Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Leisheng Yu April 12, 2022

Deep Learning for EHR-Based Diagnosis Prediction: A General Recipe

by

Leisheng Yu

Carl Yang
Advisor

Department of Mathematics

Carl Yang

Advisor

Bree Ettinger

Committee Member

Davide Fossati

Committee Member

David Zureick-Brown

Committee Member

2022

Deep Learning for EHR-Based Diagnosis Prediction: A General Recipe

By

Leisheng Yu

Carl Yang

Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Mathematics

2022

Abstract

Deep Learning for EHR-Based Diagnosis Prediction: A General Recipe
By Leisheng Yu

With the rapid accumulation of Electronic Health Records (EHRs) and the recent advance in data-driven algorithms, deep learning models have been increasingly applied to tasks in EHR-based predictive healthcare. This paper, motivated by the hierarchical structure of EHR data and the identified challenges in predictive healthcare, mathematically formulates a general architecture for learning patient representations for diagnosis prediction. With the guidance of the proposed general architecture, this work further discusses how existing works have incorporated various model designs to overcome certain challenges from four levels: diagnosis, visit, sequence, and framework-level. Through these discussions, this paper serves as a summary of existing works in modeling sequential EHR data, a cookbook for novices interested in EHR-based predictive healthcare, and a foundation for a future work, the idea of which are also introduced in this paper.

Deep Learning for EHR-Based Diagnosis Prediction: A General Recipe

By

Leisheng Yu

Carl Yang

Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Mathematics

2022

Acknowledgments

First and foremost, I would like to thank my advisor, Professor Carl Yang, who has led me into machine learning and data mining research. I joined the Emory Graph Mining group, which is led by Professor Carl Yang, in my sophomore year. Initially I knew very little about machine learning and data mining; however, now I have developed interests in doing relevant research in this field. In this process, he has offered me a lot of guidance and support: he has given me valuable advice on doing scientific research, showed me the importance of commitment and self-motivation, helped me with my graduate school application, and encouraged me to accomplish the honors thesis.

In addition, I also want to thank my committee members: Professor Davide Fossati, Professor Bree Ettinger, and Professor David Zureik-Brown. They spent time to attend my defense and provide me with valuable suggestions on improving the thesis.

Moreover, I would like to thank Yanchao Tan, who has been continuously offering me hands-on guidance on machine learning and data mining research. In the past year, she has helped me improve myself in doing research through assigning me specific tasks and including me in various types of collaborations.

Furthermore, I want to thank my close friends in Emory throughout these three years. They have been supporting me in both academics and daily life. I am deeply grateful for the friendship and mutual understanding developed in college.

Last but not least, I am grateful for the support, both emotionally and financially, from my biological family in China. Thank you!

Contents

List of Figures

Chapter 1

Introduction

Electronic Health Records (EHRs) are large-scale and systematic collections of sequences of patients' hospital visits in a chronological order, representing the longitudinal health experience of patients [9, 58]. Each hospital visit of a patient recorded in EHRs has a time stamp and contains various medical information, such as demographics, prescriptions, diagnoses, lab test results, vital signs, and output measurements [26, 21, 35]. Hence, EHR data can be viewed as temporal sequences of high dimensional clinical variables [10].

The broad adoption of electronic healthcare systems, the rapid accumulation of EHR data, and the recent advance in data-driven algorithms signal an unprecedented opportunity to employ deep learning models for predictive healthcare, which is significant for improving the quality of clinical care, assisting clinical decision making, promoting personalized medicine, and optimizing the resource allocation in hospitals [10, 12, 58, 56, 55]. Fortunately, deep neural networks, especially recurrent neural networks (RNNs), have not disappointed the researchers: deep learning models can alleviate the need for feature engineering on EHR data and have achieved state-of-the-art performance in various EHR-based tasks in predictive healthcare [44]. Most of the tasks are to predict the future health information or medical outcomes of patients

from their historical EHRs [29]: diagnosis prediction [9, 10], prescription prediction [9, 26], mortality prediction [44], length-of-stay prediction [44], and readmission time prediction [1]. EHR-based tasks in other categories include patient subtyping [3], disease subtyping [49], computational phenotyping [51], patient similarity analysis [51], and treatment recommendation [49]. Among all these tasks, diagnosis prediction will be the focus of this paper.

The key of applying deep learning models to EHR-based diagnosis prediction is to learn robust vector representations of patients [34, 26]. The learned patient representations can be fed into a classifier to perform predictions. Various model architectures have been designed to overcome specific challenges in this domain, and these models can be summarized by a general mathematical recipe. This paper, through proposing a fundamental architecture that characterizes the functions encoding patient EHR data for diagnosis prediction, aims to promote a discussion on different strategies for overcoming specific challenges and to serve as a cookbook for researchers interested in deep learning for predictive healthcare.

Before presenting the general recipe, we will first give a mathematical formulation of the diagnosis prediction task and the structure of EHR data used as input. Moreover, we will briefly summarize the challenges in modeling longitudinal EHR data, which motivates the further discussion of model designs in different levels.

