智能文档处理

智能文档处理（Intelligence Document Processing）是指利用人工智能技术对网页、数字文档或扫描文档中包含的文本和丰富的排版格式等信息进行理解、分类、提取和信息归纳的过程。它是自然语言处理（NLP）和计算机视觉（CV）交叉领域的重要研究方向。

文档版面分析

文档版面分析是指对文档版面内的图像、文本、表格信息和位置关系进行自动分析、识别和理解的过程。

启发式规则方法

自顶向下

- 方法概述: 自顶向下方法将文档图片作为整体逐步划分为不同区域，通过递归方式进行切割，直至区域分割至预定义的标准。这种方法通常使用投影分析、X-Y切割等技术来进行分割。
    - **投影分析法**: 通过计算图像在垂直或水平方向上的投影来确定文本行的边界。例如，X-Y切割算法就是通过分析垂直和水平投影来确定文本块的位置。
    - **示例**: 在文档中，通过计算每一行像素的投影密度，可以确定段落之间的空白行，从而进行段落切割。
- **优点**: 在特定格式的文档中，能够更快、更高效地分析文档。
- **缺点**: 适应性差，无法处理复杂布局的文档，尤其是那些具有不规则边界或倾斜文本的文档。

自底向上

- 方法概述: 自底向上方法以像素或组件为基本元素单位，对这些基本元素进行分组和合并，以形成更大的同质区域。这种方法通常使用图像连通组件分析、Run-length Smoothing等技术。
    - **图像连通组件分析**: 识别图像中的最小粒度元素（如像素或字符），并通过连接相邻像素或字符形成连通区域。例如，KNN算法可以用于找到每个组件的邻近组件，通过分析这些组件之间的位置和角度关系，推断出当前区域的属性。
    - **Run-length Smoothing**: 将背景像素周围的背景像素数目小于阈值时，将该背景像素修改为前景像素，以扩展同质区域。
- **优点**: 通用性强，可覆盖更多不同布局类型的文档。
- **缺点**: 耗费更多的计算时间，处理速度较慢。

混合模式

- 方法概述: 混合模式结合自顶向下和自底向上的策略，以适应不同的文档布局。例如，先通过自顶向下的方法进行初步区域分割，然后再通过自底向上的方法进行细化。
    - **示例**: Okamoto等人的混合算法首先通过分隔符和空白来切割文档块，然后在每个块内部进一步将组件合并为文本行。Smith的方法首先使用自底向上的方式定位制表符，再利用这些制表符推断列布局，最后采用自顶向下的方式推断文档的结构和文本顺序。
- **优点**: 能够结合两种方法的优势，适应更多样化的文档布局，提高分割精度。
- **缺点**: 实现复杂度较高，可能需要针对特定文档类型进行调优。

基于统计机器学习的方法

文档图像切割

- 方法概述: 使用支持向量机（SVM）、随机森林等机器学习算法对文档图像进行切割，识别出不同区域。这些算法需要大量的训练数据来学习文档的布局特征。
    - **示例**: 利用SVM对每个像素点进行分类，判断其属于文本、图像还是背景区域。
    - **特征提取**: 提取图像中的特征，如像素点的坐标、颜色、邻居关系等，作为输入特征。
- **优点**: 可以处理复杂布局和多样化的文档类型，自动化程度高。
- **缺点**: 对训练数据依赖较大，训练和推理时间较长。

区域属性分类

- 方法概述: 使用机器学习算法对分割出的区域进行分类，以确定每个区域的内容类型（如文本、图像、图表等）。
    - **示例**: 利用随机森林算法，根据区域的形状、纹理和位置特征，判断区域是文本块还是图像块。
- **优点**: 能够准确分类不同类型的区域，提高文档理解的精度。
- **缺点**: 需要高质量的标注数据进行训练，对不同文档类型的适应性可能有限。

基于深度学习的方法

DeepDeSRT

- 方法概述: 将文档版面分析视为文档图像的目标检测任务，使用迁移学习微调Faster R-CNN模型来检测和识别文档中的版面元素。
    - **实现方案**: 预训练一个Faster R-CNN模型，然后在特定的文档版面数据集上进行微调。通过目标检测框架识别出文档中的文本块、图片、表格等元素。
- **优点**: 能够处理复杂的文档版面，检测精度高。
- **缺点**: 需要大量标注数据进行训练，计算资源需求较高。

CascadeTabNet

- 方法概述: 同时完成表格检测和表格结构识别任务，使用Cascade Mask Region-based CNN High-Resolution Network（HRNet）作为模型框架。
    - **实现方案**: 使用高分辨率网络（HRNet）作为特征提取器，并结合级联的Mask R-CNN进行表格检测和结构识别。通过迭代训练来逐步提高模型的检测和识别精度。
    - **数据增强**: 使用Dilation和Smudge等数据增强技术提高模型的鲁棒性。
- **优点**: 对表格检测和结构识别的效果较好，能够处理多样化的表格布局。
- **缺点**: 模型复杂度高，训练和推理时间较长。

文档信息抽取

定义: 文档信息抽取是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。

序列标注方法

朴素方法:

- **方法概述**: 将文档信息抽取形式化为序列标注任务，使用OCR（光学字符识别）技术提取文本，然后利用序列标注模型（如Bi-LSTM-CRF）进行信息抽取。
- 实现方案:
  - **OCR**: 首先使用OCR技术将文档图像中的文字转化为可编辑的文本。
  - **Bi-LSTM-CRF**: 利用双向长短期记忆网络（Bi-LSTM）捕捉文本的上下文信息，然后使用条件随机场（CRF）进行序列标注，标记文本中的实体和关系。
- **优点**: 适用于简单和结构化的文档，可以较快实现信息抽取。
- **缺点**: 对于复杂布局和非结构化文档，性能有限，难以处理文档中的视觉和空间信息。

Chargrid:

- **方法概述**: 将文档信息抽取形式化为文档页面上字符级的实例分割任务，通过编码器-解码器架构（卷积神经网络）进行信息抽取。
- 实现方案:
  - **字符网格表示**: 将文档页面表示为一个字符网格，每个字符作为网格中的一个单元格。
  - **Chargrid-Net**: 使用卷积神经网络（CNN）作为编码器，提取字符的局部特征。然后，通过解码器进行实例分割，识别字符网格中的实体。
- **优点**: 能够保留字符的空间位置信息，提高对文档结构和布局的理解能力。
- **缺点**: 对于长文档，计算复杂度较高，需要处理大规模的字符网格。

基于深度学习的方法

VisualWordGrid:

- **方法概述**: 将文档图像的视觉信息加入WordGrid（单词级），采用多模态方法进行信息抽取。
- 实现方案:
  - **单词网格表示**: 文档图像中的单词作为网格单元，每个单元包含单词的文本和视觉信息。
  - **多模态编码器**: 使用多模态编码器分别处理文本和视觉信息，并将两者结合进行信息抽取。
- **优点**: 利用视觉信息提高信息抽取的准确性，特别是对于包含图像和文本混合的文档。
- **缺点**: 需要大量计算资源处理多模态数据，模型复杂度较高。

BERTgrid:

- **方法概述**: 融入上下文信息，将BERT与视觉信息融合，进行文档信息抽取。
- 实现方案:
  - **文本嵌入**: 使用BERT模型获取文本的上下文嵌入表示。
  - **BERTgrid表示**: 将BERT嵌入与视觉信息结合，形成一个多维度的文档表示。
- **优点**: 利用BERT强大的上下文理解能力，提高信息抽取的准确性。
- **缺点**: 对计算资源要求高，处理长文档时存在输入长度限制。

ViBERTgrid:

- **方法概述**: 在卷积层将BERTgrid与视觉信息融合，进一步提高信息抽取效果。
- 实现方案:

    - **卷积融合**: 将BERTgrid与卷积层的视觉信息进行融合，通过卷积神经网络提取联合特征。
- **优点**: 提高了文本和视觉信息的融合效果，增强模型对复杂文档的理解能力。
- **缺点**: 模型复杂度高，训练和推理时间较长。

基于图神经网络的方法:

- **方法概述**: 将OCR得到的文本段构成全连接有向图，利用图卷积进行信息抽取。
- 实现方案:
    - **图结构表示**: 将文档中的文本段作为节点，节点之间的视觉和空间关系作为边，构成全连接有向图。
    - **图卷积网络（GCN）**: 使用GCN对图结构进行卷积操作，提取节点和边的特征。
- **优点**: 能够捕捉文本和视觉信息之间的复杂关系，提高信息抽取的准确性。
- **缺点**: 图结构的构建和处理复杂度较高，对计算资源要求较大。

SPADE (SPAtial DEpendency parser):

- **方法概述**: 将文档信息抽取任务形式化为空间（两个维度）依存分析问题，缓解序列标注方法无法处理半结构化文档中的复杂布局和多层次信息的问题。
- 实现方案:
    - **空间编码器**: 使用Transformer模型对文档进行空间编码，捕捉文档中不同部分之间的空间关系。
    - **依存分析**: 构建空间依存树，表示文档中的层次结构和布局信息。
- **优点**: 能够处理复杂布局和多层次信息，适用于半结构化和非结构化文档。
- **缺点**: 模型复杂度高，训练和推理时间较长。

文档视觉问答

文档视觉问答（Document Visual Question Answering, DocVQA）是指在给定文档图像数据的基础上，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。这一任务不仅需要文本识别能力，还需要综合运用视觉信息和上下文理解能力，处理文档图像的多模态特征。

基于OCR与序列模型的方法

- **方法概述**:
    - 通过OCR技术将文档图像中的文字转化为可编辑的文本。
    - 使用序列模型（如RNN、LSTM）处理提取的文本，并生成回答。
- **实现方案**:
    - **OCR处理**: 利用OCR引擎（如Tesseract）对文档图像进行文字识别，生成文本数据。
    - **序列模型**: 将OCR提取的文本输入到序列模型中，利用上下文信息生成回答。
- **优点**:
    - 适用于结构化文本和较简单的文档问答任务。
- **缺点**:
    - 对于复杂文档和需要结合视觉信息的问答任务，效果较差。

基于深度学习的端到端模型

- **方法概述**:
  - 使用深度学习模型直接处理文档图像，结合视觉和文本信息进行问答。
- **DocVQA**:
  - **数据集概述**: DocVQA数据集包含12,767张文档图像和50,000个问题，主要涉及文档中的常见数据（如日期、标题、总量、页码）。
  - **实现方案**:
    - **视觉编码器**: 使用预训练的卷积神经网络（如ResNet）提取文档图像的视觉特征。
    - **文本编码器**: 利用OCR技术提取文本，并通过文本编码器（如BERT）处理文本信息。
    - **融合层**: 将视觉特征和文本特征融合，生成联合表示。
    - **问答模块**: 使用问答模型（如Transformer）处理联合表示，生成答案。
- **优点**:
  - 能够处理复杂文档和多模态信息，提高问答的准确性。
- **缺点**:
  - 模型复杂度高，计算资源需求大。

基于多模态融合的方法

- **方法概述**:
  - 将文档图像的视觉信息和文本信息融合，通过多模态方法进行问答。
- **VisualMRC**:
  - **数据集概述**: VisualMRC数据集包含10,197张文档图像和30,562个问题，主要关注多模态结合的文档生成式问答。
  - **实现方案**:
    - **多模态编码器**: 使用视觉编码器和文本编码器分别提取文档的视觉和文本特征。
    - **融合层**: 将视觉和文本特征进行融合，生成多模态表示。
    - **问答模块**: 利用多模态表示进行问答，生成答案。
- **优点**:
  - 有效利用视觉和文本信息，提高问答效果。
- **缺点**:
  - 多模态数据处理复杂度高，模型训练和推理时间较长。

