中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：21721584

上海大学

硕士学位论文

SHANGHAI UNIVERSITY MASTER'S DISSERTATION

MASTER'S DISSERTATION

题目	面向视觉富文档信息抽取任务的知识迁移方法研究

作　　者　李唐
学科专业　计算机应用技术
导　　师　王昊
完成日期　二〇二四年四月

姓　　名：李唐　　学号：21721584

论文题目：面向视觉富文档信息抽取任务的知识迁移方法研究

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:　　年　　月　　日

姓　　名：李唐　　学号：21721584

论文题目：面向视觉富文档信息抽取任务的知识迁移方法研究

上海大学学位论文原创性声明

本人郑重声明: 所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，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研究者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。

学位论文作者签名:

日期:　　年　　月　　日

上海大学学位论文使用授权说明

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)

学位论文作者签名:　　导师签名:

日期:　　年　　月　　日　　日期:　　年　　月　　日

上海大学工学硕士学位论文

面向视觉富文档信息抽取任务的知识迁移方法研究

作　　者: 李唐

导　　师: 王昊

学科专业: 计算机应用技术

计算机工程与科学学院

上海大学

2024 年 4 月

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

Research on Knowledge Transfer Approaches for Visually-rich Document Information Extraction

Candidate: Tang Li
Supervisor: Hao Wang
Major: Technology of Computer Application

School of Computer Engineering and Science
Shanghai University
April, 2024

摘　要

在企业数字化转型的推动下，各种视觉富文档如发票、收据、物流订单以及飞行器参数文档等呈爆发式增长。各行业有极大的需求对这些文档进行自动化的处理。尽管不同类型文档资源存在一定的差异性，考虑到相同或者类似领域的文档布局和格式具有一定相似性和相通性，开发高效的知识迁移方法，将现有文档理解模型有效地应用于新型文档的信息抽取任务，既可以降低开发成本，又可以提供重要的技术支持。本文致力于探索和研究视觉富文档信息抽取任务中的知识迁移方法，取得了以下成果：

（1）**基于跨粒度表示学习的视觉富文档关系抽取研究**：尽管视觉富文档中的语义实体大多以短语、文本框、图像块等粗粒度级别的形式出现，已有面向视觉富文档信息抽取任务的预训练语言模型，只关注文档内 token 这一细粒度层面的语义单元，而忽略了更加粗粒度的文本框单元。本研究针对以上问题，以细粒度预训练语言模型为基础，研究如何将细粒度这种丰富的信息迁移到粗粒度的语义实体中。在对粗粒度单元建模时，建立带有局部注意力机制的图神经网络，促进粗粒度的语义信息更好地被细粒度语言模型捕捉。通过多粒度、多任务的学习，提出方法在公开数据集上取得了最优的性能。

（2）**基于跨类别元学习的视觉富文档关系抽取研究**：考虑到视觉富文档自动化处理面临较多的低资源场景，如何仅依靠少量的样本，识别并抽取出新的键值实体以及它们间的关系是目前面临的又一大挑战。针对以上问题，本研究提出了一种面向跨类别关系抽取任务的元学习方法，创新性地构建了两个专门针对少样本学习场景的新型基准数据集，并提出了一种新颖的变分原型修正表示方法，该方法不仅融合二维空间的键值关系先验知识，还会降低原型表示的方差，并在多个少样本视觉富文档关系抽取数据集上进行了广泛评估，实验结果表明提出的方法达到了最优的性能，证明了该方法的有效性。

（3）**基于跨模型知识嵌入的视觉富文档表格问答研究**：大型语言模型（LLM）的发展使得通过自然语言输入来与表格数据交互成为可能。然而，现有模型对

于文档中的表格理解较为困难，准确性不高。本研究致力于视觉富文档的表格理解，通过利用现有表格解析模型的表格理解能力与 XML 中间语言，将表格知识通过 XML 代码的形式加入到提示模版中，探索 LLM 理解复杂表格的能力。但是由于研究时间有限，对小模型表格理解能力的评测以及对表格的操作指令部分还没有深入研究。

综上所述，本研究从跨粒度、跨类别、跨模型三个不同的角度，探索了视觉富文档信息抽取任务中的可行的知识迁移方法，为未来相关研究指明了道路。

关键词：视觉富文档，知识迁移，元学习，图神经网络，关系抽取，表格识别

ABSTRACT

Driven by the digital transformation of enterprises, there has been an explosive growth in various types of Visually-rich Documents such as invoices, receipts, logistics orders, and aircraft parameter documents. There is a significant demand across industries for the automated processing of these documents. Despite certain differences among document resources of various types, considering that documents in the same or similar domains often share similarities and commonalities in layout and format, the development of efficient knowledge transfer methods can effectively apply existing document understanding models to the information extraction tasks of new types of documents. This approach can not only reduce development costs but also provide important technical support. This paper is dedicated to exploring and studying knowledge transfer methods in the task of information extraction from Visually-rich Documents, and has achieved the following results:

(1) **Research on Relation Extraction from Visually-rich Documents Based on Cross-Granularity Representation Learning**:

Despite the fact that semantic entities in Visually-Rich Documents are predominantly presented at a coarse-grained level, such as phrases, text boxes, and image blocks, the pre-trained language models designed for information extraction tasks in such documents have been primarily concerned with semantic units at the fine-grained token level, thereby overlooking the coarser-grained text box units. This research tackles the aforementioned problem by leveraging fine-grained pre-trained language models as a foundation to investigate the transference of the rich information available at the fine-grained level to the coarse-grained semantic entities. A graph neural network with a local attention mechanism is constructed for modeling the coarse-grained units, which enhances the fine-grained language models' ability to capture the semantic information at a coarser granularity. Through learning across multiple granularities and

tasks, the proposed approach has demonstrated superior performance on publicly available datasets.

(2) **Research on Relation Extraction from Visually-rich Documents Based on Cross-Class Meta-Learning**: The meta-learning approach for cross-class relation extraction is aimed at low-resource scenarios, with the goal of relying on a small number of samples to extract new key-value entities and the relationships between them. The specific work is as follows: First, this study, for the first time, proposes a solution to the few-shot learning problem in the field of visual-rich document relation extraction. To promote further research in this field, we have innovatively constructed two new benchmark datasets specifically for few-shot learning scenarios. Second, this research proposes a novel variational approach that not only integrates spatial prior knowledge but also ensures the robustness of class-agnostic features, thereby reducing the variance of prototype representation. Third, our proposed approach achieves optimal performance in the few-shot visual-rich document relation extraction task and is extensively evaluated on the constructed datasets to prove its effectiveness.

(3) **Research on Table Question Answering from Visually-rich Documents Based on Cross-Model Knowledge Embedding**: The development of Large Language Models (LLMs) has made it possible to interact with tabular data through natural language input. However, existing models still struggle with understanding tables within documents, and their accuracy is not high. This study is dedicated to the understanding of tables in Visually-rich Documents. By leveraging the table comprehension capabilities of existing table parsing models and an intermediate XML language, the research explores the ability of LLMs to understand complex tables by injecting table knowledge into LLM inputs in the form of XML code. However, due to limited research time, an in-depth study of the table comprehension capabilities of smaller models and the part concerning table operation instructions has not yet been conducted.

Keywords: Visually-rich document; knowledge transfer; meta learning; graph neural network; relation extraction; table recognition

目　录

摘　要 I

ABSTRACT III

第一章　绪论 1

1.1　研究背景与意义 1

1.2　研究问题 3

1.3　研究内容 4

1.4　研究创新点 6

1.5　研究的组织结构 8

第二章　视觉富文档理解及知识迁移的相关方法 10

2.1　视觉富文档理解方法的研究现状 10

2.1.1　基于启发式规则的文档理解方法 11

2.1.2　基于统计机器学习的文档理解方法 13

2.1.3　基于深度学习的文档理解模型 14

2.2　知识迁移方法 21

2.2.1　知识迁移的概念和发展 21

2.2.2　迁移学习的主要方法 22

2.2.3　迁移学习的三种设定 22

2.2.4　迁移学习的挑战和本研究关注重点 25

第三章　基于跨粒度表示学习的视觉富文档关系抽取研究 27

3.1　研究动机 27

3.2　相关工作 29

3.3　任务定义 31

3.4　提出方法 31

3.4.1　整体框架 31

3.4.2 加入特殊分隔字符（Adding Special Tokens）.....31
3.4.3 多粒度图神经网络（Multi-grained Graph Neural Network）...35
3.4.4 关系分类器（Relation Classifier）.....36
3.5 实验.....37
3.5.1 数据集介绍.....37
3.5.2 实验环境与细节.....38
3.5.3 实验结果.....38
3.5.4 实验分析.....38
3.6 本章小结.....42
第四章 基于跨类别元学习的视觉富文档关系抽取研究.....44
4.1 研究动机.....44
4.2 相关工作.....45
4.3 任务定义.....46
4.4 视觉富文档少样本数据集.....48
4.4.1 针对关系的采样策略.....48
4.4.2 数据集细节.....49
4.5 提出方法.....51
4.5.1 感兴趣区域回归（Region of Interest Regression）.....52
4.5.2 原型矫正（Prototypical Rectification）.....54
4.5.3 基于邻近度的分类（Proximity-based Classification）.....56
4.5.4 训练函数.....56
4.6 实验.....57
4.6.1 实验环境与细节.....57
4.6.2 实验结果分析.....58
4.7 本章小结.....65
第五章 基于跨模型知识嵌入的视觉富文档表格问答研究.....66
5.1 研究动机.....66
5.2 相关工作.....67

5.3 提出方法 70
5.3.1 基线模型 70
5.3.2 跨模型知识迁移方法 71
5.4 实验具体实例 73
5.5 本章小结 74
第六章 总结与展望 76
6.1 总结 76
6.2 展望 77
插图索引 78
表格索引 80
参考文献 81
作者在攻读硕士学位期间发表的论文与研究成果 96
致 谢 97

第一章　绪论

1.1　研究背景与意义

视觉富文档（Visually-rich Documents，VRDs）是指包含丰富视觉元素的文档，这些元素不仅限于文本，还包括图像、表格、排版等多种形式的信息。视觉富文档具有信息密度高，应用场景多样等诸多特点，生活中常见的视觉富文档如图1.1所示。

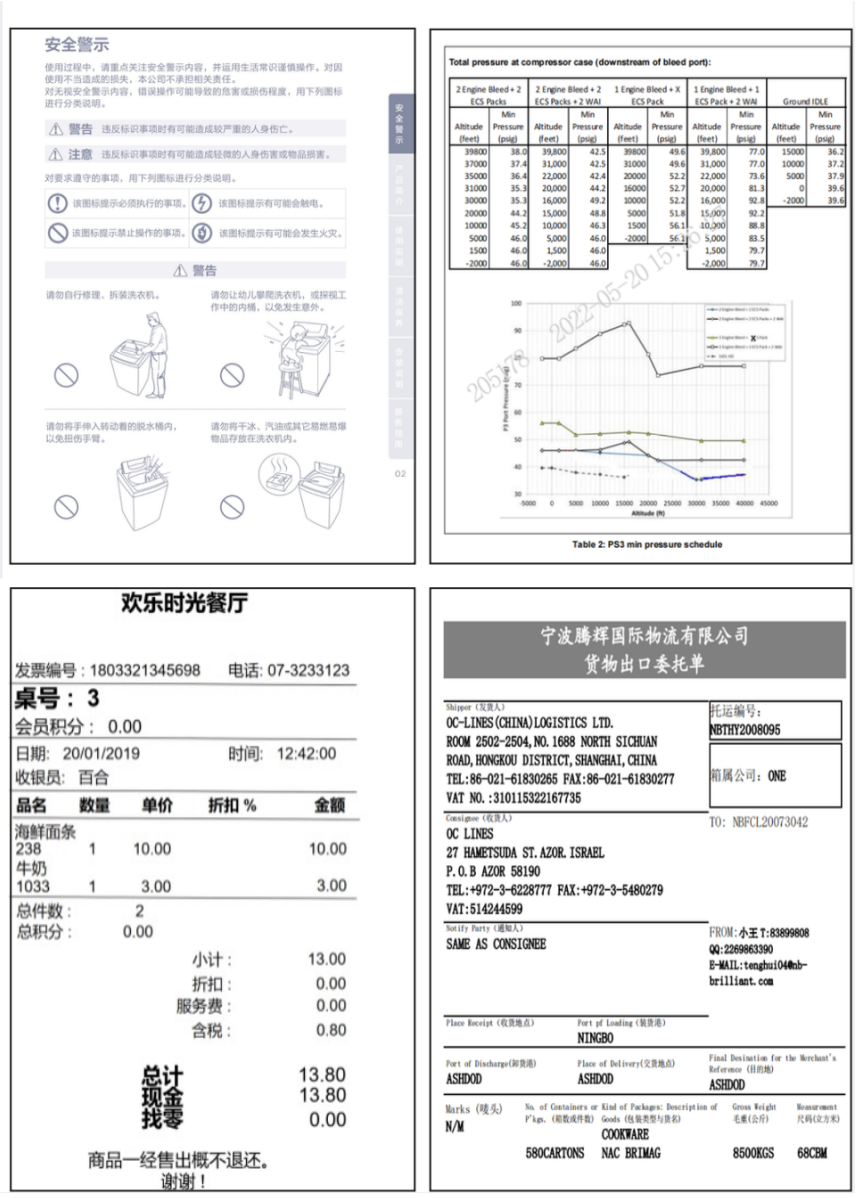

图 1.1　生活中常见的视觉富文档

这类文档在结构和内容上具有高度的复杂性和多样性，它们在多种应用领域中得到广泛应用，包括商品使用说明书、飞行器参数文档、收据、物流传单、个人简历[1]、科研论文等。随着数字化时代的到来，视觉富文档作为最常见的信息载体，用于收集、保存和展示各种模态信息，例如图像（插图、指标图、背景图等）、文本（标题、段落内容等）、排版（文本字号颜色等版式、文本缩进结构等空间、分割线等）。对视觉富文档的结构化分析和信息抽取是企业生产中的必要一环，自动化地从视觉富文档中提取关键信息，是企业数字化服务的一项关键性技术。在文档版式分析任务中，文档的视觉信息、文本信息、各版式部件间的关系信息都对分析过程具有很重要的作用。在实际应用中，视觉富文档的理解[2-3]面临着技术、语言和认知障碍的挑战。例如，现有的文档信息抽取模型需要克服由于 OCR 技术导致的阅读顺序问题，即文档中的文字内容可能按照布局坐标从上到下、从左到右的顺序排列，而这可能与人类阅读的顺序不符，影响模型的效果。因此，研究视觉富文档的信息抽取方法，旨在开发能够像人类一样阅读和理解文档的人工智能模型，这不仅需要技术上的突破，还需要对语言和认知过程的深入理解。通过构建合适的数据集、开发有效的模型和算法，可以提高对视觉富文档的理解和信息抽取的准确性和效率，进而推动相关技术和应用的发展。

视觉富文档作为本文的重要对象，其文档内的实体类别存在未知性（新的文档包含新的类别），其视觉元素具有复杂性[4]，传统的视觉富文档信息抽取方法[2-3]面临诸多挑战。知识迁移[5-6]作为一种有效的学习策略，可以跨越不同的任务和领域，提高模型的泛化能力和学习效率。知识迁移因其具备解决数据不足问题、提升模型泛化能力、促进跨领域应用等诸多意义，成为近年来学术研究的热点。知识迁移的研究不仅对于推动人工智能的理论发展具有重要意义，也对于实现更高效、更智能、更具泛化能力的人工智能系统具有实际应用价值。通过知识迁移，从而可以构建出更加灵活、适应性强的机器学习模型，以应对各种新的、未知的挑战。本文旨在研究面向视觉富文档信息抽取任务的知识迁移方法，通过知识迁移提高对视觉富文档的理解能力，并给现实世界的实际应用场景提供可能。

对于专业领域文档中表格内容的理解，非专业人士处理这些表格时会非常麻烦。大规模语言模型（Large-scale Language Model）的进步使得用自然语言输入与表格进行交互成为可能。本研究对国内外大模型进行探索，由于它们的训练过程是为自然语言量身定制的，因此，在处理表格数据时，它们的泛化能力较差，对于具有二维空间位置的表格信息有天然理解缺陷。而一些预训练的小参数表格理解模型可以弥补这一缺陷，通过小模型的知识嵌入来突破大模型对于二维空间位置的理解具有非常大的应用前景。

1.2 研究问题

根据上述研究背景，本研究主要针对视觉富文档信息抽取任务的知识迁移方法。因为在常见的视觉富文档中，键实体、值实体、表格以及键值实体间的关系多次出现且标注数据十分有限，所以，本研究主要关注以下三个研究问题：

(1) **如何将粗粒度的知识迁移到细粒度的表示中**

在文档智能模型方面，以前的工作[7-8] 只关注 token 级别细粒度的特征，忽略粗粒度的形式，例如图像块、文本框、短语等。然而视觉富文档中的布局都是以文本框这种粗粒度的信息为基础。对整个视觉富文档来说，粗粒度元素携带密集的信息并拥有一致的语义，对于文档理解至关重要。LayoutLM 系列模型[7-8]，利用段级位置特征和视觉嵌入信息，使用 token 级别嵌入表示，没有考虑文本框这些粗粒度的信息，导致 token 级别的预测错误。如何有效利用文档内粗粒度的布局信息，捕捉粗粒度元素的特征，以知识迁移的方式辅助细粒度的表示学习成为一个研究问题。

(2) **如何利用已有类别的知识帮助新类别的发现**

最近在大规模扫描文档数据集上预训练的文档理解模型，通过有效利用多模态信息，显示出了良好的性能。然而，当涉及到现实世界的应用时，文档中不同布局格式和风格对这些文档理解模型造成极大挑战。这些方法在缺乏大量有标注数据的情况下，没有自动检测和识别未知领域中新类型实体和关系的能力。相比之下，人类拥有非凡的能力，只需分析页面上的几行即可快速理解视觉富文档中的固有模式。而且，目前还没有一个既定的系统框架用于解决视觉富文

档中少样本关系学习这一现实任务。该研究领域仍然未被探索，存在尚未解决的挑战。挑战之一是在少样本场景中如何有效利用二维布局信息的问题。在视觉富文档中，键和值实体通常具有固定的位置和排列，通常表现为上下或左右排列。因此，当进行关系抽取任务，提取键值三元组时，结合二维布局空间特征可以提供互补监督信号。另一个挑战在于新类别在低资源情况下，数据样本方差大，不稳定。在多模态数据中，如何提高系统的鲁棒性是需要解决的问题。已知类别的标注数据是大量存在的，但是新类别的数据确是少量的，如何利用视觉富文档中已有类别的知识帮助新领域文档新类别的发现，既没有现成框架，更没有现成方法。

(3) **如何将表格识别模型的知识注入到大模型中**

大语言模型（LLMs）在理解表格数据方面存在一些缺陷。首先是二维数据理解的局限性：大型语言模型主要是在一维的自然语言文本上进行预训练的，而表格数据具有二维结构，这要求模型能够进行垂直和水平两个方向的推理。LLMs 在处理一维文本方面表现出色，但在理解二维表格数据时可能会遇到困难。其次是，表格结构的复杂性：表格可能包含复杂的结构，如合并的单元格、多行或多列的标题等，这些复杂性可能会影响模型对表格内容的准确理解。还有一些数据预处理的挑战，在表格理解任务之前，需要对表格进行预处理，包括表格的识别、分割、表头和表体的提取等。这些步骤可能需要结合图像处理技术和自然语言处理技术，如果预处理不充分，可能会影响模型的性能。考虑到在表格识别方面一些预训练小模型表现出色，所以可以将小模型的知识注入到大模型中，以增强大模型在表格理解任务上的性能。

1.3 研究内容

基于上述讨论，本文从三个方面研究视觉富文档信息抽取任务中存在的困难：将粗粒度元素特征迁移到细粒度中的跨粒度迁移问题、利用已有类别的知识帮助新类别发现的跨类别迁移问题、将表格识别模型的知识注入到大模型中的跨模型知识迁移问题。如图1.2所示。

（1）为了解决粗粒度元素特征捕获并进行跨粒度迁移的问题，本研究致力

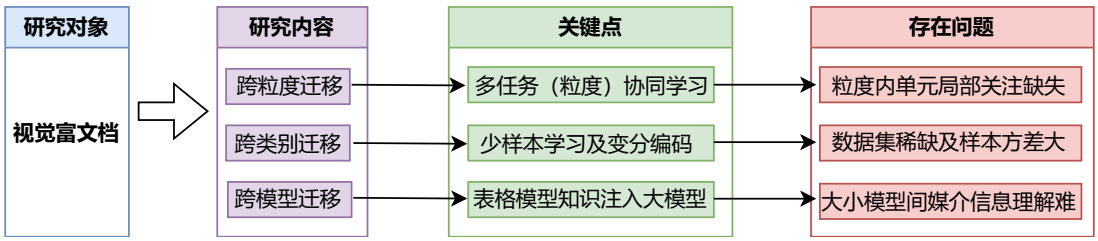

图 1.2　研究内容

于探索如何将这些细粒度的信息整合到更宏观的语义实体中。通过构建包含两个层次的图神经网络：细粒度（token）和粗粒度（bounding box）的图神经网络来分别对语义实体内的 tokens 和对 bounding box 之间的关系进行建模，从而增强了“键”和“值”实体间的信息交互。在构建细粒度的图神经网络时，本工作引入了特殊的分隔符，并设计了一种特殊的局部注意力机制。这种机制使得在语义上相关的实体在表示空间中更紧密地联系在一起，同时促进了传统细粒度预训练语言模型（如 LayoutLM[7-8]）对粗粒度语义信息的捕捉能力。通过两个层次图神经网络这种多粒度、多任务的协同学习方法，实验在 FUNSD、SEAB 和 XFUND 等公开数据集上展现了更优异的性能。

（2）为了解决视觉富文档跨类别关系抽取的研究难题，在考虑到该领域可用数据集相对稀缺的情况下，本研究对现有的有监督基准数据集进行了重新编排，并设计了一种特别针对少样本学习环境的采样策略。人类的认知过程，会先看键实体，再看其右边或者下边，这是二维空间的先验信息，如何将这些先验信息融合，增强模型在少样本关系学习任务时的表现，是一个很重要的研究内容。通过二维空间先验信息对感兴趣区域（ROI）进行建模，该区域能够有效地引导模型关注文档图像中与特定关系相关的区域。此外，在训练样本受限的情况下，一种原型矫正机制可以提升模型的泛化性能以及对新样本的适应性。这个研究点不仅可以推动视觉富文档少样本关系学习研究的进展，而且为未来在这一领域的深入探索奠定了坚实的基础。

（3）为了解决如何将表格识别模型的知识注入到大模型中的问题，本研究专注于视觉富文档中的表格理解任务，通过结合现有的表格解析模型和 XML 中间语言，将表格信息以 XML 编码的形式嵌入到大型语言模型（LLM）的提示模版中，以探究 LLM 对视觉富文档中复杂表格的理解能力。研究显示，在非

专业人士难以解读的航空航空领域表格数据，尤其是飞行器参数文档等复杂视觉文档中的表格，该方法也能取得良好的效果。实验结果表明，跨模型迁移方法可以增强大模型对表格数据的深入理解。然而，由于研究时间的限制，对于表格解析小模型的表格理解能力评估以及对表格操作指令的研究尚未进行。

1.4 研究创新点

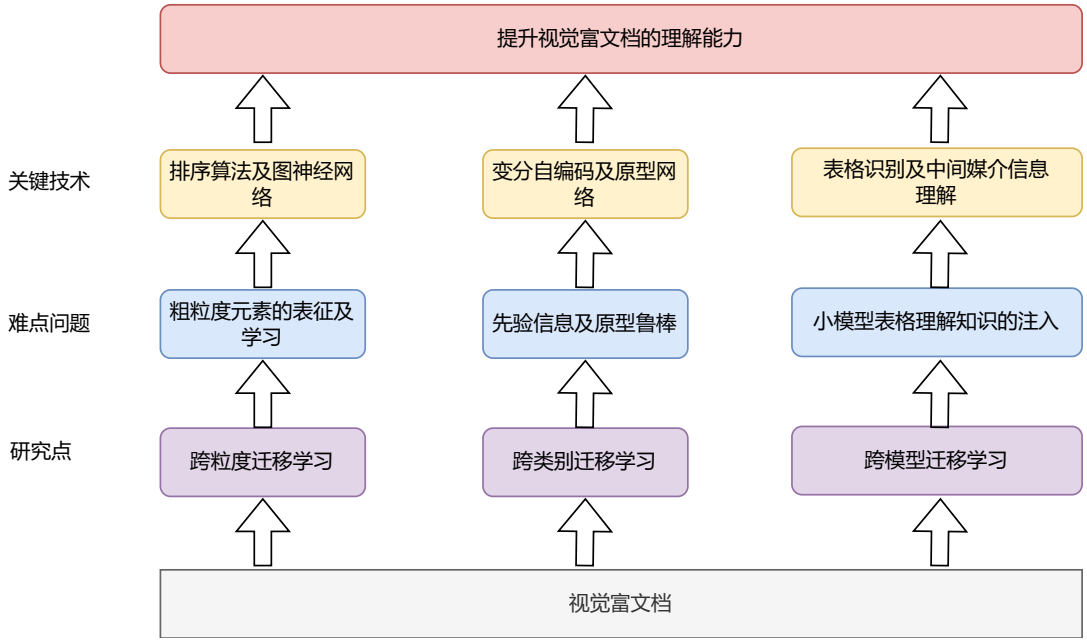

图 1.3　研究创新点

如图1.3所示，本研究主要面向视觉富文档信息抽取任务，以知识迁移方法为研究点，从跨粒度迁移、跨类别迁移、跨模型迁移三个方面展开。

其中跨粒度的迁移方法面向的是有监督场景中的关系抽取任务，目的是抽取出键值实体以及它们间的关系。其具体创新点如下：OCR 识别出的顺序不符合人类的阅读习惯。由于键实体与值实体在文档的二维空间位置上天然存在上下或者左右的关系，也就是说键实体的右边或者下面往往是值实体。本研究利用这一特点，提出阅读顺序校正算法，以确保相应的键和值关系实体彼此相邻连接。阅读顺序校正算法可以在关系抽取任务中建立与人类读者一致的阅读顺序。在细粒度上，已有的预训练模型可以很好地得到细粒度表示，本研究创新性地插入特殊的字符来注入细粒度的常识信息。在粗粒度上，将细粒度的表征迁

移到粗粒度中，构建多粒度图神经网络，利用多粒度信息来增强关系的多模态表示。本研究解决了粗粒度级别和细粒度级别的跨模态对齐问题。所以，可以适用于现有的布局感知多模态预训练模型。本研究探索了各种注意力策略，以有效建模不同粒度之间的关系。例如，采用 k 近邻图网络来捕捉粗粒度之间的关联，并利用特殊字符分割的局部注意力机制来捕捉细粒度单元之间的关联。通过注意力策略使提出的模型能够充分利用文档布局中的版式信息。实验对关系抽取任务进行广泛评估，提出的方法始终优于现有方法。研究还进行了消融实验来分析加入特殊字符和跨粒度特征迁移的影响，实验结果证明了提出方法的关键作用。

其中跨类别的迁移方法面向的是低资源场景中的关系抽取任务，目的是依靠少量的样本，抽取出新的键值实体以及它们间的关系。其具体创新点如下：第一，本项研究首次针对视觉富文档中的少样本关系抽取问题提出了框架及解决方法。为了推动该领域的进一步研究，创新性地构建了两个专门针对少样本学习场景的基准数据集。第二，本研究提出了一种新颖的变分原型矫正方法，该方法不仅融合空间先验知识，还能够保证类别原型表示的鲁棒性，降低原型表示的方差。第三，实验在构建的数据集上进行广泛评估，证明了提出的方法在少样本视觉富文档关系抽取任务中达到了最优的性能。

其中跨模型的迁移方法面向的是视觉富文档中的表格问答任务，通过已有的版面分析模型，识别到文档中的具体表格，提出了一个端到端的顺序建模框架对表格进行分析，识别表格的逻辑结构，也就是 XML 形式的中间语言。框架包含一个新颖的坐标序列解码器，该解码器将边框坐标（左、上、右和下坐标）建模为一个语言序列，并顺序解码这些坐标，利用坐标之间的依赖性来提高预测的准确性。为了解决逻辑表示缺乏局部视觉信息的问题，提出了一个辅助的视觉对齐损失函数，以增强非空单元格的逻辑表示中的局部视觉细节，从而帮助产生更好的单元格边框。其中序列解码器会产生 XML 序列和非空单元格的边框。实验证明跨模型的知识迁移方法可以避免大模型对表格内容进行回答时的错误。

1.5 研究的组织结构

本文首先介绍了研究面向视觉富文档信息抽取任务的知识迁移方法的背景及意义，然后介绍已有的视觉富文档理解模型以及知识迁移方法。从面向有监督场景中关系抽取任务的跨粒度迁移方法、面向低资源场景中关系抽取任务的跨类别迁移方法、面向视觉富文档中表格问答任务的跨模型迁移方法三个方面展开。最后总结了三个研究工作，并进行展望。本文的组织结构如下：

