UNIVERSITY OF PADUA

INFORMATION ENGINEERING DEPARTMENT (DEI)

MASTER'S DEGREE IN COMPUTER ENGINEERING

Comparing Graph Neural Network Performances Against Standard Natural Language Techniques

Prof.
Fabio Vandin

Students:
Francesco Caldivezzi
Simone D'Antimo
Riccardo Rampon

Contents

1 Motivations

Text classification is an important and classical problem in natural language processing.
The goal in the text classification task is to build systems which are able to automatically classify any kind of documents performing different operations such as document organization, news filtering, spam detection, opinion mining, and computational phenotyping. The feature space, based on the set of unique words in the documents, is typically of very high dimension, and thus document classification is not trivial. Currently much research has been made about classical text classification methods such Recurrent Neural Network (RNN), Convolution Neural Network (CNN) and many libraries such FastTex , instead, only a limited number of studies have explored the more flexible Graph Convolutional Neural Networks (GCNN).
In this paper the goal is to compare the effectiveness of classical techniques for text classification with respect to Graph Neural Networks (GNN). Comparison will be done by using different datasets, different tran-validation dataset-split and so on. The aim of this experiment is to explore the potential of the Graph Neural Networks compared with classical techniques like : Rocchio Classifier, Support Vector Machine, Recurrent Neural Network, Word Vectors, etc., and, find out if the Graph solution has better performance than the others in this particular scenario.

2 Methods

Text classification is a, supervised problem (i.e. the dataset consists of labelled examples) that, given :

- A description $d \in X$ of a document, where X is the document space.
- A fixed set of classes $C = c_1, c_2, ...c_n$ also known as tags or labels.

The goal is to find a model that, trained with the usage of a training set D, of labelled descriptions of documents i.e. $< d, c > \in X \times C$, is able to generalize. In particular, in the dataset [2] that will be used for making the comparisons that will explained later, $< d, c >$ are couples like the following :

$$< d, c > = < limbal\ autograft\ reconstruction\ cell\ carcinoma\ patient..., C04 >, \quad (1)$$

where $C04$ represents 'Neoplasms', which is a cardiovascular disease. Using a learning method, then, a classifier γ is returned, which is going to map documents into classes : $\gamma : X \to C$.
Therefore, to solve the above explained problem the following learning methods [5] will be implemented:

- Machine Learning Methods :
  - Rocchio Classifier.
  - Naive Bayes Classifier.
  - K-nearest neighbor.
  - Support Vector Machine (SVM).
  - Decision Tree.
  - Boosting/Bagging methods (Random Forest, Adaboost, . . .).
  - Conditional Random Field.
  - Word Vectors.
- Deep Learning Methods :
  - Recurrent Neural Networks (RNN).
  - Convolutional Neural Network (CNN).
  - Recurrent Neural Network (RCNN).
  - Graph Convolution Neural Network (GCNN) [6].

3 Intended Experiments

This section is going to introduce the intended experiments that will be done with the methods explained in the previous section. First of all, the dataset [2] used for the experiment that will be executed consist of three different dataset called respectively : R8, R52 and Ohsumed. In particular :

- R8 and R52 : are subsets of Reuters [3] 21578 dataset. Such dataset is composed of 12902 documents with 90 classes.
- Ohsumed : is a dataset [4] built by excrating medical abstracts from the MEDLINE [1] database. In particular, it consists of medical abstracts from the MeSH categories of the year 1991 and provides 23 cardiovascular diseases categories as classes.

The main experiments that will be run are :

- computation of Accuracy, F1 score, Precision and Recall metrics for different methods introduced in previous section.
- analysis of the different metrics depending on the dataset used for training.
- train and test with different train-test splits for datsets.
- eventually, train and test the previous techniques on other datasets.

4 References

[1] Medline website. `https://www.nlm.nih.gov/medline/medline_overview.html`.

[2] R8, r52 and ohsumed datasets. `https://www.kaggle.com/datasets/weipengfei/ohr8r52`.

[3] Reuters-21578 text categorization collection data set. `http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html`.

[4] Reuters-21578 text categorization collection data set. `http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html`.

[5] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[6] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.