KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare

Fenglong Ma
SUNY at Buffalo
Buffalo, NY, USA
fenglong@buffalo.edu

Quanzeng You
Microsoft AI & Research
Redmond, WA, USA
quanzeng.you@microsoft.com

Houping Xiao
Georgia State University
Atlanta, GA, USA
hxiao@gsu.edu

Radha Chitta
Kira Systems
Toronto, ON, Canada
radha.cr@gmail.com

Jing Zhou
eHealth Inc.
Mountain View, CA, USA
jing.zhou@ehealth.com

Jing Gao
SUNY at Buffalo
Buffalo, NY, USA
jing@buffalo.edu

ABSTRACT

The goal of diagnosis prediction task is to predict the future health information of patients from their historical Electronic Healthcare Records (EHR). The most important and challenging problem of diagnosis prediction is to design an accurate, robust and interpretable predictive model. Existing work solves this problem by employing recurrent neural networks (RNNs) with attention mechanisms, but these approaches suffer from the data sufficiency problem. To obtain good performance with insufficient data, graph-based attention models are proposed. However, when the training data are sufficient, they do not offer any improvement in performance compared with ordinary attention-based models. To address these issues, we propose KAME, an end-to-end, accurate and robust model for predicting patients' future health information. KAME not only learns reasonable embeddings for nodes in the knowledge graph, but also exploits general knowledge to improve the prediction accuracy with the proposed knowledge attention mechanism. With the learned attention weights, KAME allows us to interpret the importance of each piece of knowledge in the graph. Experimental results on three real world datasets show that the proposed KAME significantly improves the prediction performance compared with the state-of-the-art approaches, guarantees the robustness with both sufficient and insufficient data, and learns interpretable disease representations.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Applied computing** → **Health informatics**;

KEYWORDS

Healthcare informatics; medical knowledge graph; knowledge attention mechanism

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271701

ACM Reference Format:
Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271701

1 INTRODUCTION

Achieving precision medicine and improving care for individual patients are the overall objective for healthcare providers and researchers. Mining the massive and diverse Electronic Healthcare Records (EHR) provides the possibility to accomplish this goal, which attracts considerable attention. In particular, predicting the future diagnoses based on patient's historical sequential EHR data, i.e., *diagnosis prediction*, has been an intriguing yet challenging topic. The main challenge of diagnosis prediction task comes from the temporal, high dimensional and noisy EHR data. As a result, robust predictive models are necessary to achieve accurate predictions.

Recently, deep learning techniques have been adopted for diagnosis prediction tasks [9–11, 23, 33]. Med2Vec [9] generates low-dimensional representations of medical codes (i.e., diagnosis codes, procedure codes, and medication codes), but doses not consider the temporal nature of EHR data. To model the sequential relations among medical codes, state-of-the-art approaches have broadly applied recurrent neural networks (RNNs) [10, 11, 23, 33]. RETAIN [11] applies an RNN with reverse time ordered EHR sequences, which can reasonably interpret the contribution of each medical code appeared in the previous visits for the current prediction. Dipole [23] employs bidirectional recurrent neural networks (BRNNs) with different attention mechanisms, which significantly improves the predictive performance. However, training the aforementioned models with a high accuracy typically requires large amounts of data. In addition, some medical codes of rare diseases may infrequently appear in the EHR data. A more challenging task is how to train a robust prediction model with these rare codes.

To solve this challenge, GRAM [10] exploits medical ontologies and graph-based attention mechanism to learn robust medical code representations. GRAM can alleviate the difficulties of learning embeddings for rare medical codes with their ancestors to guarantee the predictive performance when there are not enough EHR data

to train deep learning models. However, *when sufficient training data are available*, each medical code can learn a satisfactory vector representation only from EHR data. In such a case, GRAM has relatively comparable performance with other RNN variants such as Dipole. Therefore, designing a robust predictive model is essential for diagnosis prediction task.

Furthermore, GRAM uses the hierarchy information for learning the representations of medical codes, then employs these embeddings to learn the representations of visits, and finally makes predictions with visit representations. In the whole process, medical ontology information is only used when learning code representations, which *implicitly* affects the final predictions. We believe that **directly exploiting medical knowledge in the whole prediction process** (i.e., *learning code representations, generating visit embeddings and making predictions*), should help the predictive models to improve the accuracy and provide better interpretation.

To tackle all the aforementioned challenges and problems, in this paper, we propose a novel, accurate and robust **k**nowledge-based **a**ttention **m**ode**l** (KAME) for predicting patients' future diagnoses, which exploits medical knowledge in the whole prediction process. Specifically, KAME first uses a given medical ontology (i.e., **knowledge graph**), such as Clinical Classifications Software (CCS)[1] or the International Classification of Diseases (ICD)[2], to learn the representations of medical codes and obtain the embeddings of medical codes' ancestors. Next, the learned medical code representations are used to embed each input visit into a low dimensional visit-level vector, and then it is fed into an RNN to generate the hidden state representation. The hidden state representation is used to calculate **knowledge attention** weights with the transformed ancestor embeddings in the knowledge graph. Here the embeddings of ancestors contain the general information of medical codes, i.e., *high-level knowledge* of the medical graph. KAME then generates a new **knowledge vector** from the relevant high-level knowledge weighted by the corresponding knowledge attention weights. The combination of the hidden state at time t and the computed knowledge vector is fed into a softmax layer to predict patient's diagnoses at time $t+1$.

We experimentally demonstrate that the proposed KAME achieves significantly higher prediction accuracy compared with the state-of-the-art approaches in diagnosis prediction, using three real world medical datasets. We then quantitatively analyze the effectiveness of the proposed KAME with sufficient and insufficient data respectively. Moreover, a case study is conducted to illustrate the interpretability and reasonableness of the designed knowledge attention mechanism in predicting patient future diagnoses. Finally, qualitative analysis demonstrates that KAME learns interpretable representations of medical codes. In summary, our main contributions are as follows:

- We propose KAME, an end-to-end, accurate and robust model to accurately predict patients' future visit information with medical ontologies, which explicitly makes use of medical knowledge in the whole prediction process.
- We design a novel knowledge-level attention mechanism, which significantly helps the proposed KAME to improve the predictive performance.
- We empirically show that the proposed KAME has strong robustness and outperforms existing methods in diagnosis prediction on three real world datasets.
- We qualitatively demonstrate the interpretability of the learned representations of medical codes and qualitatively validate the reasonableness of the designed knowledge attention mechanism.

The rest of this paper is organized as follows: We first introduce the details of the proposed KAME in Section 2. In Section 3, experiments are conducted to validate the effectiveness of the proposed KAME. We then summarize the related literatures in Section 4. Finally, conclusions of this work are presented in Section 5.

2 METHODOLOGY

In this section, we first introduce the structure of EHR data and medical ontology, and then define some notations. Finally, we describe the details of the proposed knowledge-based attention model KAME.

