**VSR：结合视觉、语义和关系的文档布局分析统一框架**

Peng Zhang 海康威视研究所

摘要：文档布局分析对于理解文档结构至关重要。在此任务中，文档的视觉和语义以及布局组件之间的关系有助于理解过程。虽然已经提出了许多利用上述信息的工作，但结果并不令人满意。

基于NLP的方法将布局分析建模为序列标记任务（语义角色标注），并且在布局建模方面表现出不足。(逻辑结构被定义为文档中逻辑组件的层次结构，例如节标题、段落和列表 )

基于CV的方法将布局分析建模为检测或分割任务，但存在模态融合效率低和布局组件之间缺乏关系建模的局限性。

为了解决上述局限性，我们提出了一个用于文档布局分析的统一框架VSR，它结合了视觉、语义和关系。VSR支持基于NLP和基于CV的方法。具体地说，我们首先通过文档图像引入视觉，通过文本嵌入映射引入语义。然后，使用双流网络提取特定于模态的视觉和语义特征，并对其进行自适应融合，以充分利用互补信息。最后，在给定的候选元件中，引入基于图神经网络的关联模块，对元件之间的关系进行建模，并输出最终结果。在三个流行的基准测试中，VSR的表现大大优于以前的模型。代码将很快发布。

**1导言**

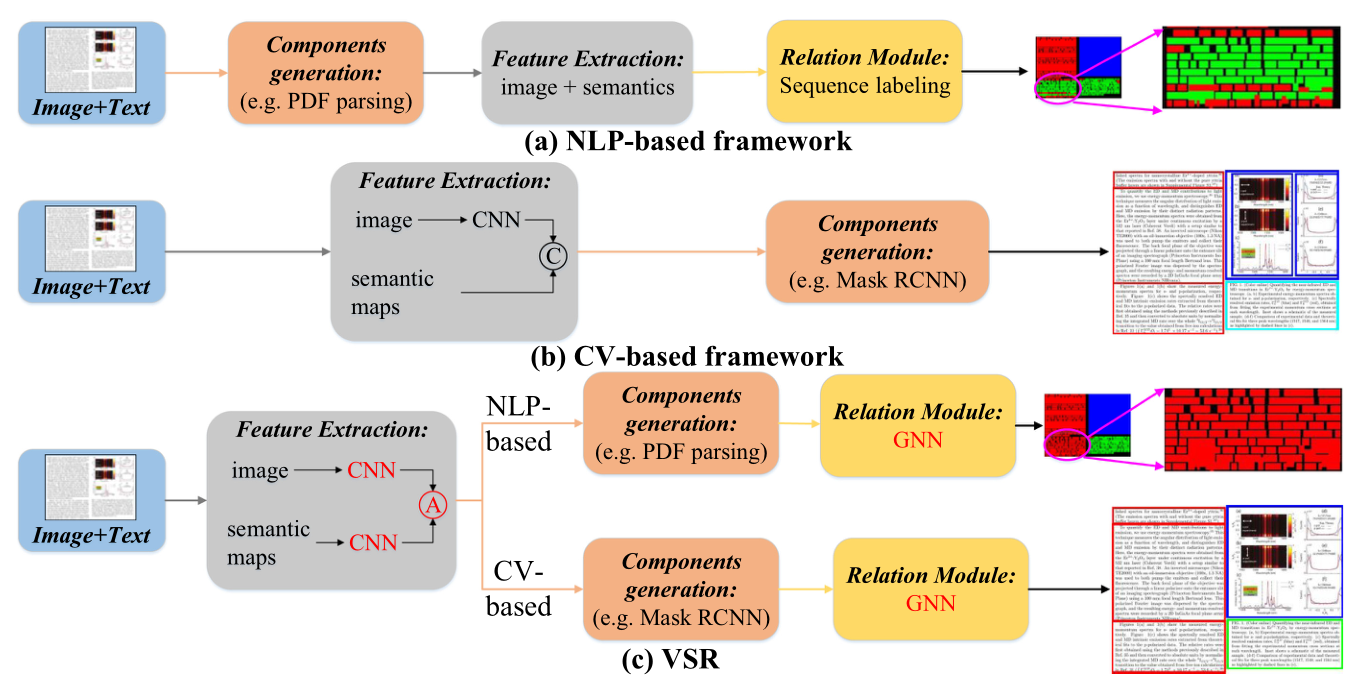
文档布局分析是自动理解文档的关键步骤，支持许多重要应用，如文档检索[4]、数字化[7]和编辑。它的目标是识别非结构化文档中感兴趣的区域，并识别每个区域的角色。由于文档布局的多样性和复杂性，此任务具有挑战性。

在计算机视觉（CV）和自然语言处理（NLP）社区中，已经提出了许多关于此任务的深度学习模型。他们中的大多数只考虑视觉特征[5，10，12，19，21，34，36，41]，或仅语义特征[6，17，27 ]。然而，来自这两种模式的信息有助于更好地认识文件的布局。

一些区域（如图形、表格）可以通过视觉特征轻松识别，而语义特征对于分离视觉上相似的区域（如摘要和段落）非常重要。因此，最近的一些努力试图将这两种模式结合起来[1,3,20,39]。这里我们将它们归纳为两类。

基于NLP的方法（图1（a））将布局分析建模为序列标记任务，并应用自底向上策略。他们首先将文本序列化为1D标记序列（在本文的其余部分中，我们假设文本可用。有一些工具可以从PDF文档（例如PDFMiner[28]）和文档图像（例如OCR引擎[30]）中提取文本）。然后，使用每个token（单词）的语义和视觉特征（如坐标和图像嵌入），它们通过序列标签模型顺序确定token（单词）标签。

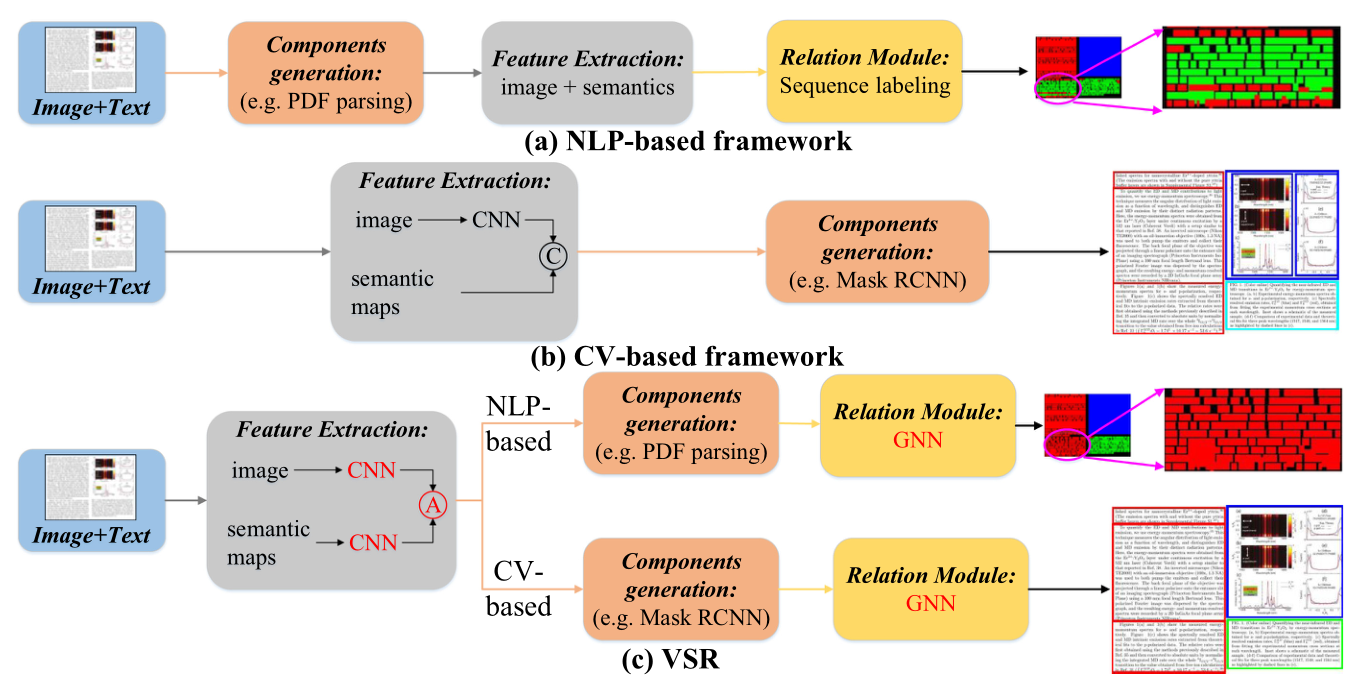
然而，基于NLP的方法在布局建模方面表现出不足的能力。例如，在图1（a）中，段落中的所有文本应具有一致的语义标签（段落），但其中一些文本被识别为图标题，图标题是相邻文本的标签。

