CAMP: Co-Attention Memory Networks for Diagnosis Prediction in Healthcare

Jingyue Gao[1,3], Xiting Wang[2], Yasha Wang[1,4,*], Zhao Yang[1,3], Junyi Gao[1], Jiangtao Wang[5], Wen Tang[6], Xing Xie[2]
[1]Key Laboratory of High Confidence Software Technologies, Ministry of Education
[2]Microsoft Research Asia; [3]School of Electronics Engineering and Computer Science, Peking University
[4]National Engineering Research Center of Software Engineering, Peking University
[5]School of Computing and Communications, Lancaster University; [6]Peking University Third Hospital
[*]The corresponding author, email: wangyasha@pku.edu.cn

***Abstract*—Diagnosis prediction, which aims to predict future health information of patients from historical electronic health records (EHRs), is a core research task in personalized healthcare. Although some RNN-based methods have been proposed to model sequential EHR data, these methods have two major issues. First, they cannot capture fine-grained progression patterns of patient health conditions. Second, they do not consider the mutual effect between important context (e.g., patient demographics) and historical diagnosis. To tackle these challenges, we propose a model called Co-Attention Memory networks for diagnosis Prediction (CAMP), which tightly integrates historical records, fine-grained patient conditions, and demographics with a three-way interaction architecture built on co-attention. Our model augments RNNs with a memory network to enrich the representation capacity. The memory network enables analysis of fine-grained patient conditions by explicitly incorporating a taxonomy of diseases into an array of memory slots. We instantiate the READ/WRITE operations of the memory network so that the memory cooperates effectively with the patient demographics through co-attention mechanism. Experiments on real-world datasets demonstrate that CAMP consistently performs better than state-of-the-art methods.**

***Index Terms*—diagnosis prediction, memory networks, attention mechanism, healthcare informatics**

I. INTRODUCTION

Nowadays, Electronic Health Record (EHR) systems are widely adopted to record longitudinal patient health data such as diagnosis, medications, and procedures, which enables the possibility of clinical predictive tasks. Predicting future diagnosis based on patient's historical records of diagnosis, i.e., *diagnosis prediction* [1], [2], has become a cornerstone of personalized healthcare. This task attracts considerable attention in both industry and the research community because of their importance in need anticipation and precision medicine [2], [3]. Although there is broad consensus on its importance, diagnosis prediction is challenging due to the sequential, high-dimensional, and noisy nature of EHR data.

With recent advances in deep learning, many studies on diagnosis prediction adopt Recurrent Neural Networks (RNNs) to model sequential EHR data. For example, Choi et al. [1] apply RNNs on reversed diagnosis sequences and Ma et al. [4] use Bidirectional RNNs for further improvement. Recently, researchers have incorporated taxonomies of diseases into RNNs [5], [6]. These methods have achieved encouraging prediction accuracy due to their ability to capture dynamic patient conditions and estimate the likelihood of future diagnosis. However, they cannot effectively address the following two challenges in diagnosis prediction.

C1: It is difficult to capture fine-grained progression patterns of patient conditions. The health conditions of a patient can be complicated: diseases are correlated with each other and there may be long-term dependencies between diseases of different categories [4]. To effectively model complex patient health conditions, we need to perform fine-grained analysis on the relationships between the diseases and their attributes (e.g., categories). However, RNNs tend to focus more on short-term memories [7], [8] and would forcefully compress historical records into one hidden state vector. Such highly abstractive features constrain the representation power of RNNs and make it difficult for RNNs to preserve fine-grained information of diagnosed diseases and long-term patient health conditions.

C2: Existing methods cannot model the mutual effect between important context and historical records. Patient demographics are considered important context in the domain of diagnosis prediction [2], [9]. However, how to model the mutual effect between patient demographics and their diagnosed diseases has not been explored, which limits the accuracy of existing methods.

Based on these observations, we propose a model called **C**o-**A**ttention **M**emory networks for diagnosis **P**rediction (**CAMP**[1]) that addresses these challenges. As shown in Figure 1, we design a three-way interaction neural architecture built upon co-attention to tightly integrate historical records, fine-grained patient conditions, and demographics. We enable the analysis of fine-grained patient conditions by explicitly incorporating taxonomies of diseases into the framework and memorizing the knowledge contained in the taxonomies with Key-Value Memory Networks (KV-MNs) [10]. Instead of relying on a compressed vector, KV-MNs store different

[1]The source code is available at CAMP. The long version of this paper can be found here, which further includes discussion about model interpretability.

