中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：20721585

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	多元异构数据融合的视觉富文档理解方法研究

作　　者　张俊伟
学科专业　计算机科学与技术
导　　师　王昊
完成日期　二〇二三年四月

姓　　名：张俊伟　　学号：20721585

论文题目：多元异构数据融合的视觉富文档理解方法研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名:

主席:

委员:

导　　师:

答辩日期:

姓　　名：张俊伟　　学号：20721585

论文题目：多元异构数据融合的视觉富文档理解方法研究

原创性声明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名: _____ 日期: _____

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定。即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

(保密的论文在解密后应遵守此规定)

签名: _____ 导师签名: _____ 日期: _____

上海大学工学硕士学位论文

多元异构数据融合的视觉富文档理解方法研究

作　　者: 张俊伟

导　　师: 王昊

学科专业: 计算机科学与技术

计算机工程与科学学院

上海大学

2023 年 4 月

A Dissertation Submitted to Shanghai University for the
Degree of Master in Engineering

Research on Visually-Rich Document Understanding for Multiple Heterogeneous Data Fusion

Candidate:
Supervisor: xxx
Major: Technology of Computer Application

Schoool of Computer Engineering and Science
Shanghai University
March, 2021

摘 要

随着国际贸易的发展与信息爆炸时代的到来，我们的日常生活中涌现出大量的视觉富文档数据，如物流表单、收据和简历等，数量急剧增长。实现视觉富文档的自动信息提取，将极大地增加社会经济效益。以国际物流单据数据为例，文档中包含大量有价值的信息，例如行业相关的实体和数字信息等。使用人工的方式提取文档中的这些关键信息会耗费大量的时间和精力。生活中的视觉富文档往往以图片、PDF、等多种形式存在，并且文档中蕴含着丰富的“视觉信息”和“排版信息”。视觉富文档的理解只考虑文本这一单一特征是不能充分理解文档的，需要综合考虑文本、视觉、以及版面结构等信息。本文通过分析视觉富文档的特点，所做的工作如下：

（1）**基于双流图网络的文档语义实体识别**: 现有多模态文档预训练模型的方法主要是基于 Transformer 架构的，例如 LayoutLM 模型，这些方法主要关注文档中例如 token 级别的细粒度元素，这使得它们很难从粗粒度元素中学习，包括自然词汇单元（如短语）和突出的视觉区域。本文针对视觉富文档理解中的语义实体识别任务，构造两个粒度的文档图神经网络，分别对每个 bounding box 内部的 token 和 bounding box 之间的文本块进行建模。在对 bounding box 建模时，尤其关注 Key 和 Value 之间的信息交互。在构建 bounding box 粒度的多模态特征时，除了使用文本特征外，将 OCR 得到的 bounding box 的坐标信息融入网络得到版面布局特征，并和图像特征相结合，使用图注意力网络对文档结构进行建模。在多个公开数据集、以及自建的海运单数据上达到了更好的效果。

（2）**基于 SPAN 图网络的文档实体关系抽取**：在进行视觉富文档关系抽取时，探究了使用不同的文档编码器，不同的特征，以及不同的粒度对文档关系抽取效果的影响。特别的，我们提出了一种构建 SPAN 粒度的图神经网络的方法，得到 SPAN 粒度的特征表示后，用于视觉富文档中的关系抽取。在多个数据集上的实验表明，所提出的模型效果远远优于基线模型中基于 token 粒度的关系抽取效果。

（3）**基于人类阅读顺序提升文档信息抽取效果**：使用 OCR 工具将视觉富文档转化为纯文本的过程中，对于版面结构复杂的文档，OCR 识别出的文本框（bounding box）存在顺序错乱的问题，本文设计出一种简单有效的算法，根据 bounding box 的

坐标重新构建文档阅读顺序，基于语义实体实体识别（SER）和关系抽取任务（RE）两个任务，在多个模型上进行实验，验证了所构建的算法的有效性。

本文研究成果对于视觉富文档的理解具有一定的意义。在公开数据集和自建国际物流货运单等多个视觉富文档数据集上的实验结果证明了本文所提出的模型和方法的有效性。

关键词：视觉富文档，阅读顺序，信息抽取，图卷积网络

ABSTRACT

With the development of international trade and the advent of the age of information explosion, there are a large number of visually rich document data in our daily life, such as logistics forms, receipts and resumes. The realization of automatic information extraction from visual rich documents will greatly increase economic and social benefits. Take the document of international logistics documents as an example. The document contains a large amount of valuable information, such as industry-related physical and digital information. Extracting this critical information from a document manually can take a lot of time and effort. The visual rich documents in daily life often exist in various forms such as pictures, PDF, etc., and the documents contain rich ”visual information” and ”typesetting information”. The understanding of visual rich documents cannot be fully understood only by considering the single feature of text. It is necessary to consider text, vision, layout structure and other information comprehensively. By analyzing the characteristics of visual rich documents, the work done in this paper is as follows:

(1) Existing methods based on document pre-training models, such as the LayoutLM model based on the Transformer architecture, focus on fine-grained elements in documents, making them difficult to learn from coarse-grained elements, including natural lexical units (such as phrases) and prominent visual areas. Based on the semantic entity recognition task in visual rich document understanding, this paper constructs two granular document graph neural networks and models the tokens inside each bounding box and text blocks between each bounding box respectively. When modeling bounding box, particular attention is paid to the information interaction between Key and Value. bounding box, in addition to text features, the coordinates information of bounding box obtained from OCR are integrated into the network to obtain the layout features of bounding Box, and combined with image features, the graph attention network is used to model the document structure. Better results were achieved on multiple public data sets, as well as self-built sea waybill data.

(2) When conducting visual rich document relation extraction, the influences of dif-

ferent document encoders, different features and different granularity on document relation extraction are explored. In particular, we propose a method to construct a SPAN granularity graph neural network. After obtaining the SPAN granularity feature representation, it can be used for relation extraction in visual rich documents. Experiments on multiple data sets show that the proposed model is much better than the token grain-based relational extraction in the baseline model.

（3）In the process of converting rich visual documents into pure text with OCR tools, the text boxes recognized by OCR for bounding box with complex layout structure have a disordered order. This paper designs a simple and effective algorithm, which reconstructs the reading order of documents according to Bounding Box coordinates.Based on two tasks of semantic entity Recognition (SER) and Relation Extraction (RE), experiments on multiple models verify the effectiveness of the constructed algorithm.

The research of this paper is of great significance to the understanding of visual rich documents. Experimental results on multiple visual rich document data sets such as public and self-built international logistics waybills demonstrate the effectiveness of the proposed model and method.

Keywords: Visual rich documents; reading order; document information extraction; graph convolutional networks

目 录

第一章　绪论

1.1　课题来源

1.2　研究背景与意义

随着经济的快速发展，视觉富文档带来了更为丰富和生动的信息呈现方式，并支持大量下游业务场景，如货运单、收据、合同文书、法律文书、电子病历和简历等，如图1.1所示。视觉富文档是指包含了视觉元素（如字体颜色、大小、样式）的文档 [1]。常见的视觉富文档如图1.2所示。与传统的纯文本不同，视觉富文档的语义不仅仅由文本决定，还与其中的视觉特征有关，例如布局、表格、图表、字体颜色和大小等，这些视觉元素对于文档的理解和分析具有重要作用。

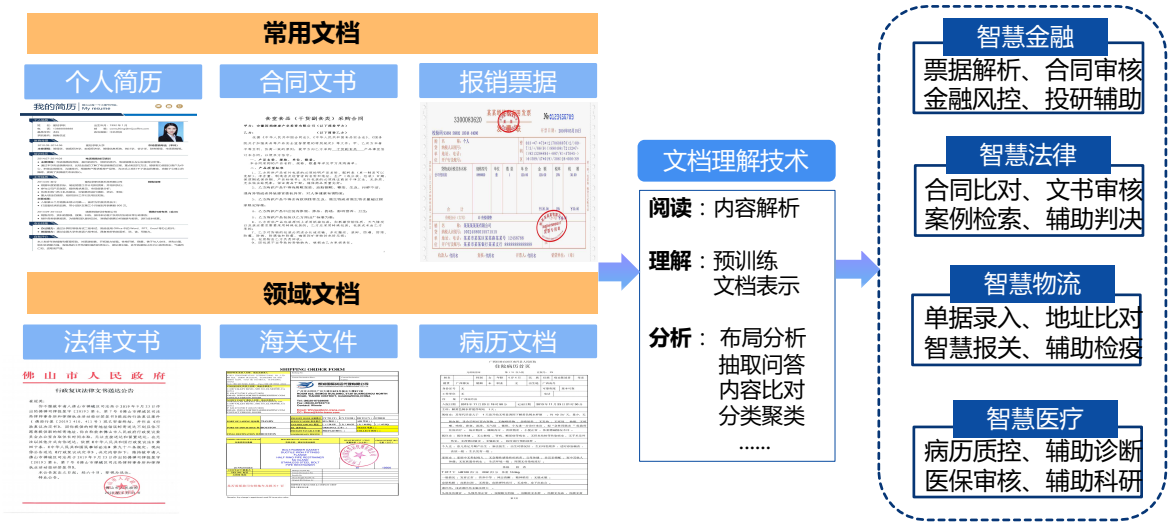
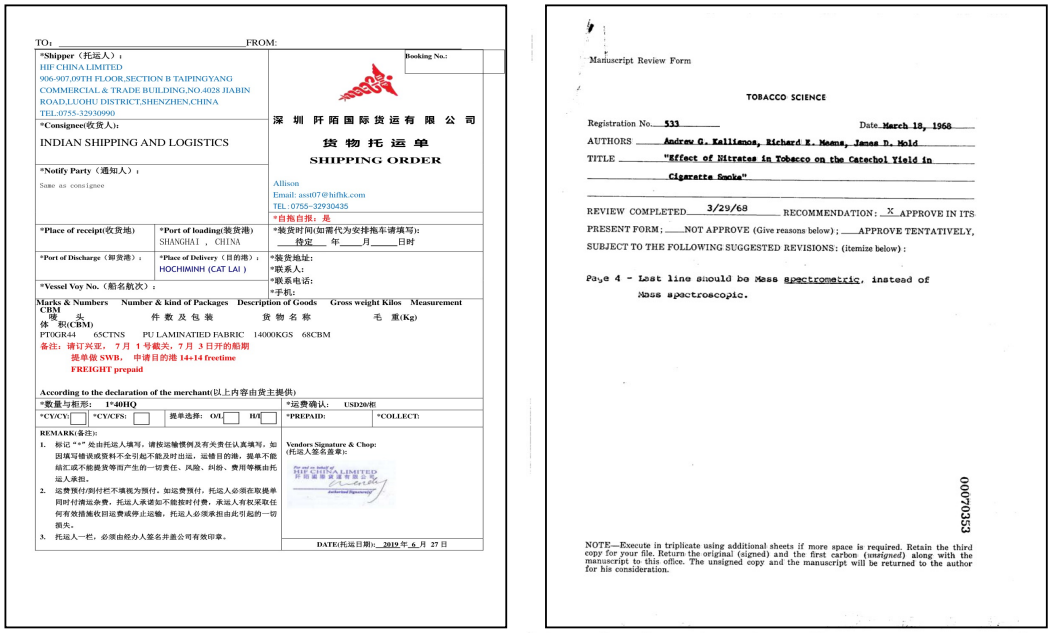


图 1.1　视觉富文档关键信息抽取应用场景

在 21 世纪，由于信息的快速增长，人们提出了信息抽取技术，以便从海量数据中快速准确地获取有用的信息。然而，目前大多数的信息抽取技术都是针对纯文本的，通常用于提取自由文本中包含的关键信息，包括实体、关系和事件等信息，并将其保存在数据库中 [2]。而处理视觉富文档需要更为复杂的技术和算法，当前视觉富文档的信息抽取技术仍在不断发展中。因此，对于视觉富文档理解的研究仍需要更进一步的探索。


TO: FROM:

*Shipper（托运人）：
HIF CHINA LIMITED
906-907,09TH FLOOR,SECTION B TAIPINGYANG COMMERCIAL & TRADE BUILDING,NO.4028 JIABIN ROAD,LUOHU DISTRICT,SHENZHEN,CHINA
TEL:0755-32930990

Booking No.:

深 圳 阡 陌 国 际 货 运 有 限 公 司
货 物 托 运 单
SHIPPING ORDER

*Consignee(收货人)：
INDIAN SHIPPING AND LOGISTICS

*Notify Party（通知人）：
Same as consignee

Allison
Email: asst07@hifhk.com
TEL: 0755-32930435

*自拖自报：是

*Place of receipt(收货地)
*Port of loading(装货港)：SHANGHAI ， CHINA
*装货时间(如需代为安排拖车请填写)：待定 年 月 日时

*Port of Discharge（卸货港）：
*Place of Delivery（目的港）：HOCHIMINH (CAT LAI)
*装货地址：
*联系人：
*联系电话：
*手机：

*Vessel Voy No.（船名航次）：

Marks & Numbers 唛 头	Number & kind of Packages 件 数 及 包 装	Description of Goods 货 物 名 称	Gross weight Kilos 毛 重(Kg)	Measurement 体 积(CBM)
PTOGR44	65CTNS	PU LAMINATIED FABRIC	14000KGS	68CBM

备注：请订兴亚， 7 月 1 号截关，7 月 3 日开的船期
提单做 SWB， 申请目的港 14+14 freetime
FREIGHT prepaid

According to the declaration of the merchant(以上内容由货主提供)

*数量与柜形： 1*40HQ
*运费确认： USD20/柜
*CY/CY: *CY/CFS: 提单选择: O/L H/L
*PREPAID: *COLLECT:

REMARK(备注):
1. 标记 "*" 处由托运人填写，请按运输惯例及有关责任认真填写，如因填写错误或资料不全引起不能及时出运，运输目的港，提单不能结汇或不能提货等而产生的一切责任、风险、纠纷、费用等概由托运人承担。
2. 运费预付/到付栏不填视为预付，如运费预付，托运人必须在取提单同时付清运杂费，托运人承诺如不能按时付费，承运人有权采取任何有效措施收回运费或停止运输，托运人必须承担由此引起的一切损失。
3. 托运人一栏，必须由经办人签名并盖公司有效印章。

Vendors Signature & Chop:
(托运人签名盖章):

DATE(托运日期): 2019 年 6 月 27 日

Manuscript Review Form

TOBACCO SCIENCE

Registration No. 533 Date March 18, 1968

AUTHORS Andrew G. Kallianos, Richard K. Means, James D. Hold

TITLE "Effect of Nitrates in Tobacco on the Catechol Yield in Cigarette Smoke"

REVIEW COMPLETED 3/29/68 RECOMMENDATION: X APPROVE IN ITS

PRESENT FORM; NOT APPROVE (Give reasons below); APPROVE TENTATIVELY, SUBJECT TO THE FOLLOWING SUGGESTED REVISIONS: (itemize below):

Page 4 - Last line should be Mass spectrometric, instead of Mass spectroscopic.

00070353

NOTE—Execute in triplicate using additional sheets if more space is required. Retain the third copy for your file. Return the original (signed) and the first carbon (unsigned) along with the manuscript to this office. The unsigned copy and the manuscript will be returned to the author for his consideration.

图 1.2 常见的视觉富文档

现实生活中各个领域使用的视觉富文档布局和样式复杂多变，没有固定的统一样式 [3]。以国际物流表单为例，不同的物流公司采用不同布局样式的物流表单。如果仅仅依靠人工处理来对文档进行关键信息抽取，将需要消耗大量的人力成本，工作内容重复而且烦躁。而基于模板匹配的方法又不具有可扩展性。因此，研究视觉富文档的关键信息自动化抽取非常重要。这样可以减少人力资源的浪费，并提高信息处理的效率和准确性。因此，迫切需要开发这样的技术以满足视觉富文档信息结构化提取的需求，这一技术具有极高的应用价值。

1.3 研究问题

针对以上问题，本文主要研究视觉富文档关键信息的自动化抽取方法、模型与技术。考虑到，视觉富文档理解中，实体、关系、阅读顺序对于文档理解尤为重要，因此，本文主要关注以下三个研究问题：

(1) 如何更好地利用视觉富文档中的多模态特征并进行有效融合

以前的工作 [4-5] 只用到了文档的文本特征，然而视觉富文档中蕴含着丰富的视觉特征，例如字体颜色和大小、文档的标题往往与内容的字体不同等等。对文本级任务来说，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。这些视觉特征可以对文档的理解起到重要的作用。

除了视觉特征外，文档的布局信息对于文档的理解也同等重要。观察视觉富文档可以发现，值（Value）的位置通常位于键（Key）的下方或者右方、发票文档中的价格往往被列在同一列中，并以“金额”作为列头。这些信息被那些仅仅依靠文本信息的模型所忽视，模型的性能也因此受到限制。因此，只用充分利用文档中的文本、视觉、以及位置等多模态特征，并采用合适的方式将三种特征进行融合，才能充分理解文档内容。

(2) **如何解决现有视觉富文档预训练模型缺乏对粗粒度元素关注的问题**

视觉富文档信息抽取（Visual Information Extraction，VIE）任务中有两个重要和基础的任务：语义实体识别（Semantic Entity Recognition，SER) 和关系抽取（Relation Extraction，RE）。语义实体识别可以看作一个分类或者序列标注任务，即预测每一个 token 的预定义类别，关系抽取任务也可以一个分类任务，即预测两个 bounding box 之间的关系。然而基于 Transformer 架构的文档预训练模型 [6-8] 主要关注细粒度 token 元素，对应图像网格（grid），使得它们很难从粗粒度的元素中学习，包括像短语这样的自然词汇单位和突出的图像区域。而 bounding box 级别的粗粒度关系对于语义实体识别和关系抽取任务都非常重要。在文档的键值对中，键（Key）的位置信息以及语义信息，对于值（Value）类别的判断是起到至关重要的作用的。关系抽取可以看做是 SPAN 粒度的分类问题，粗粒度的实体特征对关系抽取非常重要。在文档中，文本块的理解严重依赖于其周围的上下文，尤其是键值对之间的关系。[9] 发现对键值对之间的位置建模，会有效的提升语义实体识别模型的效果。即模型在预测每个 bounding box 中的 token 的类别时，不仅应该考虑 token 级别的细粒度信息，还应该考虑该 token 所在的 bounding box 对应的键值对的粗粒度关系信息。

(3) **视觉富文档的阅读顺序是否会影响信息抽取任务效果**

研究 [10-11] 发现，OCR 识别出来的视觉富文档中的 bounding box 顺序，对后续的 SER 任务和 RE 任务效果有很大的影响。版面结构复杂的视觉富文档，OCR 识别出的文本块的顺序存在乱序的情况，这会影响影响文档信息抽取任务效果。此外，

研究 [6] 表明，基于 Transformer 架构的视觉富文档预训练语言模型 [6-8] 存在无法对视觉富文档的阅读顺序进行建模的问题。所以针对上述问题，如何构建出正确的阅读顺序，提升视觉富文档的理解能力，是需要解决的问题。

1.4 研究内容

基于上述讨论，本文认为要解决视觉富文档理解中存在的问题，应该同时考虑：文档的多模态特征（文本、视觉、布局）、文档中元素的不同粒度、文档阅读顺序三个层面。如图1.3所示。

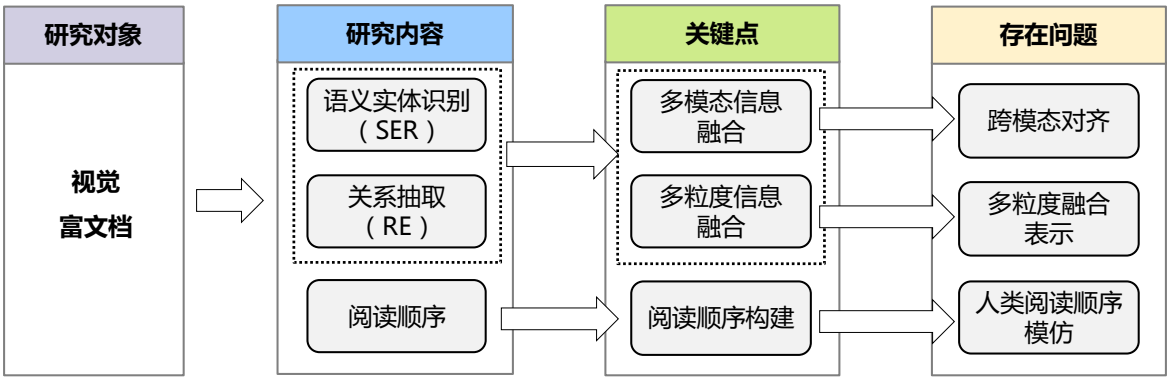


图 1.3　研究内容

（1）为了解决只使用单一文本特征存在的视觉富文档理解能力受限的问题，本文分析视觉富文档的特点，借鉴 LayoutLM 模型的方法，综合利用了文本、视觉、以及文档的版面布局信息，并探索三种模态特征有效的融合方式对视觉富文档的理解能力的影响。

（2）为了解决基于 Transformer 架构的文档预训练模型难从粗粒度元素中学习的问题，本文提出了一种新的同时结合粗细粒度的图神经网络键值匹配模型。我们构建 k 近邻的图注意力网络，以便每个输入节点只能关注其邻居节点。通过使用图模型对基于 bounding box 粒度的文档进行建模，使其更关注 bounding box 键值对之间的关系。此外，我们引入了局部熔断注意力图网络，进一步增强 bounding box 内部 token 之间的信息交互。综合实验表明，所提出的方法可以显著优于之前的方法。

（3）为了解决版面结构复杂的视觉富文档阅读顺序难以正确构建的问题，提出了一种针对 bounding box 坐标进行排序的算法。该算法可以将 OCR 识别出的非正常的 bounding box 阅读顺序，按照 bounding box 的坐标位置信息进行排序，构建出正常的

人类阅读文档的顺序，并分别在视觉富文档理解任务中的语义实体识别 (SER) 和关系抽取 (RE) 任务上做实验。实验结果表明，所构建出的阅读顺序，对于 SER 和 RE 任务的效果提升都有很大的帮助。

1.5 研究创新点

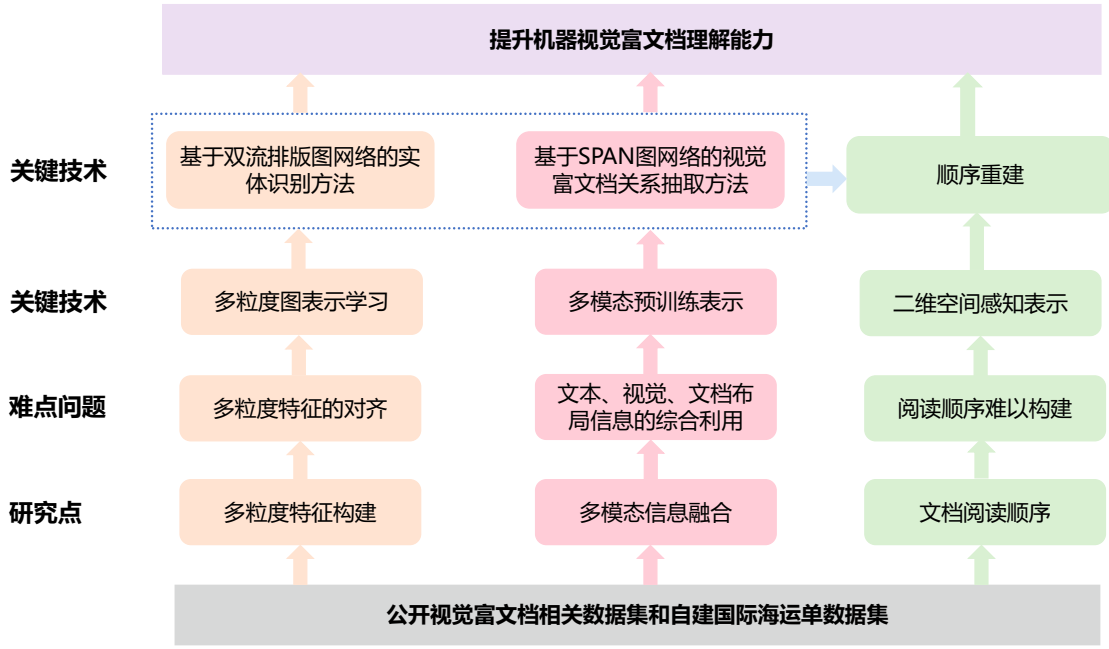


图 1.4　创新点

本文主要针对视觉富文档理解任务，从文档版面结构分析、多模态特征融合、多粒度的多模态特征融合、文档阅读顺序等方面对视觉富文档理解任务进行了研究，具体创新点如下：

(1) 早些年的视觉富文档理解方法中，只用到了文本特征。然而视觉富文档中包含着丰富的视觉特征，例如字体大小与字体的颜色，字体的样式等，都是可以利用到的视觉特征。除了文本特征和视觉特征外，文档的布局结构特征也同等重要。例如，根据文档中键值对的位置特点，值通常位于键的下方或者右方。要想提升视觉富文档的理解能力，只使用文本信息是远远不够的，必须同时利用文本，视觉，和二维文档布局特征。针对不同类型的视觉富文档版面结构复杂多变的问题，提出使用图神经网络建模文档的布局结构信息。具体的讲，经过 OCR 系统处理后，每张图像会生成一组文本块（bounding box），每个文本块包含其在图片中的坐标信息和文

本内容。为了构建一个 k 近邻的图神经网络，我们可以将每个文本块看作图中的一个节点，并使其与周围的，上、下、左、右节点相连。这样，每个节点都与它的 4 个最近邻节点相连，完成对文档结构的建模。使用图神经网络建模文档可以帮助我们更好地理解和分析文档中的多模态信息，从而提高视觉富文档理解的能力。针对诸如 LayoutLM 等视觉富文档预训练模型中，缺少对文档中键值对信息的建模，以及缺少对文档中粗粒度元素学习的问题，提出了一种同时结合粗细粒度特征的图注意力网络，分别对一个 bounding box 内部的所有 token 进行建模，以及对 bounding box 之间，尤其是键值对实体之间进行建模。在 SER 任务中，同时考虑两个粒度的多模态特征，能有效提升模型的效果。

(2) 在 RE 任务中，探究了不同编码器以及不同粒度的多模态特征对文档关系抽取效果的影响，通过构建 SPAN 粒度图神经网络得到 SPAN 粒度的多模态特征，在多个数据集上的实验表明基于 SPAN 粒度的关系抽取效果要远远优于基于 token 粒度的关系抽取效果。

（3）针对版面结构复杂的视觉富文档，OCR 识别出的 bounding box 顺序，存在难以正确表示阅读顺序的问题，LayoutLM 一族的文档预训练模型缺少对文档阅读顺序的构建。针对上述问题，提出了一种基于 bounding box 坐标进行排序的算法，能够正确构造出人类阅读文档的顺序，即从左到右，从上到下的顺序，有效的弥补了 OCR 识别结果顺序错乱的缺陷。通过在多个基线模型和我们提出的模型上进行语义实体识别和关系抽取的实验表明，所提出的阅读顺序构建算法，在两个任务上都有出色的效果提升。

1.6 研究的组织结构

本文首先介绍了视觉富文档理解的研究背景和意义，已有的视觉富文档理解模型、视觉富文档理解中的任务。本文的研究任务包括语义实体识别（SER）、关系抽取（RE）、以及文档阅读顺序构建。最后对本论文的研究工作进行了总结。本文的组织结构如下：

第一章：介绍了视觉富文档理解的研究背景和研究意义，针对当前视觉富文档理解研究所存在的问题，提出了三个主要的研究问题及研究内容，最后总结了本论文的创新点。

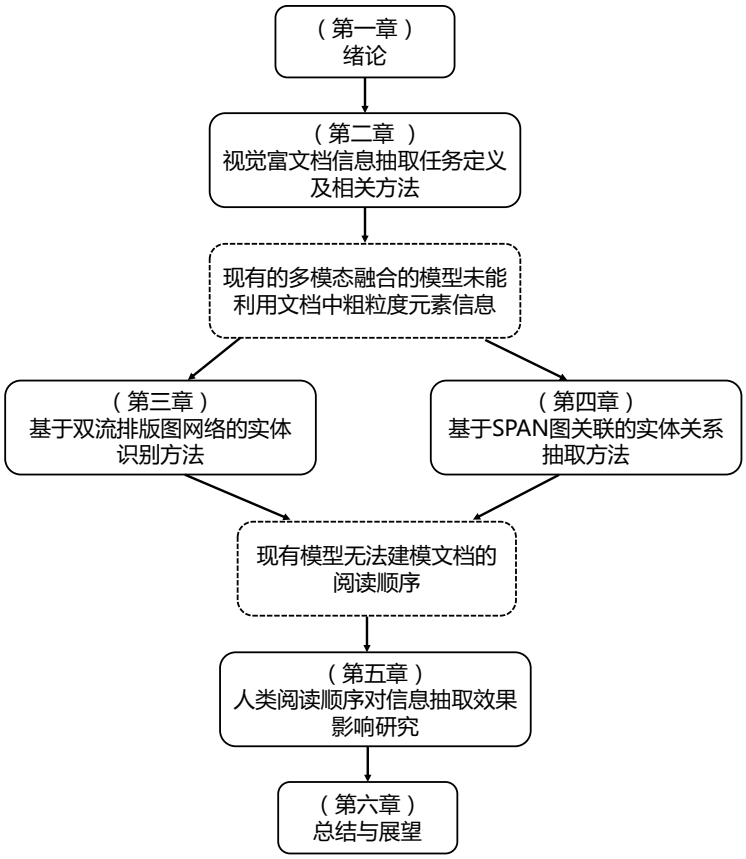


图 1.5　论文组织结构

第二章：对视觉富文档理解所包含的任务以及本文所研究的视觉富文档理解中的任务、所对应的方法和模型进行了介绍和概述。

第三章：针对视觉富文档理解中的语义实体识别（SER）任务，提出了一种同时结合粗细两种粒度多模态信息的图神经网络模型。通过构建 token 粒度与 bounding box 粒度的图神经网络并进行两种粒度多模态特征的有效融合进行视觉富文档中语义实体类别的判断。最后，通过实验验证了所提出的方法和模型的有效性。

