# 用引导对象生成图像描述

# 概要

虽然现有的图像描述模型可以使用递归神经网络(RNNs)达到预期的结果，但很难保证我们关心的对象包含在生成的描述中，例如在图像中不明显的对象。当我们关注的对象没有参与训练时，问题变得更加复杂。在这篇论文中，我们提出了一种新的方法来生成图像描述，该方法是以对象为导向的（我们称此方法为CGO）。该模型有一个约束条件：当人们关心的对象在图像中时，该对象必须参与模型的训练。CGO可以确保被关注对象在生成的描述中，而且保持语句的流畅性。我们不是从左到右生成序列，而是从一个选中的对象开始生成描述，并根据这个对象生成其他部分的描述。为了实现这一目标，我们设计了一个新的框架，它将两个LSTMs以相反的方向组合在一起。我们在MSCOCO上检验了模型的特性，在MSCOCO上，我们为图像中检测到的每个对象生成描述。使用CGO，我们可以将描述能力扩展到图像标题标签中被忽略的对象，并为图像提供一套更全面、更多样化的描述。CGO在描述新对象时有优势。我们展示了在MSCOCO和ImageNet数据集上的实验结果。评估结果表明，我们的方法在任务中的表现优于最先进的模型，平均水平为F1 75.8，从而在内容的准确性和流畅性方面都有更好的描述。

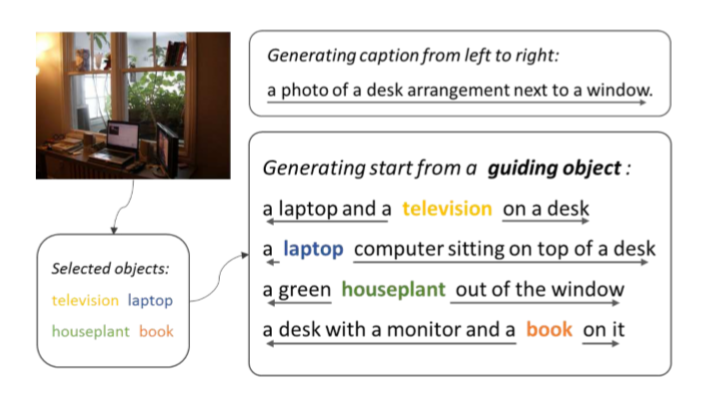


图1 现有的图像描述模型

现有的图像描述模型从左到右生成描述。我们的CGO方法开始生成一个选定的对象。CGO使我们能够将选择的对象精确地合并到描述中，并为图像生成一组多样而全面的描述。

# 介绍

为图像生成文字描述的任务，叫做图像描述，在计算机直觉中是一项挑战性任务。图像描述在生活中有很实际的应用，例如：为视觉障碍提供的机器场景感知和辅助系统。在过去的几年里，深度神经网络被广泛地应用于图像描述模型[25, 12, 40, 17, 3]，在普通数据集上（比如MSCOCO）能够流利且准确的描述图像内容。但是，它们对描述生成过程的控制是有限的。例如，一张图片可以包含很多物体，但是一个描述语句通常只包含一个或很少的物体，如图1所示。虽然现有的方法[13,36,32]可以准确地对图像中的物体进行分类或检测，但我们不能强迫语言模型去描述我们关心的物体。模型在训练中包含我们所关心的对象是很重要的，否则模型可能无法认知这个特殊的对象。在训练数据中没有出现的情况下，在描述中包含一个新的对象就更加困难了。最近的几部作品都研究了描述新奇物体的任务，但这仍然是一个悬而未决的问题。在这篇论文中，我们提出了一种新的方法来生成图像描述，该方法是以对象为导向的（我们称此方法为CGO）。CGO可以确保被关注对象在生成的描述中，而且保持语句的流畅性。所选对象可以是从图像中检测到的任何对象，即使它在图像描述模型训练数据中不可见。

近年来，编解码结构在图像描述任务中得到了广泛的应用，而递归神经网络(RNNs)也经常被用作描述图像的语言模型。在目前的方法中，描述通常是按照从左到右的单词顺序一个一个生成的。CGO是建立在编解码结构上的，但它不是生成从左到右的序列，而是基于选定的对象生成句子。我们称它们为引导对象，因为它们在生成过程中引导序列的内容。引导对象是我们想要包含在描述中的对象。它可以出现在序列中的任何位置。我们设计了一个新的框架，结合两个LSTMs[14]来生成对象周围序列的左部分和右部分。在这个过程中，重要的是两个序列的内容要一致。在CGO中，每个LSTM对序列的另一部分进行编码，然后根据编码的序列和图像的视觉特征生成一个序列。这有助于两个序列流畅地与引导对象连接。通过向LSTMs提供不同的信息序列，我们还可以为每个选中的对象生成多个不同的描述。

