Predicting ICD-10 Diagnoses Using seq2seq with GRU and Attention & Encoder BERT Models

Jessica Elrefaei
Kevin Hitt
Yaswanth Chandolu
Sai Charan Dasari
Ashwin Kumar Reddy Guduru

ISM 6251 Data Science Programming
Dr. Mohammadreza Ebrahimi
November 9, 2023

Research Question

The focus of this paper is centered on the application of deep learning architectures to process and interpret medical data, specifically diagnoses codes from the International Classification of Diseases, Tenth Revision (ICD-10). This application of deep learning is an effort to address our main research question, how could sequences of diagnosis codes be predicted for future patient visits? Within the context of Beth Israel Hospital in Boston, Massachusetts we found that of the patients who died while admitted between 2008 and 2019, 10% experienced cardiac arrest. Of those who experienced cardiac arrest, 75% of those patients passed away during their stay at the hospital. If we could analyze a patient's medical history in the form of the chronological sequence ICD-10 diagnoses, could we accurately predict the sequences of diagnoses that might follow from a subsequent hospital visit in order to improve treatment options, patient health outcomes, and decrease mortality rates?

Business Objectives

The ability to predict this would have benefits across various stakeholders, including hospitals, insurance companies, and patients. Foremost, patient care would be more proactive and would allow for healthcare providers to intervene preemptively before severe health events (like cardiac arrest). This preventative approach is an improvement to the existing, reactive nature of current medical care. In the same regard, hospital resources would be optimized by limiting lengthy hospital stays that could otherwise be prevented. While hospitals would be incentivized to employ this approach and improve their quality of care metrics, the main objective here is to improve the patient experience and ultimately prevent mortality.

Research Background

The eight papers presented below collectively dive into the domain of healthcare and medical coding with a focus on leveraging advanced technologies, particularly deep learning and natural language processing, to improve the efficiency and accuracy of processes critical to the healthcare industry. These advancements aim to address challenges such as ICD-10 coding, automated diagnosis prediction, and cancer prognosis, showcasing the potential to revolutionize healthcare quality, data management, and patient outcomes. The research demonstrates a commitment to overcoming complexities, including language variations, data insufficiency, model interpretability, and biases, while emphasizing the importance of clinical integration and collaborations between experts in deep learning and medical fields. In summary, the research serves as a foundation for the integration of cutting-edge technology into healthcare, offering substantial potential benefits across multiple domains within the industry.

CASCADENET: An LSTM Based Deep Learning Model for Automated ICD-10 Coding

This research paper presents an innovative method for classifying ICD-10 codes within diagnostic strings, employing a cascading hierarchical LSTM network approach. The assignment of ICD-10

codes plays a pivotal role in clinical applications, but the conventional manual coding process is not only time-consuming but also prone to errors. Leveraging deep learning techniques, particularly LSTM networks, emerges as a promising solution to address these challenges.

The proposed cascading hierarchical LSTM network model exhibits impressive results, achieving an accuracy rate of 72.05% for seven-digit ICD-10 codes and an even higher accuracy of 89.95% for three-digit codes. These outcomes underscore the potential of this technology in enhancing clinical applications across various domains. This includes refining the accuracy of ICD code assignments, bolstering the management of patient records, and providing valuable support for treatment decisions within the healthcare sector.

A Deep Learning Framework for Automated ICD-10 Coding

In the paper titled "Deep Learning Framework for Automated ICD-10 Coding," authored by Abdelahad Chraibi and his colleagues, an innovative approach to automating the process of assigning ICD-10 codes to electronic health records (EHRs) is presented. The primary objective of this research is to increase the efficiency and accuracy of this critical task within the healthcare domain. To tackle the challenges associated with extracting information from free-text entries in EHRs, the authors leverage the power of natural language processing (NLP) and deep learning methodologies.

By employing these advanced techniques, their proposed framework attains an impressive average accuracy rate of 84.21%. This achievement holds substantial promise for enhancing the overall quality of healthcare services and streamlining the coding procedures. The successful implementation of this approach makes a significant and valuable contribution to the realm of medical coding, potentially revolutionizing the way diagnostic coding is performed in healthcare settings.

A Comparison of Deep Learning Methods for ICD-Coding of Clinical Records

The paper delves into the utilization of neural networks and deep learning techniques to classify medical records according to ICD codes. It underscores the importance of employing specialized data sources such as MIMIC-III for training and evaluating the models. The review places particular emphasis on the use of evaluation metrics, notably micro F1 scores, as a yardstick for assessing the efficacy of different approaches.

The findings from a range of studies collectively suggest that deep learning models exhibit considerable promise in the realm of ICD code classification. Notably, techniques involving multi-code classification and recognizing hierarchical relationships among codes contribute significantly to the improvement of accuracy in this context. This research points to the potential of deep learning to enhance the classification of medical records based on ICD codes, with the promise of more accurate and efficient healthcare data management.

Automatic ICD-10 Coding Using Prescribed Drugs Data

In this research paper, a groundbreaking method for automated ICD-10 coding from electronic health records (EHRs) is introduced, leveraging the capabilities of deep learning. The article addresses the intricate nature of ICD-10 coding, often entailing challenges related to free-text input and potential ambiguities. To tackle these issues, the study harnesses a deep learning framework and harnesses a vast dataset comprising over 134,000 EHRs to construct a robust model.

