智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

# 医学知识增强的肿瘤分期多任务学习模型

张恒， 何文玢， 何军， 焦增涛， 刘红岩

引用本文:
张恒, 何文玢, 何军, 等. 医学知识增强的肿瘤分期多任务学习模型[J]. 智能系统学报, 2021, 16(4): 739–745.
ZHANG Heng, HE Wenbin, HE Jun, et al. Multi–task tumor stage learning model with medical knowledge enhancement[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(4): 739–745.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202010005

**DOI**: 10.11992/tis.202010005
**网络出版地址**: https://kns.cnki.net/kcms/detail/23.1538.tp.20210330.1322.002.html

# 医学知识增强的肿瘤分期多任务学习模型

张恒[1], 何文玢[2], 何军[1], 焦增涛[2], 刘红岩[3]
（1. 中国人民大学 信息学院, 北京 100872; 2. 医渡云 (北京) 技术有限公司, 北京 100191; 3. 清华大学 管理科学与工程系, 北京 100084）

**摘　要:** 肿瘤分期是指从病人的电子病历文本中推测肿瘤对应阶段的过程。在电子病历数据中存在类别严重不均衡现象, 因此使用深度学习方法进行肿瘤分期具有一定的挑战性。该文提出医学知识增强的多任务学习 KEMT(knowledge enhanced multi-task) 模型, 将肿瘤分期问题视作面向医疗电子病历的文本分类任务, 同时引入医生在人工预测肿瘤分期时参考的医学属性, 提出基于医学问题的机器阅读理解任务, 对上述两种任务进行联合学习。我们与医疗机构合作构建了真实场景下的肿瘤分期的数据集, 实验结果显示, KEMT 模型可以将医学知识与神经网络结合起来, 预测准确率高于传统的文本分类模型。在数据分布不均衡的条件下, 在小样本类别上的准确率提升了 4.2 个百分点, 同时模型也具有一定的解释性。

**关键词:** 肿瘤分期; 文本分类; 机器阅读理解; 多任务学习; 不均衡分类; 智慧医疗; 知识表示; 注意力机制
**中图分类号:** TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2021)04−0739−07

**中文引用格式: 张恒, 何文玢, 何军, 等. 医学知识增强的肿瘤分期多任务学习模型 [J]. 智能系统学报, 2021, 16(4): 739–745.**
**英文引用格式: ZHANG Heng, HE Wenbin, HE Jun, et al. Multi-task tumor stage learning model with medical knowledge enhancement[J]. CAAI transactions on intelligent systems, 2021, 16(4): 739–745.**

# Multi-task tumor stage learning model with medical knowledge enhancement

ZHANG Heng[1], HE Wenbin[2], HE Jun[1], JIAO Zengtao[2], LIU Hongyan[3]
(1. School of Information, Renmin University of China, Beijing 100872, China; 2. Yidu Cloud (Beijing) Technology Co., Ltd, Beijing 100191, China; 3. Department of Management Science and Technology, Tsinghua University, Beijing 100084, China)

**Abstract**: Tumor staging is the process of inferring the corresponding stage of tumors based on patients' electronic health records (EHR). The serious uneven data distribution in the types of EHRs has certain challenges on tumor stage prediction through in-depth learning. Accordingly, this paper proposes a knowledge enhanced multi-task (KEMT) model and considers tumor stage reasoning as a text classification task of EHR. It also introduces medical attributes that doctors referred to in tumor stage prediction and introduces a medical problem-based machine reading comprehension task. The tasks are jointly studied by building a real-world dataset of tumor staging with medical institutions. Experimental results show that the KEMT model combines medical knowledge with a neural network and gets a higher precision rate of prediction than the traditional text classification models. Under the condition of uneven data distribution, the accuracy of small samples is improved by 4.2%, for which the model also accounts.

**Keywords**: tumor staging; text classification; machine reading comprehension; multi-task learning; unbalanced classification; smart healthcare; knowledge representation; attention mechanism

肿瘤分期是评价肿瘤生物学行为的最重要指标之一, 是根据个体内原发肿瘤数量以及扩散程度来描述肿瘤的严重程度和侵及范围的过程[1]。医院积累的电子病历文本 (EHR) 中蕴含了大量关于肿瘤的知识, 运用机器学习和自然语言处理技

**收稿日期:** 2020−10−09.　**网络出版日期:** 2021−03−31.
**基金项目:** 国家自然科学基金项目 (U171126, 71771131).
**通信作者:** 刘红岩. E-mail: hyliu@tsinghua.edu.cn.

