计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：面向 ICD 疾病分类的深度学习方法研究
作者：张述睿，张伯政，张福鑫，杨万春
网络首发日期：2020-10-21
引用格式：张述睿，张伯政，张福鑫，杨万春. 面向 ICD 疾病分类的深度学习方法研究. 计算机工程与应用.
https://kns.cnki.net/kcms/detail/11.2127.TP.20201020.1719.016.html

网络首发时间：2020-10-21 09:52:28
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.TP.20201020.1719.016.html

面向 ICD 疾病分类的深度学习方法研究

张述睿[1,2]，张伯政[1,2]，张福鑫[2]，杨万春[3]

1. 中国人民大学 统计学院，北京 100872
2. 山东众阳健康科技集团有限公司，济南 250101
3. 山东交通学院 理学院，济南 250357

摘　要：国际疾病分类（ICD）是用于临床目的和健康管理的分类工具，是卫生统计数据的建立基础，在其庞大的分类体系中，含有与疾病健康问题和临床治疗相关的分类和对应的代码。针对在国际疾病分类的庞大标签空间中的多标签分类问题，提出一种端到端的深度学习方法。该方法首先采用改进的图注意力网络对标签空间进行建模，然后基于注意力重构的多标签分类器进行分类。在标签空间建模中，结合国际疾病分类中手术与操作分类的层次结构，构建出三种不同的图结构，利用图注意力网络将标签空间的结构信息融入到模型中，从而利用标签之间的依赖关系进行多标签文本分类。所提出的方法与实际应用场景有着紧密联系。实验表明，在临床国际疾病分类数据集上，相比于传统文本分类和其他标签空间建模方法，所提方法在分类性能上有明显的提升。

关键词： ICD 疾病分类；大标签空间；多标签；图注意力网络；深度学习；注意力重构
文献标志码:A　中图分类号:TP391　doi：10.3778/j.issn.1002-8331.2006-0032

张述睿，张伯政，张福鑫，等. 面向 ICD 疾病分类的深度学习方法研究.计算机工程与应用
ZHANG Shurui, ZHANG Bozheng, ZHANG Fuxin, et al. Towards ICD coding using deep learning approach. Computer Engineering and Applications

Towards ICD coding using deep learning approach

ZHANG Shurui[1,2], ZHANG Bozheng[1,2], ZHANG Fuxin[2], YANG Wanchun[3]

1. School of Statistics, Renmin University of China, Beijing 100872, China
2. Msunhealth, Jinan 250101, China
3. School of Sciences, Shandong Jiaotong University, Jinan 250357, China

Abstract: The International Classification of Diseases (ICD) is a classification system for health management and clinical purposes. This system is designed to map diagnoses, health conditions and therapeutic procedures to corresponding categories and assigning for these a designated code.Towards solving the multi-label classification problem in the fairly large label space of the ICD, an end-to-end deep learning approach is proposed. First the approachuses an improved graph attention network to model the label space, and then uses an attention-reconstruction based multi-label classifier for classification. During label space modeling, three different graph structures are constructed using the hierarchical structure of procedural codes in the ICD. The graph attention mechanism merges the structural information of the label space into the model to make use of label dependencies for multi-label classifica-

基金项目：山东省自然科学基金（No.ZR2017LF007）。
作者简介：张述睿(1988—),男,硕士研究生,研究领域为自然语言处理,E-mail: espresso_ml@hotmail.com;张伯政(1990-),男,工程师,研究领域为自然语言处理;张福鑫(1992-),男,工程师,研究领域为计算数学;杨万春(1982-),通信作者,男,博士研究生,副教授,研究领域为人工智能,分布式计算,E-mail: yangwch1982@126.com。

tion. The approachproposed is closely related to the actual application scenario. Experiments show that in clinical ICD dataset, the proposed method has a significant improvement in classification performance.

Key words: ICD coding; large label space; multi-label; graph attention network; deep learning; attention-reconstruction

1 引言

近年来，文本的多标签分类在自然语言处理领域中受到持续关注。在多标签分类中，一段文本可能同时具有多个对应的标签，例如一段文本可能表现出两种情感，一段新闻可能具有几种主题等，同时标签之间通常有一定的依赖关系。当标签空间变得非常庞大的时候，用深度学习模型进行文本分类就面临很大挑战。这主要是因为使用循环神经网络GRU[1]、LSTM[2][3]和多层感知机进行分类时，每个分类标签相互独立而无法建立标签之间的依赖关系。

在医疗临床中，电子病历的文本信息充斥着各种医学术语、且表述晦涩和含糊，因此对电子病历中的文本进行分类是很有挑战任务。同时由于很多深度学习模型的运作的原理都难以解释，导致了很多医学专业人士对此类方法的不信任。陈志豪等人[4]提出使用注意力机制来建立医疗问题和答案的相互影响，Mullenbach J 等人[5]采用卷积神经网络来提取病例文本中的语义信息，从而预测ICD疾病分类，但这两种方法并没有考虑到标签之间的依赖关系。檀何凤等人[6]提出了一种基于标签相关性的 K 近邻多标签分类算法，陈文实等人[7]提出使用主题模型与LSTM分别对文本的全局特征和局部特征建模，李峰等人[8]通过结合标签特征和局部标签相关性来丰富标签信息，但这三种方法没有考虑到在结构化标签空间或在超大标签空间中的应用。Steinberg E. 与 Liu P J[9]使用贝叶斯网络对标签的本体结构建模来预测 ICD 疾病分类，但是根据神经网络反向传播的特性，概率连乘的反传并不能让模型学到标签之间的依赖关系，Xie P 等人[10]提出使用树形 LSTM 的方法对结构化标签空间建模来预测 ICD 编码，但是树形LSTM递归的计算方式在超大标签空间中的计算代价非常高。图神经网络[11]是一类基于深度学习处理图域信息的方法，由于图神经网络有着很好的计算性能和可解释性[12]，因此其正受到越来越多的关注和应用。

本文结合图神经网络、图注意力网络[13]和注意力机制[14]，提出了一种解决超大标签空间中多标签分类问题的方法，并应用到国际疾病分类[15]中。本文提出的方法包括：（1）提出了一种面向 ICD 疾病分类的端到端深度学习方法。（2）采用改进的图注意力网络对标签空间建模，将标签空间的结构信息融入到模型中。（3）提出了一种基于注意力重构的编码匹配方法，其对大标签空间有着良好的适应性和可扩展性。通过与其他方法的实验对比，验证了本文所提方法的良好分类精度，在含有 3792 种标签的数据集中，前五命中的精确率达到了 0.92。

