Joint Modelling of Electronic Health Records and Clinical Notes

Chirag Nagpal
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
chiragn@cs.cmu.edu

SaiKrishna Rallabandi
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
email@domain

Abstract

Significant progress has been made with modelling Electronic Health Records using Deep Learning Methods for Clinical Tasks. In this paper we propose to jointly model a patients medical profile by learning embeddings from Natural Language Text in the form of Physicians Notes and the Patients Historical Diagnoses and Procedures represented by ICD Codes. We further go onto exploit the learnt embeddings on predicting future clinical events.

1 Introduction

Electronic Health Records (EHR) contain a course view of a Patients Medical Profile. Depending on the ustem in use a hospital tends to record various variables including the patients' demographic information, all past histories of Medical Procedures perfomred, and diseases diagnosed. The availability of this longitudanal EHR data has allowed for various Machne Learning and Data Mining techniques being applied succesully for medical tasks revolutioning Medical Informatics.

Deep Neural Models have made significant contributions to Mining of this data, with various tasks being performed, including prediction of Medical Conditons and Events which are encoded as ICD-9 codes in the subsequent admissions, predicting current conditions using Clinical Notes & Prediction of susceptibility to morbidity based on all prior data.

We observe that a significant amount of information is encoded in Patient Notes, including, the doctors impression of any Lab or Radiology Tests performed, any palliative treatments reccommeded if necessary etc. While there has been work to model this data using Neural Models, there has not been much research to glean from the Patient Notes and the ICD-9 Codes jointly in asingle model. We Propose to leverage this knowledge joinlty with the patients past history in order to predict future admissions.

Our contributions in this paper can be summarised as follows

- Learn Embeddings for each patient from the ICD-9 Codes treating the Patients profile as a Language Model.
- Learn Embeddings from the Natural Language Text in Doctors Notes in each Patients Admission.
- Exploit the learnt multimodal embeddings to predict the future admission events jointly, using deep multimodal fusion.

2 Prior Work

Deep Learning has been applied extensively in the past to clinical tasks. Lipton et al. (2015) employed LSTM RNNs (Gers et al., 1999) to model continuous time domain signals like patient vital signs. One of the first such attempts to model EHR data using Recurrent Neural Networks was the Doctor AI System (Choi et al., 2016a). Doctor AI attempted to jointly predict the future ICD events along with time to next admission using Gate Recurrent Units (Graves et al., 2009). Another work of the same author (Choi et al., 2016b) attempts to learn embeddings from the ICD-9 information that includes the Medication, Procedure and Diagnostic Codes for which they employ Skipgrams (Mikolov et al., 2013) along with ReLU activations (Nair and Hinton, 2010).

3 Dataset

We use the MIMIC-III dataset (Johnson et al., 2016), which stands for 'Medical Information

Mart for Intensive Care'. The Dataset consists of vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data of over 38,000 Patients aggregated over corresponding to over 50,000 distinct admissions agggregated over a period of 11 years. Being a one of the larger and publically available dataset, it is the most popular for clinical informatics tasks.

4 Proposed Approach

Acknowledgments

We would like to thank the instructor, Graham Neubig and all other TAs for reviewing the document.

References

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016a. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*. pages 301–318.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016b. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 1495–1504.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm .

Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence* 31(5):855–868.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data* 3.

Zachary C Lipton, David C Kale, and Randall C Wetzell. 2015. Phenotyping of clinical time series with lstm recurrent neural networks. *arXiv preprint arXiv:1510.07641* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. pages 807–814.