Temporal Self-Attention Network for Medical Concept Embedding

Xueping Peng*, Guodong Long*, Tao Shen*, Sen Wang†, Jing Jiang*, Michael Blumenstein*
* Centre for Artificial Intelligence, FEIT, University of Technology Sydney, Australia
† School of Information Technology and Electrical Engineering, The University of Queensland, Australia
Email: {xueping.peng, guodong.long}@uts.edu.au, Tao.Shen@student.uts.edu.au,
sen.wang@uq.edu.au, {jing.jiang, Michael.Blumenstein}@uts.edu.au

***Abstract*—In longitudinal electronic health records (EHRs), the event records of a patient are distributed over a long period of time and the temporal relations between the events reflect sufficient domain knowledge to benefit prediction tasks such as the rate of inpatient mortality. Medical concept embedding as a feature extraction method that transforms a set of medical concepts with a specific time stamp into a vector, which will be fed into a supervised learning algorithm. The quality of the embedding significantly determines the learning performance over the medical data. In this paper, we propose a medical concept embedding method based on applying a self-attention mechanism to represent each medical concept. We propose a novel attention mechanism which captures the contextual information and temporal relationships between medical concepts. A light-weight neural net, "Temporal Self-Attention Network (TeSAN)", is then proposed to learn medical concept embedding based solely on the proposed attention mechanism. To test the effectiveness of our proposed methods, we have conducted clustering and prediction tasks on two public EHRs datasets comparing TeSAN against five state-of-the-art embedding methods. The experimental results demonstrate that the proposed TeSAN model is superior to all the compared methods. To the best of our knowledge, this work is the first to exploit temporal self-attentive relations between medical events.**

I. Introduction

A healthcare information system (HIS) stores huge volumes of Electronic Health Records (EHRs) that contain detailed visit information about patients over a period of time [1]. The EHRs data is a multi-layer structure composed of three layers: patient, visit, and medical concept. For instance, an anonymous patient in Fig. 1 makes three visits in different days. The first and third visits recorded a diagnosis of six health conditions (denoted by diagnosis codes, e.g., ICD 585.5) while the second visit reports five disorders. A patient's healthcare journey (referred to hereafter as "patient journey"), can thus be represented by a sequence of visits occurring at different time-stamps. To standardize the healthcare procedure, medical concepts (referred to in this paper as "diseases") in each visit record are converted to an item in a standard coding system (e.g., International Classification of Diseases or ICD[1]). A medical coding system is often developed according to disease ontology and represented by a hierarchical structure, which is practical for human understanding and maintenance. This tree-based coding system includes basic medical taxonomy knowledge which can be embedded into a unified learning framework to achieve better classification performance and interpretability. In the light of this idea, a medical concept embedding method for subsequent learning tasks is highly desirable.

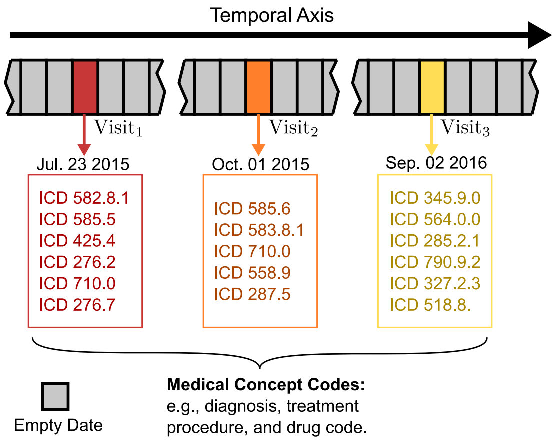

Fig. 1. An example segment of one patient's healthcare journey

Intuitively, one-hot encoding of medical concepts simply generates a binary vector that is high-dimensional and sparse. An alternative solution, inspired by Natural Language Processing (NLP), is to use word embedding approaches to learn a low-dimensional dense representation of medical concepts [2]–[5]. This method has been used in various AI-based healthcare applications [6]–[11] to improve performance. However, there are two major limitations. First, even though there is a similar multi-layer structure in a textual corpus (document, sentence, and word) compared to EHRs, intrinsic differences are still evident. For instance, two consecutive sentences in one document only have a sequential relationship, while two arbitrary visits in one patient journey may be separated by different time intervals, which is an important factor in longitudinal studies. In other words, the period of time between two visits, which have been largely disregarded in the existing works on medical concept embedding, can be modeled as auxiliary information fed into the supervised algorithms. Also, a sentence may include repeated words, whereas the medical concept in the visit is unique. Hence, the existing NLP models, such as word embedding and sentence embedding, cannot

[1] http://www.icd9data.com

be directly applied to encode the medical concepts without information loss.

Second, tree-based embedding methods cannot reflect the complex relationships between each unit of the medical concepts because of the hierarchical representation. In an EHRs dataset, there are many complicated sequential co-occurrence relationships between medical concepts that contain much richer information than tree-based taxonomy. For example, chronic kidney disease (585.5) and end stage renal disease (585.6) are separately encoded in ICD9. Both medical records and pathology support the fact that these two diseases are temporally correlated. In other words, chronic kidney disease often eventually leads to end stage renal disease. Therefore, an encoding method that considers temporal information between medical concepts over time significantly benefits prediction tasks in healthcare analytics. References [12], [13] proposed a multi-level representation learning that simultaneously incorporates visits and medical concepts using the sequential order of visits and the co-occurrence of medical concepts. Reference [14] proposed a CBOW-based medical concept embedding method enhanced by an attention mechanism to capture the temporal relation between visits. In particular, the temporal sequence of patient visits has been split into many time units (e.g., week, month, and year) so that the attention mechanism can capture the sequential information as well as the time-aware information. However, a fixed size of time units is impractical because a different diagnosis or treatment might have different awareness of time. Moreover, large time units may cause information loss because it puts several visits into one time unit. Furthermore, the time intervals between visits are used as quantitative scalars to segment time units [14] and quantify the attribute relevance [15]. Although improved performance is achieved by using the quantitative scalars on healthcare analysis tasks, these aspects of time-aware methods are arbitrary and unsmooth.

