从金融公告中自动构建股权结构知识图谱

乔美萱*，王俊，向俊夫，侯启予，李瑞轩*

南京吾道知信信息技术有限公司，*华中科技大学

CCKS2022

引言

- 各类金融公告作为典型的富格式(Rich-format)商业文档，除了包含文本内容之外，还包含各种类型和格式的表格和插图。
- 对富格式文档中插图的研究还处于早期阶段，目前还没有专门对金融公告文档中各类插图进行分析理解的工作。
- 股权结构图是金融公告中非常重要的一类框图(Diagram)，其中的节点表示机构或者个人实体，连线表示一对实体之间持股关系和比例
- 针对目前对于金融公告中的股权结构图识别效果不佳的问题，本文提出了一种更可靠的框图识别系统，用于股权结构图识别，能够很好识别各种场景下的复杂连线。
- 此外，我们利用股权结构图自身的结构特点，开发实现了一套自动生成股权结构图和对应标注数据的工具。

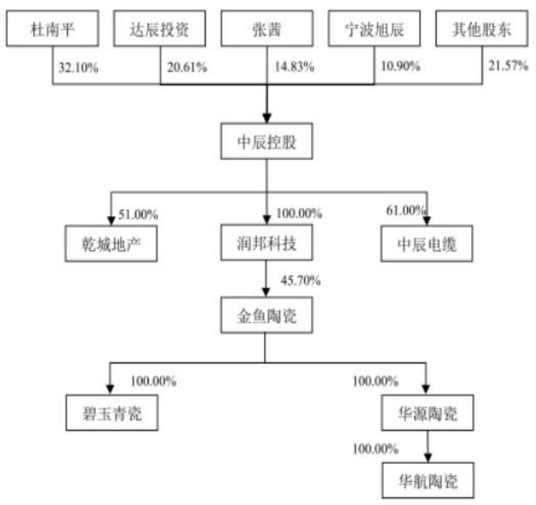


图1 股权结构图示例

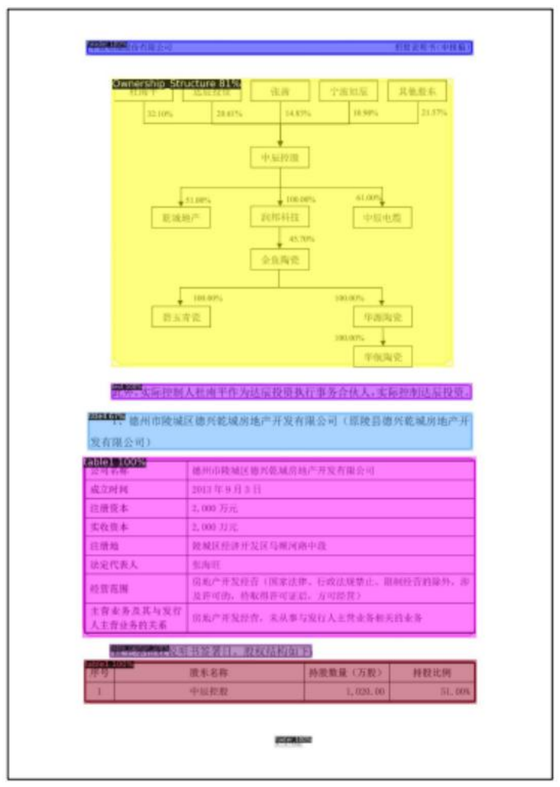


图2 股权结构图检测和定位示例

方法

- 根据对部分真实数据分布的观察，自动生成了6072个股权结构图，同时也对每张图中的节点，连线和总线的边界框，以及其中连线的起止关键点自动进行标注。
- 从招股说明书等公开披露的真实金融公告自动检测和抽取出股权结构图，如上图，最后由人工校对筛选出1849张股权结构图作为评测基准数据集。
- 基于Oriented-RCNN，使用改进的关键点检测方法对股权图的节点和连线进行检测，最后生成股权关系三元组；
- 评测标准：目标检测的mAP和股权关系的准确率F1。

结果

- 虽然自动生成的训练数据集试图尽量捕捉真实数据分布，但是真实数据毕竟还存在一些和常规场景不同的较复杂特殊的情况，所以在真实训练数据上进行精调能够进一步改进模型的性能。
- 利用在自动生成的数据集上训练得到的初始模型，对真实评测基准数据集全体进行自动标注和人工校对后再进行训练：

