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

Zachariah Zhang* NYU Langone Health zz1409@nyu.edu

Jingshu Liu* NYU Langone Health jingshu.liu@nyu.edu

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

Abstract

Clinical interactions are initially recorded and documented in free text medical notes. 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 and extremely time-consuming and expensive for hospitals. In this paper, we propose a machine learning model, BERT-XML, for large scale automated ICD coding from 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 adapt the BERT architecture for ICD coding with multi-label attention. While other works focus on small public medical datasets, we have produced the first large scale ICD-10 classification model using millions of EHR notes to predict thousands of unique ICD codes.

1 Introduction

Information embedded in Electronic Health Records (EHR) have been a focus of 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 (Savova et al., 2010; Soysal et al., 2017; Aronson and Lang, 2010; Wu 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). Yet to the best of our knowledge, application 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 with non-clinical datasets (Lee et al., 2019). In addition, these models are adapted directly the original BERT models released in (Devlin et al., 2018) which use 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. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) utilizes unsupervised pretraining procedures to produce meaningful representation of the input sequences, 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. Additionally, the attention layer provides a natural mechanism to identify part of the text that impacts final prediction.

Compare to other works using BERT for disease identification, we emphasize on the follow-

*Equal contribution

[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.

ing aspects: 1) **Large cohort pretraining.** We train BERT model from scratch on over 5 million EHR notes, and find it outperforms off-the-shelf or fine-tuned BERT using off-the-shelf vocabulary. 2) **Long input sequence.** We 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) **EHR Specific Vocabulary.** While other implementations use the vocabulary from the original BERT, we train with a vocabulary specific to EHR to build better representations of EHR notes. 4) **Extreme 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. 5) **Novel multi-label embedding initialization.** We apply an innovative initialization method as described in 3.3.2, that greatly improves training stability of the multi-label attention.

2 Related Works

2.1 CNN, LSTM based Approaches and Attention Mechanisms

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). In (Flicoteaux, 2018), authors use a text CNN as well as lexical matching to improve performance for rare ICD labels. In (Xu et al., 2018), 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.

Many works further incorporates the attention mechanisms as introduced in (Bahdanau et al., 2014), 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) uses a hierarchical combination of LSTM's to encode EHR text and then use attention with encodings of the text descriptions of ICD codes to make predictions.

While these models have achieved impressive results, they usually fall short in modeling the complexity of EHR data in terms of the number of ICD codes predicted. For example, (Shi et al., 2017) limited their predictions to the 50 most frequent codes and (Xu et al., 2018) predicted 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 of learning 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, in learning representations of words.

Recently, there have been several approaches 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., 2018). 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) uses a BERT style model for medicine recommendation by learning embeddings for ICD codes. (Sänger et al., 2019) uses BERT as well as BioBERT (Lee et al., 2019) as base models for ICD code prediction. Clinical BERT (Alsentzer et al., 2019) uses a BERT model fine-tuned on MIMIC III notes and discharge summaries and apply to downstream 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-tuned the model on the clinical notes, the authors did not con-

sider expand the base BERT vocabulary to include more relevant clinical terms. Cui (Cui et al., 2019) shows 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. Moreover, notes processed are often capped at a relatively short length. For example, Clinical BERT uses a length of 128 and (Sänger et al., 2019) truncates note length to 256. In addition most of these papers only train on the relatively small MIMIC-III (Johnson et al., 2016) dataset which contains only 60k patients. These patients are also exclusively from critical care units which only represent a small subset patients in the EHR system for most hospitals.

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 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. Self-attention avoids the vanishing gradient and inductive bias associated with sequential models such as LSTM or GRU. The model is pretrained with both a masked language model task as well as a next sentence prediction task. In the former, tokens from the input sequence are randomly replaced with a [MASK] token and the model learns to predict the removed tokens. In the latter, the model produce a binary classification to predict if one sentence follows another, in order to better model longer term dependencies.

