MNN: Multimodal Attentional Neural Networks for Diagnosis Prediction

Zhi Qiao*, **Xian Wu**, **Shen Ge** and **Wei Fan**
Tencent Medical AI Lab
{xiaobuqiao, kevinxwu, shenge, davidwfan}@tencent.com

Abstract

Diagnosis prediction plays a key role in clinical decision supporting process, which attracted extensive research attention recently. Existing studies mainly utilize discrete medical codes (e.g., the ICD codes and procedure codes) as the primary features in prediction. However, in real clinical settings, such medical codes could be either incomplete or erroneous. For example, missed diagnosis will neglect some codes which should be included, mis-diagnosis will generate incorrect medical codes. To increase the robustness towards noisy data, we introduce textual clinical notes in addition to medical codes. Combining information from both sides will lead to improved understanding towards clinical health conditions. To accommodate both the textual notes and discrete medical codes in the same framework, we propose Multimodal Attentional Neural Networks (MNN), which integrates multi-modal data in a collaborative manner. Experimental results on real world EHR datasets demonstrate the advantages of MNN in term of accuracy.

1 Introduction

The Electronic Health Records (EHR) data contains the information of patients' visits to the hospital. As shown in Fig. 1, this patient visited hospital on Nov 3rd 2015, Jan 22nd 2016 and Apr 15th 2016 respectively. For each visit, EHR data logs both the discrete medical codes and the textual clinical notes. Predicting possible diagnosis codes in future visit based on previous visits is a critical task, as it can benefit the process of diagnosis and therapy decision. As a result, many machine learning models have been developed for EHR-based phenotyping and event prediction [Choi *et al.*, 2017; Cheng *et al.*, 2016; Qiao *et al.*, 2018a]. In [Ma *et al.*, 2017; Ma *et al.*, 2018b], attention mechanism was introduced to add more interpretability to the prediction results. In [Farhan *et al.*, 2016; Liu *et al.*, 2015], embeddings of low dimensional clinical concepts were used to provide better predictions.

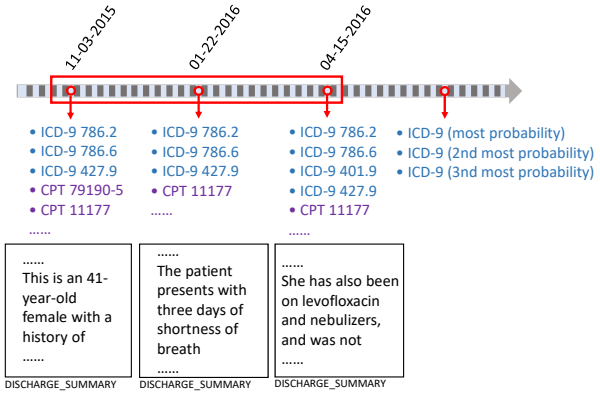

Figure 1: A Segment of Continuous Patient Records Including Discrete Medical Codes and Textual Notes of Discharge Summary

Existing studies mainly utilized discrete medical codes (e.g. diagnosis codes and procedure codes) in modeling. However, the medical codes could be incomplete and erroneous in real clinical setting. For example, some medical codes could be ignored due to missing diagnosis and some codes could be incorrect due to mis-diagnosis. To increase the robustness towards noise medical codes, in addition to discrete codes, we introduce the textual clinical notes in modeling. Notably, thanks to the latest improvement of medical data capturing method, large scale clinical textual notes now become more accessible. Such textual data could be a verification and complementation of the information carried by medical codes, thus leading to a better understanding of patients' clinical health conditions.

To model the medical codes and clinical notes in a unified framework, we propose Multimodal Attentional Neural Networks (MNN). MNN can capture both the information from discrete medical codes and the textual information from clinical notes. For clinical notes, MNN applies a convolutional neural network to mine word-level features and an attentional bidirectional recurrent neural network to mine sentence-level features which are further transformed into medical context aware text features. Meanwhile, MNN embeds multi-hot vectors of medical codes into dense latent vectors as medical code features. After that, MNN uses a deep factorization network to derive multi-modal features which explores the mutual effects of textual features and medical codes. Based on multi-modal features of each visit, we apply attentional

*Corresponding Author

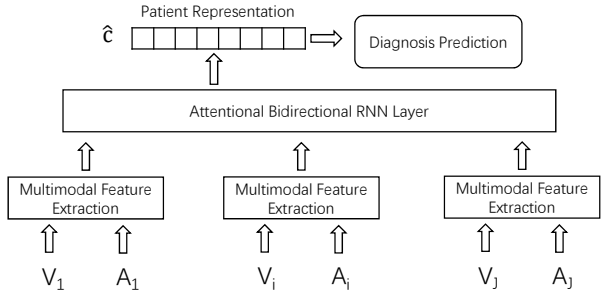

Figure 2: The architecture of MNN

bidirectional recurrent networks to model sequential clinical visits which represents the history of each patient. Finally, we predict the patient's diagnosis on his/her representation. According to the real public datasets MIMIC III, our proposed model outperforms baseline models.