组件 特征提取（视觉加语义） 关系模块：序列标注

基于CV的方法（图1（b））将布局分析建模为对象检测或分割任务，并应用自顶向下的策略。他们首先通过卷积神经网络提取视觉特征，然后通过文本嵌入映射（在句子层次[39]或字符层次[3]）引入语义特征，这些映射直接连接起来作为文档的表示。然后，使用检测或分割模型（例如，maskrcnn[13]）生成布局组件候选（坐标和语义标签）。

卷积神经网络提取视觉特征

文本嵌入映射（在句子层次[39]或字符层次[3]）引入语义特征（粒度？）



concatenated

组件生成：Mask RCNN

与基于NLP的方法相比，基于CV的方法能够更好地捕获空间信息，但仍然有3个局限性：

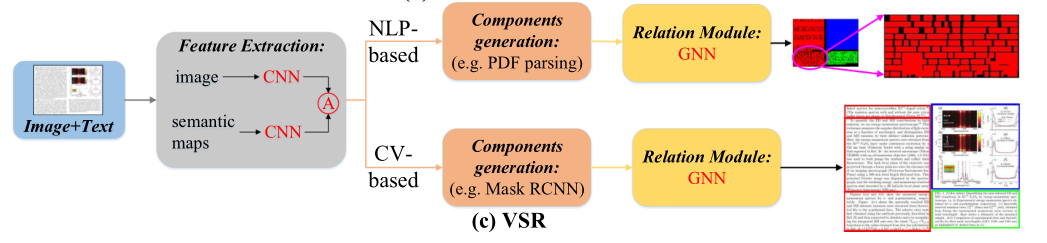
1. 语义有限。语义信息以不同的粒度嵌入到文本中，包括字符（或单词）和句子，这有助于识别不同的文档元素。例如，字符级特征更适合于识别需要较少上下文的组件（如作者），而句子级特征更适合于上下文组件（如表格标题）。在一个粒度上利用语义无法获得最佳性能。(能否利用不同粒度的语义？)
2. 简单的启发式模态融合策略。不同模式的特征对部件识别的贡献不同。视觉特征更有助于识别视觉丰富的组件（如图形和表格），而语义特征则更擅长区分基于文本的组件（摘要和段落）。简单的启发式模态拼接（concate）融合不能充分利用两种模态之间的互补信息。（采用更好的模态拼接？）

（3） 组件之间缺乏关系建模。文件中存在着密切的关系。例如，“图”和“图标题”通常同时出现，“段落”具有对齐的边界框坐标。这种关系可用于提高布局分析性能。

在本文中，我们结合视觉、语义和关系建模，提出了一个用于文档布局分析的统一框架VSR，如图1（c）所示。该框架可应用于基于NLP和基于CV的方法。

关系模块预测其语义标签。

文本标记充当组件候选



关系模块生成它们的精确坐标和语义标签。

通过检测或分割模型（例如 MASK RCNN）提出候选组件

首先，文档以图像（视觉）和文本嵌入映射（字符级和句子级语义）的形式输入VSR。然后，通过双流网络提取特定于模态的视觉和语义特征，并将其有效地组合到多尺度自适应聚合模块中。最后，结合基于GNN（图形神经网络）的关系模块，对候选组件之间的关系进行建模，并生成最终结果。

基于NLP的方法，文本标记充当组件候选，关系模块预测其语义标签。

基于CV的方法，通过检测或分割模型（例如更快的RCNN/掩码RCNN）提出候选组件，关系模块生成它们的精确坐标和语义标签。

我们的工作做出了四个关键贡献：

* 我们提出了一个统一的框架 VSR 用于文档布局分析，结合文档中的视觉、语义和关系。
* 为了有效地利用视觉和语义，我们提出了一个双流网络来提取特定于模态的视觉和语义特征，并通过自适应聚合模块自适应地融合它们。 此外，我们还探索了不同粒度的文档语义。
* 基于 GNN 的关系模块被纳入文档组件之间的关系建模，它支持基于 NLP 和基于 CV 的方法中的关系建模。
* 我们对 VSR 进行了广泛的评估，在三个公共基准测试中，VSR 与之前的模型相比显示出显着的改进。

**2相关工作**

**文档布局分析**。在本文中，我们试图从使用的模态的角度，即单峰布局分析和多峰布局分析来回顾布局分析工作。

单峰布局分析仅利用视觉特征 [19,21]（文档图像）或仅利用语义特征（文档文本）来理解文档结构。 使用视觉特征，已经提出了一些作品 [5,36] 来应用 CNN 来分割各种对象，例如文本块 [10]、文本行 [18,34]、单词 [41]、图形或表格 [12,29] ]。 同时，也有一些方法 [6,17,27] 尝试使用语义特征来解决布局分析问题。 然而，上述所有方法都严格限于视觉或语义特征，因此无法利用其他模态的补充信息。

多模态布局分析尝试结合来自视觉和语义模态的信息。 相关的方法可以进一步分为两类，基于NLP的方法和基于CV的方法。 基于 NLP 的方法将低级元素（例如tokens）和模型布局分析作为序列标记任务工作。 MMPAN [1] 用于识别表单结构。 DocBank [20] 计划作为多模态布局分析的大规模数据集，并且已经发布了多个 NLP 基线。 然而，上述方法在布局建模方面表现出不足。 基于 CV 的方法通过文本嵌入图引入文档语义，并将模型布局分析作为对象检测或分割任务。 MFCN [39] 引入了句子粒度语义并在决策级（网络末端）插入了文本嵌入图，而 dhSegmentT2[3] 引入了字符粒度语义并在输入级插入了文本嵌入图。 上述方法虽然取得了巨大的成功，但也存在以下局限性：使用的语义有限，模态融合策略简单，组件之间缺乏关系建模。

为了弥补上述局限性，我们提出了一个统一的框架VSR来利用文档中的视觉、语义和关系。

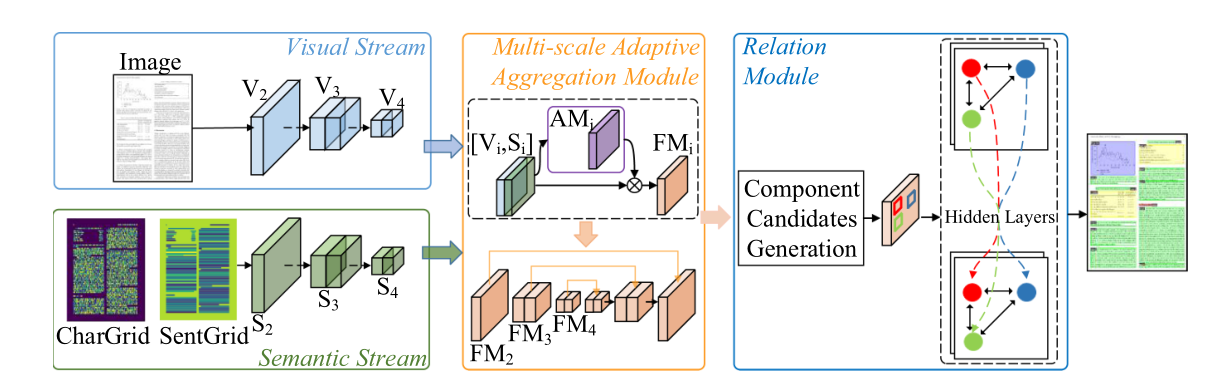
**双流网络**。双流网络被广泛用于有效地结合不同模态或表示 [2] 中的特征。 在动作识别中，双流网络用于捕获互补的时空信息[9]。 在 RGB-D 显着性检测中，完整的表示融合了 RGB 流和深度流的深层特征 [11]。 此外，双流网络用于在声音事件分类和图像识别中融合相同输入样本的不同特征 [23]。 受他们成功的启发，我们应用双流网络来捕获文档中的互补视觉和语义信息。

