Clinical data successes using machine learning

A survey by Joseph Paul Cohen, PhD
Montreal Institute for Learning Algorithms

Topics:

1. Medical Concept Representation
2. Clinical Event Prediction
Where is the Deep Learning research?

Deep EHR: Application Areas
- Representation
- Representation learning
- Concept representation
- Phenotyping
- Information extraction
- Prediction
- Deidentification

Deep EHR: Technical Methods
- Unsupervised
- RNN
- LSTM
- GRU
- CNN
- Autoencoder (~word2vec)
- RBM
- DBN
- Skip-gram (word2vec)

Mr. Smith is a 63-year-old gentleman with coronary artery disease, hypertension, hypercholesterolemia, COPD and tobacco abuse. He reports doing well. He did have some more knee pain for a few weeks, but this has resolved. He is having more trouble with his sinuses. I had started him on Flonase back in December. He says this has not really helped. Over the past couple weeks he has had significant congestion and thick discharge. No fevers or headaches but does have diffuse upper right-sided teeth pain. He denies any chest pains, palpitations, PND, orthopnea, edema or syncope. His breathing is doing fine. No cough. He continues to smoke about half-a-pack per day. He plans on trying the patches again.
Extract relationships between words and produce a latent representation
What to do with word embeddings?

- We can compose them to create paragraph embeddings (bag of embeddings).
- Use in place of words for an RNNs (More on this later!)
- Augment learned representations on small datasets

Study the compositionality of the learned latent space

[Cultural Shift or Linguistic Drift, Hamilton, 2016]

Study how the meaning between two texts varies (or hospitals, or doctors)?

[Mikolov, 2013]  
[Pennington, 2014]  
[Cultural Shift or Linguistic Drift, Hamilton, 2016]
Token representations

One-hot encoding: binary vector per token

Example:

cat = \[0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ ... \ 0\]
dog = \[0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ ... \ 0\]
house = \[1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ ... \ 0\]

Note!
If x is one hot
The dot product of Wx = a Single column of W

\[M \times N\] \[N \times 1\] = \[M \times 1\]
1. Each word is a training example
2. Each word is used in many contexts
3. The context defines each word

Mikolov, Efficient Estimation of Word Representations in Vector Space, 2013
Learning in progress
king + (woman - man) = ?

The point that is closest is queen!
Follow along online!

https://colab.research.google.com/drive/1g4zvEg921sLQK-VsBk5mMb2-h4goCGyd
window_size = 2
idx_pairs = []

for sentence in tokenized_corpus:
    indices = [word2idx[word] for word in sentence]

    for center_word_pos in range(len(indices)):
        for w in range(-window_size, window_size + 1):
            context_word_pos = center_word_pos + w
            if context_word_pos < 0 or context_word_pos >= len(indices) or center_word_pos == context_word_pos:
                continue
            context_word_idx = indices[context_word_pos]
            idx_pairs.append((indices[center_word_pos], context_word_idx))

sentence = ['paris', 'france', 'capital']
indices = [2, 5, 11]
```python
class SkipGram(nn.Module):
    def __init__(self, vocab_size, embd_size):
        super(SkipGram, self).__init__()

        self.W1 = Variable(torch.randn(embd_size, vocab_size).float(), requires_grad=True)
        self.W2 = Variable(torch.randn(vocab_size, embd_size).float(), requires_grad=True)

    def forward(self, focus):
        z1 = torch.matmul(self.W1, focus)
        z2 = torch.matmul(self.W2, z1)
        softmax = F.log_softmax(z2, dim=0)
        return softmax

model = SkipGram(vocabulary_size, 2)
```

Initialize two matrices

\[ |E| \times |V| \]

\[ |V| \times |E| \]

Encoder:

\[ |E| \times |V| \text{ dot } |V| \times 1 = |E| \times 1 \]

Decoder:

\[ |V| \times |E| \text{ dot } |E| \times 1 = |V| \times 1 \]

Softmax over context prediction

Code by github user: mbednarski
A note on the softmax function

To predict multiple classes we project to a probability distribution
Because it is on a simplex; the correction of one term impacts all
Infinite ways to generate the same output.

