DOCTOR AI: PREDICTING CLINICAL EVENTS VIA RECURRENT NEURAL NETWORKS

Edward Choi[1], **Mohammad Taha Bahadori**[1], **Andy Schuetz**[2], **Walter F. Stewart**[2],
Joshua C. Denny[3], **Bradley A. Malin**[3], **Jimeng Sun**[1]
[1]Georgia Institute of Technology, [2]Sutter Health, [3]Vanderbilt University

ABSTRACT

Large amount of Electronic Health Record (EHR) data have been collected over millions of patients over multiple years. The rich longitudinal EHR data documented the collective experiences of physicians including diagnosis, medication prescription and procedures. We argue it is possible now to leverage the EHR data to model how physicians behave, and we call our model *Doctor AI*. Towards this direction of modeling clinical behavior of physicians, we develop a successful application of Recurrent Neural Networks (RNN) to jointly forecast the future disease diagnosis and medication prescription along with their timing. Unlike traditional classification models where a single target is of interest, our model can assess the entire history of patients and make continuous and multilabel predictions based on patients' historical data. We evaluate the performance of the proposed method on a large real-world EHR data over 260K patients over 8 years. We observed Doctor AI can perform differential diagnosis with similar accuracy to physicians. In particular, Doctor AI achieves up to 79% recall@30, significantly higher than several baselines. Moreover, we demonstrate great generalizability of Doctor AI by applying the resulting models on data from a completely different medication institution achieving comparable performance.

1 INTRODUCTION

The broad adoption of Electronic Health Records (EHR) has continuously generated large amount of patient data that documents rich clinical interactions over time. This high-dimensional longitudinal data has created an opportunity to perform sophisticated temporal analysis that was not possible before. Forecasting clinical events for patients is an especially challenging, yet important task. Our goal is to develop a temporal prediction model that mimics physician practice based on the collective memory of many physicians, i.e., large amount of EHR data over a long period of time. Successfully forecasting clinical events can not only facilitate patient-specific care and timely intervention, but also potentially reduce healthcare cost.

Although related problems such as disease progression modeling have been studied by many researchers over several decades, e.g. (Heckerman, 1990; Chapman et al., 2001; Lange et al., 2015), most works do not achieve required accuracy and scalability, or need excessive expert domain knowledge, partly due to the lack of rich longitudinal EHR data and scalable computational architecture. Thanks to the recent advances in recurrent neural network, we propose *Doctor AI* system that can diagnose multiple disease conditions and prescribe relevant medications based on historical EHR data. Furthermore, the Doctor AI tries to predict when the patient will make the next visit. Our ultimate goal is to have Doctor AI help both health providers and patients.

The problem in general can be described as a multilabel marked point process modeling task. The task is different from common sequential learning tasks such as those in natural language processing as it requires prediction of multiple categories over the continuous time axis. The key challenge in this task is to find a flexible model that is capable of predicting multiple event types for patients. The two main classes of techniques, continuous-time Markov chain based models (Nodelman et al., 2002; Lange et al., 2015; Johnson & Willsky, 2013), and intensity based point process modeling techniques such as Hawkes processes (Liniger, 2009; Zhu, 2013; Choi et al., 2015b) have been proposed but they are expensive to generalize to nonlinear and multilabel settings. Furthermore,

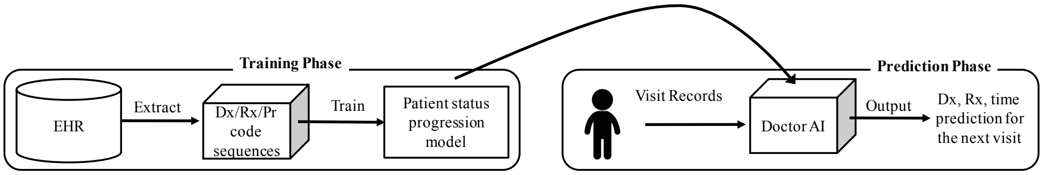

Figure 1: Doctor AI extracts clinical events as multilable point process data from EHR and learns a model for patient status dynamics. Given a new patient's record, it can forecast the patient's diagnoses (Dx), prescribed medications (Rx), and the time until his/her next visit.

they often make strong assumptions about the data generation process which might not be valid in large-scale EHR datasets.