**关系建模**。关系建模是一个广泛的主题，并且已经研究了几十年。 在自然语言处理中，序列文本之间的依赖关系是通过 RNN[15] 或 Transformer[32] 架构来捕获的。 在计算机视觉中，提出了非局部网络[35]和关系网络[16]被用来模拟像素和对象之间的长期依赖关系。此外，在文档图像处理中，探索了文本和布局之间的关系 [38] 或文档实体之间的关系 [24,40,42]。

对于多模态布局分析，基于 NLP 的方法将其建模为序列标记任务并使用 RNN 来捕获组件关系，而基于 CV 的方法将其建模为对象检测任务，但缺乏布局组件之间的关系建模。 在本文中，我们提出了一个基于 GNN 的关系模块，支持基于 NLP 或基于 CV 的方法中的关系建模。

**3方法**

**3.1架构概述**

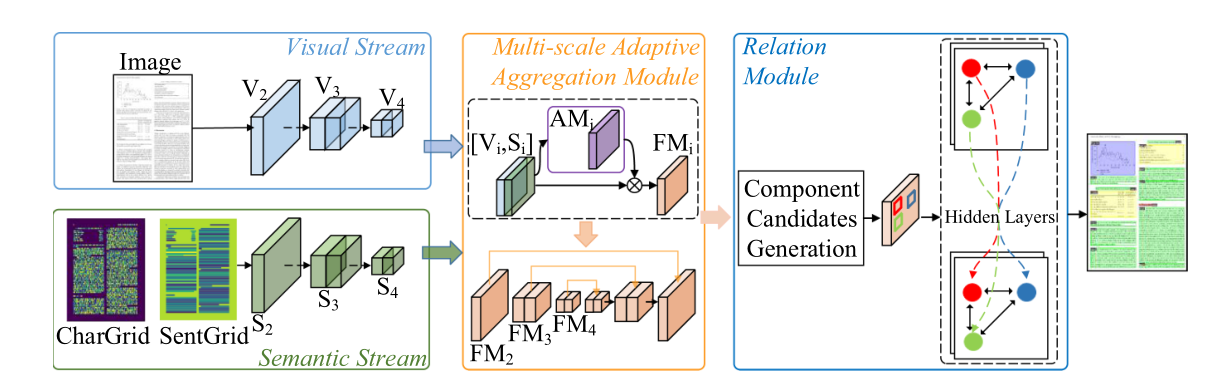
我们提出的框架由三部分组成：双流卷积网络、多尺度自适应聚合模块和关系模块（如图 2 所示）。 首先，双流卷积网络提取特定于模态的视觉和语义特征，其中视觉流和语义流分别将图像和文本嵌入图作为输入（第 3.2 节）。 接下来，我们不是简单地连接视觉和语义特征，而是通过多尺度自适应聚合模块（第 3.3 节）聚合它们。 然后，生成一组候选组件。 最后，一个关系模块被纳入这些候选者之间的模型关系并生成最终结果（第 3.4 节）。

请注意，多模态布局分析可以建模为序列标记（基于 NLP 的方法）或对象检测任务（基于 CV 的方法）。 我们的框架支持这两种建模类型。 唯一的区别是候选组件是什么以及如何生成它们。 组件候选是基于 NLP 的方法中的低级元素（例如，文本标记），可以通过解析 PDF 生成，而基于CV 中生成的高级元素（区域）的方法候选是由检测或分割模型（例如，Mask RCNN）生成的。 在本文的其余部分，我们将说明 VSR 如何应用于基于 CV 的方法，并在 DocBank 基准（第 4.3 节）的实验中展示它可以很容易地适应基于 NLP 的方法。

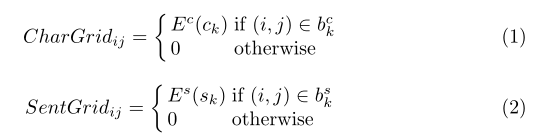
**3.2 双流卷积网络**

众所周知，CNN 擅长学习深度特征。 然而，以前的多模态布局分析工作 [3,39] 仅将其应用于提取视觉特征。 文本嵌入图直接用作语义特征。 这种单流网络设计无法充分利用文档语义。 受双流网络在各种多模式应用中取得巨大成功的推动 [9,23]，我们将其应用于提取深层视觉和语义特征。

**视觉流卷积网络**。 该流直接将文档图像作为输入，并使用 ResNet [14] 等 CNN 主干提取多尺度深度特征。 具体来说，对于输入图像x∈，提取多尺度特征图（用{V2,V3,V4,V5}表示），其中每个Vi∈。H和W是输入图像x的高度和宽度，是特征图Vi的通道维度，V0=x。

**语义流ConvNet。**类似于[3,39]，我们通过文本嵌入映射S0∈引入文档语义，这是语义流卷积网络的输入。S0与文档图像x（V0）具有相同的空间大小，表示初始通道维度。 这种类型的表示不仅对文本内容进行编码，而且还保留了文档的 2D 布局。 以前，仅使用一种粒度的语义（字符级 [3] 或句子级 3 [39]）。 然而，不同粒度的语义有助于识别不同的组件。 因此，S0 由字符和句子级别的语义组成。 接下来，我们将展示如何构建文本嵌入映射 S0。

文档页面的字符和句子表示为  和  ，其中 n 和 m 是字符和句子的总数。和 是第 k 个字符及其关联的框，其中 (x0,y0) 和 (x1,y1) 是左上角和右下角像素坐标。 同样，sk 和 bsk 是第 k 个句子及其框位置。 接下来，字符嵌入映射 CharGrid∈ 和句子嵌入映射 SentGrid∈ 可以构造如下。

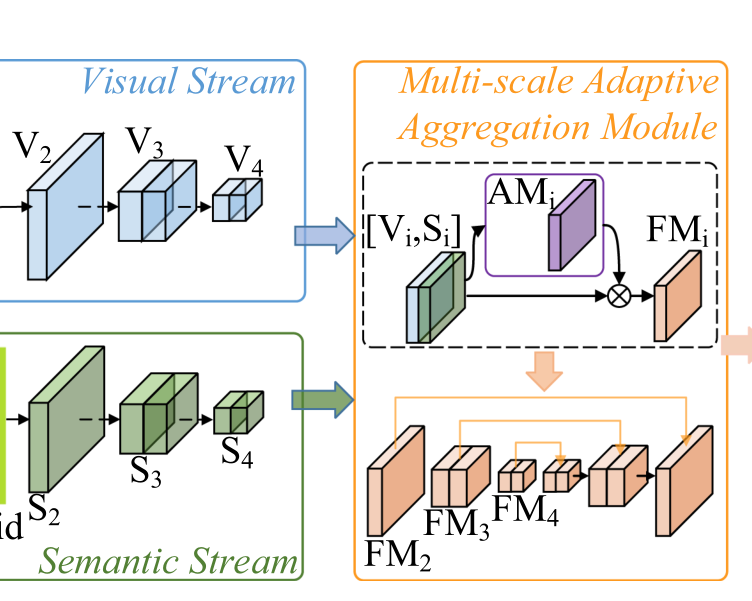


每个  中的所有像素共享相同的字符（句子）嵌入向量。  和 是 和  的映射函数。 在我们的实现中， 是一个典型的词嵌入层，我们采用预训练的语言模型 BERT [8] 作为。 最后，文本嵌入映射 S0 可以通过将 LayerNorm 归一化应用于 CharGrid 和 SentGrid 的总和来构建，如等式 3 所示。

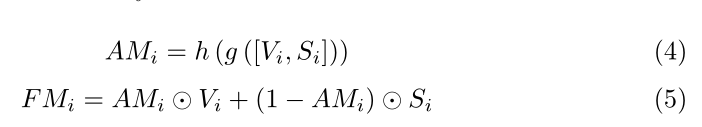


与视觉流类似，语义流卷积网络然后将文本嵌入图 S0 作为输入并提取多尺度特征 {S2,S3,S4,S5}，这些特征具有与 {V2,V3,V4,V5} 相同的空间大小和通道维度 .

