MSRA-万字综述 直击多模态文档理解

原创 付奶茶 夕小瑶的卖萌屋 2021-12-28 12:05

文 | 付奶茶

随着最近几年多模态大火的，越来越多的任务都被推陈出新为多模态版本。譬如，传统对话任务，推出了考虑视觉信息的多模态数据集；事件抽取，也推出视频形式的多模态版本；就连grammar induction（语法归纳），也有了多模态版的（详见 NAACL'2021 best paper）。

然而，**多模态大火虽是最近的事情，但它并不是近两年才有的什么新技术。如果是想要对这一领域有比较深的研究，甚至想要做出工作、有所创新，那仅仅了解多模态最近两年几个大火的多模态模型显然是不足够的。**

事实上，有些任务已经天生就是多模态很多年了。早在多模态成为焦点之前，就已经默默被研究二十来年了。比如，**智能文档**（Document AI）技术。所谓智能文档技术，也就是**自动理解、分析业务文档技术，文档内容可包含文字、图片、视频等多种形式**。由于理解多模态形式的多模态形式文的需求其实广泛长期存在，所以智能文档技术很多年来都是几个大厂的研究重点之一。近年来，深度学习技术的普及也更好地推动了例如文档布局分析、可视化信息提取、文档可视化问答、文档图像分类等智能文档算法的发展。近期，微软亚研院发表了一篇综述，简要回顾了一些有代表性的DocumentAI的模型、任务和基准数据集。小编认为这篇概述的总结体系非常扎实，是值得细细阅读的多模态相关综述，故与各位分享。

论文标题:
Document AI: Benchmarks, Models and Applications

论文链接:
https://arxiv.org/abs/2111.08609

Document AI 发展历程

作者概述智能文档的发展大致经历了以下三个阶段:

第一阶段: 启发式阶段

20世纪90年代初,研究人员主要使用基于规则的启发式(Heuristic rule-based document layout analysis)来理解和分析文档，通过手动观察文档的布局信息，从而总结出一些启发式规则。启发式规则方法主要使用固定的布局信息来处理文档.方法较为固定，定制的规则可扩展性较差,通用性较差。

> 基于启发式规则的文档的布局分析大致分为三种方式:
> (1)自顶向下:文档图像逐步划分到不同的区域,递归执行切割直到该区域被划分为预定义的标准，通常是块或列。例如projection profile,采用X-Y cut算法对文档进行剪切,通常用于文本区域和行距固定的结构化文本,对特定格式的文档进行更快、更有效的分析.但其对边界噪声敏感，对倾斜文本的处理效果不佳。
> (2)自底向上:使用像素或组件作为基本单元,将其分组并合并成一个更大的同质区域,自底向上方法虽然需要更多的计算资源，但更通用，可以覆盖更多具有不同布局类型的文档。
> (3)混合策略:将自上而下和自下而上相结合，例如Okamoto & Takahashi使用分隔符和空格来切割块，并将内部组件进一步合并到每个块中的文本行中,进而解析文档的布局。

第二阶段:机器学习阶段

- 直到从2000年来 随着机器学习技术的发展，以机器学习模型逐渐成为文档处理的主流方法。研究者设计功能模板以了解不同功能的权重，进而理解和分析文档的内容和布局。

> 基于机器学习的文档分析过程通常分为两个阶段:
> 1)对文档图像进行分割，获得多个候选区域;
> 2)对文档区域进行分类和区分，如文本块和图像。

尽管带注释的数据被用于监督学习，并且以前的方法可以带来一定程度的性能改进，但是由于缺乏定制规则和训练样本数量，通用性仍然不令人满意。此外，不同类型文档的迁移和适应成本相对较高，这使得以前的方法不适合广泛的商业应用。

第三阶段:深度学习阶段

随着深度学习的发展和大量未标注电子文档的积累，可以通过工具HTML/XML提取、PDF解析器、OCR等提取不同类型的文档中的内容，其文本内容、布局信息和基本图像信息等基本组织良好,然后对大规模深度神经网络进行预训练和微调，以完成各种下游文档AI任务.包括文档布局分析、视觉信息提取、文档视觉问答和文档图像分类等。现有的基于深度学习的智能文档模型主要分为两大类:

- 针对特定任务的深度学习模型
- 支持各种下游任务的通用预训练模型

DocumentAI的主要任务

Document AI在我们现实的应用场景主要有以下四类任务:

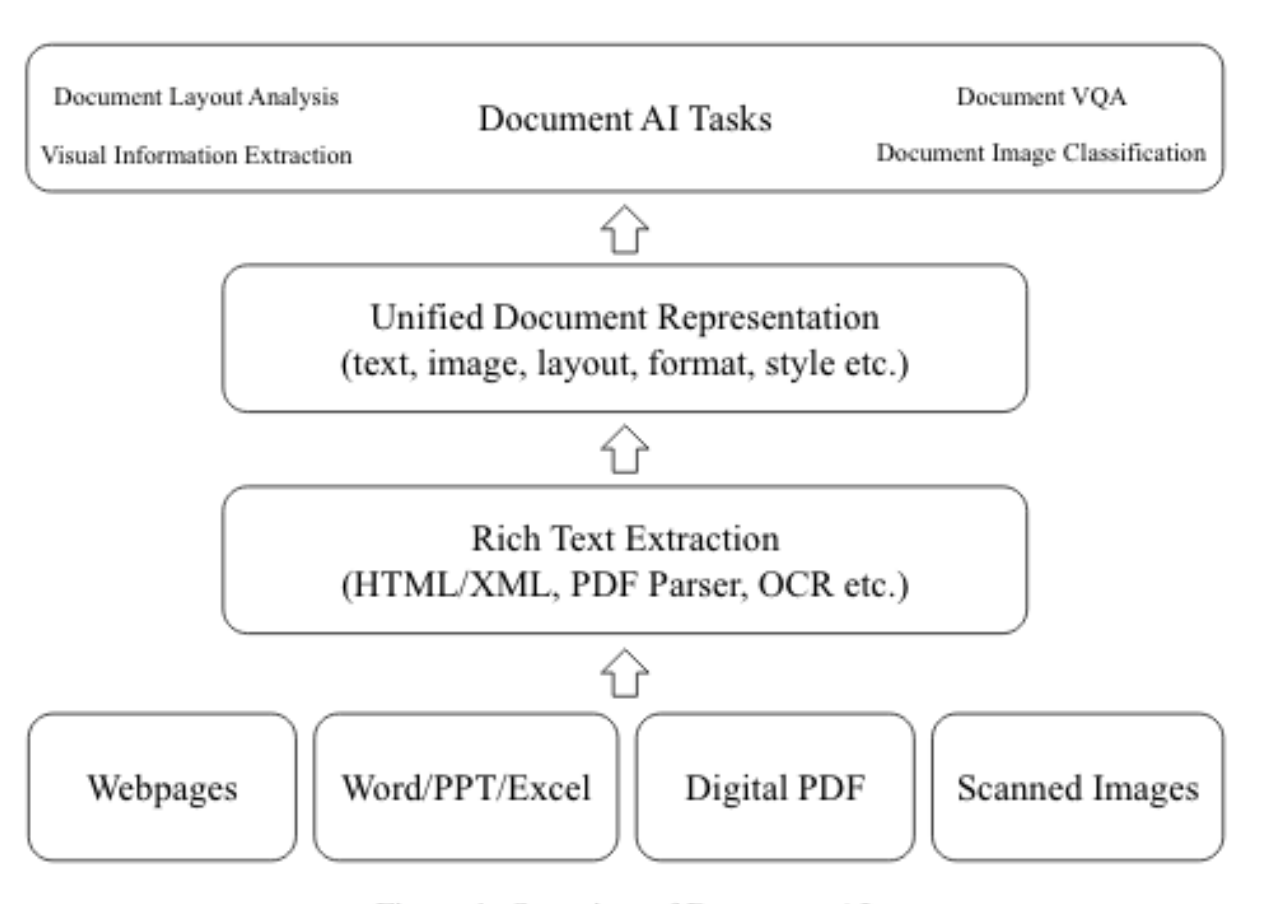


Figure 1: Overview of Document AI

Document Layout Analysis

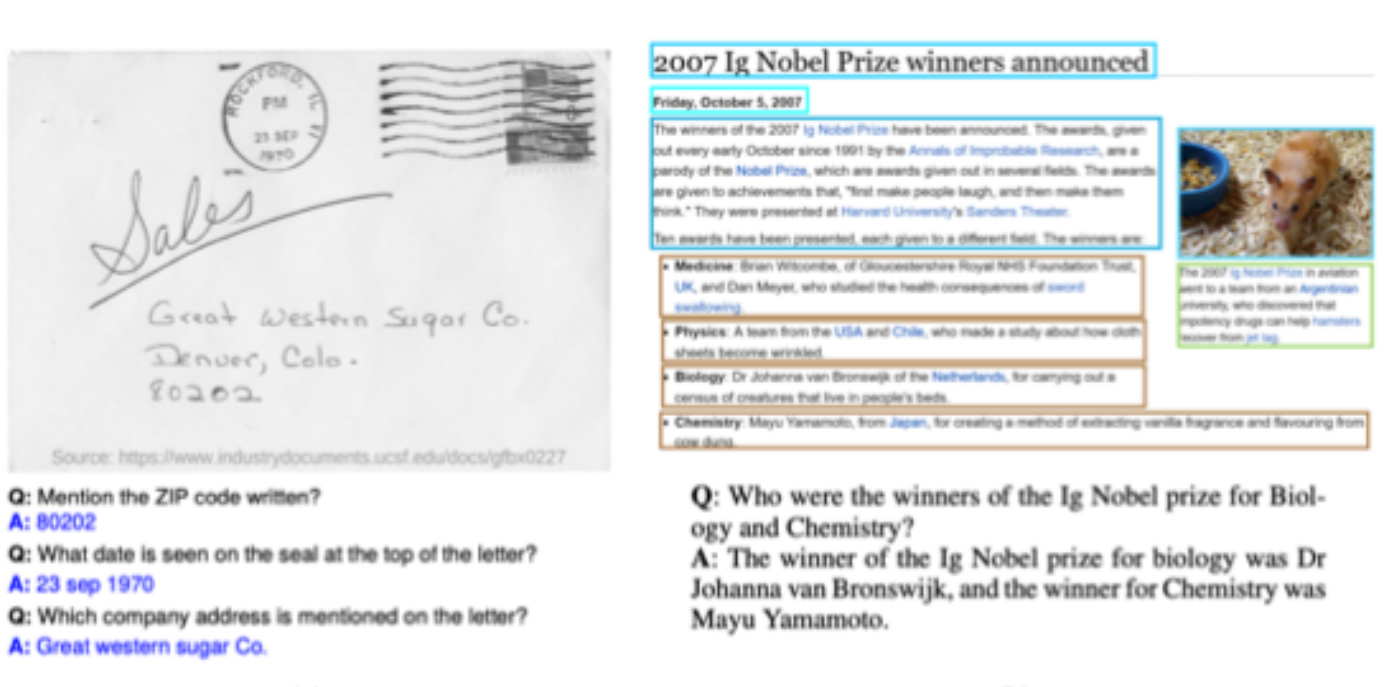
该任务主要是对文档布局中的图像、文本、表格、图等位置关系进行自动分析、识别、理解的过程.主要分为两个主任务:Visual analysis 与Semantic analysis.Visual analysis为视觉元素的分析,主要目的是检测文档的结构,确定相似区域的边界,而Semantic analysis为语义分析检测区域识别特定的文档元素,例如标题、段落、表格等。

Visual Information Extraction

该任务从文档中的大量非结构化内容中提取实体及关系.对于视觉丰富的文档建模为计算机视觉问题，通过语义分割或文本框检测来进行信息提取,将文档图像视为像素网格，将文本特征添加到视觉特征图中。根据文本信息的粒度，该任务从字符级发展到单词级，再发展到上下文级。

Document Visual Question Answering

该任务为通过判断识别文本的内部逻辑来回答关于文档的自然语言问题。文档VQA中的文本信息在任务中起着至关重要的作用，现有的有代表性的方法都是以文档图像的OCR获取的文本作为输入。获得文档文本后，将VQA任务建模为不同的问题.主流方法将其建模为机器阅读理解(MRC)问题,根据问题从给定文档中提取文本片段作为相应的答案。



Q: Mention the ZIP code written?
A: 80202
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970
Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

(a)