基于注意力机制的方法

- **方法概述**:
  - 利用注意力机制在文档图像和文本之间建立关联，提高问答性能。
- **GHMFC (Gated Hierarchical Multimodal Fusion and Contrastive Training)**:
  - **研究动机**: 特定指称可能存在多种含义，利用图像可以更好地辨别指称；文本和图像中存在多种联系，需要进行建建立细粒度的特征关联。
  - **实现方案**:
    - **特征提取**: 使用BERT和ResNet提取词项表征和图像表征，利用Conv1D卷积进一步精炼词项表征，获得短语表征。

        - **多模态相互注意力**: 利用注意力机制，建立文本指导的图像表征和图像指导的文本表征，进而融合多模态特征。
        - **对比训练**: 构建文本-图像、图像-文本的对比损失，使相同实体指称的图像表征和文本表征接近。
- **优点**:
    - 提高了文本和视觉信息的关联度，提高问答准确性。
- **缺点**:
    - 对计算资源要求较高，模型复杂度大。

基于图神经网络的方法

- **方法概述**:
    - 将OCR得到的文本段构成全连接有向图，利用图卷积进行信息抽取。
- **DRIN (Dynamic Relation Interactive Network)**:
    - **研究动机**: 以往工作大多利用注意力、门控等机制自动学习跨模态特征关联，缺少对文本、图像模态中目标的显式关联结构。
    - **实现方案**:
        - **关联图构建**: 在指称和所有候选实体之间，显式构建文本-文本、图像-文本、图像-图像关联图。
        - **图神经网络**: 利用图卷积神经网络，根据连边权重精炼节点表征，在图卷积过程中动态更新边权重。
        - **链指匹配**: 计算指称和实体节点表征的语义相似性，得到链指匹配分数。
- **优点**:
    - 能够捕捉文本和视觉信息之间的复杂关系，提高问答准确性。
- **缺点**:
    - 图结构的构建和处理复杂度高，对计算资源要求较大。

文档图像分类

文档图像分类（Document Image Classification）是指对文档图像进行分析和识别，以便将文档归类到预定义的类别中。该任务广泛应用于各种场景，如金融票据处理、医疗文档管理、法律文件分类等。文档图像分类的主要目标是通过分析图像的内容、布局和特征，将其分类为特定的类型或类别。

- **Tobacco**: 包含3,482张文档图像，涵盖多种文档类型，用于文档图像分类和检索任务。
- **RVL-CDIP**: 包含400,000张文档图像，覆盖16个类别（如信件、备忘录、电子邮件、文件夹、表单、手写文档、发票、广告、预算、新闻文章、演示文稿、科学出版物、问卷、简历、科学报告、规范），是一个用于文档图像分类的大规模基准数据集。

启发式规则方法

- **方法概述**:
    - 基于手工设计的规则和特征，对文档图像进行分类。这些规则通常依赖于文档的视觉特征，如边缘、纹理、形状和布局信息。
- **实现方案**:
    - **边缘检测**: 利用边缘检测算法（如Canny边缘检测）识别文档图像中的文本块、表格和图像。

    - **投影轮廓分析**: 通过计算图像在垂直和水平方向上的投影轮廓，识别文档的布局特征，如段落和标题的位置。
    - **模板匹配**: 使用预定义的模板匹配算法，将文档图像与标准模板进行比对，确定文档的类别。
- **优点**:
    - 对于特定格式和结构固定的文档，分类效果较好。
- **缺点**:
    - 通用性差，难以处理复杂和多样化的文档布局；需要大量的手工设计和调整规则。

基于统计机器学习的方法

- **方法概述**:
    - 使用机器学习算法（如SVM、随机森林等）对文档图像进行分类，这些算法通过学习文档的统计特征来进行分类。
- **实现方案**:
    - **特征提取**: 提取文档图像的低级特征（如颜色直方图、纹理特征、形状特征）和高级特征（如HOG、SIFT）。
    - **特征选择**: 使用特征选择算法（如PCA、LDA）选择最具区分性的特征。
    - **分类器训练**: 使用机器学习算法（如SVM、随机森林、KNN）在训练集上训练分类器，并在测试集上进行验证。
- **优点**:
    - 比启发式规则方法具有更好的适应性和泛化能力。
- **缺点**:
    - 依赖于特征工程，特征提取和选择过程复杂；对计算资源需求较高。

基于深度学习的方法

- **方法概述**:
    - 使用深度学习模型（如卷积神经网络，CNN）对文档图像进行分类，通过端到端的学习方式自动提取特征和进行分类。
- **实现方案**:
    - **卷积神经网络（CNN）**: 使用预训练的卷积神经网络（如VGG、ResNet、Inception）提取文档图像的高级特征，并进行分类。
    - **迁移学习**: 在预训练模型的基础上进行微调，利用大规模数据集（如ImageNet）预训练的模型在文档图像分类任务上进行迁移学习。
    - **数据增强**: 使用数据增强技术（如旋转、缩放、裁剪）增加训练数据的多样性，提高模型的泛化能力。
- **优点**:
    - 能够自动学习文档图像的高级特征，分类效果显著提高；适应性强，可处理复杂和多样化的文档布局。
- **缺点**:
    - 需要大量的训练数据和计算资源；模型训练时间较长。
- **Structural similarity for document image classification and retrieval (PRL 2014)**:
    - **方法概述**: 使用结构相似性（Structural Similarity, SSIM）进行文档图像分类。

    - 实现方案:
        - 通过计算图像之间的结构相似性得分，将文档图像分类到预定义的类别中。
        - SSIM通过对比图像的亮度、对比度和结构相似性来计算图像的相似度。
- **Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval (ICDAR 2015)**:
    - **方法概述**: 使用深度卷积神经网络（CNN）进行文档图像分类。
    - 实现方案:
        - 使用预训练的深度卷积神经网络（如AlexNet、VGG）提取文档图像的高级特征。
        - 在训练集上微调模型，并在测试集上进行验证。