Unlike many practitioners who use BERT models that have been already pretrained on a wide corpus, we trained BERT models from scratch on EHR Notes to address the following two major issues. Firstly, healthcare data contains a specific vocabulary that is not common within a general pretraining corpus, 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 show 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

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. Colored tokens are model predictions for [MASK] tokens

We show sample output from our BERT model 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.

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

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.

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

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

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

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

We notice randomly initialized multi-label attention takes long to start learning. Rather, we use the idea of semantic label embeddings (Pappas and Henderson, 2019) to initialize the embeddings of each label $L = \{l_0, l_1, ...l_M\}$ with the BERT encoding of the plain text description of the associated ICD code. We take the mean of the BERT embeddings of each token in the description. We find this greatly increases stability during training.

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:

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

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

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

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

$k = 0, ..., K$ 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 our pretrained EHR BERT model with others models that have been released for the purpose of EHR applications. We compare our BERT model against BioBERT (Lee et al., 2019) as well as clinical BERT (Alsentzer et al., 2019). We compare using 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 de-identified medical notes and diagnoses in ICD-10 codes from the Anonymous Institution EHR system. 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. We use a total of 7.5 million notes corresponding to visits from about 1 million patients. This data is then randomly split by patient into 70/10/20 train, dev, test sets. Notes are padded to or split to chunks of a maximum length of 512 or 1,024, depending on the model. For notes that are split, the highest predicted probability per ICD code across chunks 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 that 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.

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 for both models are provided in Table 1. We use the most frequent 20K words derived from the training set for both models. In addition, we extended the max positional embedding to 1024 to better model long term dependencies across long notes.

Models are trained for 2 complete epochs with a batch size of 32 across 4 Titan 1080 GPUs and Nvidia Apex mixed precision training. We utilize the popular HuggingFace[2] implementation of BERT for training. 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., 2018) in the downstream classification task, including the off-the-shelf BERT base uncased model and 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.

[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

4.2.2 BERT ICD Classification Models

Models are trained with Adam optimizer (Kingma and Ba, 2014) 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.

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., 2017) 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 we find that the big EHR BERT trained from scratch outperform off-the-shelf BERT models. We believe this speaks to the benefit of pretraining using a vocabulary closer to EHR notes. 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 AUC of 0.98 or higher. On ICDs with 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 Table 4.

Furthermore, we find that models trained with max length of 1024 outperform those of 512. EHR notes tend to be very long and this shows the value to have BERT models that can model longer length 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

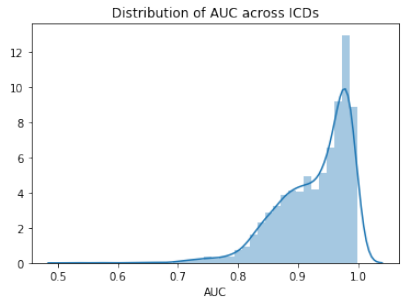

Figure 2: distributions of AUCs across ICD 10 codes

that the best non-BERT model does worse on 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 good motivation to expand our analysis to even less frequent ICD codes in future work.

4.5 Visualization

For many machine learning applications, it is important for users to be able to understand how the model comes to the predictions, especially in healthcare industry where decisions have serious implications for patients. To understand our predictions, we show the attention weights of the XML output layer for 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 for 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, there is a significant amount of probability mass that comes from far away tokens. In addition we see specialisation of different heads. We see that 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. In addition head 5 (row 1, column 1 in figure 4) models local information. We believe some of the increase in performance can be attributed to long range modeling of contextual information.

	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.

Table 3: Model Performance - Low Frequency ICDs

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

Prediction Visualization : Right Hip Fracture

physical therapy progress note referring physician : name , name , md primary care physician : name name name medical diagnosis : icd - nn - cm icd - n - cm n . 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 softmax value

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 have developed a model for ICD classification that leverages the most recent developments in NLP with BERT as well as multi-label attention. Our model achieves state of the art results using a large dataset of real EHR data across many ICDs. In addition we find that our domain specific BERT model is able to outperform open domain BERT models by modeling longer sequences as well as using a specific EHR vocabulary to overcome the WordPiece tokenizer problem. Our model is able to get impressive results on low frequency ICDs and we plan on expanding our model to more classes in future works. In addition, we plan on adapting our model to address the hierarchical nature of ICDs as well as developing memory efficient models that can support inference across longer sequences.

