

机器阅读技术概述

摘要

机器阅读技术作为人工智能领域的重要分支，旨在赋予计算机理解和处理文本信息的能力。本文系统分析了机器阅读的技术原理，深入探讨了从光学字符识别到语义理解的完整技术链条。研究发现，现代机器阅读系统主要通过**光学字符识别（OCR）技术**将图像文字转换为可编辑文本，结合**自然语言处理（NLP）技术**实现语义理解，并通过**深度学习模型**提升理解精度。本文梳理了机器阅读从早期模板匹配到深度学习的发展历程，分析了 CNN、Transformer 等核心技术架构，探讨了 DocTR、LayoutLM 等主流模型的技术特点，总结了在医疗、金融、教育等领域的广泛应用。研究表明，尽管机器阅读技术在准确性和效率方面取得了显著进展，但在低质量图像识别、复杂文档理解、多语言处理等方面仍面临挑战。未来发展趋势将聚焦于多模态融合、实时处理能力提升以及个性化阅读体验优化。

引言

在数字化转型加速的时代背景下，海量文档和图像数据的自动化处理需求日益迫切。机器阅读技术作为连接物理世界与数字智能的桥梁，能够将静态的文本图像转换为可分析、可搜索的数字信息，为各行业的智能化升级提供关键技术支撑。

机器阅读技术的发展历程可追溯至 20 世纪初期。1929 年，德国科学家古斯塔夫·蒂策尔首次提出利用光电元件识别字符的设想，开启了 OCR 技术的先河^[14]。随后，从 1960 年代的早期 OCR 系统到 1974 年雷·库兹韦尔开发的首个“全字体 OCR 系统”，再到 21 世纪深度学习技术带来的革命性突破，机器阅读技术经历了从简单字符识别到复杂语义理解的演进过程^[63]。

当前，机器阅读技术已广泛应用于医疗、金融、教育、法律等多个领域，在提高工作效率、降低人工成本、改善用户体验等方面发挥着重要作用。然而，面对低质量图像、复杂文档布局、多语言混合等复杂场景，现有技术仍存在诸多挑战。本文旨在系统梳理机器阅读的技术原理、发展现状及未来趋势，为相关研究和应用提供参考。

一、机器阅读技术的发展历程与分类体系

1.1 技术发展的四个阶段

机器阅读技术的发展可以划分为四个关键阶段，每个阶段都代表着技术范式的重要转变。

**** 早期概念阶段（1914-1930 年代）**** 标志着机器阅读思想的萌芽。1914 年，物理学家伊曼纽尔·戈德伯格发明了“读取字符并转换为电报编码”的机器，随后在 20 年代进一步创造了“统计机器”，利用光学代码识别系统检索缩微胶卷档案，成为自动化文档检索的开端。1929 年，德国科学家古斯塔夫·蒂策尔申请了美国专利 1,915,993，题为《字符电识别方法与装置》，首次提出了利用光电技术自动识别印刷字符的全新概念(17)。

**** 首批商业系统阶段（1950-1970 年代）**** 见证了 OCR 技术从实验室走向市场的关键转变。1954 年，IBM 公司的工程师雷诺·柴雷开发出首款可识别英文字母的电子 OCR 系统，通过阴极射线管扫描字符并与预设模板比对识别(62)。1959 年，IBM 正式提出“Optical Character Recognition”术语，并将其应用于银行支票处理、邮政信件分拣等领域。然而，这一时期的系统仅能识别预先定义的特定字体，通用性极差。

*****“全字体”革命阶段（1974 年）**** 带来了技术的重大突破。1974 年，发明家雷·库兹韦尔开发出世界首个“全字体 OCR（Omni-font OCR）”系统，可识别几乎任何字体，这一里程碑突破极大拓宽了 OCR 应用范围(63)。值得注意的是，库兹韦尔的初衷是为盲人制造“阅读机”，体现了技术发展背后的人文关怀。

**** 深度学习时代（2000 年至今）**** 标志着机器阅读技术的范式转移。进入 21 世纪，传统基于“模板匹配 + 特征工程”的方法逐渐被端到端深度学习模型取代。深度神经网络的引入使 OCR 在准确性上实现质的飞跃，尤其在处理手写体、自然场景文本等复杂任务时取得前所未有的成功。

1.2 技术分类体系

机器阅读技术的分类体系可以从多个维度进行构建，反映了技术发展的不同路径和应用需求。

从**技术实现角度**，机器阅读系统可分为基于规则的方法和基于统计的方法(58)。基于规则的方法依赖于人工制定的语法规则和语义规则，适用于结构化文本如法律文件或医学文档，但需要手动编写大量规则且难以处理复杂语义。基于统计的方法则主要依赖于大量文本数据训练出的统计模型，能够处理更广泛的文本类型，包括新闻文章和用户评论等。

从**任务类型角度**，机器阅读理解任务主要包含四种类型：选择型、填充型、抽取型和生成型(29)。选择型阅读理解要求模型从若干候选答案中选择正确答案；填充型阅读理解是给定一段文本并移除其中若干词或实体作为问题，模型需要在被删除的位置填入正确答案；抽取型阅读理解要求从文本中提取特定信息；生成型阅读理解则要求根据文本内容生成相应的回答或摘要。

从**技术演进路径**，机器阅读技术可分为传统特征工程方法、深层语义图匹配方法和深度神经网络方法(27)。传统特征工程方法通过手工设计特征来表示文本；深层语义图匹配方法利用图结构来建模文本的语义关系；深度神经网络方法则通过多层神经网络自动学习文本的特征表示。

从**应用场景**，机器阅读技术在不同领域呈现出专业化发展趋势。在医疗领域，主要应用于病历管理、医疗报告解读、处方识别等场景(112)；在金融领域，应用于手写支票处理、KYC 文档

自动化、欺诈检测等场景⁽¹¹⁶⁾；在教育领域，用于手写笔记转录、研究文档扫描、学生记录管理等。

二、机器阅读的核心技术原理

2.1 光学字符识别技术

光学字符识别（OCR）技术是机器阅读的基础，其核心功能是将图像中的文字转换为机器可读的文本格式。现代 OCR 系统通常由**卷积神经网络（CNNs）**和**注意力机制**组成，其中 CNN 负责特征提取，Transformer 或 RNN 负责序列建模⁽⁷⁰⁾。

OCR 系统的工作流程包含四个紧密衔接的阶段：图像预处理、文本检测、文本识别和后处理。**图像预处理**是将原始图像进行"净化"与"规整"，为深度学习模型提供高质量输入。这一阶段包括图像增强、几何校正和二值化等关键技术。图像增强的核心是"抑制噪声、强化文本与背景的对比度"，CLAHE 的局部自适应特性是处理光照不均文档的理想选择。几何校正技术用于修正拍照或扫描时产生的透视畸变和页面弯曲，现代深度学习方案通过训练 CNN 直接预测像素级位移场来实现端到端的畸变校正。

文本检测阶段需要回答"文字在哪里？"这一核心问题。在自然场景中，文本可能以任意大小、字体、颜色、方向甚至弯曲形态出现，与复杂背景融合，检测难度极大。CTPN

（Connectionist Text Proposal Network）是早期处理场景文本的开创性工作，其核心思想是"将文本视为字符序列"，通过双向 LSTM 连接水平方向连续的提议框，利用上下文信息判断是否构成完整文本行。EAST 模型在效率上实现巨大飞跃，采用单阶段全卷积网络架构，摒弃传统"提议-验证-合并"的多阶段流程，直接在特征图上进行像素级预测。