术进行挖掘与知识提取，继而自动地给出分期诊断，是一项具有研究和实用价值的工作。目前肿瘤分期的过程尚依赖于医生的诊断经验或者一些专家手动编写的规则，流程复杂并且难以广泛应用。虽然神经网络模型已经被广泛地应用于各种互联网文本挖掘的任务中并且取得了很好的效果，但是在特定的医疗文本上处理肿瘤分期问题还没有合适的模型和方法。本文提出一种将深度学习与医学知识相结合的新方法，既借用了医疗大数据的优势，又弥补了传统神经网络缺乏医学知识的缺点。

## 1　肿瘤分期问题概述

肿瘤的 TNM 分期分为 *T*(tumor), *N*(Node), *M*(Metastasis)3 个维度，*T* 分期用来表征原发肿瘤的部位以及大小，*N* 分期判断局部淋巴结受累情况，*M* 分期是指远处转移情况。医生参考 *T*、*N*、*M* 分期的结果制定更有针对性的临床诊疗方案。本文采用由美国癌症联合委员会 (AJCC) 开发的第 8 版癌症 TNM 分期系统[2] 作为标准。如表 1 所示。

**表 1　第 8 版乳腺癌分期标准 (部分)**

**Table 1　8th edition of breast cancer staging criteria (part)**

| 分期 | 临床意义 |
|---|---|
| $T_0$ | 没有证据说明存在原发肿瘤 |
| $T_{is}$ | 早期肿瘤没有扩散至相邻组织 |
| $T_1$~$T_4$ | 大小或原发肿瘤的范围 |
| $N_0$ | 无区域淋巴结转移 |
| $N_1$ | 同侧腋窝淋巴结转移,可活动 |
| $N_2$ | 同侧腋窝淋巴结转移,固定或相互融合 |
| $N_3$ | 同侧锁骨下淋巴结转移伴或不伴腋窝淋巴结转移 |
| $M_0$ | 没有远处转移 |
| $M_1$ | 有远处转移 |

在现实场景中，不同分期的样本分布严重不均衡，以 *T* 分期为例，大多数样本集中在 $T_1$、$T_2$ 两类，占总量的 80% 以上，这给运用深度学习方法解决肿瘤分期问题带来了挑战。此外，不同于通用领域的文本分类，肿瘤分期任务依赖于从文本中进行一定的医学推理，需要相当的医学背景知识，而非仅仅靠上下文就能很好地解决。

Hu 等[3] 借助法律条文作为辅助信息，处理智慧司法中的罪名判定问题，受此启发，我们在本文中首次引入医生进行诊断时所参考的医学属性，并且将其是否能从文本中推断得到作为一种标注信息。这些特征包括是否侵犯胸壁、是否橘皮样变、是否侵犯腋窝、是否炎症型癌症等。这些标注信息与最终的分期结果存在内在的联系。

在此基础上，本文提出了一种多任务学习的机制，同时预测肿瘤分期结果以及上述医学属性的存在。我们提出了针对特定医学问题的机器阅读理解任务，并使用双向注意力机制生成问题的表示与电子病历文本的表示，融合两方面的表示推断最终的分期。这些问题可以为肿瘤分期提供额外的知识，更好地对样本不均衡的类别进行区分，也实现了不同肿瘤分期之间的知识迁移。

## 2　相关研究工作

### 2.1　文本分类

Kim 等[4] 提出 TextCNN 模型，借鉴图像识别中的卷积网络捕捉 *N*-gram 信息用于文本分类。Tang 等[5] 利用门限循环网络捕捉文本的序列特征，避免训练中的梯度爆炸问题。Joulin 等[6] 提出 FastText 模型，仅使用全联接层和 *N*-gram 特征就取得了很好的效果。Johnson 等[7] 提出 DPCNN 模型，提出深度堆叠的 CNN 模型可以提高单层卷积的效果，具有更强的表征能力。Yao 等[8] 提出一种基于图卷积的模型 TexGCN 利用词与文档的贡献信息对文本节点和单词节点构建图，将文本分类看作节点分类。Sun 等[9] 使用在预训练模型 BERT 的基础上进行微调用于文本分类任务。

上述研究均是通用领域的文本分类方法，采用的多是样本分布均匀的数据集。针对肿瘤分期问题的医疗文本数据集及研究较少。医疗文本普遍存在表述不规范、使用大量医学术语、难以进行语义理解等问题，增加了分类的难度。

