Deep Learning Models for Health Care: Challenges and Solutions

Yan Liu [1] Jimeng Sun [2]

[1]Computer Science Department
Viterbi School of Engineering
University of Southern California

[2]School of Computational Science and Engineering
College of Computing
Georgia Institute of Technology

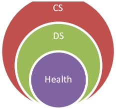
August 5, 2017

Tutorial Slides and Supplementary Materials

https://tinyurl.com/y7wuk9xt

Why healthcare?

- Healthcare is big
- Healthcare is bad
- Healthcare is challenging

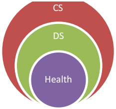


US healthcare: The COST problem

Overall spending: 3.8 trillion dollars (2014)

>

Top 10 most valuable companies combined



US healthcare: The COST problem

Overall spending: 3.8 trillion dollars (2014)

>

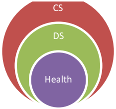
Top 10 most valuable companies combined

EXXON Google BERKSHIRE HATHAWAY Johnson & Johnson WELLS FARGO Roche Walmart

+

10× Beijing Olympics

Beijing 2008



US healthcare: The COST problem

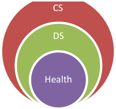
Overall spending: 3.8 trillion dollars (2014)

>

Top 10 most valuable companies combined

+

10× Beijing Olympics



US healthcare: The COST problem

Overall spending: 3.8 trillion dollars (2014)

>

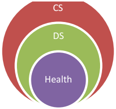
Top 10 most valuable companies combined

+

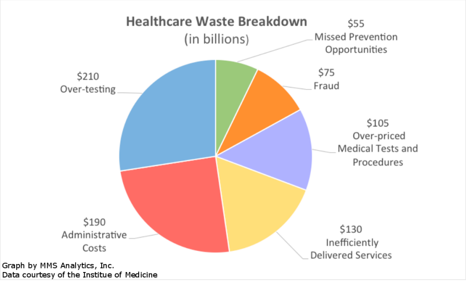
10× Beijing Olympics

+

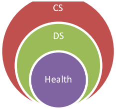
Net worth of



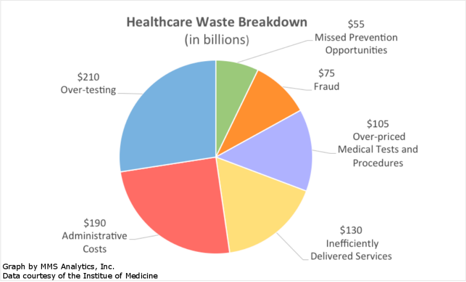
US Healthcare Waste per year



$765 billions



US Healthcare Waste per year



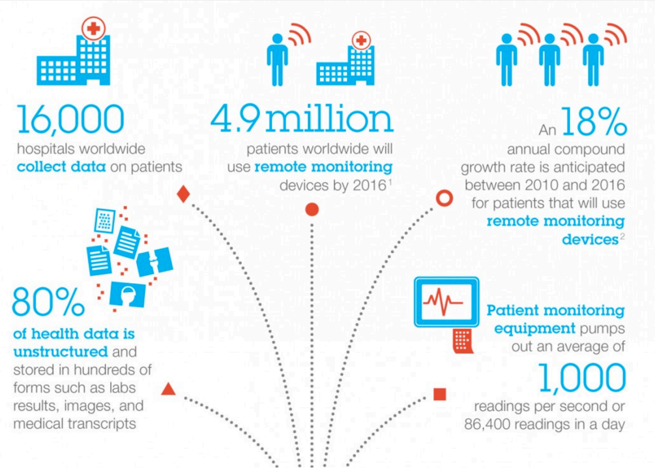
$765 billions = 50 years budget

US Healthcare Quality Issue

- 200K to 400K preventable death per year
  - Over 1000 per day

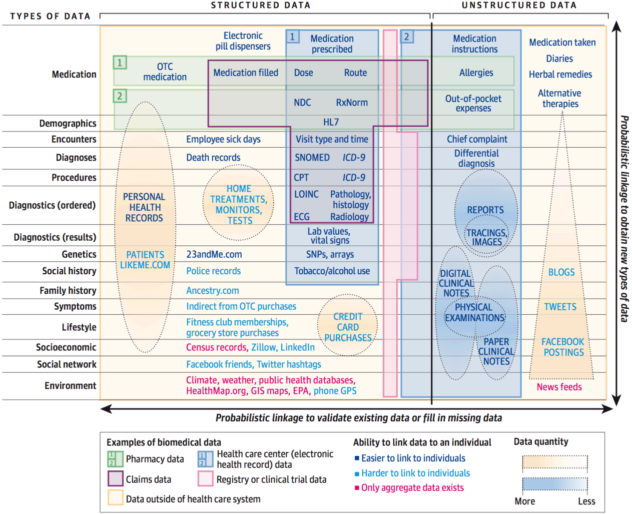
https://www.documentcloud.org/documents/781687-john-james-a-new-evidence-based-estimate-of.html#document/p1/a117333

Healthcare data is everywhere



Source from http://www.okilab.es/how-big-data-is-changing-healthcare/

Healthcare data sources

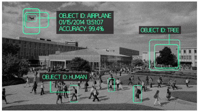


Weber, Griffin M., Kenneth D. Mandl, and Isaac S. Kohane. 2014. "Finding the Missing Link for Big Biomedical Data." *JAMA: The Journal of the American Medical Association* 311 (24): 2479–80.

Why deep learning?

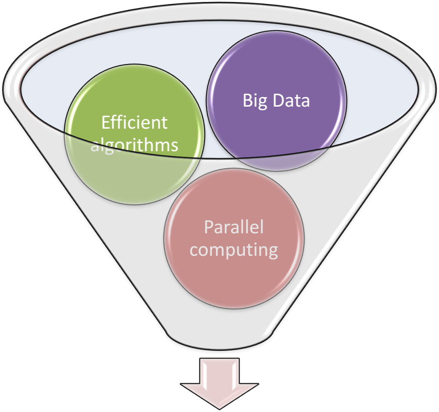
- *Speech recognition*

- *Computer vision*
  - *Image Classification*
  - *Video analysis*



- *Natural language processing*
  - *Machine translation*

Recipes for deep learning success

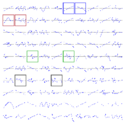


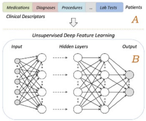
Successful deep learning models

Early Work on Deep Learning in Health Applications

Stacked Auto-encoder (SDA)

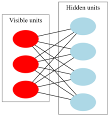
Computational phenotyping [Lasko et al., 2013; Kale et al., 2014; Che et al., 2015; Kale et al., 2015; Miotto et al., 2016]

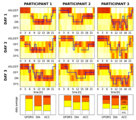




Deep neural networks (DNNs)

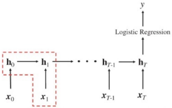
Restricted Boltzmann machine (RBM)
Multi-layer perceptron (MLP)
Condition prediction [Dabek and Caban, 2015; Hammerla et al., 2015]





Recurrent neural networks (RNNs)

Long short-term memory (LSTM) Gated recurrent unit (GRU)
Diagnosis/event prediction Lipton et al. [2015]; Choi et al. [2016]



Outline

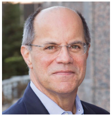
1. Lecture 1: Data Sources and Health Care Problems
   - EHR and Claims Data
   - Medical Imaging Data
   - Continuous Time Series (EEG, ECG, ICU monitoring)
   - Clinical Notes
2. Lecture 2: Challenges and Solutions of DL for Health Care
3. Future Directions

Doctor AI: Predicting Clinical Events via Recurrent Neural Networks

Edward Choi

Taha Bahadori

Andy Schuetz



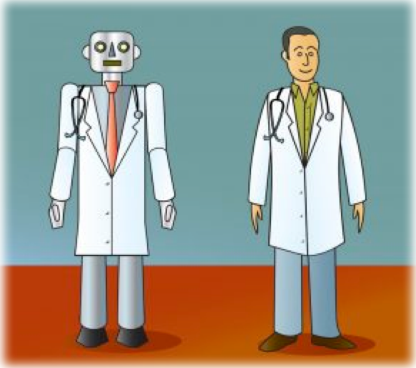
Buzz Stewart

Georgia Tech College of Computing — Computational Science and Engineering

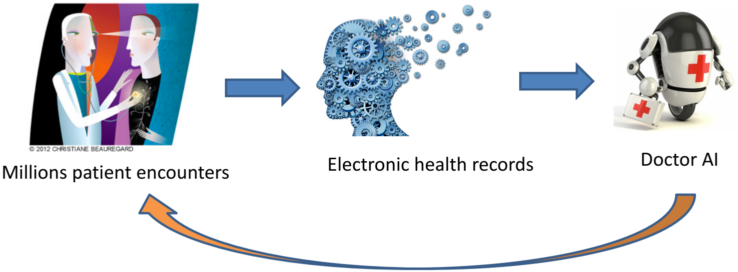
Sutter Health

Choi, Edward, et al. 2016. "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks." In *Machine Learning for Healthcare Conference*, 301–18.

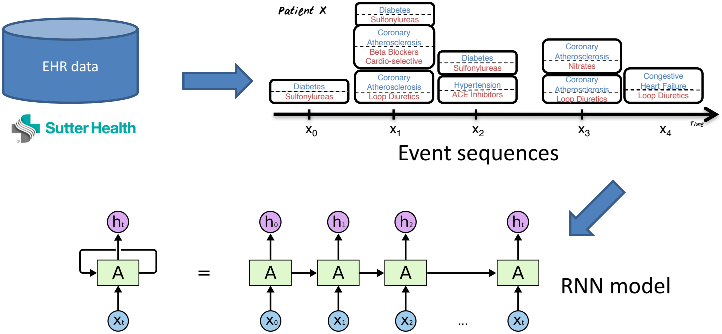
Do you want to be seen by a machine or a human for medical care?



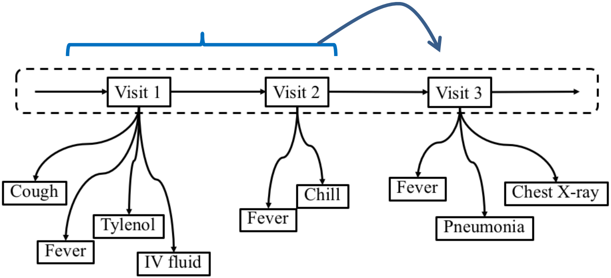
Can machine perform similarly as doctors in diagnosis?



