# 从描述到视觉概念的召回[[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

本文提出了一种自动的新方法生成图像描述：视觉检测器，语言模型和多模态相似模型从图片说明的数据集进行直接的学习。使用多实例学习训练检测测器获得语句重的不同部分，例如名词，动词，形容词。词探测器输出作为一个最大熵语言模型的输入。语言模型捕捉400000个图像描述中捕捉单词使用的统计信息。通过使用语句级特征进行重新排序获得全局的语义信息。在COCO上取得了最好的BLEU-4分数。

## 2 介绍

如何让计算机进行理解图片，是通过对图像最突出的内容进行总结获得的。这些内容也许包括目标，属性，关系。同时通过这些内容推论出一些场景知识。

共分为下面三个阶段：

通常获得视觉探测器是通过孤立的训练集进行学习，这里直接使用图像描述数据集进行训练，优势在于可以学习那些突出的内容，而不是背景。一些概念性的词语很难学习与一些视觉模式相关。

训练一个语言模型获得场景知识。例如学习一个人更可能是坐在椅子上而不是站在上面。

学习一个联合的多模态表示在图像和描述上以测量图像和文本的相似度，选择更有可能的描述。

首先作者使用弱监督学习获得探测器。直接从图像描述中进行学习有很多困难，系统难以获得监督信号，例如目标框很难被表明，或是没有，例如人群，内部。

1.使用整幅图像子区域进行推理

2.使用微调的丰富的CNN特征

3。将每个区域的特征映射到包含句子的描述中。使用了多实例学习的方法获得每个词的直接视觉特征。将每个区域的特征映射到可能相关的句子描述中。

产生句子从一系列可能的词包中需要高效的语言模型。这里用优化的方法解决这一问题。核心任务是找到具有最大得分的单词，并找到最大可能性覆盖这一单词的句子。这里使最大熵的模型对可能的句子序列进行显式搜索。这种训练会捕捉语言统计上的常识知识。

最后对候选句子根据句子特征进行线性权重的重排序。通过最小错误率的方式进行训练。通过一个基于多模态的相似性模型中的特征进行学习。这种模型学习了两个神经网络去匹配句子和图像之间的相似性。这种多模态相似性会显著提升句子的质量。

作者使用了COCO数据集，评估指标使用了BLEU、困惑度和METEOR，这里超过了人类标注的BLEU-4得分。

## 3 相关工作

这里认为有两种方法获得图像描述，一是对已有图像描述的检索，二是产生新颖的描述。这里主要介绍的是多模态空间的检索，也是主流。

这里与两种生成式模型进行了比较，并且这两种方法使用相同的测试例子，使比较成为可能。

这里作者也提到了CNN和RNN的策略，这里作者采取了不同的方法，使用CNN应用于图像区域获得单词，然后使用MIL将信息进行整合。然后通过训练先验假设最小化直接产生句子（解一个优化问题）。

### 4 词检测

## 4 词探测

第一步使产生可能在图像描述中存在的一系列单词，包括名词，动词和形容词。使用了覆盖训练数据的92%的词汇。

### 4.1 训练词检测器

给定词汇表后，这里强调我们不能使用标准的监督学习技术作为学习探测器，因为不知道对应单词的边界框。同时许多单词关联的边界框没有被定义，例如open和beautiful。一种潜在的方案是使用整幅图像的特征作为解决的途径。这里使用这种方法取得的效果并不好，代替的使用了多实例学习的方案。

对于每个属于词表的单词，MIL采取了正例包和反例包的边界框的集合包，每个包对应一幅图像i。如果一个包是正例的，那么对应的单词w在图像i的描述中，反例则不在。直观地，MIL通过迭代选择正包内的实例来执行训练，然后使用更新的正确标签重新训练检测器。

对于各个区域包的包含单词的概率是通过包内单个实例的概率计算获得的。



Pijw是图片i中给定区域j对应单词j的概率。我们计算pijw用一个多层架构获得这个概率（在fc7（通用特征）上的fc8层计算一个sigmoid的层）



全连接接一个sigmoid的二分类分类器。

这里将全连接解释为全卷积结构，获得全卷积结构，然后获得一个粗糙的空间响应图，该响应图的每个位置对应原始应用于CNN图像的重叠移位区域而获得的响应。

我们对图像进行上采样使长边是565个像素，获得一个12\*12的响应图在fc8，对应上采样的图像是一个224\*224的bounding box，位移是32.对于上面的MIL公式使用来产生一个piw针对这幅图像对于每个词。使用交叉熵优化CNN的学习，每个batch一幅图像，训练了三个epoch。初始化使用ImageNet的的网络架构。

