基于文本与图像的肺疾病研究与预测

吕 晴 [1] 赵 奎 [1] 曹吉龙 [2] 魏景峰 [3]

摘 要 通过对目前现有的肺癌检测技术研究, 发现大部分研究人员主要针对肺癌 (Computed tomography, CT) 影像进行研究, 忽略了电子病历所隐藏的肺癌信息, 本文提出一种基于图像与文本相结合的肺癌分类方法, 从现有的基于深度学习的肺癌图像分类出发, 引入了电子病历信息, 使用 Multi-head attention 以及 (Bi-directional long short-term memory, Bi-LSTM) 对文本建模. 实验结果证明, 将电子病历信息引入到图像分类模型之后, 对模型的性能有进一步的提升. 相对仅使用电子病历进行预测, 准确率提升了大约 14%, 精确率大约提升了 15%, 召回率提升了 14%. 相对仅使用肺癌 CT 影像来进行预测, 准确率提升了 3.2 %, 精确率提升了 4 %, 召回率提升了 4 %.

关键词 深度学习, 神经网络, 多头注意力机制, bi-LSTM, 肺癌

引用格式 吕晴, 赵奎, 曹吉龙, 魏景峰. 基于文本与图像的肺疾病研究与预测. 自动化学报, 2022, **48**(2): 531−538

DOI 10.16383/j.aas.c190645

Research and Prediction of Lung Diseases Based on Text and Images

LV Qing[1] ZHAO Kui[1] CAO Ji-Long[2] WEI Jing-Feng[3]

Abstract Through the study of the existing lung cancer detection technology, we found that most researchers mainly focus on the lung cancer (CT) images, ignoring the information of lung cancer hidden in the electronic medical records, this paper presents a lung cancer classification method based on the combination of image and text. Starting from the existing lung cancer image classification based on depth learning, the electronic medical record information is introduced, modeling text using Multi-head attention and (Bi-directional long short-term memory, BI-LSTM). The experimental results show that the performance of the image classification model is improved by introducing electronic medical record information. Predictions using only electronic medical records improved by about 14%, precision by about 15%, and recall by 14%. Compared to using only lung cancer CT images for prediction, the accuracy increased 3.2% , the precision increased 4% , and the recall increased 4%.

Key words Deep learning, neural network, multi-head attention, bi-LSTM, lung cancer

Citation Lv Qing, Zhao Kui, Cao Ji-Long, Wei Jing-Feng. Research and prediction of lung diseases based on text and images. *Acta Automatica Sinica*, 2022, **48**(2): 531−538

模态是指人接受信息的特定方式, 由于多媒体数据往往是多种信息的传递媒介, 多模态学习已逐渐发展为多媒体内容分析和理解的主要手段. 在医学领域, 也有研究者应用多模态学习. 针对 Alzheimer 病, 韩坤等[1] 提出结合磁共振图像 (Magnetic resonance imaging, MRI) 和正电子发射型计算机断层显像 (Positron emission computed tomography, PET) 图像模态的特征信息相融合的方法, 实验结果表明该方法在准确率上取得了较好的成绩. 为了解决传统模态医学图像缺陷, 张淑丽等[2] 提出了自由变形法对多模态的医学图像进行融合. 然而大多数研究人员主要融合多模态的医学图像, 没有加入电子病历等文本模态的数据. 调查发现, 肺癌是世界发病率和死亡率最高的疾病之一[3]. 病人在进行肺疾病诊断时, 需要 CT 检查, 影像科医生对 CT 影像进行检查描述, 但在实际的诊断和治疗过程中, 常常是由主治医生根据检查描述以及 CT 影像进行进一步的判断. 这一过场不仅增加了主治医生的工作量, 也导致了医疗资源的不合理应用.

基于此, 本文在影像 CT 基础上, 融入影像医生对 CT 影像描述的文本信息, 以及一些其他检验结果 (比如癌胚抗原测定、鳞状上皮细胞癌抗原测定等), 构建深度学习模型对肺疾病进行预测, 将影像医生给出的 CT 影像和检查描述以及其他检验结

收稿日期 2019-09-09 录用日期 2020-01-28
Manuscript received September 9, 2019; accepted January 28, 2020
国家水体污染控制与治理科技重大专项 (2012ZX07505004) 资助
Supported by National Science and Technology Major Project of Water Pollution Control and Treatment (2012ZX07505004)
本文责任编委 吴建鑫
Recommended by Associate Editor WU Jian-Xin
1. 中国科学院沈阳计算技术研究所 沈阳 110168 2. 中国医科大学附属第四医院 沈阳 110032 3. 辽宁省医疗器械检验检测院 沈阳 110000
1. Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168 2. The Fourth Affiliated Hospital of China Medical University, Shenyang 110032 3. Liaoning Provincial Medical Device Inspection and Testing Institute, Shenyang 110000

果输入到模型中, 对疾病进行判别并给出得病概率, 患病概率大的病人则交由主治医生更进一步地诊断和治疗, 以减轻主治医生的工作量, 提高工作效率.