早期的一些图像标题任务的工作是基于模板的方法[20,11]。这些方法从图像中检测视觉概念并将其直接填充到模板中。虽然这种方法可以保证被选对象出现在描述中，但生成的句子形式有限，缺乏多样性。在CGO方法中，指导对象不经过编译码过程，因此它的行为类似于基于模板的方法。同时，由于两边的序列都是由LSTMs生成的，所以与模板方法相比，句子更加流畅和多样化。这使得CGO能够更好地处理新对象的描述生成任务。

在本文中，我们首先在MSCOCO上验证了我们的方法的特点，以生成图像中每个被检测对象的描述。通常在图像描述模型的标签中只提到图像中的一小部分对象。然而，使用CGO，我们可以将描述的能力扩展到被忽略的对象上，从而为图像提供一套更全面、更多样化的描述(如图1所示)。然后我们将CGO应用到新的对象描述任务中，展示了它在面对未知对象时的优势。我们在MSCOCO数据集上测试了我们提出的方法，并展示了为ImageNet[34]对象生成的描述。实验结果表明，该方法在METEOR[8], CIDEr[37], SPICE[2]、新对象F1得分等多个评价指标上均优于现有模型。生成的描述在内容准确性和流畅性方面都得到了改进。

# 相关研究

**图像描述**。在早期的图像描述研究中，通常使用基于模板的模型[20,11]或基于检索的模型[10]。基于模板的模型从给定图像中检测视觉信息，并将其填充到模板中生成句子。因此，描述通常缺乏多样性。基于检索的模型从现有的句子中找到最相似的句子，并且不能生成新的描述。在最近的工作中，深度神经网络的编解码结构得到了广泛的应用[40,17]。在[43,12,23]中，使用了注意机制，使得语言模型在每一步都注意到图像的不同区域。在[31,33,22,45]中，使用了强化学习算法来训练语言模型，使得不可微度量被用作训练目标。

**描述的多样性**。近几年来我们主要研究描述生成过程的可控性和描述的多样性[6, 15, 35, 39, 41, 46]。基于GAN的方法[6,35]和基于VAE的方法[15,41]用于提高描述的多样性和准确性。在[26]中，生成的句子可以包含不同主题的单词。[9]提出了一种以词性引导的图像描述生成方法。与CGO不同，这些方法不能精确地控制描述中包含我们所关心的对象。[7,6]研究了如何为图像生成描述性段落。[16]为图像中的每个语义信息区域生成描述。这些方法需要数据集中的附加标签，例如可视基因组[19]。CGO方法不需要额外的标签。这些描述是用CGO基于整个图像生成的，因此这些对象之间可能有更丰富的关系。

**描述新的对象**。新对象描述任务是由Hendricks等人首先提出的。提出的模型DCC要求描述没有参与模型训练的对象。在NOC[38]中，使用联合目标来训练对象分类器和语言模型。LSTM-C[44]在NLP中应用了复制机制来在生成过程中合并新单词。NBT和DNOC[24, 42]使用语言模型生成带有占位符的模板，然后用识别出来的对象信息填充占位符。与[5,38,44,24,42]不同的是，CGO中的语言模型并没有预测到新的对象，这使得精确地在句子中包含新的单词成为可能。CBS[1]通过在波束搜索过程中添加对象约束来确保生成的句子中包含特定的对象。与CBS不同的是，在CGO中解码时，新单词不参与概率的计算。在自然语言处理领域的一些研究工作中，[29]也使用了生成带约束词的句子的方法。

# 方法

给定一个图像，CGO能够将选择的引导对象合并到生成的句子中。在这个过程中，两个LSTMs被组合在一起，在引导对象的两边产生部分序列。我们使用LSTM- L表示LSTM生成序列左侧部分的序列，使用LSTM- R表示LSTM生成序列右侧部分的序列。CGO可以灵活地应用于其他现有的RNN语言模型。

