Rare Codes Count: Mining Inter-code Relations for Long-tail Clinical Text Classification

Jiamin Chen[1,2], **Xuhong Li**[2], **Junting Xi**[1], **Lei Yu**[1], **Haoyi Xiong**[2]
[1]Beihang University, Beijing, China
[2]Baidu Inc., Beijing, China
{lixuhong, xionghaoyi}@baidu.com, {jiaminchen, drinking, yulei}@buaa.edu.cn

Abstract

Multi-label clinical text classification, such as automatic ICD coding, has always been a challenging subject in Natural Language Processing, due to its long, domain-specific documents and long-tail distribution over a large label set. Existing methods adopt different model architectures to encode the clinical notes. Whereas without digging out the useful connections between labels, the model presents a huge gap in predicting performances between rare and frequent codes. In this work, we propose a novel method for further mining the helpful relations between different codes via a relation-enhanced code encoder to improve the rare code performance. Starting from the simple code descriptions, the model reaches comparable, even better performances than models with heavy external knowledge. Our proposed method is evaluated on MIMIC-III, a common dataset in the medical domain. It outperforms the previous state-of-art models on both overall metrics and rare code performances. Moreover, the interpretation results further prove the effectiveness of our methods. Our code is publicly available[1].

1 Introduction

The International Classification of Diseases (ICD) is a worldwide diagnostic tool published and maintained by the World Health Organization (WHO). The ICD coding, a task of assigning ICD codes according to the electronic medical records (EMRs), facilitates a lot of activities in health care, such as morbidity and mortality statistical analysis, medical billing and decision support systems (W. et al., 2020; Sutton et al., 2020). Since the traditional manual EMRs coding is time-consuming and prone to error (O'malley et al., 2005), its automation has always been attracting attention since 1990s (de Lima et al., 1998). Most of the existing methods treat the automatic ICD coding as a supervised multi-label document classification task (Xie and Xing, 2018; Mullenbach et al., 2018). By learning the text representations with an RNN (Vu et al., 2020), CNN (Mullenbach et al., 2018; Liu et al., 2021) or Transformer (Biswas et al., 2021) based encoder, the model extracts the code-relevant features via a trainable query matrix and predicts the codes with multiple binary classifiers.

Rare Code Prediction. Although the introduction of deep learning methods significantly improves the overall metrics for ICD coding, the extremely long-tail distribution over labels still makes the prediction for rare diseases or procedures challenging. Taking the MIMIC-III Dataset as an example, among all the discharge summaries, the most frequent code appears 20,053 times while the codes which occur less than 100 times constitute 12% of the whole dataset. In the supervised methods, learning the distinguished representation for each code through training samples requires rich data resources, leading to better performances on frequent codes than less frequent ones. Collecting sufficient documents for rare codes can be very difficult and expensive, which makes rare code prediction a critical task in automatic ICD coding.

Regarding this subject, several research directions have been explored. For example, some unsupervised methods have been proposed (W. et al., 2020; Song et al., 2020), but there remain clear margins compared to the supervised ones. Most of the previous works with supervised methods focus on the top 50 most frequent codes and extend the model usage on infrequent codes. But their results on rare codes are far from satisfactory. A few studies concerning the few-shot literature (Wang et al., 2021; Yuan et al., 2022) are proposed to improve the rare code performance by enriching the code descriptions via external knowledge sources. However, accessing heavy external sources can be

[1]https://github.com/jiaminchen-1031/Rare-ICD

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 403–413

complicated, and it is possible to introduce unexpected bias and noise facing immense knowledge.

In this work, we propose a more efficient method by strengthening the inter-code relations to improve the rare code performances. The existing supervised learning methods with label-wise attention (Vu et al., 2020; Mullenbach et al., 2018; Liu et al., 2021; Biswas et al., 2021) can hardly capture the helpful inter-code relations for rare codes. And we consider it as the reason for their bad performance on rare codes. We show in Figure 1 the correlations between code representations in the traditional label-wise attention method and our method. As suggested in the left figure, for frequent codes, the model can well learn their strong or weak correlations with the other codes via sufficient training samples. On the contrary, for rare codes, the model fails in building the useful connections and presents irrelevance with most of the codes. By enhancing the relations, as shown in the right figure for our method, the model performance on rare codes can be effectively improved.

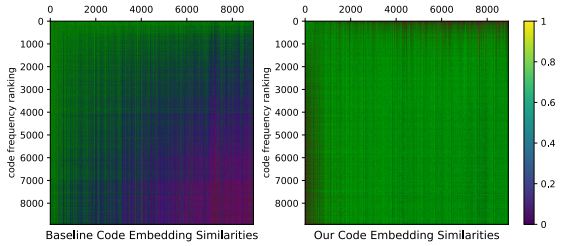

Figure 1: Code correlations by their embedding similarities in the traditional method (Vu et al., 2020) (*left*) and ours (*right*). The axis is arranged according to label frequency, where 0 indicates the most frequent and a greater value means less frequent.

