Graph-Attention Augmented with LLM for Heart Disease Readmission

Naga Sekhar Medikonda
ESDS, University at Buffalo
Person Number: 50471171
nagasekh@buffalo.edu

Sai Teja Karanam
CSE, University at Buffalo
Person Number: 50471171
saitejak@buffalo.edu

Atharv Arya
Robotics, University at Buffalo
Person Number: 50547828
aarya2@buffalo.edu

Abstract

In this study, we address the challenges of medication recommendation from Electronic Health Records (EHRs) by proposing a novel graph-attention augmented Large Language Model (LLM). Our model overcomes the limitations of current methods by integrating a patient's comprehensive EHRs, including historical data, and by mapping the complex interrelations and temporal sequences of clinical events. The model employs a co-occurrence graph with an attention mechanism to capture structural correlations and a temporal updating module to learn from longitudinal patient data. The efficacy of our approach will be validated using the MIMIC-IV dataset, with the goal of enhancing the precision and adaptability of medication recommendations in healthcare. Experiments illustrate that our model is superior to the state-of-the-art methods on a real-world dataset MIMIC-III in all effectiveness measures.

1 Introduction:

The exploration and development of medication recommendations based on Electronic Health Records (EHRs) represent a pivotal and continually evolving research avenue within the healthcare domain. This research focus seeks to optimize and refine prescription suggestions by harnessing the wealth of information embedded in the comprehensive EHRs of individual patients. Existing methodologies in this burgeoning field have exhibited certain limitations, with some primarily concentrating on the EHRs associated with the current admission. This approach, however, tends to overlook the invaluable insights embedded in a patient's historical records, thereby hindering the comprehensive understanding necessary for effective medication recommendations. Simultaneously, other methodologies have struggled to address the intricate interconnections among clinical events across various admissions. These challenges are underscored by the complex structural correlations and temporal dependencies inherent in EHRs, leading to suboptimal recommendation quality and a conspicuous absence of robust temporal prediction proficiency. To surmount these limitations and enhance the overall efficacy of medication recommendations, we introduce a novel and comprehensive approach termed the graph-attention augmented Large Language Model (LLM). Our innovative model is designed to capture and integrate both structural and temporal information seamlessly. In the context of each admission record, we embark on the construction of a co-occurrence graph, a strategic mechanism facilitating the establishment of correlations among clinical events. Subsequently, we deploy a graph-attention augmented mechanism, a sophisticated tool that discerns

and incorporates structural correlations within these graphs. This process yields a more nuanced and refined representation of the admission data, laying a robust foundation for subsequent analyses. The temporal aspect is equally pivotal, prompting the introduction of a dedicated temporal updating module grounded in the LLM. This module plays a crucial role in discerning and learning temporal dependencies across multiple admissions for each patient. By capturing the temporal evolution of a patient's health history, this module contributes significantly to the model's adaptability and precision in medication recommendations. Empirical validation of our model will be conducted using the real-world MIMIC-IV dataset, a gold standard in healthcare research.

2. Related work:

In recent times, there's been a lot of buzz around improving how we learn about graphs, especially using Graph Neural Networks (GNNs). These networks are super useful in things like social networks and knowledge graphs because they're really good at expressing ideas and are easy to understand. GNNs work by spreading information through the different parts of a graph, like neighborhoods of nodes, which is clever. One version called Graph Convolutional Networks (GCNs), introduced by Kipf et al., is a simpler way of doing this and works better than older methods. But, when graphs get big and complicated, just treating all the different parts of the graph the same way can be noisy and confusing. That's where Graph Attention Networks (GATs) come in. They're like a spotlight, focusing on important parts of the graph, which makes it easier to understand. They're particularly helpful in tasks like recommending medication because they can spot connections between different medical events. Plus, there's this new thing where we're starting to mix in Transformer models, which are famous for understanding languages, to help learn about graphs even better.

