文章编号：1003-0077(2022)06-0001-19

文档智能：数据集、模型和应用

崔 磊，徐毅恒，吕腾超，韦福如

（微软亚洲研究院 自然语言计算组，北京 100080）

摘 要：文档智能是指通过计算机进行自动阅读、理解以及分析商业文档的过程，是自然语言处理和计算机视觉交叉领域的一个重要研究方向。近年来，深度学习技术的普及极大地推动了文档智能领域的发展，以文档版面分析、文档信息抽取、文档视觉问答以及文档图像分类等为代表的文档智能任务均有显著的性能提升。该文对于早期基于启发式规则的文档分析技术、基于统计机器学习的算法以及近年来基于深度学习和预训练的方法进行简要介绍，并展望了文档智能技术的未来发展方向。

关键词：文档智能；深度学习；多模态自然语言处理

中图分类号：TP391 **文献标识码**：A

Document AI: Benchmarks, Models and Applications

CUI Lei, XU Yiheng, LYU Tengchao, WEI Furu

(Natural Language Computing Group, Microsoft Research Asia, Beijing 100080, China)

Abstract: Document AI, or Document Intelligence, is a relatively new research topic that refers to the techniques to automatically read, understand and analyze business documents. It is an important interdisciplinary study involving natural language processing and computer vision. In recent years, the popularity of deep learning technology has greatly advanced the development of Document AI tasks, such as document layout analysis, document information extraction, document visual question answering, and document image classification etc. This paper briefly introduces the early-stage heuristic rule-based document analysis, statistical machine learning based algorithms, as well as the deep learning-based approaches especially the pre-training approaches. Finally, we also look into the future direction of Document AI.

Keywords: Document AI; deep learning; multimodal NLP

0 文档智能

文档智能(Document AI, or Document Intelligence)是近年来一项蓬勃发展的研究课题，同时也是实际的工业界需求，主要是指对于网页、数字文档或扫描文档所包含的文本以及丰富的排版格式等信息，通过人工智能技术进行理解、分类、提取以及信息归纳的过程。由于布局和格式的多样性、低质量的扫描文档图像以及模板结构的复杂性，文档智能成为一项非常具有挑战性的任务并获得相关领域的广泛关注。随着数字化进程的加快，文档、图像等载体的结构化分析和内容提取成为关乎企业数字化转型成败的关键一环，自动、精准、快速的信息处理对于生产力的提升至关重要。以商业文档为例，不仅包含了公司内外部事务的处理细节和知识沉淀，还有大量行业相关的实体和数字信息。人工提取这些信息不仅耗时、费力、精度低，而且可复用性也不高，因此，文档智能技术应运而生。文档智能技术深层次地结合了人工智能和人类智能，在金融、医疗、保险、能源、物流等多个行业均有不同类型的应用。例如，在金融领域，其可以实现财报分析和智能决策分析，为企业战略的制定和投资决策提供科学、系统的数据支撑；在医疗领域，其可以实现病例的数字化，提高诊断的精准度，并通过分析医学文献和病例的关联性，定位潜在的治疗方案。在财务领域，其可以实现发票和采购单的自动化信息提取，将大量非结构化文档进行自动结构化转换，并支撑大量下游业

收稿日期：2021-10-12 **定稿日期**：2022-03-28

务场景，节省大量人工处理时间开销。

在过去的 30 年中，文档智能的发展大致经历了三个阶段，从简单的规则启发式方法逐渐进化至神经网络的方法。20 世纪 90 年代初期，研究人员大多使用基于启发式规则的方法进行文档的理解与分析，通过人工观察文档的布局信息，总结归纳一些处理规则，对固定布局信息的文档进行处理。然而，传统基于规则的方法往往需要较大的人力成本，而且这些人工总结的规则可扩展性不强，因此研究人员开始采用基于统计学习的方法。2000 年以来，随着机器学习技术的发展和进步，基于大规模标注数据驱动的机器学习模型成了文档智能的主流方法，它通过人工设计的特征模板，利用有监督学习的方式在标注数据中学习不同特征的权重，以此来理解、分析文档的内容和布局。然而，虽然传统的文档理解和分析技术基于人工定制的规则或少量标注数据进行学习，这些方法虽然能够带来一定程度的性能提升，但由于定制规则和可学习的样本数量不足，其通用性往往不尽如人意，而且针对不同类别文档的分析迁移成本较高，这距离文档智能技术的实用化和产业化还有相当一段距离。近年来，随着深度学习技术的发展，以及大量无标注电子文档的积累，文档分析与识别技术进入了一个全新的时代。图 1 是在当前深度学习框架下文档智能技术的基本框架，其中不同类型的文档通过内容提取工具（HTML/XML 抽取、PDF 解析器、光学字符识别 OCR 等）将文本内容、位置布局信息和视觉图像信息组织起来，利用大规模预训练的深度神经网络进行分析，最终完成各项下游应用任务，包括文档版面分析、文档信息抽取、文档视觉问答以及文档图像分类等。深度学习技术的出现，特别是以卷积神经网络（CNN）、图神经网络（GNN）以及 Transformer 架构[1]为代表预训练技术的出现，彻底改变了传统机器学习需要大量人工标注数据的前提，更多地依赖大量无标注数据进行自监督学习，进而通过“预训练-微调”模式来解决文档智能相关的应用任务，取得了显著性突破。

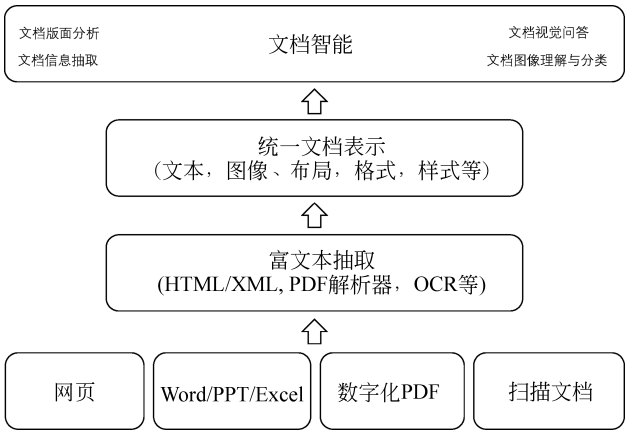


图 1 基于深度学习的文档智能技术框架

尽管深度学习极大地提高了文档智能技术的准确性，但是在实际应用中仍然有很多问题亟待解决。首先，受限于当前大规模预训练模型输入长度的限制，文档智能预训练模型通常需要将文档截断为几个部分，分别输入模型进行处理，这对于复杂长文档的多页跨页处理带来了极大的挑战。其次，由于实际场景中的扫描文档图像质量参差不齐，特别是人工标注的训练数据往往质量较高，而业务场景的文档图像由于扫描设备的清晰度、纸张褶皱和摆放位置的随意性，导致了性能不佳，因而需要利用更多数据增强技术来帮助现有模型提升性能。此外，当前文档智能各项任务通常是独立训练的，不同任务之间的关联性还未被有效地利用。例如，文档信息抽取和文档视觉问答有某些共性的语义表示，可以利用多任务学习框架更好地解决这类问题。最后，基于预训练的文档智能模型在实际应用中也遇到了计算资源和训练样本不足的问题，探索基于小模型的深度学习架构和模型压缩技术，以及少样本学习（Few-shot Learning）和零样本学习（Zero-shot Learning）技术也是当前重要的研究方向，并具有很大的实用价值。

接下来，我们首先将介绍当前主流的文档智能模型框架、任务和数据集，随后将分别重点介绍早期基于启发式规则的文档分析技术、基于传统统计机

器学习的算法模型，以及近年来基于深度学习，特别是基于多模态预训练技术的文档智能模型和算法，最后我们将展望文档智能技术的未来发展方向。

1　主流文档智能模型框架、任务及数据集

1.1　基于卷积神经网络的文档版面分析模型

近年来，卷积神经网络在计算机视觉领域取得了巨大的成功，特别是基于大规模标注数据集 ImageNet 和 COCO 的有监督预训练模型 ResNet[2] 在图像分类、物体检测以及场景分割任务上都带来了极大的性能提升。具体来讲，随着多阶段检测模型 Faster R-CNN[3] 和 Mask R-CNN[4] 等以及单阶段检测模型 SSD[5] 和 YOLO[6] 的普及，目标检测在计算机视觉中几乎成了已解决问题。文档版面分析本质上可以看作一种文档图像的物体检测任务，文档中的标题、段落、表格、插图等基本单元就是需要检测和识别的物体。Yang 等人[7] 将文档版面分析看作一个像素级分割任务，并尝试利用卷积神经网络进行像素分类取得很好的效果。Schreiber 等人[8] 首次将 Faster R-CNN 模型应用于文档版面分析中的表格识别任务，如图 2 所示，在 ICDAR 2013[9] 表格识别数据集上取得了 SOTA 的结果。然而，文档版面分析虽然是一个经典的文档智能任务，但是多年来一直受限于较小的数据集规模，仅仅套用经典计算机视觉预训练模型依然是不够的。随着大规模弱监督文档版面分析数据集 PubLayNet[10]、PubTabNet[11]、TableBank[12] 以及 DocBank[13] 的出现，研究人员可以对不同的计算机视觉模型和算法进行更为深入的比较和分析，进一步推动了文档版面分析技术的发展。

