# 图像的自动描述生成：对模型，数据集或评估措施的调查[[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)

## 1 摘要

对自然图片进行自动描述生成是一个具有挑战性的问题，近年来这个问题已经受到机器视觉领域和自然语言处理领域的广泛关注。在调查中，作者根据如何对这一问题进行概念化对现有的解决图片描述问题的方法进行了分类，也就是这些模型有的将问题描述为生成问题有的也可以描述为视觉或多模态表示空间上的检索问题。模型我们对当前模型进行了细致的讨论，重点强调了他们的优缺点。我们概述了评估机器生成图像描述质量的基准图像数据集和评估措施。最后我们对自动图像描述领域未来的发展方向进行了推断。

2 介绍

近些年的许多应用多属于计算机视觉和自然语言理解的交叉，如在一篇报纸文章的背景下解读照片，按照说明把与图表或地图结合起来，在听报告的时候理解幻灯片。网络还提供了大量的语言和视觉信息相结合的数据：标签照片，报纸插图，带有字幕的视频以及社交媒体上的多模式信息。

这项任务涉及拍摄图像，分析其视觉内容，并生成文字描述（通常是一个句子），用语言表达图像最显着的方面。从CV的观点来看，这是具有挑战性的，因为描述原则上可以谈论图像的任何视觉方面：它可以提及对象及其属性，可以谈论场景的特征（例如，室内/室外），或者表达场景中的人物和物体如何相互作用。更具挑战性的是，描述甚至可以涉及没有描述的对象（例如，即使在列车因为尚未到达而不可见的情况下，等待列车的人也可以谈论）并提供不能直接从图像提取到的的背景知识（例如，描绘的人是蒙娜丽莎）。简而言之，好的图像描述需要全面的图像理解，因此描述任务对于计算机视觉系统来说是一个很好的测试平台，比普通的CV评估更全面，通常测试例如物体检测器的精度或场景分类器在一组有限的类上。

对于图像的理解是必要的，但不足以产生一个好的描述。如果用一系列先进的检测去检测图像中的目标物体，确定属性，场景属性，人与物体之间的相互关系，结果将是一个很长的，非结构化的清单，不能用作图像的描述。好的图像描述应该是综合但是简洁的（只谈论图像中重要的事物），在形式上必须是正确的，包含具有良好的语法结构的句子。

从自然语言处理角度，这是一个自然语言生成（NLG）问题。典型NLG的任务是将非语言表达转化为人类可读的文本。通常，非语言表示是逻辑形式，数据库查询或一组数字。在图像描述中，输入是一个图像表示（例如，在前一段中列出的检测器输出），NLG模型必须变成句子。生成文本涉及一系列的步骤，传统上被称为NLP管道（Reiter＆Dale，2006）：我们需要决定哪些方面的输入要谈论（内容选择），然后我们需要组织内容（文本规划）和表述它（表面实现）。反过来，表面实现需要选择合适的词语（词汇化），如果合适的话使用代词（参考表达式生成），并将相关信息聚合在一起（聚合）。

这个任务当面向特定的用户时更具挑战性。举例一个艺术评论家相对于一个图书管理员或是游客对这方面的要求都是不同的。

这里将现有的基于图像描述的方法分为3类：传统的基于通道的模型，他们首先检测图像内容包括目标，属性，场景类型和动作基于视觉特征。然后这些模型用这种内容信息驱动自然语言产生系统输出图像描述。称为直接产生模型。

第二类时基于检索的模型，这种方法是找到近似的图像，然后通过这些图像对应的描述来迁移或是生成图像描述。涉及到两个子领域空间，视觉空间检索图像，多模态空间共同表示图像。

或是视频的理解。

## 2 图像描述模型

图像描述可以在几个不同的维度进行分析。这篇调查文章感兴趣的描述是描述图像中描述的视觉和概念信息的描述，即描述涉及所描述的实体及其属性关系以及它们所涉及的行为。在自动图像描述的范围之外的是非视觉描述，它们给出背景信息或引用图像中未描述的对象（例如，拍摄图像的位置或者谁拍了照片）。此外，与图像描述的标准方法不相关的是感知描述，其捕捉图像的全局低级视觉特征。