Inter-code Relations. As indicated in Figure 2, we present the inter-code relations in this work by co-occurrence and hierarchy. Generally, code co-occurrence is acquired by counting the co-appearing times of two diseases in the same clinical text from a group of data. Revealing this information explicitly is helpful for the model to incorporate the relations between different codes. However, for rare codes, due to the lack of related samples, their co-occurrence relations with other codes can be incomplete or biased. To alleviate this issue, we propose to introduce a parent-child structure for each code, thus being able to explore the co-occurrence under different levels and merge them to complement the co-occurrence for rare codes. Here, we extend the definition of being hierarchical, follow the ICD-9 Official Guidelines released by the U.S. Federal Government's Department of Health and Human Services, and propose a **code-organ-system** hierarchical structure: level-0 (code itself), level-1 (codes of the similar organs) and level-2 (codes of the same system).

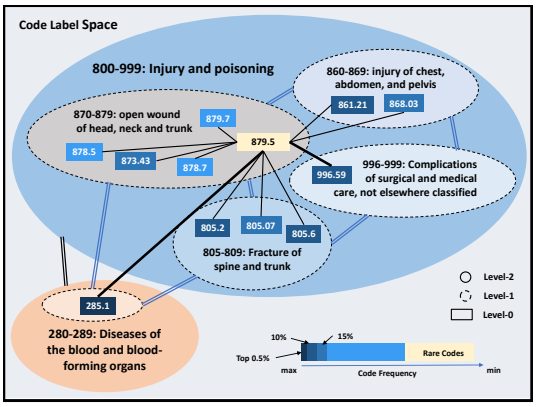

Figure 2: The inter-code relations exploited in this work. Taking the rare code 879.5 as an example, we build its relations with other codes by co-occurrence and enhance them with the code-organ-system hierarchy and the connections between different categories.

Our Contributions. In this work, we propose a novel method to improve rare code prediction by enhancing the connections between frequent and rare codes. The inter-code relations are explored via code descriptions, the code-organ-system hierarchy, and co-occurrence, which can be easily accessed without the necessity of bringing heavy external knowledge. Although quite a few studies have concentrated on inter-code relations (Tsai et al., 2021; Yan et al., 2010; Cao et al., 2020a), to our knowledge, we are the first to bond the inter-code relations specifically with rare code prediction and propose to exploit the inter-code relations under the code-organ-system hierarchy to tighten the weak connections for rare codes. We evaluate our method on the MIMIC-III-full dataset by their metrics on all codes and rare codes, where it outperforms the previous state-of-art models in automatic ICD coding.

2 Related Works

Automatic ICD Coding. Medical text categorization has been an important task in medical NLP for a quite long time. Early works adopt traditional machine learning methods for coding (Larkey and Croft, 1996; Pestian et al., 2007; Perotte et al., 2014). With the rising of neural networks, the automatic ICD coding began to be considered as

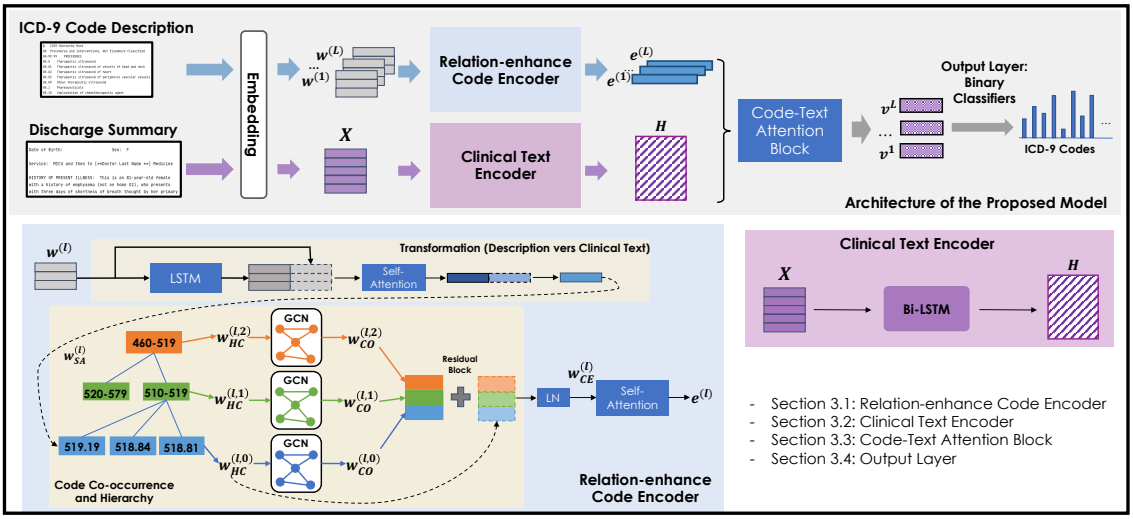

Figure 3: The architecture of our model.

