TOWARDS TRANSFORMER-BASED AUTOMATED ICD CODING: CHALLENGES, PITFALLS AND SOLUTIONS

DECEMBER 8, 2021

WEIMING REN, TIANSHU ZHU, RUIJING ZENG, TONGZI WU

AGENDA

- Background & Motivation
- Related Work
- Our Goal
- Method
- Dataset & Experiment
- Discussions
- Conclusion & Future work

BACKGROUND & MOTIVATION

- **I**nternational **C**lassification of **D**iseases (**ICD**) coding
    - Assign ICD codes to clinical discharge summaries
- Her husband, who was home with her at the time told her she was "out cold" for about two minutes. The patient continues to have cephalgias since it happened, primarily occipital, extending up into the bilateral occipital and parietal regions. The headaches come on suddenly, last for long periods of time, and occur every day. They are not relieved by Advil.

- S06.0x1A - Concussion with loss of consciousness of 30 minutes or less
- G44.311 - Acute post traumatic headache

BACKGROUND & MOTIVATION

- **I**nternational **C**lassification of **D**iseases (**ICD**) coding
  - Assign ICD codes to clinical discharge summaries
  - By professional clinical coders
  - Time-consuming and error-prone
- Automated ICD coding
  - Automatically predicts ICD codes from clinical discharge summaries
  - Machine learning:
    - Traditional ML: SVM, Logistic Regression
    - RNN: Bidirectional GRU / LSTM
    - CNN: TextCNN

RELATED WORK - CNN

- Clinical discharge summaries are unstructured text
  - Convert the raw text to **latent text representations**.
- **CAML** - **C**onvolutional **A**ttention network for **M**ulti-**L**abel classification (Mullenbach et al. 2018)
  - 1 convolutional layer + 1 attention layer
- **MultiResCNN** - **Multi**-Filter **Res**idual **CNN** (Li et al. 2019 - based on CAML)
  - Multiple filter CNN layers + residual convolutional layers + 1 attention layer
- Both achieved the **state-of-the-art** results when published.

OUR GOAL

- CNN-based ICD coding performed better compared to transformer, which has dominated NLP.
- Transformer-based ICD coding has been proved to be challenging
  - No transformer-based models have achieved state-of-the-art results.
- Investigate the pitfalls and present our solutions to transformer-based ICD coding.

METHOD

- We employ an Encoder-Decoder architecture for our ICD coding models
  - Encoder: extracts features and generate contextual representations
  - Decoder: aggregates encoder output and performs classification
- Main concerns for selecting Encoder architectures:
  - How will the input sequence length affect the ICD coding performance?
  - Deep and complex architectures? (BERT, other Transformer-based models)
  - Domain-specific (clinical/medical related) models? (BioBERT, ClinicalBERT)

ENCODER

- List of selected encoder architectures:

Encoder Model	Architecture	Input Sequence Length	Pretrained Dataset
MultiResCNN-512	CNN	512	-
MultiResCNN-2500	CNN	2500	-
BERT-512	BERT	512	Generic
ClinicalBERT-512	BERT	512	Clinical notes
XLNet-1500	XLNet	1500	Generic
ClinicalXLNet-1500	XLNet	1500	Clinical discharge summaries
Longformer-2500	Longformer	2500	Generic
Longformer-3200	Longformer	3200	Generic

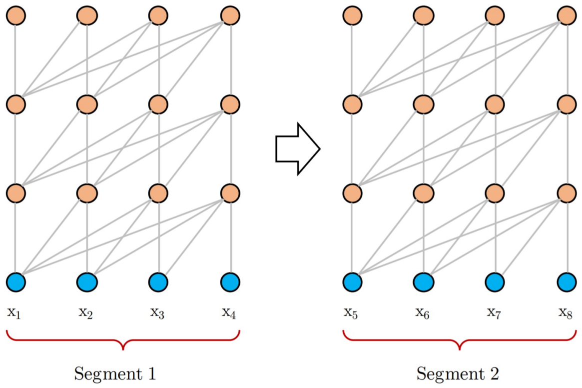
- Problem with vanilla transformer:
    - Fixed-length and limited input
    - Quadratic complexity for self-attention