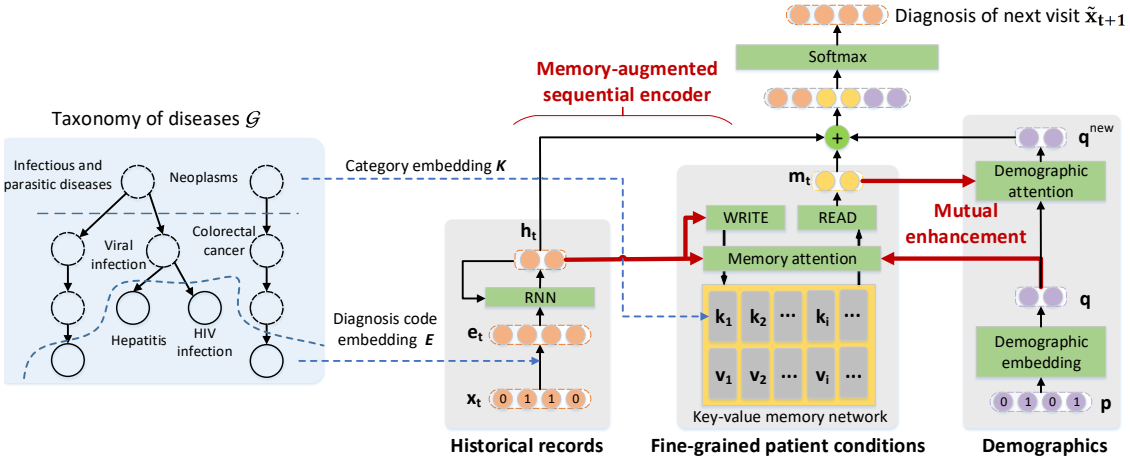

Figure 1: Framework of the proposed model for diagnosis prediction.

categories of information separately in different memory slots, which enriches the representation capacity compared with RNNs [11], [12]. We elaborately design the memory slots and the READ/WRITE operations of the memory network so that the KV-MN can 1) effectively model the disease categories and their relationships (e.g., connections through ancestors) to capture fine-grained dynamic health conditions of patients (**C1**); 2) cooperate with patient demographics in a mutual enhancement way through a co-attention mechanism (**C2**).

Experiments on real-world datasets demonstrate that CAMP consistently outperforms state-of-the-art methods in terms of different evaluation criteria. Detailed analysis of CAMP also validates the effectiveness of different components.

II. Problem Definition

We define the problem of diagnosis prediction as follows. For simplicity, all algorithms will be presented for one patient.

Input. For each patient, the input data of our model consists of a sequence of his/her historical records $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t)$, patient demographics $\mathbf{p}$ and a given taxonomy of diseases $\mathcal{G}$.

- The **historical records** of a patient contain multiple visits. Each visit $\mathbf{x}_j \in \{0,1\}^{|\mathcal{C}|}$, $j \in [1,t]$ is a multi-hot binary vector. Here $|\mathcal{C}|$ denotes the number of diseases and $\mathcal{C} = \{c_1, c_2, ..., c_{|\mathcal{C}|}\}$ is the set of entire disease codes from the EHR. $x_{j,i} = 1$ indicates that the patient was diagnosed with disease c_i in the j-th visit.
- The **patient demographics** $\mathbf{p}$ consists of patient characteristics such as age and gender, which are often recorded in the EHR. Let $\mathbf{p} \in \{0,1\}^r$ denote a multi-hot vector indicating demographics of the patient. Following [2], we construct $\mathbf{p}$ by discretizing each attribute (e.g., divide age into several groups), representing the discretized attributes as one-hot vectors, and concatenating them.
- The **disease taxonomy** $\mathcal{G}$ expresses the hierarchy of disease concepts in the form of a parent-child relationship, where diseases in $\mathcal{C}$ form the leaf nodes. As shown in Figure 1, a parent node (e.g., *viral infection*) in $\mathcal{G}$ is a disease category that summarizes the diseases described by its children (e.g., *HIV infection* and *Hepatitis*). All nodes in $\mathcal{G}$ form the set $\mathcal{D} = \mathcal{C} + \mathcal{C}'$, where $\mathcal{C}' = \{c_{|\mathcal{C}|+1}, ..., c_{|\mathcal{C}|+|\mathcal{C}'|}\}$ consists of all ancestor nodes. The L nodes at the highest hierarchical level of $\mathcal{G}$ represent the most general categories of diseases (e.g., *Infectious and parasitic diseases*). We call these nodes top-level categories. We build $\mathcal{G}$ by using well-organized taxonomies of diseases (e.g., ICD[2] and CCS[3]).