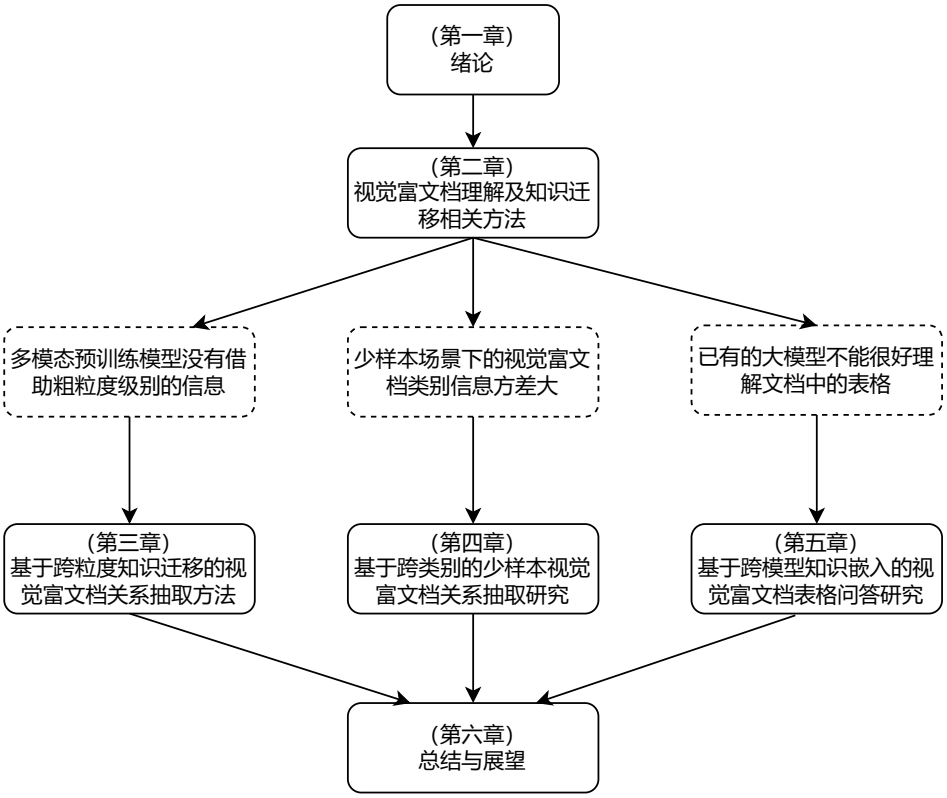

图 1.4　论文组织结构

第一章：介绍了基于视觉富文档信息抽取任务的知识迁移方法的研究背景和意义，针对当前已有研究存在的问题进行分析，提出了三个针对不同任务的研究问题与研究内容，以及对应的三个创新点。

第二章：对视觉富文档理解已有模型、知识迁移已有方法进行介绍和概述。

第三章：基于跨粒度表示学习的视觉富文档关系抽取研究。

第四章：基于跨类别元学习的视觉富文档关系抽取研究。

第五章：基于跨模型知识嵌入的视觉富文档表格问答研究。

第六章：总结与展望。概括了取得的研究成果，同时指出了当前工作的局限性。此外，也对视觉富文档领域的未来研究趋势进行了预测和讨论。

第二章　视觉富文档理解及知识迁移的相关方法

2.1　视觉富文档理解方法的研究现状

视觉富文档（Visually-Rich Documents，VRDs）出现在日常生活的方方面面，常见的有商品使用说明书、收据、发票、货运单、表单及飞行器的参数文档等。目前大量研究致力于利用光学字符识别（OCR）技术，从视觉富文档中抽取有价值信息。OCR 识别的输出通常由多个文本框（bounding box）及对应的文本段（text）组成[9-10]。如图2.1所示，OCR 识别到所有的文本框，并且对应每个文本段，例如，“姓名”、“住址”、“工作单位”等等。每个文本段是一个语义实体（文本信息），每个文本框是一个二维位置坐标（布局信息）。此外还有一些其他的方法能够从原始视觉富文档图像提取出多模态的信息。

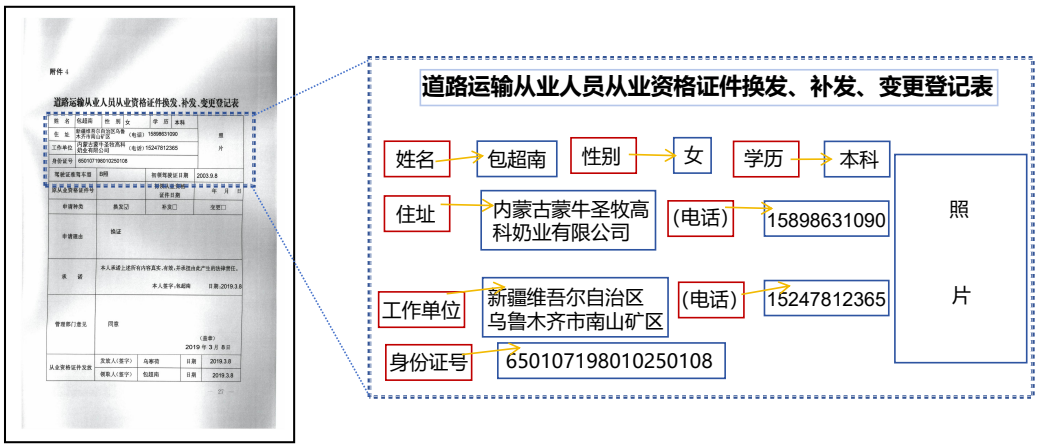

图 2.1　视觉富文档的识别结果（红色代表键实体，蓝色代表值实体，黄色代表实体间的关系）

如图2.2，现有的视觉富文档理解框架主要面向文档的版面布局分析、视觉问答、信息抽取、表格问答、文档分类等。其中，版面分析指的是要解析当前文档视觉版面结构；视觉问答指的是回答自然语言提出的问题，表格问答针对的是表格内容回答自然语言提出的问题；信息抽取指获取视觉富文档中的关键信息：包括实体识别（提取文本字段的关键信息，实体标注）以及关系抽取（实现语义实体之间的链接，实体链接）；文档分类指的是要预测每个文档图像的相

应类别。本研究重点关注信息抽取中的关系抽取及表格问答任务。关系抽取是本文第三章及第四章研究的重点。如图2.1中的黄色箭头所示，该任务旨在链接每一对有关系的语义实体。而表格问答任务，需要先对文档进行版面分析，提取出表格内容及表格的上下文，再根据大模型在处理自然语言方面的能力，去回答与表格内容有关的问题。

下面从三个角度介绍一下视觉富文档理解已有方法及模型：

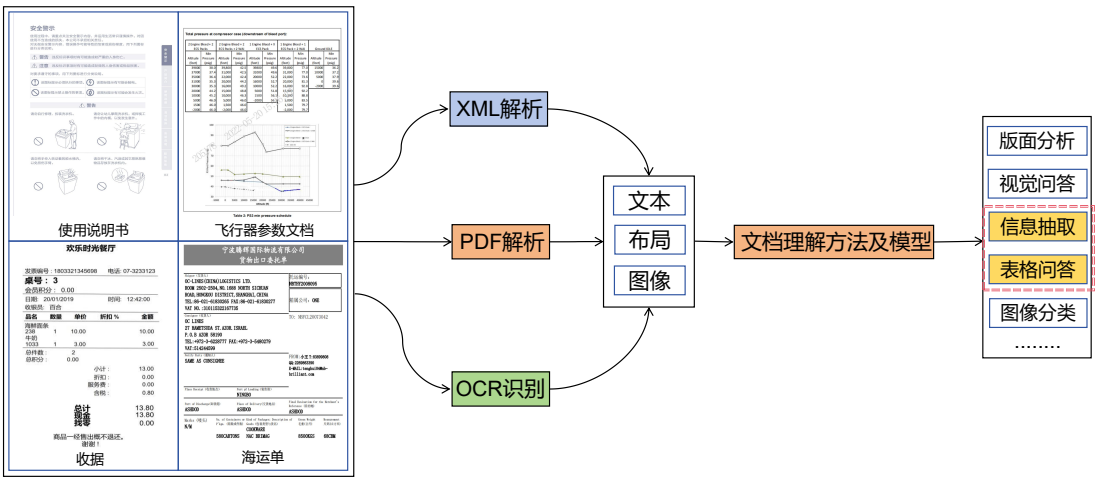

图 2.2　视觉富文档理解框架

2.1.1　基于启发式规则的文档理解方法

基于启发式规则的文档理解方法主要采用三种不同的处理策略：由整体到局部、由局部到整体以及两者结合的方法。由整体到局部的策略，通过连续递归地将文档图像细分为更小的区域，直至满足既定的划分标准，这些标准通常指的是文档中的块状结构或列状布局。相反，由局部到整体的策略则是从图像的像素或基本组件出发，通过将这些小单元组合和聚合，构建出更大的、具有一致特征的区域。由整体到局部的方法在处理特定格式的文档时，因其针对性强，往往能够实现更快速和更高效的分析。然而，由局部到整体的方法虽然在计算上更为耗时，但其适用性更广泛，能够适应各种不同布局的文档。混合方法则试图融合上述两种策略的优势，来达到更佳的处理效果。本节内容将分别从由整体到局部和由局部到整体两个视角出发，探讨包括投影轮廓分析、图像扩散分析、连通区域分析等在内的多种文档分析技术。

（1）投影轮廓分析算法是由整体到局部文档分析方法中的一种，它在文档理解领域得到了广泛的应用。该算法中的 X-Y 切割技术[11] 能够将文档分割成不同的部分，这种方法特别适合于那些具有明确定义的文本区域和固定行距的格式化文本。然而，X-Y 切割技术对文档的边界噪声非常敏感，并且在处理倾斜文本时表现不佳，因此它对文档的清晰度和质量有着较高的要求。为了解决这些问题，自适应局部投影方法[12] 被提出，它能够测量文档的倾斜角度，并尝试补偿因文本倾斜而导致的性能损失。实验结果表明，该方法在处理倾斜和非直线文本时能够取得较为精确的结果。此外，为了克服 X-Y 切割算法的局限性，研究者们提出了多种变体算法，包括利用组件边界框的投影技术，以及采用编辑成本评估作为分割过程的学习指标。这些改进在不同程度上提升了算法的性能。尽管投影轮廓分析算法在处理结构化文本方面表现出色，但它在面对布局复杂、文本倾斜或存在边界噪声的文档时，可能无法发挥最佳效果。

（2）图像扩散分析算法通过从一个选定的起点向外扩散，逐渐覆盖所有相似的区域，从而识别出页面中的特定区域。游长平滑算法[13] 作为由整体到局部的策略，是经典的图像扩散分析算法，它被用于识别和合并同质区域，将图像转换为二值图，其中像素值为 0 代表背景，而 1 代表前景。当背景像素周围的背景像素数目少于一个设定的阈值时，该像素值会被更改为 1，通过这种方式，游长平滑算法能够将相邻的前景元素合并成一个连续的区域。这个过程可以连续进行，直至字符被合并成单词，单词再被合并成文本行，最终扩展到整个同质区域。在这一基础上，后续研究进行了进一步的优化，包括引入除噪和倾斜校正等预处理步骤。还有方法[14] 对图像中的每个像素点进行扩散，生成一个新的灰度图，然后从中提取信息，这种方法即使在手写字体或文本倾斜的情况下也能保持较好的性能。

（3）连通区域分析算法是一种由局部到整体的文档分析技术，它专注于识别和推断最小单元元素之间的联系，目的是识别出同质的区域，并将这些区域归类为具有不同属性的集合。Fisher 等人[15] 利用该技术识别出每个组件的 K 个最近邻（K-Nearest Neighbors, KNN），并基于它们之间的相对位置和角度等信息来确定当前区域的属性。还有算法通过判断文档的倾斜角度，将文字元素首先

合并成线条，然后将这些线条进一步合并成更大的区域，并最终对这些区域进行属性分类。此外，还有研究尝试处理文本的倾斜问题，通过使用近似面积图技术来确定区域的潜在边界，这种方法对于处理任意倾斜角度的区域都是有效的。然而，这种方法在计算过程中需要对字符间距和行间距进行估计，因此在文档中存在较大的字体或较宽的字符间距时，模型的性能可能不会很理想。为了进一步提升性能，一些方法尝试结合连通区域分析和自动多层感知机来寻找最佳的分类器参数[16]。通过这种方式，可以在保持算法通用性的同时，针对特定文档的特征进行优化，从而提高整体的分析效果。

2.1.2 基于统计机器学习的文档理解方法

文档分析的传统流程大致分为两个主要步骤：第一阶段涉及将文档图像分割成多个潜在的区域；第二阶段则对这些区域进行属性分析，以确定它们是否属于文本、图像或其他预定义的类别。机器学习方法在这一领域也主要围绕这两个方面展开，一些研究采用机器学习技术对文档进行分割，而另一些研究则致力于在现有区域上构建特征，并利用机器学习技术对这些区域进行分类。同时，由于统计机器学习技术在性能上的显著提升，许多基于该技术的方法在文档分析的关键环节——表格检测中得到了应用。本节内容将对表格检测进行一定的讨论。与之前侧重于技术层面的讨论不同，接下来的内容将从文档分析任务的角度来介绍其发展情况。

（1）文档分割

在文档的分割过程中，相关研究[17]采用了 X-Y 裁剪算法，并结合逻辑回归对文档进行精确切分，同时去除空白区域。在识别出目标区域后，通过实验对比了 K 近邻、逻辑回归和最大熵马尔可夫模型等不同算法的效果。结果发现，在属性分类任务中，最大熵马尔可夫模型和逻辑回归算法效果均可。还有工作在文档分割任务中进一步增强了机器学习算法的应用，采用了一种基于内核的算法，从字母到单词，再到整个文本行，逐步实现由局部到整体的合并，并最终将处理结果以 XML 格式进行存储。与此同时，有工作[18]针对文档中文字可能存在的两种阅读顺序问题进行了研究。他们发现，传统的算法通常假设文字只有单一的书写方向，这在处理像汉语或日语这样既可以水平也可以垂直书写

的文字时存在局限性。为了解决这一问题，他们提出的文档分割算法能够判断并处理文本的阅读顺序，并利用支持向量机（SVM）来决定是否需要执行特定的步骤。

（2）区域分类

在处理区域属性分类问题时，大部分研究集中于探索和测试各种机器学习算法作为分类器的效果。Wei 等人[19] 通过实验对比了支持向量机、多层感知机和高斯混合模型这三种算法作为分类器的性能。实验数据显示，支持向量机和多层感知机在区域属性分类任务中的表现显著超过了高斯混合模型。另一方面，相关研究通过手动构建多个特征，并利用自动多层感知机对这些特征进行分类。此外，还有方法在文档分割方面做出了创新，并应用了动态多层感知机作为分类器[17]，提高区域分类的表现。

（3）表格检测

在表格识别的研究中，有方法通过二叉树结构对文档进行递归的自顶向下分析[20-21]，以发现可能的表格区域。随后，这些方法会根据区域的特定特征来精确定位表格。例如，研究人员利用条件随机场技术在 HTML 页面中定位表格[22]，并能够识别出表格中的标题和子标题等元素。此外，一些研究专注于从手写文档中检索表格区域[23]，他们通过支持向量机 (SVM) 来识别文档中的文字区域，并根据文本行的排列来确定表格的确切位置。还有研究者使用 SVM 来识别图像中的水平和垂直线条，并通过属性分类判断这些线条是否属于表格的一部分。

2.1.3 基于深度学习的文档理解模型

（1）基于卷积神经网络的文档版面分析模型

近年来，卷积神经网络（CNN）在计算机视觉领域取得了显著的进展，尤其是 ResNet 模型[24] 在多个任务中表现出色，如图2.3所示，ResNet 通过采用残差学习框架显著提升了性能，其利用残差块来执行任务，这种设计包含多个卷积层，并在输入与输出之间建立了一条直接的跨层连接。这种设计允许网络学习输入和输出之间的残差映射，通过简单的相加操作来生成最终的输出结果。这样的机制使得 ResNet 能够有效地捕获深层网络特征，同时解决了深度学习中常见的梯度消失和模型性能退化的问题。随着 Faster R-CNN[25]、Mask R-CNN[26] 等

多阶段检测模型的广泛采用，目标检测任务在计算机视觉领域已经得到解决。

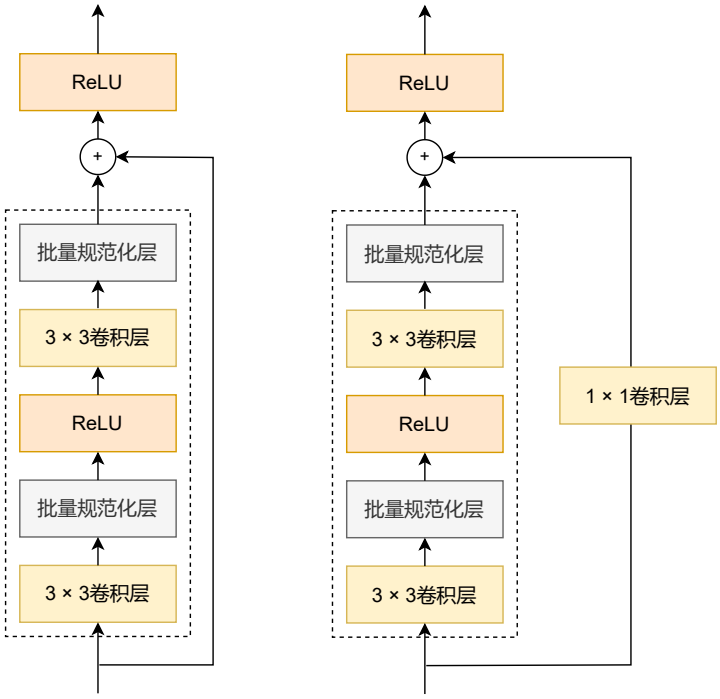

图 2.3 ResNet 结构图[24]

文档版面解析通常被视为文档图像中的对象识别问题，其中文档的标题、段落、表格、图表等构成元素是需要被识别和定位的目标。有研究[27] 将文档版面解析为像素级的分割，并成功地应用了卷积神经网络进行像素级的分类，取得了显著的成果。还有研究首次尝试将 Faster R-CNN 架构用于文档版面中的表格检测任务。尽管文档版面分析是文档理解领域中的一项成熟任务，但由于可用数据集的规模限制，仅依靠传统的计算机视觉预训练模型的潜力尚未得到充分发挥。然而，随着大规模且具有弱标注特性的文档版面分析数据集如 PUbLayNet[27] 和 DocBank[28] 的发布，研究者现在能够在更广泛的范围内对多样的计算机视觉模型和策略进行深入的评估与对比，这为文档版面解析技术的进一步发展提供了新的动力和方向。

（2）基于图神经网络的文档信息抽取模型

信息抽取是指从非结构化文本中提取有用信息并将其转化为结构化数据的过程。传统的信息抽取主要关注从纯文本中抽取实体和关系信息，而对视觉富文档数据的研究相对较少。Liu 等研究人员[9] 提出了一种基于图卷积神经网络

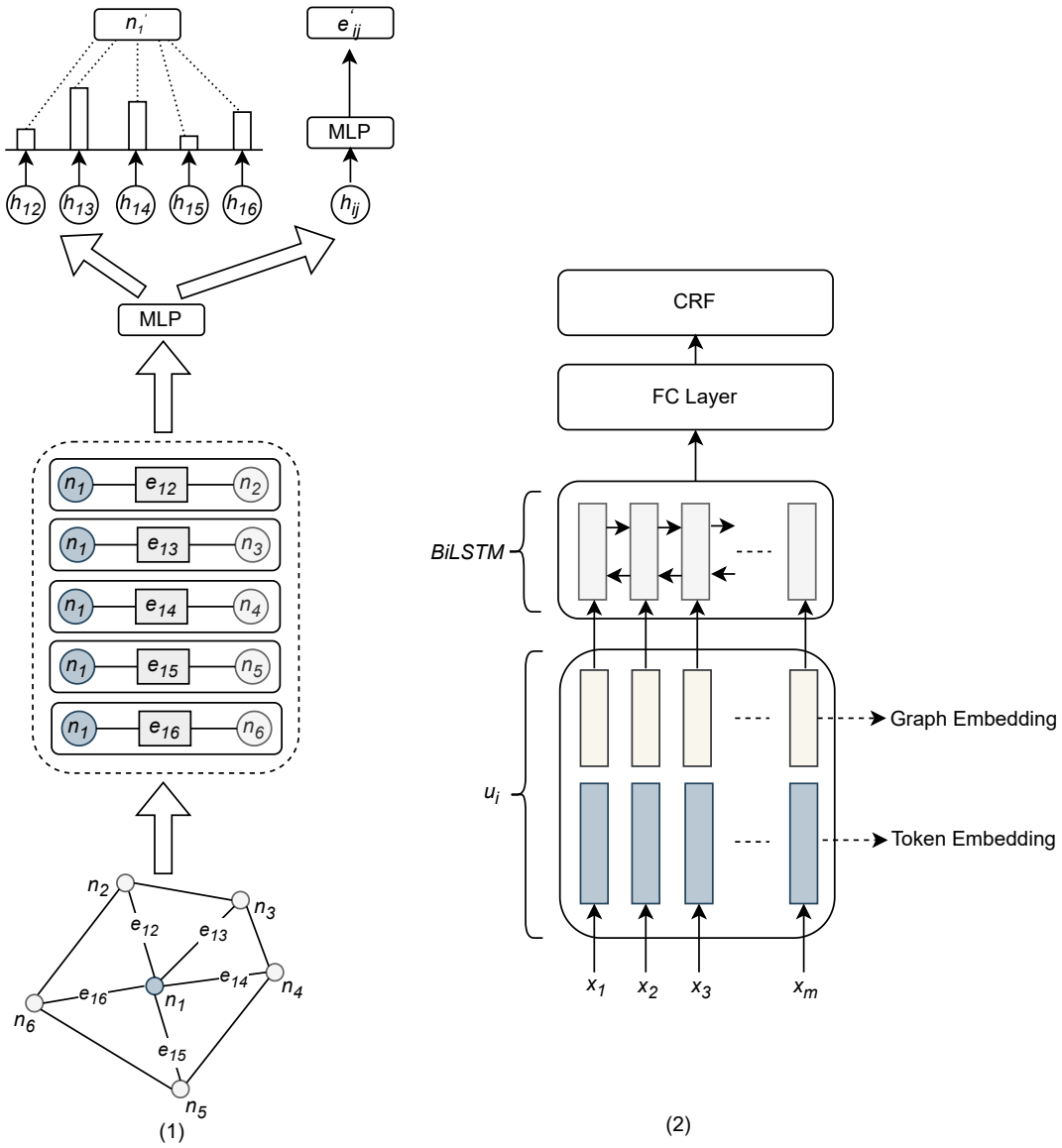

图 2.4　图卷积过程[9]

的方法对视觉富文档数据进行建模。如图2.4所示，每张图片通过光学字符识别（OCR）处理后，会生成一系列文本块，每个文本块都包含了其在图片中的位置坐标和文本内容。他们利用这些文本块构建一个全连接有向图，其中每个文本块作为一个节点，并且每个节点都与其他所有节点相连。节点的初始特征是基于文本内容，通过双向长短时记忆网络（BiLSTM）得到编码。边的初始特征则由相邻文本块与当前文本块之间的相对位置和尺寸信息构成，这些信息通过当前文本块的高度进行归一化处理，以实现仿射不变性。与仅在节点上应用卷积的传统图卷积模型不同，该方法更加注重信息抽取中的“实体-关系-实体”三元组信息，因此他们在“节点-边-节点”构成的三元组上执行卷积操作。此外，他

们还引入了自注意力机制，使网络能够在构成全连接有向图的所有有向三元组中识别出更加重要的信息，并进行特征的加权聚合。经过多层卷积处理后，节点和边的初始特征被转化为更高层的表示，从而为信息抽取任务提供了更丰富的上下文信息。该研究在增值税发票和国际采购收据的数据集上进行了效果验证。研究中采用了两套基准系统：第一套基准系统通过 Bi-LSTM+CRF 模型对每个文本块的内容进行独立解码；第二套基准系统则是将所有文本块的内容按照“从左至右、从上至下”的顺序进行拼接，然后对拼接后的文本整体应用 Bi-LSTM+CRF 模型进行解码。实验结果表明，采用图卷积网络的模型相较于基准系统在性能上有了显著的提升。特别是在那些仅依赖文本信息即可抽取的字段（例如日期），图卷积模型的性能与基准系统相当；而在需要结合视觉信息进行判断的字段（例如价格、税额），性能提升更为显著。此外，实验还发现视觉信息在模型中起到了关键作用，增强了语义相近文本之间的区分能力。同时，文本信息也在一定程度上辅助了视觉信息的识别。而自注意力机制在固定版式的文档上作用不大，但在非固定版式的文档上则能带来一定的性能提升。本研究第三章的工作也是受到该工作的启发，提出了局部注意力机制。

（3）基于 Transformer 结构的通用文档理解预训练模型

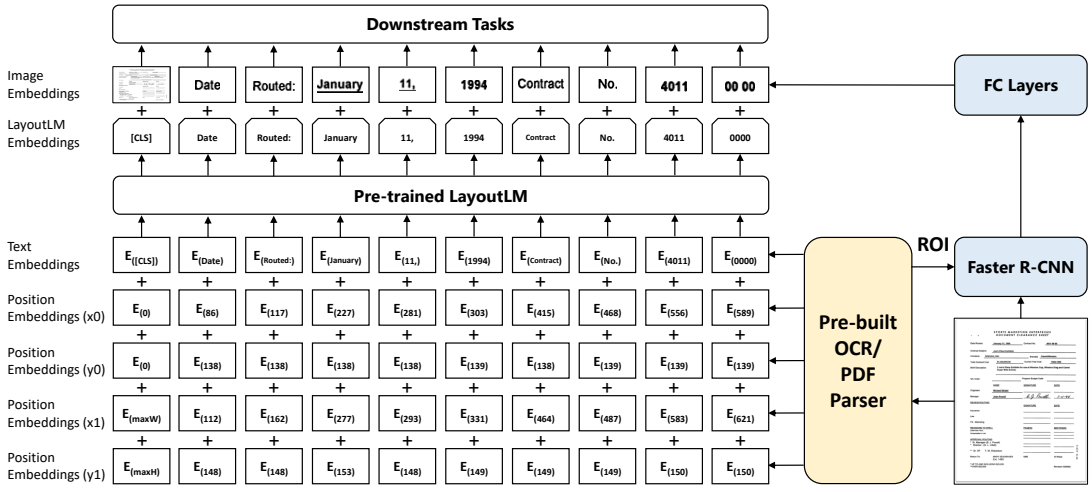

图 2.5　LayoutLM[7] 文档理解模型结构图

在许多情况下，视觉富文档中文本段的排列方式具有重要的语义信息。例如，表单通常采用键值对（key-value pair）的格式，其布局通常为横向或纵向，

并且存在特定的关联关系。通过预训练，这些与文本自然对齐的位置信息可以为后续的信息提取任务提供更丰富的语义内容。对于包含丰富格式的文档，除了文字的空间关系外，文字的视觉格式也对任务有所帮助。在文档级别的任务中，整个文档的视觉布局可以提供全局结构信息，这些视觉特征，通过视觉模型提取并与预训练阶段结合，可以有效辅助完成下游任务。

为了充分利用文档中的结构化信息，如图2.5所示，LayoutLM 通用文档预训练模型[7] 被提出，该模型将二维布局嵌入和图像嵌入整合到 BERT 模型的原始架构中，利用来自 Faster R-CNN[25] 的图像嵌入来共同处理下游任务。具体来说，它在现有的预训练模型上增加了两种新的嵌入层：二维位置嵌入层和图像嵌入层，以此来有效融合文档的结构信息和视觉信息。在图像嵌入层，每个文本框区域被视为图像编码器中的一个候选区域，用于提取该区域的局部特征。在 LayoutLM 模型的预训练阶段，设计了两个自监督的学习任务来适应其特性：一是掩码式视觉语言模型任务，二是多标签文档分类任务。广泛的实验已经证实，掩码式语言模型在自监督预训练中非常有效。掩码式视觉语言模型在传统方法的基础上进行了创新：在遮蔽特定词汇的同时，保留该词对应的二维位置嵌入信息，以此引导模型预测被遮蔽的词汇。这种设计使得模型能够依据上下文信息及视觉线索来预测被遮蔽的词，进而加深对文本位置和语义之间关系的理解。多标签文档分类任务旨在引入更高层次的语义信息，因为虽然掩码式语言模型能够捕捉词级信息，但为了获得文档级别的表示，需要专门的预训练任务。在这一任务中，每个文档都会附带多个标签，用以标注文档的类型。通过引入这一任务，模型能够利用这些标注信号来整合文档的类别信息，从而捕捉到文档的类型特征，实现更丰富的语义表示。实验结果显示，预训练阶段所融合的结构化和视觉信息能够有效地转移到下游任务中，并且在多个下游任务中实现了显著的准确率提升。与传统的基于卷积神经网络和图神经网络的模型相比，LayoutLM 作为一个通用的文档智能预训练模型，其优势在于能够适配多种不同的下游任务。

尽管 LayoutLM 在领域内获得了成功，但它仍然存在一些局限性。LayoutLMv2 是一个融合了文本、视觉和布局数据的多模态预训练框架。与前身 Lay-

outLM 相比，LayoutLMv2 的创新在于整合了视觉内容，并采纳了一种能够感知空间关系的自注意力机制。在处理输入时，LayoutLMv2 同时分析文本和图像数据，通过编码过程来构建它们与文档布局之间的联系。为了提升空间感知能力，LayoutLMv2 引入了二维相对位置编码，用以表达文本块在空间上的相互关系。这与纯文本模型中采用的一维相对位置编码相似，但扩展到了二维层面，从而更精确地捕捉文本块的空间布局。此外，LayoutLMv2 还引入了两个新的预训练任务："文本—图像对齐"与"文本—图像匹配"，这些任务旨在训练模型如何将文本内容与图像特征进行匹配和对齐。如图2.6所示，模型将输入的多模态数据转换成统一的向量表示，供后续任务使用。

