Visual Question Answering Survey

---严薪 21821130

摘要：视觉问答是最近几年非常流行的人工智能研究方向，同时也是一个非常具有挑战性的问题，需要结合计算机视觉和自然语言处理的概念，大多数现有的方法都采用双流策略，即计算图像和问题的特征，然后通过各种技术对两种特征进行融合，本文调研了发表在 CVPR2019,NIPS2018 上的多篇视觉问答方面最新论文，涉及新的数据集，方法，以及对视觉问答可解释性方面工作。

1. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge

OK-VQA 数据集是由 CMU 于 CVPR2019 中提出的旨在强调外部知识的 VQA 数据集，到目前为止，大多数的 VQA 的 benchmarks 都集中在一些问题上，比如简单的技术，视觉属性和对象检测，这些问题不需要图像内容以外的推理或知识，作者提出的这个新的数据集中，图像内容不足以回答问题，必须依赖于外部知识，数据集包含超过 14000 个需要外部知识来回答的问题对，在新的数据集上，目前最先进的 VQA 模型的性能都大幅度的下降。为了能够理解问题和图像，模型需要（1）学习回答问题所需的知识；（2）确定从外部知识源检索必要知识所需的查询方式；（3）将原始表示的知识合并起来回答问题。人类对世界的了解有很多不同的类型，比如地理知识：埃菲尔铁塔在巴黎，科学知识：人类有 23 条染色体，历史知识：乔治华盛顿是美国第一任总统。OK 数据集包含 11 个知识分类，包括品牌，科学技术，地理等分布如图 1 所示。

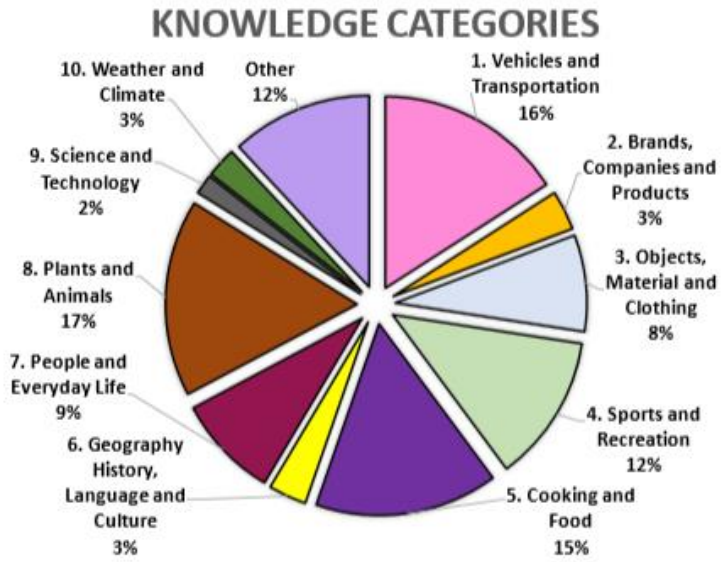


图 1

作者在这篇文章中的贡献主要有以下几点（1）引入了 OK-VQA 数据集，其中所有的问题都需要外部知识来回答（2）采用一些最先进的 VQA 模型对新数据集进行基准测试，实验结果表明，这些模型的性能在新的数据集上大幅度的下降（3）提出了一组利用非结构化知识的 baseline。

作者提出了 ArticleNet architecture 从维基百科上为每一个问题检索一些文章，通过训练网络来从检索的文章中确定答案。ArticleNet architecture 如图 2 所示：

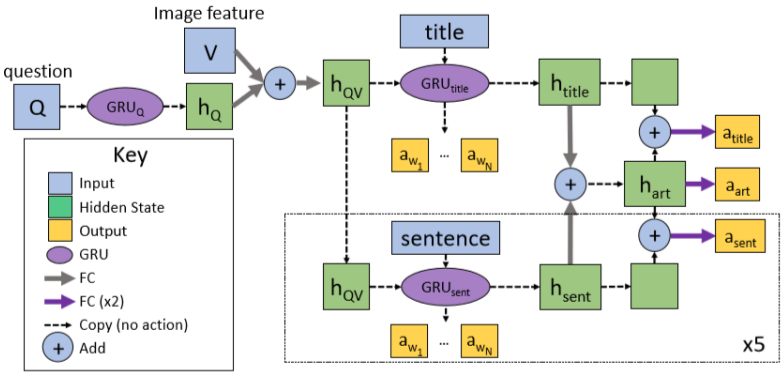


图 2

2. **Dynamic Fusion with Intra-and Inter-modlity Attention Flow for Visual Question Answering**

