ICD Coding from Clinical Text with Entity-aware Label Attention

Yunfei Yang[1], **Quan Wang**[2], **Yajuan Lyu**[2], **Yong Zhu**[2], **Zhifang Sui**[1]

Abstract

ICD coding is an important task in medical domain that assigns appropriate ICD codes to a clinical note. This task is challenging for two reasons: high-dimensional label space with a long-tail distribution and overlong clinical notes with numerous redundant content. Existing studies utilize ICD intrinsic attributes (such as code description and code hierarchy) to improve performance on infrequent codes and employ a label attention mechanism to capture the most relevant tokens for each label. However, previous methods use the raw textual descriptions which inevitably include irrelevant information and perform the label attention process independently of labels' textual descriptions though they are highly instructive and easily available. In this paper, we propose a novel **E**ntity-**A**ware **L**abel A**t**tention (EALT) mechanism. Concretely, we focus on entities instead of raw ICD descriptions as a prior knowledge to improve the performance for infrequent codes. Meanwhile, we incorporate entities into the label attention mechanism to exploit the instructive information of ICD descriptions. We demonstrate the effectiveness of our method on the widely used MIMIC datasets. Evaluation results show our model achieves state-of-the-art performance over several strong baselines.

1 Introduction

The International Classification of Diseases (ICD), which is maintained by the World Health Organization (Organization et al., 1978), is the classification standard of assigning codes of diagnoses and procedures to a medical record. ICD codes are used for a variety of purposes, including billing and providing information on diagnoses and procedures during a patient's visit (Bottle and Aylin, 2008; Choi et al., 2016; Denny et al., 2010). ICD coding is the task of assigning appropriate ICD codes to a clinical note, which is of significant importance in medical management systems. Typically, ICD coding is performed by a professional coder manually according to an electronic medical record (EMR) recorded by physicians. However, due to the increasing number of codes and numerous irrelevant information, ambiguous abbreviations in overlong clinical text, manual coding is laborious, time-consuming, and extremely prone to errors (O'malley et al., 2005; Nguyen et al., 2018).

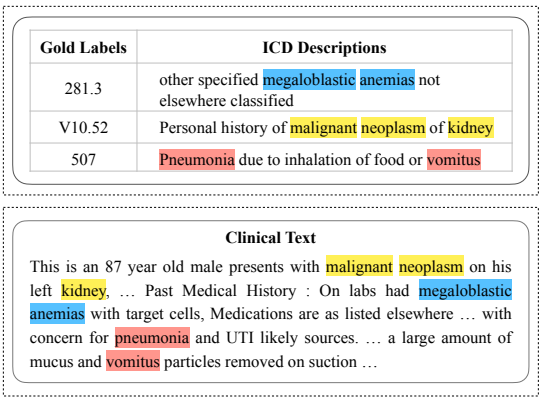

Figure 1: An example of interactions between ICD descriptions and clinical text. We divide ICD description into entity (highlight text) and non-entity tokens. "megaloblastic anemias" is an entity in the description of code "281.3", which probably indicates the presence of this code when occurring in clinical text.

To solve the expensive and time-consuming problems with manual coding, researchers have proposed many automatic ICD coding methods utilizing both machine learning and deep learning techniques (Larkey and Croft, 1996; Xie and Xing, 2018; Mullenbach et al., 2018). ICD coding can be modeled as a multi-label classification task and is challenging for the following two reasons. **First**, the label space is very high-dimensional with an extremely long-tail distribution. There are over

15,000 codes in the ICD-9 taxonomy and about 5411 of all the 8929 labels occurring less than 10 times in MIMIC-III dataset (Johnson et al., 2016; Xie et al., 2019). **Second**, the clinical notes are typically overlong (more than 1500 tokens on average) with numerous irrelevant information, misspellings and ambiguous abbreviations (Mullenbach et al., 2018). However, only a small set of tokens in a clinical note are relevant to a specific ICD code.