- Solution: variable-length transformers

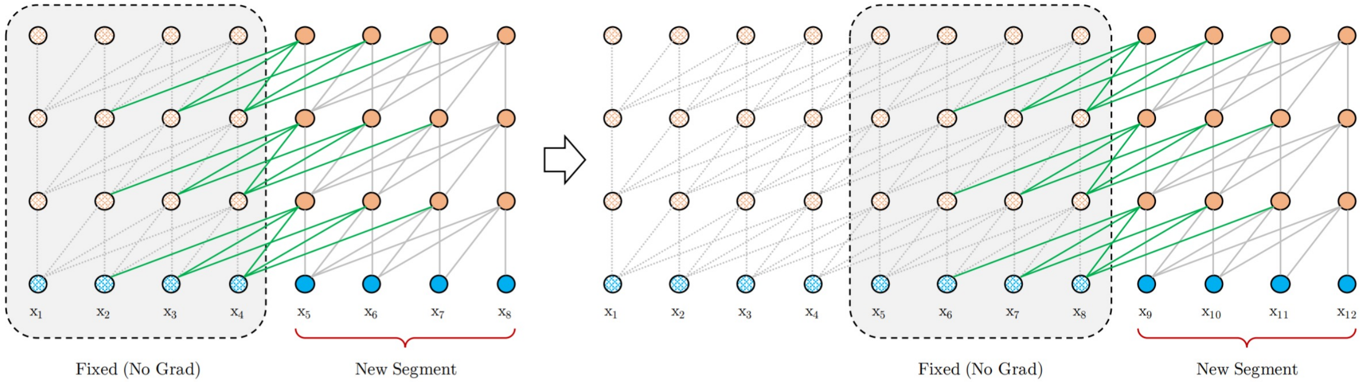
BACKGROUND: VARIABLE-LENGTH TRANSFORMER

XLNET / TRANSFORMER-XL: RECURRENCE MECHANISM FOR SELF-ATTENTION

- Vanilla Transformer: simply truncate long text, causing context fragmentation
- Transformer-XL: each previous segment is cached and reused in the next segment



Vanilla Transformer



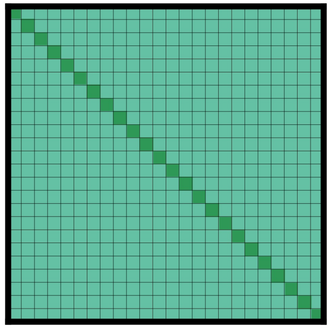
Transformer-XL

Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." *arXiv preprint arXiv:1901.02860* (2019).

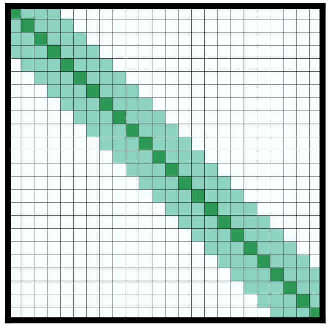
BACKGROUND: VARIABLE-LENGTH TRANSFORMER

LONGFORMER: WINDOWED/SPARSE ATTENTION

- Vanilla Transformer: A single global attention, causing terrible complexity
- Longformer: Windowed/sparse attention, more efficient



(a) Full n^2 attention



(b) Sliding window attention

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).

DECODER

- Per-label attention: let each ICD code attend to different parts of the token sequence
  - $H \in \mathbb{R}^{N \times d_f}$: last hidden state from the encoder
  - $Q \in \mathbb{R}^{d_f \times C}$: query matrix of the ICD codes

