# MULTI-LAYER REPRESENTATION LEARNING FOR MEDICAL CONCEPTS

**Edward Choi**[1], **Mohammad Taha Bahadori**[1], **Elizabeth Searles**[2], **Catherine Coffey**[2], **Jimeng Sun**[1]
[1]Georgia Institute of Technology, [2]Children's Healthcare of Atlanta

## ABSTRACT

Learning efficient representations for concepts has been proven to be an important basis for many applications such as machine translation or document classification. Proper representations of medical concepts such as diagnosis, medication, procedure codes and visits will have broad applications in healthcare analytics. However, in Electronic Health Records (EHR) the visit sequences of patients include multiple concepts (diagnosis, procedure, and medication codes) per visit. This structure provides two types of relational information, namely sequential order of visits and co-occurrence of the codes within each visit. In this work, we propose `Med2Vec`, which not only learns distributed representations for both medical codes and visits from a large EHR dataset with over 3 million visits, but also allows us to interpret the learned representations confirmed positively by clinical experts. In the experiments, `Med2Vec` displays significant improvement in key medical applications compared to popular baselines such as Skip-gram, GloVe and stacked autoencoder, while providing clinically meaningful interpretation.

## 1 INTRODUCTION

Discovering efficient representations of discrete high dimensional concepts has been a key challenge in a variety of applications recently (Bengio et al., 2013). Efficient representations for concepts is an important, if not essential, element in healthcare as well. Healthcare concepts contain rich latent relationships that cannot be represented by simple one-hot coding. To overcome this limitation, it is common in healthcare applications, to rely on carefully designed feature representations (Sun et al., 2012; Ghassemi et al., 2014; Wang et al., 2015). However, this process often involves supervision information and ad-hoc feature engineering that requires considerable expert medical knowledge and is not scalable in general.

Recently, studies have shown that it is possible to learn efficient representations of healthcare concepts without medical expertise and still significantly improve the performance of various healthcare applications (Choi et al., 2016b;c; 2015). Despite these progress, learning efficient representations of healthcare concepts, however, is still an open challenge. The difficulty stems from several aspects:

1. Healthcare data have a unique structure where the visits are temporally ordered but the medical codes within a visit form an unordered set. A sequence of visits possesses sequential relationship among them which cannot be captured by simply aggregating code-level representations.

2. Learned representations should be interpretable. While the interpretability of the model in the clinical domain is considered to be an essential requirement, some of the state-of-the art representation learning methods such as recurrent neural networks (RNN) are difficult to interpret.

3. The algorithm should be scalable enough to handle real-world healthcare datasets with millions of patients and hundred millions of visits.

The unique structure of healthcare data shares some analogy with the sentence-word structure in text. Therefore some of the recent studies in natural language processing such as (Le & Mikolov, 2014) and (Kiros et al., 2015) are somewhat related to `Med2Vec`. Le & Mikolov (2014) proposes to learn representations for a fixed number of paragraphs and words simultaneously by treating paragraphs as one of the words, although their approach does not capture the sequential order among paragraphs. Skip-thought (Kiros et al., 2015) proposes to learn representations of sentences and words by enforcing the GRU to read the sentence and regenerate the surrounding sentences. However, Skip-thought cannot be applied directly to EHR data because unlike words in sentences, the

codes in a visit are unordered. Also, the interpretation of Skip-thought model is difficult, as they rely on complex RNNs. To address such challenges in healthcare concept representation learning, we propose `Med2Vec`, a simple and robust algorithm that learns code and visit representations from a real-world health dataset of 3 million visits, without depending on expert medical knowledge. `Med2Vec` generates interpretable representations of codes and visits, and shows superior performance in visit-level prediction tasks compared baselines such as Skip-gram, GloVe and stacked autoencoder.

## 2 METHOD

**EHR structure and our notation** We denote the set of all medical codes $c_1, c_2, \ldots, c_{|\mathcal{C}|}$ in our EHR dataset by $\mathcal{C}$ with size $|\mathcal{C}|$. EHR data for each patient is in the form of a sequence of visits $V_1, \ldots, V_T$ where each visit contains a subset of medical codes $V_t \subseteq \mathcal{C}$. Without loss of generality, all algorithms will be presented for a single patient to avoid cluttered notations. The goal of `Med2Vec` is to learn two types of representations.

- Code representations: We aim to learn an embedding function $f_C : \mathcal{C} \mapsto \mathbb{R}^m_+$ that maps every code in the set of all medical codes $\mathcal{C}$ to non-negative real-valued vectors of dimension $m$. The non-negativity constraint is introduced to improve interpretability.

- Visit representations: Our second task is to learn another embedding function $f_V : \mathcal{V} \mapsto \mathbb{R}^n$ that maps every visit (a set of medical codes) to a real-valued vector of dimension $n$. The set $\mathcal{V}$ is the power set of the set of codes $\mathcal{C}$.