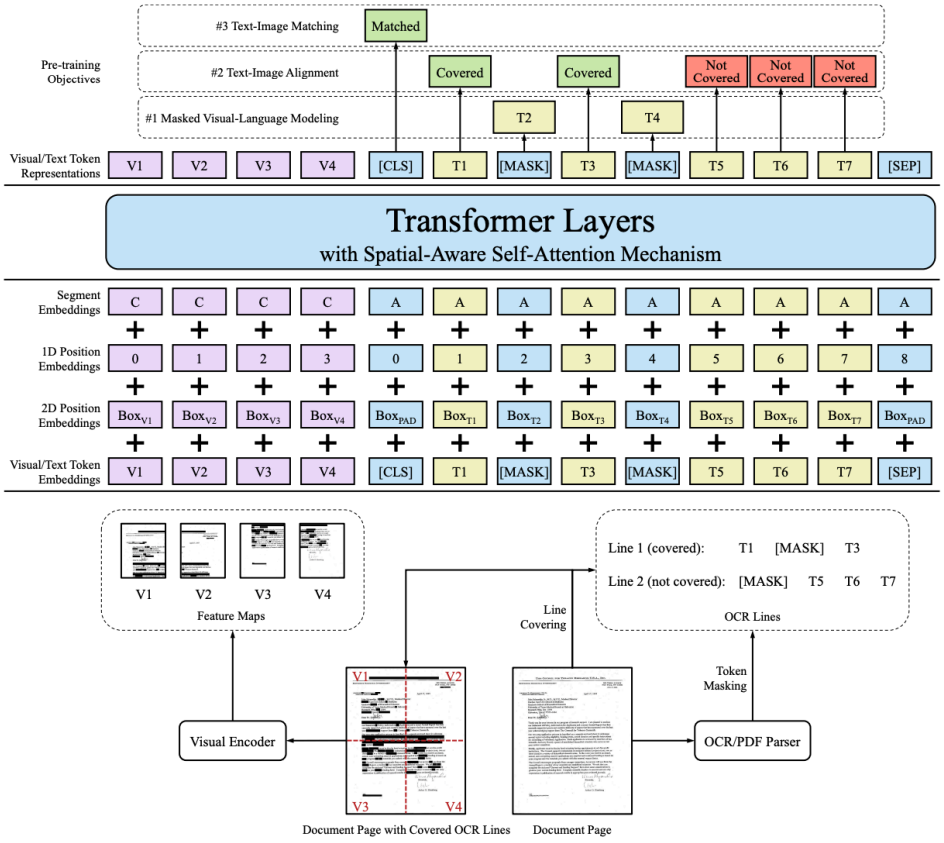

图 2.6　LayoutLMv2[8] 文档理解模型结构图

在视觉富文档中，不同国家的语言是很常见的。针对多语言的问题，LayoutXLM[10] 被提出，这是一个基于多语言文档理解任务的多模态预训练模型，其模型架构如图2.7所示。LayoutXLM 是 LayoutLMv2 的多语言版本，它采用了相

同的 Transformer 架构进行多模态预训练，并通过多语言预训练模型 InfoXLM[29] 进行初始化。该模型能够处理包含多种元素的文档，通过编码多种模态的输入，学习它们之间的关系，为文档理解任务提供信息。针对表单理解中的关键任务——键值提取，研究人员将其细化为两个子任务：语义实体识别和关系提取。基于跨模态对齐的核心理念，LayoutXLM 的预训练框架包括三个目标：多语言掩码式视觉语言模型（用于文本和布局的对齐）、细粒度的文本-图像对齐和粗粒度的文本-图像匹配。在多语言掩码式视觉语言模型的预训练任务中，模型需要依据文本上下文和布局信息预测被遮蔽的文本。文本-图像对齐任务旨在帮助模型理解文本和图像之间的细微对齐关系，通过随机选择文本行并遮盖其在文档图像中的对应区域，模型需要预测文本是否被遮盖。文本-图像匹配任务则要求模型确保文本和图像的高层语义表示相匹配，为此模型需要判断文本和图像是否源自同一文档页面。

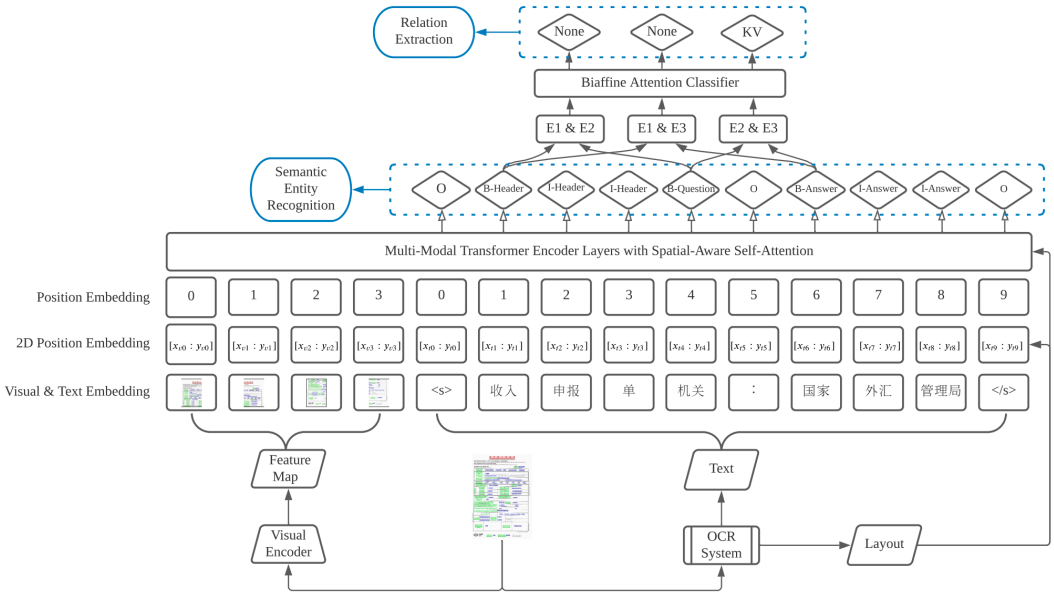

图 2.7 LayoutXLM[10] 文档理解模型结构图

2.2 知识迁移方法

2.2.1 知识迁移的概念和发展

如图2.8，在 1901 年，心理学家 Woodworth 从心理学和教育学的角度出发，对知识迁移理论进行了阐述。他们指出，知识迁移与所要完成的任务之间存在某种联系（相似性）。从人类的认知过程来说，人类具有在不同领域和问题之间进行知识迁移或转换的能力。这是因为，无论一个人是有意识还是无意识地学习某项任务，学习的结果就是会形成先验知识固化到人脑中，成为人的认知。当这个人学习新任务时，他可能会从这些经验知识中提取出对新任务学习有益的部分。总的来说，无论人类如何实现学习迁移，这种知识迁移或转换的能力，正是当前人工智能领域中机器学习[30] 所缺少的。

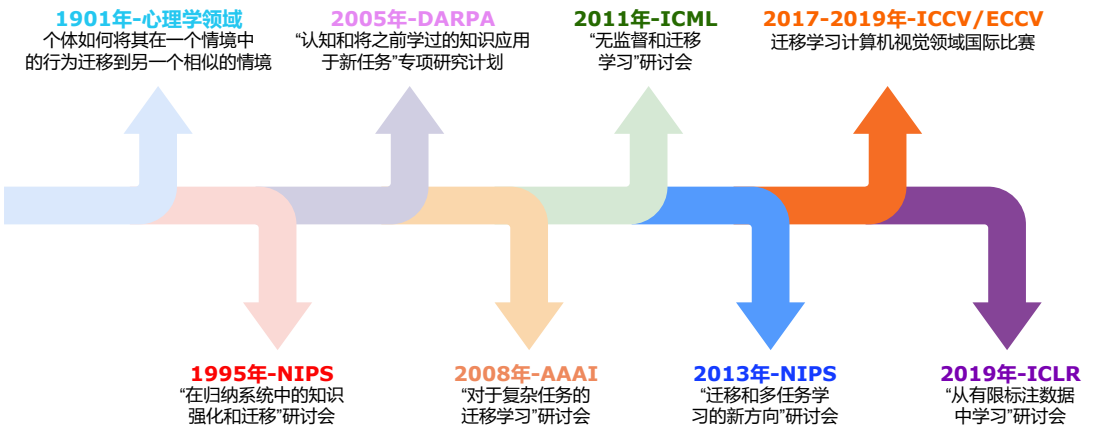

图 2.8　知识迁移的发展历史

在机器学习的研究领域，迁移学习的概念最早可以追溯到 1995 年的 NIPS-95 会议上的“学会如何学习”专题讨论[31]。自那时起，迁移学习的研究开始受到越来越多科研人员的关注[31-32]，并在该领域内占据了重要地位。迁移学习涉及的是在不同领域或不同任务之间实现知识转换的能力。简单来说，迁移学习描述的是一种学习过程对另一种学习过程产生的影响。如图2.9中所示，传统的机器学习方法与迁移学习方法在学习过程中存在明显差异。在左侧的传统机器学习中，各个学习任务基于不同的数据分布，彼此之间是独立不相连的。这意味着，面对不同的数据分布，同一学习任务需要重新进行训练。相比之下，在右侧的迁移学习中，尽管源任务与目标任务不同，但源任务之间不再是孤立的。

在迁移学习中，可以利用不同源任务的数据，发掘与目标任务相关的知识，从而辅助目标任务的学习过程。

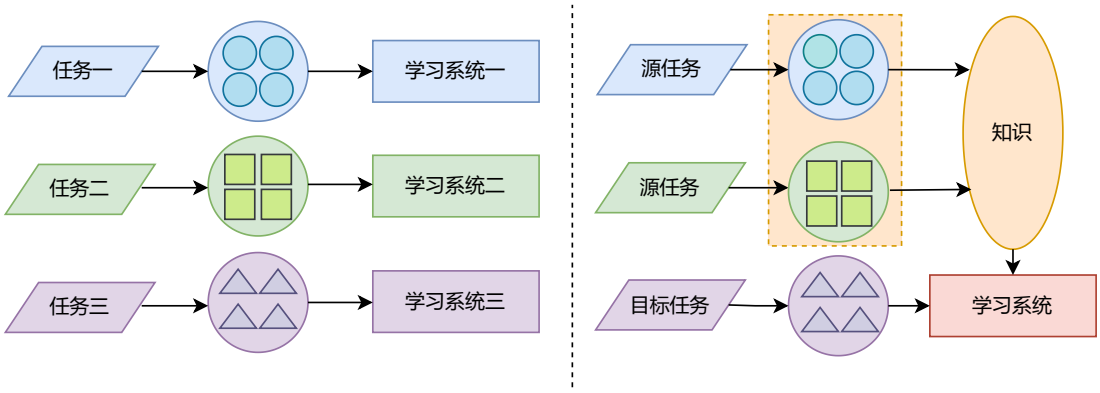

图 2.9 传统的机器学习和迁移学习的学习过程

2.2.2 迁移学习的主要方法

迁移学习的方法可以从不同的角度进行分类。从知识迁移的角度来看，主要可以分为基于实例的迁移、基于特征的迁移、基于参数的迁移和基于关系的迁移。基于实例的迁移学习是将源任务中的样本直接或间接地应用到目标任务中；基于特征的迁移学习是在源任务和目标任务之间共享特征表示；基于参数的迁移学习则是共享模型的部分参数；而基于关系的迁移学习则是在更抽象的层次上进行知识迁移。

从模型的角度来看，迁移学习的方法又可以分为基于模型的迁移和基于知识的迁移。基于模型的迁移学习通常涉及到预训练模型的使用，例如在自然语言处理领域中常用的 BERT 模型；而基于知识的迁移学习则是将知识以显式的形式表示出来，并在不同任务之间进行迁移。

2.2.3 迁移学习的三种设定

本节提供了关于迁移学习技术的分类，将其划分为几个子部分，这些子部分将迁移学习分为三种设定，即归纳式、传导式和无监督迁移学习。图2.10展示了不同迁移学习设定的概览，明确表示了三种迁移学习类型以及属于这些类别的所有可能情况。归纳式迁移学习包括两种情况，这取决于源领域和目标领

域中数据的可用性，包括回归和分类任务。传导式迁移学习也展示了两种情况：第一种情况是领域不同但任务相同，第二种情况是领域和任务都相同。在无监督迁移学习技术中，源领域和目标领域都无法获得数据。下面分别介绍三种设定：

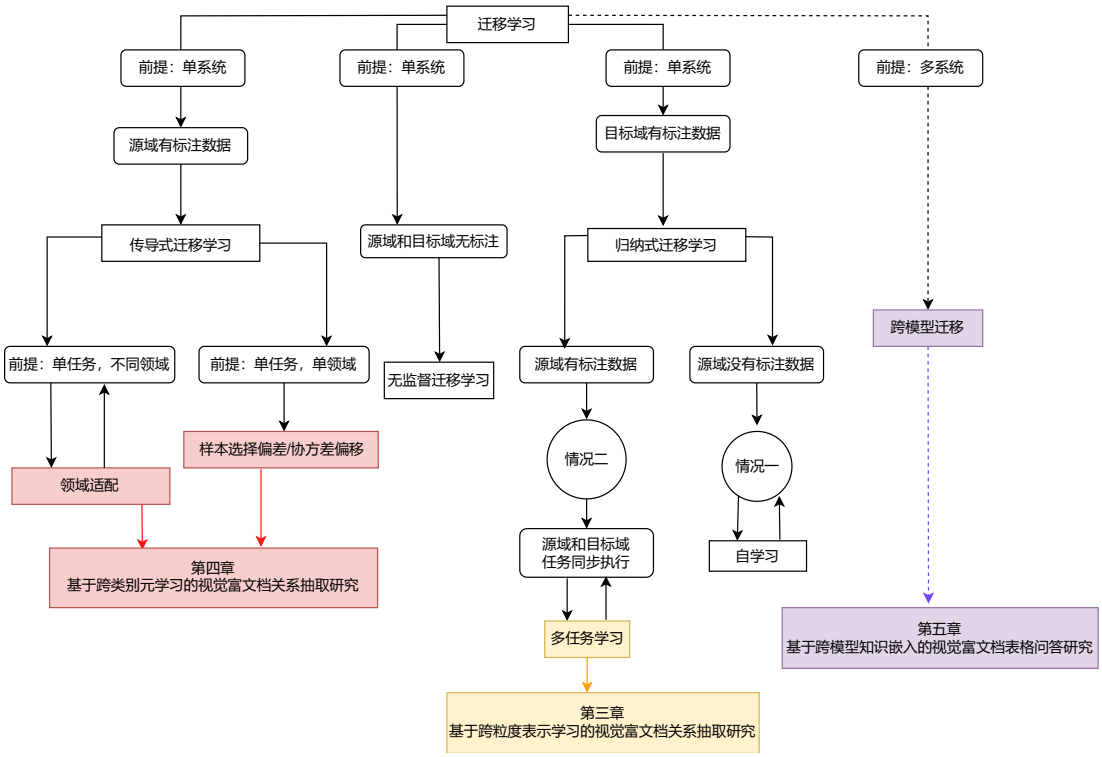

图 2.10 迁移学习框架总结以及本文关注的重点

（1）归纳式迁移学习：

归纳式迁移学习算法的思想是在目标任务与源任务不相同的情况下在目标领域中提高近似目标函数 PDF $f_T(.)$，源领域和目标领域可能相同也可能不同。根据有标注数据和未标注数据的可用性，归纳式迁移学习可以分为两种情况：（1）多任务学习；（2）自学习。多任务学习的一种特殊情况是源领域包含一个大型的有标注数据库。归纳式迁移学习方法专注于通过从源任务传递的知识来使得在目标任务中也取得不错的性能。然而，多任务学习会同时学习多个任务，包括源任务和目标任务[33]。在自学习的情况下，源领域中没有可用的标注数据，而在目标领域中有可用的标注数据。自学习是一种深度学习方法，它包含两个分类阶段。首先是对从大量未标记数据中学习到的特征进行转移，在第二阶段，这

种学习到的表示被用于标记数据以执行分类任务。在自学习中，源领域和目标领域的标签空间可能不同，这意味着源领域的知识不足以被充分使用。归纳式迁移学习包括源领域，即 D_S，以及一个学习任务，即 T_{SL}，连同目标领域 (D_T) 加上一个学习任务 (T_{TL})。归纳式迁移学习的目标是，通过利用源领域和任务中的信息，在已知 T_{SL} 不等于 T_{TL} 的情况下，改进目标领域中目标概率密度函数 PDF $f_{\mathrm{TP}}(.)$ 的学习。在归纳式迁移学习中使用的不同方法包括实例迁移[34]、特征表示迁移[35]、参数迁移[36] 和关系知识迁移[37]。

（2）传导式迁移学习：

在传导式迁移学习技术中，源领域存在大量标记数据，而在目标领域则没有标记数据。在这种设置下，源任务和目标任务是相似的，差异仅在于领域。根据源领域和目标领域之间的不同情况，传导式迁移学习中还会出现另外两种情况。在第一种情况中，认为源领域和目标领域的特征空间是不同的，即 $\dot{X}_{\mathrm{S}} \neq \dot{X}_{\mathrm{T}}$；在第二种情况中，认为源领域和目标领域之间的特征空间是相似的，但它们有不同的边缘概率密度函数，即 $P(X_{\mathrm{S}}) \neq P(X_{\mathrm{T}})$)。第二种情况讨论与领域变化有关的知识迁移。传导式迁移学习被用于脑电图信号的识别[38] 和频谱优化[39]。传导式迁移学习包括一个给定的源领域（D_S），它包含一个学习任务 T_{SL}，以及一个目标领域（D_T），它包含一个学习任务 T_{TL}。传导式迁移学习的目标是通过源领域和任务中的知识来帮助目标领域中目标概率密度函数 PDF $f_{\mathrm{TP}}(.)$ 的学习，前提是 $D_{\mathrm{S}} \neq \mathrm{D_T}$ 且 $T_{\mathrm{SL}} = T_{\mathrm{TL}}$。传导式迁移学习中使用的不同方法包括实例迁移和特征表示迁移。

（3）无监督迁移学习：

无监督迁移学习与归纳式迁移学习的相同点在于在源领域和目标领域都缺少标记数据。尽管如此，无监督迁移学习侧重于解决无监督任务，如口语语言理解[40] 和个人重新识别[41]。在训练中，源领域和目标领域都只有未标记的数据。无监督迁移学习包括一个源领域（D_S）及其学习任务 T_{SL}，以及一个目标领域（D_T）和其学习任务 T_{TL}。它的目的是通过利用源领域的知识来改进目标领域中目标概率密度函数 PDF $f_{\mathrm{TP}}(.)$ 的学习。在提到的情况下，假定 T_{SL} 等于 T_{TL}。根据无监督迁移学习的定义，在源领域和目标领域都没有标记数据。到目前为止，

用于无监督聚类的故障分类算法[42] 和图像质量评估算法[43] 是无监督迁移学习的经典应用。无监督迁移学习中使用的不同方法包括特征表示迁移，它处理关系领域的迁移学习。当给定的训练样本不足时，需要从更大的辅助数据集中转移学习到的特征表示。在自学习的工作中，具体是将聚类问题的另一个实例识别为自我教学聚类。自我教学聚类是无监督迁移学习的一个例子，其目标是在源领域有大量未标记数据的情况下，对目标领域中的较小未标记数据集合进行聚类。

2.2.4 迁移学习的挑战和本研究关注重点

根据上述内容，迁移学习目前面临着一些挑战。首先是领域适应问题，即如何将从一个领域学到的知识有效地迁移到另一个领域，尤其是在领域差异较大的情况下。其次是负迁移问题，即源任务中学到的知识可能会对目标任务的学习产生干扰。此外，如何选择和设计合适的迁移策略，以及如何评估迁移学习的效果，也是当前研究中需要解决的问题。

迁移学习的核心目标在于探索如何高效地运用一个领域的知识，以增强另一个领域学习任务的效果，进而提升机器学习算法在新领域的性能表现。具体到分类算法的研究，它着重于识别并利用源领域数据集中的样本、特征或参数等与目标领域相关联的知识，并将这些知识迁移至目标领域，以增强其分类的准确性[31]。迁移学习的未来研究方向可能包括：深入研究领域适应和负迁移的机制，发展更加有效的迁移策略；探索跨模态迁移学习，例如将视觉知识迁移到语言任务中；利用迁移学习来解决少样本学习、零样本学习等问题；以及结合元学习、强化学习和自监督学习等方法，进一步提升迁移学习的性能和效率。

迁移学习在多个领域都有广泛的应用。在计算机视觉领域，迁移学习可以用于图像分类、目标检测和图像分割等任务。通过使用在大规模数据集上预训练的模型，可以在数据量较小的新任务上获得很好的性能。在自然语言处理领域，迁移学习同样发挥着重要作用，例如在情感分析、机器翻译和问答系统等任务中。此外，迁移学习也被应用于语音识别、推荐系统、医疗影像分析等领域。

如图2.10所示，论文中第三章研究归纳式迁移学习方法，通过两种粒度协

同训练，构建两个局部注意力图神经网络进行多任务学习，学习系统可以有效提升视觉富文档在有监督条件下的关系抽取性能。在第四章研究传导式迁移学习方法，在低资源条件下，通过不同领域、不同类别的迭代训练，构建一个多模态数据元学习框架，通过对原型网络的矫正，极大提高关系类别表示的鲁棒性，提升视觉富文档在有少样本条件下的关系抽取性能。第五章，对于知识迁移的定义比较特殊，跟传统的知识迁移理论相比，是一种更泛化的迁移学习定义，因为它是针对两个模型的多系统。主要原理是将表格识别模型输出的 XML 格式代码，以一种中介媒介知识作为提示的形式迁移到大模型中。这里强调，这个定义只针对本研究，在图中用虚线标出。

第三章　基于跨粒度表示学习的视觉富文档关系抽取研究

3.1　研究动机

已有的布局感知多模态预训练语言模型在视觉富文档理解任务中表现出色，原因是这些模型在预训练过程中整合了经典的自然语言处理（NLP）和计算机视觉（CV）模型，并使用了细粒度的跨模态对齐约束损失函数来获得更好的多模态融合文档表示。但这些模型往往集中于文档中的细粒度 token 单元，例如单词和 token 级别的文档图像。这种做法导致模型难以从粗粒度语义实体中吸取知识，包括诸如短语、文本框、图像块等粗粒度级别（bounding box 级别）的语义实体。如图3.1所示，已有方法不能充分捕捉粗粒度（bounding box）级别的键值信息，这可能会造成抽取出的关系错误。本质上是由于没有进行跨粒度的知识迁移，已有方法不能同时对两种粒度的自然语言单元进行多任务学习。

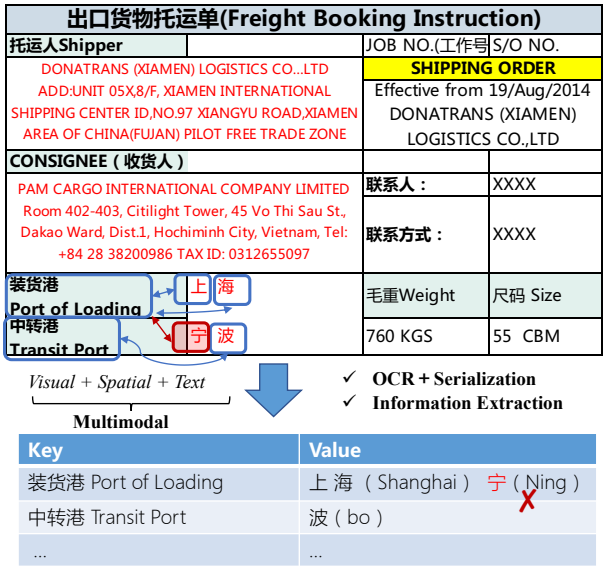

图 3.1　缺乏粗粒度级别学习的预训练模型进行关系抽取的错误案例

传统方法[44-45]通常采用两阶段的方式来进行信息抽取，将文本识别和信息抽取作为独立的阶段。在文本识别阶段，使用光学字符识别（OCR）技术来对文档

进行预处理，识别出文本框和其中的文本段，每一个文本框被叫做一个 bounding box。具体如图3.2所示，这些文本框 bounding box 的位置信息 position 与文本信息 text 一一对应。文本识别阶段大多是关于计算机视觉领域的研究。随后的信息抽取阶段专注于提取特定的感兴趣内容，例如有意义实体的文本段，并在序列化的文本输入中建立这些实体之间的关系。然而，从图像到文本的跨模态输入序列化可能会导致版式布局和视觉信息的丢失。为了克服这一缺点，最近的方法[9,46-48] 尝试结合额外的非文本特征，与文本特征共同建模文档的布局结构。研究人员普遍认为跨模态对齐对于多模态表示学习至关重要，但这些方法在实现 OCR 解析器输出的文本片段与计算机视觉模型生成的视觉表示之间的对齐有待进一步改进。

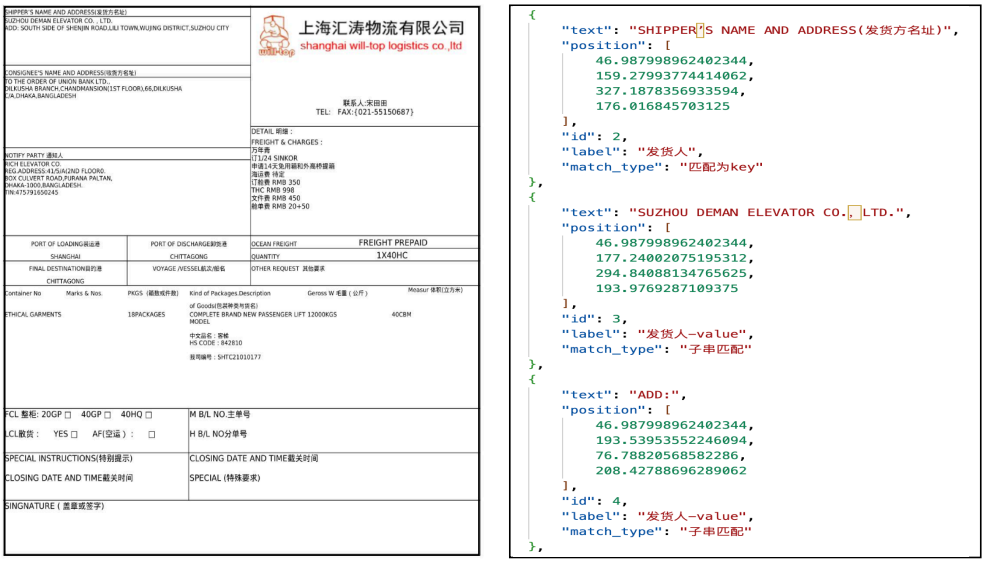

图 3.2 原始文档图像及标注数据（只截取部分）

基于上述研究缺陷，在文本识别阶段，本研究改进了原始 OCR 识别的文本框的排列顺序不规律的问题，并在重排的顺序中加入特殊的分隔字符。在信息抽取阶段，考虑到当前的预训练语言模型不能充分捕捉粗粒度的信息，如图3.3所示，本研究将已有预训练语言模型得到的 token 表示（细粒度）迁移到 bounding box 表示（粗粒度）并进行跨粒度的学习，使用图神经网络，将文档中每一个文本框 bounding box 视为图中的一个节点对文档进行表示和分析。通过该方法在 FUNSD、SEAB 和 XFUND 公开数据集上进行的广泛实验以及消融实验，成功

验证了所提出方法的有效性。

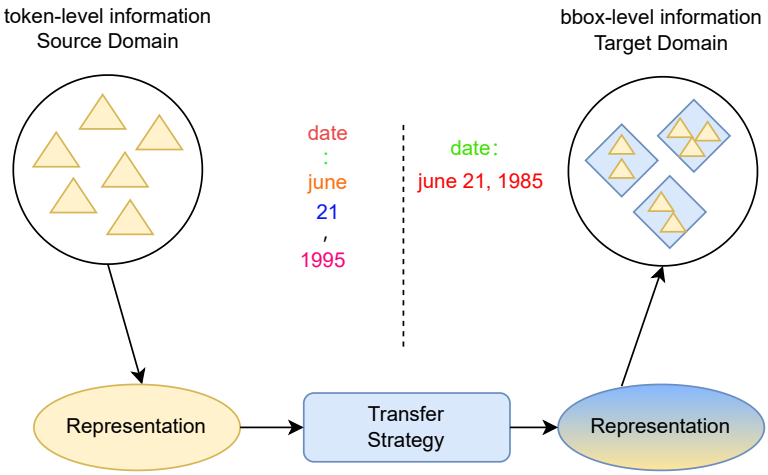

图 3.3　知识迁移策略

3.2　相关工作

关系抽取是视觉富文档理解任务的一部分。早期的文档理解工作基于单一文本模态或浅层多模态融合模型。这些模型包括基于文本的方法和基于图的方法。基于文本的方法依赖于大规模预训练语言模型的能力，例如 BERT[49]，XLM-RoBERTa[50]，InfoXLM[29]。RBERT[51] 通过将实体信息融入到预训练的语言模型中来进行关系抽取任务，具体如图3.4所示，过在目标实体对的前后添加特殊的分隔符（例如‘$’和‘#’），并将这些实体的信息融入到 BERT 模型中，利用句子以及目标实体表示来提高关系分类的性能。

但尽管如此，这些方法都忽视了布局和视觉特征且都是基于 token 细粒度的表示学习。得益于文档页面中的文本框比较分散，许多研究尝试将视觉富文档中的语义实体单元重新组织成图结构。由于图神经网络[52-53] 在处理关于非结构化数据的任务中取得了巨大成功，越来越多的研究集中于使用 GNN 来解决文档结构建模来提高文档理解能力。在基于图神经网络的方法中[9,46-48,54]，图神经网络被用来模拟文档的布局信息。通过节点（bounding box）进行粗粒度之间的信息传递，模型能够理解文档的布局。GraphIE[46] 使用图来捕获非局部和多模态特征。图卷积神经网络[9] 可以聚合节点之间无用和冗余的信息并且引入噪声，因为图中的每个节点都与其他节点相连。与完全连接图不同，PICK[47] 通过

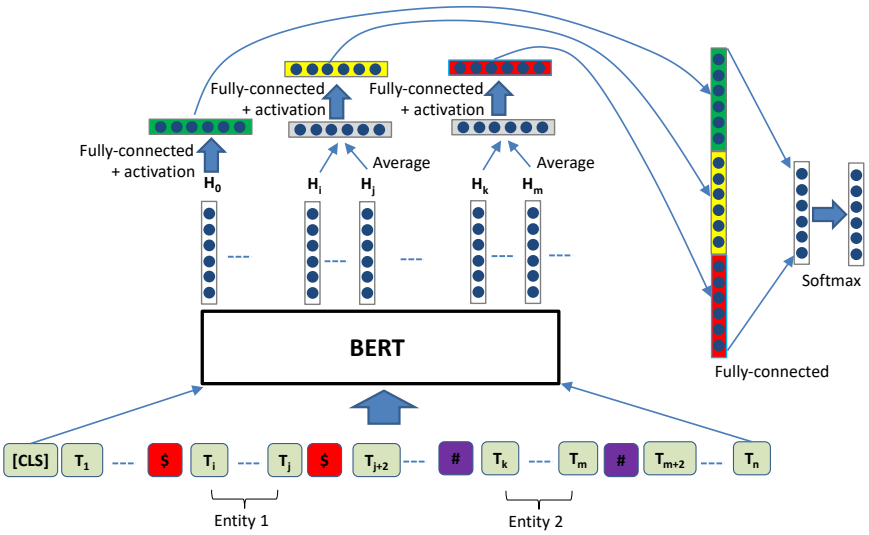

图 3.4　RBERT[51]的模型架构

图来预测节点之间的连接，并动态更新邻接矩阵的参数。受到上述工作的启发，研究提出的基于图注意力网络的方法将每对实体间的关系特征链接起来，并使用一个二分类器来预测两个实体之间是否存在关系。