2 问题描述

2.1 ICD 手术分类

本文解决的问题是国际疾病分类中的手术与操作分类(ICD-9-CM3)[16]，以下简称为 ICD 手术分类。每一种 ICD 手术分类都对应着一条由数字和字母组成的编码，以下简称为 ICD 手术编码。ICD 手术分类是医院病案信息加工、检索、汇总、统计的主要工具，在医疗、研究、教学等方面发挥重要作用。ICD 手术分类是由专业编码员负责的，是一项非常繁琐的工作任务。编码员首先查阅医生录入的手术描述，之后如果有需要的话，还要查阅病人电子病历中的某些内容，然后人工查阅分类向导，将医生录入的手术描述匹配到一个或若干个最符合的 ICD 手术编码条目上。在临床中，医生录入的手术描述经常使用缩写和简称，这使手术描述的含义变得模糊，编码员经常因为这种情况犯一些主观错误。

2.2 ICD 手术分类数据集

数据集有典型的大标签空间和多标签的特点，并且标签空间拥有层级结构，标签与标签之间有相互关系和依赖。本文中使用的数据集是从 30 余家医院的临床数据中提取的，并经过人工精细校对，其中包含 60000 条数据，每条数据由医生录入的手术描述和编码员匹配好的 ICD 手术编码条目组成。医生录入的手术描述是对手术类型或方式的简短陈述，一条手术描述可能包含一个或多个手术操作，可对应一条或多条 ICD 手术编码条目。在实际的分

类工作中，编码员主要是靠手术描述来进行 ICD 手术分类，所以在本文中仅使用手术描述来进行 ICD 手术编码条目的匹配。

例 1 数据集展示

手术描述 1：右侧背部脂肪瘤切除术

对应的 ICD 手术编码条目：

86.3x03 皮下组织病损切除术

手术描述 2：C6/7 椎间盘微创消融术+盘内臭氧注射术

对应的 ICD 手术编码条目：

80.5900x001 椎间盘射频消融术

80.5200 椎间盘化学溶解术

如例 1 所示，每一条手术描述对应一个或多个最相关的 ICD 手术编码条目，其由一串数字和字母组成的编码和与之对应的标准编码描述组成，以下简称 ICD 手术编码条目中的标准编码描述为编码描述。

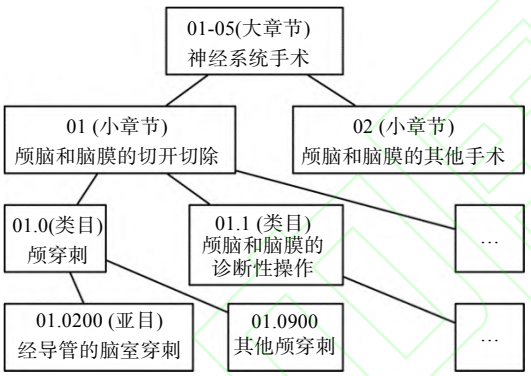


图 1　ICD 手术分类的层级结构

Fig.1　Hierarchy nature of ICD procedure coding system

图 1 所示的是 ICD 手术分类的层级结构，层级结构中每一个节点都是由一串数字和字母组成的编码和与之对应的编码描述组成的。ICD 手术分类含有 5 个层级，分别是大章节（18 个分类）、小章节（100 个分类）、类目（890 个分类）、亚目（3755 个分类）、细目（9100 个分类），共含有 13863 个分类。从大章节层级到细目层级，是一个不断细化分类的过程。ICD 手术分类的层级结构对临床 ICD 手术分类工作有重要意义，编码员在实际工作中，需要先根据手术描述中各种特征（如手术部位、针对的疾病、不同的手术方式等）通过查找目录确定一个粗略的分类寻找范围，也就是先确定章节、类目等较浅的层级的分类，再向下细分寻找。如图 1 所示，编码 01.0 颅穿刺指的是一个较大的分类范围，而它的下级节点 01.0200 和 01.0900 是指的不同的穿刺方式，是更细化的分类，上下级节点之间有着很强的依赖关系。为了充分利用标签空间的特性和标签之间的依赖关系，在本文中对标签空间进行了建模，并使用提出的方法将手术描述匹配到亚目层级和细目层级。

亚目层级有 3755 个分类，细目层级有 9100 个分类，但是在临床中，由于一些手术操作在临床中非常罕见，所以有很多分类并没有在数据集中出现，本文只对数据集中出现过的标签做分类研究，数据集中实际出现了 2237 个亚目分类，3792 个细目分类。

3　面向 ICD 手术分类的端到端模型

在本文中，将 ICD 手术分类问题看作多标签分类问题，提出了一种基于深度学习的端到端模型作为解决方案。首先使用图注意力网络对标签空间进行建模，之后使用注意力机制重构的方法使手术描述和 ICD 编码之间的语义信息得到交互，最后使用二元分类器得到对每一条 ICD 手术编码条目的预测结果，整体模型结构如图 2 所示。

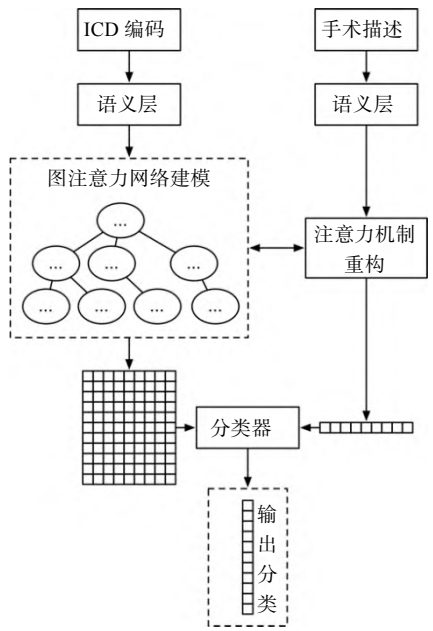


图 2　面向 ICD 手术分类的端到端模型

Fig.2　End-to-endmodel for automated ICD procedure coding

3.1　语义层

本文使用了 BiLSTM（双向长短期记忆网络）和 BERT[17]作为语义层的模型，对自然语言序列进

行向量化表达。transformer[18]模型提出使用多头注意力机制对序列进行建模，不再依赖序列的方向逐次计算，实现了对序列的并行化处理。基于transformer的模型因为其优秀的序列建模性能在自然语言处理领域被广泛应用[19]。BERT作为基于transformer的语言模型，提出先使用基础语料进行预训练，然后再适应下游任务。LSTM是被广泛使用的循环神经网络结构，其特点在于在序列建模中，将序列的每一时刻的信息依次输入到LSTM单元中，通过三种门控机制控制当前状态和记忆历史信息，并克服了RNN反向传播时梯度爆炸和梯度消失的问题，且KhandelwalU等人[20]的研究证实，LSTM对50字左右的序列有很好的辨识能力，而本文使用的数据集内医生录入的手术描述都比较简短，平均长度为13个字，而编码描述的平均长度为10个字，所以LSTM可很好地胜任当前的任务。

