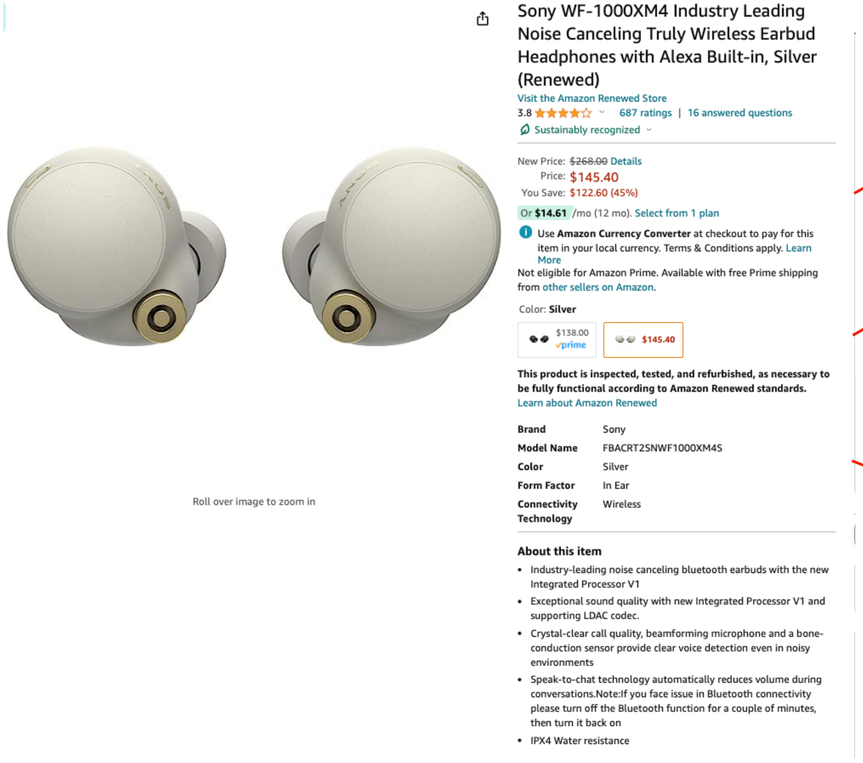


- Earbuds
- Headphone
- Computer
- Clothing
- Snacks
- …
- Personal Care

Product Advertising

Don’t Just Pay Attention, PLANT It

Transfer L2R Models to Fine-tune Attention in Extreme Multi-Label Text Classification

Debjyoti Saharoy, Javed Aslam

Khoury College of Computer Sciences, Northeastern University

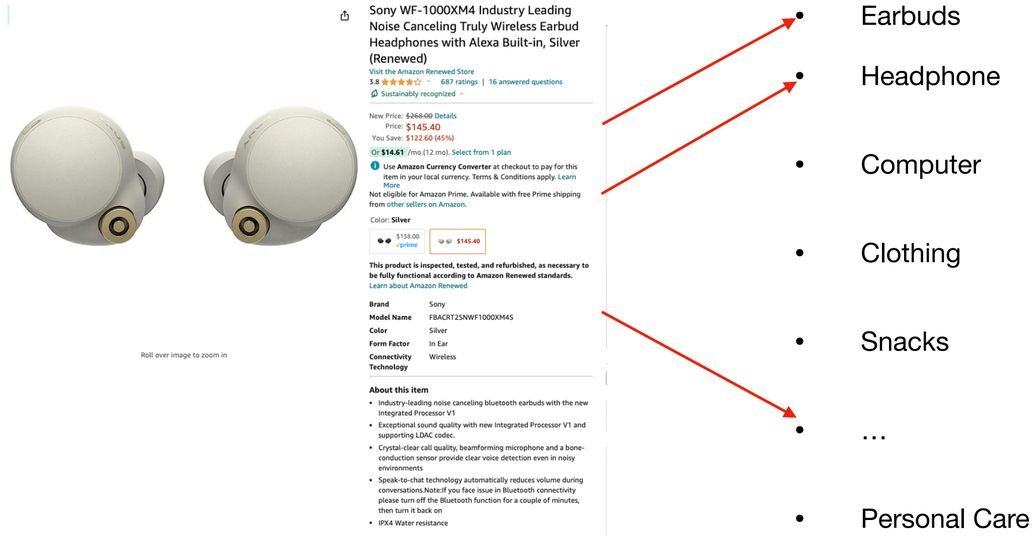
Outline

- Problem
- Shortcomings hitherto…
- Our Method
- Results and Improvements
- Conclusion

Problem: Extreme Multi-Label Text Classification (XMTC)

- Tagging a text with most relevant subset of labels from an extremely large -hundreds of thousands/millions, label set

Advertising



- Earbuds
- Headphone
- Computer
- Clothing
- Snacks
- …
- Personal Care

instance

label space

Healthcare (ICD)