To overcome the aforementioned limitations and consider the representation of time intervals, we propose a novel attention mechanism, called "Temporal Self-Attention (TeSA)", for temporal context fusion. In particular, we first transform the time intervals as vectors whose dimension is the same as that of the embedded medical concept. Therefore, it is more expressive and smoother than a time scalar for capturing temporal relationships in medical concepts. Then a proposed self-attention mechanism is utilized to capture the contextual information and temporal interval between medical concepts in context, and to apply a feature fusion gate to combine the attentive outputs with the original inputs to produce the final context-aware representations of all the medical concepts. A light-weight neural network based on TeSA, called "Temporal Self-Attention Network (TeSAN)", is also developed. TeSAN uses attention pooling to compress the output of TeSA into a vector representation. In experiments, we compare TeSAN with the state-of-the-art methods in both unsupervised and supervised learning tasks, which are clustering (i.e. nearest neighbour search) and mortality tasks, respectively. TeSAN achieves the highest normalized mutual information (NMI) and Precision at 1 (P@1) on two public medical data sets, MIMIC III and CMS, and obtains the best performance of PR-AUC and ROC-AUC for the mortality prediction task on MIMIC III data.

The remainders of this paper are organized as follows. Section II reviews related studies. In Section III, we briefly discuss some preliminary, and details about our model are presented in Section IV. In Section V, we demonstrate the experimental results conducted on two public datasets. Lastly, we conclude our study in Section VI.

II. Related Work

A. Word Embedding

Although word embedding was first introduced by Rumelhart et al. [16] in 1986, distributed representation learning of words with neural networks has only become a hot research topic since 2003 [2]–[5], [17], [18]. CBOW and the Skip-gram model [4], [17] are among two of the model families that were introduced to compute continuous vector representations of words from very large datasets. Each is based on the assumption that the order of words or a word's context do not influence the projection of the target word. However, some scholars have recently discovered that sequence and context do matter. For example, Melamud et al. [19] explored the impact of context with the Skip-gram model, finding that weighting for context improves performance with extrinsic tasks. Similarly, Liu et al. [20] showed that conditioning a target word on a subset of contexts improves both the quality of the embedding and the predictions. Ling et al. [21] extended CBOW by incorporating an attention model that considers contextual words and their positions relative to the predicted word, which results in better representations. Each of these advancements has proven effective in the field of NLP but, as discussed in Section I, the differences between documents and patient journeys mean these embedding models cannot be directly applied to medical concepts in EHRs without information loss or reduced performance.

B. Medical Concept Embedding

Borrowing ideas from word representation models [4], [17], [22], researchers in the healthcare domain have recently explored the possibility of creating representations of medical concepts. Much of this research has focused on the Skip-gram model. For example, Minarro-Gimnez et al. [6] directly applied Skip-gram to learn representations of medical text, and Vine et al. [7] did the same for UMLS medical concepts. Choi et al. [9] went a step further and used the Skip-gram model to learn medical concept embeddings from different data sources, including medical journals, medical claims, and clinical narratives. In other work [13], Choi et al. developed the Med2Vec model based on Skip-gram to learn concept-level and visit-level representations simultaneously. The shortcoming of all these models is that they view EHRs as documents in the NLP sense, which means that temporal information is ignored.

Attention mechanisms are a more recent introduction in healthcare analytics [23], [24]. Choi et al. [10] proposed a graph-based attention model that learns representations of medical concepts from medical ontologies. Rajkomar et al. [15] applied an attention-based time-aware neural network model [25] to predict patient outcomes, and Cai et al. [14] proposed MCE (Medical Concept Embedding) as a way to integrate time information into an attention model to embed medical concepts. Our work departs from Cai et al. [14] and Rajkomar et al. [15] in that TeSAN integrates the time intervals between visits with expressive and multi-dimensional vectors into the context of medical concepts to capture the temporal relationships.

III. Preliminary

This section begins by giving several definitions for medical concepts and targeted tasks. Because of the similarity between word embedding in the natural language processing literature and the code embedding of medical concept in EHRs, we then adopt some of the concepts and approaches designed for NLP tasks to apply to EHRs. We first introduce the concept of word embedding [4] to learn low-dimensional real-value distributed vector representations for medical concepts instead of discrete medical codes for downstream tasks; second, we adopt a sophisticated self-attention mechanism [26] for EHRs to capture the contextual information and temporal dependencies between the medical concepts for the context-aware medical concept representation, to achieve better empirical performance; lastly, the attention pooling [27] technique is leveraged to attentively select important elements from a set of input code embeddings, which is aimed at sequence compression or embedding via parameterized weighted sum.

A. Definitions

Definition 1 (Medical Concept): A medical concept is defined as a term or code to describe diagnosis, procedure, medication, and laboratory tests for an inpatient during a treatment process. We denote the set of medical concepts (e.g., ICD 585.5 for diagnosis, CPT 2001 for procedure) as C.

Definition 2 (Visit): A visit for an inpatient refers to a treatment process from admission to discharge, including an admission time stamp. We denote a visit as $V_{i,j} =< \boldsymbol{x}_{i,j}, \boldsymbol{t}_{i,j} >$, where *i* is the *i*-th patient, *j* the *j*-th visit of the patient, $\boldsymbol{x}_{i,j} = [x_1^{i,j}, x_2^{i,j}, ..., x_K^{i,j}]$, $\boldsymbol{t}_{i,j} = [t_1^{i,j}, t_2^{i,j}, ..., t_K^{i,j}]$, *K* is the number of medical concepts in a visit, $x_k^{i,j}$ is a medical concept and $t_k^{i,j}$ is admission time, where $k \in \{1, 2, \ldots, K\}$.

Definition 3 (Patient Journey): A patient journey consists of a sequence of visits over time, which is denoted as $J_i = [V_{i,1}, V_{i,2}, ..., V_{i,M}]$ where *M* is the total number of visits for patient *i*.

Definition 4 (Temporal Interval): Temporal interval refers to difference in days between two visits in a patient journey, denoted as $\triangle_{jl} = |t_k^{i,j} - t_q^{i,l}|$, where $j, l \in \{1, ..., M\}$ and $k, q \in \{1, ..., K\}$.

Definition 5 (Problem): Given a set of patient journeys *Js*, the problem is to learn an embedding function $f_C : C \rightarrow R^d$ that maps every code in the set of medical concept *C* to a real-value dense vector with dimension *d*.

In this paper, a patient's medical data is stored to a sequence by chronologically concatenating *M* visits in patient journey J_i. We will therefore ignore the indexes *i,j* (which index the patients and their visiting times) for simplification, if it is possible to do so without causing confusion.