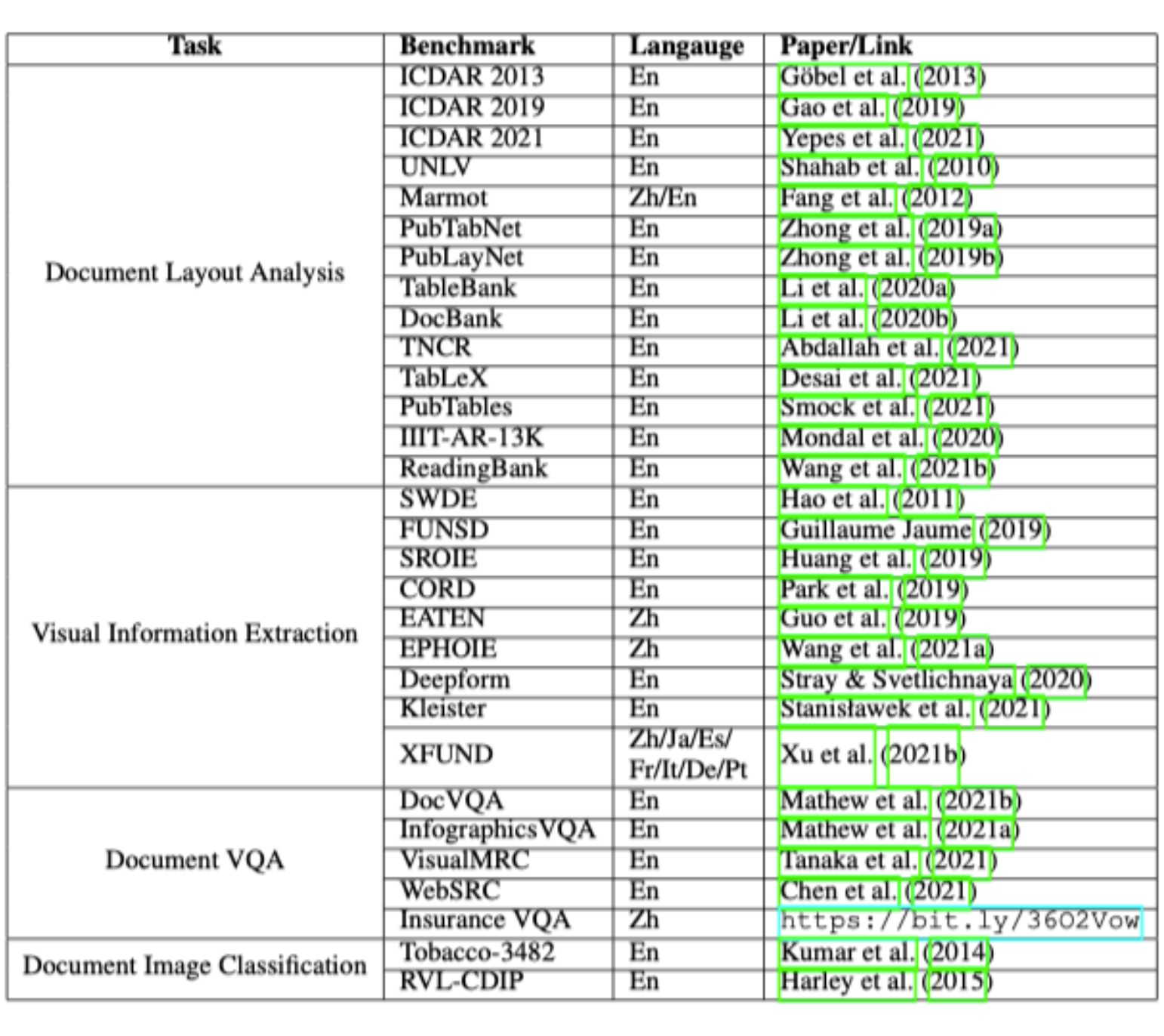
Q: Who were the winners of the Ig Nobel prize for Biology and Chemistry?
A: The winner of the Ig Nobel prize for biology was Dr Johanna van Bronswijk, and the winner for Chemistry was Mayu Yamamoto.