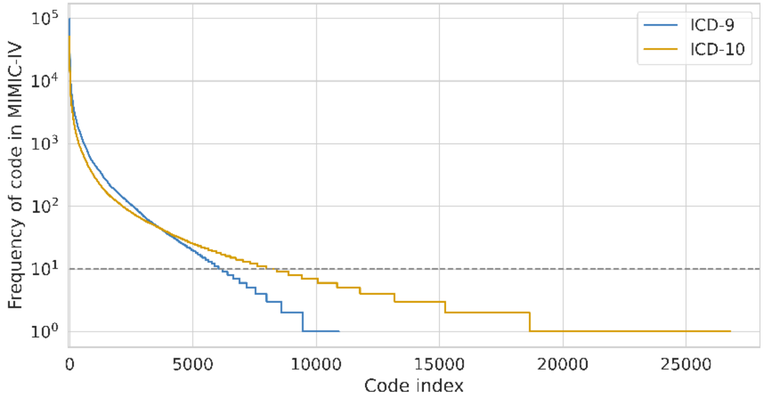
Discharge Summary

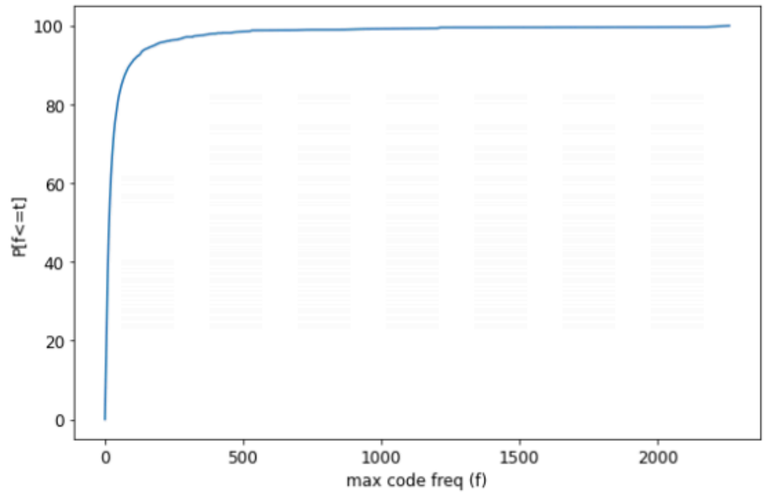
History of Present Illness:
This is a 54-year-old gentleman with a history of deep venous thrombosis and pulmonary embolism in [**2155**], history of hypertension, and atrial fibrillation who presented to the Emergency Department with extreme dyspnea on exertion and weakness.
.....
Major Surgical or Invasive Procedure:
Attempt Repair of Ruptured Aortic Aneurysm
.....
Brief Hospital Course:
The patient was brought to the cardiac catheterization laboratory and a right heart catheterization revealed a pulmonary artery pressure of 49/17, right ventricular pressure
.....
PERTINENT RADIOLOGY/IMAGING:
Electrocardiogram revealed sinus tachycardia with a rate of 120. Intervals were otherwise normal. He had poor R wave progression. No ST changes.
.....
Discharge Diagnosis:
1. Pulmonary embolism; presenting as cardiogenic shock. 2. Atrial fibrillation/atrial flutter. 3. Renal insufficiency.
.....

ICD Codes
37.21: Right Heart Cardiac Catheterization
38.7: Insertion of Intraluminal Device into Superior Vena Cava
88.52: Plain Radiography of Right Heart using High Osmolar Contrast
89.8: Autopsy
99.62: Restoration of Cardiac Rhythm

Problem: Main Challenge in XMTC

1. Datasets consists of texts with multiple lengthy narratives - However, only a small fraction of tokens are most informative with regard to assigning relevant codes
    - MIMIC-III/IV: more than 1500 tokens on average.
2. Code space is extremely high dimensional
    - 18000 and 170000 codes in ICD-9-CM and ICD-10-CM/PCS
3. Code distribution is heavily skewed
    - MIMIC-III: ~5411 out of 8929 codes appear < 10 times





Outline

- Problem
- Shortcomings hitherto…
- Our Method
- Results and Improvements
- Conclusion

Discharge Summary

History of Present Illness:
This is a 54-year-old gentleman with a history of deep venous thrombosis and pulmonary embolism in [**2155**], history of hypertension, and atrial fibrillation who presented to the Emergency Department with extreme dyspnea on exertion and weakness.
.....
Major Surgical or Invasive Procedure:
Attempt Repair of Ruptured Aortic Aneurysm
.....
Brief Hospital Course:
The patient was brought to the cardiac catheterization laboratory and a right heart catheterization revealed a pulmonary artery pressure of 49/17, right ventricular pressure
.....
PERTINENT RADIOLOGY/IMAGING:
Electrocardiogram revealed sinus tachycardia with a rate of 120. Intervals were otherwise normal. He had poor R wave progression. No ST changes.
.....
Discharge Diagnosis:
1. Pulmonary embolism; presenting as cardiogenic shock. 2. Atrial fibrillation/atrial flutter. 3. Renal insufficiency.
.....