The rest of this paper is organized as follows: In Section 2, we discuss the connection of the proposed approaches and related works. Section 3 presents the preliminary of the work. Section 4 shows the details of the proposed MNN. The experimental results are presented in Section 5. Section 6 concludes the paper.

2 Related Work

Deep Learning on EHR Discrete Data

DeepPatient [Miotto *et al.*, 2016] proposed an unsupervised representation of patient EHR data which can be applied to a large range of predictive tasks. But it has not fully solved the sequential prediction problem of EHR. Sequential prediction of clinical events based on EHR data is a hot research topic and has attracted many attention [Wu *et al.*, 2010]. Most of existing models utilized RNNs for predicting the future diagnosis. RETAIN [Choi *et al.*, 2016b] was an interpretable predictive model, which employed reverse time attention mechanism in an RNN for binary prediction task. Dipole [Ma *et al.*, 2017] employed bidirectional recurrent neural networks and introduced three attention mechanisms to measure the relationships of different visits for the prediction. TLSTM [Baytas *et al.*, 2017] was proposed to handle irregular time intervals by learning a subspace decomposition of the cell memory which enables time decay to discount the memory content according to the elapsed time. DoctorAI [Choi *et al.*, 2016a] was a straightforward approach with simple RNN for sequential patient data modeling. None of these methods collaborate additional clinical text data for diagnosis prediction.

Multimodal Data Modelling

To learn feature representations from multiple aspects, deep neural networks have been successfully applied to various tasks, including but not limited to disease diagnosis [Ma *et al.*, 2018a] and clinical prediction [Xu and Sun, 2018]. RAIM [Xu and Sun, 2018] analyzed both continuous monitoring data and discrete clinical events to predict physiological decompensation and length of stay. ML-MVC [Zhang *et al.*, 2018] was proposed to model multi-view inputs and construct a latent representation to explore the complex correlations between the features and labels of Alzheimer Disease Diagnosis.

To the best of our knowledge, few works have been done on integrating the clinical text and discrete EHR data. Simple concatenation of multiple views may make the parameter space complex due to the heterogeneity of multimodal data, and is thus bad in exploring the complementarity among data from different modalities. So it is necessary to develop an advanced method to extract deep information from the integrated multimodel data.

3 Preliminary

The EHR data of each patient can be represented as a sequence of observations. The i-th patient of N total patients can be represented by a sequence of J_i tuples $(g_j^i), j = 1, \ldots, J_i$. g_j^i represents the corresponding observation data of j-th visit of i-th patient and J_i is the total number of visits of the i-th patient. For notation simplicity, we will describe our algorithm with a single patient and omit the index i. Then the patient can be represented by a sequence of visits $\{g_1, g_2, ..., g_J\}$.

For notation purposes, let $\mathbb{D} = \{d_1, d_2, ..., d_{|\mathbb{D}|}\}$ denote the set of $|\mathbb{D}|$ disease codes, $\mathbb{M} = \{m_1, m_2, ..., m_{|\mathbb{M}|}\}$ as the set of $|\mathbb{M}|$ medical codes which consist of diseases and procedures, $\mathbb{D} \subset \mathbb{M}$. The observation data of each visit g_j contains a subset of medical codes $V_j \subseteq \mathbb{M}$ and a clinical textual note T_j. V_j can be represented as a multihot vector $x_j \in \{0, 1\}^{|\mathbb{M}|}$ where the k-th element is 1 if and only if V_j contains the code m_k. V_j can be seen as a word set.

The core task in MNN is to predict the diagnosis o_{J+1} at $J + 1$-th visit, which is a subset of disease code set $\mathbb{D}$.

4 Methodology

In this section, we will introduce the architecture of proposed Multimodal Attentional Neural Networks (MNN), which is composed of three components: the multi-modal feature extractor, attentional bidirectional recurrent neural networks and diagnosis prediction module. Figure 2 shows the architecture of our proposed MNN.