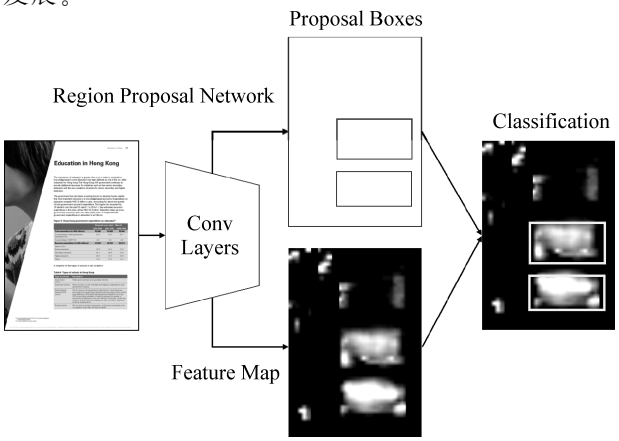
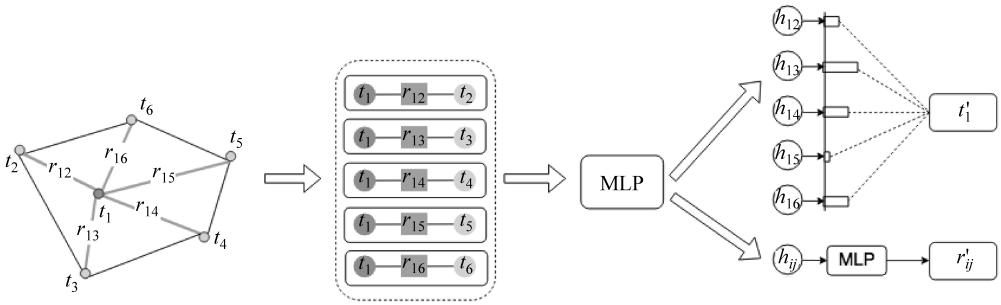


图 2　基于卷积神经网络 Faster R-CNN 的文档版面分析模型

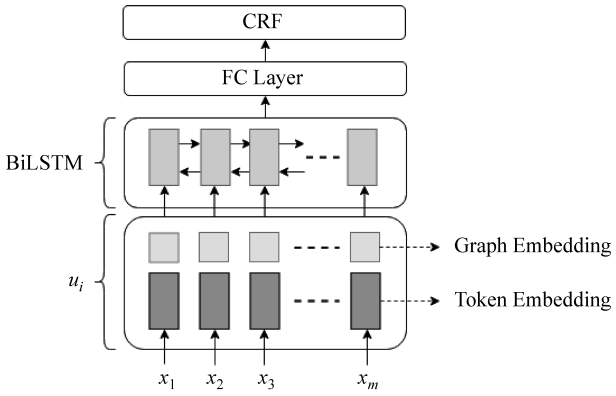
1.2　基于图神经网络的文档信息抽取模型

信息抽取是从非结构化文本中提取结构化信息的过程，其作为一个经典和基础的自然语言处理问题已经得到广泛研究。传统的信息抽取聚焦于如何从纯文本中提取实体与关系信息，却较少对视觉富文本进行研究。视觉富文本数据是指语义结构不仅由本文内容决定，也有与排版、表格结构、字体等视觉元素有关的文本数据。视觉富文本数据在生活中随处可见，例如，收据、证件、保险单等。Liu 等人[14] 提出利用图卷积神经网络对视觉富文本数据进行建模，如图 3 所示。每张图片经过 OCR 系统后会得到一组文本块，每个文本块包含其在图片中的坐标信息与文本内容。这项工作将这一组文本块构成全连接有向图，即每个文本块构成一个节点，每个节点都与其他所有节点有连接。节点的初始特征由文本块的文本内容通过 Bi-LSTM 编码得到。边的初始特征为邻居文本块与当前文本块的相对坐标与长宽信息，该特征使用当前文本块的高度进行归一化处理，具有仿射不变性。与其他图卷积模型仅在节点上进行卷积不同，这项工作更加关注在信息抽取中“个体-关系-个体”的三元信息，所以在“节点-边-节点”的三元特征组上进行卷积。除此之外，还引入了自注意力机制，让网络在全连接有向图构成的所有有向三元组中挑选更加值得注意的信息，并加权聚合特征。初始的节点特征与边特征经过多层卷积后得到节点与边的高层表征。

这项工作在两份真实商业数据上测试了所提出方法的效果，分别为增值税发票（VATI，固定版式，3 000 张）和国际采购收据（IPR，非固定版式，1 500 张）。使用了两个基准系统，基准系统Ⅰ为对每个文本块的文本内容独立做 BiLSTM＋CRF 解码，基准系统Ⅱ为将所有文本块的文本内容进行“从左到右、从上到下”的顺序拼接后，对拼接文本整体做 BiLSTM＋CRF 解码。实验表明，基于图卷积的模型在基准系统的基础上都有明显的性能提升，其中在仅依靠文本信息就可以抽取的字段（如日期）上与基准系统持平，而在需要依靠视觉信息做判断的字段（如价格、税额）上有较大的性能提升。此外，实验显示，视觉信息起主要作用，增加了语义相近文本的区分度。文本信息也对视觉信息起到一定的辅助作用。自注意力机制在固定版式数据上基本没有帮助，但是在非固定版式数据上有一定的性能提升。



(a)



(b)

图 3 基于图神经网络架构的文档信息抽取模型

1.3 基于 Transformer 结构的通用文档理解预训练模型

很多情况下，文档中文字的位置关系蕴含着丰富的语义信息。例如，表单通常是以键值对(Key-value Pair)的形式展示的。通常情况下，键值对的排布通常是左右或者上下形式，并且有特殊的类型关系。类似地，在表格文档中，表格中的文字通常是网格状排列，并且表头一般出现在第一列或第一行。通过预训练，这些与文本天然对齐的位置信息可以为下游的信息抽取任务提供更丰富的语义信息。对于富文本文档，除了文字本身的位置关系之外，文字格式所呈现的视觉信息同样可以帮助下游任务。对文本级(Token-level)任务来说，文字大小、是否倾斜、是否加粗，以及字体等富文本格式能够体现相应的语义。通常来说，表单键值对的键位(Key)通常会以加粗的形式给出。对于一般文档来说，文章的标题通常会放大加粗呈现、特殊概念名词会以斜体呈现等。对文档级(Document-level)任务来说，整体的文档图像能提供全局的结构信息，例如，个人简历的整体文档结构与科学文献的文档结构是有明显的视觉差异的。这些模态对齐的富文本格式所展现的视觉特征可以通过视觉模型抽取，结合到预训练阶段，从而有效地帮助下游任务。

为了利用上述信息，Xu 等提出了通用文档预训练模型 LayoutLM[15]，如图 4 所示。在现有的预训练模型基础上添加 2-D Position Embedding 和 Image Embedding 两种新的 Embedding 层，这样可以有效地结合文档结构和视觉信息。具体来讲，根据 OCR 获得的文本 Bounding Box，能够获取文本在文档中的具体位置。将对应坐标转化为虚拟坐标之后，计算该坐标对应在 x、y、w、h 四个 Embedding 子层的表示，最终的 2-D Position Embedding 为四个子层的 Embedding 之和。在 Image Embedding 部分，将每个文本相应的 Bounding Box 当作 Faster R-CNN中的候选框(Proposal)，从而提取对应的局部特征。特殊地，由于“[CLS]”符号用于表示整个输入文本的语义，同样使用整张文档图像作为该位置的 Image Embedding，从而保持模态对齐。

在预训练阶段，针对 LayoutLM 的特点提出两个自监督预训练任务：

- **掩码式视觉语言模型(Masked Visual-**

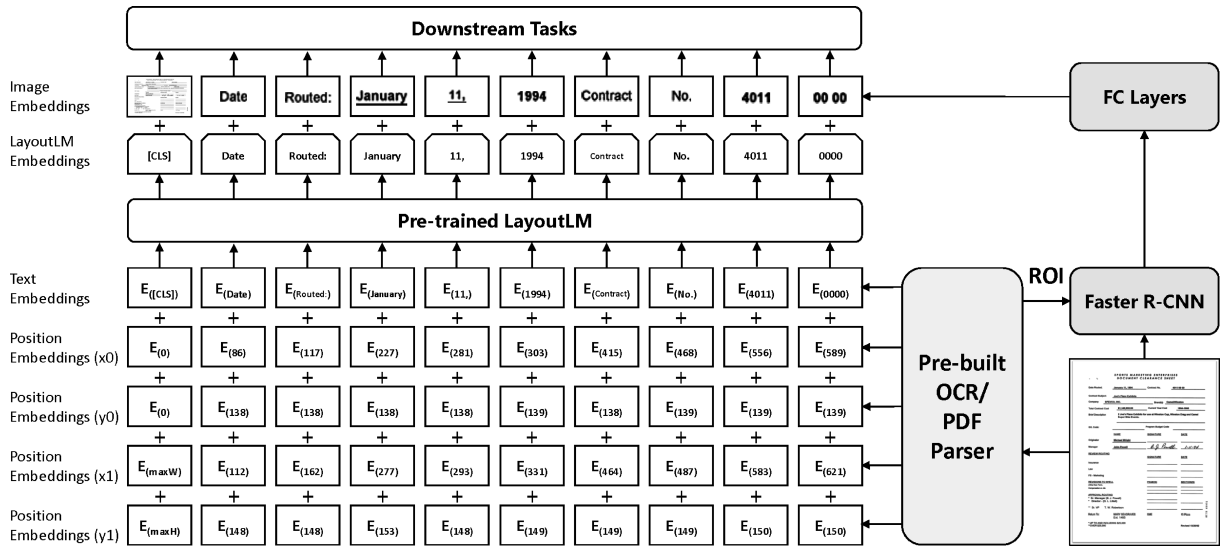


图 4　基于 Transformer 架构的通用文档理解预训练模型 LayoutLM

Language Model，MVLM）：大量实验已经证明 MLM 能够在预训练阶段有效地进行自监督学习。在此模型 MVLM 基础上进行了修改：在遮盖(Mask)当前词之后，保留对应的 2-D Position Embedding 暗示，让模型预测对应的词。在这种方法下，模型根据已有的上下文和对应的视觉暗示预测被遮罩的词，从而让模型更好地学习文本位置和文本语义的模态对齐关系。

