

TextVQA 的结构化多模态注意力机制

Chenyu Gao, Qi Zhu, Peng Wang[†], Hui Li, Yuliang Liu, Anton van den Hengel, Qi Wu[†]

Abstract—基于文本的视觉问答 (TextVQA) 是最近提出的一项挑战，要求模型读取图像中的文本并通过联合推理问题、文本信息和视觉内容来回答自然语言问题。这种新模态——光学字符识别 (OCR) 令牌的引入带来了苛刻的推理要求。大多数最先进的 (SoTA) VQA 方法在回答这些问题时失败，原因有三：(1) 文本阅读能力差；(2) 缺乏文本-视觉推理能力；(3) 选择判别性回答机制而不是生成性对应机制（尽管 M4C 已进一步解决了这个问题）。在本文中，我们提出了一种端到端结构化多模态注意 (SMA) 神经网络，主要解决上述前两个问题。SMA 首先使用结构化图表示来编码图像中出现的对象-对象、对象-文本和文本-文本关系，然后设计一个多模态图注意网络对其进行推理。最后，上述模块的输出由全局-局部注意力应答模块处理，以按照 M4C 迭代地将来自 OCR 和一般词汇的标记拼接在一起，从而产生答案。除了基于预训练的 TAP 之外，我们提出的模型在所有模型中都优于 TextVQA 数据集和 ST-VQA 数据集的两个任务上的 SoTA 模型。该模型还因表现出强大的推理能力而在 TextVQA Challenge 2020 中夺得第一名。我们在多个推理模型上广泛测试了不同的 OCR 方法，并研究了逐渐提高的 OCR 性能对 TextVQA 基准的影响。凭借更好的 OCR 结果，不同模型在 VQA 准确率上都有了显著的提高，但我们的模型受益于强大的文本-视觉推理能力。为了赋予我们的方法一个上限并为进一步的工作提供公平的测试基础，我们还提供了 TextVQA 数据集的人工注释的真实 OCR 注释，这在原始版本中是没有提供的。TextVQA 数据集的代码和真实 OCR 注释可以在 <https://github.com/ChenyuGAO-CS/SMA> 获得。

Index Terms—TextVQA, Graph Attention Network, Transformer.

Fig. 1: (左) 在问题条件图注意模块中，我们构建了一个异构图。我们用黄色表示对象，用红色表示 OCR 标记。实线框表示与问题最相关的节点，而虚线框表示其他节点。理解多种关系对于回答这个问题至关重要，e.g. “数字 12 以上的单词”是一种文本与文本的关系。(右) 抽象实体和关系。红色菱形是 OCR 标记，黄色矩形是对象。蓝色箭头是关系。

1 介绍

VQA 伊苏阿尔问答系统 (VQA) [?] 凭借深度神经网络的发展取得了长足进步。然而，最近的研究 [?], [?], [?] 表明，大多数 VQA 模型不幸在需要理解图像中文本的一类问题上失败了。VizWiz [?] 首先发现了这个问题，并发现视障人士提出的问题中近四分之一与文本阅读有关。Singh [?] 系统地研究了这个问题，并引入了一个新数据集 TextVQA，其中仅包含需要模型读取和推理图像中文本的问题。

解决 TextVQA 问题必不可少的三个关键能力是阅读、推理和回答，这也是为什么最先进的 (SoTA) VQA 模型在这个任务上表现不佳的主要原因。阅读能力依赖于光学字符识别 (OCR) 技术来准确检测和识别图像中出现的文本，这已经是计算机视觉的一个长期存在的子领域，而推理需要一个模型对图像中的视觉内容和 OCR 文本进行联合推理。SoTA VQA 模型 [?], [?] 可以通过一些复杂的机制（如注意力 [?] 和记忆网络 [?]）在视觉内容和自然语言问题上获得很强的推理能力，但它们都不能准确地读取图像中的“文本”，更不用说对它们进行推理了。TextVQA 中提供的方法 LoRRA [?] 虽然配备了 OCR 模型来读取文本，但由于缺乏对文本和视觉内容的深度推理，结果并不突出。在答题方面，几乎所有的 SoTA

• [†] PW and QW are corresponding authors.

• C. Gao, Q. Zhu and P. Wang are with the Northwestern Polytechnical University. H. Li, Y. Liu, A. van den Hengel and Q. Wu are with the University of Adelaide

Manuscript received April 19, 2005; revised August 26, 2015.

VQA 模型都选择使用判别式答题模块，因为它易于优化，并且在传统 VQA 数据集上表现更佳。然而 TextVQA 中的答案通常是从图片中检测到的 OCR token 和一般文本 token 的组合，因此答案词汇并不固定。判别式答题模块可能会限制输出的多样性。

图 1 展示了 TextVQA 中的一个涉及多种关系类型的示例。例如，“衬衫正面”、“球员衬衫”是对象与对象的链接；“球员衬衫正面印有的单词”是文本与对象的联系，“数字 12 上方的单词”是文本与文本的关系。在本文中，为了增强关系推理能力，我们引入了一个 SMA 模型来对具有多种关系类型的图进行推理。具体来说，问题自注意模块首先将问题分解为六个子组件，分别表示对象、对象-对象关系、对象-文本关系、文本、文本-文本关系和文本-对象关系。然后以对象/文本为节点构建角色感知图。节点之间的连接由相对距离决定。然后使用问题条件图注意模块更新图。在该模型中，不是使用整个问题块来指导图更新，而是仅使用从问题自注意模块中提取的某些类型的问题组件来更新相应的图组件。例如，与对象相关的问题特征用于对象节点，而与对象-文本相关的问题特征仅用于对象-文本边缘更新过程。最后，为了解决上述回答问题，我们将 M4C [?] 中的迭代答案预测机制引入到我们的全局-局部注意模块中，但我们用问题、对象和文本的总结的全局特征以及局部 OCR 嵌入替换第一步输入 $\langle \text{begin} \rangle$ 。M4C 解码第一步的原始输入是一个特殊的 token $\langle \text{begin} \rangle$ ，没有更多信息，而我们总结的全局特征包含了我们问题条件图的全面信息，为我们的模型带来了近 0.5% 的提升。中间过程使用的所有特征都是以端到端的方式自动学习的，这使我们能够灵活地适应具有不同关系类型的不同实例。唯一的监督是每个实例的 Ground-Truth 答案。

我们在最近发布的 TextVQA [?] 和 ST-VQA [?] 数据集上验证了我们提出的 SMA 模型的有效性，并且在 TextVQA 和 ST-VQA 的前两个任务上超越了之前的 SoTA 模型。我们提出的 SMA 模型还赢得了 TextVQA Challenge 2020¹。

为了进一步研究“阅读”部分和“推理”部分的贡献，我们

¹ <https://evalai.cloudcv.org/web/challenges/challenge-page/551/leaderboard/1575>

研究了如果使用固定的推理模型，OCR 性能对 TextVQA 准确率的影响有多大。我们使用 SoTA 文本检测器和识别器对 TextVQA 进行文本检测和识别。与 Rosetta OCR [?] 的结果相比，发现我们的所有模型都有所改进，LoRRA 和 M4C [?] 也是如此，但我们的 SMA 受益更多，OCR 结果更好，这证明我们的模型具有更好的文本视觉推理能力。为了完全剥离 OCR 对研究真实推理能力的影响，我们要求 AMT 工作人员对 TextVQA 数据集中出现的所有文本进行注释，从而得到 709,598 真实 OCR 注释。这些注释在原始 TextVQA 中没有给出，我们会将它们发布给社区以进行公平的比较。我们还通过提供真实 OCR 报告了 LoRRA、M4C 和我们的最佳模型的性能，以便单独测试模型的推理能力。还通过使用真实 OCR 注释给出了新的上限。

总而言之，我们的贡献有三点：

- 1) 我们提出了一种结构化多模态注意力 (SMA) 模型，该模型可以有效地推理结构化文本对象图并以生成方式产生答案。由于采用了图形推理策略，所提出的模型实现了更好的可解释性。
- 2) 我们研究 OCR 在 TextVQA 问题中的贡献，并提供人工注释的真实 OCR 标签来完善原始 TextVQA 数据集。这使得社区中的追随者只需在完美的阅读情况下评估其模型的推理能力。
- 3) 我们的 SMA 模型在 TextVQA 数据集和 ST-VQA 数据集的两个任务上均优于现有的最先进的 TextVQA 模型（基于预训练的模型 TAP [?] 除外，因为它不能直接比较），从而成为 TextVQA Challenge 2020 的冠军模型。

2 相关工作

2.1 基于文本的 VQA

横跨计算机视觉和自然语言处理领域，自大型 VQA 数据集 [?] 发布以来，视觉问答 (VQA) 引起了越来越多的关注。出现了大量的方法和数据集：引入了 CLEVR [?] 和 FigureQA [?] 等 VQA 数据集来纯粹研究视觉推理，而不考虑 OCR 标记；Wang 等人 [?] 引入了一个明确需要外部知识来回答问题的数据集。

