小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0253
2026 年 5 月 第 5 期
Vol.47 No.5 2026

TGMS:一种端到端的表格图像到标记序列的识别框架

李世琪,金大海,宫云战

(北京邮电大学 网络与交换技术全国重点实验室,北京 100876)
E-mail:jindh@bupt.edu.cn

摘 要: 由于表格样式和布局的多样性,从文档图像中识别二维结构的表格是一项复杂的任务.表格以紧凑的形式表达数据内容,提高信息传递和人类理解效率,但与人类相比,机器需要理解二维结构与内容之间的关系,因此使用机器自动识别表格面临很大的挑战.针对这一任务,提出了一种端到端的表格图像到标记序列的识别框架(TGMS:An End-to-End Framework for Table Graph to Markup Sequence).该框架首先使用卷积神经网络来进行视觉特征提取,然后采用基于分割的方法识别单元格空间位置,构建表图并利用图卷积网络和注意力机制推导逻辑关系,最后识别区域内文本并结合逻辑关系生成表格标记序列.在 ICDAR-2013、SciTSR、PubTabNet 3 个广泛使用的表格识别数据集上的实验结果表明,所提出的 TGMS 能有效完成表格识别任务.

关 键 词: 表格结构识别;表格识别;端到端;图卷积网络;注意力机制
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)05-1175-07

TGMS:an End-to-end Framework for Table Graph to Markup Sequence

LI Shiqi,JIN Dahai,GONG Yunzhan
(State Key Laboratory of Networking and Switching Technology,Beijing University of Posts and Telecommunication,Beijing 100876,China)

Abstract: Recently,due to the variety of table styles and layouts,recognizing tables with 2D structure from document images is a complex task. Tables express data content in a compact form to improve the efficiency of information transfer and human comprehension, but the relationship between the 2D structure and the content needs to be understood by machines,making it challenging to automatically recognize tables. To address these issues,an end-to-end framework for Table Graph to Markup Sequence is proposed,named TGMS. The framework first uses a convolutional neural network for visual feature extraction,and then employs a segmentation-based approach to recognize the spatial location of cells. Secondly,it uses spatial location information to recognize the text in the region and constructs a graph,and deduces logical relationships using a graph convolutional network and an attention mechanism. Finally,the last module generates a sequence of table tokens by combining the logical relationships and the text in the cell. Experimental results on three widely used form recognition datasets,ICDAR-2013,SciTSR,and PubTabNet,show that the proposed TGMS can effectively accomplish the form recognition task.

Keywords: table structure recognition;table recognition;end-to-end;graph convolutional network;attention mechanism

0 引 言

表格作为一种紧凑的表达方式在文档中广泛使用,用于快速可视化信息,在许多实际场景中管理和提取重要信息,包括财务文件、空气污染指数和电子病历的分析[1,2].随着数字化进程的推进,解析非结构化数据(例如图像和 PDF 文件)中表格的需要正在迅速增长[3-5].通常人类擅长解析表格结构、识别表格标题和解释表格单元格之间的关系,但是由于表格数据布局和样式存在很大差异,使用机器自动识别却很困难,由于这些广泛使用的文档格式都没有保留逻辑表结构,因此需要精确的表检测和表格结构识别技术来重建表格,然后将识别的内容用于后续分析工作[6].

从表中自动提取信息涉及两个基本的子任务:表结构识别和表识别.表格结构识别(Table Structure Recognition,TSR)是将一张表格图片的结构重构成机器可读格式的形式,但忽略了表格的内容.TSR 在构建完整结构前通过使用两组预测内容:空间位置和逻辑位置.空间位置表示表格单元格的布局信息,例如单元格的边界框;而逻辑位置表示单元格之间的关系信息,可以使用 HTML 形式表示.表格识别(Table Recognition,TR)则包含表格内容识别(Table Content Recognition,TCR)任务,其进一步考虑表格单元格内部文本内容,将一张表图片的结构和内容同时重构成机器可读格式的形式.机器可读格式的表格有许多潜在的应用,包括信息抽取或表格问答等.

早期的识别表格的研究工作[7-9]主要依赖人工设计的特征和启发式规则,这些方法主要应用于结构简单或格式预定义的表格.随着基于深度学习的目标检测和语义分割方法得到研究,提出了许多基于深度学习的表格识别方法,并通过实验证明功能强大[10-13].然而,这些方法依赖于包含丰富注释信息的训练数据集,并且由多个独立组件组成,因此难以维

收稿日期:2025-06-19 收修改稿日期:2025-08-18 基金项目:国家自然科学基金面上项目(62472043)资助. 作者简介:**李世琪**,男,2000 年生,硕士研究生,研究方向为表格识别;**金大海**(通信作者),男,1974 年生,博士,副教授,CCF 会员,研究方向为软件测试、缺陷检测;**宫云战**,男,1962 年生,博士,教授,CCF 会员,研究方向为软件测试、软件可靠性.

护. 近年来,受到 Transformer[14] 在自然语言处理和计算机视觉等广泛机器学习任务中取得成功的启发,人们研究了许多基于 Transformer 的表格识别方法[15-18],并在大规模表格图像数据集上取得了有竞争力的结果,但是由于 Transformer 的计算复杂度高且存在最大序列长度限制,捕获单元之间的长距离依赖关系有着较大的困难. 因此,冗长的表格可能会丢失信息,从而影响模型准确理解上下文的能力. 另外,使用全局注意力机制而未考虑表格的局部信息会导致缺乏结构约束,邻接关系建模不精确.

