ConCare: Personalized Clinical Feature Embedding via Capturing the Healthcare Context

Liantao Ma[1,3], Chaohe Zhang[1,3], Yasha Wang [1,2*] , Wenjie Ruan[4], Jiangtao Wang[4], Wen Tang[5], Xinyu Ma[1, 3], Xin Gao[1,3], Junyi Gao[1, 2]

[1]Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China
[2]National Engineering Research Center of Software Engineering, Peking University, Beijing, China
[3]School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[4]School of Computing and Communications, Lancaster University, UK
[5]Division of Nephrology, Peking University Third Hospital, Beijing, China
{malt, wangyasha}@pku.edu.cn, {wenjie.ruan, jiangtao.wang}@lancaster.ac.uk, tanggwen@126.com

Abstract

Predicting the patient's clinical outcome from the historical electronic medical records (EMR) is a fundamental research problem in medical informatics. Most deep learning-based solutions for EMR analysis concentrate on learning the clinical visit embedding and exploring the relations between visits. Although those works have shown superior performances in healthcare prediction, they fail to explore the personal characteristics during the clinical visits thoroughly. Moreover, existing works usually assume that the more recent record weights more in the prediction, but this assumption is not suitable for all conditions. In this paper, we propose ConCare to handle the irregular EMR data and extract feature interrelationship to perform individualized healthcare prediction. Our solution can embed the feature sequences separately by modeling the time-aware distribution. ConCare further improves the multi-head self-attention via the cross-head decorrelation, so that the inter-dependencies among dynamic features and static baseline information can be effectively captured to form the personal health context. Experimental results on two real-world EMR datasets demonstrate the effectiveness of ConCare. The medical findings extracted by ConCare are also empirically confirmed by human experts and medical literature.

Introduction

Performing personal health evaluation for each individual patient is always the goal that physicians pursue. Electronic Medical Records (EMR) now provide the possibility to realize these goals. EMR is a type of multivariate time series data that records patients' visits in hospitals (e.g., diagnoses, lab tests. As shown in Figure 1) and static baseline information (e.g., gender, primary disease. As shown in Figure 2). Recently deep learning-based models have demonstrated state-of-the-art performance in mining the massive EMR data (Ma et al. 2020; Lee et al. 2017; Gao et al. 2019; Liu et al. 2018; 2019). Usually, existing works incorporate multiple dynamic features (e.g., lab test values) to learn the visit embedding and the health status through the entire clinical visits by sequential models (Ma et al. 2017).

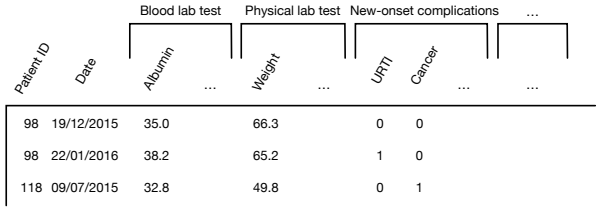

Figure 1: Dynamic medical features. The physician conducts the necessary lab tests for the patient at each visit.

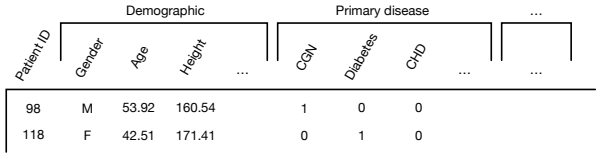

Figure 2: Static baseline information. Such characteristics are usually used to evaluate the basic condition of the patient and the prognosis in the clinical practice.

Although the state-of-the-art performance has been demonstrated in these works, the personal characteristics through clinical visits have not yet been fully taken into consideration on the healthcare prediction. Specifically, there are two research challenges, i.e., how to extract the different meanings of the particular clinical features for patients in diverse conditions, and how to evaluate the impact of irregular visit time intervals in the healthcare prediction.

- I_1**: Extracting Personal Health Context**: A certain value of a clinical feature (e.g., blood glucose) may imply different meanings to patients with diverse static baselines (e.g., diagnosis of diabetes as a primary disease, as shown in Figure 2). In order to evaluate the health status of the patient comprehensively, physicians need to take a look at the static clinical baseline information. Besides, not only the static baseline information, but also the dynamic clinical feature sequence (as shown in Figure 1) can be treated as the health context of the patient. For example, when plasma concentrations of creatinine

*Corresponding Author

and urea begin a hyperbolic rise, both of them and the GFR (i.e., Glomerular Filtration Rate) value are usually associated with systemic manifestations (uremia) for patients with chronic kidney disease (Anna Malkina 2018). Thus, considering the particular condition of the patient, the way of attending the medical features in the whole prediction process should be individualized. Some existing works try to model the relationship between clinical visits (Ma et al. 2017), dynamic features (Bai et al. 2018; Choi et al. 2016) or incorporate the static information (Lee et al. 2018). However, none of the existing models explored the interdependencies among dynamic records as well as static baseline information via a global view. In practice, it is critical to explore the inherent relationship between clinical features to build the personal healthcare context and perform the prediction individually.