文本识别阶段将输入的文本图像块转录为对应的字符序列。CRNN（Convolutional Recurrent Neural Network）是文本识别领域的里程碑，巧妙结合 CNN 与 RNN 的优势实现端到端识别。CRNN 的架构包含三个部分：卷积层从文本行图像中提取鲁棒的视觉特征，输出特征序列；循环层采用双向 LSTM 处理 CNN 输出的特征序列，捕捉序列的长距离上下文依赖；转录层使用 CTC（Connectionist Temporal Classification）损失函数，解决输入特征与输出标签无需预先对齐的核心难题。

后处理是 OCR 流水线的"最后一公里"，利用语言学知识校正原始识别结果。早期规则方法依赖拼写检查词典或 N-gram 统计语言模型，仅能修正简单拼写错误。现代大语言模型方法将 OCR 纠错任务重新定义为"机器翻译任务"，通过设计提示引导 LLM 利用世界知识与语言知识修正拼写错误、补全缺失标点、修复语法问题。

2.2 自然语言处理技术

自然语言处理（NLP）技术是机器阅读实现语义理解的关键，涵盖了从词法分析到语义理解的完整技术链条。

在**词法和句法分析**层面，NLP 技术包括分词、词性标注、命名实体识别等基础任务。中文等无天然分隔符的语言需要先进行分词处理，如将 "机器学习" 拆分为 "机器" 和 "学习"，然后标注每个词的词性（名词、动词、形容词等），为理解语义打下基础。命名实体识别则负责识别文本中的人名、地名、组织机构名、时间、日期等具有特定意义的实体。

在**语义理解**层面，现代 NLP 技术主要依托预训练语言模型实现。BERT (Bidirectional Encoder Representations from Transformers) 架构由三个主要模块组成：最底层的 Embedding 模块、中间层的 Transformer 模块和最上层的预微调模块(74)。BERT 的关键创新在于其双向训练方法，允许模型在处理每个单词时考虑来自两个方向的上下文。BERT 采用两个主要预训练目标：掩码语言建模 (MLM) 和下一句预测 (NSP)。

GPT (Generative Pre-trained Transformer) 则采用了不同的架构设计。GPT 使用 Transformer 解码器栈，但有一个关键修改：自注意力机制被掩码以防止模型看到未来的 token(76)。这种因果掩码确保在预测单词时，模型只能访问序列中的前序单词。GPT 模型通过预测序列中的下一个单词这一看似简单的任务，发展出了对语法、句法、事实知识、推理能力、文化和语境理解以及风格和语调适应的复杂理解。

注意力机制是现代 NLP 的核心技术之一。在机器阅读理解中，注意力机制主要有三类：动态注意力机制，其权重基于给定的 query、context 及其先前的 attention 进行动态更新；双向注意力机制，计算 query-to-context (Q2C) 和 context-to-query (C2Q) 两个方向的 attention 信息；多头注意力机制，通过并行运行多个注意力操作来扩展自注意力概念，每个操作关注单词间关系的不同方面(84)。

2.3 深度学习模型架构

深度学习模型架构是机器阅读技术实现突破的核心驱动力，从 CNN 到 Transformer 的演进反映了技术范式的重大转变。

CNN 架构在图像特征提取方面发挥着重要作用。CNN 通过卷积层对输入图像应用滤波器，捕捉空间层次结构和模式，每个滤波器在图像上滑动执行点积运算产生特征图。在机器阅读中，CNN 负责从文本图像中提取视觉特征，为后续的序列建模提供基础。

Transformer 架构的出现带来了革命性变化。Transformer 架构由编码器和解码器两个主要组件组成，两者都严重依赖自注意力机制，允许模型在处理每个单词时权衡序列中不同单词的重要性。自注意力机制通过计算序列中每个单词相对于其他单词的注意力权重，创建了不依赖于顺序处理的上下文综合理解。位置编码是另一个关键创新，因为 Transformer 缺乏 RNN 的固有顺序处理，架构使用正弦位置编码将单词位置信息注入模型。

在机器阅读理解领域，典型的神经机器阅读理解系统包含四个关键模块：嵌入、特征提取、上下文 - 问题交互和答案预测。**嵌入模块**将文本转换为向量表示；**特征提取模块**使用 CNN 或 Transformer 提取文本特征；**上下文 - 问题交互模块**通过注意力机制实现问题与文本的交互；**答案预测模块**基于交互结果预测最终答案。

三、主流算法与模型架构分析

3.1 文档理解模型

文档理解模型代表了机器阅读技术在处理复杂文档结构方面的最新进展，其中 **LayoutLM 系列模型**是最具代表性的技术突破。

LayoutLM 模型专门针对文档图像理解而设计，能够通过扫描文档图像对文本和布局信息进行联合建模⁽¹⁰²⁾。LayoutLM 利用来自文档布局的丰富视觉信息，并将其与输入文本对齐，极大改善了文档的语义表示。该模型使用 BERT 作为主干网络，并添加两个新的输入向量：二维位置向量和图像向量。二维位置向量用于对文档图像中的相对空间位置进行建模，文本的边界框表示为 (x_0, y_0, x_1, y_1) ；图像向量则利用 OCR 结果中每个文本的边界框，将图像分割为多个与文本一一对应的图像块，使用 Faster R-CNN 生成这些图像块的图像区域特征。

LayoutLMv2 相比 LayoutLM 有了显著改进。首先，LayoutLM 是在微调阶段与图像向量相结合，而 LayoutLMv2 在预训练阶段就将图像向量相结合，这样可以利用 Transformer 学习文本和视觉信息的交互信息。其次，LayoutLMv2 在预训练阶段不仅使用了 Masked visual-Language Model，还使用了文本图像对齐（text-image alignment）和文本图像匹配（text-image matching）策略。此外，LayoutLMv2 使用了空间感知自注意力机制（spatial-aware self-attention mechanism），能够更有效地建模文档布局中的局部不变形。

LayoutXLM 是多语言通用文档理解预训练模型，使用 SOTA 多语言预训练 InfoXLM 模型初始化。与 LayoutLMv2 把每个单词作为一个 token 不同，对于多语言情况这种策略不适用，LayoutXLM 获取到每个字符的 bbox，然后使用 SentencePiece 分词，再通过合并字符的 bbox 得到分词后的 bbox。该模型在 53 种语言的数据上进行预训练，包括 2200 万多语言文档数据集和 800 万英文扫描文档数据集。

DocTR 模型采用级联 Transformer 架构，包含四个主要组件：CNN 主干、轻量级特征细化模块、级联视觉编码器和文本解码器。首先，CNN 主干（ResNet 50）从输入图像中提取视觉特征。这种架构有效结合了卷积神经网络的局部特征提取能力与 Vision Transformer 的全局上下文建模能力，特别适合于文本丰富图像的特征提取⁽⁷²⁾。

3.2 图像字幕生成模型

图像字幕生成模型代表了机器阅读在多模态理解方面的重要进展，实现了从图像到自然语言描述的转换。

Show and Tell 模型被普遍看作是图像描述进入深度学习时代的经典之作。该模型利用机器翻译的思想，先将图像提取出抽象的特征，再将特征作为输入来生成自然语言。整个模型由 CNN 模型和 LSTM 模型两部分组成：第一步将图像输入到 CNN 模型中得到图像的特征，再输入到 LSTM 模型中得到相应的描述。

Show, Attend and Tell 模型针对 Show and Tell 的局限性进行了改进。Show and Tell 仅仅将图像抽象成单一的特征，而没有考虑到图像的空间特性。基于人眼观察事物时更多针对图像中某一区域来理解的生物视觉特性，该模型提出了 **attention** 机制，将图像生成描述的每一个单词都对应到图像的某一个区域。

