中图分类号：　　单位代号：10280
密　　级：　　学　　号：20721617

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	基于上下文感知的材料科学文献中图文信息挖掘方法研究

作　　者　夏锦桦
学科专业　软件工程
导　　师　韩越兴
完成日期　2023.03.01

姓　　名：夏锦桦　　学号：20721617

论文题目：基于上下文感知的材料科学文献中图文信息挖掘方法研究

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：

主任：

委员：

导　　师：

答辩日期：2023.6.7

姓　　名：夏锦桦　　学号：20721617

论文题目：基于上下文感知的材料科学文献中图文信息挖掘方法研究

原 创 性 声 明

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

签　名：夏锦桦　日 期：2023.6.8

本论文使用授权说明

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

（保密的论文在解密后应遵守此规定）

签 名：夏锦桦　导师签名：[illegible]　日期：2023.6.8

上海大学工程硕士学位论文

基于上下文感知的材料科学文献中图文信息挖掘方法研究

姓　　名：夏锦桦
导　　师：韩越兴
学科专业：软件工程

上海大学计算机工程与科学学院

2023 年 03 月

A Dissertation Submitted to Shanghai University for the

Degree of Master in Engineering

Research on Contextual Awareness Based Method for Mining Graph and Text Information in Materials Science Literature

MA Candidate：Jinhua Xia

Supervisor：Yuexing Han

Major：Software Engineering

School of Computer Engineering and Science, Shanghai University

03, 2023

摘　要

文献挖掘是一种利用自然语言处理、数据挖掘、机器学习等技术，从海量的文献中提取有价值的信息和知识的方法。它可以帮助学者发现相关领域的研究趋势以及隐含的知识信息等。材料领域的文献挖掘对材料的性能预测、工艺优化、新材料研发等有着重要的推动作用。常见的材料文献挖掘技术往往针对单一的数据类型进行信息提取，而科学文献中的图像、表格和非结构化文本等多种不同类型的数据中都蕴含着重要信息。这些不同类型的数据很难通过统一的方法进行挖掘，给材料文献挖掘的发展带来了一定的挑战。为了提高材料文献挖掘的全面性和准确性，本论文针对材料科学文献中的单一数值图，融合了目标检测、命名实体识别、文本相似度匹配等方法，分别提出了材料科学文献中数值图图文信息提取方法和基于图文的材料科学文献数值图坐标轴实体识别提升方法。

首先，本论文结合数值图及其对应的标题，提出了一种图片和文本相结合的文献挖掘方法。该方法首先使用 Yolov5s 截取科学文献中的单一数值图图片，并应用改进的科学文献图片检测方法来提升准确性。接着利用 PDFminer 工具解析科学文献中的文本内容。然后计算语句间的余弦相似度和 Jaccard 相似度匹配数值图对应的标题文本。其次采用 Sci-Bert 模型与 CRF 算法在标题中识别坐标轴名称。另外应用形态学操作和字符识别等技术从数值图图上提取具体的数据信息。最后将挖掘出的坐标轴名称和数据整合以获得完整的数值图信息。

其次，针对上述识别数值图坐标轴名称任务中模型识别精度低的问题，本文抓住科学文献中数值图图片和文本之间的关系，提出了一种提升识别效果的方法。该方法首先识别数值图图上的标签文本，并将其填入样本模板以生成无需标注的文本数据，达到数据增强的效果。同时，利用文本相似度匹配技术在科学文献的正文部分寻找对应的数值图描述语句，将其以扩充文本的形式与标题文本拼接，依靠上下文关联改善输入语句生成的向

量表征，从而优化模型的预测性能。

本论文通过实验验证了上述方法的有效性与可靠性，同时开发了可视化工具以供学者使用，为多元化的文献挖掘提供了新的思路，开辟了文献挖掘的新形式。同时，该方法帮助材料领域的学者进行大规模的材料科学文献中单一数值图的信息挖掘，推动了材料领域的发展，加快了新型材料的制备进程。

关键词：材料文献挖掘；深度学习；Bert；命名实体识别；图像处理

ABSTRACT

Literature mining is a method to extract valuable information and knowledge from massive literature by using natural language processing, data mining, machine learning and other technologies. It can help scholars to discover research trends and hidden knowledge in related fields. Material literature mining plays an important role in promoting material performance prediction, process optimization and new material development. Common material literature mining technologies extract information based on a single data type, but many different types of data such as figures, tables and unstructured text in scientific literature contain important information. These different types of data are difficult to be mined by the same method, which brings certain challenges to the development of literature mining in the field of materials. In order to improve the comprehensiveness and accuracy of material literature mining, this paper targets single curve graphs in material scientific literature and integrates methods such as object detection, named entity recognition, and text similarity matching. It proposes methods for extracting curve graph information from text and figures in materials scientific literature and for enhancing the recognition of coordinate axis entities in curve graphs of material scientific literature based on text and graphs.

Firstly, In association with curve graphs and their corresponding titles, this paper proposes a literature mining method based on the combination of figures and text. The method initially uses Yolov5s to intercept the single curve graph in the scientific literature, and applies an improved scientific literature figure detection method to enhance the accuracy of interception. Then it utilizes PDFminer tool to parse the text content in the literature. Next, it calculates cosine similarity and Jaccard similarity between sentences to find the title text

corresponding to the graph. After that, it adopts Sci-Bert and CRF to recognize coordinate axis names in titles. In addition, it applies morphological operations and character recognition techniques to extract specific data information from graphs. Finally, it integrates the mined axis names and data to obtain complete curve graph information.

Secondly, to address the problem of low accuracy for recognizing coordinate axis names in the above task, this paper exploits the relationship between curve graphs and text in scientific literature and proposes a method to improve the effect of recognition. The method first recognizes the label text on the graphs and fills it into the template to generate annotated text data, achieving the effect of data augmentation. At the same time, it uses text similarity matching to find the corresponding descriptive statements in the main text of scientific literature and concatenates them with the title text as extended text. In accordance with the contextual association, it improves the vector representation of the input sentence and optimizes the performance of model.

This paper verifies the effectiveness and reliability of the above method through experiments and develops a visualization tool for scholars to use, which provides a new idea for diversified literature mining and opens up a new form of literature mining. At the same time, this method helps the scholars to carry out large-scale curve graph information mining in the material scientific literature, promoting the development of materials and accelerating the preparation process of new materials.

Keywords: Material Literature Mining, Deep Learning, Bert, Named Entity Recognition, Image Processing

目录

第一章 绪论 **1**

1.1 课题来源 1

1.2 课题背景概述 1

1.3 课题研究的目的与意义 2

1.4 国内外研究现状 4

1.4.1 文本挖掘研究现状 4

1.4.2 图表挖掘研究现状 7

1.4.3 文本与图表联合挖掘研究现状 9

1.5 论文主要工作 10

1.6 论文组织结构 11

第二章 相关理论与技术概述 **13**

2.1 目标检测理论概述 13

2.1.1 目标检测 13

2.1.2 卷积神经网络架构 14

2.1.3 Yolov5 网络架构 18

2.2 文本挖掘理论概述 22

2.2.1 命名实体识别及文本匹配任务 22

2.2.2 文本表示方法 23

2.2.3 Bert 预训练模型 25

2.2.4 CRF 算法 26

2.3 评价指标概述 28

2.4 本章小结 29

第三章 材料科学文献中数值图图文信息提取方法 **31**

3.1 方法概述 31

3.1.1 文本和数值图图片的获取 32
3.1.2 数值图标题文本的匹配 34
3.1.3 数值图坐标轴名称的识别 36
3.1.4 数值图真实数据的提取 37
3.2 实验与讨论 40
3.2.1 数据准备和实验设置 41
3.2.2 单一数值图截取任务结果 42
3.2.3 数值图标题匹配任务结果 44
3.2.4 NER 任务结果 46
3.2.5 数值图图片数据提取任务结果 48
3.2.6 具体应用 49
3.3 本章小结 51

第四章 基于图文的材料科学文献数值图坐标轴实体识别提升方法 53
4.1 方法概述 53
4.1.1 预处理 54
4.1.2 寻找扩充文本 55
4.1.3 扩充数据集 56
4.1.4 命名实体识别 58
4.2 实验与讨论 60
4.2.1 数据准备和实验设置 60
4.2.2 标题扩充文本寻找结果 60
4.2.3 命名实体识别结果 62
4.3 本章小结 66

第五章 材料科学文献数值图信息提取软件设计与实现 68
5.1 开发环境 68
5.2 需求分析 68

5.3 软件架构设计 69
5.4 软件实现 70
5.4.1 界面设计 70
5.4.2 整体流程 71
5.5 本章小结 73

第六章 总结与展望 **74**
6.1 结论 74
6.2 展望 75

参考文献 **76**

作者在攻读硕士学位期间公开发表的论文 **88**

致谢 **89**

第一章　绪论

1.1　课题来源

本课题来源于国家重点研发计划（编号：2018YFB0704400，2018YFB0704402，2020YFB0704500）；国家自然科学基金面上项目（编号：52273228）；上海市自然科学基金项目（编号：20ZR1419000）；之江实验室科研攻关项目（编号：2021PE0AC02）；上海市“科技创新行动计划”启明星项目（扬帆专项）（编号：23YF1412900）。

1.2　课题背景概述

人工智能是一门研究如何让计算机模拟人类智能的技术，其在图像识别、语音识别、自然语言处理、机器人等领域都实现了突破。人工智能技术的快速发展为众多应用场景提供了技术支撑，促进了安防、金融、交通、教育、医疗、智能制造、文娱等领域的创新变革。同时，人工智能技术的赋能作用也为各行各业带来了效率和质量的提升，推动了领域学科内涵的深化，加速了各门学科在理论分析、实验发现、数据处理、研究范式等方面的进步。

人工智能技术为材料科学提供了新的工具和方法，有望为材料科学带来范式化革命。材料科学是一门涉及多个学科领域的交叉学科，为科技创新和社会发展提供了重要的基础。新型材料的设计与研发关系到人类的生存与发展，在各个领域具有广泛的应用前景，推动着科技进步和产业转型。国家的技术竞争力与经济实力很大程度上取决于高性能、高规格的新兴材料。然而，从新型材料的研发与设计至大规模的批量应用于工业化生产是一个复杂而漫长的过程，需要进行大量的实验验证，且实验成本高昂，耗时耗力，这与国家追求高水平、高质量发展的目标存在着矛盾。因此，如何使用人工智能技术更好地服务材料科学，从而加快新材料的研发与设计速度是 21 世纪的一项重大挑战。当前，各个国家都在人工智能与材料科学的跨领域学科中投入大量资源。美国于 2016 年发布《国家人工智能研究与发展战略计划》[1]，提出要高度重视人工智能与材料科学的结合，如利用人工智能设计新材料、优化材

料性能、加速材料发现等；中国于 2022 年制定了《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》，提出了以人工智能赋能材料科学的总体要求、发展目标和重点任务，包括打造人工智能重大场景、提升人工智能场景创新能力、完善人工智能场景创新生态等。

材料基因计划 [2] 是一项以人工智能为支撑的材料领域工程，旨在利用机器学习方法帮助构建和利用材料大数据，挖掘材料的内在规律和知识，预测材料的性能和行为，从而优化材料的设计和制备，辅助材料的实验和测试等。该计划基于大规模的材料数据，利用理论与计算来预测材料成分、结构、工艺和性能间的关系，并利用机器学习方法对材料的结构和性能进行分析，为工业界定制特定材料。在这一过程中，材料科学文献是一种重要的数据来源，它包含大量的实验结果和理论分析，可以为材料基因工程提供有价值的知识和指导。然而，材料科学文献中的数据大多是以自然语言或图片的形式表达，不易于机器直接处理和利用。因此，如何从材料科学文献中自动抽取出相关信息并将其转化为可使用的数据，是实现材料基因工程目标的关键技术之一。

材料文献挖掘可以帮助材料基因工程确定计算模型的构建原则、设定合适的计算输入参数、提供判断计算结果合理性的依据、发现预测材料性能的量化构效关系，是实现预测、达到新材料研发双减半战略目标的保障。同时，从材料科学文献中挖掘出的数据是满足材料基因工程数据需求的重要支撑，是构建和完善专业数据库的根基。因此，本研究结合数字图像处理和深度学习等技术，并将其应用于材料文献挖掘，为材料基因工程技术创新与发展提供数据基础。

1.3 课题研究的目的与意义

随着科学技术的飞速发展，各领域的科学文献数量呈爆炸式增长。例如在 Web of Science 中，材料主题的科学文献总量超过了 1500 万篇，2022 年份发表量也达到了 110 万。丰富的文献资源给文献挖掘提供了丰富的数据基础，但如何从大规模科学文献中自动化提取信息成为亟需解决的问题。虽然像 Google Scholar、CiteSeer 等学术数据库可以按关键字等索引帮助检索文献，但无法自动化地帮助学者提取和

收集科学文献中的数据信息。文献挖掘是一种利用自然语言处理、机器学习、信息检索等技术从科学文献中抽取出有价值的知识和信息的过程。文献挖掘在各个领域都有广泛的应用，例如在生物医学领域，文献挖掘可以帮助识别基因、蛋白质、药物等实体及其相互作用，为药物发现和疾病诊断提供支持。在材料领域，文献挖掘可以帮助提取材料的成分、结构、性能等信息，为新材料设计提供数据基础。

科学文献是由非结构化文本、数值图等多种类型的数据构成的复杂信息载体。针对文本数据的挖掘，主要面临着语言结构的复杂性、模糊性和歧义性等挑战，导致机器难以实现对文本的深层理解 [3]；针对数值图数据的挖掘，主要面临着数值图的多样性、分辨率的低质量和数据的不准确性等问题，导致提取出的数据准确率偏低。不同形式的数据在挖掘时存在着方法的差异性和难度性，造成了学者们进行文献挖掘时往往只能关注其中某一种类型的数据，而忽略了数据之间的内在联系。在科学文献中，不同形式的数据之间具有高度的相关性和互补性，例如在材料领域里，学者们常常需要从科学文献中获取他人的实验数据来辅助自己的研究，避免重复实验，从而节省成本和时间 [4]。这些实验数据往往以数值图的形式呈现在科学文献中，但数值图所描述的对象、变量名称等信息往往只在正文中出现。缺少这些文本信息，数值图所能表达的内容就仅限于数值本身，信息量大大降低，学者就无法真正理解这些数据。因此只有在文献挖掘时将不同类型的数据进行整合，才能提高所得信息的丰富性和健壮性。

文献挖掘可以分为基于规则的方法与基于学习的方法。基于规则的方法主要依赖于领域专家手工制定的规则来对科学文献文本中的关键名词、数值进行提取，以及通过边界框线截取科学文献中的图表。该方法虽然简单，但是制定规则的过程费时费力，且方法的迁移性和扩展性较差，提取效果也不理想 [5]。随着机器学习与深度学习技术的飞速发展，文献挖掘出现了新的思路和方法，基于规则的方法逐渐被基于学习的方法所取代。基于学习的方法需要学者手动标注数据集，使用模型自动学习特征，实现端对端的自动化处理，有效地提高了文献挖掘的准确性和效率 [6]。但这种方法也面临着一些难点，例如数据集问题，模型在特征学习时需要大量的标注数据，而公开的领域数据集往往无法满足学者需求，这就需要手动标注大量数据，

而标注任务只能交给领域内的学者，人力成本较高；另一个难点是从图表中获取底层数据是一个反向工程，可能面临各种复杂和变化的情况，对于机器而言十分困难[7]。

综上所述，本课题针对材料文献挖掘的特点和难点，结合目标检测、命名实体识别、文本相似度匹配等技术，提出了一种基于上下文的材料科学文献中数值图信息挖掘的方法。该方法旨在解决目前材料文献挖掘方法提取信息单一化的问题，从材料科学文献中的单一数值图图片上提取数值信息，并从对应的标题中识别数值图的坐标轴实体名称，将二者结合以提取单一数值图的全面信息。此外，基于科学文献中图片与文本之间的特点，以数据增强和改善输入语句的向量表征两个途径，提高单一数值图坐标轴实体识别的效果。本课题结合材料科学文献中的文本与图片两种类型的数据进行挖掘，为材料文献挖掘提供了新思路，同时也方便材料学者进行研究，推动了材料基因工程技术与深度学习技术的发展。

1.4 国内外研究现状

本论文的研究目标是基于材料科学文献中的单一数值图图片与文本，联合挖掘这两种数据中的信息并进行整合。为了实现这一目标，本节首先介绍了文本挖掘和图表挖掘的相关研究，分析了它们的方法和特点，然后介绍了结合文本与图表的文献挖掘相关研究，探讨了它们的意义和挑战。这些研究为本论文提供了理论基础和技术参考。

1.4.1 文本挖掘研究现状

文本挖掘是自然语言处理的一个重要分支，具有重要的理论价值和实际意义，同时也面临着诸多挑战。文本挖掘的主要目标是从非结构化或半结构化的文本中提取有用的知识，并应用于信息管理、知识发现等[8]。文本挖掘技术涉及多个学科领域，已广泛应用于生物医学、历史学、语言学等。例如在医学领域中，从临床文本中挖掘疾病的症状，从而更早地做出预防和治疗[9]；在文学领域中，对文学作品进

行挖掘，挖掘出作品的核心主题、现实意义和艺术价值等信息，为影视的二次改编提供有效参考和借鉴 [10]；在军事领域中，快速地对军事情报进行搜集、处理并分析出可用情报，不仅减轻了情报人员的负担，还提高了分析的效率 [11]；在商业领域中，通过分析商业新闻确定评论者对公司、产品、服务、人员或事件表达的情绪或态度，从而推动经济的发展 [12]；在材料领域中，获取资料中材料的名称、属性及其对应的数值关系，避免重复实验，从而加快新材料的研发 [13]。文本挖掘受到了广大学者的深入研究和讨论，目前文本挖掘技术可分为基于规则的方法和基于学习的方法。

基于规则的方法是一种依赖于领域专家制定的规则来对文本中的关键名词、数值进行提取的方法。这种方法需要学者具备某一领域的专业知识，例如生物医学中如何陈述生物相关事实的特定知识，以及生物科学家谈论的一系列事物及其彼此之间可能存在的关系等知识，然后根据这些知识以及它们所有的变体形式建立完整的字典库 [14]。Müller 等人 [15] 就生物学知识建立了基因、细胞等共计 14500 种的概念类别，然后在每个类别中填入属于该类别的对象可能的单词或短语表示，并根据类别联系创建了 33 种对象之间的关系，最终形成了一个完善的生物学的语料字典，将查找生物数据类型的查全率从 45% 提高到 95%，同时加快了对象关系抽取的速度，实现了人工搜寻到自动化获取知识的转变。一些术语字典也在不断地更新，成为标准化的数据集，为学者们的研究提供了便利 [16]。Kang 等人 [17] 在 MetaMap 的基础上，加入了缩写扩展、边界校正、术语调整、过滤等新规则，使得疾病识别在所有评估指标上都得到了改善。但这种基于规则的方法的识别效果受限于字典内容，并且一些领域内的术语变化非常频繁，例如在医学领域出现的术语中，大约三分之一的是变体，每种病症类型都有不同的同义词，且仍在不停地变化着。因此，基于规则的方法逐渐成为文本挖掘的辅助方法。

随着机器学习和深度学习技术的发展，基于学习的方法成为了文本挖掘的主流方法，且研究人员发现基于学习的方法能够更精准地挖掘出文本中的信息。Chen 等人 [18] 首先从公开字典集中摘录了药物名称和副作用目录，再对文本语料库进行了 n-gram 的频率分析筛选出链接药物名称和副作用的短语，最后定义了隐藏的状态序

列 {药物，关键字，副作用}，利用隐马尔科夫模型从消息中抽取关系，比基于规则的方法所得到的 F1 分数提高了近 20%。Saleh 等人 [19] 进行了文本情感分析的实验，首先手动将意见文本分为积极的或是消极的，以此作为数据集，并计算词频-逆文档频率（TF-IDF）和单词频率，使用向量空间模型来表示文本语句，最后利用支持向量机进行训练，分类精度达到了 80%。Vazquez 等人 [20] 利用聚类方法分析电子健康记录文本，识别慢性阻塞性肺病患者群，帮助诊断和治疗患者常见的合并症，加速和提高了医疗状况的诊断和治疗的准确性。

深度学习相比机器学习在数据集充足的情况下能够表现出更优异的学习效果。因此，基于学习的文本挖掘方法将重心从机器学习转向了深度学习。Kurniasari 等人 [21] 就基于循环神经网络（Recurrnet Neural Network, RNN）对文本进行了情感分析实验。他们改变了以往用频率等数值来表示文本，而将文本送入静态词向量模型 Word2Vec 中得到向量特征并使用 RNN 训练分类器，最终的分类结果对比朴素贝叶斯算法提升了 45% 以上。该实验证明了 RNN 在文本挖掘上的优势，此外，将文本送入词向量模型获得特征也成为了学者们的首选。长短时间记忆网络（Long short-term memory, LSTM）[22] 解决了 RNN 在训练长序列文本过程中的梯度消失问题，具备更强更好的记忆性能。L.Weston 等人 [23] 标注了 800 篇材料文献的摘要，利用静态模型将文本转换为向量并使用 LSTM 网络进行训练，从文献中提取无机材料名称、相标签和应用等信息，提取精度达到了 87%。LSTM 网络无法对从后向前的信息进行编码，为了更好地捕捉双向语义，Bi-LSTM 网络被提出，该模型使用两个 LSTM 并行运行捕获双向的信息。Ali 等人 [24] 使用 Bi-LSTM 改进了文本的特征提取，提升了文本分类任务的精度，较之 LSTM 网络在准确率上提升了 8%。Bert 预训练模型 [25] 是谷歌团队于 2018 年基于维基百科文本数据集训练得到的，该模型利用注意力机制实现了对语句上下文的动态编码，克服了静态模型无法捕捉上下文语义的缺陷，在多个自然语言处理任务中展现了优异的性能，也成为了文本挖掘任务的首选模型。Zhao 等人 [26] 在 Weston 等人工作的基础上，采用 Bert 进行文本特征提取，提高了各类实体的识别效果。许多研究者在 Bert 的基础上做出了改进并将其应用到文本挖掘任务中。例如，Sentence-Bert[27] 是一种基于孪生网

络思想的语句相似度匹配网络，它将两个句子输入到两个参数共享的 Bert 中，通过平均池化得到每个句子的句向量表示，然后计算相似度，从而降低了 Bert 语义相似度检索的时间开销。Wang 等人 [28] 使用 Sentence-Bert 来计算生物医学语句之间的相似度，然后建立欧几里得空间来表征句子表示之间的几何结构信息，从而揭示句子间的本质规律，提高下游文本任务的效果。然而，在专业领域文本上使用公共领域文本训练的模型进行词向量编码时，很可能出现词库不匹配的问题，从而影响模型的性能。为了解决该问题，一些针对专业领域下进行预训练的模型应运而生 [29, 30, 31]，它们都是在大量的科学文本下进行预训练，使得专业性的术语的向量表征变得更加准确。Avan Kumar 等人 [32] 就利用 Sci-Bert 从文献中收集制氢的催化剂和工艺参数，其结果准确率高达 99.8%。

但以上基于深度学习的文本挖掘方法都需要大规模的标注数据集作为支撑，而标注数据集需要耗费大量的时间和资源。实验证明 [33]，数据集不充足会显著地降低文本挖掘的准确度。因此，很多学者开始研究文本数据增强技术。常见的增强方法 [34] 包括对文本中的单词进行同义词替换以及随机交换等，从而获取更多与训练语料相似的语句，然而这种方法生成的只是伪造样本，随着真实的样本数据量变大，其作用会下降。Ye 等人 [35] 将 Prompt 思想应用于命名实体识别任务中，将候选实体与自定义的模板进行拼接，生成更多的样本语句，但其所指定的模板不具有普适性，只适用于指定领域的任务，且根据模板判断一个实体时需要计算所有模板的得分并比较，费时费力。如何在有限的数据集下保证文本挖掘的准确性成为了研究的热点。

1.4.2 图表挖掘研究现状

除了大量的文本内容，科学文献中的图表也是一种重要的数据信息来源，它们直观地展示作者们的实验数字结果 [36]，为学者提供更多的研究线索。因此，如何从图表中提取有用的信息也成为了一个研究热点。以往从科学文献中分割图表的方法主要有基于栅格和基于矢量两种，这两种方法都是基于规则的，因此提取的精度不高，且无法获取对应图表的标题。随着卷积神经网络的发展，检测和分类文献中的

图表变得简单且高效。Kavasidis 等人 [37] 对 VGG-16 网络进行了改进，首先将方形卷积核更改为矩形卷积核，以便提取与图表相关的行、列和间距等特征，然后加入了碰撞卷积层来建立多尺度和长范围的关系，最后加入了 CRF 层来提高检测位置的准确性。他们的方法在图表分类上达到了 94.5% 的精度，但是仍然无法匹配对应图表的标题。Siegel 等人 [38] 在成功检测图表后，利用正则表达式匹配以“Figure”或“Table”开头的语句作为图表标题，使用标准 PDF 处理库在文献页中定位这些标题，然后使用匈牙利算法计算在页中距离最近的图表进行配对，从而完成了文献中图表与对应标题的匹配。

在获得科学文献中的图表图片之后，研究者们开始着眼如何从这些图上挖掘出数据信息。然而，由于科学文献中图表的结构多样性以及作者们绘制图表时采用的不同设计惯例，图表信息提取的研究面临着巨大的挑战。图表挖掘任务主要包括文字检测与识别、刻度轴检测理解、数值提取等 [39]。其中，文字检测与识别是图表挖掘的关键步骤，因为文本在图表图像中承担着不同的角色，识别这些文本才能进一步地挖掘信息。目前，光学字符识别技术（Optical Character Recognition，OCR）已经发展成熟，能够快速高效地检测和识别图上的所有文字 [40]。在此基础上，需要对每个文本区域的角色进行分类，如坐标轴标题、轴刻度、图例标题和图例条目等。常见的分类方法基于文本的布局位置，通过计算文本的中心坐标、到图表图片边界的距离以及旋转角度等信息来判断文本角色 [41]。机器学习和深度学习技术也可以通过学习文本特征来预测角色。为了提取数值图所描绘的数值大小，必须通过分析轴线、轴刻度值来确定每个轴的尺度和范围。常见的检测轴线的方法包括投影轮廓分析 [42] 和霍夫变换 [43] 等。在确定完轴线之后，从刻度值文本之间的像素级距离进行刻度值的计算。数值提取需要先在图上找到构成图表线的像素，然后再利用计算好的刻度进行真实数据的转换。分割图表线的方法大致分为基于采样的、基于跟踪的和基于分割的三种。基于采样的方法 [44] 需要找到图表线与一组垂直线相交的点，然后将这些点拟合成直线，该方法主要针对断线和虚线。基于跟踪的方法 [45] 通过扫描图表数据区域，找到构成图表线的像素，并利用像素关系来连接每一条曲线，但当图表曲线有明显的梯度时，该方法可能会失效。基于分割的方法旨在