ICD Codes
37.21: Right Heart Cardiac Catheterization
38.7: Insertion of Intraluminal Device into Superior Vena Cava
88.52: Plain Radiography of Right Heart using High Osmolar Contrast
89.8: Autopsy
99.62: Restoration of Cardiac Rhythm

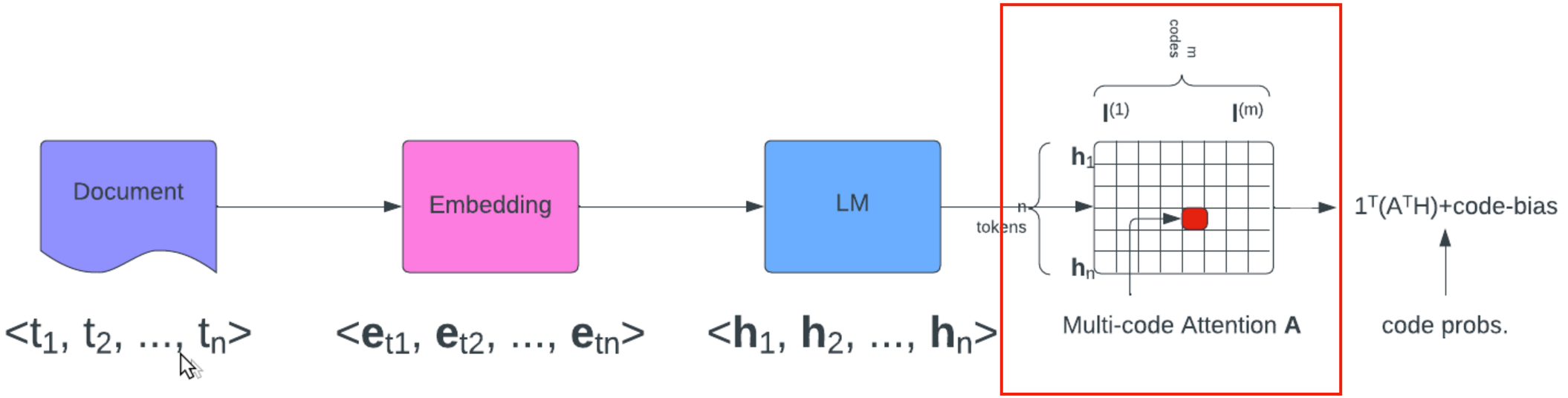
So far…

- *Attention* is crucial in any SOTA NLP model
  - Q/A[6], Translation[1,2,6], Summarization[3], Representation[4], Sentiment Analysis[5]
- Attention vital in XMTC

	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

CAML (ACL ’18)[7]

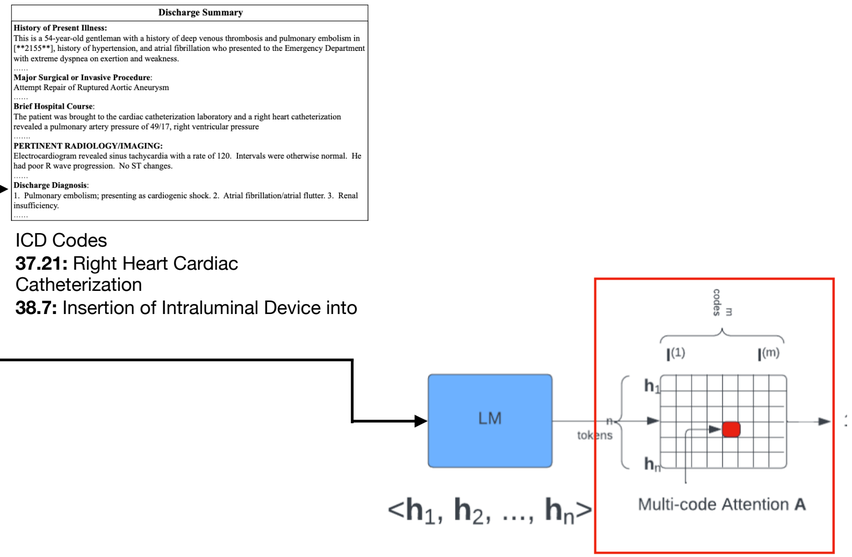
So far…: Attention!!!



- Multi-code Attention: learns code specific attention to most informative tokens
- CAML (ACL ’18)[7], MSATT-KG (CIKM ’19)[8], MultiResCNN (AAAI ’20)[9], Hyper-Core (ACL ’20)[10], LAAT (IJCAI ’20)[11], ISD (ACL ’21)[12], Effective-CAN (EMNLP ’21)[13], MSMN (ACL ’22)[14], DiscNet (ACL ’22)[15], KEPTLongformer (EMNLP ’22)[16]
- Shortcoming: learning token relevance in relation to numerous codes results in lengthy training and overfitting.

