Attention mechanism

Jimeng Sun

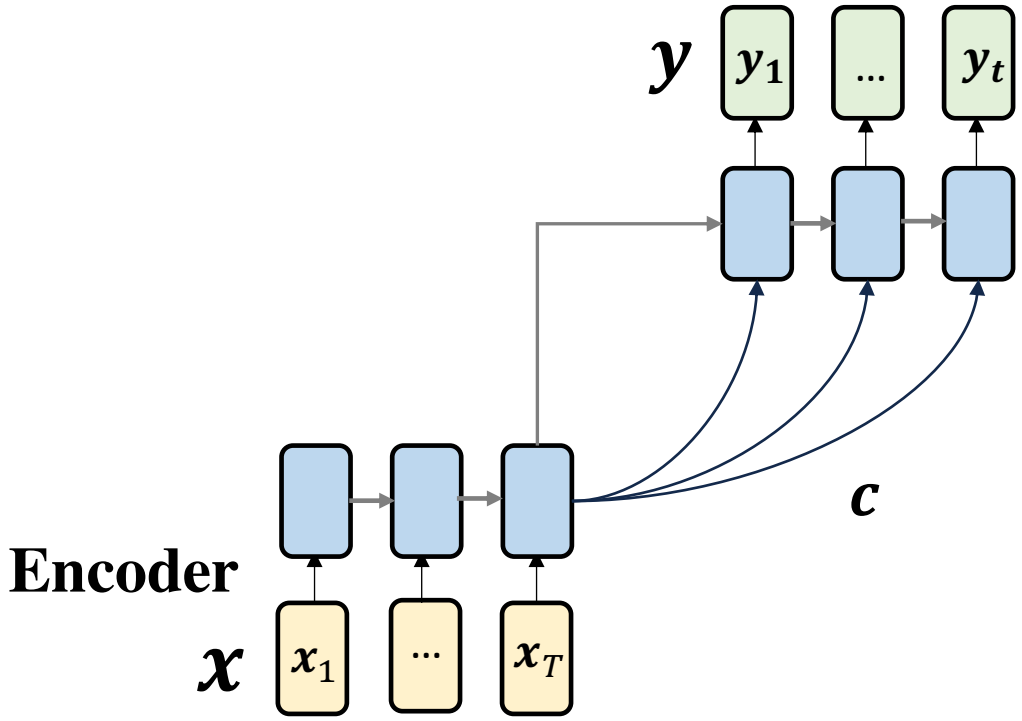
ILLINOIS

Outline

- Attention model
- Healthcare applications
  - Attention model over longitudinal EHR data (RETAIN)
  - Attention model over medical ontology (GRAM)
  - Attention model over clinical text (CAML)
  - Attention model electrocardiography (MINA)

Review: Encoder-Decoder Sequence-to-Sequence Model

$$P_\theta(\mathcal{Y}|\mathcal{X}) = \prod_{j=1}^{J+1} P_\theta(\boldsymbol{y}_j|\boldsymbol{y}_{j-1}, \cdots, \boldsymbol{y}_1, \mathcal{X})$$



Decoder

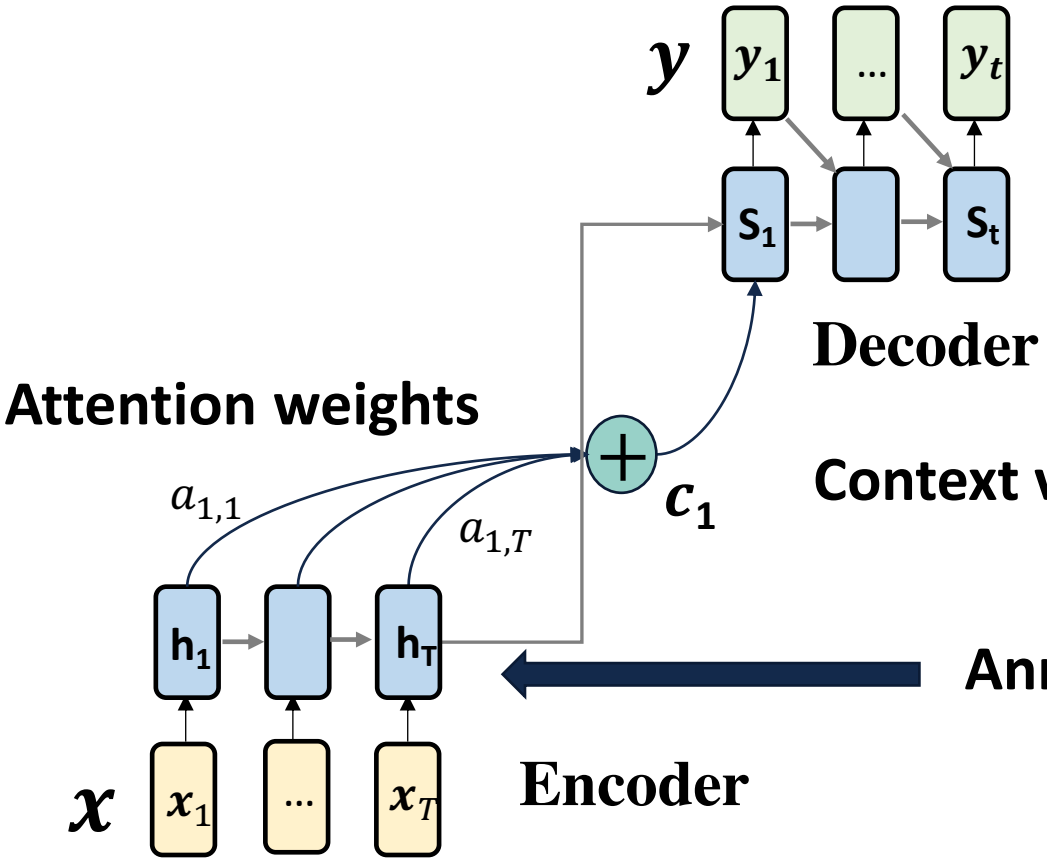
$$P(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \cdots, \boldsymbol{y}_1, \boldsymbol{c}) = g(\boldsymbol{h}_t, \boldsymbol{y}_{t-1}, \boldsymbol{c})$$

$$\boldsymbol{c} = \boldsymbol{h}_T = f(\boldsymbol{h}_{T-1}, \boldsymbol{x}_T)$$

Static context vector has limited capacity in modeling long sequences

ILLINOIS

Attention on RNN Model



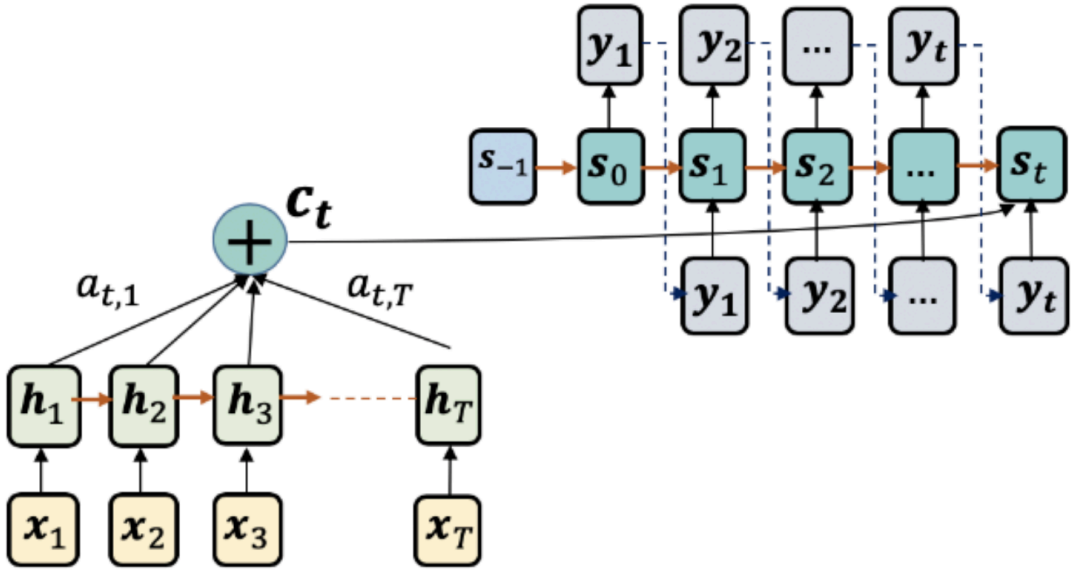
Another RNN take

Context vector becomes dynamic for each output

Annotation vectors: encode input sequence

ILLINOIS

Alignment model



- To quantify the alignment between input and output

$$e_{ij} = a(\boldsymbol{s}_{i-1}, \boldsymbol{h}_j)$$

- Dot-product attention

$$a(\boldsymbol{s}_{i-1}, \boldsymbol{h}_j) = \boldsymbol{s}_{i-1}^\top \boldsymbol{h}_j$$

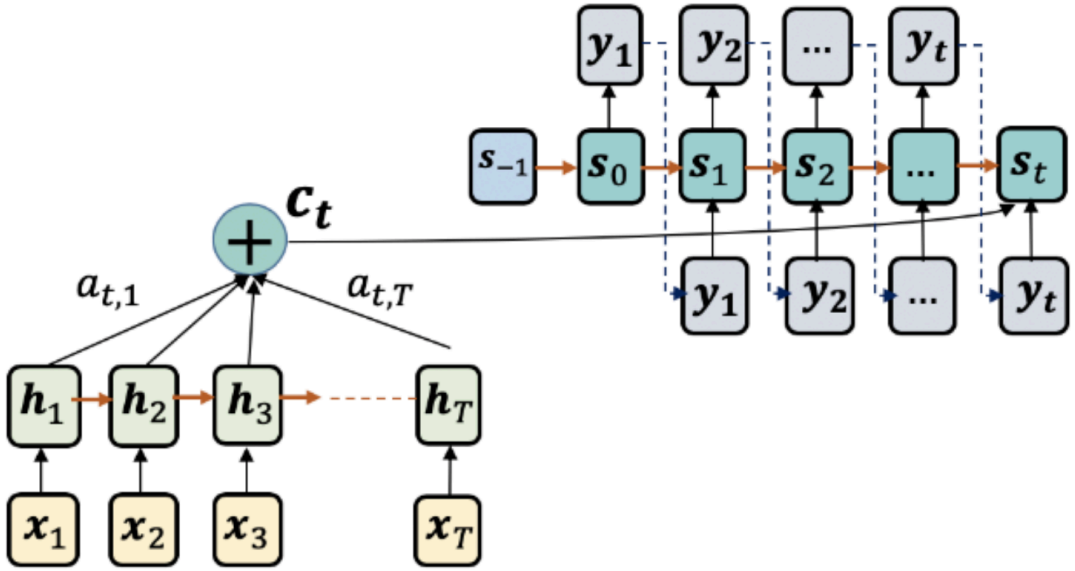
- General dot-product attention

$$a(\boldsymbol{s}_{i-1}, \boldsymbol{h}_j) = \boldsymbol{s}_{i-1}^\top \boldsymbol{W}_a \boldsymbol{h}_j$$

- Additive attention

$$a(\boldsymbol{s}_{i-1}, \boldsymbol{h}_j) = \boldsymbol{v}_a^\top \tanh(\boldsymbol{W}_a[\boldsymbol{s}_{i-1}, \boldsymbol{h}_j] + \boldsymbol{b}_a)$$

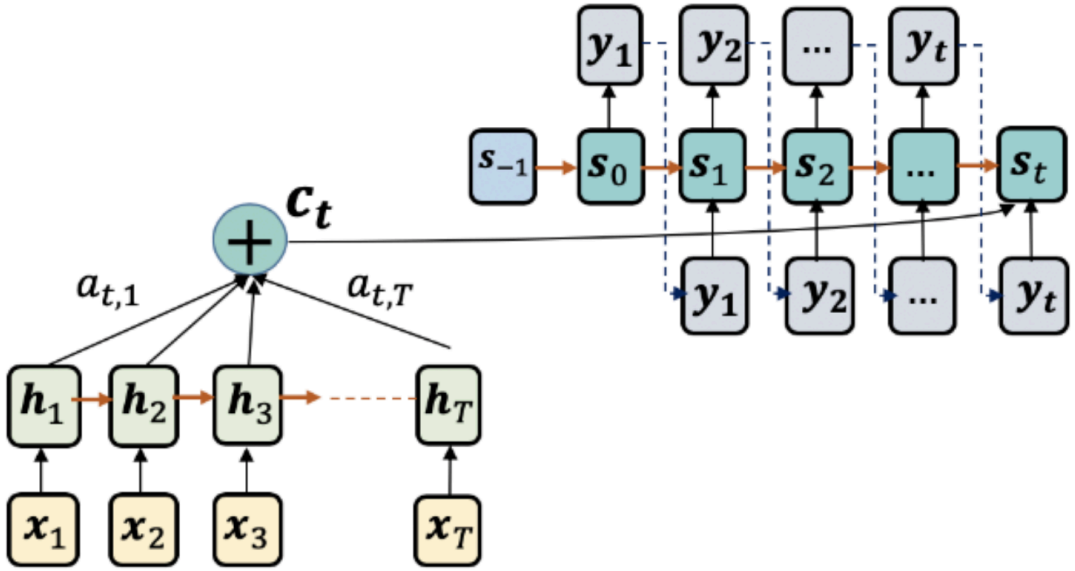
Attention weights



- Attention weight is the normalized version of the alignment model

$$a_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{T} \exp e_{ik}}$$

Decoder



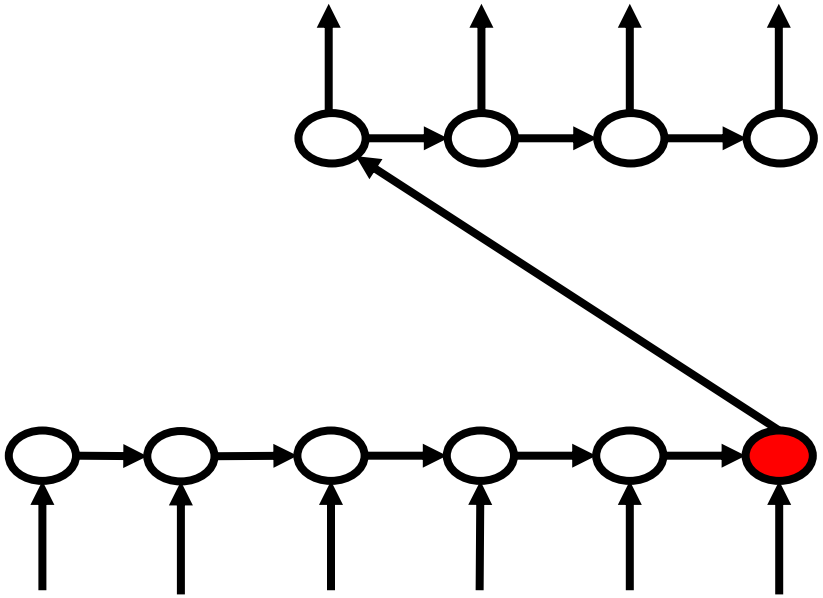
- Another RNN

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

- Context vector $\mathbf{c}_i$ is dynamic for each output

Regular Machine Translation

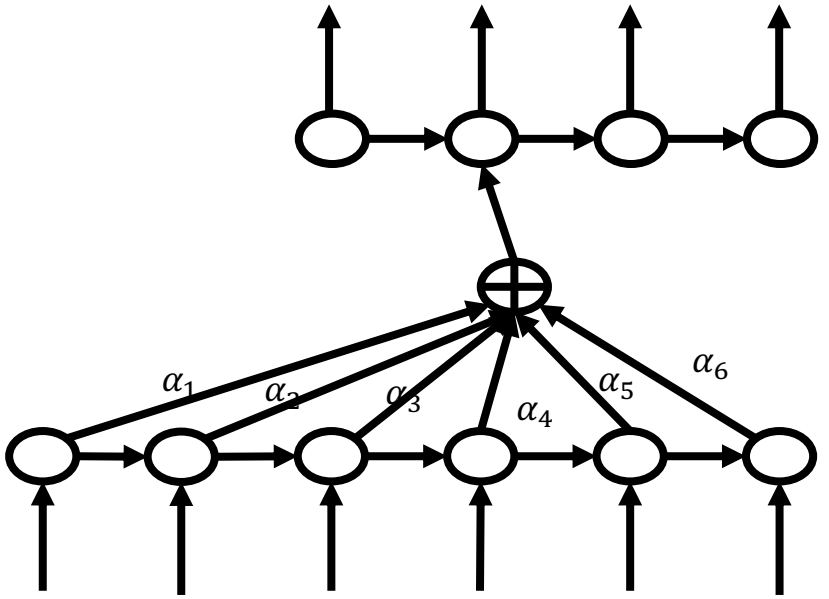
如果你不在乎谁获得了荣誉，
你所能完成的事情是**惊人的**。



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: Interpretable Deep learning model

