Deep Multimodal Fusion of Health Records and Notes for Multitask Clinical Event Prediction

Chirag Nagpal
Auton Lab
Carnegie Mellon
Pittsburgh, PA 15213
chiragn@cs.cmu.edu

Abstract

The advent of Electronic Health Records (EHR) has made it possible to leverage modern machine learning methods that rely on a multitude of data for clinically relevant tasks. However, EHRs contain multiple modalities of data from heterogeneous sources thus posing a challenge when trying to model jointly. Motivated from recent advancements in Deep Multimodal Machine Learning, we propose a Deep Multimodal architecture that jointly models Health Records from the granularity of 1) ICD Diagnostic Codes and 2) Natural Language Text from Clinical Notes. We empirically demonstrate that our approach involving Deep Multimodal Fusion along with Multitask Learning has better performance on predicting closely related clinical events like Cardiac Arrhythmia and Heart Failure.

1 Introduction

Electronic Health Records (EHR) contain a coarse view of the medical profile of a patient. Depending on the system in use, hospitals record various variables including the patients' demographic information, all past histories of medical procedures performed, and diseases diagnosed. The availability of this longitudinal EHR data offers the possibility of deploying several machine learning and data mining techniques for medical data evaluation and prediction thereby revolutionizing medical informatics.

Deep Neural Models have made significant contributions to mining of such data, with various tasks being performed, including prediction of medical conditions 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 notes and reports corresponding to the patients including the doctors impression of any lab or radiology tests performed, any palliative treatments recommended 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 notes and reports of patients alongside the ICD-9 Codes jointly using a single model. We propose to leverage this knowledge jointly with the patients past history in order to predict prevalence of a condition in the current admission. We have released the source code for the experiments at `http://github.com/<anonymous>`

2 Prior Work

Deep Learning has been applied extensively in the past to clinical tasks. [9] employed LSTM RNNs [5] 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 [2]. Doctor

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

AI attempted to jointly predict the future ICD events along with time to next admission using Gate Recurrent Units [6]. Another work of the same author [3] attempts to learn embeddings from the ICD-9 information that includes the Medication, Procedure and Diagnostic Codes for which they employ Skipgrams [10] along with ReLU activations [12]. The Deep Patient System described in [11] did attempt to extract information from the clinical texts, but this was limited to extracting tags corresponding to clinical concepts.

3 Experiments

In this section we describe, the Datasets exploited and feature extraction performes, the clinical tasks and protocol for our experiments.

Dataset: We use the MIMIC-III dataset [7], 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 aggregated 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.

Clinical Tasks: We want to empirically validate if it is possible to predict the Diagnostic Codes, given just data from the patient health records. This has significant clinical impact, certain rare and harder to diagnose diseases have a tendency to be under-reported. Using the available information, We define three predictive clinical tasks for related, cardio-circulatory conditions which are listed below, along with there corresponding ICD-9 codes in Table 1

Table 1: Clinical Tasks

	ICD CODE	CONDITION
TASK-EH	401	Essential Hypertension
TASK-HF	428	Heart Failure
TASK-HA	427	Cardiac Arrhythmia

Experimental Protocol: We proceed to utilize the previously learnt embeddings in order to train a classifier in a supervised fashion, thus for each admission corresponding to a patient, we aim to predict if the patient would be diagnosed with one of the described tasks at the end of there current admission. We perform training on an 80% of the admissions and test on a held out set of 20% of the admissions. We create the splits in a patient independent fashion, such that no single patient lands in both the splits. For any patient admission that has one of the diagnostic codes as in Table 1, we remove this code and label the said admission as a positive.

