

面向国际物流领域的视觉富文档数据集构建及信息抽取方法研究

李唐, 张俊伟, 陈夏华, 王昊*

上海大学, 计算机工程与科学学院



计算机工程与科学学院
School of Computer Engineering and Science Shanghai University



上海大学
Shanghai University

引言

视觉富文本文档主要指包含图像、文本、版式等丰富多模态信息的一种文档。目前, 面向视觉富文本文档 (VRDs) 关键信息抽取任务(KIE)的数据集并不多见, 尤其是可以作为细粒度分类任务以及少样本抽取任务评测的基准数据集**几乎没有**。本工作的主要贡献包括:

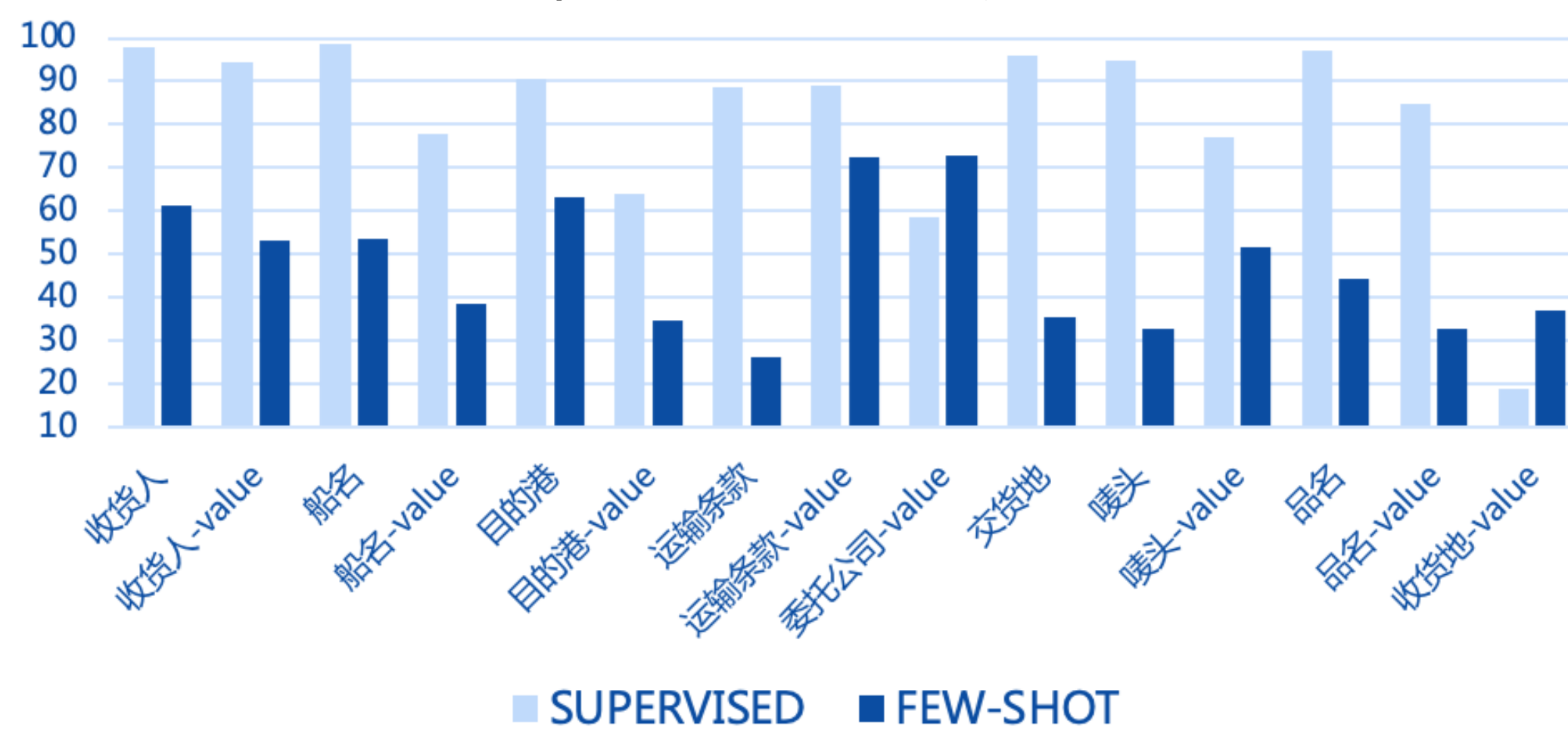
- 为面向视觉富文档的关键信息抽取任务构建了一个新的**有监督细粒度实体分类及关系抽取数据集**, 该数据集不仅包含细粒度标签信息, 也包含键值关系信息;
- 在此基础上, 构建了**支持少样本学习的FEW-VIED视觉富文档数据集**, 可以用来评测少样本学习在富文档信息任务中的表现;
- 除此之外, 我们还**提出了一种基于双注意机制的视觉富文档关键信息抽取方法 (DualVIE)**, 在多个视觉富文档数据集上取得了SOTA的效果。

