arXiv:2301.12293v1 [cs.AI] 28 Jan 2023

ACL-FIG：用于科学图分类的数据集

Zeba Karishma、Shaurya Rohatgi、Kavya Shrinivas Puranik宾夕法尼亚州立大学zebakarishma@gmail.com,{szr207, kzp5555}@psu.edu

Jian Wu旧道明大学 jwu@cs.odu.edu

C. LeeGiles宾夕法尼亚州立大学 clg20@psu.edu

摘要

目前大多数大规模学术搜索引擎都是为检索基于文本的信息而构建的。然而，还没有针对科学图表和表格的大规模检索服务。这类服务面临的一项挑战是理解科学图表语义，例如其类型和用途。一个关键障碍是缺少包含已标注科学图表和表格的数据集，而这些数据集可用于分类、问答和自动生成标题。在此，我们开发了一套处理流程，用于从科学文献中提取图表和表格；同时开发了一个基于深度学习的框架，利用视觉特征对科学图表进行分类。借助这套处理流程，我们构建了首个大规模自动标注语料库 ACL-FIG，其中包含从 ACL Anthology 的 ≈ 56K 篇研究论文中提取的 112,052 个科学图表。ACL-FIG-PILOT数据集包含 1,671 个人工标注的科学图表，分属 19 个类别。该数据集可通过 https://huggingface.co/datasets/citeseerx/ACL-fig 获取，采用 CC BY-NC 许可。

1 引言

科学论文中随处可见图表，它们用于展示实验和分析结果。我们将这些图表称为科学图表，以区别于自然图像；自然图像通常包含更丰富的色彩和渐变。科学图表以紧凑的方式呈现数值和类别数据，常常有助于研究人员获得见解并得出结论。理解科学图表有助于从网络上随时可获取的数亿篇科学论文中开发高效的检索系统 [1]。最先进的机器学习模型能够解析特定类别科学图表的图注和浅层语义。 [2] 然而，基于视觉特征对通用科学图表进行可靠分类，仍然是一项挑战。

在此，我们提出一种处理流程，用于构建经过分类且具备上下文信息的科学图表数据集。我们将该处理流程应用于 ACL Anthology 中的 55,760 篇论文（于 2021 年年中从 https://aclanthology.org/ 下载），构建了两个数据集：ACL-FIG 和 ACL-FIG-PILOT。ACL-FIG 包含 112,052 个科学图表及其图注、行内引用和元数据。ACL-FIG-PILOT（图 1 ）是 ACL-FIG 的一个未标注子集，包含 1,671 个科学图表，这些图表由人工标注为 19 个类别。ACL-FIG-PILOT 数据集被用作科学图分类基准。该处理流程为开源且可配置，使其他人能够利用预定义或新标签，从其他学术数据集中扩展这些数据集。

2 相关工作

科学图表提取 从科学论文中自动提取图表对于许多下游任务至关重要，因此人们已开发出许多框架。有人提出了一种名为 PDFMEF 的多实体提取框架，其中包含图提取模块 [3]。ImageCLEF 等共享任务 [4] 引起了人们对复合图表检测与分离的关注。 [5]提出了一个名为 PDFFIGURES 的框架，用于提取图表和