之前的方法试图使用独立的模型来解决表格识别中的不同子任务,这不可避免导致表格识别出现错误累积,多个模块缺乏协作和联合优化,使得识别性能欠佳.

针对以上问题,本文提出 TGMS:一种端到端的表格图像到标记序列的识别框架,用于联合结构识别和内容识别. 该框架采用基于分割的模块来预测单元格空间位置,在空间位置信息基础上,构建表图并通过图卷积网络来表达图的邻接关系,同时分割单元格区域实现局部于区域内的文本内容识别,最后通过逻辑结构及文本内容生成标记序列. 在 3 个公共数据集上的实验表明,所提识别框架识别性能有了明显提升,证明了所提方法的有效性. 本文的贡献总结如下:

· 提出 TGMS 统一框架,用于预测单元格空间位置、识别逻辑结构和基于空间位置局部识别单元格文本内容,融合识别信息生成表格标记序列.

· 设计了端到端的训练策略,以减少多阶段模块的错误积累,提高整体性能.

· 在 3 个广泛使用的公共数据集上进行评估. 结果表明,TGMS 在各数据集的指标中均得到了提升,提高了识别性能.

1 相关工作

1.1 表格结构识别

单元格检测与分割:受语义分割和对象检测研究方法的启发,Schreiber 等人[10] 提出的 DeepDeSRT 方法使用更快的 RCNN[19] 和 FCN[20] 进行表检测和行/列分割. Paliwal 等人[13] 提出了一种端到端的深度模型 TableNet,该模型具有一个编码器和两个解码器来进行表和列的分割. 在之后的研究工作中[21-23],将整行或整列分类为单元格或非单元格类别,而不是按像素分类. Siddiqui 等人[24] 提出的 DeepTabStR 模型,将行(列)识别问题看作对象检测问题. Zheng 等人[25] 提出了 GTE 模型,其使用基于对象检测的方法直接检测单元格,并在后处理中使用启发式规则来恢复表结构. Qiao 等人[11] 提出了 LGPMA 模型,模型在局部和全局层面应用软金字塔掩码,使模型能够更准确地检测无线表的单元边界. 以上研究深入探索了预测单元格空间位置的方法,但忽视了单元格逻辑位置的重要性. Sachin 等人[26] 提出了一种表结构识别方法,该方法结合了单元格检测和交互模块来定位单元格并预测它们与其他检测到的单元格的行和列关联.

标记序列生成:受自然语言处理(NLP)研究工作的影响,一些研究方法应用图像到序列模型[27,28],其中包括用于提取特征的编码器和用于生成标签序列的解码器,试图将表格图像转换为标记序列(如:LaTeX 或 HTML 格式),最终通过解析标记序列来识别表的结构. Deng 等人[28] 采用注意力编码器-解码器模型,建模了基于 LSTM 的表格到 LaTeX 框架,直接生成了完整的表格结构. Ye 等人[16] 提出了图像编码器双解码器方法(IEDD),在对表格图像进行编码后,首先使用一个解码器预测 HTML 结构标签(逻辑结构),然后以这些标签为条件使用另一个解码器生成单元格边界框(物理结构). 这些单元格边界框随后映射到使用 OCR 引擎或通过 PDF 解析获取的表格文本区域,将其与逻辑结构合并以生成完整的 HTML 序列.

后续研究工作提出了改进双解码器框架的不同方法. 为提升物理结构预测的准确性,Huang 等人[29] 提出了 VAST 框架,其引入了一种视觉对齐损失机制,在解码阶段显式地融合细节级视觉特征,以加强视觉与结构之间的对应关系. 为减少逻辑和物理结构生成中错误累积的影响,Shen 等人[30] 提出了 DRCC 模型,其以非自回归方式预测表格的 HTML 行标签,以半自回归方法预测单元格标签,以非自回归方式预测表格的 HTML 行标签,逐步生成每个单元格的结构信息.

尽管以上方法可以使用诸如 HTML 形式来表示表格信息,但是,标记序列包含不同样式的多样化命令,这使得表格结构可以转录成不同的标记序列. 这种一对多的映射给真值(Ground Truth)带来大量噪声,并阻碍模型训练.

基于图的方法:随着图神经网络的发展,一些研究提出可以提取表格元素(单元格或文本行),然后使用图网络来了解提取的表格元素之间的关系. Chi 等人[31] 提出 GraphTSR 模型,采用图注意力机制,融合顶点和边的特征,分别在水平和垂直方向上对相邻的顶点进行分类,判断这些顶点是否相邻. Sachin 等人[26] 提出 TabStructNet 架构,提供了一种端到端的解决方案,在网络中同时进行表元素检测和顶点关系预测. Hao 等人[32] 提出的 FLAG-Net 模型采用自注意力和图神经网络分别提取密集特征和稀疏特征,并设计门控单元聚合信息. 这些方法依赖于 OCR 注释或结果,但是 OCR 注释在实际应用程序场景中可能并不存在,同时 OCR 识别的潜在错误会从 OCR 传播到表结构识别.

1.2 表格识别

早期表格识别研究方向主要集中在非端到端的方法上,将表格识别任务分为两个不同的子任务:表结构识别和单元格内容识别,试图使用独立的模型来解决每个子问题.

Lang 等人[33] 提出了模型 TableMASTER,这是一个基于 Transformer 的模型,用于表格结构识别. 该模型将 Transformer 模型与文本行检测器相结合,以识别单个表格单元格中的文本行,其中,识别文本行时参考[9] 中的方法来提取文本内容. Nassar 等人[15] 提出另一个基于 Transformer 的模型 TableFormer,该模型可以同时识别表格结构和预测每个表单元格的边界框,利用预测的边界框从 PDF 文档中提取单元格内容.

