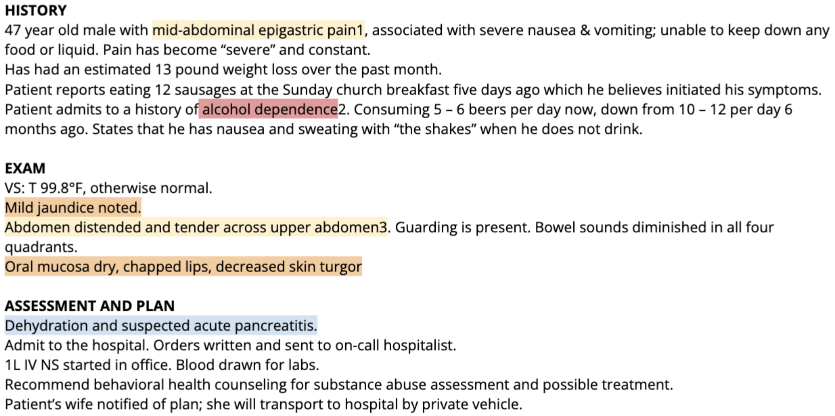
Explainable Prediction of Medical Codes from Clinical Text

Authors: J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein

Presented by: Leo Benac, Chelse Swoopes, and Kelly Zhang

Introduction

- Clinical notes
  - Free text narratives generated by clinicians during patient encounters
- International Classification of Diseases (ICD) codes
  - Metadata codes that accompany clinical notes
  - Standardized way of indicating diagnoses and procedures
  - Multiple use cases (billing, predictive modelling)



HISTORY

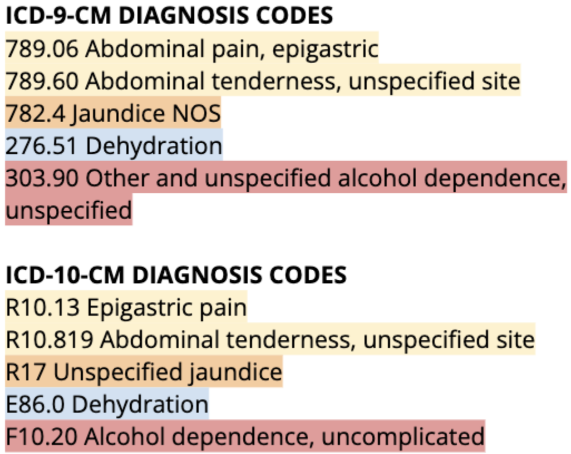
47 year old male with mid-abdominal epigastric pain1, associated with severe nausea & vomiting; unable to keep down any food or liquid. Pain has become "severe" and constant.
Has had an estimated 13 pound weight loss over the past month.
Patient reports eating 12 sausages at the Sunday church breakfast five days ago which he believes initiated his symptoms.
Patient admits to a history of alcohol dependence2. Consuming 5 – 6 beers per day now, down from 10 – 12 per day 6 months ago. States that he has nausea and sweating with "the shakes" when he does not drink.

EXAM

VS: T 99.8°F, otherwise normal.
Mild jaundice noted.
Abdomen distended and tender across upper abdomen3. Guarding is present. Bowel sounds diminished in all four quadrants.
Oral mucosa dry, chapped lips, decreased skin turgor

ASSESSMENT AND PLAN

Dehydration and suspected acute pancreatitis.
Admit to the hospital. Orders written and sent to on-call hospitalist.
1L IV NS started in office. Blood drawn for labs.
Recommend behavioral health counseling for substance abuse assessment and possible treatment.
Patient's wife notified of plan; she will transport to hospital by private vehicle.



ICD-9-CM DIAGNOSIS CODES

789.06 Abdominal pain, epigastric
789.60 Abdominal tenderness, unspecified site
782.4 Jaundice NOS
276.51 Dehydration
303.90 Other and unspecified alcohol dependence, unspecified

ICD-10-CM DIAGNOSIS CODES

R10.13 Epigastric pain
R10.819 Abdominal tenderness, unspecified site
R17 Unspecified jaundice
E86.0 Dehydration
F10.20 Alcohol dependence, uncomplicated

Scenario reference: https://www.practicefusion.com/icd-10/clinical-concepts-for-family-practice/icd-10-clinical-scenarios/

Motivation

- Limitation of ICD code annotation
  - Labor intensive
  - Coders read between 720 and 1,440 pages of Clinical Documentation per day[2]
  - Prone to error
  - Lack of connection bridging text and code

- Challenges+complexity of automatic coding
  - Features of clinical text
    - Irrelevant information, misspellings, non-standard abbreviations, large medical vocabulary
  - Label space

Summary of Contributions

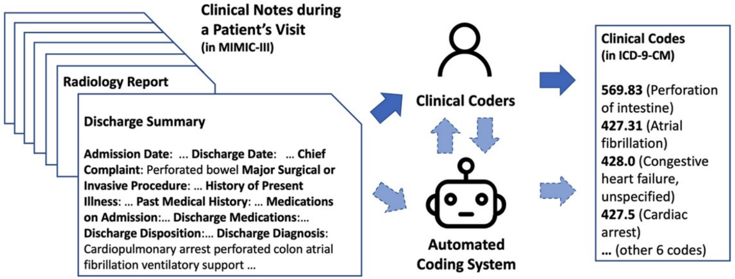
- Convolutional Attention for Multi-Label classification (CAML)
    - Attentional convolutional network
        - CNN + per-label attention mechanism
    - Important information correlated with a code’s presence may be in short snippets anywhere in the document
    - Decision support setting (clinical demain)

- Accuracy Evaluation
    - Used open-access MIMIC datasets
    - Outperformed previously proposed approaches