Approach: Recurrent Neural Network (RNN)

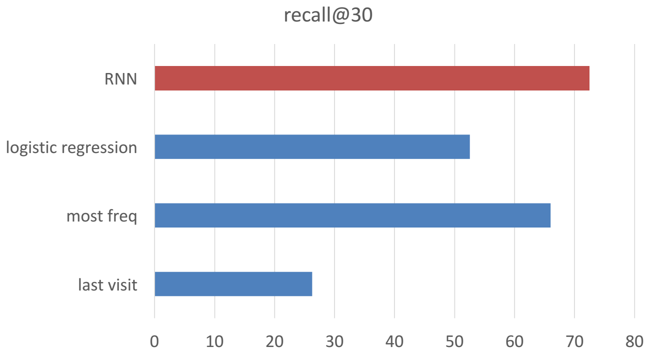


Disease Progression Modeling

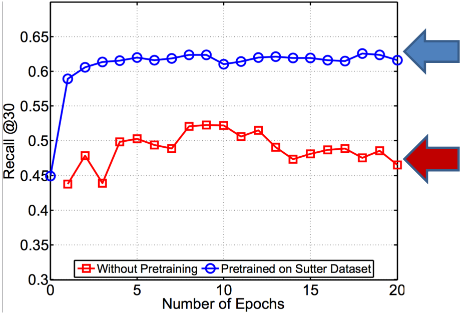


Accuracy: $\text{top-}k\text{ recall} = \dfrac{\#\text{ of true positives in the top } k \text{ predictions}}{\#\text{ of true positives}}$

RNN on predicting diagnoses in next visit



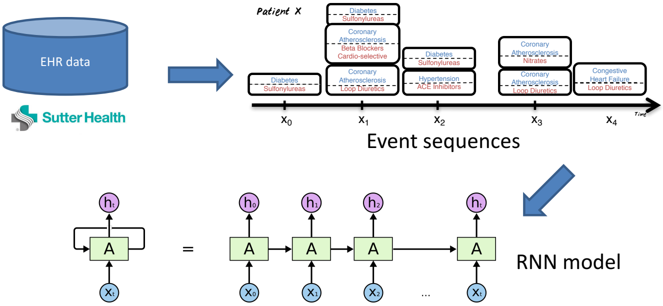
Generalize RNN model from one institution to another



Perform better when warm start the existing model

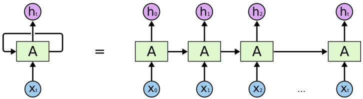
Perform worse when cold-start from random initialization

Summary: Doctor AI



- *general & accurate model for many prediction tasks*
- *Can handle sequences of variable lengths*

Choi, Edward, et al. 2016. “Doctor AI: Predicting Clinical Events via Recurrent Neural Networks.” In *Machine Learning for Healthcare Conference*, 301–18.
https://github.com/mp2893/doctorai

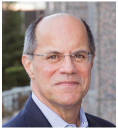


USING RECURRENT NEURAL NETWORK MODELS FOR EARLY DETECTION OF HEART FAILURE ONSET

How to model temporal relations in the EHR data


Edward Choi


Andy Schuetz


Buzz Stewart

Georgia Tech College of Computing Computational Science and Engineering

Sutter Health

Edward Choi, Andy Schuetz, Walter Stewart, Jimeng Sun. Using Recursive Neural Network Models for Early Detection of Heart Failure Onset, JAMIA 2016

MOTIVATIONS FOR EARLY DETECTION OF HEART FAILURE

Heart failure is a complex disease.

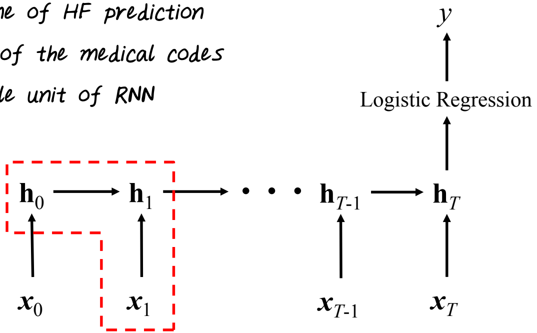
Reduces cost and hospitalization.

Improves existing clinical guidelines of HF prevention.

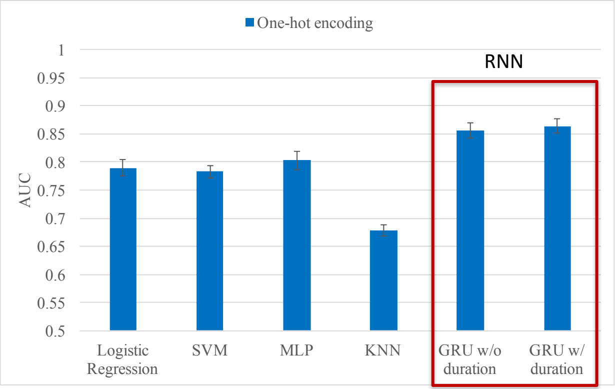
Early intervention can slow down disease progression.

Temporal model: RNN

- x_t: one-hot coded Dx, Rx, Proc at time t
- h_t: hidden state at time t
- y: binary outcome of HF prediction
- T: total length of the medical codes
- Red box: a single unit of RNN

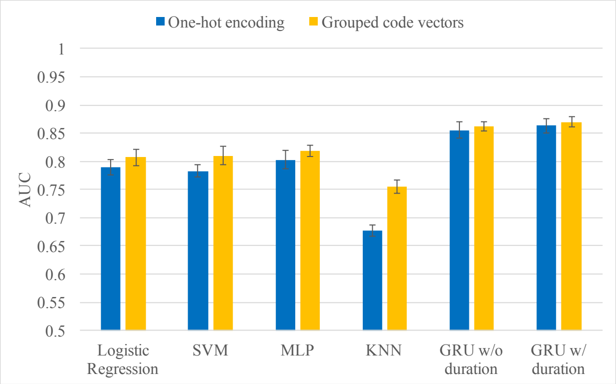


PREDICTION PERFORMANCE OF RNN



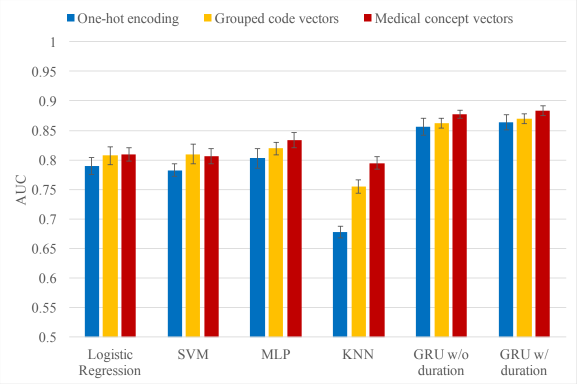
- ***RNN model achieves over 10% improvement on AUC***

PREDICTION PERFORMANCE OF RNN



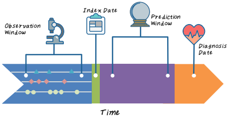
- RNN model achieves over 10% improvement on AUC
- ***Representation matters***

PREDICTION PERFORMANCE OF RNN

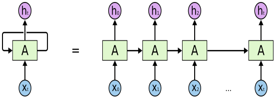


- RNN model achieves over 10% improvement on AUC
- ***Data rep. (word2vec) > knowledge rep. (medical groupers)***

Summary: Recurrent Neural Network (RNN) for heart failure onset prediction



Heart failure onset can be predicted using EHR data



Temporal information matters for HF onset prediction

Data driven representation matters

Georgia Tech College of Computing
Computational Science and Engineering

Sutter Health

Edward Choi, Andy Schuetz, Walter Stewart, Jimeng Sun. Using Recursive Neural Network Models for Early Detection of Heart Failure Onset, JAMIA 2016

DEEP LEARNING SOLUTIONS FOR CLASSIFYING PATIENTS ON OPIOID USE

Zhengping Che

Jennifer St. Sauver

Hongfang Liu

USC

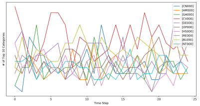
MAYO CLINIC

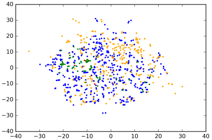
Che et al, Deep Learning Solutions for Classifying Patients on Opioid UseZhengping Che, Jennifer St. Sauver, Hongfang Liu, and Yan Liu. American Medical Informatics Assocation Annual Symposium (AMIA), 2017

Deep Learning for Opioid Use Analysis

Opioid use study on datasets from the Rochester Epidemiology Project (REP)[1] with more than 140k people

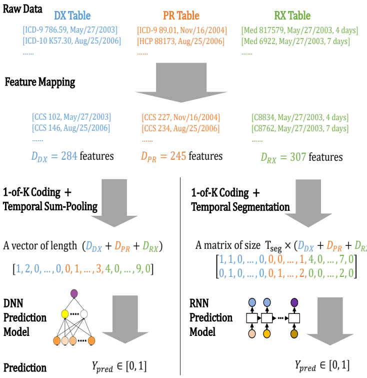
- To extract and understand risk factors and indicators for adverse opioid and opioid-related events
- To predict new opioid users and dependence and recognize misuse on opioid analgesics
- To provide health care providers with better suggestions on pain medication prescriptions





Our Framework

- Cohort selection and group identification
  - More than 110 millions of medical records in 2013-2016 are used
  - Patients are grouped into *short-term*, *long-term*, and *opioid-dependent* users
- Temporal feature processing
  - Records of *diagnoses*, *procedures*, and *prescriptions* are mapped into different coding systems via one-hot encoding
  - Sum-pooling and segmentation along the temporal dimension is applied to build the input matrix for each patient
- Multilayer DNNs and LSTMs with ReLU function are used for prediction.



Empirical Evaluations

Deep learning models outperforms other baselines with similar model size

- Classification comparisons on AUC score (*auc*) and kappa coefficient (κ)

	Short-term / Long-term					Long-term / Opioid-dependent				
	LR	SVM	RF	DNN	RNN	LR	SVM	RF	DNN	RNN
auc	0.7323	0.7327	0.6936	0.7340	**0.7536**	0.6512	0.6429	0.6999	**0.7279**	0.7144
κ	0.1090	0.0885	0.1289	0.0756 ± 0.004	**0.2076**	0.1906	0.1821	0.2342	**0.3006**	0.2542

Most important features are selected by DNN models

- Feature importance $\mathcal{I}$ are calculated from weights in all the layers in DNN

$$\mathcal{I} = \mathbf{W}^{[L]} BN^{[L]} \left(\cdots \mathbf{W}^{[2]} BN^{[2]} \left(\mathbf{W}^{[1]} BN^{[1]} (\mathbf{1}) \right) \right) \in R^{1 \times D}$$

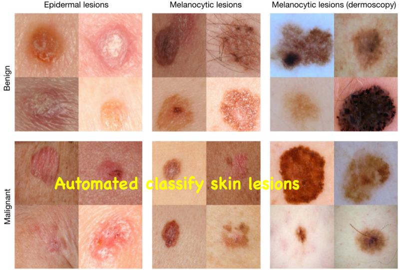
- Top related feature categories and their corresponding scores

Short-term / Long-term				Long-term / Opioid-dependent			
Table	Code	Feature Name	$\mathcal{I}$	Table	Code	Feature Name	$\mathcal{I}$
RX	C8834	Opioid Analgesics	0.2287	RX	C8834	Opioid Analgesics	0.7784
RX	C8890	Amphetamine-like Stimulants	−0.0843	DX	CCS 661	Substance-related Disorders	0.6186
RX	C8838	Non-opioid Analgesics	0.0802	PR	CCS 182	Mammography	−0.3481

Outline

Unstructured Imaging Data: Data

Stanford ENGINEERING
Computer Science



Research at Google

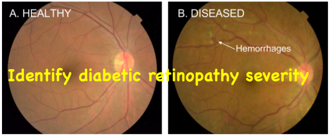
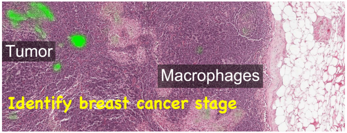


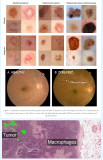
Figure 1. Examples of retinal fundus photographs that are taken to screen for DR. The image on the left is of a healthy retina (A), whereas the image on the right is a retina with referable diabetic retinopathy (B) due a number of hemorrhages (red spots) present.



1. *Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA.2016;316(22):2402-2410*
2. *A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau , S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017)*
3. *Liu, Yun, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, et al. 2017. "Detecting Cancer Metastases on Gigapixel Pathology Images." arXiv [cs.CV]. arXiv. http://arxiv.org/abs/1703.02442.*

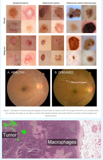
Unstructured Imaging Data: Task

Data



→ Training → Model

New Data



→ Apply Model → Prediction results

Training

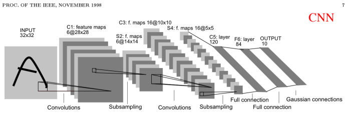
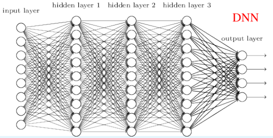


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.



DNN for detecting diabetic retinopathy

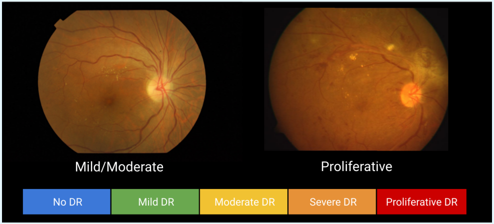


Image based detection of diabetic retinopathy (JAMA 2016)

- Train deep neural networks to find diabetic retinopathy severity from the intensities of the pixels in a fundus image.
- Received testing AUC of 0.991 on EyePACS-1 data, and testing AUC of 0.990 on Messidor-2 data.