3.1.1 BiLSTM语义层

首先通过医学教科书语料预训练的字向量[21]（Word2vec）映射手术描述和ICD手术编码描述中的每一个字到一个向量，一段手术描述或一段ICD手术编码描述变成一串字向量序列。

使用 $\boldsymbol{x}^{text}$ 来表示一条医生录入的手术描述的字向量序列，使用 $\boldsymbol{x}_i^{code}$ 来表示第 i 条编码描述的字向量序列，$\boldsymbol{x}^{text}\in\mathbb{R}^{|text|\times d}$，$\boldsymbol{x}_i^{code}\in\mathbb{R}^{|code|\times d}$，其中 d 是词向量的维度，$|text|$ 是手术描述的字数，$|code|$ 是一条编码描述的字数。

使用BiLSTM对上下文语义信息进行建模，通过在字向量序列上应用BiLSTM模型，可以将自然语言序列中的每一个字向量，结合其上下文，形成适应于本文字序列的，隐含的语义表达，公式表示为：

$$\boldsymbol{q}=\mathrm{MLP}_0\left(\mathrm{BiLSTM}_0\left(\boldsymbol{x}^{text}\right)\right) \tag{1}$$

$$\boldsymbol{h}_i=\mathrm{MLP}_1\left(\mathrm{BiLSTM}_1\left(\boldsymbol{x}_i^{code}\right)[:-1]\right) \tag{2}$$

公式(1)、(2)中，由于BiLSTM通过正向和反向各得到一层隐状态的输出，所以隐状态的维度是 $2d$，MLP表示多层感知机层，将 $2d$ 的维度映射为 d。$[:-1]$ 指的是取BiLSTM返回的隐状态序列中最后一个时刻的隐状态。通过上面的操作，得到的 $\boldsymbol{q}$ 是经BiLSTM建模后的手术描述隐含语义的向量表达，$\boldsymbol{q}\in\mathbb{R}^{|text|\times d}$，$\boldsymbol{h}_i$ 是第 i 条编码描述的向量表达，$\boldsymbol{h}_i\in\mathbb{R}^d$。

3.1.2 BERT语义层

首先使用医学教科书语料对BERT模型进行预训练，之后使用BERT模型获取手术描述和手术编码的向量表达，公式表示为：

$$\boldsymbol{q}=\mathrm{BERT}_0\left(\boldsymbol{z}^{text}\right) \tag{3}$$

$$\boldsymbol{h}_i=\mathrm{BERT}_1\left(\boldsymbol{z}_i^{code}\right)[\mathrm{CLS}] \tag{4}$$

公式(3)、(4)中，$\boldsymbol{z}^{text}$ 表示一条手术描述的字索引，$\boldsymbol{z}_i^{code}$ 表示第 i 条编码描述的字索引。字索引是将文本中每个字用一个整数来表示。输入到BERT模型的字索引的构成为“[CLS]”+文本+“[SEP]”，“[CLS]”是一个特殊索引，BERT会在“[CLS]”索引的位置输出一个向量，用来表示整句的隐含语义信息。对于医生录入的手术描述，保留BERT输出的整个向量序列，$\boldsymbol{q}\in\mathbb{R}^{|text|\times d}$。对于第 i 条编码描述，只取“[CLS]”索引对应的向量，$\boldsymbol{h}_i\in\mathbb{R}^d$。

3.2 图注意力网络建模

ICD手术分类具有层级结构，分为5个层级，如图1所示，分别是大章节、小章节、类目、亚目、细目。从大章节到细目的层次，是不断细化分类的过程，层次与层次之间有着密切的关系。将ICD手术编码描述按层级结构融入模型中，使模型得到各层级之间的依赖关系，可增加模型的分类能力。

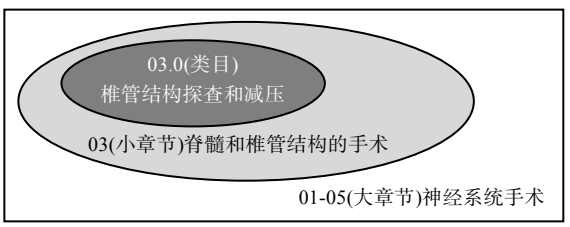


图3 ICD手术分类的上下层级节点关系

Fig.3 Dependency between nodes at different levelsofICD procedure coding system

例如上图中，“03.0（类目）椎管结构探查和减压”的上级节点有“03（小章节）脊髓和椎管结构的手术”和“01-05（大章节）神经系统手术”，把上级节点的关键信息，如“神经系统”、“脊髓和

椎管结构”融入到下级节点中，使下级节点含有 ICD 手术分类层次的上下文信息，这符合临床中编码员在分类过程中的思维方式。同样地，把下级节点的信息融入到上级节点，同样可以增强上级节点所含有的信息量，增加模型的判断能力。

本文将每一条 ICD 手术编码条目叫做节点，上级节点和下级节点之间用边连接，如图 1 中，01.0（类目）颅穿刺的下级节点的编码是 01.0200 和 01.0900，01.0200 的上级节点编码是 01.0。根据边的连接方向可使ICD手术分类的标签空间形成有向图或无向图结构，再使用图注意力网络对标签空间建模。本文提出在图注意力网络建模中构建以下三种不同的图结构。

