Medical ICD9 Code Prediction on MIMIC-III

Robert Sato
University of California, Santa Cruz
1156 High Street, Santa Cruz, CA 95064
rssato@ucsc.edu

Abstract

This work looks at the task of medical code prediction. This work originally aimed to focus on MIMIC-IV but unfortunately, all of their data has been released except the discharge summaries which are exactly what this work focuses on. This work experiments with different model architectures to apply to an extreme multi-label classification problem. We employ a recently released state of the art BERT model (clinical BERT) from pre-trained token embeddings and a classification layer on top to achieve competitive results.

1 Introduction

1.1 Motivation

Medical code prediction is a time consuming, error prone and monotonous task. This task involves assigning International Classification of Diseases (ICD) codes to notes. In this work, we focus on discharge summaries written by clinicians. The summaries are often riddled with misspellings, shorthand abbreviations and unimportant information. A medical coder must read these texts and assign, from the 6918 diagnoses codes, all codes that relate to the discharge summary. Automating this task would be beneficial for health care providers and the average person as there will be lower medical costs and more transparency around medical billing.

1.2 Novel Contributions

In this work, we tackle the extreme multi-label task of assigning multiple labels from the large label space of ICD9 codes. As ICD10 has 14,400 codes and ICD11 has 55,000 codes (as reported by the WHO), it is crucial to consider this task with the large label space. My work uses a fundamental transfer learning approach with some fine tuning to achieve close to the state of the art performance on this task. There is no domain specific knowledge, extensive pre-processing or prohibitively long training times applied in this work. I use a BERT model pre-trained on MIMIC-III for similar tasks and employ fully connected layers with dropout to achieve a micro AUC score of 95.6.

1.3 Data

The Medical Information Mart for Intensive Care (MIMIC-III) [5] database consists of a swath of critical care information. This data was recorded from 2010 to 2016 with newer editions (MIMIC-IV) being released soon. This single database holds information on tests, orders, billing and much more as shown in figure 1. This data goes through a process of de-identification and date shifting to ensure privacy for the participants. In this project, we focus on the discharge summaries and ICD9 codes assigned to these summaries. These texts range in length with an average length of 1552 words and standard deviation of 781 words. ICD9 codes for these texts also range significantly with an average of 11.7 ICD codes per text. Over 60% of the texts have less than 10 ICD codes assigned to it making the label space very sparse. Figures 2 and 3 display this information on the data.