1. *Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA.2016;316(22):2402-2410*

Unstructured Imaging Data: skin lesion (Nature 2017)

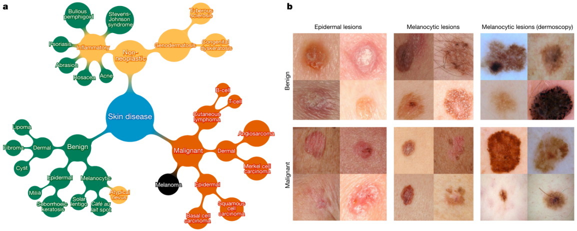
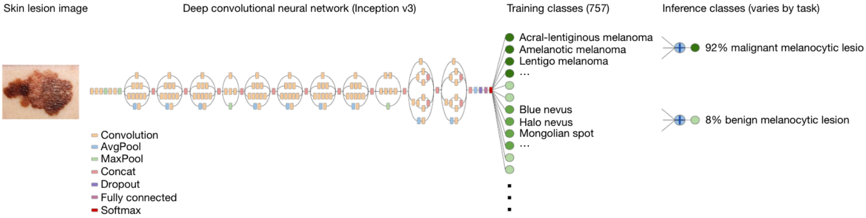


Figure 2 | A schematic illustration of the taxonomy and example test set images. a, A subset of the top of the tree-structured taxonomy of skin disease. The full taxonomy contains 2,032 diseases and is organized based on visual and clinical similarity of diseases. Red indicates malignant, green indicates benign, and orange indicates conditions that can be either. Black indicates melanoma. The first two levels of the taxonomy are used in validation. Testing is restricted to the tasks of **b**. **b**, Malignant and benign example images from two disease classes. These test images highlight the difficulty of malignant versus benign discernment for the three medically critical classification tasks we consider: epidermal lesions, melanocytic lesions and melanocytic lesions visualized with a dermoscope. Example images reprinted with permission from the Edinburgh Dermofit Library (https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html).

- Deep convolutional neural networks to perform binary classification for two use cases:
  - keratinocyte carcinomas versus benign seborrheic keratosis; and
  - malignant melanomas versus benign nevi.
- Achieved better-than-human expert accuracy (0.7210 vs. 0.6556)

1. *A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau , S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017)*

Unstructured Imaging Data: skin lesion (Nature 2017)



- Deep convolutional neural networks to perform binary classification for two use cases:
  - keratinocyte carcinomas versus benign seborrheic keratosis; and
  - malignant melanomas versus benign nevi.
- Achieved better-than-human expert accuracy (0.7210 vs. 0.6556)

1. *A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau , S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017)*

Unstructured Imaging Data: skin lesion (Nature 2017)

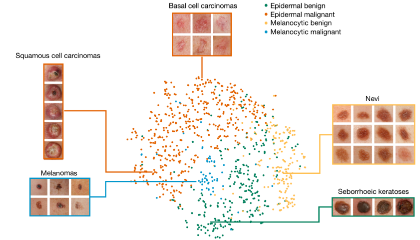


Figure 4 | t-SNE visualization of the last hidden layer representations in the CNN for four disease classes. Here we show the CNN's internal representation of four important disease classes by applying t-SNE, a method for visualizing high-dimensional data, to the last hidden layer representation in the CNN of the biopsy-proven photographic test sets (932 images). Coloured point clouds represent the different disease categories, showing how the algorithm clusters the diseases. Insets show images corresponding to various points. Images reprinted with permission from the Edinburgh Dermofit Library (https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html).

- Deep convolutional neural networks to perform binary classification for two use cases:
  - keratinocyte carcinomas versus benign seborrheic keratosis; and
  - malignant melanomas versus benign nevi.
- Achieved better-than-human expert accuracy (0.7210 vs. 0.6556)

1. *A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau , S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017)*

Outline

SLEEPNET: Automated Sleep Medicine via Deep Learning

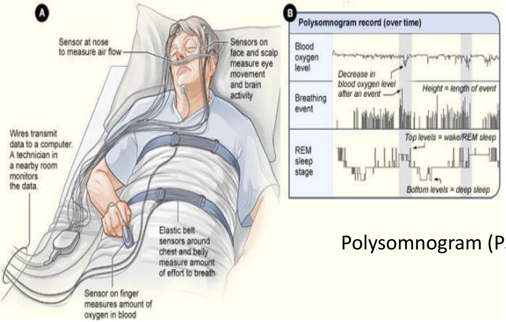
Siddharth Biswal, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Matt T Bianchi, Jimeng Sun

Georgia Tech

MGH 1811 MASSACHUSETTS GENERAL HOSPITAL

https://arxiv.org/abs/1707.08262

Motivation: Automated sleep medicine

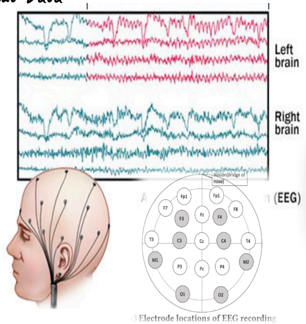


Polysomnogram (PSG) recording

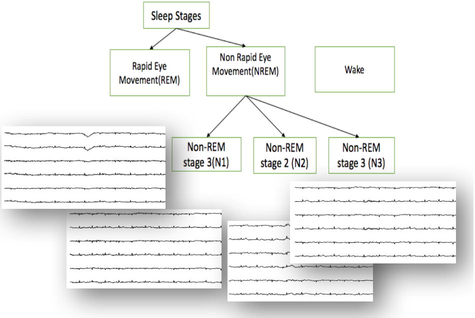
- ~50-70million people in US currently suffer sleep disorders
- Central diagnostic tool is the overnight sleep study, Polysomnogram (PSG)
- Labor intensive effort to annotating PSG
  - Automation of these could alleviate these concerns

Sleep staging

Input Data

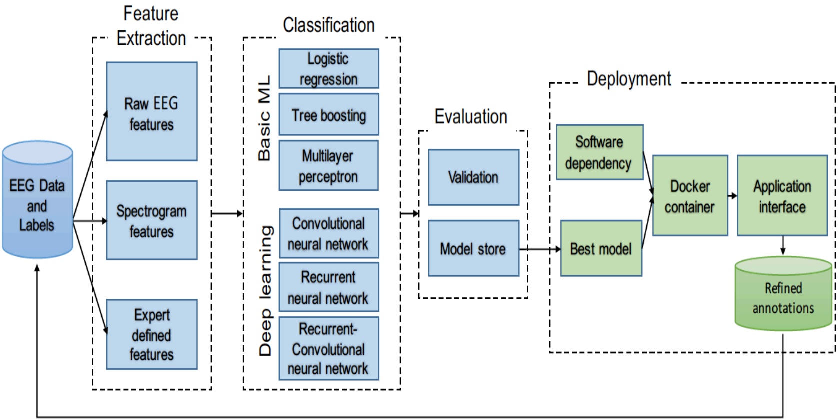


Output



- *EEG data in PSG consists of data from 6 different channels*
- *Every 30 second of EEG were annotated into one of 5 stages*
  - *Sleep stages are important for many sleep quality metrics*
- *Annotation is nontrivial even for experienced technologists*
  - *Inter rater agreement rate about 70%*

Analytic pipeline of SLEEPNET



(d) Analytic pipeline of SLEEPNET. The blue color components correspond to model training module. The green color components belongs to the model deployment module.

39

Dataset Description

Dataset Property	Number
Number of Patients	10,000
Hours of EEG data	80,000
Raw data storage	3.2 TB
Number of labeled samples	>9 million

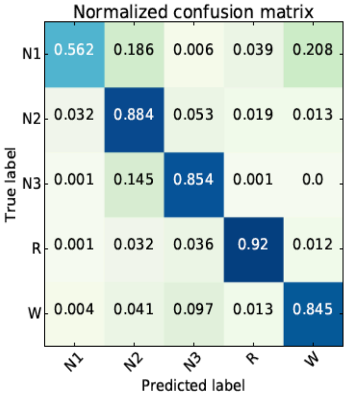
Results

	Expert Defined Features		Spectrogram Features		Waveform Features	
Model	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
LR	68.54	63.88	66.54	66.61	67.43	62.71
TB	75.67	69.47	71.61	65.37	72.36	66.37
MLP	72.23	68.41	70.23	66.71	69.56	64.21
CNN	79.45	72.63	77.83	71.45	77.31	71.47
RNN	**85.76**	**79.46**	79.21	73.83	79.46	72.46
RCNN	81.67	76.38	81.47	74.37	79.81	73.52

Table 4: Performance of different feature representations with model combinations

RNN + expert defined features perform the best

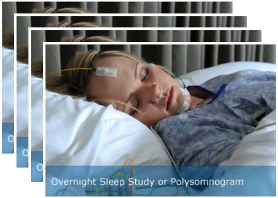
Algorithm achieves expert-level performance (avg. accuracy > 85%)



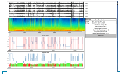
Confusion matrix for the best performing model (RNN+expert)

Summary: SLEEPNET

Automated Sleep staging based on deep learning



Large dataset of 10,000 polysomnogram studies



MGH 1811

Deployed for research

https://arxiv.org/abs/1707.08262

43

INTERPRETABLE DEEP MODELS FOR ICU OUTCOME PREDICTION

Zhengping Che **Sanjay Purushotham** **Robinder Khemani**

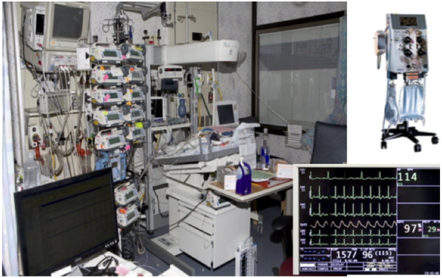
USC Children's Hospital LOS ANGELES

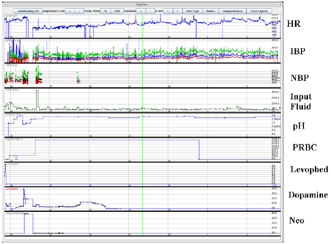
Che et al, Interpretable Deep Models for ICU Outcome Prediction. of the American Medical Informatics Association Annual Symposium (AMIA), 2016.

Time Series in Critical Care Unit (ICU)

Critical care is among the most important areas of medicine.

- >5 million patients admitted to US ICUs annually.[2]
- Cost: \$81.7 billion in US in 2005: 13.4% hospital costs, ~1% GDP.[1]
- Mortality rates up to 30%, depending on condition, care, age.[1]
- Long-term impact: physical impairment, pain, depression.





Society of Critical Care Medicine website, Statistics page.