在视觉问答任务中如何有效的融合图像特征和语义特征对模型的性能有着非常大的影响，香港中文大学在 CVPR2019 的论文中提出了一种将模态内和模态间信息流动态融合的新方法。它可以有力地捕捉语言和视觉域之间的高级交互，从而显著提高视觉问答的性能。目前大多数现有的 VQA 方法都侧重于学习视觉和语言特征之间的模态关系。双线性特征融合方法侧重于通过特征外部产品捕捉语言与视觉形态之间的高阶关系，或基于双线性注意的方法学习单词-区域对之间的情态关系，以确定回答问题的关键对。另一方面，现有的计算机视觉和自然语言处理算法侧重于学习情态间的关系。在解决 vqa 问题的统一框架中，从未联合研究过模式间和模式内关系。我们认为，对于 vqa 问题，每个模态中的模态内部关系与模态间关系是互补的，而现有的 vqa 方法大多忽略了这种互补关系。例如，对于图像形态，每个图像区域不仅应该从问题中的关联词/短语获得信息，还应该从相关图像区域获得信息，以推断问题的答案。对于问题的形式，可以通过推断其他词来更好地理解问题。这些案例激励作者提出一个统一的框架，用于建模模式间和模式内信息流。为了克服这些局限性，作者提出了一种新的动态融合方法，该方法利用内、内模式注意力流（DFAF）框架实现高效的多模式特征融合，以准确回答视觉问题。DFAF 框架整合了跨模式自我关注和跨模式共同关注机制，以实现图像和语言模式内部和之间的有效信息流动。考虑到深层神经网络编码的视觉和问题特征，DFAF 框架首先生成多式联运注意流（interfaf），以在图像和语言之间传递信息。模型结构如图 3 所示。

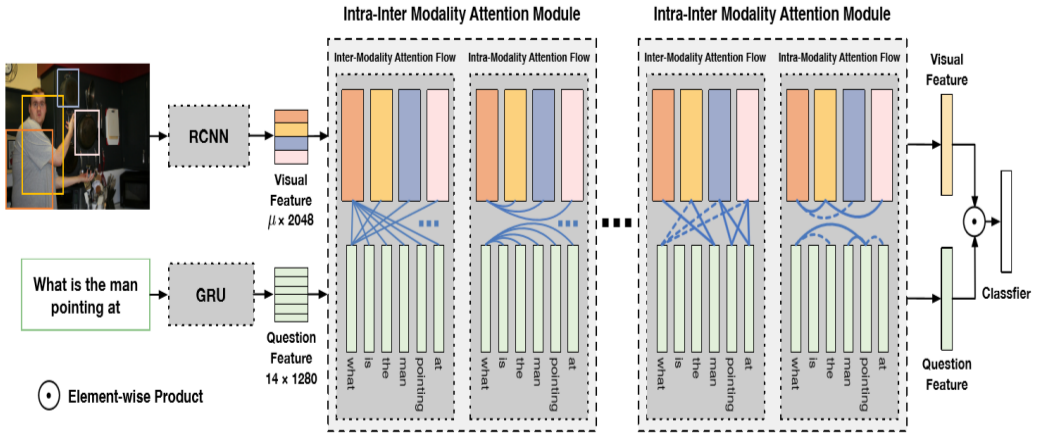


图 3

3. **Learning Conditioned Graph Structures for Interpretable Visual Question Answering**