文档大模型

文档大模型（Document Foundation Models）是近年来智能文档处理领域的核心研究方向。这些模型通常基于Transformer架构，通过预训练在大规模无标注文档数据上进行训练，从而在下游文档处理任务中实现出色的性能。文档大模型主要聚焦于文本、布局和视觉信息的融合，利用自监督学习任务来充分挖掘未标注数据中的知识。

基础架构

文档大模型的基础架构通常包括以下几个关键部分：

1. **文本编码器**：用于处理文档中的文本信息，通常使用Transformer架构，如BERT或其变体。
2. **布局编码器**：用于捕捉文档中的布局信息，如文本块的位置和顺序。常见的方法包括使用坐标嵌入或相对位置编码。
3. **视觉编码器**：用于处理文档中的视觉信息，如图像和图表。可以使用预训练的视觉模型（如ResNet或Swin Transformer）来提取图像特征。
4. **融合模块**：将文本、布局和视觉信息进行融合，常见的方法包括多模态注意力机制和联合嵌入空间。

预训练任务旨在通过自监督学习方式让模型从未标注数据中学习通用表示。这些任务通常包括：

1. **掩码语言模型（Masked Language Model, MLM）**：随机掩码输入文本的一部分，要求模型在给定上下文的情况下预测被掩码的文本。
2. **掩码视觉-语言模型（Masked Visual-Language Model, MVLM）**：结合视觉和文本信息的掩码任务，随机掩码文本和图像中的一部分，要求模型联合预测被掩码的内容。
3. **文本-图像对齐（Text-Image Alignment）**：要求模型识别文本和图像块之间的对应关系，以实现跨模态的对齐。
4. **区域掩码语言模型（Area-masked Language Model）**：扩展MLM任务，将掩码区域从一维扩展到二维，掩码整个文本块并进行预测。

主要工作

LayoutLM

- **主要思路**: LayoutLM结合文本和布局信息进行文档预训练，旨在弥合视觉模态和文本模态之间的差距。通过将文本、布局坐标（如二维位置）和图像信息结合起来，模型能够更好地理解文档的结构和内容。
- 预训练任务:

    - **Masked Visual-Language Model (MVLM)**: 随机掩码一些输入token，并保留其对应的二维位置嵌入，训练模型在给定上下文的情况下预测被掩码的token。这样不仅利用了上下文信息，还利用了相应的二维位置信息。
    - **Multi-label Document Classification**: 使用文档标签来监督预训练过程，使模型能够对来自不同领域的知识进行聚类，并生成更好的文档表示。

LayoutLMv2

- **主要思路**: LayoutLMv2在输入阶段加入图像信息，通过多模态融合增强模型的文档理解能力。不同于LayoutLM在输出端融入图像嵌入，LayoutLMv2在输入时就结合了图像信息。
- 预训练任务:
    - **Masked Visual-Language Model (MVLM)**: 与LayoutLM类似，进行掩码任务，结合文本和图像信息进行预测。
    - **Text-Image Alignment**: 通过细粒度（如行级别）的覆盖操作来对齐文本和图像信息，提升跨模态的对齐能力。
    - **Text-Image Matching**: 训练模型判断给定文本和图像是否匹配，从粗粒度和细粒度两个层面进行对齐。

StructuralLM

- **主要思路**: StructuralLM强调布局信息的作用，通过将Cell作为基本语义单元，确保同一Cell内的所有token共享相同的位置信息，以帮助模型理解文本的空间关系。
- 预训练任务:
    - **Masked Visual-Language Model (MVLM)**: 随机掩码token进行预测。
    - **Cell Position Classification**: 将文档划分为多个相同大小的区域，通过Cell的中心二维位置（边界框中心点的二维坐标）计算该Cell所属区域的索引（1~N），帮助模型更好地保留布局信息。

BROS (BERT Relying on Spatiality)

- **主要思路**: BROS使用相对位置编码，通过将位置信息融入注意力机制来增强模型对布局信息的理解，从而更鲁棒地处理位置变化。
- 预训练任务:
    - **Masked Visual-Language Model (MVLM)**: 随机掩码token进行预测。
    - **Area-masked Language Model**: 受到SpanBERT的启发，将掩码区域从一维扩展到二维，随机选择一个文本块，并通过扩展其边界确定一个区域，掩码区域内的所有token进行预测。

LAMPRET

- **主要思路**: LAMPRET融合了多种视觉模态信息，包括类型信息、属性信息和块序列信息，增强多模态预训练的效果。
- 预训练任务:
    - **Masked Visual-Language Model (MVLM)**: 掩码视觉和文本信息进行预测。
    - **Text-Image Matching**: 判断文本和图像是否匹配。
    - **Block Order Prediction**: 预测块的顺序，帮助模型理解文档结构。
    - **Image Fitting**: 训练模型将图像块组合还原。
    - **Masked Block Prediction**: 掩码块中的内容进行预测。

LayoutXLM

- **主要思路**: 基于LayoutLMv2进行多语言扩展，支持多达53种语言，通过使用数字生成的PDF文档进行预训练，确保准确的文本和布局信息。
- **预训练任务**: 类似LayoutLMv2，主要扩展到多语言环境，使模型能够处理不同语言的文档。

Donut

- **主要思路**: Donut是一种不依赖OCR的文档理解模型，通过视觉编码器和文本解码器的组合实现端到端的文档处理，避免OCR引入的错误。
- 预训练任务: **伪OCR**: 按照从左上到右下的顺序生成文本，使模型能够学习文档的整体结构。

DiT (Document Image Transformer)

- **主要思路**: 使用无标注文档图像进行自监督预训练，通过图像块和视觉token的组合实现文档理解。
- 预训练任务:
  - **Masked Language Model (MLM)**: 掩码语言模型任务。
  - **Masked Image Model (MIM)**: 掩码图像模型任务，通过掩码图像块进行预测。