**3.3多尺度自适应聚合**



来自不同模态的特征对于识别不同的对象很重要。 模态融合策略应该自适应地聚合视觉和语义特征。 因此，我们设计了一个多尺度自适应聚合模块，该模块学习注意力图以自适应地结合视觉特征 {V2,V3,V4,V5} 和语义特征 {S2,S3,S4,S5}。 在尺度 i，该模块首先连接 Vi 和 Si，然后将其输入卷积层以学习注意力图 AMi。 最后，获得聚合（连接 Vi 和 Si以及注意力图）的多模态特征FMi。 此模块中的所有操作均由以下公式



其中 [·] 表示连接操作，g(·) 是卷积层，内核大小为 是非线性激活函数。⚪表示逐元素乘法。 通过该模块，生成了一组融合的多模态特征FM={FM2,FM3,FM4,FM5}，作为文档的多模态多尺度特征。 然后，将 FPN [22]（特征金字塔网络）应用于 FM 并提供增强的表示。

AMi是连接 Vi 和 Si然后经过卷积层，再经过卷积层

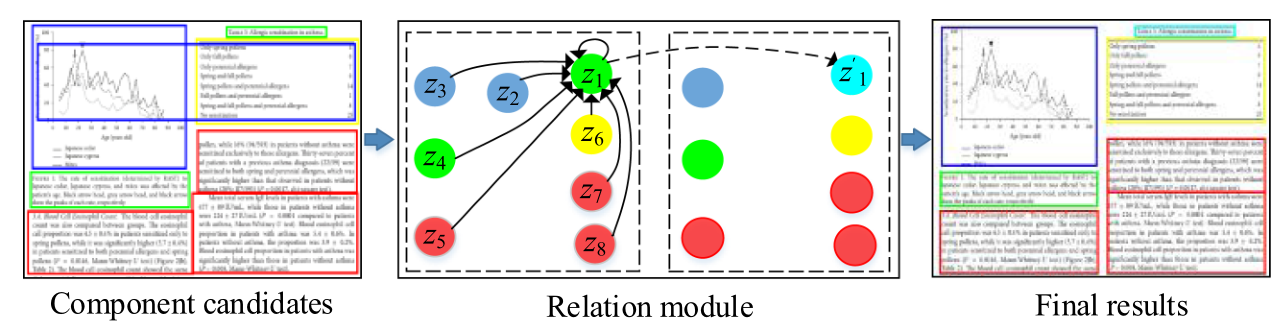
FMi是AMi（注意力图）与视觉特征逐像素乘 +（1-AMi）和语义特征逐像素乘

可以理解为，视觉特征强的时候，不需要太多语义特征，弱的时候加入强语义，AMi控制权衡

FPN [22]（特征金字塔网络）应用于 FM 并提供增强的表示。

**3.4关系模块**

给定聚合特征FM={FM2、FM3、FM4、FM5}，标准对象检测或分割模型（如Mask RCNN[26]）可以在文档中生成候选组件。以前的工作直接将这些预测作为最终结果。但是，布局组件之间存在很强的关系。例如，同一列中段落的边框应对齐；表格和表格标题经常同时出现；组件之间没有重叠。我们发现，这种关系可用于进一步细化预测，如图3所示。调整对齐边界框的回归坐标，基于组件的共现纠正错误的预测标签，并基于非重叠属性去除错误的预测。

图 3. 关系模块示意图。 它捕获候选组件之间的关系，从而提高检测结果（去除错误的图形预测(删除多出的蓝色Z2)，纠正表格标题标签（Z1绿色变浅蓝，因为表格上面为表格标签）和调整段落坐标）。 语义标签的颜色为：Figure,Paragraph,Figure Caption,Table,Table Caption。

接下来，我们将展示我们如何使用 GNN（图神经网络）对组件关系进行建模以及如何使用它来优化预测结果。

我们将文档表示为图 G=(O,E)，其中 O={o1,...,PN} 是节点集，E 是边集。 每个节点 oj 代表一个对象检测模型之前生成的候选组件，每条边代表两个候选组件之间的关系。 由于文档中的远程区域也可能具有紧密的依赖关系（例如，一个段落跨越两列），所有区域都构成了相邻关系。 因此，文档图是一个全连接图且 E⊆O×O。 我们的关系模块的关键思想是通过关注其邻居（z1，z2，...，z8→z1’，如图3 所示）来更新每个节点的隐藏表示。 通过更新的节点特征，我们可以预测其细化的标签和位置坐标。

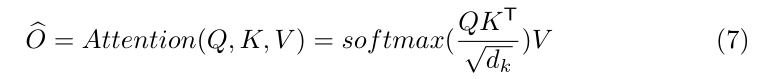
最初，用oj（候选框）=(bj , fj)表示的每个节点包括两条信息：位置坐标bj和深度特征fj=RoIAlign(FM,bj)。 为了将它们都合并到节点表示中，构建了新的节点特征如下，



其中 epos\_ j(bj) 是第 j 个节点的位置嵌入向量。

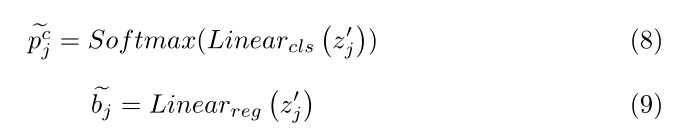
然后，不是明确指定节点之间的关系，受 [33] 的启发，我们应用自注意力机制来自动学习关系，这已经在 NLP 和文档处理中取得了巨大成功 [24, 38,40,42]。 具体来说，我们采用流行的缩放点积注意力[32]来获得足够的表达能力。 缩放点积注意力由 维度的 查询 Q 和 键 K 以及维度的值 V 组成。 输出 是通过对 V 中的所有值进行加权求和获得的，其中注意力权重是使用 Q 和 K 获得的，如等式所示。 (7). 详情请参阅[32]。

(K, V 像是 key-value 的关系从而是一一对应的，那么上式的意思就是通过 Q 中每个元素 query，与K中各个元素求内积然后 softmax 的方式，来得到 Q 中元素与 V 中元素的相似度，然后加权求和，得到一个新的向量。其中因子为了使得内积不至于太大。以上公式在文中也称为点积注意力(scaled dot-product attention)：输出是值的加权和，其中分配给每个值的权重由查询的点积与所有键确定。)



在我们的上下文中，节点特征集Z={z1,...,zN} 用作K、Q 和V，更新的节点特征集Z'={z1',...,zN'} 是输出 。 我们应用多头注意力来进一步提高节点特征的表示能力。 <https://blog.csdn.net/weixin_42691585/article/details/109031214>

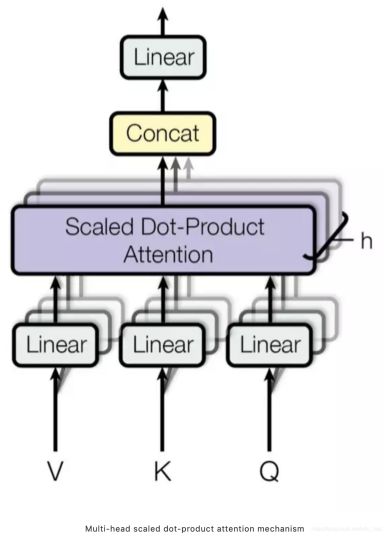
最后，给定更新的节点特征 Z'，第 j 个节点（第 j 个布局组件候选）的细化检测结果  计算为，



其中是属于第c类的概率，是它的细化回归坐标。Linear\_cls和Linear\_reg是投影层。

关系模块可以轻松应用于基于 NLP 的方法。 在这种情况下，等式（6）中的节点特征 zj 是第 j 个低级元素（例如，标记）的表示。 然后，GNN 对标记之间的成对关系进行建模并预测它们的语义标签 ()。

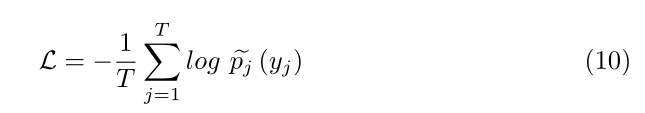
****Multi-Head Self-Attention 不是仅仅计算一次注意力****，而是多次并行地通过缩放的点积注意力。 独立的注意力输出被简单地连接并线性地转换成预期的维度。论文[6]表示，多头注意力允许模型共同关注来自不同位置的不同表示子空间的信息。 只有一个注意力的头，平均值就会抑制这一点。



**3.5 优化**

由于多模态布局分析可以建模为序列标记或对象检测任务，因此它们的优化损失不同

**布局分析作为序列标记**。损失函数被公式化为，



其中，T是低级元素的数量，yj是第 j 个元素的真实语义标签。

**布局分析作为对象检测**。损失函数由两部分生成，



其中L\_DET和L\_RM是候选生成过程和关系模块中使用的损失。 L\_DET 和 L\_RM 都由交叉熵损失（分类）和平滑 L1 损失（坐标回归）组成，如 [26] 中所定义。 超参数 λ 控制两个损失之间的权衡。