- Interpretability Evaluation
    - Physician rated the informativeness of a set of automatically generated explanations

Related Work: Automatic ICD Coding

- Structured text v. unstructured text
- Datasets
  - Subset of the full ICD label space (Wang et al., 2016)
  - Subset of medical scenarios (Zhang et al., 2017)
  - Death certificates in English + French (Névéol et al., 2017)
  - Private datasets (Subotin and Davis, 2016)
- Approaches
  - Recurrent networks with HA-GRU to classify ICD9 diagnosis codes (Baumel et al. (2018)
  - Character-aware LSTMs to generate sentence representations (Shi et al. 2017)
  - Memory networks to predict top-50 and top-100 Codes (Prakash et al. 2017)
  - Grounded Recurrent Neural Network to predict the presence of individual labels (Vani et al., 2017)



Clinical Notes during a Patient's Visit (in MIMIC-III)

Radiology Report

Discharge Summary

Admission Date: ... **Discharge Date:** ... **Chief Complaint**: Perforated bowel **Major Surgical or Invasive Procedure**: ... **History of Present Illness**: ... **Past Medical History**: ... **Medications on Admission**:... **Discharge Medications**:... **Discharge Disposition**:... **Discharge Diagnosis**: Cardiopulmonary arrest perforated colon atrial fibrillation ventilatory support ...

Clinical Coders

Automated Coding System

Clinical Codes (in ICD-9-CM)

569.83 (Perforation of intestine)
427.31 (Atrial fibrillation)
428.0 (Congestive heart failure, unspecified)
427.5 (Cardiac arrest)
... (other 6 codes)

Image reference: https://www.nature.com/articles/s41746-022-00705-7.pdf

The CAML architecture yields stronger results across all experimental conditions.

Related Work: Explainable Text Classification

- Model “rationales” through a latent variable (Lei et al., 2016)
  - Tag each word as relevant to the document label

- Compute salience of individual words (Li et al., 2016)

- Use attention to highlight salient features of the text (Rush et al., 2015; Rocktäschel et al., 2016)
  - Lack interoperability evaluations of features selected by the attention mechanism

Related Work: Attentional Convolution for NLP

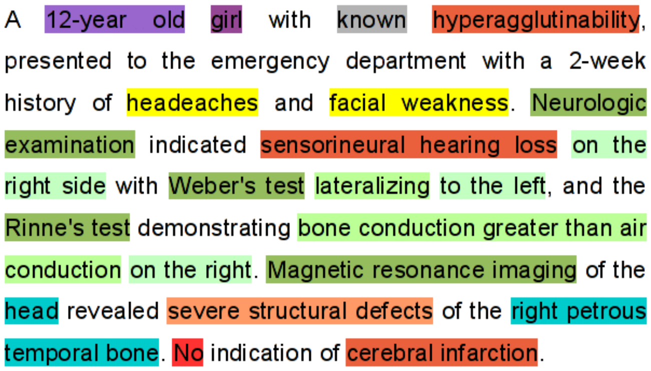
- CNNs have been applied to tasks such as sentiment classification (Kim, 2014) and language modeling

- Convolution and attention
  - Allamanis et al., 2016, Yin et al., 2016, dos Santos et al., 2016
  - Yin and Schütze, 2017

- Most relevant work (Yang et al. 2016 and Allamanis et al. 2016)
  - Use context vectors to compute attention over specific locations in the text

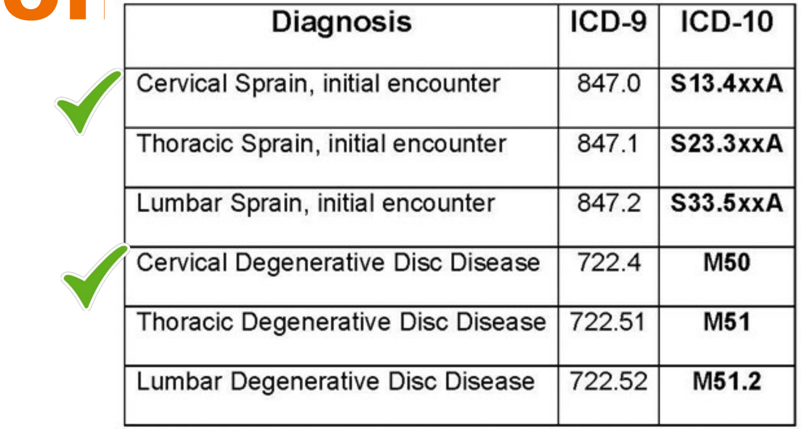
- Differentiation
  - Goal is to select locations in a document which are most important for predicting specific labels
  - Compute separate attention weights for each label in the label space

ICD Code Prediction Problem



A 12-year old girl with known hyperagglutinability, presented to the emergency department with a 2-week history of headeaches and facial weakness. Neurologic examination indicated sensorineural hearing loss on the right side with Weber's test lateralizing to the left, and the Rinne's test demonstrating bone conduction greater than air conduction on the right. Magnetic resonance imaging of the head revealed severe structural defects of the right petrous temporal bone. No indication of cerebral infarction.