How to provide interpretability of model

Edward Choi Taha Bahadori Andy Schuetz Buzz Stewart Jimeng Sun

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

ILLINOIS

Interpretable Models

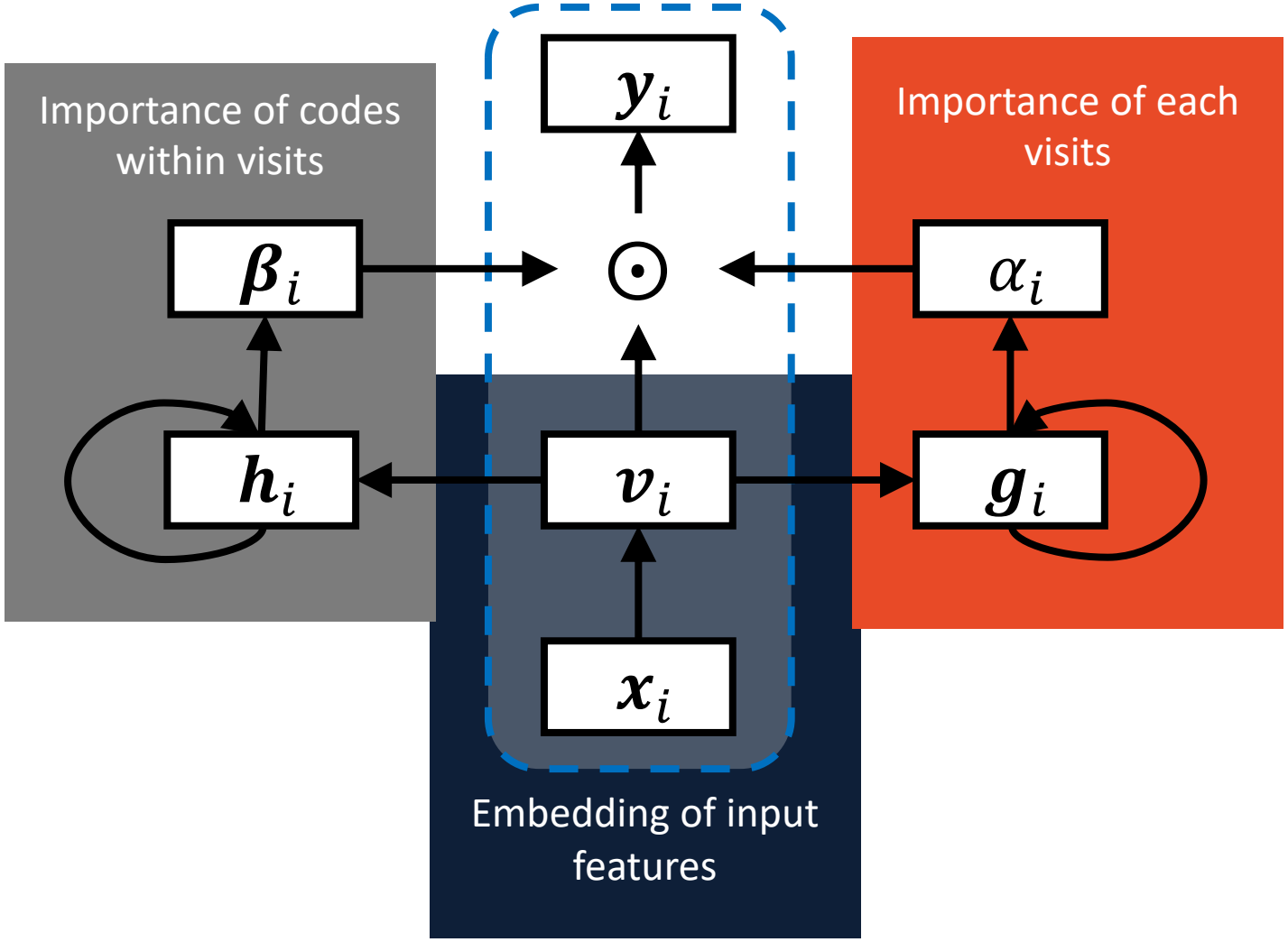
What do we mean by interpretability of a model?

Three categories of models:

- Rule based: *e.g.* decision trees
- Case based: *e.g.* nearest neighbor methods
- Risk factor based: *e.g.* sparse linear regression

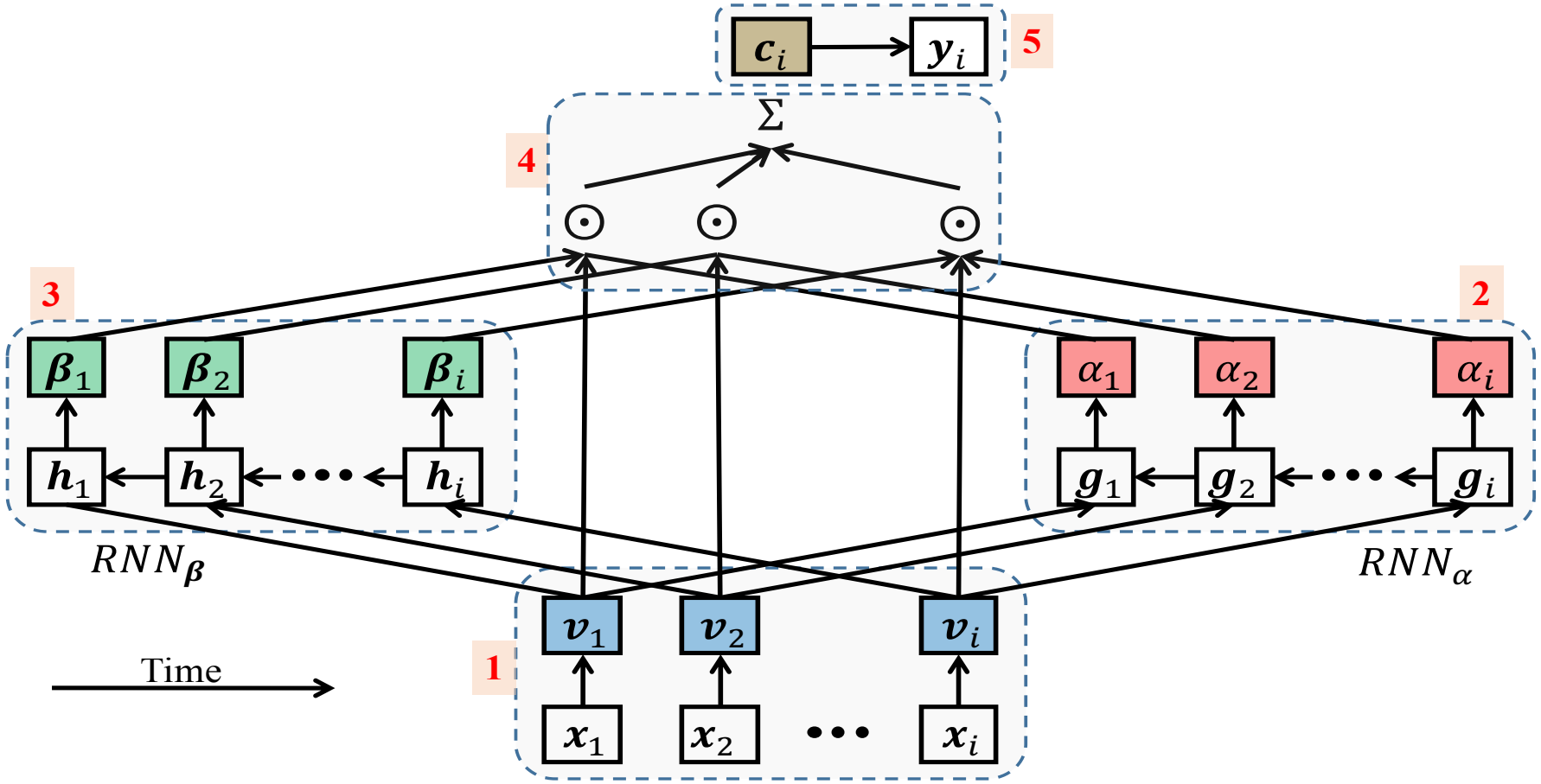
- Temporal models? Latent variable models

RETAIN: REverse Time AttentIoN model

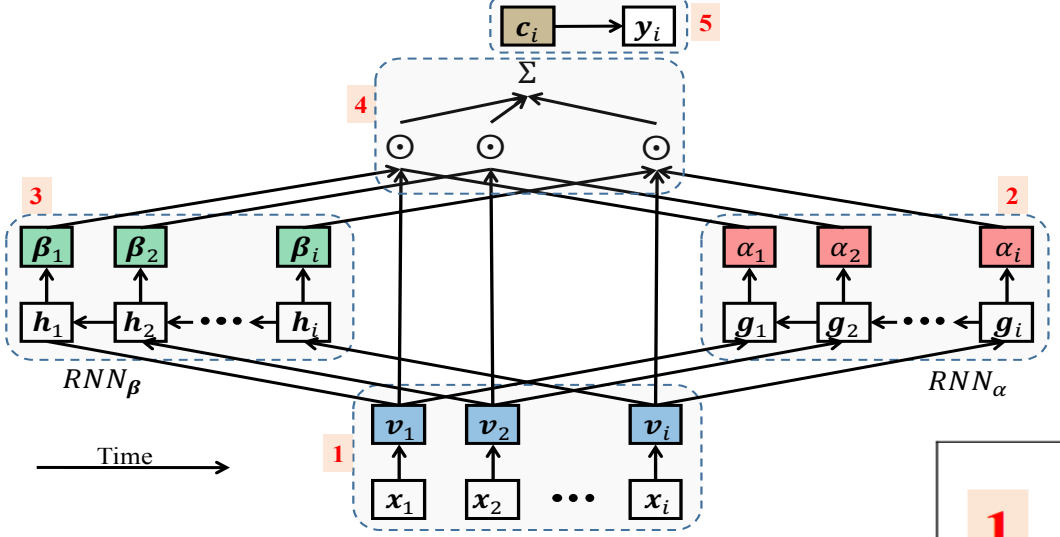


Choi, Edward, et al. 2016. "RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism." In *NIPS*

Details of RETAIN



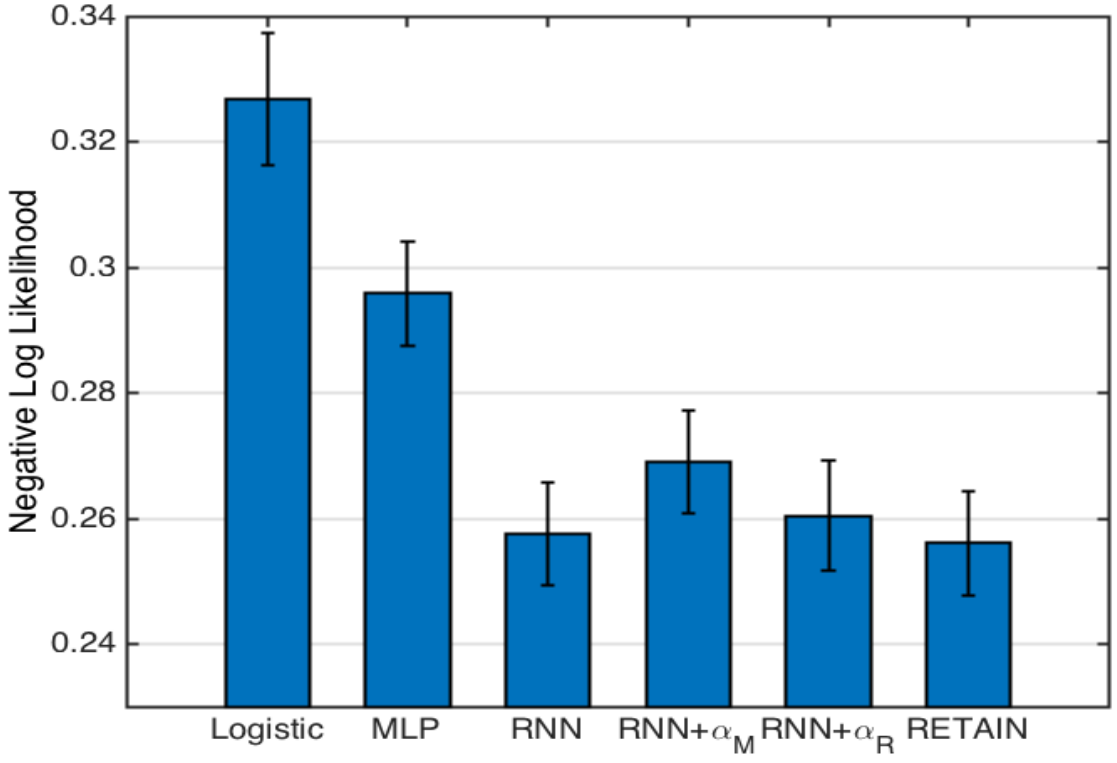
RETAIN Algorithm



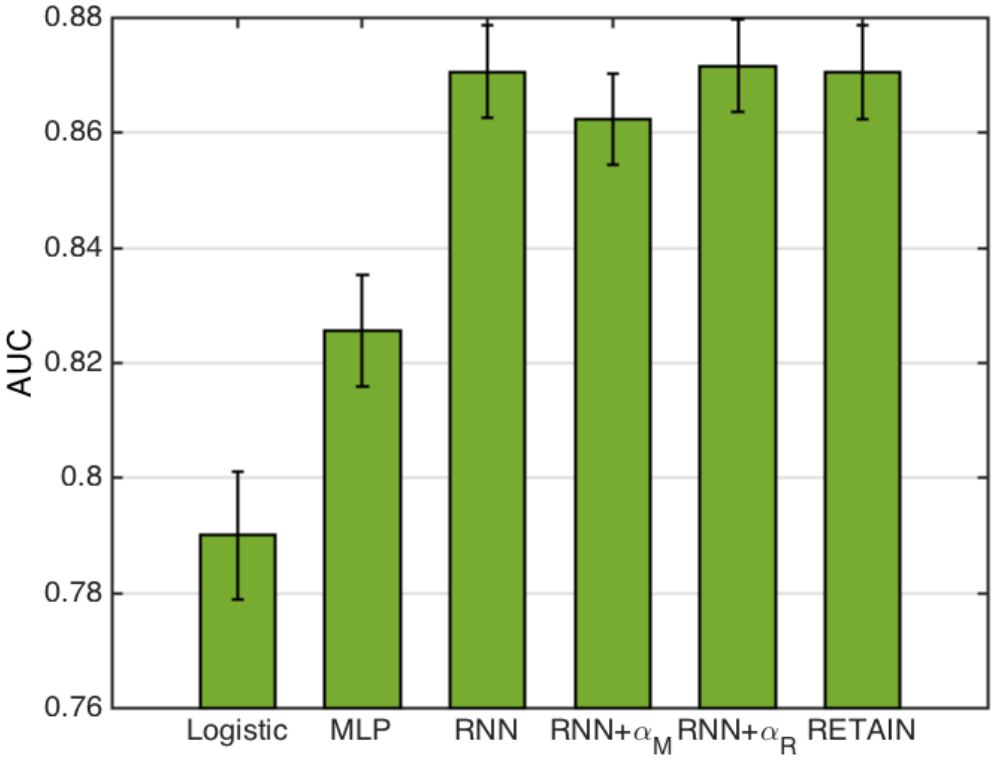
1	$\mathbf{v}_i = \mathbf{E}\mathbf{x}_i$
	Multi-hot representation of the visit is linearly projected by the embedding matrix $\mathbf{E}$.
2	$\mathbf{g}_i, \mathbf{g}_{i-1}, \ldots, \mathbf{g}_1 = \mathrm{RNN}_\alpha(\mathbf{v}_i, \mathbf{v}_{i-1}, \ldots, \mathbf{v}_1)$, $\alpha_1, \alpha_2, \ldots, \alpha_i = \mathrm{Softmax}(\mathbf{w}_\alpha^\top [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_i] + b_\alpha)$
	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	$\mathbf{h}_i, \mathbf{h}_{i-1}, \ldots, \mathbf{h}_1 = \mathrm{RNN}_\beta(\mathbf{v}_i, \mathbf{v}_{i-1}, \ldots, \mathbf{v}_1)$ $\boldsymbol{\beta}_j = \tanh\left(\mathbf{W}_\beta \mathbf{h}_j + \mathbf{b}_\beta\right) \quad \text{for} \quad j = 1, \ldots, i$
	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	$\hat{\mathbf{y}}_i = \mathrm{Softmax}(\mathbf{W}\mathbf{c}_i + \mathbf{b})$
	Using the context vector $\boldsymbol{c}_i$, we make the final prediction.

