DrCoding: Using Transformers to Predict ICD-9 Codes from Discharge Summaries

Stanford CS224N Custom Project | Option 2

Boyang Tom Jin
Department of Computer Science
Stanford University
tomjin@stanford.edu

Abstract

Manual ICD code assignment underpins the way the health care system tracks diagnoses and procedures but continues to be error-prone and inconsistent between trained human coders. With the increasing adoption of electronic medical records, the interest in automating this process continues to grow. In this project, we train recent state-of-the-art Transformer-based models on clinical discharge summaries with the aim of improving the performance on the ICD-9 assignment task over existing LSTM-based methods. Fine-tuning BioBERT pre-trained context embeddings which resulted in the best overall performance with an F1 score of 51.0%, outperforming the LSTM baseline by 12.1%. Augmenting Transformer-based models with ICD code descriptions was also shown to slightly increase the F1 score by 1.2%.

1 Introduction

International Classification of Diseases (ICD) codes are standardized identifiers for diseases, injuries, clinical procedures, and other health conditions that are assigned to patients by healthcare professionals and trained hospital coders. Published by the World Health Organization, ICD codes are utilized by over 100 countries and are wildly used in research, clinical care, and billing [1]. Up until 2015, ICD-9 was the version used by health care systems in the United States, where the healthcare coding market is a billion-dollar industry [2] [3].

Because ICD code assignment is traditionally a manual process, it is time-consuming, taking on average between 25 to 43 minutes per patient to complete, depending on the ICD version [4]. Manual ICD assignment is also prone to errors that can result from discrepancies between coders, lack of coding experience, incorrect bundling of codes, or mistakes in the transcriptions. These mistakes are highly costly, with one report suggesting that errors in ICD coding has cost the Medicare system over $260 billion dollars over a span of six years [5]. Furthermore, because ICD codes are also used by policy-makers, such as the Veterans Health Administration, to determine resource allocation, the accuracy of ICD code assignments is becoming increasingly scrutinized [6].

Automated ICD coding addresses these concerns and ensures that a consistent, standardized approach can be applied. There have been several attempts at automation which have paralleled developments in machine learning. Recent works have also tried to incorporate deep neural networks, such as Li et al. who applied a convolutional neural network on discharge summary free-text and found that it improved upon SVM approaches by at least 14% [7].

Recurrent neural network models are also popular, with several approaches using a variation on the gated recurrent unit methods [8] [9]. Although most works claim promising results, it is difficult to perform an accurate comparison as each work was evaluated using different metrics and performed

on different datasets. Datasets are also often specific to a particular hospital or institution and subject to locality-specific biases or procedures. Even studies that utilize the same datasets may use different filtering and ICD code extraction procedures.

Despite these past investigations, the ICD code prediction from clinical text remains a challenging task and an active area of research. In particular, clinical text is often unstructured and can vary between patient visits, depending on the author of the text. Moreover, medical text is often long and contains a significant amount of misspellings and field-specific terminology and abbreviations which hinders traditional bag-of-words models or vocabulary representations through word embeddings [8] [7].

While the increasing adoption of electronic medical records (EMR) provides a useful trove of patient information which can form the basis of research such as automated ICD prediction, patient data remains inherently restricted by privacy regulations. Moreover, the distribution of ICD codes is inherently skewed, with rare diagnoses having only a handful of samples at best. This ultimately reduces the amount of textual data that can be acquired for training models. This is especially problematic for training robust deep supervised learning models, where it is ideal to have more than a few thousand examples for each class [10].

Here, we apply different attention-based Transformer models on discharge summaries to predict ICD-9 diagnosis codes. For comparison purposes, we use a LSTM-baseline model with GLOVE-embeddings on the discharge summaries. Transformer models are based on the concept of scaled dot-product self-attention originally described in Vaswani et al. paper [11]. However, one limitation of Transformers is the high number of parameters used by the model, which grows quadratically with respect and the length of the input sequence and therefore restricts the length of the discharge summary that can be used as input. By truncating the discharge summary, we are potentially removing useful information that could be useful in the classification problem.