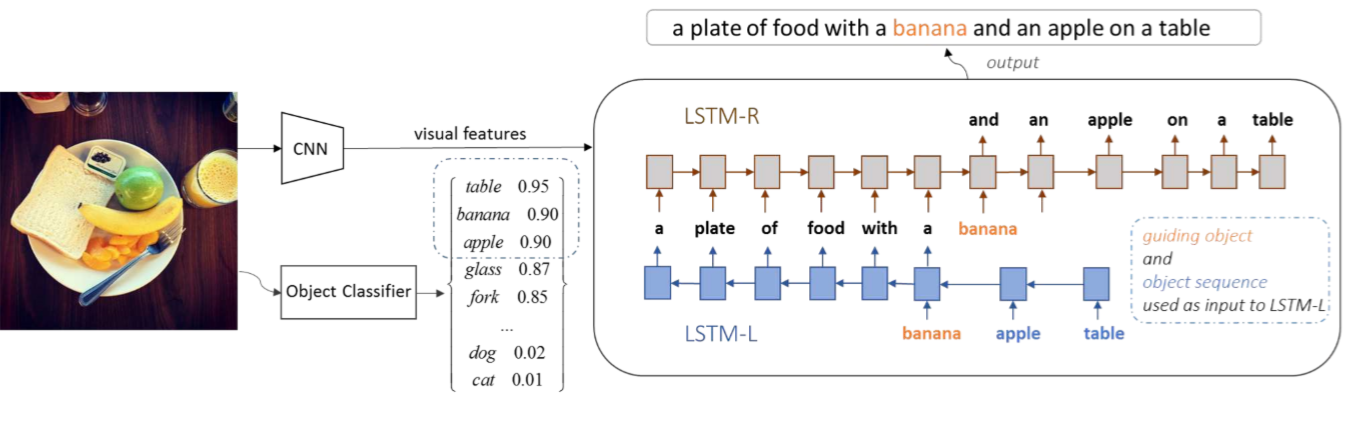


图2 CGO 方法

我们根据对象分类器的输出来选择引导对象和对象序列。对象序列用作LSTM-L的输入，提供关于假定的右侧序列的信息。LSTM-L根据视觉特征和输入对象序列生成左侧序列。然后将生成的左侧序列用作LSTM-R的输入，以生成右侧序列。然后将这两个部分序列与引导对象进行连接，得到完整的描述。

## 问题阐述

在常用的编解码模型中，卷积神经网络常用作编码器。来自CNN的图像信息的视觉特征被传递到语言模型。在解码过程中，通常使用LSTM之类的RNN网络作为语言模型。给定一个图像I,我们通常用一个序列y=(y1,y2,y3… yT)描述，用T表述序列的长度，yi代表模型词汇中的一个单词。词汇量为V。为编解码模型的参数，模型学习的目标就是找到最优的θ所以可以用公式表示为

(1)

其中 代表模型最优参数，代表标注值。当使用LSTM作为语言模型时，在每一个时间步t，它根据图像特征ft、这个时间步的输入单词xt和t-1时刻LSTM的隐藏状态ht-1来预测序列中下一个单词的概率。xt属于模型词汇表。 (2)

图像的特征ft在不同的模型设置下是不同的。在一些模型中，如NIC[40]，图像特征f只在第t=0时提供给语言模型。使用注意机制的模型会在每一步都用到图像的特征 t 如下

(3)

(4)

其中，表示第t步时的注意权重映射，⊙两个同样维度的向量/矩阵每一个元素分别相乘。计算注意权值的函数形式随注意机制的不同而不同。

如果我们希望生成的序列包含一个特定的单词，那么期望的序列就是

是特定的单词。此时，模型输出以图像I和单词为条件。模型参数被训为 (5)

其中 。和 是各部分的标签值。右边的部分序列可以是任意长度。我们将两个相反方向的LSTM结合在一起来完成两边的序列。

## LSTM-L

对于给定的图像I和单词，我们首先使用LSTM-L来生成左侧部分序列。在每一步t中，LSTM-L根据图像特征、输入单词和隐藏状态来预测前一个单词。

(6)

然而，在这个过程中存在一些问题。一个图像通常包含多个对象。这些物品可以按照不同的顺序排列。例如，“桌子上有一个苹果和一个香蕉”和“桌子上有一个香蕉和一个苹果”都是正确的描述。这两个句子可以同时出现在一个图像的标签值中。当仅提供 时，LSTM-L将不知道右侧部分序列(图3(a))。在实验中，我们发现模型往往会输出一个保守的结果，比如在这样的过程中“一个香蕉”。它通常语法正确，但缺乏变化。相反，在人类的描述中，各种对象往往会出现在左侧的部分序列中。

要描述的对象通常在我们说话之前就决定了。类似地，在图像描述中，我们可以在生成描述之前获得关于图像中对象的信息。因此，我们首先假设描述中将出现一组对象，并设置这些对象的排列顺序。然后我们可以得到一系列的对象标签，这些标签可能是在右边的序列中出现的。我们把标签序列表示为

S = 其中m是S中的对象数量，可以任意选择。 S中的对象不会出现在LSTM-L生成的序列中，但会影响序列中的内容(图3(b))。序列S作为LSTM-L的输入，在之前进行编码。LSTM-L现在根据图像I, 假设序列S和生成序列

（7）

与一般的生成过程类似，当预测的单词是结束标签<END>时，左侧序列完成，句子到达开头。在训练时，我们从一个描述标签中随机选择一个对象，然后从右侧的部分句子中提取S。将句子的左部分提供给LSTM-L作为标签值序列。对于给定的图像，我们最小化模型的交叉熵损失。

(8)

