Project Proposal based on BEHRT [4]

Birhanu Digafe, March 2023

1 Introduction

Deep learning, a subset of machine leaning, is now widely used in many healthcare applications, such as prediction of onset of heath failure [1]. Despite the availability of large amounts of electronic health records (EHR), their use is still very limited for predicting critical illnesses, especially for those conditions, where early intervention can provide dramatic positive results. As a result, currently most diagnoses happen after the patient shows visible signs of illnesses during emergency hospital visit, annual medical check, or unscheduled doctor's visit.

Earlier work [7] was based on heavy feature engineering, where patients are represented by a feature vector that include experts selected variables, such as patient demographics, lifestyle factors, laboratory tests, prescribed medications, selected morbidities, and emergency admissions. This study [7] shown improved prediction of risks of emergency admission compared to RF (Rain Forest) and GBC (Gradient Boosting Classifier).

Deep learning models, on the other hand, rely on raw patient data with minimal pre-processing or expertly selected feature variables, hence reducing the labor, time, and cost of feature engineeing tasks. For healthcare predictions, the hierarchical and the temporal dependency of the data is critical. In this project, we use BEHRT [**behrt**], based on BERT [2] (a transformer [9] based model for healthcare).

The root data element is the patient with its changing demographics at each doctor's visit. Each visit adds zero or more medical codes, such as, diagnosis codes, lab tests, prescriptions, and medical procedures to the patient data. Medical codes are normally represented by their respective standard code. For example, diagnosis codes are represented by the International Classification of Diseases (ICD-10). In EHR data, doctors visits don't follow a regular time pattern, so a timestamp is attached to each event to identify the order of visits and length of time between visits. Deep learning models use a series of computational layers to identify the important data points from this complex and sequential input based on their impact on their output prediction.

The Tansformer architecture [9] is effective for processing sequential data using parallel computations. As such, we can leverage all the compute resources that are available at our disposal, hence reducing the training time. BERT [2] (Bidirectional Encoder Representations from Transformers) is a natural language processing (NLP) model based on the Transformer architecture. BEHRT has significant similarities with the BERT model, but it applied to healthcare.

The BERTH study was conducted on 1.6 million patients from the Clinical Practice Research Datalink (CPRD) [3] database in UK that have more than five visits and have linkage to HES (a database maintained by National Health Service). CPRD is an anonymised medical records database with more than 11.3 million patients from 674 medical practices in the UK. The data also has a broad representation of the population by age, gender, and ethnicity. The study aims to simultaneously predict the likelihood of 301 conditions in patient's future visits using the patient's EHR data.

We found this study very interesting for three reasons. (1) EHR is widely adopted and quality data could be potentially available for serious studies if privacy concerns are addresses. (2) Prediction can save lives and reduce healthcare cost. (3) It uses a more modern approach and techniques learned in this project can be applied to wide range of similar or unrelated projects.

2 Scope of reproducibility

2.1 Data source

To download the CRPD data, researchers are required to submit a request form (https://cprd.com/data-access) and access is subject to approval from the Independent Scientific Advisory Committee (ISAC). In short, the data is not readily available as a public data set. For this project, we aim to use synthetic EHR data from [5], which can be downloaded after registration. This limits our ability to reproduce the results of the study. However, because the synthetic data has similar data structure, we believe we can still leverage the study and make predictions on a different set of data.

3 Methodology

3.1 Data descriptions

Instead of the original data source from CPRD, we will use the outpatient synthetic data from medysin.ai. The data consists multiple CSV data files following the Observational Medical Outcomes Partnership (OMOP) and Common Data Model (CDM).

Patient data will be obtained from the person file. Each row defines a patient, which has a unique identifier (patient id) and demographics data (date of birth, gender, ethnic, etc.). Many data elements are represented by a "concept id", which follow a standard codification system from their respective data element. Hospital visits are in the "visit occurrence" file, where each row defines a patient's visit. A visit record has a unique visit id (visit occurrence id) and links the patient (person id), and many more attributes, such as date, location, time, etc. The medical codes for each visit are defined on separate files. For example, diagnoses data is in the condition occurrence file, medical procedures are in the procedure occurrence file. For this project, we only use the patient details, visit details, and conditions (diagnosis codes).

