Peripheral Artery Disease Prediction using Medical Notes

Stanford CS224N Custom Project

Ilies Ghanzouri
Department of Mechanical Engineering
Stanford University
ghanz@stanford.edu

Abstract

In this final project, we aim to fine-tune BERT on clinical notes stored in GCP BigQuery to classify patients. First, we pre-process the data and clean it such that it can be fed into the BERT model. To date, machine learning algorithms have been applied to EHR data such as logistic regression, random forest, in classification of Peripheral Artery Disease (PAD). We will investigate if deep learning produces a more accurate classification of PAD than standard machine learning algorithms.

1 Key Information to include

- Mentor: Akshay Smit
- External Collaborators: Ross Lab at Stanford School of Medicine

2 Introduction

In recent years, breakthrough advancement in deep learning allowed decent performance improvement in solving NLP tasks such as text classification. In particular, models like Google's BERT allowed implementation to produce state of the art results with minimal task-specific fine-tuning through the usage of pretrained models. BERT showcased promising results in disease prediction. BEHRT [1] is a novel deep neural network model for EHR data that showed striking performance in a wide range of downstream tasks with small fine tuning. BEHRT was trained and tested on CPRD - one of the largest linked primary care EHR systems – for predicting the next mostly likely diseases in one's future visits. Results show that it outperformed the best deep EHR models in the literature by more than 8% in predicting for a range of more than 300 diseases.
In this project, we pursue to develop a BERT fine-tuned model on medical notes to predict Peripheral Artery Disease (PAD). Peripheral artery disease (PAD), or atherosclerotic occlusive disease of the lower extremities, affects 8-12 million American adults and more than 200 million worldwide. The prevalence of PAD is as high as 12-30% in patients over the age of 65 years and annual Medicare expenditures related to the treatment of PAD alone total $4 billion. PAD is a highly morbid condition that can lead to limb loss secondary to acute or chronically progressive lower extremity ischemia. Moreover, PAD can lead to a 6-fold increased risk of premature mortality and major adverse cardiovascular and cerebrovascular events (MACCE) .
The diagnosis of PAD can be difficult as approximately 10-30% of patients report typical symptoms of pain with walking that improves with rest. The general public is also relatively unaware of the condition – with only 25% of surveyed adults demonstrating awareness of PAD. The result of a high rate of atypical or absent symptoms, poor awareness, and ambiguous screening guidelines is that 55% of patients with PAD are undiagnosed in the primary care setting, leading to excess morbidity and mortality from lower rates of risk factor mitigation.
To date, researchers have focused on developing traditional risk scores to detect undiagnosed PAD, mainly using well-established risk factors such as smoking, hypertension, hypercholesterolemia,

kidney disease and diabetes. While these risk scores are advantageous in that they are highly interpretable, their ability to discriminate between high and lower risk patients is limited, as reflected by their achieved area under the receiver operating characteristic curves (AUROC) which range from 0.6-0.7 [2]. Furthermore relationships between risk factors may not be linear, whereas risk scores today are based on linear modeling. In contrast data from electronic health records (EHR) enable incorporation of multiple types of data outside of traditional risk factors. EHR data can capture other useful risk factors such as health care utilization, socioeconomics factors and non-obvious clinical risk factors. Furthermore, mathematical algorithms employed by machine learning techniques can improve modeling of risk of PAD across a more diverse set of risk factors.
In previous work in the lab, we have developed machine learning algorithm (logistic regression, random forest) that reached 0.89 AUC score in classification of PAD. In this paper, we aim to investigate if BERT produces a more accurate classification of PAD using clinical notes.

3 Approach

BERT (Bidirectional Encoder Representations from Transformers) pre-trained model can be fine-tuned with one additional output layer to create state-of-the-art models for text classification. We will use Huggingface Pytorch implementation to fine-tune into our dataset.
All data preprocessing and data cleaning has been done by me using SQL and Python on GCP BigQuery. Peripheral artery disease is defined as at least having two mentions of a concept code. Risk factors that are known to be associated with PAD were included in our feature matrix. The following factors were included in our models: cerebrovascular disease, heart failure, coronary artery disease, hypertension, diabetes, hyperlipidemia, body mass index (BMI), smoking status (current, ever or never), age, gender and ethnicity (five levels: Caucasian, Hispanic, Black, Asian and Other). PAD patients were derived via concept code (ICD9/10, CPT, Observation) using conditions, observations, procedure and note.nlp tables from GCP BigQuery. Patients were included if they had at least 2 mentions of codes or term mentions within their health record, and also had no exclusion codes in their health record. These concept codes were processed by deriving all of its ancestors. Terms with ambiguous or vague meanings are removed. Positive examples (cases) for learning the model were defined as patients with PAD while negative examples (controls) for the model were patients without a PAD diagnosis. Patients were excluded if they had < 1 year of data before PAD diagnosis.
As a second step to the project, I have followed the following tutorial (link): BERT Fine-Tuning Tutorial with PyTorch for Text Classification on The Corpus of Linguistic Acceptability (COLA) Dataset. and adapted it to my dataset. PAD patients are labeled as 1 and 0 otherwise.

4 Experiments

4.1 Data

Data are derived from the STAnford Medicine Research Data Repository (STARR). Data include de-identified EHR clinical practice data at Stanford from 1998-2020 featuring over 4 million patients, $>$ 75 million visits, $>$ 65 million notes, $>$ 67 million procedures, $>$ 350 million labs and $>$ 55 million prescriptions. These data have already been converted to the a common data model known as the Observational Medical Outcomes Partnership. The OMOP CDM is a way to represent data in such a way that it uses standardized definitions for different variables and concepts in the EHR. 6583 patients are identified with PAD and 20,000 as non-PAD patients.