a multi-label classification task. Mullenbach et al. (2018) propose a convolutional neural network text encoder and an attention layer to capture the important features of each code. Vu et al. (2020) further develop the label-wise attention layer with randomly initialized label representations and propose to use LSTM as text encoder. Various CNN (Li and Yu, 2020; Liu et al., 2021), RNN (Vu et al., 2020) and Transformer (Biswas et al., 2021) variants have also been used to encode the clinical documents. Some works (Xie et al., 2019; Cao et al., 2020a) propose to use GCN to integrate the code hierarchy and co-occurrence into the representation learning. Yuan et al. (2022) propose to enrich the code semantic information by introducing its synonyms from the **U**nited **M**edical **L**anguage **S**ystem (UMLS). Concerning the benchmark for evaluation, although some ICD-10 datasets have been collected and used in previous works (Cao et al., 2020b; Koopman et al., 2015), MIMIC-III for ICD-9 codes is still the only available dataset for clinical documents up till now.

Few-shot Learning for Long-tail ICD Codes. Few-shot learning targets at achieving good performances to the classes where a few samples are available (Medina et al., 2020). Due to the long tail of medical document dataset, some methods concerning ICD coding have also been proposed with similar strategies of few-shot learning. Current strategies can be divided into two types. The first works on improving the training process to achieve a better performance, such as proposing a novel optimization mechanism (Li et al., 2017) and modifying the loss function (Lin et al., 2017). For ICD coding, some works introduce different weights to the loss terms to help rare code prediction, such as Focal Loss (Lin et al., 2017) in Effective-CAN (Liu et al., 2021) and label-distribution-aware margin (Cao et al., 2019) in TransICD (Biswas et al., 2021).

The other type aims at learning a similarity function between frequent and few-shot labels (Vinyals et al., 2016). Matching networks (Geng et al., 2020) give predictions by searching the few-shot labeled support set through cosine similarities. In ICD coding, this strategy is usually achieved by introducing external knowledge to obtain the similarities. Vu et al. (2020) use the code formulation rules to apply a hierarchical joint learning mechanism. Some works bring in code relations from reliable sources, such as Wikipedia (Wang et al., 2021), code synonyms from UMLS (Yuan et al., 2022) and knowledge graphs (Xie et al., 2019).

3 Approach

In this section, we introduce the whole architecture of our model, which is illustrated in Figure 3. First, both the medical records and code descriptions are tokenized and embedded via a shared Word2Vec (Mikolov et al., 2013). Then we adopt a dual encoder architecture to encode code descriptions and medical records respectively. The embedding of code descriptions is put into a **Relation-enhanced Code Encoder** (Section 3.1) to strengthen the connections between codes, especially between the rare codes and frequent codes. We exploit the co-occurrence and the hierarchical structure of ICD codes via a series of modules in-

side the proposed code encoder. In parallel, the embedding of clinical texts is fed into a **Clinical Text Encoder** (Section 3.2) for contextualization. The outputs of these two encoders interact in the **Code-Text Attention Block** (Section 3.3), where the important words are highlighted and combined to generate a new code-specific vector with the representation of each code served as the query. The combination of the weighted words for each code is finally fed into the corresponding binary classifier in the **Output Layer** (Section 3.4) to calculate the probabilities.

3.1 Relation-enhanced Code Encoder

This encoder aims to identify the useful relations between different codes and enhance them via the built representations. We obtain the code descriptions from the World Health Organization (2016). Through the pretrained Word2Vec, the description for the code l is transformed into $\mathbf{w}^{(l)} \in \mathbb{R}^{N_c \times d_e}$. We denote N_c the number of words in each code description, L the total number of code space and d_e the embedding size.

Contextual Transformation. Unlike discharge summaries, the code descriptions are usually noun phrases instead of sentences. During the pretraining of Word2Vec, we construct their embeddings directly from contextual information. As the contexts of a word in the sentence and the noun phrase are different, a gap exists in their embeddings between the words in code descriptions and clinical texts. To solve the gap, we propose the following module to align the words in them. The major differences between the clinical texts and code descriptions are the word order and writing style. Therefore, we feed the word embeddings of code descriptions into an LSTM, concatenate the output and input, then put the concatenated results into a self-attention layer to combine the most important temporal features, and finally generate an overall representation $w_{\text{SA}}^{(l)}$ for code l:

$$\tilde{w}^{(l)} = [\mathbf{w}^{(l)} \oplus \overrightarrow{\text{LSTM}}(\mathbf{w}^{(l)})] \quad , \tag{1}$$

