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:TP183　　**文献标识码**:A　　**文章编号**:1006-8961(2025)11-3506-18

论文引用格式:Zhu J H, Gao L C, Zhao W Q, Peng S, Hu P F and Du J. 2025. Offline recognition of handwritten mathematical expressions: a survey. Journal of Image and Graphics, 30(11):3506-3523(朱建华, 高良才, 赵文祺, 彭帅, 胡鹏飞, 杜俊. 2025. 离线手写数学公式识别综述. 中国图象图形学报, 30(11):3506-3523)[DOI:10.11834/jig.240425]

离线手写数学公式识别综述

朱建华[1],高良才[1*],赵文祺[1],彭帅[1],胡鹏飞[2],杜俊[2]

1. 北京大学王选计算机研究所,北京　100871; 2. 中国科学技术大学语音及语言信息处理国家工程研究中心,合肥　230026

摘　要:手写数学公式在教育和科技等领域具有广泛的应用,如何将其准确识别并转换成MathML或LaTeX等格式的结构化表达式即手写公式识别,成为文字识别领域一个备受关注的研究问题。由于手写公式具有嵌套层次结构、书写风格多样等特点,这个研究问题仍极具挑战性。目前,手写公式识别的研究工作主要分为基于语法规则的传统方法和基于深度学习的方法。本文在系统回顾传统公式方法的识别流程与问题分析之后,重点梳理总结了基于深度学习的手写公式识别方法,围绕视觉特征提取、视觉与文本特征对齐和文本输出回归3个公式识别子任务,针对语义不变的视觉特征学习、"缺乏覆盖"、输出不平衡和建模公式二维结构4个问题,综述了过往工作进行的相关改进与优化。另外,针对当下热门的多模态大模型,在手写数学公式识别数据集上也对其进行了测试,并补充了其在印刷体公式识别中的表现。最后,结合手写公式识别目前面临的挑战和困难,对未来的发展方向和研究趋势进行了展望。

关键词:手写数学公式识别(HMER);密集卷积网络;注意力机制;双向训练;树结构

Offline recognition of handwritten mathematical expressions: a survey

Zhu Jianhua[1], Gao Liangcai[1*], Zhao Wenqi[1], Peng Shuai[1], Hu Pengfei[2], Du Jun[2]

1. Wangxuan Institute of Computer Technology, Peking University, Beijing　100871, China; 2. National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei　230026, China

Abstract: The rapid development of online learning and communication has gradually increased the application of handwritten mathematical expressions in fields such as education and technology. The capability to accurately recognize handwritten mathematical expressions and convert them into structured formats such as MathML or LaTeX has become a critical issue. However, due to the inherent challenges introduced by the diverse handwriting styles and the complex two-dimensional structure of mathematical expressions, this problem remains highly challenging. Current research on handwritten mathematical expression recognition (HMER) is primarily divided into two categories: traditional and deep learning-based methods. Among these approaches, deep learning-based methods have attracted considerable attention due to their capability to model the problem end-to-end using encoder-decoder architectures. These methods typically involve three key steps: feature extraction, alignment, and regression, each of which presents its own set of difficulties. Feature extraction refers to the process by which the encoder extracts high-level semantic information from the input image. One of the primary challenges in this step is learning semantic invariant features. Owing to differences in writing habits, the same symbol can be written in vastly different styles by different individuals. This stylistic variability can lead to significant challenges in

收稿日期:2024-07-29;**修回日期**:2025-02-18;**预印本日期**:2025-02-25

*通信作者:高良才　glc@pku.edu.cn

基金项目:国家自然科学基金项目(62376012)

Supported by:National Natural Science Foundation of China(62376012)

recognizing symbols accurately. For instance, symbols such as "x" and "X" or "C" and "c" can appear very similar in handwritten form, which causes difficulty for the model to distinguish between them. In addition, the same symbol may vary in size depending on its position within the expression. For example, a comma or a period in a superscript or subscript position may be much smaller than the same symbol in the main body of the expression. Therefore, the vision encoder must learn to extract features that are invariant to these stylistic, size, and shape differences. This capability is crucial for ensuring that the model can accurately recognize symbols regardless of how they are written or where they appear in the expression. Alignment refers to the process by which the decoder gradually aligns the extracted visual features with the corresponding text features using attention mechanisms. A key challenge in this step is the "lack of coverage" problem. Ideally, the model should convert each text structure on the image into an appropriate text representation without repeating or omitting any parts. However, in practice, models often suffer from over-parsing (repeating parts of the expression) or under-parsing (omitting parts of the expression). This issue arises because the model lacks global alignment information, which leads to difficulty in determining which parts of the image have already been processed. To address this concern, some models introduce coverage vectors. These vectors keep track of the attention distribution over previous time steps, which ensure that the model does not repeatedly attend to the same region of the image. This feature helps mitigate over-parsing and under-parsing issues, which improves the overall accuracy of the model. Regression refers to the process by which the decoder generates the output sequence in a self-regressive manner. This step involves two main challenges: the output imbalance problem and the modeling of 2D structure relationships. The output imbalance problem arises because models typically predict the output sequence from left to right (L2R), which can lead to more accurate predictions for the prefix of the sequence compared with the suffix. The reason is that the capability of the model to capture dependencies between symbols diminishes as the distance between them increases. To address this problem, some models employ bidirectional training strategies, where the model is trained to predict the sequence in L2R and right-to-left directions. This approach helps balance the accuracy of the prefix and suffix predictions. Moreover, mathematical expressions often have a nested——2D structure, which can be challenging to model using traditional sequence-based approaches. Some models attempt to address this issue by incorporating tree-structured decoders, which explicitly model the hierarchical relationships between symbols. However, these methods often require highly complex training procedures and may not always outperform sequence-based approaches. This study builds on the introduction of the abovementioned "feature extraction", "alignment", and "regression" steps and the existing challenges and difficulties. It then reviews relevant improvements made in previous works on these challenges. For example, in the feature extraction step, models such as DenseWAP and ABM have introduced multi-scale feature extraction techniques to better capture fine-grained details in the input image. In the alignment step, models including WAP and CoMER have introduced coverage vectors and attention refinement modules to address the lack of coverage problem. In the regression step, models such as BTTR and ABM have employed bidirectional training strategies to improve the modeling of long-range dependencies. After discussing these improvements, this study supplements the experimental performance of some models on printed datasets. These experiments show that the main bottleneck of existing models is the diversity of writing strokes in handwritten expressions. While models perform well on printed datasets, where the style and size of symbols are consistent, their performance drops significantly on handwritten datasets due to the variability in writing styles. This drop in performance highlights the need for further research into improving the robustness of models to different writing styles. In addition to discussing traditional and deep learning-based methods, this study also conducts tests on HMER datasets using large multimodal models (LMMs) such as GPT-4V and Qwen-VL-Max. These models, which are trained on vast amounts of multimodal data, have shown impressive performance in various vision and language tasks. However, their performance on HMER tasks is still suboptimal compared with those of specialized models. This performance is likely due to the lack of handwritten mathematical expression data in their training sets and the difficulty of capturing the complex 2D structure of mathematical expressions. The results of these experiments suggest that, while LMMs have potential in HMER, their performance can still be improved. Finally, this study points out future directions for development in the field of HMER. One promising direction is the integration of tree-structured prediction with sequence-based approaches. While tree-structured decoders offer a more interpretable way to model the hierarchical relationships in mathematical expressions, they often require highly complex training procedures

and may not always outperform sequence-based methods. Future research could explore ways to combine the strengths of both approaches through pre-training the decoder on large-scale LaTeX datasets and fine-tuning it on handwritten expression data. Another direction is the development of more robust data augmentation techniques to improve the capability of the model to handle diverse writing styles. Furthermore, the creation of larger and more diverse datasets, including multi-line expressions and mixed handwritten and printed text, could help advance the field. While significant progress has been made in the field of HMER, many challenges still need to be addressed. By overcoming these challenges and exploring new directions, researchers can continue to improve the accuracy and robustness of HMER models. This enhancement increases their implementation in real-world applications such as education, document digitization, and scientific research.

Key words: handwritten mathematical expression recognition (HMER); dense convolutional network; attention mechanism; bidirectional training; tree structure

0 引 言

光学字符识别(optical character recognition, OCR)是模式识别、计算机视觉等领域的热门话题。由于手写公式具有嵌套层次结构、书写风格多样等特点手写数学公式识别(handwritten mathematical expression recognition, HMER)一直是OCR领域的一个研究难点。相比需要手工设计特征的传统方法,深度学习(deep learning, DL)(LeCun等, 2015; Liu等, 2023)能够受益于大规模数据,自动学习数据中的结构和语义特征信息,在计算机视觉、自然语言处理等多个领域取得了突破性进展,也在手写数学公式识别任务中大放异彩。

在深度学习的语境下,离线手写数学公式识别(offline handwritten mathematical expression recognition, Offline HMER)任务的简单定义是给定手写数学公式图像作为输入,模型需要预测输出公式对应的结构化标记语言(如LaTeX、MathML等)序列。为了之后叙述方便,本文首先介绍数学公式中常见的4种数学符号类型(Chan和Yeung, 2000b)。

1)基础符号(basic symbols)。例如1、2、3等数字符号,a、b、c等字母符号,这些符号是数学公式的基础组成部分。

2)绑定符号(binding symbols)。例如分数线"—"、求和符号∑等,这些符号将周围的子公式绑定,形成一个新的公式。

3)栅栏符号(fence symbols)。例如"("和")"、"{"和"}"等,这些符号可以将数学公式拆分组合成多个子公式。

4)运算符号(operator symbols)。例如"+"、"−"等,这些符号代表它们相邻符号之间的操作。

随着深度学习技术的发展,离线手写文字识别,例如单个字词识别或文本行识别,已经取得了长足进步。但是离线手写数学公式识别由于手写公式本身的特点或深度学习模型的局限,依然存在诸多研究挑战。总体而言,主要存在以下4类研究难点:

1)数学公式具有嵌套层次结构。不同于一般的文本数据,数学公式不仅具有水平方向的语义结构,而且在竖直方向上存在语义结构。例如,在手写或印刷公式图像中,分式中的分子和分母具有空间位置的上下关系。同时,为了建模公式的嵌套层次结构,LaTeX等标记语言存在着诸多语法规范。如标识幂指数位置的^的后面通常是由栅栏符号"{"和"}"控制的子表达式,栅栏符号"{"和"}"总是成对出现。这种语法规范往往只存在标记语言文本中,而在输入的图像中却没有这些语法规范的显式信息的提示,因此模型在将图像上的文本结构转换为对应的LaTeX文本时,会预测出不符合语法规范的LaTeX文本,如"{"并没有对应的"}"与之匹配。

2)书写风格多样。由于书写习惯不同,不同人在书写公式符号时会存在显著的风格差异。这种风格差异可能导致相同字符的书写结果差异很大,同时不同符号在书写时又极其相似,如X和x、C和c、S和∫等符号。另外,手写的公式符号由于其在公式所处的位置不同会出现明显的大小差异,对于一些较小的公式符号例如",",或处于幂指数位置的细粒度数学符号,现有的模型很难识别准确。另外,目前手写数学公式的数据集较小,如常用的CROHME的训练集大小只有8 836幅,严重限制了现有深度学习方法的性能。

3)缺乏覆盖(lack of coverage)问题。在公式识别任务中,希望深度学习模型能够不重复且不遗漏地将图像上的每个文本和结构转换为合适的公式描

述(如LaTeX或MathML)。然而现有的深度学习方法依然在不同程度上受到缺乏覆盖问题的影响。这个问题有两种表现形式:过转换(over-parsing)和欠转换(under-parsing)。过转换意味着手写公式图像的某些部分被重复地转换多次,而欠转换则表示某些部分没有被转换到。造成这个问题的主要原因是模型缺乏全局的对齐信息来判断某个区域是否被转换过。

4)长序列建模时存在输出不平衡问题。基于深度学习的方法在推理阶段采用自回归模型从左向右(left to right ,L2R)逐个预测符号。随着距离的增加,这种方法捕捉到的当前符号与之前符号的依赖关系减弱,因此导致输出不平衡的问题,即预测序列的前缀序列比后缀更加准确。除了自回归预测所产生的问题,长序列建模还受到模型所采用解码器的限制,例如循环神经网络(recurrent neural network,RNN)模型自身存在的长期依赖问题。此外,缺乏覆盖问题也对其产生影响,当预测序列变长时,会出现字符漏识别或重复字符的情况,而这些错误的字符预测进一步致使后缀序列的预测准确率大幅降低。

根据实现方法的不同,本文将现有的离线手写公式识别方法主要分为基于语法规则的传统方法和基于深度学习的方法;而且根据识别任务和研究问题的不同,对深度学习方法进行了细分介绍。另外,补充了现有手写数学公式识别模型在印刷体公式识别中的表现,也在手写数学公式数据集上使用多模态大模型进行了测试。最后,结合手写公式识别目前面临的挑战和困难,对未来的发展方向和研究趋势进行了展望。

1 基于语法规则的传统方法

在传统方法中,手写数学公式识别通常包含符号识别(symbol recognition)和结构分析(structural analysis)两个步骤。首先,符号识别通常需要先将输入的公式图像进行分割,再识别成对应的数学符号;然后,结构分析将识别得到的符号排列,找到一个合适的数学表达式的结构,从而构造出符合语法规范的完整的数学公式。相较于基于深度学习的方法,传统方法往往需要人工定义规则,并非数据驱动,因此很难从大规模的标注数据中获益,也没有很好的通用性和泛化性。具体介绍如下。

1.1 符号识别

在数学公式的分割阶段,1991年,Okamoto(1991)提出一种基于像素点的横向和纵向裁剪算法。在此基础上,Ha等人(1995)进一步发展了这一技术,将输入从像素点转变为公式中字符元素的边界框进行分割,并提出一种递归式裁剪方法。这些分割方法均依赖于人工设定的阈值。

此外,在符号识别的过程中,过往工作采用隐马尔可夫模型(hidden Markov model,HMM)(Winkler,1996;Kosmala等,1999;Álvaro等,2014)、弹性匹配(elastic matching)(Chan和Yeung,1998;Vuong等,2010)或支持向量机(support vector machines,SVM)(Keshari和Watt,2007)等方法来解决符号的分类问题。使用隐马尔可夫模型的优点是在训练过程中不需要显式的符号分割信息,从而实现符号分割、符号识别和结构分析的联合优化,然而,这种方法的计算成本较高。

1.2 结构分析

传统方法中,一些研究依赖于二维的随机上下文无关语法(stochastic context-free grammar,SCFG)(Chou,1989)来解析数学公式的结构,并结合使用CYK(Cocke-Younger-Kasami)算法。然而,当CYK算法应用于二维语法结构时,其效率极低。相较之下,Fateman等人(1996)提出的自左向右的递归下降算法则显著加快了解析过程。Zanibbi等人(2002)基于二维语法进一步发展了树转换方法。此方法首先利用基准结构树描述输入符号的二维排列,通过将各种符号如数字、字母及绑定符号进行分组形成Lexed BST,然后将其转换成操作符树,以此表达输入表达式中的操作符顺序及其作用范围,从而实现对数学公式的高效解析。

此外,还有一些研究选择不使用二维语法结构对数学公式进行分析。例如,Chan和Yeung(2000a)采用定义条目语法(define clause grammar,DCG)递归解析一维数学表达式,从而洞察数学公式的结构。Toyota等人(2006)则是首先将数学结构表示为一维形式,随后通过公式描述语法进行解析。总的来说,设计一个能够完整表达数学公式符号多样性和结构关系的语法系统是具有挑战性的,语法的约束同样可能限制算法的可扩展性。

2 基于深度学习的方法

深度学习方法已经广泛应用于多种光学字符识别(OCR)任务,包括场景文本识别和手写文本识别,并且性能大大超过了传统方法。在手写数学公式识别领域,基于深度学习的方法也已成为主流,这些方法通常采用编码器—解码器(encoder-decoder)架构来构建模型,如图1所示。

在基于编码器—解码器架构的深度学习模型中,手写数学公式识别任务通常分为3个步骤进行:1)公式图像视觉特征提取:编码器处理输入图像,将其转换成表征其内在属性的潜在特征向量;2)对齐:解码器利用注意力机制,逐步将提取的视觉特征与文本特征进行对齐;3)回归:解码器依据对齐的信息生成预测序列。在采用自回归方法(auto regressive)训练期间,为了产生下一个预测词,模型会把之前单词的真实标签作为输入传递给解码器,并通过梯度下降法最小化交叉熵损失(cross entropy loss),以优化模型性能。

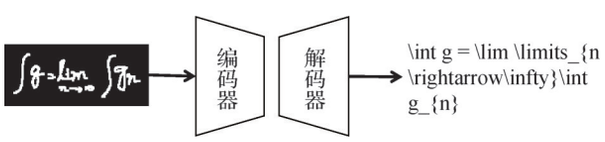


图1 基于深度学习的手写数学公式识别

Fig. 1 Handwritten mathematical expression recognition based on deep learning

在提取图像视觉特征的过程中,为了减少书写风格多样性带来的影响,编码器需要专注于学习图像中的语义不变信息;在对齐阶段,为防止图像某些区域的特征被重复处理或忽略,解码器需利用全局对齐信息确保特征图中的每一个文本结构都能被准确且完整地转换成相应的公式文本;在回归阶段,面对包含栅栏符号(如"{"和"}")的长数学公式时,模型必须建立这些符号间的依赖关系,以正确解析出数学公式的结构。这一系列步骤确保了公式识别的准确性和效率。

如表1所示,根据解码方式的不同,深度学习方法可分为基于序列解码的方法和基于树结构的方法。本文将通过这两种解码方式的不同,对之前的深度学习研究在视觉特征提取、对齐和回归3个步骤中解决的问题进行综述。

2.1 基于序列解码的方法

Zhang等人(2017)提出一个端到端的深度学习模型WAP(watch, attend and parse)解决手写数学公式识别问题,编码器采用了类似于VGGnet(Visual Geometry Group network)(Simonyan和Zisserman,2015)的全卷积神经网络结构,负责将输入的图像转换成潜在的特征向量。解码器则使用GRU(gated recurrent unit)模型,基于提取的视觉特征生成预测的LaTeX序列。WAP模型避免了传统方法中符号分割不准确的问题,并且免去了预定义公式语法的需要,因此成为后续研究中的一个重要基准模型。在此基础上,Zhang等人(2018)进一步提出DenseWAP模型,该模型在编码器中将VGGnet替换为DenseNet,如图2所示。DenseNet模型的特点是,在每个DenseNet块中,每一层的输入包括前面所有层的输出,这种结构促进了层间信息的流动。因此,DenseNet提高了特征提取的效能,广泛应用于后续研究中编码器的主干网络。

2.1.1 学习语义不变特征:捕捉多尺度信息

在数学公式的书写中,相同符号在公式不同位置会有明显的大小变化,同时不同符号之间也存在明显的大小差异,如较小的",",和较大的"∑"等符号之间的大小差异。此外,由于不同人的书写风格各异,同一符号在不同手稿中可能出现大小不一的情况。这些因素使得模型在处理时必须能够捕捉到多尺度的信息,以便学习到那些不受符号大小影响的语义不变特征。因此,开发能够适应符号尺度多样性的技术,是提高模型性能的一个关键方向。这种多尺度信息的捕捉能力不仅有助于更准确地解析手写公式,还能提高模型对于不同书写习惯的适应性。

Zhang等人(2018)提出的DenseWAP模型,缓解了池化操作带来的分辨率降低问题,增强了模型对于如逗号(",")和句点(".")这类小符号的捕获能力。DenseWAP模型在基于DenseNet的编码器中添加了一个能捕捉细粒度语义信息的多尺度分支模型。该多尺度分支在进行最终的平均池化层(average pooling)前延伸出来,保持特征图的原始分辨率,使得编码器能够提取到高分辨率的视觉特征图。这些高分辨率特征图与通过常规路径获得的低分辨率特征图一同被送入GRU解码器,以预测LaTeX序列。此种结合高低分辨率特征图的方法使得模型能够有效捕捉多尺度的视觉信息,从而减少了在识别