### 2.2　不均衡分类

难度由于医疗电子病历数据的严重不均衡，直接应用深度学习模型效果不佳。不平衡分类问题在机器学习领域受到广泛关注，由此产生了小样本学习等研究领域。

不平衡分类的解决办法中，一种是数据层面的改进，采用过采样技术与欠采样技术对数据集进行平衡。通过复制样本或者消减样本达到总体平衡。另一种是从模型层面改进，通过引入外部知识，帮助神经网络对样本量较少的类别也能够很好地学习。本文主要探讨第 2 种。

Hu 等[3] 提出一个多任务学习的罪名预测模型，针对法律文书类别不均衡的问题，引入 10 个有判别作用的区分性属性 (盈利、死亡情节、暴力行为等) 作为判定罪名的中间依据，通过联合学习罪名预测任务与相关属性预测任务提升了预测

准确率。Elhoseiny 等[10] 提出引入类标签的文本描述在文本特征和视觉特征之间建立一种映射关系, 提升了小样本分类的效果。此类方法可以自动地学习标签或属性的向量表示, 但是这种向量只从各属性在文本中的贡献中学习得到, 对分类的增益较弱。

本文借鉴了上述思想, 引入医学属性对应的文本描述作为启发信息, 并将其作为问题进行机器阅读理解模型的训练, 模型学习的是多个具有实际意义的医学属性与文本的关系, 即将肿瘤分期拆解为对多个医学属性是否存在的判断, 相当于在文本与分期结果中引入了一层中间映射, 且增加了监督信息。即使是样本较少的类别, 也更加容易进行学习, 由此减弱了类别不均衡带来的影响 。

### 2.3　机器阅读理解

机器阅读理解技术是自然语言处理的重要研究领域, 其目标是给定一段文本, 给出答案或者指出答案的位置。本文借鉴机器阅读理解的思想, 将医学问题对应结果的预测视作一个多标签二分类问题。

Cui 等提出了双向注意力机制[11], 计算了问题−上下文 (Q2C) 和上下文−问题 (C2Q) 两个方向的注意力信息, 双向注意力机制为许多机器阅读理解模型所采用。

Seo 等[12] 在 BiDAF 模型中提出双向注意力流, 获取注意力矩阵以后, 没有把上下文和问题编码为固定大小的向量, 而是由后续的编码模块继续处理, 减少早期加权求和造成的信息损失。实验表明双向注意力对结果的提升尤为重要。本文将双向注意力引入肿瘤分期任务, 来捕捉上下文和问题间的关系, 并对注意力的形式做了改进。

## 3　医学知识增强的多任务学习肿瘤分期模型

### 3.1　肿瘤分期相关医学属性

本文选取了医生在推断肿瘤分期时重点观察的医学属性, 如表 2 所示, 这些医学属性与分期结果有一定的对应关系, 可以作为肿瘤分期的推断依据。本文针对每个医学属性定义“阅读理解问题”, 然后基于病历文本回答该问题, 即文本中是否蕴含了该属性及其相关特征, 结果要么为“是”, 要么为 “否”。所以本文将此任务转化为一个给定问题的机器阅读理解问题。

### 3.2　问题定义

肿瘤分期。给定一个电子病历文本, 记作序列 $D=\{w_1, w_2, \cdots, w_N\}$, 其中 $N$ 为文本的长度, $w_i$ 是文本的第 $i$ 个元素, 肿瘤分期任务的目标是根据 $D$ 推测其相应的分期结果 $y_T$、$y_N$、$y_M$, 且 $y_T \in \{T_{is}, T_1, T_2\}$, $y_N \in \{N_0, N_1, N_2, N_3\}$, $y_M \in \{M_0, M_1\}$。

**表 2　医学属性及对应“问题”描述 (部分)**

**Table 2　Description of medical attributes and corresponding “questions” (part)**

| 医学属性 | 问题描述 |
|---|---|
| 侵犯胸壁 | 存在直接侵犯胸壁(包括肋骨、肋间肌、前锯肌)等现象 |
| 橘皮样变 | 存在橘皮样变, 患侧乳房皮肤水肿 |
| 淋巴结转移 | 同侧腋窝淋巴结转移 |
| 侵犯皮肤 | 存在肿瘤侵犯皮肤 |
| 炎症型 | 肿瘤属于炎症型 |
| 原位癌症 | 原位癌, 早期肿瘤未侵及相邻组织 |
| 小型 | 微小浸润癌, 最大直径≤1 mm |
| 中型 | 20 mm < 肿瘤最大直径≤50 mm |
| 大型 | 肿瘤最大直径>50 mm |