• **多标签文档分类（Multi-label Document Classification，MDC）**：MLM 能够有效地表示词级别的信息，但是对于文档级的表示，需要文档级的预训练任务来引入更高层的语义信息。在预训练阶段使用 IIT-CDIP 数据集为每个文档提供了多标签的文档类型标注，同时引入 MDC 多标签文档分类任务。该任务使得模型可以利用这些监督信号去聚合相应的文档类别，并捕捉文档类型信息，从而获得更有效的高层语义表示。

实验结果表明，在预训练中引入的结构和视觉信息，能够有效地迁移到下游任务中。最终在多个下游任务中都取得了显著的准确率提升。与传统的基于卷积神经网络和图神经网络模型不同，通用文档智能预训练模型的优势在于可以支持不同类型的下游应用。

1.4　文档智能主流任务和数据集

文档智能涉及自动阅读、理解和分析文档的相关技术，在实际场景的应用中主要包括四大类任务，分别是：

• **文档版面分析**：指对文档版面内的图像、文本、表格信息和位置关系所进行的自动分析、识别和理解的过程。

• **文档信息抽取**：指从文档中大量非结构化内容中抽取实体及其关系的技术。与传统的纯文本信息抽取不同，文档的构建使得文字由一维的顺序排列变为二维的空间排列，因此文本信息、视觉信息和位置信息在文档信息抽取中都是极为重要的影响因素。

• **文档视觉问答**：指给定文档图像数据，利用 OCR 技术或其他文字提取技术自动识别影像资料后，通过判断所识别文字的内在逻辑，回答关于图片的自然语言问题。

• **文档图像分类**：指针对文档图像进行分析识别从而归类的过程。

对于这四种主要的文档智能任务，学术界和工业界也开源了大量相关的基准数据集，如表 1 所示。这也极大地推动了相关领域的研究人员构建新的算法模型，特别是当前基于深度神经网络的模型在这些任务上都有不俗的表现。接下来，本文将分别详细介绍在过去不同时期的经典模型和算法，包括基于启发式规则的文档分析技术、基于统计机器学习的文档分析技术和基于深度学习的通用文档智能模型，为大家提供参考。

表 1 文档智能领域主流任务(文档版面分析、文档信息抽取、文档视觉问答、文档图像分类)开源数据集

任务	数据集	支持语言	参考文献/链接
文档版面分析	ICDAR 2013	英文	[9]
	ICDAR 2019	英文	[16]
	ICDAR 2021	英文	[17]
	UNLV	英文	[18]
	Marmot	中文/英文	[19]
	PubTabNet	英文	[11]
	PubLayNet	英文	[10]
	TableBank	英文	[12]
	DocBank	英文	[13]
	TNCR	英文	[20]
	TabLex	英文	[21]
	PubTables	英文	[22]
	IIIT-AR-13K	英文	[23]
	ReadingBank	英文	[24]
文档信息抽取	SWDE	英文	[25]
	FUNSD	英文	[26]
	SROIE	英文	[27]
	CORD	英文	[28]
	EATEN	中文	[29]
	EPHOIE	中文	[30]
	DeepForm	英文	[31]
	Kleister	英文	[32]
	XFUND	中文/日文/西班牙文/法文/意大利文/德文/葡萄牙文	[33]
文档视觉问答	DocVQA	英文	[34]
	InfographicsVQA	英文	[35]
	VisualMRC	英文	[36]
	WebSRC	英文	[37]
	保险文本视觉问答	中文	https://bit.ly/36O2Vow
文档图像分类	Tobacco-3482	英文	[38]
	RVL-CDIP	英文	[39]

2 基于启发式规则的文档分析技术

基于启发式规则的文档分析技术大致可分为自顶向下、自底向上和混合模式三种方式。自顶向下方式将文档图片作为整体逐步将其划分为不同区域,以递归方式进行切割,直至区域分割至预定义的标准,通常为块或列。自底向上以像素或组件为基本元素单位,对基本元素进行分组、合并以形成更大的同质区域。自顶向下方式在特定格式下的文档中能够更快、更高效地分析文档。而自底向上方式虽需要耗费更多的计算时间,但通用性更强,可覆盖更

多不同布局类型的文档。混合方式则将其两者相结合以尝试产生更好的效果。

本节从自顶向下和自底向上两种角度出发，介绍基于 Projection Profile、Image Smearing、Connected Components 等方式的文档分析技术。

2.1 Projection Profile

Projection Profile 作为一种自顶向下的分析方式被广泛应用于文档分析。Nagy 等人[40]使用 Projection Profile 中的 *X*-*Y* 切割算法对文档进行切割，这一方式适用于具有固定文本区域和行距的结构化文本，但该方式对边界噪声敏感且无法在倾斜的文本上提供良好性能，对文档质量要求较高。Itay 等人[41]使用自适应局部投影方式计算文档的倾斜度，以尝试消除文本倾斜导致的性能下降，实验证明模型在倾斜和弯曲文本上得到了较为准确的结果。此外，还有很多 *X*-*Y* 切割算法的变体被提出以解决现存的缺陷，O′Gorman[42]将 *X*-*Y* 切割算法扩展至使用组件边界框的投影，Sylwester 等人[43]使用了编辑成本评估指标以指导模型进行分割，所有这些方法均在一定程度上提高了模型的性能。

Projection Profile 分析算法适用于结构化文本，尤其是曼哈顿（Manhattan）布局文档。在布局复杂、文本倾斜或包含边界噪声的文档上可能无法展现出良好的性能。

2.2 Image Smearing

Image Smearing 分析法指从一个位置向四周渗透，逐渐扩展至所有同质区域，以此确定页面当中的一个区域。Wong 等人[44]采用自顶向下策略，使用游长平滑算法（Run-length Smoothing Algorithm，RLSA）判断同质区域。将图像二值化后，像素值 0 表示背景，1 表示前景，当 0 周围的 0 数目小于指定阈值 *C* 时，该位置的 0 修改为 1，游长平滑算法通过这一操作将距离相近的前景内容合并为整体。这种方式可以逐步将字符合并为单词，单词合并为文本行，继而将范围不断延伸至整个同质区域。Fisher 等人[45]在此基础上对其做进一步改进，增加了除噪、倾斜矫正等预处理。此外，游长平滑算法的阈值 *C* 修改为依据动态算法进行调整，进一步提升模型的适应能力。Esposito 等人[46]采用了类似的方法，但操作对象由像素改为了字符框。Shi 等人[47]则是对图片中的每一个位置像素进行扩展，得到一个新的灰度图，随后进行抽取，在手写字体、文本倾斜等情况下仍能表现出良好的性能。

2.3 Connected Components

Connected Components 分析法作为一种自底向上的技术，推测最小粒度元素之间的关系，用于寻找同质区域，最终将区域分类为不同属性。Fisher 等人[45]采用 Connected Components 技术，找到每个组件的 *K* 近邻（*K* Nearest Neighbors，KNN）组件，通过互相之间的位置、角度等关系来推断当前区域属性。Saitoh 等人[48]判断并根据文档的倾角将文字合并成线，继而将线合并为区域，随后将其分类为不同的属性。Kise 等人[49]同样尝试解决文本的倾斜问题，作者采用了近似面积 Voronoi 图（Approximated Area Voronoi Diagram）来获得区域的候选边界，这一操作对于任意倾角的区域有效。但由于计算过程中需要估计字符间距和行内间距，因此当文档中包含大字体及宽字间距等情况时，模型并不能发挥出良好性能。此外，Bukhari 等人[50]也尝试在使用 Connected Components 的基础上使用 AutoMLP 以便寻找分类器最佳参数，进一步提升性能。

2.4 其他方法

除上文所述外，还有一些其他的启发式规则方法，例如，Baird 等人[51]采用自顶向下的方式按空白将文档进行切割划分区域。Xiao 等人[52]使用了 Delaunay Triangulation 算法进行文档分析，Bukhari 等人[53]在此基础上将其应用于书写随意的手写文档。此外还有一些混合算法，Okamoto 等人[54]通过分隔符和空白来切割块，在每个块中进一步将内部组件合并为文本行。Smith[55]将文档分析分成两部分，首先使用自底向上的方式来定位制表符，借助制表符推断列布局。随后在列布局上采用自顶向下的方式来推断结构和文本顺序。

3 基于统计机器学习的文档分析技术

传统的文档分析过程通常分为两阶段：①将文档图片切割，得到多个不同候选区域；②对区域进行属性分类，将其判别为文本、图像等规定类。基于机器学习的方法也通常从这两个角度入手，部分研究工作尝试使用机器学习算法参与文档的切割，其余则尝试在已生成的区域上构造特征，使用机器学习算法对区域进行分类。此外，由于统计机器学习技

术带来的性能上的提升，较多基于统计机器学习的方法在表格检测任务中被尝试使用，因表格检测是文档分析的一个重要子任务，本节也会对其进行一些介绍。因此与前文基于技术角度的阐述方式不同的是，从下文开始将会从文档分析中的任务角度来对其发展情况做出介绍。