Datasets and Tasks

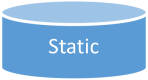
Children's Hospital Los Angeles (CHLA)
398 patients stay > 3 days
Static features (age, weight, etc.): 27 variables
Temporal features (Blood gas, ventilator signals,injury markers, etc.): 21 variables

MIMIC III Dataset
19714 patients stay for 2 days
All temporal features (input fluids, output fluids, lab tests, prescription): 99 variables

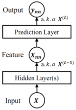
PhysioNet Challenge Part of MIMIC II dataset

Task Prediction task (mortality, ventilator free days, and disease code), computational phenotyping, anomaly detection, disease subtyping

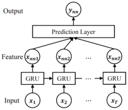
Deep learning model: DNN + GRU



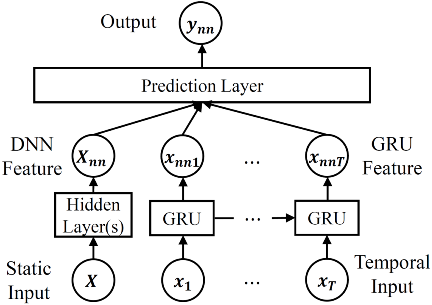
- *Static + (flattened) temporal features*
  - *DNN*



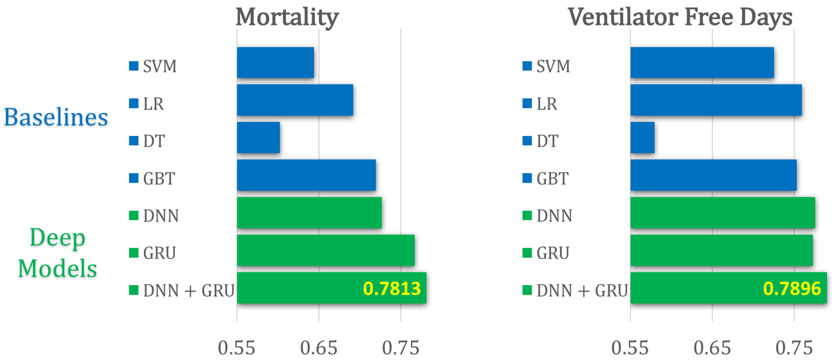
- *Temporal features only*
  - *GRU*



- Static + temporal features
  - **DNN + GRU** (combination)



Experiment Results



SVM: support vector machine; LR: logistic regression;
DT: decision tree; GBT: gradient boosting tree.
Results are based on 5-fold cross-validation.

Outline

1. Lecture 1: Data Sources and Health Care Problems
   - EHR and Claims Data
   - Medical Imaging Data
   - Continuous Time Series (EEG, ECG, ICU monitoring)
   - Clinical Notes
2. Lecture 2: Challenges and Solutions of DL for Health Care
3. Future Directions

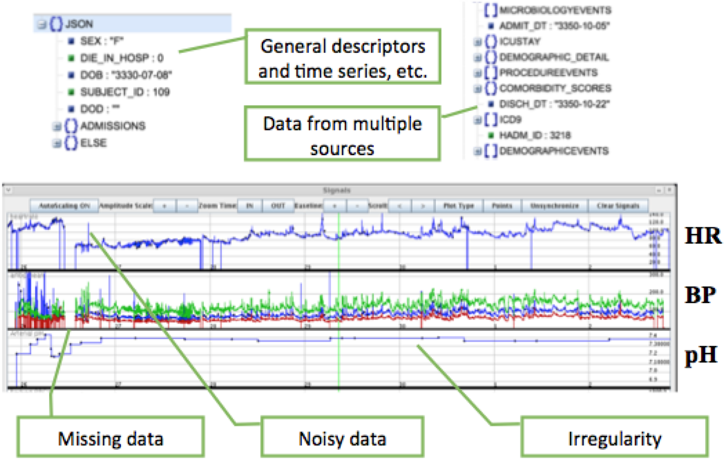
Deep Neural Networks for Analyzing Clinical Notes

Examples of some recent development:

- LSTM for i2b2/VA relation classification challenge [Luo, 2017]
- Convolutional neural networks for medical text classification [Hughes et al., 2017]
- Bidirectional RNN for medical event detection [Jagannatha and Yu, 2016]
- RNN with attention for adverse drug reaction [Pandey et al., 2017]
- Condensed memory networks for clinical diagnostic Inferencing [Prakash et al., 2016]
- Neural attention models for classification of radiology reports [Shin et al., 2017]

Outline

Example of Health Care Data

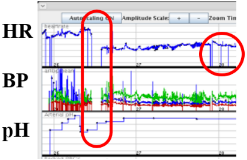


Machine learning challenges for health applications

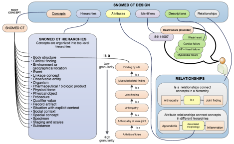
Small sample size

- Rare diseases
- Small clinics

Missing value

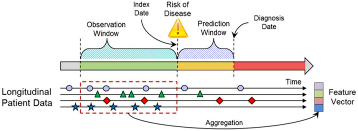


Medical domain knowledge



- Medical ontology

Interpretation



- Explain the prediction

Outline

VARIATIONAL RECURRENT ADVERSARIAL DEEP DOMAIN ADAPTATION



Sanjay Purushotham **Wilka Carvalho** **Tanachat Nilanon**

USC

Purushotham et al, Variational Recurrent Adversarial Deep Domain Adaptation. International Conference on Learning Representations (ICLR 2017)

88

Motivation - Big Small Data

Limited amount of data available to train age-specific or disease-specific models

- A toy example: predicting mortality across adults and children in ICU
-

Target	Model Trained on Adults	Model trained on Children
Children	0.56	0.70

- Training models for each age group independently is not ideal due to limited amount of data

Question: How do we adapt models from Adults (source domain) to Children (target domain)?

Problem Formulation

Case study: mortality prediction for patients across different age groups

- Input: N multivariate time series example: $x^i = (x^i_t)_{t=1}^{T^i}$
- Source domain (e.g. adult): $\{x^i, y_i\}_{i=1}^{n}$, target domain (e.g., child): $\{x^j\}_{j=n+1}^{N}$
- Output: mapping function $f^{target}(x^i) \approx y_i$

Problem definition: unsupervised domain adaptation for multivariate time series

Related Work

Domain adaption for non-time series data

- Domain discrepancy reduction [Ben-David et al., 2007]
- Instance re-weighting [Jiang and Zhai, 2007]
- Subspace alignment [Fernando et al., 2013]
- Deep learning approaches [Ganin and Lempitsky, 2014; Tzeng et al., 2015], domain adversarial neural networks (DANN) [Ganin et al., 2016]

Domain adaption for sequence or time series data

- Dynamic Bayes networks [Huang and Yates, 2009]
- Recurrent neural networks [Socher et al., 2011]

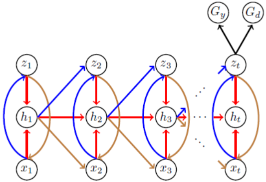
Our solution:

Deep learning model with adversarial training and variational methods for domain invariant representation while transferring temporal dependencies

Variational Adversarial Deep Domain Adaptation (VADDA) [ICLR 2017]

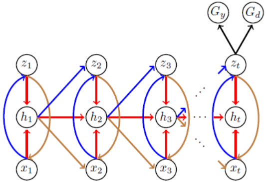
VRNN objective function [Chung et al, 2016]

$$\mathcal{L}_r(x_t^i;\theta_e,\theta_g) = E_{q_{\theta_e}(z^i_{\leq T^i}|x^i_{\leq T^i})}\sum_{t=1}^{T^i}(-D(q_{\theta_e}(z_t^i|x^i_{\leq t},z^i_{<t})||p(z_t^i|x^i_{<t},z^i_{<t}))+\log p_{\theta_g}(x_t^i|z^i_{\leq t},x^i_{<t}))$$



Inference: $z_t^i \sim q(z_t^i|x^i_{\leq t}, z^i_{<t})$
Generation: $x_t^i \sim p(x_t^i|z^i_{\leq t}, x^i_{<t})$
Recurrence: $h_t = RNN(x_t, z_t, h_{t-1})$

Variational Adversarial Deep Domain Adaptation (VADDA) [ICLR 2017]



Inference: $z_t^i \sim q(z_t^i | x_{\le t}^i, z_{<t}^i)$
Generation: $x_t^i \sim p(x_t^i | z_{\le t}^i, x_{<t}^i)$
Recurrence: $h_t = RNN(x_t, z_t, h_{t-1})$

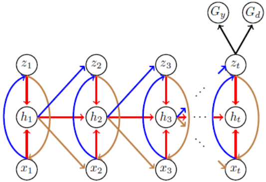
VRNN objective function [Chung et al, 2016]

$$\mathcal{L}_r(x_t^i; \theta_e, \theta_g) = E_{q_{\theta_e}(z^i_{\le T^i} | x^i_{\le T^i})} \sum_{t=1}^{T^i} (-D(q_{\theta_e}(z_t^i | x_{\le t}^i, z_{<t}^i) || p(z_t^i | x_{<t}^i, z_{<t}^i)) + \log p_{\theta_g}(x_t^i | z_{\le t}^i, x_{<t}^i))$$

Source classification loss with regularizer

$$\min_{\theta_e, \theta_g, \theta_y} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T^i} \mathcal{L}_r(\mathbf{x^i}; \theta_e, \theta_g) + \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y(\mathbf{x^i}; \theta_y, \theta_e) + \lambda \mathcal{R}(\theta_e)$$

Variational Adversarial Deep Domain Adaptation (VADDA) [ICLR 2017]



VRNN objective function [Chung et al, 2016]

$$\mathcal{L}_r(x_t^i; \theta_e, \theta_g) = E_{q_{\theta_e}(z^i_{\le T^i} | x^i_{\le T^i})} \sum_{t=1}^{T^i} (-D(q_{\theta_e}(z_t^i | x_{\le t}^i, z_{<t}^i) || p(z_t^i | x_{<t}^i, z_{<t}^i)) + \log p_{\theta_g}(x_t^i | z_{\le t}^i, x_{<t}^i))$$

Source classification loss with regularizer

$$\min_{\theta_e, \theta_g, \theta_y} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T^i} \mathcal{L}_r(\mathbf{x}^\mathbf{i}; \theta_e, \theta_g) + \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y(\mathbf{x}^\mathbf{i}; \theta_y, \theta_e) + \lambda \mathcal{R}(\theta_e)$$

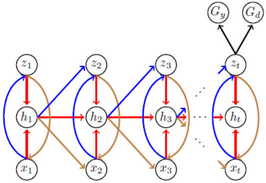
Domain regularizer [Ganin et al, 2016]

$$\mathcal{R}(\theta_e) = \max_{\theta_d} \left[-\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_d(\mathbf{x}^\mathbf{i}; \theta_d, \theta_e) - \frac{1}{n'} \sum_{i=n+1}^{N} \mathcal{L}_d(\mathbf{x}^\mathbf{i}; \theta_d, \theta_e) \right]$$

Variational Adversarial Deep Domain Adaptation (VADDA) [ICLR 2017]

VRNN objective function [Chung et al, 2016]

$$\mathcal{L}_r(x_t^i;\theta_e,\theta_g) = E_{q_{\theta_e}(z^i_{\le T^i}|x^i_{\le T^i})}\sum_{t=1}^{T^i}(-D(q_{\theta_e}(z_t^i|x^i_{\le t}, z^i_{<t})||p(z_t^i|x^i_{<t}, z^i_{<t})) + \log p_{\theta_g}(x_t^i|z^i_{\le t}, x^i_{<t}))$$



Inference: $z_t^i \sim q(z_t^i|x^i_{\le t}, z^i_{<t})$
Generation: $x_t^i \sim p(x_t^i|z^i_{\le t}, x^i_{<t})$
Recurrence: $h_t = RNN(x_t, z_t, h_{t-1})$

Source classification loss with regularizer

$$\min_{\theta_e,\theta_g,\theta_y} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T^i}\mathcal{L}_r(\mathbf{x^i};\theta_e,\theta_g) + \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_y(\mathbf{x^i};\theta_y,\theta_e) + \lambda\mathcal{R}(\theta_e)$$

Domain regularizer [Ganin et al, 2016]

$$\mathcal{R}(\theta_e) = \max_{\theta_d}\left[-\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_d(\mathbf{x^i};\theta_d,\theta_e) - \frac{1}{n'}\sum_{i=n+1}^{N}\mathcal{L}_d(\mathbf{x^i};\theta_d,\theta_e)\right]$$

Overall Objective function

$$E(\theta_e,\theta_g,\theta_y,\theta_d) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T^i}\mathcal{L}_r(\mathbf{x^i};\theta_e,\theta_g) + \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_y(\mathbf{x^i};\theta_y) - \lambda(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_d(\mathbf{x^i};\theta_d) + \frac{1}{n'}\sum_{i=n+1}^{N}\mathcal{L}_d(\mathbf{x^i};\theta_d)))$$