三类模型：直接生成模型，从视觉空间中检索模型，以及从多模态空间中检索模型。

### 2.1 从视觉输入生成描述

这组研究的一般方法是首先通过分析其视觉内容来预测给定图像的最可能的意义，然后生成反映该意思的句子。此类别中的所有模型都使用以下通用管线架构来实现此目的：

1.使用计算机视觉技术对场景类型进行分类，检测图像中存在的对象，预测它们之间的属性及其关系，并识别发生的行为。

2.接下来是生成阶段，将检测器输出转换为单词或短语。然后使用自然语言生成技术（例如，模板，n元语法，语法规则）将这些结合在一起以产生图像的自然语言描述。

这种方法基于图像到描述的明确映射，后面小节的研究包含隐式视觉和语言模型。显式流水线架构针对当前的问题而定制，因为它依赖于预定义的场景，对象，属性和操作的语义类集合，因此会限制生成的描述。此外，这样的架构关键地假定每个语义类别的检测器的准确性，这在实践中并不总是被满足。



如图所示的方法通过图像进行检测获取目标，属性，介词，然后构建一个条件随机场，预测标签产生句子。

描述生成的方法在两个主要方面有所不同：（a）它们从哪个图像表示中获得描述;（b）它们如何处理句子生成问题。在表示方面，现有的模型依靠空间关系，语料库关系或是空间视觉属性。另外还有一些人通过抽象出的图像表示形式表示图像的各个方面。后来，又提出了密集标记的图像标注，包含对象，属性，动作，场景注释等等。在后来还有的方法不适用实现标记好的属性和方法，而是直接从图像和相关描述中使用多实例学习（弱监督的方法训练物体检测器。然后将检测器返回的单词输入语言模型以生成句子。

解释：总体感觉是在不停增加图像的信息量，然后提高效果，或是使用弱监督的方法结合图像和描述学习。

第一个明确表示图像结构如何与其描述结构相关的框架是Elliott和Keller（2013）提出的Visual Dependency Representations（VDR）方法。这种方法捕捉的是图像中一幅图像中各个目标的空间关系，以依赖图的形式。该图可以与图像描述的语法依赖树相关。刚开始这种方法基于人工的标注的数据集，后来更多的方法被用到了这种方法上，例如目标检测器的输出或是抽象场景下的标签。Lin等人明确提出了表示图像结构并将其用于描述生成的想法。（2015），他们将图像解析成与VDR类似的场景图形，并表示场景中物体之间的关系。然后他们使用语义语法从场景图生成。

现有的方法在语句生成这部分表现也不相同。一种方法是n-gram方法，可以使用维基百科来训练获得这个模型。这些方法首先确定图像中的区域之间的属性和关系作为区域 - 介词 - 区域三元组。在语言模型给定的情况下，n-gram语言模型被用来组成一个流畅的图像描述。后来Fang等人的方法（2015）是相似的，但是使用最大熵语言模型而不是n-gram模型来生成描述。这为作者提供了更大的灵活性来处理作为其模型核心的单词检测器的输出。

RNN也可以被认为是语言模型，他在给定已产生词基础上通过给出下一个产生词的概率获得产生的词。而在图像描述领域RNN既使用语言信息也使用视觉信息。

另一种方法使用句子模板来产生描述。这些是（通常是手动的）预先定义的句子框架，其中开放时隙需要用对象，关系或属性的标签填充。例如使用隐马尔科夫模型选择可能的对象，动词，介词和场景类型来填充句子模板。通过在Gigaword外部语料库中找到最可能的对象标签配对来生成动词。Elliott和Keller（2013）的生成模型将图像解析为VDR，然后遍历VDR以填充句子模板的槽。这种方法还通过在训练时间学习VDR和句法依赖树之间的关联来进行有限的内容选择;这些关联允许在测试时为描述选择最合适的动词。

还有通过更加语言学的复杂方法生成。如米切尔等。（2012）过度生成句法结构完整的句子片段，然后使用树替换语法重新组合这些句子片段。库兹涅佐娃（Kuznetsova）等人采取了相关的方法（2014），从现有描述的训练集中学习树片段，然后将这些片段在测试时间组合以形成新的描述。最近Ortiz等人提出了另一种语言表达模型。 （2015年）。作者将图像描述建模为VDR句对上的机器翻译，并在语言约束下使用整数线性规划执行显式内容选择和表面实现。

