PROJECT REPORT

MACHINE LEARNING UML501

SUBMITTED BY

SHIVESH NARULA (102003128)

ARNAV SINGH (102003136)

BE THIRD YEAR

SUBMITTED TO:

Mr. NIYAZ WANI



Submission Date-16/11/2022

**Table of Contents**

Objective

Introduction

Dataset Description

Modules Used

Methodology

Results

Objective

In this machine learning project, we will develop a Language Translator App using a many-to-many encoder-decoder sequence model. We will train our model using LSTM which will convert English to French language where English will be input text and French will be the target text. For this, we will be using English-French dataset.

Introduction

When ever we are in a foreign country or even in some distant parts in our country we struggle with communication with other people because of language barrier.

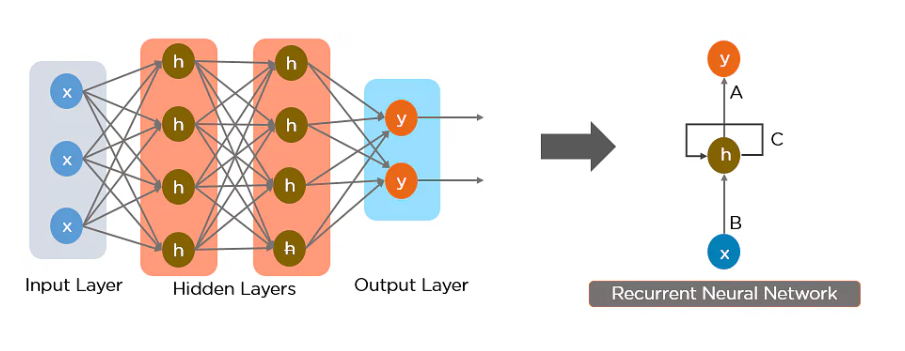
With this project we aim to minimize this problem by creating a French to English translator that takes English sentences as output and returns their French translation.

This could further be expand to work on voice recognition as well.

RNN

**Recurrent Neural Network(RNN)** A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation, natural language processing (nlp), speech recognition, and image captioning.

While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depend on the prior elements within the sequence.

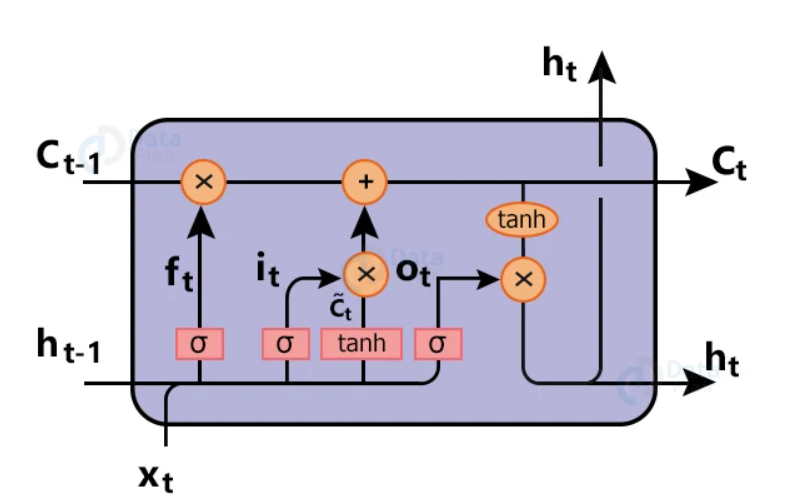


LSTM (Long Short Term Memory)

 LSTM stands for long short-term memory networks, used in the field of [Deep Learning](https://intellipaat.com/blog/tutorial/machine-learning-tutorial/introduction-deep-learning/). It is a variety of [recurrent neuralnetworks(RNNs)](https://intellipaat.com/blog/tutorial/artificial-intelligence-tutorial/recurrent-neural-network/) that are capable of learning long-term dependencies, especially in sequence prediction problems. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images. This finds application in speech recognition, machine translation, etc. LSTM is a special kind of RNN, which shows outstanding performance on a large variety of problems.