注意，损失只计算生成的左侧部分序列，即在t=k之前的步长的输出。



图3

(a)在引导对象(“橙子”)两侧独立生成的句子是不连贯的。

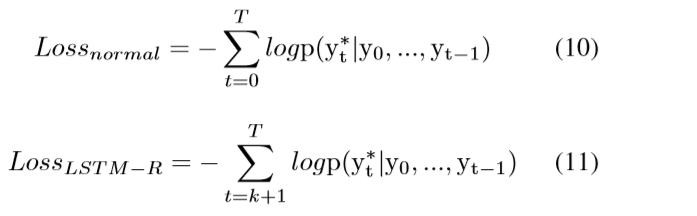
(b)将对象标签序列作为LSTML-L的输入，提供关于右侧序列的信息。当输入序列不同时，LSTM-L产生的左侧序列不同。

## LSTM-R

从LSTM-L中获得左侧部分序列后，LSTM-R将该序列作为输入，完成句子的其他部分。模型现在被训练成

(9)

实际上，我们不需要将表单中的描述标签处理为右侧部分序列。相反，我们可以简单地按照训练一个正常的LSTM的过程来生成从左到右的句子。给定图像和选定的训练损失是不同的。



其中和 代表连个训练过程中的损失函数。注意比 有更严格的限制。生成完整句子的过程可以看作是输入序列长度为零的特殊情况。另一方面，经过完整序列训练的LSTM使我们能够更灵活地使用模型。当图像中没有检测到任何对象，或者描述中没有要求包含任何对象时，我们可以使用LSTM-R作为正常的语言模型，从步长t = 0开始。在这种情况下，这个过程被简化为正常的过程，并从左到右生成句子。我们的方法可以应用于所有类型的RNN语言模型的图像描述。在推理过程中，可以使用各种解码方法，包括贪婪采样和波束搜索方法。

## 新词汇的嵌入

在编码过程中，将输入字 表示为one-hot向量，然后嵌入学习参数。在第t步，将嵌入向量作为语言模型的输入。当进行推理时，语言模型不会生成训练过程中没有出现的单词。在CGO方法中，当选择一个新的对象作为引导对象时，我们可以简单地使用与该对象相似的另一个被看到对象的嵌入向量。相似的对象可以根据WordNet[28]或word2vec[27]或GloVe[30]的嵌入向量之间的距离来选择。在正常的从左到右生成过程中，使用来自相似对象的嵌入向量不能强制语言模型生成新单词。然而，在CGO中，由于新单词是直接合并到生成的句子中，不需要经过解码过程，我们不需要语言模型来预测新单词。新单词只在编码过程中使用，相似对象的嵌入在编码过程中完成就足够了。

## 模型细节

**描述模型**。在我们的实验中，我们使用自下而上和自上而下的注意力模型(Up-Down)[3]作为我们的基本模型。LSTM-L和LSTM-R都是自上而下的模型。在我们的实验中，我们使用来自[4]的预先训练的模型特征。它是从建立在ResNet-101[13]上的一个更快的RCNN模型中提取的，并在MSCOCO和视觉基因组上进行了预训练。

**对象分类器**。现有的目标检测模型或目标分类器可以识别给定图像中的目标。在我们的实验中，我们沿用了之前的工作[5,1]，使用多标签分类器来确定一个物体是否出现在图像中。我们对MSCOCO对象检测数据集中的80个对象类别进行了分类。我们在分类器中使用与语言模型中相同的特性。

# 实验过程和结果

在本节中，我们将展示CGO将人为选择的对象合并到描述中的能力。在第4.1节和4.2节中，我们通过生成图像中每个被选中对象的描述来展示CGO的特性。在第4.2节中，我们展示了生成的描述的多样性。在第4.3和4.4小节中，我们将CGO方法应用于新的对象描述任务。

**数据集**。模型在包含123287张图像的MSCOCO数据集上进行训练和评估。有80个对象类别标记用于对象检测，每个图像都用5个人为生成的描述标记用于图像描述。我们按照前面的工作[12]对描述标签进行预处理，所有标签都转换为小写和标识符。出现次数少于5次的单词被过滤掉，其他的单词形成9487大小的词汇表。我们在4.1和4.2小节中使用了Karpathy的split[17]，这在图像描述研究中得到了广泛应用。113287张图片用于训练集，5000张图片用于验证集，5000张图片用于测试集。详情见第4.3节。在4.4小节中，我们在ILSVRC2012验证集上对模型进行了测试，该验证集包含1000个类，每个图像都标有其类别。

**训练过程。**在实验中，利用随机梯度下降法对目标分类器进行了优化。学习率设置为1e-4，每10个epoch衰减0.1。分类器训练了20个epoch。使用Adam[18]对语言模型进行了优化。学习率设置为1e-4，每20个epoch衰减10次。LSTM-L训练了80个epoch，LSTM-R训练了40个epoch。