4.1 Multi-Modal Feature Extractor

The multi-modal feature extractor integrates different types of inputs to extract multi-modal features, which are consisted of three parts, medical code feature extraction, clinical text feature extraction, and deep feature mixture. Figure 3 shows the high-level overview of the multimodal feature extraction module.

Medical Code Feature Extraction

The discrete medical codes are generally considered as multihot vectors with binary features, which could be embedded into dense spaces of real values employing an embedding procedure. Here, we first denote the latent feature matrix of medical codes as $\mathbb{W} \in \mathbb{R}^{|\mathbb{M}| \times l}$, where l is the latent feature dimension and $|\mathbb{M}|$ represents the size of medical code set. Equivalently, the latent feature matrix also can be considered as a linear embedding before activation function. $\mathbb{W}_{z\bullet} \in \mathbb{R}^l$ represents the latent feature vector of z-th medical code. Next, we embed medical codes of each visit x_j into a dense vector π_j by following equation.

$$\pi_j = \texttt{ReLU}(\mathbb{W}^{\mathrm{T}} x_j + b_\pi) \tag{1}$$

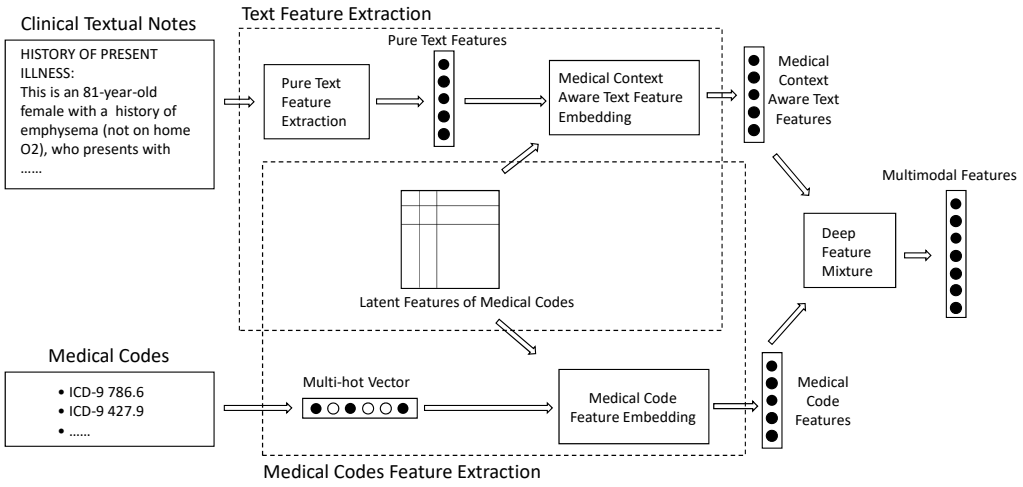

Figure 3: Multimodal Feature Extraction Module of MNN

Clinical Text Feature Extraction

The clinical text feature extraction consists of two parts, one is pure text feature extraction based on original clinical text information, and the other is medical context aware text feature embedding which correlates text data with medical codes to make compensation for discrete medical codes.

Pure Text Feature Extraction

Since clinical notes have a hierarchical structure (words form sentences, sentences form a document), we likewise construct a document representation by first building representations of sentences and then aggregating them into a document representation. For sentence representations, we use a specific Convolutional neural networks by using multiple filters with various window sizes to capture different granularities of word-level features, which is also used in [Kim, 2014]. For document representations, in order to leverage the effects of different importance in different contexts, we use bidirectional recurrent neural networks with attention mechanism to ensemble the hidden states into a final document representation.

In textual feature extractor, each word in the text is represented as a word embedding vector. The embedding vector for each word is initialized with the pre-trained word embedding on the given dataset. For the k-th word in u-th sentence of the j-th visit notes, the corresponding r-dimensional word embedding vector is denoted as $\omega_{k;u;j} \in \mathbb{R}^r$. Thus, u-th sentence of j-th note, with n words can be represented as:

$$\omega_{[1:n];u;j} = \omega_{1;u;j} \oplus \omega_{2;u;j} \oplus \omega_{3;u;j} ... \oplus \omega_{n;u;j} \qquad (2)$$

where $\oplus$ is the concatenation operator. There exist κ convolutional filters, each of which has a different window size. For example, the t-th convolutional filter with window size ϖ^t takes the contiguous sequence of ϖ^t words in the sentence as input and outputs one sentence feature. In order to show the procedure clearly, we take the contiguous sequence of ϖ^t words starting with the γ-th word as example, the filter operation can be represented as:

$$\begin{aligned} &\rho^t_{u;j} = \{\rho^t_{1;u;j}, ..., \rho^t_{\gamma;u;j}, ...\} \\ &s.t.\ \ \rho^t_{\gamma;u;j} = \texttt{ReLU}(W^t_\omega \cdot \omega_{[\gamma:\gamma+\varpi^t-1];u;j}) \end{aligned} \qquad (3)$$

where W^t_ω represents the weight of the t-th filter. The filter can also be applied to the rest of words. We use max-pooling operation to take the maximum value, extracting the most important information for $\rho^t_{u;j}$. Then we get a feature vector with κ dimensions for u-th sentence of the j-th visit notes.