机器阅读理解。将表 2 中的每种医学属性对应的问题描述当作问题, 设每个问题由 $M$ 个字符组成, 假设一共有 $K$ 个医学属性, 对应 $K$ 个问题任务目标是根据 $D$ 推测每个问题对应的答案 $p=\{p_1, p_2, \cdots, p_k\}$, 且有 $p_i \in \{0, 1\}$。

### 3.3　模型介绍

本文借鉴 Hu 等[3] 提出的 Attribute-based LSTM 和 Seo 等[12] 提出的双向注意力机制, 提出医学知识增强的多任务学习 (KEMT) 模型, 包括输入层、文本编码层、双向注意力层和输出层, 如图 1 所示。

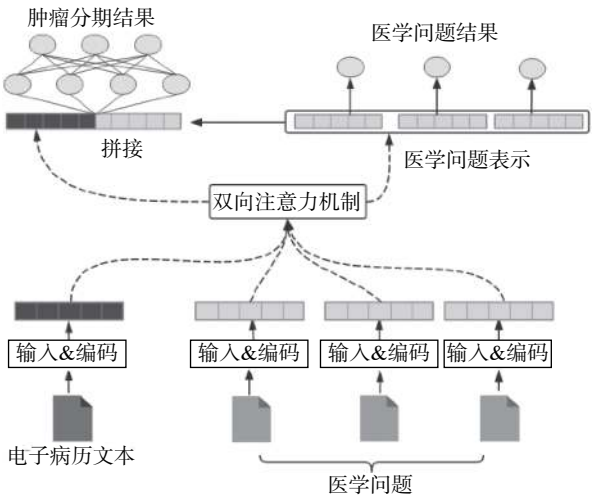


**图 1　模型结构**

**Fig. 1　Model structure**

输入层。负责将输入文本 $D$ 转化为向量序列。由于医疗文本切词复杂，模型的效果随切词粒度不同存在很大差异。本文使用字符级的表示，能更好地捕捉上下文语义，避免未登录词 (OOV) 现象。记 $E \in R^{|V|\times d}$ 为输入层字符嵌入矩阵，$|V|$为字典的大小，即所有病历文本中出现的不同字符数，$d$ 为输入层字符向量的维度，$N$ 为本段文本的字符数。

经过输入层后，输入文本转化为字符向量序列 $X=\{x_1, x_2, \cdots, x_N\}$。

编码层。对电子病历文本和问题文本进行分别编码，编码层结构如图 2 所示。

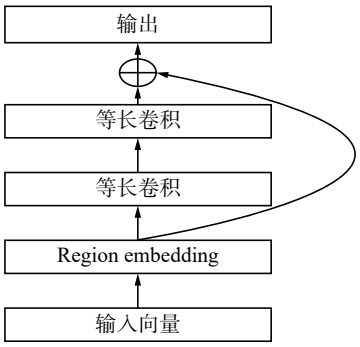


**图 2　编码层**

**Fig. 2　Encoder layer**

编码层中，模型借鉴 DPCNN[7] 中的 Region embedding 方式对输入文本片段进行嵌入表示，在后面的多层卷积中，使用两层等长卷积代替传统的窄卷积，使得每一个位置的向量都包含了上下文的信息。在卷积块的输入与输出间使用残差连接。

$$z' = z + f(z) \tag{1}$$

式中：$z$ 为输入卷积层的向量；$f$ 代表两层等长卷积；$z'$为卷积层的输出向量；编码层也可以采用其他自然语言处理模型，如 BERT，并不限定采用 CNN 模型，主要目的是提取文本的基本特征。

注意力层本文将病历文本经过编码后获得的表示记为 $\boldsymbol{C}$，且 $\boldsymbol{C} \in \mathbf{R}^{d\times N}$，$d$ 为向量的维度，$N$ 为病历文本的长度。每个问题经过编码后的表示记为 $\boldsymbol{Q} \in \mathbf{R}^{d\times M}$，$M$ 表示问题 $\boldsymbol{Q}$ 的长度。首先计算文本表示 $\boldsymbol{C}$ 与问题 $\boldsymbol{Q}$ 的注意力分数矩阵 $\boldsymbol{S}$，其第 $i$ 行第 $j$ 列的取值$S_{i,j}$如式 (2) 所示。

$$S_{i,j} = f(\boldsymbol{c}_i, \boldsymbol{q}_j) = W_0[\boldsymbol{q}_j; \boldsymbol{c}_i; \boldsymbol{q}_{j\odot c_i}] \tag{2}$$