Clinical Text



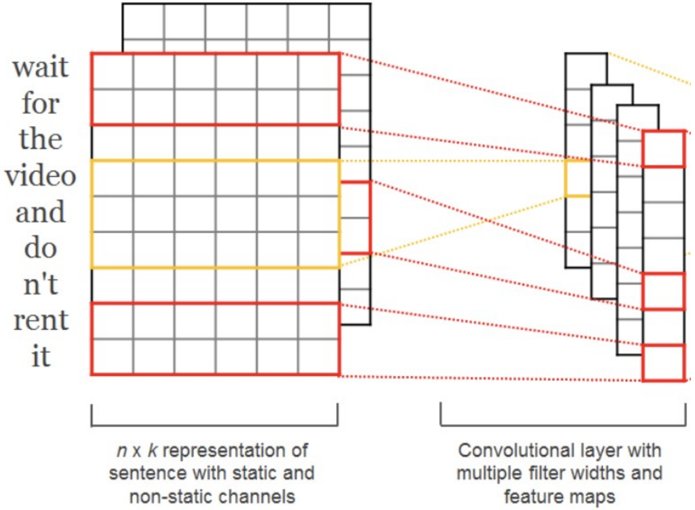
	Diagnosis	ICD-9	ICD-10
✓	Cervical Sprain, initial encounter	847.0	**S13.4xxA**
	Thoracic Sprain, initial encounter	847.1	**S23.3xxA**
	Lumbar Sprain, initial encounter	847.2	**S33.5xxA**
✓	Cervical Degenerative Disc Disease	722.4	**M50**
	Thoracic Degenerative Disc Disease	722.51	**M51**
	Lumbar Degenerative Disc Disease	722.52	**M51.2**

ICD Codes

Multi-Label Classification Problem

- “Match” snippets of clinical text to a label
- Use code descriptions to help make text snippets relevant to the predicted codes (esp. for labels with little data)

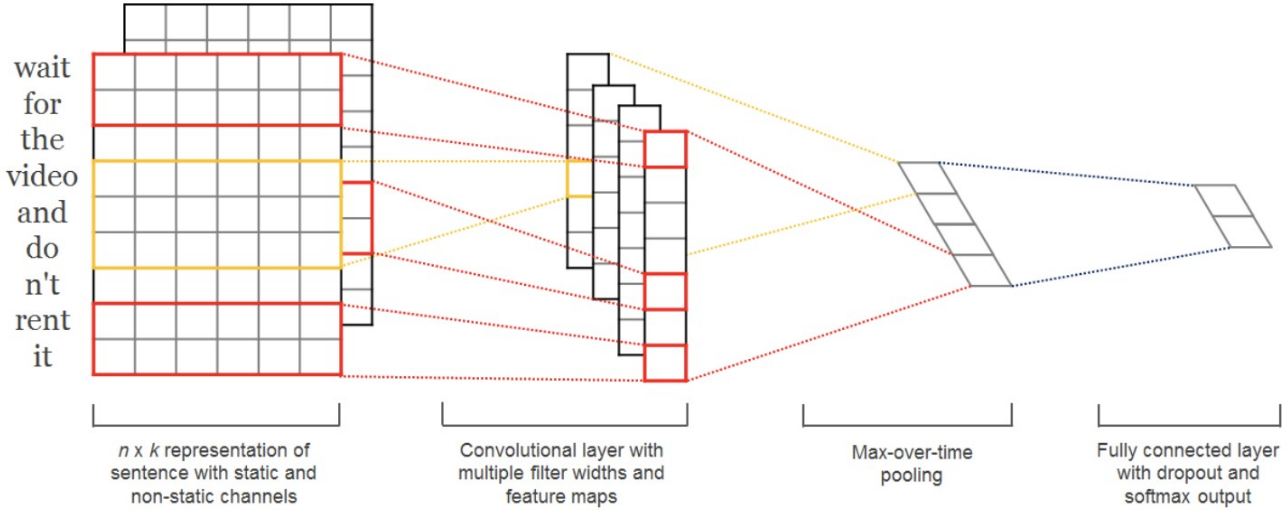
Typical Convolutional N-Gram Model


wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps

- Word embeddings
- Convolutional filter of “N” consecutive words (N-gram)
- Each convolutional filter “detects” certain “word snippet” motifs

Image taken from https://people.csail.mit.edu/yoonkim/data/sent-cnn-slides.pdf

Typical Convolutional N-Gram Model

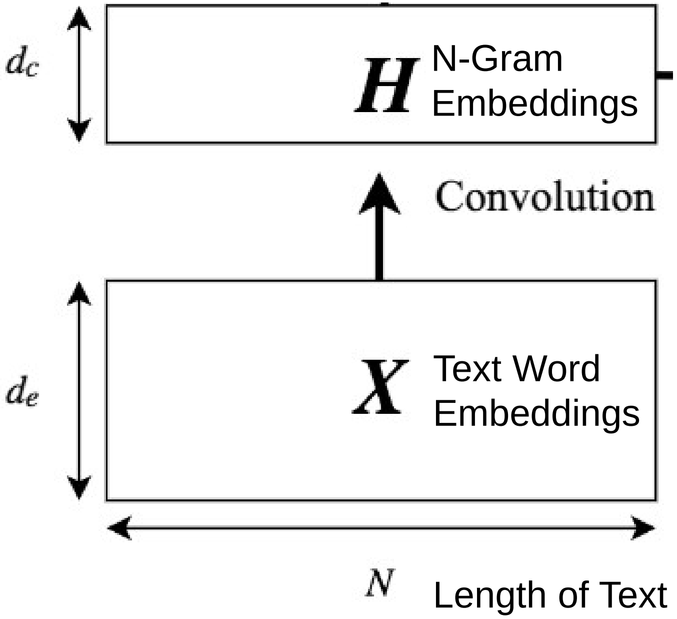

wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Max pooling → Was the "N-gram" detected at all in this piece of text?