LLaVAR

- **主要思路**: 增强多模态大语言模型对文本图像的理解，通过收集指令数据并进行端到端指令微调，提高模型对文本图像的理解能力。
- 预训练任务: **使用海量指令数据和对话数据进行预训练和微调**: 通过对话数据和指令数据进行训练，增强模型的理解能力。

UReader

- **主要思路**: 通过仅微调少量参数，实现高效的多任务文档理解，采用低分辨率编码器处理高分辨率图像。
- 预训练任务: **统一的下游任务微调**: 借鉴InstructBLIP的指导微调格式，使模型能够高效处理各种任务。

TextMonkey

- **主要思路**: 提供灵活的文档理解框架，避免传统OCR引入的错误，通过关键模块（Shifted Window Attention、Image Resampler、Token Resampler）实现高效的文档处理。
- 预训练任务: **结合多种自监督预训练任务**: 实现更精细的视觉和文本信息处理，通过Swin Transformer结构和token resampling技术来优化文档处理。

智能文档生成

智能文档生成是指利用多种模态数据（如文本、图像等）来生成完整的文档内容的技术。随着人工智能技术的发展，内容生成形式不断演变，从专业生成内容（PGC）发展为用户生成内容（UGC），然后逐步进化为人工智能辅助用户生成内容（AIUGC）和人工智能生成内容（AIGC）。这种模式的升级带来了内容生成数量的快速增长。

在信息时代，大量的多模态数据被广泛应用于社交媒体、在线新闻等各个领域。多模态文档类型丰富，包括论文、网页文档、海报、技术手册、PPT、公文等多种形式。由于人工生成多模态文档的能力难以满足大数据时代急剧增长的需求，智能文档生成成为AIUGC和AIGC中具有巨大潜力的研究领域之一。其可以：

1. **丰富文档内容**:
   - 智能文档生成能够整合多种模态的信息，生成更加丰富、多样化的文档内容，从而提升文档的信息量和吸引力。

2. **提高生成效率**:
    - 传统的文档生成过程需要人工编写和编辑，耗时耗力。而智能文档生成技术可以自动化这一过程，大幅降低人工成本和资源开销，提高文档生成的效率。
3. **推动智能化应用**:
    - 智能文档生成技术是人工智能领域的重要应用之一，其研究和应用将推动人工智能技术在文档处理、自然语言理解、计算机视觉等领域的发展，促进智能化应用在各行业的广泛应用和普及。

智能文档生成技术涉及多个领域和技术，包括但不限于自然语言处理（NLP）、计算机视觉（CV）、多模态融合等领域。主要技术包括：

1. **文档元素生成技术**:
    - 通过自然语言处理生成文档中的文本内容，通过计算机视觉生成或识别文档中的图像、图表等元素。
2. **文档布局生成技术**:
    - 根据文档的主题和内容自动生成合理的文档布局。这个过程涉及到多模态信息的融合，以确保文档布局美观且功能性强。
3. **多任务集成系统**:
    - 综合运用多种技术，实现从内容生成到布局生成，再到最终文档输出的全流程自动化。这样的系统能够根据用户的需求进行调整，提高文档生成的定制化和灵活性。

智能文档生成技术的不断发展，有望解决传统文档生成过程中效率低、成本高的问题，推动文档处理技术迈向新的高度。

文档布局生成

文档布局生成是指将元素的位置和大小以有意义的排列方式显示的过程。在常见的2D场景中，布局通常表示为元素的边界框，用5个属性来描述一个元素：(c, x, y, w, h)，其中c表示元素类别；(x, y)表示元素位置；(w, h)表示元素大小。应用场景包括自然图片布局设计、应用程序UI设计、室内布局设计、制作PPT模板、论文排版、商业海报设计、3D形状设计等。

- 图像无关的文档布局生成：图像无关的文档布局生成任务仅需要考虑生成元素之间的关系，关注布局规范化的要求，如元素的对齐和重叠度等。这类任务主要包括：
    - **Gen-T**: 根据元素类别生成布局
    - **Gen-TS**: 根据元素类别和大小生成布局
    - **Gen-TR**: 根据元素类别和关系生成布局
    - **Refinement**: 更新调整元素属性
    - **Completion**: 补全布局中缺失属性
    - **U-Gen**: 无条件生成
- 图像感知的文档布局生成：图像感知的文档布局生成任务不仅要考虑生成的各个元素之间的空间关系，还要考虑元素层和图像层的关系，例如海报设计。这类任务需要根据已知的图像生成文本、logo等元素，避免图像中主要内容的遮挡，保证语义和细节的完整性。

常用评价指标：

1. **Maximum IoU (mIoU)**: 计算生成布局与真实布局的IoU值。
2. **Overlap**: 度量生成布局内任何一对边界框之间的重叠面积。
3. **Alignment**: 计算生成布局内相邻元素之间的对齐距离。

4. **Constraint Violation Rate (Vio.%)**: 度量生成布局中元素违反约束的比率。

图像无关布局生成

图像无关的布局生成任务侧重于生成不依赖于背景图像的文档布局，仅关注各个元素之间的关系、位置和大小。这种方法广泛应用于生成如论文版面、PPT模板、商业海报等结构化的文档布局。其主要特点包括：

- **规范化要求高**：需要保证布局的对齐、元素之间的间距和重叠度等。
- **多样性强**：需要生成适应各种不同文档类型和内容的布局。
- **可扩展性好**：能够处理不同类型的文档布局生成任务。

核心问题：

1. **布局规范性**：生成的布局需要满足特定的规范，如对齐要求、元素间距、避免重叠等。
2. **数据稀缺**：高质量的布局标注数据稀缺，难以通过传统方法获取大量标注数据进行训练。
3. **多样性生成**：需要生成多样化的布局，以适应不同文档类型和内容的需求。

图像无关布局生成方法主要采用基于深度学习的生成模型，通过学习现有布局的特征，生成新的合理布局。主要的解决思路包括：

1. **深度学习模型**：利用深度学习模型，如VAE（变分自编码器）、Transformer、GAN（生成对抗网络）等，学习布局的分布特征。
2. **条件生成**：基于元素类别、大小、关系等条件生成布局，确保生成的布局符合特定的约束条件。
3. **数据增强**：通过生成合成数据来扩充训练集，增强模型的泛化能力。

