A Review on Med-BERT: Pretrained Contextualized Embeddings for Disease Prediction

Introduction

In clinical settings, the ability to predict a patient's future disease progression is crucial in informing physician decision making and providing timely care delivery to prevent adverse events and over healthcare utilization. Oftentimes, administrative healthcare claims data is the predominance source for predictive modeling given its standardized schema and coding requirements. However, there is a considerable amount of information lost when patient journeys are converted to structured data. In contrast, electronic health records (EHR) data presents itself as a richer source in facilitating clinical discovery. It is real-time, patient centered, and more accurately captures the diagnostic and treatment histories with finer granularity including elements such as medications, immunization records, imaging, lab testing results, etc.

This tech review covers an analytical capability named Med-BERT. This tool leverages EHR data coupled with NLP techniques to deliver high performance in clinical risk predictions. Specifically, the Med-BERT framework utilizes bidirectional encoder representations from transformers (BERT) to create contextualized embeddings pretrained on structured EHR data to boost the accuracy of deep learning-based predictive models.

Model Paradigm, Training Data, and Evaluation

On a high-level, Med-BERT follows the pretraining fine-tuning paradigm. The pretrained model generates contextualized clinical embeddings without any prediction labels. These token embeddings are used in downstream prediction tasks and deep learning classification layers. Generally, deep learning models greatly enhances diagnostic performance in disease progression. However, a requirement to adopting such deep learning methods lies in the large and high-quality datasets with annotated labels, ones that would enable the model to uncover the complex semantics in the clinical domain and not underfit in the training stage. EHR data, though rich in capturing various subtle medical information, poses challenges when being used directly in a deep learning framework due to data accessibility, governance, and skewed distribution in rare diseases, etc. To address these challenges, Med-BERT utilizes transfer learning techniques[i] to pretrain first using large amount of unannotated EHR data. The pretraining will capture the intrinsic structures, heuristics, and semantics embedded in the data. These elements can then be plugged into specific datasets or prediction models to fine-tune. Conceptually, Med-BERT creates a mapping between natural language and structured EHR under the pretraining fine-tuning paradigm.

Unlike conventional BERT that is usually trained on unstructured text, Med-BERT is trained on the combined version 9 and 10 International Classification of Diseases (ICD) codes[ii]. They are a set of alpha-numeric designations that provide a uniform terminology for medical professional to communicate diseases, symptoms, abnormalities, and other patient diagnostics. There is a total of more than 82K codes forming the vocabulary of the Med-BERT training data. To better represent the EHR data modality, Med-BERT also included visits and predicted prolonged length of stay in hospital (Prolonged LOS)[iii] as additional sources of contextual information, both of which provide another lens to evaluate the severity and complication level

of the patient's health condition, and very conveniently, neither requires human annotation. The diagnosis code and visits are converted to embeddings such that the transformer structure can capture the intercorrelations between codes and code ordering through serialization. Med-BERT effectively captures the data modality of structured EHR via a sequence of visits, each with a list of ordered or unordered set of diagnosis codes. Unlike the input modality of the original BERT, which is represented as a 1-D sequence of words, structured EHR is recorded in a multilayer and multi-relational modality in Med-BERT. This composes of code embeddings of individual diagnosis codes, serialization embeddings on code ordering, and visit embeddings to capture the sequence of patient encounters. The pretraining inherits the masked language model from the original BERT, in this case, predictions of diagnosis code are made given existing context. Diverging from the original BERT, Med-BERT chose the Prolonged LOS amongst a few candidate quality-of-care measures, to replace the question-answer pairs for the classification task. This clinical problem is not specific to any disease progression models and hence increases the generalizability of the pretrained model. The prolonged LOS reflects the bidirectional structure because it is not only indicative of a patients' historical health status but also impacts their subsequent journey through healthcare.

The pretrained Med-BERT was evaluated on two disease prediction tasks: the prediction of heart failure among patient with diabetes (DHF) and the prediction of onset of pancreatic cancer (PaCa), both of which are based on established phenotyping algorithms and capture clinical complexity beyond the existence of diagnosis codes. Layering Med-BERT on existing state-of-the art predictive models, including GRU[iv], Bi-GRU[v], and RETAIN[vi] as base models, boosted performance across all sample sizes in phenotyped cohorts. In addition, Med-BERT also proved to be more effective when compared against static embeddings such as t-W2V[vii]. It's worth noting that untrained Med-BERT performs poorly and are considered worse than baseline logistic regression. This is potentially due to overfitting given that untrained Med-BERT is an over-parameterized model (around 17 million parameters) with a many configurations.

Conclusion

Med-BERT proves to be particularly effective in capturing complex clinical semantics and variational medical context. When coupled with deep learning algorithms, it provides intricate and meaningful structures as input and injects knowledge that are comparable to domain expertise in tasks that involve uncovering deep patterns in underlying disease progressions. Masked LM[viii] and Prolonged LOS serve as two sources of reinforcement to establish sequential dependencies in patient care in a bidirectional and cumulative manner. These two features also significantly reduce the burden of data annotation because labels for both can be generated in an unsupervised way.

There are a few limitations with the current Med-BERT framework. It underperforms when benchmarked against baseline logistic regression models for smaller samples ($n<500$). The current vocabulary is only restricted to diagnosis codes and does not include other important clinical sources, such as time, medications, procedures, or lab tests. Lastly, there are certain degrees of temporal and clinical information loss due to the fact that length of time intervals between visits are ignored and no meaningful clinical episode groupings are attempted to categorize the visits. However, the framework should still be considered an effective methodologically innovation amongst various other NLP techniques applied to large EHR data.

[i] Pan, S. J. & Yang, Q. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359 (2009).
[ii] ICD-10 | CMS. http://www.cms.gov/Medicare/Coding/ICD10 (last accessed May 2021)
[iii] Bo M, Fonte G, Pivaro F, Bonetto M, Comi C, Giorgis V, Marchese L, Isaia G, Maggiani G, Furno E, Falcone Y, Isaia GC. Prevalence of and factors associated with prolonged length of stay in older hospitalized medical patients. Geriatr Gerontol Int. 2016 Mar;16(3):314-21. doi: 10.1111/ggi.12471. Epub 2015 Mar 9. PMID: 25752922.
[iv] Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In NIPS 2014 Workshop on Deep Learning, December 2014 (NIPS, 2014).
[v] Zhao, R. et al. Machine health monitoring using local feature-based gated recurrent unit networks. IEEE Trans. Ind. Electron. 65, 1539–1548 (2017).
[vi] Choi, E. et al. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. Adv. Neural Inf. Process. Syst. 29, 3504–3512 (2016)
[vii] Xiang, Y. et al. Time-sensitive clinical concept embeddings learned from large electronic health records. BMC Med. Inf. Decis. Mak. 19, 58 (2019).
[viii] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. 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), 4171–4186 (ACL, 2019).