# 视觉基因组：使用众包密集的图像注释连接语言和视觉[[1]](#footnote-1)

赵淼译

(北京交通大学电子与信息工程学院，电子科学与技术，15650761201)

## 摘要

尽管在诸如图像分类等感知任务方面取得了进展，但计算机在诸如图像描述和问题应答等认知任务上仍然表现不佳。认知是任务的核心，它不仅涉及到认知，还涉及到我们的视觉世界。然而，用于处理认知任务图像中丰富内容的模型仍在使用为感知任务设计的相同数据集进行训练。为了在认知任务中取得成功，模型需要理解图像中物体之间的相互作用和关系。这篇论文里展示了是视觉基因组数据集对图像中的关系进行建模。我们收集每个图像中对象、属性和关系的密集注释，以学习这些模型。具体地说，我们的数据集包含超过108 K的图像，其中每个图像平均有35个对象、26个属性和21个成对的对象之间的关系。我们在区域描述中对对象、属性、关系和名词短语进行了规范化，并对单词-Net同步进行了回答。总之，这些注释代表了图像描述、对象、属性、关系和问题答案对的最密集和最大的数据集。

## 1 介绍

计算机视觉的圣杯是对视觉场景的完全理解：一个能够命名和检测对象的模型，描述它们的属性，并识别它们之间的关系。理解场景将使重要的应用程序，如图像搜索、问题回答和机器人交互。近年来在实现这一目标方面取得了很大进展，包括图像分类（Perronnin等人，2010年;Simonyan和Zisserman 2014;Krizhevsky et al . 2012;Szegedy等人，2015年）和目标检测（Girshick等人，2014年;Sermanet et al . 2013;Girshick 2015;任正非et al . 2015 b)。一个重要的贡献因素是大量数据的可用性，这些数据驱动着今天在计算视觉理解方面的进步的统计模型。虽然进展令人兴奋，但我们离实现全面的场景理解还有很长的路要走。如图1所示，现有的模型将能够检测照片中的离散物体，但无法解释它们之间的相互作用或它们之间的关系。这样的解释往往是认知性质的，将知觉信息整合到一个场景中物体之间的关系（Bruner 1990;费尔斯通和肖勒2015)。例如，一个最先进的模型（kar病变和fei-fe2015）在图1中描述了一个MS-COCO的图像，因为两个男人站在大象旁边。但是缺少的是进一步了解每个物体的位置，每个人在做什么，人与大象之间的关系，等等。如果没有这样的关系，这些模型无法区分这张图片和其他大象旁边的人的图像。

彻底地理解图像,我们相信三个关键元素需要被添加到现有的数据集:对于视觉概念真实的语言对应(Kiros et al . 2014年),一个更完整的描述和QAs为每个图像基于多个图像区域(Johnson et al . 2015年),和一个形式化的表示图像的组件(海斯1978年)。

一个从视觉概念到文本的基本信息

对每张图片都包括对图片内多个区域的 完整的descriptions和QAs的数据集

对图片中的每一个元素都有的一个 形式化的表述方式

在绘制视觉世界的完整信息的精神下，我们引入了视觉基因组数据集。视觉基因组数据集的第一个版本使用了108,077图像来自YFCC100M（Thomee等人，2016）和MS-COCO（Lin等人2014）的交集。第5节提供了关于数据集的更详细的描述。我们强调以下三个关键要素的贡献和贡献，这些要素将视觉基因组与现有的数据集区分开来。



图1从知觉意识到对图像的认知理解所需要的数据的概述。我们提供了大量的图像数据集，其中包含了大量的区域描述、对象、属性和关系。一些地区描述的例子（例如，女孩喂养大象和一个男人在女孩身后拍照）被显示（上）。对象（如大象）、属性（如大型）和关系（如喂食）显示（底部）。我们的数据集还包含与图像相关的问题答案对（未显示）

视觉基因组数据集将关系和属性视为注释空间的一等公民，除了传统的对象焦点之外。对关系和属性的认识是对视觉场景的完全理解的重要部分，在很多情况下，这些元素是场景故事的关键（例如，狗追逐男人和追逐狗的人之间的区别）。视觉基因组数据集是第一批提供对象交互和属性的详细标记的，将视觉概念根植于语言。

一张图片通常是一个丰富的风景，在一个总结句子中不能被完全描述。图1中的场景包含多个故事:一个人拍照的大象,一个女人喂养一头大象,一条河的背景中郁郁葱葱的理由,等。现有的数据集,如Flickr 30 k(年轻et al . 2014年)和MS-COCO林(et al . 2014年)关注图像的高层描述。相反，对于视觉基因组数据集中的每一个图像，我们收集了超过50个不同区域的描述，为场景提供了更密集、更完整的描述。此外，受byVQA（Antol等人2015）的启发，我们还根据每个图像的描述平均收集了17个问题答案对。基于区域的问题答案可以用来联合开发NLP和视觉模型，这些模型可以回答来自描述的问题或来自图像的问题，或者两者都可以。