相关工作主要包括：

1. **READ (Recursive Autoencoders for Document Layout Generation)**
    - **任务类型**: 学习训练数据中的布局结构，生成结构类似的新文档布局。
    - **任务难点**: 生成结构化布局、多样化布局和满足约束的布局。
    - 解决思路:
        - 结合递归神经网络与变分自编码器（VAE），将文档的结构化表示映射到一个紧凑的代码空间。
        - 使用数据增强技术，通过READ合成的文档布局增强训练数据，提高文档分析任务中的检测性能。
    - 实现细节:
        - **层次结构构建**: 用树状结构反映文档中各元素之间的相对位置和关系，递归合并相邻元素。
        - **空间关系编码器（SRE）**: 将文档的层次结构输入到一个单层神经网络中，映射为一个n维向量。
        - **空间关系解码器（SRD）**: 从高斯分布中采样得到的随机向量，使用解码器将其解码为新的文档层次结构。
2. **LayoutTransformer**
    - **任务类型**: 布局补全（由部分布局元素生成完整布局）。
    - 解决思路:
        - 使用Transformer模型进行布局生成，分属性建模，使得注意力模块更容易集中到关键属性上。

        - 采用自回归机制和"教师强制"策略，在训练和验证阶段指导模型生成布局。
    - 实现细节:
        - 使用自回归方式生成布局元素属性（如类别、位置、大小），通过逐步预测的方式生成完整布局。
3. **BLT (Bidirectional Layout Transformer)**
    - **任务类型**: 以类别为条件生成布局、以类别和大小为条件生成布局、无条件生成。
    - 解决思路:
        - 采用双向Transformer进行可控布局生成，克服传统自回归模型的局限性。
        - 采用分层采样和解码策略，每次选取一个语义组进行掩码预测，确保模型每次只预测相同语义含义的属性。
    - 实现细节:
        - 双向Transformer模型通过非自回归的方式生成布局，能够灵活处理不同条件下的布局生成任务。
4. **LayoutFormer++**
    - 解决思路:
        - 基于Transformer的编码器-解码器框架，通过序列化约束和解码空间限制来生成图形布局。
        - 处理各类布局生成的子任务，如布局补全、布局调优、基于元素类型的布局生成、基于元素大小的布局生成等。
    - 实现细节:
        - **解码空间限制策略**: 在推理阶段剪除预测分布中可能违反用户约束的选项。
        - **序列化约束方案**: 将用户约束表示为预定义格式的一系列标记，建模为序列到序列的转换问题。
5. **LDGM (Unifying Layout Generation with a Decoupled Diffusion Model)**
    - **方法的通用性**: 基于解耦扩散模型，将现有版面生成任务进行统一，实现更加通用的版面生成。
    - **任务类型**: 条件生成、布局细化、布局补全、无条件布局生成。
    - 解决思路:
        - 将布局生成任务统一为扩散过程和去噪过程，设定每个元素有三种状态：精确属性、粗粒度属性和缺失属性。
        - 利用扩散模型的思想，通过mask-and-replace策略进行解耦加噪。
    - 实现细节:
        - **解耦加噪策略**: 对于不同的属性分开进行加噪，同时采用不同的加噪策略。
        - **模型框架**: 基于Transformer的网络结构进行预测，所有输入进行量化，并使用相对位置编码建模元素间关系。

图像感知的布局生成

图像感知的布局生成任务需要在已有图像的基础上生成合理的文本、Logo等元素布局。这种方法广泛应用于广告设计、海报制作等需要视觉与文本结合的场景。其主要特点包括：

- **内容感知**：需要理解图像内容，并基于图像内容生成与之匹配的文本和其他元素布局。
- **避免遮挡**：生成的布局需要避免重要内容的遮挡，确保图像和文本的和谐。

- **视觉美感**：生成的布局不仅要符合语义逻辑，还需要具备美学上的吸引力。

核心问题：

1. **图像内容理解**：需要对输入图像进行语义理解，以便在生成布局时考虑图像内容。
2. **遮挡处理**：生成的文本和其他元素不能遮挡图像中重要的部分。
3. **美学评估**：生成的布局需要在视觉上美观，符合设计规范。

图像感知的布局生成方法主要采用深度学习模型，通过学习现有设计的特征，生成与图像内容匹配的新布局。主要的解决思路包括：

1. **图像特征提取**：利用卷积神经网络等模型提取图像的多尺度特征，捕捉图像的语义信息和视觉显著性。
2. **多模态融合**：结合图像特征和文本特征，通过多模态模型进行布局生成。
3. **对抗生成网络（GAN）**：使用GAN生成高质量布局，同时通过判别器确保生成布局的真实性和美观性。

主要工作包括：

1. **ContentGAN (Content-aware generative modeling of graphic design layouts)**
    - 任务目标：根据用户输入的视觉和文本内容生成符合内容语义的多模态布局（杂志布局）。
    - 任务难点：
        - **内容感知**：需要理解图像和文本内容，识别设计的主题、风格和目的。
        - **布局多样性**：需要生成多样化的布局，而不是依赖预定义的模板或手工规则。
        - **数据依赖性**：需要大规模、多样化的图形设计数据集，并对布局进行精细的语义标注。
    - 解决思路：
        - 采用多模态嵌入网络学习图像中的视觉特征和文字中的文本特征及设计属性特征，三者融合后作为布局生成网络的输入。
        - 使用条件生成对抗网络（GAN）进行布局生成。
    - 实现细节：
        - **多模态嵌入网络**：使用预训练的VGG16模型作为图像编码器，word2vec方法作为文本编码器，高维度的设计属性编码器。
        - **布局生成网络**：生成器负责将随机向量映射到布局样本，判别器区分生成布局和真实布局，编码器将实际布局映射到潜在空间的特征向量。
2. **Harmonious Textual Layout Generation**
    - 任务定义：在自然图像上生成和谐的文本布局，确保文本位置和大小的最优选择。
    - 任务难点：
        - **视觉美学规则复杂**：需要考虑图像内容、文本位置和大小等多种因素。
        - **高质量数据收集困难**：需要大量高质量的标注数据，收集和标注成本高。
    - 解决思路：
        - 结合语义特征和视觉感知原理，提出高效的布局学习网络。
        - 利用显著性网络生成视觉显著性图，引导文本区域生成。
    - 实现细节：
        - **显著性网络**：基于编码器-解码器架构，生成视觉显著性图。
        - **文本区域建议**：使用扩散方程生成文本驱动的概率图，基于此生成文本锚点。