1.1 Longitudinal Patient EHR Data

In general, patient EHR data can be viewed as sequences of unordered sets. Specifically, a patient in EHR data can be modeled as a time-stamped sequence of n tuples: (v_i, t_i) for $i = 1, ..., n$. The vector v_i corresponds to the i-th visit in the sequence, and each visit $v_j \in \{0, 1\}^{|C|}, j \in [1, n]$ is a multi-hot binary vector, where $|C|$ denotes the number of unique diagnoses and $C = \{c_1, c_2, ..., c_C\}$ is the set of all unique diagnosis

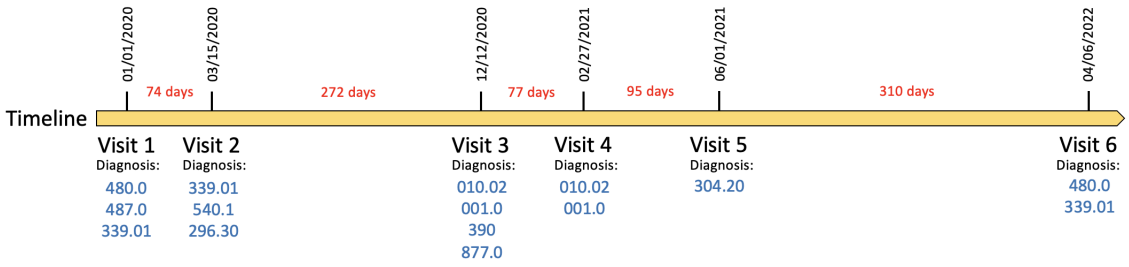

Figure 1.1: The hierarchical structure of longitudinal patient EHR data

in the EHR data. $v_{i,j} = 1$ indicates that the patient was diagnosed with c_j in the i-th hospital visit. The time stamp t_i denotes the time of the i-th hospital visit.

Thus, patient EHR data have a hierarchical structure: each patient is a sequence of hospital visits and each hospital visit is an unordered set of diagnoses. Figure 1.1 gives a graphical illustration.

Another crucial characteristic of EHR data is that every diagnosis in C has a corresponding code in International Classification of Diseases, Ninth Revision[1] (ICD-9), which is a well-organized tree-structure medical ontology consisting of "parent-child" relations between diseases. Figure 1.2 presents a subgraph of this disease taxonomy. The diagnosis codes used in Figure 1.1 are from ICD-9.

1.2 EHR-Based Diagnosis Prediction

The task of EHR-based diagnosis prediction can be split into two categories.

The first category is called disease progression modeling [10], which is to predict what diseases will the patient be diagnosed with in the next hospital visit given the historical EHR data. This task is not limited to the prediction of one specific disease, instead, it aims to forecast all diseases that the the patient will be diagnosed with in the future. Therefore, this is a multi-label classification task.

The second category is called disease risk prediction [28]. Different from disease

[1] https://www.cdc.gov/nchs/icd/icd9cm.htm

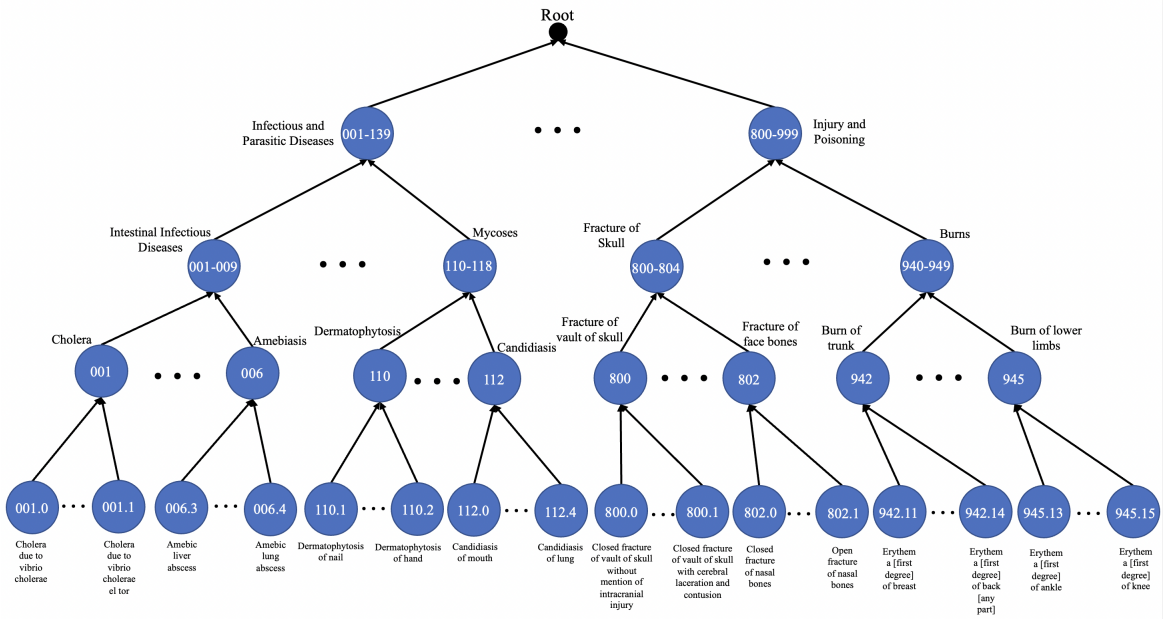

Figure 1.2: The tree structure of ICD-9 Ontology

progression modeling, disease risk prediction aims to predict whether a patient will be diagnosed with a single target disease in his/her next hospital visit. For example, this target disease can be heart failure [10, 12, 56, 51], renal failure [30], Alzheimer's disease [55], or septic shock [58]. Therefore, this is a binary classification task.