- I_2**: Capturing the Impact of Time Interval**: The patient goes to the hospital only when he feels sick, and the physician prescribes lab examinations when it is necessary. Therefore, medical records are produced irregularly in clinical practice. It is assumed by many existing works (Pham et al. 2016; Baytas et al. 2017; Ma, Xiao, and Wang 2018; Bai et al. 2018) that the more recent clinical records weight more than previous records in general on the healthcare prediction. However, under certain circumstance, historical records also contain valuable clinical information, which may not be revealed in the latest record (e.g., the blood glucose level was extremely abnormal). For instance, the upper respiratory tract infection (URTI) record a few years ago has almost no influence on the current healthcare prediction. However, the historical diagnosis of cerebrovascular disease indicates that the patient has been suffering from chronic cerebrovascular damage, so it is continuously the risk factor during the rest of the life (Somers et al. 2008). Thus building a more adaptive time-aware mechanism to flexibly learn the impact of the time interval for each clinical feature is urgently needed.

To jointly tackle the above issues, in this paper, we propose a multi-channel healthcare predictive model, which can learn the representation of health status and perform the health prediction by more deeply considering the personal health context. `ConCare` evaluates the health status of patients mainly from the perspective of clinical features, rather than visits. It embeds the time series of each feature separately. The time decay effects of different features can be extracted separately and flexibly via corresponding learnable time-aware parameters. The model explicitly extracts the interdependencies among time series of dynamic features as well as static baseline information, to learn the personal health context of patients in a global view. `ConCare` re-encodes each feature by looking at other features for clues that can help lead to a better understanding for this feature, so as to depict the health status more individually. Our main contributions are summarized as follows:

- We propose a novel health status representation framework called `ConCare` by fully considering the personal patient's health context. The health context is formed by capturing the interdependencies between clinical features which are extracted separately. To the best of our knowledge, we are the first research to jointly consider static baseline information, sequential dynamic features and the impact of the time interval as personal health context in the clinical representation learning.
- Specifically, 1) We explicitly extract interdependencies between clinical features to learn the personal health context and regenerate the feature embedding under the context, by a multi-head self-attention mechanism (addressing I_1). The cross-head decorrelation is utilized to encourage the diversity among heads. 2) We propose a multi-channel medical feature embedding architecture, which learns the representation of different feature sequences via separate GRUs, and adaptively captures the effect of time intervals between records of each feature by time-aware attention (addressing I_2).
- We conduct the mortality prediction task on two real-world datasets (i.e., MIMIC-III dataset and end-stage renal disease dataset) respectively to verify the performance. The results[1] show that `ConCare` significantly and consistently outperforms the baseline approaches in both tasks. We also reveal several interesting medical implications. 1) We provide the overall time-decay ratios for diverse biomarkers by the learnable parameter in time-aware attention. 2) We provide the adaptive cross-feature interdependencies, which further suggests possible medical research between specific features. The obtained medical knowledge has been positively confirmed by human experts and clinical literature.

Related Work

Exploring Relationship Among Clinical Records

Most existing works only focus on exploring the relationship between clinical visits, in similar ways as general time series analysis and natural language processing tasks. For example, Dipole (Ma et al. 2017) uses bidirectional RNN architecture and the attention mechanism to capture the relationships of different visits for the prediction. SAnD (Song et al. 2018) employs self-attention mechanism, positional encoding, and dense interpolation strategies to incorporate temporal order on clinical prediction tasks.

There are also a few novel research works that try to model the relationship between dynamic features rather than just the visits. For example, (Bai et al. 2018) uses the self-attention mechanism to combine all diagnosis records produced in the visit to form the visit embedding, but it fails to extract the relationship in a global sequential view. (Gupta et al. 2018) embeds the feature sequences by a pre-trained TimeNet, which cannot capture the unique characteristics for different features, respectively. RETAIN (Choi et al. 2016) employs two RNNs to learn time attention as well as feature attention, and then sums up the weighted visit embedding to perform the prediction, but it lacks advanced fea-

[1]We release our code and case studies at GitHub https://github.com/Accountable-Machine-Intelligence/ConCare

ture extraction, and its prediction accuracy is limited (Ma, Xiao, and Wang 2018; Ma et al. 2018a).

Besides the utilization of dynamic sequential data, several novel solutions also try to incorporate the static baseline information. For example, (Lee et al. 2018) proposes a medical context attention-based RNN that utilizes the derived individual information from conditional variational autoencoders. However, none of them explore the interdependencies between static baseline information and dynamic records from a global view. The proposed model in this paper, `ConCare`, can adaptively capture the relationship between clinical features and perform individualized prediction for patients in diverse health contexts.