3.2 Model descriptions

The model uses four embeddings: disease, positional encoding, age, and visit segment. The last two are unique to BEHRT. Age is included because many diseases are related or influenced by the patient's age. Visit segments are either A or B and are used to differentiate two consecutive visits for a patient as they are assigned alternatively to each visit.

To apply language models to electronic medical records, diagnosis codes are depicted as words, visits as sentences, and the patient's medical history as a document. As shown in Figure 2, two additional codes are added: "CLS" (the start of the patient's medical code, similar to a new paragraph) and "SEP" (separates diagnosis codes between visits, similar to a comma). This allows the new algorithm to use BERT [2] with multi-head attention, positional encoding, and MLM (Masked Language Model). Each visit's sequence number if encoded and used as positional encoding.

The transformed input, the summation of all the embeddings, is feed to the BERTH transformer-based architecture. First, the input is pre-trained using the MLM task where 86.5% of the disease codes were unchanged, 12% replaced with a mask, and the remaining 1.5% were changed with a random noisy codes. The model learns the network parameters including the embedding representation of the disease codes and the masked disease codes. The precision score [6]of the pre-training MLM task is calculated over all patients and labels.

The disease prediction task is a downstream tasks and uses the same weights pre-trained by the MLM. The inputs are passed via a pooling layer followed by a sigmoid function to produce a multi-label prediction for each disease code.

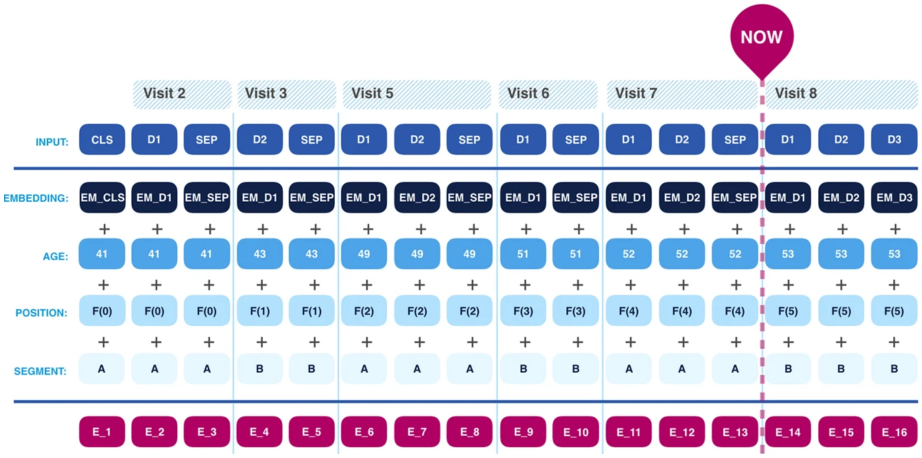
NOW
Visit 2
Visit 3
Visit 5
Visit 6
Visit 7
Visit 8
INPUT:
CLS D1 SEP D2 SEP D1 D2 SEP D1 SEP D1 D2 SEP D1 D2 D3
EMBEDDING:
EM_CLS EM_D1 EM_SEP EM_D1 EM_SEP EM_D1 EM_D2 EM_SEP EM_D1 EM_SEP EM_D1 EM_D2 EM_SEP EM_D1 EM_D2 EM_D3
AGE:
41 41 41 43 43 49 49 49 51 51 52 52 52 53 53 53
POSITION:
F(0) F(0) F(0) F(1) F(1) F(2) F(2) F(2) F(3) F(3) F(4) F(4) F(4) F(5) F(5) F(5)
SEGMENT:
A A A B B A A A B B A A A B B B
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_10 E_11 E_12 E_13 E_14 E_15 E_16

Figure 1: Source:https://www.nature.com/articles/s41598-020-62922-y

3.3 Hypothesis

The study makes disease predictions for three scenarios: next visit (T1), disease prediction in the next six months (T2), and disease prediction in the next 12 months (T3). For T1, only patients that have three or more visits are included. Additionally, a random index ”j” (3 ¡ j ¡ Np, where Np is the patient’s max number of visits) is selected and the model will try to predict the disease codes for the visit j+1. For T2 and T3, patients must have 6 and 12 months worth of EHR data. The j index will be selected randomly from (3, M), where M is the highest index after which there are 6 or 12 months visits for the patient. It should be noted, that this will reduce the amount of data that can be used for training.