1 数据预处理

本文收集的电子病历数据, 主要分为三部分: 检查描述、CT 影像和检验结果.

对检查描述研究发现, 虽出自不同医生之手, 但是对医学名词写法相同, 只是在电子病历输入的时候, 存在错别字、同音异字等问题. 如 "双肺实质未见明显异常密度, 双肺门不大, 纵膈内未见明确肿大淋巴结 … 肺癌不除外纵隔淋巴结增大, 肝脏内见斑片状高密度影, 门静脉周围间隙增宽." 数据中除了含有少见的医学专有名词 "纵隔淋巴结"、"斑片状高密度影"外, 还有错别字 "隔". 本文使用预定义词库的方法, 解决医学常见缩略语的分词问题, 然后使用 Multi-head attention 与 Bi-LSTM 对文本进行编码, 减少同音异字或者语法错误带来的文本理解上的问题.

CT 影像数据是通过成像设备进行采集的, 但是由于成像设备参数、外界环境的干扰, 会导致采集的 CT 图像数据有差异, 这些问题都会影响模型的准确率. 本文使用去噪和归一化等图像处理技术对 CT 图像进行处理.

其他检验结果主要是痰液细胞学、胸水检查、血常规检查和肿瘤标记物筛查等. 痰液与胸水细胞学检查, 主要是判断痰液与胸水中是否存在肿瘤细胞; 血常规检查包括白细胞、红细胞和血小板以及细胞酸碱性等; 肺癌筛选的肿瘤标记物主要有癌胚抗原 (Carcinoembryonic antigen, CEA)、癌抗原 CA125 (Cancer antigen 125, CA125)、细胞角蛋白 19 片段 (Cytokeratin fragment 19, CYFRA 21-1) 等.

考虑到数据由文本数据和图像数据两部分组成, 因此分别对两部分数据进行处理.

1.1 文本数据预处理

1.1.1 检查描述数据预处理

深度学习出现后, 基于神经网络的词嵌入模型成为了主流, GloVe[4] 使用词共现矩阵学习更广泛的共现概率. CoVe[5] 通过神经翻译的编码器向词嵌入中添加含有上下文背景的表征向量, 令模型学习上下文背景化的语义. BERT (Bidirectional encoder representation from transformers) 使用多层 Transformer[6] 编码器学习词汇前后的语义依赖关系, 并通过遮罩语言模型 (Masked language model, MLM) 解决了模型的输入在多层 Transformer 结构中可以看到自己的 "镜像问题". ERNIE[7] 提出了知识融合与对话语言模型的概念, 针对中文通用领域的自然语言处理任务对 BERT 进行了优化.

本文使用 jieba 分词, 考虑医学短文本中特有的专有名词、缩写语多的特点, 在分词过程中加入了医学词库, 医学词库的建立一方面是通过网络爬取医学专业词汇, 另一方面通过影像科医生总结出常见的肺部 CT 描述词汇. 文本数据中有大量的词虽然出现频率很高, 却对分类预测没有帮助, 比如在 "检查描述"中常出现 "无"、"可"、"检查"这类词在实际训练中不能体现不同病历差异性的作用, 更加重了学习器的负担, 一般称其为 "停用词". 因此在分词的时候, 需要将这些停用词去掉. 分词之后的文本数据还需向量化, 本论文使用 (Word to vector, word2vec) 模型来训练词向量, 并在模型中加入位置词向量与 Multi-head attention 来更好地表征文本语义.

1.1.2 检验结果数据预处理

检验结果主要是痰液细胞学、胸水检查、血常规检查和肿瘤标记物筛查等, 检验项目如表 1 所示, 电子病历中的检查结果会给出参考范围、检查名称、状态和结果值, 由于不同检查项目的量纲不同, 所以结果值有很大的差异, 因此, 本文使用状态值来作模型的输入, 将正常的状态映射为 0, 非正常状态 (高或低) 映射为 1, 然后输入到模型里面.