Heart Failure Results

- Negative Log Likelihood on Test Set

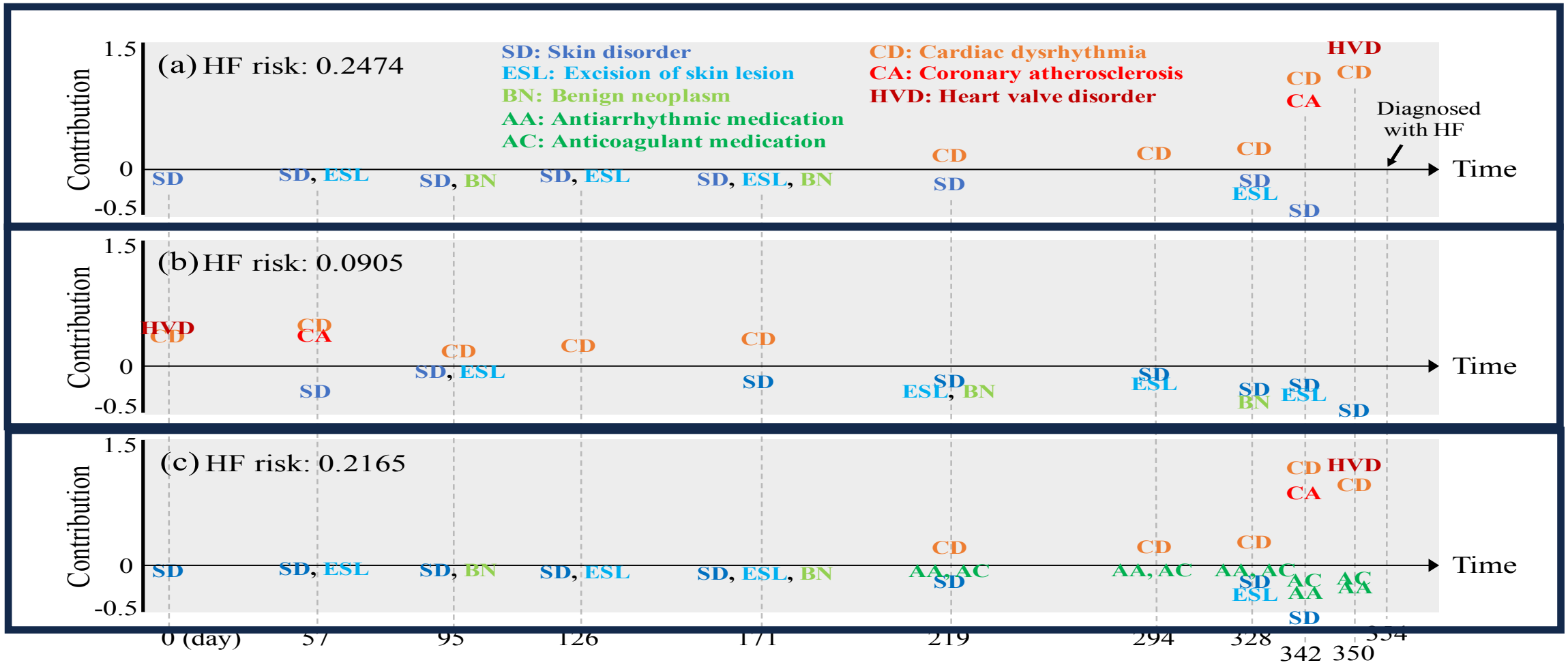


- Classification AUC



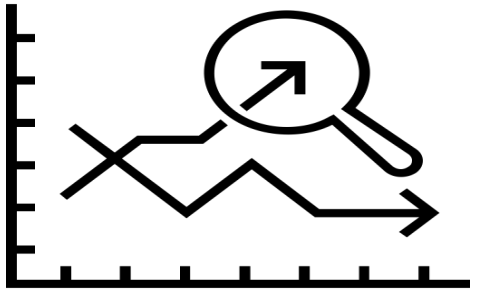
ILLINOIS

Interpretation of RETAIN model



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*

GRAM: Graph-based Attention Model for Healthcare Representation Learning

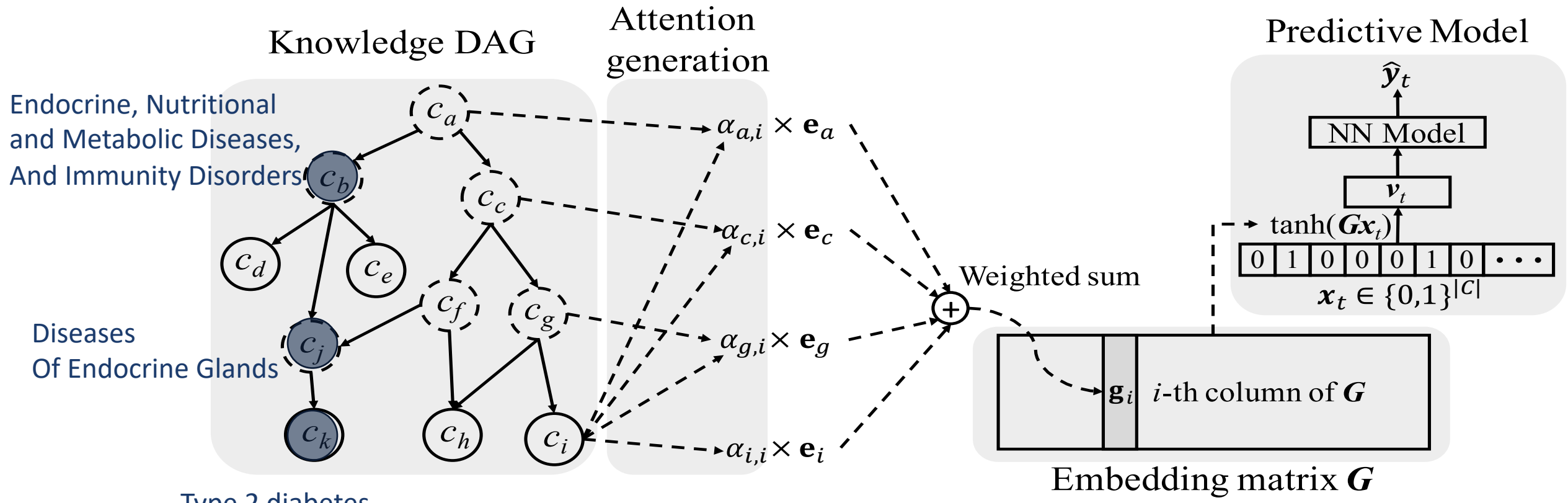
Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, Jimeng Sun

KDD' 17

ILLINOIS

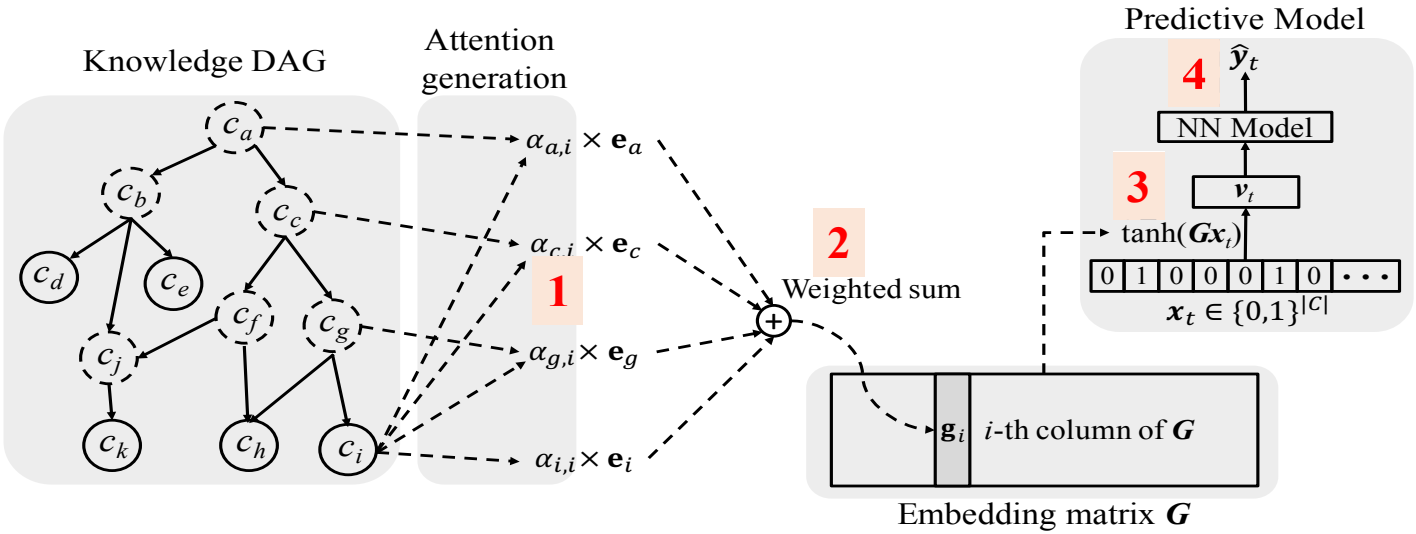
GRAM: Learn representations of medical codes leveraging medical ontologies

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



ILLINOIS

GRAM: Algorithm



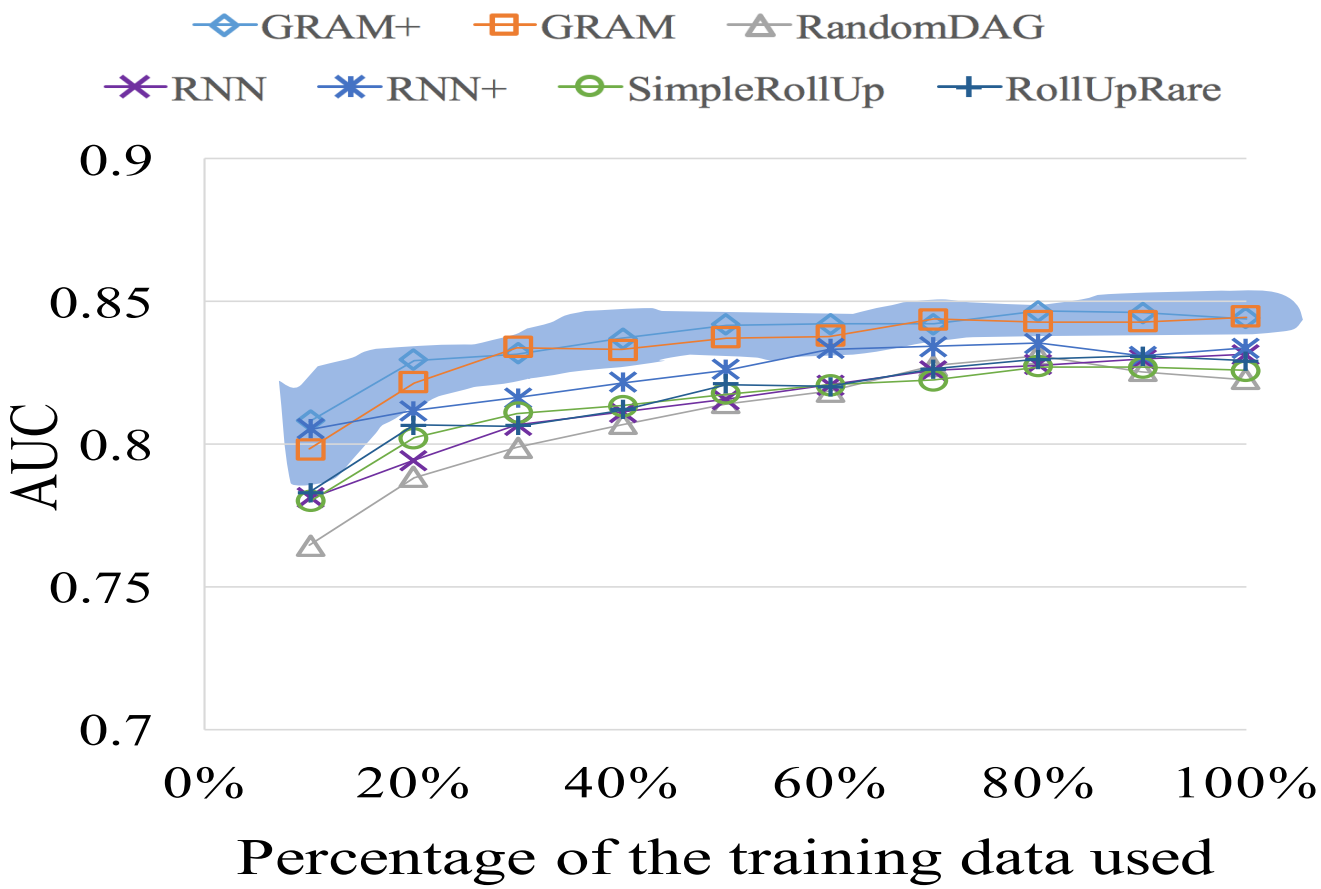
1	$\alpha_{ij} = \frac{\exp(f(\mathbf{e}_i, \mathbf{e}_j))}{\sum_{k \in \mathcal{A}(i)} \exp(f(\mathbf{e}_i, \mathbf{e}_k))}$ where $f(\mathbf{e}_i, \mathbf{e}_j) = \mathbf{u}_a^\top \tanh(\mathbf{W}_a \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_j \end{bmatrix} + \mathbf{b}_a)$
	Attention weights are generated for all pairs of basic embeddings and its ancestors .
2	$\mathbf{g}_i = \sum_{j \in \mathcal{A}(i)} \alpha_{ij} \mathbf{e}_j,$
	Final representation 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 ***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.

ILLINOIS

GRAM provides accurate prediction

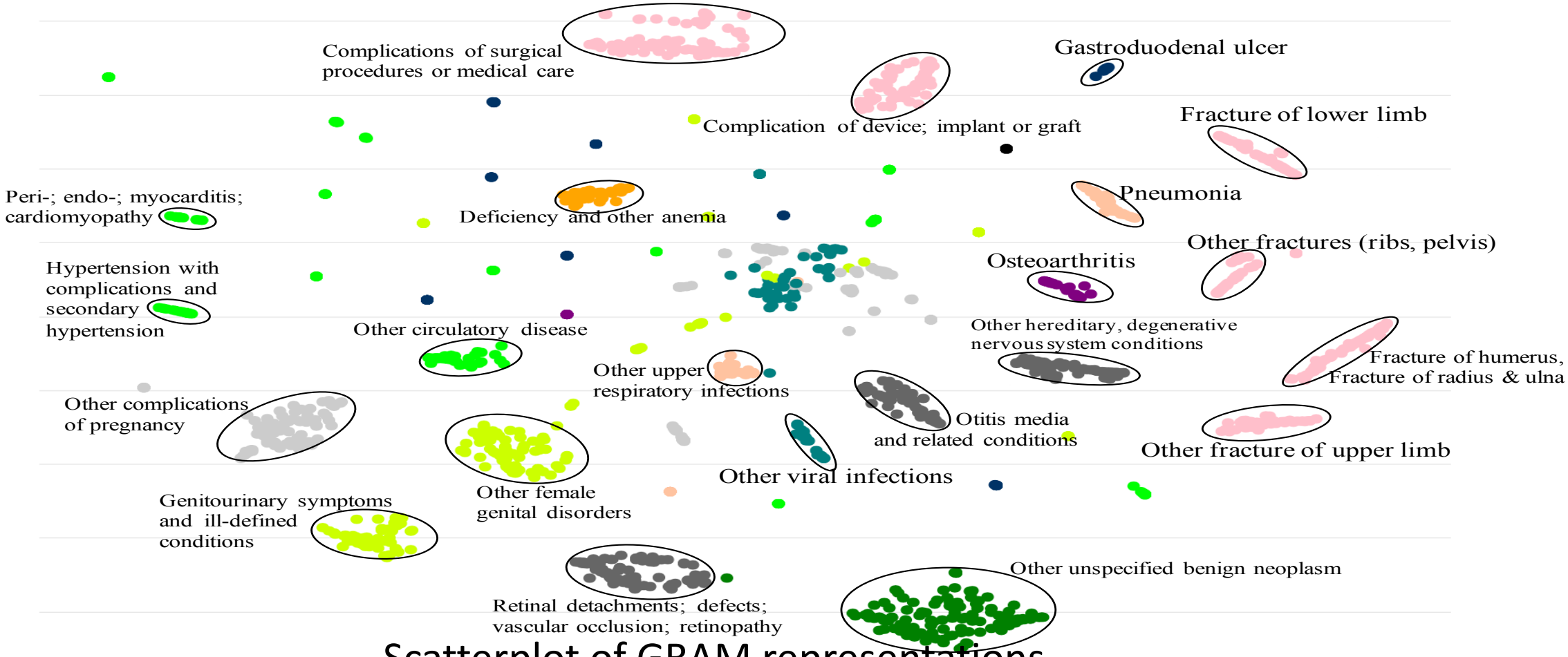
GRAM shows better predictive performance under data constraints