Output: Given the historical records of a patient, his/her demographics, and a disease taxonomy, the output of our model is the predicted diagnosis of the next visit: $\tilde{\mathbf{x}}_{t+1}$.

III. The Proposed Model

Figure 1 shows the framework of CAMP, which is a three-way interaction architecture that tightly integrates historical records, fine-grained patient conditions, and demographics. In particular, CAMP predicts future diagnosis with two major components: 1) **Memory-augmented sequential encoder** that captures fine-grained dynamic health conditions of a patient by augmenting RNN-based models with external memory networks; 2) **Co-attention-based mutual enhancement** where the representations of patient health conditions and demographics are mutually improved via co-attention mechanism. In the following, we describe the design of these two components and illustrate how they can be jointly optimized.

A. Memory-Augmented Sequential Encoder

Our memory-augmented sequential encoder consists of diagnosis embedding, the RNN, and the memory network.

1) Diagnosis Embedding: The goal of diagnosis embedding is to encode hierarchical medical knowledge in the representations of diseases and their categories. The enriched embeddings help to handle data insufficiency and enhance model accuracy. Given taxonomy $\mathcal{G}$ (shown in Figure 1), we learn robust embeddings for each node in $\mathcal{G}$ by using a state-of-the-art method, GRAM [5]. GRAM represents a node as a

[2]https://www.cdc.gov/nchs/icd/index.htm
[3]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

combination of itself and its ancestors in $\mathcal{G}$ via a graph-based attention mechanism. With embeddings of all nodes available, we construct two embedding matrices:

- Diagnosis code embedding matrix $\mathbf{E} \in \mathbb{R}^{d_1 \times |\mathcal{C}|}$, which contains embeddings of all leaf nodes (i.e., diseases).
- Category embedding matrix $\mathbf{K} \in \mathbb{R}^{d_1 \times L}$, which contains embeddings of all top-level disease categories.

d_1 is the embedding size. L is the number of top-level nodes.

2) RNN: Recurrent neural networks (RNNs) have been proven effective in modeling the temporal dependency in sequences. Here we choose Gated Recurrent Unit (GRU) [13], a RNN variant tackling the problem of vanishing gradient.

Let d_2 denote the hidden size of GRU, current hidden state vector $\mathbf{h}_t \in \mathbb{R}^{d_2}$ of GRU can be computed recursively:

$$\mathbf{h}_t = \text{GRU}(\mathbf{h}_{t-1}, \mathbf{e}_t; \Theta), \tag{1}$$

where GRU$(\cdot)$ is the GRU unit, $\mathbf{e}_t = \mathbf{E}\mathbf{x}_t$ is the embedded diagnosis of t-th visit, $\mathbf{h}_{t-1}$ denotes the previous hidden state vector, and Θ represents all parameters of the GRU unit. Researchers have shown that RNNs tend to capture disease progression in the short term and fail to remember patient health conditions over the long term [14]. Thus, we consider $\mathbf{h}_t$ as a representation of short-term patient conditions.

3) Memory Network: To overcome the shortcoming of RNNs in capturing patient health conditions over the long term, we design a memory network that can preserve fine-grained information of long-term health conditions.

Modeling fine-grained patient conditions with key-value memory networks. To model fine-grained patient conditions, we adopt a key-value memory network (KV-MN), which memorizes information by using a large array of external memory slots. The external memories enrich the representation capability compared with hidden vectors of RNNs and enable the KV-MN to capture long-term data characteristics [15]. Specifically, we incorporate the knowledge contained in the disease taxonomy into the memory slots and design each memory slot so that it memorizes patient health conditions on a specific disease category. Compared with RNNs that capture the overall health conditions of a patient, the KV-MN decomposes patient conditions into different disease categories and thus preserves more fine-grained information.

In KV-MNs, a memory slot is represented by a key vector and an associated value vector. Next, we introduce our design of the key vectors, the value vectors, and READ/WRITE operations used to manipulate the memory.

