CS598 DL4H in Spring 2022

Joseph Allen and Sonal Sharma
jaallen2@illinois.edu sonals3@illinois.edu

Group ID: X, **Paper ID:** 215
Presentation link: TODO
Code link: https://github.com/josepha1/DLH_Final_Project

Introduction

Our team is trying to replicate **Contextual Embeddings from Clinical Notes Improves Prediction of Sepsis** by Fatemeh Amrollahi, Supreeth P.Shashikumar, Fereshteh Razmi and Shamim Nemati for our Final Project.

Sepsis, a life-threatening organ dysfunction, is a clinical syndrome triggered by acute infection and affects over 1 million Americans every year. Early detection of sepsis and timely antibiotics administration is known to save lives. In this work the authors have designed a sepsis prediction algorithm based on data from electronic health records (EHR) using a deep learning approach. The findings indicate that incorporation of clinical text features via a pre-trained language representation model(ClinicalBert in this case) can improve early prediction of sepsis and reduce false alarms.

The proposed method replaces the traditional tf-idf features with contextual embedded representations learned using Clinical-BERT, a state-of-the-art model for word and document embedding and specifically trained on a corpus of Biomedical and clinical texts. The resulting document-level representations were then concatenated with physiological, laboratory, and demographic information to make predictions of sepsis using a long-short term memory (LSTM) recurrent neural network.

Scope of reproducibility

Contextual word embedding models such as ELMo and BERT have dramatically improved performance for many natural language processing (NLP) tasks. However, these models have been minimally explored in specialty corpora, such as clinical text; moreover, in the clinical domain, no publicly-available pre-trained BERT models yet exist. ClinicalBert addresses this need by exploring and releasing BERT models for clinical text.
The paper claims that using a domain-specific model with LSTM for prediction yields performance improvements for sepsis prediction.

Addressed claims from the original paper

- ClinicalBERT outperforms performance of traditional BERT model on Clinical predictions
- Combining both structural clinical data with ClinicalBERT embeddings (Model IV) achieved the best AUC performance.
- ClinicalBERT generated more meaningful representations of clinical notes and enabled the model to predict onset of sepsis more accurately

Methodology

We tried to re-implement the approach from the paper description as the author's code wasn't accessible to us. Publicly available MIMIC-III dataset was used to reproduce results similar to the paper.

3.1 Model descriptions

The paper uses ClinicalBERT and LSTM models for sepsis prediction and performance improvements.

ClinicalBert learns deep representations of clinical text using two unsupervised language modeling tasks: masked language modeling and next sentence prediction. A clinical note input to ClinicalBert is represented as a collection of tokens. In ClinicalBert, a token in a clinical note is computed as the sum of the token embedding, a learned segment embedding, and a position embedding. ClinicalBERT improves over BERT on the MIMIC-III corpus of clinical notes .

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points, but also entire sequences of data.

3.2 Data descriptions

Publicly available MIMIC-III (Medical information for intensive care) dataset of critically ill patients, which includes anonymized physiological and clinical data, as well as clinical notes from over 50,000 intensive care unit (ICU) admissions collected between 2001 and 2012 was used as part of the study. Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more.

MIMIC-III is a relational database consisting of 26 tables. Five tables are used to define and track patient stays: ADMISSIONS; PATIENTS; ICU STAYS; SERVICES; and TRANSFERS. Another five tables are dictionaries for cross-referencing codes against their respective definitions: D_CPT; D_ICD_DIAGNOSES; D_ICD_PROCEDURES; D_ITEMS; and D_LABITEMS.

Access to the PhysioNet database was requested and obtained following the steps mentioned in piazza.

3.3 Hyperparameters

TODO

3.4 Implementation

3.5 Computational requirements

4 Results

4.1 Result 1

4.2 Result 2

4.3 Additional results not present in the original paper

5 Discussion

TODO

References

TODO