方法

- 有监督数据集构建**: 我们收集了大量国际物流海运单数据, 从中抽取了约四千份高质量单页文档; 预先设计了包含包含了20个键以及其对应值的标签体系 (包含如“发货人”和“发货人-value”等共40种类别的实体提及); 招募了20名志愿者, 3名质检员, 并利用百度PPOCRLabel半自动标注工具, 以及百度飞浆OCR工具识别文本及其对应位置; 然后, 对每一个文本框进行人工标注。该数据集包含了近**4000**份文档, 涵盖**40种**细粒度的键值分类及键值关系标注, 且具有较高的标注质量。

类别	样本数
收货人-value	18421
发货人-value	14717
通知人-value	8788
委托公司-value	1048
收货地-value	226
...	...

- 少样本数据集构建**: 目前, 很少有专门针对少样本场景视觉富文档信息抽取评测的少样本数据集。我们在现有有监督数据集基础上, 重新划分了数据子集并重新组织, 通过掩码抽样策略, 将其迁移到少样本的环境中进行实证研究, 目的是用很少的例子来识别看到的实体类型, 该数据集具有多个层次的标记: 粗粒度表格单元格 (cell级)、中粒度文本框(bounding box 级)和细粒度单词(token级)。与纯文本少样本信息抽取任务不同的是, 我们关注视觉富文档信息抽取任务中**布局及位置信息是否对于迁移学习是至关重要**, 并使用ProtoNet进行了测试, 这需要进一步研究。我们将模型与传统的有监督方法进行了比较, 发现在少样本的设定下模型还存在着较大的改进空间。如果忽略掉位置信息, 发现模型的表现大幅下降。



- Dual注意力文档信息抽取框架**: 2020年, 微软亚研院提出一种面向富文本文档信息抽取任务的多模态融合文档预训练语言模型LayoutLM, 它将BERT预训练语言模型与文档布局、视觉特征等其他多模态信息融合, 具有通用文档信息抽取的能力。最近, 我们在LayoutLM的基础上提出了一种基于异构图文档信息抽取方法, 通过对token与bounding box两个粒度的语义关联应用不同策略的图注意力机制(DualVIE)。评测结果显示, 我们提出的模型在自己构建的海运单数据集以及FUNSD等多个公开数据集上的关键信息抽取任务中, 都获得了SOTA的效果。

