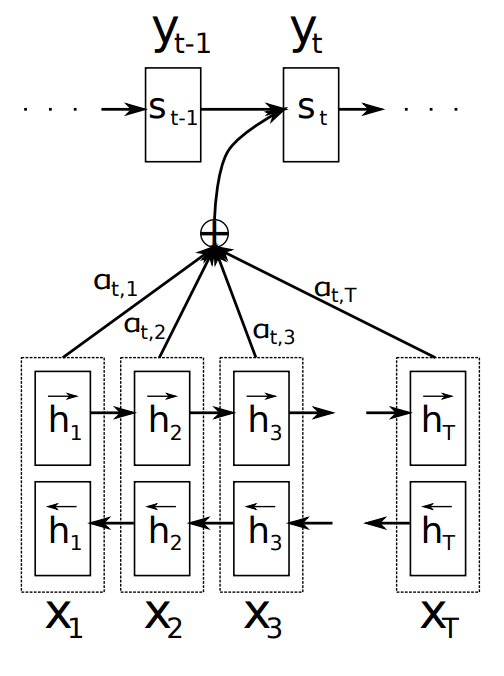


Figure 3

Example of the Encoder-Decoder model with Attention

Taken from “Neural Machine Translation by Jointly Learning to Align and Translate,” 2015.

In Bahdanau attention mechanism, all the hidden states of the encoder (including both forward and backward) and the decoder are used to generate the context vector. This mechanism aligns the input and output sequences that are parameterized by a feed-forward network which helps to pay attention to the most relevant information in the source sequence. The model predicts a target word based on the background vectors associated with the source position and the target words previously produced.

**Attention Layer**

Attention layer contains-

* Alignment score
* Attention weights
* Context vector

Alignment score-

The alignment score maps how well the inputs around position “**j**” and the output at position “**i**” match. The score is based on the previous decoder’s hidden state, **s₍ᵢ₋₁₎** just before predicting the target word and the hidden state, **hⱼ** of the input sentence



Figure 4

The decoder determines which part of the source sentence it needs to pay attention to, instead of making the encoder translate all the source sentence information into a fixed-length vector.

For predicting the first target word, we use the last hidden state of the encoder for the first hidden state of the decoder.

In our example, to predict the second target word, “**wheat**”, we will generate a high score for the input word “**trigo**”.