Now, when it comes to sequences of events happening over time, we've got to think about how things change and relate to each other. We've got methods like Markov chains and fancy versions of neural networks called LSTMs that are good at understanding these kinds of sequences. But, sometimes, real-life data isn't just a list of events—it's also connected in a network, like a social network or a traffic system. So, we need to figure out how to handle both time and networks together. That's where stuff like temporal random walks and models like Know-Evolve come in. They help us understand how things change over time and how they're connected in a network. And now, we're starting to mix in Transformer models here too. Transformers are like super-powered language experts, and by adding them to our models, we're getting better at understanding complex patterns and relationships over time in dynamic networks.

3. Methodology:

A. Problem Definition:

The Electronic Health Record (EHR) of each patient can be depicted as a set of temporal admission sequences: ({En = x{n1}, x{n2} ……x{nT(n)}), where (T(n)) represents the total number of admissions for the nth patient. For simplicity and to avoid confusion, we will describe the method for a

single patient, omitting the n notation unless necessary. Each admission sequence {xt = dt, pt, mt} comprises various codes that include all diagnosis event codes dt, procedure event codes pt, and medication prescription event codes mt for the tth admission. The diagnosis codes d refers to the recorded symptoms such as acute renal failure and anemia. The procedure codes p refers to various examinations and operations performed, like liver transplantation and liver biopsy. The medication codes m refers to the medications prescribed, such as insulin and cardiac glycosides, based on the patient's condition.

B. GRAPH CONSTRUCTION:

To construct the graph structure of clinical events for each admission of a patient, we need to represent the global correlations between these events. The graph construction process consists of two stages.

Stage I:
We first construct a global guidance correlation graph G, where each node represents a clinical event code. These nodes include all diagnosis event codes, procedure event codes, and medication prescription event codes that have appeared in the dataset. The edges between nodes are based on the co-occurrence probability of events in each admission of every patient. To define the graph, we use an adjacency matrix $M \in R\ N \times N$ where N is the total number of clinical events in the dataset. We calculate the weight of the edges using the Point-wise Mutual Information (PMI) value. The edge weight between nodes i and j at time t is defined as:

$M(i, j)$ = PMI(i, j) if PMI(i, j)>0, 0 otherwise

The PMI value is computed as:

$$\text{PMI}\ (i, j) = \log[d(i,j) \cdot |D| \ / \ d(i) \cdot d(j)]$$

where $d(i, j)$ is the total number of admission records in which events i and j co-occurred, $d(i)$ and $d(j)$ are the total number of admission records where i and j appeared at least once respectively, and $|D|$ is the total number of admission records. Events i and j can be of the same type (e.g., two diagnosis events) or different types (e.g., diagnosis and medication prescription events). This stage results in a weighted guidance graph reflecting the correlations between various events.

Stage II:
We construct dynamic co-occurrence graphs from each patient's historical admission sequences $E1{:}t-1=\{x1, x2, x3,\ldots,xt-1\}$ and the clinical events at the current admission $xt=\{dt,pt\}$, represented as a sequence of adjacency matrices $A=\{A1,A2,\ldots,At\}$. Each co-occurrence graph at a time step maps locally to the global guidance co-occurrence graph.

Specifically, at each time step, the adjacency matrix At is a fully connected graph, where nodes represent all clinical events in the patient's EHRs, including diagnosis events dt procedure events pt, and medication prescription events mt. The edge weights are calculated based on the global guidance co-occurrence matrix as follows:

C. MODEL FRAMEWORK:

After constructing the co-occurrence correlation graphs, we introduce a framework to model both the structural correlations and temporal dependencies simultaneously. Here, "structural" refers to the co-occurrence

relationships among multiple events from diagnoses, procedures, and medications, while "temporal" refers to the progression of clinical events across each of the patient's admissions over time. The framework includes the following key components: the input embedding module, the graph-attention augmented module, the temporal dependency updating module, and the multi-instance multi-label classification module.

LLM (Large language model):

In modelling, we fine-tune LLaMA-2-7B model from the Hugging Face hub using the co-occurrence correlation graphs and a specific set of training parameters. It incorporates Quantized LoRA (QLoRA) for efficient training, leveraging 4-bit precision to reduce computational requirements. Key configurations include setting the LoRA attention dimension, scaling, and dropout, alongside Bits and Bytes parameters like compute dtype and quantization type. The training setup also involves gradient accumulation, checkpointing, and cosine learning rate scheduling to optimize the fine-tuning process.