在当前研究中，文档理解通常需要先进行 OCR 扫描，解析图像中的文本段和文本框的位置，然后将文本框及其对应内容按照 OCR 解析的默认顺序输入模型。然而，与普通文档图像不同，票据等文档通常不能按照传统的“从左到右”或“从上到下”的顺序排列。阅读顺序校正器[55-56]旨在捕捉文档的正确阅读顺序。通常，人类倾向于按从左到右、从上到下的顺序阅读文档。最近，为了在大量复杂文档上进行阅读顺序检测，研究人员提出了一种多模态网络[57]。XYLayoutLM[58]是一种增强型 XY Cut 算法，作为排序文本框以生成合理阅读顺序的增强策略，从而提高模型性能。而 Doctrack[59]是通过人类眼动仪标注的视觉富文档阅读顺序数据集，很好地模拟了人类的阅读顺序。受到上述工作的启发，本研究提出面向关系抽取任务的阅读顺序校正算法，旨在将有关系的实体在重排的阅读顺序序列中前后接近。

3.3 任务定义

本章节专注于有监督场景中的视觉富文档关系抽取。视觉富文档关系抽取主要指键值关系抽取。键值关系抽取与传统自然语言处理中的关系抽取类似，关系抽取是将键实体与其对应的值实体链接起来（如图3.1，目标任务是将“装货港”与“上海”链接起来），视觉富文档的键值关系抽取需要预测文档内任意两个语义实体之间的关系。

数学上可以定义如下：给定一个视觉富文档 D 和一组语义实体 $E = \{e_1, e_2, \ldots, e_n\}$，该任务旨在预测一组键值关系 $R = \{r_1, r_2, \ldots, r_m\}$，其中每个关系 r_i 由一个键实体 e_k 和一个值实体 e_v 组成，表明键实体与值实体相关联。

3.4 提出方法

3.4.1 整体框架

如图3.5所示，本研究提出的整体框架主要分为三个模块：（1）加入特殊分隔字符模块：文档的阅读顺序对于文档理解至关重要，本模块提出了阅读顺序重排算法，对于重排后的序列加入特殊分隔字符。本模块对输入序列做处理，从输入端进行改进。（2）多（跨）粒度图神经网络模块：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，本模块对于 token 粒度的图神经网络进行当前 token 与特殊分隔字符 token 的距离比较，提出了 token 粒度的局部注意力；对于 bounding box 粒度的图神经网络进行 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。（3）关系分类器模块：基于实体之间的关系得分，具体做法是判断每个文档中任意两个实体之间是否存在关系。在这种情况下，可以将关系预测作为二元分类任务，并使用二元交叉熵损失。

3.4.2 加入特殊分隔字符（Adding Special Tokens）

（1）阅读顺序校正算法

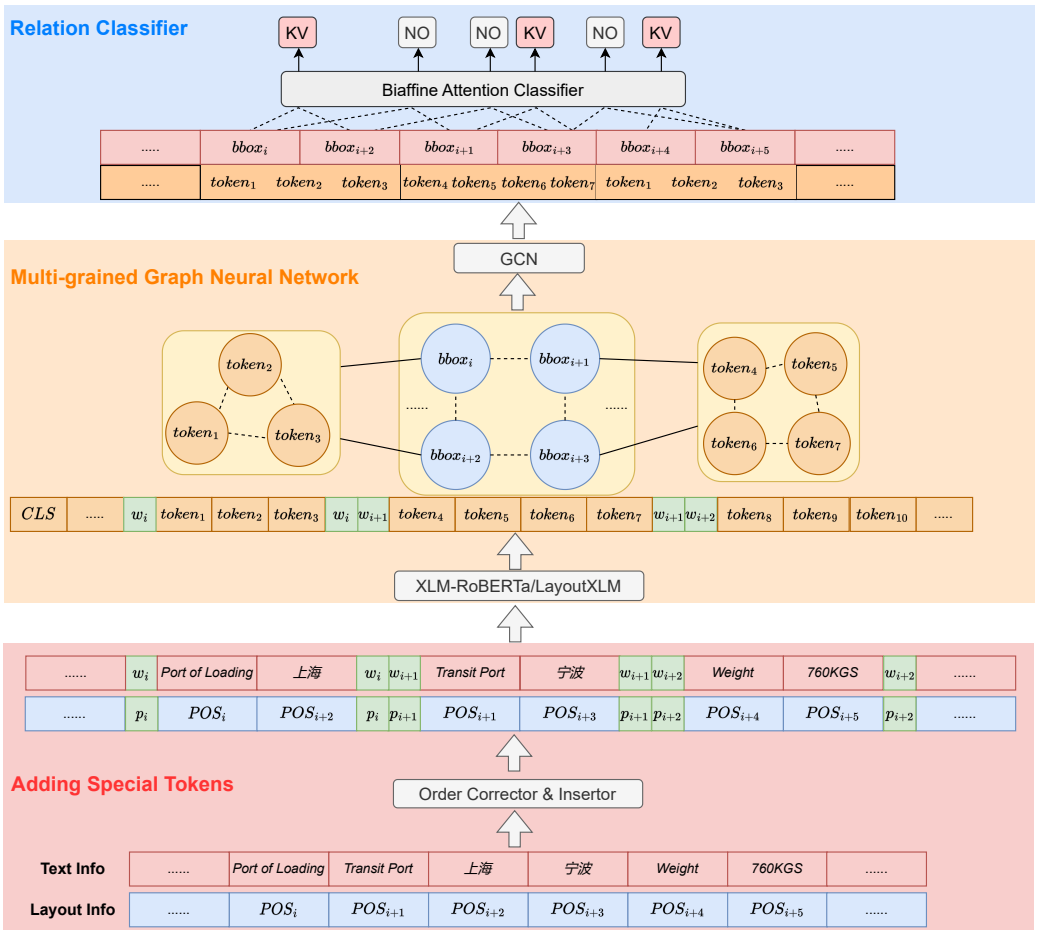

图 3.5　跨粒度迁移模型结构

在对纯文本进行关系抽取任务时，RBERT[51] 为每个目标实体之间插入特殊的分隔符，并利用句子以及目标实体表示进行关系抽取。受到这种方法的启发，同样需要将两个有关系的实体两边插入分隔字符，这就要求两个有关系的实体在重排的阅读顺序序列中前后接近。一个正确的阅读顺序可以实现这一点，由于视觉富文档内有关系的实体之间的位置具有上下或者左右之间的关系，于是可以充分利用这一特点。对于具有许多关键实体和值实体的文档 D，键实体很容易区分，因为它们的语义信息与它们自己的标签非常相似。如图3.5所示，第一个展示的文本框对应的文本段是“Port of Loading”，跟其标注的语义信息“装货港”非常相似。但值实体很难识别，就像“上海”和“宁波”一样，因为文本信息和布局信息在语义上太相似而无法区分。这一点在后面展示具体实体间关系性能的实验也可以看出来。基于这个问题，提出了一种阅读顺序校正算法，如算法1所示，保证对应的键值对在阅读顺序上前后接近。具体将阅读顺序序列

Algorithm 1 用于关系抽取的文档阅读顺序校正算法

Input: OCR 识别出的原始 bbox 序列 B，且每个 bbox 的文本段为 text，文本框为空间坐标:(x_1, y_1, x_2, y_2)，语义鉴别器 M

Output: 校正后的阅读顺序序列 B_r

初始化每一个 bbox 的 flag==0，定义一个缓冲区 $buffer$，用来保存有位置关系的一组 bbox，再定义一个 bbox 的集合 G 用来存放已经排序的 bbox，定义一个 c 用来保存当前的 bbox 经过语义鉴别器鉴别出的位置关系类型 ▷flag==0 表示没有被重排
for each bbox in B **do**
  if flag == 0 **then**
    if M(text) is O **then**
      将当前 bbox 插入到 G 中，且将 flag 改为 1 ▷ 该 bbox 经过语义鉴别器鉴别属于“没有位置关系”类型
    else
      将当前 bbox 加入到 $buffer$ 中，bbox 的 flag 改为 1,$buffer$ 中的 c=M(text) ▷ 该 bbox 经过语义鉴别器鉴别属于“有位置关系”类型，记为 c
      for each bbox in B **do**
        if flag == 0 **then**
          if M(text) is c **then**
            if c 属于上下关系 **then**
              flag == 1，将 bbox，$buffer$，G 都传入给 Algorithm 2 中处理
            else
              flag == 1，将 bbox，$buffer$，G 都传入给 Algorithm 3 中处理
            end if
          end if
        end if
      end for
    end if
  end if
end for

中的每个 bounding box（后面简写成 bbox）打上标记 flag，目的为了避免重复排序。算法将所有的关系分成三种：一种是像“装货港”类型的左右关系；一种是像“收货人”类型的上下关系；最后一种是像“出口货物托运单”类型的没有关系。根据每个 bbox 的语义类别，判断它的位置是左右关系还是上下关系还是没有关系。如果是左右关系，如算法2所示，将具备左右关系的键值实体对应的 bbox 重排；如果是上下关系，如算法3所示，将具备上下关系的键值实体对应的 bbox 重排。如果没有关系，则直接插入到重排序列中。最后得到的用来存放已经排序的 bbox 的集合 G 就是校正后的阅读顺序序列 B_r。

注意，由于具有关系的两个实体在二维空间位置上彼此相对接近，所以该算法是为关系提取而设计的。例如，读取“装货港”后，下一个边界框读取顺序可以校正为“上海”，而不是“中转港”。

Algorithm 2 左右位置关系排序算法

Input: bbox，缓冲区 $buffer$，缓冲区类型 c，已排序的 bbox 集合 G
Output: 修改后的 $buffer$ 和 G

if c is x **then**
  计算 $buffer$ 中所有 bbox 中的 x_1 和 x_2 的平均值 avg_x_1,avg_x_2，得到:$overlap = (avg\_x_2 - x_1) / (x_2 - avg\_x_1) * 100$
  if $overlap > \sigma_x$ **then**
    将当前的 bbox 放入到 $buffer$ 中
  else
    将当前的 $buffer$ 中的所有 bbox 的 flag=1，并且将它们放入到 G 中
  end if
else
  将当前 $buffer$ 中的所有 bbox 的 flag 改为 1，并且将他们都放入到 G 中，将当前 bbox 添加到 $buffer$ 中，且将 c 改为 y
end if

Algorithm 3 上下位置关系排序算法

Input: bbox，缓冲区 $buffer$，缓冲区类型 c，已排序的 bbox 集合 G
Output: 修改后的 $buffer$ 和 G

if c is y **then**
  计算 $buffer$ 中所有 bbox 中的 y_1 和 y_2 的平均值 avg_y_1,avg_y_2，得到:$overlap = (avg\_y_2 - y_1) / (y_2 - avg\_y_1) * 100$
  if $overlap > \sigma_y$ **then**
    将当前的 bbox 放入到 $buffer$ 中
  else
    将当前的 $buffer$ 中的所有 bbox 的 flag=1，并且将它们放入到 G 中
  end if
else
  将当前 $buffer$ 中的所有 bbox 的 flag 改为 1，并且将他们都放入到 G 中，将当前 bbox 添加到 $buffer$ 中，且将 c 改为 x
end if

（2）常识知识注入

记原始阅读顺序文本框序列 B，

$$B = [bd_1, ..., bd_i, bd_{i+1}, bd_{i+2}, ..., bd_n] \tag{3.1}$$

bd_i 表示第 i 个边界框，其布局信息表示为 $POS_i = (x_i^1, y_i^1, x_i^2, y_i^2)$。对于校正后的阅读顺序序列 B_r，键值实体前后接近。这样可以让模型捕获两个实体的本地化信息，在链接的键值实体的开头和结尾，插入特殊的分隔字符 ω，特殊 token 的布局信息是根据键值实体布局信息生成的。如图3.5所，对于 B_r：

$$B_r = [..., \omega_i, bd_i, bd_{i+2}, \omega_i, \omega_{i+1}, bd_{i+1}, ...] \tag{3.2}$$

两个特殊标记 ω_i 围绕着 bd_i 和 bd_{i+2}，它们的布局信息共享为

$$p_i = (Min(x_i^1, x_{i+2}^1), Min(y_i^1, y_{i+2}^1), Max(x_i^2, x_{i+2}^2), Max(y_i^2, y_{i+2}^2)) \tag{3.3}$$

这意味着特殊的分隔字符的二维坐标可以包围有关系的键值实体。

3.4.3 多粒度图神经网络（Multi-grained Graph Neural Network）

根据之前的工作[10]，本研究采用 XLM-RoBERTa/LayoutXLM 作为主干网络来获得 token 细粒度级别的表示，LayoutXLM 包含有多种模态的信息，包括文本、布局和视觉。在预训练阶段，LayoutXLM 通过计算各种模态之间的损失函数来建立深度的跨模态交互。请注意，由于图像信息就是布局位置信息产生的，方便起见，图3.5中没有展示图像信息。对于 token 细粒度级别，给定一个图 $G = (V, E)$，G 代表需要建模的图神经网络，其中 $v_i \in V$ 和 $i \in \{1, 2, ..., n\}$。v_i 表示文本框分词之后的细粒度单元。通过忽略离 ω 特殊字符节点比较远的节点信息，以确保键值实体的文本框之内细粒度单元的相关性。这样可以将有关系的实体信息融入到 LayoutXLM 模型中，使得 LayoutXLM 专注于本地化信息。所以，对于键值文本框，鼓励键值文本框之内的细粒度单元（token）之间的交互，同时防止特殊字符 ω 以外的信息，具体如图3.3，鼓励 “date”、“:”、“june”、“21”、“,”、“1995” 之间的交互。这会产生有位置关系文本框的完全连接的子图。$e_{i\tilde{i}}$ 表示细粒度单元（token）在 v_i 和 $v_{\tilde{i}}$ 之间的边，由以下公式计算得到，

$$e_{i\tilde{i}} = \begin{cases} 0, & \text{if } distance(v_i, v_{\tilde{i}}) > distance(v_{\tilde{i}}, \omega_i) \\ 1, & \text{if } distance(v_i, v_{\tilde{i}}) <= distance(v_{\tilde{i}}, \omega_i) \end{cases} \tag{3.4}$$

这意味着键值文本框中的 token 细粒度单元节点已连接，但它们均不与键值文本框外的任何 token 细粒度单元连接。对于所有 token 细粒度单元，使用局部注意力机制执行图卷积：

$$h_i{}^{(t+1)} = \delta(\sum_{\tilde{i}=1}^{N} \alpha_{i\tilde{i}} W h_{\tilde{i}}^{(t)} \otimes e_{i\tilde{i}}) \tag{3.5}$$

其中 $h_i{}^{(t+1)}$ 是通过 $h_i{}^{(t)}$ 的聚合操作得到的。$h_{\tilde{i}}$ 是节点 v_i 的领居节点 $v_{\tilde{i}}$ 的隐藏层。$\otimes$ 表示逐元素乘法运算。δ 是一个激活函数。$\alpha_{i\tilde{i}}$ 是注意力系数来表示节点

i 关于节点 $\tilde{i}$ 的重要性。注意力系数的计算方式为：

$$\alpha_{i\tilde{i}} = \frac{exp(\delta(V^T[Wh_i \oplus Wh_k]))}{\sum_{n\in N} exp(\delta(V^T[Wh_i \oplus Wh_n]))} \tag{3.6}$$

对于 bounding box 粗粒度级别，具体如图3.3，鼓励 “date :” 和 “june 21 , 1995” 之间的交互。粗粒度单元的最初表征通过 bbox 内部的 token 表征平均化得到，然后对 bbox 级别实施 K 近邻图卷积神经网络以学习最终的表征。与之前的工作[9] 不同，他们构建了一个完全连接的图神经网络，本研究提出了一个 K 最近邻（KNN）图注意力卷积神经网络，通过自注意力层仅关注其邻居节点，尤其是二维空间布局位置上接近的节点。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：

$$\tilde{h}_i^{(t+1)} = \delta(\sum_{k\in Nt(M,i,K)} \gamma_{ik} W\tilde{h}_k^{(t)}) \tag{3.7}$$

k 代表 v_i 的邻居节点之一。γ_{ik} 代表节点之间的相关性。$Nt(M,i,K)$ 是 v_i 的 k 近邻索引的集合。$\tilde{h}_k^{(t)}$ 是节点 v_k 的相邻节点在时间步 t 的隐藏层表示。$\tilde{h}_i^{(t+1)}$ 由相邻节点的特征以及它自己的特征获得。W 是线性层，δ 是激活函数。研究还使用多头注意力机制来提高性能，具体图卷积使用 2 层，使用四头注意力机制。任意 k 到 i 的 γ_{ik} 系数为：

$$\gamma_{ik} = \frac{exp(\delta(V^T[W\tilde{h}_i \oplus W\tilde{h}_k]))}{\sum_{m\epsilon M} exp(\delta(V^T[W\tilde{h}_i \oplus W\tilde{h}_m]))} \tag{3.8}$$

W 和 V 变量是可训练参数。然后将不同图卷积层的表示连接起来形成最后的表示。通过知识迁移的方法，将跨粒度的图神经网络进行多任务学习表示，提高关系抽取的性能。

3.4.4 关系分类器（Relation Classifier）

此外，提出方法利用每个语义实体的标注信息，将实体标注映射到嵌入表示中。我们连接实体表示和标注嵌入表示：

$$e_i = h_i \oplus l_i \tag{3.9}$$

其中，h_i 是键实体的第一个 token 或值实体的最后一个 token 的表示。l_i 是标注嵌入表示。使用双仿射变换注意力之后，多层感知机可以用来丢弃对当前关系决策没有价值的信息。对键值实体分别使用两个多层感知机来生成每个关系链接中键和值实体的不同表示：

$$h_i^{key} = F(W^{key}e_i + b^{key}) \tag{3.10}$$

$$h_i^{value} = F(W^{value}e_j + b^{value}) \tag{3.11}$$

$$Score(i,j) = h_i^{key}W_1h_j^{value} + (h_i^{key} \oplus h_j^{value})W_2 \tag{3.12}$$

因为这种双仿射机制可以更好地捕获实体对之间的关系。

3.5 实验

3.5.1 数据集介绍

实验使用 XFUND[10]，FUNSD[60] 和 SEAB[61] 数据集。FUNSD 数据集是一个专门用于理解噪声扫描文档中表单的公共数据集。它的目的是从这些表单中提取和结构化文本内容。数据集包含 199 个真实的、完全注释的扫描表单，这些表单来自不同的领域，如市场报告、广告和科学报告等。这些文档通常噪声较大，外观差异显著，因此对表单理解是一个挑战性的任务。FUNSD 数据集可用于多种任务，包括文本检测、光学字符识别（OCR）、空间布局分析以及实体链接（关系抽取）等。SEAB 数据集由 4515 张海运单图像组成，包含 22 种关系。关系包括发货人、收货人、货物体积、重量等。数据集的样例如图3.2所示。XFUND 数据集是一个多语言表单理解基准数据集，它包含了 7 种不同语言的表单数据，并为每种语言手动标注了键值对。这些语言包括中文（ZH）、日语（JA）、西班牙语（ES）、法语（FR）、意大利语（IT）、德语（DE）和葡萄牙语（PT）。XFUND 数据集的目的是为了评估和改进多语言环境下的文档理解技术，特别是在处理视觉富文档时跨越语言障碍的能力。本章仅关注原始论文[10] 中定义的关系抽取任务，旨在预测任意两个给定语义实体（数据集有“header”、“question”、“answer”、“other” 四个语义实体）之间的关系。

3.5.2 实验环境与细节

该模型使用 Adam 优化器进行训练，学习率为 0.0005，批量大小设置为 16，最大文档序列长度设置为 512，dropout 参数设置为 0.1，将 σ_x 设置为 0.3 来衡量两个有关系实体左右有多接近，σ_y 设置为 0.5 来衡量两个有关系实体上下有多接近。

3.5.3 实验结果

如表3.1所示，与传统的关系抽取方法相比，通过引入特殊分隔字符和多（跨）粒度图神经网络，实现了更高的性能。XLM-RoBERTa 和 InfoXLM 在这些数据集上效果不佳，可能是因为缺乏多模态信息。值得注意的是，在 XLM-RoBERTa 和 InfoXLM 中，虽然没有考虑图像特征，但在添加特殊字符并进行跨粒度的多任务学习之后，关系抽取性能得到了很大的提高。同时，LayoutXLM 在同时使用文本、布局和图像信息时取得了更好的性能。提出的方法取得了最佳结果，并且明显优于仅仅使用 LayoutXLM 预训练模型。注意，研究所做实验在这些真实数据集中使用与 LayoutXLM 论文[10] 中相同的评估指标。

表 3.1　在 XLM-RoBERTa[50]、InfoXLM[29] 和 LayoutXLM[10] 三个基线模型上的消融实验。**+ST** 代表在基线模型基础上增加了特殊字符（Adding Special Token），**+MG** 代表经过跨粒度知识迁移，在基线模型上加入了多粒度神经网络（Multi-grained Graph Neural Network）

Method	Modality	SEAB	FUNSD	ZH	JA	ES	FR	IT	DE	PT
XLM-RoBERTa[50]	Text	46.22	47.76	51.05	58.01	52.95	49.65	53.05	50.41	39.82
XLM-RoBERTa+ST	Text	46.75	48.34	51.65	58.48	53.49	49.98	53.74	52.14	40.31
XLM-RoBERTa+ST+MG	Text	47.37	49.93	52.23	60.08	54.25	51.07	54.58	52.97	42.11
InfoXLM[29]	Text	50.11	49.68	52.14	60.00	55.16	49.14	52.81	52.62	41.70
InfoXLM+ST	Text	50.54	49.89	53.42	61.13	55.83	50.31	54.08	54.12	43.56
InfoXLM+ST+MG	Text	51.34	50.51	54.51	61.79	55.96	51.48	54.90	54.78	43.87
LayoutXLM[10]	Text+Layout+Vision	67.31	74.33	70.73	69.63	68.96	63.53	64.15	65.51	57.18
LayoutXLM+ST	Text+Layout+Vision	67.73	74.81	71.60	69.85	71.21	64.68	68.31	66.36	58.42
LayoutXLM+ST+MG	Text+Layout+Vision	**68.54**	**75.54**	**72.51**	**70.33**	**72.06**	**65.42**	**69.18**	**67.05**	**59.21**

3.5.4 实验分析

（1）预定义实体类别的性能：

如图3.6所示，实验发现键（Key）实体的性能不及值（Value）实体，表明对于键（Key）实体的识别要比值（Value）实体容易得多。这进一步说明对键实体的识别可以帮助值实体的识别，因为值实体在二维空间位置上通常在键实体的右方或者下方。同时，这也解释了为什么要去构建面向关系抽取任务的阅读顺序校正算法。

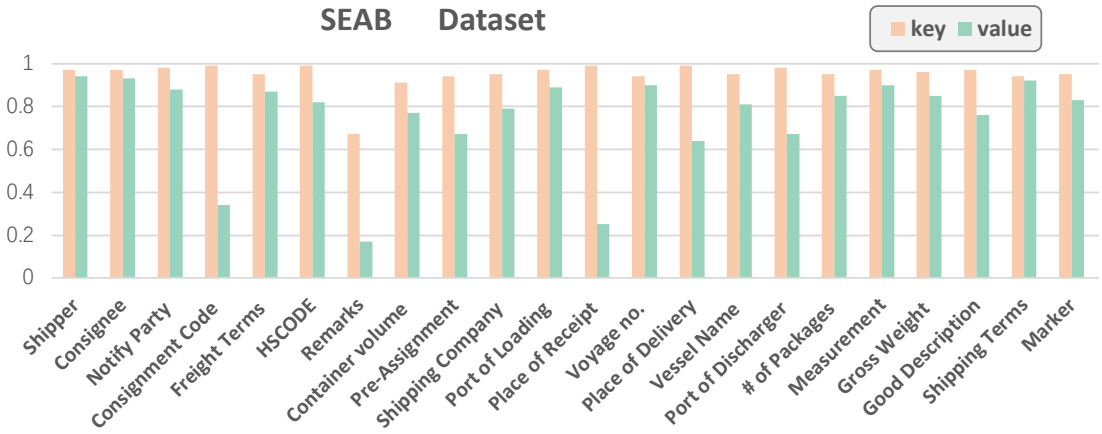

图 3.6　SEAB 数据集中预定义键值实体的抽取性能

（2）阅读顺序的影响：

经过阅读顺序校正算法后，提出方法在校正后的序列中添加特殊分隔字符。由于键值实体在二维空间中存在由左到右或者从上到下的关系，所以如果能在键值实体两边加上特殊分隔字符（special token），自然可以使模型更关注键值实体自身的关系，这可以被认为是一种局部注意力。根据 Doctrack[59]，一个使用眼动追踪技术来模拟人眼运动轨迹的视觉富文档阅读顺序数据集，我们设计了两个评估指标，衡量重新构建后的阅读顺序与人类阅读顺序的接近程度。实验计算重新构建后的顺序和人眼运动顺序之间的相似度，假设人类眼动阅读顺序为

$$Idx_{eye} = [idx_1^e, ...idx_i^e, ...idx_n^e] \tag{3.13}$$

重新构建的阅读顺序是

$$Idx_{re} = [idx_1^r, ...idx_i^r, ...idx_n^r] \tag{3.14}$$

全局相似度度量为

$$M_{global} = 1/n \sum_{i}^{n} |idx_i^e - idx_i^r| \tag{3.15}$$

局部相似度度量为

$$M_{localized} = 1/N \sum_{i}^{i \notin O} |idx_i^e - idx_i^r| \tag{3.16}$$

表 3.2　已有的阅读顺序排列算法与 Doctrack 数据集[59] 的相似度度量比较

Dataset	Reading Order	Global	Localized
XFUND[10]	Default OCR	4.36	3.37
XFUND[10]	XYCut[58]	2.68	1.73
XFUND[10]	ours	**2.66**	**1.31**
SEAB[61]	Default OCR	8.36	7.41
SEAB[61]	XYCut[58]	**2.43**	1.84
SEAB[61]	ours	2.46	**1.65**

局部阅读顺序相似度度量只关注键值实体的前后顺序。所以 N 是标注为键值关系实体的文本框数量，而 n 是文档中所有文本框的数量。在以上公式中，idx_i 表示第 i 个文本框在阅读顺序序列中排第几个，idx_i 介于 1 到 n 之间。如表3.2所示，提出的阅读顺序重排算法与人类阅读顺序在局部阅读顺序相似度度量上非常接近。

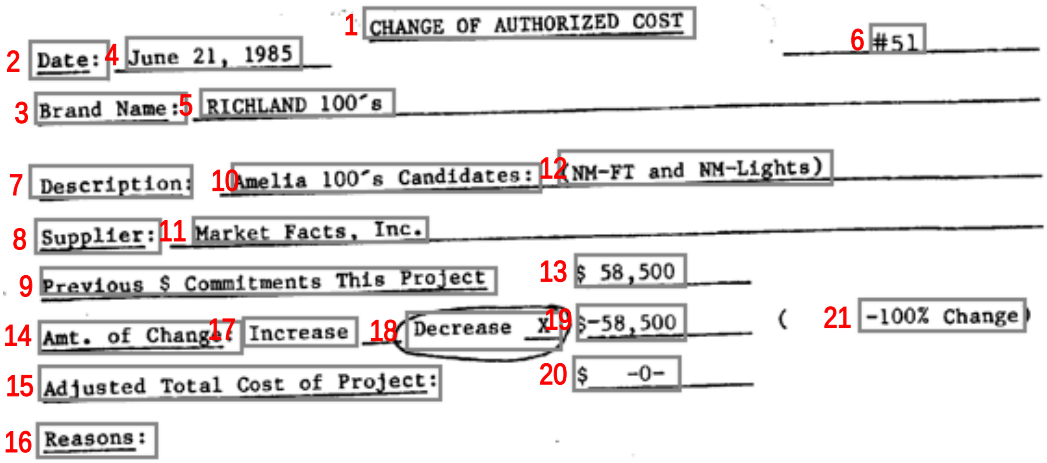

图 3.7　通过 OCR 识别到文档的原始阅读顺序

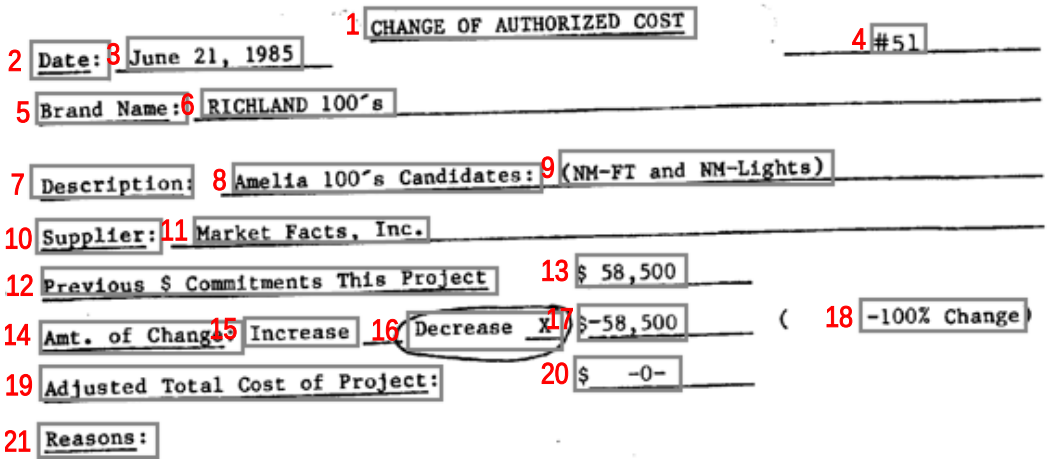

图 3.8　经过阅读顺序校正算法之后的文档阅读顺序

如图3.7和3.8所示，这是有没有经过阅读顺序校正算法之后的阅读顺序的比较。图中红色的数字就代表公式中的 idx_i。可见，经过阅读顺序校正算法之后的序列非常接近人类的眼动顺序。而全局阅读顺序相似度度量在关系抽取这项任务中似乎没有衡量的意义。