$$\alpha_{\text{SA}}^{(l)} = \text{softmax}(\tilde{w}^{(l)} \cdot W_{\text{Att}} + b_{\text{Att}}) \quad , \tag{2}$$

$$w_{\text{SA}}^{(l)} = \alpha_{\text{SA}}^{(l)} \cdot \tilde{w}^{(l)} \cdot W_{\text{SA}} \quad , \tag{3}$$

where $\alpha_{\text{SA}}^{(l)}$ refers to the self-attention weight, $W_{\text{SA}} \in \mathbb{R}^{(d_e+u_{\text{SA}})\times d_e}$, $W_{\text{Att}} \in \mathbb{R}^{(d_e+u_{\text{SA}})\times 1}$ and $b_{\text{Att}} \in \mathbb{R}^1$ are shared trainable vectors for all codes with u_{SA} the hidden size of LSTM and softmax is applied at the row level.

Code Hierarchy. Due to the lack of samples, the collected co-occurrence relations for the rare codes can be incomplete or biased. To further exploit the inter-code relations for rare codes, we introduce the code-organ-system hierarchical structure, where three levels are defined: level-0 (itself), level-1 (similar organs), and level-2 (same system). Observed from sufficient samples, the codes concerning similar organs or systems have some intrinsic co-occurrence links. These links can be utilized to enrich the connections for rare codes and make them more reliable and less biased via shared embeddings. We define the embedding of each level as the average of all the codes belonging to same categories. Thus, we obtain three embeddings which describe code l from different levels:

$$w_{\text{HC}}^{(l,p)} = \begin{cases} w_{\text{SA}}^{(l)}, & p = 0 \\ \dfrac{1}{|C_p|} \displaystyle\sum_{i \in C_p} w_{\text{SA}}^{(i)}, & p \in \{1, 2\} \end{cases} \quad , \tag{4}$$

where $p \in \{0, 1, 2\}$ is the level number and C_p the category which code l belongs to.

Code Co-occurrence. After obtaining the embeddings at different levels for each code, we adopt three GCNs to exploit the co-occurrence on these three levels. The inputs of these GCNs are the corresponding level embedding $w_{\text{HC}}^{(l,0)}$, $w_{\text{HC}}^{(l,1)}$, $w_{\text{HC}}^{(l,2)}$ and the adjacency matrix $A^{(1)}$, $A^{(2)}$, $A^{(3)}$ based on the co-appearing times for the three levels. The co-occurring times are sampled from a group of data, which is later analysed in Section 4.6. We use a standard convolution computation (Kipf and Welling, 2017):

$$E_{i+1} = \text{ReLU}(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} E_i W_i) \quad , \tag{5}$$

where $\tilde{A} = A + I$, I is the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. Thus, we have the output node representations $[w_{\text{CO}}^{(1,p)}; ...; w_{\text{CO}}^{(L,p)}]$ with $w_{\text{CO}}^{(l,p)}$ for code l at level p.

Code Embedding. We merge the three level representations in this module and obtain a specific representation $\mathbf{e}^{(l)}$ for code l in the label space:

$$w_{\text{CE}}^{(l,p)} = \text{LayerNorm}(w_{\text{CO}}^{(l,p)} + w_{\text{HC}}^{(l,p)}) \quad , \tag{6}$$

$$\mathbf{e}^{(l)} = \sum_{p=0}^{2} \alpha_{\text{CE}}^{(l,p)} w_{\text{CE}}^{(l,p)} \quad , \tag{7}$$

where $\alpha_{\text{CE}}^{(l,p)}$ is calculated using the same self-attention method as Eq. 2.

3.2 Clinical Text Encoder

A sequence of words from electronic medical records is transformed into word embedding via the same Word2Vec with embedding size d_e. Assuming the number of words N_w, the input word embedding matrix can be written as

$$\mathbf{X}_w = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_w}] \in \mathbb{R}^{N_w \times d_e} \quad . \quad (8)$$

Then we use a Bidirectional LSTM to capture the contextual information of the words. Although the Transformer-based models have taken a great leading place in various NLP applications in recent years, they are not that favorable for this task. We argue this point for the following reasons and verify it with empirical evidence: Unlike other NLP tasks, its vocabulary is domain-specific and thus low-dimensional. Using a Transformer-based encoder may add difficulty and redundancy to the training, costing more time and space. Besides, the sentences in EMR-like documents are not necessarily long and quite concentrated in their meanings. The long dependency issue is not very phenomenal in this case.

Finally, we compute the document representation by concatenating the output $\overrightarrow{\mathbf{h}_i}$ of $\overrightarrow{\text{LSTM}}$ and $\overleftarrow{\mathbf{h}_i}$ of $\overleftarrow{\text{LSTM}}$ for word $\mathbf{x}_i$. All the representations of words in the document formulate the document representation $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{N_w}] \in \mathbb{R}^{N_w \times 2u}$ with $\mathbf{h}_i = \overrightarrow{\mathbf{h}_i} \oplus \overleftarrow{\mathbf{h}_i}$ and u the hidden size.