B. Medical Concept Embedding

Medical concept embedding is a fundamental processing unit in deep neural network-based EHRs. It transfers each discrete medical concept into a distributed real-value vector representation. Formally, given a sequence or set of medical concepts $\boldsymbol{x} = [x_1, x_2, ..., x_n] \in \mathbb{R}^{|C| \times n}$, where x_i is a one-hot vector, $|C|$ is the vocabulary size of the medical concept codes, and n is the sequence length. A word embedding method (typically in the NLP literature, e.g. word2vec [4], [17]) is applied to the sequence, which outputs a sequence of low dimensional vectors $\boldsymbol{c} = [c_1, c_2, ..., c_n] \in \mathbb{R}^{d \times n}$, where d is the embedding dimension of c_i. This process can be formally written as $\boldsymbol{c} = W^{(e)}\boldsymbol{x}$, where $W^{(e)} \in \mathbb{R}^{d \times |C|}$ is the embedding weight matrix that can be fine-tuned during the training phase. Following the idea of word embedding and context modeling in NLP, this paper is an attempt to embed medical concepts in low-dimensional vectors.

C. Attention Mechanism

1) Vanilla Attention: Given an input context of medical concepts $\boldsymbol{c} = [c_1, c_2, ..., c_n]$ composed of concept embeddings and a vector representation of a query $q \in \mathbb{R}^d$, vanilla attention [23] computes the alignment score between q and each concept c_i using a compatibility function $f(c_i, q)$. A softmax function then transforms the alignment scores $\alpha \in \mathbb{R}^n$ to a probability distribution $p(z|\boldsymbol{c}, q)$, where z is an indicator of which concept is important to q. A large $p(z = i|\boldsymbol{c}, q)$ means that c_i contributes important information to q. This attention process can be formalized as

$$\alpha = [f(c_i, q)]_{i=1}^n, \tag{1}$$

$$p(z|\boldsymbol{c}, q) = softmax(\alpha). \tag{2}$$

The output s is the weighted average of sampling a concept according to its importance, i.e.,

$$s = \sum_{i=1}^{n} p(z = i|\boldsymbol{c}, q) \cdot c_i. \tag{3}$$

Additive attention [23], [28] is commonly-used attention mechanism in which the compatibility function $f(\cdot)$ is parameterized by a multi-layer perceptron (MLP), i.e.,

$$f(c_i, q) = w^T \sigma(W^{(1)} c_i + W^{(2)} q + b^{(1)}) + b, \tag{4}$$

where $W^{(1)} \in \mathbb{R}^{d \times d}$, $W^{(2)} \in \mathbb{R}^{d \times d}, w \in \mathbb{R}^d$ are learnable parameters, and $\sigma(\cdot)$ is an activation function. In contrast to additive attention, multiplicative attention [29], [30] uses cosine similarity as the compatibility function for $f(x_i, q)$, i.e.,

$$f(x_i, q) = \langle\ W^{(1)} x_i,\ W^{(2)} q\ \rangle, \tag{5}$$

In practice, although additive attention is expensive in time cost and memory consumption, it usually achieves better empirical performance for downstream tasks.

To improve the context modeling capability of the attention module, a **multi-dimensional (multi-dim) attention mechanism** [31] that uses a feature-wise alignment score has recently been proposed. The alignment score from the attention compatibility function is computed for each feature; the score of a concept pair is a vector rather than a scalar, so the score might be large for some features but small for others to model more subtle context and dependency relationship. Formally, $P_{ki} \triangleq p(z_k = i|\boldsymbol{c}, q)$ denotes the attention probability of i-th element on k-th feature dimension, where the attention score is obtained from a multi-dim compatibility function by replacing the weight vector w with a weight matrix in Eq.(4). For simplicity, we ignore the subscript k if this does not cause confusion. The attention result can be written as $s = \sum_{i=1}^{n} P_{\cdot i} \odot c_i$. In the remaining paper, we use the multi-dim compatibility function by default for rich expressive power and better performance for downstream tasks.

2) Self-attention mechanism: The self-attention mechanism [26], [31]–[33] can produce context-aware representations by exploring the contextual relationships between two medical concepts c_i and c_j from the same context $\boldsymbol{c}$. It is naturally compatible with medical concept embedding because unlike the commonly-used recurrent neural network, the self-attention mechanism is order-insensitive, making it suitable for all the medical concepts in a single patient visit. The query q in the attention compatibility function (e.g., multi-dim compatibility function) is replaced by c_j , i.e.,

$$f(c_i, c_j) = W^T \sigma(W^{(1)} c_i + W^{(2)} c_j + b^{(1)}) + b. \tag{6}$$

Similar to P in multi-dim attention, each input medical concept c_j is associated with a probability matrix P_j such that $P^j_{ki} \triangleq p(z_k = i|\boldsymbol{c}, c_j)$. The output representation for each c_j is

$$s_j = \sum_{i=1}^{n} P^j_{\cdot i} \odot c_i \tag{7}$$

The final output of self-attention is $\boldsymbol{s} = [s_1, s_2, \dots, s_n]$, each of which is the medical-concept-context embedded representation for each medical concept. However, a fatal defect in previous self-attention mechanisms applied to NLP tasks is that they cannot model the relative time interval between the medical concepts from different patient visits, even if equipped with positional encoding [26].

3) Attention Pooling: Attention pooling [27], [34] explores the importance of each medical concept to the entire context given a specific task. This is used to compress a sequence of medical concept embeddings from a visit or a patient to a single context-aware vector sequence embedding for downstream classification or regression. In particular, q is removed from the common compatibility function which is formally written as the following equation.

$$f(c_i) = W^T \sigma(W^{(1)} c_i + b^{(1)}) + b. \tag{8}$$

The multi-dim attention probability matrix P is defined as $P_{ki} \triangleq p(z_k = i|\boldsymbol{c})$. The final output of the attention pooling, which is used as the sequence encoding, has a similar form as the aforementioned attention mechanism, i.e.,

$$s = \sum_{i=1}^{n} P_{\cdot i} \odot c_i \tag{9}$$

IV. Proposed Model

We first introduce the "temporal self-attention (TeSA)" as a fundamental self-attention module. Then, we present the "temporal self-attention network (TeSAN)" for medical concept embedding, which uses TeSA as its context fusion module. Table I lists the notations used in the study.

TABLE I
NOTATIONS FOR HISANTH.

