Neural Machine Translated communication system to diminish language barrier.

Nishant Banjade
Jain(deemed-to-be) university

ABSTRACT

Language Barrier is one the most gruelling challenges we face in our day-to-day life. Person can be skillful in his/her work but may fail in many places just because of the language barrier. Even in colleges, offices or in remote works, language plays a predominant role in communication.
To diminish this issue, we use Neural machine translation with attention to decipher this issue. The model is basically direct to convert one source language to another targeted language using encoder and decoder architecture. The model encodes the message sent by the sender to a vector of fixed length and decoder generates the translated message which is received by the receiver in their communication system(chat application) automatically. We use the Bahdanau Attention mechanism in our communication system. Using an automatic language translation system, people can be connected around the world and talk in their own mother tongue through the neural machine translated chat application.

1. INTRODUCTION

Neural Machine Translation is the process of converting a sequence of words from a given language to sequence of words to another language using deep neural networks.
The model is designed to train a single and large neural network in order to predict correct translation by reading the given sentence. Following architecture uses an attention model in order to ameliorate the translation where the input sentence may contain long term dependencies.

While speaking, a human's brain remembers only the important keyword while someone speaks and understands the theme of the sentence. So, in the attention model, the machine remembers only the important keyword from the sentence and uses it to decode the sentence. After we pass the source sentence into the encoder, the neural network reads the sentence and encodes it into the fixed-length vector. After this, a decoder outputs translation using the encoded vector. It uses conditional probability from the encoder to get maximum probability in output given a source sentence.

One problem in having a fixed length vector is that it may not be efficient for the longer sentences. It may be difficult to squeeze the longer sentence in a fixed length vector. So, it won't takes whole input sentences into a single vector instead, It encodes the input sentence into a separate sequence of vectors and chooses a subset of these vectors pliant
while decoding the translation. This helps the model to handle longer sentences. In each time stamp, a model generates a word in a translation, it

searches for the most favourable word from the source sentence and concentrates it with the previous generated word and so on. The model then predicts a target word based on the context vectors kindred with these source locus and all the previous generated target words.

After getting the translation with great performance, we are ready to deploy our model to our chat application system where users are allowed to select their favourable language. For this proposed research project, we use only spanish-english language pair as a prototype. To communicate with the person that talks in Spanish, we send the message in English language which is converted by encoder-decoder attention model into Spanish language and the receiver gets the message in Spanish. Likewise, if the receiver replies in Spanish, the sender gets the message in English language.

2. BACKGROUND: NEURAL MACHINE TRANSLATION

If we observe through the perspective of classical probability, our task is to find target message y given a source message x. Mathematically, we represent it as argmax(y) p(y | x). Basically, we try to maximize the conditional probability of sentence pairs using parallel training corpus. Once the model learns the conditional probability model, corresponding translation can be generated by searching the sentence that has higher conditional probability. Once the translation is generated, we save the translated model and deploy into the web application for the communication system. 2.1 Gated Recurrent Unit Vanishing gradient problem used to be a major problem in the past while training sequence models. If the input sentence has long term dependencies, the model may not generalize properly. So, in order to remove this problem, the Gated Recurrent unit came into action. To solve the vanishing gradient problem of a standard RNN, GRU uses update gate and reset gate. update gate helps the model to determine how much of the previous information needs to be passed in the future. Reset gate is used from the model to determine how much of the past informative information to be forgotten. Current memory content uses both forget and reset gates to store relevant information.

$$
\begin{aligned}
z &= \sigma(W_z \cdot x_t + U_z \cdot h_{(t-1)} + b_z) \\
r &= \sigma(W_r \cdot x_t + U_r \cdot h_{(t-1)} + b_r) \\
\tilde{h} &= tanh(W_h \cdot x_t + r * U_h \cdot h_{(t-1)} + b_z) \\
h &= z * h_{(t-1)} + (1 - z) * \tilde{h}
\end{aligned}
$$