**4.1 数据集**

所有三个基准都提供文档图像及其原始 PDF。因此，可以通过解析 PDF 直接获得文本，从而可以探索多模态技术。为了与每个基准测试的现有解决方案进行比较，我们使用与每个基准测试相同的评估指标。

**文章区域**[31]由 822 个文档样本组成，并标注了 9 个区域类（标题、作者、摘要、正文、图、图标题、表格、表格标题和参考文献）。注释采用对象检测格式，评估指标为平均精度 (mAP)。

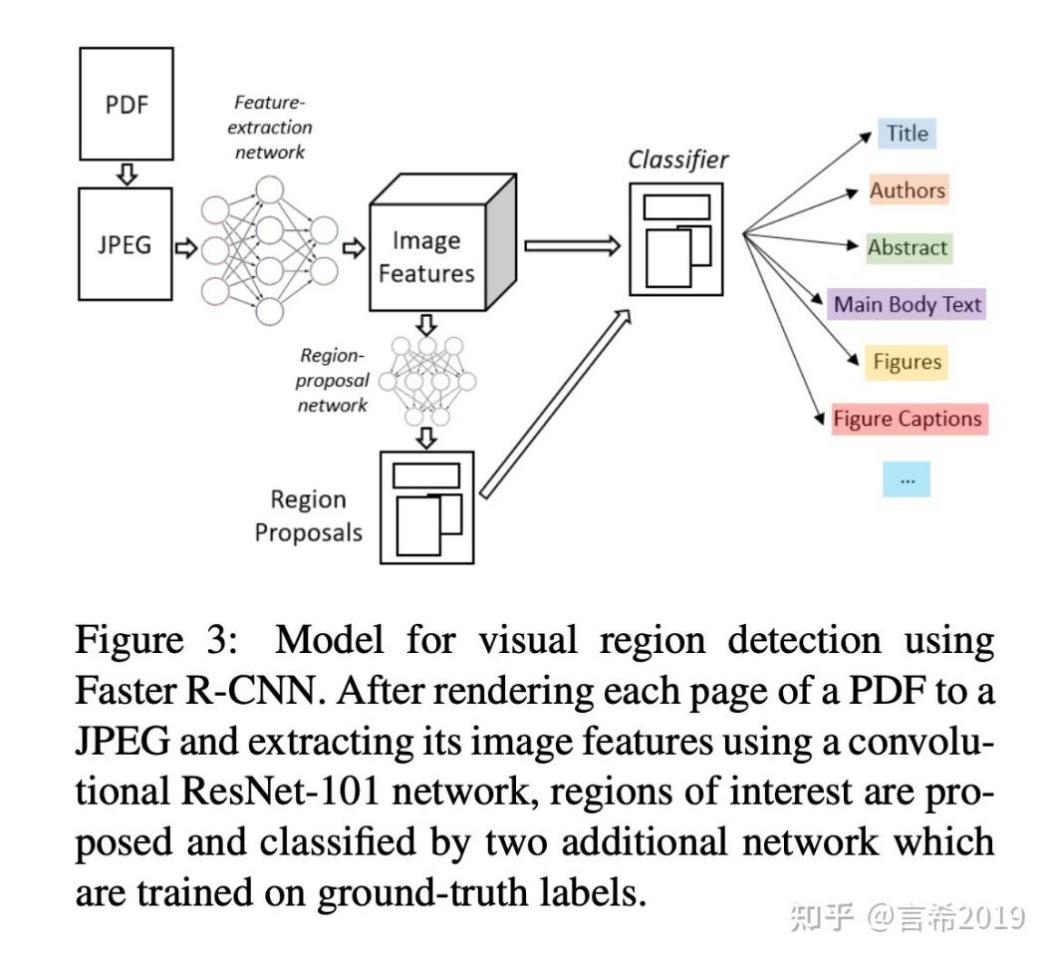
（在ROI（感兴趣区域）中加入对应page信息和box（感兴趣的区域）位置和大小信息

上下文信息被纳入模型的分类和回归阶段，用于文章页面信息和建议的 ROI 边界框。

* 页面上下文由文章中当前图像的页码和文章中的页数组成，两者都标准化为数据集中文章的平均页面长度 (8.22)。
* 边界框上下文由建议的感兴趣区域的位置和大小组成，归一化为图像的尺寸。 两者都被附加到用于批量 ROI 的池化特征向量中，用于发送到分类和边界框回归模型。

如果两个提议的区域重叠，它们的大小和位置可以帮助确定哪个更可能更接近真实的文档区域

faster rcnn(roi层很简单，就是将前面卷积层得到的特征图分为7\*7块(原文)，每块中取最大或者平均值。这样得到7\*7的一个新特征图)



即使添加非常简单的上下文信息（页码和感兴趣区域的位置和大小），性能也比基线提高了 50% 以上

<https://github.com/cxsoto/article-regions>

包含一个脚本，用于从 PubMed Central （医学中心）下载相应的文章 PDF

正在探索其他上下文特征，包括 ROI 过采样（即查看区域周围环境）、随机特征图采样（即查看整页或整篇文章中的模式）和全区域位置信息（即 其他预测 ROI 在文章中的位置）

**PubLayNet**[43]是IBM最近发布的大规模文档数据集。它由 36 万个文档样本组成，并标注了 5 个区域类（文本、标题、列表、图形和表格）。注释也是对象检测格式。他们使用与 COCO 比赛中使用的相同的评估指标，即平均精度 (AP) @ 并集交集 (IOU) [0.50:0.95]。

**DocBank**[20] 是微软提出的。它包含 500K 文档样本，具有 12 个区域类（摘要、作者、标题、方程、图表、页脚、列表、段落、参考、节、表格和标题）。它提供令牌级别的注释，并使用 F1 分数作为官方评估指标。此外，它还提供对象检测注释，支持对象检测方法。

**4.2 实现细节**

文档图像直接用作可视流的输入。对于语义流，我们从文本中提取嵌入映射（SentGrid和CharGrid）作为输入，其中SentGrid由预训练的BERT模型[8]生成，CharGrid由单词嵌入层获得（**Embedding**）。它们都具有相同的通道维度大小（CS0=64）。ResNeXt-101[37]被用作提取视觉和语义特征的主干（除非另有规定），这些特征随后通过多尺度自适应聚合和特征金字塔网络进行融合。

对于基于CV的多模态布局分析方法，将融合特征输入RPN(区域提案网络)，然后再输入RCNN，以生成候选组件。在RPN中，采用7个锚定比率（0.02、0.05、0.1、0.2、0.5、1.0、2.0）来处理大小和比例不同的文档元素。在关系模块中，每个候选对象的维度设置为1024，并使用2层16个头的多头注意来建模关系。在所有实验中，我们将等式（11）中的λ设为1。

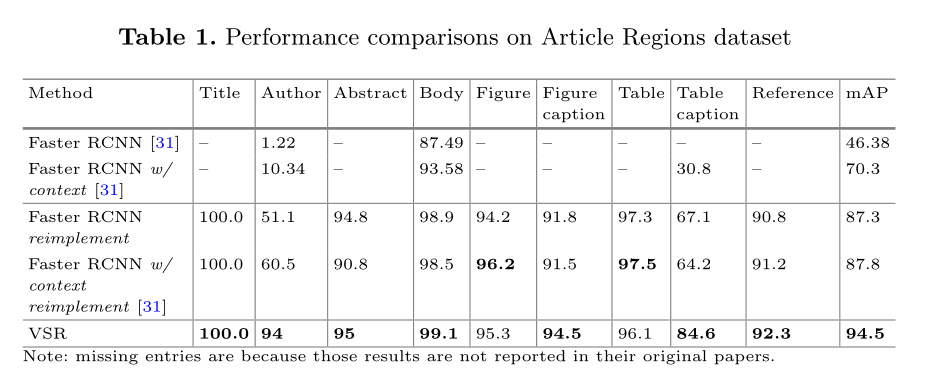
对于基于NLP的多模式布局分析方法，从PDF解析的低级元素（例如标记）用作组件候选，关系模块预测其语义标签。