阅读和推理图像中的文本对于视觉理解具有重要价值，因为文本包含丰富的语义信息，而这正是 VQA 主要关注点。近年来，出现了一些数据集和基线方法，旨在研究视觉和文本内容的联合推理能力。例如，Textbook QA [?] 根据中学教科书中的文本、图表和图像提出多模态问题。FigureQA [?] 需要根据合成的科学风格图形（如线图、条形图或饼图）回答问题。DVQA [?] 在 VQA 框架中评估条形图理解能力。在这些数据集中，文本是机器打印的，以标准字体显示，质量很好，这减轻了文本识别的难度。Vizwiz [?] 是第一个需要文本信息来回答问题的数据集，给出在自然场景中捕获的图像。然而，由于图像质量差，58% 的问题“无法回答”，这使得该数据集不适合训练有效的 VQA 模型和系统地研究该问题。

最近，TextVQA [?] 和 ST-VQA [?] 同时被提出，以强调在 VQA 过程中从自然场景图像中读取文本的重要性。TextVQA 中提出了 LoRRA，它在图像对象和 OCR 文本上使用简单的 Updn [?] 注意框架来推断答案。然后通过使用基于 BERT [?] 的词嵌入和基于多模态分解高阶池化的特征融合方法改进模型，并在 TextVQA 挑战赛中获胜。与一旦需要读取文本就允许回答任何问题的 TextVQA 相比，ST-VQA 中的所有问题都可以直接通过图像中的文本明确地回答。ST-VQA 采用堆叠注意力网络 (SAN) [?] 作为基线，通过简单地将文本特征与图像特征连接起来进行答案分类。先前模型（如 LoRRA [?]）中的回答模块遇到了两个瓶颈。一个严重的缺点是它们

将动态 OCR 空间视为不变索引，另一个是无法生成由多个单词组成的长答案。MM-GNN [?] 专注于构建可视化图、语义图和数字图来表示场景文本的信息。M4C [?] 首先通过变压器解码器和动态指针网络解决了这两个问题。继 M4C 之后，LaAP [?] 进一步预测 OCR 位置作为最终预测答案的证据；SA-M4C [?] 用新颖的空间感知自注意力层取代了原有的自注意力层；Singh 等 [?] 提出了一个带有端到端 PixelM4C 模型的 OCR 数据集（900k 个带注释的任意形状词），该模型将 Mask TextSpot-ter (MTS) v3 [?] 与 M4C [?] 连接起来，可以同时提取 OCR 和执行基于文本的任务。MTXNet [?] 专注于文本和视觉解释，提出了一个用于生成解释的数据集 TextVQA-X，并发现现有的 TextVQA 模型可以轻松地调整以生成多模态解释。Yong 等 [?] 致力于提出预训练任务，以学习文本词、视觉对象和场景文本的更好对齐表示，并提高基于文本的任务（如 TextVQA 和 Textcaps [?]）的性能。在这项工作中，我们专注于明确建模对象、文本和对象-文本对之间的关系，并取得了比以前的 TextVQA 方法更好的性能和可解释性。

2.2 视觉和语言中的图形网络

图网络因其在结构特征学习方面的表达能力而备受关注。它们不仅可以捕获节点本身的特征，还可以对图中节点之间的邻域属性进行编码，这对于 VQA 以及其他需要结合空间和语义信息结构的视觉和语言任务至关重要。例如，Teney 等人 [?] 分别在图像场景对象和疑问词上构建图形，以利用这些表示中的结构信息。该模型在一般 VQA 任务中显示出显着的改进。Narasimhan 等人 [?] 通过图卷积网络 (GCN) 考虑事实列表以选择正确答案，对事实 VQA 任务 [?] 执行更精细的关系探索。工作 [?] 学习 VQA 中输入图像的问题特定图形表示，通过空间图卷积捕获对象与相关邻居的交互。MUREL [?] 更进一步，为关系推理建立了所有区域对之间的空间语义成对关系模型，此外还为区域的视觉内容和问题之间的交互提供了丰富的矢量表示。

我们的工作也使用图作为表示，但与以前使用全连接图连接所有对象的方法不同，我们的任务需要同时考虑图像中的视觉元素和文本信息，而这些元素和信息本质上是异构的。我们构建了一个角色感知图，该图考虑了节点（例如对象和文本）和边（“对象-对象”、“文本-文本”和“对象-文本”）的不同角色，从而为答案推断提供了更好的跨模态特征表示。

2.3 场景文本检测与识别

自然场景图像中的 OCR（包括文本检测和识别）结果在 TextVQA 中起着重要作用。然而由于文本模式极其多样且背景高度复杂，自然场景中的 OCR 本身是一项非常具有挑战性的任务，在计算机视觉界引起了广泛关注。大多数工作分别解决这个问题，通过设计高性能的文本检测器 [?], [?], [?], [?] 和复杂的单词识别器 [?], [?], [?], [?], [?]，将它们结合起来最终获取图像中的文本内容。最近也提出了一些工作，在一个框架中同时解决文本检测和识别 [?], [?], [?], [?], [?]。在深度神经网络和大规模数据集的驱动下，性能和速度都取得了实质性的进展。方法从仅处理水平文本改进到处理任意形状的文本。巨大的改进为基于视觉信息和文本线索的 VQA 奠定了基础。LoRRA 采用 Rosetta OCR [?] 进行文本识别，这是一个两步模型，基于 Faster-RCNN 的框架用于文本检测，全卷积模型带有 CTC 损失，用于单词识别。Rosetta 无法很好地读取不规则文本（有方向、扭曲）。在我们的模型中，首先使用 SoTA 无序列框离散化 (SBD) 模型 [?] 进行场景文本检测，然后使用基于稳健变压器的网络 [?] 进行单词识别（表示为“SBD-Trans OCR”），这实现了更好的 OCR 结果并改善了 TextVQA 任务。

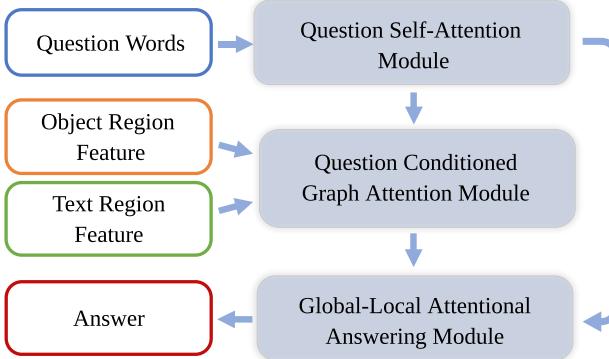


Fig. 2: 我们的 SMA 模型中有三个模块。问题自注意力模块将问题分解为引导信号，引导图形注意力模块对图形进行推理，以及全局-局部注意力回答模块生成答案。

3 方法

在本节中，我们将介绍结构化多模态注意力 (SMA) 模型。从高层次来看，SMA 由三个模块组成，如图 2 所示：(1) 问题自注意力模块，将问题分解为构建的对象文本图中的六个不同角色的子组件 w.r.t.。(2) 问题条件图注意力模块，该模块在上述问题表示的指导下对图进行推理，并推断出不同节点的重要性及其关系。(3) 全局-局部注意力回答模块，可以生成多个单词拼接在一起的答案。我们的模型是一个端到端框架，唯一的监督是每个实例的真实答案。所有子特征都是在训练过程中自动学习的，这使我们能够灵活地适应具有不同类型关系的不同实例。我们将在以下章节中详细介绍每个模块。

Notation 在本文的其余部分，矩阵用粗体大写字母表示，列向量用粗体小写字母表示。 \circ 表示元素乘积。 $[;]$ 表示连接。

3.1 问题自注意力模块

由于 TextVQA 问题可能不仅包含对象和文本节点的信息，还包含它们之间的四类关系（对象-对象、对象-文本、文本-文本和文本-对象），我们的问题自注意力模块（见图 3）首先将问题分为六个子组件。它类似于领域知识引导的多头注意力，因为不同的问题组件对应不同的注意力头，它们是细粒度的，具体应用于每个节点和边。虽然这是受到 [?], [?] 的启发，但我们的模块更细粒度，并且是为 TextVQA 任务精心设计的。