2.1 Basic Notations

We denote the set of medical codes from the EHR data as $c_1, c_2, \cdots, c_{|C|} \in C$, and $|C|$ is the number of unique medical codes. P denotes the number of patients in the EHR data. For the p-th patient who has $T^{(p)}$ visit records, his/her clinical records can be represented by a sequence of visits $V_1, V_2, \cdots, V_{T^{(p)}}$. Each visit V_t contains a subset of medical codes ($V_t \subseteq C$), and is denoted by a binary vector $\mathbf{x}_t \in \{0,1\}^{|C|}$, where the i-th element is 1 if V_t contains the medical code c_i. For simplicity, we drop the superscript (p) when it is unambiguous.

A medical ontology $\mathcal{G}$ contains the hierarchy of various medical concepts with the *parent-child* relationship, which is a directed acyclic graph (DAG) and referred to as *knowledge graph* in this paper. The nodes of $\mathcal{G}$ include leaves and their ancestors. Each leaf node is a medical code in C, and each ancestor node belongs to the set $\mathcal{N} = \{n_1, n_2, \cdots, n_{|\mathcal{N}|}\}$, where $|\mathcal{N}|$ is the number of ancestor codes in $\mathcal{G}$. The ancestors of the leaf node c_i are represented by $q(c_i)$, which consists of all the intermediate nodes of the path from root of $\mathcal{G}$ to leaf c_i. For each visit $\mathbf{x}_t$, it contains multiple medical codes, and Q_t denotes the union of $q(c_i)$ for each of the medical code c_i in $\mathbf{x}_t$. Similar to V_t, Q_t can also be represented by a binary vector $\mathbf{f}_t \in \{0,1\}^{|\mathcal{N}|}$, where the j-th element is 1 if Q_t contains the ancestor code n_j.

With the above notations, the inputs of the proposed KAME model are a medical knowledge graph $\mathcal{G}$, a time-ordered sequence of each patient visits $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{T-1}$, and a time-ordered sequence of medical code ancestors in patient visits $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_{T-1}$. For the t-th visit, we aim to predict the next visit information. Thus, the outputs are $\mathbf{x}_2, \mathbf{x}_3, \cdots, \mathbf{x}_T$.

2.2 The Proposed Model KAME

Figure 1 shows the overview of the proposed KAME. Using the given knowledge graph $\mathcal{G}$, we can obtain the embedding matrix $\mathbf{M}$ of medical codes and the matrix $\mathbf{A}$ of ancestor code embeddings

[1] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
[2] http://www.icd9data.com/

with graph-based attention mechanism [10]. Given the t-th visit information of a patient $\mathbf{x}_t$, it is embedded into a vector representation $\mathbf{v}_t$ with the learned $\mathbf{M}$. The embedded vector $\mathbf{v}_t$ is fed into a Recurrent Neural Network (RNN), which produces a hidden state $\mathbf{h}_t$ as the representation of the t-th visit. With the corresponding ancestor set $\mathbf{f}_t$ for the t-th visit, the learned ancestor embedding matrix $\mathbf{A}$ can be mapped into a new matrix, called latent knowledge embeddings denoted by $\mathbf{L}^t$ via a function θ. Along with $\mathbf{h}_t$ and $\mathbf{L}^t$, we are able to generate a knowledge vector $\mathbf{k}_t$ using a knowledge-based attention mechanism, which will be detailed in the following sections. From the hidden state $\mathbf{h}_t$ and the knowledge vector $\mathbf{k}_t$, a knowledge attentional vector $\mathbf{s}_t$ can be obtained, which is used to predict the information of the $(t+1)$-th visit, i.e., $\hat{\mathbf{y}}_t$. It is obvious that the proposed model can be trained end-to-end.

Knowledge Graph Embedding

In order to learn reasonable and correct representations of medical codes, we employ the state-of-the-art graph embedding approach GRAM [10]. Through balancing the ontology information in relation to the data volume, GRAM can learn the robust representations even when the data volume is constrained.

In the knowledge graph $\mathcal{G}$, each medical code or leaf node c_i has a basic learnable embedding vector $\mathbf{e}_i$ $(1 \le i \le |C|)$, and each ancestor code n_j also has an embedding vector $\mathbf{a}_j$ $(1 \le j \le |\mathcal{N}|)$. The final embedding vector of the i-th medical code denoted as $\mathbf{m}_i$ can be obtained by combining the basic embedding $\mathbf{e}_j$ and its ancestors via graph-based attention mechanism. The details can be found in [10].

By concatenating the vector representation $\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_{|C|}$ of all the medical codes, GRAM generates the embedding matrix $\mathbf{M} \in \mathbb{R}^{d \times |C|}$, where d is the dimensionality size and $\mathbf{m}_i$ is the i-th column of $\mathbf{M}$. GRAM only uses the medical code embeddings $\mathbf{M}$ in the final prediction and ignores the ancestor code information. Actually, the ancestor codes contain general or coarse-grained information about the medical codes, which may help the predictive model to improve the prediction performance. Thus, the proposed KAME not only generates the medical code embeddings $\mathbf{M} \in \mathbb{R}^{d \times |C|}$, but also the ancestor code embeddings $\mathbf{A} \in \mathbb{R}^{d \times |\mathcal{N}|}$, where each ancestor code embedding vector $\mathbf{a}_i$ is the i-th column of $\mathbf{A}$. The ancestor code embeddings $\mathbf{A}$ will be used in the knowledge attention layer as shown in Figure 1.

Visit Embedding

Given the t-th visit information $\mathbf{x}_t \in \{0,1\}^{|C|}$, the vector representation $\mathbf{v}_t \in \mathbb{R}^d$ is obtained by multiplying medical code embeddings $\mathbf{M}$ with one-hot vector $\mathbf{x}_t$ as follows:

$$\mathbf{v}_t = \tanh(\mathbf{M}\mathbf{x}_t). \tag{1}$$

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) provide a very efficient and elegant way of modeling sequential healthcare data [10, 11, 23, 33]. Note that we use "RNNs" to denote any Recurrent Neural Network variants, such as Long-Short Term Memory (LSTM) [16] and Gated Recurrent Unit (GRU) [8]. In our implementation, we use GRU to adaptively capture dependencies among patient visit information. A GRU has two gates, a reset gate r and an update gate z. The reset gate r determines the combination of the new input and the previous memory, which allows the hidden layer to drop irrelevant information. The update gate z controls how much information should be kept around from the previous hidden state. Accordingly, the mathematical formulation of GRU can be described as follows:

$$\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_z\mathbf{v}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z),\\
\mathbf{r}_t &= \sigma(\mathbf{W}_r\mathbf{v}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r),\\
\tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h\mathbf{v}_t + \mathbf{r}_t \circ \mathbf{U}_h\mathbf{h}_{t-1} + \mathbf{b}_h),\\
\mathbf{h}_t &= \mathbf{z}_t \circ \mathbf{h}_{t-1} + (\mathbf{1} - \mathbf{z}_t) \circ \tilde{\mathbf{h}}_t.
\end{aligned} \tag{2}$$