The study yields highly promising outcomes, showcasing an impressive average accuracy rate of 84.21%. Additionally, the model achieves a precision score of 84.52%, recall of 82.61%, and an F1 score of 82.79%. The research focuses intently on mitigating noise and bias within the dataset, resulting in a significantly enhanced machine coding system performance. This pioneering approach demonstrates its potential to not only elevate the quality of healthcare but also greatly enhance the efficiency of coding processes in this domain.

Applying Deep Learning Model to Predict Diagnosis Code of Medical Records

This research paper delves into the pivotal task of predicting ICD-10 codes from the wealth of clinical notes present in medical records. To accomplish this, the study harnesses the power of deep learning models, with a particular focus on convolutional neural networks (CNN) and cutting-edge natural language processing techniques, aiming to automate and optimize this intricate process.

Through rigorous training and testing of the CNN model using clinical data, the authors have achieved noteworthy levels of precision, recall, and F-scores, with cardiology standing out as the domain with the most outstanding performance.

The potential applications of this research in the realm of healthcare are nothing short of transformative. They encompass the automation of medical coding, early identification of disease risks, and the enhancement of treatment plans through the generation of data-driven insights. This research has the potential to spark a revolution in medical record keeping and healthcare efficiency, ultimately leading to improved patient outcomes and reduced costs across the healthcare landscape.

Sequential Diagnosis Prediction with Transformer and Ontological Representation

In this paper, the author presented SETOR, an innovative end-to-end robust transformer-based model designed for the task of sequential diagnosis prediction within the realm of healthcare analytics. Through a series of experiments conducted on real-world healthcare datasets, including the widely used MIMIC-III and MIMIC-IV, our research reveals that SETOR consistently

outperforms alternative approaches for diagnosis prediction, particularly when dealing with extensive datasets.

SETOR employs a unique approach that involves the incorporation of medical ontology, realized through graph-embedding and ontological encoding, aimed at addressing the common challenge of insufficient data in healthcare analytics. Furthermore, it adeptly manages the irregular intervals between patient visits using neural ordinary differential equations (ODEs). This combination of medical ontology and ODEs yields substantial improvements in predictive performance and data handling, effectively showcasing SETOR's remarkable effectiveness within the field of healthcare analytics.

Deep Learning in Cancer Diagnosis and Prognosis Prediction: A Minireview on Challenges, Recent Trends, and Future Directions

This article delves into the complexities and challenges associated with the integration of deep learning techniques in the context of cancer diagnosis and prognosis prediction. It underscores the critical factors that must be considered, including the quality of available data, the intricacies of model complexity, the interpretability of results, the potential introduction of biases, and the seamless integration of these technologies into clinical practices.

Recent developments in the field of oncology have seen the application of deep learning methodologies across various cancer types, facilitating advancements in diagnosis, prognosis, and classification. Looking ahead, the direction of research in this domain entails the development of more interpretable models, the incorporation of deep learning into the clinical decision-making process, the exploration of personalized medicine approaches, and the synergistic fusion of deep learning with genomics and medical imaging for more accurate cancer diagnosis and prognosis prediction.

This article underscores the pivotal role that collaboration between experts in deep learning and cancer research plays in driving progress within this exciting and transformative field.

Interpretable Deep Learning to Map Diagnostic Texts to ICD-10 Codes

This paper introduces an innovative and interpretable deep learning method designed to associate diagnostic text with ICD-10 codes. The proposed approach utilizes a Convolutional Neural Network (CNN) to extract relevant features from the text and employs a Recurrent Neural Network (RNN) to capture sequential dependencies within the data. The final step involves a multi-label classifier responsible for predicting the appropriate ICD-10 codes for the given diagnostic text.

Compared to other existing methods, this approach proves to be superior when dealing with multilingual ICD-10 coding, and it produces results that are interpretable. The model's output

demonstrates a clear alignment between the original diagnostic text and the assigned ICD-10 codes, resulting in F-measures of 0.838 for French, 0.963 for Hungarian, and 0.952 for Italian. The challenges in clinical document coding are multifaceted, including language variations, the informality of spontaneous writing in medical records, the complexity of large-scale classification tasks, and the intricacies of establishing accurate alignments between text and codes.

Dataset

Our dataset for this analysis originated from the Medical Information Mart for Intensive Care (MIMIC-IV) database, a patient centric source from admissions at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. MIMIC-IV draws its data from a pair of hospital-based database systems: an electronic health record (EHR) system implemented hospital-wide and a specialized clinical information system dedicated to intensive care units. This source has been deidentified according to the Health Insurance Portability and Accountability Act (HIPAA). While this data is feature rich with qualities like demographics, timestamps, and vital signs, the files were merged and grouped by patient, (patients, admissions, diagnoses files) to result in a simpler dataset appropriate for our models. The final dataset includes patient identifiers and a corresponding list of their ICD-10 codes. While the lists of codes for each patient represent their diagnosis history and are grouped logically by visit, there is no further ordering of the codes chronologically or otherwise.