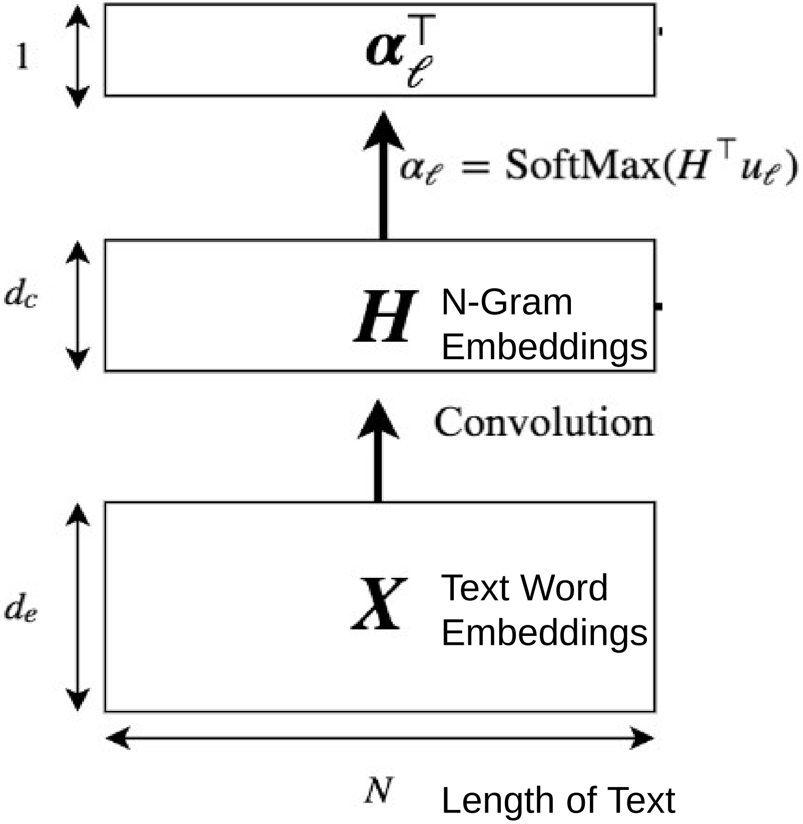
- Word embeddings
- Convolutional filter of "N" consecutive words (N-gram)
- Each convolutional filter "detects" certain "word snippet" motifs

Image taken from https://people.csail.mit.edu/yoonkim/data/sent-cnn-slides.pdf

Convolutional Attention Model



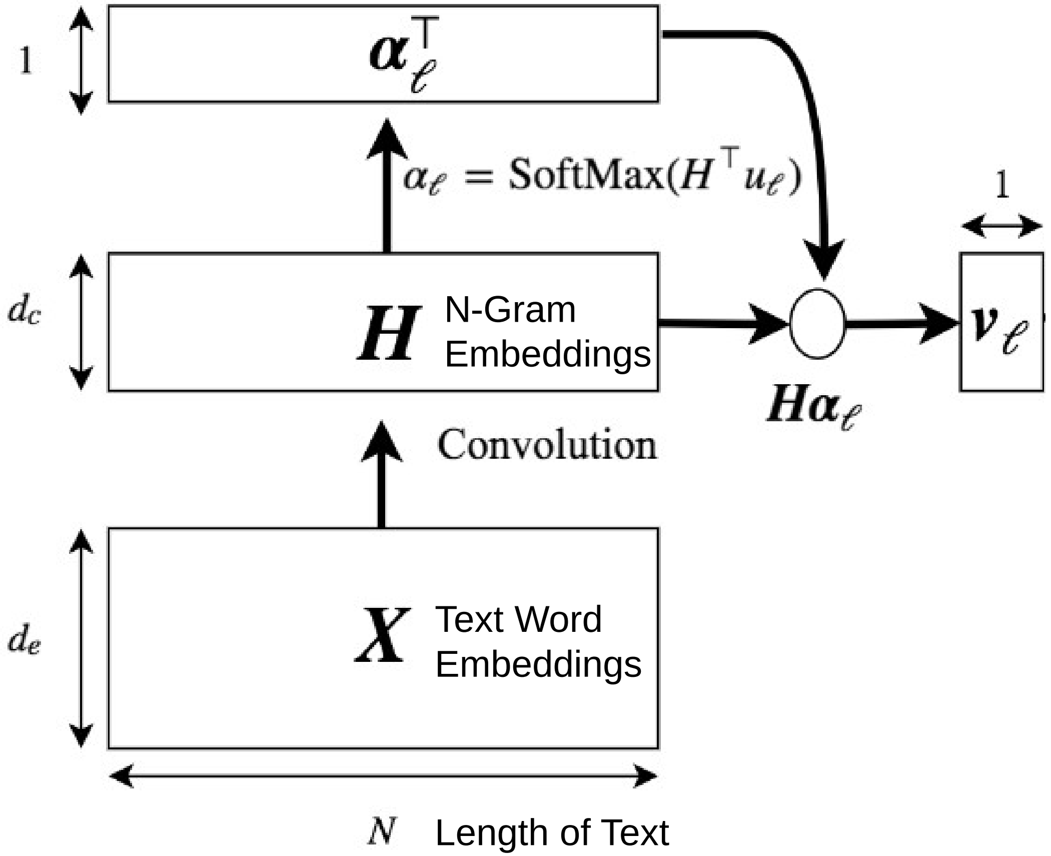
Convolutional Attention Model



$$\boldsymbol{\alpha}_\ell = \text{SoftMax}(\boldsymbol{H}^\top \boldsymbol{u}_\ell),$$

