BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining

Zachariah Zhang
NYU Langone Health
zz1409@nyu.edu

Jingshu Liu
NYU Langone Health
jl7722@nyu.edu

Narges Razavian
NYU Langone Health
narges.razavian@nyumc.org

Abstract

ICD coding is the task of classifying and coding all diagnoses, symptoms and procedures associated with a patient's visit. The process is often manual, extremely time-consuming and expensive for hospitals as clinical interactions are usually recorded in free text medical notes. In this paper, we propose a machine learning model, BERT-XML, for large scale automated ICD coding of EHR notes, utilizing recently developed unsupervised pretraining that have achieved state of the art performance on a variety of NLP tasks. We train a BERT model from scratch on EHR notes, learning with vocabulary better suited for EHR tasks and thus outperform off-the-shelf models. We further adapt the BERT architecture for ICD coding with multi-label attention. We demonstrate the effectiveness of BERT-based models on the large scale ICD code classification task using millions of EHR notes to predict thousands of unique codes.

1 Introduction

Information embedded in Electronic Health Records (EHR) have been a focus of the healthcare community in recent years. Research aiming to provide more accurate diagnose, reduce patients' risk, as well as improve clinical operation efficiency have well-exploited structured EHR data, which includes demographics, disease diagnosis, procedures, medications and lab records. However, a number of studies show that information on patient health status primarily resides in the free-text clinical notes, and it is challenging to convert clinical notes fully and accurately to structured data (Ashfaq et al., 2019; Guide, 2013; Cowie et al., 2017).

Extensive prior efforts have been made on extracting and utilizing information from unstructured EHR data via traditional linguistics based methods in combination with medical metathesaurus and semantic networks (Savova et al., 2010; Aronson and Lang, 2010; Wu et al., 2018a; Soysal et al., 2018). With rapid developments in deep learning methods and their applications in Natural Language Processing (NLP), recent studies adopt those models to process EHR notes for supervised tasks such as disease diagnose and/or ICD[1] coding (Flicoteaux, 2018; Xie and Xing, 2018; Miftahutdinov and Tutubalina, 2018; Azam et al., 2019; Wiegreffe et al., 2019).

Yet to the best of our knowledge, applications of recently developed and vastly-successful self-supervised learning models in this domain have remained limited to very small cohorts (Alsentzer et al., 2019; Huang et al., 2019) and/or using other sources such as PubMed publication (Lee et al., 2020) or animal experiment notes (Amin et al., 2019) instead of clinical data sets. In addition, many of these studies use the original BERT models as released in (Devlin et al., 2019), with a vocabulary derived from a corpus of language not specific to EHR.

In this work we propose BERT-XML as an effective approach to diagnose patients and extract relevant disease documentation from the free-text clinical notes with little pre-processing. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) utilizes unsupervised pretraining procedures to produce meaningful representation of the input sequence, and provides state of the art results across many important NLP tasks. BERT-XML combines BERT pretraining with multi-label attention (You et al., 2018), and outperforms other baselines without self-supervised pretraining by a large margin. Ad-

[1] ICD, or International Statistical Classification of Diseases and Related Health Problems, is the system of classifying all diagnoses, symptoms and procedures for a patient's visit. For example, I50.3 is the code for Diastolic (congestive) heart failure. These codes need to be assigned manually by medical coders at each hospital. The process can be very expensive and time consuming, and becomes a natural target for automation.

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 24–34

ditionally, the attention layer provides a natural mechanism to identify part of the text that impacts final prediction.

Compare to other works on disease identification, we demonstrate the effectiveness of BERT-based models on automated ICD-coding on a large cohort of EHR clinical notes, and emphasize the following aspects: 1) **Large cohort pretraining and EHR Specific Vocabulary.** We train BERT model from scratch on over 5 million EHR notes and with a vocabulary specific to EHR, and show that it outperforms off-the-shelf or fine-tuned BERT using off-the-shelf vocabulary. 2) **Minimal pre-processing of input sequence.** Instead of splitting input text into sentences (Huang et al., 2019; Savova et al., 2010; Soysal et al., 2018) or extracting diagnose related phrases prior to modeling (Azam et al., 2019), we directly model input sequence up to 1,024 tokens in both pre-training and prediction tasks to accommodate common EHR note size. This shows superior performance by considering information over longer span of text. 3) **Large number of classes.** We use the 2,292 most frequent ICD-10 codes from our modeling cohort as the disease targets, and shows the model is highly predictive of the majority of classes. This extends previous effort on disease diagnose or coding that only predict a small number of classes. 4) **Novel multi-label embedding initialization.** We apply an innovative initialization method as described in Section 3.3.2, that greatly improves training stability of the multi-label attention.

