EHR prediction based BEHRT - Project Draft Report [5]

Birhanu Digafe, April 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 [8] 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, and focused on selected morbidity and emergency admissions. This study [8] shown improved prediction of risks of emergency admission compared to RF (Random 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 engineering tasks. For healthcare predictions, the hierarchical and the temporal dependency of the data is critical. For example, medical codes generally follow a hierarchical classification system to capture the general and specific nature of diagnosis codes. Also, the onset of a medical condition may cause other conditions after a period of time. In this project, we attempt to reproduce BEHRT (BERT for Electronic Health Records) [5], which is based on BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding) [3], which in-turn is based on the Transformer architecture [9].

The core input data is the patient with its changing demographics at each doctor's visit. Each visit adds zero or more medical codes, such as, disease diagnosis, 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, such as ICD-10. In EHR data, doctor visits don't follow a regular time pattern, so a timestamp is attached to each event to identify the order of events and length of time between events. Deep learning models use a series of computational layers to identify the important data points from this complex and sequential data to gain understanding of diseases, and predict future conditions based on medical history.

The Transformers 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 to reduce the training time or analyze large data. BERT (Bidirectional Encoder Representations from Transformers) [3] is a natural language processing (NLP) system and is designed to pre-train deep bidirectional representations from unlabelled text. BEHRT uses the same methods as BERT to pre-train medical conditions.

The BEHRT study was conducted on 1.6 million patients from the Clinical Practice Research Datalink (CPRD) [4] database in UK that have more than five visits and have linkage to HES (a database maintained by National Health Service). CPRD is an anonymized 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.

In this project, we attempt to reproduce BEHRT using synthetic data due restrictions of accessing the original data source. We use the same techniques as BERT (and BEHRT)

to pre-train medical diagnosis codes applying MLM (Masked Language Model) and next visit prediction as a downstream task.

2 Scope of reproducibility

2.1 Data source

To download the original source 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 used synthetic EHR data from [6], which can be downloaded after registration. The data follows OMOP format.

2.2 Model Parameters

Generally, we used the BEHRT model parameters to pre-train BERT. These parameters may need to be fine tuned to avoid over-fitting. More research is required to find the optimal hyper-parameters, e.g. using Bayes Optimization.

2.3 Source code

The source code of BEHRT is published on github [2], but there are some references to private codes for pre-processing the data. We made best effort to pre-process the data following the paper and the general input forms required by BERT. One potential diffrence from the is the representation of age. The code implies using the age (in years), but the code that calculate age is not available. For many patients, the same age may be present multiple times due to repeated visits. To gain better temporal representation, we used the age in months as the full date of birth and visit dates are available in the data. The results, however, were not significantly different.

Another issue is the ordering of condition codes within a visit. As BERT is a language model ordering should be important. BEHRT does not explicitly mention this. We sorted the condition codes within a specific visit.

In summary, the limited quality volume of data and incomplete information limited our ability to reproduce the exact results of the research. Nevertheless, we achieved an acceptable performance as shown in the results section.

3 Methodology

3.1 Data descriptions

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

Patient data is obtained from the person file (person.csv), which contains a unique identifier (patient id) and demographics data (date of birth, gender, ethnicity, etc.). Hospital visits are stored in the $visit_occurrence.csvfile, whichcontainsauniquevisitid, visitdate, pers$

In this project, we only used patient details (pid, dob), visit details (pid, visit date), and condition occurrence (visit date, diagnosis code). We created a pre-processing code to generate pickle files for BERT modeling. Key points are: (1) We calculated the frequency of occurrence of each condition across all patients and visits, and removed those infrequent conditions (3000 times of less). This brought the number of conditions to 296, which is close BEHRT (301). Therefore, our conditions vocabulary is small. (2) We removed patients with less than 5 visits, so we can have some medical history for next visit prediction. (2) We used 256 as the maximum sequence length. BERT uses fixed sequence length. The max length in our data set is 447. Therefore, we removed data for some patients and padded with a special character for those that have less data.

3.2 Model descriptions

The system uses the BERT encoder architecture to generate a vector representation for each condition code based on its occurrence in patients' medical history. This involves a multi-head self-attention layer, layer normalization, and a dense neural network layer with a non-linear activation function. Our model parameters are similar to BEHRT (12 attention heads, 6 hidden layers), with about 4.5 million parameters in total.

We fine-tuned the system using two tasks: Masked Language Model (MLM) and Next Visit Prediction. For MLM, we mask 15% of known condition codes and force the model to predict them. We evaluated the prediction accuracy. The resulting vector representation has 288 dimensions, and we used PCA for a 2D representation and K-means for spatial grouping of condition codes.

3.3 Data Representation

Electronic medical records are processed using language models, with diagnosis codes represented as words, visits as sentences, and a patient's medical history as a document. To facilitate this, two additional codes ("CLS" and "SEP") are added to the sequence to represent the start of a patient's medical code and to separate diagnosis codes between visits. The algorithm then uses BERT with multi-head attention, positional encoding, and MLM.