Previous state-of-the-art models utilize ICD intrinsic attributes to improve performance on infrequent codes. For example, (Mullenbach et al., 2018) utilizes ICD descriptions as a regularization on the model parameters. The regularizer tries to learn similar representations between documents and corresponding label descriptions. However, as they use raw textual descriptions, which inevitably include irrelevant information with a diagnosis disease. Meanwhile, since they leverage golden label of a document, the description can only use in training process.

To make each label attend to the most predictive tokens in a document, (Mullenbach et al., 2018; Li and Yu, 2020; Vu et al., 2020) utilize a label attention mechanism to select the most informative tokens for each label adaptively. However, these methods perform the label attention process independently of labels' textual descriptions though they are highly instructive and easily available.

In this paper, we propose a novel **E**ntity-**A**ware **L**abel A**t**tention (EALT) model that incorporates entities into the label attention mechanism. Specifically, to improve the performance for infrequent codes, we focus on entities (such as diseases and symptoms) instead of raw ICD descriptions, as entities often contain key tokens to predict a code and provide a prior knowledge which leads to a bias towards corresponding labels. Meanwhile, irrelevant information in raw ICD descriptions could be excluded naturally. As Figure 1 shows, take code "281.3" for example, "megaloblastic anemias" is an entity in the description of code "281.3", which probably indicates the presence of code "281.3" when occurring in clinical text. However, other tokens in ICD descriptions such as "elsewhere" and "classified" have a weak correlation with the code's presence. We utilize an entity-aware label attention mechanism to select the most informative tokens for each label adaptively. To exploit the highly instructive and easily available labels' textual descriptions, we introduce a label-specific entity indicator which enables the model to distinguish the entity tokens for each label explicitly and employ different query mechanisms to emphasize those entity tokens during label attention process.

The main contributions of this paper are summarized as follows:

- We leverage entities in ICD descriptions as a prior knowledge to improve the performance for infrequent codes.
- We propose an entity-aware label attention mechanism to select the most informative tokens for each label adaptively. The mechanism incorporates entities in labels' textual descriptions into label attention process.
- Evaluation results show that our model achieves state-of-the-art performance on the widely used MIMIC datasets.

2 Related Work

Automatic ICD Coding. Automatic ICD coding is a long-standing task in the healthcare domain, it has been studied using machine learning method early in the 1990s (Larkey and Croft, 1996; de Lima et al., 1998). (Larkey and Croft, 1996) formulate the ICD coding as a single label classification task, they propose an ensemble model consisting of three classifiers: bayesian independence, k-nearest neighbors and relevance feedback. (Perotte et al., 2014) propose a hierarchical support vector machine (SVM) to utilize ICD hierarchical strutures. Recently, deep learning methods have been increasingly used to solve the ICD coding problem. These methods can be mainly divided into CNN-based (Mullenbach et al., 2018; Xie et al., 2019; Li and Yu, 2020) and RNN-based (Xie and Xing, 2018; Vu et al., 2020). (Mullenbach et al., 2018) utilizes convolution neural network (CNN) with attentions to select most relevant information in clinical notes. (Li and Yu, 2020) propose a CNN architecture that combines the multi-filter CNN and residual CNN (He et al., 2016), which can capture important information with different lengths. (Vu et al., 2020) utilizes a bidirectional Long-Short Term Memory (BiLSTM) to handle both the various lengths and the interdependence of the ICD code related text fragments.

Label attention mechanism. Attention mechanism has been very popular in most of the NLP