这种生成式的方法引入了评价的困难，因为这里对于语言学的困难的注意影响了对于图像进行理解。因此这里提出了一种方法，可以独立于生成方面评估图像和句子之间的映射关系。遵循这种方法的模型将图像描述概念化为检索问题：它们通过检索和排列具有候选描述的一组相似图像来将图像与描述相关联。然后可以直接使用这些候选描述（描述传送）或者可以从候选（描述生成）合成新的描述。这种检索可能是视觉空间上的和或是视觉文本空间相结合。

### 2.2 视觉空间检索的描述

这一类是通过检索相似图像来实现图像描述，将相似的图像空间迁移描述到图像中。这种方法需要大量含有描述的图像。

在算法角度，分为以下三步：

1.通过视觉特征表示特定图像

2.根据相似度量从训练集中检索候选图像集

3. 对候选图像根据视觉信息或文本信息，进行重新排序。或是有选择的结合候选片段描述的片段根据一些特定的规则或框架。

遵循这种方法的第一个模型是Ordonez et al的Im2Text模型(2011)。GIST（Oliva&Torralba，2001）和Tiny Image（Torralba，费格斯，弗里曼，2008），他们采取不同的描述子衡量相似度。大多数基于检索的模型都考虑利用这一步作为基准。重排序部分是利用各种检测器，目标，场景，行人等等来捕捉视觉内容，图像被这些检测器和分类器的响应进行表示，重排序通过这些语义特征训练分类器进行的。

这一模型第一次由Kuznetsova et al. (2012)提出，它通过运行检测器和分类器作用在Im2Text模型作为重排序，提取图像在语义内容上的表示，同时执行图像检索步骤对图像中每一个视觉实体以收集相关的短语。举例，如果狗被检测到了，这里会从训练集中返回相应的短语。更特定化的，这一步被用来三种不同种类的短语。基于训练图像和查询图像中检测到的区域的相似性，从训练集合中提取名词和动词短语。根据查询图像与训练图像在全局场景的相似度，为每个场景收集介词短语。

Gupta等人的方法（2012）是另一种基于短语的方法。这里使用RGB，SIFT等基本的描述符作为图像特征，然后依据这些特征寻找相似图像中的文本信息。这一过程不依赖于检测器对描述符的处理，只是将根据描述符寻找不同类型的短语，在每种短语下做一个相似性处理。通过相似度和谷歌搜索计数作一个联合概率分布获得最后的结果获得三种三元组，在这三个顶级的三元组下在固定模板下产生描述，另加入一些语法规则。

Patterson等人（2014）是第一个在计算机视觉领域提出大型场景属性数据集。该数据集包括来自707个场景类别的14,340个图像，其被从与材料，表面属性，照明，可供件和空间布局相关的102个判别性属性列表中的某些属性注释。这允许他们从这个数据集中训练属性分类器。在他们的论文中，作者还表明，这些属性分类器的响应可以被用作全局图像描述符，它比标准全局图像描述符（如GIST）更好地捕获语义内容。作为一个应用，他们通过用自动提取的场景属性替换全局特征来扩展Im2Text的基线模型，给出更好的图像检索和描述结果。

Mason和Charniak（2014）的描述生成方法将描述生成描述为一个抽取摘要的问题，最后排序过程仅考虑文本信息选择输出描述。首先使用场景描述符识别视觉相似图像，然后通过检索图像描述的非参数密度估计来估计在查询图像的描述中观察词的条件概率。然后使用两种不同的抽取摘要技术来确定最终的输出描述。

后面的两种方法采用CNN做为特征提取器，其中一种使用k最邻作为相似图像选择，然后通过这些图像描述的平均n-gram作为描述。

### 2.3 多模态空间检索描述

这里将图像描述作为一个检索问题，但是在作为多模态空间下进行检索。

这部分步骤如下所示：

1.学习一个通用的多模态空间对于视觉和文本数据用一个训练的特征描述对

2.给定一副检索图像，使用联合表示空间表现交叉检索

相对于视觉空间的检索模型而言，这种模型将视觉和语句特征嵌入到共同的多态空间中这种方法可以进行双向转换。

首先讨论Hodosh等人的开创性论文，他们提出使用CCA（典型相关分析）的改进版KCCA，摒弃了以前使用（目标，动作，场景）的三元组来表示共同的语义空间。CCA采用语言句子对的形式作为训练集，因此输入来自于两个空间。然后找到一个线性预测到一个公共空间，在KCCA中，核函数将原始项目映射到高阶空间，以便捕获将图像和文本关联所需的模式。KCCA之前曾被证明可以成功地将图像（Hardoon，Szedmak，＆Shawe- Taylor，2004）或图像区域（Socher＆Fei-Fei，2010）与单个词或一组标签联系起来。