In order to better utilize the input, we turn to using the Reformer model, which utilizes a locality-sensitive hashing attention scheme and reversible layers [12]. The Reformer model allows us to incorporate the full length of the input sequence while improving the training performance over a vanilla Transformer. However, the Reformer model does not utilize any pre-trained contextual embeddings, which limits their efficacy in the face of limited data. To rectify this, we fine-tune a BERT-based model after loading pre-trained weights from BioBERT, a biomedical language model [13].

Having been trained on biomedical text corpora, BioBERT enables us to take advantage of medical domain-specific terms whose meanings may not be accurately captured by the original BERT model. We also attempt to incorporate metadata from ICD code descriptions to try to improve the performance of these Transformer models.

Of the models assessed, the best performance was achieved by a fine-tuned BioBERT model with an F1 score of 51.0%. A Reformer model trained from scratch came slightly behind the BioBERT models with an F1 score of 49.0%. The metadata augmentation scheme improved the Reformer model by 1.2%.

2 Related Work

Early approaches to the ICD prediction problem applied K-nearest neighbors and Bayesian classifiers on discharge summaries, after representing the summaries with hand-crafted features [14]. While these approaches utilized the full discharge summary, they were still relatively simple in nature. Later approaches included using hierarchical classifiers, which reported the best F-measure of 39.5% based on predictions over 7042 ICD codes [15], and rule-based systems that tried to align the ICD code descriptions on top of discharge summaries [16]. This latter model reported an F1 score of 82.0% over 45 ICD codes, but the study utilized only 978 radiology reports as input which have more formal structure than discharge summaries. Nevertheless, these rule-based systems are intuitive approaches to the problem as they are most similar to the process that a human would take. As such, this work serves as inspiration to the metadata-augmented models utilized in this study.

With the rise in the popularity of deep learning, several studies have attempted to apply different variations of neural networks on the ICD coding task. One popular approach has been to use recurrent neural networks, driven by idea that there is rich contextual and temporal information contained

within the free-text of clinical notes. Both long-short term memory (LSTM) and gated recurrent networks (GRUs) seem to be a popular choice to overcome the long-distance relationships within clinical text, which commonly averages over a thousand words in length. Performance with these models has varied, with studies reporting F1 scores between 42.0% and 70.8% on predictions over 10 and 19 ICD codes respectively [17] [18]. One interesting RNN variant is one study that utilized hierarchical attention as part of a GRU-architecture. While the results are not significantly better than an SVM baseline, the authors suggested that this model is invaluable because the attention-weights provide valuable insights into which parts of the clinical note that was feeding into the ICD prediction [8]. These insights would be valuable when trying to achieve adoption of a Transformer-based automated ICD coding system by the medical community.

Transformers are still a relatively recent concept and as such, few studies have applied Transformer-based models to the ICD-prediction problem. One such study that did used a transfer-learning approach to fine BERT and BioBERT pre-trained contextual embeddings and achieved the best F1 score of 82.9% based on translated German animal experiment summaries. The authors found minor differences between using BERT and BioBERT [19]. One healthcare-associated Transformer model was introduced by Shang et al, who used graph-based neural networks to produce ontology embeddings that could feed into a fine-tuned BERT model to produce state-of-the-art performance on medication recommendations [20]. While the ICD prediction problem has been extensively studied, these latest findings motivate us to apply Transformers to the discharge summaries to try to improve upon past models.

3 Approach

3.1 Baseline

The lack of a publicly available evaluation dataset for discharge notes makes it difficult to perform comparisons with past work. As such, for our baseline, we created a bi-directional LSTM model that was inspired by previous deep learning models built for ICD code predictions. A dropout and a fully-connected layer were applied on the conjoined outputs of the bi-directional LSTM. The baseline model was applied on the same processed dataset that was used for the Transformer models to ensure a common starting point.

3.2 Reformer Transformer

The Transformer model is based on the concept of the scaled dot-product self-attention, whereby an input sequence is attended to itself in order to discover dependencies within itself. This can be represented by $softmax(\frac{QK^T}{\sqrt{d_k}}V)$, where Q, K, and V in this case are matrices derived from a discharge summary [11]. Note that the scaling factor here is used to prevent vanishing gradients when a large dimensionality is used.