The key idea of this paper is to learn an effective representation of the patient status over time and to leverage such representation to predict future clinical events of the patients such as diagnoses and medication prescriptions and their timings. To learn such patient representations we propose to use recurrent neural networks (RNN), considering the fact that patients have different length of medical records and that recurrent neural networks have been shown to be particularly successful for representation learning in sequential data, *e.g.* (Graves, 2013; Graves & Jaitly, 2014; Sutskever et al., 2014; Kiros et al., 2014; Zaremba & Sutskever, 2014).

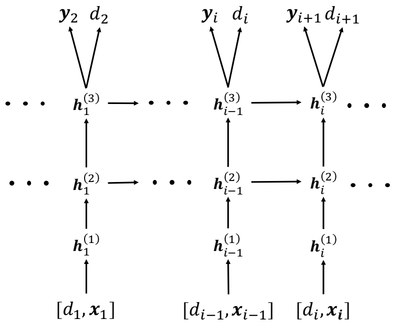

Figure 2: Joint forecast of next visits' time and the codes assigned during each visit. After an embedding layer, the recurrent units (here two layers) learn the status of the patient at each timestamp as a real-valued vector. Given the status vector, we use two dense layers to generate the codes observed in the next timestamp and the duration until next visit. We use Gated Recurrent Units (GRU) (Chung et al., 2014) as it is shown to achieve similar performance to Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997; Graves et al., 2009) while being simpler.

In particular, we make the following main contributions in this paper[1]: (i) We demonstrate a successful application of RNNs in representing the status of patients, predicting the future clinical events and the timing of the events. The trained RNN is able to achieve above 64% (79%) recall in its top 10 (30) predicted diagnosis codes, demonstrating great potential as a computerized differential diagnosis guide. (ii) We propose an efficient initialization scheme for RNNs using Skip-gram embedding (Mikolov et al., 2013) and show that it improves the performance of the RNN in both accuracy and speed. (iii) We empirically confirm that RNNs can be used to transfer knowledge across multiple medical institutions. This suggests that hospitals with insufficient patient records can adopt the models learned from larger data of other health institutions to improve the quality of their clinical service.

2 Experiments

Initializing the RNN with Skip-gram vectors Instead of using a binary vector as an input to the RNN and projecting it to a latent space via the embedding layer, we can use the sum of the representations for each medical code (*e.g.* diagnosis, medication, procedure codes) occurring within a visit and plug it into the RNN directly. The code representations are trained by applying Skip-gram (Mikolov et al., 2013) to a sequence of visits the patient made over time. The learned vectors will capture the hidden relationships among diverse medical codes, thus provide more detailed information to the RNN than a simple binary vector.

Experimental results We use a health records dataset provided by Sutter Health. It includes EHR records for 263,706 patients, where the record for each patient includes the medical codes he has been assigned and the time of the visit. There are 38,594 unique codes in the dataset. Table 1

[1] The long version of the paper can be found at (Choi et al., 2015a).

Table 1: Accuracy of algorithms in forecasting future medical activities. The RNN initialized with Skip-gram vectors is denoted as RNN-IR.

	Dx Only Recall @k			Rx Only Recall @k			Dx,Rx,Time Recall @k			
Algorithms	$k = 10$	$k = 20$	$k = 30$	$k = 10$	$k = 20$	$k = 30$	$k = 10$	$k = 20$	$k = 30$	R^2
Last visit	29.17			13.81			26.25			—
Most freq.	56.63	67.39	71.68	62.99	69.02	70.07	48.11	60.23	66.00	—
Logistic (L=1)	22.97	32.20	36.58	28.01	39.75	43.79	17.66	26.12	31.23	0.0013
MLP (L=1)	26.09	39.19	48.04	32.27	51.12	61.50	19.49	30.80	38.13	0.0017
Logistic (L=5)	26.04	39.17	48.19	32.39	51.06	61.03	18.79	29.13	35.63	0.0013
MLP (L=5)	26.14	39.41	48.28	32.39	51.18	61.66	19.32	30.77	38.08	0.0002
Feature Ext.	26.12	39.33	48.20	32.27	51.12	61.65	19.60	30.80	38.13	0.0022
RNN-1	63.12	73.11	78.49	67.99	79.55	**85.53**	53.86	65.10	71.24	0.2519
RNN-2	63.32	73.32	78.71	67.87	79.47	85.43	53.61	64.93	71.14	0.2528
RNN-1-IR	63.24	73.33	78.73	**68.31**	**79.77**	85.52	54.37	65.68	71.85	0.2492
RNN-2-IR	**64.30**	**74.31**	**79.58**	68.16	79.74	85.48	**54.96**	**66.31**	**72.48**	**0.2534**