Weight representing “match” between n-grams in document (H) and label embedding

Convolutional Attention Model



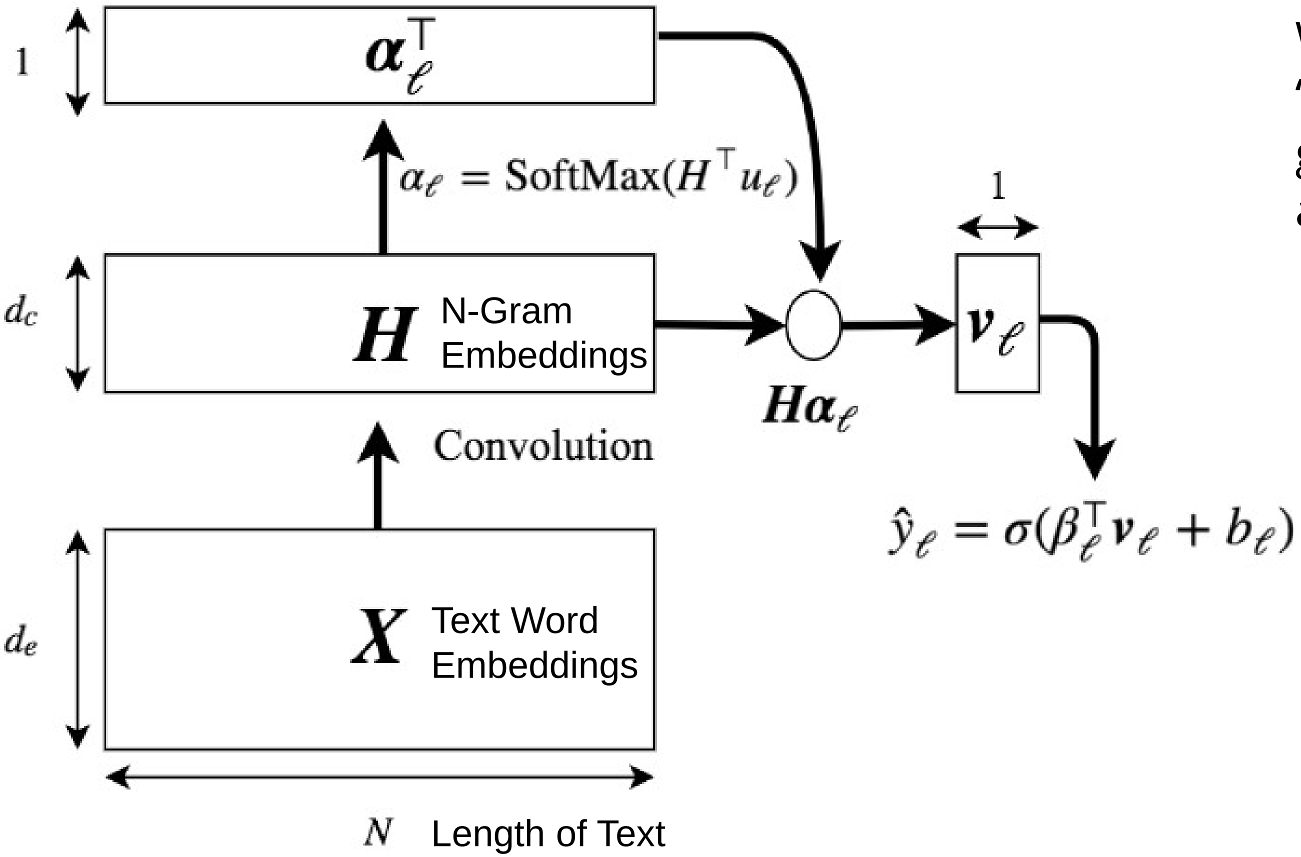
$$\boldsymbol{\alpha}_\ell = \mathrm{SoftMax}(\boldsymbol{H}^\top \boldsymbol{u}_\ell),$$

Weight representing "match" between n-grams in document (H) and label embedding

$$\boldsymbol{v}_\ell = \sum_{n=1}^{N} \alpha_{\ell,n} \boldsymbol{h}_n.$$

N-grams weighted by weights alpha form document vector v

Convolutional Attention Model


1
$\boldsymbol{\alpha}_\ell^\top$
$\alpha_\ell = \mathrm{SoftMax}(H^\top u_\ell)$
d_c
$\boldsymbol{H}$
N-Gram Embeddings
Convolution
d_e
$\boldsymbol{X}$
Text Word Embeddings
N
Length of Text
$\boldsymbol{H\alpha}_\ell$
1
$\boldsymbol{v}_\ell$
$\hat{y}_\ell = \sigma(\beta_\ell^\top \boldsymbol{v}_\ell + b_\ell)$

$$\boldsymbol{\alpha}_\ell = \mathrm{SoftMax}(\boldsymbol{H}^\top \boldsymbol{u}_\ell),$$

Weight representing "match" between n-grams in document (H) and label embedding

$$\boldsymbol{v}_\ell = \sum_{n=1}^{N} \alpha_{\ell,n} \boldsymbol{h}_n.$$

N-grams weighted by weights alpha form document vector v

Training Loss Function

$$L_{\mathrm{BCE}}(\boldsymbol{X}, \boldsymbol{y}) = -\sum_{\ell=1}^{\mathcal{L}} y_\ell \log(\hat{y}_\ell) \\ + (1 - y_\ell) \log(1 - \hat{y}_\ell),$$

Binary Cross-Entropy Loss

- Binary classification for each possible label

$$L(\boldsymbol{X}, \boldsymbol{y}) = L_{\mathrm{BCE}} + \lambda \frac{1}{n_y} \sum_{\ell : y_\ell = 1}^{\mathcal{L}} \|\boldsymbol{z}_\ell - \boldsymbol{\beta}_\ell\|_2,$$

Regularize to make Labels Representations Close to Code Descriptions

Training Loss Function

$$L_{\mathrm{BCE}}(\boldsymbol{X}, \boldsymbol{y}) = -\sum_{\ell=1}^{\mathcal{L}} y_\ell \log(\hat{y}_\ell) + (1 - y_\ell) \log(1 - \hat{y}_\ell),$$