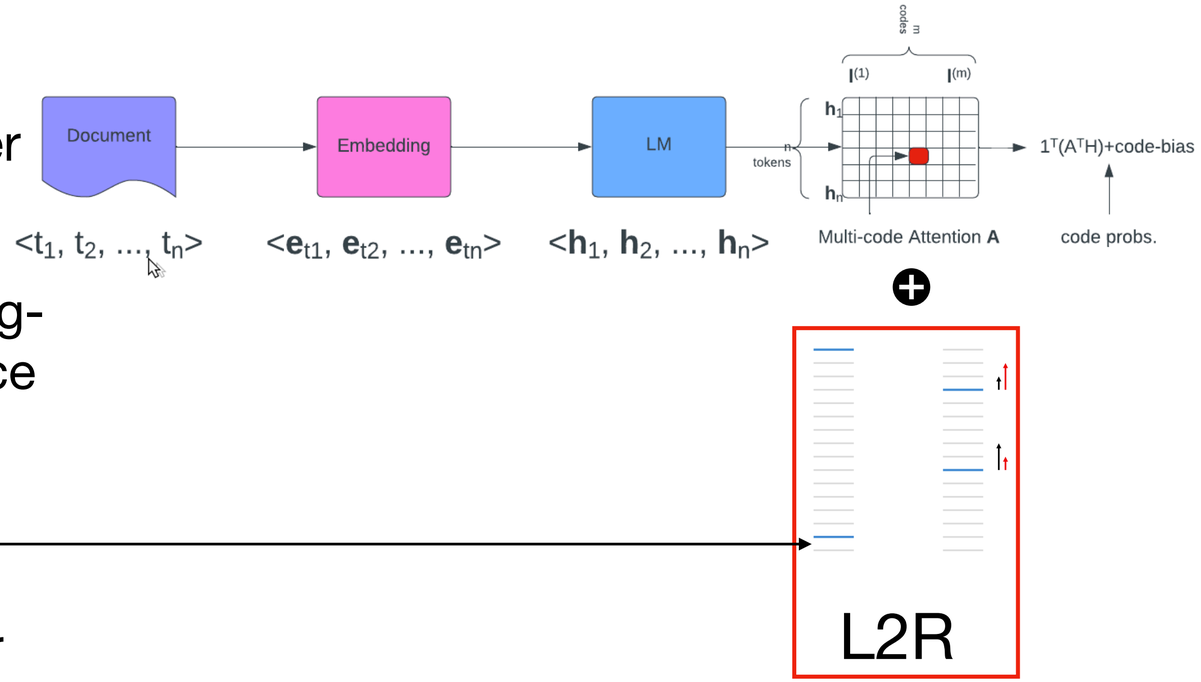
Outline

- Problem
- Shortcomings hitherto…
- Our Method
- Results and Improvements
- Conclusion



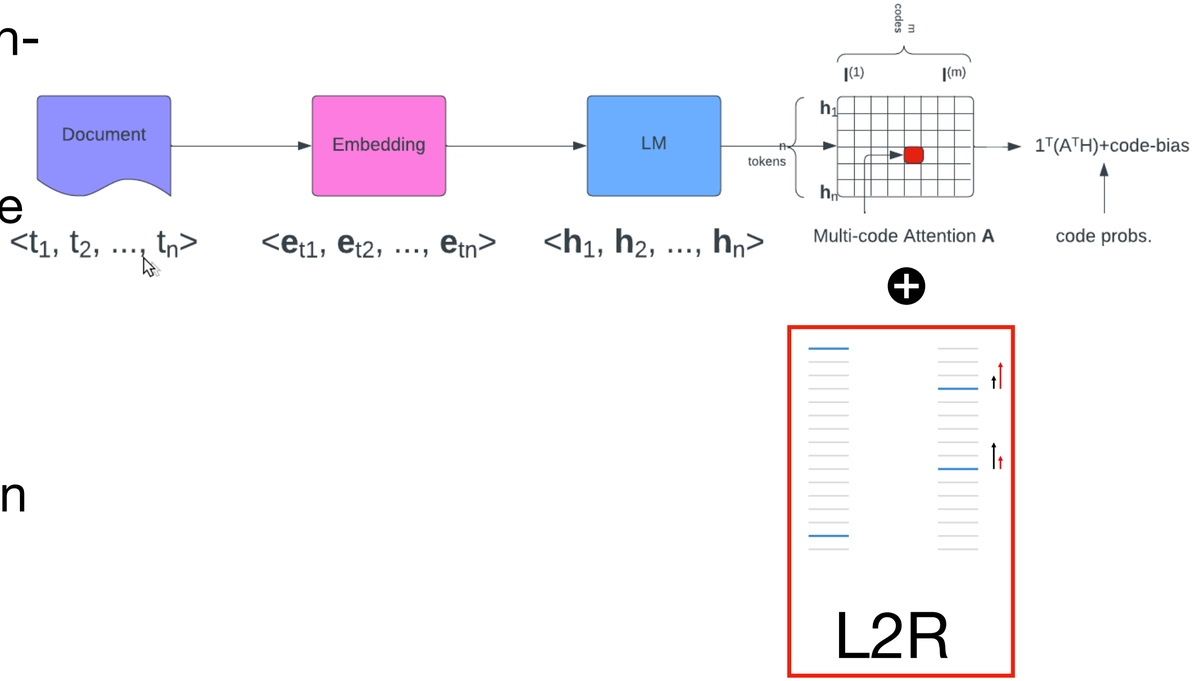
Our Method: PLANT

- **P**retrained & **L**everaged **A**tte**NT**ion: novel transfer learning to fine-tune attention in XMTC
- Capture token-code dynamics: pre-train Learning-to-Rank (L2R) model that ranks token significance by code relevance
- L2R activations as PLANTed attention
- Benefit: Start with well-informed attn wgts rather than random, then fine-tune