3.1 文档分割

在文档切割过程中，Baechler 等人[56]结合 *X-Y* 裁剪算法，使用逻辑斯蒂回归对文档进行切割，丢弃空白部分。在得到相应区域后，实验比较了 *K* 近邻、逻辑斯蒂回归(Logistic Regression，LR)和最大熵马尔可夫模型(Maximum Entropy Markov Models，MEMM)等算法作为分类器的性能优劣，实验表明，最大熵马尔可夫模型和逻辑斯蒂回归在属性分类任务上可以展现出较好的性能。Esposito 等人[57]在文档分割过程中进一步加强机器学习算法在其中的参与程度。在自底向上的过程中，从字母到单词到文本行逐渐合并的过程中使用了一种基于内核的算法[58]，并将结果转换成 XML 结构存储。之后使用文档组织算法(Document Organization Composer，DOC)对文档进行分析。Wu 等人[59]则致力于文字同时存在两种阅读顺序的问题，此前的算法均假定文字只有一种书写方向，但遇到诸如汉语或日语等可以水平或者竖直方向书写的文字时无法正常地工作。该算法将文档分割分为四个步骤，用于判断并处理文本，并使用了支持向量机以决定是否执行步骤。

3.2 区域分类

在区域属性分类问题上，大量工作主要致力于尝试不同机器学习算法作为分类器输出结果。其中，Wei 等人[60]实验比较了支持向量机、多层感知机(Multi-Layer Perceptron，MLP)和高斯混合模型(Gaussian Mixture Models，GMM)几种机器学习算法作为分类器时的性能优劣，实验结果表明，支持向量机和多层感知机在区域属性上的分类性能明显优于高斯混合模型。Bukhari 等人[61]手动构造了多个特征，对区域抽取相应特征后使用 AutoMLP 算法进行分类，在阿拉伯语数据集中得到了 95% 的分割准确率。Baechler 等人[56]在文档分割上做了进一步改进，使用了金字塔形算法，在中世纪手稿上进行了三个不同级别的分析，最后使用动态多层感知机(Dynamic Multi-Layer Perceptron，DMLP)作为分类器。

3.3 表格检测

除上述方式之外，基于统计机器学习技术在表格识别领域存在大量研究。Wang 等人[62-64]使用了二叉树对文档进行自上而下的分析来查找表格候选区，继而根据区域特征确定最终表格区域。Pinto 等人[65]则使用了条件随机场在 HTML 页面中抽取表格区域，并确定表格中的标题、子标题等内容。Silva 等人[66]使用隐马尔可夫(Hidden Markov Models，HMMs)抽取表格区域。Chen 等人[67]在手写文档中检索表格区域，并使用支持向量机识别其中的文字区域，随后依据文本行确定表格所在位置。Kasar 等人[68]同样使用了支持向量机技术，首先识别图中水平和竖直的垂直线，随后使用支持向量机对每条线的属性进行分类，判断该线条是否属于表格。Barlas 等人[69]使用多层感知机对文档中的 Connected Components 进行分类，判断其是否为文本。Bansal 等人[70]使用 leptonica 库[71]对文档进行分割，随后对每一个区域构造包含周围环境信息的特征。使用 Fixed-point Model[72]对每一个区域进行分类，用以识别文档中的表格区域。它使得模型在分类过程中不再孤立地对区域进行分类，而是学习区域相互之间的关系。Rashid 等人[73]采用了与前一份工作相同的思路，但将操作粒度缩小为单词级别，对每一个词进行分类，之后使用 AutoMLP 来判断该词是否属于表格。

4 基于深度学习的文档智能技术

近年来，深度学习方法已经成为许多机器学习问题的解决范式。在众多研究领域，深度学习方法被证明是十分有效的。最近，预训练模型的流行也进一步发掘了深度神经网络的性能。而文档智能领域的发展也体现出同样的趋势。本节中我们将现存的模型分为针对特定任务的深度学习模型和支持多种下游任务的通用预训练模型进行介绍。

4.1 针对特定任务的深度学习模型

4.1.1 文档版面分析

文档版面分析包含两个主要的子任务：文档视觉结构分析和文档语义结构分析[74]。文档视觉分析的主要目的是检测文档结构并确定其同类区域的边界。而文档语义结构分析是需要为这些检测到的

区域标记具体的文档类别，如标题、段落、表格等。PubLayNet[10]是一个大规模的文档版面分析数据集，通过自动解析 PubMed 的 XML 文件构建了超过 360 000 个文档图片。DocBank[13]通过 arXiv 网站的 PDF 文件和 LaTeX 文件的对应关系自动构建了一个可扩展的文档版面分析数据集，同时支持对基于文本的方法和基于图像的方法进行评测。IIIT-AR-13K[23]提供了 13 000 的人工标注的文档图片用于版面分析。

1.1 节中介绍了将较为经典的卷积神经网络应用在文档版面分析领域的工作[2-8]，但随着对文档版面分析的性能要求逐渐提高，越来越多的科研工作针对文档这一领域对目标检测算法进行了针对性的改进。Yang 等人[7]将文档语义结构分析任务视为一个逐像素的分类问题。他们提出了一个同时考虑视觉和文本信息的多模态神经网络。Viana 等人[75]提出了一个用于移动和云服务的文档布局分析的轻量级模型。该模型使用图像的一维信息进行推理，并与使用二维信息的模型进行比较，在实验中取得了较高的准确性。Chen 等人[76]介绍了一种基于卷积神经网络(CNN)的手写历史文件图像的页面分割方法。Oliveira 等人[77]提出了一个基于 CNN 的多任务逐像素预测模型。Wick 等人[78]提出了一个用于历史文件分割的高性能全卷积神经网络(FCN)。Grüning 等人[79]提出了一种针对历史文献的两阶段文本行检测方法。Soto 等人[80]将上下文信息纳入 Faster R-CNN 模型。该模型利用文章元素内容的局部不变性质，提高了区域检测性能。

4.1.2　表格检测与表格结构识别

在文档版面分析中，表格理解是一项富有挑战性的任务。与标题、段落等文档元素相比，表格的格式通常较为多变，结构也较为复杂。因此，有大量的相关工作围绕表格展开，其中最为主要的两个子任务分别是表格检测和表格结构识别。表格检测是指确定文档中的表格的边界；表格结构识别是指将表格的语义结构，包括行、列、单元格的信息按照预定义的格式抽取出来。

近年来，有许多针对表格理解这一任务提出的数据集。UNLV[18]和 Marmot[19]是较早的表格识别数据集。ICDAR 会议在表格检测与识别上举办的多次竞赛提供了优质的表格数据集[9,16]。但这些传统表格数据集通常较小，难以发挥现代深度神经网络的优势，因此研究工作 TableBank[12]利用 LaTex 和 Office Word 来自动构建了一个大规模的表格理解数据集。此后，PubTabNet[11]提出了一个大规模表格数据集并提供了表格结构及单元格内容辅助表格识别。TNCR[20]在提供表格标注的同时提供了表格类别的标注。

针对表格理解这一任务的特性，许多目标检测方法在表格理解领域都能取得较好的效果。Faster R-CNN[3]在表格检测任务上直接应用就能取得非常好的性能。在此基础上，Siddiqui 等人[81]通过将可变形卷积应用在 Faster R-CNN 上获得了更好的性能。CascadeTabNet[82]使用了 Cascade R-CNN[83]模型同时完成表格检测和表格结构识别。TableSense[84]通过增加单元格特征、添加采样算法来显著提高表格检测能力。

除了上述两个主要的子任务，针对已解析后表格的理解也逐渐成为新的挑战。TaPas[85]是较早的将预训练技术引入表格理解任务的模型。通过引入额外的位置编码层，TaPas 可以使 Transformer[1]编码器接受结构化的表格输入。经过在大量的表格数据上进行掩码式预训练后，TaPas 在多种下游语义分析任务中显著超过了传统方法。继 TAPAS 后，TUTA 模型[86]引入了二维坐标树来表示结构化表格的层级信息，并针对这一结构提出了基于树结构的位置表示方式和注意力机制来显示建模层次化表格。结合不同层级的预训练任务，TUTA 在多个下游数据集上取得了进一步的性能提升。

4.1.3　文档信息抽取

文档信息抽取是指从大量非结构化富文本文档内容中抽取语义实体及其之间关系的技术。文档信息抽取任务，文档类别不同，抽取的目标实体也不尽相同。FUNSD[26]是一个文档理解数据集，其包含 199 张表单，每张表单中包含表单实体的键值对。CORD[28]是一个票据理解数据集，并包含 8 个大类共 54 小类种实体标签。Kleister[32]是一个针对长文档实体抽取任务的文档理解数据集，包含有协议和财务报表等长文本文档。DeepForm 数据集[31]是一个针对电视和有线电视政治广告披露表格的英文数据集。EATEN 数据集[29]是针对中文证件的信息抽取数据集，Yu 等人[87]在其 400 张子集上进一步添加了文本框标注。EPHOIE[30]数据集是一个针对中文文档数据的信息抽取数据集。XFUND[33]是随着 LayoutXLM 模型提出的针对 FUNSD 数据集的多语言扩展版本，包含有除英文以外的七种主流语言的富文本文档。

由于富文本文档具有丰富的视觉信息，所以很

多研究工作将文档信息抽取任务建模为计算机视觉任务,通过语义分割或文本框回归等任务进行信息抽取。考虑到文档信息抽取中文本信息同样具有重要作用,通常的框架是将文档图片视为像素网格,并在该特征图上添加文本特征来获得更好的特征表示。根据添加文本特征级别的不同,这一方法的基本发展顺序呈现出了从字符级别到单词级别再到上下文级别的趋势。Chargrid 模型[88]利用一个基于卷积的编码器-解码器网络,通过将字符进行 Onehot 编码来将文本信息融合到图像中。Visual-WordGrid 模型[89]实现了 Wordgrid,通过将字符级文本信息换成单词级的 Word2Vec 特征,并融合了一定的视觉信息,提高了抽取任务的性能。BERTgrid 模型[90]通过使用 BERT 获得了上下文文本表示,进一步提升了性能。ViBERTgrid 模型[91]在 BERTgrid 的基础上将 BERT 的文本特征较早地在卷积阶段与图像特征进行融合,从而获得了较好的效果。