该模型使用两种 **attention** 机制：**soft-attention** 和 **hard-attention**。**soft-attention** 在生成每个单词时，先使用深度网络计算各个区域对应的权重，再将权重乘上对应区域的特征，将结果输入到 LSTM 来得到这一时刻的单词。由于各个区域的权重不相同，因此叫做 **soft-attention**。而 **hard-attention** 是将每个单词对应到固定尺寸的区域，这种 **attention** 机制更加符合生物视觉的机制，因此得到的效果也比 **show and tell** 更好，但缺点是每个单词都强制对应一个区域。

3.3 多模态阅读模型

多模态阅读模型代表了机器阅读技术的前沿发展方向，能够同时处理文本、图像、音频等多种模态的信息。

CLIP (Contrastive Language-Image Pre-training) 模型是一个多模态视觉和语言模型，其设计动机是克服计算机视觉模型训练时对象类别的固定数量限制。**CLIP** 直接从原始文本中学习图像，通过在 4 亿（图像，文本）对上联合训练。这种规模的预训练使得下游任务的零样本迁移成为可能。**CLIP** 使用图像编码器和文本编码器来获取视觉特征和文本特征，两个特征都被投影到具有相同维度的潜在空间中，它们的点积给出相似性分数[\(109\)](#)。

CLIP 的架构设计具有创新性。视觉编码器可以是 ResNet 或 Vision Transformer，文本编码器是一个 Transformer 模型。模型通过对比学习的方式训练，使得匹配的图像 - 文本对在潜在空间中的距离更近，而不匹配的对距离更远。这种训练方式赋予了 **CLIP** 强大的零样本学习能力，可以在没有见过的类别上进行图像分类。

DALL-E 模型在多模态理解和生成方面展现了强大能力。**DALL-E** 是一个基于 Transformer 的模型，能够根据文本描述生成图像。该模型通过在大规模图像 - 文本对上进行训练，学习了视觉概念和语言描述之间的复杂映射关系。**DALL-E** 的创新之处在于其能够理解文本中的抽象概念，并生成相应的图像，例如根据 "一只穿着宇航服的猫" 这样的描述生成对应的图像。

多模态模型的发展还体现在**统一架构**的探索上。**T5 (Text-to-Text Transfer Transformer)** 提出了 "大一统思想"，将所有 NLP 任务统一建模为 "文本到文本" 的形式，从而实现模型结构、训练流程、任务适配的全面统一，极大提升了模型的通用性、可迁移性和工程实用性[\(100\)](#)。**T5** 模型是编码器 - 解码器 Transformer，其中编码器处理输入文本，解码器生成输出文本，在多任务混合的无监督和监督任务上进行预训练，每个任务都被转换为文本到文本格式。

四、应用场景与技术集成

4.1 行业应用案例

机器阅读技术在各个行业的广泛应用体现了其巨大的实用价值和商业潜力。

在**医疗领域**，机器阅读技术正在革新传统的医疗信息处理方式。在病历管理方面，医院的病历资料通常包含大量的文字信息，使用 OCR 技术可以将纸质病历快速转换为电子文本，方便医生查阅、检索和分析患者的病史、诊断结果等信息，提高医疗服务的效率和质量。在医疗报告解读方面，对于医学检验报告、影像诊断报告等，OCR 技术能够识别其中的文字内容，结合自然语言处理技术，帮助医生更快速地理解报告信息，做出准确的诊断[\(112\)](#)。在处方识别方面，通过扫描医疗处方，OCR 可以提取打印信息并导入药房的软件和应用程序中，与现有数据库或库存系统进行验证，甚至用于指导拣选机器人[\(115\)](#)。

在**金融领域**，机器阅读技术在自动化处理和风险控制方面发挥着关键作用。在支票处理方面，银行使用 OCR 扫描和验证手写支票，生成式 AI 帮助预测欺诈行为或错误。在 KYC（了解你的客户）文档自动化方面，OCR 从身份证件中提取客户详细信息，生成式 AI 验证数据并标记不一致之处。在欺诈检测方面，基于 OCR 数据训练的 AI 模型能够比传统方法更快地检测交易模式中的异常或文档伪造迹象[\(116\)](#)。在发票处理方面，OCR 能够自动提取发票号码、日期、供应商详细信息和行项目详细信息等信息，实现应付账款流程的简化[\(118\)](#)。

在**教育领域**，机器阅读技术为个性化学习和智能教育提供了技术支撑。在手写笔记转录方面，学生和教育工作者可以将手写笔记数字化，并使用 AI 进行总结或翻译。在研究文档扫描方面，OCR 从研究论文中提取数据，而生成式模型帮助生成摘要或创建引用。在学生记录管理方面，自动化帮助机构维护准确、可访问和安全的记录。

在**法律领域**，机器阅读技术在文档管理和法律研究方面展现出巨大潜力。在旧案例文件数字化方面，OCR 帮助扫描和数字化历史文档，使其可搜索以便更快地进行案例研究。在合同分析方面，AI 工具可以审查合同、提取条款、总结条款并建议编辑，减少数小时的法律劳动。在法律研究自动化方面，结合 OCR 与 AI 使律师能够快速检索相关先例，提高诉讼准备度。

4.2 技术集成方案

机器阅读技术的广泛应用离不开与云计算、边缘计算等技术的深度集成，形成了多样化的技术部署方案。

在**云计算集成**方面，主流云服务提供商都推出了完整的机器阅读解决方案。****AWS 智能文档处理 (IDP) **** 利用光学字符识别 (OCR)、计算机视觉、自然语言处理 (NLP) 和机器学习来自动化处理数据。AWS 还提供了针对 SAP 系统的集成方案，展示了如何使用 AWS AI/ML 和生成式 AI 服务自动化非结构化文档处理以及后续的审计和分析，包括使用 Amazon Textract 提取、分类和处理文档，使用 AWS SDK for SAP ABAP 进行 SAP Clean Core 扩展，以及使用 Amazon Bedrock 构建智能审计聊天机器人助手来构建审计摘要并提高业务用户生产力[\(121\)](#)。

Azure AI 文档智能服务是微软 Azure AI 服务的一部分，是一个强大的基于云的服务，使用机器学习自动化从文档中提取数据。该服务支持在边缘、本地和云端识别表单，Azure AI 服务中提供容器支持。Azure AI 内容理解 REST API 可以从文档、图像、音频和视频文件中的多模态内容获取结构化数据[\(125\)](#)。

在**边缘计算集成**方面，机器阅读技术正在向实时、本地化处理方向发展。边缘计算环境下的 OCR 应用能够在设备端直接进行文本识别，减少了数据传输延迟和隐私风险。例如，在移动设备上的实时文本识别应用，可以在不将图像上传到云端的情况下直接识别图像中的文字，提高了应用的响应速度和用户体验。

在**API 服务集成**方面，机器阅读技术已经形成了成熟的 API 生态系统。ABBYY 提供了高性能的 OCR 和文档处理 API，能够轻松地将非结构化文档转换为干净、结构化的数据，该 API 专为速度、准确性与现有技术栈的无缝集成而构建⁽¹²³⁾。腾讯云提供了 DeepSeek API 接口，可以通过腾讯云提供的 SDK 进行调用，同时腾讯云还额外封装了 DeepSeek OpenAI 对话接口，兼容了 OpenAI 的接口规范，这意味着开发者可以直接使用 OpenAI 官方提供的 SDK 来调用。

