csoc-ig report

Shambhavi

June 2025

SEQUENCE MODEL BASICS

1 Introduction

The code involves executing sequence-based modeling at two levels:

1. RNN
2. LSTM or GRU

2 Preprocessing

I found there are 207 null values in the second column, so i dropped them because the data was immense and that negligible amount would not affect the data.

3 Learnings

As I started word embedding using word2vec, it took a lot of time just running and my RAM was running out of storage. To make it faster, i try to change some values like vector size, min count and workers. I create a class RNN using PyTorch. I used tanh for hidden layers and sigmoid for the last layer. As there was a lot of data, the RAM was out of memory, so I used gc (garbage collection) and del to remove unnecessary data. I first used BCEWithLogitsLoss for the loss function but that was giving a negative loss. I realized that the targets were 1 and 2, this was causing problems, so i changed them to 0 and 1 after that the loss function worked well. Smilarly I created Class LSTM using LSTM function of PyTorch library.I kept the epoch to be 10 and learning rate is 0.01.

4 Accuracy

Accuracy of RNN model is about 65 percent and its F1 score is 0.75. Accuracy of LSTM model is about 80 percent and its F1 score is 0.90. RNNs struggle with long sequences because their memory fades over time. LSTMs use memory cells

and gates to selectively retain important past information.This makes LSTMs ideal for tasks like sentiment analysis and speech recognition.

5 Summary

LSTM works better than RNN