The paper is organized as follows: We summarize related works in Section 2. In Section 3 we define the problem and describe the BERT-based models and several baseline models. Section 4 provides experiment data and model implementation details. We also show the performances of different model and examples of visualization. The last Section concludes this work and discusses future research areas.

2 Related Works

2.1 CNN, LSTM based Approaches and Attention Mechanisms in ICD-coding

Extensive work has been done on applying machine learning approaches to automatic ICD coding. Many of these approaches rely on variants of Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs). Flicoteaux (2018) uses a text CNN as well as lexical matching to improve performance for rare ICD labels. In Xu et al.(2019), authors use an ensemble of a character level CNN, Bi-LSTM, and word level CNN to make predictions of ICD codes. Another study Xie and Xing (2018) proposes a tree-of-sequences LSTM architecture to simultaneously capture the hierarchical relationship among codes and the semantics of each code. Miftahutdinov and Tutubalina (2018) propose an encoder-decoder LSTM framework with a cosine similarity vector between the encoded sequence and the ICD-10 codes descriptions. A more recent study Azam et al. (2019) compares a range of models including CNN, LSTM and a cascading hierarchical architecture in prediction class with LSTM and show the hierarchical model with LSTM performs best.

Many works further incorporates the attention mechanisms as introduced in Bahdanau et al. (2015), to better utilize information buried in longer input sequence. In Baumel et al. (2018), the authors introduce a Hierarchical Attention bidirectional Gated Recurrent Unit(HA-GRU) architecture. Shi et al. (2017) use a hierarchical combination of LSTMs to encode EHR text and then use attention with encodings of the text description of ICD codes to make predictions.

While these models have impressive results, some fall short in modeling the complexity of EHR data in terms of the number of ICD codes predicted. For example, Shi et al. (2017) limit their predictions to the 50 most frequent codes and Xu et al. (2019) predict 32. In addition, these works do not utilize any pretraining and performance can be limited by size of labeled training samples.

2.2 Transformer Modules

Unsupervised methods to learn word representations has been well established within the NLP community. Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) learn vector representations of tokens from large unsupervised corpora in order to encode semantic similarities in words. However, these approaches fail to incorporate wider context into account as the pretraining only considers words in the immediate neighbourhood.

Recently, several approaches are developed to learn unsupervised encoders that produce contextualized word embedding such as ElMo (Peters et al., 2018) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019).

These models utilize unsupervised pretraining procedures to produce representations that can transfer well to many tasks. BERT uses self-attention modules rather than LSTMs to encode text. In addition, BERT is trained on both a masked language model task as well as a next sentence prediction task. This pretraining procedure has provided state of the art results across many important NLP tasks.

Inspired by the success in other domains, several works have utilized BERT models for medical tasks. Shang et al. (2019) use a BERT style model for medicine recommendation by learning embeddings for ICD codes. Sänger et al. (2019) use BERT as well as BioBERT (Lee et al., 2020) as base models for ICD code prediction. Clinical BERT (Alsentzer et al., 2019) uses a BERT model fine-tuned on MIMIC III (Johnson et al., 2016) notes and discharge summaries and apply to downstream tasks. Si et al. (2019) compare traditional word embeddings including word2vec, GloVe and fastText to ELMo and BERT embeddings on a range of clinical concept extraction tasks.

Transformer based architectures have led to a large increase in performance on clinical tasks. However, they rely on fine tuning off-the-shelf BERT models, whose vocabulary is very different from clinical text. For example, while clinical BERT (Alsentzer et al., 2019) fine-tune the model on the clinical notes, the authors did not expand the base BERT vocabulary to include more relevant clinical terms. Cui et al. (2019) show that pretraining with many out of vocabulary words can degrade quality of representations as the masked language model task becomes easier when predicting a chunked portion of a word. Si et al. (2019) show BERT models pretrained on the MIMIC-III data dominate those pretrained on non-clinical datasets on clinical concept extraction tasks. This further motivates our hypothesis that pretraining on clinical text will improve the performance on ICD-coding task.