最近,随着深度学习的进一步研究以及表格数据的扩展,研究方向逐渐转向端到端方法. Zhong 等人[34] 提出一种编码器双解码器模型 EDD,能够同时识别每个单元格的表结构和内容,除此以外,Zhong 等人还构建了 PubTabNet 数据集,以专注于表格识别任务.

2　TGMS 模型

2.1　模型概述

为了达到从表格图像重建表格图的目的,本文首先描述了所提出的用于表格图像到标记序列的 TGMS 的主要框架,如图 1 所示. TGMS 模型旨在识别图像中的表格结构,根据识别到的结构信息指导单元格内容的准确识别,最终生成表格标记序列. TGMS 模型由 5 部分组成:特征提取模块,单元格空间位置预测模块,单元格逻辑位置预测模块,单元内容识别模块,序列生成模块. 本节将首先给出问题的形式化定义,再详细介绍所提出的模型.

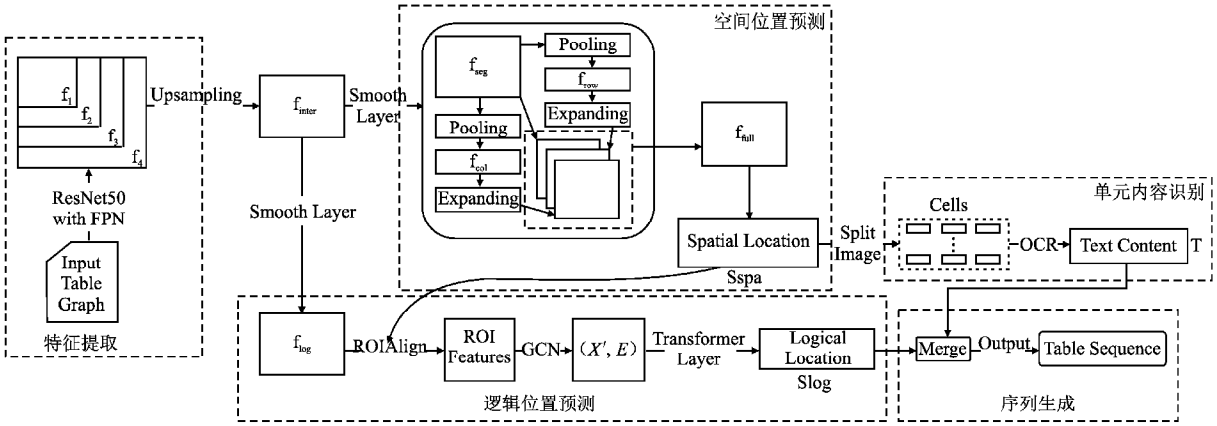

图 1　TGMS 框架

Fig. 1　TGMS framework

问题定义:给定图像 X,TSR 任务的目标是从 X 中识别出表格的结构,其中,$S=[s_1,s_2,\cdots,s_N]$,s_i 是生成的第 i 个表格标记序列. TCR 任务的目标是从 X 中识别出每个单元格的真实文本内容,其中 $T=[[t_{11},t_{12},\cdots,t_{1n}][t_{21},t_{22},\cdots,t_{2n}]\cdots[t_{m1},t_{m2},\cdots,t_{mn}]]$,$t_{ij}$ 是第 i 行第 j 列识别出的文本内容.

2.2　特征提取模块

表格识别包括表格结构识别和单元格文本内容识别两个子任务,两个子任务同时需要丰富的空间信息和语义信息,即满足更大的感受野要求的同时还需要更高的分辨率. 值得注意的是,这两个要求在实现上是相互矛盾的. 本文提出的特征提取模块以 ResNet-50[35] 作为主干网络,其堆叠多个卷积层和最大池化层,并在层与层之间使用残差连接来学习表格图像的深层特征,共有 50 层并分成 5 个阶段(stage),通过顺序执行,可以从每个阶段提取出不同层次的特征表示. 相比较而言,f_2 作为第 2 阶段的输出特征图具有小感受野和高分辨率的特点,即视觉特征丰富而语义特征不足,而 f_5 作为第 5 阶段的输出特征图的特点则是大感受野和低分辨率,即视觉特征不足而语义特征丰富. 这些在不同阶段提取到的特征可以在后续任务中使用.

2.3　单元格空间位置预测模块

基于分割的方法在表格的行与列方向上具有显著的统计特性,因此广泛应用于单元格检测问题. 本文使用基于分割的空间位置解码器(Spatial Location Decoder,SLD)来检测表格单元格的边界框,即通过识别出表格的分割线信息,进而识别出每个单元格在图像中所处的位置. 对于每个输入表格图像 $X\in R^{3\times H\times W}$(其中 H 和 W 分别表示输入图像的高度和宽度),单元格空间位置预测模块调用特征提取模块 ResNet-50 这 4 个特征图 f_1,f_2,f_3 和 f_4,步幅 $s=4,8,16,32$ 来构建特征金字塔[36]. 计算过程如公式(1)所示:

$$f_{inter}=u_{\times 4}(C(f_1,u_{\times 2}(f_2),u_{\times 4}(f_3),u_{\times 8}(f_4))) \tag{1}$$