1.2 图像数据预处理

在计算机辅助诊断领域中, 主要针对肺部 CT 影像进行肺癌良恶性的诊断. Sun 等[8] 使用了单层的 CNN (Convolutional neural networks) 和 SDAE (Stacked denoised autoencoder) (3 个 DAE (DialAnExchange)) 以及 DBN (Deep belief nets) (4 层 RBM (Restricted Boltzmann machine)) 解决了肺节点的良恶性分类问题. Xiao 等[9] 增加了一个卷积层, 使用 CNN (2 个卷积层、2 个池化层、2 个全连接层) 和 DBN (2 层 RBM) 实现了肺节点的良恶性分类, 其效果有明显的提高. Cheng 等[10] 提出将肺节点兴趣区的多个参数与肺节点兴趣区一起输入到 SDAE 模型, 仅使用肺节点中间切片的 Single 模型与使用所有肺节点切片的 All 模型进行对比, 实验结果表明 All 模型相比 Single 模型, 在准确率上大约有 11 % 的提升, 而 AUC 大约有 5 % 的提升. Nibali 等[11] 将深度残差网络模型与迁移学习应用到肺癌分类中, 由于深度残差模型, 在加深网络深度的同时, 减少了梯度消失的可能, 因此, 通过深度残差网络模型以 ImageNet 图像集为源域进行迁

表 1 检验项目
Table 1 Examine items

	参考范围	检验名称	状态	结果值
血常规检查	0 ~ 0.1	嗜碱性粒细胞	正常	0.01
	0.05 ~ 0.5	嗜酸性粒细胞	正常	0.07
	0 ~ 1	嗜碱性粒细胞比率	正常	0.20 %
	110 ~ 160	血红蛋白	正常	128 g/L
	100 ~ 300	血小板	正常	135109̂/L
	3.5 ~ 5.5	红细胞	正常	4.25
	37 ~ 50	红细胞分布宽度	正常	43.90 %
	4 ~ 10	白细胞	正常	6.18109̂/L
	86 ~ 100	红细胞平均体积	正常	88.2 fL
痰液检查	无肿瘤细胞	痰液细胞	正常	无肿瘤细胞
肿瘤标记物	5 μg/ml	CEA (Carcinoembryonic antigen)	正常	2.31
	30 U/ml	CA125 (Cancer antigen 125)	正常	13.70 U/ml
	8.20 U/ml	CA72-4 (Cancer antigen 72-4)	正常	1.34 U/ml
	16.3 ng/ml	NSE (Neuron-specific enolase)	正常	15.18 ng/ml
	1.5 ng/ml	SCC (Squamous cell carcinoma)	正常	0.8 ng/ml
	2.0 ng/ml	CYFRA21-1 (Cytokeratin fragment 19)	高	7.31 ng/ml
胸水检验	0.38 ~ 2.1	甘油三脂	正常	0.74 mmol/L
	0.8 ~ 1.95	高密度脂蛋白	正常	1.31 mmol/L
	3.8 ~ 6.1	葡萄糖	高	10.11 mmol/L
	2 ~ 4	低密度脂蛋白	正常	2.02 mmol/L
	109 ~ 271	乳酸脱氢酶	正常	205.2 U/L
	0 ~ 6.8	直接胆红素	正常	3.49 μmol/L
	3.6 ~ 5.9	总胆固醇	低	3.54 mmol/L
	20 ~ 45	球蛋白	正常	31.7 g/L

移学习分类, 使得分类准确率为 89.9 %, AUC (Area under curve) 为 0.946. Shen 等[12] 提出了一种具有多级裁剪结构的 CNN 模型, 该模型可以获取不同尺度的图像特征, 从而加强模型的分类效果, 该模型的准确率为 87.1 %, AUC 为 0.93.

通过对已有方法对比发现, 分类准确率有明显的提高, 但是分类效果还不是很高. 一方面是由于模型过于简单, 另一方面, 没有根据目标数据进行有针对性的调整, 所以模型仍有更大的改进空间.

由于 CT 图像使用不同的扫描以及重建方法, 会产生一些不需要的杂质和噪点, 比如像结节一样的球状结构, 这些干扰信息与感兴趣区域之间存在某种相似性. 如果不去除噪声, 后面对特征提取的质量将受到严重影响, 从而影响模型的准确性. 本文实验分析发现高斯滤波器的去噪效果比均值滤波等的效果更好, 而且高斯滤波器对边缘信息的保留能力也更佳. 除此之外, 为了加快模型收敛, 将图像像素归一化或标准化, 在本文中, 对去噪之后的图像, 将像素的值归一化为 0 到 255 的整数. 处理后的图像采用残差神经网络为基础构建模型, 具体模型将在实验的图像模型部分给出.

2 实验

模型结构如图 1 所示, 整个模型的主要由三部分构成, 分别是文本部分、图像部分和多层感知器 (Multilayer perceptron, MLP), 文本部分输入的是电子病历的文本信息 (影像医生给出的 CT 描述信息), 图像部分输入的是影像检查的 CT 图像, 多层感知器输入的是其他检查结果. 将文本部分的输出、图像部分的输出和多层感知器的输出拼接起来, 然后经过全连接层, 最后输出结果. 模型的损失函数是交叉熵:

$$L = -\frac{1}{n}\sum[y\ln(a) + (1-y)\ln(1-a)] \qquad (1)$$

其中, a 是真实值, y 是预测值.