The input to these two categories of diagnosis prediction is the same: a collection of patient records described in Section 1.1. The output of disease progression modeling is a multi-hot vector o_{N+1} representing the set of predicted diagnoses in the next visit; the counterpart of disease risk prediction is a probability of whether the patient will be diagnosed with the target disease in the next visit.

1.3 Challenges in Modeling Longitudinal EHR Data

Challenges originating from various aspects of predictive healthcare have been motivating the continuous efforts in refining the design of deep learning models for diagnosis prediction. These challenges mainly come from four sources: the characteristics of EHR data, the nature of tasks in predictive healthcare, clinical common sense, and

the inherent defects of RNN models.

Challenges originating from the characteristics of EHR data include temporality [10, 56], high dimensionality [10, 56], irregular time intervals between consecutive hospital visits [3, 58, 56, 51, 60], data insufficiency [31, 12, 56, 51, 40, 26], heterogeneity [60], and noisiness [37, 58, 60]. Challenges brought by the nature of tasks in predictive healthcare are model interpretability [10, 1, 58, 56, 51, 57] and trustworthiness [57, 5]. Ample challenges come from clinical common sense: incorporating patient demographics [16, 5], location factors [5], disease progression stages [17], structural information [38] as well as clinical relations [49, 27] among diseases, and prior medical knowledge [11]; modeling fine-grained progression patterns of patient health conditions [16] and non-stationary disease progression [28]. The failure of RNNs in capturing long-term dependency [16, 60] and being significantly accelerated with distributed/parallel computing schemes [34] are also problems that researchers have been trying to solve in modeling sequential EHR data.

More detailed descriptions of the aforementioned challenges and various strategies proposed to overcome them will be presented in later chapters.

Chapter 2

General Recipe

In this chapter, we propose a mathematical formulation of a fundamental architecture for encoding longitudinal patient EHR data and learning patient representations. The learned patient representations can be used for diagnosis prediction. The formulation of this general recipe is motivated by the hierarchical nature of sequential EHR data [12, 34].

To enable a more convenient representation of an unordered set of diagnosis codes, we expand $\mathbf{v}_i^n$, a multi-hot vector denoting the i-th hospital visit of the n-th patient in the dataset, as a sum over a set of unique one-hot vectors,

$$\mathbf{v}_i^n = \mathbf{c}_a + \mathbf{c}_b + ... + \mathbf{c}_z \tag{2.1}$$

where $\mathbf{c}_a, \mathbf{c}_b, ..., \mathbf{c}_z \in C$ denote unique diagnosis codes in C, which is the set of all unique diagnosis codes in EHR data. With this expression, we can formulate the general architecture for learning the representation of the n-th patient from longitudinal EHR data as follows,

$$\mathbf{z}^n = h\{g[\sum_{i=1}^{N_{v_1^n}} f(\mathbf{c}_i)], g[\sum_{i=1}^{N_{v_2^n}} f(\mathbf{c}_i)], ..., g[\sum_{i=1}^{N_{v_{T(n)-1}^n}} f(\mathbf{c}_i)], g[\sum_{i=1}^{N_{v_{T(n)}^n}} f(\mathbf{c}_i)]\} \tag{2.2}$$

where $\mathbf{v}_j^n$ denotes the j-th visit of the n-th patient, $T(n)$ denotes the total number of observed hospital visits of the n-th patient, $N_{v_j^n}$ denotes the number of diagnoses in the j-th visit of the n-th patient, c_i denotes the i-th diagnosis in a hospital visit, $f(\cdot)$ is the function mapping a diagnosis code $\mathbf{c}_i$ (one-hot vector) to a dense and continuous vector, $g(\cdot)$ together with the sum operator learns the representation of a hospital visit by aggregating the constituent diagnosis representations, and $h(\cdot)$ is the function that learns the final representation of a patient by aggregating his/her sequence of visit representations. $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ can either be parameterized or non-parameterized, depending on the desired level of expressive power and the target challenges to overcome.

After obtaining the final representation, $\mathbf{z}^n$, of the n-th patient, we can feed it to a classifier that generates the prediction. This classifier is often the final layer of the model, and it is typically designed as follows,

$$\hat{y} = \sigma(FC(\mathbf{z}^n)) \tag{2.3}$$

where $\hat{y}$ is the predicted label, FC denotes a fully connected layer, and σ corresponds to the activation function. If the category of diagnosis prediction that we want to perform is disease progression modeling, which is a multi-label classification problem, σ should be chosen as the Softmax function; conversely, if disease risk prediction, which is a binary classification task, is the goal, then we should use the Sigmoid function as the activation function.

Cross-entropy is a loss function typically used for classification tasks. Specifically, for disease risk prediction, the loss function is as follows,

$$l(\hat{y}, y) = -(y \cdot log(\hat{y}) + (1 - y) \cdot log(1 - \hat{y})) \tag{2.4}$$

and the loss function for disease progression modeling is as follows,

$$l(\hat{y}, y) = \frac{1}{K} \sum_{k=1}^{K} -(y_k \cdot log(\hat{y}_k) + (1 - y_k) \cdot log(1 - \hat{y}_k)) \tag{2.5}$$

where K denotes the number of labels, y_k denotes the ground truth of label k, and $\hat{y}_k$ denotes the predicted result for label k.