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.

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

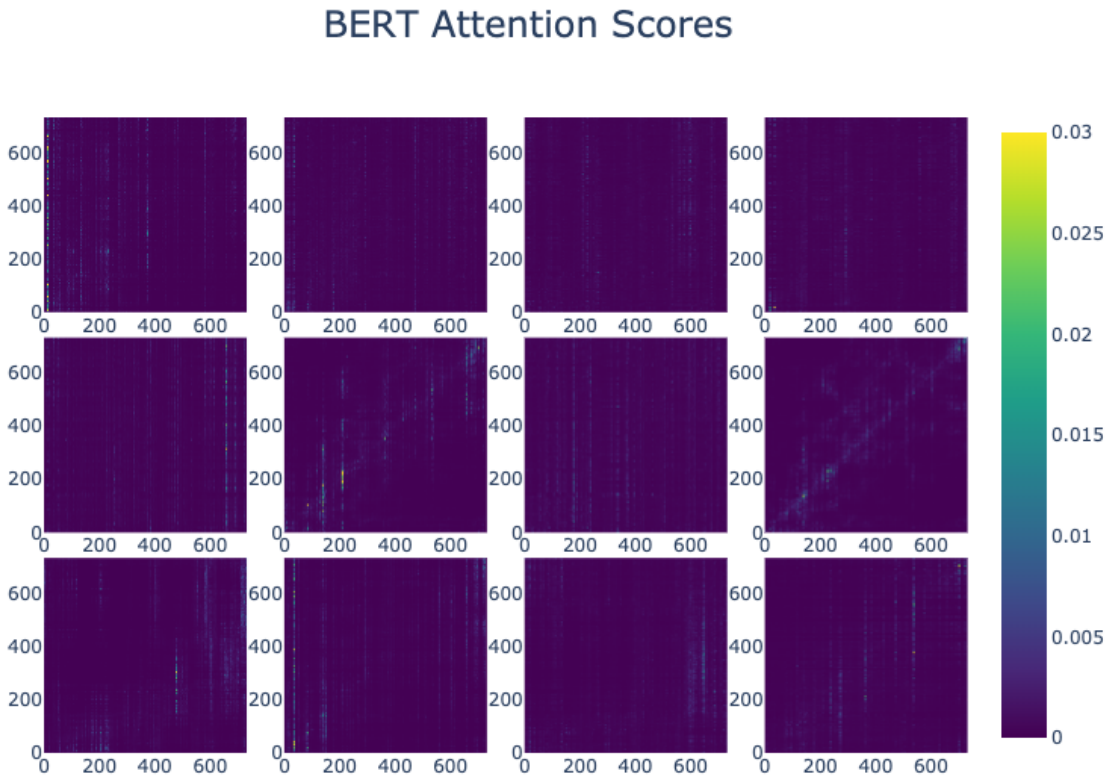

Figure 4: The attention weights of each head for each head in the last layer of the BERT encoder. Brighter color denotes higher attention score. We see some heads specialize in modeling local information(row 2, column 2) while some specialize in passing global information (row 1, column 1)

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

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*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv: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. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

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.

Nikolaos Pappas and James Henderson. 2019. Gile: A generalized input-label embedding for text classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.

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. *arXiv preprint arXiv:1802.05365*.

Mario Sänger, Leon Weber, Madeleine Kittner, and Ulf Leser. 2019. Classifying german animal experiment summaries with multi-lingual bert at clef ehealth 2019 task.

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. *arXiv preprint arXiv:1906.00346*.

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*.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2017. 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.

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, et al. 2018. 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 Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2017. Starspace: Embed all the things! *arXiv preprint arXiv:1709.03856*.

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, Pengtao Xie, and Eric Xing. 2018. Multimodal machine learning for automated icd coding. *arXiv preprint arXiv:1810.13348*.

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*.

Upon acceptance, the appendices come after the references, as shown here.