第四章：提出了一种基于 SPAN 图神经网络的模型用于视觉富文档中的关系抽取任务（RE）。通过对 bounding box 粒度的实体关系进行精细化建模，更好的实现键值关系的抽取。在多个数据集上验证了所提出方法和模型的有效性。

第五章：介绍了文档阅读顺序构建的国内外研究概况及相关技术。提出一种简单有效的基于 bounding box 坐标排序进行人类阅读顺序构建的算法，在多个视觉富文档数据集以及多个任务上验证了所提出的算法的有效性。

第六章：总结与展望。归纳总结本文的工作和目前工作的不足之处，并对未来视觉富文档的研究方向进行了展望。

第二章　视觉富文档理解现有方法及模型

2.1　相关任务定义

在现实生活场景中，视觉富文档（Visually Rich Document，VRD）有很多种，比如发票、货运单、问卷调查表、申报材料等。仅靠文字理解文档是不够的，这些文档包含丰富的视觉信息和排版信息，可以帮助我们理解文档内容。近年来，许多工作集中在根据 OCR（光学字符识别）的结果从 VRD 中提取关键信息，OCR 识别结果包含许多个文本框（bounding box）[4, 12]。每个文本框包含：(1) 从语义和空间的角度来看属于一起的一组单词；(2) 文本框在图像中的坐标信息。我们将这些边界框和文本框内的描述，如文本，称为语义实体，每个实体包含词组和位置坐标。

如图2.1所示，视觉富文档理解（Visually-rich Document Understanding，VDU）中的下游任务，包括文档布局分析、文档信息抽取、文档视觉问答、文档图像分类等。在本文中，我们重点关注视觉富文档信息抽取，以及文档的阅读顺序对文档信息抽取结果的影响。

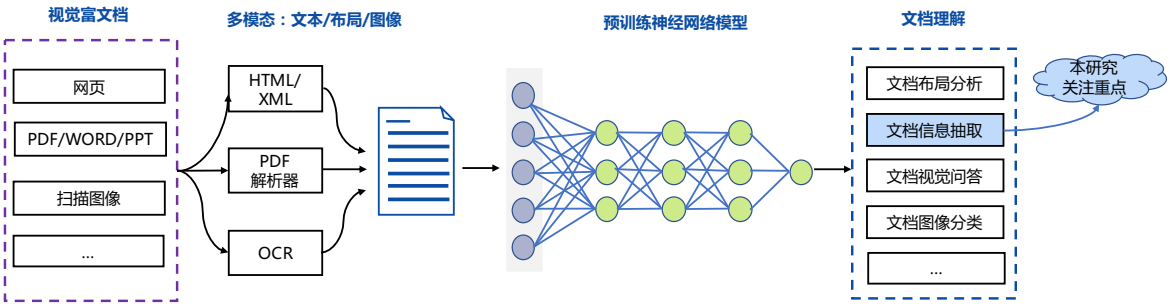


图 2.1　现有基于深度学习的视觉富文档理解技术框架

视觉富文档信息抽取（Visually-rich document Information Extraction，VIE）又被称为视觉信息提取，是指视觉富文档上进行信息抽取的任务，它通常包含两个子任务，语义实体识别（SER）和实体关系抽取（RE）。类似于传统自然语言处理 (NLP) 中的命名实体识别 (NER) 和关系抽取（RE），SER 任务旨在为 VRD 中的语义实体分配预定义的标签 [12-13]，实体关系抽取预测这些语义实体之间的关系 [8, 14-15]。与传统 NLP 中的 NER 和 RE 任务相比，来自 VRD 的 VIE 是一项更具挑战性的任务。首先，纯文本中的普通（命名）实体不像 VRD 中的语义实体那样包含布局信息。其

次，视觉富文档中蕴含着丰富的视觉特征，如果只是单纯的考虑文本信息，模型的关键信息抽取效果会大大受限。最后，面向文本中的关系抽取预测两个给定提及之间的关系，而 VRD 中的关系抽取需要预测文档中任意两个语义实体之间的关系。如图2.2所示，SER 任务是用标签“Answer”标记“夏艳辰”、用“Question”标记“姓名”。然而，在没有实体关系抽取的情况下，“夏艳辰”可以回答哪个问题仍然未知。与语义实体识别任务（SER）相比，实体关系抽取任务（RE）的探索较少，但它的好处至少包括：（1）提供更接近人类对 VRD 理解的额外结构信息，（2）当预定义的标签集合发生更改时，更容易转移到其他领域。因此，在本研究中，我们不仅关心语义实体识别任务，同时考虑语义实体关系抽取的任务，即发现两组文本块之间的关系，如图2.2中的黄色链接所示。

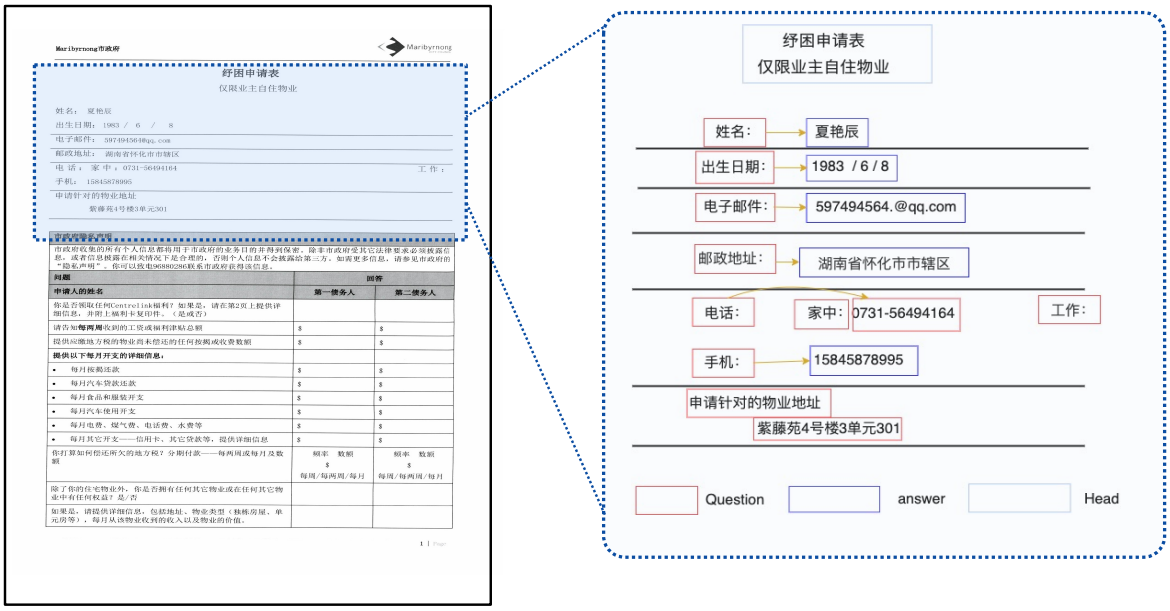


图 2.2　XFUN 中文数据集样例 (不同的颜色的矩形框代表不同的语义实体，矩形框之间的箭头表示语义实体之间的关系)

2.2　文字识别方法及模型

文字识别（Optical Character Recognition，OCR）是一种将图像或印刷体字符转换为文本的技术。该技术涉及到数字图像处理、模式识别、自然语言处理、人工智能等多个领域。OCR 软件可扫描文本或图像文件，将其中的字符识别出来，然后将它们转换为计算机可读的文本格式，如 Word 文档或文本文件。这样，用户就可以通

过计算机对这些文本进行编辑、搜索和分析，以提高工作效率。OCR 在商业、政府、教育等领域都有广泛的应用，例如数字化档案、自动化数据输入、识别票据、自动识别手写文字等。

文本识别分为两个步骤：文本检测和文本识别。随着深度学习的快速发展，基于深度学习神经网络的文字识别技术也得到了快速发展，并且取得了显著的进步。下面将对两个步骤分别进行介绍。

（1）文本检测

文本检测是指从自然场景图片或视频中检测出文本区域，它是 OCR、图像处理、计算机视觉等领域的重要基础任务之一。近年来，深度学习技术的发展使得文本检测在精度和效率上都取得了很大进展。

现阶段的基于深度学习的文本检测方法可以分为基于候选区域和基于图像分割两大类。由于目前文本检测技术较为成熟，亦非本文关注的重点，因此简要介绍如下：

基于候选区域的文本检测方法：这种方法首先生成可能包含文本的候选区域，然后对这些区域进行分类，以确定哪些区域包含文本。这种方法通常使用卷积神经网络（CNN）作为分类器，因为 CNN 在图像分类任务中表现出色。在生成候选区域方面，常用的方法包括滑动窗口和锚框（anchor boxes）。

基于候选区域的文本检测方法通常以 RCNN 网络为框架。[16] 提出了两阶段范式（首先成候选框，再对候选框进行分类 + 微调）。RCNN 是一种基于候选区域的目标检测方法，它首先使用选择性搜索（Selective Search）等方法生成可能包含目标的候选区域，然后将这些候选区域提取为固定大小的特征图，并使用卷积神经网络（CNN）进行分类和回归。RCNN 的主要优点是可以处理各种尺度、形状和方向的目标，并且在训练过程中可以利用更多的上下文信息，提高检测性能。然而，RCNN 也存在一些缺点，例如计算量大、训练过程复杂等。[17] 设计了 CTPN （Connectionist Text Proposal Network）。该方法结合了卷积神经网络（CNN）和循环神经网络（RNN）的优点，旨在实现一种能够直接检测文本行的端到端文本检测算法。CTPN 主要思想是将图像中的文本行作为一个序列，并使用 RNN 对序列进行建模，同时使用 CNN 提取特征并作为 RNN 的输入。具体来说，CTPN 首先将图像中的像素沿水平方向划分为多个列，然后在每个列上使用 CNN 提取特征，并在每个位置处预测一个文本行

的得分以及对应的文本框。这些得分和文本框被视为一个序列，并输入到 RNN 中进行建模，以生成文本行的边界框。最后，CTPN 使用非极大值抑制（NMS）对检测结果进行筛选和去重。相比于其他基于候选区域的方法，CTPN 的优点在于可以直接检测文本行，不需要额外的文本行合并操作，并且能够处理任意方向的文本。同时，CTPN 的检测速度也比较快，可以实现实时文本检测。

基于图像分割的文本检测方法：这种方法通过对整个图像进行像素级别的分割，将图像中的文本区域与其他区域区分开来。这种方法通常使用全卷积神经网络（FCN）或 U-Net 等语义分割网络作为模型。相比于基于候选区域的方法，基于图像分割的方法具有更好的像素级别精度和鲁棒性，但是计算复杂度更高。值得注意的是，现阶段的文本检测方法通常是基于两种方法的结合，例如先使用基于候选区域的方法生成候选区域，然后使用基于图像分割的方法对这些区域进行进一步的筛选和细化。

SSD[18] 将特征图上每个位置的 b-box 的输出离散化成了一系列不同尺度和纵横比的默认框。预测时，网络对每个默认的框的每个目标类别进行打分，并通过调整 box 来更好的适应目标的形状。此外，本网络通过结合多种分辨率的不同尺度的特征图来适应不同大小的目标。TextSnake[19] 是一种适用于任意形状的文本检测方法，其将文本表示为一个“蛇形”（Snake）曲线。该算法使用一个神经网络来输出文本的蛇形曲线表示，然后使用一组形态学变换将其转换为文本区域。SegLink[20] 通过在每个像素点处同时预测文本的存在和文本的边界信息，实现了一种直接从图像中提取文本信息的端到端文本检测算法。SegLink 的主要思想是利用 CNN 对图像进行特征提取，然后在特征图上对每个像素点同时预测文本的存在和文本的边界信息。具体来说，SegLink 将每个像素点看作一个中心点，然后在以该中心点为中心的若干个方向上，预测文本的存在概率和边界信息。这些预测结果被视为线段的端点，并通过线段合并和筛选操作得到最终的文本框。SegLink 相比于其他文本检测算法的优点在于可以直接从图像中提取文本信息，避免了候选区域生成和合并的过程，提高了检测的效率和准确率。同时，SegLink 也能够处理任意方向的文本，具有较好的鲁棒性。然而，SegLink 也存在一些缺点，例如在图像中存在较小的文本时，检测精度较低。

EAST（Efficient and Accurate Scene Text）(Zhou et al. [21]) 是一种基于深度学习的文本检测模型，旨在准确高效地检测场景中的文本。EAST 模型最初由香港中文大

学提出，已经被广泛应用于文本检测领域。EAST 模型的主要思想是将文本检测任务转化为一个回归问题，通过预测文本区域的四个顶点坐标来定位文本区域。具体来说，EAST 模型包括一个基于深度学习的卷积神经网络（CNN）和一个合并层，以及一个后处理模块。在卷积神经网络中，EAST 模型使用多尺度特征图和横向连接，以获得更好的文本检测结果。合并层用于将卷积神经网络输出的多个尺度特征图进行融合，以进一步提高检测性能。后处理模块则用于过滤非文本区域，以及进一步调整文本区域的位置和大小。相对于其他文本检测模型，EAST 模型具有高效、准确的优点，已经在许多实际场景中得到了广泛的应用，如车牌识别、图像翻译、场景文本识别等。[22] 提出了 PixelLink，PixelLink 是一种基于实例分割的文本检测方法，其使用一个联合实例分割和文本检测网络来检测和定位文本。该算法将每个文本实例视为一个独立的对象，并使用像素链接（PixelLink）技术将它们与相邻像素链接起来，形成文本区域。这种只看该像素与其周围邻居而忽略了更多的上下文信息的方式可能会使 PixelLink 产生一些误检。针对 PixelLink 的缺点，[23] 提出 PSENet 文本检测算法，基于像素级别的分割，对任意形状的文本进行定位，使用渐进的尺度扩展算法，对相邻文本实例进行识别。以 ResNet 为主干网络，对于弯曲的文本达到了很好的效果。但是该算法的缺点是检测速度比较慢，不能很好的满足实时性的需求 [23]。CRAFT[24] 是一种基于字符区域感知的文本检测方法，其可以检测任意方向和形状的文本。该算法使用一个神经网络来输出字符区域，并将其组合成文本区域。该算法还可以对字符进行分组，以提高文本识别的准确性。

（2）文本识别

在文本识别算法研究的早期阶段，主要的方法是先将文本行中的字符进行分割，然后使用分类算法对每个单独字符进行识别。然而，当图像背景复杂时，这种方法对单个字符的识别效果很差。

场景文本识别是指从场景图像中自动检测和识别出文本信息的过程。深度学习时代，场景文本识别模型通常使用卷积神经网络（CNN）将图像编码为特征空间。两个主要的技术是连接主义时空分类 [25]（CTC）和编码器-解码器框架 [26]，主要的区别在于文本内容解码模块。常用的文本识别模型框架如图2.3所示。

连接主义时空分类（Connectionist Temporal Classification，CTC）是一种常用的文本内容解码方法，它是一种无监督的序列学习算法。CTC 可以将变长的输入序列映

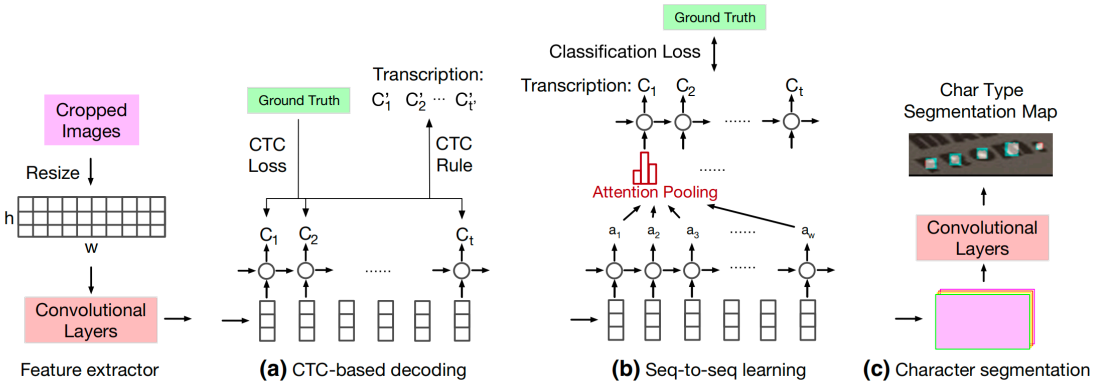


图 2.3　文本识别模型的框架 [27]：(a) 序列标记模型，并在训练和推理中使用 CTC 进行对齐 [28]；(b) 序列到序列模型，并可以使用交叉熵直接学习 [29]；(c) 基于分割的方法 [30]。

射到固定长度的输出序列，并且可以对缺失的标签进行处理。CTC 适用于不需要对文本进行字符级别的分割和识别的场景，例如场景文本识别中的行识别任务。CTC 解码模块采用的是语音识别，在时域中数据是连续的。CTC 在 OCR 领域的首次应用可以追溯到 [31] 的手写识别系统。现在这项技术在场景文本识别中被广泛采用 [28, 32-36] 。最初的尝试可以称为卷积递归神经网络（CRNN），这些模型是在 CNN 的基础上堆叠 RNN 组成的。[33] 是第一个 CRNN 模型。它的实现原理是在输入图像上应用 CNN 模型以生成卷积特征片，然后将其送入 RNN。[28] 则采用全卷积方法将输入图像作为一个整体进行编码以生成特征片，利用了 CNN 不受输入的空间大小限制的特性。

编码器-解码器框架则是一种基于序列到序列学习的方法，它使用编码器将输入序列映射到一个固定长度的向量表示，并使用解码器从该向量表示中生成输出序列。在场景文本识别中，编码器可以是卷积神经网络，解码器可以是循环神经网络（RNN）或者加上注意力机制（attention mechanism）。[37] 提出了用于无词典场景文本识别的递归递归神经网络和注意力建模。该模型首先通过递归卷积层传递输入图像，提取编码后的图像特征，然后利用隐式学习字符级语言统计的递归神经网络将其解码为输出字符。基于注意力的机制执行软特征选择，以更好地使用图像特征。[29] 观察到现有的基于注意的方法存在注意漂移问题，并提出对注意分数施加局部监督以减弱注意漂移。

CTC 和编码器-解码器框架最初都是为一维序列输入数据设计的，因此适用于直线和水平文本的识别，这些文本可以被 CNN 编码为一连串的特征帧而不丢失重要信息。然而，定向和弯曲文本中的字符分布在一个二维空间。为了适应 CTC 和编码

器-解码器框架，在特征空间中有效地表示定向和弯曲的文本仍然是一个挑战，其解码需要一维的输入。对于定向和弯曲的文本，直接将特征压缩成一维的形式可能会失去相关的信息，并带来来自背景的噪音，从而导致较差的识别精度。总体而言，CTC 适用于不需要字符级别分割和识别的场景，而编码器-解码器框架则适用于需要字符级别分割和识别的场景。当然，这两种方法也可以结合使用，例如先使用 CTC 进行行识别，再使用编码器-解码器框架进行字符级别的分割和识别。

[30] 将识别任务转化为语义分割，并将每个字符类型视为一个类。该方法对形状不敏感，因此对不规则文本有效，但缺乏端到端训练和序列学习，容易出现单字符错误，特别是在图像质量较低的情况下。他们还首先通过填充和转换测试图像来评估他们的识别方法的鲁棒性。

2.3 版面分析方法及模型

文档布局分析（Document Layout Analysis，DLA）是一个致力于从文档图像中提取语义信息的重要研究领域。作为文档理解系统的一个关键预处理步骤，DLA 可以为文档检索、内容分类和文本识别等多种应用提供参考信息。由于文档布局的多样性和复杂性，此任务极具挑战性。

在早期阶段，经典的文档布局分析策略可以大致分为两类：自下而上和自上而下。通常，自下向上的方法从较小粒度的数据级别动态地获得文档分析结果。自下而上的分析策略使用文档中的局部特征，例如像素分布，以确定单个像素或单词。然后，再对检测到的元素进行扩充，形成更大的文档区域，直到它们达到分析级别。虽然自下而上的策略具有较高的准确性，能够处理复杂的布局。然而，大多数方法的时间复杂度很高，需要很大的空间。经典的自下而上策略可以细分为五个核心类别：连通成分分析 [38]、纹理分析 [39]、基于学习的分析 [40]、Voronoi 图 [41] 和 Delaunay 三角剖析 [42]。

自上而下的策略迭代地将整个页面划分为小文档元素，如列、块、和单词。类似地，自上而下的策略可以细分为四类：基于文本的分析 [43]、RunLength Smealing Algorithm（RLSA)[44]、DLA 投影轮廓 [45] 和结合空间的分析 [46]。通常，自顶向下的策略在常规文档布局上非常有效。然而，尽管它可以处理常规的矩形文档布局，但自下而上的策略和自上而下的策略不足以有效地处理复杂的布局。

最近，随着深度学习的快速发展，DLA 研究取得了重大进展。受全连接卷积神经网络（FCN）[47] 的启发，提出了许多基于深度学习的方法。这些方法将 DLA 视为一项特殊的分割任务 [48]。然而，设计一个高效的 DLA 系统仍然是一个难题。准确预测文档中的区域类别受到以下因素的限制：缺乏足够的训练数据和合适的端到端模型。

在 [49] 的工作中，DLA 任务被认为是一种特殊的语义分割任务，其目标是对分割对象进行像素级理解。基于深度学习的 DLA 方法主要受到完全卷积网络（FCN）[47] 的启发。Heet 等人 [50] 使用具有多尺度特征的 FCN 进行语义页面分割，并执行辅助元素轮廓检测任务。[51] 采用改进的 FCN 来处理历史文档分割。Liet 等人 [52] 提出了一种新颖的端到端深度神经网络，名为 DeepLayout，用于页面布局分析。

低维特征包含更详细的信息，高维特征包含更多信息 [53]。为了生成更稳健的特征，FCN 使用跳跃连接结构融合低维特征和高维特征。然而，跳跃连接结构不能利用高维特征，因为所包含的高维类别语义信息被低维特征所混淆。因此，[54] 提出了一种动态分割特征融合模块。该模块实现了图像细节的恢复，同时维护了类别语义信息。

现有的基于深度学习的文档版面分析按照使用的模态特征类别大致可分为两种：基于 NLP 的方法将版面分析任务看作是序列标签分类任务（sequence labeling），但是该类方法在版面建模上表现出不足，无法捕获空间信息；基于 CV 的方法则将版面分析看作是目标检测或分割任务（object detection or segmentation），该类方法的不足表现在：缺乏细粒度的语义、简单的拼接方式、未利用关系信息。

单模态布局分析仅利用视觉特征 [55-56]（文档图像）或仅利用语义特征（文档文本）来理解文档结构。使用视觉特征，已经提出了一些工作 [51, 57]，以将 CNN 应用于分割各种对象，例如文本块 [58]、文本行 [59-60]、单词 [61]、图形或表格 [50, 62]。同时，还有一些方法 [63-65] 试图使用语义特征来解决布局分析问题。然而，所有上述方法都严格限于视觉或放射特征，因此无法利用其他模态的补充信息。

多模态布局分析试图结合视觉和发散模态的信息。相关方法可进一步分为两类，基于 NLP 的方法和基于 CV 的方法。基于 NLP 的方法用于低级元素（例如，token）和模型布局分析，作为序列标记任务。[66] 用于识别表单结构。[67] 被提议作为多模态布局分析的大规模数据集，并且已经发布了几个 NLP 基线。然而，上述方法显示

出嵌入建模的能力不足。基于 CV 的方法通过文本嵌入映射引入文档语义，并将模型布局分析作为对象检测或分割任务。[68] 引入句子粒度语义，并在决策层（网络末端）插入文本嵌入映射，而 [69] 引入字符粒度语义并在输入层插入文本嵌入图。尽管取得了巨大的成功，但上述方法也存在以下局限性：所使用的语义有限、简单的模态融合策略以及缺少组件之间的关系建模。

2.4 多模态表示学习

基于大规模预训练模型的文本表示：预训练模型是指在大规模数据集上进行了预训练的机器学习模型。预训练模型在计算机视觉、自然语言处理、语音识别等领域中广泛应用，已经成为了这些领域中的一个重要技术。预训练模型的基本思想是，在大规模数据集上预先训练一个深度神经网络模型，并在后续任务中将其作为初始参数，再进行微调或者 fine-tuning，从而获得更好的效果。通过预训练，模型可以学习到更多的知识和特征，提高了模型的泛化能力和效果。预训练模型的训练方法可使用自监督学习技术（如自回归的语言模型和自编码技术），可训练单语言、多语言和多模态的模型。此类模型可经过微调之后，可以在机器翻译、情感分析等多个下游任务中应用。

ELMO[70]（Embedding From Language Models），是一种基于深度双向语言模型的词向量表示方法，由斯坦福大学和 AllenNLP 团队共同提出。由于 word2vec 是一种静态词嵌入方法，它不会随着上下文场景的变化而变化。但一词多义在现实中太常见了，因此这种静态嵌入的局限性在很多场景显得力不从心。而 ELMo 就是为解决这些问题提出的，它的提出意味着从词嵌入 (Word Embedding) 时代进入了语境词嵌入 (Contextualized Word-Embedding) 时代。与传统的静态词向量表示方法（如 word2vec 和 GloVe）不同，ELMo 词向量是动态的，它会根据上下文的不同而变化。ELMo 模型基于双向语言模型（Bidirectional Language Model，BiLM）进行训练。BiLM 能够利用上下文信息预测当前单词的概率，即在正向和反向两个方向上分别构建一个语言模型。在预测时，ELMo 模型会将输入的文本经过 BiLM 得到每个词的多层表示，这些表示不同的深度代表不同层级的语言特征，比如词义、句法和语境信息。ELMo 模型会将这些表示按照一定的权重进行加权求和，得到一个动态的词向量表示。相比于静态词向量，ELMo 模型的优势在于它能够更好地捕捉上下文信息，因为它能

够利用双向语言模型在句子中的左右两个方向上进行建模，从而更好地表达词语在不同上下文中的含义。ELMo 词向量已被广泛应用于各种自然语言处理任务，包括文本分类、问答系统、命名实体识别等。

GPT[71]（Generative Pre-trained Transformer）是一种基于 Transformer 架构的自然语言处理模型，由 OpenAI 团队于 2018 年提出。GPT 模型的主要思路是在大规模语料库上进行预训练，然后在各种 NLP 任务上进行微调。GPT-3 是目前最大的自然语言处理模型之一，它是通过预训练来学习大规模语料库中的语言模式，并在各种 NLP 任务上微调。GPT-3 包含数十亿个参数，具有很强的语言生成和理解能力，可以进行文本生成、翻译、问答等多种任务。GPT-3 的预训练过程使用了无监督的语言模型预测任务，通过预测单词、句子和段落的下一个词、句子和段落来学习语言模式，从而能够生成连贯的文本。此外，GPT-3 还支持零样本学习，即可以在没有任何新数据的情况下学习新任务，这使得它在自然语言处理领域具有很高的潜力。

BERT[72]，全称为“Bidirectional Encoder Representations from Transformers”，是一种自然语言处理（NLP）模型，由 Google 公司在 2018 年发布。BERT 模型使用了 Transformer 模型架构，采用了预训练的方法，在海量文本数据上进行了训练，可以有效地理解自然语言的含义和语境。与传统的自然语言处理模型不同，BERT 模型是一种双向的预训练语言模型，这意味着在处理自然语言时，它会同时考虑上下文的信息。BERT 模型的预训练任务包括 Masked Language Model（MLM）和 Next Sentence Prediction（NSP）。在 MLM 任务中，BERT 模型需要预测句子中被 MASK 掉的单词是什么；在 NSP 任务中，模型需要判断两个句子是否是相邻的，并给出一个二元分类的结果。BERT 模型可以用于多种 NLP 任务，例如文本分类、命名实体识别、语义相似度计算、问答系统等。