Experiments

Case Study: Acute Hypoxemic Respiratory Failure

- Datasets
  - Pediatric ICU: Child-AHRF
    - 398 patients at Children's Hospital Los Angeles (CHLA) Group 1: children (0-19 yrs)
  - MIMIC-III : Adult-AHRF
    - 5527 patients Group 2: working-age adult (20 to 45 yrs); Group 3: old working-age adult (46 to 65 yrs, Group 4: elderly (66 to 85 yrs); Group 5: old elderly (> 85 yrs)
- Input features - 21 time series variables (e.g., blood gas, ventilator signals, injury markers, etc.) for 4 days
- Prediction tasks - Mortality label

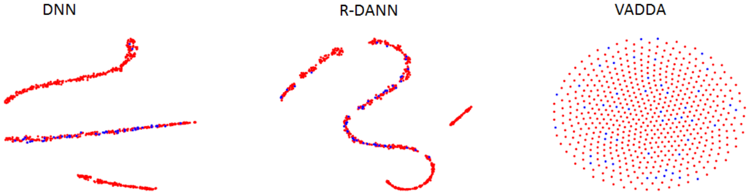
Classification Accuracy

Baselines:

- Non-domain adaptation: Logistic regression, Adaboost, Deep Neural Networks
- Deep Domain adaptation: DANN, R-DANN, VFAE [Louizos et al, 2015])

Source-Target	LR	Adaboost	DNN	DANN	VFAE	R-DANN	VRDDA
3- 2	0.555	**0.562**	0.569	0.572	0.615	0.603	**0.654**
4- 2	0.624	**0.645**	0.569	0.589	0.635	0.584	**0.656**
5- 2	0.527	**0.554**	0.551	0.540	0.588	0.611	**0.616**
2- 3	**0.627**	0.621	0.550	0.563	0.585	0.708	**0.724**
4- 3	**0.681**	0.636	0.542	0.527	0.722	**0.821**	0.770
5- 3	0.655	**0.706**	0.503	0.518	0.608	0.769	**0.782**
2- 4	0.585	**0.591**	0.530	0.560	0.582	0.716	**0.777**
3- 4	**0.652**	0.629	0.531	0.527	0.697	**0.769**	0.764
5- 4	0.689	**0.699**	0.538	0.532	0.614	0.728	**0.738**
2- 5	**0.565**	0.543	0.549	0.526	0.555	0.659	**0.719**
3- 5	0.576	**0.587**	0.510	0.526	0.533	0.630	**0.721**
4- 5	**0.682**	0.587	0.575	0.548	0.712	0.747	**0.775**
5- 1	0.502	**0.573**	0.557	0.563	0.618	0.563	**0.639**
4- 1	**0.565**	0.533	0.572	0.542	**0.668**	0.577	0.636
3- 1	0.500	0.500	0.542	0.535	0.570	0.591	**0.631**
2- 1	**0.520**	0.500	0.534	0.559	0.578	0.630	**0.637**

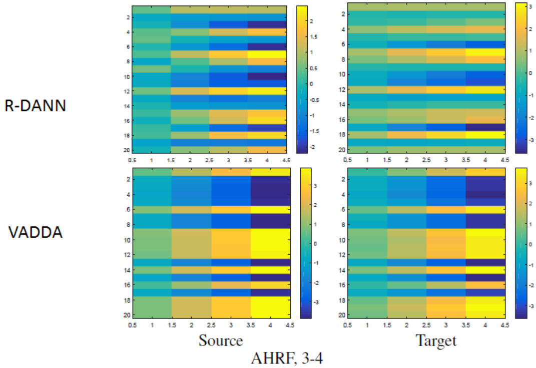
Domain-invariant Representations



t-SNE projections for the latent representations for domain adaptation from Adult-AHRF to Child-AHRF

VADDA has better distribution mixing than DANN

Temporal Dependencies across Domains



Memory cell state neuron activations of the R-DANN and VADDA

Activation patterns of VADDA are more consistent across time-steps than for R-DANN

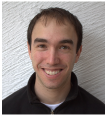
Outline

RECURRENT NEURAL NETWORKS FOR MULTIVARIATE TIME SERIES WITH MISSING VALUES


Zhengping Che


Sanjay Purushotham


Kyunghyun Cho

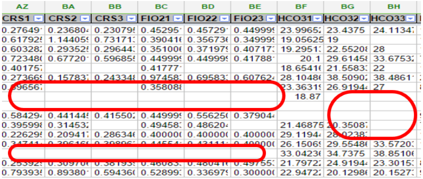

David Sontag

USC NYU

Che et al, Recurrent Neural Networks for Multivariate Time Series with Missing Values. arXiv:1606.01865

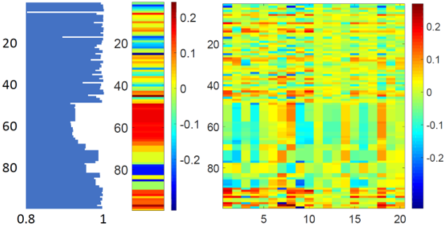
Missing Values are Useful

Missingness comes from various reasons.



AZ	BA	BB	BC	BD	BE	BF	BG	BH
CRS1	CRS2	CRS3	FIO21	FIO22	FiO23	HCO31	HCO32	HCO33
0.27649	0.23680	0.23079	0.45295	0.45729	0.44999	23.9965	23.4375	24.1134
0.61792	1.14405	0.73171	0.39041	0.35673	0.34999	19.05625	19	
0.60328	0.29352	0.29644	0.35100	0.37197	0.40717	19.2951	22.5520	28
0.72348	0.67720	0.59685	0.44999	0.44999	0.41788	20.1	29.6145	33.6753
0.40175			0.41777			18.6541	21.5583	22
0.27366	0.15783	0.24334	0.97458	0.69583	0.60762	28.1048	38.5090	38.4861
0.59656			0.35808			23.3631	26.9194	27
						18.87		
0.58429	0.44144	0.41550	0.44999	0.55625	0.37904			
0.39599	0.31453		0.49458	0.48620		21.46875	20.3508	
0.22629	0.20941	0.28634	0.40000	0.40000	0.40000	29.1194	28.0238	
0.34741	0.39616	0.39896	0.44554	0.43111	0.40000	26.1506	29.5548	33.5720
						33.0423	34.7375	38.8510
0.25392	0.30970	0.38193	0.46083	0.48041	0.49755	21.7972	24.9194	23.3015
0.79393	0.89380	0.59436	0.52899	0.33697	0.30000	22.9472	20.1298	20.1527

Missingness provides rich information about patients health condition.



Time-Series Inputs with Missing Values

Given time series data with missing values $\mathbf{X}$, we have two representations for missingness.

- *Masking* $\mathbf{M}$:
  Whether a variable is missing or not.
- *Time Interval* $\boldsymbol{\Delta}$:
  How long a variable has been missing.

Time Step	0	1	2	3	4	5	6	7
Value	47	55	N/A	49	N/A	N/A	47	50
Masking	1	1	0	1	0	0	1	1
Duration	1	1	1	2	1	2	3	1

There exist three solutions with no modification on the predictive models.

- Mean imputation of missing values (**Mean**)
- Forward imputation of missing values (**Forward**)
- Simple concatenation of indicators (**Simple**)

Deep Learning Models for Time-Series with Missing Values

- **Mean**: Replacing each missing observation by the mean of the variable $\tilde{x}$ across the training examples [Shao et al., 2009].

$$x_t^d \leftarrow m_t^d x_t^d + (1 - m_t^d)\tilde{x}^d$$

Input	47	55	$\tilde{x}$	49	$\tilde{x}$	$\tilde{x}$	47	50

- **Forward**: Assuming each missing value is the same as its last measurement $x_{t'}$ and using forward imputation [Unnebrink and Windeler, 2001].

$$x_t^d \leftarrow m_t^d x_t^d + (1 - m_t^d)x_{t'}^d$$

Input	47	55	55	49	49	49	47	50

- **Simple**: Concatenating the measurement x, masking m, and/or time interval δ.
  - Similar ideas are used in RNN models: [Choi et al., 2015](x and time t), Pham et al. [2016](x and δ), and [Lipton et al., 2016](x and m).

$$x_t^{(n)} \leftarrow \left[x_t^{(n)}; m_t^{(n)}; \delta_t^{(n)}\right]$$

Input	47	55	55	49	49	49	47	50
Masking	1	1	0	1	0	0	1	1
Duration	1	1	1	2	1	2	3	1

GRU-R [Che et al., 2016a]

Decay Term γ: A flexible transformation on $\mathbf{\Delta}$ jointly learned with deep model.

$$\gamma_t = \exp\{-ReLU(\mathbf{W}_\gamma \delta_t + \mathbf{b}_\gamma)\}$$

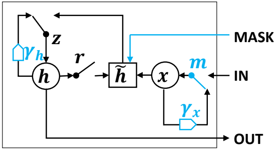
GRU-D model

- Decay on the last observations:

$$x_t^d \leftarrow m_t^d x_t^d + (1 - m_t^d)\gamma_{\boldsymbol{x}_t}^d x_{t'}^d + (1 - m_t^d)(1 - \gamma_{\boldsymbol{x}_t}^d)\tilde{x}^d$$

- Decay on the hidden states:

$$\boldsymbol{h}_{t-1} \leftarrow \boldsymbol{\gamma}_{\boldsymbol{h}_t} \odot \boldsymbol{h}_{t-1}$$



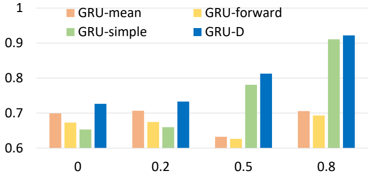
The update functions for GRU are:

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z \boldsymbol{x}_t + \boldsymbol{U}_z \boldsymbol{h}_{t-1} + \boldsymbol{V}_z \boldsymbol{m}_t + \boldsymbol{b}_z) \boldsymbol{r}_t = \sigma(\boldsymbol{W}_r \boldsymbol{x}_t + \boldsymbol{U}_r \boldsymbol{h}_{t-1} + \boldsymbol{V}_r \boldsymbol{m}_t + \boldsymbol{b}_r)$$

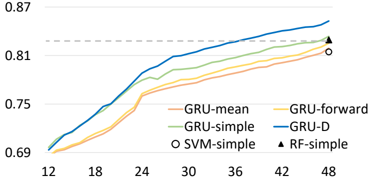
$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}\boldsymbol{x}_t + \boldsymbol{U}(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{V}\boldsymbol{m}_t + \boldsymbol{b}) \boldsymbol{h}_t = (\boldsymbol{1} - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t$$

Empirical Evaluation

Evaluations on synthetic dataset with different missing rates



Evaluations for mortality early prediction

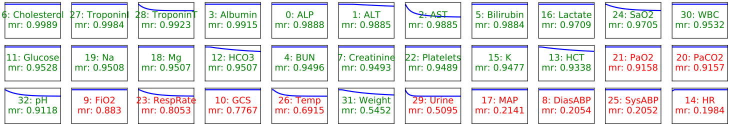


AUC score on mortality prediction

	Models	MIMIC-III	PhysioNet
Non-RNN	LR-forward	0.7589	0.7423
	SVM-forward	0.7908	0.8131
	RF-forward	0.8293	0.8183
	LR-simple	0.7715	0.7625
	SVM-simple	0.8146	0.8277
	RF-simple	0.8294	0.8157
RNN	LSTM-mean	0.8142	0.8025
	GRU-mean	0.8192	0.8195
	GRU-forward	0.8252	0.8162
	GRU-simple	0.8380	0.8155
Ours	GRU-D	**0.8527**	**0.8424**

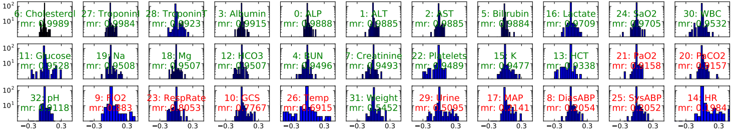
Empirical Evaluation

Input decay plots of all 33 variables for mortality prediction on PhysioNet dataset



- Get a few important variables, e.g., weight, arterial pH, temperature, and respiration rate, etc.

Histograms of of hidden state decay for mortality prediction on PhysioNet dataset



- Parameters related to variables with smaller missing rate are more spread out.