The general architecture for patient representation learning, the final prediction layer, and the generic loss function together constitute the general recipe of deep learning for EHR-based diagnosis prediction. In later chapters, we will investigate how existing works design the three hierarchical encoding functions—$f(\cdot)$, $g(\cdot)$, and $h(\cdot)$—to overcome the challenges specified in Section 1.3. Moreover, we will also explore some existing designs on the framework level, outside of $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$. For example, the framework-level designs can be relevant to loss functions or optimization tricks.

Chapter 3

Diagnosis-Level Representation Learning

This chapter focuses on the design of $f(\cdot)$ in Equation 2.2. The progress of learning the representations for diagnoses in the EHR data has been constantly motivated by various challenges, especially data insufficiency. External knowledge enhancement and utilizing different types of clinical information can be the key. We will see how existing works in this domain have moved from simple embedding approaches to graph-based methods in the following sections.

3.1 Simple Embedding Table

Most of the early works in leveraging deep learning to model sequential EHR data chose to encode diagnosis codes to dense and continuous vector representations via a simple embedding table [9, 10, 1, 37, 38, 34]. Mathematically, this strategy can be formulated as follows,

$$\mathbf{d}_i = \mathbf{W}_{emb}\mathbf{c}_i \tag{3.1}$$

where $\mathbf{c}_i \in \mathbb{R}^{|C|}$ is a one-hot vector denoting a specific diagnosis code in C, $\mathbf{W}_{emb} \in \mathbb{R}^{m \times |C|}$ is an embedding matrix to be learnt, and $\mathbf{d}_i \in \mathbb{R}^m$ is the diagnosis embedding obtained after passing through this linear embedding layer. Since there are thousands of unique diagnosis codes in the EHR data, the one-hot vector $\mathbf{c}_i \in \mathbb{R}^{|C|}$ is extremely high-dimensional and sparse. However, after being multiplied with $\mathbf{W}_{emb} \in \mathbb{R}^{m \times |C|}$, the one-hot vector can be mapped to a dense and continuous embedding of a lower dimension m. Therefore, this simple linear embedding approach can manage to overcome the challenge of high dimensionality in EHR data.

However, deep learning models typically require a large amount of training data [56], and unfortunately, a single EHR dataset can have a significant data insufficiency issue, because a hospital is normally only visited by patients living in nearby areas and a patient may visit multiple hospitals in one period of time. The phenomenon that a diagnosis code appearing in the test set does not exist in the training data, which leads to unsatisfactory representation learning for this diagnosis code and thus unfavorable performance, is common in healthcare applications. Moreover, this simple embedding table method learns embeddings of different diagnosis codes independently, ignoring the inherent structural information among diseases. For example, introverted personality (301.21) and Schizotypal personality disorder (301.22) have the same parent or ancestors in the ICD-9 hierarchy, which means that their embeddings should be relatively close.

Motivated by the two challenges mentioned above, several models have been proposed to make use of ICD-9, a well-organized public medical ontology.

3.2 Incorporating Medical Ontology

GRAM [11] is a predecessor of incorporating ICD-9 taxonomy into modeling sequential EHR. Since the diagnosis codes recorded in most EHR datasets correspond to

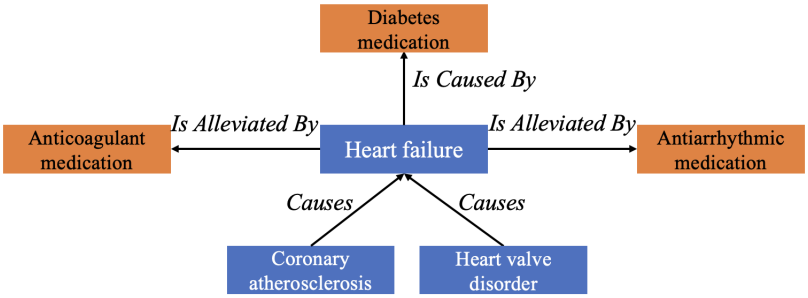

Figure 3.1: A subgraph of MKG consisting of heterogeneous medical entities and relations

the leaf nodes in the ICD-9 hierarchy, GRAM proposed to obtain the embedding for each diagnosis code in the EHR data by aggreagting the embeddings of corresponding ancestors:

$$\mathbf{d}_i = \sum_{j \in \mathcal{A}(i)} \alpha_{ij} \mathbf{e}_j \tag{3.2}$$

where $\mathcal{A}(i)$ denotes the indices of diagnosis code $\mathbf{c}_i$ and $\mathbf{c}_i$'s ancestors, $\mathbf{e}_j$ denotes the basic embedding of either a diagnosis code in C or an ancestor in the ICD-9 hierarchy, and α_{ij} denotes the weight learned by an attention mechanism and $\sum_{j \in \mathcal{A}(i)} \alpha_{ij} = 1$. Many works following GRAM have adopted this approach for learning diagnosis code representations [31, 56, 16, 52].

Besides the parent-child structural information among diseases, rich information contained in external medical knowledge graphs (MKGs) is also valuable to be incorporated into the modeling process. From Figure 3.1, we can observe that there are different types of clinical relations attached to the edges in MKGs. Therefore, some existing works have proposed to utilize MKGs to further alleviate data insufficiency.

