E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2024)07-1960-10

论文引用格式:Zeng S L, Li Z X, Zhang J X, Ding L F and Zhao C R. 2024. Chemical structure recognition method based on attention mechanism and encoder-decoder architecture. Journal of Image and Graphics, 29(07):1960-1969(曾水玲, 李昭贤, 张嘉雄, 丁龙飞, 赵才荣. 2024. 结合注意力机制和编码器—解码器架构的化学结构识别方法. 中国图象图形学报, 29(07):1960-1969)[DOI:10.11834/jig.230367]

结合注意力机制和编码器—解码器架构的化学结构识别方法

曾水玲[1,2*],李昭贤[1],张嘉雄[1],丁龙飞[1],赵才荣[3]

1. 吉首大学通信与电子工程学院,吉首　416000; 2. 南京理工大学江苏省社会安全图像与视频理解重点实验室,南京　210094; 3. 同济大学电子与信息工程学院,上海　201804

摘　要:**目的**　化学结构识别是化学和计算机视觉领域的一个重要问题,传统光学化学结构识别技术在复杂化学结构识别任务中易发生信息丢失或误识别的现象,同时又因为化学物质的结构多样性常导致其无法解析,识别效果不佳。而基于深度学习的模型通常具有网络结构复杂度高、上下文信息易丢失和识别率低的问题。为此,提出一种结合注意力机制和编码器—解码器架构的化学结构识别方法。**方法**　首先,使用改进的ResNet50(residual network)作为特征提取器抓取表征信息;其次,使用BLSTM(bi-directional long-short term memory)作为行编码器为ResNet50提取的表征信息加强空间信息;最后,使用去填充模块和基于覆盖注意力机制的LSTM(long short-term memory)网络作为模型解码器,对化学结构图像进行解码,将编码结果解码为SMILES(simplified molecular input line entry system)序列。**结果**　在Indigo、ChemDraw、CLEF(Conference and Labs of the Evaluation Forum)、JPO(Japanese Patent Office)、UOB(University of Birmingham)、USPTO(United States Patent and Trademark Office)、Staker、ACS(American Chemistry Society)、CASIA-CSDB(Institute of Automation of Chinese Academy of Sciences—Chemical Structure Database)和Mini CASIA-CSDB数据集上,所提方法识别准确率分别为71.1%、70.21%、45.8%、30.3%、53.02%、58.21%、43.39%、46.3%、84.42%和85.78%,高于SwimOCSR、Image2Mol和ChemPix模型得分。**结论**　与其他模型相比,本文方法通过少量训练集能够获得较高的识别准确率。

关键词:化学结构识别;编码器—解码器;注意力机制;残差网络;SMILES(simplified molecular input line entry system)

Chemical structure recognition method based on attention mechanism and encoder-decoder architecture

Zeng Shuiling[1,2*], Li Zhaoxian[1], Zhang Jiaxiong[1], Ding Longfei[1], Zhao Cairong[3]

1. School of Communication and Electronic Engineering, Jishou University, Jishou　416000, China; 2. Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, Nanjing　210094, China; 3. College of Electronics and Information Engineering, Tongji University, Shanghai　201804, China

Abstract: **Objective**　Emerging digital and intelligent technologies have ushered in a new era of text recognition and interpretation. These advancements have greatly facilitated the ability to recognize and comprehend textual content originating

收稿日期:2023-06-20;**修回日期**:2023-10-17;**预印本日期**:2023-10-23

*通信作者:曾水玲　zengflsl@163.com

基金项目:国家自然科学基金项目(61966014);湖南省自然科学基金项目(2024JJ7413);江苏省社会安全图像与视频理解重点实验室开放课题项目(202212);吉首大学校级科研项目(JGY2023071,Jdy23042);湖南省研究生科研创新项目(QL20230255,CX20221107)

Supported by:National Natural Science Foundation of China(61966014); Natural Science Foundation of Hunan Province, China(2024JJ7413)

from a variety of sources, including paper documents, photographs, and diverse contexts. One particularly noteworthy application of these technologies is in the field of chemical structure image recognition, where portable devices such as mobile phones and tablet PCs have become indispensable tools, playing a vital role in converting hand-drawn chemical structure images into machine-readable formats. They translate these intricate structures into human-readable representations, simultaneously highlighting relevant physical properties, chemical characteristics, and elemental compositions. These innovative models for chemical structure recognition serve as a bridge between hand-drawn representations and machine-interpretable data. This capability has made it feasible to electronically document complex scenarios, such as those encountered in classrooms and academic meetings. Notably, ongoing research has focused on developing encoder-decoder-based methods for mathematical expression recognition, which have shown promising results. However, the pivotal role of the quality and quantity of training data in shaping the performance of deep neural networks needs to be acknowledged. The current challenge lies in the absence of a comprehensive, high-quality dataset that is specifically tailored for chemical structure image recognition. This data deficiency poses a significant hurdle, impacting the optimization, generalization, and robustness of the models. Furthermore, the computational demands of real-time offline recognition on mobile devices remain a practical limitation. **Method** To address the aforementioned issues, we developed a chemical structure recognition model based on an encoder-decoder architecture. This model is capable of generating corresponding character representations, such as SMILES, from given chemical structure images. In the context of image-related tasks, the effectiveness of the encoder in extracting features from images and the decoder's ability to decode feature sequences directly impact the performance of the recognition task. The encoder is designed to efficiently model the input images, while the decoder should be able to comprehensively extract various features from the images, obtain accurate feature distributions, and encode them to establish feature maps. Therefore, we designed a feature extraction network based on ResNet-50 in the encoder, which adequately captures the two-dimensional structural information of chemical structure images. Furthermore, to enhance the effectiveness of information in feature maps, we introduced a row encoder based on bi-directional long-short term memory (BLSTM), reinforcing the spatial feature distribution weight through row encoding of feature maps. The decoder should be capable of accurately decoding the sequence information from the encoder's output. To align input sequence information with output characters and improve the model's memory and decoding capabilities for long sequences, we incorporated a coverage-attention mechanism into the decoder. Ultimately, the model can generate corresponding representations from input chemical structure images. **Result** For an objective evaluation of the performance of our model in this study, we conducted training on the Image2Mol and ChemPix models using the CASIA-CSDB (Institute of Automation, Chinese Academy of Sciences Chemical Structure Database) dataset. Subsequently, we performed performance testing on a range of datasets, including Indigo, ChemDraw, Conference and Labs of the Evaluation Forum (CLEF), Japanese Patent Office (JPO), University of Birmingham (UOB), United States Patent and Trademark Office (USPTO), Stacker, American Chemistry Society (ACS), CASIA-CSDB, and Mini CASIA-CSDB. Results demonstrated that our model achieved higher recognition accuracy when trained on small datasets and exhibited robust generalization capabilities. Furthermore, we compared our model with untrainable models such as SwimOCSR, MSE-DUDL, ChemGrapher, Image2Graph, and MolScribe. The comparison revealed that our model also exhibited commendable performance when compared with models trained on millions of images. **Conclusion** A chemical structure recognition method is introduced based on an encoder-decoder architecture. The method allows for the generation of SMILES strings from given chemical structure images. Experimental results demonstrate that the model achieves higher recognition accuracy when trained on small datasets and exhibits strong generalization capabilities.