***up* 图：**所有边由下级节点指向其上级节点，并包括每个节点自身形成的环边，形成由亚目到大章节方向的有向图结构。

***down* 图：**所有边由上级节点指向其下级节点，并包括每个节点自身形成的环边，形成由大章节到亚目方向的有向图结构。

***undirected* 图：**结合 *up* 图和 *down* 图形成不区分方向的无向图结构。

在完成上述三种图结构的构建之后，使用图注意力网络进行建模，即通过 3 种图结构分别进行注意力计算，最后将计算的结果级联。

首先进行图注意力权重的计算。

$$\alpha_{ij}=\frac{\exp(\text{leakyReLU}(\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{W}\boldsymbol{h}_i\oplus\boldsymbol{W}\boldsymbol{h}_j)))}{\sum_{k\in N_i}\exp(\text{leakyReLU}(\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{W}\boldsymbol{h}_i\oplus\boldsymbol{W}\boldsymbol{h}_k)))} \tag{5}$$

(5)式中，α_{ij} 是一个标量，表示第 i 个节点与第 j 个节点的注意力权重。$\boldsymbol{a}$ 是一条由可训练参数组成的向量，$\boldsymbol{a}\in\mathbb{R}^{2d}$。T 表示向量或矩阵转置，$\oplus$ 表示矩阵级联。N_i 表示与 i 节点相邻的且边的方向指向 i 的所有节点的集合，包括 i 节点自身。leakyReLU 是带泄露线性整流函数。$\boldsymbol{W}\in\mathbb{R}^{d'\times d}$ 是一个可训练的权重矩阵。所有节点使用相同的权重矩阵进行计算，用来获取更强的特征表达能力。

得到注意力权重后，通过相邻节点线性加权求和的方法，重构图结构中的每一个节点。在注意力重构过程中，为了使模型在高维度语义空间内捕捉到更丰富的特征，使用了文献[18]中提出的多头注意力机制，即把注意力重构的过程重复若干次，之后把所有的结果进行级联。

$$\boldsymbol{h}_i'=\Big\|_{m=1}^{M}\sum_{j\in N_i}\alpha_{ij}^{m}\boldsymbol{W}^{m}\boldsymbol{h}_j \tag{6}$$

公式(6)中，M 表示注意力机制头的数量，α_{ij}^{m} 表示第 m 个头的注意力权重，$\boldsymbol{W}^{m}$ 是第 m 个头的可训练权重矩阵，$\|$表示矩阵级联。在注意力机制重构的过程中，对每一个节点，都用与其相邻的节点重新表示，并且将 M 个头的计算结果进行级联，则最终 $\boldsymbol{h}_i'\in\mathbb{R}^{Md'}$。

分别使用 *up* 图、*down* 图和 *undirected* 图对标签空间建模，标签空间中的每个节点得到三种图结构的计算结果。如第 i 条编码描述得到的是 $\boldsymbol{h}_i^{up}$、$\boldsymbol{h}_i^{down}$ 和 $\boldsymbol{h}_i^{undirected}$，他们的维度都是 Md'。将它们级联得到最终的节点语义表达，之后通过多层感知机层将维度映射回 d'，如式(7)所示。

$$\boldsymbol{h}_i''=\text{MLP}_3(\boldsymbol{h}_i^{up}\oplus\boldsymbol{h}_i^{down}\oplus\boldsymbol{h}_i^{undirected}) \tag{7}$$

3.3 基于注意力重构的分类器

通过注意力机制使手术描述的隐藏表达 $\boldsymbol{q}$ 和编码描述的隐藏表达 $\boldsymbol{h}_i''$ 的语义信息进行融合，并用二元分类器得到对每一条ICD手术编码条目的预测结果，所有 ICD 手术编码条目的预测使用的是同一个二元分类器。

$$\beta_{ti}=\frac{\exp(\boldsymbol{q}_t\boldsymbol{h}_i'')}{\sum_{k=1}^{L}\exp(\boldsymbol{q}_k\boldsymbol{h}_i'')} \tag{8}$$

$$\hat{\boldsymbol{q}}_i=\sum_{t=1}^{L}\beta_{ti}\boldsymbol{q}_t \tag{9}$$

$$p_i=\text{MLP}_5\left(\text{MLP}_4\left(\hat{\boldsymbol{q}}_i\oplus\boldsymbol{h}_i''\right)\right) \tag{10}$$

公式(8)中 β_{ti} 为 $\boldsymbol{q}$ 中第 t 个字的隐藏表达向量与第 i 条编码描述的隐藏表达向量 $\boldsymbol{h}_i''$ 的注意力权重。公式(9)使用 β_{ti} 对 $\boldsymbol{q}_t$ 进行加权，即根据 $\boldsymbol{h}_i''$ 的语义信息对 $\boldsymbol{q}$ 进行线性加权求和，使 $\boldsymbol{q}$ 和 $\boldsymbol{h}_i''$ 的语义信息得到交互，得到的 $\hat{\boldsymbol{q}}_i$ 是根据第 i 条编码描述重构后的手术描述的隐藏表达向量，$\hat{\boldsymbol{q}}_i\in\mathbb{R}^{d'}$。公式(10)把根据 $\boldsymbol{h}_i''$ 重构后的 $\hat{\boldsymbol{q}}_i$ 与第 $\boldsymbol{h}_i''$ 进行级联，使他们之

间的信息再次得到交互，并通过两层多层感知机 MLP，将级联之后向量映射成一个标量，MLP_5 使用 sigmoid 函数进行非线性激活。p_i 是一个值介于 $(0,1)$ 之间的标量。假如 $\hat{\boldsymbol{q}}_i$ 和 $\boldsymbol{h}_i''$ 含有的语义信息比较相似，则他们之间的欧式距离比较近，p_i 输出的值会比较大。p_i 指的是当前手术描述是属于第 i 条 ICD 手术编码的概率。

通过对模型输出的注意力权重 β_{ti} 进行可视化，可发现模型从手术描述中筛选出匹配到一条 ICD 手术编码的线索，同时过滤掉与这条 ICD 手术编码相关程度较低的信息。例 2 中对注意力权重进行了可视化并在下方坐标轴标出注意力权重的值 β_{ti}，黄色代表较高注意力权重，绿色代表中等注意力权重，深紫色代表低注意力权重。图 4 和图 5 分别对单标签和多标签两种情况的注意力权重进行了可视化。

例 2 注意力权重可视化

手术描述 1：右侧硬膜外血肿钻孔引流术

01.0900 其他颅的穿刺

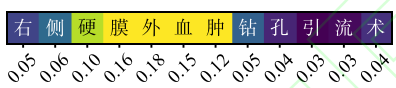
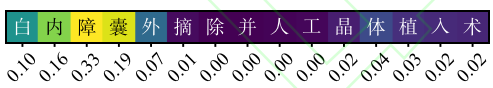


