Extensible Pre-training and Language Modeling for Electronic Health Records

Hanxin Zhang and Andrey Rzhetsky

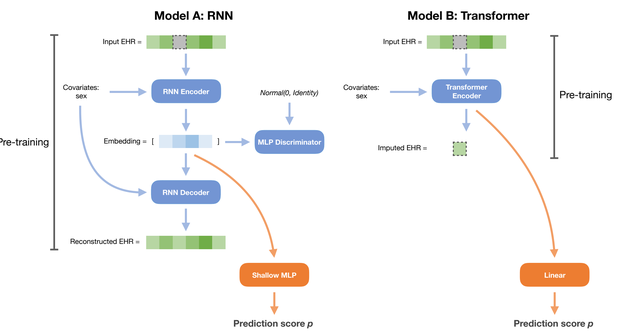
Department of Medicine, the University of Chicago

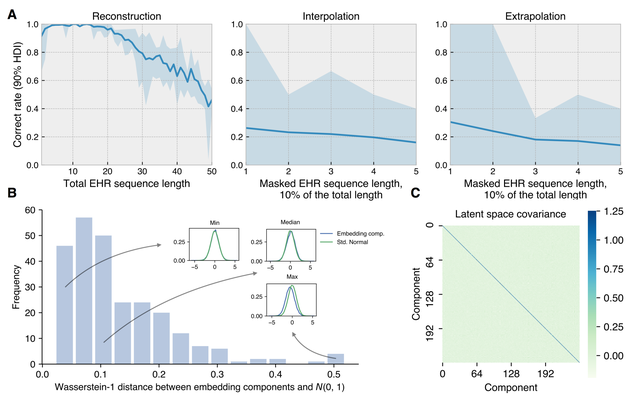
Abstract

We introduce a pre-trained language model designed specifically for sequence data of electronic health records. Unlike previous successful implementations, e.g., BERT, the model adopts an encoder-decoder architecture with latent embedding space regularized to be standard gaussian by a discriminator. The encoder-decoder structure, along with the well-shaped embedding space, confer multiple functions on the model including health records imputation, prediction, patient similarity estimation, approximate matching, rarity detection, language modeling, and sample generation. It is also easy to mount and fine tune the model modules to other targets and achieve potential downstream tasks. We show the great importance of regularized training of such sequence modeling problems. The present work extends the deep natural language modeling techniques to the more general realm of sequence understanding and learning and promises versatility in thorough research of biomedical data.

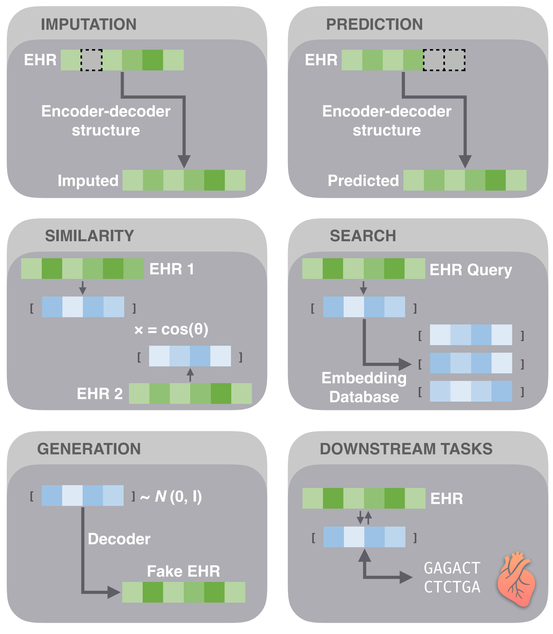
Methods

The model adopts an encoder-decoder architecture that embeds EHR sequences into a latent space regularized by an adversarial training procedure. The discriminator constrains the latent space to standard gaussian (consisting of random vectors with every element following an independent standard gaussian distribution) by matching embeddings to randomly generated standard gaussian random vectors. Every EHR entry documents the health history of a patient; and if the embedding achieved, it could represent the hidden health state of them.





