IT412: Natural Language Processing

Assignment 10: Deep Learning based Text Classification

Instructor: Prasenjit Majumder

1 Text Classification

Classification is a supervised machine learning problem. Classification can be binary i.e given mail is spam or not, multi-class i.e categorize a given news as sports, entertainment, politics etc. Aim of **Text Classification** is to automatically classify a given document. Text classification is the task of assigning a set of predefined categories to free-text. Text classifiers can be used to organize, structure, and categorize pretty much anything. Few examples of text classification are span detection, sentiment analysis, fake news detection etc.

1.1 Approaches to text classification

1.1.1 Traditional approaches

There are many statistical classifiers like Naive Bayes, Logistic Regression, Support Vector Machine(SVM) etc.

1.1.2 Deep learning approaches

Deep learning models are able to beat traditional classifiers. RNN and CNN are used for sentence classification.

1. RNN based text classification

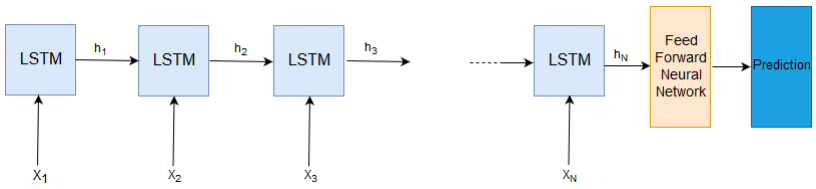

Figure 1: RNN based text Classification

Fig 1 shows block diagram of RNN based text classification. Given a sentence containing N words, each word is converted into it's corresponding embedding represented by X_i which are input to RNN's. Hidden state of each cell are passed to next cell, final hidden state is passed to feed forward neural network and last layer contains K number of neurons, where K is number of classes.

2. CNN based text classification

   Fig 2 shows block diagram on CNN based text classification. Given a sentence containing N words, each word is converted into it's corresponding embedding represented by X_i, where $X_i \in R^k$, where k is word embedding size. Each X_i are concatenated and represented as $R^{N \times k}$. After this different filters of size $R^{h \times k}$ are convolved to get feature maps, where h is the size of filter. Once feature map corresponding to a given filter is obtained max-pooling is applied. Finally obtained vector is passed through feed-forward neural network and final layer is according to number of classess.

2 Implementaion

2.1 Dataset

For assignment purpose you are provided with clickbait dataset, where each headline is classified as Clickbait / no-clickbait. Table 1 shows example of dataset.

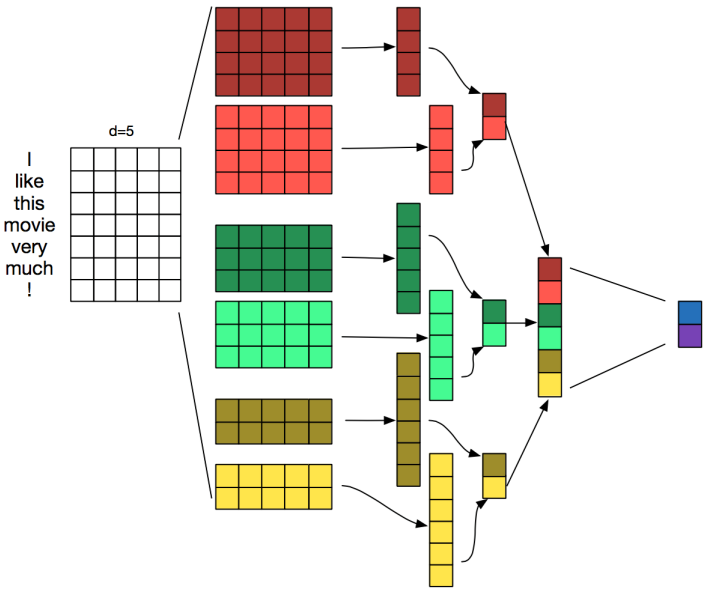

Figure 2: CNN based text classification

Table 1: Example of dataset

headline	**truth**
Watch People Get Their Mind Blown To The Possibility Of Winnie The Pooh Being A Girl	clickbait
American surgeon Michael E. DeBakey dies at age 99	no-clickbait
27 Important Questions All Teachers Have Asked	clickbait

2.2 Implementation Excercise

1. Train binary text classifier using CNN with pre-trained embedding.
2. Train binary text classifier using RNN/BiRNN/BiRNN with attention with pre-trained embedding.
3. For both trained models, calculate Accuracy and F1 Score for both models on test set.

Note: Hyperparameters can be as per your choice.

3 References

- Convolutional Neural Networks for Sentence Classification
- Comparative Study of CNN and RNN for Natural Language Processing
- Understand LSTM

3.1 Codes

- https://pytorch.org/tutorials/beginner/text_sentiment_ngrams_tutorial.html
- https://github.com/AnubhavGupta3377/Text-Classification-Models-Pytorch/tree/master/Model_TextCNN
- https://github.com/prakashpandey9/Text-Classification-Pytorch

3.2 Dataset

List of few publicly available dataset.

- *Topic classification*: Reuters news, 20 Newsgroup
- *Sentiment Analysis*: Amazon Product Review, IMDB reviews, Twitter Airline Sentiment
- Spam detection
- Hate speech and Offensive language