表 1　基于深度学习的手写数学公式识别方法

Table 1　HMER methods based on deep learning

解码方式	方法	解码器	改进方向	机制
序列解码	WAP	RNN	改进“缺乏覆盖”问题	1)DL 方法在 HMER 首次使用;2)在 RNN 解码器引入覆盖注意力
	DenseWAP	RNN	学习语义不变特征:捕捉多尺度信息	1)引入 DenseNet 编码器;2)多尺度分支捕捉细粒度特征
	WS WAP	RNN	学习语义不变特征:辅助任务	引入符号分类的辅助任务
	PALv2	RNN	学习语义不变特征:辅助任务	印刷体和手写体混合的对抗学习
	PrimCLR	RNN	学习语义不变特征:辅助任务	使用对比学习预训练编码器和 BiLSTM 上下文编码器
	CCLSL	RNN	学习语义不变特征:辅助任务	引入印刷体公式图像做编码器端的互学习
	SAM	RNN	学习语义不变特征:辅助任务	利用符号共现概率,区分相似公式符号
树解码	DenseWAP-TD	RNN	树解码	预测二维树结构而非序列解码
	TDv2	RNN	树解码	同一 LaTeX 有不同的树结构转换方式
	SAN	RNN	树解码	构造一系列语法规则重构树结构预测过程
序列解码	ABM	RNN	改进输出不平衡问题	双向解码的互学习
	CAN	RNN	学习语义不变特征:辅助任务	引入符号计数的辅助任务
	BTTR	Transformer	改进输出不平衡问题	1)引入 Transformer 解码器替代 RNN 解码器;2)双向训练、双向解码
	CoMER^{T}	Transformer	改进“缺乏覆盖”问题	引入覆盖注意力到 Transformer 解码器
	GCN	Transformer	弱监督/无监督学习	通用类别识别任务作为辅助任务

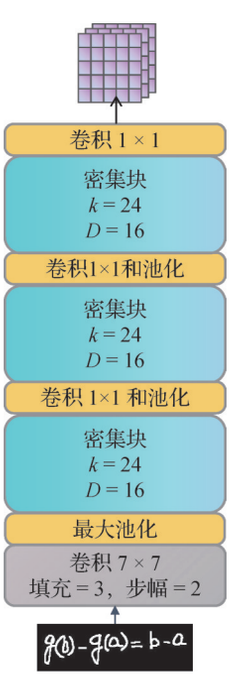


图 2　DenseNet 编码器

Fig. 2　Architecture of DenseNet encoder

细小符号时可能出现的语义信息损失。

后续的 ABM（attention aggregation based bi-directional mutual learning network）模型（Bian 等，2022），将多尺度信息捕捉的功能进一步迁移到解码器中。在构建覆盖向量（coverage vector）的过程中，ABM 模型利用不同大小的卷积核（如 5 × 5 和 11 × 11）处理过去累积的对齐信息，以实现对不同分辨率视觉特征的多尺度对齐。与 DenseWAP 模型相比，ABM 模型在减少运算量和细粒度语义信息损失的同时，还能够覆盖更广泛的感受野，关注更多的全局信息。这种使用不同大小卷积核捕捉多尺度信息的策略，在随后的 CAN（coumting-aware network）模型（Li 等，2022）中也得到进一步应用和发展。

2.1.2　学习语义不变特征：弱监督/无监督学习方法

如前所述，在基于编码器—解码器的模型中，公式识别任务包括图像视觉特征提取、对齐和回归 3 个主要环节。通常的训练方法是通过使用交叉熵损失函数（cross entropy loss）优化解码器输出的回归任务。然而，解码器产生的梯度在反向传播过程中可能不会精确地优化识别和对齐这两个任务，从而

可能未达到联合优化的最佳效果。

为了解决这一问题，部分研究引入了弱监督信息到编码器端，以帮助编码器更有效地提取高层视觉语义特征，从而优化识别任务。此外，这些弱监督学习过程中获得的全局信息也被送入到解码器，以进一步优化对齐任务。这种方法通过在整个模型架构中引入额外的监督信号，提高各任务的协同优化效果，从而提升整体模型的性能。

受到目标检测领域使用图像类别信息作为弱监督信号来训练检测器的启发，Truong等人(2020)提出WS WAP(weakly supervised WAP)模型。该模型利用公式图像中存在的符号集合作为弱监督信息，通过一个新增的符号分类器处理编码器提取的高维视觉特征，从而生成一个C维预测向量，其中，C代表公式中符号的种类数量。在WS WAP模型中，图像内所有出现的符号通过One-Hot编码形式作为真实标签，模型针对C维预测向量的每一个维度使用二分类的交叉熵损失函数计算损失，从而监督编码器判断数学公式中的符号是否出现在图像中。这种方法通过联合优化符号分类和回归任务显著提升了模型的性能。然而，WS WAP模型虽然利用了公式符号的存在信息，提高识别的准确性，但是只使用了公式符号是否存在这一信息，忽略了部分公式符号在图像中出现多次的情况。这一限制意味着模型在处理包含重复符号的公式图像时可能不够准确，未能完全捕捉到公式的全部结构信息。

CAN模型(Li等，2022)改进了这一问题。CAN模型引入一个多尺度计数模块(multi-scale counting module)，使用符号计数任务作为辅助任务与公式识别任务进行联合优化，如图3所示。这一多尺度计数模块由两个分支构成，每个分支使用不同大小的卷积核(3×3和5×5)捕捉编码器提取的高维视觉特征中的不同尺度信息。每个分支后续连接了通道注意力模块(channel attention)、1×1卷积层、ReLU(rectified linear unit)激活层和求和池化(sum pooling)层。这些组件的输出通过加和操作形成一个维度为C的计数向量(counting vector)，其中，C表示公式中符号的种类数量。在训练过程中，CAN模型使用每个符号在图像中实际出现的次数作为真实标签，这与生成的计数向量一起用来计算平滑的L1回归损失，从而优化模型对符号数量的预测能力。这种方法不仅解决了WS WAP模型中的问题，即未能准确处理符号重复出现的情况，而且通过将计数向量作为额外的全局信息输入解码器，进一步增强了对齐任务的性能。

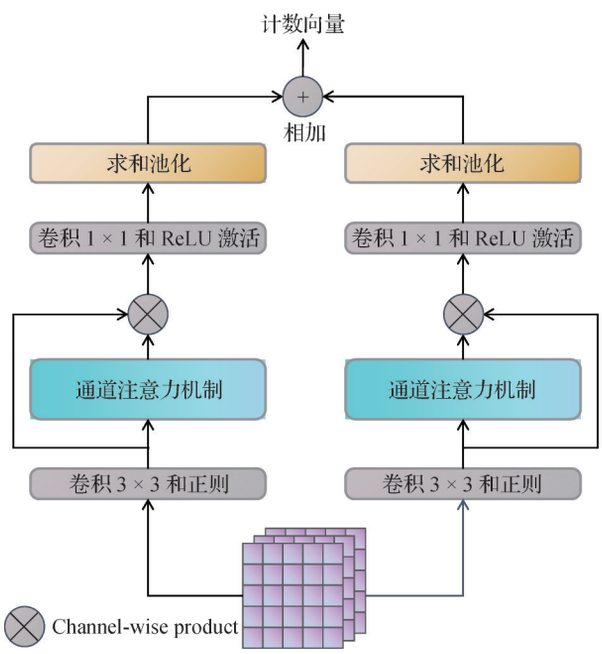


图3　CAN多尺度计数模块结构

Fig. 3　Architecture of CAN multi-scale counting module

GCN(general category network)模型(Zhang等，2023)则继续探索了辅助任务在手写数学公式识别的有效性。GCN提出通用类别识别任务(general category recognition task，GCRT)，根据数学符号用途，将公式符号分为23个不同的类别。GCN模型通过学习属于不同通用类别的符号之间的差异，可以更好地区分那些可能看起来相似的符号。GCN同时比较了符号计数任务等其他辅助任务对HMER任务的作用，验证了提出的通用类别识别任务的有效性。

由于书写风格的多样性与印刷体数学公式的标准形式之间存在明显的差异，最小化这两者在视觉特征分布上的差异是一个有效的策略，用以监督视觉编码器学习到更加通用的、语义不变的特征。Wu等人(2020)提出的PALv2(paired adversarial learning)模型则通过对抗学习的方法在解码器端间接实现了这一点，如图4所示。PALv2模型的训练过程中不仅处理手写公式图像，还同时处理与之内容相同的印刷公式图像。模型引入了一个判别器，该判别器的任务是判断解码器输出的字符是来自手写源还是印刷源。通过这种方式，模型被激励去学习那

些与书写或印刷方式无关的、核心的语义特征。然而，由于PALv2模型中的判别器是作用于解码器端，其产生的梯度需要通过解码器传回编码器，这可能导致梯度在传递过程中的效能降低，从而不足以直接强化编码器在捕捉语义不变特征上的能力。这意味着虽然模型能够在一定程度上缩小手写与印刷公式在视觉特征上的差异，但对编码器的直接影响可能较为有限。

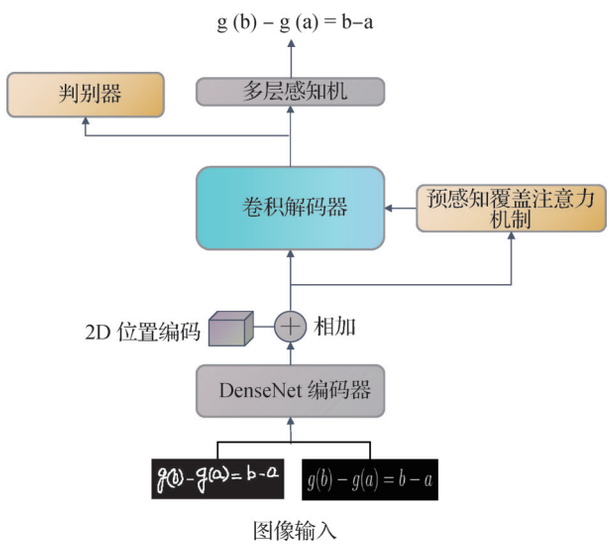


图4　PALv2模型架构

Fig. 4　Architecture of PALv2 model

CCLSL（combination of contrastive learning and supervised learning）（Lin 等，2022）模型通过在视觉编码器后加入了一个简单的投影层（projection head）来直接监督编码器学习语义不变特征。这个投影层是一个包含单个隐藏层的多层感知机（multi-layer perception，MLP），它的功能是将视觉编码器捕获的手写公式和印刷公式的视觉特征图投影到一个固定大小的低维向量空间中。CCLSL 模型使用带温度系数的交叉熵损失函数来最小化手写公式和印刷公式对应特征的分布的差异。这种方法尽管实现了对视觉编码器学习抽取语义不变特征的能力的直接监督，但是经过投影层后得到的低维向量损失了公式的结构布局信息。此外，CCLSL模型并未在解码器端引入任何损失函数来最小化手写与印刷公式符号概率分布的差异，这意味着解码器没有直接受益于印刷体数据训练带来的信息增益。