$$A = softmax(HQ) \in \mathbb{R}^{N \times C}$$

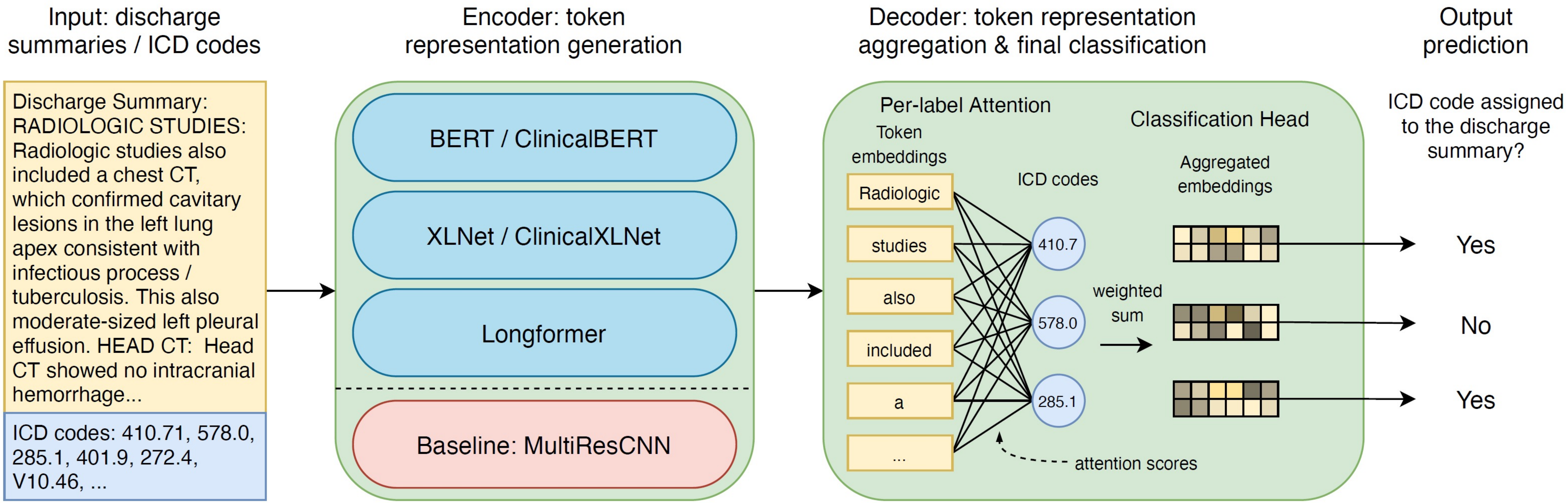
$$V = A^T H \in \mathbb{R}^{C \times d_f}$$

- Classification head: final linear layer
- Learning objective: binary cross-entropy loss

$$\mathcal{L} = -\sum_i y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)$$