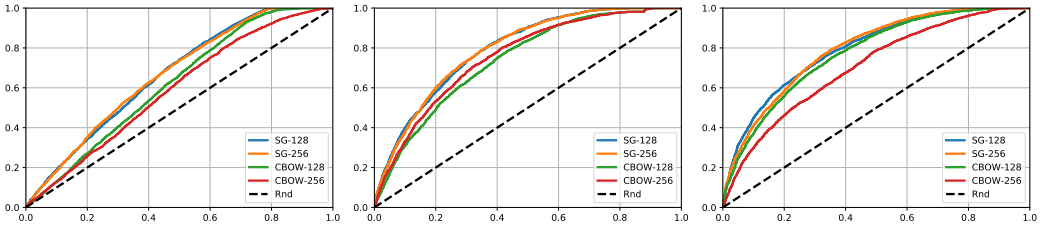


Figure 1: ROC Plots of LR Model trained on Different Learnt Representations

Clinical Report Embeddings: Clinical reports in MIMIC are broadly grouped into 16 categories, and consist mostly of some natural language text recorded by the medical practitioner, that includes multiple medical concepts and some metadata about the patient, including admission units, serial numbers, name of Caregivers involved, Dates.

We first build regular expressions to strip all the metadata from the clinical reports in order to learn embeddings. We then proceed to utilise, Continuous Bag of Words (CBOW) and Skipgram (SG) models to learn the per word embeddings. We observed, and corroborated from previous research [1]

that since these reports were hand generated, there were significant instances of misspellings, and thus the use of Character Level, or Subword Level Embeddings may perform better, and be robust to these idiosyncrasies of the data.

For both CBoW & SG we used a context window of 5 and extracted embeddings of dimensionality 128 & 256. Since each admission has multiple reports that correspond to 16 different tests[1], we aggregate the word vector representations for each individual test by averaging over them. These averaged vectors for each test are then concatenated together to represent the patients stay in a continuous space.

Table 2 & Figure 1 represents the performance of a Logistic Regression Classifier trained on the various different embeddings we extract in terms of Area Under the Receiver Operator Curve (AU-ROC). We observe that Skipgrams with a dimensionality of 256 outperformed all other representations, although the improvement was only of a small margin.

Table 2: AU-ROC for Various Embeddings

EMBEDDING	TASK-EH	TASK-HF	TASK-HA
CBOW-128	0.6673	0.7885	0.7867
CBOW-256	0.6685	0.7856	0.7871
SG-128	0.6756	0.7970	0.7991
SG-256	**0.6803**	**0.7999**	**0.8006**

EHR Features: For each of the Patient admissions, we also extract the Diagnostic and Procedures codes which follow the ICD-9 convention. When featurising, we truncate the ICD-9 codes to there Top Level 3 Digits, and then proceed to represent this explicitly as One Hot Encoded feature representations.

4 Baselines

We Compare our models against the following baselines:

Logistic Regression With $\ell 2$ Regularisation : A simple baseline, which is useful as a diagnostic tool for determining the hardness of the Learning Task. We apply an $\ell 2$ penalty on the weights vector with α, the regularisation parameter set to 10^{-4}

Random Forest Ensemble: We use a Random Forest Estimator with 100 trees, with Gini Index as the criteria for splitting.

Multilayer Perceptron (MLP): We employ a Feedforward MLP with two hidden layers of Equal Dimensionality with Sigmoid Activations. The final layer is trained to minimize Cross Entropy loss between the output and the true labels. We train a separate MLP for each task.