2.1 文本模型

在文本方面, 以 Bi-LSTM 和 Multi-head at-

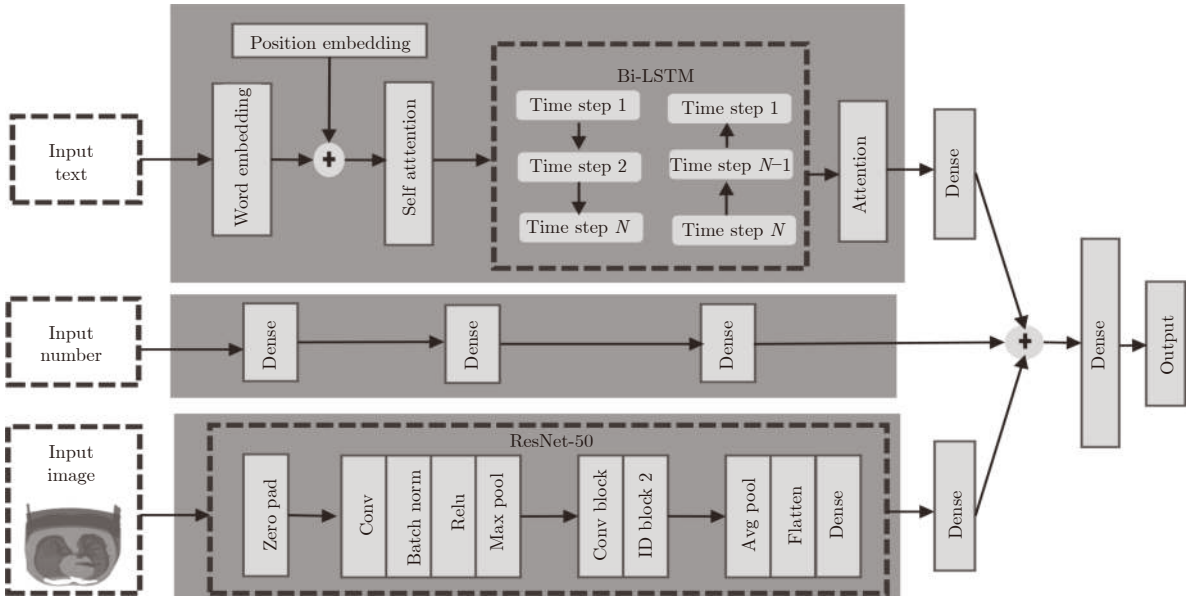

图 1 模型结构图

Fig. 1 Model structure

tention 为核心对文本建模, 模型的输入层为词向量加位置向量, 同时在模型的输入层后面引入 Multi-head attention. 最后将多个特征进行拼接和融合, 使模型进一步提高特征表达能力.

2.1.1 Word Embedding

本文使用词粒度的词向量. 考虑到文本语料相对比较少, 训练出来的词向量语义不够丰富, 而腾讯预训练词向量大约超过 800 万中文词汇数据, 与其他公开的预训练词向量相比, 具有比较好的覆盖性和新鲜度, 因此本文使用腾讯预训练向量.

由于病例中的词语所在的位置不同而代表不同的语义, 在词向量基础上, 加入位置向量, 能够使模型区别出不同位置的单词. 因此, 模型的输入也会将位置向量 (Position embedding) 作为辅助词向量输入. 在语言序列中, 相对位置至关重要, 而 Position embedding 本身是绝对值位置的信息, 因此, 本文将 Position embedding 定义为如下:

$$\begin{aligned}\boldsymbol{PE}_{2i}(p) &= \sin\left(\frac{p}{10\,000^{2i/d_{pos}}}\right)\\ \boldsymbol{PE}_{2i+1}(p) &= \cos\left(\frac{p}{10\,000^{2i/d_{pos}}}\right)\end{aligned} \tag{2}$$

$\boldsymbol{PE}$ 代表 Position embedding, p 代表词的位置, d_{pos} 代表维度, 公式将词位置信息使用三角函数映射到 d_{pos} 维度上.

2.1.2 Multi-head Attention

Multi-head attention 本质是进行多次 Self-attention 计算, 它可以使模型从不同表征子空间获取更多层面的特征, 从而使模型能够捕获句子更多的上下文信息.