作者在 NIPS2018 的文章中提出了一种基于图的视觉问答方法，基于图学习模块，该模块能够学习基于输入图像和问题的特定图形表示，以及最新的图形卷积概念，旨在学习捕获问题特定交互的图像表示。该模型在 VQA2.0 数据集上取得了 66.18%的准确率且具有较好的解释性。最近的计算机视觉工作一直在探索更高层次的图像表示，特别是使用对象检测器和基于图形的结构来更好地理解语义和空间图像。将图像表示为图形可以显式地建模交互，从而通过先进的图形处理技术在图形项（如图像中的对象）之间无缝地传递信息。这种基于图形的技术已经成为最近 VQA 工作的焦点，用于抽象图像理解或对象计数，达到最先进的性能。尽管如此，所提出的技术的一个重要缺点是，输入图结构经过严格设计，图像是特定的，而不是问题的特定，并且不容易从抽象场景转换为真实图像。此外，很少有方法能够解释模型的行为，这是深度学习模型中经常缺乏的一个基本方面。本文提出了一种新颖的、可解释的、基于图形的视觉问答方法，通过引入一个场景结构形式的先验图来解决这个问题，它被定义为一个根据问题的上下文从观察中学习到的图表。边界框对象检测被定义为图形节点，而基于问题的图形边缘则通过基于注意的模块学习。这不仅可以识别与问题相关联的图像中最相关的对象，还可以识别最重要的交互作用（例如相对位置、相似性），而不需要任何手工制作的图表结构的脚本。学习图形结构允许学习相关邻居使用图形卷积影响的问题特定对象表示。学习图形结构不仅为 VQA 任务提供了强大的预测能力，而且通过检查最重要的图形节点和边来解释模型的行为。模型结构如图 4 所示

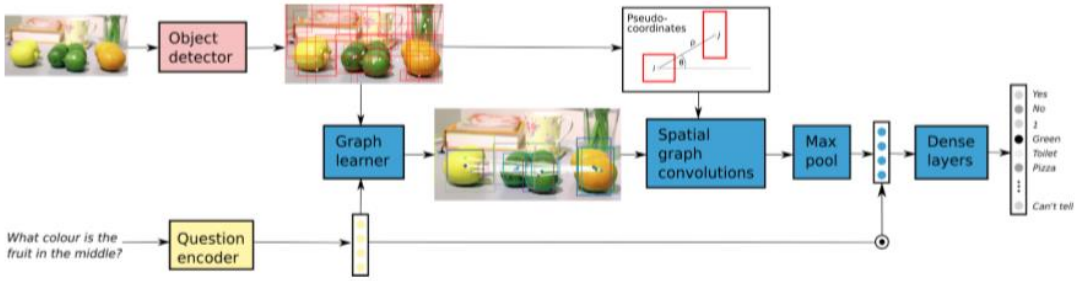


图 4

4. **Generating Question Relevant Captions to Aid Visual Question Answering**

作者在 ACL2019 的文章中提出了一种改进 VQA 性能的新方法，通过生成描述来帮助提高 VQA 性能，该模型使用现有的描述数据集进行训练，通过使用基于在线梯度的方法自动确定与问题相关的描述。很少有 VQA 研究从图像中挖掘文本特征，从而简洁地编码回答问题所需的信息。这些信息可以丰富视觉特征，因为这些内容没有结构约束，并且可以包含多个对象的属性和关系。作者探索了一种生成问题相关描述的新方法，其中包含与特定 VQA 问题直接相关的信息。为了整合描述信息，我们提出了一种新的题注嵌入模块，该模块给出了一个可视化问题的问题和图像特征，识别出题注中的简单关键词，并生成了一个针对答案预测的题注嵌入。此外，标题嵌入还用于调整每个对象的可视自上而下的注意权重。此外，生成与问题相关的标题可以确保图像和问题信息都被编码到它们的联合表示中，从而降低从问题偏差中学习的风险，并在仅从问题中获得高精度时忽略图像内容。模型的结果如图 5 所示

