王烤, 吴钦木 . 基于跨模态特征融合的肺病分类方法[J]. 智能计算机与应用,2025,15(7):56-60. DOI:10.20169/j.issn.2095-2163.250708

基于跨模态特征融合的肺病分类方法

王 烤, 吴钦木
(贵州大学 电气工程学院, 贵阳 550025)

摘 要: 目前大多使用卷积神经网络对单模态医学图像进行特征提取和决策,未能充分考虑跨模态和上下文的语义相关性,在处理特征时可能会受到冗余信息的影响,本文提出一种基于跨模态特征融合的肺病分类方法。首先,采用预训练模型 ResNet50 和 BERT 分别对医学图像和文本进行特征提取;其次,设计一个跨模态注意力机制模块对图像和文本的特征进行融合;最后,在 Bi-LSTM 的门控单元后引入稀疏门来减少融合特征的冗余信息,提高模型的计算效率及鲁棒性。实验结果表明,本文所提出模型的准确率、召回率、*AUC* 分别是 96.38%、96.22%、0.957,与先进的模型相比,准确率提升了 2.60%、召回率提升了 2.50%、*AUC* 提升了 1.33%,引入稀疏门改进后的模型测试平均推理时间减少了 60%。

关键词: 医学图像; 跨模态注意力机制; 稀疏门
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)07-0056-05

A cross-modal feature fusion-based approach to lung disease classification

WANG Kao, WU Qinmu
(School of Electrical Engineering, Guizhou University, Guiyang 550025, China)

Abstract: Most of the current use of convolutional neural networks for feature extraction and decision making of single-modality medical images fails to fully consider cross-modal and contextual semantic relevance, while ignoring the possibility of being affected by redundant information when processing features. In this paper, we propose a lung disease classification method based on cross-modal feature fusion. Firstly, we use the pre-trained models ResNet50 and BERT to extract features from medical images and text respectively, then we design a cross-modal attention mechanism module to fuse the features of images and text, and finally, we introduce sparse gates after the gating unit of Bi-LSTM to reduce the redundant information of the fused features and improve the computational efficiency and robustness of the model. computational efficiency and robustness of the model. The experimental results show that the accuracy, recall, and *AUC* of the proposed model are 96.38%, 96.22%, and 0.957, respectively, which are 2.60%, 2.50%, and 1.33% higher than those of the state-of-the-art model, and the test inference time of the model improved by introducing the sparse gate is reduced by 60%.

Key words: medical images; cross-modal attention mechanisms; sparse gate

0 引 言

X 射线照片是肺部疾病识别的低成本高效益的诊断工具,胸部 X 射线诊断需要高度熟练的放射科医师,而人工检测肺部疾病是一个耗时的过程,容易出现主观差异,可能会延误诊断和治疗。计算机辅助诊断在临床上具有很大潜力,可以在短时间内准确诊断肺部疾病,帮助医生减轻肺部疾病诊断的工作负担,降低了误诊率,大大提升了诊断的效率。深度学习因对自动特征提取和分类问题的普遍适用性而得到了广泛的研究[1]。基于卷积神经网络(CNN)的评估被广泛用于图像分类和对象检测[2]。

基金项目: 国家自然科学基金(51867006,52267003); 贵州省科学技术计划项目(黔科合支撑[2022] 一般 264)。
作者简介: 王 烤(1998—),男,硕士, 主要研究方向:控制理论与应用,深度学习,图像处理。
通信作者: 吴钦木(1975—),男,博士,教授,博士生导师,主要研究方向:控制理论与应用,运动控制,电动汽车传动控制等。Email:qmwu@gzu.edu.cn。
收稿日期: 2023-11-01

预先训练的神经网络模型，例如：AlexNet 模型、VGG 模型和 ResNet50 模型，通过 Softmax 分类器将所选的射线照片图像分类为正常和异常。Anthimopoulos 等[3]使用 CNN 分析医学图像，确定不同器官的疾病严重程度；Kawahara 等[4]研究局部信息和全局上下文信息的组合，并设计了不同尺度的图像分析架构；Setio 等[5]利用 3D CNN 来增强分类性能。虽然以上方法在处理医学图像时具有较好的效果，但未能考虑多模态的情况。