UNIVERSITY OF TORONTO

OVERALL ARCHITECTURE



EXPERIMENT

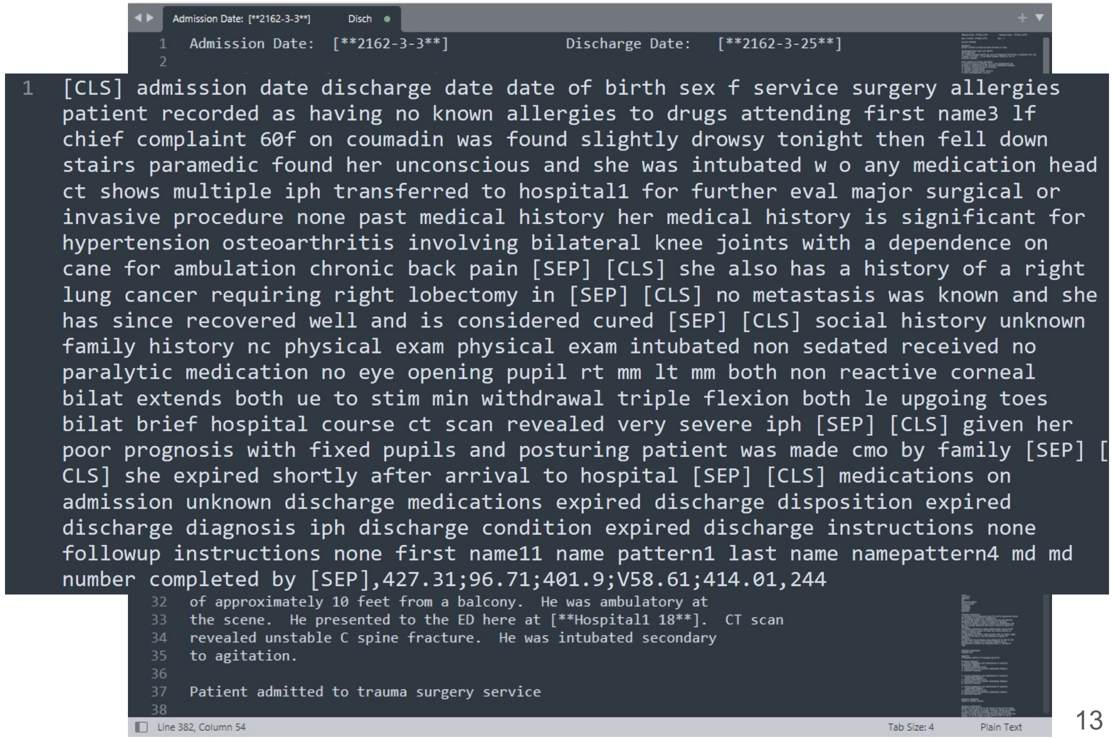
DATASET

- **MIMIC-III: Medical Information Mart for Intensive Care**
  - Focusing only on discharge summaries
  - Including medical history, medications, laboratory reports, hospital course, diagnoses, follow up plans
  - Every summary related to one or more ICD-9 codes
- **Preprocessing**
  - Removed numbers and punctuations
  - Lowercases

EVALUATION METRICS

- Micro/Macro AUC and F1 scores
- Precision at K (P@K)
  - P@8 and P@15 for full codes, P@5 for top 50 codes

AVG. Word Count	1513.51
Discharge Summaries	52,722
Top 50 Codes Summaries	8,067
Code Types	8,929
AVG. Codes Per Summary	15.9



EXPERIMENT

IMPLEMENTATION AND HYPER-PARAMETER SETTINGS

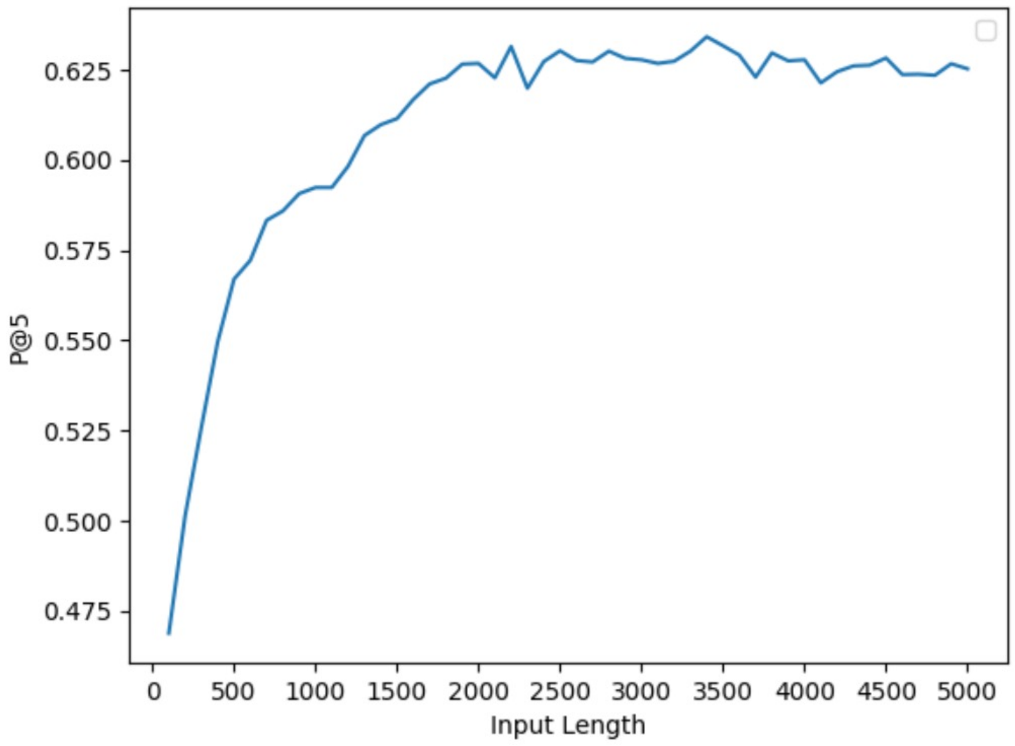
- **Reproduced MultiResCNN**
  - All settings are the same
  - N=2500 originally, N=512 for comparison with BERT
- **6 pre-trained Transformer-based models**
  - Same architectural hyper-parameters as designed