这种方法需要保存两个核矩阵，只对数据量小的情况使用，后来开始使用神经网络代替构建联合空间形成图像描述。

Socher等人（2014）使用神经网络来构建句子和图像矢量表示，然后将这些表示映射到一个共同的嵌入空间。他们采用了组合语句进行向量的表示。图像和文字的表现形式是以单一形态学习的，最终映射到一个共同的多模式空间。们使用一个DT-RNN（依赖树递归神经网络）来组成语言向量来抽象语义上不相关的词序和句法差异。这导致了50维词嵌入。对于图像空间，作者使用一个在ImageNet数据上训练的九层神经网络，使用无监督的预训练。图像嵌入是通过取最后一层（4,096维）的输出而得出的。然后通过最大边缘目标函数将这两个空间投影到多模态空间中，直观地训练一对正确的图像和句子向量以获得高的内积。作者表明，他们的模型比以前使用的KCCA方法（如Hodosh和Hockenmaier（2013））要好。

后来的人对于映射的粒度进行修改，将图像和句子片段进行嵌入和顶层信息一起获得嵌入空间，还有基于短语，基于概念的方法。

Karpathy等人（2014）扩展了以前的多模式嵌入模型。它们的模型不是直接将整个图像和句子映射到一个共同的嵌入空间，而是将更多的细粒度单元，即图像（对象）和句子（依存树片段）的片段嵌入共同的空间。他们的最终模型整合了全（句子和图像层面）以及细粒度的信息，并且优于以前的方法，如DT-RNN（Socher et al。，2014）。 Pinheiro等人采用了类似的方法。 （2015），他们提出了一个双线性基于短语的模型，学习图像表示和句子之间的映射。然后使用受约束的语言模型从这个表示生成。 Ushiku等人提出了一个概念相关的方法。 （2015）：作者使用一个公共的子空间模型，将与同一个短语相关的所有特征向量映射到空间的附近区域。为了生成，使用基于波束搜索的解码器或模板。

描述生成系统很难进行评价，因此通常视为检索，排序任务，但这受到数据集和可用性的影响。最近的一些方法，在多模态空间中不能只是排序，还要生成。

Kiros等人（2015）引入了图像描述排序和生成的通用编码器 - 解码器框架。编码器首先构建联合多模态空间。这个空间可以用来排列图像和描述。第二阶段（解码器）然后使用共享的多模式表示来产生新颖的描述。使用长期短期记忆（LSTM）递归神经网络编码句子，使用深度卷积网络（CNN）编码图像特征。在Kiros等人的（2015）编码器 - 解码器模型中，视觉空间被投影到LSTM隐藏状态的嵌入空间中;为了学习图像的排序及其描述，最小化成对排序损失。解码器，一个基于神经网络的语言模型，能够从这个多模态空间中产生新的描述。

然后几人基本上实在CNN，RNN的架构里进行改进，具体变化是使用形式不同，例如可以将静态图像与前一个字作为副本直接作为输入。或是增加语义信息。

Karpathy和Fei-Fei（2015）通过提出一个更简单的架构和目标函数的深度视觉 - 语义对齐模型来改进以前的模型。假定句子的部分是指图像中特定但未知的区域。他们的模型试图推断图像句子和区域之间的对齐，并且基于图像区域上的卷积神经网络，用在句子上的双向RNN 和使两种模式对齐的结构化目标。单词和图像区域被映射到通用的多模式嵌入。多模式递归神经网络结构使用推断的对奇来学习并产生新的描述。这里，图像被用作递归神经网络中的第一状态的条件，然后产生图像描述。

后面还有通过马尔可夫模型将图像特征和句法短语空间获得的短语进行嵌入生成句子。

最后弱监督学习和可移植性较少被研究。前面训练数据都是全注释，代价太高，是否可以使用弱监督的方法提高准确性。

### 2.4 方法比较

这三种方法各有优缺点，第一种方案将任务作为生成模型可以产生一些新的句子描述图像，但是对视觉内容的准确程度依赖高，这里明确的使用计算机视觉技术来获得图像最可能的含义，准确性有限。生成方面需要复杂的自然语言处理，极为复杂。