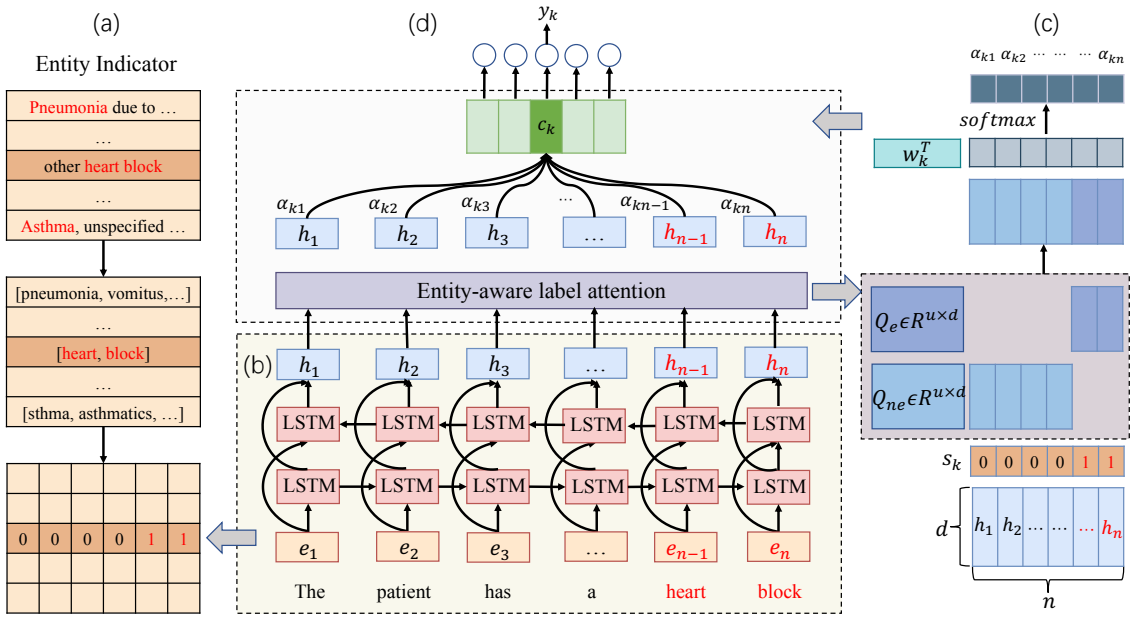

Figure 2: The architecture of our EALT model. (a) shows a Entity Indicator that identifies entity tokens in a input text for each label. Label-specific document representation c_k is computed as weighted sums of hidden states. (c) illustrates the way of computing summation weights. w_k is the parameter of label l_k, Q_e and Q_{ne} are query matrices for entity and non-entity tokens respectively, s_k is a multi-hot vector obtained from Entity Indicator.

tasks, there are many variants of attention mechanism proposed for different purposes (Luong et al., 2015; Kim et al., 2017; Vaswani et al., 2017). (Mullenbach et al., 2018) introduces label attention mechanism on ICD coding task. Label attention can help the model attend to most important part in a clinical document and significantly improves the performance. (Vu et al., 2020) proposes a new label attention model which is an extension of structured self-attention mechanism (Lin et al., 2017). Our attention mechanism is inspired by (Yamada et al., 2020). This work extends the transformer using an entity-aware self-attention mechanism and adopts different query mechanisms based on the type of the tokens (words or entities). The difference between this work and our method is that the token type in this work is query-insensitive, however, we utilize a label-specific entity indicator to determine whether a token is an entity or non-entity for a specific code.

Medical Entities. Medical entities play a key role in biomedical domain. There are many high-quality knowledge bases consisting of the huge number of entities and semantic relations between them, such as UMLS (Bodenreider, 2004), MeSH (Rogers, 1963) and ICD taxonomy. Entities are also used in many medical NLP tasks, for example, Medical Entity Normalization aims to link mentions of named entities in natural language text to entities in a curated knowledge-base (Leaman et al., 2013; Zhu et al., 2020). As ICD descriptions often contain irrelevant information for a diagnosis disease, we focus on entities instead of raw ICD descriptions to improve the performance for infrequent codes.

3 Method

3.1 Problem Definition

We formulate the ICD coding task as a multi-label text classification problem (McCallum, 1999). The input clinical document X can be represented as n word sequence $X = \{x_1, x_2, ..., x_n\}$. Let $L = \{l_1, l_2, ..., l_{|L|}\}$ denote the set of ICD-9 codes. The problem can be defined as follows: given a document W and codes set L, the objective is to train $|L|$ binary classifiers, in which each classifier is to determine the value of $y_k \in \{0, 1\}, k = 1, 2, ...|L|$, where y_k is a prediction for label l_k.