4.3 产品化实践

机器阅读技术的产品化实践体现了技术从实验室到市场的成功转化。

在**企业级解决方案**方面，IBM 自动化文档处理解决方案结合了 AI 与深度学习和低代码工具，帮助企业消除手动文档处理。该解决方案能够更快速、轻松和准确地对业务文档进行分类和提取信息。ABBYY FlexiCapture 是另一个成功的企业级产品，该产品利用在 Microsoft Azure 上安全托管的 ABBYY FlexiCapture 云服务，提供 SDK 供开发者设计具有智能捕获功能的自己的应用程序，被全球 10,000 多家领先公司所信赖⁽¹²⁰⁾。

在**消费级应用**方面，机器阅读技术正在改变人们的日常工作和生活方式。Google Docs、Microsoft Word 等主流办公软件都集成了 OCR 功能，能够直接识别和编辑扫描文档中的文字。在移动应用领域，各种 OCR 扫描应用如 CamScanner、Adobe Scan 等已经成为用户处理纸质文档的重要工具。这些应用不仅能够识别文字，还能够进行实时翻译、格式转换等高级处理。

在**垂直领域产品**方面，机器阅读技术正在形成专业化的解决方案。在金融领域，出现了专门处理银行对账单、发票、收据等金融文档的 OCR 产品；在医疗领域，出现了专门处理病历、处方、检验报告等医疗文档的识别系统；在教育领域，出现了专门处理试卷、作业、笔记等教育文档的智能工具。

五、技术挑战与解决方案

5.1 图像质量相关挑战

图像质量问题是机器阅读技术面临的最基础也是最具挑战性的问题之一，直接影响着整个系统的识别准确率和用户体验。

模糊图像识别挑战是最常见的问题。在 OCR 购物小票识别等应用中，因打印质量差、拍摄角度倾斜或对焦不准导致的模糊图像是一个关键技术难题。模糊图像会降低字符边缘清晰度，影响文本检测与识别准确率。具体表现为：高斯或运动模糊使字符粘连或断裂，低对比度导致边

缘信息丢失，透视畸变造成文本行错位(127)。这些问题在实际应用中经常出现，特别是在移动设备拍摄的场景下，手抖、光线不足、拍摄角度不当等因素都会导致图像质量下降。

光照不均问题是另一个严重影响识别效果的因素。在实际扫描或拍摄过程中，由于光源位置、强度分布不均等原因，文档的不同区域可能呈现出明显的亮度差异。这种光照不均会导致同一文档中有的区域过亮而有的区域过暗，严重影响 OCR 系统的二值化处理效果。传统的全局阈值方法在处理光照不均图像时效果很差，而自适应阈值方法如 **Sauvola** 方法虽然能够动态计算每个像素的阈值，但需要合理设定窗口大小和参数值。

低分辨率图像处理是移动应用中经常遇到的挑战。这些图像通常以低分辨率捕获，并存在噪声、光照不均或透视畸变等波动，使得基于相机的 OCR 比传统 OCR 更具挑战性。低分辨率不仅会导致字符细节丢失，还会增加字符间的混淆概率，例如 "i" 和 "l"、"o" 和 "0" 等相似字符在低分辨率下很难区分。

针对这些挑战，技术解决方案正在不断演进。在**图像预处理**方面，先进的图像增强算法能够自动识别和校正图像缺陷。例如，基于深度学习的图像超分辨率技术能够将低分辨率图像放大并恢复丢失的细节；基于 **Retinex** 理论的图像增强算法能够有效改善光照不均问题，增强图像的对比度和清晰度。在**识别算法优化**方面，现代 OCR 系统采用了更鲁棒的特征提取和匹配算法，能够在图像质量较差的情况下仍保持较高的识别准确率。例如，基于深度学习的方法通过大量低质量图像数据的训练，提高了模型对各种图像缺陷的适应能力。

5.2 文档结构理解挑战

复杂文档结构的理解是机器阅读技术面临的另一个重大挑战，特别是在处理具有不规则布局、多元素混合的文档时。

表格和图表处理挑战是文档理解中的难点。许多文档包含大量的表格和图表，这些元素的结构复杂且变化多样。传统的 OCR 系统在处理表格时常常遇到单元格合并、跨行跨列、斜线分割等复杂情况，导致表格结构识别错误或文本提取顺序混乱。对于图表，虽然其中可能包含重要的文字信息如标题、标签等，但这些文字的位置、方向、大小都可能不规则，增加了识别难度。

多栏布局处理挑战在报纸、杂志等出版物中尤为常见。多栏文本的阅读顺序识别是一个复杂问题，特别是当页面中还包含图片、广告等元素时，传统的按行扫描方法无法正确判断文本的逻辑顺序。此外，不同栏之间的文本可能存在跨栏引用、脚注等复杂关系，需要更高级的语义理解能力。

不规则文档布局挑战在历史文档、手写文档等特殊场景中更为突出。许多 OCR 系统，包括 Tesseract 和 EasyOCR 等开源引擎，缺乏对历史材料的优化，在处理非典型布局几何形状或质量下降的扫描件时经常失败。这些文档可能存在页面破损、文字褪色、背景污渍等问题，同时布局可能完全不规则，没有统一的格式标准。

针对这些挑战，技术解决方案正在向**布局感知**方向发展。**Nougat** 模型针对文档布局的特殊性，对 Swin Transformer 的相对位置偏置进行了优化。当窗口大小从预训练的 12 调整为 7

时，通过双三次插值重新计算位置偏置表，确保局部注意力机制适应文档元素的分布特点(130)。现代文档理解系统采用布局感知模型、OCR 引擎和大语言模型的组合，能够检测结构元素（表格、列、标题）并理解它们之间的关系，即使在不规则或视觉复杂的布局中也能实现精确的语义提取(131)。

DocBed 等多阶段 OCR 解决方案采用了更智能的处理策略，不是将整个扫描文档输入到 OCR 引擎，而是先智能地将文档分割成具有相似语义内容的较小部分，并按顺序将这些片段传递给 OCR 引擎以保持阅读顺序(134)。这种方法能够更好地处理复杂布局，提高整体识别准确率。

5.3 语言理解相关挑战

语言理解是机器阅读技术的核心目标，也是最具挑战性的技术难题之一。

多语言处理挑战在全球化背景下日益突出。随着国际化应用的增加，机器阅读系统需要处理多种语言的混合文本。不同语言在文字系统、书写方向、语法结构等方面存在巨大差异，例如中文、日文、韩文等亚洲语言通常采用表意文字，而阿拉伯语、希伯来语等语言从右向左书写。此外，同一文档中可能包含多种语言的混合，如英文技术文档中夹杂着中文注释，这种情况对语言识别和处理提出了更高要求。

语义理解深度不足是当前机器阅读技术的主要瓶颈。虽然现代 OCR 系统能够准确识别文字，但在理解文本的深层含义方面仍存在很大差距。例如，对于“这件事我办不了”这句话，机器可能能够正确识别每个字，但很难理解说话人的真实意图是拒绝还是能力不足。语义理解需要结合上下文、语境、文化背景等多方面信息，这对机器阅读系统提出了极高的要求。

上下文推理能力有限是另一个重要挑战。在实际阅读中，人们能够根据上下文信息进行推理和推断，理解文本中隐含的信息。例如，在阅读“小明今天没去上学，因为他生病了”这句话时，人们能够推断出小明是因为生病而缺课。但对机器而言，这种推理过程需要复杂的知识表示和推理机制支持。

针对这些挑战，技术发展呈现出几个重要趋势。首先是**预训练语言模型**的广泛应用。BERT、GPT 等大规模预训练模型通过在海量文本数据上的学习，获得了强大的语言理解能力，能够处理多种语言并理解文本的深层含义。其次是**多模态融合技术**的发展。通过结合文本、图像、音频等多种模态的信息，机器阅读系统能够获得更丰富的上下文信息，提高理解的准确性。例如，在处理包含图表的文档时，系统可以同时分析文本和图像内容，更好地理解文档的整体含义。