对比学习（contrastive learning）方法推动了自监督预训练模型在计算机视觉领域的广泛应用。自监督预训练方法可以通过学习大量未标记的数据提高泛化性能。该方法首先使用未标记的数据对视觉编码器进行预训练，然后通过有监督训练方法进行微调，从而迁移到下游任务中。这种预训练方法能够帮助视觉编码器学习到图像中的语义不变信息。

PrimCLR 模型（Guo 等，2022）通过对比学习方法预训练了一个基于 ResNet 的视觉特征编码器和基于 BiLSTM 的上下文编码器，并迁移到数学公式识别任务中来，如图 5 所示。该模型在预训练阶段，采用了局部语义不变的数据增强技术（local semantic-invariant data augmentation）如投影变换、高斯模糊等构造对比学习中的正例数据。这些方法能够在不破坏公式内在结构的情况下保留图像的局部语义信息。这与传统的数据增强方法如裁剪（cropping）和翻转（flipping）不同，后者可能会改变公式的原始结构。这些数据增强后的图像与原始图像同处一个批次（大小为 B），经过视觉特征编码器和上下文编码器后，投影到一个固定维度的特征图分布。同一个批次里的图像互为正例，而其他批次的图像为负例。PrimCLR 最小化正例的特征图分布之间的差异，增大与负例之间的差异来对视觉特征编码器和上下文编码器进行预训练。为了保留空间位置信息，投影后的特征图在每个通道上保持了二维结构。PrimCLR 进一步将这些二维特征图分割成多个块（patch），并在每一个块上应用带有温度系数的交叉熵损失函数（InfoNCE loss）计算损失。

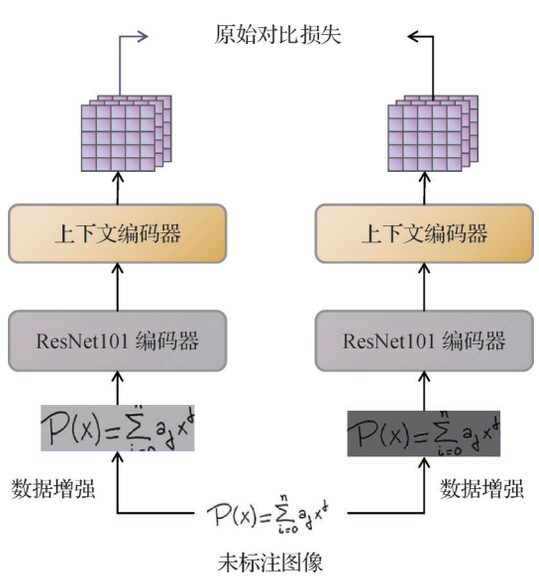


图5　PrimCLR 预训练模型架构

Fig. 5　Pretrain architecture of PrimCLR model

SAM模型(Liu等,2023)专门针对现有模型在识别相近数学符号时出现的混淆问题,引入了一种新的方法。该方法利用训练集中数学符号的共现概率作为弱监督信息,辅助模型更准确地学习和区分这些符号。为此,SAM模型设计了一个可插入式的语义感知模块(semantic aware module,SAM),该模块结构包含两个相同的分支。这两个分支各自处理不同的输入:视觉特征分支接收视觉编码器提取的视觉特征和解码器生成的注意力图;分类特征分支则处理解码器产生的分类向量。这种设计允许模块从视觉和分类信息的结合中学习如何区分外观相似的符号。在模型训练过程中,SAM模块通过计算投射后的视觉特征向量和分类向量之间的余弦相似度,并将此相似度与符号的共现概率进行对比,使用L2回归损失函数优化这一相似度计算,确保模型能有效地利用共现概率信息。这种方法不仅提高模型对相似符号的识别能力,还增强了模型在复杂数学公式解析中的整体性能,通过直观地利用数学符号间的语义关系减少识别错误。

2.1.3 改进输出不平衡问题

基于深度学习的单向自回归模型在推理阶段使用从左到右(L2R)的方向逐一预测符号,这样的方法可能会造成输出不平衡的问题:输出序列的前缀通常比后缀更加准确。表2是ABM模型(Bian等,2022)在进行单向训练时在CROHME 2014测试集上的开头2个或5个符号和结尾2个或5个符号的识别准确率。可以看到,L2R模型更擅长处理开头的符号,而R2L(right to left)模型则在结尾符号上取得更高的准确率。这个问题主要是由从左到右单向解码引起,其使用条件概率$P(\vec{y}_j \mid \vec{y}_{<j})$进行建模,使得下一个符号的输出概率只依赖于前面预测的结果,而没有利用到两个方向上的信息。

表2 ABM模型在CROHME 2014测试数据集上识别前2个或5个字符和最后2个或5个字符的准确率

Table 2 Recognition accuracy of prefixes-(2, 5) (the first two or five symbols) and suffixes-(2, 5) (the last two or five symbols) on CROHME 2014 test dataset with ABM model

/%

模型	开头2个字符	结尾2个字符	开头5个字符	结尾5个字符
单向训练—从左到右(L2R)	85.29	80.02	72.21	67.14
单向训练—从右到左(R2L)	81.22	84.47	67.01	71.17

为了缓解输出不平衡的问题,充分利用双向语言信息,BTTR模型(bidirectionally trained transformer)(Zhao等,2021)在基于Transformer的解码器上采用了双向训练策略,如图6所示。BTTR模型在使用Vanilla Transformer(Vaswani等,2017)作为解码器的情况下实现了双向语言建模。BTTR模型从目标LaTeX序列中生成两个方向的目标序列(L2R和R2L),然后将它们放在同一个批次(batch)中计算训练损失,并进行梯度反向传播。在推理阶段,利用BTTR模型具有语言双向建模能力的优势,使用联合估计搜索(approximate joint search)算法进一步提升性能。基本思想包含以下4个步骤:

1)对从左到右(L2R)和从右到左(R2L)两个方向进行束搜索,以此得到两个大小为k的候选集,以及其相应的分数$S(\boldsymbol{y})$。

2)将L2R候选序列翻转为R2L方向,将R2L序列翻转为L2R方向,记为$\boldsymbol{y} \overset{\text{reverse}}{\rightarrow} \boldsymbol{y}_{\text{rev}}$。

3)使用BTTR模型计算翻转后的序列$\boldsymbol{y}_{\text{rev}}$在模型中的损失$S(\boldsymbol{y}_{\text{rev}})$。

4)将相同序列的在两个方向上得到的分数相

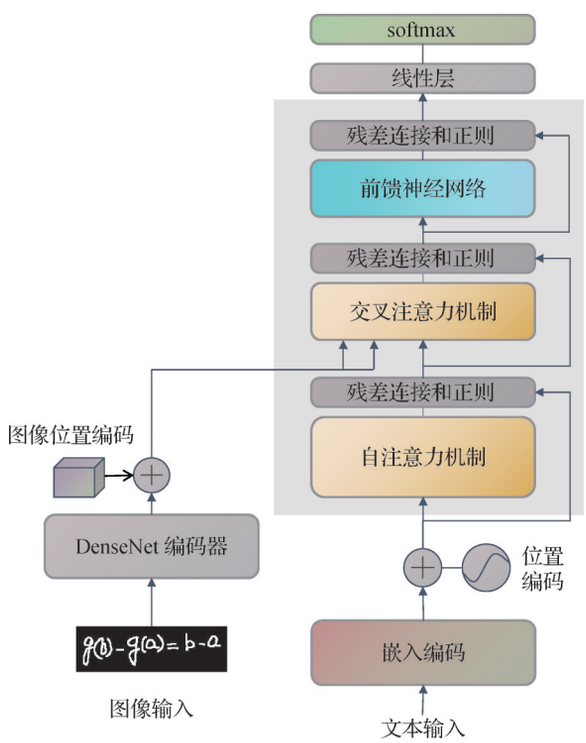


图6 BTTR模型架构

Fig. 6 Architecture of BTTR model

加，得到最终分数 $S(\boldsymbol{y}) + S(\boldsymbol{y}_{\text{rev}})$，并以此为依据，将最终分数最高的序列作为输出序列。

ABM 模型（Bian 等，2022）同样使用了双向训练的策略，如图 7 所示。该模型采用两个具有相同结构但解码方向相反的 GRU 解码器进行解码。在训练过程中，从目标 LaTeX 序列生成两个方向的目标序列（L2R 和 R2L），并使用这两个目标序列与对应解码器的输出计算损失以进行梯度回传。同时，ABM 模型引入 KL（Kullback-Leibler）散度损失减小两个解码器预测分布的差异，以使其能够互相学习解码信息。