（3）注意力机制参数设定：

在 token 细粒度级别，使用加入特殊分隔字符的方法来实现局部注意力，这种方法可以根据当前 token 与特殊分隔符的距离建立图神经网络，而在 token 级别中的 KNN 方法，表示根据与当前 token 最近的 K 个 token 字符做连接。如表3.3所示，KNN 方法的表现不如加入特殊分隔字符的方法。而在 bounding box 粗粒度级别，使用 k 近邻的图神经网络来捕捉键值实体之间的关系信息。实验表明，当 k 为 4 时效果最好，这可能是因为与当前 bounding box 节点相联的上侧、下侧、左侧和右侧的 4 个节点彼此之间的关系最为紧密。

（4）可视化分析：

为了直观地解释加入特殊分隔字符（ST）与多（跨）粒度图神经网络（MG）的有效性，可视化实验从 SEAB 数据集中随机选取了 50 个实例（”装货港” 键值关系和” 转运港” 键值关系），并在训练后将它们编码为隐藏层的嵌入表示。然后，使用主成分分析（PCA）将它们映射到同一特征嵌入空间中。如图3.9（左）所示，这是仅使用 LayoutXLM 作为预训练模型直接预测的结果，可视化结果显示，两种关系表示的散度很高，特征嵌入分散。为了解决这个问题，本研究提出了跨粒度的图神经网络。如图3.9（中）所示，加入了跨粒度图神经网络的

表 3.3　不同粒度局部化方法在不同参数下的比较

Settings	Grain-level	SEAB	FUNSD
ST	token	**67.73**	**74.81**
KNN(k = 1)	token	67.54	74.33
KNN(k = 2)	token	67.48	74.45
KNN(k = 3)	token	67.41	74.57
KNN(k = 1)	bbox	67.43	74.37
KNN(k = 2)	bbox	67.64	74.47
KNN(k = 3)	bbox	67.71	74.84
KNN(k = 4)	bbox	**68.12**	**75.05**

LayoutXLM 的特征空间分布变得密集，主要是由于 bounding box 粗粒度的节点更关注其最近的节点，关注局部的信息。再如图3.9（右）所示，在 token 细粒度级别加入了特殊的分割字符（ST）之后，不同的关系表征就会更容易区分，不同关系在特征空间中的表征被进一步拉开。

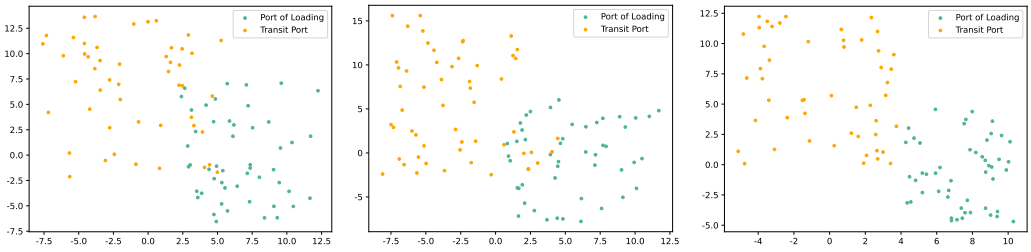

图 3.9　以 LayoutXLM 为主干网络的嵌入空间可视化消融实验

3.6　本章小结

本章提出了一种新颖的跨粒度、多任务学习的局部注意力模型，用于视觉富文档中的关系提取。该模型利用了在 token 细粒度级别添加的特殊分隔字符及对应的根据特殊字符距离判断的图神经网络和 bounding box 粗粒度级别的 K 最近邻图神经网络。此外，鉴于版面结构复杂多样，很难表示 OCR 结果的正确读取顺序，为了加入特殊分隔字符还解决了视觉富文档中原始 OCR 结果的前后顺序问题提出了一种面向关系抽取任务的顺序校正算法，以构建正确的阅读顺序。提出的基于跨粒度知识迁移的视觉富文档关系抽取方法在三个真实世界数据集上得到验证。在配备有提出方法的主干模型（LayoutXLM）上进行实验，

实验分析证明了所提出方法有效的原因。

第四章　基于跨类别元学习的视觉富文档关系抽取研究

4.1　研究动机

如图4.1所示，视觉富文档在我们的生活中非常常见，包含有宽泛的领域，比如金融，药物，海运物流传单，商品使用说明书以及化学等领域。其中，键实体（Key）与值实体（Value）之间的关系在视觉富文档中普遍存在。最近的文档理解人工智能模型[7-8,10,62-64]在大规模的扫描文档数据集上进行了预训练，通过利用多模态信息，表现出了良好的性能。但涉及到实际应用时，由于视觉富文档布局格式及样式的多样性和复杂性，如何使预训练模型能够适应这些不同布局格式和未知的实体关系类别[65]是一个重大的挑战。这些预训练模型在低资源情况下，缺乏自动检测和识别不熟悉领域中新型实体间关系的固有能力。具体如图，已有大量已知的日期、公司实体间的键值关系这些有标注的清单数据，现在需要去学习新领域的药品名称、毛重、易燃成分间的键值关系。相比之下，人类在这项任务中具有非凡的能力，通过分析几行内容便可以快速理解各种键值关系。文档模型本身在检测和识别新领域中的实体类型和关系上存在缺陷，特别是在缺少标注数据的情况下。本章节的研究动机之一就是让模型像人一样，在已经学会识别已知键值关系的情况下，仅依靠少样本快速理解新文档中的新键值关系。

虽然针对低资源条件下的视觉富文档理解任务取得了进展[54,66-67]，但是目前还没有工作去建立视觉富文档少样本关系抽取这一现实任务的系统框架。本研究首次在有监督数据集（CORD[68]和 SEAB 数据集）基础上重构了两个少样本数据集，建立了多模态数据中跨类别元学习的关系抽取框架。为了让机器像人一样去快速理解视觉富文档中的键值关系，有效利用空间布局特征，考虑到键实体和值实体通常具有固定的位置和排列，本研究在提取键值实体及其关系的三元组时，会结合二维空间位置特征提供的监督互补信号。另外一个研究动机在于，由于少样本学习中样本方差较大，迫切需要一个多模态融合机制，提高

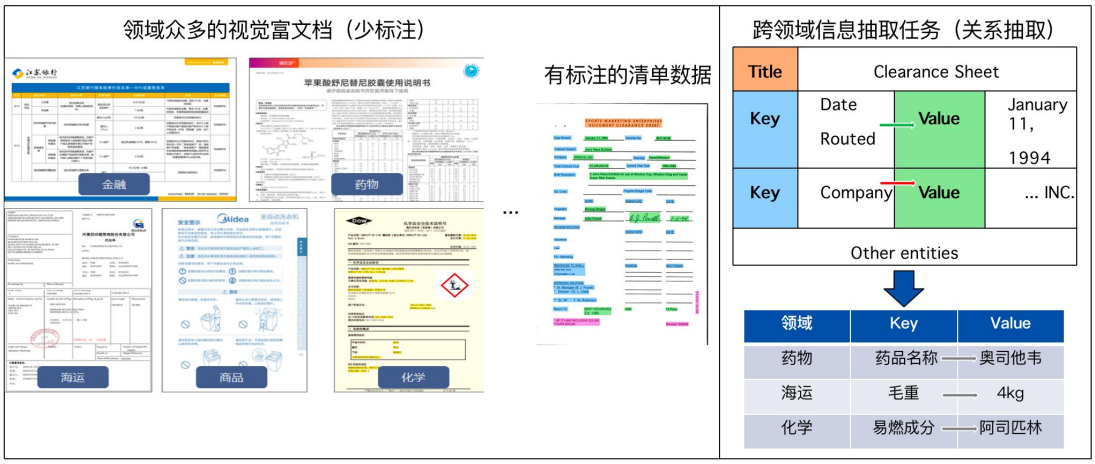

图 4.1 多领域视觉富文档的低资源应用场景

关系类别原型表征的鲁棒性。虽然已经有研究证明多模态信息可以极大增强监督学习信号[62]，但如何仅仅通过少样本实例来对齐并聚合不同模态的特征仍然是一个未解决的问题，凸显了研究人员在该方向上进一步探索的必要性。注意，本研究认为，少样本学习与元学习是相同的概念，在学术界也是这样认为的。并且，元学习本身就是一种知识迁移方法。

4.2 相关工作

在自然语言处理领域中，最近关于元（少样本）学习关系抽取的研究主要关注使用单一文本模态信息[69-70]。这些工作都会遵循 N-way K-shot 的少样本设定[71]，其中一个关系实例需要根据每个类的 K 个实例来判断并作分类。评估指标基本按照 FewRel[69]、FewRel2.0[72] 和 Few-Shot TACRED[73] 这样的少样本关系抽取基准集（benchmark）进行。虽然已有模型取得了超过人类的性能，但是相关工作[73-74] 已经证明现有的基准集与现实世界的应用场景相差甚远，远远没有现实世界来得复杂。所以，在多模态少样本关系抽取领域内还有很多挑战，其中包括减少对实体提及信息的依赖问题。还有一些挑战包括如何从句子级别的关系抽取过渡到文档级别的关系抽取[75]，因为这两者需要不同的模型架构，从而出现新的问题。

鉴于少样本关系抽取的目标是通过训练已知的关系类别来快速适应到仅有少量样本的未知关系类别，度量学习成为该任务的主要范式。经典的方法包括

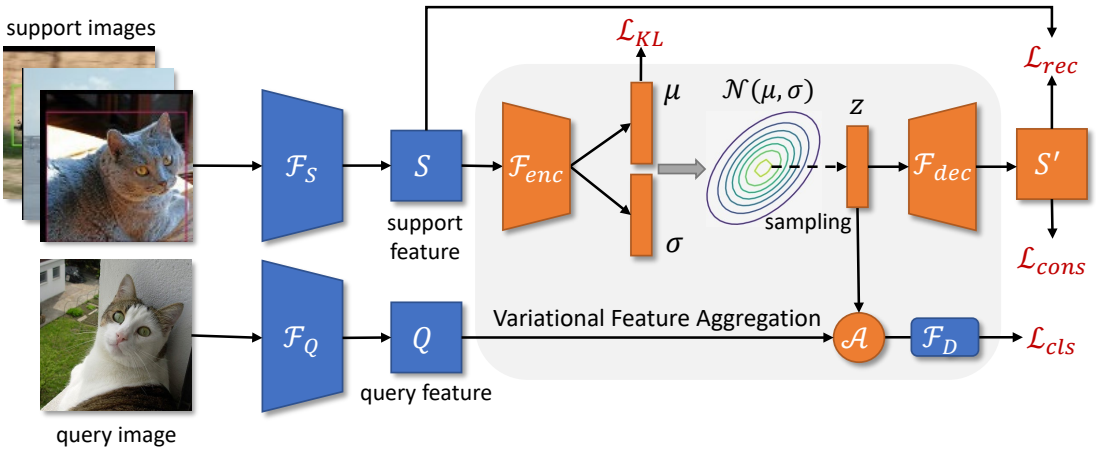

图 4.2　VFA 模型[78] 架构图

原型网络及其衍生方法[71,76]，例如原型矫正方法[77]。这些方法可以学习每个类别的原型表征，并根据测试过程中的待预测样本的表征与原型表征的相似度进行分类。如图4.2所示，在计算机视觉领域，有研究通过变分特征聚合进行少样本目标检测[78]。在多模态领域，有一些以视觉富文档为研究对象的文档级别关系表示学习框架[65]被提出。在视觉富文档中，已有工作还可以通过原型矫正[77]的方法，来增强关系原型的表征。本研究工作受到在图像领域变分特征聚合[78]方法的启发，首次在多模态数据中使用变分自编码器拟合数据分布特征。

4.3　任务定义

如图4.3所示，本章任务沿用了之前工作中的少样本设定[69-70]，旨在通过使用少样本示例对已知关系类别进行训练，目的是在新的视觉富文档 $\mathcal{D}$ 中联合提取语义实体的键和值，该图展示了本工作和前人工作在少样本关系学习时在一次迭代过程中的不同点[75]。在测试任务中，少样本学习目的是根据给定的查询集文档来抽取出该文档的实体和关系类型组成的三元组。注意，测试任务中的关系类别和训练任务不重合。传统方法通常依赖现成的 OCR 引擎从原始文档图像中提取文本，并且仅依赖文本特征来提取关系三元组。相比之下，本工作模拟人的视角，并利用多模态信息来有效地提取关系三元组。在图中，本研究使用布局简单的收据（CORD 数据集）来说明这一想法，但现实世界的场景要复杂得多，更具挑战性。

重要的数学符号如表4.1所示。

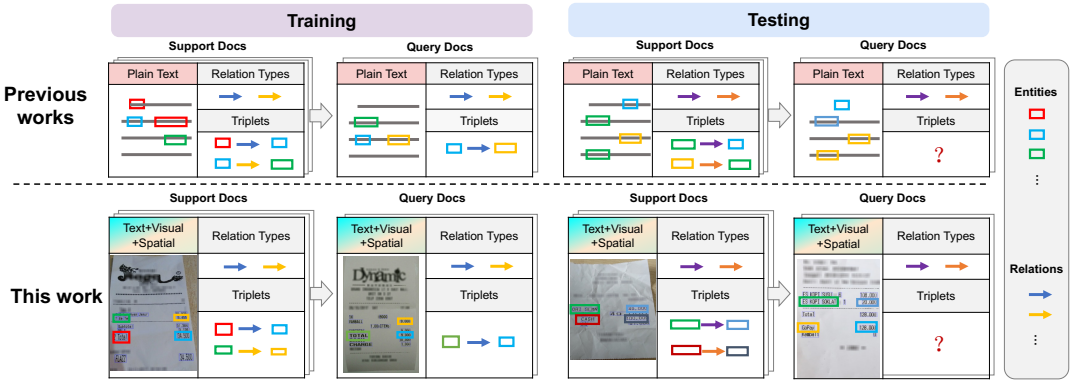

图 4.3　多模态数据的少样本学习与纯文本少样本学习（前人工作）的不同

文档 $\mathcal{D}$ 通常由扫描图像和一组文本框组成，扫描图像定义为 $\mathcal{I}$，一组文本框组成的列表定义为 $\mathcal{B}=\{b_1,\ldots,b_N\}$，其中 b_i 对应一段由 OCR 识别到的文本段，有 N 个 tokens，这 N 个 tokens 对应文本框的二维位置坐标 (x_0,y_0,x_1,y_1)，(x_0,y_0) 和 (x_1,y_1) 代表这个文本框的左上角和右下角的坐标。这些文本段和文本框是在文档预处理阶段通过利用 OCR 引擎获得的。每个经过 OCR 扫描的视觉富文档都是由语义实体列表组成，每个实体由一组词和文本框组成。定义 bounding box 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 的对角线上的坐标。所使用的数据集都有每个实体的标签和实体之间的关系标签信息。我们将每个注释文档表示为 $\mathcal{D}=\{[b_1,...,b_n],[l_1,...,l_n],[(b_1,b_{h_1}),...,(b_m,b_{h_m})]\}$，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

表 4.1　重要的数学符号及其含义

标记	含义
$\mathcal{D}$，$\mathcal{I}$	一个给定的文档及对应的图片
$\mathcal{B}$	由 OCR 引擎识别得到的文本框（bounding box）的序列
$\mathcal{T}$	由 OCR 引擎识别得到的文本段（text segment）的序列
L	所有的实体类型的集合
$\mathcal{R}$	所有的关系类型的集合
t_i,b_i,r_i	第 i 个 token，bounding box 和关系类型标注
$\mathbf{h}_i,\mathbf{H}$	具有多模态特征的输入隐藏层嵌入表示
$\mathbf{h}'_i,\mathbf{H}'$	重构之后的隐藏层嵌入表示
$\tilde{\mathbf{h}}_i,\tilde{\mathbf{H}}$	具有二维空间先验的隐藏层嵌入表示
x,y	x 轴和 y 轴的坐标
$\mathcal{S},\mathcal{Q}$	支持集，查询集

的是，实体可能与多个实体存在关系，或者与其他任何实体没有关系。测试集实例通常包括训练集中不存在的关系类型。

为了简单性和实用性，本研究仅考虑给定文档中单个关系中的键和值实体作为支持集/测试集实例。然后采用一个坍塌型序列标注模型[79-80] 从文档 $\mathcal{D}$ 中联合提取给定关系类型 $\mathcal{R}$ 的键和值实体集 ($\mathcal{E}_k$ 和 $\mathcal{E}_v$)。这是通过扩展先前的命名实体识别任务中实体类型的标注空间来实现的。例如，在“Cash”（CORD）和“Consignee”（SEAB）实体类型基础上加入后缀“-Key”和“-Value”，在本章中，只考虑使用基于 token 的分类方法来抽取实体提及，所以使用标准的“BIO”标注模式（“Begin, Inside, Other”）。也就是说，“Consignee-Key-B”代表实体提及“Consignee-Key”的首个 token，而“O”代表“Other”类，不属于任何的实体提及。

4.4 视觉富文档少样本数据集

4.4.1 针对关系的采样策略

在基于跨类别元学习的视觉富文档关系抽取这一现实任务的系统框架内，如何合理采样满足少样本学习 N-way K-shot 的设定是一个问题。在采样过程中，应该仔细选择并组织文档中的关系实例，以确保 N-way 的设定能够维持，同时捕捉必要的上下文信息。本研究的采样策略旨在让文档级别的上下文和 N-way 设定这两个约束之间取得平衡，从而为本章元学习框架提供支持。我们使用包含原始文档多个副本的扩展数据集，从这个扩展数据集中采样，在每次采样步骤后将一个转换后的文档添加到支持集或查询集中，直到支持集和查询集达到设定的关系类别数量（N-way）和每个类别实例的数量（K-shot）。采样方法的整体流程在算法4中。

具体来说，提出的元学习框架需要在重构的少样本数据集（包含有多个副本的扩展数据集）上进行，采用 N-way K-shot 的设定来训练少样本关系学习模型，这需要迭代地采样 N 个类，每个类 K 个关系实例来组成支持集及查询集。这是因为在文档中通常包含多个关系/实体，这可能会超出 N-way 的限制。为了保持 N-way 设定，如图4.4所示，会单独对每个文档创建关系特定的副本，这意味着每

Algorithm 4 Relation-wise N-way K-shot 采样策略

Input: 文档 $\hat{\mathcal{D}}$, N, K, K';
Output: 支持集 $\mathcal{S}$, 查询集 $\mathcal{Q}$;
$\mathcal{S} \leftarrow [\,], \mathcal{Q} \leftarrow [\,]$; ▷ 初始化
for $j \leftarrow 1$ **to** $N/2$ **do**
  $\mathcal{S}[j] \leftarrow \{\}$; ▷$N/2$ 是关系类型的数量
  $\mathcal{Q}[j] \leftarrow \{\}$; ▷$N$ 是实体类型的数量
end for
repeat
  从拓展后的掩码数据集中随机采样 $(\mathcal{D}^{(i)}, \mathcal{R}^{(i)})$ ▷ $\mathcal{D}^{(i)} \in \hat{\mathcal{D}}$
  if $|\mathcal{S}| < N/2$ & $|\mathcal{S}[j]| < K$ **then**
    $\mathcal{S}[j] \leftarrow \mathcal{S}[j] \cup (\mathcal{D}^{(i)}, \mathcal{R}^{(i)})$; ▷ 加入支持集
  end if
  if $|\mathcal{Q}| < N/2$ & $|\mathcal{Q}[j]| < K'$ and $(\mathcal{D}^{(i)}, \mathcal{R}^{(i)}) \notin \mathcal{S}[j]$ **then**
    $\mathcal{Q}[j] \leftarrow \mathcal{Q}[j] \cup (\mathcal{D}^{(i)}, \mathcal{R}^{(i)})$; ▷ 加入查询集
  end if
until $|\mathcal{S}| = |\mathcal{Q}| = N/2$ **and** $\{\forall j \mid |\mathcal{S}[j]| = K$ **and** $|\mathcal{Q}[j]| = K'\}$
return $\mathcal{S}$, $\mathcal{Q}$;

个副本只包含特定关系类型中的实体，其他边界框简单地被重新标记为 Other 类型。数学公式为: $\mathcal{D}_{\mathcal{R}} = \mathcal{M}(\mathcal{D}, \bar{\mathcal{R}})$, 其中 $\mathcal{D}_{\mathcal{R}}$ 代表与 $\mathcal{R}$ 无关的实体被遮掩的副本。遮掩操作 $\mathcal{M}$ 选择性地保留仅与 $\mathcal{R}$ 相关的关系类型，并抑制或消除其余部分，记为 $\bar{\mathcal{R}}$。对于每次迭代，随机选择 N 个关系类别，并从包含每个关系类别的多个副本的扩展数据集中采样 K 个示例来构建支持集 $\mathcal{S}_{train} = \left\{\mathcal{D}^{(i)}, \mathcal{R}^{(i)}\right\}_{i=1}^{N\times K}$ 以及 K' 个实例用于查询集 $\mathcal{Q}_{train} = \left\{\mathcal{D}^{(j)}, \mathcal{R}^{(j)}\right\}_{j=1}^{N\times K'}$，确保支持集 $\mathcal{S}$ 和查询集 $\mathcal{Q}$ 不重叠 ($\mathcal{S} \cap \mathcal{Q} = \emptyset$)。此外，在每次迭代采样后，向支持集（Support Set）或查询集（Query Set）中添加一个或多个关系实例，直到支持集和查询集达到所需的关系实体类别数量（N-way）和每个类别的实例数（K-shot）。在训练阶段，使用支持集和查询集 ($\mathcal{S}_{train}$, $\mathcal{Q}_{train}$) 来训练少样本学习系统，其中两个集合的监督信号都是可见的。在测试阶段，需要在查询集 $\mathcal{Q}_{test}$ 中预测新类别，并根据真实值评估性能。

4.4.2 数据集细节

本节提供了一个概述，介绍了 Few-CORD 和 Few-SEAB 数据集。我们在 CORD[68] 和 SEAB[61] 数据集的基础上构建了两个少样本基准数据集，用于视觉富文档的少样本关系抽取。这两个数据集包含现实生活中的实例。表格4.2提供

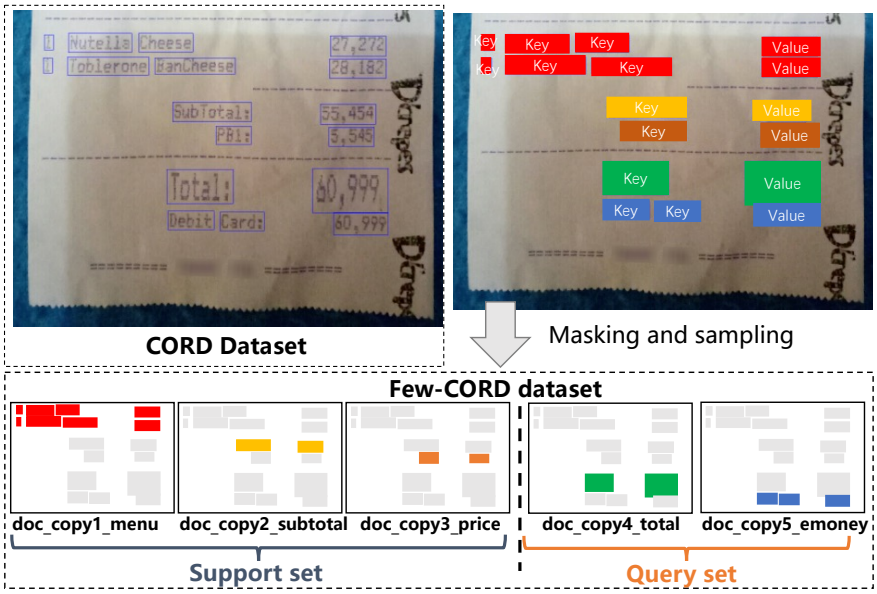

图 4.4　拷贝，遮掩及采样示例

了原先的有监督数据集和重构的少样本数据集的详细信息。表格内容包括文档数量，文本框数量，实体/关系类型的数量（A/B 中，A 代表实体数量，B 代表关系数量，为 2:1 的比例）。Few-CORD 数据集是基于 CORD[68] 数据集构建的，这是一个专门为收据中的关键信息提取而设计的数据集。原始数据集包含 800 份收据，用于训练集，100 份用于验证集，以及 100 份用于测试集。在本文工作中，将验证集和测试集合并形成一个 200 份的测试集（之所以这样做，是为了跟随前人[69] 的工作）。此外，通过将关系实体类别转换为键值对来修改标签结构，这一过程使得 CORD 数据集可以用于少样本关系学习任务。我们将 CORD 中的 32 个实体（16 种键值对）划分到训练集和测试集中，其中训练集包含 18 个已知实体类别（9 种关系类型），测试集包含 14 个新的实体类别（7 种关系类型）。另一个 Few-SEAB 数据集是基于 SEAB[61] 数据集构建的，该数据集包含很多海运单文档，与第三章使用的数据集一致。原始的 SEAB 数据集包含 44 个细分实体类别，总共 22 个键值对关系。我们将 22 种键值对分为 12 个已知的关系用于训练集，10 个新的键值对关系类型用于测试集。SEAB 是手动注释的，在图4.5中展示了所有类型，包含 4 种粗分类型和 22 种细分类型。注意，在跨类别元学习任务中，具体关系类型的划分在表格4.3中展示，关系类型根据它们的粗分类型分配了不同的颜色。

表 4.2 有监督数据集和少样本数据集的统计数据

数据集	训练集			测试集		
	文档数	文本框	类型	文档数	文本框	类型
CORD	800	18,915	32/16	200	4,466	32/16
SEAB	3,562	249,255	44/22	953	73,873	44/22
Few-CORD	1,211	32,160	18/9	702	15928	14/7
Few-SEAB	20,831	575,391	24/12	4,048	146,312	20/10

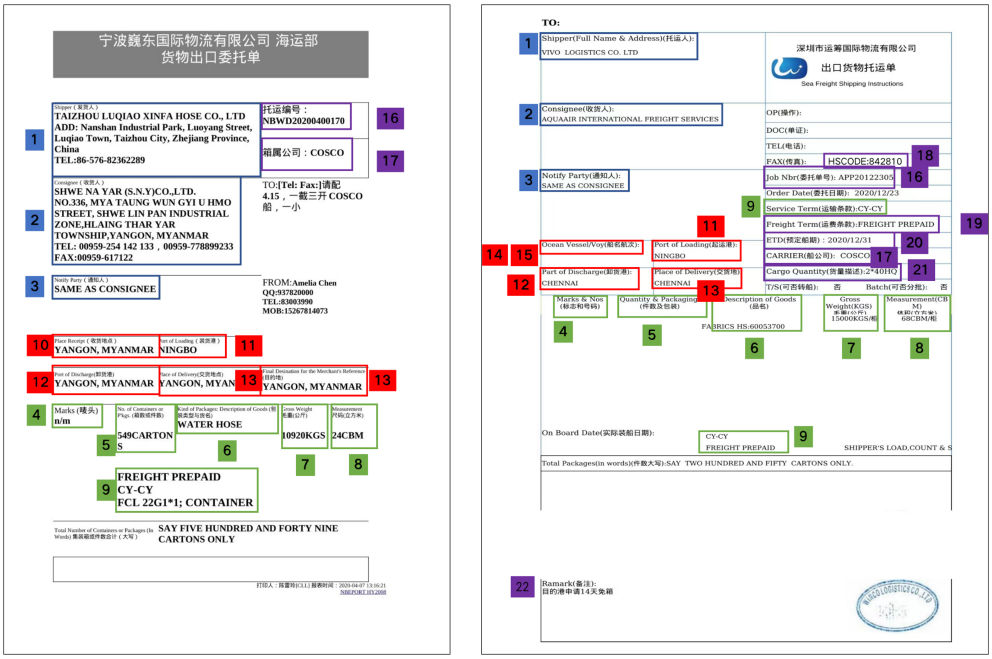

图 4.5 SEAB 数据集中具体文档的注释实例，包含了所有实体和关系类型

4.5 提出方法

人类即使在很少接触到具体实例的情况下，也会拥有迅速掌握新类型关系的能力。因为人类可以有效地利用各种文档图像中的布局信息，学习到关键实体和值实体之间明确的二维空间关系。这种二维空间先验其实是智能模型学习

表 4.3 Few-SEAB 数据集中关系类型的划分，每个数字对应一个关系类型。

数据集	训练集	测试集
SEAB	ALL	ALL
Inter	1,5,7,8,9,10,11,12,13,14,15,18	2,3,4,6,16,17,19,20,21,22
Intra	1,2,3,9,10,11,12,13,14,15,16,19	4,5,6,7,8,17, 18,20,21,22

的宝贵线索。然而，现有方法难以有效利用这些特征。此外，人们有能力在只看到几个实例后，推断出从未见过的新类别。他们可以通过建立不同类别之间的联系，超越语言界限[81]和上下文限制[78]，学习高维的、与类别无关的特征。这种独特的能力使他们能够在多样化的环境中理解并推广知识，以识别新的类别关系和实体。因此，本研究引入新颖的方法，该方法模仿人类在学习关系模式中的行为，专注于整合二维空间位置先验，并旨在学习高维特征，弥补现有模型与人类认知之间的差距，从而实现多模态数据中的少样本关系学习。

模型结构如图4.6所示，包括三个关键组成部分：（1）感兴趣区域回归模块（ROI Regression），使用构建的黄金窗口学习感兴趣区域（ROIs）。（2）原型矫正模块（Prototypical Rectification），通过变分方法矫正有偏的原型表征。（3）基于邻近度的分类模块（Proximity-based Classification），预测最终的关系类型。该模型可以通过将注意力集中到相关区域并通过原型矫正来学习高维度的类别无关特征，帮助模型快速适应新的关系类别，通过更鲁棒的表示，提高跨类别元学习框架的知识迁移能力。