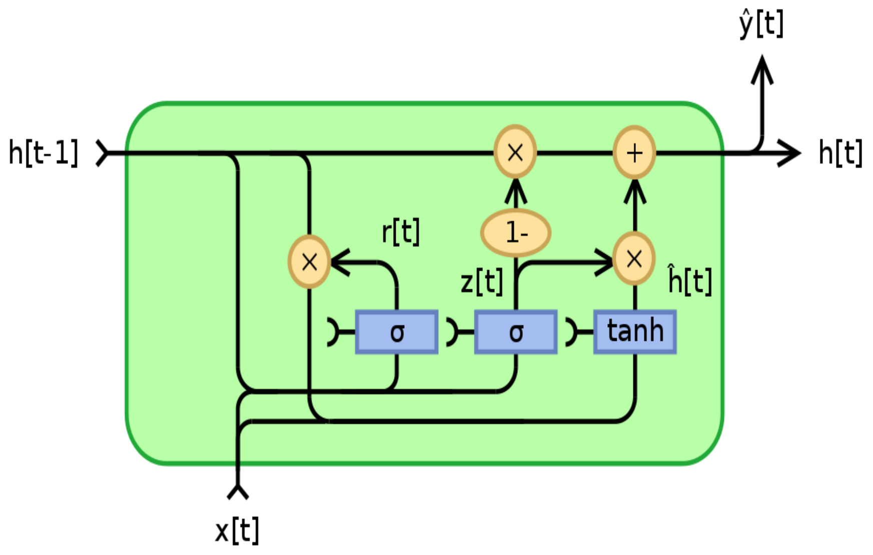
xt = input vector

z = update gate vector

r = reset gate vector

ht = output vector

W, U & b : parameter matrices and vector



2.1 Encoder Architecture

The encoder-decoder architecture is responsible for translation of the sentence.

The Encoder takes the input as the source tokens, after it is passed into the embedding layer for the dense representation of the vector, and is finally passed into the GRU.

2.2 Bahdanau Attention Layer

Attention has 3 layers in it’s architecture. It consists of Alignment score, Attention weights, context vector. Let us briefly discuss it.
The context vector depends on a sequence of annotation (h1, h2, … hx) by using it, an encoder maps the sent message. The context vector is the summation of all these annotations.
Pseudocode

Key points

- X = Input sequence
- FC = fully connected dense layer
- enc_output = Encoder output
- H = Hidden state

- score = FC(tanh(FC(enc_output) + FC(H)))
- attention weights = softmax(score, axis = 1). Max_length is the length of our input.
- context vector = sum(attention weights * enc_output, axis = 1).
- embedding output = The input to the decoder X is passed through an embedding layer.
- merging = concat(embedding output, context vector)
- This merged vector is then given to the GRU

2.3 Decoder

The Decoder has Fully connected layer, GRU layer and embedding layer.It uses a context vector which is the sum of attention weights and the output given by the Encoder, output from the Decoder’s previous time step and the previous hidden state of the Decoder. We use Adam optimizer for the optimization task. In keras , we can define Adam optimization as

optimizer = tf.keras.optimizers.Adam()

2.4 Loss

For computing the loss between an actual and the predicted sentences or say in our case, message, we use SparseCategoricalCrossentropy loss. As it uses an integer rather than vector, it is computationally as well as memory efficient. When using SparseCategoricalCrossentropy the targets are represented by the index of the category (starting from 0). SparseCategoricalCrossentropy is more useful and efficient when you have a lot of categories.

