IT412: Natural Language Processing

Assignment 5: Fake News Classification

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn text classification using deep learning and representing words using Seq2Seq models.

1 Problem description

Text classification is the task of assigning a text to its proper category. The sequence of words in a sentence can be represented by using Seq2Seq models. These models capture the context of a sentence and represents the sentence in low dimensional sentence. A fully connected network can then be trained to learn the classification.

- LSTM: Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems). A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

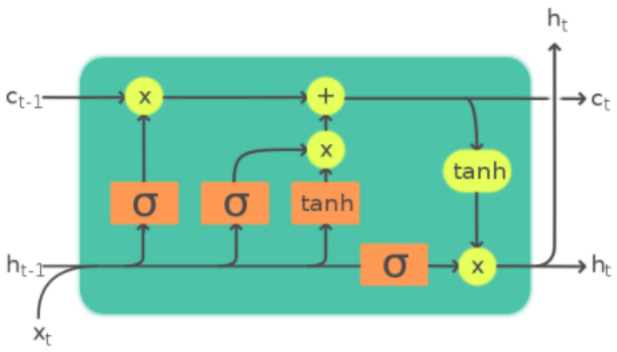

Figure 1: LSTM Cell

2 Implementation

2.1 Dataset

- For this assignment we will use Fake news dataset from Kaggle.
  https://drive.google.com/file/d/1q0ZimHCtMlhftljfVuy0i-wcgPnDUYjl/view?usp=sharing
- train.csv: A full training dataset with the following attributes:
  1. id: unique id for a news article
  2. title: the title of a news article
  3. author: author of the news article
  4. text: the text of the article; could be incomplete
  5. label: a label that marks the article as potentially unreliable 1: unreliable 0: reliable
- test.csv: A testing training dataset with all the same attributes at train.csv without the label.
- submit.csv: A sample submission

2.2 Exercise

- Use the train.csv file for this experiment.
- Divide the dataset into 90:10 ratio.
- Use the title of the news articles and model it using LSTM.
- Use Adam optimizer as optimization function and for calculating loss you can Cross Entropy loss. Select appropriate batch size and learning rate for training your model.
- Use accuracy for evaluation of the classifier

3 References

- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_files.html
- https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html
- https://pytorch.org/docs/stable/data.html
- https://www.kaggle.com/awalber94/fake-news-prediction-using-lstm-neural-network
- https://colab.research.google.com/drive/1n9Iz3R5oEnN1kGBwdhKL7h4RrFSBH9OO?usp=sharing

4 Submission

- You have to submit your assignment in notebook with proper comments and explanation of your approach.
- Report the accuracy for the approach
- The submission deadline for this assignment is **20th September 2021 11 pm**