我们的模型是在PyTorch框架下实现的。它由SGD优化器进行训练，batchsize=2，动量=0.9，重量衰减=10−4. 初始学习速率设置为10−3，在Article Regions数据集上每10个epoch除以10，在其他两个基准上除以3个epoch。关于文章区域的模型培训持续30个epoch，而关于其他两个基准的培训持续6个epoch。所有实验均在特斯拉-V100 GPU上进行。 源代码将在不久的将来发布。

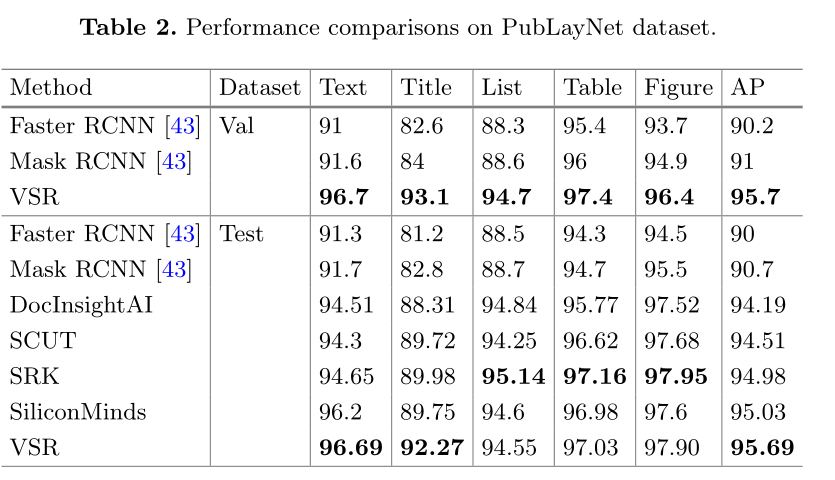
**4.3 结果**

**文章区域**。我们比较了两种模型在该数据集上的VSR性能：更快的RCNN和更快的带上下文的RCNN[31]。带上下文的更快RCNN除了文档图像外，还添加有限的上下文（页码、感兴趣区域位置和大小）作为输入。

在表1中，我们首先显示了他们原始论文中报告的地图[31]。为了公平比较，我们使用VSR中使用的相同主干（ResNet101）和颈部配置重新实现了这两个模型。我们还报告了重新实施后的性能。我们可以看到，我们重新实现的模型比其原始模型具有更高的映射。我们认为这主要是因为我们在RPN中使用了多个锚定比率，从而在不同大小的文档元素上实现了更好的检测结果。VSR充分利用了视觉、语义和组件之间的关系，在大多数类上显示最高的映射。在图形和表格类别上，VSR取得了可比的结果，轻微的性能下降将在第4.4节中进一步讨论。



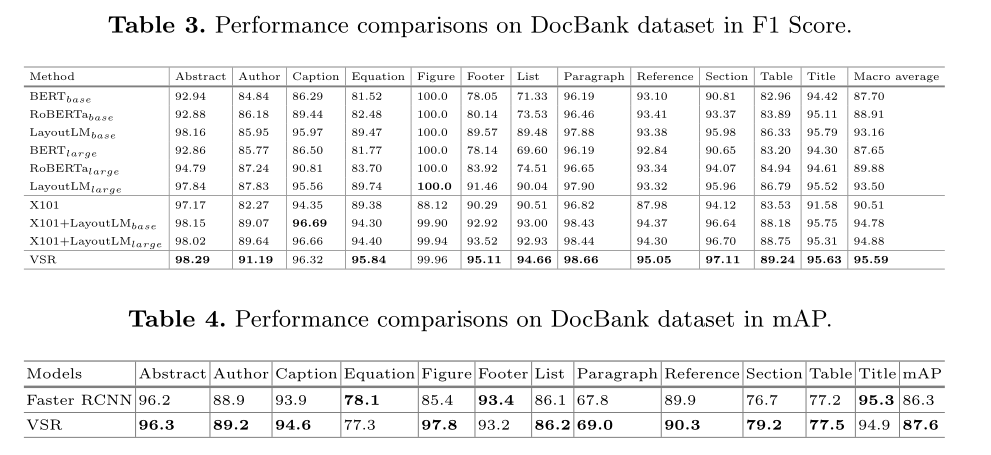
**PubLayNet**.在表2中，我们将VSR在该数据集上的性能与两种纯基于图像的方法（更快的RCNN[26]和掩码RCNN[13]）进行了比较。虽然这两个模型在验证数据集上显示了有希望的结果（AP>90%），但VSR提高了所有类的性能，并将最终AP提高了4.7%。VSR在与文本相关的类（文本、标题和列表）上显示了巨大的性能改进，因为除了文档图像之外，它还利用了文档语义。在测试数据集（也称为ICDAR221布局分析识别竞赛排行榜4）上，VSR超过所有参赛团队并排名第一，与Mask RCNN基线相比，AP增加4.99%。



**DocBank。**此数据集提供令牌和检测注释。因此，我们可以将布局分析任务视为序列标记任务或对象检测任务，然后在这两种情况下将VSR与现有解决方案进行比较。

**布局分析作为序列标记**。使用令牌级注释，我们将VSR与BERT[8]、RoBERTa[25]、LayoutLM[38]进行比较，将更快的RCNN与ResNeXt-101[37]和表3中的集成模型（ResNeXt-101+LayoutLM）进行比较。尽管标题和图形的F1最高分数分别通过集成模型（ResNeXt-101+LayoutLM）和LayoutLM实现，但VSR在较小的差距下实现了可比的结果(≤0.37%). 更重要的是，VSR在所有其他课程中得分最高。这表明VSR在文档布局分析任务上明显优于BERT、RoBERTa和LayoutLM体系结构。

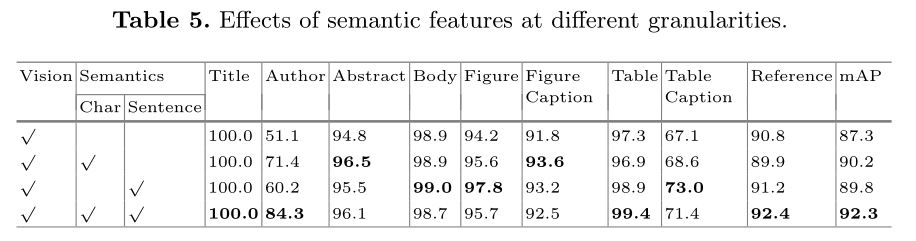
**布局分析作为目标检测**。由于VSR和带有ResNeXt-101的快速RCNN都可以提供目标检测结果，因此我们使用mAP作为评估指标，在目标检测格式中进一步比较了它们。表4中的结果表明，除了等式、页脚和标题外，VSR在大多数类上都优于更快的RCNN。总体而言，VSR在最终地图中显示1.3%的增长。



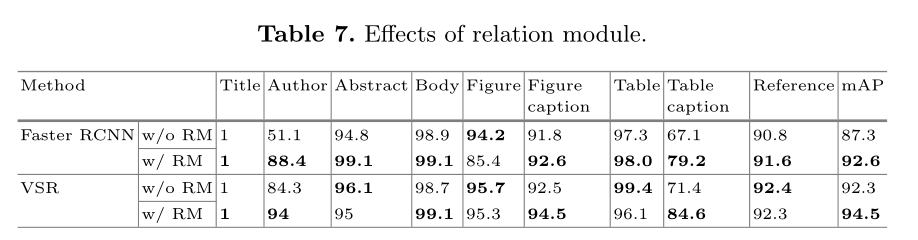
**4.4 消融研究**

VSR引入了多粒度语义、自适应聚合的双流网络和关系模块。现在，我们将探讨每种方法如何在Article Regions数据集上提高VSR的性能。

**多粒度语义特征的影响**。为了了解多粒度语义特征是否确实提高了VSR的性能，我们比较了表5中的4个版本的VSR（仅视觉、视觉+字符、视觉+句子、视觉+字符+句子）。在这里，字符与语义特征在两个不同的粒度上是相关的。我们可以看到，在每个粒度上单独引入文档语义可以提高分析性能，同时将两者结合起来可以获得最高的映射。这与人类理解文档的方式是一致的。人类可以更好地从字符/单词（例如，作者）和句子（例如，表格标题）中识别需要少量上下文的区域。



**关系模块的影响**。为了验证关系模块（RM）的有效性，我们比较了表7中两个版本的更快的RCNN和VSR，即有RM和没有RM。由于标签和位置坐标都可以在RM中细化，单峰快速RCNN和VSR在加入关系模块后都显示出一致的改进，分别增加了5.3%和2.2%。可视示例如图4所示。然而，对于图形组件，在引入RM后，性能可能会略有下降。原因是，在删除重复预测的同时，我们的关系模块也可能会冒删除正确预测的风险。但是，我们仍然看到整体表现有所改善，显示出引入关系的好处。



