Towards applying a time-aware Bert model to structured EHR records to generate contextualized concept embeddings

Chao Pang, PhD[1], Krishna S. Kalluri, MS[1], Xinzhuo Jiang, MS[1], Karthik Natarajan, PhD[1]

[1]Columbia University, New York, NY

Abstract

The aim of this study was to combine the time-aware concept embedding model with the state of the art language model BERT to generate both time-aware and context aware concept embeddings. The preliminary results show promise that the two attention mechanisms can be incorporated.

Research Category (please highlight or circle which category best describes your research)

Methodological research

Background

Embedding algorithms, initially developed by the Natural Language Processing (NLP) community for obtaining low dimensional vector representations of words, have been largely adopted in medical informatics to learn concept embeddings from the structured EHRs (1–3). Despite the massive success of these embedding algorithms, almost all do not properly address how to handle time within EHRs, e.g. two adjacent concepts in an EHR sequence could occur far away from each other on the timeline of the patient's medical history. It wasn't until 2018 that Cai *et al* developed the time-aware concept embedding algorithm (4) to address this issue, in which they added a time-attention layer on top of the word2vec continuous bag of word model (CBOW), allowing the algorithm to observe where co-occurring concepts often appear on the timeline for target concepts and to pay more attention to specific time buckets. In the same year, Google published its revolutionary work on the new deep learning language model called Bidirectional Encoder Representations from Transformers (BERT) that incorporates several layers of transformer encoders to generate contextualized word embeddings via a self-attention mechanism (5). The powerful idea behind BERT is that it can aggregate meaning from a sentence and modify the vector representation of each word on the fly based on the context words around it. So far, there have been several attempts to apply BERT to EHRs, all of which seemed to improve the performance in comparison to the classic embedding algorithms (6,7). Inspired by the previous works, we assessed the feasibility of combining the time attention model with BERT to create a new embedding algorithm that is both time-aware and context-aware in this study.

Methods

We used a deep learning library called keras-transformer that provides the basic layers (encoder layer and self-attention layer) for building the BERT model (8). We implemented a time attention layer based on the architecture described in Cai *et al* (4). In addition, the transformer encoder was modified to combine time-attention and self-attention weights. The model was implemented in Tensorflow 2.2.0 and the code is publicly available at https://github.com/ChaoPang/keras-transformer. The model was trained on EHR data from Columbia University Irving Medical Center's Observational Medical Outcomes Partnership (OMOP) common data model (CDM) (9). To speed up training, we only included two years of data and limited the data to three domain tables (conditions, drugs, and procedures). For validation, the time attention weight distributions were extracted from the model and visualized in a Jupyter notebook for the top five conditions found in the training data. Bertviz was used to visualize the self-attention weights extracted for a selected training example related to COVID-19 (10). In addition, we extracted the "raw concept embeddings" from the first layer of the model (as opposed to the output, which is the contextualized embeddings) and applied PCA to extract 3D features for visualization in Tensorflow Projector (11).

Results

The total number of patients in the train set was 589,150 with their medical histories ranging from 2018

to 2020. The number of concepts per person was 70 at the 95% percentile. Based on this, we selected a context window of 100 concepts, and a time window of 101 weeks (50 preceding weeks/50 following weeks plus the index week). We padded an “unused token” to the concept sequence and its corresponding timestamp sequence for lengths that were less than 100. The model was trained for 20 epochs with a batch size of 256 using a distributed strategy on two GeForce RTX 2080 Ti GPUs.

Figure.1 shows time attention distributions for the most frequent conditions, where x-axis represents time intervals in weeks and y-axis represents the corresponding attention weights. The model learned to pay more attention to the index week for acute conditions such as *disease due to coronaviridae*, whereas it pays the uniform attention throughout time for chronic conditions such as Type 2 Diabetes. In **Figure.2**, the self-attention mechanism is demonstrated for a given sequence of concepts ordered chronologically from top to bottom. The sequence is mirrored between the left and right (the left text is cut off by Bertviz) and the connection between them indicates the attention weights, the bolder the connection the more attention weight between the two. What’s currently shown is the attention that the condition concept, *Disease due to Crorонaviridae*, pays to the surrounding context. The bolder connection to the condition, *Acute lower respiratory tract infection*, suggests that there is a strong correlation between the two. In **Figure 3,** the nearest neighbors are highlighted for the seed concept, *Disease due to Crorонaviridae*, based on the cosine similarities computed from their embeddings, among which the top 5 closest concepts are *Acute lower respiratory tract infection, Disease caused by SARS-COV2, Disorder of respiratory system, Ceftriaxone, and Viral pneumonia*.