In these equations, $\circ$ denotes the element-wise multiplication, $\sigma()$ is the activation function, $\mathbf{z}_t \in \mathbb{R}^g$ is the update gate at time t, $\mathbf{r}_t \in \mathbb{R}^g$ is the reset gate at time t, $\tilde{\mathbf{h}}_t \in \mathbb{R}^g$ represents the intermediate memory, $\mathbf{h}_t \in \mathbb{R}^g$ is the hidden state, and g is the dimensionality of hidden states. Matrices $\mathbf{W}_z \in \mathbb{R}^{g \times d}$, $\mathbf{W}_r \in \mathbb{R}^{g \times d}$, $\mathbf{W}_h \in \mathbb{R}^{g \times d}$, $\mathbf{U}_z \in \mathbb{R}^{g \times g}$, $\mathbf{U}_r \in \mathbb{R}^{g \times g}$, $\mathbf{U}_h \in \mathbb{R}^{g \times g}$ and vectors $\mathbf{b}_z \in \mathbb{R}^g$, $\mathbf{b}_r \in \mathbb{R}^g$, $\mathbf{b}_h \in \mathbb{R}^g$ are parameters to be learned.

Knowledge-based Attention Mechanism

The benefit of employing the medical knowledge graph $\mathcal{G}$ is not only to learn the robust vector representations of medical codes, but also learn the coarse-grained information of ancestor codes. Correct vector representations of medical codes can help RNN to generate the accurate vector representation for next visit, i.e., the hidden state $\mathbf{h}_t$. Moreover, the embeddings of ancestor codes $\mathbf{A}$ contain the relevant high-level medical code information, which provides additional features for the learning model. With $\mathbf{h}_t$ and $\mathbf{A}$, it is expected that the predictive model can improve its performance on the task of future diagnosis prediction.

Now, we describe the details of computing the knowledge attention representations. We first map the ancestor embeddings $\mathbf{A}$ to space $\mathbf{L}^t \in \mathbb{R}^{g \times |\mathcal{N}|}$ as follows:

$$\mathbf{L}^t_n = \mathbf{f}_{tn}(\mathbf{W}_k\mathbf{A}_n + \mathbf{b}_k), \tag{3}$$

where $\mathbf{L}^t_n \in \mathbb{R}^g$ is the n-th column of $\mathbf{L}^t$, $\mathbf{f}_{tn}$ is the n-th element of the one-hot ancestor vector $\mathbf{f}_t$, $\mathbf{W}_k \in \mathbb{R}^{g \times d}$ and $\mathbf{b}_k \in \mathbb{R}^g$ are parameters to be learned. In such a way, $\mathbf{L}^t$ encodes the relevant high-level knowledge of the previous visit.

Next, we compute the knowledge vector $\mathbf{k}_t$ by combining $\mathbf{L}^t$ and $\mathbf{h}_t$. In particular, we propose a knowledge based attention mechanism to compute $\mathbf{k}_t$ as follows:

$$\begin{aligned}
\mathbf{k}_t &= \sum_{n=1}^{|\mathcal{N}|} \alpha_{tn}\mathbf{L}^t_n,\\
\text{s.t.} \quad & \alpha_{tn} \ge 0,\ n = 1, \cdots, |\mathcal{N}|.
\end{aligned} \tag{4}$$

where α_{tn} is the attention weight on the embedding $\mathbf{L}^t_n$ when calculating $\mathbf{k}_t$. The attention weight in Eq. (4) is calculated by the following Softmax function,

$$\alpha_{tn} = \frac{\exp(\mathbf{h}_t^\top \mathbf{L}^t_n)}{\sum_{j=1}^{|\mathcal{N}|} \exp(\mathbf{h}_t^\top \mathbf{L}^t_j)}. \tag{5}$$

Knowledge-based Diagnosis Prediction

Given the knowledge vector $\mathbf{k}_t$ and the current hidden state $\mathbf{h}_t$, we employ a simple concatenation layer to combine the information

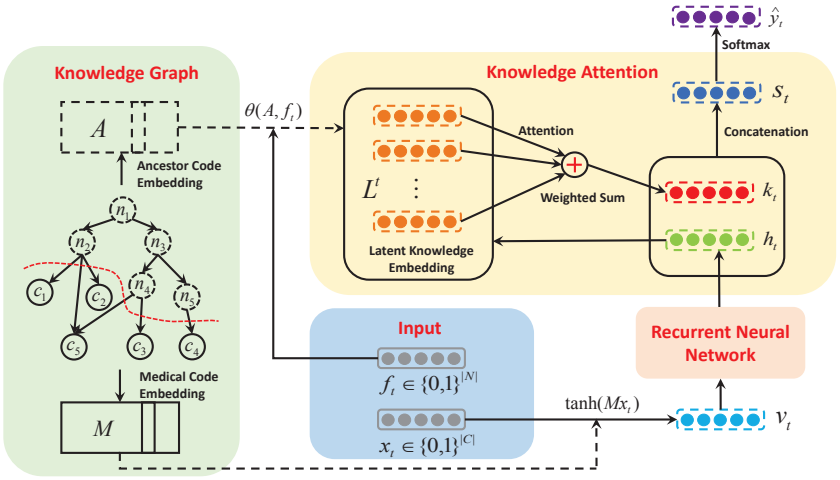


Figure 1: The Proposed KAME Model.

from both vectors to generate a knowledge attentional vector $\mathbf{s}_t \in \mathbb{R}^{2g}$ as follows:

$$\mathbf{s}_t = [\mathbf{h}_t; \mathbf{k}_t]. \tag{6}$$

Therefore, $\mathbf{s}_t$ contains both information from previous visits and the relevant high-level knowledge from $\mathcal{G}$. $\mathbf{s}_t$ is fed through the softmax layer to produce the $(t+1)$-th visit information defined as:

$$\hat{\mathbf{y}}_t = \text{Softmax}(\mathbf{W}_c \mathbf{s}_t + \mathbf{b}_c), \tag{7}$$

where $\mathbf{W}_c \in \mathbb{R}^{|C| \times 2g}$ and $\mathbf{b}_c \in \mathbb{R}^{|C|}$ are the learnable parameters.

Objective Function

Based on Eq. (7), we use the cross-entropy between the ground truth visit $\mathbf{y}_t$ and the predicted visit $\hat{\mathbf{y}}_t$ to calculate the loss for each patient from all the timestamps as follows:

$$\begin{aligned} &\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T; \mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_T) \\ &= -\frac{1}{T-1} \sum_{t=1}^{T-1} \left(\mathbf{y}_t^\top \log(\hat{\mathbf{y}}_t) + (\mathbf{1} - \mathbf{y}_t)^\top \log(\mathbf{1} - \hat{\mathbf{y}}_t) \right). \end{aligned} \tag{8}$$

Note that in our implementation, we take the average of the individual cross entropy error for multiple patients. Algorithm 1 describes the overall training procedure of the proposed KAME.