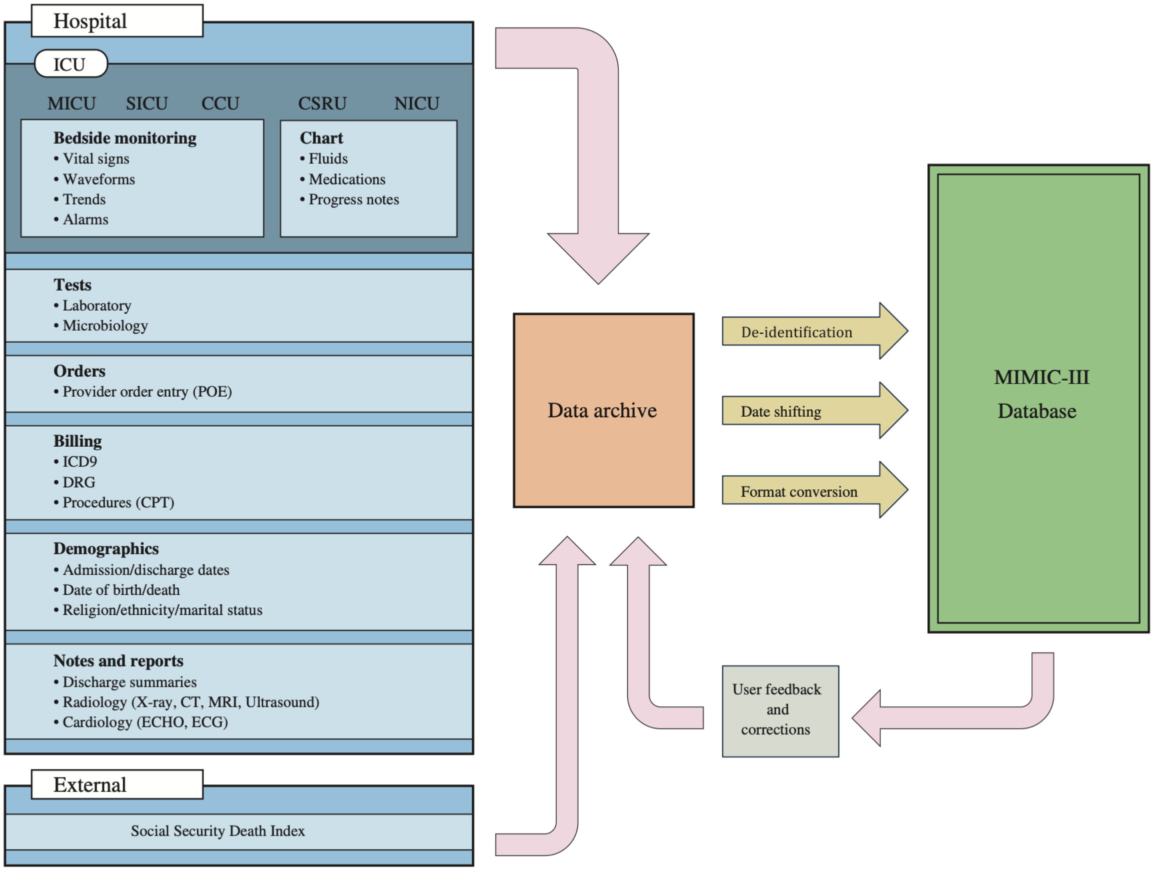

Figure 1: MIMIC-III database

2 Related Work

The current state of the art is Read, Attend, and Code (RAC) [1] model which uses auxiliary information on the individual ICD codes to help in the classification process. It uses an attention module to give cross attention scores between the input texts and the individual encoded ICD code titles. This work uses information that may not always be available in extreme multi-label classification problems and introduces overhead that does help in the task. Also, I recently discovered that the method of using the pre-trained ClinicalBERT on this task of medical code prediction has also been done. In the work by Biseda et al. [6], they focus on a restricted label space of top 50 most common labels. The focus of their work appeared to be to deploy their ICD code predictor. Also, by focusing on the top 50 labels, from the 6,918 possible, they simplified the problem greatly while making it much less practical for use in hospitals. I view the sparsity of the label space to be the main challenge of this task and posit that this is a novel direction.

3 Technical Approach

My approach involves using the pre-trained ClinicalBERT [2]. This pre-trained BERT model was trained on MIMIC-III discharge summaries and clinical notes. This model has 108 million parameters

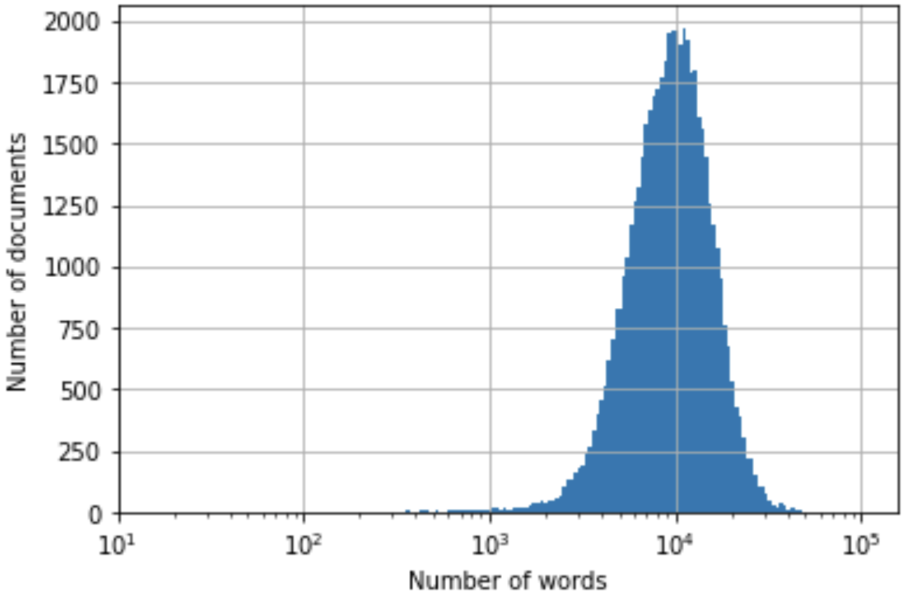

Figure 2: Most input texts are over 1000 words long.