式中：⊙表示逐元素相乘，且 $\boldsymbol{S} \in \mathbf{R}^{N\times M}$，$\boldsymbol{q}_j$ 和 $\boldsymbol{c}_i$ 分别表示问题描述的第 $j$ 个字符向量和病历文本的第 $i$ 个字符向量。$W_0$ 是一个可以训练的权重。

将病历文本看作回答问题的上下文信息，将 $\boldsymbol{S}$ 相似度矩阵每一行经过 softmax 层可以得到上下文−问题 (context-to-query) 方向的注意力，因为 $\boldsymbol{S}$ 中每一行表示的病历文本中第 $i$ 个字符与问题中每个字符间的相似度。将得到的 C2Q 注意力与 $\boldsymbol{Q}$ 做点积，如式 (3) 所示：

$$\boldsymbol{A} = \text{softmax}(\boldsymbol{S}, \text{axis} = \text{row}) \cdot \boldsymbol{Q}^{\mathrm{T}} \tag{3}$$

式中：$\boldsymbol{A}$ 为 $N\times d$ 的矩阵，即用 $\boldsymbol{Q}$ 中的所有词表示病历文本的每一个词。得到 $\boldsymbol{A}$ 以后与病历文本表示 $\boldsymbol{C}$ 进行拼接，得到融合问题信息的文本表示 $\bar{\boldsymbol{C}}_k$，$\bar{\boldsymbol{c}}$ 为 $\bar{\boldsymbol{C}}_k$ 的一行，如式 (4) 所示：

$$\bar{\boldsymbol{c}} = [\boldsymbol{c}; \boldsymbol{a}; \boldsymbol{c} \odot \boldsymbol{a}] \tag{4}$$

式中：$\boldsymbol{a}$ 为 $\boldsymbol{A}$ 的一行，将 $K$ 个 $\boldsymbol{Q}$ 分别经过注意力机制得到的向量表示做平均池化操作，得到最终的文本表示 $\bar{\boldsymbol{C}}$，如式 (5) 所示：

$$\bar{\boldsymbol{C}} = \text{avgpool}\left(\left[\bar{C}_1\ \bar{C}_2\ \cdots\ \bar{C}_K\right]\right) \tag{5}$$

将 $\boldsymbol{S}$ 相似度矩阵每一列经过 softmax 层可以得到 query-to-context (Q2C) 方向的注意力，计算的是对每一个问题中的词，文本中哪些词和它最相关，计算方法是取相似度矩阵中最大的一列，对其进行 softmax 归一化然后计算病历文本向量的加权和，如式 (6) 所示：

$$\bar{q} = \text{softmax}(\max(\boldsymbol{S}), \text{axis} = \text{col}) \cdot \boldsymbol{C}^{\mathrm{T}} \tag{6}$$

得到融合了上下文信息的问题表示 $\bar{q}$，且 $\bar{q} \in R^d$。$K$ 个 $\bar{q}$ 组成整体的问题表示为 $\bar{\boldsymbol{Q}} \in R^{K\times d}$。

输出层。通过全联接层和 sigmoid 函数，将 $\bar{\boldsymbol{Q}}$ 映射到对应的答案空间内，获得问题对应的答案，计算出问题所对应的医学属性是否在文中存在的概率，如式 (7) 所示：

$$p_i = \text{sigmoid}\left(W_i \bar{Q}_i + b_i\right) \tag{7}$$

式中：$p_i$ 是问题 $i$ 所对应的医学属性是否在文中存在的概率，本文把其视作一个二分类问题，$W_i$ 和 $b_i$ 是输出层的权重和偏置。

对于病历文本的表示 $\bar{\boldsymbol{C}}$，使用最大池化获取全局的表示$e = [e_1\ e_2\ \cdots\ e_d]$，其中 $d$ 为向量的维度。

$$e_i = \max\left(\bar{C}_{1,i}, \bar{C}_{2,i}, \cdots, \bar{C}_{N,i}\right),\quad \forall i \in [1, d] \tag{8}$$

为了更好地融合两部分信息，将 $\bar{\boldsymbol{Q}}$ 经过平均池化与 $\boldsymbol{e}$ 进行拼接，作为最终的全文表示表示输入给分类器，如式 (9)、(10) 所示。

$$r = \text{avgpool}\left(\bar{\boldsymbol{Q}}\right) \tag{9}$$

$$y = \text{softmax}(W_y[r; \boldsymbol{e}] + b_y) \tag{10}$$

这里 $r$ 是 $K$ 个问题向量的平均池化，$r$ 和 $\boldsymbol{e}$ 是采用拼接的方式输入给最终的预测层，$W_y$ 和 $b_y$ 是分类输出层的权重和偏置，$y$ 为最终在各个分期类别上的概率。