Self-attention 本质是一种信息编码方式, 类似于 CNN 中的卷积, Self-attention 的定义如下所示:

$$\begin{aligned}&\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) =\\ &\text{softmax}\left(\begin{bmatrix}v_1\\ v_2\\ \vdots\\ v_n\end{bmatrix}\left[v_1^{\mathrm{T}}, v_2^{\mathrm{T}}, \cdots, v_n^{\mathrm{T}}\right]\right)\begin{bmatrix}v_1\\ v_2\\ \vdots\\ v_n\end{bmatrix} =\\ &\text{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})\boldsymbol{V}\end{aligned} \tag{3}$$

$\boldsymbol{Q}$ 是 Query, 代表 Query 向量, $\boldsymbol{K}$ 是 Key, 代表 Key 向量, $\boldsymbol{V}$ 是 Value, 代表 Value 向量. W_q 矩阵, W_k 矩阵和 W_v 矩阵将输入的词向量映射成 $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$, 然后按照公式进行加权求和, 对文本信息进行编码.

将 Self-attention 执行 k 次, 然后将结果拼接起来, 就得到了 Multi-head attention.

2.1.3 Bi-LSTM

词向量经过 Multi-head attention 的时候, 由于 Self-attention 是对输入信息的上下文的向量进行计算编码信息, 没有考虑到输入信息的词序, 所以, 在模型的输入层加入了 Position embedding, 除此之外, 还在 Multi-head attention 的后面加入了 Bi-LSTM. LSTM (Long short-term memory)[13] 是为了缓解 RNN 的梯度消失而提出的, LSTM 单元有三个门, 分别是遗忘门 $\boldsymbol{f}_t$, 输入门 $\boldsymbol{i}_t$ 和输出门 $\boldsymbol{o}_t$ [14]. 假设在 t 时刻, 输入为 $\boldsymbol{x}_t$, 而 $t-1$ (上一时刻) 的隐藏层的输出为 $\boldsymbol{h}_{t-1}$, 其中 $\boldsymbol{C}_{t-1}$ 为 $t-1$ (上

一时刻) 的细胞状态值, 则在 t 时 LSTM 的各个状态值:

$$\begin{aligned}\boldsymbol{f}_t &= \sigma\left(\boldsymbol{W}_f \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f\right)\\ \boldsymbol{i}_t &= \sigma\left(\boldsymbol{W}_i \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i\right)\\ \tilde{\boldsymbol{C}}_t &= \tanh\left(\boldsymbol{W}_C \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_C\right)\\ \boldsymbol{C}_t &= \boldsymbol{f}_t \times \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \times \tilde{\boldsymbol{C}}_t\\ \boldsymbol{o}_t &= \sigma\left(\boldsymbol{W}_o \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o\right)\\ \boldsymbol{h}_t &= \boldsymbol{o}_t \times \tanh\left(\boldsymbol{C}_t\right)\end{aligned} \tag{4}$$

通过以上计算, 最终得到 t 时刻 LSTM 隐层状态的输出值. 由于 LSTM 对句子只是从前向后单向建模, 无法进行从后向前的编码信息. 因此, 本文使用 Bi-LSTM (双向 LSTM), 可以更好地捕捉双向的语义信息.

2.1.4　Soft Attention

Soft attention 即传统的 Attention mechanism, 通过保留 Bi-LSTM 编码器对输入序列的中间输出结果, 然后计算每个中间结果与其他结果的点积, 最后加权求和.

$$\begin{aligned}\boldsymbol{M} &= \tanh(\boldsymbol{H})\\ \boldsymbol{\alpha} &= \mathrm{softmax}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}\right)\\ \boldsymbol{r} &= \boldsymbol{H}\boldsymbol{\alpha}^{\mathrm{T}}\end{aligned} \tag{5}$$

$\boldsymbol{H}$ 是 Bi-LSTM 隐藏层的输出结果, $\boldsymbol{w}$ 是需要学习的参数. 第二个 Attention 机制的实现是通过计算每个中间结果与其他结果的点积, 其中中间结果是通过保留 Bi-LSTM 编码器对输入序列的中间输出的结果, 最后再进行加权求和. 这一层的 Attention 能够观察到序列中的每个词与输入序列中一些词的对齐关系. 本文使用的是乘法注意力机制, 其中使用高度优化的矩阵乘法实现乘法注意力机制, 那么整体计算成本和单次注意力机制的计算成本并不会相差很大, 同时又提升了模型的特征表达能力.

2.2　多层感知机 (Multilayer Perceptron, MLP)

模型的第三部分是多层感知器 (MLP), MLP 主要包含输入层、隐藏层和输出层. 实验验证, 隐藏层不能过多, 一方面, 层数越多, 参数越多, 容易过拟合, 另一方面, 到了一定的层数, 增加更深的隐藏层, 分类效果也不会提升太多, 反而有时会下降. 因此, MLP 部分设置三个隐藏层, 具体参数如表 2 所示.