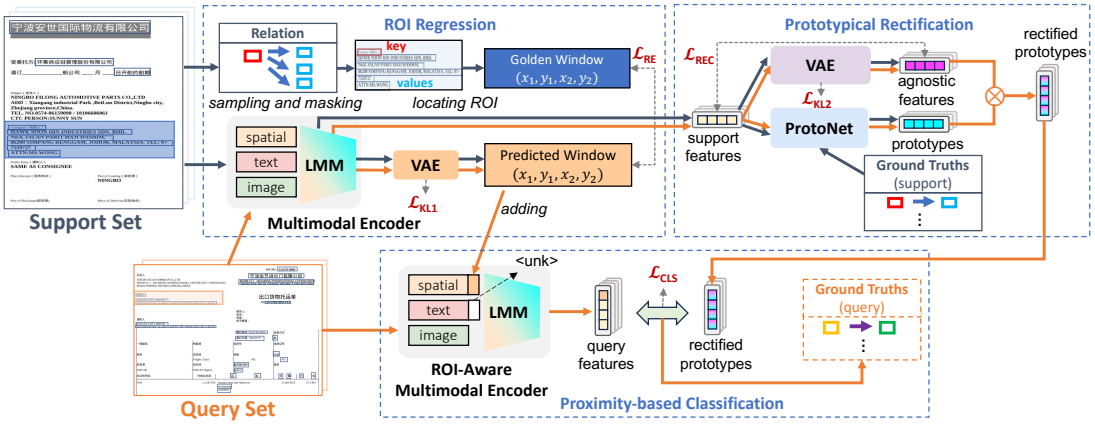

图 4.6　跨类别迁移学习模型结构

4.5.1　感兴趣区域回归（Region of Interest Regression）

通过第三章的初步研究，揭示在视觉富文档的二维布局空间中键实体和值实体之间具有强相关性。通常，键和值在这种空间排列中往往表现出独特的双驼峰分布（见图4.7和图4.8），这可以被视为潜在的二维空间先验。

为了充分利用分布规律性，本工作引入了“黄金窗口”的概念，它是一个大

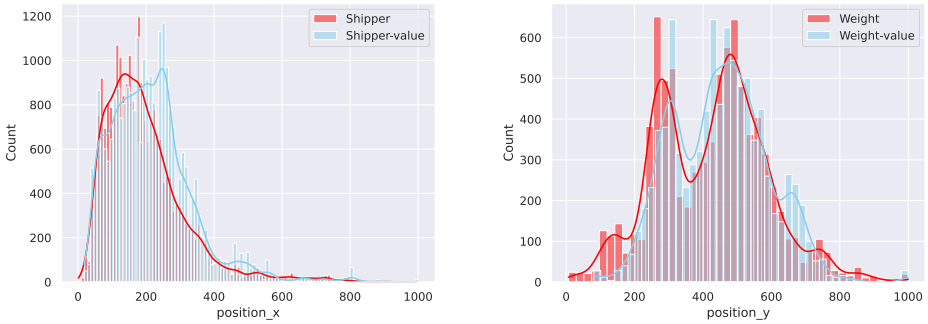

图 4.7 SEAB 数据集中的“Shipper”和“Weight”类型的双驼峰分布

图 4.8 CORD 数据集中的“Menu”和“Total”类型的双驼峰分布

的框，包括属于特定类型的键和值，如图4.6，在支持集中，蓝色方框（黄金窗口）覆盖了当前迭代过程随机采样得到的关系所属区域。在训练过程中，这个黄金窗口充当了一个明确的监督信号，引导模型专注于相关的感兴趣区域（ROI）。首先，使用 LayoutLM[7] 和 LayoutLMv2[8] 作为多模态编码器，从中提取包括文本、空间（一维位置和二维布局）和视觉的 token 级特征 $\mathbf{H}$。当文档的输入序列被分成 n 个 tokens，$[t_1, t_2, \ldots, t_n]$，将多模态编码器的最终隐藏层输出作为中间表示 $\mathbf{h}_i \in \mathbb{R}^l$。

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n] = \text{LMM}\left(\begin{bmatrix} t_1, & t_2, & \ldots & t_n \\ b_1, & b_2, & \ldots & b_n \\ \mathcal{I}_{b_1}, & \mathcal{I}_{b_2}, & \ldots & \mathcal{I}_{b_n} \end{bmatrix} \right) \tag{4.1}$$

然后，使用变分自编码器（VAE）[82] 来对局部 ROI 的分布进行建模。具体地，VAE 中解码器的输出对应于重构的黄金窗口坐标 (x_1, y_1, x_2, y_2)。这个回归任务通过显式关注键值关系在视觉富文档中的几何布局，提高了预测键值关系的准确性。变分自编码器有抽象特征，定义为 $\mathbf{z}$，我们假设 $\mathbf{z}$ 是根据先验分布生

成的隐藏变量，$\mathcal{S}$ 从条件分布 $p(\mathcal{S}|\mathbf{z})$ 中产生。感兴趣区域回归的训练目标是将变分支持集特征 $\mathcal{S}$ 转换成表示一个特定关系类别二维空间分布信息 $\mathcal{N}$，最小化 $p(\mathbf{z}|\mathcal{S})$ 和后验 $q(\mathbf{z}|\mathcal{S})$ 之间的 KL 散度（the Kullback Leibler）：

$$\mathcal{L}_{\text{KL1}} = D_{\text{KL}}\Big(q(\mathbf{z}|\mathcal{S}_{\text{train}})||p(\mathbf{z}|\mathcal{S}_{\text{train}})\Big) \tag{4.2}$$

这是通过最大化证据下限（ELBO）来实现的：

$$ELBO = \mathbb{E}_{q(\mathbf{z}|\mathcal{S}_{\text{train}})}\Big[\log p(\mathcal{S}_{\text{train}}|\mathbf{z})\Big] - D_{\text{KL}}\Big(q(\mathbf{z}|\mathcal{S}_{\text{train}})||p(\mathbf{z})\Big) \tag{4.3}$$

$$\mathcal{L}_{\text{RE}} = \max(0, ||b^*{}_{\text{gold}} - b_{\text{pred}}|| - \xi). \tag{4.4}$$

ξ 表示预定义的松弛变量，用于维持回归任务的约束程度，并允许训练数据出现噪声。所以该模块最终的训练函数为：

$$\mathcal{L}_{\text{ROI}} = \mathcal{L}_{\text{RE}} + \alpha\mathcal{L}_{\text{KL1}} \tag{4.5}$$

其中 α 作为权重系数，值为 2.5×10^{-4}。在得到预测的黄金窗口框 b_{pred} 后，将其添加到原始多模态编码器输出 $\mathbf{H}$ 的 token 序列的末尾，为了多模态对齐，在相应的文本模态中填充 t_{spe} 字符，例如，[UNK]。这样，就可以用一个包含 $n+1$ 个 token 的序列 $[t_1, t_2, \ldots, t_n, t_{spe}]$ 表示 ROI-Aware 多模态编码器的最后隐藏层输出：

$$\tilde{\mathbf{H}} = \text{LMM}\Big(\begin{bmatrix} t_1, & t_2, & \ldots & t_n, & t_{spe} \\ b_1, & b_2, & \ldots & b_n, & b_{\text{pred}} \\ \mathcal{I}_{b_1}, & \mathcal{I}_{b_2}, & \ldots & \mathcal{I}_{b_n}, & \mathcal{I}_{b_{\text{pred}}} \end{bmatrix}\Big) \tag{4.6}$$

4.5.2 原型矫正（Prototypical Rectification）

通过多模态编码器，可以直接得到支持集的特征，再利用原型网络[71] 来计算每个关系类别的 M 维表征 $\mathbf{p}_c \in \mathbb{R}^M$，取 token 嵌入表示的平均值来获得类别 c 的原型 $\mathbf{p}_c$：

$$\mathbf{p}_c = \frac{1}{|S_{\text{train}}|} \sum_{\mathcal{D} \in S_{\text{train}}} \frac{1}{|\mathcal{D}|} \sum_{t \in \mathcal{D}} f_{\mathbf{p}}(\mathbf{h}_t) \tag{4.7}$$

原型网络仅仅依靠支持集中的实例，将其编码为类别相关的特征向量。但是在多模态数据的少样本场景中，由于视觉富文档布局格式及样式的多样性和复杂性，这种方法并不能准确估计出类别中心的表示[77,83-84]。低资源条件下，数据稀疏，支持集实例的方差变化很大。这就造成原型网络仅仅依靠平均表示的方式很难代表类别的分布多样信息。

该模块受到变分特征学习[78,85-86]最新研究进展的启发，提出了变分矫正机制，将类别分布信息融合到原型中。这种机制能够将支持集的特征转换为带有类别分布信息的表示。同样利用变分自编码器来约束从支持实例中学习到的潜在表示，采用类别无关特征的聚合方法促进不同类之间的特征聚合，从而鼓励模型学习类别无关特征减少原型表征偏置。这里变分自编码器重构任务的输出与输入表示具有相同的维度，都是 768。重构损失函数 $\mathcal{L}_{\text{REC}}$ 可以被定义成输入 $\mathbf{H}$ 和重构特征 $\mathbf{H}'$ 之间的 L_2 距离，

$$\mathcal{L}_{\text{REC}} = ||\mathbf{H} - \mathbf{H}'|| \tag{4.8}$$

$$\mathcal{L}_{\text{KL2}} = D_{\text{KL}}\Big(q(\mathbf{z}|\mathcal{S}_{\text{train}})||p(\mathbf{z})\Big) \tag{4.9}$$

$\mathbf{z}$ 是变分自编码器的隐变量，变分特征的学习过程为，

$$\tilde{\mathcal{L}}_{\text{PR}} = \mathcal{L}_{\text{REC}} + \beta \mathcal{L}_{\text{KL2}} \tag{4.10}$$

其中 β 是权重系数，值为 2.5×10^{-4}。

通过矫正后的原型特征 $\tilde{\mathbf{p}}_c$ 能够捕获该类的更多通用特征，使得矫正后的原型特征对支持集实例的方差具有鲁棒性。其中，$\mathbf{p}_c$ 代表类别 k 的原型特征，从模拟的类别分布为 $\mathcal{N}(\mu_{\text{v}}, \Sigma_{\text{v}})$ 并从中采样，获得一个变分特征 $\mathbf{z}_{\text{v}} = f_{\text{v}}(\mathbf{z}) = f_{\text{v}}(\mu_{\text{v}} + \Sigma_{\text{v}})$，然后使用如下公式将原型和变分特征结合起来，

$$\tilde{\mathbf{p}}_c = \mathcal{A}(\mathbf{p}_c, \mathbf{z}_{\text{v}}) = \mathbf{p}_c \otimes sigmoid(\mathbf{z}_{\text{v}}), c \in C \tag{4.11}$$

该方法使用 sigmoid 函数将变分特征映射到 0 到 1 之间的范围，然后与原型特征 $\mathbf{p}_c$ 逐元素相乘，即在原型特征基础上增加了变分校正。

4.5.3 基于邻近度的分类（Proximity-based Classification）

为了利用感兴趣区域信息，该模块通过将预测的黄金窗口 b_{pred} 的坐标附加到序列的末尾来增强多模态输入序列，所以从空间布局模态上序列就具备了感兴趣区域信息。为了保证输入序列在文本、布局、视觉模态之间的对齐，需要在文本模态的末尾加入一个特殊的 token，<UNK>。因此，最终所得到的序列从初始的 n 个 token 拓展到 $n+1$ 个 token，标记成 $[t_1, t_2, \ldots, t_n, <\text{UNK}>]$。然后采用这个序列得到最后一个隐藏层输出 $\tilde{\mathbf{H}}$，也就是说该编码器是一个可以感知感兴趣区域的多模态编码器。隐藏层输出融合了感兴趣区域信息同时也捕获了足够的多模态信息。对于每个查询集 $\mathcal{Q}_{\text{test}}$ 中的实例，通过计算查询集的表征 $\tilde{\mathbf{h}}_i$ 和矫正之后的原型 $\tilde{\mathbf{p}}_c$ 的欧几里得距离，来实现基于邻近度的分类，

$$\mathbf{d}_c = d_{l2-norm}(\tilde{\mathbf{p}}_c, \tilde{\mathbf{h}}_i). \tag{4.12}$$

然后，使用 softmax 函数对所有类别的预测概率进行归一化。该模型将标签预测为最接近输入标记的原型，

$$p(\mathbf{r} = c) = \frac{exp(-\mathbf{d}_c)}{\sum_{c' \in C} \exp(-\mathbf{d}_{c'})} \tag{4.13}$$

最后使用交叉熵损失作为多分类任务的损失函数，

$$\mathcal{L}_{\text{CLS}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} c_j \log p_{ij} \tag{4.14}$$

其中 c_j 是输入序列中第 t_i 个 token 的标注。

4.5.4 训练函数

算法5描述了一个完整的模型训练过程。在训练过程中，每个 episode 都会更新参数。在测试过程中，模型将 token 预测为最接近的原型类别，预测新型实体时，无需计算损失或更新模型参数。所以，少样本关系学习框架的最终损失

Algorithm 5 训练过程

Input: 支持集 $\mathcal{S}_{\text{train}}$, 查询集 $\mathcal{Q}_{\text{train}}$;
Output: A trained few-shot extractor F_{RL} for key and value entities;

for episode $e = 0$ **to** num_train_episodes **do**
  $\mathcal{S}_{\text{train}}^{(e)}, \mathcal{Q}_{\text{train}}^{(e)} \leftarrow$ Algorithm4; ▷ 执行采样策略
  if Support Set $\mathcal{S}_{\text{train}}^{(e)}$ **then**
    借助多模态编码器得到支持集表示 **H**;
    给定文本框的 label 生成黄金窗口 b^*_{gold} ;
    H 执行感兴趣区域回归, 生成预测窗口 b_{pred};
    计算 $\mathcal{L}_{\text{KL1}}$ **and** $\mathcal{L}_{\text{RE}}$; ▷compute $\tilde{\mathcal{L}}_{\text{ROI}}$
    将 **H** 输入到原型网络中得到原型 **p**;
    将 **H** 输入到变分模块中, 得到 $\mathbf{H}'$ and $\tilde{\boldsymbol{z}_v}$;
    计算 $\mathcal{L}_{\text{KL2}}$ **and** $\mathcal{L}_{\text{REC}}$; ▷compute $\tilde{\mathcal{L}}_{\text{PR}}$
  end if
  if 查询集 $\mathcal{Q}_{\text{train}}^{(e)}$ **then**
    借助多模态编码器得到查询集表示 **H**;
    H 执行感兴趣区域回归, 生成预测窗口 b_{pred};
    将预测出的窗口 b_{pred} 加到 $\tilde{\mathbf{H}}$ 中得到融合先验信息的感兴趣区域特征; ▷ 融合先验信息
    利用 $\tilde{\boldsymbol{z}_v}$ 将原型 **p** 矫正成 $\tilde{\mathbf{p}}$;
    计算 $\tilde{\mathbf{H}}$ 和 $\tilde{\mathbf{p}}$ 之间的距离，输出预测结果;
    计算 $\mathcal{L}_{\text{CLS}}$ 和最终的损失函数; ▷compute $\mathcal{L}_{\text{final}}$
    反向传播，利用 AdamW 更新模型参数;
  end if
end for
return F_{RL};

函数为，

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{REC}} + \alpha\mathcal{L}_{\text{KL1}} + \mathcal{L}_{\text{RE}} + \beta\mathcal{L}_{\text{KL2}} + \mathcal{L}_{\text{CLS}}, \tag{4.15}$$

其中权重系数为 $\alpha = \beta = 2.5 \times 10^{-4}$。

4.6 实验

4.6.1 实验环境与细节

（1）主干网络（backbone）：

本实验跟随了前人在纯文本模态的少样本数据集[87] 上所采取的方法。对于主干架构(backbone), 使用预训练的文档智能模型, 如 LayoutLM 和 LayoutLMv2。这些模型，因为其可以提供稳健有效的特征，是本研究方法的基础。此外，主干架构还使用 Bert 模型，舍弃多模态的信息来评估提出方法的性能。

（2）硬件与超参数：

本实验在配备有两个 NVIDIA 3090 GPU 的机器上使用 AdamW 优化器优化模型，学习率为 1×10^{-5}。在训练中，通过 10000 次迭代对模型进行微调，并通过 500 次的测试迭代的平均得分来评估性能。

（3）评估方法：

根据前人在纯文本模态下的少样本数据集[87] 上所采取的评估方法，本实验通过多次计算测试集上的 micro-F1 分数，采用迭代累加评估。注意，每一次迭代都包含一个带有 K 个标注样本的支持集和一个没有任何标注的 $K\prime$ 个样本的查询集。

（4）基线模型

本实验比较了少样本学习中经典的基线模型：ProtoNet[71,76] 是一个基线原型系统，通过使用原型网络学习训练集中的示例，将每个 token 表示分配给最近的原型表示。NNShot 和 StructShot[88] 是最经典的基于 token 级最近邻分类的方法，与 ProtoNet 不同，NNShot 根据与当前预测的 token 距离最近的 token 标注来确定查询集实例的关系分类。StructShot 在推理阶段引入了额外的 Viterbi 解码，以提高整体性能。VFA[78] 被用于计算机视觉领域中的少样本目标检测，本研究复现了该方法并在多模态数据下做了相关实验。

4.6.2 实验结果分析

（1）主要结果

表4.4和表4.5展示了在构建的少样本数据集 Few-CORD 和 Few-SEAB 上的 f1 值得分。加粗和下划线表示每组中性能最好的结果和性能第二好的结果。**ROI-Aware** 和 **ProtoRec** 是本章提出的两个主要方法。**+CF** 代表在黄金窗口内部用浅颜色填充，在颜色填充实验中会有说明。**+SW** 指的是通过窗口缩放来进行对抗学习的方法。

从表格纵向角度看，以 BERT 为主干网络的模型比以 LayoutLM 和 LayoutLMv2 为主干网络的模型性能差很多。这主要是由于 BERT 在捕获非文本信息方面弱于多模态预训练语言模型。相比之下，包括视觉信息的 LayoutLM 和 LayoutLMv2 模型被证明更有效。由于 LayoutLM 本身在预训练阶段并没有对视觉特征进行建模，实验使用 ResNet 作为视觉编码器。实验结果显示，LayoutLM

表 4.4　在 Few-CORD 数据集上的性能（平均 F1 值）

Method	LLM	Proto	VAE	Few-CORD				
				1-SHOT	2-SHOT	3-SHOT	4-SHOT	5-SHOT
ProtoNet	BERT	✓	✗	32.32	35.14	38.87	40.08	42.95
NNShot	BERT	✓	✗	29.19	32.34	35.16	36.28	38.04
StructShot	BERT	✓	✗	**33.54**	34.95	37.41	38.31	40.38
VFA	BERT	✗	✓	30.08	31.34	32.86	34.89	37.14
ProtoRec	BERT	✓	✓	33.30	**35.40**	**39.01**	**40.53**	**43.14**
ProtoNet	LayoutLM	✓	✗	70.25	74.10	77.02	79.31	80.40
NNShot	LayoutLM	✓	✗	68.20	72.70	73.76	75.24	76.67
StructShot	LayoutLM	✓	✗	71.38	73.88	74.52	77.24	77.83
VFA	LayoutLM	✗	✓	68.39	69.79	71.83	73.18	74.98
VFA+ROI	LayoutLM	✗	✓	68.71	70.07	72.31	73.91	75.65
ROI-Aware	LayoutLM	✓	✓	71.33	74.96	77.84	80.78	81.22
ProtoRec+ROI	LayoutLM	✓	✓	**73.21**	**76.19**	**78.42**	**81.35**	**81.54**
ProtoNet	LayoutLMv2	✓	✗	70.30	74.22	77.16	79.35	80.52
NNShot	LayoutLMv2	✓	✗	68.17	72.80	73.88	75.39	76.84
StructShot	LayoutLMv2	✓	✗	71.45	73.95	74.62	77.50	78.11
VFA	LayoutLMv2	✗	✓	68.87	70.16	72.05	73.63	75.56
VFA+ROI	LayoutLMv2	✗	✓	69.21	71.39	72.63	74.15	76.01
ROI-Aware	LayoutLMv2	✓	✓	71.59	75.83	77.87	80.92	81.40
ProtoRec+ROI	LayoutLMv2	✓	✓	**73.37**	**76.90**	**78.54**	**81.46**	**81.88**
ProtoRec+ROI+CF	LayoutLMv2	✓	✓	73.32	76.96	78.63	81.40	81.85
ProtoRec+ROI+SW	LayoutLMv2	✓	✓	**73.56**	**77.39**	**78.80**	**81.88**	**82.18**

不及 LayoutLMv2 的表现，主要由于在微调阶段加入视觉模态会造成其与其他模态之间的不对齐，进而影响 LayoutLM 的性能。

从表格横向角度看，随着给定样本数（表示为 K）的增加，所有方法的性能都会提升。这样的提升可以归因于一个事实：随着 K 的增加，原型表征变得更接近真实的数据值，从而获得更好的性能。整体来看，无论使用何种主干模型，原型矫正（**ProtoRec**）和感兴趣区域回归方法（**ROI-aware**）都能产生稳定的提升。而且，原型矫正和感兴趣区域回归方法的组合取得了最好的性能。注意，由于 BERT 缺乏二维位置信息，在 BERT 的主干模型中使用空间感知方法（使用 Layout-aware Encoder）是不可行的，所以实验只展示了 BERT 加上原型矫正方法的结果。

（2）语义相似度

表 4.5　在 Few-SEAB 数据集上的性能（平均 F1 值）

Method	LLM	Proto	VAE	Few-SEAB				
				1-SHOT	2-SHOT	3-SHOT	4-SHOT	5-SHOT
ProtoNet	BERT	✓	✗	26.88	28.59	30.08	31.74	34.11
NNShot	BERT	✓	✗	25.16	26.10	27.92	28.80	29.95
StructShot	BERT	✓	✗	27.30	28.63	29.20	30.19	31.75
VFA	BERT	✗	✓	24.83	25.61	27.11	28.65	29.06
ProtoRec	BERT	✓	✓	**27.45**	**29.04**	**30.35**	**32.15**	**34.28**
ProtoNet	LayoutLM	✓	✗	60.95	64.02	66.31	69.53	73.12
NNShot	LayoutLM	✓	✗	58.80	61.75	62.27	64.40	66.89
StructShot	LayoutLM	✓	✗	61.15	63.14	63.50	65.28	68.10
VFA	LayoutLM	✗	✓	58.14	60.23	63.14	64.36	66.12
VFA+ROI	LayoutLM	✗	✓	58.76	60.61	63.91	65.08	66.79
ROI-Aware	LayoutLM	✓	✓	61.79	64.92	66.45	69.84	73.19
ProtoRec+ROI	LayoutLM	✓	✓	**62.77**	**65.54**	**66.59**	**69.95**	**73.28**
ProtoNet	LayoutLMv2	✓	✗	61.18	64.10	66.43	69.80	73.37
NNShot	LayoutLMv2	✓	✗	59.09	61.81	62.35	64.49	67.14
StructShot	LayoutLMv2	✓	✗	61.30	63.34	63.57	65.37	68.18
VFA	LayoutLMv2	✗	✓	58.77	60.59	63.70	65.03	67.18
VFA+ROI	LayoutLMv2	✗	✓	59.20	61.57	63.97	65.28	68.26
ROI-Aware	LayoutLMv2	✓	✓	61.94	65.02	66.48	69.86	73.52
ProtoRec+ROI	LayoutLMv2	✓	✓	**63.05**	**65.59**	**66.83**	**70.18**	**73.70**
ProtoRec+ROI+CF	LayoutLMv2	✓	✓	62.98	65.47	66.80	70.21	73.65
ProtoRec+ROI+SW	LayoutLMv2	✓	✓	**63.43**	**65.87**	**67.30**	**70.56**	**73.96**

原型共享[89] 可以帮助识别类之间的原型相似性。本研究旨在探索实体之间关系的语义相似性，对实体类相似性进行实证研究，从而更好地解释跨类别迁移学习的有效性。为此，实验按照第三章有监督学习的方式，专门使用 BERT 和 LayoutLMv2，在有监督训练集上进行训练，获得所有实体的嵌入表示。具体地，实验为每种细粒度类型随机选择了 100 个实体嵌入表示实例，并对它们进行平均，然后计算每个实体类型的中心表示之间的点积来测量它们的相似度。图4.9展示了 BERT 和 LayoutLMv2 生成的实体表示的语义相似度热力图。通过对 BERT 获得的结果与 LayoutLMv2 获得的结果进行比较，说明对多模态数据进行知识迁移的挑战性。而在每个模型内部，实验发现共享同一粗粒度类型的实体类型往往具有更大的相似性，更易于知识转移。

（3）迁移学习能力

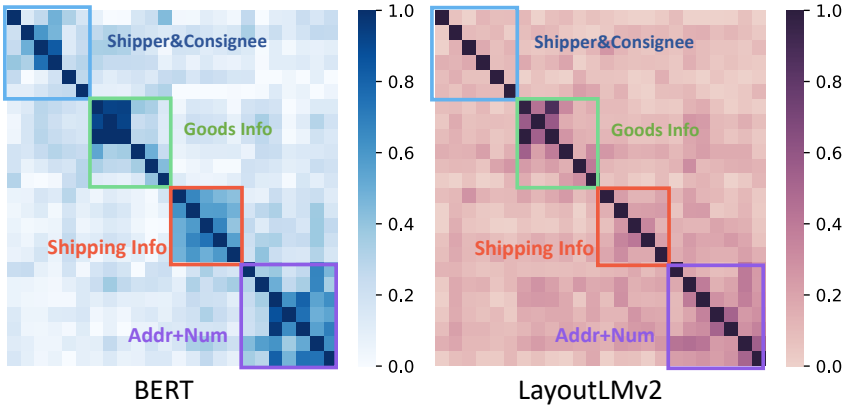

图 4.9 由 BERT 和 LayoutLMv2 生成的实体表示的语义相似度热力图

评估模型的迁移学习能力对于少样本学习至关重要。为了进行迁移学习的实验，本实验遵循 Few-NERD[87] 的工作，对有监督数据集 SEAB 进行了关系类型划分，并采用不同的划分方式构建了两种模式，即 Few-SEAB（Inter）和 Few-SEAB（Intra）。我们将整个实体集划分为 4 个粗粒度的不相交子集，例如，“Shipper and Consignee”，“Goods Information”，“Shipping Information”，和 “Address+Numbers”。换句话说，实验根据粗粒度类型把实体类型粗划分，构建了 Few-SEAB（Intra）数据集 $\langle \hat{\mathcal{D}}_{\text{train}}, \hat{\mathcal{D}}_{\text{test}} \rangle$，划分原则是不同集合中的实体属于不同的粗粒度类型。例如，“Shipper and Consignee”，“Goods Information” 仅出现在 $\hat{\mathcal{D}}_{\text{train}}$ 中，“Shipping Information” 和 “Address+Numbers” 只出现在 $\hat{\mathcal{D}}_{\text{test}}$ 中。Intra 这种模式确保训练集和测试集在知识相似度方面最小。相比之下，虽然 Few-SEAB（Inter）数据集中的细粒度实体类型在 $\hat{\mathcal{D}}_{\text{train}}$ 和 $\hat{\mathcal{D}}_{\text{test}}$ 中是相互不相交的，但是粗粒度类型是共享的，这意味着 $\hat{\mathcal{D}}_{\text{train}}$ 和 $\hat{\mathcal{D}}_{\text{test}}$ 两者应该包含所有细粒度类型。如表4.3所示，1-22 依次代表 Shipper, Consignee, Notify Party, Marker, Number of Packages, Good Description, Gross Weight, Measurement, Shipping Terms, Place of Receipt, Port of Loading, Port of Discharger, Place of Delivery, Vessel Name, Voyage no, Consignment Code, Shipping Company, HSCODE, Freight Terms, Pre-Assignment, Case Size, Remarks。表4.6和表4.7给出了 SEAB 数据集在 Intra 和 Inter 两种迁移学习模式下的结果比较，Inter 模式相比 Intra 模式的性能要好。因为 Inter 模式要求训练集和测试集中的实体类型按照粗划分的角度相交, 共享粗划分的实体类型。而 Intra 模式要求训练集和测

表 4.6　在 Few-SEAB 数据集上的 Inter 模式下的比较

Method	**LLM**	**Proto**	**VAE**	**Inter**				
				1-shot	2-shot	3-shot	4-shot	5-shot
ProtoNet	BERT	✓	✗	33.03	36.73	37.08	38.36	39.54
NNShot	BERT	✓	✗	34.35	34.68	35.75	35.93	36.47
StructShot	BERT	✓	✗	**36.45**	36.63	36.83	37.10	37.29
VFA	BERT	✗	✓	31.65	33.87	35.48	36.11	36.69
ProtoRec	BERT	✓	✓	34.80	**37.10**	**37.45**	**38.71**	**39.86**
ProtoNet	LayoutLM	✓	✗	66.75	69.31	70.86	74.88	77.84
NNShot	LayoutLM	✓	✗	63.83	65.50	66.32	68.43	70.88
StructShot	LayoutLM	✓	✗	67.69	69.35	69.64	71.64	72.35
VFA	LayoutLM	✗	✓	64.45	65.89	67.34	70.63	72.90
VFA+ROI	LayoutLM	✗	✓	65.32	66.17	67.81	71.23	73.12
ROI-Aware	LayoutLM	✓	✓	68.80	70.39	71.48	75.60	78.17
ProtoRec+ROI	LayoutLM	✓	✓	**69.67**	**71.18**	**71.88**	**75.83**	**78.64**
ProtoNet	LayoutLMv2	✓	✗	67.21	69.55	71.23	75.14	78.22
NNShot	LayoutLMv2	✓	✗	63.87	66.20	66.87	68.87	71.26
StructShot	LayoutLMv2	✓	✗	68.16	69.73	70.75	71.63	74.27
VFA	LayoutLMv2	✗	✓	64.89	66.46	68.05	70.98	73.59
VFA+ROI	LayoutLMv2	✗	✓	65.43	67.04	68.78	71.61	74.04
ROI-Aware	LayoutLMv2	✓	✓	70.11	71.19	72.64	75.80	78.67
ProtoRec+ROI	LayoutLMv2	✓	✓	**71.59**	**72.76**	**73.32**	**75.97**	**79.59**

试集中的实体类型按照粗划分的角度不相交，不共享粗划分的实体类型。

（4）原型表示的鲁棒性实验

在测试阶段，模型需要参考 K-shot 实例的平均特征得到的类原型。如图4.10所示，使用原型矫正和区域回归的方法估计的类原型比基线模型更加稳健和准确。实验通过在有监督任务获得所有实体类型的真实类中心。我们分别使用三个主干模型来计算不同方法的原型与真实中心的相对距离。提出的原型矫正方法和区域回归方法在 K 很少时与真实类中心的距离更近，原因是提出的变分方法可以充分利用已知类的分布来估计新类的分布。从分布中采样的特征可以矫正原型表示，这对支持集实例的方差是稳健的。基线模型对支持集实例的方差很敏感。注意，实验将 5 个样本（K=5）估计的原型与类的实际质心之间的距离定义为 1。