A correction of one sends gradients to others

We can learn unseen classes through a process of elimination.

https://github.com/ieee8023/NeuralNetwork-Examples/blob/master/general/simplex-softmax.ipynb
To predict multiple class we can project the output onto a simplex and compute the loss there.
Open Access Subset of PubMedCentral

Subset of files that are available open access

- Papers available in PDF or XML
- 1.25 million biomedical articles and 2 million distinct words
- Available over FTP for bulk download (CompSci Friendly!)
- Metadata includes journal name and year

Example XML data

```xml
<sec sec-type="methods">
<title>Patients and methods</title>
</sec>
<title>Patients</title>
<p>Between September 1995 and September 1998, 413 patients with abnormal breast findings were referred for histological evaluation to the Department of Gynecology of the Friedrich-Schiller University, Jena, Germany. Patients had been selected and referred because of the presence of breast lesions detected by palpation and/or mammography and/or sonography. In addition, MR mammography was performed in all patients. We excluded five patients with invasive cancer who had a history of core-needle or fine-needle biopsy cancer within 2 weeks before referral, because the presence of haemorrhage may mimic false-positive findings on MR mammography. In addition, five patients who did not keep still during MR mammography were excluded. The</p>
</sec>
<title>Imaging</title>
</sec>
<p>Analysis of the sonograms taken in patients with histologically confirmed carcinoma were excluded from analysis because the value of sonography for detection of premalignant disease is</p>
```

Journal names

- Breast Cancer Res.
- Genome Biol.
- Arthritis Res.
- BMC Cell Biol.
...
Representations are biased by the data.

We can use this to our advantage to control the domain.
The full name of diabetes is "diabetes mellitus"

A type of peptic ulcer

Colon cancer is associated with breast cancer!

Articles say diabetics are at increased risk of hypertension (high blood pressure)

One symptom is an ulcer

<table>
<thead>
<tr>
<th>Specific</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EHR (Mayo Clinic)</strong></td>
<td><strong>PubMed</strong></td>
</tr>
<tr>
<td>diabetes</td>
<td>mellitus,</td>
</tr>
<tr>
<td>controlled,</td>
<td>nonalcoholic,</td>
</tr>
<tr>
<td>hyperlipidemia,</td>
<td>obesity,</td>
</tr>
<tr>
<td>peptic ulcer disease</td>
<td>mellitus,</td>
</tr>
<tr>
<td>peptic ulcer</td>
<td>polycystic</td>
</tr>
<tr>
<td>duodenal,</td>
<td>gastritis,</td>
</tr>
<tr>
<td>pernicious</td>
<td>alcoholism,</td>
</tr>
<tr>
<td>gastroduodenal,</td>
<td>rheumatic,</td>
</tr>
<tr>
<td>diverticular</td>
<td>ischaemic,</td>
</tr>
<tr>
<td>diverticular</td>
<td>nephropathy</td>
</tr>
<tr>
<td>colon cancer</td>
<td>breast,</td>
</tr>
<tr>
<td>ovarian,</td>
<td>mcf,</td>
</tr>
<tr>
<td>prostate,</td>
<td>cancers,</td>
</tr>
<tr>
<td>postmenopausally,</td>
<td>tumor Suppressing,</td>
</tr>
<tr>
<td>cancer</td>
<td>downregulation</td>
</tr>
</tbody>
</table>
Indiana University Hospital Reports

Chest X-ray images from the Indiana University hospital network

Indiana University Chest X-ray Collection
Affiliation: Indiana University

ABSTRACT

Comparison: None.
Indication: Positive TB test
Findings: The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.
Impression: Normal chest x-XXXX.

NOTE: The data are drawn from multiple hospital systems.
Show MeSH
Related in: MedlinePlus Request Collection

1000 reports available in XML format!
The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.

Normal chest x-XXXX.
Hyperparameters!

Depending on the configuration of the model the embeddings can vary:

We can vary: the dimension of the embedding, learning rate, token words, window size, etc..
Size of embedding space matters

<table>
<thead>
<tr>
<th></th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>spine 0.000000</td>
</tr>
<tr>
<td>1</td>
<td>seen 6.632637</td>
</tr>
<tr>
<td>2</td>
<td>evidence 8.230038</td>
</tr>
<tr>
<td>3</td>
<td>infiltrate 13.233962</td>
</tr>
<tr>
<td>4</td>
<td>findings 17.762381</td>
</tr>
<tr>
<td>5</td>
<td>disease 19.574038</td>
</tr>
<tr>
<td>6</td>
<td>focal 22.267534</td>
</tr>
<tr>
<td>7</td>
<td>aortic 23.346777</td>
</tr>
<tr>
<td>8</td>
<td>suspicious 25.558247</td>
</tr>
<tr>
<td>9</td>
<td>acute 29.176380</td>
</tr>
<tr>
<td>10</td>
<td>within 29.798399</td>
</tr>
</tbody>
</table>