2.3　图像模型

本文的图像卷积部分在 ResNet-50 结构基础上, 基于 ImageNet 数据集预训练, 然后微调构建的模型. 模型的结构如图 2 所示, ResNet 中有 2 个基本的 block,一个是 Identity block, 输入和输出的 dimension 是一样的, 所以可以串联多个; 另一个是 ConvBlock, 输入和输出的 Dimension 是不一样的, 所以不能连续串联, 它的作用是为了改变特征向量的 Dimension.

表 2　MLP 参数设置

Table 2　The parameter of MLP

Name	节点个数	激活函数
Hidden1	65	Sigmoid
Hidden2	131	Sigmoid
Hidden3	263	Sigmoid

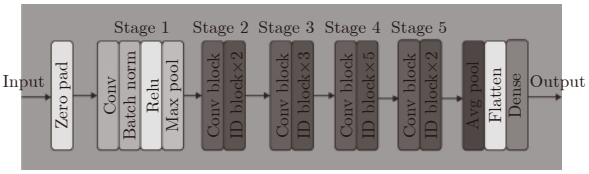

图 2　图像模型结构图

Fig. 2　Image model structure

图像中包含足够的区分信息是卷积神经网络能够学习不同肺癌特征的重要条件[15]. 图像的大小会影响网络区分不同特征的能力, 太小会使一些不明显的特征提取不到, 太大会受计算机内存的限制, 因此必须选择大小合适的图像尺寸, 由于本文使用的是 ResNet-50 (Residual neural network) 网络, 输入的图像尺寸需要调整为 224×224 .

2.4　实验设置

实验中所用的计算机硬件配置为 Centos 系统, CPU 为 Intel(R) Xeon(R) CPU E5-2630, GPU 为 NVIDIA Tesla M4 显卡, 深度学习框架为 Keras 2.2.4, 后端为 Tensorflow 1.13.

在本论文中, 主要有两个实验, 第一个是分别测试 Multi-head attention, Bi-LSTM 和 Soft attention 层在文本深度模型的效果, 第二个是测试文本深度模型、图像深度模型、MLP 和文本图像混合模型.

为了验证模型的优点和比较模型的表现能力, 在第二个实验中, 主要实现了以下几个模型: 一个基线模型为 ImageNet 预训练的 VGG-19 (Visual geometry group), 三个单模态模型为图像深度模型 (Img-net)、多层感知器 (MLP) 和文本深度模型 (Text-net), 以及多模态模型 Img+Text, Img+MLP 和 MLP+Text. Text-net 网络去掉下面的图像卷积部分, 添加一个全连接层, 损失函数为交叉熵的输出层. Img-net 网络去掉上面的文本深度模型, 添加

全连接层之后加上代价函数为交叉熵的输出层. MLP 是一个多层感知机网络, 只使用检查结果进行预测. TI-net 网络是文本图像混合模型, 输入为图像、文本和其他数值, 数据经过各自的模型之后, 拼接起来, 经过一个全连接层之后输出. 为了减少模型之间的扰动, 对于单模型 Text-net, Img-Net 和 MLP 三个网络分别用各自的输入进行预训练, 而对于多模态模型, 使用预训练的单模型的网络权重作为初始化, 再对多模态模型进行微调.

实验数据共有 3 785 个样本. 本文主要研究的是一个二分类问题, 即判断病人是否患有肺癌, 与一般分类问题不同, 疾病诊断分类问题的数据集往往存在不均衡问题, 因此需要对不均衡的样本进行处理. 由于本文的数据量比较大, 因此, 使用采样的方法来平衡数据集, 以 1:2 的比例对全量数据进行采样, 数据的比例分布如表 3 所示.

表 3 正负样本比例

Table 3 Positive and negative sample ratio

正样本	1 262
负样本	2 523

为了验证模型的效果, 将原始数据按照 8:2 的比例切分出训练集和验证集, 并将训练集在 3 个模型上进行训练, 然后在验证集上评价模型. 防止模型结果的偶然性, 在训练模型的时候, 采用 k-fold 交叉验证的形式来训练模型, 实验结果显示 k 取值为 7 的时候效果比较好一些. 训练集和验证集中, 文本的最大长度设置为 80, 词向量的维度为 200, 优化器为 Adam, 初始学习率为 0.01, 衰减因子为 0.0001, 训练轮次为 2 000 次, 为了防止过拟合, 使用 EarlyStopping 来提前停止训练, 评价指标采用准确率, 精确率和召回率.