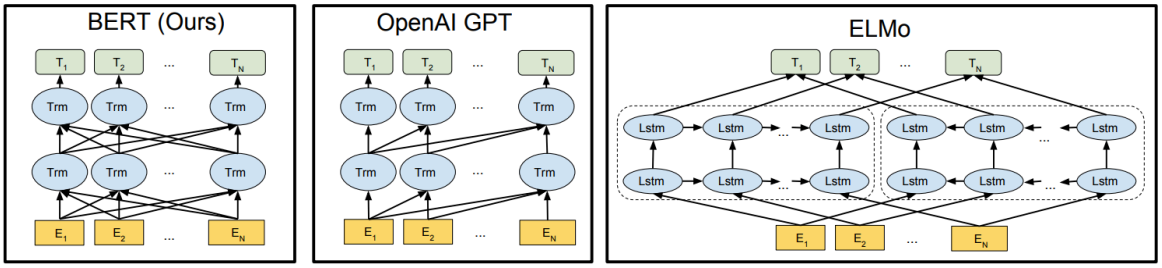


图 2.4 预训练模型结构的差异。BERT 使用了一个双向的 Transformer[72]

图神经网络结构表示：图可以用来表示包括社会科学（社会网络、自然科学）、蛋白质相互作用网络和知识图谱等许多其他研究领域在内的各个系统。图神经网络（Graph Neural Networks，简称 GNN）是一种神经网络模型，它旨在对图结构数据进行机器学习和深度学习任务。GNN 可以有效地从图中提取节点和边的特征信息，同时考虑它们在全局结构中的相互作用，从而实现对图数据的高效分析和处理。它可以处理各种类型的图形数据。

GNN 可以理解为一个节点嵌入（node embedding）和图嵌入（graph embedding）的过程。节点嵌入的目的是将每个节点映射到低维向量空间中，以便在此向量空间中对节点进行操作和计算。图嵌入的目的是将整个图映射到向量空间中，以便在此向量空间中对整个图进行操作和计算。这些操作和计算可以包括分类、预测、聚类、链接预测等任务。

GNN 的核心思想是从每个节点的邻居节点中汇总信息并更新节点的表示。这个过程可以被描述为一个逐层传播（message passing）的过程。每个节点的邻居节点发送信息到该节点，该节点收集所有信息，并将其与自身的表示结合起来进行更新。这个过程可以被看作是一种局部信息汇聚的过程，可以从整个图的角度来理解，这种局部信息汇聚可以导致全局的图信息。

GNN 是近年来非常活跃的研究领域，有许多优秀的模型被提出，如 GCN（Graph Convolutional Networks）、GAT（Graph Attention Networks）等，它们在许多任务上都取得了很好的效果。。

对于图，我们有以下特征定义：对于图 $G = (V, E)$，V 为节点的集合，E 为边的集合，对于每个节点 i，均有其特征 x_i，可以用矩阵 X_{N*D} 表示。其中 N 表示节点数，D 表示每个节点的特征数，也可以说是特征向量的维度。

图卷积神经网络（Graph Convolutional Networks, GCNs）是最为基础和常用的图神经网络模型之一，也是许多复杂模型的核心组成部分。GCNs 的基本思想是将普通卷积操作应用到图数据上，通过对当前节点和其邻居节点的特征进行卷积运算，来提取拓扑图的特征空间。这样，GCNs 可以有效地从图数据中提取节点和边的特征信息，并将它们融合到整个图结构中，从而实现对图数据的高效建模和分析，其过程如图2.5所示。

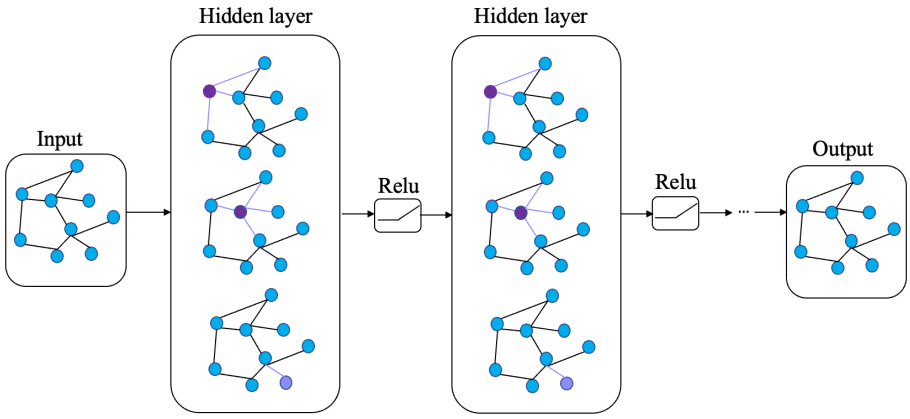


图 2.5　图卷积过程

图注意力网络（Graph Attention Network，GAT）是一种基于注意力机制的图神经网络（Graph Neural Network，GNN）模型。GAT 通过引入注意力机制来实现节点的信息聚合，这使得 GAT 可以针对不同节点关系赋予不同的权重。多头注意力机制的计算过程如图2.6（左）所示。具体地，GAT 通过对邻居节点的特征向量进行加权平均来更新节点的表示向量。权重的计算是通过注意力系数来完成的，注意力系数是由两个节点之间的特征向量计算得出的。通过注意力系数的引入，GAT 可以根据节点之间的相互关系，对邻居节点的特征向量进行不同程度的加权平均，以此来更新节点的表示向量。GAT 模型中的注意力机制是通过自注意力（self-attention）机制实现的，这个机制也常被用于自然语言处理中的文本表示学习。GAT 的自注意力机制可以将一个节点的表示向量与其邻居节点的表示向量进行计算，并给予不同邻居节点不同的权重。具体来说，对于一个节点 i，它的表示向量 h_i 可以表示为以下形式：

$$h_i = \delta\left(\sum_{j\epsilon N_i} \alpha_{ij} W h_j\right) \tag{2.1}$$

其中 N_i 表示节点 i 的邻居节点集合，α_{ij} 表示节点 i 与邻居节点 j 之间的注意力系数，W 表示一个可学习的权重矩阵，σ 表示一个激活函数。在 GAT 中，注意力系数 α_{ij} 可以通过以下方式计算：

$$e_{ij} = \vec{a}\left(W\vec{h}_i, W\vec{h}_j\right) \tag{2.2}$$

$$\alpha_{ij} = \text{softmax}_\text{j}\left(e_{ij}\right) = \frac{exp\left(e_{ij}\right)}{\sum_{k\epsilon N_i} exp\left(e_{ik}\right)} \tag{2.3}$$

其中 $\vec{a}$ 是一个可学习的参数向量。这个公式可以理解为，节点 i 与邻居节点 j 之间的注意力系数是由 i 和 j 的特征向量与参数向量 $\vec{a}$ 的加权和经过激活函数得到的，其中加权系数是通过 softmax_j 归一化得到的。

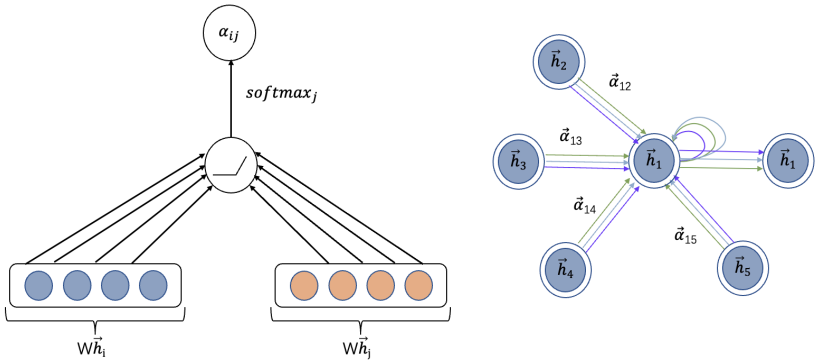


图 2.6　单头图注意力（左）和多头图注意力（右）

通过自注意力机制的引入，GAT 可以对图中的节点进行自适应的特征聚合，从而获得更具表征性的节点表示向量。GAT 模型具有较好的表现，在很多图结构数据上都取得了优秀的性能。

多头注意力机制是一种深度学习模型中的注意力机制，它在自然语言处理、图像识别等领域中广泛应用。多头注意力机制的基本思想是将输入的特征矩阵分成若干个头，每个头都有自己的注意力权重矩阵，通过对这些头的特征矩阵和对应的注意力权重矩阵进行加权和，得到一个最终的表示向量。这个最终的表示向量能够更好地捕捉输入中不同方面的信息。在自然语言处理领域中，多头注意力机制常常被用于序列建模任务，例如机器翻译、文本分类、问答系统等。在图像识别领域中，多头注意力机制可以用于图像中目标的定位和分类。多头注意力机制的优点是能够从多个不同的角度对输入进行建模，从而更好地理解输入中的信息。缺点是需要耗费更多的计算资源和时间。

多头注意力机制的计算过程如图2.6（右）所示。多头注意力系数的计算公式如下：

$$\vec{h}_i' = \delta\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\epsilon N_i}\alpha_{ij}^k W^k \vec{h}_j\right) \tag{2.4}$$

图像特征提取：ResNet[73] 残差神经网络是由何凯明等人在 2015 年提出的一种

深度卷积神经网络，其目的是解决深度神经网络训练过程中的梯度消失和模型退化问题。

传统的深度卷积神经网络通常由若干个卷积层和池化层组成，这些层的作用是将输入数据映射到一系列高级别的特征表示。这些特征表示在深层神经网络中逐渐抽象和复杂化，从而使网络能够对输入数据进行更准确的分类或预测。然而，当网络变得更深时，训练过程变得更加困难。在传统的深度神经网络中，梯度会逐渐消失，导致训练变得困难。此外，当网络变得更深时，其测试性能可能会出现退化，即增加网络的深度反而会导致测试性能下降。

ResNet 通过引入残差块（residual block）来解决这些问题。残差块是一个包含多个卷积层的模块，其结构与普通的卷积块类似，但还包含了一条从输入到输出的跨层连接（shortcut connection）。这种跨层连接允许网络学习残差函数，即通过将输入与输出进行相加来产生输出。通过这种方式，ResNet 可以更轻松地学习网络中的深层特征，并且可以避免梯度消失和模型退化问题。

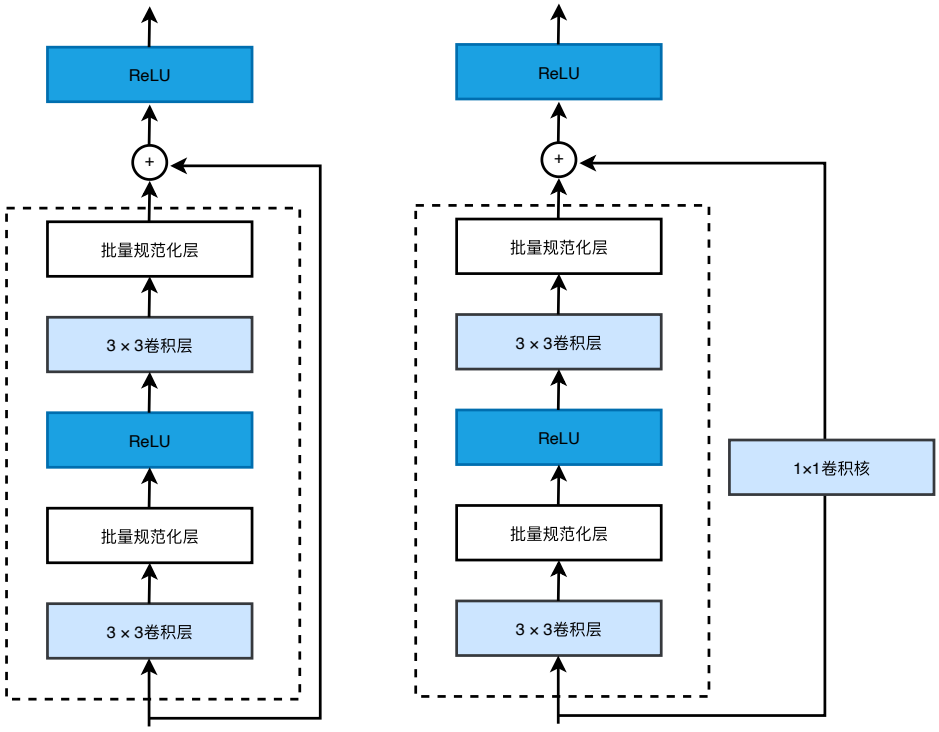


图 2.7 ResNet 单元结构

序列预测建模：循环神经网络（Recurrent Neural Network，RNN）[74] 是一种具有循环连接的神经网络，可以用于处理序列数据，例如自然语言文本、音频信号和

时间序列数据等。与传统神经网络不同的是，循环神经网络在网络中引入了一个隐状态（hidden state），该状态可以捕捉先前输入的信息，从而在处理序列数据时具有记忆能力。

循环神经网络的核心思想是在时间上共享参数，以便在每个时间步骤使用相同的网络结构来处理序列数据。具体来说，每个时间步骤的输入都会与上一个时间步骤的隐状态进行连接，这样可以将之前的信息传递到当前时间步骤。在每个时间步骤，循环神经网络会将输入和隐状态作为网络的输入，计算输出并更新隐状态。这种递归的过程可以一直进行下去，直到处理完整个序列。

但是循环神经网络也存在一些缺点，其中一些常见的缺点包括：（1）梯度消失和梯度爆炸：由于循环神经网络的循环连接，每个时间步骤的梯度会被反复地传递，导致梯度的累积效应，从而可能导致梯度消失或梯度爆炸的问题。（2）计算成本高：循环神经网络的计算成本相对较高，因为每个时间步骤都需要进行前向计算和反向传播，而且时间步骤的数量可能很大。（3）难以并行化：由于循环神经网络的循环连接，每个时间步骤必须按顺序进行计算，从而难以有效地并行化计算，限制了循环神经网络的计算效率。为了解决这些问题，研究人员提出了许多改进的循环神经网络模型，例如长短时记忆网络（LSTM），在一定程度上缓解了循环神经网络的缺点。

LSTM（Long Short-Term Memory）[75] 是一种循环神经网络（Recurrent Neural Network，简称 RNN）的变体，它可以有效地解决传统 RNN 中梯度消失（Vanishing Gradient）和梯度爆炸（Exploding Gradient）的问题，同时能够长期记忆之前输入的信息。LSTM 模型的核心思想是引入了一个称为“门”的机制，包括输入门、遗忘门和输出门，这些门会控制信息在 LSTM 模型中的流动，从而有效地处理长序列输入数据。

在 LSTM 模型中，每个时间步上都有三个关键的输入：输入数据、前一个时间步的输出，以及前一个时间步的记忆状态。这些输入被送入 LSTM 模型的三个门中，门通过权重来控制输入数据、上一时刻输出和记忆状态的加权和，从而产生一个新的输出和记忆状态，这些新的输出和记忆状态被传递到下一个时间步。LSTM 模型被广泛应用于自然语言处理、语音识别、图像分类等领域，是目前非常流行的深度学习模型之一。

条件随机场（Conditional Random Field，CRF）是一种概率图模型，常用于序列

标注、分割等任务。与隐马尔可夫模型（HMM）类似，CRF 也是一种针对序列数据的模型，但是相比 HMM 更为灵活，能够处理更加复杂的特征。

CRF 的基本思想是给定观测序列，求解最可能的状态序列。具体来说，CRF 假设状态序列与观测序列之间存在一个条件概率分布，即 $P(Y|X)$，其中 Y 表示状态序列，X 表示观测序列。CRF 模型的训练目标就是最大化给定观测序列下的条件概率分布，即求解 $\arg\max\big(P(Y|X)\big)$。

线性条件随机场（Linear Conditional Random Field，LCRF）是一种用于序列标注和结构化预测的概率模型。LCRF 是基于条件随机场（CRF）模型的改进，其最大的特点是假设特征函数满足线性组合的形式。

在 LCRF 中，对于给定的输入序列，需要对每个位置上的标签进行预测。为了实现这一目标，LCRF 将序列中每个位置的标签视为一个随机变量，并定义一个概率分布来表示每个标签的可能性。该分布由输入序列和当前标签的条件特征表示，通过线性组合计算得到。

LCRF 的训练通常采用极大似然估计法，即最大化训练集上所有样本的条件对数似然。具体地，可以采用梯度下降等优化算法来求解模型参数。在测试阶段，LCRF 可以通过动态规划算法来计算给定输入序列的最优标签序列，可以广泛应用于自然语言处理、计算机视觉、生物信息学等领域，如命名实体识别、词性标注、句法分析、图像分割等任务。

现有视觉富文档理解方法: [76] 提出了一个案例，说明基于规则的方法对行业从业者的重要性。[77] 采用为每个模板设计的特定配置规则。[78] 提供了一种基于模板匹配的算法来解决文档理解问题，为了应对不同的情况，需要构建和维护大量的模板。然而基于人工规则的模板匹配法需要为不同类型、不同布局结构的文档设计不同的规则，这需要研究人员花费大量时间和精力。基于规则字典的信息抽取方法需要手动整理规则，仅适用于特定的领域，通用性较差。即使在应用了规则的情况下，信息的多样性和格式的不统一也会导致很难设计出完备的抽取规则。目前的基于模板匹配的关键信息结构化抽取算法主要是基于位置坐标制作模板，需要进行多次模板对齐，然后再提取关键信息。这种方法需要大量前期投入，而且在文档格式复杂、类别繁多的情况下效果较差。此外，对于存在模糊或扭曲情况的文档或图像，基于位置的模板抽取系统表现不佳。尽管一些研究提出了基于文字流和分布式环境下的

表格信息抽取方法，但是它们仅适用于特定类型的表格和文档，并且需要预先制定特定的模板或关键词信息。因此，基于模板或文字流的表格信息抽取方法在面对不存在完整表格的表单时效果较差，且通用性较低。

随着深度学习的快速发展，许多基于深度学习的 VIE 方法已经出现，并且在准确性和能里方面都显著优于传统的基于规则和基于模板的方法。这些方法将 VIE 作为一个 token 分类问题，并使用不同的深度学习模型来预测每个文档 token 的字段类型。早期的工作 [79] 通常采用语言模型从纯文本中提取实体，这些基于 NLP 的方法通常操作于文本序列，并且不包含视觉和布局信息。后来的研究开始倾向于探索布局信息的融合，一系列方法将文档表示为带有文本标记的二维网格，以获得上下文嵌入。在此之后，一些研究者认识到了多模态融合的必要性，并通过整合视觉和布局信息来提高性能。这些方法主要包括应用图神经网络和 Transformer 来融合多模态的信息。根据不同的文档表示，大致可以分为三种模型：基于序列、基于图和基于网格。

基于图的方法：基于图的方法将每个文档建模为一个图，每个 bounding box 为图中的一个节点。每个节点的初始特征表示可以为只包含文本的单一模态特征，也可以是同时结合文本、视觉、布局等多种模态的特征。然后利用图神经网络或自注意力机制 [80] 更新图中每个节点的特征表四。之后，[81] 将最后一层的图节点嵌入输入到分类器进行类别判断。而其他一些方法 [4, 13, 82-86] 将每个节点嵌入到相应文本段中的所有标记嵌入，然后输入序列标记模型提取字段值。在 [82]、[4] 的模型中，视觉特征和非顺序信息的相对重要性存在争议，图神经网络对文档的建模具有提取实体任务的良好性能。但是 [82] 需要先验知识和广泛的人力努力来预定义特定任务的边缘类型和图的相邻矩阵。但是，在文档结构很复杂的时候，设计这些有效的边缘特征和图的邻接矩阵是比较困难耗时的。[4] 直接定义一个全连通图，然后使用一个自注意机制来定义全连通节点上的卷积。该方法可能会忽略节点的噪声，导致节点信息的无用性和冗余性的聚合。最近的基于图的方法，如 PICK[13]、TRIE[85] 和 VIES[86]，使用图卷积，充分利用文档的文本、图像、位置等特征，为 VIE 获得更丰富的表示，这已经在公共 SROIE 数据集上取得了优越的性能。

基于网格的方法：基于网格的方法，如 Chargrid[87]、U-Net[88]、CUTIE[5]、BERT-grid[89] 将每个文档表示为 token 嵌入的二维网格，然后使用标准实例分割模型从二

维网格中提取字段值。这些网格虽然使用了文本和布局特征，但是却忽略了视觉特征。因此，VisualWordGrid[90] 将这些网格表示与文档图像特征相结合，产生了更好的的多模态二维文档表示，可以同时结合文档的文本、视觉和布局信息。然而，这些基于网格的方法无法实现与 LayoutLM、PICK、TRIE 和 VIES 等最先进的方法相当的性能，原因有两个：1) 这些模型和方法没有利用最先进的上下文词嵌入技术来提取足够强的标记嵌入；2) 虽然 BERTgrid 将 BERT 合并到网格表示中，但预训练的 BERT 参数在模型训练过程中是固定的，因此没有充分利用基于 BERT 的标记嵌入的潜力。最近有工作提出 ViBERTgrid[91]，改进了这两个问题，同时将文档的文本、布局和视觉信息编码到二维特征映射中，与 BERTgrid 不同的是，ViBERTgrid 网络中的 BERT 和 CNN 参数是联合训练的，在很多数据集上取得了显著的提升。从多模态融合的发展过程可以看出，多模态的丰富信息从单一模态逐步扩展到多种模态，并不断丰富各种模态的表示。

2.5 版式感知文档理解模型

注意力机制: 注意力机制受到了人类注意力的启发，就像我们在阅读时会将注意力集中在重要的信息上一样。在模型的训练过程中，输入的不同部分具有不同的权重，注意力机制能够自动学习到这些权重，并在推理过程中使用它们来加权计算，以提高模型的性能。最开始 attention 机制在 CV 领域被提出来，但后面广泛应用在 NLP 领域。

编码器-解码器框架（Encoder-Decoder Framework）是一种常见的神经网络结构，被广泛应用于自然语言处理、图像处理、语音识别等领域。在这种框架中，输入数据首先经过编码器（Encoder）进行编码，将输入转换成一个高维向量表示。编码器通常使用卷积神经网络（CNN）或者循环神经网络（RNN）等结构来实现。接着，解码器（Decoder）接收编码器的输出向量作为输入，并在输出端生成一个与原始输入相同维度的输出序列。在自然语言处理任务中，解码器通常是一个 RNN，它通过逐步生成一个词序列来完成翻译、摘要等任务。

首先，该框架对输入数据进行编码时，会将所有输入信息都压缩到一个固定长度的向量中，这可能导致信息的损失 [92] 和模糊。这种情况在处理长文本和长序列数据时尤为明显，因为很难将所有信息都表示在一个向量中。其次，编码器-解码器

框架在处理复杂的任务时，存在梯度消失和梯度爆炸的问题。这是因为在训练过程中，误差会通过解码器向编码器传递，这可能导致深度神经网络中的梯度无法正确传递，影响了模型的训练效果。最后，编码器-解码器框架的推理速度通常比较慢，因为在解码器生成每一个输出时，都需要进行一次前向计算，这会大大增加模型的计算时间和复杂度。

Transformer: Transformer[93] 模型是一种用于自然语言处理 (NLP) 和其他序列到序列 (sequence-to-sequence) 任务的深度学习模型，由 Google 在 2017 年提出。它的主要特点是使用了自注意力机制 (self-attention mechanism) 来进行编码和解码，而不是使用传统的循环神经网络 (RNN) 或卷积神经网络 (CNN)。

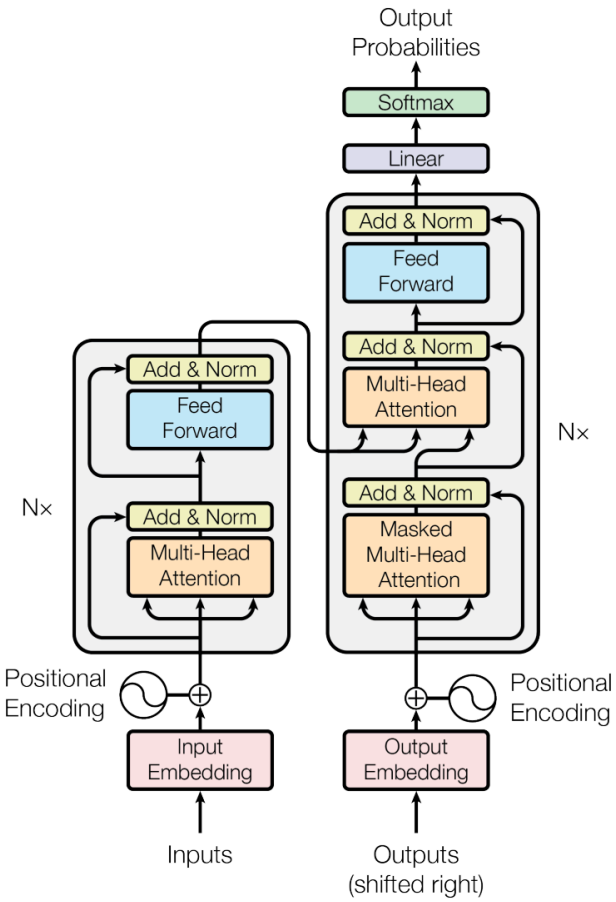


图 2.8 Transformer 结构示意图 [93]

自注意力机制允许 Transformer 模型在每个时间步中对输入的整个序列进行编码，而不仅仅是单个时间步。这使得 Transformer 模型能够同时考虑输入序列中的所有信息，而不像 RNN 需要通过循环迭代逐步处理序列。此外，自注意力机制也能够

为每个输入位置分配不同的权重，以捕捉不同输入位置的重要性

Transformer 模型通常由一个编码器和一个解码器组成，其中编码器和解码器都包含多层自注意力机制和前馈神经网络。在编码器中，输入序列经过多层自注意力机制和前馈神经网络，生成一组高维向量表示；在解码器中，目标序列先经过一个自注意力机制，然后与编码器的输出结合起来，再经过多层自注意力机制和前馈神经网络生成最终的输出序列。

Transformer 模型的整体结构如图2.8所示，由 encoder-decoder 两部分组成，左边是 encoder 部分，右边是 decoder 部分，下图所展示的只是“一层”结构，实际上，其内部可以并行的包含多个这样的结构。

基于序列的方法需要首先将文档图像序列化为一维文本序列，然后使用 NLP 中现有的序列标注模型（例如 [72, 94-97]）来提取字段值。为了减少序列化的影响，早期的方法 [98-100] 试图把二维位置信息编码标记嵌入，但这仍然依赖于文本段序列化的准确性，所以很难应用于复杂布局的文档序列化，而这个步骤又是十分重要的。虽然目前 OCR 减轻了处理图像的负担，但理解不同布局上的文本块之间的语义关系仍然是一个具有挑战性的问题。为了解决这个问题，现有的方法使用了一个预训练语言模型来利用其有效的文本表示。$Post-OCR$ 通过将 VIE 任务作为序列标记问题来微调 BERT。BERTgrid[89] 在其图像分割任务中使用 BERT 将文本信息合并到图像像素中。然而，由于 BERT 是为文本序列设计的，它们人为地将分布在二维空间中的文本块转换为单一的文本序列，从而丢失空间布局信息。

LayoutLM：最近，微软亚洲研究院 [6] 提出了利用文本块的空间信息在大规模文档数据上进行预先训练的 LayoutLM。他们在视觉夫文档理解的几个下游任务上实现较好的性能，显示了预训练方法的有效性。LayoutLM 是微软团队在 2020 年提出的一种自然语言处理模型，主要应用于文档识别、文档理解和表格识别等领域。LayoutLM 模型基于预训练的 transformer 架构，并在此基础上加入了视觉布局信息，能够同时处理文本和图像信息，模型架构如图2.9所示。

LayoutLM 模型通过将文本和视觉布局信息相结合，实现了对多种文档类型的识别和理解，包括 PDF、Microsoft Word 文档和 PPT 演示文稿等。在训练阶段，LayoutLM 使用大量的文档图像和对应的标注数据进行预训练，以学习文本和视觉信息的联合表示。在应用阶段，LayoutLM 模型可以对输入的文档图像进行分析和理解，提取出

其中的文本和表格信息，并进行分类、提取等多种任务。

相比于传统的文本识别模型，LayoutLM 模型不仅考虑了文本本身的信息，还考虑了文本在页面布局中的位置和结构等信息，这些信息能够帮助模型更准确地识别和理解文档中的内容。LayoutLM 模型已经在各种文档理解任务中得到了广泛应用，包括文本识别、表格识别、布局分析等。