For every sentence feature vector $s_{u;j}$ (where $u = 1, \ldots, Q$) of a clinical note, we use a bidirectional GRU [Cho *et al.*, 2014] to get the annotations of sentences by summarizing information from both directions for sentences, and therefore incorporate the contextual information in the annotations. The bidirectional GRU contains the forward GRU which reads the note document from $s_{1;j}$ to $s_{Q;j}$ and a backward GRU which reads from $s_{Q;j}$ to $s_{1;j}$:

$$\begin{aligned} \overrightarrow{h^s_j} &= \overrightarrow{\texttt{GRU}}(s_{u;j}), u = 1...Q \\ \overleftarrow{h^s_j} &= \overleftarrow{\texttt{GRU}}(s_{u;j}), u = Q...1 \end{aligned} \qquad (4)$$

where $\overrightarrow{h^s_j} \in \mathbb{R}^r$ and $\overleftarrow{h^s_j} \in \mathbb{R}^r$.

We obtain an annotation for a given sentence by concatenating the forward backward hidden states, i.e., $h^s_j = [\overrightarrow{h^s_j}, \overleftarrow{h^s_j}]$, which summarizes the information of the whole sentence centered around a given word. Not all sentences contribute equally to the meaning representation of the note. Hence, we introduce an attention mechanism to extract such sentences that are important to the meaning of the note.

$$\alpha_{u;j} = W^T_d h^s_{u;j} + b_d \qquad (5)$$

where $W_d \in \mathbb{R}^{2r \cdot 1}$ and $b_d \in \mathbb{R}$. Eq.5 is followed by $\hat{\alpha}_j = \texttt{softmax}(\alpha_{[1:Q];j})$. Next, we aggregate the representation of those informative sentences to form a note vector $\lambda_j = \sum_{u=1}^{Q} \hat{\alpha}_{u;j} h^s_{u;j}$ for the j-th patient visit and $\lambda_j \in \mathbb{R}^{2r}$.

Medical Context Aware Text Feature Embedding

Based on the inpatient states, doctors flag proper medical codes for patients. In real settings, some latent diseases or

symptoms generally just occur in the clinical notes. Due to mis-diagnosis and missing diagnosis, sparse medical codes maybe incomplete. Hence, it's crucial to correlate clinical text information with medical codes to enrich patients' presentation.

We have used the parameter matrix $\mathbb{W}$ to represent medical code latent features. Based on the pure text feature λ_j of clinical notes, we can calculate the correlation of clinical notes with each medical code as,

$$\varepsilon_{z;j} = \mathbb{W}_{z\bullet} W_\varepsilon \cdot \lambda_j + b_\varepsilon, \quad z = 1, ..., |\mathbb{M}| \tag{6}$$

where $W_\varepsilon \in \mathbb{R}^{l \times r}$ denotes parameter matrix to capture the correlation, and $b_\varepsilon \in \mathbb{R}$ is bias. Then we can get the medical code distribution $\hat{\varepsilon}_j \in \mathbb{R}^{|\mathbb{M}|}$ of clinical note via,

$$\hat{\varepsilon}_j = \texttt{softmax}([\varepsilon_{1;j}, \varepsilon_{2;j}, \ldots, \varepsilon_{|\mathbb{M}|;j}]) \tag{7}$$

After we get medical code distribution, we can obtain medical context aware text embedding feature as in medical code embedding process.

$$\tau_j = \texttt{ReLU}(\mathbb{W}^{\mathrm{T}} \hat{\varepsilon}_j + b_\tau) \tag{8}$$

where $\mathbb{W} \in \mathbb{R}^{|\mathbb{M}| \times l}$ is medical code latent feature matrix employed in subsection in 4.1.1 and $b_\tau \in \mathbb{R}^l$ are specific parameters.

Deep Feature Mixture

After we get the textual feature representation τ_j and medical code feature representation π_j, we use deep feature mixture module to generate the final multi-modal feature representation.