Transformers traditionally consists of several encoder and decoder layers, where each layer consists of a multi-head self-attention layer and a fully-connected feed-forward network. In addition, there is a residual connection around each sublayer followed by layer normalization. In our classification task, we forego the decoder layers and instead make use of a series of stacked encoder layers followed by a non-linearity, fully-connected layer, and a softmax to produce an output probability for each ICD code. The Reformer Transformer model suggests three areas of improvements: use of reversible layers, use of locality-sensitive hashing for self-attention, and splitting activations in the feed-forward layers [12].

For this model, we will use the pytorch-reformer package from Github [21]. However, because this package only contains the base language model, we built a separate text classification model on the base Reformer module by applying a non-linearity, fully-connected layer, and softmax after the stacks of self-attention layers. Our Reformer model allows us to scale better than an equivalent vanilla Transformer models and intake longer discharge summaries.

3.3 BioBERT

This model fine-tunes pre-trained BioBERT context embeddings using the discharge summary data on the ICD prediction task. Like the original BERT model, BioBERT inserts a [CLS] token at

the beginning of each training sample [13] [22]. Classification tasks are then done by passing the representation of the [CLS] token through an output layer. Furthermore, one of the innovations of a BERT language model is the use of a word tokenizer, which allows out-of-vocabulary words to be represented by word pieces rather than just being assigned to a default unknown token. This feature is especially useful in this project because misspellings and non-standard abbreviations frequently occur in the discharge summaries.

We expect this model to have the best performance of all models as we do not have enough data to properly train our own Transformer model. For our project, we will fine-tune a 'HuggingFace' BERT transformer but adapt and load in the pre-trained BioBERT embeddings instead of the default BERT embeddings [23] [13].

Metadata

3.3.1 Metadata Attention

In this model, we replace self-attention with attention between the discharge summary and 32-length ICD-9 code descriptions in a sliding window approach. Specifically, the Q and K in $softmax(\frac{QK^T}{\sqrt{d_k}}V)$ will represent the discharge summary and ICD discharge summary respectively. This model is driven by the intuition that each label should naturally attend most closely with parts of the input sequence that the label represents. In this case, we forked both the 'pytorch-reformer' Reformer implementation and the 'HuggingFace' Transformer model and replaced the existing attention schemes with the metadata attention scheme to compare the performance.

3.3.2 Metadata Augmentation

Here, we extracted 92 keywords from the ICD code descriptions. For each discharge summary, we aligned the keywords over the summary to produce a one-hot vector of keyword hits. This vector is joined with the output of the Transformer model and passed through a fully-connected layer to produce the class predictions. This model can be considered a hybrid between the rule-based model favored by past ICD-prediction models and the Transformer model.

This model is loosely based off of past augmented Transformer models such as the model proposed by Shang et al. to perform medication recommendations which uses the embedding produced by an ontology tree as input to BERT [20].

4 Experiments

4.1 Data

We used the MIMIC-III dataset, a restricted-access electronic medical record dataset of over 40,000 patients who stayed in the Beth Israel Deaconess Medical Center between 2001 and 2012 [24]. For this study, we specifically extracted the discharge summaries, which averaged 1400 words in length. The corresponding ICD-9 diagnosis codes and code descriptions were also extracted and joined with the text into a primary dataset. The data was further filtered to include only the top 50 base ICD-9 codes which resulted in 51,295 total samples. This data was split 0.64/0.16/0.20 into training, validation, and test sets.

An additional smaller secondary dataset of 44,450 samples was produced by further extracting only the hospital course section, which is a more concise summary of the patient visit averaging 460 words. This dataset omits information such as medication lists, raw diagnostic test results, and past medical history which either contain an abundance of technical abbreviations, numeric measurements, or background information not relevant to the current patient visit.

Because the discharge summaries were in an HIPAA-compliant form-like format, additional pre-processing steps were taken to reshape into a sequence of sentences, remove all numbers and name-placeholders, and convert to lowercase. Each discharge summary was also padded or truncated according to a tunable target length parameter before being tokenized according to a vocabulary built from the training data.