Moreover, existing BERT implementations often require segmenting the notes. For example, Clinical BERT caps at a length of 128 and Sänger et al. (2019) truncate note length to 256. This poses question on how to combine segments from the same document in down-stream prediction tasks, as well as difficulty in learning long-term relationship across segments. Instead, we extend the maximum sequence length to 1,024 and can accommodate common clinical notes as a single input sequence.

3 Methods

3.1 Problem Definition

We approach the ICD tagging task as a multi-label classification problem. We learn a function to map a sequence of input tokens $x = [x_0, x_1, x_2, ..., x_N]$ to a set of labels $y = [y_0, y_1, ...y_M]$ where $y_j \in [0, 1]$ and M is the number of different ICD classes. Assume that we have a set of N training samples $\{(x_i, y_i)\}_{i=0}^{N}$ representing EHR notes with associated ICD labels.

3.2 BERT Pre-training

In this work, we use BERT to represent input text. BERT is an encoder composed of stacked transformer modules. The encoder module is based on the transformer blocks used in (Vaswani et al., 2017), consisting of self-attention, normalization, and position-wise fully connected layers. The model is pretrained with both a masked language model task as well as a next sentence prediction task.

Unlike many practitioners who use BERT models that have been pretrained on general purpose corpora, we trained BERT models from scratch on EHR Notes to address the following two major issues. Firstly, healthcare data contains a specific vocabulary that leads to many out of vocabulary(OOV) words. BERT handles this problem with WordPiece tokenization where OOV words are chunked into sub-words contained in the vocabulary. Naively fine tuning with many OOV words may lead to a decrease in the quality of the representation learned as in the masked language model task as shown by Cui (Cui et al., 2019). Models such as Clinical BERT may learn only to complete the chunked word rather than understand the wider context. The open source BERT vocabulary contains an average 49.2 OOV words per note on our dataset compared with 0.93 OOV words from our trained-from-scratch vocabulary. Secondly, the off-the-shelf BERT models only support sequence lengths up to 512, while EHR notes can contain thousands of tokens. To accommodate the longer sequence length, we trained the BERT model with 1024 sequence length instead. We found that this longer length was able to improve performance on downstream tasks. We train both a small and large architecture model whose configurations are given in table 1. More details on pretraining are described in Section 4.2.1.

We show sample output from our BERT model

Masked Language Model Example

review of systems : gen : no weight loss or gain , good general state of health , no weakness , no fatigue , no fever , good exercise tolerance , able to do usual activities . heent : head : no headache , no dizziness , no lightheadness eyes : normal vision , no redness , no blind spots , no floaters . ears : no earaches , no fullness , normal hearing , no tinnitus . nose and sinuses : no colds , no stuffiness , no discharge , no hay fever , no nosebleeds , no sinus trouble . mouth and pharynx : no cavities , no bleeding gums , no sore throat , no hoarseness . neck : no lumps , no goiter , no neck stiffness or pain . ln : no adenopathy cardiac : no chest pain or discomfort no syncope , no dyspnea on exertion , no orthopnea , no pnd , no edema , no cyanosis , no heart murmur , no palpitations resp : no pleuritic pain , no sob , no wheezing , no stridor , no cough , no hemoptysis , no respiratory infections , no bronchitis .

Figure 1: Example of masked language model task for BERT trained on EHR notes. Highlighted tokens are model predictions for [MASK] tokens

in Figure 1. Our model successfully learns the structure of medical notes as well as the relationships between many different types of symptoms and medical terms.

3.3 BERT ICD Classification Models

3.3.1 BERT Multi-Label Classification

The standard architecture for multi-label classification using BERT is to embed a [CLS] token along with all additional inputs, yielding contextualized representations from the encoder. Assume $H = \{h_{cls}, h_0, h_1, ...h_N\}$ is the last hidden layer corresponding to the [CLS] token and input tokens 0 through N, h_{cls} is then directly used to predict a binary vector of labels.

$$\mathbf{y} = \sigma(\mathbf{W_{out}} \mathbf{h_{cls}}) \tag{1}$$

where $y \in R^M$, W_{out} are learnable parameters and $\sigma()$ is the sigmoid function.