(b)

Document Image Classification

该任务是对文档图像进行分析和识别，并将其划分为不同类别的过程，如科学论文、简历、发票、收据等。最早文档图像分类问题的方法与自然图像分类方法基本一样,例如基于CNN的文档图像分类方法使用经过ImageNet训练的Alexnet作为初始化对文档图像进行模型适配。

下图为作者整理的四大Task的Benchmark:



Task	Benchmark	Langauge	Paper/Link
Document Layout Analysis	ICDAR 2013	En	Göbel et al. (2013)
	ICDAR 2019	En	Gao et al. (2019)
	ICDAR 2021	En	Yepes et al. (2021)
	UNLV	En	Shahab et al. (2010)
	Marmot	Zh/En	Fang et al. (2012)
	PubTabNet	En	Zhong et al. (2019a)
	PubLayNet	En	Zhong et al. (2019b)
	TableBank	En	Li et al. (2020a)
	DocBank	En	Li et al. (2020b)
	TNCR	En	Abdallah et al. (2021)
	TabLeX	En	Desai et al. (2021)
	PubTables	En	Smock et al. (2021)
	IIIT-AR-13K	En	Mondal et al. (2020)
	ReadingBank	En	Wang et al. (2021b)
Visual Information Extraction	SWDE	En	Hao et al. (2011)
	FUNSD	En	Guillaume Jaume (2019)
	SROIE	En	Huang et al. (2019)
	CORD	En	Park et al. (2019)
	EATEN	Zh	Guo et al. (2019)
	EPHOIE	Zh	Wang et al. (2021a)
	Deepform	En	Stray & Svetlichnaya (2020)
	Kleister	En	Stanisławek et al. (2021)
	XFUND	Zh/Ja/Es/ Fr/It/De/Pt	Xu et al. (2021b)
Document VQA	DocVQA	En	Mathew et al. (2021b)
	InfographicsVQA	En	Mathew et al. (2021a)
	VisualMRC	En	Tanaka et al. (2021)
	WebSRC	En	Chen et al. (2021)
	Insurance VQA	Zh	https://bit.ly/36O2Vow
Document Image Classification	Tobacco-3482	En	Kumar et al. (2014)
	RVL-CDIP	En	Harley et al. (2015)

Table 1: Benchmark datasets for document layout analysis, visual information extraction, document visual question answering and document image classification.

Document AI 主流模型

Documents layout analysis with convolutional neural networks

文档布局分析可以看作是对文档图像进行目标检测的任务。将文档中的标题、段落、表格、图表等基本单元是需要检测和识别的对象。Yang等人将文档布局分析作为像素级的分割任务，利用卷积神经网络进行像素分类，取得了较好的效果.