由于富文本文档中的信息仍以文本作为主体,很多研究工作将文档信息抽取任务作为特殊的自然语言理解任务。Majumder 等人[92]根据抽取目标的类别来生成目标备选,在表单任务上取得了较好的效果。TRIE 模型[93]联合文本检测识别与信息抽取,让两个阶段的任务互相促进,从而获得更好的信息抽取效果。Wang 等人[94]通过三种不同模态信息的融合来预测文本片段之间的关系,实现了对表单的层次化抽取。

非结构化的富文本文档由多个邻接的文本片段组成,所以通常使用图网络对非结构化富文本文档进行表示。文档中的文本片段建模为图中的节点,而文本片段之间的关系则可建模为边,这样整个文档就可以被表示为一个图网络。在 1.2 节中,我们介绍了图神经网络在富文本文档中进行信息抽取的代表性工作[14]。在此基础上,逐渐有更多的研究工作基于图神经网络展开。Wang 等人[95]将文档建模为有向图,通过依存分析的方法对文档进行信息抽取。Riba 等人[96]使用基于图神经网络的模型来进行发票中表格的信息抽取。Wei 等人[97]通过在预训练模型的输出表示上使用图卷积神经网络来建模文本布局,提高了信息抽取的性能。Cheng 等人[98]通过将文档表示为图结构并使用基于图的注意力机制,结合 CRF 在小样本学习上取得了较好的性能。PICK 模型[87]通过引入一个可基于节点进行学习的图来表示文档,在发票抽取任务中取得了较好的性能。

4.1.4 文档图像分类

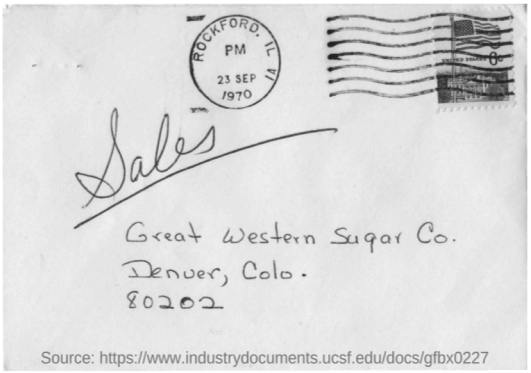
文档图像分类是指对文档图像进行归类标记的任务。RVL-CDIP[39]是该任务中的代表性数据集。该数据集包含 16 个文档图像类别共 400 000 张灰度图片。Tabacco-3482[38]选取了 RVL-CDIP 的一个子集进行评测,共包含 3 482 张文档灰度图片。

由于文档图像分类仍然属于图像分类的范畴,所以针对自然图片的分类算法同样能较好地解决文档图像分类的问题。Afzal 等人[99]介绍了一种基于深度卷积神经网络(CNN)的文档图像分类方法来进行文档图像分类。为了克服小数据集样本不足的问题,他们使用了经过 ImageNet 训练的 Alexnet 网络进行初始化,从而迁移到文档图像领域。Afzal 等人[100]尝试将 GoogLeNet、VGG、ResNet 等在自然图片领域获得成功的模型通过迁移学习的方式在文档图片上进行训练。Tensmeyer 等人[101]通过对模型参数和数据处理的调整,使 CNN 模型不借助从自然图片的迁移学习就能获得优于此前模型的性能。Das 等人[102]提出了一个基于不同区域分类的深度卷积神经网络框架用于文档图像分类。该方法通过对文档的不同区域分别进行分类,最终融合多个不同区域的分类器在文档图像分类上获得了明显的性能提升。Sarkhel 等人[103]通过引入金字塔形的多尺度结构来抽取不同层级的特征。Dauphinee 等人[104]通过对文档图片进行字符识别(OCR)获得文档的文本,并对图像特征和文本特征进行组合,进一步提升了分类性能。

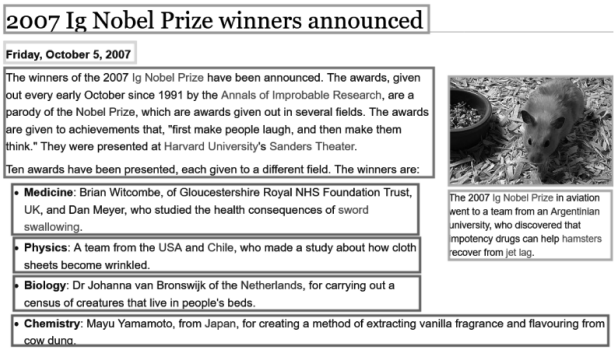
4.1.5 文档视觉问答

文档视觉问答是一个针对文档图片的高层理解任务。具体来说,给定一张文档图片和一个有针对性的问题,模型需要根据图片给出该问题的正确答案。具体的例子如图 5 所示。针对文档的视觉问答工作最早出现在数据集 DocVQA[34]中,该数据集包含了超过 12 000 个文档和对应的 5 000 个问题。后来,出现了针对文档中图表的视觉问答工作 Info-graphicVQA[35]。针对 DocVQA 数据集的答案较短、文档主题较单一的缺陷,有研究人员提出了 Vi-sualMRC[36]数据集。除了文档图片,针对网页视觉问答的 WebSRC[37]数据集也受到了广泛关注。

不同于传统 VQA 任务,文档视觉问答中的文档文本对任务具有关键作用,所以现存的代表性方法都将文档图片进行字符识别(OCR)处理得到的文档文本作为重要的信息。在得到文档文本后,针对不同数据的特点,视觉问答任务被建模为不同的



Q: Mention the ZIP code written?
A: 80202
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970
Q: Which company address is mentioned on the letter?
A: Great western sugar Co.



Q: Who were the winners of the Ig Nobel prize for Biology and Chemistry?
A: The winner of the Ig Nobel prize for biology was Dr Johanna van Bronswijk, and the winner for Chemistry was Mayu Yamamoto.

图 5　文档视觉问答任务示例
来自于 DocVQA 和 VisualMRC 数据集

问题。对于 DocVQA 数据来说，绝大部分的问题答案都是作为文本片段存在于文档文本中的，所以主流的方法都将其建模为了机器阅读理解问题（Machine Reading Comprehension，MRC）。通过为模型提供视觉特征和文档文本，模型根据问题在给定的文档文本上进行文本片段的抽取来作为问题答案。而对于 VisualMRC 数据集，问题的答案通常不蕴含在文档文本片段中，需要给出较长的抽象回答，因此在这种情况下，可行的方法是使用文本生成式的方法生成问题的答案。

4.2　支持多种下游任务的通用预训练模型

以上针对特定任务的深度学习方法，在针对特定文档理解任务上能够取得较好的性能，然而这些方法主要面临两个限制：①这些模型通常依赖于有限的标记数据，而忽视了挖掘大量无标注数据中的知识。对于文档理解任务，尤其是其中的信息抽取任务来说，详细标注的数据是昂贵且消耗时间的。另一方面，富文本文档在现实生活中大量使用，因而存在着大量的未标注文档，而这些大量的未标注数据可以使用自监督预训练加以利用。②富文本文档不仅有大量的文本信息，同时也包含丰富的版面和视觉信息。已有的针对特定任务的模型由于数据量的限制，通常只能通过预训练的 CV 模型或 NLP 模型来获取对应模态的特征，而且大部分工作只利用单一模态的信息或者两种特征的简单组合，而不是深度交互。Transformer[1] 在迁移学习领域的成功证明了深度上下文化（Contextualizing）对于序列建模的重要性，因此将文本和其他模态进行深度交互融合是一个较为明显的趋势。

富文本文档主要包含三种模态信息：文本、布局以及视觉信息，并且这三种模态在富文本文档中有天然的对齐特性。因此，如何对文档进行建模并且通过训练达到跨模态对齐是一个重要的问题。LayoutLM[15] 以及后续提出的 LayoutLMv2[105] 模型的提出正是针对这一方向进行的研究工作。在 1.3 节中，我们详细介绍了 LayoutLM 这一通用文档理解预训练模型，通过将文本和布局进行联合预训练，LayoutLM 在多种文档理解任务上取得了显著的性能提升。在此基础上，又有许多后续的研究工作对这一框架进行了针对性的改进。LayoutLM 在预训练过程中没有引入文档视觉信息，从而在 DocVQA 这类需要较强视觉感知能力的任务上效果欠佳。对此，LayoutLMv2[105] 通过将视觉特征信息融入预训练过程中，明显提高了模型的图像理解能力。具体来说，在结构方面，LayoutLMv2 引入了空间感知自注意力机制，并将视觉特征作为输入序列的一部分。在预训练目标方面，LayoutLMv2 在掩码视觉语言模型（Masked Visual-Language Model）之外又提出了文本-图像对齐（Text-Image Alignment）和文本-图像匹配（Text-Image Match）任务。通过在这两方面的改进，模型对于视觉信息的感知能力大大提高，并在包括 DocVQA 在内的六种下游任务中获得了显著性能提升。

LayoutLM提出之后,许多研究工作针对这一框架进行了针对性的改进,其中针对位置表达方式的改进是一个主要方向。许多工作将Embedding表示的位置编码改为了正余弦方式,其中有代表性的是BROS[106]和StructuralLM[107]。BROS[106]在绝对位置编码中使用了正弦函数,同时又在自注意力机制中通过正弦函数引入了文本相对位置信息,提高了模型对空间位置的感知能力。StructuralLM[107]在绝对位置表示方式上通过在文本块内共享相同的位置信息,帮助模型理解同一文本实体内的文本信息,从而对信息抽取任务有进一步的帮助。