六、发展趋势与前景展望

6.1 技术发展趋势

机器阅读技术正处于快速发展期，呈现出多个重要的技术发展趋势。

多模态融合是最显著的发展趋势之一。随着计算机视觉、自然语言处理、语音识别等技术的不断成熟，机器阅读正在从单一文本处理向多模态信息融合方向发展。未来的机器阅读系统将能够同时理解文本、图像、音频、视频等多种信息模态，并在它们之间建立语义关联。例如，在处理一份包含图表的技术报告时，系统不仅能够识别文本内容，还能够理解图表的含义，并将两者结合起来形成完整的理解。最新的研究进展包括 CLIP、DALL-E、Stable Diffusion、Flamingo 和 GPT-4 等，这些模型在零样本分类、图像生成和上下文学习方面提供了前所未有的泛化能力(108)。

实时处理能力提升是另一个重要趋势。随着边缘计算、5G 通信等技术的发展，机器阅读正在向实时、在线处理方向发展。未来的移动设备将具备强大的本地 OCR 和文本理解能力，能够在拍摄的同时完成文字识别和语义理解，为用户提供即时的信息服务。例如，在参观博物馆时，用户可以通过手机摄像头实时识别展品说明，并获得详细的语音讲解。

个性化阅读体验正在成为新的发展方向。基于用户的阅读习惯、知识背景、兴趣偏好等个性化信息，机器阅读系统将能够提供定制化的阅读服务。例如，为不同教育水平的用户提供相应难度的文本摘要，为专业人士提供深度的技术分析，为普通读者提供通俗易懂的解释说明。这种个性化服务不仅能够提高阅读效率，还能够增强用户的阅读体验。

智能化推理能力增强是技术发展的核心目标。未来的机器阅读系统将具备更强的逻辑推理和知识推理能力，能够理解文本中的因果关系、逻辑结构和隐含信息。例如，在阅读一份科学论文时，系统不仅能够提取实验方法和结果，还能够理解实验设计的逻辑、结果的科学意义以及与前人工作的关系。这种深度理解能力将使机器阅读系统在科学研究、技术创新等领域发挥更大作用。

6.2 应用前景分析

机器阅读技术的应用前景广阔，将在多个领域带来革命性变化。

在**智能制造领域**，机器阅读技术将成为工业 4.0 的重要支撑。通过识别和理解生产线上的各种文档、标签、指示牌等信息，结合计算机视觉技术对生产过程进行实时监控和分析，智能系统能够实现生产过程的自动化控制和优化。例如，在汽车制造过程中，系统可以自动识别零部件上的二维码标签，追踪零部件的来源和质量信息，确保产品质量的可追溯性。

在**智慧城市建设**中，机器阅读技术将发挥重要作用。通过对城市环境中的各种文本信息进行自动识别和分析，如交通标志、广告牌、公共设施标识等，智能城市系统能够更好地理解和响应用户需求。例如，智能交通系统可以实时识别交通标志和路况信息，为自动驾驶车辆提供导航服务；智能环境监测系统可以识别污染源相关的文字信息，及时发现和处理环境问题。

在**数字人文研究**领域，机器阅读技术正在开启全新的研究范式。通过对海量历史文献、古籍、档案等进行自动化处理和分析，研究者能够发现人工阅读难以察觉的模式和规律。例如，通过对大量古代文献的文本挖掘，研究者可以分析语言的演变规律、文化的传播路径、社会的发展趋势等。这种基于大数据的研究方法正在改变传统的人文研究模式。

在**医疗健康领域**，机器阅读技术将推动精准医疗的发展。通过自动分析患者的病历、检查报告、医学影像等多源信息，结合医学知识图谱和临床决策支持系统，医生能够获得更全面、准确的诊断信息。特别是在处理大量医学文献时，机器阅读系统能够快速识别和提取相关的研究成果、治疗方案、药物信息等，为临床决策提供有力支持。

6.3 面临的机遇与挑战

机器阅读技术的发展既面临着前所未有的机遇，也面临着诸多挑战。

从**机遇**角度看，首先是**数据资源的丰富**。随着数字化进程的加速，全球范围内积累了海量的文本和图像数据，为机器阅读技术的训练和优化提供了丰富的资源。据统计，互联网上每天产生的数据量已经达到了 ZB 级别，其中包含了大量的文档、网页、社交媒体内容等文本信息。这些数据为机器学习模型提供了丰富的训练样本，有助于提升模型的泛化能力和准确性。

其次是**计算能力的提升**。云计算、GPU 计算、量子计算等技术的发展为机器阅读技术提供了强大的计算支撑。特别是 GPU 集群的大规模应用，使得训练超大规模神经网络模型成为可能。例如，GPT-4 等模型的训练需要使用数千个 GPU 进行数月的计算，这种计算需求在十年前是不可想象的。计算能力的提升不仅加快了模型训练速度，还使得更复杂、更强大的模型架构成为可能。

再次是**应用需求的增长**。在数字化转型的大背景下，各行各业对文档自动化处理的需求急剧增长。企业希望通过自动化技术提高工作效率、降低成本；政府部门希望通过智能化手段提升公共服务质量；个人用户希望获得更便捷、智能的信息服务。这种广泛的应用需求为机器阅读技术提供了巨大的市场空间。

从**挑战**角度看，首先是**技术瓶颈**。尽管机器阅读技术在许多方面取得了突破，但在一些关键问题上仍存在瓶颈。例如，在处理低质量图像、复杂文档布局、多语言混合文本等方面，现有技术的准确率和鲁棒性还需要进一步提升。特别是在语义理解的深度方面，机器阅读系统与人类的理解能力仍有较大差距。

其次是**数据质量问题**。训练数据的质量直接影响模型的性能，但在实际应用中，数据质量问题普遍存在。例如，标注数据可能存在错误或不一致；训练数据与实际应用场景存在分布差异；隐私数据的使用受到法规限制等。这些问题都对技术的发展和应用造成了挑战。

再次是**伦理和法律问题**。随着机器阅读技术的广泛应用，隐私保护、数据安全、算法公平性等问题日益突出。例如，在处理包含个人隐私信息的文档时，如何确保数据不被泄露；在使用机器学习模型进行决策时，如何确保算法的公平性和可解释性；在自动化处理法律文档时，如何确保处理结果的合法性和准确性。这些问题需要技术开发者、政策制定者和社会各界共同关注和解决。

结论

机器阅读技术作为人工智能领域的重要分支，已经从早期的字符识别发展为涵盖图像理解、语义分析、知识推理等多个层面的综合性技术体系。本文通过系统分析机器阅读的技术原理、发展历程、主流模型和应用实践，得出以下主要结论：

首先，**技术发展呈现出明显的阶段性特征**。从 1929 年蒂策尔提出光电字符识别概念，到 1974 年库兹韦尔发明全字体 OCR 系统，再到 21 世纪深度学习技术带来的革命性突破，机器阅读技术经历了从机械模仿到智能理解的演进过程。特别是近年来，以 BERT、GPT、LayoutLM 为代表的预训练模型和以 CLIP、DALL-E 为代表的多模态模型，标志着机器阅读技术进入了智能化、多模态化的新阶段。

其次，**技术架构日趋成熟和多样化**。现代机器阅读系统形成了以 OCR 技术为基础、以深度学习模型为核心、以多模态融合为特色的技术架构体系。在图像识别层面，CNN、Transformer 等架构实现了高精度的文字检测和识别；在语义理解层面，预训练语言模型展现出强大的语言理解能力；在多模态处理层面，统一架构和跨模态注意力机制实现了文本、图像、音频等信息的综合理解。