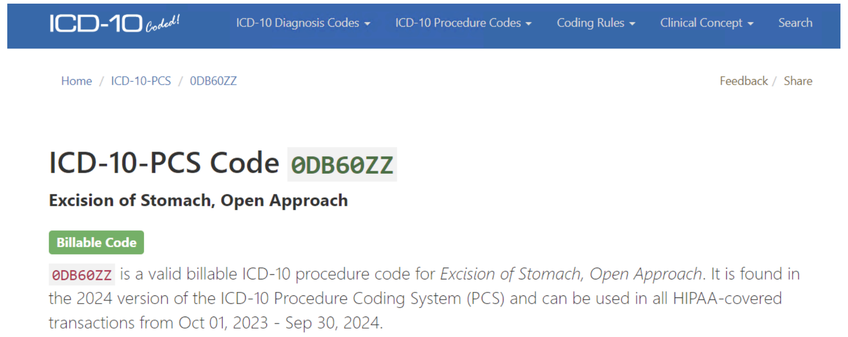
Our Method: PLANT

- **P**retrained & **L**everaged **A**tte**NT**ion: capture token-label dynamics with L2R model.
- Goal of L2R: Embed tokens to enhance relevance in sorting by taking dot products with label embeddings.
- Key step 1: Bootstrap “How much can I infer certain code from a token?”-> Mutual-Information Gain
- Key step 2: Max “How good is the code specific token list wrt top k tokens?” -> LambdaRank[17] /w nDCG@k



Our Method: Why PLANT works?

```
+ ('antecolic', 0.10208515)
+ ('blumgart', 0.07633876)
+ ('pancreatojejunostomy', 0.054873332)
+ ('gastrojejunostomy', 0.050927494)
+ ('fiducials', 0.05050926)
+ ('pancreaticojejunostomy', 0.04120218)
+ ('bloc', 0.039802585)
+ ('spirrometry', 0.038191367)
+ ('pancreaticoduodenectomy', 0.03692116)
+ ('splenoportal', 0.03632862)
+ ('gastrectomy', 0.03271102)
+ ('polya', 0.029407317)
+ ('modifier', 0.029353755)
+ ('hepaticojejunostomy', 0.028313298)
+ ('transgastric', 0.028199965)
+ ('duodenectomy', 0.023473764)
+ ('instruct', 0.023241024)
+ ('jejunostomy', 0.021264277)
+ ('gastrojejunal', 0.020096961)
+ ('resectable', 0.019506555)
```

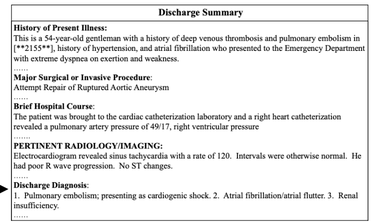



```
+ ('antecolic', 117.187805)
+ ('pancreatojejunostomy', 116.8118)
+ ('blumgart', 105.97807)
+ ('undone', 103.58473)
+ ('splenoportal', 102.52783)
+ ('duodenectomy', 101.54767)
+ ('pancreaticoduodenectomy', 99.660324)
+ ('witzel', 99.13038)
+ ('adenomatosis', 98.637375)
+ ('spirrometry', 98.34472)
+ ('jejunostomy', 97.52427)
+ ('transgastric', 97.08511)
+ ('gastrojejunostomy', 97.03317)
+ ('pancreaticojejunostomy', 96.57304)
+ ('pancreatoduodenectomy', 95.53003)
+ ('jtube', 94.991776)
+ ('tubefeed', 93.88675)
+ ('cleard', 93.40699)
+ ('gastrogastric', 93.329994)
+ ('bloc', 92.97085)
+ ('esophagogastrectomy', 92.6966)
+ ('duodenojejunostomy', 92.550865)
+ ('fiducials', 92.35274)
+ ('trophic', 92.18849)
+ ('instruct', 92.13136)
+ ('efferent', 92.1006)
+ ('gastrojejunal', 91.979706)
+ ('basion', 91.517975)
+ ('hpb', 91.27107)
+ ('modifier', 90.64114)
+ ('enteroatmospheric', 90.56534)
+ ('gastrectomy', 90.19148)
+ ('ethylsuccinate', 90.10019)
+ ('dpca', 90.03682)
+ ('attaching', 90.016174)
+ ('periesophageal', 89.87409)
+ ('punctured', 89.7901)
+ ('malecot', 89.70523)
+ ('hepaticojejunostomy', 89.70446)
+ ('folloiwng', 89.670815)
```

- Key Step 1: Mutual-Info gain 👍
- Key Step 2: As goes L2R so goes XMTC 👍

Outline

- Problem
- Shortcomings hitherto…
- Our Method
- Results and Improvements
- Conclusion



Discharge Summary

History of Present Illness:
This is a 54-year-old gentleman with a history of deep venous thrombosis and pulmonary embolism in [**2155**], history of hypertension, and atrial fibrillation who presented to the Emergency Department with extreme dyspnea on exertion and weakness.
.....