3.3 Code-Text Attention Block

After the above modules, we obtain a text representation $\mathbf{H} \in \mathbb{R}^{N_w \times 2u}$, introduced in Section 3.2, and the code representations $\mathbf{e} = [\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, ..., \mathbf{e}^{(L)}] \in \mathbb{R}^{L \times d_c}$ in Section 3.1. The traditional label-wise attention (Vu et al., 2020; Mullenbach et al., 2018) generate the attention weights from text encoding H. Here we propose a more code-constrained version, involving both text and code representations:

$$\mathbf{A} = \text{softmax}(\mathbf{e} \cdot \tanh(\mathbf{H}W_H)) \quad , \quad (9)$$

$$\mathbf{V} = \mathbf{A} \cdot \mathbf{H} \quad , \quad (10)$$

where W_H is a matrix $\in \mathbb{R}^{2u \times d_c}$ which maps the document representation to the code embedding space to avoid dimension mismatch. $\mathbf{A} \in \mathbb{R}^{L \times N_w}$ denotes a code-specific weight, which is measured by how similar the documents are represented with each code. Afterwards, $\mathbf{A}$ is multiplied with the document $\mathbf{H}$ to generate a code-specific representation $\mathbf{V} \in \mathbb{R}^{L \times 2u}$.

3.4 Output Layer

With the code-specific representation $\mathbf{V}$, we calculate the probability for each code by passing each vector $\mathbf{V}_l \in \mathbb{R}^{2u}$ into a fully connected layer following an activation function sigmoid to produce the binary prediction score for code l. The training objective is to minimise the binary cross entropy between the prediction score $\hat{y}_l$ and target y_l:

$$\sum_{l \in \mathbb{L}} -y_l \log(\hat{y}_l) - (1 - y_l) \log(1 - \hat{y}_l) \quad . \quad (11)$$

4 Experiments

4.1 Dataset

The Medical Information Mart for Intensive Care III, denoted MIMIC-III, is a large open-source dataset (Johnson et al., 2016), containing the medical records of over forty thousand patients in the Beth Israel Deaconess Medical Center between 2001 and 2012. In terms of our problem, we use specifically the discharge summaries and their corresponding ICD-9 codes. Following the previous works (Shi et al., 2017; Mullenbach et al., 2018), we rearrange the data, where in total 52,726 discharge summaries with 8,929 ICD-9 codes served as labels are formulated. We split the dataset with the settings of Mullenbach et al. (2018). The data statistics are shown in Table 1.

MIMIC-III-Full			
	Train	Dev	Test
# Doc.	47,723	1631	3372
Avg words per Doc.	1484	1785	1792
Avg codes per Doc.	16.1	23.2	20.9

Table 1: Statistics of MIMIC-III-Full dataset.

4.2 Implementation Details

We follow the preprocessing schema of Vu et al. (2020) in our experiments. The CBOW is utilized during the pretraining stage, with the embedding size $d_e = 128$ on the processed text. All the documents are truncated with maximum 4,000 words. A data augmentation strategy (**w/ Sentence Permutation**) are applied in the experiment, where multiple sentences in the same document are shuffled in a random order to generate a new sample for training (Kim and Ganapathi, 2021). In our experiment, we use the 3-fold augmentation, i.e. increasing the training set three times. To show that our method

Model	AUC		F1		P@k	
	Macro	Micro	Macro	Micro	k=8	k=15
CAML (Mullenbach et al., 2018)	89.5	98.6	8.8	53.9	70.9	56.1
DR-CAML (Mullenbach et al., 2018)	89.7	98.5	8.6	52.9	69.0	54.8
TransICD (Biswas et al., 2021)	89.7	98.5	8.4	51.1	67.9	53.3
MultiResCNN (Li and Yu, 2020)	91.0	98.6	8.5	55.2	73.4	58.4
HyperCore (Cao et al., 2020a)	93.0	98.9	9.0	55.1	72.2	57.9
LAAT (Vu et al., 2020)	91.9	98.8	9.9	57.5	73.8	59.1
JointLAAT (Vu et al., 2020)	92.1	98.8	10.7	57.5	73.5	59.0
MSMN (Yuan et al., 2022)	95.0	**99.2**	10.3	58.4	75.2	59.9
Baseline (Label-wise attention)	89.4	98.6	9.0	56.2	73.9	58.8
Ours	95.0	**99.2**	10.3	58.0	**75.3**	59.9
Ours w/ Sentence Permutation	**95.2**	**99.2**	**10.8**	58.2	75.1	59.9
Ours w/ Enriched Descriptions	**95.2**	**99.2**	**10.8**	**58.6**	**75.3**	**60.3**

Table 2: Results on MIMIC-III-full, i.e. all codes. We compare our models with all the baselines by their values reported in the original papers for overall metrics.

can be complementary with other existing methods, we conduct experiments with the synonyms enriched descriptions (**w/ Enriched Descriptions**) in MSMN (Yuan et al., 2022) as well.