从背景像素中分割出整条图表线，包括基于颜色的启发式 [46] 和使用深度神经网络进行学习特征 [47]，该方法也是最常用的方法。

现有的从数值图上提取数据的方法分为半自动和全自动两种。半自动化工具 [48, 49] 虽然非常精确，但在提取数据时需要人工标注图中的像素点和坐标轴并输入横纵坐标的刻度差，这样的方法显然费时费力，且效率不高。Cliche 等人 [50] 通过标记检测、OCR 等操作实现了自动化提取散点图的数值信息，但其不具有普适性，且提取出的仅仅是数值大小，缺少其他描述性的信息。Hsu 等人 [51] 融合了数值和轴标签设计出了为数值图自动化生成文字说明的方法，描述了图中的一些大小关系以及对象信息等。Zhou 等人 [52] 设计了编码-解码框架来提取条形图信息，并加入注意力机制提高模型的精度和鲁棒性。对于数值图中最复杂的曲线图，Figureseer 工具 [47] 可以在提取数据大小的同时分析数值图上的内容，包括图例条目的关联信息等。

然而，以上方法只挖掘了图片本身的信息，都没有考虑到在科学文献中与图片有关的文本中的信息。这些文字往往会对图片进行详细的描述，可能包含着图片上未描述的重要信息，如图片的含义和数值图坐标轴的名称等。在科学文献中，图表与对应的描述文字相辅相成，帮助读者更快地理解，把这些信息有机结合将会获得更丰富的信息，提升文献挖掘的全面性和多元性。

1.4.3 文本与图表联合挖掘研究现状

部分学者兼顾了科学文献中文本与图表两种不同形式的数据，将两者提取的信息进行了整合。Jiang 等人 [53] 从科学文献中的表格中按行列关系抽取了元素，并对文本内容进行命名实体识别，最终结合两者获得了高温合金的化学成分和特性数据，最后通过这些数据对合金进行分析和预测。Jensen 等人 [54] 自动提炼了科学文献中的表格和正文文本，得到了沸石合成和拓扑数据，挖掘含锗沸石的潜在机会。从表格和文本两种类型的数据中提取的信息会比仅从单一数据中提取的信息更加全面，从而提高数据的可用性和健壮性。然而，这两项工作仅针对 XML 或 HTML 格式的科学文献，不具有普适性。此外，他们解析表格的方法也较为简单，仅依靠 XML

或 HTML 的结构提取表格中的每一行文本，并将对应的行列数据进行抽取。Safder 等人 [55] 实现了在 PDF 格式的科学文献中提取数值图和文本的信息。他们训练了卷积神经网络模型截取科学文献中的 AUC 结果图片，应用“梯形法则”提取 AUC 图片的具体数值结果。同时，他们通过转换 PDF 文件获得了全文文本，然后匹配出每一幅结果图的标题文本。然而，他们仅针对 AUC 这一种数值图片，并且没有进一步提炼标题文本中的信息，只是展示出了标题。

这些文献挖掘方法兼顾了科学文献中不同的类型数据，抓住了数据间的联系从而提升了挖掘信息的全面性，推动了文献挖掘多元化的发展。但他们的方法普适性和迁移性较差，且 Safder 等人对于文本信息没有进一步地处理，削弱了信息的可读性。因此在联合挖掘不同类型数据的基础上，如何加强挖掘方法的普适性和挖掘信息的可读性成为了研究重点。

1.5 论文主要工作

为了实现能够结合科学文献中多种类型的数据进行文献挖掘的目标，本文提出了一种基于科学文献中单一数值图图片和文本内容的联合挖掘方法。该方法能够从材料科学文献中截取单一数值图的图像、标题、坐标轴信息和数值信息。本文的主要工作和创新点包括：

（1）设计并实现从材料科学文献中提取单一数值图信息的方法。该方法采用计算机视觉技术与文本挖掘技术相结合的方式，首先使用 Yolov5s 模型对科学文献中单一的数值图图片进行检测和截取，并应用改进的科学文献图片检测方法来提升截取的准确性。然后使用文本相似度匹配算法寻找对应的数值图标题，并结合 Sci-Bert 与 CRF 模型来识别标题中数值图的坐标轴名称实体。最后结合图像处理技术设计了数值图数据提取算法，结合坐标轴名称获取更全面的数值图信息。该方法相比于其他方法，在材料文献挖掘的数据多样性上有了突破。

（2）设计基于图文的提升坐标轴实体识别效果的方法。该方法利用科学文献中数值图图片与文本内容之间的联系，首先识别出数值图图上的坐标轴标签文本，并将其填充入设定好的实体模板，生成无需标注的文本数据以达到数据集扩充的目的。

然后使用文本相似度匹配算法寻找在正文中与数值图相关的描述语句，将其作为扩充文本来改变标题文本的上下文表征，从而提升命名实体识别的准确性。利用该方法，模型在识别坐标轴名称时的效果有所提升，缓解了数据集规模小导致的模型精度下降的问题。

（3）基于前两个算法开发材料科学文献数值图信息提取软件。该软件可在本地执行，用户在无联网、无额外运行环境的情况下，即可进行大规模的文献挖掘任务，获取材料科学文献中数值图的数据、坐标轴名称等信息。该软件降低了材料学者收集科学文献数据的人力成本，推动了材料领域的发展。

1.6 论文组织结构

本论文以作者在攻读硕士研究生阶段的成果为基础，针对以上提出的问题和材料科学文献的特点，研究和设计了从材料科学文献中提取数值图的图文信息的方法和技术，实现了材料文献挖掘的功能，并通过实验验证了本论文提出方法的有效性。

本论文的其他各章内容安排如下：

第二章简要介绍本研究涉及的相关理论基础知识，首先介绍了目标检测的相关概念及卷积神经网络的相关知识，并阐述了本研究在截取单一数值图图片任务中使用到的 Yolov5 网络。然后解释了文本挖掘任务中的命名实体识别和文本匹配的概念，介绍了在计算机中文本数据的表示方法以及在文本挖掘中使用到的 Bert 预训练模型和 CRF 算法。最后罗列了本论文在实验中使用到的模型评价指标。

第三章紧紧围绕材料科学文献的特性，提出了一种基于科学文献中的图片与文本两种不同类型的数据，挖掘科学文献中数值图信息的算法。首先在 PDF 格式的文献中截取出单一的数值图图片，并匹配其对应的图片标题。然后结合图片和标题两种不同类型的数据，挖掘出两者中的关键信息并结合。最后通过实验验证了本章方法的有效性。

第四章针对第三章方法中的命名实体识别任务数据集少导致的坐标轴实体识别精度低的问题，提出了一种基于图文的提升识别准确性的方法。首先通过识别图上的标签文本，以模板的方式扩充数据集。然后寻找正文中对应的数值图描述语句，将

其与标题文本进行拼接送入模型以生成更准确的向量表征，从而提升模型的识别效果。最后通过实验对比说明了本章方法的有效性。

第五章介绍了材料文献数值图信息提取可视化软件，通过开发环境、需求分析、架构设计和软件实现四个方面介绍了软件的结构和使用流程。

第六章对全文进行了总结和回顾，并对未来的研究方向提出了一定的展望和设想。

第二章　相关理论与技术概述

计算机视觉和自然语言处理技术的快速发展为材料文献挖掘的研究开辟了新的可能性。本研究基于现有的先进技术，针对材料科学文献的特征，提出了一种从材料科学文献中提取单一数值图信息的方法，实现了多源数据融合的材料文献挖掘框架。为了阐述本研究的理论基础和方法原理，本章将分别回顾与本研究相关的目标检测和文本挖掘技术，并介绍一些常用的算法和模型以及实验中使用的评价指标。

2.1　目标检测理论概述

近年来随着深度学习的发展，各领域学者发现使用深度学习模型在进行目标检测任务时总能获得令人满意的效果。大量神经网络模型不断涌现，出现了如 Faster R-CNN[56]、SSD[57]、Detr[58]、Yolo[59] 等经典网络。本节介绍了目标检测任务的概念，并详细说明了经典的卷积神经网络（convolutional neural networks, CNN）模型架构和本论文使用到的 Yolov5s 网络。

2.1.1　目标检测

目标检测是计算机视觉领域的一个重要问题，其目的是在图像中识别和定位感兴趣的目标，给出它们的类别和位置信息。该问题通常涉及目标定位和识别检测两个子任务，即在图像中精确地划分目标的边界框，并判断目标属于哪个类别。目标检测技术在人脸识别、无人驾驶、遥感影像分析、医学图像检测、OCR 等领域有着广泛的应用。

目标检测技术基于深度学习的方法可以分为两大类：两阶段方法和单阶段方法。两阶段方法是较早的方法，其主要包括两个步骤：第一步是生成一些候选框（region proposals），第二步是对这些候选框进行分类和回归，R-CNN 系列网络就是典型的两阶段方法。单阶段方法不依赖于候选框，而是直接在整个图像上进行预测，这样可以显著降低计算量，提高速度，Yolo 系列网络就是代表性的单阶段方法。

目标检测任务中常用交并比（Intersection of Union, IoU）来评价预测结果的质

量，IoU 是指预测边框（Predicted bounding box）和实际边框（Ground-truth bounding box）的交集与并集的比值，其计算公式如式 2.1所示。

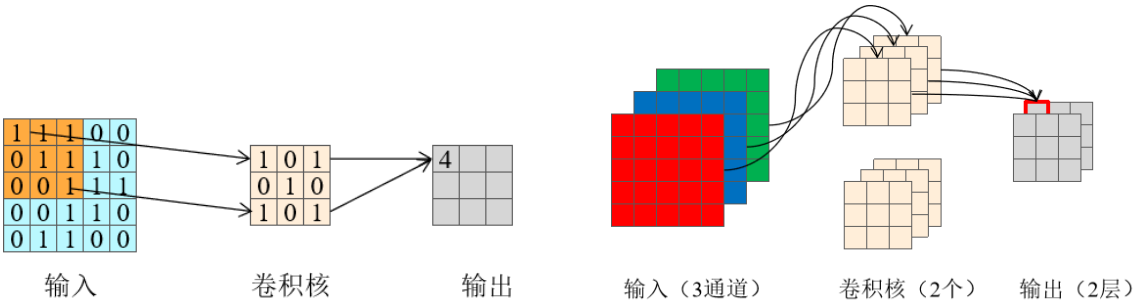
$$IoU = \frac{|S_P \cap S_G|}{|S_P \cup S_G|}, \tag{2.1}$$

2.1.2 卷积神经网络架构

卷积神经网络是最常用的深度学习模型之一。CNN 是一种结合了人工神经网络和深度学习的方法，通过反向传播算法优化参数，实现对图像的高层特征提取和分类。CNN 相比于传统的神经网络，不仅具有较好的容错性、自适应性和自学习能力，还具有自动提取特征、权值共享以及输入图像与网络结构匹配等优势。CNN 通常采用标准的反向传播算法进行训练，模型复杂度相对较低，训练效率较高，网络结构主要由卷积层、池化层、激活层和全连接层组成。

卷积是 CNN 最基本的核心操作，其作用是通过卷积核对输入进行特征提取，得到输出特征图。单通道的卷积的计算过程如图 2.1（a）所示，对于一个输入的图像或特征矩阵，卷积核以一定的步长在其上滑动，每一步都将卷积核和输入的对应区域进行逐元素相乘并求和，得到输出特征图的一个像素值。卷积核的通道数要与输入的通道数一致，对于三通道的输入矩阵，卷积核也是由三个二维矩阵叠加而成。卷积核的个数可以是多个，每一个卷积核可以得到输出特征图的一层。如图 2.1（b）所示，如果有 2 个卷积核，那么输出的特征图的深度就为 2。由于卷积操作会导致输出特征图的尺寸缩小，而且边缘区域的信息会被忽略，为了避免这种情况，可以在输入矩阵的边缘进行填充（padding）操作，即在边缘处补 0，这样可以保持输出特征图的尺寸不变，并且增加边缘区域的信息量。卷积层具有局部连接和权值共享两个重要的特性。局部连接是指在卷积操作时，相邻区域的信息会被融合在一起，使得模型可以捕捉到更高层次的抽象特征，并且保留了空间位置信息。权值共享是指在卷积操作时，卷积核内的权值是固定不变的，并且会与输入图像的所有通道进行计算，这样可以保证图像中每个像素都有一个权重系数，并且减少了模型的参数量。

池化是一种模拟人的视觉系统对数据进行降维的操作，其目的是用更简洁的特征表示图像。池化层的主要作用有：（1）对特征进行降维，减少参数量；（2）实现模



（a）单通道卷积过程　　（b）多通道卷积过程

图 2.1: 单通道和多通道的卷积计算过程

型的尺度不变性、旋转不变性、平移不变性；（3）扩大感知。常见的池化操作可分为四种，包括最大池化、平均池化、全局最大池化和全局平均池化。如图 2.2所示，图中最大池化和平均池化的步长均为 2，池化窗口大小均为 2*2。最大池化是从特征图中选取池化窗口中的最大值作为输出特征图的一个像素值，通过设定的步长遍历整个特征图。平均池化是从特征图中计算池化窗口内的平均值作为输出特征图的一个像素值。全局最大池化是从整个特征图中选取最大值作为输出，全局平均池化是从整个特征图中计算平均值作为输出，通常用于代替分类器中的全连接层或密集连接层。

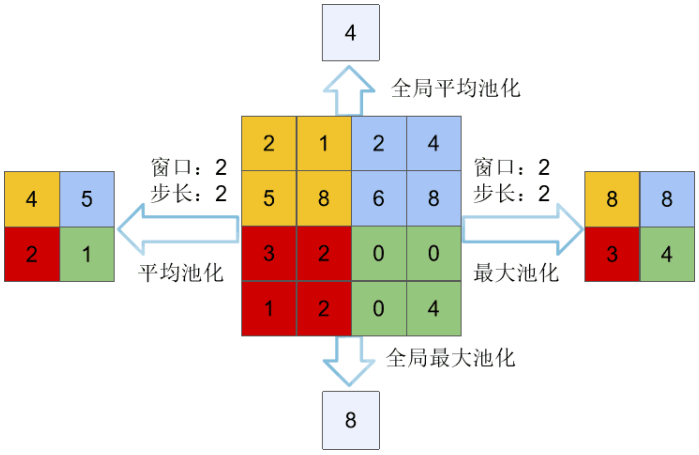
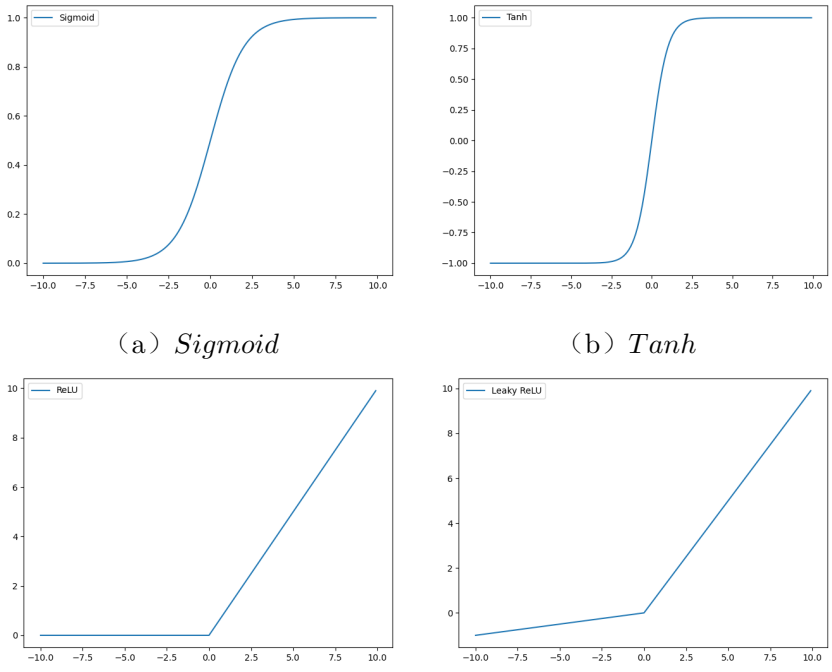


图 2.2: 四种池化操作

激活层是一种引入非线性变换的操作，其作用是增强网络的非线性拟合能力。如果只使用卷积池化等线性操作，那么网络就无法描述数据间的非线性特征关系，

因此需要激活函数来增加网络的表达能力。如图 2.3所示，网络中常用的激活函数有 *Sigmoid*、*Tanh*、*ReLU* 等。*Sigmoid* 函数计算公式如式 2.2所示，该函数可以将任意输入映射到区间 [0,1] 上，因此它适合用于输出预测概率的模型，例如二分类问题。*Tanh* 函数计算公式如式 2.3所示，该函数可以将任意输入映射到区间 [-1,1] 上，在输入接近 0 时变化较快，在两端时变化较慢。这两个函数都存在梯度消失问题，即在输入过大或过小时，梯度趋近于 0，导致网络难以训练。*ReLU* 函数有效地解决了这个问题，如式 2.4所示，该函数在输入大于 0 时保持梯度为 1，在输入小于 0 时将输出置为 0，从而增加网络的稀疏性。但是，*ReLU* 函数也存在一个缺点，就是当输入小于 0 时，神经元会完全失活（dead），导致权重无法更新。为了解决这个问题，*Leaky ReLU* 函数被提出，如式 2.5所示，该函数在输入小于 0 时不再输出 0，而是输出一个很小的值 a 乘以输入，从而避免神经元死亡。



（a）*Sigmoid*　（b）*Tanh*

（c）*ReLU*　（d）*Leaky ReLU*

图 2.3: 不同的激活函数

$$Sigmoid(x) = \frac{1}{1+e^{-x}}, \tag{2.2}$$

$$Tanh(x) = \frac{e^{x} - e^{-x}}{e^{\dot{x}} + e^{-x}}, \tag{2.3}$$

$$ReLU(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases}, \tag{2.4}$$

$$Leaky\ ReLU(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{if } x < 0 \end{cases}, \tag{2.5}$$

全连接层通常位于 CNN 的输出端，其作用是将提取到的高层特征表示映射为样本的类别标签，起到分类的作用。全连接层的结构如图 2.4所示，假设一个 n 维的特征向量 $X = (x_1, x_2, ..., x_n)$，其每一元素 x_j 都会与多个神经元相连，并进行加权求和和激活操作，最终得到一个 t 维的输出向量 $C = (c_1, ..., c_t)$，其中 t 为任务的类别数，每一个值 c_i 表示该类别的预测概率，概率最高的值对应的类别就是最终的预测结果，从而完成分类任务。

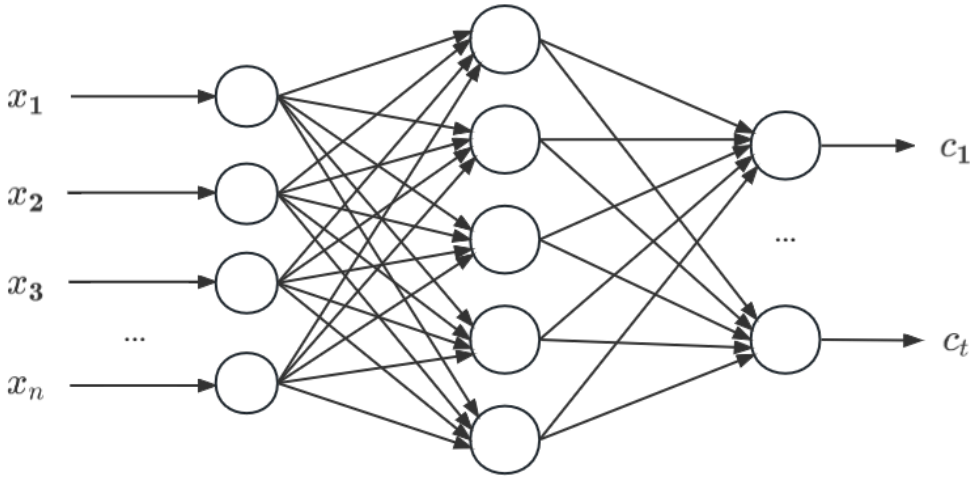


图 2.4: 全连接层

2.1.3 Yolov5 网络架构

Yolo（You only look once）[59] 是一种基于单阶段方法的经典目标检测模型，而 Yolov5 是其第五代版本。Yolov5 相较于之前的网络有着更高的精度，且图像推理速度最快可达 0.007s。因此，该网络可以在保证检测精度的同时，拥有较快的检测速度，满足实际任务的需求。Yolov5 的网络架构如图 2.5所示，主要包括 Backbone，Neck 和 Prediction 三个部分。其中 Backbone 是模型的特征提取部分，主要用来提取图像的特征；Neck 负责特征融合，融合从 Backbone 获得的不同尺度的图像特征；而 Prediction 为模型的最终输出，输出网络预测的目标信息。

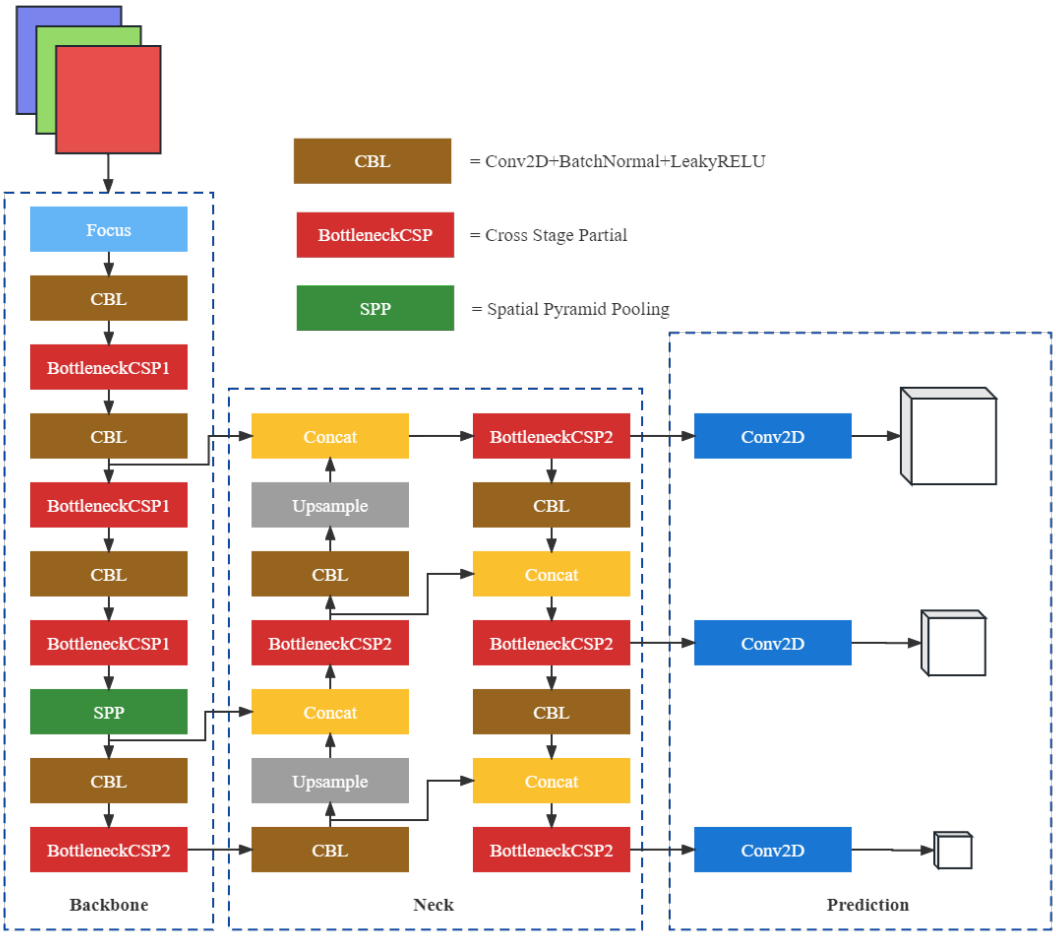


图 2.5: Yolov5 架构图

Backbone 主要由 Focus 模块和 CBL 模块组成。Focus 模块的作用是将输入图

像切分为四个子图像，并将它们拼接在一起。如图 2.6所示，把高分辨率的特征图隔行、隔列进行采样，拆分成多个低分辨率的特征图，再进行拼接，这样可以减少输入图像的尺寸，降低网络参数量，提高处理速度。CBL 模块是一个包含卷积、归一化和激活操作的基本单元。BottleneckCSP 模块主要是利用残差的思想，如图 2.7（a）所示，通过在卷积层的输入和输出之间添加跳跃链接的操作从而实现特征融合。输入 x 经过两个或多个卷积层得到输出 $f(x)$，与 x 进行相加运算得到最终输出 $H(x)$，使得网络在反向传播时，梯度可以直接传递到更前面的层。CSP 模块将输入的特征进行卷积和残差操作，将两者输出的特征进行拼接，最后经过归一化、激活、CBL 操作，从而能够在保留细粒度特征的同时防止网络退化。在 Yolov5 中有两种 CSP 模块，如图2.7（b）所示，通过调整该模块的深度，可以得到不同大小的网络。表 2.1对比了四种不同网络模型的大小，其中 CSP1 模块按照图2.5中从上到下排序，从对比结果可以看出 Yolov5s 中的所有 CSP 模块都是最浅的，因此参数量相较于另外 3 个网络少了数倍。Spatial pyramid pooling(SPP) 模块也被称为空间金字塔池化，如图 2.8（a）所示，其作用是通过使用不同大小的池化窗口来提取特征，并将提取后的特征进行拼接，其目的在于对于任意尺寸的输入生成固定大小的输出。而 Yolov5 中的 SPP 模块使用的是最大池化，如图 2.8（b）所示，将经过 CBL 操作后得到的特征输入使用不同大小的最大池化层计算每个窗口的最大值，然后将池化前后的特征进行拼接并输入到 CBL 模块，从而增加网络的感受野。