### 3.4　损失函数

本模型采用联合学习的方法，损失函数分为

两部分。一部分为肿瘤分期的预测概率与真实值之间的交叉熵损失 $\mathcal{L}_c$:

$$\mathcal{L}_c = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i) \tag{11}$$

式中: $y_i$ 代表肿瘤分期的真实结果; $\hat{y}_i$ 是网络预测得到的概率分布; $C$ 为对应肿瘤分期的种类数 ($T$ 分期为 5, $N$ 分期为 4, $M$ 分期为 2)。另外一部分, 对于第 $j$ 个问题的预测结果, 利用式 (12) 计算二分类交叉熵损失 $\mathcal{L}_{q,j}$:

$$\mathcal{L}_{q,j} = -(p_j \log(\hat{p}_j) + (1-p_j)\log(1-\hat{p}_j)) \tag{12}$$

$$\mathcal{L}_q = \frac{1}{K}\sum_{j=1}^{K} \mathcal{L}_{q,j} \tag{13}$$

$\mathcal{L}_q$ 为所有问题对应的损失加和。模型整体的损失函数由上述两个损失函数加和而成:

$$L = \mathcal{L}_c + \alpha \cdot \mathcal{L}_q \tag{14}$$

其中 $\alpha$ 是超参数, 用来平衡损失函数中两部分的比重。

## 4　实验设置及结果分析

### 4.1　数据集构建

目前尚未有公开的适用于肿瘤分期数据集, 于是我们与医疗 AI 公司医渡云合作构建了实验数据集, 主要来自医渡云医学专家基于临床经验撰写的部分病历内容, 包括病人的病理诊断, 现病史信息等。针对 $T$ 分期、$N$ 分期、$M$ 分期 3 种标准构建了 3 个数据集详情如表 3 所示。

**表 3　各数据集信息统计**

**Table 3　Statistics of data sets**

| 数据集 | 类数 | 训练/测试/总数 | 平均长度 | 最大长度 | 字典大小 |
|---|---|---|---|---|---|
| $T$分期 | 5 | 6542/1496/8038 | 406 | 2297 | 1826 |
| $N$分期 | 4 | 9224/2286/11510 | 395 | 2566 | 1883 |
| $M$分期 | 3 | 3819/4738 | 550 | 2560 | 1780 |

在搜集的肿瘤电子病历数据中, 具有显著的类别分布不均衡的现象, 以 $T$ 分期的数据集为例, 如表 4 所示共分为 5 类, 较高的 $T$ 下标值意味着更大的肿瘤和/或更广泛地扩散到附近的组织 ($T_{is}$ 指没有更深入地侵入其他组织的原位癌, $T_{is}$ 是 Tissue 的缩写)。可以看到 $T_1$、$T_2$ 类别的样本较多, $T_3$、$T_4$、$T_{is}$ 样本较少。所以我们在预处理阶段使用上采样的方法, 复制样本数较少类别的样本, 使各类别的样本数均与样本数最多的种类一致。

### 4.2　评价指标与基准模型

本文采用文本分类中常用的精确率 (Precision), 召回率 (Recall), $F_1$ 值作为模型评价指标。

本文选取多种经典的文本分类模型作为基准模型, 分别是:

TextCNN: Kim 等[4] 提出的 TextCNN;

BLSTM: 双向的 LSTM 加 max-pooling;

FastText: Joulin 等[6] 提出的浅层模型;

DPCNN: Johnson 等[7] 提出的多层卷积网络。

**表 4　$T$ 分期数据分布**

**Table 4　Data distribution of $T$ stage**

| 类别 | 数量 | 百分比/% |
|---|---|---|
| $T_1$ | 2597 | 32 |
| $T_2$ | 4245 | 53 |
| $T_3$ | 493 | 6 |
| $T_4$ | 294 | 4 |
| $T_{is}$ | 409 | 5 |

### 4.3　实验参数设置

本文使用 PyTorch[13] 实现了所有的模型, 设置最大训练轮次为 100 轮。使用 Adam[14] 作为模型优化算法, 初始学习率设置为 0.001, Dropout[15] 的大小设置为 0.5, batch 的大小设置为 64, 损失函数里的权重参数 $\alpha$ 设置为 0.5。输入向量的维度设置为 128 维, 采用标准正态分布随机初始化, 文本最大长度设置为 512。对基准模型中的 TextCNN 模型, 卷积核大小设置为 (3、4、5), BLSTM 的隐藏层大小设置为 128 维。