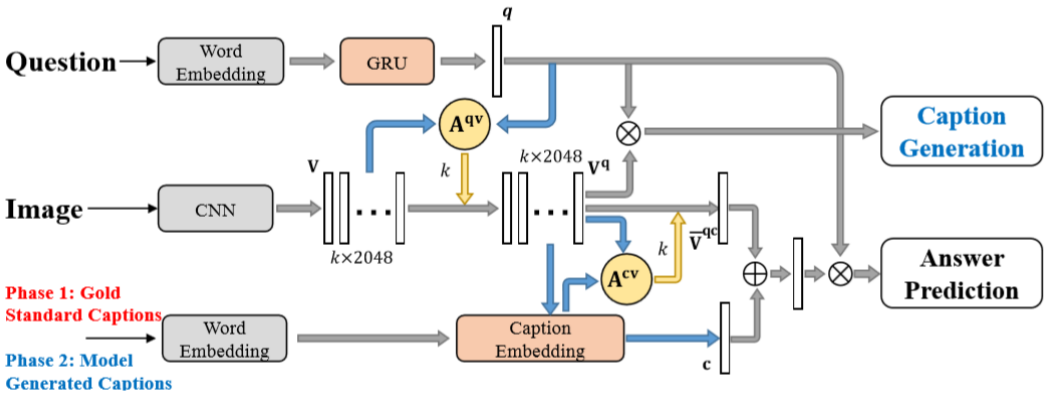
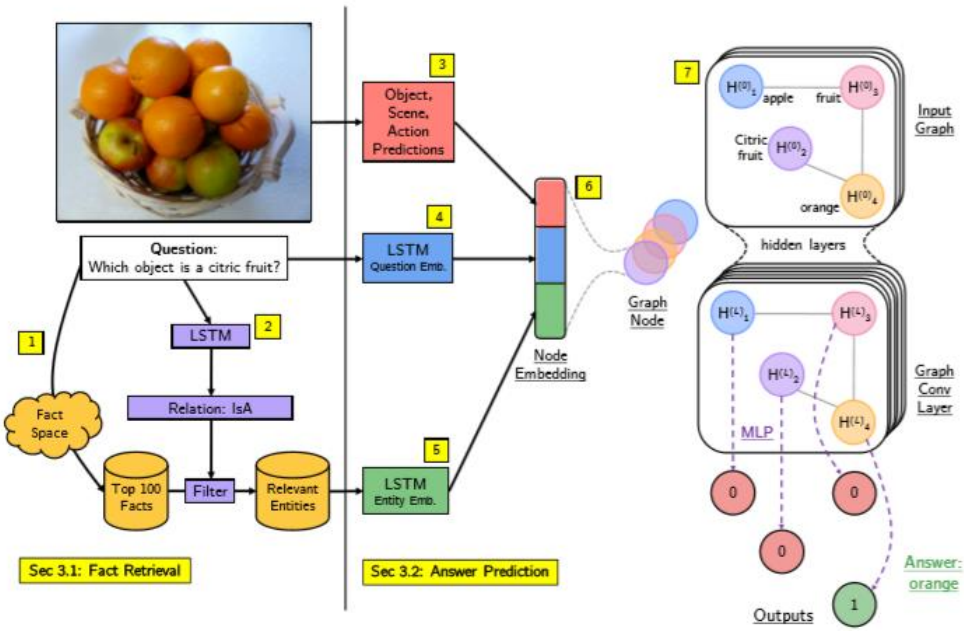


图 5

5. **Out of Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering**

准确回答有关给定图像的问题需要结合观察和一般知识。虽然这对人类来说是轻而易举的，但是用一般知识进行推理仍然是一个算法上的挑战。为了推进这一方向的研究，最近引入了一项新颖的“基于事实的”视觉问题回答（FVCA）任务，以及一组通过关系将两个实体（即两个可能的答案）联系起来的大量精选事实。对于一个问题-图像对，采用深度网络技术来连续减少大量事实，直到最后一个剩余事实的两个实体之一被预测为答案。我们观察到一个连续的过程，一次考虑一个事实，形成一个局部决策是次优的。相反，我们开发了一个实体图，并通过联合考虑所有实体，使用一个图卷积网络来“推理”正确答案。我们在具有挑战性的 FVCA 数据集上表明，与最先进的技术相比，这将导致精度提高约 7%。模型结构如图 6 所示



参考文献

[1] Marino K, Rastegari M, Farhadi A, et al. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge[J]. arXiv preprint arXiv:1906.00067, 2019.

[2] Gao P, Jiang Z, You H, et al. Dynamic Fusion With Intra-and Inter-Modality Attention Flow for Visual Question Answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6639-6648.

[3] Norcliffe-Brown W, Vafeias S, Parisot S. Learning conditioned graph structures for interpretable visual question answering[C]//Advances in Neural Information Processing Systems. 2018: 8334-8343.

[4] Wu J, Hu Z, Mooney R J. Generating Question Relevant Captions to Aid Visual Question Answering[J]. arXiv preprint arXiv:1906.00513, 2019.

[5] Narasimhan M, Lazebnik S, Schwing A. Out of the box: Reasoning with graph convolution nets for factual visual question answering[C]//Advances in Neural Information Processing Systems. 2018: 2654-2665.