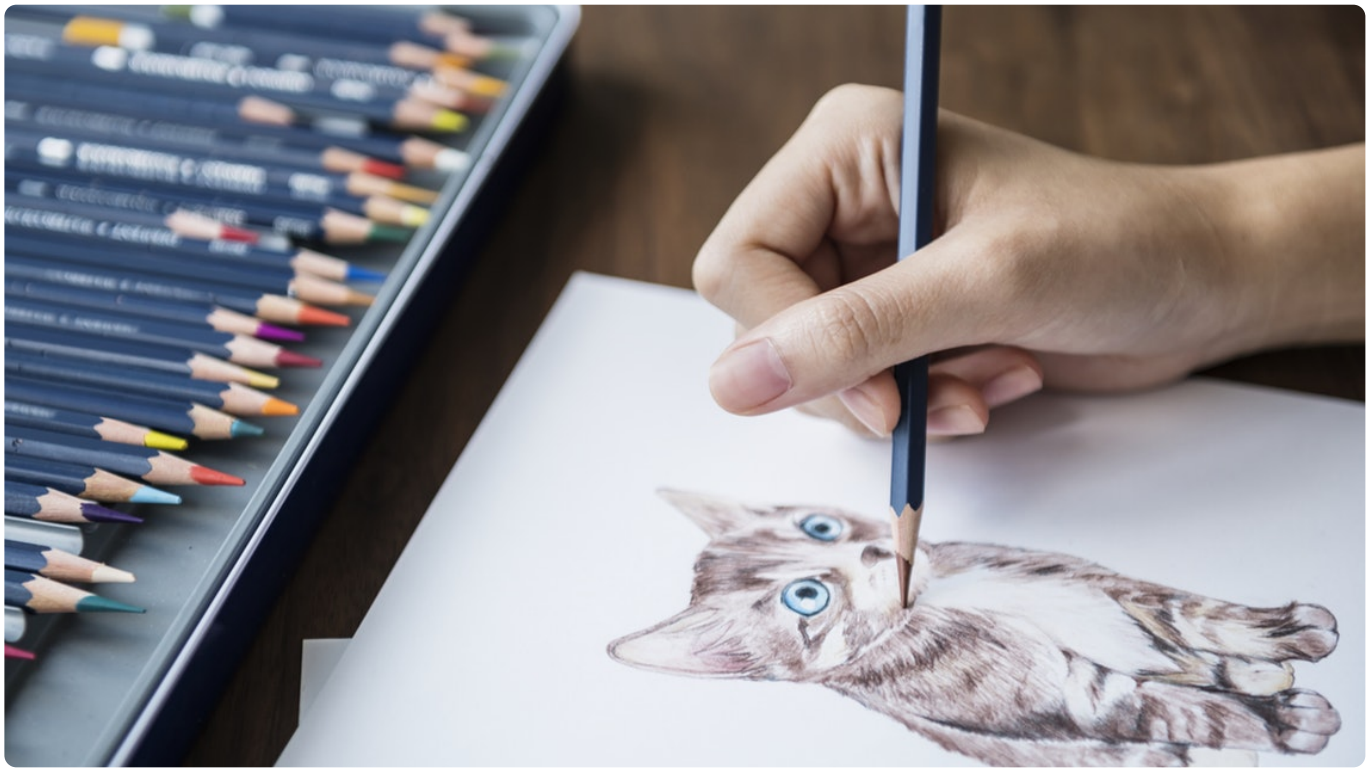
150 | 计算机视觉高级话题（二）：视觉问答

2018-09-17 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 04:50 大小 2.22M

今天，我们继续分享计算机视觉领域的高级话题，聊一聊“**视觉问答**”（Visual Question Answering）这个话题。

我们在前面曾经提到过“问答系统”（Question Answering），可以说这是人工智能领域最核心的问题之一。传统的问答系统主要是针对文字而言的，问题和答案都是以文字的形式表达的。当然，问答所针对的内容，有可能来自一个外在的知识库，例如维基百科。

我们今天要讨论的视觉问答，特别是“自由形式”（Free-Form）或者“开放形式”（Open Ended）的视觉问答，主要指的是根据一个图片进行自由的基于自然语言的问答。例如，我们可以问一个图片中是否存在一只猫；或者可以问图片里的天气是不是阴天等等。

视觉问答的挑战

那么，为什么视觉问答会在最近几年里得到很多学者的关注呢？我们有必要先来分析一下视觉问答所面临的挑战。

首先，**视觉问答需要对图片中的细节加以理解。**例如，我们问图片中的匹萨用了哪种奶酪，那就代表着我们的系统必须能够识别匹萨中的奶酪，而这往往意味着非常微观的一些细节的物体的识别。