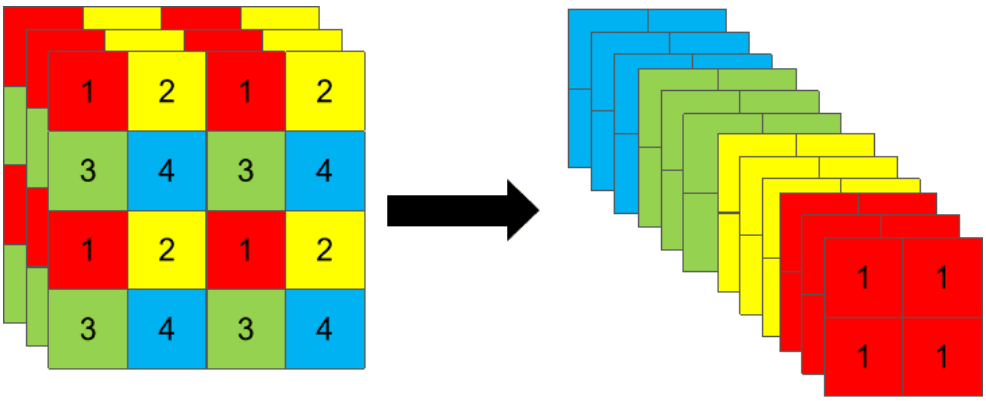
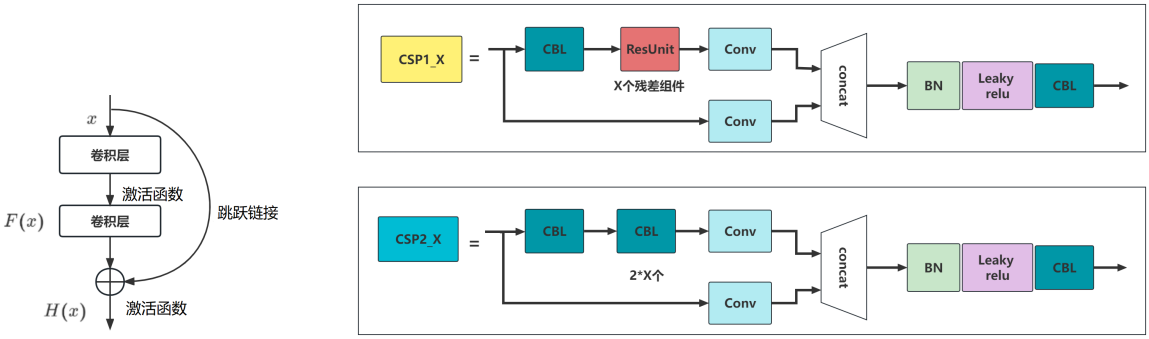


图 2.6: Focus 操作流程

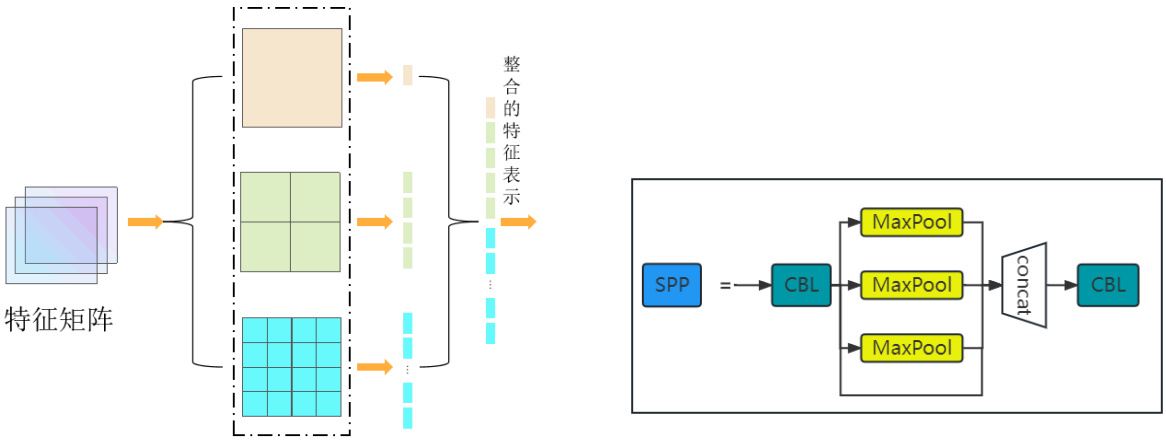


（a）残差思想　　（b）Yolov5 中两种不同结构的 CSP 模块

图 2.7: 残差思想及 Yolov5 网络中的实现方法

表 2.1: Yolov5 不同网络的大小对比，CSP1 模块按照图2.5中从上到下排序。

模型名称 / 模块	Yolov5s	Yolov5m	Yolov5l	Yolov5x
第一个 CSP1	CSP1_1	CSP1_2	CSP1_3	CSP1_4
第二个 CSP1	CSP1_3	CSP1_6	CSP1_9	CSP1_12
第三个 CSP1	CSP1_3	CSP1_6	CSP1_9	CSP1_12
所有 CSP2	CSP2_1	CSP2_2	CSP2_3	CSP2_4
总参数量 (M)	7.2	21.2	46.5	86.7



（a）空间金字塔池化思想　　（b）Yolov5 中的 SPP 模块结构

图 2.8: 空间金字塔池化思想及 Yolov5 中的实现方法

Neck 主要借鉴了 feature pyramid networks（FPN）[60] 以及 path aggregation network（PAN）[61] 的思想。如图2.9所示，FPN 使用一种自顶向下的方式，在不同尺度上生成具有高级语义的特征图，通过上采样的方式将高层的小特征图放大，并与低层的大特征图进行融合，这样既保留了高层的强语义特征，又利用了低层的高分辨率信息。但这种方法只增强了语义信息，而没有传递定位信息。PAN 在此基础上增加了一条自底向上的路径，利用下采样的方式将低层的定位特征传递到高层。

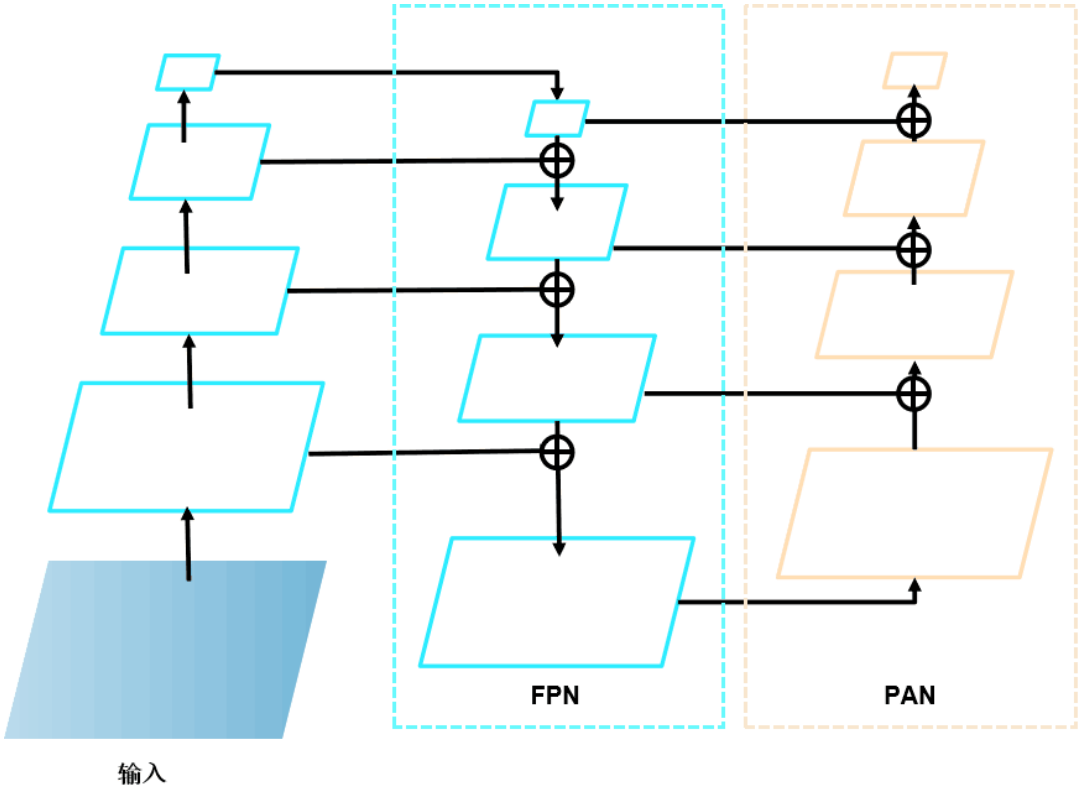


图 2.9: FPN 和 PAN 思想

在 Prediction 阶段会得到不同尺度下的输出特征，包含预测目标的类别、置信度以及边界框的坐标。在确定最终的检测结果时，首先会根据目标的置信度进行阈值的筛选，只保留高于阈值的候选框。此外，还需要解决单个目标有多个候选框重叠在一起的问题，Yolov5 使用非极大值抑制（Non-Maximum Suppression, NMS）算法解决此问题。具体步骤如下：首先将候选框按照置信度大小降序排列，选择置信度最高的候选框作为参考框，计算其他候选框与参考框的 IoU 值，并设定阈值，如

果两个框的 IoU 值大于阈值，则认为它们都检测到了同一个目标，因此删除置信度较低的候选框。对每个候选框重复此过程，最终确定具有高置信度且不重复的目标位置。

2.2 文本挖掘理论概述

文本挖掘是一种从未经处理的文本中提取有价值的信息和知识的过程，从而支持决策的效率和质量。文本挖掘的任务包括文本分类和聚类，命名实体识别，情感分析和文本匹配等 [62]。本节将介绍命名实体识别和文本相似度匹配任务的定义，并回顾一些文本挖掘的理论和技术，为后续章节的方法介绍做准备。

2.2.1 命名实体识别及文本匹配任务

命名实体识别（Named Entity Recognition，NER）是文本挖掘的核心任务之一，旨在从文本中识别出任务指定的实体。如图 2.10所示，在通用领域中，实体通常指人名、地名、机构名等专有名词，在特定领域如材料领域中，实体包括材料名称、工艺名称、工艺参数和材料性能等。NER 是信息提取、问答系统、知识图谱等领域的重要基础技术 [63]。NER 任务通常会被建模成序列标注任务，即模型的输入是一个待识别的文本序列，模型的输出是该文本序列对应的标签序列。

（a）公共领域文本的 NER　　（b）材料文本的 NER

图 2.10: NER 任务示例

文本相似度匹配任务是一种评估两个文本在语义上是否相似的任务，是自然语言处理的基础任务之一。它可以帮助学者理解文本的含义，以及文本之间的关系，从而为其他任务提供基础和支持 [64]。文本相似度匹配的任务输入是一对文本，例如两个句子，两个段落，或者一个问题和一个答案。任务输出是一个 [0,1] 区间内的小数，表示两个文本的相似程度，或者一个二元标签，表示两个文本是否相似。文本

相似度匹配任务可以应用于信息检索、问答系统、机器翻译等领域。

2.2.2 文本表示方法

计算机无法直接处理自然语言文本，而只能对数值形式的数据进行操作。因此，在文本挖掘任务中，需要先将文本转换为计算机可理解和处理的向量形式 [65]。按照转换方式的不同，文本表示可以分为离散式和分布式两种。离散式的文本表示将词典中的每个词或短语对应到一个高维的稀疏向量，如 one-hot 编码、词袋模型、TF-IDF 等。这种方法的优点是易于实现，能够反映词汇的统计特征，能够降低高频词汇的影响，突出特征词汇的作用。但离散式的文本表示也存在以下缺点：（1）维度过高：由于离散式的文本表示需要一个包含所有词汇的词典，而词典的规模往往很大，导致向量的维度过高，增加了计算复杂度和存储开销。（2）稀疏性：离散式的文本表示只有少数元素为非零值，而大部分元素为零值，导致向量的信息密度低，表达能力弱。（3）语义缺失：离散式的文本表示只考虑了词汇的词频或词权重信息，而忽略了词汇的语义和语法信息，导致向量无法反映语义相似度，无法处理词汇的多义性和一义多词性。

随着深度学习技术的发展，神经网络在文本表示方面的应用逐渐成熟，分布式表示就此取代了离散式表示。2013 年，Mikolov 等人 [66] 提出了 Word2Vec 语言模型，Word2Vec 使用一个两层的神经网络模型，从大规模的文本语料中学习词汇的向量表示。训练好的模型可以检测同义词或者为部分句子提供合适的词汇，Word2Vec 模型有两种主要形式：CBOW 和 Skip-Gram。如图 2.11所示，CBOW 模型是根据上下文词汇来预测中心词汇，即给定一个词汇的周围词汇，输出这个词汇的概率分布。而 Skip-gram 模型则是根据中心词汇来预测上下文词汇。例如给定一个句子“I drive my car to the store.”，如果窗口大小为 2，那么 CBOW 模型的一个训练样本可以是（context: [I, drive, car, the, store], center: my），而 Skip-gram 模型则为（center: my, context: [I, drive, car, the, store]）。CBOW 模型更适用小规模的数据集，它对高频词汇更加敏感，且计算速度会更快，而 Skip-gram 模型更适合大规模的数据集，且对低频词汇更加敏感，表达更精确。

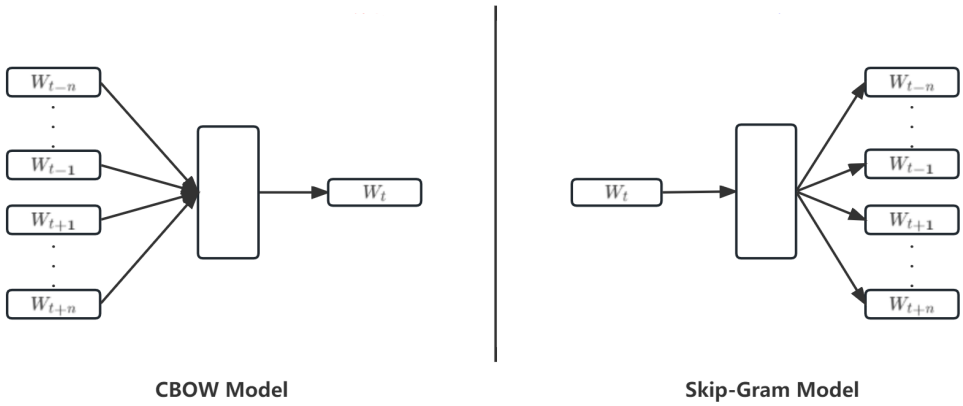


图 2.11: Word2Vec 模型

Word2Vec 的提出，推动了文本挖掘技术的发展，静态词向量模型也成为文本表示的主流。在 Word2Vec 的基础上，一些新的词向量模型也被提出，以解决不同的问题。例如 GloVe[67] 是一种全局词向量模型，它不仅利用了词语的局部上下文信息，还利用了词语的全局共现信息，即词语在整个语料库中出现的频率和关系。FastText[68] 是一种子词向量模型，它不仅考虑了词语的整体表示，还考虑了词语的子词（n-gram）表示，可以处理一些生僻词和拼写错误的词语。Mat2vec[69] 是一种专门用在材料科学领域的词向量模型，它可以从材料科学文献中提取词语的语义和语法信息，以及识别词语与材料组成和性质之间的关系。

静态词向量模型虽然在自然语言处理领域取得了一定的成果，但也面临着如下的问题 [70]：（1）无法表达单词的多义性，即一个单词不管上下文如何变化，该单词的向量表示都是同一的。（2）无法捕捉单词之间的复杂关系，例如同义词、反义词、上下位关系等。（3）无法利用预训练模型的优势，即在大规模语料库上学习通用的语言知识，然后在特定任务上进行微调。为了克服这些问题，动态词向量模型应运而生。动态词向量模型能够根据上下文动态适应性地调整文本的向量表示，在一定程度上解决了单词多义性的问题，并且利用预训练模型的优势可以提高自然语言处理任务的性能。预训练模型的出现将自然语言处理带入了一个全新时代，经典的预训练模型以 Bert 为代表。

2.2.3 Bert 预训练模型

Bert 是一种基于 Transformer[71] 的预训练语言模型，它使用了掩码语言模型和下一句预测模型两种任务来学习文本的双向表示。Bert 在多个自然语言处理任务中取得了显著的性能提升，从而使得动态词向量模型成为了文本表示的首选模型。Bert 的结构如图 2.12所示，输入文本首先经过分词操作转为单词序列，然后根据预训练好的词汇表查找每个子词在表中的位置作为单词索引，将索引序列送入嵌入层得到三种嵌入向量：词向量（Token embeddings）、句子向量（Segment embeddings）以及位置向量（Position embedding），并将它们相加得到最终的输入向量，以获得更丰富的特征。

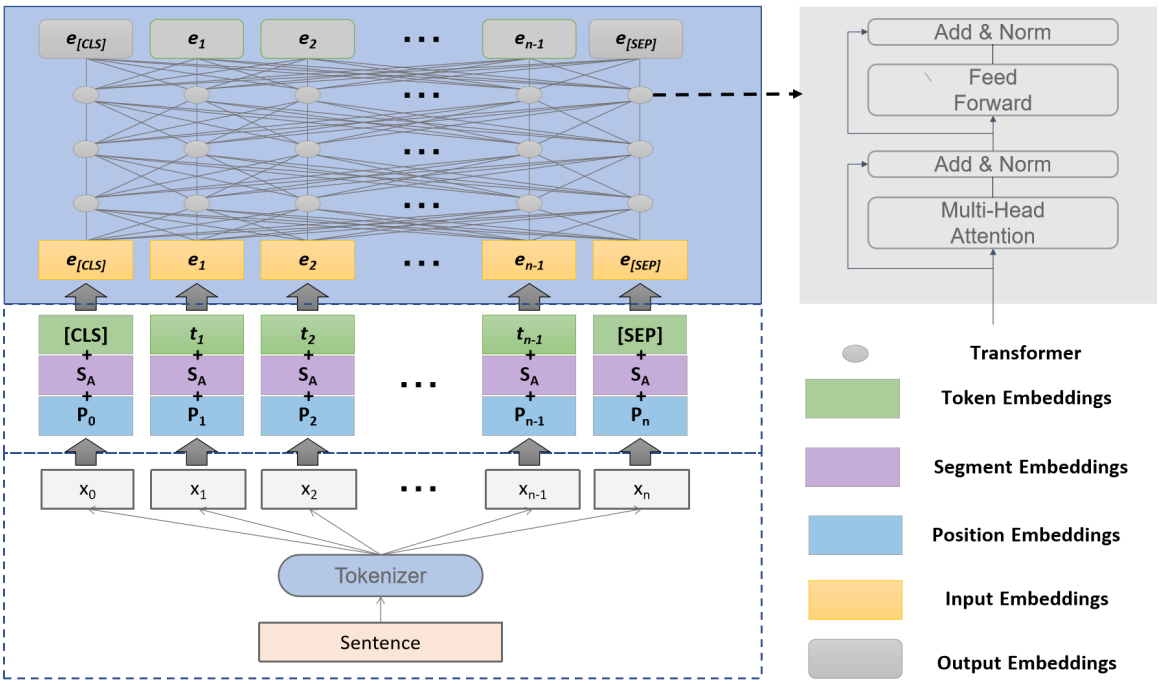


图 2.12: Bert 模型结构。其中“[CLS]”和“[SEP]”分别标识一句话的开头和结尾。

Bert 内部核心是一个多层的双向 Transformer 的 Encoder 部分。Encoder 由若干个相同的子层组成，每个子层包含两个部分：多头自注意力机制和全连接前馈网络。多头自注意力机制是将单个自注意力机制分解为多个子空间，并在每个子空间上执行自注意力机制的一种方法，它能够更好地捕捉不同层次和不同角度特征。它

的基本思想是将输入的 Q、K、V 分别映射到 h 个低维空间，然后在每个低维空间上计算缩放点积注意力，将最后结果拼接起来并通过一个线性变换得到输出。多头注意力机制可以增加模型的多通道性能，让模型在不同的表示子空间里学习到相关的信息，让每个单词在编码时考虑到其他单词的信息，捕捉序列中的依赖关系，从而提高模型的表达能力和泛化能力。全连接前馈网络是由两个线性变换和一个激活函数组成，通过对每个位置上的特征进行进一步编码以提取更高层次的语义信息，在增加模型的非线性能力的同时提高模型的表达能力。每个子层后面都有一个残差连接和一个层归一化操作，以保证梯度流动和信息传递。残差连接可以让输入直接加到输出上，从而避免梯度消失或爆炸，层归一化可以对每个隐藏状态进行规范化，从而加速收敛和稳定训练。

输入语句经过多层的 Transformer 之后会得到句子内每个单词的语句表征，根据研究者的任务对向量进行微调即可获取最终结果。比如 NER 任务，对于一个长度为 $n+1$ 的输入语句 $X=(x_0,x_1,...,x_n)$，Bert 会输出一个同样长度的语句向量 $E=(e_0,e_1,...,e_n)$，利用激活函数将每一维的编码映射到标记空间上，即 $\mathbb{R}^E \mapsto \mathbb{R}^K$，其中 K 是标记的数量，取决于标签的数量和标记方案，而每一维的向量值即为每个单词对应实体标签的概率。

2.2.4 CRF 算法

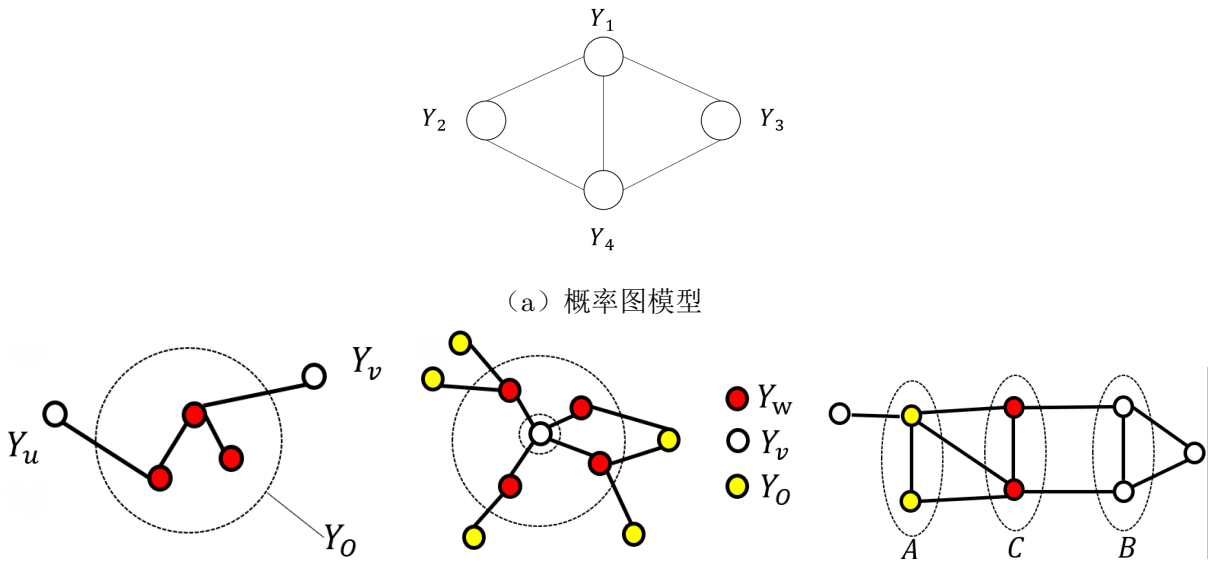
条件随机场（Conditional Random Field, CRF）[72] 是一种基于概率图模型的序列标注模型，它假设给定输入序列时，输出序列的联合概率分布满足马尔科夫性质，即每个输出变量只依赖于它相邻的变量。如图 2.13（a）所示，CRF 可以用无向有环图来表示，其中节点表示随机变量，边表示变量之间的依赖关系。在概率图模型中，存在三种马尔科夫性质：（1）成对马尔科夫性。如图 2.13（b）所示，如果图中没有边相连的两个节点 u 和 v 对应的变量 Y_u 和 Y_v 在给定其他所有节点 O 对应的变量集合 Y_O 的条件下是独立的，则称满足成对马尔科夫性，其表达如式2.6所示。（2）局部马尔科夫性。如图 2.13（c）所示，如果与图中任意节点 v 相连的节点集合 W 对应的变量集合 Y_W 与网络中剩余节点集合 O 对应的变量集合 Y_O 在给

定节点 v 对应的变量 Y_v 的条件下是独立的，则称满足局部马尔科夫性，其表达如式2.7所示。（3）全局马尔科夫性。如图 2.13（d）所示，如果网络中任意被节点集合 C 分开的两个节点集合 A 和 B 对应的变量集合 Y_A 和 Y_B 在给定节点集合 C 对应的变量集合 Y_C 的条件下是独立的，则称满足全局马尔科夫性，其表达如式2.8所示。如果一个无向图满足任一马尔科夫性质，则称该无向图表示了一个马尔科夫随机场，即条件随机场。

$$P\left(Y_u, Y_v \mid Y_O\right) = P\left(Y_u \mid Y_O\right) P\left(Y_v \mid Y_O\right) , \tag{2.6}$$

$$P\left(Y_v, Y_O \mid Y_W\right) = P\left(Y_v \mid Y_W\right) P\left(Y_o \mid Y_W\right) , \tag{2.7}$$

$$P\left(Y_A, Y_B \mid Y_C\right) = P\left(Y_A \mid Y_C\right) P\left(Y_B \mid Y_C\right) , \tag{2.8}$$



（a）概率图模型

（b）成对马尔科夫性　（c）局部马尔科夫性　（d）全局马尔科夫性

图 2.13: 概率图模型和马尔科夫性质介绍

条件随机场可以利用全局信息和复杂特征来建模输出序列的条件概率分布，而不是像传统的分类器那样只考虑局部信息和独立假设，从而能够捕捉特征之间的依赖关系，更有效地利用上下文信息，提高序列预测任务的性能和准确度。条件随机场

有多种形式，其中最常用的是线性链条件随机场，它适用于序列标注问题，如 NER、词性标注等。线性链条件随机场是一个对数线性模型，它使用了一系列特征函数和权重参数来定义输出序列的概率分布。线性链条件随机场的学习方法可以采用极大似然估计或正则化的极大似然估计，优化目标是最大化对数似然函数或加入正则项后的对数似然函数。线性链条件随机场的推断方法通常是维特比算法（Viterbi Algorithm），它能够找到使对数似然函数达到最大值时的标签序列作为最终的预测结果。