形式化地，给定一个问题 Q ，其中有 T 个词 $q = \{q_t\}_{t=1}^T$ ， $\{\mathbf{x}_t^{bert}\}_{t=1}^T$ 是使用预先训练的 BERT [?] 得到的。分解的问题特征 $(\mathbf{s}^o, \mathbf{s}^{oo}, \mathbf{s}^{ot}, \mathbf{s}^t, \mathbf{s}^{tt}, \mathbf{s}^{to})$ 被视为问题表示分解的 w.r.t. 对象节点 (\mathbf{o})、对象-对象 (\mathbf{oo}) 边、对象-文本 (\mathbf{ot}) 边、文本节点 (\mathbf{t})、文本-文本 (\mathbf{tt}) 边和文本-对象 (\mathbf{to}) 边。以 \mathbf{s}^o 为例，计算如下：

$$\begin{aligned} \mathbf{a}_t^o &= \text{softmax}(\text{MLP}_{obj}^a(\mathbf{x}_t^{bert})), \quad t = 1, \dots, T, \\ \mathbf{s}^o &= \sum_{t=1}^T \mathbf{a}_t^o \mathbf{x}_t^{bert}. \end{aligned} \quad (1)$$

其他特征以相同的方式计算，并且这些分解的问题特征在第 3.2 节中执行问题条件图注意时用作指导信号。

我们还对平均隐藏状态应用了另一系列变换，以生成两组问题自注意力权重，它们将作为先验概率用于 3.3 节中的最终特征组合。更具体地说， $\{\mathbf{w}^o, \mathbf{w}^{oo}, \mathbf{w}^{ot}\}$ 和 $\{\mathbf{w}^t, \mathbf{w}^{tt}, \mathbf{w}^{to}\}$ 分别用于生成对象特征 \mathbf{g}_{obj} 和文本特征 \mathbf{g}_{text} 。

所有上述子特征都是以端到端训练方式学习的，无需任何手工编码的监督，设计简单，适应性强。注意权重是在语言自

注意模块中学习的，它们将发挥什么样的作用取决于它们查询的查询键的类型（在我们的例子中，键是节点和边的表示）。

3.2 问题条件图注意模块。

问题条件图注意模块（如图 4 所示）是我们网络的核心，它针对图像的对象和文本生成异构图，然后对其进行推理。

Role-aware Heterogeneous Graph Construction. ‘角色’表示不同类型的节点。我们在图像 I 的对象节点和文本节点上构建一个角色感知的异构图 $\mathcal{G} = \{\mathcal{O}, \mathcal{T}, \mathcal{E}\}$ ，其中 $\mathcal{O} = \{o_i\}_{i=1}^N$ 是 N 对象节点集， $\mathcal{T} = \{t_i\}_{i=N+1}^{N+M}$ 是 M 文本节点集， $\mathcal{E} = \{e_{ij}\}$ 是边集。在我们的图中，边表示两个特定节点之间的关系，每个节点可以连接到 $k = 5$ 对象节点和 $k = 5$ 文本节点。很明显，我们的图中的节点和边具有不同的角色，因此我们称之为异构图。‘角色感知’意味着我们明确使用每个节点的角色信息来构建图。我们又可以根据边的作用，将边进一步分为四组： \mathcal{E}^{oo} 对应 \mathbf{oo} 边， \mathcal{E}^{ot} 对应 \mathbf{ot} 边， \mathcal{E}^{tt} 对应 \mathbf{tt} 边， \mathcal{E}^{to} 对应 \mathbf{to} 边。具体参见图 5。

然后，我们根据两个节点的相对空间关系构建它们之间的边表示。这里我们再次以构建一条 \mathbf{oo} 边为例。假设节点 o_i 的中心坐标、宽度和高度表示为 $[x_i^c, y_i^c, w_i, h_i]$ ，另一个节点 o_j 的左上坐标、右下坐标、宽度和高度表示为 $[x_j^{tl}, y_j^{tl}, x_j^{br}, y_j^{br}, w_j, h_j]$ ，则关联的边表示定义为 $e_{ij} = [\frac{x_j^{tl}-x_i^c}{w_i}, \frac{y_j^{tl}-y_i^c}{h_i}, \frac{x_j^{br}-x_i^c}{w_i}, \frac{y_j^{br}-y_i^c}{h_i}, \frac{w_j \cdot h_j}{w_i \cdot h_i}]$ 。

Question Conditioned Graph Attention. 我们使用第 3.1 节中分解的问题特征 \mathbf{s} 来推理上一节中构建的角色感知图。

我们将推理过程表述为一种注意力机制，但我们不是使用单个问题特征来应用全局注意力权重，而是根据不同问题特征的角色来更新图的不同部分。例如，与对象相关的问题表示 \mathbf{s}^o 用于指导对象节点上的注意力权重，而 \mathbf{s}^{to} 用于指导文本-对象边注意力权重。考虑到图中有六个角色，我们分别计算对象节点 (\mathbf{p}^o)、文本节点 (\mathbf{p}^t)、对象-对象边 (\mathbf{p}^{oo})、对象-文本边 (\mathbf{p}^{ot})、文本-文本边 (\mathbf{p}^{tt}) 和文本-对象边 (\mathbf{p}^{to}) 的注意力权重。该机制可以表述为：

$$\mathbf{p}^m = \text{Att}_m(\{\mathbf{x}^{obj}\}, \{\mathbf{x}^{text}\}, \{e_{ij}\}, \mathbf{s}^m), \quad (2)$$

其中 Att_m 是使用问题特征和图中的特定节点/边计算注意力权重的注意力机制，以及 $m = \{o, oo, ot, t, tt, to\}$ 。 \mathbf{x}^{obj} 和 \mathbf{x}^{text} 分别表示从孤立对象和文本区域中提取的特征，然后将其输入到图注意力模块中以生成问题条件特征。现在我们描述如何根据不同类型的注意力来计算 Att_m 。

1) 节点表示。一个对象节点由 Faster R-CNN 检测器的 2048 D 外观特征和 4 D 边界框特征（带有对象的相对边界框坐标 $[\frac{x_i^{tl}}{W}, \frac{y_i^{tl}}{H}, \frac{x_i^{br}}{W}, \frac{y_i^{br}}{H}]$ ）表示，其中 W 和 H 表示图像的宽度和高度。给定对象的外观特征 $\{\mathbf{x}_{fr,i}^o\}_{i=1}^N$ 和边界框特征 $\{\mathbf{x}_{bbox,i}^o\}_{i=1}^N$ ，对象节点的表示通过以下公式计算：

$$\hat{\mathbf{x}}_i^{obj} = \text{LN}(\mathbf{W}_{fr}^o \mathbf{x}_{fr,i}^o) + \text{LN}(\mathbf{W}_b^o \mathbf{x}_{bbox,i}^o), \quad (3)$$

其中 $\text{LN}(\cdot)$ 是层归一化； \mathbf{W}_{fr}^o 和 \mathbf{W}_b^o 是需要学习的线性变换参数。

对于文本节点，我们还采用了多种特征的组合（称为 Multi-Feats）来丰富 OCR 区域的表示，如 [?]：1) 300 D FastText 特征 $\{\mathbf{x}_{ft,i}^t\}_{i=N+1}^{N+M}$ 由预先训练的 FastText [?] 嵌入生成，2) 2048 D 外观特征 $\{\mathbf{x}_{fr,i}^t\}_{i=N+1}^{N+M}$ 由与对象节点相同的 Faster R-CNN 检测器生成，3) 604 D 字符金字塔直方图 (PHOC) [?] 特征 $\{\mathbf{x}_{p,i}^t\}_{i=N+1}^{N+M}$ 和 4) 4 D 边界框特征 $\{\mathbf{x}_{bbox,i}^t\}_{i=N+1}^{N+M}$ 。除了多项特征之外，我们还引入了一个 512 D CNN 特征 $\{\mathbf{x}_{rec,i}^t\}_{i=N+1}^{N+M}$ （称为 RecogCNN），它是从基于 Transformer