3.2 Model Description

As shown in Figure 2, Our proposed model consists of several modules: (a) The label-specific entity indicator that identifies entity tokens in a input text for each label. The main procedure includes

	MIMIC-III-full			MIMIC-III-50		
	training	validation	testing	training	validation	testing
samples	47719	1631	3372	8067	1574	1730
Unique codes	8692	3012	4035	50	50	50
Avg.# tokens per document	1434	1724	1731	1478	1739	1763
Avg.# codes per document	15.68	17.31	17.59	5.69	5.87	6.03

Table 1: The detailed statistics of MIMIC-III-full and MIMIC-III-50 datasets.

recognizing entities in labels' textual description, tokenizing entities and matching with document. (b) BiLSTM encoder: A bidirectional Long-Short Term Memory (Hochreiter and Schmidhuber, 1997) encoder is utilized to capture contextual information in a clinical document. (c) Entity-aware label attention layer: The goal is to learn a label-specific document representation c_k which is computed as weighted sums of hidden states. (d) Output layer: After obtain c_k, we exploit a single layer full-connected neural network followed by a sigmoid activation function to compute the probability and determine the value of $y_k \in \{0, 1\}$, where y_k is a prediction for label l_k.

Label-specific Entity Indicator. Given a input sequence $X = \{x_1, x_2, ..., x_n\}$ and codes set $L = \{l_1, l_2, ..., l_{|L|}\}$, the indicator will output $|L|$ multi-hot vectors, denoted as $S = \{s_1, s_2, ..., s_{|L|}\}$, where $\{s_k\}_{k=1}^{|L|} \in R^n$, the i^{th} element is 1 if and only if x_i is an entity token for code l_k. Specifically, for each $l_k \in L$, we first do a named entity recognition process on its textual description simply using spacy[1] tools and obtain the entities in textual description. Note that if there is only one token in the description, we treat it as an entity directly. Then we tokenize the entities and obtain a token set t_k. As Figure 3 shows, in a $|L| \times n$ 0-1 matrix, the values in k^{th} row represent the value of vector s_k. If t_k contains the token x_i, then the value of k^{th} row and i^{th} column (denoted as s_{ki}) is 1, otherwise the value is 0.

Embedding Layer. Embedding layer is the first layer in our model, this layer takes word sequences $X = \{x_1, x_2, ..., x_n\}$ as input, each word $x \in V$ is mapped to $e \in R^{d_e}$ using embedding matrix E, where V denotes the whole vocabulary and $E \in R^{|V| \times R^{d_e}}$ denotes the pre-trained weight matrix of V. d_e is the embedding size. We use a pre-trained word vector released from (Vu et al., 2020), which is trained on all processed discharge summaries in the MIMIC-III-full dataset with $d_e = 100$ using CBOW Word2Vec method (Mikolov et al., 2013). After input sequence pass through this layer, we obtain a word embedding matrix $E_X = \{e_1, e_2, ..., e_n\}$ as output.

[1] https://spacy.io/, we use en_ner_bc5cdr_md-0.3 model, which is trained on the BC5CDR corpus.

	x_1	x_2	...	x_i	...	x_n
s_1	0	1	0	1	1	0
s_2	0	0	1	1	0	0
...	1	1	0	0	0	0
$s_{\|L\|}$	0	0	1	0	1	1

Figure 3: An example output of entity indicator. For a $|L|$ label and n word input, it can be represented as a $|L| \times n$ 0-1 matrix, a value 1 in k^{th} row and i^{th} column denotes token x_i is an entity token for label l_k.

Bidirectional LSTM Layer. To capture contextual information of the input clinical document, we exploit a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to encode the clinical text. This layer takes the embedding matrix $E_X = \{e_1, e_2, ..., e_n\}$ as input, which is obtained from embedding layer. Then we compute the contextual representations of $i^{th}, i \in 1, 2, ...n$ word as:

$$h_i^f = \overrightarrow{LSTM}(h_{i-1}^f, e_i)$$

$$h_i^b = \overleftarrow{LSTM}(h_{i+1}^b, e_i)$$

where h_i^f and h_i^b represent forward and backward hidden state of the i^{th} word, $\overrightarrow{LSTM}$ and $\overleftarrow{LSTM}$ denote bidirectional LSTMs. Let the dimension of h_i^f and h_i^b be d_h, then we concatenate the bidirectional hidden state:

$$h_i = [h_i^f; h_i^b]$$

where $h_i \in R^d, d = 2 * d_h$ denotes the final hidden representation of token x_i.