## 描述每个选择的对象。

为了展示我们的方法的特点，我们为图像中选择的每个对象生成一个描述，以获得描述每个图像中不同对象的所有句子。我们根据对象分类器的输出，选择概率最大的k个对象作为引导对象。我们分别用k=1、3、5、10来测试模型。我们计算每个图像的描述集合中涉及的不同对象类别的平均数量。我们还计算每个对象类别的recall数。也就是说，图像中出现的对象是否在描述集中被提及。需要注意的是，一个物体是否出现在图像中取决于物体检测标签，因为描述标签通常只包含图像中出现的对象的一小部分。

基础模型和CGO都是在MSCOCO Karpathy的[17]上训练和测试的。图4的左栏为实例。在表1中，我们展示了使用beam search (beam size = b)作为基线模型生成的结果，以及使用CGO生成的结果。average of recalls是所有80个对象类别的宏观average recall。

基线模型的平均对象类别数和recall分别为1.98和0.66,b = 10。使用CGO，平均数量和recall分别提高到2.92和0.75 (k = 10)。我们还计算了描述标签的平均数量和recall。当k = 5时(每幅图像有5个描述标签)，CGO的对象recall为0.73，高于描述标签(0.61)，说明CGO可以描述caption标签中被忽略的对象。注意，虽然基线模型可以用较大的波束尺寸描述更多的对象类别，但是它不能控制在过程中描述哪些对象。

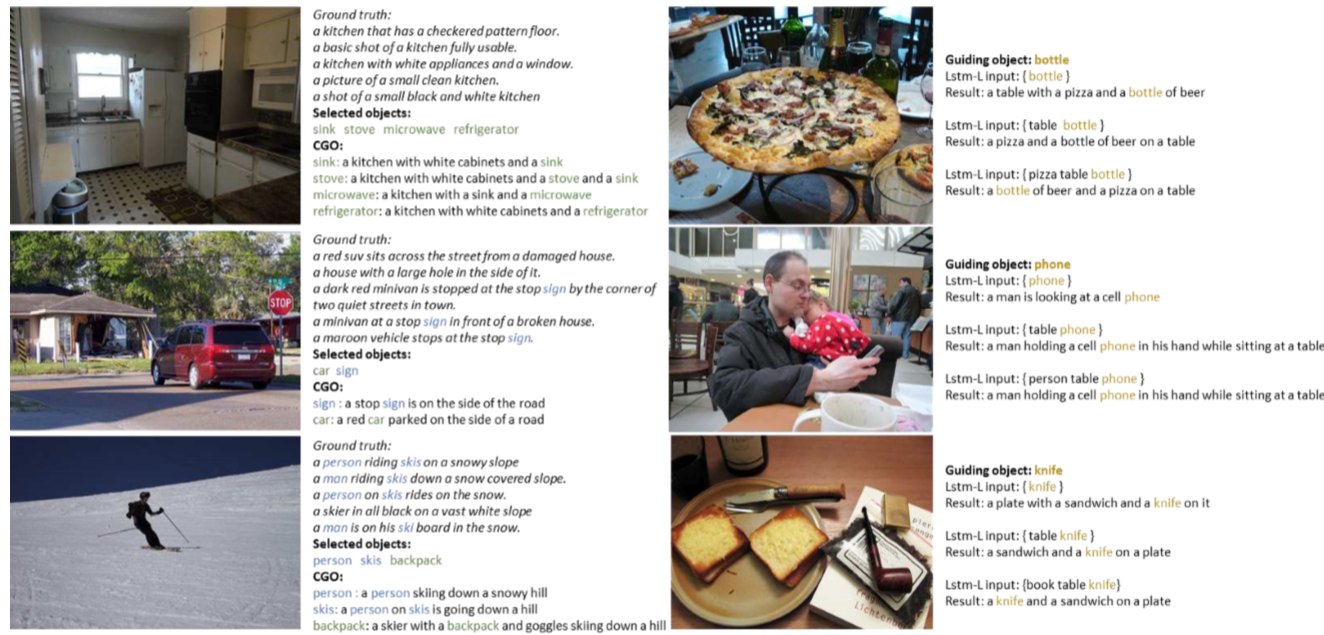


图 4

所选对象的描述示例显示在左侧列中。CGO结果中每一行开头的对象用于该描述的指导词。标签值中包含的对象是蓝色的，其他的是绿色的。一个引导对象的不同描述的例子显示在右栏。