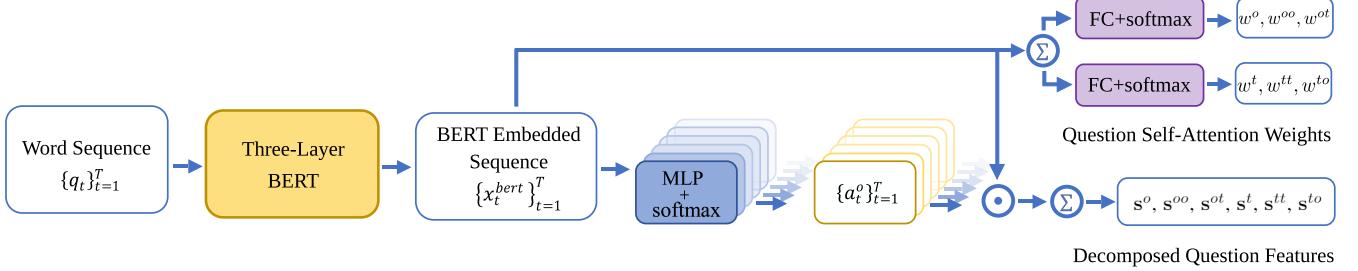


Fig. 3: 问题自注意力模块概述。输入一个问题的词序列，我们得到两种注意力权重：考虑整个图中的先验概率的问题自注意力权重和对应节点或边的细粒度分解问题特征。

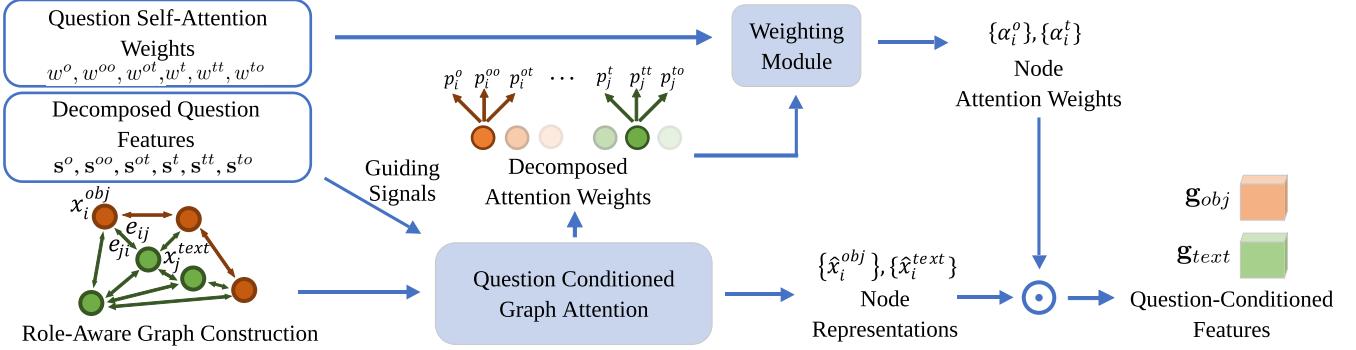


Fig. 4: 问题条件图注意力模块概述。该模块构建一个异构图，其混合节点以不同的颜色显示。引导信号有助于产生注意力权重，将其与节点表示融合，我们得到问题条件特征。

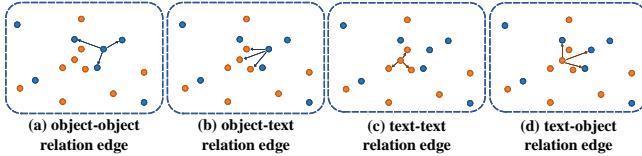


Fig. 5: 四种边结构示意图。蓝点表示对象节点，橙点表示文本节点。这里我们设置 $k = 3$ 以简化操作，一个节点只会与其 k 最近的邻居有边。

的文本识别网络 [?] 中提取出来的。文本节点的表示通过以下公式计算：

$$\begin{aligned} \mathbf{x}_i^m &= \mathbf{W}_{ft}^t \mathbf{x}_{ft,i}^t + \mathbf{W}_{fr}^t \mathbf{x}_{fr,i}^t + \mathbf{W}_p^t \mathbf{x}_{p,i}^t + \mathbf{W}_{rec}^t \mathbf{x}_{rec,i}^t, \\ \hat{\mathbf{x}}_i^{text} &= \text{LN}(\mathbf{x}_i^m) + \text{LN}(\mathbf{W}_b^t \mathbf{x}_{bbox,i}^t), \end{aligned} \quad (4)$$

其中 \mathbf{W}_{ft}^t 、 \mathbf{W}_{fr}^t 、 \mathbf{W}_p^t 、 \mathbf{W}_{rec}^t 和 \mathbf{W}_b^t 是需要学习的线性变换参数。

2) 节点注意力权重。由于对象节点注意权重和文本节点注意权重的计算过程相似，我们选择 p^o 进行说明：给定一个对象节点的表示 $\{\hat{\mathbf{x}}_i^{obj}\}_{i=1}^N$ ，在 s^o 的指导下计算对象节点的注意权重：

$$\begin{aligned} p_i^{o'} &= \mathbf{w}_o^\top [\text{ReLU}(\mathbf{W}_s^o s^o) \circ \text{ReLU}(\mathbf{W}_x^o \hat{\mathbf{x}}_i^{obj})], \\ p_i^o &= \text{softmax}(p_i^{o'}), \quad i = 1, \dots, N, \end{aligned} \quad (5)$$

其中 \mathbf{W}_s^o 、 \mathbf{W}_x^o 和 \mathbf{w}_o 是需要学习的线性变换参数。

对于文本节点 (p^t) 的注意权重，我们执行相同的计算，但使用独立的参数。

3) 边缘注意权重。边注意力权重需要考虑两个节点之间的关系。由于不同边类型 p^{oo} 、 p^{ot} 、 p^{tt} 和 p^{to} 的注意力权重计算过程类似，我们仅展示如何计算 p^{oo} 。

主要有两个步骤。首先，对于每个节点 o_i ，我们计算与 o_i 相连的所有 oo 边上的注意权重 $\mathbf{q}_i^{oo} = \{q_{ij}^{oo}\}_{j \in \mathcal{N}_i^{oo}}$ ：

$$\begin{aligned} \hat{\mathbf{x}}_{ij}^{oo} &= \text{MLP}([e_{ij}; \hat{\mathbf{x}}_i^{obj}]), \\ q_{ij}^{oo'} &= \mathbf{w}_{oo}^\top [\text{ReLU}(\mathbf{W}_s^{oo} s^{oo}) \circ \text{ReLU}(\mathbf{W}_x^{oo} \hat{\mathbf{x}}_{ij}^{oo})], \\ q_{ij}^{oo} &= \text{softmax}(q_{ij}^{oo'}), \quad j \in \mathcal{N}_i^{oo}, \end{aligned} \quad (6)$$

其中 \mathbf{W}_s^{oo} 和 \mathbf{W}_x^{oo} 分别将 oo 边相关问题表示 s^{oo} 和嵌入边特征 $\hat{\mathbf{x}}_{ij}^{oo}$ 映射到相同维度的向量中。注意权重 \mathbf{q}_i^{oo} 通过 softmax 层在 o_i 的邻域 \mathcal{N}_i^{oo} 上进行归一化。

在第二步中，我们计算所有对象节点的 oo 边缘注意权重 $\mathbf{p}^{oo} = \{p_i^{oo}\}_{i=1}^N$ ：

$$\begin{aligned} \tilde{\mathbf{x}}_i^{oo} &= \sum_{j \in \mathcal{N}_i^{oo}} q_{ij}^{oo} \hat{\mathbf{x}}_{ij}^{oo}, \\ p_i^{oo'} &= \mathbf{w}_{oo'}^\top [\text{ReLU}(\mathbf{W}_s^{oo'} s^{oo}) \circ \text{ReLU}(\mathbf{W}_x^{oo'} \tilde{\mathbf{x}}_i^{oo})], \\ p_i^{oo} &= \text{softmax}(p_i^{oo'}), \quad i = 1, \dots, N, \end{aligned} \quad (7)$$

其中 $\tilde{\mathbf{x}}_i^{oo}$ 被视为问题条件下的 oo 边缘特征 w.r.t. 对象节点 o_i 。我们使用上述相同方程式计算 p^{ot} 、 p^{tt} 和 p^{to} ，但使用单独的初始边缘特征、问题表示和转换参数。

Weighting Module 上述图注意模块通过相应的自注意问题部分作为指导，为每个对象和文本节点输出三个注意权重。对于每个对象节点 o_i ，我们有 p_i^o 、 p_i^{oo} 和 p_i^{ot} 。类似地，对于每个文本节点 t_i ，我们有 p_i^t 、 p_i^{tt} 和 p_i^{to} 。现在我们将它们与问题自注意力权重结合在一起。对于每个对象节点，最终权重得分计算为三部分的加权和：

$$a_i^o = w^o p_i^o + w^{oo} p_i^{oo} + w^{ot} p_i^{ot}, \quad i = 1, \dots, N, \quad (8)$$