Encoder Model	Architecture	Input Sequence Length	Pretrained Dataset
MultiResCNN-512	CNN	512	-
MultiResCNN-2500	CNN	2500	-
BERT-512	BERT	512	Generic
ClinicalBERT-512	BERT	512	Clinical notes
XLNet-1500	XLNet	1500	Generic
ClinicalXLNet-1500	XLNet	1500	Clinical discharge summaries
Longformer-2500	Longformer	2500	Generic
Longformer-3200	Longformer	3200	Generic

EXPERIMENT

INPUT LENGTH SENSITIVITY EXPERIMENT

- CNN Input Length Experiment



- Fixed vocabulary & WordPiece Tokenizer

  => partitions new words into sub-words: "Ibuprofen" -> "ibu" "pro" "fen"

  => less information if N remains the same

  => larger N should be assigned

EXPERIMENT

BASELINES AND RESULTS

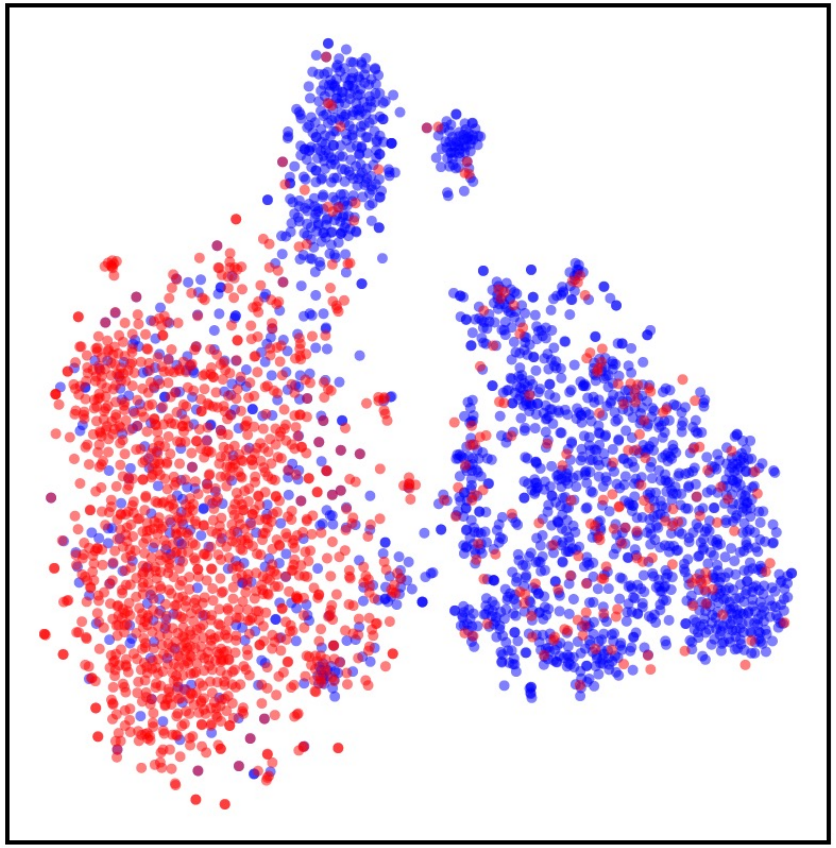
Table 3: MIMIC-III baseline results (top-50 codes). Bolded results indicate the best, while underlined results indicate the second best.