为了解决计算复杂度过高的问题,单元格空间位置预测模块首先使用 1×1 卷积层(即 $f_{seg}\in R^{256\times H\times W}$)将输入通道从 1024 减少到 256. 同时考虑到表格数据是按行和列排列的,单元格空间位置预测模块采用分割-聚合模块来利用行和列表示的统计信息. 具体来说,行和列特征,$f_{row}\in R^{256\times H}$ 和 $f_{cot}\in R^{256\times W}$ 是分别通过使用 $1\times W$ 和 $H\times 1$ 平均池化层获得的. 然后扩展行和列特征表示 f_{row} 和 f_{col} 以与像素特征连接,计算过程如公式(2)所示:

$$f_{full}=C(f_{row},f_{col},f_{seg}) \tag{2}$$

然后进一步得到 $R^{K\times H\times W}\in\hat{y}_{full}$,其中 K 表示识别的种类数量(包括“背景”、“单元格”和“边界”). 在训练阶段,除了 $\hat{y}_{full}$,本文使用 f_{row} 和 f_{col} 来预测行和列的分割图,$\hat{y}_{row}\in R^{K\times H}$ 和 $\hat{y}_{col}\in R^{K\times W}$,可以看作是一种统计正则化. 在测试阶段,使用 $\hat{y}_{full}$ 计算分割图上每个连通分量的最小矩形边界框,以此来获取单元格的空间位置 S_{spa}.

训练期间单元格空间位置预测的损失函数计算过程如公式(3)所示:

$$L_{spa}(p(x),\hat{p}(x))=-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}w_c\cdot p(x_{i,k})log(\hat{p}(x_{i,k})) \tag{3}$$

其中,$p(x)$ 和 $\hat{p}(x)$ 分别是掩码 x 的真实概率分布和预测概率分布,$p(x_{i,k})$ 和 $\hat{p}(x_{i,k})$ 分别是像素 i 属于掩码 x 的 k 类的真实概率和预测概率;N 和 K 是像素编号和类编号,w_k 是类 k 的权重系数.

经过单元格空间位置预测模块,可以得到用于表示每个单元格在图像中的位置的特征向量(如向量 S_{spa_1} 表示第 1 个单元格边界框坐标为(0,0),(100,0),(0,50),(100,50),单位是像素 px).

2.4　单元格逻辑位置预测模块

要准确地获取单元格的结构信息,不仅要得到单元格空间位置,还必须在此基础上获取其逻辑位置. 常用的单元格逻

辑位置预测方法使用图神经网络(Graph Neural Network, GNN)来建立每个单元格之间的邻接关系,将每个单元格视作图上的顶点,通过判断单元格与单元格之间的位置关系来表达图上节点之间是否用边相连. 但该方法依靠图中的邻接边传播,难以对距离较远的单元格之间的信息进行建模(特别是表格跨页、合并单元格等情况). 针对表格全局单元格之间关系的获取问题,TGMS 提出逻辑位置解码器(Logical Location Decoder,LLD)来着重获取表格的逻辑信息. 逻辑位置解码器 LLD 主要包含两个模块:用于提取单元格局部结构依赖的图卷积网络(Graph Convolutional Network,GCN)以及用于全局建模的 Transformer 块.

本文将表格表示为一个单元格相邻关系的无向图,从检测到的单元格边界框中选择候选表格单元格来初始化表格图. 本文使用 $G=(V,E)$ 表示这个图,其中 $V=\{v_1,v_2,\cdots,v_n\}$ 表示图中节点集合,v_i 代表表格中第 i 个单元格,$E\in R^{n\times n}$ 是一个邻接矩阵,表示所有的边集合,$e_{i,j}$ 为 0 代表表格中第 i 个单元格与第 j 个单元格不相邻,否则相邻. 通过图卷积网络来学习该表格图表示.

图 G 的构造方法如下:在训练过程中,节点集合 V 由所有候选表单元格组成,这些单元格与任一真实表单元格的交并比(Intersection-over-Union,IoU)大于 0.5. IoU 计算过程如公式(4)所示:

$$IoU(R_1,R_2)=\frac{a(R_{12})}{a(R_1)+a(R_2)-a(R_{12})} \tag{4}$$

其中,$a(\cdot)$ 表示区域的面积,R_{12} 表示两个区域 R_1 和 R_2 的交集.

节点 v_i 包括两部分特征:空间位置特征 x_i^s 和视觉外观特征 x_i^v,其中 $x_i^s\in R^{256}$ 是由 SLD 提取得到;而使用 RoIAlign[37] 操作(输出大小为 2×2),根据第 i 个表格单元格的边界框 b_i,从平滑多尺度特征图 $f_{log}\in R^{256\times H\times W}$ 中得到 $x_i^v\in R^{1024}$. 对于邻接矩阵 E,通过计算两个节点之间的欧几里得距离得到每个元素 $e_{i,j}=\{e_{i,j}^{row},e_{i,j}^{col}\}$,计算过程如公式(5)、公式(6)所示:

$$e_{i,j}^{row}=e^{-\left(\frac{b_i^y-b_j^y}{\alpha\frac{H}{H}}\right)^2} \tag{5}$$

$$e_{i,j}^{col}=e^{-\left(\frac{b_i^x-b_j^x}{\alpha\frac{H}{H}}\right)^2} \tag{6}$$

其中,α 是调整因子,其随着行(列)数的增加而增加,从而保证同样大小的单元格在行(列)数增加的同时还能保持较强的相邻关系.

在图 G 构建完成后,使用图卷积网络 GCN 进行消息传递,计算过程如公式(7)所示:

$$X'=ReLU(GCN(X,E)) \tag{7}$$

其中,X 和 X' 分别表示输入和输出节点特征矩阵,E 为节点的邻接矩阵. 相对于不支持对边进行多维表示的标准 GCN 架构,本文采用一对并行 GCN 来更新节点特征矩阵,分别用于行和列索引预测.