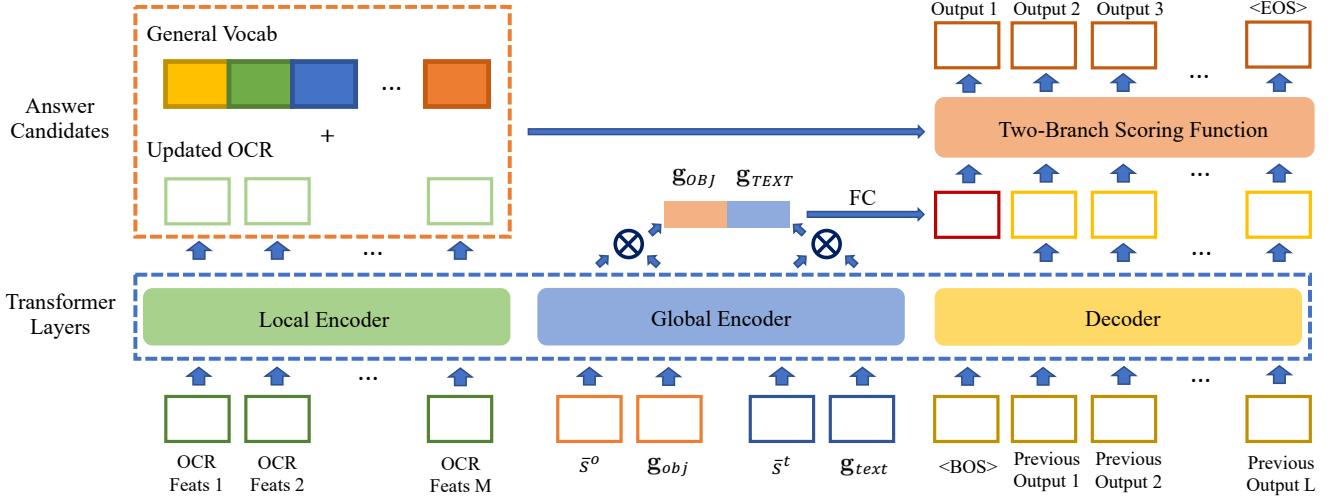


Fig. 6: 全局-局部注意力应答模块概述。相同的 Transformer 层被分成三个功能不同的部分。局部编码器更新局部 OCR 嵌入。全局编码器的结果用于预测第一个时间步的答案。通用词汇表和更新的 OCR 构成了答案候选，从我们使用双分支评分函数在每个时间步中选择答案。

其中 $w^{o,oo,ot}$ 在 3.1 部分获得。类似地，每个文本节点的最终权重为：

$$\alpha_i^t = w^t p_i^t + w^{tt} p_i^{tt} + w^{to} p_i^{to}, i = N+1, \dots, N+M. \quad (9)$$

注意 $\sum_{i=1}^N \alpha_i^o = 1$ ，因为我们有 $w^o + w^{oo} + w^{ot} = 1$ 、 $\sum_{i=1}^N p_i^o = 1$ 、 $\sum_{i=1}^N p_i^{oo} = 1$ 和 $\sum_{i=1}^N p_i^{ot} = 1$ 。同样，我们也有 $\sum_{i=N+1}^{N+M} \alpha_i^t = 1$ 。权重 $\{\alpha_i^o\}_{i=1}^N$ 和 $\{\alpha_i^t\}_{i=N+1}^{N+M}$ 实际上衡量了对象/文本节点与问题之间的相关性，并用于生成问题条件下的对象和文本特征：

$$\mathbf{g}_{obj} = \sum_{i=1}^N \alpha_i^o \cdot \hat{\mathbf{x}}_i^{obj}, \quad \mathbf{g}_{text} = \sum_{i=N+1}^{N+M} \alpha_i^t \cdot \hat{\mathbf{x}}_i^{text}. \quad (10)$$

3.3 全局-局部注意力应答模块

受到 M4C [?] 中迭代答案预测机制的启发，我们将其引入到我们的 SMA 中，通过修改第一个解码步骤的输入作为我们的全局-局部注意回答模块（如图 6 所示）。M4C 的第一个解码步骤的输入是一个特殊的 token < begin >，而我们将其替换为问题、对象和文本的总结性全局特征以及局部 OCR 嵌入 (\mathbf{g}_{obj} 和 \mathbf{g}_{text})，其中包含了我们的问题条件图的综合信息。

全局图特征 \mathbf{g}_{obj} 和 \mathbf{g}_{text} 是通过融合我们的问题条件图的全局特征生成的。具体来说，首先将与对象相关的问题特征和与文本相关的问题特征连接在一起：

$$\bar{s}^o = [\mathbf{s}^o; \mathbf{s}^{oo}; \mathbf{s}^{ot}], \quad \bar{s}^t = [\mathbf{s}^t; \mathbf{s}^{tt}; \mathbf{s}^{to}]. \quad (11)$$

\bar{s}^o 、 \bar{s}^t 、 \mathbf{g}_{obj} 、 \mathbf{g}_{text} 与局部 OCR 节点嵌入一起转发到 Transformer 层，并更新为 \tilde{s}^o 、 \tilde{s}^t 、 $\tilde{\mathbf{g}}_{obj}$ 、 $\tilde{\mathbf{g}}_{text}$ 。这里，OCR 嵌入已经由图注意模块 *i.e.* 更新，它们是 $\{\alpha_i^o \cdot \hat{\mathbf{x}}_i^{text}\}_{i=N+1}^{N+M}$ ，与公式 10 略有不同。在 Transformer 更新过程中，这些全局特征和局部 OCR 特征可以自由地相互关注。

然后，我们将更新后的特征 $\tilde{\mathbf{g}}_{obj}$ 和 $\tilde{\mathbf{g}}_{text}$ 与它们各自的问题表示融合，如下所示：

$$\mathbf{g}_{OBJ} = \tilde{\mathbf{g}}_{obj} \circ \tilde{s}^o, \quad \mathbf{g}_{TEXT} = \tilde{\mathbf{g}}_{text} \circ \tilde{s}^t. \quad (12)$$

用于预测第一个时间步 \mathbf{p}_{ans}^1 中的答案概率的方程可以写成：

$$\mathbf{p}_{ans}^1 = f_{pred}(\mathbf{W}_g[\mathbf{g}_{OBJ}; \mathbf{g}_{TEXT}]), \quad (13)$$

#	Method	OCR system	Output module	Accu. on val
1	Baseline	Rosetta-ml	classifier	29.16
2	Baseline+oo	Rosetta-ml	classifier	29.34
3	Baseline+ot	Rosetta-ml	classifier	29.58
4	Baseline+tt	Rosetta-ml	classifier	29.73
5	Baseline+to	Rosetta-ml	classifier	30.14
6	Baseline+ all	Rosetta-ml	classifier	30.26

TABLE 1: TextVQA 数据集上问题条件图注意模块关键组件的消融研究。如第 3.2 节所述，我们的图中有四种边（关系），分别是 oo、ot、tt 和 to 边。剥离问题条件图注意模块中的这四种关系可得到一条基线。我们将四种边注意分别添加到基线中并评估相应的准确性。

其中 \mathbf{W}_g 是线性变换， f_{pred} 是双分支评分函数，它解决了 TextVQA 任务中的答案可以在不同问题中变化的动态文本的困境。其余时间步的输入和答案空间设置与 M4C 中的相同。

Training Loss 考虑到答案可能来自两个来源，我们使用多标签二元交叉熵 (bce) 损失：

$$pred = \frac{1}{1 + \exp(-y_{pred})}, \quad (14)$$

$$\mathcal{L}_{bce} = -y_{gt} \log(pred) - (1 - y_{gt}) \log(1 - pred),$$

其中 y_{pred} 是预测， y_{gt} 是真实目标。

4 实验

我们在两个具有挑战性的 TextVQA 基准上评估了我们的模型，包括 TextVQA [?] 和 ST-VQA [?] 的所有三个任务，并在 TextVQA 和 ST-VQA 的前两个任务上取得了 SoTA 性能。在我们的实验中，我们发现 OCR 的准确性可能会限制模型的推理能力，因此我们手动标记了 TextVQA 数据集中出现的所有文本，*i.e.*，我们提供了 OCR 部分的 ground-truth，以便研究社区可以充分研究 Text-VQA 模型的推理能力，而不必考虑文本识别的影响。

#	Method	Question enc. pretraining	OCR system	OCR token representation	Output module	Accu. on val	Accu. on test
1	LoRRA [?]	GloVe	Rosetta-ml	FastText	classifier	26.56	27.63
2	LoRRA	GloVe	Rosetta-en	FastText	classifier	29.35	-
3	LoRRA	GloVe	SBD-Trans	FastText	classifier	29.73	-
4	DCD ZJU (ensemble) [?]	-	-	-	-	31.48	31.44
5	MSFT VTI [?]	-	-	-	-	32.92	32.46
6	M4C [?]	BERT	Rosetta-ml	Multi-feats	decoder	37.06	-
7	M4C	BERT	Rosetta-en	Multi-feats	decoder	39.40	39.01
8	M4C	BERT	SBD-Trans	Multi-feats	decoder	40.24	-
9	MM-GNN [?]	-	Rosetta-ml	-	classifier	31.44	31.10
10	LaAP-Net [?]	BERT	Rosetta-en	Multi-feats	decoder	40.68	40.54
11	SA-M4C [?]	BERT	Google-OCR	Multi-feats	decoder	45.40	44.60
12	PixelM4C [?]	BERT	MTS v3	Multi-feats	decoder	42.12	-
13	TAP * w/o extra data [?]	BERT	Microsoft-OCR	Multi-feats	decoder	49.91	49.71
14	TAP * [?]	BERT	Microsoft-OCR	Multi-feats	decoder	54.71	53.97
15	SMA w/o dec. (Model 6 in Tab.1)	GloVe	Rosetta-ml	FastText	classifier	30.26	-
16	SMA w/o dec.	GloVe	Rosetta-en	FastText	classifier	32.28	-
17	SMA w/o dec.	GloVe	Rosetta-en	Multi-feats	classifier	35.03	-
18	SMA with M4C dec.	GloVe	Rosetta-en	Multi-feats	decoder	38.91	-
19	SMA	GloVe	Rosetta-en	Multi-feats	decoder	39.36	-
20	SMA	BERT	Rosetta-en	Multi-feats	decoder	40.24	40.21
21	SMA	BERT	Rosetta-ml	Multi-feats + RecogCNN	decoder	37.74	-
22	SMA	BERT	Rosetta-en	Multi-feats + RecogCNN	decoder	40.39	40.86
23	SMA	BERT	SBD-Trans	Multi-feats + RecogCNN	decoder	43.74	44.29
24	SMA with ST-VQA Pre-training	BERT	SBD-Trans	Multi-feats + RecogCNN	decoder	44.58	45.51

