多单元解码器与互学习在表格结构与字符识别中的应用

川胜隆哉[1]

Preferred Networks, Inc.，日本东京都千代田区大手町1-6-1。

kat.nii.ac.jp@gmail.com

https://researchmap.jp/t.kat

摘要。从科学论文和财务报告等文档中提取表格内容并将其转换为大型语言模型可处理的格式是知识信息处理中的重要任务。端到端方法不仅识别表格结构，还识别单元格内容，其性能已达到使用外部字符识别系统的最先进模型的水平，并具有进一步改进的潜力。此外，这些模型通过引入局部注意力机制，现在能够识别包含数百个单元格的长表格。然而，这些模型在从表头到表尾的单向方向上识别表格结构，并且对每个单元格的内容识别是独立进行的，因此没有机会从相邻单元格中检索有用信息。本文提出了一种多单元内容解码器和双向互学习机制，以改进端到端方法。该方法在两个大型数据集上进行了验证，实验结果表明，即使在包含大量单元格的长表格中，其性能也与最先进的模型相当。

关键词：深度学习，表格识别，Transformer，互学习

1Introduction

提供高质量知识给大型语言模型（LLMs）的信息检索技术正受到关注。许多研究人员致力于将扫描和图像文档转换为机器可读格式，如HTML代码[37,42,15]和LaTeX代码[3,12]。这一举措既有直接也有间接的好处。首先，由于过去的文献大多以印刷形式存在，将其转换为结构化的电子文档是必要的，这是直接的好处。其次，从人机交互的角度来看，实现能够识别出版文档布局中隐藏意义的智能是重要的，这是间接的好处。

在本文中，我们将专注于表格识别，这包括两类任务，即结构识别和单元格内容识别。一个简单的表格具有水平和垂直边框，每个单元格包含字符。复杂的

表格可能包含垂直或水平合并的单元格，和/或涉及隐形边框的单元格。人们即使在没有明确边界的情况下，也能从单元格布局中理解表格结构，这是一个具有挑战性的问题。

近年来，随着Transformer[31]模型在语言和视觉识别任务中的成功应用，许多基于Transformer的方法[21,36,18,19]被提出用于表格识别任务。由于我们可以利用外部光学字符识别（OCR）系统来解析单元格内容，因此我们可以主要关注表格结构识别。该任务利用图像特征与HTML标记嵌入表示之间的交叉注意力来顺序预测HTML标记。在以往的研究[21,36,18,19,20]中，标记预测是从表头到表尾、从左到右单向进行的。这限制了提前关注表格结构的机会。

Ly 和 Takasu [19] 报告称，端到端学习表格结构和单元格内容识别任务可能提高整体表格识别性能。此外，表格可能包含数百个单元格，而序列预测方法可能会表现不佳，通过引入局部注意力机制 [18] 可以改善这一问题。这些先前的研究在结构识别后独立地对每个检测到的单元格进行内容识别。这限制了从相邻单元格获取有用信息的机会。

作为解决问题的方案，我们改进了端到端的方法[18,19,20]，并提出了一种方法，该方法参考了邻近单元格的识别结果，并结合了关注前后单元格的学习机制。前者通过引入一个推断多个单元格的单元格解码器，并配置一个层次解码器和一个HTML解码器用于结构识别来实现。后者通过正向解码器（从左到右读取表格结构）和反向解码器（从右到左读取表格结构）之间的相互学习[38]来实现。所提出方法的有效性通过两个大规模表格图像数据集进行了验证。

本文的主要贡献如下：1) 我们提出了一种细胞解码器，能够推断多个细胞并从周围细胞中获取有用信息。2) 我们提出了一种双向互学习机制，以迫使所提出的模型同时关注前后的细胞。3) 在所有实验结果中，我们提出的方法表现优于现有最先进模型。

2相关工作

一般来说，表格识别任务分为两个子任务，即表格结构识别和单元格内容识别。当然，最终的输出是一个HTML [37,42,15] 或 LaTeX [3,12] 文档，因此无需区分这两个子任务。然而，最好使用单独的模型来识别标签（或命令）和其他可见字符。对于单元格内容识别任务，现有的高精度OCR系统 [17] 已经可用，因此之前的研究 [11,13,32,26,28,30] 主要集中在表格结构识别上。

表格结构识别已经研究了很长时间，早期的方法基于手工设计的特征和启发式规则[11,13,32]，但它们的应用局限于简单表格或具有预定义模式的表格。随着深度学习的发展，自动学习表格结构模式的方法[39,27,26,28,30]已成为主流。这些研究可以分为基于目标检测和分割的方法[30,27]，以及基于序列标记预测的方法[21,36]。

对于检测和分割方法，Schreiber 等人 [30] 提出了一个双系统，使用 Faster R-CNN [29] 和全卷积网络 [16] 分别进行表格检测和表格结构识别。Raja 等人 [28] 提出了一种两阶段模型，在识别单元格位置后估计它们之间的关系。Qiao 等人 [27] 通过结合文本、单元格、行和列识别任务，使用 Mask R-CNN [5] 在 ICDAR 竞赛 [37] 中获得了第一名。