Notation	Description
C	Set of unique medical concepts
$\|C\|$	The number of unique medical concepts
$V_{i,j}$	The *j*-th visit of the *i*-th patient
$\boldsymbol{x}_{i,j}$	Set of medical concepts in $V_{i,j}$
$x^{i,j}_k$	The *k*-th medical concept in $\boldsymbol{x}_{i,j}$, $k \in \{1, \dots, K\}$
$\boldsymbol{c}_{i,j}$	Set of medical concept embeddings in $\boldsymbol{x}_{i,j}$
$c^{i,j}_k$	The *k*-th medical concept embedding in $\boldsymbol{c}_{i,j}$, $k \in \{1, \dots, K\}$
$\boldsymbol{J}_i$	A patient journey consisting of a sequence of visits over time
$\triangle$	number of days between two visits in a patient journey
d	The embedding dimension
K	The number of medical concepts in a visit
M	The total number of visits for a patient

A. Temporal Self-attention

As discussed in the previous section, the self-attention mechanism is unlike the commonly-used recurrent neural network which is order-insensitive and suitable for the medical concepts in a single patient visit. However, a flattened patient journey is a sequence of medical concepts with time stamps, so previous self-attention mechanisms applied to NLP tasks cannot model the relative time interval between the medical concepts from different patient visits. Inspired by previous work on masked self-attention [31], [35], which achieves state-of-the-art performance on many NLP tasks, we propose a novel attention mechanism, called "Temporal self-attention (TeSA)", in which the attention mechanism captures the contextual information and temporal relationships between medical concepts.

Temporal self-attention is composed of a self-attention block to explore the contextual relationship and temporal interval and a fusion gate to combine the output and input of the attention block. Its structure is shown in Figure 2. We rewrite the self-attention in Eq.(6) as a temporal-dependent format:

$$f(c_i, c_j, \triangle_{ij}) = W^T \sigma(W^{(1)} c_i + W^{(2)} c_j + W^{(3)} e_{\triangle_{ij}} + b^{(1)}) + b \tag{10}$$

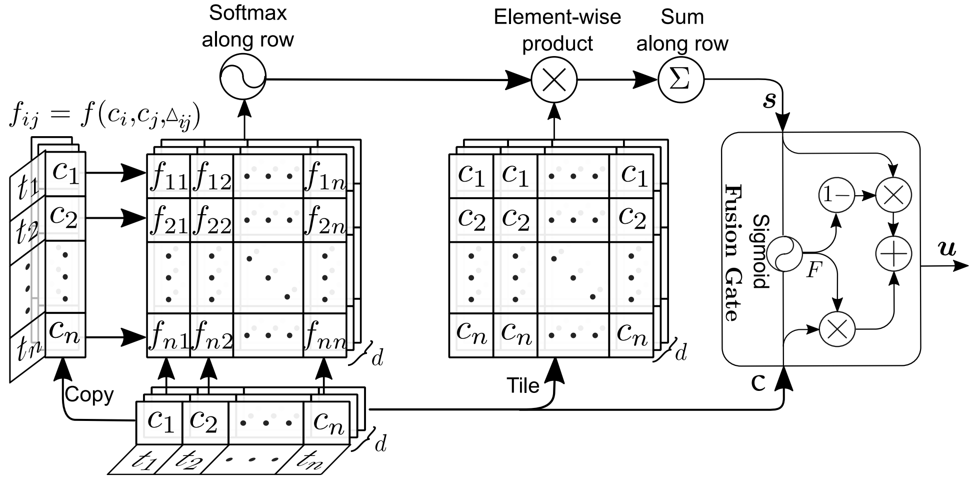

Fig. 2. Temporal self-attention mechanism. The inputs are embedded medical concepts $(c_1, c_2, \ldots, c_n)$ with corresponding visit time stamps $(t_1, t_2, \ldots, t_n)$; $\triangle_{ij}$ can be obtained by giving t_i and t_j. f_{ij} is formally defined as Eq.(10); Softmax along row produces a probability distribution for each element of the multi-dim embedded medical concepts; Element-wise product outputs the weighted element-wise multi-dim medical concepts; Sum along row produces the weighted multi-dim medical concepts whose dimension size is the same as the size of the input of embedded medical concepts. The fusion gate merges the weighted output $\boldsymbol{s}$ and the input of embedded medical concepts $\boldsymbol{c}$ to produce output $\boldsymbol{u}$.

where $\triangle_{ij}$ is the temporal days' interval between t_i and t_j as defined in Def. (4), and $e_{\triangle_{ij}} \in \mathbb{R}^d$ is the temporal interval embedding, which is a learnable parameter. A temporal interval embedding layer is added before $e_{\triangle_{ij}}$ is taken as input to the TeSA module and the size of the embedding matrix is $\mathbb{R}^{n_{days} \times n}$, where n_{days} is the number of days when the dataset spans.

Given input context $\boldsymbol{c}$ and a temporal interval matrix $\triangle$, we compute $f(c_i, c_j, \triangle_{ij})$ according to Eq.(10), and follow the standard procedure of self-attention to compute the probability matrix P_j for each $j \in [n]$. Each output s_j in $\boldsymbol{s}$ is computed as in Eq.(7).

The final output $\boldsymbol{u} \in \mathbb{R}^{d \times n}$ of TeSA is obtained by combining the output $\boldsymbol{s}$ and the input $\boldsymbol{c}$ of the temporal self-attention block. This yields an encoded temporal interval and a context-aware vector representation for each medical concept. The combination is accomplished by a dimension-wise fusion gate, i.e.,

$$F = sigmoid(W^{(f_1)}s + W^{(f_2)}c + b^{(f)}) \tag{11}$$

$$u = F \odot s + (1 - F) \odot c \tag{12}$$

where $W^{(f1)}, W^{(f2)} \in \mathbb{R}^{d \times d}$ and $b^{(f)} \in \mathbb{R}^d$ are the learnable parameters of the fusion gate.

B. Temporal Self-attention Network

We propose a light-weight network, "Temporal Self-Attention Network (TeSAN)" for medical concept embedding. Its architecture is shown in Figure 3.

Given an input sequence of concept representation $\boldsymbol{c}$, which is from concatenated visits in one patient journey, TeSAN first applies the TeSA block to capture the contextual relationship and temporal interval information. The multi-dimensional attention pooling block takes the TeSA output as input to produce $h_i \in \mathbb{R}^d$ computed by Eq. 8 and 9. The context embedding result of TeSAN is exploited to predict target concept c_i.

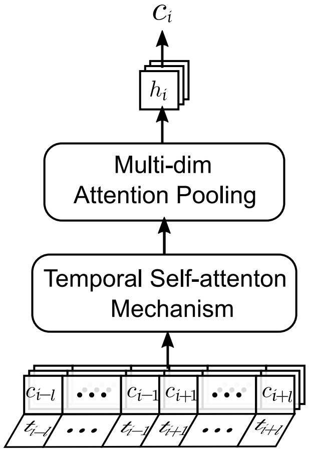

Fig. 3. Temporal self-attention network (TeSAN). Note that c_i and t_i are from concatenated visits in one patient journey, and l is the size of the skip window.