**Med2Vec Architecture** Figure 1 depicts the architecture of `Med2Vec`. Given a visit $V_t$, we use a multilayer perceptron (MLP) to to generate the corresponding visit representation $\boldsymbol{v}_t$. First, visit $V_t$ is represented by a binary vector $\boldsymbol{x}_t \in \{0,1\}^{|\mathcal{C}|}$ where the $i$-th entry is 1 only if $c_i \in V_t$. Then $\boldsymbol{x}_t$ is converted to an intermediate visit representation $\boldsymbol{u}_t \in \mathbb{R}^m = ReLU(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{b}_c)$ using the code weight matrix $\boldsymbol{W}_c \in \mathbb{R}^{m \times |\mathcal{C}|}$ and the bias vector $\boldsymbol{b}_c \in \mathbb{R}^m$. The rectified linear unit is defines as $ReLU(\boldsymbol{v}) = \max(\boldsymbol{v}, \boldsymbol{0})$. Note that $\max()$ applies element-wise to vectors. We use the rectified linear unit (ReLU) as the activation function to enable interpretability.

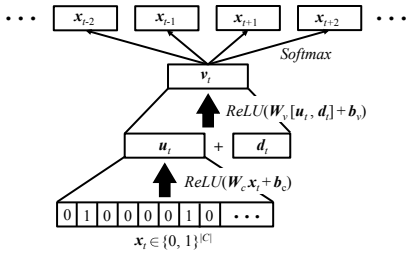


Figure 1: The architecture of Med2Vec.

We concatenate the demographic information $\boldsymbol{d}_t \in \mathbb{R}^d$, where $d$ is the size of the demographic information vector, to the intermediate visit representation $\boldsymbol{u}_t$ and create the final visit representation $\boldsymbol{v}_t \in \mathbb{R}^n = ReLU(\boldsymbol{W}_v[\boldsymbol{u}_t, \boldsymbol{d}_t] + \boldsymbol{b}_v)$ using the visit weight matrix $\boldsymbol{W}_v \in \mathbb{R}^{n \times (m+d)}$ and the bias vector $\boldsymbol{b}_v \in \mathbb{R}^n$, where $n$ is the predefined size of the visit representation. We use ReLU once again as the activation function.

**Learning visit representations** By exploiting the sequential information of visits, we can learn efficient visit representations. Specifically, given a visit representation $\boldsymbol{v}_t$, we train a softmax classifier that predicts the medical codes of the visits within a context window. We minimize the cross entropy error as follows,

$$\min_{\boldsymbol{W}_s, \boldsymbol{b}_s} \frac{1}{T} \sum_{t=1}^{T} \sum_{-w \le i \le w, i \neq 0} -\boldsymbol{x}_{t+i}{}^\top \log \hat{\boldsymbol{y}}_t - (\boldsymbol{1} - \boldsymbol{x}_{t+i})^\top \log(\boldsymbol{1} - \hat{\boldsymbol{y}}_t), \quad \text{where } \hat{\boldsymbol{y}}_t = \frac{\exp(\boldsymbol{W}_s \boldsymbol{v}_t + \boldsymbol{b}_s)}{\sum_{j=1}^{|\mathcal{C}|} \exp(\boldsymbol{W}_s[j,:]\boldsymbol{v}_t + \boldsymbol{b}_s[j])} \tag{1}$$

where $\boldsymbol{W}_s \in \mathbb{R}^{|\mathcal{C}| \times n}$ and $\boldsymbol{b}_s \in \mathbb{R}^{|\mathcal{C}|}$ are the weight matrix and bias vector for the softmax classifier, $w$ the predefined context window size, $\exp$ the element-wise exponential function, and $\boldsymbol{1}$ denotes an all one vector. We have used MATLAB's notation for selecting a row in $\boldsymbol{W}_s$ and a coordinate of $\boldsymbol{b}_s$.

**Learning code representations** We learn code representations based on the co-occurrence of codes within a visit. The natural choice for learning representations based on intra-visit code co-occurrence would be to use Skip-gram (Mikolov et al., 2013) for training $\boldsymbol{W}_c$. However, to enable precise interpretation of the code representations by each coordinate, we enforce non-negativity to the code representations by using $ReLU(\cdot)$ function. Then the code representations to be learned is

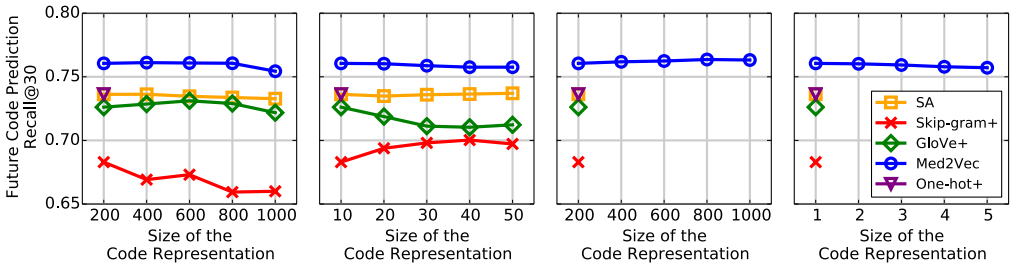


Figure 2: Prediction performance of Med2Vec and baseline models.