Key words: chemical structure recognition; encoder-decoder; attention mechanism; residual network; SMILES (simplified molecular input line entry system)

0 引　言

化学结构识别在新药研发、环境监测等领域有着广泛的应用(Beard 和 Cole, 2020; Bukhari 等, 2019;刘成林 等,2023)。然而,传统的化学结构识别工作需要专业人员参与,任务量大且容易出错。因此,如何快速识别化学结构已经成为目前研究的

重点。

早期的化学结构图像识别研究主要是基于规则的光学化学结构图像识别方法，如Kekulé（McDaniel和Balmuth，1992）和CliDE（Ibison等，1993）可以将化学结构图像转换为连接表或其他适用于化学结构数据库的计算机可读格式。然而，这些方法存在众多缺点，如无法识别模糊或结构复杂的化学结构图像，各组件逻辑混乱关联复杂，系统难以改进等。

为解决上述模型面对复杂化学结构图像难以识别的问题，Filippov和Nicklaus（2009）提出了光学化学结构识别软件OSRA（optical structure recognition application），该方法使用标签字典来解析原子和超原子标签，且用户可以通过修改字典的方式识别特定结构。之后基于OSRA，Tharatipyakul等人（2012）结合文本挖掘技术提出了化学信息提取系统ChemEx。Smolov等人（2011）通过将分割结果细化为符号层和图形层提出了二维化学结构图像识别工具包Imago。Frasconi等人（2014）提出了光学分子结构识别方法MLOCSR（molecular optical chemical structure recognition），该方法使用模式识别技术、概率知识表示和推理的流水线集成，使用马尔可夫逻辑概率推理引擎（Domingos和Richardson，2007）对化学结构进行识别，这种方法的主要优点是可以将多个模型集成在一起，提高识别准确性。2019年，NIH（National Institutes of Health）的研发团队开发了基于Java的分子矢量化工具MOLVec（molecular vectorize）模型（Peryea等，2019），这是目前基于规则的化学结构图像识别工具中效果最优的，但识别速度和准确率依旧不甚理想。

随着硬件算力的发展，基于深度学习的光学化学结构识别方法越来越多。例如，分子图像提取工具MSE-DUDL（molecular structure extraction from documents using deep learning）（Staker等，2019）模型和光学分子图像识别工具ChemGrapher（Oldenhof等，2020）模型由分割网络和预测网络组成。分割网络部分由U-Net组成，用于分割文档中的化学结构图像；预测网络部分由卷积神经网络（convolutional neural network，CNN）和循环神经网络（recurrent neural network，RNN）组成，用于将化学结构图像解码为简化分子线性输入规范（simplified molecular input line entry system，SMILES）字符串。化学图像识别模型DECIMER（deep learning for chemical image recognition）（Rajan等，2020）的编码器由CNN和RNN组成。CNN部分结构类似于InceptonV3网络，用于提取公式图像特征；RNN部分则对特征图进行行编码。解码器部分由一个基于注意力机制的RNN解码器组成，通过对特征序列的解码实现对图像中公式内容的识别。随后，Rajan等人（2021）在DECIMER的基础上运用Transformer替换了原来的RNN解码器，提出了DECIMER1.0模型，增强了模型对长序列信息的解码能力。Weir等人（2021）提出的手绘碳氢化合物结构识别工具ChemPix由卷积神经网络编码器和长短期记忆神经网络解码器组成，结合图形增强、退化等技术，可实现由手绘化合物结构照片转化为机器可读的SMILES。分子图像转化器Img2Mol（image to molecular）（Clevert等，2021）模型在编解码器架构（Deng等，2017；杨晨等，2023）的瓶颈处加入了一个预训练的自编码器，有效地提高了识别准确率。基于图像到图生成的分子结构识别模型MolScribe（robust molecular structure recognition with image-to-graph generation）（Qian等，2023）使用Swim Transform作为编码器，Transformer作为解码器，在复杂化学结构图像识别中获得不错效果。

上述化学结构识别模型均通过增大模型参数和训练数据量的方式提高对化学结构图像识别的准确率，而未研究编码器部分对化学结构图像表征信息和空间特征信息综合编码能力以及解码器阶段特征信息保留能力对化学结构图像识别任务的识别准确率的影响。因此，本文提出了一种结合注意力机制和编码器—解码器架构的化学结构识别方法，在编码器阶段从抓取图像表征信息和空间特征信息的角度设计了结合上下文语义的特征提取网络；在解码器阶段引入位置感知的注意力机制和覆盖机制来帮助序列信息和预测字符对齐并去除重复序列信息。最后，多种数据集测试实验结果表明本文方法具有良好的识别准确率。