再次，**应用场景不断拓展和深化**。机器阅读技术已经广泛应用于医疗、金融、教育、法律等多个行业，在提高工作效率、改善用户体验、推动行业数字化转型等方面发挥着重要作用。特别是在云计算、边缘计算、API 服务等技术的支撑下，机器阅读技术正在形成完整的产业生态，为各行业提供了丰富的技术解决方案。

最后，**技术挑战依然存在但前景光明**。尽管机器阅读技术在准确性和效率方面取得了显著进展，但在低质量图像识别、复杂文档理解、多语言处理、深度语义理解等方面仍面临挑战。未来的发展将聚焦于多模态融合、实时处理、个性化服务、智能化推理等方向，随着技术的不断进步和应用需求的持续增长，机器阅读技术有望在更广泛的领域实现突破性应用。

展望未来，机器阅读技术将在人工智能技术体系中占据越来越重要的地位。随着大语言模型、多模态技术、边缘计算等技术的不断成熟，机器阅读将从“认字”走向“理解”，从“阅读”走向“思考”，为人类社会的智能化发展做出更大贡献。同时，我们也需要关注技术发展带来的伦理、法律、社会等问题，确保技术的健康、可持续发展。

参考资料

[1] 机器阅读理解技术拓展-洞察研究 - 豆丁网

https://www.docin.com/touch_new/preview_new.do?id=4786840547

[2] The History of OCR <https://www.veryfi.com/ocr-api-platform/history-of-ocr/>

[3] NLP 作业:机器阅读理解(MRC)综述_机器阅读理解模型-CSDN 博客

<https://blog.csdn.net/zuzhiang/article/details/103114373>

[4] 机器阅读理解系统行业深度调研及发展战略咨询报告 - 豆丁网 <https://www.docin.com/p-4857614805.html>

[5] 浅述机器阅读理解_基于知识的机器阅读理解任务-CSDN 博客

<https://blog.csdn.net/sannian3/article/details/102522687>

- [6] 机器阅读理解挑战-洞察分析-金锄头文库 <https://m.jinchutou.com/shtml/view-596048263.html>
- [7] NLP 机器阅读理解-CSDN 博客 <https://blog.csdn.net/zag666/article/details/128233296>
- [8] 机器阅读理解与文本摘要-深度研究.docx - 人人文库
<https://m.renrendoc.com/paper/387970776.html>
- [9] 机器阅读理解挑战-全面剖析.docx - 金锄头文库
<https://m.jinchutou.com/shtml/f6d90984f15aa360a35ac121ddf54357.html>
- [10] 机器如何在阅读理解中拿高分?达摩院总结 MRC 演进之路-阿里云开发者社区
<https://developer.aliyun.com/article/748261>
- [11] 机器阅读理解-深度研究.docx - 人人文库 <https://m.renrendoc.com/paper/387616620.html>
- [12] 神经机器阅读理解(MRC)技术全景:从任务分类到 BERT 革命-CSDN 博客
https://blog.csdn.net/m0_63679833/article/details/149490343
- [13] 机器阅读理解的进展与挑战-全面剖析.docx - 人人文库
<https://www.renrendoc.com/paper/403580914.html>
- [14] OCR 技术:从理论探索到智能识别的演进历程_ocr 最早提出来的契机是什么?-CSDN 博客
https://blog.csdn.net/2401_83623586/article/details/149321648
- [15] 从像素到篇章:深入剖析光学字符识别(OCR)的技术原理_ocr 技术-CSDN 博客
https://blog.csdn.net/m0_63309778/article/details/150947985
- [16] 细数 OCR 文字识别技术半世纪发展史-楚识科技 - CHOOSE
<https://www.whchoose.com/sys-nd/397.html>
- [17] OCR 文字识别领域先驱:保罗·汉德尔 - 腾讯云开发者社区-腾讯云
<https://cloud.tencent.com.cn/developer/news/2421514>
- [18] OCR 技术的前世今生:从概念到广泛应用_搜狐网
https://www.sohu.com/a/914719248_122044713
- [19] OCR 文字识别技术的发展与市场应用历程研究 - CHOOSE
<https://www.whchoose.com/sys-nd/381.html>
- [20] 揭开神秘面纱:OCR 技术如何重塑信息处理新格局揭开神秘面纱:OCR 技术如何重塑信息处理新格局 在这个日新月异的信息时代, - 掘金
<https://juejin.cn/post/7328603400303001611>
- [21] 文本数据的机器学习自动分类方法-CSDN 博客
https://blog.csdn.net/qq_34941023/article/details/78177810
- [22] Teaching Machines to Read: Advances in Text Classification Techniques
<https://encord.com/blog/text-classification/>
- [23] Taxonomy Classification using Machine Learning Based Model
<https://engrxiv.org/preprint/view/3967/version/5525>

- [24] 机器阅读理解系统行业深度调研及发展战略咨询报告 - 豆丁网 <https://www.docin.com/p-4857614805.html>
- [25] 机器阅读理解技术拓展-洞察研究.docx - 金锄头文库 <https://m.jinchutou.com/shtml/view-595589449.html>
- [26] 机器阅读理解-第 2 篇-全面剖析.docx - 人人文库 <https://m.renrendoc.com/paper/403901085.html>
- [27] 自然语言处理重点 第 11 章 机器阅读理解 复习_机器阅读理解源码复现-CSDN 博客 https://blog.csdn.net/weixin_44151034/article/details/128524616
- [28] 论文阅读 Neural Machine Reading Comprehension: Methods and Trends_kbmrc 相关论文-CSDN 博客 https://blog.csdn.net/weixin_44549377/article/details/103449382
- [29] 机器阅读理解的使用机器学习方法_mob64ca13f6bbea 的技术博客_51CTO 博客 https://blog.51cto.com/u_16213577/11924910
- [30] 机器阅读理解-深度研究.docx - 人人文库 <https://m.renrendoc.com/paper/387616620.html>
- [31] 文档图像分类、信息提取、信息结构化之 LayoutLM、LayoutLMv2、LayoutXLM —— 论文阅读笔记_layoutlm 提取字段-CSDN 博客 https://blog.csdn.net/m0_38007695/article/details/115803829
- [32] Document classification using Machine Learning <https://www.abbyy.com/ocr-sdk/features/classification/>
- [33] A Survey on Machine Reading Comprehension: Tasks, Evaluation Metrics and Benchmark Datasets <https://arxiv.org/pdf/2006.11880>
- [34] A Survey on Machine Reading Comprehension Systems <https://www-cambridge-org.accedys.udc.es/core/services/aop-cambridge-core/content/view/23FBCE30CB4325538E7DD08A7924315F/S1351324921000395a.pdf/a-survey-on-machine-reading-comprehension-systems.pdf>
- [35] Optical Character Recognition (OCR) <https://www.semanticscholar.org/paper/Optical-Character-Recognition-%28OCR%29-Mariner/611d0cf307894c6ca73ffa849946ba8960f364ab>
- [36] HDLTex: Hierarchical Deep Learning for Text Classification <https://arxiv.org/pdf/1709.08267>
- [37] A Survey of Machine Reading Comprehension Methods https://www.colips.org/conferences/ialp2022/proceedings/papers/IALP2022_P032.pdf
- [38] CHARACTER RECOGNITION - A REVIEW <https://dacemirror.sci-hub.ru/journal-article/93a19181310f11509e3ecfbc4d47e60b/govindan1990.pdf#navpanes=0&view=FitH>
- [39] Text Classification Algorithms: A Survey <https://www.sanjanamendu.com/files/pubs/kowsari2019text.pdf>
- [40] a survey on optical character recognition system <https://arxiv.org/pdf/1710.05703>