TABLE 2: 更多消融模型和与之前工作的比较。与我们的 SMA 相比, TAP 模型需要额外的预训练阶段和一组预训练任务。第 13 行和第 14 行中的 TAP 分别使用 TextVQA 数据集的训练集和额外的大规模训练数据进行预训练。

4.1 实现细节

继 M4C [?] 之后, 从紧跟 Faster R-CNN [?] 模型的 RoI-Pooling 层的 fc6 层中提取基于对象和 OCR 区域的外观特征。该模型在 Visual Genome [?] 上进行预训练, 然后在 TextVQA [?] 上对 fc7 层进行微调。对象区域的最大数量 $N = 36$ 。对于文本节点, 我们使用四种独立的 OCR 方法来识别字符串。

- 1) Rosetta-ml OCR。Rosetta 系统的多语言版本 [?].
- 2) Rosetta-en OCR。Rosetta 系统的英语版本。
- 3) SBD-Trans OCR。SoTA 无序列框离散化 (SBD) 模型 [?] 用于场景文本检测, 而基于稳健变压器的网络 [?] 用于单词识别。它们的训练过程将在附录 F 中详细说明。
- 4) 真实值 OCR。其收集过程将在第 4.3 节中详细介绍。

我们从一张图像中识别最多 $M = 50$ 个 OCR 标记, 并基于它们生成丰富的 OCR 表示。如果上述任何一个低于最大值, 我们将对其余部分应用零填充。我们将问题的最大长度设置为 $T = 20$, 并通过预训练的 BERT [?] 的前三层将它们编码为 768 D 特征序列, 其参数在训练期间进一步微调。我们的应答模块使用 4 层的 Transformer 和 12 个注意力头。其他超参数与 BERT-BASE [?] 相同。解码步骤的最大数量设置为 $L = 12$ 。

我们在 PyTorch 中实现所有模型, 并在 4 NVIDIA GeForce 1080Ti GPU 上进行实验, 批次大小为 96。除用于问题编码的三层 BERT 和用于区域特征编码的 fc7 层 (它们的学习率为 $1e - 5$) 外, 所有层的学习率均设置为 $1e - 4$ 。我们在 14000 和 19000 迭代中将学习率乘以 0.1, 优化器为 Adam。在每次 1000 迭代中, 我们都会在验证集上计算一个 VQA 准确度指标 [?], 并根据该指标选择性能最佳的模型。为了优雅地捕捉文本识别中的错误, ST-VQA 数据集 [?] 采用平均归一化编辑相似度 (ANLS) 作为其官方评估指标。

4.2 TextVQA 的结果与分析

TextVQA 数据集 [?] 从 OpenImages 数据集 [?] 中抽样 28,408 图像。问题分为训练、验证和测试部分, 大小分别为 34,602、5,000 和 5,734, 每个问题-图像对都有 10 个人工提供的真实答案。

4.2.1 消融研究

Ablations on Relationship Attentions. 我们进行了一项消融研究, 以调查所提出的问题条件图注意模块 *i.e.* 的关键组件, 即四种类型 (**oo**、**ot**、**tt**、**至**) 的关系注意。为了关注推理能力, 我们在没有丰富的 OCR 表示和迭代回答模块的情况下对其进行评估。测试的架构变化及其结果如表 1 所示。在 基线 模型中, 图 2 中的角色感知图具有隔离的 OCR 和对象节点, 省去了节点之间的交互和匹配问题特征。尽管相似, **至** 和 **ot** 本质上是不同的。文本-对象边以文本为中心, 以相邻的对象节点为上下文来更新文本特征; 而对象-文本关系以对象为中心, 以相邻的文本节点为上下文来更新对象特征。两类边在特征更新中扮演着不同的角色。实验结果表明, 四种建模关系都提高了准确率, 尤其以 **至** 关系注意力机制带来的提升最大。这与注释者倾向于通过描述打印文本的对象来指代特定文本的观察结果一致。总体而言, 以文本为起源的关系 (**至** 和 **tt**) 比以对象为起源的关系 (**oo** 和 **ot**) 更重要, 这验证了文本在该文本 VQA 任务中的关键作用。Baseline+all 模型结合了所有关系并获得了最佳性能。

Ablations on Answering Modules. 从表 2 中的模型 17 和 19 可以看出, 我们提出的生成式回答模块大大超越了基于判别分类器的回答模块 (验证准确率方面为 4.33%)。我们还通过将我们的回答模型的修改与表 2 中的 18 和 19 中 M4C [?] 提出的原始答案生成结构进行比较来验证我们的回答模型的修改, 这表明我们的修改可以带来近 0.5% 的准确率提升。

Ablations on Features for Question and OCR. 从模型 16 到 17, 通过将 FastText 特征替换为 OCR 标记的多功能特征, 获得了近 3% 的改进。多功能特征是一个特征包, 包括 M4C [?] 中提出的 FastText、Faster R-CNN、PHOC 和 BBox 特征。Glove 和 BERT 特征用于编码问题进行评估, 后

Fig. 7: 边缘注意力和 SMA 的分解问题注意力可视化。给出了三个需要关系推理进行问答的代表性示例，它们需要不同类型的边关系。例如，**至** 表示前一个节点为 **t ext** 而后一个节点为 **o bject** 的关系。对于每个示例，我们突出显示具有最高注意力权重的节点或边，其中节点用框表示，边用从前一个节点指向后一个节点的箭头显示。对于框/节点，黄色表示对象，蓝色表示文本。实线表示具有最高注意力权重的，而虚线表示正常。对于分解的问题注意力，颜色较深的突出显示文本区域具有更高的注意力权重。所有这些都是由 SMA 使用 Ground-Truth OCR 预测的。可以看出，问题注意力模块成功地找出了所需的关系。

#	Methods	Val Accuracy (%)	
		w/o GT OCR	w GT OCR
1	LoRRA [?]	29.35	35.07
2	M4C [?]	39.40	47.91
3	SMA (Ours)	40.39	50.07
4	OCR UB	44.98	68.81
5	Human	85.01	85.01

TABLE 3: 在三种模型上使用 GT OCR 进行评估。在用 GT OCR 替换 Rosetta-en OCR 后，它们都得到了大幅提升。还提供了 OCR UB 指标以供参考，该指标衡量某个 OCR 系统的上限 vqa 准确率。

者在验证准确率上优于 0.88%（参见表 2 中的模型 19 和 20）。通过比较表 2 中的模型 20 和 22，我们可以看到通过为 OCR 添加 RecogCNN 特征，0.15% 在验证集上得到了进一步改进，0.65% 在测试分割上得到了进一步改进。这验证了 RecogCNN 特征与 Multi-Feats 是互补的。请注意，RecogCNN 是在文本识别任务上训练的，而 Faster R-CNN 是针对一般物体检测进行训练的。FastText 和 PHOC 是从识别的 OCR 字符序列中提取的，而 RecogCNN 是从文本视觉块中提取的。

Visualization. 对于每种关系，我们将注意力权重最高的关系及其对应的分解问题注意力可视化，以探索它们在答案预测中的贡献，并为解释我们的模型提供更好的见解（见图 7）。在第一个例子中，有几辆自行车，问题询问其中哪辆是正确的。找到所要求的自行车需要 **oo** 关系推理。还需要另一个关系 **至**，因为自行车上的号码正是我们要弄清楚的。在第二个例子中，我们需要 **ot** 关系来定位号码为 20 的球员。然后使用 **至** 关系来推理这个球员的姓氏。类似地，在最后一个例子中，提取了两个不同的 **oo** 关系来精确定位右边蓝色头发的球员的位置。然后使用 **ot** 关系来获取球员的号码。所有示例都验证了我们模型的关系推理能力。