3.3.2 BERT-XML

Multi-Label Attention

One drawback of using the standard BERT multi-label classification approach is that the [CLS] vector of the last hidden layer has limited capacity, especially when the number of labels to classify is large. We experiment with the multi-label attention output layer from AttentionXML (You et al., 2018), and find it improves performance on the prediction task. This module takes a sequence of contextualized word embeddings from BERT $H = \{h_0, h_1, ...h_N\}$ as inputs. We calculate the prediction for each label y_j using the attention mechanism shown below.

$$\mathbf{a_{ij}} = \frac{\mathbf{exp}(\langle \mathbf{h_i}, \mathbf{l_j} \rangle)}{\sum_{\mathbf{i=0}}^{\mathbf{N}} \mathbf{exp}(\langle \mathbf{h_i}, \mathbf{l_j} \rangle)} \tag{2}$$

$$\mathbf{c_j} = \sum_{\mathbf{i=0}}^{\mathbf{N}} \mathbf{a_{ij}} \mathbf{h_i} \tag{3}$$

$$\mathbf{y_j} = \sigma(\mathbf{W_a relu}(\mathbf{W_b c_j})) \tag{4}$$

Where l_j is the vector of attention parameters corresponding to label j. W_a and W_b are shared between labels and are learnable parameters.

Semantic Label Embedding

The output layer of our model introduces a large number of randomly initialized parameters. To further leverage our unsupervised pretraining, we use the BERT embeddings of the text description of each ICD code to initialize the weights of the corresponding label in the output layer. We take the mean of the BERT embeddings of each token in the description. We find this greatly increases the stability of the optimization procedure as well decreases convergence time of the prediction model.

3.4 Baseline Models

3.4.1 Logistic Regression

A logistic regression model is trained with bag-of-words features. We evaluated L1 regularization with different penalty coefficients but did not find improvement in performance. We report the vanilla logistic regression model performance in table 2.

3.4.2 Multi-Head Attention

We then trained a bi-LSTM model with a multi-head attention layer as suggested in (Vaswani et al., 2017). Assume $H = \{h_0, h_1, ..., h_n\}$ is the hidden layer corresponding to input tokens 0 through n from the bi-LSTM, concatenating the forward and backward nodes. The prediction of each label is calculated as below:

$$\mathbf{a_{ik}} = \frac{\mathbf{exp}(\langle \mathbf{h_i}, \mathbf{q_k} \rangle)}{\sum_{\mathbf{i=0}}^{\mathbf{n}} \mathbf{exp}(\langle \mathbf{h_i}, \mathbf{q_k} \rangle)} \tag{5}$$

$$\mathbf{c_k} = (\sum_{\mathbf{i=0}}^{\mathbf{n}} \mathbf{a_{ik}} \mathbf{h_i}) / \sqrt{\mathbf{d_h}} \tag{6}$$

$$\mathbf{c} = concatenate[\mathbf{c_0}, \mathbf{c_1}, ..., \mathbf{c_K}] \quad (7)$$

$$\mathbf{y} = \sigma(\mathbf{W_a c}) \quad (8)$$

$k = 0, ..., K$ is the number of heads and d_h is the size of the bi-LSTM hidden layer. q_k is the query vector corresponding to the kth head and is learnable. $W_a \in R^{M \times K d_h}$ is the learnable output layer weight matrix. Both the query vectors and the weight matrices are initialized randomly.

3.4.3 Other EHR BERT Models

We compare the BERT model pretrained on EHR data (EHR BERT) with other models released for the purpose of EHR applications, including BioBERT (Lee et al., 2020) and clinical BERT (Alsentzer et al., 2019). We compare to the BioBERT v1.1 (+ PubMed 1M) version of the BioBERT model and Bio+Discharge Summary BERT for Clinical BERT. We use the standard multi-label output layer described in section 3.3.1. We choose to compare only with Alsentzer et al. (2019) and not Huang et al. (2019) as they are trained on very similar datasets derived from MIMIC-III using the same BERT initialization.

4 Experiments

4.1 Data

We use medical notes and diagnoses in ICD-10 codes from the NYU Langone Hospital EHR system. These notes are de-identified via the Physionet De-ID tool (Neamatullah et al., 2008), with all personal identifiable information removed such as names, phone numbers, and addresses of both the patients and the clinicians. We exclude notes that are erroneously generated, student generated, belongs to miscellaneous category, as well as notes that contain fewer than 50 characters as these are often not diagnosis related. The resulting data set contains a total of 7.5 million notes corresponding to visits from about 1 million patients, with a median note length of around 150 words and 90th percentile of around 800 tokens. Overall about 50 different types of notes presents in the data. Over 50% of the notes are progress notes, following by telephone encounter (10%) and patient instructions (5%).