学者们已经提出了各种方法来处理多模态医学数据。例如，一些研究利用卷积神经网络（CNN）和递归神经网络（RNN），对医学图像和时间序列数据进行建模[6]。Tan 等[7]提出了一种在大数据时代应用的多模态医学图像融合算法。虽然通过结合多个不同模态的医学图像，利用深度学习算法将信息进行融合，可以提高诊断的准确性，但忽略了网络模型的诊断速度。Dastider 等[8]在卷积神经网络之后引入长短期记忆网络（Long Short-Term Memory，LSTM），对肺部疾病的严重程度进行预测，分类性能有了很大的提高；吕晴等[9]提出了一种基于图像与文本相结合的肺癌分类方法，使用多头注意力机制以及双向长短期记忆网络（Bidirectional Long Short-Term Memory，Bi-LSTM）对电子病历信息建模，进一步提升了医学图像分类模型的性能，但因其计算复杂度高，参数量较多，增加了计算负担；洪欣等[10]提出了基于 Bi-ConvLSTM 时序特征提取的阿尔兹海默症预测模型，通过时序卷积双向长短时记忆模型及注意力机制，在大脑影像的分层切面上进行时序特征提取。

尽管这些方法取得了一定的成果，但这些方法没有充分利用模态之间的关联信息，同时在处理大规模数据时存在计算效率低下的问题。本文引入跨模态注意力机制和稀疏门改进多模态医学数据分析。首先，通过引入注意力机制自适应地为不同模态数据分配权重，更好地捕捉到不同模态之间的相关性，从而提高多模态数据的融合效果；其次，将稀疏门应用于 Bi-LSTM 的门控单元中，过滤掉不重要的信息，有效地减少参数数量，从而减少模型的计算复杂度并提高模型的鲁棒性。

本文创新之处在于将跨模态注意力机制和稀疏门应用于多模态医学数据分析中，充分利用不同模态之间的关联信息，提高模型在处理大规模数据时的计算效率。相比于现有模型，本文的模型能够更好地挖掘多模态数据的潜在特征，从而提高模型的性能和应用效果。

1　基本原理

本文针对医学图像和医学报告进行特征提取和融合，提出了一种基于跨模态特征融合的肺病分类方法，即一种结合跨模态注意力机制与稀疏门的双向长短期记忆网络模型（A Bidirectional Long Short-Term Memory Network Model Combining Cross-Modal Attention Mechanism and Sparse Gate，Bi-LSTM-CMASG），以提高特征处理的准确性和效率。

首先，采用 ResNet50 模型对医学图像进行特征提取；其次，使用 BERT（Bidirectional Encoder Representations from Transformers，BERT）模型对文本进行特征提取，捕捉到文本数据中的上下文信息和语义关联，采用跨模态注意力机制将图像和文本的特征进行融合；最后，训练和测试融合后的特征。模型结构如图 1 所示。

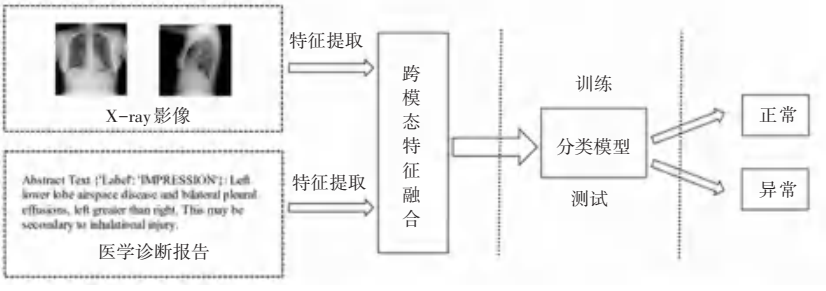

图 1　模型结构框图

Fig. 1　Block diagram of model structure

1.1　跨模态特征融合

本文采用基于跨模态注意力机制的特征融合方法，将图像与文本两种模态信息进行融合，公式如下：

$$\mathrm{CMA}(I,M)=\mathrm{Softmax}\left(\frac{I\cdot M^{\mathrm{T}}}{\sqrt{d}}\right)I \tag{1}$$