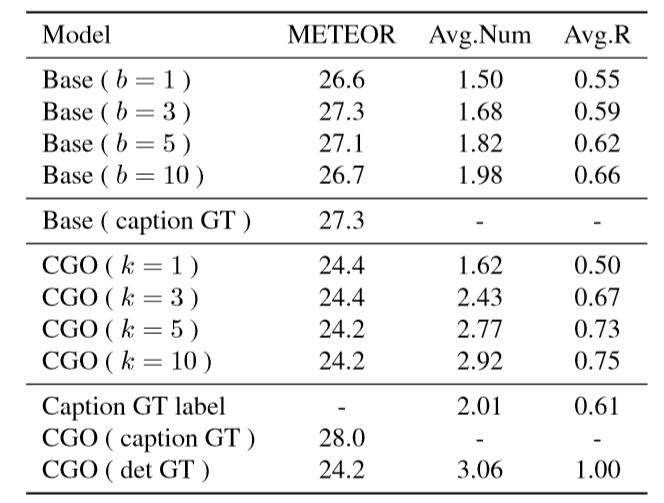


表1

Base表示用作基线的基础模型。b表示使用顶级b波束搜索生成。对于CGO，我们使用对象分类器预测的最前面的k个对象作为指导对象。“Caption GT label”是数据标签的统计结果。Base(CaptionGT)显示至少有一个标签会出现在描述标签中。CGO(Caption GT)显示了引导对象出现在描述标签中的分数。CGO(det GT)显示了我们为图像中每个对象都生成描述的结果。Avg.Num表示图像描述中对象类别的平均数量。AVG.R代表average recall。

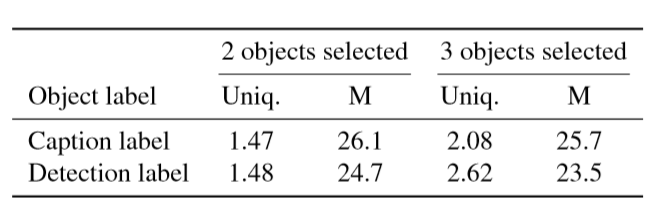


表 2

从描述标签和目标检测标签中选择引导对象的结果。M为METEOR分数。Uniq.表示每个选定导向对象的唯一描述的平均数量

CGO的METEOR评分在24.2左右，低于基础模型(26.7)。评价方法仅将结果与给定描述签进行比较。即使图像中出现的对象描述正确，如果这些对象没有出现在给定的描述标签中，那么得分也会很低。虽然不能用这一分数来精确地评价几总体的流利程度，但它为我们提供了一个粗略的参考。我们也评估了那些出现在给定标签上的引导对象的描述，METEOR得分为28.0。这表明，当引导对象位于标题标签的域内时，生成的句子是流畅的。

图5显示了7个对象的recall作为示例。与基线模型和给定描述标签相比，CGO可以显著提高不显眼对象例如“cup”(从0.15到0.69)和“bowl”(从0.21到0.65)的recall。

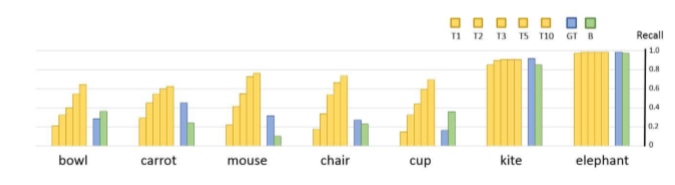


图 5

单个对象类别的recall示例。T1∼T10表示不同选定对象数目下CGO的结果。GT表示ground truth描述标签的统计结果。B为基础模型的结果。使用Adam[18]对guage模型进行优化。学习率设置为1e-4，每20个epoch衰减10次。LSTM-L训练了80个epoch，LSTM-R训练了40个epoch。