This data is then randomly split by patient into 70/10/20 train, dev, test sets. For the models with a maximum length of 512 tokens, notes exceeding the length are split into segments of every 512 tokens until the remaining segment is shorter than the maximum length. Shorter notes, including the ones generated from splitting, are padded to a length of 512. Similar approach applies to models with a maximum length of 1,024 tokens. For notes that are split, the highest predicted probability per ICD code across segments is used as the note level prediction.

We restrict the ICD codes for prediction to all codes that appear more than 1,000 times in the training set, resulting in 2,292 codes in total. In the training set, each note contains 4.46 codes on average. For each note, besides the ICD codes assigned to it via encounter diagnosis codes, we also include ICD codes related to chronic conditions as classified by AHRQ (Friedman et al., 2006; Chi et al., 2011), that the patient has prior to a encounter. Specifically, if we observe two instances of a chronic ICD code in the same patient's records, the same code would be imputed in all records since the earliest occurrence of that code. Notes without the in-scope ICD codes are still kept in the dataset, with all 2,292 classes labeled as 0.

4.2 BERT-Based Models

4.2.1 BERT Pretraining

We trained two different BERT architectures from scratch on EHR notes in the training set. Configurations of both models are provided in Table 1. We use the most frequent 20K tokens derived from the training set for both models. Our vocabulary is select based on the most frequent tokens in the training set. In addition, we extended the max positional embedding to 1024 to better model long term dependencies across long notes. More details given in sections 4.

Models are trained for 2 complete epochs with a batch size of 32 across 4 Titan 1080 GPUs and Nvidia Apex mixed precision training for a total training time of 3 weeks. We found that after 2 epochs the training loss becomes relatively flat. We utilize the popular HuggingFace[2] implementation of BERT. Training and development data splits are the same as the ICD prediction model. Number of epochs is selected based on dev set loss. We compare the pretrained models with those released in the original BERT paper (Devlin et al., 2019) in the downstream classification task, including the off-the-shelf BERT base uncased model and

[2]https://github.com/huggingface/pytorch-transformers

	EHR BERT models	
	small	**big**
hidden size	512	768
# layers	8	12
# attention heads	8	12
intermediate size	2048	3072
activation function	gelu	gelu
hidden dropout	.1	.1
attention dropout	.1	.1
max len	1024	1024

Table 1: configurations for from scratch BERT models. Big configuration matches the base BERT configuration from original paper but has larger max positional embedding

that after fine-tuning on EHR data. The original BERT models only support documents up to 512 tokens in length. In order to extend these to the same 1024 length as other models, we randomly initialize positional embeddings for positions 512 to 1024.

4.2.2 BERT ICD Classification Models

Models are trained with Adam optimizer (Kingma and Ba, 2015) with weight decay and a learning rate of 2e-5. We use a warm-up proportion of .1 during which the learning rate is increased linearly from 0 to 2e-5. After which the learning rate decays to 0 linearly throughout training. We train models for 3 epochs using batch size of 32 across 4 Titan 1080 GPUs and Nvidia mixed precision training. Learning rate and number of epochs are tuned based on AUC of the dev set. All of the ICD classification models optimizes the Binary Cross Entropy loss with equal weights across classes.

4.3 Baseline Models

All baseline models use a max input length of 512 tokens. The multi-headed attention model utilizes pretrained input embeddings with the StarSpace (Wu et al., 2018b) bag-of-word approach. We use the notes in training set as input sequence and their corresponding ICD codes as labels and train embeddings of 300 dimensions. Input embeddings are fixed in prediction task because of memory limitation. Additionally, a dropout layer is applied to the embeddings with rate of 0.1. We use a 1-layer bi-LSTM encoder of 512 hidden nodes with GRU, and 200 attention heads.

The multi-headed attention model is trained with Adam optimizer with weight decay and an initial learning rate of 1e-5. We use a batch size of 8 and trained it up to 2 epochs across 4 Titan 1080 GPUs. Hyperparameters including learning rate, drop out rate and number of epochs are tuned based on AUC of the dev set.

4.4 Results

For each model we report macro AUC and micro AUC. We found that all BERT based models far outperform non-transformer based models. In addition, the big EHR BERT trained from scratch outperforms off-the-shelf BERT models. We believe this speaks to the benefit of pretraining using a vocabulary closer to the prediction task. In addition we find that adding multi-label attention outperforms the standard classification approach given the large number of ICD codes.