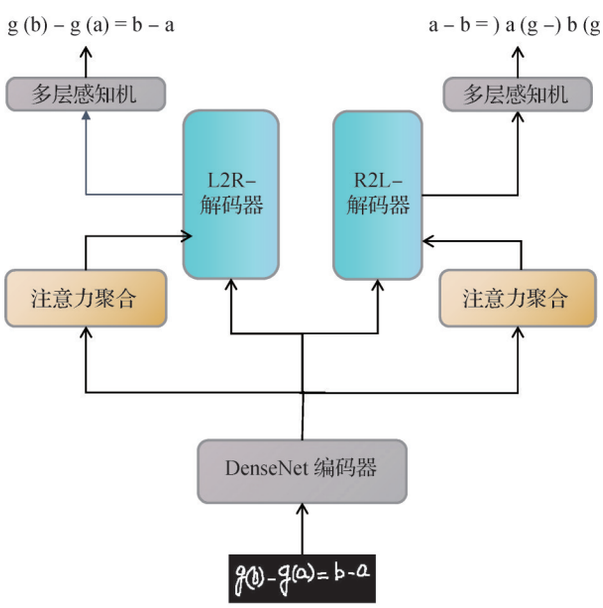


图 7　ABM 模型架构

Fig. 7　Architecture of ABM model

2.1.4　改进“缺乏覆盖”问题

在公式识别任务中，涉及编码器和解码器的联合优化，所以希望模型能够不重复且不遗漏地将图像上的每个文本结构转换为合适的文本。然而现有的深度学习方法都在不同程度上受到缺乏覆盖问题的影响。这个问题有两种表现形式，过转换和欠转换。过转换意味着手写公式图像的某些部分被重复地转换多次，而欠转换则表示某些部分没有被转换到。如图 8 左栏图像，模型重复解析了第 1 个字母“e”，而漏掉了要转换的“^”符号。造成缺乏覆盖问题的主要原因是模型缺乏全局的对齐信息来判断某个区域是否被转换过。

因此，为了缓解缺乏覆盖的问题，WAP 模型（Zhang 等，2017）使用覆盖向量引导解码器在图像尚未解码的区域分配更高的注意概率。这个覆盖向量能够记录之前所有时间步中的注意力分布信息，从而引导当前时间步向尚未被注意到的图像区域进行匹配，避免解码器在一个图像区域重复匹配多次，同时减少未匹配的图像区域。图 8 展示了 WAP 模型在使用覆盖向量前后模型预测的结果，可以看到，使用覆盖向量后过转换和欠转换的问题有所缓解。

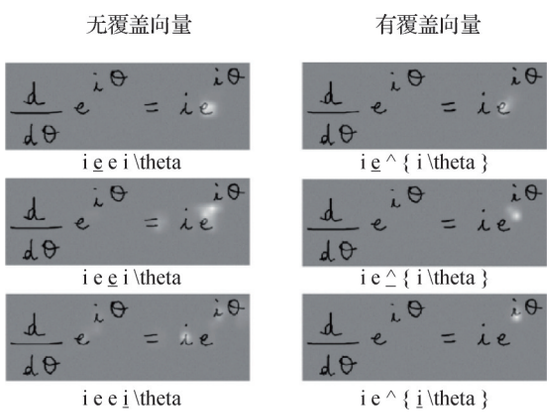


图 8　是否使用覆盖向量的注意力可视化

Fig. 8　Visualization of attention with and without the coverage vector

下面介绍 t 时刻覆盖向量的计算和注意力权重矩阵的更新算法。

设视觉编码器的输出视觉特征为 $\boldsymbol{X}_f \in \mathbf{R}^{L \times d_m}$，其中，$L = h_0 \times w_0$，$d_m$ 为可调整的超参数。t 时刻前的注意力权重 a_k 的加和记为覆盖向量 $\boldsymbol{c}_t$。对覆盖向量 $\boldsymbol{c}_t$ 进行卷积和线性变换后，得到覆盖矩阵 $\boldsymbol{F}_t$。详细的计算步骤为

$$\boldsymbol{c}_t = \sum_{k=1}^{t-1} a_k \in \mathbf{R}^L \tag{1}$$

$$\boldsymbol{F}_t = cov(\boldsymbol{c}_t) \in \mathbf{R}^{L \times d_a} \tag{2}$$

式中，$cov(\cdot)$ 表示经过卷积核大小为 11 的卷积操作和线性变换的复合，d_a 是可调整的超参数。

在得到覆盖矩阵 $\boldsymbol{F}_t$ 后，需要对 t 时刻的注意力权重进行更新。对于每一个位置在 $i \in [0, L)$ 的视觉特征，通过注意力机制计算相似度 $e_{t,i}$，进而得到整体的 e_t。e_t 在经过 softmax 层后，即可得到 t 时刻的注意力权重 a_t。具体计算为

$$e_t = \tanh(\boldsymbol{H}_t \boldsymbol{W}_h + \boldsymbol{X}_f \boldsymbol{W}_x + \boldsymbol{F}_t)\boldsymbol{v}_a \tag{3}$$

$$a_t = softmax(e_t) \in \mathbf{R}^L \tag{4}$$

式中，$\boldsymbol{W}_h$和$\boldsymbol{W}_x$都是可训练的参数矩阵，$\boldsymbol{v}_a$是可训练的参数向量，H_t为RNN产生的隐藏层状态。

在使用Transformer作为解码器主干网络的后续工作BTTR模型（Zhao等，2021）中，由于Transformer具有并行化特性，无法迭代地更新覆盖向量。因此BTTR模型使用了单词位置编码和二维归一化的图像位置编码，为Transformer解码器提供显式的位置信息，从而缓解缺乏覆盖问题。

CoMER模型（Zhao和Gao，2022）在BTTR模型的基础上和不损害Transformer并行解码的特性的同时，重新引入覆盖信息到基于Transformer的解码器中。其中，基本思路是，使用Transformer解码器中的多头注意力机制（multi-head attention）中的注意力权重计算覆盖向量。但Zhao和Gao（2022）指出，如果按照式（1）—式（4）的做法，将每一个时刻t和每个位置$i \in [0, L)$的注意力权重进行卷积操作，最终产生的覆盖矩阵$\boldsymbol{F} \in \mathbf{R}^{T \times L \times h \times d_a}$，空间复杂度太高。Zhao和Gao（2022）在意识到式（3）的tanh操作是整个过程的瓶颈后，巧妙地将相似度e_t的计算切分为两个部分。具体为

$$e'_t = \tanh(H_t\boldsymbol{W}_h + \boldsymbol{X}_f\boldsymbol{W}_x)\boldsymbol{v}_a + \boldsymbol{F}_t\boldsymbol{v}_a = \tanh(H_t\boldsymbol{W}_h + \boldsymbol{X}_f\boldsymbol{W}_x)\boldsymbol{v}_a + R \tag{5}$$

式中，等号右边第1项只与隐藏层状态H_t和视觉特征图$\boldsymbol{X}_f$有关，等号右边第2项则只与覆盖矩阵$\boldsymbol{F}_t$有关，用修正项R表示。这样，如果能设计一个新的简单模块拟合这个修正项R，就能将相似度的空间复杂度降低到O(TLh)。在具体实现中，CoMER模型引入了一个注意力修正模块（attention refinement module）实现修正项R的模拟计算，如图9所示。

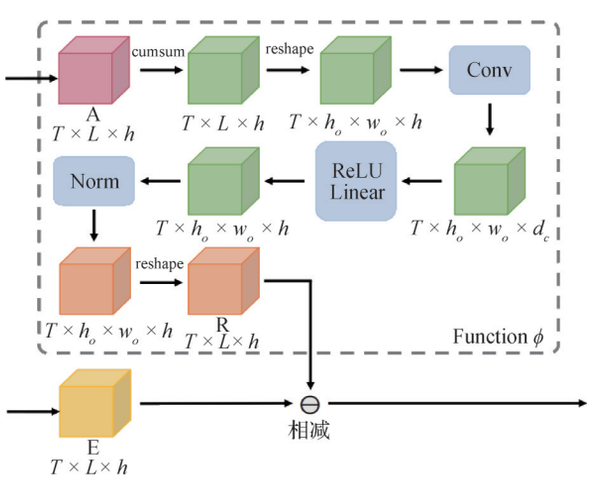


图9　CoMER中的注意力修正模块

Fig. 9　Attention refinement module in CoMER

LaTeX作为标记语言，在栅栏符号等的影响下，容易解析为一个树状表达式。因此利用数学公式本身存在的二维结构进行建模，为模型的预测过程提供可解释性，解决栅栏符号带来的依赖问题就成为一个重要的改进思路。

Zhang等人（2020）提出DenseWAP-TD模型，将直接回归LaTeX序列的GRU解码器替换为基于二维树状结构的解码器。如图10所示，在GRU解码器的每个时间步中，通过预测每个节点的字符类别、该节点的父节点以及与父节点之间的关系，构建数学公式的二维结构信息，从而避免了直接预测“{”和“}”等栅栏符号。但是TD模型在建模树结构时，没有考虑字符串到二维树状结构的转换方式有多种可能，而只限定了一种树状结构的解码过程，限制了模型的泛化能力。

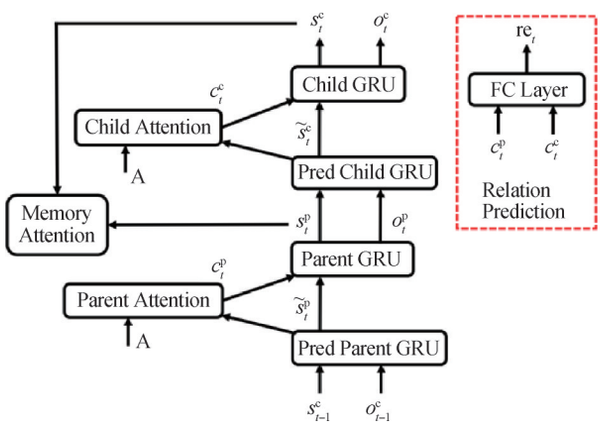


图10　TD模型树结构解码器（Zhang等，2020）

Fig. 10　Architecture of TD tree decoder (Zhang et al., 2020)

2.2　基于树解码的方法

TDv2模型（Wu等，2022）在训练过程中对同一LaTeX字符串使用不同的转换方式，削弱了上下文依赖关系，使得解码器具有更强的泛化性。同时，TDv2模型使用两个结构相同的节点分类模块（node classification module）和分支预测模块（branch prediction module）分别预测当前节点的符号类别和分支关系（如左方、上方、结尾等），简化了解码过程。

SAN（syntax-aware network）模型（Yuan等，2022）将LaTeX序列转为解析树，并设计了一系列的语法规则，将LaTeX序列的预测问题转换为树遍历的过程。具体而言，模型需要从开始符号（start symbol）产生3个过程，分别用$\sigma S, E, \varepsilon$表示，分别指下一步应当产生终结符（terminal symbol）——普通数学符