Vanishing Gradient: While training our models, in order to get the best output we have to minimize the loss i.e. errors after every time step. This can be achieved by propagating backward and calculating the gradients, that is loss with respect to weights applied to every vector at different time steps. We repeat this process until we get an optimal set of weights for which the error is minimum. After reaching at some time step gradient value becomes so less that it approximates to zero or gradient vanishes. After reaching that limit, the network stops training. This leads to the problem of vanishing gradient.

These are some issues which are resolved by the LSTM networks. Instead of a single neural network layer, LSTM has three gates along with hidden and cell states.



**Cell Memory state ( ct ):** Cell state is actually what makes LSTM a unique network. Cell state holds the memory for over a long period of time. Data can be removed or added in cell state depending upon the layer requirements.

**Hidden state ( ht ):** hidden state is basically output of the previous block. We decide what to do with the memory looking at the previous hidden state output and current input. And also we don’t want output after every timestep until we reach the last input of our sequence.

**Forget Gate ( ft ):** Forget Gate is used to check what data we want to neglect away from the cell state. This is done using a sigmoid layer. This Gate looks at hidden output from previous time steps and current input, after that it outputs number 0 which means neglect the data, or 1 means keep the data.

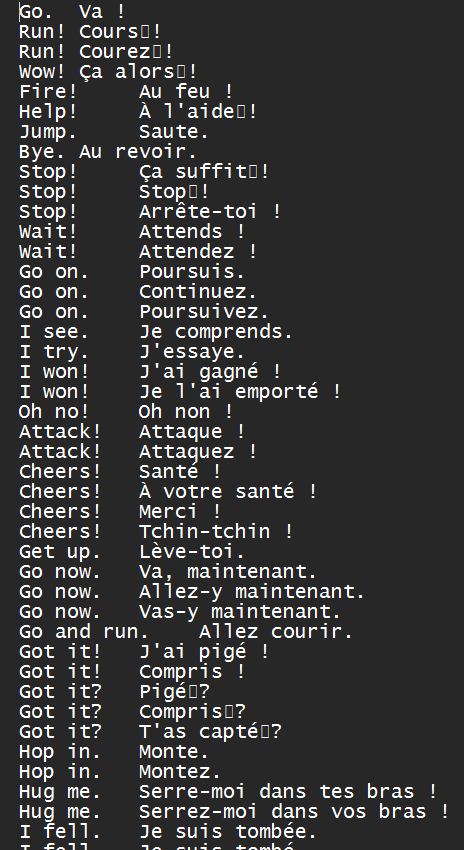
**Input Gate:** We want to check what new information we are going to store in the cell memory state ( ct ). So data will pass through the sigmoid function which decides which values to update ( it ). Next a tanh function creates a vector of new candidates ( čt ) for our state.

Dataset Description

The English to French dataset used in this model is available on Kaggle.

The dataset contains 660,291 words and sentences in English with their French

translation



Modules Used

**Tensorflow**: TensorFlow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks.The TensorFlow platform helps you implement best practices for data automation, model tracking, performance monitoring, and model retraining.

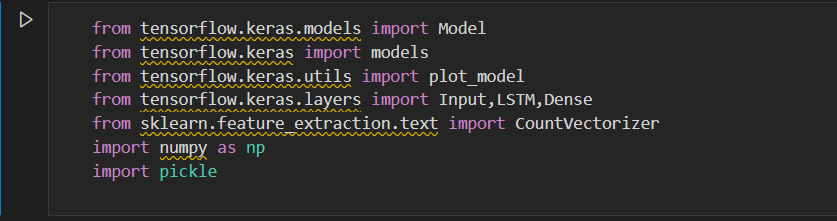
**Pickle:** Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script.