1 网络模型的改进

本文提出的结合注意力机制和编码器—解码器架构的化学结构识别模型如图1所示，主要包含3个部分：ResNet50特征提取网络、双向行编码器Row BLSTM（row bidirectional long short term memory）和

覆盖—注意力解码器(coverage-attention based LSTM, C-A LSTM)。首先,输入灰度图像 $\boldsymbol{I} \in \mathbf{R}^{1\times H\times W}$,经过 ResNet50 提取到特征图 $\boldsymbol{E} \in \mathbf{R}^{C\times H\times W}$;然后,将特征图输入到双向行编码器,得到特征编码后的特征序列 $\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_T]$,其中 T 是序列长度;最后,通过覆盖—注意力解码器生成 SMILES 预测结果。

1.1 特征提取器

与普通文本图像不同,化学结构图像拥有复杂的表征信息和繁杂的空间特征信息,这要求编码器在提取浅层表征信息和抓取深层空间特征信息的同时还能够摒除无意义的冗余信息。为此,本文基于 ResNet50 网络,通过优化网络结构,设计了特征提取网络,用于抓取化学结构特征。

为提高网络的表征信息提取能力和降低过拟合的风险,去除了 ResNet50 网络的第 3 阶段和第 4 阶段,即 Stage3 和 Stage4。另外,由于化学结构拥有复杂的空间特征,这要求编码器必须对 ResNet50 提取的表征信息进行行编码,以强化特征图中的深层空间特征表示。因此,去除了 ResNet50 网络的平均池化层和全连接网络层。通过以上操作后,输入的 1 × 256 × 256 灰度图像经过改进的 ResNet50 处理,将生成一幅富含上下文信息的 512 × 32 × 32 的特征图。改进后的 ResNet50 的网络结构如图 2 所示。

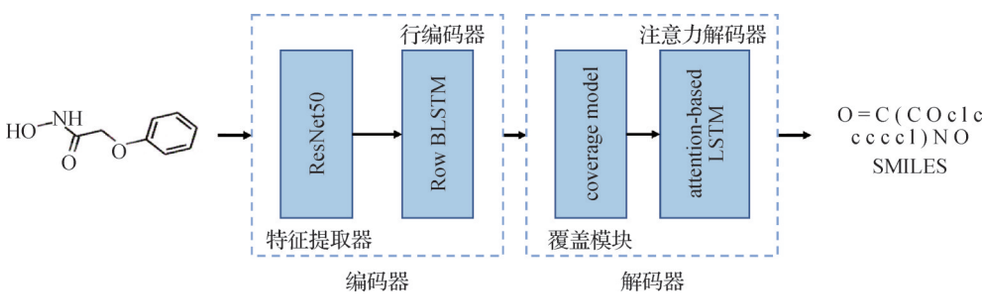

图 1 结合注意力机制和编码器—解码器框架的化学结构识别模型

Fig. 1 Chemical structure recognition framework combining attention mechanism and encoder-decoder model

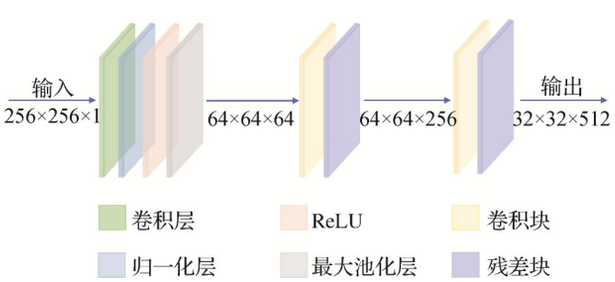

图 2 改进后的 ResNet50 的网络结构

Fig. 2 The optimized network structure of ResNet50

1.2 行编码器

双向长短期记忆网络 BLSTM(Zhang 等,2016)是一种特殊的循环神经网络 RNN,有效缓解了 RNN 在学习随机序列时易出现的梯度消失问题。本文采用 BLSTM 变体(Hamdi 等,2022)作为行编码器对特征图进行行编码。

在 BLSTM 中,每个时间步骤看做一个位置,同时从左向右和从右向左自下而上遍历序列,从而捕捉到当前位置和相对于当前位置的上下文信息,为特征图赋予空间特征,如图 3 所示。计算过程为

$$\hat{\boldsymbol{E}}_{h,w}=f_{\mathrm{BLSTM}}\left(\hat{\boldsymbol{E}}_{h-1,w},\boldsymbol{E}_{h,w}\right) \tag{1}$$

式中,f_{BLSTM} 为双向行编码器,$\boldsymbol{E}_{h,w}$ 为 ResNet50 输出的特征图,$\hat{\boldsymbol{E}}_{h,w}$ 为通过双向行编码器编码后的特征图,h,w 分别表示特征图的行数和列数,$\hat{\boldsymbol{E}}_{h-1,w}$ 为上一行特征进行行编码之后的特征。本文对每一行的特征都使用一个可训练的初始化隐藏状态来进行编码,并将输出的特征图传递给解码器进行解码。

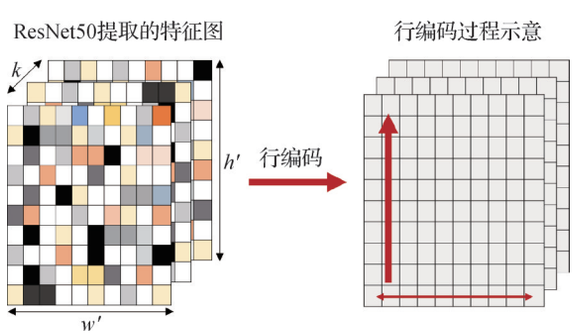

图 3 行编码器编码示意

Fig. 3 Illustration of row encoder encoding

1.3 覆盖—注意力编码器

传统 RNN 在处理长序列数据会对某些位置进行重复关注,难以准确地对齐输入序列和输出序列,

导致模型在生成输出时易出现重复内容。结合注意力机制的RNN(Hochreiter和Schmidhuber, 1997)能够有效地缓解以上问题,但在处理长序列和重复序列时效果依旧不佳。为此,提出了C-A LSTM做为解码器,如图4所示。