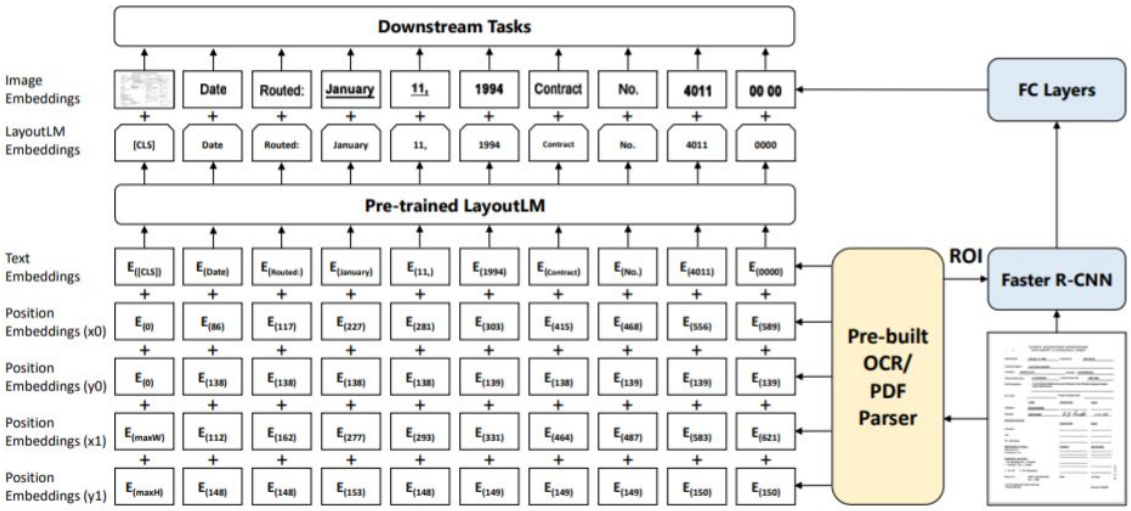


图 2.9　LayoutLM 模型结构示意图 [6]

LayoutLM 在 BERT 模型结构基础上，新增了两个输入特征：2-D 位置特征与图像特征。

1. **2-D 位置特征**：2-D 位置特征的目的在于编码文档中的相对空间位置关系。一个文档可视为一个坐标系统，其左上角即为坐标原点 $(0,0),(0,0)$。对于一个单词，其包围盒能够以坐标 (x_0,y_0,x_1,y_1) 进行表示，其中 (x_0,y_0) 表示左上角坐标，(x_1,y_1) 表示右下角坐标。x_0 与 x_1 共享嵌入层参数 X，y_0 与 y_1 共享嵌入层参数 Y。特别地，整个文档图像的包围盒为 $(0,0,W,H)$，W 与 H 分别表示文档图像的宽与高。
2. **图像特征**：根据文本框的坐标，LayoutLM 利用 ROI 操作从 Faster R-CNN 的输出特征图中生成图像区域特征，与单词一一对应。对于特殊的 $[CLS]$ 标记（$[CLS]$ 标记的输出接分类层，用于文档分类任务，详情可见 BERT 模型），则采用整图的平均特征作为该标记的图像特征。应该注意的是，LayoutLM 在预训练阶段并没有采用图像特征；图像特征仅在下游任务阶段可以选择性地加入，而生

成图像特征的 Faster R-CNN 模型权重来自于预训练模型且不作调整。

LayoutLM 预训练采用了两种预训练任务，通过这两种预训练任务的联合训练，LayoutLM 模型可以从大规模的未标注文档数据中学习到丰富的文档布局和对象表示，为后续的文档布局分析和信息提取任务提供了有力的基础，这两种任务分别是：

- **Task 1：掩码视觉语言模型 MVLM**。在预训练阶段，随机掩盖掉一些单词的文本信息，但仍保留其位置信息，然后训练模型根据语境去预测被掩盖掉的单词。通过该任务，模型能够学会理解上下文语境并利用 2-D 位置信息，从而连接视觉与语言这两个模态。
- **Task 2：多标签文档分类 MDC**。文档理解的许多任务需要文档级别的表征。由于 IIT-CDIP 数据中的每个文档图像都包含多个标签，LayoutLM 利用这些标签进行有监督的文档分类任务，以令 [CLS] 标记输出更为有效的文档级别的表征。但是，对于更大规模的数据集，这些标签并非总可获取，因此该任务仅为可选项，并且实际上在后续的 LayoutLMv2 中被舍弃。

尽管 LayoutlM 取得了成功，但它有三个限制，首先，LayoutlM 在 BERT 的位置嵌入基础之上加入 x 轴和 y 轴的位置信息，但忽略了序列中位置与二维空间之间的区别。其次，它的预训练方法与 BERT 基本相同，而 BERT 没有明确地考虑文本块之间的空间关系。最后，在其下游任务中，LayoutlM 只做了需要对文本块进行序列化的序列标注任务 (例如，BIO 标记)。这些限制表明，LayoutlM 不仅不能充分利用空间信息，而且不能解决实际文本块序列化困难的情况下的 VIE 问题。Pramanik 等人 [101] 和 LayoutLMv2[7] 进一步扩展了布局管理的思想，通过整合预训练阶段的图像信息，学习更强的模态文档表示。

LayoutLMv2：LayoutLMv2 模型是一种多模态预训练模型，结合了文本、图像和布局信息。相较于 LayoutLM 模型，LayoutLMv2 模型的主要改进在于引入了图像信息，并且使用了空间感知自注意力机制。在输入阶段，LayoutLMv2 模型会同时考虑文本和图像信息。模型将输入的文本和图像一起编码，以建立文本、图像和布局信息之间的关系。具体来说，模型使用了一个基于 ResNet50 的图像编码器来提取图像特征，然后使用一个双向 Transformer 编码器来对文本和图像进行编码。这个编码器包含多个 Transformer 层，每个层都包含多头自注意力机制和前向神经网络。在空间感知自注意力机制中，LayoutLMv2 模型通过引入二维的相对位置编码来表示文本

块之间的相对位置。这种方法类似于纯文本预训练模型中使用的一维相对位置编码，但是在 LayoutLMv2 模型中，相对位置编码是在二维空间中表示的。这样可以更好地捕捉文本块之间的相对位置关系，有助于模型更准确地理解文档的结构和布局。

此外，LayoutLMv2 模型还使用了两种新的预训练任务——“文本—图像对齐”和“文本—图像匹配”。这些任务旨在让模型学习如何将文本和图像进行对齐和匹配。通过这些任务的训练，模型可以更好地理解文本和图像之间的关系，并提高文本识别和布局分析的准确性。

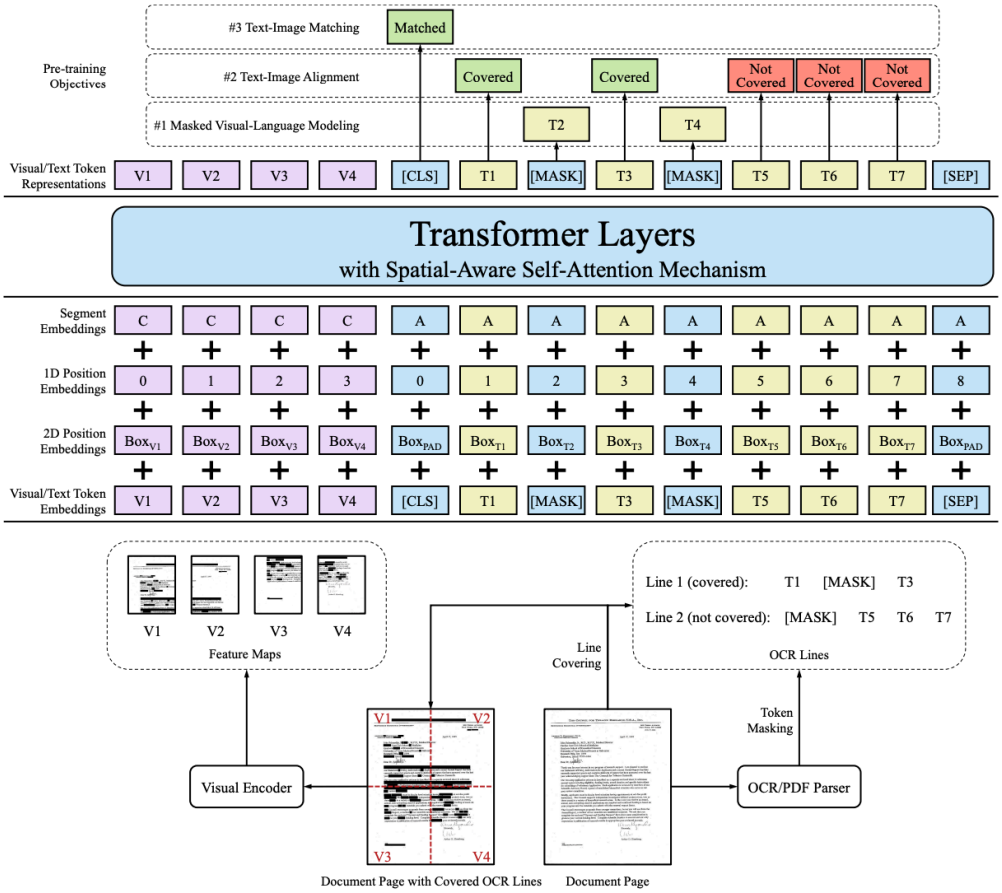


图 2.10　LayoutLMv2 模型结构示意图 [7]

如图2.10所示，模型将文本、图像、布局三种模态的输入转换成向量表示，然后再交给编码器网络，最终输出的表示向量可以供下游任务使用。

LayoutXLM：大多数跨语言的预训练模型是基于纯文本数据的，例如 mBERT、XLM、XLM-RoBERTa、InfoXLM 和 mT5 等。因此不能直接应用于文档理解任务，因为文档中包含的元素比单纯的文本更加丰富。例如，文档可以包含图像、表格、

图表等等，这些元素需要结合上下文才能进行理解。针对这一问题，微软亚洲研究院的研究员提出了一种基于多语言文档理解任务的多模态预训练模型 LayoutXLM。LayoutXLM 模型架构如图2.11所示。LayoutXLM 是 LayoutLMv2 模型的多语言扩展

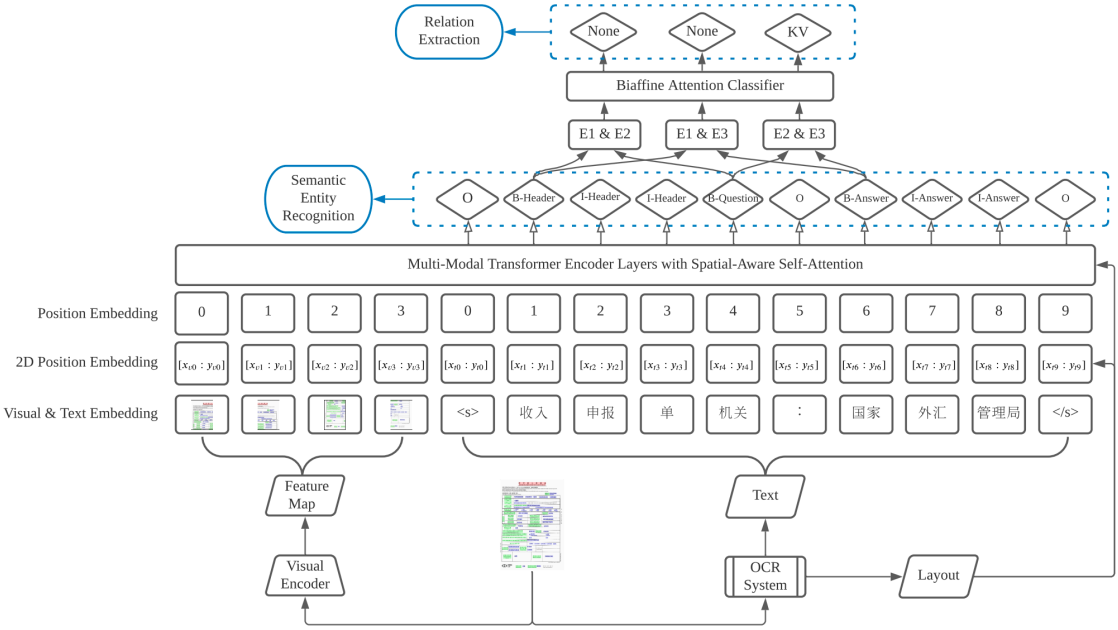


图 2.11　LayoutXLM 模型结构示意图 [8]

版，采用与之相同的 Transformer 架构进行多模态预训练，并使用多语言预训练的 InfoXLM 模型进行初始化。该模型能够处理包含多种元素的文档，例如文本、图像、表格和图表等。通过将多个模态的输入元素进行编码，LayoutXLM 能够学习到这些元素之间的关系，并为文档理解任务提供有用的信息。研究员们选择了多语言的 PDF 文件作为训练数据来源。因为，PDF 文件可以大大方便收集和预处理的步骤：不同于扫描文档图片，PDF 文件可以免去在自然图像中筛选扫描文件的麻烦；另一方面，也可以用 PDF 解析器直接提取准确的文本和相应的布局信息，从而节省运行 OCR 工具的时间。针对键值提取这一表单理解中最关键的任务之一，与 FUNSD 类似，研究员们将这项任务定义为两个子任务，即语义实体识别和关系提取。

LayoutLMv2 的预训练目标在对视觉丰富的文档进行建模时已显示出其有效性。因此，研究员们很自然地将这个预训练框架调整为多语言文档预训练。按照跨模态对齐这一核心思路，LayoutXLM 文档理解预训练框架包含三个预训练目标——多语言掩码式视觉语言模型（文本-布局对齐）、文本-图像对齐（细粒度的文本-图像对齐）和

文本-图像匹配（粗粒度的文本-图像对齐）。多语言掩码式视觉语言模型（Multilingual Masked Visual-Language Model）：掩码式视觉语言模型（MVLM）最初是在 LayoutLM 中提出的，在 LayoutLMv2 中也有使用，其目的是对视觉丰富的文档中的丰富文本进行建模。在这个预训练目标中，模型需要根据其文本上下文和所有的布局暗示来预测被遮盖的文本。与 LayoutLM/LayoutLMv2 类似，研究员们用多语言掩码式视觉语言模型（MMVLM）来训练了 LayoutXLM。

- **文本—图像对齐（Text-Image Alignment）：**文本-图像对齐（TIA）任务是为了帮助模型捕捉文本和图像之间的细粒度对齐关系。研究员们随机选择了一些文本行，然后在文档图像上覆盖其对应的图像区域，模型需要对每个文本预测其是否被覆盖。
- **文本—图像匹配（Text-Image Match）：**对于文本-图像匹配（TIM）任务，研究员们的设计目标是使文本和图像之间的高层语义表示保持一致。为此，需要要求模型预测文本和图像是否来自同一个文档页面。

LayoutLMv3: 为了克服文本和图像在预训练目标上的差异，促进多模态表征学习，微软亚洲研究院的研究员们提出了 [102]，以统一的文本和图像掩码建模目标来预训练多模态模型，即 LayoutLMv3 学习重建语言模态的遮盖词 ID，并对称地重建图像模态的遮盖图像块 ID。在模型架构设计上，LayoutLMv3 不依赖复杂的 CNN 或 Faster R-CNN 网络来表征图像，而是直接利用文档图像的图像块，从而大大节省了参数并避免了复杂的文档预处理（如人工标注目标区域框和文档目标检测）。简单的统一架构和训练目标使 LayoutLMv3 成为通用的预训练模型，可适用于以文本为中心和以图像为中心的文档 AI 任务。

Lilt: 大部分的预训练模型和下游任务都必须构建在同样的语言上（一般是英文），目前存在的一个难点是如何能够在大量的英文数据集上进行预训练，然后训练出来的模型在跨语言数据上仍然能够用于微调。为了解决这一问题，[103] 提出了 Lilt 模型，一种简单而有效的独立于语言的布局转换器，用于理解结构化文档，可以在单一语言的结构化文档上进行预训练，然后使用相应的现成单语言或多语言预训练文本模型直接在其他语言上进行微调。

第三章　基于双流排版图网络的实体识别方法

3.1　研究动机

现有的基于 Transformer 架构的文档预训练模型虽然在文档信息抽取上取得了不错的效果，但是这些模型 [6-8] 主要关注文档中细粒度 token 元素，如单词和细粒度的文档图像块，使得它们很难从粗粒度的元素中学习，包括像短语这样的自然词汇单位和突出的图像区域。如图3.1所示，模型缺少对 boundign box 粒度的键值信息的建模和文档中粗粒度的元素的学习，导致抽取结果出错。

视觉富文档在经过 OCR 处理之后会得到一组文本行或文本块，每一个文本块称为一个 bounding box，如图3.4（右）所示，这些文本块包含待提取的关键文本信息以及所对应的坐标，当然并不是每个 bounding box 都包含有用的信息，有的文本块仅仅包含无用的冗余文本信息，而对于整个 VIE 提取系统来说，每个视觉富文档所对应的文本块的个数是不确定的。由于图神经网路有较好的灵活性，可以为每个 bounding box 建模为图中的一个节点，这样即使每个文档中的 bounding box 个数不同，也能采用统一的形式对文档进行建模。

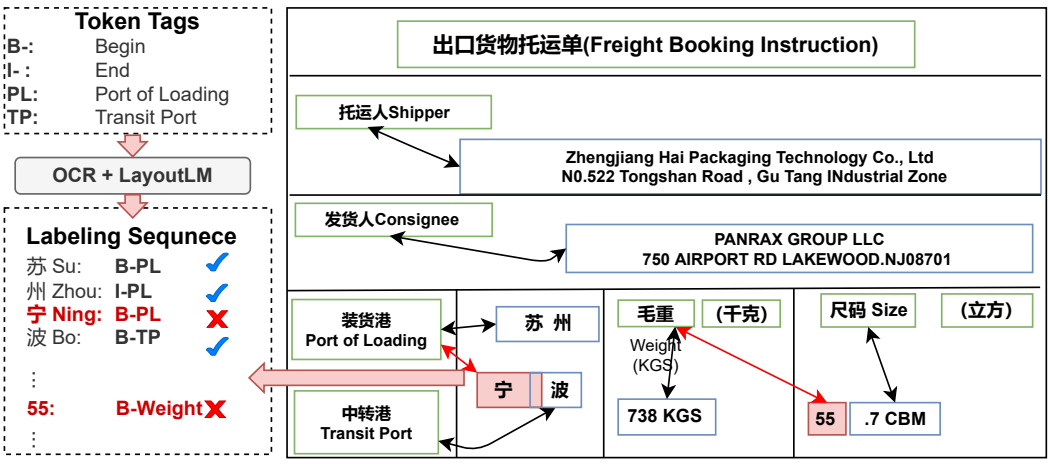


图 3.1　由于基于 Transformer 架构的文档预训练模型缺少对键值信息的建模以及缺少文档中粗粒度的元素的学习，导致抽取结果出错

本章提出一种基于双流排版图神经网络视觉富文档关键信息抽取方法，同时关注文档中粗细两种粒度的元素信息，并采用有效的方式经两种粒度的信息融合。本

章先概括了视觉富文档信息抽取方法的发展路线，然后分析目前方法中存在的缺陷，并针对这些缺点进行了针对性的改进。最后通过在多个数据集上的实验以及消融实验验证了所提出的模型和方法的有效性。

3.2 相关工作

针对视觉富文档视觉信息抽取任务 (VIE) 的典型方法如图3.2所示。传统方法使用手工特征（例如正则表达式和模板匹配）来提取关键信息，如图3.2(a) 所示。然而，这种解决方案 [78, 104] 只使用文本和位置信息来提取实体，需要大量的特定知识和人为设计的规则，而且不具有可扩展性。大多数现代方法认为 VIE 是一个序列标签问题，由 NER 解决，如图3.2(b) 所示。与典型的 NER 任务相比，要让机器从复杂文档中毫无歧义地判断出文本块类别，难度要大得多。其中一个主要原因是，这样的框架只在纯文本上运行，而没有结合文档的视觉信息和布局信息以获得更丰富的表示。比如海运单数据的发货人和收货人这两个关键信息，若忽略了 bounding box 的位置信息以及上下文语义信息，仅从文本内容上考虑的话，是无法准确判断出实体类别的，因为这两种实体在文本内容和语义上非常相似。因此模型不能只使用文本特征，需要同时考虑文本特征，图像特征与布局特征等。

替代方法 [4, 13, 82] 通过使用图卷积运算预先定义一个图来组合文本和视觉信息，如图3.2(c) 所示。在文献 [4, 82] 中，对视觉特征和非序列信息的相对重要性进行了讨论，基于文档的图神经网络建模在提取文档中的关键信息具有良好的性能。但 [82] 需要引入先验知识预先确定任务图中边的具体类型和图的邻接矩阵。然而，设计图的边的类型和邻接矩阵是具有挑战性的、主观的，耗时的，尤其是当文档结构不明确时。[4] 直接定义一个全连通图，然后使用自注意力机制来定义全连通图上的卷积操作。这种方法可能忽略了节点学习到的额外噪声，并导致节点聚集过多的无用和冗余的信息。[13] 提出一种动态更新图中邻接矩的值的方法，该方法的虽然不用人工设计图的边的类型和邻接矩阵，但是模型仍然只关注文档中的细粒度元素，缺少对粗粒度元素的学习，如图3.2(d) 所示。

最近，VIE 任务中的一些研究试图充分利用复杂文档中未开发的特征。[6] 受 BERT[72] 启发，提出了 LayoutLM 方法，在大规模的文档上进行预训练。尽管该方法使用文本、图像和位置特征来预训练模型，并在文档图像理解的下游任务（如 KIE）

上表现良好，但它没有考虑两个文本框之间的潜在关系，而且不能学习文档中的粗粒度的元素信息。此外，该模型需要足够的数据和时间来有效地预训练模型。

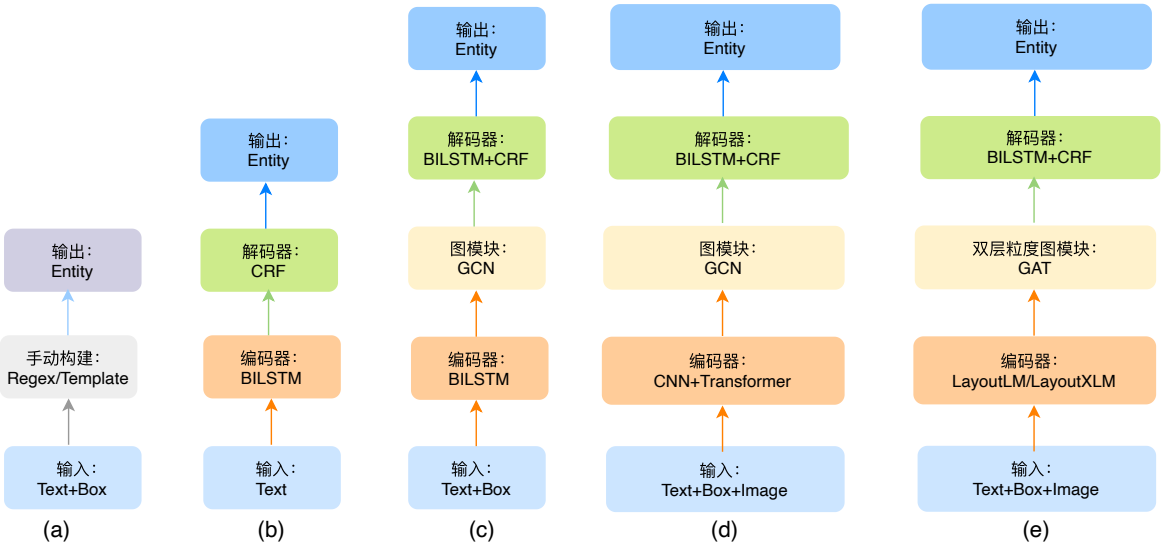


图 3.2　典型的架构和我们的关键信息提取方法：(a) 基于手工制作特征的方法；(b) 基于自动提取特征的方法；(c) 使用更丰富的特征的方法；(d) 使用单一粒度的文档元素；(e) 我们提出的模型

3.3　任务定义

每个扫描的视觉丰富的文件都是由语义实体列表组成的，定义为：

$$D = \{[b_1, ..., b_n], [l_1, ..., l_n], [(b_1, b_{h1}), ..., (b_m, b_{hm})]\}$$

其中，每个实体由一组词和边界框的坐标组成，定义为：$b_i = \{[w_i^1, ..., w_i^m], [x_i^1, x_i^2, y_i^1, y_i^2]\}$，其中 $[w_i^1, ..., w_i^m]$ 是第 i 个 bounding box 中的单词，$[x_i^1, x_i^2, y_i^1, y_i^2]$ 是第 i 个 bounding box 中的对角上的坐标。语义实体识别任务要求模型标记抽取出所有定义的语义实体并且将其分类到正确的类别中。

3.4　提出模型

3.4.1　整体框架

为了解决基于全连接图神经网络构建文档出现的节点学习到的过多的冗余信息，以及目前主流的基于 Transformer 架构的文档预训练模型中缺少对文档中键值对信息

的建模和难以学习文档中粗粒度元素信息的问题，本文提出基于双流排版图网络的实体识别方法，模型整体框架如图3.3所示。

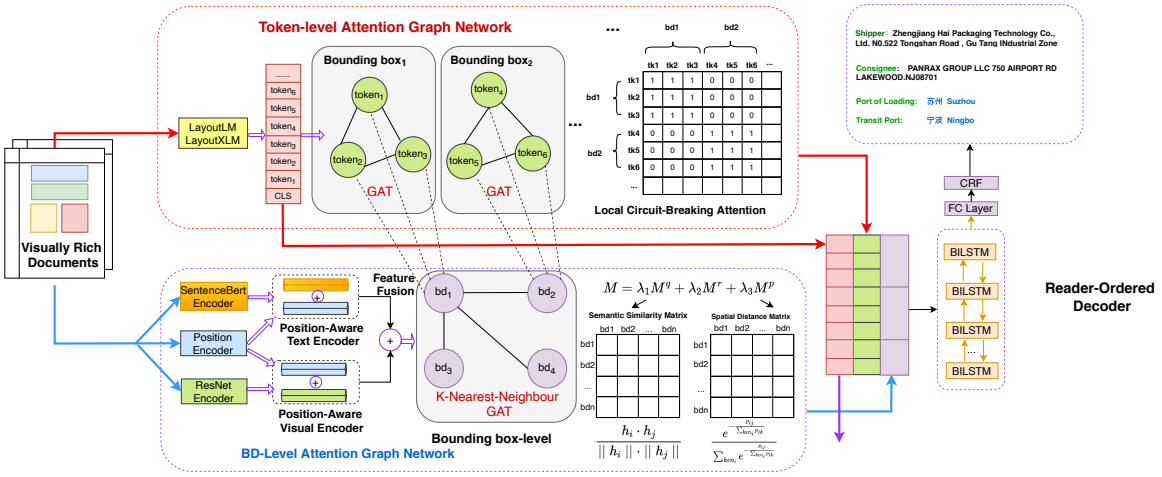


图 3.3　双流排版图网络，包含 token 与 bounding box 两个粒度的图网络的双流结构

每个视觉富文档图像经过文本检测与识别 OCR 系统后会得到一组文本行或文本块 (称为 bounding box)，每个 bounding box 包含其在图片中的相对坐标信息与其文本内容。基于图卷积神经网络的信息抽取方法将每个 bounding box 构成一个图中的一个结点，为了避免构建全连接图 [4]，每个节点学习到过多的噪音，我们构建一个 k 近邻的图神经网络，即每个 bounding box 只与其最近的上方，下方，左方，右方，总共四个节点相连接。基于 boudnging box 粒度的图神经网络首先使用 SentenceBert[105] 提取文本特征，同时使用 ResNet 提出视觉特征，这两种多模态特征再与 bounding box 的坐标特征 (每个 bounding box 的 width（宽度）和高度（height）) 嵌入相加得到包含三种模态的特征作为图神经网络节点的初始节点表示。

为了促进一个 bouding box 内部所有 token 之间的信息交互，我们提出使用 LayoutLM 编码文档，获得 token 粒度的多模态表示，然后为每一个 bounding box 内部的 token 构建全连接的图神经网络。

以上两个分支再经过图卷积或者图注意力操作后，再进行两个粒度的融合。具体而言，对第 i 个 bounding box，其内部包含的 token 为：$\{token_{i1}, token_{i2}, \ldots, token_{in}\}$，将每个 token 粒度的特征 token_{ik}，和第 i 个 bouding box 的特征相拼接，使用双向 LSTM（BiLSTM）以及 CRF [106] 模型对其进行解码，对 CRF 求解最佳路径后即可得到文本实体提取结果，即可完成对关键信息的结构化识别。整体模型架构如图**??**所示。