Descriptive Statistics

Table 1 shows a summary of descriptive statistics of the dataset. Table 2 shows the average age of patients based on their ethnicity. Table 3 shows the top 10 ICD-10 codes by patient count. Figures 1, 2, and 3 show the distribution of gender, the age distribution, and the racial distribution, respectively.

Table 1

Descriptive Statistics of the Dataset

Total Number of Patients (ICD-9 and ICD-10)	180,640
Number of Patients with ICD-10 Codes	80,213
Number of Unique ICD-10 Codes	16,757
Number of ICD-10 code patients who died while admitted	3,360
Number of ICD-10 code patients who died who had cardiac arrest in their final visit (cardiac arrest subpopulation)	363

Total number of codes (non-distinct) in cardiac arrest subpopulation	17,379
Longest total sequence of codes across all visits for a cardiac arrest patient	409
Shortest total sequence of codes across all visits for a cardiac arrest patient	6
Maximum number of ICD-10 codes in one visit for a cardiac arrest patient	39
Minimum number of ICD-10 codes in one visit for a cardiac arrest patient	1

Table 2

Average Age by Race

Race	Mean Age
HISPANIC/LATINO - MEXICAN	48.82
WHITE - BRAZILIAN	50.09
ASIAN - KOREAN	52.02
HISPANIC/LATINO - DOMINICAN	52.26
HISPANIC/LATINO - SALVADORAN	52.40
HISPANIC/LATINO - GUATEMALAN	52.93
HISPANIC/LATINO - PUERTO RICAN	53.91
HISPANIC/LATINO - HONDURAN	54.05
BLACK/AFRICAN	54.93
BLACK/AFRICAN AMERICAN	55.90

OTHER	56.39
ASIAN	56.64
ASIAN - ASIAN INDIAN	57.14
HISPANIC OR LATINO	57.69
AMERICAN INDIAN/ALASKA NATIVE	58.43
ASIAN - SOUTH EAST ASIAN	58.44
BLACK/CARIBBEAN ISLAND	58.58
PORTUGUESE	59.02
HISPANIC/LATINO - COLUMBIAN	59.58
HISPANIC/LATINO - CENTRAL AMERICAN	59.87
BLACK/CAPE VERDEAN	59.92
PATIENT DECLINED TO ANSWER	60.52
SOUTH AMERICAN	60.57
UNABLE TO OBTAIN	60.80
WHITE	62.29
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	62.51
UNKNOWN	63.03

HISPANIC/LATINO - CUBAN	63.07
WHITE - OTHER EUROPEAN	63.18
ASIAN - CHINESE	63.52
WHITE - EASTERN EUROPEAN	65.00
WHITE - RUSSIAN	75.71
MULTIPLE RACE/ETHNICITY	85.00

Table 3

Top 10 ICD-10 Codes by Patient Count

Diagnosis Code	Patient Count	Diagnosis Description
I10	31521	Essential (primary) hypertension
E785	27903	Hyperlipidemia, unspecified
Z87891	21356	Personal history of nicotine dependence
K219	19067	Gastro-esophageal reflux disease without esophagitis
F329	16476	Major depressive disorder, single episode, unspecified
F419	14223	Anxiety disorder, unspecified
I2510	13438	Atherosclerotic heart disease of native coronary artery without angina pectoris
N179	13394	Acute kidney failure, unspecified
Y929	11706	Unspecified place or not applicable

