Fine-Tuning Transformer-XL on Clinical Natural Language Processing

Stanford CS224N Custom Project

Xianghao Zhan
Department of Bioengineering
Stanford University
xzhan96@stanford.edu

Yiheng Li
Department of Biomedical Data Science
Stanford University
yyhhli@stanford.edu

Abstract

Although many applications based-on mining clinical free text have been developed, the state-of-the-art transformer-based models have not been applied in clinical NLP. We aim to address the long-range dependencies in clinical free text caused by different sections with the latest Transformer-XL model by fine-tuning it on MIMIC-III clinical text. Having requested and cleaned the MIMIC-III clinical text based on self-developed rules, we prepared the data for training classifiers on diagnostic code prediction of 8 common cardiovascular diseases. We used huggingface API to fine-tune and evaluate Transformer-XL model on MIMIC-III dataset and compared the results with baseline methods including bag-of-words and TF-IDF. And the Transformer-XL outperformed the Bag-of-Words and TF-IDF on 3 of 6 tasks, on which we have already got the results. Furthermore, the Transformer-XL was only fine-tuned for 1 epoch, and therefore we believe there is a promising potential for a better fine-tuned Transformer-XL to better predict the diagnostic codes accurately. The better accuracy of diagnostic codes aids in the structuring of free-text clinical notes, which can be better and easier for downstream machine learning tasks, such as survival predictions and multi-modality data fusion, because, the structured diagnostic codes can be fed into machine learning models than the unstructured data.

1 Key Information to include

- External collaborators (if you have any): NA
- External mentor (if you have any): Olivier Gevaert
- Sharing project: NA

2 Introduction

Over the past years, natural language processing (NLP) has been used in many clinical data mining applications such as correcting erroneous diagnostic codes [1], predicting cancer staging information [2], which has benefited the healthcare system [3, 4]. The characteristics of clinical free-text notes such as infrequent clinical terms, acronyms and abbreviations can lead the conventional NLP techniques, which are generally based on general English corpus, hard to be immediately and directly transplanted to clinical applications. To cope with this issue, adaptations of general natural language processing have been developed by many researchers. For instance, a biomedical word2vec word-embedding model has been trained on a corpus of PubMed and MIMIC-III notes [5]; a doc2vec model has been trained on a corpus of cardiovascular outpatient progress notes [1]. However, the development of models able to successfully capture extra long dependencies in clinical text is still a challenge. In the previous study of clinical text mining, such methods as continuous bag-of-words with average pooling directly used the averaged word2vec embeddings as the text-level embeddings can lose the

sequence information. While LSTM models can capture sequence information in the text, they are also relatively hard to capture long-range dependencies which are frequent in clinical notes as the notes usually contain personal medical history, family medical history and prescriptions in different sections. Recently, researchers have applied pre-trained transformers such as BERT on clinical notes and showed good performances on clinical NLP challenges but the long-range dependency and the cutting of long clinical notes in transformer fine-tuning still leaves space for improvement.

To be more specific, the vanilla transformers, including BERT, suffered from a drawback related to fixed-length input. By segmenting the input sentences into fixed-length chunks, and then modeling each chunk independently, the model assumes that each segments are independent. And the model provides no information interaction between the boundaries of each segment of the text chunks. This limitation may lead to two issues: first, token at the beginning of each segment do not have sufficient context for proper optimization; and secondly, the model is limited by fix-length context and have no ability to capture information across an extra-long sentence.

In this study, we aimed to develop a fine-tuned Transformer-XL model situated in the clinical note text mining field with an aim to capture extra-long dependencies for variable lengths for better clinical text mining performance in the tasks of diagnostic code predictions. More experiments including language modeling and name entity recognition were designed but due to extreme challenges of experimental failures with the Transformer-XL models. Therefore, we have analyzed the failures and challenges we encountered in great details in the discussion session for further work to better improve the project.