对于顺序令牌预测方法，可以利用简单的图像描述模型进行细胞检测，因为细胞的顺序是唯一确定的。Ye等人[36]和Nassar等人[21]提出了带有两种解码器的Transformer模型，用于表格结构识别和细胞定位。Peng等人[25]通过引入卷积主干，实现了与使用深度卷积编码器模型相媲美的性能，同时显著减少了参数。

在2020年代，研究人员正在研究端到端模型，这些模型能够同时学习表格结构和单元格内容识别任务[3]。Zhong等人[42]提出了一种模型，该模型使用ResNet[6]编码器和两个LSTM[8]解码器来识别表格结构和单元格内容，但其性能不如使用外部OCR的模型。

Ly和Takasu [19] 提出了一种多任务模型，该模型能够检测表格结构、单元格位置以及单元格内容。他们的模型采用带有全局上下文注意力机制的ResNet编码器[2]和两个Transformer解码器。第一个解码器按顺序推断HTML标记，然后第二个解码器逐一读取单元格内容。该模型在性能上可与使用外部OCR的模型相媲美。他们还提出了弱监督学习，以降低准备边界框训练数据的成本[20]，并引入了局部注意力机制[1]，以有效识别包含大量单元格的表格[18]。

在2021年，ICDAR 2021举办了科学文献解析竞赛[37]。竞赛包括文档布局识别任务A和表格识别任务B。任务B要求将表格图像转换为带有单元格内容的HTML标签。为此任务提供了PubTabNet[42]数据集和最终评估数据集。训练数据集包括HTML标记、单元格文本和单元格边界框。共有30支队伍提交了30份作品，其中大多数前十名的解决方案利用了单独的OCR模型、额外的注释和集成技术。

TabRecSet [35] 是一个包含旋转和扭曲表格的双语数据集，适用于三个任务：表格检测、结构识别和单元格内容识别。此类表格的检测不在本文的讨论范围内。

3Background

与先前的工作[18,19,20]类似，我们提出的模型采用了ResNet编码器和一个由多个注意力块[31]组成的HTML解码器来推断表示表格结构的HTML标记。此外，还利用了一个额外的解码器来推断单元格内容。编码器和两个解码器通过端到端的方式同时进行训练。在本节中，我们将介绍第4节中描述的所提出方法使用的一些技术。

3.1Encoder

先前的研究[21,36,18]采用卷积神经网络（CNN）来提取图像特征，并将这些特征输入到解码器中。CNN在识别小字符的同时，能保留如字符位置等局部信息，减少图像特征的大小，从而提升解码器的计算效率和性能。

卷积层的数量对识别性能有贡献，

并且已经探索了许多衍生物来增加这一数量。ResNet [6]，由大量包含简单跳跃连接的多卷积层残差块组成，已被广泛使用。此外，提出了带有分组卷积的ResNeXt [34]和在所有卷积层之间具有更复杂跳跃连接的DenseNet [9]。

CNN的一个弱点是其识别全局上下文的能力较差，这是因为它过于关注局部特征。为此，提出了一个全局上下文注意力（GCA）模块[2]，其定义如式（1）所示。

$$\boldsymbol{y}_{ij} = \boldsymbol{x}_{ij} + W_3 \max(0, \mathrm{LayerNorm}(W_2 \sum_i \sum_j \mathrm{SoftMax}(W_1 \boldsymbol{x}_{ij})\ \boldsymbol{x}_{ij})),$$

其中，i, j 是像素索引，x, y 分别是输入和输出像素。

W1, W2, W3 是三个线性层的权重矩阵。LayerNorm 表示层归一化。softmax 函数的定义如下。

$$\mathrm{SoftMax}(\boldsymbol{z}_{ij}) = \frac{\exp \boldsymbol{z}_{ij}}{\sum_m \sum_n \exp \boldsymbol{z}_{mn}},$$

其中 m, n 是像素索引。GCA 模块应放置在某些残差（或密集）模块之间。

3.2Decoder

Transformer [31] 在语言建模和视觉识别任务中均表现出色。与包括长短期记忆 [8] 在内的递归神经网络相比，Transformer 本身不涉及递归，从而允许对序列输入和输出数据进行并行处理。需要注意的是，除非预测长度固定，否则会进行递归的顺序推断。

然而，Transformer不需要递归来识别序列的上下文，避免了梯度消失问题，并提供了更好的性能。

Transformer 的核心思想被称为缩放点积注意力。设 X 为一个长度为 lx 和 dx 通道的序列，Y 为另一个输入序列。对于自注意力机制，X 和 Y 是相同的序列，Transformer 在处理 X 时会关注 X 的其他部分。对于交叉注意力机制，X 和 Y 属于不同的领域，Transformer 在处理 X 时会关注 Y 的某些部分。这些机制使得 Transformer 能够学习序列数据的上下文以及视觉和语言领域之间的关系。