3.3 Incorporating External Medical Knowledge Graph

Motivated by the success of graph neural networks (GNNs) in recent years [24, 19], some existing works proposed to leverage TransE [4] together with a message passing

module to learn the embeddings of the diagnosis codes from external MKGs [56, 51]. Then the learnt diagnosis embeddings can be either directly aggregated to obtain the visit-level embeddings or first concatenated with the embeddings learned from the ICD-9 ontology and then aggregated.

Although the ICD-9 ontology and the MKGs can provide rich information in terms of parent-child as well as causal relationships between diseases, the clinical relations between diseases can be overlooked if disease co-occurrences are not incorporated in the modeling process. Some diseases that are not relevant in the ICD-9 taxonomy can frequently be diagnosed with simultaneously; in contrast, some diseases close in the hierarchy can rarely co-occur in a single hospital visit: taking into account the clinical relations can be beneficial for disease progression modeling.

3.4 Incorporating Disease Co-Occurrence Graph

Chet [27] constructs a directed co-occurrence graph following the idea that there is an edge between two diagnoses if their number of co-occurrences is larger than a threshold of relative frequency. RGNN [26] constructs a time-aware diagnosis graph based on the rule that diagnoses in visit v_i are connected to the diagnoses in visits v_{i-1} and v_{i+1}, and the edge weight corresponds to the time interval between two consecutive visits. These two works both proposed to leverage a GNN model to learn the node embeddings on the constructed co-occurrence graph.

Heterogeneity is an important feature of EHR data: fully utilizing various types of information in EHR can better assist the prediction. Therefore, MiME [12] constructs a co-occurrence graph among diseases and treatments, and GMAN [52] includes the co-occurrence among diseases and symptoms to learn the embeddings of diagnoses in a more comprehensive way.

Chapter 4

Visit-Level Representation Learning

This chapter investigates how existing works in sequential EHR modeling have designed the function $g(\cdot)$ in Equation 2.2, and some ideas that can potentially be borrowed from the domain of set representation learning.

Since a hospital visit can be modeled as an unordered set of diagnoses codes, the function $g(\cdot)$ should be permutation invariant [53], which means $g(\cdot)$ needs to be indifferent to the ordering of the constituent diagnoses. The simplest way of modeling an unordered set in a permutation invariant way is to add up the embeddings of all the elements in the set. Most of the existing works in modeling sequential EHR data resorted to this sum operation to obtain visit embeddings [9, 10, 11, 31, 37, 12, 38, 27]. This sum operation is similar and closely related to Equation 3.1:

$$\mathbf{x}_i = \mathbf{W}_{emb}\mathbf{v}_i \tag{4.1}$$

where $\mathbf{v}_i$ is a multi-hot vector denoting the i-th hospital visit, $\mathbf{W}_{emb}$ is same as the one in Equation 3.1, and $\mathbf{x}_i$ is the learnt visit embedding. The problem of this straightforward approach is that it assumes that every diagnosis in the set contributes

equally to the visit embedding, which fails to provide interpretations of the predicted result to clinical doctors.

Interpretability of the employed model is crucial for healthcare applications. In order to trust the prediction of the model, clinical doctors must be able to understand the rationale behind the model prediction: why does the model make such a prediction [1]?

4.1 Attention Pooling

The attention mechanism has been introduced to model a hospital visit with interpretability. Some of the existing works proposed to leverage attention pooling to aggregate diagnosis embeddings. In this way, the hospital visit, a composite object, can be represented as a weighted sum of the diagnosis embeddings, where the weights are attention weights learned by certain attention mechanisms [10, 1, 34]. Moreover, to capture unique disease progression patterns, Timeline [1] calculates another set of weights to be combined with attention weights, by utilizing a disease-specific progression function. Through observing the weight value corresponding to each diagnosis, domain experts can tell which diagnoses in a visit are the most important for the prediction.

Although some of the aforementioned works utilized attention pooling to obtain visit embeddings, this department, the design of $g(\cdot)$, has not received much attention. However, following Deep Sets [53], a work that proposed a general architecture for deep learning models operating on set-structured data, there has been a branch of research focusing on leveraging deep learning models to learn set representations. The ideas proposed in the domain of set representation learning can potentially be exploited in modeling hospital visits. Therefore, several existing works leveraging deep learning techniques for set representation learning are briefly introduced in the

following section.

4.2 Set Representation Learning

The most important principle for representation learning on unordered sets is permutation invariance. PointNet [36] applies multi-layer perceptron (MLP) and feature transformations on the elements in the set and used max-pooling to aggregate information. RepSet [43] measures the similarity between the input set and each one of the hidden sets by bipartite matching to learn the representation. [23] proposed a differentiable Expectation-Maximization model to represent a set of objects and [13] provided an optimal transport based way to learn set representation. Some of the existing permutation invariant approaches can also manage to tackle other nontrivial challenges in set representation learning. Specifically, to incorporate the interactions among elements in the set, SetRank [33] leverages a stack of induced multi-head self attention blocks, and DMPS [42] constructs a latent graph from the set and performs message passing on it. Moreover, to weight elements in the set, Set Transformer [25] employs multi-head attention for both processing every elements and pooling.