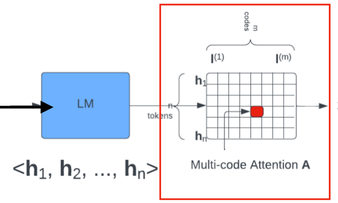
Major Surgical or Invasive Procedure:
Attempt Repair of Ruptured Aortic Aneurysm
.....

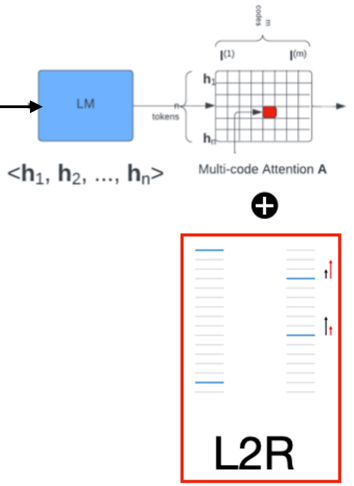
Brief Hospital Course:
The patient was brought to the cardiac catheterization laboratory and a right heart catheterization revealed a pulmonary artery pressure of 49/17, right ventricular pressure
.....

PERTINENT RADIOLOGY/IMAGING:
Electrocardiogram revealed sinus tachycardia with a rate of 120. Intervals were otherwise normal. He had poor R wave progression. No ST changes.
.....

Discharge Diagnosis:
1. Pulmonary embolism; presenting as cardiogenic shock. 2. Atrial fibrillation/atrial flutter. 3. Renal insufficiency.
.....

ICD Codes
37.21: Right Heart Cardiac Catheterization
38.7: Insertion of Intraluminal Device into





Results and Improvements: Metrics

- Precision: Out of all the items the model predicted as positive, how many were actually correct?
- Recall: Out of all the actual positive items, how many did the model correctly identify as positive?
- F1: balance conservatism (high precision but low recall) and liberalism (high recall but low precision)

Results: Improvements

Model	AUC		F1	
	Macro	Micro	Macro	Micro
MSMN	75.3	76.2	17.1	17.2
KEPTLongformer	82.7	83.3	30.4	32.6
PLANT (Ours)	**95.6***	**96.0***	**82.6***	**84.2***

MIMIC-III-rare50

Model	AUC		F1		P@k		
	Macro	Micro	Macro	Micro	P@5	P@8	P@15
CAML/DR-CAML	88.4	91.6	57.6	63.3	61.8	-	-
MSATT-KG	91.4	93.6	63.8	68.4	64.4	-	-
MultiResCNN	89.9	92.8	60.6	67.0	64.1	-	-
HyperCore	89.5	92.9	60.9	66.3	63.2	-	-
LAAT/JointLAAT	92.5	94.6	66.6	71.6	67.5	54.7	35.7
ISD	**93.5**	**94.9**	67.9	71.7	68.2	-	-
Effective-CAN	92.0	94.5	66.8	71.7	66.4	-	-
MSMN	92.8	94.7	68.3	72.5	**68.0**	-	-
PLANT (Ours)	93.1	**94.9**	**68.7**	**72.8**	67.2	**55.0***	**36.3***

MIMIC-III-top50

Model	AUC		F1		P@k		
	Macro	Micro	Macro	Micro	P@5	P@8	P@15
CAML/DR-CAML	89.7	98.6	8.8	53.9	-	70.9	56.1
MSATT-KG	91.0	99.2	9.0	55.3	-	72.8	58.1
MultiResCNN	91.0	98.6	8.5	55.2	-	73.4	58.4
HyperCore	93.0	98.9	9.0	55.1	-	72.2	57.9
LAAT/JointLAAT	92.1	98.8	10.7	57.5	81.3	73.8	59.1
ISD	93.8	99.0	11.9	55.9	-	74.5	-
Effective-CAN	92.1	98.9	10.6	58.9	-	75.8	60.6
MSMN	95.0	99.2	10.3	58.4	-	75.2	59.9
DiscNet	**95.6**	**99.3**	**14.0**	58.8	-	76.5	61.4
PLANT (Ours)	90.4	98.9	10.1	**59.4***	**84.0***	**77.1***	**61.7***