一系列密集的描述图像和视觉像素之间的明确的对应关系(即边界框的对象)和文本描述符(即关系、属性),视觉基因组数据集将是第一个图像能够提供一个结构化的数据集的形式化表示的图像,在广泛应用于知识库的形式表示在NLP(周et al . 2007;成国栋et al . 2005;Culotta和索伦森2004;Socher et al . 2012年)。例如，在图1中，我们可以正式表达妇女和食物之间的关系（女人，食物）。把所有的对象和关系放在一个场景中，我们可以将每个图像表示为一个场景图（Johnson等人，2015）。场景图表示已被证明可以改善语义图像检索（Johnson等人，2015;舒斯特尔等人，2015年）和图像字幕（far哈迪等人，2009年;Chang et al . 2014;古普塔和戴维斯2008)。此外，在可视基因组数据集中，每个图像中的所有对象、属性和关系都被规范化为相应的WordNet（Miller 1995）ID（称为synset ID）。这个映射连接了视觉基因组中的所有图像，并提供了一种有效的方法来持续地查询数据集中的相同概念（对象、属性或关系）。它还可以帮助训练能够从多个图像中学习上下文信息的模型（图2、3)。

摘要本文介绍了视觉基因组数据集，旨在对下一代计算机模型进行训练和基准测试，以实现全面的场景理解。本文的内容如下：在第2部分中，我们提供了数据集的每个组件的详细描述。第3部分提供了相关数据集的文献回顾和相关的识别任务。第4部分讨论了我们在收集这个数据集的过程中部署的众包策略。第5部分是数据分析统计数据的集合，展示了可视化基因组数据集的关键属性。最后但并非最不重要的是，第6节提供了一组实验结果，将视觉基因组作为基准。

首先，给定一张图片，worker需要手工撰写“region description”，并且用bounding box框出这个region。

得到region description之后，worker开始从这些文本中抽object（通常是名词），同时再用bounding box严格地框住这些object。

有了object之后，再根据region description、图片本身等数据，人工抽取这些object的attribute和region内部多个object之间的relationship。将这些东西combine在一起就得到了一个region graph

有了region graph和其他各种数据之后，找到各个region graph中公共的object（可以通过包围盒的overlap来判断该标注的object是否对应了同一个object）。用这些公共的object作为桥梁，将这些region graph连接在一起，就形成了scene graph。

与此同时，再找一伙worker，给图片标QA pairs。QA Pairs 有两种。

5.1. 第一种是freeform QA pairs。把图片给worker，让worker针对这个图片提出至少8个问题（必须是以who, what, where, when, why and how开头的问题）。同时这个8个问题中必须包含这六类中的三类。

5.2. 第二种是region-based QA pairs。把region给worker，让worker针对这个region提出一个问题。

再找一伙worker做double check。如果符合要求，就把所有这些东西添加到数据集中。

然后就得到了Visual Genome

知识表示：这里的场景图以一种类似自然语言理解中的知识图的形式发挥作用。

我们分析了这个数据集的各个组成部分，并给出了一些实验的基线结果，例如属性分类、关系分类、描述生成和问题回答。然而，我们的数据集可以使用更多的应用程序和实验。在本节中，我们将注意到我们的数据集能够启用的一些潜在应用程序。密集的图像字幕，我们看到了很多图片字幕试图用一个标题来描述整个图像。然而，这些文字说明并没有详尽地描述场景的每一个部分。这个应用程序的自然扩展，是视觉基因组数据集所支持的，是创建描述场景部分的密集字幕模型的能力。视觉问题回答作为一项独立的任务被研究过，我们引入了一个数据集，将我们所有的问题答案与描述和场景图结合在一起。未来的工作可以建立监督模型，利用视觉基因组的各种组成部分来解决问题的答案。当我们看到图像字幕的激增（Kiros等人2014）和问答（Antol等人2015）的模型时，在创建更全面的评估指标来衡量这些模型的表现方面，几乎没有什么工作。这些模型通常使用BLEU、苹果酒或流星和其他类似的指标来评估，这些指标不能有效地衡量这些模型对图像的理解程度（Chen等人，2015）。视觉基因组场景图可以作为图像理解的测量。生成的描述和答案可以与图像的地面实况图相匹配，以评估其相应的模型。摘要在信息检索和自然语言处理中，对关系提取关系的提取进行了广泛的研究（周等人，2007年;成国栋et al . 2005;Culotta和索伦森2004;Socher et al . 2012年)。视觉基因组是第一个大规模的视觉关系数据集。这个数据集可用于研究从图像中提取的视觉关系（Sadeghi等人2015），它之间的相互作用也可以用来研究动作识别和物体间的空间定位。语义图像检索先前的工作已经表明，场景图可以用来改善语义图像搜索。我们可以使用区域描述和区域图来探索进一步的方法。还可以探索基于注意力的搜索方法，其中查询指定的兴趣区域也在检索到的图像中进行本地化。在视觉基因组为认知图像理解最密集的视觉数据集的同时，完成一组注释，它仍然是不完整的。在大多数图像中，为每一个对象或一对对象收集一组详尽的属性和关系是不可行的。这引发了两个新的研究问题。在计算机视觉方面，我们需要开发新的评估指标，因为缺乏完整的注释集，所以不会对模型进行惩罚。在人机交互中，我们需要设计新的界面和工作流，激励人们去注释视觉常识。

1. 赵淼 电信研1706班 电子科学与技术 17120036 15650761201 [↑](#footnote-ref-1)