Model	AUC		F1		P@5
	Macro	Micro	Macro	Micro	
LR	82.9	86.4	47.7	53.3	54.6
CNN	87.6	90.7	57.6	62.5	62.0
BiGRU	82.8	86.8	48.4	54.9	59.1
CAML(Mullenbach et al., 2018)	87.5	90.9	57.6	62.5	62.0
DR-CAML(Mullenbach et al., 2018)	88.4	91.6	57.6	63.3	61.8
MSATT-KG(Xie et al., 2019)	91.4	93.6	63.8	68.4	64.4
MultiResCNN(Li and Yu, 2020)	89.9	92.8	60.6	67.0	64.1
HyperCore(Cao et al., 2020)	89.5	92.9	60.6	66.3	63.2
LAAT (Vu et al., 2020)	<u>92.3</u>	<u>94.4</u>	<u>66.4</u>	<u>70.5</u>	<u>67.0</u>
EALT(ours)	**92.7**	**94.6**	**68.4**	**71.9**	**67.2**

Table 2: Results on the MIMIC-III-50 test set. **Bold** text denotes the best results, while the second best is <u>underlined</u>.

Entity-aware label attention Layer. To attend the most informative tokens for each label and exploit the instructive effect of entities in labels' textual description, we incorporate an entity-aware label attention layer. Our attention mechanism is an extension of label attention from (Vu et al., 2020) and entity-aware self-attention from (Yamada et al., 2020). This layer takes document representation H as input and outputs $|L|$ label-specific vectors. Specifically, given a document representation $H = \{h_1, h_2, ..., h_n\}$ from bidirectional LSTM layer, where $\{h_i\}_{i=1}^{n} \in R^d$, each of the output vector $c_k \in R^d, k = 1, 2, ...|L|$ is obtained from the weighted sum of the input vectors. The k^{th} label-specific vector is computed as:

$$
\begin{aligned}
c_k &= \sum_{i=1}^{n} w_i h_i \\
\alpha_{ki} &= softmax(v_{ki}) \\
v_{ki} &= w_k^T Q h_i
\end{aligned} \tag{1}
$$

where $Q \in R^{u \times d}$, $w_k \in R^u$ represents query matrix and the vector parameter for label l_k. With the conjecture that it is instructive to enable the model to distinguish the entity tokens explicitly, we adopt two different query matrices Q_e and Q_{ne} for entity and non-entity tokens when calculating similarity score between labels and tokens. Specifically, the similarity score of k^{th} label and i^{th} token can be computed as:

$$
v_{ki} = \begin{cases} w_k^T Q_e h_i & \text{if } s_{ki} = 1. \\ w_k^T Q_{ne} h_i & \text{if } s_{ki} = 0. \end{cases} \tag{2}
$$

where $Q_e, Q_{ne} \in R^{u \times d}$ and $s_{ki} \in \{0, 1\}$ is obtained from label-specific entity indicator. Finally, we will obtain a matrix $C \in R^{|L| \times d}$ consisting of $|L|$ label-specific vectors $c_i, c_2, ..., c_{|L|}$, matrix C is the output of attention layer.

Output Layer. This layer takes the label-specific document representation C as input, and outputs a probability vector $o \in R^{|L|}$. For the k^{th} label-specific document representation c_k, we employ a single linear layer followed by a sigmoid activation function to compute the probability for k^{th} label. Then we use output probability o_k to determine the prediction result $\tilde{y} \in \{0, 1\}$. As we treat ICD coding task as a multi-label text classification problem (McCallum, 1999), the training objective is to minimize the binary cross entropy between the predicted label $\tilde{y}$ and the target y:

$$
L(X, y, \theta) = -\sum_{j=1}^{|L|} [y_j log(\tilde{y_j}) + (1 - y_j) log(1 - \tilde{y_j})]
$$

where θ denotes all the trainable parameters, X is the input sequence, and y represents true labels.