（5）颜色填充实验

表 4.7　在 Few-SEAB 数据集上的 Intra 模式下的比较

Method	LLM	Proto	VAE	Intra				
				1-SHOT	2-SHOT	3-SHOT	4-SHOT	5-SHOT
ProtoNet	BERT	✓	✗	24.39	26.68	27.86	28.35	29.87
NNShot	BERT	✓	✗	23.46	24.43	25.10	25.57	26.20
StructShot	BERT	✓	✗	**25.28**	25.87	26.19	26.86	27.22
VFA	BERT	✗	✓	21.90	23.31	24.58	25.42	25.94
ProtoRec	BERT	✓	✓	25.06	**27.18**	**28.51**	**28.92**	**30.09**
ProtoNet	LayoutLM	✓	✗	55.67	58.23	60.92	64.20	65.89
NNShot	LayoutLM	✓	✗	53.74	56.37	58.83	62.13	63.48
StructShot	LayoutLM	✓	✗	55.53	57.90	58.96	62.41	64.81
VFA	LayoutLM	✗	✓	51.48	54.67	55.86	59.34	61.22
VFA+ROI	LayoutLM	✗	✓	51.96	55.04	56.12	59.76	61.46
ROI-Aware	LayoutLM	✓	✓	56.80	58.31	61.20	64.38	66.95
ProtoRec+ROI	LayoutLM	✓	✓	**57.94**	**59.13**	**61.79**	**64.85**	**67.05**
ProtoNet	LayoutLMv2	✓	✗	56.16	58.52	61.65	64.96	66.31
NNShot	LayoutLMv2	✓	✗	54.70	57.23	58.93	62.27	63.02
StructShot	LayoutLMv2	✓	✗	56.82	58.24	59.21	62.54	63.27
VFA	LayoutLMv2	✗	✓	52.42	55.61	56.74	60.38	61.69
VFA+ROI	LayoutLMv2	✗	✓	52.95	56.38	57.61	60.84	61.93
ROI-Aware	LayoutLMv2	✓	✓	57.69	59.80	61.91	64.87	67.14
ProtoRec+ROI	LayoutLMv2	✓	✓	**58.79**	**60.44**	**62.37**	**65.13**	**67.53**

为了充分利用提出模型中的二维空间先验感知能力，本实验通过用不同颜色填充（**C**olor **F**illing，简写成 CF）感兴趣区域来整合有关 ROI 的信息。经过几次尝试，实验发现使用绿色、黄色、蓝色和红色这些浅颜色产生了一致的性能，如 **ProtoRec+ROI+CF** 在表4.4和表4.5中所示。实验通过用不同颜色在视觉上突出显示 ROI 的方法增强了模型捕获和区分文档中重要特征的能力。这一有趣的发现为未来研究指明了方向，强调了 ROI 在文档上下文中强烈视觉对比的重要性。

（6）窗口缩放的对抗学习实验

利用感兴趣区域回归模块预测黄金窗口的设计使模型能够专注于文档中可能出现键值实体的位置。然而，使用黄金窗口作为监督信号可能不是窗口回归的最优训练目标。实验引入了松弛变量 ξ 来缩放窗口（**S**hrink **W**indow，简写成 SW）。表4.4和表4.5中的 **ProtoRec+ROI+SW** 行已经展示了通过缩放窗口进行对

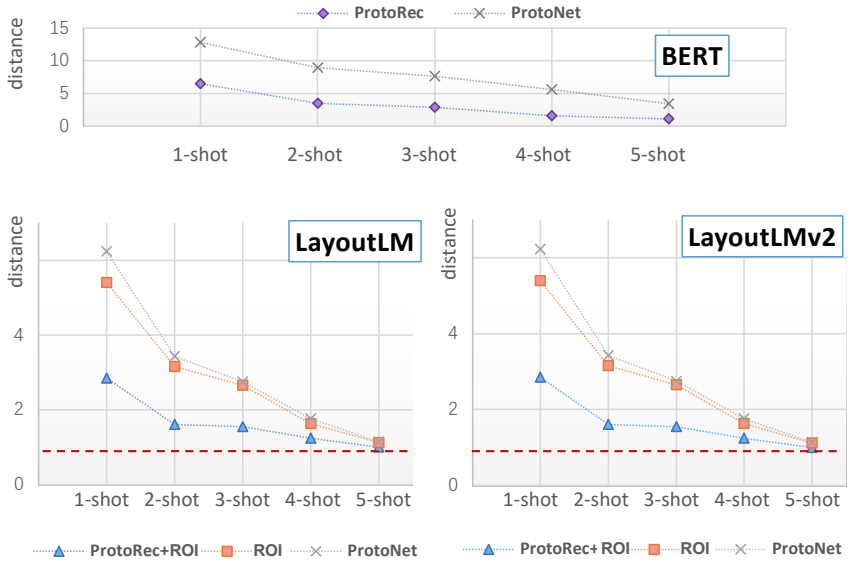

图 4.10　由 K 个样本实例估计出的原型到类的真实点的平均距离

抗性学习的有效性。

表 4.8　窗口缩放的影响

ξ	FEW-CORD				FEW-SEAB			
	1-SHOT	3-SHOT	5-SHOT	AVG	1-SHOT	3-SHOT	5-SHOT	AVG
0.25	**+0.33**	+0.17	+0.24	+0.25	+0.13	+0.08	+0.21	+0.14
0.50	+0.24	**+0.26**	**+0.30**	**+0.27**	+0.33	+0.29	**+0.35**	+0.32
0.75	-0.03	-0.10	+0.11	-0.01	**+0.38**	**+0.47**	+0.26	**+0.37**
1.00	-0.23	-0.17	-0.27	-0.22	+0.12	+0.07	+0.13	+0.11
1.25	-0.36	-0.19	-0.34	-0.30	-0.21	-0.18	+0.05	-0.11

如表4.8所示，实验通过调整松弛变量 ξ 的值，比较黄金窗口的不同缩放程度对结果的影响，这种调整使模型能够更好地适应和泛化到不同的文档布局和结构。实验观察到 0.50 的缩放程度，可以提高整体性能。此外，还观察到较小的松弛变量在 FEW-CORD 文档中的性能优于 FEW-SEAB 文档。这是符合预期的，因为 FEW-CORD 文档具有更紧密的布局，而 FEW-SEAB 文档布局则相对松散。

4.7 本章小结

元学习是和知识迁移方法高度相关的，本章讨论了基于跨类别元学习的视觉富文档关系抽取的研究主题。鉴于该领域数据集非常有限，本章重新组织了现有的有监督基准数据集，并设计了一种专门为少样本学习设定量身定制的采样算法。受到人类认知的启发，该工作还提出了新颖的变分方法把二维空间先验和类别无关特征结合起来，以提高模型在少样本关系学习任务中的性能。具体利用二维位置空间先验对感兴趣区域（ROI）窗口进行建模，感兴趣区域（ROI）窗口将模型的注意力引导到文档图像内给定关系的相关区域。此外，因为低资源的条件，还引入了一种原型矫正机制，以增强学习系统的泛化性和适应新的关系实例的能力。通过在重构的数据集上进行广泛实验，已经证明了加入二维位置空间先验方法和原型矫正方法在视觉富文档中少样本关系抽取任务中的有效性。本章工作极大地促进了在视觉富文档中进行少样本关系抽取的研究进展，为进一步探索这一研究领域铺平了道路。

第五章　基于跨模型知识嵌入的视觉富文档表格问答研究

5.1　研究动机

表格类型的数据在视觉富文档中很常见，它可以通过网格化的布局直观地展示结构化信息，还通过使用不同的单元格样式，如加粗、斜体或背景色，突出显示重要的数据点或趋势。而人类需要大量的时间和精力来分析和操作表格，特别是在非专业人士处理专业领域的表格时会造成非常多的困扰。大规模语言模型（Large-scale Language Model）的进步使得用自然语言输入与表格进行交互成为可能。随着法律、金融、医疗大模型的落地，各个行业对垂直大模型的需求不断增长。本章对飞行器垂直领域大模型进行简单探讨，使用 LangChain 搭建本地化知识库。由于项目原因，本章讨论的某飞行器参数文档（主要为 PDF，篇幅较长，几十数百页均有，简单命名为 FPD）为涉密文件，其包含的敏感信息会进行遮挡。这类飞行器参数文档大部分内容由字段、图、表组成，本研究拟开发一个针对该飞行器垂直领域的知识问答系统，目的是通过对话交互，快速从分布在很多文档内不同位置的信息中提取汇总出正确答案，包括术语速查、公式问答、表格问答等功能。由于研究对象的敏感性，针对文档的问答任务如图5.1所示（已遮挡）。

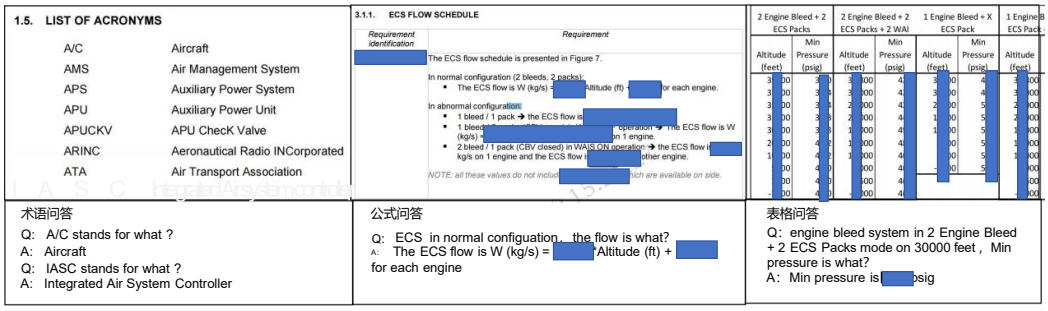

图 5.1　面向飞行器参数文档的问答任务示例

在表格问答任务中，针对异形、特定领域的表格，经常出现无线、少线的情况，单元格也经常出现空白、跨多行/列、多行文本等情况。这些场景给仅仅

使用大模型进行视觉富文档表格问答带来极大挑战。再者，LangChain 对本地文档解析的过程是一行一行解读的，虽然 LangChain 技术可以很好地完成术语速查、公式问答等任务，但是表格类型的数据却不符合这一解析方式，表格回答往往答非所问。考虑到大模型在代码理解上的能力，再结合上述内容，本研究将飞行器参数文档中的表格作为输入，使用预训练的表格识别小模型生成 XML 格式的代码，以 XML 格式的代码作为中间媒介进行知识嵌入，使得大模型理解表格内容。

5.2 相关工作

起初，研究人员尝试通过复杂的 Excel 公式或手动编程来处理表格数据[90]。这促使人们寻求一种更高效的手段来理解表格中的数据。在自然语言处理领域内，生成式预训练模型（GPTs）[91-92] 和大型语言模型（LLMs）[93-94] 已经彻底革新了语言数据挖掘的方式。延续这一研究方向，研究者们同样探索了适用于视觉[95] 和语音[96] 等不同模态的大模型。技术上，这些模型生成文本的能力为处理表格数据提供了新的方法和视角。尽管如此，要将标准的大型语言模型（例如 ChatGPT①、ChatGLM②、星火③等）直接应用于表格数据的处理仍然面临挑战，主要原因有两个：（1）表格全局信息的理解：生成式预训练模型受限于它们处理 token 的长度，限制了它们一次性读取大型表格的能力。因此，生成式预训练模型在捕捉和解析整个表格时面临挑战。（2）将模型扩展到表格数据领域：它们的训练过程是为自然语言量身定制的，因此，在处理表格数据时，它们的泛化能力较差。

近期，一些研究工作致力于将自然语言处理技术融入到表格数据的分析过程中[97-100]。NL2SQL[97-99]，即从自然语言到 SQL 语句的转换，该工作致力于将自然语言表述转换为能够操作关系型数据库的 SQL 查询语句。此外，SheetCopilot[100] 尝试将自然语言转换为 VBA（Visual Basic for Applications，微软 Excel 中使用的内置脚本语言），以此来利用电子表格软件提供的丰富功能集。但是这些编程

① https://openai.com/blog/chatgpt
② https://chatglm.cn/
③ https://xinghuo.xfyun.cn

语言的非结构化特性引入了额外的复杂度，这使得后续实现自动化的处理变得极其困难。根据这些工作，本章研究关注表格识别预训练小模型，将表格转换成 XML 格式的编程语言结构。鉴于大型语言模型显示出理解编程语言的能力，本研究工作致力于将小模型生成的 XML 代码以中间知识的方式，迁移到大模型中，实现跨模型的迁移。

而表格识别方法可以大致分为三类：基于分割和合并的方法、基于检测和分类的方法，以及基于图像到文本生成的方法。以下是对这些方法的简要介绍：

（1）基于分割和合并的方法：

这类方法通常包括两个阶段。第一阶段是检测行和列，然后通过行和列的交叉点，将表格分割成多个文本块。第二阶段是将文本块合并以恢复结构。一些工作专注于更好地分割行和列。例如，DeepDeSRT[101] 和 TableNet[102] 调整了全卷积神经网络来分割行和列。还有一些合并的方法，用来识别包含跨越行或列的单元格。SPLURGE[103] 提出表的拆分和合并的思想，该方法设计了一个合并模型来合并跨多个列或行的单元格。TRUST[104] 引入了一个基于端到端的 Transformer，包含基于查询的分割模块和基于顶点的合并模块。分割模块用于提取行和列分隔符的特征，并将行和列特征进一步输入到基于顶点的合并模块中，以预测相邻单元格之间的关系。

（2）基于检测和分类的方法：

这种方法的基本思路是首先检测单元格，然后对单元格之间的行和列关系进行分类。根据单元格之间的连接构建一张图，以获得表格结构。最近的研究方法将单元格检测和单元格关系分类两个任务整合到一个网络中。TableStruct-Net[105] 和 FLAG-NET[106] 都利用了 Mask R-CNN[26] 网络来获取单元格区域和单元格的视觉特征。他们都利用了 DGCNN 架构[107] 来模拟检测到几何上相邻的单元格之间的交互。

（3）基于图像到文本生成的方法：

这类方法将表格结构（如 XML 或 LaTex 等）视为一个序列，并采用端到端的图像到文本范式来识别表格结构。Deng 等人使用经典的 IM2MAKEUP 框架[108] 来识别表格的逻辑结构，其中 CNN 被用来提取视觉特征，而带有注意力

机制的 LSTM 用于生成表格的 LaTex 代码。Zhong 等人[109] 尝试使用 Encoder-Dual-Decoder（EDD）架构来生成逻辑结构和单元格内容。在解码阶段，他们使用了两个基于注意力的循环神经网络，一个负责解码表格结构代码，另一个负责解码表格内容。TableMaster[110] 和 TableFormer[111] 利用 Transformer 解码器来改进 EDD 的解码器。此外，他们使用回归解码器来预测边界框。由于缺乏局部视觉信息，这些方法预测的边界框准确性较低。基于此，本研究将边界框预测视为一个坐标序列生成任务，并配合视觉对齐损失函数以产生更准确的边界框。

本研究还需结合大模型的能力对垂直领域文档进行问答，对于飞行器参数文档（FPD），还需要将其转化为本地知识库。LangChain 是一个用于开发由语言模型驱动的应用程序框架。主要功能有：调用语言模型、将不同数据源接入到语言模型的交互中、允许语言模型与运行环境交互。LangChain 的可组合性允许用户通过 LangChain 表达式语言轻松创建任意链，提供了数据编排框架的所有优势。LangChain 已经被应用于多个领域，包括智能客服、内容创作、问答系统等，展现其在人机交互和应用开发中的潜力。LangChain 中提供的模块有 Modules：支持的模型类型和集成；Prompt：提示词管理、优化和序列化；Memory：在链/代理调用之间持续存在的状态；Chain：结构化的调用序列；Agents：代理，其中 LLM 在给定高级指令和工具的情况下，反复决定操作，执行操作并观察结果，直到高级指令完成。Callbacks：回调，允许记录和流式传输任何链的中间步骤，从而轻松观察、调试和评估应用程序的内部。

5.3 提出方法

5.3.1 基线模型

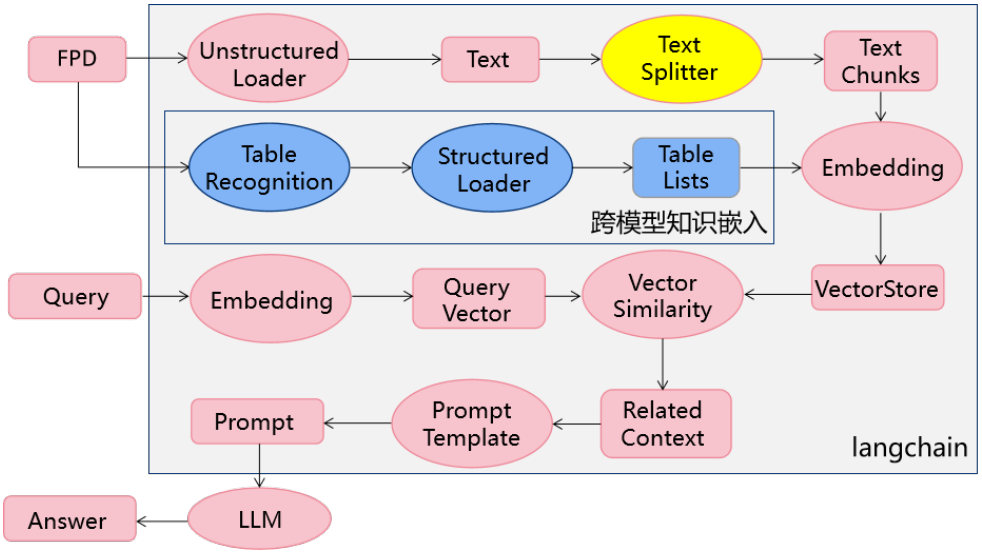

图 5.2 基于跨模型迁移的表格问答框架

如图5.2所示，红色部分和黄色部分是基本的 LangChain 架构，FPD 是飞行器参数文档，Unstructured Loader 模块首先读取本地飞行器参数文档，一行一行地加载为非结构化文本。由于文档较长，需要将文本按照字符、长度或语义进行拆分。再根据用户提问对文档内容进行字符匹配或语义检索。将匹配文本、用户提问加入提示（Prompt）模版。最后将提示发送给 LLM 获得基于文档内容的回答。具体的实现原理如图5.3所示。

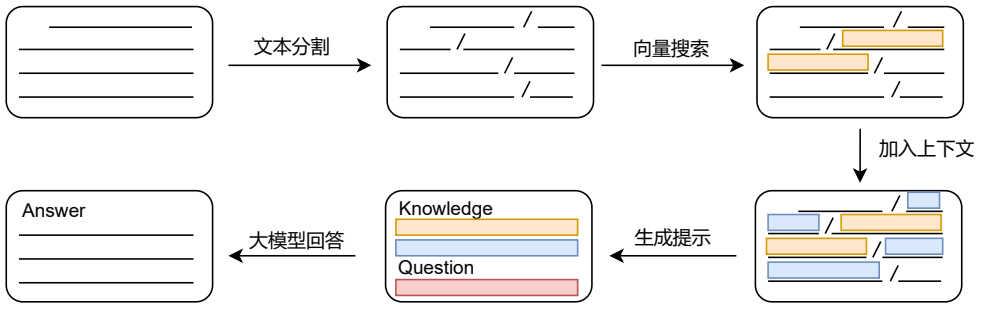

图 5.3 大模型问答原理

5.3.2 跨模型知识迁移方法

如图5.2所示，图中蓝色部分是跨模型知识嵌入的具体方法，是本研究关注的核心内容。提出方法旨在利用表格识别小模型提取表格图像的逻辑结构，从而将非结构化的表格图像转换为机器可读的格式。逻辑结构表示单元格的行/列关系（例如同行、同列）和单元格的跨度信息。机器可读的格式在本章是指 XML 格式的中间知识。具体地，文档经过版面分析模型，识别到文档中的表格图像，表格识别模型对表格图像生成结构化的 XML 序列。TableRecognition 模块借鉴了前人的工作[112]，本研究将原论文中 HTML 格式代码替换成 XML 格式，具体如图5.4所示，这是一个端到端的序列建模框架，由三个主要模块构成：CNN 图像编码器、XML 序列解码器和一个坐标序列解码器。给定一个表格图像，通过 CNN 图像编码器提取特征，然后将特征输入到 XML 序列解码器和坐标序列解码器中，分别产生 XML 序列和非空单元格的二维位置坐标。XML 序列解码器中非空单元格的表示将触发坐标序列解码器。坐标序列解码器将边框坐标（左、上、右和下坐标）建模为一个语言序列，并顺序解码这些坐标，利用坐标之间的依赖性来提高预测的准确性。为了解决逻辑表示缺乏局部视觉信息的问题，模型会训练辅助的视觉对齐损失，以增强非空单元格逻辑表示中的局部视觉表示，从而帮助产生更好的单元格二维坐标。

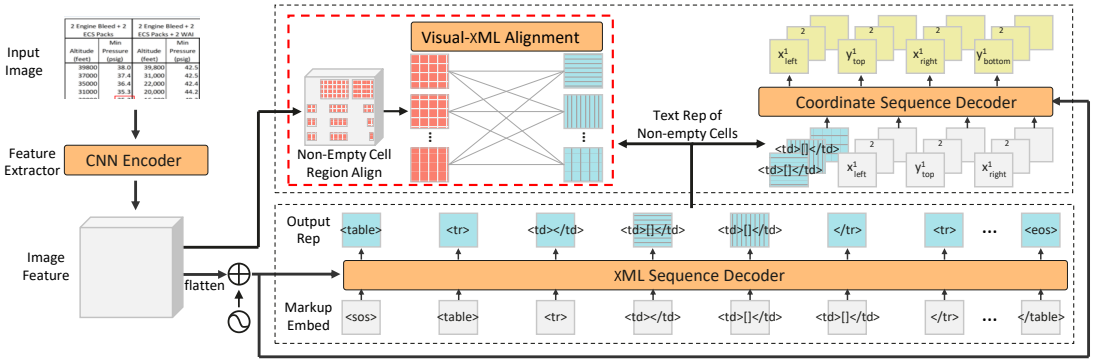

图 5.4 表格识别模型架构[112]

(1) CNN 图像编码器

模型使用全局注意力机制中的 ResNet[113] 作为 CNN 图像编码器。生成的图像特征来自第四阶段最后一个卷积层的输出。编码器的输入是大小为 $H \times W \times 3$

的 RGB 图像，编码器的输出大小为 $H\times W\times d/256$ 的特征。

（2）XML 序列解码器

表格的逻辑结构包含了单元格数量、行、列、邻接性、跨距等信息。表格识别模型生成的 XML 代码可以用来表示表格的逻辑结构，然后 XML 代码输出到 Structured Loader 模块中，被分词成 token 序列，为了减少序列长度，提出方法使用合并的标签来表示非跨距单元格。具体来说，该方法使用 $< td >< /td >$ 和 $< td >[]< /td >$ 来分别表示空单元格和非空单元格。对于跨行或跨列的单元格，XML 被标记为 $< td, colspan = "n" \ rowspan = "n" >$ 和 $< /td >$。方法规定使用第一个 token 来表示一个跨距单元格。XML 序列解码器是一个 Transformer，存储键和值的目的是为了增加二维位置编码的扁平化特征，Transformer 的输出是一个 XML 序列，通过自回归解码。第 t 个时间步的输出是一个分布：$p\left(s_t \mid \mathbf{M}, s_{1:t-1}\right)$。在训练中，采用交叉熵损失：

$$\mathcal{L}_s = -\log p\left(S^* \mid \mathbf{M}\right) = -\sum_{t=2}^{n} \log p\left(s_t^* \mid s_{1:t-1}^*, \mathbf{M}\right) \tag{5.1}$$

其中 S^* 是目标表格的 XML 标注。在训练和测试阶段，起始标记 s_1^* 或 s_1 是一个固定的 token，记为 $< sos >$。

（3）坐标序列解码器

与 XML 序列解码器类似，坐标序列解码器采用扁平化的特征和位置编码来存储键和值。起始标记的嵌入表示是 XML 序列解码器中 s_i^{nc} 的表示，标记为 f_i^{nc}。第 t 个时间步的输出是一个分布：$p\left(c_t \mid c_{1:t-1}, f_i^{nc}, \mathbf{M}\right)$，其中 $\{c_t\}_{t=1}^{4}$ 是离散随机变量。坐标序列解码器也通过交叉熵损失函数进行训练：

$$\mathcal{L}_c = -\frac{1}{K}\sum_{i=1}^{K}\sum_{t=1}^{4} \log p\left(c_t^{i*} \mid c_{1:t-1}^{i*}, f_i^{nc}, \mathbf{M}\right), \tag{5.2}$$

其中 K 是非空单元格的数量，4 个 c 分别对应单元格左、上、右和下坐标的标注。表征 f_i^{nc} 来自 XML 序列解码器，它包含上下文信息和视觉信息，可以帮助坐标序列解码器表示不同的单元格。

本章内容借助了 LangChain 的本地知识库技术与已有的表格识别模型[112]。注意，本研究根据表格识别模型，利用内部飞行器参数文档微调，生成准确的

XML 格式代码知识。通过这种跨模型知识嵌入的方式，使得视觉富文档中的表格以大模型能够理解的知识形式作为提示模版，从而避免大模型对表格内容的回答错误。

5.4 实验具体实例

针对飞行器参数文档内容进行提问：在规定模式下，给定海拔高度，最小压强应为多少?

下面展示了有关方法的具体实例:

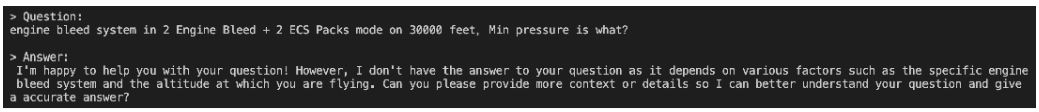

图 5.5　通用大模型的回答

仅使用大模型而不使用 LangChain 采集本地文档的情况下，如图 5.5所示，大模型回答错误。这是因为飞行器参数文档还没有以本地知识库的形式导入，诸如 ChatGPT、ChatGLM、星火等通用大模型自然不知道这种特定领域的问题。像术语速查、公式问答这种简单任务通用大模型也难以胜任。

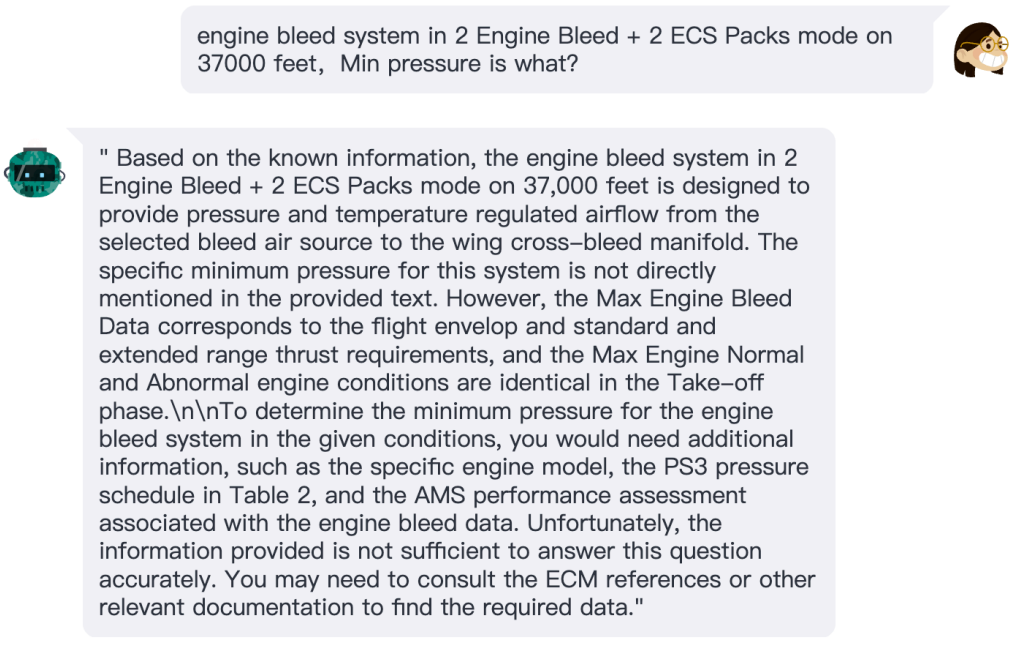

图 5.6　结合了本地知识库的回答

在使用通用大模型并使用 LangChain 采集本地文档的情况下，如图5.6所示，会出现答非所问的现象，仅仅依靠大模型理解不了飞行器参数文档中的表格数据。但是对于术语速查、公式问答这些自然语言问题，通用大模型是可以胜任的。

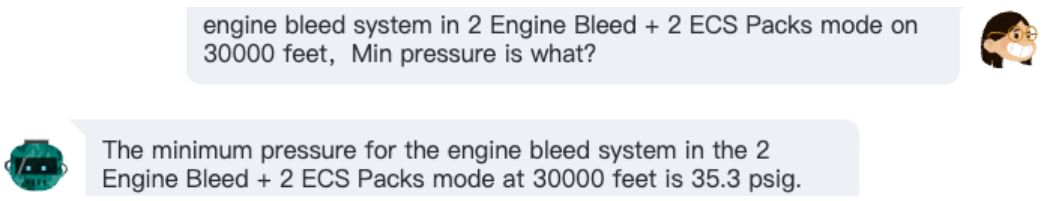

图 5.7　利用本地知识库并结合跨模型知识迁移方法的回答

当使用通用大模型和 LangChain 技术并加入跨模型的知识嵌入方法时，如图5.7所示，大模型结合了表格识别小模型生成的 XML 格式代码之后，对表格内容可以回答正确。但是并不是所有通用大模型都可以理解小模型生成的 XML 代码。本章研究由于时间有限，只评测了 ChatGPT3.5、ChatGLM1.0、星火这三个版本的通用大模型。如表5.1所示，研究通过多次评测，整理出各个方法对三项任务的实验结果，只有 ChatGPT 可以理解 XML 格式的代码。

表 5.1　各种方法的比较

通用大模型	术语速查	公式问答	表格问答
ChatGPT	✗	✗	✗
ChatGLM + Langchain	✓	✓	✗
星火 + Langchain	✓	✓	✗
ChatGPT + Langchain	✓	✓	✗
ChatGLM + Langchain + TableRecognition	✓	✓	✗
星火 + Langchain + TableRecognition	✓	✓	✗
ChatGPT + Langchain + TableRecognition	✓	✓	✓