E119	10302	Type 2 diabetes mellitus without complications

Figure 1

Gender Distribution

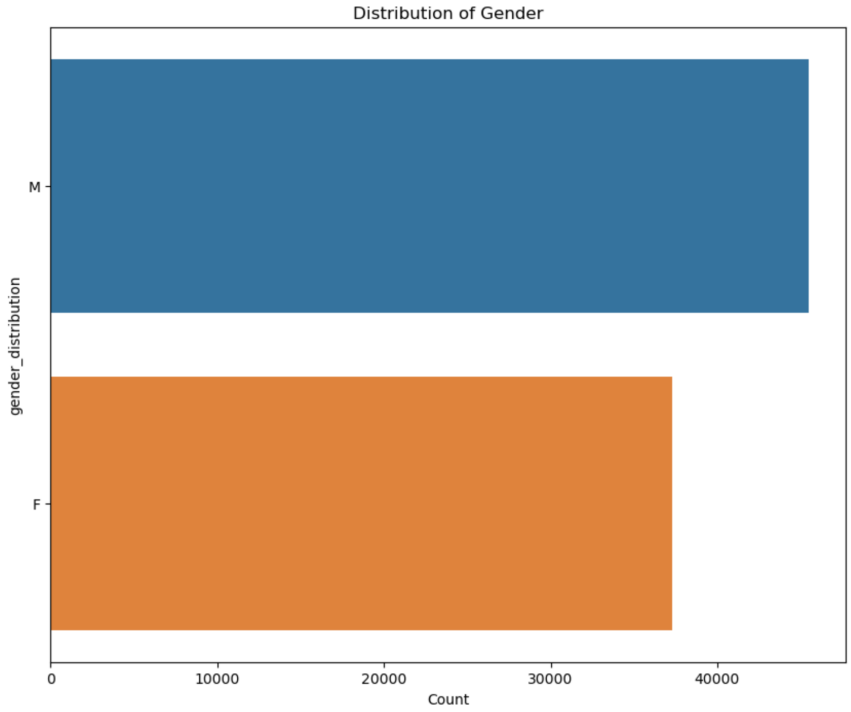


Figure 2

Age Distribution

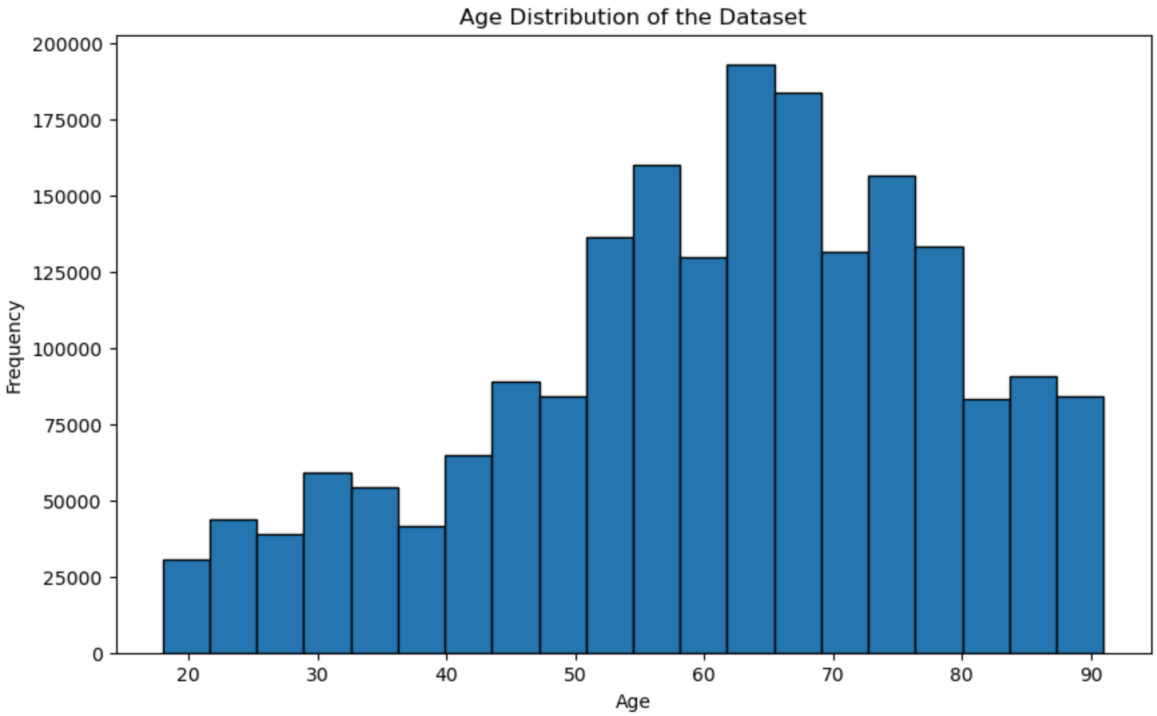
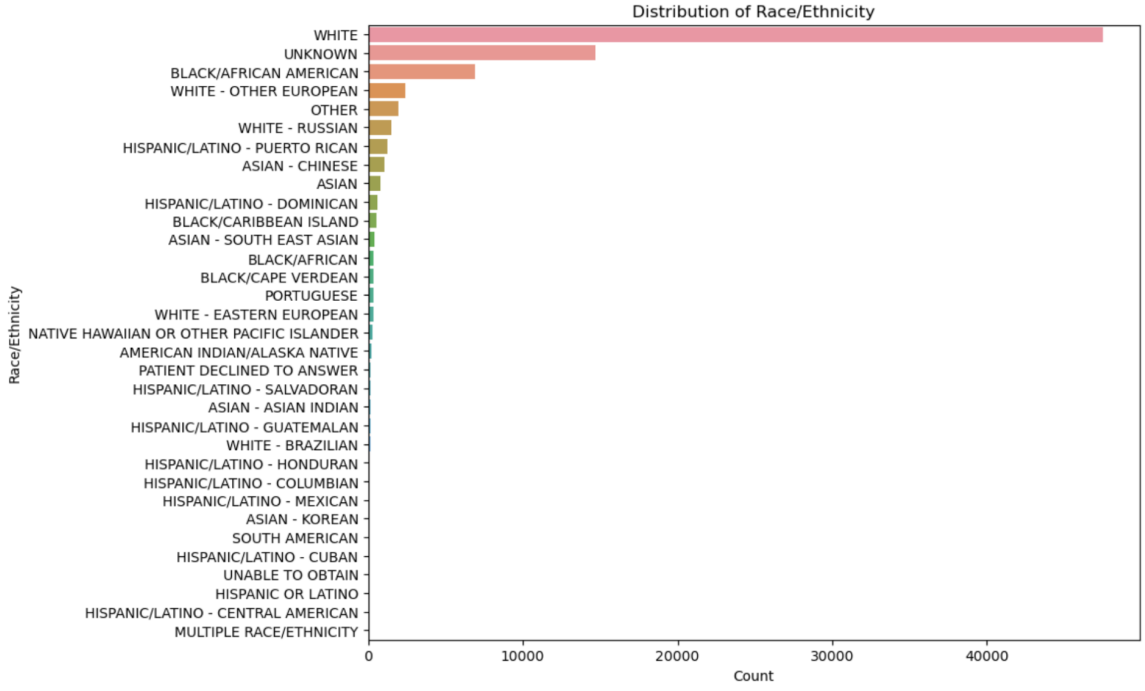