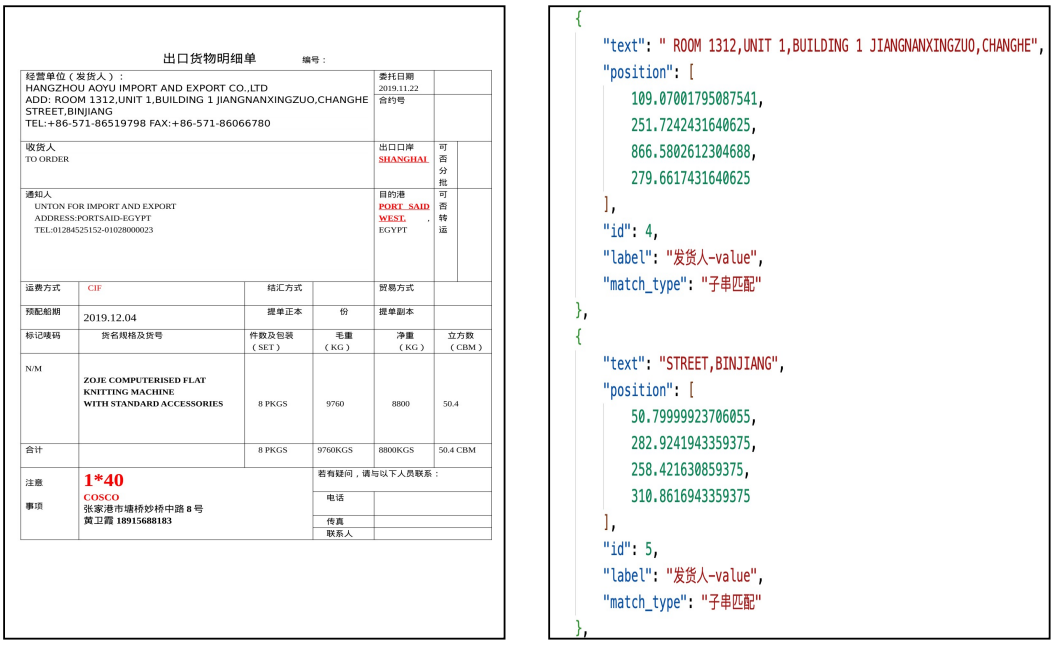


图 3.4　模型的输入，包括图像（左）和 word group（右）

3.4.2　Token 粒度的图神经网络流

本文提出的模型中 token 粒度的图神经网络包括两部分：编码器和图注意力网络层。接下来对着两部分内容分别做介绍。

统一多模态编码器：我们使用 LayoutXLM[8] 模型作为骨干网络提取 token 粒度的嵌入表示。得到的 token 粒度的表示，包含语义，布局以及图形特征。文档 D 的编码如下: $C_{0:n} = LayoutXLM((b_1||....||b_n)$，$||$ 表示连接操作，将 n 个 bounding box 的文本连接在一起进行编码，C_0 表示 Cls 位置的编码。得到整个文档的 token 粒度表示后，按照第 i 个 bounding box 包含的 token 数和其在标注文件中所处的索引位置，进行分割，得到第 i 个 bounding box 中的 token 的特征表示为:

$$C_i = C_{1:n}[start : start + l_i], 0 \leq i \leq n \tag{3.1}$$

局部熔断图注意力网络：如图3.5所示，为了表示 bounding box 内部的相关性，促进一个 bounding box 内部的 token 之间的信息交互。我们为每一个 bounding box 构建一个完全连通的图，这意味着其每个 token 节点之间都是相互连接的。而 bounding box 与 bounding box 之间是不连通的。对于每一个 bounding box，给定一个图 $G = (V, E)$，$v_i \epsilon V$，$i = (1, 2 \ldots n)$ 表示 bounding box 中的一个 token 节点，$e_{ij} \epsilon E$ 表示 token 之间的边。

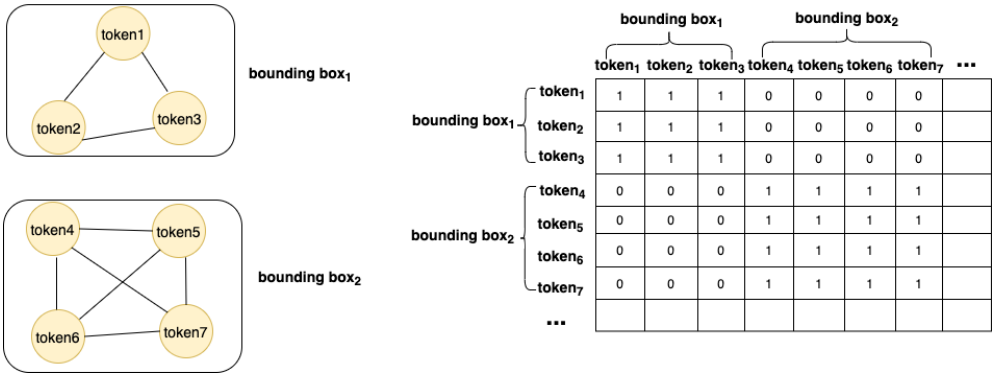


图 3.5 bounding box 内部熔断注意矩阵的构造

$$e_{jk} = \begin{cases} 1, & \text{if } v_j \in bd_i \quad \text{and} \quad v_k \in bd_i \\ 0, & \text{if } v_j \notin bd_i \quad \text{or} \quad v_k \notin bd_i \end{cases} \tag{3.2}$$

如公式3.2所示和图3.5所示，对于一个 bounding box 内部的 token 之间，有边相连接，邻接矩阵中对应位置的值设为 1，对与不同 bounding box 的 token 之间没有边相连接，邻接矩阵中对应位置的值设为 0。初始的 token 节点特征是通过 LayoutLM 得到的。

$$h_j{}^{(t+1)} = \delta\left(\sum_{\tilde{k}=1}^{N} \alpha_{jk} W h_k^{(t)} e_{jk}\right) \tag{3.3}$$

图注意力网络的计算方式如公式3.3和公式3.4所示，给定一个节点 v_j 及其隐藏层表示 $h_j^{(t+1)}$，通过自注意机制计算节点 v_j 的输出嵌入表示，其中 $h_j^{(t+1)}$ 是由 $h_j^{(t)}$ 聚合更新得到的。$h_k^{(t)}$ 是节点 v 的所有的邻接节点的隐藏层表示。δ 是激活函数，α_{jk} 是节点 k 和节点 j 之间的注意力系数。α_{jk} 的计算方式如公式3.3所示，其中 W 和 V 是可训练参数。我们应用非线性的 LeakyReLU 来避免“dyingReLU”问题。

$$\alpha_{j\tilde{j}} = \frac{exp\left(\delta\left(V^T[Wh_j \oplus Wh_k]\right)\right)}{\sum_{n \in N} exp\left(\delta\left(V^T[Wh_j \oplus Wh_n]\right)\right)} \tag{3.4}$$

与 [93] 类似，我们也使用多头注意力来提高模型的性能。K 个注意机制独立执行，并将它们的特征相加求均值。最终表示如公式3.5所示。

$$\vec{h}_i' = \delta\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \epsilon N_i} \alpha_{ij}^k W^k \vec{h}_j\right) \tag{3.5}$$

3.4.3 Bounding box 粒度的图神经网络流

文本编码与位置编码：由于文档中的文本内容以 2D 结构呈现，因此有必要使用布局信息对文本进行编码。按照 LayoutLM[6] 的做法，我们将所有坐标标准化并离散为 [0,1000] 范围内的整数，并使用两个嵌入层分别嵌入 x 轴特征和 y 轴特征。给定标准化的第 i 个文本标记的边界框 $box_i = (x_{i0}, x_{i1}, y_{i0}, y_{i1}, w_i, h_i)$，其中 w_i 是第 i 个文本框的水平长度，h_i 是垂直长度。然后，把通过两个布局嵌入层得到的六个边界框特征 $(x_{i0}, x_{i1}, y_{i0}, y_{i1}, w_i, h_i)$ 进行拼接，构建最终 2D 布局嵌入 p_i:

$$p_i = [Embx(x_{i0}, x_{i1}, w_i)]||Emby(y_{i0}, y_{i2}, h_i)], 1 \leq i \leq n \tag{3.6}$$

其中 || 表示拼接操作。$Embx$ 和 $Emby$ 是两个嵌入层。如图所示，我们使用预先训练的 Sentence-BERT 模型 [105] 将语义区域中包含的纯文本嵌入到特征向量中，该模型可以派生出语义丰富的句子嵌入表示。最终的句子嵌入量计算如下:

$$s_i = SentenceEmb(t_i) + Proj(P_i), 1 \leq i \leq n \tag{3.7}$$

其中 $SentenceEmb$ 表示 Sentence-Bert, $Proj$ 表示线性层。

粗粒度视觉信息编码：给定文档图像 I，缩放到 224×224 大小，使用 ResNet 作为视觉编码器的主干网络，从整个图像中提取特征。然后，使用 ROI Align 根据 bounding box 的坐标获得关注的区域。第 i 个 bounging box 的视觉特征为:

$$v_i = ROIAlign(ConvNet(X)) + Proj(p_i), 1 \leq i \leq n \tag{3.8}$$

其中 $Proj$ 表示线性层，P_i 在文本编码阶段得到的 bounding box 位置嵌入。

模态融合层: 在得到文本，位置，以及视觉特征的编码后，最后输入到图神经网络节点中的特征为:

$$h_i = s_i + p_i, 1 \leq i \leq n \tag{3.9}$$

最近邻键值图关联网络：我们采用如图3.6的方式，对视觉富文档的 bounding box 级别的键值关联进行建模。不同于以往的工作 [4] 构建全连接的图神经网络，我们设计了 K 近邻图注意力网络，通过遵循自注意力机制，每个节点只关注其周围最近的四个节点，尤其是 Key 和 Value 键值节点之间的信息交互，尽量避免节点学习到过

多的噪声。为了更好的表示节点特征，基于物理位置和节点相似度两个层面的考虑，我们构造了 bounding box 之间基于欧式距离的邻接矩阵 M^p 和基于 bounding box 特征相似度的邻接矩阵 M^q。因为我们发现键值对在文档中的距离较近，而且当一个键的值对应多个 bounding box 时，属于同一个键值对的高维特征的余弦相似度大于不同键值对之间的特征相似度。而且 M^q 中的元素值随着图神经网络的学习不断更新。给定节点 i, $j \epsilon N_i$ 是 i 的邻居节点。

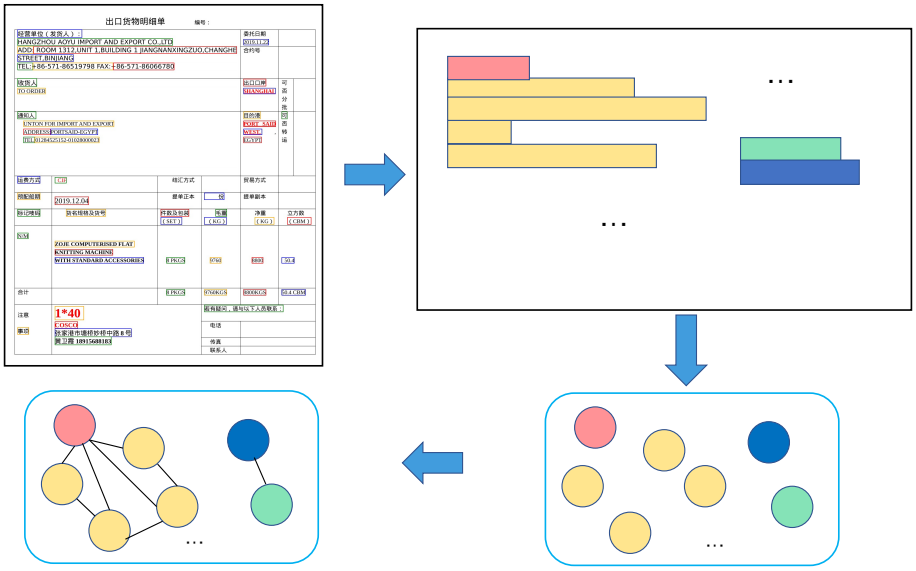


图 3.6 基于图结构的文档建模方法, 不同颜色的条形框代表不同的实体类别，每个文档框抽象为图中的节点，构建 K 近邻图网络。

M^p_{ij} 表示节点 i 和节点 j 之间经过归一化后的的欧式距离。

$$\mathbf{M}^p_{i\tilde{i}} = \frac{p_{i\tilde{i}}}{\sum_{m \in M} p_{im}} \quad \text{w.r.t.} \quad p_{i\tilde{i}} = \frac{exp(-p_{i\tilde{i}})}{\sum_{m \in M} exp(-p_{im})} \tag{3.10}$$

$$\mathbf{M}^q_{i\tilde{i}} = \frac{\tilde{h}_i \cdot \tilde{h}_{\tilde{i}}}{\left\|\tilde{h}_i\right\| \cdot \left\|\tilde{h}_{\tilde{i}}\right\|} \tag{3.11}$$

给定一个节点 v_i 和特征融合后的编码器表示 $\tilde{h}_i$，每层的输出可以表示为:

$$h_i^{(t+1)} = \delta(\sum_{k \in N_i} \gamma_{ik} W h_k^{(t)}) \tag{3.12}$$

其中 $\gamma_{ik} = \lambda_1 \mathbf{M}^r_{ik} + \lambda_2 \mathbf{M}^p_{ik} + \lambda_3 \mathbf{M}^q_{ik}$；$\lambda_1$、$\lambda_2$、$\lambda_3$ 满足: $\lambda_1 + \lambda_2 + \lambda_3 = 1$, $\mathbf{M}^p_{ik} \in \mathbf{M}^p, \mathbf{M}^q_{ik} \in \mathbf{M}^q$。$\mathbf{M}^p_{ik} \in \mathbf{M}^p, \mathbf{M}^q_{ik} \in \mathbf{M}^q$ 是上面提到的构建的两个矩阵。$h_i^{(t+1)}$ 是节点 i 经过邻居节点聚合得到的第 $t+1$ 层的节点表示。W 是可学习参数，$h_k^{(t)}$ 是节点 i 的

邻居节点在第 t 层的隐藏层表示。δ 是激活函数。

解码层：对于每个 token，我们将两个粒度的图神经网络得到的图形嵌入与原始的 LayoutXLM 的输出拼接起来。具体的说，每个 token 的表示包括三部分，第一部分是 LayoutXLM 的 token 输出，第二部分是 token 粒度的图神经网络的输出，第三部分是 bounding box 粒度的图神经网络的输出。特别的，对于每个 bounding box, 将从 bounding box 级别的得到的图神经网络表示与该 bounding box 中的每个 token 表示拼接在一起。三部分的输出被拼接在一起得到每个 token 的最终表示后，将其送入标准的 BiLSTM-CRF 进行实体提取。首先被送入 Bi-LSTM 网络进行编码，输出被进一步传递给全连接网络，最后是 CRF 网络。

$$z_{1:n} = BiLSTM(\vec{H}_{1:n}; 0, \theta_{lstm}) \tag{3.13}$$

其中 $\vec{z}_j \in \mathbb{R}^{n \times d}$ 由上面所说的三部分的特征得到。最后使用条件随机场为序列生成一组条件概率。$z_{1:n}^{final} = \left[z_1^{final}, z_2^{final}, \ldots, z_n^{final}\right]$, and the probability disritibution of a label $\hat{y} = \left[\hat{l}_1, \hat{l}_2 ..., \hat{l}_n\right]$ 整个模型参数通过最小化以下损失函数联合训练:

$$p(\hat{y} \mid z) = \frac{exp(\sum_{i=1}^{n} W_{(l_{i-1}, l_i)} z_i^{final} + b_{(l_{i-1}, l_i)})}{\sum_{y' \in Y(s)} exp(\sum_{i=1}^{n} W_{(l'_{i-1}, l'_i)} z_i^{final} + b_{(l'_{i-1}, l'_i)})} \tag{3.14}$$

CRF 层的解码是搜索具有最高条件概率的输出序列 y^*。

$$y^* = \underset{y \in \hat{Y}}{argmax} - \sum_{i=1}^{n} log(p(y_i \mid z_i)) \tag{3.15}$$

3.5 实验

3.5.1 数据集介绍

PaddleOCR 模型架构：在实际使用中，OCR 模型一般需要在嵌入式设备上运行，因此模型的大小也是一个重点考虑因素，需要在确保模型效能的同时，尽力减小模型尺寸。PaddleoCR 是由百度公司开源的超轻量 OCR (Optical Character Recognition) 系统，主要由文本检测、检测框矫正和 CRNN 文本识别三部分组成，模型结构如图3.7所示。

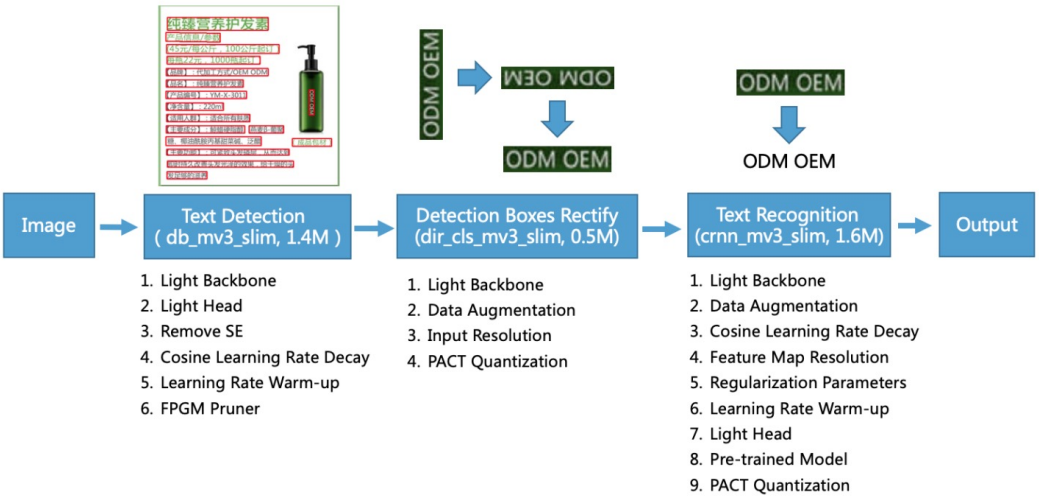


图 3.7　模 PaddleOCR 模型结构 [107]

文本检测的目的是定位图像中文本区域的位置，在 PaddleOCR 中使用 Differantiable Binarization（DB）作为简单分割网络的文本检测器。为了尽量提高其有效性和效率，采用以下六种策略：轻型骨干网络、轻型头部网络、余弦学习率衰减、学习率预热和 FPGM 剪枝，大大缩小了文本检测的模型。检测框在识别检测到的文本之前，需要将文本框转换为水平矩形框，以便进行后续的文本识别。由于检测框架由四点组成，容易通过几何变换来实现，但是可将指定的框反转，因此需要一个分类器来确定文本的方向。如果重新确定了一个框，则需要进一步翻转。训练文本方向分类器是一项简单的图像分类任务。PaddleOCR 采用了以下四种策略来提高模型的能力和缩小模型的大小：轻型骨干网、数据增强、输入分辨率和 PACT 量化。在 PaddleOCR 中的文本识别中，采用 CRNN 作为文本识别器，CRNN 在文本识别中有着广泛的应用和实用价值。CRNN 集特征提取和序列建模于一体，采用 Temporal Classification (CTC) 损失来避免预测和标记之间的不一致性。为了提高文本识别器的可建模性和减小模型大小，PaddleOCR 采用了以下九种策略：轻型骨干网络、数据增强、余弦学习率衰减、特征映射解析、正则化参数、学习率预热、轻型头部网络、预训练模型和 PACT 量化。

通过 Paddl1eOCR 能够实现文本的检测与识别。PaddleOCR 已经提供适用于中文的预训练模型 chinese_text_detection _db _server。

chinese_ocr_db_crnn_server 模块可以用于识别图片当中的汉字。其基于 chinese_text_detection_db_server 检测得到的文本框，继续识别文本框中的中文文字。之

后对检测文本框进行角度分类。最终识别文字算法采用 CRNN（Convolutional Recurrent Neural Network）即卷积递归神经网络。其是 DCNN 和 RNN 的组合，专门用于识别图像中的序列式对象。与 CTC loss 配合使用，进行文字识别，可以直接从文本词级或行级的标注中学习，不需要详细的字符级的标注。该模块是一个通用的 OCR 模型，支持直接预测。

PPOCRLabel 和 LabelImg 工具简介: PPOCRLabel 是 PaddleOCR 的一款开源的半自动化标注工具，使用 python 编写，支持矩形框标注和四点标注法，导出格式可以直接用于 PPOCR 模型训练。工具本身提供预训练模型包括 Chinese、English、French、German、Korean 和 Japanese 模型。使用预训练模型可以快速的进行大批量的自动标注，在完成自动标注之后，可以根据需要进行修改。在使用过程中，发现模型会出现一些漏标或是将一些专有名词分开的情况，需要人工修改。完成修改后，点击 “check”，将会覆盖掉之前的结果。该标注工具以文件夹为单位，最终会在文件夹下生成一个 txt，包含文本信息，及矩形框的四个顶点坐标。LabelImg 是一个图像标注工具，python 编写，用 QT 作为图形界面，标注完成保存可以选择保存为 ImageNet 使用的 PASCAL VOC 格式文件，或 XML 文件。保存内容包括类别和标注框的位置信息，分别为左上点和右下点坐标。

我们应用提出的模型从现实生活中的真实视觉富文档数据集中尝试提取信息并评估其结果。它们是国际货运单数据集（Freight-BI）、FUNSD、XFUN。

（1）国际货运单数据集由 4,800 张海运单图片组成。借助上述的 OCR 标注工具，标注每个 bouding box 的坐标，并为每个 bounding box 标注了预定义的实体类型，我们预定义了 20 种实体类型，包括发货人地址，发货人电话，收货人地址，收货人电话，起运港，卸货港等。际货运单文档版面布局复杂多样，没有固定的模板。文档中存在许多噪声，例如运单的最下方通常包含复杂格式的子表，以及运单中存在许多内容相似但是语义不同的字段等。数据集样例如图3.8所示。我们将数据集的 70% 作为训练集，30% 作为测试集。

表 3.1　Funsd 和 Freight-BI 数据集的统计。包括关于键 (Key)、值 (Value)、边界框 (Bounding box) 和 Token 数的平均数和标准差

Dataset	Training	Testing	Entities	Key	Value	BD	Token
FUNSD	149	50	4	17.8 ± 14.4	21.7 ± 12.1	47.9 ± 26.3	234.4 ± 104.7
Freight-BI	3,375	1,125	20	18.7 ± 28.2	26.5 ± 8.9	71.6 ± 28.2	441.0 ± 213.7

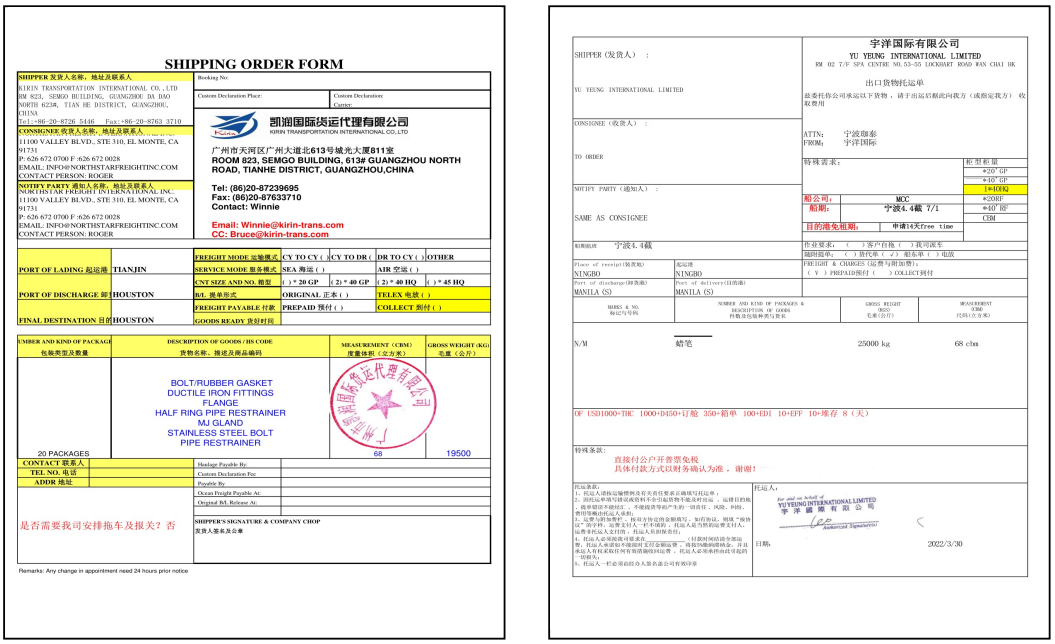


图 3.8　国际货运单数据集样例

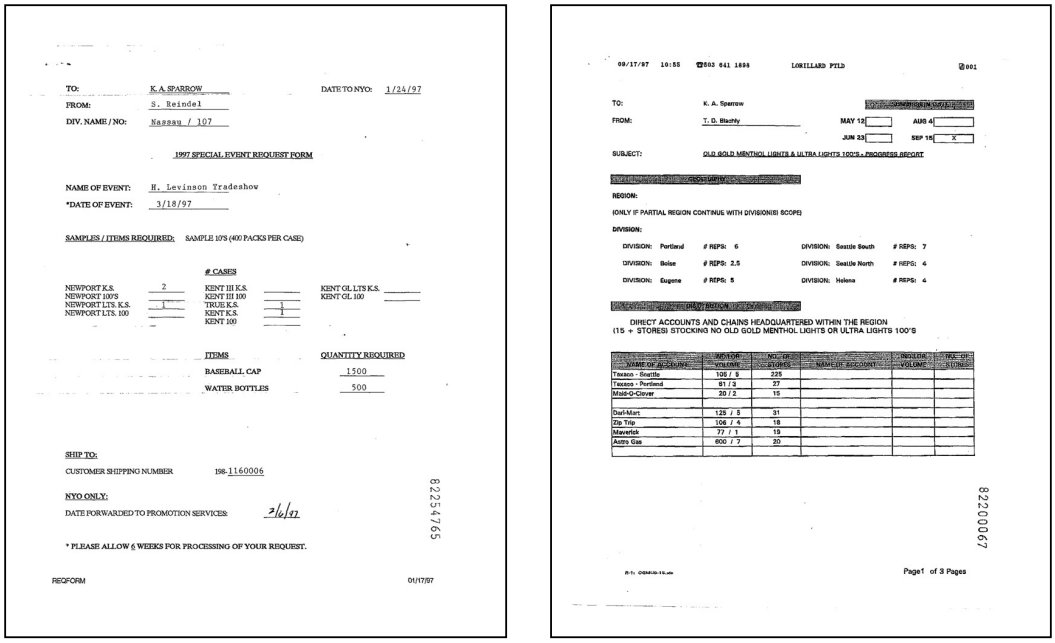


图 3.9　Funsd 数据集样例

（2）FUNSD[12] 是一个有噪声的，用于文档理解任务的英文数据集。它由 199 个真实、完整、带注释的扫描表单图像组成。数据集分为 149 个训练样本和 50 个测试样本。数据集样例如图3.9所示。它适用于各种任务，但本文是实体分类任务，包括:“Answer”、“Header”、“Question”、“Other” 四种类别。

（3）XFUN[8] 是一个多语言的用于文档理解的数据集。近几年，许多针对文档理解任务的评估数据集被提出，如 PublayNet[108]、FUNSD[12]、SROIE、TableBank[109]、

DocBank[67]、DocVQA[110] 等。这些数据集成功地帮助评估了神经网络模型，并显示了深度学习模型与人类之间的性能差距，极大地促进了文档理解研究的进展。然而，这些评估和基准都只关注于英文文档，限制了对非英文文档理解任务的研究。为此，微软亚洲研究院的研究员们按照英文表单理解数据集的标注方式扩展到了其他 7 种语言，包括中文、日文、西班牙文、法文、意大利文、德文和葡萄牙文，提出了一个新的多语言表单理解基准测试数据集 XFUN。数据集样例如图3.10所示。每种语言由 199 个真实、完整、带注释的扫描表单图像组成。数据集划分为 149 个训练样本和 50 个测试样本。它适用于各种任务，但本文是实体分类任务。包括: “Answer”、“Header”、“Question”、“Other” 四种类别。XFUN 数据集的统计信息如表3.2所示。