Outline

DEEP COMPUTATIONAL PHENOTYPING

Zhengping Che

David Kale

Wenzhe Li

Taha Bahadori

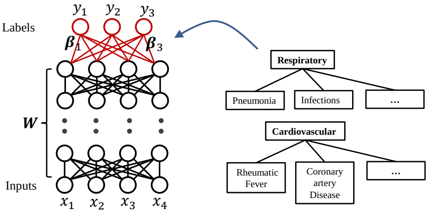
USC

Che et al, Deep Computational Phenotyping. Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD), 2015.

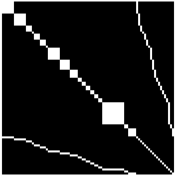
90

Label Sparsity and Structured Domain Knowledge

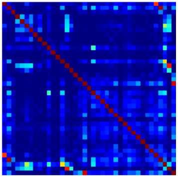
- Many diagnoses occur in $< 1\%$ of patients.
  How do we handle sparsity in our labels?
- Ontologies (e.g., ICD-9 diagnostic codes) describe relationships between diseases.
  How can we incorporate (structured) domain knowledge?
- **Solution:** Multi-task net + Graph Laplacian regularization.



Our solutions



Tree-based priors



Co-occurrence priors

Multi-task Neural Nets + Graph Laplacian Regularization [Che et al., 2015]

Assume:

- K outputs (labels) with parameters $\{\boldsymbol{\beta}_k\}_{k=1}^K$, $\boldsymbol{\beta}_k \in \mathbb{R}^{D^{(L)}}$
- Label similarity matrix $\boldsymbol{A} \in \mathbb{R}^{K \times K}$ where $A_{ij} \in [0, 1]$.

Define Graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{C} - \boldsymbol{A}$ with $\boldsymbol{C}$ a diagonal matrix $C_{kk} = \sum_{k'=1}^K A_{kk'}$, then

$$\mathrm{tr}(\boldsymbol{\beta}^\top \mathbf{L} \boldsymbol{\beta}) = \sum_{1 \le k, k' \le K} \mathrm{A}_{k,k'} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}\|_2^2$$

where $\mathrm{tr}(\cdot)$ represents the *trace* operator.

Regularized loss function for supervised training of multi-task neural network:

$$\mathcal{L} = -\sum_{i=1}^N \sum_{k=1}^K \Big[y_{ik} \log \sigma(\boldsymbol{\beta}_k^\top \mathbf{h}_i) + (1 - y_{ik}) \log(1 - \sigma(\boldsymbol{\beta}_k^\top \mathbf{h}_i)) \Big] + \frac{\rho}{2} \mathrm{tr}(\boldsymbol{\beta}^\top \mathbf{L} \boldsymbol{\beta})$$

Experiment Results

Impact of priors on phenotype classification

PICU data (AUROC across 67 labels and 19 categories from ICD-9 codes)

	Tasks	*No Prior*	*Co-Occurrence*	*ICD-9 Tree*
Subsequence	*All*	0.7079 ± 0.0089	$\mathbf{0.7169 \pm 0.0087}$	0.7143 ± 0.0066
	Categories	0.6758 ± 0.0078	$\mathbf{0.6804 \pm 0.0109}$	0.6710 ± 0.0070
	Labels	0.7148 ± 0.0114	$\mathbf{0.7241 \pm 0.0093}$	0.7237 ± 0.0081
Episode	*All*	0.7245 ± 0.0077	$\mathbf{0.7348 \pm 0.0064}$	0.7316 ± 0.0062
	Categories	0.6952 ± 0.0106	$\mathbf{0.7010 \pm 0.0136}$	0.6902 ± 0.0118
	Labels	0.7308 ± 0.0099	$\mathbf{0.7414 \pm 0.0064}$	0.7407 ± 0.0070

MED2VEC: MULTI-LAYER REPRESENTATION LEARNING FOR MEDICAL CONCEPTS

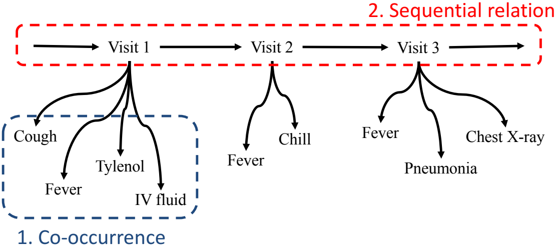
E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, J. Sun, (2016)

Multi-layer Representation Learning for Medical Concepts

KDD'16

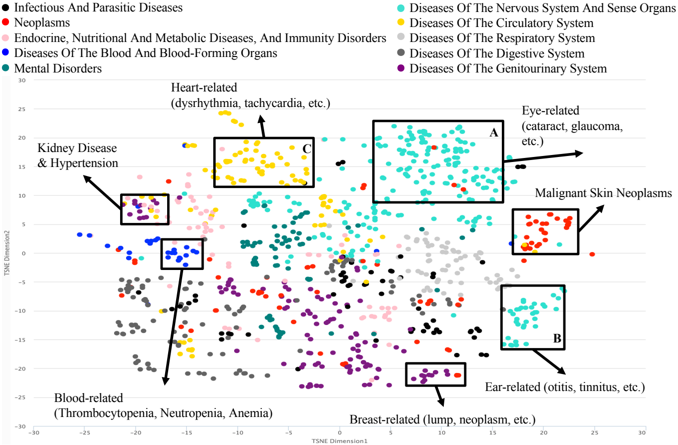
Med2Vec: two-layered representation learning

- *Abstraction in patient records*



- *Objective function: the sum of*
  1. *Negative intra-visit Skip-gram*
     - *Because Skip-gram objective function is to be maximized*
  2. *Inter-visit multi-label classification loss*

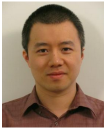
Med2Vec encoding is well aligned with medical knowledge



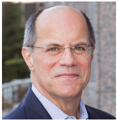
GRAM: GRAPH-BASED ATTENTION MODEL FOR HEALTHCARE REPRESENTATION LEARNING

Edward Choi

Taha Bahadori



Le Song



Buzz Stewart

Georgia Tech College of Computing
Computational Science and Engineering

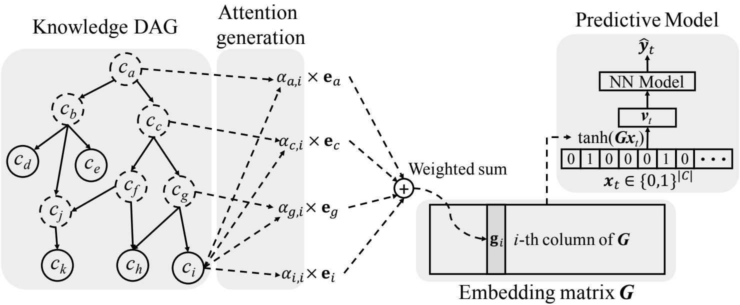
Sutter Health

KDD'17

65

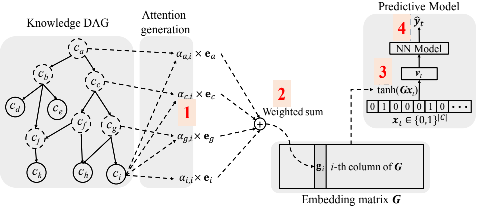
GRAM: Learn representations of medical codes leveraging medical ontologies

- *Method: Generate a medical code representation vector by combining the representation vectors of its ancestors using the attention mechanism*



Model structure of GRAM

66

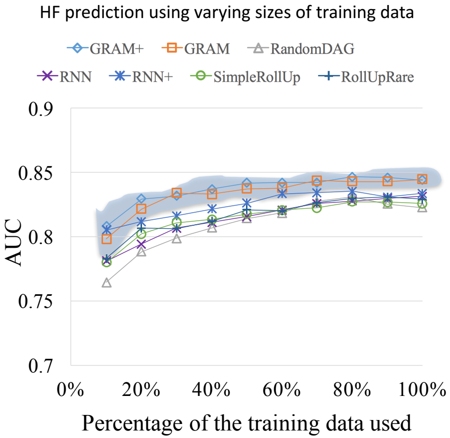


GRAM algorithm

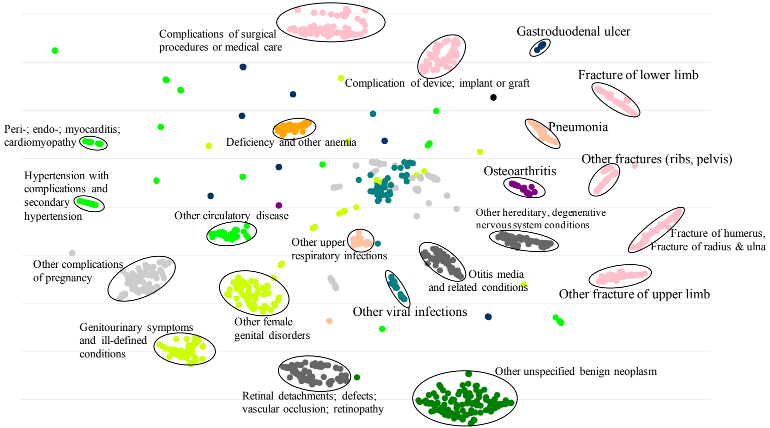
1	$\alpha_{ij} = \frac{\exp(f(\mathbf{e}_i, \mathbf{e}_j))}{\sum_{k \in \mathcal{A}(i)} \exp(f(\mathbf{e}_i, \mathbf{e}_k))}$ where $f(\mathbf{e}_i, \mathbf{e}_j) = \mathbf{u}_a^\top \tanh(\mathbf{W}_a \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_j \end{bmatrix} + \mathbf{b}_a)$
	Attention weights are generated for all pairs of basic embeddings $\mathbf{e}_i$ and its ancestors $\mathbf{e}_j$.
2	$\mathbf{g}_i = \sum_{j \in \mathcal{A}(i)} \alpha_{ij} \mathbf{e}_j,$
	Final representation $\mathbf{g}_i$ is the weighted sum of attention weights and basic embeddings.
3	$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t = \tanh(\mathbf{G}[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t])$
	Sequence of visit representations are obtained using the Embedding matrix $\boldsymbol{G}$.
4	$\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_t = \text{RNN}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t, \theta_r),$ $\widehat{\mathbf{y}}_t = \widehat{\mathbf{x}}_{t+1} = \text{Softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b}),$
	Performing sequential diagnoses prediction, outcomes are generated by RNN and Softmax.