除了对位置布局信息这一模态的改进之外,很多研究工作针对图像信息做了进一步的改进。LayoutLMv2的图像输入分辨率较低,这在某种程度上限制了模型对视觉信息的进一步挖掘。为此,许多研究工作针对视觉这一模态进行了优化和加强。LAMPRET[108]通过为模型提供更多的视觉模态信息如字体、字号、插图等,对网页文档进行建模,帮助模型对丰富的网页数据进行建模和理解。SelfDoc[109]采用了双流(Two-Stream)结构,针对给定的富文本文档数据,首先使用预先训练好的文档实体检测模型,通过目标检测将文档中所有的语义单元识别出来,然后使用OCR对识别的区域进行光学字符识别。针对识别出的图像区域和文本序列,模型分别使用了Sentence-BERT[110]和Faster-RCNN[3]进行了特征抽取,编码为特征向量,并使用一个跨模态的编码器进行编码,最终获得了多模态的表示来服务于下游任务。DocFormer[111]采用分离式多模态结构(Discrete Multi-Modal),在每层使用位置信息分别结合文本和图像模态使用自注意力机制。DocFormer首先使用ResNet对图像信息进行编码获得较高分辨率的图像特征,同时将文本信息以嵌入(Embedding)的形式编码为文本特征向量。位置信息向量分别与图像和文本信息相加,并单独传入Transformer层,每层分别编码之后重新相加。在这种机制下,不仅获取了高清图像信息,减小了输入序列,而且不同模态通过位置信息进行了对齐,使模型更好地建模了富文本文档的模态对齐关系。

许多模型在模态信息表示之外,又针对不同的模态设计了更丰富的预训练任务。例如,BROS[106]除了掩码式视觉语言模型(MVLM)之外,提出了基于区域的掩码式语言模型(Area-masked Language Modeling)。基于区域的掩码会对一个随机选择的区域内的所有文本块进行掩码操作。其可以被解释为将SpanBERT[112]中的针对一维文本的区间掩码操作扩展为二维空间中文本块的区间掩码。具体来说,该操作由以下四个步骤组成:①随机选择一个文本块,②通过扩大文本块的区域来确定一个最终区域,③确定属于该区域的文本块,④对文本块的所有文本进行掩码并预测它们。LAMPRET[108]额外引入的网页实体顺序排序任务,让模型通过对实体排布顺序的预测来学习空间位置进行预测。与此同时,模型还利用了图像匹配预训练任务,通过去除网页中的图像,并通过检索的方式进行匹配,提高了模型对多模态数据的语义理解能力。StructuralLM[107]提出的单元位置分类任务是对文档中文本块的相对空间位置进行建模。给定一组扫描的文件,该任务旨在预测文件中文本块的位置。首先,富文本文档被分成N个相同大小的区域。然后,模型通过文本块的中心二维位置,计算出该文本块所属的区域。这一研究工作较早地提出了针对位置信息进行掩码预测式学习。SelfDoc[109]和DocFormer[111]针对图像这一模态优化加强了输入的同时,也引入了对应的预训练任务,SelfDoc针对图像特征进行了掩码并预测,从而帮助模型学习建模视觉信息。DocFormer引入了一个解码器来对图像信息进行重建。在这种情况下,这项任务类似于自动编码器的图像重建,但又包含了文本和位置等多模态特征。在有图像和文本特征的情况下,图像重建需要两种模式的协作,加强了不同模态之间的交互。

在模型初始化方面,许多模型利用已有的更加强大的预训练语言模型进一步提高性能,同时也可以拓展模型的能力。例如,LAMBERT[113]通过使用RoBERTa[114]作为预训练初始化获得了更好的性能。除了语言理解之外,很多模型着眼于扩展模型的语言生成能力。它们的一个共同特点是都使用了编码-解码(Encoder-Decoder)范式。TILT[115]通过将Layout编码层引入T5[116]模型并结合文档数据预训练,使模型能够处理文档领域的生成任务。LayoutT5和LayoutBART[36]在文档视觉问答任务微调阶段在T5和BART[117]模型的基础上引入文本位置编码,来帮助模型理解并生成问题答案。

这些模型虽然在英文数据上取得了成功,但对于非英语世界来说文档理解任务同样重要。LayoutXLM[33]最早在多语言富文本文档上进行多语言预训练的研究工作。LayoutXLM基于LayoutLMv2的模型结构,通过使用53种语言进行预训练,扩展了LayoutLM

的语言支持。与此同时，相比于纯文本的跨语言模型，LayoutXLM 在迁移能力上具有明显优势，这证明了不仅多语言文本之间可以进行跨语言学习，而且多语言富文本文档之间也可以进行文档布局的迁移学习。

富文本文档通常可分为两类：第一类是固定布局的文件，如扫描的文档图像和数字原生的 PDF 文件，其布局和风格信息是预先渲染的，与软件、硬件或操作系统无关。这一特性使得现有的基于布局的预训练方法（LayoutLM）很容易适用于文档理解任务。第二类是基于标记语言的文档，如 HTML/XML 等，其布局和风格信息需要根据软件、硬件或操作系统进行交互和动态渲染以实现可视化。对基于标记语言的文档，二维布局信息并不以明确的格式存在，而是通常需要针对不同的设备动态呈现，例如移动/桌面/台式机，这使得目前基于布局的预训练模型难以应用。为此，MarkupLM[118] 在一个单一的框架中联合预训练文本和标记语言，用于基于标记语言的文档理解任务。与固定布局的文档不同，MarkupLM 为通过标记结构进行的文档表示学习提供了另一种视角，因为在预训练中不能直接使用二维位置信息和文档图像信息，而 MarkupLM 利用基于树形的标记结构来模拟文档中不同单元之间的关系，提高了标记语言文档理解问题的准确性。

除了通用多模态预训练模型之外，基于 ViT 视觉 Transformer[119-126] 的图像预训练技术近来取得了很大进展，研究人员通过有监督预训练方法或者自监督预训练等技术将视觉 Transformer 模型应用到图像分类、物体识别、场景分割等领域，取得了显著的进展。受自监督预训练视觉 Transformer 模型 BEiT[123] 的启发，Li 等提出一种自监督文档图像 Transformer 模型 DiT[127]，通过利用海量无标注文档图像数据进行大规模自监督预训练，在文档图像分类、文档版面分析、表格检测等任务均取得了最佳的结果。与自然图像理解领域不同，由于文档图像理解的研究并不存在类似于 ImageNet 这样的大规模人工标注数据集，因此无须人工标注数据的自监督预训练技术在文档智能领域将发挥越来越重要的作用。

5　未来发展方向

商业文档的自动阅读和分析具有明显的应用价值，是自然语言处理和计算机视觉交叉领域的一个重要研究方向。因此我们分别从自然语言处理、计算机视觉以及多模态融合的角度来梳理一下文档智能的未来发展方向。

从自然语言处理的角度出发，近年来以 BERT[128] 为代表的大规模自监督预训练成为自然语言处理的主流研究方向。与此同时，在大规模预训练模型基础上，以 GPT-3[129] 为代表的提示学习（Prompt Learning）研究方法；为文本预训练模型的应用给出一种新型的范式，能够达到低计算量与性能调优的平衡，受到了广泛关注。GPT-3 通过上下文学习（In-context Learning）的方法在零样本（Zero-shot）和少样本（Few-shot）学习中展现出与 BERT 完全不同的结论和性能，因此应该探究在文档智能领域大模型的性质，以及如何利用大模型进行文档智能下游任务的微调，例如 Parameter-efficient 相关的方法也是非常重要的。

文档智能中有大量以文档图片为载体的信息抽取和问答任务，如表单/发票理解等。由于这些任务所需的数据，人工标注代价很高，对自监督预训练模型有很强的需求。除此之外，如何降低模型参数微调（Fine-tuning）计算量也是这些任务亟待解决的问题，因此文档图像的提示学习技术也是未来十分重要的一个研究方向。

从计算机视觉的角度出发，以 ViT 视觉 Transformer[119] 为代表的大规模预训练技术近年来也成为计算机视觉的主流研究方向。由于文档图像理解领域不存在类似 ImageNet 这种大规模人工标注数据集，但无标注的文档图像却大量存在，因此自监督文档图像预训练模型对于文档智能领域的发展至关重要。文档智能领域中图像理解任务大多与版面分析相关，如光学字符识别（OCR）、文档对象识别，特别是表格识别等。传统的研究方法通常依赖任务相关的标注数据来解决，相信随着视觉自监督预训练模型的发展和成熟，对于标注数据的依赖会越来越小。

作为自然语言处理和计算机视觉的交叉领域，文档智能更多地应用了多模态融合技术。以 LayoutLM[15] 为代表的多模态文档智能预训练模型成为文档智能的主流研究方向。当前多模态融合主要采用将不同模态的信息通过跨模态对齐任务进行联合学习和预训练，取得了不错的效果。文档智能领域中的多数任务都会同时利用文本信息和图像信息，因此如何挖掘文本与图像之间的关联成为文档智能理解的重要任务。与此同时，不同模态之间的

互补性也将决定文档智能任务的精确度和可扩展性。

展望未来,除了解决文档多页跨页、训练数据质量参差不齐、多任务关联性较弱以及少样本零样本学习等问题,还应该特别关注文字检测识别 OCR 技术与文档智能技术的结合,因为文档智能下游任务的输入通常来自于自动文字检测和识别算法,文字识别的准确性往往对于下游任务有很大的影响。此外,如何将文档智能技术与现有人类知识以及人工处理文档的技巧相结合,也是未来值得探索的一个研究课题。

6 结语

信息处理是数字化转型的基础和前提,如今对处理能力、处理速度和处理精度也都有越来越高的要求。以商业领域为例,电子商业文档就涵盖了采购单据、行业报告、商务邮件、销售合同、雇佣协议、商业发票、个人简历等大量繁杂的信息。机器人流程自动化(Robotic Process Automation,RPA)行业正是在这一背景下应运而生,其利用人工智能技术帮助大量人工从繁杂的电子文档处理任务中解脱出来,并通过一系列配套的自动化工具提升生产力,RPA 的关键核心之一就是文档智能分析技术。过去的 20 年间,文档智能分析技术主要经历了三个阶段,从最初的基于启发式规则,过渡到基于统计机器学习的方法,到近来基于深度学习的方法,极大地提升了分析性能和准确率。与此同时我们也观察到,以 LayoutLM 为代表的大规模自监督通用文档智能预训练模型也越来越多地受到人们的关注和使用,逐步成为构建更为复杂算法的基本单元,后续研究工作也层出不穷,促使文档智能领域加速发展。

参考文献

[1] Vaswani A. Attention is all you need[C]//Proceedings of the 31st International Conferance on Neural Information Processing Systems, 2017: 5998-6008.