Remark. *The proposed KAME is the generalization of the state-of-the-art diagnosis prediction model GRAM [10]. When removing the proposed knowledge-based attention component (i.e., deleting $\mathbf{k}_t$), then the proposed KAME is reduced to GRAM.*

Algorithm 1 KAME Optimization Algorithm.

Randomly initialize basic embedding matrix of medical codes $\mathbf{E} = \{\mathbf{e}_i\}_{i=1}^{|C|}$, embedding matrix of ancestor codes $\mathbf{A}$, attention parameter used in GRAM Δ, RNN parameter Ω, latent knowledge embedding parameters $\mathbf{W}_k$ and $\mathbf{b}_k$, softmax parameters $\mathbf{W}_c$ and $\mathbf{b}_c$;
repeat
  $\mathbf{X} \leftarrow$ random patient from dataset
  for visit V_t in $\mathbf{X}$ **do**
    for medical code c_i in V_t **do**
      Refer $\mathcal{G}$ to find c_i's ancestors $q(c_i)$;
      Update Q_t according to $q(c_i)$;
      Obtain the medical code representation $\mathbf{m}_i$;
    end for
    Obtain the ancestor code representations $\mathbf{A}$;
    Calculate the visit embedding $\mathbf{v}_t$ according to Eq. (1);
    Compute the hidden state $\mathbf{h}_t$ according to Eq. (2);
    Calculate the knowledge vector $\mathbf{k}_t$ according to Eq. (3) and (4);
    Obtain the knowledge attentional vector $\mathbf{s}_t$ according to Eq. (6);
    Make prediction $\hat{\mathbf{y}}_t$ using Eq. (7);
  end for
  Calculate the prediction loss $\mathcal{L}$ using Eq. (8);
  Update parameters according to the gradient of $\mathcal{L}$;
until convergence

3 EXPERIMENTS

In this section, we conduct experiments on three real world medical claim datasets to evaluate the performance of the proposed KAME. Compared with the state-of-the-art predictive models, KAME yields better performance on different evaluation strategies.

3.1 Data Description

The real world datasets used in this experiments are the Medicaid dataset, the Diabetes dataset and the MIMIC-III dataset.

Medicaid Dataset

The Medicaid dataset consists of insurance claims over the years 2011 and 2012, which has 99, 159 patients and 2, 034, 485 visits. The patient visits were grouped by week [23], and we chose patients who made at least ten visits.

Diabetes Dataset

The Diabetes dataset is a subset of the Medicaid dataset, corresponding to patients who have been diagnosed with diabetes (i.e., Medicaid members who have the ICD9 diagnosis code 250.xx in their claims). There are 17, 584 patients with 466, 024 visits.

MIMIC-III Dataset
The MIMIC-III dataset is a publicly available EHR dataset, which consists of medical records of 7,499 intensive care unit (ICU) patients over 11 years. For the MIMIC-III dataset, we chose the patients who made at least two visits.

We choose these three representative datasets to extensively evaluate different aspects of the models: (1) The number of patients and visits in the Medicaid dataset is big enough to validate the performance of the proposed KAME with long visit records. (2) The MIMIC-III dataset consists of very short visits, and the number of patients is small. With this dataset, we can validate the performance of KAME with insufficient training data. (3) The number of patients and visits in the Diabetes dataset is smaller than that of the Medicaid dataset and bigger than that of the MIMIC-III dataset. This dataset is used to validate the performance of all the state-of-the-art diagnosis prediction approaches on a specific disease. With these three different types of datasets, we can fully and correctly validate the performance of all the diagnosis prediction approaches.

The goal of diagnosis prediction task is to predict the diagnosis information of the next visit. In the experiments, we aim to predict diagnosis categories instead of the real diagnosis codes. Predicting category information not only improves the training speed and predictive performance, but also guarantees the sufficient granularity of all the diagnoses [10, 23]. We use the nodes in the second hierarchy of the ICD9 codes[3] as the category labels, such as the category label of diagnosis code "250.1: Diabetes with ketoacidosis" is "Diseases of other endocrine glands (249-259)". Actually, the hierarchy of CCS[4] can also be used as category labels [10]. These two kinds of grouping methods can obtain similar predictive performance. Table 1 lists more details about the three datasets.

Table 1: Statistics of the Medicaid Dataset, the Diabetes Dataset and the MIMIC-III Dataset.

Dataset	Medicaid	Diabetes	MIMIC-III
# of patients	99,159	17,584	7,499
# of visits	2,034,485	466,024	19,911
Avg. # of visits per patient	20.52	26.50	2.66
# of unique ICD9 codes	9,701	7,437	4,880
Avg. # of ICD9 codes per visit	2.78	3.39	13.06
Max # of ICD9 codes per visit	41	37	39
# of category codes	157	155	171
Avg. # of category codes per visit	2.30	2.92	10.16
Max # of category codes per visit	23	22	30

3.2 Experimental Setup

In this subsection, we first introduce the state-of-the-art approaches for diagnosis prediction task in healthcare, and then outline the measures used for predictive performance evaluation. Finally, we describe the implementation details.

Baseline Approaches
To validate the predictive performance of the proposed approach KAME, we compare it with the following four state-of-the-art approaches:

GRAM [10]. GRAM is the first work that uses a medical knowledge graph to learn the medical code representations and predict the future visit information with recurrent neural networks. A time-ordered visit sequence $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ is first transformed into visit vectors by the medical code embedding matrix $\mathbf{M}$, and then visit vectors are fed to the GRU with a single hidden layer, which in turn predict the future visit information.

Dipole [23]. Dipole uses bidirectional recurrent neural networks and three attention mechanisms to predict patient visit information, which can achieve the best performance compared with other diagnosis prediction models. In the experiments, the attention mechanism we selected is the local-based one. A time-ordered visit sequence $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ is first embedded into visit vectors by a multilayer perceptron (MLP) with the rectified linear unit (ReLU), and then visit vectors are fed to the bidirectional GRUs. Finally, the concatenated outputs from GRUs with attention mechanism are used to generate latent vectors to make the predictions with a single softmax layer.

RNN+. RNN+ adds location-based attention model into RNN [23]. The difference between RNN+ and Dipole is that RNN+ only uses one directional GRU to make the prediction.

RNN. We directly embed visit information $\mathbf{x}_t$ into the vector representation $\mathbf{v}_t$, and then feed this embedding to the GRU. The hidden state $\mathbf{h}_t$ produced by the GRU is used to predict the $(t+1)$-th visit information.