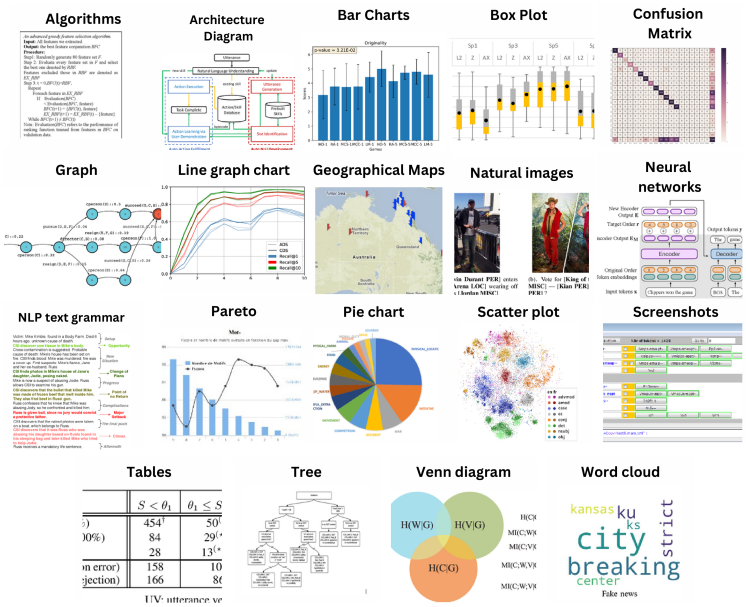

图 1：ACL-FIG-PILOT 中各类型图表示例。

表 1：科学图表分类数据集。

数据集	标签	#图表	图像来源
DeepChart	5	5,000	网络图像
FigureSeer[1]	5	30,600	网络图像
Prasad等人	5	653	网络图像
Revision	10	2,000	网络图像
FigureQA[3]	5	100,000	合成图表
DEEPFIGURES	2	1,718,000	科学论文
DocFigure[2]	28	33,000	科学论文
ACL-FIG-PILOT	**19**	**1,671**	科学论文
ACL-FIG（推断）[4]	-	**112,052**	科学论文

[1] 仅有 1000 张图像公开。[2] 未公开。[3] 科学风格合成数据。
[4] ACL-FIG-inferred 不包含人工标注标签。

研究论文中的图注。作者扩展了这项工作，构建了一个更稳健的框架，称为 PDFFIGURES2[6]。之后，有人提出了 DEEPFIGURES，以引入深度神经网络模型 [2]。

科学图分类 科学图分类 [7, 8]有助于机器理解图表。早期研究采用视觉词袋表示和支持向量机分类器 [7]。 [9] 使用霍夫变换识别文档图像中的条形图。 [10] 使用人工设计特征对科学文档中的图表进行分类。 [11] 结合了卷积神经网络（CNN）和深度置信网络，其表现优于基于特征的分类器。

图分类数据集目前已有多个用于图分类的数据集，例如 DocFigure [12],FigureSeer [10], Revision [7], 以及 [13] 提出的数据集（表 1）。FigureQA 是一个与我们的数据集类似的公开数据集，包含基于 10 万多张合成科学图像 [14]构建的一百多万个问答对，涵盖五种风格。我们的数据集与 FigureQA 不同，因为其中的图表是直接从研究论文中提取的。尤其是，DEEPFIGURES 的训练数据来自 arXiv 和 PubMed，仅标注为“图”或“表格”，且不包含细粒度标签。我们的数据集包含细粒度标签、行内上下文，并且采集自不同领域。

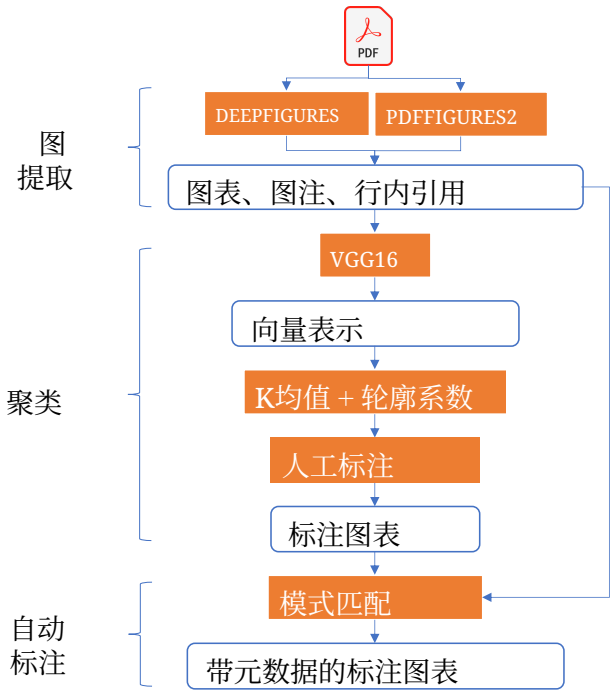

图 2：数据生成流水线概览。

3 数据挖掘方法论

ACL Anthology 是一个规模庞大、维护良好的 PDF语料库，具有干净的元数据，涵盖可免费获取全文的计算语言学论文。以往的图分类研究使用一组预定义类别（例如 [14], ），但这些类别可能只涵盖部分图类型。我们采用无监督方法来确定图类别，以克服这一限制。分配类别标签后，系统会自动为每个图表添加元数据、图注和行内引用。处理流程包含三个步骤：图提取、聚类和自动标注（图 2）。