- [41] GVdoc: Graph-based Visual Document Classification
<https://preview.aclanthology.org/abstract-close-tag/2023.findings-acl.329.pdf>
- [42] Survey of Machine Reading Comprehension Based on Neural Network
<https://www.mendeley.com/catalogue/c2c000e6-dcb4-3132-9bc5-cd5a98b8c5d3/>
- [43] Transforming paper documents into XML format with WISDOM++
<http://www.di.uniba.it/~malerba/publications/ijdar.pdf>
- [44] A Survey on Neural Machine Reading Comprehension <https://arxiv.org/pdf/1906.03824>
- [45] A view on the past and future of character and document recognition
<https://www.mendeley.com/catalogue/6e1469be-105a-3bc4-b3b1-6e517a16b79b/>
- [46] English Machine Reading Comprehension Datasets: A Survey
<https://preview.aclanthology.org/icon-2018-ingestion/2021.emnlp-main.693.pdf>
- [47] Optical Character Recognition: An illustrated guide to the frontier
https://sites.ecse.rpi.edu/~nagy/PDF_chrono/2000_Nartker_Rice_SPIE3967-DR&R2000.pdf
- [48] Semantic classifier approach to document classification <https://arxiv.org/pdf/1701.04292>
- [49] A survey on Datasets for Machine Reading Comprehension <https://sci-hub.st/downloads/2019-08-25/bb/k2019.pdf#navpanes=0&view=FitH>
- [50] A Detailed Analysis of Optical Character Recognition Technology
<https://typeset.io/pdf/a-detailed-analysis-of-optical-character-recognition-34vci3pvybz.pdf>
- [51] Advanced document analysis and automatic classification of PDF documents
https://www.researchgate.net/publication/294583381_Advanced_document_analysis_and_a_utomatic_classification_of_PDF_documents
- [52] Machine Reading Comprehension: a Literature Review <https://arxiv.org/pdf/1907.01686>
- [53] 机器阅读理解技术简介_机器阅读理解,多项式选择任务-CSDN 博客
<https://blog.csdn.net/wlqkycg/article/details/120648796>
- [54] Teaching Machines to Read: Advances in Text Classification Techniques
<https://encord.com/blog/text-classification/>
- [55] Autocategorization <https://www.synaptica.com/autocategorization/>
- [56] Machine reading <https://www.thomsonreuters.com/en/artificial-intelligence/machine-reading.html>
- [57] 机器阅读理解技术拓展-洞察研究 - 豆丁网
https://www.docin.com/touch_new/preview_new.do?id=4786840547
- [58] 机器阅读理解系统行业深度调研及发展战略咨询报告 - 豆丁网 <https://www.docin.com/p-4857614805.html>
- [59] NLP 作业:机器阅读理解(MRC)综述_机器阅读理解模型-CSDN 博客
<https://blog.csdn.net/zuzhiang/article/details/103114373>
- [60] 浅述机器阅读理解_基于知识的机器阅读理解任务-CSDN 博客
<https://blog.csdn.net/sannian3/article/details/102522687>

- [61] 机器阅读理解与文本摘要-全面剖析.docx - 人人文库
<https://m.renrendoc.com/paper/403898952.html>
- [62] OCR 技术的前世今生:从概念到广泛应用_搜狐网
https://www.sohu.com/a/914719248_122044713
- [63] 从像素到篇章:深入剖析光学字符识别(OCR)的技术原理_ocr 技术-CSDN 博客
https://blog.csdn.net/m0_63309778/article/details/150947985
- [64] A Picture is Worth a Thousand Words – and Heaps More! The Past, Present and Future of OCR Technology <https://mathco.com/article/a-picture-is-worth-a-thousand-words-and-heaps-more-the-past-present-and-future-of-ocr-technology/>
- [65] AI OCR is Revolutionizing Receipt Scanning <https://tabscanner.com/ai-ocr-receipt-scanning-history/>
- [66] History and the future: Deep-learning-based OCR https://blog.advance.ai/blog/history-and-the-future-deep-learning-based-ocr?hs_amp=true
- [67] Optical Character Recognition: A Deep History of Transforming Text to Data
<https://www.downelink.com/optical-character-recognition-a-deep-history-of-transforming-text-to-data/>
- [68] Technologies in machine-learning-based OCR and its further directions
https://blog.advance.ai/blog/technologies-in-machine-learning-based-ocr-and-its-further-directions?hs_amp=true
- [69] From object detection to text detection and recognition: A brief evolution history of optical character recognition <https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.1547>
- [70] OCR 大模型 - CSDN 文库 <https://wenku.csdn.net/answer/1ejd77m95s>
- [71] Conv-Transformer Architecture for Unconstrained Off-Line Urdu Handwriting Recognition(pdf) <https://dll.seecs.nust.edu.pk/wp-content/uploads/2022/12/Conv-Transformer-Architecture-for-Unconstrained-off-line-Urdu-Handwriting-Recognition.pdf>
- [72] MixTex: Unambiguous Recognition Should Not Rely Solely on Real Data(pdf)
<https://arxiv.org/pdf/2406.17148v2.pdf>
- [73] Model Details https://www.modelscope.cn/models/iic/cv_LightweightEdge_ocr-recognitoin-general_damo/summary
- [74] 自然语言处理——07 BERT、ELMO、GTP 系列模型-CSDN 博客
<https://blog.csdn.net/kongbaidemumu/article/details/150775393>
- [75] A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT(pdf) <https://arxiv.org/pdf/2302.09419.pdf>
- [76] Transformer vs BERT vs GPT: Complete Architecture Comparison
<https://mljourney.com/transformer-vs-bert-vs-gpt-complete-architecture-comparison/>
- [77] BERT vs. GPT - Which One Is the Best to Use? | UPDF https://updf.com/chatgpt/bert-vs-gpt/?srltid=AfmBOophaptrmz8AhdRZNK-Y4Ya4DJQL2EE8X0zLAWtSwb26Yc_cBH7K
- [78] MRC 机器阅读理解 - CSDN 文库 <https://wenku.csdn.net/answer/7rzo9gkzmf>