C. Loss Function

The loss function is inspired by the Word2Vec [4], [17] model by using negative sampling to maximize

$$J = \log \sigma({c_i}^T h_i) + \sum_{j=1}^{r} \mathbb{E}_{c_j \sim P(c)}[\log \sigma(-{c_j}^T h_i)], \tag{13}$$

where σ is a Sigmoid function, r is the number of negative samples, and $P(c)$ is the noise distribution [4].

V. Experiments

The proposed model is evaluated on two public datasets via the unsupervised and prediction tasks. The source code of TeSAN is available at https://github.com/Xueping/tesan/.

A. Dataset Description

We conducted comparative studies on two public datasets listed as follows:

- **MIMIC III** [36] is an open-source, large-scale, de-identified EHRs data set consisting of clinical logs of patients admitted to intensive care units with serious conditions. The diagnosis codes in this dataset follow the ICD9 standard. The statistics of the dataset are provided in Tab. II.
- **CMS** is a publicly available[2] synthetic claims dataset, which includes four types of files: inpatient, outpatient, carrier and beneficiary summary. For our experiment, we chose only a subset of inpatient files between 2008 and 2010 as one of our two datasets. The basic statistical information is shown in Table II.

TABLE II
STATISTICS OF DATASETS.

Datasets	MIMIC III	CMS
# of patients	46,520	755,214
# of visits	58,976	1,332,822
Avg. # of visits per patient	1.27	1.76
# of unique diagnosis codes	6,985	7,873
# of unique procedure codes	2,032	10,726

B. Tasks of Clustering and Nearest Neighbour Search

1) Ground Truth: Two clustering and nearest neighbour search (NNS) [14] tasks were conducted to evaluate the quality of the medical concept embedding results. We selected the ground truth by using two well-organized ontologies, the ICD9 standard and Clinical Classifications Software (CCS)[3]. The ICD9 standard has a hierarchical structure [37] consisting of 19 categories. We used the high level nodes as the clustering labels. We obtained 19 categories for the MIMIC III and CMS datasets. Medical concepts under the same subroot were considered as near neighbours for the nearest neighbour search. We obtained 555,873 near neighbour pairs for MIMIC III and 869,144 for CMS. This ground truth set is named **ICD**. CCS provides a way to classify diagnoses and procedures into a limited number of categories by aggregating individual ICD9 codes into broad diagnosis and procedure groups to facilitate statistical analysis and reporting[4]. CCS aggregates ICD9 diagnosis codes into 285 mutually exclusive categories. For clustering, we obtained 265 categories for MIMIC III and 267 for CMS. For the nearest neighbour search, we obtained 61,630 near neighbour pairs for MIMIC III and 89,546 for CMS. We refer to this ground truth set as **CCS**.

2) Baseline Methods: We compared our model with five baseline models that are state-of-the-art embedding methods as listed below. All baseline models were trained with their source codes.

- **CBOW** [17] learns the representations by averaging the context within a sliding window to predict the target vector.
- **Skip-gram (Sg)** [17] predicts the target vector based on context, using each target word as an input to predict words within that context.
- **GloVe** [13] An unsupervised learning algorithm for obtaining vector representations for words.
- **med2vec** [13] A multi-level embedding model for simultaneously embedding medical concepts and visits.
- **MCE** [14] A CBOW model with time-aware attention model to embed medical concepts with temporal information.
- **TeSAN** Our proposed temporal self-attention network for medical concept embedding to capture the contextual relationship and temporal interval.

3) Experimental Set-Up: All infrequent medical concepts were removed and the threshold empirically set to 5. Patients whose number of hospital visits was less than 4 in CMS were empirically discarded. Following the original Word2vec [4], [17], the same negative sampling strategy as used in Skip-gram, CBOW and TeSAN, and the number of negative samples in MIMIC III and CMS was set to 10 and 5 respectively. All models were trained with 30 epochs for MIMIC III and 20 epochs for CMS. The dimension d of the medical concept embedding was set to 100. The batch size is 64 for MIMIC III and 128 for CMS.

4) Results: We used the clustering and nearest neighbour search tasks to evaluate the embedding results on two public datasets: MIMIC III and CMS. We chose K-Means as the clustering algorithm, and used clustering performance indicator called Normalized Mutual Information (NMI), to evaluate the learned representations for the medical concepts. The skip window of our model was empirically set to 6 for MIMIC III and 7 for CMS. We used the two ground truth sets to evaluate the embedding performance of the proposed model and other baselines.

TABLE III
CLUSTERING PERFORMANCE (NMI) OF THE MODELS ON TWO DATASETS W.R.T. GROUND TRUTHS, ICD AND CCS (%).

Model	MIMIC III		CMS	
	ICD	CCS	ICD	CCS
CBOW	25.34	53.09	08.52	41.70
Sg	26.02	52.97	07.65	35.61
GloVe	17.68	46.57	06.50	33.45
med2vec	5.25	33.65	3.69	17.66
MCE	8.26	37.37	04.27	31.88
TeSAN	**32.84**	**58.33**	**14.69**	**45.63**

a) Overall Performance: Normalized mutual information for clustering performance is reported in Table III, and precision@1 (P@1) for NNS is shown in Table IV, where we highlight the best results. From the two tables, we find that the TeSAN model obtains the best performance in medical concept embedding compared to most state-of-the-art models on medical concept embedding. Our model outperformed the

[2] https://www.cms.gov
[3] https://www.hcup-us.ahrq.gov
[4] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf

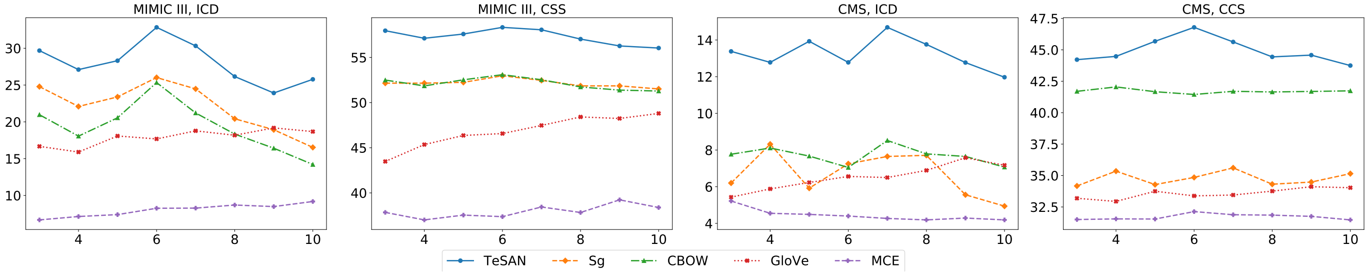

Fig. 4. NMI (%) of clustering performance on two datasets w.r.t. two ground truths, ICD and CCS. The window size varies from 3 to 10.