Model	AUC Micro	AUC Macro	F1 Micro	F1 Macro	P@K 5
CAML	0.909	0.875	0.614	0.532	0.609
DR-CAML	0.916	0.884	0.633	0.576	0.618
MultiResCNN-512	0.878	0.839	0.569	0.484	0.570
MultiResCNN-2500	0.923	0.895	0.655	0.594	0.620
BERT-512	0.865	0.831	0.539	0.458	0.545
ClinicalBERT-512	0.887	0.858	0.589	0.507	0.581
XLNet-1500	0.904	0.875	0.622	0.542	0.609
Longformer-2500	<u>0.928</u>	<u>0.901</u>	<u>0.678</u>	<u>0.606</u>	<u>0.642</u>
Longformer-3200	**0.931**	**0.905**	**0.689**	**0.631**	**0.651**

LATENT SPACE VISUALIZATION

- To interpret our experimental results, we visualized the aggregated embedding for ICD code 401.9 (Unspecified essential hypertension)
  - 3372 total instances, in which 1441 instances contain this code
  - Taking the corresponding row from the value matrix V as the latent feature
  - Visualization is done using t-SNE

(a) MultiResCNN-512

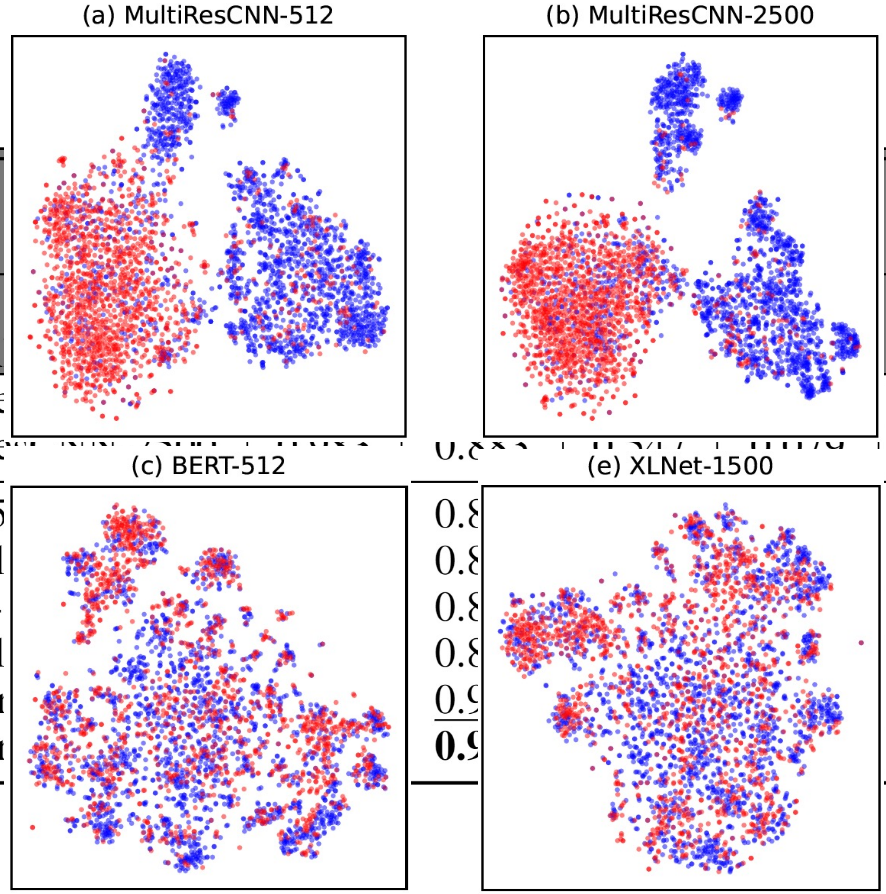


DISCUSSION

HOW WILL THE INPUT SEQUENCE LENGTH AFFECT THE ICD CODING PERFORMANCE?

- For both CNN-based and transformer-based models, increasing input sequence length results in performance improvement
  - More information input to the model
  - Each token can attend to more context

- BERT & XLNet model failed at producing embeddings with enough discrepancy
  - Potential underfitting
  - Wrong hyperparameter settings