Because τ_j and π_j come from different feature domains. We need to embed them into a uniform feature space considering cross domain features. Compressed Interaction Network (CIN) [Lian, 2018] proposed to apply vector-wise level computation to extract explicit interaction features, which has been proven to be effective in extracting cross domain features. In our model, inter-domain interactions are applied at vector-wise level and we also use DNN to extract implicit interaction features.

First, we concatenate two representation together as $\chi_j = [\tau_j, \pi_j] \in \mathbb{R}^{2l}$. For interaction layer, $\tilde{\xi_{jk}^H}$ is calculated via,

$$\tilde{\xi_{jk}^H} = \sum_{\epsilon=1}^{2l} \sum_{\zeta=1}^{2l} W_{\epsilon,\zeta}^k (\chi_j \otimes \chi_j) \tag{9}$$

where $W_{\epsilon,\zeta}^k \in \mathbb{R}^{2l \times 2l}$ is the parameter matrix, $\otimes$ denotes the outer product, and k represents k-th feature map to extract correlation among features. Here we assume there are in total q feature maps, and hence we can get $\tilde{\xi_j^H} \in \mathbb{R}^q$.

For implicit interaction features, we employ fully-connected layer having the following formula:

$$\tilde{\xi_j^L} = \texttt{ReLU}(W_\chi \chi_j + b_\chi) \tag{10}$$

where $W_\chi \in \mathbb{R}^{q \times 2l}$, $b_\chi \in \mathbb{R}^q$ are parameters. And, we can get $\tilde{\xi_j^L} \in \mathbb{R}^q$.

Since both parts about explicit and implicit interaction feature learning can be a complement to each other, an intuitive way to make the model stronger is to combine these two structures by concatenation, i.e., $\Lambda_j = [\tilde{\xi_j^L}, \tilde{\xi_j^H}]$.

4.2 Attentional Bidirectional RNN

Recurrent Neural Networks (RNN) provide a very elegant way of modeling sequential healthcare data. Here, we employ Bidirectional Recurrent Neural Networks (BiRNN) [Schuster and Paliwal, 1997] in the proposed model which can be trained using all the available input visits' information from two directions to improve the prediction performance.

A BiRNN consists of a forward and backward RNN. The forward RNN $\overrightarrow{f}$ reads the input visit sequence from Λ_1 to Λ_J and calculates a sequence of forward hidden states $(h_1^f, h_2^f, \ldots, h_J^f)$ ($h_i^f \in \mathbb{R}^p$ and p is the dimensionality of hidden states). The backward RNN $\overleftarrow{b}$ reads the visit sequence in the reverse order, i.e., from Λ_J to Λ_1, resulting in a sequence of backward hidden states $(h_1^b, h_2^b, \ldots, h_J^b)$ ($h_i^b \in \mathbb{R}^p$).

$$(h_1^f, h_1^b), (h_2^f, h_2^b), ..., (h_J^f, h_J^b) = \texttt{BiRNN}(\Lambda_1, \Lambda_2, ..., \Lambda_J) \tag{11}$$

Bidirectional Attention

We separately compute the specific attention weights for forward and backward hidden states by adopting the approach similar to [Bahdanau *et al.*, 2015]. In particular, α_r^f is computed via $\alpha_r^f = \texttt{softmax}([e_1^f, e_2^f, ..., e_J^f])$, where $e_j^f = \texttt{ReLU}(W^f h_j^f + b^f)$; α_r^b is computed via $\alpha_r^b = \texttt{softmax}([e_1^b, e_2^b, ..., e_J^b])$, where $e_j^b = \texttt{ReLU}(W^b h_j^b + b^b)$. And $W^f \in \mathbb{R}^{1 \times p}$, $W^b \in \mathbb{R}^{1 \times p}$, $b^f \in \mathbb{R}$ and $b^b \in \mathbb{R}$ are the parameters to be learned.

At each step, we can compute the correlation weight of forward-backward by,

$$e^{fb} = [e_1^{fb}, e_2^{fb}, ..., e_J^{fb}],$$
$$\text{where } e_j^{fb} = \texttt{sigmoid}(h_j^f W^{fb} h_j^b + b^{fb})$$

where $W^{fb} \in \mathbb{R}^{p \times p}$ and $b^{fb} \in \mathbb{R}$ are the parameters to be learned.

Based on e^{fb}, α_r^f, α_r^b and forward & backward hidden states, we can derive contextual state as $c = \sum_{j=1}^J e^{fb} \cdot \alpha_r^f h_j^f + (1 - e^{fb}) \cdot \alpha_r^b h_j^b$. Concatenated with the last hidden states h_J^f and h_J^b, we have the patient representation $\hat{c} = \texttt{ReLU}(W_d[h_J^f, h_J^b, c])$.