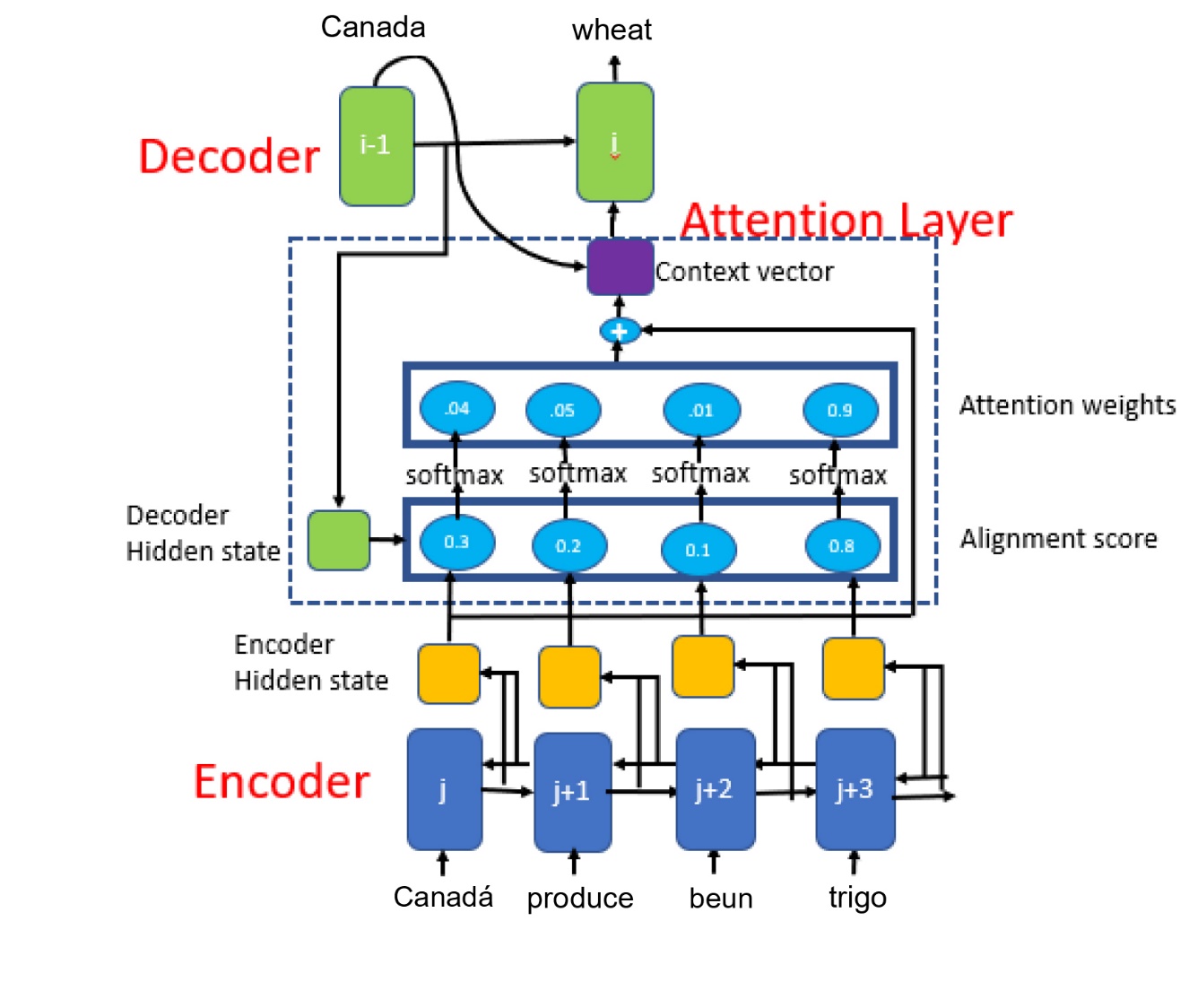


Figure 5

Bahdanau et al. attention mechanism

Attention weights-

We are using softmax activation function to the alignment scores to obtain the attention weights.

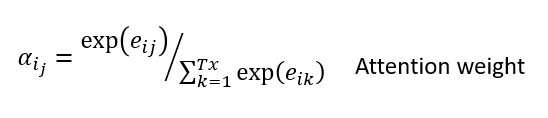


Figure 6

The Softmax activation function will have the probabilities whose sum will be equal to 1, which will help to represent the weight of the effect for each input sequence. The higher the attention weight of the input sequence, the higher its effect on predicting the target term.

In our example, we see a higher attention weight value for the input word “**trigo**” to predict the target word “**wheat**”.

Context Vector-

The context vector is used to calculate the final output of a decoder. The context vector ci is the weighted sum of the attention weights and the hidden encoder states (h1, h2, ..., htx) that map to the input sentence.

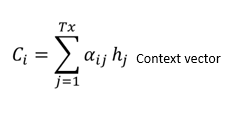


Figure 7

**Predicting the Target Word**

To predict the target word, the decoder uses-

* Context vector (**𝒸ᵢ**)
* Decoder’s output from the previous time step (**yᵢ₋₁**)
* Previous decoder’s hidden state (**sᵢ₋₁**)

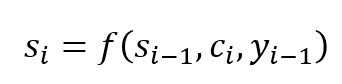


Figure 8

Decoder’s hidden state at time step i

**Methodology**

1. Importing all the essential Libraries.
2. Downloading the dataset.

Dataset**-** <http://www.manythings.org/anki/>

This is an English-Spanish dataset containing all the translation pairs.

