Project Draft - Team 91

You need to use the report template and fill out the following sections, each of which we will score based on the clarity and appropriateness of your writing (percentage of total grade for each component shown). All the information must be in the Jupyter notebook.

Introduction (2)

(A clear, high-level description of what the original paper is about and what is the contribution of it)

Longitudinal Electronic Health Records (EHRs) successfully used for clinical disease and outcome prediction using Deep Learning models. State-of-the-art (SOTA) models outperform traditional ML models by using pretrain-finetune methods in EHR-based predictive modeling. However, their pre-training objectives are limited in predicting fraction of ICD codes within each visit. In real life scenarios, patients have multiple diseases which can be correlated and can contribute to disease progression and change in outcome. Additionally, generalizing the same model on out-of-domain data in different medical settings with limited computing resources is a major challenge today.

The paper proposes TransformEHR, which is a generative encoder-decoder model with a transformer that is pre-trained using a new strategy, to predict the complete set of diseases and outcomes of patients at a future visit from previous visits. The model is generalizable and can be finetuned for various clinical prediction tasks with limited data.

TransformEHR uses encoder-decoder transformer architecture. The encoder processes the input embeddings and generates a set of hidden representations. The model performs cross-attention over hidden representations from the encoder and assigns an attention weight for each representation. The decoder generates ICD codes following the sequential order of code priority within a visit. It includes the date of each visit as a feature to integrate temporal information. The model uses 3 unique components compared to state-of-the-art models (BERT - Bidirectional Encoder Representations from Transformers) i.e. Visit masking, Encoder-decoder architecture and time embedding.

As per author, during the pretraining with a larger set of longitudinal EHR data, TransformEHR model learned the probability distribution of ICD codes through correlation of cross attention. Later It was fine-tuned to the predictions of a single disease or outcome.

Scope of reproducibility (2)

(no specific instructions given in the rubric)

The hypothesis we are planning to test is whether using the TransformEHR model can effectively predict the complete set of diseases and outcomes of a patient from past visits (i.e. Disease or outcome agnostic prediction (DOAP) task). The result will be compared to the state-of-the-art model (BERT - Bidirectional Encoder Representations from Transformers) that is usually trained to predict a fraction of ICD codes within each visit. The model will use a transformer architecture and seek to outperform bidirectional encode-only models.

The other hypothesis could be that the TransformEHR model can perform better for single disease outcomes with pre-training than without pre-training. After carefully reviewing the large data requirement of the VHA dataset for pretraining, we realized that testing this hypothesis may not be feasible with limited computational resources.

The ablations planned are as follows:

1. To evaluate the effectiveness of 3 of the unique components of the TransformEHR model, which are visit masking, encoder-decoder architecture, and time embedding.
2. To assess the performance of an Encoder-only architecture model compared to an encoder-decoder architecture.
3. Impact of the inclusion and exclusion of certain temporal features such as date of each visit.

Methodology (8)

Data

(no specific instructions given in the rubric)

Data Descriptions

To recreate the paper and its task, we will be using data from MIMIC-IV. The MIMIC-IV dataset comprises data from intensive care unit patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Although the dataset spans from 2008 to 2019, the implementation of ICD-10CM began in October 2015. As per authors implementation plan, to align with the cohorts from the Veterans Health Administration (VHA), only patients with ICD-10CM records were selected, resulting in a dataset of 29,482 patients.

We are able to fetch the data from https://physionet.org/content/mimiciv/2.2/icu/#files-panel, which seems to be close to what the paper used. There are a total of 299,712 patient records present in the dataset, however, when we filter the dataset on ICD-10 diagnosis codes, we are expecting to see the total patient record data come down to ~30K patients as mentioned in the paper.

Implementation code

Model

(no specific instructions given in the rubric)

Model Descriptions

Implementation code

Training

(no specific instructions given in the rubric)

Computational requirements

As we are going to use some of the existing code, we will be using the following:

- Operating systems:
  - Ubuntu 20.04.5 LTS
  - GPU (usage may be needed)
  - Google colab environment
- Python 3.8.11 with libraries:
  - NumPy (currently tested on version 1.20.3)
  - PyTorch (currently tested on version 1.9.0+cu111)
  - Transformers (currently tested on version 4.16.2)
  - tqdm==4.62.2
  - scikit-learn==0.24.2
  - Pyhealth (v1.1.6 release - latest version)

Implementation code

Evaluation

(no specific instructions given in the rubric)

Metrics Descriptions

Implementation code

Results

(no specific instructions given in the rubric)

Results

Analyses

Plans

References

Yang, Z., Mitra, A., Liu, W. *et al.* TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat Commun* **14**, 7857 (2023). https://doi.org/10.1038/s41467-023-43715-z. 2023 Nov 29;14(1):7857. doi: 10.1038/s41467-023-43715-z. PMID: 38030638; PMCID: PMC10687211.

https://www.nature.com/articles/s41467-023-43715-z
https://physionet.org/content/mimiciv/2.2/icu/#files-panel
https://github.com/whaleloops/TransformEHR