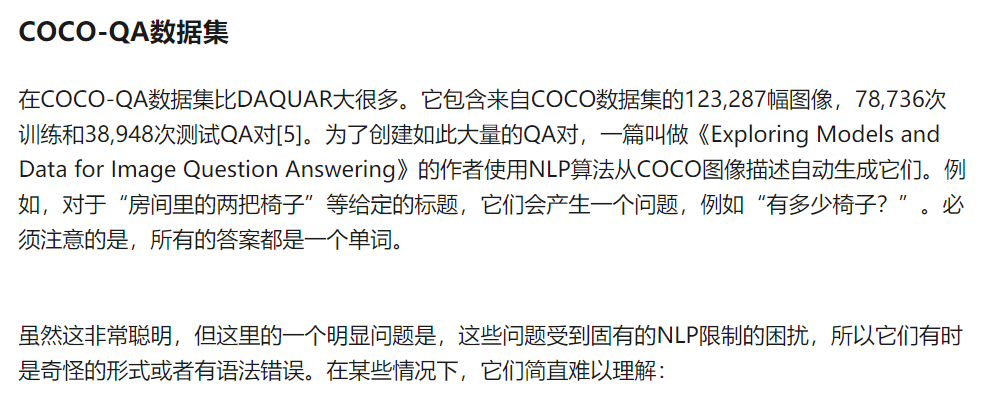
VQA的主要区别在于，搜索和推理部分必须在图像的内容上执行。因此，为了回答****在某处是否有人类****这样的问题，这个系统必须能够探测到物体。如果想要回答****是否下雨****这样的问题，系统就需要对一个场景进行分类。回到我们之前提到的问题，要回答出****哪个球员在踢球****，那么常识性推理以及可能的知识推理是必要的。这些任务中许多（包括对象识别、对象检测、场景分类等）已经在计算机视觉（CV）领域得到了解决，在过去的几年中取得了令人瞩目的成果。

一个好的VQA系统必须能够解决典型的自然语言处理和计算机视觉任务的问题，以及关于图像内容的推理。这显然是一个多学科的人工智能研究问题，涉及到计算机视觉、自然语言处理和知识表示（KR）等领域。

作为需要视觉理解与推理能力的，介于Vision与NLP间的视觉问答VQA，是一个有趣而又充满挑战的问题。它的进步不仅依赖于计算机视觉的发展和自然语言处理的能力，还需要对图像的理解——视觉基础能力，如识别，检测等，同时学习到知识与推理的能力。然而，这条路还有很长的距离要走，或许目前大部分方法只不过是对训练数据的拟合分类，并没有理解，也可能包含着语言先验（见CVPR2018论文Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering）。因此，一个真正理解图像，能够学习到知识和推理能力的模型才是最终目标。





几种方法的主要区别在于它们如何将文本特征和图像特征结合起来

[强人工智能](https://baike.baidu.com/item/%E5%BC%BA%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD/10403678" \t "https://baike.baidu.com/item/%E5%BC%BA%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD/_blank)观点认为有可能制造出真正能推理（Reasoning）和解决问题（Problem\_solving）的智能机器，并且，这样的机器将被认为是有知觉的，有自我意识的。可以独立思考问题并制定解决问题的最优方案，有自己的价值观和世界观体系。有和生物一样的各种本能，比如生存和安全需求。在某种意义上可以看作一种新的文明。