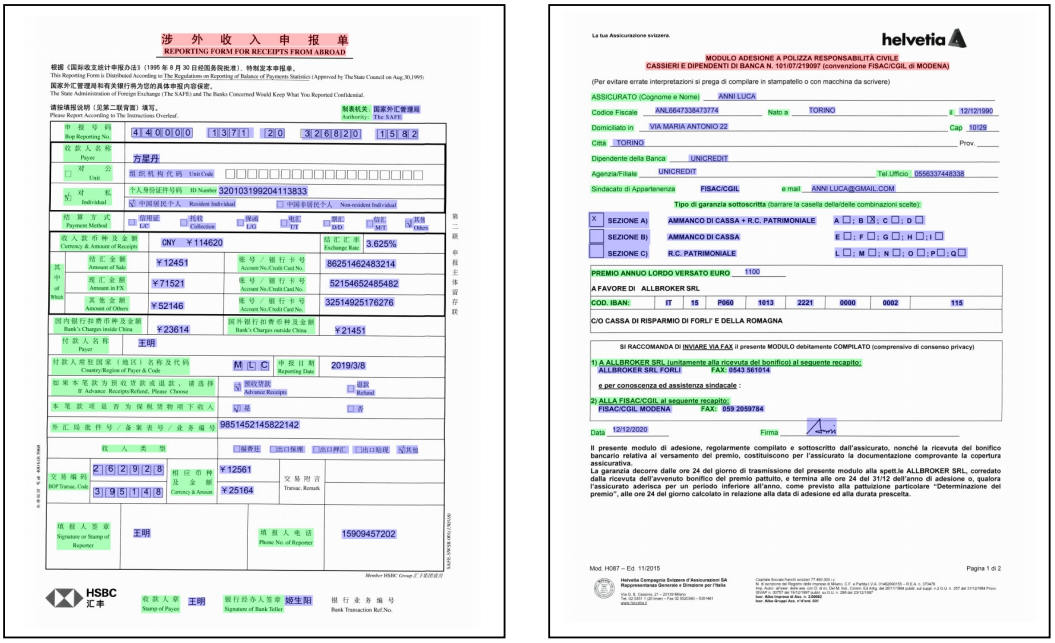


图 3.10　XFUN 中两种不同语言的视觉富文档：(a) 中文；(b) 意大利语

表 3.2　XFUN 数据集的统计信息。表中的每个数字表示每个类别中的实体数量

lang	split	header	question	answer	other	total
ZH	training	229	3,692	4,641	1,666	10,228
	testing	58	1,253	1,732	586	3,629
JA	training	150	2,379	3,836	2,640	9,005
	testing	58	723	1,280	1,322	3,383
ES	training	253	3,013	4,254	3,929	11,449
	testing	90	909	1,218	1,131	3,501
FR	training	183	2,497	3,427	2,709	8,816
	testing	66	1,023	1,281	1,131	3,501
IT	training	166	3,762	4,932	3,355	12,215
	testing	65	1,230	1,599	1,135	4,029
DE	training	155	2,609	3,992	1,876	8,632
	testing	59	858	1,322	650	2,889
PT	training	185	3,510	5,428	2,531	11,654
	testing	59	1,288	1,940	882	4,169

3.5.2　实验环境与细节

本章所做实验使用的服务器（硬件环境）为：2 张 Geforce RTX 3090 Ti 显卡，系统内存大小为 128G。软件环境为：Ubuntu 20.04.2 操作系统，Pytorch 1.8.1 版本，transformer 4.5.0 版本，CUDA 11.4 版本，python 3.8.10 版本。

在编码器部分，token 级别的嵌入表示由 LayoutLM 得到。Bounding box 级别的文本特征由 Sentence-BERT 实现，产生文本嵌入。图像特征提取器由 ResNet101 实现，生成图像嵌入，然后，使用 ROI Align 根据 bounding box 的坐标获得关注的区域。位置特征由两个独立的位置嵌入层编码得到。然后通过拼接的方式将文本嵌入、图像嵌入，位置嵌入结合起来进行特征融合，作为 bounding box 级别的图神经网络的输入。在我们的实验中，图卷积的层数 l=2，注意力机制的头数为 8，alpha 为 0.005。Decoder 由 BiLSTM 和 CRF 层组成。在 BiLSTM 层中，隐藏大小设置为 768，层数为 2，并对模型参数进行随机初始化。使用 F1 值来评估模型的的性能。网络模型是用 Pytorch 框架实现。使用 Adam 优化器来训练网络，最多可训练 50 个 epochs。学习率设置为 5e-5，dropout 设置为 0.5，batchsize 设置为 4。

表 3.3　基于双流排版图网络的实体识别方法实验环境

部件	**参数**
操作系统	Ubuntu 20.04.2
系统内存	128G
CPU 处理器	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz
GPU 处理器	Geforce RTX 3090 Ti
Python 版本	1.8.1
Pytorch 版本	3.8.10
CUDA 版本	11.4
transformer 版本	4.5.0

3.5.3　实验结果与分析

模型评价指标：评价指标使用的是 SER 任务中常用的评价指标，通过衡量实体类型的 F1 分数（F1-score）来评价模型的好坏。混淆矩阵（Confusion Matrix）如前述表 2-2 所示。精确率（Precision, P）、召回率（Recall, R）和 F1 分数（F1-score, F1）计算公式如：

（1）精确率（Precision, P）：

$$P = \frac{TP}{TP + FP} \tag{3.16}$$

（2）回率（Recall, R）

$$R = \frac{TP}{TP + FN} \tag{3.17}$$

（3）F1 分数（F1-score, F1）

$$F1 = \frac{2 \times P \times R}{P + R} \tag{3.18}$$

表 3.4　混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

为了验证我们提出的方法的性能, 与以下几种基线模型相比较：

BERT: 因为 XFUN 是一个多语言的数据集，所以我们使用多语言的 BERT，获得 token 粒度的文本特征。

BERT+ResNet: 使用 BERT 获得 token 粒度的图象特征后，再使用 ResNet 模型得到 token 粒度的图象特征，将两种模态的特征进行融合后，输入分类器进行 token 类别的判断。

BERT+BILSTM+CRF: 使用 BERT 获得 token 粒度的图象特征后，再经过一层双向的 LSTM 网络，提升 token 特征的表示能力，使用 CRF 模型进行解码。

XLM-RoBERTa：与 BERT 一样，XLM-RoBERTa 通过在大规模的无监督数据上进行预训练，学习到了丰富的语言表示。但不同于 BERT 只支持单一语言，XLM-RoBERTa 旨在处理多语言文本，具有跨语言的能力。XLM-RoBERTa 通过在多种语言中进行预训练，从而能够在不同语言之间进行跨语言的知识迁移，从而在资源稀缺的低资源语言中表现出色。

LAMBERT[111]: LAMBERT 也是基于 Transformer 编码器架构，与经典语言模型的唯一区别是使用边界框坐标作为附加输入，而且没有使用文档的图像信息。Lambert 的一个缺点是只支持英文文档，不支持多语言文档。

GraphIE[112]: 一个基于图的信息抽取框架。主要通过图卷积神经网络在连接的节点之间传播上下文信息，以改进非局部或非顺序的信息提取能力，从而改进单词级别的预测能力。GraphIE 应用于三种不同的任务：文本，社交媒体和文档可视化信息提取。

TRIE[85]: TRIE 是第一个将检测、识别、信息抽取整合成一个 end-to-end 框架的工作，其中，检测部分是 FPN+Faster-RCNN，识别部分是 attention+LSTM，信息抽取部分是 BiLSTM-CRF。通过联合训练，文本阅读与信息抽取能够相互促进，提升准确率。

MatchVIE[9]: MatchVIE 是一种基于图神经网络的关键值匹配模型。该方法利用实体间的匹配相关性进行视觉信息提取，绕过了对各种语义的识别，只关注实体间的强相关性。

LayoutLM 和 **LayoutXLM** 模型在 2.4 小节中已经介绍过，这里不再介绍。

表 3.5　Precession(Prec), Recall(Rec), 和 F1 值在 FUNSD 和 Freight-BI 数据集，SER 任务上的得分比较。

Method	**FUNSD（英文）**			**Freight-BI**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
BERT	45.61	59.14	51.50	63.46	69.30	66.26
BERT+BILSTM+CRF	50.06	57.71	55.61	67.61	70.57	69.06
BERT+ResNet	47.32	61.98	53.67	67.75	72.54	72.07
LAMBERT	41.47	52.57	46.37	75.56	83.34	79.42
LAMBERT+ResNet	44.09	57.67	49.97	78.41	86.09	82.07
XLM-RoBERTa	63.31	70.18	66.57	-	-	-
GraphIE	-	-	72.12	-	-	-
TRIE	-	-	78.86	-	-	-
MatchVIE	-	-	81.33	-	-	-
LayoutLM	77.51	83.56	80.42	83.92	87.22	85.54
LayoutXLM	77.57	80.64	79.07	-	-	-
DUAL-VIE	**83.17**	**83.75**	**83.46**	**86.59**	**90.57**	**88.54**

表 3.6　Precession(Prec), Recall(Rec), 和 F1 值在 XFUN 的 ZH（中文）和 JA（日文）数据集，SER 任务上的得分比较。

Method	**ZH（中文）**			**JA（日文）**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
BERT	42.49	55.41	48.10	41.22	47.25	44.03
BERT+BILSTM+CRF	43.71	56.52	49.29	39.45	53.61	45.45
BERT+ResNet	47.57	57.82	52.19	42.15	50.62	46.00
XLM-RoBERTa	46.74	59.14	65.26	71.70	84.62	77.63
LayoutLM	66.24	73.94	69.88	60.57	68.96	64.50
LayoutXLM	83.32	90.69	86.85	72.14	87.01	78.88
DUAL-VIE	**85.16**	**90.95**	**87.96**	**73.61**	**86.05**	**79.35**

表 3.7　Precession(Prec), Recall(Rec), 和 F1 值在 XFUN 的 ES（西班牙文）和 FR（法文）数据集，SER 任务上的得分比较。

Method	**ES（西班牙文）**			**Fr（法文）**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
BERT	40.43	47.75	43.78	47.00	52.50	49.60
BERT+BILSTM+CRF	46.75	44.32	45.51	45.73	57.62	50.99
BERT+ResNet	41.58	48.41	44.74	46.36	54.64	50.16
XLM-RoBERTa	58.55	60.20	59.36	62.78	71.98	67.07
LayoutLM	62.73	70.54	66.41	71.91	78.21	74.93
LayoutXLM	74.11	76.31	75.19	78.15	78.82	78.84
DUAL-VIE	**75.32**	**77.14**	**76.22**	**79.97**	**90.85**	**80.41**

表 3.8　Precession(Prec), Recall(Rec), 和 F1 值在 XFUN 的 IT（意大利文）和 DE（德文）数据集，SER 任务上的得分比较。

Method	**IT（意大利文）**			**DE（德文）**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
BERT	41.55	52.13	46.24	47.73	49.37	48.54
BERT+BILSTM+CRF	42.74	50.94	46.48	47.20	54.63	50.65
BERT+ResNet	43.98	51.11	47.28	49.45	57.15	53.02
XLM-RoBERTa	64.19	70.00	66.96	66.25	70.18	68.16
LayoutLM	70.04	79.59	74.51	68.88	78.74	73.49
LayoutXLM	76.99	83.39	80.06	77.34	81.78	79.50
DUAL-VIE	**78.62**	**83.42**	**80.95**	**78.43**	**82.90**	**80.60**

表 3.9　Precession(Prec), Recall(Rec), 和 F1 值在 XFUN 的 PT（葡萄牙文），SER 任务上的得分比较。

Method	**PT（葡萄牙文）**		
	Prec	**Rec**	**F1**
BERT	46.13	56.15	50.65
BERT+BILSTM+CRF	46.58	56.20	50.94
BERT+ResNet	47.50	57.45	52.00
XLM-RoBERTa	62.46	68.64	65.40
LayoutLM	64.78	74.77	69.42
LayoutXLM	75.09	81.68	78.24
DUAL-VIE	**78.68**	**81.70**	**80.16**

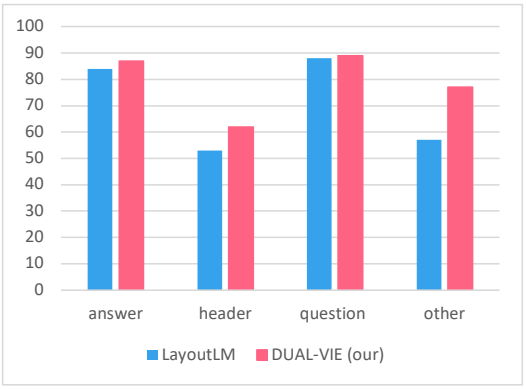


图 3.11　模型在 FUNSD 数据集上的效果比较

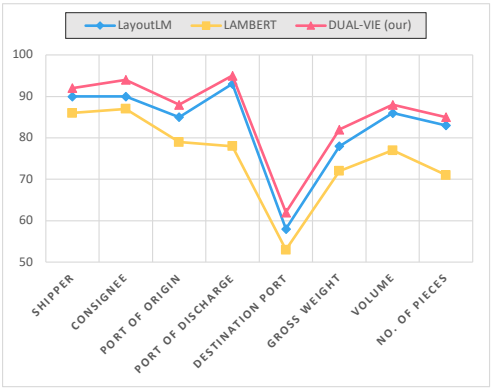


图 3.12　模型在 Freight-BI 数据集上的效果比较

如表3.5到表3.9中结果所示，与传统的只使用文本的信息提取方法相比，引入图形特征后的性能更高。值得注意的是，在 LamBert 模型中，没有考虑图像特征，而在我们的实验中，加入图像特征后，模型效果得到了很大的改善。LAMBERT 在 FUNSD 数据集中的效果不是太好，可能是因为数据集的大小问题。同时，LayoutLM 和 LayoutXLM 在同时使用文本、布局和图像信息时取得了较好的性能，且 LayoutXLM 的性能优于 LayoutLM。由于考虑了不同的粒度特征，我们的方法在各个数据集上取得了最好的效果，并以明显的优势超过了 LayoutLM（在 FUNSD 数据集上为 3.02%，在 Freight-BI 数据集上为 3.0%。在 XFUN 的各个语言数据集上，所提出的模型均优于 LayoutXML 模型，验证了粗粒度的元素特征对 SER 任务的有效性。

消融实验：如表3.10所示，我们分析了模型的各个组成部分对模型结果的影响，包括 token 粒度图神经网络、bounding box 粒度图神经网络。我们设定了不考虑这两个模块时模型的 F1 值的变化。如果没有 token 粒度图神经网络，该模型就不能很好地进行 bounding box 和对应 token 之间的信息交互，结果在 FUNSD 数据集上为 0.78，在 Freight-BI 数据集上为 0.74。由于 bounding box 的粒度图神经网络可以对键值对之间的关系进行建模。如果没有它，在 FUNSD 数据集上的 F1 得分下降为 0.66，在 Freight-BI 数据集上的 F1 得分下降为 0.82。

表 3.10　我们提出的模型在 FUNSD 和 Freight-BI 数据集，SER 任务上的消融实验

Model	**FUNSD**	**Freight-BI**
full model	**83.46**	**88.54**
w/o token-level CB-GAT	82.68	87.80
w/o BD-level KNN-GAT	82.80	87.72

结果可视化展示：本文对海运单数据集关键信息抽取模型在已知和未知类别的测试集上进行了批量测试，并将测试结果保存至 excel 中以便检索。其中，部分测试结果如图3.13所示。保存的模型在测试集上识别效果如图3.14所示，其中不同颜色的框表示不同的实体类别。

文档名称	发货人地址	发货人电话	发货人传真
上海瑞浩.xls	RM 901-903, 9/F TOWER 2 LIPPOCTR NO 89 QUEENSWAYHONG KONG	+852-25253830	+852-25256930
上海汇涛.xls	SOUTH SIDE OF SHENJIN ROAD,LILI TOWN,WUJING DISTRICT,SUZHOU CITY		
上海柏联.xls	RM2805, SHANGHAI INFORMATION TOWER,NO.1403 OF MINSHENG ROAD, SHANGHAI 200135, CHINA		
青岛浩航.XLS	1, FOURTH FLOOR NEW CITY DIJING 38 BUILDING,HUTANG TOWN, WUJING DISTRICT, CHANGZHOU, JIANGSU, CHINA		
深圳阡陌.pdf	09TH FLOOR,SECTION B TAIPINGYANGCOMMERCIAL & TRADE BUILDING,NO.4028 JIABINROAD,LUOHU DISTRICT,SHENZH	0755-32930990	
南海运通.xls	8/F,XIAMEN INTERNATIONAL SHIPPING CENTER D,NO.97 XIANGYU ROAD,XIAMEN AREA OF CHINA(FUJIAN) PILOT FRI		
深圳顺丰.docx	No.100, Hengtong Road ,Economic DevelopmentZone ,Wujiang City,Jiangsu Province,China		
上海鼎熙.DOCX	NO. 417 SAN XIANG RD, KUNSHAN CITY JIANGSU PROVINCE, P. R CHINA		
宁波达倍国际物流有限公司.docx	25F.,NO.168,EAST BAIZHANGROAD,INTERNATIONALEXHIBITION CENTER,NINGBO CHINA		
中太世达国际物流有限公司LCH	NO.26xiqinRD., national hi-tech industrialdevelopment zone,wuxi,jiangsu,PRC	0510-85342899-6152	0510-85343586
昆山敏亚.DOC	NO.1000 NANBANG ROAD,KUNSHAN JIANGSU PROVINCE 215300,CHINA		
波合一国际货运代理有限公司.D	Nanwo road,Panwan town,SheyangCounty,Yancheng City,Jiangsu Province,China	15061181502	0515-82758889
常州隆和.doc	28 LIJIANG RD,XIXIASHU TOWN,XINBEI DISTRICT,CHANGZHOU,213181.JIANGSU, P.R.CHINA	0519-83441318	
上海贸联.xlsx	NO.39 SONGBEI ROAD, SUZHOU INDUSTRIAL PARK,JIANGSU PROVINCE, P.R.OF CHINA		
上海信昱晟.rtf	ADDRESS:RM409,288 WUHUA RD., AIJIA BUILDING3#,SHANGHAI,CHINA.	86-21-66670083	86-21-51685466
海泷特国际物流有限公司 HCM	7C,SHIMEI MANSION,445 JIANG NING RD. ,JING ANDISTRICT,SHANGHAI,CHINA.	86-21-51523068	86-21-51523066
上海永津.rtf	NO. 31, TANGNING ROAD, XINGWANG INDUSTRIAL CITY,YUHANG, HANGZHOU, ZHEJIANG, CHINA.		
BOOKINH (2).xls	112 Qinghu East Road, Meilong Avenue, Mingzhi Streets, Longhua New District, Shenzhen	8615989430162	
ORDER.All-Ways-43137第一个!	NO.128 QIXING EAST ROAD,ZICHUAN ECONOMIC DEVELOPMENT ZONE,ZIBO,CHINA.255188	86 13589382778	
booking-Po17599,17530-1.pdf	Team 20, Maxi Village, Matang Town,Rudong County, Jiangsu Province, China	+8651384477728	+8651384471262
ER All-Ways Ocean - GBI-IN#	P, Q, R, 12/F, WAH LIK IND., CENTRE,459-469 CASTLE PEAK ROAD,TSUEN WAN, N.T. HONG KONG	(852) 2770-3739	
BOOKING-NJEG21032 修改1.xls	18TH FLOOR, INVESTMENT MANSION, NO.414, SOUTH ZHONG SHAN ROAD, NANJING, CHINA		
BOOKINH JURA PO106341.pdf	RM. 809-810, SHATIN GALLERIA, 18-24 SHAN MEI STREET,FOTAN, N.T. HONG KONG	02086796388-6136	
BOOKINH 12051#.xls	Liwu Village,Yuanzhou Town,Boluo County,Huizhou Guangdong China	18925728183	
BOOKINH (3).xls	No. 11, Xing Ye East Road, Zone AShishan Industrial ParkNanhai, Foshan, Guangdong CHINA	13590552988	
BOOKINH(1).xls	Room 309, NO.82 Xingang Dong Road, Haizhu District, Guangzhou, China	020-37574104	
志泰.pdf	BIYU INDUSTRIAL PARK, JIANGZAO TOWN,ZHUJI CITY,ZHEJIANG PROVINCE,CHINA	0575-87715282	0575-87715282
鑫轮进出口.pdf	ROOM446, DAGONGGGUAN, NO.135, CANGSONG ROAD,NINGBO CITY,ZHEJIANG PROVINCE	86-574-87478681	86-574-88658755
新托书模板--8379.doc	NO.88,SHANGHAI,N.STR,LUQIAO,318050,TAIZHOU,ZHEJIANG,CHINA.		
新托书模板111.pdf	No. 19, Jinfu 2nd Road, Huanan Ind. Zone, Liaobu Town, Dongguan,Guangdong, 523430, China		
新托书模板.pdf	NO.5, ALLEY 9, LANE 48, SAN CHUN ST., SHU LIN DIST., NEW TAIPEI CITY 23865, TAIWAN	886-2-26896015	886-2-26896130
新版托书模板(1)(1).pdf	QUJIAWA VILLAGE, HI-TECH ZONE, YANTAI,SHANDONG, CHINA		
新托书模板.doc	BUILDING NO.87,THE 1ST INDUSTRIAL ZONE, LISONGLANG GONGMING STREET,GUANGMING NEW DISTRICT, SHENZHE	0755-27694669	
中天.pdf	NO.15 XINGGONG 2 ROAD,NINGHAI COUNTY,NINGBO,ZHEJIANG,CHINA	0086-574-6520781	0086-574-65563118
L-21013123 - booking form.do	12F., NO. 421, SUNG SHAN RD., XINYI DIST., TAIPEI, TAIWAN 11083		
FINE SHINE美国托书.doc	5 Jin shajiang Road,Tongji Block,Jimo, 266228,800, PYEONGCHEON-RO,BUCHEON-SIGYEONGGI-DO, REPUBLIC OF KO	82 32 672 4022	82 32 677 4022
home always货代新托书模板.pdf	FORMDONGQIAO INDUSTRIAL ZONE,HAISHU, NINGBO CHINA		
-Booking-NINGBO BOHONGCHUA	4F,BUILDING 1,NO.,25 YUXIU ROAD,ZHUANGSHI STREET, ZHENHAI DISC.,NINGBO	86-574-87611718	86-574-87612598
HAG012 订舱.pdf	Address:No.1 Industrial Concentration Zone,Yugou Town,Huaiyin District,Huai an CityPost code:234222	0517-84743098	
terial Industry Co.,ltd-booking f	NO.35,SANMING ROAD,HECHENG STREET,GAOMING DISTRICT,FOSHAN CITY,CHINA	0757-88622261/88620007	
GSN BOOKING3.18.pdf	No.50 Zhoutang East Street,Xiaolin Town,Cixi,Zhejiang,China 315321	86-574-63699997	
FOR SHIPMENT qingdao.doc	ROOM 1202, 45 EAST HUAIHAI RDSHANGHAI 200021, CHINA		
HPS-CM20210407托书.pdf	No.522 Tongshan Road, Gu Tang Industrial Zone, Yiwu, Zhejiang, 322000, China		
new booking (1).xlsx	6-2-201 WEIDUOLIYA YUHUA DISTRICT SHIJIAZHUANG HEBEI CHINA		
NEW BOOKING.doc	DONGMEN INDUSTRIAL,FENGZHOU,NANAN,362330QUANZHOU,FUJIAN,CHINA		
petit clair booking -4.14.docx	D BLD, FEIHUANGDA INDUSTRIAL PARK, JINXIU RD, CUNTOU NEW DISTRICT, HUMEN TOWN, DONGGUAN, GUANGD	86-769-89339813	86-769-89339813
SC210304093-26'28 BOOKING.x	UNIT 5, 9/F., HEWLETT CENTRE,54 HOI YUEN ROAD,KWUN TONG, KOWLOON, HONG KONG		
ny2384-booking.pdf	No.288 North Zhenghan Street,Xietang Industrial Park,Jindong District Jinhua 321036,China	0579-82982055	0579-82982055
INNNNNNNNNNEW booking form.p	MMERCIAL BUILDING,NO.269,CHONGDE ROAD,CHENCUN COMMUNITY,QINGYANG STREET,JINJIANG CITY,QUANZHOU C		
Y2247 Booking Form-新格式.dc	HONGTANG INDUSTRIAL ZONE ANINGBO,ZHEJIANG CHINA		
New BookingForm (PO#16678).pd	No.8 Ln.8 Xinli E. St., Toufen City, Miaoli county 35154, TAIWAN	+886376610 93	
0011 BOOKING FROM3M 513 BLK B 5/F NEW MANDARIN PLAZA 14 SCIENCE MUSEUM ROAD TSIM SHA TSUI EAST KOWLOON HONG			

图 3.13 部分海运单视觉富文档关键信息结构化结果

图 3.14　模型预测结果可视化展示

3.6 本章小结

本章提出了一种新的用于视觉信息提取的基于双流排版图网络的实体识别方法。该模型利用 token 粒度和 bounding box 两个粒度的结合来提取视觉富文档中的语义实体。我们研究了在每个粒度上聚合文本、视觉和位置特征的方法的有效性。此外，我们的模型可以关注潜在的候选键值对。在多个 VIE 数据集上的实验结果表明，粗粒度元素特征对于 VIE 任务非常重要，特别是对于具有复杂布局的视觉富文档。

第四章　基于 SPAN 图关联的实体关系抽取方法

4.1　研究动机

在视觉富文档理解的相关任务中，关系抽取任务要求模型预测任意两个语义实体之间的关系，从而帮助自动化处理文档。例如，在发票处理任务中，货物名称和价格之间的“包含”关系，公司名称和发票日期之间的“开具”关系等等。通过这种方式，关系抽取可以帮助自动化文档处理过程，提高效率和准确性。

然而，目前的模型，例如 LayoutXLM 使用的是头尾实体中第一个 token 做关系分类，使得模型难以捕获实体的全局信息。关系抽取可以看做是 SPAN 粒度的分类问题，粗粒度的实体特征对关系抽取非常重要。

所以本章基于提出 SPAN 图关联的视觉富文档关系抽取方法，探究了分别使用 token 和 SPAN 粒度多模态特征以及不同的文档编码器在文档关系抽取任务上的效果。综合实验表明，所提出的模型和方法明显优于现有的基于 token 粒度模型的关系抽取效果。

4.2　相关工作

关系抽取（Relation Extraction，RE）是自然语言处理领域的一个重要任务，其目标是从文本中提取出实体之间的语义关系。长期以来，人们对关系理解进行了深入研究。先前的工作集中于识别同一句子中两个命名实体之间的关系。基于此，句子级和文档级的关系理解 [113-115] 吸引了研究人员，因为在自然语言处理中，理解语言中不同层次之间的关系是一个重要的问题，能够帮助提高语言处理的准确度和效率。[116] 提出一种基于卷积神经网络（CNN）的关系分类模型。它使用了池化卷积神经网络（PCNN）来对实体的上下文信息进行建模，同时引入了注意力机制来加强模型对于关键信息的关注。该模型在多个数据集上都取得了很好的效果。[117] 提出一种基于多实例多标签分类的关系抽取模型，它使用了双向门控循环单元（BiGRU）、卷积神经网络（CNN）和注意力机制来提取实体之间的关系。该模型能够处理多个实体之间的关系，并且在多个数据集上都取得了很好的效果。[118] 是一种基于 BERT

和图卷积网络（GCN）的关系抽取模型，它能够同时抽取多个实体之间的关系。该模型使用 BERT 来学习实体的上下文信息，然后使用 GCN 来对实体之间的关系进行建模。该模型在多个数据集上都取得了很好的效果。

以上是纯文本中关系抽取的任务和模型介绍。在视觉富文档中的关系抽取与传统的关系抽取有所不同，任务更为复杂。在视觉富文档中的关系抽取，需要检测任意两个 bunding box 之间的关系。视觉富文档经过 OCR 工具转换为纯文本后，两个实体可能会离得很远，但实际上，它们在文档中的位置可能非常接近，甚至在文档中是对齐的。在这个任务中，关系属于预定义的集合，所以它适合得到我们关注的结构化知识。

VRD 中的关系不是传统纯文本级关系，而是源自文本、布局和图像的多模态组合，其中的每种特征都在构建关系时发挥重要的作用。受自然语言处理系统的启发[72, 119]，VRD 理解的解决方案通常遵循预训练加微调机制。然而，现有的大多数相关模型只能处理预训练集合中包含的特定语言 (通常是英语) 的文档数据，这是非常有限的。为了解决这个问题，[103] 提出了一个简单而有效的独立于语言的布局转换器 (LiLT)，用于结构化文档理解。LiLT 可以在单一语言的结构化文档上进行预训练，然后直接使用现成的单语言或多语言预训练文本模型对其他语言进行微调。在八种语言上的实验结果表明，LiLT 可以在各种广泛使用的下游基准测试中获得具有竞争力甚至更好的性能，这使得文档布局结构的预训练能够带来与语言无关的好处。

GNN 被广泛应用于 NER 和表格提取等任务中，[15] 提出将 GNN 应用于提取键值对的任务中，不仅对文档图片中的实体进行分类，而且还会对实体间的关系进行预测。给定一个输入文档，模型需要完成的任务包括：（a）单词分组：检测文档实体，即将相同语义的单词进行分组；（b）实体分类：将检测到的实体分为预设的类别；（c）关系预测：发现实体间配对关系。[120] 提出 FUDGE，一种可视化的表单理解方法，通过结合文本片段（图形顶点）和以迭代的方式修剪边缘来编辑图形结构，以获得最终的文本实体和关系。可以应用于文本识别困难的表单。

在 OCR 下游的信息抽取当中，经常会需要涉及到对实体之间的关系进行建模。表格信息抽取是判断实体是否属于同一行或同一列的关系。字段信息抽取是判断内容实体和关键字实体之间是否存在对应的关系。阅读顺序检测更是判断实体之间是否存在上下文关系。但是以往的建模只能够获取到 token 的 embedding 向量，在预训

练的时候并没有强化对节点和节点之间关系的建模。但是同时可以注意到一个额外的问题，节点和节点之间的关系是难以用“一个固定的模型”表述的，比如两个具有行关系的节点可能在描述列关系的时候就无关了。所以，不应该强行的赋予节点之间的联系，应该让模型在训练过程中自动的学习节点和节点之间的联系，并应用到具体的下游任务当中。基于以上问题，[121] 试图提出一个新的预训练任务，以期望模型在能够获取到 token 的嵌入表示外，同时获取到表示节点和节点之间关系的嵌入表示。