4.2 Evaluation method

Binary cross-entropy loss, averaged across samples, was used during the training process. For evaluation, the micro-F1 score based on the predicted versus actual ICD-9 codes was the primary metric used to assess the performance of all models:

$$F1 = \frac{1}{n}\sum_{i-1}^{n}\frac{2 * (|P_i \cap A_i|)}{|P_i| + |A_i|}$$

where A_i denotes the actual ICD codes and P_i denotes the predicted ICD codes

4.3 Experimental details

All experiments were run on a NV6-type Microsoft Azure machine which had 80GB of RAM, 6 vCPUs, and a NVIDIA M60 GPU.

Tokenized discharge summaries averaging 1400 words were used as input, although the exact length of the discharge summary was tuned as a hyperparameter. Whenever a discharge summary needed to be truncated, the discharge summaries were truncated from the tail as this was found to have better performance. Note that for all non-BERT models, we use a word-based tokenizer based on a vocabulary of 50,000 words generated from the discharge notes corpus.

4.3.1 Baseline

We used a bidirectional LSTM encoder containing a single forward and a single backward layer with 256 hidden units. The baseline also uses a 100-dimensional GLOVE embedding applied on input sequence batch sizes of 32. Hyperparameter tuning was primarily applied on the learning rate, where we found that using a rate of 0.001 works well to train the baseline eight epochs.

4.3.2 Reformer

For performance considerations, we used four multi-attention heads and six Transformer encoding layers, with a dropout of 0.1 applied in each encoding layer. In the Reformer model, we also used four rounds of LSH hashing as this was found to have 99.9% accuracy relative to full attention in the Reformer paper. These models were trained over a period of 12 epochs.

4.3.3 BioBERT

We loaded the BioBERT-Base v1.1 pre-trained weights into a standard 12-layer BERT model with 12 attention heads in each layer. The size of the BERT model restricted us to using maximum batch sizes of 16 for 256-length input sequences and 2 for 512-length input sequences during the fine-tuning process. Batches beyond this length resulted in out-of-memory issues. A smaller learning rate of 0.00003 was found to work better, which was in line with the original author recommendations [25].

4.4 Results

Table 1 and Figure 1 show that the overall best performance was achieved by the BERT model with an F1 score of 51.0% on length 256 sequences, which was 2% higher than the top-performing Reformer model at 49.0%.

As shown in Figure 3, longer input sequences performed better in the Reformer model, with a 6.5% and 3.6% improvement in F1 score when using a 2048-length sequence versus a 512-length and 256-length sequence respectively. This interestingly did not come at a cost in speed however, as the 2048-length sequence processed 3.5% more words/sec than the corresponding 256-length sequence (27,689 words/sec vs 26,713 words/sec), although twice as many iterations were required to achieve 12 epochs due to the proportionally smaller batch size.

Table 1: Top F1 scores on primary dev dataset

Model	Input Length	F1	Words/Sec
LSTM with GLOVE	1024	0.389	**300,041**
BioBERT	256	**0.510**	2265
Reformer	2048	0.490	27,689
Reformer	512	0.454	27,420
Reformer	256	0.425	26,713