3 Related Work

Many of the previous studies on clinical NLP tasks have been based on context-independent word embeddings and vectorization such as word2vec[5], TF-IDF[1], etc. But they are insufficient to capture the sequential information in the clinical free text. Particularly, there are lots of long range dependencies in clinical text. Different sections can have varying lengths and there can be long range dependencies between sections (e.g. personal medical history and clinical symptom descriptions) which can be separated by other sections. LSTM deals with long range dependencies in a stochastic manner, which could also suffer from gradient vanishing or explosion problems. Vanilla transformer models also may not capture long range dependencies due to the blocking of documents. Therefore we aim to adopt the newly developed Transformer-XL[6] model to capture extra-long dependencies in clinical text.

The research of Transformer-XL[6] proposed a novel architecture, Transformer-XL, which managed to address the limitations of vanilla transformers mentioned above. Transformer-XL consists two key novel ideas: first, it involves segment-level recurrence, which cached and reused hidden states from last batch and therefore allows information to flow over the boundaries of text chunks; second, it keeps the sequence information coherent by a novel positional encoding scheme because directly including segment-level recurrence may lead the position encoding to fail. With these two novelties, Transformer-XL benefited from the key advantages of both Recurrent Neural Network (RNN) and transformers with both recurrence and attention mechanisms. As a result, Transformer-XL learns dependency that is 80% longer than RNNs and 450% longer than vanilla Transformers. And it has achieved better performance in language modeling tasks with a bpc/perplexity of 0.99 on en-wiki8, 1.08 on text8, 18.3 on WikiText-103, 21.8 on One Billion Word, and 54.5 on Penn Treebank (without fine-tuning). Furthermore, this algorithm is up to 1800+ times faster than vanilla transformers.

4 Approach

In this work we fine-tuned the pre-trained Transformer-XL model on clinical free text in MIMIC-III dataset[7] in sentence classification tasks based on eight diagnostic codes related to common cardiovascular diseases. In addition, we intended to firstly fine-tune the Transformer-XL model on causal language modeling and evaluate the perplexity on the causal language modeling task against a baseline given by distilGPT2 but the experiments failed due to GPU RAM explosion and the time expenditure of over 90 hours to train for one epoch. Finally, the Transformer-XL model was directly fine-tuned on the diagnostic code prediction tasks.

5 Experiments

5.1 Data

In this study, we used the MIMIC-III dataset[7] (Medical Information Mart for Intensive Care III) data set of de-identified health-related data of 40,000 intensive care unit stays at Beth Israel Deaconess Medical Center. The MIMIC-III data set has been regarded as the benchmark data set for many NLP tasks including time-of-stay prediction, diagnostic code prediction, etc. [8, 9, 10]. In order to request data from the MIMIC-III public dataset, we have firstly passed the required courses of personal health information (PHI), on The Collaborative Institutional Training Initiative (CITI).

5.2 Evaluation method

In the evaluation of diagnostic code prediction, the following metrics were calculated and compared:accuracy (rate of correct prediction in all predictions), AUROC (area under the receiver operating curve) and AUPRC (area under the precision recall curve). AUROC is the area under the curve with the x-axis denoting the false positive rate and y-axis denoting the true positive rate (TPR). AUPRC is the area under the curve with the x-axis denoting the recall and y-axis denoting the precision (precision: rate of true positive predictions in all positive predictions), recall: rate of true positive predictions in all positive cases). AUROC has been a widely used metric in evaluating binary classifiers without dependence on the decision threshold on predicted class probability. The AUPRC was also used in this study because it is more sensitive to prevalence and can better reflect model performance in a imbalanced data set [11].

The baseline models we developed in this study involved bag-of-words (BOW) and term frequency-inverse document frequency (TF-IDF). Bag-of-words (BOW) [12] is a word-count based word vectorization algorithm which is commonly used in document classification. The method counts the frequency of each term in the text and uses the frequency of individual term as the feature. The number of features is the same as the number of all distinct terms in the training set and the feature values are proportional to the occurrences of the distinct terms.TF-IDF [13] is an algorithm with normalized BOW word vectors to emphasize the different importance of terms. The feature in TF-IDF is the ratio of term frequency (TF) and inverse document frequency (IDF). The value of a word vector increases proportionally to the term frequency but is offset by the number of texts that contain the term. The feature dimensionality of TF-IDF was the same as that of BOW.