Key vectors. We set the key vectors as the embeddings of the top-level nodes in taxonomy $\mathcal{G}$. This ensures that each memory slot corresponds to a disease category. In particular, the i-th key vector $\mathbf{k}_i \in \mathbb{R}^{d_1}$ is set to the i-th column of the category embedding matrix $\mathbf{K}$. Since $\mathbf{K}$ is computed by using graph-based attention (Section III-A1), it captures the hierarchical information of the taxonomy, e.g., relationships between different diseases. $\mathbf{K}$ is shared by all patients and fixed during the processing of diagnosis sequences.

Value vectors. Let $\mathbf{v}_i$ denote the value vector associated with $\mathbf{k}_i$. Each value vector $\mathbf{v}_i$ memorizes information about patient conditions on one disease category, which helps predict future diagnosis regarding this category. We form a value memory matrix $\mathbf{V} \in \mathbb{R}^{d_v \times L}$ by combining all L value slots. Different from $\mathbf{K}$, $\mathbf{V}$ is patient-specific and is continuously updated according to the input diagnosis sequence. In this way, we capture the dynamic patient conditions on each disease category. Two types of operations, READ and WRITE, are designed to manipulate the value vector.

READ operation. With fine-grained information of historical diagnosis stored in $\{(\mathbf{k}_1, \mathbf{v}_1), \cdots, (\mathbf{k}_L, \mathbf{v}_L)\}$, we obtain long-term patient health conditions from these slots by using the READ operation. Since the patients do not equally suffer from all categories of diseases, we use the short-term representation $\mathbf{h}_t$ as a query to attentively visit the memory network. The attention weight $a_{t,i}$ of $(\mathbf{k}_i, \mathbf{v}_i)$ is calculated according to the correlation between $\mathbf{h}_t$ and $\mathbf{k}_i$:

$$a_{t,i} = \frac{\exp(\mathbf{k}_i^\top \text{MLP}(\mathbf{h}_t))}{\sum_{j=1}^{L} \exp(\mathbf{k}_j^\top \text{MLP}(\mathbf{h}_t))}, \tag{2}$$

where MLP$(\cdot)$ is a transformation layer. Larger $a_{t,i}$ suggests that there is a larger probability that the patient suffers from diseases from the i-th category. The long-term patient health conditions $\mathbf{m}_t$ can thus be represented:

$$\mathbf{m}_t = \sum_{i=1}^{L} a_{t,i} \mathbf{v}_i. \tag{3}$$

WRITE operation. To memorize information of recent diagnosis in the memory network, we update the value matrix $\mathbf{V}$ according to the short-term representation $\mathbf{h}_t$. Inspired by [16], we employ an *erase*-followed-by-*add* update mechanism. This mechanism allows us to erase unnecessary information in the memory and add new information with respect to patient health conditions dynamically.

We first derive an erase vector and an add vector from $\mathbf{h}_t$:

$$\begin{aligned} \mathbf{erase}_t &= \text{sigmoid}(\mathbf{W}_1 \mathbf{h}_t + \mathbf{b}_1), \\ \mathbf{add}_t &= \tanh(\mathbf{W}_2 \mathbf{h}_t + \mathbf{b}_2), \end{aligned} \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_v \times d_2}$, $\mathbf{b}_1 \in \mathbb{R}^{d_v}$, $\mathbf{W}_2 \in \mathbb{R}^{d_v \times d_2}$, and $\mathbf{b}_2 \in \mathbb{R}^{d_v}$ are parameters of the erase layer and the add layer. Here sigmoid$(\cdot)$ and tanh$(\cdot)$ are chosen as the activation functions of the erase layer and the add layer following [16]. Since memory slots that are associated with patient health conditions should be emphasized during the update, the WRITE operation is performed attentively by considering the attention weight $a_{t,i}$:

$$\mathbf{v}_i \leftarrow \mathbf{v}_i \odot (\mathbf{1} - a_{t,i} \mathbf{erase}_t) + a_{t,i} \mathbf{add}_t, \tag{5}$$

where $\odot$ is the Hadamard product and $\mathbf{1}$ is a d_v-dimensional column vector of all 1's. By learning the parameters of the erase and add layers, our model can determine which signals to weaken or strengthen based on recent diagnosis.

B. Co-Attention-Based Mutual Enhancement

To model the mutual effect between patient demographics and the memory network, we design a co-attention mechanism that consists of **attention for memory slots** and **attention for**

patient demographics. In this way, CAMP can accurately predict future diagnosis with mutually enhanced representations of long-term memory and patient demographics.