Chapter 5

Sequence-Level Representation Learning

The design of $h(\cdot)$ is the focus for most of the existing works in diagnosis prediction, because temporality and sequentiality are the most significant and explicit characteristics of EHR data, and diagnosis prediction is a sequential prediction problem in essence. Since RNNs are popular for modeling sequential data, the variants of RNNs, Long Short-Term Memory (LSTM) [20] and Gated Recurrent Units (GRU) [8] are part of the first wave of success that deep learning had achieved in modeling longitudinal EHR data.

5.1 Recurrent Neural Networks

Some works directly applied RNN models like LSTM or GRU, without any modifications, to encode the sequence of hospital visits [9, 1, 11, 12]. Although this simple architecture can capture the sequentiality of EHR data, it has many limitations. For example, RNNs are like black box, which means domain experts cannot understand why the model makes a certain prediction. Moreover, RNNs are known to suffer from capturing long-term dependency because of its forgetfulness when the sequence

is long.

To overcome RNNs' drawback in modeling long-term dependency, Dipole [29] utilizes bidirectional RNN to consume the input sequence from two opposite directions; some other works [16, 60] resorted to augmenting RNN models with an external memory network which can store detailed information of all the hidden states in the long sequence, thus the fine-grained progression patterns of patient health conditions can also be captured.

Some existing works also proposed multi-thread GRUs, which means employing multiple GRUs simultaneously, to overcome specific challenges. ConCare [32] utilizes a GRU for each type of sequential information in EHR (including lab test results and historical diagnoses) to make good use of the heterogeneity of EHR data. Chet [27] splits the diagnoses in each hospital visit into three types and applies a GRU for each type, in order to capture the fine-grained progression patterns.

Lack of interpretability is the major challenge of applying RNNs to diagnosis prediction. Similar to what we have introduced in Chapter 4, some works integrated the attention mechanism with RNN models [10, 29, 37]: applying self-attention on the sequence of hidden states generated by the RNN model to get attention weights. After obtaining the attention weights, the weighted aggregation of hidden states can serve as the context vector:

$$\mathbf{c}^n = \sum_{i=1}^{T(n)} \alpha_i \mathbf{h}_i \tag{5.1}$$

where $T(n)$ denotes the number of visits for the n-th patient, $\mathbf{h}_i$ denotes the hidden state of the i-th visit generated by the RNN, α_i corresponds to the attention weight associated with $\mathbf{h}_i$, and $\mathbf{c}^n$ denotes the context vector for the n-th patient. Concatenating the context vector with the hidden state at the last time stamp, we can arrive at the final patient representation. Following this scheme, the clinical doctors will be able to identify the visits that contribute the most to the prediction by inspecting the attention weights.

As the attention mechanism has demonstrated its potential in providing interpretability and RNN models need to process elements in the sequence one by one, recent works began to follow the Transformer architecture [44, 28, 34, 40], which means no recurrence in modeling the sequence of visits [46], so that the computation can be significantly accelerated with distributed computing schemes [34].

5.2 Transformers

SAnD [44] is an early work modeling sequential EHR solely based on attention mechanisms: it leverages a self-attention mechanism coupled with a dense interpolation to enable sequence modeling. Several works following SAnD also chose to employ the self-attention mechanism to aggregate visit information [28, 34, 40].

Although both RNNs and Transformer can model the sequential characteristics of EHR data, they do not take the sequence of time stamps as input, which means they cannot capture the irregular time intervals. However, irregular time intervals between consecutive visits are important information that needs to be incorporated into the modeling process because they can indicate certain health status of a patient. For example, a patient will probably visit the hospital more frequently if he/she is having a severe health problem; on the contrary, if a patient has not been visiting the hospital for a long period of time, then it implies that he/she is in a good shape, without any abnormalities. Thus, many existing works injected time interval information into sequence models in various ways.

5.3 Time-Aware Sequence Models

Some works having RNNs as backbones incorporated the time interval information via different ways. T-LSTM [3] converts the time intervals into weights discounting the cell memory of LSTM through a decay function. ATTAIN [58] adjusts the mem-

ory of LSTM to retrospect memories of all previous events and discount them by weights generated from attention mechanism and time intervals. StageNet [17] takes the concatenation of time interval information and the visit embedding as the input to LSTM. DG-RNN [51] encodes time interval as a vector, taking the encoded time and visit embedding as the input to LSTM.

There are also various strategies for models based on Transformer to include time interval information. Specifically, HiTANet [28] and RAPT [40] embed time information into the visit representation which is the input to sequence models. ConCare [32] includes time interval as a factor in calculating attention weights for hospital visits.

Chapter 6

Framework-Level Design

In order to overcome certain challenges, some works have been "thinking outside of the box," not focusing on the design of $f(\cdot)$, $g(\cdot)$, or $h(\cdot)$ but making modifications to loss functions or optimization procedures.

Incorporating prior medical knowledge into prediction can be challenging, because prior medical knowledge usually takes the form of discrete arbitrary rules which are difficult to be converted to continuous values. To overcome this challenge, PRIME [30] uses a log-linear model to estimate the desired distribution of diseases according to the given medical knowledge, and applies posterior regularization.

To capture the phenomenon that the diagnosis codes in EHR data all correspond to the leaf nodes of ICD-9 hierarchy, MHM [38] makes predictions and calculates loss at each level of the ICD-9 ontology, resulting in a layer-wise loss function.