图 4　注意力权重的可视化（1）

Fig.4　Attention weights visualization(1)

手术描述 2：白内障囊外摘除并人工晶体植入术

13.5900 晶状体其他囊外摘出术



13.7000 置入人工晶状体

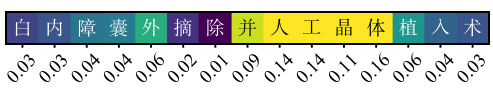
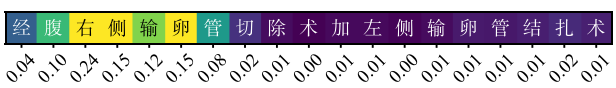


图 5　注意力权重的可视化（2）

Fig.5　Attention weights visualization(2)

手术描述 3：经腹右侧输卵管切除术加左侧输卵管结扎术

66.4x00 单侧输卵管全部切除术



66.9200 单侧输卵管破坏或闭合

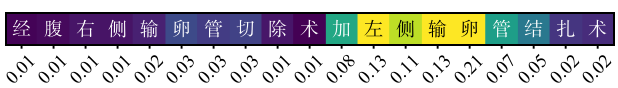


图 6　注意力权重的可视化（3）

Fig.6　Attention weights visualization(3)

例 2 中，手术描述 3“经腹右侧输卵管切除术加左侧输卵管结扎术”与 ICD 手术编码条目 66.4x00 和 66.9200 相匹配。图 6 上半部分展示的是该条手术描述与 ICD 手术编码条目 66.4x00 的注意力权重向量。通过观察可发现手术描述中与 ICD 手术编码条目 66.4x00“单侧输卵管全部切除术”相关的“经腹右侧输卵管切除术”所处部分的注意力权重较高，而其余不相关部分的注意力权重较低。经过统计，“经腹右侧输卵管切除术”的注意力权重之和约为 0.91，而“加左侧输卵管结扎术”的注意力权重之和约为 0.09。图 6 下半部分展示的是该条手术描述与 ICD 手术编码条目 66.9200 的注意力权重向量。可见“左侧输卵管结扎术”八个字所处部分的注意力权重较高，而与该编码关联度低的部分的注意力权重较低。经过统计，“经腹右侧输卵管切除术加”的注意力权重之和约为 0.25，而“左侧输卵管结扎术”这九个字的注意力权重之和约为 0.74。在使用手术描述对上述两个 ICD 编码条目进行分类的过程中，有关文字区域与无关文字区域的权重之比平均达到 4：1 以上。证明模型能够通过学习，明确手术描述中与每个 ICD 编码条目高度相关的部分，过滤低相关部分，提升了分类时中间数据的信噪比，且与人类的阅读方式和判断原则基本一致。

3.4　优化目标

在模型训练中，通过最小化交叉熵损失函数，可以得到模型的参数，损失函数的定义为：

$$\boldsymbol{L}=\sum_{i=1}^{N}\mathrm{CE}\left(p_i,y_i\right) \tag{11}$$

公式(11)中，CE 是交叉熵损失函数，y_i 是正确标记的分类，当手术描述对应的标记是第 i 条 ICD 编码时 $y_i=1$，其他情况下 $y_i=0$。N 是所有在标签空间中的 ICD 手术编码的数量，也就是标签的个数。

3.5　层次搜索

本文使用的数据集来自临床。在临床中，有些手术种类很普遍，也有一些手术种类非常少见，导致数据集中有很多标签类别出现在正样本中的次数非常少，所以数据集中标签类别的分布属于偏态长尾分布。文献[9]针对在大标签空间中且标签分布不均的情况，提出使用标签空间的本体结构对标签的概率进行因子分解，并应用到了 ICD 疾病分类中，小幅提升了分类性能，公式表达为：

$$\mathrm{P}\left(\boldsymbol{h}_i|\ \boldsymbol{x}\right)=\mathrm{P}\left(\boldsymbol{h}_i,\mathrm{ancestors}\left(\boldsymbol{h}_i\right)|\ \boldsymbol{x}\right)$$
$$=\prod_{\tilde{\boldsymbol{h}}_i\in\left\{\boldsymbol{h}_i\cup\mathrm{ancestors}\left(\boldsymbol{h}_i\right)\right\}}P\left(\tilde{\boldsymbol{h}}_i|\ \boldsymbol{x},\mathrm{parents}\left(\tilde{\boldsymbol{h}}_i\right)\right)\qquad(12)$$

公式(12)中，$\boldsymbol{h}_i$ 表示第 i 条手术编码条目，$\boldsymbol{x}$ 表示医生录入的手术描述文本，$\mathrm{ancestors}\left(\boldsymbol{h}_i\right)$ 表示 $\boldsymbol{h}_i$ 的所有祖先节点，$\mathrm{parents}\left(\tilde{\boldsymbol{h}}_i\right)$ 表示 $\tilde{\boldsymbol{h}}_i$ 的父级节点。公式(12)表示每一个分类条目的输出的概率，是由它本身的概率连乘它所有上级节点的概率得到。通过该方法，可以利用标签空间的结构让低频率标签和高频率标签之间的概率信息得到交互。本文在模型预测的过程中，从 ICD 手术分类层级结构的大章节向细目的方向迭代地搜索。每层只保留前 k 个概率最大的 ICD 手术编码条目，直到达到需要预测的层级。在层次搜索的过程中每一层只保留前 k 个概率最大的条目，这样就避免了低概率节点的无用计算，增加模型预测速度。

4 实验

在实验中，采用的数据集含有 60000 条数据，其中 70%作为训练集，10%作为验证集，20%作为测试集。在验证集上调整超参数，用测试集进行模型评估。在语义层使用了维度为 200 的 BiLSTM 和文献[17]中的 BERTBASE 模型。在图注意力网络建模的多头注意力机制部分，使用的注意力头的数量分别为 2、4、8、16。使用 Adam 优化器[22]对公式(11)进行优化，初始学习率被设置为 0.001。在语义层后使用 dropout 方法[23]防止模型过拟合，初始 dropout 率被设置为 0.3，在每个 *mini batch* 的计算过程中，只进行一次图注意力网络的正向传播计算，然后用这一次正向传播计算输出的编码描述的隐藏表达进行当前 *mini batch* 内的计算步骤。