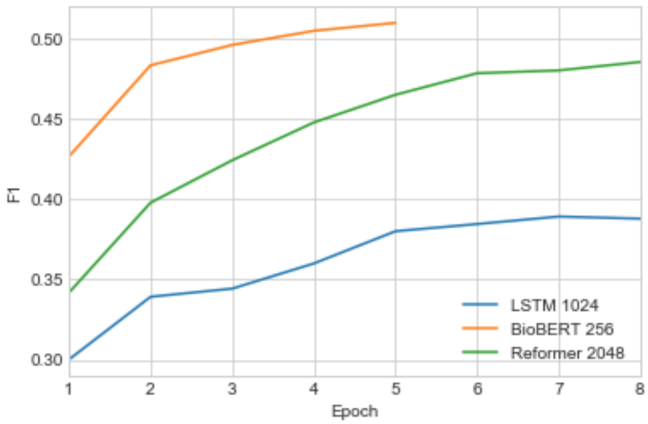

Figure 1: Performance of models on the dev dataset over varying training epochs

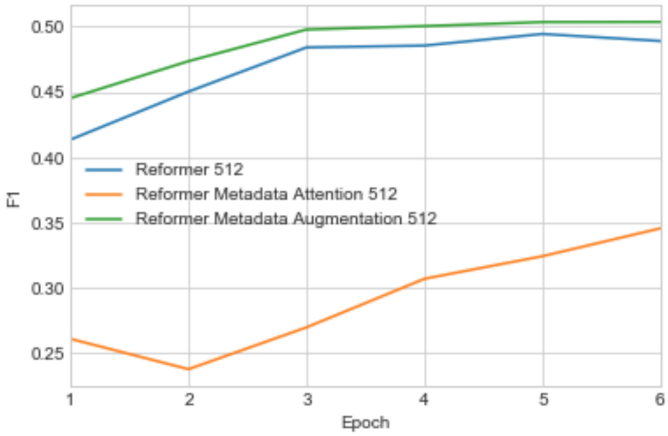

Figure 2: Performance of the Reformer using different metadata incorporate trials. Metadata augmentation refers to joining a one-hot vector of ICD description keywords with the Reformer output. Metadata attention refers to attending the ICD code descriptions with the input text.

Table 2: Model Size Comparisons

Model	Model Size	Number of Parameters
LSTM Baseline	24 MB	5.8M
Reformer	59 MB	22.8M
BioBERT	414 MB	113M

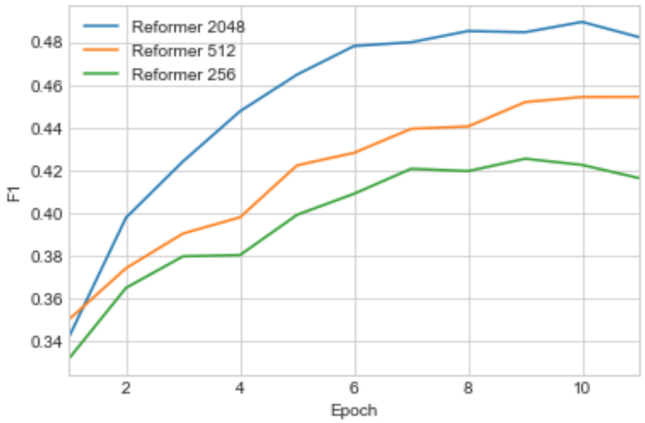

Figure 3: Performance of the Reformer over increasing input sequence lengths

5 Analysis

5.1 Performance

Interestingly, the BioBERT model exceeded the performance of all other models on the primary dataset, despite using shorter input sequences. The Reformer model using 2048 words came close, but was perhaps handicapped by the limited number of training examples used to train the model. Nevertheless, this study shows that the Reformer model is an ideal substitute for task-specific training, since the time and resources it takes to train a model is ten times lower than it is to fully train a BERT model from scratch. That said, if pre-existing contextual embeddings exist, they should still be used as they provide a substantial bump in performance.

The higher F1 score achieved by augmenting the Reformer model with keywords extracted from the ICD code descriptions in Figure 2 indicates the benefit of incorporating context-rich metadata. This technique parallels how a human might look at the dataset, providing the model with some intuition on what to look out for.

5.2 Attention

Attention was shown to be useful within the context of this problem as it is able to identify relationships between key words and concepts. For instance, an analysis on the attention relationships between two sequences on the fine-tuned BioBERT model showed there is a degree of attention between the words “infection” and “pneumonia” (Figure 4), which can help lead to identify specific pneumonia or bacterial-associated ICD codes.

However, replacing the self-attention scheme to attend between the ICD code descriptions and the text was not successful as the model would immediately overfit to the descriptions, regardless of dropout that was used. Although the intention was to focus on relevant portions of the text using the ICD code descriptions, the ICD descriptions were in general quite short, consisting of only a few words (eg. “Acute kidney failure, unspecified”). This caused problems in creating the sliding window, as even a window as small as 32 words would contain a significant amount of padding characters.