3.1 图提取

为减轻单一图提取器可能造成的偏差，我们使用 PDFFIGURES2 [6] 和 DEEPFIGURES[2] 提取图表，这两种工具的工作方式各不相同。PDFFIGURES2 首先识别图注和正文，因为这两者相对容易准确识别。随后，通过查找与图注相邻且不与正文重叠的矩形边界框，即可定位包含图表的区域。DEEPFIGURES 使用远程监督学习方法，从大量 LaTeX 和 XML 格式的科学文档中推导图表标签。该模型基于 TensorBox，将 Overfeat检测架构应用于使用 ResNet-101 [2]生成的图像嵌入。我们使用公开提供的模型权重[1] 进行提取，这些权重是在 400 万个推导出的图表和 100 万个推导出的表格上训练的。模型输出图表和表格的边界框。除非另有说明，我们统称图表和表格为“图表”。我们使用多进程处理 PDF。每个进程都按以下步骤提取图表。该系统在配备 24 核处理器的 Linux 服务器上平均每分钟处理 200 篇论文。

1. 从任务队列中获取论文标识符。2. 从文件系统中调取论文。3. 从论文中提取图表和图注。4. 使用检测到的边界框，从渲染后的 PDF 中裁剪图表。5. 将裁剪后的图表保存为 PNG 格式，并将元数据保存为 JSON格式。

[1]https://github.com/allenai/deepfigures-open

3.2 聚类方法

接下来，我们使用无监督方法自动为提取出的图表标注。我们使用在 ImageNet [16]上预训练的 VGG16 [15], 提取视觉特征。为满足 VGG16 的输入要求，所有输入图表均缩放至尺寸 224×224。特征取自倒数第二个隐藏（全连接）层，该层包含 4096 个特征。我们采用主成分分析将维度降至 1000。

接下来，我们使用 k-均值聚类对由 1000 维向量表示的图表进行聚类。我们比较了两种启发式方法来确定最佳簇数，包括肘部法和轮廓分析 [17]。肘部法考察 解释方差，这是一种用于量化组间方差与总方差之差的度量，并将其作为簇数的函数进行分析。曲线的拐点（肘部）决定簇数。

轮廓分析通过测量簇之间的距离来确定簇数。它会考虑方差、偏度以及高低差异等多个因素，通常比肘部法更受青睐。轮廓图展示了一个簇中的每个点与相邻簇中各点的接近程度，从而让我们能够直观地评估簇数。

3.3 将图表关联到元数据

该模块将图表与元数据相关联，包括图注、行内引用、图类型、图边界坐标、图注边界坐标以及图中文字（出现在图表上的文本，仅适用于 PDFFIGURES2 的结果）。图类型由上述聚类步骤确定。行内引用通过 GROBID 获取（见下文）。其他元数据字段由图提取器输出。PDFFIGURES2 和 DEEPFIGURES 提取相同的元数据字段，但“图像文本”和“无区域图注”（未找到图区域的图注）除外，这两项仅适用于 PDFFIGURES2 的结果。

行内引用是包含对图表或表格的引用的文本片段。行内引用有助于理解文本与其所指对象之间的关系。处理论文后，GROBID 会输出一个 TEI文件（一种 XML文件），其中包含带标记的全文和引用。我们使用正则表达式定位行内引用，并

提取包含引用标记的句子。

4 结果

4.1 图提取

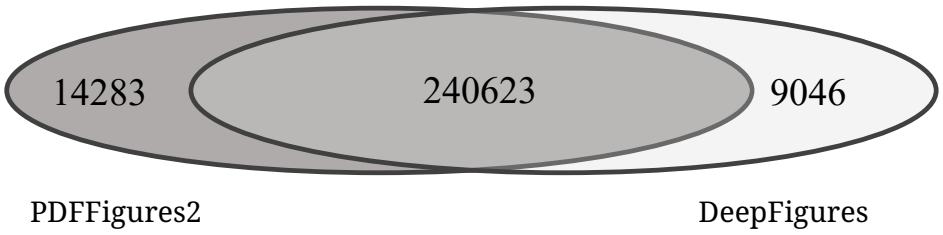

图 3：提取的图像数量。

图 3 展示了 PDFFIGURES2 和 DEEPFIGURES 提取的图表数量，表明这两个软件包提取的图表有大量重叠。不过，其中任一软件包都提取出了一些（≈ 5%）未被另一个软件包提取的图表。通过检查任一软件包提取的随机图表样本，我们发现 DEEPFIGURES 往往会漏掉两个图表上下相邻的情况。我们合并了两个软件包提取的所有图表，构建了 ACL-FIG 数据集，共包含 263,952 个图表。所有提取的图像均使用标准 OpenCV 库转换为 100 每英寸点数。压缩前，数据总大小为 ~ 25GB。行内引用使用 GROBID 提取。约 78% 的图表包含行内引用。