HF prediction using varying sizes of training data



ILLINOIS

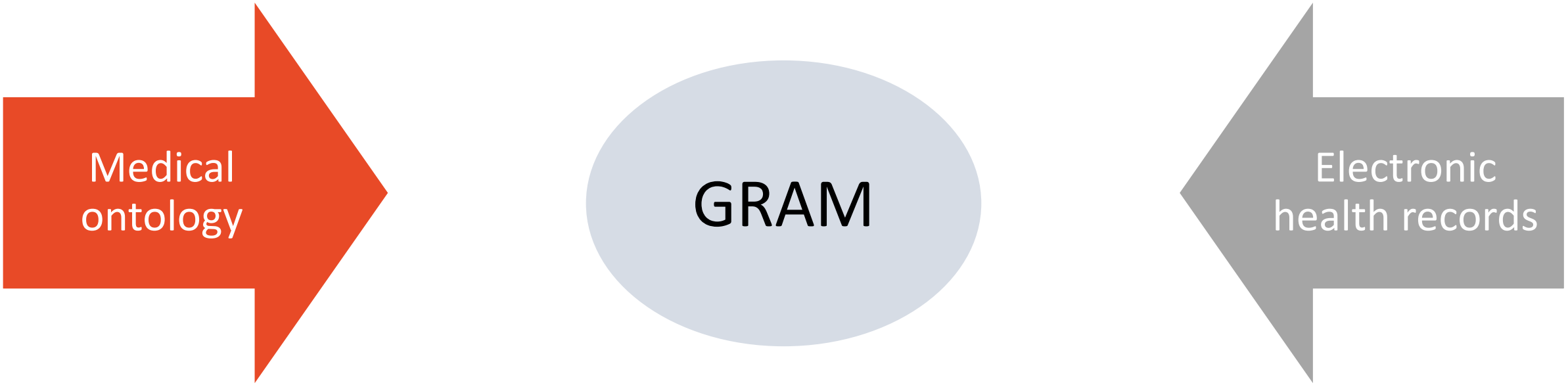
GRAM learns representations well aligned with knowledge ontology



Scatterplot of GRAM representations

ILLINOIS

GRAM: Summary



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

Choi, Edward, et al."GRAM: Graph-based Attention Model for Healthcare Representation Learning." KDD, 2017.

ILLINOIS

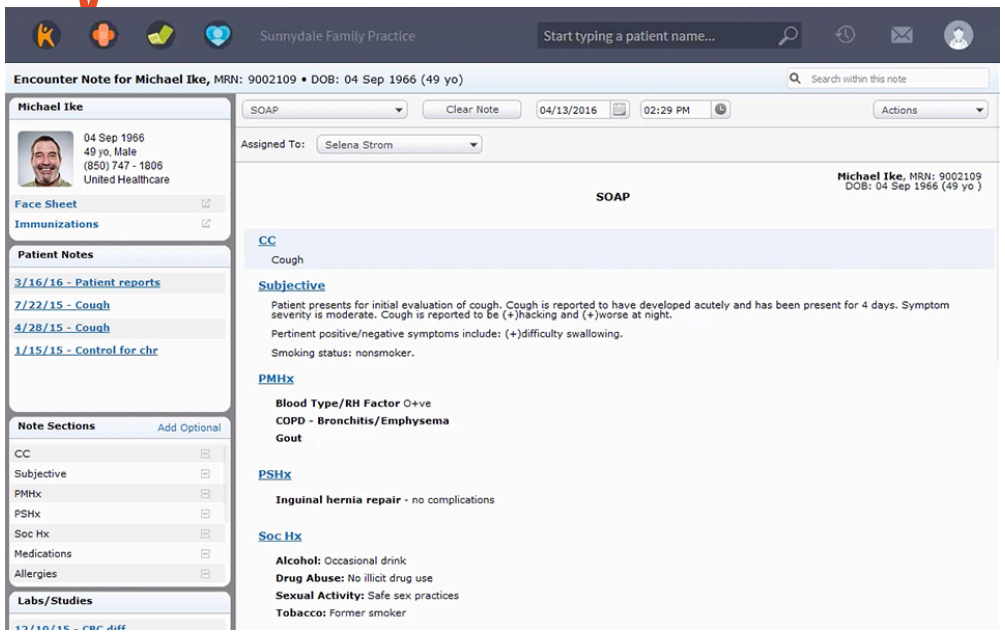
Explainable Prediction of Medical Codes from Clinical Text

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jacob Eisenstein

NAACL'18

ILLINOIS

Clinical Coding Problem



Electronic Health Records contain rich clinical texts

Features	ICD-9
Possible Codes	14,000
Characters	3–5

ICD: taxonomy of diagnoses & procedures

Human coding laborious, error-prone [Birman-Deych et al., 2005]

Clinical Coding Problem, as a CS researcher

- Highly multi-label classification
  - 14K ICD9, 68K ICD10 labels

- Testbed for document representations

- Documents are long and loosely structured

Prior approaches

- Predict a subset of labels
- Sub-domain focus, e.g. radiology
- Private datasets

This work

- Predict all labels
- General ICU setting
- Open-access data

The dataset

- Open-access, de-identified
- 47k admissions -> 47k documents for training

https://mimic.physionet.org/

The dataset

MIMIC

- Open-access, de-identified
- 47k admissions -> 47k documents for training
- **Loosely structured:**

```
Admission Date:  [**2118-6-2**]       Discharge Date:  [**2118-6-14**]

Date of Birth:                      Sex:  F

Service:  MICU and then to [**Doctor Last Name **] Medicine

HISTORY OF PRESENT ILLNESS:  This is an 81-year-old female
with a history of emphysema (not on home O2), who presents...
```

ILLINOIS

The dataset

- Open-access, de-identified
- 47k admissions -> 47k documents for training
- **Loosely structured:**

```
Admission Date:  [**2118-6-2**]       Discharge Date:  [**2118-6-14**]

Date of Birth:                      Sex:  F

Service:  MICU and then to [**Doctor Last Name **] Medicine

HISTORY OF PRESENT ILLNESS:  This is an 81-year-old female
with a history of emphysema (not on home O2), who presents...
```

Long: Median post-processed document length: 1,341

The dataset

- Open-access, de-identified
- 47k admissions -> 47k documents for training
- **Loosely structured:**

```
Admission Date:  [**2118-6-2**]       Discharge Date:  [**2118-6-14**]

Date of Birth:                    Sex:  F

Service:  MICU and then to [**Doctor Last Name **] Medicine

HISTORY OF PRESENT ILLNESS:  This is an 81-year-old female
with a history of emphysema (not on home O2), who presents...
```

Many labels:

```
519.1: 'Other disease…'
491.21: 'Obstructive …'
518.81: 'Acute respir…'
486: 'Pneumonia, orga…'
276.1: 'Hyposmolality…'
244.9: 'Unspecified h…'
31.99: 'Other operati…'
.
.
.
```

Long: Median post-processed document length: 1,341

ILLINOIS

The dataset

MIMIC

- Open-access, de-identified
- 47k admissions -> 47k documents for training
- **Loosely structured:**

Median # labels: 14

Many labels:

```
Admission Date:  [**2118-6-2**]       Discharge Date:  [**2118-6-14**]

Date of Birth:                    Sex:  F

Service:  MICU and then to [**Doctor Last Name **] Medicine

HISTORY OF PRESENT ILLNESS:  This is an 81-year-old female
with a history of emphysema (not on home O2), who presents...
```

```
519.1: 'Other disease…'
491.21: 'Obstructive …'
518.81: 'Acute respir…'
486: 'Pneumonia, orga…'
276.1: 'Hyposmolality…'
244.9: 'Unspecified h…'
31.99: 'Other operati…'
.
.
.
```

Long: Median post-processed document length: 1,341

ILLINOIS

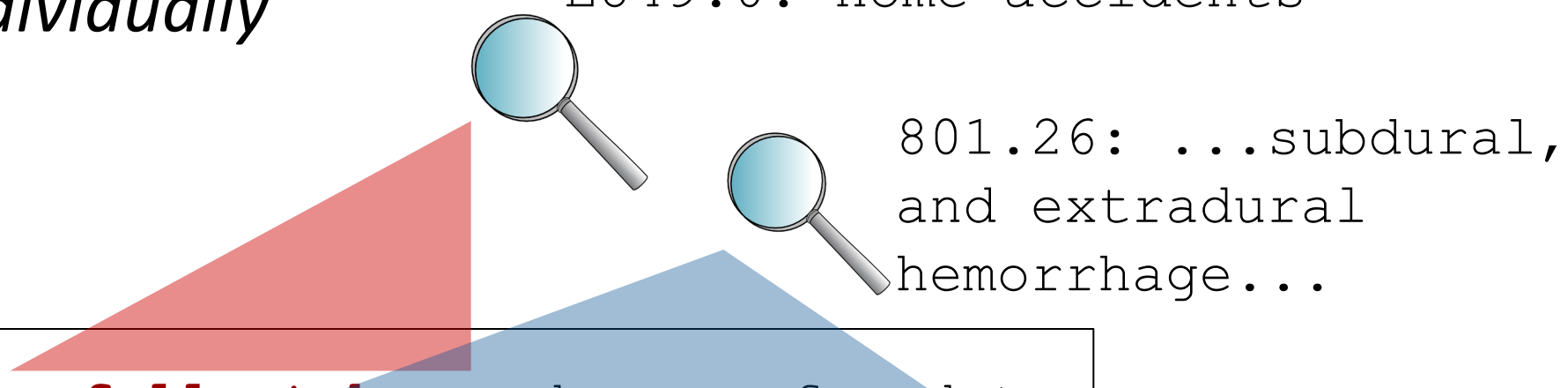
Modeling consideration

- Focus on the parts that matter

...who sustained a fall at home she was found to have a large acute on **chronic subdural hematoma** with extensive midline shift...

Modeling consideration

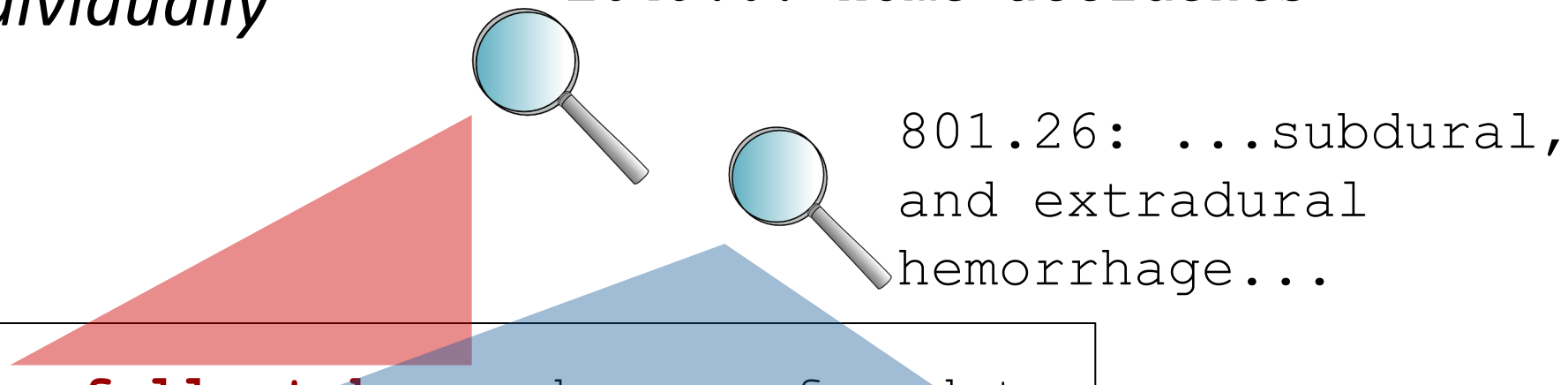
- Focus on the parts that matter
- Treat labels *individually*



...who sustained **a fall at home** she was found to have a large acute on **chronic subdural hematoma** with extensive midline shift...

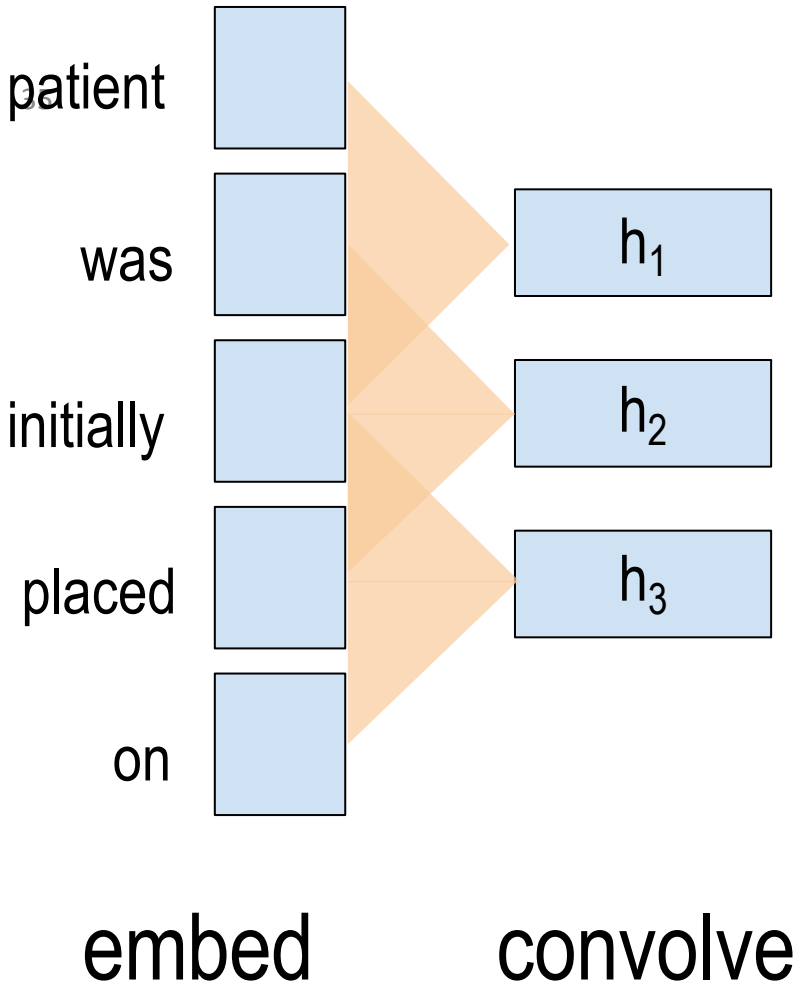
Modeling consideration

- Focus on the parts that matter
- Treat labels *individually*
- Be fast!