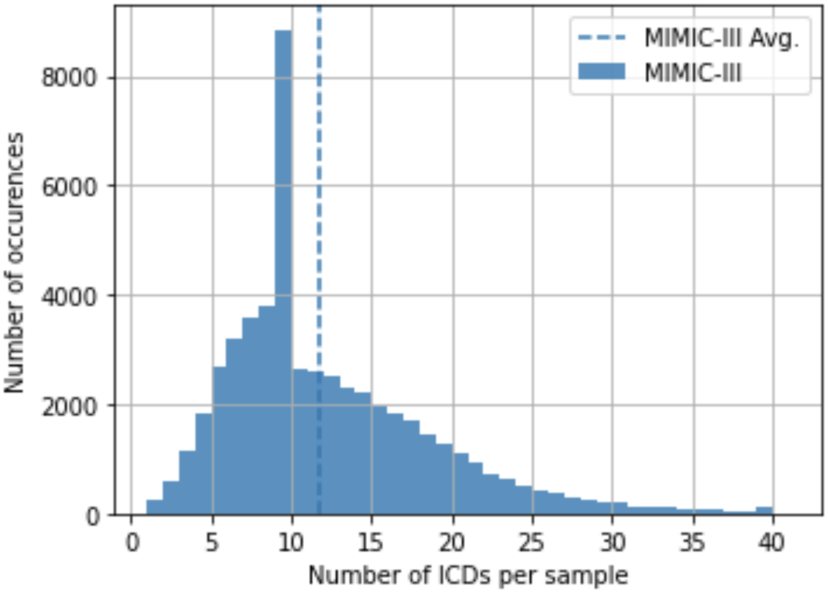

Figure 3:

and was trained for 18 days. It was trained on a number of tasks including natural language inference, named entity recognition, concept extraction, named entity extraction and de-identification. This pre-trained model was stacked with fully connected layers and dropout layers. The final output layer was a fully connected layer with output nodes equal to the number of ICD9 codes with sigmoid activation at each node. The data was preprocessed by removing extra white space, removing special characters and tokenizing the input.

Table 1: Sample table title

Part		
Model	Micro AUC-ROC	Micro Average Precision
Trainable Embedding LSTM	68.4	-
BERT base [7]	78.8	4.9
Logistic Regression [4]	93.7	-
ClinicalBERT [2] + Dense	95.6	15.4
RAC [1]	99.2	-

4 Experiments

To gain a benchmark, I developed a LSTM model which used a trainable embedding layer to convert the discharge summary text into embeddings. This LSTM model used a similar dense layer for multi-label classification. The rest of the experiments used BERT embeddings from their corresponding pre-trained BERT model. The base BERT [7] model used input sequences truncated or padded to 512 to match the original model. Clinical-BERT's inputs were preprocessed similarly to sequence length of 128. The best performing model had ClinicalBERT stacked with a dense layer with 768 ReLU nodes, a 50% dropout layer, and the final 6,918 sigmoid node output layer. This model was trained with Adam optimizer at a learning rate of 1e-5 and decay of 1e-6. As this a task with a large, sparse label space, we focus micro AUC-ROC. This model achieved a final micro AUC-ROC of 95.6. Comparison with other models can be seen in Table 1. Our model outperforms the state of the art at 2018 [4] which focused on utilizing logistic regression and domain knowledge to tackle this task.

5 Conclusion

This model outperforms numerous state of the art models from just a few years ago. While it is unable to compete with RAC, this model uses less information and was trained for a much shorter period. This work clearly demonstrates the power of transfer learning and making easily publicly accessible large, well-trained models. As I, and the majority of the population cannot get their hands on expensive GPUs (or extensive time on cloud computing services), the fact that a 100 million parameter model trained for 18 days can be used by the public is truly extraordinary. This experience was a great introduction to natural language processing and helped me learn many fundamental processes and standards in the field.

References

[1] Kim, B. (2021, July 10). Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. arXiv:2107.10650

[2] Alsentzer, E. (2019, June 20). Publicly Available Clinical BERT Embeddings. arXiv:1904.03323

[3] Reys, A. (2020, July 29). Predicting Multiple ICD-10 Codes from Brazilian-Portuguese Clinical Notes. arXiv:2008.01515

[4] Mullenbach, J. (2018, April 16). Explainable Prediction of Medical Codes from Clinical Text. arXiv:1802.05695

[5] Johnson, A., Pollard, T., Shen, L. et al. (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 https://doi.org/10.1038/sdata.2016.35

[6] Biseda B. et al. (2020) Prediction of ICD Codes with Clinical BERT Embeddings and Text Augmentation with Label Balancing using MIMIC-III arXiv:2008.10492

[7] Devlin J. et al. (2018) Pre-training of Deep Bidirectional Transformers for Language arXiv:1810.04805