We analyze the performance by ICD in figure 2. We achieve very high performance in many ICD classes: 467 of them have an AUC of 0.98 or higher. On ICDs with a low AUC value, we notice that the model can have trouble delineating closely related classes. For example, ICD G44.029-"Chronic cluster headache, not intractable" has a rather low AUC of 0.57. On closer analysis, we find that the model commonly misclassifies this ICD code with other closely related ones such as G44.329-"Chronic post-traumatic headache, not intractable". In future iterations of the model we can better adapt our output layer to the hierarchical nature of the classification problem. Detailed performance of the EHR-BERT+XML model on the test set for the top 45 frequent ICD codes is included in Appendix A.

Furthermore, we find that models trained with max length of 1024 outperform those of 512. EHR notes tend to be long and this shows the value of modeling longer sequences for EHR applications. However, training time for the longer sequence models is roughly 3.5 times that of the shorter ones. In order to scale training and inference to longer patient histories with multiple notes it is necessary to develop faster and more memory efficient transformer models.

In addition, while the BERT based models do better than standard models on average, we see very pronounced gains in lower frequency ICDs. Table 3 compares the macro AUC for all ICD codes with fewer than 2000 training examples (757 ICDs in total) of the best BERT and non-BERT models. Note that the best non-BERT model does worse on

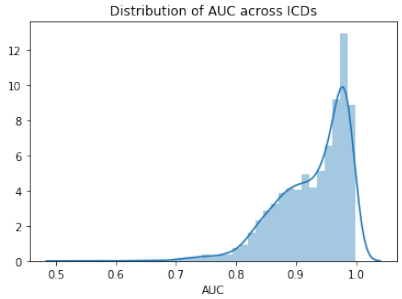

Figure 2: distributions of AUCs across ICD 10 codes

Prediction Visualization : Right Hip Fracture
physical therapy progress note ... right hip pain mnn . nnn nnn . nn treatment diagnosis : r hip pain , s / p labral repair with [UNK] primary insurance : [UNK] group subscriber number : @ subnum @ secondary insurance : n / a primary language spoken : english [UNK] nn [UNK] interpreter present : no any relevant changes to medical status : no recent falls : no precautions : see surgical protocol in media file , currently phase ii

Figure 3: visualization of XML-BERT attention layer. Darker colors correspond to higher attention weights.

this set compare to its performance on all ICDs, while the best BERT model performs better on average on the lower frequency ones. This further illustrates the value of the unsupervised pretraining and provides a good motivation to expand our analysis to even less frequent ICD codes in future works.

4.5 Visualization

For many machine learning applications, it is important to enable users to understand how the model comes to the predictions, especially in healthcare industry where decisions have serious implications for patients. To understand the model predictions, we can visualize the attention weights of the XML output layer of each of the classes. In figure 3 we show attention weights corresponding to a note coded with right hip fracture. The model successfully identify key terms such as 'right hip pain', 'hip pain' and 's/p labral'.

In addition, we examine the attention weights between tokens in the BERT encoder. In figure 4 we show the attention scores between each word of the note of the final layer of the BERT encoder of a note with 735 tokens. We observe that, while probability mass tends to concentrate between sequentially close tokens, a significant amount of probability mass also comes from far away tokens. In addition we see specialisation of different heads. For example, head 0 (row 1, column 1 in figure 4) tends to capture long range contextual information such as the note type and encounter type which are typically listed at the beginning of each note; while head 5 (row 1, column 1 in figure 4) tends to model local information. We believe the increase in performance can partially be attributed to the ability to model long range contextual information.

5 Conclusion

Automatic ICD coding from medical notes has high value to clinicians, healthcare providers as well as researchers. Not only does auto-coding have high potential in cost- and time-saving, but more accurate and consistent ICD coding is necessary to facilitate patient care and improve all downstream healthcare EHR based research.

We demonstrate the effectiveness of models leveraging the most recent developments in NLP with BERT as well as multi-label attention on ICD classification. Our model achieves state of the art results using a large set of real world EHR data across many ICD classes. In addition we find that domain specific pretrained BERT model outperforms BERT models trained on general purpose corpora. We note that the off-the-shelf WordPiece tokenizer can naively split domain-specific yet OOV words and resulting in a BERT model focusing on word completion, while using a specific EHR vocabulary seem to help overcome the problem. Lastly, we also observe the benefit of modeling longer sequences.