4.1 评估指标

本文在实验中采用的评估指标分别是 $P@k$、*macro* $F1$ 和加权平均的 F1 值（*weighted* $F1$）。$P@k$ 又可称为 top k 精确率或 precision at k，$P@k$ 的定义为：

$$P@k=\frac{1}{k}\sum_{l=1}^{k}\mathbf{1}_{\boldsymbol{Y}}\left(rank\left(l\right)\right)\qquad(13)$$

公式(13)中，$\mathbf{1}$ 是指示函数，$\boldsymbol{Y}$ 是正确标注的 ICD 编码所对应的索引的集合，$rank\left(l\right)$ 是模型输出的第 l 个概率最大的 ICD 手术编码条目的索引。$P@k$ 评估指标指的是，模型预测的前 k 个概率最大的结果里面含有正确标注的标签的比例。

$F1$ 值通常在二分类问题上作为评估指标，在对本文实验结果的评估中，先把每种标签的分类结果看作一个二分类问题并求出 $F1$ 值：

$$F1=2\times\frac{precision\times recall}{precision+recall}\qquad(14)$$

公式(14)中 *precision* 是精确率，*recall* 是召回率。在多标签分类问题上，需要对所有标签的 $F1$ 值通过一些方法进行平均，本文中使用了 *macro* $F1$ 和加权平均的 $F1$ 值（*weighted* $F1$）这两种方法进行平均。*macro* $F1$ 的定义为：

$$macro\ F1=\frac{1}{N}\sum\nolimits_{s\in\boldsymbol{S}}F1\left(\hat{y}_s,y_s\right)\qquad(15)$$

公式(15)中，$\boldsymbol{S}$ 表示所有标签种类的集合，s 是其中的一种标签。N 是标签类别的数量，y_s 和 $\hat{y}_s$ 分别表示验证集中标记的标签为 s 的集合和模型对应的预测结果。*macro* $F1$ 的计算过程是求出所有标签类别的 $F1$ 值并直接求平均值，所以验证集中所有标签类别无论出现的次数多少获得的权重都是相同的。加权平均的 $F1$ 值（*weighted* $F1$）的定义为：

$$weighted\ F1=\frac{1}{M}\sum\nolimits_{s\in\boldsymbol{S}}|y_s|F1\left(\hat{y}_s,y_s\right)\qquad(16)$$

公式(16)中，M 为验证集的总数量，$|y_s|$ 表示验证集中标记的标签类别为 s 的个数，在加权平均的 $F1$ 值的计算过程中，出现次数比较多的标签会获得更高权重。

4.2 实验结果

在对比实验中，对 ICD 手术分类亚目层级和细目层级的预测结果分别进行评估。所有的评估都是在划分出的 20%测试集上进行的，测试集含有 12000 条数据，共对 8 种方法进行实验并评估。在 ICD 手术分类中，亚目层级有 2237 个分类可用，细目层级有 3792 个分类可用。实验分为两个阶段，第一阶段所有方法都分别使用 BiLSTM 语义层和 BERTBASE 语义层进行实现，并设图注意力网络中注意力机制头的个数（以下简称 GAT 头数）为 4。在进行完第一阶段的实验和评估之后，在表现比较好的方法上调整 GAT 头数，分别以 GAT 头数为 2、

4、8、16 进行实验并进行评估。实验的方法包括：（1）noGAT，MLP 在这个尝试中直接移除图注意力网络（GAT）和注意力重构的过程，在语义层之后直接通过 MLP 做分类预测。（2）Steinberg E etal. 通过文献[9]中提出的，在语义层之后使用标签空间的本体结构对标签的概率进行因子分解做分类预测。（3）CAML，该方法是文献[5]中提出的，直接在字向量序列上使用卷积神经网络和注意力机制做分类预测，不包含标签空间建模和语义层。（4）GAT, MLP 只使用无向图的 GAT 对标签空间建模，并直接用 MLP 做分类预测。（5）GAT, Att 使用无向图的 GAT 对标签空间建模，保留手术描述注意力重构的过程。（6）treeLSTM, Att 通过文献[10]中的方法使用双向树形 LSTM 的方式对标签空间建模，并使用本文中提出的基于注意力重构的分类器进行分类预测（7）catGAT, Att 通过 3 种不同方向的图结构使用图注意力网络对标签空间进行建模，之后使用基于注意力重构的分类器进行分类预测。（8）catGAT, AttHS, 在 catGAT, ATT 方法的基础上，在模型预测时使用层次搜索的方法。

表 1　BiLSTM 语义层实验结果对比（*P@k*）

Table 1　Performance of BiLSTM semantic layer (*P@k*)

方法	语义层	GAT 头数	*P@5*		*P@10*	
			亚目	细目	亚目	细目
noGAT, MLP	BiLSTM	-	0.6487	0.6092	0.6689	0.6248
Steinberg E et al.	BiLSTM	-	0.6789	0.6319	0.6980	0.6502
GAT, MLP	BiLSTM	4	0.7432	0.6935	0.7683	0.7196
GAT, Att	BiLSTM	4	0.8693	0.8169	0.8902	0.8370
treeLSTM, Att	BiLSTM	-	0.9332	0.8918	0.9503	0.8991
catGAT, Att	BiLSTM	4	0.9599	0.9141	0.9723	0.9229
catGAT, AttHS	BiLSTM	4	**0.9677**	**0.9235**	**0.9813**	**0.9309**

表 2　其他语义层实验结果对比（*P@k*）

Table 2　Performance of other semantic layer (*P@k*)

方法	语义层	GAT 头数	*P@5*		*P@10*	
			亚目	细目	亚目	细目
noGAT, MLP	BERTBASE	-	0.6776	0.6264	0.6969	0.6430
Steinberg E et al.	BERTBASE	-	0.7097	0.6600	0.7299	0.6789
CAML	N/A	-	0.6900	0.6631	0.7192	0.6859
GAT, MLP	BERTBASE	4	0.7798	0.7209	0.8056	0.7449
GAT, Att	BERTBASE	4	0.8761	0.8194	0.8949	0.8367
treeLSTM, Att	BERTBASE	-	0.8788	0.8171	0.8945	0.8310
catGAT, Att	BERTBASE	4	0.9610	0.9179	0.9712	0.9288
catGAT, AttHS	BERTBASE	4	**0.9708**	**0.9259**	**0.9836**	**0.9350**