### 4.4　实验结果与分析

改进后的 KEMT 模型与上述基准模型对比如表 5 所示。

**表 5　$T$ 分期实验结果**

**Table 5　Results of $T$ stage experiment**　%

| 指标模型 | MP | MR | $M$-$F_1$ |
|---|---|---|---|
| FastText | 80.2 | 74.0 | 76.5 |
| TextCNN | 87.9 | 82.7 | 85.6 |
| BLSTM | 89.2 | 85.0 | 87.2 |
| DPCNN | 88.9 | 90.7 | 89.6 |
| KEMT | **94.7** | **91.8** | **93.1** |

从表 5 可以看出, 本文提出的 KEMT 模型的各指标均超过了基准模型, 比基准模型的最好结果分别提升了 5.8%、1.7%、3.5%。为了说明我们的模型在小样本类别上的有效性, 图 3 展示了各个类别上的效果对比。

如表 6 所示, KEMT 模型在 Macro-$F_1$ 值上超过了基准模型在小样本类别上的值, 显示出模型在样本数量极度不均匀的情况下, 对小样本类别

也有不错的分类效果。基准模型中 $F_1$ 值最大的为 $T_2$(93.8%), 最小值为 $T_4$(83.1%), 相差 10.7 个百分点, 而 KEMT 模型中 $F_1$ 最大值 $T_2$(95.2%) 和最小值 $T_4$(91%) 相差 4.2 个百分点。以上结果均显示出 KEMT 模型的效果在各类别上更均衡。

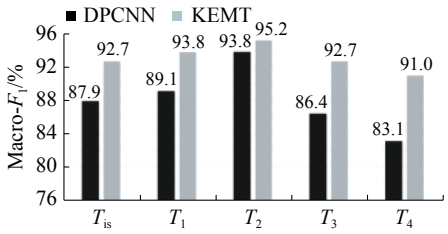


**图 3　KEMT 与 DPCNN 的 $F_1$ 对比**
**Fig. 3　$F_1$-score of KEMT and DPCNN**

**表 6　小样本类别 Macro-$F_1$**
**Table 6　Macro-$F$ of category**　%

| 类别(占比) | DPCNN | KEMT |
|---|---|---|
| $T_3$(6%) | 83.1 | **92.5** |
| $T_4$(4.7%) | 85.3 | **90.9** |
| $T_{is}$(2.3%) | 88.1 | **92.7** |
| 极差 | 10.7 | **4.2** |

为了说明模型的有效性, 接下来采用同样的方法对 $N$ 分期和 $M$ 分期数据集进行实验。实验结果如表 7 和表 8 显示, KEMT 模型在 $N$ 分期与 $M$ 分期标准下均取得了良好的效果。

**表 7　$N$ 分期实验结果**
**Table 7　Results of $N$ stage experiment**　%

| 指标模型 | MP | MR | $M$-$F_1$ |
|---|---|---|---|
| FastText | 86.5 | 84.9 | 85.7 |
| TextCNN | 89.3 | 88.1 | 88.6 |
| BLSTM | 87.9 | 88.4 | 88.2 |
| DPCNN | 91.9 | 91.0 | 91.4 |
| KEMT | **95.3** | **92.1** | **93.3** |

**表 8　$M$ 分期实验结果**
**Table 8　Results of $M$ stage experiment**　%

| 指标模型 | MP | MR | $M$-$F_1$ |
|---|---|---|---|
| FastText | 86.4 | 85.9 | 86.2 |
| TextCNN | 88.0 | 87.4 | 87.7 |
| BLSTM | 89.0 | 88.7 | 88.9 |
| DPCNN | 90.8 | 89.9 | 90.3 |
| KEMT | **93.7** | **92.8** | **93.1** |

### 4.5　有效性说明

为了说明注意力机制的有效性, 本文还设计了两组消融实验:

1)w/o attention, 即去掉模型中的注意力机制模块。则模型退化为将病历文本和问题分别编码。

2)w/o concatenation, 即保留双向注意力模块, 但直接用文本表示 $r$ 进行最终的分类。

从表 9 可以看到, 移除注意力模块以及医学领域知识后, 模型的 Macro-$F_1$($M$-$F_1$) 值分别下降了 5% 和 4%, 由此可见, 双向注意力机制和医学领域知识对于模型的效果是有显著影响的。

**表 9　注意力机制有效性**
**Table 9　Effectiveness of attention mechanism**

| 指标 | MP | MR | $M$-$F_1$ |
|---|---|---|---|
| KEMT | **94.7** | **91.8** | **93.1** |
| w/o attention | 90.6 | 86.5 | 88.3 |
| w/o concatenation | 91.3 | 87.7 | 89.2 |