- [79] Machine reading <https://www.thomsonreuters.com/en/artificial-intelligence/machine-reading.html>
- [80] 深入理解 CTC Loss:原理、计算与代码实现-CSDN 博客 https://blog.csdn.net/CITY_OF_MO_GY/article/details/149887060
- [81] MindSpore/docs https://gitee.com/mindspore/docs/blob/ac413710bd9da35821a2d87894db5e4dc889cfaf/tutorials/application/source_zh_cn/cv/cnnctc.ipynb
- [82] Connectionist Temporal Classification <https://apache.googlesource.com/mxnet/+/refs/heads/docs/example/ctc/README.md>
- [83] CNN and CTC for Recognizing Text from Images <https://www.mindspore.cn/tutorials/application/en/r2.2/cv/cnnctc.html>
- [84] 机器阅读理解系列文章-BiDAF(Bi-Directional Attention Flow for Machine Comprehension)-CSDN 博客 <https://blog.csdn.net/ljp1919/article/details/89101850>
- [85] Machine Reading Comprehension Based On Multi-headed attention Model(pdf) <https://sci-hub.st/downloads/2019-06-19/6a/xu2018.pdf#navpanes=0&view=FitH>
- [86] BI-DIRECTIONAL ATTENTION FLOW FOR MACHINE COMPREHENSION(pdf) <https://arxiv.org/pdf/1611.01603v5>
- [87] Single-paragraph Extractive Machine Reading Comprehension Model Based on Multi-layer Aligned Attention <https://www.computer.org/csdl/proceedings-article/icaml/2022/626500a107/1Lkfx8YZvNe>
- [88] Research on machine reading comprehension task based on BiDAF with self-attention <https://jns.nju.edu.cn/EN/10.13232/j.cnki.jnju.2019.01.013>
- [89] CRNN 详解_crnn 算法-CSDN 博客 <https://blog.csdn.net/bestriVERN/article/details/91050960>
- [90] CRNN Model Research Articles <https://discovery.researcher.life/topic/centers-alpha/22027771>
- [91] 深度学习 | BERT 详解-CSDN 博客 https://blog.csdn.net/weixin_40920183/article/details/117393610
- [92] A Pairwise Probe for Understanding BERT Fine-Tuning on Machine Reading Comprehension(pdf) <https://arxiv.org/pdf/2006.01346>
- [93] Automated Scoring for Reading Comprehension via In-context BERT Tuning(pdf) <https://arxiv.org/pdf/2205.09864v2.pdf>
- [94] BERT Based Multilingual Machine Comprehension in English and Hindi(pdf) <https://arxiv.org/pdf/2006.01432v1>
- [95] 开源论文神器 DocsGPT!多类型文档查询,快速搭建属于你的问答式的文档检索服务-CSDN 博客 https://blog.csdn.net/2401_83063795/article/details/136845094
- [96] Optimizing File Uploads in ChatGPT Enterprise <https://help.openai.com/en/articles/10029836>

- [97] An extensive benchmark study on biomedical text generation and mining with ChatGPT(pdf) <https://academic.oup.com/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btad557/51441415/btad557.pdf>
- [98] What is a GPT? | Microsoft Azure <https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-gpt>
- [99] File inputs <https://platform.openai.com/docs/guides/pdf-files>
- [100] happy_llm 学习 06—Encoder-Decoder PLM_下载 llm 的 encoder-CSDN 博客 https://blog.csdn.net/weixin_65615457/article/details/148921298
- [101] idT5: Indonesian Version of Multilingual T5 Transformer(pdf) <https://arxiv.org/pdf/2302.00856v2.pdf>
- [102] 文档图像分类、信息提取、信息结构化之 LayoutLM、LayoutLMv2、LayoutXLM —— 论文阅读笔记_layoutlm 提取字段-CSDN 博客 https://blog.csdn.net/m0_38007695/article/details/115803829
- [103] Enhancing Visually-Rich Document Understanding via Layout Structure Modeling(pdf) <https://arxiv.org/pdf/2308.07777.pdf>
- [104] LayoutLM <https://boinc-ai.gitbook.io/transformers/api/models/multimodal-models/layoutlm>
- [105] 图像描述(image caption)历年突破性论文总结_faster rcnn image caption-CSDN 博客 <https://blog.csdn.net/kdongyi/article/details/100738678>
- [106] Image Captioning using Facial Expression and Attention(pdf) <https://arxiv.org/pdf/1908.02923>
- [107] Show and Tell - Dss Solutions <https://dsssolutions.com/2025/02/03/show-and-tell-2/>
- [108] In search of the next generation of multimodal datasets(pdf) <https://arxiv.org/pdf/2304.14108v5>
- [109] CLIP https://huggingface.co/docs/transformers/main/en/model_doc/clip
- [110] 多模态预训练模型 | 菜鸟教程 <https://m.runoob.com/nlp/multimodal-pre-trained-models.html>
- [111] Daily Papers - Hugging Face <https://huggingface.co/papers?q=Multimodal%20document%20retrieval>
- [112] 初窥强大，AI 识别技术实现图像转文字(OCR 技术)_ai ocr-CSDN 博客 https://blog.csdn.net/weixin_42108731/article/details/145482636
- [113] Healthcare Automation <https://ocrsolutions.com/market-solutions-healthcare-automation/>
- [114] Streamline Medical Claims Processing with OCR <https://ocrsolutions.com/medical-claims/>
- [115] ABBYY Solutions for Healthcare Automation <https://www.abbyy.com/solutions/healthcare/sdk-for-ocr-and-data-capture-in-healthcare/>

- [116] Top Use Cases of OCR Generative AI Across Industries <https://techbulletin.com/ocr-generative-ai/>
- [117] An AI OCR service that reads text from documents <https://www.abbyy.com/vantage/ocr-container/>
- [118] OCR for Financial Documents <https://www.klippa.com/en/ocr/financial-documents/>
- [119] Advanced Document Processing with Azure AI and OpenAI: 8+ Wk Implementation - Microsoft 市场 https://azuremarketplace.microsoft.com/zh-cn/marketplace/consulting-services/dataartnewyork1670255734487.document_processing_openai
- [120] ABBYY FlexiCapture <https://www.abbyy.com/flexicapture/>
- [121] Guidance for SAP Intelligent Document Processing and Insights using Generative AI on AWS <https://aws.amazon.com/cn/solutions/guidance/sap-intelligent-document-processing-and-insights-using-generative-ai-on-aws/>
- [122] 基于腾讯云 DeepSeek 的 AI 智能阅读助手开发实践-腾讯云开发者社区-腾讯云 <https://cloud.tencent.com/developer/article/2506217>
- [123] ABBYY OCR & Document Processing API <https://www.abbyy.com/ai-document-processing/api/>
- [124] Leveraging APIs for Seamless Machine Learning Model Integration <https://blogs.infosys.com/digital-experience/artificial-intelligence/leveraging-apis-for-seamless-machine-learning-model-integration.html>
- [125] Quickstart: Azure AI Content Understanding REST APIs - Azure AI services | Microsoft Learn <https://docs.microsoft.com/en-us/azure/ai-services/content-understanding/quickstart/use-rest-api>
- [126] MLaaS APIs for Text Classification <https://www.focalworks.com/blog/mlaas-apis-text-classification/>
- [127] OCR 购物小票识别中如何处理模糊或倾斜图像?_编程语言-CSDN 问答 <https://ask.csdn.net/questions/8832919>
- [128] 9 Biggest OCR Limitations and How to Overcome Them <https://www.docuclipper.com/blog/ocr-limitations/>
- [129] How to Improve OCR Accuracy – 4 Easy Hacks, Tips, and Tricks <https://www.klippa.com/en/blog/information/improve-ocr-accuracy/>
- [130] Nougat 模型解释:为什么它能准确理解复杂文档布局-CSDN 博客 https://blog.csdn.net/gitblog_00065/article/details/151633758
- [131] Document Layout Intelligence <https://www.businesswaretech.com/document-layout-intelligence>
- [132] Handling Document Understanding with Varying Layouts <https://forum.uipath.com/t/handling-document-understanding-with-varying-layouts/709533>
- [133] Fixing Document Layout Problems in MS Word: A Simple Guide <https://mswtutor.com/fixing-document-layout-problems/>

[134] DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts(pdf)
<https://arxiv.org/pdf/2202.01414v1>

[135] MULTIMODALQA: COMPLEX QUESTION ANSWERING OVER TEXT, TABLES AND IMAGES——全文详细翻译-CSDN 博客
https://blog.csdn.net/weixin_60302661/article/details/144316704

[136] Towards Solving Multimodal Comprehension(pdf) <https://arxiv.org/pdf/2104.10139v1>

[137] A Multi-Modal Context Reasoning Approach for Conditional Inference on Joint Textual and Visual Clues <https://aclanthology.org/2023.acl-long.601/>

（注：文档部分内容可能由 AI 生成）