2.3 评价指标概述

评价指标是用来度量不同模型在特定任务中的性能和效果的，它们能够为模型提供客观、量化、可比较的性能衡量标准，同时也为深度学习模型的应用和推广提供一个可信的依据和参考。针对分类任务，可以使用 N 和 P 分别表示实际结果中的负类和正类，使用 T 和 F 分别表示模型预测结果与实际结果相符和不符的情况，将其两两组合可以得到如表 2.2所示的混淆矩阵。

表 2.2: 混淆矩阵

	标签为真	标签为假
预测为真	TP	TN
预测为假	FP	FN

基于混淆矩阵，可以定义如下三种常用的评价指标：

（1）精确率（Precision）：

$$Precision = \frac{TP}{TP + FP}, \tag{2.9}$$

精确率表示预测结果中真正属于正类样本占所有预测为正类样本的比例，模型的精确率高说明该模型对正类的判别能力越强。例如，一个模型在所有样本中预测出了 100 只猫，其中 90 只是真正的猫，那么该模型的精确率就是 90%。

（2）召回率（Recall）：

$$Recall = \frac{TP}{TP + FN}, \tag{2.10}$$

召回率代表预测结果中真正属于正类样本占所有真正属于正类样本的比例，模型的召回率高说明该模型对正类的覆盖能力越强。例如，如果一个模型在所有样本中预测出了 100 只猫，其中 90 只是真正的猫，但实际上有 200 只猫，那么它的召回率就是 45%。

（3）F1 分数（F1-score）：

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{2.11}$$

F1 分数是一种结合精确率和召回率的综合评价指标。它可以兼顾精确率和召回率，其值越高，说明模型越优秀。

此外，为了能够对 NER 模型性能做出全面的评价，本论文采用了宏平均（macro-averaging）评测方法分别计算三种基准指标的平均值，得到 $Macro-P$、$Macro-R$ 和 $Macro-F1$。宏平均评测方法的计算公式如下：

$$Macro - P = \frac{\sum_{i=1}^{n} P_i}{n}, \tag{2.12}$$

$$Macro - R = \frac{\sum_{i=1}^{n} R_i}{n}, \tag{2.13}$$

$$Macro - F1 = \frac{2 * Macro - P * Macro - R}{Macro - P + Macro - R}, \tag{2.14}$$

其中 n 为命名实体类别数量。

本论文还使用了模型参数量作为评判模型优劣的指标。神经网络的参数量直接反映了模型的复杂度，直接影响模型推断时间和性能。这项性能关系到模型是否适用于实际场景，关乎算法的应用和推广。在相同准确率下，如果模型参数量越小，那么模型效果越优秀。

2.4 本章小结

本章首先介绍了目标检测的相关理论，解释了目标检测中常用的卷积神经网络的基本原理，并对本论文中使用的 Yolov5 网络结构进行了详细说明，阐述了其中的

关键技术。然后对文本挖掘任务作了定义，包括本论文涉及的 NER 和文本匹配任务，介绍了文本这一数据类型的表示方法，并对本论文使用到的 Bert 预训练模型和 CRF 算法进行了深入说明。最后对本论文中使用的相关评价指标进行了描述。

第三章　材料科学文献中数值图图文信息提取方法

近年来，随着深度学习研究的发展，神经网络在材料文献挖掘中所展现出来的优势日益显现。但由于材料科学文献中图像和文本两种数据类型的差异较大，导致材料文献挖掘研究通常只关注单一类型的数据，这限制了材料文献挖掘的进步。本章针对材料科学文献中单一数值图图片与其对应的图片标题语句这两种不同类型的数据，提出了一种基于深度学习和图像处理的材料科学文献数值图图文信息提取的方法。通过这个方法，融合材料科学文献中图片与文本两种不同的数据，挖掘出单一数值图包含的数值信息和坐标轴实体信息，开辟了多元化材料文献挖掘的新道路，并为材料学者大规模提取材料科学文献中数值图全面信息提供了帮助，推动新兴材料的发展。

3.1　方法概述

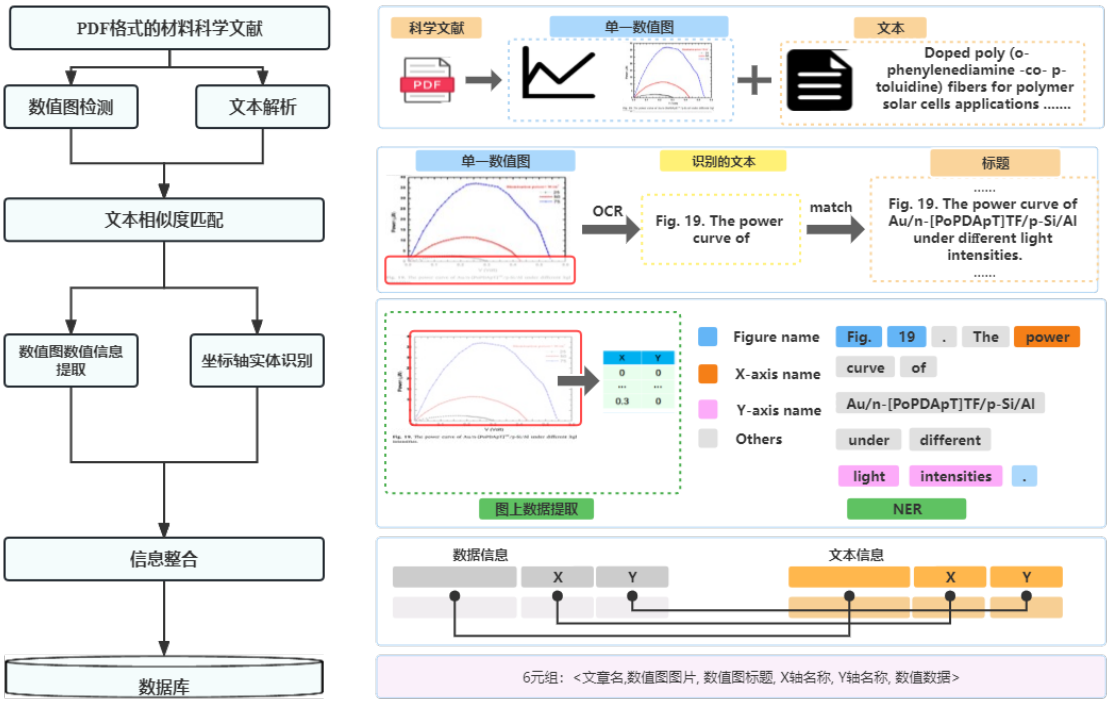


图 3.1: 自动化提取材料科学文献中数值图图片以及标题文本信息的流程

本方法的完整流程如图 3.1所示，首先提取 PDF 格式的材料科学文献中的文本内容，训练深度学习模型在科学文献中检测并截取数值图图片；然后使用文本相似度匹配技术在文本中匹配数值图所对应的图片标题；接着使用命名实体识别技术在标题中识别数值图所对应的 X 轴和 Y 轴实体名称；最后运用图像处理技术实现自动化地获得图上的数据信息，结合提取的实体名称生成全面的数值图信息。

3.1.1 文本和数值图图片的获取

本论文所挖掘的科学文献均为 PDF 格式，因为 PDF 格式是学术数据库中常见的文献格式，且可以同时获取图片和文本两种形式的数据。为了解析 PDF 格式文件中的文本，本文采用了 PDFminer 工具 [73]，它可以提取文本的位置、字体、颜色等信息，并把文本数据转化成 txt 格式。然而，由于格式排版和图片字符等干扰因素，通过 PDFminer 获取的 txt 格式的文本往往是不连贯和不完整的。因此，本文对提取出的文本内容进行了整合处理。具体步骤如下：（1）根据每一行的结束字符来连接句子，将末尾不是结束标识符（如句号、问号、感叹号）的语句与下一句进行拼接，并将一些截断的单词再次拼接起来。（2）若末尾文本为“-”时，将该字符删除，并将前后单词进行拼接。（3）去除一行中单词数很少的语句，这些一般都是无用的干扰字符，将这些字符去除可以加快后续查找文本的速度。

本论文还需要从科学文献中截取数值图图片，为了后续提取数值任务的需要，本文针对的数值图都是单一数值图，即每幅图片只包含一幅数值图，而不是多幅数值图拼接在一起。本论文使用了 Yolov5s 这一目标检测网络来截取数值图图片，它是 Yolov5 系列中参数量最少的网络模型，具有较高的检测精度和速度，适合大规模的任务。首先将 PDF 格式的科学文献按页保存为图片，然后利用 Yolov5s 模型在每一页上进行目标检测，得到预测的单一数值图候选框。

根据科学文献中图片摆放的特点，本文对单一的数值图检测方法做出了改进。科学文献中的图片下方往往会带有图片说明，即图片标题，标题文本中均带有关键标识符（如“Fig.”，“Figure”，“图”等），因此使用 paddle OCR 识别候选框上的文本，并根据正则表达式匹配标识符。若匹配成功则认为截取的数值图是正确的，

若匹配失败则扩展候选框的下边界再次进行匹配。这样既能保证图片截取的正确性，又能帮助匹配数值图所对应的标题，方便后续任务进行。另外，由于图片标题都是在数值图的下方，在使用 paddle OCR 时仅对候选框的下半区域进行识别，这样既加快了识别的效率，又增加了 OCR 的准确性。最后，为了避免候选框过窄而影响图片标题的识别以及最后的数据提取结果，对候选框的宽度 w_{box} 进行校验修改。若 w_{box} 小于 1/2 倍的 PDF 论文的页宽度，则左右拓宽候选框，使 w_{box} 达到 1/2 倍的论文页宽度；若 w_{box} 大于等于 1/2 倍的 PDF 论文的页宽度，则将 w_{box} 扩大为论文页宽度。整体的算法流程如算法 1所示。

算法 1: 科学文献中图片检测算法

Input: 文献页图片 P; Yolov5s 网络预测出的候选框集合 C_{box}; 关键标识符集合 C_f;

Output: 单一数值图图片;

```
获取文献页图片的宽度 w_P;
for i ∈ C_box do
    if w_i < 1/2 w_P then
        set w_box = 1/2 w_P;
    else
        set w_box = w_P;
    end
    选定新的候选框的下半部分 i_bottom;
    在 i_bottom 上使用 paddle ocr 进行识别获得文本集合 C_ocr;
    if C_f ∩ C_ocr ≠ ∅ then
        截取数值图图片;
    else
        扩展候选框的下边界，使得 h_box 变大至 1.1 倍;
        将新的候选框元素加入 C_box;
    end
end
```

3.1.2 数值图标题文本的匹配

为了提取单一数值图的坐标轴名称信息，本文除了需要截取数值图图片，还需要从文本中找到数值图对应的图片标题。虽然在上一节的数值图图片截取过程中，确保了截取的图片上带有其所对应的图片标题，但无法保证该标题的完整性，且无法保证利用 OCR 识别出的文本准确率。因此，本文还需要在 PDFminer 解析出的文本中找到该图片中的标题文本。首先利用 OCR 识别出图片上的图片标题语句，它可能是不准确、不完整的。然后计算该语句和文本中的所有语句的相似度，并认为相似度最高的语句即是数值图所对应的准确图片标题。

本论文使用 Sentence-Bert 网络计算语义相似度，图 3.2比较了其与 Bert 网络在计算语句相似度时的差异。Bert 在计算句子间的相似度时需要先将语句拼接，然后输入模型生成编码向量，最后通过激活函数获得相似度，这种方式对长文本很不利，拼接之后的长文本送入 Bert 会导致丢失与任务相关的全局信息。而 Sentence-Bert 借用了孪生网络的思想，利用两个参数共享的 Bert 得到语句的向量表示，再利用池化操作使得两个向量的维度统一，最后计算两个向量的相似度。Sentence-Bert 只需对每个句子进行一次计算就能够获得该句的最终语句表征，而不需要每次输入模型进行计算。此外，使用两个 Bert 在一定程度上避免了长语句的信息丢失。通过计算两个表征向量的余弦相似度来获得语句相似度，余弦相似度可以从方向上区分两个向量的差异，而对绝对数值不敏感，这样可以避免因为向量的长度不同而导致距离偏大。余弦相似度也可以快速地计算高维数据的相似度，节省计算资源和时间。对于任意两个语句 S_i 和 S_j，相似度的计算方法如下：

$$u = f(S_i), v = f(S_j) \ , \tag{3.1}$$

$$cos(S_i, S_j) = \frac{u \cdot v}{|u| \times |v|} = \frac{\sum_{t=1}^{n} (u_t \times v_t)}{\sqrt{\sum_{t=1}^{n} (u_t)^2} \times \sqrt{\sum_{t=1}^{n} (v_t)^2}} \ , \tag{3.2}$$

其中 u 和 v 分别是语句 S_i 和 S_j 经过 Bert 和池化操作后得到的向量表征，$f(\bullet)$ 表示模型的输出，u_t 和 v_t 表示 u 和 v 的每一维的数值。$cos(S_i, S_j)$ 表示两个语句的余弦相似度，余弦相似度值越接近 1，就表明两个语句越相似。

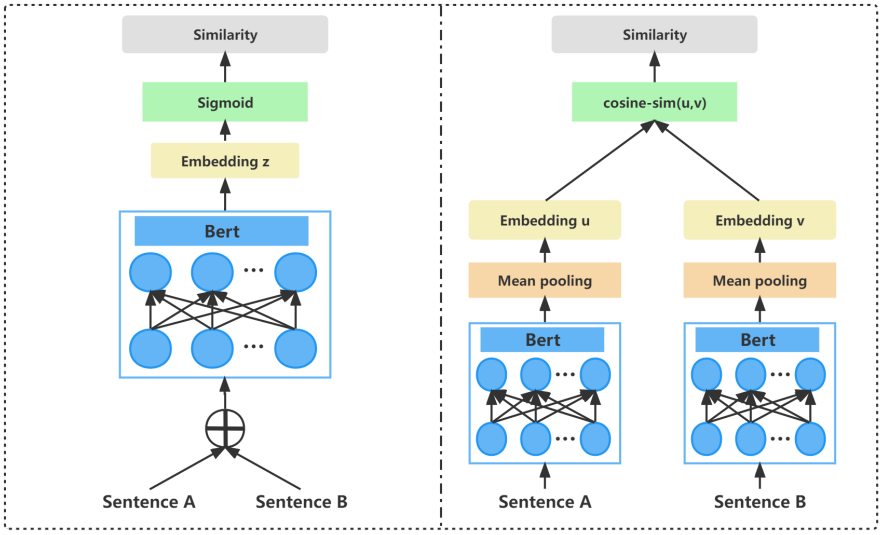


图 3.2: Bert（左）和 Sentence-Bert（右）之间文本匹配的差异

为了提高材料文本的表征效果，本文使用了 Sci-Bert 来代替 Sentence-Bert 中的两个 Bert 来生成语句向量。Sci-Bert 是一个专门针对科学文本进行二次训练的预训练模型，它使用了 114 万篇科学文献作为语料库，来增强科学领域的专有名词和术语的表示能力。相比于 Bert 模型，Sci-Bert 可以更准确地表征科学文本中的单词，而不会将其拆分为语料库中拥有的词根单元。例如，对于材料领域的单词“protease”，由于该单词未出现在 Bert 的语料库中，Bert 在分词时会按照词库中存在的词根将该单词拆分为“pro”、“##te”和“##ase”，而 Sci-Bert 的词库中存在该单词，在分词时会将其完整保留。因此，使用 Sci-Bert 模型来编码语句向量可以避免语料迁移的问题，提高材料文本生成向量的准确性。

考虑到本任务只是为了在文本中寻找对应数值图的标题文本，该文本与图上识别出的文本有着较大的重复率。因此本文加入了 Jaccard 相似度 [74] 来衡量两个语句的重复单词数量。Jaccard 相似度通过计算两个语句的单词集合的交并比来评价相似性。对于两个语句 S_i 和 S_j，它们的 Jaccard 相似度 $J(S_i, S_j)$ 计算方式如下：

$$J\left(S_i, S_j\right) = \frac{\left|W_{S_i} \cap W_{S_j}\right|}{\left|W_{S_i} \cup W_{S_j}\right|}, \tag{3.3}$$

其中 W_{S_i} 和 W_{S_j} 分别表示语句 S_i 和 S_j 的单词集合。$|\bullet|$ 代表统计集合内的元素

数量。

Jaccard 相似度可以反映两个语句中词频的差异，从而弥补余弦相似度在语义不平滑空间中的不足。本文将余弦相似度和 Jaccard 相似度进行加权平均，得到最终的语句相似度 $Sim(S_i, S_j)$，计算方式如下：

$$Sim\left(S_i, S_j\right) = (1-\lambda)\cos\left(S_i, S_j\right) + \lambda J\left(S_i, S_j\right) , \tag{3.4}$$

其中，λ 是一个介于 0 和 1 之间的权重系数，用于调节余弦相似度和 Jaccard 相似度的贡献程度。本文通过实验确定了 λ 的最优值。

3.1.3 数值图坐标轴名称的识别

为了提取数值图的坐标轴信息，即 X 轴与 Y 轴数据对应的变量名称，本文对上一节获取的图片标题文本进行了命名实体识别。首先对标题语句进行分词，预训练模型会根据训练的语料库对文本单词进行切割，若出现不在词库上的单词，则将该单词切割为若干个词根。然后将单词序列送入预训练模型生成句子的向量表征，本工作依然选用 Sci-Bert 来生成更确切的材料文本的向量。最后添加 CRF 来对获得的向量表征进行预测，获得最优的标签序列。其模型架构如图 3.3所示。对于一个标题语句，首先将其切割为长度为 n 的单词序列 $X = (x_1, ..., x_n)$，送入到 Sci-Bert 之后生成长度为 n 的向量 $E = (E_1, ..., E_n)$，最后将该向量送入 CRF 层得到预测标签序列 $Y = (y_1, ..., y_n)$，每一种预测的序列的得分计算公式如下：

$$s(X, Y) = \sum_{i=1}^{n} P_{x_i, y_i} + \sum_{i=0}^{n} A_{y_i, y_{i+1}} , \tag{3.5}$$

其中，P_{x_i,y_i} 表示句子中单词 x_i 被预测为标签 y_i 的概率，而 $A_{y_i,y_{i+1}}$ 计算的是预测标签之间的影响程度，表示 y_i 转移到 y_{i+1} 的概率。其归一化后的表示如公式 3.6所示，并采用极大似然法计算正确的实体序列概率，计算方法如式 3.7所示。

$$p(Y, X) = \frac{\exp[s(X, Y)]}{\sum_{Y'} \exp\left[s\left(X, Y'\right)\right]} , \tag{3.6}$$

$$\log(p(Y,X)) = s(X,Y) - \log\left(\sum_{Y'}\exp\left[s\left(X,Y'\right)\right]\right), \tag{3.7}$$

其中，Y' 表示句子所有可能标注序列的集合。

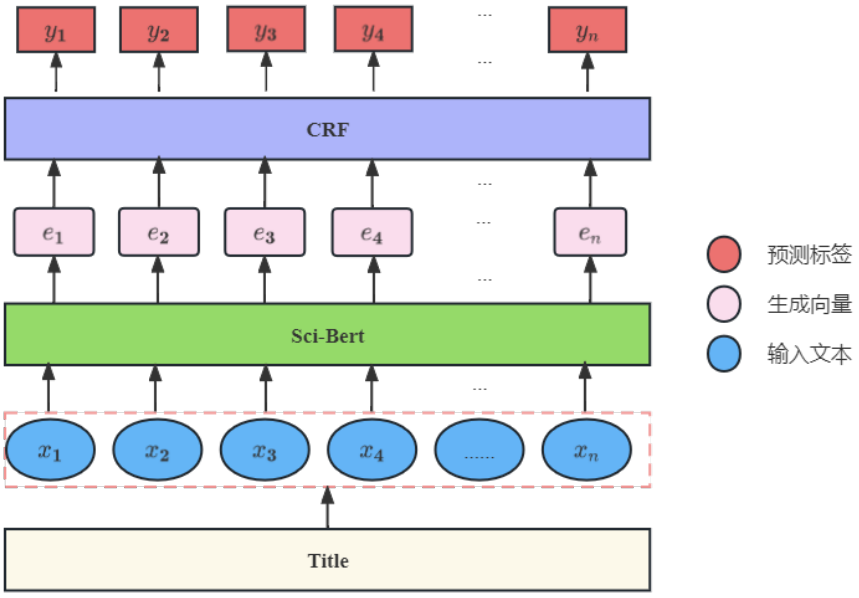
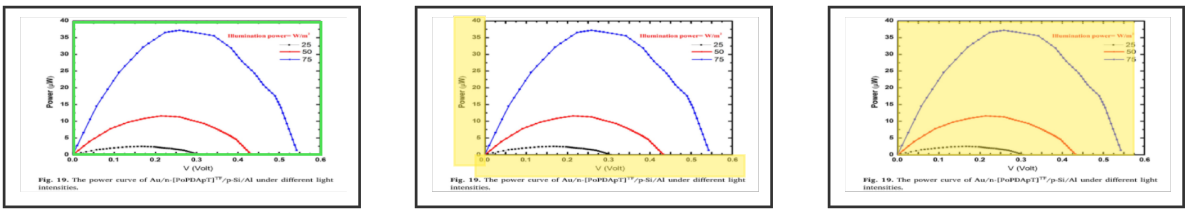


图 3.3: NER 网络架构

3.1.4 数值图真实数据的提取

本文还需对数值图图片所包含的数据信息进行提取，包括解析图片上的所有内容。如图 3.4所示，数值图图片主要包括坐标轴、坐标轴图例和绘图区域。因此本节利用图像处理和 OCR 等技术分别处理这三个部分，整合得到数值图的数据信息完成反向工程。

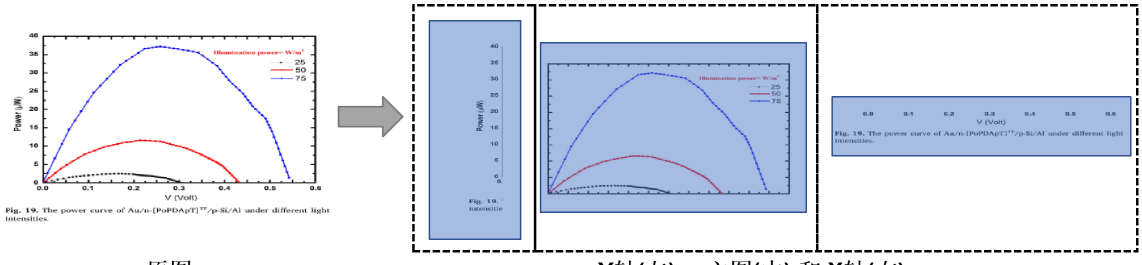
坐标轴位置可以确定绘图区域的边界，并且可以作为寻找数值图例的基础。数值图上的坐标轴在像素层面上可以将其看作 1*N 或 N*1 的矩形，因此通过形态学操作 [76] 中的开运算找到图片上所有的符合条件的矩形作为坐标轴候选。开运算是一个基于几何运算的滤波器，通过设定不同大小的结构从图上分割出对应的特征。开运算通过先腐蚀后膨胀的操作，能够去除像素中孤立的小点和毛刺，并保证总的位



（a）坐标轴（绿色框线）　（b）坐标轴图例（黄色区域）　（c）绘图区域（黄色区域）

图 3.4: 数值图的组成图示，该数值图图片来源于科学文献 [75]。

置和形状不变。考虑到坐标轴的刻度值总是贴近坐标轴且一条轴上的每个刻度值都是水平或垂直对齐的，以 Y 轴为例，Y 轴的左侧通常会有刻度且这些刻度数字大多处在同一竖直位置。因此使用 paddle OCR 识别图上的所有文本，根据文本框的位置分别找到一系列水平对齐和垂直对齐的数字文本确立为 X 轴和 Y 轴的轴刻度值，并找到最贴近两个轴刻度值的候选坐标轴作为 X 轴和 Y 轴。找到两个坐标轴直线位置后，可以将图片按轴线进行拆解。如图 3.5所示，一幅曲线图可以拆分为 Y 轴刻度值图，X 轴刻度值图以及主曲线图。同时，两条坐标轴的交点即可确立为该数值图的坐标轴原点 $O = (a, b)$，a 和 b 分别表示原点在整张图上的水平、竖直位置。



原图　　Y轴(左) 、主图(中) 和 X轴(右)

图 3.5: 数值图图片拆分的结果。整个图片分为三部分：Y 轴区域、X 轴区域和主要内容区域。

坐标轴的图例信息就在裁剪出的 Y 轴刻度值图和 X 轴刻度值图上，一般包括刻度数字以及单位等文本。Y 轴的刻度数字位于 Y 轴的左侧的最右边，而 X 轴的刻度数字位于 X 轴的下方的最上方。识别这些刻度数字并通过这些数字计算出单位像素点所对应的数值。以 Y 轴刻度为例，整体算法流程如算法 2所示，为了规避个别数字识别出错，计算每两个相邻数字间的差值并选取其中出现频率最高的差值作

为正确的刻度差，记作 D_y。用计算出该刻度差的两个数字文本框的高度差作为像素差，记作 P_y，算出最终单位竖直的像素所对应的刻度 $Scale_y$。同理，可以计算出 X 轴的真实刻度值 $Scale_x$。此外，刻度值的下方可能会出现单位字符，Y 轴单位通常最靠近 Y 轴刻度字符的左边，X 轴单位通常最靠近 X 轴刻度字符的下面，将这些字符识别出来作为每条坐标轴的单位。

算法 2: 数值图的单位像素刻度数值计算方法

Input: Y 轴的刻度数字集合 Y;

Output: 单位竖直像素刻度数值 $Scale_y$;

按水平高度降序排列 Y 中的刻度元素;

创立一个集合 F 用来存放刻度数字之间的差值;

for $y \in Y$ **do**

    计算其与后一个元素的数值差 D_y;

    以 (D_y, y) 的形式存入 F;

end

统计 F 中出现次数最多的 D_y;

计算得到该 D_y 值的 y 与其后一个元素的文本框像素差值 P_y;

统计所有 P_y 的平均值 $\bar{P}_y$;

最终得到单位竖直像素所表示的刻度数值 $Scale_y = D_y/\bar{P}_y$

主图部分包含着每条曲线的像素信息，通过这些曲线像素可以计算出最终的数值图的数值信息。但主图上往往会存在背景框线、曲线条例等干扰信息，因此在提取曲线像素时必须清除这些干扰项。首先，使用 OCR 识别出图上的字符，扩大这些字符框并将框内的所有像素置为白色以此来尽可能地消除干扰像素。然后，对主图进行闭运算和骨架提取操作。闭运算是对图像进行先膨胀、后腐蚀的运算，能够填平小孔，弥合小裂缝，使得曲线线条不会出现断裂，并且闭运算附带二值化的操作也可以消除一些图上的背景框线。而骨架提取操作是将膨胀腐蚀后的图像与原图像进行减法运算，使得曲线线条更加精细。接着根据图上像素点的颜色来分离每一条曲线，本文按照 RGB 值设定了 800 多种颜色类别，计算图上每个像素点与这些颜色的距离并规定最近的颜色即为该点的类别。最后，根据每个类别的像素点数量