Figure 1. Visualization of time attention distributions

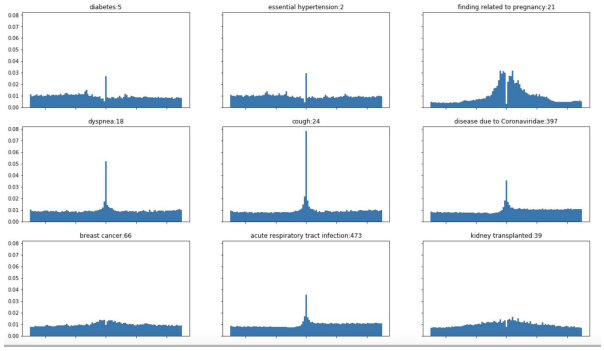


Figure 2.self-attention mechanism

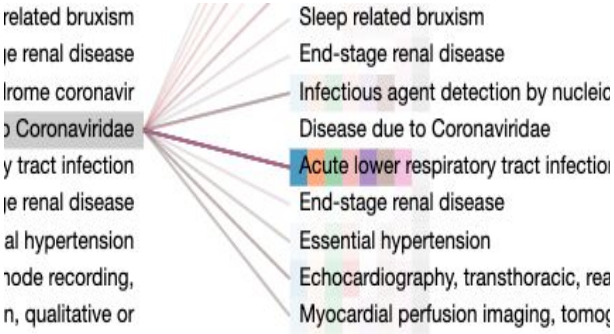
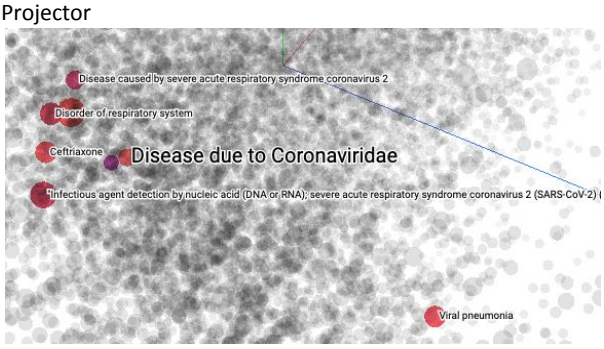


Figure 3. Visualization of embeddings in Tensorflow Projector



Conclusion

To the best of our knowledge, this was the first attempt ever to combine time attention with self-attention mechanisms to develop a model that is both time-aware and context-aware. The preliminary experiments showed encouraging results, however, additional validation is needed to assess the algorithms performance and utility to improve phenotyping efforts.

References

1. Xiang Y, Xu J, Si Y, Li Z, Rasmy L, Zhou Y, et al. Time-sensitive clinical concept embeddings learned from large electronic health records. BMC Med Inform Decis Mak. 2019;

2. Glicksberg BS, Miotto R, Johnson KW, Shameer K, Li L, Chen R, et al. Automated disease cohort selection using word embeddings from Electronic Health Records. In: Pacific Symposium on Biocomputing. 2018.

3. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer NP, et al. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. Pac Symp Biocomput [Internet]. 2018 Apr 4 [cited 2020 Mar 16];25:295–306. Available from: http://arxiv.org/abs/1804.01486

4. Cai X, Gao J, Ngiam KY, Ooi BC, Zhang Y, Yuan X. Medical concept embedding with time-aware attention. In: IJCAI International Joint Conference on Artificial Intelligence. 2018.

5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. 2019.

6. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep [Internet]. 2020;10(1):7155. Available from: https://doi.org/10.1038/s41598-020-62922-y

7. Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. In: IJCAI International Joint Conference on Artificial Intelligence. 2019.

8. Mavreshko K. Keras-Transformer [Internet]. [cited 2020 Jun 1]. Available from: https://github.com/kpot/keras-transformer

9. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. In: Studies in Health Technology and Informatics. IOS Press; 2015. p. 574–8.

10. Vig J. A multiscale visualization of attention in the transformer model. In: ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations. 2019.

11. Tensorflow. Tensorflow Projector [Internet]. [cited 2020 Jun 26]. Available from: https://projector.tensorflow.org/