1) Attention for Memory Slots: It often happens that patients with certain demographics are vulnerable to some diseases while others are not. For instance, HFMD (hand, foot, and mouth disease) typically occurs in children instead of adults [17]. It inspires us to consider patient demographics $\mathbf{p}$ when computing the attention weight $a_{t,i}$ of the i-th disease category. Specifically, we first obtain the demographics embedding $\mathbf{q} \in \mathbb{R}^{d_3}$ from $\mathbf{p} \in \{0,1\}^r$ as:

$$\mathbf{q} = \mathbf{W}_p\mathbf{p} + \mathbf{b}_p, \tag{6}$$

where $\mathbf{W}_p \in \mathbb{R}^{d_3 \times r}$ and $\mathbf{b}_p \in \mathbb{R}^{d_3}$ are parameters of the embedding layer and d_3 is the embedding size. Then, we use the concatenation of $\mathbf{h}_t$ and $\mathbf{q}$ as the query to visit the memory network. Thus, Equation (2) is replaced with:

$$a_{t,i} = \frac{\exp(\mathbf{k}_i^\top \text{MLP}(\mathbf{h}_t \oplus \mathbf{q}))}{\sum_{j=1}^{L} \exp(\mathbf{k}_j^\top \text{MLP}(\mathbf{h}_t \oplus \mathbf{q}))}, \tag{7}$$

where $\oplus$ is the concatenation operator. We can manipulate the memory network better with the enhanced memory attention mechanism that takes patient demographics into consideration.

2) Attention for Patient Demographics: Given the demographics embedding $\mathbf{q}$, the long-term memory representation $\mathbf{m}_t$ can serve as an important context about historical diagnosis and help decide which latent features in $\mathbf{q}$ are more important for prediction. Thus, we leverage $\mathbf{m}_t$ to derive an attention vector $\boldsymbol{\beta} \in \mathbb{R}^{d_3}$ on $\mathbf{q}$ for enhancement:

$$\boldsymbol{\beta} = \text{ReLU}(\mathbf{W}_3\mathbf{m}_t + \mathbf{W}_4\mathbf{q} + \mathbf{b}_3), \tag{8}$$

where $\mathbf{W}_3 \in \mathbb{R}^{d_3 \times d_v}$, $\mathbf{W}_4 \in \mathbf{R}^{d_3 \times d_3}$, and $\mathbf{b}_3 \in \mathbb{R}^{d_3}$ are parameters. Conditioned on the historical diagnosis, $\boldsymbol{\beta}$ is used to enhance the original representation of demographics:

$$\mathbf{q}^{new} = \boldsymbol{\beta} \odot \mathbf{q}. \tag{9}$$

C. Joint Learning

Given the short-term representation $\mathbf{h}_t$, the long-term representation $\mathbf{m}_t$ of patient health conditions and the enhanced representation of patient demographics $\mathbf{q}^{new}$, we generate a joint representation of patient by concatenating the three representations. The concatenated vector is fed through a softmax layer to predict the diagnosis of next visit $\tilde{\mathbf{x}}_{t+1}$:

$$\tilde{\mathbf{x}}_{t+1} = \text{softmax}(\mathbf{W}_x(\mathbf{h}_t \oplus \mathbf{m}_t \oplus \mathbf{q}^{new}) + \mathbf{b}_x). \tag{10}$$

Here $\mathbf{W}_x \in \mathbb{R}^{|\mathcal{C}| \times (d_2 + d_v + d_3)}$ and $\mathbf{b}_x \in \mathbb{R}^{|\mathcal{C}|}$ are parameters to be learn. We calculate the cross-entropy loss between the ground truth $\mathbf{x}_{t+1}$ and the predicted $\tilde{\mathbf{x}}_{t+1}$:

$$\mathcal{L} = -\frac{1}{T-1}\sum_{t=1}^{T-1}(\mathbf{x}_{t+1}^\top \log(\tilde{\mathbf{x}}_{t+1}) + (\mathbf{1} - \mathbf{x}_{t+1})^\top \log(\mathbf{1} - \tilde{\mathbf{x}}_{t+1})). \tag{11}$$

The loss of all patients can be calculated by averaging $\mathcal{L}$ and all parameters in the neural architecture can be jointly optimized in an end-to-end way.

IV. Experiments

A. Experimental Settings

Table I: Statistics of two datasets.