注意力层首先根据式(3)从X、Y生成查询Q、键K和值V。

$$\begin{aligned}\boldsymbol{q}_i &= W_Q\ \boldsymbol{x}_i,\\ \boldsymbol{k}_j &= W_K\ \boldsymbol{y}_j,\\ \boldsymbol{v}_j &= W_V\ \boldsymbol{y}_j,\end{aligned}$$

其中，i, j 是序列索引，q, k, v, x, y 分别是 Q, K, V, X, Y 的元素。WQ, WK, WV 是投影矩阵。

注意力层的输出Z由公式（4）定义。

$$Z_i = W_Z\ \mathrm{SoftMax}(\frac{QK^\top}{\sqrt{d_k}})V,$$

其中 WZ 是输出投影矩阵，dk 是 k 的维度。这是缩放点积注意力的机制 [31]。

在实际应用中，注意力层被分为若干组，每组独立进行注意力计算，并在最后将输出结果合并。这种方法被称为多头注意力机制 [31]。通过上述机制，Transformer 能够聚焦于 Y 的特定值，并将这些值融入到 X 序列中。

3.3 局部注意力

尽管Transformer在识别长序列方面相比循环神经网络具有更优越的能力，但在处理极长序列时仍表现不佳。局部注意力（Local Attention，LA）[1]是一种专为处理这种长序列而设计的Transformer技术。

设 Mij 为聚焦于 X 的第 i 个元素中的第 j 个元素的掩码。局部注意力层的输出由式 (5) 定义，涉及因果掩码以防止后续元素的泄漏。

$$Z_i = W_Z\ \mathrm{SoftMax}(\frac{QK^\top}{\sqrt{d_k}} + M_{ij})V.$$

掩码矩阵M由公式(6)给出。

$$M_{ij} = \begin{cases} 0 & 0 \le i - j \le w,\\ -\infty & \text{otherwise},\end{cases}$$

其中，i, j 是序列索引，w 是滑动窗口的宽度。

3.4 位置编码

Transformer [31] 本身在识别序列中每个元素的位置方面能力较差，因此必须显式提供位置信息。与直接输入简单位置值不同，提出了两种方法，即位置嵌入 [4] 和位置编码 [31]。一般来说，后者在小规模训练数据集上表现更佳。

位置编码的输出 \(p(n) \) 由公式 (7) 定义。

$$\boldsymbol{p}(n) = \begin{pmatrix} \vdots \\ \sin \frac{n}{10000^{\frac{2k}{d}}} \\ \cos \frac{n}{10000^{\frac{2k}{d}}} \\ \vdots \end{pmatrix}, \text{ where } k \in \left[0, \frac{d}{2}\right). \tag{7}$$

p(n) 必须直接添加到具有 d 个通道的特征向量 x(n) 中。

如果序列 X 具有二维位置 (i, j)，Zhao 等人 [40] 提出的二维位置编码可能是一个更好的选择。它将水平和垂直坐标归一化到 [0, 1] 范围内，使用公式 (7) 分别进行编码，然后将它们组合起来得到一个单一的向量。第 i, j 个像素的位置编码由公式 (8) 给出。

$$\boldsymbol{p}_{2\mathrm{D}}(i,j) = \begin{pmatrix} \boldsymbol{p}\left(\frac{i}{H}\right) \\ \boldsymbol{p}\left(\frac{j}{W}\right) \end{pmatrix}, \tag{8}$$

其中，H、W 分别表示用于位置归一化的高度和宽度。在本文中，我们省略了这种归一化处理。

3.5 相互学习

集成学习常用于通过平均或互补多个推理模型的输出来提升机器学习的泛化性能和拟合精度。然而，由于参数数量庞大，特别是对于深度学习方法，其计算成本比单一模型更高。

为了仅使用单一模型实现类似效果，知识蒸馏[7]可能是一个替代方案。这是一种利用大型、复杂的神经网络（即集成模型）作为教师，小型、简单的模型作为学生，以获得比仅使用真实数据训练学生模型更高的性能的技术。

互学习[38]可能是另一种解决方案。在此方案中，多个学生模型同时进行训练，彼此互相教授，而无需事先训练一个教师模型。具体而言，每个学生模型使用真实数据进行监督学习，并最小化Kullback–Leibler（KL）散度[14]，以使彼此的分类输出分布相匹配。

3.6Metrics

Zhong等人[42]提出了一种基于树编辑距离的相似度（TEDS）度量方法，用于评估表格结构和单元格内容识别的性能。在将识别结果和真实值转换为HTML标签的树结构后，根据公式（9）计算TEDS得分。