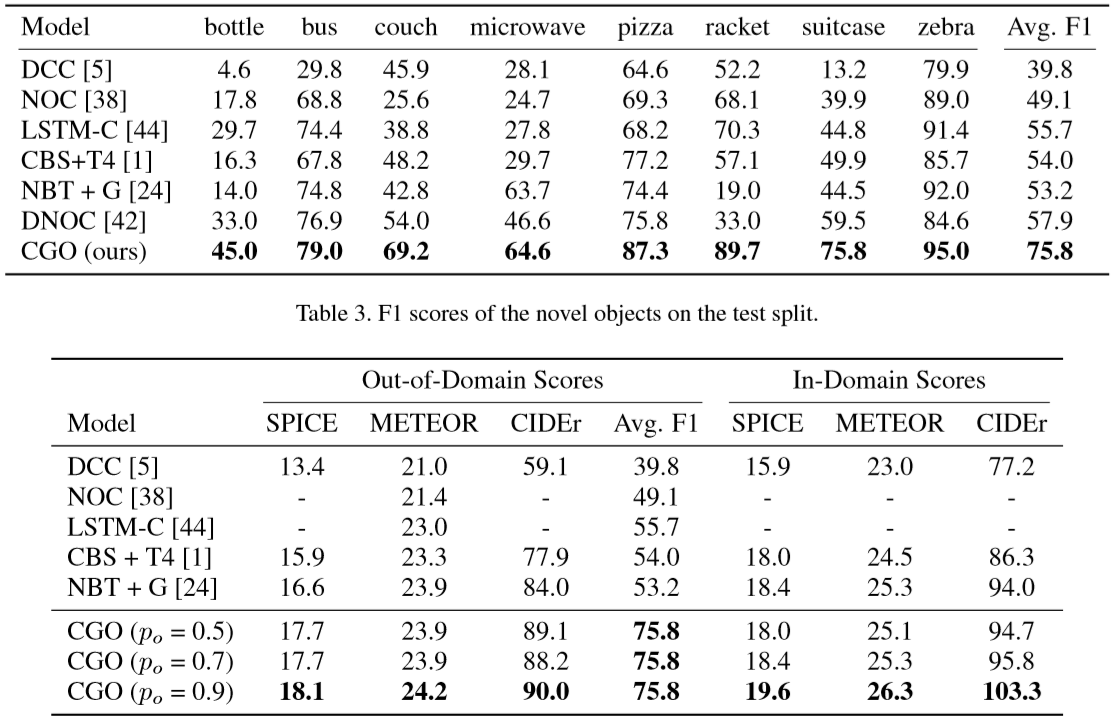


表 4

使用image caption metrics来评估域内和域外图像生成的描述。是选择域内引导对象的阈值。当对象分类器预测的对象出现的概率超过时，将其作为指导对象。当多个对象满足要求时，我们选择概率最大的对象。

## 每个对象生成多样化描述

在这一部分中，我们展示了CGO使用选定的引导对象生成不同描述的能力。我们从图像中随机选择一个对象作为引导对象，然后选择n=1或2个其他对象组成LSTM-L输入序列。当n=1时，输入序列可以是<guide object>或<guide object, Object1>，‘Object1’表示为LSTM-L输入序列选择的对象。n = 2时，我们用长度为1、2和3的3个不同的输入序列进行测试。

结果如表2所示。当我们使用从对象检测标签中选择的对象时，在2种不同的输入情况下，不同描述的平均数量是1.48，在3种不同的输入情况下，不同描述的平均数量是2.62。这说明即使有选定的引导对象，CGO也可以产生不同的描述。图4的右栏展示了示例。

## 新对象的描述

在这一部分中，我们演示了CGO在新对象描述任务中的有效性。在[5]之后，选择了“巴士、瓶子、披萨、微波炉、沙发、手提箱、球拍、斑马”8个对象类别作为新奇对象。在训练时，如果图像的描述标签包含新对象，则将其排除在MSCOCO训练集之外。MSCOCO验证集的一半用作验证集，另外一半作为测试集。用F1分数来评估包含新对象的准确性。对于每个新的对象类别，如果生成的描述和标签同时包含该对象，则视为正确。F1的平均分数是8个类别的宏观平均分数。image caption metrics用于评价生成的句子质量，包括SPICE[2]、METEOR[8]和CIDEr[37]。对域外图像(包含一个新对象)和域内图像的描述分别进行评估。

与之前的工作类似[5,1]，对象分类器的标签是来自描述标签的。训练分类器时使用全部数据集，包括包含新对象的图像。如果一个新对象出现在图像中，它将被用作引导对象。我们根据对象分类器的结果来确定一个新的对象是否出现在图像中。选择使验证集F1分数最大的对象作为引导对象。对于新词汇，我们简单地将其单词嵌入向量，替换为相同超类别下的其他域内对象，如“bottle”→“cup”。结果如表3和表4所示。描述的例子如图6所示。

与现有模型相比，CGO显著提高了新对象的F1分，平均F1分为75.8分。实际上，输出的F1分数直接取决于分类结果的准确性，就像基于模板的模型一样。注意，在语言模型中使用不同的RNN模型或不同的CNN特征并不影响F1结果。另一方面，CGO利用LSTM语言模型生成流畅的描述。METEOR评分提高到24.2的域外的图像和26.3的域内的图像。我们用不同的概率阈值来测试主要目标的CGO。当目标分类器预测的概率超过阈值时，使用域内对象作为引导对象。当分类器对图像中包含的对象不确定时，生成过程简化为通常的从左到右的生成过程。

由于对象分类器与语言模型是独立的，使用更高级的模型如对象检测模型可能会进一步提高F1分数。在CGO方法中，我们只保证一个被选中的对象被提及，但这并不影响它的实用性。在很多情况下，新奇的词汇并不多见，在生活中我们可以对一个图像使用多个描述。此外，可以将CGO与其他方法(如CBS[1])结合使用，以在输出中包含更多对象，同时确保在描述中包含引导对象。