3.4 Implementation

The system was implemented using python. The optimizer used for both MLM and disease prediction is BertAdam [2].

The BEHRT source code is publicly available at (https://github.com/deepmedicine/BEHRT). We will recreate this code to adapt to the new data model and experiment new ideas.

3.5 Computational requirements

The original study used NVIDIA Titan Xp Graphical Processing Units (GPU) for pre-training, training, and testing, which is expensive and powerful. The synthetic data only has 241,095 patients and the visit data close to 3M records. We hope we can leverage a less powerful computer to perform similar experiments on the synthetic data.

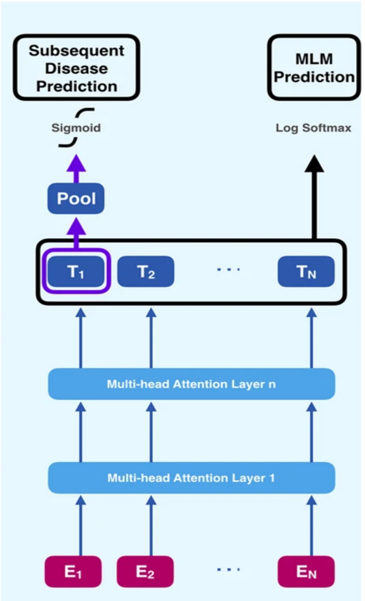
Subsequent Disease Prediction
MLM Prediction
Sigmoid
Log Softmax
Pool
T_1
T_2
. . .
T_N
Multi-head Attention Layer n
Multi-head Attention Layer 1
E_1
E_2
. . .
E_N

Figure 2: Source: https://www.nature.com/articles/s41598-020-62922-y

4 Results

The APS and AURO scores for each the three tasks are compared with Deepr and RETAIN algorithm.

Model name	Next Visit	Next 6 Months	Next 12 months —
	APS ||*AOROC*	APS ||*AOROC*	APS ||*AOROC*|
BEHRT	0.462 ||0.954	0.525 ||0.958	0.506 ||0.955|
Deepr	0.360 ||0.942	0.393 ||0.943	0.393 ||0.943|
RETAIN	0.382 ||0.921	0.417 ||0.927	0.413 ||0.928|

4.1 Analysis

The result shows BEHRT achieved significant improvements over the other methods in all the predictive tasks. We anticipate that due low volume and quality of EHR data we will not match the results in the study. Nevertheless, the experience will be highly educational.

5 Plans

We plan to download the data, format the data in the appropriate data format, review and modify the code, train and evaluate, and present our findings in the final report.

The original study used Bayesian Optimization [8] to find optimal hyper parameters (number of layers, hidden state size, etc.). Due to the limited time and resource we will try to reuse similar parameters as the ones found in the study.

As an ablation to the study, we will test removing the age to measure the impact of incorporating the age in the prediction results.

6 Citations

References

[1] Edward Choi et al. "Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction". In: (2017). arXiv: `1602.03686 [cs.LG]`.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`. URL: `https://aclanthology.org/N19-1423`.

[3] Emily Herrett et al. "Data Resource Profile: Clinical Practice Research Datalink (CPRD)". In: *International Journal of Epidemiology* 44.3 (June 2015), pp. 827–836. ISSN: 0300-5771. DOI: `10.1093/ije/dyv098`. eprint: `https://academic.oup.com/ije/article-pdf/44/3/827/14153119/dyv098.pdf`. URL: `https://doi.org/10.1093/ije/dyv098`.

[4] Yikuan Li et al. "BEHRT: Transformer for Electronic Health Records". In: *Scientific Reports* 10.1 (Apr. 2020). DOI: `https://doi.org/10.1038/s41598-020-62922-y`.

[5] medysin.ai. *https://app.medisyn.ai/download*. `"https://app.medisyn.ai/download"`. [Online; Download synthetic EHR data].

[6] David MW Powers. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061* (2020).

[7] Fatemeh Rahimian et al. "Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records". In: *PLoS medicine* 15.11 (2018), e1002695.

[8] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical bayesian optimization of machine learning algorithms". In: *Advances in neural information processing systems* 25 (2012).

[9] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).