2.5 Dataset

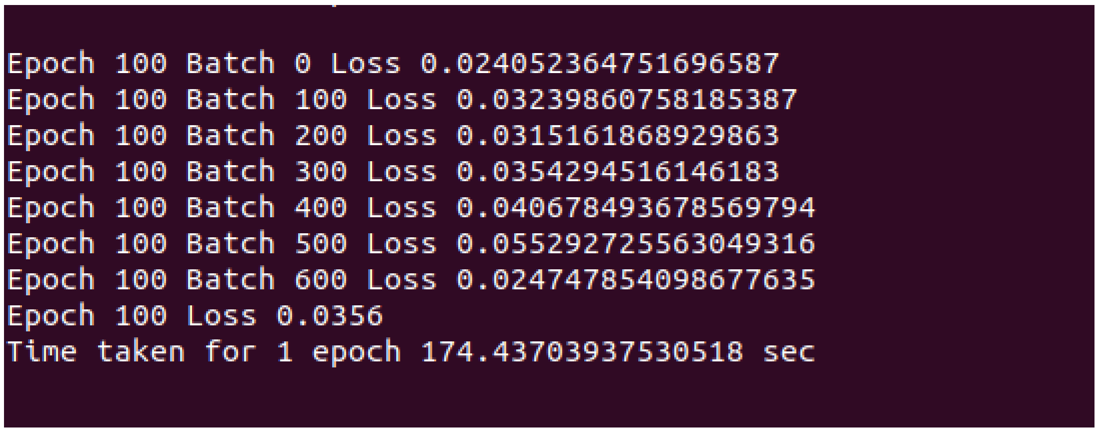
To train our English-Spanish model, we use the English-Spanish dataset from kaggle. The dataset consists of about 102904 unique english texts & 113322 spanish texts having two columns.

2.6 Train the dataset

After the encoder returns the encoder output sequence and the hidden states, encoder output, it’s hidden state, and the decoder’s input is passed to the Decoder. Decoder now predicts the word and again it’s hidden state is passed back in order to calculate the loss. We are using teacher forcing for training where we are passing the actual word to the decoder at each time step.

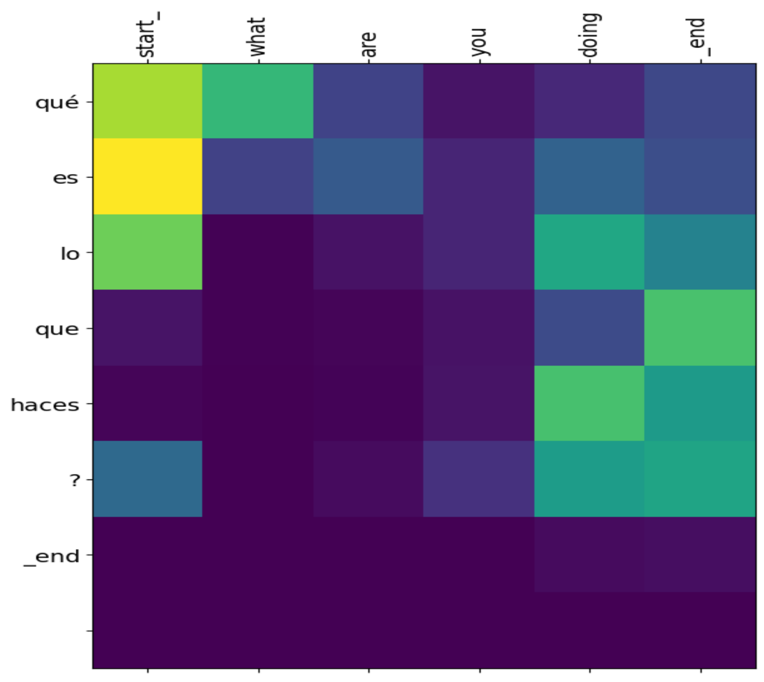
Basically, predicted words from the previous state act as the input in the present state. As long as these things are working, we apply Gradient descent , Adam optimization and most important back propagation. We are using 100's epochs which took about 4.5 hour to train using Ryzen 5 processor, tensorflow GPU and 8 GB primary memory where we got loss of only 0.03 which is very good accuracy.