经典 Transformer[14] 提出的自注意力机制可以通过计算输入向量之间的相关性,进而选择性地关注整个序列中的重要信息,从而完成对全局信息的捕获. 本文所采用的 Transformer 块分为编码器和解码器两个部分,编码器将 GCN 的输出节点特征矩阵 X' 作为输入,并通过多头点积注意力层对其进行细化,而解码器接收 Transformer 编码器生成的输出特征作为输入,使用多头注意力通过将不同子空间中注意力头的分数组合在一起,来推断出表格的逻辑结构信息 S_{log}.

2.5 单元内容识别模块

单元内容识别模块(Cell Content Recognition,CCR)的作用是将图像中的文本信息准确识别出来,用于表格序列的生成. 与自然场景中的文本识别任务不同,识别单元格内的文本内容需要同时关注文本所在的单元格位置(例如所在的单元格内文本是否多行). 单元格内容识别模块根据前述工作提取出的单元格空间位置信息,将输入图像进行裁剪,对裁剪后的局部区域内的图像执行 OCR 操作,将局部区域内识别的文本视作文本块整体,可以有效解决多行文本识别时出现多个文本块的问题.

2.6 序列生成模块

经过单元格逻辑位置预测模块和单元内容识别模块后,可以得到表格的逻辑结构信息 S_{log} 和文本信息 T. 序列生成模块的任务是利用 S_{log} 和 T,生成表格图像相对应的 HTML 表结构序列并在其中适当位置置入单元格文本信息,以生成完整的标记序列.

训练期间序列生成的损失函数计算如公式(8)所示:

$$L_{seq}=-\sum_{t\in T}logP(y_t|y_{<t}) \tag{8}$$

其中,T 是整个序列长度,y_t 是真值标签 token,$y_{<t}$ 是前 $t-1$ 个时间步的输出序列.

2.7 损失函数

训练期间 TGMS 的总损失是所有模块损失的加权和,计算过程如公式(9)所示:

$$L=\lambda_1L_{spa}+\lambda_2L_{seq} \tag{9}$$

其中 L_{spa} 是单元格空间位置预测损失,L_{seq} 是序列生成损失,λ_1 和 λ_2 是是相应的权重系数.

3 实验与结果分析

3.1 实验设置

本节首先介绍了数据集与实验设置. 接着,指出实验所使用的评价指标. 最后,将提出的 TGMS 模型与其他基准模型进行对比实验.

3.1.1 数据集

为验证所提出的 TGMS 模型在不同场景下的性能,TGMS 在 3 个广泛使用的公共数据集上进行实验,分别是 ICDAR-2013、SciTSR、PubTabNett 数据集. 具体描述如下:

1) ICDAR-2013 数据集[38]:该数据集由 156 个具有跨单元格和其他不同样式的表格组成. 行和列的最大索引分别为 57 和 12.

2) SciTSR 数据集[31]:该数据集包含 12000 张训练图像和 3000 张从科学文献 PDF 中裁剪的测试图像. 为了评估不同方法在复杂表上的性能,作者还将测试集中的所有 716 个复杂表提取为测试子集,称为 SciTSR-COMP.

3) PubTabNet 数据集[34]:该数据集是一个大规模的表格图像数据集,从 PubMed Central Open Access Subset(PMCOA)收集科学文章创建,包含超过 568k 个样本,以及以 HTML 格式呈现的表格结构的相应注释、文本内容以及每个非空表格单元格的边界框. 该数据集包含大量具有多行(列)

单元格、空单元格等的三行表,由 500777 张训练图像,9115 张验证表图像以及 9064 张测试图像组成.

3.1.2　环境设置与实验参数设置

本实验使用的开发环境是 python 3.7.0,开发框架是 pytorch1.7.0 + cuda9.2,GPU 为显存大小为 32G 的 Nvidia V100.考虑到 GPU 显存大小有限,将输入的表图像大小调整为 800×800 像素.本文使用 Adam[39] 优化器进行训练,实验的部分参数设置如表 1 实验识部分参数设置所示.

表 1　实验识部分参数设置

Table 1　Experimental parameter settings

参　数	值
batch size	8
learning rate	0.0001
weight decay	0.0005
epochs	20
max steps	250000
max length	1376

与基准方法相比,LGPMA 学习率设置为 0.01,batch size 设置为 4;EDD 将学习率设置为 0.001,batch size 设置为 10;ReS2TIM 学习率采用 0.00005;SEM 设置最小学习率 0.00001,最大学习率 0.0001,并使用公式(10)进行更新:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{T_{cur}}{T_{\max}}\pi\right)\right) \quad (10)$$

本文实验采用 5 个不同的随机种子进行测试,并报告平均结果.每次测试保留在训练集上效果最好的模型,然后在测试集上进行测试.

3.1.3　评价指标

1)单元邻接关系度量[38](Cell Adjacency Relationship Metric):生成每个内容单元格与其最近的水平和垂直邻居之间的所有邻接关系的列表,邻接关系是一个元组,其中包含两个单元格的文本内容、方向和中间空白单元格(如果存在)的数量,而在空白单元格之间或空白单元格与内容单元格之间不会生成邻接关系.这个邻接关系的一维列表可以通过使用精确率(precision)、召回率(recall)和 F1-measure 值进行度量.