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

其中 Ta 和 Tb 是 HTML 树，EditDist 是编辑距离函数，|T| 是 T 中的节点数。TEDS 有两种版本，即结构 TEDS 和总体 TEDS。

前者计算的是不包括单元格内容的HTML树，仅表示表格结构的识别性能。后者计算的是包括单元格内容的完整HTML树，表示总体的识别性能。

此外，Zhong [42] 将表格分为两类，即简单表格和复杂表格。前者是指没有单元格进行垂直或水平合并的表格，而后者则是指其他类型的表格。

4Proposal

该提案包括一个ResNet编码器和两个局部注意力Transformer解码器。这两个解码器分别推断表格结构和单元格内容。此外，一个额外的输出层估计单元格的边界框。

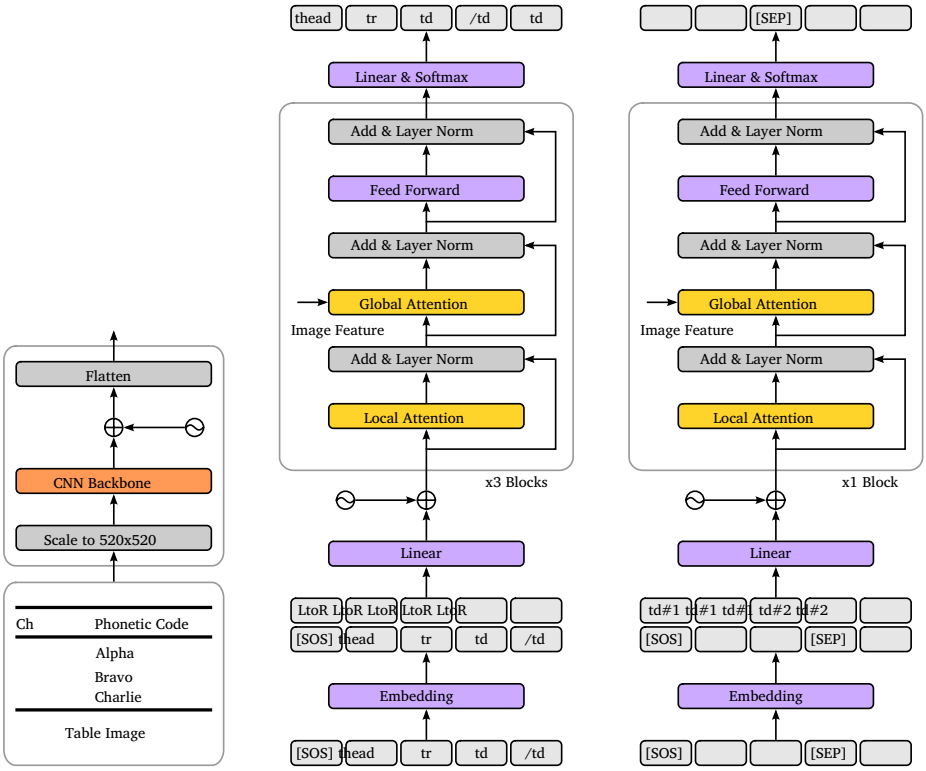
与之前的研究[18,19]相比，主要有两个不同之处：1) 引入了多细胞解码器，2) 在HTML解码器中引入了双向互学习机制。此外，还采用了2D位置编码。我们根据互学习、多任务学习和多细胞解码器的特点，将提出的方法命名为MuTabNet。图1展示了网络架构。

4.1Encoder

编码器由一个CNN主干网络和2D位置编码组成。CNN从520x520像素的图像中提取65x65像素的图像特征。对于CNN，我们采用了具有26个卷积层和三个GCA块的TableResNetExtra [36]。经过2D位置编码后，图像特征被展平为一维特征，具有512个通道，用于解码器中的交叉注意力。

4.2HTML解码器

HTML解码器由一个嵌入层、三个局部注意力块和两个输出层组成。每个注意力块通过块内的自注意力层接收表格结构序列。然后，注意力块通过交叉注意力层将图像特征融入表格结构序列中。



(a) CNN backbone. (b) HTML decoder. (c) Cell decoder.

图1：提出的网络架构。

并将序列通过一个前馈层输出。在模块内插入了多个跳跃连接和层归一化。最后一个注意力模块的输出通过两个输出层转换为HTML标记和单元格边界框。

在训练过程中，解码器从输入的HTML标记中预测左移或右移的HTML标记。移位方向由一个额外的独热向量指定。在推理过程中，解码器预测下一个标记，并迭代扩展输入序列以获得完整的HTML序列。

除了HTML标记外，解码器还接受一些特殊标记，即SOS、EOS和PAD。SOS是一个触发顺序推断的标记，并插入在标记序列的开头。EOS是一个停止推断的标记，并插入在标记序列的末尾。PAD在EOS之后插入，以均衡小批次中标记的长度。

继先前研究[18,19]之后，HTML序列被简单地标记化为HTML标签，除了表示单元格开始的标签。如果标签包含colspan，则将其标记化为‘’。

或rowspan属性。否则，该标签仅被标记化为‘’。需要注意的是，第5.1节中描述的FinTabNet [41]和PubTabNet [42]是公开可用的，并且已经应用了这种标记化方法。然后，我们将和紧随其后的标记合并为一个标记。