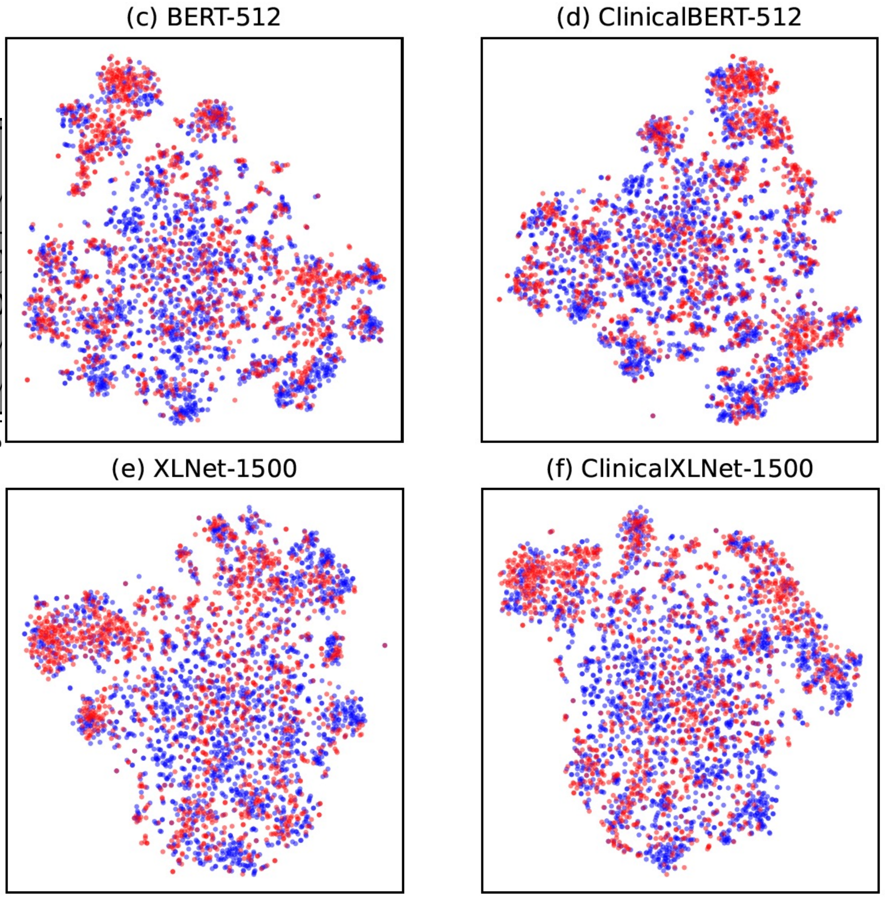


Model		P@K 8	P@K 15
CAML		0.709	0.561
DR-CA		0.690	0.548
MultiRe		0.659	0.504
MultiRe	0.8	0.724	0.572
BERT-5	0.8	0.556	0.414
Clinical	0.8	0.589	0.442
XLNet-	0.8	0.640	0.484
Clinical	0.8	0.644	0.488
Longfor	0.9	0.734	0.546
Longfor	**0.9**	**0.746**	**0.592**

DISCUSSION

WILL THE APPLICATION OF DOMAIN-SPECIFIC LANGUAGE MODELS, SUCH AS CLINICAL/MEDICAL RELATED LANGUAGE MODELS, INCREASE THE MODEL PERFORMANCE?

- Both ClinicalBERT and ClinicalXLNet models outperform the original BERT and XLNet models
    - Finetuning a transformer from a checkpoint that is pre-trained on domain-specific datasets benefits the performance of the ICD coding task
- Applying domain-specific language models is not enough to avoid the underfitting problem



M	P@K 8	P@K 15
C	0.709	0.561
D	0.690	0.548
M	0.659	0.504
M	0.724	0.572
F	0.556	0.414
C	0.589	0.442
X	0.640	0.484
C	0.644	0.488
L	0.734	0.546
L	**0.746**	**0.592**