4.3 Diagnosis Prediction

The patient representation $\hat{c}$ is fed through the softmax layer to produce the $(J+1)$-th visit diagnosis o_{J+1} defined as:

$$\hat{o}_{J+1} = \texttt{softmax}(W_s \hat{c} + b_s) \tag{12}$$

where $W_s \in \mathbb{R}^{3p \times |\mathbb{D}|}$ and $b_s \in \mathbb{R}^{|\mathbb{D}|}$ are the parameters to be learned.

Based on Eq.12, we use the cross-entropy between the ground truth diagnosis o_{J+1} and the predicted diagnosis $\hat{o}_{J+1}$ to calculate the loss for all the patients as follows:

$$L = -\frac{1}{N} \sum_{i=1}^{N} o_{J+1} log(\hat{o}_{J+1}) + (1 - o_{J+1}) log(1 - \hat{o}_{J+1}) \tag{13}$$

Data	Methods	Recall@10	Recall@20	Recall@30	Precision@10	Precision@20	Precision@30
Baseline Methods	Dipole	0.35795	0.48998	0.58314	0.31139	0.22091	0.17459
	Retain	0.34074	0.48274	0.57046	0.30422	0.21624	0.17131
	DoctorAI	0.33271	0.47628	0.56098	0.29876	0.20221	0.16542
	PacRNN	0.36212	0.49877	0.60821	0.32123	0.22791	0.17981
	RNN-multimodal	0.33679	0.47961	0.56479	0.30137	0.20576	0.16799
Variant MNN	MNN-text	0.35131	0.50998	0.60031	0.31987	0.23077	0.17981
	MNN-code	0.35989	0.51678	0.60466	0.32321	0.23249	0.18143
	MNN-avg	0.36556	0.51762	0.61087	0.32987	0.23344	0.18311
	MNN	**0.37489**	**0.51846**	**0.61619**	**0.33333**	**0.23789**	**0.18439**

Table 1: Performance Comparison on MIMIC III Data where the size of predicted diagnosis space is 700

5 Experiments

We evaluate our model MNN on the publicly available real-world data sets. We show that MNN outperforms our baselines.

5.1 Dataset

We use a publicly available multimodal EHR data, MIMIC-III released on PhysioNet [Goldberger *et al.*, 2000]. The data set constituted of 46,520 patients contains deidentified comprehensive clinical medical codes and rich clinical textual notes from intensive care units (ICU) at the Beth Israel Deaconess Medical Center between 2001 and 2012. In the dataset, the average time interval between two consecutive visits is 349.5 days, the 1/4-quantile is 39 days. We remove the patient with less than three visits, and after this filtering, the average sequence length is 3.87.

Data Preprocess

Each visit is represented by a set of medical codes, including disease codes (ICD 9) and procedure codes, and corresponding dispatching textual notes. To reduce the size of feature set and avoid information overload, we group codes into coarse-grained categories as [Choi *et al.*, 2016a]. For both disease and procedure codes, we extract the top-3 digits, yielding 700 disease groups and 740 procedure groups, and the size of predicted diagnosis space is also 700. For clinical textual notes, we first reorganize text as sequential sentences, then preprocess the sentences to generate sequential word sets for each sentence.

Baseline Methods

In order to verify the performance gain by introducing clinical text, medical codes and attention mechanism, we create three variants for our proposed model MNN: modelling just clinical text data (**MNN-text**), modelling just medical code data (**MNN-code**) and modelling via integrating average outputs of recurrent neural networks without using attention mechanism(**MNN-avg**).

We then compare our methods with the baseline approaches for diagnosis prediction. Here we list the models compared in our experiments.

- **DoctorAI**: [Choi *et al.*, 2016a] embeds visits into vector representations and then feeds them into the GRUs. The hidden states of the GRUs are directly used to predict the diagnosis of the future visit.
- **RETAIN**: [Choi *et al.*, 2016b] proposes an interpretable predictive model in healthcare with reverse time attention mechanism.
- **Dipole**: [Ma *et al.*, 2017] uses attention-based bidirectional recurrent neural networks for diagnosis prediction.
- **PacRNN**: [Qiao *et al.*, 2018b] embeds medical codes with attentional RNN, then uses Bayesian Personalized Ranking (BPR) regularized by disease co-occurrence to rank probabilities of patient-specific diseases.
- **RNN-multimodal**: embeds concatenated text features and medical code features into basic RNN without employing deep feature mixture, then directly uses average outputs of RNN for final diagnosis prediction.