1. Preparing the dataset-
2. Converting the Unicode file to ASCII

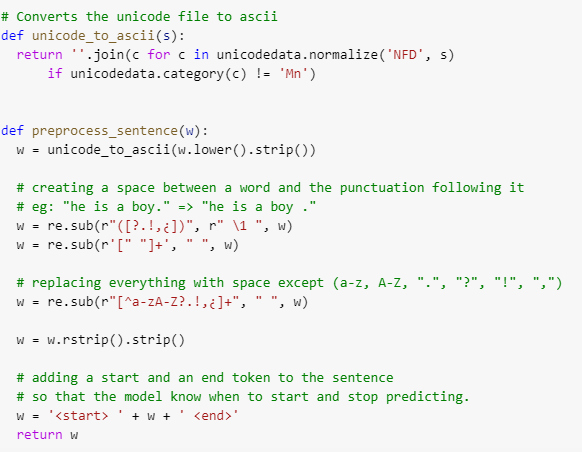


Figure 9

1. Adding a start and end token to each sentence.
2. Cleaning the sentences by removing special characters.
3. Creating a word index and reverse word index (dictionaries mapping from word → id and id → word)
4. Padding each sentence to a maximum length.

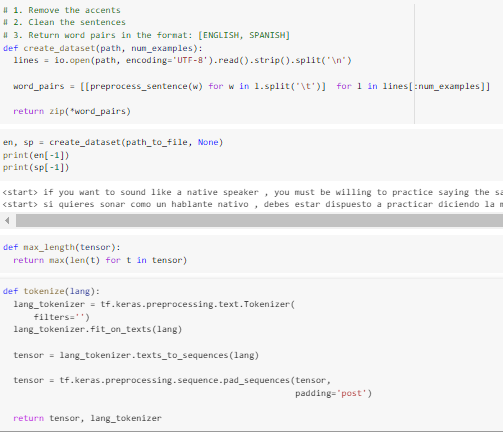


Figure 10

## 

## Encoder and Decoder model

## encoder\_inputs [max\_encoder\_time, batch\_size]: source input words.

## decoder\_inputs [max\_decoder\_time, batch\_size]: target input words.

## decoder\_outputs [max\_decoder\_time, batch\_size]: target output words, these are decoder\_inputs shifted to the left by one time step with an end-of-sentence tag appended on the right.

The input is put through an encoder model which gives us the encoder output of shape (batch\_size, max\_length, hidden\_size) and the encoder hidden state of shape (batch\_size, hidden\_size).