2)树编辑距离分数[34](Tree-EditDistance-based Similarity,TEDS):由于 HTML 格式的表格可以使用树结构来进行表示,因此[34]首先提出可以将预测结构和真实结构表示为 HTML 标记的树结构,使用树编辑距离[40]来测量预测表和真实表之间的相似性,并使用该指标对 PubTabNet 数据集进行了实验.TEDS 的计算过程如公式(11)所示:

$$TEDS(T_a, T_b) = 1 - \frac{EditDist(T_a, T_b)}{max(|T_a|, |T_b|)} \quad (11)$$

其中 EditDist 表示树编辑距离,T_a 和 T_b 表示树形结构 HTML 格式的表格,而|T|是 T 中的节点数.

3.2　对比方法

为验证所提出方法的有效性,本文与以下基准方法进行对比实验:

1)ReS2TIM[41]:一种基于图的表格识别框架,将每个单元格在 4 个方向上的邻居构建一个单元关系网络,使用一个加权图表示检测到的关系,然后使用该图推断语法表结构.

2)EDD[34]:一种基于注意力的编码器双解码器架构,包含结构解码器和单元格解码器,将表格图像转换为 HTML 代码.

3)GTE[25]:一种视觉引导的系统化框架,用于表格检测和单元格结构识别,该框架基于表格的自然单元格包含约束,设计一种新惩罚机制,结合单元位置预测来训练表格识别网络.

4)LGPMA[11]:使用软金字塔掩码学习机制提取局部和全局特征图,集成金字塔掩码重新评分模块,设计一种表结构恢复管道来获得最终的表格结构.

5)SEM[42]:由分割器、嵌入器和合并器 3 部分组成.通过分割表格网格、融合视觉与文本特征,采用注意力机制实现自回归式网格合并.

3.3　实验结果

实验结果如表 2 和表 3 所示,其中加粗字体表示最优值,下划线表示次优值.从表格中可以观察到,本文所提模型在所有的数据集上都优于基准方法.

表 2　TGMS 在 ICDAR 2013、SciTSR、SciTSR-COMP 数据集的结果

Table 2　Results of TGMS on ICDAR 2013、SciTSR、SciTSR-COMP datasets

方法	ICDAR-2013			SciTSR			SciTSR-COMP		
	P	R	F1	P	R	F1	P	R	F1
Res2TIM	73.4	74.7	74.0	-	-	-	-	-	-
GTE	94.4	92.7	93.5	-	-	-	-	-	-
LGPMA	<u>96.7</u>	**99.1**	<u>97.9</u>	<u>98.2</u>	<u>99.3</u>	<u>98.8</u>	<u>97.3</u>	<u>98.7</u>	<u>98.0</u>
SEM	-	-	-	97.7	96.5	97.1	96.8	94.7	95.7
TGMS	**98.5**	<u>98.8</u>	**98.6**	**99.0**	**99.4**	**99.2**	**98.2**	**99.1**	**98.6**

表 3　TGMS 在 PubTabNet 数据集上的结果

Table 3　Results of TGMS on PubTabNet

方法	TEDS
EDD	88.3
GTE	93.0
LGPMA	<u>94.6</u>
SEM	93.7
TGMS	**96.4**

实验首先在 ICDAR-2013 数据集和 SciTSR 数据集上进行,评估指标采用单元邻接关系度量(即使用精确率、召回率和 F1-measure 值作为评估指标).在该数据集上与 Res2TIM、GTE、LGPMA 和 SEM 这 4 个方法进行对比.实验结果如表 2 所示,其中基线方法数据来自论文中报告的结果.表中可以看到,所提出的 TGMS 的 F1-measure 值均优于基准方法.在 SciTSR 数据集上,TGMS 的 F1-measure 值达到了 99.2%,而在 SciTSR-COMP 数据集上,由于是从 SciTSR 数据集中选取了相对复杂的表格,识别性能略有下降,F1-measure 值达到了 98.6%,下降了 0.6%,但 TGMS 下降幅度仍低于其他基准方

法. 在 ICDAR 数据集上,本文提出的方法 F1-measure 值为 98.6%,在所有方法中取得了最优结果,但在召回率指标上,TGMS 稍低于 LGPMA 方法,说明 TGMS 识别略微保守.

此外,在较为复杂的 SciTSR-COMP 数据集上,TGMS 性能并没有明显下降,说明具有更好的空间位置感知能力.

表 3 是在更为复杂的 PubTabNet 数据集上进行实验的结果.

Zhong 等人[34] 论文中提出方法 EDD,构建出数据集 PubTabNet,首次提出新的评估标准 TEDS 来同时考虑表格结构和文本内容. 因此,在 PubTabNet 数据集上的实验并没有采用与 ICDAR 和 SciTSR 数据集一样的评估标准. 本文提出的 TGMS 在 PubTabNet 数据集 TEDS 上达到了 96.4%,优于其他方法,与获得次优结果的 LGPMA 方法相比,分数提高了 1.8%.

3.4 可视化结果

在本节展示 TGMS 在在 PubTabNet 数据集上可视化结果. 如图 2 和图 3 所示. 图 2 是检测到表格单元格边界框的原始图像,虚线框表示检测到表格单元格的边界框.