4 Experiments

4.1 Datasets

We evaluate our proposed method on the large, freely-available critical care database MIMIC-III (**M**edical **I**nformation **M**art for **I**ntensive **C**are III) (Johnson et al., 2016). MIMIC-III dataset comprises de-identified electronic medical records from over forty thousand patients, each record is annotated by clinical coders with a set of ICD-9-CM codes. Following previous work, we focus on two common experiment settings called **MIMIC-III-full** and **MIMIC-III-50**. Table 1 shows the detailed statistics of the two setting datasets.

Model	AUC		F1		P@N	
	Macro	Micro	Macro	Micro	P@8	P@15
LR	56.1	93.7	1.1	27.2	54.2	41.1
CNN	80.6	96.9	4.2	41.9	58.1	44.3
BiGRU	82.2	97.1	3.8	41.7	58.5	44.5
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
MSATT-KG(Xie et al., 2019)	91.0	**99.2**	9.0	55.3	72.8	58.1
MultiResCNN(Li and Yu, 2020)	91.0	98.6	8.5	52.2	73.4	58.4
HyperCore(Cao et al., 2020)	**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
EALT(ours)	92.2	98.9	**10.2**	**57.8**	**74.4**	**59.3**

Table 3: Results on the MIMIC-III-full test set. **Bold** text indicates the best results, while the second best is underlined.

MIMIC-III-full. In full-label setting, there are totally 8921 unique ICD-9 codes consisting of 6918 diagnosis codes and 2003 procedure codes which occurred in 52722 discharge summaries. We split the dataset using patient ID for comparison with previous work (Mullenbach et al., 2018). Specifically, we used 47719 discharge summaries for training, 1631 and 3372 for validation and testing, respectively.

MIMIC-III-50. In top-50 label setting, we conduct experiment using top 50 most frequent codes, and filter each dataset down to the instances that have at lease one code appearing in top-50 frequent codes set. Finally, there are 8067 discharge summaries for training, 1574 for validation, and 1730 for testing.

4.2 Preprocessing

Since the clinical text contains many non-alphabetic characters and is not well organized, we tokenize the text using NLTK (Bird, 2006) and remove all the numbers and punctuations, then we convert all tokens to lowercase. Besides, we remove some sections that contain rarely useful information after previous preprocessing, such as admission date, discharge date and birth date. Following the previous work (Xie et al., 2019) and (Vu et al., 2020), we set the maximum length of a token sequence to 4000 and truncated the sequence whose length exceeds the maximum length.

4.3 Evaluation Metrics

Following previous work, we report a variety of metrics including macro-averaged AUC (area under the ROC curve), micro-averaged AUC, macro F1, micro F1 and P@N (precision of N highest scored labels). We use P@5 in top-50 label setting, P@8 and P@15 in the full label setting. A macro-average score computes the metric independently for each label and then take the average, while a micro-average score aggregates the contributions of all labels to compute the average metric.

4.4 Hyper-parameters

To select optimal value of hyper-parameters, we perform a grid search on validation set for top-50 label setting. As the cost of grid search on all hyper-parameters for full-label setting is costly, most of hyper-parameters were chosen following privious work (Vu et al., 2020). Since the dataset is unbalance and the output probability is trongly biased towards negative predictions, we treat the threshold t as a hyper-parameter to make a prediction, where $t \in \{0.375, 0.4, 0.45, 0.5\}$. We use Adamw (Loshchilov and Hutter, 2019) to optimize the model parameters. For both top-50 and full label setting, the word embedding size d_e is 100, the hidden size of LSTM is 512, the vector parameter size for each label is 512, the batch size is 8, the initial learning rate is 5e-4, the threshold t is 0.375, the dropout after word embedding is 0.3 and 0.35 for top-50 label setting and full-label setting, respectively.

4.5 Baselines

We compare our proposed model with following baselines.