For the text encoder, we set LSTM hidden size as 256 with 2 layers. The dropout rate is 0.3. We add an extra linear layer after LSTM output with the dimension of 256. Our model is trained with AdamW (Loshchilov and Hutter, 2019) at a learning rate of $1e^{-3}$ on a single NVIDIA Tesla A100 (40GB). The batch size is set as 32. The early stopping mechanism is applied, in which the training will be stopped if there is no improvement of the micro-F1 score on the validation set in ten successive epochs. We run the experiment with 3 random seeds and report the average.

4.3 Evaluation Metrics

To make a comparison with other previous works on ICD-9 prediction, we evaluate the model performance by **F1**, **AUC** (area under the ROC curve) and **P@k**. F1 and AUC are calculated in two manners: macro-averaged, i.e. a simple average of all labels, and micro-averaged, i.e. aggregating the contributions of all classes to compute the average. P@k denotes the precision of top-k predicted results. K is conventionally set as the average number of labels for each document. However, it is not applicable for this task, since the number of labels for each document varies widely.

The macro-averaged metrics cannot thoroughly represent the rare code performance due to the huge gap in prediction performance between rare codes and frequent codes. Our experiment is designed to validate the idea that by enhancing the relations between frequent codes and rare codes, our method can achieve better performance on rare codes. Considering the insufficiency of distinguishing overall results and rare code performances, we select the codes with label frequency in the training set between 2-10 as rare codes, and report the results on predicting them. Since the macro- and micro- averaged results are very similar under this rare code setting (due to the similarity in label frequency), we only report their micro metrics.

4.4 Baselines

Our model is compared against the following SOTA models, chosen by their task settings:

CAML (Mullenbach et al., 2018), i.e. the Convolutional Attention network for Multi-Label classification, utilizes CNN as text encoder and propose a label attention for prediction. Meanwhile, they propose a **DR-CAML** version, where the ICD-9 code descriptions are used for regularization to improve the performance on rare codes, with a similar purpose to ours.

TransICD (Biswas et al., 2021) adopts an Transformer Encoder for discharge summaries.

MultiResCNN (Li and Yu, 2020) encodes the text with multi-filter residual CNN.

HyperCore (Cao et al., 2020a) takes also code co-occurrence and hierarchy into consideration. It embeds both code and text into hyperbolic space and calculates their similarities.

LAAT (Vu et al., 2020) applies also LSTM text encoder. A hierarchical joint structure, i.e. **JointLAAT**, is proposed to solve the imbalance label

distribution, thus ameliorate the performances on rare codes.

MSMN (Yuan et al., 2022) enriches the code descriptions with synonyms from UMLS and encodes the clinical text with Bi-LSTM.

Baseline is designed by replacing the code-text attention and the relation-enhanced code encoder with label-wise attention (Vu et al., 2020).

4.5 Results

MIMIC-III-Full. Table 2 shows the results on MIMIC-III for all codes, where our model outperforms all the other baselines. By adopting the code descriptions with the synonyms from UMLS, our model performs better and outperforms MSMN in almost all metrics. With the simple code description, our method still produces comparable, even better results without the necessity of bringing external knowledge source. Moreover, the large margin between baseline and ours validates the idea that by enhancing the inter-code relations, model can produce better results than traditional label-wise attention method.

Rare Codes. We collect all the codes with an appearing times in the training set between 2 and 10, and observe specifically the model performance on these codes. As shown in Table 3, our model achieves the best results among all the baselines. With Sentence Permutation, the improvements are more significant. It is interesting to observe that CAML and DR-CAML mess up all the predictions for these rare codes. Though JointLAAT is proposed at the intention of improving its few-shot performances, where the final prediction is based on the prediction on the codes starting with the same first three characters, the results actually degrade compared with LAAT. It is because of a low recall, since excluding the first level codes affects the prediction on its child codes as well. Our model exploits the inter-code relations from co-occurrence and enhances them for rare codes via hierarchy, thus producing better results and leaving lower margin to the overall results.

4.6 Adjacency Matrix

It is commonly recognized that more training data can lead to a better performance. However, for clinical documents, collecting them and tagging the ICD codes can be quite difficult due to its privacy and difficulty of processing. In our model, we apply a GCN module with adjacency matrix (ADJ) based on co-appearing relations. The above experiments are conducted with ADJ sampled from training set, which is denoted "Training". We are wondering if the performances can be further ameliorated with ADJ of more samples, or derived from some prior medical knowledge. This idea is analysed by proposing another two ADJs. "Full" denotes the ADJ sampled from the full dataset, including training, validation and testing sets. Besides, we add the co-appearing relations in MIMIC-IV (Johnson et al., 2021) as well and denote this matrix "w/ MIMIC-IV".