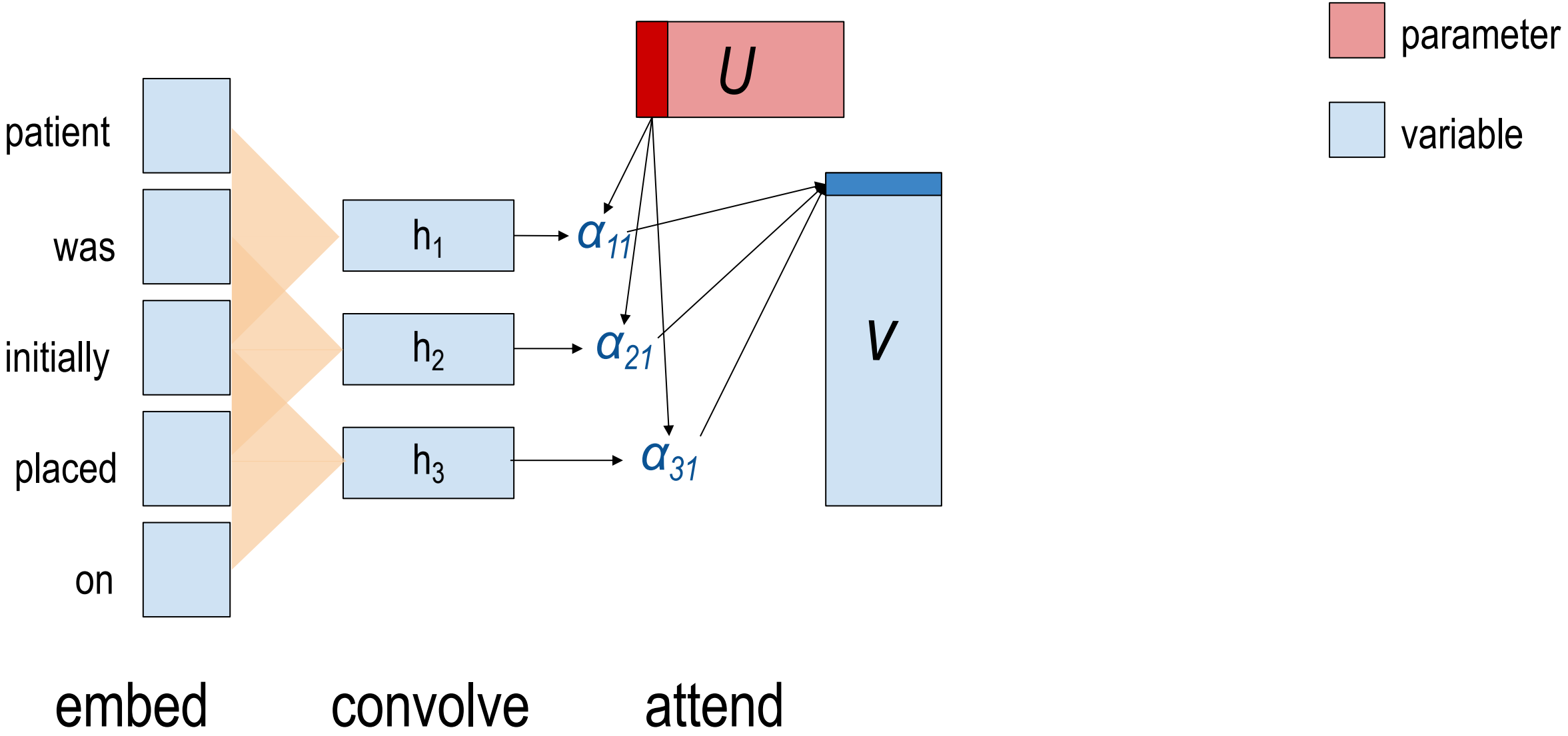


...who sustained **a fall at home** she was found to have a large acute on **chronic subdural hematoma** with extensive midline shift...

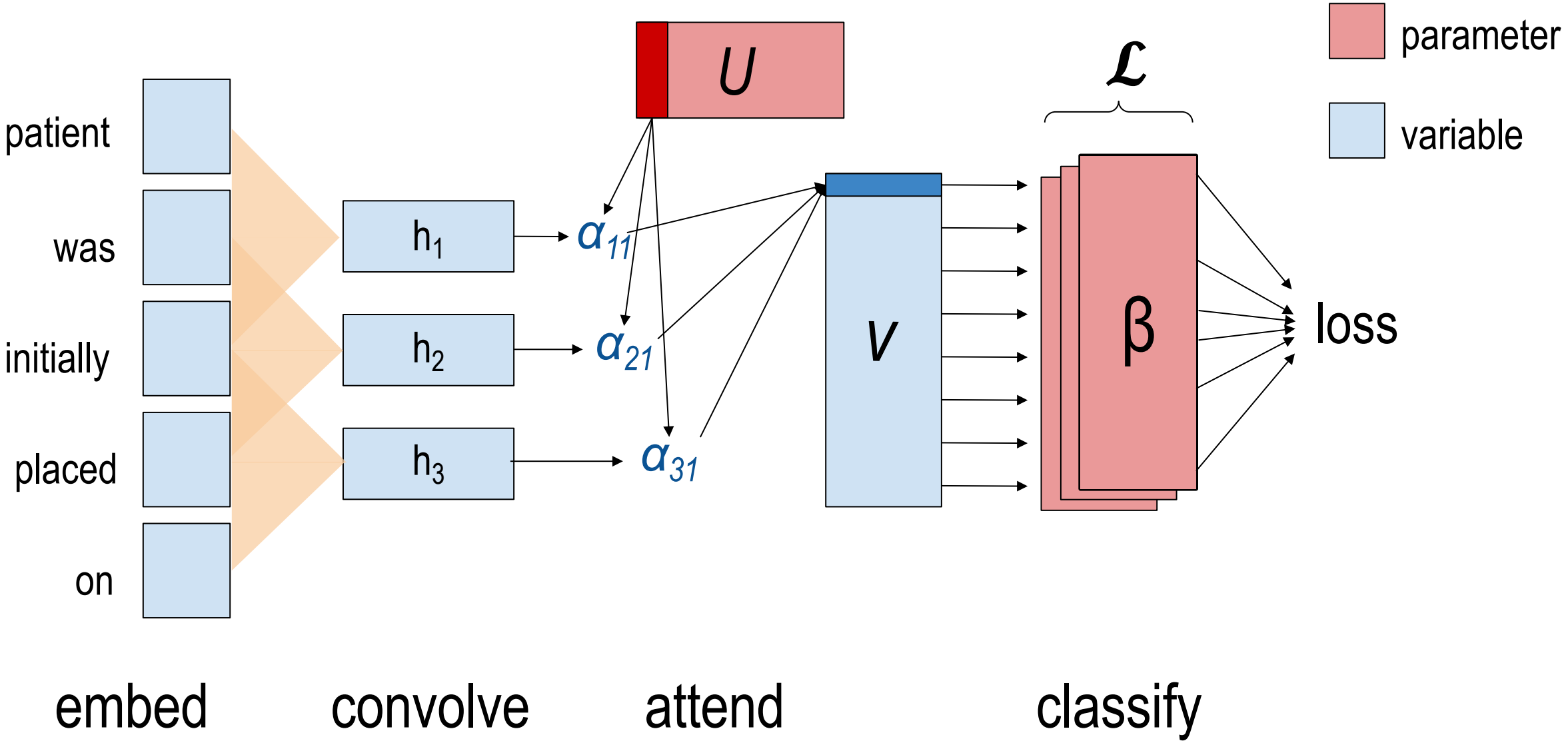
Convolution



Convolution + Attention



Convolution + Attention = CAML model



Dealing with the long tail

- Huge label space (nearly 9,000 total)

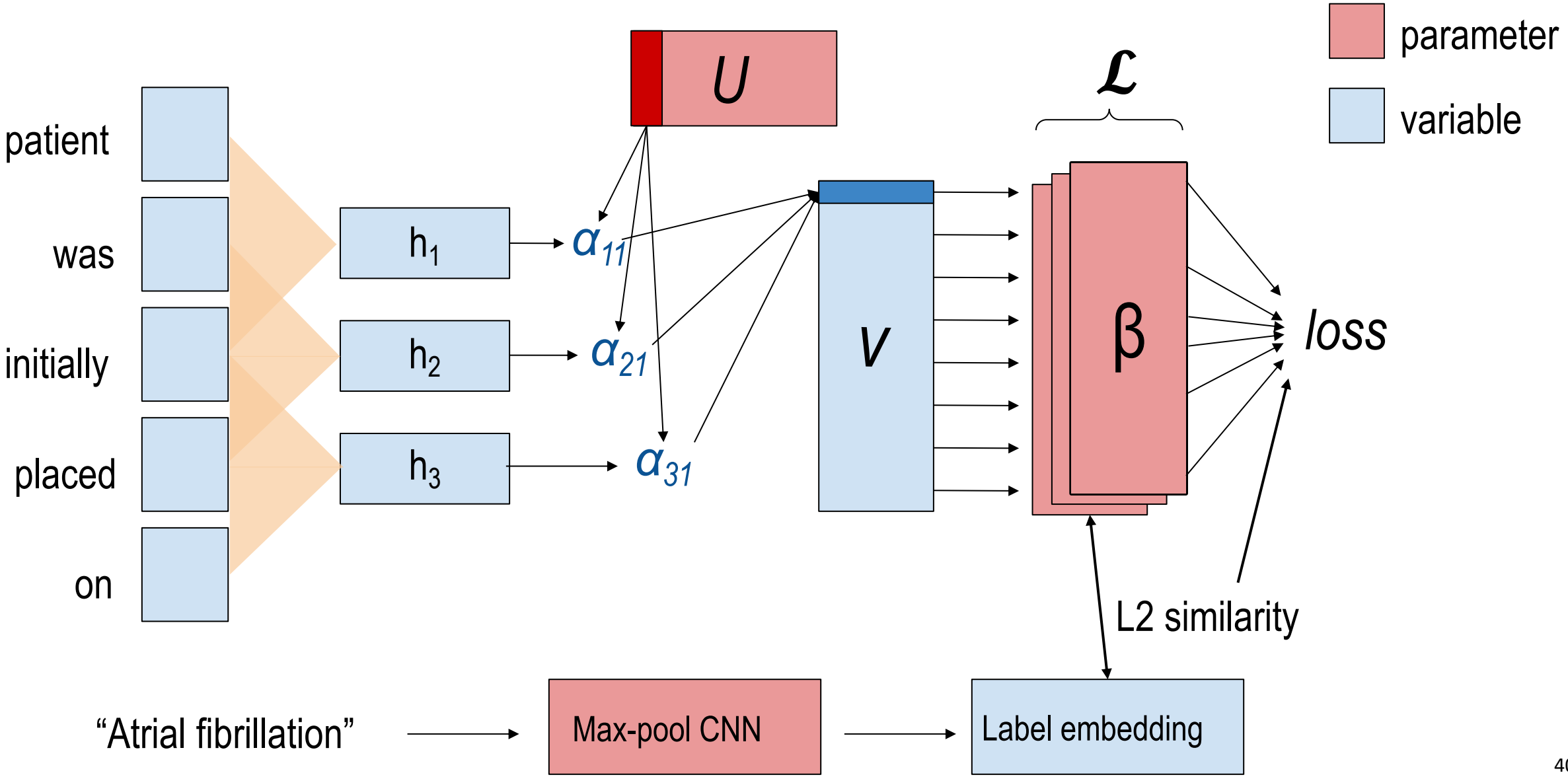
Dealing with the long tail

- Huge label space (nearly 9,000 total)
- Many labels are similar

250.00: “Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled”

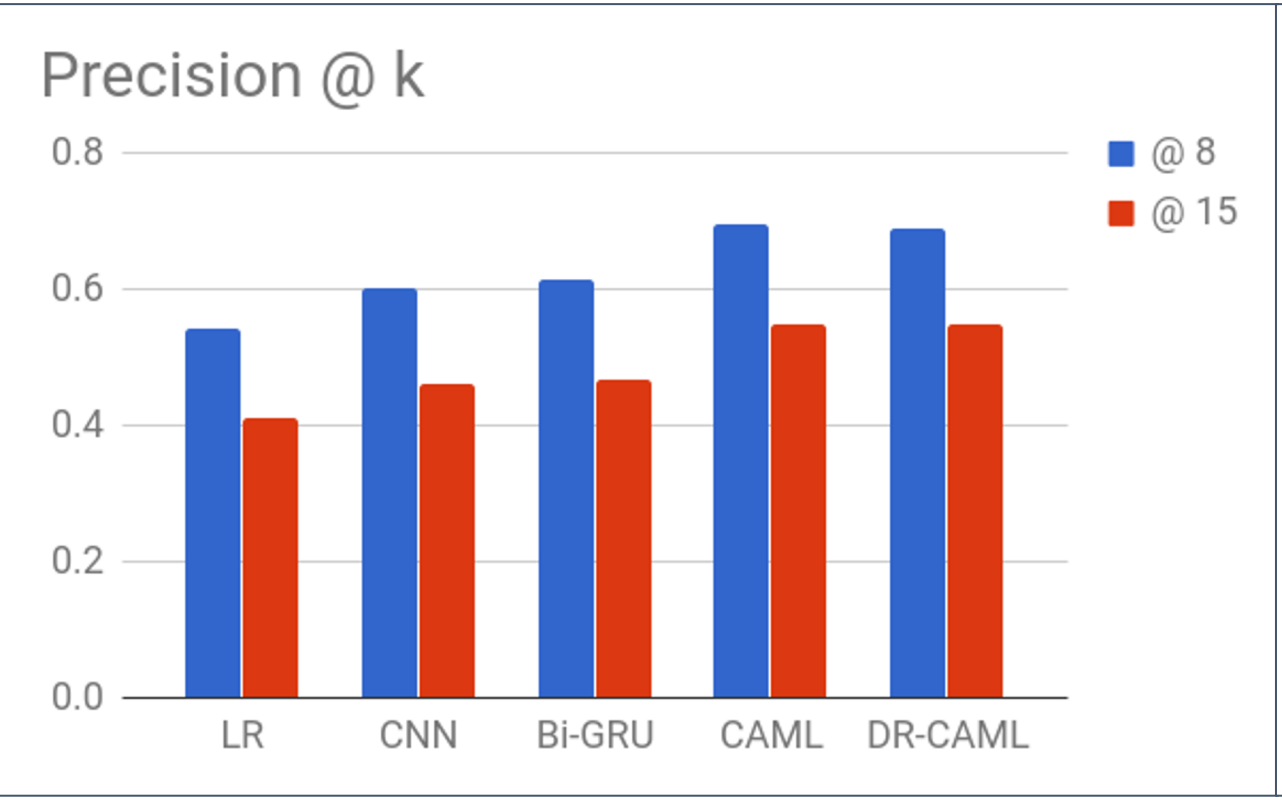
250.02: “Diabetes mellitus without mention of complication, type II or unspecified type, uncontrolled”

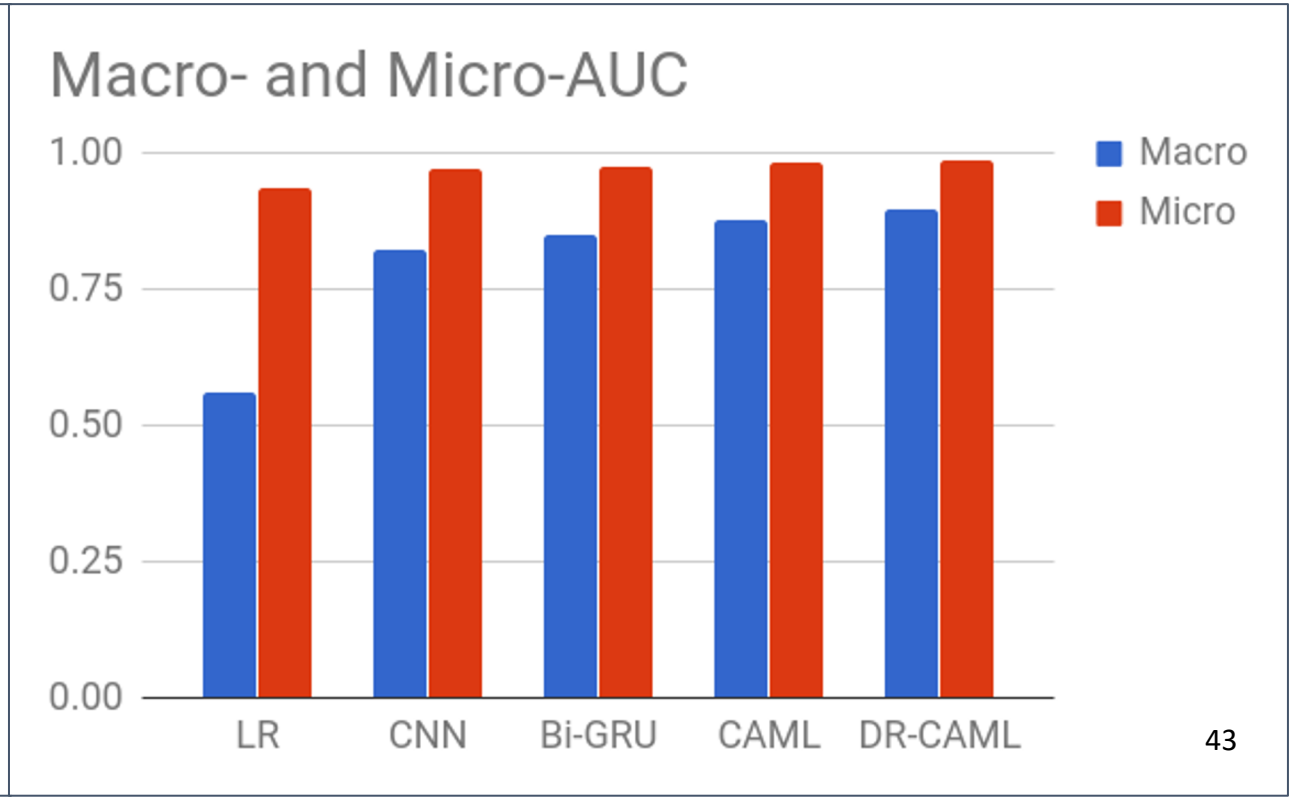
Dealing with the long tail: DR-CAML



Experiment results

- Enable future comparison
- Precision @ k: decision support use-case





But, is the model really interpretable?