大小筛选正确的曲线，假设获得的曲线像素点集合 $C_{all} = (c_1, c_2, ..., c_i, ..., c_j)$，每一个 c_i 代表一种颜色的像素点集合，按如下方法筛选出最终的曲线集合 C_{true}：

$$c_{max} = max(|c_1|, ..., |c_j|)\,, \tag{3.8}$$

$$f(c) = \{c_t \in C_{all} \mid |c_t| \geq \epsilon * |c_{\max}|\}\,, \tag{3.9}$$

$$C_{true} = f(c)\,, \tag{3.10}$$

其中 c_{max} 表示像素点最多的曲线颜色集合，ϵ 表示阈值，这里设定为 0.6，$|\bullet|$ 代表统计集合内的元素数量。规定 C_{all} 内的所有满足集合元素个数大于 $\epsilon * |c_{max}|$ 的颜色集合 c_t 即为真正的数值图曲线，集合内的元素即是一条曲线的全部像素位置。利用之前计算出的坐标轴原点的位置和两个坐标轴的单位像素刻度 $Scale_y$ 和 $Scale_x$ 将该条曲线对应的真实数据信息计算出来，最终获得整幅数值图的数据信息。

通过上述操作之后，提取的数值图数据信息仅是具体的数值大小，光是有了这些数字还是没有真正地解析数值图，我们将这些数字与上一节识别出的坐标轴名称以及数值图自带的单位字符进行结合，获得带有单位和描述对象的数值图信息。

3.2 实验与讨论

为了方便显示在大规模科学文献下本方法的优势，本文选择增材制造 [77] 领域内的材料科学文献来测试和验证。增材制造是一种基于数字模型，通过逐层叠加材料来制造实体零件的新型制造技术。它与传统的减材制造和等材制造相比，具有以下优点：（1）可以制造出结构复杂、性能高和成本低的零件；（2）可以实现个性化、定制化、智能化的生产；（3）可以节约资源和时间，保护环境；（4）可以促进创新设计和跨学科融合。增材制造是一项颠覆性的制造技术，其具有广泛的应用前景，涵盖了航空航天、汽车、医疗、建筑、教育等多个领域。它被认为是未来工业革命的重要驱动力之一。因此，对增材制造领域的科学文献进行文献挖掘，以获取其他学

者在科学文献中公开的关键数据信息，并进一步优化制备工艺，是一项具有重要价值的研究。

本文利用深度学习和图像处理对增材制造领域内的材料科学文献中的数值图进行图文信息的挖掘和处理，通过可视化的数据分析以此验证所提方法的有效性。所有实验均在 Intel(R) Core(TM) i7-10700 CPU 2.9GHz*16 和 32G RAM 的计算机上运行。

3.2.1 数据准备和实验设置

本文从 Elsevier 学术数据库下载了 6000 多篇 PDF 格式的增材制造领域的材料科学文献。首先，本文将每篇 PDF 格式的科学文献按页拆分成论文页图片，并选取了 600 多张带有单一数值图的论文页，对这些论文页上的数值图进行位置标注。利用 paddle OCR 识别出图上的标题文本，并在正文中摘录与之对应的图注标题，对每条标题语句对中的图片标号以及坐标轴实体进行标注。最终，本文获得了 756 幅单一数值图图片作为数值图截取任务的数据集以及 1000 条已标注的数值图标题文本作为 NER 任务的数据集。

本文在数值图截取任务中标注的数据集都是单一数值图的单幅图片，并按照 9：1 划分为训练样本和测试样本。此外，本文采用 BIO 标注方法对图片标题文本进行标签标注。BIO 方法是一种用于识别多单词实体的标注方式，在这种方法中，在实体的开头（B），内部（I）或外部（O）都有特殊标记来表示该实体。如图 3.6所示，本文将数值图标题中的图片标号、X 轴和 Y 轴实体分别用“Fig”、“X”和“Y”表示，并将实体的第一个单词标注为“B-（实体类别）”，实体的其余单词都标注为“I-（实体类别）”，非实体单词则被标记为“O”。本文将这 1000 句标注好的图片标题文本按照 8：2 划分为训练集、测试集，各种类别的实体数量统计结果如表3.1所示。

本章实验中的模型参数设置如下：在数值图截取实验中图像的大小为 640*640，训练 batch size 为 16，训练轮次为 1000，初始学习率为 0.001，衰减策略为每隔 20 步衰减 0.5；在 NER 实验中语句的最长长度为 128，训练 batch size 为 32，训练轮次为 5，初始学习率为 0.00005，丢弃率为 0.01。

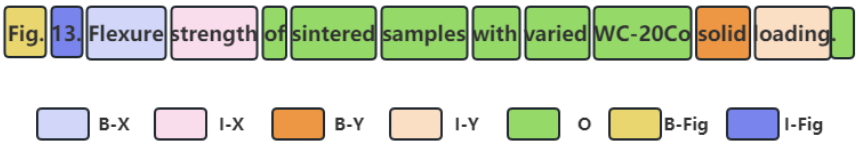


图 3.6: 坐标轴命名实体识别标注方法示例

表 3.1: 坐标轴命名实体识别数据集统计

实体类别	训练集	测试集
标号	700	204
X 轴实体	696	209
Y 轴实体	765	219

3.2.2 单一数值图截取任务结果

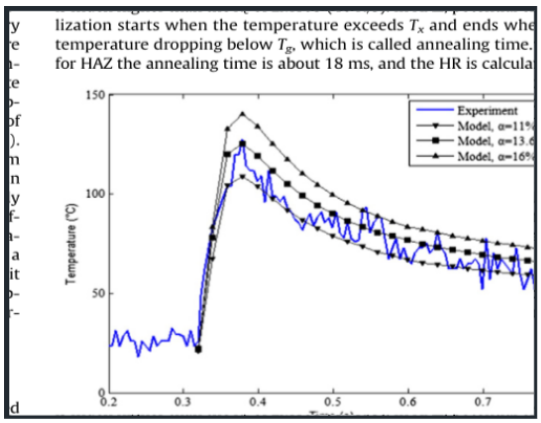
本文对比了 Faster R-CNN[56]、SSD[57]、Detr[58] 和 Yolov5s[59] 四种网络在数值图截取任务上的结果，并定性地展示了截取的数值图图片。Faster R-CNN 是一种典型的两阶段目标检测网络，它将目标分类和定位当作两个任务分开进行。相比之下，SSD 与 Yolov5s 是一阶段网络。Detr 则是将 transformer 运用到目标检测中的网络，它简化了之前目标检测的一些传统步骤，如 NMS、anchor 等。本文在实验中只有当检测出的图片是完整的单一数值图并且附带图片标题时，才认为该目标是正确的。

四种网络的预测精度和参数量如表 3.2所示，在原始的网络预测中，Yolov5s 的截取精度达到了 87.8%，是四个网络中最高的，其参数量也是最少的，甚至比 Faster R-CNN 少了数十倍，保证了截取的效率。再加入了 3.1.1 节提到的科学文献图片检测方法后 Yolov5s 的精确率达到了 96.4%，也是四种网络中最高的。实验结果证明了科学文献图片检测方法的有效性和健壮性，它在每种网络上都能提升原有的精度，在 SSD 模型上提升效果最明显，达到了 9.3%。图 3.7以定性的结果展示了科学文献图片检测方法的优越性。以 SSD 网络为例，原始网络预测出的图片不仅缺少数值图的标题，而且数值图本身也不完整，这样的结果是不符合任务要求的。在加入了改

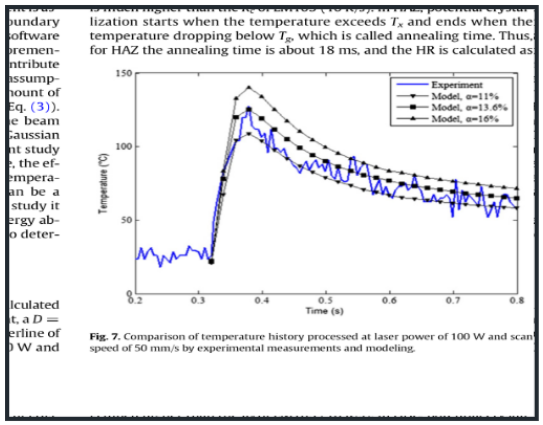
进的数值图检测方法后，该网络能够截取更完整的数值图，并且包含数值图的标题，更符合任务需要，体现了该方法的优势。此外，由于数值图带有大量空白区域，模型在学习图像特征时容易出现偏差，会将科学文献中的公式、期刊标志、页眉或页脚等部分误识别为数值图。而在加入了科学文献图片检测方法后，这些情况将会减少或消除，体现了该方法的健壮性。

表 3.2: 4 个网络截取数值图的精确率和网络参数对比

模型	精确率	精确率（加入科学文献图片检测方法）	网络参数量
Faster R-CNN	86.2%	94.2%	136,873,389
SSD	83.7%	93.0%	23,612,246
Detr	87.2%	89.8%	41,279,495
Yolov5s	**87.8%**	**96.4%**	**7,276,605**



（a）原始网络预测结果



（b）改进后的预测结果

图 3.7: SSD 网络截取数值图结果，（a）为原始网络预测的位置结果，（b）为加入改进方法后对同一数值图的预测结果。该数值图图片来源于科学文献 [78]。

表 3.3对比了本方法与其他方法在截取任务中的精度。Pdfimages[79] 是针对 PDF 格式文件的解析工具，该工具利用 PDF 的文件信息截取图片，且无法对多数值图和单一数值图进行分类，因此准确率非常低。由此也证明了基于规则的方法无法满足高精度任务的需求。PSiegel 等人 [47] 使用 ResNet 作为基础网络进行目标检测，

其截取精度达到了 80.6%。Younas 等人 [80] 使用特征金字塔与 Detr 融合，将精度提升至了 88.6%。Naiman 等人 [81] 则是在 Yolov5 的基础上加入了灰度特征作为判断规则进行目标检测，将精度提升至了 93.5%。而本文则是在 Yolov5s 网络的基础上改进了数值图检测方法进行目标检测，最终的精度也高于其他三种方法，达到了 96.4%。图3.8展示了本方法在论文页上截取单一数值图图片的效果。

表 3.3: 各种截取数值图方法精度比较

	Pdfimages	Siegel	Younas	Naiman	Ours
精确率	9.8%	80.6%	88.6%	93.5%	**96.4%**

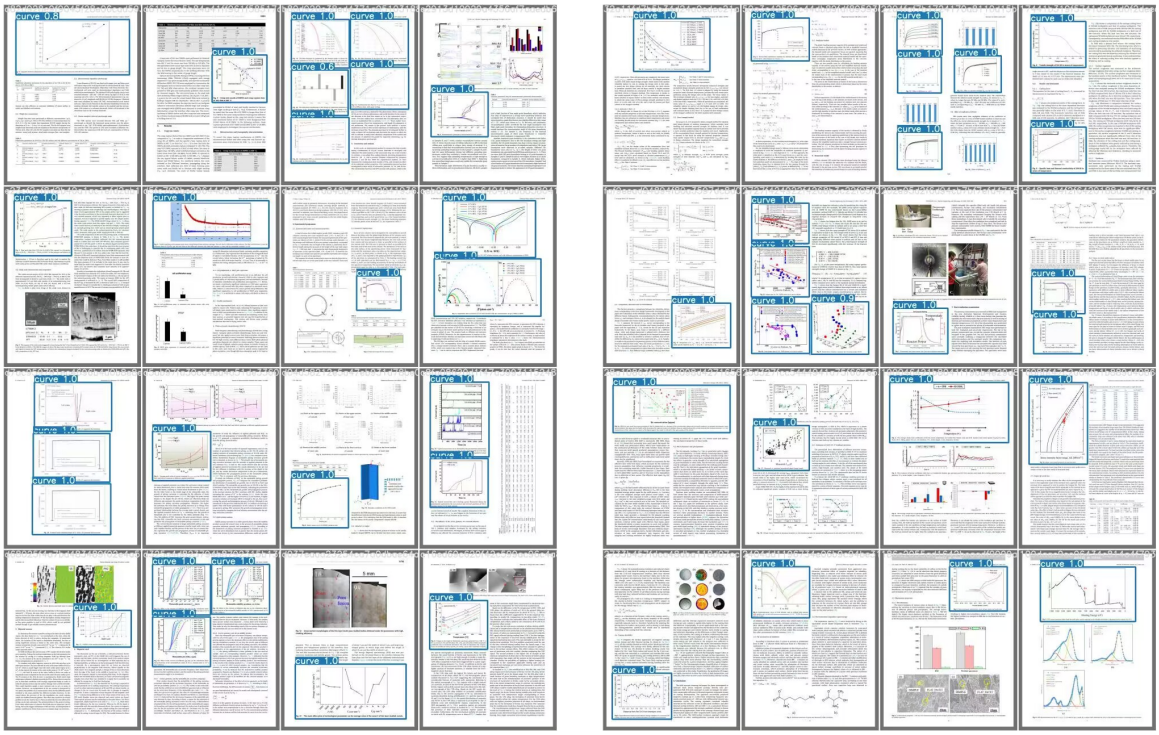


图 3.8: 本方法在论文页上的单一数值图截取效果

3.2.3 数值图标题匹配任务结果

本文首先测试了对于式 3.4 中 λ 的取值对标题匹配的精确率影响，以 0.1 为间隔进行十次测试，结果如图3.9所示，发现当 $\lambda = 0.5$ 时效果最佳。此外，结果表明以 Jaccard 相似度（$\lambda = 1$ 时）为偏重指标的匹配效果优于以余弦相似度为偏重指标的

匹配效果。接着测试了四种寻找数值图标题文本的方法的精度，结果如表3.4所示。其中直接识别图上的标题所获得的标题的精确率仅有 43.6%，这是因为数值图上字符大小以及位置都是不确定的，OCR 识别出的标题可能会被打断或缺失，说明了在文本中寻找标题语句的必要性。因此在数值图上识别出带有图片标号的文本之后，通过文本相似度在 PDFminer 解析出的 txt 中寻找相似度最高的语句作为标题文本。表3.4对比了三种不同的相似度计算方法匹配出的标题的精度。由于余弦相似度只关注了文本向量方向上的一致性，缺乏了对语义信息的考量，匹配出的标题精确率只有 89.2%。而本任务的实质是以不完整的句子找寻其完整的语句形式，Jaccard 相似度关注的是文本语句的重合度，更适合本任务，用该指标作为相似度计算方法寻找出的标题精度达到了 93.5%。但由于 OCR 识别出的文本可能只是标题的一部分，因此将二者结合作为相似度指标更贴合本任务，精确率达到了 95.8%。

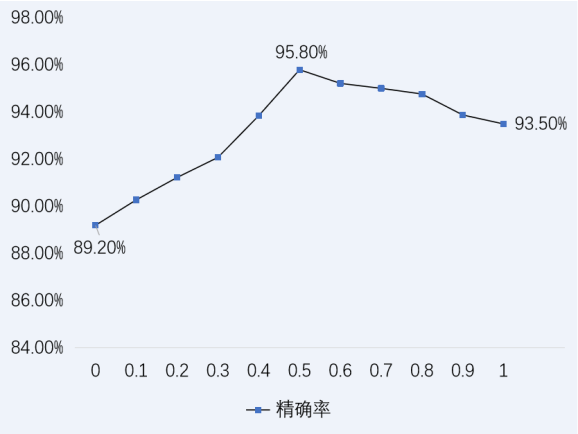


图 3.9: 不同 λ 对匹配结果的影响

表 3.4: 四种寻找数值图标题方法的结果

方法	精确率
paddle OCR 直接识别图上的标题	43.6%
利用余弦相似度匹配出的标题（$\lambda = 0$）	89.2%
利用 Jaccard 相似度匹配出的标题（$\lambda = 1$）	93.5%
余弦相似度 +Jaccard（$\lambda = 0.5$）	**95.8%**

3.2.4 NER 任务结果

本节首先对比了不同网络模型在识别坐标轴实体名称任务上的效果，结果如表3.5所示，包括 Lstm、Bi-Lstm、Bert、Bio-Bert[30]、Clinical-Bert[31]、Sci-Bert，每种模型都加入了 CRF 层来优化预测的标签序列。

表 3.5: 不同网络提取标题中坐标轴实体名称的结果

模型	实体名称	精确率	召回率	F1 得分
Lstm+CRF	Fig	0.9657	1.0000	0.9826
	X	0.5550	0.6372	0.5932
	Y	0.5281	0.6099	0.5661
Bi-Lstm+CRF	Fig	0.9755	1.0000	0.9876
	X	0.5972	0.6522	0.6235
	Y	0.5420	0.6269	0.5814
Bert+CRF	Fig	1.0000	1.0000	1.0000
	X	0.6810	0.7566	0.7168
	Y	0.6652	0.7209	0.6920
Bio-Bert+CRF	Fig	0.9947	1.0000	0.9973
	X	0.7449	0.7725	0.7584
	Y	0.7249	0.7721	0.7477
Clinical-Bert+CRF	Fig	1.0000	1.0000	1.0000
	X	0.7638	0.8042	0.7835
	Y	**0.7401**	**0.7814**	**0.7602**
Sci-Bert+CRF	Fig	1.0000	1.0000	1.0000
	X	**0.7913**	**0.8624**	**0.8253**
	Y	0.7354	0.7628	0.7489

Bi-Lstm 作为双向的 Lstm 模型，相比于 Lstm 可以同时利用前向和后向的信息，从而更好地捕捉时序数据的上下文关系，因此该网络的效果优于 Lstm，在各个实体预测的效果上都有所提升。但静态词向量模型生成的文本向量无法根据上下文动态调整，捕捉单词的语义变化，而动态词向量模型弥补了这一点。在本任务中

预训练模型 Bert 的效果超过基于静态词向量训练的 Bi-Lstm 网络约 8%。Bio-Bert、Clinical-Bert 以及 Sci-Bert 三个预训练模型均以公共语料库为基础，在一些特定领域的语料文本上进行了二次训练，因此在材料科学文献的文本上进行 NER 任务时效果会优于 Bert。其中 Sci-Bert 与 Clinical-Bert 分别在实体“X”和“Y”上取得了最高的准确率。在每一类的识别精度上，对于数值图的图片标号实体“Fig”，每个网络的识别准确率几乎都达到了 100%。这是因为在大多数科学文献中，图片标号的形式几乎是一致的。而对于实体“X”和“Y”，由于每个领域学者在绘制数值图时坐标轴对象都是不固定的，且名称术语也是多样化的，导致在这两个类别上的识别精度下降。

表 3.6: Softmax 和 CRF 分类器的效果对比

模型	Macro-P	Macro-R	Macro-F1
Lstm+Softmax	0.6600	0.7363	0.6954
Lstm+CRF	0.6829	0.7490	0.7140
Bi-Lstm+Softmax	0.6929	0.7466	0.7187
Bi-Lstm+CRF	0.7049	0.7597	0.7308
Bert+Softmax	0.7758	0.8083	0.7917
Bert+CRF	0.7821	0.8258	0.8029
Bio-Bert+Softmax	0.8099	0.8411	0.8252
Bio-Bert+CRF	0.8215	0.8482	0.8345
Clinical-Bert+Softmax	0.8280	0.8517	0.8397
Clinical-Bert+CRF	0.8346	0.8619	0.8479
Sci-Bert+Softmax	0.8377	0.8618	0.8496
Sci-Bert+CRF	**0.8422**	**0.8751**	**0.8581**

表3.6展示了 Softmax 和 CRF 作为标签分类器的效果对比。CRF 是一种基于全局无向转移概率图的方法，能够有效地考虑单词前后的上下文关系。例如，假设已经知道前面一个单词标签为 B-X，则下一个单词的预测标签大概率是 I-X。因此全局的概率转移建模在 NER 任务更加合理。从每个网络的对比结果中，可以发现 CRF 层确实提升了网络的预测精度。此外，Bert 系列的模型使用的是全局的自注意

力机制，因此在 Lstm 和 Bi-Lstm 上使用 CRF 层代替 Softmax 的提升效果会比在预训练模型上的提升效果更加明显。最后表3.7展示了坐标轴实体识别的定性结果，在三种不同语句序列的标题文本中都能够准确地提取出 X 轴和 Y 轴的坐标轴名称，这表明即使句子结构不同，网络仍然能够有效地进行预测。

表 3.7: 坐标轴实体识别结果实例

标题文本	图片标号	X 轴名称	Y 轴名称
Fig. 7. Load-displacement diagrams: numerical vs experimental tests, and dense vs porous stem concept.	Fig. 7.	Load	displacement
Fig. 4 –Effect of admixture on the thermal diffusivity of the samples.	Fig. 4	admixture	thermal diffusivity
Fig. 11 – Comparing the stress strain curves resulted by FEM with the experimental data.	Fig. 11	strain	stress

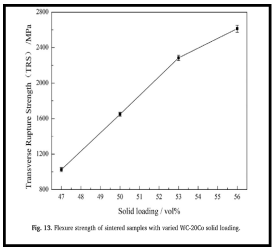
3.2.5 数值图图片数据提取任务结果

本节按照每一步的流程对提取数值图数据信息任务进行了测试评估。首先测试了寻找坐标轴位置以及计算单位像素所对应的真实刻度值方法的准确率。规定当找到的坐标轴与真实坐标轴的 IoU 大于 0.5 时，认为该轴是正确的。在此标准下，寻找坐标轴位置方法的准确率为 90.2%。对于真实刻度值的评估，本文比较计算出的真实刻度值与实际刻度值之间的差异，如果差异低于 5%，则认为该刻度值是有效的。在此原则下，计算真实刻度值方法的准确率达到 92.6%。

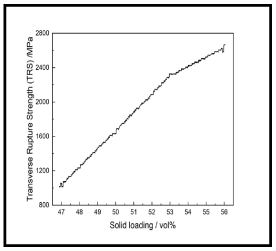
为了评估最终提取出的数值的准确率，本文采用如下定义：对于曲线上的点 x_i，如果计算出的数据值 y_i 与数值图所对应的真实值 y_i' 的差值小于 5%，则认为该点计算正确。若整条曲线的正确率高于 95%，则判定所提取的数据是正确的。最终计算出的准确率为 28.2%。

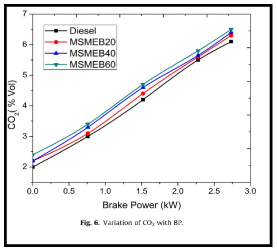
图 3.10展示了从数值图上提取数值的定性结果，将提取出的数值信息绘制成新

的数值图并与原图进行对比，OCR 识别出的刻度轴单位作为坐标轴信息添加在对应的坐标轴上。图（a）展示了单一曲线的数值图识别结果，可以发现原数值图可以很好地被还原，证明了提取的数值的准确性。图（b）和图（c）展示了多曲线的复杂数值图识别结果，可以看到文字以及背景框线等干扰问题也能够被完全清除，每条曲线的数值都被正确地提取出来。同时，本方法也存在着局限性，如图（d），若图上出现相同颜色的曲线，提取出的数值就会出错。

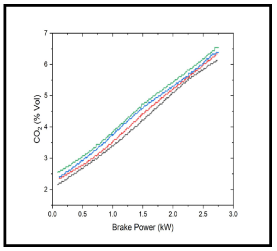


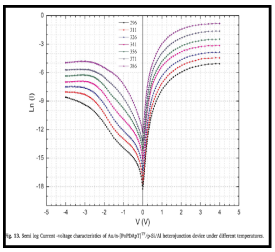
（a）单曲线的情况，来源于科学文献 [82]。



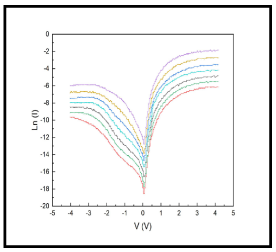


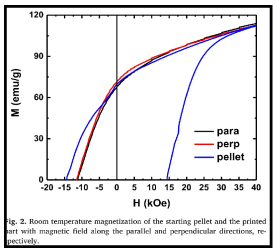
（b）多曲线的情况，来源于科学文献 [83]。





（c）背景复杂的情况，来源于科学文献 [75]。





（d）失败的情况，来源于科学文献 [84]。

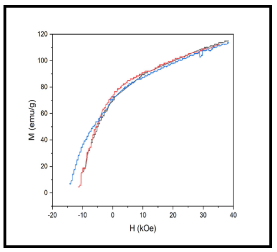


图 3.10: 数值图数据提取的定性比较。原始数值图显示在左侧，还原的数值图显示在右侧，以便进行直观的比较。很好地再现了图（a）、（b）和（c）。由于图（d）存在颜色相同的曲线，导致部分数据提取失败。

3.2.6 具体应用

本节首先展示了从 PDF 格式的材料科学文献到最后获得数值图信息的完整流程，如图 3.11所示。首先从科学文献中解析出文本内容并截取出单一数值图图片，然后从截取的数值图图片中自动提取数值图上的数值信息，包括图上坐标轴附近的单位信息；同时从对应的标题语句中识别出坐标轴的名称。最初的图片结果仅仅能够得知数值、单位等信息，并不知道这些数值描绘的具体对象，这样的挖掘结果是不全面的。补充了 X 轴和 Y 轴名称之后，数值图的信息将会更加丰富，证明了从多

