Automatic Encoding of ICD-9 Using Clinical Notes from MIMIC-III Discharge Summary

Junge Zhang

Xiaoyi Zhang

Abstract

ICD, International Classification of Disease, determines the scope and type of patients' disease. To accurately assign ICD codes is important as it is closely related to the reimbursement and patient care. With the growing power of machine learning and deep learning techniques in medical fields, studies have appeared to automatically encode ICD codes to save the human labor cost. In our study, we focus on applying state-of-art BERT and RoBERTa models to assign multi-label ICD-9 codes for patients based on discharge summary data from MIMIC-III, as transfer learning and BERT dominate the NLP field in recent years. However, due to the 512 token limit described in the original BERT paper, it is not sufficient to apply a single BERT or RoBERTa model on discharge summaries that are typically very long sequences. In our work, we have shown how we break this limit by applying multiple predefined BERT or RoBERTa models for the same sequence. We also prove BERT and RoBERTa can provide a strong predictive performance for ICD-9 encoding than other traditional deep learning techniques like LSTM and GRU.

1. Purpose

The goal of this study is to use natural language understanding models to automatically encode ICD-9 codes to discharge summaries in the MIMIC-III dataset.

2. Introduction

To accurately assign ICD codes as easily and as accurate as possible is very important as it determines reimbursement (Larkey and Croft, 1996). More importantly, ICD codes are also closely related to patient care. However, the assignment process requires huge human labor cost in the past because of manual labeling of ICD codes. To address this issue, research for automatic encoding of ICD codes has started. Early work incorporates traditional machine learning methods for this task. This includes building K-nearest-neighbor classifier to assign ICD-9 codes based on similar documents (Larkey and Croft, 1996) and train decision tree to construct a rule-based system for ICD-9 codes (Farkas and Szarvas, 2008).

With deep learning becomes more popular for medical data, related work begins to focus on more predicative deep learning techniques. By treating sequence data as features, Recurrent Neural Network is the main technique for this study area. Typical work includes using LSTM for multi-label classification task (Ayyar and Bear, 2016), bidirectional GRU to train all ICD labels together (Blanco et al., 2019) and tree-LSTM to model the hierarchical structure of ICD codes (Chen and Ren, 2019). Furthermore, with the appearance of modern NLP techniques, some work combine these NLP concepts including training LSTM with attention mechanism (Shi et al., 2017) and applying sequence-to-sequence model as a multilingual approach (Atutxa et al., 2019).

However, with transfer learning and BERT begin to dominate the NLP field, our study inclines to mainly refer these state-of-art deep learning techniques for ICD-9 encoding. Recent related work has obtained outstanding results including transfer learning with a CNN architecture (Zeng et al., 2019) and ClinicalBERT trained on first 128 tokens with pre-defined BERT (Alsentzer et al., 2019). However, as the medical record data are mostly very long sequences, including the MIMIC-III discharge summary data we use for ICD-9 encoding, our work primary focuses on how to apply BERT on such long sequence data while breaking the 512 token limit discussed in the original BERT paper (Delvin et al., 2019). The primary work we refer to is applying multiple BERT models trained on segmented sequence and then combine the output for i2b2 data (Mulyar et al., 2019). Our study extends the idea to ICD-9 encoding as a multi-label classification task. Moreover, as RoBERTa is introduced as a robust version of BERT (Liu et al., 2019), we also experiment the same framework with RoBERTa for this task.

3. Hypothesis

We hypothesize that general language models, namely BERT and RoBERTa in this study, are capable to accurately encode ICD-9 codes to clinical notes and outperform language models such as LSTM and GRU. The reasoning behind the hypothesis is the superior performance of BERT and RoBERTa to LSTM and GRU on the SuperGLUE (Wang et al., 2019), which is a widely-recognized NLU benchmark including various tasks with a focus on the understanding of long passages, and hence our hypothesis that BERT and RoBERTa outperforming LSTM/GRU is a natural extension from general to domain-specific text data.

4. Data

In this study we train and evaluate language models on the discharge summaries in the MIMIC-III dataset (Johnson et al., 2016). We obtain 47377 notes from unique patients, and for patients with multiple admissions (i.e. discharge summaries), we only keep the earliest note to prevent data leakage of our model (e.g. predicting based on memorized patients instead of on input notes). We randomly split the obtained notes into train, validation, and test sets by 0.85:0.05:0.10.

In practice, a discharge summary is usually encoded with multiple ICD codes, and in our training set there are 12939 unique ICD-9 codes, making our task a multi-label, multi-class classification problem. For ICD-9 codes in the validation or test set but not found in training set, we discard them from the hypothesis space, as it is impossible for a model to learn to predict unseen labels. Since many of the ICD-9 codes appear for very limited times in our dataset as shown in Fig 1, there is insufficient information for our model to predict these codes, and thus we only keep the most frequent 1000 ICD-9 codes from the training set to form our hypothesis space.