Handling the Time Interval between Visits

Although most of the existing works (Ma et al. 2018b; Choi et al. 2017) simply treat the clinical visits in equal intervals, several novel works (Pham et al. 2016; Ma, Xiao, and Wang 2018) try to model the importance of clinical visits with time intervals, by attaching a fixed time-decay ratio to decay the hidden memory of the previous visit. Those works omit the different characteristics between features. For example, T-LSTM (Baytas et al. 2017) handles irregular time intervals of visits in longitudinal patient records by enabling time decay to discount the cell memory content in LSTM. In order to capture the characteristics of different disease codes, Timeline (Bai et al. 2018) develops time decay factors for diseases to form visit representation and feeds it into an RNN for prediction. But its time-aware effect is still disrupted during the historical visit embedding process due to the rapid memory forgetting of RNN. And Timeline can only handle the disease codes as features rather than biomarkers.

Therefore, existing works simply assume that recent records play more important roles than previous records. However, according to the clinical practice, some historical clinical events also strongly indicate the health status under certain circumstances while it may not be revealed in the latest record. The time-aware mechanism should take the characteristics of features into consideration and meanwhile, flexibly retain the vital historical information. `ConCare` can capture the impact of time interval in diverse feature sequences by a learnable time-aware attention.

Problem Formulation

We assume that the patient's dynamic clinical records (as shown in Figure 1) consist of T visits to the hospital. The number of features in each visit record is N. As a result, such a clinical sequence can be formulated as a "longitudinal patient matrix" $record$ where one dimension represents medical features and the other one denotes visit timestamps (Lee et al. 2017):

$$record = \begin{pmatrix} r_{11} & \cdots & r_{1T} \\ \vdots & \ddots & \vdots \\ r_{N1} & \cdots & r_{NT} \end{pmatrix}. \quad (1)$$

The static baseline data (as shown in Figure 2), including demographic attributes and historical primary diseases, is denoted as $base$. The objective of healthcare prediction is using EMR data (i.e., $record$ and $base$) to predict whether a patient suffers from the target health risk during the period of the treatment procedure, denoted as $y \in \{0, 1\}$. This problem is posed as a binary classification under a certain time window (e.g., 24 hours), namely, $\widehat{y} = \texttt{ConCare}(record, base)$.

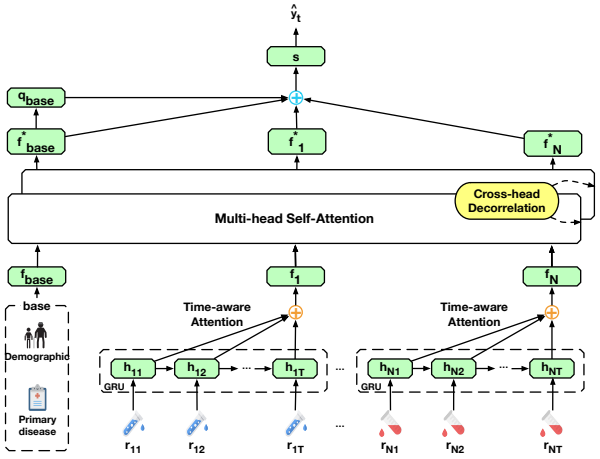

Figure 3: The Framework of `ConCare`.

Solution

Figure 3 shows the framework of the proposed `ConCare`. The model treats the clinical information of the patient from the perspective of features rather than visits. We extract the context vector of each dynamic feature and static baseline information separately. Such feature embedding vector are then re-encoded by taking the information of all features as healthcare context. The framework comprises of the following sub-modules:

- The multi-channel time series embedding module with time-aware attention is developed to separately learn the representation of each dynamic feature.
- The feature encoder is adopted to combine all the static information and dynamic records based on self-attention.

The individualized prediction finally is obtained from all regenerated feature embeddings with an attention queried by static baseline information. We will present the details in the following subsections.