其中 CMA（Cross-Modal Attention）表示跨模态注意力机制；I 和 M 分别表示不同的模态；M 的作用

就是用来增强对 I 的表示；· 表示矩阵的点积；$1/\sqrt{d_l}$ 表示缩放因子。

跨模态注意力机制的融合涉及到多个模态（如图像和文本）之间的交互。一种常见的跨模态注意力机制融合方法是通过计算不同模态之间的注意力权重来实现，利用两个模态之间的相似得分作为权重来融合特征，$A_{i,j}$ 表示第 i 个模态对第 j 个模态的注意力权重，一个简单的跨模态注意力机制的融合，公式如下：

$$A_{i,j} = \frac{\exp(S_{i,j})}{\sum_{k=1}^{N} \exp(S_{i,j})} \tag{2}$$

其中，$S_{i,j}$ 是两个模态之间的相似度得分，N 表示模态的总数量。

即对于每个模态 i，通过计算其与其他模态 j 之间的相似度得分，然后将这些得分转换为概率分布，从而计算出 i 对 j 的注意力权重。这种跨模态注意力机制的融合方法可以帮助模型更好地利用不同模态之间的信息，提高模型的性能。

通过对图像和文本的特征进行提取，通过跨模态注意力机制将图像特征和文本特征进行跨模态的特征融合，最终得到较好的特征进行分类模型的训练和测试，提高模型分类效果。

1.2 稀疏门

稀疏门（Sparse Gate）是一种新型的门控机制，其主要作用是控制信息的流动，提高神经网络的表示能力和泛化能力。相对于传统的门控机制，稀疏门能够更有效地过滤噪声信息，提高网络的鲁棒性和可解释性，稀疏门是通过引入稀疏矩阵的方式来实现的。假设输入数据为 $X = [x_1, x_2, \cdots, x_n] \in R^{n \times d}$，其中，$n$ 表示序列长度，d 表示每个时间步的特征维度，则稀疏门可以表示如下：

$$g_{i,j} = \sigma(a_j(x_i) + b_j) \cdot \boldsymbol{s}_{i,j} \tag{3}$$

其中，a_j 是一个函数，用于提取输入 x_i 中的特征；b_j 是偏置项；σ 是 Sigmoid 函数；$\boldsymbol{s}_{i,j}$ 是一个二值化矩阵，表示稀疏性。

当 $\boldsymbol{s}_{i,j} = 1$ 时，表示第 i 个时间步的第 j 个特征参与运算；当 $\boldsymbol{s}_{i,j} = 0$ 时，表示第 i 个时间步的第 j 个特征被忽略。为了实现自适应的稀疏门，可以将 $\boldsymbol{s}_{i,j}$ 看作是概率变量，然后通过最大化模型的边际似然或最小化重构误差的方式来学习参数。

1.3 基于稀疏门的双向长短期记忆网络（SGBi-LSTM）

基于稀疏门的双向长短期记忆网络（SGBi-LSTM）是将 Bi-LSTM 与稀疏门相结合的网络。在稀疏门和 Bi-LSTM 结合的模型中，先使用 Bi-LSTM 对序列进行编码，再通过稀疏门对编码后的序列进行稀疏化。因为 Bi-LSTM 能够充分利用输入序列中的信息，对输入序列进行建模，从而得到序列数据。稀疏门则能够在保留序列重要信息的前提下，进一步减少序列中的冗余信息，从而提高模型的效率。在 Bi-LSTM 中的输入门、遗忘门和输出门之后分别加上一个稀疏门，可以控制模型中每个时间步的信息流动，进一步提高模型的效率。Bi-LSTM 是由前向和后向两个方向的 LSTM 单元组成，基于稀疏门的长短期记忆网络框架如图 2 所示。

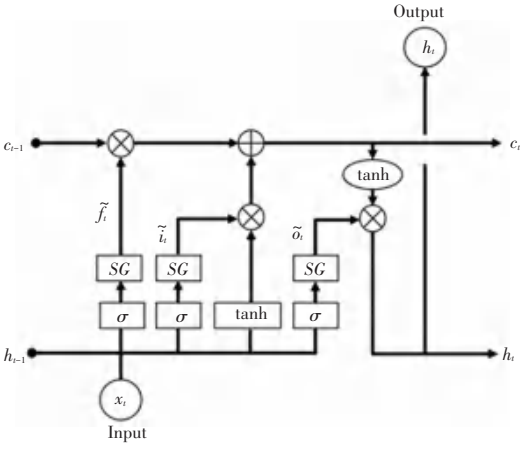

图 2 基于稀疏门的长短期记忆网络框架

Fig. 2 Framework of Long Short-Term Memory Network with Sparse Gate

在每个时间步 t，稀疏门的输入为 Bi-LSTM 在该时间步的输出 h_t，稀疏门的输出为一个二值向量 $g_t \in [0,1]$，表示是否对 h_t 进行稀疏化。稀疏门的输出 g_t 会与 Bi-LSTM 中的输入门、遗忘门和输出门的输出进行点积，从而得到稀疏化后的门的输出。设输入门、遗忘门和输出门在时间步 t 的输出分别为 i_t，f_t，o_t，则稀疏化后的门的输出公式如下：

$$\tilde{i}_t = i_t \odot g_t,\ \tilde{f}_t = f_t \odot g_t,\ \tilde{o}_t = i_t \odot g_t \tag{4}$$