4.3 任务定义

每个经过 OCR 扫描得到的视觉富文档都是由语义实体列表组成的，每个实体由一组词和定义的边界框组成。边界框的坐标，定义为: $b_i = [w_i^1, ..., w_i^m], [x_i^1, x_i^2, y_i^1, y_i^2]$，其中，$[w_i^1, ..., w_i^m]$ 是每个 bounding box 内的词组，$[x_i^1, x_i^2, y_i^1, y_i^2]$ 是每个 bounding box 的对角线上的 x 坐标和 y 坐标。我们所使用的数据集中的文档都被标注有每个实体的标签和实体之间的关系。我们将每个注释文档表示为 $D = \{[b_1, ..., b_n], [l_1, ..., l_n], [(b_1, b_{h_1}), ..., (b_m, b_{h_m})]\}$，其中 $l \in L$ 是每个实体的标签，L 是预定义的实体标签集。(b_i, b_{h_i}) 指实体 b_i 和 b_{h_i} 之间的关系。值得注意的是，实体可能与一个以上的实体存在关系，或者与任何其他实体没有关系。

4.4 提出模型

我们首先增量地构建了关系候选集合，产生所有可能实体键值对。对于每一对头尾实体的表示，我们尝试了两种不同粒度的表示，分别是基于 token 粒度和基于 SPAN 粒度的表示，以探究不同粒度的表示对关系抽取效果的影响。对于 token 粒度的表示，是每个头实体中的第一个 token 向量和每个尾实体中的第一个 token 向量的拼接。对于 SPAN 粒度的表示，我们采用图神经网络，通过图卷积操作，使得每个 SPAN 中所有的 token 表示都聚集到该 SPAN 中第一个 token 表示上，此时得到的第一个 token 的表示就是 SPAN 粒度的表示。为了融入标签信息，将头实体和尾实体的表示与头尾实体的标签类型进行拼接。在分别经过两个线性层后，头部和尾部的表征被拼接起来，然后被送入一个双仿射网络，根据双仿射网络预测出头尾实体的关系分数。

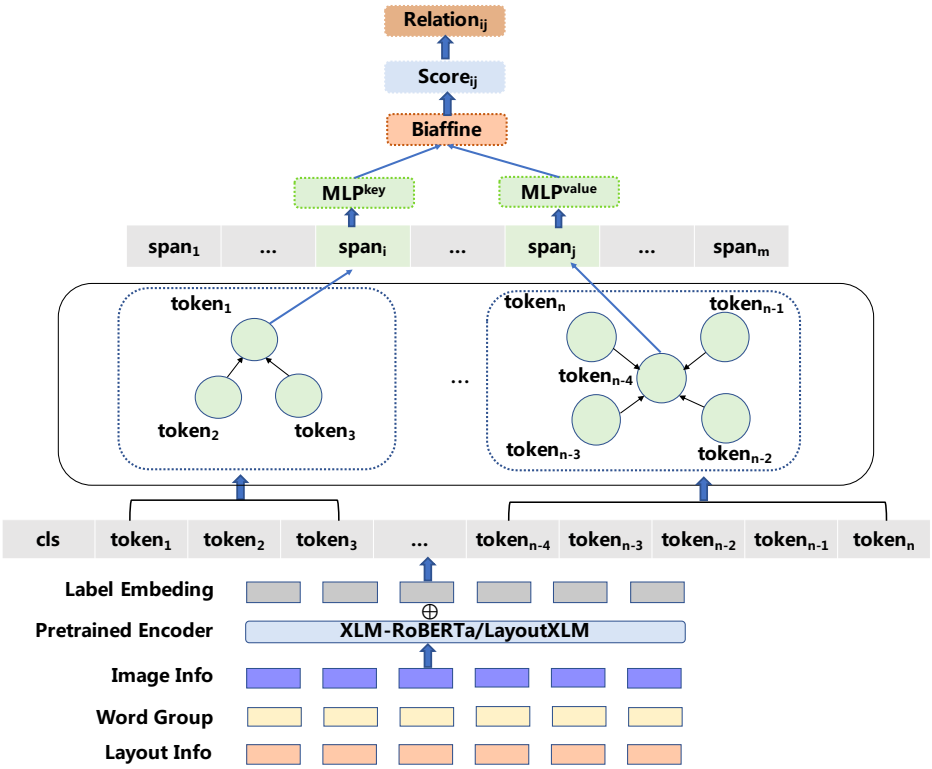


图 4.1　基于 SPAN 图网络的文档关系抽取模型

4.4.1　实体表示

在输入层，为了获得更好的实体表征，我们比较了不同的方式来编码语义实体的信息，包括词组、图像、和布局特征。在这项工作中，我们使用了基于文本和基于文档的预训练模型，包括来自 XLM-RoBERTa、LayoutXLM 的 token 粒度的实体表示，其中使用 XLM-RoBERTa 编码文档只包含文本特征，后者同时包含文本、视觉、以及布局特征。在得到 token 粒度的表示后，我们通过构建 SPAN 粒度的图神经网络，得到 SPAN 粒度的实体表示。此外，我们还利用了每个语义实体的标签，例如“问题”、“答案”等。我们将实体标签映射为标签嵌入，然后，我们把实体表示和标签嵌入拼接起来作为每个语义实体的文档编码器的输入，如公式4.1所示。

$$e_i = b_i \oplus l_i \tag{4.1}$$

其中，l_i 表示实体标签嵌入，b_i 是指语义实体的表示，它可以不同的文档编码器模型中得到，如 XLM-RoBERTa、LayoutXLM。

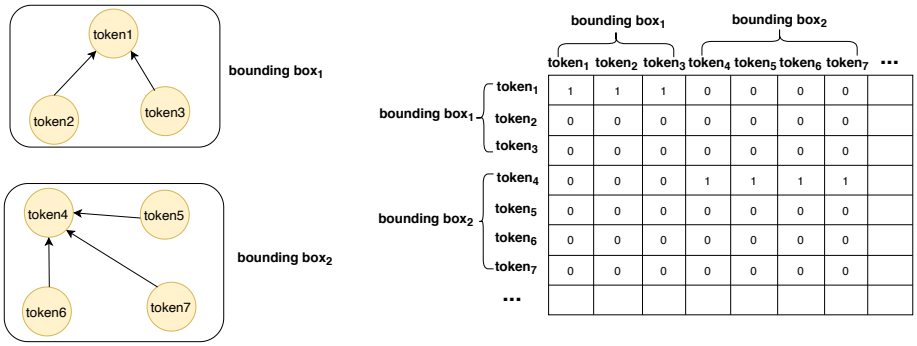


图 4.2 一个 SPAN 内部的 token 连接

4.4.2 图神经网络

在使用 XLM-RoBERTa、LayoutXLM 文档编码器对文档进行编码后，得到 token 粒度的表示。为了 SPAN 粒度的实体表示，我们使用图注意力网络对文档进行进一步编码。具体而言，对于每个 SPAN 里的 token, 使用图神经网络将一个 SPAN 内的所有 token 粒度的信息，聚集到该 SPAN 内第一个 token 的身上，这样每一个 SPAN 内的第一个 token 就代表了整个 SAPN 粒度的表示。图注意力的计算过程如公式4.2和公式4.3所示。

$$\alpha_{j\tilde{j}} = \frac{exp\left(\delta\left(V^T[Wh_j \oplus Wh_k]\right)\right)}{\sum_{n\in N} exp(\delta(V^T[Wh_j \oplus Wh_n]))} \tag{4.2}$$

$$\vec{h}_i' = \delta\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\epsilon N_i}\alpha_{ij}^k W^k \vec{h}_j\right) \tag{4.3}$$

借鉴 [122] 的做法，我们首先通过产生给定语义实体的所有可能键值对来构建关系候选集。对于每一对头尾实体，其表示是 SPAN 粒度的实体表示和实体类型嵌入的拼接。我们首先应用 MLP 模块删除与当前关系决策无关的琐碎信息，得到头尾实体的不同表示，如公式4.4和4.5所示，其中 F 表示激活函数。然后使用双仿射网络计算头尾实体之间的关系分数，如公式4.6所示。

$$h_i^{key} = F\left(W^{key}e_i + b^{key}\right) \tag{4.4}$$

$$h_i^{value} = F\left(W^{value}e_j + b^{value}\right) \tag{4.5}$$

$$Score\,(i,j) = Biaffine\left(h_i^{key}W_1 h_j^{value} + (h_i^{key} \oplus h_j^{value})W_2\right) \tag{4.6}$$

4.4.3 文档编码

为了更好地对视觉富文档中的语义实体的信息进行编码，我们比较了不同的文档编码器，包括 XLM-RoBERTa 、基于 Transformer 架构的文档预训练模和 GNN。具体地说，我们将实体的表示送入文档编码器，并获得编码器的输出作为实体的上下文表示。对于 GNN 编码器，使用 LayoutXLM 编码文档得到 token 粒度的表示，接着使用图注意力网络构建 SPAN 粒度的实体表示。

4.4.4 关系解码器

基于实体之间的关系得分，我们的做法是判断每个 VRD 中任意两个实体之间是否存在关系，这种方法类似于语义角色标记 (Semantic Role Labeling, SRL)。在这种情况下，我们将关系预测作为二元分类任务，并使用二元交叉熵损失 [12]。

4.5 实验

4.5.1 实验环境与细节

本章所做实验使用的服务器（硬件环境）为：2 张 Geforce RTX 3090 Ti 显卡，系统内存大小为 128G。软件环境为：Ubuntu 20.04.2 操作系统，Pytorch 1.8.1 版本，transformer 4.5.0 版本，CUDA 11.4 版本，python 3.8.10 版本。

在我们的实验中，图卷积的层数 l=2，注意力机制的头数为 4，alpha 为 0.005。使用 Adam 优化器来训练网络，初始学习率设置为 5e-5，dropout 设置为 0.5，常规实验训练 140 轮，训练时 batchsize 设置为 8，测试时 batchsize 设置为 16。

表 4.1　基于 SPAN 图关联的视觉富文档关系抽取方法实验环境

部件	**参数**
操作系统	Ubuntu 20.04.2
系统内存	128G
CPU 处理器	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz
GPU 处理器	Geforce RTX 3090 Ti
Python 版本	1.8.1
Pytorch 版本	3.8.10
CUDA 版本	11.4
transformer 版本	4.5.0

4.5.2 实验结果分析

表 4.2 Precession(Prec), Recall(Rec), 和 F1 值在 FUNSD 和 XFUN 的 ZH（中文）语言数据集，RE 任务上的得分比较。

Method	**Funsd**			**ZH（中文）**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
XLM-RoBERTa	24.74	26.66	25.66	39.94	64.33	49.28
XLM-RoBERTa+ResNet	27.34	25.56	26.42	48.34	54.26	51.13
LayoutXLM	48.35	61.60	54.18	59.80	83.13	69.56
ours	**49.06**	**62.73**	**55.06**	**64.79**	**76.97**	**70.36**

表 4.3 Precession(Prec), Recall(Rec), 和 F1 值在 XFUN 的 JA（日文）和 ES（西班牙文）数据集，RE 任务上的得分比较。

Method	**JA（日文）**			**ES（西班牙文）**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
XLM-RoBERTa	43.26	68.03	52.89	41.57	61.74	49.69
XLM-RoBERTa +ResNet	47.76	68.87	56.41	42.13	63.36	50.61
LayoutXLM	58.11	76.83	66.17	58.53	80.53	67.79
ours	**60.79**	**76.29**	**67.67**	**65.82**	**77.94**	**71.37**

表 4.4 Precession(Prec), Recall(Rec), 和 F1 值在 XFUN 的 FR（法文）和 IT（意大利文）数据集，RE 任务上的得分比较。

Method	**FR（法文）**			**IT（意大利文）**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
XLM-RoBERTa	40.68	64.90	50.01	42.67	64.14	51.25
XLM-RoBERTa+ResNet	40.18	67.05	50.25	42.99	67.15	52.42
LayoutXLM	52.60	76.90	62.47	54.02	77.20	63.56
ours	**53.96**	**76.18**	**63.17**	**64.33**	**73.18**	**68.47**

表 4.5 Precession(Prec), Recall(Rec), 和 F1 值在 XFUN 的 DE（德文）和 PT（葡萄牙文）数据集，RE 任务上的得分比较。

Method	**DE（德文）**			**PT（葡萄牙文）**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
XLM-RoBERTa	35.58	55.28	43.29	28.96	59.31	38.92
XLM-RoBERTa+ResNet	45.28	58.05	50.88	33.93	54.80	41.91
LayoutXLM	56.23	71.47	62.94	47.42	68.91	56.18
ours	**59.66**	**71.47**	**65.03**	**50.19**	**69.25**	**58.20**

通过在 Funsd 数据集以及 XFUN 的七种不同语言的数据上的实验表明，所提出的基于 SPAN 图关联的视觉富文档关系抽取方法明显优先基于 token 粒度的关系抽取效果，此外，在文本特征的基础上，同时融合图形特征，关系抽取效果有明显提升，这表明使用更丰富的特征以及采用有效的方式将不同模态的特征融合，可以有效提升文档关系抽取效果。

4.6 本章小结

本章对视觉富文档理解任务中的关系抽取（RE）子任务进行了研究。通过实验研究了融合不同的模态特征以及文档中元素的粒度对关系抽取效果的影响。提出了一种基于 SPAN 粒度的图神经网络用于关系抽取。实验表明所提出的基于 SPAN 粒度的关系抽取效果在 XFUN 的各个语言数据集上的效果要明显优于基于 Transformer 架构的文档预训练模型中基于 token 粒度的关系抽取效果。

第五章　人类阅读顺序对信息抽取效果影响研究

5.1　研究动机

目前学术界中，针对多模态文档理解的模型方案，通常都需要先经过对图像进行 OCR 扫描，解析出图中的文本和文本框位置，再将得到的文本和文本框坐标，按照 OCR 解析出的默认顺序，将文本框及其对应的内容输入给模型。然而，对于布局结构复杂的视觉富文档，诸如票据、表单、卡证等数据，由于其版面存在丰富的层次结构，OCR 的结果往往存在不正常的阅读顺序。如图5.3所示。针对视觉富文档的理解任务，通常需要考虑到文本框的阅读顺序问题，传统的 OCR 技术无法解决这一问题，OCR 结果的坐标顺序通常无法按照“从左至右”或者“从上到下”进行排序。一个合理的文本框阅读顺序（proper reading order），可以帮助模型更好得理解图像信息。

因此本章拟解决的问题是：构建合理的文档阅读顺序，提升视觉富文档信息抽取任务的效果。

5.2　相关工作

阅读顺序检测 [123-127] 旨在捕获文档的正确阅读顺序。这个任务的目的是将 OCR 识别到的文本块，按正常的阅读顺序重新进行排列，解决内容折行、数据分组的问题。一般来说，人类倾向于以从左到右和从上到下的方式阅读文件。然而，由于 OCR 工具对复杂文档提取的标记，这种简单的排序规则可能会失败。人类对信息加工时在很大程度上依赖于视觉，通过眼球的运动获取视觉信息，眼球运动主要包括注视（fixation）和扫视（saccade），前者是指眼球停止运动并开始获取信息，后者则是指眼球在注视点间的快速移动。其中注视包括注视时长、注视顺序等信息。这些眼动信息是人类阅读时获得信息的重要途径也是分析人类阅读习惯的重要途径。认知神经科学从语言处理的角度研究构成人类大脑心理语言处理过程的生物学和认知过程和方面，而自然语言处理则教机器阅读、分析、翻译和生成人类语言序列。我们通过在数据中加入人类阅读文档时的阅读顺序和关注点来提高模型的认知能力从而增

强对富文档的理解。人类阅读文档的顺序如图5.2顺序，基本按照“从左到右”，“从上到下的”的规则。

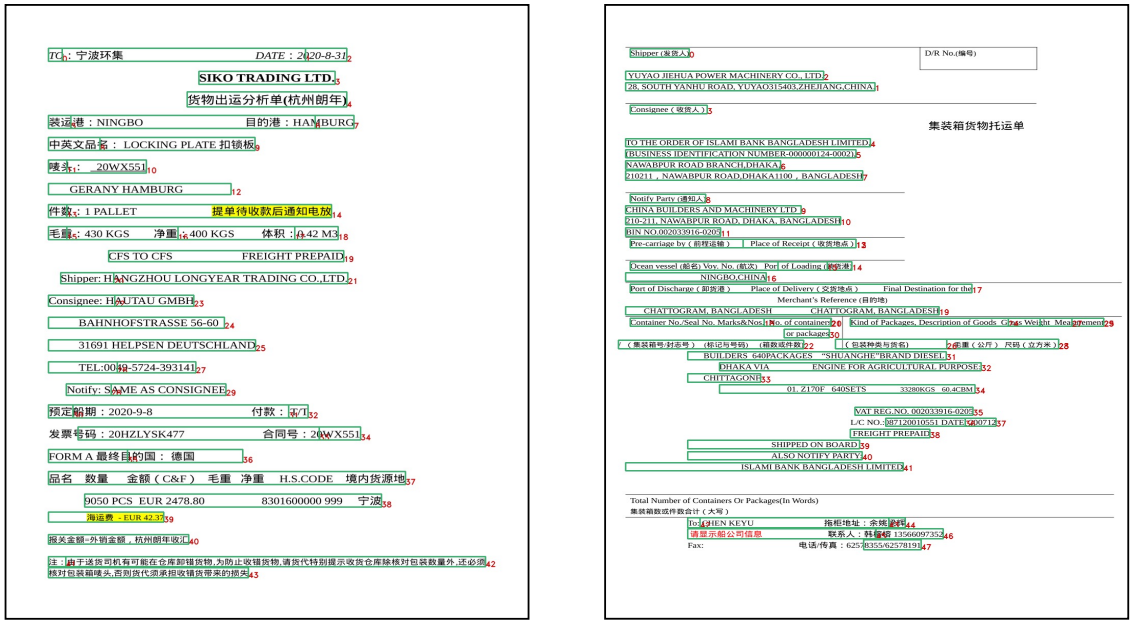


图 5.1 人类阅读文档的眼动顺序（左），排序算法构建出的阅读顺序（右）

图 5.2 人类阅读文档的眼动轨迹

一种解决方案是采用基于规则的方法，利用图像中的几何信息和排版规则来确定文本框的阅读顺序。这种方法需要先对图像进行分析，提取出文本框的位置和大

小等信息，然后利用排版规则和几何信息来推断文本框的阅读顺序。该方法的优点是准确性高，但需要针对不同类型的文档制定不同的规则，对于复杂的文档类型会带来一定的挑战。阅读顺序检测最早是在 [128] 提出的，他们使用定性矩形关系的命题语言从文档图像中检测阅读顺序。这也被认为是第一个基于规则的读取顺序检测系统。随着机器学习方法的发展，[129] 提出了一种使用贝叶斯框架的概率分类器，并重建布局组件的单个或多个链。同时，[130] 应用了 ILP 学习算法来引入两个谓词的定义并建立排序关系。之后，[131] 研究了使用主要特定知识检测逻辑结构组件之间的阅读顺序关系的问题。[132] 提出了一种基于抽象论证的无监督策略，用于识别文档页面组件的正确阅读顺序。该方法基于关于人类阅读文档时的行为的经验假设。

尽管 LayoutLM[6]、LayoutLMv2[7] 和 LayoutXLM[8] 等尝试已被用于以多模式方式处理文档理解，但它们仍面临两个限制：（1）它们依赖 OCR 工具中的标记和框，而不探索阅读顺序的影响。正确的读取顺序是指组织良好的可读 token 序列，它可能不是唯一的。直观地说，输入 token 的读取顺序对许多任务至关重要，如语言翻译 [80] 和 VQA[133]。例如，当我们打乱单词时，句子的意思可能会改变，从而导致语言翻译过程中的错误。常见的解决方案是使用位置嵌入来表示输入标记的顺序。然而，我们发现，广泛使用相对位置嵌入的多模态模型仍然存在不正确的读取顺序。属性读取顺序隐式地包括布局信息，这是 VRDU 任务中基本需要的。（2）它们通常利用 Transformer 中的固定长度的绝对或者相对位置嵌入。一旦模型被训练，它就不能处理具有较长 token 序列的测试数据。尽管位置嵌入扫描上的双线性插值可以应用于较长序列，但性能并不令人满意。最近，条件位置编码 CPE 被提出用于处理图像分类任务中可变长度的输入。它将输入标记重塑为 2D 特征，并使用卷积从输入标记中动态提取局部相邻上下文。然而，CPE 不能在多模态网络中使用到文档理解任务中的一维文本特征。

目前，一些工作集中在利用深度学习模型来进行文本框的阅读顺序预测。这种方法通常基于序列到序列的模型，通过学习文本框之间的相对位置关系来预测其阅读顺序。一些研究者还尝试将图像分割和目标检测技术与深度学习模型相结合，以提高文本框阅读顺序预测的准确性。[134] 提出了一种端到端 OCR 文本重组模型，其中他们使用具有注意力映射的图形神经网络来编码具有视觉布局特征的文本块，并使用基于注意力的序列解码器将 OCR 文本排序为适当的序列。[10] 提出了一个用于

读取顺序检测的基准数据集 ReadingBank，包含 500,000 个文档和图像。此外还提出了一种新的基于预训练的阅读顺序检测方法 LayoutReader。其显著优于从左到右、从上到下的启发式算法以及一些较强的基线，且可以很容易地适应任何 OCR 引擎，从而可以改善下游任务的读取顺序。[11] 提出 VRDU 中存在两个未充分探索的限制，即基于 Transformer 架构的文档预训练模型存在不正确的文档读取顺序和欠缺处理较长序列的能力，从而提出 Augmented XY Cut 算法对文本框进行排序生成合理的阅读顺序，从而改进模型性能。

5.3 阅读顺序构建算法介绍

如图 5.3（左）所示，是 XFUN 中文数据集中的一个样例所示，由于数据集中存在较多的内容折行，OCR 识别引擎很多时候不会按照会按从左到右，从上到下的阅读顺序顺序返回。如图中所示，OCR 输出的文本块的顺序为图上标识的数字（6、5、7、64...,16），显然这个顺序既不符合阅读顺序，也会打乱语义和排版信息。如果不进行调整，会导致模型无法准确的得到正确的语义信息和结构信息。经过阅读顺

图 5.3 OCR 出来的不正确的阅读顺序（左），经过阅读顺序构建算法构建出的人类阅读顺序（右）

序构建算法排序后的坐标顺序如图5.3（右）所示。文档阅读顺序构建算法流程图如图5.4所示。

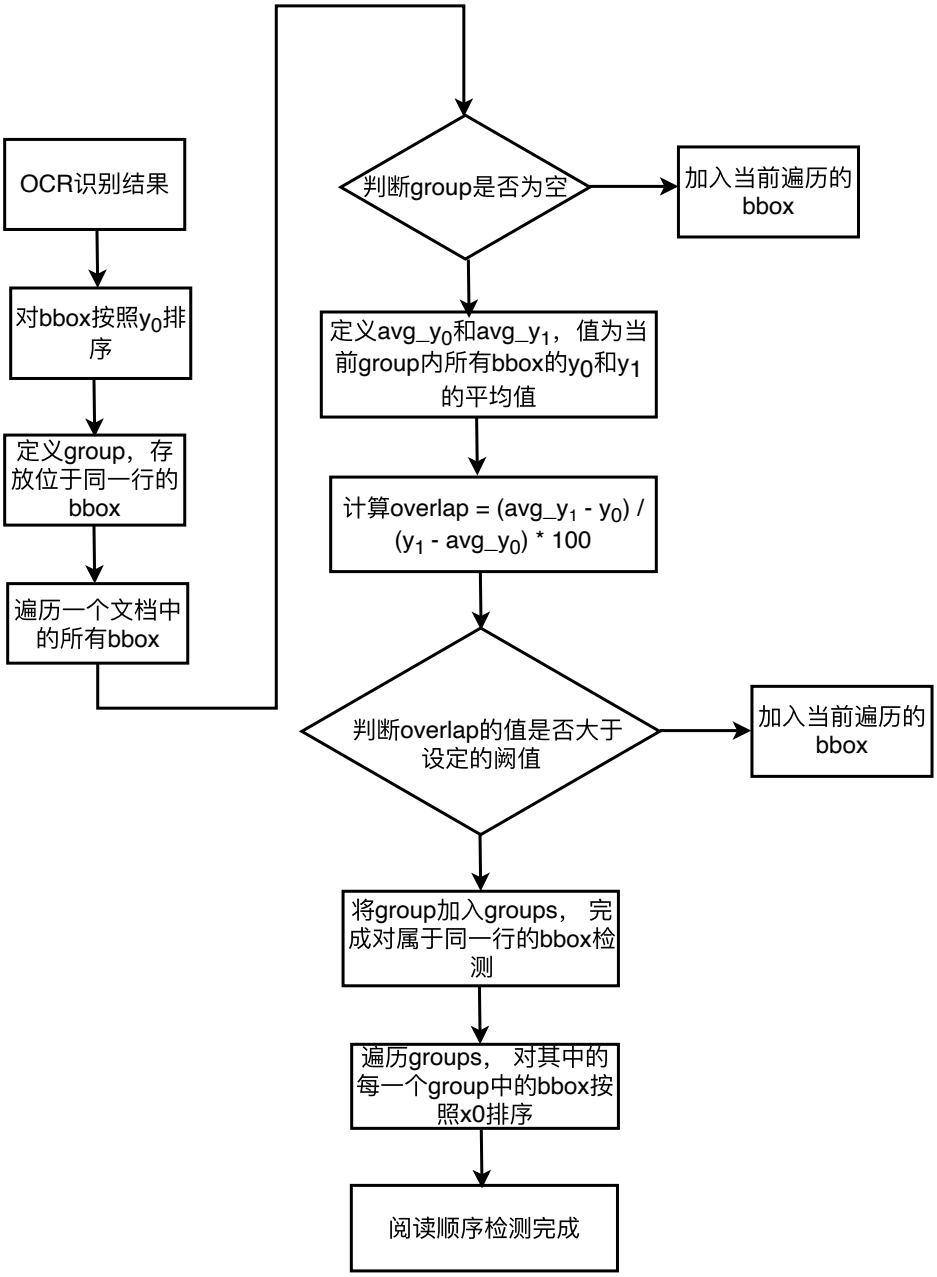


图 5.4　阅读顺序构建算法流程图

Algorithm 1: 人类文档阅读顺序构建算法

输入: OCR 识别得到的每个文本框 ($bbox$)，每个文本框的坐标为：(x_0, y_0, x_1, y_1)，文本框集合定义为 B

输出: 构建人类文档阅读顺序后的 $bbox$

将所 B 按照坐标 y_0 从小到大排序

定义 $group$ 和 $groups$，$group$ 中存放检测到的属于同一行的 $bbox$，$groups$ 中存放所有的 $group$

for B 中的每一个 $bbox$, **do**

  if $group$ 为空 **then**

    将当前遍历的 $bbox$ 加入 $group$

  end

  else

    计算当前 $group$ 中所有的 $bbox$ 的 y_0 和 y_1 的平均值，即为 avg_y_0，avg_y_1

    得到当前遍历的 bbox 的坐标 y_0 和 y_1, 计算 $overlap = (avg\_y_1 - y_0) / (y_1 - avg\_y_0) * 100$

    if $overlap$ > 设定的阈值 σ **then**

      将当前遍历的 $bbox$ 加入 $group$

    end

    else

      将 $group$ 加入 $groups$ 中，并将 bbox 的值赋值给 $gruop$

    end

  end

end

5.4 实验

5.4.1 数据集介绍

本章所用数据集第四章所用数据集相同，在这里不再介绍。

5.4.2 实验结果与分析

表 5.1 FUNSD 数据集，SER 和 RE 任务上，构建阅读顺序前后的 F1 值得分比较。

Method	**默认的阅读顺序**		**阅读顺序重建后**	
	SER	**RE**	**SER**	**RE**
BERT	49.60	-	51.13	-
BERT+BILSTM+CRF	51.62	-	52.96	-
BERT+ResNet	52.25	-	54.49	-
XLM-RoBERTa	66.57	25.66	67.10	48.22
LayoutLM	78.82	-	79.59	-
LayoutXLM	79.07	54.18	79.52	69.86
ours	80.15	55.06	80.77	72.76

表 5.2 XFUN 的 JA（日文）数据集，SER 和 RE 任务上，构建阅读顺序前后的 F1 值得分比较。

Method	**默认的阅读顺序**		**阅读顺序重建后**	
	SER	**RE**	**SER**	**RE**
BERT	44.03	-	45.44	-
BERT+BILSTM+CRF	45.45	-	45.69	-
BERT+ResNet	46.00	-	46.85	-
XLM-RoBERTa	77.63	52.89	79.68	60.25
LayoutLM	64.50	-	65.91	-
LayoutXLM	78.88	66.17	80.26	71.84
ours	79.35	67.76	81.63	75.85