Multi-Channel Clinical Sequence Embedding

In `ConCare`, we aim to capture the interdependencies between features based on self-attention mechanism (Vaswani et al. 2017). Since the self-attention architecture contains no recurrence, in order to incorporate information about the order of the sequence, researchers simply utilize the fixed positional encoding to provide relative position information for timestamps (Song et al. 2018). However, such positional embedding capability is limited, especially for the absolute position understanding, but the logical order of the clinical sequence actually matters in the medical domain. `ConCare`

thus embeds the time series of each feature separately by multi-channel GRU:

$$h_{n,1}, ..., h_{n,T} = GRU_n(r_{n,1}, ..., r_{n,T}), \quad (2)$$

where, the time series of feature n is denoted as $r_{n,:} = (r_{n,1}, ...,r_{n,T}) \in R^T$. Then, the hidden representations is summarized across the whole time span. To capture the impact of time intervals in each sequence, we propose a time-aware attention mechanism here. Generally, an attention function can be described as mapping a query and a set of key-value pairs to an output (Vaswani et al. 2017). The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. First, the $Query$ vector is generated by the hidden representation at the last time step T, and the Key vectors are generated by each hidden representation:

$$q^{emb}_{n,T} = W^q_n \cdot h_{n,T}, \quad (3)$$

$$k^{emb}_{n,t} = W^k_n \cdot h_{n,t}, \quad (4)$$

where $q^{emb}_{n,T}$ and $k^{emb}_{n,t}$ are the $Query$ vector and the Key vector respectively. W^q_n and W^k_n are the corresponding projection matrices to obtain the query and key vectors. Then we design the time-aware attention weights as follow:

$$\alpha_{n,1}, \alpha_{n,2}, ..., \alpha_{n,T} = Softmax(\zeta_{n,1}, \zeta_{n,2}, ..., \zeta_{n,T}), \quad (5)$$

where

$$\zeta_{n,t} = tanh(\frac{q^{emb}_{n,T} \cdot k^{emb}_{n,t}}{\beta_n \cdot log(e + (1 - \sigma(q^{emb}_{n,T} \cdot k^{emb}_{n,t})) \cdot \Delta t)}). \quad (6)$$

This is an alignment model that can quantify how much each hidden representation contributes to the densely summarized representation for each feature. Δt is the time interval to the latest record. σ is the $sigmoid$ function. β_n is a feature-specific learnable parameter trained to control the influence of the time interval on the corresponding feature. The attention weight $\alpha_{n,t}$ will be significantly decayed, if:

- the time interval Δt is long, which means that such value is recorded a long time ago. It is obviously that, the most recent (i.e., $\Delta t = 0$) value of any feature will only be decayed slightly (i.e., $log(e) = 1$).
- the time-decay ratio β_n is high which means that for particular clinical feature only recent recorded value matters. The clinical feature whose influence persists (i.e., β_n is low) will be decayed just slightly.
- the historical record does not actively respond to the current health condition (i.e., $q^{emb}_{n,T} \cdot k^{emb}_{n,t}$ is small).

Finally, based on the learned weights, we can derive time-aware contextual feature representation as $f_n = \sum_{i=1}^{T} \alpha_{n,t} \cdot h_{n,t}$. Furthermore, the demographic baseline data is embedded into the same hidden space of f_n

$$f_{base} = W^{emb}_{base} \cdot base, \quad (7)$$

where W^{emb}_{base} is an embedding matrix. Thus, all the data of the patient can be represented by a matrix F (i.e., a sequence of vectors, where each vector represents one feature of the patient over time): $F = (f_1, \cdots, f_N, f_{base})^\top$.

Learning the Context and Re-encoding the Feature

We capture the interdependencies among dynamic features through visits as well as static baseline information, and further re-encode the feature embedding under the personal context based on self-attention. As the `ConCare` processes each feature, self-attention allows it to look at other features for clues that can help lead to a better encoding for this feature. For example, when the model is processing the feature "blood glucose", self-attention may allow it to associate it with "diagnosis of diabetes" in the static baseline information. Besides, the multi-head mechanism enhances the attention layer with multiple representation subspaces. Mathematically, given current feature representations F, the refined new representations are calculated as:

$$\begin{aligned} u_n &= MultiHeadAttention(F) \\ &= [head_1(f_n) \oplus head_2(f_n) \oplus ... \oplus head_m(f_n)]W^O, \end{aligned} \quad (8)$$

where $head_m$ is m-th attention head, $\oplus$ is the concatenation operation and W^O is a linear projection matrix. Considering both efficiency and effectiveness, the scaled dot product is used as the attention function (Vaswani et al. 2017). This following $softmax$ score determines how much each feature will be expressed at this certain feature. Specifically, $head_m$ is the weighted sum of all value vectors and the weights are calculated by applying attention function to all the query, key pairs:

$$\alpha_1, \alpha_2, ..., \alpha_{N+1} = \text{Softmax}(\frac{q \cdot k_1}{\sqrt{d_k}}, \frac{q \cdot k_2}{\sqrt{d_k}}, ..., \frac{q \cdot k_{N+1}}{\sqrt{d_k}}), \quad (9)$$

$$head_m(g_n) = \sum_{i=1}^{N+1} \alpha_i \cdot v_i, \quad (10)$$

where q, k_i and v_i are the query, key, and value vectors and d_k is the dimension of k_i. Moreover, q, k_i and v_i are obtained by projecting the input vectors into query, key and value spaces, respectively (Wang et al. 2019). They are formally defined as:

$$q, k_i, v_i = W^q \cdot f, W^k \cdot f_i, W^v \cdot f_i, \quad (11)$$

where W^q, W^k and W^v are the projection matrices and each $head_m$ has its own projection matrices. As shown in Eq.9 and Eq.10 each $head_m$ is obtained by letting f attending to all the $fearure$ positions, thus any feature interdependencies between f and f_i can be captured.