Binary Cross-Entropy Loss

- Binary classification for each possible label

$$L(\boldsymbol{X}, \boldsymbol{y}) = L_{\mathrm{BCE}} + \lambda \frac{1}{n_y} \sum_{\ell : y_\ell = 1}^{\mathcal{L}} \|\boldsymbol{z}_\ell - \boldsymbol{\beta}_\ell\|_2,$$

Regularize to make Labels Representations Close to Code Descriptions

Number of true labels for data example

Representation of label based on code description

Representation of label learned by model

Dataset

- MIMIC 3
- Input: Sequence of words describing the stay of a patient in the ICU

  (Max length of 2500 words)
- Output: ICD-9 codes describing diagnoses and procedures
- 9,000 unique ICD-9 codes:
  - 7,000 diagnosis codes and 2,000 treatment codes
- Preprocessing (100 dimension embeddings)

	MIMIC-III full
# training documents	47,724
Vocabulary size	51,917
Mean # tokens per document	1,485
Mean # labels per document	15.9
Total # labels	8,922

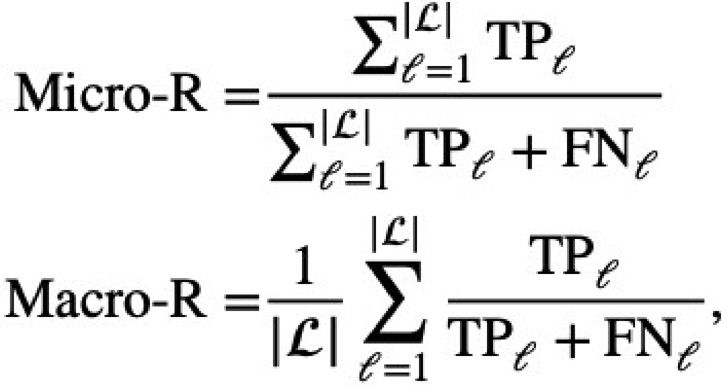
Baselines

- Single Layer CNN
- Logistic Regression
- Bidirectional Gated Recurrent Unit

Evaluation Metrics

- Micro-averaged and Macro-averaged F1 and AUC
- Precision@n

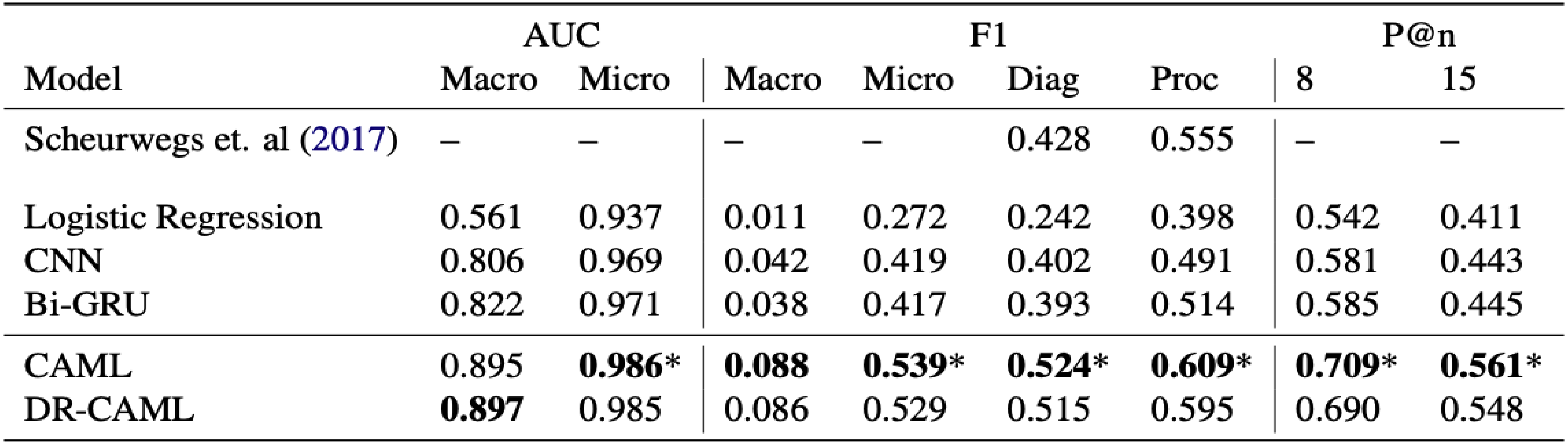
  (Most common n labels in the ground truth data)



$$\text{Micro-R} = \frac{\sum_{\ell=1}^{|\mathcal{L}|} \text{TP}_\ell}{\sum_{\ell=1}^{|\mathcal{L}|} \text{TP}_\ell + \text{FN}_\ell}$$

$$\text{Macro-R} = \frac{1}{|\mathcal{L}|} \sum_{\ell=1}^{|\mathcal{L}|} \frac{\text{TP}_\ell}{\text{TP}_\ell + \text{FN}_\ell},$$