Model	Micro AUC	Micro F1
CAML	50.0	0.00
DR-CAML	50.0	0.00
TransICD	79.9	5.54
MultiResCNN	81.7	0.65
LAAT	88.8	3.23
JointLAAT	87.2	0.79
Baseline	80.4	4.23
Ours	**94.5**	6.72
Ours w/ SP	**94.5**	**7.15**

Table 3: Results on rare codes, i.e. codes whose frequency in training set is between 2-10. Since this metric is not considered in the original papers for baselines, we select the models with released codes and re-implement their experiments.

Table 4 shows the results on full codes and rare codes with various adjacency matrices. Since the macro AUC is not very sensitive, we just list the macro/micro F1 and micro AUC. We notice that having a more complete and reasonable adjacency matrix can help the model prediction, since both of the F1 metrics get better.

MIMIC-III-Full			
	F1		AUC
	Macro	Micro	Micro
Training	10.3	58.0	**99.2**
Full	10.6	**58.2**	99.1
w/ MIMIC-IV	**10.7**	**58.2**	99.1
Rare Codes			
	F1		AUC
Training	6.72		94.6
Full	7.13		**94.8**
w/ MIMIC-IV	**7.35**		94.2

Table 4: Influence of different adjacency matrices.

4.7 Model Interpretability

Being able to explain the model decision with conformance to human understanding is an important criterion in healthcare. To prove the effectiveness of our model on rare code prediction, we provide in Figure 4 some spans from the same discharge summaries where the tokens with high attention scores locate for predicting *Pneumonia in other systemic mycoses* (484.7), an infrequent code with only 17 documents.

We notice that the baseline model puts high attentions to the words like "discharge", "date" and "marrow", whose relevance with *Pneumonia in other systemic mycoses* is hard to find because almost all the documents contain them. Our model can capture the closely related words like "pneumonia", "aspergillus" and "lobe", indicating a better code representation.

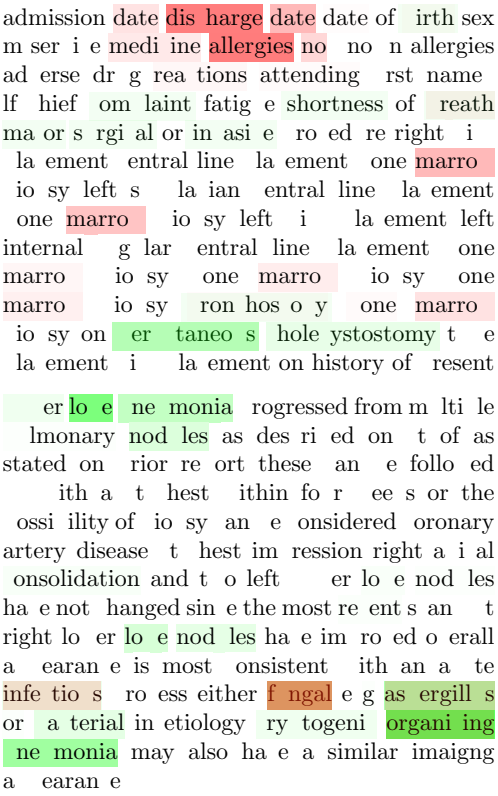

Figure 4: The spans with high attentions (darker means higher) of the same text for predicting *Pneumonia in other systemic mycoses* (484.7 with 17 documents). The red shows the interpretation for the baseline and the green for ours.

5 Ablation Study

We conduct the ablation study concerning the effectiveness of the contextual transformation, code hierarchy and code co-occurrence modules inside the relation-enhanced code encoder. We measure the F1 metrics, which are more sensitive and representative to different models, between our ordinary version and those with a module removed. The results without the entire code encoder, which means using only the embeddings of code descriptions as query, are also listed.

As shown in Table 5, removing the transformation module causes a significant decrease, indicating that the gap we described between descriptions and discharge summaries do exist. Besides, we notice that removing the code hierarchy or the code co-occurrence module has also degraded the model performances, showing that the code hierarchy and co-occurrence are useful to the model. Since the three GCNs may bring extra training difficulties, further tuning the training process might still be helpful to improve the performances.

Model	F1	
	Macro	Micro
Ours	**10.3**	**58.0**
w/o transformation	9.8	57.2
w/o hierarchy	10.2	57.9
w/o co-occurrence	10.1	57.7
w/o code encoder	7.5	51.0

Table 5: Results of ablation study on all codes.