号的集合、关系符号(relation symbol)，如上、下、左、右等以及空字符的概率。其中，当$S \rightarrow E$的概率较高时，则使用关系预测分支输出每个关系符号的概率；当$S \rightarrow \sigma S$概率较高时，则输出对应的数学符号的概率。同时，为了更好地利用语法信息，SAN模型引入了一个新的语法感知注意力模块(syntax-aware attention module)，该模块并没有像过往工作使用先前所有注意力向量的和来计算覆盖注意力，而是只使用从解析树根节点到当前节点的路径中的注意力，从而减少无关噪声的影响。

2.3 多模态大模型

ChatGPT在海量数据上预训练Transformer，并通过人类反馈进行的强化学习(reinforcement learning from human feedback，RLHF)进一步优化和调整模型的表现，使其产生的回答更加符合人类的期待和偏好。后续的多模态大模型(large multimodal models)GPT-4V则将从语言模态拓展到视觉模态，并在多个视觉任务上取得了显著突破。使用大模型时，情景学习(in-context learning)方法能够显著提高其回答的质量和辅助模型实现多种任务。因此，可以通过设计特殊的提示(prompt)，并将其与图像一并输入给多模态大模型，使大模型完成OCR任务。

然而，GPT-4V在OCR任务上仍然面临多语言和复杂场景下性能受限、推理和更新的花销巨大等多个挑战(Shi等，2023)，在性能上无法与专用OCR模型相媲美。之后的开源工作则多针对复杂场景下输入图像的分辨率进行相关的改进工作。TextMonkey(Liu等，2024)使用Qwen-VL的LLM(large language model)模块作为解码器，使用ViT-BigG进行视觉编码，通过移位窗口注意力机制(shifted window attention)加强了视觉特征中跨窗口之间的关联，并扩张了输入图像的分辨率，从而能够更好地适用文档理解任务。在没有专用OCR模型或方法的辅助的情况下，TextMonkey在多个OCR任务上(如以场景文本为中心和以文档文中心)上的性能取得了显著提升，并在OCRbench上超越了过往的开源多模态大模型的表现。

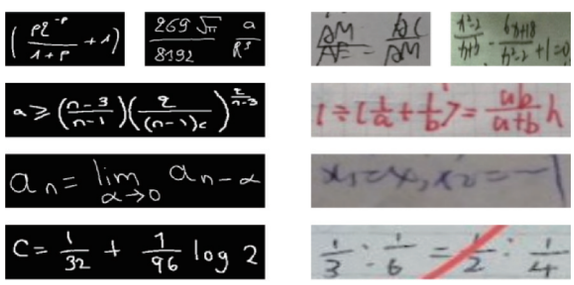
UReader(Ye等，2023)同样使用已预训练好的多模态大模型，通过充分利用各种视觉位置的语言理解数据集直接对多模态大模型进行指令微调。UReader提出一个形状自适应裁剪模块，将高分辨率图像切割成多个局部图像，从而可以利用多模态大模型的低分辨率编码器处理高分辨率图像，能够尽量避免由于缩放而导致的模糊和失真问题。

尽管多模态大模型本身具备处理手写数学公式识别任务的潜力，但由于部分多模态大模型的训练数据并未公开，而且，过往专门针对手写数学公式识别任务的研究也很少涉及多模态大模型在该领域的表现。因此，本文在4.2节中特别评测了若干可用的多模态大模型在手写数学公式识别任务上的性能，旨在为未来的研究提供参考。

3 数据集

3.1 数据集介绍

离线手写数学公式识别任务常用数据集包括CROHME和HME100K数据集，如图11所示。



(a) CROHME (b) HME100K

图11 CROHME和HME100K数据集的部分手写数学公式图像

Fig. 11 Partial handwritten mathematical expression images from CROHME and HME100K datasets ((a) CROHME; (b) HME100K)

CROHME数据集包括来自多年举办的在线手写数学公式识别比赛(CROHME)的数据，是目前最广泛使用的手写数学公式数据集。CROHME数据集的训练集有8 836个样本，而CROHME 2014、2016、2019测试集各自有986、1 147、1 199个样本。在CROHME数据集中，每个手写数学表达式以InkML格式存储，记录了手写笔画的轨迹坐标。在使用深度学习模型时，需要将InkML文件中的手写笔画轨迹信息转换为图像格式进行训练和测试。

HME100K数据集(Yuan等，2022)由真实场景中的手写数学公式组成，训练集和测试集分别包含74 502幅和24 607幅图像。此外，由于数据集是在真实场景条件下由相机拍摄生成，图像具有颜色变化、模糊、扭曲和复杂背景等特点，因此HME100K数

据集更加具有挑战性。

3.2 评估方法

公式识别率(ExpRate)(Zanibbi 等,2013)和结构识别准确率(StruRate)是手写数学公式识别领域中常用的评估指标。ExpRate 反映了公式中符号和结构被完全正确识别的公式占总公式数的百分比。为了增强对模型性能的评估,通常还会计算容错条件下的准确率,即≤ 1 error 和≤ 2 error,分别表示在最多容忍一个或两个符号错误时的公式识别准确率。

在计算 ExpRate 的过程中,模型的预测公式与标注公式首先被表示为由基本单元组成的有向标签图(label graph)。随后,该有向图被转换为邻接矩阵,邻接矩阵进一步被拆解为可供计算的子项,从而得到最终的公式识别率。

结构识别准确率(StruRate)则专注于衡量公式结构的正确识别情况,反映了模型在不考虑符号正确与否的情况下,对公式结构的识别能力。具体而言,StruRate 衡量的是模型预测公式的结构树与标注公式结构树的相似度,而不涉及对应符号的匹配。例如,“2 + 2” 和 “a = b”,虽然符号不同,但由于它们的结构相同,因此在 StruRate 的计算中会被视为具有相同的结构。

3.3 数据增强方法

不同于手写文本,手写数学公式存在复杂的二维结构,因此如果在保持长宽比不变的情况下,将公式图像归一化到相同的大小,会导致部分符号和子表达式的大小明显小于其他符号,这些符号或子表达式通常位于上标(superscript)或下标(subscript)。Li 等人(2020)提出 Scale Augmentation 的方法,在模型训练时,对训练数据中的图像的长宽随机使用不同的比例因子进行缩小和扩大,并使用零填充的方法将图像的大小固定在 256 × 1 024 像素。由此,在保持公式图像长宽比的同时,也达到了数据增强的效果。

Chen 等人(2022)利用公式的树结构信息和边界框(bounding box)信息对公式图像在符号、子公式、图像 3 个层次上进行了数据增强。在符号层次的数据增强上,使用符号的替换和删除,分别缓解手写符号风格多样和手写符号的间隔不同的问题;在子公式层次的数据增强上,使用子公式的符号替换和分解进一步丰富训练图像的多样性;在图像层次的数据增强方面,使用图像偏移和矩形的图像裁剪分别缓解数学公式二维结构带来的空间位置模糊和真实场景下的遮挡问题。

4 实 验

4.1 性能

在 CROHME 2014/2016/2019 测试集上各模型的识别性能表现如表 3 所示。为保证公平比较,所有模型均未使用数据增强,且训练数据均为 CROHME 数据集中 8 836 个训练样本。

对于传统方法,本文提供基于树形语法的 3 种传统手写数学公式识别方法在 CROHME 2014 中的结果作为基准,标记为 I、VI 和 VII(Mouchère 等,2014)。对于 CROHME 2016(Mouchère 等,2016),本文提供表现最佳的 TOKYO 方法作为基准。对于 CROHME 2019,本文提供 Univ. Linz 方法(Mahdavi 等,2019)作为基准。对于基于深度学习的方法,本文提供了以上介绍过的所有相关模型的结果。

如表 3 所示,尽管基于树结构解码的方法在手写公式识别任务中具有一定的可解释性,其实际性能仍整体落后于基于序列解码的方法。其原因可能在于,树结构预测任务涉及符号分类和符号关系预测这两个子任务,属于一个多任务学习问题。相比之下,单纯的序列预测任务更为简单,因此模型的优化难度较低。在树结构解码的情况下,模型需要同时学习公式的符号和结构信息,这使得训练和优化过程更为复杂,从而导致性能下降。另一方面,许多现有方法通过引入辅助任务与手写数学公式识别(HMER)任务进行联合优化,取得了显著的性能提升。辅助任务可以帮助模型更好地学习不同层次的信息,增强其对复杂结构和符号关系的理解能力。相比基准模型,这种联合优化策略不仅提升了模型的预测精度,还在处理复杂公式时展现出更强的鲁棒性。这表明,通过合理设计辅助任务和联合训练机制,模型在 HMER 任务中的性能仍有较大的提升空间。

在规模更大的 HME100K 数据集上训练和测试,部分基于深度学习方法的模型的性能如表 3 最后一列所示。相较 CROHME 数据集,现有方法受益于 HME100K 更大的训练集规模,性能表现更加出色。

表3 在CROHME 2014/2016/2019和HME100K测试集上的识别性能表现

Table 3 Recognition performance on the CROHME 2014/2016/2019 and HME100K test sets