其中，$\odot$ 表示点积。

在每个时间步中，稀疏门可以根据输入序列的特征，自适应地选择保留哪些位置的信息，从而提高模型的效率。

2 实验与分析

2.1 数据集及预处理

NLMCXR 数据集来源于美国国立卫生研究院（National Institutes of Health）的国家医学图书馆

(National Library of Medicine)，主要用于医学图像的自动诊断和疾病分类等研究[11]。该数据集包含了 7 470 张胸部 X 光片(包括正、侧位)和对应的 3 955 份诊断报告，使用 20%的样本作为测试集。

2.2　实验设置

本文使用 Python 语言在 GPU 加速环境下进行实验，采用 Pytorch 深度学习框架，电脑配置为 Windows11 系统、128 G 内存、RTX Geforce4090 24 G 显存。实验的参数设置见表 1。

表 1　实验参数设置

Table 1　Experimental parameters

参数	值
损失函数	交叉熵
优化器	Adam
学习率	0.000 1
批量大小(Batch_size)	256
轮次(Epoch)	200

2.3　实验结果与分析

本文采用准确率、召回率、*AUC* 作为肺病分类模型评价指标。为了说明本文所提出的模型的有效性和优越性，本文将进行基线模型对比和消融实验对比。

首先，在同一数据集上分别与传统的基线模型 ResNet50 以及现有 Inceptionv3+Bi-LSTM -Attention 模型[12]、LungNet22 模型[13]做对比实验，ROC 曲线图如图 3 所示，相关指标对比结果见表 2。从图 3 可以直观地看出本文提出的模型的 ROC 曲线明显优于其他基线模型和现有模型；从表 2 可见 ResNet50 模型的残差模块虽然能防止模型过拟合，但综合性能表现较差，准确率、召回率、*AUC* 均表现出最差的效果。Inceptionv3+Bi-LSTM-Attention 使用了 Bi-LSTM 以及注意力机制提升了模型的性能，但效果不明显，本文提出的模型使用了跨模态注意力机制以及引入稀疏门的 Bi-LSTM，提升了模型的总体性能，各评价指标均优于其他模型，准确率较 ResNet50、Inceptionv3 + Bi - LSTM - Attention、LungNet22 分别提升了 12.49%、4.24%、2.60%，召回率较 ResNet50、Inceptionv3+Bi-LSTM-Attention、LungNet22 分别提升了 11.21%、4.17%、2.50%。*AUC* 也有明显的提升，与最新的 LungNet22 的模型相比也提升了 1.33%，验证了本文模型的有效性及优越性。

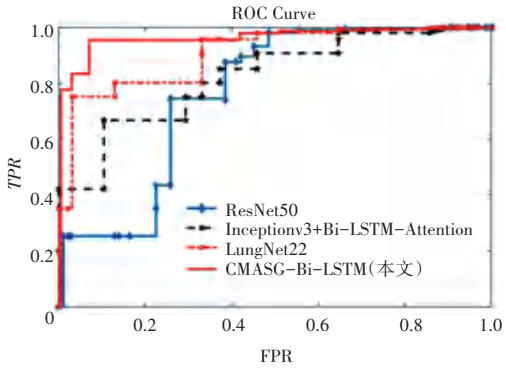

图 3　本模型与其他基线模型对比的 ROC 曲线图

Fig. 3　ROC Curve for comparison between the proposed model and baseline models

表 2　NLMCXR 数据集上本文算法与其他基线模型算法比较

Table 2　Comparison of proposed algorithm and other traditional model algorithms on Action NLMCXR dataset

模型	准确率/%	召回率/%	*AUC*
ResNet50	85.68	86.56	0.846 9
Inceptionv3+Bi-LSTM-Attention[12]	92.46	92.37	0.905 5
LungNet22 [13]	93.94	93.87	0.934 2
CMASG-Bi-LSTM (本文)	96.38	96.22	0.957 0