Note that Med2Vec [9] and RETAIN [11] are not listed as baselines in the following experiments because the performance of these two approaches is worse than that of Dipole [23]. Med2Vec focuses on the learning of medical code representations, and RETAIN aims to interpret the prediction results with a two-level attention model.

Evaluation Measures
We evaluate the performance for all the diagnosis predication approaches from two aspects: visit-level and code-level evaluation. Thus, the evaluation measures are the same: *visit-level precision@k* and *code-level accuracy@k*.

For the visit-level evaluation, *visit-level precision@k* is defined as the correct medical codes in top k divided by $\min(k, |\mathbf{y}_t|)$, where $|\mathbf{y}_t|$ is the number of category labels in the $(t+1)$-th visit. We report the average values of *visit-level precision@k* in the experiments.

In the code-level evaluation, given a visit V_t which contains multiple category labels, if the target label is in the top k guesses, then we get 1 and 0 otherwise. Thus, *code-level accuracy@k* is defined by the number of correct label predictions divided by the total number of label predictions.

We vary k from 5 to 30. *Visit-level precision@k* aims to evaluate the coarse-grained performance, and *code-level accuracy@k* is proposed to evaluate the fine-grained performance. For all the measures, the greater values, the better performance.

Implementation Details
As in [10], we also use CCS-multi-level diagnoses hierarchy[5] as the knowledge graph. We implement all the approaches with Theano

[3] http://www.icd9data.com
[4] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt
[5] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixCMultiDX.txt

Table 2: The Accuracy@k of Diagnosis Prediction Task.

Dataset	Model	Visit-Level Precision@k						Code-Level Accuracy@k					
		5	10	15	20	25	30	5	10	15	20	25	30
Medicaid	KAME	**0.6107**	**0.7475**	**0.8168**	**0.8606**	**0.8920**	**0.9154**	**0.5461**	**0.7037**	**0.7808**	**0.8305**	**0.8667**	**0.8940**
	GRAM	0.5832	0.7189	0.7902	0.8367	0.8717	0.8976	0.5279	0.6842	0.7630	0.8146	0.8528	0.8819
	Dipole	0.5943	0.7226	0.7892	0.8340	0.8680	0.8942	0.5406	0.6903	0.7637	0.8130	0.8503	0.8791
	RNN+	0.5964	0.7210	0.7919	0.8397	0.8746	0.9011	0.5402	0.6867	0.7642	0.8166	0.8550	0.8845
	RNN	0.5448	0.6737	0.7503	0.8036	0.8433	0.8740	0.4914	0.6370	0.7200	0.7782	0.8222	0.8564
Diabetes	KAME	**0.5881**	**0.7313**	**0.8054**	**0.8523**	**0.8859**	**0.9107**	**0.5147**	**0.6939**	**0.7779**	**0.8293**	**0.8666**	**0.8949**
	GRAM	0.5596	0.7048	0.7822	0.8326	0.8684	0.8962	0.4958	0.6776	0.7617	0.8158	0.8546	0.8848
	Dipole	0.5697	0.7015	0.7765	0.8267	0.8640	0.8921	0.5110	0.6771	0.7585	0.8120	0.8520	0.8824
	RNN+	0.5680	0.7007	0.7769	0.8279	0.8649	0.8943	0.5086	0.6740	0.7569	0.8118	0.8519	0.8838
	RNN	0.5515	0.6851	0.7639	0.8179	0.8575	0.8877	0.4984	0.6611	0.7459	0.8024	0.8445	0.8765
MIMIC-III	KAME	**0.7103**	**0.6568**	**0.6967**	**0.7562**	**0.8091**	**0.8470**	**0.3167**	**0.5100**	**0.6379**	**0.7240**	**0.7862**	**0.8303**
	GRAM	0.6998	0.6447	0.6847	0.7439	0.8007	0.8424	0.3123	0.5026	0.6296	0.7142	0.7798	0.8266
	Dipole	0.6220	0.5839	0.6310	0.6953	0.7556	0.8059	0.2774	0.4556	0.5801	0.6671	0.7354	0.7902
	RNN+	0.6158	0.5803	0.6243	0.6912	0.7542	0.8017	0.2760	0.4548	0.5751	0.6647	0.7350	0.7867
	RNN	0.6580	0.6186	0.6637	0.7254	0.7836	0.8272	0.2941	0.4836	0.6106	0.6961	0.7629	0.8119

0.9.0 [37]. For training models, we use Adadelta [42] with a min-batch of 50 patients. We randomly divide the datasets into the training, validation and testing sets based on the number of patients in a 0.75:0.10:0.15 ratio. The validation set is used to determine the best values of parameters in the 100 training iterations. The regularization (l_2 norm with the coefficient 0.001) and the drop-out strategies (the drop-out rate is 0.5) are used for all the approaches. In order to fairly compare the performance, we set the same $d = 128$ and $g = 128$ for all the baselines and the proposed KAME.

3.3 Results of Diagnosis Prediction

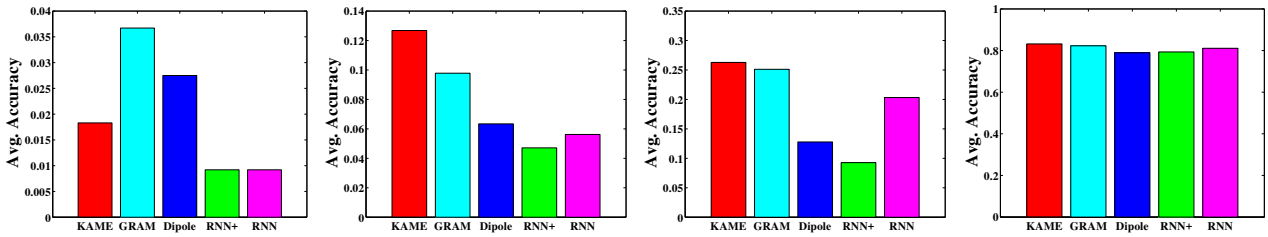
Table 2 shows both the visit-level precision and code-level accuracy of the proposed KAME and baselines with different k's on three real world datasets for diagnosis predication task. From Table 2, we can observe that the performance of the proposed KAME, including both visit-level precision and code-level accuracy, is better than that of all the baselines on the three datasets.

On the Medicaid dataset, compared with GRAM, the visit-level precision improves 4.7% and code-level accuracy improves 3.4% when $k = 5$. These results suggest that adding knowledge attention layer when predicting diagnoses is effective. Comparably, Dipole and RNN+ do not use external knowledge in the diagnosis prediction task. They directly learn the medical code embeddings from the input data with location-based attention mechanism. Compared with GRAM, the performance of both Dipole and RNN+ is better. The results also suggest that with sufficient data, even without external knowledge, attention-based models can still learn reasonable medical code embeddings to make accurate predictions. However, compared with the proposed KAME, the precision and accuracy of these two approaches are lower, which again confirms that considering general or high-level information can improve the prediction performance. The performance of RNN is the worst since this approach does not use any attention mechanism or external knowledge. The visit-level precision and code-level accuracy of KAME increase 12.1% and 11.1% respectively compared with RNN when $k = 5$.