2.5 实验结果

实验 1 的结果如表 4 所示, 主要用来测试 Multi-head attention, Bi-LSTM 和 Soft attention 层的效果, Text-net 网络使用了所有的层, Text-net1 去掉了 Multi-head attention 层, Text-net2 去掉了 Bi-LSTM 层, Text-net3 去掉了 Soft attention 层, 从表中结果可以看出, Text-net 模型比其他三个模型都要好. 对比 Text-net、Text-net1 和 Text-net2 可以看出, 加入 Multi-head attention 准确率提升了 7 %, 加入 Bi-LSTM 准确率提升了 3 %, 所以加入 Multi-head attention 层比 Bi-LSTM 层效果更好. 对比 Text-net 和 Text-net3, 加入 Soft-attention 层后, 模型准确率提升了 4 %, 这是因为 Bi-LSTM 层只对文本进行序列建模, 缺乏层次信息, 后面加入 Soft-attention, 可以将 Bi-LSTM 编码后的信息, 进行层次信息建模.

实验 2 的结果如表 5 所示, 从表 5 可以看出, 基线模型 VGG-19 的准确率为 92.53 %, 而 Img-Net (ResNet-50) 的准确率为 93.85 %, 从图像深度卷积方面来看, 显然 ResNet-50 模型的效果更好. 从单模态模型与多模态模型方面来说, 对比 Img-net、Img+Text、Img+MLP 和 TI-net 模型, 可以看出, 增加 CT 检验信息准确率提升了 1 %, 增加检验结果准确率提升了 2 %, 同时增加 CT 检验信息和检验结果, 准确率提升了 3.2 %, 精确率提升了 4 %, 召回率提升了 4 %. 从实验结果上可以看出, 基于多模态数据的模型效果优于单模型的效果, 并且对比单模型的结果可以看出, Img-net 效果远比 Text-net 和 MLP 的效果好, 这说明, CT 影像仍是肺癌诊断的主要信息, 而检查描述和检验结果作为补充信息加入到模型中, 可以很好地提升模型的精确度.

3 结论

本文提出了一种基于文本和图像的肺疾病分类算法, 详细介绍了本文提出的文本图像混合深度模型, 从基于深度学习的肺癌图像分类出发, 引入了 CT 影像描述信息和电子病历的检验项目, 并使用 Multi-head attention 以及 Bi-LSTM 对文本建模, 提取文本信息. 实验结果证明, 将文本信息和检验信息引入到模型后, 与传统单纯的图像模型相比, 本文提出的算法具有更好的识别效果和更强

表 4 实验 1 的结果

Table 4 The result of experiment 1

Model name	**Train (%)**			**Test (%)**		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Text-net	**83.12 ± 0.02**	**80.10 ± 0.05**	**81.12 ± 0.02**	**81.21 ± 0.01**	**79.82 ± 0.03**	**80.15 ± 0.01**
Text-net1	76.87 ± 0.02	75.29 ± 0.01	75.11 ± 0.03	74.91 ± 0.02	73.41 ± 0.02	74.07 ± 0.03
Text-net2	80.49 ± 0.03	78.16 ± 0.04	78.82 ± 0.03	78.43 ± 0.02	77.15 ± 0.01	78.59 ± 0.02
Text-net3	79.73 ± 0.02	77.19 ± 0.02	76.92 ± 0.01	78.19 ± 0.02	76.79 ± 0.03	75.57 ± 0.02

表 5 实验 2 的结果

Table 5 The result of experiment 2

Model Name	Train (%)			Test (%)		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
TI-Net	**97.08 ± 0.03**	**95.69 ± 0.01**	**94.37 ± 0.02**	**96.90 ± 0.04**	**95.17 ± 0.03**	**93.71 ± 0.01**
Img+MLP	95.15 ± 0.03	93.90 ± 0.02	93.17 ± 0.03	94.76 ± 0.02	92.89 ± 0.03	91.78 ± 0.01
Img+Text	94.71 ± 0.02	92.13 ± 0.03	91.26 ± 0.04	93.17 ± 0.04	90.88 ± 0.03	89.99 ± 0.03
MLP+Text	89.88 ± 0.04	87.67 ± 0.01	86.92 ± 0.02	87.78 ± 0.03	84.23 ± 0.03	84.57 ± 0.04
Img-Net	93.85 ± 0.03	91.84 ± 0.02	90.83 ± 0.03	92.67 ± 0.02	89.77 ± 0.03	88.93 ± 0.01
VGG-19	92.53 ± 0.02	89.16 ± 0.03	88.57 ± 0.01	90.94 ± 0.02	87.10 ± 0.03	87.04 ± 0.02
MLP	86.75 ± 0.03	85.21 ± 0.02	85.12 ± 0.03	84.86 ± 0.02	82.37 ± 0.03	81.59 ± 0.01
Text-Net	83.12 ± 0.04	80.10 ± 0.05	81.12 ± 0.02	81.21 ± 0.03	79.82 ± 0.03	80.15 ± 0.02

的泛化能力.

