Project Draft

Zerui Tian and Yixing Zheng
{zeruit2, yixingz3}@illinois.edu

Group ID: 64, Paper ID: 79
Code link: https://github.com/yixingz3/DL4H_team_project

1 Introduction

This work aims to provide a new effective disease inference method utilizing symptoms extracted from Electronic Medical Records (EMR) data. The relationship between symptoms and diseases is represented by the term frequency-inverse document frequency (TF-IDF) model. And a bidirectional recurrent neural network (Bi-LSTM) is utilized to model the symptom sequences in EMR data. This combination of models shows a significant improvement in disease inference - 4% to 10% on average improvement from the two baseline models (Guo et al., 2018).

2 Scope of reproducibility

This paper introduced a new disease inference method utilizing the combined model of TF-IDF and Bi-LSTM that outperforms two similar existed models on the task disease inference based on symptoms. Utilizing the TF-IDF and Bi-LSTM combination will provide a higher accuracy than utilizing DeepLabeler and WordVec + Bi-LSTM with the same extracted symptoms data. Besides the TF-IDF, WordVec is also a method of embedding symptoms and converting them into vectors. This method is also considered in this paper, but it does not perform as well as TF-IDF in term of accuracy.
We decide to choose this claim as it is the major contribution of the paper and we would like to verify its validity with our own implementation of the combined model. In addition, in the course Deep Learning for Healthcare, we learned the word embedding method of WordVec, but this paper uses TF-IDF to achieve better results, and we have a strong interest in it. We wanted to try out different word embedding methods and compare them or combine them to achieve better results.
In the meanwhile, to replicate this article we need the relevant knowledge of natural language processing. We are passionate about this field and we want to apply the knowledge we learn in the course to solve some relevant problems.

2.1 Addressed claims from the original paper

Clearly itemize the claims you are testing:

- **TF-IDF + Bi-LSTM** provides higher accuracy on disease inference than **DeepLabeler** and **WordVec + Bi-LSTM** given the same extracted symptoms input data.

3 Methodology

We couldn't locate any code and the original data set that was extracted from EMR records using MetaMap by the authors and used to train the proposed and baseline models. Therefore, we will re-implement the proposed model based on the paper's description as well as the baseline models. So that we can compare and validate the claim with our implementation and potentially a different data set with a very similar structure.

3.1 Model descriptions

For the draft, we have re-implemented the TF-IDF + Bi-LSTM model. TF-IDF is proposed to represent text documents as vectors of identifiers. We can use TF-IDF to model the relationship between symptoms and diseases. Now each symptoms is transformed to vector representation. For symptoms S_i, it can be represented as follows:

$$S_i = (W_{i,1}, W_{i,2}, \ldots, W_{i,d})$$

And $W_{i,j}$ is the strength of the association between symptoms i and diseases j used TF-IDF method.

$$W_{i,j} = TF_{i,j} * \mathbf{log}\frac{N}{D_i}$$

Here N is the number of all diseases we take into consideration and D_i denotes how many diseases associate with the symptoms i. $TF_{i,j}$ is the number of symptom i in the discharge summaries correlated with disease j.

Since RNNs have the problem of vanishing gradients, it is difficult to handle long sequences of data. The special case of RNN, LSTM (Long Short-Term Memory), is improved by RNN. It can avoid the gradient disappearance of conventional RNN, so it has been widely used in the industry. LSTM has three gates, namely forget gate, input gate and output gate. These three gates cooperate with each other and can solve the problem of gradient disappearance to a certain extent. And bidirectional LSTM can better extract information. So the model adopts bidirectional LSTM as the neural network structure.

About the evaluation metrics, the measurements are defined as follows:

$$MiP = \frac{\sum_{i,j} y_i^j \hat{y}_i^j}{\sum_{i,j} \hat{y}_i^j}$$

$$MiR = \frac{\sum_{i,j} y_i^j \hat{y}_i^j}{\sum_{i,j} y_i^j}$$

$$MiF1 = \frac{2 * MiP * MiR}{MiP + MiR}$$

Here y_i^j stands for the true label i of sample j, $\hat{y}_i^j$ stands for the predicting label i of sample j.