On the Diabetes dataset, the proposed KAME still outperforms all the state-of-the-art diagnosis prediction approaches. Compared with the Medicaid dataset, the data are relatively insufficient in the Diabetes dataset. Thus, the performance (both visit-level precision and code-level accuracy) of GRAM is competitive to that of RNN+ and Dipole, but still worse than that of KAME. This shows that the performance of models with knowledge graph is comparable to models with attention mechanisms on the Diabetes dataset.

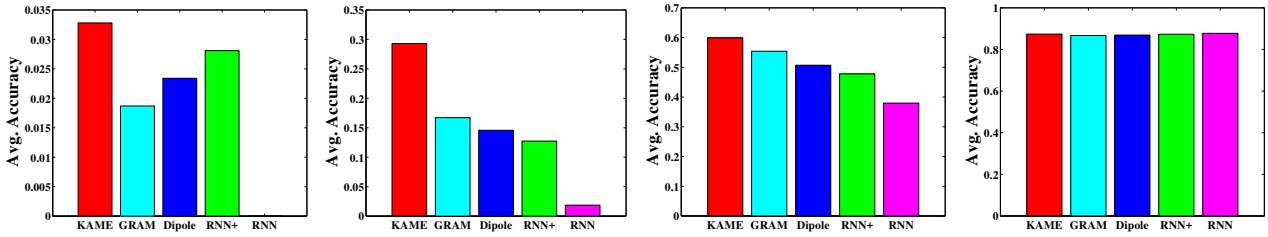
Since the number of visits for each patient on the MIMIC-III dataset is much smaller than that on the Medicaid and Diabetes dataset, the data are significantly insufficient, i.e., less labels are observed in the training data. On this insufficient dataset, KAME still outperforms all the baselines. In the four baselines, GRAM achieves the best performance, which shows that employing knowledge graph is effective with significant data insufficiency. The precision and accuracy of both Dipole and RNN+ are lower than those of RNN. This demonstrates that training attention models on the previous visits needs more data. However, instead of adding attention mechanisms on the past visits, the proposed KAME aims to extract knowledge from the given knowledge graph with attention mechanism.

As expected, the values of precision and accuracy increase with larger k values, except the visit-level precision on the MIMIC III dataset. The reason is that there are some labels without sufficient training data, and they obtain lower probabilities in the predictions compared with those well trained. Thus, for the visits that contain some labels without sufficient training data, the number of correct predictions when k is 10 or 15 may be the same with that when $k = 5$. However, they are divided by a bigger $\min(k, |\mathbf{y}_t|)$, which leads to the observation that the average performance is worse than that with $k = 5$. All the results in Table 2 can significantly and strongly validate the robustness of KAME on different types of datasets.



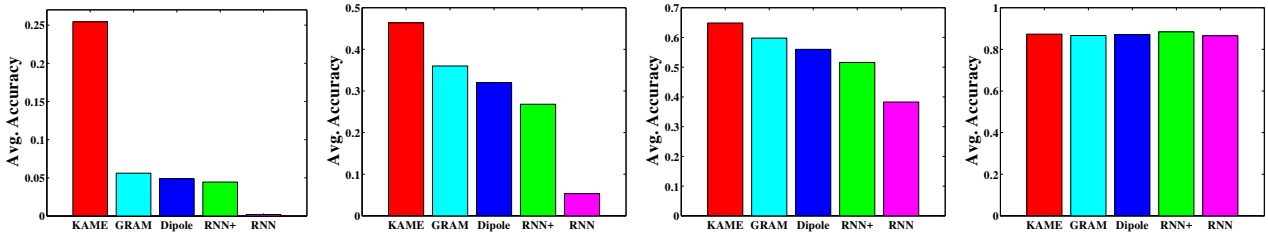
(a) 0-25 (b) 25-50 (c) 50-75 (d) 75-100

Figure 2: Code-Level Accuracy@20 of Diagnosis Prediction on the MIMIC-III Dataset.



(a) 0-25 (b) 25-50 (c) 50-75 (d) 75-100

Figure 3: Code-Level Accuracy@20 of Diagnosis Prediction on the Diabetes Dataset.



(a) 0-25 (b) 25-50 (c) 50-75 (d) 75-100

Figure 4: Code-Level Accuracy@20 of Diagnosis Prediction on the Medicaid Dataset.

3.4 Data Sufficiency Analysis

In order to analyze the influence of data sufficiency on the predictions, we conduct the following experiments on the MIMIC-III, Diabetes and Medicaid datasets, respectively. We first rank all the category labels appeared in the training set based on their frequency, and then divide them into four groups: 0-25, 25-50, 50-75 and 75-100. The category labels in the 0-25 group are the most rare ones in the training set, while the labels in the 75-100 group are the most common ones. Finally, we calculate the accuracy of labels in each group. Figures 2, 3 and 4 show the code-level accuracy@20 on the MIMIC-III, Diabetes and Medicaid datasets, respectively. X-axis denotes all the approaches, and Y-axis is the average accuracy of the approaches. Note that similar results can be obtained when $k = 5, 10, 15, 25$ or 30.

From Figure 2, we can observe that the accuracy of the proposed KAME is higher than that of baselines in the groups 25-50, 50-75 and 75-100. For the group 0-25, GRAM outperforms other approaches, which shows that with insufficient data, GRAM still learns reasonable medical code embeddings to improve the predictions. Similar observations also can be found in other groups, i.e., the performance of GRAM is better than that of other baselines.

On the other hand, when the training data on the Diabetes and Medicaid datasets is sufficient, the proposed KAME still significantly outperforms baselines in the groups 0-25, 25-50 and 50-75. Especially in the group 0-25 on the Medicaid dataset, GRAM achieves the highest average accuracy among baselines that is 0.0561, but the accuracy of KAME is 0.2543, which improves 353.3%. This results demonstrates the effectiveness of the proposed knowledge-based attention mechanism with insufficient training EHR data. As shown in Figures 3 and 4, the difference of average accuracy between GRAM and attention-based models drops, i.e., RNN+ and Dipole, which shows that the attention mechanism starts to play a more important role under sufficient data. These observation also can be found in Table 2.

From Figures 2, 3 and 4, we can conclude that through adopting the medical knowledge graph, the proposed KAME uses knowledge-based attention mechanism in the prediction step, which infers the

general knowledge information to improve the predictive performance. Thus, the final prediction performance of KAME is better than that of baselines as shown in Table 2.