表1 精调模型和初始模型在真实测试数据集上的对比

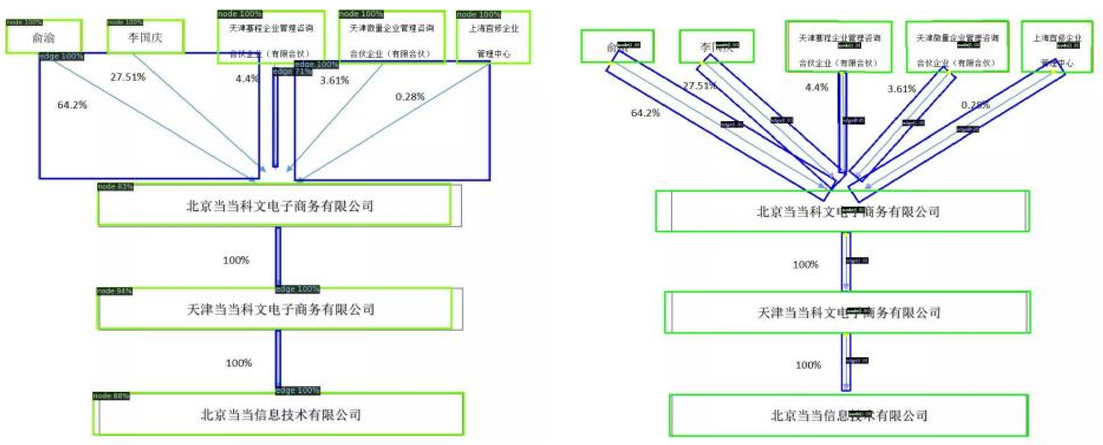
目标	指标 (%)	Arrow R-CNN		本文方法	
		初始	精调	初始	精调
节点	Precision	92.4	92.0	88.2	97.8
	Recall	94.8	95.3	97.7	98.6
	AP	89.5	90.1	90.6	90.8
连线	Precision	22.0	34.9	77.4	85.9
	Recall	27.9	38.5	77.6	90.3
	AP	20.5	32.9	67.4	85.1
总线	Precision	N/A	N/A	68.4	74.6
	Recall	N/A	N/A	75.0	93.2
	AP	N/A	N/A	66.9	88.5
连线起止点	Precision	34.8	42.4	97.9	97.8
	Recall	36.5	40.8	87.0	93.2
	AP	31.8	32.7	80.7	89.5
mAP		57.1	64.5	73.1	87.4

表2 从股权结构图中抽取结构化数据的结果对比

指标 (%)	Arrow R-CNN		本文方法	
	初始	精调	初始	精调
Precision	32.0	34.6	83.5	87.3
Recall	16.7	19.8	76.2	83.4
F1	21.9	25.2	79.7	85.3

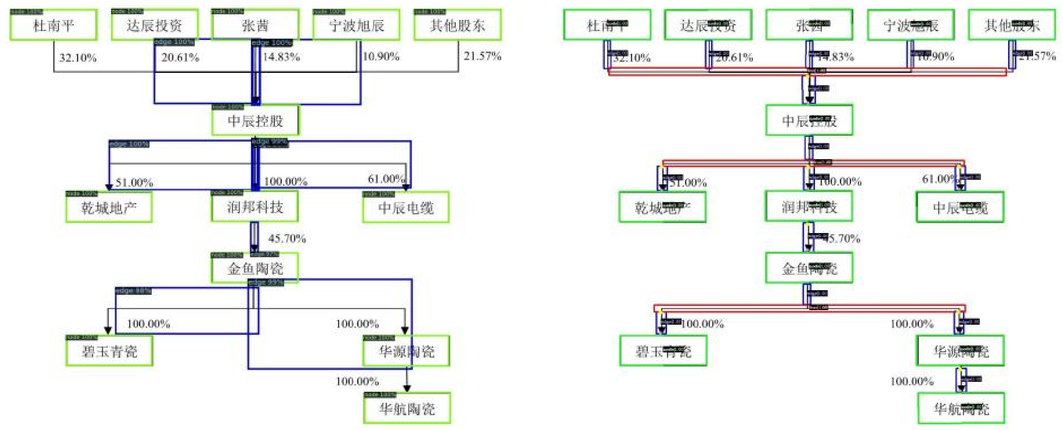
讨论

- 实验结果显示我们提出的方法对于Arrow R-CNN在性能上有显著的进步，尤其是对连线的识别有巨大的改进：
- 未来我们计划将本文的方法进行扩展，并将其应用到金融文档中更多的其他类型的框图的识别上。



(a) 只能输出水平边界框的Arrow R-CNN检测结果　(b) 可以输出旋转边界框的本文方法检测的结果

图3 对包含的倾斜连线场景的检测结果对比



(a) Arrow R-CNN将折线进行整体检测的结果　(b) 本文方法对折线进行分段检测的结果

图4 对于包含总线结构的一对多或者多对一的连线场景检测结果对比