- Physician evaluation: select the text snippet(s) that explain the code
- Informative and Highly Informative
- 100 random label-document samples
- Baselines: LogReg, CNN, cosine similarity

Physician evaluation example*

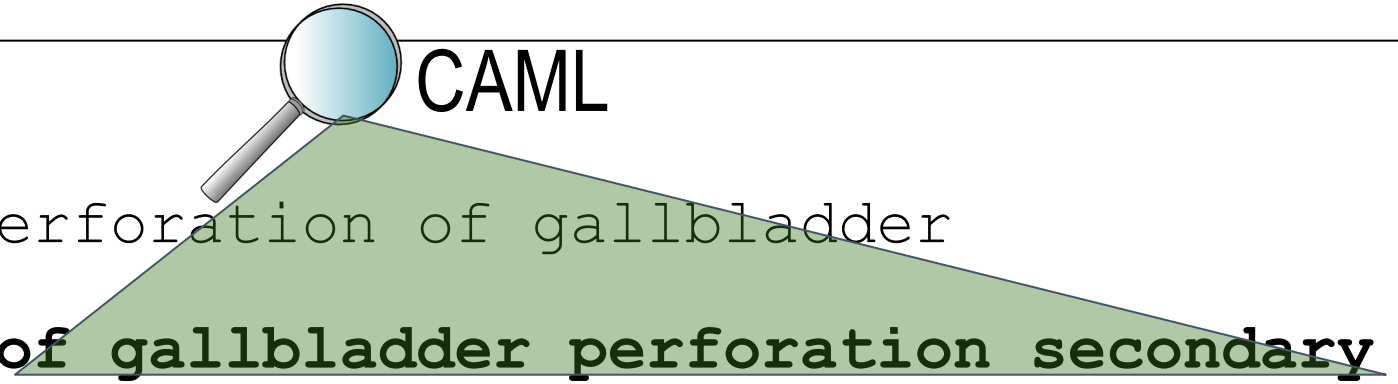
Code: 575.4

Full descriptions: Perforation of gallbladder

“. . in the setting of gallbladder perforation secondary to acute acalculous cholecystitis after . . . . . inhalation hospital1 times a day metronidazole mg tablet sig one tablet po tid times . . . . . to have an infection in your gallbladder requiring iv antibiotics and tube placement for . . .”

*not exact format used

Physician evaluation example*



Code: 575.4

Full descriptions: Perforation of gallbladder

“. . in the setting **of gallbladder perforation secondary** to acute acalculous cholecystitis after . . . . . inhalation hospital1 times a day metronidazole mg tablet sig one tablet po tid times . . . . . to have an infection in your gallbladder requiring iv antibiotics and tube placement for . . .”

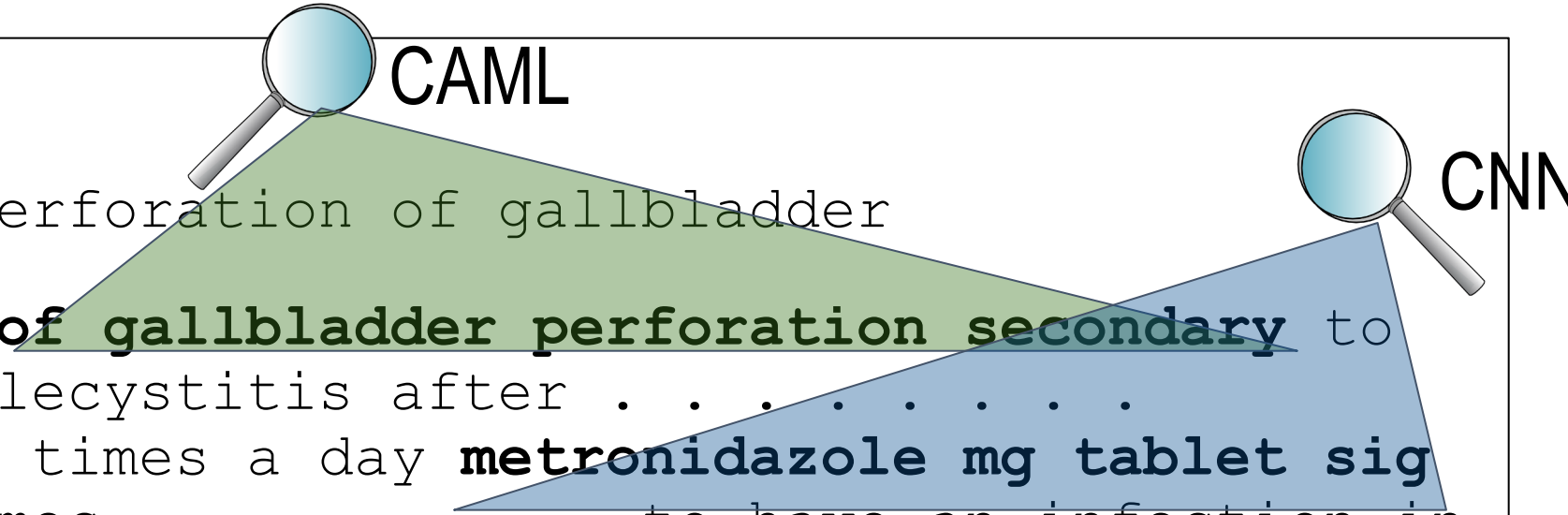
*not exact format used

Physician evaluation example*

Code: 575.4

Full descriptions: Perforation of gallbladder

“. . in the setting **of gallbladder perforation secondary** to acute acalculous cholecystitis after . . . . . inhalation hospital1 times a day **metronidazole mg tablet sig** one tablet po tid times . . . . . to have an infection in your gallbladder requiring iv antibiotics and tube placement for . . .”



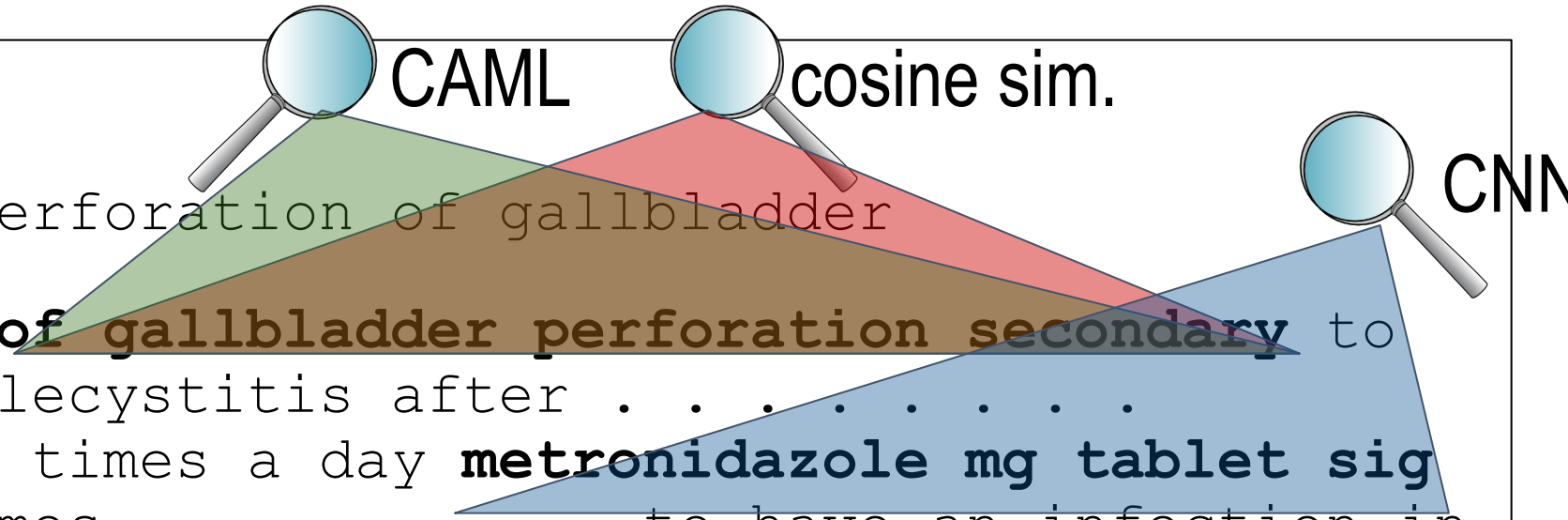
*not exact format used

Physician evaluation example*

Code: 575.4

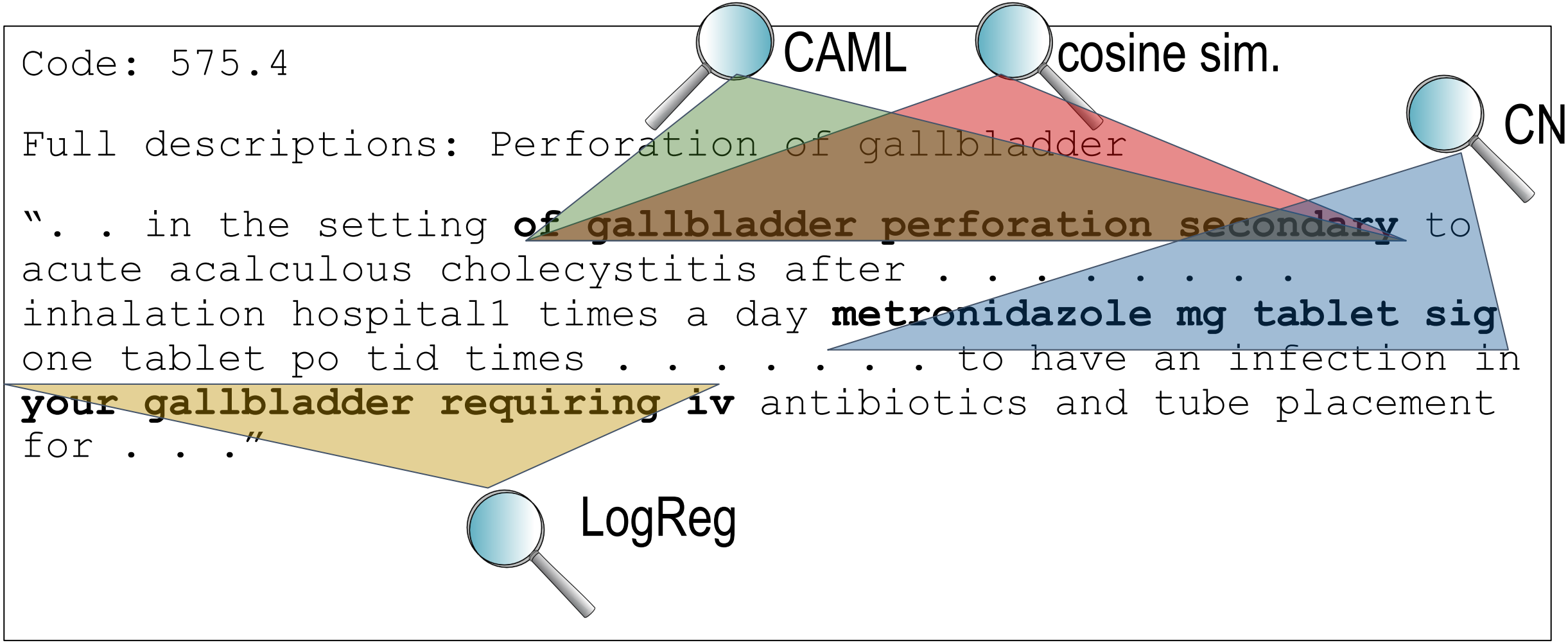
Full descriptions: Perforation of gallbladder

"**. . in the setting of gallbladder perforation secondary** to acute acalculous cholecystitis after . . . . . **inhalation hospital1 times a day metronidazole mg tablet sig** one tablet po tid times . . . . . to have an infection in your gallbladder requiring iv antibiotics and tube placement for . . ."