[2] He K. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[3] Ren S. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2016: 1137-1149.

[4] He K. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017.

[5] Liu Wei. SSD: Single shot multibox detector[G].Lecture Notes in Computer Science, 2016,9905: 21-37.

[6] Redmon J, Ali F. YOLOv3: An incremental improvement[EB/OL]. http://pjreddie.com/media/files/papers/YQL0v3.pdf[2021-08-29]

[7] Yang X. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4342-4351.

[8] Schreiber S. Deep DeSRT: Deep learning for detection and structure recognition of tables in document images [C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, 2017: 1162-1167.

[9] Göbel M C. ICDAR 2013 table competition[C]//Proceedings of the 12th International Conference on Document Analysis and Recognition, 2013: 1449-1453.

[10] Xu Z, Tang J, Antonio J Y. PubLayNet: Largest dataset ever for document layout analysis[C]//Proceedings of the International Conference on Document Analysis and Recognition, 2019. 1015-1022.

[11] Xu Z. Image-based table recognition: data, model, and evaluation[G]. Lecture Notes in Computer Science, 2020,12366: 564-80.

[12] Li M. TableBank: Table benchmark for image-based table detection and recognition[C]//Proceedings of the 12th Language Resources and Evaluation Conference. Marseille: European Language Resources Association, 2020: 1918-1925.

[13] Li M. DocBank: A benchmark dataset for document layout analysis[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: International Committee on Computational Linguistics, 2020: 949-960.

[14] Liu X J. Graph convolution for multimodal information extraction from visually rich documents.[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 32-39.

[15] Xu Y. Layout L M: pre-training of text and layout for document image understanding[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2020: 1192-1200.

[16] Gao L. ICDAR 2019 competition on table detection and recognition[C]//Proceedings of the International Conference on Document Analysis and Recognition, 2019: 1510-1515.

[17] Yepes A J, Xu Z, Douglas B. ICDAR 2021 competition on scientific literature parsing[C]//Proceedings of the Competition on Scientific Literature Parsing, 2021:605-617.

[18] Shahab A. An open approach towards the benchmarking of table structure recognition systems[C]//Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. New York, NY, USA: Association for Computing Machinery, 2010: 113-120.

[19] Fang J. Dataset, ground-truth and performance metrics for table detection evaluation[C]//Proceedings of the 10th IAPR International Workshop on Document Analysis Systems, 2012: 445-449.

[20] Abdallah A. TNCR: Table net detection and classification dataset[J]. arXiv preprint arXiv:2106.15322, 2021.

[21] Desai H. TabLeX: A benchmark dataset for structure and content information extraction from scientific tables[G]. Lecture Notes in Computer Science, 2021: 554-569.

[22] Smock B, Rohith P, Robin A. PubTables-1M: Towards a universal dataset and metrics for training and evaluating table extraction models[J]. arXiv preprint arXiv: 2110.00061v2, 2021.

[23] Mondal A, Peter L, Jawahar C V. IIIT-AR-13K: A new dataset for graphical object detection in documents[G]. Lecture Notes in Computer Science, 2020,12116: 216-230.

[24] Wang Z. Layout Reader: Pre-training of text and layout for reading order detection[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 4735-4744.

[25] Hao Q. From one tree to a forest: a unified solution for structured web data extraction[C]//Proceedings of the 34th International ACM SIGIR Conference on Research and development in Information Retrieval. Association for Computing Machinery, New York, NY, USA,2011: 775-784.

[26] Jaume G, Hazim K E, Jean Philippe T. FUNSD: A dataset for form understanding in noisy scanned documents[C]//Proceedings of the ICDARW, 2019: 1-6.

[27] Huang Z. ICDAR 2019 competition on scanned receipt OCR and information extraction[C]//Proceedings of the International Conference on Document Analysis and Recognition, 2019:1516-1520.

[28] Park S. CORD: A consolidated receipt dataset for post-OCR parsing[DB/OL].https://github.com/clovaai/cord[2021-08-29]

[29] Guo H. EATEN: Entity-aware attention for single shot visual text extraction[C]//Proceedings of the International Conference on Document Analysis and Recognition, 2019: 254-259.

[30] Wang J. Towards robust visual information extraction in real world: new dataset and novel solution[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 2738-2745.

[31] Stray J, Stacey S. Project DeepForm: extract information from documents[BD/OL]. https: wandb. ai/deepform/politicd-ad-extraction[2021-08-29]

[32] Stanisławek T. Kleister: Key information extraction datasets involving long documents with complex layouts[G].Lecture Notes in Computer Science, 2021: 564-79.

[33] Xu Y. XFUND: A benchmark dataset for multilingual visually rich form understanding[G].ACL 2022 Findings,2022: 3214-3224.

[34] Mathew M. DocVQA: A dataset for VQA on document images[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2021: 2200-2209.

[35] Mathew M. Infographics VQA[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2582-2591.

[36] Tanaka R, Kyosuke N, Sen Y. VISUALMRC: Machine reading comprehension on document images[C]//Proceedings of the AAAI Conference Artifical Intelligence, 2021,35(15): 13878-13888.

[37] Chen X. WebSRC: A dataset for web-based structural reading comprehension[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021:4173-4185.

[38] Kumar J, Peng Y, Doermann D. Structural similarity for document image classification and retrieval[J]. Pattern Recognit. Lett. 2014(43): 119-126.

[39] Harley A W, Alex U, Konstantinos G D. Evaluation of deep convolutional nets for document image classification and retrieval[C]//Proceedings of the 13th International Conference on Document Analysis and Recognition, 2015: 991-995.

[40] Nagy G, Sharad C S. Hierarchical representation of optically scanned documents[C]//Proceedings of the 7th International Conference on Pattern Recognition, 1984: 347-349.

[41] Itay B Y. Line segmentation for degraded handwritten historical documents[C]//Proceedings of the 10th International Conference on Document Analysis and Recognition. 2009: 1161-1165.

[42] O′gorman L. The document spectrum for page layout analysis[J]. IEEE Transactions on pattern analysis and machine intelligence,1993(15): 1162-1173.

[43] Sylwester D, Sharad S. A trainable, single-pass algorithm for column segmentation[C]//Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995: 615-618.

[44] Wong K Y, Richard G C, Friedrich M W. Document

analysis system[J].IBM journal of research and development, 1982,26(6): 647-656.

[45] Fisher J L, Stuart C H, Donald P, et al. A rule-based system for document image segmentation[C]// Proceedings of 10th International Conference on Pattern Recognition, 1990: 567-572.

[46] Esposito F. An experimental page layout recognition system for office document automatic classification: an integrated approach for inductive generalization [C]//Proceedings of 10th International Conference on Pattern Recognition, 1990: 557-562.

[47] Shi Z, Venu G. Line separation for complex document images using fuzzy runlength[C]//Proceedings of 1st International Workshop on Document Image Analysis for Libraries, 2004: 306-312.

[48] Saitoh T, Michiyoshi T, Toshifumi Y. Document image segmentation and text area ordering[C]//Proceedings of 2nd International Conference on Document Analysis and Recognition, 1993: 323-329.

[49] Kise K, Akinori S, Motoi I. Segmentation of page images using the area Voronoi diagram[J].Computer Vision and Image Understanding, 1998(70): 370-382.

[50] Bukhari S S. Document image segmentation using discriminative learning over connected components[C]// Proceedings of the 9th IAPR International Workshop on Document Analysis Systems,2010: 183-190.

[51] Baird H S, Susan E J, Steven J F. Image segmentation by shape-directed covers[C]//Proceedings of 10th International Conference on Pattern Recognition, 1990: 820-825.

[52] Xiao Y, Hong Y. Text region extraction in a document image based on the Delaunay tessellation[J]. Pattern Recognition,2003 (36): 799-809.

[53] Bukhari S S, Faisal S, Thomas M B. Script-independent handwritten textlines segmentation using active contours[C]//Proceedings of the 10th International Conference on Document Analysis and Recognition, 2009: 446-450.

[54] Okamoto M, Makoto T. A hybrid page segmentation method[C]//Proceedings of the 2nd International Conference on Document Analysis and Recognition, 1993: 743-746.

[55] Smith R W. Hybrid page layout analysis via tab-stop detection[C]//Proceedings of the 10th International Conference on Document Analysis and Recognition, 2009: 241-245.

[56] Baechler M, Rolf I. Multi resolution layout analysis of medieval manuscripts using dynamic MLP[C]// Proceedings of the International Conference on Document Analysis and Recognition. 2011: 1185-1189.

[57] Esposito F. Machine learning for digital document processing: from layout analysis to metadata extraction[M]. Machine learning in document analysis and recognition. Springer, 2008: 105-138.

[58] Dietterich T G, Richard H L, Tomás L P. Solving the multiple instance problem with axis-parallel rectangles[J]. Artificial intelligence,1997(89): 31-71.

[59] Wu C C,Chou C H, Fu C. A machine-learning approach for analyzing document layout structures with two reading orders[J]. Pattern recognition, 2008 (41): 3200-3213.