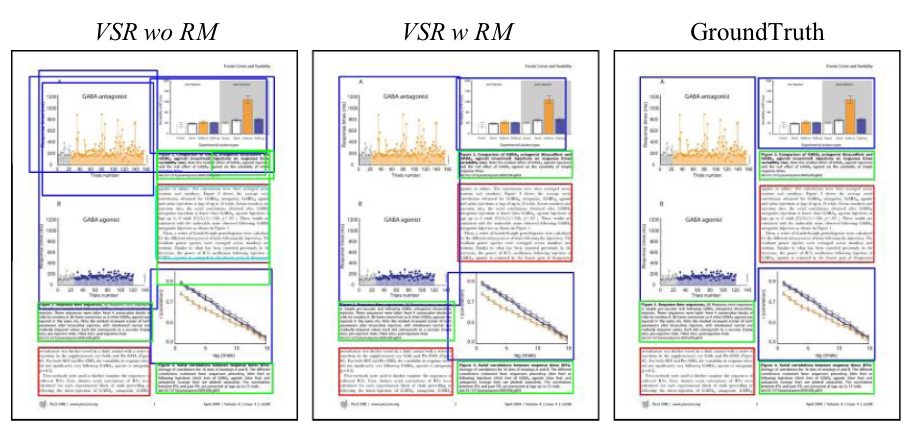


图4：VSR（w/wo）和RM之间的定性比较。引入RM可有效消除重复预测，并提供更准确的检测结果（标签和坐标）。语义标签的颜色有：图形、主体、图形标题。

**限制**。如上所述，除了文档图像，VSR还需要文档中文本的位置和内容。因此，VSR的泛化能力可能不如单峰VSR，我们将在将来讨论。

**5 结论**

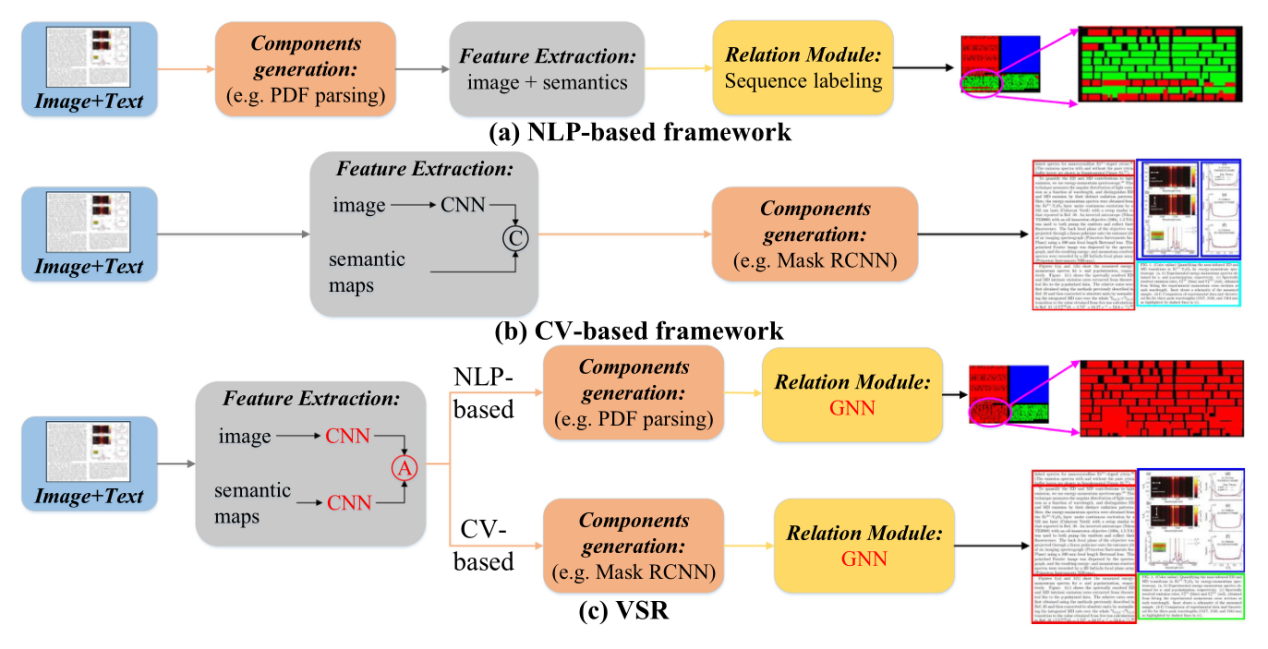
在本文中，我们提出了一个结合视觉、语义和关系的多模态布局分析的统一框架VSR。我们首先从字符和句子粒度的角度介绍文档的语义。然后，使用两流卷积网络提取特定于模态的视觉和语义特征，并在自适应聚合模块中进一步融合。最后，对于给定的候选组件，采用关系模块对它们之间的关系进行建模，并输出最终结果。在三个基准上，VSR的性能显著优于其单峰和多峰单流。在未来，我们将研究VSR的预训练模型，并将其扩展到其他任务，如信息提取。

缺点： 语义信息是pdf解析出来的，如果没有pdf，很难利用

**总结：**

现有文档版面分析方法大致可分为两种：

1. 基于NLP的方法将版面分析任务看作是序列标注任务（sequence labeling），但是该类方法在版面建模上表现出不足，无法捕获空间信息；
2. 基于CV的方法则将版面分析看作是目标检测或分割任务（object detection or segmentation），该类方法的不足表现在（1）不同细粒度的语义（2）简单的拼接方式（3）未利用关系信息。



**VSR**: Vision(Cascade Mask R-CNN) + Semantic(text embedding) + Relation(GNN)

主要包括三个模块：

双流卷积网络(two-stream ConvNets)、多尺度自适应聚合模块、关系学习模块。

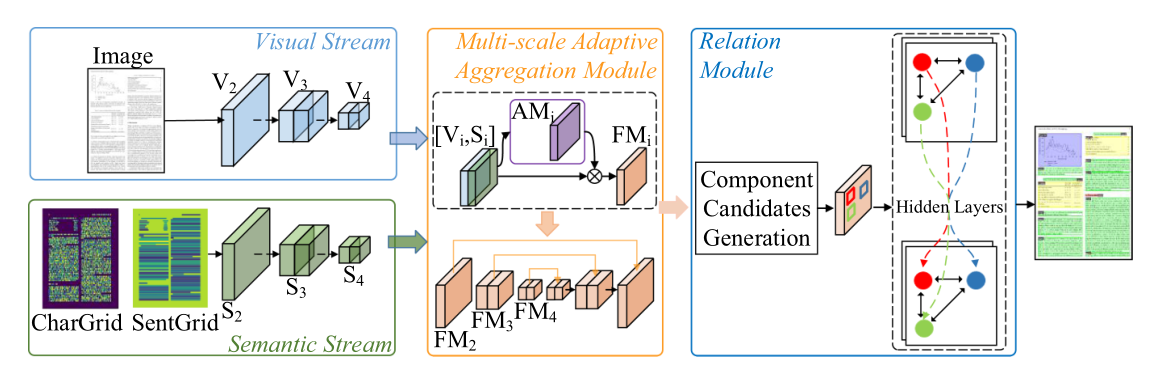
•1.双流卷积网络提取视觉和语义特征；

•2.相对于简单的拼接，多尺寸自适应聚合模块来得到视觉和语义双模态信息表示；

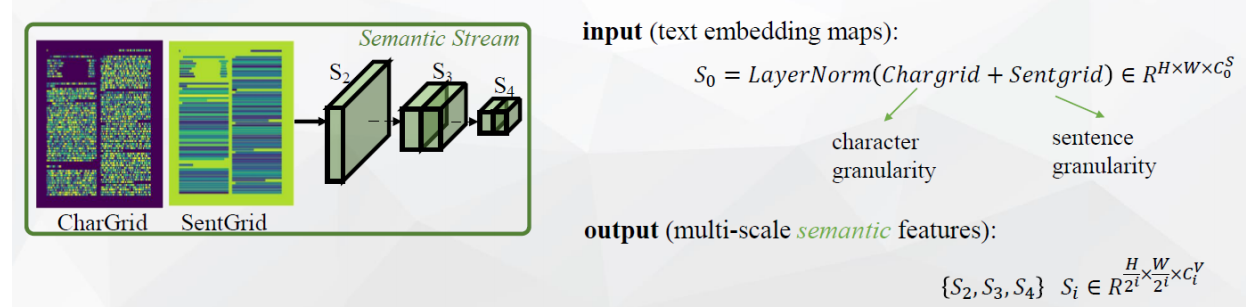
•3.通过Mask R-CNN得到各个候选框；•4.最后，关系学习模块来学习各组件候选间的关系，并生成最终结果。