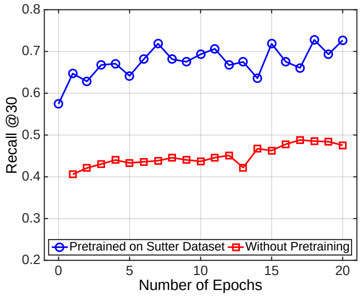

Figure 3: The impact of pre-training on improving the performance on smaller datasets. In the first experiment, we first train the model on a small dataset (red curve). In the second experiment, we pre-train the model on our large dataset and use it for initializing the training of the smaller dataset. This procedure results in more than 20% improvement in the performance.

compares the results of different algorithms with RNN based Doctor AI. We report the results in three settings: when we are interested in (1) only predicting disease codes (Dx), (2) only medication codes (Rx), and (3) jointly predicting Dx, Rx, and time to next visit. The results confirm that the proposed approach is able to outperform the baseline algorithms by a large margin. Note that the recall values for the joint task are lower than those for single Dx or Rx prediction because the hypothesis space is larger for the joint prediction task.

The superior performance of RNN based approaches can be attributed to the efficient representation that they learn for patients at each visit (Bengio et al., 2013; Schmidhuber, 2015). RNNs are able to learn succinct vector representations of patients by accumulating the relevant information from their history and the current set of codes. Comparing RNN-based and most frequent past pattern algorithm with the lagged multilayer perceptron algorithm, we postulate that the status of the patients in this dataset depends on more than 5 lags. This can be because this dataset is collected for study of heart failure which shows long-term dynamics.

Transferring knowledge across hospitals. As we observed from the previous experiments, the dynamics of clinical events are complex, which requires models with a high representative power. However, many institutions have not yet collected large scale datasets, and training such models could easily lead to overfitting. To address this challenge, we resort to the recent advances in domain adaptation techniques for deep neural networks (Mesnil et al., 2012; Bengio, 2012; Yosinski et al., 2014; Hoffman et al., 2014; Paine et al., 2014).

A clinical dataset of 7,653 patients from Vanderbilt University was chosen to conduct the experiment. This dataset differs from the Sutter dataset in that it consists of demographically and diagnostically different patients. The number of unique diagnosis code in this dataset is 1092, which is a subset of Sutter dataset. From the dataset, we extracted sequences of 3-digit ICD-9 codes. We chose 5,000 patients for training, 2,683 for testing. We performed two experiments. First, we trained the model only on the target dataset. Second, we initialized the coefficients of the model with the values learned from Sutter data, then we refined the coefficients with the target dataset. Figure 3 shows the vast improvement of the prediction performance induced by the knowledge transfer from the Sutter data. It is interesting that the model trained on the Sutter dataset without any refinement achieves a significantly higher recall, and further refinements improve the recall up to 15%.

REFERENCES

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. *Unsupervised and Transfer Learning Challenges in Machine Learning*, 7:19, 2012.

Yoshua Bengio, Aaron Courville, and Pierre Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

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

Edward Choi, Nan Du, Robert Chen, Le Song, and Jimeng Sun. Constructing disease network and temporal progression model via context-sensitive hawkes process. In *ICDM*, 2015b.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, pp. 1764–1772, 2014.

Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *PAMI*, 2009.

David Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In *UAI*, 1990.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pp. 3536–3544, 2014.

Matthew J Johnson and Alan S Willsky. Bayesian nonparametric hidden semi-markov models. *The Journal of Machine Learning Research*, 14(1):673–701, 2013.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Jane M Lange, Rebecca A Hubbard, Lurdes YT Inoue, and Vladimir N Minin. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics*, 71(1):90–101, 2015.

Thomas Josef Liniger. *Multivariate hawkes processes*. PhD thesis, Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 18403, 2009, 2009.

Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian J Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, et al. Unsupervised and transfer learning challenge: a deep learning approach. *ICML Unsupervised and Transfer Learning*, 27:97–110, 2012.

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

Uri Nodelman, Christian R Shelton, and Daphne Koller. Continuous time bayesian networks. In *UAI*, pp. 378–387. Morgan Kaufmann Publishers Inc., 2002.

Tom Le Paine, Pooya Khorrami, Wei Han, and Thomas S Huang. An analysis of unsupervised pre-training in light of recent advances. *arXiv preprint arXiv:1412.6597*, 2014.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104–3112, 2014.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.

Wojciech Zaremba and Ilya Sutskever. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014.

Lingjiong Zhu. *Nonlinear Hawkes Processes*. PhD thesis, New York University, 2013.