此外，我们为常见的序列模式分配了一些特殊标记。

如果数据集中的所有标题单元格都包含加粗文本，我们会在后处理过程中移除标签并进行补全。这些方法遵循了先前的研究**[18,19]**。

4.3单元解码器

细胞解码器由一个嵌入层、一个局部注意力块和一个输出层组成。根据先前的研究[18,19]，嵌入层逐个接受细胞字符。这是因为细胞内容通常由短句或未知词或数字组成，难以利用预训练的语言模型。

细胞解码器的基本结构与HTML解码器类似，但有以下不同之处。首先，在细胞内容之间插入一个特殊标记SEP，以触发移动到下一个细胞。其次，细胞解码器接受细胞内容及其对应的HTML特征的组合，这些特征是从HTML解码器的输出中提取的。这些改进使得该方案能够依次读取多个细胞的内容，同时参考之前细胞的信息。

之前的研究[18]利用局部注意力机制用于HTML解码器，而全局注意力机制用于单元格解码器。这是因为单元格解码器独立处理每个单元格，且单元格内容通常较短。另一方面，在我们提出的多单元格解码器中，单元格内容的序列往往较长。因此，我们采用了局部注意力机制。

4.4 双向互学习

我们提出了一种受深度互学习[38]启发的方法，用于训练HTML解码器。在此方法中，两个等效的解码器被联合训练，以从左到右（LtoR）或从右到左（RtoL）的方向预测表格结构。为了减少模型参数，我们通过结合一个额外的one-hot向量来确定方向，实现了在单一解码器中的互学习。

方向与嵌入的HTML标记。设 $\overrightarrow{x}$ 和 $\overleftarrow{x}$ 分别为从左到右和从右到左的序列，p(x)为真实概率，q(x)为预测概率。从左到右解码器的损失 $\overrightarrow{L}$ 由公式(10)定义。

$$\overrightarrow{\mathcal{L}} = -\frac{1}{N}\sum_{n=1}^{N} p(\overrightarrow{\boldsymbol{x}_n}) \log q(\overrightarrow{\boldsymbol{x}_n}) + \frac{1}{N}\sum_{n=1}^{N} q(\overleftarrow{\boldsymbol{x}_n}) \log \frac{q(\overleftarrow{\boldsymbol{x}_n})}{q(\overrightarrow{\boldsymbol{x}_n})}. \tag{10}$$

表1：表格图像数据集的统计数据。

Dataset	Training	Validation	Evaluation
FinTabNet	91,596	10,635	10,656
PubTabNet	500,777	9,115	9,064
PubTabNet250	114,111	2,161	

5Experiments

为了评估多单元解码器和双向互学习的效果，我们在两个公开的表格数据集上进行了实验。

5.1Datasets

我们使用了两个大型数据集，FinTabNet [41] 和 PubTabNet [42]。此外，我们还使用了名为 PubTabNet250 [18] 的子集进行消融研究。表1显示了这些数据集的统计数据。

FinTabNet 是一个大型表格图像数据集，包含 HTML 标签和单元格边界框，提取自 S&P 500 公司的年度报告。该数据集包含 112,000 个表格，并分为训练集、验证集和评估集。需要注意的是，原始的 FinTabNet 将验证集和评估集混淆了。根据先前研究 [21,41]，我们将包含 10,656 张图像的验证集视为评估集。

PubTabNet 是一个数据集，通过收集来自 PubMed 中央开放获取子集的科学文章构建而成，包含 568k 个表格及其对应的结构、单元格内容注释和单元格边界框。PubTabNet 提供了训练集和验证集，评估集则提供给 ICDAR 竞赛 [37]。我们根据第 3.6 节中的描述，将表格分类为简单表格和复杂表格。

PubTabNet250 Ly和Takasu[18]从PubTabNet中提取了包含250个或更多HTML标记的表格，并创建了一个名为PubTabNet250的子集。他们还引入了包含至少500、600和700个标记的表格子集。这些子集最初[18]用于展示局部注意力机制的有效性。我们也在第5.4节中利用这些子集进行了消融研究，大约将每个模型的训练时间从179小时减少到45小时。

表2：FinTabNet评估集上的表格识别结果。

		TEDS (%)	
Model		Structure	Total
EDD	[42]	90.60	
GTE	[41]	87.14	
GTE (PT)	[41]	91.02	
TableFormer	[21]	96.80	
VAST	[10]	98.63	98.21
Ly et al.	[20]	98.72	95.32
Ly and Takasu	[19]	98.79	
Ly and Takasu	[18]	98.85	95.74
MuTabNet		98.87	97.69