        - **评分网络**：评估候选文本区域的美学质量，从中选择最佳布局。
3. **PDA-GAN (Unsupervised Domain Adaption with Pixel-level Discriminator)**
    - 任务定义：生成图像感知的广告海报图形布局。
    - 任务难点：域间差异导致源域数据和目标域数据之间存在显著差异。
    - 解决思路：结合无监督域自适应技术，使用像素级鉴别器实现跨域精细特征对齐。
    - 实现细节：
        - **像素级鉴别器**：通过细粒度特征空间对齐，缩小源域图像和目标域图像之间的差距。
        - **布局生成网络**：基于DETR的结构，包括多尺度CNN和Transformer编码器-解码器。
4. **PosterLayout**
    - 任务定义：自动在给定画布上布置预定义元素，如文本、logo和底纹。
    - 任务难点：处理层间关系，避免重要内容遮挡。
    - 解决思路：
        - 使用基于CNN-LSTM的增强生成对抗网络（GAN）进行布局生成。
        - 设计序列形成算法模拟人类设计师的设计过程。
    - 实现细节：
        - **设计序列形成（DSF）**：将布局中的元素按重要性排序，逐步重组布局。
        - **生成对抗网络（DS-GAN）**：使用CNN-LSTM模型处理视觉特征和设计序列，生成器负责生成布局，判别器区分生成布局和真实布局。
5. **LayoutPrompter**
    - 任务定义：基于大模型强大的上下文学习能力，生成内容感知的布局。
    - 解决思路：利用显著性图捕捉图像关键内容，将图像显著性信息序列化，输入LLM进行提示学习。
    - 实现细节：
        - **输入-输出序列化**：将用户约束和布局表示为序列，利用LLM中的布局相关知识。
        - **动态示例选择**：选择与测试样本的约束具有类似布局的提示示例。
        - **布局排序和渲染**：根据指标组合衡量布局质量，选择最优布局作为输出。

文档元素生成

文档元素生成任务主要指在文档中生成各种类型的元素，如文本、标题、图像、表格等。这些元素不仅仅是简单地插入，还需要根据文档的内容和布局进行合理的生成和摆放，从而提升文档的可读性和美观性。

CapOnImage: Context-driven Dense-Captioning on Image (EMNLP 2022)

任务类型：在图片对应文本框的具体位置上生成合适的文案。

任务难点：

- **上下文信息理解**：不同位置的文字描述需要充分利用周围的视觉上下文来生成最合适的描述，模型需建立图片-位置框-文案之间的强依赖关系。
- **文案冗余问题**：模型可能会为临近位置生成相似的多个文案，需要避免生成重复的文案。

解决方案：该工作提出了一种自动化的文案生成方式，利用多模态技术，综合考虑图片本身的信息（如商品类型、位置和背景色）、商品文本信息、文本框位置布局以及多个框之间的相互逻辑关系等信息自适应地生成合适的文案。

模型整体框架：

- 基于多层Transformer的多模态模型，将商品主图、当前布局边界框位置、前后边界框位置以及商品标题等信息进行嵌入后，输入到多层Transformer模型中。
- 提出了邻居增强位置编码模块和多级预训练任务，通过自回归的方式生成预测文案。

具体实现：

- **邻居增强位置编码模块**：
  - 该模块利用相邻文本框的位置作为上下文信息，增强当前文本框位置编码的表达能力。
  - 每个文本框表示为2D坐标，选择最近的前一个文本框和最近的下一个文本框作为邻居位置进行编码，将邻居位置与当前位置嵌入连接起来，并添加segment embedding。
- **多级预训练任务**：
  - **Caption Generation (CG)**：在整个预训练过程中持续进行，用自回归方式生成描述文本。
  - **Caption Matching (CM)**：逐步引入不同级别的负样本，通过渐进预训练策略提升模型能力。

AnyText: Multilingual Visual Text Generation And Editing (ICLR 2024)

任务类型：在图像中呈现准确和连贯的文本。

研究出发点： 尽管目前的合成图像技术高度先进，能够生成高保真的图像，但在生成图像中的文本区域时，效果并不理想，生成文本通常包含模糊、不可读或不正确的字符。

任务难点：

- **多语言文本生成**：现有文本编码器大多针对拉丁文字训练，难以处理其他语言的文本生成，特别是中文、日文和韩文等非拉丁文字。
- **文本与背景的无缝融合**：实现生成文本与图像背景的无缝融合是一个技术挑战，涉及到精确的文本定位和字符笔画信息的编码。
- **文本区域的专门监督**：大多数模型缺乏对文本区域的专门监督，导致生成文本的准确性和清晰度不足。

模型介绍：

- **AnyText**是一个基于扩散的多语言视觉文本生成和编辑模型，能够将指定的文本从提示符渲染到指定的位置，并生成视觉上吸引人的图像。还可以对输入图像内指定位置的文本内容进行修改，同时保持与周围文本样式的一致性。