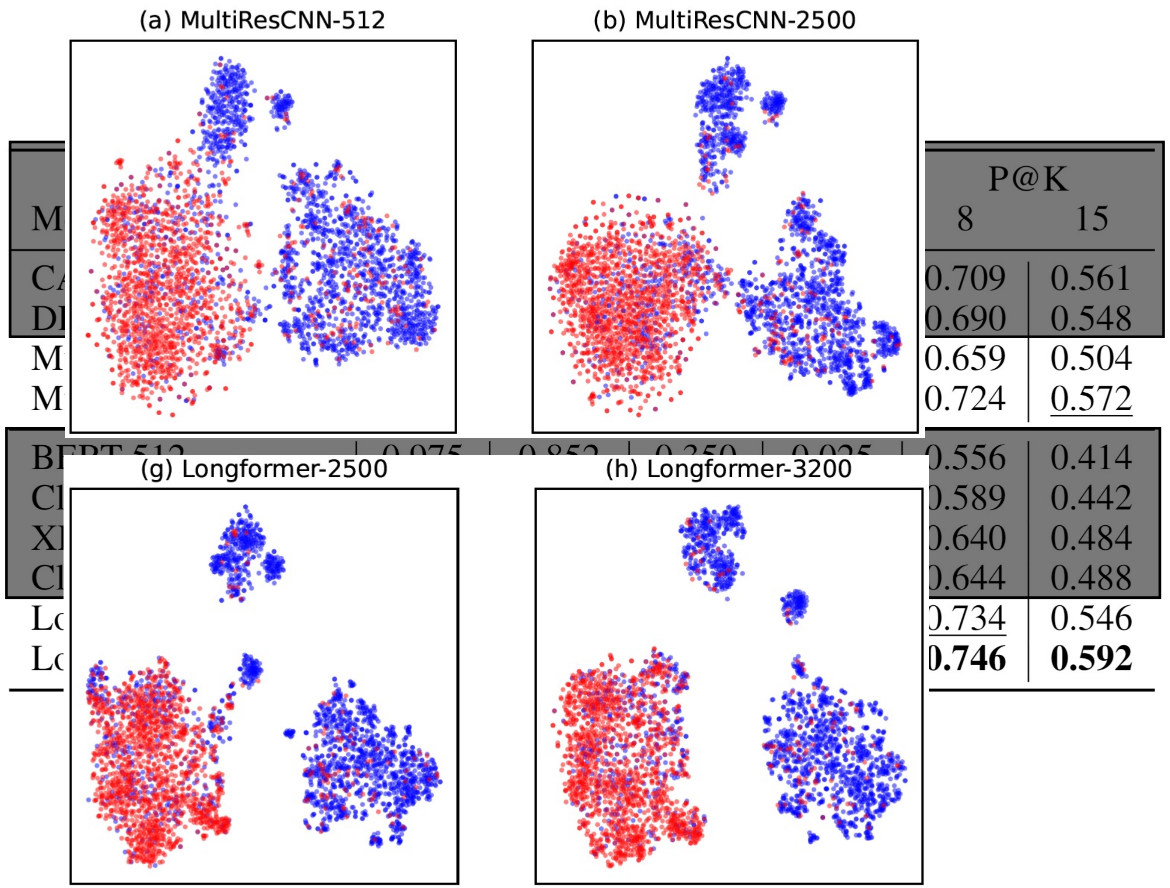
UNIVERSITY OF TORONTO

DISCUSSION

WILL THE DEEP AND COMPLEX ARCHITECTURE OF THE TRANSFORMER-BASED MODELS BENEFIT THE OUTPUT REPRESENTATION OF THE CLINICAL NOTES AND RESULT IN ENHANCED ICD CODING PERFORMANCE?

- Our Longformer model performs better than the MultiResCNN model under the same input sequence length restriction

- Major performance improvement: better output text representations
  - Clusters are denser
  - Margins between clusters are larger



CONCLUSION

- We identified three key characteristics for transformer-based automated ICD coding
  - Input sequence length impacts heavily on model performance
  - Pretraining using clinical texts benefits model performance
  - Deep transformer architecture generates better contextual representations
- Our Longformer model reaches state-of-the-art performance and can act as a strong baseline for transformer-based ICD coding and other related clinical NLP tasks

Limitations

- Diagnoses often show up in the last part of discharge summaries – directly truncate the input text may not be a good idea
- Long document transformers are still computationally expensive – are there more efficient solutions?

FUTURE WORK

- Improve preprocessing methods for raw text
  - Try different text splitting strategies

- Pre-train Longformer using clinical texts
  - Try different unsupervised pre-training methods

- Extend and validate our model to other clinical NLP tasks
  - E.g. discharge readmission prediction

UNIVERSITY OF TORONTO

THANK YOU FOR YOUR ATTENTION!