Impact of multimodal pretraining on patient similarity drivers

— Exposé —

by Alexander Reinicke

Introduction

Within the field of personalized medicine, determining patient similarity can play a crucial role. It enables healthcare professionals to identify patients with comparable characteristics, diagnoses and diseases and allows them to predict outcomes and determine the best possible care [Par+18]. Due to the sheer volume of medical data from thousands of patients across many hospitals, determining similar patients without computer assistance is often infeasible or outright impossible for any given healthcare professional. However, with the increasing availability of large medical datasets such as MIMIC-III, which contain rich multimodal data of thousands of patients, machine learning *(ML)* models are well-positioned to extract meaningful patterns from such data including patient similarity [JPM16]. This allows healthcare professionals to make informed decisions for their patients by leveraging the massive amount of medical data available. Deep learning based ML models can encode diverse medical information into embeddings, represented by a continuous vector space, where the proximity between embeddings generally reflect the similarity of patients. Understanding which characteristics of the patients (e.g. phenotypes, mortality, age, gender, etc.) most affect these embeddings can provide valuable insights into the underlying drivers of patient similarity.

Using multimodal data for pretraining, especially in the medical field, is a rather new strategy. Consequently, the impact of such pretraining on the generated embeddings remains underexplored [Azh+22]. Pretraining on large amounts of multimodal patient data could enhance the models ability to capture subtle patterns between different patients that may benefit patient similarity. On the other hand, a lack of pretraining may cause the model to find it difficult to learn complex patterns from scratch. This project aims to to evaluate how pretraining on multimodal and unimodal data affects the quality of patient similarity determinations.

State of the Art

Today, patient similarity analysis benefits from recent advances in machine learning, particularly in the use of embeddings to capture complex relationships between patients based on their medical data and characteristics [Zhu+16]. Transformer-based models such as BERT have been adapted to healthcare, benefiting from pretraining on large medical datasets to improve performance. Pretrained models such as BioBERT and ClinicalBERT have demonstrated superior performance when fine-tuned on clinical data in tasks like clinical prediction compared to models which have not been pretrained [HAR20]. Additionally, multimodal models that combine structured data (e.g. lab values, vital signs) with unstructured data (e.g. clinical notes) have shown promise in capturing more complex representations of patients [KYM23]. Despite

these advances, the question on how different types of pretraining affect the quality of patient embeddings remains an open question.

Goals

This is an exploratory project. The principal aim of this project is to determine which patient properties (e.g. phenotypes, mortality, age, gender, etc.) affect patient similarity for machine learning models the most. We will perform this evaluation using multiple models that have been pretrained on various data included in the MIMIC-III dataset:

- A model which has been pretrained on multimodal data (Text and Time-Series)
- A model which has only been pretrained on textual data
- A model which has not been pretrained

We will identify the driving properties for patient similarity for each model. A comparison between the results of each model allows us to then perform a qualitative analysis on what impact, if any, multimodal pretraining has on the driving properties affecting patient similarity.

Approach

The patient reports utilized for training and evaluation will be obtained from the publicly accessible MIMIC-III dataset [JPM16], which comprises deidentified medical data of thousands of patients who were admitted to an Intensive Care Unit. MIMIC-III contains data of multiple modalities including clinical notes, vitals, lab values, etc.

We will perform our evaluation with models that have been pretrained on different data as described in the Goals section. The contrastive pretraining presented in [KYM23] will be utilized, as it is able to align and combine the various modalities present in medical data, including vitals, lab values, notes, and other data. The models will be pretrained on the data using the publicly available source code that was used in [KYM23].

After pretraining models, we will generate embeddings for each patient and model. These embeddings will then be clustered. Patient embeddings, represented as vectors, that are closer together (i.e. vectors with less distance between them) are considered to be more similar by the models than embeddings that are farther apart.

We will also generate a cluster for each property type under investigation where all patients that have been assigned the same value are placed into the same cluster. The similarity clusters generated by the models and the property clusters will then be compared and their overlap determined (e.g. by using the Rand Index). This allows us to see whether patients considered to be similar have also generally been assigned the same property. Thereafter we can determine which patient properties affect similarity the most and which the least.

To assess the influence of multimodal pretraining versus unimodal and no pretraining, the results of each model are compared against each other and qualitatively evaluated.

In the future, this data analysis may also be extended to medical data sets other than MIMIC-III, for example, a dataset from the German hospital Charité, in order to ascertain whether the results obtained from this project are applicable in general or only to this particular dataset.

References

[JPM16] Alistair Johnson, Tom Pollard, and Roger Mark. “MIMIC-III Clinical Database”. Version 1.4. In: *PhysioNet* (2016). DOI: https://doi.org/10.13026/C2XW26.

[Zhu+16] Zihao Zhu et al. “Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding”. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016, pp. 749–758. DOI: 10.1109/ICDM.2016.0086.

[Par+18] E. Parimbelli et al. “Patient similarity for precision medicine: A systematic review”. In: *Journal of Biomedical Informatics* 83 (2018), pp. 87–96. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2018.06.001. URL: https://www.sciencedirect.com/science/article/pii/S1532046418301072.

[HAR20] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. 2020. arXiv: 1904.05342 [cs.CL]. URL: https://arxiv.org/abs/1904.05342.

[Azh+22] Zarif L. Azher et al. “Assessment of Emerging Pretraining Strategies in Interpretable Multimodal Deep Learning for Cancer Prognostication”. In: *bioRxiv* (2022). DOI: 10.1101/2022.11.21.517440. eprint: https://www.biorxiv.org/content/early/2022/11/24/2022.11.21.517440.full.pdf. URL: https://www.biorxiv.org/content/early/2022/11/24/2022.11.21.517440.

[KYM23] Ryan King, Tianbao Yang, and Bobak J. Mortazavi. “Multimodal Pretraining of Medical Time Series and Notes”. In: *Proceedings of the 3rd Machine Learning for Health Symposium*. Ed. by Stefan Hegselmann et al. Vol. 225. Proceedings of Machine Learning Research. PMLR, Oct. 2023, pp. 244–255. URL: https://proceedings.mlr.press/v225/king23a.html.