Results

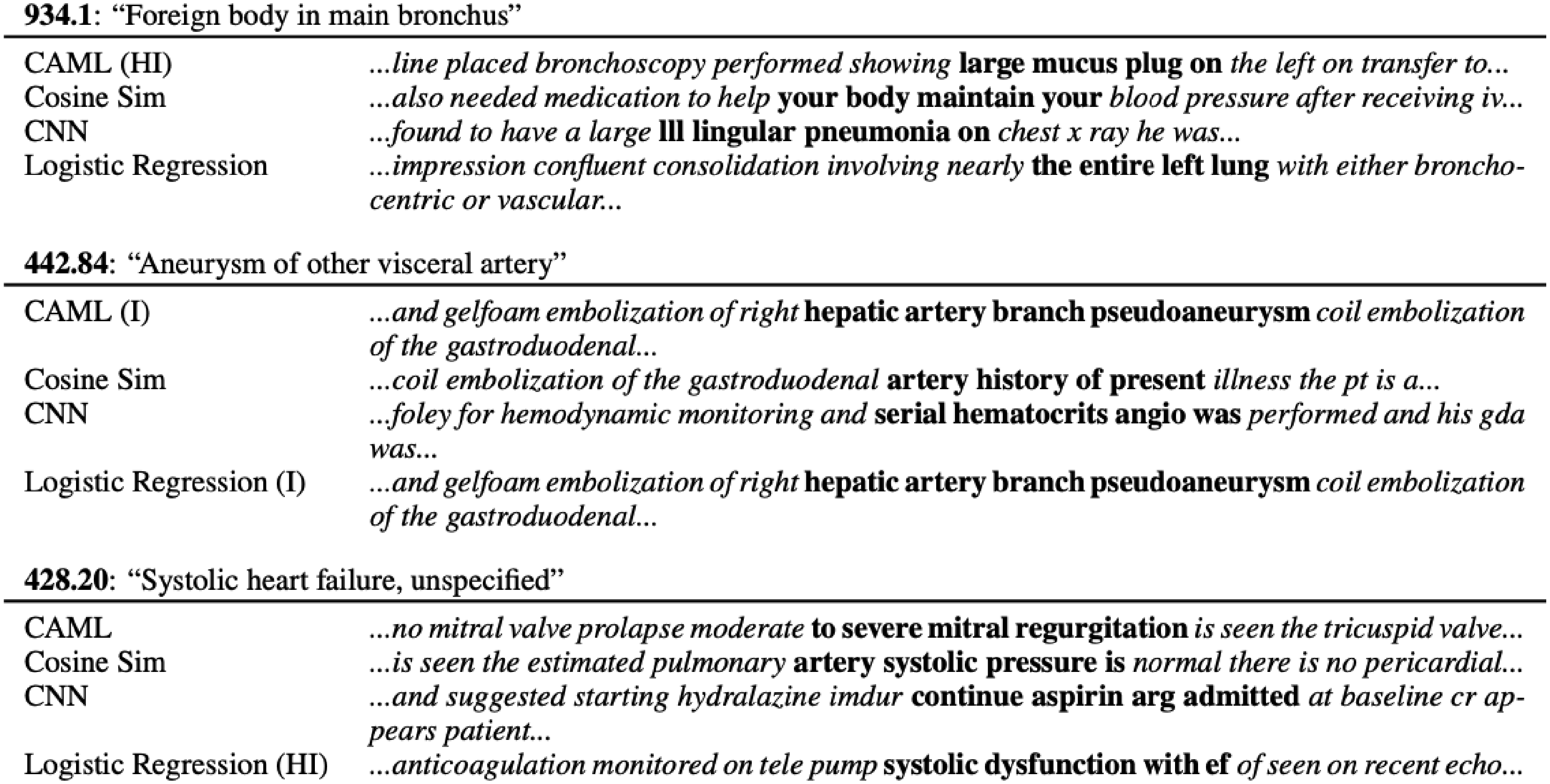


	AUC		F1				P@n	
Model	Macro	Micro	Macro	Micro	Diag	Proc	8	15
Scheurwegs et. al (2017)	–	–	–	–	0.428	0.555	–	–
Logistic Regression	0.561	0.937	0.011	0.272	0.242	0.398	0.542	0.411
CNN	0.806	0.969	0.042	0.419	0.402	0.491	0.581	0.443
Bi-GRU	0.822	0.971	0.038	0.417	0.393	0.514	0.585	0.445
CAML	0.895	**0.986***	**0.088**	**0.539***	**0.524***	**0.609***	**0.709***	**0.561***
DR-CAML	**0.897**	0.985	0.086	0.529	0.515	0.595	0.690	0.548

- The CAML model gives the strongest results on all metrics
- DR-CAML is the regularized version using lambda = 0.01
- The Bi-GRU architecture is comparable to the vanilla CNN
- Logistic Regression worse than all neural architectures.

Interpretability Evaluation

- Sample of 100 predicted codes
- Using he most important k-gram (k=4) + five words on each side for context
- Asked the Physician how informative it is



934.1: “Foreign body in main bronchus”

CAML (HI)	*...line placed bronchoscopy performed showing* **large mucus plug on** *the left on transfer to...*
Cosine Sim	*...also needed medication to help* **your body maintain your** *blood pressure after receiving iv...*
CNN	*...found to have a large* **lll lingular pneumonia on** *chest x ray he was...*
Logistic Regression	*...impression confluent consolidation involving nearly* **the entire left lung** *with either bronchocentric or vascular...*

442.84: “Aneurysm of other visceral artery”

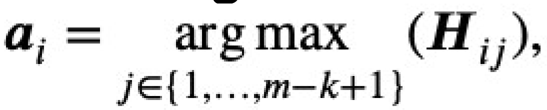
CAML (I)	*...and gelfoam embolization of right* **hepatic artery branch pseudoaneurysm** *coil embolization of the gastroduodenal...*
Cosine Sim	*...coil embolization of the gastroduodenal* **artery history of present** *illness the pt is a...*
CNN	*...foley for hemodynamic monitoring and* **serial hematocrits angio was** *performed and his gda was...*
Logistic Regression (I)	*...and gelfoam embolization of right* **hepatic artery branch pseudoaneurysm** *coil embolization of the gastroduodenal...*

428.20: “Systolic heart failure, unspecified”

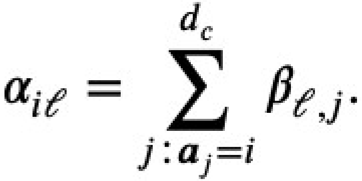
CAML	*...no mitral valve prolapse moderate* **to severe mitral regurgitation** *is seen the tricuspid valve...*
Cosine Sim	*...is seen the estimated pulmonary* **artery systolic pressure is** *normal there is no pericardial...*
CNN	*...and suggested starting hydralazine imdur* **continue aspirin arg admitted** *at baseline cr appears patient...*
Logistic Regression (HI)	*...anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo...*

How to pick the most influential 4-gram for each label?