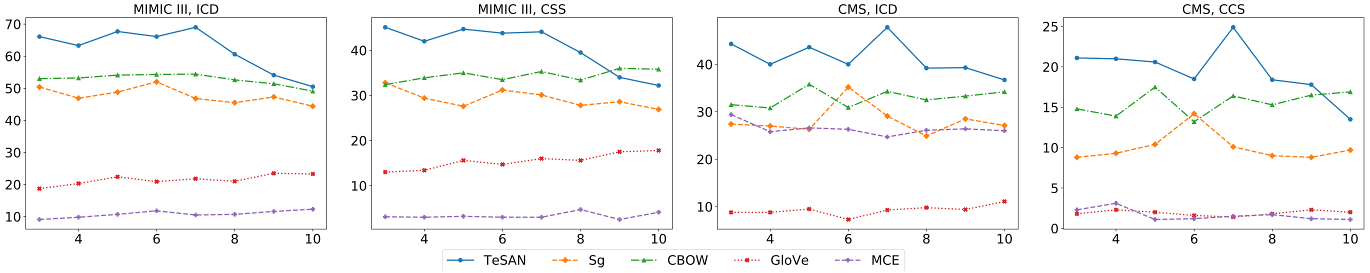

Fig. 5. P@1 (%) of NNS performance on two datasets w.r.t. two ground truths, ICD and CCS. The window size varies from 3 to 10.

TABLE IV
NNS PERFORMANCE (P@1) OF THE MODELS ON TWO DATASETS W.R.T. GROUND TRUTHS, ICD AND CCS (%).

Model	**MIMIC III**		**CMS**	
	ICD	**CCS**	**ICD**	**CCS**
CBOW	54.3	33.5	34.3	16.4
Sg	52.0	35.2	29.1	10.1
GloVe	20.9	14.7	9.3	1.4
med2vec	11.8	4.8	10.7	2.8
MCE	11.8	3.0	24.7	1.5
TeSAN	**66.1**	**43.8**	**47.8**	**24.9**

best baseline model for NMI by 6.82% on ICD and 5.24% on CCS over MIMIC III, and by 6.17% on ICD and 3.93% on CCS over CMS; for P@1 by 11.8% on ICD and 8.6% on CCS over MIMIC III, and by 13.5% on ICD and 8.5% on CCS over CMS. The superior performance of TeSAN over the other models can be explained by the introduction of the temporal self-attention model and the incorporation of the contextual information and temporal interval from the data, which creates a better learner of the medical concept embeddings.

We find that the performance of med2vec and MCE are the worst in the clustering task and the NNS task in MIMIC III dataset, which indicates the importance of the skip window, since med2vec does not use the skip window and CME uses the skip window based on the weeks of the time unit.

All models achieve better performance on the ground truth of CCS than ICD for the clustering task, whereas the performance of all models on ICD is better than CSS for the NNS task. This might be explained by the fact that each of the well-organized ontologies have particular advantages for different tasks. We also find that the performance of all models on MIMIC III is better than those on CMS. There are two possible reasons: one is that the number of epochs is larger for MIMIC III than for CMS, and the other is that MIMIC III is drawn from real world healthcare data whereas CMS is synthesized data.

b) Performance of varying skip window sizes: To take the effects of the context window on the performance of the baseline and proposed models into consideration, we vary the size of the context window to compare performance. In this work, we only compare the proposed model TeSAN with other baselines; the exception is med2vec, due to lack of a parameter for window size. The window size is adjusted from 3 to 10.

The results on the clustering on both datasets are summarized in Fig. 4. The performance of most models is decreased, as an increase in window size induces noise. However, because GloVe makes use of global co-occurrences and MCE obtains a bigger skip window, neither is sensitive on increasing window size. As the window size is increased, GloVe and MCE achieve better performance with the larger window size. Moreover, the TeSAN model always outperforms the other models in terms of NMI on the MIMIC III dataset and in P@1 on CMS, which demonstrates that the integration of the proposed embedding model captures more comprehensive relationships between medical concepts.

Figure 5 is the summary of results on the NNS task over two datasets. The TeSAN model outperforms the baseline models in terms of P@1 on the ground truth of CCS when the size of the skip window is not more than 8, which demonstrates that the attention mechanism benefits embedding in a smaller window. The performance of TeSAN slowly increases to the highest value and then quickly decreased, whereas the performance of GloVe and MCE is relatively stable over an increasing window size, which follow the same trend as their performance in the clustering task.

c) Ablation Study.: We performed a detailed ablation study to examine the contributions of the proposed model components to unsupervised tasks. There are three replaceable components in this model:

- **Normal_Sa:** we replaced the temporal self-attention module with a normal self-attention module;
- **Interval:** we only considered interval information in the temporal self-attention module;
- **Multi_Sa:** we only considered contextual information in the temporal self-attention module;
- **TeSAN:** is the proposed model.

All models were trained with 30 epochs for MIMIC III and 20 epochs for CMS. The skip window of all models was empirically set to 6 for MIMIC III and 7 for CMS. Table V and VI respectively show the performance for clustering and nearest neighbour search for the ablated models and our proposed model.

From the two tables, we find that the TeSAN model obtains the best performance on medical concept embedding compared to the ablated models. Moreover, we note that Multi_Sa outperforms Normal_Sa, which gives us the confidence to apply multiple dimensional self-attention to learn the representation for medical concepts. It is clear that the Interval model provides comparable information with the learning embeddings of medical concepts to the performance of the Multi_Sa and Normal_Sa model. In particular, TeSAN outperforms the best ablated model for NMI by 1.22% on ICD and 0.3% on CCS over MIMIC III, and by 2.58% on ICD and 1.88% on CCS over CMS. It outperforms the best ablated model for P@1 by 1.3% on ICD and 1.38% on CCS over MIMIC III, and by 6.2% on ICD and 6.3% on CCS over CMS.

TABLE V
CLUSTERING PERFORMANCE (NMI) OF THE MODELS ON TWO DATASETS W.R.T. GROUND TRUTHS, ICD AND CCS (%)

Ablation	**MIMIC III**		**CMS**	
	ICD	**CCS**	**ICD**	**CCS**
Normal_SA	30.66	55.99	11.99	43.13
Interval	30.63	57.37	10.81	42.75
Multi_SA	31.66	58.03	12.11	43.75
TeSAN	**32.84**	**58.33**	**14.69**	**45.63**

TABLE VI
NNS PERFORMANCE (P@1) OF THE MODELS ON TWO DATASETS W.R.T. GROUND TRUTHS, ICD AND CCS (%)

Ablation	**MIMIC III**		**CMS**	
	ICD	**CCS**	**ICD**	**CCS**
Normal_SA	60.3	38.0	37.1	13.3
Interval	64.8	42.0	37.0	16.3
Multi_SA	63.7	40.5	41.6	18.6
TeSAN	**66.1**	**43.8**	**47.8**	**24.9**

The ability of TeSAN to outperform the ablated models benefits from the introduction of the temporal self-attention model and the incorporation of contextual information and the temporal interval from the data, which enables better embeddings of medical concepts to be learnt.