/%

解码方式	模型	CROHME 2014			CROHME 2016			CROHME 2019			HME100K		
		ERate*	≤1*	≤2*	ERate*	≤1*	≤2*	ERate*	≤1*	≤2*	ERate*	≤1*	≤2*
传统方法	I	37.22	44.22	47.26	–	–	–	–	–	–	–	–	–
	VI	25.66	33.16	35.90	–	–	–	–	–	–	–	–	–
	VII	26.06	33.87	38.54	–	–	–	–	–	–	–	–	–
	TOKYO	–	–	–	43.94	50.91	53.70	–	–	–	–	–	–
	Univ. Linz	–	–	–	–	–	–	41.49	54.13	58.88	–	–	–
序列解码	WAP*	46.55	61.16	65.21	44.55	57.10	61.55	–	–	–	–	–	–
	DenseWAP*	50.10	–	–	47.50	–	–	–	–	–	61.85	70.63	77.14
	DenseWAP-MSA*	52.80	68.10	72.00	50.10	63.80	67.40	47.70	59.50	63.30	–	–	–
序列解码	WS WAP	53.65	–	–	51.96	64.34	70.10	–	–	–	–	–	–
	PALv2	48.88	64.50	69.78	49.61	64.08	70.27	–	–	–	–	–	–
	PrimCLR	51.90	–	–	54.80	–	–	54.90	–	–	–	–	–
树解码	DenseWAP-TD	49.10	64.20	67.80	48.50	62.30	65.30	51.40	66.10	69.10	62.60	79.05	85.67
	TDv2	53.62	–	–	55.18	–	–	58.72	–	–	–	–	–
	SAN	56.20	72.60	79.20	53.60	69.60	76.80	53.50	69.30	70.10	67.10	–	–
序列解码	ABM	56.85	73.73	81.24	52.92	69.66	78.73	53.96	71.06	78.65	65.93	81.16	87.86
	CAN	57.26	74.52	82.03	56.15	73.58	81.43	55.96	72.73	80.57	68.09	83.22	89.91
序列解码	BTTR	53.96	66.02	70.28	52.31	63.90	68.61	52.96	65.97	69.14	64.10	–	–
	$CoMER^T$	58.38	74.48	81.14	56.98	74.44	81.87	59.12	77.45	83.87	68.12	84.20	89.71
	GCN	60.00	–	–	58.94	–	–	61.63	–	–	–	–	–

注：ERate*表示ExpRate，≤1* 表示≤1 error，≤2* 表示≤2 error。为确保公平比较，模型均未使用数据增强。模型中，上标T表示本文未使用数据增强的复现结果，上标*表示原始论文在CROHME数据集上使用Ensemble方法的结果。“–”表示未提供相关信息。

4.2 多模态大模型性能表现

本文同样比较了多模态大模型在多个手写公式识别数据集上的表现，结果如表4所示。评测数据集包括CROHME 2014、CROHME 2016、CROHME 2019和HME100K。为了评估模型在不同公式复杂度下的表现，依据括号深度这一衡量公式复杂度的标准，从各数据集中分层随机抽取了100幅公式图像，共计400幅图像用于评测。

为了最大程度发挥多模态大模型的识别能力，本文针对不同模型使用了相应的提示语(prompt)。Qwen-VL-Max多模态大模型的提示语为“请使用LaTeX格式书写图像中公式的表达式，并将其置于\boxed{}”，GPT-4和Gemini-pro模型的提示语为“Please write out the expression of the formula in the image using LaTeX format and place it inside \boxed{}”。为了便于比较，还列出了复现的CoMER模型的识别性能。结果显示，现有多模态大模型在手写数学公式识别任务上的表现均不理想，且模型之间的性能差距不大。这可能归因于多模态大模型在训练过程中缺乏手写字符的训练集图像。此外，面对数学公式这种具有复杂二维结构的输入，多模态大模型显得力不从心，难以准确捕捉和解析其内部的结构关系。

4.3 补充实验

为了尽可能排除手写公式风格多样性带来的干扰，本文利用部分模型的官方实现代码，并按照原论

表4　多模态大模型在CROHME 2014/2016/2019测试集和HME100K测试集上的识别性能表现

Table 4　Recognition performance of large multimodal models on the CROHME 2014/2016/2019 test sets and the HME100K test set

/%

模型	CROHME 2014			CROHME 2016			CROHME 2019			HME100K		
	ExpRate	≤1*	≤2*	ExpRate	≤1*	≤2*	ExpRate	≤1*	≤2*	ExpRate	≤1*	≤2*
CoMER^{T}	67	76	79	51	71	81	57	74	82	75	88	91
GPT-4V	16	26	34	10	17	20	15	18	27	10	16	22
Qwen-VL-Max	15	20	26	4	5	9	5	8	11	12	16	20
Gemini-pro	16	26	30	14	18	24	12	20	25	7	14	20

注：≤1*表示≤1 error，≤2*表示≤2 error。模型中，上标T表示专用模型，用于与多模态大模型比较，结果未使用数据增强。

文提供的参数配置，在印刷体公式数据集上进行了性能评估，结果如表5所示。该印刷体数据集由CROHME 2014手写数学公式数据集的LaTeX公式渲染生成，并确保每幅印刷体公式图像与对应的手写公式图像在分辨率上保持一致。训练集和测试集的划分与CROHME数据集相同。实验结果显示，TDv2模型和CoMER模型在印刷体数据集上取得最优性能。表明在排除书写风格多样性干扰的情况下，模型对公式二维结构的建模能力以及对视觉模态和语言模态的对齐能力成为处理该任务的关键因素。这一结果进一步验证了相关模型在处理纯粹的公式结构时的有效性，表明未来的改进方向不仅应关注对手写风格的鲁棒性，还应深入探索如何更好地建模公式的结构特征。

表5　在CROHME 2014印刷体数学公式测试集上的识别性能表现

Table 5　Recognition performance on the CROHME 2014 test set with printed mathematical expression

模型	ExpRate/%
DenseWAP	89.04
TDv2	93.51
BTTR	85.32
ABM	91.67
CAN	91.68
SAN	88.64
CoMER^{T}	93.35

注：为确保公平比较，模型均未使用数据增强。模型中，上标T表示本文未使用数据增强的复现结果。

5　结　语

5.1　待解决的问题与未来发展方向

尽管深度学习技术在手写公式识别任务中取得了显著进展，但该领域仍面临若干关键挑战，是未来的研究方向。本文总结了当前主要的难点，并讨论了可能的改进策略，以期为后续研究提供参考。

1）现有模型在捕捉和转译细粒度数学符号的视觉信息方面仍存在不足。编码器在处理细节丰富的小符号时表现出较为有限的能力，导致解码器在生成公式时可能无法准确还原这些信息。因此，提升模型对细粒度视觉信息的捕捉能力，以及确保解码器能够精确转译这些细节，成为优化的关键方向。未来的研究可以探索通过引入更高分辨率的输入数据，或设计专门针对小符号处理的神经网络结构，以提高符号级别的识别精度。

2）手写数学公式的书写风格多样且符号大小不一，这给模型的准确性带来了较大的挑战。模型在预测视觉相似的符号时往往容易混淆，例如BTTR模型在某些情况下将“x”误识别为“X”。为了提升模型对不同风格手写公式的鲁棒性，未来可以进一步研究数据增强技术，从而使模型更好地适应风格多变的手写数据。同时，也可以开发更细致的分类网络以更有效地区分相似符号。

3）大多数现有方法依赖于直接回归LaTeX序列，然而由于训练数据集的规模有限，模型难以充分捕捉公式的内部结构，尤其是在处理长距离符号匹配（如栅栏符号对）的情况下表现不佳。尽管部分方法尝试通过建模树结构处理这些问题，但尚未显示出显著优于直接预测LaTeX序列的方法。未来的研

究可以考虑在大规模纯文本LaTeX数据集上对解码器进行预训练，并将其迁移到手写数学公式识别任务（HMER）中，或在模型训练时联合进行树结构预测和LaTeX序列生成任务。推理过程中，还可以引入基于树结构的评分机制，以更好地捕捉公式的逻辑结构和层次关系，从而提升模型的公式解析能力。

4）当前的手写数学公式识别研究主要集中于单行公式的识别任务，现有的数据集也多以单行手写数学公式为主。多行手写数学公式的数据集仍然十分稀缺，限制了对更复杂场景中公式识别的研究和发展。同时，在实际应用中，手写公式通常与印刷体文字描述共存于同一文档或页面中，形成混合的密集文本场景。然而，目前还缺少针对这种包含印刷体文字和手写公式共存的复杂文本的识别数据集。这种缺失阻碍了模型在真实应用场景中的适应性和鲁棒性研究。未来的研究方向可以致力于构建更丰富的数据集，涵盖多行公式和手写公式与印刷体文字混合的场景，以推动手写数学公式识别技术向更加实际、复杂的应用迈进。同时，开发能够同时处理这两种信息模式的模型将为手写公式识别领域带来更多创新和进展。

5）正如引言中提到的，手写数学公式识别任务面临4个主要研究难点：数学公式的嵌套结构、手写风格的多样性、缺乏覆盖问题以及长序列预测中的输出不平衡问题。这4个研究难点中，为了确保模型能够正确预测数学公式的嵌套结构（语法规范），可以采用树结构预测方法，或将树结构预测与序列预测相结合。缺乏覆盖问题则可以通过引入覆盖注意力机制来缓解，长序列预测的输出不平衡问题可以通过双向训练策略加以解决。尽管许多现有方法已尝试通过引入辅助任务和多尺度模型结构应对这些挑战，但手写风格多样性和相似字符的区分问题仍然没有得到有效解决。这一问题的根源在于手写数据集的规模较小，同时也提供了启示：可以借鉴其他手写识别任务（如手写中文识别和手写英文识别）中的方法，改进手写数学公式识别中的风格多样性问题，进一步提升模型的鲁棒性和准确性。

综上所述，尽管深度学习方法在手写公式识别中取得了长足的进展，但未来的研究仍需在细粒度符号识别、模型对风格多样性的适应性、公式结构的建模与解析方面以及构建更多真实场景下的数据集等方向上持续探索与改进。这些方向的研究有望进一步提升模型在实际应用中的性能和鲁棒性。

5.2 总结

本文首先系统回顾了传统手写公式识别方法的识别流程和存在的问题。在此基础上，重点探讨了基于深度学习的手写公式识别技术，详细梳理了围绕公式图像的视觉特征提取、视觉与文本特征的对齐以及文本输出的回归这3个子任务的研究进展。此外，文中还介绍了使用多模态大模型在手写数学公式识别数据集上的测试结果，并补充了这些典型方法在印刷体公式数据集上的表现。最后，针对手写公式识别目前所面临的挑战和困难，本文对未来的发展方向和研究趋势进行了前瞻性的展望，这些探讨将为研究人员提供宝贵的参考信息，推动手写公式识别技术的进一步发展。

参考文献（References）

Álvaro F, Sánchez J A and Benedí J M. 2014. Recognition of on-line handwritten mathematical expressions using 2D stochastic context-free grammars and hidden Markov models. Pattern Recognition Letters, 35: 58-67 [DOI: 10.1016/j.patrec.2012.09.023]

Bian X H, Qin B, Xin X Z, Li J W, Su X F and Wang Y F. 2022. Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 113-121 [DOI: 10.1609/aaai.v36i1.19885]

