Indic Language NLP model Tamil-English

Mentor:
Mr Mahendra Rajan
Mr Gokul
Faculty Mentor:
Dr S P Rajamohana

Done By :
John Alphonse Raja V
Deepak Pandian A

Date: 08.07.2022

Problem statement

- Replicating the process to build the model (AI4Bharat Indic Trans) will help validating the approach and potentially helps understanding the end-to-end process of data collection, pre-processing, parameters involved.

Proposed Approach

- Download corpus used by AI4Bharat
- Convert to csv form with it’s equivalent translations
- Preprocess the text
- Load the dataset to GPU
- Train the model
- Test the model once every 5 epoch

Text preprocessing

1 | Tokenization:

Breaks the raw text into words

Example:

```
"Let's see how it's working."
```

```
['Lets', 'see', 'how', 'its', 'working']
```

2 | Vocabulary

Set of unique words used in the text corpus

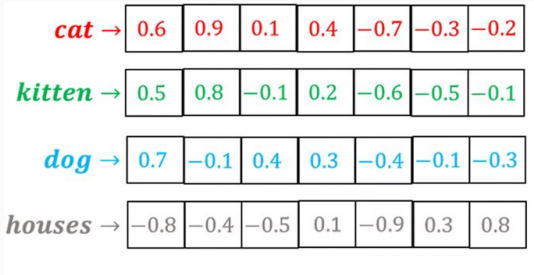
Example:

```
'bob ate apples, and pears', 'fred ate apples!'
```

```
['bob', 'ate', 'apples', 'and', 'pears', 'fred']
```

3 | Word Embedding

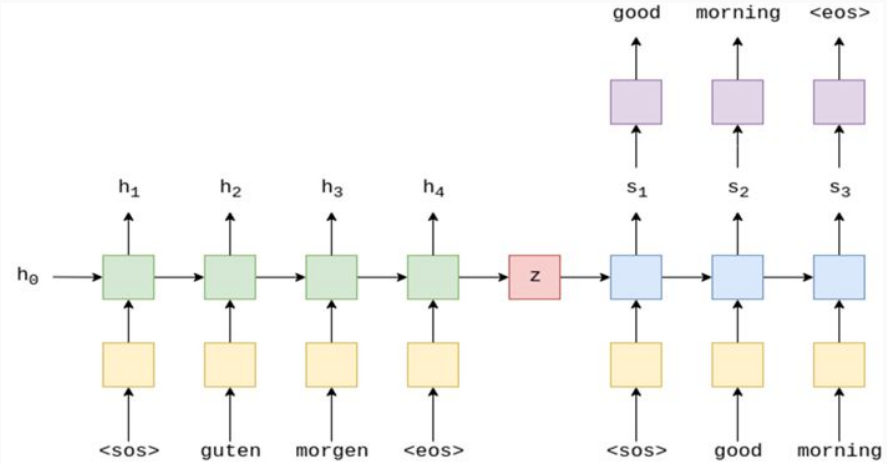
Representation of words in the form of a real-valued vector



Seq2Seq

Encoder
The encoder iterates through the input sentence one token at a time, at each time step outputting “hidden state” vector.

Decoder
The decoder generates the response sentence in a token-by-token fashion. It uses the encoder’s context vectors to generate the next word in the sequence.



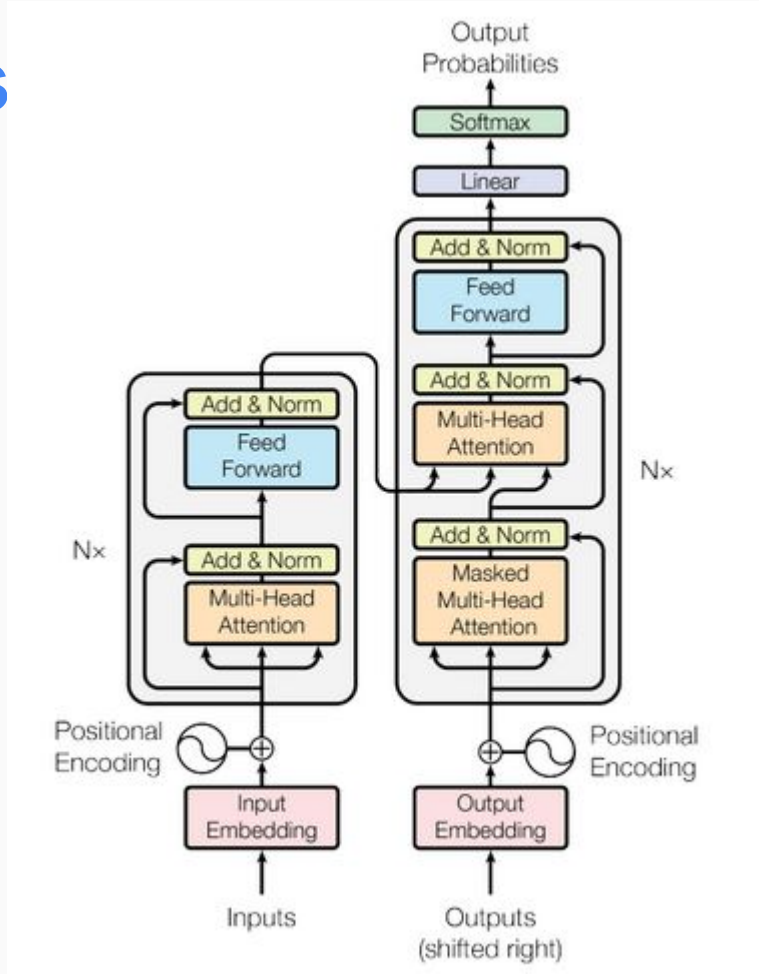
Drawbacks

- The output sequence relies heavily on the context defined by the hidden state in the final output of the encoder, making it challenging for the model to deal with long sentences.
- Trained it on only 1000 sentences which took an hour for 150 epochs
- In the case of long sequences, there is a high probability that the initial context has been lost by the end of the sequence.
- Takes a longer time to train and does not use GPU’s effectively.

Transformers

- Transformers were introduced in 2017 by a team at Google Brain and are increasingly the model of choice for NLP problems.

- Unlike RNNs, transformers do not necessarily process the data in order. Rather, the attention mechanism provides context for any position in the input sequence.



Work Done and Challenges faced

Work Done:

- Trained the model on 20,000 sentences for 220 epochs (6 hours)
- Was able to predict correct translation of tamil sentence in train dataset
- Was able to partially predict correct translation in unseen sentence about four to five words
- Got a BLEU score above 75

Challenges faced:

- Need more GPU’s to train on large dataset
- Takes longer time to train in GPU
- Colab runtime gets disconnected

Deploying the Model Locally

- Converted the code to change the model into evaluation mode
- Downloaded the trained model from colab
- Developed frontend using flutter
- Language model act’s as backend
- Interfaced the language model with flutter using flask api