GRAM provide accurate prediction

GRAM shows better predictive performance under data constraints

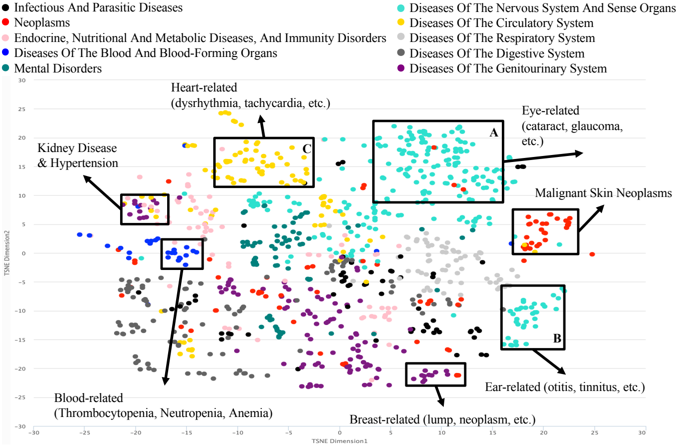


GRAM learns representations well aligned with knowledge ontology

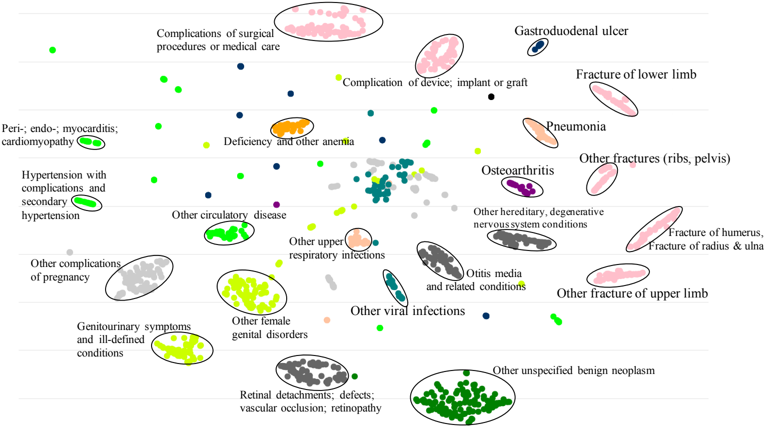


Scatterplot of GRAM representations

Med2Vec encoding is well aligned with medical knowledge



GRAM learns representations well aligned with knowledge ontology



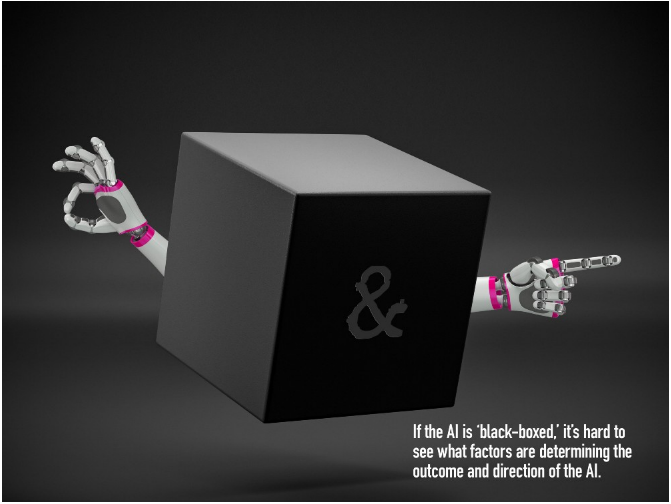
Scatterplot of GRAM representations

GRAM: Graph-based Attention Model for Healthcare Representation Learning

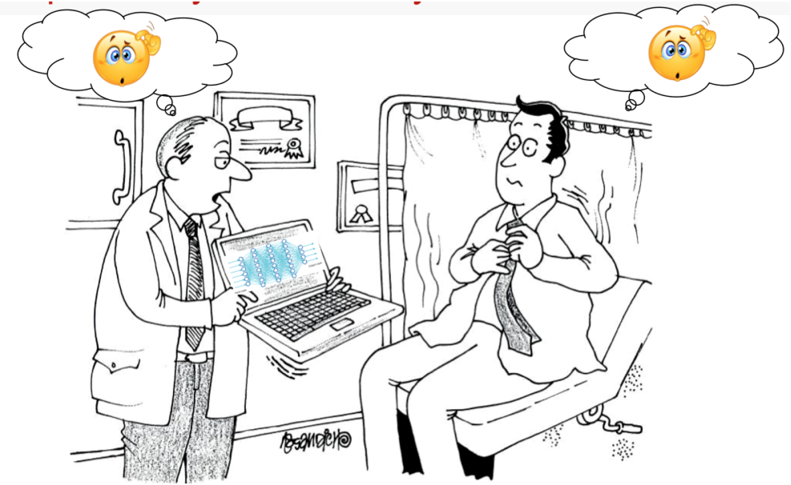
- Robust representation against data insufficiency
- *Interpretable*: Well aligned with medical knowledge

Outline

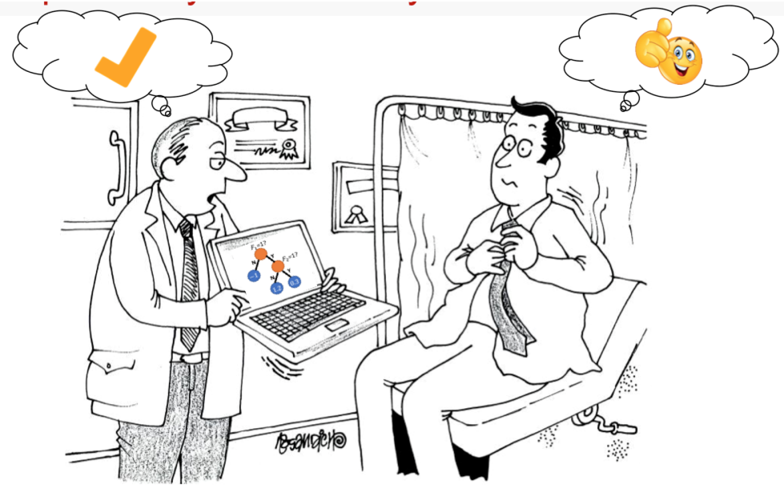
Deep Learning as Blackbox



Importance of Explainable Artificial Intelligence - I



Importance of Explainable Artificial Intelligence - I



Importance of Explainable Artificial Intelligence - II

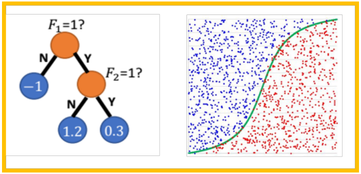
How can I trust any machine learning algorithm? [Ribeiro et al, 2016]

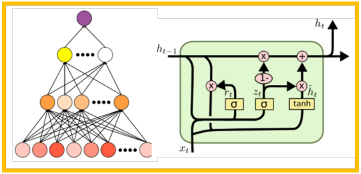

(a) Husky classified as wolf


(b) Explanation

Interpretable Model is Necessary

Interpretable predictive models are shown to result in faster adoptability of machine learning models.





- Simple and commonly use models
- Easy to interpret, mediocre performance

- Deep learning solutions
- Superior performance, hard to explain

Can we learn interpretable models with robust prediction performance?

Ongoing Work on Explainable Machine Learning Models

Direct Interpretation

- [Garson, 1991]: estimating feature importance directly from network weight connections
- [Hechtlinger, 2016]: computing output gradients with respect to input features
- [Itti et al., 1998; Mnih et al., 2014; Xu et al., 2015]: attention models

Indirect Interpretation

- [Provost et al., 1997]: sensitivity analysis of feature contributions to a neural network's output
- [Ribeiro et al., 2016]: local interpretability for black-box models
- [Che et al., 2016b]: mimicking the blackbox through the prediction scores
- [Maaten and Hinton, 2008; Simonyan et al., 2013; Yosinski et al., 2014; LeCun et al., 2015; Mnih et al., 2015; Mahendran and Vedaldi, 2015]: visualizing the hidden units

INTERPRETABLE DEEP MODELS FOR ICU OUTCOME PREDICTION


Zhengping Che


Sanjay Purushotham


Robinder Khemani

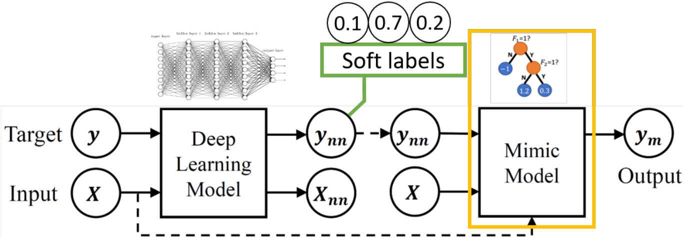
USC

Children's Hospital LOS ANGELES

Che et al, Interpretable Deep Models for ICU Outcome Prediction. of the American Medical Informatics Association Annual Symposium (AMIA), 2016.

Interpretable Mimic Learning Framework [Che et al., 2016b]

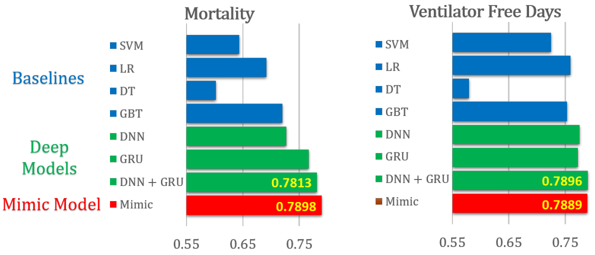
- Main ideas:
  - Borrow the ideas from knowledge distillation [Hinton, et al., 2015] and mimic learning [Ba, Caruana, 2014].
  - Use **Gradient Boosting Trees (GBTs)** to mimic deep learning models.
- Training Pipeline:



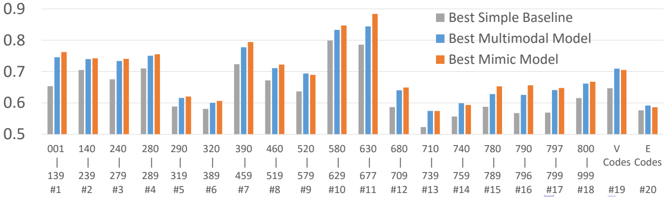
- Benefits: Good performance, less overfitting, interpretations.

Quantitative Evaluation

AUROC score of prediction on patients with acute hypoxemic respiratory failure.

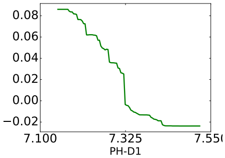


AUROC score of 20 ICD-9 diagnosis category prediction tasks on MIMIC-III dataset.



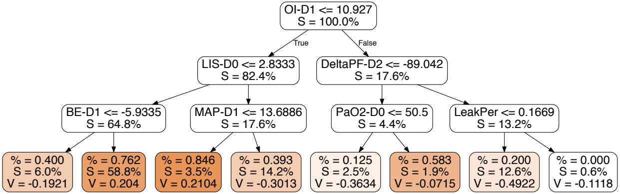
Model/Feature Interpretation

Partial dependency plot for mortality prediction on patients with acute hypoxemic respiratory failure.



- pH value in blood should stay in a normal range around 7.35-7.45.
- Our model predicts a higher mortality change when the patient pH value below 7.325.

Most Useful Decision Trees for ventilator free days prediction.



Useful features:

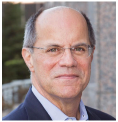
- Lung injury score
- Oxygenation index
- PF ratio change

RETAIN: INTERPRETABLE DEEP LEARNING MODEL


Edward Choi


Taha Bahadori


Andy Schuetz


Buzz Stewart

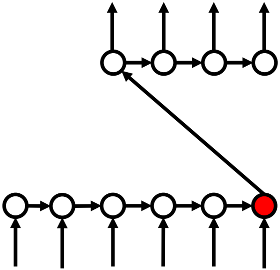
Georgia Tech College of Computing
Computational Science and Engineering

Sutter Health

Choi, Edward, et al. 2016. "RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism." In *NIPS*

Regular Machine Translation

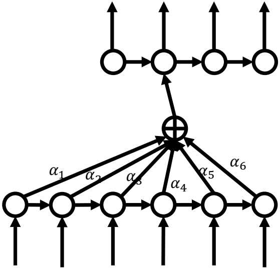
如果你不在乎谁获得了荣誉，
你所能完成的事情是**惊人的**。



It is amazing what you can accomplish
if you do not care who gets the credit

Neural Attention Mechanism

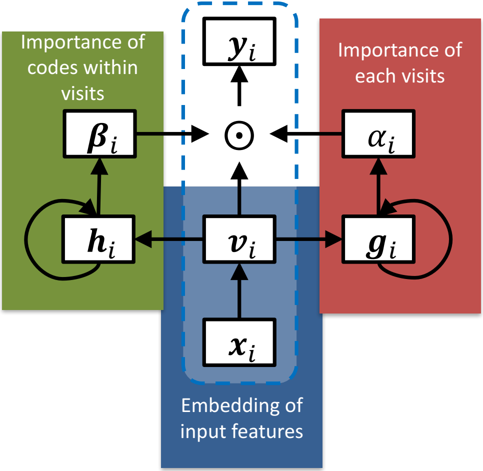
如果你不在乎谁获得了荣誉，
你所能完成的事情是惊人的。



It is amazing what you can accomplish
if you do not care who gets the credit

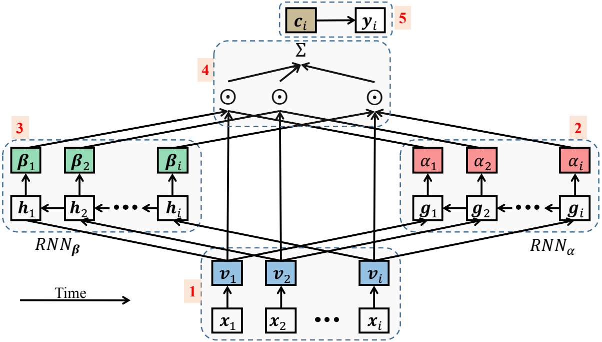
Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. "Neural Machine Translation by Jointly Learning to Align and Translate." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1409.0473.