In order to provide model trustworthiness, which means offering an uncertainty level of the prediction made by the model, INPREM [57] and UNITE [5] incorporates a variational inference loss.

Because of the high rate of human errors and missing diagnoses in EHR data, PacRNN [37] formulates the disease progression modeling task as an events recommendation problem, utilizing a pairwise ranking method regularized by disease co-occurrence to

rank probabilities of potential diagnoses.

It is common that some diagnoses are extremely rare in EHR data. Thus, TC-EMNet [60] provides an idea that the model can be trained with the loss (MSE) of reconstructing the observation from the learnt representation when the labels associated with the target disease risk prediction task are not sufficient.

For methods incorporating MKGs, one problem is that MKGs are noisy and suffer from the issue of missing edges. To tackle this problem, MendMKG [49] pre-trains the graph attention network [47] applied to the MKG in a self-supervised manner: first learn the node embeddings based on the observed edges in the MKG, then use the learnt embeddings to predict missing edges, and finally the edges assigned with a high importance score by the unlabeled EHR data reconstruction task are added to the MKG. This iterative mutual enhancement between MKG and EHR data guides the completion of MKG, which in turn results in better learnt patient representations. Similarly, RAPT [40] pre-trains a time-aware Transformer model via three tasks—similarity prediction, masked prediction, and reasonability check—in order to overcome data insufficiency. With a similar purpose, MetaPred [55] incorporates a meta-learning framework in which the model parameters are first adjusted on the training data from the source domains and then further tuned on the simulated target domain.

Chapter 7

Conclusion

This work summarized the existing challenges in Electronic Health Record-based predictive healthcare, and provided the mathematical formulation of a general architecture for leveraging deep learning models to learn patient representations for diagnosis prediction. Motivated by the specified challenges and guided by the proposed general architecture, this work further discussed how existing works had devised different strategies from four levels: diagnosis-level, visit-level, sequence-level, and framework-level. Through these discussions, this paper could serve as a cookbook for novices interested in applying deep learning to EHR-based diagnosis prediction. Interestingly, the proposed general recipe in this work is not limited to the task of diagnosis prediction; instead, any tasks in predictive healthcare for which learning a high-quality patient representation is desired can gain insights from this recipe.

Chapter 8

Future Work

This paper also serves as a foundation for a future work. Currently this future work is at an embryonic stage and the basic ideas are introduced in this chapter from four levels.

8.1 Diagnosis-Level Representation Learning

For learning the representations of diagnosis codes, we plan to jointly utilize the disease co-occurrence graph and ICD-9 ontology. Specifically, motivated by existing works in learning graph embeddings with the guidance of a hierarchical structure [59, 50, 15, 39], we plan to utilize a GNN-based model that can learn node embeddings on the disease co-occurrence graph and simultaneously preserve the ICD-9 hierarchy in the learnt embeddings.

8.2 Visit-Level Representation Learning

Motivated by how existing works leveraged counterfactual reasoning and information bottleneck for learning robust representations in the domain of recommender systems [48], we plan to obtain counterfactuals by randomly dropping out diagnoses in a

hospital visit and facilitate balanced learning between the factual and counterfactual domains. In this way, the learnt embeddings can be more robust against the noise in EHR data.

8.3 Sequence-Level Representation Learning

Motivated by the recent advance of neural ordinary differential equations and relevant success in time series modeling [6, 14, 41, 22, 7], we plan to leverage neural ODEs with deep learning models to aggregate the sequence of hospital visit embeddings. In this way, the issues including irregular time intervals and sporadical observations can be properly handled.

8.4 Framework-Level Design

Firstly, inspired by [54] and the message passing characteristics of graph neural networks (GNNs) [19, 24], we plan to incorporate a message passing module that can take into account similar patients in the training data after obtaining patient representations following the proposed general architecture in Chapter 2. In this way, the model can mimic real doctors who depend on past experience with similar patients to attend a new patient.

Since some recent works in recommender systems have demonstrated that weights generated by attention mechanisms are not suitable for providing interpretability [45], and counterfactual explanations have great potential in opening a black-box deep learning model [18, 2, 45], we plan to device an external counterfactual explainer to provide interpretations for the predictions made by our model.

Bibliography

[1] Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *KDD*, 2018.

[2] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In *NeurIPS*, 2021.

[3] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *KDD*, 2017.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.

[5] Chacha Chen, Junjie Liang, Fenglong Ma, Lucas Glass, Jimeng Sun, and Cao Xiao. Unite: Uncertainty-based health risk prediction leveraging multi-sourced data. In *WWW*, 2021.

[6] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

[7] Ricky TQ Chen, Brandon Amos, and Maximilian Nickel. Learning neural event functions for ordinary differential equations. In *ICLR*, 2021.

[8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[9] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, 2016.

[10] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*, 2016.

[11] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *KDD*, 2017.

[12] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *NeurIPS*, 2018.

[13] Dan dan Guo, Long Tian, Minghe Zhang, Mingyuan Zhou, and Hongyuan Zha. Learning prototype-oriented set representations for meta-learning. In *ICLR*, 2022.

[14] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *NeurIPS*, 2019.

[15] Bahare Fatemi, Siamak Ravanbakhsh, and David Poole. Improved knowledge graph embedding using background taxonomic information. In *AAAI*, 2019.