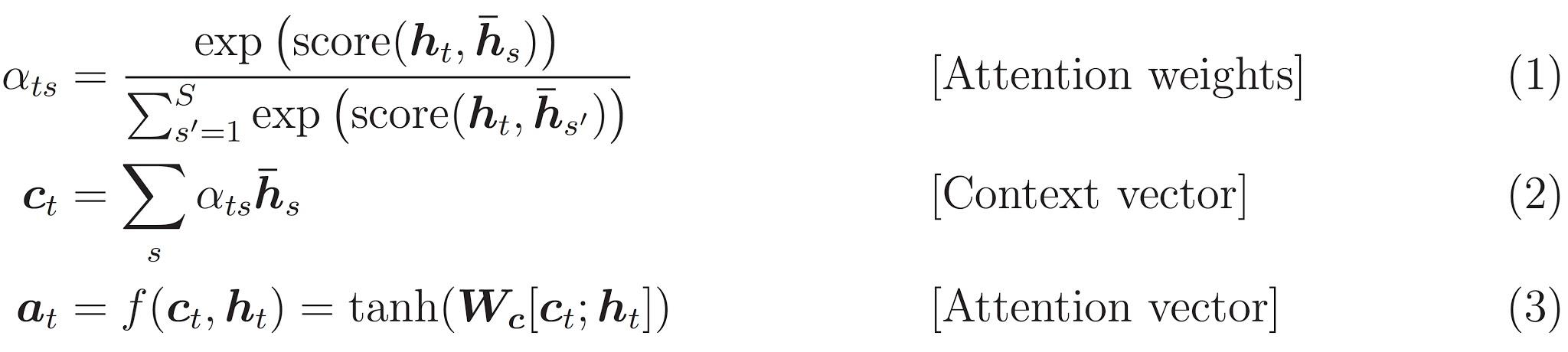


Figure 11

Notations used are as follows-

1. FC = Fully connected (dense) layer
2. EO = Encoder output
3. H = hidden state
4. X = input to the decoder

And the pseudo-code:

1. score = FC(tanh(FC(EO) + FC(H)))
2. attention weights = softmax(score, axis = 1). Softmax by default is applied on the last axis but here we want to apply it on the 1st axis, since the shape of score is (batch\_size, max\_length, hidden\_size).

Max\_length is the length of our input. Since we are trying to assign a weight to each input, softmax should be applied on that axis.

1. context vector = sum(attention weights \* EO, axis = 1). Same reason as above for choosing axis as 1.
2. embedding output = The input to the decoder X is passed through an embedding layer.
3. merged vector = concat(embedding output, context vector).