类型数据进行文献挖掘的优势。同时，该方法可以让材料学者进行大规模的材料科学文献数据提取工作，学者通过理解和分析这些数据来推动材料领域研究的快速发展。

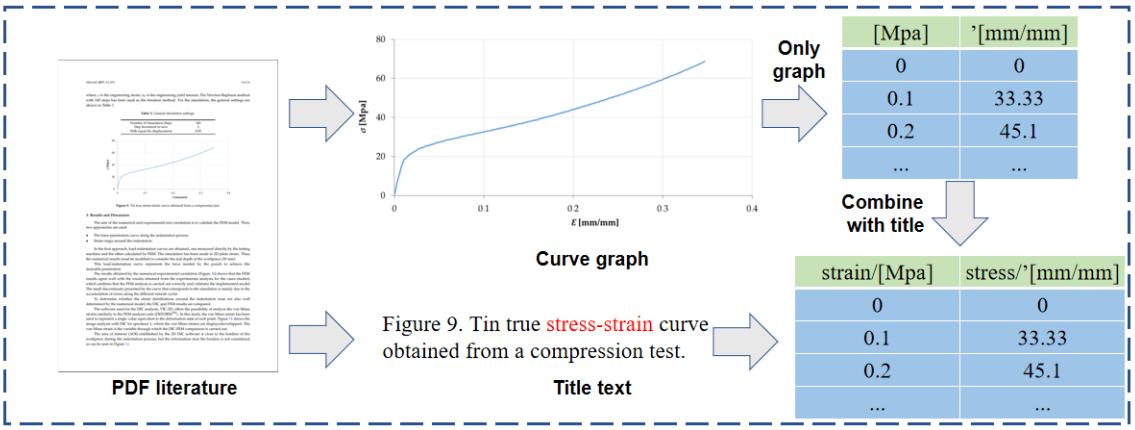


图 3.11: 从 PDF 材料科学文献中自动提取数值图信息结果的整体流程，来源于科学文献 [85]。从数值图图片上只能获得数值及少有的单位信息，在数值图对应的标题文本中提取出坐标轴实体可以丰富提取出的信息，更好地帮助材料学者进行研究。

此外，本节对下载的 6000 多篇增材制造的 PDF 科学文献进行了完整的测试，以证明本方法在大规模的材料科学文献上进行文献挖掘的有效性。图3.12按年份统计了科学文献数量以及截取出的单一数值图图片数量，能够发现学者们更倾向于在文章中使用数值图来展示自己的数据结果，平均每一篇科学文献中至少有一幅单一的数值图，2020 年该比例达到了 1.8，这也说明了从科学文献中的数值图提取信息的必要性。接着本节逐年统计了识别出的数值图坐标轴实体名称，图3.13分别展示了 2017 年以及 2021 年坐标轴实体的词频分布。在 2017 年的科学文献中，增材制造领域内的学者们对相关材料的“Stress”，“Strain”和“Temperature”等属性研究最为频繁，而到了 2021 年，“Density”，“Energy”成为了新的热门研究属性。该数据表明了本方法可以快速帮助学者捕捉领域内的研究热点，从而帮助推进相关领域的研究。

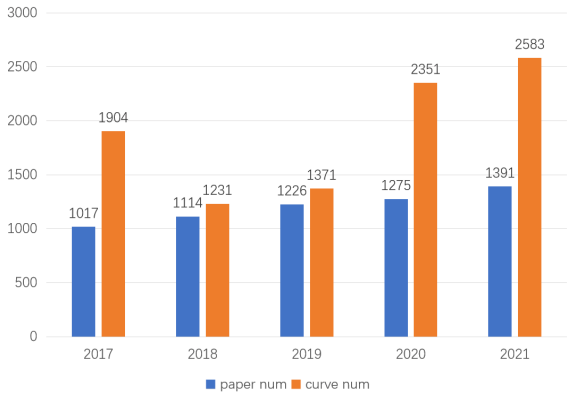
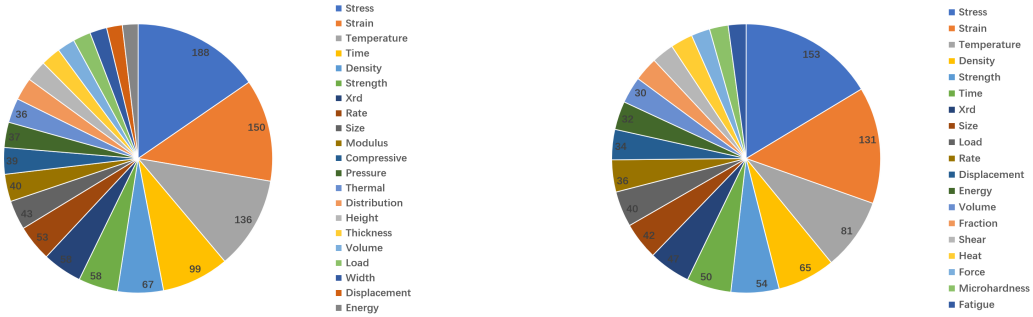


图 3.12: 从不同年份的科学文献中截取的单一数值图的数量。文章全部来自 Elsevier 数据库。蓝色柱状条代表挖掘的科学文献数量，黄色柱状条代表截取出的单一数值图图片的数目。



(a)2017 (b)2021

图 3.13: 2017 和 2021 年份的科学文献中数值图坐标轴名称统计结果

3.3 本章小结

本章提出了一种自动化提取材料科学文献中单一数值图图片信息的方法，包括单一数值图的截取，数值图标题语句的匹配，标题中坐标轴实体的识别以及数值图数值信息的提取。该方法可以处理大规模的材料科学文献，方便学者获取关键信息，在此基础上理解，分析和总结数据，推动领域研究的新发展。科学文献中的数据形式包含非结构化文本，图片以及表格等，若从多种形式的数据中挖掘信息并进行整合，将会获得更丰富的信息，实现多元化的文献挖掘。本方法就是同时提取科学文献中单一的数值图图片及其所对应标题文本两种不同形式数据的信息，将二者整合之后获得更全面的科学文献中数值图信息。此外，本方法还改进了科学文献图片检

测的方法，利用科学文献中图片的特点帮助网络更好地完成截取任务。在匹配数值图对应标题的任务中，本方法结合了余弦相似度与 Jaccard 相似度，同时关注了语句在高维空间中的方向差距和语句中单词的重复率，从而更准确地匹配出对应的标题文本。希望该工作将推动文献挖掘这一领域的继续研究，同时通过挖掘出的信息帮助学者推动专业领域的发展。

虽然当前的方法已经得到了初步的效果，但远远还未达到工业化的程度。尤其是在提取数值图数值信息的这一工作上，仅利用最基本的图像处理的知识，还无法处理图中有相同颜色的曲线图。另外，科学文献中的数值图的样式还有很多，如多坐标轴和刻度数字为对数等情况，目前能处理的数值图种类比较有限。最后，科学文献文本中能够提取的数值图信息远远不止坐标轴名称以及数值，需要考虑更全面、更多样的信息。

第四章　基于图文的材料科学文献数值图坐标轴实体识别提升方法

在上一章所述的方法中，对于材料科学文献中单一数值图的标题进行 NER 任务时存在着以下的问题：文本数据集规模小导致模型识别坐标轴实体时的精度不够高，且标注任务困难，人工标注成本较高。传统的文本数据增强方法仅在现有的数据样本中进行替换修改，其增强效果有限，无法让模型学到新的知识。针对此问题，结合科学文献中的正文部分会存在与数值图对应的描述语句以及数值图图上坐标轴区域会带有标签信息这两个特点，本章提出了一种提升数值图坐标轴实体识别效果的方法。该方法通过上下文搜寻数值图标题语句对应的正文描述语句作为标题的扩充文本，并通过识别图上的坐标轴名称信息进行数据集扩充，从而提升在数据集规模较小情况下的坐标轴实体识别效果的方法。

4.1　方法概述

本工作的具体流程如图4.1所示，主要包含四个部分：对于 PDF 格式科学文献的预处理，在正文中寻找数值图对应的的描述语句，根据数值图图片上的信息扩充数据集以及结合描述语句的坐标轴实体识别。首先，本工作基于第三章的工作得到 PDF 格式的材料科学文献中的单一数值图图片及其对应标题和全文的文本，然后使用 NLTK 工具包按句整理科学文献的文本，数值图图片和文本是后续任务的基础。接着，在文本集合中利用文本相似度找到正文中对数值图的描述语句作为扩充文本。然后使用 OCR 识别数值图图片上坐标轴附近的文本标签，根据之前标注好的标题语句格式创建标签模板，将文本标签填充入模板以达到数据集扩充的目的。最后将标题文本和扩充文本拼接成为新的输入语句，利用 Sci-Bert 生成语句向量，通过 CRF 优化预测的标签序列。

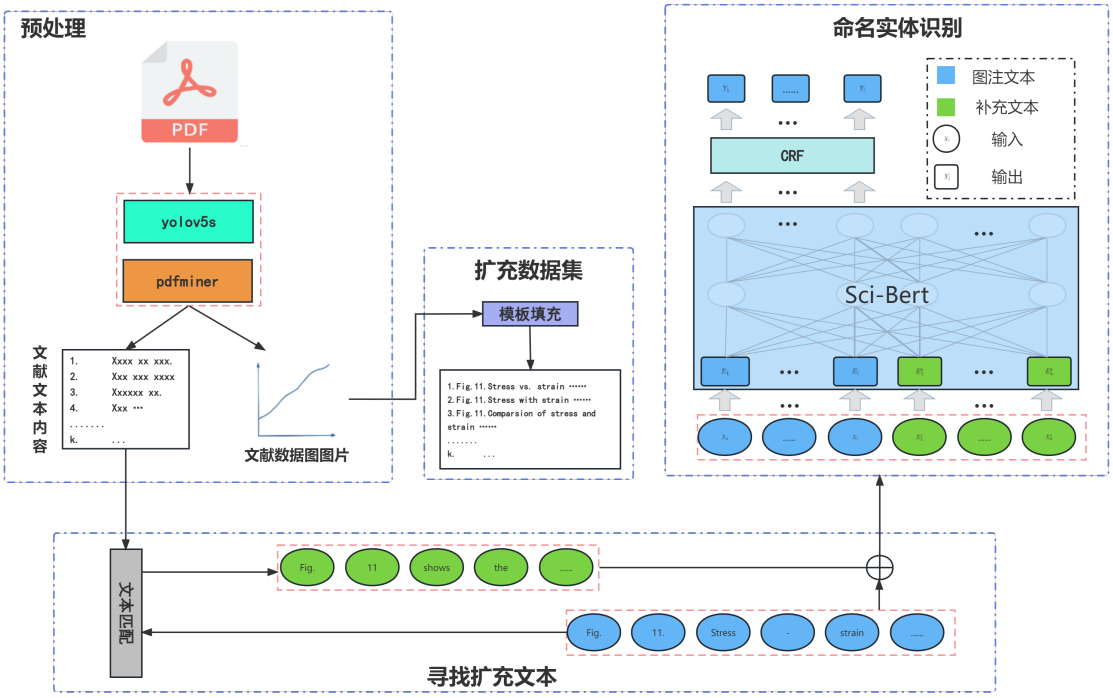


图 4.1: 基于图文的数值图坐标轴实体识别效果提升方法流程

4.1.1 预处理

预处理过程大致如 3.1.1和 3.1.2节所述，不同之处在于在前一章的任务中只需从文本中匹配数值图的标题，而本章需要匹配文献正文中对数值图进行描述的语句。由于 PDFminer 是利用文字位置等信息从 PDF 文件中解析文本，图片的标题往往独立于正文部分，以单独语句的形式保存在解析出的 txt 文件中；而正文则是以段落的形式被解析，数值图的描述语句与其他语句杂糅在一起，这为匹配任务带来了困难。因此需要对 PDFminer 解析后的文本内容做更精细的处理。

本工作在 3.1.1节处理后的文本中使用 NLTK 工具包 [86] 进行分句，使得科学文献的文本以句子集合的形式呈现。NLTK 是自然语言处理领域中经常使用的开源工具，用于自然语言处理的研究和开发。其应用场景非常广泛，包括文本分类、分词分句、词性标注等。NLTK 的分句功能是基于 Punkt Tokenizer 进行训练的，使用无监督的学习方法训练缩写词、搭配和句子开始的词标志的模型，将文本划分为句子列表。经过 NLTK 分句之后，科学文献中的每一句话都被分开，方便了后续寻

找正文中的数值图描述语句。

4.1.2 寻找扩充文本

除了图表自身的标题语句，科学文献的正文部分通常也会对图表进行描述和解释，这些语句相比标题更为详尽和具体，有助于读者快速把握图表的主要信息。此外，已有的大量研究证明，在上下文嵌入模型的输入中加入目标语句的上下文可以提高 NER 任务的性能 [87, 88]。因此，本节采用了该思想，在正文中寻找与每张数值图相对应的描述语句，并将其作为标题语句的补充文本，从而增强数值图坐标轴实体识别任务中输入标题的上下文关联性，进而提升识别的准确率。

算法 3: 数值图正文描述语句寻找方法

Input: 文本语句集合 C_s; 数值图标题语句 S_{title};

Output: 数值图正文描述语句 S_{text};

计算 S_{title} 经过 Sentence-Bert 之后得到的向量表征 E_{title};

在 C_s 中删除元素 S_{title}:$C_s = C_s - \{S_{title}\}$;

for 语句 $s \in C_s$ **do**

  计算 s 经过 Sentence-Bert 之后得到的向量表征 E_s;

  计算 E_s 与 E_{title} 的余弦向量作为 s 与 S_{title} 的余弦相似度 $cos(s, S_{title})$;

  以空格为分隔符将语句 s 和 S_{title} 分割成单词集合 W_s 和 $W_{S_{title}}$;

  计算 s 与 S_{title} 的 Jaccard 相似度 $J(s, S_{title})$;

  计算语句 s 与 S_{title} 的复合相似度:$Sim\,(s, S_{title}) = \lambda \cos\,(s, S_{title}) + (1-\lambda) J\,(s, S_{title})$;

  将复合相似度 $Sim(s, S_{title})$ 和对应的语句 s 以 $(Sim(s, S_{title}, s)$ 的形式保存进集合 C_{Sim};

end

以 $Sim(s, S_{title})$ 降序排列 C_{Sim};

得到 C_{Sim} 第一个元素的语句 s 作为数值图正文描述语句 S_{text};

在 3.1.2中介绍了如何利用文本相似度匹配出数值图的标题文本，本节采用同样的方法去寻找数值图在正文中的描述语句，以余弦相似度与 Jaccard 相似度为衡量

指标，计算文本中每一语句与标题语句的相似度，相似度最高的语句即为数值图在正文中的描述语句。但在文本中会存在标题语句干扰匹配结果，因此在匹配之前需要在文本中删除标题语句再进行匹配，具体的流程如算法 3所示。

4.1.3 扩充数据集

为了保证模型的准确性，材料领域下的 NER 任务往往需要大量的标注数据进行训练。但材料文本的标注任务又面临着非常大的困难，材料领域专业性较强，需要标注人员具备一定的专业知识；此外，数值图坐标轴命名实体的歧义性和实体的长尾部分也是一大难点 [89]。小规模数据集下的 NER 任务成为了研究热门，一些在有限数据集下提升 NER 识别效果的方法被提出，例如数据增强 [90]、对比学习 [91]、Prompt 思想 [35] 等。NER 的数据增强方法包括同义词替换、实体替换等，然而该方法生成的只是伪造样本，随着真实的样本数据量变大，其作用会减弱；Prompt 思想是一种基于模板的方式，将所有实体与手工设定的模板进行拼接，使用模型对每一个模板打分，通过所有模板的得分预测出最终的实体类别。如图4.2所示，传统的 NER 任务采用序列标注模型对句子中的每个单词进行标签预测，以得分最高的类别作为该单词的实体类别。而基于模板的方法首先需要根据任务设定模板，再将语句中的每个单词填入模板，使用模型对每一个模板进行评分，将得分最高的模板中的实体标签作为该单词的实体类别。例如图中单词“ACL”获得最高得分（0.9）的模板中的实体标签为“organization”，则将“ACL”预测为“organization”类别。

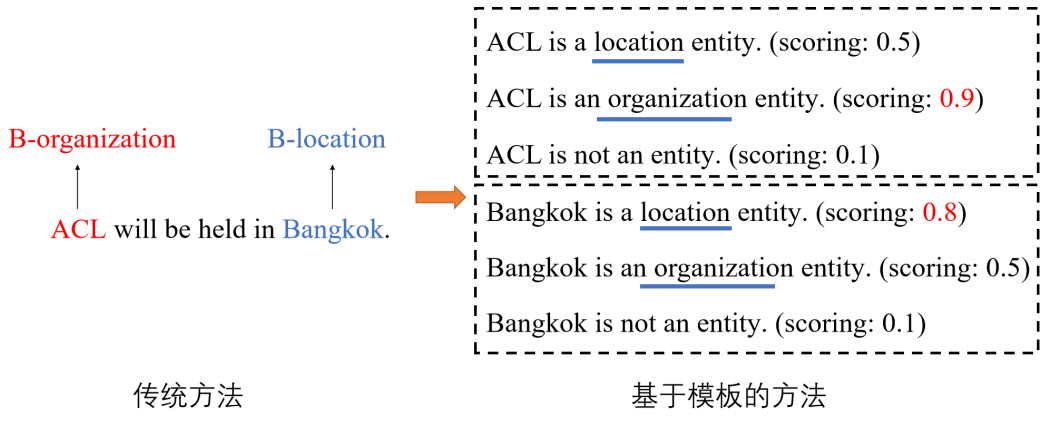
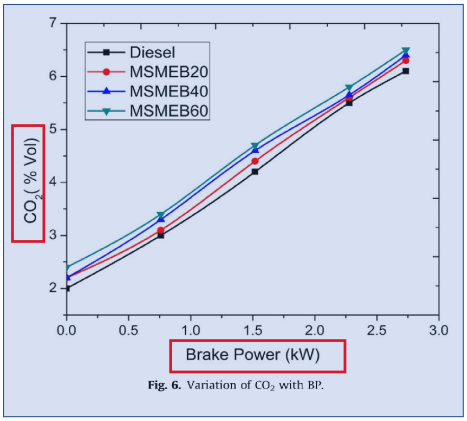
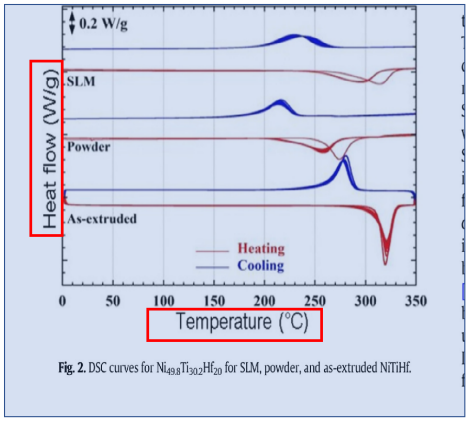


图 4.2: 基于模板的 NER 任务

本文结合数据增强和 Prompt 思想相结合的方法，利用数值图图片上的坐标轴文本标签作为候选实体，并根据之前标注好的文本数据集设定模板，将这些候选实体嵌入到模板中，生成无需标注的文本数据，从而实现数据集的扩充。如图 4.3红色矩形框所示，为了便于读者快速理解科学文献中的数值图，作者在绘制图片时往往会在坐标轴附近添加文本信息，包括坐标轴名称和单位信息等。这些文本标签可以补充标题中未能包含的信息，如 4.3右图的文本标签就是对标题中的“DSC”进行了更具体的说明。因此，将这些信息纳入到坐标轴实体识别中，不仅能增加数据量，还能弥补标题文本中可能存在的信息缺失，提高数据集的完整性，从而改善训练效果。





标题：Fig. 6. Variation of CO2 with BP.　　标题：Fig. 2. DSC curves for Ni49.8Ti30.2Hf20 for SLM, powder, and as-extruded NiTiHf.

图 4.3: 数值图图上文本标签示例，数值图图片来源于科学文献 [83, 92]。

本文采用第三章的方法，对 OCR 识别出的数值图图片标号，X 轴和 Y 轴标签文本进行处理。首先对文本进行清洗，去除文本中的单位文字，如图4.3所示，文本标签中的单位文字一般会用“()”或“/”进行区分，因此利用该特性将单位字符删除。然后根据之前标注好的数据集进行了统计分析，设定了如表4.1所示的六种模板格式，将实体填入到模板中，生成无需标注的数据文本。特别地，当文本标签仅存在单位文字时，即进行清洗后文本标签为空时，便舍弃该次填充，整体流程如算法4所示。

表 4.1: 不同模板形式

	模板
1	[Fig] Effect of [X] on [Y].
2	[Fig] Relationship between the [Y] and [X].
3	[Fig] The [Y] with different [X].
4	The [Y] as a function of [X] is shown in [Fig].
5	[Fig] Change in [Y] with the [X].
6	[Fig] Dependence of [Y] on [X].

算法 4: 基于数值图文本标签和模板的数据增强方法流程

Input: 数值图图片集合 C_{curve}; 模板集合 $C_{template}$;

Output: NER 任务数据样本;

for $c \in C_{curve}$ **do**
  以第三章的方法分割数值图图片 c;
  用 paddle OCR 识别图片标号文本、X 轴和 Y 轴的标签文本，获得 $Entity_{fig}$、$Word_x$ 和 $Word_y$;
  对 $Word_x$ 和 $Word_y$ 进行清理，去除文本中“()”内的内容以及“/”后的内容;
  for $t \in C_{template}$ **do**
    按照设定好的实体位置填入 $Entity_{fig}$、$Entity_x$ 和 $Enity_y$;
  end
end

4.1.4 命名实体识别

本节基于4.1.2节得到的扩充文本改进了数值图坐标轴名称识别模型的网络架构。如图4.4所示，首先将长度为 t 的数值图标题语句 S_{title} 与长度为 u 的扩充文本语句 S_{text} 进行拼接，构成新的输入语句 S_{expand}。然后对输入语句 S_{expand} 经过分词，得到长度为 $t+u$ 的单词序列 $X_{expand} = (x_1, ...x_t, x'_1, ...x'_u)$。接着利用 Sci-Bert 对单词序列进行编码，生成上下文关联的语句向量 $E_{expand} = (e_1, ...e_t, e'_1, ..., e'_u)$。Sci-Bert 通过 Transformer 来对序列进行编码，通过多头自注意力机制可以捕捉到语句序列间的上下文关系，因此加入了扩充文本之后生成的语句向量会更加关注词与词

之间的关联性。最后将编码向量输入 CRF 模块，CRF 预测得到最终的预测序列 $Y_{title} = (y_1, ..., y_t)$。

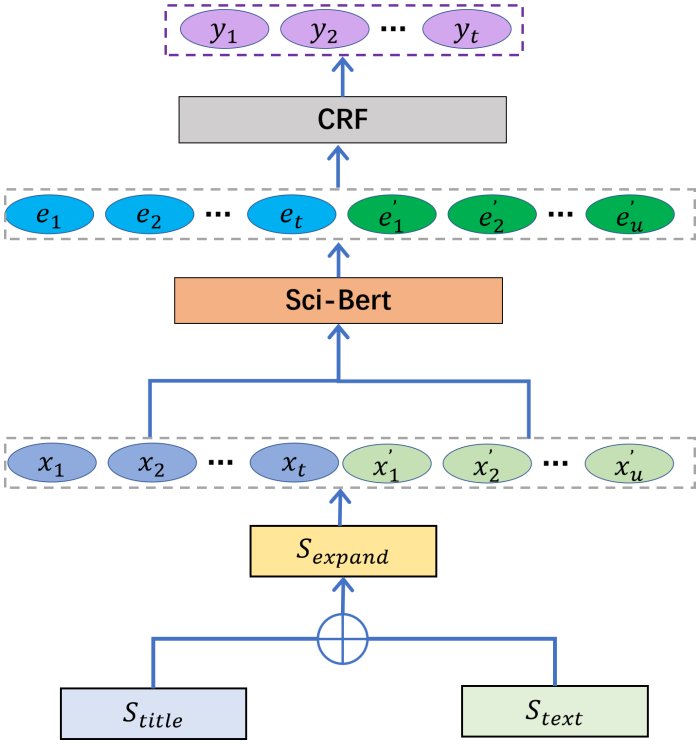


图 4.4: 加入扩充文本的 NER 模型框架

CRF 预测序列得分的方法已在式3.7介绍，原始模型的最终损失函数为负对数似然损失（Negative Log-Likelihood Loss, NLL），其形式如式4.1所示。而在加入扩充文本之后，预测的目标仍是原标题文本的标签序列，因此其损失函数可以表示为式4.2所示。

$$L_{NLL} = -log(p(Y_{title}, X_{title})), \tag{4.1}$$

$$L_{NLL-exp} = -log(p(Y_{title}, X_{expand})), \tag{4.2}$$

其中 $p()$ 表示向量经过 CRF 层采用极大似然法计算的实体序列概率。

4.2 实验与讨论

4.2.1 数据准备和实验设置

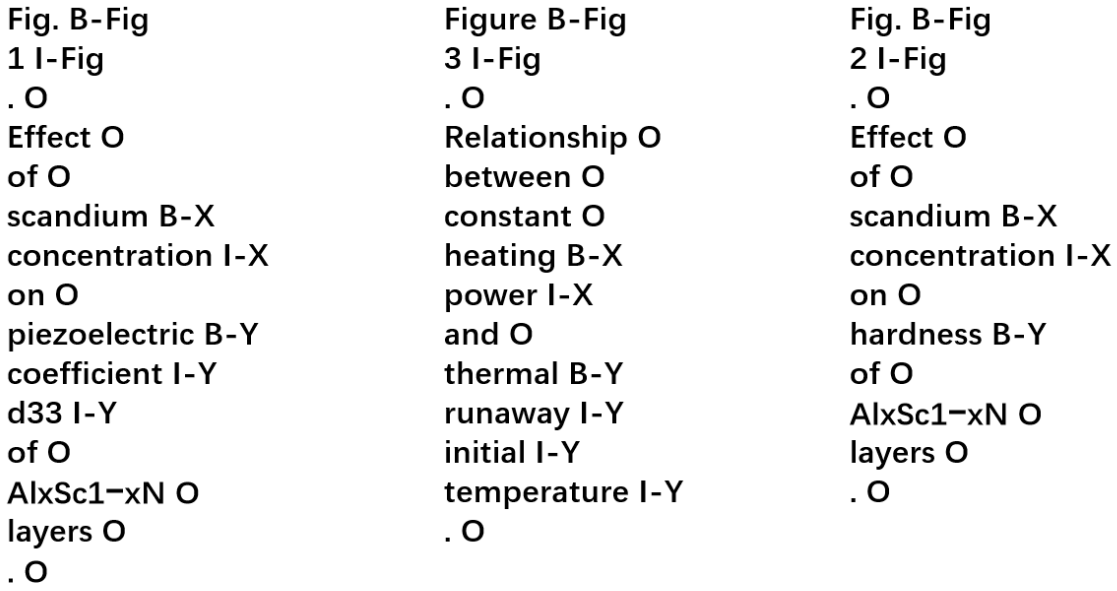


图 4.5: 标题坐标轴实体识别数据集示例

本章的数据集基于第三章的 NER 任务数据集，数据形式如图4.5所示，采用（单词，标签实体）的形式对每一句标题语句进行标注并保存，以相同的比例划分训练集和测试集用来与之前的识别效果做对比。在扩充数据集时，只对 3.2.1 节所标注的 756 幅单一数值图图片进行识别扩充。