### 4.6　样例阐释

本文选取了一个直观的样例, 来对于注意力机制如何帮助预测分期结果进行了说明。该样例的真实分期标签和 KEMT 模型预测的结果均为 $T_4$, 一个显著的特征是病人的电子病历中是否有隐含医学属性“橘皮样变”的出现。将 “橘皮样变”这个属性对应的注意力用热力图可视化出来。背景颜色越深的词, 具有的注意力权重值更大, 通过热力图显示, 可以清楚地看到, 注意力机制可以捕捉与医学属性相关的关键模式。如图 4 所示。


**图 4　注意力机制热力图**
**Fig. 4　Heat-map of attention mechanism**

## 5　结束语

本文充分利用医生诊断肿瘤分期时所依据的医学属性, 将属性对应的文本描述作为问题, 提出了面向医学问题的机器阅读理解任务和知识增强的多任务学习 (KEMT) 肿瘤分期模型, 实现了医学问题答案预测和肿瘤分期两种任务之间的知识迁移。实验结果表明该方法一定程度上解决了数据集不均衡带来的分类效果不佳的问题。

然而本文仍有需要改进的地方, 比如医生实际运用的知识更复杂, 本文对于分期的划分目前还是粗粒度的, 在每一种分期下还有更细粒度的划分, 如果要达到更精细的分类, 需要制定更精细的医学属性信息。

近来，图神经网络和预训练模型兴起，在多项任务中有巨大潜力，下一步我们也将探索这些新方法运用到肿瘤分期问题中，希望能够引入更多有效的医学知识，提升肿瘤分期问题的模型效果。

## 参考文献：


[1] 姚云峰. 肿瘤分期与疗效评价 [J]. 中国医学前沿杂志(电子版), 2010, 2(4): 70–75.
YAO Yunfeng. Evaluation of tumor stage and curative effect[J]. Chinese journal of the frontiers of medical science (electronic version), 2010, 2(4): 70–75.

[2] 周斌, 季科, 辛灵, 等. 美国肿瘤联合会乳腺癌分期系统(第 8 版) 更新内容介绍及解读 [J]. 中国实用外科杂志, 2017, 37(1): 10–14.
ZHOU Bin, JI Ke, XIN Ling, et al. Updates and interpretations of the 8th edition of AJCC breast cancer staging system[J]. Chinese journal of practical surgery, 2017, 37(1): 10–14.

[3] HU Zikun, LI Xiang, TU Cunchao, et al. Few-shot charge prediction with discriminative legal attributes[C]//Proceedings of the 27th International Conference on Computational Linguistics. New Mexico, USA, 2018: 487−498.

[4] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1746−1751.

[5] TANG Duyu, QIN Bing, LIU Ting. Document modeling with gated recurrent neural network for sentiment classification[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1422−1432.

[6] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017: 427−431.

[7] JOHNSON R, ZHANG Tong. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 562−570.

[8] YAO Liang, MAO Chengsheng, LUO Yuang. Graph convolutional networks for text classification[C]//Proceedings of 32rd AAAI Conference on Artificial Intelligence. Hawaii, USA: 7370−7377.

[9] SUN Chi, QIU Xipeng, XU Yige, et al. How to fine-tune BERT for text classification?[C]//Proceedings of the 18th China National Conference on Chinese Computational Linguistics. Kunming, China, 2019: 194−206.

[10] ELHOSEINY M, SALEH B, ELGAMMAL A. Write a classifier: zero-shot learning using purely textual descriptions[C]//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 2584−2591.

[11] CUI Yiming, CHEN Zhipeng, WEI Si, et al. Attention-over-attention neural networks for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 593−602.

[12] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[EB/OL]. (2016-11-05) [2019-10-12] https://arxiv.org/abs/1611.01603.

[13] PASZKE A, GROSS S, CHINTALA S, et al. Automatic differentiation in PyTorch[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, USA, 2017.

[14] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22) [2019-12-12] https://arxiv.org/pdf/1412.6980.pdf.

[15] 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.


## 作者简介：


张恒，硕士研究生，主要研究方向为自然语言处理，医疗数据挖掘。

何文玢，硕士研究生，主要研究方向为运动康复、医学数据分析、医学 AI 产品设计等。

刘红岩，教授，博士生导师，CCF 数据库专业委员会委员，主要研究方向为大数据管理与分析、数据/文本挖掘、商务智能、个性化推荐系统、医疗数据分析。发表学术论文近百篇，出版学术专著 2 部。