This merged vector is then given to the GRU

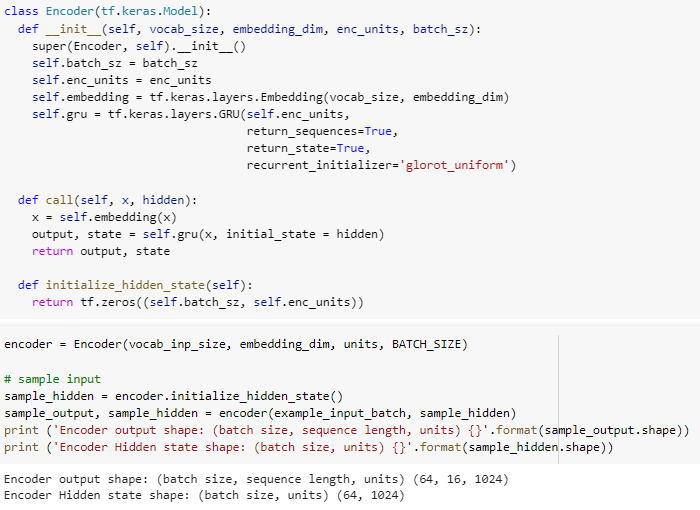


Figure 12



Figure 13

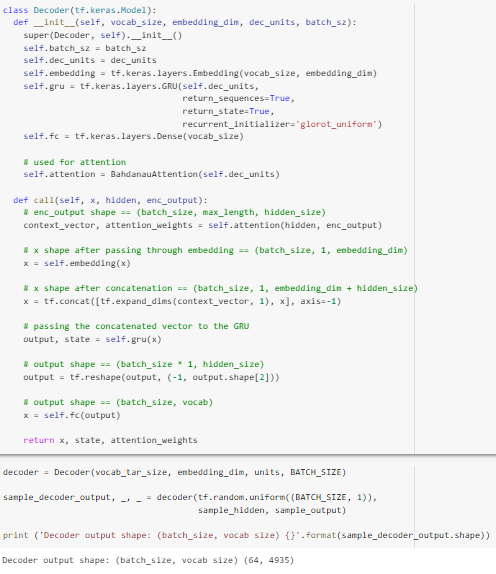


Figure 14

1. Evaluation/Result

Example 1-

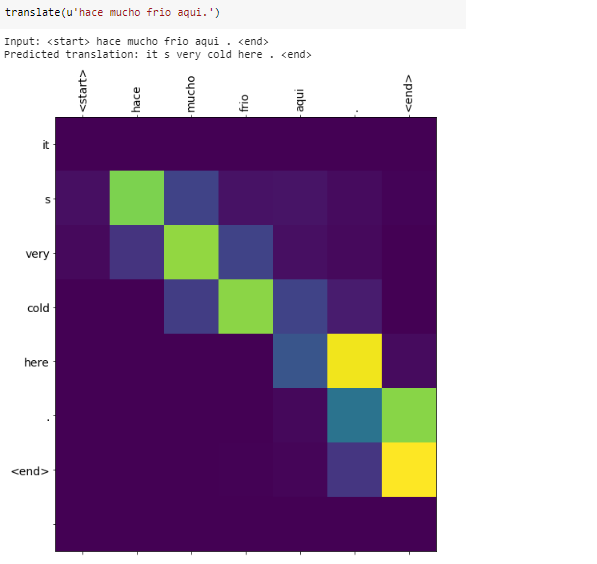


Figure 15

Example 2-

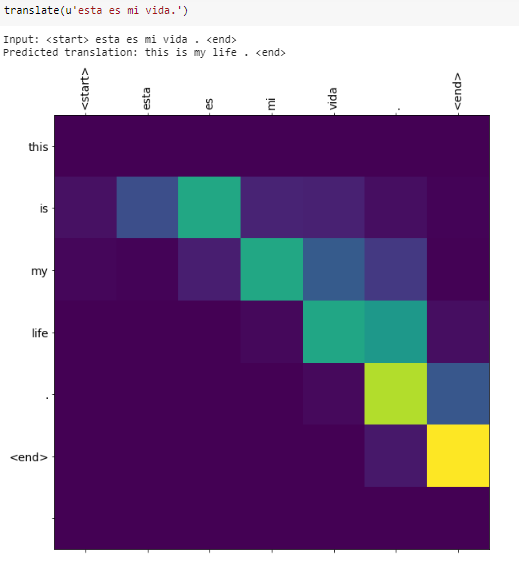


Figure 16

Example 3-

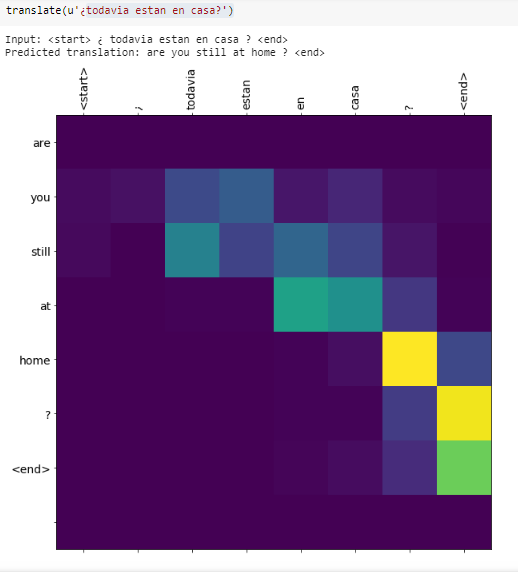


Figure 17

**Conclusion**

In this model, I presented an attention based NMT machine which is capable of translating thousands of Spanish words to English. We can try the same mechanism for large no. of data as well for more appropriate results. We will try to collect more data and do more work on neural models to increase its translation efficiency.

**References**

1. [Neural machine translation by jointly learning to align and translate Dzmitry Bahdanau.](https://arxiv.org/pdf/1409.0473.pdf)
2. [Effective Approaches to Attention-based Neural Machine Translation: Minh-Thang Luong Hieu Pham Christopher D. Manning.](https://arxiv.org/pdf/1508.04025.pdf)
3. [Attention based syllable level Neural Machine Translation System for Myanmar to English Language pair.](http://aircconline.com/ijnlc/V8N2/8219ijnlc01.pdf)
4. [Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. NIPS.](https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf)
5. [Neural machine translation with attention in Tensorflow](A.%09https:/www.tensorflow.org/tutorials/text/nmt_with_attention)
6. <https://github.com/tensorflow/nmt>