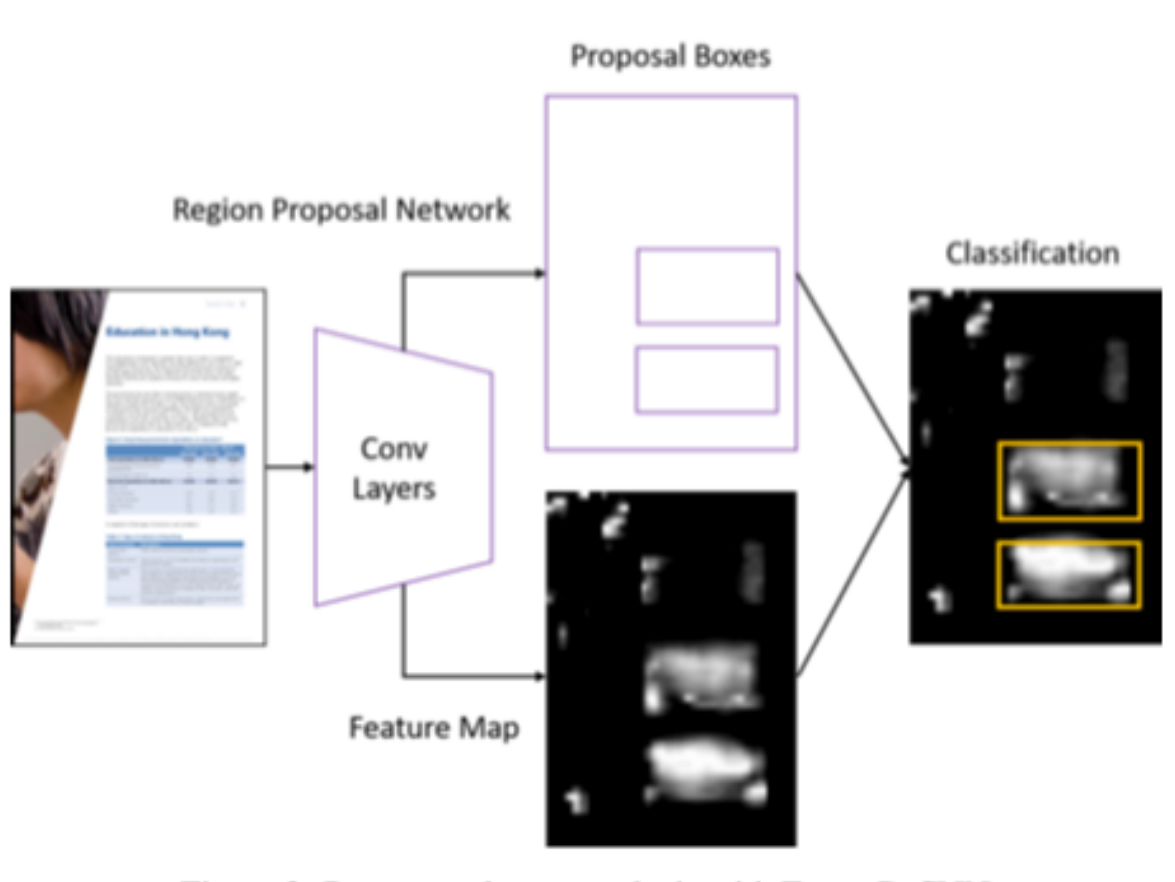


Figure 2: Document layout analysis with Faster R-CNN

Visual information extraction with graph networks

对于视觉信息丰富的文档的结构不仅仅由文本内容的结构决定,与布局、排版、格式、表/图结构等视觉元素同样相关.例如收据、证书、保险文件等.Liu等人提出的利用图卷积神经网络建模视觉元素丰富的文档,首先通过OCR系统获得一组Text Blocks,每一个Text Block包含其在图像中与文本内容的坐标信息,将其构成一个完全连通的有向图,即每个Text Blocks构成一个节点,通过Bi-LSTM获取节点的初始特征,边的初始特征是相邻文本块与当前文本块之间的相对距离以及这两个文本块的长宽比。对“节点-边缘-节点”三元特征集进行卷积,实验表明，视觉信息发挥了主要作用，增加了文本识别相似语义的能力,对视觉信息也起到一定的辅助作用。