Evaluation Metrics

The performance of algorithms in predicting diagnoses was evaluated using the Top-k recall and Top-k precision. Top-k recall & precision mimic the behavior of doctors conducting differential diagnosis, where doctors list most probable diagnoses and treat patients accordingly to identify the patient status. Therefore, a machine with a high Top-k recall & precision translates to a doctor with an effective diagnostic skill. This makes Top-k recall & precision attractive performance metrics for our problem. In our experiments, we separately set k to be 10, 20, and 30 for both Recall and Precision.

Implementation details

We first learn the word vectors via an unsupervised neural language model which is a popular method to improve performance in absence of a large supervised training set. The word2vec vectors were trained on the medical textual notes. The vectors have dimensionality of 128.

In all experiments, the learning rate is set to be 0.001, embedding size $l = 64$ and hidden state size $r = 128$ for our methods. We also use regularization (l2 norm with the coefficient 0.001), drop-out strategies (with the drop-out rate 0.5) and batch size 20 for all methods.

We implement all the models with Tensorflow 1.4 [Abadi *et al.*, 2015].

5.2 Experimental Results

Diagnosis Prediction Results

Table 1 shows the accuracies of the proposed MNN model and baselines on MIMIC III datasets for the diagnosis prediction task.

	Real Codes	Description	DoctorAI	Description	MNN	Description
1	599	Other disorders of urethra and urinary tract	496	Chronic airway obstruction, not elsewhere classified	599*	**Other disorders of urethra and urinary tract**
2	428	Heart failure	038	Septicemia	487	With pneumonia
3	584	Acute renal failure	682	Other cellulitis and abscess	428*	**Heart failure**
4	410	Acute myocardial infarction	599*	**Other disorders of urethra and urinary tract**	995	Certain adverse effects not elsewhere classified
5	425	Cardiomyopathy	V43	Organ or tissue replaced by other means	250*	**Diabetes mellitus**
6	250	Diabetes mellitus	280	Iron deficiency anemias	585	Chronic kidney disease
7	135	Sarcoidosis	250*	**Diabetes mellitus**	038	Septicemia
8	799	Other causes of morbidity and mortality	511	Pleurisy	584*	**Acute renal failure**
9	486	Pneumonia, organism unspecified	135*	**Sarcoidosis**	707	Chronic ulcer of skin
10			789	Other symptoms involving abdomen and pelvis	135*	**Sarcoidosis**

Table 2: Case Study: Comparison of predicted next diagnosis for a real patient in MIMIC III data. (The example patient has 9 diseases in next clinical visits with 9 different diagnosis codes and the diagnosis codes are ordered by priority where the order does have an impact on the reimbursement for treatment. The diagnosis codes marked with an asterisk are correctly predicted.)

We note that the accuracy of DoctorAI is somewhat lower than others on real public datasets. The main reason is that DoctorAI is the only one without using attention mechanism. It predicts the diagnosis depending on the last hidden state of the RNN, which lacks this capability to memorize all the past information, making it focus on the recent visits information only. RETAIN, Dipole, PacRNN and our methods can take all the visits into consideration. By assigning different attention weights to each visit, these methods achieve better performance than DoctorAI. MNN gets better performance on all of recall and precision measurements than baseline methods. Compared with Retain, Dipole and PacRNN, MNN utilized extra clinical text data along with medical code data.

RNN-multimodal also utilizes multimodal data, but gets lower performance than some of baseline methods. The main reasons are 1) the heterogeneity of multimodal data making the parameter space so complex that could not be settled down via simple concatenation of heterogeneous features; 2) using average outputs as patient representation lacks consideration on different impacts of different visits for diagnosis outcomes. Compared with RNN-multimodal, MNN can get higher accuracy, due to the effectiveness of our proposed multimodal features extractor which can model different types of inputs and learn significant features using different kinds of data, and also the advantages of attention mechanism for deriving context vector that captures relevant information from historical multiple visits to help prediction.

In order to verify the performance gain by introducing clinical text, medical codes and attentional mechanism separately, we also implement experiments to compare MNN with three variants (MNN-text, MNN-code, and MNN-avg). From Table 1, we can find that just using text information or using medical code data cannot get good performance compared to the fully powered MNN. Using just average process to combining sequential data without considering attention mechanism, MNN-avg is still less accurate than our full MNN.