其次，在同一数据集上分别用 Bi-LSTM、CMA+Bi-LSTM、SG+Bi-LSTM、CMA+SG+Bi-LSTM 结合 CNN 特征提取器(ResNet50)和文本特征提取器 BERT 进行消融实验，实验结果见表 3。可以看出本文的模型在准确率和 *AUC* 均表现最优。模型 SG+Bi-LSTM 在 Bi-LSTM 的基础上引入了稀疏门，其测试平均推理时间(T_infer)降低了 60%，充分发挥了稀疏门的作用，降低了运算量，提升了运算速度；CMA +Bi-LSTM 在 Bi-LSTM 的基础上引入了跨模态注意力机制，充分融合了多模态的特征，明显提升了模型性能，准确率、召回率以及 *AUC* 较 Bi-LSTM 分别提升了 4.58%、4.19%、5.79%，证明本文模型引入跨模态注意力机制以及利用稀疏门改进后的 Bi-LSTM 的有效性。

表 3　本文的模型与其他消融模型比较结果

Table 3　Model in this paper was compared with other ablation models

模型	准确率/%	召回率/%	*AUC*	推理时间/s
Bi-LSTM	92.16	92.35	0.904 6	1.321
SG+Bi-LSTM	92.74	92.55	0.920 4	0.793
CMA+Bi-LSTM	95.86	96.14	0.946 2	1.684
CMA+SG+BiLSTM(本文)	96.38	96.22	0.957 0	1.095

3 结束语

本文提出了一种结合跨模态注意力机制与稀疏门的双向长短期记忆网络模型,有效地融合跨模态特征,并提高模型的性能和效率。首先,利用跨模态注意力机制动态地调整每个模态的权重,从而更好地利用不同模态的信息;其次,利用稀疏门减少冗余信息的影响,提高模型的计算效率及鲁棒性;最后,通过与传统模型 ResNet50 以及现有模型 Inceptionv3+Bi-LSTM-Attention、LungNet22 作对比实验以及消融实验,验证了本模型的有效性和优越性。此外,本文在提高模型解释性方面未进行充分探索,限制了对模型诊断过程的解释性分析,未来研究将提升模型的解释性和实用性。

参考文献

[1] WANG Y, CHEN Y, YANG N, et al. Classification of mice hepatic granuloma microscopic images based on a deep convolutional neural network[J]. Applied Soft Computing, 2019, 74: 40-50.

[2] AKCAY S, KUNDEGORSKI M E, WILLCOCKS C G, et al. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery [J]. IEEE Transactions on Information Forensics and Security, 2018, 13(9): 2203-2215.

[3] ANTHIMOPOULOS M, CHRISTODOULIDIS S, EBNER L, et al. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1207-1216.

[4] KAWAHARA J, BENTAIEB A, HAMARNEH G. Deep features to classify skin lesions[C]//Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). Piscataway, NJ: IEEE, 2016: 1397-1400.

[5] SETIO A A A, CIOMPI F, LITJENS G, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks [J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1160-1169.

[6] BEAM A L, KOMPA B, SCHMALTZ A, et al. Clinical concept embeddings learned from massive sources of multimodal medical data[J]. Pacific Symposium on Biocomputing, 2020, 25: 295.

[7] TAN W, TIWARI P, PANDEY H M, et al. Multimodal medical image fusion algorithm in the era of big data [J]. Neural Computing and Applications, 2020 (3): 1-21.

[8] DASTIDER A G, SADIK F, FATTAH S A. An integrated autoencoder-based hybrid CNN-LSTM model for COVID-19 severity prediction from lung ultrasound [J]. Computers in Biology and Medicine, 2021, 132: 104296.

[9] 吕晴,赵奎,曹吉龙,等. 基于文本与图像的肺疾病研究与预测[J]. 自动化学报,2022,48(2):531-538.

[10] 洪欣,黄铠洋,杨晨晖. 基于 Bi-ConvLSTM 时序特征提取的阿尔兹海默症预测 CTISS 模型[J]. 中国图象图形学报,2023,28(4):1146-1156.

[11] XUE Z, YANG F, RAJARAMAN S, et al. Cross dataset analysis of domain shift in CXR lung region detection[J]. Diagnostics, 2023,13(6):1068.

[12] 朱铭康,卢先领. 基于 Bi-LSTM-Attention 模型的人体行为识别算法[J]. 激光与光电子学进展,2019,56(15):153-161.

[13] SHAMRAT F M J M, AZAM S, KARIM A, et al. LungNet22: A fine-tuned model for multiclass classification and prediction of lung disease using X-ray images[J]. Journal of Personalized Medicine, 2022, 12(5): 680.