Figure 3

Racial and Ethnic Distribution



Methodology

Overview

Two models were used for the analysis. The first was a sequence-to-sequence encoder decoder GRU-based model with a Bahdanau attention mechanism, implemented with Pytorch and Jupyter Notebooks. This model was a text-generative AI model which used prior sequences of ICD-10 codes to output the next predicted sequences of ICD-10 codes. The second model was a Bidirectional Encoder Representation Transformer (BERT) model, also implemented in Pytorch and Jupyter Notebooks. This is an encoder only model with multi-headed attention and positional encoding which classifies if a sequence of ICD-10 codes follows the previous sequence of ICD-10 codes, essentially treating diagnosis prediction as a binary text classification problem.

Data Preprocessing

Extensive data preprocessing was conducted. Four CSV files (Patients, Diagnoses, Diagnosis Descriptions, and Admissions) from the original MIMIC-IV database were joined on multiple keys (subject_id, hadm_id, icd_code) and sorted by date and sequence number in order to create a chronological order to the ICD-10 codes. This joined dataframe which originally contained 180, 640 patients was filtered for just patients with ICD-10 codes (ICD-9 codes are also present), reducing the sample to 80,213 patients. Next, a list of patients who died on their last visit was created, and the dataframe was filtered on this list to extract all of the records for these patients, reducing the population to 3,360 patients. Next, strings of ICD-10 codes were created for each subject starting from their first visit to their last visit. A new subpopulation was found for just those patients who experienced cardiac arrest in their final visit, classified under ICD-10 codes I462, I468, and I469, further reducing the population sample to 363 patients. A single list of two string elements was then created for each patient, where all of the codes until the second to last visit comprised the first element in the list, and all of the codes from the last visit comprised the second element in the list. Start of Sentence (SOS) and End of Sentence (EOS) tokens were appended to the start and end of the strings and the max sequence length was set to 250 characters.

['I63132 J189 J690 E43 N179 G92 E870 G8191 R1310 G20 R4701 I10 D638 F0280 E785 H9193 E860 Z6820 R918 R197 H409 Z781 R0902 J690 E43 G20 I69391 Z931 Z681 I10 E785 R1310', 'J189 J9691 E43 E870 G92 G20 R1310 I69351 I69391 E46 I469 I10 E785 J988']

seq2seq Methodology

The seq2seq model implemented in Pytorch comprised Encoder and Decoder Recurrent Neural Networks (RNNs), specifically GRUs, along with a Bahdanau Attention mechanism in the decoder layer which incorporates an additional context vector into the final output. An 80/20 training test split was created using sci-kit learn and dataloader libraries. Testing and training modules were created, a learning rate scheduler was added, and loss was used as an evaluation metric as well as

visual inspection of the output sequences of codes. The training parameters for Model 1 and Model 2 are listed in Table 4 below.

Table 4

Training Parameters for seq2seq Models

	seq2seq Model 1	seq2seq Model 2
Learning Rate	0.0001	0.001
Number of Training Epochs	20	20
Weight Decay	1e-6	1e-5
Optimizer	Adam	Adam
Step Size	20	30
Gamma	0.8	0.95
Dropout Rate	0.1	0.1

BERT Methodology

The BERT model implemented in Pytorch comprised Encoder-only architecture, multihead attention, positional embedding, normalization and a feed forward layer. An 80/20 training test split was created using sci-kit learn and dataloader libraries. Testing and training modules were created. Loss, perplexity, and accuracy scores were used as evaluation metrics and the losses were used as evaluation metrics. The training parameters for Model 1 and Model 2 are listed in Table 5.

Table 5

Training Parameters for BERT Models

	BERT Model 1	BERT Model 2
Dimensions	768	768
Number of Layers	12	12
Number of Attention Heads	12	24
Hidden Layers	768*4	768*4
Dropout Rate	0.1	0.2

Optimizer	Adam	Adam
Learning Rate	1e-4	1e-5
Weight Decay	0.01	0.001
Betas	(0.9, 0.999)	(0.9, 0.999)
Warm Up Steps	10,000	1,000

Results

The results of the models are presented in a structured manner. In the Seq-2-Seq model, each set of sequences consists of an input sequence (marked as '>'), the actual target sequence (marked as '='), and the generated output sequence (marked as '<'). This format allows for a clear understanding of the model's performance in sequence generation.

For the BERT model, the results show a pattern as the number of training epochs increases. As the epochs progress, the model's losses consistently decrease, which is a positive indicator of learning and optimization. Conversely, the accuracy of the model fluctuates periodically as the epochs increase, which can be visually observed through graphical representations. One notable trend in the BERT model's performance is the reduction in perplexity scores. Initially starting at 4, the perplexity score steadily decreases to 1.3. This decreasing perplexity indicates that the model is learning and performing effectively in generating sequences. This trend is a valuable insight into the model's progress and its ability to capture complex patterns in the data.