C. Mortality Prediction Task

We predicted impending inpatient death, defined as the latest discharge disposition of “hospital expire” [38]–[40]. Note that there is no corresponding “hospital expire” flag in the CMS dataset, so the mortality prediction was only conducted on MIMIC III data.

1) Baseline Methods: First, we applied Gated Recurrent Units (GRU) [12] with the following embedding strategies to map visit embedding sequence $v_1, \ldots, v_M$ to a patient representation h:

- **CBOW+:** For the visit embedding, we simply mean the CBOW embeddings of the medical concepts within the visit.
- **Skip-gram (Sg+):** We performed the same process as CBOW+ but used Skip-gram vectors instead of CBOW vectors.
- **GloVe+:** [13] The same process as CBOW+, but using GloVe vectors instead of CBOW vectors.
- **med2vec:** We used Med2Vec [13] to learn visit embedding where the dimension is the same as other embedding strategy.
- **MCE+:** [14] The same process as CBOW+, but using MCE vectors instead of CBOW vectors.
- **TeSAN+:** The same process as CBOW+, but using the vectors of the proposed model.

We applied logistic regression to the patient representation h to obtain a value between 0 (Survivor) and 1 (Death). All models were trained end-to-end. We reported the Area under the Precision-Recall Curve (PR-AUC) and Area under the Receiver Operating Characteristic (ROC-AUC) in the experiment, as PR-AUC is considered a better measure for imbalanced data like ours [40], [41]. All models were trained with 50,00 steps; the batch size is 128 and the RNN cell type is GRU.

2) Results: Table VII shows the test loss, PR-AUC and ROC-AUC of all models on dataset MIMIC III. We find that TeSAN again consistently outperforms all baseline models. Achieving high specificity in mortality prediction is relatively easy as there are many more negative samples than positive ones. However, correctly identifying positive cases while ignoring negative ones requires a model differentiate between positive cases. This means attending to the details of patient records, such as the relationship between the diagnosis codes and temporal intervals. This is why TeSAN demonstrates a significant improvement in PR-AUC and ROC-AUC. Also, we note that GloVe shows very poor PR-AUC and ROC-AUC, and we observe that medical concepts with low frequencies are assigned near-zero vectors in this model, which might explain its poor performance.

3) Ablation Study: We performed a similar ablation study with unsupervised learning tasks to examine the contributions of the proposed model components to the prediction task.

TABLE VII
MORTALITY PREDICTION PERFORMANCE ON MIMIC III DATASET.

Model	test loss	test PR-AUC	test ROC-AUC
CBOW+	0.6765	0.5251	0.7784
Sg+	0.6764	0.5276	0.7785
GloVe+	0.6834	0.4172	0.6548
med2vec	0.6772	0.5217	0.7690
MCE+	0.6767	0.5204	0.7630
TeSAN+	**0.6736**	**0.5544**	**0.8064**

- **Normal_Sa+:** For the visit embedding, we simply mean the Normal_Sa embeddings of the medical concepts within the visit.
- **Interval+:** We perform the same process as Normal_Sa+, but use Interval vectors instead of Normal_Sa vectors.
- **Multi_Sa+:** We perform the same process as Normal_Sa+, but use Multi_Sa vectors instead of Normal_Sa vectors.
- **TeSAN:** We perform the same process as Normal_Sa+, but use our proposed model vectors instead of Normal_Sa vectors.

Table VIII shows the mortality prediction performance for the ablated models and our proposed model. As can be seen from the table, the proposed model achieves the best performance compared to the ablated models on medical concept embedding. We observe the same trend as the ablated performance of unsupervised tasks, in which Multi_Sa+ outperforms Normal_Sa+. TeSAN outperforms the best ablated model by 2.75% on PR-AUC and 2.54% on ROC-AUC over MIMIC III.

TABLE VIII
MORTALITY PREDICTION PERFORMANCE ON MIMIC III DATASET.

Ablation	test loss	test PR-AUC	test ROC-AUC
Normal_SA+	0.6762	0.5233	0.7759
Interval+	0.6764	0.5191	0.7778
Multi_SA+	0.6759	0.5269	0.7808
TeSAN+	**0.6736**	**0.5544**	**0.8064**

D. Visualization

We present the visualized sample medical concepts with 3-dimension T-SNE results from the learned 100-dimension embedding vectors using our proposed medical embedding model. We selected three out of 19 categories in the high level of ICD9 standard and used the high level nodes as the clustering labels.

Figure 6 shows the 2D T-SNE results of three categories of sample medical concepts trained on MIMIC III, in which the red dots represent "congenital anomalies", the green dots represent "certain conditions originating in the perinatal period", and the blue dots represent "symptoms, signs, and ill-defined conditions". We find that the majority of the red and blue dots can be grouped in dense areas, whereas some green dots mix with the blue dots but are mainly clustered in a sparse area.

Figure 7 shows the 2D T-SNE results three categories of sample medical concepts trained on CMS, in which the red dots represent "diseases of the blood and blood-forming organs", the green dots represent "diseases of the respiratory system", and the blue dots represent "diseases of the genitourinary system". We observe that although some green and red dots are intermingled, the red dots are grouped into a long dense arc area, and most green dots are grouped in another long dense area. The blue dots are clustered into a dense area close to red dots.

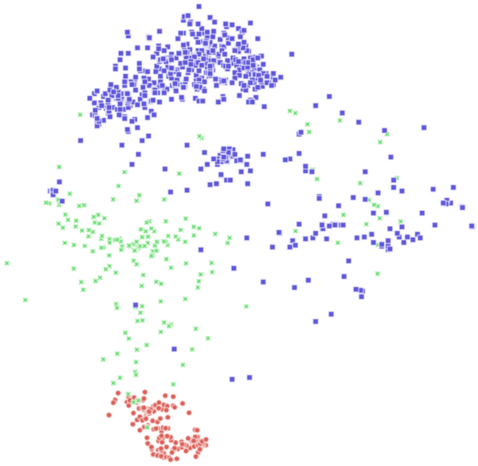

Fig. 6. Visualisation of 3 diagnosis categories in MIMIC III dataset.