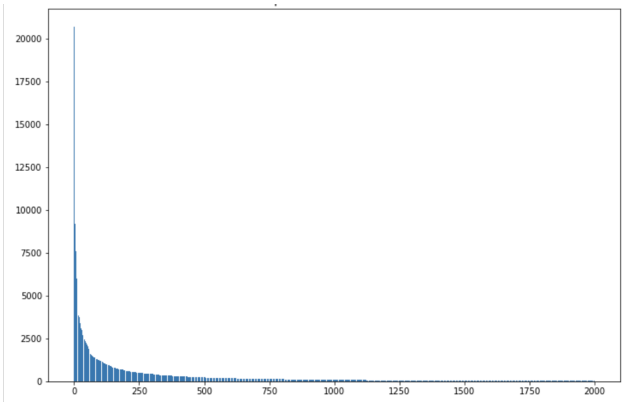

Figure 1: The frequency distribution of ICD-9 codes in the MIMIC-III discharge summaries. The x-axis gives the count of ICD-9 codes, and the y-axis stands for the frequency of the codes. After the most frequent 1000 ICD-9 codes, the frequency becomes very low for prediction tasks.

5. Methods

5.1 Baseline

We train an LSTM and a GRU for our baselines. For fair comparison with BERT/RoBERTa models, we input the first 1024 tokens into the models after removing the special characters. The models output logits for 1000 labels, and the optimization is done by Adam optimizer with learning rate $= 10^{-3}$ and weight decay $= 10^{-5}$ on binary cross entropy loss (implemented as `BCEWithLogitLoss` module on pytorch). The models preformace are evaluated in ranking metrics, namely recall at k and precision at k.

5.2 BERT and RoBERTa

We train the BERT and RoBERTa models with the huggingface transformers package, and we implement the `bert_base_uncased` predefined BERT model and `roberta_base` RoBERTa predefined model. To break the 512 token limit of BERT for the long discharge summary data (mostly with over 1000 tokens but less than 1500 tokens), we first split the sequence data into half. Then, we apply the BERT or RoBERTa tokenizer with a limit of 512 tokens to tokenize each half sequence to get two sets of token IDs and attention masks. Each set of token ID and attention mask is put into the predefined BERT or RoBERTa model to output two hidden states. We then concatenate the last hidden state of the special token, "CLS" as the final output of the predefined models. The concatenated result is input to a full connected layer with 1000 output to perform the multi-label classification task. We train both models with the batch size = 4, and learning rate $= 5 * 10^{-5}$ on binary cross entropy loss (implemented as `BCEWithLogitLoss` module on pytorch). Due to the capacity limit of NYU HPC cluster, we first train both models for 5 epochs and then retrain the models for another 5 epochs. Fig 2, Fig 3, Fig 4, Fig 5 show the loss curves for BERT and RoBERTa for the 10 epochs.