Dataset	DPH	MIMIC-III
# of patients	46,074	7,499
# of visits	447,505	19,911
Avg. # of visits per patient	9.71	2.66
# of unique ICD codes	6,059	4,880
Avg. # of ICD codes per visit	2.42	13.06
Max. # of ICD codes per visit	27	39
# of unique CCS group codes	238	272
Avg. # of CCS group codes per visit	2.32	11.23
Max. # of CCS group codes per visit	24	34
# of top-level codes	17	17

1) Datasets: We conduct experiments on two real-world EHR datasets (Statistics are shown in Table I):

DPH Dataset consists of medical records of 46,074 patients collected by Peking University People's Hospital from 2009 to 2014. We filter out sequences that are too short in length. Only patients with at least 5 visits are preserved. It helps evaluate how prediction models perform on long diagnosis sequences.

MIMIC-III Dataset[4] is a public EHR dataset containing medical records of 7,499 ICU patients over 11 years. We only choose patients with at least two visits. Since MIMIC-III consists of very short visits and the number of patients is small, it helps evaluate the performance of prediction approaches on high-risk patients with insufficient training data.

We group the ICD codes with CCS single-level diagnosis grouper[5] and replace the original ICD codes with their group codes following [5]. We use CCS-multi-level diagnosis hierarchy[6] as the taxonomy of diseases.

2) Models for comparison: We select six competitive models for comparison, which can be divided into three groups.

G1: Models that utilize only historical records. Models in G1 handle diagnosis sequences without incorporating auxiliary information, including **RNN**, **RNN+** [4], and **Dipole** [4].

G2: Demographics-aware model. The model **MCA-RNN** [2] utilizes patient demographics in diagnosis prediction.

G3: Taxonomy-aware models. Models in G3 incorporate taxonomies of diseases for diagnosis prediction, which consist of **GRAM** [5] and **KAME** [6].

3) Evaluation Criteria: The following criteria are used.

Recall@K is defined as the number of correct codes in top K of $\tilde{\mathbf{x}}_{t+1}$ divided by the number of all correct codes.

MAP@K (mean average precision) is another widely used metric that considers the orders of correctly predicted codes.

We vary K in $\{5, 10, 15\}$ for a more thorough evaluation.

4) Implementation Details: We treat visits of each patient as a sample and randomly split the dataset into training (75%), validation (10%) and testing (15%) sets as [6]. We report

[4]https://mimic.physionet.org/

[5]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt

[6]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixCMultiDX.txt

Table II: Performance of models on two datasets. Best results are highlighted in bold. The symbol * means that the improvement is significant with p-value < 0.001 by t-test.

Dataset	Group	Method	Recall@K			MAP@K		
			K=5	K=10	K=15	K=5	K=10	K=15
DPH	**G1**	RNN	0.672±0.002	0.769±0.002	0.814±0.001	0.547±0.001	0.577±0.001	0.585±0.001
		RNN+	0.667±0.001	0.765±0.002	0.811±0.001	0.542±0.001	0.572±0.001	0.581±0.001
		Dipole	0.665±0.002	0.756±0.002	0.799±0.002	0.542±0.003	0.570±0.002	0.578±0.002
	G2	MCA-RNN	0.677±0.002	0.783±0.001	0.824±0.002	0.549±0.002	0.581±0.001	0.588±0.002
	G3	GRAM	0.679±0.002	0.778±0.001	0.822±0.001	0.554±0.001	0.583±0.001	0.591±0.001
		KAME	0.680±0.002	0.777±0.001	0.821±0.001	0.556±0.002	0.585±0.002	0.593±0.002
	Ours	CAMP	**0.694±0.001***	**0.791±0.001***	**0.833±0.001***	**0.567±0.002***	**0.596±0.001***	**0.604±0.001***
MIMIC-III	**G1**	RNN	0.288±0.002	0.432±0.002	0.528±0.002	0.245±0.002	0.333±0.001	0.380±0.002
		RNN+	0.289±0.002	0.431±0.003	0.527±0.003	0.247±0.002	0.334±0.002	0.381±0.003
		Dipole	0.282±0.002	0.423±0.002	0.520±0.002	0.239±0.002	0.323±0.002	0.370±0.002
	G2	MCA-RNN	0.291±0.001	0.438±0.001	**0.539±0.001**	0.248±0.002	0.340±0.002	0.391±0.001
	G3	GRAM	0.293±0.002	0.437±0.002	0.535±0.003	0.252±0.002	0.341±0.002	0.389±0.002
		KAME	0.292±0.002	0.438±0.002	0.535±0.002	0.249±0.002	0.339±0.002	0.387±0.003
	Ours	CAMP	**0.297±0.001***	**0.443±0.001***	**0.539±0.002**	**0.256±0.001***	**0.347±0.001***	**0.396±0.001***