On the other hand, the current work has several limitations. Most importantly, while we have found that modeling longer term dependencies improves performance, it comes at a large cost of training time. Doubling the input length roughly triples the training and inference time. For many applications this increase in computational demand may offset the gain in model performance. This motivates further exploration on efficient variants of the self-attention modules to accommodate longer input length in similar tasks. Additionally, adding XML to the BERT architecture generates significant yet rather marginal performance improvement (Micro-AUC improvement of 0.002 for EHR BERT Big model with maximum input length of 1024). This also increases the computation complexity

	AUC	
	Micro	**Macro**
Logistic Reg (max length 512)	0.932	0.815
Multi-head Attn (max length 512)	0.941	0.859
BERT (max length 512)	0.954	0.895
BERT (max length 1024)	0.955	0.898
Finetuned BERT (max length 1024)	0.958	0.903
BioBERT	0.960	0.908
clinical BERT	0.961	0.904
EHR BERT Small (max length 512)	0.959	0.897
EHR BERT Small (max length 1024)	0.965	0.918
EHR BERT Small + XML (max length 1024)	0.968	0.924
EHR BERT Big (max length 512)	0.964	0.917
EHR BERT Big (max length 1024)	0.968	0.925
EHR BERT Big + XML (max length 512)	0.967	0.919
EHR BERT Big + XML (max length 1024)	**0.970**	**0.927**

Table 2: Test set model performance. The largest confidence interval calculated was only 4e-5 so all results shown are statistically significant.

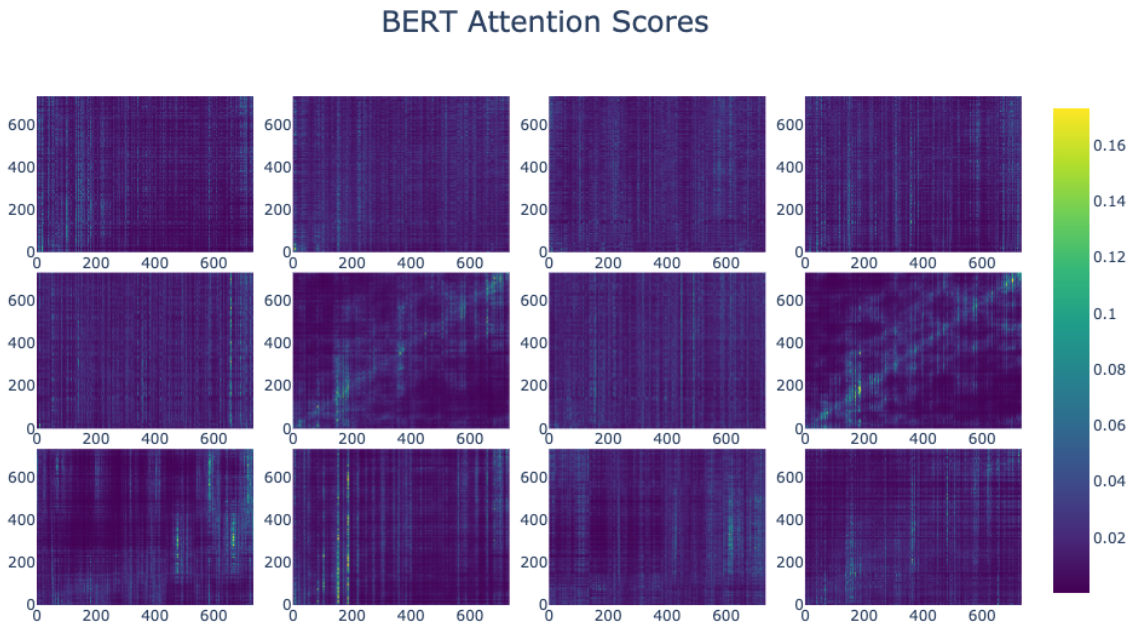

Figure 4: The attention weights of each head for each head in the last layer of the BERT encoder. Brighter color denotes higher attention score. We see some heads specialize in modeling local information(row 2, column 2) while some specialize in passing global information (row 1, column 1). Suggest print in color.

Macro AUC - Low Frequency ICDS	
Multi-head Att	0.825
Big EHR BERT + XML	0.933

Table 3: Macro AUC of the best non transformer model and the best BERT model compared using only ICDs with fewer than 2000 examples. Note that the non pre-trained model performs worse on this section of the dataset while the BERT model performs just as good.

and more efficient alternatives, such as hierarchical based methods as evaluated in Azam et al. (2019), are promising candidates.