模型特色：

- **多语言支持**：能够生成多种语言的文本，包括中文、英文、日文、韩文等。
- **多行文本生成**：用户可以指定在图像的多个位置生成文本。
- **变形区域书写**：能够生成水平、垂直甚至曲线或不规则区域内的文本。
- **文本编辑能力**：提供了修改图像中指定位置文本内容的功能，同时保持与周围文本风格的一致性。
- **即插即用**：可以无缝集成到现有的扩散模型中，提供生成文本的能力。

具体实现：

- **辅助潜在模块**：通过编码辅助信息如文本字形、位置和掩码图像到潜在空间，辅助文本生成和编辑。
- **文本嵌入模块**：采用OCR模型对文本笔画信息进行编码，并与来自分词器的图像标题嵌入进行融合，实现与背景无缝融合的文本生成。

- **文本感知损失**：在图像空间引入文本感知损失，进一步提高书写准确性。

CF-Font: Content Fusion for Few-shot Font Generation (CVPR 2023)

任务类型：自动的少样本字体生成，即仅根据少量目标字体的参考图像，将字体图像从已有字体转换到目标域，生成新字体的所有字符。

任务难点：

- **内容与风格的解耦**：现有方法多通过内容和风格的解耦实现字体生成，但完全解耦内容和风格特征是一个难题。
- **少样本学习**：少样本学习对模型的泛化能力提出了挑战，需要从有限的参考数据中提取有用的风格特征并应用于内容特征的转换。
- **字体内容特征的选择**：单一选择某一种字体作为内容特征的代表，可能导致生成的字体包含不完整或不需要的笔画。

模型整体结构：

- **内容融合模块（CFM）**：通过线性融合多个基准字体的内容特征，有效缓解内容与风格解耦不完全带来的影响。
- **投影字符损失（PCL）**：设计了一种基于概率分布距离的损失函数，更加关注字符的全局形状。
- **迭代风格向量优化（ISR）**：通过轻量级的迭代优化，进一步优化参考图像的风格表示向量，提升生成字体的质量。

具体实现：

- **内容融合模块（CFM）**：通过聚类算法选择代表性的基准字体，计算内容融合权重，对基准字体的内容特征进行加权平均，得到目标字体的融合内容特征。
- **投影字符损失（PCL）**：将字符图像的二维形状投射到一维空间，通过计算这些一维投影分布之间的距离来衡量图像的重建误差。
- **迭代风格向量优化（ISR）**：初始风格向量通过风格编码器提取，每个字符的风格向量取平均，进行迭代优化，生成的风格向量用于当前字符和其他字符的生成，提高推断效率。

多任务系统

多任务集成系统旨在统一处理多种设计任务，以提高文档生成的效率和质量。其由多个构成元素和任务分类组成，包括布局生成、文案生成、字体属性预测、文本/图片/元素填充等。这些任务相互关联，共同作用于多模态文档的生成。

文档的构成元素主要包括：

1. **文档布局**：确定文档中各个元素的位置和大小。
2. **文本属性**：预测文本的字体、颜色、大小等属性。
3. **文本内容**：生成文本内容，如标题、描述、标语等。
4. **图像**：包括图像的选择、编辑和摆放。
5. **元素之间相互影响**：各元素之间的相互关系和影响需要综合考虑，以保证整体文档的和谐美观。

相关任务主要可以分类为：

1. **布局生成**：根据文档的结构和内容生成合理的布局。
2. **文案生成**：生成符合文档内容的文本内容。
3. **字体属性预测**：预测文本的字体、颜色等属性，使其与整体文档风格一致。

4. **文本/图片/元素填充**：将文本、图片等元素合理地填充到文档中。

AutoPoster: A Highly Automatic and Content-aware Design System for Advertising Poster Generation (MM 2023) 是一个高度自动化的内容感知广告海报生成系统，仅需产品图片和产品描述作为输入，通过以下四个关键阶段生成不同尺寸的海报：

1. **图像清理和重定向**：检测并去除图像中的现有图形元素，重新定位产品图像以匹配海报尺寸。
2. **布局生成**：使用内容感知布局生成模型（如CGL-GAN和ICVT）生成图形元素的布局。
3. **标语生成**：采用CapOnImage模型，根据视觉和文本上下文生成不同位置的标语内容。
4. **风格属性预测**：使用多任务风格属性预测器，联合预测视觉风格属性（包括色彩相关属性和字体相关属性）。

具体实现：

- 图像清理和重定向：
  - 使用物体检测模型检测图像中的标志、标语和底图等图形元素。
  - 使用修复模型去除检测到的图形元素，并采用Outpainting方法无缝扩展图像区域。
  - 使用显著性检测模型生成显著性图，并裁剪图像达到目标的长宽比。
- **布局生成**：通过CGL-GAN和ICVT模型生成图形元素的布局。
- **标语生成**：结合视觉特征和文本描述信息，生成初步标语内容并确定其具体位置和样式。
- **风格属性预测**：通过视觉Transformer提取图像特征，并通过多层自注意力和交叉注意力机制预测图形元素的风格属性。

FlexDM: Towards Flexible Multi-modal Document Models (CVPR 2023) 模型通过多任务学习和域内预训练的方法处理多种设计任务，包括布局生成、文本填充、图片填充、元素填充、字体颜色设计等。其主要特点和实现方法如下：

1. **多模态掩码预测**：通过掩码预测策略（随机mask策略、元素mask策略、属性mask策略）来预测缺失的文档元素或属性。
2. **网络结构**：包括Encoder、Transformer blocks、Decoder，将每个元素的多个属性字段映射成固定维度向量。
3. **显式多任务学习**：在训练过程中，从目标任务中随机抽取任务进行训练，通过随机掩码的方式预训练模型。

具体实现：

- **多模态掩码预测**：在不完整文档中使用[MASK]填充的字段，通过模型预测生成完整文档。
- **网络结构**：通过Encoder映射元素的多个属性字段，Transformer blocks进行长距离依赖关系捕捉，Decoder重构文档内容。
- **显式多任务学习**：通过随机掩码预训练，结合具体任务（元素掩码预测、属性掩码预测）训练模型，提高模型的通用性和精度。