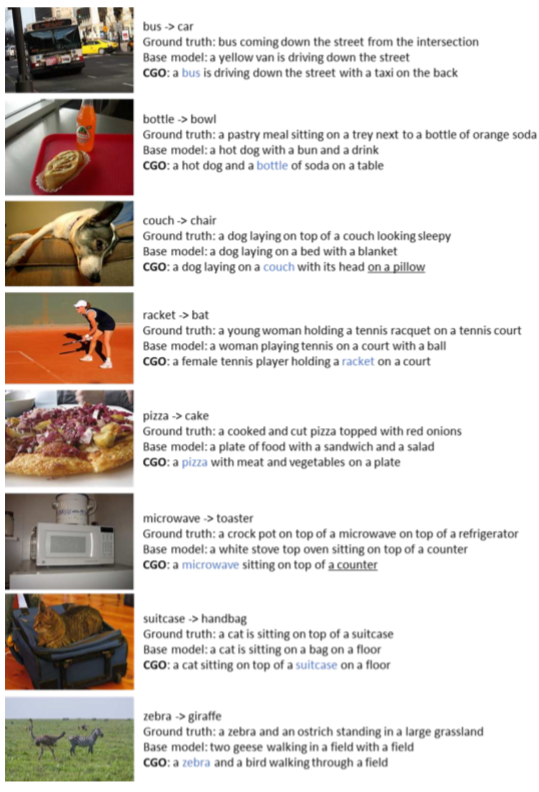


图 6

为域外对象生成的描述示例(蓝色)。O1→O2表示我们使用O1作为引导对象，O1是使用域内对象O2的词向量进行编码的。突显出错误。

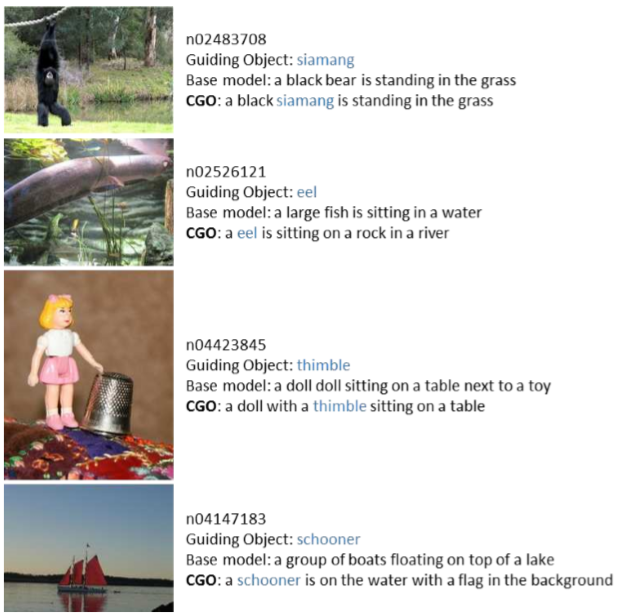


图 7

为ImageNet对象生成标题的示例(蓝色)。使用CGO，一个对象可以被包含在描述中，即使它没有出现在模型词汇表中。

## 描述ImageNet对象

与之前的工作类似[44,38]，我们展示了描述ImageNet[34]对象的方法的结果。从MSCOCO描述标签选择词表中没有出现的对象对于在MSCOCO模型训练是新颖的。我们使用Karpathy的分割模型来生成描述。示例如图7所示，更多的结果示例可在附录中找到。与4.3小节的过程类似，将新的对象用相似的现有对象替换并嵌入词向量中。例如“帆船”→“船”。

# 结论

我们提出了一种新的图像描述方法，句子生成过程从一个选定的引导对象开始。我们的CGO允许我们在生成的句子中包含特定的对象，并以全面和多样的方式描述图像。

# 致谢

项目获得国家自然科学基金第61701277、61771288号拨款和国家“十三五”重点发展项目第2016YFB0801301号拨款支持。

# References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. arXiv preprint arXiv:1612.00576, 2016.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision, pages 382–398. Springer, 2016.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6077–6086, 2018.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In CVPR, 2018.

[5] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–10, 2016.

[6] Dai Bo, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In IEEE International Conference on Computer Vision, 2017.

[7] Moitreya Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In Proceedings of the European Conference on Computer Vision (ECCV), pages 729–744, 2018.

[8] Michael Denkowski and Alon Lavie. Meteor universal: Language speciﬁc translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation, pages 376–380, 2014.

[9] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander Schwing, and David A Forsyth. Diverse and controllable image captioning with part-of-speech guidance. arXiv preprint arXiv:1805.12589, 2018.

[10] Jacob Devlin, Cheng Hao, Fang Hao, Saurabh Gupta, Deng Li, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. Computer Science, 2015.

[11] Desmond Elliott and Arjen de Vries. Describing images using inferred visual dependency representations. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 42–52, 2015.

[12] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Doll´ar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1473–1482, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[14] Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[15] Unnat Jain, Ziyu Zhang, and Alexander G Schwing. Creativity: Generating diverse questions using variational autoencoders. In CVPR, pages 5415–5424, 2017.

[16] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4565–4574, 2016.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137, 2015.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1):32–73, 2017.

[20] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12):2891–2903, 2013.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[22] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In Proc. IEEE Int. Conf. Comp. Vis, volume 3, page 3, 2017. [23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 6, page 2, 2017.

[24] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7219– 7228, 2018.

[25] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632, 2014.

[26] Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. Show and tell more: Topic-oriented multi-sentence image captioning. In IJCAI, pages 4258–4264, 2018.