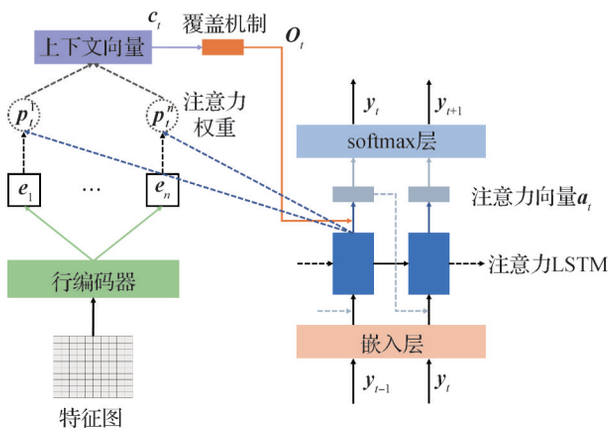

图4 覆盖—注意力解码器

Fig. 4 Coverage-attention based LSTM decode

覆盖—注意力机制解码器解码过程是在每个时间步骤t中,根据编码得到的特征向量$\boldsymbol{e}_n$和隐藏状态$\boldsymbol{h}_t$计算注意力权重$\boldsymbol{p}_t$,对所有特征向量计算加权平均值得到上下文向量$\boldsymbol{c}_t$。然后,使用覆盖机制对$\boldsymbol{c}_t$进行处理得到覆盖处理后的上下文向量$\boldsymbol{O}_t$。之后,运用覆盖处理后的上下文向量$\boldsymbol{O}_t$和隐藏状态$\boldsymbol{h}_t$计算出注意力向量$\boldsymbol{a}_t$。最后,将注意力向量通过softmax层,生成预测分布$\boldsymbol{y}_t$,具体为

$$\boldsymbol{y}_t = f_{\text{softmax}}\left(W_a \times \boldsymbol{a}_t\right) \tag{2}$$

式中,W_a为全连接层,$\boldsymbol{a}_t$为每一个时间步骤t中注意力LSTM网络的输出,由覆盖机制处理后的上下文向量$\boldsymbol{O}_t$和注意力LSTM输出的当前隐藏状态$\boldsymbol{h}_t$计算得到,具体为

$$\boldsymbol{a}_t = \tanh\left(\boldsymbol{W}_c\left[\boldsymbol{O}_t ; \boldsymbol{h}_t\right]\right) \tag{3}$$

式中,$\boldsymbol{W}_c[\cdot]$是一个权重矩阵,为$\boldsymbol{O}_t$和$\boldsymbol{h}_t$拼接组成的矩阵附加权重,隐藏状态$\boldsymbol{h}_t$计算过程为

$$\boldsymbol{h}_t = f_{\text{LSTM}}\left(\boldsymbol{h}_{t-1}, \left[f_{\text{Embed}}\left(\boldsymbol{y}_{t-1}\right), \boldsymbol{a}_{t-1}\right]\right) \tag{4}$$

式中,$\boldsymbol{y}_{t-1}$为上一个时刻输出的预测分布,$f_{\text{Embed}}[\cdot]$为嵌入层,$\boldsymbol{a}_{t-1}$为上一个时刻产生的注意力向量。

式(3)中覆盖模块处理后的上下文向量$\boldsymbol{O}_t$由上下文向量$\boldsymbol{c}_t$通过一个卷积运算得到,具体为

$$\boldsymbol{O}_t = f_{\text{conv}}\left(\boldsymbol{c}_t\right) \tag{5}$$

式中,$f_{\text{conv}}(\cdot)$表示卷积运算,上下文向量$\boldsymbol{c}_t$由注意力权重$\boldsymbol{p}_t$和编码后的特征$\boldsymbol{e}_i$计算得到,具体为

$$\boldsymbol{c}_t = \sum_{i=1}^{n} \boldsymbol{p}_t^i \boldsymbol{e}_i \tag{6}$$

式中,$\boldsymbol{p}_t$由注意力LSTM网络输出的隐藏状态$\boldsymbol{h}_t$和编码后的特征$\boldsymbol{e}_i$计算得到,具体为

$$\boldsymbol{p}_t^i = \frac{\exp\left(\tanh\left(\boldsymbol{w}_h \boldsymbol{h}_t, \boldsymbol{w}_e \boldsymbol{e}_i\right)\right)}{\sum_{j=1}^{n} \exp\left(\tanh\left(\boldsymbol{w}_h \boldsymbol{h}_t, \boldsymbol{w}_e \boldsymbol{e}_j\right)\right)} \tag{7}$$

式中,$\boldsymbol{w}_h$和$\boldsymbol{w}_e$为权重矩阵,用于将隐藏状态$\boldsymbol{h}_t$和编码后特征向量$\boldsymbol{e}_i$映射到相同空间,实现相似性比较。

1.4 去填充模块

由于各物质元素构成不同使对应的化学结构大小长度各不相同,导致本文使用数据集中化学结构图像尺寸不同。因此,网络在训练前需要在数据预处理阶段将图像填充至相同尺寸,以保证训练时每个批次的图像大小相同。但填充后的图像会增加冗余信息,干扰模型对化学结构的识别和解码。为此,在改进解码器的同时添加了去填充模块,用于去除编码器输出特征图中填充的部分,使模型只关注有效的信息,从而提高化学结构的识别和解码的准确率,提高模型的性能和鲁棒性。图5为去填充模块的示意图,其中红框部分是填充的像素。

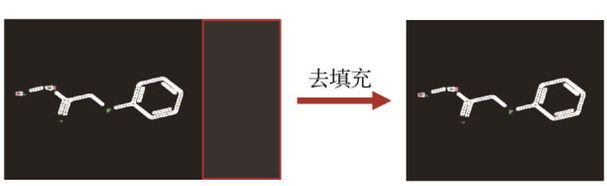

图5 去填充模块示意图

Fig. 5 Schematic diagram of homogeneous trimming

2 实验与分析

2.1 实验数据集与评估指标

2.1.1 数据集

CASIA-CSDB(Institute of Automation of Chinese Academy of Science—Chemical structure Database, C-C)是由中国科学院自动化研究所模式识别国家重点实验室创建的公开数据集(Ding等,2022),包含了来自6 500多篇出版物和50个数据库的480 668个具有苯环、氨基、羧基、羟基、醛、硫基等多种常见官能团的化学结构样本图像及其对应的SMILES字符串。另外,为了满足快速设计和评估的需求,从