5.3 Experimental details

As soon as we got the datasets, we found out that the clinical free text documented in MIMIC-III was filled with unexpected special tokens, characters and structures. We firstly cleaned the dataset by designing several rules: 1. Detection and removal of personal health information (PHI) with unknown charactor ‘unk‘; 2.Deletion of repetitive section names, such as "admission date", "discharge date"; 3. Lower casing of the entire corpus; 4. Detection and removal of special characters; 5. Deletion of sentences shorter than 10 words; 6. Replacement of repetitive new line token with a single space.

To prepare for the diagnostic code classification tasks, we extracted all of the discharge summaries from the MIMIC-III dataset. Because diagnostic codes are related to encounters, we only extracted the discharge summaries (59,652 notes; 41,127 patients) related to their associated encounters. We split the train/val/test based on patients IDs to prevent information leakage across train/val/test. Finally, we got 32,901/4,113/4,113 notes in the train/val/test sets. We focused on 8 common cardiovascular diseases and formulated 8 binary classification tasks to test the Transformer-XL. The ICD-9 diagnostic codes and diseases and their prevalence in train/val/test sets were shown in the Table. 1. These eight codes represent a large variance of prevalence, with the largest prevalence (28.8%) approximately twenty times the smallest prevalence (1.4%).

5.4 Results

The results of the classification of eight diagnostic codes were shown in Table 2 (the pending cells were due to still-running programs, as it took 8 hours for fine-tuning the pre-trained Transformer-XL for one epoch). According to the results (we focus on AUPRC since the data was highly imbalanced), on the prediction of diagnostic code – 425, 427, 416 – Transformer-XL outperformed the BOW and

Table 1: The prevalence of eight common cardiovascular diseases and ICD-9 codes in the training, validation and test sets of the MIMIC-III dataset.

Code	Description	Training	Validation	Test
410	Acute myocardial infarction	8.85%	9.08%	8.98%
414	Chronic ischemic heart disease	23.41%	24.22%	24.43%
416	Pulmonary heart disease	4.37%	4.53%	4.66%
425	Cardiomyopathy	3.46%	3.62%	3.12%
427	Atrial fibrillation flutter	27.97%	28.81%	28.55%
428	Heart failure	22.30%	22.72%	22.23%
440	Atherosclerosis	3.07%	2.94%	2.97%
456	Esophageal Varices	1.67%	1.59%	1.40%

Table 2: The prevalence of eight common cardiovascular diseases and ICD-9 codes in the training, validation and test sets of the MIMIC-III dataset.

Code	Disease	Method	Accuracy	AUROC	AUPRC
410	Acute myocardial infarction	BOW	0.903	0.526	0.094
		TF-IDF	0.910	0.550	0.099
		Transformer-XL	0.910	0.452	0.054
414	Chronic ischemic heart disease	BOW	0.731	0.541	0.264
		TF-IDF	0.756	0.548	0.263
		Transformer-XL	0.790	0.501	0.194
416	Pulmonary heart disease	BOW	0.944	0.513	0.048
		TF-IDF	0.953	0.563	0.053
		Transformer-XL	0.955	0.500	0.523
425	Cardiomyopathy	BOW	0.962	0.491	0.029
		TF-IDF	0.969	0.537	0.031
		Transformer-XL	0.969	0.500	0.516
427	Atrial fibrillation flutter	BOW	0.688	0.542	0.307
		TF-IDF	0.714	0.547	0.309
		Transformer-XL	0.714	0.503	0.626
428	Heart failure	BOW	0.754	0.525	0.233
		TF-IDF	0.778	0.540	0.235
		Transformer-XL	0.778	0.459	0.193
440	Atherosclerosis	BOW	0.962	0.518	0.034
		TF-IDF	0.970	0.567	0.040
		Transformer-XL	(Pending)	(Pending)	(Pending)
456	Esophageal Varices	BOW	0.978	0.458	0.014
		TF-IDF	0.986	0.536	0.014
		Transformer-XL	(Pending)	(Pending)	(Pending)

TF-IDF baselines. However, on certain other tasks (task code: 410, 414, 428), the Transformer-XL fine-tuned for 1 epoch on the diagnostic code was not performing better than the BOW/TF-IDF baseline with logistic regression. On these tasks, AUPRC was very low, which indicated that the model was very reluctant to make positive predictions. We assume that it may be because the dataset is extremely imbalanced, which is very detrimental for training. As the most of the time, the model is only learning the negative cases. And we only fine-tuned the Transformer-XL for 1 epoch due to the fact that each epoch of fine-tuning took 8 hours and there were eight prediction tasks in this project. In the future, more epochs of fine-tuning of the pre-trained Transformer-XL may lead to better results.