表 5.3 XFUN 的 ES（西班牙文）数据集，SER 和 RE 任务上，构建阅读顺序前后的 F1 值得分比较。

Method	**默认的阅读顺序**		**阅读顺序重建后**	
	SER	**RE**	**SER**	**RE**
BERT	43.78	-	44.28	-
BERT+BILSTM+CRF	45.51	-	45.75	-
BERT+ResNet	44.74	-	46.48	-
XLM-RoBERTa	59.36	49.69	62.14	60.55
LayoutLM	66.41	-	71.39	-
LayoutXLM	75.19	67.79	76.25	76.45
ours	76.22	71.37	78.43	78.24

表 5.4　XFUN 的 FR（法文）数据集，SER 和 RE 任务上，构建阅读顺序前后的 F1 值得分比较。

Method	**默认的阅读顺序**		**阅读顺序重建后**	
	SER	**RE**	**SER**	**RE**
BERT	49.60	-	53.66	-
BERT+BILSTM+CRF	50.99	-	54.72	-
BERT+ResNet	50.16	-	55.65	-
XLM-RoBERTa	67.07	50.01	69.48	58.77
LayoutLM	74.93	-	80.12	-
LayoutXLM	78.84	62.47	79.37	71.20
ours	80.41	63.17	81.39	75.06

表 5.5　XFUN 的 IT（意大利文）数据集，SER 和 RE 任务上，构建阅读顺序前后的 F1 值得分比较。

Method	**默认的阅读顺序**		**阅读顺序重建后**	
	SER	**RE**	**SER**	**RE**
BERT	46.24	-	48.26	-
BERT+BILSTM+CRF	46.48	-	48.84	-
BERT+ResNet	47.28	-	50.25	-
XLM-RoBERTa	66.96	51.25	67.88	62.84
LayoutLM	74.51	-	79.11	-
LayoutXLM	80.06	63.56	80.50	77.15
ours	80.95	68.47	81.04	80.63

表 5.6　XFUN 的 DE（德文）数据集，SER 和 RE 任务上，构建阅读顺序前后的 F1 值得分比较。

Method	**默认的阅读顺序**		**阅读顺序重建后**	
	SER	**RE**	**SER**	**RE**
BERT	48.54	-	51.21	-
BERT+BILSTM+CRF	50.65	-	51.76	-
BERT+ResNet	53.02	-	53.81	-
XLM-RoBERTa	68.16	43.29	69.46	57.33
LayoutLM	73.49	-	74.43	-
LayoutXLM	79.50	62.94	80.14	67.28
ours	80.60	65.03	81.76	71.58

表 5.7　XFUN 的 PT（葡萄牙文）文数据集，SER 和 RE 任务上，构建阅读顺序前后的 F1 值得分比较。

Method	**默认的阅读顺序**		**阅读顺序重建后**	
	SER	**RE**	**SER**	**RE**
BERT	50.65	53.07	51.63	69.30
BERT+BILSTM+CRF	50.94	-	53.07	-
BERT+ResNet	52.00	-	54.52	-
XLM-RoBERTa	65.40	38.92	66.29	45.78
LayoutLM	69.42	-	72.80	-
LayoutXLM	78.24	56.18	79.41	67.43
ours	80.16	58.20	81.34	69.24

表 5.8　XFUN 的 ZH（中文）数据集，SER 和 RE 任务上，构建阅读顺序前后的 F1 值得分比较。

Method	**默认的阅读顺序**		**阅读顺序重建后**	
	SER	**RE**	**SER**	**RE**
BERT	48.10	-	52.65	-
BERT+BILSTM+CRF	49.29	-	53.67	-
BERT+ResNet	52.19	-	54.25	-
XLM-RoBERTa	65.26	49.28	68.74	57.10
LayoutLM	69.88	-	74.88	-
LayoutXLM	86.85	69.56	90.00	74.84
ours	87.96	70.36	91.15	82.05

通过在不同数据集上，使用不同的文档编码器编码文档，其中多语言 BERT 模型和 XLM-RoBERTaBASE 只用到了文档的文本特征，BERT+Resnet 使用了文本和图形特征。其余模型同时使用了文本、视觉、布局特征。通过构建阅读顺序，与默认 OCR 识别出的结果相对比，在各个数据集上和模型上的实验结果均有较大的提升。从而验证了所构建的阅读顺序算法的有效性。

5.5　本章小结

本章针对视觉富文档中 OCR 结果的阅读顺序问题进行了研究。针对视觉富文档版面结构复杂多样，OCR 结果难以表示正确的阅读顺序，提出了一种针对 bounding box 坐标进行排序的算法以构建正确的阅读顺序的方法，弥补了 OCR 识别的结果的

部分缺陷。在多语言版本的 XFUN 数据集，小票收据数据集已经自建的海运单数据集上验证了才所提出的阅读顺序构建算法的有效性。

第六章　总结与展望

6.1　总结

随着国际贸易的发展和信息技术的快速进步，我们现在面对着大量的视觉富文档数据，如物流表单、收据和简历等。这些文档中包含了丰富的信息，包括行业相关的实体和数字信息等，这些信息对于经济和社会效益的提高至关重要。然而，手动提取这些关键信息需要大量的时间和精力，因此，实现自动化信息抽取技术对于提高效率和降低成本至关重要。以国际物流单据文档为例，自动化信息抽取技术可以帮助我们快速准确地提取文档中的关键信息，如货物名称、数量、价格和发货地点等，从而大大提高我们的工作效率和准确性。视觉富文档的格式复杂、种类繁多，为了探索通用的视觉富文档关键信息识别的方法，本文以国际货运单、文档扫描件等数据及为例，并结合图像特征、文本内容和文档布局特点，研究了针对不同类别视觉富文档的视觉信息抽取模型。除此之外，探究了 OCR 的文档识别顺序对关键信息抽取的影响，同时提出一种简单有效的排序算法，根据 bounding box 的坐标构建出正确的阅读顺序。

本文的主要工作和贡献包括以下几个方面：

贡献一：视觉富文档中蕴含着丰富的视觉特征，例如字体大小、颜色、样式等。之前的一些模型只用到了文本特征进行关键信息抽取，使得模型理解文档的能力大大受限。此外，文档中的布局信息也对文档的理解至关重要。所以本文将文本、视觉、布局三种模态的特征进行有效融合，大大提升了模型理解视觉富文档的能力。

贡献二：为了解决基于 Transformer 架构的文档预训练模型只使用文档中基于 token 的细粒度元素，缺少对文档中粗粒度元素学习的问题，对于 SER 任务，本文提出了一种新的同时结合粗细粒度的图神经网络键值匹配模型，我们建立 bounding box 粒度的图注意力网络，学习文档中的粗粒度元素。通过基于 bounding box 级别的图关联建模，更关注 bounding box 键值对之间的关系。此外，我们引入了 token 级别的图注意力网络 (Graph Attention Network)，进一步增强 bounding box 内部 token 之间的信息交互。。对于 RE 任务，提出一种基于 SPAN 图关联的视觉富文档关系抽取方法，可以有效弥补基于 Transformer 架构的预训练文档模型使用每个 bounding box 的

单个 token 用于关系抽取中，存在的单个 token 难以表示整个 bounding box 信息的问题。综合实验表明，所提出的模型和方法可以显著优于以前的方法。

贡献三：为了解决视觉富文档阅读顺序难以正确构建的问题，提出了一种针对 bounding box 坐标进行排序的算法。该算法可以将 OCR 识别出的非正常阅读顺序，按照 bounding box 的坐标位置信息进行排序，构建出正常的阅读顺序。实验表明，所构建出来的阅读顺序，对于 SER 和 RE 任务的效果都有很大的提升。

6.2 展望

本文针对财经类文档的信息抽取问题进行了研究并提出了相应模型和方案，但依旧存在诸多不足和可以改进之处：

（1）本文在进行视觉富文档关键信息抽取时，虽然利用了文本、视觉、布局三种模态的特征。但是视觉特征中还有很多没有利用的特征，例如字体的颜色和大小。如果能融入更多的视觉特征，相信能进一步提高视觉富文档的理解能力。

（2）本文在进行文本、视觉、布局三种模态的信息进行融合时，采用的是直接拼接的方式。但是有没有更好的方式进行三种模态的融合，比如使用注意的方式，是接下来需要继续开展的工作。

插图索引

表格索引

参考文献

[1] ORAL B, EMEKLIGIL E, ARSLAN S, et al. Information extraction from text intensive and visually rich banking documents[J]. Information Processing & Management, 2020, 57(6): 102361.

[2] 晏文坛. 半结构化中文简历的信息抽取[D]. 华南理工大学.

[3] 杨茜. 基于视觉特征的多类型表单关键信息识别研究[D]. 北京交通大学.

[4] LIU X, GAO F, ZHANG Q, et al. Graph convolution for multimodal information extraction from visually rich documents[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). 2019: 32-39.

[5] ZHAO X, NIU E, WU Z, et al. Cutie: Learning to understand documents with convolutional universal text information extractor[A]. 2019.

[6] XU Y, LI M, CUI L, et al. Layoutlm: Pre-training of text and layout for document image understanding[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1192-1200.

[7] XU Y, XU Y, LV T, et al. 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 (Volume 1: Long Papers). 2021: 2579-2591.

[8] XU Y, LV T, CUI L, et al. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding[A]. 2021.

[9] TANG G, XIE L, JIN L, et al. Matchvie: Exploiting match relevancy between entities for visual information extraction[A]. 2021.

[10] WANG Z, XU Y, CUI L, et al. Layoutreader: Pre-training of text and layout for reading order detection[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 4735-4744.

[11] GU Z, MENG C, WANG K, et al. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4583-4592.

[12] JAUME G, EKENEL H K, THIRAN J P. Funsd: A dataset for form understanding in noisy scanned documents[C]//2019 International Conference on Document Analysis and Recognition Workshops (ICDARW): volume 2. IEEE, 2019: 1-6.

[13] YU W, LU N, QI X, et al. Pick: processing key information extraction from documents using improved graph learning-convolutional networks[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 4363-4370.

[14] ZHANG Y, BO Z, WANG R, et al. Entity relation extraction as dependency parsing in visually rich documents[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 2759-2768.

[15] CARBONELL M, RIBA P, VILLEGAS M, et al. Named entity recognition and relation extraction with graph neural networks in semi structured documents[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 9622-9627.

[16] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[17] TIAN Z, HUANG W, HE T, et al. Detecting text in natural image with connectionist text proposal network[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer, 2016: 56-72.

[18] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016: 21-37.

[19] LONG S, RUAN J, ZHANG W, et al. Textsnake: A flexible representation for detecting text of arbitrary shapes[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 20-36.

[20] SHI B, BAI X, BELONGIE S. Detecting oriented text in natural images by linking segments[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2550-2558.

[21] ZHOU X, YAO C, WEN H, et al. East: an efficient and accurate scene text detector[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 5551-5560.

[22] DENG D, LIU H, LI X, et al. Pixellink: Detecting scene text via instance segmentation[C]// Proceedings of the AAAI conference on artificial intelligence: volume 32. 2018.

[23] WANG W, XIE E, LI X, et al. Shape robust text detection with progressive scale expansion network [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9336-9345.

[24] BAEK Y, LEE B, HAN D, et al. Character region awareness for text detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9365-9374.

[25] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd international conference on Machine learning. 2006: 369-376.

[26] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.

[27] LONG S, HE X, YAO C. Scene text detection and recognition: The deep learning era[J]. International Journal of Computer Vision, 2021, 129: 161-184.

[28] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11): 2298-2304.

[29] CHENG Z, BAI F, XU Y, et al. Focusing attention: Towards accurate text recognition in natural images[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5076-5084.

[30] LIAO M, ZHANG J, WAN Z, et al. Scene text recognition from two-dimensional perspective[C]// Proceedings of the AAAI conference on artificial intelligence: volume 33. 2019: 8714-8721.

[31] GRAVES A, LIWICKI M, BUNKE H, et al. Unconstrained on-line handwriting recognition with recurrent neural networks[J]. Advances in neural information processing systems, 2007, 20.

[32] SU B, LU S. Accurate scene text recognition based on recurrent neural network[C]//Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I 12. Springer, 2015: 35-48.

[33] HE P, HUANG W, QIAO Y, et al. Reading scene text in deep convolutional sequences[C]// Proceedings of the AAAI conference on artificial intelligence: volume 30. 2016.

[34] LIU W, CHEN C, WONG K Y K, et al. Star-net: a spatial attention residue network for scene text recognition.[C]//BMVC: volume 2. 2016: 7.

[35] GAO Y, CHEN Y, WANG J, et al. Reading scene text with attention convolutional sequence modeling[A]. 2017.

[36] YIN F, WU Y C, ZHANG X Y, et al. Scene text recognition with sliding convolutional character models[A]. 2017.

[37] LEE C Y, OSINDERO S. Recursive recurrent nets with attention modeling for ocr in the wild[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2231-2239.

[38] O'GORMAN L. The document spectrum for page layout analysis[J]. IEEE Transactions on pattern analysis and machine intelligence, 1993, 15(11): 1162-1173.

[39] JOURNET N, EGLIN V, RAMEL J Y, et al. Text/graphic labelling of ancient printed documents [C]//Eighth International Conference on Document Analysis and Recognition (ICDAR'05). IEEE, 2005: 1010-1014.

[40] MARINAI S, GORI M, SODA G. Artificial neural networks for document analysis and recognition [J]. IEEE Transactions on pattern analysis and machine intelligence, 2005, 27(1): 23-35.

[41] KISE K, SATO A, IWATA M. Segmentation of page images using the area voronoi diagram[J]. Computer Vision and Image Understanding, 1998, 70(3): 370-382.

[42] ESKENAZI S, GOMEZ-KRÄMER P, OGIER J M. The delaunay document layout descriptor[C]// Proceedings of the 2015 ACM Symposium on Document Engineering. 2015: 167-175.

[43] SAABNI R, ASI A, EL-SANA J. Text line extraction for historical document images[J]. Pattern Recognition Letters, 2014, 35: 23-33.

[44] ALAEI A, PAL U, NAGABHUSHAN P. A new scheme for unconstrained handwritten text-line segmentation[J]. Pattern Recognition, 2011, 44(4): 917-928.

[45] SHAFAIT F, BREUEL T M. The effect of border noise on the performance of projection-based page segmentation methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(4): 846-851.

[46] SHAFAIT F, VAN BEUSEKOM J, KEYSERS D, et al. Background variability modeling for statistical layout analysis[C]//2008 19th International Conference on Pattern Recognition. IEEE, 2008: 1-4.

[47] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[48] CHEN K, SEURET M, HENNEBERT J, et al. Convolutional neural networks for page segmentation of historical document images[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): volume 1. IEEE, 2017: 965-970.

[49] RAVÌ D, BOBER M, FARINELLA G M, et al. Semantic segmentation of images exploiting dct based features and random forest[J]. Pattern Recognition, 2016, 52: 260-273.

[50] HE D, COHEN S, PRICE B, et al. Multi-scale multi-task fcn for semantic page segmentation and table detection[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): volume 1. IEEE, 2017: 254-261.

[51] WICK C, PUPPE F. Fully convolutional neural networks for page segmentation of historical document images[C]//2018 13th IAPR International Workshop on Document Analysis Systems (DAS). IEEE, 2018: 287-292.

[52] LI Y, ZOU Y, MA J. Deeplayout: A semantic segmentation approach to page layout analysis[C]// Intelligent Computing Methodologies: 14th International Conference, ICIC 2018, Wuhan, China, August 15-18, 2018, Proceedings, Part III 14. Springer, 2018: 266-277.

[53] CHEN D, CAO X, WEN F, et al. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3025-3032.

[54] WU X, HU Z, DU X, et al. Document layout analysis via dynamic residual feature fusion[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.

[55] LI K, WIGINGTON C, TENSMEYER C, et al. Cross-domain document object detection: Benchmark suite and method[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12915-12924.

[56] LI X H, YIN F, XUE T, et al. Instance aware document image segmentation using label pyramid networks and deep watershed transformation[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019: 514-519.

[57] CHEN K, SEURET M, LIWICKI M, et al. Page segmentation of historical document images with convolutional autoencoders[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015: 1011-1015.

[58] GATOS B, LOULOUDIS G, STAMATOPOULOS N. Segmentation of historical handwritten documents into text zones and text lines[C]//2014 14th International Conference on Frontiers in Handwriting Recognition. IEEE, 2014: 464-469.

[59] LEE J, HAYASHI H, OHYAMA W, et al. Page segmentation using a convolutional neural network with trainable co-occurrence features[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019: 1023-1028.

[60] VO Q N, LEE G. Dense prediction for text line segmentation in handwritten document images[C]// 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016: 3264-3268.

[61] ZAGORIS K, PRATIKAKIS I, GATOS B. Segmentation-based historical handwritten word spotting using document-specific local features[C]//2014 14th International Conference on Frontiers in Handwriting Recognition. IEEE, 2014: 9-14.

[62] SIEGEL N, LOURIE N, POWER R, et al. Extracting scientific figures with distantly supervised neural networks[C]//Proceedings of the 18th ACM/IEEE on joint conference on digital libraries. 2018: 223-232.

[63] CONWAY A. Page grammars and page parsing. a syntactic approach to document layout recognition[C]//Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93). IEEE, 1993: 761-764.

[64] KRISHNAMOORTHY M, NAGY G, SETH S, et al. Syntactic segmentation and labeling of digitized pages from technical journals[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, 15(7): 737-747.

[65] SHILMAN M, LIANG P, VIOLA P. Learning nongenerative grammatical models for document analysis[C]//Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1: volume 2. IEEE, 2005: 962-969.

[66] AGGARWAL M, SARKAR M, GUPTA H, et al. Multi-modal association based grouping for form structure extraction[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 2075-2084.

[67] LI M, XU Y, CUI L, et al. Docbank: A benchmark dataset for document layout analysis[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020: 949-960.

[68] YANG X, YUMER E, ASENTE P, et al. 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: 5315-5324.

[69] BARMAN R, EHRMANN M, CLEMATIDE S, et al. Combining visual and textual features for semantic segmentation of historical newspapers[J]. Journal of Data Mining & Digital Humanities, 2021, 2021(ARTICLE): 1-26.

[70] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[Z]. 2018.

[71] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[M]. OpenAI, 2018.

[72] KENTON J D M W C, TOUTANOVA L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT. 2019: 4171-4186.

[73] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[74] MEDSKER L R, JAIN L. Recurrent neural networks[J]. Design and Applications, 2001, 5: 64-67.

[75] GRAVES A. Long short-term memory[J]. Supervised sequence labelling with recurrent neural networks, 2012: 37-45.

[76] CHITICARIU L, LI Y, REISS F. Rule-based information extraction is dead! long live rule-based information extraction systems![C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 827-832.

[77] DENGEL A R, KLEIN B. smartfix: A requirements-driven system for document analysis and understanding[C]//Document Analysis Systems V: 5th International Workshop, DAS 2002 Princeton, NJ, USA, August 19–21, 2002 Proceedings 5. Springer, 2002: 433-444.

[78] SCHUSTER D, MUTHMANN K, ESSER D, et al. Intellix–end-user trained information extraction for document archiving[C]//2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013: 101-105.

[79] DAI Z, YANG Z, YANG Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2978-2988.

[80] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[81] LOHANI D, BELAÏD A, BELAÏD Y. An invoice reading system using a graph convolutional network[C]//Asian Conference on Computer Vision. Springer, 2018: 144-158.

[82] QIAN Y, SANTUS E, JIN Z, et al. Graphie: A graph-based framework for information extraction [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 751-761.

[83] 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.

[84] LUO C, WANG Y, ZHENG Q, et al. Merge and recognize: a geometry and 2d context aware graph model for named entity recognition from visual documents[C]//Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs). 2020: 24-34.

[85] ZHANG P, XU Y, CHENG Z, et al. 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.

[86] WANG J, LIU C, JIN L, et al. Towards robust visual information extraction in real world: new dataset and novel solution[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 2738-2745.

[87] KATTI A R, REISSWIG C, GUDER C, et al. Chargrid: Towards understanding 2d documents [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4459-4469.

[88] DANG T A N, NGUYEN D T. End-to-end information extraction by character-level embedding and multi-stage attentional u-net[A]. 2021.

[89] DENK T I, REISSWIG C. Bertgrid: Contextualized embedding for 2d document representation and understanding[C]//Workshop on Document Intelligence at NeurIPS 2019. 2019.

[90] KERROUMI M, SAYEM O, SHABOU A. Visualwordgrid: Information extraction from scanned documents using a multimodal approach[C]//International Conference on Document Analysis and Recognition. Springer, 2021: 389-402.

[91] LIN W, GAO Q, SUN L, et al. Vibertgrid: a jointly trained multi-modal 2d document representation for key information extraction from documents[C]//International Conference on Document Analysis and Recognition. Springer, 2021: 548-563.

[92] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep learning based natural language processing[J]. ieee Computational intelligenCe magazine, 2018, 13(3): 55-75.

[93] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[94] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of NAACL-HLT. 2016: 260-270.

[95] DO H H, PRASAD P, MAAG A, et al. Deep learning for aspect-based sentiment analysis: a comparative review[J]. Expert systems with applications, 2019, 118: 272-299.

[96] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50-70.

[97] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[A]. 2019.

[98] PALM R B, WINTHER O, LAWS F. Cloudscan-a configuration-free invoice analysis system using recurrent neural networks[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): volume 1. IEEE, 2017: 406-413.

[99] SAGE C, AUSSEM A, ELGHAZEL H, et al. Recurrent neural network approach for table field extraction in business documents[C]//International Conference on Document Analysis and Recognition. 2019.

[100] HWANG W, KIM S, SEO M, et al. Post-ocr parsing: building simple and robust parser via bio tagging[C]//Workshop on Document Intelligence at NeurIPS 2019. 2019.

[101] PRAMANIK S, MUJUMDAR S, PATEL H. Towards a multi-modal, multi-task learning based pre-training framework for document representation learning[A]. 2020.

[102] HUANG Y, LV T, CUI L, et al. Layoutlmv3: Pre-training for document ai with unified text and image masking[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 4083-4091.

[103] WANG J, JIN L, DING K. Lilt: A simple yet effective language-independent layout transformer for structured document understanding[A]. 2022.

[104] SIMON A, PRET J C, JOHNSON A P. A fast algorithm for bottom-up document layout analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(3): 273-277.

[105] REIMERS N, GUREVYCH I. Sentence-bert: Sentence embeddings using siamese bert-networks [A]. 2019.

[106] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, et al. Conditional random fields as recurrent neural networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1529-1537.

[107] DU Y, LI C, GUO R, et al. Pp-ocr: A practical ultra lightweight ocr system[A]. 2020.

[108] ZHONG X, TANG J, YEPES A J. Publaynet: largest dataset ever for document layout analysis[C]// 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019: 1015-1022.

[109] LI M, CUI L, HUANG S, et al. Tablebank: Table benchmark for image-based table detection and recognition[C]//Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 1918-1925.

[110] MATHEW M, KARATZAS D, JAWAHAR C. Docvqa: A dataset for vqa on document images [C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 2200-2209.

[111] GARNCAREK Ł, POWALSKI R, STANISŁAWEK T, et al. Lambert: Layout-aware language modeling for information extraction[C]//Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I. Springer, 2021: 532-547.

[112] QIAN Y. A graph-based framework for information extraction[D]. Massachusetts Institute of Technology, 2019.

[113] FENG J, HUANG M, ZHAO L, et al. Reinforcement learning for relation classification from noisy data[C]//Proceedings of the aaai conference on artificial intelligence: volume 32. 2018.

[114] ZENG D, LIU K, CHEN Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 1753-1762.

[115] ZHANG N, CHEN X, XIE X, et al. Document-level relation extraction as semantic segmentation [A]. 2021.

[116] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]// Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. 2014: 2335-2344.

[117] JIANG X, WANG Q, LI P, et al. Relation extraction with multi-instance multi-label convolutional neural networks[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 1471-1480.

[118] LIU X, LUO Z, HUANG H. Jointly multiple events extraction via attention-based graph information aggregation[A]. 2018.

[119] DAI Z, YANG Z, YANG Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[A]. 2019.

[120] DAVIS B, MORSE B, PRICE B, et al. Visual fudge: Form understanding via dynamic graph editing [C]//Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16. Springer, 2021: 416-431.

[121] LI X, ZHENG Y, HU Y, et al. Relational representation learning in visually-rich documents[C]// Proceedings of the 30th ACM International Conference on Multimedia. 2022: 4614-4624.

[122] GIANNIS B, JOHANNES D, THOMAS D, et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. Expert Systems with Application, 2018, 114(DEC.): 34-45.

[123] CECI M, BERARDI M, PORCELLI G, et al. A data mining approach to reading order detection[C]//Ninth International Conference on Document Analysis and Recognition (ICDAR 2007): volume 2. IEEE, 2007: 924-928.

[124] FERILLI S, GRIECO D, REDAVID D, et al. Abstract argumentation for reading order detection [C]//Proceedings of the 2014 ACM symposium on Document engineering. 2014: 45-48.

[125] LI L, GAO F, BU J, et al. An end-to-end ocr text re-organization sequence learning for rich-text detail image comprehension[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. Springer, 2020: 85-100.

[126] MALERBA D, CECI M, BERARDI M. Machine learning for reading order detection in document image understanding[J]. Machine learning in document analysis and recognition, 2008: 45-69.

[127] MALERBA D, CECI M. Learning to order: A relational approach[C]//Mining Complex Data: ECML/PKDD 2007 Third International Workshop, MCD 2007, Warsaw, Poland, September 17-21, 2007, Revised Selected Papers 3. Springer, 2008: 209-223.

[128] AIELLO M, SMEULDERS A. Bidimensional relations for reading order detection[Z]. 2003.

[129] CECI M, BERARDI M, PORCELLI G, et al. A data mining approach to reading order detection[C]//Ninth International Conference on Document Analysis and Recognition (ICDAR 2007): volume 2. IEEE, 2007: 924-928.

[130] MALERBA D, CECI M. Learning to order: A relational approach[C]//Mining Complex Data: ECML/PKDD 2007 Third International Workshop, MCD 2007, Warsaw, Poland, September 17-21, 2007, Revised Selected Papers 3. Springer, 2008: 209-223.

[131] MALERBA D, CECI M, BERARDI M. Machine learning for reading order detection in document image understanding[J]. Machine learning in document analysis and recognition, 2008: 45-69.

[132] FERILLI S, GRIECO D, REDAVID D, et al. Abstract argumentation for reading order detection [C]//Proceedings of the 2014 ACM symposium on Document engineering. 2014: 45-48.

[133] YU Z, YU J, CUI Y, et al. Deep modular co-attention networks for visual question answering [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6281-6290.

[134] LI L, GAO F, BU J, et al. An end-to-end ocr text re-organization sequence learning for rich-text detail image comprehension[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. Springer, 2020: 85-100.

作者在攻读硕士学位期间发表的论文与研究成果

发表的学术论文

1. Junwei Zhang, Hao Wang, Xiangfeng Luo. Dual-VIE: Dual-Level Graph Attention Network for Visual Information Extraction, Proceeding of the 19th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2022, Shanghai, China, November 10–13, 2022, 422–434, Springer. (CCF-C)
2. 面向国际物流领域的视觉富文档数据集构建及信息抽取方法研究, CCKS

作者在攻读硕士学位期间所作的项目

1. 项目来源：xxx

   项目名称：xxx

   项目编号：xxx

   执行期限：2018.6-2021.05

致　谢

在我完成硕士学位论文之际，我想表达我最深切的谢意和感激之情。

首先，我要感谢我的导师，您在整个学习过程中一直支持我、鼓励我、引导我。感谢您对我的指导和建议，让我在研究方向上更加明确、深入。您对我认真、严谨、负责的态度深深地影响了我，也为我的学术道路奠定了坚实的基础。

同时，我还要感谢我的同学们，感谢你们一直以来的支持和帮助。在这里我要特别感谢那些给我提供了实验平台、实验数据和实验设备的同学，你们的支持使我得以顺利地完成我的研究工作。

最后，我要感谢我的家人。感谢你们对我的支持和理解，是你们一直支持我、鼓励我，让我有勇气和毅力去追求我的梦想。

完成硕士学位论文是一次艰苦的历程，但也是我人生中的一次宝贵经历。在这个过程中，我深刻地感受到了知识的重要性和持续不断的努力所带来的成就感。在未来的日子里，我将继续学习、努力工作，将所学所得转化为实际行动，为社会的发展和人类的进步贡献自己的力量。

我在这次研究中，深刻地认识到了自己的不足和需要改进的地方，同时也明确了自己的优势和擅长之处。这是一次对我的学术研究和个人成长都有很大帮助的经历。在未来的工作和学习中，我会更加注重细节、更加认真负责地对待每一个任务，同时也会不断地开拓自己的视野，学习更多的知识和技能。

我相信，在未来的工作中，我将不断挑战自己、超越自己，用自己的实际行动证明自己的能力和价值。在此，我要向所有支持和帮助过我的人致以最诚挚的谢意，祝愿大家生活愉快、工作顺利！