3.5 Case Study for Knowledge Attention

To demonstrate the additional benefits of applying the proposed knowledge attention mechanism in diagnosis prediction task, we analyze the attention weights learned from the proposed approach KAME with two examples from the MIMIC-III dataset shown in Table 3. In Table 3, the first column represents the medical codes of the t-th visit, the second column denotes the knowledge (i.e., the ancestors of the medical codes in V_t) with high attention weights which are calculated by Eq. (5), and the third column is the $(t+1)$-th visit's medical codes. We intend to show the relationships between the knowledge attention weights and the predictions.

In the first example, we can observe that the knowledge "*Coronary atherosclerosis*" has the highest attention weight, which is related to heart disease. This potentially helps the model to predict ("*Other complications due to other cardiac device, implant, and graft (996.72)*") at the $(t+1)$-th visit. The second example shows that the proposed KAME can calculate the correct attention weight with the knowledge ("*Secondary malignancies*") in the knowledge graph, which makes KAME predict "*Secondary malignant neoplasm of pleura (197.2)*" and "*Secondary malignant neoplasm of retroperitoneum and peritoneum (197.6)*" with high confidence.

This case study demonstrates that we can learn an accurate attention weight for each piece of knowledge, and the experimental results in Section 3.3 also illustrate that learning over the knowledge graph with the proposed knowledge attention mechanism can significantly improve the performance of the diagnosis prediction task in healthcare.

3.6 Interpretable Representation Analysis

To qualitatively demonstrate the interpretability of the learned medical code representations by all the predictive models on the Diabetes dataset, we randomly select 2000 medical codes and then plot on a 2-D space with t-SNE [26] shown in Figure 5. Each dot represents a diagnosis code. The color of the dots represents the highest or first disease categories in CCS multi-level hierarchy. Ideally, the dots with the same color should be in the same cluster, and there are margins among different clusters.

From Figure 5, we can observe that KAME and GRAM learn interpretable disease representations that are in accord with the hierarchies of the given knowledge graph $\mathcal{G}$. In addition, the predictive performance of KAME is much better than that of GRAM shown in Table 2, which proves that the proposed knowledge attention mechanism does not affect the interpretability of medical codes. In addition, it significantly improves the prediction accuracy. Figure 5(c), 5(d) and 5(e) confirm that without knowledge graph, simply using the co-occurrence or supervised predictions cannot easily learn interpretable representations.

4 RELATED WORK

In this section, we review the work about mining electronic healthcare records with deep learning techniques, especially for diagnosis prediction. We then introduce some work on attention mechanism and graph representation learning.

4.1 Deep Learning for EHR Data

Gaining knowledge from the massive EHRs [25, 27, 31, 34, 35, 41, 43] is a hot research topic in healthcare informatics. Recently, deep learning techniques have shown their superior ability for mining EHR data. Recurrent neural networks (RNNs) can be used for diagnosis classification [20], patient subtyping [3], modeling disease progression [30], and mining time series healthcare data with missing values [6, 21]. Convolutional neural networks (CNNs) are used for predicting unplanned readmission [28] and risk [7, 24] with EHRs. Stacked denoising autoencoders (SDAs) are employed to detect the characteristic patterns of physiology in clinical time series data [5].

Diagnosis prediction is one of the important tasks in EHR data mining, which aims to predict the future visit information according to historical visit records of patients. Med2Vec [9] is an unsupervised method for learning the representations of medical codes, which can be used to predict the future visit information. However, this method ignores long-term dependencies of medical codes among visits. RETAIN [11] is an interpretable predictive model, which employs a reverse time attention mechanism in an RNN for binary prediction task. Dipole [23] applies bidirectional recurrent neural networks (BRNNs) and attention mechanisms to predict patient visit information. GRAM [10] is a graph-based attention model for healthcare representation learning, which uses medical ontologies to learn robust representations and an RNN to model patient visits.

Among the aforementioned predictive models, GRAM is the most relevant model to our proposed KAME. Actually, KAME is a generalization of GRAM. Compared with GRAM, KAME not only uses graph-based attention model to learn medical representations, but also employs *knowledge-based attention* mechanism to generate *knowledge vectors* and makes predictions according to the learned knowledge vectors to improve the predictive performance.

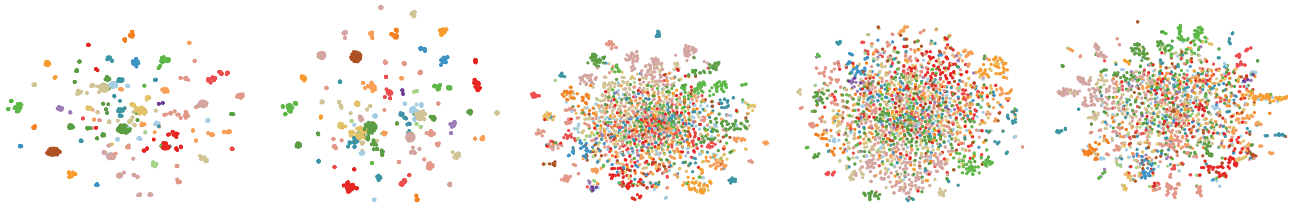
4.2 Attention & Graph Representation

Attention-based neural networks have been successfully used in many tasks [1, 2, 12, 15, 18, 22, 39, 40], such as neural machine translation [2, 22], computer vision [40], speech recognition [12] and healthcare [10, 11, 23, 33]. In healthcare, most of existing work aims to learn attention weights between the current visit and all the previous ones, or medical codes and their ancestors. However, the proposed KAME calculates attention weights between knowledge graph and the current visit. The goal of KAME is to learn general or high-level knowledge representations to help the final predictions.

Learning the representations of graphs is a hot research topic which motivates various methods, such as DeepWalk [29], Node2Vec [14], LSHM [17], LINE [36], Metapath2Vec [13], and Struc2Vec [32]. All the aforementioned models focus on learning good representations for graph data, while the proposed KAME is a diagnosis predictive model, and we aim to improve the predictive performance with the given knowledge graph as supplementary information.

Table 3: Case Study for the Proposed Knowledge Attention Mechanism.

Visit t	Knowledge and Attention Weight	Visit $t+1$
Acute myocardial infarction of other anterior wall, initial episode of care (410.11) Coronary atherosclerosis of native coronary artery (414.01) Pure hypercholesterolemia (272.0)	Coronary atherosclerosis (0.5968) Coronary atherosclerosis and other heart disease (0.236) Acute myocardial infarction (0.0919) Hypopotassemia (0.0220) Diseases of the heart (0.0126)	Other complications due to other cardiac device, implant, and graft (996.72) Coronary atherosclerosis of native coronary artery (414.01) Acute myocardial infarction of other anterior wall, initial episode of care (410.11) Hypopotassemia (276.8)
Other specified diseases of pericardium (423.8) Malignant neoplasm of upper lobe, bronchus or lung (162.3) Esophageal reflux (530.81) Polyneuropathy due to drugs (357.6) Injury due to war operations by guided missile (E993.1)	Secondary malignancies (0.9777) Neoplasms (0.0223)	Secondary malignant neoplasm of pleura (197.2) Secondary malignant neoplasm of retroperitoneum and peritoneum (197.6) Malignant neoplasm of other parts of bronchus or lung (162.8)



(a) KAME (b) GRAM (c) Dipole (d) RNN+ (e) RNN

Figure 5: t-SNE Scatterplots of Medical Codes Learned by Predictive Models on the Diabetes dataset.