**Sklearn:** Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

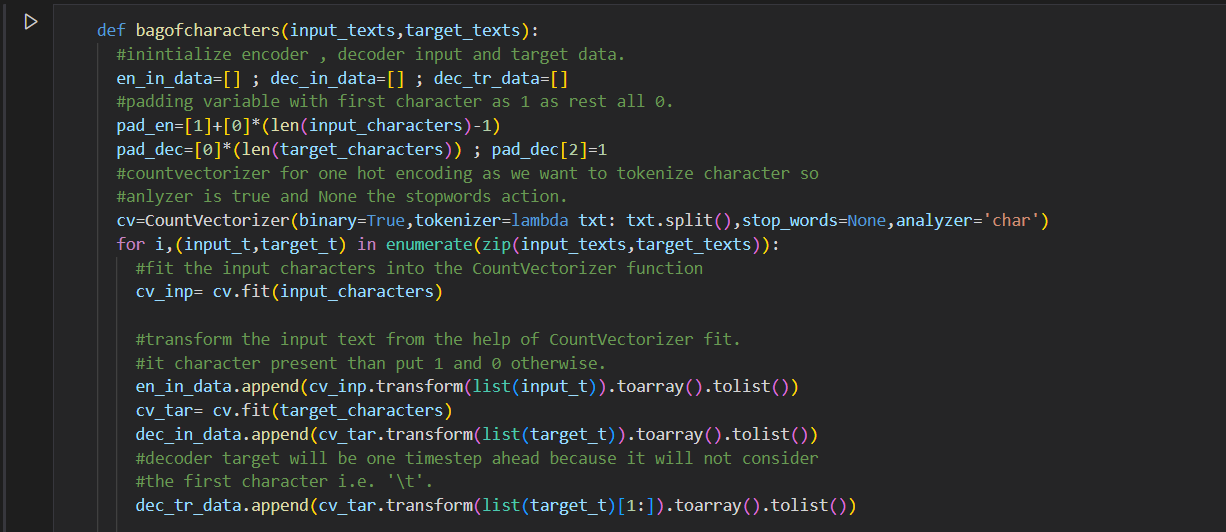
**Numpy:** NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