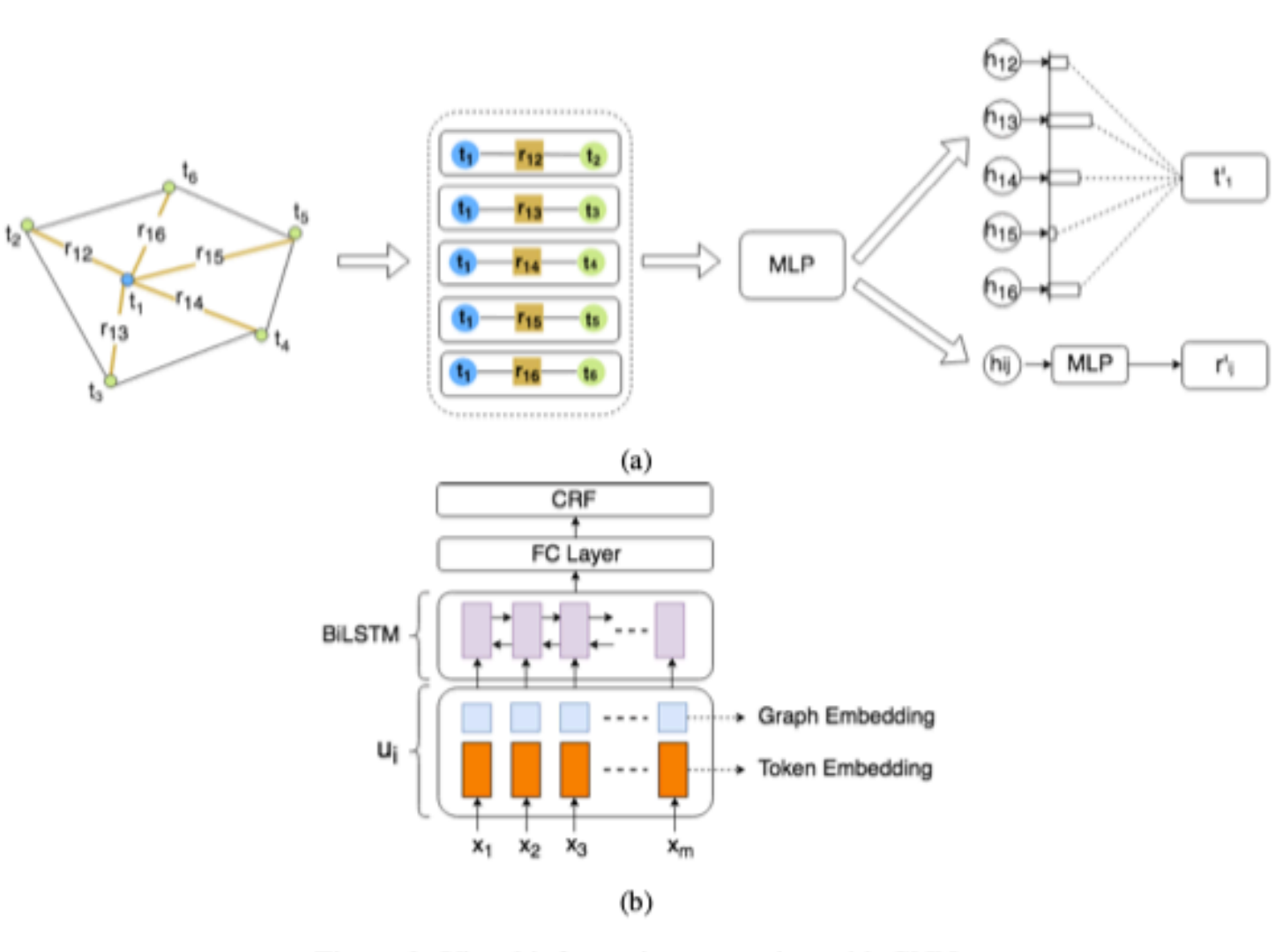


Figure 3: Visual information extraction with GNN

General-purpose multimodal pre-training with the transformer

文本块的空间关系通常包含丰富的语义信息。例如,在表格中本块通常以网格布局排列，标题通常出现在第一列或第一行。不同文档类型之间的布局不变性是通用预训练的一个关键属性。通过预训练与文本自然对齐的位置信息可以为下游任务提供更丰富的语义信息。对于视觉信息丰富的文档,其视觉信息如字体类型、大小、样式等明显的视觉差异，其可以通过视觉编码器提取出来，结合到预训练阶段，从而有效地改善下游任务。为了利用布局和视觉信息，2020年Xu提出通用文档预训练模型LayoutLM，在已有预训练模型的基础上，增加了2-Dposition embedding 和image embedding.首先根据OCR得到的文本边界框得到文本在文档中的坐标。将对应的坐标转换为虚坐标后，模型计算出x、y、w、h四个embedding sublayers对应的坐标表示，最终的二维位置嵌入是四个子层的embedding之和。在imageembedding 中，模型将每个文本对应的边框作为Faster R-CNN,提取相应的局部特征。特别是，由于[CLS]符号用于表示整个文档的语义，因此模型还使用整个文档的image作为image embedding以保持多模态对齐,Layout模型在三个下游任务，表单理解，票据理解，文档图像分类，都取得了显著的准确率提升。

LayoutLM的两个自监督预训练任务: Task1:Masked Visual-Language：随机mask除了2D postionembedding, 以及其他文本的text embdedding, 让模型预测mask的 token. Task2:Multi-Label Document Classification: 在给定一组扫描文档的情况下，利用文档标签对训练前的过程进行监督，使模型能够对来自不同领域的知识进行聚类，生成更好的文档级表示.该模型的相关实验表明，利用布局和视觉信息的预训练可以有效地转移到下游任务中。