2.7 Loss and training



```
Epoch 100 Batch 0 Loss 0.024052364751696587
Epoch 100 Batch 100 Loss 0.03239860758185387
Epoch 100 Batch 200 Loss 0.0315161868929863
Epoch 100 Batch 300 Loss 0.0354294516146183
Epoch 100 Batch 400 Loss 0.040678493678569794
Epoch 100 Batch 500 Loss 0.055292725563049316
Epoch 100 Batch 600 Loss 0.024747854098677635
Epoch 100 Loss 0.0356
Time taken for 1 epoch 174.43703937530518 sec
```

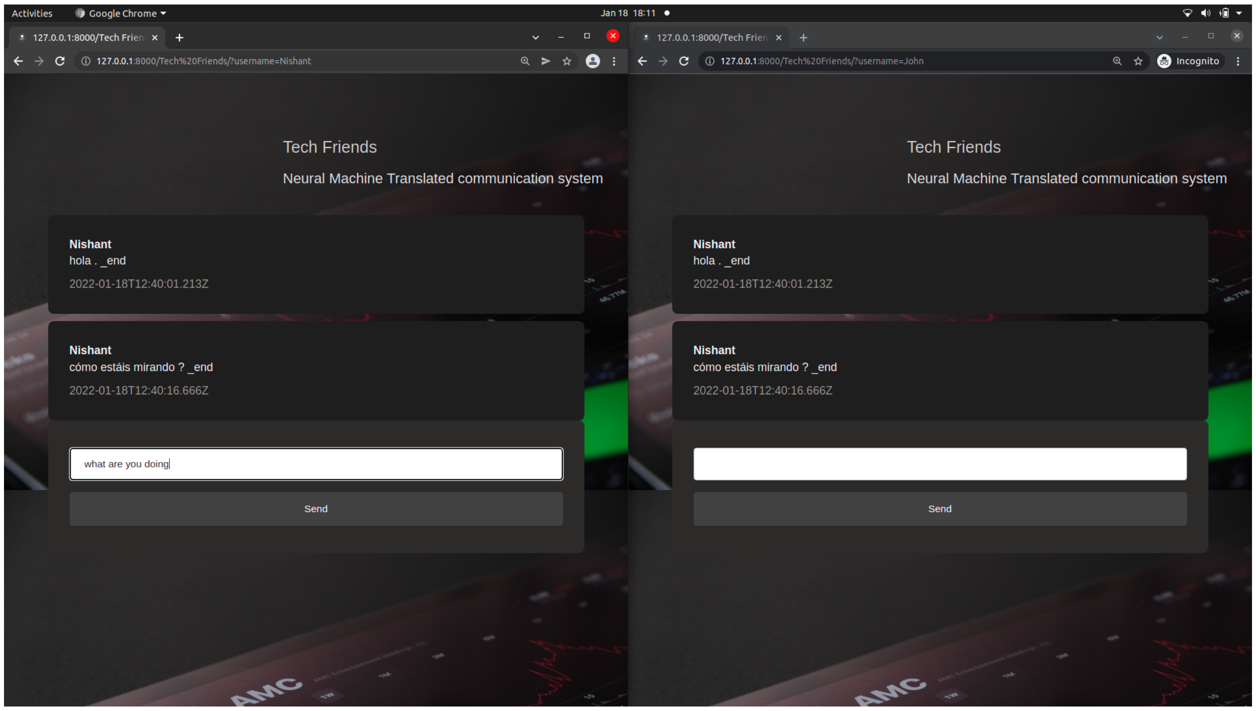
2.8 Attention Plot



3. Web application integration

One of the beautiful ways to represent and demonstrate the idea of Machine learning model is deployment of the model in web application.
As per the objective of the model, we are developing a neural machine translation communication system. So, whenever a person types in English, his/her message gets converted to spanish message which helps spanish people to communicate properly.