6 Discussion

In this project, we fine-tuned Transformer-XL on the predictions of 8 diagnostic codes on MIMIC-III dataset, and the Transformer-XL outperformed the Bag-of-Words and TF-IDF on 3 of 6 tasks, on which we have already got the results. Furthermore, the Transformer-XL was only fine-tuned for 1 epoch, and therefore we believe there is a promising potential for a better fine-tuned Transformer-XL to better predict the diagnostic codes accurately. The better accuracy of diagnostic codes aids in

the structuring of free-text clinical notes, which can be better and easier for downstream machine learning tasks, such as survival predictions and multi-modality data fusion, because the structured diagnostic codes can be fed into machine learning models than the unstructured data.

During the experiment, we have encountered lots of failures and challenges and based on these failures, we discuss the limitations of the current work and what can be addressed by further studies in the future.

Firstly, the Transformer-XL model lacks transparency and publicly available resources at present. Although the original developers discuss the potential to expand the scope of utility of Transformer-XL, no pre-trained models or detailed tutorials on fine-tuning the Transformer-XL models are given by the developers, which kind of hinders the development of the community.

Secondly, as we fine-tuned the Transformer-XL on the classification of eight CVD diagnostic codes independently, the fine-tuned models may not generalize well to other tasks such as extrapolation to other symptom codes (R-codes in ICD10). Therefore, it may be worthwhile for future researchers to fine-tune the Transformer-XL with multiple heads in a round-robin manner. For instance, in the study proposed by Mulyar et al[14], the BERT model could be trained with multiple heads involving not only the sentence classification heads, but also the name entity recognition heads and the language modeling heads. Via multi-task learning in a similar manner, the Transformer-XL model can be better fine-tuned to be able to adapt to multiple different types of clinical NLP tasks.

Thirdly, in this finalized version of this study, we only fine-tuned the Transformer-XL model based on eight diagnostic code classification tasks. We initially aimed to apply causal language modeling as the fine-tuning process on the clinical free-text notes but there wasn't any tutorial of fine-tuning Transformer-XL on custom datasets. Although we managed to transfer the fine-tuning protocols designed for distilGPT2 to Transformer-XL, the intrinsic difference between these two types of models led to failure of training because many of the trainer and trainer argument settings for distilGPT2 were incompatible to Transformer-XL while the instructions on huggingface lack details of fine-tuning. In the future, as time allows us to debug all the errors, the Transformer-XL can be fine-tuned on language modeling tasks with much more notes than the discharge summaries, which may be able to further boost the power of clinically fine-tuned Transformer-XL models.

Fourthly, as we were to train the distilGPT2 in causal language modeling to form a baseline (the pre-trained distilGPT2 model reached a perplexity of 211.5 on the MIMIC-III data), the time expenditure was enormous (fine-tuning for 1 epoch took 90 hours) and even the evaluation of the distilGPT2 took 2 hours on the 10% MIMIC-III dataset. In the future, it is worthwhile for researchers to seek for simpler models with less complexity and fewer parameters to accelerate the training, fine-tuning and evaluation via further knowledge distillation, for instance. More powerful GPU with more RAM and stronger computational power could also be adopted to further the work of fine-tuning large pre-trained language models.

Fifthly, during the project, we applied huggingface package and found that it easily and implicitly uses up the cache storage on the virtual machine, which leads the training/fine-tuning and even the connection of IDE to the virtual machine to fail. In the future, it may be worthwhile for the developers to add in better instructions on the cache usage and how users may be able to adapt the cache using strategies and manage the space on disk.