*via t-sne

Dim = 2

<table>
<thead>
<tr>
<th></th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>spine 0.000000</td>
</tr>
<tr>
<td>1</td>
<td>severe 3.304547</td>
</tr>
<tr>
<td>2</td>
<td>changes 3.469765</td>
</tr>
<tr>
<td>3</td>
<td>anterior 3.474130</td>
</tr>
<tr>
<td>4</td>
<td>mild 3.731760</td>
</tr>
<tr>
<td>5</td>
<td>also 3.828588</td>
</tr>
<tr>
<td>6</td>
<td>thoracic 3.838329</td>
</tr>
<tr>
<td>7</td>
<td>noted 3.871284</td>
</tr>
<tr>
<td>8</td>
<td>calcifications 3.934170</td>
</tr>
<tr>
<td>9</td>
<td>vertebral 4.004414</td>
</tr>
<tr>
<td>10</td>
<td>stable 4.022851</td>
</tr>
</tbody>
</table>

Dim = 100
Time series medical records

Tasks:

- Predict probability of event in future
- Predict duration until next visit
- Find similar patients/events/visits

We need to define what events are!
ICD (International Classification of Diseases)

ICD is the foundation for the identification of health trends and statistics globally, and the international standard for reporting diseases and health conditions. It is the diagnostic classification standard for all clinical and research purposes. ICD defines the universe of diseases, disorders, injuries and other related health conditions, listed in a comprehensive, hierarchical fashion (http://www.who.int/)

1893 - Causes of Death (International Statistical Institute)

1975 - ICD-9 - (WHO)

1990 - ICD-10 - (WHO)

2022 - ICD-11 - (WHO)

There are alternative standards but they can require a fee.
Example ICD-9 codes:

786 Symptoms involving respiratory system and other chest symptoms
786.0 Dyspnea and respiratory abnormalities
786.1 Stridor
786.2 Cough
786.3 Hemoptysis
786.4 Abnormal sputum
786.5 Chest pain
786.6 Swelling, mass or lump in chest
786.7 Abnormal chest sounds
786.8 Hiccough
786.9 Other

Example ICD-9 codes (786.5)

Cardialgia (see also Pain, precordial) 786.51
Diaphragmalgia 786.52
chest 786.59
anginoid (see also Pain, precordial) 786.51
chest (central) 786.50
atypical 786.59
midsternal 786.51
musculoskeletal 786.59
noncardiac 786.59
substernal 786.51
wall (anterior) 786.52
costochondral 786.52
diaphragm 786.52
heart (see also Pain, precordial) 786.51
intercostal 786.59
over heart (see also Pain, precordial) 786.51
pericardial (see also Pain, precordial) 786.51
pleura, pleural, pleuritic 786.52
precordial (region) 786.51
respiration 786.52
retrosternal 786.51
rib 786.50
substernal 786.51
respiration 786.52
Pleuralgia 786.52
Pleurodynia 786.52
Precordial pain 786.51
chest 786.59
Prinzmetal-Massumi syndrome (anterior chest wall) 786.52
painful 786.52

Lots of grouping!
ICD-10 is very detailed

V97.33XD: **Sucked into jet engine**, subsequent encounter

V00.15: **Heelies Accident**
Applicable To Rolling shoe, Wheeled shoe, Wheelies accident.

Supertypes:
V00-Y99  External causes of morbidity
V95-V97  Air and space transport accidents
V00  Pedestrian conveyance accident
V00.1  Rolling-type pedestrian conveyance accident
Time series medical records

Sequence of medical codes over time

Tasks:
- Predict probability of code in future
- Predict duration until next visit
- Find similar patients/codes/visits
MLPs on time series data

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>787.2</td>
<td>682.1</td>
</tr>
<tr>
<td>358.2</td>
<td></td>
</tr>
</tbody>
</table>

-5 years  +1 year

Multi-hot vectors!
Word2Vec for time series patient visits with ICD codes.

Embeddings learned for codes and demographics.
Baseline methods

**One-hot:** In order to compare with the raw input data, the binary vector $\mathbf{x}_t$ for the visit is used.

**Stacked autoencoder (SA):** Using the binary vector $\mathbf{x}_t$ concatenated with patient demographic information as the input the SA is trained to minimize the reconstruction error. Then will be used to generate visit representations.