Type in English


Activities Google Chrome Jan 18 18:11

127.0.0.1:8000/Tech%20Friends/?username=Nishant

Tech Friends

Neural Machine Translated communication system

Nishant
hola . _end
2022-01-18T12:40:01.213Z

Nishant
cómo estáis mirando ? _end
2022-01-18T12:40:16.666Z

what are you doing

Send

127.0.0.1:8000/Tech%20Friends/?username=John

Incognito

Tech Friends

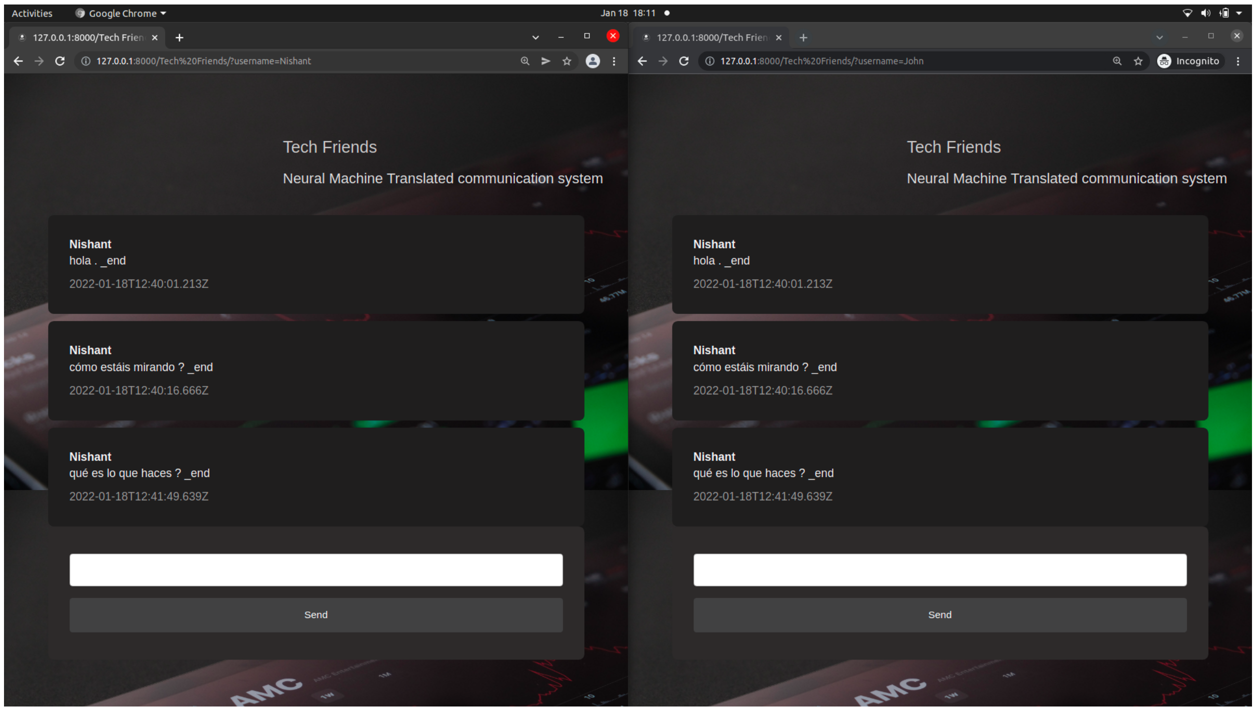
Neural Machine Translated communication system

Nishant
hola . _end
2022-01-18T12:40:01.213Z

Nishant
cómo estáis mirando ? _end
2022-01-18T12:40:16.666Z

Send

Receive in Spanish


Activities Google Chrome Jan 18 18:11

127.0.0.1:8000/Tech%20Friends/?username=Nishant

Tech Friends

Neural Machine Translated communication system

Nishant
hola . _end
2022-01-18T12:40:01.213Z

Nishant
cómo estáis mirando ? _end
2022-01-18T12:40:16.666Z

Nishant
qué es lo que haces ? _end
2022-01-18T12:41:49.639Z

Send

127.0.0.1:8000/Tech%20Friends/?username=John

Incognito

Tech Friends

Neural Machine Translated communication system

Nishant
hola . _end
2022-01-18T12:40:01.213Z

Nishant
cómo estáis mirando ? _end
2022-01-18T12:40:16.666Z

Nishant
qué es lo que haces ? _end
2022-01-18T12:41:49.639Z

Send

This chat room helps to chat with any Spanish people in their own language. Input the sentence in English, our attention model translates the model in spanish. Thus, if this propose is used & provided good support, the language barrier from the world will end soon and everyone will be proudly speaking their mother tongue language.