For future works, we plan on expanding our model to more classes with fewer records as we observe the model performing as well on low frequency ICD codes as on the high frequency ones. To address limitations discussed above, we plan on adapting our model to utilize the hierarchical nature of the ICD codes as well as developing memory efficient models that can support inference across long sequences.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. 2019. Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Hamza A Ashfaq, Corey A Lester, Dena Ballouz, Josh Errickson, and Maria A Woodward. 2019. Medication accuracy in electronic health records for microbial keratitis. *JAMA ophthalmology*.

Sheikh Shams Azam, Manoj Raju, Venkatesh Pagidimarri, and Vamsi Chandra Kasivajjala. 2019. Cascadenet: An lstm based deep learning model for automated icd-10 coding. In *Future of Information and Communication Conference*, pages 55–74. Springer.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. 2018. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Mei-ju Chi, Cheng-yi Lee, and Shwu-chong Wu. 2011. The prevalence of chronic conditions and medical expenditures of the elderly by chronic condition indicator (cci). *Archives of gerontology and geriatrics*, 52(3):284–289.

Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. 2017. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186.

Rémi Flicoteaux. 2018. Ecstra-aphp@ clef ehealth2018-task 1: Icd10 code extraction from death certificates. In *CLEF (Working Notes)*.

Bernard Friedman, H Joanna Jiang, Anne Elixhauser, and Andrew Segal. 2006. Hospital inpatient costs for adults with multiple chronic conditions. *Medical Care Research and Review*, 63(3):327–346.

Beacon Nation Learning Guide. 2013. Capturing high quality electronic health records data to support performance improvement. *Implementation Objective*, 2:16.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

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. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zulfat Miftahutdinov and Elena Tutubalina. 2018. Deep learning for icd coding: Looking for medical concepts in clinical documents in english and in french. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 203–215. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):32.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Mario Sänger, Leon Weber, Madeleine Kittner, and Ulf Leser. 2019. Classifying german animal experiment summaries with multi-lingual bert at clef ehealth 2019 task 1. In *CLEF (Working Notes)*.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5953–5959. AAAI Press.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2018. Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sarah Wiegreffe, Edward Choi, Sherry Yan, Jimeng Sun, and Jacob Eisenstein. 2019. Clinical concept extraction for document-level coding. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 261–272.

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, et al. 2018a. Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association*, 25(5):530–537.

Ledell Yu Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018b. Starspace: Embed all the things! In *AAAI*, pages 5569–5577.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated icd coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076.

Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K Khanna, Jacek B Cywinski, Kamal Maheshwari, et al. 2019. Multimodal machine learning for automated icd coding. In *Machine Learning for Healthcare Conference*, pages 197–215. PMLR.

Ronghui You, Suyang Dai, Zihan Zhang, Hiroshi Mamitsuka, and Shanfeng Zhu. 2018. Attentionxml: Extreme multi-label text classification with multi-label attention based recurrent neural networks. *arXiv preprint arXiv:1811.01727*.

A ICD Performance for frequent ICDs

ICD-10	Count	AUC	ICD-10	Count	AUC	ICD-10	Count	AUC
I10	391298	0.877	M81.0	46528	0.868	I73.9	24992	0.898
E78.5	291430	0.863	Z00.00	43136	0.988	F41.1	26032	0.842
I25.10	131280	0.904	I48.0	40032	0.923	E11.65	22912	0.882
E11.9	132150	0.874	Z51.11	42336	0.970	F17.200	21408	0.849
K21.9	133422	0.816	G47.33	38592	0.846	Z23	23296	0.977
E55.9	114322	0.839	N40.0	34688	0.896	M17.0	19648	0.885
E03.9	91072	0.840	J45.909	34496	0.835	M54.5	21296	0.973
E66.9	80454	0.838	E66.01	30080	0.877	C50.912	21984	0.944
E78.00	72740	0.862	N18.3	28784	0.888	M06.9	18160	0.913
F41.9	71836	0.835	I48.2	26592	0.936	C50.911	22544	0.945
F32.9	68172	0.824	Z95.0	24592	0.930	C50.919	22880	0.950
I48.91	61056	0.922	G62.9	25632	0.853	R53.83	19616	0.968
G89.29	49600	0.838	M17.9	22992	0.854	I35.0	17536	0.917
J44.9	48224	0.881	E78.2	24096	0.876	Z51.12	20784	0.963
M19.90	47968	0.830	I34.0	21600	0.900	J45.20	18848	0.856

Table 4: Individual ICD Performance for most frequent ICDs, Big EHR BERT + XML. Count is the total positive examples we have observed in our test set.