4.2 Evaluation method

Model performance will be evaluated on discrimination (AUROC) in order to compare it with previous AUCs results obtained with machine learning algorithm (random forest).

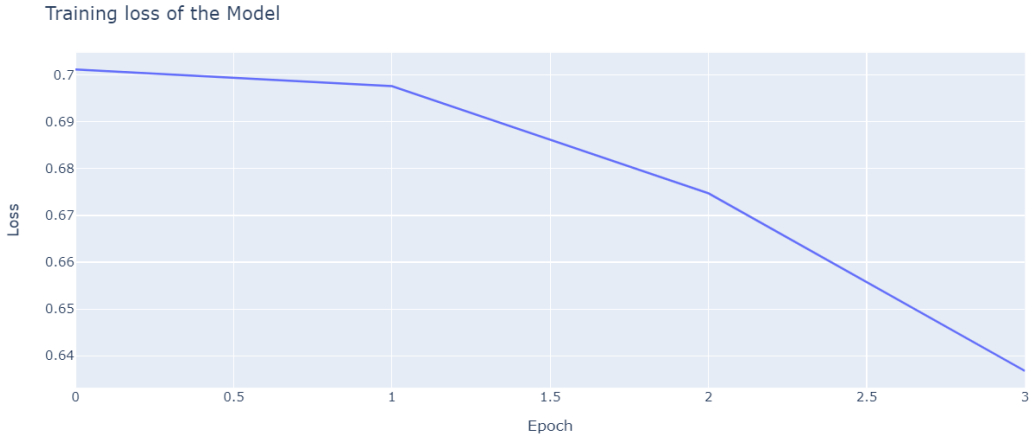
4.3 Experimental details

BERT (bert-base-uncased) has been fine-tuned to our own dataset. 4 epochs of training are used for fine-tuning BERT. Due to current hardware issues on Nero, we were only able to train our model on CPU. The MAX_LEN tokenization is set to 512 tokens (maximum sentence length for BERT). We

use BertTokenizer from transformers library of Hugging Face in order to tokenize our medical notes. The batch size is varied between 4 and 16 (BERT authors recommend a batch size of 16 or 32). The learning rate is set to $2e-5$ and eps $= 1e-8$ (Adam optimization parameter). We used 6500 PAD Cases and 6500 PAD Controls (13,000 in total). 90% are used for training and 10% for the dev set.

4.4 Results

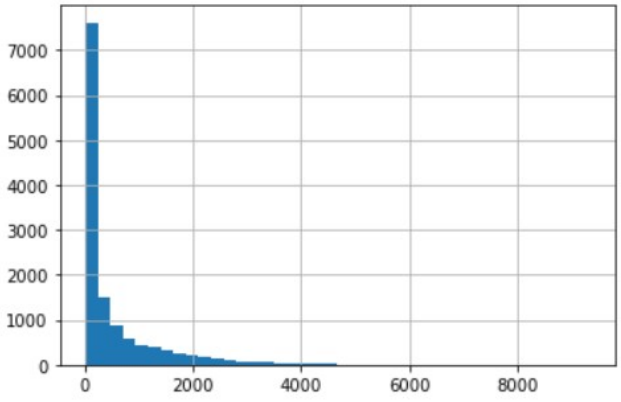
Figure 1: Training loss of BERT-base-uncased



1000 medical notes are used on the test set and the test accuracy is 0.622. We notice an improvement of the accuracy from 0.5315 (project milestone) due to the fact that we increase MAX_LEN (from 64 to 512) and the number of training examples (from $1,000$ to $13,000$).
Moreover, the obtained AUC on the current fine-tuned BERT model is 0.6643. This AUC score is similar to traditional risk scores models ($0.6-0.7$) and way lower than machine learning algorithm we developed before (0.89 AUC using random forest). *Due to Nero server issues, I was not fortunate to fine-tune on a biomedically pretrained model (BioBERT) on time and present results before the deadline. The BioBERT fine-tuning is still currently running (training takes around* 16 *hours).*

5 Analysis

Figure 2: Length distribution of medical notes



The current AUC score must be improved. Is it similiar to traditional risk scores models [2] ($0.6-0.7$). The maximum sentence length that BERT can handle is 512 tokens. Currently, notes have been cut to use only the first 512 tokens which might have cut off valuable medical note information and impeded BERT to efficiently learn. As we can see on Figure 2, most medical notes have a length less than

256. However, one way to improve our model would be to consider 1024 tokens in order to have a more robust and general model. *Moreover, as mentioned above, we must fine-tune on a biomedically pretrained model (BioBERT) to verify if it leads to some AUC improvement. The BioBERT fine-tuning is still currently running (training takes around* 16 *hours).*

6 Conclusion

In this paper, we have implemented a fine-tuned BERT model on our dataset to predict PAD. Currently, the standard BERT model was unable to achieve a higher accurate classification (AUROC) of PAD using clinical notes. In future work, we must try to use a linear-complexity transformers to process longer sequences (Big Bird [3], Longformer [4] or Linformer [5]) . Another step is to clean the data and remove unrelevant phrases portion from medical notes by using Medical Named-entity recognition on Spacy (by keeping disease entities).

References

[1] Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records, 2019.

[2] Sue Duval, Joseph M Massaro, Michael R Jaff, William E Boden, Mark J Alberts, Robert M Califf, Kim A Eagle, Sr Ralph B D'Agostino, Alison Pedley, Gregg C Fonarow, Joanne M Murabito, P Gabriel Steg, Deepak L Bhatt, Alan T Hirsch, and on behalf of the REACH Registry Investigators. An evidence-based score to detect prevalent peripheral artery disease (pad). *Vascular Medicine*, 17(5):342–351, 2012. PMID: 22711750.

[3] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2021.

[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[5] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.