其次，**视觉问答还需要我们对图片的上下文进行理解。**例如，我们可以问图片中有几辆自行车。这个问题其实不仅需要我们对图片中的自行车进行理解，还需要能够计数，这显然是一种更加复杂的理解任务。

除此以外，**我们还需要对图片中的物体进行推理。**例如，我们问图片中的匹萨是不是素食匹萨。那这个问题就需要对匹萨的种类进行分类，这是一个最基本的推理。

当然，视觉问答的挑战还远远不止这些。但从这些例子我们已经可以看出，视觉问题是一个综合性的人工智能问题。

不少视觉问答的数据集除了纯粹的图片作为输入以外，还有一个图片的“标题”（Caption）。这个图片标题往往提供了不少的信息，也算是帮助研究者在一定程度上降低了任务的难度。

如果需要对视觉问答的总体情况有一个更加深入的理解，推荐你阅读我在文末列出的参考文献 [1]。

视觉问答建模

接下来我们来聊一个视觉问答的基础模型 [1]。这个模型需要对问题、图片以及图片标题分别进行建模，从而能够进行问答。

针对问题，模型利用所有问题中的重要词进行了“词包”（Bag of Words）的表达，并且得到了一个 1030 维度的输入表征。类似地，针对图片标题，模型也进行了词包表达，得到了一个 1000 维度最高频词的表征。最后，作者们利用了 VGG 网络来提取图片的特征，得到了一个 4096 维度的图像表征。一种更加简单的方法则是先利用神经网络的隐含层，针对

每一种特征单独训练，然后把第一层中间层给串联起来。串联起来之后，这就是所有特征的一种联合的表达了。那么我们可以再经过一层隐含层学习到各个表征之间的相互关系。

文章中还讨论了另外一种模型，那就是利用 LSTM 来把问题和图像结合到一起，来最后对回答进行预测。

在这样的模型架构下，回答的准确度大概在 55% 左右。如何来理解这个准确度呢？在同样的一个数据集中，如果针对所有的问题回答都是“是”（Yes）所达到的准确度大概是 20% 多。

在最初的模型被开发出来以后的几年时间里，针对视觉问答的各类模型如雨后春笋般爆发式地增长。其中一个大类的模型利用了“关注”（Attention）机制。在深度模型中，**关注机制是一种相对来说复杂一些的“加权”模式。**也就是说，我们希望对某一些神经元或者是隐含变量更加关注一些。这个机制在视觉问答中的一种应用就是，针对不同的问题，我们希望让模型学习到图片的哪一部分来负责回答。

在一篇论文中 [2]，作者们提出了一种更加高级的“关注”机制，那就是“**层次同关注**”（Hierarchical Co-Attention）。

这个机制是什么样的呢？针对某一个回答，我们不仅要学习到究竟需要模型“看到”图片的某一个局部，这也就是我们刚才说到的“加权”，还需要针对问题，也就是文字，进行“加权”。这里的一个观察是，有时候一个问题中的核心其实就是几个关键词，这些关键词直接影响了回答。这就是“同关注”这一概念。

文章中还提出了另外一个概念，那就是“层次关注”，是指问题的文字，在单词、短语以及整个提问三个层次来进行建模。可以说，这种方法在语义的局部以及整体上更能找到问题的核心所在。

最后，需要提及一点，最近的一些研究又把视觉问答和“推理”（Reasoning），特别是“神经编程”（Neural Programming）联系起来，让回答问题变成自动生成程序的某种特殊形式 [3]。

小结

今天我为你讲了计算机视觉高级话题之一的视觉问答的概念。

一起来回顾下要点：第一，我们讲了视觉问答所面临的三大主要挑战；第二，我们讨论了对视觉问答进行建模的一些基本思路。

最后，给你留一个思考题，你觉得当前视觉问答的主要瓶颈是什么？

欢迎你给我留言，和我一起讨论。

参考文献

1. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh. **VQA: Visual Question Answering**. The IEEE International Conference on Computer Vision (ICCV), pp. 2425-2433, 2015.
2. Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. **Hierarchical question-image co-attention for visual question answering**. Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS' 16), Daniel D. Lee, Ulrike von Luxburg, Roman Garnett, Masashi Sugiyama, and Isabelle Guyon (Eds.). Curran Associates Inc., USA, 289-297, 2016.
3. Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, Ross B. Girshick. **Inferring and Executing Programs for Visual Reasoning**. ICCV 2017: 3008-3017.

上一篇　149 | 计算机视觉高级话题（一）：图像物体识别和分割

下一篇　151 | 计算机视觉高级话题（三）：产生式模型

精选留言 (1)

写留言

sky
2018-09-19

我认为目前的瓶颈在于，当前的模型都只对图像中的物体进行建模，而没有对图像中模型之间的关系进行建模，而机器问答很重要的一点是了解物体之间的关系。目前感觉对图像中物体间的关系，不管是空间关系还是语义关系进行建模都非常地难。