[16] Jingyue Gao, Xiting Wang, Yasha Wang, Zhao Yang, Junyi Gao, Jiangtao Wang, Wen Tang, and Xing Xie. Camp: Co-attention memory networks for diagnosis prediction in healthcare. In *ICDM*, 2019.

[17] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In *WWW*, 2020.

[18] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. Prince: Provider-side interpretability with counterfactual explanations in recommender systems. In *WSDM*, 2020.

[19] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, 2017.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–1780, 1997.

[21] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:1–9, 2016.

[22] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *NeurIPS*, 2020.

[23] Minyoung Kim. Differentiable expectation-maximization for set representation learning. In *ICLR*, 2022.

[24] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[25] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019.

[26] Sicen Liu, Tao Li, Haoyang Ding, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Jun Yan, and Yi Zhou. A hybrid method of recurrent neural network and graph neural network for next-period prescription prediction. *International Journal of Machine Learning and Cybernetics*, 11:2849–2856, 2020.

[27] Chang Lu, Tian Han, and Yue Ning. Context-aware health event prediction via transition functions on dynamic disease graphs. In *AAAI*, 2022.

[28] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *KDD*, 2020.

[29] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *KDD*, 2017.

[30] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. Risk prediction on electronic health records with prior medical knowledge. In *KDD*, 2018.

[31] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *CIKM*, 2018.

[32] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *AAAI*, 2020.

[33] Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen. Setrank: Learning a permutation-invariant ranking model for information retrieval. In *SIGIR*, 2020.

[34] Xueping Peng, Guodong Long, Tao Shen, Sen Wang, Jing Jiang, and Chengqi Zhang. Bitenet: Bidirectional temporal encoder network to predict medical outcomes. In *ICDM*, 2020.

[35] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5:1–13, 2018.

[36] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

[37] Zhi Qiao, Shiwan Zhao, Cao Xiao, Xiang Li, Yong Qin, and Fei Wang. Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction. In *IJCAI*, 2018.

[38] Zhi Qiao, Zhen Zhang, Xian Wu, Shen Ge, and Wei Fan. Mhm: Multi-modal clinical data based hierarchical multi-label diagnosis prediction. In *SIGIR*, 2020.

[39] Ziyue Qiao, Pengyang Wang, Yanjie Fu, Yi Du, Pengfei Wang, and Yuanchun Zhou. Tree structure-aware graph representation learning via integrated hierarchical aggregation and relational metric learning. In *ICDM*, 2020.

[40] Houxing Ren, Jingyuan Wang, Wayne Xin Zhao, and Ning Wu. Rapt: Pre-training of time-aware transformer for learning robust healthcare representation. In *KDD*, 2021.

[41] Yulia Rubanova, Ricky TQ Chen, and David Duvenaud. Latent odes for irregularly-sampled time series. In *NeurIPS*, 2019.

[42] Yifeng Shi, Junier Oliva, and Marc Niethammer. Deep message passing on sets. In *AAAI*, 2020.

[43] Konstantinos Skianis, Giannis Nikolentzos, Stratis Limnios, and Michalis Vazirgiannis. Rep the set: Neural networks for learning set representations. In *International conference on artificial intelligence and statistics*, 2020.

[44] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *AAAI*, 2018.

[45] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. Counterfactual explainable recommendation. In *CIKM*, 2021.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[47] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[48] Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan Kuruoglu, and Yefeng Zheng. Information theoretic counterfactual learning from missing-not-at-random feedback. In *NeurIPS*, 2020.

[49] Xiao Xu, Xian Xu, Yuyao Sun, Xiaoshuang Liu, Xiang Li, Guotong Xie, and Fei Wang. Predictive modeling of clinical events with mutual enhancement between longitudinal patient records and medical knowledge graph. In *ICDM*, 2021.

[50] Carl Yang, Jieyu Zhang, and Jiawei Han. Co-embedding network nodes and hierarchical labels with taxonomy based generative adversarial networks. In *ICDM*, 2020.

[51] Changchang Yin, Rongjian Zhao, Buyue Qian, Xin Lv, and Ping Zhang. Domain knowledge guided deep learning with electronic health records. In *ICDM*, 2019.

[52] Quan Yuan, Jun Chen, Chao Lu, and Haifeng Huang. The graph-based mutual attentive network for automatic diagnosis. In *IJCAI*, 2021.

[53] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *NIPS*, 2017.

[54] Daochen Zha, Kwei-Herng Lai, Kaixiong Zhou, and Xia Hu. Towards similarity-aware time-series classification. In *SDM*, 2022.

[55] Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *KDD*, 2019.

[56] Xianli Zhang, Buyue Qian, Yang Li, Changchang Yin, Xudong Wang, and Qinghua Zheng. Knowrisk: an interpretable knowledge-guided model for disease risk prediction. In *ICDM*, 2019.

[57] Xianli Zhang, Buyue Qian, Shilei Cao, Yang Li, Hang Chen, Yefeng Zheng, and Ian Davidson. Inprem: An interpretable and trustworthy predictive model for healthcare. In *KDD*, 2020.

[58] Yuan Zhang. Attain: Attention-based time-aware lstm networks for disease progression modeling. In *IJCAI*, 2019.

[59] Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *AAAI*, 2020.

[60] Zicong Zhang, Changchang Yin, and Ping Zhang. Temporal clustering with external memory network for disease progression modeling. In *ICDM*, 2021.