5.5　本章小结

本章研究针对的是飞行器参数文档这一垂直领域，考虑到项目落地，研究开发了一个文档问答系统。虽然 LangChain 技术和表格识别模型是前人已有的

工作，但是本研究首次将它们结合到一起。由于文档涉密，论文中对一部分内容进行遮掩。针对大语言模型不理解二维空间表格的难题，研究工作通过小模型知识迁移的方式，利用微调之后的表格识别小模型解析文档中的表格，生成 XML 格式的代码，这样就识别出表格的逻辑结构，包括行（<tr>）、列（<th> 或 <td>），以及表头（<th>）和单元格内容（<td>）。此外，结合大模型的能力，提出的方法还可以理解表格的上下文信息，这样表格数据作为问题的一部分作为提示，本方法可以使用这些信息来回答相关问题，比如总结表格内容、提取特定数据或者解释表格的结构。需要注意的是，本方法是基于对文本和结构化数据的理解，而不是执行复杂的数据运算或统计分析。对于后一种对表格操作方面的需求，可能需要专门的指令理解方法。另外，由于数据集和时间的限制，对于表格识别模型生成的 XML 序列还没有有效评估，这是在未来需要做的工作。由于本章部分工作涉及保密条例，因此，后续实验及其它问题在此不做展开，仅就方法进行稍微讨论。最后，多模态大模型的进展后续还可以多多关注。

第六章　总结与展望

6.1　总结

随着数字化时代的到来，视觉富文档作为新兴的信息传播方式，正日益受到广泛关注和应用。与纯文本文档相比，视觉富文档拥有更为复杂的版式设计和丰富的视觉元素，如颜色、字体、边框等，这些视觉信息对文档的理解和认知至关重要。然而，由于不同领域文档的版式多样化，人工处理这些文档不仅效率低下，而且成本高昂。因此，研究智能文档模型和开发自动化抽取技术具有重要的实际应用价值。受益于已有的数据集，如何通过知识迁移的方法，使得出现新领域新类别的视觉富文档时，已有的预训练模型也能够胜任是需要解决的问题。本文还综合考虑文档内粗粒度的文本内容、版式布局和视觉特征，改进在文档智能领域的前沿模型。针对视觉富文档内表格数据复杂多样的特点，本文还尝试探索大型语言模型在理解表格方面的潜力。

本文的主要工作和贡献包括以下几个方面：

贡献一：基于跨粒度表示学习方法的研究针对有监督场景中的关系抽取任务，提出了一种创新的跨粒度迁移方法，旨在提高键值实体及其关系的抽取准确性。主要创新包括：读取顺序校正、注入常识的细粒度表示、多粒度图神经网络解决跨粒度迁移。通过广泛评估和消融实验，证明了所提方法在关系抽取任务上优于现有方法，特殊字符和粗粒度特征迁移对提升性能有关键作用。

贡献二：基于跨类别元学习方法的研究针对的是低资源场景中的关系抽取任务，在视觉富文档的自动化处理中，经常遇到资源稀缺的情况，这给从有限的样本中识别和提取新的键值对及其相互关系带来了巨大的挑战。为了解决这一问题，本研究创新性地设计了新的变分原型矫正方法。这种方法整合了二维空间的先验知识，有效减少原型表示的不确定性。在构建的多个少样本视觉富文档关系抽取数据集上进行了全面的测试和评估，实验结果证明，提出方法表现出了最佳的性能，验证了其有效性。

贡献三：跨模型的研究聚焦于提升 LLM 对视觉富文档中表格的理解能力，

提出了新颖的方法：结合现有的表格解析模型和 XML 中间语言，将表格数据以 XML 编码的形式嵌入到 LLM 的提示中，实现模型对复杂表格的理解。本研究在航空航天的垂直领域，通过项目实践将表格小模型生成的知识迁移到大模型，提升大模型对表格数据的问答效果。

6.2 展望

本文针对视觉富文档信息抽取任务中的关系抽取和表格问答进行了研究并提出了相应的知识迁移方法，但依旧存在诸多不足和可以改进之处：

（1）在进行有监督的视觉富文档关系抽取任务中，研究提出的基于跨粒度表示学习方法，尽管采用了文本、视觉和布局三种模态的特征，但视觉特征的潜力尚未完全挖掘，尚未充分利用视觉特征中字体的颜色和大小这些内容。本研究期望通过整合更多的视觉特征，从而显著提升视觉富文档的理解能力。还有，对于文本、视觉和布局这三种模态的信息目前采用的是简单的直接拼接方法，当然存在更优的融合策略，例如利用注意力机制来实现。所以，探索和实现更先进的模态融合技术，是未来研究工作的重点。最后，该研究对于阅读顺序的重建是基于规则的方法，在未来有必要讨论基于深度学习的智能排序策略。

（2）在进行少样本视觉富文档关系抽取任务时，研究提出的跨类别元学习的方法，在实验部分没有办法证明提出方法在处理不同领域的时候是否具有扩展性。本研究证明了提出方法可以提升少样本在同一领域内跨类别抽取的性能。但是面对跨领域的问题，例如将海运单领域跨越到药物说明书领域，这种跨领域的问题是未来需要去做的事情。

（3）大规模语言模型通过表格识别小模型知识迁移的方式可以理解和解析文档中的表格，包括那些以 XML 格式编写的代码。本研究中表格数据是作为提示模版，虽然可以回答表格相关的问题，比如总结表格内容、提取特定数据或者解释表格的结构。然而，本研究的方法是基于对文本和结构化数据的理解，而不是执行复杂的数据运算或统计分析。对于后一种需求，可能需要专门的指令理解方法，这个需要在未来逐步探索。

插图索引

图 1.1 生活中常见的视觉富文档 1
图 1.2 研究内容 5
图 1.3 研究创新点 6
图 1.4 论文组织结构 8
图 2.1 视觉富文档的识别结果（红色代表键实体，蓝色代表值实体，黄色代表实体间的关系） 10
图 2.2 视觉富文档理解框架 11
图 2.3 ResNet 结构图[24] 15
图 2.4 图卷积过程[9] 16
图 2.5 LayoutLM[7] 文档理解模型结构图 17
图 2.6 LayoutLMv2[8] 文档理解模型结构图 19
图 2.7 LayoutXLM[10] 文档理解模型结构图 20
图 2.8 知识迁移的发展历史 21
图 2.9 传统的机器学习和迁移学习的学习过程 22
图 2.10 迁移学习框架总结以及本文关注的重点 23
图 3.1 缺乏粗粒度级别学习的预训练模型进行关系抽取的错误案例 27
图 3.2 原始文档图像及标注数据（只截取部分） 28
图 3.3 知识迁移策略 29
图 3.4 RBERT[51] 的模型架构 30
图 3.5 跨粒度迁移模型结构 32

图 3.6 SEAB 数据集中预定义键值实体的抽取性能 39
图 3.7 通过 OCR 识别到文档的原始阅读顺序.....40
图 3.8 经过阅读顺序校正算法之后的文档阅读顺序 41
图 3.9 以 LayoutXLM 为主干网络的嵌入空间可视化消融实验 42
图 4.1 多领域视觉富文档的低资源应用场景 45
图 4.2 VFA 模型[78] 架构图 46
图 4.3 多模态数据的少样本学习与纯文本少样本学习（前人工作）的不同 47
图 4.4 拷贝，遮掩及采样示例.....50
图 4.5 SEAB 数据集中具体文档的注释实例，包含了所有实体和关系类型.51
图 4.6 跨类别迁移学习模型结构 52
图 4.7 SEAB 数据集中的“Shipper”和“Weight”类型的双驼峰分布 53
图 4.8 CORD 数据集中的“Menu”和“Total”类型的双驼峰分布 53
图 4.9 由 BERT 和 LayoutLMv2 生成的实体表示的语义相似度热力图 61
图 4.10 由 K 个样本实例估计出的原型到类的真实点的平均距离 64
图 5.1 面向飞行器参数文档的问答任务示例 66
图 5.2 基于跨模型迁移的表格问答框架.....70
图 5.3 大模型问答原理 70
图 5.4 表格识别模型架构[112] 71
图 5.5 通用大模型的回答 73
图 5.6 结合了本地知识库的回答 73
图 5.7 利用本地知识库并结合跨模型知识迁移方法的回答 74

表格索引

表 3.1 在 XLM-RoBERTa[50]、InfoXLM[29] 和 LayoutXLM[10] 三个基线模型上的消融实验。**+ST** 代表在基线模型基础上增加了特殊字符（Adding Special Token），**+MG** 代表经过跨粒度知识迁移，在基线模型上加入了多粒度神经网络（Multi-grained Graph Neural Network）.....38

表 3.2 已有的阅读顺序排列算法与 Doctrack 数据集[59] 的相似度度量比较..40

表 3.3 不同粒度局部化方法在不同参数下的比较.....42

表 4.1 重要的数学符号及其含义.....47

表 4.2 有监督数据集和少样本数据集的统计数据.....51

表 4.3 Few-SEAB 数据集中关系类型的划分，每个数字对应一个关系类型。.....51

表 4.4 在 Few-CORD 数据集上的性能（平均 F1 值）.....59

表 4.5 在 Few-SEAB 数据集上的性能（平均 F1 值）.....60

表 4.6 在 Few-SEAB 数据集上的 Inter 模式下的比较.....62

表 4.7 在 Few-SEAB 数据集上的 Intra 模式下的比较.....63

表 4.8 窗口缩放的影响.....64

表 5.1 各种方法的比较.....74

参考文献

[1] 晏文坛. 半结构化中文简历的信息抽取[D]. 华南理工大学, 2018.

[2] 林泽柠, 汪嘉鹏, 金连文. 视觉信息抽取的深度学习方法综述[J]. 中国图象图形学报, 2023, 28: 2276-2297.

[3] 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.

[4] 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.

[5] 刘建. 基于知识迁移的跨领域深度推荐方法研究[D]. 苏州大学, 2020.

[6] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[A]. 2020. arXiv: 1910.10683.

[7] XU Y, LI M, CUI L, et al. Layoutlm: Pre-training of text and layout for document image understanding[C/OL]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery, 2020: 1192–1200. https://doi.org/10.1145/3394486.3403172.

[8] XU Y, XU Y, LV T, et al. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding[C/OL]//the Annual Meeting of the Association for Computational Linguistics. 2021. https://arxiv.org/abs/2012.14740.

[9] LIU X, GAO F, ZHANG Q, et al. Graph convolution for multimodal information extraction from visually rich documents[C/OL]//NAACL. 2019: 32-39. https://arxiv.org/abs/1903.11279.

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

[11] NAGY G, SETH S C. Hierarchical representation of optically scanned documents[Z]. 1984.

[12] BAR-YOSEF I, HAGBI N, KEDEM K, et al. Line segmentation for degraded handwritten historical documents[C]//2009 10th International Conference on Document Analysis and Recognition. IEEE, 2009: 1161-1165.

[13] WONG K Y, CASEY R G, WAHL F M. Document analysis system[J/OL]. IBM Journal of Research and Development, 1982, 26(6): 647-656. DOI: 10.1147/rd.266.0647.

[14] SHI Z, GOVINDARAJU V. Line separation for complex document images using fuzzy runlength[C/OL]//First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings. 2004: 306-312. DOI: 10.1109/DIAL.2004.1263259.

[15] FISHER J L, HINDS S C, D'AMATO D P. A rule-based system for document image segmentation[C]//[1990] Proceedings. 10th International Conference on Pattern Recognition: volume 1. IEEE, 1990: 567-572.

[16] BUKHARI S S, AL AZAWI M I A, SHAFAIT F, et al. Document image segmentation using discriminative learning over connected components[C]// Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. 2010: 183-190.

[17] BAECHLER M, INGOLD R. Multi resolution layout analysis of medieval manuscripts using dynamic mlp[C]//2011 International Conference on Document Analysis and Recognition. 2011: 1185-1189.

[18] WU C C, CHOU C H, CHANG F. A machine-learning approach for analyzing document layout structures with two reading orders[J]. Pattern recognition, 2008, 41(10): 3200-3213.

[19] WEI H, BAECHLER M, SLIMANE F, et al. Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents[C]//2013 12th international conference on document analysis and recognition. IEEE, 2013: 1220-1224.

[20] WANG Y, HARALICK R, PHILLIPS I. Improvement of zone content classification by using background analysis[C]//Fourth IAPR International Workshop on Document Analysis Systems.(DAS2000). 2000.

[21] WANG Y, PHILLIPS I T, HARALICK R M. Table detection via probability optimization[C]//DAS '02. Berlin, Heidelberg: Springer-Verlag, 2002: 272–282.

[22] PINTO D, MCCALLUM A, WEI X, et al. Table extraction using conditional random fields[C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. 2003: 235-242.

[23] CHEN J, LOPRESTI D. Table detection in noisy off-line handwritten documents[C]//2011 International Conference on Document Analysis and Recognition. 2011: 399-403.

[24] 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.

[25] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[26] HE K, GKIOXARI G, DOLLáR P, et al. Mask r-cnn[A]. 2018. arXiv: 1703.06870.

[27] 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.

[28] LI M, XU Y, CUI L, et al. Docbank: A benchmark dataset for document layout analysis[A]. 2020.

[29] CHI Z, DONG L, WEI F, et al. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 3576-3588.

[30] CAVALLANTI G, CESA-BIANCHI N, GENTILE C. Linear algorithms for online multitask classification[J]. Journal of Machine Learning Research, 2010, 11 (97).

[31] PAN S J, YANG Q. A survey on transfer learning[J/OL]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359. DOI: 10.1109/TKDE.2009.191.

[32] PAN W, XIANG E W, LIU N N, et al. Transfer learning in collaborative filtering for sparsity reduction[C]//AAAI'10: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta, Georgia: AAAI Press, 2010: 230–235.

[33] SIMÕES R, MALTAROLLO V G, OLIVEIRA P R, et al. Transfer and multitask learning in qsar modeling: Advances and challenges[J/OL]. Frontiers in Pharmacology, 2018, 9. https://api.semanticscholar.org/CorpusID:3565150.

[34] ALOTHMAN B. Similarity based instance transfer learning for botnet detection [J]. Int. J. Intell. Comput. Res.(IJICR), 2018, 9(2018): 880-889.

[35] WANG W, WANG H, ZHANG C, et al. Transfer feature representation via multiple kernel learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 29. 2015.

[36] KUMAGAI W. Learning bound for parameter transfer learning[C/OL]//LEE D, SUGIYAMA M, LUXBURG U, et al. Advances in Neural Information Processing Systems: volume 29. Curran Associates, Inc., 2016. https://proceedings.neurips.cc/paper_files/paper/2016/file/7f53f8c6c730af6aeb52e66eb74d8507-Paper.pdf.

[37] WANG D, LI Y, LIN Y, et al. Relational knowledge transfer for zero-shot learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 30. 2016.

[38] XIE L, DENG Z, XU P, et al. Generalized hidden-mapping transductive transfer learning for recognition of epileptic electroencephalogram signals[J]. IEEE transactions on cybernetics, 2018, 49(6): 2200-2214.

[39] YAO Q, YANG H, YU A, et al. Transductive transfer learning-based spectrum optimization for resource reservation in seven-core elastic optical networks[J]. Journal of Lightwave Technology, 2019, 37(16): 4164-4172.

[40] SIDDHANT A, GOYAL A, METALLINOU A. Unsupervised transfer learning for spoken language understanding in intelligent agents[C]//Proceedings of the AAAI conference on artificial intelligence: volume 33. 2019: 4959-4966.

[41] LV J, CHEN W, LI Q, et al. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7948-7956.

[42] GEORGE D, SHEN H, HUERTA E. Classification and unsupervised clustering of ligo data with deep transfer learning[J]. Physical Review D, 2018, 97(10): 101501.

[43] BIANCO S, CELONA L, NAPOLETANO P, et al. On the use of deep learning for blind image quality assessment[J/OL]. Signal, Image and Video Processing, 2018, 12. DOI: 10.1007/s11760-017-1166-8.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] GAL R, ARDAZI S, SHILKROT R. Cardinal graph convolution framework for document information extraction[C]//Proceedings of the ACM Symposium on Document Engineering 2020. 2020: 1-11.

[49] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J/OL]. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019. https://aclanthology.org/N19-1423.

[50] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8440-8451.

[51] WU S, HE Y. Enriching pre-trained language model with entity information for relation classification[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 2361-2364.

[52] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[A]. 2016.

[53] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. stat, 2017, 1050: 20.

[54] CHENG M, QIU M, SHI X, et al. One-shot text field labeling using attention and belief propagation for structure information extraction[C/OL]//Proceedings of the 28th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2020: 340–348. https://doi.org/10.1145/3394171.3413511.

[55] 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. 2007: 924-928.

[56] 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]//European Conference on Computer Vision. Springer, 2020: 85-100.

[57] WANG Z, XU Y, CUI L, et al. LayoutReader: Pre-training of text and layout for reading order detection[C/OL]//MOENS M F, HUANG X, SPECIA L, et al. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 4735-4744. https://aclanthology.org/2021.emnlp-main.389. DOI: 10.18653/v1/2021.emnlp-main.389.

[58] 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.

[59] WANG H, WANG Q, LI Y, et al. DocTrack: A visually-rich document dataset really aligned with human eye movement for machine reading[C/OL]//BOUAMOR H, PINO J, BALI K. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023: 5176-5189. https://aclanthology.org/2023.findings-emnlp.344. DOI: 10.18653/v1/2023.findings-emnlp.344.

[60] 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.

[61] ZHANG J, WANG H, LUO X. Dual-vie: Dual-level graph attention network for visual information extraction[C/OL]//PRICAI 2022: Trends in Artificial Intelligence. 2022: 422-434. https://doi.org/10.1007/978-3-031-20862-1_31.

[62] HUANG Y, LV T, CUI L, et al. Layoutlmv3: Pre-training for document ai with unified text and image masking[C/OL]//Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2022: 4083–4091. https://doi.org/10.1145/3503161.3548112.

[63] HONG T, KIM D, JI M, et al. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents[C/OL]// AAAI Conference on Artificial Intelligence. 2021. https://api.semanticscholar.org/CorpusID:237485613.

[64] GARNCAREK Ł, POWALSKI R, STANISŁAWEK T, et al. Lambert: Layout-aware language modeling for information extraction[C/OL]//Document Analysis and Recognition –ICDAR 2021. 2021: 532-547. https://link.springer.com/chapter/10.1007/978-3-030-86549-8_34.

[65] LI X, ZHENG Y, HU Y, et al. Relational representation learning in visually-rich documents[C/OL]//ACM MM. 2022: 4614–4624. https://arxiv.org/abs/2205.02411.

[66] WANG Z, SHANG J. Towards few-shot entity recognition in document images: A label-aware sequence-to-sequence framework[C/OL]//MURESAN S, NAKOV P, VILLAVICENCIO A. Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational

Linguistics, 2022: 4174-4186. https://aclanthology.org/2022.findings-acl.329. DOI: 10.18653/v1/2022.findings-acl.329.

[67] WANG Z, ZHAO K, WANG Z, et al. Formulating few-shot fine-tuning towards language model pre-training: A pilot study on named entity recognition[C/OL]// GOLDBERG Y, KOZAREVA Z, ZHANG Y. Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 3186-3199. https://aclanthology.org/2022.findings-emnlp.232. DOI: 10.18653/v1/2022.findings-emnlp.232.

[68] PARK S, SHIN S, LEE B, et al. Cord: A consolidated receipt dataset for post-ocr parsing[C/OL]//Document Intelligence Workshop at NeurIPS. 2019. https://api.semanticscholar.org/CorpusID:207900784.

[69] HAN X, ZHU H, YU P, et al. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation[C/OL]//the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4803-4809. https://aclanthology.org/D18-1514.

[70] BALDINI SOARES L, FITZGERALD N, LING J, et al. Matching the blanks: Distributional similarity for relation learning[C/OL]//The 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2895-2905. https://aclanthology.org/P19-1279.

[71] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2017: 4080–4090.

[72] GAO T, HAN X, ZHU H, et al. FewRel 2.0: Towards more challenging few-shot relation classification[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019: 6250-6255. https://aclanthology.org/D19-1649.

[73] SABO O, ELAZAR Y, GOLDBERG Y, et al. Revisiting few-shot relation classification: Evaluation data and classification schemes[J/OL]. Transactions of the Association for Computational Linguistics, 2021, 9: 691-706. https://aclanthology.org/2021.tacl-1.42. DOI: 10.1162/tacl_a_00392.

[74] BRODY S, WU S, BENTON A. Towards realistic few-shot relation extraction [C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 5338-5345. https://aclanthology.org/2021.emnlp-main.433. DOI: 10.18653/v1/2021.emnlp-main.433.

[75] POPOVIC N, FÄRBER M. Few-shot document-level relation extraction [C/OL]//North American Chapter of the Association for Computational Linguistics. 2022. https://api.semanticscholar.org/CorpusID:248512966.

[76] GARCIA V, BRUNA J. Few-shot learning with graph neural networks[C/OL]// Proceedings of the 2018 Conference on Learning Representations. 2018. https://arxiv.org/abs/1711.04043.

[77] LIU Y, HU J, WAN X, et al. Learn from relation information: Towards prototype representation rectification for few-shot relation extraction[C/OL]//Findings of NAACL. 2022: 1822-1831. https://aclanthology.org/2022.findings-naacl.139.

[78] HAN J, REN Y, DING J, et al. Few-shot object detection via variational feature aggregation[C/OL]//AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. https://doi.org/10.1609/aaai.v37i1.25153.

[79] WANG P, BLUNSOM P. Collapsed variational Bayesian inference for PCFGs [C/OL]//HOCKENMAIER J, RIEDEL S. Proceedings of the Seventeenth Conference on Computational Natural Language Learning. Sofia, Bulgaria: Association for Computational Linguistics, 2013: 173-182. https://aclanthology.org/W13-3519.

[80] ISLAM R, FOULDS J. Scalable collapsed inference for high-dimensional topic models[C/OL]//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). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 2836-2845. https://aclanthology.org/N19-1291. DOI: 10.18653/v1/N19-1291.

[81] WANG J, JIN L, DING K. LiLT: A simple yet effective language-independent layout transformer for structured document understanding[C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 7747-7757. https://aclanthology.org/2022.acl-long.534.

[82] KINGMA D P, WELLING M. Auto-encoding variational bayes[C/OL]// International Conference on Learning Representations. 2013. https://browse.arxiv.org/pdf/1312.6114.pdf.

[83] YANG K, ZHENG N, DAI X, et al. Enhance prototypical network with text descriptions for few-shot relation classification[C/OL]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 2273–2276. https://doi.org/10.1145/3340531.3412153.

[84] GAO T, HAN X, LIU Z, et al. Hybrid attention-based prototypical networks for noisy few-shot relation classification[C/OL]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. 2019. https://doi.org/10.1609/aaai.v33i01.33016407.

[85] XIA C, XIONG C, YU P, et al. Composed variational natural language generation for few-shot intents[C/OL]//COHN T, HE Y, LIU Y. Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, 2020: 3379-3388. https://aclanthology.org/2020.findings-emnlp.303. DOI: 10.18653/v1/2020.findings-emnlp.303.

[86] TRAN V K, NGUYEN L M. Adversarial domain adaptation for variational neural language generation in dialogue systems[C/OL]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018: 1205-1217. https://aclanthology.org/C18-1103.

[87] DING N, XU G, CHEN Y, et al. Few-NERD: A few-shot named entity recognition dataset[C/OL]//The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021). 2021: 3198-3213. https://arxiv.org/abs/2105.07464.

[88] YANG Y, KATIYAR A. Simple and effective few-shot named entity recognition with structured nearest neighbor learning[C/OL]//WEBBER B, COHN T, HE Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics, 2020: 6365-6375. https://aclanthology.org/2020.emnlp-main.516. DOI: 10.18653/v1/2020.emnlp-main.516.

[89] RYMARCZYK D, STRUSKI L, TABOR J, et al. Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification[C/OL]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2021: 1420–1430. https://doi.org/10.1145%2F3447548.3467245.

[90] LU G, WANG H, YANG S, et al. Catch: Collaborative feature set search for automated feature engineering[C/OL]//WWW '23: Proceedings of the ACM Web Conference 2023. New York, NY, USA: Association for Computing Machinery, 2023: 1886–1896. https://doi.org/10.1145/3543507.3583527.

[91] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[A]. 2020. arXiv: 2005.14165.

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

[93] CHEN Z, JIANG F, CHEN J, et al. Phoenix: Democratizing chatgpt across languages[A]. 2023.

[94] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[A]. 2023.

[95] GONG T, LYU C, ZHANG S, et al. Multimodal-gpt: A vision and language model for dialogue with humans[A]. 2023.

[96] HUANG R, LI M, YANG D, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head[A]. 2023. arXiv: 2304.12995.

[97] HU C, FU J, DU C, et al. Chatdb: Augmenting llms with databases as their symbolic memory[A]. 2023. arXiv: 2306.03901.

[98] ZHONG V, XIONG C, SOCHER R. Seq2sql: Generating structured queries from natural language using reinforcement learning[A]. 2017. arXiv: 1709.00103.

[99] LI J, HUI B, CHENG R, et al. Graphix-t5: Mixing pre-trained transformers with graph-aware layers for text-to-sql parsing[A]. 2023. arXiv: 2301.07507.

[100] LI H, SU J, CHEN Y, et al. Sheetcopilot: Bringing software productivity to the next level through large language models[A]. 2023. arXiv: 2305.19308.

[101] SCHREIBER S, AGNE S, WOLF I, et al. Deepdesrt: Deep learning for detection and structure recognition of tables in document images[C/OL]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): volume 01. 2017: 1162-1167. DOI: 10.1109/ICDAR.2017.192.

[102] PALIWAL S, D V, RAHUL R, et al. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images [A]. 2020. arXiv: 2001.01469.

[103] TENSMEYER C, MORARIU V I, PRICE B, et al. Deep splitting and merging for table structure decomposition[C/OL]//2019 International Conference on Document Analysis and Recognition (ICDAR). 2019: 114-121. DOI: 10.1109/ICDAR.2019.00027.

[104] GUO Z, YU Y, LV P, et al. Trust: An accurate and end-to-end table structure recognizer using splitting-based transformers[A]. 2022. arXiv: 2208.14687.

[105] RAJA S, MONDAL A, JAWAHAR C V. Table structure recognition using top-down and bottom-up cues[C]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 70-86.

[106] LIU H, LI X, LIU B, et al. Show, read and reason: Table structure recognition with flexible context aggregator[C/OL]//MM '21: Proceedings of the 29th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2021: 1084–1092. https://doi.org/10.1145/3474085.3481534.

[107] QASIM S R, KIESELER J, IIYAMA Y, et al. Learning representations of irregular particle-detector geometry with distance-weighted graph networks[J/OL]. The European Physical Journal C, 2019, 79(7). http://dx.doi.org/10.1140/epjc/s10052-019-7113-9.

[108] DENG Y, KANERVISTO A, LING J, et al. Image-to-markup generation with coarse-to-fine attention[A]. 2017. arXiv: 1609.04938.

[109] ZHONG X, SHAFIEIBAVANI E, JIMENO YEPES A. Image-based table recognition: Data, model, and evaluation[C]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 564-580.

[110] YE J, QI X, HE Y, et al. Pingan-vcgroup's solution for icdar 2021 competition on scientific literature parsing task b: Table recognition to html[A]. 2021. arXiv: 2105.01848.

[111] NASSAR A, LIVATHINOS N, LYSAK M, et al. Tableformer: Table structure understanding with transformers[A]. 2022. arXiv: 2203.01017.

[112] HUANG Y, LU N, CHEN D, et al. Improving table structure recognition with visual-alignment sequential coordinate modeling[A]. 2023. arXiv: 2303.06949.

[113] LU N, YU W, QI X, et al. Master: Multi-aspect non-local network for scene text recognition[J/OL]. Pattern Recognition, 2021, 117: 107980. http://dx.doi.org/10.1016/j.patcog.2021.107980.

作者在攻读硕士学位期间发表的论文与研究成果

一、发表论文

[1] Hao Wang, **Tang Li**, Chenhui Chu, Nengjun Zhu, Rui Wang*, Pinpin Zhu. Towards Human-Like Machine Comprehension: Few-Shot Relational Learning in Visually-Rich Documents. In Proceeding of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 16557–16569, Torino, Italy. International Committee on Computational Linguistics. (第二作者, 导师一作, CCF-B).

[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, 第十六届全国知识图谱与语义计算大会 (CCKS), 2022, poster, (第一作者).

二、参与项目

1. 国家自然科学基金青年项目“基于人机协同拟态学习的富文档隐性模式与知识获取研究”(项目编号：62306173), 参与, 2024-01-01 至 2026-12-31。

2. 上海市科技创新行动扬帆计划项目, “药物警戒中跨语言多知识驱动的不良事件抽取方法研究”(项目编号：21YF1413900), 参与, 2021-05 至 2024-04。

3. 横向研发项目-企业委托, “基于多模态预训练模型的国际物流表单结构化抽取方法研究”, 参与, 2021-10 至 2022-10。

致　谢

在这篇关于视觉富文档的硕士论文即将画上句号之际，我满怀感激地向我的导师、家人、朋友以及所有在这段学术旅程中给予我支持的人表达我的诚挚感谢。

首先，我要感谢我的导师，王老师和朱老师，他不仅以其深厚的专业知识和严谨的学术态度指导我的研究，更以其无私的关怀和鼓励陪伴我度过了每一个挑战。两位老师的洞察力和建议对于我的研究工作至关重要，他们的智慧和耐心是我不断前进的动力。

我还要感谢课题组的同学们，除了科研，他们还是我生活上的好伙伴，他们在我遇到困难时总是伸出援手，与我分享知识，共同探讨问题，他们的友谊和合作精神是我宝贵的财富。

对于我的家人，尤其是我的父母和兄弟姐妹，他们的爱和支持是我最坚强的后盾。在我面临压力和挑战时，是他们的鼓励和理解让我能够坚持下去。

此外，我要感谢上海大学计算机工程与科学学院提供的资源和环境，以及所有为我的学术研究提供帮助的教职工和技术人员。

最后，我要感谢所有参与我论文评审的专家和学者，他们的宝贵意见让我的研究更加完善。

感谢所有给予我帮助和启发的人，是你们让这段旅程充满了意义和价值。