A person's medical history is represented as a sequence of visits, with each visit having a unique sequence number used for positional encoding. The age sequence, unique to BEHRT, has the same length as condition tokens and stores the person's age at the time of the condition. A novel idea introduced by Transformers is the use of positional encoding to capture sequence information while avoiding sequential learning. Each condition code is assigned an order number, with conditions of a single visit having the same number. A segment encoding is added to differentiate conditions between two subsequent visits.

For example, for a patient with three visits [[D1, D2], [D1, D3], [D1]] Input tokens = ['CLS', 'D1', 'D2', 'SEP', 'D1', 'D3','D4', 'SEP', 'D1, 'SEP'] Position tokens = [0, 0, 0, 0, 1, 1, 1, 1, 2, 2] Segment token = [0, 0, 0, 0, 1, 1, 1, 1, 0, 0]

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

3.4 Implementation

The transformed input, the summation of all the embeddings, is feed to the BERT transformer-based architecture. First, the input is pre-trained using the MLM task where 86.5% of the condition 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 [7]of the pre-training MLM task is calculated over all patients and labels.

The disease prediction task is a downstream task that uses the same weights as the pre-training model created by the MLM. The input data is split into 80% for training and 20% for validation. For each patient, a random visit is selected, and the selected visit along with prior visits are used as input. The next visit's multi-hot vector representation of its condition codes is used as the label that the task will attempt to predict.

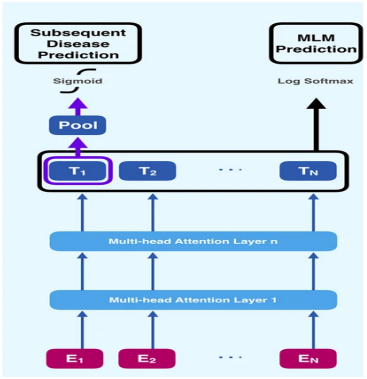

Figure 2: Source

The system was implemented using python. The optimizer used for both MLM and disease prediction is BertAdam [3].

The BEHRT source code is publicly available at (https://github.com/deepmedicine/BEHRT). We created three Jupyter notebooks: pre-processing input csv files, BERT pre-training, and next visit prediction task. The synthetic data, notebooks, and supporting module files will be available on github as part of the final project submission.

3.5 Computational requirements

We used Google Colab with Python 3 Google Compute Engine back-end (GPU) which as 12.7GB system RAM and 15GB GPU RAM. We were able to both tasks with a batch size of 64. Larger batch sizes caused out of memory issues.

4 Results

The APS and AUROC scores of the model, BEHRT, Deepr, and RETAN are presented below.

Model Name	Next Visit		Next Visit-6months	
	APS	AUROC	APS	AUROC
BEHRT	0.462	0.954	0.525	0.958
Deepr	0.360	0.942	0.393	0.943
RETAIN	0.382	0.921	0.413	0.928
Our	0.46	0.71		

4.1 Analysis

The accuracy of the MLM task is almost the same as BEHRT. The AUROC score (0.71), however, is low compared to other methods. We also plotted the conditions codes in a 2D map. The purpose is to see whether similar conditions are shown closer in the map. We only conducted visual inspection to see the overall distribution of the condition codes and examining the name of the conditions that are close to each other in the map. For example, diabetes mellitus type I and II are very close. For through analysis, cosine-similarity score can be computed and review the top 10 similar conditions.

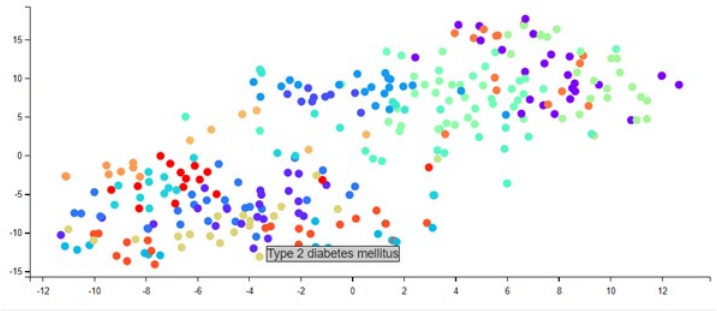

Figure 3: Disease mapping

The prediction of next visit task also performed well with AUROC of 0.71, but it lags behind other models. We suspect over-fitting and insufficient data could be the main reasons.

4.2 Ablation study

We removed the age repeated the same experiments. The results were not that significantly different. We will conduct and report the findings in the final report.

We also experimented with different model parameters, such as number of attention heads, hidden size, and sequence length. BERT tries to encode 30,000 words and our vocabulary is much smaller and should work with lesser number of parameters. We will report

5 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] deepmedicine. *https://github.com/deepmedicine/BEHRT*. "https://github.com/deepmedicine/BEHRT". [BEHRT code].

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

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