Knowledge graph representation learning is relevant to the proposed model, but they are totally different. The goal of knowledge graph representation learning is to learn the representations of nodes, entities and relations, such as TransE [4], TransH [38] and TransR [19]. These approaches are used for link prediction or entity classification. The proposed KAME is different from them in that it aims to design intuitive attention mechanisms on the given knowledge graph and learn meaningful and interpretable medical code representations for making accurate predictions.

5 CONCLUSIONS

Diagnosis prediction is a core task in healthcare informatics. The state-of-the-art diagnosis prediction approaches employ recurrent neural networks to model sequential EHR data and adopt attention mechanisms to improve the prediction accuracy and interpretability. However, these models suffer from the problem of robustness for different types of data and ignore the importance of employing general knowledge in the medical ontologies to improve the predictive performance.

In this paper, we propose a new diagnosis prediction model, named KAME, which can fully utilize the information of medical ontologies to improve the prediction accuracy. By learning from the given knowledge graph, KAME not only obtains the accurate embeddings of medical codes, but also directly derives the general knowledge from the ancestor codes. With the learned medical code embeddings and RNNs, KAME can remember the hidden information of all the previous visits. Through calculating attention weights between the hidden information and the general knowledge, KAME can obtain a novel knowledge vector, which largely helps the predictive model to improve the performance. Moreover, the learned attention weights allow us to reasonably interpret the importance of each piece of knowledge. Experimental results on three real world medical datasets prove the effectiveness and robustness of the proposed KAME for diagnosis prediction task. An experiment is conducted to show that the proposed KAME outperforms baselines with both sufficient and insufficient data. The representations of medical codes are visualized to illustrate the interpretability of KAME. Finally, a case study demonstrates the reasonableness of the proposed knowledge-based attention mechanism.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and suggestions, and NVIDIA Corporation with the donation of the Titan Xp GPU. This work is supported in part by the US National Science Foundation under grants IIS-1553411 and IIS-1747614. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2015. Multiple Object Recognition with Visual Attention. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*.

[3] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient Subtyping via Time-Aware LSTM Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 65–74.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NIPS'13)*. 2787–2795.

[5] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. 507–516.

[6] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2016. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *arXiv preprint arXiv:1606.01865* (2016).

[7] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM'16)*. 432–440.

[8] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv preprint arXiv:1409.1259* (2014).

[9] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. 1495–1504.

[10] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 787–795.

[11] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *Advances in Neural Information Processing Systems (NIPS'16)*. 3504–3512.

[12] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based Models for Speech Recognition. In *Advances in Neural Information Processing Systems (NIPS'15)*. 577–585.

[13] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. 2017. Metapath2Vec: Scalable Representation Learning for Heterogeneous Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 135–144.

[14] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, 855–864.

[15] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems (NIPS'15)*. 1693–1701.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[17] Yann Jacob, Ludovic Denoyer, and Patrick Gallinari. 2014. Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM'14)*. 373–382.

[18] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor Forcing: A New Algorithm for Training Recurrent Networks. In *Advances In Neural Information Processing Systems (NIPS'16)*. 4601–4609.

[19] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. 2181–2187.

[20] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. 2015. Learning to Diagnose with LSTM Recurrent Neural Networks. In *International Conference on Learning Representations (ICLR'16)*.

[21] Zachary C Lipton, David C Kale, and Randall Wetzel. 2016. Modeling Missing Data in Clinical Time Series with RNNs. In *Proceedings of Machine Learning for Healthcare (MLHC'16)*.

[22] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. 1412–1421.

[23] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 1903–1911.

[24] Fenglong Ma, Gao Jing, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. 2018. Risk Prediction on Electronic Health Records with Prior Medical Knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18)*. ACM, 1910–1919.

[25] Fenglong Ma, Chuishi Meng, Houping Xiao, Qi Li, Jing Gao, Lu Su, and Aidong Zhang. 2017. Unsupervised Discovery of Drug Side-Effects from Heterogeneous Data Sources. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 967–976.

[26] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[27] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2017. Deep Learning for Healthcare: Review, Opportunities and Challenges. *Briefings in Bioinformatics* (2017), bbx044.

[28] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2016. Deepr: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics* (2016).

[29] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. ACM, 701–710.

[30] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'16)*. 30–41.

[31] Alvin Rajkomar, Eyal Oren, Andrew M. Dai Kai Chen, Nissan Hajaj, Peter J. Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin E. Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, Justin Tansuwanand De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael Howell, Claire Cui, Greg Corrado, and Jeff Dean. 2018. Scalable and accurate deep learning for electronic health records. *arXiv preprint arXiv:1801.07860* (2018).

[32] Leonardo F.R. Ribeiro, Pedro H.P. Saverese, and Daniel R. Figueiredo. 2017. Struc2Vec: Learning Node Representations from Structural Identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 385–394.

[33] Qiuling Suo, Fenglong Ma, Giovanni Canino, Jing Gao, Aidong Zhang, Pierangelo Veltri, and Agostino Gnasso. 2017. A Multi-task Framework for Monitoring Health Conditions via Attention-based Recurrent Neural Networks. In *Proceedings of the AMIA 2017 Annual Symposium (AMIA'17)*.

[34] Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, and Aidong Zhang. 2017. Personalized Disease Prediction Using A CNN-Based Similarity Learning Method. In *Proceedings of The IEEE International Conference on Bioinformatics and Biomedicine (BIBM'17)*. 811–816.

[35] Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, and Aidong Zhang. 2018. Deep Patient Similarity Learning for Personalized Healthcare. *IEEE Transactions on NanoBioscience* (2018), 219–227.

[36] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web (WWW'14)*. 1067–1077.

[37] The Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688* (2016).

[38] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14)*. 1112–1119.

[39] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*. 2048–2057.

[40] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 4651–4659.

[41] Ye Yuan, Guangxu Xun, Fenglong Ma, Qiuling Suo, Hongfei Xue, Kebin Jia, and Aidong Zhang. 2018. A Novel Channel-aware Attention Framework for Multi-Channel EEG Seizure Detection via Multi-view Deep Learning. In *Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI'18)*. IEEE, 206–209.

[42] Matthew D Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv preprint arXiv:1212.5701* (2012).

[43] Shiyue Zhang, Pengtao Xie, Dong Wang, and Eric P Xing. 2017. Medical Diagnosis From Laboratory Tests by Combining Generative and Discriminative Learning. *arXiv preprint arXiv:1711.04329* (2017).