**Sum of Skip-gram vectors (word2vec):** First learn the code-level representations with Skip-gram only. Then for the visit-level representation, add the representations of the codes within the visit.
Evaluation: Predicting codes of the next visit

\[
\text{top-k recall} = \frac{\text{# of true positives in the top } k \text{ predictions}}{\text{# of true positives}}
\]
Mr. Smith is a 63-year-old gentleman with coronary artery disease, hypertension, hypercholesterolemia, COPD and tobacco abuse. He reports doing well. He did have some more knee pain for a few weeks, but this has resolved. He is having more trouble with his sinuses. I had started him on Flonase back in December. He says this has not really helped. Over the past couple weeks he has had significant congestion and thick discharge. No fevers or headaches but does have diffuse upper right-sided teeth pain. **He denies any chest pains, palpitations, PND, orthopnea, edema or syncope.** His breathing is doing fine. No cough. He continues to smoke about half-a-pack per day. He plans on **trying the patches again.**

- Notes are still written together with codes selected for each visit.
- Often explaining details which do not have equivalent codes.
- Natural language is very difficult to parse with certainty.
Predicting codes from notes

<ICD-9 787.0
Nausea and vomiting>

Converting text to codes can

- Adapt old databases
- Correct errors
- Upgrade ICD versions

... with upset stomach
<done>
RNNs, Different types of sequential prediction tasks

<table>
<thead>
<tr>
<th>Output</th>
<th>State</th>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>one to one</td>
<td>one to many</td>
<td>many to one</td>
</tr>
<tr>
<td>many to many</td>
<td>many to many</td>
<td></td>
</tr>
</tbody>
</table>

Input:
- Cat
- "This is a cat"
- "It's hairy and I'm allergic to it"
- "Ceci est un chat"
- "This is a cat"
- "Meow Meow"

Taken from http://vision.stanford.edu/teaching/cs231n/slides/2016/winter1516_lecture10.pdf and Francis Dutil
An RNN applies a function over a sequence of inputs \([x_1, x_2, \ldots, x_T]\) which produces a sequence of outputs \([y_1, y_2, \ldots, y_T]\) and each input produces an internal state \([h_1, h_2, \ldots, h_T]\).
RNNs

- A simple RNN
  \[ h_t = \tanh(U x_t + W h_{t-1}) \]
  \[ y_t = f(V h_t) \]

- W, U and V are the parameters of the network.

- The weights are shared over time.
Unrolling the RNN over time

The weights are shared over time.

Image du blog de Christopher Olah, slide from l’École d’automne 2018, César Laurent
Unrolling the RNN over time

The weights are shared over time.
Unrolling the RNN over time

The weights are shared over time.
Predicting future events

Predicting next time step

A multi-hot vector
Vanishing gradients
The shade of gray shows the influence of the input of the RNN at time 1. It decreases over time, as the RNN gradually forgets its first input.

Issue addressed by:
- LSTM
- GRU
- Attention
- Recurrent batch norm
- Weight regularization
- Layer norm

More reading: [Pascanu 2013]
We can get creative with RNN designs

Stacked RNNs

Bi-directional RNN

Slide from l’École d’automne 2018, César Laurent
RNNs for next visit prediction (Doctor AI)

- Treat codes equally: ICD diagnosis codes, procedure codes, and medication codes
- Grouped codes into higher-order categories

\[ y = \text{High level medical codes } R^{1778} \]
\[ d = \text{time since last visit} \]

\[ y_i, \hat{y}_i, \hat{y}_{i+1}, \hat{d}_i, \hat{d}_{i+1}, \ldots, h_i^{(3)}, h_{i-1}^{(3)}, h_i^{(2)}, h_{i-1}^{(2)}, \ldots, x_i, x_{i-1}, x_i, x_i^{(1)}, d_1, \ldots \]

\[ R^{40000} \]

Medical codes

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Dx,Rx,Time Recall @k</th>
</tr>
</thead>
<tbody>
<tr>
<td>k = 10</td>
<td>26.25</td>
</tr>
<tr>
<td>k = 20</td>
<td>48.11 60.23 66.00</td>
</tr>
<tr>
<td>k = 30</td>
<td>53.61 64.93 71.14 0.2528</td>
</tr>
</tbody>
</table>

- Logistic: 36.04 46.32 52.53 0.0726
- MLP: 38.82 49.09 55.74 0.1221
- RNN-1: 53.86 65.10 71.24 0.2519
- RNN-2: 53.61 64.93 71.14 0.2528
- RNN-1-IR: 54.37 65.68 71.85 0.2492
- RNN-2-IR: 54.96 66.31 72.48 0.2534

Table 1: Basic statistics of the the clinical records dataset.