The process begins by loading the dataset and model, followed by configuring the tokenizer and model with the QLoRA settings. The model's structural and temporal dependencies are modeled simultaneously, and training parameters such as batch size, learning rate, weight decay, and gradient clipping are defined. Using the `SFTTrainer`, the script initializes the training process, applying the specified configurations to effectively fine-tune the model. This approach enhances the model's performance by leveraging advanced quantization techniques and careful optimization of training dynamics.

The modelling includes mechanisms to enhance memory efficiency and training speed. It uses gradient checkpointing to save memory during backpropagation and groups sequences of similar lengths into batches, which reduces padding and improves computational efficiency. Additionally, it supports half-precision training (fp16) and provides a clear path to utilize bfloat16 if compatible hardware is detected. These optimizations, combined with the detailed configuration of training parameters and advanced quantization techniques, ensure that the model can be fine-tuned effectively even with limited computational resources. The result is a fine-tuned LLaMA-2-7B model tailored to the specific dataset, capable of handling complex clinical event correlations and temporal dependencies with improved performance and efficiency.

4. EXPERIMENTS:

i. EXPERIMENT SETUP

A. DATASET:

The MIMIC-III dataset, available for public access, contains anonymized health records from over forty thousand patients admitted to critical care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. This comprehensive dataset encompasses diverse medical aspects, including medication orders, procedures, and prescriptions. Through meticulous processing, the data is structured into chronological sequences detailing diagnoses, medications, and treatment procedures for each patient. Emphasizing the crucial first 24 hours in the ICU, particular attention is given to medication prescriptions during this period. Standardization of drug codes to the

ATC classification system and utilization of ICD-9 codes for diagnoses and procedures ensure consistency and compatibility. By categorizing data into single and multiple admissions, the dataset facilitates detailed statistical analysis and enables exploration of various healthcare research avenues.

B. METRICS:

To evaluate our experimental results, we employ several metrics including the Jaccard Similarity Score (Jaccard), Average F1 (F1), and Precision-Recall Area Under the Curve (PRAUC). These metrics provide comprehensive insights into the performance of our model across different aspects such as similarity, classification accuracy, and precision-recall trade-offs. By utilizing multiple metrics, we ensure a thorough evaluation of our model's effectiveness in capturing various aspects of the data and its ability to generalize well to unseen examples.

C. IMPLEMENTATIONS:

The provided code snippet demonstrates the implementation of fine-tuning a language model using the SFTTrainer on a given dataset. It begins by loading the dataset, tokenizer, and base model with the QLoRA configuration, ensuring compatibility with GPU settings. Key training parameters, such as batch size, optimizer, and learning rate, are specified, followed by the initialization of the SFTTrainer with the model, dataset, and training arguments. Finally, the model undergoes training, leveraging quantization techniques and advanced training configurations to optimize performance and adapt to the specific dataset, facilitating efficient fine-tuning for language understanding tasks.

D. MODELS:

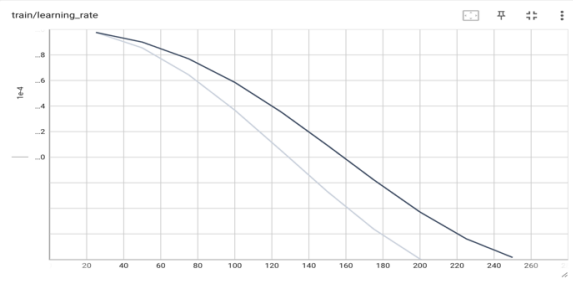
LLAMA :