LR & CNN & BiGRU. These three baselines were employed by (Mullenbach et al., 2018) for ICD codes prediction on MIMIC datasets, which

represent a bags-of-words **l**ogistic **r**egression (LR) model, a single layer one-dimensional **c**onvolutional **n**eural **n**etwork (CNN) with max-pooling model (Kim, 2014), and a **b**idirectional **g**ated **r**ecurrent **u**nit (BiGRU) model (Cho et al., 2014).

CAML. (Mullenbach et al., 2018) proposes a **C**onvolutional **A**ttention **n**etwork for Multi-Label classification (CAML) model for ICD coding. CAML utilizes a single layer CNN to extract document features and employs a label-dependent attention mechanism to learn the most informative representation for each label.

DR-CAML. **D**escription **R**egularized CAML (DR-CAML) is an extension of the CAML model, employing label description to regularize the final loss function. Compare to CAML, DR-CAML has achieved a significant improvement on MIMIC-III-50 datasets, while it performs worse on most metrics than CAML on MIMIC-III-full datasets.

MultiResCNN. A novel CNN architecture that combines the multi-filter CNN (Kim, 2014) and residual CNN (He et al., 2016) was proposed by (Li and Yu, 2020). MultiResCNN captures various text patterns with different lengths via the multi-filter CNN and utilizes a residual convolutional layer to enlarge the receptive field.

MSATT-KG. (Xie et al., 2019) develop a model with **m**ulti-**s**cale feature **att**ention and structured **k**nowledge **g**raph propagation (MSATT-KG). The model incorporates a multi-scale feature attention to adaptively select the most informative n-gram features that produced by a densely connected convolutional neural network, and leverages graph convolutional neural network (Kipf and Welling, 2016) to capture the hierarchical relationships among medical codes. MSATT-KG achieves the state-of-the-art performences on most metrics over MIMIC-III-full and MIMIC-full-50 datasets.

HyperCore. A novel **Hyper**bolic and **Co**-graph **Re**presentation model was proposed by (Cao et al., 2020). HyperCore incroporates hyperbolic representation to learn code hierarchy and utilizes a GCN (Kipf and Welling, 2016) to exploit code co-occurrence correlation.

LAAT. (Vu et al., 2020) proposes a new **la**bel **at**tention (LAAT) model on ICD coding, and utilizes a bidirectional Long-Short Term Memory (BiLSTM) encoder to handle the different sizes of a single text fragment. LAAT achieved new state-of-the-art results on MIMIC datasets.

Model	AUC		F1		P@N	
	macro	micro	macro	micro	8	15
EALT	**92.9**	**98.9**	**9.1**	**58.4**	**74.6**	**58.9**
- entity aware	92.6	98.8	8.7	58.1	74.3	58.4
- label specific	92.7	98.8	8.1	57.4	73.5	58.4

Table 4: Ablation results on MIMIC-III-full validation dataset. - entity aware denotes removing entity-aware mechanism. - label-specific denotes a variant of our label-specific entity mechanism, in which we replace the label-specific entity indicator with a label-insensitive entity indicator.

4.6 Results

MIMIC-III-50 As shown in Table 2, our EALT model outperformed all the baselines on five evaluation metrics. Compare with the previous strong state-of-the-art model LAAT[2], our method improved the Macro-AUC by 0.4, the Micro-AUC by 0.2, the Macro-F1 by a considerable margin **2.0**, the Micro-F1 by **1.4** and p@5 by 0.2. Compare with other recent baselines, take HyperCore (Cao et al., 2020) for example, our model improved the Macro-AUC, Micro-AUC, Macro-F1, Micro-F1, p@5 by 3.2, 1.7, 8.4, 5.6 and 4.0, respectively.

MIMIC-III-full Table 3 shows the results on the MIMIC-III-full dataset. Following previous work (Mullenbach et al., 2018), we treat precision@8 as main metric, as it measures the ability of the system to return a small high-confidence subset of codes in a large label space. Our model outperformed the state-of-the-art model by 0.6 on precision@8. Moreover, our model improved the Macro-F1 by 0.3, Micro-F1 by 0.3, precision@15 by 0.2. We achieved second best performance on Macro-AUC and Micro-AUC.