讨论

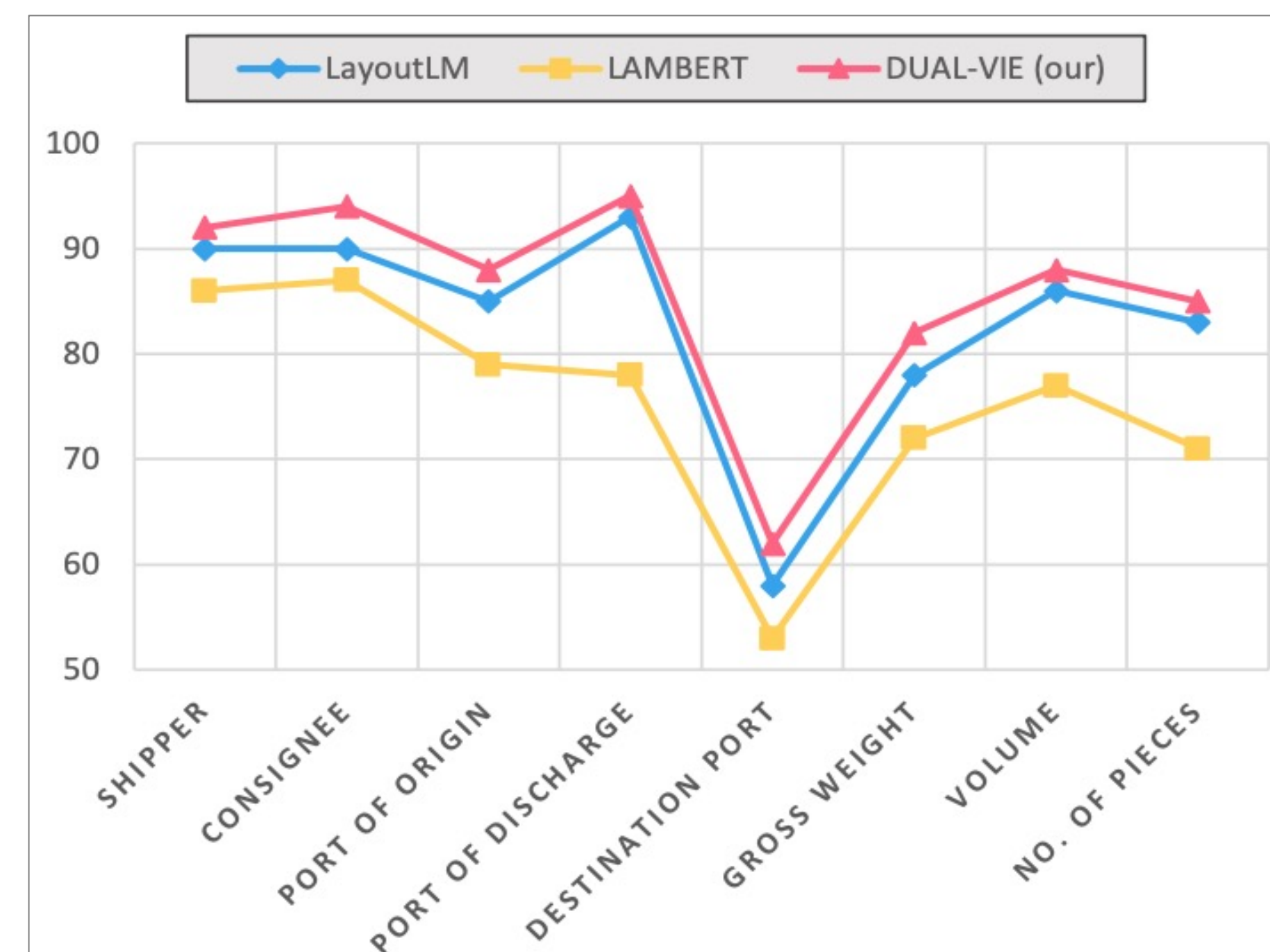
现有方法获得的视觉富文档信息抽取模型具有领域局限性, 如果迁移到一个新的领域往往效果很差, 这仍需未来进一步探索。

CCKS2022

图(1): 国际物流领域的海运单文档及标注样例

类别	文本+位置+图像	文本+图像
收货人	61.43	26.46
收货人-value	53.34	45.08
船名	53.63	22.65
船名-value	38.49	13.90
目的港	63.15	27.67
目的港-value	34.78	10.42
运输条款	26.30	16.25
运输条款-value	72.38	29.07
唛头	32.78	20.79
唛头-value	51.52	18.62
收货地-value	37.19	12.10

图(2): 少样本全模态与不含位置时的比较



图(3): 海运单数据集上各方法效果细粒度比较

*DualVIE: Dua-level Graph Attention Network for Visual Information Extraction, *The 19th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, 2022



wang-hao@shu.edu.cn

2022全国知识图谱与语义计算大会
China Conference on Knowledge Graph and Semantic Computing
秦皇岛 8.24-8.27