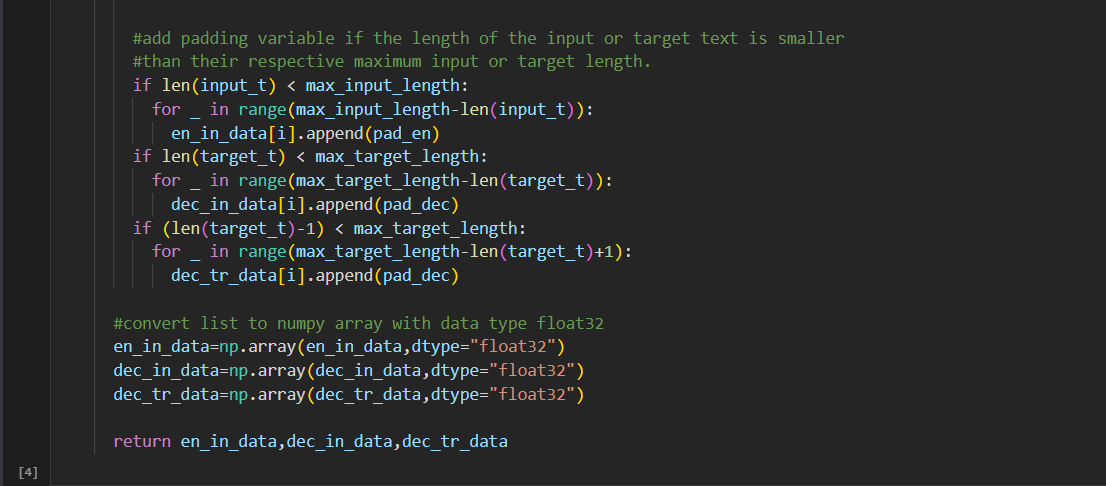
Methodology

Importing Libraries

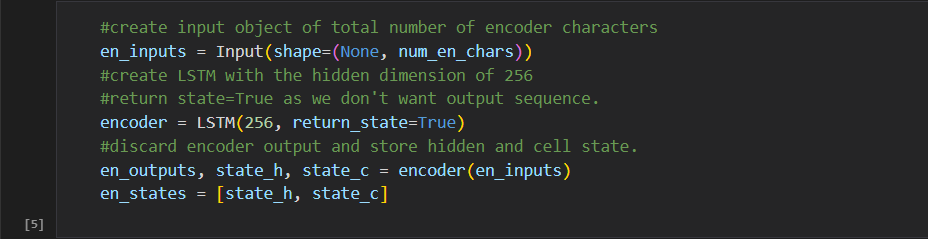


One hot Encoding

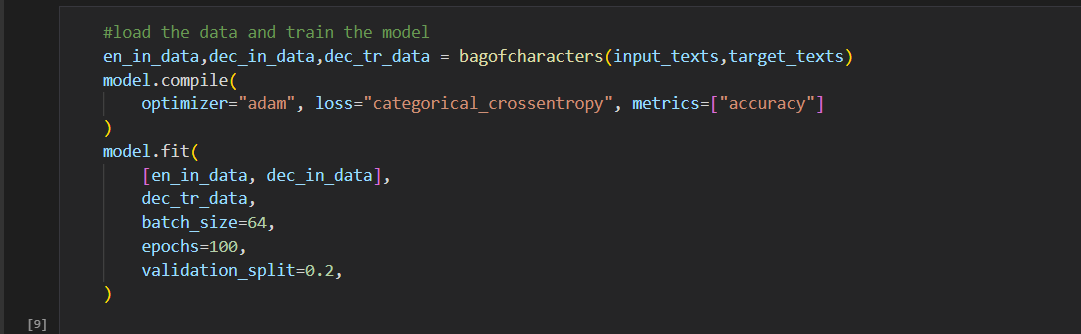