CASIA-CSDB 数据集的 8 个不同 mol 分区中随机选择了 20% 的化学结构样本图像，共计 97 309 个，构建了 Mini CASIA-CSDB（Mini C-C）数据集，如表 1 所示。表中括号部分为使用镜像和旋转的方法对原始数据集的困难和复杂部分进行数据增强生成的图片数量，其目的为增强模型的泛化性和鲁棒性。数据增强图像如图 6 所示。

表 1　数据集信息

Table 1　Database information

数据集	训练集/幅	验证集/幅	测试集/幅
CASIA-CSDB	395 204（+237 122）	36 240	49 224
Mini CASIA-CSDB	80 781（+48 472）	8 286	8 242

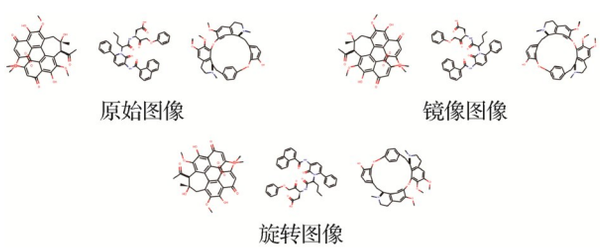

图 6　数据增强部分示例

Fig. 6　Data augmentation example

2.1.2　数据集评估指标

本文使用精确匹配分数作为主要评价指标，并使用字符错误率和相似度度量 BLEU-4 来衡量翻译的 SMILES 序列与真实 SMILES 序列之间的相似度。

1）精准匹配率（exact matchscore，EM）。EM 指的是精确匹配真实 SMILES 字符串的预测 SMILES 字符串的比率。设 T 表示样本总数，R 表示与理想 SMILES 字符串完全匹配的预测 SMILES 字符串数。EM 的计算过程为

$$f_{\mathrm{EM}} = \frac{R}{T} \tag{8}$$

2）字符错误率（character error rate，CER）。CER 的计算基于 Levenshtein 距离概念，搜索以删除、插入和替换操作将预测的 SMILES 字符串转换为真实的 SMILES 字符串的最小次数。删除、插入和替换操作的数量分别用 D、I 和 S 表示。N 是预测 SMILES 字符串长度与理想 SMILES 字符串长度之间的最大值。CER 的具体计算过程为

$$f_{\mathrm{CER}} = \frac{S + D + I}{N} \tag{9}$$

3）双语评价替换（bilingual evaluation understudy，BLEU）。BLEU 值是评价机器翻译质量的指标，其中 BLEU-4 值通常用于衡量预测序列的质量。BLEU-4 定义为 n-gram 匹配精度分数的几何平均值乘以一个简单的惩罚因子 BP，以防止生成非常短的句子获得高分，计算过程为

$$BLEU = BP \times \exp\sum_{i=0}^{n} \xi \log\left(P_n\right) \tag{10}$$

式中，$\xi = 1/4$，P_n 为修正后 n-gram 精度的几何平均值。BP 是惩罚因子，计算过程为

$$BP = \begin{cases} 1 & c > r \\ \exp\left(1 - \dfrac{r}{c}\right) & c \leqslant r \end{cases} \tag{11}$$

式中，c 为翻译文本长度，r 为参考文本长度。

2.1.3　测试数据集

为方便与其他化学结构识别模型进行性能比较，选取了 10 个类别测试集，如表 2 所示。

首先，使用 CASIA-CSDB 和 Mini CASIA-CSDB 的测试集验证模型的性能。随后，使用 Molscribe 模型提供的 Indigo、ChemDraw 和 ACS（American Chemistry Society）测试集测试模型对合成化学结构图像和其他绘画风格化学结构图像的识别性能。最后，使用 CLEF（Conference and Labs of the Evaluation Forum）、JPO（Japanese Patent Office）、UOB（University of Birmingham）、USPTO（United States Patent and Trademark Office）和 Staker 测试集验证模型对科学出版物中化学结构图像的识别性能。

表 2　测试集信息

Table 2　Test dataset information

测试数据集	来源	图像总量/幅	手性比率/%
CASIA-CSDB	出版物和数据库	49 224	17.23
Mini CASIA-CSDB	出版物和数据库	8 287	14.52
Indigo	Indigo 绘制	5 719	20.20
ChemDraw	ChemDraw 绘制	5 719	20.20
CLEF	美国专利局	992	32.70
JPO	日本专利局	450	0
UOB	化学结构图录	5 740	0
USPTO	美国专利局	5 719	20.20
Staker	美国专利局	50 000	17.30
ACS	期刊	331	19.30

2.2 实验结果

2.2.1 对比实验

实验在一个3.6 GHz 10核CPU、32 GB RAM、RTX 3070 8 GB的GPU和Windows 10 21H1操作系统的机器上进行。输入图像尺寸设置为256×256像素，批处理大小设置为32。初始学习率设置为0.1，当训练准确率停止提高时，学习率降低为一半。连续5次降低学习率而训练准确率不变时停止训练。另外，在训练时引入随机数为模型设置了权重扰动。

为防止测试数据集中图像尺寸不统一导致显存占用过大的问题，将图像维度缩放至256×256像素，不做其他任何处理。

为了验证本文方法的性能，与目前公开的10种化学结构识别模型（基于规则的MolVec和OSRA，以及基于深度学习的Img2Mol、DECIMER、SwimOCSR（Xu等，2022）、MSE-DUDL（Staker等，2019）、ChemGrapher、Image2Graph（Clevert等，2021）、ChemPix和MolScribe）进行比较。为客观比较不同化学结构图像识别模型的性能，统一使用CASIA-CSDB数据集进行训练，并使用2.1.3节中介绍的10种测试集进行模型性能对比实验，比较结果如表3所示，其中，C-C和Mini C-C分别表示CASIA-CSDB和Mini CASIA-CSDB测试集。评估指标为精准匹配率EM。