表 3　BiLSTM 语义层实验结果对比（*F*1 值）

Table 3　Performance of BiLSTM semantic layer (*F*1 score)

方法	语义层	GAT 头数	*macro F1*		*weighted F1*	
			亚目	细目	亚目	细目
noGAT, MLP	BiLSTM	-	0.2209	0.1598	0.5009	0.3980
Steinberg E et al.	BiLSTM	-	0.2598	0.2290	0.5333	0.4501
GAT, MLP	BiLSTM	4	0.4132	0.3763	0.6455	0.5876
GAT, Att	BiLSTM	4	0.4698	0.4318	0.7129	0.6201
treeLSTM, Att	BiLSTM	-	0.4891	0.4433	0.7423	0.6321
catGAT, Att	BiLSTM	4	0.5102	0.4609	0.7543	0.6488
catGAT, AttHS	BiLSTM	4	**0.5132**	**0.4676**	**0.7631**	**0.6525**

表 4　其他语义层实验结果对比（*F*1 值）

Table 4　Performance of other semantic layer (*F*1 score)

方法	语义层	GAT 头数	*macro F1*		*weighted F1*	
			亚目	细目	亚目	细目
noGAT, MLP	BERTBASE	-	0.2398	0.1748	0.5279	0.4107
Steinberg E et al.	BERTBASE	-	0.2798	0.2393	0.5566	0.4501
CAML	N/A	-	0.2741	0.2410	0.5498	0.4569
GAT, MLP	BERTBASE	4	0.4257	0.3842	0.6526	0.5927
GAT, Att	BERTBASE	4	0.4761	0.4333	0.7140	0.6129
treeLSTM, Att	BERTBASE	-	0.4534	0.4200	0.6910	0.5784
catGAT, Att	BERTBASE	4	0.5102	0.4683	0.7519	0.6512
catGAT, AttHS	BERTBASE	4	**0.5152**	**0.4739**	**0.7613**	**0.6599**

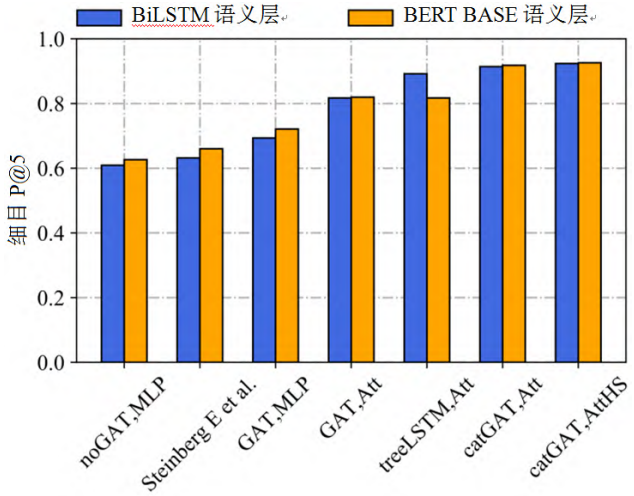


图 7　不同语义层细目 *P@5* 实验结果对比

Fig.7　Performance in terms of detailed entriesof different semantic layers (*P@5*)

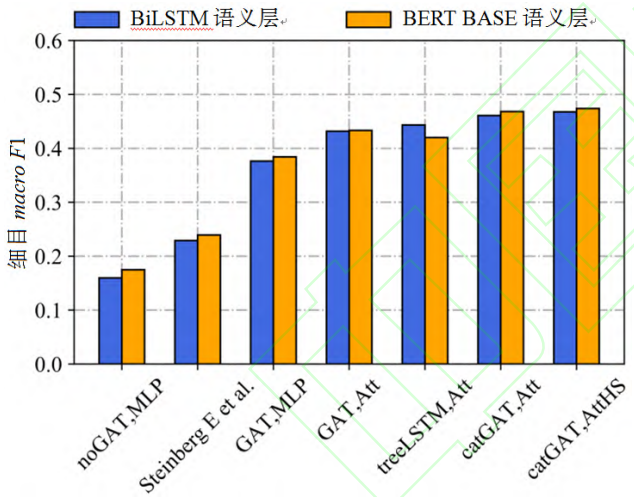


图 8 不同语义层细目 *macro F*1 实验结果对比

Fig. 8 Performance in terms of detailed entriesof different semantic layers (*F*1)

在实验中，除了不包含语义层的 CAML 方法，其他每种方法都使用两种语义层进行对比评估。通过 7 和图 8 可看出，BERTBASE 语义层在各项评估标准上的整体表现比 BiLSTM 语义层略微占优势，其中在没有标签空间建模和注意力重构的方法上优势较明显，但在本文提出的方法 catGAT, AttHS 上，使用 BERTBASE 语义层带来提升并不突出。从表 1、表 2 中可看出，使用图注意力网络（GAT）和基于注意力重构的分类器大幅提高了模型在 *P@k* 评估指标上的表现。在表 3、表 4 中，*macro F*1 值和加权平均的 *F*1 值（ *weighted F*1 ）差别比较大，这是由于数据集中标签类别的分布属于偏态分布，有些种类的手术在临床中很普遍，也有一些手术非常少见。没有使用标签空间建模的 noGAT, MLP, Steinberg E et al.和 CAML 方法的 *macro F*1 值比较低，说明标签空间建模对应对数据集标签分布不均衡的情况有重要意义。

表 5　GAT 注意力机制头数实验结果对比

Table 5　Comparison of the results of GAT attention mechanism head number experiment

方法	语义层	GAT 头数	细目 macroF1	细目 P@5
catGAT, AttHS	BiLSTM	2	0.4585	0.9190
	BiLSTM	4	0.4676	0.9235
	BiLSTM	8	0.4679	0.9225
	BiLSTM	16	0.4620	0.9210
	BERTBASE	2	0.4676	0.9200
	BERTBASE	4	0.4739	0.9259
	BERTBASE	8	**0.4759**	**0.9270**
	BERTBASE	16	0.4740	0.9252