所有实验均在 Intel(R) Core(TM) i7-10700 CPU 2.9GHz*16 和 32G RAM 的计算机上进行。实验参数设置如下：语句最大长度为 256，batch size 为 32，训练轮次为 5，初始学习率为 0.00005，丢弃率为 0.01。

4.2.2 标题扩充文本寻找结果

本节首先比较了对于式 3.4 中不同 λ 的取值匹配的数值图对应的正文描述语句的精确率，以 0.1 为间隔分别进行了十次测试。其结果如图4.6所示，可以发现单一

的相似度指标匹配出的文本的准确性比复合的相似度匹配出的文本的准确率低，其中仅使用 Jaccard 相似度（$\lambda = 1$ 时）的准确性只有 76.53%。这是因为其只考虑了单词的比例，忽略了语句的语义以及结构信息，而本任务与上一章匹配任务不同，需要匹配的正文描述语句是对数值图的详细描述，与数值图标题在词频上仅有较小的重合。因此在本任务上 Jaccard 相似度相较于余弦相似度效果更差（80.20%），余弦相似度更多关注的是语句之间的语义相似性。根据实验结果，本章方法中匹配数值图描述语句任务的 λ 值设置为 0.4。

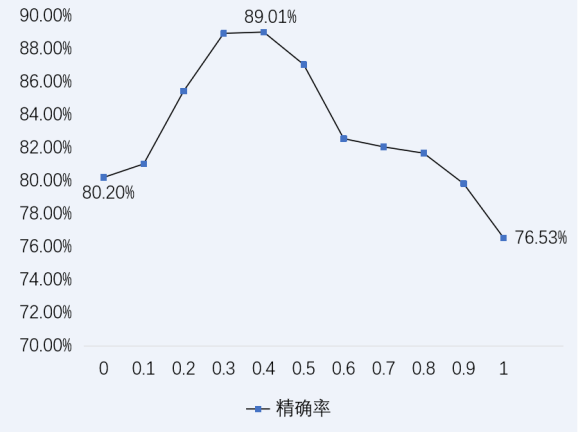


图 4.6: 不同 λ 对寻找扩充文本的影响效果

此外，本节还定性比较了对于同一标题不同 λ 匹配到的扩充文本的情况，对比结果如表4.2所示。可以发现，仅使用余弦相似度或 Jaccard 相似度匹配到的语句都会有偏差，例如案例 2 中，仅使用余弦相似度（$\lambda = 0$ 时）匹配出的文本甚至是科学文献中的参考文献内容，与正确的数值图描述语句相差甚远；而仅使用 Jaccard 相似度（$\lambda = 1$ 时）匹配出的语句仅仅是在单词重复率上接近，也未匹配成功。这是因为两个相似度都存在一定的局限性，余弦相似度仅计算了语句向量的距离，不能很好地代表语义相关性；而 Jaccard 只是计算了语句间的单词重复率。在用复合指标的情况下（$\lambda = 0.4$ 时），在 3 个案例中都能寻找出准确的数值图在正文中的描述语句。

表 4.2: 寻找扩充文本的定性结果，文本数据来源于科学文献 [93, 94, 95]。

案例 1	Fig. 11. Stress-strain curve for printed CA scaffolds.
λ=0	For viscoelastic materials such as CA, there is a lag in the strain response.
λ=1	Fig. 11 shows the results (stress vs. strain plots), where a linear region, indicative of elastic behavior is evident.
λ=0.4	Fig. 11 shows the results (stress vs. strain plots), where a linear region, indicative of elastic behavior is evident.
案例 2	Fig. 2. Effect of graphene nanosheet on wear at steel-cast iron contact as additive in PAO4 base oil.
λ=0	Zhu, Investigation of the tribology behaviour of the graphene nanosheets as oil additives on textured alloy cast iron surface, Appl.
λ=1	investigated the effects of graphene nanosheets in oil at a contact between a GCr15 steel ball on an RTCr2 alloy cast iron plate [61].
λ=0.4	Wear was reduced 50% after the addition of graphene nanosheets in PAO base oil, and up to 90% on the textured surfaces (Fig. 2)
案例 3	Fig. 5. Absorbance spectra as a function of wavelength for Co-Al LDH, Co-Al/G LDH, Co-Al/A LDH and Co-Al/U LDH thin films.
λ=0	Fig. 5 depicts the absorption spectra of Co-Al LDH thin films, Co-Al/G LDH, Co-Al/A LDH and Co-Al/U LDH on the absorption spectra optical properties.
λ=1	Fig. 6. Transmittance spectra as a function of wavenumbers for Co-Al LDH, Co-Al/G LDH, Co-Al/A LDH and Co-Al/U LDH thin films.
λ=0.4	Fig. 5 depicts the absorption spectra of Co-Al LDH thin films, Co-Al/G LDH, Co-Al/A LDH and Co-Al/U LDH on the absorption spectra optical properties.

4.2.3 命名实体识别结果

在前一章的坐标轴实体识别任务结果中发现，模型对于图片标号的识别精度非常高，而坐标轴实体的识别准确性较低，因此本节在统计识别结果时只针对 X 轴实体和 Y 轴实体两个类别。本节首先验证了本章的数据集扩充方法的有效性，对比了不同的文本数据增强方法在 Lstm、Bert 以及 Sci-Bert 上的识别效果，三种模型分别代表了循环神经网络、公共领域文本的预训练模型以及科学文本的预训练模型，具

有较强的代表性。此外，每个模型都添加了 CRF 层以优化预测的序列标签，每一种增强方法仅应用在训练集上，且将每一种增强方式生成的样本数量规定为 500，保证训练规模一致。结果如表4.3所示，其中标签替换表示对于已标注的 X 轴和 Y 轴实体，随机选择另一标注好的标题文本中的 X 轴和 Y 轴实体进行替换；而同义词替换表示对于已标注的坐标轴实体，从 WordNet[96] 检索同义词并进行替换。可以发现标签替换的方式可以缓解数据量不足的问题，提升模型的识别效果，但其替换的范围仅在训练集内部，使得模型无法进一步学习新的知识；而同义词替换的方式在 Lstm 和 Sci-Bert 上效果反倒不如原始模型，这是因为在材料文献中，一些专业名词是特有的，不可替换的，用语义上的同义词来替换它是不合理的，增加了生成假样本的风险；通过识别数值图图上文本标签填充模板的增强方式在 Lstm 和 Sci-Bert 上都达到了最好的提升，文本标签在一些情况下会与标题中的坐标轴实体不同，因此通过该方法可以提升模型的泛化能力。

表 4.3: 不同文本增强数据方法对模型识别效果的影响

模型	增强方法	Macro-P	Macro-R	Macro-F1
Lstm	原始数据	0.5416	0.6236	0.5797
	标签替换	0.5607	0.6331	0.5947
	同义词替换	0.5104	0.6082	0.5550
	我们的方法	**0.5866**	**0.6592**	**0.6209**
Bert	原始数据	0.6731	0.7388	0.7044
	标签替换	0.6804	**0.7593**	0.7177
	同义词替换	**0.6957**	0.7467	**0.7201**
	我们的方法	0.6883	0.7531	0.7192
Sci-Bert	原始数据	0.7634	0.8126	0.7871
	标签替换	0.7791	0.8393	0.8081
	同义词替换	0.7583	0.8019	0.7759
	我们的方法	**0.8033**	**0.8451**	**0.8237**

接着验证了本方法在坐标轴实体识别任务中加入扩充文本对于模型识别效果的影响，对比了 Lstm、Bert 和 Sci-Bert 在加入扩充文本前后的模型识别效果。实验

中，在训练与测试过程中设置了两种不同的条件，即带有扩充文本（w/ text）与不带有扩充文本（w/o text）。每个模型在训练时都进行了4.1.3节的数据扩充，且在模型之后都加入了 CRF 层来优化预测序列，其结果如表4.4所示。可以发现，当训练阶段没有加入扩充文本，而预测阶段加入扩充文本时，相较于原始模型效果有所提升。这证明了扩充文本即使没有参与训练，其在预测阶段对于标题生成的向量表征的上下文影响也有助于提升识别精度；而当训练时加入扩充文本且预测时不使用时，模型的效果反而略有所下降，在三个模型上的平均 F1 得分都下降了 1% 以上；而当在训练与预测阶段同时加入扩充文本时，模型的识别效果会达到最佳，在 Sci-Bert 上的平均 F1 得分达到了 84.2%。

表 4.4: 扩充文本对于坐标轴实体识别效果的影响

模型	训练方法	测试方法					
		w/o text			w/ text		
		Macro-P	Macro-R	Macro-F1	Macro-P	Macro-R	Macro-F1
Lstm	w/o text	0.5866	0.6592	0.6209	0.5911	0.6628	0.6249
	w/ text	0.5794	0.6386	0.6076	0.5937	0.6713	0.6301
Bert	w/o text	0.6883	0.7531	0.7192	0.6895	0.7601	0.7231
	w/ text	0.6693	0.7201	0.6938	0.7013	0.7776	0.7375
Sci-Bert	w/o text	0.8033	0.8451	0.8237	0.8087	0.8621	0.8345
	w/ text	0.7911	0.8164	0.8036	**0.8179**	**0.8676**	**0.8420**

表4.5展示了坐标轴实体识别的定性结果，以 Sci-Bert+CRF 为基础模型，对比了在加入扩充文本前后的坐标轴实体识别结果，其中“w/o”和“w/”分别表示不加入和加入扩充文本两种方式。可以发现，在加入了扩充文本之后，模型识别出的坐标轴实体更全面。例如在第一句标题中，改进后的模型识别出的 Y 轴实体相较于原始模型更加准确，并且弥补了之前 X 轴实体未被识别出的问题。在第三句标题中改进的模型识别出的 X 轴实体比原始结果更加具体。此外，改进后的模型对于一些材料领域的专业数值图的坐标轴名称也能识别准确，比如在第二句标题文本中的“XRD”代表的是衍射图谱，因此正确的坐标轴实体仅为“XRD”，而原始模型将语句中的“ ”识别为 X 轴实体，改进后的模型修改了这一错误。

表 4.5: 坐标轴实体识别结果实例

标题文本	X 轴		Y 轴	
	w/o	w/	w/o	w/
Fig. 9. State and estimate evolution under noisy demand.	-	noisy demand	State	State and estimate evolution
Fig. 7. XRD profifiles for each build orientation indicating significant α content with very small amounts of β phase.	α	-	XRD	XRD
Fig. 8. Shrinkage of the mortar as a function of age in days.	age	age in days	Shrinkage	Shrinkage

为了验证通过输入语句的上下文联系可以提升 NER 模型的识别效果，本文分别在 WNUT16[97]、WNUT17[98] 以及 CoNLL-03[99] 数据集上进行了实验。三个数据集的统计信息如表4.6所示，其中 WNUT16 和 WNUT17 来源于社交媒体的用户评论，实体类别包括公司名、人名、产品名等；而 CoNLL-03 数据集是 NER 最流行的数据集，实体类别包括人名、地名、组织名以及杂项实体。本实验采用 Bert 来生成语句向量以适应公共文本下的 NER 任务，加入 CRF 来优化预测标签，并以 WikiText [100] 数据集作为搜索扩充文本的资源，在文本数据集中以相似度最高的语句作为扩充文本。最终在三个数据集上的 NER 结果如表4.7所示，以 Bert+CRF 识别的 F1 得分作为 baseline，Nguyen 等人的方法 [101] 使用 RoBert 在推特文本上进行了训练，因此其效果较 Bert 有所提升；而 Nie 等人的方法 [102] 引入了额外的词级别的特征用来缓解数据集稀疏的问题，且他们的增强方法仅针对社交评论数据集，因此在本实验中只测试了在 WNUT16 和 WNUT17 数据集的效果，其在 WNUT16 数据集上的效果优于 Nguyen 等人的方法，提升了 3%；Li 等人的方法 [103] 将 Dice 损失代替标准的交叉熵损失，以优化 NER 数据不平衡问题，该方法在三个数据集上的效果都有所提升；而 LUKE 模型 [87] 是一种基于 Bert 的实体感知模型，它采用新的预训练任务来训练，通过训练预测被掩盖的实体，从而提升预训练模型在实体识别上的效果，其在三个数据集上都取得了非常好的效果；在本方

法中，模型训练与测试阶段都加入了扩充文本以保证识别效果达到最高，最终的结果优于 LUKE 模型，且在数据量相对少的 WNUT16 和 WNUT17 中提升明显，在 WNUT17 中提升了 5%。

表 4.6: 3 种公共数据集的数据统计

	训练集	测试集	实体类别
WNUT16	2394	3849	10
WNUT17	3394	1287	6
CoNLL-03	14987	3684	4

表 4.7: 不同模型在公开数据集上的识别结果（以 F1 得分展示）

	WNUT16	WNUT17	CoNLL-03
Bert+CRF	49.52%	53.74%	90.06%
Nguyen 方法	52.10%	56.50%	91.14%
Nie 方法	55.01%	50.36%	-
Li 方法	53.03%	54.29%	**93.33**%
LUKE	54.04%	55.22%	92.42%
本方法	**56.34**%	**60.02**%	93.27%

4.3 本章小结

本章提出了一种用于提升数值图标题坐标轴实体识别效果的方法，包括通过识别数值图坐标轴的标签文本并利用模板的方式自动化生成数据集以及通过加入扩充文本增强标题语句生成向量的准确性。该方法可以有效地提升原有模型在坐标轴命名实体识别任务上的效果，从而使得第三章的材料文献挖掘方法更加准确。本方法首先利用材料科学文献中数值图标题文本与数值图图片的关联性，使用 paddle OCR 识别出数值图图上坐标轴附近的文本标签并进行清理，同时，根据之前标注好的数据集设立模板，将清理好的实体填充入模板以达到扩充数据集的效果。实验证明通过这种方式扩充数据集相较于同义词替换、标签替换等方法来得更有效，可以让模型学到更多正确有用的信息。其次，依靠科学文献中正文部分与数值图图片的联系，

找出正文中对数值图的描述语句作为标题的扩充文本，将二者拼接送入 NER 任务中，通过预训练模型的自注意力机制使得生成的语句向量更关注于有用的信息，从而更准确地预测出坐标轴实体。此外，通过在公共数据集上的测试证明了通过增强上下文关联可以有效地提升识别的准确性。希望该工作能够推动 NER 的研究，从而推动自然语言处理的发展。

虽然本方法有效地提升了数值图标题坐标轴实体识别的效果，但其还存在着一定的缺陷。通过数值图图片上的坐标轴文本标签来扩充数据集具有一定的局限性，对于其他 NER 任务是不适用的。另外，通过加入扩充文本的方法来提升模型的效果，需要花费时间从额外的文本中找出关联语句，这在一定程度上弱化了本方法的工业化程度，且对于特定领域的任务需要提前准备好充足的文本数据以供搜寻相似文本也是非常困难的。因此，本方法在实用性上还有很大的改进空间。

第五章　材料科学文献数值图信息提取软件设计与实现

文献挖掘可以帮助学者快速地掌握领域的发展趋势，通过自动化地挖掘大规模的文献信息，学者可以方便地获得研究数据，从而推动行业发展。而在材料科学文献中，数值图的信息挖掘可以获取科学文献中关键的实验数据信息，包括具体数值与描述对象等信息，从而避免重复实验，节省人力成本。本章通过开发一个离线的自动化提取材料科学文献数值图信息的软件，可以为材料学者展开自身研究提供便利。

本章节在前面两个章节的基础上，对材料科学文献数值图信息提取软件进行具体的设计与实现，使得使用者只需输入 PDF 格式的材料科学文献，即可获取科学文献中数值图的数值信息、描述对象等信息。

5.1　开发环境

本软件的开发是基于 Windows 操作系统，主要是利用 Python 语言进行开发，利用 PyQt5 实现软件的可视化界面。PyQt5 是由一组 Python 模块构成的图形界面开发库。PyQt5 提供了一个设计良好的窗口控件集合，具有很好的移植性，可以在多个操作系统上运行。为了实现软件的本地化运行，使用 PyInstaller 模块进行软件的打包，PyInstaller 是一个 Python 第三方库，它可以将使用 Python 语言开发的项目打包成独立可执行程序，用户无需安装 Python 解释器和项目环境即可运行。

5.2　需求分析

本软件主要目的是向学者提供一个快速且直观的材料科学文献数值图信息挖掘接口，其主要功能需求包括查看科学文献、截取科学文献中单一数值图、提取数值图信息以及批处理四个功能，其用例图如图5.1所示。

查看科学文献功能：学者输入单篇材料科学文献后，软件界面中将会展示该篇科学文献，并提供页面切换功能，以供学者浏览整篇科学文献。

截取科学文献中单一数值图：学者输入单篇材料科学文献后，该功能将会调用训练好的检测模型来截取该篇科学文献中的所有单一的数值图图片，并将截取的图片展示在界面上以供使用者查阅。

提取数值图信息：学者在一篇科学文献中截取完单一的数值图图片后，调用该功能可以实现数值图的信息提取，包括数值图的标题语句和在正文部分的描述文本，数值图的数据信息及其描述对象（坐标轴名称）。

批处理：调用截取科学文献中单一数值图和提取数值图信息功能，实现批量材料科学文献的自动化处理。

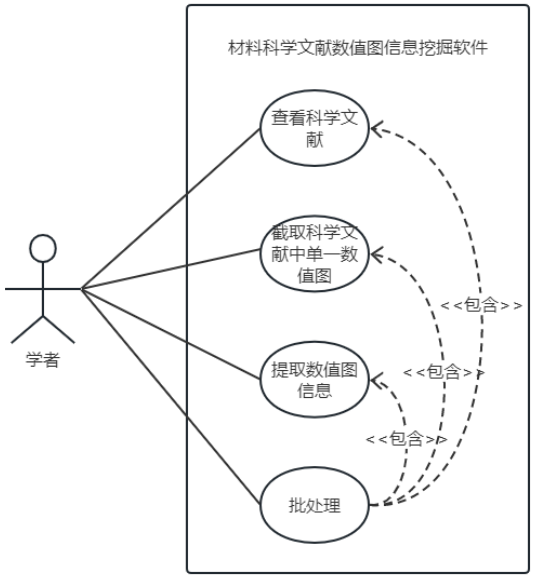


图 5.1: 软件需求分析用例图

5.3 软件架构设计

本节将对系统的架构设计进行详细阐述。系统的整体架构设计如图5.2所示，系统的结构可以划分为三个层次：应用层、服务层和数据层。应用层主要负责提供用户友好的可视化操作界面，以及实现用户使用功能的接口。服务层主要负责处理用户的操作请求，并根据请求调用相应的功能模块，最终返回用户期望的结果信息。数据层主要负责存储和管理操作过程中涉及到的数据信息。

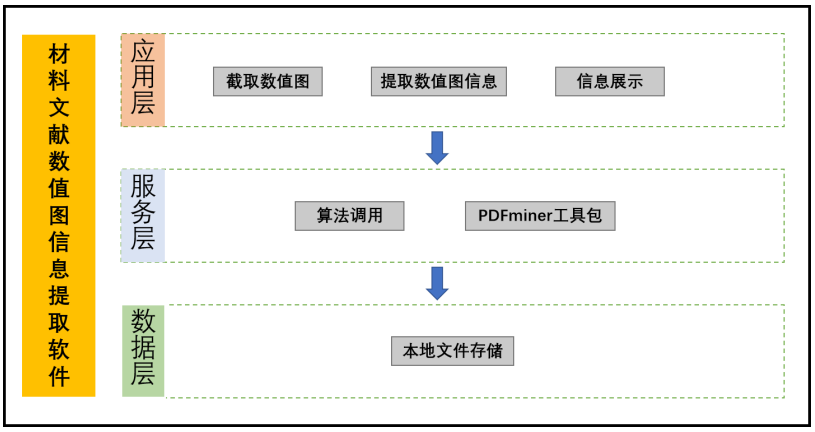


图 5.2: 软件架构图

为了实现用户的离线使用，系统的应用层设计为本地可执行程序的形式，用户在安装软件后即可直接运行程序，无需依赖网络连接和配置运行环境等。用户在运行程序后，可以通过可视化界面进行各项操作，如截取科学文献中单一数值图和提取数值图信息等，并且最终结果也会在界面中呈现。

系统的服务层主要负责实现算法的调用，包括数值图图片数据提取算法和数值图标题坐标轴实体识别算法。此外，在处理 PDF 格式的材料科学文献时，需要借助 PDFminer 工具包对文件进行解析，以获取所需的文献文本内容。

系统的数据层主要负责实现对于算法输出结果的保存，包括数值图图片、数值图数据、数值图标题和数值图坐标轴名称等，所有数据均存储在本地磁盘中，以便学者获取和存储结果。

5.4 软件实现

5.4.1 界面设计

软件界面主要分为四个区域。如图5.3所示，区域 1 主要用来展示用户输入的 PDF 格式的材料科学文献，用户可以通过下方的按键自由切换文献页面来浏览科学文献。区域 2 主要用来显示操作过程中的日志记录，以便用户更详细地知晓中间过程和完成进度，如处理进度、数值图提取信息的中间结果等。区域 3 主要是用户与

软件的交互按钮，用户通过按钮请求不同的功能请求，包括加载 PDF 格式的材料科学文献和提取科学文献中的数值图信息。区域 4 主要是用来显示从科学文献中截取的数值图图片，用户可以浏览所有截取的数值图图片，并按照需求点击按钮提取数值图的信息，而提取的数值图信息等结果会保留在本地文件夹中，用户可以通过按钮进入文件夹下查看所有结果。

图 5.3: 软件初始界面

5.4.2 整体流程

软件的整体使用流程如图5.4所示，用户通过点击区域 3 的“打开 PDF”按钮选择需要处理的 PDF 材料科学文献，在选择文件时仅支持 PDF 格式文件或文件夹（批处理时处理文件夹下的所有 PDF 文件）。如图5.5（a）所示，在选择完成后，软件加载该篇科学文献并在区域 1 显示，同时区域 2 会显示加载日志，如图5.5（b）所示。

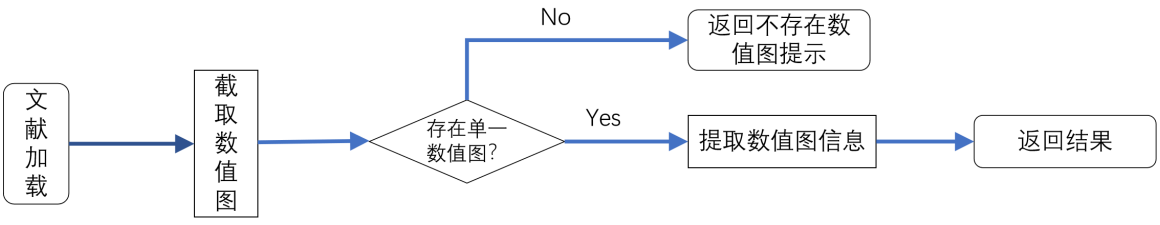
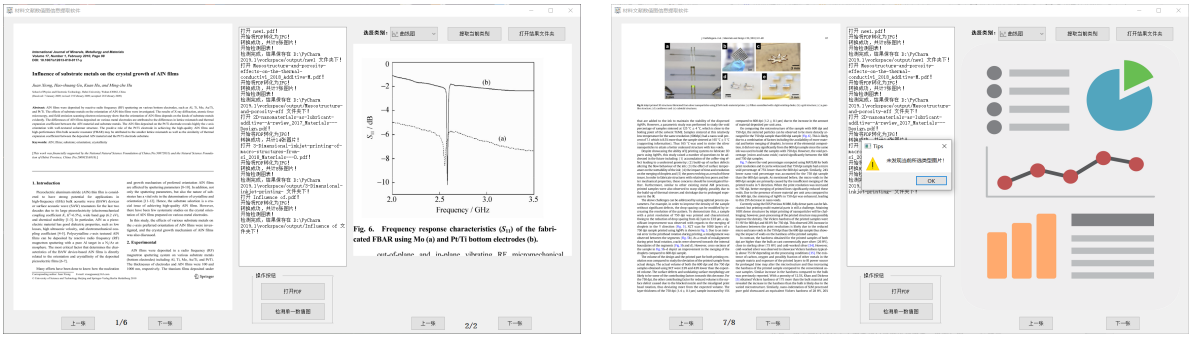


图 5.4: 软件操作整体流程

（a）选择 PDF 文件界面　　（b）文献加载界面

图 5.5: 软件文献加载图示

在加载文献后，用户可以点击“检测单一数值图”按钮，软件会调用 Yolov5s 模型对每一页文献进行检测，并将检测出的数值图图片显示在区域 4，处理的结果日志也会在区域 2 中显示，如图5.6（a）所示。而当未检测到数值图时，软件会弹出对话框提示用户，如图5.6（b）所示。



（a）数值图检测成功图示　　（b）数值图检测失败图示

图 5.6: 软件数值图检测图示

对于检测出的数值图图片，用户可以点击区域 4 的“提取当前类别”按钮进行所有数值图图片信息的提取。当提取完成后，软件会以对话框的方式提示用户任务完成，并自动打开结果文件夹供用户查看结果。结果文件夹将会包括单一数值图图片及其对应的数据结果文件夹，数据结果将以 excel 文件形式保存，图上每一种颜色的曲线对应一个 excel 文件，文件中的数据将以（X，Y）的格式记录，并且会在首行显示数值图对应的 X 轴和 Y 轴坐标轴名称，如图5.7所示。而对于信息提取失

败的数据图，其对应的文件夹内容将为空。

图 5.7: 软件数值图信息提取图示

批处理功能要求用户在选择处理文献时选择文件夹即可实现，软件将会自动处理文件夹下所有 PDF 格式的文件。此外，在处理完每一篇科学文献后，软件会按条记录截取的数值图条目，包括文件名、数值图标题、数值图保存路径及其对应的坐标轴名称，处理记录的汇总也方便了用户进行整理和分析。

5.5 本章小结

本节使用 Python 程序设计语言开发了材料科学文献数值图信息提取软件，将数值图截取算法和信息提取算法集成进软件中使得用户可以方便地使用。本章主要从开发环境、需求分析、软件架构设计和软件实现四个方面对软件进行了展开分析，证明了软件的实用性和有效性。

第六章　总结与展望

6.1　结论

材料科学文献中多元化的数据类型导致从文献中挖掘信息难度较大，以往的材料文献挖掘方法仅针对单一的数据类型，无法将不同类型数据之间的信息整合起来。从多种数据类型中提取信息并将其整理关联，有助于材料文献挖掘的进一步发展，推动领域进步，促进新兴材料的研发。本文综合目标检测技术、自然语言处理、机器学习等领域知识，重点研究了利用计算机技术实现材料科学文献中数值图图文信息挖掘的方法，并针对材料文本数据量少和标注困难的问题，抓住材料科学文献图文之间的联系以提升挖掘信息的准确性。本文研究成果及贡献如下：

（1）针对材料科学文献中数值图对应的标题文本和图片，本文第三章分别从这两个类型的数据中挖掘信息。通过 Yolov5s 截取文献中数值图图片，并提出改进的科学文献图片检测方法来提升截取精度，使用图像处理等知识还原数值图真实数值信息；使用 Sci-Bert+CRF 模型来识别对应数值图标题语句中的坐标轴实体，并将二者信息结合获得丰富的数值图数据信息。该方法实现了多元数据类型的联合挖掘，打造了材料文献挖掘的新形式。

（2）为了提升坐标轴实体识别的效果，本文第四章通过科学文献中数值图图上标签文本及正文中数值图的描述语句，提出了一种通过模板填充来扩充数据集的方法，达到数据增强的效果，并将正文描述语句作为 NER 的扩充文本，通过预训练模型中的自注意力机制来强化语句的上下文关系，从而增强输入语句的向量表征的正确性，提升模型的识别效果。

（3）将算法和模型封装，以可视化软件的形式提供给学者使用，学者在本地无需联网和安装运行环境即可实现大规模的材料科学文献中数值图信息挖掘，从而推动材料领域的发展，加快新型材料的研发。

本论文对以上方法都进行了实验，实验结果表明这些方法具有较好的准确率以及鲁棒性，具有较高的实用价值。这为快速提取材料科学文献中的实验数据、工艺

数据等关键信息提供了帮助，为后续的材料基因工程提供数据基础。

6.2 展望

本论文的方法能够有效地从图文两个类型的数据中提取材料科学文献数值图的信息，但是由于材料科学文献格式的多样性，科学文献中数值图绘制风格和文本描述风格不一致，导致了目前方法仍然有很大的局限性。为了进一步提高本论文方法的有效性，本论文可以在以下几个方面进一步深入研究：

（1）优化数值图数值提取的方法。基于基础的图像处理技术提取的数值图数据信息准确率偏低且该方法适用性较差。因此利用深度学习进行数据提取的方式亟需发展，让计算机自主学习数值图的特征可以提高数值信息提取的准确性。

（2）融合多种数据类型以提升材料文献挖掘的全面性。材料科学文献中的数据类型包括图片、表格和文本，而图片中又包含数值图、材料图像等类别，仅仅从数值图和文本两个方面提取的信息不足以代表整篇文献。因此将多元化的数据类型中提取的信息进行整合，从而提升挖掘信息的全面性。

（3）充分利用提取的数据信息进行材料的分析及预测。提取完的大规模数据需要得到有效地利用才能真正地推动材料学科的发展，这就需要将领域知识和机器学习方法相结合。

（4）引入多模态学习技术。本论文的方法虽然结合了图文两种类型数据，但都是从单一类型中提取后再进行结合，多模态的学习方式可以更好地融合不同模态之间的知识，因此可以将基于多模态学习的技术应用于材料文献提取，结合非结构化文本、图像和表格等模态的数据，更好地完成提取任务。

以上优化的方向仍需要不断的探索和实验才能取得一定的成果，材料文献挖掘的未来任重且道远。

参考文献

[1] 于成丽, 胡万里, and 刘阳, "美国发布新版《国家人工智能研究与发展战略计划》," 保密科学技术, vol. 108, no. 2, pp. 35–37, 2019.