5.2Implementation

提出的模型在PyTorch中使用mmcv [22]、mmdet [23]和mmocr [24]框架实现，并在总共四块NVIDIA V100 GPU上以批量大小8进行训练。我们使用了Ranger [33]优化器。学习率在前25个周期初始化为0.001，接下来的三个周期和最后两个周期分别降至0.0001和0.00001。

每个表格图像都被归一化并缩减为520x520像素，必要时在边缘填充零。细胞边界框被归一化，使其最小值为0，最大值为1。

HTML 标记和单元格内容被转换为 512 维的嵌入表示。HTML 和单元格解码器中的四个注意力块具有相同的 8 头、512 通道架构，本地注意力的滑动窗口大小默认设置为 300，这与之前的工作 [18] 一致。表格结构序列和单元格内容序列的最大长度分别设置为 800 和 8000，包括特殊标记。我们采用贪婪搜索进行序列预测。

为了确保与之前的研究进行公平比较，我们没有使用数据增强或集成学习技术。我们也没有利用早停法。

5.3 实验结果

我们比较了在FinTabNet和PubTabNet上训练的所提出模型的性能与现有模型所宣称性能的差异。

FinTabNet 我们使用 TEDS 指标评估了结构识别和整体识别的实验结果。表 2 比较了 TEDS 分数。

表3：PubTabNet验证集上的表格识别结果。

		TEDS (%)		
Model		Simple	Complex	Total
EDD	[42]	91.20	85.40	88.30
TabStruct-Net	[28]			90.10
TableFormer	[21]	95.40	90.10	93.60
SEM	[39]	94.80	92.50	93.70
LGPMA&OCR	[27]			94.60
VCGroup	[36]			96.26
VCGroup&ME	[36]			96.84
VAST	[10]			96.31
Ly et al.	[20]	97.89	95.02	96.48
Ly and Takasu	[19]	97.92	95.36	96.67
Ly and Takasu	[18]	98.07	95.42	96.77
MuTabNet		98.16	95.53	96.87

提案与先前模型之间的测试集。提案的表现优于先前模型，得分分别为98.87%和97.69%。使用4块GPU进行推理的时间为3.78小时。

该提案的总TEDS评分低于VAST的评分[10]，这可能是因为VAST利用了外部OCR进行单元格内容识别。相比之下，VAST的结构TEDS评分低于端到端方法的评分[20,19,18]，包括该提案。

PubTabNet 我们在验证集上使用TEDS指标评估了表格识别的实验结果。表3比较了该提案与之前方法的得分。该提案在简单表格、复杂表格和所有表格上的得分分别为98.16%、95.53%和96.87%，均优于所有之前的方法。使用4个GPU的推理时间为3.23小时。

我们还对我们的提案在评估集上进行了评估。表4将该提案的得分与ICDAR竞赛[37]的前10名解决方案进行了比较。在这两个集合上取得的高分表明了该提案具有较高的泛化性能。推理时间为3.13小时。

该提案的得分高于VAST[10]的得分。PubTabNet包含大量训练数据，所提出的模型似乎在单元内容识别任务上得到了良好的训练。

需要注意的是，VCGroup&ME [36] 利用了单元格内容中文字行的额外边界框标注以及三种模型的集成学习。所提出的模型在所有非端到端模型中表现最佳。

表4：PubTabNet评估集上的表格识别结果。

		TEDS (%)		
Model		Simple	Complex	Total
LTIAYN	[37]	97.18	92.40	94.84
anyone	[37]	96.95	93.43	95.23
PaodingAI	[37]	97.35	93.79	95.61
TAL	[37]	97.30	93.93	95.65
DBJ	[37]	97.39	93.87	95.66
YG	[37]	97.38	94.79	96.11
XM	[39]	97.60	94.89	96.27
VCGroup	[36]	97.90	94.68	96.32
Davar-Lab-OCR	[37]	97.88	94.78	96.36
Ly et al.	[20]	97.51	94.37	95.97
Ly and Takasu	[19]	97.60	94.68	96.17
Ly and Takasu	[18]	97.77	94.58	96.21
MuTabNet		98.01	94.98	96.53

尽管我们的模型没有使用这些技术，但使用了额外的注释和集成学习。

5.4 消融研究

我们进行了额外的消融实验，使用PubTabNet250数据集进行训练，并使用PubTabNet子集进行评估。

多细胞解码器与互学习的有效性

我们评估了所提出方法的有效性，即多细胞（MC）解码器和双向互学习（BML）。我们在训练集上训练了两个模型，并计算了如表5所示的验证分数。我们选择了先前的实验结果[18]作为基线，使用了完全相同的模型架构和数据集，除了MC和BML。表中的LA指的是局部注意力。

由于之前的研究[18]主要关注长表格的性能，我们也计算了包含至少500、600和700个结构标记的表格的TEDS分数。MC解码器在所有表格长度上都优于基线，而BML进一步提升了表格识别性能。