Sixthly, it is also worthwhile for researchers to develop Transformer-XL models for clinical applications from scratch because the power of fine-tuning can still be limited by the Transformer-XL tokenizer. Because in the clinical text, there are usually many unique terms and abbreviations such as "hx" for history, "afib" for atrial fibrillation, "hf" for heart failure, etc. To develop a completely new and clinically adaptive tokenizer based on clinical text may help design a better pre-trained model for clinical NLP applications.

Finally, as the majority of the issues we encountered in this project are related to the limitation of GPU RAM and/or disk memory, this leads us to ponder over the current trend in NLP to go more complicated and the preference for more complicated model structures and more parameters. We believe that it may also be necessary for researchers to develop simpler structures and models for those developers who do not have access to expensive and strong computational resources. The state-of-the-art complicated large pre-trained language models such as GPT3 can be too overwhelming and even intimidating for many research groups and developers to use for any further fine-tuning tasks and application development. As Leonardo Da Vinci puts it, simplicity is the ultimate sophistication.

The multiple failures during this project really motivates us to seek for simple, fast, interpretable and readily understandable NLP models which works for a specific problem. For instance, in a latest publication of international conference on machine learning (ICML 2021)[15], the researchers from IBM Watson borrowed the inspiration of adapting a biological neural network of fruit fly brains to develop word vectors for language modeling and the results turned out to be pretty well. Although the performance was not the best when compared with other deep-neural-net-based models (Spearman rank correlation on SCWS dataset: 49.1 given by fruit-fly-model vs 56.8 given by BERT; accuracy on WiC dataset: 57.7 given by fruit-fly-model vs 61.2 given by BERT; accuracy for document classification task on TREC-6 dataset: 90.4 given by fruit-fly-model vs 94.0 given by BERT), the simplicity and speed of the model seems extremely attractive for researchers as it abandons millions of model parameters. In another study, Zhan et al[1] compared the conventional word vectorization method such as BOW and TF-IDF and new neural-net-based embeddings including word2vec and doc2vec by building simple logistic regression (LR) models to classify ICD-10 diagnostic codes and found the fully interpretable model TF-IDF/LR showed the highest AUROC and AUPRC. Therefore, it is also of great interest for further work to emphasize on simpler, smaller and faster models in NLP tasks.

7 Conclusion

We fine-tuned and evaluated a newly developed NLP model Transformer-XL on MIMIC-III dataset. For the task of predicting CVD symptoms, we compared the model's performance with baseline models including bag-of-words and TF-IDF. fine-tuned Transformer-XL could not outperform the baseline models, due to training obstacles of dataset imbalance and computational resource limitations. Use either down-sampling negative samples or up-sampling positive samples, or larger GPU memories and multiple GPUs to train Transformer-XL would be potential strategies for better performance, as future plan.

References

[1] Xianghao Zhan, Marie Humbert-Droz, Pritam Mukherjee, and Olivier Gevaert. Structuring clinical text with ai: old vs. new natural language processing techniques evaluated on eight common cardiovascular diseases. *medRxiv*, 2021.

[2] Khushbu Gupta, Ratchainant Thammasudjarit, and Ammarin Thakkinstian. Nlp automation to read radiological reports to detect the stage of cancer among lung cancer patients. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 138–141, 2019.

[3] Xing Wei and Carsten Eickhoff. Embedding electronic health records for clinical information retrieval. *arXiv preprint arXiv:1811.05402*, 2018.

[4] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[5] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9, 2019.

[6] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[7] 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.

[8] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*, 2018.

[9] Vincent Major, Monique S Tanna, Simon Jones, and Yin Aphinyanaphongs. Reusable filtering functions for application in icu data: a case study. In *AMIA Annual Symposium Proceedings*, volume 2016, page 844. American Medical Informatics Association, 2016.

[10] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer methods and programs in biomedicine*, 177:141–153, 2019.

[11] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[12] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[13] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[14] Andriy Mulyar and Bridget T McInnes. Mt-clinical bert: Scaling clinical information extraction with multitask learning. *arXiv preprint arXiv:2004.10220*, 2020.

[15] Yuchen Liang, Chaitanya K Ryali, Benjamin Hoover, Leopold Grinberg, Saket Navlakha, Mohammed J Zaki, and Dmitry Krotov. Can a fruit fly learn word embeddings? *arXiv preprint arXiv:2101.06887*, 2021.