[2] 宿彦京, 付华栋, 白洋, 姜雪, and 谢建新, "中国材料基因工程研究进展," 金属学报, vol. 56, no. 10, pp. 1313–1323, 2020.

[3] 王慧芳, 曹靖, and 罗麟, "电力文本数据挖掘现状及挑战," 浙江电力, vol. 38, no. 3, pp. 1–7, 2019.

[4] X. Li, L. Lu, J. Li, X. Zhang, and H. Gao, "Mechanical properties and deformation mechanisms of gradient nanostructured metals and alloys," *Nature Reviews Materials*, vol. 5, no. 9, pp. 706–723, 2020.

[5] Z.-Z. Hu, M. Narayanaswamy, K. Ravikumar, K. Vijay-Shanker, and C. H. Wu, "Literature mining and database annotation of protein phosphorylation using a rule-based system," *Bioinformatics*, vol. 21, no. 11, pp. 2759–2765, 2005.

[6] 吴刚勇, 张千斌, 吴恒超, and 顾冰, "基于自然语言处理技术的电力客户投诉工单文本挖掘分析," 电力大数据, vol. 21, no. 10, pp. 68–73, 2018.

[7] F. Zhou, Y. Zhao, W. Chen, Y. Tan, Y. Xu, Y. Chen, C. Liu, and Y. Zhao, "Reverse-engineering bar charts using neural networks," *Journal of Visualization*, vol. 24, no. 6, pp. 419–435, 2021.

[8] 林立涛 and 王东波, "古籍文本挖掘技术综述," 科技情报研究, vol. 5, no. 7, pp. 78–91, 2023.

[9] S. Reddy, R. Bhaskar, S. Padmanabhan, K. Verspoor, C. Mamillapalli, R. Lahoti, V. Makinen, S. Pradhan, P. Kushwah, and S. Sinha, "Use and validation of text mining and cluster algorithms to derive insights from corona virus disease-2019 (covid-19) medical literature.," *Computer methods and programs in biomedicine update*, vol. 1, no. 12, p. 100010, 2021.

[10] 郭金龙, 许鑫, and 陆宇杰, "人文社会科学研究中文本挖掘技术应用进展," 图书情报工作, vol. 56, no. 08, p. 10, 2012.

[11] 郭继光 and 黄胜, "基于大数据的军事情报分析与服务系统架构研究," 中国电子科学研究院学报, vol. 12, no. 4, pp. 389–393, 2017.

[12] P. Zerbino, A. Stefanini, and D. Aloini, "Process science in action: A literature review on process mining in business management," *Technological Forecasting and Social Change*, vol. 172, no. 3, p. 121021, 2021.

[13] O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti, and G. Ceder, "Opportunities and challenges of text mining in materials research," *Iscience*, vol. 24, no. 3, p. 102155, 2021.

[14] D. Rebholz-Schuhmann, A. Oellrich, and R. Hoehndorf, "Text-mining solutions for biomedical research: enabling integrative biology," *Nature Reviews Genetics*, vol. 13, no. 12, pp. 829–839, 2012.

[15] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS biology*, vol. 2, no. 11, p. e309, 2004.

[16] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno, "Text processing through web services: calling whatizit," *Bioinformatics*, vol. 24, no. 2, pp. 296–298, 2008.

[17] N. Kang, B. Singh, Z. Afzal, E. M. van Mulligen, and J. A. Kors, "Using rule-based natural language processing to improve disease normalization in biomedical text," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 876–881, 2013.

[18] H. Sampathkumar, X.-w. Chen, and B. Luo, "Mining adverse drug reactions from online healthcare forums using hidden markov model," *BMC medical informatics and decision making*, vol. 14, no. 1, pp. 1–18, 2014.

[19] M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. Ureña-López, "Experiments with svm to classify opinions in different domains," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14799–14804, 2011.

[20] R. Vazquez Guillamet, O. Ursu, G. Iwamoto, P. L. Moseley, and T. Oprea, "Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records," *Health informatics journal*, vol. 24, no. 4, pp. 394–409, 2018.

[21] L. Kurniasari and A. Setyanto, "Sentiment analysis using recurrent neural network," *Journal of Physics: Conference Series*, vol. 1471, no. 1, pp. 12–18, 2020.

[22] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[23] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain, "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," *Journal of chemical information and modeling*, vol. 59, no. 9, pp. 3692–3702, 2019.

[24] F. Ali, S. El-Sappagh, and D. Kwak, "Fuzzy ontology and lstm-based text mining: a transportation network monitoring system for assisting travel," *Sensors*, vol. 19, no. 2, p. 234, 2019.

[25] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[26] X. Zhao, J. Greenberg, Y. An, and X. Hu, "Fine-tuning bert model for materials named entity recognition," in *2021 IEEE International Conference on Big Data*, (Los Alamitos, CA, USA), pp. 3717–3720, IEEE Computer Society, Dec. 2021.

[27] H. Zhou, W. Huang, M. Li, and Y. Lai, "Relation-aware entity matching using sentence-bert," *CMC-Comput. Mater. Contin*, vol. 71, no. 2, pp. 1581–1595, 2022.

[28] D. Zhao, J. Wang, H. Lin, Y. Chu, Y. Wang, Y. Zhang, and Z. Yang, "Sentence representation with manifold learning for biomedical texts," *Knowledge-Based Systems*, vol. 218, no. 5, p. 106869, 2021.

[29] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*,

(Hong Kong, China), pp. 3615–3620, Association for Computational Linguistics, Nov. 2019.

[30] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[31] A. Lamproudis, A. Henriksson, and H. Dalianis, "Evaluating pretraining strategies for clinical bert models," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (Marseille, France), pp. 410–416, European Language Resources Association, June 2022.

[32] A. Kumar, S. Ganesh, D. Gupta, and H. Kodamana, "A text mining framework for screening catalysts and critical process parameters from scientific literature-a study on hydrogen production from alcohol," *Chemical Engineering Research and Design*, vol. 184, no. 2, pp. 90–102, 2022.

[33] J. Li, B. Chiu, S. Feng, and H. Wang, "Few-shot named entity recognition via meta-learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 9, pp. 4245–4256, 2020.

[34] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of big Data*, vol. 8, no. 3, pp. 1–34, 2021.

[35] H. Ye, N. Zhang, S. Deng, X. Chen, H. Chen, F. Xiong, X. Chen, and H. Chen, "Ontology-enhanced prompt-tuning for few-shot learning," in *Proceedings of the ACM Web Conference 2022*, (New York, NY, USA), pp. 778–787, Association for Computing Machinery, Apr. 2022.

[36] U. E. Chigbu, "Visually hypothesising in scientific paper writing: Confirming and refuting qualitative research hypotheses using diagrams," *Publications*, vol. 7, no. 1, p. 22, 2019.

[37] I. Kavasidis, C. Pino, S. Palazzo, F. Rundo, D. Giordano, P. Messina, and C. Spampinato, "A saliency-based convolutional neural network for table and chart detection in digitized documents," in *Image Analysis and Processing–ICIAP 2019: 20th International Confer-*

ence, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20, (Cham, Switzerland), pp. 292–302, Springer International Publishing, Sept. 2019.

[38] N. Siegel, N. Lourie, R. Power, and W. Ammar, "Extracting scientific figures with distantly supervised neural networks," in *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, (New York, NY, USA), pp. 223–232, Association for Computing Machinery, May 2018.

[39] K. Davila, S. Setlur, D. Doermann, B. U. Kota, and V. Govindaraju, "Chart mining: A survey of methods for automated chart analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3799–3819, 2020.

[40] S. Thakare, A. Kamble, V. Thengne, and U. Kamble, "Document segmentation and language translation using tesseract-ocr," in *2018 IEEE 13th International Conference on Industrial and Information Systems*, (Rupnagar, India), pp. 148–151, IEEE, Dec. 2018.

[41] W. Dai, M. Wang, Z. Niu, and J. Zhang, "Chart decoder: Generating textual and numeric information from chart images automatically," *Journal of Visual Languages & Computing*, vol. 48, no. 12, pp. 101–109, 2018.

[42] X. Lu, J. Wang, P. Mitra, and C. L. Giles, "Automatic extraction of data from 2-d plots in documents," in *Ninth International Conference on Document Analysis and Recognition*, (Curitiba, Brazil), pp. 188–192, IEEE, Nov. 2007.

[43] A. Kaur, D. Dani, and N. Mishra, "Improving web accessibility of graphs for visually impaired," *Int. J. Comput. Sci. Inf. Technol*, vol. 2, no. 5, pp. 1979–1981, 2011.

[44] W. Huang, C. L. Tan, and W. K. Leow, "Associating text and graphics for scientific chart understanding," in *Eighth International Conference on Document Analysis and Recognition*, (Seoul, Korea (South)), pp. 580–584, IEEE, Aug. 2005.

[45] R. R. Nair, N. Sankaran, I. Nwogu, and V. Govindaraju, "Automated analysis of line plots in documents," in *2015 13th international conference on document analysis and recognition*, (Tunis, Tunisia), pp. 796–800, IEEE, Aug. 2015.

[46] S. Ray Choudhury, S. Wang, and C. L. Giles, "Curve separation for line graphs in scholarly documents," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital*

Libraries, (New York, NY, USA), pp. 277–278, Association for Computing Machinery, June 2016.

[47] N. Siegel, Z. Horvitz, R. Levin, S. Divvala, and A. Farhadi, "Figureseer: Parsing result-figures in research papers," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, (Cham, Switzerland), pp. 664–680, Springer International Publishing, Sept. 2016.

[48] O. Aydin and M. Y. Yassikaya, "Validity and reliability analysis of the plotdigitizer software program for data extraction from single-case graphs," *Perspectives on Behavior Science*, vol. 45, no. 1, pp. 239–257, 2022.

[49] S. Wang, L. Xu, Q. Wang, J. Li, B. Bai, Z. Li, X. Wu, P. Yu, X. Li, and J. Yin, "Postoperative complications and prognosis after radical gastrectomy for gastric cancer: a systematic review and meta-analysis of observational studies," *World journal of surgical oncology*, vol. 17, no. 3, pp. 1–10, 2019.

[50] M. Cliche, D. Rosenberg, D. Madeka, and C. Yee, "Scatteract: Automated extraction of data from scatter plots," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, (Cham Switzerland), pp. 135–150, Springer International Publishing, Dec. 2017.

[51] W. Chai and G. Wang, "Deep vision multimodal learning: Methodology, benchmark, and trend," *Applied Sciences*, vol. 12, no. 13, p. 6588, 2022.

[52] F. Zhou, Y. Zhao, W. Chen, Y. Tan, Y. Xu, Y. Chen, C. Liu, and Y. Zhao, "Reverse-engineering bar charts using neural networks," *Journal of Visualization*, vol. 24, no. 2, pp. 419–435, 2021.

[53] W. Wang, X. Jiang, S. Tian, P. Liu, D. Dang, Y. Su, T. Lookman, and J. Xie, "Automated pipeline for superalloy data by text mining," *npj Computational Materials*, vol. 8, no. 1, p. 9, 2022.

[54] Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, and E. Olivetti, "A machine learning approach to zeolite synthesis enabled by automatic literature data extraction," *ACS central science*, vol. 5, no. 5, pp. 892–899, 2019.

[55] I. Safder, H. Batool, R. Sarwar, F. Zaman, N. R. Aljohani, R. Nawaz, M. Gaber, and S.-U. Hassan, "Parsing auc result-figures in machine learning specific scholarly documents for semantically-enriched summarization," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2004347, 2022.

[56] F. Sultana, A. Sufian, and P. Dutta, "A review of object detection models based on convolutional neural network," *Intelligent computing: image processing based applications*, vol. 1157, no. 1, pp. 1–16, 2020.

[57] J. Yi, P. Wu, and D. N. Metaxas, "Assd: Attentive single shot multibox detector," *Computer Vision and Image Understanding*, vol. 189, no. 12, p. 102827, 2019.

[58] S. Zuo, Y. Xiao, X. Chang, and X. Wang, "Vision transformers for dense prediction: A survey," *Knowledge-Based Systems*, vol. 253, no. 11, p. 109552, 2022.

[59] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia Computer Science*, vol. 199, no. 1, pp. 1066–1073, 2022.

[60] C. Wang and C. Zhong, "Adaptive feature pyramid networks for object detection," *IEEE Access*, vol. 9, no. 3, pp. 107024–107032, 2021.

[61] H. Yu, X. Li, Y. Feng, and S. Han, "Multiple attentional path aggregation network for marine object detection," *Applied Intelligence*, vol. 53, no. 2, pp. 2434–2451, 2023.

[62] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, p. 1, 2020.

[63] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: State-of-the-art," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–39, 2021.

[64] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM computing surveys*, vol. 54, no. 3, pp. 1–40, 2021.

[65] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–35, 2021.

[66] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *PloS one*, vol. 14, no. 8, p. e0220976, 2019.

[67] S. M. Mohammed, K. Jacksi, and S. Zeebaree, "A state-of-the-art survey on semantic similarity for document clustering using glove and density-based algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 1, pp. 552–562, 2021.

[68] A. Al Badawi, L. Hoang, C. F. Mun, K. Laine, and K. M. M. Aung, "Privft: Private and fast text classification with homomorphic encryption," *IEEE Access*, vol. 8, no. 1, pp. 226544–226556, 2020.

[69] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature*, vol. 571, no. 7763, pp. 95–98, 2019.

[70] P. J. Worth, "Word embeddings and semantic spaces in natural language processing," *International Journal of Intelligence Science*, vol. 13, no. 1, pp. 1–21, 2023.

[71] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, no. 3, pp. 15908–15919, 2021.

[72] C. Sutton and A. Mccallum, "An introduction to conditional random fields," *Foundations and Trends® in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2010.

[73] T. De Smedt and W. Daelemans, "Pattern for python," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2063–2067, 2012.

[74] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant jaccard similarity," *Information Sciences*, vol. 483, no. 10, pp. 53–64, 2019.

[75] M. S. Zoromba and A. Al-Hossainy, "Doped poly (o-phenylenediamine -co- p-toluidine) fibers for polymer solar cells applications," *Solar Energy*, vol. 195, no. 3, pp. 194–209, 2020.

[76] M. Goyal, "Morphological image processing," *IJCST*, vol. 2, no. 4, p. 59, 2011.

[77] C. Sun, Y. Wang, M. D. McMurtrey, N. D. Jerred, F. Liou, and J. Li, "Additive manufacturing for energy: A review," *Applied Energy*, vol. 282, no. 5, p. 116041, 2021.

[78] Y. Shen, Y. Li, C. Chen, and H.-L. Tsai, "3d printing of large, complex metallic glass structures," *Materials & Design*, vol. 117, no. 2, pp. 213–222, 2017.

[79] P. A. Praczyk and J. Nogueras-Iso, "Automatic extraction of figures from scientific publications in high-energy physics," *Information Technology and Libraries*, vol. 32, no. 4, pp. 25–52, 2013.

[80] J. Younas, S. A. Siddiqui, M. Munir, M. I. Malik, F. Shafait, P. Lukowicz, and S. Ahmed, "Fi-fo detector: Figure and formula detection using deformable networks," *Applied Sciences*, vol. 10, no. 18, p. 6460, 2020.

[81] J. P. Naiman, P. K. G. Williams, and A. Goodman, "Figure and figure caption extraction for mixed raster and vector pdfs: Digitization of astronomical literature with ocr features," in *Linking Theory and Practice of Digital Libraries*, (Cham Switzerland), pp. 6442–6454, Springer International Publishing, Sept. 2022.

[82] X. Zhang, Z. Guo, C. Chen, and W. Yang, "Additive manufacturing of wc-20co components by 3d gel-printing," *International Journal of Refractory Metals and Hard Materials*, vol. 70, no. 2, pp. 215–223, 2018.

[83] V. Telgane, S. Godiganur, H. Srikanth, and S. Patil, "Performance and emission characteristics of a ci engine fueled with milk scum biodiesel," *Materials Today: Proceedings*, vol. 45, no. 5, pp. 284–289, 2021.

[84] K. Gandha, L. Li, I. Nlebedim, B. K. Post, V. Kunc, B. C. Sales, J. Bell, and M. P. Paranthaman, "Additive manufacturing of anisotropic hybrid ndfeb-smfen nylon composite bonded magnets," *Journal of Magnetism and Magnetic Materials*, vol. 467, no. 1, pp. 8–13, 2018.

[85] C. Bermudo, L. Sevilla, and G. Castillo López, "Material flow analysis in indentation by two-dimensional digital image correlation and finite elements method," *Materials*, vol. 10, no. 674, pp. 1–16, 2017.

[86] M. Wang and F. Hu, "The application of nltk library for python natural language processing in corpus research," *Theory and Practice in Language Studies*, vol. 11, no. 9, pp. 1041–1049, 2021.

[87] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, (Online), pp. 6442–6454, Association for Computational Linguistics, Nov. 2020.

[88] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu, "Improving named entity recognition by external context retrieving and cooperative learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 1800–1812, Association for Computational Linguistics, Aug. 2021.

[89] A. Smith, V. Bhat, Q. Ai, and C. Risko, "Challenges in information-mining the materials literature: A case study and perspective," *Chemistry of Materials*, vol. 34, no. 11, pp. 4821–4827, 2022.

[90] Y. Wang, L. Zhang, Y. Yao, and Y. Fu, "How to trust unlabeled data? instance credibility inference for few-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6240–6253, 2021.

[91] T. Ma, H. Jiang, Q. Wu, T. Zhao, and C.-Y. Lin, "Decomposed meta-learning for few-shot named entity recognition," in *Findings of the Association for Computational Linguistics: ACL 2022*, (Dublin, Ireland), pp. 1584–1596, Association for Computational Linguistics, May 2022.

[92] M. Elahinia, N. Shayesteh Moghaddam, A. Amerinatanzi, S. Saedi, G. P. Toker, H. Karaca, G. S. Bigelow, and O. Benafan, "Additive manufacturing of nitihf high temperature shape memory alloy," *Scripta Materialia*, vol. 145, no. 1, pp. 90–94, 2018.

[93] M. Abdel-Aziz, M. S. Zoromba, M. Bassyouni, M. Zwawi, A. Alshehri, and A. Al-Hossainy, "Synthesis and characterization of co-al mixed oxide nanoparticles via thermal decomposition route of layered double hydroxide," *Journal of Molecular Structure*, vol. 1206, no. 13, p. 127679, 2020.

[94] H. Xiao and S. Liu, "2d nanomaterials as lubricant additive: A review," *Materials & Design*, vol. 135, no. 9, pp. 319–332, 2017.

[95] H. Huang and D. Dean, "3-d printed porous cellulose acetate tissue scaffolds for additive manufacturing," *Additive Manufacturing*, vol. 31, no. 8, p. 100927, 2020.

[96] N. Gardner, H. Khan, and C.-C. Hung, "Definition modeling: literature review and dataset analysis," *Applied Computing and Intelligence*, vol. 2, no. 1, pp. 83–98, 2022.

[97] B. Strauss, B. Toma, A. Ritter, M.-C. de Marneffe, and W. Xu, "Results of the WNUT16 named entity recognition shared task," in *Proceedings of the 2nd Workshop on Noisy User-generated Text*, (Osaka, Japan), pp. 138–144, The COLING 2016 Organizing Committee, Dec. 2016.

[98] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, "Results of the WNUT2017 shared task on novel and emerging entity recognition," in *Proceedings of the 3rd Workshop on Noisy User-generated Text*, (Copenhagen, Denmark), pp. 140–147, Association for Computational Linguistics, Sept. 2017.

[99] F. Reiss, H. Xu, B. Cutler, K. Muthuraman, and Z. Eichenberger, "Identifying incorrect labels in the CoNLL-2003 corpus," in *Proceedings of the 24th Conference on Computational Natural Language Learning*, (Online), pp. 215–226, Association for Computational Linguistics, Nov. 2020.

[100] M. Guo, Z. Dai, D. Vrandečić, and R. Al-Rfou, "Wiki-40B: Multilingual language model dataset," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, (Marseille, France), pp. 2440–2452, European Language Resources Association, May 2020.

[101] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing: System Demonstrations, (Online), pp. 9–14, Association for Computational Linguistics, Oct. 2020.

[102] Y. Nie, Y. Tian, X. Wan, Y. Song, and B. Dai, "Named entity recognition for social media texts with semantic augmentation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, (Online), pp. 1383–1391, Association for Computational Linguistics, Nov. 2020.

[103] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 465–476, Association for Computational Linguistics, July 2020.

作者在攻读硕士学位期间公开发表的论文

攻读硕士期间论文发表情况：

1. “A Literature Mining Method Based on Curve Graph Information in Literature”（导师第一，学生第二，CIBDA 会议）

攻读硕士期间软著获取情况：

1. “软件名称：文献中图表信息的提取软件 V1.0”，创作人：韩越兴，夏锦桦，登记号：2021SR1218880，申请人：上海大学，开发完成日期：2021 年 5 月 1 日，登记日期：2021 年 8 月 17 日。
2. “软件名称：文献曲线图提取软件 V1.0”，创作人：韩越兴，夏锦桦，登记号：2022SR1365018，申请人：上海大学，开发完成日期：2022 年 5 月 4 日，登记日期：2022 年 9 月 21 日。

致谢

光阴荏苒，日月如梭，我的学生生涯也要就此告以段落了。回想求学岁月，对曾经帮助过我的老师同学不禁心生感激。

首先，我要感谢我的导师韩越兴老师。他在整个研究过程中都给予了我极大的支持和鼓励，不仅在研究方向上引领着我，在人际交往和事件处理上也提供了指导帮助。在我撰写论文的过程中，韩老师还提供了宝贵的意见和建议，韩老师就像一位引航者，引导我顺利走完科研与学术的旅程，这会是我人生美妙的一段旅程。无论是科研态度、学术能力方面，还是人际沟通与相处方面，这些都注定令我受益终生，祝愿韩老师家庭幸福美满，学术成就斐然，桃李满天下！

同时，感谢研究组内的陈侨川老师和张瑞老师的帮助。两位老师同样启发着我探索与发现科研的新领地，并以诙谐幽默的方式为实验组带来一份欢声笑语，这些都将是我人生的宝贵财富。祝愿两位老师学术成绩硕果累累，在科研的道路上宽越走越远，越来越宽！

我还要感谢三年来陪伴我的同学们。课题组的王璐学姐、刘宇虹同门、李睿祺同门、池洳婷学妹、万冠新学弟以及已经毕业的张宏坤学长，他们为我的研究生生活带来了许多色彩，也使得课题组更加有凝聚力。同班的李韶杰和茅威洋同学，在我三年生活中给予了不少支持和帮助，鼓励着我前行。

最后，我要感谢我的家人，他们无条件给予我经济支持与生活关怀，他们给了我一个稳定的港湾，使我全身心地投入学习中，在我完成这篇论文的过程中也给予了我巨大的支持和鼓励，他们的陪伴和支持是我完成这篇论文的重要动力。

在此，感谢所有对本论文有帮助的人，祝愿各位一帆风顺。