4.2 自动图表标注

提取结果包括 151,900 个表格和 112,052 个图表。我们仅使用 k-均值算法对图表进行聚类。我们将 k 从 2 逐步增加到 20，每次递增 1，以确定簇的数量。我们使用

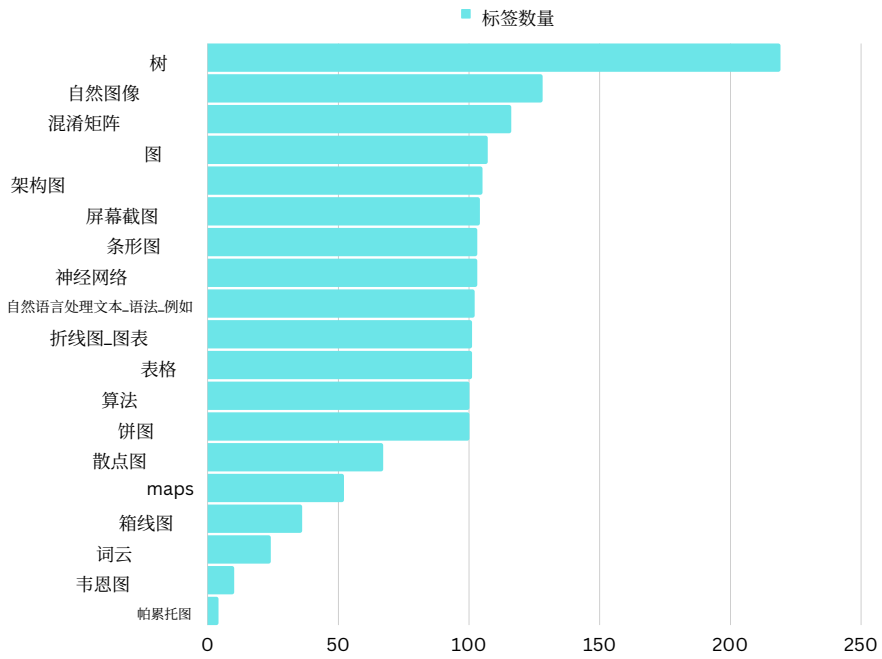

图4：ACL-FIG-PILOT数据集中各图类的分布。

肘部法和轮廓分析对结果进行了分析。肘部法曲线中未观察到明显的肘点。轮廓图展示了簇的数量与轮廓系数之间的关系，在 $k = 15$ 处出现明显的转折点，此时轮廓系数达到全局最大值。因此，我们将图表分为 15 个簇。

为验证聚类结果，我们从每个簇中随机抽取100幅图进行目视检查。检查过程中，我们发现了三种新的图类型：词云、帕累托图和韦恩图。随后，我们使用所有经过人工检查的图构建了ACL-FIG-PILOT数据集。两位标注人员对这些簇进行了人工标注和检查。我们使用科恩卡帕系数衡量一致率；ACL-FIG-PILOT数据集的系数为 $\kappa - 0.78$ （高度一致）。为求完整，我们添加了100个随机选取的表格。因此，ACL-FIG-PILOT数据集共包含1671幅图和表格，标注为19个类别。图4展示了所有类别的分布。

5 监督式科学图表分类

基于ACL-FIG-PILOT数据集，我们训练了监督式分类器。数据集按8:2的比例划分为训练集和测试集。我们研究了三种基线模型。模型1是一个三层卷积神经网络，使用分类交叉熵损失函数和Adam优化器进行训练。该模型包含三个典型的卷积层，每层后接一个最大池化层和一个丢弃层，以及三个全连接层。维度从 32×32 降至 16×16 ，再降至 8×8。最后一个全连接层将编码向量分为19类。该分类器的准确率为59%。

模型 2 基于 VGG16 架构训练，但将原始网络中的最后三个全连接层替换为一个长短期记忆层，随后接入用于分类的全连接层。该模型的准确率达到 $\sim$ 79%，比模型 1 高 20%。

模型 3 是视觉变换器（ViT） [18], ，它将图形划分为固定大小的图像块。随后对每个图像块进行线性嵌入，并添加位置嵌入。得到的向量序列被输入标准变换器编码器。ViT 模型表现最佳，准确率达到 83%。

6 结论

我们基于 ACL Anthology 收录的论文设计了一套处理流程，并用它构建了一个自动标注科学图表的语料库，其中包含相关元数据和上下文信息。这个语料库名为 ACL-FIG，共包含 ≈ 250k