RETAIN: REverse Time AttentIoN model

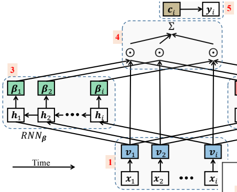


Choi, Edward, et al. 2016. “RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism.” In *NIPS*

Details of RETAIN

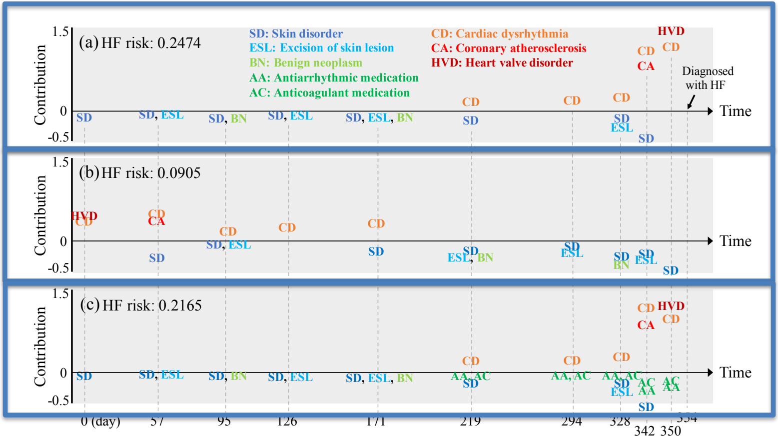


RETAIN Algorithm



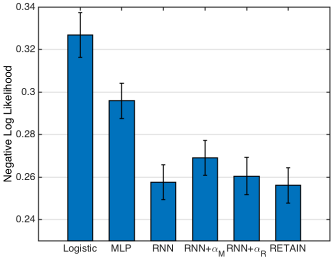
1	$\mathbf{v}_i = \mathbf{E}\mathbf{x}_i$
	Multi-hot representation of the visit is linearly projected by the embedding matrix $\mathbf{E}$.
2	$\mathbf{g}_i, \mathbf{g}_{i-1}, \ldots, \mathbf{g}_1 = \text{RNN}_\alpha(\mathbf{v}_i, \mathbf{v}_{i-1}, \ldots, \mathbf{v}_1)$, $\alpha_1, \alpha_2, \ldots, \alpha_i = \text{Softmax}(\mathbf{w}_\alpha^\top [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_i] + b_\alpha)$
	RNN_α generates α_i, the scalar attention weight for the i-th visit. The visit representations $\boldsymbol{v}_i$'s are fed to the RNN_α in reverse order.
3	$\mathbf{h}_i, \mathbf{h}_{i-1}, \ldots, \mathbf{h}_1 = \text{RNN}_\beta(\mathbf{v}_i, \mathbf{v}_{i-1}, \ldots, \mathbf{v}_1)$, $\boldsymbol{\beta}_j = \tanh\left(\mathbf{W}_\beta \mathbf{h}_j + \mathbf{b}_\beta\right) \quad \text{for} \quad j = 1, \ldots, i$
	RNN_β generates $\boldsymbol{\beta}_i$, the vector attention weight for the medical codes in the i-th visit. $\boldsymbol{v}_i$'s are fed to the RNN_β in reverse order as well.
4	$\mathbf{c}_i = \sum_{j=1}^{i} \alpha_j \boldsymbol{\beta}_j \odot \mathbf{v}_j$
	The attention weights α_i and $\boldsymbol{\beta}_i$ are combined with the visit representation $\boldsymbol{v}_i$ to obtain the context vector $\boldsymbol{c}_i$.
5	$\widehat{\mathbf{y}}_i = \text{Softmax}(\mathbf{W}\mathbf{c}_i + \mathbf{b})$
	Using the context vector $\boldsymbol{c}_i$, we make the final prediction.

Interpretation of RETAIN model

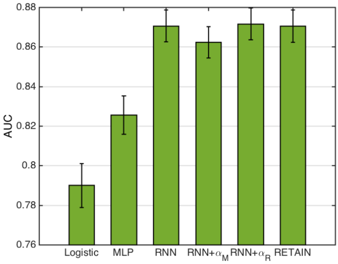


Heart Failure Results

Negative Log Likelihood on Test Set

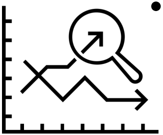


Classification AUC



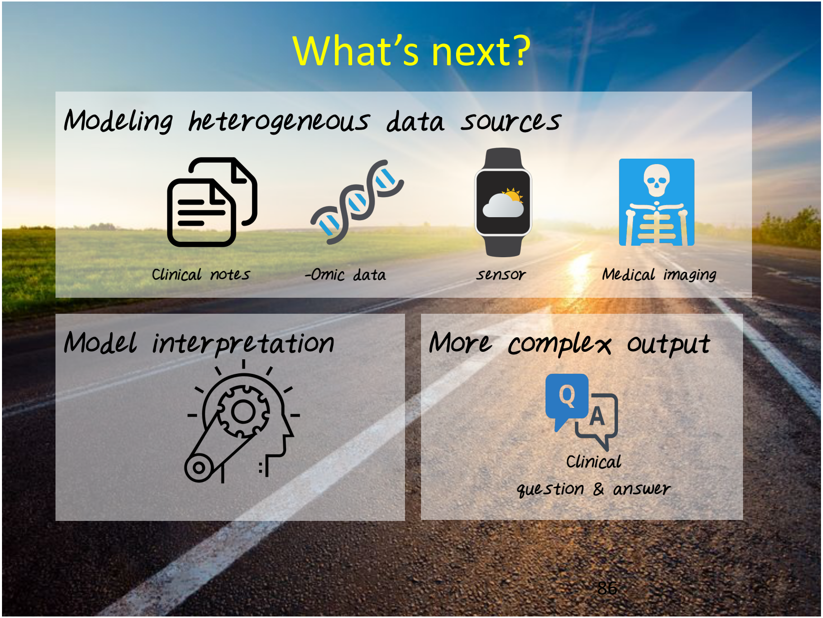
Retain: Interpretable Deep learning model

- Challenge: Deep learning models are often difficult to interpret



- RETAIN is a temporal attention model on electronic health records
  - Great predictive power
  - Good interpretation

Choi, Edward, et al. 2016. "RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism." In *NIPS*

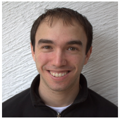

What's next?
Modeling heterogeneous data sources
Clinical notes
-Omic data
sensor
Medical imaging
Model interpretation
More complex output
Q
A
Clinical
question & answer
86

Acknowledgment Yan Liu @ USC

Funding agencies: CCF-1539608, IIS-1254206, Samsung, USC Coulter Foundation, USC Ming Hsieh Institute

Collaborators:

Children's Hospital Los Angeles, Mayo Clinic, IBM, Samsung



David Sontag (MIT)

Kyunghyun Cho (NYU)

Jimeng Sun @ Georgia Tech
Healthcare Analytics

SunLab
Enlighten the World

Collaborators & Sponsors

CDC | Children's Healthcare of Atlanta | MGH 1811 | Vanderbilt | Northwestern University | ucb | amazon

NSF | Sutter Health We Plus You | The University of Texas at Austin | Emory University | Google

IIS#1418511
& CCF#1533768

Geisinger Health System | Microsoft Azure

Government | *Provider* | *University* | *Company*

Contacts and Additional Information

Yan Liu
yanliu.cs@usc.edu

Jimeng Sun
jsun@cc.gatech.edu

Tutorial websites: https://tinyurl.com/y7wuk9xt

References I

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144.

Che, Z., Kale, D., Li, W., Bahadori, M. T., and Liu, Y. (2015). Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516. ACM.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2016a). Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*.

Che, Z., Purushotham, S., Khemani, R., and Liu, Y. (2016b). Interpretable deep models for icu outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 371. American Medical Informatics Association.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. (2016). Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318.

Choi, E., Bahadori, M. T., and Sun, J. (2015). Doctor ai: Predicting clinical events via recurrent neural networks. *arXiv preprint arXiv:1511.05942*.

Dabek, F. and Caban, J. J. (2015). A neural network based model for predicting psychological conditions. In *International Conference on Brain Informatics and Health*, pages 252–261. Springer.

References II

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. (2013). Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967.

Ganin, Y. and Lempitsky, V. (2014). Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1).

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert*, 6(4):46–51.

Hammerla, N. Y., Fisher, J., Andras, P., Rochester, L., Walker, R., and Plötz, T. (2015). Pd disease state assessment in naturalistic environments using deep learning. In *AAAI*, pages 1742–1748.

Hechtlinger, Y. (2016). Interpretation of prediction models using the input gradient. *arXiv preprint arXiv:1611.07634*.

Huang, F. and Yates, A. (2009). Distributional representations for handling sparsity in supervised sequence-labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 495–503. Association for Computational Linguistics.

References III

Hughes, M., Li, I., Kotoulas, S., and Suzumura, T. (2017). Medical text classification using convolutional neural networks.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

Jagannatha, A. N. and Yu, H. (2016). Bidirectional RNN for medical event detection in electronic health records. *Proc Conf*, 2016:473–482.

Jiang, J. and Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *ACL*, volume 7, pages 264–271.

Kale, D., Che, Z., Liu, Y., and Wetzel, R. (2014). Computational discovery of physiomes in critically ill children using deep learning. In *Workshop DMMI in AMIA*.

Kale, D. C., Che, Z., Bahadori, M. T., Li, W., Liu, Y., and Wetzel, R. (2015). Causal phenotype discovery via deep networks. In *AMIA Annual Symposium Proceedings*, volume 2015, page 677. American Medical Informatics Association.

Lasko, T. A., Denny, J. C., and Levy, M. A. (2013). Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

References IV

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzell, R. (2015). Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

Lipton, Z. C., Kale, D. C., and Wetzel, R. (2016). Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. *arXiv preprint arXiv:1606.04130*.

Luo, Y. (2017). Recurrent neural networks for classifying relations in clinical notes. *J. Biomed. Inform.*, 72:85–95.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Mahendran, A. and Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5188–5196.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094.

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

References V

Pandey, C., Ibrahim, Z., Wu, H., Iqbal, E., and Dobson, R. (2017). Improving RNN with attention and embedding for adverse drug reactions. In *Proceedings of the 2017 International Conference on Digital Health*, DH '17, pages 67–71, New York, NY, USA. ACM.

Pham, T., Tran, T., Phung, D., and Venkatesh, S. (2016). Deepcare: A deep dynamic memory model for predictive medicine. *arXiv preprint arXiv:1602.00357v1*.

Prakash, A., Zhao, S., Hasan, S. A., Datla, V., Lee, K., Qadir, A., Liu, J., and Farri, O. (2016). Condensed memory networks for clinical diagnostic inferencing.

Provost, F. J., Fawcett, T., et al. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD*, volume 97, pages 43–48.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Shao, J., Jordan, D. C., and Pritchett, Y. L. (2009). Baseline observation carry forward: reasoning, properties, and practical issues. *Journal of biopharmaceutical statistics*, 19(4):672–684.

Shin, B., Chokshi, F. H., Lee, T., and Choi, J. D. (2017). Classification of radiology reports using neural attention models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4363–4370.

References VI

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Socher, R., Lin, C. C., Manning, C., and Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076.

Unnebrink, K. and Windeler, J. (2001). Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. *Statistics in medicine*, 20(24):3931–3946.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.