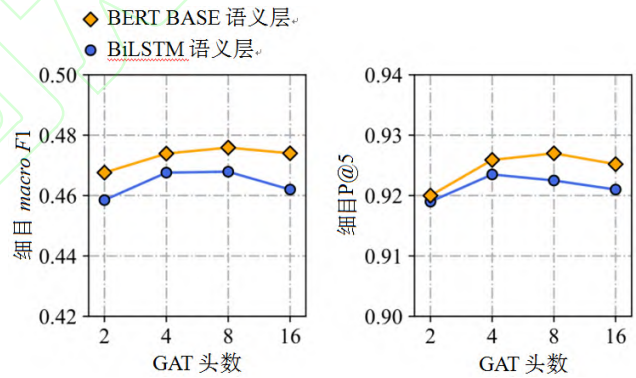


图 9　GAT 注意力机制头数实验结果对比

Fig. 9　Performance of different numbers of GAT attention heads

在进行完第一阶段的实验和评估之后，在表现最佳的方法 catGAT, AttHS 的基础上，调整 GAT 头数，并评估在细目 *macroF*1 和细目 *P@5* 上的表现差异。如表 5 和图 9 所示，将 GAT 头数调整为 2 之后，在上述两项评估指标上都有略微下降，GAT 头数调整为 8 之后，较 GAT 头数为 4 时的变化并不明显，而调整成 16 后在上述两相评估指标上都有一定程度下降，可见 GAT 头数的增大使得模型参数冗余并带来了不利影响。综上所述，本文提出的 catGAT, AttHS 方法，使用 BERTBASE 语义层，使用 8 个 GAT 头，在亚目和细目 *P@k* 、*macro F*1 值和加权平均的 *F*1 值（ *weighted F*1 ）这三种评估指

标上都取得了最佳的成绩。

5 结束语

本文提出了一种面向 ICD 手术分类的端到端深度学习方法。该方法主要采用了标签空间建模、注意力重构和层次搜索。在标签空间建模和层次搜索过程中，将标签空间的结构信息融入到模型当中。基于注意力重构的分类器能够在文本中寻找与标签相关的线索来进行分类。所提出的方法可应用于标签空间有本体结构或层级结构的数据集上，并且基于注意力重构的分类器在大标签空间内有很好的可扩展性。实验结果表明所提方法在亚目和细目层级的 ICD 手术分类上获得了较好分类表现。但是实验结果中 *macroF*1 评估指标表明，标签类别分布不均衡的问题对模型分类性能影响很大。在实际临床中，大量且高质量的标记数据获取代价很大，需要组织大量人力进行数据标注和校对。而且现实中很多情况下数据的分布往往是不均衡的。所以，怎样利用少量的数据样本在出现频率较低的标签类别上取得良好的分类效果是一个值得研究的问题，小样本学习和单样本学习是针对于此类问题的解决方法。除此之外，虽然高质量的标记数据获取代价很大，但大量的无标记的数据是比较容易获得的。近年来，半监督学习和自监督学习的研究让无标记数据得到利用。怎样将上述方法应用到国际疾病分类中，是下一步的研究方向。

参考文献:

[1] Chung J, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[C]//NIPS 2014 Deep Learning and Representation Learning Workshop, 2014.

[2] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[3] 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述[J]. 计算机应用, 2018, 38(S2):6-11+31. Yang Li, Wu Yuxi, Wang Junli, et al. Research on recurrent neural network[J]. Journal of Computer Applications, 2018, 38(S2):6-11+31.

[4] 陈志豪, 余翔, 刘子辰,等. 基于注意力和字嵌入的中文医疗问答匹配方法[J]. 计算机应用, 2019, 39(6): 1639-1645. Chen Zhihao, Yu Xiang, Liu Zichen, et al. Chinese medical question answer matching method based on attention mechanism and character embedding[J]. Journal of Computer Applications,2019, 39(6):1639-1645.

[5] Mullenbach J, Wiegreffe S, Duke J, et al. Explainable prediction of medical codes from clinical text[J]. arXiv preprint arXiv:1802.05695, 2018.

[6] 檀何凤, 刘政怡. 基于标签相关性的 K 近邻多标签分类方法[J]. 计算机应用, 2015(10):53-57.Tan Hefeng, Liu Zhengyi. Multi-label K nearest neighbor algorithm by exploiting label correlation[J]. Journal of Computer Applications, 2015(10):53-57.

[7] 陈文实, 刘心惠, 鲁明羽. 面向多标签文本分类的深度主题特征提取[J]. 模式识别与人工智能, 2019, 32(9): 785-792. Chen Wenshi, Liu Xinhui, Lu Mingyu. Feature extraction of deep topic model for multi-label classification[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(9):785-792.

[8] 李锋, 杨有龙. 基于标签特征和相关性的多标签分类算法[J]. 计算机工程与应用, 2019, 55(4):48-55. Li Feng, Yang Youlong. Multi-Label Classification Algorithm Based on Label-Specific Features and Label Correlation[J]. Computer Engineering and Applications. 2019, 55(4): 48-55.

[9] Steinberg E, Liu P J. Using Ontologies To Improve Performance In Massively Multi-label Prediction Models[J]. arXiv preprint arXiv:1905.12126, 2019.

[10] Xie P, Xing E. A neural architecture for automated icd coding[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1066-1076.

[11] Zhou J, Cui G, Zhang Z, et al. Graph neural networks: A review of methods and applications[J]. arXiv preprint arXiv:1812.08434, 2018.

[12] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[13] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.

[14] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[15] 北京协和医院世界卫生组织国际分类家族合作中心.疾病和有关健康问题的国际统计分类：第十次修订本[M]. 董景五,主译.第二版.北京:人民卫生出版社,2008.Beijing Union Medical College Hospital, World Health Organization. International Statistical Classification of Diseases and Related Health Problems: The Tenth Revision [M]. Dong

Jingwu. 2nded. Beijing: People's Medical Publishing House, 2008 .

[16] 刘爱民.国际疾病分类第九版临床修订本手术与操作 ICD-9-CM-3[M].2011 版.北京:人民军医出版社,2013. Liu Aimin. International Classification of Diseases Clinical Modification of 9th Revision Operations and Procedures ICD-9-CM-3[M].2011ed.Beijing:People's Military Medical Press,2013.

[17] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[19] Wolf T, Debut L, Sanh V, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing[J]. ArXiv, 2019: arXiv: 1910.03771.

[20] Khandelwal U, He H, Qi P, et al. Sharp nearby, fuzzy far away: How neural language models use context[J]. arXiv preprint arXiv:1805.04623, 2018.

[21] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[22] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[23] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.