performance according to predictions for the last visit of patients in the testing set. For fair consideration, all models are optimized using Adam [18] with an initial learning rate of 0.001 and the batch size is fixed to 100. The coefficient of L_2 norm regularization is fixed to 0.001. The size r of patient demographics vector is 7 (2 genders + 5 age groups) in the DPH dataset and 11 (2 genders + 5 age groups + 4 admission types) in the MIMIC-III dataset. We tune hyper-parameters of models on the validation set. We initialize $\mathbf{V}$ following [16]. Each experiment is repeated ten times and we report the average and standard deviation as the result. To ensure reproducibility, detailed instructions on running our model has been provided along with the source code.

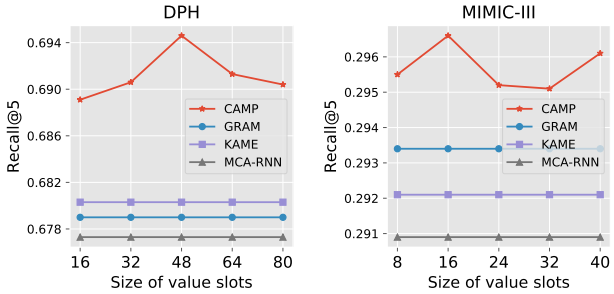

Figure 2: Recall@5 of CAMP and baselines on two datasets with different sizes of memory value slots (d_v).

B. Performance Comparison

The diagnosis prediction results of CAMP and all six baselines on two datasets are given in Table II. We further conduct paired t-tests showing whether the improvements of CAMP are statistically significant (e.g., p-value < 0.001). Three observations are made from Table II.

First, our model CAMP outperforms all state-of-the-art models. On the DPH dataset, CAMP achieves 2.1% higher Recall@5 and 2.0% higher MAP@5 over all baselines. Moreover, the improvements in terms of most criteria are statistically significant. This demonstrates the effectiveness of our proposed framework of co-attention memory networks, which allows CAMP to capture complicated patient health conditions from diagnosis sequences. The superiority of CAMP also stems from its design that jointly models the mutual effect between important context (i.e., medical knowledge and patient demographics) and historical records while baselines fail to do so. We also observe that the overall improvement of CAMP on the DPH dataset is more significant than that on the MIMIC-III dataset. This is ascribed to the fact that the average length of diagnosis sequences on the former is much longer than the latter, which makes it easier to highlight the strength of memory networks in handling long sequences. Even for the MIMIC-III dataset with short sequences, our method also achieves stable accuracy gain compared with the baselines.

Second, demographics-aware model (MCA-RNN) performs better than the models that consider only historical records (RNN, RNN+ and Dipole), achieving 1.4% higher Recall@K and 1.6% higher Map@K on the MIMIC-III dataset. It demonstrates that patient demographics are important contextual information for modeling the diagnosis sequences and help improve the prediction performance.

Third, taxonomy-aware models (**G3**) generally achieve better performance than the models that consider only historical records. The mean improvements of Recall@K on two datasets are 1.1% and 1.4%. The mean improvements of MAP@K on two datasets are 1.5% and 2.1% respectively. We ascribe these improvements to the fact that they learn better disease embeddings that capture intrinsic characteristics with medical knowledge and thus predict future diagnosis more accurately.

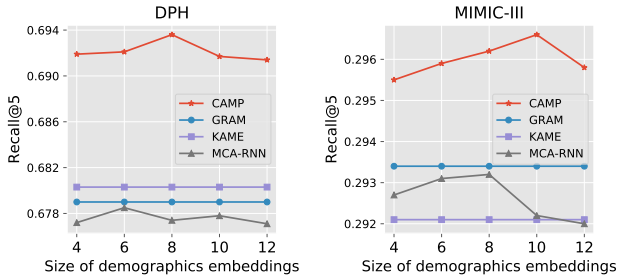

Figure 3: Recall@5 of CAMP and baselines on two datasets with different sizes of demographics embeddings (d_3).