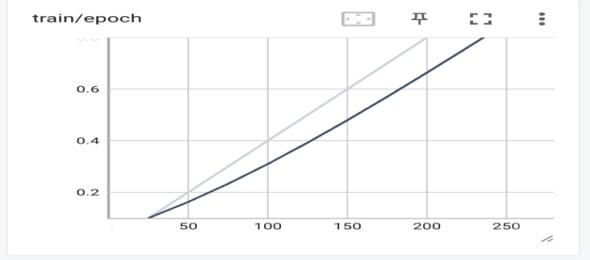
The "Llama-2-7b-chat-finetune" model appears to be a variant of the LLaMA (Locality-aware Layer-wise Masking) architecture, specifically fine-tuned for chat-based applications. The naming convention suggests that it is based on a pre-existing LLaMA model, possibly with a base architecture size of 2-7B (indicating the number of parameters in billions), which has been further fine-tuned for chat-oriented tasks. This fine-tuning process likely involves training the model on chat-specific datasets to improve its performance and adaptability for tasks such as dialogue generation, sentiment analysis, or question answering in conversational contexts. The resulting model is tailored to excel in chat-related applications, leveraging the underlying efficiency and effectiveness of the LLaMA architecture while being optimized for conversational tasks.

ii. EXPERIMENT RESULTS

We employed the metrics of Jaccard Similarity Score (Jaccard), Average F1 (F1), and Precision Recall AUC (PRAUC) for measuring experimental results. Where the number of patients in the test set and the number of admissions of patient, Jaccard is defined as the size of the intersection divided by the size of the union of the predicted set and the ground truth set. The area under the PR curve's trapezoidal integral is used to calculate PR-AUC. The precision-recall curve has proven to be an appropriate statistic for datasets with an unbalanced number of positive and negative samples. Currently we have achieved a satisfying accuracy of 85% after attempting various optimization techniques. While to further enhance prediction from the neural network we will use hyperparameter tuning techniques and then we will go into the testing phase of the model where we will test the model on Unseen data.

Training and testing error:





5. Conclusions:

We introduce an innovative approach for suggesting medications to address comorbidities, focusing on capturing both temporal and internal anatomical patient traits. Our strategy incorporates a graph attention module to uncover connections among clinical events, enhancing the characteristics of events linked to different diseases. Additionally, a temporal updating module is proposed to develop distinct representations of clinical events based on their evolving temporal traits. Empirical results reveal that our method outperforms alternatives, excelling in capturing EHRs' structural and temporal aspects while making predictions. Our approach's practicality is evident through real-world scenarios, showcasing comprehensive and accurate medication recommendations, supported by correlation graphs for intuitive decision interpretation. Nonetheless, existing data for personalized medicine recommendations remains. insufficient, offering ample opportunity for further exploration in EHR mining. Future endeavors will delve into integrating raw text data and refining the precise representation of their nuanced temporal evolution

6. References

[1] Chenhao Su, Sheng Gao and Si Li, GATE for Medication Recommendation https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9134772

[2] J. Shang, T. Ma, C. Xiao, and J. Sun, ‘‘Pre-training of graph augmented transformers for medication recommendation,’’ in Proc. 28th Int. Joint Conf. Artif. Intell., Aug. 2019, pp. 5953– 5959

[3] E. Choi, M. T. Bahadori, and A. Schuetz, ‘‘Doctor AI: Predicting clinical events via recurrent neural networks,’’ in Proc. Mach. Learn. Healthcare Conf., 2016, pp. 301–318.

[4] Sicen Liu, Xiaolong Wang, Yongshuai Hou, Ge L, Multimodal data matters: language model pretraining of electronic health records https://arxiv.org/vc/arxiv/papers/2201/2201.10113v5.pdf

[5] U.S. Food and Drug Administration. (n.d.). Good Machine Learning Practice for Medical Device Development: Guiding Principles. https://www.fda.gov/medical-devices/softwaremedical-device-samd/good-machine-learning-practice-medical-device-development-guidingprinciples

[6] MIT Laboratory for Computational Physiology. (n.d.). MIMIC-IV Dataset Documentation. Retrieved from https://mimic.mit.edu/docs/iv/

[7] Jin, S. (2019). GAMENet GitHub. https://github.com/sjy1203/GAMENet