Reproducibility Project Proposal for CS598 DL4H in Spring 2023

Manu Vinod Shesha
manuv3@illinois.edu

Group ID: 96
Paper ID: 187

1 Original paper

Automated ICD-9 Coding via A Deep Learning Approach, published by Min Li, Zhihui Fei, Min Zeng, Fang-Xiang Wu, Yaohang Li, Yi Pan, and Jianxin Wang [3].

2 Problem Description

The paper aims to automate the extraction of ICD-9 (Ninth Revision of International Classification of Diseases) codes from patient discharge summary, through application of state-of-art (general) text processing technique called Document-to-Vector (D2V) in combination of more general Convolution Neural Network. This extracted codes are used to further do billing, as well as raising insurance claims with the provider. Today, this procedure to extract medical codes from patient discharge summary is largely a manual effort, undertaken by hospital's medical record department personnel. This has two problems:

- The process is very slow and inefficient, causing delay in the patient discharge process.
- The process requires specialized knowledge making it costly, and sometimes error prone.

In addition, any accurate computational system to automate this procedure faces following problems:

- The discharge summaries don't have a fixed format, and hence any approach which depends on consistent syntactic structure of documents is unsuitable
- Health care and, in general, medical field has lot of concepts and terminologies and most of researcher, programmers may not have complete understanding of the domain to build tools which are based on semantic mapping of the concepts (like knowledge graph).
- Health care professionals like doctors have different ways to describe same medical condition. This inconsistency cannot be addressed by conventional approaches.

The authors demonstrate that the advances in the field of Deep Learning can be leveraged successfully in automating the extraction of ICD-9 codes from the discharge reports.

3 Novelty in the approach

This work by the authors is an extension to the line of work done by other researchers, who have applied various Machine Learning techniques like Support Vector Machine (SVM), k-Nearest Neighbors, and other conventional Natural Language Processing (NLP) techniques to similar problems, and achieved varying degree of success. The authors describe a novel approach to this problem through usage of DL techniques. Specifically, the authors combine two approaches to produce vector embedding of documents, which were further used in multi-label classification:

- CNN (Convolution Neural Network) to discover, and extract **local features** in text. We can understand the intuitive reasoning behind it. The local context of the text , like phrases describing a medical concept, should be important in deriving related ICD codes. Reference architecture has been described in this paper [1].
- D2V (Document to Vector) [2] embedding technique to capture the **global features** of the document. D2V extends upon Word2Vec, and (unlike CNN) takes order of words into account. It should be noted that D2V unsupervised learning approach.

The authors point out that CNN alone suffers with certain deficiencies:

- The documents are of different lengths, which will need to be padded with zeroes or truncated before applying them to CNN. This can lead to loss of information.

- CNN ignores the order of words, which can also lead to loss of symantic information.

Both these deficiencies are avoided by D2V because, by design, it takes the order of words in account, and also considers all the words in the document to generate the embedding. Hence it can act as a complementary component to CNN.

As per the authors, at the time of publication of their research paper, no one else had applied D2V technique (in combination with CNN) for the purpose of ICD-9 codes extraction from medical text like discharge report. So the novelty of their method is to combine the two different techniques (CNN and D2V) to extract features from two levels (global and local), and produce dense embedding representing the document. Following inferences were drawn based on outcome:

- The model performed better (based on micro-F measure) in comparison to baseline models flat-SVM and hierarchical-SVM.

- Each of the components: CNN and D2V provide important contribution to the overall accuracy of the model (, with CNN being more important). The authors found that model accuracy degrades significantly when either of these components are removed.

- The model performance improves when more data is used to train it, typical of DL-based models.

3.1 Method

- Embedding generation steps:
  - For a given discharge report, generate a D2V embedding, to capture the semantic information of the whole document. The authors used Gensim Doc2Vec to implement this part.
  - For a given discharge report, generate a document embedding by passing the document through a multi-channel single layer CNN architecture.
    - ∗ The vector of each CNN filter will pass through a max-pooling layer.
    - ∗ The max-pooled output of each kernel filter together form the output embedding of the CNN layer.
  - Concatenate the two vectors generated in above steps, to create the final embedding vector representing the given document.

- Classification steps: Pass the embedding to a fully connected feed-forward layer (with number of neurons equal to the number of ICD-9 codes under classification scope), followed by a softmax layer. This will give the probability of each ICD-9 code per document.

In our reproduction work, we will use Gensim for D2V embedding, and Pytorch (instead of Tensorflow used in original paper) to implement CNN.

4 Hypotheses in-scope for this reproduction study

- In scope:
  - It will be verified that the proposed model performs better (based on micro-F measure) than the baseline models flat-SVM and hierarchical-SVM, by same percentage, as observed by authors in their work. On MIMIC-III dataset, their model achieves micro-average F-measure of 0.408, significantly outperforming flat-SVM (0.253) and hierarchy-based SVM (0.335).

- Out-of scope:
  - The baseline models: flat-SVM and hierarchical-SVM, will not be built. Instead, the team will use the accuracy values provided for these models by the authors, while comparing the results.
  - The hypothesis that model performance improves with inclusion of more training data, will not be verified. In original paper, author test this hypothesis by training/testing the model on MIMIC-II dataset, which has lesser number of discharge documents, and got degraded micro-average F-measure compared to model trained/tested on larger corpus in MIMIC-III dataset.

5 Additional ablations

The team will verify that each of the components: CNN and D2V, are important to the overall accuracy of the model, with CNN being more important, by building additional two networks, each with one missing component. This will be done to verify authors claim that each of these components play a different, yet, complementary role in capturing the features of a document. The authors observed that CNN part is the most effective component as micro-average F-measure drops to 0.308 without CNN part. D2V is also important as the micro-average F-measure drops to 0.399 without D2V part. Same will be verified.

6 Access to necessary data

The in-scope work uses discharge reports from MIMIC-III (Medical Information Mart for Intensive Care) database. As per Physionet website (which maintains MIMIC database): *MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.*

The team has fulfilled the criteria to access this open source data:

- members' identities have been credentiated on PhysioNet
- mandatory training (Data or Specimens Only Research) to access this data, has been completed

So, the team has required access to necessary data.

7 Computational feasibility of reproduction work

Both the input data and model are big. On the whole, MIMIC-III is about 7 GB in size. In line with the original work, the team is considering total of more than 52000 discharge reports, with average number of words per report as 1524, and total number of ICD-9 codes as 6984. More specifically for CNN team will be using documents with 700 words, with each word embedding as 100. Roughly, this will lead to following number of parameters:

- Number of parameters in CNN layer : 76992
- Number of parameters in fully-connected layer: 2241864

Since, this model is going to be big, requiring significant amount of computing and memory, cloud computing environment will be leveraged, to train the model. Most probably GCP (Google Cloud) AI platform will be used to run Gensim and Pytorch installed custom Jupyter notebook as described here.

Prototyping will be done locally on dev machine with CUDA drivers enabled Linux machine with following configuration:

- Intel® Core™ i7-10750H × 12 processor
- 16.0 GiB memory
- NVIDIA GeForce GTX 1650 graphics

8 Coding

All the code will be built from scratch. To the best of team's understanding, original work of authors or any other reproduction work is not publicly available.

References

[1] Yoon Kim. Convolutional neural networks for sentence classification, 2014.

[2] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents, 2014.

[3] Min Li, Zhihui Fei, Min Zeng, Fang-Xiang Wu, Yaohang Li, Yi Pan, and Jianxin Wang. Automated icd-9 coding via a deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(4):1193–1202, 2019.