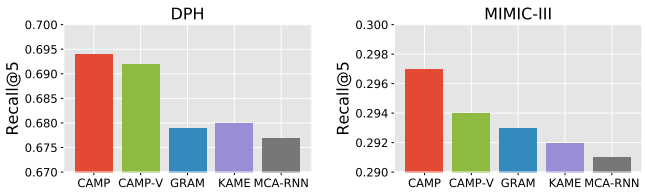

Figure 4: Recall@5 of CAMP, CAMP-V (the variant without co-attention-based mutual enhancement) and baselines.

C. Detailed Analysis of CAMP

In this section, we study the effect of the external memories and the demographics embeddings by varying the size of memory value slots d_v and the size of demographics embeddings d_3. We further study the effect of co-attention-based mutual enhancement by comparing CAMP with one variant. Three most competitive baselines are selected for comparison. Due to space limitations, we only show the results of Recall@5. The results regarding other criteria are similar. Based on the observation, we draw following conclusions.

Effectiveness of the external memories. Figure 2 shows that our method achieves the best performance on DPH (or MIMIC-III) when d_v is equal to 48 (or 16). Smaller d_v leads to insufficient representation of fine-grained patient conditions and larger d_v may result in over fitting. This demonstrates the importance of using external memories, which contain fine-grained information about long-term patient health conditions regarding each category of diseases.

Effectiveness of and the demographics embeddings. Figures 3 shows that too small or too large values of d_3 will hurt the prediction performance of CAMP. This illustrates the importance of properly encoding patient demographics.

Effectiveness of co-attention-based mutual enhancement. We compare CAMP with one variant: CAMP-V. which disables the interaction between patient demographics and the memory network. In CAMP-V, $\mathbf{q}$ (instead of $\mathbf{q}^{\mathbf{new}}$) is utilized in the final prediction layer. The memory attention is calculated by using Equation (2) instead of Equation (7). As shown in Figure 4, CAMP performs better than CAMP-V on two datasets, which confirms our assumption that modeling the mutual effects between long-term patient conditions and demographics results in improved prediction performance.

V. Conclusions

In this paper, we propose a model named co-attention memory networks (CAMP) for diagnosis prediction. The model adopts a three-way interaction architecture to tightly integrate historical records, fine-grained patient conditions, and demographics. The analysis of fine-grained patient conditions is enabled by explicitly incorporating taxonomies of diseases into a memory network. We elaborately design the memory network to ensure that it cooperates with patient demographics in a mutual enhancement manner. Experiments on real-world datasets demonstrate that CAMP consistently performs better than state-of-the-art methods.

VI. Acknowledgment

This work is supported by the National Science and Technology Major Project (No. 2018ZX10201002).

References

[1] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *NIPS*, 2016, pp. 3504–3512.

[2] W. Lee, S. Park, W. Joo, and I.-C. Moon, "Diagnosis prediction via medical context attention networks using deep generative modeling," in *ICDM*. IEEE, 2018, pp. 1104–1109.

[3] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016, pp. 301–318.

[4] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *KDD*. ACM, 2017, pp. 1903–1911.

[5] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: graph-based attention model for healthcare representation learning," in *KDD*. ACM, 2017, pp. 787–795.

[6] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "Kame: Knowledge-based attention model for diagnosis prediction in healthcare," in *CIKM*. ACM, 2018, pp. 743–752.

[7] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *ICLR*, 2015.

[8] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *AAAI*, 2018, pp. 4091–4098.

[9] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *KDD*. ACM, 2016, pp. 1495–1504.

[10] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," *arXiv preprint arXiv:1606.03126*, 2016.

[11] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *ICML*, 2016, pp. 1378–1387.

[12] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, and H. Zha, "Sequential recommendation with user memory networks," in *WSDM*. ACM, 2018, pp. 108–116.

[13] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *Computer Science*, 2014.

[14] C. Xiao, T. Ma, A. B. Dieng, D. M. Blei, and F. Wang, "Readmission prediction via deep contextual embedding of clinical concepts," *PloS one*, vol. 13, no. 4, p. e0195024, 2018.

[15] F. Liu and J. Perez, "Gated end-to-end memory networks," in *EACL*, vol. 1, 2017, pp. 1–10.

[16] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in *WWW*, 2017, pp. 765–774.

[17] L.-Y. Chang, T.-Y. Lin *et al.*, "Clinical features and risk factors of pulmonary oedema after enterovirus-71-related hand, foot, and mouth disease," *The Lancet*, vol. 354, no. 9191, pp. 1682–1686, 1999.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.