总结一下卷积网络特征图对应的区域形成一个包，一个包对应一幅图像，然后对应许多单词，单词属于句子，则是正例，否则反例，训练出神经网络针对包对应每个单词的概率获得单词。这样可以学的那些正例图像的共同区域并获得最大值。

### 4.2 产生词分数对于训练图像

对新的测试图像i，我们上采样和进行前向传播通过CNN获得piw。对于词典中的每一个单词都是独立检测的，这样就需要给出一个分数，因此输出需要进行重新矫正。这里对于准确值设置了全局的阈值t，输出所有在图像级别概率piw大于阈值的单词表里的单词和原始的最大分数pijw（针对特征图上特征点对应的区域，获得相应的分数）。这种方法不需要任何边界框标注，但是能够产生稳定的目标定位和将图像区域与更抽象的概念联系起来。（这里的难点是将抽象的概念与图像中特定的子区域对应起来，但是没有合适的标记框，因此使用了一个框集合的多实例方法，获得一个大致的概念对应框）。

## 5 语言产生

我们将生成过程作为搜索由视觉检测到的单词集合的最可能的句子进行搜索。语言模型是这个过程的核心，因为它定义了单词序列的概率分布。请注意，尽管是一个统计模型，但LM可以编码非常有意义的信息，例如在马匹后面更容易跟随着跑步而不是马匹。这些信息可以帮助识别假字检测并编码一种常识性知识。

### 5.1 统计学模型

为了产生候选描述，我们使用了一个最大熵的方法基于检测到的单词。这一模型估计词wl在w1，w2，…,wl-1为条件，并且这个具有高可能性的集合未在句子中被提及。

对未使用的词进行限制的动机是鼓励所有的词被使用，同时避免重复。前15个最常出现的闭集合被从集合V中移除以避免那些在每个图像中都被检测到的单词。这里的单词会受噪声影响。因此，当正在预测句子标记的结束时，剩余词的集合可能仍然包含具有高检测可信度的一些词。

在上述检测方法，词的条件概率可以定义为：



这里s表示句子的开始，wj属于词典V并上s，fk和拉姆达分别代表第k个最大熵的特征和它的权重。这里总结了相关的特征。这些特征来自与我们的基准系统。已经被证明分数特征特别高效，它会根据对应视觉探测器获得log似然性。这里作者探索了bigram特征和连续的log双向特征，他们对BLEU，METOR或是其他指标都没有显著的提升。

训练模型时，目标是以在一组检测的行为条件下的log似然性：



S表示句子的索引，#s表示句子的长度。噪声对比度假设（NCE）技术用来加速训练。在生成过程中，我们使用非标准化的NCE似然估计，它比确切的可能性更有效，并且产生非常相似的输出。 然而，我们报告的所有PPLX数据都是通过详尽的归一化计算得出的。 ME特性在[31]中的哈希表中实现。在我们的实验中，我们在NCE训练中使用N-gram特征高达4个和15个对比样本。

### 4.2 产生模型

在产生过程中，作者使用了一个从左到右的流搜索。这里保持一个局部假设的长度为l的堆栈。在搜索的每一步中，堆栈中的每条路径都会扩展一组可能的单词，并存储所得到的长度为l + 1的路径。 顶部k长度l + 1路径被保留，其他路径被剪掉。...

将可能的扩展标记为=s，包括100个最常用的单词，还有待提及的属性单词集以及训练数据中观察到的最后一个单词 在假设中的词。修剪基于部分路径的可能性。 当生成<= s>时，<= s>的完整路径将从堆栈中移除并作为完整句子放置。 该过程继续，直到达到最大句子长度L.

在获得完成句子集合C之后，我们形成如下M-最佳列表。 给定要提及的T图像属性的目标数量，将覆盖至少T个对象的C中的序列添加到M-最佳列表，按降序排列对数似然。 如果少于M个序列覆盖了在C中找到的至少T个对象，我们将T减1，直到找到M个序列。

## 6 句子重排序

前面的模型产生M个最优句子集合。最后用MERT对M个句子进行重排序。

MERT使用在整个句子上计算的特征的线性组合，如表2所示.MERT模型使用BLEU度量标准在验证集的M最佳列表上训练，并应用于测试集的M最佳列表。 最后，选择重新排序后的最佳序列作为图像的标题。 除了标准的MERT特性外，我们还介绍了一种新的多模态语义相似度模型，下面将对此进行讨论。