Cross-Head Decorrelation

Heads for self-attention are expected to capture dependencies from different aspects. However, in practice, heads may tend to learn similar dependencies according to (Vaswani et al. 2017). To overcome this challenge, we encourage diverse or non-redundant representations (Cogswell et al. 2015; Chu et al. 2019) by minimizing the cross-covariance of hidden activations across different heads. We utilize the cross-head decorrelation module to expand the model's ability to

focus on different features, based on (Cogswell et al. 2015) which reduces the redundancy of the normal neural network layer. According to Eq.8, we get u_t as the multi-head attention for f_t, which is the concatenation of the heads. For simplicity, here we use u to denote u_t as a general case. The covariances between all pairs of activations i and j of u form a matrix C:

$$C_{i,j} = \frac{1}{B}\sum_{b=1}^{B}(u_i^b - \mu_i)(u_j^b - \mu_j), \tag{12}$$

where B is the batch size and u_i^b is the $i-th$ activation of u at $b-th$ case in the batch. $\mu_i = \frac{1}{B}\sum_{b=1}^{B} u_i^b$ is the sample mean of activation i over the batch. Covariance between diverse heads is expected to be minimized. The diagonal of C is then subtract from the matrix norm to build the cross-head decorrelation loss term:

$$\mathcal{L}_{decorrelation} = \frac{1}{2}(\|C\|_F^2 - \|diag(C)\|_2^2), \tag{13}$$

where $\|\cdot\|_F$ is the frobenius norm,and the $diag()$ operator extracts the main diagonal of a matrix into a vector. After obtaining the refined representation of each position by the multi-head attention mechanism, we add a position-wise fully connected feed-forward network sub-layer. This feed-forward network transforms the features non-linearly and is defined as $FeedForward(r_n) = max(0, u_n \cdot W_1 + b_1) \cdot W_2 + b_2$. We also employ a residual connection (He et al. 2016) around each of the two sub-layers, followed by layer normalization (Ba, Kiros, and Hinton 2016). As shown in Fig.3, the outputs of this subsection from F are denoted as $F^* = (f_1^*, f_2^*, ..., f_N^*, f_{base}^*)^\top$.

Healthcare Prediction

A dense health status representation is expected to perform the final prediction. Here, we introduce an individualized characterization attention summarization. The $Query$ is obtained by f_{base}^* and Keys are formed by F^* as:

$$q_{base}^{fin} = W_{base}^{fin} \cdot f_{base}^*, \tag{14}$$

$$k_n^{fin} = W_n^{fin} \cdot f_n^*, \tag{15}$$

where W_{base}^{fin} and W_n^{fin} are the projection matrix respectively. Similar to the first subsection, the attention weights are calculated as:

$$\alpha_{base}^{fin}, \alpha_1^{fin}, ..., \alpha_N^{fin} = \text{Softmax}(\zeta_{base}^{fin}, \zeta_1^{fin}, ..., \zeta_N^{fin}), \tag{16}$$

$$\zeta_n^{fin} = tanh(q_{base}^{fin} \cdot k_n^{fin}). \tag{17}$$

The health status representation s and the prediction result $\widehat{y}$ can be obtained by:

$$s = \sum_{i=1}^{N} \alpha_i^{fin} \cdot f_i^* + \alpha_{base}^{fin} \cdot f_{base}^*, \tag{18}$$

$$\widehat{y} = \sigma(W^{fin} \cdot s + b^{fin}), \tag{19}$$

where W^{fin} and b^{fin} are the weight matrix and bias term, respectively. And the final loss can be denoted as the combination of cross-entropy loss and decorrelation loss

$$\mathcal{L} = \mathcal{L}_{cross-entropy} + \mathcal{L}_{decorrelation}. \tag{20}$$

Experiment

We conduct the mortality prediction experiments on MIMIC-III dataset [2] and end-stage renal disease (ESRD) dataset. The source code of `ConCare`, statistics of datasets and case studies are available at the GitHub repository[3].