从实验结果可以发现，本文方法在各类测试集性能测试中取得了较优的识别率。需要说明的是，表中（a）模型均使用数百万数量级图像的训练集训练得到，而本文方法训练集仅有60万幅图像（包括数据增强的23万幅）；Img2Mol和ChemPix在使用CASIA-CSDB数据集进行训练后，测试集识别率得分低于本文方法；本文方法在少量数据集训练情况下表现出良好的表现。

为更直观地展现本文方法的性能，图7展示了在不同测试集上的识别准确率和基准值比较结果。图中结果表示，本文方法在多数测试集上表现出较好的识别准确率；在CLEF、UOB测试集上表现不佳，通过分析发现这是由于测试集中噪声过大而训练集中缺少噪声图像导致的；在JPO测试集中的表现，则是由于测试集中图像充斥大量非化学信息的文字标注，导致模型出现了误识别现象。

2.2.2 消融实验

特征提取网络的特征抓取能力对本文方法的性能影响巨大。为此，本文基于CASIA-CSDB和Mini

表3 不同模型识别结果EM对比

Table 3 Comparison of recognition (EM) results of different models

/%

化学结构识别模型		合成得到的数据集		真实文档和专利提取得到的数据集						本文训练用数据集的测试集	
		Indigo	ChemDraw	CLEF	JPO	UOB	USPTO	Staker	ACS	C-C	Mini C-C
基于规则模型	MolVec	95.4	87.9	82.8	67.8	80.6	88.4	0.8	47.4	97.3	93.5
	OSRA	95.0	87.3	84.6	55.3	78.5	87.4	0.0	55.3	96.9	93.7
基于深度学习的模型	Img2Mol (a)	58.9	46.4	18.3	16.4	68.7	26.3	17.0	23.0	–	–
	DECIMER (a)	69.6	86.1	62.7	55.2	88.2	41.1	40.8	46.5	–	–
	SwimOCSR (a)	74.0	79.6	30.0	13.8	44.9	27.9	–	27.5	–	–
	MSE-DUDL (a)	–	–	–	–	–	–	77.0	–	–	–
	ChemGrapher (a)	–	–	–	–	70.6	–	–	–	–	–
	Image2Graph (a)	–	–	51.7	50.3	83.9	55.1	–	–	–	–
	MolScribe (a)	97.5	93.8	88.9	76.2	87.9	92.6	86.9	71.9	–	–
	Img2Mol (b)	20.3	18.6	0.2	0.0	30.1	11.2	6.2	5.2	55.6	52.1
	ChemPix (b)	30.2	31.3	0.3	0.1	25.2	9.62	4.65	4.95	48.6	50.1
本文模型	基准 (c)	67.61	66.38	46.61	37.23	65.86	48.85	29.17	35.22	52.1	51.1
	本文	71.10	70.21	45.80	30.30	53.02	58.21	43.39	46.30	84.42	85.78

注："–"表示该数据集不可用或论文原文未提供数据。(a) 表示原始论文提供的数据或模型无法重新训练；(b) 表示可使用CASIA-CSDB数据集重新训练；(c) 为基准识别率，由对比模型识别结果取均值得到。

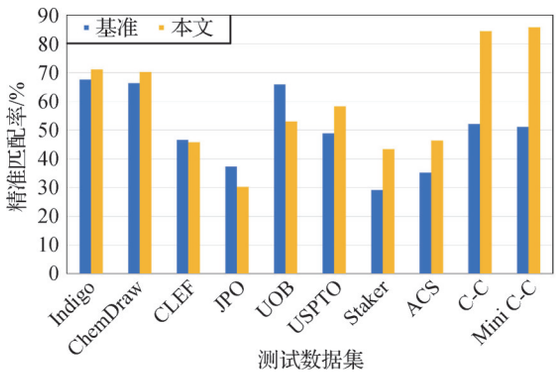

图 7　不同测试集的识别结果

Fig. 7　Recognition results of different test

CASIA-CSDB 数据集，以 ChemPix 模型中的 CNN 为基准模型，进行一系列对比实验以比较不同深度的 ResNet 和 DenseNet（dense convolutional network）的性能表现。表 4 列出了 ResNet18、ResNet50、ResNet152；DenseNet121、DenseNet169 和 DenseNet201 各特征提取网络对原始模型性能的影响。由于 ResNet18 和 ResNet34 以及 ResNet50 和 ResNet101 结构相似且实验结果无较大差异，因此表 4 中未列出 ResNet34 和 ResNet101 的数据。

实验结果表明，在两个数据集上的 ResNet50、DenseNet169 和 DenseNet201 的性能测试结果相近，其中，DenseNet169 表现较差，而 DenseNet201 较 ResNet50 提升有限且网络层数深、训练时间过长、计算成本高昂，因此，本文选择 ResNet50 作为特征提取网络是较合理的。

为进一步说明对 ResNet50 进行结构优化的合理性，表 5 展示了 Stage3 和 Stage4 对 ResNet50 的性能影响。通过实验可知，去除 Stage4 后 EM 度量提高了 0.64% 和 1.61%，CER 值分别上升了 0.19% 和下降了 0.39%，BLEU-4 指数提升了 0.06% 和 0.89%；去除 Stage4 和 Stage3 后在 C-C 数据集上 EM 度量、CER 值和 BLEU-4 指数相较单独去除 Stage4 均有微小下降，但较原始 ResNet50 网络仍有一定提升，在 Mini C-C 数据集上则出现了全面提升；考虑到去除 Stage4 和 Stage3 能够极大地降低网络参数，提高网络运算速率，因此本文对 ResNet50 的改进是合理的。