MIMIC-III-full

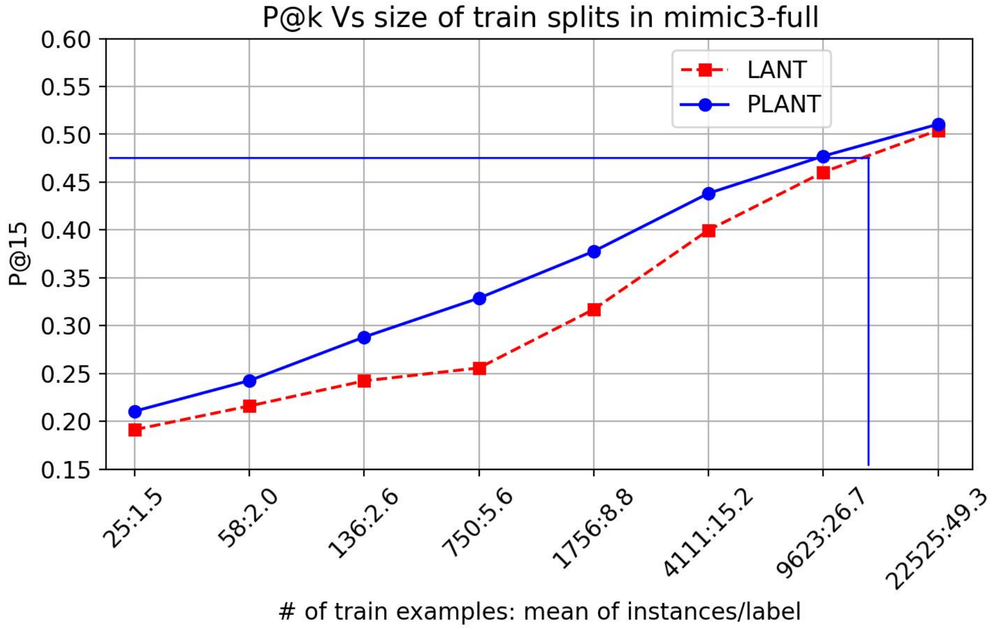
Model	AUC		F1		P@k		
	Macro	Micro	Macro	Micro	P@5	P@8	P@15
CAML/DR-CAML	91.1	98.5	16.0	55.4	-	66.8	52.2
MultiResCNN	94.5	**99.0**	21.1	56.9	-	67.8	53.5
LAAT/JointLAAT	**95.4**	**99.0**	20.3	57.9	-	68.9	54.3
PLANT (Ours)	94.8	**99.0**	19.6	57.1	**78.1***	**70.6***	**55.6***

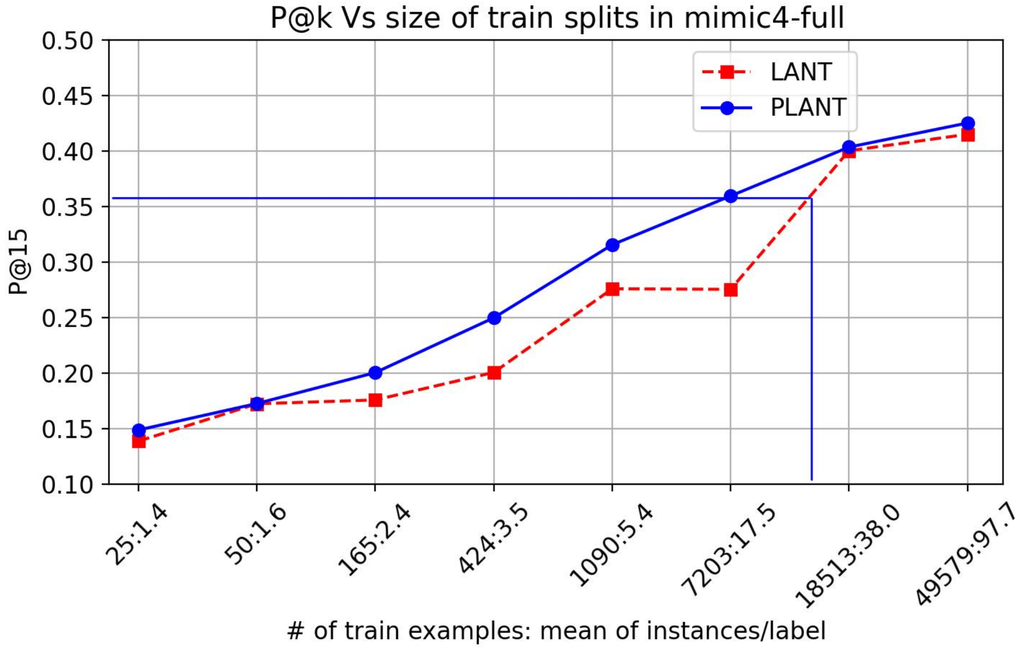
MIMIC4-IV-full

- *Precision:* Out of all the items the model predicted as positive, how many were actually correct?
- *Recall:* Out of all the actual positive items, how many did the model correctly identify as positive?
- *F1:* balance conservatism (high precision but low recall) and liberalism (high recall but low precision)

* indicates that the performance difference between PLANT and the next best is significant (, using the Approximate Randomization test).

Results: Juice is worth the squeeze!





P@15 for PLANT vs LANT (Learned Attention) with different # of training examples on MIMIC-III-full and MIMIC-IV-full.