如果直接通过视觉空间解决问题，则总会产生语法正确的描述，但是需要大量的数据集，随着数据集减少，描述的准确性降低。

最后多模式空间问题具有类似生成类的优势，因为他能通过预定义的大型检索池获得合适的描述，但是这方面那难以进行准确度量，同时也需要大量的数据集，同时这种方法可以应用与双向的问题。

### 3 数据集介绍

#### 3.1 图像描述数据集

这部分数据集有许多是从物体识别数据集扩展出来的，如Pacal1K（常用的基准数据集）源自Pascal2008，是物体识别数据集。

视觉和语言数据库源自Pascal2010动作识别数据集，主要是行动者和其行为以及背景，还有相关性的表示。

Flickr数据集针对特定对象和动作的用户进行查询。

抽象场景数据集，由剪切画组成。

VQA数据集，场景问答数据集

IAPR-TC12数据集包含20000个描述图像（最初由搜索引擎检索提供）。包含完整的像素级分割。

MS COCO数据集由123287个图像组成，80个图像类别。

NYU数据集包含3D对象分割的1449个室内场景。

### 3.3 图像标题数据集

这部分比较抽象，所描述的内容不能在图像中直观的看到，包含许多文化，历史背景。

BBC数据集 图像及标题 3361篇文章，1个图象，1个标题。

D´ej`a-Image Captions 数据集 4000000幅图像，180000个几乎相同的标题。，进行了词法化和停用词处理。

这些数据集许多实在亚马逊众包网站上标注的。Amazon Mechanical Turk，涉及标注质量的问题。

### 3.4 评估措施

专家评审

•描述在语法上是正确的（Yang et al。，2011; Mitchell et al。，2012; Kuznetsova et al。，2012; Elliott＆Keller，2013等）。

•描述没有不正确的信息（Mitchell et al。，2012）。

•描述与此图像相关（Li等，2011; Yang等，2011）。

•描述是创造性地构建的（Li et al。，2011）。

•描述与人类相似（Mitchell et al。，2012）。

自动测量

（BLEU ROUGE Meteor）（为了测试机器翻译或是文本摘要而涉及的 CIDEr（专门用来评测图像描述）

这些评价指标可能并不准确，很多显示评测好的方法，实际使用时效果并不好。没有任何系统胜过人类评估。Meteor 被认为是最强有力的测量方法，系统在一个和两个提交文件（这取决于参考文献的数量）上打败了人类文本。用CIDEr测量的系统超过人类七到五次;根据ROUGE和BLEU，该系统几乎总是胜过人类，进一步证实了这些评估措施的不适用性。然而，库兹涅佐娃等（2014）发现，BLEU与Meteor在图像标题生成方面与人类判断的相关性更强。

最近的系统（从2014年开始）已经融合了大量描述数据集（Flickr8K / 30K，MS COCO）的使用，并采用与人类判断（Meteor，CIDEr）相关性较好的评估措施。然而，BLEU的使用尽管有其局限性，但仍然普遍存在。文献中也没有普遍使用人的评价。

### 4 未来方向

数据集：在数据集之间具有较大差异性，在数据集之间的迁移性较差，因此，可能需要收集更大型的数据集或是降低监督数量以获得更多的无监督模型。

评价指标：这里评价系统会高估图像描述的结构，而人工成本过高。这方面需要改进。

多样性和创新：如何产生多样性和独特的描述，预测潜在的语义。Chen和Zitnick（2015）及相关方法通过耦合描述和可视化表示生成来解决这些限制。Jas和Parikh（2015）介绍了图像特异性的概念，认为图像描述的形式不统一，某些图像比其他图像更具体。由于人们倾向于从不同的角度描述非特定的场景，非特定图像的描述往往会有很大的差异。这个概念及其对描述系统和措施的影响应进一步详细研究。

### 4.4 进一步的任务

视觉问答：多句话的生成。通过问答不断获得图像的信息。这里从特定领域的问答到自由问答。

多语言存储库是另一个方向。今后的工作应该研究在单一描述模型之间传递多模态特征是否导致与单语基准相比更好的描述。在多语种多模式环境中研究不同的模型和新任务将是有趣的，使用更大，更多的语法多样化的多语言描述语料库。

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