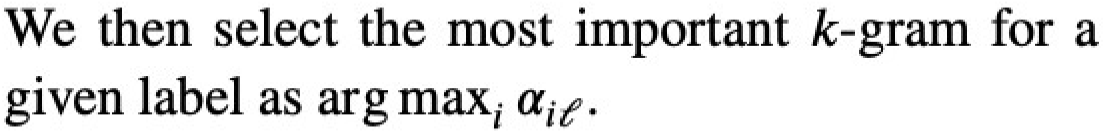
- **CAML** Looks at the SoftMax output $\boldsymbol{\alpha}\ell$ to get the best 4-gram for a code prediction ℓ
- **Logistic regression** takes the sum of the coefficients of the weight matrix for ℓ, over the words in the k-gram. The top-scoring k-gram is then returned.
- **Code descriptions** calculate a word similarity (cosine similarity) metric between each stemmed k-gram and the stemmed ICD-9 code description.
- **Max-Pooling CNN**



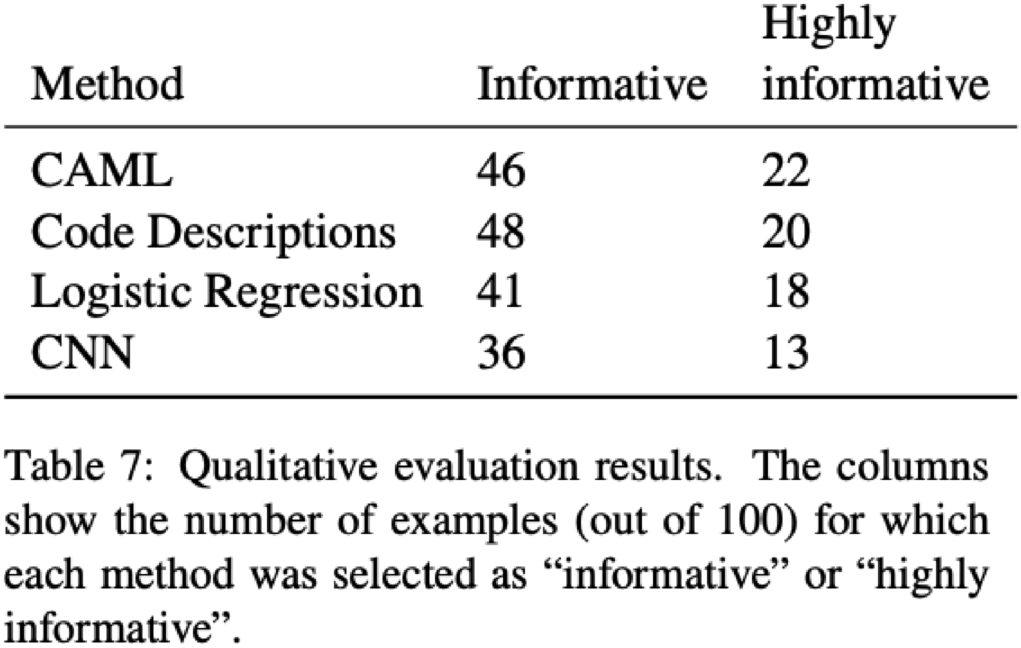
$$\boldsymbol{a}_i = \mathop{\arg\max}_{j\in\{1,\ldots,m-k+1\}} (\boldsymbol{H}_{ij}),$$



$$\alpha_{i\ell} = \sum_{j:\boldsymbol{a}_j=i}^{d_c} \beta_{\ell,j}.$$



We then select the most important k-gram for a given label as $\arg\max_i \alpha_{i\ell}$.



Method	Informative	Highly informative
CAML	46	22
Code Descriptions	48	20
Logistic Regression	41	18
CNN	36	13

Table 7: Qualitative evaluation results. The columns show the number of examples (out of 100) for which each method was selected as “informative” or “highly informative”.

- CAML selects the greatest number of “highly informative” explanations, and selects more “informative” explanations than both the CNN baseline and the logistic regression model.
- Cosine Similarity metric also performs well the examples in Table 1 demonstrates the strengths of CAML.

Conclusion

Concluding Thoughts

- Yields strong improvements over previous metrics across several formulations of the ICD-9 code prediction task
- Provides satisfactory explanations for its predictions per stakeholder evaluation

Future Work

- Extensible
  - To other multi-label document tagging task
- Linguistic perspective
  - Integrate document structure of discharge summaries in MIMIC-III
  - Better handle non-standard writing and other sources of out-of-vocabulary tokens
- Application perspective
  - Build models that leverage hierarchy of ICD codes
  - Predict codes for future visits

Discussion

1. What are potential limitations of CAML?

2. The explanations provided from the attention mechanism are in the form of extracted snippets of text from the input document. Does this provide enough information? What else could be provided with this explanation?

3. Strengths and weaknesses of their evaluation and results (for both accuracy and interpretability)

4. What additional applications could CAML be useful for?

References

1. Mullenbach, James, et al. "Explainable prediction of medical codes from clinical text." arXiv preprint arXiv:1802.05695 (2018).

2. Scriven, Sarah. “An Introduction to Clinical Coding.” Federation for Informatics Professionals (FEDIP), FEDIP, 14 Mar. 2022, https://www.fedip.org/post/what-is-clinical-coding.