4.7 Ablation Study

We conduct the ablation studies on MIMIC-III-full dataset, since this setting is more difficult and has both characteristics of high-dimensional label space with a long-tail distribution and overlong clinical notes. We demonstrate the effect of our meth-

[2]We read the source code of LAAT and observed that the top-50 labels used in LAAT is slightly different from other baselines. For a fair comparison, we rerun the code on the widely used top-50 dataset. We use optimal hyper-parameters listed in original paper and report the results in Table 2.

ods in two aspects: effect of **entity-aware** label attention mechanism and effect of **label-specific** entity indicator.

Effect of entity-aware label attention. We remove entity-aware mechanism, thus the model is simplified as LAAT. In this setting, the similarity score between labels and tokens is computed as Equation 1. As shown in Table 4, our entity-aware label attention mechanism improved macro F1 by 0.4, improved micro F1 by 0.4, improved precision@8 by 0.3 and improved precision@15 by 0.5.

Effect of label-specific entity tokens. To illustrate the effect of label-specific entity indicator, we introduce a variant of our entity-aware label attention mechanism, denoted as label-insensitive entity indicator. In this setting, the similarity score between labels and tokens is computed as:

$$v_{ki} = \begin{cases} w_k^T Q_e h_i & \text{if } \exists\ k, \text{s.t. } s_{ki} = 1. \\ w_k^T Q_{ne} h_i & \text{if for } \forall\ k,\ s_{ki} = 0. \end{cases}$$

In this variant, the model ignores the interactions between labels and entities. Meanwhile, other labels' entity tokens may bring some noise when computing the similarity score between those tokens and label l_k. Therefore, the performance of this variant is worse than simply removing the entity-aware mechanism on some metrics. Compare with label-insensitive variant, our label-specific setting improved macro F1 by 1.0, improved micro F1 by 1.0, improved precision@8 by 0.9 and improved precision@15 by 0.5.

5 Conclusion

In this paper, we proposed a novel entity-aware label attention model that incorporates entities into the label attention mechanism. To better leverage the information in ICD descriptions, we focus on entities instead of raw ICD description, as key tokens correlated with a code's presence often present in entities and raw ICD descriptions typically includes irrelevant information. We utilize an entity-aware label attention mechanism to select the most informative tokens for each label adaptively. The mechanism exploits the highly instructive and easily available labels' textual descriptions by introducing a label-specific entity indicator which enables the model to distinguish the entity tokens for each label explicitly and employing different query mechanisms to emphasize those entity tokens during label attention process. Evaluation results show that our model achieves state-of-the-art performance on the widely used MIMIC datasets.

References

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Alex Bottle and Paul Aylin. 2008. Intelligent information: a national system for monitoring clinical performance. *Health services research*, 43(1p1):10–31.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. 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.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318.

Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. 2010. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

S Hochreiter and J Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

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*, pages 289–297.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In *AAAI*, pages 8180–8187.

Luciano RS de Lima, Alberto HF Laender, and Berthier A Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 132–139.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Andrew Kachites McCallum. 1999. Multi-label text classification with a mixture model trained by em. In *AAAI 99 workshop on text learning*. Citeseer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Anthony N Nguyen, Donna Truran, Madonna Kemp, Bevan Koopman, David Conlan, John O'Dwyer, Ming Zhang, Sarvnaz Karimi, Hamed Hassanzadeh, Michael J Lawley, et al. 2018. Computer-assisted diagnostic coding: effectiveness of an nlp-based approach using snomed ct to icd-10 mappings. In *AMIA Annual Symposium Proceedings*, volume 2018, page 807. American Medical Informatics Association.

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.

World Health Organization et al. 1978. *International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index*. World Health Organization.

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

Frank B Rogers. 1963. Medical subject headings. *Bulletin of the Medical Library Association*, 51(1):114–116.

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

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. *arXiv preprint arXiv:2007.06351*.

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.

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*, pages 649–658.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Ming Zhu, Busra Celikkaya, Parminder Bhatia, and Chandan K Reddy. 2020. Latte: Latent type modeling for biomedical entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9757–9764.