4.2.2 与以前的工作比较。

我们将我们的方法与之前的方法进行了比较，并取得了卓越的成果。使用相同的问题、对象和 OCR 特征，我们的单一模型在测试集上比 M4C（模型 20 VS. 7）好 1.20%。通过应用先进的 OCR 系统 SBD-Trans（更多详细信息请参见第 4.3 节），我们的最终模型 24 在测试集上实现了 44.29%。通过 ST-VQA 预训练，我们进一步在测试集上实现了 45.51%，这是目前最先进的技术（基于预训练的 TAP [?] 除外），大大超越了之前的最佳模型 M4C、i.e. 6.5%，并赢得了 2020 TextVQA 挑战赛（更多详细信息请参阅附录 E）。

4.3 OCR 性能对 TextVQA 的影响

然而，在实验过程中，我们发现有时即使我们的模型可以关注与问题相关的正确 OCR 区域，我们仍然无法正确回答。这主要是因为 OCR 模型无法识别 OCR 标记。这说明 OCR 的性能、i.e. 和阅读能力极大地影响了模型的推理能

Fig. 8: LoRRA、M4C 和 SMA 与不同 OCR 系统在 TextVQA 数据集上的表现。Hmean 衡量 OCR 模型的能力。x 轴显示四个逐渐增加的 OCR (Rosetta-ml、Rosetta-en、SBD-Trans 和 Ground-Truth) 的 Hmean 值。y 轴表示推理模型的验证准确率。

Methods	Precision	Recall	Hmean
Rosetta-ml	0.4789	0.1959	0.2781
Rosetta-en	0.5106	0.1966	0.2839
SBD-Trans	0.5958	0.4683	0.5244

TABLE 4: 如果 OCR 结果的边界框与对应的真实边界框重叠超过总面积的 50%，并且给出的标记相同，则认为该 OCR 结果匹配。可以清楚地看到，SBD-Trans OCR 比 Rosetta-ml OCR 和 Rosetta-en OCR 更准确。

力。有关示例，请参阅附录 A。为了充分研究这个问题并为我们的 SMA 模型（以及之前的模型）设置上限，我们提供了 TextVQA 图像中所有 OCR 标记的人工注释。注释可在 <https://github.com/ChenyuGAO-CS/SMA> 获得

4.3.1 人工注释的真相 OCR。

我们提供了 TextVQA 训练和验证集的真实 OCR 注释，因为它为研究人员提供了一个公平的测试基础，使他们可以专注于文本视觉推理部分，而无需额外调整 OCR 模型。更多详细信息请参阅附录 B。

4.3.2 使用 Ground-Truth OCR 的结果。

我们使用真实 OCR 评估了 LoRRA、M4C 和 SMA 的性能。结果如表 3 所示。两者都有很大幅度的提高：LoRRA 从 29.35% 上升到 35.07%，M4C 从 39.40% 上升到 47.91%，而 SMA 在验证集上从 40.39% 上升到 50.07%，方法是用真实结果替换 Rosetta-en 结果。准确率的大幅提升（9.68% vs 5.72% 和 8.51%）表明我们的模型具有更好的推理和应答能力。OCR UB 是如果可以使用受 [?] 启发的 OCR 源中的单个或多个标记构建答案，则可以获得的上限准确率。使用 GT OCR，在验证中上限可以提升到 68.81%。然而，人类的表现与 SMA 之间仍然存在很大差距，这有很大的潜力等待我们去挖掘。

4.3.3 不同 OCR 的模型性能

在图 8 中，我们比较了不同模型与四种 OCR 系统 (Rosetta-ml、Rosetta-en、SBD-Trans 和 Ground-Truth) 的准确率。很明显，随着 OCR 系统的精度提高，LoRRA、M4C 和 SMA 的准确率都得到了提高。我们提出的 SMA 模型表现最佳。

作为参考，我们还分别提供了基于手动标记的地面实况的 Rosetta-ml、Rosetta-en 和 SBD-Trans 的 OCR 准确率，如表 4 所示。我们发现 Rosetta-ml 未能表现出令人满意的结果，Rosetta-en 表现略好，而 SBD-Trans 表现更好，尤其是在“召回率”和“Hmean”方面。

#	Method	Task 1 ANLS	Task 2 ANLS	Task 3 ANLS
1	SAN+STR [?]	0.135	0.135	0.135
2	VTA [?]	0.506	0.279	0.282
3	M4C [?]	—	—	0.462
4	MM-GNN [?]	—	0.203	0.207
5	LaAP-Net [?]	—	—	0.485
6	SA-M4C [?]	—	—	0.504
7	TAP * w/o extra data [?]	—	—	0.543
8	TAP * [?]	—	—	0.597
9	SMA (Ours)	0.508	0.483	0.486

TABLE 5: 在 ST-VQA 数据集上进行评估。与我们的 SMA 相比, TAP 模型需要额外的预训练阶段和一组预训练任务。第 7 行和第 8 行中的 TAP 分别使用 ST-VQA 数据集的训练集和额外的大规模训练数据进行预训练。



Qi Zhu received the B.E. degree in Computer Science from Northwestern Polytechnical University, China. Her research interest are computer vision and machine learning.



Peng Wang is a Professor at School of Computer Science, Nothwestern Polytechnical University, China. He was with School of Computer Science, the University of Adelaide for about four years. His research interests are computer vision, machine learning and artificial intelligence. He received a Bachelor in electrical engineering and automation, and a PhD in control science and engineering from Beihang University (China) in 2004 and 2011, respectively.



Hui Li received the PhD degree in computer science from the University of Adelaide (Australia). She is now a postdoctoral researcher at the Australian Centre for Robotic Vision (ACRV). Her research interests include deep learning, scene text recognition and visual question answering.



YuLiang Liu received the B.S. degree in electronic and information engineering from the South China University of Technology, China, in 2016, where he received the Ph.D. degree with Deep Learning and Vision Computing Lab (DLVC-Lab), under the supervision of Prof. Lianwen Jin. He is now a postdoc in University of Adelaide, under the supervision of Prof. Chunhua Shen. He is working on scene text understanding, handwritten character recognition, document analysis, and deep learning-based text detection and recognition.



Anton Van Den Hengel is a Professor at the University of Adelaide and the founding Director of The Australian Centre for Visual Technologies (ACVT). He received a PhD in Computer Vision in 2000, a Master Degree in Computer Science in 1994, a Bachelor of Laws in 1993, and a Bachelor of Mathematical Science in 1991, all from The University of Adelaide.



Qi Wu is a Senior Lecturer (Assistant Professor) in the University of Adelaide and he is an Associate Investigator in the Australia Centre for Robotic Vision (ACRV). He is the ARC Discovery Early Career Researcher Award (DECRA) Fellow between 2019-2021. He obtained his PhD degree in 2015 and MSc degree in 2011, in Computer Science from the University of Bath, United Kingdom. His educational background is primarily in computer science and mathematics. He works on the Vision and Language problems, including Image Captioning, Visual Question Answering, Visual Dialog etc. His work has been published in prestigious journals and conferences such as TPAMI, CVPR, ICCV, AAAI and ECCV.

4.4 在 ST-VQA 数据集上进行评估

ST-VQA 数据集 [?] 包含 23,038 图像和 31,791 问答对。有三个 VQA 任务, 即强语境化、弱语境化和开放词汇。对于强语境化任务, 作者为每个图像提供了一个 100 词词典; 在弱语境化任务中, 作者为所有图像提供了一个 30,000 词词典; 对于开放词典任务, 不提供候选答案。由于 ST-VQA 数据集没有正式的训练和验证分组, 我们遵循 M4C [?] 随机选择 17,028 图像作为训练集, 并使用剩余的 1,893 图像作为验证集。

对于第一个任务, 我们使用了我们模型的单步版本 (SMA w/o dec.), 而对于第二个和第三个任务, 我们使用了建议的完整 SMA 模型, 但时间步长不同 (任务 2 为 3, 任务 3 为 12)。与排行榜上的方法相比, 我们为前两个任务设置了新的 SoTA, 并为任务 3 提供了可比的结果 (见表格 5)。

5 结论

我们引入了结构化多模态注意力 (SMA), 这是一种基于图像中的文本回答问题的新颖模型架构, 它在 TextVQA 和 ST-VQA 数据集上创造了新的最佳性能。SMA 由三个关键模块组成: 问题自注意力模块引导图形注意力模块学习节点和边缘注意力, 最后的应答模块结合了上述图形注意力模块的注意力权重和问题引导特征, 以迭代方式得出合理的答案。除了 SMA 模型之外, 我们还对多个 OCR 系统进行了彻底的实验, 并分析了它们对整体性能的影响程度。还提供了一组人工注释的 TextVQA 真实 OCR 集, 以设置新的上限并帮助社区评估不同模型的真实文本视觉推理能力, 而不会受到 OCR 准确性不佳的影响。

References

Chenyu Gao received the Bachelor degree in software engineering from the Northwestern Polytechnical University, China, in 2019. She is currently pursuing the MSc degree from School of Software Engineering, Northwestern Polytechnical University. Her research interests include computer vision, machine learning and artificial intelligence.