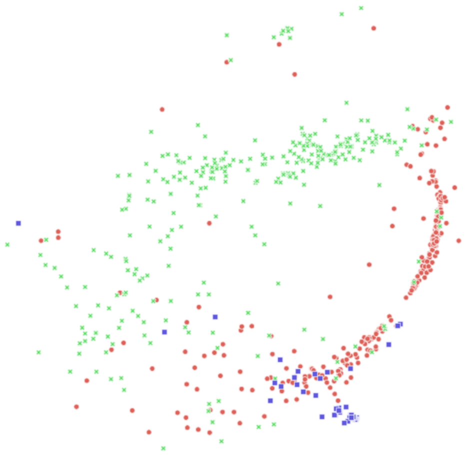

Fig. 7. Visualisation of 3 diagnosis categories in CMS dataset.

VI. CONCLUSION

This paper proposes a novel embedding model, the temporal self-attention network **TeSAN**. First, the model uses the self-attention mechanism to capture the contextual information and temporal relationships between medical concepts to compress the context into a vector representation by exploiting the temporal self-attention module (TeSA). Our model then applies the learning context vector to predict the target medical concept to learn representation for each medical concept. We conducted two types of tasks in unsupervised learning and prediction to evaluate the performance of our proposed model against baseline methods.We also executed ablation studies to examine the contributions of the proposed model components. The experimental study demonstrates that the proposed model outperforms the baseline methods over two public datasets in tasks of clustering, nearest neighbour search, and mortality prediction.

VII. Acknowledgements

This research was funded by the Australian Government through the Australian Research Council (ARC) under grants 1) LP160100630 partnership with Australia Government Department of Health and 2) LP150100671 partnership with Australia Research Alliance for Children and Youth (ARACY) and Global Business College Australia (GBCA). We also acknowledge the support of NVIDIA Corporation and MakeMagic Australia with the donation of GPUs.

References

[1] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE J Biomed Health Inform*, vol. 22, no. 5, pp. 1589–1604, 2018.

[2] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *JMLR*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[3] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *ICML*. ACM, 2008, pp. 160–167.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[5] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[6] J. A. Minarro-Giménez, O. Marin-Alonso, and M. Samwald, "Exploring the application of deep learning techniques on medical text corpora." *Stud Health Technol Inform*, vol. 205, pp. 584–588, 2014.

[7] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza, "Medical semantic similarity with a neural language model," in *CIKM*. ACM, 2014, pp. 1819–1822.

[8] T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh, "Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm)," *J Biomed Inform*, vol. 54, pp. 96–105, 2015.

[9] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 41, 2016.

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

[11] X. Zhang, J. Chou, and F. Wang, "Integrative analysis of patient health records and neuroimages via memory-based graph convolutional network," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 767–776.

[12] E. Choi, C. Xiao, W. Stewart, and J. Sun, "Mime: Multilevel medical embedding of electronic health records for predictive healthcare," in *NeurIPS*, 2018, pp. 4552–4562.

[13] 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 *SIGKDD*. ACM, 2016, pp. 1495–1504.

[14] X. Cai, J. Gao, K. Y. Ngiam, B. C. Ooi, Y. Zhang, and X. Yuan, "Medical concept embedding with time-aware attention," in *IJCAI*, 2018, pp. 3984–3990.

[15] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, p. 533, 1986.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv:1301.3781*, 2013.

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[19] O. Melamud, D. McClosky, S. Patwardhan, and M. Bansal, "The role of context types and dimensionality in learning word embeddings," *arXiv:1601.00893*, 2016.

[20] L. Liu, F. Ruiz, S. Athey, and D. Blei, "Context selection for embedding models," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4816–4825. [Online]. Available: http://papers.nips.cc/paper/7067-context-selection-for-embedding-models.pdf

[21] W. Ling, Y. Tsvetkov, S. Amir, R. Fermandez, C. Dyer, A. W. Black, I. Trancoso, and C.-C. Lin, "Not all contexts are created equal: Better word representations with variable attention," in *EMNLP*, 2015, pp. 1367–1372.

[22] K. Jha, Y. Wang, G. Xun, and A. Zhang, "Interpretable word embeddings for medical domain," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1061–1066.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[24] W. Lee, S. Park, W. Joo, and I.-C. Moon, "Diagnosis prediction via medical context attention networks using deep generative modeling," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1104–1109.

[25] L. Liu, T. Zhou, G. Long, J. Jiang, L. Yao, and C. Zhang, "Prototype propagation networks (ppn) for weakly-supervised few-shot learning on category graph," *IJCAI*, 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[27] Y. Liu, C. Sun, L. Lin, and X. Wang, "Learning natural language inference using bidirectional lstm model and inner-attention," *arXiv:1605.09090*, 2016.

[28] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *arXiv:1503.02364*, 2015.

[29] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," in *NIPS*, 2015, pp. 2440–2448.

[30] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv:1509.00685*, 2015.

[31] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," in *AAAI*, 2018.

[32] M. Hu, Y. Peng, Z. Huang, X. Qiu, F. Wei, and M. Zhou, "Reinforced mnemonic reader for machine reading comprehension," *arXiv:1705.02798*, 2017.

[33] Z. Li, Y. Wei, Y. Zhang, and Q. Yang, "Hierarchical attention transfer network for cross-domain sentiment classification," in *AAAI*, 2018.

[34] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv:1703.03130*, 2017.

[35] T. Shen, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Bi-directional block self-attention for fast and memory-efficient sequence modeling," *arXiv:1804.00857*, 2018.

[36] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[37] S. Wang, X. Li, L. Yao, Q. Z. Sheng, G. Long *et al.*, "Learning multiple diagnosis codes for icu patients with local disease correlation mining," *TKDD*, vol. 11, no. 3, p. 31, 2017.

[38] J. Kellett and A. Kim, "Validation of an abbreviated vitalpac early warning score (views) in 75,419 consecutive admissions to a canadian regional hospital," *Resuscitation*, vol. 83, no. 3, pp. 297–302, 2012.

[39] Y. P. Tabak, X. Sun, C. M. Nunez, and R. S. Johannes, "Using electronic health record data to develop inpatient mortality predictive model: Acute laboratory risk of mortality score (alarms)," *Journal of the American Medical Informatics Association*, vol. 21, no. 3, pp. 455–463, 2013.

[40] H. Yamana, H. Matsui, K. Fushimi, and H. Yasunaga, "Procedure-based severity index for inpatients: development and validation using administrative database," *BMC health services research*, vol. 15, no. 1, p. 261, 2015.

[41] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *ICML*. ACM, 2006, pp. 233–240.