6 Conclusions

The multi-label clinical text classification is an important task in the domains of both healthcare and natural language processing. In this paper, we reveal the existing problem of traditional methods in capturing the inter-code relations for rare codes. Hereby, we propose to strengthen the relations, thus improving the model performance on rare code prediction. We exploit the inter-code relations by encoding code descriptions and incorporating co-occurrence under a code-organ-system hierarchical structure in order to enhance the connections for rare codes. Our model is then evaluated on the commonly used MIMIC-III dataset and outperforms the other baselines on both rare codes and full codes. The visualisations further demonstrate the advantage of our method on providing more human-understandable explanations. We conduct as well an analysis concerning the design of adjacency matrices and the ablation study to better understand the different components in our method.

Limitations

In this work, we adopt three GCNs to exploit the inter-code relations under different levels. However, this may bring extra training difficulty and the risk of over-parameterization to the model. Besides, during the preprocessing stage, we adopt a word-level tokenizer and CBOW to obtain their embeddings for MIMIC-III texts and code descriptions. However, this might not be enough to represent the words since medical documents have some special characteristics, but we do not take them into consideration. We tried in our work with other pretraining strategies, such as ClinicalBERT (Alsentzer et al., 2019), BERT (Devlin et al., 2019), BioBERT (Lee et al., 2020) and BioWordVec (Zhang et al., 2019). We added as well the BPE tokenizer (Sennrich et al., 2016) in order to capture the meaningful medical sub-word units. However, the results are all far from satisfactory.

References

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

Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. 2021. Transicd: Transformer based code-wise attention model for explainable ICD coding. In *Artificial Intelligence in Medicine - 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15-18, 2021, Proceedings*, volume 12721 of *Lecture Notes in Computer Science*, pages 469–478. Springer.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1565–1576.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020a. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, Online. Association for Computational Linguistics.

Pengfei Cao, Chenwei Yan, Xiangling Fu, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020b. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 294–301, Online. Association for Computational Linguistics.

Luciano R. S. de Lima, Alberto H. F. Laender, and Berthier A. Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management, Bethesda, Maryland, USA, November 3-7, 1998*, pages 132–139. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Dynamic memory induction networks for few-shot text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1087–1094, Online. Association for Computational Linguistics.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and S Horng IV. 2021. Mimic-iv-ed. *PhysioNet*.

Alistair Johnson, Tom Pollard, and R Mark III. 2016. Mimic-iii clinical database. *Physio Net*, 10:C2XW26.

Byung-Hak Kim and Varun Ganapathi. 2021. Read, attend, and code: Pushing the limits of medical codes prediction from clinical notes by machines. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2021, 6-7 August 2021, Virtual Event*, volume 149 of *Proceedings of Machine Learning Research*, pages 196–208. PMLR.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Bevan Koopman, Guido Zuccon, Anthony N. Nguyen, Anton Bergheim, and Narelle Grayson. 2015. Automatic ICD-10 classification of cancers from free-text death certificates. *Int. J. Medical Informatics*, 84(11):956–965.

Leah S. Larkey and W. Bruce Croft. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 289–297. ACM.

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

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8180–8187. AAAI Press.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. 2017. Meta-sgd: Learning to learn quickly for few shot learning. *CoRR*, abs/1707.09835.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Carlos Medina, Arnout Devos, and Matthias Grossglauser. 2020. Self-supervised prototypical transfer learning for few-shot classification. *CoRR*, abs/2006.11325.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

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

Adler J. Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Gray Weiskopf, Frank D. Wood, and Noemie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *J. Am. Medical Informatics Assoc.*, 21(2):231–237.

John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards automated ICD coding using deep learning. *CoRR*, abs/1711.04075.

Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric P. Xing. 2020. Generalized zero-shot text classification for ICD coding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4018–4024. ijcai.org.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):1–10.

Shang-Chi Tsai, Chao-Wei Huang, and Yun-Nung Chen. 2021. Modeling diagnostic label correlation for automatic ICD coding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4043–4052. Association for Computational Linguistics.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3630–3638.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3335–3341. ijcai.org.

Aaron Sonabend W., Winston Cai, Yuri Ahuja, Ashwin N. Ananthakrishnan, Zongqi Xia, Sheng Yu, and Chuan Hong. 2020. Automated ICD coding via unsupervised knowledge integration (UNITE). *Int. J. Medical Informatics*, 139:104135.

Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Huasheng Liang, Qiang Yan, Evangelos Kanoulas, and Maarten de Rijke. 2021. Few-shot electronic health record coding through graph contrastive learning. *CoRR*, abs/2106.15467.

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, Melbourne, Australia. Association for Computational Linguistics.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 649–658. ACM.

Yan Yan, Glenn Fung, Jennifer G. Dy, and Rómer Rosales. 2010. Medical coding classification by leveraging inter-code relationships. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 193–202. ACM.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland. Association for Computational Linguistics.

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