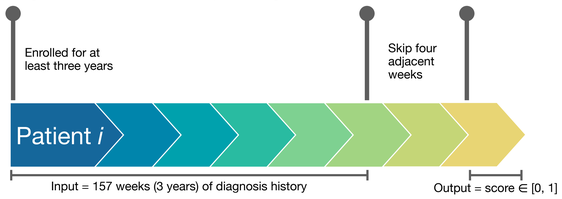
Applications

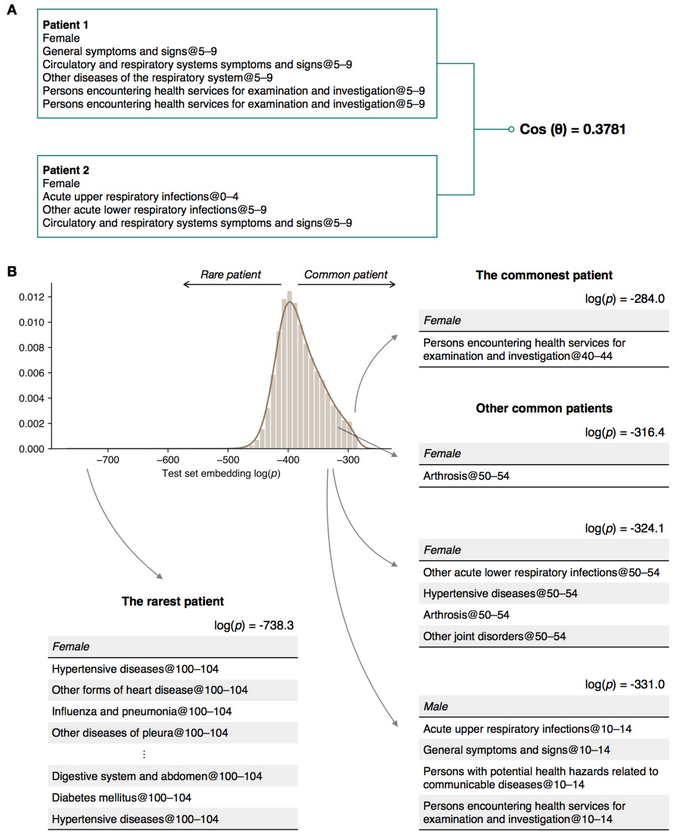


We could then employ such representation of health to multiple downstream tasks. Masked and unobserved entries of an EHR sequence could possibly be imputed and predicted by the reconstruction process. Besides, the vector representation could use to compute the similarity between EHR sequences and patients. A search system is also practicable. For a newly admitted patient, we could match its health embedding to existing samples in our database, and see their diagnosis, prescription, operation, prognosis, and other medical information for reference. More interestingly, since the latent embedding space is regularized and thus standard gaussian, we are able to forge samples y feeding standard gaussian vectors to the trained decoder.

A Downstream Task Example: Predictive Model for Heart Attack and Stroke

Cardiovascular diseases, followed by cancers are the most common causes of death in the world. Specifically of the circulatory system diseases, ischemic heart diseases (heart attacks) and stroke lead the mortality. Cardiovascular diseases develop progressively with deteriorating symptoms and signs, so we assume heart attack and stroke are predictable given prior medical history of patients. The project intends to build a predictor that signals the risk of first-time heart attack OR stroke 4 weeks before the onset for ALL patients. We use The IBM MarketScan DX (diagnosis) data set, incorporating 150+ million unique patients.





Pretraining and Fine-tuning

Records in the MarketScan data set are coded in ICD-9. We mapped them to actual conditions through a hierarchical mapping covering all ICD codes. Circulatory system diseases are finely mapped, while others are coarsely projected to large-group diseases.

Pre-training

- 10% of the words in each sequence are masked.
- The RNN model was trained for 5 epochs (2M iterations) on 100M unique sequences. Batch size = 256.
- The Transformer model was trained for 5 epochs (1.3M iterations) on 64M unique sequences. Batch size = 256.

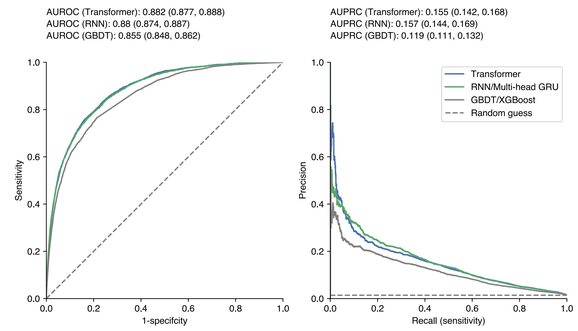
Fine-tuning

- The RNN model: 15 epochs (2M iterations) on 38M unique sequences. Batch size = 256.
- The Transformer model: 15 epochs (2M iterations) on 38M unique sequences. Batch size = 256.

For your reference: Google BERT was pre-trained for 1M iterations, and tuned for another 1M iterations.

AT THE FOREFRONT UChicago Medicine

Results and evaluation

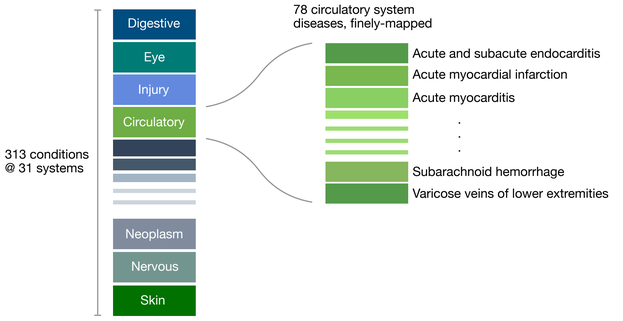


Models were evaluated on a large validation set consisting of 0.18M samples. A test set of the same size was reserved but not used yet.

- The RNN model is on a par with the Transformer model though it is 14 times smaller: 0.78M vs. 11M parameters.
- The deep learning approaches outperform the ensemble tree model (GBDT).

For your reference: Google BERT-large has 340M parameters.

Age is the most predictive factor for the all-inclusive test set. If we pair ever positive sample with a negative counterpart in the same age and sex, the statistics become around 70\% for accuracy, AUROC, and AUPRC.



Acknowledgements

IGSB Institute for Genomics & Systems Biology

NIH National Institutes of Health

DARPA