个对象，其中约 42% 是图表，约 58% 是表格。我们还构建了 ACL-FIG-PILOT，这是 ACL-FIG 的一个子集，包含 1671 个科学图表，并带有 19 个经过人工核验的标签。我们的数据集包括从真实世界数据中提取的图表，类别数量也多于现有数据集，例如 DEEPFIGURES 和 FigureQA。

我们的处理流程的一个局限是使用了基于 ImageNet 预训练的 VGG16。未来，我们将通过在科学图表上重新训练更先进的模型（例如 CoCa） [19], ，来改进图表表示。另一个局限是，确定簇的数量需要通过目视检查。我们将考虑采用基于密度的方法，以实现聚类模块的完全自动化。

参考文献

[1] Madian Khabsa 和 C. Lee Giles。公共网络上的学术文档数量。*PLoS ONE*，9(5):e93949，2014 年 5 月。[2] Noah Siegel 、 Nicholas Lourie 、 Russell Power 和 Waleed Ammar。利用远程监督神经网络提取科学图表。第 *18* 届 *ACM/IEEE JCDL* 会议论文集，2018 年 5 月。[3] Jian Wu、Jason Killian、Huaiyu Yang、Kyle Williams、Sagnik Ray Choudhury、Suppawong Tuarob、Cornelia Caragea 和 C. Lee Giles。Pdfmef：面向学术文档和语义搜索的多实体知识抽取框架。第 *8* 届国际知识获取会议论文集，2015 年。[4] Alba García Seco de Herrera、Henning Müller 和 Stefano Bromuri。ImageCLEF 2015 医学分类任务概述。载于*CLEF*，2015 年。[5] Christopher Clark 和 S. Divvala。超越文本：从计算机科学论文中提取图表、表格和图注。载于*AAAI* 学术大数据研讨会，2015 年。[6] Christopher Clark 和 S. Divvala。Pdffigures 2.0：从研究论文中挖掘图表。*2016* 年 *IEEE/ACM* 联合数字图书馆会议（*JCDL*），第 143–152 页，2016 年。[7] M. Savva、Nicholas Kong、Arti Chhajta、Li Fei-Fei、Maneesh Agrawala 和 J. Heer。Revision：图表图像的自动分类、分析与重新设计。第 *24* 届 *ACM* 用户界面软件与技术年度研讨会论文集，2011 年。[8] Sagnik Ray Choudhury 和 C. Lee Giles。数字图书馆中图表信息抽取的架构。第 *24* 届万维网国际会议论文集，2015 年。[9] Y. Zhou 和 C. Tan。用于文档图像中条形图检测与识别的霍夫技术。*2000* 年国际图像处理会议论文集（目录号 *00CH37101*），第 2 卷：605–608，2000 年。[10] Noah Siegel、Zachary Horvitz、Roie Levin、S. Divvala 和 Ali Farhadi。FigureSeer：解析研究论文中的结果图表。载于*ECCV*，2016 年。[11] Binbin Tang、Xiao Liu、Jie Lei、Mingli Song、Dapeng Tao、Shuifa Sun 和 Fangmin Dong。DeepChart：结合深度卷积网络和深度置信网络的图表分类。信号处理，124:156–161，2016 年。[12] K. V. Jobin、Ajoy Mondal 和 C. V. Jawahar。DocFigure：一个用于科学文档图分类的数据集。载于*2019* 年国际文档分析与识别会议研讨会（*ICDARW*），第 74–79 页，2019 年。[13] V. Karthikeyani 和 S. Nagarajan。用于识别便携式文档格式（PDF）文件中图表类型的机器学习分类算法。国际计算机应用期刊，39:1–5，2012 年。[14] Samira Ebrahimi Kahou、Vincent Michalski、Adam Atkinson、Akos Kadar、Adam Trischler 和 Yoshua Bengio。FigureQA：一个用于视觉推理的带注释图表数据集，2018 年。[15] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*arXiv* 预印本 *arXiv:1409.1556*，2014 年。[16] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：一个大规模分层图像数据库。载于*2009* 年 *IEEE CVPR* 会议，第 248–255 页，2009 年。[17] Peter J. Rousseeuw。轮廓系数：用于解释和验证聚类分析的图形辅助工具。计算与应用数学杂志，20:53–65，1987 年。[18] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。图像胜过 16×16 个词：用于大规模图像识别的 Transformer。*arXiv* 预印本 *arXiv:2010.11929*，2020 年。[19] Jiahui Yu、Zirui Wang、Vijay Vasudevan、Legg Yeung、Mojtaba Seyedhosseini 和 Yonghui Wu。CoCa：对比式图像描述器是图文基础模型。*CoRR*，abs/2205.01917，2022 年。