In general, the structure of this paper is to first extract symptoms, then use TF-IDF to represent symptoms as vectors, and then input them into bidirectional LSTM for prediction and evaluation.

3.2 Data descriptions

The data set we used to train and evaluate our model implementation is the IMDb data set due to the lack of original data set at the moment. We have contacted the authors of the paper in the hope to retrieve the original data set that was used by them, but have not received any responses yet. On the other hand, we have retrieved the MIMIC-III data set that was used for the paper after being processed by MetaMap, a software that extracts the symptoms from the discharge summary in MIMIC-III. As of now, we have submitted the request for using MetaMap, but the application is still pending. In our final submission, we hope to either receive the data from the authors or obtained a license for MetaMap so we can process the data ourselves and train our model implementations with a data set that is similar to the original data set.

3.3 Hyperparameters

TODO

3.4 Implementation

We have re-implemented the model ourselves and the code can be found from the following link: https://github.com/yixingz3/DL4H_team_project. Our GitHub repository has a README file that states the files and data set used for our implementation. In each .ipynb file that corresponds to different model implementations, we also included in-line comments describing the process we use for training the model and validating the result.

3.5 Computational requirements

We use the Google CoLab to perform the computation. And as of now, we have only computed the TF-IDF + Bi-LSTM model with it. We couldn't find any specifications for GPU and Memory usage, but the computation took 5 minutes to complete.

For the resources usage and run-time, we were not able to gauge an estimate properly due to the lack of original implementation and source of data. We considered the size of the original data set and determined that it was manageable, which stays valid with our CoLab run with the sample data set. Additionally, at the current stage, we are only training a small subset of the data, so the resources usage and run-time for training the model stays within our capability, even without specifications. We have also obtained data points in terms of how much time it might take to train with a larger data set, and we might be able to use it as a reference for the other two baseline models as well.

Model	GPU	Memory	Time
DeepLabeler	CoLab	CoLab	TBD
WordVec + Bi-LSTM	CoLab	CoLab	TBD
TF-IDF + Bi-LSTM	CoLab	CoLab	5 Mins

4 Results

4.1 TF-IDF + Bi-LSTM provides higher accuracy on disease inference than the other two baseline models

At the moment, we have only completed the computation for the TF-IDF + Bi-LSTM model with a smaller simple data set and thus unable to compare its accuracy with the baseline models. However, we have obtained an average accuracy result of 0.8312 on the test data.

4.2 Analyses

The accuracy result we have from our implementation of the TF-IDF + Bi-LSTM model shows a similar result to the paper (0.854). We believe that the variance of data set would be mainly responsible for the gaps here. Additionally, even with different and smaller amount of data, our implementation performed better than the two baseline models, which is an observation we would expect to see when we build and train the baseline models using the same data set.

4.3 Plans

On the model side, we plan to implement the two baseline models (DeepLabeler and WordVec + Bi-LSTM) so we can validate the claim in the paper. If we might have more availabilty, we might consider implementing the proposed model from the authors' other paper that published later as one more baseline model (Guo et al., 2020). If there is a chance, we will try more word embedding methods and different network structures such as convolutional neural networks to expand the model.
On the data side, we are expecting to obtain licence for the data processing tool MetaMap, so that we could extract the symptoms from MIMIC-III and built our full and final data set. We have also contacted the original authors of the paper asking for their help to provide the data set. Currently, we are not optimistic that we will receive the help we need. With the attempt has been made, we are hoping for the best.

4.4 Additional results not present in the original paper

TODO

5 Discussion

TODO

5.1 What was easy

TODO

5.2 What was difficult

TODO

5.3 Recommendations for reproducibility

TODO

6 Communication with original authors

TODO

References

Donglin Guo, Guihua Duan, Ying Yu, Yaohang Li, Fang-Xiang Wu, and Min Li. 2020. A disease inference method based on symptom extraction and bidirectional long short term memory networks. *Methods*, 173:75–82.

Donglin Guo, Min Li, Ying Yu, Yaohang Li, Guihua Duan, Fang-Xiang Wu, and Jianxin Wang. 2018. Disease inference with symptom extraction and bidirectional recurrent neural network. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 864–868. IEEE.