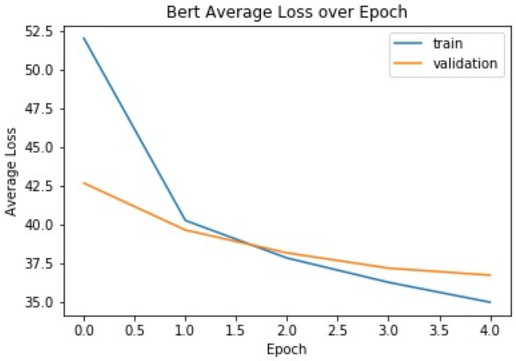

Figure 2: Loss curve of BERT for the first five epochs

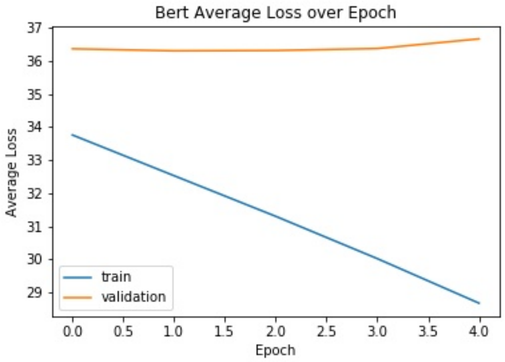

Figure 3: Loss curve of BERT for the retrain five epochs

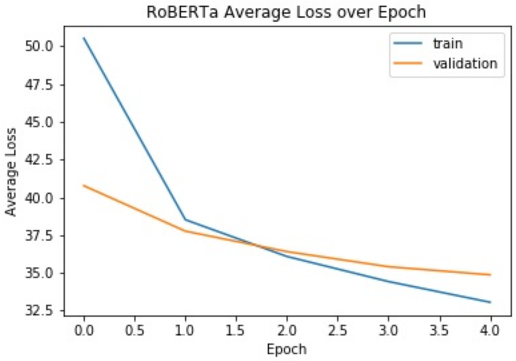

Figure 4: Loss curve of RoBERTa for the first five epochs

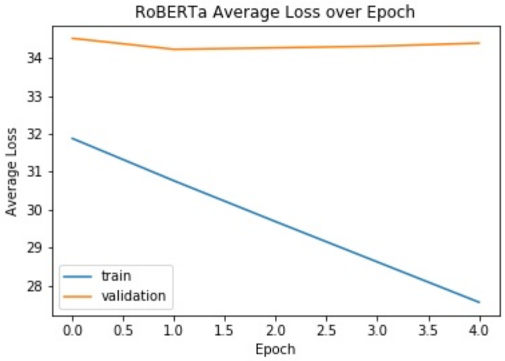

Figure 5: Loss curve of RoBERTa for the retrain five epochs

6. Results

We use precision at k and recall at k to evaluate the performance of our models. While precision is the percentage of true positive predictions out of all predicted positive labels, recall is the percentage of true positive predictions out of all real positive labels. Our models output logits from the fully connected layer and we input the logits into a sigmoid layer. The order of the values from the sigmoid layer determines top k predicted labels, which define ”at k” in our case. Table 1 shows the evaluation results for all of our models.

Table 1: Evaluation Results

	Precision at 5	**Precision at 10**	**Recall at 5**	**Recall at 10**
LSTM	0.2555	0.1891	0.1233	0.1825
GRU	0.2523	0.1810	0.1217	0.1746
BERT	0.8596	0.8337	0.7974	0.6227
RoBERTa	0.7037	0.6770	0.8222	0.6293

Our evaluation results have shown BERT and RoBERTa perform significantly better regarding the multi-label automatic ICD-9 encoding task. Meanwhile, BERT performs slightly better for precision at k and RoBERTa performs slightly better for recall at k.

7. Discussion

In our study, our hypothesis is that general language models, namely BERT and RoBERTa can more accurately encode ICD-9 codes to clinical notes and outperform other deep learning models such as LSTM and GRU. From our results evaluated based on precision at k and recall at k, BERT and RoBERTa have shown significant higher predicative power than LSTM and GRU for assigning ICD-9 codes. With higher precision at k for BERT and better recall at k for RoBERTa, they both can capture the important information flow of the discharge summary and assign majority of ICD-9 labels correctly to patients. Although our study doesn’t show any inconsistent results and completely proves our hypothesis, there are two main deficiencies. First, the results are not able to achieve higher results over 90%. The possible reason is the discharge summary sequences are not further processed with some punctuation or typos also tokenized and processed through our BERT and RoBERTa model. The other deficiency is both BERT and RoBERTa models require long running time, over 48 hours each for total 10 epochs. The reason is we train two predefined BERT or RoBERTa models to get hidden state outputs of each half sequence. But we believe the long training time is worth for stronger predicative power.

8. Teamwork

Junge Zhang: literature review, train BERT and RoBERTa models, evaluate BERT and RoBERTa models.
Xiaoyi Zhang: literature review, data preprocessing, train baseline models, evaluate basline models.
All the codes are shared and reviewed by both of us.

References

E. Alsentzer, J. Murphy, W. Boag, W. Wang, D. Jin, T. Naumann, and A. McDermott. Publicly Available Clinical BERT Embeddings. *arXiv*, 1907:11692, 2019.

A. Atutxa, A. Ilarraza, K. Gojenola, M. Oronoz, and O. Perez-de Viñaspre. Interpretable Deep Learning to Map Diagnostic Texts to ICD-10 Codes. *International Journal of Medical Informatics*, 129:49–59, 2019.

S. Ayyar and O. Bear. Tagging Patient Notes with Icd-9 Codes. 2016.

A. Blanco, O. Perez-de Viñaspre, A. Pérez, and G. Casilas. Boosting ICD Multi-label Classification of Health Records with Contextual Embeddings and Label-granularity. *Computer Methods and Program in Biomedicine*, 188, 2019.

Y. Chen and J. Ren. Automatic ICD Code Assignment Utilizing Textual Descriptions and Hierarchical Structure of ICD Code. pages 348–353, 2019.

J. Delvin, M. Chang, K. Lee, and K Toutanonva. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 1810:04085, 2019.

R. Farkas and G. Szarvas. Automatic Construction of Rule-based ICD-9-cm Coding Systems. *BMC Bioinformatics*, 2008.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

L. Larkey and W. Croft. Automatic Assignment of ICD9 Codes to Discharge Summaries. 02 1996.

Y. Liu, M. Ott, N. Goyal, J Du, M. Joshi, Chen D., and et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 1907:11692, 2019.

E. Mulyar, E. Schumacher, M. Rouhizadeh, and M Dredze. Phenotyping of Clinical Notes with Improved Document Classification Models Using Contextualized Neural Language Models. *arXiv*, 1910:13664, 2019.

H. Shi, P. Xie, Z. Hu, M. Zhang, and E. Xing. Towards Automated ICD Coding Using Deep Learning. *arXiv*, 1711:04075, 2017.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275, 2019.

M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang. Automatic ICD-9 Coding via Deep Transfer Learning. *Neurocomputing*, pages 43–50, 2019.