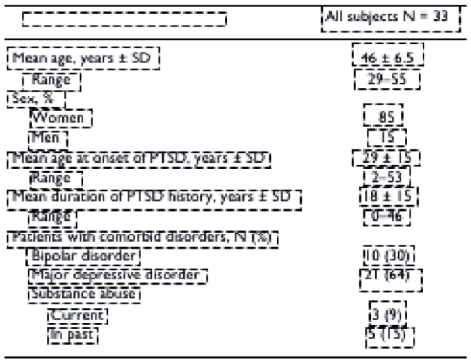

图 2 PubTabNet 上的可视化结果

Fig. 2 Visualization results on PubTabNet

	All subjects N=33
Mean age, years±SD	46±6.5
Range	29–55
Sex, %	
Women	85
Men	15
Mean age at onset of PTSD,years±SD	29±15
Range	2–53
Mean duration of PTSD history,years±SD	18±15
Range	0–46
Patients with comorbid disorders,N(%)	
Bipolar disorder	10(30)
Major depressive disorder	21(64)
Substance abuse	
Current	3(9)
In past	5(15)

图 3 在 Web 浏览器上查看的表格预测 HTML 代码

Fig. 3 Predicted HTML code of the table viewed on the Web browser

图 3 是在 Web 浏览器上查看的表格预测 HTML 代码. 根据结果显示,TGMS 可以预测包括空单元格和跨列单元格在内的复杂表格结构.

3.5 消融实验

在完成了相应的对比实验后,本节从 PubTabNet 数据集中选择 60000 张训练图像和 1000 张验证图像用于消融实验,以系统分析本文提出的 TGMS 方法,其结果如表 4 所示,其中(w/o)代表删除该模块后的模型. 从结果中可以看出,删除不同的模块后,相对于 TGMS 会出现不同程度的识别性能下降.

表 4 消融实验结果

Table 4 Ablation experiment results

模 型	TEDS
LLD-GCN(w/o)	85.2
LLD-attention(w/o)	88.6
CCR-SLD(w/o)	90.4
TGMS	**96.4**

LLD-GCN(w/o)代表在逻辑位置解码器 LLD 中删除 GCN 构建单元格局部信息的过程,直接对空间位置解码器 SLD 的输出添加注意力机制,以此识别表格的逻辑结构,识别结果出现了显著下降. 这说明没有局部结构信息的情况下,全局注意力并不能准确提取出表格结构.

LLD-attention(w/o)代表在 LLD 中删除全局注意力模块,只利用 GCN 提取出的单元格局部信息来捕获表格逻辑结构信息,识别性能下降. 说明失去全局信息后,逻辑结构识别时会出现无法识别跨行或不规则表格结构.

CCR-SLD(w/o)代表在单元内容识别时删除前置步骤获得的空间位置信息. 识别性能出现下降说明在识别文本内容时,如果不关注所在的单元格位置,会导致识别出的多个文本块不能置于表格中正确的行(列),即出现单元格和多个文本块不能正确对应问题.

4 结论和未来工作

本文提出了一种端到端表格图像到标记序列的识别框架 TGMS. 首先在空间位置预测模块利用特征提取模块提取到的图像特征构建特征金字塔,以此获取单元格空间特征向量;其次,在逻辑位置预测模块通过构建表图来提取表格内部的局部结构关系,使用注意力机制提取全局单元格依赖;然后,利用单元格空间位置信息准确识别表格文本内容;最后,根据表格逻辑结构和文本内容得到完整的表格序列. 实验结果表明,TGMS 在 3 个公共数据集上的指标都得到了提高,说明了方法的有效性. 本文提出的方法需要获得文档中表格的准确位置,在后续的研究工作中,将进一步探索 TGMS 与文档表检测模块的集成,以实现端到端的表检测和识别.

References:

[1] KONG L J, BAO Y C, WANG Q W, et al. A summary of table detection and recognition algorithms based on deep learning[J]. Computer and Network, 2021, 47(2): 65-73.

[2] Yoon J, Zhang Y, Jordon J, et al. Vime: extending the success of self-and semi-supervised learning to tabular domain[C]//Advances in Neural Information Processing Systems, 2020: 11033-11043.

[3] Ramel J Y, Crucianu M, Vincent N, et al. Detection, extraction and representation of tables[C]//7th International Conference on Document Analysis and Recognition, 2003: 374-378.

[4] Yildiz B, Kaiser K, Miksch S. pdf2table: a method to extract table information from PDF files[C]//Indian International Conference on Artificial Intelligence, 2008, doi: US591830 A.

[5] Hassan T, Baumgartner R. Table recognition and understanding from PDF files[C]//9th International Conference on IEEE, 2007, doi: 10.1109/ICDAR.2007.4377094.