References

1 Han Kun, Pan Hai-Wei, Zhang Wei, Bian Xiao-Fei, Chen Chun-Ling, He Shu-Ning. Alzheimer′s disease classification method based on multimodal medical images. *Journal of Tsinghua University (Natural Science)*, 2020: 1−9
(韩坤, 潘海为, 张伟, 边晓菲, 陈春伶, 何舒宁. 基于多模态医学图像的Alzheimer病分类方法. 清华大学学报(自然科学版), 2020: 1−9)

2 Zhang Shu-Li, Li Jing-Yu, Mu Chuan-Bin, Liu Yanan, Meng Xin, Yang Dian. Free-form fusion method for multi-modal medical images. *Computer Programming Skills and Maintenance*, 2019, **8**: 139−140, 155
(张淑丽, 李靖宇, 穆传斌, 刘雅楠, 孟欣, 杨滇. 多模态医学图像的自由变形法融合策略. 电脑编程技巧与维护, 2019, **8**: 139−140, 155)

3 Tian Juan-Xiu, Liu Guo-Cai, Gu Shan-Shan, Ju Zhong-Jian, Liu Jin-Guang, Gu Dong-Dong. Deep learning in medical image analysis and its challenges. *Acta Automatica Sinica*, 2018, **44**(3): 401−424
(田娟秀, 刘国才, 谷珊珊, 鞠忠建, 刘劲光, 顾冬冬. 医学图像分析深度学习方法研究与挑战. 自动化学报, 2018, **44**(3): 401−424)

4 Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. 1532−1543

5 McCann B, Bradbury J, Xiong C, et al. Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems*, 2017: 6294−6305

6 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing System*, 2017: 5998−6008

7 Sun Y, Wang S, Li Y, et al. ERNIE: Enhanced representation through knowledge integration. arXiv preprint arXiv: 1904.09223, 2019

8 Sun W, Zheng B, Qian W. Computer aided lung cancer diagnosis with deep learning algorithms. *SPIE Medical Imaging*, 2016

9 Xiao Huan-Hui, Yuan Cheng-Lang, Feng Shi-Ting. Research progress of computer aided diagnosis in cancer based on deep learning. *International Journal of Medical Radiology*, 2019, **42**(1): 22−25

10 Cheng J Z, Ni D, Chou Y H, et al. Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans. *Scientific Reports*, 2016, **6**: 24454

11 Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. *International Journal of Computer Assisted Radiology and Surgery*, 2017, **12**: 1799−1808

12 Shen W, Zhou M, Yang F, et al. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 2017, **61**: 663−673

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

14 Chen Bin, Zhou Yong, Liu Bing. Event-triggered word extraction method based on convolutional long-term and short-term memory networks. *Computer Engineering*, 2019, **45**(1): 153−158
(陈斌, 周勇, 刘兵. 基于卷积长短期记忆网络的事件触发词抽取方法. 计算机工程, 2019, **45**(1): 153−158)

15 Litjens G., Sánchez C., Timofeeva, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep*, 2016, **6**: 2628

吕 晴 中国科学院沈阳计算技术研究所硕士研究生. 2017 年获得曲阜师范大学信息科学与工程专业学士学位. 主要研究方向为医学图像处理.
E-mail: lvqing17@mails.ucas.ac.cn
(**LV Qing** Master student at Shenyang Institute of Computing Technology, Chinese Academy of Sciences. She received her bachelor degree in information science and engineering from Qufu Normal University in 2017. Her main research interest is medical image processing.)

赵 奎 中国科学院沈阳计算技术研究所研究员. 2017 年获得中国科学院大学硕士学位. 主要研究方向为人工智能, 大数据, 物联网. 本文通信作者.
E-mail: zhaokui@sict.ac.cn
(**ZHAO Kui** Professor at Shenyang Institute of Computing Technology, Chinese Academy

of Sciences. He received his master degree from University of Chinese Academy of Sciences in 2017. His research interest covers artificial intelligence, big data, and the internet of things. Corresponding author of this paper.)

曹吉龙 中国医科大学附属第四医院信息中心主任. 2013 年获得东北大学硕士学位. 主要研究方向为医疗信息化, 医疗健康物联网, 医疗信息安全. E-mail: jlcao@cmu.edu.cn

(**CAO Ji-Long** Director at the Information Center, the Fourth Affiliated Hospital of China Medical University. He received his master degree from Northeastern University in 2013. His research interest covers hospital information, health internet of things, and medical information security.)

魏景峰 辽宁省医疗器械检验检测院高级工程师. 2011 年获得中国医科大学生物医学工程专业硕士学位. 主要研究方向为源医疗器械检验, 电磁兼容检测, 检测实验室质量体系管理. E-mail: 13898154351@163.com

(**WEI Jing-Feng** Senior engineer at Liaoning Medical Device Testi Institute. He received his master degree in biomedical engineering from China Medical University in 2011. His research interest covers medical electrical equipment test, electromagnetic compatibility test, and quality management of testing laboratories.)