denoted as a matrix $\boldsymbol{W}'_c = ReLU(\boldsymbol{W}_c) \in \mathbb{R}^{m \times |\mathcal{C}|}$. From a sequence of visits $V_1, V_2, \ldots, V_T$, the code-level representations can be learned by maximizing the following log-likelihood,

$$\min_{\boldsymbol{W}'_c} \quad \frac{1}{T}\sum_{t=1}^{T}\sum_{i:c_i \in V_t}\sum_{j:c_j \in V_t, j \neq i} \log p(c_j|c_i), \quad \text{where} \quad p(c_j|c_i) = \frac{\exp\left(\boldsymbol{W}'_c[:,j]^\top \boldsymbol{W}'_c[:,i]\right)}{\sum_{k=1}^{|\mathcal{C}|}\exp\left(\boldsymbol{W}'_c[:,k]^\top \boldsymbol{W}'_c[:,i]\right)}. \quad (2)$$

**Unified training** The single unified framework can be obtained by adding the two objective functions (2) and (1) as follows,

$$\min_{\boldsymbol{W},\boldsymbol{b}} \frac{1}{T}\sum_{t=1}^{T}\Big\{ -\sum_{i:c_i \in V_t}\sum_{j:c_j \in V_t, j \neq i} \log p(c_j|c_i) + \sum_{-w \leq k \leq w, k \neq 0} -\boldsymbol{x}_{t+k}^\top \log \hat{\boldsymbol{y}}_t - (\mathbf{1} - \boldsymbol{x}_{t+k})^\top \log(\mathbf{1} - \hat{\boldsymbol{y}}_t)\Big\}$$

By combining the two objective functions we learn both code representations and visit representations from the same source of patient visit records, exploiting both intra-visit co-occurrence information as well as inter-visit sequential information at the same time.

## 3 Experiments

We use the EHR dataset provided by Children's Healthcare of Atlanta, which consists of 550,339 patients, 3,359,240 visits and 28,840 unique medical codes. ICD9 codes, NDC codes and CPT codes are used respectively for diagnosis, medication and procedure codes. We devide the dataset by 4:1 ratio, where we use the former to train `Med2Vec` and the latter to evaluate the learned representations. For evaluation, we predict the medical codes that will occur in the next visit using the visit representations. Specifically, given two consecutive visits $V_i$ and $V_j$, the medical codes $c \in V_j$ will be the target $\boldsymbol{y}$, the medical codes $c \in V_i$ will be the input $\boldsymbol{x}$, which will be converted to a visit representation via the trained model. We use softmax to predict $\boldsymbol{y}$ given $\boldsymbol{x}$. The predictive performance will be measured by Top-$k$ Recall, which mimics the differential diagnosis conducted by doctors. We set $k = 30$ to cover even the complex cases of CHOA dataset, as over 167,000 visits are assigned with more than 20 medical codes. The baseline models are simple sum of one-hot vectors (One-hot+), the sum of Skip-gram vectors (Skip-gram+), the sum of GloVe vectors (GloVe+) and stacked autoencoder (SA). We conduct the experiment with various values for different hyperparameters: the size of the code representation, the size of the visit representation, the size of the visit context window and the number of training epochs. The results in Fig.2 confirms that the representations learned by `Med2Vec` consistently show superior performance in various settings.

We also trained a logistic regression model on top of the learned visit representations to predict the severity level of the patient at the given visit. The learned model can be then approximately interpreted to find out which diseases positively activate the severity of the patient by computing the code representation that maximizes the logistic regression output. Then the obtained code representation can be interpreted by finding the medical codes that have the strongest value in each coordinate. We have found that diseases related to congenital chromosome anomalies (*e.g.* Down's syndrome) and congenital paralysis (*e.g.* congenital quadraplegia) are the strongest influencer of patient severity, which was positively confirmed by the medical experts in CHOA.

We also performed additional visit-level evaluation, code-level evaluation and interpretation of the learned representations, all of which can be found in the full version (Choi et al., 2016a) of this extended abstract.

## REFERENCES

Yoshua Bengio, Aaron Courville, and Pierre Vincent. Representation learning: A review and new perspectives. *PAMI*, 2013.

Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. *arXiv preprint arXiv:1511.05942*, 2015.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, and Jimeng Sun. Multi-layer representation learning for medical concepts. *arXiv preprint arXiv:1602.05568*, 2016a.

Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*, 2016b.

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. 2016c. To be submitted to AMIA CRI.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *KDD*, 2014.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *NIPS*, 2015.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

Jimeng Sun, Fei Wang, Jianying Hu, and Shahram Edabollahi. Supervised patient similarity measure of heterogeneous patient records. *KDD Explorations*, 2012.

Yajuan Wang, Kenney Ng, Roy J Byrd, Jianying Hu, Shahram Ebadollahi, Zahra Daar, Christopher deFilippi, Steven R Steinhubl, and Walter F Stewart. Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records. In *EMBC*, 2015.