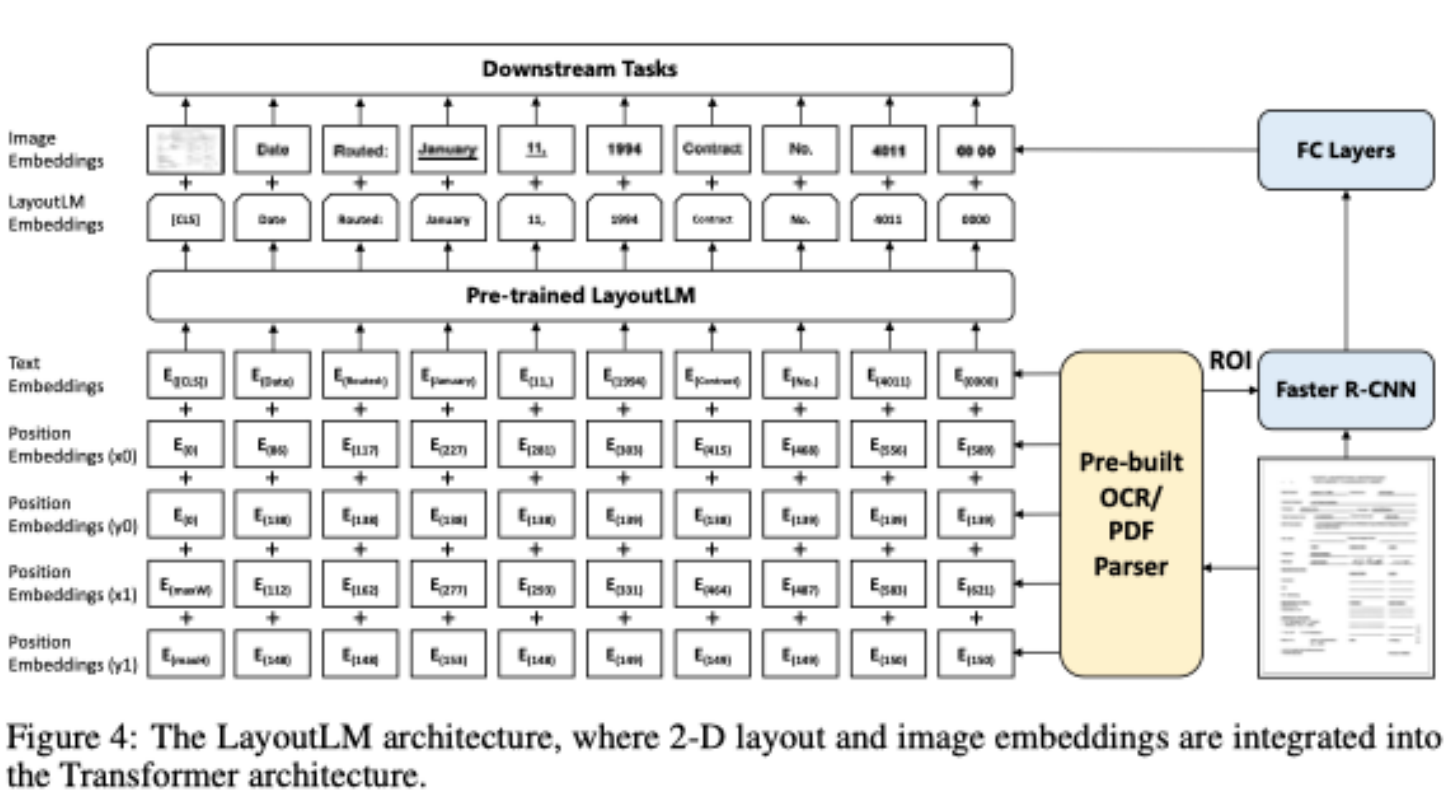


Figure 4: The LayoutLM architecture, where 2-D layout and image embeddings are integrated into the Transformer architecture.

小结

除了这篇文章介绍的之外LayoutLM等经典模型，最近DocumentAI的研究工作中几个后起之秀也非常值得关注。例如LayouLM后出现的LayoutLMv2以及LayoutXML，将跨模态对齐的思路贯彻在模型训练的过程中。不仅仅利用文本和布局信息，将图像信息也融合到文档多模态的框架内。除此之外，跨模态文档理解模型ERINE-Layout，提出阅读顺序预测和细粒度图文匹配两个与训练任务，除了跨模态予以对齐能力外，增加了布局理解能力。我们可以看到，在预训练时代下，DocumentAI正在逐渐向“多模态文档理解”方向前进，从模态之间的对齐到预测，DocumentAI将会怎样找寻可以建模的更多元素，挖掘视觉与文本、布局之间的精细关系，变得更加值得期待了。

卖萌屋相关阅读：《别再搞纯文本了！多模文档理解更被时代需要！》

萌屋作者：付奶茶
新媒体交叉学科在读Phd,卖萌屋十级粉丝修炼上任小编,目前深耕多模态,希望可以和大家一起认真科研,快乐生活!

作品推荐

1. 在斯坦福，做 Manning 的 phd 要有多强?
2. 史上最大多模态图文数据集发布!