5 Proposed Models

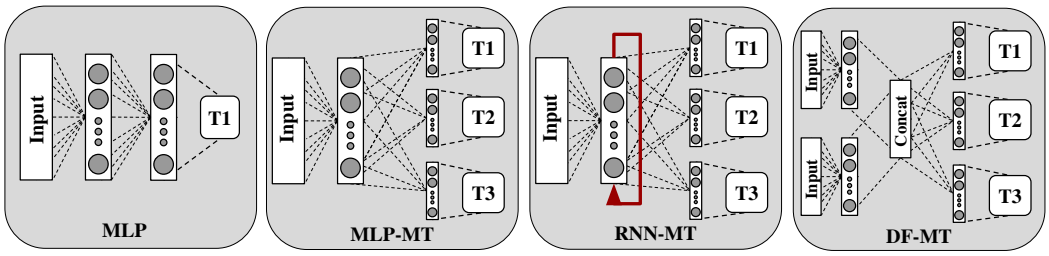


Figure 2: Proposed Models

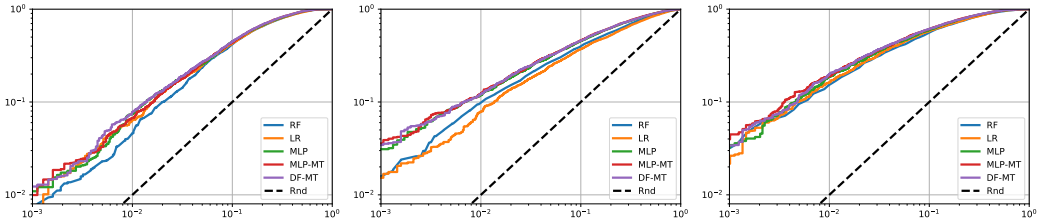
[1]Note that this includes discharge summaries, that encapsulate terms suggestive of the diagnoses at discharge. When evaluating the Proposed Models in the next section we ignore the Discharge Summaries.

Multitask Multilayer Perceptron (MLP-MT): The MLP-MT consists of a single hidden layer with sigmoid activations, followed by a hidden layer each for each task. Thus, these final layers act as a softmax for each task. As compared to MLP, MLP-MT reduces the number of parameters to learn since, it shares intermediate representation amongst each task. We hypothesize that Multitask Learning would allow model to jointly leverage knowledge amongst tasks, which would be useful in scenarios where the target classes are related.

Recurrent Multitask Network (RNN-MT) : The RNN-MT improves over MLP-MT by replacing the intermediate hidden layer with a Recurrent Unit. The hypothesis being that the recurrent unit will be able to better model patients with multiple admissions by treating it as a time series, where each time step is represented by the feature vector in that admission.

Late Fusion Multitask Network (DF-MT) : As opposed to all other models, that concatenate feature representations corresponding to the modalities at the input, we perform Late fusion, which first applies a Linear transformation with Sigmoid activation, individually to each modality and then concatenate the resulting output before performing another Non-Linear Transformation with a Softmax for each task.

We train all our Neural Models in PyTorch for 100 Epochs on the Training Dataset using the Adam Optimizer [8] with a learning rate set to 10^{-4}



TASK-EH TASK-HF TASK-HA

Figure 3: Log-Log ROC Plots of the Proposed Models

Table 3: AU-ROC for the Proposed Models

MODEL	TASK-EH	TASK-HF	TASK-HA
LR	0.8158	0.7593	0.8502
RF	0.8176	0.7405	0.8553
MLP	0.8249	0.7927	0.8688
MLP-MT	0.8281	0.7979	0.8731
RNN-MT	0.7911	0.7751	0.8519
DF-MT	**0.8296**	**0.7957**	**0.8732**

6 Results

Figure 3 & Table 3 presents the ROC Plots and AUC Scores of the proposed. We observed that Deep Multimodal Fusion (DF-MT) outperformed the other approaches in AUC-ROC scores, although on visual inspection from ROC plots, we observe that in the Low FPR range, there was not much difference in performance between the DF-MT and MLP-MT.

Although, for temporal even modeling, recurrent models perform well, in our case, the use of the recurrent intermediate layer (RNN-MT) significantly reduced performance. We also experimented with LSTM and GRU [4] hidden units, however, did not see any improvement. We attribute this to the fact that most of the Patients in the MIMIC Dataset have just single admissions. This, combined with the fact that Recurrent Units require greater number of parameters to be learnt, which requires greater amount of training data.

7 Conclusion

In this paper, we experiment with better approaches to model Multimodal data in Electronic Health Records for clinical tasks. We observe Late Fusion techniques outperform simple feature concatenation when used to model multiple related output classes, in a Multitask Setting.

References

[1] Sandeep Ayyar. Tagging patient notes with icd-9 codes. 2017.

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

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

[4] 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.

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

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

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

[8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

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

[11] Riccardo Miotto, Li Li, Brian A 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, 2016.

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