BML在结构TEDS评分中的效果尚不明确，但在总TEDS评分中效果显著。BML可能仍提升了隐式结构识别的表现，并对细胞内容识别产生了影响，这需要精确的内容定位。

表5：采用所提方法的表格识别结果。

Methods			TEDS (%)							
			Structure				Total			
LA	MC	BML	250+	500+	600+	700+	250+	500+	600+	700+
							93.86	91.16	90.63	88.65
✓							94.28	92.99	91.29	89.61
✓	✓		96.60	96.71	96.75	96.67	95.02	94.59	93.73	93.14
✓	✓	✓	97.02	96.70	96.35	96.65	95.81	95.11	94.05	94.02

表6：不同窗口尺寸的表格识别结果。

	TEDS (%)								
	Structure				Total				
Size	250+	500+	600+	700+	0+	250+	500+	600+	700+
100	96.96	96.69	96.98	96.60	75.91	95.70	95.19	94.99	94.35
200	96.79	96.53	96.30	95.83	75.79	95.46	94.66	93.80	92.69
300	97.02	96.70	96.35	96.65	83.15	95.81	95.11	94.05	94.02
400	96.83	96.85	96.48	96.51	82.58	95.40	95.08	94.00	93.50
500	96.97	96.74	97.03	96.54	81.14	95.51	94.46	93.88	92.65

Ly和Takasu [18] 报告称，HTML解码器的窗口大小为300时效果最佳，而单元解码器则利用了全局注意力。在本研究中，我们确定了MC解码器的最佳窗口大小。表6显示了验证集的TEDS分数随着窗口大小从100到500的变化情况，而HTML解码器的窗口大小固定为300。

通常，窗口大小为300时得分最高，但在包含超过500个token的表格中，窗口大小为100时得分最高。含有许多单元格的表格往往每个单元格的字符较少，因此较短的窗口可能已足够。

需要注意的是，我们使用了PubTabNet250数据集进行训练，

对于结构化标记较少的表格，其性能低于表3中的得分。

我们从泛化性能的角度出发，选择了300的窗口大小作为整个PubTabNet数据集中包含较少标记的表格的最佳值。

6Conclusion

我们基于Transformer改进了一个端到端的表格识别模型，以实现与使用外部OCR的最先进模型相媲美的性能。

系统。提出的模型由一个ResNet编码器和两个解码器组成，分别用于结构识别和单元格内容识别。在第一个解码器推断出结构标记后，第二个解码器读取每个单元格中的文本。

我们提出了一种用于单元格内容识别的多单元格解码器，以利用来自相邻单元格的有用信息。此外，我们还提出了双向互学习方法，以迫使模型同时关注前后的单元格。使用两个公开数据集的实验结果证明了所提出方法的有效性。

在未来的工作中，我们将进一步考虑包含识别表格意义任务的多任务模型，这使得能够深入理解包括表格内容在内的印刷文档，并为大语言模型（LLMs）和问答系统提供高质量的科学知识。

参考文献

1. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: 长文档Transformer (2020)

2. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: GCNet: 非局部网络与挤压激励网络及超越。在：IEEE/CVF国际计算机视觉研讨会（ICCVW）。第1971–1980页 (2019)

3. Deng, Y., Rosenberg, D., Mann, G.: 端到端神经科学表格识别的挑战。在：国际文档分析与识别会议（ICDAR）。第894–901页 (2019)

4. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: 卷积序列到序列学习。在：国际机器学习会议（ICML）。第1243–1252页 (2017)

5. He, K., Gkioxari, G., Doll´ar, P., Girshick, R.: Mask R-CNN。在：IEEE国际计算机视觉会议（ICCV）。第2980–2988页 (2017)

6. He, K., Zhang, X., Ren, S., Sun, J.: 用于图像识别的深度残差学习。在：IEEE计算机视觉与模式识别会议（CVPR）。第770–778页 (2016)

7. Hinton, G., Vinyals, O., Dean, J.: 神经网络中的知识蒸馏 (2015)

8. Hochreiter, S., Schmidhuber, J.: 长短期记忆。神经计算 9, 1735–80 (1997)

9. Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: 密集连接的卷积网络。在：IEEE计算机视觉与模式识别会议（CVPR）。第2261–2269页 (2017)

10. Huang, Y., Lu, N., Chen, D., Li, Y., Xie, Z., Zhu, S., Gao, L., Peng, W.: 通过视觉对齐序列坐标建模改进表格结构识别。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第11134–11143页 (2023)

11. Itonori, K.: 基于文本块排列和表格线位置的表格结构识别。在：国际文档分析与识别会议（ICDAR）。第765–768页 (1993)

12. Kayal, P., Anand, M., Desai, H., Singh, M.: ICDAR 2021科学表格图像识别到LaTeX竞赛。在：国际文档分析与识别会议（ICDAR）。第754–766页 (2021)

13. Kieninger, T.G.: 基于鲁棒块分割的表结构识别。

在：Photonics West ’98。第3305卷，第22–32页（1998年）

14. Kullback, S., Leibler, R.A.: 关于信息和充分性。数学统计年鉴22(1)（1951年）

15. Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: TableBank: 基于图像的表格检测和识别基准。在：语言资源和评估会议（LREC）。第1918–1925页（2020年）