表 6 列出了不同模块对模型性能的影响，本文基础模型的 EM 度量分别为 72.09% 和 73.06%，CER 指标分别为 16.98% 和 16.68%，BLEU-4 值分别为 84.01% 和 85.91%；在加入 ResNet50 模块后，EM 度量得分提高到 79.02% 和 80.71%，CER 下降到 6.35% 和 5.97%，BLEU-4 提升到 92.25% 和 92.12%；加入行编码器后，EM 度量得分达到 82.01% 和 83.91%，CER 指标减少到 5.98% 和 5.26%，BLEU-4 值提高到 93.56% 和 94.32%；在解码器中添加覆盖—注意力机制后，EM 度量增长到 84.21% 和 84.51%，CER 指标降低至 5.41% 和 4.29%，BLEU-4 值提高到 94.71% 和 95.11%；添加去填充模块后，EM 提升至 84.42% 和 85.78%，CER 指标进一步下降至 5.22% 和 4.10%，BLEU-4 值上升为 94.76% 和 95.21%。

表 4　不同特征提取网络对原始模型性能的影响

Table 4　Impact of different feature extraction networks on the performance of the base model

/%

模块	CASIA-CSDB 数据集			Mini CASIA-CSDB 数据集		
	EM	CER	BLEU-4	EM	CER	BLEU-4
CNN	72.09	16.98	90.01	73.06	16.68	93.91
ResNet18	76.02	6.45	92.25	76.51	6.52	94.01
ResNet50	78.54	6.35	92.95	79.22	6.49	94.18
ResNet152	77.99	6.65	91.92	77.19	6.44	93.98
DenseNet121	78.21	6.35	92.35	77.96	6.55	94.02
DenseNet169	78.49	6.29	92.98	78.89	6.48	94.20
DenseNet201	78.62	6.28	93.08	79.31	6.45	94.25

通过多组对比实验结果及分析，可得出结论，即所设计的 4 个子模块对本文方法的识别准确率均有提升作用。

表 5　不同层级下 ResNet50 性能对比实验

Table 5　Comparison experiment of ResNet50 performance at different stages

/%

数据集	ResNet50		EM	CER	BLEU-4
	Stage3	Stage4			
CASIA-CSDB	√	√	78.54	6.35	92.95
	√	×	79.18	6.54	93.01
	×	×	79.02	6.32	92.95
Mini CASIA-CSDB	√	√	77.21	6.81	91.02
	√	×	78.82	6.42	91.91
	×	×	80.07	5.97	92.12

注：“√”表示使用该模块，“×”表示未使用。

表6　不同模块对模型性能的影响

Table 6　The impact of different modules on network performance

数据集	模块				EM/%	CER/%	BLEU-4/%
	ResNet50	行编码器	覆盖—注意力解码器	去填充模块			
CASIA-CSDB	×	×	×	×	72.09	16.98	84.01
	√	×	×	×	79.02	6.35	92.25
	√	√	×	×	82.01	5.98	93.56
	√	√	√	×	84.21	5.41	94.71
	√	√	√	√	**84.42**	**5.22**	**94.76**
Mini CASIA-CSDB	×	×	×	×	73.06	16.68	85.91
	√	×	×	×	80.71	5.97	92.12
	√	×	×	×	83.91	5.26	94.32
	√	√	×	×	84.51	4.29	95.11
	√	√	√	√	**85.78**	**4.10**	**95.21**

注：加粗字体表示各数据集各列最优结果，“√”表示使用该模块，“×”表示未使用。

3　结　论

针对当前公开的化学结构识别模型训练时间长、参数量大、提取的特征缺乏上下文信息、空间信息、注意力机制重复覆盖等问题，提出了一种结合注意力机制和编码器—解码器架构的化学结构识别方法。该方法在编码器中采用改进的ResNet50残差网络做为模型的特征提取网络，提高了原始模型的特征提取能力，使用BLSTM做为行编码器为ResNet50提取的特征做特征强化，增强空间特征在特征图中的信息量，从而提高化学结构识别准确率；在解码器中加入去填充模块用于去除预处理中填充的无用信息，减少干扰，加入覆盖—注意力机制增强LSTM解码器记忆能力和长序列解码出现的信息复用问题，减少翻译错误。实验结果表明，提出的化学结构识别方法在Indigo、ChemDraw、CLEF、JPO、UOB、USPTO、Staker、ACS、CASIA-CSDB和Mini CASIA-CSDB数据集上相较SwimOCSR、Image2Mol和ChemPix模型表现更优。

然而，本文方法仍有许多不足之处，如手性分子识别率不佳，超大规模分子识别率低，模型性能受图像质量影响等。未来的研究重点包括如何加强解码器网络对长序列的解码能力，如何改进Transformer网络并运用于化学结构图像识别任务，以及改进模型分类层，通过优化分类结果的方式增加模型的鲁棒性。

致谢：本文多项实验使用中国科学院自动化研究所模式识别国家重点实验室的研究团队采集的CASIA-CSDB和Mini CASIA-CSDB公开数据集，在此表示感谢。

参考文献（References）

Beard E J and Cole J M. 2020. ChemSchematicResolver: a toolkit to decode 2D chemical diagrams with labels and R-groups into annotated chemical named entities. Journal of Chemical Information and Modeling, 60(4): 2059-2072 [DOI: 10.1021/acs.jcim.0c00042]

Bukhari S S, Iftikhar Z and Dengel A. 2019. Chemical structure recognition (CSR) system: automatic analysis of 2D chemical structures in document images//Proceedings of 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney, Australia: IEEE: 1262-1267 [DOI: 10.1109/icdar.2019.00-41]

Clevert D A, Le T, Winter R and Montanari F. 2021. Img2Mol—accurate SMILES recognition from molecular graphical depictions. Chemical Science, 12(42): 14174-14181 [DOI: 10.1039/D1SC01839F]

Deng Y T, Kanervisto A, Ling J and Rush A M. 2017. Image-to-markup generation with coarse-to-fine attention//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org: 980-989