### 6.1 深度多模态相似性模型

这里通过DMSM对图像和文本之间的相似性进行建模。这里使用两个神经网络来学习将图像和文本片段映射到共同的向量表示空间中。通过测量图像和文本之间的相似性是它们对应向量的余弦相似度。这种余弦相似度被默认用于重新排列的句子。DMSM与单模态深度结构化语义模型（DSSM）密切相关。最初提出的DSSM是为了模拟文本搜索查询和文档之间的语义相关性，并且扩展了这项工作以用深卷积网络计算的图像矢量代替原始DSSM中的查询矢量。

DMSM由一对神经网络组成，其中一个用于将每个输入模态映射到共同的语义空间，这些空间被共同训练。 在训练中，数据由一组图像/标题对组成。 在训练期间最小化的损失函数表示给定相应图像的标题的负对数后验概率。

图像模型：使用与学习词包相同的CNN网络获得图像模型。这里提取了fc7的表示然后附加了3个全连接层带有tanh层获得一个最终的与文本模型大小一致的向量表示。然后在DMSM中学习附加的全连接层。

文本模型：DMSM的文本部分按照与原始DSSM相同的方式将文本片段映射为语义向量。 通常，文本片段可以是完整的标题。 在[16]之后，我们将标题中的每个单词转换为一个字母 - 三字符计数矢量，该矢量使用与上下文有关的字母的计数分布来表示一个单词。该表示具有减少输入层的大小同时泛化到一些罕见，不可见和不正确的拼写单词。 然后[41]，这种表示通过深层卷积神经网络向前传播，以产生最后一层的语义向量。

目标和训练：我们根据图像和文本模型获得的表示yQ和yD，将相关性R定义为图像或查询（Q）与文本片段或文档（D）之间的余弦相似度：。 对于给定的图像文本对，我们可以计算文本与图像相关的后验概率



这是使用验证集确定的平滑因子，在我们的实验中为10。 D表示应该与查询（图像）进行比较的所有候选文档（标题）的集合。 我们发现将D限制为一个匹配文档D +和固定数目N的随机选择的不匹配文档D􀀀工作得相当好，尽管使用噪声对比估计可以进一步改善结果。因此对于每幅图像我们可以选择一个相关文本片段和N个不相关片段计算后验概率。N设置为50.通过最小化后验概率训练模型参数。



## 7 实验结果

### 7.1 数据集

COCO数据集和PASCAL 数据集。

### 7.2 词检测

多实例学习将如果一个词包含在至少一个实际标注的句子中则视为正例。这里的挑战在于一些概念上相似的词语被孤立的考虑了。尝试使用引理词语没有获得显著的性能提高。

这里使用描述的不同部分的AP和准确率对各方法进行了评估。这里有两个基准，一是对词语随机分类，二是使用AlexNet或是VGG对全局特征进行分类。

如表3所示，MIL方法提高了语句各部分的基准，表现出了更好的的定位性能。对名词和形容词获得了一个巨大的提升，这通常对应图像中一些子区域中的具体对象。对于部分缺少视觉信息的单词准确率较低，如一些形容词，代词，介词。相比之下，高的AP分数是典型的视觉信息（红色：AP 66.4，她：AP 45.6）或与特定物体相关

### 7.3 描述生成

度量标准：句子生成过程使用自动度量和人类研究进行度量。 我们使用三种不同的自动指标：PPLX，BLEU [37]和METEOR [1]。 PPLX（困惑）测量语言模型的不确定性，对应于在给定语言模型的情况下对每个单词进行编码需要多少比特。 较低的PPLX表示较好的分数。 BLEU [37]广泛用于机器翻译，并测量假设与参考或参考文献之间共同的N-gram（最多4克）的比例; 我们在这里比较4个随机选择的参考。 METEOR [1]测量单精度和回忆，扩展精确的单词匹配，包括基于WordNet同义词和词干标记的相似词。 我们另外报告了从MSCOCO字幕挑战中获得的指标的绩效，其中包括BLEU-1到BLEU-4，METEOR，CIDEr [44]和ROUGE-L [27]的评分。

已知所有这些自动度量只与人类判断大致相关[10]。 因此，我们包括人类评估，以进一步探索我们的模型的质量。每个任务提供一个人（机械特克员工）的图像和两个标题：一个是自动生成的，另一个是人类标题。 要求人选择哪个标题更好地描述图像，或者在质量相同时选择“相同”选项。



结果显示这里的处理很先进。这里一方面可以看出finetune的提升。Ft代表fine tuen。Score代表加入word detector的特征。基线是使用AlexNet和所有表一中的特征通果MELM产生句子。加入score对BLEU和METOR有一个明显的提升。加入BLEU和DMSM后BLEU分数和METOR明显上升。

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