CS 598 DLH Project Proposal

Amit Jangid, Constantin Kappel, Daniel Sanchez

March 24th 2024

ajangid2@illinois.edu, normank2@illinois.edu, daniel43@illinois.edu

1 Citation to original Paper

We picked a publication by Hur et al. [3] titled "*Unifying Heterogeneous Electronic Health Records Systems via Text-Based Code Embedding*".

2 General Problem

While EHR (Electronic Health Records) provide an attractive data source to do research on exposures and disease, there are many heterogenous medical code formats used by different healthcare providers. Thus, clinical studies using EHRs are difficult to scale up due to data incompatibilities. The technical hurdle is that previous systems learned hidden representations (latent embeddings) for each code system which projected the same medical concept with different encodings into different, incompatible semantic spaces. The paper [3] learns from the unstructured textual descriptions instead of the medcodes themselves, a strategy which the authors termed *description embedding* (*DescEmb*). Using *DescEmb* the authors were able to pool differently structured EHRs, namely *MIMIC-III* and *eICU*, into one larger dataset and achieved higher accuracy.

3 Scientific Approach

1. First, we will try to replicate the results of [3] as a baseline.
2. We plan on using `pyhealth`
   applicable. Especially for handling EHR data this should be very helpful.
3. Thirdly, we intend to test a hypothesis described in more detail in section 4. In short, we would like to replace the BERT[1]-derived encoders by one which was pretrained from scratch on biomedical data.

4 Hypotheses to be tested

The authors of [3] conducted experiments with several BERT-based architectures, such as BioBERT, ClinicalBERT and BlueBERT, which were partially trained or fine-tuned on medical literature. In contrast, Gu et al. [2] found that pre-training on generic NLP corpi, such as derived from Wikipedia,

[1] Bidirectional Encoder Representations from Transformers **(author?)** [1]

actually perform worse than ones which have been initialized with random values and only pre-trained on PubMed. They called their model PubMedBERT and were able to demonstrate superior performance to the models employed by [3].

We would like to test if PubMedBERT is capable of matching or even superceding the performance of the other models utilized, which were not fully pre-trained on pure corpi from the biomedical domain.

5 Ablations planned

A true ablation study with PubMedBERT and e.g. ClinicalBERT or BioBERT would mean we would have to pre-train all these architectures from scratch on both, generic NLP corpi as well as PubMed. This will not be computationally feasible. Instead, we will cite the work of [2] where applicable and rather try to replicate their results with different, already pre-trained architectures.

6 Data access

7 Feasibility of the computation

8 Existing Code

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[2] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, January 2022. arXiv:2007.15779 [cs].

[3] Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Young-Hak Kim, and Edward Choi. Unifying Heterogeneous Electronic Health Records Systems via Text-Based Code Embedding, March 2022. arXiv:2108.03625 [cs].