Chan K F and Yeung D Y. 1998. Elastic structural matching for online handwritten alphanumeric character recognition//Proceedings of the 14th International Conference on Pattern Recognition. Brisbane, Australia: IEEE [DOI: 10.1109/icpr.1998.711993]

Chan K F and Yeung D Y. 2000a. An efficient syntactic approach to structural analysis of on-line handwritten mathematical expressions. Pattern Recognition, 33(3): 375-384 [DOI: 10.1016/s0031-3203(99)00067-9]

Chan K F and Yeung D Y. 2000b. Mathematical expression recognition: a survey. International Journal on Document Analysis and Recognition, 3(1): 3-15 [DOI: 10.1007/pl00013549]

Chou P A. 1989. Recognition of equations using a two-dimensional stochastic context-free grammar//Proceedings of SPIE 1199, Visual Communications and Image Processing IV. Philadelphia, USA: SPIE [DOI: 10.1117/12.970095]

Fateman R J, Tokuyasu T, Berman B P and Mitchell N. 1996. Optical character recognition and parsing of typeset mathematics. Journal of Visual Communication and Image Representation, 7(1): 2-15 [DOI: 10.1006/jvci.1996.0002]

Guo H Y, Wang C, Yin F, Liu H Y, Wu J W and Liu C L. 2022. Primi-

tive contrastive learning for handwritten mathematical expression recognition//Proceedings of the 26th International Conference on Pattern Recognition. Montreal, Canada: IEEE: 847-854 [DOI: 10.1109/icpr56361.2022.9956214]

Ha J, Haralick R M and Phillips I T. 1995. Understanding mathematical expressions from document images//Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, Canada: IEEE: 956-959 [DOI: 10.1109/ICDAR. 1995. 602060]

Keshari B and Watt S. 2007. Hybrid mathematical symbol recognition using support vector machines//Proceedings of the 9th International Conference on Document Analysis and Recognition. Curitiba, Brazil: IEEE: #4377037 [DOI: 10.1109/ICDAR.2007.4377037]

Kosmala A, Rigoll G, Lavirotte S and Pottier L. 1999. On-line handwritten formula recognition using hidden Markov models and context dependent graph grammars//Proceedings of the 5th International Conference on Document Analysis and Recognition. Bangalore, India: IEEE: #791736 [DOI: 10.1109/ICDAR.1999.791736]

LeCun Y, Bengio Y and Hinton G. 2015. Deep learning. Nature, 521(7553): 436-444 [DOI: 10.1038/nature14539]

Li B H, Yuan Y, Liang D K, Liu X, Ji Z L, Bai J F, Liu W Y and Bai X. 2022. When counting meets HMER: counting-aware network for handwritten mathematical expression recognition//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 197-214 [DOI: 10.1007/978-3-031-19815-1_12]

Li Z, Jin L W, Lai S X and Zhu Y C. 2020. Improving attention-based handwritten mathematical expression recognition with scale augmentation and drop attention//Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), 175-180 [DOI:10.1109/ICFHR2020.2020.00041]

Lin Q Q, Huang X N, Bi N, Suen C Y and Tan J. 2022. CCLSL: combination of contrastive learning and supervised learning for handwritten mathematical expression recognition//Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer: 577-592 [DOI: 10.1007/978-3-031-26284-5_35]

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 [DOI: 10.11834/jig.221112]

Liu Y L, Yang B, Liu Q, Li Z, Ma Z Y, Zhang S and Bai X. 2024. TextMonkey: an OCR-free large multimodal model for understanding document. [EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2403.04473.pdf

Mahdavi M, Zanibbi R, Mouchère H, Viard-Gaudin C and Garain U. 2019. ICDAR 2019 CROHME + TFD: competition on recognition of handwritten mathematical expressions and typeset formula detection//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: #247 [DOI: 10.1109/ICDAR.2019.00247]

Mouchère H, Viard-Gaudin C, Zanibbi R and Garain U. 2014. ICFHR 2014 competition on recognition of on-line handwritten mathematical expressions (CROHME 2014)//Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition. Hersonissos, Greece: IEEE: #138 [DOI: 10.1109/ICFHR.2014.138]

Mouchère H, Viard-Gaudin C, Zanibbi R and Garain U. 2016. ICFHR2016 CROHME: competition on recognition of online handwritten mathematical expressions//Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition. Shenzhen, China: IEEE: #116 [DOI: 10.1109/ICFHR.2016.0116]

Okamoto M. 1991. Recognition of mathematical expressions by using the layout structure of symbols//Proceedings of the 1st International Conference Document Analysis and Recognition. Saint Malo, France: [s.n.]: 242-250

Shi Y X, Peng D Z, Liao W H, Lin Z N, Chen X H, Liu C Y, Zhang Y Y and Jin L W. 2023. Exploring OCR capabilities of GPT-4V (ision): a quantitative and in-depth evaluation. [EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2310.16809.pdf

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR

Truong T N, Nguyen C T, Phan K M and Nakagawa M. 2020. Improvement of end-to-end offline handwritten mathematical expression recognition by weakly supervised learning//Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition. Dortmund, Germany: IEEE: #42 [DOI: 10.1109/ICFHR2020.2020.00042]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vuong B Q, He Y L and Hui S C. 2010. Towards a web-based progressive handwriting recognition environment for mathematical problem solving. Expert Systems with Applications, 37(1): 886-893 [DOI: 10.1016/j.eswa.2009.05.091]

Vuong B Q, Hui S C and He Y L. 2008. Progressive structural analysis for dynamic recognition of on-line handwritten mathematical expressions. Pattern Recognition Letters, 29(5): 647-655 [DOI: 10.1016/j.patrec.2007.11.017]

Winkler H J. 1996. HMM-based handwritten symbol recognition using on-line and off-line features//Proceedings of 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Atlanta, USA: IEEE: #550767 [DOI: 10.1109/ICASSP.1996.550767]

Wu C J, Du J, Li Y Q, Zhang J S, Yang C, Ren B and Hu Y Q. 2022. TDv2: a novel tree-structured decoder for offline mathematical expression recognition//Proceedings of the 36th AAAI Conference

on Artificial Intelligence. Palo Alto, USA: AAAI: 2694-2702 [DOI: 10.1609/aaai.v36i3.20172]

Wu J W, Yin F, Zhang Y M, Zhang X Y and Liu C L. 2020. Handwritten mathematical expression recognition via paired adversarial learning. International Journal of Computer Vision, 128(10): 2386-2401 [DOI: 10.1007/s11263-020-01291-5]

Yang C, Du J, Zhang J S, Wu C J, Chen M J and Wu J J. 2022. Tree-based data augmentation and mutual learning for offline handwritten mathematical expression recognition. Pattern Recognition, 132(c): #108910 [DOI: 10.1016/j.patcog.2022.108910]

Ye J B, Hu A W, Xu H Y, Ye Q H, Yan M, Xu G H, Li C L, Tian J F, Qian Q, Zhang J, Jin Q, He L, Lin X and Huang F. 2023. UReader: universal OCR-free visually-situated language understanding with multimodal large language model//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore, Singapore: Association for Computational Linguistics: #187 [DOI: 10.18653/v1/2023.findings-emnlp.187]

Yuan Y, Liu X, Dikubab W, Liu H, Ji Z L, Wu Z Q and Bai X. 2022. Syntax-aware network for handwritten mathematical expression recognition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: #451 [DOI: 10.1109/CVPR52688.2022.00451]

Zanibbi R, Blostein D and Cordy J R. 2002. Recognizing mathematical expressions using tree transformation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(11): 1455-1467 [DOI: 10.1109/tpami.2002.1046157]

Zanibbi R, Mouchère H and Viard-Gaudin C. 2013. Evaluating structural pattern recognition for handwritten math via primitive label graphs//Proceedings of SPIE 8658, Document Recognition and Retrieval XX. Burlingame, USA: SPIE: #2008409 [DOI: 10.1117/12.2008409]

Zhang J S, Du J and Dai L R. 2018. Multi-scale attention with dense encoder for handwritten mathematical expression recognition//Proceedings of the 24th International Conference on Pattern Recognition. Beijing, China: IEEE: #8546031 [DOI: 10.1109/ICPR.2018.8546031]

Zhang J S, Du J, Zhang S L, Liu D, Hu Y L, Hu J S, Wei S and Dai L R. 2017. Watch, attend and parse: an end-to-end neural network based approach to handwritten mathematical expression recognition. Pattern Recognition, 71: 196-206 [DOI: 10.1016/j.patcog.2017.06.017]

Zhang J S, Du J, Yang Y X, Song Y Z, Wei S and Dai L R. 2020. A tree-structured decoder for image-to-markup generation//Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: PMLR: 11076-11085 [DOI: 10.5555/3524938.3525965]

Zhang X Y, Ying Han, Tao Y, Xing Y L and Feng G F. 2023. General category network: handwritten mathematical expression recognition with coarse-grained recognition task//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: 1-5 [DOI: 10.1109/ICASSP49357.2023.10097228]

Zhao W Q and Gao L C. 2022. CoMER: modeling coverage for transformer-based handwritten mathematical expression recognition//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: #23 [DOI: 10.1007/978-3-031-19815-1_23]

Zhao W Q, Gao L C, Yan Z Y, Peng S, Du L and Zhang Z Y. 2021. Handwritten mathematical expression recognition with bidirectionally trained transformer//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: #37 [DOI: 10.1007/978-3-030-86331-9_37]

作者简介

朱建华，男，硕士研究生，主要研究方向为公式识别。
E-mail：zhujianhuapku@pku.edu.cn
高良才，通信作者，男，副教授，主要研究方向为模式识别。
E-mail：glc@pku.edu.cn
赵文祺，男，硕士研究生，主要研究方向为公式识别。
E-mail：wenqizhao@stu.pku.edu.cn
彭帅，男，博士研究生，主要研究方向为自动数学推理。
E-mail：pengshuaipku@pku.edu.cn
胡鹏飞，男，硕士研究生，主要研究方向为公式识别。
E-mail：pengfeihu@mail.ustc.edu.cn
杜俊，男，副教授，主要研究方向为信号处理与模式识别。
E-mail：jundu@ustc.edu.cn