Datasets and Prediction Tasks

- **MIMIC-III Dataset.** We use ICU data from the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database (Johnson et al. 2016). We perform the in-hospital mortality prediction for patients based on patients' demographic data and events produced during ICU stays (Harutyunyan et al. 2017). We fix a test set of 15% of patients and divide the rest of the dataset into the training set and validation set with a proportion of $0.85 : 0.15$.
- **Real-World ESRD Dataset.** We perform the mortality prediction on an end-stage renal disease dataset. The cleaned dataset consists of 656 patients with static baseline information and 13,091 dynamic records. There are 1196 records with positive labels (i.e., died within 12 months) and 10,804 records with negative labels. The training set is further split into 10 folds to perform the 10-fold cross-validation.

We assess performance using the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and the minimum of precision and sensitivity Min(Se,P+). AUPRC is the most informative and the primary evaluation metric when dealing with a highly imbalanced and skewed dataset (Davis and Goadrich 2006; Choi et al. 2018) like the real-world EMR data.

Implementation Details and Baseline Approaches

The training was done in a machine equipped with CPU: Intel Xeon E5-2630, 256GB RAM, and GPU: Nvidia Titan V by using Pytorch 1.1.0. For training the model, we used Adam (Kingma and Ba 2014) with the mini-batch of 256 patients and the learning rate is set to $1e-3$. To fairly compare different approaches, the hyper-parameters of the baseline models are fine-tuned by grid-searching strategy. We include several state-of-the-art models as our baseline approaches.

- GRU_α is the basic GRU with an addition-based attention mechanism.
- RETAIN (NeurIPS 2016) (Choi et al. 2016) utilizes a two-level neural attention mechanism to detect influential visits and significant variables.
- T-LSTM (SIGKDD 2017) (Baytas et al. 2017) handles irregular time intervals by enabling time decay. We modify it into a supervised learning model.
- MCA-RNN (ICDM 2018) (Lee et al. 2018) utilizes the derived individual patient information from conditional variational auto-encoders to construct a medical context attention-based RNN.

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

[3] https://github.com/Accountable-Machine-Intelligence/ConCare

- Transformer$_e$ (NeurIPS 2017) (Vaswani et al. 2017) is the encoder of the Transformer, in the final step, we use to flatten and FFNs to make the prediction.
- SAnD$_*$ (AAAI 2018) (Song et al. 2018) models clinical time-series data solely based on masked self-attention. When performing prediction at every time step, we use causal padding (Van Den Oord et al. 2016) for the convolutional layer to prevent using future information. We re-implement SAnD by using $r_{t-k+1:t}$ to build input embedding at the measurement position t, instead of the one proposed in the original paper $r_{t:t+k-1}$, to avoid the violation of causality.

For a fair comparison, although most of the comparative approaches did not take the static baseline information into consideration which is greatly beneficial for improving the performance of healthcare prediction, we feed such characteristics as additional input (i.e., concatenate with the raw input) for them at each visit.

Results of Risk Prediction

Table 1 shows the performance of all approaches on two datasets. The number in () denotes the standard deviation of bootstrapping for 100 times on the MIMIC-III dataset and the standard deviation of 10-fold cross-validation on the ESRD dataset. The results indicate that ConCare significantly and consistently outperform other baseline methods.

We find that ConCare outperforms the approaches that only utilize the embedding of the health status in visits (i.e., ConCare$_{MC-}$ and all comparative approaches). ConCare also outperforms the approaches which incorporate the static information. It indicates that capturing the interdependencies among clinical features (including static baseline information and dynamic features) and regenerating the feature embedding under the personal health context is critical for evaluating the health status.

Moreover, ConCare outperforms the positional encoding-based approaches (i.e., ConCare$_{PE}$, Transformer-Encoder, SAnD). It demonstrates the superior of multi-channel GRU encoder than the conventional positional encoding which is difficult to precisely embed the positional information. ConCare also outperforms the time-aware approaches (i.e., T-LSTM), which demonstrates that capturing the time-decay impact of each feature separately in a global view is superior to directly decaying the hidden memory of entire visits. The superior performance of ConCare than the ConCare$_{DE-}$ (i.e., without the decorrelation loss) verifies the efficacy of the decorrelation loss which can encourage the diversity among heads and improve the performance.

Findings and Implications

This section will discuss the findings and implications of ConCare in the experiments.

Decay Rates For Different Features Figure 4 shows the decay rates (i.e. the β in Eqn. 6) learned adaptively for different features, which depicts how the importance of previous values of features fades through time. The darker boxes

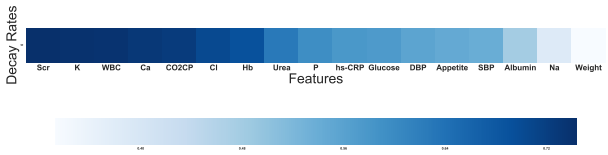

Figure 4: Decay Rates For Different Features

mean the importance of previous values of features fades quickly (i.e., the short-term patterns of the features matter), and vice versa. The figure indicates that ConCare attends more on the short-term of serum creatinine (Scr), K, White Blood Cell Count (WBC), Ca, Carbon-dioxide Combining Power (CO2CP), Cl, hemoglobin (Hb). According to the medical commonsense, the above features are relatively fast-changing indicators, reflecting the patient's infection status or dialysis adequacy, etc. Conversely, the weight, albumin, Na and systemic blood pressure (SBP) need to be attended in the long-term aspect. According to medical research (Meijers et al. 2008), these features are usually related to nutrition intake and reflect the patient's condition over a period of time.