Seq-2-Seq Model-1

Reading lines...
Read 140 sentence pairs
Trimmed to 87 sentence pairs
Counting words...
Counted words:
prior 716
last 733
0m 3s (- 1m 12s) (1 5%) 6.8819
> K9189 N360 Y842 Y92230 Z21 Z8546 F17210
= A047 K55029 D65 J9601 R6521 K7200 A419 N179 E872 M79A3 D62 F05 N360 Y842 Z66 Z515 Z21 I468 I10 E875 E162 E8339 E8351 F17210 Z8546 Z933
< T50995A Z923 J9690 G936 A0471 B690 J984 I9751 B690 I9751 K560 Y92000 Z992 E1122 E1122 I272 A408 A4101 Z833 L89610 J939 E1169 Z85038 F05 D709 M4712 I428 A408

Z6825 I25810 Z8701 R402322 C715 Z8572 L89610 L89153 R0489 K810 R1310 A408 I96 H5462 T8119XA Z1611 N183 N183 I2699 R1310 J441 F79 Y92019 S2243XA Y92019 H409 H409 G936 A0471 B690 J984 T508X5A I272 I2510 L98429 K651 K651 K651 R9401 B258 I69122 Z95810 Z6825 E039 D696 Z8701 K56609 R402322 B690 J984 I9751 B690 I9751 K560 Y92000 J15211 L02213 Y92234 Y92234 J90 Z923 J9690 R64 C250 M79A3 M79A3 E8809 B690 J984 I9751 B690 J984 I9751 B690 I9751 K560 Y92000 Z992 E1122 E1122 I272 A408 A4101 Z833 L89610 J939 E1169 Z85038 F05 D709 M4712 I428 A408 Z6825 I25810 Z8701 R402322 C715 Z8572 L89610 L89153 R0489 K810 R1310 A408 I96 H5462 T8119XA Z1611 N183 N183 I2699 R1310 J441 F79 Y92019 S2243XA Y92019 H409 H409 G936 A0471 B690 J984 T508X5A I272 I2510 L98429 K651 K651 K651 R9401 B258 I69122 Z95810 Z6825 E039 D696 Z8701 K56609 R402322 B690 J984 I9751 B690 I9751 K560 Y92000 J15211 L02213 Y92234 Y92234 J90 Z923 J9690 R64 C250 M79A3 M79A3 E8809 B690 J984 I9751 B690 J984 I9751 B690 I9751 K560 Y92000 Z992 E1122 E1122 I272 A408 A4101 Z833 L89610 J939 E1169 Z85038 F05 D709 M4712 I428 A408 Z6825 I25810 Z8701 R402322 C715 Z8572 L89610 L89153 R0489 K810 R1310 A408 I96 H5462 T8119XA Z1611 N183 N183 I2699 R1310 J441 F79 Y92019 S2243XA Y92019 H409 H409 G936 A0471 B690 J984 T508X5A I272 I2510 L98429 K651

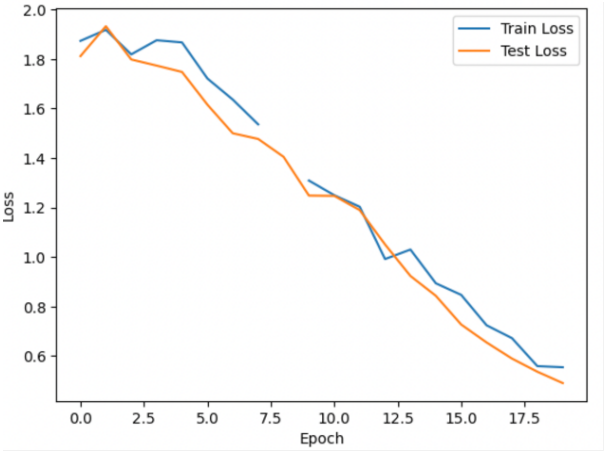
Seq-2-Seq Model-2

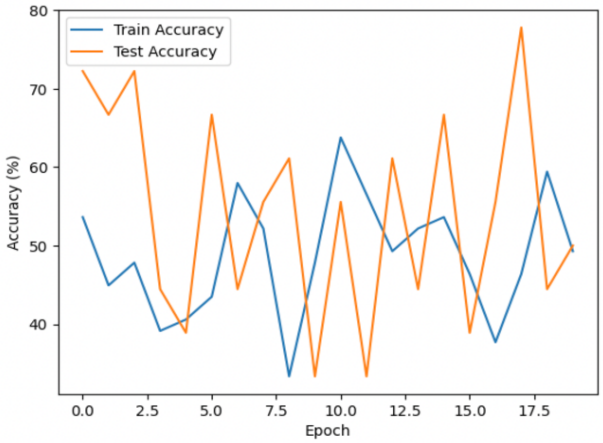
Reading lines...
Read 140 sentence pairs
Trimmed to 87 sentence pairs
Counting words...
Counted words:
prior 716
last 733
0m 4s (- 1m 31s) (1 5%) 6.8143
> G935 G950 I252 N400 V499XXS Z9181
= G935 J9600 G9589 D62 Z66 R1310 Q758 I469 D72829 N401 R338 Z87891
< I5023 Y838 Z923 I4581 R34 R471 J704