- Patients from Sutter Health Palo Alto Medical Foundation

Choi, Doctor AI: Predicting Clinical Event via Recurrent Neural Networks, 2016
Pretraining RNNs (Doctor AI)

MIMIC II has 2,695 patients with 2+ visits

Pretrained using a larger Sutter Health dataset

~300,000 patients
Related work


Che, Zhengping et al. “Recurrent Neural Networks for Multivariate Time Series with Missing Values.” Nature Scientific Reports. 2018
## Medical Natural Language Inference

<table>
<thead>
<tr>
<th>#</th>
<th>Premise</th>
<th>Hypothesis</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALT, AST, and lactate were elevated as noted above</td>
<td>patient has abnormal Lfts</td>
<td>entailment</td>
</tr>
<tr>
<td>2</td>
<td>Chest x-ray showed mild congestive heart failure</td>
<td>The patient complains of cough</td>
<td>neutral</td>
</tr>
<tr>
<td>3</td>
<td>During hospitalization, patient became progressively more dyspnic requiring BiPAP and then a NRB</td>
<td>The patient is on room air</td>
<td>contradiction</td>
</tr>
<tr>
<td>4</td>
<td>She was not able to speak, but appeared to comprehend well</td>
<td>Patient had aphasia</td>
<td>entailment</td>
</tr>
<tr>
<td>5</td>
<td>T1DM: x 7yrs, h/o DKA x 6 attributed to poor medication compliance, last A1c [** 3-23 **]: 13.3 % 2</td>
<td>The patient maintains strict glucose control</td>
<td>contradiction</td>
</tr>
<tr>
<td>6</td>
<td>Had an ultimately negative esophagogastroduodenoscopy and colonoscopy</td>
<td>Patient has no pain</td>
<td>neutral</td>
</tr>
<tr>
<td>7</td>
<td>Aorta is mildly tortuous and calcified</td>
<td>the aorta is normal</td>
<td>contradiction</td>
</tr>
</tbody>
</table>

MedNLI dataset derived from the MIMIC-III dataset
# Medical Natural Language Inference

Studying the errors we can see the limits of the model.

<table>
<thead>
<tr>
<th>Category</th>
<th>Premise</th>
<th>Hypothesis</th>
<th>Predicted</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical</td>
<td>On weaning to 6LNC, his O2 decreased to 81-82%</td>
<td>He has poor O2 stats</td>
<td>neutral</td>
<td>entailment</td>
</tr>
<tr>
<td>Reasoning</td>
<td>WBC 12, Hct 41.</td>
<td>WBC slightly elevated</td>
<td>contradiction</td>
<td>entailment</td>
</tr>
<tr>
<td>World</td>
<td>The infant emerged with spontaneous cry. No known sick contacts</td>
<td>The infant was still born.</td>
<td>entailment</td>
<td>contradiction</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Received GI cocktail for h/o GERD, esophageal spasm</td>
<td>Patient has no angina</td>
<td>neutral</td>
<td>entailment</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>No CP or fevers.</td>
<td>Received a proton pump inhibitor</td>
<td>neutral</td>
<td>entailment</td>
</tr>
<tr>
<td>Medical</td>
<td>EKG showed T-wave depression in V3-5, with no prior EKG for comparison.</td>
<td>Patient has a normal EKG</td>
<td>neutral</td>
<td>contradiction</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Mother developed separation of symphysis pubis and was put in traction.</td>
<td>She has orthopedic injuries</td>
<td>neutral</td>
<td>entailment</td>
</tr>
<tr>
<td>Negation</td>
<td>Head CT was negative for bleed. Denied headache, sinus tenderness, or congestion</td>
<td>The patient has intracranial hemorrhage</td>
<td>neutral</td>
<td>contradiction</td>
</tr>
</tbody>
</table>

Table 9: Representative errors made by different models
"While code-based representations of clinical concepts and patient encounters are a tractable first step towards working with heterogeneous EHR data, they ignore many important real-valued measurements associated with items such as laboratory tests, intravenous medication infusions, vital signs, and more." [Shickel et al., 2018]
"While some researchers downplay the importance of interpretability in favor of significant improvements in model performance, we feel advances in deep learning transparency will hasten the widespread adoption of such methods in clinical practice." [Shickel et al, 2018]
"Many studies claim state-of-the-art results, but few can be verified by external parties. **This is a barrier** for future model development and one cause of the slow pace of advancement."

[Shickel et al, 2018]
Ching, T., et al. **Opportunities And Obstacles For Deep Learning In Biology And Medicine.** Journal of The Royal Society Interface. 2018

Shickel, B. et al. **Deep EHR : A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record.** IEEE Journal of Biomedical and Health Informatics, 2018