*not exact format used

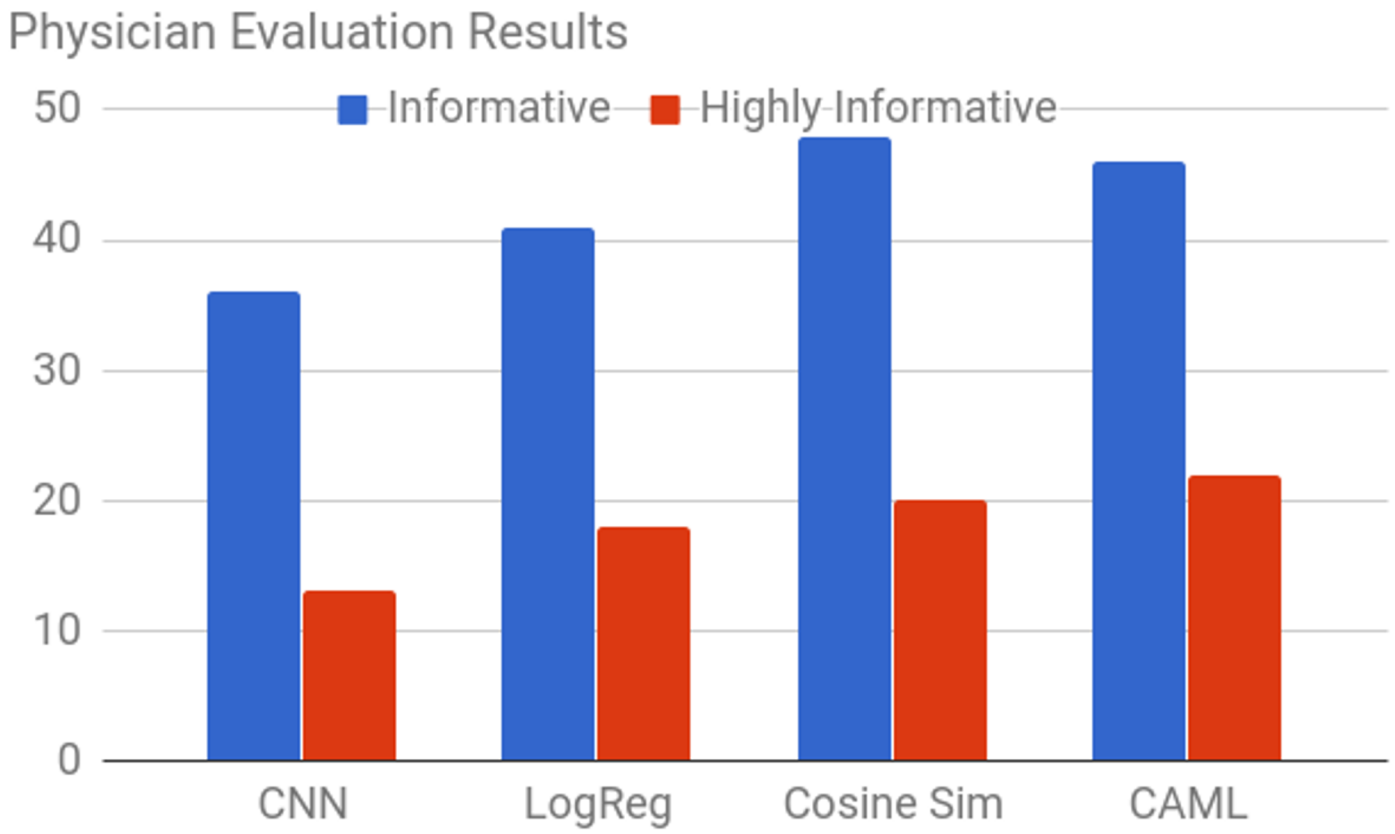
Physician evaluation example*



*not exact format used

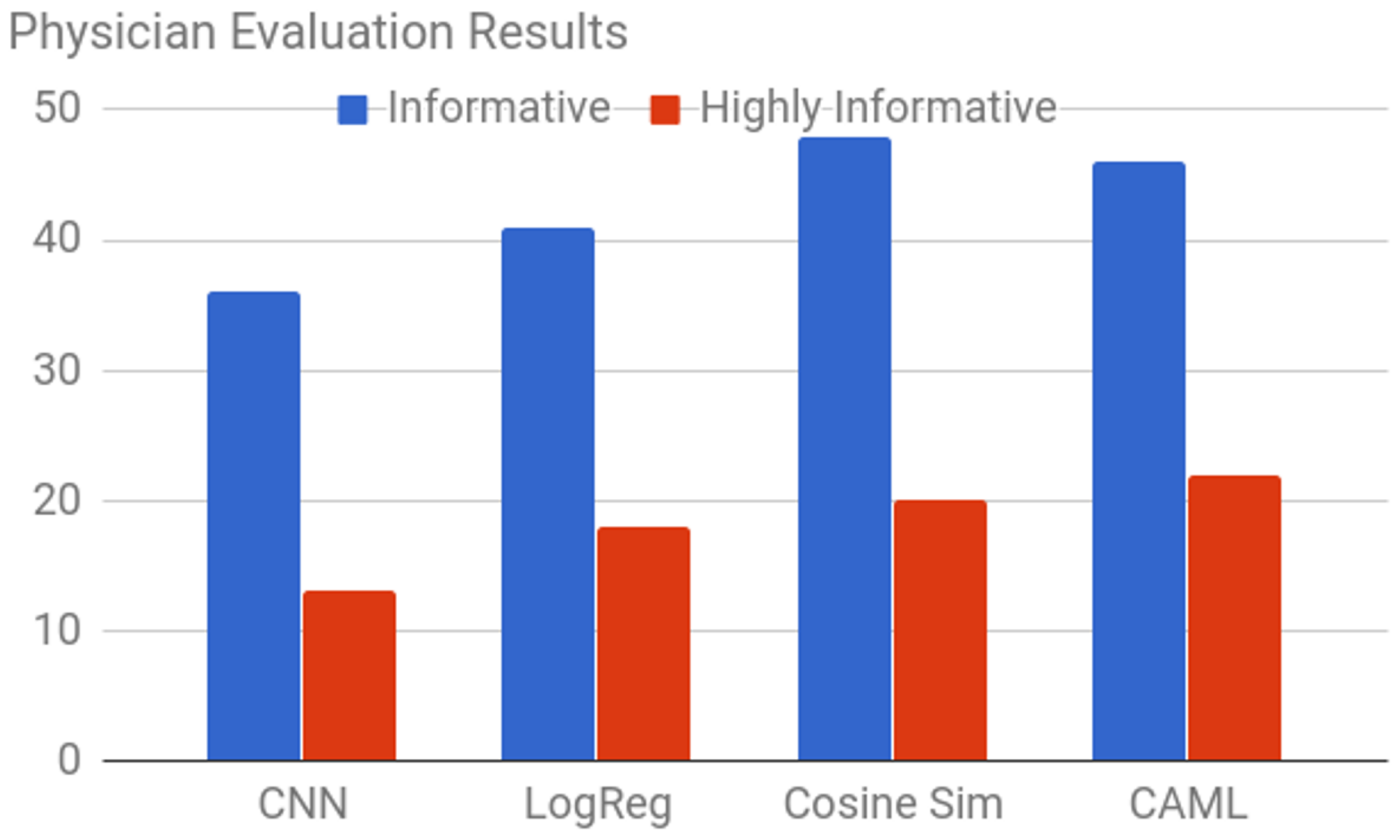
Physician evaluation results

- Improves upon CNN, LogReg



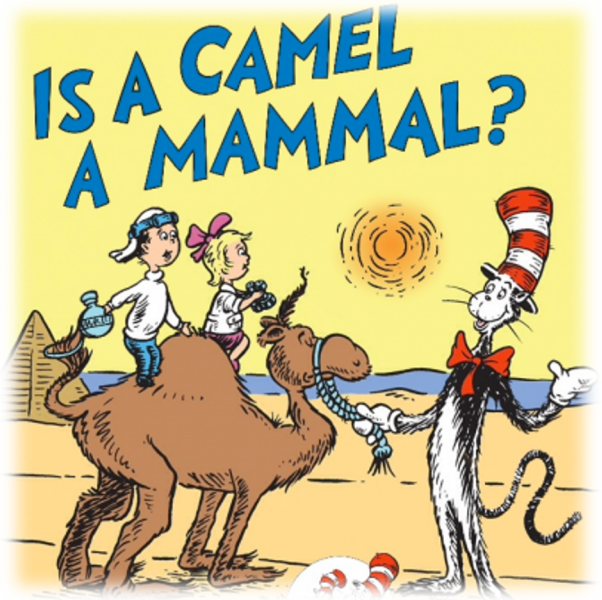
Physician evaluation results

- Improves upon CNN, LogReg
- More experts needed!



CAML: Explainable Prediction of Medical Codes from Clinical Text

- ICD coding is valuable and challenging
- Convolution + attention works well
- Attention can explain the predictions



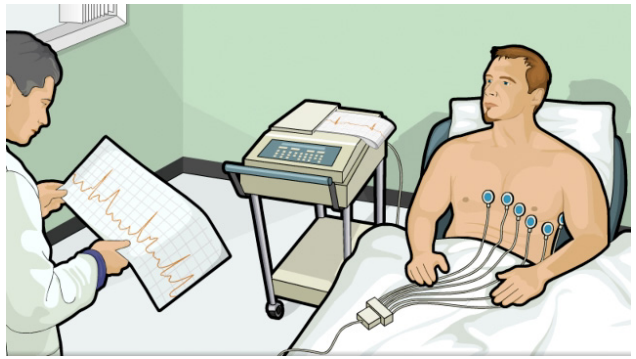
MINA: Multilevel Knowledge-Guided Attention for Modeling Electrocardiography Signals

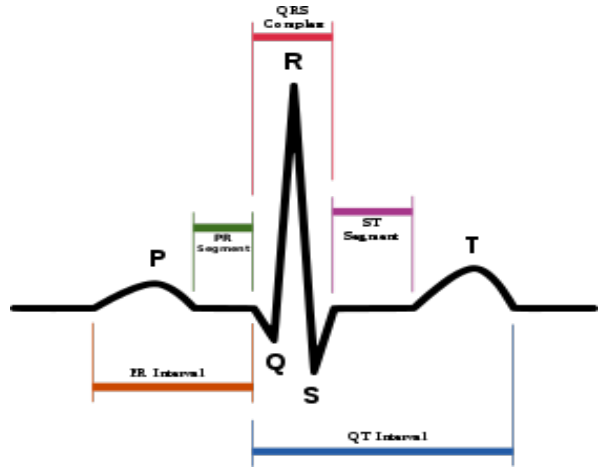
Shenda Hong, Cao Xiao, Tengfei Ma, Hongyan Li, Jimeng Sun

ICJAI'19

ILLINOIS

Motivation





Electrocardiography (ECG) is a commonly used non-invasive diagnostic tool for heart diseases

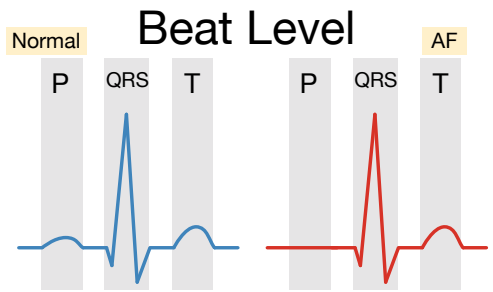
Deep learning models showed initial success in modeling ECG

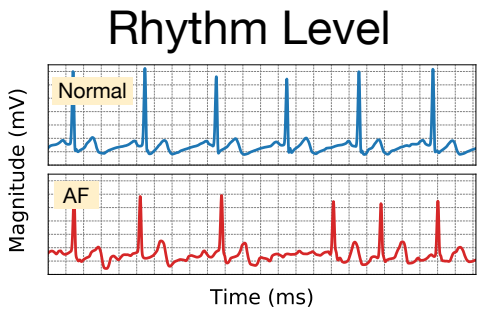
- Convolutional neural networks (CNN)
- Recurrent neural networks (RNN)
- Attention mechanism

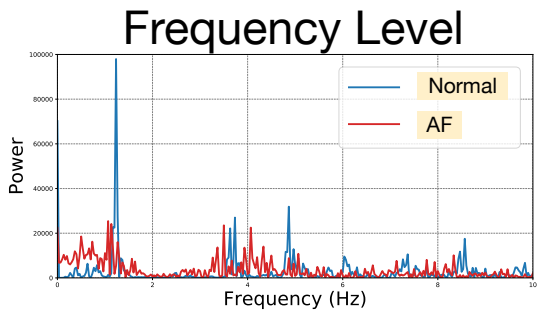
ILLINOIS

Challenges

Incorporate **knowledge** across different levels

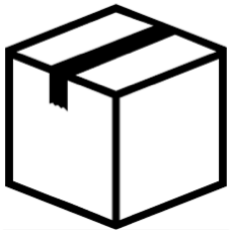




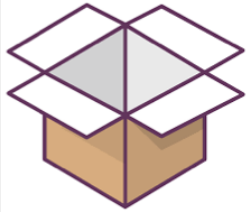


🤔

Provide **interpretable** results

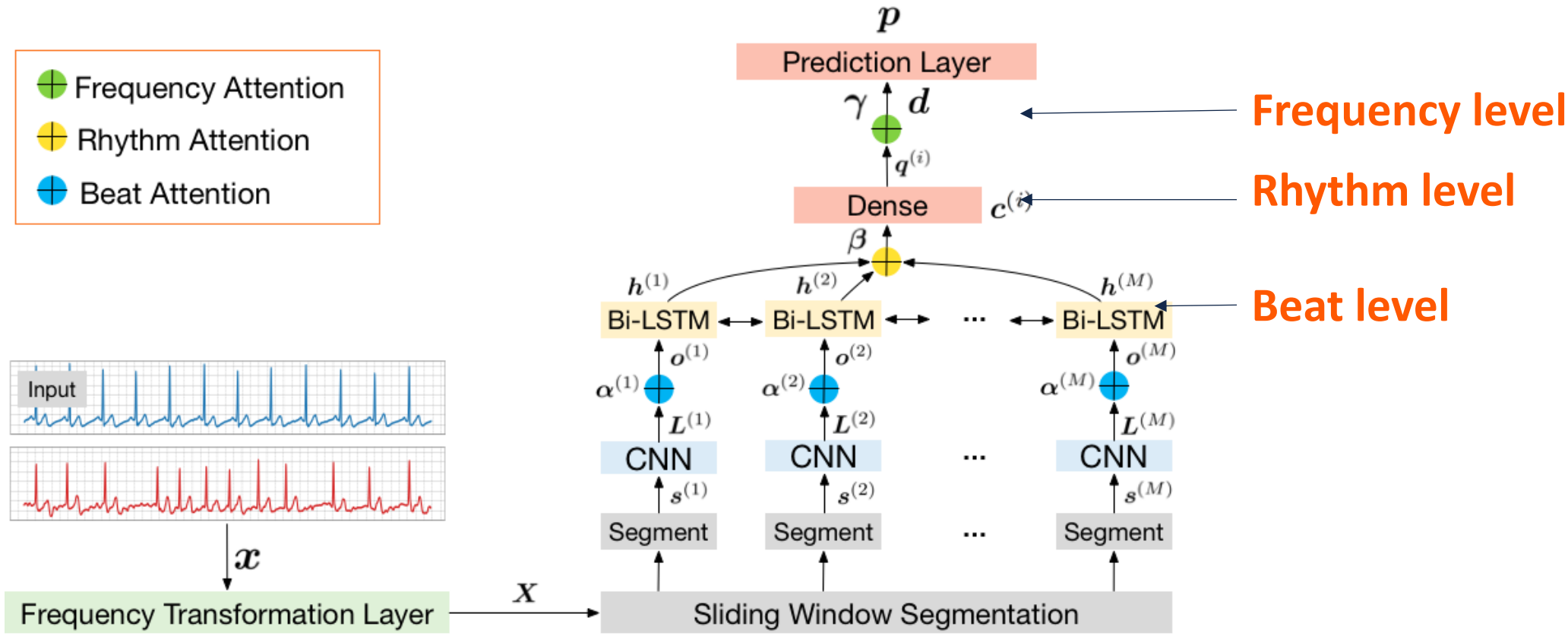


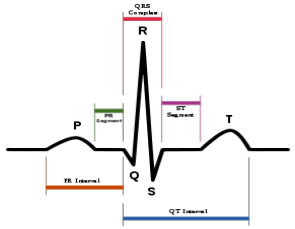
😰



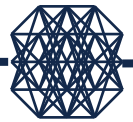
😎

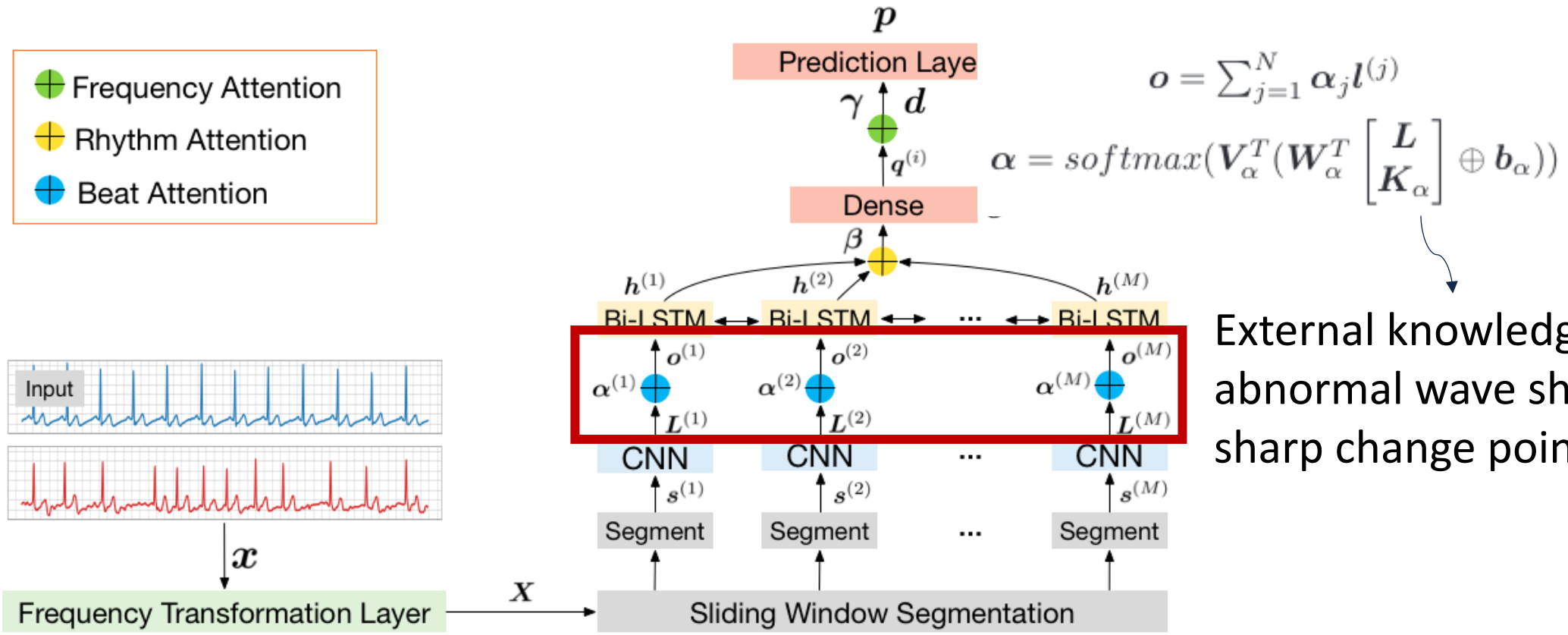
Multilevel Deep Neural Network





Beat-level attention

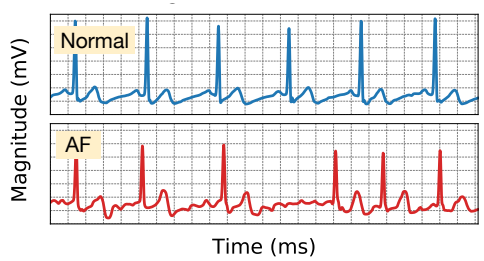




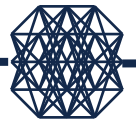
$$o = \sum_{j=1}^{N} \alpha_j l^{(j)}$$

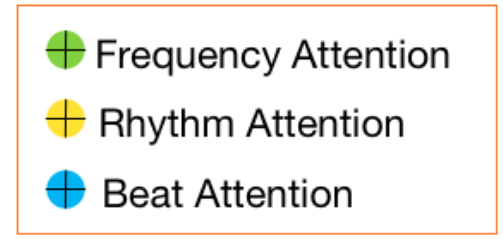
$$\alpha = softmax(V_\alpha^T(W_\alpha^T \begin{bmatrix} L \\ K_\alpha \end{bmatrix} \oplus b_\alpha))$$