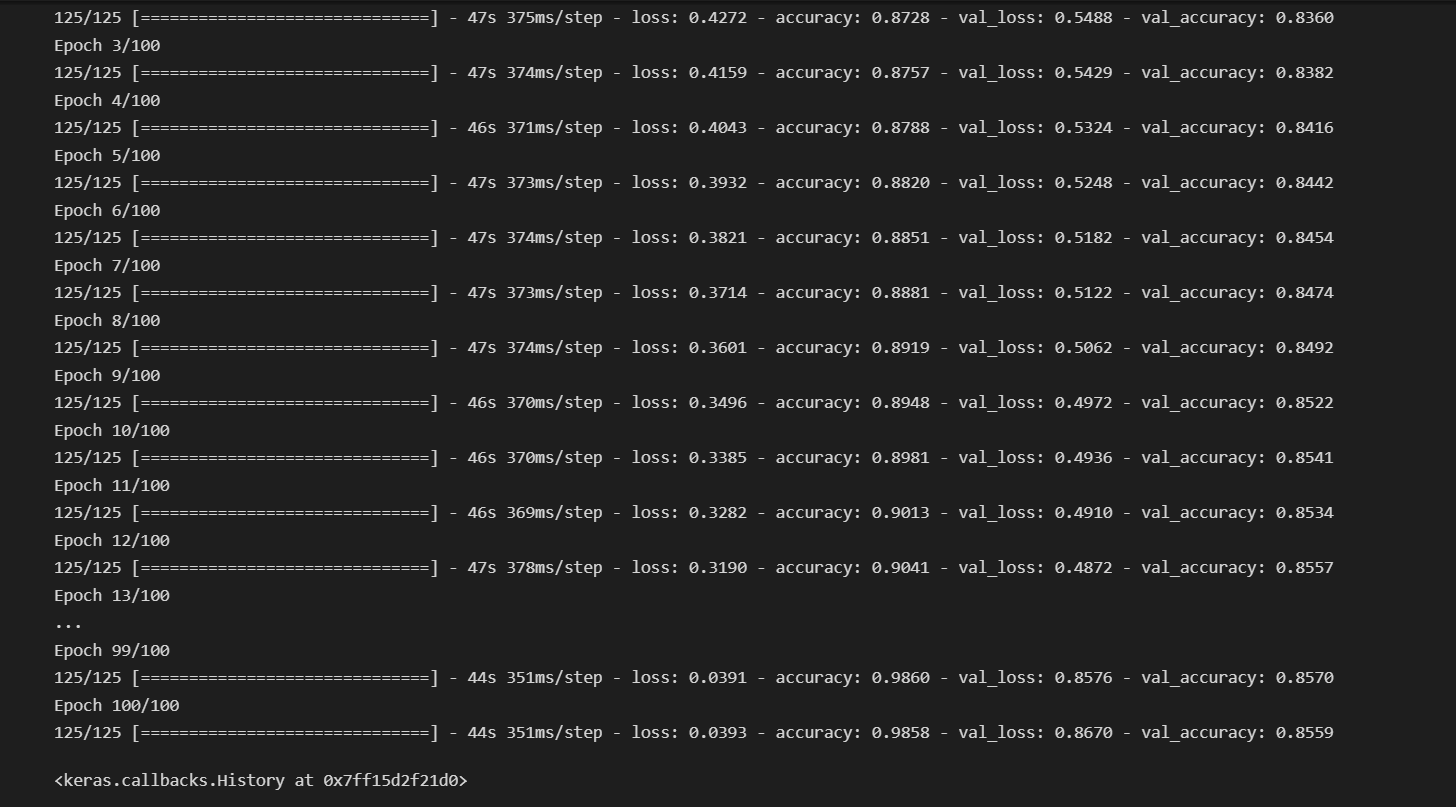


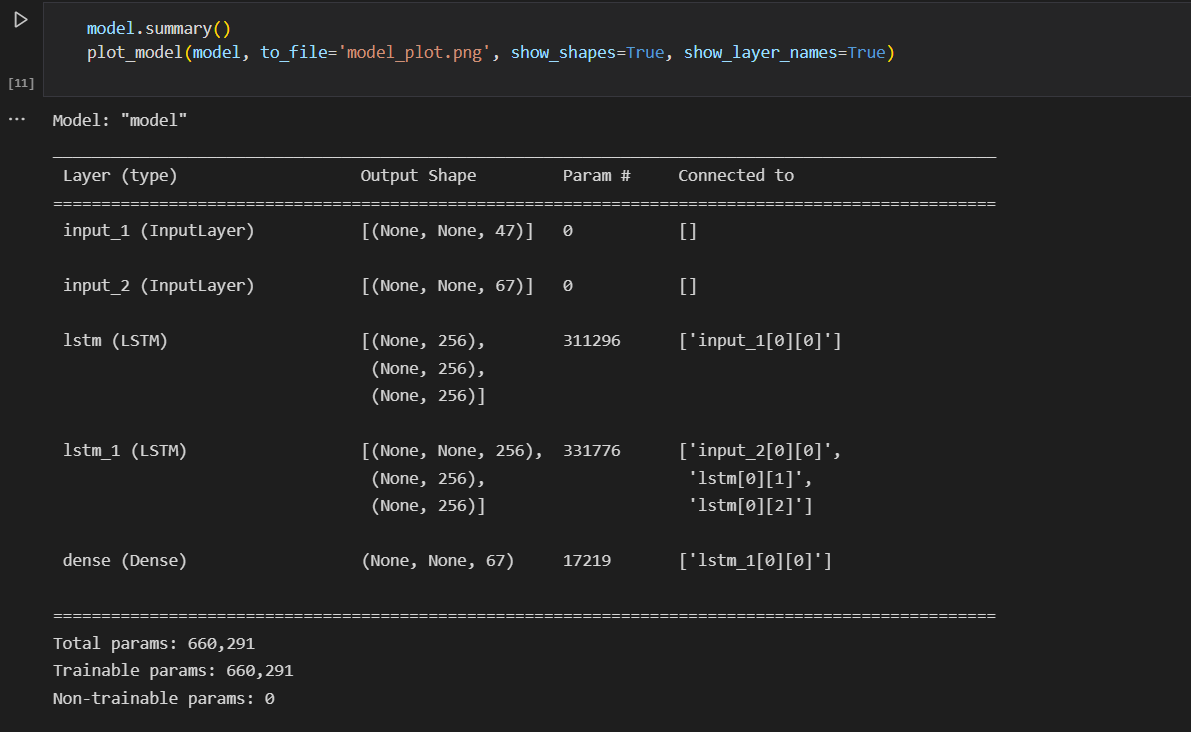
Creating LSTM



Loading and Training the Model







Results