Ding L F, Zhao M B, Yin F, Zeng S L and Liu C L. 2022. A large-scale database for chemical structure recognition and preliminary evaluation//Proceedings of the 26th International Conference on Pattern

Recognition (ICPR). Montreal, Canada: IEEE: 1464-1470 [DOI: 10.1109/icpr56361.2022.9956654]

Domingos P and Richardson M. 2007. Markov logic: a unifying framework for statistical relational learning//Getoor L and Taskar B, eds. Introduction to Statistical Relational Learning. Cambridge, USA: MIT Press: 339-371 [DOI: 10.7551/mitpress/7432.003.0014]

Filippov I V and Nicklaus M C. 2009. Optical structure recognition software to recover chemical information: OSRA, an open source solution. Journal of Chemical Information and Modeling, 49(3): 740-743 [DOI: 10.1021/ci800067r]

Frasconi P, Gabbrielli F, Lippi M and Marinai S. 2014. Markov logic networks for optical chemical structure recognition. Journal of Chemical Information and Modeling, 54(8): 2380-2390 [DOI: 10.1021/ci5002197]

Hamdi Y, Boubaker H, Rabhi B, Qahtani A M, Alharithi F S, Almutiry O, Dhahri H and Alimi A M. 2022. Deep learned BLSTM for online handwriting modeling simulating the Beta-Elliptic approach. Engineering Science and Technology, an International Journal, 35: #101215 [DOI: 10.1016/j.jestch.2022.101215]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Ibison P, Jacquot M, Kam F, Neville A G, Simpson R W, Tonnelier C, Venczel T and Johnson A P. 1993. Chemical literature data extraction: the CLiDE project. Journal of Chemical Information and Computer Sciences, 33(3): 338-344 [DOI: 10.1021/ci00013.a010]

Liu C L, Jin L W, Bai X, Li X H and Yin F. 2023. Frontiers of intelligent document analysis and recognition: review and prospects. Journal of Image and Graphics, 28(8):2223-2252 (刘成林, 金连文, 白翔, 李晓辉, 殷飞. 2023. 文档智能分析与识别前沿:回顾与展望. 中国图象图形学报, 28(8): 2223-2252)

McDaniel J R and Balmuth J R. 1992. Kekule: OCR-optical chemical (structure) recognition. Journal of Chemical Information and Computer Sciences, 32(4): 373-378 [DOI: 10.1021/ci00008a018]

Oldenhof M, Arany A, Moreau Y and Simm J. 2020. ChemGrapher: optical graph recognition of chemical compounds by deep learning. Journal of Chemical Information and Modeling, 60(10): 4506-4517 [DOI: 10.1021/acs.jcim.0c00459]

Peryea T, Katzel D, Zhao T, Southall N and Nguyen D T. 2019. MOLVEC: open source library for chemical structure recognition// Abstracts of Papers of the American Chemical Society. San Diego, USA: ACS: #258

Qian Y J, Guo J, Tu Z K, Li Z N, Coley C W and Barzilay R. 2023. MolScribe: robust molecular structure recognition with image-to-graph generation. Journal of Chemical Information and Modeling, 63(7): 1925-1934 [DOI: 10.1021/acs.jcim.2c01480]

Rajan K, Zielesny A and Steinbeck C. 2020. DECIMER: towards deep learning for chemical image recognition. Journal of Cheminformatics, 12(1): #65 [DOI: 10.1186/s13321-020-00469-w]

Rajan K, Zielesny A, Steinbeck C. 2021. DECIMER 1.0: deep learning for chemical image recognition using transformers. Journal of Cheminformatics, 13: 1-16 [DOI: 10.1186/s13321-021-00538-8]

Smolov V, Zentsev F and, Rybalkin M. 2011. Imago: open-source toolkit for 2D chemical structure image recognition//Proceedings of the 20th Text REtrieval Conference, Gaithersburg, USA: NIST Special Publication: 296-500

Staker J, Marshall K, Abel R and McQuaw C M. 2019. Molecular structure extraction from documents using deep learning. Journal of Chemical Information and Modeling, 59(3): 1017-1029 [DOI: 10.1021/acs.jcim.8b00669]

Tharatipyakul A, Numnark S, Wichadakul D and Ingsriswang S. 2012. ChemEx: information extraction system for chemical data curation. BMC Bioinformatics, 13(Suppl 17): #S9 [DOI: 10.1186/1471-2105-13-S17-S9]

Weir H, Thompson K, Woodward A, Braun A and Martínez T J. 2021. ChemPix: automated recognition of hand-drawn hydrocarbon structures using deep learning. Chemical Science, 12(31): 10622-10633 [DOI: 10.1039/D1SC02957F]

Xu Z P, Li J H, Yang Z P, Li S L and Li H L. 2022. SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer. Journal of Cheminformatics, 14(1): #41 [DOI: 10.1186/s13321-022-00624-5]

Yang C, Du J, Xue M B and Zhang J S. 2023. An encoder-decoder based generation model for online handwritten mathematical expressions.Journal of Image and Graphics, 28(8): 2356-2369 (杨晨, 杜俊, 薛莫白, 张建树. 2023. 用于在线手写公式合成的编解码网络. 中国图象图形学报, 28(8): 2356-2369)

Zhang H W, Wang M, Hong R C and Chua T S. 2016. Play and rewind: optimizing binary representations of videos by self-supervised temporal hashing//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, the Netherlands: Association for Computing Machinery: 781-790

作者简介

曾水玲，女，教授，主要研究方向为人工智能和模式识别。E-mail：zengflsl@163.com

李昭贤，男，硕士研究生，主要研究方向为深度学习和图像识别。E-mail：lzx9241@163.com

张嘉雄，男，硕士研究生，主要研究方向为深度学习和图像检测。E-mail：1265492924@qq.com

丁龙飞，男，硕士研究生，主要研究方向为深度学习和图像识别。E-mail：271949947@qq.com

赵才荣，男，教授，主要研究方向为模式识别、机器学习、计算机视觉和图像内容理解。E-mail：zhaocairong@tongji.edu.cn