Appendix

A OCR 性能对 TextVQA 的影响

此项分析是在 TextVQA 的验证集上进行的。在图 9 中，我们可视化了 SMA 使用 Rosetta-en OCR 和 Ground-Truth OCR 的预测结果。我们可以发现，当与问题相关的 OCR 区域被机器正确检测到时，基于 Rosetta-en OCR 和 Ground-Truth OCR 的 SMA 在大多数情况下都可以关注到最相关的 OCR 区域。然而，轻微的识别错误仍可能导致错误的答案。如果与问题相关的 OCR 区域根本无法被机器检测到，那么该问题就没有机会被正确回答。这说明 OCR 的性能、*i.e.*、阅读能力极大地影响了模型的推理能力。

B 人工标注的地面实况 OCR

我们提供真实 OCR 注释的原因是它将为该领域提供公平的测试基础，以便研究人员可以专注于文本视觉推理部分，而无需调整 OCR 模型。我们要求 Amazon Mechanical Turk (AMT) 工作人员对 TextVQA 数据集中出现的所有文本进行注释，以便完全剥离 OCR 的影响并在公平的测试基础上调查真实的推理能力。

为了减轻工人的劳动强度，我们首先用训练好的文本检测和识别模型生成一组 OCR 结果。然后 AMT 工人检查机器生成的边界框和文本是否正确，之后是一个四分支过程：第一，如果边界框完全错误，工人被要求直接删除它；第二种情况，如果文本检测正确但识别错误，工人给出正确的文本；第三，如果边界框的位置不够准确，则对其进行轻微修改；最后，当机器遗漏了某个文本时，工人绘制边界框并同时为文本区域提供标签。

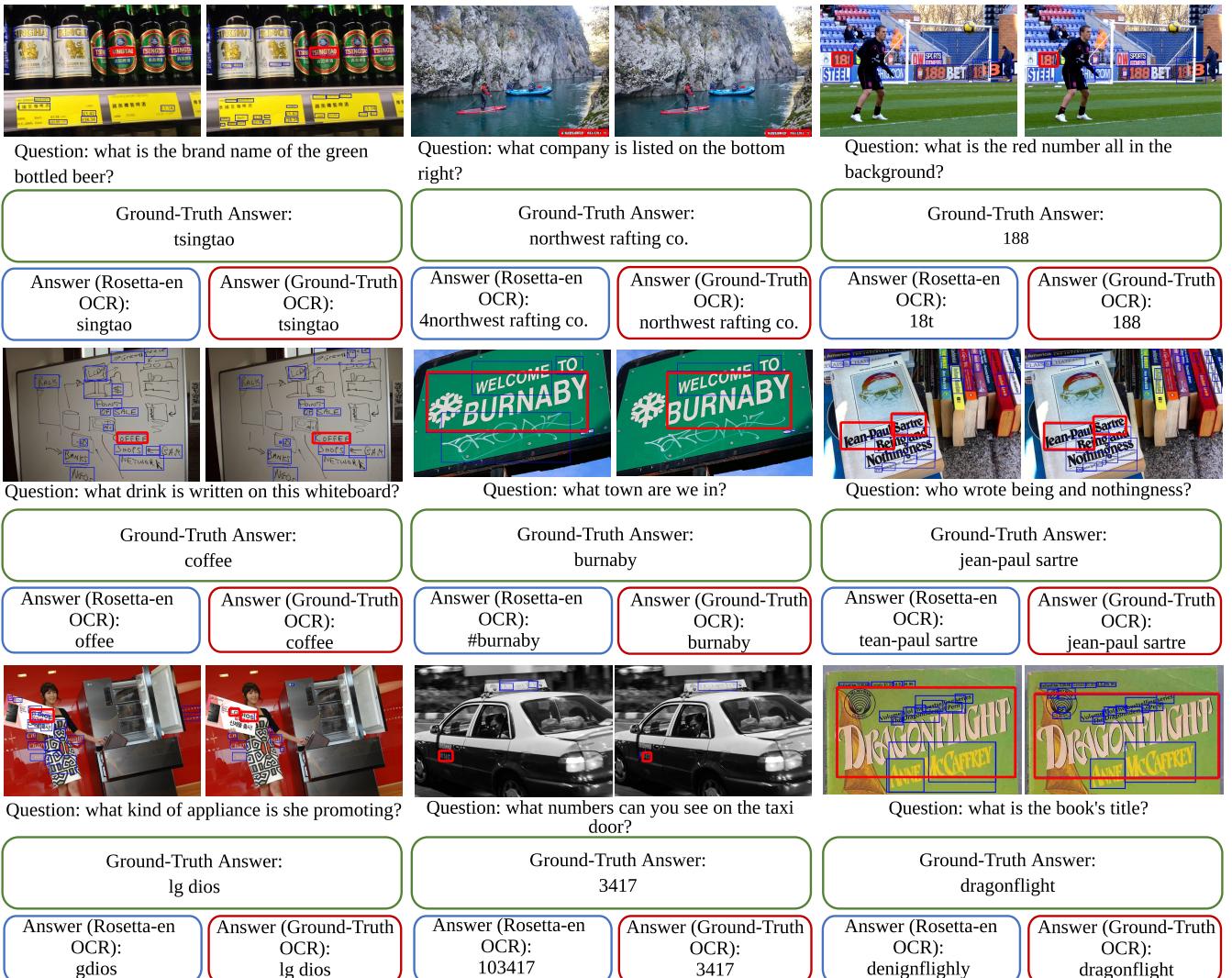


Fig. 9: SMA 对 TextVQA 例的预测。我们从 TextVQA 中选择了多个答案位于 OCR 标记中的示例。绿色虚线框包含 TextVQA 数据集中每个问题的 10 个 Ground-Truth 答案中出现频率最高的答案。在图像中，OCR 注释在蓝色框中，得分最高的 OCR 在红色框中。蓝色虚线框和红色虚线框中的答案分别由 SMA 使用 Rosetta-en OCR 和 Ground-Truth OCR 生成。由于 OCR 结果有缺陷，即使正确的推理过程也无法得出完全正确的答案。



Fig. 10: 基于 Ground-Truth OCR 的 TextVQA 验证集上 SMA 的典型失败案例。我们用蓝框突出显示得分最高的 OCR。(1) SMA 无法区分“小”和“大”。(2) 时钟指针角度和“xx:xx”格式答案之间的差距压倒了推理模型。(3) 需要常识知识。(4) 预期答案超出了考虑的 50 OCR 标记的范围。

C SMA 的失败案例。

我们在图 10 中展示并分析了我们模型的一些失败示例。在第一个例子中，我们的模型正确推断出颜色“白色”，但是，在尺寸属性“小”上失败。这个问题需要一些关于尺寸的常识性知识。第二个询问时间的例子非常棘手，因为时钟指针角度和所需的“xx: xx”格式的答案之间存在差距，这要求模型具有复杂的推理能力。第三个例子对于我们的模型来说很难回答，因为它需要的背景知识很难从当前有限的数据集中掌握。在第四张图中，由于所需答案超出了指定的 50 OCR 标记的范围，因此无法正确回答。

D TextVQA 挑战赛 2020

在 TextVQA Challenge 2020 中，不允许使用集成模型。亚军模型 SA-M4C [?] (测试集上为 44.80%) 从 M4C [?] 中的 vanilla Transformer 升级为具有空间感知的自注意层。此外，他们的方法还使用了一些技巧，例如更好的检测器主干、波束搜索解码、两个额外的自注意层、更好的 OCR 系统（谷歌 OCR 系统）以及与 ST-VQA 数据集 [?] 的联合训练。相比之下，我们仅使用单个 SMA 模型，以 SBD-Trans OCR 系统和 ST-VQA 数据集作为附加训练数据，从而实现了 45.51% 最终测试准确率——TextVQA 数据集上的新 SoTA。

E SBD-Trans 训练数据。

SBD 模型在 60k 数据集上进行预训练，该数据集包含来自 LSVT [?] 训练集的 30,000 图像、来自 MLT 2019 [?] 训练集的 10,000 图像、来自 ArT [?] 的 5,603 图像（包含 SCUT-CTW1500 [?] 和 Total-text [?, ?] 的所有图像）以及从一堆数据集 (RCTW-17 [?]、ICDAR 2013 [?]、ICDAR 2015 [?]、MSRA-TD500 [?]、COCOText [?] 和 USTB-SV1K [?]) 中选择的 14,859 图像）。该模型最终在 MLT 2019 [?] 训练集上进行了微调。基于鲁棒 Transformer 的网络在以下数据集上进行训练：IIIT 5K-Words [?]、Street View Text [?]、ICDAR 2013 [?]、ICDAR 2015 [?]、Street View Text Perspective [?]、CUTE80 [?] 和 ArT [?]。