J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704
J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704
J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704
J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704 J704

BERT Model-1

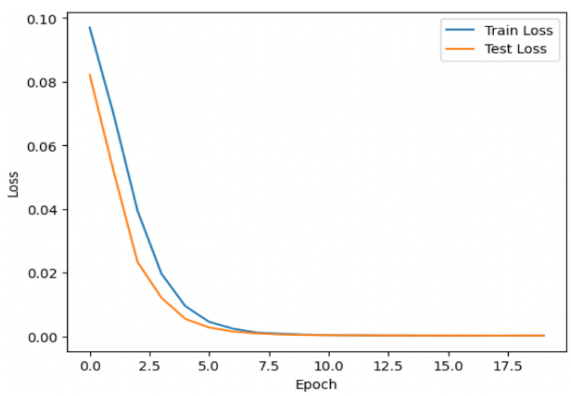
EP_train:0: 33%|| 1/3 [00:09<00:18, 9.46s/it]
{'epoch': 0, 'iter': 0, 'avg_loss': 1.8296449184417725, 'avg_acc': 56.25, 'loss': 1.8296449184417725}
EP_train:0: 100%|| 3/3 [00:18<00:00, 6.29s/it]
EP_train:1: 100%|| 3/3 [00:19<00:00, 6.47s/it]
EP1, train: avg_loss=1.91853133837382, total_acc=44.927536231884055
Precision: 0.5000
Recall: 0.6667
F1-Score: 0.5714
EP_train:19: 33%|| 1/3 [00:08<00:17, 8.61s/it]
{'epoch': 19, 'iter': 0, 'avg_loss': 0.5541626214981079, 'avg_acc': 56.25, 'loss': 0.5541626214981079}
EP_train:19: 100%|| 3/3 [00:18<00:00, 6.19s/it]
EP19, train: avg_loss=0.5544461806615194, total_acc=49.27536231884058
Precision: 0.6000
Recall: 1.0000
F1-Score: 0.7500
EP_test:19: 100%|| 1/1 [00:01<00:00, 1.60s/it]
{'epoch': 19, 'iter': 0, 'avg_loss': 0.49049076437950134, 'avg_acc': 50.0, 'loss': 0.49049076437950134}
EP19, test: avg_loss=0.49049076437950134, total_acc=50.0
Precision: 0.5000
Recall: 1.0000
F1-Score: 0.6667
EP19, test: perplexity=1.4541428089141846

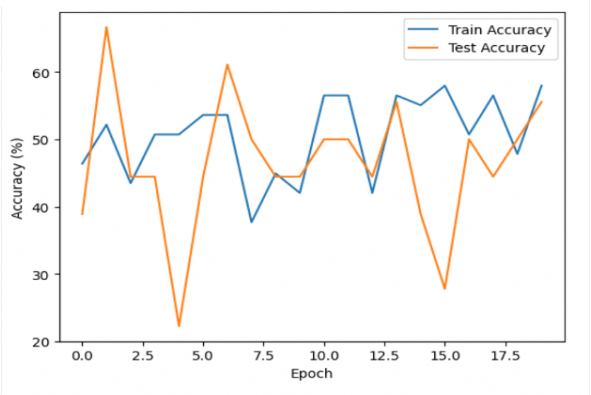




BERT Model-2

EP_train:0: 33%|| 1/3 [00:18<00:37, 18.58s/it]
{'epoch': 0, 'iter': 0, 'avg_loss': 0.10677383840084076, 'avg_acc': 56.25, 'loss': 0.10677383840084076}
EP_train:0: 100%|| 3/3 [00:40<00:00, 13.44s/it]
EP0, train: avg_loss=0.09691423674424489, total_acc=46.3768115942029
Precision: 0.6000
Recall: 1.0000
F1-Score: 0.7500
EP_test:0: 100%|| 1/1 [00:02<00:00, 2.96s/it]
{'epoch': 0, 'iter': 0, 'avg_loss': 0.08213409781455994, 'avg_acc': 38.88888888888889, 'loss': 0.08213409781455994}
EP0, test: avg_loss=0.08213409781455994, total_acc=38.888888888888886
Precision: 0.3889
Recall: 1.0000
F1-Score: 0.5600
EP19, test: avg_loss=0.00017938799283001572, total_acc=55.55555555555556
Precision: 0.5556
Recall: 1.0000
F1-Score: 0.7143
EP19, test: perplexity=1.0001143217086792





Conclusions

Parameter tuning of seq2seq model 2, especially increasing the step size and gamma of the learning rate scheduler and increasing the weight decay of the learning rate, resulted in moderately lower training and testing loss compared to seq2seq model 1. It predicted legitimate sequences of codes, but they were overall not the true sequences and further training is required. Parameter tuning of BERT model 2, especially the decrease in warm up steps, increase in attention heads, and increase in dropout rate, resulted in improvement over BERT model 1 with significantly lower loss and perplexity. However, it also had slightly lower F1 scores compared to the first model. Further training, again, is required.