**视觉分支**

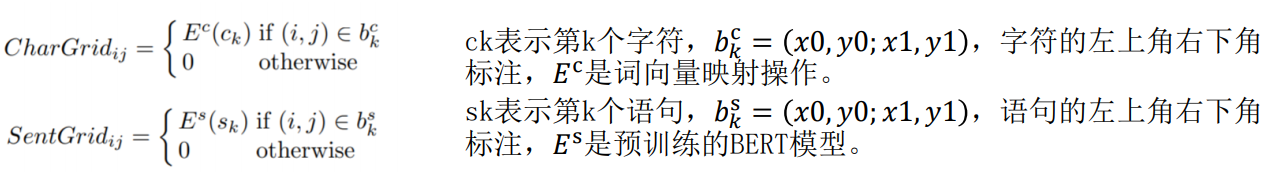
ResNet系列的主干网络提取的多层级特征V2，V3，V4，V5，分别是1/2^2，1/2^3...下采样倍率



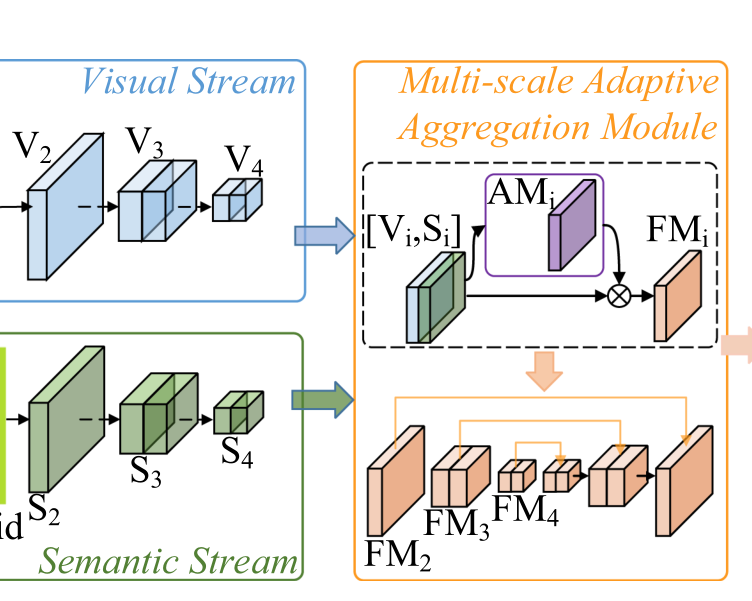
**语义分支：**



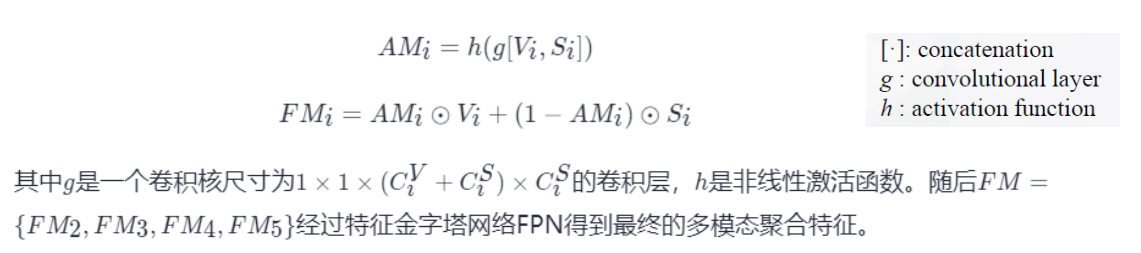
CharGrid以及SentGrid都是通过pdf解析获得的



**融合模块**



在得到各模态的特征图表示后，通过一个卷积层得到AMi，再基于attention map对各模态特征进行聚合得到聚合特征图FMi



**关系模块**

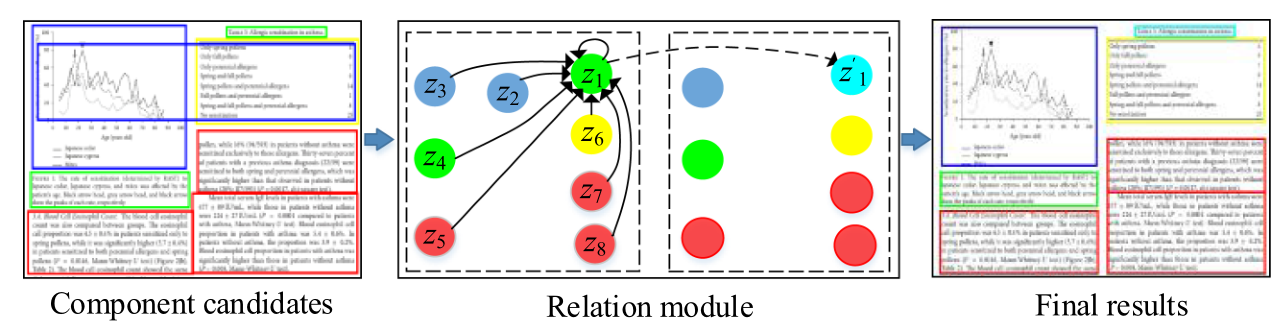
通常检测分割框架会直接输出结果（Mask R-CNN），这里利用一些布局上的结构关系，使用GNN作refine，在得到FM后，可以通过RPN网络很容易得到ROI（Region of Interest）作为布局部件候选集。

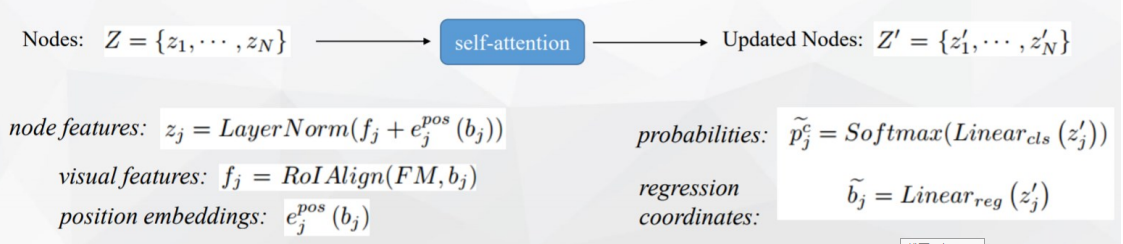
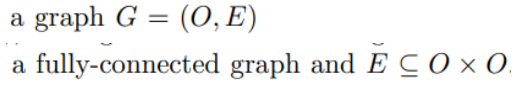
•（1）利用空间位置关系，调整文本框坐标，例如某些段落的空间坐标的对齐性质；

•（2）根据部件间的共现关系（比如表格和表格标题一般会同时出现）修正预测标签；

•（3）各component间不重叠的特性移除多余框。VSR中关系学习模块对各部件候选间的关系进行建模，并最终得到版面分析的结果。

将一篇文档作为一张图graph，而每个部件候选component candidate作为节点node。每个节点的特征表示由多模态特征表示和位置信息表示组成，通过self-attention机制来学习关系，计算节点即部件候选的分类概率、边框坐标回归：

全连接：由于文档中的远程区域也可能具有紧密的依赖关系（例如，一个段落跨越两列），所有区域都构成了相邻关系。



**conclusion：**

**Highlights**

–结合视觉和语义做版面分析，文本语义用character和sentence两种粒度来表示；

–采用two-stream convnet来分别提取视觉和语义特征，再通过attention将两个模态特征聚合，最后基于聚合特征得到component candidates（组件候选）；

–Graph attention networks来学习各component candidates间的关系。

–在多个数据集上达到SOTA，其中ICDAR2021的PubLayNet刷到冠军

**Limitations**

–需要预先获取文本内容和位置信息，这里利用了PDF的解析结果，对于纯图片输入可能需要过一个通用文本检测识别器，相比一般的目标检测分割方法性价比低。适合PDF提取版面。