[60] Wei H. Evaluation of SVM, MLP and GMM classifiers for layout analysis of historical documents[C]// Proceedings of the 12th International Conference on Document Analysis and Recognition, 2013: 1220-1224.

[61] Bukhari S S. Layout analysis for Arabic historical document images using machine learning[C]//Proceedings of the International Conference on Frontiers in Handwriting Recognition, 2012: 639-644.

[62] Wang Y, Robert H, Ihsin T P. Improvement of zone content classification by using background analysis [C]//Proceedings of the 4th IAPR International Workshop on Document Analysis Systems. 2000.

[63] Wang Y, Ihsin T P, Robert Haralick. Automatic table ground truth generation and a background-analysis-based table structure extraction method[C]//Proceedings of 6th International Conference on Document Analysis and Recognition. 2001: 528-532.

[64] Wang Y, Ihsin T P, Robert M H. Table detection via probability optimization[C]//Proceedings of the International Workshop on Document Analysis Systems, 2002: 272-282.

[65] Pinto D. Table extraction using conditional random fields[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, 2003: 235-242.

[66] Silva E, Ana C. Learning rich hidden Markov models in document analysis: table location [C]//Proceedings of the 10th International Conference on Document Analysis and Recognition, 2009: 843-847.

[67] Chen J, Daniel L. Table detection in noisy off-line handwritten documents[C]//Proceedings of the International Conference on Document Analysis and Recognition, 2011: 399-403.

[68] Kasar T, et al. Learning to detect tables in scanned document images using line information[C]//Proceedings of the 12th International Conference on Document Analysis and Recognition, 2013: 1185-1189.

[69] Barlas P. A typed and handwritten text block segmentation system for heterogeneous and complex documents[C]//Proceedings of the 11th IAPR International Workshop on Document Analysis Systems, 2014: 46-50.

[70] Bansal A, Gaurav H, Sumantra D R. Table extrac-

tion from document images using fixed point model [C]//Proceedings of the Indian Conference on Computer Vision Graphics and Image Processing, 2014: 1-8.

[71] Bloomberg D S. Multiresolution morphological approach to document image analysis[C]//Proceedings of the International Conference on Document Analysis and Recognition, Saint-Malo, France, 1991.

[72] Li Q. Fixed-point model for structured labeling[C]//Proceedings of the International Conference on Machine Learning, 2013: 214-221.

[73] Rashid S F. Table recognition in heterogeneous documents using machine learning [C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, 2017: 777-782.

[74] Binmakhashen G M, Sabri A M. Document layout analysis: a comprehensive survey[J]. ACM computing surveys, 2019(52): 1-36.

[75] Viana M P, Dário A B O. Fast CNN-based document layout analysis[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017: 1173-1180.

[76] Chen K. Convolutional neural networks for page segmentation of historical document images [C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, 2017: 965-970.

[77] Oliveira S A, Benoit S, Frederic K. dhSegment: A generic deep-learning approach for document segmentation [C]//Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition, 2018: 7-12.

[78] Wick C, Frank P. Fully convolutional neural networks for page segmentation of historical document images[C]//Proceedings of the 13th IAPR International Workshop on Document Analysis Systems, 2018: 287-292.

[79] Grüning, T. A two-stage method for text line detection in historical documents[J]. International Journal on Document Analysis and Recognition, 2019(22): 285-302.

[80] Soto C, Shinjae Y. Visual detection with context for document layout analysis [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong: Association for Computational Linguistics, 2019: 3462-3468.

[81] Siddiqui S A. Decnt: Deep deformable CNN for table detection[J]. IEEE Access, 2018(6): 74151-74161.

[82] Prasad D. Cascade TabNet: 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.

[83] Cai Z, Nuno V. Cascade R-CNN: Delving into high quality object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6154-6162.

[84] Dong H. Tablesense: Spreadsheet table detection with convolutional neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 69-76.

[85] Herzig J. TaPas: Weakly supervised table parsing via pre-training [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[86] Wang Z. TUTA: Tree-based transformers for generally structured table pre-training[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021.

[87] Yu W. PICK: Processing key information extraction from documents using improved graph learning-convolutional networks[C]//Proceedings of the 25th International Conference on Pattern Recognition, 2021: 4363-4370.

[88] Katti A R. Chargrid: Towards understanding 2D documents[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 4459-4469.

[89] Kerroumi M. VisualWordGrid: Information extraction from scanned documents using a multimodal approach [G]. Lecture Notes in Computer Science, 2021: 389-402.

[90] Denk T I, Christian R. BERTgrid: Contextualized embedding for 2D document representation and understanding [C]//Proceedings of Document Intelligence Workshop of 33rd Conference on Neural Information Processing Systems,2019.

[91] Lin W. ViBERTgrid: A jointly trained multi-modal 2D document representation for key information extraction from documents[C]//Proceedings of Lecture Notes in Computer Science, 2021: 548-63.

[92] Majumder B P. Representation learning for information extraction from form-like documents[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6495-6504.

[93] Zhang P. Trie: End-to-end text reading and information extraction for document understanding[C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 1413-1422.

[94] Wang Z. DocStruct: A multimodal method to extract hierarchy structure in document for general form understanding[C]//Proceedings of the Association for Computational Linguistics,2020.

[95] Wang H W. Spatial dependency parsing for semi-structured document information extraction[C]// Proceedings of the Association for Computational Linguistics,2021.

[96] Riba P. Table detection in invoice documents by graph neural networks[C]//Proceedings of the International Conference on Document Analysis and Recognition, 2019: 122-127.

[97] Wei M, He Y, Zhang Q. Robust layout-aware IE for visually rich documents with pre-trained language models[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020: 2367-2376.

[98] Cheng M. One-shot text field labeling using attention and belief propagation for structure information extraction[C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 340-348.

[99] Afzal M Z. Deep doc classifier: document classification with deep convolutional neural network[C]// Proceedings of the 13th International Conference on Document Analysis and Recognition, 2015: 1111-1115.

[100] Afzal M Z. Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification[C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, 2017: 883-888.

[101] Tensmeyer C, Tony M. Analysis of convolutional neural networks for document image classification [C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, 2017: 388-393.

[102] Das A. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks[C]//Proceedings of the 24th International Conference on Pattern Recognition, 2018: 3180-3185.

[103] Sarkhel R, Arnab N. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019.

[104] Dauphinee T, Nikunj P, Mohammad R. Modular multimodal architecture for document classification [J].arXiv preprint arXiv:1912.04376, 2019.

[105] Xu Y. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online: Association for Computational Linguistics, 2021: 2579-2591.

[106] Hong T. BROS: A pre-trained language model for understanding texts in document[C]//Proceedings of ICLR, 2021:1-17.

[107] Li C. Structural LM: Structural pre-training for form understanding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 6309-6318.

[108] Wu T. LAMPRET: Layout-aware multimodal pretraining for document understanding[J]. arXiv preprint arXiv:2104.08405, 2021.

[109] Li P. SelfDoc: Self-supervised document representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 5652-5660.

[110] Reimers N, Iryna G. Sentence-BERT: Sentence embeddings using siamese BERT-networks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong: Association for Computational Linguistics, 2019: 3982-3992.

[111] Appalaraju S. DocFormer: End-to-End transformer for document understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021:973-983.

[112] Joshi M. SpanBERT: Improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics, 2020 (8): 64-77.

[113] Garncarek Ł. LAMBERT: Layout-aware language modeling for information extraction[G]. Lecture Notes in Computer Science, 2021: 532-547.

[114] Liu Y. RoBERTa: A robustly optimized BERT pretraining approach[J]. arXiv preprint arXiv: 1907.11692, 2019.

[115] Powalski R. Going full-TILT boogie on document understanding with text-image-layout transformer [G].Lecture Notes in Computer Science, 2021: 732-747.

[116] Raffel C. Exploring the limits of transfer learning with a unified text-to-text transformer[J]Journal of Machine Learning Research, 2020(21): 1-67.

[117] Lewis M. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7871-7880.

[118] Li J. MarkupLM: Pre-training of text and markup language for visually-rich document understanding [C]//Proceedings of the 60th ACL, 2022: 6078-6087.

[119] Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale[C]//

Proceedings of the ICLR, 2021: 1-22.
[120] Touvron H. Training data-efficient image transformers and distillation through attention[C]//Proceedings of the International Conference on Machine Learning, 2021: 10347-10357.
[121] Liu Z. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.
[122] Chen X. An empirical study of training self-supervised vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 9640-9649.
[123] Bao H. BEiT: BERT pre-training of image transformers[J]. arXiv preprint arXiv: 2106.08254, 2021.
[124] El-nouby A. XCiT: Cross-covariance image transformers[C]//Proceedings of the NeurIPS, 2021: 1-14.
[125] He K. Masked autoencoders are scalable vision learners[G]. Masked Autoencoders Are Scalable Vision Learners. CVPR, 2022: 16000-16009.
[126] Zhou J. iBOT: Image BERT pre-training with online tokenizer[J]. arXiv preprint arXiv: 2111.07832, 2021.
[127] Li J. DiT: Self-supervised pre-training for document image transformer[J]. arXiv preprint arXiv: 2203.02378, 2022.
[128] Devlin J. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the NAACL, 2019: 4171-4186.
[129] Brown T B. Language models are few-shot learners[C]//Proceedings of the NeurIPS, 2020, 2: 1877-1901.

崔磊(1986—)，通信作者，博士，高级研究员，主要研究领域为自然语言处理。
E-mail: lecu@microsoft.com

吕腾超(1995—)，硕士，研究员，主要研究领域为自然语言处理。
E-mail: tengchaolv@microsoft.com

徐毅恒(1998—)，博士研究生，主要研究领域为自然语言处理。
E-mail: t-yihengxu@microsoft.com