Through our research, model development, model implementation, and model evaluation we have explored the capacity of seq2seq with GRU and Attention, alongside Encoder BERT models, to predict future ICD-10 diagnosis codes. By combining natural language processing techniques with traditional methods, we successfully generated sequences of ICD-10 codes as well as binary predictions as output from the models. While the respective loss values for both sets of models were minimized across progressive epochs, the other evaluation metrics, specifically accuracy, give us pause as to the efficacy of the implementation in a real-world business or medical context. For instance, while the decreasing perplexity scores for the BERT models during testing indicate the model is understanding the context of the medical codes, the varying levels of accuracy would not satisfy benchmarks or standards for healthcare providers. Therefore, this model shows the potential to achieve the stated business objectives of predicting sequences of ICD-10 codes and thus improving patient health outcomes and decreasing hospital costs, but it needs further development before it is ready for deployment.

Research Contributions and Novelty

The seq2seq model is the first generative AI model which starts with prior sequences of ICD-10 codes and generates the next sequences of predicted ICD-10 codes at a patient level. In addition, the transformer model is the only BERT model to our knowledge which classifies next ICD-19 codes, not just disease groupings like the SETOR model, therefore showing a greater level of

diagnosis specificity. Overall, these models treat diagnosis predictions from a novel framework of natural language processing (NLP) instead of more traditional methods of binary and multiclass classification.

Limitations and Future Research

One of the major limitations of the MIMIC-IV dataset is that it has intervisit (between hospital visit) order chronology but no intravisit (within hospital visit) order chronology. This is because for every hospital visit, we know when they are admitted, and when they are discharged, and there are no other timestamps in between. This means for example that a patient can be admitted to the hospital for two weeks, acquire 40 diagnoses and the codes will be input into the record for that time period with no discernible order. Most visits however or short, only last a day, but this still nonetheless inherently limits the accuracy of the models. In addition, the longitudinal nature of this data is limited and biased, since it only represents critical events of patients at a single hospital. The majority of patients have complex medical histories where they are visiting many different doctors and facilities for a long period of time on an outpatient and inpatient basis for chronic as well as acute conditions.

Future research should employ true claims and electronic health records (EHR) data which would have every code associated with a specific date and time and all the facilities a patient visited in order to hopefully construct a more accurate and complete sequence of diagnoses. In addition, a larger, more homogeneous sample with longer, full sequences, should be used, such as all outpatient and inpatient visits of MS patients in the two years before their MS diagnosis so that stronger patterns would be present in the data. Multimodal data could be used, such as also incorporating CPT codes. Finally, the temporal aspect of the data should be addressed - there are uneven intervals between visits, and it is important to be able to predict not just what will happen next but when it will happen.

References

Masud, J. H. B., Kuo, C.-C., Yeh, C.-Y., Yang, H.-C., & Lin, M.-C. (2023). Applying Deep Learning Model to Predict Diagnosis Code of Medical Records. Retrieved from https://www.mdpi.com/2075-4418/13/13/2297

Peng, X., Long, G., Shen, T., Wang, S., & Jiang, J. (2021). Sequential Diagnosis Prediction with Transformer and Ontological Representation. Retrieved from https://arxiv.org/abs/2109.03069

Tufail, A. B., Ma, Y.-K., Kaabar, M. K. A., Martínez, F., Junejo, A. R., Ullah, I., & Khan, R. (2021). Deep Learning in Cancer Diagnosis and Prognosis Prediction: A Minireview on Challenges, Recent Trends, and Future Directions. Retrieved from https://www.hindawi.com/journals/cmmm/2021/9025470/

Kohei Arai, Rahul Bhatia (2020). Lecture Notes in Networks and Systems. doi:10.1007/978-3-030-12385-7, Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-12385-7

Mantas, J. (n.d.). Public Health and Informatics: Proceedings of MIE 2021. Retrieved from https://books.google.com/books?hl-en&Ir=&id=81A2EAAAQBAJ&oi

Moons, E., Khanna, A., Akkasi, A., & Moens, M.-F. (2020). A Comparison of Deep Learning Methods for ICD Coding of Clinical Records. Retrieved from https://www.mdpi.com/2076-3417/10/15/5262

Dokumentov, A., Shaalan, Y., Khumrin, P., Khwanngern, K., Wisetborisut, A., Hatsadeang, T., Karaket, N., Achariyaviriya, W., Auephanwiriyakul, S., Theera-Umpon, N., & Siganakis, T. (2021). Automatic ICD-10 Coding Using Prescribed Drugs Data. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8580462

Aitziber Atutxa, Pakhomov, S. V., Pérez, A., Koopman, B., Zeng, M., Duarte, F., Lopez-Gazpio, I., Lang, D., Organization, W. H., Cho, K., Sutskever, I., Klein, G., Sennrich, R., Hieber, F., Pestian, J. P., Uzuner, ”O., Bossy, R., Névéol, A., Mullenbach, J. (2019, May 22). Interpretable deep learning to map diagnostic texts to ICD-10 codes. International Journal of Medical Informatics. https://www.sciencedirect.com/science/article/abs/pii/S1386505618310670?via%3Dihub

Johnson, A. *et al.* (2023) *Mimic-IV*, *MIMIC-IV v2.2*. Available at: https://physionet.org/content/mimiciv/2.2 (Accessed: 08 November 2023).