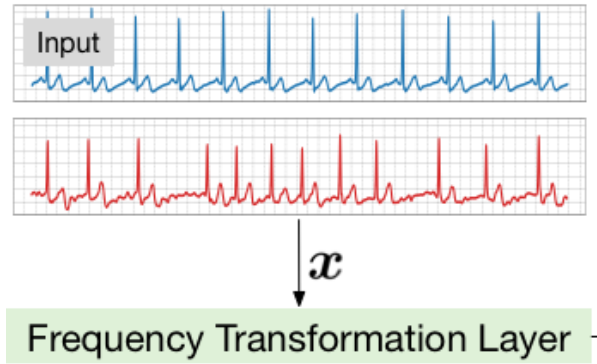
External knowledge: abnormal wave shapes, sharp change points



Rhythm level attention



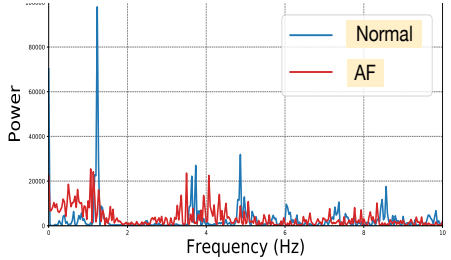




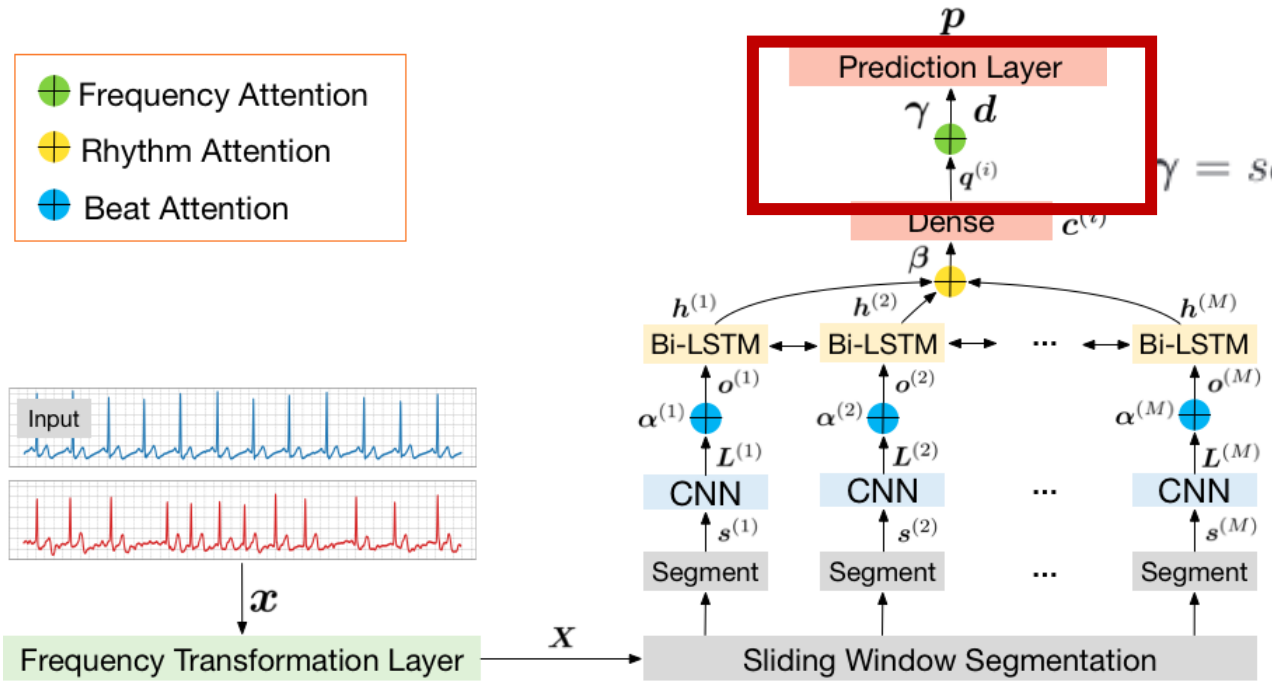
$$c = \sum_{k=1}^{M} \beta_k h^{(k)}$$

$$\beta = softmax(V_\beta^T (W_\beta^T \begin{bmatrix} H \\ K_\beta \end{bmatrix} \oplus b_\beta))$$

severe fluctuation



Frequency level attention



$$\boldsymbol{d} = \sum_{i=1}^{F} \gamma_i \boldsymbol{q}^{(i)}$$

$$\gamma = softmax(\boldsymbol{V}_\gamma^T(\boldsymbol{W}_\gamma^T \begin{bmatrix} \boldsymbol{Q} \\ \boldsymbol{K}_\gamma \end{bmatrix} \oplus \boldsymbol{b}_\gamma))$$

power spectral density

Experiments

PhysioNet Challenge 2017 databases

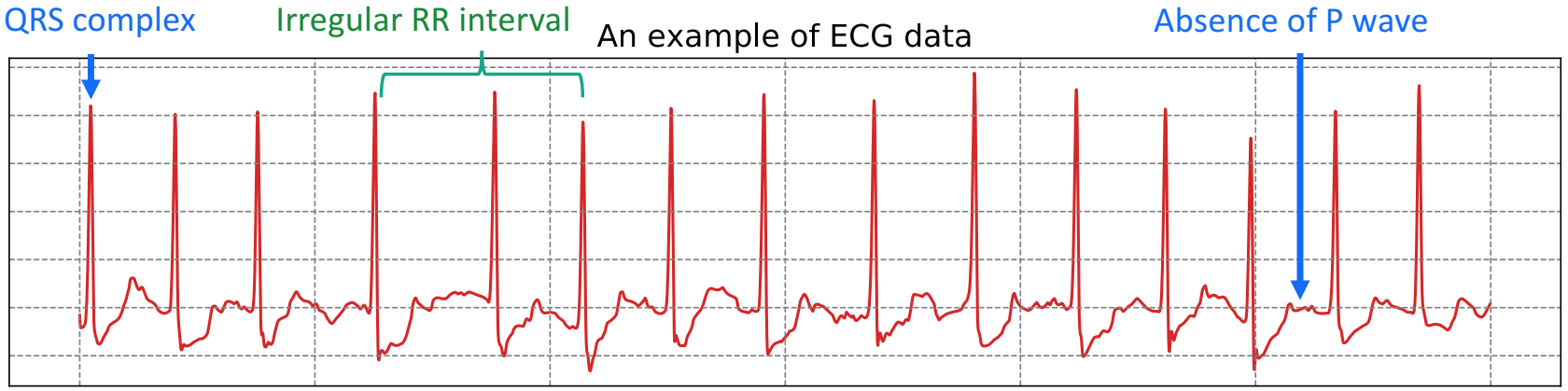
- ECG recordings lasting from 9s to just over 60s and sampled at 300Hz by the AliveCor device
- 738 from AF patients and 7790 from controls as predefined by the challenge

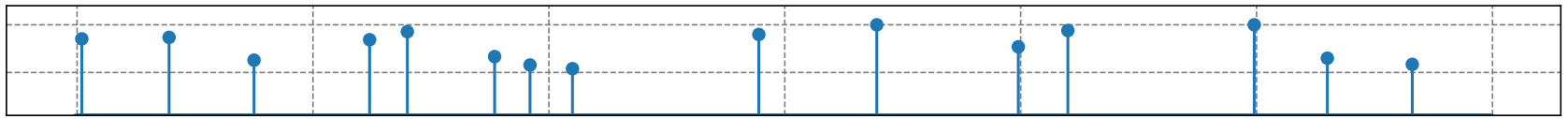
Discriminate records of AF patients from those of controls

	ROC-AUC	PR-AUC	F1
ExpertLR	0.9350 ± 0.0000	0.8730 ± 0.0000	0.8023 ± 0.0000
ExpertRF	0.9394 ± 0.0000	0.8816 ± 0.0000	0.8180 ± 0.0000
CNN	0.8711 ± 0.0036	0.8669 ± 0.0068	0.7914 ± 0.0090
CRNN	0.9040 ± 0.0115	0.8943 ± 0.0111	0.8262 ± 0.0215
ACRNN	0.9072 ± 0.0047	0.8935 ± 0.0087	0.8248 ± 0.0229
MINA	$\mathbf{0.9488} \pm 0.0081$	$\mathbf{0.9436} \pm 0.0082$	$\mathbf{0.8342} \pm 0.0352$

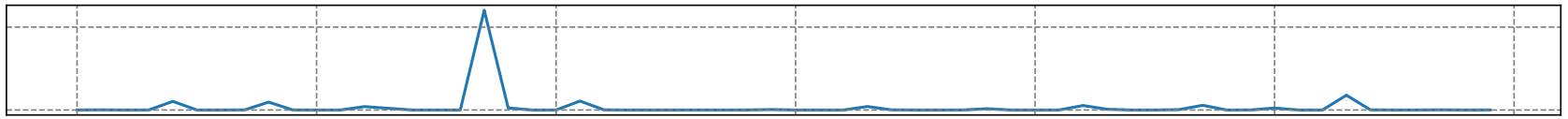
ILLINOIS

MINA provides interpretable results



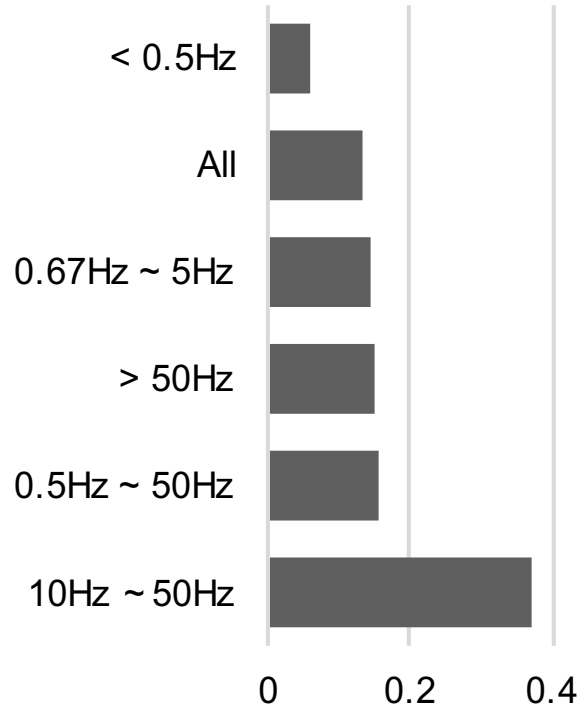


Beat level attention points the location of QRS complex and absent P waves.



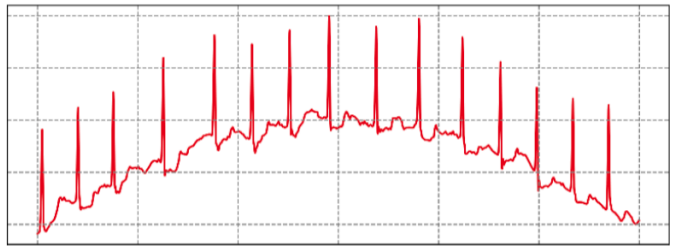
Rhythm level attention shows the location of abnormal RR interval.

Frequency level attention
shows QRS complex is dominant between 10Hz ~ 50 Hz.

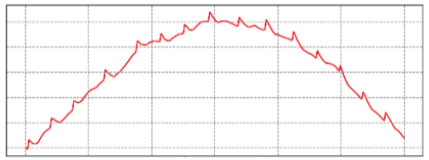


ILLINOIS

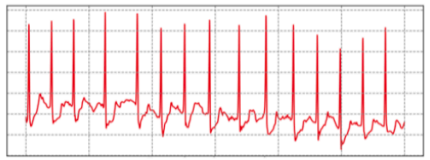
Experiment: MINA handles well low frequency noise



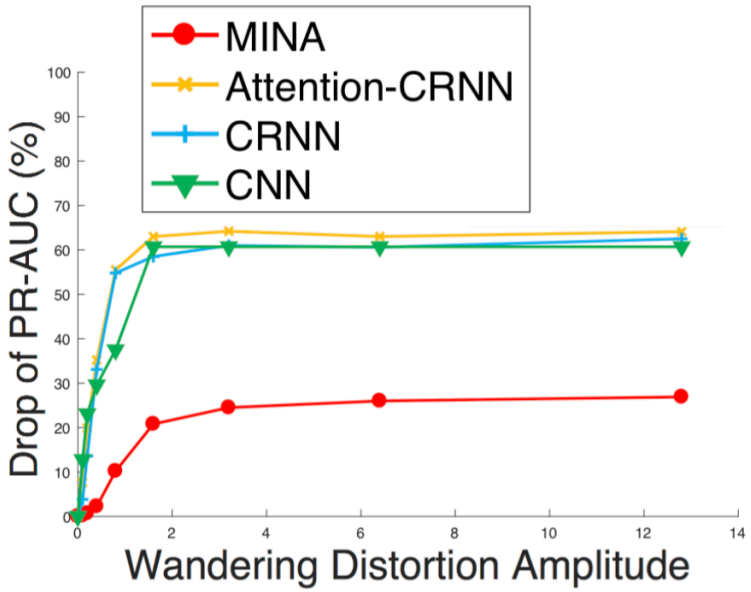
(a) Baseline wandering distortion ECG signal



(b) Channel 1



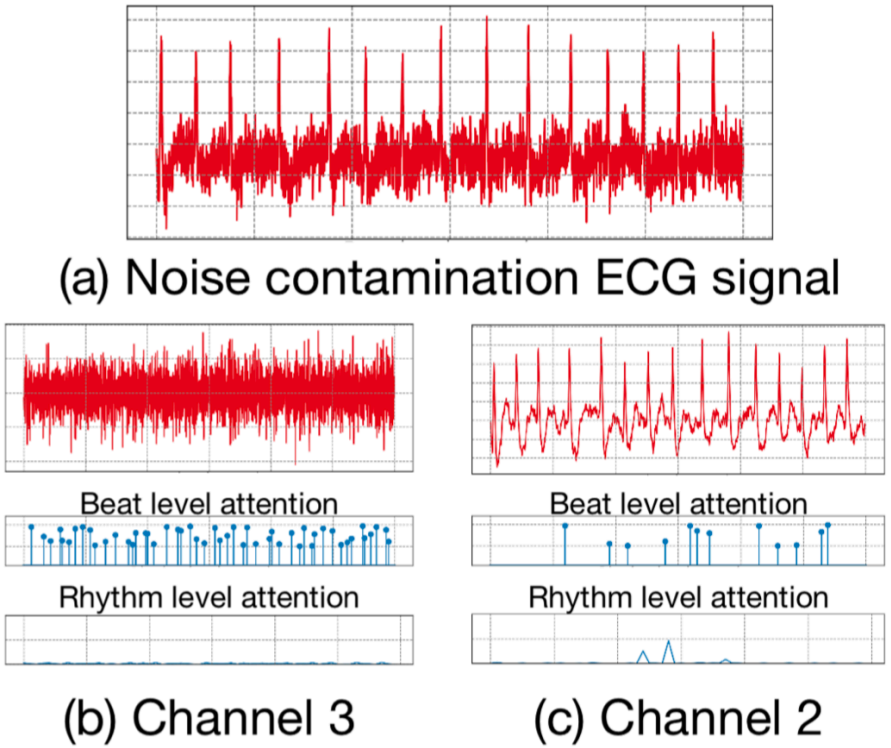
(c) Channel 2



(d) Performance drop comparison

ILLINOIS

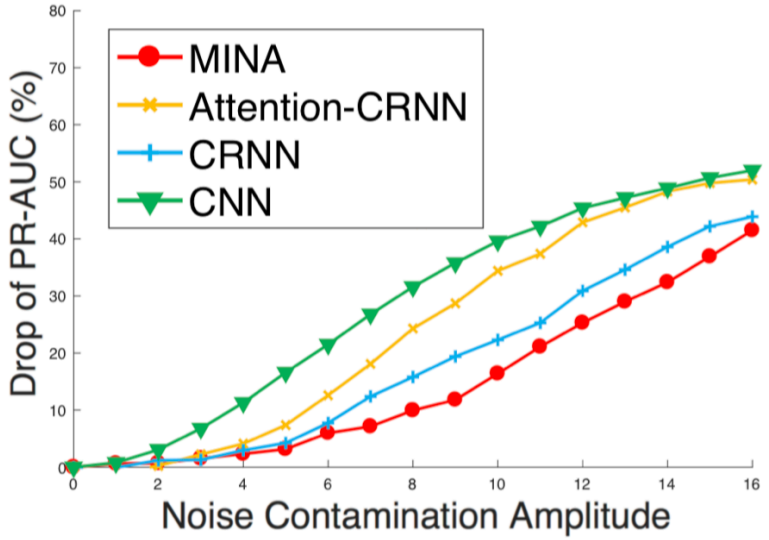
Experiment: MINA handles well high frequency noise



(a) Noise contamination ECG signal

(b) Channel 3

(c) Channel 2



(d) Performance drop comparison