The description keywords not surprisingly improved performance as it provides the model with hints on which ICD code to assign. However, the fact that only a small improvement of 1.2% was observed meant that the model was already able to pick up most of this knowledge on its own.

5.3 Common Sources of Errors

5.3.1 Lack of Information

An ICD code was assigned to a patient but the discharge summary did not contain enough information to ascertain the ICD code. This occurs either because the relevant information was removed when the

summary was truncated to the input length or because the information was contained in other parts of the patient's electronic medical records, such as their lab results or past patient history.

Example: The discharge note of this patient was focused on the acute respiratory failure, whereas hypertension and osteoporosis were pre-existing conditions not mentioned in the input to the model.
Actual Diagnosis: ['Unspecified essential hypertension ', 'Urinary tract infection, site not specified ', 'Acute respiratory failure ', 'Osteoporosis, unspecified ']
Predicted Diagnosis: ['Congestive heart failure, unspecified ', 'Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled ', 'Other and unspecified hyperlipidemia ', 'Unspecified essential hypertension ', 'Acute respiratory failure ']

5.3.2 Closely Related ICD Codes

The model was not able to differentiate between similar ICD codes.

Example: The model correctly predicted an occurrence of pneumonia within the discharge summary, but the specific type of pneumonia was incorrect.
Predicted: ['Unspecified essential hypertension ', 'Esophageal reflux ', 'Pneumonia, organism unspecified ', 'Tobacco use disorder ', 'Acute respiratory failure ']
Actual: ['Tobacco use disorder ', 'Pneumonitis due to inhalation of food or vomitus ']

5.3.3 Similar Prefixes

The BioBERT model occasionally sometimes struggle with deciphering similar terminology which could be a result of the high number of out-of-vocabulary words. On average, approximately, 25% of the BioBERT input sequence was made of word-pieces. Combined with the fact that misspellings frequently occur, this means that some key technical words may not be represented correctly, especially given the relatively few number of examples to learn from.

Example: Here, the model predicted several words with the suffix "hyper", which was also a word-piece Although BERT differentiate ill-defined words based on the surrounding sentence context, even looking at sentence context may not help the model in this case, as most of the terms below are associated with blood (eg. hypertension = 'high blood pressure', hyperlipidemia = 'high cholesterol in the blood') [26].
Predicted: ['Anemia, unspecified ', 'Other and unspecified hyperlipidemia ', 'Unspecified essential hypertension ', 'Hyposmolality and/or hyponatremia ', 'Unspecified acquired hypothyroidism ']
Actual: ['Hyposmolality and/or hyponatremia ']

5.4 Input Length Effect

It is unsurprising that increasing the input sequence length led to better results (Figure 3), given that the discharge summaries are often written in a methodical manner such that one event of a patient's visit is described in detail (eg. medication, specific tests done) before moving onto another event. This means that pertinent information related to diagnosis is often spread across the full length of the discharge summary. This may explain why the BERT performance on a smaller discharge summary dataset comprised only of the hospital course information was actually worse, as this smaller dataset contained key pieces of information.

6 Conclusion

Transformer models with pre-trained contextual embeddings have previously been shown to improve performance on a variety of natural language processing tasks and the ICD-code prediction problem is no exception. In this study, we showed that the a fine-tuned BioBERT Transformer-based model outperforms an LSTM-baseline by 12.1% with respect to the F1 score. Reformer models were

also demonstrated to be highly scalable, training on input sequences over 1000 words with little performance overhead, something that is unobtainable with BERT models. Future work should look at expanding the discharge summary dataset to examine whether additional samples will improve the performance of a Transformer model trained from scratch. There should also be efforts to create an authoritative discharge summary dataset post-cleaning and standardization, similar to SQuAD. Such a dataset would help the biomedical field in comparing models on medical-specific tasks outside of just ICD-code prediction such as length-of-stay prediction, medication prediction, and patient-status question-answering.