16. Long, J., Shelhamer, E., Darrell, T.: 用于语义分割的全卷积网络。在：IEEE计算机视觉与模式识别会议（CVPR）。第3431–3440页（2015年）

17. Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., Bai, X.: MASTER: 用于场景文本识别的多方面非局部网络。模式识别117，107980（2021年）

18. Ly, N.T., Takasu, A.: 一种基于端到端局部注意力的表格识别模型。在：国际文档分析与识别会议（ICDAR）。第20–36页（2023年）

19. Ly, N.T., Takasu, A.: 一种基于图像的表格识别的端到端多任务学习模型。在：国际计算机视觉、成像与计算机图形理论与应用联合会议（VISIGRAPP）。第626–634页（2023年）

20. Ly, N.T., Takasu, A., Nguyen, P., Takeda, H.: 使用弱监督方法重新思考基于图像的表格识别。在：国际模式识别应用与方法会议（ICPRAM）。第872–880页（2023年）

21. Nassar, A., Livathinos, N., Lysak, M., Staar, P.: TableFormer: 使用变换器的表格结构理解。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）。第4604–4613页（2022年）

22. OpenMMLab: MMCV, https://github.com/open-mmlab/mmcv

23. OpenMMLab: MMDetection, https://github.com/open-mmlab/mmdetection

24. OpenMMLab: MMOCR, https://github.com/open-mmlab/mmocr

25. Peng, A., Lee, S., Wang, X., Balasubramaniyan, R., Chau, D.H.: 高性能变换器用于表格结构识别需要早期卷积。在：表格表示学习研讨会（2023年）

26. Prasad, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.: CascadeTabNet: 一种用于图像文档表格检测和结构识别的端到端方法。在：IEEE/CVF计算机视觉与模式识别会议研讨会（CVPRW）。第2439–2447页（2020年）

27. Qiao, L., Li, Z., Cheng, Z., Zhang, P., Pu, S., Niu, Y., Ren, W., Tan, W., Wu, F.: LGPMA: 使用局部和全局金字塔掩码对齐的复杂表格结构识别。在：国际文档分析与识别会议（ICDAR）。第99–114页（2021年）

28. Raja, S., Mondal, A., Jawahar, C.V.: 使用自上而下和自下而上的线索进行表格结构识别。在：计算机视觉 – ECCV。第70–86页（2020年）

29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: 使用区域提议网络实现实时目标检测。在：神经信息处理系统进展。第28卷（2015年）

30. Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: DeepDeSRT: 用于文档图像中表格检测和结构识别的深度学习。在：国际文档分析与识别会议（ICDAR）。第1162–1167页（2017年）

31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: 注意力就是你所需要的。在：神经信息处理系统进展。第30卷，第6000-6010页 (2017)

32. Wang, Y., Phillips, I.T., Haralick, R.M.: 表格结构理解及其性能评估。模式识别 37(7), 1479-1497 (2004)

33. Wright, L.: Ranger - 一个协同优化器 (2019), https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer

34. Xie, S., Girshick, R.B., Doll´ar, P., Tu, Z., He, K.: 深度神经网络的聚合残差变换。在：IEEE计算机视觉与模式识别会议 (CVPR)。第5987-5995页 (2017)

35. Yang, F., Hu, L., Liu, X., Huang, S., Gu, Z.: 一个用于端到端表格识别的大规模数据集。科学数据 10(1), 110 (2023)

36. Ye, J., Qi, X., He, Y., Chen, Y., Gu, D., Gao, P., Xiao, R.: PingAn-VCGroup在ICDAR 2021科学文献解析任务B中的解决方案：表格识别为HTML (2021)

37. Yepes, A.J., Zhong, P., Burdick, D.: ICDAR 2021科学文献解析竞赛。在：国际文档分析与识别会议 (ICDAR)。第605-617页 (2021)

38. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: 深度互学习。在：IEEE/CVF计算机视觉与模式识别会议 (CVPR)。第4320-4328页 (2018)

39. Zhang, Z., Zhang, J., Du, J., Wang, F.: 分割、嵌入与合并：一个准确的表格结构识别器。模式识别 126, 108565 (2022)

40. Zhao, W., Gao, L., Yan, Z., Peng, S., Du, L., Zhang, Z.: 基于双向训练变换器的手写数学表达式识别。在：国际文档分析与识别会议 (ICDAR)。第570-584页 (2021)

41. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: 全局表格提取器 (GTE)：一个联合表格识别与单元格结构识别的框架，利用视觉上下文。在：IEEE冬季应用计算机视觉会议 (WACV)。第697-706页 (2021)

42. Zhong, X., ShafieiBavani, E., Yepes, A.J.: 基于图像的表格识别：数据、模型与评估。在：计算机视觉 – ECCV。第564-580页 (2020)