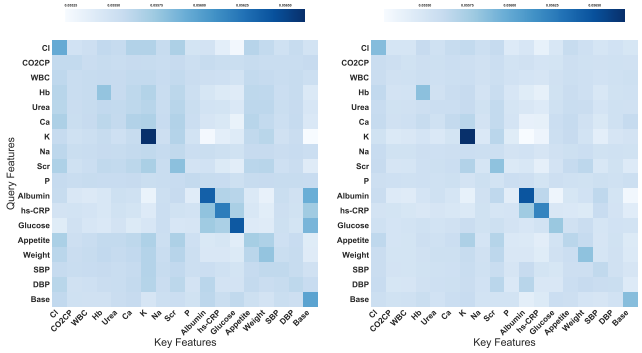

Figure 5: Cross-Feature Interdependencies: Patients Died with (Left) / without (Right) Diabetes

Cross-Feature Interdependencies Figure 5 shows cross-feature interdependencies of all patients who died with/without diabetes respectively. The average attention weights of one head calculated by the self-attention module are shown. The ordinates of the two figures are the $Query$ features and the abscissas are the Key features. The boxes in the figures show when a $Query$ feature makes a query, how much each Key feature respond to the $Query$. Most of the clinical features are more likely to respond to themselves, which denoted by the diagonal of two matrices. It is common medical knowledge that the glucose of a patient is strongly related to diabetes. By comparing the two figures, in the box of Glucose-Glucose position, the model pays much more attention to the glucose in patients who died with diabetes. Besides, ConCare figures out that there are relatively high interdependencies between albumin, hyper-sensitive C-reactive protein (hs-CRP), glucose and the static information (including age, diagnosis of diabetes) for patients suffering from diabetes. This is

Table 1: Results of the Healthcare Prediction Tasks

	MIMIC-III Dataset (Bootstrapping)			ESRD Dataset (10-Fold Cross Validation)		
Methods	AUROC	AUPRC	min(Se, P+)	AUROC	AUPRC	min(Se, P+)
GRU_{α}	.8628 (.011)	.4989 (.022)	.5026 (.028)	.8066 (.004)	.3502 (.009)	.3770 (.006)
RETAIN	.8313 (.014)	.4790 (.020)	.4721 (.022)	.7986 (.005)	.3386 (.009)	.3699 (.011)
MCA-RNN	.8587 (.013)	.5003 (.028)	.4932 (.024)	.8021 (.015)	.3451 (.041)	.3731 (.025)
T-LSTM	.8617 (.014)	.4964 (.022)	.4977 (.029)	.8101 (.015)	.3508 (.052)	.3721 (.045)
Transformer_{e}	.8535 (.014)	.4917 (.022)	.5000 (.019)	.8082 (.027)	.3502 (.062)	.3719 (.037)
SAnD_{*}	.8382 (.007)	.4545 (.018)	.4885 (.017)	.8002 (.026)	.3371 (.036)	.3591 (.053)
$\texttt{ConCare}_{PE}$	.8566 (.008)	.4811 (.024)	.5012 (.020)	.8124 (.025)	.3561 (.047)	.3761 (.037)
$\texttt{ConCare}_{MC-}$	.8594 (.008)	.4902 (.024)	.4947 (.025)	.8101 (.023)	.3498 (.066)	.3766 (.064)
$\texttt{ConCare}_{DE-}$	.8671 (.009)	.5231 (.028)	.5080 (.023)	.8162 (.033)	.3525 (.063)	**.3864** (.034)
ConCare	**.8702** (.008)	**.5317** (.027)	**.5082** (.021)	**.8209** (.036)	**.3606** (.084)	.3853 (.071)

highly consistent with the medical research (Milan Manani et al. 2015) and medical experience.

Conclusion

In this work, we proposed a novel medical representation learning framework, ConCare, which can explicitly extract the personal healthcare context and perform health prediction individually. Specifically, it extracts the clinical features by multi-channel GRU with a time-aware attention mechanism. The interdependencies among static baseline information and dynamic features are captured to build the health context and re-encode the clinical information. We conducted experiments on two real-world datasets. ConCare demonstrated significant prediction performance improvement across both tasks. It provides the time-decay ratios for different features respectively and indicates the interdependencies between features as interpretability. All extracted medical findings have been positively confirmed by experts and medical literature. The results also remind some possible medical research opportunities for deeply analyzing the relationship between some clinical features.