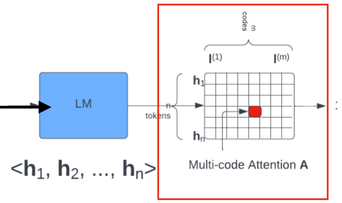
Outline

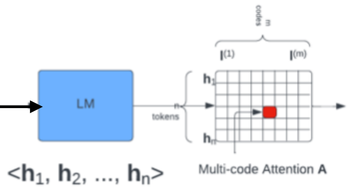
- Problem
- Shortcomings hitherto…
- Our Method
- Results and Improvements
- Conclusion

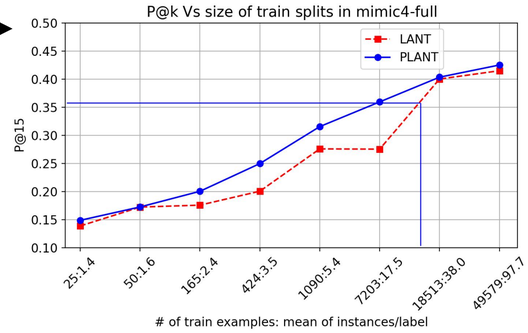
Discharge Summary

History of Present Illness:
This is a 54-year-old gentleman with a history of deep venous thrombosis and pulmonary embolism in [**2155**], history of hypertension, and atrial fibrillation who presented to the Emergency Department with extreme dyspnea on exertion and weakness.
.....
Major Surgical or Invasive Procedure:
Attempt Repair of Ruptured Aortic Aneurysm
.....
Brief Hospital Course:
The patient was brought to the cardiac catheterization laboratory and a right heart catheterization revealed a pulmonary artery pressure of 49/17, right ventricular pressure
.....
PERTINENT RADIOLOGY/IMAGING:
Electrocardiogram revealed sinus tachycardia with a rate of 120. Intervals were otherwise normal. He had poor R wave progression. No ST changes.
.....
Discharge Diagnosis:
1. Pulmonary embolism; presenting as cardiogenic shock. 2. Atrial fibrillation/atrial flutter. 3. Renal insufficiency.
.....

ICD Codes
37.21: Right Heart Cardiac Catheterization
38.7: Insertion of Intraluminal Device into







Conclusion

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

2. Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4263–4272, Brussels, Belgium. Association for Computational Linguistics.

3. Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xi- aodong He, and Bowen Zhou. 2020. Self-attention guided copy mechanism for abstractive summarization. In Proceedings of the 58th annual meeting of the association for computational linguistics, pages 1355–1362.

4. Douwe Kiela, Changhan Wang, and Kyunghyun Cho. 2018. Dynamic meta-embeddings for improved sen- tence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Lan- guage Processing, pages 1466–1477, Brussels, Bel- gium. Association for Computational Linguistics.

5. Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention net- work for aspect-based sentiment analysis. In Pro- ceedings of the 58th Annual Meeting of the Asso- ciation for Computational Linguistics, pages 3229– 3238, Online. Association for Computational Lin- guistics.

6. Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. arXiv preprint arXiv:1807.03819.

7. James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable predic- tion of medical codes from clinical text. In Proceed- ings of the 2018 Conference of the North American Chapter of the Association for Computational Lin- guistics: Human Language Technologies, Volume 1 (Long Papers), pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

8. Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. Ehr coding with multi-scale feature at- tention and structured knowledge graph propagation. In Proceedings of the 28th ACM International Con- ference on Information and Knowledge Management, CIKM '19, page 649–658, New York, NY, USA. As- sociation for Computing Machinery.

9. Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In proceedings of the AAAI conference on artificial intelligence, volume 34, pages 8180–8187.

10. Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Sheng- ping Liu, and Weifeng Chong. 2020. HyperCore: Hy- perbolic and co-graph representation for automatic ICD coding. In Proceedings of the 58th Annual Meet- ing of the Association for Computational Linguistics, pages 3105–3114, Online. Association for Computa- tional Linguistics.

References

11. Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for icd coding from clinical text. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelli- gence, IJCAI'20.

12. Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic ICD coding via interactive shared representation networks with self-distillation mech- anism. In Proceedings of the 59th Annual Meet- ing of the Association for Computational Linguistics and the 11th International Joint Conference on Natu- ral Language Processing (Volume 1: Long Papers), pages 5948–5957, Online. Association for Computa- tional Linguistics.

13. Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. 2021. Effective con- volutional attention network for multi-label clinical document classification. In Proceedings of the 2021 Conference on Empirical Methods in Natural Lan- guage Processing, pages 5941–5953, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

14. Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In Proceedings of the 60th Annual Meeting of the As- sociation for Computational Linguistics (Volume 2: Short Papers), pages 808–814, Dublin, Ireland. As- sociation for Computational Linguistics.

15. Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. 2022. Automatic ICD cod- ing exploiting discourse structure and reconciled code embeddings. In Proceedings of the 29th Inter- national Conference on Computational Linguistics, pages 2883–2891, Gyeongju, Republic of Korea. In- ternational Committee on Computational Linguistics.

16. Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few- shot icd coding. In Proceedings of the Conference on Empirical Methods in Natural Language Process- ing. Conference on Empirical Methods in Natural Language Processing, volume 2022, page 1767. NIH Public Access.

17. Christopher JC Burges. 2010. From ranknet to lamb- darank to lambdamart: An overview. Learning, 11(23-581):81.