Case Study

Table 2 shows top-10 prediction comparison between MNN and DoctorAI. Our method has better ranking performance (3 correct ones in top-5) than DoctorAI (1 in top-5). It is worth noting that ICD code 428 (Heart failure) is correctly diagnosed by MNN in the rank positions of 4, which is not observed in the historical visits of this patient. Actually, some cardiovascular related diseases occurred in the patient's history which demonstrates that our method is capable of predicting new diseases via modelling disease evolving. ICD code 584 (Acute renal failure) is also correctly diagnosed by MNN in the rank positions of 8, which is not observed in the historical visits of this patient. The patient has been diagnosed as disorders of urethra and urinary tract in the historical visits and kidney related discussions occur in clinical notes multiple times. With text information strengthen and disease evolving modelling, three kidney related diseases (585, 487 and 584) are predicted by MNN and ranked higher. Collaborate clinical textual notes can help for correct predictions.

6 Conclusions

In this paper, in order to increase the robustness towards medical code data, we introduce textual clinical notes to improved understanding towards clinical health conditions. We propose Multimodal Attentional Neural Networks (MNN) which can model the historical clinical multimodal data, including both medical code data and textual note data in a unified fashion. Experimental results on real world EHR dataset demonstrate the good performance of MNN.

Acknowledgements

This work was supported by National Key R&D Program of China, No. 2018YFC0117000.

We also thank the anonymous reviewers for their valuable comments.

References

[Abadi *et al.*, 2015] Martín Abadi *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Software available from tensorflow.org.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*, 2015.

[Baytas *et al.*, 2017] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. Patient Subtyping via Time-Aware LSTM Networks. In *KDD*, pages 65–74, 2017.

[Cheng *et al.*, 2016] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *SDM*, pages 432–440, 2016.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*, pages 1724–1734, 2014.

[Choi *et al.*, 2016a] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *MLHC*, pages 301–318, 2016.

[Choi *et al.*, 2016b] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *NIPS*, pages 3504–3512, 2016.

[Choi *et al.*, 2017] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

[Farhan *et al.*, 2016] Wael Farhan, Zhimu Wang, Yingxiang Huang, Shuang Wang, Fei Wang, and Xiaoqian Jiang. A Predictive Model for Medical Events Based on Contextual Embedding of Temporal Sequences. *JMIR Medical Informatics*, 4(4):e39, 2016.

[Goldberger *et al.*, 2000] Ary Goldberger, Luis Amaral, Leon Glass, Jeffrey Hausdorff, Plamen Ivanov, Roger Mark, Joseph Mietus, George Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[Kim, 2014] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *EMNLP*, pages 1746–1751, 2014.

[Lian, 2018] Jianxun Lian. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. In *KDD*, pages 1754–1763, 2018.

[Liu *et al.*, 2015] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework. In *KDD*, pages 705–714, 2015.

[Ma *et al.*, 2017] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *KDD*, pages 1903–1911, 2017.

[Ma *et al.*, 2018a] Feanglong Ma, Quanzeng You, Houping Xia, and Jing Gao. KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare. In *KDD*, pages 743–752, 2018.

[Ma *et al.*, 2018b] Tengfei Ma, Cao Xiao, and Fei Wang. Health-ATM: A Deep Architecture for Multifaceted Patient Health Record Representation and Risk Prediction. In *SDM*, pages 261–269, 2018.

[Miotto *et al.*, 2016] Riccardo Miotto, Li Li, Brian Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6:26094, 05 2016.

[Qiao *et al.*, 2018a] Zhi Qiao, Ning Sun, Xiang Li, Eryu Xia, Shiwan Zhao, and Yong Qin. Using Machine Learning Approaches for Emergency Room Visit Prediction Based on Electronic Health Record Data. *Studies in health technology and informatics*, 247:111–115, 01 2018.

[Qiao *et al.*, 2018b] Zhi Qiao, Shiwan Zhao, Cao Xiao, Xiang Li, Fei Wang, and Yong Qin. Pairwise-Ranking based Collaborative Recurrent Neural Networks for Clinical Event Prediction. In *IJCAI*, pages 3520–3526, 2018.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K. Paliwal. Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal Processing*, 45(11):2673–2681, 1997.

[Wu *et al.*, 2010] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6):S106–S113, 2010.

[Xu and Sun, 2018] Yanbo Xu and Jimeng Sun. RAIM Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data. In *KDD*, pages 2565–2573, 2018.

[Zhang *et al.*, 2018] Changqing Zhang, Ehsan Adeli, Tao Zhou, and Xiaobo Chen. Multi-Layer Multi-View Classification for Alzheimer's Disease Diagnosis. In *AAAI*, pages 4406–4413, 2018.