References

[1] World Health Organization et al. International classification of diseases (icd) information sheet. *found at http://www.who.int/classifications/icd/factsheet/en*, 2014.

[2] Jessica Germaine Shull. Digital health and the state of interoperable electronic health records. *JMIR medical informatics*, 7(4):e12712, 2019.

[3] Grand View Research. U.s. medical coding market worth $7.0 billion by 2025.\ 2019.$

[4] Mary H Stanfill, Kang Lin Hsieh, Kathleen Beal, and Susan H Fenton. Preparing for icd-10-cm/pcs implementation: Impact on productivity and quality. *Perspectives in health information management*, 11(Summer), 2014.

[5] Council for Medicare Integrity. Error rate drops, but medicare still lost $31.6 billion to preventable billing errors in fy 2018.\ 2018.$

[6] Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.

[7] Min Li, Zhihui Fei, Min Zeng, Fang-Xiang Wu, Yaohang Li, Yi Pan, and Jianxin Wang. Automated icd-9 coding via a deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(4):1193–1202, 2018.

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

[9] Ssu-ming Wang, Yu-hsuan Chang, Lu-cheng Kuo, Feipei Lai, Yun-nung Chen, Fei-yun Yu, Chih-wei Chen, Zong-wei Li, and Yufang Chung. Using deep learning for automatic icd-10 classification from free-text data. *European Journal of Biomedical Informatics*, 16(1), 2020.

[10] Dan C Cireşan, Ueli Meier, and Jürgen Schmidhuber. Transfer learning for latin and chinese characters with deep neural networks. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2012.

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

[12] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv*, 2020.

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

[14] Leah S Larkey and W Bruce Croft. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297, 1996.

[15] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237, 2014.

[16] Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Talukdar, and Steven Carroll. Automatic code assignment to medical text. In *Biological, translational, and clinical language processing*, pages 129–136, 2007.

[17] Priyanka Nigam. Applying deep learning to icd-9 multi-label classification from medical records. Technical report, Technical report, Stanford University, 2016.

[18] Sandeep Ayyar, OB Don, and W Iv. Tagging patient notes with icd-9 codes. In *Proceedings of the 29th Conference on Neural Information Processing Systems*, pages 1–8, 2016.

[19] Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert. *CLEF (Working Notes)*, 2019.

[20] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv*, 2019.

[21] Phil Wang. Reformer-pytorch. `https://github.com/lucidrains/reformer-pytorch`, 2020.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018.

[23] HuggingFace. Transformers. `https://github.com/huggingface/transformers`, 2020.

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

[25] Google Research. bert. `https://github.com/google-research/bert/blob/master/README.md`, 2020.

[26] Sharad Jones, Carly Fox, Sandra Gillam, and Ronald B Gillam. An exploration of automated narrative analysis via machine learning. *PloS one*, 14(10), 2019.

[27] Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv*, 2019.

A Appendix (optional)

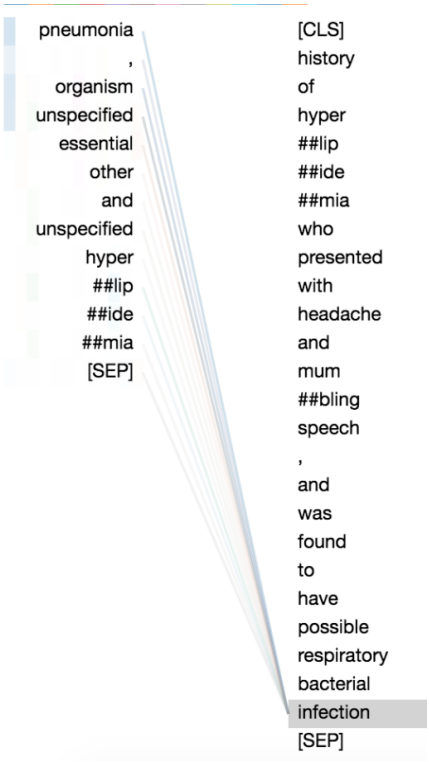

Figure 4: Sample attention relationships for one attention head between two representative discharge summary sentences as visualized using BertViz [27]