Acknowledgments

This work is supported by the National Science and Technology Major Project (No. 2018ZX10201002), and the fund of the Peking University Health Science Center (BMU20160584). WR is supported by ORCA PRF Project (EP/R026173/1).

References

Anna Malkina. 2018. Chronic kidney disease. [Online; October 2018].

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bai, T.; Zhang, S.; Egleston, B. L.; and Vucetic, S. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 43–51. ACM.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 65–74. ACM.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, 3504–3512.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 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*, 787–795. ACM.

Choi, E.; Xiao, C.; Stewart, W.; and Sun, J. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems*, 4547–4557.

Chu, X.; Lin, Y.; Wang, Y.; Wang, L.; Wang, J.; and Gao, J. 2019. Mlrda: A multi-task semi-supervised learning framework for drug-drug interaction prediction. In *The 28th International Joint Conference on Artificial Intelligence*, 4518–4524. Morgan Kaufmann.

Cogswell, M.; Ahmed, F.; Girshick, R.; Zitnick, L.; and Batra, D. 2015. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*.

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240. ACM.

Gao, J.; Wang, X.; Wang, Y.; Yang, Z.; Gao, J.; Wang, J.; Tang, W.; and Xie, X. 2019. Camp: Co-attention memory networks for diagnosis prediction in healthcare. In *ICDM*, 1036–1041. IEEE.

Gupta, P.; Malhotra, P.; Vig, L.; and Shroff, G. 2018. Using features from pre-trained timenet for clinical predictions. In *The 3rd International Workshop on Knowledge Discovery in Healthcare Data at IJCAI*.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; and Galstyan, A. 2017. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

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

Lee, C.; Luo, Z.; Ngiam, K. Y.; Zhang, M.; Zheng, K.; Chen, G.; Ooi, B. C.; and Yip, W. L. J. 2017. Big healthcare data analytics: Challenges and applications. In *Handbook of Large-Scale Distributed Computing in Smart Healthcare*. Springer. 11–41.

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

Liu, L.; Shen, J.; Zhang, M.; Wang, Z.; and Tang, J. 2018. Learning the joint representation of heterogeneous temporal events for clinical endpoint prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liu, L.; Li, H.; Hu, Z.; Shi, H.; Wang, Z.; Tang, J.; and Zhang, M. 2019. Learning hierarchical representations of electronic health records for clinical outcome prediction. In *AMIA Annual Symposium*.

Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; and Gao, J. 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*, 1903–1911. ACM.

Ma, F.; Gao, J.; Suo, Q.; You, Q.; Zhou, J.; and Zhang, A. 2018a. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1910–1919. ACM.

Ma, F.; You, Q.; Xiao, H.; Chitta, R.; Zhou, J.; and Gao, J. 2018b. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 743–752. ACM.

Ma, L.; Gao, J.; Wang, Y.; Zhang, C.; Wang, J.; Ruan, W.; Tang, W.; Gao, X.; and Ma, X. 2020. Adacare: Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Ma, T.; Xiao, C.; and Wang, F. 2018. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, 261–269. SIAM.

Meijers, B. K.; Bammens, B.; Verbeke, K.; and Evenepoel, P. 2008. A review of albumin binding in ckd. *American journal of kidney diseases* 51(5):839–850.

Milan Manani, S.; Virzì, G. M.; Clementi, A.; Brocca, A.; de Cal, M.; Tantillo, I.; Ferrando, L.; Crepaldi, C.; and Ronco, C. 2015. Pro-inflammatory cytokines: a possible relationship with dialytic adequacy and serum albumin in peritoneal dialysis patients. *Clinical kidney journal* 9(1):153–157.

Pham, T.; Tran, T.; Phung, D.; and Venkatesh, S. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 30–41. Springer.

Somers, V. K.; White, D. P.; Amin, R.; Abraham, W. T.; Costa, F.; Culebras, A.; Daniels, S.; Floras, J. S.; Hunt, C. E.; Olson, L. J.; et al. 2008. Sleep apnea and cardiovascular disease: An american heart association/american college of cardiology foundation scientific statement from the american heart association council for high blood pressure research professional education committee, council on clinical cardiology, stroke council, and council on cardiovascular nursing in collaboration with the national heart, lung, and blood institute national center on sleep disorders research (national institutes of health). *Journal of the American College of Cardiology* 52(8):686–717.

Song, H.; Rajan, D.; Thiagarajan, J. J.; and Spanias, A. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, Z.; Ma, Y.; Liu, Z.; and Tang, J. 2019. R-transformer: Recurrent neural network enhanced transformer. *arXiv preprint arXiv:1907.05572*.