[6] Jane Hoffswell, Zhicheng Liu. Interactive repair of tables extracted from PDF documents on mobile devices[C]//Proceedings of the CHI Conference on Human Factors in Computing Systems, 2019: 1-13.
[7] Itonori K. Table structure recognition based on textblock arrangement and ruled line position[C]//International Conference on Document Analysis & Recognition, 1993, doi: 10. 1109/ICDAR. 1993. 395625.
[8] Kieninger T G. Table structure recognition based on robust block segmentation[C]//Document Recognition V, 1998: 22-32.
[9] Wang Y, Phillips I T, Haralick R M. Table structure understanding and its performance evaluation[J]. Pattern Recognition, 2004, 37(7): 1479-1497.
[10] Schreiber S, Agne S, Wolf I, et al. DeepDeSRT: deep learning for detection and structure recognition of tables in document images[C]//14th IAPR International Conference on Document Analysis and Recognition(ICDAR), 2017, doi: 10. 1109/ICDAR. 2017. 192.
[11] Qiao L, Li Z, Cheng Z, et al. Lgpma: complicated table structure recognition with local and global pyramid mask alignment[C]// International Conference on Document Analysis and Recognition, 2021: 99-114.
[12] Prasad D, Gadpal A, Kapadni K, et al. CascadeTabNet: an approach for end to end table detection and structure recognition from image-based documents[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 572-573.
[13] Paliwal S S, Vishwanath D, Rahul R, et al. Tablenet: deep learning model for end-to-end table detection and tabular data extraction from scanned document images[C]//International Conference on Document Analysis and Recognition(ICDAR), 2019: 128-133.
[14] Ashish V, Noam S, Niki P, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, 2017: 5998-6008.
[15] Nassar A, Livathinos N, Lysak M, et al. Tableformer: table structure understanding with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 4614-4623.
[16] Jiaquan Y, Xianbiao Q, Yelin H, et al. PingAn-VCGroup's solution for ICDAR 2021 Competition on Scientific literature parsing task B: table recognition to HTML[J]. Computing Research Repository, 2021, doi: 10. 48550/arXiv. 2105. 01848.
[17] Nam T L, Atsuhiro T, Phuc N, et al. Rethinking image-based table recognition using weakly supervised methods[C]//International Conference on Pattern Recognition Applications and Methods, 2023: 872-880.
[18] Nam Tuan Ly, Atsuhiro Takasu. An end-to-end multi-task learning model for image-based table recognition[C]//Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory andApplications, 2023: 626-634.
[19] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.
[20] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640-651.
[21] Khan S A, Khalid S M D, Shahzad M A, et al. Table structure extraction with bi-directional gated recurrent unit networks[C]//International Conference on Document Analysis and Recognition(ICDAR), 2019: 1366-1371.
[22] Siddiqui S A, Khan P I, Dengel A, et al. Rethinking semantic segmentation for table structure recognition in documents[C]//International Conference on Document Analysis and Recognition(ICDAR), 2019, doi: 10. 1109/ICDAR. 2019. 00225.
[23] Tensmeyer C, Morariu V I, Price B, et al. Deep splitting and merging for table structure decomposition[C]//International Conference on Document Analysis and Recognition(ICDAR), 2019, doi: 10. 1109/ICDAR. 2019. 00027.
[24] Siddiqui S A, Fateh I A, Rizvi S T R, et al. DeepTabStR: deep learning based table structure recognition[C]//International Conference on Document Analysis and Recognition(ICDAR), 2020, doi: 10. 1109/ICDAR. 2019. 00226.
[25] Zheng X, Burdick D, Popa L, et al. Global table extractor(GTE): a framework for joint table identification and cell structure recognition using visual context[C]//Workshop on Applications of Computer Vision, 2021, doi: 10. 1109/WACV48630. 2021. 00074.
[26] Sachin Raja, Ajoy Mondal, Jawahar C V. Table structure recognition using top-down and bottom-up cues[C]//European Conference on Computer Vision, 2020, doi: 10. 1007/978-3-030-58604-1_5
[27] Li M, Cui L, Huang S, et al. Tablebank: table benchmark for image-based table detection and recognition[C]//Proceedings of the 12th Language Resources and Evaluation Conference, 2020: 1918-1925.
[28] Deng Y, Rosenberg D, Mann G. Challenges in end-to-end neural scientific table recognition[C]//International Conference on Document Analysis and Recognition(ICDAR), 2019, doi: 10. 1109/ICDAR. 2019. 00148.
[29] Huang Y, Lu N, Chen D, et al. Improving table structure recognition with visual-alignment sequential coordinate modeling[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 11134-11143.
[30] Huawen S, Xiang G, Jin W, et al. Divide rows and conquer cells: towards structure recognition for large tables[C]//International Joint Conference on Artificial Intelligence, 2023: 1369-1377.
[31] Chi Z, Huang H, Xu H D, et al. Complicated table structure recognition[J]. arXiv preprint arXiv: 1908. 04729, 2019.
[32] Hao L, Xin L, Bing L, et al. Show, read and reason: table structure recognition with flexible context aggregator[C]//ACM International Conference on Multimedia, 2021: 1084-1092.
[33] Lang Cao, Hanbing Liu. TableMaster: a recipe to advance table understanding with language models[J]. CoRR, 2025, doi: 10. 48550/arXiv. 2501. 19378.
[34] Zhong X, Shafieibavani E, Yepes A J. Image-based table recognition: data, model, and evaluation[C]//European Conference on Computer Vision, 2020, doi: 10. 1007/978-3-030-58589-1_34.
[35] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[36] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.
[37] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.
[38] Göbel M, Hassan T, Oro E, et al. ICDAR 2013 table competition[C]//12th International Conference on Document Analysis and Recognition, 2013: 1449-1453.
[39] Diederik P Kingma, Jimmy Ba. Adam: a method for stochastic optimization[C]//International Conference on Learning Representations, 2014, doi: 10. 48550/arXiv. 1412. 6980.
[40] Pawlik M, Augsten N. Tree edit distance: robust and memory-efficient[J]. Information Systems, 2016, 56: 157-173, doi: 10. 1016/j. is. 2015. 08. 004.
[41] Xue W, Li Q, Tao D. Res2tim: reconstruct syntactic structures from table images[C]//International Conference on Document Analysis and Recognition(ICDAR), 2019: 749-755.
[42] Zhenrong Z, Jianshu Z, Jun D, et al. Split, embed and merge: an accurate table structure recognizer[J]. Pattern Recognition, 2022, 126: 108565-108565, doi: 10. 1016/j. patcog. 2022. 108565.

附中文参考文献:

[1] 孔令军,包云超,王茜雯,等. 基于深度学习的表格检测识别算法综述[J]. 计算机与网络,2021,47(2):65-73.